

Table of Content

[Table of Content](#)

[Chapter 1 - SQL](#)

[Chapter 2 - Python](#)

[Chapter 3 - Pandas](#)

[Chapter 4 - Numpy](#)

[Chapter 5 - Case Study and Guesstimate](#)

[Chapter 6 Linear Regression](#)

[Chapter 7 - Logistic Regression](#)

[Chapter 8 - NLP](#)

[Chapter 9 - Decision Tree and Random Forest](#)

[Chapter 10 - Random Forest](#)

[Chapter 11 - K-means](#)

[Chapter 12 - KNN](#)

[Chapter 13 - Power BI](#)

[Chapter 14 - Forecasting](#)

[Chapter 15 - Data Preprocessing](#)

[Chapter 16 - Statistics](#)

[Chapter 17 - MS Excel](#)

[Chapter 18 - Big Data Technologies](#)

[What is Big Data, and where does it come from? How does it work?](#)

[Chapter 19 - Amazon Web Services](#)

[Chapter 20 - Regular Expression](#)

[Chapter 21 - CLTV](#)

[Chapter 22 - Shell Scripting](#)

[Chapter - 23 Machine Learning Topics](#)

Chapter 1 - SQL

We will start with SQL, the bread and butter of anyone in the analytics domain. Now, we will try to crack the concepts of SQL in the next 200 questions.

How to use the book to be more efficient in the interviews?

Suppose we explain to you the concepts of ranking in SQL, then try to solve a few questions on the same topic and try to look for application of the concept. In total there are only a few concepts in any language or technology, the rest is the application of the basics. We will try to cover as much as possible and will make sure that you reach such a level that you can apply the learned concepts.

P.S. - If a question is getting repeated then believe that the concept is important and that you need to again learn/solve it.

One last thing before we start is that we will focus very little on the definitions and more on the type of questions asked in the interview and the ways to tackle these questions. Pardon the spelling or grammatical mistakes. And we do assume that you have a basic knowledge of the concepts discussed here, wherever necessary we will put some links which you can use to complete the basics of the topics

In SQL we will mostly focus on the topics below:-

- Basic syntaxes
- Group by and order by
- Aggregate functions
- case when statements
- Wild Cards
- Subqueries
- Rank
- Dense Rank
- Row number
- Joins
- Unnesting a column
- Working with JSON data types
- Working with different date formats
- Working with Array data type
- Regular expression in SQL
- Optimization of SQL queries on big data
- Partitions
- Create and insert a statement
- CTE
- SQL vs no SQL DB

-direct interview questions asked in an analytics interview

In the first few questions, we will try to cover the base of the topic so that a fresher or less experienced person can start with the topic. Try not to skip any questions.

1. What is SQL?

A. SQL stands for Structured Query Language and it is used to derive results from a table present in a database. For example, suppose you are the CEO of Flipkart.com and you have 3 tables in your database that goes by the name of inventory, supply, and demand.

Now, SQL can help you in understanding the total number of orders taken by the application, total number of buyers, number of canceled orders, number of people churning out of the website, and many other simples as well as complex metrics. All this information is derived from the respective table by an analyst using a query language on the tables and databases maintained by the team.

2. Who maintains the database, queries, website, frontend, backend, and other services in an application?

This question is remotely related to SQL, but it's very important to understand your responsibilities and the people around you.

Frontend Team - The front end i.e. the UI and UX part is mostly looked after by the web developers and front-end developers

Backend Team - The backend of how the input from the user will be stored in tables and databases is maintained by the backend team

Data Engineers - They pull the raw data stored in some cloud platform and sync it in your system in a very structured and tabular format

Analysts - They take the data from these tables as columns and provide business solutions to the clients

Business Intelligence Engineer - It is a combination of the data engineer and analysts

Data Scientists - They take up the data, clean it, test different machine learning models, automation, etc., and then deploy the model on the application. The role is in between a Data Engineer and an analyst

3. What is a database?

A database is an organized collection of data, stored and retrieved digitally from a remote or local computer system. Databases can be vast and complex, and such databases are developed using fixed design and modeling approaches.

In general, there are multiple tables in a particular database

4. What is DBMS?

DBMS stands for Database Management System and is used to identify, manage and create a database that provides administered access to data.

Data is stored as a file in DBMS, there is no connection between the data but the data is normalized. It does not support distributed system. In general, a very small quantity of data is stored in DBMS.

5. What is RDBMS?

RDBMS stands for Relational Data Base Management System, In DBMS, the data is stored as a file, whereas in RDBMS, data is stored in the form of tables. In RDBMS, data normalization is not possible, it deals with a high volume of data. In RDBMS, the redundancy of data is reduced by using indexes and keys.

6. What is the difference between DBMS and RDBMS?

We have already gone through the definitions of both DBMS and RDBMS. Let's look into the specific differences between the two:

DBMS	RDBMS
Data stored in file format	Data stored in table format
No connection between the database	Tables are linked together
No support for distributed data	Distributed data is supported
Normalization is practiced	Normalization is not common in practice
Data redundancy is common	Redundancy of data is reduced
Data stored in small quantity	Data stored in huge volume
Ex. XML, Microsoft Access	Ex. Oracle, SQL Server, Amazon Redshift

7. What are the examples of DBMS and RDBMS?

DBMS - XML, Microsoft Access, etc.

RDBMS - Oracle, SQL Server, etc.

8. What is the difference between SQL and MySQL?

SQL is a standard language for retrieving and manipulating structured databases. On the

contrary, MySQL is a relational database management system, like SQL Server, Oracle, or IBM DB2, that is used to manage SQL databases.

9. What are Tables and Fields?

A table is an organized collection of data stored in the form of rows and columns. Columns can be categorized as vertical and rows as horizontal. The columns in a table are called fields while the rows can be referred to as records.

10. What is the basic syntax of SQL?

Basically below are the important keywords that are used to write SQL code:-

```
SELECT  
FROM  
JOIN  
WHERE  
ORDER BY  
GROUP BY  
HAVING  
LIMIT
```

The rest of the functions and keywords will be discussed as when we come across the concepts

11. Write any query to show how the code is written in SQL?

Assuming some intuitive names of tables and columns, here is a query to find out the number of products sold to date on product name at Amazon.

There are two tables i.e.

orders [order_id, guest_id, price, status, product_id]
Product [product_id, department, product_name]

The names written inside the square bracket are the names of the columns.

Now let's write the query, don't worry if you do not understand the code below(which is highly unlikely but definitely acceptable)

```
SELECT pro.product_name, COUNT(DISTINCT order_id) as number_of_orders  
FROM amazon.orders ord  
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)  
WHERE ord.status = 'paid'  
GROUP BY pro.product_name
```

ORDER BY 2 DESC

Now let's see we have two tables that have only one common field i.e. product_id and it is joined on the same. We put a condition of status = 'paid' so that we can filter out things of our interest. COUNT takes the count of each instance where the order_id is distinct i.e. UNIQUE. ORDER BY is used to make sure that the output is sorted in a descending order with the maximum sold product_name at the top.

Now we will try to cover some basics.

12. What are the Constraints in SQL?

Constraints are used to specify the rules concerning data in the table. It can be applied for single or multiple fields in an SQL table during the creation of the table or after creating using the ALTER TABLE command. The constraints are:

NOT NULL - Restricts NULL value from being inserted into a column.

CHECK - Verifies that all values in a field satisfy a condition.

DEFAULT - Automatically assigns a default value if no value has been specified for the field.

UNIQUE - Ensures unique values are inserted into the field.

INDEX - Indexes a field providing faster retrieval of records.

PRIMARY KEY - Uniquely identifies each record in a table.

FOREIGN KEY - Ensures referential integrity for a record in another table.

13. What is the SELECT command?

The SELECT command is used to specify the columns which you want to keep in the final output. There are two ways to write the column names:-

- Directly using the column's name without specifying the table name

Example

```
SELECT user_id, product_id  
FROM amazon.orders  
WHERE status = 'unpaid'
```

Here we have directly used the column names because there is only one table and each table will have unique column names

-Now suppose you join two tables i.e. orders and product, and we know that each of these tables has a column named product_id, so if you write code like the one below

```
SELECT product_id, COUNT(DISTINCT order_id) as number_of_orders  
FROM amazon.orders ord  
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)  
WHERE ord.status = 'paid'  
GROUP BY product_id  
ORDER BY 2 DESC
```

In the above example, the query engine will throw an error like ‘Ambiguous column name’ , this is because the column is present in both the tables and the query engine will not be able to decide where to pull the data from. In this case, we specify the table name before the column name

```
SELECT pro.product_id, COUNT(DISTINCT order_id) as number_of_orders  
FROM amazon.orders ord  
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)  
WHERE ord.status = 'paid'  
GROUP BY pro.product_id  
ORDER BY 2 DESC
```

Bonus tip - When you are working in an organization then it is expected that you specify the table name for each of the columns as a good coding practice

14. Explain the use of the FROM clause?

We are starting with the very basics because there are many people who start from the very basics and need a lot of guidance. We want to cater to all the audience at once.

FROM basically specifies the name of the database and tables from where the data is to be pulled. If there are multiple tables from where you need to pull the data then your FROM command will be followed by JOIN statement and there are many types of JOINs which we will discuss in the upcoming questions.

A very simple example of the FROM statement is given below:-

```
SELECT *  
FROM amazon.order  
Limit 20
```

Here * means all the columns of the table and LIMIT means the top 20 rows of the table will be extracted.

15. What and how is * used in SQL?

* is mostly used in the SELECT statement and it has two use cases:-

- It is used to get all the columns from the table
- It might also lead to error, suppose you have the following statement now tell me whether the below query will throw an error?

```
SELECT *
FROM amazon.orders ord
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)
WHERE ord.status = 'paid'
```

It will surely throw an error because it will fetch all the columns from both the tables and it will find two columns with the same name i.e. product_id, thus you need to keep an alias of this column

Example to fix this issue

```
SELECT pro.product_id as product_id_1, order_id, guest_id, product_name
FROM amazon.orders ord
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)
WHERE ord.status = 'paid'
```

Remember - If the column name is unique then there is no need to write the table name, but if there are two column names with the same name then you need to either remove one column or keep the name different.

Don't worry about practicing the questions, you just need to go through all these questions first and then take up any online platform to practice the questions.

16. What is the use of the WHERE clause in SQL?

The WHERE clause is very a very important clause and you need to use it very frequently to optimize your query. Remember this line '**The WHERE clause restricts the number of rows and SELECT clause restricts the number of Columns'**

Suppose there are 100 Million rows in the orders table in amazon database. Now you are interested only in the order details of the last one month then you do not need to read all the 10 years of data, right?

You can use the following logic

```
SELECT *
FROM amazon.orders ord
WHERE year = 2022 and month = 1
```

Similarly, you can put many conditions in the where clause which we will cover in upcoming questions.

17. What is the use of the ORDER BY clause in SQL?

The ORDER BY clause is used to organize the output in a specific order, by default it takes the ascending clause. Same example as above

```
SELECT pro.product_id, COUNT(DISTINCT order_id) as number_of_orders
FROM amazon.orders ord
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)
WHERE ord.status = 'paid'
GROUP BY pro.product_id
ORDER BY 2 DESC
```

Here 2 represents the second column i.e. the one where we are counting the distinct number of orders as number_of_orders. We want to have the output with the product name with the maximum orders on the top.

Even better, let's have the top 10 product name with maximum order sold

```
SELECT pro.product_id, COUNT(DISTINCT order_id) as number_of_orders
FROM amazon.orders ord
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)
WHERE ord.status = 'paid'
GROUP BY pro.product_id
ORDER BY 2 DESC
LIMIT 10
```

The LIMIT clause will make sure that at max 10 rows are there in the output

18. What is the use of GROUP BY clause in SQL?

The GROUP BY clause is a very important clause and there will be a lot of questions where this concept will be used. GROUP BY is simply used to aggregate the rows on a set of columns.

Example - Suppose there is a table student with three columns student_name, class, total_marks

Suppose you want to know the total marks scored by all the students of each class then you will take the SUM() of total_marks and group it on class

```
SELECT class, SUM(total_marks) as Total_marks_of_class  
FROM student  
GROUP BY class
```

19. Write a SQL query to get the average marks scored by each class.

```
SELECT class, AVG(total_marks) as Avg_marks  
FROM student  
GROUP BY class
```

The above query will have only 10 rows i.e. from class 1st to 10th and will have 2 columns i.e. class and Avg_marks

20. Can we use the WHERE command with an aggregate function?

No, we can not use the WHERE command with an aggregate function like SUM, AVG, etc. The HAVING clause was added to SQL because the WHERE keyword cannot be used with aggregate functions.

21. What is the use of the HAVING clause in SQL?

The HAVING Clause enables you to specify conditions that filter which group results appear in the results. The WHERE clause places conditions on the selected columns, whereas the HAVING clause places conditions on groups created by the GROUP BY clause.

The HAVING clause must follow the GROUP BY clause in a query and must also precede the ORDER BY clause if used.

```
SELECT pro.product_id, COUNT(DISTINCT order_id) as number_of_orders
FROM amazon.orders ord
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)
WHERE ord.status = 'paid'
GROUP BY pro.product_id
HAVING count(order_id) > 100
```

The above code will get you only those product_id for which the number of orders are more than 100.

22. What is the use of the LIMIT clause in SQL?

We have already discussed it in the questions above, LIMIT is just used to make a condition on the number of rows to display as the output

LIMIT 100 at the end of the code will make sure that a max of 100 rows are displayed

15. What is the LIKE clause in SQL?

The LIKE clause is used very extensively in queries across the floor, it is useful because it gives extra power to the where clause and can help you at places where you think you are helpless in filtering out on some condition.

Ex.

```
SELECT pro.product_id, COUNT(DISTINCT order_id) as number_of_orders
FROM amazon.orders ord
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)
WHERE ord.status = 'paid' and product_name like 'iphone 10'
GROUP BY pro.product_id
HAVING count(order_id) > 100
```

23. Use the LIKE clause to filter all the products that start with a and end with e

```
SELECT pro.product_id, COUNT(DISTINCT order_id) as number_of_orders
FROM amazon.orders ord
INNER JOIN amazon.product pro on (ord.product_id = pro.product_id)
WHERE ord.status = 'paid' and product_name like 'a%ee'
GROUP BY pro.product_id
HAVING count(order_id) > 100
```

24. What are wildcards in SQL?

The wildcards are used to make sure that you are able to filter out specific conditions. There are two wildcards that are used in SQL

One is _ and the other is %

_ is used where you are not sure about only one specific position in the values, for example, you are not sure if the name is kamal or Komal, then you will use

Where name like 'k_mal'

Similarly % is used wherever you are unsure about zero, one or more characters like you just know that the name of a person is nitin kamal, but it could be nitin Singh kamal or nitin kumar kamal or just nitin kamal (three spaces between the first and the last name, in this case you will use something like

Where name like 'nitin%kamal'

% will make sure all the rows where the name has nitin followed by kamal is picked

25. One more important interview question on this line is when you are asked to pick the product name iphone but the iphone can be written as

Iphone
IPHONE
IphonE
Etc.

How to write that query?

Select * from order
Where lower(product_name) like 'iphone'

Here you converted everything in lower case and then checked for lower case iPhone, so all the matches are done and you don't have to worry about the different cases in the column.

Remember one thing, it's very unlikely that you will be asked a question like this in your interviews but you are expected to put this condition blindly everywhere there is a like statement on a string. This just shows that you write clean and production-ready code.

Even if the interviewer says that you need not keep everything lower, then also you get a brownie point because you know the importance of writing a robust code.

Believe me, on the floor you will definitely find n number of instances where a person has spent countless hours to fix a bug due to this silly issue [From a former bug fixer]

26. How to create a table with a single field as primary key?

```
CREATE TABLE Students (
    Roll_number INT NOT NULL
    Name VARCHAR(255)
    PRIMARY KEY (Roll_number)
);
```

27. How to create a table with multiple fields as primary key?

```
CREATE TABLE Students (
    ID INT NOT NULL
    LastName VARCHAR(255)
    FirstName VARCHAR(255) NOT NULL,
    CONSTRAINT PK_Student
    PRIMARY KEY (ID, FirstName)
);
```

28. What is a Primary Key?

The PRIMARY KEY constraint uniquely identifies each row in a table. It must contain UNIQUE values and has an implicit NOT NULL constraint.

A table in SQL is strictly restricted to having one and only one primary key, which is comprised of single or multiple fields (columns).

29. What is a UNIQUE constraint?

A UNIQUE constraint ensures that all values in a column are different. This provides uniqueness for the column and helps identify each row uniquely. Unlike the primary key, there can be multiple unique constraints defined per table. The code syntax for UNIQUE is quite similar to that of PRIMARY KEY and can be used interchangeably.

30. What is a Foreign Key?

A FOREIGN KEY comprises of single or collection of fields in a table that essentially refers to the PRIMARY KEY in another table. Foreign key constraint ensures referential integrity in the relation between two tables.

The table with the foreign key constraint is labeled as the child table, and the table containing the candidate key is labeled as the referenced or parent table.

31. What is a Join?

A join is an SQL operation performed to establish a connection between two or more database tables based on matching columns, thereby creating a relationship between the tables. Most complex queries in an SQL database management system involve joining commands.

32. What are the different types of joins?

There are a lot of joins and at times these are a bit confusing, we will try to understand the most used joins first:-

- A. Left Join
- B. Right Join
- C. Inner Join
- D. Outer Join

33. What is an inner join? Give its syntax.

Retrieves records that have matching values in both tables involved in the join. This is the widely used join for queries. It can be applied by both the keywords i.e. either JOIN or INNER JOIN

```
SELECT *  
FROM T_1  
JOIN T_2;
```

```
SELECT *  
FROM Table_A  
INNER JOIN Table_B;
```

34. What is Left Join? Give its syntax.

Left join retrieves all the elements in the left table and then looks for matches in the right-hand table. So, if there is a match then the corresponding value is populated else it populates Null.

```
SELECT *
FROM T_1 A
LEFT JOIN T_2 B
ON A.col = B.col;
```

35. What is Right Join? Give its syntax.

Right join retrieves all the elements in the right table and then looks for matches in the left-hand table. So, if there is a match then the corresponding value is populated else it populates Null.

```
SELECT *
FROM T_1 A
RIGHT JOIN T_2 B
ON A.col = B.col;
```

36. What is Full Join? Give its syntax.

Full Join retrieves all the elements from both the tables and populates where there is a matched value else Null is populated.

```
SELECT *
FROM Table_A A
FULL JOIN Table_B B
ON A.col = B.col;
```

37. What is self-join?

A self-join is used when you need to join the same table itself. As the name suggests, it is the join of table A on the table the same table A on the same column or a different column. A self JOIN is a case of regular join where a table is joined to itself based on some relation between its own column.

You can use any type of join for self-join i.e. Inner, Left, Right, Outer, etc.

In the next question, we will go through the classic example of self-join.

38. In an EmployeeDetails table, we have 3 columns:-

Emp_id, Emp_Name, and Mgr_id

The Manager id is nothing but the employee id of some employee. Example

1, A, 2

2, B, 3

3, C, 4

B is the manager of A and C is the manager of B.

Output required

Employee Name	Manager Name
A	B
B	C
C	D

Write a code in SQL to get the desired output.

Ans

```
SELECT e1.emp_Id EmployeeId, e1.emp_name EmployeeName,
       e1.emp_mgr_id ManagerId, e2.emp_name AS ManagerName
  FROM  tblEmployeeDetails e1
        JOIN tblEmployeeDetails e2
      ON e1.emp_mgr_id = e2.emp_id
```

39. Now in the above example, we have used Inner join as a part of the self join, the above will query will work fine for all the employees who have a Manager, but the Managing Director, CEO's, etc. won't necessarily have a Manager. What part of the query will you change to make sure all the employees of the company is present in the output.

Ans.

Instead of Inner Join, go for a left join. Thus, you will have all the employees name from this table and you can extract the name of

```

SELECT e1.emp_Id EmployeeId, e1.emp_name EmployeeName,
       e1.emp_mgr_id ManagerId, e2.emp_name AS ManagerName
  FROM  tblEmployeeDetails e1
        LEFT JOIN tblEmployeeDetails e2
          ON e1.emp_mgr_id = e2.emp_id

```

40. What is a Cross-Join?

Cross join can be defined as a cartesian product of the two tables included in the join. The table after join contains the same number of rows as in the cross-product of the number of rows in the two tables.

```

SELECT stu.name, sub.subject
  FROM students AS stu
CROSS JOIN subjects AS sub;

```

41. What is an Index?

A database index is a data structure that provides a quick lookup of data in a column or columns of a table. It enhances the speed of operations accessing data from a database table at the cost of additional writes and memory to maintain the index data structure.

42. We have two columns(Revenue and Cost Price) in a table like below, get me the Profit column

Revenue	Cost Price	Profit
100	Null	100
200	20	180
300	50	250
Null	50	-50

Ans.

Select nvl(Revenue,0)-nvl(Cost,0) as Profit
From Table

43. How do the SQL commands flow at the back end?

Ans.

Order of execution for an SQL query

- 1) FROM, including JOINS
- 2) WHERE
- 3) GROUP BY
- 4) HAVING
- 5) WINDOW Functions
- 6) SELECT
- 7) DISTINCT
- 8) UNION
- 9) ORDER BY
- 10) LIMIT AND OFFSET

But the reality isn't that easy nor straight forward. The SQL standard defines the order of execution for the different SQL query clauses. Said that modern databases are already challenging that default order by applying some optimization tricks which might change the actual order of execution, though they must end up returning the same result as if they were running the query at the default execution order.

44. Write a SQL query to get the second highest query using sub query.

```
SELECT
  Name, MAX(salary) AS salary
FROM Table_name
WHERE salary < (SELECT MAX(salary) FROM Table_name);
```

45. Write a SQL query to find all the student names Nitin in a table

```
select name
from student
where lower(name) like '%nitin%'
```

46. Write a query to get all the student with name length 10, starting with K and ending with z.

```
select name
from student
where length(name)=10 and lower(name) like 'k%z'
```

47. Write a SQL query to get the second highest query using Ranking

Note: Dense_rank() has been used to handle duplicate salaries if there are any.

With result as
{

```
select salary,  
dense_rank() over (order by salary desc) as salaryrank  
from employees  
}
```

```
select top 1 salary  
from result  
where salaryrank = 2
```

48. Can you use HAVING command without any aggregate function in SQL?

No it's not necessary for having to use aggregate functions and even without group by having can exist.

Eg: This query works well in PostgreSql

```
select 1 having 1 = 1;
```

49. You have data on people have applied for a lottery ticket. The data consists of their name and ticket number. You have to choose winners by selecting the people present in the alternate rows (the first winner starting from row number 3). Write a query to make things easy to select the winners.

(Hint- Choose alternate rows, beginning from row number 3)

```
select *  
from (select name, ROW_NUMBER() over (order by ticket_no) as srNo from db) t  
where (t.srNo % 2) = 1
```

50. We have the following values

```
10000  
10000  
20000  
30000  
30000  
30000
```

What would be the result of row number, rank, and dense rank ?

A) Row_Number() assigns a sequential integer to each row within the partition of a result set.

Ans:

```
1000 1  
1000 2
```

2000 3
3000 4
3000 5
3000 6

B) Rank() assigns a rank to each row within a partition of a result set. Rows in each partition receive the same ranks if they have the same values. But the ranks will be skipped here.

Ans:

1000 1
1000 1
2000 3
3000 4
3000 4
3000 4

C) Dense_Rank() differs from Rank() as it assigns consecutive ranks and ranks won't be skipped.

Ans:

1000 1
1000 1
2000 2
3000 3
3000 3
3000 3

51. Find all the students who either are male or live in Mumbai (have Mumbai as a part of their address).

Select name
From students
Where lower(gender) in ('male', 'm')
Or lower(address) = '%mumbai%'

52. Suppose there are two columns in employee table i.e. emp id and email get all the unique domains like gmail.com, yahoo.com, outlook.com, etc.

select substr(email, instr(email, '@') + 1, length(email)) as Domain
from emp1;

53. Can you join two table without any common column?

Yes we can do cross join without any common column.

Eg: We have Roll Number, Name of Students in Table A and their Class (let's say 5th) in Table B.

We will use cross join to append class against each student.

```
SELECT B.CLASS,A.ID,A.NAME  
FROM A, B  
WHERE 1=1
```

54 to 58. Give the output for the following

```
SELECT 'NITIN'+1  
SELECT 'NITIN'+'1'  
SELECT(SELECT 'NITIN')  
SELECT '1'+1  
SELECT 1+'1'
```

Ans.

SELECT 'NITIN'+1: Error (string to, int datatype conversion)

SELECT 'NITIN'+'1': NITIN1

SELECT(SELECT 'NITIN') : NITIN

SELECT '1'+1: 2 ('1' can be converted to 1)

SELECT 1+'1': 2

59. select case when null=null then 'Amit' else 'Rahul' end from dual.

What will be the output of the above query?

The Null value has a memory reference.2 Null values cannot have same memory Reference. So output will be 'Rahul'.

60. What is the difference between COUNT(*) and COUNT(ColName)?

COUNT(*) : It will return total number of records in table.

COUNT(ColName) : It will return total number of records where value for that ColName is Not-Null.

Eg: Table A

ID, Name, Dept

1,'A','D1'

2,'B',NULL

3,'C','D5'

COUNT(*) : 3
COUNT(Dept) : 2
COUNT(ID) : 3

61. Help me create a table with all the employee Names and Manager Names.

Employee_Name	Employee_Id	Manager_Id
A	1	2
B	2	3
C	3	2
D	4	3

```
SELECT e1.Employee_Name, e2.Employee_Name As Manager_Name
FROM Employee e1
INNER JOIN Employee e2
WHERE e1.Employee_id = e2.Manager_id
```

62. Write a query for collecting the names of children who pursuing their graduation in their residential city.

NAME	AGE	ADDRESS	COLLEGE
.....

(Hint- Look for the common city name in the columns of ADDRESS and COLLEGE)

Example

Nitin 26 Patna Patna
Ankit 27 Bangalore Pune

Result – Nitin

Ans.

```
select Name
from common data
where Substring_index( address,'',-1 ) =substring_Index (College,'',-1)
```

63. What is indexing in SQL?

Ans.

An index can be used to efficiently find all rows matching some column in your query and then walk through only that subset of the table to find exact matches. If you don't have indexes on any column in the WHERE clause, the SQL server has to walk through the whole table

and check every row to see if it matches, which may be a slow operation on big tables.

Creating an index involves the CREATE INDEX statement, which allows you to name the index, to specify the table and which column or columns to index, and to indicate whether the index is in an ascending or descending order.

Basic syntax

```
CREATE INDEX index_name ON table_name;
```

Single Column Index

```
CREATE INDEX index_name  
ON table_name (column_name);
```

Unique Index

```
CREATE UNIQUE INDEX index_name  
on table_name (column_name);
```

64. List the different types of relationships in SQL.

There are different types of relations in the database:

One-to-One – This is a connection between two tables in which each record in one table corresponds to the maximum of one record in the other.

One-to-Many and Many-to-One – This is the most frequent connection, in which a record in one table is linked to several records in another.

Many-to-Many – This is used when defining a relationship that requires several instances on each sides.

Self-Referencing Relationships – When a table has to declare a connection with itself, this is the method to employ.

65. Demonstrate how to write a query to show details of an HR whose name starts with M.

Name	Designation	Joining Date	Salary
Ashutosh Singla	HR	2019-12-15	80,000
Chandan Garg	Admin	2019-10-09	60,000
Himanshi Kaur	HR	2019-03-13	75,000
Mohit Dharm	HR	2019-05-27	85,000
Meena Batra	Accountant	2019-07-11	50,000
Omkar Singh	Accountant	2019-01-05	45,000
Pratham Garg	Admin	2019-05-27	65,000
Rakesh Sharma	HR	2018-06-30	75,000
Sharadha Gupta	Accountant	2019-09-18	50,000

```
select *
from table_name
where designation = 'HR' AND name LIKE 'M%';
```

66. What is OLTP?

OLTP, or online transactional processing, allows huge groups of people to execute massive amounts of database transactions in real-time, usually via the internet. A database transaction occurs when data in a database is changed, inserted, deleted, or queried.

67. What are the differences between OLTP and OLAP?

OLTP stands for online transaction processing, whereas OLAP stands for online analytical processing. OLTP is an online database modification system, whereas OLAP is an online database query response system.

68. What is the usage of the NVL() function?

You may use the NVL function to replace null values with a default value. The function returns the value of the second parameter if the first parameter is null. If the first parameter is anything other than null, it is left alone.

This function is used in Oracle, not in SQL and MySQL. Instead of NVL() function, MySQL have IFNULL() and SQL Server have ISNULL() function.

69. Explain character-manipulation functions? Explains its different types in SQL.

Change, extract, and edit the character string using character manipulation routines. The function will do its action on the input strings and return the result when one or more characters and words are supplied into it.

The character manipulation functions in SQL are as follows:

- A) CONCAT** (joining two or more values): This function is used to join two or more values together. The second string is always appended to the end of the first string.
- B) SUBSTR**: This function returns a segment of a string from a given start point to a given endpoint.
- C) LENGTH**: This function returns the length of the string in numerical form, including blank spaces.
- D) INSTR**: This function calculates the precise numeric location of a character or word in a string.
- E) LPAD**: For right-justified values, it returns the padding of the left-side character value.
- F) RPAD**: For a left-justified value, it returns the padding of the right-side character value.
- G) TRIM**: This function removes all defined characters from the beginning, end, or both ends of a string. It also reduced the amount of wasted space.
- H) REPLACE**: This function replaces all instances of a word or a section of a string (substring) with the other string value specified.

70. Get the number of duplicate names and their frequency

Table – Employee

Name

Nitin
Amit
Gaurav
Nitin
Amit

Output

Nitin – 2
Amit – 2
Gaurav – 1

Ans.

— if you want only the duplicate names

```
select name, count(name)
from employees
group by name having count(name) > 1
```

—returning all names and their frequency

```
select name, count(name) as frequency
from employees
group by name
```

71. Write the SQL query to get the third maximum salary of an employee from a table named employees.

```
SELECT * FROM(
  SELECT employee_name, salary, DENSE_RANK()
  OVER(ORDER BY salary DESC)r FROM Employee)
 WHERE r=&n;
```

To find 3rd highest salary set n = 3

72. Define records and fields in a table

A table is a collection of data components organized in rows and columns in a relational database. A table can also be thought of as a useful representation of relationships. The most basic form of data storage is the table. An example of an Employee table is shown below.

ID	Name	Department	Salary
1	Rahul	Sales	24000
2	Rohini	Marketing	34000
3	Shylesh	Sales	24000
4	Tarun	Analytics	30000

A Record or Row is a single entry in a table. In a table, a record represents a collection of connected data. The Employee table, for example, has four records.

A table is made up of numerous records (rows), each of which can be split down into smaller units called Fields(columns). ID, Name, Department, and Salary are the four fields in the Employee table above.

73. What is the use of FETCH command?

The FETCH command cannot be used alone. It has to be used in conjunction with the OFFSET command.

It is used to return a set of number of rows.

The OFFSET argument is used to identify the starting point to return rows from a result set. Basically, it exclude the first set of records.

Example

```
SELECT * FROM Employee  
ORDER BY Salary  
OFFSET 5 ROWS  
FETCH NEXT 10 ROWS ONLY;
```

The above query will skip the first 5 rows and return the next 10 rows.

ORDER BY clause is mandatory to be used with OFFSET and FETCH. OFFSET value must be greater than or equal to 0. It cannot be negative.

74. What are UNION, MINUS and INTERSECT commands?

The UNION operator is used to combine the results of two tables while also removing duplicate entries.

The MINUS operator is used to return rows from the first query but not from the second query.

The INTERSECT operator is used to combine the results of both queries into a single row.
Before running either of the above SQL statements, certain requirements must be satisfied –

Within the clause, each SELECT query must have the same amount of columns.

The data types in the columns must also be comparable.

In each SELECT statement, the columns must be in the same order.

75. How to fetch alternate records(even rows) from a table?

— To fetch even records

```
Select *, Row_Number() Over(order by salary) as rowno from employees
```

— To fetch odd number of records

```
select * from  
Table A  
WHERE ID % 2 == 0
```

76. You always have a big data i.e. millions of rows in your tables, how would you partition it for optimum performance?

Ans.

MySQL partitioning is about altering – ideally, optimizing – the way the database engine physically stores data. It allows you to distribute portions of table data across the file system based on a set of user-defined rules. In this way, if the queries you perform access only a fraction of table data and the partitioning function is properly set, there will be less to scan and queries will be faster. Partitioning makes the most sense when dealing with millions of data.

Horizontal partitioning means that all rows matching the partitioning function will be assigned to different physical partitions.

Vertical partitioning allows different table columns to be split into different physical partitions.

RANGE Partitioning:

This type of partition assigns rows to partitions based on column values that fall within a stated range. The values should be contiguous, but they should not overlap each other. The VALUES LESS THAN operator will be used to define such ranges in order from lowest to highest.

LIST partitioning:

It is similar to RANGE, except that the partition is selected based on columns matching one of a set of discrete values. In this case, the VALUES IN statement will be used to define matching criteria.

HASH Partitioning:

In HASH partitioning, a partition is selected based on the value returned by a user-defined expression. This expression operates on column values in rows that will be inserted into the table. A HASH partition expression can consist of any valid MySQL expression that yields a nonnegative integer value. HASH is used mainly to evenly distribute data among the number of partitions the user has chosen.

LINEAR HASH:

Instead of using the modulo described above, when MySQL uses LINEAR HASH a powers-of-two algorithm is employed to calculate the partition where the data is to be stored. Syntactically, LINEAR HASH is exactly the same as HASH, except for the addition of the word LINEAR.

77. Suppose in class, you have $3n$ boys and $2n$ girls with their names tabulated along with their weight and gender. Write a SQL query to separate students alphabetically who are over-weight (55kg for girls, 75kg for boys)

Ans.

```
select *  
from table_name  
where gender = 'Male' and weight > '75kg'  
union all  
select *  
from table_name  
where gender = 'Female' and weight > '55kg'
```

78. What is indexing in SQL? Explain with proper example.

Ans.

An index can be used to efficiently find all rows matching some column in your query and then walk through only that subset of the table to find exact matches. If you don't have indexes on any column in the WHERE clause, the SQL server has to walk through the whole table

and check every row to see if it matches, which may be a slow operation on big tables.

Creating an index involves the CREATE INDEX statement, which allows you to name the index, to specify the table and which column or columns to index, and to indicate whether the index is in an ascending or descending order.

Basic syntax

```
CREATE INDEX index_name ON table_name;
```

Single Column Index

```
CREATE INDEX index_name  
ON table_name (column_name);
```

Unique Index

```
CREATE UNIQUE INDEX index_name  
on table_name (column_name);
```

79. Arrange the employees with respect to their Joining date, most experienced employee coming on the top followed by others.

(Hint- In case of same date, preference should be HR>Admin>Accountant)

Ans.

```
SELECT*FROM Employees  
ORDER BY Joining_Date, FIELD( Designation, 'HR','Admin','Accountant');
```

80. What is Cursor? How to use a Cursor?

After any variable declaration, DECLARE a cursor. A SELECT Statement must always be coupled with the cursor definition.

To start the result set, move the cursor over it. Before obtaining rows from the result set, the OPEN statement must be executed.

To retrieve and go to the next row in the result set, use the FETCH command.

To disable the cursor, use the CLOSE command.

Finally, use the DEALLOCATE command to remove the cursor definition and free up the resources connected with it.

81. List the different types of relationships in SQL.

There are different types of relations in the database:

One-to-One – This is a connection between two tables in which each record in one table corresponds to the maximum of one record in the other.

One-to-Many and Many-to-One – This is the most frequent connection, in which a record in one table is linked to several records in another.

Many-to-Many – This is used when defining a relationship that requires several instances on each sides.

Self-Referencing Relationships – When a table has to declare a connection with itself, this is the method to employ.

82. How to create a temp table in SQL Server?

Temporary tables are created in TempDB and are erased automatically after the last connection is closed. We may use Temporary Tables to store and process interim results. When we need to store temporary data, temporary tables come in handy.

The following is the syntax for creating a Temporary Table:

```
CREATE TABLE #Employee (id INT, name VARCHAR(25))
INSERT INTO #Employee VALUES (01, 'Ashish'), (02, 'Atul')
```

83. Write an SQL Query find number of employees whose DOB is between 01/07/1965 to 31/12/1975. Collect the data separately for different gender.

Use column name as sex, DOB table name Employees.

(Hint- Imagine a hypothetical data)

Ans.

```
Select Count(*) , Sex
from Employees
WHERE DOB BETWEEN CAST('1965-07-01' as Date) AND CAST('1965-12-31' as Date)
```

Note: When you use Between operator with Date Values, It is better to use with CAST Function to get better results.CAST Function helps us to convert the Type of column or expression to the Date Type.

84. How can you create an empty table from an existing table? Write the steps and explain the working.

Take the following table for instance and create a new table named as cstudent.

Ans.

```
create table cstudent  
AS  
(select * from student  
where 1=0  
)
```

85. What is the use of IFNULL and ISNULL in SQL?

Ans.

IFNULL() and ISNULL() are two functions in sql to check the existence of null values in a particular column and replace with a value depending on the type of DB used.

ISNULL() in Sql server:- Checks for the null values and if it's present it returns with an alternate value.

```
ISNULL(column_name, value_to_be_replaced)
```

ISNULL() in MYSQL :- Checks the column value if it is null or not null. It returns a boolean result.

IFNULL in MYSQL :- Checks for the null values and if it's present it returns with an alternate value.

```
IFNULL(column_name, value_to_be_replaced)
```

86. You have got some data in the Table 1 and Table 2

Write a SQL query to create a Table 3 that contains the following columns- Id, First_Name, Last_Name, Salary

(Hint- The columns should be in the prescribed order)

Ans.

```
CREATE TABLE table3  
AS  
SELECT t1.id , t1.first_name ,t1.last_name ,t2.salary  
FROM  
table1 t1 JOIN table2 t2 ON t1.id = t2.id
```

87. NoSQL vs SQL

In summary, the following are the five major distinctions between SQL and NoSQL:

Relational databases are SQL, while non-relational databases are NoSQL.

SQL databases have a specified schema and employ structured query language. For unstructured data, NoSQL databases use dynamic schemas.

SQL databases scale vertically, but NoSQL databases scale horizontally.

NoSQL databases are document, key-value, graph, or wide-column stores, whereas SQL databases are table-based.

SQL databases excel in multi-row transactions, while NoSQL excels at unstructured data such as documents and JSON.

88. What is the difference between NOW() and CURRENT_DATE()?

NOW() returns a constant time that indicates the time at which the statement began to execute. (Within a stored function or trigger, NOW() returns the time at which the function or triggering statement began to execute).

The simple difference between NOW() and CURRENT_DATE() is that NOW() will fetch the current date and time both in format 'YYYY-MM_DD HH:MM:SS' while CURRENT_DATE() will fetch the date of the current day 'YYYY-MM_DD'.

89. Extract the information about all department managers who were hired between the 1st of January 2020 and the 1st of January 2021.

Ans.

```
SELECT *
FROM
dept_manager
WHERE
Emp_no IN (SELECT
Emp_no
FROM
employees
WHERE
Hire_date BETWEEN '2020-01-01' AND '2021-01-01');
```

90. If a table contains duplicate rows, does a query result display the duplicate values by default? How can you eliminate duplicate rows from a query result?

Ans.

Yes, the query will display duplicate rows.

To eliminate duplicate rows, you can try the following query:

```
WITH cte as (
Select contact_id, first_name, last_name, email,
```

```
Row_number() OVER (PARTITION BY  
first_name,last_name,email  
ORDER BY  
first_name,last_name,email)row_num  
FROM Contacts  
)  
DELETE from cte where row_num > 1
```

91. How to create a stored procedure using SQL Server?

A stored procedure is a piece of prepared SQL code that you can save and reuse again and over.

So, if you have a SQL query that you create frequently, save it as a stored procedure and then call it to run it.

You may also supply parameters to a stored procedure so that it can act based on the value(s) of the parameter(s) given.

Stored Procedure Syntax

```
CREATE PROCEDURE procedure_name
```

```
AS
```

```
sql_statement
```

```
GO;
```

Execute a Stored Procedure

```
EXEC procedure_name;
```

92. Which join is used to join a table with itself

- a. Inner join
- b. full join
- c. self join

Ans:

Self Join

Explanation: Self-join will create a virtual table that is a copy of the table itself to join which last for the moment of operation and doesn't occupy any extra space, as the table is a copy of the existing table.

93. What is the difference between CHAR and VARCHAR2 datatype in SQL?

Both Char and Varchar2 are used for characters datatype but varchar2 is used for character strings of variable length whereas Char is used for strings of fixed length. For example, char(10) can only store 10 characters and will not be able to store a string of any other length whereas varchar2(10) can store any length i.e 6,8,2 in this variable.

94. List the total numbers of products of each brand(Take the name of the table as well as the column name by yourself)

```
select count(Product_Brand),Product_Brand  
from Product_Master  
group by Product_Brand
```

95. Does the virtual table created occupy space for the operation to joining with itself?

Ans.

Yes, a virtual table created in a database does occupy space in memory or on disk, depending on the specific database implementation. This space is used to store the data and metadata associated with the virtual table, such as column names, data types, and any indexes or constraints defined on the table.

If you are performing a self-join operation on a virtual table, then the database engine will need to allocate additional space in memory to store the results of the join operation. The amount of space required will depend on the size of the virtual table and the specific join operation being performed.

It's worth noting that virtual tables are typically implemented as views or temporary tables, which means that the space they occupy is generally reclaimed by the database engine when the table is no longer needed. However, it's still important to be aware of the space requirements of virtual tables when designing and optimizing database queries.

96. What are Constraints?

Constraints in SQL are used to specify the limit on the data type of the table. It can be specified while creating or altering the table statement. The sample of constraints are:

NOT NULL
CHECK
DEFAULT
UNIQUE

PRIMARY KEY
FOREIGN KEY

97. What is the difference between DELETE and TRUNCATE statements?

DELETE and TRUNCATE are both SQL commands used to remove data from a table, but they work in different ways and have different effects.

The main difference between DELETE and TRUNCATE is that DELETE removes rows one by one, while TRUNCATE removes all rows from a table at once. Here are some more details:

DELETE: The DELETE statement is used to remove one or more rows from a table based on some condition. It can be used with a WHERE clause to specify which rows to delete, or without a WHERE clause to delete all rows in the table. When you use DELETE, the rows are removed one by one, and the table's indexes and triggers are updated as each row is deleted. This means that DELETE can be slower than TRUNCATE when deleting large amounts of data.

TRUNCATE: The TRUNCATE statement is used to remove all rows from a table without removing the table itself. When you use TRUNCATE, the entire table is emptied in one operation, and any indexes, triggers, or constraints on the table are also removed. Because TRUNCATE removes all rows at once, it is usually much faster than DELETE for large tables, since it doesn't need to update indexes or triggers for each individual row.

Another difference between DELETE and TRUNCATE is that DELETE can be rolled back using a transaction, while TRUNCATE cannot be rolled back. This means that if you accidentally delete the wrong rows with DELETE, you can use a transaction to undo the change, but if you accidentally truncate a table, the data is permanently lost.

In summary, DELETE is used to remove individual rows from a table based on a condition, while TRUNCATE is used to remove all rows from a table at once. TRUNCATE is usually faster than DELETE for large tables, but it cannot be rolled back like DELETE can.

DELETE	TRUNCATE
Delete command is used to delete a row in a table.	Truncate is used to delete all the rows from a table.
You can rollback data after using delete statement.	You cannot rollback data.
It is a DML command.	It is a DDL command.
It is slower than truncate statement.	It is faster.

98. What is CROSS JOIN UNNEST in Presto ?

CROSS JOIN UNNEST – It is simply used to flatten an array, flattening means converting an Array, Map or Row in a flat relation by converting it into multiple rows (one row for every value in array)

Sample table

Name	Emp_id	Subject(Array)	Phone
X	123	[C, JAVA, SQL]	[123,456]
Y	4231	[Hive, Presto]	[542.654]
Z	322	[Ruby, Perl]	[12343]
Q	421	[Python, R]	[765,987]

```
Select Name,Emp_id,expertise  
from Employee  
CROSS JOIN UNNEST(Subject) as t(expertise)
```

If there are multiple arrays

```
Select Name,Emp_id,expertise,phone_num  
from Employee  
CROSS JOIN UNNEST(Subject,Phone) as t(expertise,phone_num)
```

99. What is LATERAL VIEW explode in Hive?

LATERAL VIEW explode in Hive is the same as CROSS JOIN UNNEST in Presto.

```
Select Name,Emp_id,expertise  
from Employee  
LATERAL VIEW explode(Subject) myTable1 as expertise
```

If there are two columns to be unnested then

```
Select Name,Emp_id,expertise,Phone  
from Employee  
LATERAL VIEW explode(Subject) myTable1 as expertise  
LATERAL VIEW explode(Phone) myTable2 as Phone
```

100-103. Suppose there are two tables, X and Y, X has just one column A and Y has B.

These are the two tables

X(A)	Y(B)
1	1
2	2
3	3
	4
	5

Questions – How many rows will be populated if you do

- X left join Y
- X inner join Y
- X cross join Y
- X right join Y

The answer is 3,3,3 and 5

104.

In the table below, some duplicate records might be present by mistake. Sort out a way to locate them and find a way to delete them.

empid	empname	managerid	deptid	Salary
1	Emp1	0	1	6000
2	Emp2	0	5	6000
3	Emp3	1	1	2000
13	Emp13	2	5	2000
11	Emp11	2	1	2000
9	Emp9	1	5	3000
8	Emp8	3	1	3500
7	Emp7	2	5	NULL
3	Emp3	1	1	2000

```
WITH cte as (
    Select *,Row_number() OVER (PARTITION BY
        empid,empname,managerid,deptid,salary
    ORDER BY
        empid,empname,managerid,deptid,salary) row_num
    FROM Employees)
Delete from cte
where row_num > 1
```

105. What is the difference between clustered and non-clustered index in SQL?

The differences between the clustered and non clustered index in SQL are :

Clustered index is used for easy retrieval of data from the database and its faster whereas reading from non clustered index is relatively slower.

Clustered index alters the way records are stored in a database as it sorts out rows by the column which is set to be clustered index whereas in a non clustered index, it does not alter the way it was stored but it creates a separate object within a table which points back to the original table rows after searching.

One table can only have one clustered index whereas it can have many non clustered index.

106. Pivot a table in SQL without using pivot function

Suppose there are two columns

Age Name

25 Nitin

30 Amit

27 Rishab

29 Ankush

Convert into

Name Nitin. Amit. Rishab. Ankush

Age. 25. 30. 27. 29

Ans.

```

With CTE As
(SELECT Age, Name, ROW_NUMBER() OVER() As Row_Number
FROM table)
SELECT
MAX(CASE Name WHEN 'Nitin' Then Age End) Nitin,
MAX(CASE Name WHEN 'Amit' Then Age End) Amit,
MAX(CASE Name WHEN 'Rishab' Then Age End) Rishab,
MAX(CASE Name WHEN 'Ankush' Then Age End) Ankush
FROM CTE
GROUP BY Row_Number

```

107. What is the function of OFFSET command? You might know the definition but create a scenario where you have to make use of it. Also, can the same process be done with any other method?

Ans. OFFSET command is used to skip rows from the results which has been fetched from the query. Suppose we have a table and we want to fetch all the rows apart from the first 5 rows, then we can use OFFSET command. An alternate method to do similar kind of operation would be to use the Row_Number() window function. This function assigns a unique row number to every record in the table starting from 1. Using this column we can impose the condition with the help of the 'WHERE' clause.

108. Get all employee detail from EmployeeDetail table whose “FirstName” not start with any single character between ‘a-p’

Ans.

```

Select *
from EmployeeDetail
where FirstName LIKE '[^a-p]%'

```

109. What do you understand by query optimization?

The phase that identifies a plan for evaluation query which has the least estimated cost is known as query optimization. The advantages of query optimization are as follows:

- The output is provided faster
- A larger number of queries can be executed in less time
- Reduces time and space complexity

110. Important conditions for joining two tables on a key?

Primary conditions for joining two tables on keys:

1. The key column should contain non-NULL values
2. The datatype of both tables's column should be similar

111.What are Entities and Relationships?

Entities: A person, place, or thing in the real world about which data can be stored in a database. Tables store data that represents one type of entity. For example – A bank database has a customer table to store customer information. The customer table stores this information as a set of attributes (columns within the table) for each customer.

Relationships: Relation or links between entities that have something to do with each other. For example – The customer name is related to the customer account number and contact information, which might be in the same table. There can also be relationships between separate tables (for example, customer to accounts).

112. Assume the name of table and columns

There are two tables with a common column. Which one will take more processing time

- Outer Join
- Full Outer Join
- Cartesian Join

Ans.

Cartesian Join will take the most processing time.

Suppose, table 1 has 10 rows and table 2 has 20 rows, with 5 rows common.

Cartesian join will give $10 \times 20 = 200$ rows.

Full outer join will give $(10 + 20 - 5) = 25$ rows(Left outer + Right Outer – inner).

Left outer join will give 10 and Right Outer will give 20 rows.

This is a scenario when you are generally matching on some columns which are acting as primary key or have unique values.

If all the values in the columns which you are matching are same, then probably, they would all take the same time as they would produce the same output.

113. What is NTILE with syntax? With example and use case

Ans.

NTILE() function distributes the rows in an ordered partition into specific number of groups. It assigns each group a bucket number starting from one.

For Example.

NTILE(10) will divide the 100 rows into 10 groups, with each group consisting of 10 rows.

If the groups are not equally divided, the function will set more rows to the starting groups and less to the following groups.

Now, suppose we have employee table with 2 columns as empname and salary and having 6 records. The following query will divide those 6 rows into 3 buckets and number them as 1, 2 and 3 according to their salary.

Query:

```
SELECT empname,  
NTILE (3) OVER (  
ORDER BY salary DESC  
) buckets  
FROM employees;
```

114. What is Normalization and what are the advantages of it?

Normalization in SQL is the process of organizing data to avoid duplication and redundancy.

Some of the advantages are:

- Better Database organization
- More Tables with smaller rows
- Efficient data access
- Greater Flexibility for Queries
- Quickly find the information
- Easier to implement Security
- Allows easy modification
- Reduction of redundant and duplicate data
- More Compact Database
- Ensure Consistent data after modification

115. Explain different types of Normalization.

There are many successive levels of normalization. These are called normal forms. Each consecutive normal form depends on the previous one. The first three normal forms are usually adequate.

Normal Forms are used in database tables to remove or decrease duplication. The following are the many forms:

First Normal Form:

When every attribute in a relation is a single-valued attribute, it is said to be in first normal form. The first normal form is broken when a relation has a composite or multi-valued property.

Second Normal Form:

A relation is in second normal form if it meets the first normal form's requirements and does not contain any partial dependencies. In 2NF, a relation has no partial dependence, which means it has no non-prime attribute that is dependent on any suitable subset of any table candidate key. Often, the problem may be solved by setting a single column Primary Key.

Third Normal Form:

If a relation meets the requirements for the second normal form and there is no transitive dependency, it is said to be in the third normal form.

116. Find the Nth largest salary from employee table.

Best way to get the Nth highest salary from a table.

Table name – Employee

Columns – Emp_id, salary,depar

Ans.

Use OFFSET and LIMIT to get the Nth highest salary.

We can use Dense_Rank() or Sub Query to get the same output.

```
Select Salary from Employee  
ORDER BY Salary Desc  
LIMIT n-1 ,1
```

In the above query, replace "N" with the value of the Nth largest salary you want to find. For example, if you want to find the 2nd largest salary, replace N with 2.

The query first orders the salaries in descending order using the ORDER BY clause. The LIMIT clause then skips the first N-1 rows and selects the next row, which corresponds to the Nth largest salary.

117. How the triggers will execute if two or more triggers?

A trigger is a stored procedure in database which automatically invokes whenever a special event in the database occurs. For example, a trigger can be invoked when a row is inserted into a specified table or when certain table columns are being updated.

118. Get all the employee detail from EmployeeDetail table

whose “FirstName” starts with A and contain 5 letters

Ans.

Since we know that FirstName will start from A so using like ‘%a’ is wrong/redundant. Also the total length of the FirstName is 5 letters , so rest of the 4 letters are _ (4 underscore). The underscore character (_) represents a single character to match a pattern. More than one (_) underscore characters can be used to match a pattern of multiple characters.

```
select *  
from EmployeeDetail  
WHERE FirstName LIKE 'A____'
```

119. Could you tell output or result of following SQL statements?

(Hint- In some cases, there may be an error. So, try to locate them and answer accordingly)

```
select 5  
  
select '5'  
  
select count ('5')  
  
select count (5)  
  
select count (*)
```

Ans.

```
SELECT 5 :- 5  
SELECT '5' :- 5  
SELECT COUNT(5) :- 1  
SELECT COUNT('5') :- 1  
SELECT COUNT(*) :- 1
```

Point to be known is that there is no difference between last 3 queries, they will always return identical results, wherever they are used.

120. How do you maintain database integrity where deletions from one table will automatically cause deletions in another table?

You can create a Trigger that will automatically delete elements in the second table when elements from the first table are removed.

121. Find the number of people who are from Delhi and have arrived in Patna in the last 7 days

Ans.

```
select count(*)
from table
where (resident_place= 'Delhi' and arrival_place = 'Patna' ) and
(datetime_col >= DATE(NOW()) + INTERVAL -7 DAY and
datetime_col < DATE(NOW()) + INTERVAL 0 DAY)
```

122. How do you handle ties in dense ranking?

Dense rank automatically handles ties in the ordering column by assigning the same rank value to all tied rows, without leaving gaps in the ranking values.

123. Does the data stored in the stored procedure increase access time or execution time?

Ans.

A stored procedure is a set of Structured Query Language (SQL) statements with an assigned name, which are stored in a relational database management system as a group, so it can be reused and shared by multiple programs. Data stored in stored procedures can be retrieved much faster than the data stored in SQL database. Data can be precompiled and stored in Stored procedures. This reduces the time gap between query and compiling as the data has been precompiled and stored in the procedure.

124. How many Aggregate functions are available in SQL?

SQL aggregate functions provide information about a database's data. AVG, for example, returns the average of a database column's values.

SQL provides seven (7) aggregate functions, which are given below:

AVG(): returns the average value from specified columns.

COUNT(): returns the number of table rows, including rows with null values.

MAX(): returns the largest value among the group.

MIN(): returns the smallest value among the group.

SUM(): returns the total summed values(non-null) of the specified column.

FIRST(): returns the first value of an expression.

LAST(): returns the last value of an expression.

125. When do we use Coalesce() function?

Any amateur SQL person would think Why to use COALESCE , when we have ISNULL function

in SQL. This is because when we have multiple values , and not sure they might be null , then we opt for COALESCE. It return the first non-null value in a list.

For e.g.

```
emp_vehicle_count = NULL  
emp_mobile_count = NULL  
emp_count = 200  
SELECT COALESCE(emp_vehicle_count, emp_mobile_count, emp_count);  
Output : 200
```

126. You have a table with the following 3 columns

customer_id, order_id, order_date		
123.	987.	2/9/2020
123	1234	4/9/2020
456	1211	1/1/2019
456	2344	8/8/2020

We want to know the number of customer who have their first two orders in the last 180 days. In the above example 123 will qualify but 456 won't qualify

Ans.

```
select count(DISTINCT(customer_id))from  
(  
-- To obtain the order id  
select *, ROW_NUMBER() over (PARTITION BY customer_id order by order_date  
asc) as order_number from trx_table)  
as a  
where a.order_date >= DATE_SUB(NOW(), INTERVAL 180 DAYS) and a.order_number  
in (1,2)
```

127. What is a RANK() function?

Rank()

```
Select *,  
Rank() Over(order by salary)as rank  
from employees
```

Rank() function will assign same ranks to the observations who have same value(over which it has been ordered, in this case salary) whereas Row_Number() will assign a unique Row Number to every observation starting from 1 till the count of the rows.

128. How do we use the DISTINCT statement? What is its use?

The SQL DISTINCT keyword is combined with the SELECT query to remove all duplicate records and return only unique records. There may be times when a table has several duplicate records. The DISTINCT clause in SQL is used to eliminate duplicates from a SELECT statement's result set

129. What is the ACID property in a database?

ACID stands for Atomicity, Consistency, Isolation, Durability. It is used to ensure that the data transactions are processed reliably in a database system.

Atomicity: Atomicity refers to the transactions that are completely done or failed where transaction refers to a single logical operation of a data. It means if one part of any transaction fails, the entire transaction fails and the database state is left unchanged.

Consistency: Consistency ensures that the data must meet all the validation rules. In simple words, you can say that your transaction never leaves the database without completing its state.

Isolation: The main goal of isolation is concurrency control.

Durability: Durability means that if a transaction has been committed, it will occur whatever may come in between such as power loss, crash or any sort of error.

130. What are the different types of a subquery?

There are two types of subquery namely, Correlated and Non-Correlated.

Correlated subquery: These are queries which select the data from a table referenced in the outer query. It is not considered as an independent query as it refers to another table and refers the column in a table.

Non-Correlated subquery: This query is an independent query where the output of subquery is substituted in the main query.

131. Write an SQL query to fetch the current date-time from the system.

TO fetch the CURRENT DATE IN SQL Server

SELECT GETDATE();

TO fetch the CURRENT DATE IN MYSQL

SELECT NOW();

TO fetch the CURRENT DATE IN Oracle

SELECT SYSDATE FROM DUAL();

132 to 137

There are two tables:

Patients Table				
Patient ID	Patient Name	Sex	Age	Address
01	Sheela	F	23	Flat no 201, Vasavi Heights, Yakutap
02	Rehan	M	21	Building no 2, Yelahanka
03	Anay	M	56	H No 1, Panipat
04	Mahira	F	42	House no 12, Gandhinagar
05	Nishant	M	12	Sunflower Heights, Thane

PatientsCheckup Table

Patient ID	BP	Weight	Consultation Fees
01	121/80	67	300
02	142/76	78	400
03	151/75	55	300
04	160/81	61	550
05	143/67	78	700

132. Write a query to fetch top N records using the TOP/LIMIT, ordered by ConsultationFees.

TOP Command – SQL Server

```
SELECT TOP N *
FROM PatientsCheckup
ORDER BY ConsultationFees DESC;
```

LIMIT Command - MySQL

```
SELECT TOP N *
FROM PatientsCheckup
ORDER BY ConsultationFees DESC;
```

133. Write a SQL query to create a table where the structure is copied from other table.

Create an Empty Table

```
SELECT *
FROM PatientsCheckup
ORDER BY ConsultationFees DESC
LIMIT N;
```

134. Write a query to fetch even and odd rows from a table.

If you have an auto-increment field like PatientID then you can use the MOD() function:

Fetch even rows using MOD() function:

```
CREATE TABLE NewPatientsTable
SELECT *
FROM Patients
WHERE 1=0;
```

Fetch odd rows using MOD() function:

```
SELECT *
FROM Patients
WHERE MOD(PatientID,2)=0;
```

In case there are no auto-increment fields then you can use the Row_number in SQL Server or a user-defined variable in MySQL. Then, check the remainder when divided by 2.

135. Write an SQL query to fetch duplicate records from Patients, without considering the primary key.

```
SELECT *
FROM Patients
WHERE MOD(PatientID,2)=1;
```

136. Write a query to fetch the number of patients whose weight is greater than 68.

```
SELECT COUNT(*)
FROM PatientsCheckup
WHERE Weight > '68';
```

137. Write a query to retrieve the list of patients from the same state.

```
SELECT DISTINCT P.PatientID, P.PatientName, P.State
FROM Patients P, Patient P1
WHERE P.State = P1.State AND P.PatientID != P1.PatientID;
```

138. Write a SQL query to fetch consultation fees – wise count and sort them in descending order.

```
SELECT DISTINCT P.PatientID, P.PatientName, P.State
FROM Patients P, Patient P1
WHERE P.State = P1.State AND P.PatientID != P1.PatientID;
```

139. Write a SQL query to retrieve patient details from the Patients table who have a weight in the PatientsCheckup table.

```
SELECT ConsultationFees, COUNT(PatientId) CFCount
FROM PatientsCheckup
GROUP BY ConsultationFees
ORDER BY CFCount DESC;
```

140. Write a SQL query to retrieve the last 2 records from the Patients table.

```
SELECT * FROM Patients P
WHERE EXISTS
(SELECT * FROM PatientsCheckup C WHERE P.PatientID = C.PatientID);
```

141. Write a SQL query to find all the patients who joined in the year 2022.

–USING BETWEEN

```
SELECT * FROM Patients WHERE
PatientID <=2 UNION SELECT * FROM
(SELECT * FROM Patients P ORDER BY P.PatientID DESC)
AS P1 WHERE P1.PatientID <=2;
```

– USING YEAR

```
SELECT * FROM Patients
WHERE RegDate BETWEEN '2021/01/01' AND '2021/12/31';
```

142. Write a SQL query to fetch 50% records from the PatientsCheckup table.

```
SELECT *
FROM PatientsCheckup WHERE
PatientID <= (SELECT COUNT(PatientD)/2 FROM PatientsCheckup);
```

143. Write a query to update the patient names by removing the leading and trailing spaces.

```
UPDATE Patients
SET PatientName = LTRIM(RTRIM(PatientName));
```

144. Write a query to add email validation to your database.

```
SELECT email FROM Patients WHERE NOT REGEXP_LIKE(email,
'[A-Z0-9._%+-]+@[A-Z0-9.-]+.[A-Z]{2,4}', 'i');
```

145. Write a query to find all patient names whose name:

Begin with A

Ends with S and contains 3 alphabets

Staying in the state Telangana

```
SELECT * FROM Patients WHERE PatientName LIKE 'A%';
SELECT * FROM Patients WHERE PatientName LIKE '__S';
SELECT * FROM Patients WHERE State LIKE 'Telangana%';
```

146. Write a SQL query to fetch details of all patients excluding patients with name “Sheela” and “Anay”.

```
SELECT * FROM Patients WHERE PatientName NOT IN ('Sheela', 'Anay');
```

147. Write a query to fetch the total count of occurrences of a particular character – ‘x’ in the PatientName.

```
SELECT PatientName, PatientID
LENGTH(PatientName) - LENGTH(REPLACE(PatientName, 'x', '')) 
FROM Patients;
```

148. Write a query to retrieve the first three characters of PatientName from the Patients table.

```
SELECT SUBSTRING(PatientName, 1, 3) FROM Patients;
```

149. Write a query to combine Address and state into a new column – NewAddress.

```
SELECT CONCAT(Address, ' ', State) AS 'NewAddress' FROM Patients;
```

150. Write a query to fetch PatientIDs which are present in:

- a. Both tables
- b. One of the table.

Let us say, patients present in Patients and not in the PatientsCheckup table.

–Present IN BOTH TABLES

```
SELECT PatientId FROM Patients
WHERE PatientId IN
(SELECT PatientId FROM PatientsCheckup);
```

– Present IN One OF the TABLE

```
SELECT PatientId FROM Patients
WHERE PatientId NOT IN
(SELECT PatientId FROM PatientsCheckup);
```

151. Write a query to find the number of patients whose RegDate is between 01/04/2021 to 31/12/2022 and are grouped according to state.

```
SELECT COUNT(*), State FROM Patients WHERE RegDate BETWEEN '01/04/2021' AND
'31/12/2022' GROUP BY State
```

152. Write a query to fetch all records from the Patients table; ordered by PatientName in ascending order, State in descending order.

```
SELECT * FROM Patients ORDER BY PatientName ASC, State DESC;
```

153. How can you create an empty table from an existing table? Write the steps and explain the working.

Take the following table for instance and create a new table named as cststudent.

NAME	MARKS	ROLL NUMBER
Ashutosh	87	1001
Bhavya	92	1002
Garima	69	1003
Pratham	75	1004
Sushant	90	1005

Ans.

```
create table cststudent
AS
(select * from student
where 1=0
)
```

154. For what purpose we are using view? How is it created to get the job done?

Ans.

Views are virtual tables that can be a great way to optimize your database experience. Not only are views good for defining a table without using extra storage, but they also accelerate data analysis and can provide your data extra security.

```
CREATE VIEW V_Customer
AS SELECT *
FROM Customer;
```

155. Query the list of CITY names from STATION that does not start with vowels and do not end with vowels. Your result cannot contain duplicates.

Input Format

The STATION table is described as follows:

STATION	
Field	Type
ID	NUMBER
CITY	VARCHAR2(21)
STATE	VARCHAR2(2)
LAT_N	NUMBER
LONG_W	NUMBER

```
SELECT DISTINCT CITY
FROM STATION
WHERE CITY regexp '^[^aeiou]' and CITY regexp '[^aeiou]$';
```

156. State the differences between views and tables.

<u>Views</u>	<u>Tables</u>
A view is a virtual table that is extracted from a database	A table is structured with a set number of columns and a boundless number of rows
A view does not hold data itself	A table contains data and stores it in databases
A view is utilized to query certain information contained in a few distinct tables	A table holds fundamental client information and cases of a characterized object
In a view, we will get frequently queried information	In a table, changing the information in the database changes the information that appears in the view

157. What is the difference between BETWEEN and IN operators in SQL?

The BETWEEN operator is used to represent rows based on a set of values. The values may be numbers, text, or dates. The BETWEEN operator returns the total number of values that exist between two specified ranges.

The IN condition operator is used to search for values within a given range of values. If we have more than one value to choose from, then we use the IN operator.

158. What are the different types of a subquery?

There are two types of subquery namely, Correlated and Non-Correlated.

Correlated subquery: These are queries which select the data from a table referenced in the outer query. It is not considered as an independent query as it refers to another table and refers the column in a table.

Non-Correlated subquery: This query is an independent query where the output of subquery is substituted in the main query.

159. Name the operator which is used in the query for pattern matching?

LIKE operator is used for pattern matching, and it can be used as -.

% – It matches zero or more characters.

For example- select * from students where studentname like 'a%'

_ (Underscore) – it matches exactly one character.

For example- select * from student where studentname like 'abc_'

160. Write a SQL query to find the 10th highest employee salary from an Employee table

```
SELECT TOP (1) Salary FROM
(
    SELECT DISTINCT TOP (10) Salary FROM Employee ORDER BY Salary DESC
) AS Emp ORDER BY Salary
```

Or

```
SELECT DISTINCT Salary FROM Employee ORDER BY Salary DESC LIMIT 1 OFFSET 9;
```

161. Given a table dbo.users where the column user_id is a unique numeric identifier, how can you efficiently select the first 100 odd user_id values from the table?

```
SELECT TOP 100 user_id
FROM dbo.users
WHERE user_id % 2 = 1
ORDER BY user_id
```

162. What will be the output of the below query, given an Employee table having 10 records?

```
BEGIN TRAN
TRUNCATE TABLE Employees
ROLLBACK
SELECT * FROM Employees
```

This query will return 10 records as TRUNCATE was executed in the transaction. TRUNCATE does not itself keep a log but BEGIN TRANSACTION keeps track of the TRUNCATE command.

163. How do you get the last id without the max function?

```
select id from table order by id desc limit 1  
select top 1 id from table order by id desc
```

164. What is the difference between IN and EXISTS?

IN:

Works on List result set
Doesn't work on subqueries resulting in Virtual tables with multiple columns
Compares every value in the result list
Performance is comparatively SLOW for larger resultset of subquery

EXISTS:

Works on Virtual tables
Is used with co-related queries
Exits comparison when match is found
Performance is comparatively FAST for larger resultset of subquery

165. Which is better CTE or temp table?

If you will have a very large result set, or need to refer to it more than once, put it in a #temp table. If it needs to be recursive, is disposable, or is just to simplify something logically, a CTE is preferred. Also, a CTE should never be used for performance.

166. Why are CTEs slower than temp tables?

Sometimes the complexity of stacked CTEs just creates too many opportunities for the optimizer, and it can't reach one efficiently that would be best to satisfy the whole query. Sometimes out of date stats contribute to this, leading to bad choices

167. Write an SQL query to fetch the Emplds that are present in EmployeeDetails but not in EmployeeSalary.

Ans. Using subquery-

```
SELECT EmpId FROM  
EmployeeDetails  
where EmpId Not IN  
(SELECT EmpId FROM EmployeeSalary);
```

168. Write an SQL query to fetch the position of a given character(s) in a field.

Ans. Using the 'Instr' function-

```
SELECT INSTR(FullName, 'Snow')  
FROM EmployeeDetails;
```

169. Write a query to fetch only the first name(string before space) from the FullName column of the EmployeeDetails table.

Ans. In this question, we are required to first fetch the location of the space character in the FullName field and then extract the first name out of the FullName field.

For finding the location we will use the LOCATE method in MySQL and CHARINDEX in SQL SERVER and for fetching the string before space, we will use the SUBSTRING OR MID method.

MySQL – using MID

```
SELECT MID(FullName, 1, LOCATE(' ', FullName))  
FROM EmployeeDetails;
```

SQL Server – using SUBSTRING

```
SELECT SUBSTRING(FullName, 1, CHARINDEX(' ', FullName))  
FROM EmployeeDetails;
```

170. Write an SQL query to find the count of the total occurrences of a particular character – ‘n’ in the FullName field.

Ans. Here, we can use the ‘Length’ function. We can subtract the total length of the FullName field from the length of the FullName after replacing the character – ‘n’.

```
SELECT FullName,  
LENGTH(FullName) - LENGTH(REPLACE(FullName, 'n', ''))  
FROM EmployeeDetails;
```

171. Write an SQL query to fetch employee names having a salary greater than or equal to 5000 and less than or equal to 10000.

Ans. Here, we will use BETWEEN in the ‘where’ clause to return the EmpId of the employees with salary satisfying the required criteria and then use it as a subquery to find the fullName of the employee from the EmployeeDetails table.

```
SELECT FullName  
FROM EmployeeDetails  
WHERE EmpId IN  
(SELECT EmpId FROM EmployeeSalary  
WHERE Salary BETWEEN 5000 AND 10000);
```

172. Write an SQL query to fetch duplicate records from EmployeeDetails (without considering the primary key – EmpId).

Ans. In order to find duplicate records from the table, we can use GROUP BY on all the fields and then use the HAVING clause to return only those fields whose count is greater than 1 i.e. the rows having duplicate records.

```
SELECT FullName, ManagerId, DateOfJoining, City, COUNT(*)  
FROM EmployeeDetails  
GROUP BY FullName, ManagerId, DateOfJoining, City  
HAVING COUNT(*) > 1;
```

173. What is the difference between COALESCE() & ISNULL()?

COALESCE(): COALESCE function in SQL returns the first non-NULL expression among its arguments. If all the expressions evaluate to null, then the COALESCE function will return null.

Syntax:

```
SELECT column(s), COALESCE(expression_1, . . . , expression_n)
FROM table_name;
```

ISNULL(): The ISNULL function has different uses in SQL Server and MySQL. In SQL Server, ISNULL() function is used to replace NULL values.

Syntax:

```
SELECT column(s), ISNULL(column_name, value_to_replace)
FROM table_name;
```

174. Write SQL Query to find duplicate rows in a database? and then write SQL query to delete them?

```
SELECT * FROM emp a
WHERE rowid = (SELECT MAX(rowid)
FROM EMP b
WHERE a.empno=b.empno)
```

to Delete:

```
DELETE FROM emp a
WHERE rowid != (SELECT MAX(rowid) FROM emp b WHERE a.empno=b.empno);
```

175. There is a table which contains two columns Student and Marks, you need to find all the students, whose marks are greater than average marks i.e. list of above-average students.

Answer: This query can be written using subquery as shown below:

```
SELECT student, marks  
FROM table  
WHERE marks > SELECT AVG(marks) from table)
```

176. How to Show the Max marks and min marks together from student table?

Answer:

```
Select max (marks) from Student  
Union  
Select min (marks) from Student;
```

177. Give the order of SQL SELECT.

Answer: The order of SQL SELECT clauses is:

SELECT
FROM
WHERE
GROUP BY
HAVING
ORDER BY.

Only the SELECT and FROM clauses are mandatory.

178. What do we need to check in Database Testing?

Answer: In Database testing, the following thing is required to be tested:

- Database connectivity
- Constraint check
- Required application field and its size
- Data Retrieval and processing with DML operations
- Stored Procedures
- Functional flow

179. How many types of Privileges are available in SQL?

Answer: There are two types of privileges used in SQL, such as

System privilege: System privilege deals with the object of a particular type and provides users the right to perform one or more actions on it. These actions include performing administrative tasks, ALTER ANY INDEX, ALTER ANY CACHE GROUP creates/ALTER/DELETE TABLE, CREATE/ALTER/DELETE VIEW, etc.

Object privilege: This allows us to perform actions on an object or object of another user(s) viz. table, view, indexes, etc. Some of the object privileges are EXECUTE, INSERT, UPDATE, DELETE, SELECT, FLUSH, LOAD, INDEX, REFERENCES, etc.

180. What are the types of transaction levels in SQL SERVER?

There are four transaction levels in SQL SERVER.

- Read committed
- Read uncommitted
- Repeatable read
- Serializable

181. What is log shipping?

Answer: Log shipping is the process of automating the backup of database and transaction log files on a production SQL server, and then restoring them onto a standby server. Enterprise Editions only supports log shipping. In log shipping, the transactional log file from one server is automatically updated into the backup database on the other server.

182. What is a T-SQL?

T-SQL is an abbreviation for Transact Structure Query Language. It is a product by Microsoft and is an extension of SQL Language which is used to interact with relational databases. It is considered to perform best with Microsoft SQL servers. T-SQL statements are used to perform the transactions to the databases. T-SQL has huge importance since all the communications with an instance of an SQL server are done by sending Transact-SQL statements to the server. Users can also define functions using T-SQL.

Types of T-SQL functions are :

Aggregate functions.

Ranking functions. There are different types of ranking functions.

Rowset function.

Scalar functions.

183. What is query optimization?

Query optimization is the process of selecting the most efficient execution plan for a database query. When a database management system (DBMS) receives a query, it must determine the most efficient way to execute that query, taking into account factors such as the size of the data set, the available resources, and the query's complexity.

The goal of query optimization is to reduce the time and resources needed to execute a query, while ensuring that the results are accurate and complete. To achieve this, the DBMS analyzes the query and considers various execution strategies, such as which tables to access first, which indexes to use, and which join algorithms to apply.

Query optimization can be a complex and resource-intensive task, particularly for large or complex queries. However, optimizing queries can have a significant impact on the overall performance and efficiency of a database system. By selecting the most efficient execution plan for a query, the DBMS can reduce the amount of time and resources required to execute the query, and thus improve the overall responsiveness and throughput of the system.

184. What is the purpose of Query Optimization?

The major purposes of SQL Query optimization are:

1. Reduce Response Time: The major goal is to enhance performance by reducing the response time. The time difference between users requesting data and getting responses should be minimized for a better user experience.

2. Reduced CPU execution time: The CPU execution time of a query must be reduced so that faster results can be obtained.

3. Improved Throughput: The number of resources to be accessed to fetch all necessary data should be minimized. The number of rows to be fetched in a particular query should be in the most efficient manner such that the least number of resources are used.

185. What are the metrics for analyzing query performance for SQL Query Optimization?

1. Execution Time: The most important metrics to analyze the query performance is the execution time of the query. Execution time/Query duration is defined as the time taken by the query to return the rows from the database

2. Statistics IO - IO is the major time spent accessing the memory buffers for reading operations in case of query. It provides insights into the latency and other bottlenecks for executing the query. By setting STATISTICS IO ON, we get the number of physical and logical reads performed to execute the query

186. What is the difference between cross join and natural join?

The cross join produces the cross product or Cartesian product of two tables whereas the natural join is based on all the columns having the same name and data types in both the tables.

187. What is SQL Injection?

SQL injection is a sort of flaw in website and web app code that allows attackers to take control of back-end processes and access, retrieve, and delete sensitive data stored in databases. In this approach, malicious SQL statements are entered into a database entry field, and the database becomes exposed to an attacker once they are executed. By utilising data-driven apps, this strategy is widely utilised to get access to sensitive data and execute administrative tasks on databases. SQLi attack is another name for it.

The following are some examples of SQL injection:

Getting access to secret data in order to change a SQL query to acquire the desired results.
UNION attacks are designed to steal data from several database tables.
Examine the database to get information about the database's version and structure

188. What is BLOB and TEXT in MySQL?

BLOB stands for Binary Huge Objects and can be used to store binary data, whereas TEXT may be used to store a large number of strings. BLOB may be used to store binary data, which includes images, movies, audio, and applications.

BLOB values function similarly to byte strings, and they lack a character set. As a result, bytes' numeric values are completely dependent on comparison and sorting.

TEXT values behave similarly to a character string or a non-binary string. The comparison/sorting of TEXT is completely dependent on the character set collection.

189. Explain database white box testing and black box testing.

The white box testing method mainly deals with the internal structure of a particular database, where users hide specification details. The white box testing method involves the following:

As the coding error can be detected by testing the white box, it can eliminate internal errors. To check for the consistency of the database, it selects the default table values.

This method verifies the referential integrity rule.

It helps perform the module testing of database functions, triggers, views, and SQL queries.

The black box testing method generally involves interface testing, followed by database integration. The black box testing method involves the following:

Mapping details

Verification of incoming data

Verification of outgoing data from the other query functions

190. Explain the difference between OLTP and OLAP.

OLTP: It stands for online transaction processing, and we can consider it to be a category of software applications that are efficient for supporting transaction-oriented programs. One of the important attributes of the OLTP system is its potential to keep up the consistency. The OLTP system often follows decentralized planning to keep away from single points of failure. This system is generally designed for a large audience of end users to perform short transactions. The queries involved in such databases are generally simple, need fast response time, and, in comparison, return in only a few records. So, the number of transactions per second acts as an effective measure for those systems.

OLAP: It stands for online analytical processing, and it is a category of software programs that are identified by a comparatively lower frequency of online transactions. For OLAP systems, the efficiency of computing depends highly on the response time. Hence, such systems are generally used for data mining or maintaining aggregated historical data, and they are usually used in multidimensional schemas.

191. What is a stored procedure? Give an example.

A stored procedure is a prepared SQL code that can be saved and reused. In other words, we can consider a stored procedure to be a function consisting of many SQL statements to access the database system. We can consolidate several SQL statements into a stored procedure and execute them whenever and wherever required.

A stored procedure can be used as a means of modular programming, i.e., we can create a stored procedure once, store it, and call it multiple times as required. This also supports faster execution when compared to executing multiple queries.

192. What are the types of views in SQL?

In SQL, the views are classified into four types. They are:

Simple View: A view that is based on a single table and does not have a GROUP BY clause or other features.

Complex View: A view that is built from several tables and includes a GROUP BY clause as well as functions.

Inline View: A view that is built on a subquery in the FROM clause, which provides a temporary table and simplifies a complicated query.

Materialized View: A view that saves both the definition and the details. It builds data replicas by physically preserving them.

193. Mention different types of replication in SQL Server?

In SQL Server, three different types of replications are available:

Snapshot replication

Transactional replication

Merge replication

194. What is a data warehouse?

A data warehouse is a large store of accumulated data, from a wide range of sources, within an organization. The data helps drive business decisions.

195. What do you mean by data integrity?

Data integrity is the assurance of accuracy and consistency of data over its whole life cycle. It is a critical aspect of the design, implementation, and usage of systems that store, process, or retrieve data.

Data integrity also defines integrity constraints for enforcing business rules on data when it is entered into a database or application.

196. What is a No SQL database? What are some examples of NO SQL database?

If you are already working in the analytics domain then you might have heard words like PostgreSQL, MongoDB, MySQL, Oracle, etc. We also do have a fair bit of idea about the fact that we store our tables and databases in these places. But why can't we have everything under one hood?

We should have heard people saying that ABC is a NoSQL database and we should store these data in it rather than a traditional Relational database, what does that mean?

Does NoSQL mean that you don't have to write SQL on top of it? If yes, then what is the use? Let's try to understand NoSQL vs Relational Database

NoSQL stands for Not Only SQL

Not Only SQL – Every item in the database stands on its own

197. What is a Relational Database? What are some examples of Relational databases?

Relational databases are like predefined table structures where you can surely add as many rows as possible but adding a new column is a pain as you have to change the schema of the table, thus it is a vertically scaling database. Whereas NoSQL is built in such a way that you can add any number of rows but at the same time you can add any additional information pertaining to a row.

Example – Suppose you have an employee table and you have a NoSQL database, in this table, there will be only 2 columns i.e. a primary key and a values column where you can add as many details as possible, for one employee you can add only name and phone number whereas

for another employee you can add the complete profile ranging from name to spouse's work status, how?

In general, the value thing is stored in a JSON format which again is a key-value pair and you can add any number of attributes corresponding to an employee

Relational Database – Vertical scaling Ex. PostgreSQL, SQLite, MySql

NoSQL – Vertical and Horizontal scaling – Each item in the database only has two things – unique key and values

198. Which type of data is stored in NO SQL and Relational Databases?

Different formats in which you can store data in your NoSQL database

1. Document Database – JSON
2. Key-Value store
3. Graphical database

199. NO SQL is a vertically scalable or horizontally scalable database? What about Relational Database? How does this scaling make database life easier?

As discussed, Scaling in NoSQL can work in a horizontal way

Let's take an example – A company XYZ runs a NoSQL database which has 10000 servers, All these servers work as a partition i.e. your data which you want to fetch is not present in only one server, it can be present in any partition.

Won't it be too cumbersome for the database to check all the partitions to find a detail? How do find where is the data stored?

200. When should we use NO SQL and where should we use Relational Database?

NoSQL databases are key-value stores

It's not just a key-value pair – It is a key-hash-value where each key is associated with a hash value and this hash value works as a partition identifier 😊

Suppose the hash value is in the range of 0 to 1000 and all your data is stored in one DB. Now, if you want to double your DB performance, you can add another server, and now

Server 1 – Hash value 0 to 500

Server 2 – Hash value 501 to 1000

This range of 0 to 1000 is called a keyspace. Keyspace tells you where to store new items and where to find the existing ones.

NoSQL is schemaless – anyways you are storing everything in a JSON format and you can increase or decrease the amount of information in each field. In a relational database, you have to define the schema of the table and have to adhere to it

Example of NoSQL –

AWS – DynamoDB

Google Cloud – Big Table

Azure – CosmoDB

You can also run NoSQL DB yourself by using software like Cassandra, CouchDB, MongoDB.

201. Disadvantage of Temporary keyword in Presto?

In Presto, the "temporary" keyword can be used to create temporary tables or views that only exist for the duration of the session. While temporary tables can be useful for storing intermediate results or working with subsets of data, there are some potential disadvantages to using them.

One disadvantage of temporary tables is that they can increase the complexity of the query and make it more difficult to maintain or troubleshoot. Additionally, temporary tables can consume additional resources and storage space, especially if they are not dropped or cleaned up properly after the session.

Another potential disadvantage of temporary tables is that they may not be suitable for use in some distributed systems or parallel computing environments, where data is spread across multiple nodes or clusters. In these environments, temporary tables may not be able to be shared or accessed by all nodes, which can lead to inconsistent or incomplete results.

Finally, using temporary tables can also impact the performance of the system, especially if the tables are used excessively or if they are not optimized for the specific use case. Therefore, it is important to carefully consider the use of temporary tables in Presto and ensure that they are used appropriately and efficiently.

Row Number interview questions

201. What is row number in SQL?

Row number is a function in SQL that assigns a sequential integer value to each row in a result set. The row number function is used to generate a unique identifier for each row in the result set.

202. How is row number different from rank and dense rank?

Rank and dense rank are functions in SQL that also assign a unique value to each row in a result set based on the ordering of a specified column. However, row number simply assigns a sequential integer value to each row in the result set, without taking into account the ordering of any column.

203. What is the syntax for calculating row number in SQL?

The syntax for calculating row number in SQL is:

```
SELECT column1, column2, ..., ROW_NUMBER() OVER (ORDER BY ordering_column)
as row_number_column
FROM table_name;
```

204. Can you give an example of when you would use row number?

Row number can be used to generate a unique identifier for each row in a result set. For example, if you have a table of customer orders and you want to generate a unique order number for each order, you can use row number to assign a unique integer value to each row in the result set.

205. How can you use row number to group data?

Row number can be used to group data by adding a PARTITION BY clause to the query. The PARTITION BY clause specifies the column or columns that should be used to group the data, and the row number function is applied within each group. For example, if you have a table of employee salaries and you want to group the salaries by department, you can use row number with a PARTITION BY clause on the department column.

206. How do you reset the row number to 1 for each group in SQL?

To reset the row number to 1 for each group in SQL, you can use the PARTITION BY clause in conjunction with the row number function. The PARTITION BY clause specifies the column or columns that should be used to group the data, and the row number function is applied within each group. For example, if you have a table of employee salaries and you want to reset the row number to 1 for each department, you can use row number with a PARTITION BY clause on the department column.

207. Can you explain the performance implications of using row number?

Using row number can have performance implications for large data sets, as it requires SQL to scan the entire result set and assign a sequential integer value to each row. This can result in slower query performance compared to using simpler functions such as COUNT or SUM.

208. How would you handle NULL values when using row number?

By default, row number treats NULL values in the ordering column as distinct values and assigns them their own unique row number. If you want to exclude NULL values from the row numbering, you can add a WHERE clause to the query to filter the data before applying the row number function.

209. How can you use row number to perform pagination in SQL?

To perform pagination in SQL using row number, you can use the ROW_NUMBER() function with an OVER() clause to assign a row number to each row in the result set. You can then use a combination of the WHERE and BETWEEN clauses to filter and retrieve a specific subset of rows based on their row number.

210. How can you use row number to delete duplicate rows in SQL?

To delete duplicate rows in SQL using row number, you can use the ROW_NUMBER() function with an OVER() clause to assign a row number to each row in the result set. You can then use a common table expression (CTE) and the DELETE statement to delete the duplicate rows based on their row number.

Dense Rank interview questions

211. What is dense rank in SQL?

Dense rank is a function in SQL that assigns a unique rank value to each row in a result set, based on the ordering of a specified column. The dense rank function differs from the regular rank function in that it does not leave gaps in the ranking values when there are ties in the ordering column.

212. How is dense rank different from rank and row number?

Rank and row number are similar functions to dense rank in that they also assign a unique value to each row based on the ordering of a specified column. However, rank leaves gaps in the ranking values when there are ties in the ordering column, while row number simply assigns a sequential number to each row in the result set. Dense rank, on the other hand, assigns a unique value to each row, but does not leave gaps in the ranking values when there are ties.

213. What is the syntax for calculating dense rank in SQL?

The syntax for calculating dense rank in SQL is:

```
SELECT column1, column2, ..., dense_rank() OVER (ORDER BY ordering_column)
      as dense_rank_column
   FROM table_name;
```

214. Can you give an example of when you would use dense rank?

Dense rank can be used to assign a unique rank value to each row in a result set based on the ordering of a column, while also avoiding gaps in the ranking values when there are ties. For example, if you have a list of salespeople and their sales amounts, you can use dense rank to assign a unique ranking value to each salesperson based on their sales amounts, without leaving gaps in the ranking values for tied sales amounts.

215. How do you handle ties in dense ranking?

Dense rank automatically handles ties in the ordering column by assigning the same rank value to all tied rows, without leaving gaps in the ranking values.

216. Can you explain the difference between dense rank and percent rank?

Dense rank and percent rank are both functions used in SQL to assign a unique rank value to each row based on the ordering of a column. However, while dense rank assigns a unique integer value to each row without leaving gaps in the ranking values when there are ties, percent rank assigns a fractional value between 0 and 1 to each row based on its rank position relative to the total number of rows in the result set.

217. How can you use dense rank to group data?

Dense rank can be used to group data by assigning a unique rank value to each row based on the ordering of a column. For example, if you have a table of employee salaries and you want to

group the salaries into salary ranges based on their ranking, you can use dense rank to assign a unique ranking value to each salary, and then group the salaries by their ranking value.

218. How do you calculate dense rank for a subset of data in SQL?

To calculate dense rank for a subset of data in SQL, you can add a WHERE clause to the query to filter the data before applying the dense rank function. For example, if you want to calculate dense rank only for salespeople in a certain region, you can add a WHERE clause to the query to filter the data before applying the dense rank function.

219. Can you explain the performance implications of using dense rank?

Using dense rank can have performance implications for large data sets, as it requires SQL to scan the entire result set and sort it based on the ordering column. This can result in slower query performance compared to using simpler functions such as row number or rank.

220. How would you handle NULL values when using dense rank?

By default, dense rank treats NULL values in the ordering column as distinct values and assigns them their own unique ranking value. If you want to exclude NULL values from the ranking, you can add a WHERE

LIKE command interview questions

221. What is the LIKE command in SQL?

The LIKE command in SQL is a comparison operator used to search for specific patterns in a column of a table. It is commonly used with the WHERE clause in a SQL statement.

222. How does the LIKE command work?

The LIKE command matches a specified pattern in a column of a table. The pattern can contain special characters known as wildcards that represent one or more characters. The LIKE command compares each row in the column to the pattern and returns the rows that match the pattern.

223. What are the wildcards used in the LIKE command?

The wildcards used in the LIKE command are:

% (percent sign): represents zero, one, or multiple characters

_ (underscore): represents a single character

Can you give an example of using the LIKE command?

Yes, for example, the following SQL statement selects all the rows from the customers table where the country column starts with the letter 'U':

```
SELECT * FROM customers  
WHERE country LIKE 'U%';
```

224. How can you use the LIKE command with multiple patterns?

You can use the OR operator with multiple LIKE conditions to search for data that matches any of the patterns. For example, the following SQL statement selects all the rows from the customers table where the country column starts with 'U' or 'C':

```
SELECT * FROM customers  
WHERE country LIKE 'U%' OR country LIKE 'C%';
```

225. How can you use the LIKE command to search for data that contains special characters?

You can use the escape character (usually a backslash) before the special character to search for data that contains special characters. For example, the following SQL statement selects all the rows from the customers table where the city column contains a percentage sign (%):

```
SELECT * FROM customers  
WHERE city LIKE '%\%%';
```

226. Can you use the LIKE command with a NOT operator to exclude specific patterns?

Yes, you can use the NOT operator with the LIKE command to exclude specific patterns. For example, the following SQL statement selects all the rows from the customers table where the country column does not start with the letter 'U':

```
SELECT * FROM customers  
WHERE country NOT LIKE 'U%';
```

227. How can you use the LIKE command to search for data that starts with or ends with a specific pattern?

You can use the % wildcard to search for data that starts with or ends with a specific pattern. For example, the following SQL statement selects all the rows from the customers table where the city column ends with the letters 'ford':

```
SELECT * FROM customers  
WHERE city LIKE '%ford';
```

228. How can you improve the performance of a LIKE command query?

You can improve the performance of a LIKE command query by:

- Using a specific pattern rather than a generic one
- Using indexes on the column being searched
- Avoiding leading wildcards (e.g. '%value') as they cannot use indexes
- Using the full-text search feature if available

229. Can you use the LIKE command with parameters in a stored procedure?

Yes, you can use the LIKE command with parameters in a stored procedure. You can pass the pattern as a parameter and use it in the LIKE command. For example:

```
CREATE PROCEDURE search_customers  
    @pattern VARCHAR(50)  
AS  
BEGIN  
    SELECT * FROM customers  
    WHERE name LIKE '%' + @pattern + '%';  
END
```

Date Format Conversion in Presto

230. What is a regular expression and why would you use it in SQL?

A regular expression is a pattern of characters that can be used to match and manipulate text. In SQL, regular expressions can be used to search for, replace, or validate text data.

231. How do you use the REGEXP_LIKE function in SQL? Can you provide an example of when you might use this function?

The REGEXP_LIKE function in SQL is used to determine if a string matches a regular expression pattern. For example, you might use REGEXP_LIKE to check if a string contains a specific sequence of characters, such as a phone number or email address. Here's an example:

```
SELECT *
FROM employees
WHERE REGEXP_LIKE(email,
'^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}+$')
```

This query would return all employees whose email addresses match the regular expression pattern for a valid email address.

232. How do you use the REGEXP_REPLACE function in SQL? Can you provide an example of when you might use this function?

The REGEXP_REPLACE function in SQL is used to replace parts of a string that match a regular expression pattern with a new string. For example, you might use REGEXP_REPLACE to clean up messy data or to modify specific parts of a string. Here's an example:

```
SELECT REGEXP_REPLACE(description, '[^a-zA-Z0-9 ]', '') AS
cleaned_description
FROM products
```

This query would return a new column that contains the description of each product with all non-alphanumeric characters removed.

233. How do you use character classes in regular expressions? Can you provide an example of when you might use a character class in SQL?

Character classes in regular expressions are used to match any character from a specified set. For example, the expression [aeiou] would match any vowel. In SQL, you might use character classes to search for specific patterns of text within a larger string, such as searching for all words that start with a vowel. Here's an example:

```
SELECT *
FROM articles
WHERE REGEXP_LIKE(title, '^[aeiouAEIOU]\w+')
```

This query would return all articles whose titles begin with a vowel.

234. How do you use quantifiers in regular expressions? Can you provide an example of when you might use a quantifier in SQL?

Quantifiers in regular expressions are used to specify how many times a character or group should be matched. For example, the expression a{2,4} would match between 2 and 4 consecutive instances of the letter 'a'. In SQL, you might use quantifiers to search for repeated patterns within a string, such as searching for all phone numbers that contain exactly 10 digits. Here's an example:

```
SELECT *
FROM customers
WHERE REGEXP_LIKE(phone_number, '^\\d{10}$')
```

235. How do you use anchors in regular expressions? Can you provide an example of when you might use an anchor in SQL?

Anchors in regular expressions are used to match text at specific locations within a string. For example, the ^ anchor matches the beginning of a string, while the \$ anchor matches the end of a string. In SQL, you might use anchors to search for patterns that occur at specific positions within a string, such as searching for all strings that end with a specific word. Here's an example:

```
SELECT *
FROM documents
WHERE REGEXP_LIKE(text, 'marketing$')
```

236. How do you use alternation in regular expressions? Can you provide an example of when you might use alternation in SQL?

Alternation in regular expressions is used to match one pattern or another. For example, the expression cat|dog would match either the word 'cat' or the word 'dog'. In SQL, you might use alternation to search for multiple patterns within a string, such as searching for all articles that contain the word 'data' or 'analytics'. Here's an example:

```
SELECT *
FROM articles
WHERE REGEXP_LIKE(title, 'data|analytics')
```

237. How do you escape special characters in regular expressions? Can you provide an example of when you might need to escape a special character in SQL?

Special characters in regular expressions have a specific meaning, but sometimes you might need to match these characters literally. In SQL, you can use the backslash () character to escape a special character and match it literally. For example, if you want to match a period (.) character in a string, you can use the following regular expression:

```
SELECT *
FROM my_table
WHERE my_column REGEXP '\\.';
```

-- Note the double backslash to escape the period character
This will match any string that contains a period character.

239. How do you use regular expressions to validate input in SQL? Can you provide an example of how you might validate an email address or a phone number using regular expressions in SQL?

Regular expressions can be used to validate input in SQL by matching a string against a pattern. For example, you can use regular expressions to validate an email address or a phone number. Here are examples of regular expressions to validate an email address and a phone number in SQL:

Email address validation:

```
SELECT *
FROM my_table
WHERE my_column REGEXP '^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$';
```

This regular expression matches any string that starts with one or more alphanumeric characters, followed by an @ symbol, followed by one or more alphanumeric characters or hyphens, followed by a period character, followed by two or more alphabetic characters.

240. How can you use regular expressions to extract data from a string in SQL? Can you provide an example of how you might extract specific values from a string using regular expressions in SQL?

```
SELECT *
FROM my_table
WHERE my_column REGEXP '^\\+[0-9]{1,3}[-. ]?\\(?[0-9]{3}\\)\\)?[-. ]?[0-9]{3}[-. ]?[0-9]{4}$';
```

This regular expression matches any string that represents a valid phone number, including optional country code, area code, and separators.

241. How can you use regular expressions to extract data from a string in SQL? Can you provide an example of how you might extract specific values from a string using regular expressions in SQL?

Regular expressions can be used to extract data from a string in SQL by matching a substring that matches a pattern and extracting it using functions like REGEXP_SUBSTR. Here's an example of how you might extract a specific value from a string using regular expressions in SQL:

```
SELECT REGEXP_SUBSTR('Hello, my name is John.', 'my name is ([A-Za-z]+).',  
1, 1, 'i') as name;
```

This will extract the name "John" from the string "Hello, my name is John." by matching the pattern "my name is ([A-Za-z]+)." The regular expression includes a capture group ([A-Za-z]+) to extract the name, and the fourth argument to REGEXP_SUBSTR specifies the index of the capture group (1 in this case). The "i" flag in the last argument specifies that the regular expression should be case-insensitive.

JSON Interview Questions

242. What is JSON?

JSON (JavaScript Object Notation) is a lightweight data interchange format that is easy for humans to read and write and easy for machines to parse and generate.

243. How does Presto handle JSON data?

Presto has built-in support for querying JSON data. It can directly access JSON files stored on Hadoop Distributed File System (HDFS) or Amazon S3, and it can also query JSON data stored in Hive tables.

244. What is the syntax for querying JSON data in Presto?

In Presto, you can use the json_extract() function to extract data from a JSON object. The syntax is as follows:

```
SELECT json_extract(json_column, '$.field_name') FROM table_name
```

Here, json_column is the name of the column containing the JSON data, and field_name is the name of the field you want to extract.

245. How does Presto handle nested JSON data?

Presto can handle nested JSON data by using dot notation to navigate through the object hierarchy. For example:

```
SELECT json_extract(json_column, '$.parent_field.child_field') FROM  
table_name
```

This query would extract the value of the child_field that is nested inside the parent_field.

246. Can you join JSON data with other tables in Presto?

Yes, you can join JSON data with other tables in Presto. To do this, you can use the json_extract() function to extract the relevant fields from the JSON data, and then join on those fields as you would with any other table.

247. How does Presto handle JSON arrays?

Presto can handle JSON arrays by using the json_extract_scalar() function to extract scalar values from the array. For example:

```
SELECT json_extract_scalar(json_column, '$[0]')  
FROM table_name
```

This query would extract the first element of the JSON array stored in json_column.

248. How can you flatten nested JSON data in Presto?

To flatten nested JSON data in Presto, you can use the json_extract() function with the flatten() function to extract all fields at all levels of the JSON hierarchy.

For example:

```
SELECT flatten(json_extract(json_column, '$.*')) FROM table_name
```

This query would extract all fields at all levels of the JSON hierarchy stored in json_column.

249. How can you convert JSON data to a relational format in Presto?

To convert JSON data to a relational format in Presto, you can use the json_tuple() function to extract multiple fields from the JSON object and return them as separate columns.

For example:

```
SELECT json_tuple(json_column, '$.field1', '$.field2', '$.field3') FROM  
table_name
```

This query would extract the field1, field2, and field3 fields from the JSON object stored in json_column and return them as separate columns.

250. How can you use the json_table() function in Presto?

The json_table() function in Presto allows you to extract data from a JSON object and return it as a table. You can specify the structure of the output table using the COLUMNS clause.

For example:

```
SELECT * FROM json_table(json_column, '$.items[*]' COLUMNS (id varchar,  
name varchar))
```

This query would extract the id and name fields from an array of objects stored in json_column and return them as a table with columns id and name.

251. How can you use the json_arrayagg() function in Presto?

The json_arrayagg() function in Presto allows you to aggregate rows into a JSON array.

For example:

```
SELECT json_arrayagg(json_column) FROM table_name
```

This query would aggregate all rows in table_name into a JSON array.

Chapter 2 - Python

We have multiple books on Python floating around the internet. Some are good, but the majority is good in bits and pieces. In this book we want to cover the complete syllabus which you need before digging into any Machine Learning algorithm.

Once you are good with the following topics, you can go ahead with building models, which will again require some expertise. But, building a model without knowing the basics of a language is not a very favourable way to learn something.

Why do we want to go through everything, we can learn while we practice, right?

Correct, but going through the important topics and concepts will save a lot of time for you which you can utilise in building your favourite model.

We will start with very basics, will spend 200 questions on reaching to a decent state where you won't be uncomfortable with the types of questions generally asked to check you on Python.

Then we will spend 100 questions each on Pandas and Numpy to make you comfortable with the two libraries. Apart from that you will have a fair bit of practice in different machine learning algorithms. So, all in all you can expect to solve somewhere around 700 interview questions. Rest if you want to explore more then you can visit our website i.e. www.thedatamonk.com to practice more questions.

Don't get disheartened by seeing simple questions at the beginning, also don't expect a lot of definition questions as well. All the best, may these 400 questions are the go to questions for you :)

Also, you need to have at least some basics in Python or any coding language as the book will start with 2 on 10 rather than 0 on 10. If you want to grab the very basics then do check out something like tutorials point or w3school.com or any other coding website/youtube channel.

252. What is Python? What is the difference between Python and Java?

Python is a programming language. Things that make Python a really popular language is it's easy-to-learn syntax and its capability to build products in almost all the spheres of computer science. You can create softwares, websites, machine learning algorithms, games, etc. in Python.

It's very easy to define a variable in Python?

What is a variable in Python?

A variable is a place where you store a value. In other languages you need to specify the data type with the variable.

Java

```
Int x =  
10; Int  
y = 20;  
String s = "ABC"
```

But in Python, you don't have to specify the data type, you can directly define a variable like below

```
X = 10  
Y = 20  
print(X*Y)
```

Python is case sensitive y != Y

253. Sum of two numbers in Python

Let's start with taking sum of two numbers in Python, we will first take two variables and add it, easy

```
a = 10  
b = 20  
print("Sum = ",a+b)
```

```
In [1]: a = 10  
        b = 20  
        print("Sum = ",a+b)  
  
Sum = 30
```

254. How to take input from user?

Ans. I don't want to put the numbers in a variable rather I would like to take the input of two numbers from the user.

```
X = input("Enter first number = ")  
Y = input("Enter second number = ")
```

Done

```
X = input("Enter First Number = ")
Y = input("Enter Second Number = ")

Enter First Number = 40
Enter Second Number = 30
```

255. Take input from user and print the sum. This will test two concepts, first the concept of type casting and secondly how to print the sum and not concatente the numbers.

Ans. We already know how to take input from user and how to take the sum, why do you want to waste a question?

I will do the following and will get the sum

```
X = input("Enter First Number = ")
Y = input("Enter Second Number = ")
print("Sum = ", X+Y)
```

The result :-

```
X = input("Enter First Number = ")
Y = input("Enter Second Number = ")
print("Sum = ", X+Y)

Enter First Number = 20
Enter Second Number = 30
Sum = 2030
```

This is not what you wanted, right?

So, right now the variable X and Y has numbers stored as a string and the command X+Y will act as a concatenation and will print the two numbers side by side, this is not what we wanted right?

So, let's convert X and Y in integer and try again

```
X = input("Enter First Number = ")
Y = input("Enter Second Number = ")
a = int(X)
b = int(Y)
```

```
print("Sum = ", a+b)
```

```
X = input("Enter First Number = ")
Y = input("Enter Second Number = ")
a = int(X)
b = int(Y)
print("Sum = ", a+b)
```

```
Enter First Number = 20
Enter Second Number = 30
Sum = 50
```

So in three questions you understood a few things:-

1. How to declare variables?
2. How to add two variables?
3. How to take input from the user?
4. How to typecast the numbers?

Good start !!

256. What are the data types in Python?

Ans. Data Types in Python

It's very important to understand and remember the data types and how to define these in Python. It's a very common starter question in any interview

Python by default has 6 Data Types:-

- A. Number
- B. String
- C. List
- D. Tuple
- E. Dictionary
- F. Set
- G. Boolean

Data Frame and Arrays are also there but those are in Pandas and Numpy packages which we will discuss later.

257. What is number data type? What is type() function in Python?

Number:

- A. Integers - Example, X = 10
- B. Float - Example, X = 10.0
- C. Complex - Example, X = 10j

```
In [5]: #Data Type Number
X = 10
Y = 10.334
Z = 10j
print(type(X))
print(type(Y))
print(type(Z))

<class 'int'>
<class 'float'>
<class 'complex'>
```

Using the above example you can note that **type(variable)** can be used to find out the data type of the variable

258. What is Boolean data type in Python?

There is this data type which is boolean and stores only True or False

```
In [6]: #Boolean
x = 100 > 68
print(x)
print(type(x))

True
<class 'bool'>
```

Here we defined a condition which is true, so x stores a value i.e. True.

Data type of True is boolean, so when you do type(x) then it gives the output as Bool

259. What is string data type in Python?

We all know string, it's a sequence of characters. Now, unlike other languages, in Python you can declare string inside a single or double quote.

```
In [10]: #String
x = 'The Data Monk'
y = "The Data Monk"
z = input()

print(x,'oo',y,'pp',z)
print(type(x))
print(type(y))

23
The Data Monk oo The Data Monk pp 23
<class 'str'>
<class 'str'>

In [11]: print(type(z))

<class 'str'>
```

260. How to concat variables?

Separate the variables with commas, as show in screenshot

There are two important things to remember about String.

First, string in Python is a zero indexed data type i.e. the index of character starts with 0

Index of
KAMAL is K =
0
A = 1
M = 2
A = 3
L = 4

Length of string KAMAL is 5

```
In [13]: #Index number and length in a string variable
x = 'TheDataMonk'
print(x[2])
len(x)

e

Out[13]: 11
```

261. What do you mean when you say “Strings are immutable”?

Strings in Python are immutable i.e you can not change the defined string.
You can not change a part of the string, as it is immutable.

```
In [17]: #String are immutable
x = 'qwerty'
print(x[2])
x[2] = 'p'
print(x)

e

-----
TypeError                                 Traceback (most recent call last)
<ipython-input-17-155e9dc0538b> in <module>
      2 x = 'qwerty'
      3 print(x[2])
----> 4 x[2] = 'p'
      5 print(x)

TypeError: 'str' object does not support item assignment
```

String revision:-

3 very simple things to keep in mind, obviously you don't have

-In string the index is from 0, you can access the character of the string by specifying it within the two square brackets i.e x[3] - It will give you the 4th character

-You can also pull character from the reverse side, but here you do not have to follow index but the position x[-3] will give you the 3rd character from the right hand side

```
In [27]: #When you fetch the value from reverse side, the index number does not start from 0, it's the position which
#we are talking about
x = 'qwerty'
print(x[3])
print(x[-4])

r
e
```

-Strings are mutable i.e. you can not change a pre existing string, but you can change the complete variable

X = 'qwerty' X[2] = 'f'

The above code will throw an error as you are trying to change the already defined string. But you can write like this

```
X = 'Nitin'
X = 'Kamal'
print(X)
```

The above will print Kamal as it will reassign the variable to the new string.

262. What is a list in Python?

List is a set of data. Few very important points:-

1. List is defined within a square bracket and values are separated by a comma

```
X = ['Nitin',2,'Kamal',True]
```

2. A list can have different data types. In above example we have string, integer and boolean in the same list. There are very data types in other language which have this kind of property. This makes Python special

```
In [31]: #List - It is our 3rd Data Type
Team = ['RCB', 'MI', 'CSK']
print(Team[2])
print(Team[1:3])
print(type(Team))

CSK
['MI', 'CSK']
<class 'list'>
```

3. Like any other data type, the index of list starts from 0 only
4. You can have duplicate entries in a list as well
5. Team[1:3] will take the 2nd and 3rd value of the list. The same way we can extract multiple character from a string

```
X = 'Patna'
print(X[1:4])
```

Output = tna

```
print(X[2:2]) Output = t
```

263. Are lists mutable ?

Lists are mutable i.e. you can change the values already present in the list. It's a decent interview question to check if someone knows the basic

String is immutable
Lists are mutable

```
In [32]: #List - It is our 3rd Data Type
Team = ['RCB', 'MI', 'CSK']
print(Team[2])
print(Team[1:3])
print(type(Team))
Team[2] = 'Rajasthan Royals'
print(Team)

CSK
['MI', 'CSK']
<class 'list'>
['RCB', 'MI', 'Rajasthan Royals']
```

Since list is mutable, there are a few ways in which you can alter the values of a list

```
In [43]: #Alter the values of a list in different ways
Team = ['RCB', 'MI', 'CSK']
print(Team)
Team[2] = 'KKR'
print(Team)
Team.append('Punjab')
print(Team)
Team.insert(0, 'Pune')
print(Team)
Team.reverse()
print(Team)

['RCB', 'MI', 'CSK']
['RCB', 'MI', 'KKR']
['RCB', 'MI', 'KKR', 'Punjab']
['Pune', 'RCB', 'MI', 'KKR', 'Punjab']
['Punjab', 'KKR', 'MI', 'RCB', 'Pune']
```

264. What are some important functions used with list?

A. By directly specifying the index of the list

Team[2] = 'KKR' will replace the 3rd indexed value with KKR

B. By using append() function, append always adds the value at the end

C. use insert(index_number, Value)

D. Reverse the list, use the Team.reverse() command to first reverse the list and then print the list. You can not do

```
print(Team.reverse())
```

Again, you do not need to learn the syntaxes, just understand what is possible and what is not. Learn the basics like what is mutable, immutable, how to define list, append function, etc.

265. What is a dictionary in Python?

Now we are growing our knowledge base. List was the first step, but dictionary is a very good and a bit new concept for non-coding students. But, we will sail through :)

```
In [51]: #Dictionary - Let's conquer this
Team = {'MI':5,
        'CSK':4,
        'KKR':1,
        'RCB':0,
        44 : 56
       }
print(Team)
#print(Team[0])
print(Team['KKR'])
print(Team.get('CSK'))
print(type(Team))
Team['RCB'] = 99
print(Team)

{'MI': 5, 'CSK': 4, 'KKR': 1, 'RCB': 0, 44: 56}
1
4
<class 'dict'>
{'MI': 5, 'CSK': 4, 'KKR': 1, 'RCB': 99, 44: 56}
```

Few very important points:-

1. A dictionary is a set of key-value pair i.e. every value which you put in a dictionary need to have a key. A key is nothing but an index
2. A dictionary is defined inside curly braces {} and values are separated by comma
3. <key><Value>
4. A dictionary is mutable, you can change the value. But you need to write the index of the value Team['Danapur11'] = 432

This will add a value to your dictionary

5. There are two ways in which you can retrieve the value of a key i.e
 - A. by mentioning the key like Team['KKR']
 - B. by using the get() function, Team.get('KKR') - Both works the same way

266. Is dictionary zero indexed? Can we pull something like Team[0] from the above example?

The whole purpose of having a dictionary is that you can have your own index i.e. key. So, to answer the question, Dictionary is not zero indexed.

You can not use the basic index thing example, you can not use Team[0] to pull the first value because you have already specified an index to all the values

267. Lists and strings can have duplicate entries, but can a dictionary have duplicate entries like same key and value declared twice?

Dictionary can not have duplicate entries, in case there are duplicate entries, it will take the last one.

```
In [58]: dup = {'KKR': 00,
              'RCB': 33,
              'KKR': 22}
print(dup)

{'KKR': 22, 'RCB': 33}
```

Once again, Dictionary is a set of keys and values which can not have duplicate entries and values can be accessed by its key only

Before we move forward, let's make sure that you have understood the concepts. Answer the following questions

1. What are the datatypes of Python?
2. Is string mutable?
3. Is list mutable?
4. Is a dictionary mutable?
5. What is mutable and immutable?
6. How to define a string, list, and dictionary?
7. Can we use multiple data type values in a list and dictionary?
8. Two ways in which you can add a value to a list
9. Ways in which you can add value in a dictionary
10. What is the output for the following

```
S =
'TheDataMonk'
print(S[2])
print(S[-3])
print(S[2:5])
```

. Is duplicate entry possible in list and dictionary?

List = Yes,

Dictionary = No

Let's move to Tuple, a very important data type. There are only a few important data types i.e. dictionary, tuple, data frame (pandas) and array(numpy)

268. What is a tuple?

You thought immutable data types were over? No !! Tuple is immutable !!

Duplicate entries can be present in a Tuple.

```
In [59]: #Tuple - an immutable data type
x = ('The', 'Data', 2, 3, 4, 'Monk', True, 'Data', 'Data')
print(x[3])
print(type(x))
print(x.count('Data'))
```

3
<class 'tuple'>
3

269. What are the salient features of tuples?

A few important points to consider:-

1. Tuple is immutable i.e. you can not change values in a tuple
2. Tuple values are accessible by index numbers. See, you do not have to remember which data type is accessible by index and which are not. Only dictionary is not accessible by simple index numbers because we actually create a index inside the dictionary itself. Rest all are the same
3. There is a count function which can help count the number of elements in the tuple
4. Tuple can have duplicates. Only dictionary can not have duplicates because a dictionary have a key and value pair, how can the computer understand anything is two values have the same keys/index

Dup1 = {'a' = 22, 'b' = 22}

Dup2 = {'a' = 22, 'b' = 33, 'a' = 24}

Both these will get created, it's just that Dup2 will have only one value for the key 'a' i.e. 22 as it is the first one to appear

269. What is a set in Python?

Set is a revolutionary data type. Easy to understand and very helpful.

```
In [66]: #Set, unorder and no duplicate present. Set does not support indexing
stock = {"icici","hdfc",'sbi',"ruchi soya","icici"}
print(stock)
#print(stock[2])
for i in stock:
    print(i)

{'hdfc', 'ruchi soya', 'icici', 'sbi'}
hdfc
ruchi soya
icici
sbi
```

269. What are some salient features of set?

A few very important points for interviews:-

1. Set are unordered (we will come to in a minute)
2. Set can not have duplicate values
3. Set is not indexed i.e. you can not access an element of set using stock[2]. Set do not have a key also like dictionary, so how to access an element of set?

We need to write a loop to access elements of a set.

Writing a simple for look (we will also deal with loops in the coming chapter)

In Java we have a simple looking for loop for(int i = 0, i<=5 ; i++)

But in Python we define loop in a more concise manner, it might take you some practice to get used to it.

For i in stock: This statement means 'for each element of stock' starting with 0 print(i) - print each element one by one

270. A few data types are ordered and other are not ordered. What does that mean?

Ordered data types are those data types whose values are accessible by index and these values are always stored in the manner in which you write it. Like list, tuple and string

Unordered does not care about the way you write it, why?

Dictionary and set are unordered i.e. in dictionary you actually provide a hard coded index, so computer does not care about any ordering. A set is simply arrogant enough to consider any sort of ordering.

Still confused?

See the example below

```
In [62]: #Sets and dictionaries do not support indexing
letters = 'Nitin'

string_letters = str(letters)
lists_letters = list(letters)
tuples_letters = tuple(letters)
sets_letters = set(letters)

print("String: ", string_letters)
print() # for new line
print("Lists: ", lists_letters)
print() # for new line
print("Tuples: ", tuples_letters)
print() # for new line
print("Sets: ", sets_letters)

String: Nitin

Lists: ['N', 'i', 't', 'i', 'n']

Tuples: ('N', 'i', 't', 'i', 'n')

Sets: {'n', 'N', 't', 'i'}
```

It's just that I can type the name 'Nitin' and 'The Data Monk' quicker than other names, it saves some time (no obsession with the names)

Set gives random result, removing the duplicate. Remember when I told you that Python is case sensitive? Thus n and N are treated differently.

Revision Time

Ordered - String, List and Tuple

Unorder - Dictionary and Set

Mutable - List, Sets and Dictionary

Immutable - String, Tuple

Indexed - This is synonym of ordered i.e. String, List and Tuple

Now we will deal with some specific functions, loops, and creating user defined functions. All these things play a very vital role in your Data Science journey.

271. What is the function range() ?

Range(10) will get you numbers from 0 to 9. But you need to put this range in some data type. Suppose you want to put this in a list.

```
In [68]: print(range(10))
print(list(range(0,10)))
print(range(2))

range(0, 10)
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
range(0, 2)
```

See, you have created a list of numbers from 0 to 10

2. Can we create a list with different data types?? Confused?

Basically can we have a list,dictionary, string, and set inside a list?

Let's try

```
In [71]: dic2 = {1,2,3,4}
str2 = "string"
list2 = ['q','w','e','r','t']
set2 = {"kbc","Tu-Tu Mai Mai"}
tup2 = (123,456,765)

list_final = [dic2,str2,list2,set2,tup2]
list_final

Out[71]: [{1, 2, 3, 4},
 'string',
 ['q', 'w', 'e', 'r', 't'],
 {'Tu-Tu Mai Mai', 'kbc'},
 (123, 456, 765)]
```

We defined a simple dictionary,string,list,set,tupple and then put it in one list. This can be done. You can clearly see the advantages of list as well as Python.

You do not have to worry a lot about data types, you can create a huge dataset with different type of data.

Interview question?

Q. If you do not want to change the data of a variable, which data type should you use?

A. Use set in this case

272. What is type conversion?

Type Conversion

```
list()  
set()  
dict()  
tuple()  
float()  
int()
```

Above are the functions which are used to convert a variable in some other data type. Why do we need it?

Because you can have similar operations on different data types. Example below

```
In [72]: x = '2'  
y = 3  
print(x+y)  
  
-----  
TypeError Traceback (most recent call last)  
<ipython-input-72-f6b8f74c965c> in <module>  
      1 x = '2'  
      2 y = 3  
----> 3 print(x+y)  
      4 x = int(x)  
      5 print(x+y)  
  
TypeError: can only concatenate str (not "int") to str  
  
In [73]: x = int(x)  
print(x+y)  
5
```

Similarly you can change one type of data to another as per your need.

273. What is collections in Python?

Collections are basically the container data type i.e. List, Tuple, Dictionary, and Set.

Set - Square bracket, mutable, duplicate values can be stored, and indexed i.e. elements can be accessed using index number

Tuple - Ordered and immutable in nature

Set - Unordered and declared in curly braces, unindexed. You can use loop to access data. No duplicate entry present

Dictionary - Key-Value pair and is mutable in nature

These are the 4 different containers present in Python by default. But there are downsides of each of the data types. To sort this out, Python has a Collection module that contains a specialized collection data structures. Below are the types of collections. We will not go deep into this as you will need it for a specific purpose.

Let's solve some questions that are readily asked in product-based companies like Moonfrog Labs, Zynga, Ola, Housing.com , etc.

These are some standard questions asked in the interviews to check the competency of the candidate in the basic logics and syntax of Python. Solve each of these questions on your own in Jupyter notebook.

We will make sure that you are ready for any sort of companies in the futures. These questions are available on almost all the coding platform but so is other 1000 coding questions, you just need to do the questions that are put over here.

Let's start with the logical questions

274. Write a Python program to get the factorial of any number.

```
a = 10

def fact(n):
    if (n == 1 or n==0):
        return 1
    else:
        return n*fact(n-1)
print(fact(5))
```

Or

```
import math
print(math.factorial(5))
```

275. Get the square of a number if the number is odd and cube if even

```
def sq(n):
```

```
if (n%2 == 0):
    return n*n*n
else:
    return n*n

print(sq(3))
print(sq(2))
```

276. Sum of square of first n natural number

```
x = int(input("Enter a natural number = "))

def sum_of_natural(x):
    ss = 0
    for i in range(1,x):
        ss = ss+(i*i)
    return ss

print("Sum of square of natural numbers are = " , sum_of_natural(x))
```

277. Check if a number is Armstrong number in Python. An armstrong number is a number that equals the sum of cube of all the digits of the number. Ex. 153 = $1^3+5^3+3^3$

```
num = int(input("Enter a number = "))
s = 0
x = num
while (x >0 ):
    digit = x%10
    s = s+(digit*digit*digit)
    x = x//10
print(s)

print("Armstrong" if s == num else "Not Armstrong")
```

278. Write a program to get the area of a rectangle (to know if the candidate understands the concept of function)

```
def area(l,b):
    return 2*l*b
```

```
a = int(input("Enter Length = "))
b = int(input("Enter breadth = "))
print(area(a,b))
```

279. Check if a number is a prime number in Python.

```
x = int(input("Enter a number = "))

if x > 1:
    for i in range(2,int(x/2)+1):
        if (x%i == 0):
            print("Not a Prime Number")
            break
    else:
        print("Prime Number")
else:
    print("Not a prime number, infact it is a negative number")
```

280. Write a program to get the Prime number in a list of numbers with starting and end point

```
x = int(input("Enter a starting point = "))
y = int(input("Enter an ending point = "))

def prime(x,y):
    prime = []
    for i in range(x,y):
        if (i == 0 or i == 1):
            continue
        else:
            for j in range(2,int(i/2)+1):
                if(i%j == 0):
                    break
            else:
                prime.append(i)
    return prime
print("The list of prime numbers are = ", prime(x,y))
```

281. Print the nth number in the fibonacci series.

Fibonacci Series 0,1,1,2,3,5,8,13...

```
x = int(input("Enter the n-th fibonacci series number = "))
```

```
def fib(x):
    if x<=0:
        print("Incorrect number")
    elif x==1:
        return 0
    elif x==2:
        return 1
    else:
        return fib(x-1)+fib(x-2)

print("Fibonacci number on the n-th place is ", fib(x))
```

282. Check if a given number is a perfect square or not

```
n = int(input("Enter a number = "))
```

```
def sq(n):
    s = int(math.sqrt(n))
    return s*s == n
```

```
print("The status of number = " , sq(n))
```

283. Print ASCII value of a character in python

```
x = input("Take a character")
```

```
print("The ASCII Value of the character is ", ord(x))
```

A few questions on Arrays in Python.

284. Finding sum of elements of array

```
a = [1,4,5,7]
s = 0
```

```
for i in a:
    s = s+i;
```

```
print("Sum of array is = ",s)
```

285. Largest element of an array

```

a = [1,5,7,3,4,9,18,222]
num = len(a)

def lar(a):
    largest = a[0]
    for i in range(2,num):
        if largest <= a[i]:
            largest = a[i]
    return largest

print("The largest element is = ",lar(a))

```

286. Rotate an array

```

x = [1,2,5,6,7]
y = len(x)
z = []
for i in range(0,(y)):
    z.append(x[y-i-1])

print(z)

```

287. Split the array at a particular point

```

a = [1,4,6,12,54]
d = 3
x = []
y = []
for i in range(0,len(a)):
    if i<d:
        x.append(a[i])
    else:
        y.append(a[i])

print(x)
print(y)
print(x+y)

```

288. Create an array using array library

```

import array
ar = array.array('i',[1,2,3])

```

```
for i in range(0,3):
    print(ar[i])
```

289. What is append(), pop() and remove() function for array?

```
ar.append(44)
for i in range(0,len(ar)):
    print(ar[i])
```

Append adds 44 at the very end of the array ar

```
ar.pop(2)
```

pop() method takes the index and removes the element from that position. In the above ar array the elements are 1,2,3,44

So, it will remove the element 3 from the array

```
ar.remove(44)
```

remove () method is used to remove the element by value. So, the above function will remove the value 44 from the array

290. How can you reverse an array?

```
Ar = array.array('i',[1,2,3,4])
```

```
ar.reverse()
```

```
for i in range(0,len(ar)):
    print(ar[i])
```

The reverse() function can directly be used to reverse the elements of an array

Basic questions to practice lists

291. Interchange first and last element in a list

```
l = [1,2,3,5,6]
```

```
temp = l[0]
l[0] = l[len(l)-1]
l[len(l)-1] = temp
print(l)
```

292. Use another method or way to interchange the first and the last elements of a list in Python. This was asked in Flipkart interview in 2021

```
l = [4,5,6,7,8]
print("List",l)

def inter(l):
    l[0],l[-1] = l[-1],l[0]
    return l

print("Interchanged List" , inter(l))
```

293. In a list, Swap two positions in a list (Meesho interview question)

```
p1 = 2
p2 = 5
l = [1,2,3,4,5,6,7]
print("Original List",l)

def swap(l,p1,p2):
    l[p1],l[p2] = l[p2],l[p1]
    return l

print("Swaped list", swap(l,p1,p2))
```

294. Get the length of a list and check if a particular element is present in the list?

To check if a particular element is there in the list

```
l = ['a','b','d']
count = l.count('d')
print(count)
```

Length of the list

```
A = len(l)
```

295. Convert a list in a set and use in function

```
l_set = set(l)
if 'b' in l_set:
    print("Mil gaya")
```

296. Use a loop to find if an element is present in a set or not?

Use a naive method of loop

```
for i in range(0,len(l)):
    if l[i] == 'd':
        print("Found at index = ",i)
```

297. Reverse a list

```
website = ['GreatLearning', 'Udacity', 'The Data Monk']
print('Original List:', website)
```

```
# Reversing a list
# Syntax: reversed_list = systems[start:stop:step]
reversed_list = website[::-1]
```

```
# updated list
print('Updated List:', reversed_list)
```

```
#5b Name
website = ['GreatLearning', 'Udacity', 'The Data Monk']
print('Original List:', website)

# Reversing a list—*
# Syntax: reversed_list = systems[start:stop:step]
reversed_list = website[::-1]

# updated list
print('Updated List:', reversed_list)
```

Original List: ['GreatLearning', 'Udacity', 'The Data Monk']
Updated List: ['The Data Monk', 'Udacity', 'GreatLearning']

298. sum of elements in a list

```
l = [1,3,5,6]
```

```
s = 0
```

```
for i in range(0,len(l)):
    s = s+l[i]
```

```
print(s)
```

299. Another method to print the sum of all the elements in the list

```
list1 = [9,8,7,6]
```

```
def sumOfList(list, size):  
    if (size == 0):  
        return 0  
    else:  
        return list[size - 1] + sumOfList(list, size - 1)
```

```
total = sumOfList(list1, len(list1))
```

```
print("Sum of all elements in given list: ", total)
```

300. Multiple all the numbers in a list

```
l = [1,3,5,6]
```

```
m = 1
```

```
for i in range(0,len(l)):  
    s = s*l[i]
```

```
print(s)
```

301. Count occurrence of an element in a list

```
def ct(l, x):  
    count = 0  
    for i in l:  
        if (i == x):  
            count = count + 1  
    return count
```

```
# Driver Code  
l = [8, 6, 8, 10, 8, 20, 10, 8, 8]  
x = 8  
print('{} has occurred {} times'.format(x, countX(l, x)))
```

302. Cumulative sum in a list

```
I = [1,2,3,4]
x = []

s = 0

for i in range(0,len(I)):
    s = s+I[i]
    x.append(s)

print(x)
```

303. Break a list in N parts using the yield clause

```
I = [1,2,3,4,5,7,8,9]
N = 3

def br(I,N):
    for i in range(0,len(I),N):
        yield I[i:i+N]

x = list(br(I,N))
print(x)
```

String practice questions

304. Reverse a word

```
x = "amit"
w = ""

for i in x:
    w = i+w
print("reverse word = ",w)
```

305. Check if a string is palindrome. Palindrome example - Nitin or 12321

```
str = input("Enter a string")

def pal(str):
    return str == str[::-1]
```

```
print("String is a palindrome ? = ",pal(str))
```

2nd method

```
def pal_2(str):
    for i in range(0,int(len(str)/2)):
        if str[i] != str[len(str)-i-1]:
            return False
    return True
```

```
str = input("Take a name ")
a = pal_2(str)
print(a)
```

```
def pal_2(str):
    for i in range(0,int(len(str)/2)):
        if str[i] != str[len(str)-i-1]:
            return False
    return True

str = input("Take a name ")
a = pal_2(str)
print(a),
```

```
Take a name nitin
True
```

Or

3rd method

```
x = input("A name ")
st = ""

for i in x:
    st = i+st

if x == st:
    print("Palindrome")
else:
    print("Not a palindrome")
```

306. check if a string is present in another string

```
x = "The Data Monk"
```

```
y = "Dataaa"
```

```
print(y in x)
```

```
#check if a string is present in another string
x = "The Data Monk"
y = "Dataaa"

print(y in x)
False
```

Or Use find method to check is a string is present in another string

```
def check(string, sub_str):
    if (string.find(sub_str) == -1):
        print("NO")
    else:
        print("YES")
```

```
string = "The Data Data Monk"
sub_str ="Data"
check(string, sub_str)
```

307. How to get word frequency in a particular sentence using collections?

```
from collections import Counter
```

```
sent = "The Data Monk will help you in making a career in Data Science"
print(sent)
```

```
res = Counter(sent.split())
print(res)
```

```
print("The word frequency is = " + dict(res))
```

```
#Word frequency in a string

from collections import Counter

sent = "The Data Monk will help you in making a career in Data Science"
print(sent)

res = Counter(sent.split())
print(res)

print("The word frequency is = " + dict(res))

The Data Monk will help you in making a career in Data Science
Counter({'Data': 2, 'in': 2, 'The': 1, 'Monk': 1, 'will': 1, 'help': 1, 'you': 1, 'making': 1, 'a': 1, 'career': 1,
'Science': 1})
```

308. What is the one big advantage of set that you can use to remove duplicate alphabets from a string?

Remember set never contains a duplicate value, so if you put ‘Nitin’ in a string and convert this string into a set followed by printing the set, then all you will get is nit that too in a jumbled order(unless specified)

```
s = "nitin"
x = set(s)
print(x)

{'i', 't', 'n'}
```

309. What is the use of the following command

`"".join(str)`

It joins two adjacent elements in iterable with any symbol defined in "" (double quotes) and returns a single string

`x = "The Data Monk"" is a website"`

`y = "".join(x)`

`print(x)`

```
x = "The Data Monk"" is a website"
y = "".join(x)

print(x)
```

The Data Monk is a website

310. Using the above two questions remove all the duplicate characters from a string

```
from collections import OrderedDict

def duplicate(str):
    return "".join(set(str))

def remove_dupe(str):
    return "".join(OrderedDict.fromkeys(str))

if __name__ == "__main__":
    st = "The data monk and the monk"
    print("Without order = ", duplicate(st))
    print("Ordered = ", remove_dupe(st))
```

```
from collections import OrderedDict

def duplicate(str):
    return "".join(set(str))

def remove_dupe(str):
    return "".join(OrderedDict.fromkeys(str))

if __name__ == "__main__":
    st = "The data monk and the monk"
    print("Without order = ", duplicate(st))
    print("Ordered = ", remove_dupe(st))
```

```
Without order = hamtTeonkd
Ordered = The datmonk
```

311. Split a sentence in all the words

```
x = "The Data Monk is a website"
print(x.split(" "))
```

```
#Split a sentence in all the words
x = "The Data Monk is a website"
print(x.split(" "))
```

```
['The', 'Data', 'Monk', 'is', 'a', 'website']
```

312. Find words with length greater than 3

```
x = "The Data Monk is a website"
```

```

d = 3

def extract(x,d):
    t = x.split(" ")
    fin = []
    for i in t:
        if len(i)>d:
            fin.append(i)
    return fin

print(extract(x,d))

```

```

#Find words with length greater than 3

x = "The Data Monk is a website"
d = 3

def extract(x,d):
    t = x.split(" ")
    fin = []
    for i in t:
        if len(i)>d:
            fin.append(i)
    return fin

print(extract(x,d))

```

['Data', 'Monk', 'website']

313. rotate a string at a particular

```

s = "The Data Monk"
x = ""

for i in range(0,len(s)):
    x = s[i]+x
print(x)

```

```

#rotate a string at a particular

s = "The Data Monk"
x = ""

for i in range(0,len(s)):
    x = s[i]+x
print(x)

```

knoM ataD ehT

314. Print words of even length

```
x = "The Data Monk is a website"
```

```
def even(x):
    t = x.split(" ")
    ev = []
    for i in t:
        if (len(i) % 2 == 0):
            ev.append(i)
    return ev
```

```
print(even(x))
```

```
#Print words of even length

x = "The Data Monk is a website"

def even(x):
    t = x.split(" ")
    ev = []
    for i in t:
        if (len(i) % 2 == 0):
            ev.append(i)
    return ev

print(even(x))
```

```
['Data', 'Monk', 'is']
```

Let's try out some questions on Dictionary

315.

Print all the keys of a dictionary

```
d = {'a':100,'b':500,'c':1000}
```

```
def s(d):
    for i in d:
        print(i)
```

```
s(d)
```

```
#Print all the keys of a dictionary
d = {'a':100,'b':500,'c':1000}

def s(d):
    for i in d:
        print(i)

s(d)
```

a
b
c

316. Print all the values of a dictionary

```
d = {'a':100,'b':500,'c':1000}
```

```
def s(d):
    for i in d:
        print(d[i])
```

```
s(d)
```

```
#Print all the values of a dictionary
d = {'a':100,'b':500,'c':1000}

def s(d):
    for i in d:
        print(d[i])

s(d)
```

100
500
1000

317. Sum of all the values of a dictionary

```
d = {'a':100,'b':500,'c':1000}
```

```
def s(d):
    l = []
    for i in d:
        l.append(d[i])
    su = sum(l)
    return su
```

```
print("Sum of the values in the dictionary is. = ", s(d))
```

318. Remove a key from dictionary

```
d = {'a':100,'b':500,'c':1000}
print(d)
del d['b']
print(d)
```

```
#Remove a key from dictionary
print(d)
del d['b']
print(d)
```

```
{'a': 100, 'b': 500, 'c': 1000}
{'a': 100, 'c': 1000}
```

319. Merge two dictionaries

```
a = {'x':12,'y':34}
b = {'z':34,'a':76}
```

```
def merge(a,b):
    return(a.update(b))
```

```
print(merge(a,b))
print(a)
```

320. Flatten a dictionary

```
d = {'name': ['Nitin','Kamal','Afeem'],
      'code': [1,6,12]}
print(d)
```

```
from itertools import product
```

```
flat = dict(zip(d['name'],d['code']))
print(str(flat))
```

```

# Flatten a dictionary

d = {'name': ['Nitin', 'Kamal', 'Afeem'],
      'code': [1, 6, 12]}
print(d)

from itertools import product

flat = dict(zip(d['name'], d['code']))

print(str(flat))

```

{'name': ['Nitin', 'Kamal', 'Afeem'], 'code': [1, 6, 12]}
{'Nitin': 1, 'Kamal': 6, 'Afeem': 12}

Tuple is a collection of Python objects much like a list.

The sequence of values stored in a tuple can be of any type, and they are indexed by integers.

321. How to define a tuple and get the size of tuple variable

```

tup = {"amit", 33, "sachin", 100, "virat", "amitabh"}
print(tup)
import sys
print("Size of the variable tup is =", sys.getsizeof(tup))

```

#How to define a tuple and get the size of tuple variable

```

tup = {"amit", 33, "sachin", 100, "virat", "amitabh"}
print(tup)
import sys
print("Size of the variable tup is =", sys.getsizeof(tup))

```

{33, 100, 'amitabh', 'amit', 'virat', 'sachin'}
Size of the variable tup is = 728

322. Create a tuple with the 4 elements, then create a list of tuple with cube of each element

```
tup = [1, 2, 3, 4]
```

```
a = [(i, pow(i, 3)) for i in tup]
```

```
print(a)
```

```
tup = [1, 2, 3, 4]
a = [(i, pow(i, 3)) for i in tup]
print(a)
[(1, 1), (2, 8), (3, 27), (4, 64)]
```

323. Adding tuple to a list

```
a = [1, 2, 3, 4]
b = (5, 6)

#Adding tuple to a list
print(b+tuple(a))

#Adding list to a tuple
print(a+list(b))
```

```
#Adding tuple to a list
a = [1, 2, 3, 4]
b = (5, 6)

#Adding tuple to a list
print(b+tuple(a))

#Adding list to a tuple
print(a+list(b))
```

```
(5, 6, 1, 2, 3, 4)
[1, 2, 3, 4, 5, 6]
```

324. Extract digits from a tuple.

```
from itertools import chain

list_1= [(97,98),(4,93),(0,1)]

# Extract digits from Tuple list by using map() and chain.from_iterable()
# We have used set() to remove the duplicate values

x = map(lambda a: str(a), chain.from_iterable(list_1))
```

```

res = set()
for sub in x:
    for a in sub:
        res.add(a)

# printing result
print("The extracted digits : " + str(res))

```

325. Removeing list of length 2 from the tuple

```
list_1 = [(1,2,3),(3,4,5),(2,3),(7,8,9)]
```

```
print ("The original list is = ",list_1 )
```

```
# Put the value of k
K = 2
```

```
res = [i for i in list_1 if len(i) != K]
```

```
# Final list
```

```
print("Final list : " + str(res))
```

```
#Removeing list of length 2 from the tuple
list_1 = [(1,2,3),(3,4,5),(2,3),(7,8,9)]
print ("The original list is = ",list_1 )
# Put the value of k
K = 2
res = [i for i in list_1 if len(i) != K]
# Final list|
print("Final list : " + str(res))
```

```
The original list is = [(1, 2, 3), (3, 4, 5), (2, 3), (7, 8, 9)]
Final list : [(1, 2, 3), (3, 4, 5), (7, 8, 9)]
```

326. Flatten tuple of a list

```
tup_1 = ([1,2,3], [4,5,6,7], [8])
```

```
# printing original tuple
print("The original tuple : " , (test_tuple))
```

```
# Flatten tuple of List to tuple using sum() and tuple() method
res = tuple(sum(tup_1, []))

# printing result
print("The flattened tuple : " + str(res))
```

```
tup_1 = ([1,2,3], [4,5,6,7], [8])

# printing original tuple
print("The original tuple : " , (test_tuple))

# Flatten tuple of List to tuple using sum() and tuple() method
res = tuple(sum(tup_1, []))

# printing result
print("The flattened tuple : " + str(res))
```

```
The original tuple : ([5, 6], [6, 7, 8, 9], [3])
The flattened tuple : (1, 2, 3, 4, 5, 6, 7, 8)
```

327. What is split() function in Python ?

The split() function breaks a string based on set criteria.
You can use it to split a string value from a web form.
Or you can even use it to count the number of words in a piece of text.

```
word = "The Data Monk is a website"
```

```
word = word.split(" ")
print(word)
```

```
# split()
word = "The Data Monk is a website"
word = word.split(" ")
print(word)

['The', 'Data', 'Monk', 'is', 'a', 'website']
```

328. What is reduce() function in Python?

Python's reduce() function iterates over each item in a list, or any other iterable data type, and returns a single value. It's one of the methods of the built-in functools class of Python.

```
from functools import reduce
def add_str(a,b):
```

```
return a+'+b  
a = ['the','Data','Monk', 'is', 'a', 'website']  
print(reduce(add_str, a))
```

```
from functools import reduce  
def add_str(a,b):  
    return a+' '+b  
a = ['the','Data','Monk', 'is', 'a', 'website']  
print(reduce(add_str, a))
```

the Data Monk is a website

329. What is eval() ?

eval() function in Python is a very important function that is used to evaluate a mathematical operation even in its string format

```
A = "3*5-4"  
print(A)  
print(eval(A))
```

```
a = "4/6*5"  
print(a)  
print(eval(a))
```

4/6*5
3.333333333333333

330. What is enumerate() ?

The enumerate() function returns the length of an iterable and loops through its items simultaneously. Thus, while printing each item in an iterable data type, it simultaneously outputs its index.

Assume that you want a user to see the list of items available in your database. You can pass them into a list and use the enumerate() function to return this as a numbered list.

```
cricket_player = ['Sachin','Ganguly','Dravid The God']
for i, j in enumerate(cricket_player):
    print(i, j)
```

```
cricket_player = ['Sachin','Ganguly','Dravid The God']
for i, j in enumerate(cricket_player):
    print(i, j)|
```

```
0 Sachin
1 Ganguly
2 Dravid The God
```

331. What is round() in Python?

You can round up the result of a mathematical operation to a specific number of significant figures using round()

```
raw_average = (4/3+5/2+7/3)
rounded_average=round(raw_average, 2)
print("The raw average is:", raw_average)
print("The rounded average is:", rounded_average)
```

332. What is max() in Python ?

The max() function returns the highest ranked item in an iterable. Be careful not to confuse this with the most frequently occurring value, though.

```
b = {1:"amit", 2:"sumit", 3:"alpha", 4:"beta", 5:"zamma"}
print(max(b.values()))
```

```
b = {1:"amit", 2:"sumit", 3:"alpha", 4:"beta", 5:"zamma"}
print(max(b.values()))
```

```
zammaa
```

```
c = [1,2,3,4]
print(max(c))
```

The above code will print 4

333. What is min() in Python?

```
b = {1:"amit", 2:"sumit", 3:"alpha", 4:"beta", 5:"zamma"}  
print(min(b.values()))
```

```
c = [1,2,3,4]  
print(min(c))
```

```
b = {1:"amit", 2:"sumit", 3:"alpha", 4:"beta", 5:"zamma"}  
print(min(b.values()))
```

alpha

```
c = [1,2,3,4]  
print(min(c))
```

1

334. What is a map() function in Python ?

the map() function lets you iterate over each item in an iterable, map() operates on each item independently. You can use it to manipulate an array containing any data type.

```
b = [9, 32, 14, 116]  
a = [14, 5, 72, 12]  
# add to add the two  
def add(a, b):  
    return a+b  
# Pass the function and the two lists into the map()  
a = sum(map(add, b, a))  
print(a)
```

```
b = [9, 32, 14, 116]  
a = [14, 5, 72, 12]  
# add to add the two  
def add(a, b):  
    return a+b  
# Pass the function and the two lists into the map()  
a = sum(map(add, b, a))  
print(a)
```

274

335. What is getattr() in Python?

Python's getattr() returns the attribute of an object. It accepts two parameters: the class and the target attribute name.

336. What is append() in Python?

It works by writing new data into a list without overwriting its original content.

```
a = [1,2,3]
b = [4,5]
```

```
for i in a:
    c = i*3
    b.append(c)
print(b)
```

```
a = [1,2,3]
b = [4,5]

for i in a:
    c = i*3
    b.append(c)
print(b)|
```

```
[4, 5, 3, 6, 9]
```

337. What is range() in Python?

Create a list of integers ranging between specific numbers without explicitly writing them out.

```
for i in range(0,6):
    print(i)
```

```
for i in range(0,8,2):
    print(i)
```

The third argument in the above program is the step size. So, it will take i as 0, then take the second step, then the fourth and so on.

```
for i in range(0,6):
    print(i)
```

0
1
2
3
4
5

```
for i in range(0,8,2):
    print(i)
```

0
2
4
6

338. What is the use of slice() function in Python?

Slice() as the name suggests is used to cut or slice the elements on the index value.

```
a = "The Data Monk"
b = [1,2,3,4,5]

print(a[slice(0,4)])
print(b[slice(0,3)])
```

```
a = "The Data Monk"  
b = [1,2,3,4,5]  
  
print(a[slice(0,4)])  
print(b[slice(0,3)])|
```

The
[1, 2, 3]

339. What is the format() command in Python?

format() is a very useful command in Python that is used to fill the values or variable in a string.

```
a = "The Data Monk"  
b = "website"  
  
c = "{} is a {}"  
c = c.format(a,b)  
print(c)
```

```
a = "The Data Monk"  
b = "website"  
  
c = "{} is a {}"  
c = c.format(a,b)  
print(c)|
```

The Data Monk is a website

340. What is strip() method in Python ?

Python's strip() removes leading characters from a string. It repeatedly removes the first character from the string, if it matches any of the supplied characters.

```
a = " The Data Monk "  
print(a)  
print(a.strip())
```

```
a = " The Data Monk "
print(a)
print(a.strip())
```

The Data Monk
The Data Monk

341. What is the use of abs() function in Python?

abs() function is used to make the negative numbers or computations

```
ab = -45
print(abs(ab))
```

```
b = 30
print(abs(ab-b))
```

```
ab = -45
print(abs(ab))

b = 30
print(abs(ab-b))
```

45
75

342. What is the use of upper() and lower() function in Python?

upper() and lower() function in Python is used to convert all the alphabet in lower and upper case simultaneously. This is very much needed when you don't know all the cases in a particular column, example suppose in one of the columns the country name is India and you are not sure if all the cases in the country name is the same, it could be india, India,inDia,etc. So, it is best to keep the country name in lower or upper case and then check

Like lower(countr_name) like 'india'

```
a = "nltin"  
a = a.lower()  
b = a.upper()  
print(a)  
print(b)
```

343. What is the join() function in Python?

The join() function lets you merge string items in a list.

```
a = ["The", "Data", "Monk"]  
a = " ".join(a)  
print(a)
```

```
a = ["The", "Data", "Monk"]  
a = " ".join(a)  
print(a)
```

The Data Monk

344. What is replace() function in Python?

Python's replace() method lets you replace some parts of a string with another character.

```
a = "Upgrad is the best platform for analytics job preparation"  
a = a.replace("Upgrad","The Data Monk")  
print(a)
```

```
a = "Upgrad is the best platform for analytics job preparation"  
a = a.replace("Upgrad","The Data Monk")  
print(a)
```

The Data Monk is the best platform for analytics job preparation

345. What is sorted() function in Python?

The sorted() function works by making a list from an iterable and then arranging its values in descending or ascending order:

```
f = {1, 5, 11, 2}
sort = {"A":8, "Z":5, "X":9, "T":3}
#Descending
print(sorted(f, reverse=True))
print(sorted(sort.values()))
```

```
f = {1, 5, 11, 2}
sort = {"A":8, "Z":5, "X":9, "T":3}
#Descending
print(sorted(f, reverse=True))
print(sorted(sort.values()))
```

```
[11, 5, 2, 1]
[3, 5, 8, 9]
```

346. Write a Python program to print set of duplicates in a list.

```
list=[1,2,3,1,5,6,2,5,7]
print(set[x for x in list if list.count(x)>1])) Output: {1,2,5}
```

347. Write a Python program to extract all the digits from a text

```
import re

def extract_digits(input_string):
    # use regular expression to match digits
    digits = re.findall(r'\d', input_string)
    return digits

# example usage
input_string = "The Data Monk has 7989 members and 2000+ questions on
www.TheDataMonk.com"
digits = extract_digits(input_string)
print(digits)
```

```

import re

def extract_digits(input_string):
    # use regular expression to match digits
    digits = re.findall(r'\d', input_string)
    return digits

# example usage
input_string = "The Data Monk has 7989 members and 2000+ questions on www.TheDataMonk.com"
digits = extract_digits(input_string)
print(digits)

['7', '9', '8', '9', '2', '0', '0', '0']

```

In this program, the `extract_digits` function takes an input string as an argument and returns a list of all the digits in the string. The function uses the `re.findall` method to search for all occurrences of digits in the input string using a regular expression.

The regular expression `r'\d'` matches any digit character in the input string. The `findall` method returns all the matches as a list.

Visualization in Python

Library to use – There are a lots of good visualization libraries, but `matplotlib` library is the most preferred one to start with because of its simple implementation.

So, We will mostly concentrate on `matplotlib` library.

Importing the library and giving it the standard alias as `plt`.

```
import matplotlib.pyplot as plt
```

Following are the two important functions which will come handy in this book:-

To display a chart you should use – `plt.show()`

To save the chart as an image, use the code – `plt.savefig("Filename.png")`

348. What are the most used package for visualization in Python?

Popular plotting libraries in Python are:-

1. `Matplotlib` – Best to start with. It provides easy implementation and gives a lot of freedom
2. `Seaborn` – It has a high level interface and great default styles
3. `Plotly` – To create interactive plots
4. `Pandas Visualization` – Easy interface, built on `Matplotlib`

349. How to create a line chart?

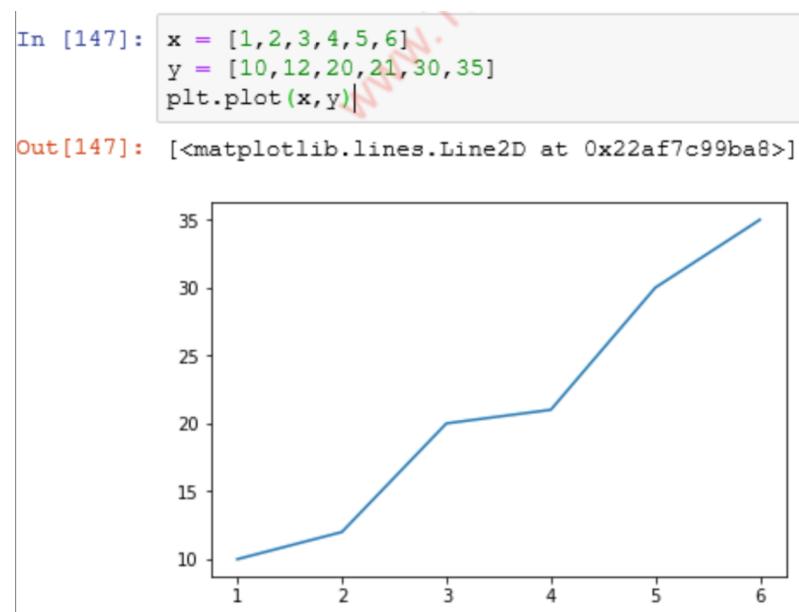
A line chart or line graph is a type of chart which displays information as a series of data points called ‘markers’ connected by straight line segments.

So, a line plot is a very basic plot which is used to show observations collected after a regular interval. The x-axis represents the interval and the y-axis represents the values.

Lets plot our first graph

```
import matplotlib.pyplot as plt
x = [1,2,3,4,5,6]
y = [10,12,20,21,30,35]
plt.plot(x,y)
```

Here is what you will get



Graph 1 – Basic Line Chart

350. Plot a sin graph using line plot

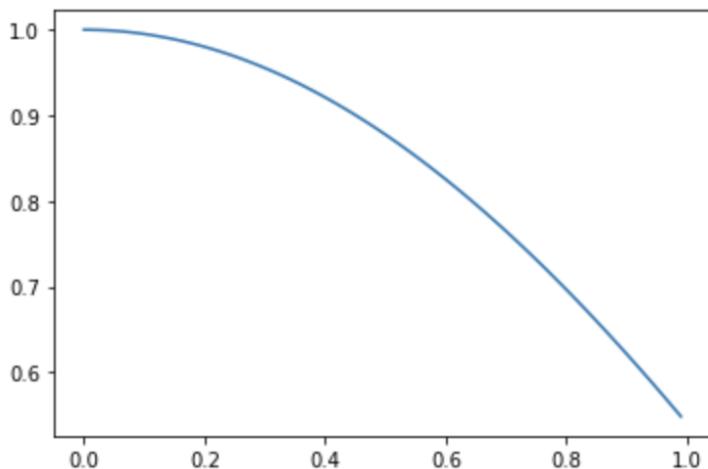
```
import matplotlib.pyplot as plt
from numpy import cos
x = [x*0.01 for x in range(100)] y = cos(x)
plt.plot(x,y)
plt.show()
```

Here is what you get as a cos graph

```

import matplotlib.pyplot as plt
from numpy import cos
x = [x*0.01 for x in range(100)]
y = cos(x)
plt.plot(x,y)
plt.show()

```



351. Graph 2 – Cos graph using line plot

You know how to plot a line graph, but there is one important thing missing in the graph i.e. the x and y-axis, and the plot title. Let's create another line plot for number of students in a class for the following data

c = [1,2,3,4,5,6]

student = [40,52,50,61,70,78]

Following commands are used to put x-axis label, y-axis label, and chart title

```

plt.xlabel("Label")
plt.ylabel("Label")
plt.title("Title")

```

The code is given below

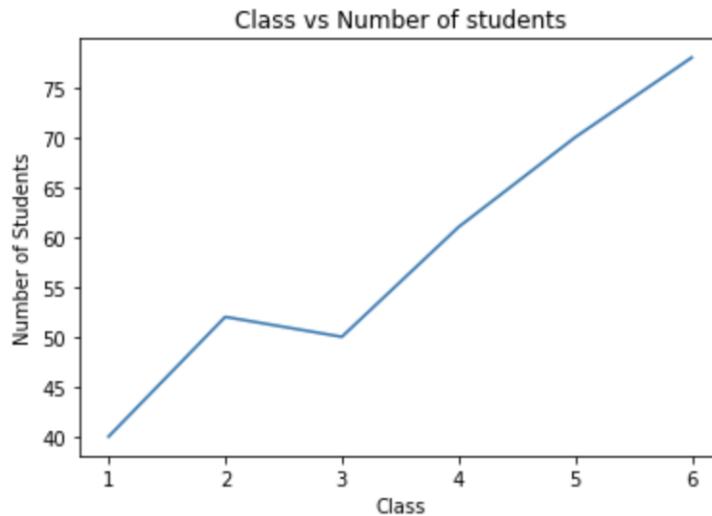
```

c = [1,2,3,4,5,6]
student = [40,52,50,61,70,78]
plt.xlabel("Class")
plt.ylabel("Number of Students")
plt.title("Class vs Number of students")
plt.plot(c, student)

```

```
In [12]: c = [1,2,3,4,5,6]
student = [40,52,50,61,70,78]
plt.xlabel("Class")
plt.ylabel("Number of Students")
plt.title("Class vs Number of students")
plt.plot(c, student)

Out[12]: [<matplotlib.lines.Line2D at 0x7fdb6b0ad550>]
```



352. Class vs Number of Students chart with proper labels and plot title

You want to change the color of the line?

Try the following code instead to make the line green in color

```
plt.plot(c,student,color='g')
```

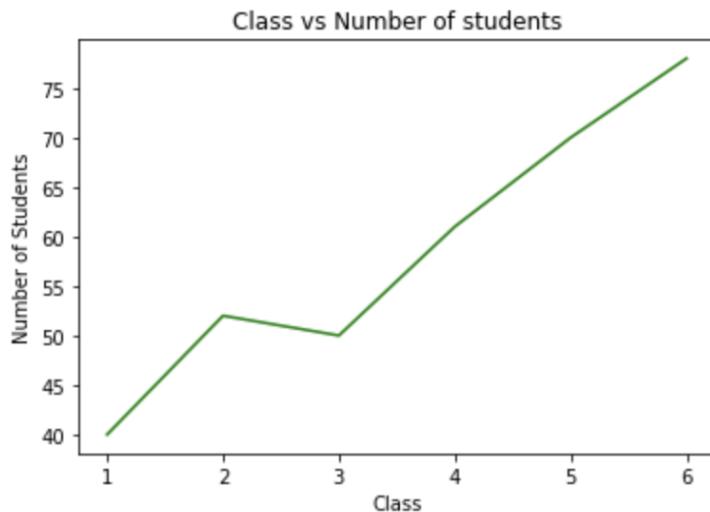
Graph 4 – Adding color to the same graph

```

c = [1,2,3,4,5,6]
student = [40,52,50,61,70,78]
plt.xlabel("Class")
plt.ylabel("Number of Students")
plt.title("Class vs Number of students")
plt.plot(c,student,color='g')

```

[<matplotlib.lines.Line2D at 0x7fdb6b13f460>]



353.

Multi Line Chart

Graph 5 – Adding multiple lines to a graph

To add a legend, you have to give label to each of the line which you want to plot and after that you specify a location to the legend

The code is self explanatory and is given below:-

```

c = [1,2,3,4,5,6]
student = [40,52,50,61,70,78]
avg_marks = [34,43,54,44,50,55]
num_of_teachers = [10,12,13,10,15,10]
plt.xlabel("Class")
# plt.ylabel("Number of Students")
plt.title("Class vs Number of students")
plt.plot(c,student,color='orange',label='Student')
plt.plot(c,avg_marks,color='red',label='Marks')
plt.plot(c,num_of_teachers,color='green',label='Teachers')
plt.legend(loc="upper left")

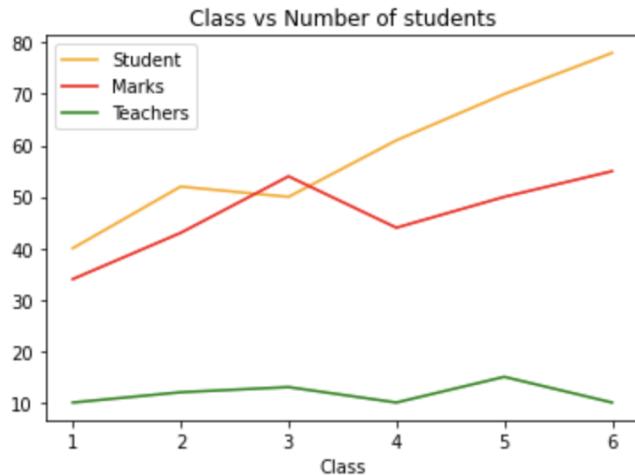
```

```

c = [1,2,3,4,5,6]
student = [40,52,50,61,70,78]
avg_marks = [34,43,54,44,50,55]
num_of_teachers = [10,12,13,10,15,10]
plt.xlabel("Class")
# plt.ylabel("Number of Students")
plt.title("Class vs Number of students")
plt.plot(c,student,color='orange',label='Student')
plt.plot(c,avg_marks,color='red',label='Marks')
plt.plot(c,num_of_teachers,color='green',label='Teachers')
plt.legend(loc="upper left")

```

<matplotlib.legend.Legend at 0x7fdb681df670>



354. Bar Chart

"A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally."

After the line chart, the second basic but highly used chart is the bar chart To create a bar chart – plt.bar(x,y)

We will plot few graphs first and then you can put labels, title, and legends later.

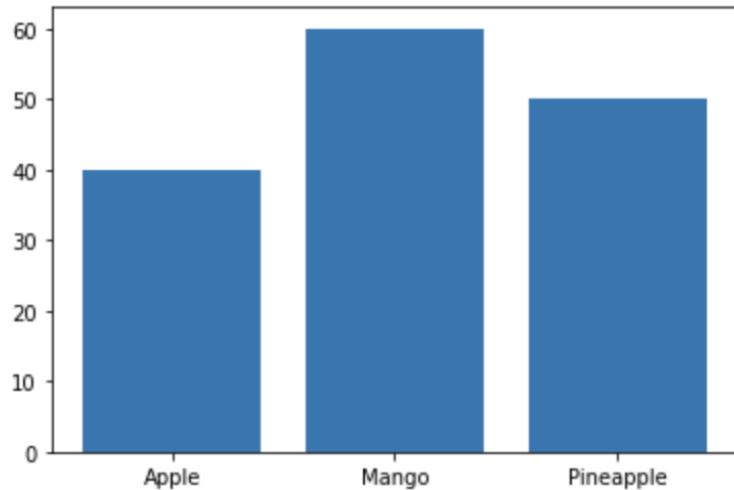
```

import matplotlib.pyplot as plt
a = ['Apple','Mango','Pineapple']
b = [40,60,50]
plt.bar(a,b)

```

```
a = ['Apple','Mango','Pineapple']
b = [40,60,50]
plt.bar(a,b)
```

```
<BarContainer object of 3 artists>
```



Graph 6 – A simple bar chart

355. Use random values between 1 and 100 to create the same graph.

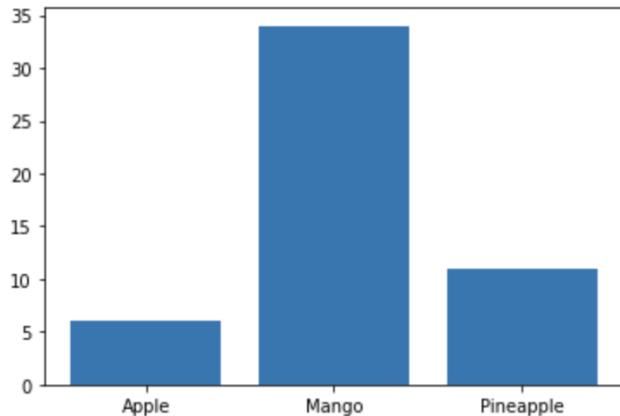
```
import matplotlib.pyplot as plt
from random import seed
from random import randint
seed(123)
x = ['Apple','Mango','Pineapple']
y = [randint(0,100),randint(0,100),randint(0,100)]
plt.bar(x,y)
```

```

import matplotlib.pyplot as plt
from random import seed
from random import randint
seed(123)
x = ['Apple', 'Mango', 'Pineapple']
y = [randint(0,100), randint(0,100), randint(0,100)]
plt.bar(x,y)

<BarContainer object of 3 artists>

```



356. Bar chart with random values

Adding color, labels, and title to the random values bar chart

Stacked 100% bar chart with sub component

When you have to show components of components like the graph below

Example of 100% bar chart

```

Import numpy as np
x = ["a","b","c","d"]
y1 = np.array([3,8,6,4])
y2 = np.array([10,2,4,3])
y3 = np.array([5,6,2,5])
snum = y1+y2+y3
# normalization
y1 = y1/snum*100.
y2 = y2/snum*100.
y3 = y3/snum*100.
plt.figure(figsize=(4,3))
# stack bars
plt.bar(x, y1, label='y1')
plt.bar(x, y2 ,bottom=y1,label='y2')
plt.bar(x, y3 ,bottom=y1+y2,label='y3')

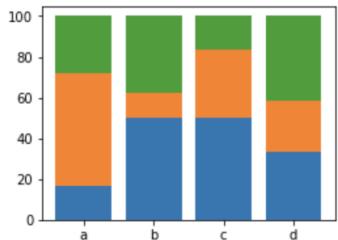
```

```

import numpy as np
x = ["a", "b", "c", "d"]
y1 = np.array([3,8,6,4])
y2 = np.array([10,2,4,3])
y3 = np.array([5,6,2,5])
snum = y1+y2+y3
# normalization
y1 = y1/snum*100.
y2 = y2/snum*100.
y3 = y3/snum*100.
plt.figure(figsize=(4,3))
# stack bars
plt.bar(x, y1, label='y1')
plt.bar(x, y2 ,bottom=y1,label='y2')
plt.bar(x, y3 ,bottom=y1+y2,label='y3')

```

<BarContainer object of 4 artists>



357. A 100% stacked bar chart

4. Histogram

Histograms are density estimates. A density estimate gives a good impression of the distribution of the data. The idea is to locally represent the data density by counting the number of observations in a sequence of consecutive intervals (bins).

To plot a histogram use this code –

```
plt.hist(x,y)
```

A simple histogram plot

```
q = [1,2,34,5,44,66,66,90,33,45,2,1,2,3,4]
```

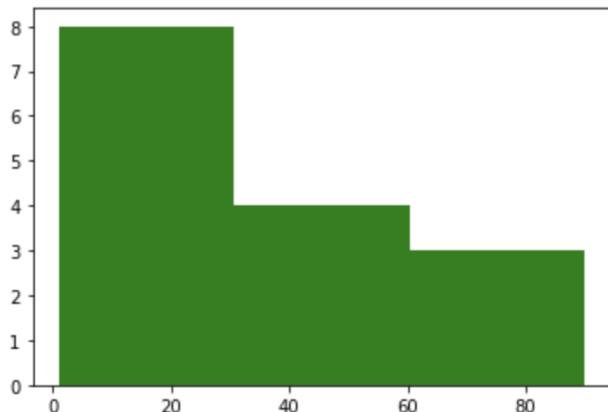
```
plt.hist(q,bins = 3,color='green')
```

```

q = [1,2,34,5,44,66,66,90,33,45,2,1,2,3,4]
plt.hist(q,bins = 3,color='green')

(array([8., 4., 3.]),
 array([ 1.          , 30.66666667, 60.33333333, 90.        ]),
 <BarContainer object of 3 artists>
)

```



358. A simple histogram - create a list using random variables and plot it in 4 bins

```

import random
my_rand = random.sample(range(1,30),20)
print(my_rand)
print(type(my_rand))
plt.hist(my_rand,bins=4,color='orange')

```

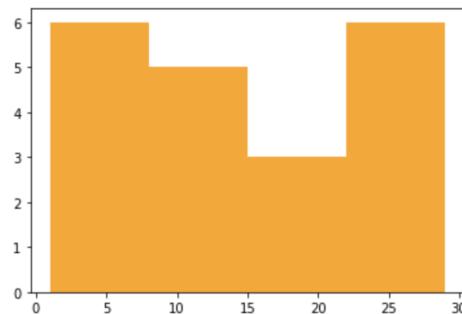
```

import random
my_rand = random.sample(range(1,30),20)
print(my_rand)
print(type(my_rand))
plt.hist(my_rand,bins=4,color='orange')

[25, 14, 9, 4, 2, 13, 18, 23, 11, 21, 29, 6, 5, 20, 27, 22, 12, 26, 3, 1]
<class 'list'>

(array([6., 5., 3., 6.]),
 array([ 1., 8., 15., 22., 29.]),
 <BarContainer object of 4 artists>
)

```



359. A histogram made with random variables

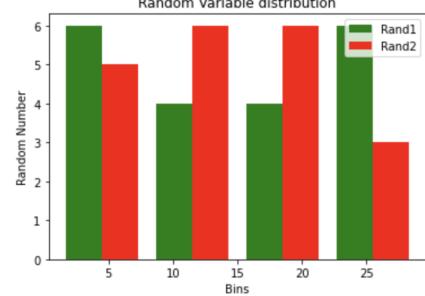
In Histogram also you can add more than one data points to make parallel bars.

```
import random
my_rand = random.sample(range(1,30),20)
my_rand2 = random.sample(range(1,25),20)
print(my_rand)
print(type(my_rand))
plt.hist([my_rand,my_rand2],bins=4,color=['green','red']) legend = ['Rand1','Rand2']
plt.legend(legend)
plt.xlabel("Bins")
plt.ylabel("Random Number")
plt.title("Random Variable distribution")
```

```
import random
my_rand = random.sample(range(1,30),20)
my_rand2 = random.sample(range(1,25),20)
print(my_rand)
print(type(my_rand))
plt.hist([my_rand,my_rand2],bins=4,color=['green','red'])
legend = ['Rand1', 'Rand2']
plt.legend(legend)
plt.xlabel("Bins")
plt.ylabel("Random Number")
plt.title("Random Variable distribution")
```

```
[29, 14, 25, 3, 20, 13, 26, 1, 11, 15, 4, 2, 23, 5, 17, 24, 22, 16, 7, 10]
<class 'list'>
```

```
Text(0.5, 1.0, 'Random Variable distribution')
```



360. Parallel histogram

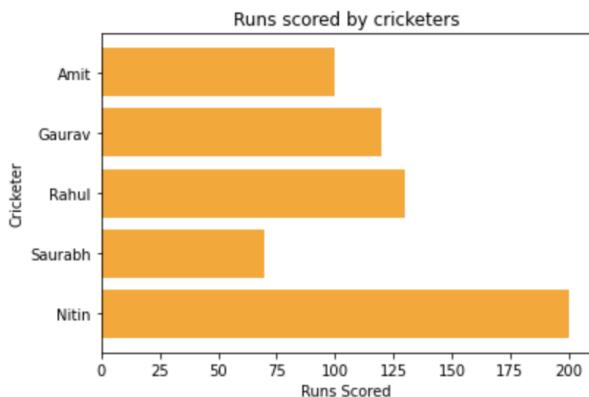
Horizontal Histogram

```
import numpy as np
import matplotlib.pyplot as plt
name = ['Nitin','Saurabh','Rahul','Gaurav','Amit']
run = [200,70,130,120,100]
plt.barh(name,run,color='orange')
plt.xlabel("Runs Scored")
plt.ylabel("Cricketer")
plt.title("Runs scored by cricketers")
plt.show()
```

```

import numpy as np
import matplotlib.pyplot as plt
name = ['Nitin', 'Saurabh', 'Rahul', 'Gaurav', 'Amit']
run = [200, 70, 130, 120, 100]
plt.barr(name, run, color='orange')
plt.xlabel("Runs Scored")
plt.ylabel("Cricketer")
plt.title("Runs scored by cricketers")
plt.show()

```



361. A horizontal histogram

Line Histogram

Now let's create a line histogram with some random data

```

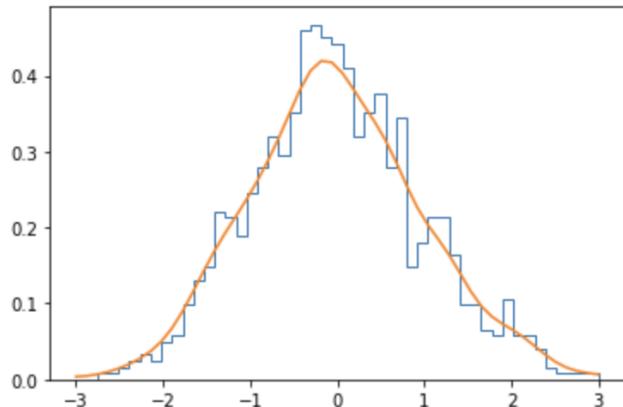
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats
noise = np.random.normal(0, 1, (1000, ))
density = stats.gaussian_kde(noise)
n, x, _ = plt.hist(noise, bins=np.linspace(-3, 3, 50), histtype=u'step', density=True)
plt.plot(x, density(x))
plt.show()

```

```

import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats
noise = np.random.normal(0, 1, (1000, ))
density = stats.gaussian_kde(noise)
n, x, _ = plt.hist(noise, bins=np.linspace(-3, 3, 50), histtype='step', density=True)
plt.plot(x, density(x))
plt.show()

```



362. A line histogram

Variable Width histogram

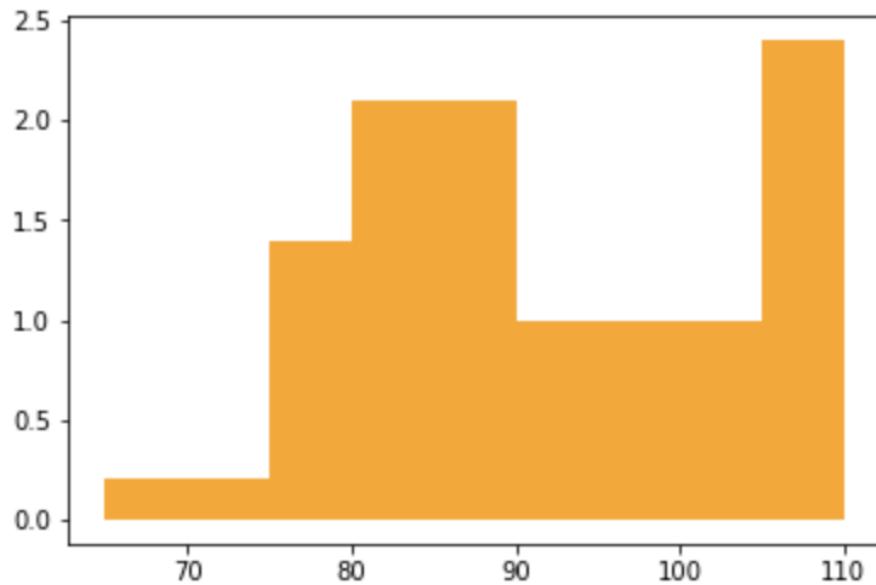
This is how a variable column width histogram looks like

Let's create one with our dataset

```

import numpy as np
import matplotlib.pyplot as plt
freqs = np.array([2, 7, 21, 15, 12])
bins = np.array([65, 75, 80, 90, 105, 110])
widths = bins[1:] - bins[:-1]
heights = freqs.astype(np.float)/widths
plt.fill_between(bins.repeat(2)[1:-1], heights.repeat(2), facecolor='orange')
plt.show()

```



363. A variable width histogram

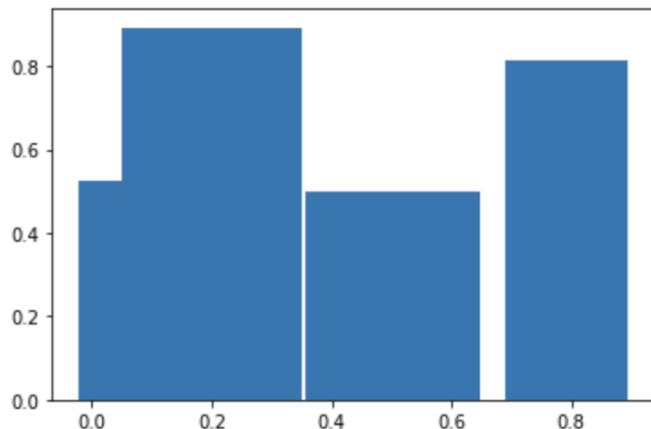
One more example below

```
import numpy as np
import matplotlib.pyplot as plt
x = np.sort(np.random.rand(6))
y = np.random.rand(5)
plt.bar(x[:-1], y, width=x[1:] - x[:-1])
plt.show()
```

```

import numpy as np
import matplotlib.pyplot as plt
x = np.sort(np.random.rand(6))
y = np.random.rand(5)
plt.bar(x[:-1], y, width=x[1:] - x[:-1])
plt.show()

```



Graph 15 – Variable width histogram

364. Area Chart

Below is how an area chart looks like:

Let's create a basic area chart with some dummy data

```

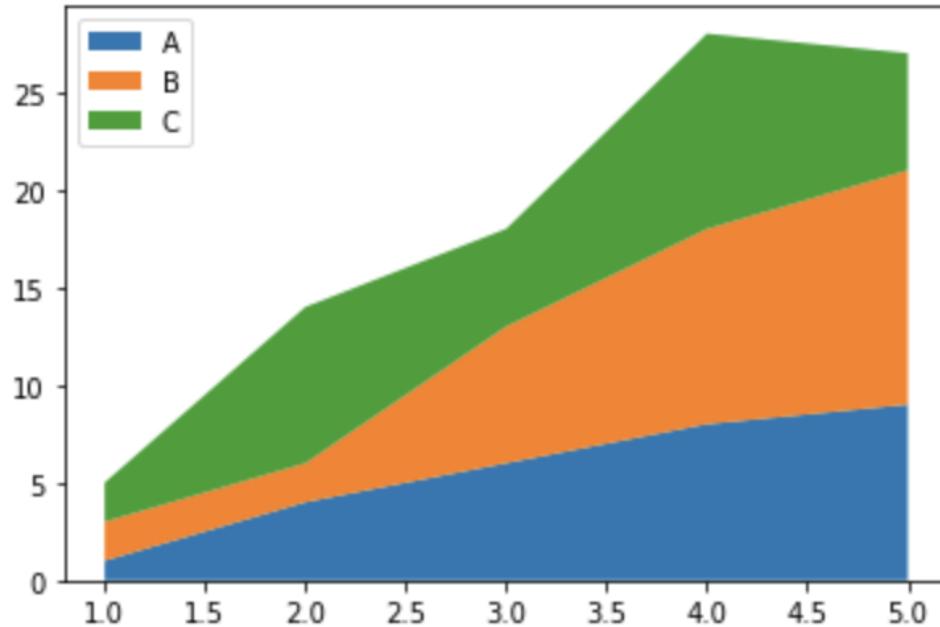
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Data
x=range(1,6)
y=[[1,4,6,8,9], [2,2,7,10,12], [2,8,5,10,6]]
# Plot
plt.stackplot(x,y, labels=['A','B','C'])
plt.legend(loc='upper left')
plt.show()

```

```

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Data
x=range(1,6)
y=[ [1,4,6,8,9], [2,2,7,10,12], [2,8,5,10,6] ]
# Plot
plt.stackplot(x,y, labels=['A','B','C'])
plt.legend(loc='upper left')
plt.show()

```



365 A basic area chart

You already know how to add x-labels, y-labels, title, etc. Go ahead and add these in the graph above

366. Box and Whisker Plot

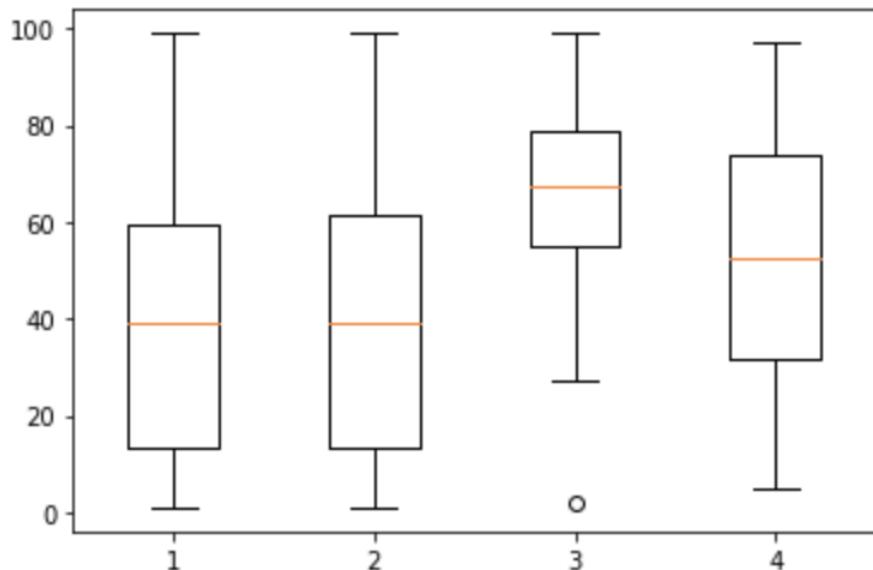
A box and whisker plot, or boxplot for short, is generally used to summarize the distribution of a data sample.

The x-axis is used to represent the data sample, where multiple boxplots can be drawn side by side on the x-axis if desired.

Box plot is one of the most common type of graphics. It gives a nice type of summary of one or more numeric variables. The line that divides the box in the two half is the median of the numbers.

The end of the boxes represents

```
seed(123)
a = random.sample(range(1,100),20)
b = random.sample(range(1,100),20)
c = random.sample(range(1,100),20)
d = random.sample(range(1,100),20)
list_Ex = [a,b,c,d]
plt.boxplot(list_Ex)
```



Graph 17 – A basic Box-Whisker graph

367. Now we will try to make the graph look better by adding color to the plot. The box-plot shows median, 25th and 75th percentile, and outliers. You should try to give different color to these points to make the plot more appealing.

When you plot a boxplot, you can use the following 5 attributes of the plot:-
a. box – To modify the color, line width, etc. of the central box

b. whisker – To modify the color and line width of the line which connects the box to the cap i.e. the horizontal end of the box plot

c. cap – The horizontal end of the box

d. median – The center of the box

e. flier

The box denotes the 1st and 3rd Quartile and it is called IQR i.e. the Inter Quartile Range. The lower fence is at $Q1 - 1.5 \times IQR$ and the upper fence is at $Q3 + 1.5 \times IQR$. Any point which falls above or below it is called fliers or outliers

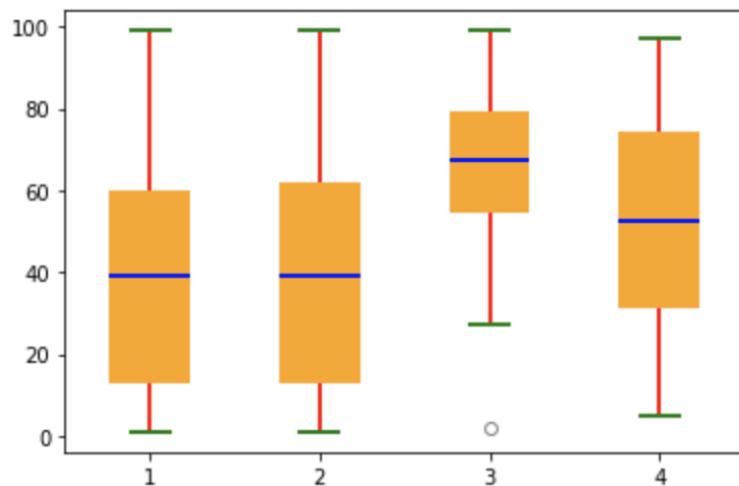
368. Following is the code with some fancy colors to help you understand each term individually.

```
bp=plt.boxplot(list_Ex,patch_artist = True)
for box in bp['boxes']:
    box.set(color='orange',linewidth=2)
for whisker in bp['whiskers']:
    whisker.set(color = 'red',linewidth=2)
for cap in bp['caps']:
    cap.set(color='green',linewidth=2)
for median in bp['medians']:
    median.set(color='blue',linewidth=2)
for flier in bp['fliers']:
    flier.set(marker='o',color = 'black', alpha=0.5)
```

```

bp=plt.boxplot(list_Ex,patch_artist = True)
for box in bp['boxes']:
    box.set(color='orange',linewidth=2)
for whisker in bp['whiskers']:
    whisker.set(color = 'red',linewidth=2)
for cap in bp['caps']:
    cap.set(color='green',linewidth=2)
for median in bp['medians']:
    median.set(color='blue',linewidth=2)
for flier in bp['fliers']:
    flier.set(marker='o',color = 'black', alpha=0.5)

```



The graph produced is below:-

369. Box Whisker Chart

Box-plot practice

Following is one more code with the help of which you can replicate a Gaussian distribution

```

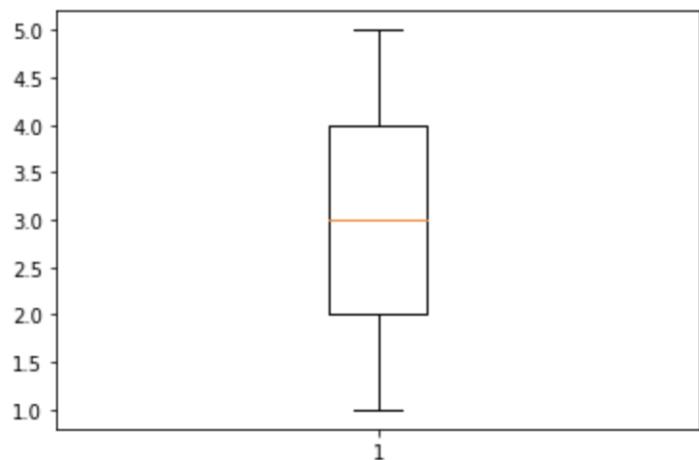
from numpy.random import seed
from numpy.random import randn
from matplotlib import pyplot
seed(1)
# random numbers drawn from a Gaussian distribution x = [randn(1000), 5 * randn(1000), 10 *
randn(1000)] # create box and whisker plot
pyplot.boxplot(x)
# show line plot
pyplot.show()

```

```

from numpy.random import seed
from numpy.random import randn
from matplotlib import pyplot
seed(1)
pyplot.boxplot(x)
# show line plot
pyplot.show()

```



Graph 19 – A Box-Whisker Plot

370. Scatter plot

Scatter plot is an easy to make but interesting visualization which gives a clear picture of how the data is distributed.

Let's take example of 10 innings played by Sachin, Dhoni, and Kohli and see how their scores are distributed. The code is fairly easy to understand

```

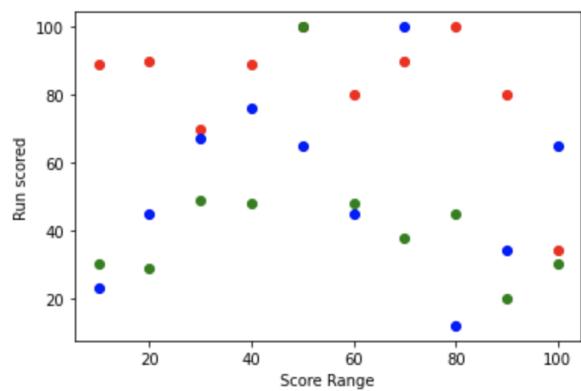
sachin = [89, 90, 70, 89, 100, 80, 90, 100, 80, 34]
kohli = [30, 29, 49, 48, 100, 48, 38, 45, 20, 30]
dhoni = [23,45,67,76,65,45,100,12,34,65]
run = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
plt.scatter(run, sachin, color='red')
plt.scatter(run, kohli, color='green')
plt.scatter(run,dhoni,color='blue')
plt.xlabel('Score Range')
plt.ylabel('Run scored')
plt.show()

```

```

sachin = [89, 90, 70, 89, 100, 80, 90, 100, 80, 34]
kohli = [30, 29, 49, 48, 100, 48, 38, 45, 20, 30]
dhoni = [23,45,67,76,65,45,100,12,34,65]
run = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
plt.scatter(run, sachin, color='red')
plt.scatter(run, kohli, color='green')
plt.scatter(run,dhoni,color='blue')
plt.xlabel('Score Range')
plt.ylabel('Run scored')
plt.show()

```



You can also add legend in the plot by using the following command

```
legend = ['sachin','kohli','dhoni'] plt.legend(legend)
```

The plot will now look like this

371. A scatter plot

Below is one more scatter plot where you give weighted area and the size of the circle will be on the basis of the circle

```
import numpy as np
np.random.seed(123)
x = random.sample(range(1,100),40)
y = random.sample(range(1,100),40)
colors = np.random.rand(N)
area = (30*np.random.rand(N))**2
plt.scatter(x,y,s=area,c=colors,alpha=0.5)
plt.show()
```

This code generates a scatter plot of 40 random points with random colors and sizes. The `np.random.seed(123)` line sets the random seed to ensure that the random values are reproducible. The `s` parameter controls the size of the points, and the `c` parameter controls the color of the points. The `alpha` parameter controls the transparency of the points. The scatter plot is displayed using the `plt.show()` function.

372. Pie Chart

Create a pie chart for the number of centuries scored by Sachin, Dhoni, Dravid, and Kohli.

```
labels = 'Sachin','Dhoni','Kohli','Dravid'
size = [100,25,70,50]
colors = ['pink','blue','red','orange']
explode = (0.1,0,0,0) plt.pie(size,explode=explode,labels=labels,colors=colors,autopct='%.1f%%',shadow=True,startangle=140)
plt.axis('equal') plt.show()
```

`explode` is used to set apart the first part of the pie chart. Everything else in the code is self explanatory. Below is the plot

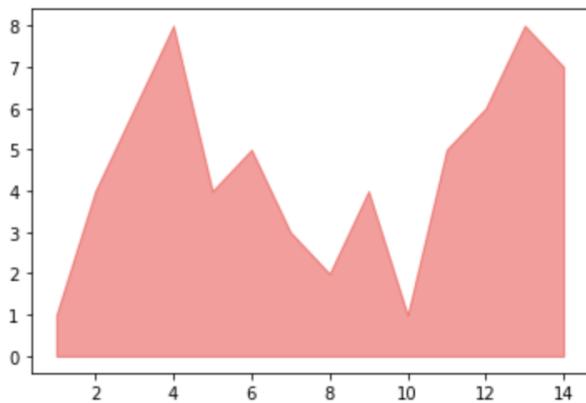
373. Stacked Bar graph

A cool area graph

```
import numpy as np
import matplotlib.pyplot as plt # create data
x=range(1,15)
y=[1,4,6,8,4,5,3,2,4,1,5,6,8,7]
```

```
# Change the color and its transparency
plt.fill_between( x, y, color="red", alpha=0.4)
plt.show()
# Same, but add a stronger line on top (edge) plt.fill_between( x, y, color="red", alpha=0.2)
plt.plot(x, y, color="red", alpha=0.6)
```

```
import numpy as np
import matplotlib.pyplot as plt # create data
x=range(1,15)
y=[1,4,6,8,4,5,3,2,4,1,5,6,8,7]
# Change the color and its transparency
plt.fill_between( x, y, color="red", alpha=0.4)
plt.show()
```



The parameter alpha is used to give weight age to the density of color. 0.4 is given to the edge and 0.2 is given to the fill

374. What are the type of information that we can display using plot?

There are four types of information which we can display using any plot:-

1. Distribution
2. Comparison
3. Relationship
4. Composition

375. What are the charts that show the distribution in your data set?

Distribution shows how diversely the data is distributed in your data set.

- a Histogram – If you have few data point
- b. Line Histogram – When you have a lot of data points
- c. Scatter plot – When you have to show the distribution of 2-3 variables

376. What are the charts that show a comparison metrics?

Comparison – When you have to compare something over 2 or more categories

- a. Variable width chart – When you have to compare two variables per item
- b. Tables with embedded charts – When there are many categories, basically a matrix of charts
- c. Horizontal or Vertical Histogram – When there are few categories in a data set
- d. If you want to compare something over time
 - i. Line Chart
 - ii. Bar Vertical Chart
 - iii. Many categories line chart

377. What are the types of relationship charts?

Relationship Charts – When you want to see the relationship between two or more variables then you have to use relationship charts

- a. Scatter Plot
- b. Scatter plot bubble chart

378. What are comparison charts?

Composition Charts – When you have to show a percentage or composition of variables.

- a. Pie Chart – Elementary plot when there are 3-6 categories
- b. Stacked 100% bar chart with sub-component – When you have to show components of components
- c. Stacked 100% bar chart – When you have to look into the contribution of each component.
- d. Stacked area chart – When relative and absolute difference matters

Tricky Interview Questions

Python for Analytics

379. How do you handle missing values in a data set? What are some common techniques for imputing missing data?

Missing data can be handled by various techniques in Python. Some of the common techniques for imputing missing data are:

-Dropping missing data: If the number of missing values is very small in the dataset, then dropping those rows or columns might not affect the analysis.

Mean/median/mode imputation: Replace missing data with the mean, median or mode value of the feature/column.

Forward/Backward filling: Use the previous or the next value to fill the missing data.

Interpolation: Using statistical or mathematical techniques, interpolate the missing values.

K-nearest neighbors imputation: Use the values of k-nearest neighbors to impute the missing values.

380.How do you handle categorical data in a data set, and what are some common techniques for encoding or transforming it?

Categorical data can be handled by the following techniques in Python:

Label encoding: Assigning a unique numerical value to each category.

One-Hot encoding: Create dummy variables for each category.

Binary encoding: Convert each category into binary code.

Target encoding: Replace the categories with the mean of the target variable.

Frequency encoding: Replace the categories with the frequency of occurrence.

381.What are some common machine learning algorithms that you have experience with, and how do you use them in Python?

Some common machine learning algorithms are:

Linear Regression: Used for regression problems.

Logistic Regression: Used for classification problems.

Decision Trees: Used for classification and regression problems.

Random Forest: Ensemble method of decision trees for classification and regression problems.

K-Nearest Neighbors: Used for classification and regression problems.

Naive Bayes: Used for classification problems.

Support Vector Machines: Used for classification and regression problems.

382. How do you assess the performance of a machine learning model, and what are some common metrics used to do so?

The performance of a machine learning model can be assessed by various evaluation metrics in Python, some of which are:

Mean Absolute Error (MAE): measures the average absolute difference between the predicted and actual values.

Mean Squared Error (MSE): measures the average squared difference between the predicted and actual values.

Root Mean Squared Error (RMSE): square root of the average squared difference between the predicted and actual values.

R-squared: measures the proportion of variance in the dependent variable that is predictable from the independent variables.

Classification accuracy: measures the percentage of correct predictions.

Confusion matrix: gives a table of the actual vs. predicted values.

383.What is cross-validation, and how is it used in machine learning?

Cross-validation is a technique used to evaluate the performance of a machine learning model by dividing the dataset into multiple parts, training the model on some parts, and testing the model on the remaining parts. The purpose of cross-validation is to ensure that the model is not overfitting the data. The common types of cross-validation are k-fold cross-validation and leave-one-out cross-validation.

384.What is regularization, and how is it used to prevent overfitting in machine learning?

Regularization is a technique used to prevent overfitting in machine learning models by adding a penalty term to the objective function. The penalty term is a function of the magnitude of the coefficients, and it encourages the model to choose simpler coefficients. The common types of regularization are L1 regularization (Lasso) and L2 regularization (Ridge).

385.How do you handle imbalanced data sets in machine learning, and what are some common techniques for addressing this issue?

category in the data set is represented by far fewer observations than the other(s). This can lead to biased models that do not perform well on new data. Some common techniques for addressing this issue include:

Undersampling the majority class to balance the distribution of observations across classes

Oversampling the minority class to create synthetic data points for that class

Using a combination of undersampling and oversampling techniques

Using algorithms specifically designed to handle imbalanced data, such as decision trees with cost-sensitive learning or support vector machines with different misclassification penalties

386.How do you use Python for data visualization, and what are some common libraries or tools that you have used?

Matplotlib, a popular plotting library with a wide range of customization options

Seaborn, a library for statistical data visualization with built-in support for many common plot types and themes

Plotly, a library for interactive and dynamic visualizations

Bokeh, a library for interactive visualizations that can be easily integrated with web applications

Pandas, a data manipulation library that includes some basic visualization functions.

387.How do you handle time series data in Python, and what are some common techniques for analyzing or forecasting time series data?

Time series data is a common type of data in many fields, and Python offers a range of tools

and techniques for analyzing and forecasting time series data. Some common techniques include:

Visualization of the time series to identify trends, patterns, and seasonality

Decomposition of the time series into its trend, seasonal, and residual components to better understand its underlying structure

Smoothing techniques such as moving averages or exponential smoothing to remove noise and highlight trends or seasonality

Autoregressive Integrated Moving Average (ARIMA) models for forecasting

Exponential Smoothing (ES) models for forecasting

Seasonal ARIMA (SARIMA) models for forecasting seasonality in the data

Prophet, a library for time series forecasting developed by Facebook.

Visualization Interview Questions in Python

388. What are some common Python libraries used for data visualization, and how do they differ in terms of functionality and ease of use?

Some common Python libraries used for data visualization are Matplotlib, Seaborn, Plotly, Bokeh, and ggplot. They differ in terms of functionality and ease of use. Matplotlib is the most basic and widely used library for creating static visualizations, while Seaborn provides more advanced statistical visualization techniques. Plotly and Bokeh are powerful libraries for creating interactive visualizations, and ggplot is a library that follows the grammar of graphics approach for creating sophisticated plots.

389. How do you create a line chart or scatter plot in Python using a given data set?

To create a line chart or scatter plot in Python, you can use the Matplotlib library. First, import Matplotlib and then use the plot function to create a line chart or scatter plot. You can customize the appearance of the plot by adding labels and changing colors, and then use the show function to display the plot.

390. How do you customize the appearance of a plot in Python, such as changing the color, title, axis labels, or legend?

To customize the appearance of a plot in Python, you can use various functions provided by the Matplotlib library. For example, you can use the title function to add a title to the plot, the xlabel and ylabel functions to add labels to the axes, and the legend function to add a legend to the plot. You can also change the color of the lines or markers by specifying a color parameter.

391. What is a heatmap in data visualization, and how do you create one in Python?

A heatmap is a data visualization technique that displays data as a color-coded matrix, with different colors representing different values. In Python, you can create a heatmap using the Matplotlib library and the imshow function. You can customize the color scheme and the axis

labels, and add a colorbar to the plot.

392. What is a histogram in data visualization, and how do you create one in Python?

A histogram is a data visualization technique that displays the distribution of a variable in the form of bars, where the height of each bar represents the frequency or proportion of values falling in a particular interval. In Python, you can create a histogram using the Matplotlib library and the hist function. You can customize the number of bins, the range of values, and the color of the bars.

393. How do you handle missing data or outliers in a data set when creating a visualization, and how do they affect the final result?

When handling missing data or outliers in a data set, you can choose to either remove the data points or impute the missing values using techniques like mean, median, or mode imputation. Missing data or outliers can affect the final result of a visualization by skewing the distribution or introducing bias. It's important to carefully evaluate the impact of missing data or outliers on the visualization and take appropriate steps to handle them.

394. What are some common techniques for visualizing high-dimensional data sets in Python, such as using dimensionality reduction or clustering algorithms?

When visualizing high-dimensional data sets, you can use techniques like dimensionality reduction or clustering algorithms. For example, principal component analysis (PCA) can be used to reduce the dimensionality of the data, while clustering algorithms like k-means can be used to group similar data points together. Visualization techniques like parallel coordinates plots or scatter matrix plots can also be used to visualize high-dimensional data sets.

395. How do you create interactive visualizations in Python, such as using tools like Bokeh or Plotly?

To create interactive visualizations in Python, you can use tools like Bokeh or Plotly. These libraries allow you to create dynamic visualizations that respond to user interactions, such as hovering over data points or clicking on elements. You can create various types of interactive visualizations like scatter plots, line charts, and bar charts.

396. What are some common techniques for visualizing geographic data in Python, such as using choropleth maps or scatter plots with latitudes and longitudes?

When visualizing geographic data in Python, you can use techniques like choropleth maps or scatter plots with latitudes and longitudes. Choropleth maps display data on a geographic map, with different colors representing different values, while scatter plots display data points on a map using latitudes and longitudes. Libraries like Geopandas and Folium provide convenient

functions for creating these types of visualizations.

397. How do you create animated visualizations in Python, such as using the Matplotlib animation module or GIFs?

To create animated visualizations in Python, you can use the Matplotlib animation module or create GIFs using external libraries. Here is an overview of the process:

Import the necessary libraries: To use the Matplotlib animation module, you will need to import Matplotlib and the animation module. For creating GIFs, you can use external libraries like imageio.

Create a figure and set the initial plot: Create a figure using Matplotlib and set the initial plot. This can be a line chart, scatter plot, or any other type of visualization.

Define the animation function: Define an animation function that updates the data in the plot for each frame. This function will be called repeatedly to generate the animation.

Create the animation object: Create an animation object using the FuncAnimation method of the animation module. This object takes the figure, the animation function, the number of frames, and the interval between frames as arguments.

Display or save the animation: You can display the animation in a Matplotlib window or save it as a video file using the save method of the animation object. For creating a GIF, you can use the write_gif method of the imageio library.

Here is an example of using the Matplotlib animation module to create a simple line chart animation:

```
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.animation import FuncAnimation

# Create the figure and set the initial plot
fig, ax = plt.subplots()
line, = ax.plot([], [])

# Define the animation function
def animate(i):
    x = np.linspace(0, 10, 1000)
    y = np.sin(x - i/10)
    line.set_data(x, y)
    return line,
```

```
# Create the animation object
ani = FuncAnimation(fig, animate, frames=100, interval=50)

# Display the animation
plt.show()
```

In this example, the animation function updates the data in the plot by shifting the phase of a sine wave for each frame. The FuncAnimation object takes the figure, the animation function, 100 frames, and an interval of 50 milliseconds between frames. Finally, we display the animation using the show method of Matplotlib.

Python Tricky Interview Questions

398. What is the difference between a list and a tuple in Python, and when would you use each?

In Python, a list is a mutable collection of ordered elements, while a tuple is an immutable collection of ordered elements. You would use a list when you need to modify the collection (add, remove, or change elements), while a tuple is useful when you need to represent a fixed sequence of values that cannot be modified. For example, you might use a list to represent a shopping list that you can add items to or remove items from, while you might use a tuple to represent a date with fixed values for the year, month, and day.

399. Can you explain the difference between is and == in Python?

In Python, the `is` operator checks whether two objects are the same object in memory, while the `==` operator checks whether two objects have the same value. So `a is b` is True if `a` and `b` are the same object in memory, while `a == b` is True if `a` and `b` have the same value. For example, if you have two lists `a` and `b` with the same values, `a == b` will be True, but `a is b` will be False, because they are different objects in memory.

400. What is the difference between a shallow copy and a deep copy in Python, and when would you use each?

In Python, a shallow copy creates a new object that references the same memory as the original object, while a deep copy creates a new object with its own memory that is a complete copy of the original object. You would use a shallow copy when you only need a new object that shares some data with the original object, while you would use a deep copy when you need a new object that is completely independent of the original object. For example, if you have a list of lists, a shallow copy will create a new list of references to the same inner lists, while a deep copy will create a new list of completely independent inner lists.

401. Can you explain what a closure is in Python and provide an example?

In Python, a closure is a function that remembers the values of the variables in its enclosing scope, even when the function is called outside that scope. A closure can be created by defining a nested function that references a variable from the enclosing scope, and returning that nested function. For example:

```
def make_adder(n):
    def adder(x):
        return x + n
    return adder

add5 = make_adder(5)
add10 = make_adder(10)

print(add5(3)) # prints 8
print(add10(3)) # prints 13
```

In this example, `make_adder` returns a nested function `adder` that remembers the value of the variable `n` from its enclosing scope.

402. How does Python's global interpreter lock (GIL) impact multi-threaded programming in Python?

Python's Global Interpreter Lock (GIL) is a mechanism that ensures that only one thread executes Python bytecode at a time. This means that Python cannot take advantage of multiple cores or processors to execute multiple threads in parallel. However, I/O-bound tasks, such as network operations or disk access, can be run in separate threads without being impacted by the GIL.

403. What is the difference between a generator and a list comprehension in Python, and when would you use each?

In Python, a list comprehension is a way to create a new list by applying an expression to each element of an existing list, while a generator is an iterable that generates values on-the-fly as they are requested. You would use a list comprehension when you need to create a new list that you will use multiple times, while you would use a generator when you need to generate a sequence of values one at a time and do not need to store the entire sequence in memory. For example, you might use a list comprehension to create a list of squared numbers: `squares = [x**2 for x in range(10)]`. You might use a generator to iterate over the lines of a large file: `lines = (line for line in open('large_file.txt'))`.

404. How does Python handle memory management, and what are some common memory management issues that can arise?

Python uses automatic memory management with a garbage collector that periodically frees memory that is no longer being used. The garbage collector tracks references to objects, and when an object no longer has any references, it is considered garbage and its memory is freed. Some common memory management issues in Python include creating circular references, which prevent objects from being garbage-collected, and holding onto references to large objects for longer than necessary, which can cause memory usage to grow unnecessarily.

405. Can you explain the difference between a module and a package in Python, and how would you import and use each?

In Python, a module is a file containing Python code, while a package is a directory containing one or more Python modules. Modules can be imported into other modules or scripts using the `import` statement, while packages can be imported using either the `import` statement or the `from ... import` statement. To use a module or package, you would typically import it at the top of your Python script or module, and then use its functions, classes, or variables as needed.

406. What is the difference between a class method and a static method in Python, and when would you use each?

In Python, a class method is a method that is bound to the class rather than an instance of the class, and receives the class itself as its first argument (usually called `cls`). A static method is a method that is also bound to the class rather than an instance, but does not receive the class as an argument. You would use a class method when you need to modify or access class-level variables or properties, while you would use a static method when you need to perform a task related to the class but not specific to any instances of the class.

407. Can you explain the difference between the range and xrange functions in Python 2, and why was xrange removed in Python 3?

In Python 2, `range` and `xrange` are two functions for generating a sequence of integers. `range` returns a list containing all the integers in the specified range, while `xrange` returns an iterable that generates the integers on-the-fly as they are requested. In Python 3, the `xrange` function was removed, and the `range` function was updated to behave like the `xrange` function in Python 2. This means that `range` in Python 3 now returns an iterable that generates the integers on-the-fly, rather than a list containing all the integers. This change was made to reduce memory usage and improve performance for large ranges.

408. What is the difference between "global" and "nonlocal" in Python?

"`global`" is a keyword in Python that allows you to access and modify global variables from within a function. "`nonlocal`" is a keyword that allows you to access and modify variables from the nearest enclosing scope (excluding the global scope). In other words, "`global`" is used to access and modify variables in the global scope, while "`nonlocal`" is used to access and modify

variables in a specific enclosing scope.

409. What is the difference between "deepcopy" and "shallow copy" in Python?

"deepcopy" is a method in the "copy" module that creates a new object with a complete copy of the original object, including all nested objects. "shallow copy" is a method that creates a new object with a copy of the original object's references, but not the objects themselves. In other words, a shallow copy creates a new object that refers to the same objects as the original object, while a deepcopy creates a new object with a completely independent copy of all objects.

410. What is a decorator in Python?

A decorator is a function that takes another function as input and returns a new function that usually extends or modifies the original function's behavior. Decorators are a way to modify or extend the behavior of functions without having to modify the original function's code directly.

411. What is a lambda function in Python?

A lambda function is an anonymous function in Python that can be defined in a single line of code. Lambda functions are often used as a shorthand for simple functions that are only used once and do not require a separate function definition.

412. What is the difference between a module and a package in Python?

A module is a single file that contains Python code, while a package is a collection of modules that are organized into a directory hierarchy. A package can contain sub-packages, which in turn can contain modules, allowing for a more organized and modular code structure.

413. What will be the output of the code below in Python 2?

```
def div1(x,y):
    print "%s/%s = %s" % (x, y, x/y)
def div2(x,y):
    print "%s//%s = %s" % (x, y, x//y)
div1(5,2)
div1(5.,2)
div2(5,2)
div2(5.,2.)
```

Output:

```
$pythonmain.py
5/2 = 2
5.0/2 = 2.5
```

```
5//2 = 2  
5.0//2.0 = 2.0
```

Statistics using Python

414. Write a Python program to calculate mode without using inbuilt function Input

```
5 1 1 1 2 3
```

Output

Mode-1

Solution/Approach

```
#Mode  
X = int(input()) L=[]  
for i in range(X):  
    Y = int(input())  
    L.append(Y)  
q = max(set(L),key=L.count) print("Mode "+str(q))
```

```
In [6]: x = int(input())  
L=[]  
for i in range(x):  
    y = int(input())  
    L.append(y)  
q = max(set(L),key=L.count)  
print("Mode "+str(q))
```



```
5  
1  
1  
1  
2  
2  
Mode 1
```

415. How to webscrapping in Python?

Python is a popular programming language for web scraping, as it has powerful libraries such as BeautifulSoup and Scrapy that make it easy to extract data from websites. Here's a brief overview of how to perform web scraping in Python:

Install the necessary libraries: Before you start web scraping, you'll need to install the libraries you'll be using. For example, you can use pip to install BeautifulSoup or Scrapy.

Inspect the webpage: To extract data from a webpage, you'll need to know the structure of the HTML code. You can do this by inspecting the page using your browser's developer tools. Look for the HTML elements that contain the data you're interested in.

Use BeautifulSoup or Scrapy to extract the data: Once you've identified the HTML elements you want to extract data from, you can use BeautifulSoup or Scrapy to extract the data. For example, with BeautifulSoup, you can use its find() and find_all() methods to search for specific HTML elements, and then extract the text or attributes from those elements.

Parse the data: Once you've extracted the data, you may need to clean it up or convert it to a different format. For example, you can use regular expressions or string manipulation to remove unwanted characters or formatting.

Store the data: Finally, you'll want to store the data in a useful format. You can write the data to a file or a database, or use it in another application.

Here's some example code that uses BeautifulSoup to extract the title of a webpage:

```
import requests
from bs4 import BeautifulSoup

url = 'https://www.example.com'
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

title = soup.find('title').text
print(title)
```

This code sends a GET request to https://www.example.com, gets the HTML response, and uses BeautifulSoup to find the title element and extract its text. The code then prints the title to the console.

416. What is the difference between "is" and "=="? In Python?

The "is" operator checks if two objects are the same object in memory. In other words, it checks if the two operands refer to the same object in memory. If they do, then "is" returns True; if not, it returns False.

The "==" operator, on the other hand, checks if the values of the two operands are equal. It compares the contents of the objects, rather than the memory addresses. If the values are equal, then "==" returns True; if not, it returns False.

Here's an example to illustrate the difference:

```
a = [1, 2, 3]
b = a
c = [1, 2, 3]

print(a is b) # True, since a and b refer to the same object in memory
print(a is c) # False, since a and c are different objects in memory
print(a == c) # True, since the values of a and c are the same
```

In this example, "a" and "b" both refer to the same list object in memory, so "a is b" returns True. However, "a" and "c" are different list objects with the same values, so "a is c" returns False. Finally, "a == c" returns True, since the values of "a" and "c" are the same, even though they are different objects in memory.

417. How is a variable allocated memory in Python?

In Python, variables are dynamically typed, which means that the type of a variable is determined at runtime, rather than at compile-time. When a variable is assigned a value, Python automatically allocates memory for the variable and assigns a memory address to the variable. The memory allocation and deallocation in Python are managed by the Python interpreter, so you don't have to worry about managing memory manually as you would in languages like C++.

Python uses a memory management technique called reference counting to keep track of the number of references to an object in memory. When an object is created, it is given a reference count of one. When a variable is assigned to that object, the reference count is incremented by one. When a variable is reassigned or deleted, the reference count is decremented. When the reference count reaches zero, the object is no longer being used and its memory is freed.

Here's an example to illustrate how memory allocation works in Python:

```
a = 42
b = "hello"
c = [1, 2, 3]

print(id(a)) # prints the memory address of a
print(id(b)) # prints the memory address of b
print(id(c)) # prints the memory address of c
```

In this example, Python automatically allocates memory for variables "a", "b", and "c" and assigns a memory address to each variable. The id() function returns the memory address of the object that a variable is referring to.

It's important to note that Python handles memory allocation and deallocation automatically, but you can still run into memory-related issues if you're not careful. For example, if you create a

large number of objects that are never released, you can run out of memory and crash your program. To avoid these issues, you can use tools like memory profiling and garbage collection to monitor and manage memory usage in your Python programs.

418. Define a class named car with 2 attributes, “color” and “speed”. Then create an instance and return speed.

```
class Car :  
    def __init__(self, color, speed):  
        self.color = color  
        self.speed = speed  
car = Car('red','100mph')  
car.speed  
#=> '100mph'
```

419. What is the difference between instance, static and class methods in python?

In Python, there are three types of methods that can be defined in a class: instance methods, static methods, and class methods. Here's a brief overview of each:

Instance methods: Instance methods are the most common type of method in Python classes. They are defined with the "self" parameter and operate on the instance of the class. They can access and modify instance variables, and are called using the dot notation on an instance of the class.

Static methods: Static methods are methods that do not operate on the instance of the class, but instead operate on the class itself. They are defined using the "@staticmethod" decorator and do not take any implicit parameters (like "self" in instance methods). Static methods can be called using the dot notation on the class itself, rather than an instance of the class.

Class methods: Class methods are methods that operate on the class itself, rather than on the instance of the class. They are defined using the "@classmethod" decorator and take the class itself as an implicit parameter (usually named "cls"). Class methods can be used to create alternate constructors for the class, or to modify class variables.

Here's an example to illustrate the difference between instance, static, and class methods:

```
class MyClass:  
    class_variable = 0  
  
    def __init__(self, instance_variable):  
        self.instance_variable = instance_variable  
  
    def instance_method(self):
```

```

print("This is an instance method.")
print("Instance variable:", self.instance_variable)
print("Class variable:", MyClass.class_variable)

@staticmethod
def static_method():
    print("This is a static method.")
    print("Class variable:", MyClass.class_variable)

@classmethod
def class_method(cls):
    print("This is a class method.")
    print("Class variable:", cls.class_variable)

# Creating an instance of the class
obj = MyClass("hello")

# Calling instance method
obj.instance_method()

# Calling static method
MyClass.static_method()

# Calling class method
MyClass.class_method()

```

In this example, "instance_method" is an instance method that operates on the instance of the class, "static_method" is a static method that operates on the class itself, and "class_method" is a class method that also operates on the class itself. Each method is called using a different syntax, depending on whether it

420. What is the difference between “func” and “func()”?

In Python, "func" refers to the function object itself, while "func()" refers to the result of calling the function.

When you define a function in Python, you are creating a new object that can be called later to perform a specific task. The name you give to the function is just a reference to that object. For example, consider the following function:

```

def say_hello():
    print("Hello, world!")

```

In this example, "say_hello" is a function object that can be called by using parentheses after

the function name:

```
say_hello() # calls the function and prints "Hello, world!"
```

So, "say_hello" refers to the function object itself, while "say_hello()" calls the function and returns its result (in this case, the function simply prints "Hello, world!" to the console).

It's important to note that not all functions in Python return a value. Some functions may simply perform some action, like printing output to the console or modifying a global variable, without returning a specific value. In these cases, calling the function with parentheses simply executes the function and has no return value.

Here's an example to illustrate this:

```
def increment_global_variable():
    global x
    x += 1

x = 0
increment_global_variable() # calls the function, increments x by 1
print(x) # prints 1
```

In this example, "increment_global_variable" is a function that modifies a global variable ("x") without returning a value. When the function is called with parentheses, it simply modifies the global variable and returns nothing.

421. Does Python call by value or call by reference?

In Python, everything is passed by assignment, which means that neither call by value nor call by reference is used.

When a variable is passed as an argument to a function in Python, a new reference to the object it points to is created and passed to the function. The function then receives a reference to the object, but not the original object itself. This means that modifications made to the object inside the function will be visible outside the function, since both the original reference and the function's reference point to the same object.

However, if the function reassigns the reference to a new object, this change will not be visible outside the function. For example:

```
def modify_list(lst):
    lst.append(4) # modifies the list in place
    lst = [1, 2, 3] # reassigns lst to a new list
```

```
my_list = [0]
modify_list(my_list)
print(my_list) # prints [0, 4], not [1, 2, 3, 4]
```

In this example, "my_list" is a list that is passed to the "modify_list" function. The function modifies the list in place by appending the value 4 to it, which is visible outside the function. However, when the function reassigns "lst" to a new list, this change is not visible outside the function, since the original reference to the list ("my_list") is still pointing to the original object.

So, to summarize, Python passes everything by assignment, which means that the function receives a reference to the original object, but not the object itself. If the function modifies the object in place, these changes will be visible outside the function, but if the function reassigns the reference to a new object, this change will not be visible outside the function.

422. Given two strings, string1 and string2, determine if there exists a one to one character mapping between each character of string1 to string2.

Example:

```
string1 = 'qwe'
string2 = 'asd'
string_map(string1, string2) == True
#q = a, w = s, and e = d
```

Note: This example would return False if the letters were repeated; for example, string1 = 'donut' and string2 = 'fatty'. This is because the letter t from fatty attempts to map to two different outcomes (t = n or t = u).

Python Pattern printing interview questions

423. Write a program to print the following pattern:

1
12
123
1234
12345

```
for i in range(1, 6):
    for j in range(1, i+1):
        print(j, end="")
    print()
```

424. Write a program to print the following pattern:

A
AB
ABC
ABCD
ABCDE

```
for i in range(1, 6):
    for j in range(65, 65+i):
        print(chr(j), end="")
    print()
```

425. Write a program to print the following pattern:

1
22
333
4444
55555

```
for i in range(1, 6):
    for j in range(1, i+1):
        print(i, end="")
    print()
```

426. Write a program to print the following pattern:

E
DE
CDE
BCDE
ABCDE

```
for i in range(5):
    for j in range(i, 5):
        print(chr(69-j), end="")
    print()
```

427. Write a program to print the following pattern:

```
1
2 3
4 5 6
7 8 9 10
```

```
num = 1
for i in range(1, 5):
    for j in range(i):
        print(num, end=" ")
    num += 1
print()
```

428. Write a program to print the following pattern:

```
A B C D E F
A B C D E
A B C D
A B C
A B
A
```

```
for i in range(6, 0, -1):
    for j in range(65, 65+i):
        print(chr(j), end=" ")
    print()
```

429. Write a program to print the following pattern:

```
1
0 1
1 0 1
0 1 0 1
1 0 1 0 1
```

```
for i in range(1, 6):
    for j in range(i):
        if (i+j) % 2 == 0:
            print(1, end=" ")
        else:
            print(0, end=" ")
```

```
print()
```

Python Special Number interview questions

430. Write a program to check if a number is prime or not.

```
def is_prime(num):
    if num < 2:
        return False
    for i in range(2, int(num**0.5) + 1):
        if num % i == 0:
            return False
    return True
```

The function takes an integer as input and returns True if the number is prime, and False otherwise. We iterate from 2 to the square root of the number and check if there is any number that divides the given number. If we find any such number, then it's not a prime number, and we return False. If we don't find any such number, then it's a prime number, and we return True.

431. Write a program to find all prime numbers in a given range.

```
def find_primes(start, end):
    primes = []
    for num in range(start, end + 1):
        if is_prime(num):
            primes.append(num)
    return primes
```

The function takes a range of integers as input and returns a list of all the prime numbers in that range. It calls the is_prime function to check if a number is prime or not.

432. Write a program to check if a number is an Armstrong number or not.

```
def is_armstrong(num):
    n = len(str(num))
    sum = 0
    for digit in str(num):
        sum += int(digit)**n
    return sum == num
```

The function takes an integer as input and returns True if the number is an Armstrong number, and False otherwise. An Armstrong number is a number that is equal to the sum of its digits

raised to the power of the number of digits in the number. For example, 153 is an Armstrong number because $1^3 + 5^3 + 3^3 = 153$.

433. Write a program to check if a number is a perfect number or not.

```
def is_perfect(num):
    divisors = []
    for i in range(1, num):
        if num % i == 0:
            divisors.append(i)
    return sum(divisors) == num
```

The function takes an integer as input and returns True if the number is a perfect number, and False otherwise. A perfect number is a positive integer that is equal to the sum of its proper divisors (excluding the number itself). For example, 6 is a perfect number because its proper divisors are 1, 2, and 3, and $1 + 2 + 3 = 6$.

434. Write a program to find the factorial of a number.

```
def factorial(num):
    if num == 0:
        return 1
    else:
        return num * factorial(num - 1)
```

The function takes an integer as input and returns the factorial of the number. The factorial of a number is the product of all the positive integers from 1 to that number. For example, the factorial of 5 is $5 * 4 * 3 * 2 * 1 = 120$.

435. Write a program to check if a number is a palindrome or not.

```
def is_palindrome(num):
    num_str = str(num)
    reversed_str = num_str[::-1]
    return num_str == reversed_str

# example usage
print(is_palindrome(121)) # True
print(is_palindrome(123)) # False
```

The function takes an integer as input and returns True if the number is a palindrome, and False otherwise. A palindrome number is a number that remains the same when its digits are reversed. For example, 121 is a palindrome number.

436. Write a program to find the sum of digits of a number.

```
def sum_of_digits(num):
    sum = 0
    while num > 0:
        digit = num % 10
        sum += digit
        num = num // 10
    return sum

# example usage
print(sum_of_digits(123))  # 6
print(sum_of_digits(4567)) # 22
```

437. Write a program to check if a number is an automorphic number or not.

```
def is_automorphic(num):
    square = num * num
    while num > 0:
        if num % 10 != square % 10:
            return False
        num = num // 10
        square = square // 10
    return True

# example usage
print(is_automorphic(5))  # True, because 5*5 = 25 and 5 is present at the
end of 25
print(is_automorphic(6))  # False, because 6*6 = 36 but 6 is not present
at the end of 36
```

438. Write a program to check if a number is a Harshad number or not.

```
def is_harshad(num):
    sum_of_digits = 0
    original_num = num
    while num > 0:
        digit = num % 10
        sum_of_digits += digit
        num = num // 10
    return original_num % sum_of_digits == 0

# example usage
print(is_harshad(18))  # True, because 1+8=9 and 18 is divisible by 9
print(is_harshad(23))  # False, because 2+3=5 and 23 is not divisible by 5
```

439. Write a program to check if a number is a Smith number or not.

```
# helper function to find prime factors of a number
def prime_factors(num):
    i = 2
    factors = []
    while i * i <= num:
        if num % i:
            i += 1
        else:
            num //= i
            factors.append(i)
```

```

if num > 1:
    factors.append(num)
return factors

def is_smith(num):
    # find sum of digits of the number
    sum_of_digits = sum(int(digit) for digit in str(num))

    # find prime factors of the number
    prime_factors_list = prime_factors(num)

    # find sum of digits of prime factors
    sum_of_factors = sum(sum(int(digit) for digit in str(factor)) for
factor in prime_factors_list)

    return sum_of_digits == sum_of_factors

# example usage
print(is_smith(666))  # True, because 6+6+6=18 and prime factors are [2,
3, 3, 37] and 2+3+3+3+7=18
print(is_smith(123))  # False, because 1+2+3=6 but prime factors are [3,
41] and 3+4+1=8

```

Output base Python interview questions

440. What is the difference between print() and return in Python?

`print()` is a function used to output information to the console. It does not return a value that can be used in other parts of the code. `return` is a keyword used to indicate the value that should be returned from a function. It is used to exit the function and pass a value back to the caller.

441. Can you give an example of using f-strings in Python?

```

name = "John"
age = 30
print(f"My name is {name} and I am {age} years old.")

```

The output of this code would be: My name is John and I am 30 years old.

442. How would you handle errors in Python code when printing output?

When printing output in Python, it is important to handle errors properly to avoid crashing the

program. One way to handle errors is to use a try-except block to catch any exceptions that might occur. For example:

```
try:  
    print("Hello" + 5)  
except TypeError:  
    print("Error: cannot concatenate string and integer")
```

This code attempts to concatenate a string and an integer, which will raise a `TypeError`. The `try-except` block catches the error and prints a more helpful error message to the console.

443. What is the purpose of `sys.stdout` and how can it be used?

`sys.stdout` is a file object that represents the standard output stream in Python. It can be used to redirect output to a file or to capture output in a variable. For example:

```
import sys  
  
# Redirect output to a file  
sys.stdout = open('output.txt', 'w')  
print("Hello, world!")  
sys.stdout.close()  
  
# Capture output in a variable  
import io  
output = io.StringIO()  
print("Hello, world!", file=output)  
print(output.getvalue())
```

444. Can you explain the difference between `sys.stdout.write()` and `print()`?

`sys.stdout.write()` is a method that writes a string to the standard output stream without adding a newline character. `print()` is a function that writes one or more objects to the standard output stream, separating them with a space and adding a newline character at the end. `print()` is more commonly used for printing output in Python.

445. How would you write a Python function that writes output to a file instead of to the console?

To write output to a file instead of to the console, you can open a file for writing and use the `file` argument of the `print()` function to write to the file instead of to `sys.stdout`. For example:

```
def write_to_file(filename, text):
    with open(filename, 'w') as f:
        print(text, file=f)
```

This function takes a filename and a string of text as arguments, and writes the text to the file with the given filename.

446. How would you use the logging module in Python to output messages to a file?

To use the logging module to output messages to a file, you can create a FileHandler object and add it to a Logger object. For example:

```
import logging

# Create a FileHandler object and set its level to INFO
handler = logging.FileHandler('output.log')
handler.setLevel(logging.INFO)

# Create a Logger object and add the FileHandler to it
logger = logging.getLogger('my_logger')
logger.addHandler(handler)

# Use the Logger object to output messages to the file
logger.info("This message will be written to output.log")
```

447. How can you redirect the output of a Python script to a file instead of to the console?

You can redirect the output of a Python script to a file by using the command-line shell to redirect the standard output stream to a file. For example, to redirect the output of a script called myscript.py to a file called output.txt, you can run the following command:

```
python myscript.py > output.txt
```

This will redirect the standard output stream to the file output.txt.'

448. Can you explain the difference between using print() and logging for outputting information in a Python application?

print() is a built-in Python function that writes information to the standard output stream (usually

the console). It is useful for debugging and quick testing, but it has some limitations. For example, it can be difficult to disable or redirect the output of `print()` in a large application, and it does not provide features like timestamps, levels, or formatting.

The logging module is a more powerful way to output information in a Python application. It provides a flexible and configurable logging system that can write messages to different destinations (such as files or databases) with different levels of severity, timestamps, and formatting. It also provides features like logging levels (DEBUG, INFO, WARNING, ERROR, CRITICAL) and loggers (which can be used to separate log messages by module or subsystem).

449. How can you use the pprint module to print formatted output in Python?

The `pprint` module is a built-in Python module that can be used to pretty-print Python objects in a more human-readable format. It provides a `pprint()` function that takes an object as input and prints it to the console in a formatted way. For example:

```
import pprint

data = {'name': 'John', 'age': 30, 'city': 'New York'}
pprint.pprint(data)
```

This code creates a dictionary called `data` and pretty-prints it using the `pprint()` function. The output will be:

```
{'age': 30, 'city': 'New York', 'name': 'John'}
```

`pprint` can be useful when working with large or complex data structures, as it can make it easier to read and understand the output.

Chapter 3 - Pandas

450. What is Pandas ?

Pandas is a powerful, flexible, open source and easy to use data analysis and manipulation tool. It aims to be the fundamental building block for data analysis , data manipulation tasks.

451. What is python pandas used for ?

Pandas is a open source library of python programming language. Which is mostly use for Data manipulation and analysis.

452. Write Steps to install Pandas on Windows.

These are the following steps :

- The initial step would be to download Python on windows
- Run the Python executable installer
- Install pip on Windows
- Install Pandas in Python using pip
“pip install pandas”

Or install Anaconda, open Jupyter notebook and write
import pandas as pd

453. What are the key features of pandas library ?

These are various features in pandas library :

- Memory Efficient
- Reshaping
- Merge and join
- Time Series
- Data Alignment

454. What is pandas dataframe ?

Pandas dataframe is a 2- dimensional heterogeneous data structure with labeled axes (rows and columns). Pandas dataframe consists of three principle components Data, rows and columns.

455. How to Import Pandas Library and also check the version of Library.

```
# Load the Pandas library with alias pd  
  
import pandas as pd  
  
print(pd.__version__)
```

```
import pandas as pd  
print(pd.__version__)
```

```
1.2.4
```

Output: 1.2.4

456. Mention the different types of Data Structures in Pandas?

Pandas provide two data structures, which are supported by the pandas library, Series, and DataFrames. Both of these data structures are built on top of the NumPy.

A Series is a one-dimensional data structure in pandas, whereas the DataFrame is the two-dimensional data structure in pandas.

457. How to read the different – different format files using pandas??

```
# Load the Pandas library with alias pd import pandas as pd print(pd.__version__)  
# Reading the Comma Separated file pd.read_csv( 'filename.csv')  
# Reading the tab-separated file pd.read_table ('filename.tsv')  
# Reading the Excel File using Pandas
```

```
Pd.read_excel( 'filename.xlsx')  
# Reading the Html file using pandas Pd.read_html('filename.html')
```

458. Define Series in Pandas?

A Series is defined as a one-dimensional array that is capable of storing various data types. The row labels of series are called the index. By using a 'series' method, we can easily convert the list, tuple, and dictionary into series. A Series cannot contain multiple columns.

459. How to create a Series from a numpy array ?

Series: Pandas Series is nothing but a Single Column of the Excel Sheet or we can say that Series is a 1D array capable of holding the data of any type(str, int, float etc)

```
#Load the numpy library with alias np
import numpy as np
# Creating the List
mylist = [1,2,3,4,5]
# Converting the List into Series
Ser1= pd.Series ( Mylist)
print ( Ser1 )
```

```
import numpy as np
# Creating the List
mylist = [1,2,3,4,5]
# Converting the List into series
Ser1= pd.Series ( mylist)
Ser1
```



```
0    1
1    2
2    3
3    4
4    5
dtype: int64
```

```
# Series Creation using Numpy array
# importing numpy library import numpy as np
# data
arr= np.arange(10)
# array to Series Conversion Ser2= pd.Series(arr)
print( Ser2)
```

```
import numpy as np

# data
arr= np.arange(10)

# array to Series Conversion
Ser2= pd.Series(arr)
print( Ser2)
```

Output :

Output :

```
0    0
1    1
2    2
3    3
4    4
5    5
6    6
7    7
8    8
9    9
dtype: int32
```

460.. Create a series using disctionary

```
# Series using Dictionary
Mydic = { 1: 'Monday', 2: 'Tuesday', 3: 'Wednesday', 4: 'Thursday', 5: 'Friday', 6 : 'Saturday'}
Ser3= pd.Series(Mydic)
print(Ser3)
```

```
Mydic = { 1: 'Monday', 2: 'Tuesday',
          3: 'Wednesday', 4: 'Thursday',
          5: 'Friday', 6 : 'Saturday'}

Ser3= pd.Series(Mydic)

print(Ser3)
```

Output:

Output :

```
1      monday
2      Tuesday
3    Wednesday
4    Thursday
5     Friday
6   Saturday
dtype: object
```

461. What is the describe() method in pandas ?

The describe method is used for calculating mean, min, max and standard deviation of each column of the dataset. It analyzes both numeric and object series.

Syntax : dataframe.describe()

462. What is a Data Frame?

A data frame is used to create a two-dimensional array with labeled axes i.e row number and column number. It consists of the following properties:

- The columns can be heterogeneous types like int and bool.
- It can be seen as a dictionary of Series structure where both the rows and columns are indexed. It is denoted as "columns" in the case of columns and "index" in case of rows.

463. Define the different ways a DataFrame can be created in pandas?

We can create a DataFrame using following ways:

- Lists
- Dict of ndarrays

464. Create a dataframe form a List

```
#Create a data frame using a list
xx = ['Xx','Yy']
yy = pd.DataFrame(xx)
print(yy)
print(type(yy))
```

```
In [6]: #Create a data frame using a list
xx = ['Xx', 'Yy']
yy = pd.DataFrame(xx)
print(yy)
print(type(yy))

0
0  Xx
1  Yy
<class 'pandas.core.frame.DataFrame'>
```

465. Create a DataFrame from dict of ndarrays:

```
info = {'State_id' :[101, 102, 103],'Name' :['Bihar','M.P.','Karnataka',]}
info = pd.DataFrame(info)
print (info)
```

```
In [8]: info = {'State_id' :[101, 102, 103], 'Name' :['Bihar', 'M.P.', 'Karnataka', ]}
info = pd.DataFrame(info)
print (info)

   State_id      Name
0      101      Bihar
1      102      M.P.
2      103  Karnataka
```

466. Describe how you will get the names of columns of a DataFrame in Pandas?

```
info = {'State_id' :[101, 102, 103],'Name' :['Bihar','M.P.','Karnataka',]}
info = pd.DataFrame(info)
print (info)
for col in info.columns:
    print(col)
list(info.columns)
sorted(info)
```

```
In [13]: info = {'State_id' :[101, 102, 103], 'Name' :['Bihar','M.P.', 'Karnataka',]}
info = pd.DataFrame(info)
print (info)
for col in info.columns:
    print(col)
list(info.columns)
sorted(info)

      State_id      Name
0        101      Bihar
1        102       M.P.
2        103   Karnataka
State_id
Name

Out[13]: ['Name', 'State_id']
```

467. How to Covert Json(Javascript object notation) data into Dataframe ?

Json is widely used data format for data interchange on the web. We can convert json files into dataframe by using pandas library.

```
#Let Suppose this is our Json data obj = """
{"name": "Wes",
"places_lived": ["United States", "Spain", "Germany"],
"pet": null,
"siblings": [{"name": "Scott", "age": 30, "pets": ["Zeus", "Zuko"]},
 {"name": "Katie", "age": 38,
 "pets": ["Sixes", "Stache", "Cisco"]}]
}
"""

# importing all necessary libraries import pandas as pd
import json
# loading the file result = json.loads(obj)
# dataframe creation
df = pd.DataFrame(result['siblings'], columns=['name', 'age', 'pets']) print(df)
```

```

obj = """
{
  "name": "Wes",
  "places_lived": ["United States", "Spain", "Germany"],
  "pet": null,
  "siblings": [{"name": "Scott", "age": 30, "pets": ["Zeus", "Zuko"]},
               {"name": "Katie", "age": 38,
                "pets": ["Sixes", "Stache", "Cisco"]}]
}
"""

# importing all necessary libraries
import pandas as pd
import json

# Loading the file
result = json.loads(obj)

# dataframe creation
df = pd.DataFrame(result['siblings'], columns=['name', 'age', 'pets'])
print(df)

```

Important Pandas functions at a glance

Pandas is no doubt one of the most important library for any Analytics professional. In 8/10 cases your data will be stored in a data frame and you need a good understanding of the capability of Pandas to understand this data

I will always advice people to create their own data set. Don't go after already available data set on the internet.

By doing this you will get fluent in Excel.

By creating data I mean to say, use functions like RANDBETWEEN, RAND, IF ELSE, etc.
Use them extensively

So, We created my own data in Excel, which looks like this

A	B	C	D	E	F
Inventory	Sales	Markup	Markdown	Population	size
13132	23222	1	0	120265	471
15887	25332	0	1	132004	450
12987	22343	1	0	144149	494
15753	25444	0	1	131259	537
14677	24323	1	1	104146	569
15694	25432	0	1	129172	562
16758	26422	0	1	101456	478
14968	24333	1	1	93882	426
17956	27444	0	1	106301	882
18764	28654	0	1	133373	988
19045	29555	0	1	98195	923
10127	20344	1	0	115019	522
16314	26322	0	1	149626	509
15433	25332	0	1	142754	539
10659	20344	1	0	93673	495
17172	27543	0	1	146503	997

468 .Read a file using read_csv()

```
import pandas as pd
xyz = pd.read_csv('/Users/Nitin/Downloads/Kaggle/THD_Inventory.csv')
Get the top and bottom data set
```

469. head() and tail()

```
xyz.head(5)
xyz.tail(5)
Get the complete information about all the columns in the table
```

470.xyz.info()

```
xyz.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
---  --          -----          int64 
 0   Inventory   100 non-null    int64 
 1   Sales        100 non-null    int64 
 2   Markup       100 non-null    int64 
 3   Markdown    100 non-null    int64 
 4   Population   100 non-null    int64 
 5   size         100 non-null    int64 
dtypes: int64(6)
memory usage: 4.8 KB
```

471. You can also take a look at the shape and size of the dataset. We have 100 rows and 6 columns in the original dataset

```
xyz.shape
xyz.size
```

```
In [100]: print(xyz.shape)
print(xyz.size)
```



```
(100, 6)
600
```

472. Choose n number of random sample from the dataset

```
xyz.sample(n=5)
```

473. Get all the standard mathematical analysis of each column of data set

```
xyz.describe()
```

Use xyz.describe().T will transpose the table

```
#Give all the statistical details of the column  
xyz.describe()
```

	Inventory	Sales	Markup	Markdown	Population	size
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	15289.550000	25714.680000	0.480000	0.660000	120465.280000	631.060000
std	2785.356715	3094.94162	0.502117	0.476095	17516.276993	204.633243
min	10127.000000	18222.000000	0.000000	0.000000	90276.000000	411.000000
25%	13131.000000	23379.500000	0.000000	0.000000	106223.500000	468.750000
50%	15282.500000	25432.000000	0.000000	1.000000	120317.000000	537.000000
75%	17595.750000	28107.250000	1.000000	1.000000	135721.500000	852.250000
max	19975.000000	32111.000000	1.000000	1.000000	149626.000000	997.000000

474. Find number of distinct values – This is an important function as it will directly tell you how many categorical variables are there in your dataset

```
xyz.unique()
```

```
xyz.unique()  
  
Out[112]: Inventory      100  
          Sales         70  
          Markup        2  
          Markdown      2  
          Population    100  
          size          85  
          dtype: int64
```

475. How to find if there is any variable/column with missing values in it?

Use `isna().any()` – This will check if there is even one null value in a column or not

```
xyz.isna().any()
```

```
xyz.isna().any()  
Out[109]: Inventory    False  
          Sales      False  
          Markup     False  
          Markdown  False  
          Population False  
          size       False  
          dtype: bool
```

We don't have any NULL value

476. isnull() – This function shows all the row and column with a boolean for the data present in that cell. TRUE means null value, FALSE means Not a null value

```
xyz.isnull()
```

```
xyz.isnull()  
Out[110]:  
   Inventory  Sales  Markup  Markdown  Population  size  
0      False  False  False  False  False  False  
1      False  False  False  False  False  False  
2      False  False  False  False  False  False  
3      False  False  False  False  False  False  
4      False  False  False  False  False  False  
...  
95     False  False  False  False  False  False  
96     False  False  False  False  False  False  
97     False  False  False  False  False  False  
98     False  False  False  False  False  False  
99     False  False  False  False  False  False
```

100 rows × 6 columns

477. Find the number of null values in each column, this set of function tells you if you can ignore a column or not

Unfortunately we do not have any NULL value in our data set

```
xyz.isnull().sum()
```

```
xyz.isnull().sum()
```

```
Out[111]: Inventory      0  
          Sales        0  
          Markup       0  
          Markdown    0  
          Population   0  
          size         0  
          dtype: int64
```

478. Get the name of all the columns

```
xyz.columns
```

479. Get the nsmallest or nlargest values from a column

```
xyz.nsmallest(10,'Sales')
```

```
In [115]: #nsmallest and nlargest values  
xyz.nsmallest(10,'Sales')
```

```
Out[115]:
```

	Inventory	Sales	Markup	Markdown	Population	size
88	10249	18222	1	0	125941	433
89	13997	20000	1	0	117449	469
31	10823	20223	1	0	122176	497
11	10127	20344	1	0	115019	522
14	10659	20344	1	0	93673	495
43	10808	20532	1	0	124737	515
48	10575	20643	1	0	116828	421
64	11897	22123	1	0	92924	427
59	11783	22342	1	0	117924	556
71	11556	22342	1	0	92037	549

480. Now comes loc and iloc – There are a few interviewers who tries to check your basics with loc and iloc

loc – When you use loc function then you need to specify the name of the columns

```
xyz.loc[1:5,['Sales','Size']]
```

This will fetch row number 1,2,3,4 for the column Sales and Size

iloc – When you use iloc function then you need to specify the index of the columns
xyz.iloc[1:5,2:4]

This will fetch row number 1 to 4 from the 2nd and 3rd column index (Remember Python is zero indexed)

481. Slice the date – It means cutting the dataset vertically or horizontally

xyz[1:5]

This will fetch row number 1,2,3,4 from all the columns

482. Group by in Pandas – Very useful pandas function

xyz[['Markdown', 'inventory']].groupby(['Markdown']).mean()

```
In [131]: xyz[['Markdown', 'Markup', 'Inventory']].groupby(['Markdown', 'Markup']).mean()
Out[131]:
```

		Inventory
Markdown	Markup	
0	1	12153.794118
1	0	17546.788462
	1	14520.928571

483. Sort the complete data frame according to one column

xyz.sort_values(by = 'Sales', ascending = False)

```
In [133]: #Sorting in Pandas by column name..Remember here we are using sort_values  
xyz.sort_values(by = 'Sales', ascending = False)
```

Out[133]:

	Inventory	Sales	Markup	Markdown	Population	size
90	12615	32111	1	0	107160	461
84	15418	30988	0	1	120369	434
85	18735	30987	0	1	107231	960
96	10538	30000	1	0	99892	519
77	17691	30000	0	1	140996	932
...
14	10659	20344	1	0	93673	495
11	10127	20344	1	0	115019	522
31	10823	20223	1	0	122176	497
89	13997	20000	1	0	117449	469
88	10249	18222	1	0	125941	433

100 rows × 6 columns

484. Query in data frame

```
xyz.query('Population > 10000')[5]
```

This gets 5 complete rows with Population more than 10000

```
In [142]: xyz.query('Population > 100000')[:5]
```

Out[142]:

	Inventory	Sales	Markup	Population	size
0	13132	23222	1	120265	471
1	15887	25332	0	132004	450
2	12987	22343	1	144149	494
3	15753	25444	0	131259	537
4	14677	24323	1	104146	569

485. Get unique values from a column

```
xyz['Column Name'].unique()
```

```
In [149]: #Get the unique values from a particular column  
xyz['Markup'].unique()
```

```
Out[149]: array([1, 0])
```

486. If you want. to know how many space columns are taking into your computer then use memory_usage

```
xyz.memory_usage()
```

438. Write a file to csv

```
xyz.to_csv('File Name.csv')
```

487. How do you select a column in a DataFrame?

To select a column in a DataFrame, you can use the bracket notation or the dot notation. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')
```

```
# Using bracket notation  
column1 = df['column1']
```

```
# Using dot notation  
column2 = df.column2
```

488. How do you select multiple columns in a DataFrame?

To select multiple columns in a DataFrame, you can pass a list of column names inside the bracket or use the loc function. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')
```

```
# Using bracket notation  
columns = df[['column1', 'column2']]
```

```
# Using loc function  
columns = df.loc[:, ['column1', 'column2']]
```

489. How do you filter rows based on a condition in a DataFrame?

To filter rows based on a condition in a DataFrame, you can use boolean indexing. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')  
  
# Filter rows where column1 is greater than 10  
filtered_df = df[df['column1'] > 10]
```

490. How do you merge two DataFrames in Pandas?

To merge two DataFrames in Pandas, you can use the merge function. For example:

```
import pandas as pd
```

```
df1 = pd.read_csv('data1.csv')  
df2 = pd.read_csv('data2.csv')  
  
merged_df = pd.merge(df1, df2, on='key')
```

491. How do you handle missing values in a DataFrame?

To handle missing values in a DataFrame, you can use the fillna function or the dropna function. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')  
  
# Replace missing values with 0  
df = df.fillna(0)  
  
# Remove rows with missing values  
df = df.dropna()
```

492. How do you group data in a DataFrame?

To group data in a DataFrame, you can use the groupby function. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')  
  
# Group by column1 and calculate the mean of column2 for each group  
grouped_df = df.groupby('column1')['column2'].mean()
```

493. How do you sort a DataFrame by one or more columns?

To sort a DataFrame by one or more columns, you can use the `sort_values` function. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')
```

```
# Sort by column1 in ascending order and then by column2 in descending order
sorted_df = df.sort_values(['column1', 'column2'], ascending=[True, False])
```

494. How do you apply a function to a DataFrame?

To apply a function to a DataFrame, you can use the `apply` function. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')
```

```
# Apply the square function to column1
```

```
df['column1_squared'] = df['column1'].apply(lambda x: x**2)
```

495. How do you rename columns in a DataFrame?

To rename columns in a DataFrame, you can use the `rename` function. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')
```

```
# Rename column1 to new_column1 and column2 to new_column2
```

```
df = df.rename(columns={'column1': 'new_column1', 'column2': 'new_column2'})
```

496. How do you create a new column in a DataFrame?

To create a new column in a DataFrame, you can simply assign a new column name to the DataFrame with the values that you want. For example:

```
import pandas as pd
```

```
df = pd.read_csv('data.csv')
```

```
# Create a new column that is the sum of column1 and column2
```

```
df['column3'] = df['column1'] + df['column2']
```

497. How do you remove a column from a DataFrame?

To remove a column from a Pandas DataFrame, you can use the drop method. Here's an example:

```
import pandas as pd
```

```
# create a sample DataFrame
df = pd.DataFrame({'A': [1, 2, 3], 'B': [4, 5, 6], 'C': [7, 8, 9]})
```

```
# drop column C from the DataFrame
df = df.drop('C', axis=1)
```

In this example, the drop method is used to remove the column named "C" from the DataFrame. The axis=1 argument is used to indicate that we want to drop a column (as opposed to a row, which would be axis=0).

498. How do you export data to a CSV file using Pandas?

To export data to a CSV file using Pandas, you can use the to_csv method. Here's an example:

```
import pandas as pd
```

```
# create a sample DataFrame
df = pd.DataFrame({'A': [1, 2, 3], 'B': [4, 5, 6], 'C': [7, 8, 9]})
```

```
# export the DataFrame to a CSV file
df.to_csv('my_data.csv', index=False)
```

In this example, the to_csv method is used to export the DataFrame to a file named "my_data.csv". The index=False argument is used to exclude the DataFrame's index from the output. If you want to include the index, you can omit this argument or set it to True.

499. Suppose you have a DataFrame named sales_data that contains columns Year, Month, and Sales. Write a code snippet to calculate the total sales for each year and output the results in a new DataFrame with columns Year and Total Sales.

A: Here's one way to do it:

```
# group by year and sum the sales
sales_by_year = sales_data.groupby('Year')['Sales'].sum()

# create a new DataFrame with year and total sales columns
result = pd.DataFrame({'Year': sales_by_year.index, 'Total Sales': sales_by_year.values})
```

The groupby method is used to group the data by year, and the sum method is applied to the

Sales column to calculate the total sales for each year. The resulting object is a Series with the year as the index and the total sales as the values. Then, a new DataFrame is created using a dictionary that maps column names to the Series values. The resulting DataFrame has two columns: Year and Total Sales.

500. Suppose you have a DataFrame named customer_data that contains columns CustomerID, Name, Age, and City. Write a code snippet to filter the data to only include customers who are 18 years old or older and who live in either "New York" or "San Francisco".

A: Here's one way to do it:

```
# filter the data to only include customers who are 18 or older
customer_data = customer_data[customer_data['Age'] >= 18]

# filter the data to only include customers who live in New York or San Francisco
customer_data = customer_data[(customer_data['City'] == 'New York') | (customer_data['City'] == 'San Francisco')]
```

In the first step, the DataFrame is filtered to only include customers who are 18 years old or older by using boolean indexing with the `>=` operator. The resulting DataFrame only contains rows where the Age column is greater than or equal to 18.

In the second step, the DataFrame is further filtered to only include customers who live in either "New York" or "San Francisco". This is done by using boolean indexing with the `|` (or) operator and checking if the City column is equal to either "New York" or "San Francisco". The resulting DataFrame only contains rows where the City column is "New York" or "San Francisco".

501. How do you select a subset of rows and columns from a DataFrame based on certain conditions?

To select a subset of rows and columns based on certain conditions, you can use boolean indexing. For example, to select all rows where the "age" column is greater than 30 and the "gender" column is "female", you can use the following code:

```
df_subset = df[(df['age'] > 30) & (df['gender'] == 'female')]
```

502. How do you group a DataFrame by a certain column and perform an aggregation function on each group?

To group a DataFrame by a certain column and perform an aggregation function on each group, you can use the `groupby` method. For example, to group a DataFrame named `df` by the "category" column and calculate the sum of the "sales" column for each group, you can use the

following code

```
df_grouped = df.groupby('category')['sales'].sum()
```

503. How do you merge two DataFrames based on a common column?

To merge two DataFrames based on a common column, you can use the merge method. For example, to merge two DataFrames named df1 and df2 based on the "id" column, you can use the following code:

```
merged_df = pd.merge(df1, df2, on='id')
```

504. How do you create a new column in a DataFrame that is the result of a calculation involving other columns?

To create a new column in a DataFrame that is the result of a calculation involving other columns, you can simply use basic arithmetic operations. For example, to create a new column named "total" that is the sum of the "quantity" and "price" columns, you can use the following code:

```
df['total'] = df['quantity'] * df['price']
```

505. How do you rename a column in a DataFrame?

To rename a column in a DataFrame, you can use the rename method. For example, to rename the "old_name" column to "new_name" in a DataFrame named df, you can use the following code:

```
df = df.rename(columns={'old_name': 'new_name'})
```

506. How do you sort a DataFrame by one or more columns?

To sort a DataFrame by one or more columns, you can use the sort_values method. For example, to sort a DataFrame named df by the "age" column in descending order and then by the "name" column in ascending order, you can use the following code:

```
df_sorted = df.sort_values(['age', 'name'], ascending=[False, True])
```

507. How do you check if a DataFrame has any missing values?

To check if a DataFrame has any missing values, you can use the isnull method followed by the any method. For example, to check if a DataFrame named df has any missing values, you can use the following code:

```
has_missing = df.isnull().any().any()
```

508. How do you fill in missing values in a DataFrame with a certain value or method?

To fill in missing values in a DataFrame with a certain value or method, you can use the `fillna` method. For example, to fill in all missing values in a DataFrame named `df` with the mean value of the "age" column, you can use the following code:

```
df['age'] = df['age'].fillna(df['age'].mean())
```

509. How do you write a pandas DataFrame to a CSV file?

To write a pandas DataFrame to a CSV file, you can use the `to_csv` method. For example, to write a DataFrame named `df` to a file named "output.csv" with the first column as the index and using a comma as the delimiter, you can use the following code:

```
df.to_csv('output.csv', index=True, sep=',')
```

510. What are some ways to handle missing data in pandas?

Some ways to handle missing data in pandas include: dropping the rows or columns with missing data using `dropna()`, filling in missing data with a specific value using `fillna()`, and interpolating missing data using `interpolate()`.

511. How can you handle duplicates in a pandas DataFrame?

Duplicates in a pandas DataFrame can be handled by dropping them using `drop_duplicates()`, selecting only the first occurrence of each duplicate using `duplicated()` with `keep='first'`, or selecting only the last occurrence of each duplicate using `duplicated()` with `keep='last'`.

512. How can you merge two pandas DataFrames together?

Two pandas DataFrames can be merged together using `merge()`, which joins the DataFrames on a specified column or index.

513. What are some ways to improve the performance of pandas operations?

Some ways to improve the performance of pandas operations include: using vectorized operations instead of loops, minimizing memory usage by selecting the appropriate data types for columns, and using the `chunksize` parameter to process large DataFrames in smaller pieces

514. How can you use pandas to group data by a certain column and perform aggregate functions on the groups?

To group data by a certain column in pandas and perform aggregate functions on the groups,

you can use the groupby() function followed by a specific aggregation function such as sum(), mean(), max(), or min().

515. How can you pivot a pandas DataFrame?

To pivot a pandas DataFrame, you can use the pivot() or pivot_table() function, which reorganizes the DataFrame so that the values of one column become the new column names, and the values of another column become the new row values.

516. How can you use pandas to create a time series plot?

To create a time series plot in pandas, you can use the plot() function with the specified time series column as the index and the value column as the y-axis.

517. How can you use pandas to handle categorical data?

Pandas can handle categorical data using the Categorical data type, which can be created using the pd.Categorical() function. Categorical data can be used for more efficient storage and better performance in operations that involve grouping or sorting by categories.

518. What are some common methods for data manipulation in pandas?

Some common methods for data manipulation in pandas include: selecting columns using indexing or the loc[] and iloc[] functions, filtering rows using boolean indexing, and applying functions to columns or rows using the apply() function.

519. How can you use pandas to perform statistical analysis on a dataset?

Pandas can be used to perform statistical analysis on a dataset using functions such as describe(), which provides summary statistics for each column, or corr(), which calculates the correlation between columns. Pandas also provides many functions for data visualization, such as plot() and hist(), which can help with exploratory data analysis.

520. How can you use pandas to handle time series data?

Time series data can be handled in pandas using the DatetimeIndex data type, which allows for efficient time-based indexing and slicing. Time series data can also be resampled or aggregated using functions such as resample() or rolling(), and shifted in time using shift(). Additionally, pandas provides functions for calculating rolling windows statistics, handling time zones, and parsing date strings.

521. What are some techniques for reshaping data in pandas?

Techniques for reshaping data in pandas include: pivoting data using pivot() or pivot_table(), melting data using melt(), stacking and unstacking data using stack() and unstack(), and transforming wide format data to long format using melt() and stack().

522. How can you use pandas to handle text data?

Text data can be handled in pandas using functions such as str.split() to split strings into columns, str.extract() to extract specific patterns from strings, and str.replace() to replace patterns in strings. Text data can also be normalized using functions such as str.lower() or str.upper(), and cleaned using regular expressions.

523. What are some ways to handle outliers and anomalies in a pandas DataFrame?

Ways to handle outliers and anomalies in a pandas DataFrame include: dropping rows with extreme values using quantiles or z-scores, capping extreme values using quantiles or truncation, and transforming data using functions such as logarithmic or power transformations.

524. How can you use pandas to handle large datasets that don't fit in memory?

Pandas can handle large datasets that don't fit in memory using functions such as read_csv() or read_table() with the chunksize parameter, which reads the data in smaller chunks and concatenates them into a single DataFrame. Alternatively, pandas can handle large datasets using the Dask library, which provides parallel processing and lazy evaluation.

525. How can you use pandas to handle imbalanced datasets?

Imbalanced datasets can be handled in pandas using techniques such as oversampling, undersampling, or a combination of both. The imbalanced-learn library provides functions for generating synthetic data using oversampling or undersampling techniques.

526. How can you use pandas to handle multi-index DataFrames?

Multi-index DataFrames can be handled in pandas using functions such as groupby() with multiple columns, set_index() with multiple columns, or pivot_table() with multiple columns. Multi-index DataFrames can also be flattened using functions such as reset_index().

527. What are some techniques for feature engineering in pandas?

Feature engineering techniques in pandas include: creating new columns using functions such as apply(), transform(), or groupby(), encoding categorical data using techniques such as one-hot encoding, label encoding, or frequency encoding, and scaling or normalizing numerical data using functions such as StandardScaler() or MinMaxScaler().

528. How can you use pandas to handle data from multiple sources or files?

Pandas can handle data from multiple sources or files using functions such as concat() or merge() to combine multiple DataFrames, or read_csv() or read_table() with multiple files specified as a list.

529. What are some advanced visualization techniques in pandas?

Advanced visualization techniques in pandas include: creating faceted plots using FacetGrid(), plotting multiple axes on a single figure using subplots(), creating interactive plots using hvplot() or bokeh, and creating customized plots using matplotlib.

530. How can you create a scatter plot in pandas?

You can create a scatter plot in pandas using the plot method with the kind='scatter' parameter. For example, if you have a DataFrame df with columns 'x' and 'y', you can create a scatter plot as follows:

```
df.plot(kind='scatter', x='x', y='y').
```

531. How can you create a bar chart in pandas?

You can create a bar chart in pandas using the plot method with the kind='bar' parameter. For example, if you have a DataFrame df with a column 'category' and a column 'count', you can create a bar chart as follows: df.plot(kind='bar', x='category', y='count').

532. How can you create a line plot in pandas?

You can create a line plot in pandas using the plot method with the kind='line' parameter. For example, if you have a DataFrame df with a column 'x' and a column 'y', you can create a line plot as follows: df.plot(kind='line', x='x', y='y').

533. How can you create a histogram in pandas?

You can create a histogram in pandas using the plot method with the kind='hist' parameter. For example, if you have a DataFrame df with a column 'x', you can create a histogram as follows: df.plot(kind='hist', y='x').

534. How can you create a box plot in pandas?

You can create a box plot in pandas using the plot method with the kind='box' parameter. For example, if you have a DataFrame df with a column 'category' and a column 'value', you can create a box plot as follows: df.plot(kind='box', x='category', y='value').

535. How can you create a heatmap in pandas?

You can create a heatmap in pandas using the heatmap function from the seaborn library. For example, if you have a DataFrame df with columns 'x', 'y', and 'z', you can create a heatmap as follows:

```
import seaborn as sns
import pandas as pd

pivot_df = pd.pivot_table(df, values='z', index='x', columns='y')
sns.heatmap(pivot_df)
```

536. How can you customize the appearance of a plot in pandas?

You can customize the appearance of a plot in pandas using various parameters in the plot method. For example, you can set the title of a plot using the title parameter, set the axis labels using the xlabel and ylabel parameters, and set the size of the figure using the figsize parameter. Additionally, you can use various functions from the matplotlib library to customize the appearance of the plot further.

537. How can you create multiple plots on a single figure in pandas?

You can create multiple plots on a single figure in pandas by calling the plot method on the same DataFrame multiple times with different parameters. For example, if you have a DataFrame df with columns 'x', 'y1', and 'y2', you can create a figure with two subplots, one for each column 'y1' and 'y2', as follows:

```
import matplotlib.pyplot as plt

fig, ax = plt.subplots(2, 1)
df.plot(kind='line', x='x', y='y1', ax=ax[0])
df.plot(kind='line', x='x',
```

538. How can you convert a series of strings to a series of numbers in pandas, where some of the strings may contain non-numeric characters or missing values?

To convert a series of strings to a series of numbers in pandas, you can use the to_numeric() function. This function will try to convert each element in the series to a number. To handle missing values or non-numeric characters, you can specify the errors parameter. For example, to convert a series s to numeric values and replace any non-numeric values with NaN:

```
s = pd.to_numeric(s, errors='coerce')
```

539. How can you identify and handle multicollinearity in a pandas DataFrame?

Multicollinearity can occur when there are high correlations between the predictor variables in a regression model. One way to identify multicollinearity in a pandas DataFrame is to compute the correlation matrix using the corr() function. You can then look for high correlation coefficients between pairs of predictor variables. To handle multicollinearity, you can use techniques such as principal component analysis (PCA) or ridge regression to reduce the impact of the correlated variables on the model.

540. How can you select rows from a pandas DataFrame that satisfy a condition based on the values in multiple columns?

To select rows from a pandas DataFrame that satisfy a condition based on the values in multiple columns, you can use the loc[] function. For example, to select all rows where the values in column A are greater than 1 and the values in column B are less than 10:

```
df.loc[(df['A'] > 1) & (df['B'] < 10)]
```

541. How can you compute the rolling standard deviation of a time series in pandas, where the rolling window size varies over time?

To compute the rolling standard deviation of a time series in pandas, where the rolling window size varies over time, you can use the rolling() function. For example, to compute the rolling standard deviation of a time series s using a window size that varies between 5 and 10:

```
s.rolling(window=np.random.randint(5, 11), min_periods=1).std()
```

542. How can you handle imbalanced data in a pandas DataFrame for a classification problem?

To handle imbalanced data in a pandas DataFrame for a classification problem, you can use techniques such as oversampling, undersampling, or SMOTE (Synthetic Minority Over-sampling Technique). You can also use cost-sensitive learning, where you assign different costs to misclassifying different classes.

543. How can you handle data with a mix of categorical and continuous variables in a pandas DataFrame for a regression problem?

To handle data with a mix of categorical and continuous variables in a pandas DataFrame for a regression problem, you can use techniques such as one-hot encoding, which creates binary variables for each possible category in a categorical variable, and scaling the continuous variables using techniques such as standardization or normalization.

544. How can you perform anomaly detection in a pandas DataFrame using statistical methods?

To perform anomaly detection in a pandas DataFrame using statistical methods, you can use techniques such as the Z-score method, which identifies values that are a certain number of standard deviations away from the mean, or the interquartile range (IQR) method, which identifies values that are outside a certain range based on the IQR.

545. How can you perform sentiment analysis on text data in a pandas DataFrame?

To perform sentiment analysis on text data in a pandas DataFrame, you can use natural language processing (NLP) techniques such as tokenization, stemming, and lemmatization to preprocess the text, and then use machine learning models such as Naive Bayes or Support Vector Machines to classify the sentiment.

546. How can you use pandas to perform feature selection and feature extraction for a machine learning model?

To use pandas to perform feature selection and feature extraction for a machine learning model, you can use techniques such as mutual information, which measures the dependence between a feature and the target variable, or principal component analysis (PCA), which identifies linear combinations of features that explain the most variance in the data.

547. How can you use pandas to perform hyperparameter tuning for a machine learning model?

To use pandas to perform hyperparameter tuning for a machine learning model, you can use techniques such as grid search or randomized search to search over a range of hyperparameters and evaluate the performance of the model using cross-validation. You can also use techniques such as Bayesian optimization to search for hyperparameters more efficiently.

548. How can you optimize the memory usage of a pandas DataFrame?

Use the appropriate data types for each column to minimize memory usage, such as using the int8 or uint8 data type for columns with small integer values.

Remove any unnecessary columns or rows that are not needed for analysis or modeling.

Use the chunksize parameter in methods like `read_csv` to read in large datasets in smaller chunks, reducing memory usage.

Use the `astype` method to convert columns to the appropriate data type after loading the data.

Use the `memory_usage` method to check the memory usage of a DataFrame and identify potential areas for optimization.

549. How can you handle missing values in time series data in a way that preserves the time series structure?

To handle missing values in time series data while preserving the time series structure, you can use techniques such as interpolation, forward filling, and backward filling. These techniques fill in missing values with estimates based on the adjacent data points. However, you need to be careful when applying these techniques because they can introduce biases into the data. For example, forward filling can introduce a bias towards the past and backward filling can introduce a bias towards the future.

Another technique is to use imputation methods that take into account the time series structure, such as linear regression imputation or time series imputation models like ARIMA or Prophet.

550. How can you handle imbalanced data in a multi-class classification problem in a way that maintains class balance during training?

To handle imbalanced data in a multi-class classification problem, you can use techniques such as oversampling, undersampling, or a combination of both. Oversampling involves increasing the number of samples in the minority class, while undersampling involves decreasing the number of samples in the majority class. You can also use techniques like SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling) to generate synthetic samples for the minority class.

Another technique is to use class weights during model training to give more importance to the minority class. You can also use ensemble methods like bagging or boosting to improve the performance of the minority class.

551. How can you use pandas to perform time series forecasting using machine learning models?

To perform time series forecasting using machine learning models in pandas, you can use techniques like moving averages, exponential smoothing, ARIMA (Autoregressive Integrated Moving Average), or Prophet. These models take into account the temporal dependence of the data and use it to make predictions about future values.

You can also use more advanced machine learning models like LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit) to capture longer-term dependencies in the data. These models are part of the deep learning family and require more computational power to train.

552. How can you use pandas to handle very large datasets that don't fit into memory on a single machine?

To handle very large datasets in pandas that don't fit into memory on a single machine, you can use techniques such as out-of-core processing or distributed computing. Out-of-core processing involves processing the data in smaller chunks that can fit into memory and then combining the results.

Distributed computing involves splitting the data across multiple machines and processing it in parallel. You can use distributed computing frameworks like Dask, Apache Spark, or Ray to perform distributed computations on large datasets.

553. How can you use pandas to perform natural language processing on text data in a DataFrame?

To perform natural language processing on text data in a pandas DataFrame, you can use techniques such as tokenization, stemming, and lemmatization to preprocess the text data. You can also use techniques like bag-of-words or TF-IDF (Term Frequency-Inverse Document Frequency) to convert the text data into numerical features that can be used for modeling.

More advanced techniques include word embeddings like Word2Vec or GloVe, which represent words as vectors in a high-dimensional space, and deep learning models like CNNs (Convolutional Neural Networks) or LSTMs that can process sequential data like text.

554. How can you use pandas to perform unsupervised learning on a DataFrame with mixed data types?

Unsupervised learning on a DataFrame with mixed data types can be performed using techniques such as clustering, dimensionality reduction, and anomaly detection. One approach is to first preprocess the data by scaling and encoding the features appropriately, and then using techniques such as K-means clustering, Principal Component Analysis (PCA), or t-SNE to identify patterns in the data.

555. How can you use pandas to perform deep learning on a DataFrame with image data?

Pandas is not typically used for deep learning on a DataFrame with image data, as deep learning typically requires specialized libraries such as TensorFlow or PyTorch. However, pandas can be used to preprocess and clean the data before feeding it into a deep learning model.

556. How can you use pandas to perform distributed computing across a cluster of machines?

Pandas is not designed for distributed computing across a cluster of machines, as it is primarily designed for in-memory data processing on a single machine. However, there are some tools that integrate with pandas to enable distributed computing, such as Dask and Apache Arrow.

557. How can you use pandas to handle streaming data in real-time?

Pandas is not designed for handling streaming data in real-time, as it is designed for batch processing of data. However, there are some libraries that integrate with pandas to enable streaming data processing, such as Apache Kafka and Apache Spark. These libraries can be used to ingest and process streaming data, and then store the results in a pandas DataFrame for further analysis.

Chapter 4 - Numpy

558. What is NumPy, and what are its main features?

NumPy (short for Numerical Python) is a popular Python library used for numerical computing. It provides efficient and convenient ways to work with arrays, matrices, and other multidimensional data structures, as well as a range of mathematical and statistical functions for working with this data.

Some of the main features of NumPy include:

N-dimensional array objects: NumPy's core feature is its array object, which is an N-dimensional array that provides a powerful way to store and manipulate large amounts of numerical data.

Broadcasting: NumPy allows for operations on arrays of different shapes and sizes, by automatically broadcasting (replicating) values to match the shapes of the operands.

Mathematical functions: NumPy provides a wide range of mathematical functions for manipulating arrays, including arithmetic operations, trigonometric functions, logarithms, and more.

Linear algebra functions: NumPy provides linear algebra functions such as matrix multiplication, matrix inversion, and eigenvalues and eigenvectors.

Fourier transform: NumPy has functions for computing fast Fourier transforms (FFT), which are important for signal processing and other applications.

Integration with other libraries: NumPy is often used in combination with other scientific computing libraries such as SciPy, Matplotlib, and Pandas.

Efficient memory management: NumPy's implementation is optimized for efficient memory use, which is particularly important when working with large arrays of data.

Overall, NumPy provides a solid foundation for numerical and scientific computing in Python, making it a popular choice for data analysts, scientists, and engineers.

559. How do you create a NumPy array from a Python list?

You can create a NumPy array from a Python list by using the `numpy.array()` function. Here's an example:

```
import numpy as np
```

```
my_list = [1, 2, 3, 4, 5]
my_array = np.array(my_list)
print(my_array)
```

In this example, we import the NumPy library and create a Python list called `my_list`. We then pass this list as an argument to the `numpy.array()` function to create a new NumPy array called `my_array`. Finally, we print the contents of `my_array` to the console using the `print()` function.

Note that when we create a NumPy array from a Python list, NumPy will automatically infer the data type of the elements in the array based on the types of the elements in the list. If the list contains a mix of data types (e.g., integers and floats), NumPy will choose the most general data type that can represent all the elements in the list.

560. How do you perform element-wise arithmetic operations between two NumPy arrays?

You can perform element-wise arithmetic operations between two NumPy arrays using the standard arithmetic operators (+, -, *, /, **, etc.). Here's an example:

```
import numpy as np

# create two NumPy arrays
arr1 = np.array([1, 2, 3])
arr2 = np.array([4, 5, 6])

# perform element-wise addition
result = arr1 + arr2
print(result)

# perform element-wise multiplication
result = arr1 * arr2
print(result)
```

Output
[5 7 9]
[4 10 18]

In this example, we create two NumPy arrays `arr1` and `arr2`, which both have three elements. We then perform element-wise addition and multiplication on these arrays using the `+` and `*` operators, respectively. The resulting arrays are stored in the `result` variable and printed to the console using the `print()` function.

Note that when performing element-wise arithmetic operations between two NumPy arrays, the arrays must have the same shape. If the arrays have different shapes, NumPy will attempt to broadcast the arrays to a common shape before performing the operation. If broadcasting is not possible, NumPy will raise a ValueError.

561. How do you calculate the dot product of two NumPy arrays?

You can calculate the dot product of two NumPy arrays using the `numpy.dot()` function. Here's an example:

```
import numpy as np
```

```
# create two NumPy arrays
arr1 = np.array([1, 2, 3])
arr2 = np.array([4, 5, 6])
```

```
# calculate the dot product
result = np.dot(arr1, arr2)
print(result)
```

Output

32

In this example, we create two NumPy arrays `arr1` and `arr2`, which both have three elements. We then calculate the dot product of these arrays using the `numpy.dot()` function and store the result in the `result` variable. Finally, we print the result to the console using the `print()` function.

Note that the dot product is only defined for arrays with the same number of dimensions, so both arrays must be one-dimensional or two-dimensional. If both arrays are two-dimensional, the dot product is calculated as a matrix multiplication. If one or both arrays are one-dimensional, the dot product is calculated as the sum of the element-wise product of the two arrays.

562. What is broadcasting in NumPy, and how does it work?

Broadcasting is a feature of NumPy that allows arrays with different shapes to be used in arithmetic operations. When an arithmetic operation is performed between two arrays, NumPy compares their shapes element-wise starting from the trailing dimensions, and broadcasts the smaller array to have the same shape as the larger one.

The broadcasting rules are as follows:

If the arrays have the same number of dimensions, but different shapes, the array with the smaller shape is broadcast to the larger shape by adding dimensions of size 1 to its shape until the shapes match.

If the arrays have different numbers of dimensions, the array with fewer dimensions is broadcast to have the same number of dimensions as the other array, by adding dimensions of size 1 to its shape.

If the size of a dimension in either array is 1, the array is broadcast along that dimension to match the size of the corresponding dimension in the other array.

Here's an example to illustrate how broadcasting works:

```
import numpy as np

# create a 2D array
arr1 = np.array([[1, 2, 3],
                 [4, 5, 6]])

# create a 1D array
arr2 = np.array([1, 2, 3])

# add the 2D array and the 1D array
result = arr1 + arr2
print(result)
```

Output

```
[[2 4 6]
 [5 7 9]]
```

In this example, we create a 2D NumPy array arr1 with shape (2, 3) and a 1D NumPy array arr2 with shape (3,). When we add these two arrays using the + operator, NumPy broadcasts the 1D array arr2 to a 2D array with shape (2, 3) by adding a new dimension of size 1 to its shape. The resulting arrays are then added element-wise, and the resulting 2D array is stored in the result variable and printed to the console using the print() function.

Broadcasting can save memory by eliminating the need to create copies of arrays with different shapes. It can also make the code more concise and readable by reducing the need for explicit reshaping or tiling of arrays. However, it is important to be aware of the broadcasting rules to avoid unexpected results.

563. What are some common statistical functions available in NumPy, and how do you use them?

NumPy provides a wide range of statistical functions that can be used to analyze and manipulate numerical data. Here are some common statistical functions in NumPy

```
numpy.mean()  
numpy.median()  
numpy.std()  
numpy.min()  
numpy.max()  
numpy.var()
```

564. Define the mean function in numpy and give a simple example

numpy.mean(): computes the arithmetic mean of a given array.

```
import numpy as np
```

```
# create a 1D array  
arr = np.array([1, 2, 3, 4, 5])  
  
# compute the mean  
mean = np.mean(arr)  
print(mean)
```

Output
3

565. Define the median function in numpy and give a simple example

numpy.median(): computes the median of a given array.

```
import numpy as np
```

```
# create a 1D array  
arr = np.array([1, 2, 3, 4, 5])  
  
# compute the median  
median = np.median(arr)  
print(median)
```

Output
3

566. Define the standard deviation function in numpy and give a simple example

numpy.std(): computes the standard deviation of a given array.

```
import numpy as np

# create a 1D array
arr = np.array([1, 2, 3, 4, 5])

# compute the standard deviation
std = np.std(arr)
print(std)
```

Output
1.4142135623730951

567. Define the var function in numpy and give a simple example

numpy.var(): computes the variance of a given array.

```
import numpy as np
```

```
# create a 1D array
arr = np.array([1, 2, 3, 4, 5])

# compute the variance
var = np.var(arr)
print(var)
```

Output
2

568. Define the min function in numpy and give a simple example

numpy.min(): returns the minimum value of a given array.

```
import numpy as np
```

```
# create a 1D array
arr = np.array([1, 2, 3, 4, 5])

# get the minimum value
min_val = np.min(arr)
print(min_val)
```

569. Define the max function in numpy and give a simple example

numpy.max(): returns the maximum value of a given array

```
import numpy as np

# create a 1D array
arr = np.array([1, 2, 3, 4, 5])

# get the maximum value
max_val = np.max(arr)
print(max_val)
```

Output
5

570. How do you select elements from a NumPy array based on a conditional expression?

You can select elements from a NumPy array based on a conditional expression by using boolean indexing. Boolean indexing allows you to select elements from an array based on a boolean mask that has the same shape as the original array.

Here's an example of how to select elements from a NumPy array based on a conditional expression:

```
import numpy as np

# create a 1D array
arr = np.array([1, 2, 3, 4, 5])

# create a boolean mask based on a condition
mask = arr > 2

# use the boolean mask to select elements from the array
selected = arr[mask]

print(selected)
```

Output
[3,4,5]

In this example, we first create a 1D NumPy array containing the values [1, 2, 3, 4, 5]. We then create a boolean mask based on the condition $arr > 2$, which returns a boolean array of the same shape as arr , with the value True at each index where the condition is satisfied and False otherwise.

We then use this boolean mask to select elements from the original array by passing it inside square brackets after the array. The resulting selected array contains only the elements of the original array where the corresponding value in the boolean mask is True.

You can also use boolean indexing to assign new values to the selected elements. For example:

```
import numpy as np
```

```
# create a 1D array
```

```
arr = np.array([1, 2, 3, 4, 5])
```

```
# create a boolean mask based on a condition
```

```
mask = arr > 2
```

```
# assign new values to the selected elements
```

```
arr[mask] = 0
```

```
print(arr)
```

Output

```
[1 2 0 0 0]
```

In this example, we use the same boolean mask as in the previous example to select elements from the original array. However, instead of creating a new array with the selected elements, we assign a new value of 0 to each selected element in the original array using the same boolean mask. The resulting arr array contains the same elements as the original array, but with the values 3, 4, and 5 replaced by 0 based on the boolean mask.

571. How do you reshape a NumPy array?

You can reshape a NumPy array using the reshape() function. The reshape() function returns a new array with the same data as the original array but with a new shape.

Here's an example of how to reshape a NumPy array:

```
import numpy as np
```

```
# create a 1D array
```

```
arr = np.array([1, 2, 3, 4, 5, 6])
```

```
# reshape the array to a 2D array with 2 rows and 3 columns
```

```
new_arr = arr.reshape(2, 3)
```

```
print(new_arr)
```

Output

```
[[1 2 3]
 [4 5 6]]
```

572. Convert a multidimensional array to 1D array

In this example, we first create a 1D NumPy array containing the values [1, 2, 3, 4, 5, 6]. We then use the `reshape()` function to reshape the array to a 2D array with 2 rows and 3 columns. The resulting `new_arr` array has the shape (2, 3) and contains the same elements as the original array, but arranged in a 2D format.

It's important to note that the total number of elements in the reshaped array must be the same as the total number of elements in the original array. In other words, the product of the dimensions of the reshaped array must be equal to the size of the original array. Otherwise, you will get a `ValueError` when you try to reshape the array.

You can also use the `reshape()` function to convert a multi-dimensional array to a 1D array. For example:

```
import numpy as np

# create a 2D array
arr = np.array([[1, 2, 3], [4, 5, 6]])

# reshape the array to a 1D array
new_arr = arr.reshape(-1)

print(new_arr)
```

Output

```
[1 2 3 4 5 6]
```

In this example, we create a 2D NumPy array containing the values [[1, 2, 3], [4, 5, 6]]. We then use the `reshape()` function to reshape the array to a 1D array using the -1 argument, which tells NumPy to automatically calculate the size of the array based on the original shape. The resulting `new_arr` array contains the same elements as the original array, but in a 1D format.

573. How do you perform matrix operations in NumPy, such as matrix multiplication and inversion?

NumPy provides a set of functions for performing matrix operations, including matrix multiplication, inversion, and determinant calculation. Here are some examples:

Matrix Multiplication

You can perform matrix multiplication using the `dot()` function or the `@` operator.

```
import numpy as np

# create two 2x3 matrices
A = np.array([[1, 2, 3], [4, 5, 6]])
B = np.array([[7, 8, 9], [10, 11, 12], [13, 14, 15]])

# compute the matrix product of A and B
C = A.dot(B)
# or equivalently, C = A @ B

print(C)
```

Output

```
[[ 66  72  78]
 [156 171 186]]
```

574. What is Matrix Inversion?

You can perform matrix inversion using the `inv()` function.

```
import numpy as np

# create a 3x3 matrix
A = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

# compute the inverse of A
A_inv = np.linalg.inv(A)

print(A_inv)
```

Output

```
[-1.23333333  0.46666667  0.3      ]
 [ 1.16666667 -0.33333333 -0.16666667]
 [-0.1       0.06666667  0.03333333]]
```

In this example, we create a 3x3 NumPy array A, and then compute its inverse A_inv using the inv() function from the linalg module.

575. What is Determinant calculation?

You can calculate the determinant of a matrix using the det() function.

```
import numpy as np

# create a 3x3 matrix
A = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

# compute the determinant of A
det = np.linalg.det(A)

print(det)
Output

-9.51619735393e-16
```

In this example, we create a 3x3 NumPy array A, and then compute its determinant using the det() function from the linalg module.

Note that the linalg module provides many other matrix operations as well, such as eigenvalue and eigenvector calculation, singular value decomposition, and QR factorization, among others.

576. How do you save and load NumPy arrays from disk?

NumPy is a popular Python library used for numerical computing, and it provides several methods for saving and loading NumPy arrays from disk. Here are some examples:

Saving NumPy arrays: NumPy provides the numpy.save() and numpy.savez() methods for saving NumPy arrays to disk. The numpy.save() method can be used to save a single NumPy array, while the numpy.savez() method can be used to save multiple NumPy arrays in a single compressed file. Here's an example of how to use the numpy.save() method:

```
import numpy as np

# create a NumPy array
arr = np.array([1, 2, 3, 4, 5])

# save the array to disk
```

```
np.save('my_array.npy', arr)
```

Loading NumPy arrays: NumPy provides the `numpy.load()` method for loading a saved NumPy array from disk. Here's an example of how to use the `numpy.load()` method to load a previously saved array:

```
import numpy as np

# load a saved NumPy array from disk
arr = np.load('my_array.npy')

# print the contents of the array
print(arr)
```

The above code will load the array saved in the 'my_array.npy' file, and print its contents.

It is worth noting that NumPy arrays can also be saved and loaded in other formats, such as CSV or text files, using methods like `numpy.savetxt()` and `numpy.loadtxt()`. However, using the `numpy.save()` and `numpy.load()` methods is generally faster and more efficient for working with large numerical arrays.

577. How do you concatenate two or more NumPy arrays horizontally?

NumPy provides the `numpy.hstack()` method for concatenating two or more arrays horizontally. This method stacks the arrays side by side along the second axis (i.e., columns). Here's an example of how to use the `numpy.hstack()` method:

```
import numpy as np

# create two NumPy arrays
arr1 = np.array([[1, 2], [3, 4]])
arr2 = np.array([[5, 6], [7, 8]])

# concatenate the arrays horizontally
result = np.hstack((arr1, arr2))

# print the result
print(result)
```

578. How do you concatenate two or more NumPy arrays vertically?

NumPy provides the `numpy.vstack()` method for concatenating two or more arrays vertically.

This method stacks the arrays on top of each other along the first axis (i.e., rows). Here's an example of how to use the `numpy.vstack()` method:

```
import numpy as np

# create two NumPy arrays
arr1 = np.array([[1, 2], [3, 4]])
arr2 = np.array([[5, 6], [7, 8]])

# concatenate the arrays vertically
result = np.vstack((arr1, arr2))

# print the result
print(result)
```

579. How do you concatenate two or more NumPy arrays arbitrarily?

umPy also provides the `numpy.concatenate()` method for concatenating arrays along an arbitrary axis. This method takes a tuple of arrays as input, and an optional axis parameter to specify the axis along which to concatenate the arrays. Here's an example of how to use the `numpy.concatenate()` method:

```
import numpy as np

# create three NumPy arrays
arr1 = np.array([[1, 2], [3, 4]])
arr2 = np.array([[5, 6], [7, 8]])
arr3 = np.array([[9, 10], [11, 12]])

# concatenate the arrays along the second axis
result = np.concatenate((arr1, arr2, arr3), axis=1)

# print the result
print(result)
```

The above code will concatenate the `arr1`, `arr2`, and `arr3` arrays along the second axis, and print the resulting array.

580. How do you create a masked array in NumPy, and what is its purpose?

In NumPy, a masked array is an array that has some entries marked as invalid or masked, such that they are ignored in various calculations. The purpose of masked arrays is to allow for efficient handling of missing or invalid data in numerical calculations.

```
import numpy as np

# create a NumPy array
arr = np.array([1, 2, 3, -999, 5])

# create a masked array
mask = arr != -999
masked_arr = np.ma.masked_array(arr, mask)

# print the masked array
print(masked_arr)
```

In the above code, we first create a NumPy array with some missing or invalid data represented by the value -999. We then create a mask for the array using a boolean expression, such that any element in the array with the value -999 is marked as invalid. Finally, we create a masked array using the `numpy.ma.masked_array()` function, which takes the original array and the mask as input.

The resulting masked array contains the same data as the original array, but with the masked values marked as invalid. We can perform various operations on the masked array, such as calculating the mean or sum, and the masked values will be ignored in these calculations.

For example, to calculate the mean of the masked array:

```
# calculate the mean of the masked array
mean = np.ma.mean(masked_arr)

# print the mean
print(mean)
```

Here's an example of how to create a masked array in NumPy:

```
import numpy as np

# create a NumPy array
arr = np.array([1, 2, 3, -999, 5])
```

```
# create a masked array
mask = arr != -999
masked_arr = np.ma.masked_array(arr, mask)

# print the masked array
print(masked_arr)
```

581. What is a shallow copy in a NumPy array?

In NumPy, copying an array creates a new array that is a copy of the original array. However, there are two different ways to copy an array: shallow copying and deep copying. The difference between the two is in how they copy the data in the array.

A shallow copy of an array creates a new array that is a view of the original array. This means that the new array shares the same data as the original array, and any modifications to the new array will also affect the original array. In other words, a shallow copy creates a new array object, but the data in the new array still points to the same data as the original array. NumPy provides the `numpy.view()` method to create a shallow copy of an array.

Here's an example of how to create a shallow copy of a NumPy array:

```
import numpy as np
```

```
# create a NumPy array
arr = np.array([1, 2, 3, 4, 5])

# create a shallow copy of the array
arr_shallow = arr.view()

# modify the shallow copy
arr_shallow[0] = 100

# print both arrays
print(arr)
print(arr_shallow)
```

In the above code, we create a shallow copy of the original array `arr` using the `view()` method. We then modify the first element of the shallow copy, which also changes the first element of the original array. Finally, we print both arrays to confirm that they have been modified.

582. What is a deep copy in NumPy?

A deep copy of an array creates a new array that is a completely separate copy of the original array. This means that the new array has its own data, which is a copy of the data in the original array. Any modifications to the new array will not affect the original array. NumPy provides the `numpy.copy()` method to create a deep copy of an array.

Here's an example of how to create a deep copy of a NumPy array:

```
import numpy as np
```

```
# create a NumPy array
arr = np.array([1, 2, 3, 4, 5])

# create a deep copy of the array
arr_deep = arr.copy()

# modify the deep copy
arr_deep[0] = 100

# print both arrays
print(arr)
print(arr_deep)
```

In the above code, we create a deep copy of the original array `arr` using the `copy()` method. We then modify the first element of the deep copy, which does not affect the original array. Finally, we print both arrays to confirm that only the deep copy has been modified.

583. How do you generate random numbers in NumPy, and what are some common distributions you can sample from?

In NumPy, you can generate random numbers using the `numpy.random` module, which provides functions for various probability distributions. Here's an example of how to generate random numbers from a normal distribution:

```
import numpy as np

# generate 10 random numbers from a normal distribution
mu, sigma = 0, 0.1 # mean and standard deviation
rand_nums = np.random.normal(mu, sigma, 10)

# print the random numbers
```

```
print(rand_nums)
```

In the above code, we use the `numpy.random.normal()` function to generate 10 random numbers from a normal distribution with mean μ and standard deviation σ . We then print the random numbers.

Here are some common probability distributions you can sample from using NumPy's random module:

Uniform distribution: `numpy.random.uniform()`

Normal (Gaussian) distribution: `numpy.random.normal()`

Binomial distribution: `numpy.random.binomial()`

Poisson distribution: `numpy.random.poisson()`

Exponential distribution: `numpy.random.exponential()`

Gamma distribution: `numpy.random.gamma()`

For example, to generate 10 random numbers from a uniform distribution between 0 and 1:

```
# generate 10 random numbers from a uniform distribution
rand_nums = np.random.uniform(0, 1, 10)
```

```
# print the random numbers
```

```
print(rand_nums)
```

In the above code, we use the `numpy.random.uniform()` function to generate 10 random numbers from a uniform distribution between 0 and 1. We then print the random numbers.

584. How do you sort a NumPy array in ascending or descending order?

In NumPy, you can sort a NumPy array in ascending or descending order using the `numpy.sort()` function. The `numpy.sort()` function returns a sorted copy of the input array.

Here's an example of how to sort a NumPy array in ascending order:

```
import numpy as np
```

```
# create a NumPy array
arr = np.array([3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5])
```

```
# sort the array in ascending order
sorted_arr = np.sort(arr)
```

```
# print the sorted array
print(sorted_arr)
```

In the above code, we create a NumPy array arr and then sort it in ascending order using the numpy.sort() function. We then print the sorted array.

To sort a NumPy array in descending order, you can use the numpy.sort() function with the kind parameter set to 'quicksort' and the order parameter set to None, and then reverse the resulting array using the numpy.flip() function. Here's an example:

```
import numpy as np

# create a NumPy array
arr = np.array([3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5])

# sort the array in descending order
sorted_arr = np.sort(arr, kind='quicksort', order=None)[::-1]

# print the sorted array
print(sorted_arr)
```

In the above code, we create a NumPy array arr and then sort it in descending order using the numpy.sort() function with the kind parameter set to 'quicksort' and the order parameter set to None. We then reverse the resulting array using the numpy.flip() function, which gives us the sorted array in descending order. Finally, we print the sorted array.

585. How do you perform element-wise logical operations between two NumPy arrays?

In NumPy, you can perform element-wise logical operations between two NumPy arrays using the bitwise logical operators & (and), | (or), ^ (xor), and ~ (not). These operators work element-wise on the input arrays, and return a new NumPy array with the same shape as the input arrays.

Here's an example of how to perform element-wise logical operations between two NumPy arrays:

```
import numpy as np

# create two NumPy arrays
arr1 = np.array([True, False, True, False])
arr2 = np.array([True, True, False, False])

# perform element-wise logical operations
and_arr = arr1 & arr2 # element-wise and
or_arr = arr1 | arr2 # element-wise or
```

```

xor_arr = arr1 ^ arr2 # element-wise xor
not_arr = ~arr1 # element-wise not

# print the results
print(and_arr)
print(or_arr)
print(xor_arr)
print(not_arr)

```

In the above code, we create two NumPy arrays `arr1` and `arr2`, and then perform element-wise logical operations on them using the `&`, `|`, `^`, and `~` operators. We store the results in new NumPy arrays `and_arr`, `or_arr`, `xor_arr`, and `not_arr`, respectively. Finally, we print the results.

Note that the input arrays must have the same shape for the element-wise logical operations to work. If the input arrays have different shapes, you can use NumPy's broadcasting rules to perform the operations.

586. How do you compute the Fourier transform of a signal using NumPy?

In NumPy, you can compute the Fourier transform of a signal using the `numpy.fft.fft()` function. The Fourier transform is a mathematical technique that allows you to decompose a signal into its frequency components. The `numpy.fft.fft()` function computes the complex Fourier transform of an input signal, and returns an array of complex numbers representing the frequency spectrum of the input signal.

Here's an example of how to compute the Fourier transform of a signal using NumPy

```

import numpy as np
import matplotlib.pyplot as plt

# create a time vector
t = np.linspace(0, 1, 1000)

# create a signal
f = 10 # signal frequency
signal = np.sin(2 * np.pi * f * t)

# compute the Fourier transform
spectrum = np.fft.fft(signal)

# plot the signal and spectrum
fig, (ax1, ax2) = plt.subplots(nrows=2, sharex=True, figsize=(8, 6))

```

```

ax1.plot(t, signal)
ax1.set_ylabel('Amplitude')
ax1.set_title('Signal')

freqs = np.fft.fftfreq(len(signal), t[1] - t[0])
ax2.plot(freqs, np.abs(spectrum))
ax2.set_xlabel('Frequency (Hz)')
ax2.set_ylabel('Amplitude')
ax2.set_xlim(0, 2 * f)
ax2.set_title('Frequency Spectrum')

plt.tight_layout()
plt.show()

```

In the above code, we first create a time vector `t` using the `numpy.linspace()` function. We then create a signal by computing a sinusoidal wave with frequency `f` using the `numpy.sin()` function. We then compute the Fourier transform of the signal using the `numpy.fft.fft()` function, and store the resulting complex array in `spectrum`.

To plot the signal and its spectrum, we use the `matplotlib.pyplot.subplots()` function to create a figure with two subplots. We plot the signal on the first subplot using the `matplotlib.pyplot.plot()` function, and plot the frequency spectrum on the second subplot using the `matplotlib.pyplot.plot()` function. We also use the `numpy.fft.fftfreq()` function to compute the frequency values corresponding to the Fourier transform output.

Finally, we use the `matplotlib.pyplot.show()` function to display the figure.

Note that the output of the `numpy.fft.fft()` function is a complex array, and the elements of the array correspond to the frequency components of the signal. The magnitude of each element represents the amplitude of the corresponding frequency component, and the phase angle of each element represents the phase shift of the corresponding frequency component. To get a more meaningful visualization of the frequency spectrum, it is common to plot the magnitude of the complex array using the `numpy.abs()` function.

587. How to use NumPy with SciPy?

Using NumPy with SciPy:

SciPy is a library that extends NumPy with additional functionality for scientific computing. It provides modules for optimization, interpolation, integration, signal and image processing, and more.

To use NumPy with SciPy, you can simply import both libraries and use the functions provided by each library. For example, to perform a least-squares fit of a polynomial to some data:

```
import numpy as np
from scipy.optimize import curve_fit

x = np.array([1, 2, 3, 4, 5])
y = np.array([2.3, 3.2, 4.5, 5.1, 6.7])

def func(x, a, b, c):
    return a * x**2 + b * x + c

popt, pcov = curve_fit(func, x, y)
```

In this example, we use NumPy to create the arrays `x` and `y` to hold our data. We then define a function `func` that takes an array `x` and some coefficients `a`, `b`, and `c`, and returns the value of a quadratic polynomial. Finally, we use the `curve_fit` function from SciPy to perform a least-squares fit of the polynomial to the data.

588. How to use NumPy with matplotlib?

Using NumPy with Matplotlib:

Matplotlib is a plotting library that provides a variety of functions for creating visualizations of data.

To use NumPy with Matplotlib, you can create NumPy arrays to hold your data, and then use Matplotlib functions to plot the data. For example, to create a line plot of a sine wave

```
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0, 2*np.pi, 100)
y = np.sin(x)

plt.plot(x, y)
plt.show()
```

In this example, we use NumPy to create an array `x` that ranges from 0 to 2π , and an array `y` that contains the sine of each value in `x`. We then use the `plot` function from Matplotlib to create a line plot of `y` versus `x`, and the `show` function to display the plot.

Coding Questions

589. Write a NumPy code snippet to create an array of zeros with shape (3, 4).

```
import numpy as np  
zeros_array = np.zeros((3, 4))
```

590. Write a NumPy code snippet to create an array of ones with shape (2, 5).

```
import numpy as np  
  
ones_array = np.ones((2, 5))
```

591. Write a NumPy code snippet to create an array of evenly spaced values between 0 and 10 with a step size of 2.

```
import numpy as np  
  
evenly_spaced_array = np.arange(0, 11, 2)
```

592. Write a NumPy code snippet to create a random array with shape (2, 3) and values between 0 and 1.

```
import numpy as np  
  
random_array = np.random.rand(2, 3)
```

593. Write a NumPy code snippet to calculate the sum of all elements in a two-dimensional array.

```
import numpy as np  
  
arr = np.array([[1, 2], [3, 4]])  
sum_all_elements = arr.sum()
```

594. Write a NumPy code snippet to calculate the mean of all elements in a one-dimensional array.

```
import numpy as np  
arr = np.array([1, 2, 3, 4, 5])  
mean_all_elements = arr.mean()
```

595. Write a NumPy code snippet to calculate the standard deviation of all elements in a one-dimensional array.

```
import numpy as np  
  
arr = np.array([1, 2, 3, 4, 5])  
  
std_all_elements = arr.std()
```

596. Write a NumPy code snippet to calculate the dot product of two one-dimensional arrays.

```
import numpy as np  
  
a = np.array([1, 2, 3])  
b = np.array([4, 5, 6])  
  
dot_product = np.dot(a, b)
```

597. Write a NumPy code snippet to reshape a one-dimensional array into a two-dimensional array with 3 rows and 2 columns.

```
import numpy as np  
  
arr = np.array([1, 2, 3, 4, 5, 6])  
  
reshaped_arr = arr.reshape((3, 2))
```

598. Write a NumPy code snippet to find the index of the maximum value in a one-dimensional array.

```
import numpy as np  
  
arr = np.array([1, 5, 3, 8, 2])  
  
max_value_index = arr.argmax()
```

Data Cleaning using Numpy

599. How do you remove missing or null values from a NumPy array?

NumPy arrays do not have a built-in way to represent missing or null values. However, you can use the `numpy.nan` value to represent missing or null values. To remove these values from a NumPy array, you can use the `numpy.isnan()` function to create a Boolean mask that identifies the missing values, and then use this mask to filter the array using Boolean indexing. For example:

```
import numpy as np

a = np.array([1, 2, 3, np.nan, 4, 5, np.nan])
a = a[~np.isnan(a)]
```

The above code creates a NumPy array `a` with two missing values represented by `numpy.nan`. The `numpy.isnan(a)` function creates a Boolean mask that is True for elements that are missing, and False for elements that are not missing. The `~` operator negates the mask, so it is True for elements that are not missing, and False for elements that are missing. We use this mask to filter the array `a` using Boolean indexing to remove the missing values.

600. How can you identify and remove outliers in a NumPy array?

Outliers are extreme values that are much larger or smaller than the other values in a dataset. They can be identified using various statistical methods, such as the z-score or the interquartile range (IQR).

To remove outliers from a NumPy array using the z-score method, you can compute the z-score of each element in the array, and then filter out the elements that have a z-score outside of a certain threshold. The z-score of an element is the number of standard deviations it is away from the mean of the array. For example:

```
import numpy as np

a = np.array([1, 2, 3, 10, 4, 5, 20, 30])
z = np.abs((a - np.mean(a)) / np.std(a))
threshold = 2
a = a[z <= threshold]
```

The above code creates a NumPy array `a` with two outliers, the values 10 and 20. We compute the z-score of each element in the array using the formula `(a - np.mean(a)) / np.std(a)`. We then

compute a threshold of 2 standard deviations, and use the Boolean mask $z \leq \text{threshold}$ to filter the array a using Boolean indexing to remove the outliers.

601. What are some common techniques for normalizing data in a NumPy array?

Normalizing data in a NumPy array is the process of scaling the values of the array so that they are in a common range, typically between 0 and 1 or -1 and 1. Normalization can help to improve the performance of machine learning algorithms, as well as the interpretability of the data.

Two common techniques for normalizing data in a NumPy array are min-max scaling and z-score scaling.

Min-max scaling scales the values of the array to a range between 0 and 1. To min-max scale a NumPy array, you can use the formula:

```
a_min = np.min(a)
a_max = np.max(a)
a_scaled = (a - a_min) / (a_max - a_min)
```

Z-score scaling scales the values of the array to a standard normal distribution with a mean of 0 and a standard deviation of 1. To z-score scale a NumPy array, you can use the formula:

```
a_mean = np.mean(a)
a_std = np.std(a)
a_scaled = (a - a_mean) / a_std
```

602. How do you sort a NumPy array, and what are some of the options for customizing the sort?

You can sort a NumPy array using the `np.sort()` function, which returns a sorted copy of the original array. By default, the elements are sorted in ascending order along the last axis of the array.

Here's an example of using `np.sort()` to sort a one-dimensional array in ascending order:

```
import numpy as np

arr = np.array([3, 1, 4, 1, 5, 9, 2, 6, 5, 3])

sorted_arr = np.sort(arr)

print(sorted_arr) # output: [1 1 2 3 3 4 5 5 6 9]
```

If you want to sort the array in descending order, you can use the `[::-1]` slice to reverse the order of the sorted elements:

```
desc_sorted_arr = np.sort(arr)[::-1]  
  
print(desc_sorted_arr) # output: [9 6 5 5 4 3 3 2 1 1]
```

You can also use the `np.argsort()` function to return the indices that would sort the array, rather than the sorted array itself:

```
idx = np.argsort(arr)  
  
print(idx) # output: [1 3 6 0 8 2 4 9 7 5]
```

You can customize the behavior of the `sort` function by specifying optional arguments. For example, you can use the `axis` argument to sort along a particular axis of a multi-dimensional array, or use the `kind` argument to specify the sorting algorithm (e.g., 'quicksort', 'mergesort', or 'heapsort'). You can also use the `order` argument to sort structured arrays by a particular field, or use the `na_position` argument to specify the handling of NaN values during the sort.

603. What are some functions in NumPy that are commonly used for data cleaning?

Functions commonly used for data cleaning in NumPy:

- `numpy.nan_to_num`: replaces NaN values with zeros or other specified values.
- `numpy.interp`: linearly interpolates missing values in an array.
- `numpy.delete`: removes specified rows or columns from an array.
- `numpy.unique`: returns unique values in an array.
- `numpy.isin`: identifies values in an array that are present in a specified list or array.
- `numpy.where`: replaces values in an array that meet a specified condition with another value.

604. How can you handle duplicate values in a NumPy array?

Handling duplicate values in a NumPy array:

- `numpy.unique`: returns only the unique values in an array.
- `numpy.delete`: removes duplicates from an array.
- `numpy.concatenate`: concatenates arrays along a specified axis.

Considerations when concatenating arrays include ensuring that the arrays have compatible dimensions and data types.

- `numpy.argwhere`: returns the indices of duplicate values in an array.
- `numpy.bincount`: counts the number of occurrences of each value in an array.

605. What is the difference between slicing and indexing in NumPy, and how are they used for data cleaning?

Slicing is used to extract a portion of an array, whereas indexing is used to access a specific element or set of elements in an array.

Slicing is specified using the syntax [start:stop:step], where start is the starting index, stop is the ending index, and step is the step size between elements.

Indexing is specified using the syntax [row, column] or [index].

606. How can you concatenate two NumPy arrays, and what are some of the considerations when doing so?

In NumPy, you can concatenate two arrays along a specified axis using the concatenate() function. The syntax for concatenating two arrays is as follows:

```
np.concatenate((arr1, arr2), axis=0)
```

Here, arr1 and arr2 are the arrays that you want to concatenate, and axis is the axis along which you want to concatenate the arrays. If axis=0, the arrays will be concatenated vertically, while if axis=1, the arrays will be concatenated horizontally.

Some considerations when concatenating two NumPy arrays are:

The shapes of the arrays should match along the concatenation axis.

If the arrays have different data types, the resulting array will have the data type that can accommodate both types.

The concatenate() function creates a new array, so it can be memory-intensive if you are working with large arrays.

You can also use other functions like hstack() and vstack() to concatenate arrays along the horizontal and vertical axes, respectively. These functions are shortcuts for concatenating arrays along a specific axis. For example:

```
np.hstack((arr1, arr2)) # concatenates arrays horizontally
```

```
np.vstack((arr1, arr2)) # concatenates arrays vertically
```

These functions have the same considerations as concatenate().

607. How can you reshape a NumPy array, and what are some common use cases for doing so in data cleaning?

In NumPy, you can reshape an array using the `reshape()` method, which returns a new array with the specified dimensions. Reshaping an array can be useful in data cleaning and preprocessing when you need to change the shape of your data to fit a specific model or analysis.

For example, suppose you have a one-dimensional array with 12 elements representing the sales of a product over the course of a year:

```
import numpy as np
```

```
sales = np.array([10, 12, 15, 8, 20, 14, 17, 9, 11, 13, 16, 18])
```

You can reshape this array into a two-dimensional array with four rows and three columns representing the sales for each quarter of the year:

```
sales_quarterly = sales.reshape((4, 3))
```

This will result in an array with the following shape:

```
array([[10, 12, 15],  
       [ 8, 20, 14],  
       [17,  9, 11],  
       [13, 16, 18]])
```

Some common use cases for reshaping in data cleaning include converting data from one format to another, preparing data for visualization, and preparing data for analysis with machine learning models. For example, some machine learning models require that the input data be in a specific shape or format, and reshaping the data can be a necessary step in preparing it for analysis.

608. How do you create a NumPy array with specific dimensions and data types?

To create a NumPy array with specific dimensions and data types, you can use the `numpy.array()` function, which takes a list or tuple of values as input and returns a NumPy array. Here's an example:

```
import numpy as np
```

```
# Create a 2x3 array of integers
```

```
arr_int = np.array([[1, 2, 3], [4, 5, 6]], dtype=np.int32)

# Create a 2x3 array of floating-point values
arr_float = np.array([[1.0, 2.0, 3.0], [4.0, 5.0, 6.0]], dtype=np.float32)
```

In this example, we create two NumPy arrays with different data types (int32 and float32) and the same dimensions (2x3).

You can also use other NumPy functions to create arrays with specific dimensions and data types, such as:

`numpy.zeros(shape, dtype=None)`: Returns a new array of the given shape and data type, filled with zeros.

`numpy.ones(shape, dtype=None)`: Returns a new array of the given shape and data type, filled with ones.

`numpy.full(shape, fill_value, dtype=None)`: Returns a new array of the given shape and data type, filled with a specified value.

`numpy.random.rand(shape)`: Returns a new array of the given shape, filled with random values between 0 and 1.

These functions take a `shape` parameter that specifies the dimensions of the array, and an optional `dtype` parameter that specifies the data type of the array. For example:

```
# Create a 3x3 array of zeros
arr_zeros = np.zeros((3, 3), dtype=np.int32)

# Create a 2x2 array of ones
arr_ones = np.ones((2, 2), dtype=np.float32)

# Create a 4x4 array filled with the value 5.0
arr_fives = np.full((4, 4), 5.0, dtype=np.float32)

# Create a 5x5 array of random values
arr_random = np.random.rand(5, 5)
```

609. How do you access specific elements in a NumPy array?

To access specific elements in a NumPy array, you can use indexing. Indexing works by specifying the row and column of the element you want to access. For a one-dimensional array, you can use a single index to access an element. For example:

```
import numpy as np

a = np.array([1, 2, 3, 4, 5])
```

```
print(a[0]) # Output: 1  
print(a[2]) # Output: 3
```

For a two-dimensional array, you can use two indices to access an element. For example:

```
b = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])  
print(b[0, 0]) # Output: 1  
print(b[1, 2]) # Output: 6
```

You can also use slicing to access subsets of the array. Slicing works by specifying the start index, end index, and step size. For example:

```
print(a[1:4]) # Output: [2, 3, 4]  
print(b[:2, 1:]) # Output: [[2, 3], [5, 6]]
```

In this example, `a[1:4]` returns elements 1 through 3 of `a`, and `b[:2, 1:]` returns the first two rows and columns 1 through the end of `b`.

610. How can you perform basic arithmetic operations on NumPy arrays?

To perform basic arithmetic operations on NumPy arrays, you can simply use the arithmetic operators (+, -, *, /, //, %, **) between two or more arrays. These operations are performed element-wise, so the arrays must have the same shape.

54. How can you calculate basic descriptive statistics (e.g., mean, median, standard deviation) for a NumPy array?

To calculate basic descriptive statistics for a NumPy array, you can use various functions from the `numpy` module. For example:

`numpy.mean`: calculates the arithmetic mean of an array or a specific axis.
`numpy.median`: calculates the median of an array or a specific axis.
`numpy.std`: calculates the standard deviation of an array or a specific axis.
`numpy.var`: calculates the variance of an array or a specific axis.
`numpy.min`: finds the minimum value in an array or a specific axis.
`numpy.max`: finds the maximum value in an array or a specific axis.
`numpy.percentile`: calculates the specified percentile of an array or a specific axis.

611. How can you create a mask for a NumPy array based on specific conditions?

To create a mask for a NumPy array based on specific conditions, you can use Boolean indexing. For example, to create a mask for all elements in an array that are greater than 5:

```
mask = array > 5
```

This will create a Boolean array with the same shape as the original array, where the value at each element is True if the corresponding element in the original array is greater than 5, and False otherwise.

612. How can you apply a function to specific elements in a NumPy array?

To apply a function to specific elements in a NumPy array, you can use the `numpy.vectorize` function to create a vectorized version of the function. For example, to apply the `sin` function to each element in an array:

```
import numpy as np

array = np.array([0, np.pi/2, np.pi])

sin_array = np.vectorize(np.sin)(array)
```

This creates a new array where each element is the sine of the corresponding element in the original array.

613. How can you combine two or more NumPy arrays to create a new array?

To combine two or more NumPy arrays to create a new array, you can use functions such as `numpy.concatenate`, `numpy.vstack`, or `numpy.hstack`. For example, to concatenate two arrays along the first axis:

```
import numpy as np

array1 = np.array([[1, 2], [3, 4]])
array2 = np.array([[5, 6], [7, 8]])

combined_array = np.concatenate((array1, array2), axis=0)
```

This creates a new array where the rows of `array1` are stacked on top of the rows of `array2`.

614. How can you save a NumPy array to a file, and how can you load a saved array back into Python?

To save a NumPy array to a file, you can use the `numpy.save` or `numpy.savetxt` functions. For example, to save an array to a `.npy` file:

```
import numpy as np

array = np.array([1, 2, 3, 4, 5])

np.save('array.npy', array)
```

To load a saved array back into Python, you can use the `numpy.load` function:

```
loaded_array = np.load('array.npy')
```

615. How can you calculate dot products and matrix multiplication using NumPy?

NumPy is a powerful library for numerical computing in Python, and it provides convenient functions for calculating dot products and matrix multiplication.

To calculate the dot product between two vectors, you can use the `dot` function from NumPy. For example, to calculate the dot product between two 1-dimensional arrays `a` and `b`, you can use the following code:

```
import numpy as np

a = np.array([1, 2, 3])
b = np.array([4, 5, 6])
dot_product = np.dot(a, b)

print(dot_product)
```

This will output the result 32, which is the dot product of the two vectors.

To perform matrix multiplication, you can use the `dot` function as well. For example, to multiply two matrices `A` and `B`, you can use the following code:

```
import numpy as np

A = np.array([[1, 2], [3, 4], [5, 6]])
B = np.array([[7, 8], [9, 10]])
C = np.dot(A, B)

print(C)
```

Output

```
[[ 25  28]
 [ 57  64]
 [ 89 100]]
```

This is the result of multiplying the matrix A (which has dimensions 3x2) by the matrix B (which has dimensions 2x2), resulting in a matrix C with dimensions 3x2.

616. How can you apply linear algebra operations (e.g., inverse, determinant) to a NumPy array?

NumPy provides several functions for performing linear algebra operations on arrays, including finding the inverse and determinant of a matrix.

To find the inverse of a matrix, you can use the `inv` function from NumPy's `linalg` module. For example, to find the inverse of a 2x2 matrix A, you can use the following code:

```
import numpy as np

A = np.array([[1, 2], [3, 4]])
A_inv = np.linalg.inv(A)

print(A_inv)
```

Output

```
[[ -2.  1. ]
 [ 1.5 -0.5]]
```

To find the determinant of a matrix, you can use the `det` function from NumPy's `linalg` module. For example, to find the determinant of a 3x3 matrix B, you can use the following code:

```
import numpy as np

B = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
det_B = np.linalg.det(B)

print(det_B)
```

Output
0.0

In this case, the determinant of matrix B is zero, which means that the matrix is singular and does not have an inverse.

Numpy Tricky Interview Questions

617. What happens if you try to reshape a NumPy array with the wrong number of elements?

If you try to reshape a NumPy array with the wrong number of elements, you will receive a ValueError. For example, if you try to reshape an array with 10 elements into a shape of (4, 4), which requires 16 elements, you will receive an error.

49. How can you avoid modifying a NumPy array when performing a mathematical operation on it?

To avoid modifying a NumPy array when performing a mathematical operation on it, you can use the copy() method to create a copy of the original array before performing the operation. This ensures that the original array is not modified. For example, if you have an array a and want to perform an operation on it without modifying it, you can use b = a.copy() to create a copy of a and then perform the operation on b.

618. What is the difference between NumPy's broadcasting rules and Python's broadcasting rules?

The main difference between NumPy's broadcasting rules and Python's broadcasting rules is that NumPy allows for broadcasting between arrays of different shapes, while Python requires that the two objects being broadcast have the same shape. NumPy's broadcasting rules allow for arrays with different shapes to be used in arithmetic operations, as long as they meet certain criteria.

619. How can you modify the data type of a NumPy array?

To modify the data type of a NumPy array, you can use the astype() method to create a new array with the desired data type. For example, if you have an array a of integers and want to convert it to an array of floating point numbers, you can use b = a.astype(float).

620. How can you perform element-wise comparison between two NumPy arrays with different shapes?

To perform element-wise comparison between two NumPy arrays with different shapes, NumPy's broadcasting rules can be used to align the shapes of the arrays. For example, if you have arrays a and b of different shapes and want to compare them element-wise, you can use c = a == b. NumPy will automatically align the shapes of a and b using its broadcasting rules, and return an array of the same shape as the largest input array, with a boolean value at each

position indicating whether the corresponding elements of a and b are equal.

621. What is the difference between NumPy's views and copies, and how can you determine which one you have?

NumPy's views and copies are two different ways of accessing the data in a NumPy array. A view refers to a different way of looking at the same data, while a copy refers to a completely new set of data. Views are created using slicing, while copies are created using the `copy()` method. To determine whether you have a view or a copy of a NumPy array, you can use the `base` attribute, which returns `None` for a copy and the original array for a view.

622. How can you create a custom data type in NumPy, and what are some use cases for doing so?

To create a custom data type in NumPy, you can use the `dtype()` function and pass in a string representing the desired data type. For example, to create a data type that represents complex numbers with 64-bit floating point components, you can use `dtype('complex128')`. Custom data types are useful for representing complex data structures or specialized data formats.

623. What are some best practices for optimizing performance when working with large NumPy arrays?

Some best practices for optimizing performance when working with large NumPy arrays include using views instead of copies when possible, avoiding loops over arrays and instead using built-in functions, and using NumPy's built-in functions for operations that are common in scientific computing.

624. How can you handle missing or invalid data in a NumPy array?

To handle missing or invalid data in a NumPy array, you can use the `numpy.ma` module to create a masked array. A masked array is an array that contains additional information about which elements are valid and which are invalid or missing. The `numpy.ma` module provides functions for creating masked arrays and performing operations on them. For example, to create a masked array from an existing array `a` with missing values represented by `NaN`, you can use `m = np.ma.masked_invalid(a)`. This will create a masked array `m` with `NaN` values masked out.

625. What are some ways to create a NumPy array, and when would you use each one?

There are several ways to create a NumPy array, and the choice depends on the type and shape of the data you are working with. Some common ways to create NumPy arrays are:

Using the `array` function to create an array from a Python list or tuple.

Using the `zeros` function to create an array of zeros with a specified shape.

Using the `ones` function to create an array of ones with a specified shape.

Using the `empty` function to create an array of uninitialized values with a specified shape.

Using the `arange` function to create an array of values with a specified range and spacing.

Using the linspace function to create an array of values with a specified range and number of elements.

Using the random module to create an array of random values with a specified distribution and shape.

626. How can you access and modify individual elements of a NumPy array?

To access and modify individual elements of a NumPy array, you can use indexing. Indexing works in a similar way to Python lists, with the first element of the array having an index of 0. For example, to access the element in the first row and second column of a 2-dimensional array A, you can use the following code:

```
import numpy as np
```

```
A = np.array([[1, 2], [3, 4]])
print(A[0, 1]) # prints 2
```

```
A[0, 1] = 5
print(A) # prints [[1, 5], [3, 4]]
```

627. What is the difference between slicing and indexing in NumPy, and how can you use them to extract subsets of an array?

Slicing and indexing are two ways to extract subsets of a NumPy array. Indexing refers to accessing individual elements of an array using their position, while slicing refers to extracting a subset of an array along one or more dimensions. To slice a NumPy array, you can use the : operator to specify a range of indices. For example, to extract the first two rows of a 2-dimensional array A, you can use the following code:

```
import numpy as np
```

```
A = np.array([[1, 2], [3, 4], [5, 6]])
B = A[:2, :]

print(B) # prints [[1, 2], [3, 4]]
```

628. How can you perform mathematical operations on a NumPy array, and what are some common functions for doing so?

You can perform mathematical operations on a NumPy array using element-wise operations or matrix operations. Element-wise operations apply an operation to each element of an array,

while matrix operations apply an operation to the entire array or a subset of the array. Some common functions for performing mathematical operations on NumPy arrays are:

add: add two arrays element-wise.

subtract: subtract one array from another element-wise.

multiply: multiply two arrays element-wise.

divide: divide one array by another element-wise.

dot: perform matrix multiplication between two arrays.

629. What is broadcasting in NumPy, and how can you use it to perform element-wise operations on arrays with different shapes?

Broadcasting is a feature in NumPy that allows you to perform element-wise operations on arrays with different shapes. When performing an element-wise operation between two arrays with different shapes, NumPy automatically broadcasts the smaller array to match the shape of the larger array. For example, to add a scalar value to each element of a 1-dimensional array A, you can use the following code:

```
import numpy as np

A = np.array([1, 2, 3])
B = A + 1

print(B) # prints [2, 3, 4]
```

630. How can you reshape a NumPy array, and what are some common use cases for doing so?

To reshape a NumPy array, you can use the reshape function. The reshape function allows you to change the shape of an array without changing the underlying data. Some common use cases for reshaping NumPy arrays are:

Flattening a 2-dimensional array into a 1-dimensional array using reshape(-1).

Reshaping an array to match the shape of another array using reshape_like.

Changing the number of dimensions of an array using reshape.

631. What is a masked array in NumPy, and how can you use it to handle missing data?

A masked array in NumPy is an array that contains a mask that specifies which elements of the array are valid and which are invalid. The mask is an array of the same shape as the data array, where each element is either True if the corresponding element in the data array is invalid, or False if the corresponding element is valid. Masked arrays are useful for handling missing data,

where some elements of the array may be invalid or undefined. You can create a masked array in NumPy using the `ma` module, and you can apply mathematical operations to the array while taking the mask into account. For example, to calculate the sum of a masked array `A` while ignoring the masked elements, you can use the following code:

```
import numpy as np
import numpy.ma as ma

A = np.array([1, 2, 3, -999])
mask = (A == -999)
A = ma.masked_array(A, mask=mask)

print(A.sum()) # prints 6
```

632. How can you stack and concatenate NumPy arrays, and what are some use cases for doing so?

To stack and concatenate NumPy arrays, you can use the `stack` and `concatenate` functions. The `stack` function allows you to stack arrays along a new axis, while the `concatenate` function allows you to concatenate arrays along an existing axis. Stacking and concatenating NumPy arrays can be useful for combining data from multiple sources, or for creating multi-dimensional arrays. Some common use cases for stacking and concatenating NumPy arrays are:

Stacking a sequence of 1-dimensional arrays into a 2-dimensional array using `np.stack`.
Concatenating multiple 2-dimensional arrays along the rows or columns using `np.concatenate`.
Combining a sequence of multi-dimensional arrays into a single multi-dimensional array using `np.stack`.

633. What are some best practices for optimizing performance when working with large NumPy arrays?

When working with large NumPy arrays, there are several best practices for optimizing performance:

- Use vectorized operations to perform mathematical operations on arrays, rather than iterating over the elements of the arrays.
- Avoid creating unnecessary copies of arrays, as this can be a performance bottleneck.
- Use views of arrays instead of copying data whenever possible, as this can reduce memory usage.
- Use the `dtype`

Numpy Advance Interview Questions

634. What are some of the performance benefits of using NumPy over pure Python when working with numerical data?

NumPy is much faster than pure Python when working with numerical data because it is optimized for numerical operations, particularly operations on arrays of data. NumPy performs calculations using precompiled C code, which is much faster than Python code. NumPy also allows for vectorized operations, which can be much faster than using loops in Python.

635. What is a view in NumPy, and how does it differ from a copy?

In NumPy, a view is a reference to the original data, whereas a copy creates a new data object. Views and copies differ in how they handle changes made to the data. When you modify a view, you modify the original data as well. When you modify a copy, you create a new data object that is not related to the original data. Views are more efficient because they do not require copying the data, but they can be more error-prone because changes made to the view can affect the original data.

636. How can you use NumPy to perform linear algebra operations, such as matrix multiplication and solving systems of equations?

NumPy provides many functions for performing linear algebra operations, such as matrix multiplication and solving systems of equations. These functions are part of the linalg module in NumPy. Some of the commonly used functions include dot, inv, and solve.

637. What are some of the built-in functions in NumPy for generating random numbers, and how can you use them to simulate data?

NumPy provides many functions for generating random numbers, such as random, randint, and normal. These functions can be used to simulate data for a wide variety of applications, such as Monte Carlo simulations and statistical analysis.

638. What is the difference between a structured array and a record array in NumPy, and what are some use cases for each?

Structured arrays and record arrays are both ways of storing data in NumPy arrays. Structured arrays are arrays where each element can have a different data type, while record arrays are arrays where each element is a record with named fields. Structured arrays are useful for

working with data that has different data types, while record arrays are useful for working with structured data, such as data in a database.

639. How can you use NumPy to perform Fourier transforms, and what are some applications of Fourier analysis in signal processing and image processing?

NumPy provides a fft module that can be used to perform Fourier transforms. Fourier analysis is used in signal processing and image processing to analyze and manipulate signals and images in the frequency domain. Some applications of Fourier analysis include noise reduction, compression, and filtering.

640. How can you use NumPy to perform interpolation, and what are some use cases for doing so?

641. What are some of the limitations of NumPy, and how can you work around them?

NumPy provides functions for performing interpolation, such as interp1d and interp2d. Interpolation is useful for estimating values of a function at points where no data is available or for smoothing noisy data. Some applications of interpolation include data analysis and image processing.

642. What are some best practices for organizing and structuring code when working with NumPy, especially when dealing with large, complex arrays?

Some limitations of NumPy include its inability to handle data that is larger than available memory, its lack of support for distributed computing, and its inability to handle missing data. These limitations can be worked around by using other libraries or techniques, such as distributed computing frameworks, database systems, and statistical methods for handling missing data.

643. Write a NumPy code snippet to create an array of 100 random integers between 0 and 10.

```
import numpy as np  
  
random_integers = np.random.randint(0, 11, size=100)  
print(random_integers)
```

644. Write a NumPy code snippet to calculate the element-wise sum of two arrays of the same shape.

```
import numpy as np

arr1 = np.array([1, 2, 3, 4])
arr2 = np.array([5, 6, 7, 8])
sum_arr = arr1 + arr2
print(sum_arr)
```

645. Write a NumPy code snippet to compute the inner product of two one-dimensional arrays

```
import numpy as np

arr1 = np.array([1, 2, 3])
arr2 = np.array([4, 5, 6])
inner_prod = np.inner(arr1, arr2)
print(inner_prod)
```

646. Write a NumPy code snippet to find the indices of the maximum and minimum values in a two-dimensional array.

```
import numpy as np

arr = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
max_index = np.unravel_index(np.argmax(arr), arr.shape)
min_index = np.unravel_index(np.argmin(arr), arr.shape)
print(max_index, min_index)
```

647. Write a NumPy code snippet to reshape a one-dimensional array into a two-dimensional array with 4 rows and 5 columns.

```
import numpy as np

arr1d = np.arange(20)
arr2d = arr1d.reshape(4, 5)
print(arr2d)
```

648. Write a NumPy code snippet to calculate the correlation coefficient between two arrays of the same length.

```
import numpy as np

arr1 = np.array([1, 2, 3, 4, 5])
arr2 = np.array([2, 3, 4, 5, 6])
corr_coef = np.corrcoef(arr1, arr2)
print(corr_coef)
```

649. Write a NumPy code snippet to create a diagonal matrix with the elements 1, 2, and 3 on the diagonal.

```
import numpy as np

diagonal_arr = np.diag([1, 2, 3])
print(diagonal_arr)
```

650. Write a NumPy code snippet to sort a one-dimensional array in ascending order.

```
import numpy as np

arr = np.array([5, 2, 8, 1, 9])
sorted_arr = np.sort(arr)
print(sorted_arr)
```

651. Write a NumPy code snippet to calculate the element-wise product of two arrays of the same shape.

```
import numpy as np

arr1 = np.array([1, 2, 3, 4])
arr2 = np.array([5, 6, 7, 8])
prod_arr = arr1 * arr2
print(prod_arr)
```

652. Write a NumPy code snippet to create a mask that selects all elements of an array that are greater than 5.

```
import numpy as np
```

```
arr = np.array([3, 8, 2, 10, 4, 6])
mask = arr > 5
print(arr[mask])
```

Chapter 5 - Case Study and Guesstimate

We will divide this section into two parts and will further divide it into domains. So, we will target 60 questions from the case study and 40 from guesstimates.

Case study segregation:-

- E-commerce
- Digital Analytics
- Data Analysis

What is the importance of a case study round in an analytics interview?

It is true that an analytics interview will mostly revolve around your technical abilities, you will be asked questions on SQL, Python, and some new technologies. Then why work on your case study solving abilities? There are two answers to this question:-

1. A good case study round will prove that you have a good ability to think through any problem i.e. you can take up any scenario and can break it up into parts to come up with a good solution
2. Suppose you did not do well in your technical round then this case study round can increase your chances of getting through the interview

Let's start with our 13 point approach:-

1. Understand the problem

You need to clearly understand the problem. It is necessary because you will definitely be given the case study which you have not done in the past. Suppose, you are working in an e-commerce domain then you could be asked something on the lines of telecommunication industry or pharmaceutical or search engine. Now, your ability to understand the question will form the foundation of the case study round.

Bonus point – Once the question is asked, don't repeat the question as it is. You can summarise the question but don't kill time in repeating the question

2. Jot down the important points

When you are taking a telephonic round then you need to take up all the important points. These written points will help you whenever you are stuck in the round. At least try to write down the numbers and data given in the round. Try to comprehend things in your notes

3. Repeat your understanding

Now, once you have written the points and understood the problem then now is the time to summarise the problem. We can not stress more on the fact that you need to make sure that you have understood the problem and that the interviewer is aligned with your understanding.
SUMMARISE

4. Clear out the final solution

Be very crisp but do clear out the final solution which you need to provide to the interviewer, if possible write it down in bold in your notes. Keep referring to the final outcome at least twice in your interview

5. Ask Questions

Never will the interviewer provide you a complete case study, they will always keep a room for at least a few questions. Keep asking thoughtful questions and not just for the sake of asking questions

6. Mention assumptions

You will have to ask questions to come up with assumptions, solid assumptions. Make sure that mention your assumptions boldly. For example, in one of the case studies a person was asked a case study to formulate the strategy to expand KFC in a new Japan. Now the first assumption the candidate mentioned was that this is the same multinational KFC brand that is famous for its fried chicken. Why is this assumption important? Because has it not been the same KFC but a local brand then the whole case study will change its flow. This new KFC might have budget concerns and can not burn cash to acquire the market. So, it's very important to make and validate assumptions

7. Keep a check on time

Every interview is time-bound, be sure that you respect the allotted time and increase or decrease the pace accordingly. A standard case study round in an analytics interview can range from 15 minutes to 30 minutes in a product-based company but in a consulting or investment firm, it can go up to 50 to 60 minutes.

8. Organise your answer

You cannot solve a case study if your thoughts are not organized. Make sure to have clear horizontal and vertical cuts in your answer. If you are solving a problem to design a dashboard for a firm then you need to have main cuts like Supply, Inventory, Revenue, Operations and then go deep into each point.

9. Talk with numbers and data

You need to be quick with your calculation. Make sure to have a simple number for your calculations. Even if you are dead sure that the population of India is 133 crore, please assume and mention that the population is 100 cr.

Always ask for numbers and data associated with the case study, what is the market cap of the company in question, what is the user base, what were the numbers last year, anything and everything under the hood which is relevant to the case study. Avoid asking the capital of South Africa 😊

10. Listen to the interviewer

There is only one person in the room who has a better idea of the solution to the problem and believe me, it's not you. You need to get the answer out of the interviewer, have you ever played 20 questions? There you think of a word and the person in front of you have to ask you 20

objective questions to boil down to the answer or that word. Play that game but don't get too excited !!

And don't straight away start celebrating like Saurav Ganguly in the Natwest series 😊

11. Come up with corner cases

Now you have to always think of the odd cases, cases that might take the interviewer off guard and get you brownie points. Things like 'is there a new release in your app which is not working for a particular less branded operating system?'

12. Have room for ideas from the interviewer

Now, sometimes you might get into an argument or get too personal with your point, do make sure to listen to the interviewer and accept it if you think he is making a good point, he is your interviewer, not your boyfriend/girlfriend that you need not accept even if you are slightly off the topic

STAR methodology to tackle any case study on personal

S – What was the Situation ? ML model with a time crunch

T – What was the Task at hand? Had to work on with limited resource to predict XYZ with ABC accuracy

A – What Action did you take? After I missed out on this opportunity, we did rigorous EDA and identified many good variables

R – What was the Result ? Very stable and robust model with better accuracy

653.What are the metrics to evaluate a website? A search engine?

Metrics to evaluate a website

1. Number of visitors
2. Bounce Rate
3. Average time on page
4. Click-through rate
5. Conversion rate

Metrics to evaluate a search engine

1. Recall: Measuring the ability of a search engine to find the relevant material in the index
2. Precision: measuring its ability to place that relevant material high in the ranking.

2. How would you measure the success of private stories on Instagram, where only certain close friends can see the story?

Start by answering: What is the goal of the private story feature on Instagram? You can't evaluate "success" without knowing what the initial objective of the product was, to begin with.

One specific goal of this feature would be to drive engagement. A private story could potentially increase interactions between users, and grow awareness of the feature.

Now, what types of metrics might you propose to assess user engagement? For a high-level overview, we could look at:

Average stories per user per day

Average Close Friends stories per user per day

However, we would also want to further bucket our users to see the effect that Close Friends stories have on user engagement. By bucketing users by age, date joined, or another metric, we could see how engagement is affected within certain populations, giving us insight on success that could be lost if looking at the overall population.

654. Punjabi By Nature is going through a rough phase and is wasting 30-40 Kgs of biryani on some days of the week.

Management is thinking to find a method to sell the left over biryani or reduce the amount of biryani prepared.

You are the manager and you are supposed to come up with a methodology for the same

Ans.

Assumptions:-

- PBN is a restaurant and not a home kitchen.
- They cook biryani only once a day.

Here is kind of a mindmap I created for this problems with my idea's and suggestions

- Reduce the amount cooked/ Sell leftover
- Selling Leftovers
- Leveraging food delivery apps
- Exclusive Partnership with food delivery apps with good offers during closing hours
- Free testers with other orders of delivery app
- Leveraging footfall in restaurant
- x% off for customers on donating food for the needy
- Exclusive buy 1 get 1 on specific days of weeks where wastage is more historically
- Marketing
- selling them nearby competitor stores with publicity pamphlets
- Estimating the cooking amount

- Based on data
- Reduced quality/ quantity of food/ Prices
- Quality
- Recent change in the cook for biryani
- Customer reviews on popular food critic website
- Quantity
- Change in portion size
- Prices
- Change in prices from history
- Competitors nearby
- New competitors nearby or on delivery apps
- Prices and offers compared to competitors
- Change in customers eating habits
- Footfall and ordering trend for all the competitors

Historic data for the following columns can be asked for further analysis – (Order history, Cook information, Customer feedback from various websites, portion sizes per plate, price per plate, Competitors order history)

Leveraging above columns, a suitable model can be trained giving a probabilistic estimate for the amount of biryani to be cooked everyday.

- Non data based
- X% off on reserving table or an order a day before
- Cooking biryani twice a day instead of once allows more control on quantity cooked

655. TVF recently went to app/website only mode. How do you think it makes money to survive and pay the actors ?

TVF was started in 2010 with an aim to entertain people by making videos revolving around daily life experiences of the normal people in a sarcastic way.

Talking about the business model of TVF, they make money from YouTube by making videos and by displaying ads during the entire video. Today TVF is having an approx. 2.5-3 Cr views/Month which actually contributes into 15-20 lac rupees. Now as per the videos published by TVF , the cost of making a video would also be some what high. We can see every video from TVF is sponsored by a big name or a start-ups for e.g. Permanent Roommates season 1 was sponsored by commonfloor, then season 2 was sponsored by OLA cabs, The famous show TVF pitchers was sponsored by Kingfisher. These companies pay money to TVF because they know their product will be reaching out to millions of people

Secondly , the people involved in TVF are also great stand-up comedians they make money by doing stand-up comedy shows for corporate organization and by doing shows at Institutions , they are actually paid a good amount of money for doing a stand-up comedy at various places.

By introducing the mobile app of TVF , they are trying to attract people on their own platform. Let suppose we consider money made from YouTube , some portion of the money is taken by YouTube as a share. By having an app of TVF they will convert that share into profit for TVF. Then after when they are sure that some people are regularly engaging in their mobile app they will start charging the user to view the content just like other platform.

656. How to increase the marketing of online games?

First, let us understand the current state of marketing and brand awareness of our online game. We would have data on our customer base (no. of active users, new users added every month and users leaving every month). We would also have data on our marketing campaigns (media vehicle – YT,Social Media, Radio, TV etc.) – clicks, impressions, cost, conversions and revenue made through these channels.

To design our marketing, we would first identify the various segments we cater to (from our customer base). Once done with that, we find out the segment growing the most, most sizeable, most revenue-generating, with least drop-off. If our aim is to increase our customer base, then we focus on growing segment; if our aim is to reduce attrition, we focus on the most sizeable segment which has the highest drop-off. We come up with one target persona for our online games and position ourselves accordingly.

Some ways of increasing our reach-

1. Posting ads and articles on gaming blogs and websites.
2. Partnerships with popular streamers/gamers on YT and FB.
3. Sponsor gaming tournaments hosted on our app (will increase user base, even if momentarily).
4. Improving SEO ranking and work on SEM (Will work if search engine is a major media).
5. Click-through ads and posts on social media (FB and Instagram) – targeting those who follow related pages and gamers and their gaming streams.

657. Given a set of webpages and changes on the website, how will you test the new website feature to determine if the change works positively?

Generally the traffic coming to your website will be divided into 2 different groups, control group(who will be shown the original version) and variation group(who will be shown the modified version). Now, depending upon how the users react to these 2 different scenarios , some metrics can be determined to measure the effect of the change.

658. What is root cause analysis? How to identify a cause vs. a correlation? Give examples.

Root cause analysis is about going to the depths of the problem and identifying

the real issue which is creating the problem in the first case.

Consulting firms like Mckinsey use methodologies like Issue trees, in which they start from a high-level problem at hand and start breaking those problems into smaller issues and understanding the cause behind all those issues.

For Example, a company wants to investigate about the declining profits from a particular product, they can start from breaking this problem in 2 parts, i.e increasing costs

and decreasing sales, then further scrutinizing the reasons for the occurrence of the both and continuing the process until they get to the root cause of the problem.

Correlation does not imply causation in every case.

Consider a example, the onset of summer causes an increase in the sale of ice-creams and sunglasses. If you consider the data points, you will find that the sale of both the quantities is highly correlated, but increase in the sale of sunglasses is not the cause of increase in the sale of ice-creams or vice-versa. The cause for both the cases is the onset of summer.

659. Can you choose a strategy for increasing the number of songs listened by the user on an online application? How will you decide the types of playlists to suggest him?

Mainly I will tackle this problem like any recommendation engine.

Machine learning algorithms in recommender systems are typically classified under two main categories—

1. content-based – strategy relies on analyzing factors and demographics that are directly associated with the user or product, such as the age, sex and demographic of the user or a song genre or playlist,
2. collaborative filtering – Collaborative Filtering takes consumer behavior data and utilizes it to predict future behavior. This consumer behavior leaves a trail of data, generated through implicit and explicit feedback, based on the user's listening history, in tandem with songs enjoyed by users who seem to have a similar history
9. If you are having 4GB RAM in your machine and you want to train your model on a 10GB data set. How would you go about this problem?
Batch processing can accomplish this.

There might also be situations when a lot of data points add little or no value to the model. In such cases, we can sample a smaller proportion of the data to run the model on, without compromising on performance

660. How would you measure the success of acquiring new users through a 30-day free trial at Netflix?

More context: Netflix is offering a promotion where users can enroll in a 30-day free trial. After 30 days, customers will automatically be charged based on their selected package. How would you measure acquisition success, and what metrics would you propose to measure the success of the free trial?

One way we can frame the concept specifically to this problem is to think about controllable inputs, external drivers, and then the observable output. Start with the major goals of Netflix:

Acquiring new users to their subscription plan.

Decreasing churn and increasing retention.

Looking at acquisition output metrics specifically, there are several top-level stats that we can look at, including:

Conversion rate percentage

Cost per free trial acquisition

Daily conversion rate

With these conversion metrics, we would also want to bucket users by cohort. This would help us see the percentage of free users who were acquired, as well as retention by cohort.

661. How would you measure the success of Facebook Groups?

Start by considering the key function of Facebook Groups. You could say that Groups are a way for users to connect with other users through a shared interest or real-life relationship.

Therefore, the user's goal is to experience a sense of community, which will also drive our business goal of increasing user engagement.

What general engagement metrics can we associate with this value? An objective metric like Groups monthly active users would help us see if Facebook Groups user base is increasing or decreasing. Plus, we could monitor metrics like posting, commenting, and sharing rates.

There are other products that Groups impact, however, specifically the Newsfeed. We need to consider Newsfeed quality and examine if updates from Groups clog up the content pipeline and if users prioritize those updates over other Newsfeed items. This evaluation will give us a better sense of if Groups actually contribute to higher engagement levels.

662. Describe how you would build a model to predict Uber ETAs after a rider requests a ride.

Common machine learning case study problems like this are designed to explain how you would build a model. Many times this can be scoped down to specific parts of the model building process. Examining the example above, we could break it up into:

How would you evaluate the predictions of an Uber ETA model?

OR

What features would you use to predict the Uber ETA for ride requests?

Our recommended framework breaks down a modeling and machine learning case study to individual steps in order to tackle each one thoroughly. In each full modeling case study, you will want to go over:

- Data processing
- Feature Selection
- Model Selection
- Cross Validation
- Evaluation Metrics
- Testing and Roll Out

663. You're a Data Scientist / Business Analyst working for a new eCommerce company called A&B Co. (similar to Amazon) and you've been asked to prepare a presentation for the Vice President of Sales and the Vice President of Operations that summarizes sales and operations thus far. The summary should include (at a minimum) a summary of current state the business, current customer satisfaction, and a proposal of 2-3 areas where the company can improve. Here are some facts:

**It's currently September 2018 (e.g., you can ignore all data after September 2018)
The company's inception was January 2017 (so you can ignore all data before January 2017)**

Company is US-based, but launched in Brazil (which is why some information is in Portuguese)

You can assume all orders are delivered (so ignore the order state field)

Your presentation should not have more than 10 slides of content, and the presentation itself should only take ~15 minutes.

Ans. Build a framework to answer the questions. If you're not sure what the questions are, create questions for yourself to answer. It makes the process of digging for data so much easier. It's hard to come up with an answer if you don't know what the questions are. This point seems like a no-brainer, but it's good to make sure that you've created a structure to answer the key points. If you're not sure what the question at hand is, you might need to play with the data a bit to understand what seems to be the problem at hand. For this particular case the ask is really clear, we need to create a summary which includes current state the business, current customer satisfaction, and a proposal for an area where the company can improve. The questions we came up with (and some answers listed below) given the ask are:

What should be shown to reflect current state of business? (Probably something to do with product growth and amount of money)

Revenue → how much money are we making?

Volume of sales → how many orders are we getting?

Customer summary based on spend + behavior

How should we present customer satisfaction?

Is customer satisfaction a problem? Do initial analysis to figure it out?

Does customer satisfaction reflect our current business state? (e.g. if business is down, does customer satisfaction also go down?, if business is boomin' is customer satisfaction higher?)

Should I focus on sales area of improvement or customer satisfaction area of improvement?

This depends on what bullet points 1 + 2 yield, also depends on customer churn

Questions don't always have analytical answers, sometimes questions lead to more questions. This is of course a high-level framework, but the above should give you an idea of how we went through solving this! (the details of course are provided in the code/presentation material)

664. Profit of a company selling mobile back cover is declining. List out all the possible reasons

Following is the way in which discussion proceeded with the interviewer:-

1. The demand itself has declined i.e. customers are not using cover that much. Asked to think more by the interviewer
2. Maybe the competitor is also facing loss which again means that the demand is low. Competitors are making a decent profit
3. Bad Marketing – The company is not putting stalls or shops in a crowded place. The interviewer told that the company was making a decent profit 6 months back
4. Maybe the footfall of the mall or place decreased. Could be(first positive response)
5. Maybe a popular mobile phone shop has shifted somewhere else. Could be(again a so-so response)
6. Maybe the other companies have reduced the price of their product which is why customers are drifting to these companies. The interviewer seemed pleased
7. New technology in the cover market to make covers more durable and the company we are talking about is using the same old technology. Seemed good enough point
8. Since we are talking about back covers, there could be new or trending designs which are not produced by the company
9. The company has not registered on different e-commerce websites and the website they are present on is not doing good business. He looked satisfied with the point

665. How would you construct a feed to show relevant content for a site that involves user interactions with items?

There are seven pillars of User interaction

1. Valuable
2. Useful
3. Usable
4. Findable
5. Credible
6. Desirable
7. Accessible

We can do so using building a recommendation engine. The easiest we can do is to show contents that are popular other users, which is still a valid strategy if for example the contents are news articles. To be more accurate, we can build a content based filtering or collaborative filtering.

If there's enough user usage data, we can try collaborative filtering and recommend contents other similar users have consumed. If there isn't, we can recommend similar items based on vectorization of items (content based filtering).

666. How would you design the people you may know feature on LinkedIn or Facebook?

1. Find strong unconnected people in weighted connection graph
2. Define similarity as how strong the two people are connected
3. Given a certain feature, we can calculate the similarity based on
-friend connections (neighbors)
- Check-in's people being at the same location all the time. -same college, workplace
4. Have randomly dropped graphs test the performance of the algorithm
5. News Feed Optimization
6. Affinity score: how close the content creator and the users are
7. Weight: weight for the edge type (comment, like, tag, etc.).
8. Emphasis on features the company wants to promote
9. Time decay: the older the less important

667. How would you predict who someone may want to send a Snapchat or Gmail to?

For each user, assign a score of how likely someone would send an email to the rest is feature engineering:

- Number of past emails
- How many responses
- The last time they exchanged an email
- Whether the last email ends with a question mark
- Features about the other users, etc.
- People who someone sent emails the most in the past conditioning on time decay.

668. How would you suggest to a franchise where to open a new store?

- Build a master dataset with local demographic information available for each location.
- Local income levels
- Proximity to traffic, weather, population density, proximity to other businesses
- A reference dataset on local, regional, and national macroeconomic conditions (e.g. unemployment, inflation, prime interest rate, etc.)
- Data on the local franchise owner-operators, to the degree the manager identify a set of KPIs acceptable to the management that had requested the analysis concerning the most desirable factors surrounding a franchise -Quarterly operating profit, ROI, EVA, pay-down rate, etc.
- Run econometric models to understand the relative significance of each variable
- Run machine learning algorithms to predict the performance of each location candidate

669. In a search engine, given partial data on what the user has typed, how would you predict the user's eventual search query?

- Based on the past frequencies of words shown up given a sequence of words, we can construct conditional probabilities of the set of next sequences of words that can show up (n-gram).
- The sequences with highest conditional probabilities can show up as top candidates.
- To further improve this algorithm, we can put more weight on past sequences which showed up more recently and near your location to account for trends show your recent searches given partial data

670. Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

- frequency of donations
 - Amount of donation
 - Graduation year
 - Major subject
 - Business domain
 - Network strength of the alumuni
- Construct a supervised regression (or binary classification) algorithm.

671. You're Uber and you want to design a heat map to recommend to drivers where to wait for a passenger. How would you approach this?

- Based on the past pickup location of passengers around the same time of the day, day of the week (month, year), construct
- Based on the number of past pickups
- Account for periodicity (seasonal, monthly, weekly, daily, hourly)
- Special events (concerts, festivals, etc.) from tweets
- Public holidays

672. How will you change the UI of the game to increase the number of people buying coins from the shop section (Clash Of Clans)?

Ans. This was a business case study, we had the discussion on the following points:-

1. Project the offer directly on the home page instead of clicking on the shop button
2. Put the price to buy elixir and gold near the collector(the pink and yellow images)
3. After every attack resulting in a loss, give an option to buy back the points
4. Get the country-wise data and see the most engaging time of the player and project the offers accordingly
5. Give a variable discount to the players on the basis of their day-to-day performance to allure them in buying coins on the day they performed well

673. Taj Group of Hotels is planning to start a new branch, What are the parameters it should consider to find the appropriate place?

The following points were discussed:-

- a. Find out the place where people have mostly searched for 5 or 7 star hotels
- b. Find the place where the average annual income is high, may be Bangalore, Pune, Delhi, Hyderanad, etc.
- c. Look for that place which is known for tourism as it will attract foreign customers
- d. Look for that area which has good facilities around like popular restaurants, pubs, malls, etc.
- e. Look for that city where there are all the necessary facilities like airport near the city, railway station, etc.
- f. Look for that city where you can get good service from third party vendors for basic services like laundry, service employees, security service, etc.

674. List the e-commerce metrics

Visitor – If you are reading this article then you are a visitor to our website

Unique Visitor – If you visit our website 30 times today then also you are only one unique visitor

Visits – Now you have made 30 visits to the website

Repeat visitor – Suppose The Data Monk defines the logic that a visit span is 30 days. If you come two times in the last 30 days then you are a repeat visitor

Return visitor – Suppose you came last time in Oct'21 and it's currently Jan'22 then you are a return visitor

Views – number of times you visit a particular page is the number of views. Even if you refresh the webpage then also the number of views will increase

Session – If the admin of the website had defined a session to be of 30 minutes then accordingly a session id will be generated. Ex. you start at 11:00 in the morning and surf for 10 minutes, then come back at 4 in the evening, then a new session id will be generated for you

Bounce – If you close the website on the first page itself then it is termed as a bounce, though you will still be a visitor and your number of views will be 1

Click-Through Rate – The click-through rate refers to the percentage of people who click through a certain link compared to the total number of people who saw the link. For example, if 1,000 people see a Google ad and 10 of them click on that ad, that ad has a click-through rate of 1 percent. You should aim for a click-through rate on your Google Ads of about 3-5 percent

Impression – Number of times your advertisement or web link is shown to the customer on the online space. For example, when you google something then you get some sponsored links, you may get around 4-5 images at the top. It does not matter if you click on these images or not, but there are impressions of these products. Suppose this same page is shown to 1000 people and only 10 people clicked on the one which is priced at Rs. 21900, then the CTR for the product is 1% with 1000 impressions and 10 clicks

Load Time – Load time refers to the amount of time it takes a page to appear when a user lands on it. About 30 percent of users will leave a website whose pages take longer than 3 seconds to load, so that number should be your benchmark. Optimize slower loading pages to maximize their speed.

Time on Page – Time spent by users on a particular page.

Session Time – What is the average time for which a user explores your website or app

Conversion – If you are selling a product then what is the conversion i.e. number of people entering the buy flow and the number of people actually buying something on the web portal.

675. The case study was to increase the number of conversions of freemium customers to premium customers for a telecom company in India.

To solve or approach a case study, you need to have an initial state in mind and convey this to the interviewer.

Assumptions – I assumed that the company sells iPhone and is also providing free calling plans to the customers for the first 6 months. But the customers are not continuing the plan once the free period is over and this thing is alarming for the company

So, I gave a shape to the complete case study problem, from here I have one and only one specific problem to solve.

Jot down all the reasons which you think could be contributing in the low conversion rate

We discussed the following points:-

1. High price of the premium service which is driving customers away from the offered premium service
2. Bad or ineffective marketing strategy

3. Privacy breach in the past quarter (Assumption, be a bit creative while framing the assumptions)
4. Lack of premium features which are provided by other vendors at the same price

Once you got the top 4 reasons which you need to tackle, there after you need to frame the questions which when answered will get you the answer to the actual problem.

Following are few such questions:-

1. What are the previous breach in the system and how did the company countered the same
2. Competitor analysis of Plan Cost
3. Measurement of impact of promotional plans
4. Feature comparison between the company and the competitors

Now the final recommendation, after drilling down the data should look something like this:-

1. Reduction in the price of premium service
2. Providing more features in the same cost or adding extra features in extra cost
3. Providing a good security service to assure no further data/security breach
4. Provide offers to the customers

Here, as you can see, we approached in a very methodological order by first defining the problem, then solving a specific part of problem by asking questions to the data and finally proposing solution.

676. Flipkart has recently observed that people are falling out of its buy funnel i.e. a list of pages following which a customer makes a payment.

The pages are – Listing page, Description Page, Delivery option page, Payment Page and Redirection page.

How would you identify the problem faced by the customers

Ans.

The problem – people are falling out of Flipkart's buy funnel (different pages mentioned, across the funnel)

Clarification questions – Have there been any UI/UX or any changes made to any of the pages in recent times? Are customers of all profiles/locations falling out of the funnel, or is it specific to any? What is the horizon of this fall-out (timeframe since this has been happening)

Approach – We have to calculate drop-off rate, bounce rate, click rates across page elements, error alerts sent to customers, avg. time spent across all these pages, for customers. This data would help us narrow down the page which is problematic. Based on the above tracked metrics, we can identify the problem too. If the error messages are high, then we need to fix some bug in the back-end. If the drop-off rate is high, we have to check our 'items in the cart' reminder alert system and customers' response (one probable solution). If the click rates across elements have dropped, we have to check when the UI was updated and run some A/B tests to solve for the problem.

677. For a particular e-commerce company, there has been a decline in the number and value of items stored in the cart. Tell all the possible reasons for this decline?

Some clarification questions-

1. What category do these items belong to? General or a particular category only?
2. Were there any changes made (in the UI or otherwise) on the platform?
3. What is the magnitude and timeline of the decline?
4. Items stored in the cart – Does this mean that checkout has increased? Or they do not checkout at all?

Reasons-

1. External – A competitor has these items in a better price or quality, hence users are switching to them. Another possible reason is that the general demand of those items (if they are of a particular category), is dropping.
 2. Internal – Our delivery time, item price and quality are not up to the customers' expectations.
 3. Internal – The path to checkout, till adding items to the cart, is not too clear to customers.
- Some UI related issue.

678. KFC wants to open its first branch in India, what data points should be considered before it entering in a new market?

Factors –

1. External – Existing competition performance and macro-economic conditions of India (inflation, raw items prices, consumer spending sentiment)
2. Customer – Taste, existing preferences and preferences/openness related to similar food items.
3. Internal – Small scale or large scale? (one outlet or multiple outlets launched simultaneously?)
4. Location – Population density, income groups and city development/growth factors.

679. For a particular e-commerce business the number of supplier and user both have grown by 5% week over week, but the revenue has declined by 5 %, what could be the reason?

Ans.

Problem – number of supplier and user both have grown by 5% week over week, but the revenue has declined by 5 %, what could be the reason

Question – what is the horizon of the decline? Have we made any updates to the app in recent times (related to UI etc.)

1. Low conversion rate (people are not adding items to cart, possibly due to UI related or even pricing issues)
2. High drop-off rates (people don't buy after adding to cart)
3. Bugs/Glitches in the app due to higher onboarding (simply looking out for technical issues, if any)
4. Users find competition options better. Stickiness is low (we have one-time users, who do not buy again for very long).

there could be a possibility that the demand was more in the Northern part of India but supply was in Souther part, resulting in decline of revenue

680. What are the factors to consider if you work in the sales department of Samsung and you want to start a store in one of the most crowded malls of Bangalore?

Think through all the points like

- Foot fall
- Customers
- Supply and demand issue (connect with warehouse)
- Break even point of profit
- Exit points if things fail

Ans.

Foot fall

1. Variations during weekdays and weekends (any particular hours of max crowding?)
2. Uptick during mall promotional activities (a general campaign during a holiday which invites people to the mall)
3. Variations as per the mall floor (does 1st floor have more relevant neighboring shops than the ground floor and hence attracts more potential customers?)
4. Maximum capacity during any day and average footfall per store, per type.

Customers

1. Are they coming from different parts of Bangalore or just neighboring ones? Is the mall accessible?
2. The profile of a customer – do college kids visit more or a mix of families and young working professionals? Any idea of the breakup?
3. Is the locality posh and upscale?

Supply and demand issue (connect with warehouse)

1. How far is the warehouse from the store? Can it ensure that there is no stock-out ever?
2. What is the capacity of that warehouse? Can it cater to the market demand and replenish stock in a periodic manner?
3. Demand generation activities – Do mall visitors respond to promo activities done inside (or outside) the mall?

Break even point of profit

1. Cost headers – Rent, salaries, inventory holding, promo activities (in and beyond the mall, including discounts), supply chain (trucking).
2. Revenue headers – sales made as per SKU (Did low-priced handsets sell more or high-priced ones sold less but contributed more to the overall revenue?)

Exit points if things fail

1. Relocation within the mall or an area near it?
2. Identify the main reason of failure – was it due to competitors' presence?

Amazon's Leadership Questions and Answer

Below are the 14 leadership principles of Amazon and the interviewer will keep on asking you the example of these principles in your past or current experience, so you need to understand these principles and frame examples around it

680. Customer Obsession

This world is customer centric, you make an app for customers, you provide services to customers, you do everything for your customer. If you are working in a client facing company then your client is your customer.

Give an example when you did not meet the expectation of your client. What did you do next? How did you react?

You can't just go and say that you were perfect and did not commit any mistake. May be you will remember a blunder which you did in the last MBR, you stated that but you can't answer 'How did you mend that mistake'

My example – I once created an ML model but since it was a high priority task with time crunch so I missed an important variable thus went with half cooked model.

How I mended – Looked for more variables, did rigorous EDA and came up with far better results

681. Invent and Simplify

How innovative are you?

Have you ever took up a complex problem and solved it with a fresh and simple approach? Well, There was a situation when we were trying to build an application to check if two migrated reports are identical or not, we tried to build some pipeline but it was taking a lot of time. There was a freely available software which used to break PDF in screenshots and another application which can match two pictures. Build a simple program to automate these two processes. Simple solution

682. Are Right, A lot

Leaders have a good judgment and they are mostly right.

In your previous role did you ever face an issue which you solved using your judgement or past experience?

There was a time when my colleague were trying to build a time series model but they were unable to figure out the reason for low accuracy. I asked them to do EDA and figure out some variables which might be a reason for the dip and hiccups. Then I suggested to use a model like ARIMAX which can take both time series and Regression !!

683. Hire and Develop the Best

Do you hire the best or do you hire and develop the best?

Tell me one instance when you realize that a person in your team is not the best fit for the role. What did you do?

I once had a person to whom I was giving KT in my last organization, he was not able to grasp some concepts. What I did? I made extensive documentation and videos of the complete process of KT, advised him a couple of courses on internet. The output – He was able to pick up things very quickly and was able to deliver everything at a much better speed.

684. Think Big

Tell me one instance when you thought of a big opportunity, pressed for it and delivered something beautifully

I once had the courage to talk the client of a BIG MNC about the fault in their web home page. There was an issue of cannibalization of lower pages. I worked individually on the project to justify the issue and presented with a different design

685. Frugality

Tell me one time when you tried something out of your own personal time and solved a very small but regular problem.

In my last organization people used to misplace their Key fobs once they were in their notice period. So, I created a simple tracking Excel sheet which resulted in less number of Key fobs misplacement

686. Earn Trust

Tell me the last time you earned trust with your last client. Why was there a trust issue and how did you resolve?

There was some trust issue with the Client, they were not sure of our analytical capabilities and we were monitored by their Analytics consultant. We took this as an opportunity, did cross team sessions and hands-on exercises to build capability

687. Deliver Results

What is the most difficult situation you ever faced in your life? How did you deal with it?
Created a dashboard where a client can plug and play around the UI so that we can conclude on the design of the dashboard. Thus saved a lot of time for the team

688. What are the KPIs for Nykaa?

Nykaa is an Indian online beauty and wellness retailer. As such, some potential key performance indicators (KPIs) for the company could include metrics such as website traffic, conversion rate, customer satisfaction, repeat customer rate, and revenue growth. Other financial metrics such as gross margin, inventory turnover, and customer acquisition cost may also be important for the company to track.

One of the important KPI for Nykaa would be customer retention rate, as it is a direct measure of how well the company is performing in terms of customer satisfaction and loyalty. Another key KPI would be Average Order Value (AOV), as it is a measure of how much each customer is spending on average with each purchase.

Nykaa could also track the performance of its various product categories, brand wise sales and marketing initiatives to understand which are performing well and which are not and make necessary changes.

Ultimately, the specific KPIs that Nykaa chooses to focus on will depend on its goals and objectives as a business, as well as on the specific challenges and opportunities it faces in the competitive Indian e-commerce market.

689. What are the KPIs for Zomato?

some potential KPIs for a food delivery company like Zomato could include metrics such as order fulfillment rate, delivery time, customer satisfaction, and repeat customer rate. Other financial metrics such as revenue growth, profit margin, and customer acquisition cost may also be important for the company to track. Ultimately, the specific KPIs that Zomato chooses to focus on will depend on its goals and objectives as a business.

690. What are the KPIs for Uber and Ola?

Uber, as a ride-hailing company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Uber may track include:

Rider Satisfaction: Uber measures rider satisfaction through surveys and ratings provided by riders at the end of each trip.

Driver Satisfaction: Uber tracks driver satisfaction through surveys and ratings provided by drivers to assess the quality of the driver-partner experience.

Trip Volume: Uber measures the number of trips taken on its platform to track growth and assess the overall demand for its services.

Trip duration: Uber tracks the duration of each trip to measure the efficiency of its routing and dispatching systems.

Revenue: Uber tracks its revenue as a measure of financial performance, including fare and commission revenues.

Cost per trip: Uber tracks the cost of each trip, including driver pay and incentives, to measure its efficiency and profitability.

Cancellation Rate: Uber tracks the number of cancellations made by riders and drivers, to measure the reliability of its service.

Occupancy Rate: Uber measures the number of riders per vehicle, to measure the efficiency of its use of resources.

Pick-up and Drop-off Time: Uber tracks the time taken by drivers to reach the pick-up location and drop-off location of the riders, to measure the efficiency of its service.

Fleet Utilization: Uber measures how often its vehicles are on the road, to measure the efficiency of its use of resources.

These are some of the key performance indicators that Uber may track, but it may also track other metrics depending on its specific goals and objectives.

691. What are the KPIs of Amazon or e-commerce company?

Amazon, as an e-commerce and retail company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Amazon may track include:

Sales: Amazon tracks its sales revenue as a measure of overall financial performance.

Gross Margin: Amazon tracks the difference between its revenue and cost of goods sold to measure profitability.

Net Promoter Score (NPS): Amazon measures customer satisfaction through the Net Promoter Score, which measures how likely customers are to recommend the company to others.

Customer Acquisition Cost (CAC): Amazon tracks how much it costs to acquire a new customer to measure the efficiency of its marketing and advertising efforts.

Customer Retention Rate: Amazon tracks the percentage of customers that make repeat purchases to measure customer loyalty.

Order Fulfillment Rate: Amazon tracks the percentage of orders that are fulfilled on time to measure the efficiency of its logistics and supply chain operations.

Inventory Turnover: Amazon tracks the number of times its inventory is sold and replaced over a certain period to measure efficiency in inventory management.

Logistics and Delivery Performance: Amazon tracks the performance of its logistics and delivery network, including delivery times, package tracking, and customer satisfaction.

Website Traffic: Amazon tracks the number of visitors to its website to measure the effectiveness of its digital marketing efforts.

Return Rate: Amazon tracks the percentage of products that are returned by customers to measure the quality of its products and customer satisfaction.

These are some of the key performance indicators that Amazon may track, but it may also track other metrics depending on its specific goals and objectives.

692. What are the KPIs of Google?

Google, as a technology company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Google may track include:

Revenue: Google tracks its revenue as a measure of overall financial performance.

Advertising Revenue: Google tracks the revenue generated from advertising, as advertising is a major source of revenue for the company.

Search Query Volume: Google tracks the number of search queries made on its search engine to measure the popularity of its search platform.

Click-through Rate (CTR): Google tracks the number of clicks on ads divided by the number of ad impressions to measure the effectiveness of its advertising.

Bounce Rate: Google tracks the percentage of users who leave a website after only visiting one page, to measure the engagement and relevance of its search results.

Time on Site: Google tracks the amount of time users spend on a website to measure the relevance and usefulness of its search results.

Pageviews: Google tracks the number of pages viewed by users to measure the engagement of its search results.

Market Share: Google tracks its market share in search engine, advertising and other related market to measure its competitiveness.

Ad Quality: Google tracks the quality of its ads, such as the relevance, appropriateness, and usefulness of the ads, to measure the effectiveness of its advertising platform.

Return on Ad Spend (ROAS): Google tracks the return on investment (ROI) of its advertising campaigns, to measure the efficiency of its advertising.

These are some of the key performance indicators that Google may track, but it may also track other metrics depending on its specific goals and objectives.

693. What are the KPIs for Apple?

Apple, as a technology company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Apple may track include:

Revenue: Apple tracks its revenue as a measure of overall financial performance.

Product Sales: Apple tracks the sales of its various product lines, such as iPhones, iPads, Macs, and accessories, to measure the success of its product offerings.

Gross Margin: Apple tracks the difference between its revenue and cost of goods sold to measure profitability.

Market Share: Apple tracks its market share in various product categories, such as smartphones and personal computers, to measure its competitiveness.

Customer Satisfaction: Apple tracks customer satisfaction through surveys and ratings provided by customers to measure the quality of its products and services.

Brand Loyalty: Apple tracks the percentage of customers that continue to buy Apple products over time to measure customer loyalty.

Return Rate: Apple tracks the percentage of products that are returned by customers to measure the quality of its products and customer satisfaction.

Inventory turnover: Apple tracks the number of times its inventory is sold and replaced over a certain period to measure efficiency in inventory management.

Product Development: Apple tracks the development of new products and services to measure the efficiency and effectiveness of its research and development efforts.

Operating Margin: Apple tracks the operating margin, which is the difference between revenue and operating expenses as a percentage of revenue, to measure the efficiency of its operations.

These are some of the key performance indicators that Apple may track, but it may also track other metrics depending on its specific goals and objectives.

694. What are the KPIs of Tesla?

Tesla, as a electric vehicle and energy company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Tesla may track include:

Revenue: Tesla tracks its revenue as a measure of overall financial performance.

Vehicle Sales: Tesla tracks the sales of its electric vehicles (EVs) to measure the success of its product offerings.

Gross Margin: Tesla tracks the difference between its revenue and cost of goods sold to measure profitability.

Market Share: Tesla tracks its market share in the electric vehicle market to measure its competitiveness.

Production Volume: Tesla tracks the number of vehicles produced and delivered to measure the efficiency of its manufacturing operations.

Delivery Time: Tesla tracks the time it takes to deliver vehicles to customers to measure the efficiency of its logistics and supply chain operations.

Energy Storage and Generation: Tesla tracks the production and storage of energy through its solar panels and energy storage systems to measure the efficiency of its renewable energy operations.

Customer Satisfaction: Tesla tracks customer satisfaction through surveys and ratings provided by customers to measure the quality of its products and services.

Battery Production: Tesla tracks the production of batteries to measure the efficiency of its manufacturing operations, and also to forecast the demand for its electric vehicles.

Research and Development: Tesla tracks the development of new products and services to measure the efficiency and effectiveness of its research and development efforts.

695. What are the KPIs for Walmart?

Walmart, as a retail company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Walmart may track include:

Revenue: Walmart tracks its revenue as a measure of overall financial performance.

Same-store Sales: Walmart tracks the sales of stores open for at least one year to measure the performance of its existing store base.

Gross Margin: Walmart tracks the difference between its revenue and cost of goods sold to measure profitability.

Market Share: Walmart tracks its market share in retail industry to measure its competitiveness.

Customer Satisfaction: Walmart tracks customer satisfaction through surveys and ratings provided by customers to measure the quality of its products and services.

Inventory turnover: Walmart tracks the number of times its inventory is sold and replaced over a certain period to measure efficiency in inventory management.

Online Sales: Walmart tracks the sales made through its online platform to measure the success of its e-commerce efforts.

Same-day Delivery and Pickup: Walmart tracks the use of its same-day delivery and pickup services to measure the efficiency and popularity of these services.

Employee Turnover: Walmart tracks the rate at which employees leave the company to measure the effectiveness of its human resources practices.

Operating Margin: Walmart tracks the operating margin, which is the difference between revenue and operating expenses as a percentage of revenue, to measure the efficiency of its operations.

These are some of the key performance indicators that Walmart may track, but it may also track other metrics depending on its specific goals and objectives.

696. What are the KPIs for Microsoft?

Microsoft, as a technology company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Microsoft may track include:

Revenue: Microsoft tracks its revenue as a measure of overall financial performance.

Operating Income: Microsoft tracks its operating income as a measure of profitability

Market Share: Microsoft tracks its market share in various product categories, such as operating systems, office software, and cloud services, to measure its competitiveness.

Customer Satisfaction: Microsoft tracks customer satisfaction through surveys and ratings provided by customers to measure the quality of its products and services.

Cloud Services: Microsoft tracks the usage and revenue generated from its cloud services, such as Azure and Office 365, to measure the success of its cloud computing efforts.

Productivity and Business Processes: Microsoft tracks the adoption and usage of its productivity and business software, such as Office and Dynamics, to measure the success of these product lines.

Gaming: Microsoft tracks the revenue and usage of its gaming products, such as Xbox, to measure the success of this product line.

Research and Development: Microsoft tracks the development of new products and services to measure the efficiency and effectiveness of its research and development efforts.

Talent and Human Capital: Microsoft tracks the employee satisfaction, retention, and diversity to measure the effectiveness of its human resources practices.

Cybersecurity: Microsoft tracks the security of its products and services to measure the effectiveness of its cybersecurity efforts.

These are some of the key performance indicators that Microsoft may track, but it may also track other metrics depending on its specific goals and objectives.

697. What are the KPIs for iTC?

iTC is a conglomerate company which has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that iTC may track include:

Revenue: iTC tracks its revenue as a measure of overall financial performance.

Net Profit Margin: iTC tracks the net profit margin which is the difference between revenue and expenses as a percentage of revenue, to measure the efficiency of its operations.

Market Share: iTC tracks its market share in various product categories, such as tobacco, paper, FMCG and hotels to measure its competitiveness.

Customer Satisfaction: iTC tracks customer satisfaction through surveys and ratings provided by customers to measure the quality of its products and services.

Return on Investment (ROI): iTC tracks the return on investment of its various business units to measure the efficiency of its operations.

Employee Turnover: iTC tracks the rate at which employees leave the company to measure the effectiveness of its human resources practices.

Occupancy Rate: iTC tracks the occupancy rate of its hotels to measure the demand for its services.

Research and Development: iTC tracks the development of new products and services to measure the efficiency and effectiveness of its research and development efforts.

Occupational Health and Safety (OHS): iTC tracks the safety of its employees and the performance of its safety programs to measure the effectiveness of its OHS efforts.

Sustainability: iTC tracks its performance on environmental, social and governance (ESG) initiatives to measure the effectiveness of its sustainability efforts.

698. What are the KPIs of Pfizer?

Pfizer, as a pharmaceutical company, has several key performance indicators (KPIs) that it uses to measure the success and efficiency of its operations. Some of the main KPIs that Pfizer may track include:

Revenue: Pfizer tracks its revenue as a measure of overall financial performance.

Product Sales: Pfizer tracks the sales of its various product lines, such as drugs, vaccines and consumer healthcare products, to measure the success of its product offerings.

Gross Margin: Pfizer tracks the difference between its revenue and cost of goods sold to measure profitability.

Market Share: Pfizer tracks its market share in various product categories, such as drugs and vaccines, to measure its competitiveness.

Research and Development (R&D): Pfizer tracks the efficiency and effectiveness of its R&D efforts to measure the pipeline of new products and the potential for future revenue.

Clinical trial success rate: Pfizer tracks the success rate of its clinical trials to measure the efficiency and effectiveness of its R&D efforts.

Regulatory Approvals: Pfizer tracks the number of regulatory approvals it receives for its products to measure the efficiency of its regulatory affairs efforts.

Patent expirations: Pfizer tracks the expiration of patents on its products to measure the potential for future revenue and the need for new product development.

Manufacturing efficiency: Pfizer tracks the efficiency of its manufacturing operations to measure the cost-effectiveness of its operations and to ensure compliance with regulations.

Collaborations: Pfizer tracks the success of its collaborations and partnerships with other companies and organizations in the pharmaceutical industry to measure the effectiveness of its business development efforts.

These are some of the key performance indicators that Pfizer may track, but it may also track other metrics depending on its specific goals and objectives.

699. How would you go about designing a recommendation system for an e-commerce website, and what data sources and algorithms would you consider?

Designing a recommendation system for an e-commerce website typically involves the following steps:

Define the problem and success criteria: The first step is to define the business problem and what success looks like. This includes determining what types of recommendations to provide, how to measure their effectiveness, and what data sources to use.

Collect and preprocess data: Next, we need to collect and preprocess data from various sources, such as customer browsing and purchase histories, product catalogs, and customer demographics.

Feature engineering: We need to identify and create features that describe the characteristics of the products and users, such as product categories, brands, price, and user preferences.

Algorithm selection: There are several algorithms we can use to generate recommendations, including collaborative filtering, content-based filtering, and hybrid approaches. We need to select the algorithm that is best suited for the specific use case and data available.

Model training and evaluation: We train the recommendation model on historical data and evaluate its performance on a holdout set of data. We can use metrics such as precision, recall, and F1 score to evaluate the model's performance.

Deployment and monitoring: Once the model is trained, we deploy it in production and monitor its performance over time. We can use A/B testing to compare the performance of the recommendation model to that of a baseline.

Some common data sources and algorithms used in recommendation systems include:

Data Sources:

Customer purchase history

Customer browsing behavior

Product catalogs and descriptions

Customer demographic information

Algorithms:

Collaborative filtering

Content-based filtering

Matrix factorization

Association rules

Deep learning-based methods

In summary, designing a recommendation system for an e-commerce website involves identifying the problem and success criteria, collecting and preprocessing data, feature engineering, selecting an appropriate algorithm, training and evaluating the model, and deploying and monitoring it in production.

700. Imagine you've been given a dataset on customer churn in a telecom company. What kind of analysis would you conduct to identify the key factors that contribute to customer churn, and what recommendations would you make to reduce it?

To identify the key factors that contribute to customer churn in a telecom company, I would conduct the following analysis:

Data cleaning and preparation: The first step would be to clean and prepare the data by removing missing values, handling outliers, and transforming the data into a format that can be used for analysis.

Exploratory data analysis: I would perform exploratory data analysis (EDA) to gain insights into the dataset and understand the distribution of the data, correlations between variables, and identify any patterns or trends.

Feature engineering: Next, I would engineer features that could potentially be related to customer churn, such as the length of time the customer has been with the company, the number of services they have subscribed to, their payment history, and their customer service interactions.

Model training and evaluation: I would then train a predictive model to identify the key factors that contribute to customer churn. Some models that can be used for this purpose include logistic regression, decision trees, and random forests. I would use evaluation metrics such as accuracy, precision, recall, and F1 score to assess the performance of the model.

Feature importance analysis: Once the model is trained, I would perform a feature importance analysis to identify the most important factors that contribute to customer churn. This can be done by examining the weights or coefficients of the features in the model or by using techniques such as permutation feature importance or SHAP (SHapley Additive exPlanations).

Based on the results of the analysis, I would make the following recommendations to reduce customer churn in the telecom company:

Improve customer service: The analysis may reveal that poor customer service is a key factor contributing to churn. The company can take steps to improve customer service, such as providing better training to customer service representatives or offering incentives to customers who have had a negative experience.

Offer incentives to retain customers: The analysis may reveal that customers who are offered incentives, such as discounts or special offers, are less likely to churn. The company can offer these incentives to customers who are at risk of churning.

Simplify product offerings: The analysis may reveal that customers are overwhelmed by the number of product offerings and find it difficult to choose the right product for their needs. The company can simplify its product offerings or provide personalized recommendations to help customers choose the right product.

Improve billing and payment processes: The analysis may reveal that billing and payment issues are a key factor contributing to churn. The company can improve its billing and payment processes, such as by offering more flexible payment options or by providing better billing statements that are easier to understand.

In summary, to reduce customer churn in a telecom company, I would conduct an analysis of the data to identify the key factors contributing to churn, and make recommendations based on the results of the analysis.

701. Suppose you're working on a project to improve customer engagement for a mobile app. What metrics would you use to measure engagement, and what techniques would you employ to increase it?

When working on a project to improve customer engagement for a mobile app, the following metrics can be used to measure engagement:

Active users: The number of unique users who have interacted with the app during a given time period.

Retention rate: The percentage of users who return to the app after their initial visit.

Session length: The amount of time users spend on the app during a session.

Frequency of use: The number of times users open the app in a given time period.

In-app events: The number of actions taken by users within the app, such as making a purchase or sharing content.

To increase customer engagement for a mobile app, the following techniques can be employed:

Personalization: Customizing the user experience based on user preferences and behavior.

Push notifications: Sending relevant and timely notifications to users to encourage them to return to the app.

Gamification: Adding game-like features to the app to increase user engagement and motivation.

Incentives: Offering rewards or discounts to users who perform certain actions within the app.

User feedback: Gathering feedback from users to identify pain points and areas for improvement.

A/B testing: Testing different versions of the app to see which features and design elements resonate best with users.

By analyzing these metrics and employing the above techniques, it is possible to increase customer engagement and retention for a mobile app.

702. You're tasked with building a fraud detection system for a bank. What features and data sources would you consider, and what models and techniques would you use to identify fraudulent activity?

When building a fraud detection system for a bank, the following features and data sources can be considered:

User behavior: Patterns of transactions and activities can help to identify unusual behavior.

Account information: Suspicious activity can be detected by analyzing account information, such as the number of accounts opened, the age of the account, and the type of account.

Geolocation data: The location of transactions can be analyzed to identify unusual patterns or locations.

Device information: The device used for a transaction can be analyzed to detect unusual patterns or to identify known fraudulent devices.

External data sources: External data sources, such as public records or credit bureaus, can be used to verify information and detect fraud.

To identify fraudulent activity in a bank, the following models and techniques can be used:

Rule-based systems: Rule-based systems use a set of predefined rules to identify fraudulent activity based on specific patterns.

Anomaly detection: Anomaly detection models use statistical techniques to identify transactions that deviate significantly from normal patterns.

Machine learning: Machine learning models use algorithms to identify fraudulent activity based on historical data.

Social network analysis: Social network analysis can be used to detect fraudulent activity by identifying patterns of connections between accounts.

Human experts: Human experts can be used to manually review transactions and flag suspicious activity.

By combining these models and techniques with the relevant features and data sources, it is possible to build a comprehensive fraud detection system that can identify and prevent fraudulent activity in a bank.

703. How would you go about developing a predictive maintenance system for a fleet of trucks, and what data sources and algorithms would you use to anticipate maintenance needs?

When developing a predictive maintenance system for a fleet of trucks, the following data sources and algorithms can be considered:

Sensor data: Sensor data, such as engine temperature, oil pressure, and tire pressure, can be used to monitor the health of the trucks and detect potential issues.

Maintenance records: Historical maintenance records can provide insights into the types of problems that have occurred in the past and the frequency of maintenance needs.

Driving behavior data: Data on driving behavior, such as speed, acceleration, and braking patterns, can be used to anticipate maintenance needs based on wear and tear on the truck.

External data sources: External data sources, such as weather and road condition data, can be used to anticipate maintenance needs based on the impact of external factors on the trucks.

To anticipate maintenance needs for a fleet of trucks, the following algorithms can be used:

Regression models: Regression models can be used to predict the likelihood of maintenance needs based on historical data.

Time series models: Time series models can be used to analyze patterns over time and anticipate maintenance needs based on historical trends.

Classification models: Classification models can be used to predict the likelihood of specific types of maintenance needs based on historical data.

Clustering algorithms: Clustering algorithms can be used to group trucks based on similar characteristics and identify maintenance needs specific to those groups.

By combining these algorithms with the relevant data sources, it is possible to build a comprehensive predictive maintenance system that can anticipate maintenance needs and prevent costly breakdowns in a fleet of trucks.

704. You've been asked to build a chatbot for a customer service team. What kind of data would you need to train the chatbot, and what technologies and techniques would you use to ensure it provides accurate and helpful responses?

To build a chatbot for a customer service team, the following data and technologies can be considered:

Data sources: The chatbot can be trained on data from various sources such as customer service call transcripts, online chat transcripts, and customer feedback. The data should be pre-processed, labeled, and structured in a way that can be used to train the chatbot.

Natural Language Processing (NLP): NLP techniques can be used to analyze the input text and understand the customer's intent, context, and sentiment. These techniques can include text tokenization, named entity recognition, part-of-speech tagging, and sentiment analysis.

Machine Learning algorithms: Machine learning algorithms can be used to train the chatbot to recognize and respond to customer queries. This can include supervised learning algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forests, as well as unsupervised learning algorithms such as clustering and topic modeling.

Dialogue management: Dialogue management techniques can be used to manage the conversation flow and ensure that the chatbot provides accurate and helpful responses. This can include rule-based approaches or reinforcement learning techniques.

Integration with backend systems: The chatbot needs to be integrated with the backend systems such as CRM, ERP, and knowledge management systems, to provide accurate and personalized responses to customer queries.

Continuous improvement: Continuous learning and improvement techniques can be used to optimize the performance of the chatbot over time. This can include user feedback, chatbot analytics, and continuous training of the chatbot using new data.

Overall, building a chatbot for a customer service team requires a combination of data, technologies, and techniques to ensure that the chatbot provides accurate and helpful responses to customers. It requires an iterative approach of testing, refining, and improving the chatbot's performance over time.

705. You're working on a project to optimize the pricing strategy for a ridesharing company. What data sources would you use to inform your analysis, and what techniques would you employ to set optimal prices in real time?

To optimize the pricing strategy for a ridesharing company, the following data sources can be used:

Ride data: Data on the rides that have been taken in the past can be used to understand demand patterns and price sensitivity. This data can be analyzed to identify patterns of demand and supply for different times of the day, week, or year.

Market data: Market data on competitor pricing, economic trends, and other external factors that affect demand and supply can also be used.

Customer data: Data on customer preferences, demographics, and behavior can be analyzed to understand how different customers respond to different price points.

Operational data: Data on operational costs, such as fuel and vehicle maintenance, can be used to determine the minimum fare needed to cover costs.

To set optimal prices in real time, the following techniques can be used:

Dynamic pricing: Dynamic pricing algorithms can be used to adjust prices in real time based on demand and supply. These algorithms can take into account historical demand patterns, current demand, and competitor pricing to set prices that maximize revenue.

Predictive modeling: Predictive models can be used to forecast demand for different times and locations. These models can be used to set prices that maximize revenue and ensure that there are enough drivers to meet demand.

A/B testing: A/B testing can be used to test different pricing strategies and identify the most effective one. This can involve testing different prices for different times of the day, different locations, or different customer segments.

Machine learning: Machine learning algorithms can be used to analyze large amounts of data to identify patterns and predict demand. These algorithms can be used to develop more accurate pricing models and identify opportunities to increase revenue.

706. You have been given a dataset containing millions of online transactions for an e-commerce platform. How would you analyze the data to identify customer behavior patterns and preferences, and what insights would you extract to optimize the platform's user experience and increase sales?

To analyze the data and extract insights to optimize the platform's user experience and increase sales, I would follow these steps:

Data Preparation: The first step would be to clean and preprocess the data, including removing duplicates, missing values, and outliers. This would ensure the data is ready for analysis.

Exploratory Data Analysis: I would perform exploratory data analysis to gain insights into customer behavior patterns and preferences. This would involve analyzing different variables, such as customer demographics, products viewed, time of day, device used, etc., to identify trends and patterns.

Customer Segmentation: Based on the insights gained from exploratory data analysis, I would segment customers based on their behavior and preferences. This would allow us to better understand our customer base and tailor our marketing efforts to specific segments.

Recommender Systems: I would build a recommender system to recommend products to customers based on their behavior and preferences. This would not only increase sales but also improve the overall user experience.

Predictive Analytics: I would use predictive analytics to forecast future sales and identify potential areas of growth. This would allow us to make data-driven decisions and stay ahead of the competition.

A/B Testing: I would conduct A/B testing to test different pricing strategies, product recommendations, and marketing campaigns. This would help us identify what works best for our customers and improve the platform's user experience.

Overall, the goal would be to use data analysis and insights to optimize the platform's user experience and increase sales.

707. You are working on a project to build a recommendation engine for a streaming media platform. How would you design and implement the recommendation engine, and what algorithms and data sources would you use to provide personalized recommendations to users?

Designing and implementing a recommendation engine for a streaming media platform typically involves several steps:

Define the business problem and goals: The first step is to define the business problem and goals. For example, the goal may be to increase user engagement by providing personalized recommendations to users.

Gather and preprocess data: The next step is to gather and preprocess data from various sources such as user profiles, viewing history, and product catalog. This data needs to be cleaned, transformed, and formatted in a way that can be used by the recommendation engine.

Choose a recommendation algorithm: There are several types of recommendation algorithms such as collaborative filtering, content-based filtering, and hybrid filtering. Depending on the business problem and data available, you need to choose the appropriate algorithm to use.

Train and test the recommendation model: Once the recommendation algorithm has been chosen, the next step is to train and test the recommendation model using historical data. This involves splitting the data into training and testing sets, training the model on the training set, and evaluating its performance on the testing set.

Integrate the model into the platform: After the model has been trained and tested, it needs to be integrated into the streaming media platform. This involves setting up the infrastructure and APIs to serve personalized recommendations to users.

Monitor and improve the recommendation engine: Finally, it's important to monitor the performance of the recommendation engine and continuously improve it over time. This can be done by collecting feedback from users and measuring key performance metrics such as click-through rates and conversion rates.

To provide personalized recommendations to users, you may use various data sources such as user demographics, viewing history, search queries, ratings and reviews, and social media activity. The recommendation algorithm used will depend on the type of data available and the business problem you are trying to solve. For example, collaborative filtering may be used to recommend content based on users' viewing history and similar users' behavior, while content-based filtering may be used to recommend content based on the characteristics of the content itself. Hybrid filtering combines both collaborative filtering and content-based filtering to provide more accurate recommendations.

To improve the recommendation engine, you may also use techniques such as A/B testing and reinforcement learning. A/B testing involves testing different versions of the recommendation engine with a subset of users to see which version performs better. Reinforcement learning involves training the recommendation model to maximize a reward signal such as click-through rate or revenue.

708. You are tasked with developing a predictive model for a healthcare provider to identify patients who are at risk of developing chronic conditions. What data sources and features would you use to train the model, and what algorithms and techniques would you employ to improve the accuracy of the predictions and inform targeted interventions?

Developing a predictive model for identifying patients at risk of chronic conditions is a challenging but critical task in healthcare. Some data sources and features that could be used to train the model include electronic health records, medical claims data, socioeconomic data, genetic data, and lifestyle data.

To improve the accuracy of predictions and inform targeted interventions, various algorithms and techniques could be employed, such as:

Feature engineering: This involves selecting and transforming relevant features that are most predictive of the outcome.

Ensemble models: Ensemble models can be used to combine the predictions of several different models to improve the overall accuracy of the prediction.

Deep learning models: Deep learning models, such as neural networks, can be used to identify complex patterns in the data that might be missed by traditional models.

Risk stratification: Patients can be stratified based on their risk of developing chronic conditions, allowing targeted interventions to be deployed to those most at risk.

Validation and testing: The model should be validated on a separate dataset to ensure that it is accurate and generalizable.

Overall, the goal is to develop a model that can accurately identify patients at risk of developing chronic conditions, so that targeted interventions can be deployed to prevent or manage these conditions.

709. You are working on a project to optimize a supply chain for a consumer goods company. What data sources and analytics tools would you use to identify bottlenecks and inefficiencies in the supply chain, and what recommendations would you make to improve performance and reduce costs?

To optimize a supply chain for a consumer goods company, the following data sources and analytics tools can be used:

Supply chain data: Data related to suppliers, manufacturers, distributors, and retailers can be collected to identify bottlenecks and inefficiencies in the supply chain.

Logistics data: Data related to transportation, warehousing, and inventory can be analyzed to identify areas where delays or overstocking are occurring.

Market data: Data related to customer demand, seasonal trends, and competitor activity can be analyzed to anticipate demand and plan production accordingly.

Analytics tools: Tools like network optimization software, simulation tools, and machine learning algorithms can be used to identify the root causes of inefficiencies and optimize the supply chain.

Based on the analysis, the following recommendations can be made to improve supply chain performance and reduce costs:

Reduce lead times: By working closely with suppliers and manufacturers, lead times can be reduced, which can help to optimize inventory levels and reduce transportation costs.

Optimize transportation: By using network optimization software, the most efficient transportation routes and modes can be identified, which can help to reduce transportation costs and improve delivery times.

Improve forecasting: By using machine learning algorithms to analyze market data, demand can be forecasted more accurately, which can help to reduce inventory costs and prevent stockouts.

Implement automation: By implementing automation technologies in the warehouse and in transportation, the cost of labor can be reduced, and accuracy can be improved.

Implement data-driven decision-making: By using analytics tools and data to make supply chain decisions, the supply chain can be continuously optimized, and inefficiencies can be identified and addressed quickly.

710. You have been given a large dataset of sensor data from a fleet of industrial equipment. How would you analyze the data to identify patterns of performance and maintenance needs, and what techniques and models would you use to inform a predictive maintenance strategy?

To analyze the sensor data and inform a predictive maintenance strategy for the industrial equipment fleet, I would follow these steps:

Data exploration: First, I would explore the data to understand its characteristics and identify any patterns, anomalies, or missing values. I would also identify the types of sensors, their readings, and the frequency of data collection.

Feature engineering: I would extract relevant features from the sensor data that could be used to identify patterns and maintenance needs. This could include statistical features such as mean, standard deviation, and range, as well as time-based features such as trends, seasonality, and anomalies.

Data preprocessing: I would preprocess the data to ensure it is in a suitable format for modeling. This could involve handling missing or erroneous values, scaling and normalizing the data, and converting it to a suitable format for the chosen algorithm.

Modeling: I would use a suitable algorithm, such as regression or clustering, to identify patterns of performance and maintenance needs from the sensor data. I would also explore machine learning techniques, such as supervised and unsupervised learning, to predict future maintenance needs based on historical data.

Evaluation: I would evaluate the performance of the predictive maintenance model using suitable metrics such as accuracy, precision, and recall. I would also compare the performance of different models to select the best one.

Implementation: Finally, I would implement the predictive maintenance strategy by integrating the model into the equipment management system. This would involve setting thresholds for maintenance needs and generating alerts for maintenance personnel when those thresholds are exceeded.

Overall, the goal of this project would be to improve equipment uptime, reduce maintenance costs, and increase productivity by anticipating and preventing equipment failures before they occur.

711. You are tasked with developing a fraud detection system for a financial services company. What features and data sources would you use to train the model, and what algorithms and techniques would you employ to improve the accuracy of fraud identification and reduce false positives?

To develop a fraud detection system for a financial services company, I would consider several features and data sources to train the model. Some possible features and data sources that could be useful are:

Transaction amount and frequency: Unusually large transactions or a high frequency of transactions can be an indicator of fraudulent activity.

Geographic location: Transactions originating from a location that is different from the customer's typical location can be an indicator of fraudulent activity.

Time of day: Transactions that occur at unusual times of day can be an indicator of fraudulent activity.

Card and account details: Any changes to card or account details can be an indicator of fraudulent activity.

User behavior: Any deviation from a user's typical behavior, such as changes in purchasing patterns, can be an indicator of fraudulent activity.

In terms of algorithms and techniques, I would consider the following:

Machine learning algorithms: Supervised learning algorithms like logistic regression, decision trees, and random forests can be trained on historical data to identify patterns of fraudulent behavior. Unsupervised learning algorithms like clustering and anomaly detection can be used to identify unusual behavior that may indicate fraud.

Feature engineering: Careful feature engineering can help to improve the accuracy of fraud detection models. For example, transforming transaction amount and frequency data into log scale can help to reduce the influence of outliers.

Ensemble methods: Ensemble methods like bagging, boosting, and stacking can be used to improve the accuracy of fraud detection models by combining the predictions of multiple models.

Data preprocessing: Preprocessing techniques like normalization and scaling can help to improve the performance of machine learning models on the data.

Explainability: In order to ensure the model is transparent and can be explained, techniques like LIME and SHAP values can be used to explain how the model made its predictions.

Overall, building a successful fraud detection system requires a combination of domain expertise, data engineering, and machine learning techniques.

712. You are working on a project to optimize marketing spend for a consumer packaged goods company. What data sources and analytics tools would you use to inform decisions around channel allocation, audience targeting, and ad creative, and how would you measure the impact of the marketing campaigns on sales and customer engagement?

To optimize marketing spend for a consumer packaged goods company, I would consider the following data sources and analytics tools:

Sales data: To understand how marketing campaigns are impacting sales, I would use sales data from the company's point of sale (POS) systems. This data would give me insight into how different marketing campaigns are affecting sales in different regions and channels.

Customer data: To understand who is buying the company's products, I would use customer data from the company's CRM system. This data would allow me to identify key customer segments and target marketing campaigns to those segments.

Market data: To understand market trends and competitive dynamics, I would use market data from third-party sources. This data would allow me to benchmark the company's performance against its competitors and identify opportunities for growth.

Web analytics data: To understand how customers are interacting with the company's online presence, I would use web analytics data from the company's website and social media channels. This data would allow me to identify customer preferences and optimize the company's digital marketing campaigns.

Attribution modeling: To measure the impact of marketing campaigns on sales and customer engagement, I would use attribution modeling techniques. This would allow me to assign credit to different marketing channels and campaigns for driving conversions and customer engagement.

To inform decisions around channel allocation, audience targeting, and ad creative, I would use a combination of data analysis and machine learning techniques. Specifically, I would use clustering algorithms to identify customer segments and target marketing campaigns to those segments. I would also use regression analysis to identify which marketing channels and creative are driving the highest ROI.

Finally, to measure the impact of the marketing campaigns on sales and customer engagement, I would use A/B testing to compare the performance of different marketing campaigns and channels. I would also use correlation analysis to identify relationships between marketing campaigns and sales/customer engagement.

713. You have been given a dataset of social media activity for a large retailer. How would you analyze the data to identify trends and sentiment among customers, and what recommendations would you make to improve the retailer's social media presence and customer engagement?

To analyze the social media activity dataset, I would follow these steps:

Define the problem and objectives: The first step is to define the problem and objectives of the analysis. In this case, the objective could be to understand customer sentiment towards the retailer and identify opportunities to improve engagement.

Collect and preprocess the data: The next step is to collect and preprocess the data. This involves gathering data from various social media platforms and cleaning and transforming the data to make it suitable for analysis.

Perform exploratory data analysis (EDA): EDA is crucial to understand the data and identify patterns, trends, and relationships. The EDA could involve visualizations such as word clouds, sentiment analysis, and trend analysis.

Perform sentiment analysis: Sentiment analysis is a powerful tool to understand customer sentiment towards the retailer. The sentiment analysis could involve classifying the data into positive, negative, or neutral sentiments.

Identify influencers and brand advocates: Identifying influencers and brand advocates could help the retailer to leverage their social media presence and improve engagement.

Identify opportunities to improve customer engagement: Based on the analysis, the retailer could identify opportunities to improve customer engagement, such as responding to negative feedback, improving customer service, and creating targeted marketing campaigns.

Monitor and measure results: To ensure the success of the social media strategy, it is important to monitor and measure the results regularly. This could involve tracking engagement metrics such as likes, comments, shares, and click-through rates.

Overall, analyzing social media data could provide valuable insights into customer sentiment and behavior, and help retailers to improve engagement and drive sales.

714. You work for an e-commerce company that sells a variety of products online. The company has a website, mobile app, and social media accounts, and it uses various online marketing channels to drive traffic to its website and app. The company has been experiencing declining sales over the past few months and has brought you in to help identify the problem and recommend solutions.

Question:

How would you use digital analytics to help the e-commerce company identify the problem and recommend solutions to increase sales?

Ans

Potential approach:

Set up tracking and data collection:

The first step would be to ensure that the company has proper tracking and data collection mechanisms in place for all its online assets. This includes the website, mobile app, and social media accounts. This would involve implementing tags and pixels for web analytics and setting up mobile app tracking SDKs.

Identify key performance indicators (KPIs):

Once tracking is set up, the next step would be to identify the KPIs that will help the company understand how its online assets are performing. KPIs could include website traffic, conversion rates, revenue per visit, app downloads, app engagement, social media engagement, and more.

Conduct a website and app audit:

Using web analytics tools, conduct a website and app audit to identify areas where the user experience can be improved. Look at metrics like bounce rates, exit rates, and time on site to identify pages or screens that may be causing users to leave the site or app.

Analyze the customer journey:

Using a combination of web analytics and mobile app analytics tools, analyze the customer journey to identify areas where users are dropping off or experiencing friction. Look at metrics like click-through rates, abandonment rates, and time to complete tasks to identify areas where the user experience can be improved.

Analyze marketing channels:

Using marketing analytics tools, analyze the performance of each marketing channel that the company is using to drive traffic to its website and app. Look at metrics like click-through rates, conversion rates, and cost per acquisition to identify channels that may be underperforming.

Conduct a customer segmentation analysis:

Using data from the website, mobile app, and social media accounts, conduct a customer segmentation analysis to identify different customer segments and their behaviors. Look at metrics like average order value, frequency of purchase, and customer lifetime value to identify the most valuable customer segments.

Develop a testing and optimization plan:

Using the insights gathered from the previous steps, develop a testing and optimization plan to improve the user experience and increase sales. This could involve testing new landing pages, optimizing the checkout process, improving the app onboarding process, and testing new marketing channels.

Implement and measure:

Finally, implement the testing and optimization plan and measure the impact on key performance indicators. Use web analytics, mobile app analytics, and marketing analytics tools to track the impact of changes on KPIs and make adjustments as needed.

Overall, the key to using digital analytics to identify and solve e-commerce problems is to have a comprehensive understanding of the customer journey and to use data to inform decisions about user experience and marketing strategies.

715. KPIs for Telecommunication companies

Key Performance Indicators (KPIs) are critical for measuring the success of any business, including telecommunication companies. Some KPIs that telecommunication companies may use include:

Average Revenue per User (ARPU): This metric measures the average amount of revenue a company generates per customer.

Churn Rate: This metric measures the percentage of customers who cancel their service or stop using a product over a given period.

Customer Satisfaction (CSAT) Score: This metric measures how satisfied customers are with a company's products and services.

Network Availability: This metric measures the percentage of time that a company's network is available to customers.

Average Speed of Answer (ASA): This metric measures the average time it takes for a customer to speak with a customer service representative.

Average Handle Time (AHT): This metric measures the average time a customer spends on a call with a customer service representative.

First Call Resolution (FCR): This metric measures the percentage of customer issues that are resolved on the first call.

Cost per Acquisition (CPA): This metric measures the cost of acquiring a new customer.

Average Revenue per Account (ARPA): This metric measures the average amount of revenue a company generates per account.

Lifetime Value (LTV): This metric measures the total amount of revenue a customer will generate over the entire duration of their relationship with the company.

These KPIs can help a telecommunication company measure and improve its performance in areas that are critical to its success.

716. KPIs for Human Resource department

Key Performance Indicators (KPIs) are important for measuring the effectiveness and success

of any department, including Human Resources. Here are some KPIs that a Human Resources department may use:

Time-to-Fill: This metric measures the time it takes to fill a job vacancy, from the date the job is posted to the date an offer is accepted.

Turnover Rate: This metric measures the percentage of employees who leave the company over a given period.

Employee Engagement: This metric measures the level of employee engagement and satisfaction with the company.

Absenteeism Rate: This metric measures the percentage of employee absences, including planned and unplanned absences.

Cost per Hire: This metric measures the cost of recruiting and hiring a new employee.

Training and Development ROI: This metric measures the return on investment (ROI) of training and development programs.

Employee Retention: This metric measures the percentage of employees who remain with the company over a given period.

Diversity and Inclusion: This metric measures the company's progress in promoting diversity and inclusion in the workplace.

HR to Employee Ratio: This metric measures the number of HR employees relative to the total number of employees in the company.

Performance Appraisal Completion Rate: This metric measures the percentage of employees who receive a performance appraisal within a given period.

These KPIs can help a Human Resources department measure its performance, identify areas for improvement, and make data-driven decisions to optimize employee satisfaction, engagement, and productivity.

717. KPIs for Data Platform of any Technology company

Key Performance Indicators (KPIs) are important for measuring the success of any data platform. Here are some KPIs that a data platform may use:

Data Availability: This metric measures the percentage of time that data is available to users.

Data Latency: This metric measures the time it takes for data to be available in the platform from the source system.

Data Quality: This metric measures the accuracy, completeness, and consistency of the data in the platform.

Data Usage: This metric measures the volume and frequency of data usage by users.

Query Performance: This metric measures the speed and efficiency of data queries and processing.

System Availability: This metric measures the percentage of time that the data platform is available to users.

Data Security: This metric measures the level of security and protection of data in the platform.

Cost per Query: This metric measures the cost of processing and querying data in the platform.

Data Governance Compliance: This metric measures the adherence to data governance policies and regulations.

User Satisfaction: This metric measures the level of satisfaction of users with the data platform, its functionality, and usability.

These KPIs can help a data platform measure its performance, identify areas for improvement, and make data-driven decisions to optimize data availability, quality, security, and usage.

718. KPIs for Finance Department

Key Performance Indicators (KPIs) are important for measuring the effectiveness and success of any department, including Finance. Here are some KPIs that a Finance department may use:

Cash Flow: This metric measures the cash inflows and outflows of the company.

Revenue Growth: This metric measures the rate of growth in revenue over a given period.

Gross Profit Margin: This metric measures the percentage of revenue that remains after the cost of goods sold is subtracted.

Operating Expenses Ratio: This metric measures the percentage of revenue that is spent on operating expenses.

Return on Investment (ROI): This metric measures the return on investment for a particular project, asset, or business unit.

Debt-to-Equity Ratio: This metric measures the amount of debt relative to equity on the company's balance sheet.

Accounts Receivable Turnover: This metric measures the number of times accounts receivable turns over in a given period.

Days Sales Outstanding (DSO): This metric measures the average number of days it takes to collect payment from customers.

Net Profit Margin: This metric measures the percentage of revenue that remains after all expenses, including taxes, have been paid.

Budget Variance: This metric measures the difference between the actual and budgeted amounts for a particular expense or revenue item.

These KPIs can help a Finance department measure its performance, identify areas for improvement, and make data-driven decisions to optimize revenue, profitability, and financial stability.

719. KPIs for Marketing Department

Key Performance Indicators (KPIs) are important for measuring the effectiveness and success of any department, including Marketing. Here are some KPIs that a Marketing department may use:

Return on Investment (ROI): This metric measures the return on investment for a particular marketing campaign or initiative.

Customer Acquisition Cost (CAC): This metric measures the cost of acquiring a new customer.

Cost per Lead (CPL): This metric measures the cost of generating a new sales lead.

Conversion Rate: This metric measures the percentage of website visitors who take a desired action, such as making a purchase or filling out a form.

Website Traffic: This metric measures the number of website visitors over a given period.

Social Media Engagement: This metric measures the level of engagement on social media channels, such as likes, shares, and comments.

Brand Awareness: This metric measures the level of awareness and recognition of the company's brand.

Customer Lifetime Value (CLV): This metric measures the total value of a customer over the lifetime of their relationship with the company.

Net Promoter Score (NPS): This metric measures the likelihood that customers would recommend the company to others.

Marketing Qualified Leads (MQLs): This metric measures the number of leads that are likely to become customers based on their engagement with marketing campaigns.

These KPIs can help a Marketing department measure its performance, identify areas for improvement, and make data-driven decisions to optimize customer acquisition, engagement, and loyalty.

720. KPIs for Pharmaceutical client

Key Performance Indicators (KPIs) are important for measuring the effectiveness and success of any industry, including the pharmaceutical industry. Here are some KPIs that a pharmaceutical company may use:

Sales Growth: This metric measures the rate of growth in sales over a given period.

Return on Investment (ROI): This metric measures the return on investment for a particular drug or product.

Research and Development (R&D) Spending: This metric measures the percentage of revenue spent on R&D for the development of new drugs.

Time to Market: This metric measures the time it takes to bring a new drug to market from the beginning of the research and development process.

New Product Pipeline: This metric measures the number of new drugs in the pipeline for future release.

Market Share: This metric measures the percentage of total sales within a specific market segment.

Patient Adherence: This metric measures the percentage of patients who adhere to the prescribed treatment regimen.

Product Recall Rate: This metric measures the number of products that need to be recalled due to safety or quality concerns.

Manufacturing Efficiency: This metric measures the efficiency of the manufacturing process for producing drugs.

Employee Turnover: This metric measures the percentage of employees who leave the company within a given period.

These KPIs can help a pharmaceutical company measure its performance, identify areas for improvement, and make data-driven decisions to optimize revenue, profitability, and patient outcomes.

721. KPIs for Insurance client

Key Performance Indicators (KPIs) are important for measuring the effectiveness and success of any industry, including the insurance industry. Here are some KPIs that an insurance company may use:

Loss Ratio: This metric measures the ratio of losses incurred to premiums earned.

Combined Ratio: This metric measures the sum of the loss ratio and the expense ratio, which represents the total cost of underwriting and operating expenses.

Claims Processing Time: This metric measures the time it takes to process and settle insurance claims.

Customer Retention Rate: This metric measures the percentage of customers who renew their insurance policies.

Customer Satisfaction Score (CSAT): This metric measures the level of satisfaction of customers with the insurance products and services.

Premium Growth Rate: This metric measures the rate of growth in premiums over a given period.

Net Promoter Score (NPS): This metric measures the likelihood that customers would recommend the insurance company to others.

Loss Adjustment Expense Ratio: This metric measures the expenses related to adjusting and settling claims as a percentage of earned premiums.

Underwriting Profit Margin: This metric measures the percentage of premium revenue that remains after underwriting expenses are deducted.

Policy Cancellation Rate: This metric measures the percentage of policies cancelled by customers or the insurance company.

These KPIs can help an insurance company measure its performance, identify areas for improvement, and make data-driven decisions to optimize revenue, profitability, and customer satisfaction.

722. KPIs for Football Match

Key Performance Indicators (KPIs) for a football match can vary depending on the objectives and strategies of the team, but some common KPIs for individual players and the team as a whole could include:

Shots on goal: This measures the number of attempts a team or player makes to score a goal. It helps assess how well a team is creating and taking opportunities to score.

Pass completion rate: This measures the percentage of passes that a team or player successfully completes. It helps evaluate how well a team is able to maintain possession of the ball and move it around the field.

Tackles won: This measures the number of successful tackles made by a team or player. It helps assess how well a team is defending and preventing the opposition from advancing.

Possession: This measures the amount of time a team has possession of the ball compared to their opponents. It helps evaluate how well a team is controlling the game and dictating the pace.

Dribbles completed: This measures the number of successful dribbles by a player. It helps assess the ability of a player to take on and beat defenders.

Interceptions: This measures the number of times a player or team intercepts the opposition's passes. It helps assess how well a team is disrupting the opposition's play.

Corners won: This measures the number of corners a team earns. It helps evaluate how well a team is creating chances and putting pressure on the opposition's defense.

Clean sheets: This measures the number of times a team keeps a clean sheet, meaning they don't concede any goals. It helps assess how well a team is defending and preventing the opposition from scoring.

These KPIs can be tracked and analyzed to provide insights into the performance of individual players and the team as a whole, and can be used to identify areas for improvement and develop strategies for future matches.

723. KPIs for a Chess Match

Key Performance Indicators (KPIs) for a chess match can vary depending on the objectives and strategies of the player, but some common KPIs for individual players could include:

Accuracy: This measures the percentage of moves made by a player that are considered optimal or near-optimal based on analysis. It helps evaluate how well a player is playing and making the best possible decisions in the game.

Material advantage: This measures the value of the pieces a player has captured or retained compared to their opponent. It helps assess how well a player is managing the game and gaining an advantage.

Control of the center: This measures the number of pieces a player has in the center of the board. It helps evaluate how well a player is controlling the board and positioning their pieces for success.

Development: This measures the number of pieces a player has developed from their starting positions. It helps assess how well a player is progressing in the game and making use of their resources.

Tempo: This measures the number of moves a player has made compared to their opponent. It helps assess how well a player is managing the pace of the game and putting pressure on their opponent.

Time management: This measures how efficiently a player is using their time during the game. It helps assess how well a player is managing their time and making the best use of it to achieve their objectives.

These KPIs can be tracked and analyzed to provide insights into the performance of individual players and their strategies in the game, and can be used to identify areas for improvement and develop strategies for future matches.

724. Company A, a technology company that specializes in mobile app development, was looking to expand its market share in the gaming industry. After conducting research and analyzing potential acquisition targets, Company A identified Company B, a smaller mobile game development company, as a potential acquisition target.

Company B had a portfolio of successful mobile games and a loyal user base, which made it an attractive target for Company A. The two companies entered into negotiations, and after several rounds of discussions, Company A acquired Company B for \$10 million.

Following the acquisition, Company A integrated Company B's portfolio of mobile games into its existing app development business. This enabled Company A to diversify its product offerings and expand its presence in the gaming industry.

To ensure a smooth transition, Company A retained key members of Company B's development team, who were instrumental in the creation of the acquired mobile games. Company A also invested in marketing and promotional campaigns to increase awareness of the newly acquired mobile games.

The acquisition was successful, with Company A achieving significant growth in its mobile gaming business. The integration of Company B's mobile games helped Company A expand its user base and increase revenue. Additionally, the acquisition allowed Company A to enter new markets and gain a competitive advantage in the mobile gaming industry.

Overall, the acquisition of Company B was a strategic move that enabled Company A to achieve its goal of expanding its market share in the gaming industry. The acquisition allowed Company A to acquire valuable assets and resources, which it could leverage to drive growth and improve its competitiveness in the market.

725. KPIs for Mobile Game

Key Performance Indicators (KPIs) for a mobile game can vary depending on the objectives and strategies of the game developer, but some common KPIs for mobile games could include:

Downloads: This measures the number of times the game has been downloaded from app stores. It helps evaluate the game's popularity and the effectiveness of marketing and promotion efforts.

Retention rate: This measures the percentage of users who continue to play the game after a certain period of time, such as a week or a month. It helps evaluate how engaging the game is and how well it retains players.

Daily active users (DAU): This measures the number of unique users who play the game on a daily basis. It helps evaluate the game's daily engagement and popularity.

Monthly active users (MAU): This measures the number of unique users who play the game in a month. It helps evaluate the game's monthly engagement and popularity.

Revenue: This measures the amount of money the game generates from in-app purchases, subscriptions, or advertising. It helps evaluate the game's monetization strategy and profitability.

Average revenue per user (ARPU): This measures the average amount of revenue generated per user. It helps evaluate the game's monetization strategy and the effectiveness of in-app purchases and other revenue streams.

Session length: This measures the amount of time users spend playing the game in a single session. It helps evaluate how engaging the game is and how well it retains players.

Churn rate: This measures the percentage of users who stop playing the game over a certain period of time, such as a week or a month. It helps evaluate how well the game retains players and the effectiveness of retention strategies.

These KPIs can be tracked and analyzed to provide insights into the performance of the mobile game and its engagement with users. They can be used to identify areas for improvement and develop strategies for increasing user retention, monetization, and overall success of the game.

726. KPIs for Electric Vehicle

Key Performance Indicators (KPIs) for an electric vehicle can vary depending on the objectives and strategies of the manufacturer, but some common KPIs for electric vehicles could include:

Range: This measures the distance that an electric vehicle can travel on a single charge. It helps evaluate the vehicle's performance and suitability for different use cases, such as daily commuting or long-distance travel.

Battery efficiency: This measures how efficiently the battery of the electric vehicle is used to power the vehicle. It helps evaluate the energy efficiency of the vehicle and its ability to maximize range.

Charging time: This measures the time it takes to charge an electric vehicle's battery from empty to full. It helps evaluate the convenience and practicality of the vehicle for daily use and long-distance travel.

Charging infrastructure: This measures the availability and accessibility of charging stations for electric vehicles. It helps evaluate the convenience and practicality of owning and using an electric vehicle.

Sales: This measures the number of electric vehicles sold over a certain period of time. It helps evaluate the popularity and market demand for the vehicle.

Market share: This measures the percentage of the total market for vehicles that is occupied by electric vehicles. It helps evaluate the competitiveness of the electric vehicle industry and the potential for growth.

The total cost of ownership: This measures the total cost of owning and operating an electric vehicle over its lifetime, including purchase price, maintenance costs, and fuel/charging costs. It

helps evaluate the economic feasibility of owning and using an electric vehicle compared to traditional gasoline-powered vehicles.

These KPIs can be tracked and analyzed to provide insights into the performance of electric vehicles and the electric vehicle industry. They can be used to identify areas for improvement and develop strategies for increasing market adoption and profitability.

727. KPIs for stores

Key Performance Indicators (KPIs) for stores can vary depending on the type of store, but some common KPIs for stores could include:

Sales: This measures the amount of revenue generated by the store over a certain period of time. It helps evaluate the store's overall performance and revenue growth.

Conversion rate: This measures the percentage of visitors to the store who make a purchase. It helps evaluate the effectiveness of the store's merchandising, pricing, and sales strategies.

Average transaction value: This measures the average amount spent by customers per transaction. It helps evaluate the store's ability to upsell and cross-sell products and services.

Foot traffic: This measures the number of people who visit the store over a certain period of time. It helps evaluate the store's visibility and popularity.

Inventory turnover: This measures the rate at which the store sells its inventory over a certain period of time. It helps evaluate the efficiency of the store's inventory management and purchasing strategies.

Gross margin: This measures the difference between the store's revenue and the cost of goods sold. It helps evaluate the store's profitability and pricing strategy.

Customer satisfaction: This measures the satisfaction level of customers with the store's products, services, and overall experience. It helps evaluate the store's customer service and reputation.

These KPIs can be tracked and analyzed to provide insights into the performance of the store and to identify areas for improvement. They can be used to develop strategies for increasing revenue, customer loyalty, and overall success of the store.

728. How OTT platform earns money?

Over-the-top (OTT) platforms are streaming services that provide content directly to users over

the internet, bypassing traditional cable and satellite TV providers. There are several ways that OTT platforms earn money:

Subscription fees: Many OTT platforms charge a monthly or annual subscription fee for access to their content. This is a primary source of revenue for many platforms, and the fees can vary depending on the platform and the level of access to content.

Advertising: Some OTT platforms, such as Hulu and Peacock, offer both subscription and advertising-based models. Ad-supported models are typically free for users, and the platform earns revenue from advertisers paying for ad space.

Transactional Video on Demand (TVOD): OTT platforms may also offer a TVOD model, where users can pay to rent or purchase individual movies or episodes of TV shows.

Original content: Many OTT platforms invest in producing their own original content, which can attract subscribers and generate revenue through licensing deals, merchandising, and other forms of ancillary revenue.

Partnerships and licensing deals: OTT platforms can generate revenue through partnerships with other companies, such as providing content to other platforms or licensing content to other companies.

Data and analytics: OTT platforms can also generate revenue by collecting and selling data on user behavior and preferences to advertisers and other companies.

Overall, OTT platforms have multiple revenue streams, including subscription fees, advertising, TVOD, original content, partnerships, and data and analytics. Each platform's revenue model may differ depending on its business strategy, audience, and content offerings.

729. KPI for Meesho

Meesho is a social commerce platform that allows individuals and small businesses to sell products through social media channels like Facebook and Instagram. Some possible KPIs (key performance indicators) for Meesho could include:

Gross Merchandise Value (GMV): This is the total value of all products sold through the Meesho platform. It is a critical KPI for any e-commerce company as it directly impacts revenue and profitability.

Customer Acquisition Cost (CAC): This is the amount of money Meesho spends to acquire a new customer. A low CAC is critical for long-term profitability.

Average Order Value (AOV): This is the average value of each order placed on the Meesho platform. A higher AOV can increase revenue and profitability.

Conversion Rate: This is the percentage of people who visit the Meesho website or app and end up making a purchase. Improving the conversion rate can increase revenue and profitability.

Active Users: This is the number of people who regularly use the Meesho platform to buy or sell products. Increasing the number of active users can drive revenue growth.

Return on Investment (ROI): This is the ratio of the revenue generated by Meesho to the amount of money invested in the company. A high ROI is critical for long-term profitability.

Customer Retention Rate: This is the percentage of customers who continue to use the Meesho platform over time. A high retention rate is important for long-term revenue growth and profitability.

These are just a few examples of KPIs that Meesho could use to measure its performance. The specific KPIs that are most important for Meesho will depend on its business goals and objectives.

Guesstimates

Firstly, let us list down what stuff might be relevant for the interviewer to give you a good evaluation, in decreasing order of priority.

1. A Structured approach
2. What general numbers should I remember?
3. Math: Numbers galore
4. Handy Formulae
5. Miscellaneous

1. A Structured approach

So, the torture has started. You have been asked a question which doesn't seem to have any relevance in ANY Universe. What do you think should ensue? Chaos, Panic ! Tears roll down your eyes, piss down your pants. Well, this is exactly what must be avoided. So, some ways to add method to the madness, are as follows:

Clarify, Clarify Clarify !!!! " The first step to solving a problem is to know it." So it is essential to have a COMPLETE picture of what you need to estimate

Contraceptive Example:

Curious Interviewer: "What is the number of people using contraception in a night?"

Much more Curious Interviewee: "I have the following clarifying questions" (asks them one by one).

"Which month is this night in?" This matters, since; marriages in India are concentrated in December, due to religious reasons, ultimately leading to an increased number of contraceptive users.

"Is the required number to be calculated for the World or India?" The interviewer might have purposefully left out the fact that the question was India specific.

"What are the different contraceptives?" No matter how sincerely you had taken your sex education class in school, you might want to clarify this point too.

Try having 'n' methods to estimate:

So, let's say, you have started using a particular approach and midway into the discussion you realize that a different approach will give a much better estimate.

What do you do? You are faced with the Sunk Cost Fallacy and hence continue with your inefficient approach.

So, the solution to this is fairly obvious. Have a list of 'n' approaches upfront and subsequently choose one which seems most apt.

Contraceptive Example: You could estimate this from the Demand side (number of people who WISH to have protected sex) or from the Supply Side (number of people who CAN have protected sex). Now, in rural populations, the Supply<Demand, and in urban populations,

Supply=Demand. So, the calculations for urban populations can be done from both the Supply and the Demand side. However, for rural populations the Supply side will give the right estimate. Go Old School:

Write a formula ! : You want to find a number which is basically a combination of other numbers. So, write the relationship between your required number and the other numbers (basically, write a formula). Now, all that is left is finding the numbers on the RHS independently and Presto! You have your guesstimate.

Contraceptive Example:

Going from the Demand Side. "Number of contraceptives used = Number of couples having sex per night*Fraction having protected sex* Number of contraceptives used per couple per night"

Backward Traceability: The idea is to write the calculations; the tree diagrams etc. in a chronological manner such that if at any point in time, you want to go back and check your calculations or approach you can do it without any fuss.

2. Awareness in General:

What general numbers should I remember?

The following numbers can be memorized for your country (here India) *Only ballpark figures are mentioned

GDP = 1.8 trillion USD

Population = 1.2 billion ~ 1 billion

Land Area = 3 million km²

GDP growth rate = 5%

Average size of a family = 3.6 ~ 4

Number of households = 330 million ~ 300 million Population growth rate = 1.5 % (World = 1%)

Sex ratio = 1:1

Rural: Urban population = 70:30

Population Distribution by Age:

India has a young population. It has more than 50% of its population below the age of 25.

0-15: 30 %

15-25 : 20 %

25-50 : 30 %

50+ : 20%

Population Distribution by Income:

Upper Middle Class (>32,000 pm): 10% Middle Class (16,000-32,000 pm): 30% Lower Middle

Class (8000-16,000 pm): 40% Below poverty line (<8000pm): 20%

Mumbai population = 20 million

Kolkata, Delhi population (Take Approx same for all metros) =15 million

3. Math: Numbers galore

Number of Zeros : 1 lakh = 10^5 , 1 million = 10^6 , 1 crore = 10^7 , 1 billion = 10^9 , 1 trillion = 10^{12}

Percentages: Situations arise when you have to multiply percentages. So it is good to have this well practised. Example: In a population 80% males and 60% females wear watches. Then, assuming a 1:1 sex ratio we get $80\% * 50\% + 60\% * 50\% = 40\% + 30\% = 70\%$ of the population wears watches.

4. Handy Formulae

Market Size: Estimating the market size would basically mean, how many new products will be required in the next year.

$$\begin{aligned}\# \text{ Products required per year} &= \# \text{ existing products that get obsolete} + \# \text{ new products required} \\ &= Q/n + r*Q\end{aligned}$$

Where, Q = existing number of products in the market

n = average age of the product

r = average growth rate of the product~ GDP growth rate of the country (5% for India)

Example: "What is the market size of squash rackets in India?" The average age of a racket ($n=1.5$ yrs), average growth rate for the racket ($r= 5\%$) and the number of existing rackets in India ($Q = 1$ million has to be found by guesstimation). Market size = $1 \text{ million} * (1/1.5 + 0.05) = 0.72 \text{ million}$

Occupancy: This is valid for any situation in which there are a particular number of places and a partial number of them are occupied. Thus, like a bus, theatre, stadium etc. have 'n' seats and a fraction of them are occupied.

Example : Avg occupancy of a particular bus is 70%, then if there are 100 seats, at any point in time on an average 70 seats will be occupied.

5. Miscellaneous

Example : "Q: What is the market of roses in India?". Here, you must think that roses are not just sold as a flower but also is a raw material for the production of rose water. Hence, it is important to include this hidden application in your guesstimate.

Example: During the course of investigating any costs for a guesstimate of total costs, you might encounter a situation like, what is the cost of potatoes per Kg and you have no freaking idea.

Solution: Estimate the weight of a samosa (30g) and the cost of the cheapest one that you have eaten (say, Rs 5) and assume a % of this samosa's cost which would come from the potato (say, 20% : This number is low because oil is an essential component in samsosa making which is definitely expensive) . Thus, the cost of potatoes in Rs per Kg then is = $(5 * 20\%) / (0.030) \sim 35$.

730.Number of Maggi sold in a day in India

I took a bottom-up approach.

Considering an ordinary, urban household with 4 individuals Number of Maggi needed per month = 10

Therefore, per head consumption = $(10/4) = 2.5$ Maggi per person

Population = 1.3 billion

Urban population: 70% of total population

Above poverty line population: 40% of total population

Therefore, net population to consider: $1300 \times 0.7 \times 0.4 = 364$ million.

Population distribution: (Age-wise)

0 – 10 (consume less than 2.5 packets per month, say 2 packets): 20% of the population

{which equals to $(364 \times 0.2 \times 2)$ million packets per month = 145.6 million packets per month}

10 – 60 (consume 3 packets per month): 65% of the population {which equals to $(364 \times 0.65 \times 3)$ million kg per month = 709.8 million packets per month}

60+ (consume less than 2.5 packets per month, 2 packets): 15%

{which equals to $(364 \times 0.15 \times 2)$ million packets per month = 109.2 million packets per month}

Total approximate consumption = $(145.6 + 709.8 + 109.2)$ million packets/month = 964.6 million packets/month

Assuming a month of 30 days, per day consumption = $(964.6 / 30)$ million packets per day = 32.15 million packets per day.

731. How many t-shirts e-commerce companies selling in India per day?

We can approach this problem in two ways:

Demand side

Supply side

I am going to solve using demand of t-shirts in the market

Total population of india : 1 bn (approx)

Reach to internet : 40% = 400 Mn

Reach of ecommerce companies to deliver products : $3/4$ th = 300Mn Let's assume 50% are male and 50% are female

Lets solve for male population first:

Now i have divided males in the four categories on the basis of age because demand of t-shirts for different age groups will be different

0–15 yr = 45 Mn, on an average, individual own 4 t shirts -> $4 \times 45 = 180$ Mn 16–22 yr = 23 Mn, on an average individual own 4 t shirts -> $4 \times 23 = 92$ Mn 23–50 yr = 65 Mn, on an average individual own 3 t shirts -> $3 \times 65 = 195$ Mn

50 - 80 yr = 18 Mn, on an average individual have 2 t shirts -> $2 \times 18 = 36$ Mn

Total t shirts own by men : $180 + 92 + 195 + 36 = 503$ Mn ~ 500 Mn

Let's solve for female population now:

0–15 yr = 45 Mn, on an average individual own 2 t shirts -> $2 \times 45 = 90$ Mn 16–22 yr = 23 Mn, on an average individual own 4 t shirts -> $4 \times 23 = 92$ Mn 23–30 yr = 15 Mn, on an average individual own 3 t shirts -> $3 \times 15 = 45$ Mn 30 - 80 yr = 67 Mn -> we can neglect this section. Only few ladies prefer to use t-shirts in this age group.

Total t shirts own by females : $90 + 92 + 45 = 227$ Mn ~ 230 Mn

Total t shirts own by men + women = $500 + 230 = 730$ Mn

Average life of a t shirt = 2 year

Demand per year = 365 Mn ~ 360 Mn

Online portals provide coupons and offers but because of trust factor and fitting issues, people in India still prefer to buy offline. So I am assuming 30% of people buy t-shirt from ecommerce portal and 70% are buying from market.

Total number of t-shirts sold through ecommerce platform per year in India = $.3 * 360 = 108 \text{ Mn} \sim 100 \text{ Mn}$ per year

Number of t-shirts sold in India per day (From ecommerce portal) = $100 * 10^6 / 365 \sim 27,000$

732. What is the number of laptops sold in Bangalore on an average routine day?

Laptop is a costly product. I am assuming that people buy laptop only when they needed. That's why I am going to calculate potential market of laptops in India.

Total population of Bangalore = 18Mn ~ 20Mn

Let's divide population on the basis of age

0–18 Yr - 30% of 20 Mn = 6 Mn -> We can neglect this age group because generally they don't need personal laptop and when needed, they prefer to use others laptop.

19–22 Yr - 10% of 20 Mn = 2Mn -> $0.6 * 2 = 1.2 \text{ Mn}$ (This is the college age group. Most of the college students need a laptop. Assumed 60% of them own a laptop)

22–50 Yr = 40% of 20 Mn = 8 Mn. 22–50 age group is the working class of the society. I have divided this class into 3 major categories.

White collar employees (25%)

Blue collar employees (50%)

Small business owners (25%)

Assumed 80% and 30% people in the category of white collar employees and Small business owners respectively own a laptop or PC. We can neglect blue collar employees.

80% white collar own a laptop or PC -> 1.6 Mn Small business owners own laptops or PC -> 0.6 Mn

50–80 Yr = 20% = 4 Mn -> we can ignore this age group

Total laptop + PC users in Bangalore = $1.2 + 1.6 + 0.6 = 2.4 \text{ Mn}$

Corporate offices/Schools/Computer centers generally have desktop. Lets assume 60% are desktops.

Laptops = 40% -> 0.9 Mn

Average life of a laptop = 5 years (in India)

Number of sold per day in Bangalore = $0.9 \text{ Mn} / 365 * 5 \sim 500 \text{ laptops}$

733. What are the number of smart phones sold in India per year?

Population of India : 1200 mn

Population above poverty line: 70% 840 mn

Population below 14 years: 30%

Hence, proxy figure: 588 mn

Rural Population (70%) : 410 mn

Rural Households: 82 Mn

Rural Mobile Penetration: Avg 2 per household- 164 Mn

In rural areas assume that a new mobile is bought once in 3 years. Hence, new mobiles bought
In current year- 55 Mn
Urban (30%) :176 Mn
Assume Avg No of Mobiles per person : 1.5
Urban Mobile Penetration: 265 Mn
Assuming that a new mobile is bought once in 1.5 years. Hence new mobiles in current year-
176 Mn
Total New Mobiles: 231 mn
Assuming 3 out of 10 new mobiles are smart phones
No. of smart phones sold=70 Mn

734. What is the total number of people who get a new job in India in a year?

Observations:

35 million students enroll in India(Undergraduate, graduate, doctorate, diploma)
72% of 35 million graduate every year = 25 million
Students completing 10th grade = 20 million
Students completing 12th grade= 15 million Unemployed graduates of the previous year= 15 million (Since 60% of 25 million graduates are unemployed) GDP growth rate is 7%

Calculations:

40% of 25 million graduates are only employed= 10 million
Assuming 500,000 of the previous year's graduates get a new job 100,000 starts working after 12th grade due to poverty, poor grades etc An estimate of 50,000 starts working after 10th grade due to poverty,poor grades etc
10,000 people already on workforce end up with a new job

Total= 10 million + 500,000 + 100,000 + 50,000 + 10,000
= 10.66 million (approx) Note:

Migrants working in India are negligible
Due to urbanization, very few go for work without completing their 10th grade
Increased feminism has a significant effect on the estimates

735. How many red colored Swift cars are there in Delhi?

The approach to such problems follows a MECE approach. MECE expands to Mutually Exclusive Collectively Exhaustive, which trivially means breaking your problem down to Non-overlapping segments which add up together to give your final solution.

Let's solve the guesstimate Population of Delhi: 20 Mn

Children or college going = 20% of 20 Mn -> 4 Mn

Senior citizens = 20% of 20 Mn -> 4 Mn

Working people = 60% of 20 Mn -> 12 Mn

let there are 5 brands of car and each brand have 10 cars which are equally distributed. So in total, we have 50 models of cars running in the streets. This does not include luxury cars.

Working class people, let's assume half are married and half remain unmarried. So married -> 6 Mn and unmarried -> 6 Mn

Married couples:-

Number of married couples = 6 Mn/2 -> 3 Mn

I am assuming 10% belong to the rich class and prefer luxury cars and 20% cannot afford a car. The rest 70% has one car each.

70% of 3 Mn = 2.1 Mn

There is the equal distribution of above mentioned 50 cars among these 2.1 couples again. So the number of Swift Cars right now is 2.1 Mn / 50 = 0.042 Mn. I am assuming Swift car comes in 10 colors. Hence number of red swift cars in married couples is 0.0042 Mn -> 42,000

Unmarried couples:-

Out of 6 Mn unmarried couples, Only 10% can afford mid range non luxury cars. Hence no of cars = 6 lakh. These are again divided into 50 models as above and each model has 10 colors. So number of red colored swift cars among unmarried people = 6 lakh / 500 -> 12,000

Senior citizens

Out of 2 Mn families(4 Mn people), 20% i.e. 0.4 Mn families own a car. Again, as above, these cars are divided into 50 models with each model having 10 colors. So 4 lakh/500 -> 8,000

Total number of red colored swift cars in Delhi = 42,000 + 12,000 + 8,000 - > 62,000

736. How many paan shops are there in India?

We will approach this problem by solving both demand and supply sides First i will calculate the demand for the Pans in India

Total population of India = 1.2 bn or 1200 Mn (approx)

Males = 700 Mn

Female = 600 Mn (900 women per 1000 men)

Ratio of women consuming pan in India on the regular basis is very small. Let's assume 2% of total women population consume pan on the regular basis

Number of pans consumed by females = 2% of total female population = 600 Mn*.02 -> 12 Mn

Let's solve for male population now:

Consumption of pans varies with the age group.

For example, Old person will consume less pans compare to young guys because of health issues. I have divided males on the basis of age group

0–15 -> we can neglect this section. Only few boys below age 15 consume pans

16–22 -> 15% of 700 Mn = 105 Mn -> ~ 5 Mn (16-22 is the college age & as per my personal experience in the college, Avg. 5 students consume pan out of the 100)

23–50 -> 35% of 700 Mn = 240 Mn -> 20% people consume pan -> 0.2*240 -> 48 Mn

51–80 -> 20% of 700 Mn =140 Mn -> .05*140 -> 7 Mn (In the older age, because of health issues, few people consume pan. Here assumed 5% will consume pan

Total demand of pans in India per day = 12 Mn + 5 Mn + 48 Mn + 7Mn = 72 Mn

Let's calculate the supply of pans per day

Time taken to ready 1 pan = 2 mins -> 30 Pans in 1 hour

Let's assume pan shop open for 10 hours per day -> 30*10 = 300 pans per day

Number of pan shops in india = 72 Mn/ 300 = 2,40,000 ~ 2.5 lakh

737. What is the maximum number of human beings that can survive on the earth?

Assumptions

Total usable land for human on earth is the 30% of total area

Average number of persons living in a house = 4

Average house includes 2 rooms, 1 washroom, 1 bathroom and 1 kitchen Schools / hospitals / Roads / Railway line and other public utilities cover 25% of total land area

Diameter of earth = 13,000 km -> land area = 30% of total area = $.3 \times 3.14 \times (13,000/2)^2 = 4 \times 10^8 \text{ km}^2$

Total land area = $4 \times 10^8 \text{ km}^2$

Area of 1 room = $5 \times 5 \text{ m}^2$ -> Total area of two rooms = 50 m^2 Area of 1 washroom = $1 \times 1 = 1 \text{ m}^2$

Area of 1 bathroom = $2 \times 2 = 4 \text{ m}^2$

Area of 1 kitchen = $2 \times 2 = 4 \text{ m}^2$

Total area of 1 house = 60 m^2

Food will be required for the survival, hence each family will need land for raw food material production. Based on the experience, 300 m^2 is the

sufficient land to grow raw food material like wheat, rice etc for the survival of 4 persons

Total land required per family (House + food production) = 360 m^2

25% of total useful land will be used by public utilities like schools / hospitals etc = $0.25 \times 4 \times 10^8 = 10^8 \text{ km}^2$

Land left = $3 \times 10^8 \text{ km}^2$

Total number of families those can survive on earth = $3 \times 10^8 \times 10^6 \text{ m}^2 / 360 \text{ m}^2 = 8.33 \times 10^{11}$

Person per family = 4

Total number of persons those can survive on the earth = $4 \times 8.33 \times 10^{11} = 3.3 \text{ trillion}$

738. How to calculate the number of taxis operating in Mumbai per day?

Here, Taxi means the yellow white one. Not ola/uber etc.

Solution is dependent on two things . First is the population i.e. how many person are willing to use taxi per day which will directly give the total trips required per day to satisfy the population and second from the supply side i.e. Number of trips per taxi per day.

Demand of taxis

Step 1 - Population of Mumbai

552 MP's for 1.26 Billion people = 2.3 million people/ MP

MP's elected from Mumbai + Thane + Kalyan (your choice of city) = 8 Population of Mumbai = $8 \times 2.3 = 18.4 \text{ million} = 1.8 \text{ Crore} \sim 2 \text{ Crore}$

Step 2: No of ways one could commute in Mumbai 1) Walk

2) Bike/bicycle

3) Auto

4) Car

5) Taxi

6) Bus

7) Local Train

Step 3 : People preference for each transportation Considering population distribution in Mumbai as below: 1) Upper Class: 10%

2) Middle Class: 70%

3) Lower Class: 20%

Preference distribution in each class will dependent on affordability, distance to travel and convenience

1) Upper Class (10%) = 7% Cars + 2% super bikes + 1% taxi(others)

2) Middle Class (70%) = 0.5% Walk + 10% Bike + 6% Car + 8% Taxi + 3.5% Auto + 17% Bus + 25% Local

3) Lower Class (20%) = 0.5% Walk + 12% Local + 7% Bus + 0.5% Bike(Others)

Preference distribution:

1) Walk : 1%

2) Bike/bicycle: 12.5%

3) Auto: 3.5%

4) Car: 13%

5) Taxi: 9%

6) Bus: 24%

7) Local Train: 37%

Step 4: Population traveling in the taxis

9% of population travels in taxi hence resulting in 18 Lakh people. 18 lakh people can travel alone or in group of 4 at max. Let's assume avg 3 people per trip, we arrive at 6 lakh taxi trips per day.

Supply of taxis

Taxi's are normally used for trips which range between 5km - 15 kms of travel. Considering traffic signals, waiting time (at peek hours) and travel time, One an average, one ends up doing 10 trips (pickup and drop) a day.

Therefore, (6lakh trips)/(10trips per taxi) = 60,000 Taxi's

739. How to calculate the number of weddings in India per year?

India population ~ 1.2 billion

Rural - 72% and Urban - 28% (Original stats)

Assumption

Rural areas, the age of marriage (in average) is between 15 - 35 years and in urban areas = 20 - 35 years

India is a young country. 0 - 35 years has around 65% of the total population. Assuming that population is uniformly distributed.

Rural population

So, the percentage of people in rural areas fit for marriage is $(35-15)/35 * 65 = 40\%$

Rural population fit for marriage is = $0.72 \cdot 40 \cdot 1.2$ billion = 345.6 million Assuming the sex ratio as 50% male and 50% female (its close to 49%) Number of marriages in rural areas = 345.6 mill/2 = 172.8 million

If only one marriage per women, number is $172.8 / 20 = 8.64$ million

Urban population

Urban area population fit for marriage = $0.28 \cdot 30 \cdot 1.2$ billion = 100 million Assuming the sex ratio as 50% male and 50% female (its close to 49%) Number of marriages in urban areas = 100 mill /2 = 50 million

If only one marriage per women, number is $50 / 15 = 3.3$ million

Total number of marriages in India per year = 8.64 Mn + 3.3 Mn ~ 12 Million

As per resources, there are nearly 10 Million marriages happen each year in India. Our answer is close to the actual numbers.

740. How many KGs of paint are used in the USA annually?

The assumption here is we are calculating the paint used to coat buildings.

Population of US= 320 million

Types of Buildings can be classified broadly as:

1. Homes
2. Large offices
3. Schools + Universities
4. Commercial buildings (restaurants, DMVs etc.)

Now let us estimate the paint for each category. Once we establish the approach in one category we should be able to replicate it across others.

Some basic metrics & assumptions needed for the calculation :

- A. Population of USA = 320 million
- B. Amount of paint needed per 500 sqft: 1 gallon (we will convert to KG at the end)
- C. The sqft. that needs to be painted for each building will be 4 times the floor-sqft, considering the walls to be painted on 4 sides, inside & out, minus windows. This is an approximation.

Number of buildings of each type:

1. Homes = There is one home per every 4 people on an avg i.e. 80 million homes
2. Offices = There is one large office building per 500 people on avg = 0.64 million large offices
3. Schools + Universities = for every 10000 ppl there are 3 schools (elementary, middle and high) + 1 university i.e. 4 buildings per every 10000 people i.e a total of $320 \text{ mil} / 10000 = 32000 * 4 = 128000$ buildings
4. Similarly, Commercial buildings = 10 buildings for every 10000 ppl on an avg i.e. $32000 * 10 = 320000$ buildings

Now let us estimate the sqft. to be painted for each category: (review the assumptions & metrics we listed earlier)

1. Homes get painted once in 10 years, so apprx. 8 million homes get painted every year.

10% homes are very large i.e. 0.8 mil = 5000 sqft

10% homes are small i.e. 0.8 mil= 600 sqft

Remaining i.e. 6.4 mil = avg 2000 sqft

Total sqft = $5000 * 0.8 + 0.8 * 600 + 2000 * 6.4 = 4000 + 480 + 12800 \text{ mil sft} = 17,280 \text{ mil sqft}$

Walls to be painted = $4 * 17280 \text{ sqft} = 69,120 \text{ mil sqft}$ to be painted

2. Offices avg 10000 sqft per building

Offices get painted once in 5 years; so 0.15 mil offices get painted every year

Office wall space = 40000 sqft per building

Walls to be painted = $40000 * 0.15 \text{ mil sqft} = 6000 \text{ mil sft}$ to be painted

3. Schools & universities ; assuming same size as office buildings Schools get painted once in 5 years; so 25600 schools get painted every year

Walls to be painted = $40000 \text{ sqft} * 25600 \text{ buildings} = 1024 \text{ mil sqft}$ to be painted

4. Commercial buildings ; assuming an average commercial building is of size 5000 sqft

Commercial buildings get painted once in 5 years; so 64000 buildings get painted every year

Total walls to be painted = $20000 * 64000 = 1280 \text{ mil sqft}$ Total of all calculations:

Total sqft to be painted = 77424 mil sqft

1 gallon of paint is good for 500 sqft. Hence total gallons needed = 154.8 million gallons or 17.2 million KGs

741. How many street lights are there in India?

Lets start with the formula we can use to estimate this:

(Area covered by streets in India) * (Percentage of street area containing street lights) * (Avg. number of street light in the area/Km²)

Let's float some numbers now:

Total area of India is: 3.28~3.3 million Km² 1. Area covered by streets: .66 million Km² (Real Estate+Farm Lands+Streets)

Assuming real estate to be most of the area followed by farm land (including barren land), I'll divide the area into 50%, 30% & 20%. So, we have total street area = $3.3 * 20\% = .66 \text{ million Km}^2$.

2. Percentage of street area covered by street lights: 30%

For this I would assume 70%-30% division of people in villages and cities.

Lets take into account 60% in tier 1 (10% of overall cities) cities, 40% in tier 2 (70% of overall cities), 10% in tier 3 (10% of overall cities) and 10% in other small towns (another 10%).

So, we have 6%, 28%, 1%, 1%. Adding them up and rounding them off to lower number owing to poor infrastructure and an increase in assumption of total area in India. We assume total as 30%.

3. Avg. number of street lights per unit area: 2100

Assuming avg Width of street containing 1 fleet of street lights is: 10 m = .010 Km

Number of Street lights per Km = 2 (Factor of two coz number of street lights at one spot)* [1000 m/ Distance between street lights (50 meters)] = 40.

So, Avg Street lights per unit area: $20 / .010 = 2000 / \text{Km}^2$

I will spike this number up by 5% considering flyovers, Multiple lane expressways making it to be: $2100 / \text{Km}^2$.

So, Putting it in formula: (.66 million Km²)*(30%)*(2100/Km²)= 415.8 million?

742. How many cups of tea were consumed in Mumbai last month?

First, clarify the question. Then, start solving.

As a first step, inform the interviewers that each day of the week is being considered equally. Tea consumption might likely decrease during the weekend as people do not go to the office—so you might consider that as well. We shall go with the first assumption.

The population of Mumbai is 18 crore; we shall round it up to 2 crores. 20% of this population is assumed to be children who do not drink tea. Another assumption is that of the remaining population, 20% are habitual drinkers, 30% are regular drinkers, 20% are occasional drinkers, and 10% are non-drinkers.

The habitual drinkers may be said to have three cups of tea in a day. Regular drinkers may be said to have one cup of tea in a day. The tea consumption of occasion drinkers maybe once a week, and that of non-drinkers none at all.

Calculating proportions-

Habitual – $3 \times 0.2 \times 7 = 4.2$

Regular – $1 \times 0.3 \times 7 = 2.1$

Occasional – $1 \times 0.2 \times 1 = 0.2$

Non – 0

Total = 6.5

Total cups per week = 6.5×1.6 crore = 10.4 crore

743. How many iPhones are currently being used in India?

Clarify with the interviewers whether the question is about only a single version of the iPhone or all versions put together. Here, we shall assume that all iPhones put together are being talked about.

The first step toward solving this query will be segmentation. There are many ways in which India's population can be segmented. Here, we shall first assume that only people who have attained a working age and are under the age of retirement own an iPhone. Children and old

citizens do not own an iPhone. This removes 20% of the population as children and 20% as senior citizens.

The next assumption will be that only the upper stratum of India's income range can afford an iPhone. This metric assumes that only 5% of the eligible citizens from the previous filter can own an iPhone.

Now, it is not necessary that every member of this upper stratum will own an iPhone. Other options, such as OnePlus, Samsung, etc., are also available. However, a fair assumption would be that 50% of the eligible population from the previous filter owns an iPhone.

Calculating the proportion of the population that owns an iPhone –

$$0.6 \times 0.05 \times 0.5 = 0.015$$

$$\text{Total iPhones in India} = 0.015 \times 130 \text{ crore} = 1.95 \text{ crore}$$

744. What is the size of the laptop market in the USA?

Make important clarifications such as the unit of measurement. Here, we shall assume that the unit of measurement is the number of laptops sold in a year.

The first step will be to make clear the USA's population, which may be taken to be 300 million. Next, assume the proportion of this population that owns a laptop. The last determination will be the average span of the life of a laptop in the USA.

The USA population may be segmented into retirees, students, stay-at-home population, and working population. The working population may be said to be 50% of the total population. The retirees maybe 30% and students maybe 20% of the whole population.

Among the working population, it is assumed everyone owns a laptop. Among the retirees, a fair assumption would be that nobody owns a laptop. Among students, while younger classes do not require laptops, older classes do. So, half the students may be assumed to own a laptop.

Calculating the proportions of the population that own a laptop –

$$\text{Working population} - 0.5 \times 1 = 0.5$$

$$\text{Students} - 0.2 \times 0.5 \times 1 = 0.1$$

$$\text{Total} - 0.6.$$

The average age of a laptop may be said to be 5 years. So, 1/5 of the total calculated population will change their laptops every year.

The market size of laptops in the USA is $350 \text{ million} \times 0.6 \times 0.2 = 42 \text{ million}$.

745. What number of tennis balls can fit inside a room?

Ans. First of all, you need to know the size of the tennis ball. You can do one of the two things: ask the interviewer or assume its size.

Now, calculate the volume of the room and divide this volume by the volume of tennis balls. You need to consider that the balls are round and a regular arrangement will leave empty space due to their shape.

Suppose, the room has only 4 seats. The room may fit 5 chairs in the vertical direction and 10 chairs in the horizontal direction. It seems as if this arrangement can be repeated 10 times to fill the room. This means that the room can roughly fill 500 seats.

The total space occupied by the seat should be considered here (sp). Here $sp = (4 \times 2 \times 1) \text{ ft} = 8 \text{ ft}$. This means that the room's volume is approximately $sp \times \text{number of seats} = 8 \times 500 = 4000 \text{ cubic ft}$.

The tennis ball seems to occupy 4 cubic inches of area, the number of balls = volume of room/area occupied by balls = 1000 balls.

Since tennis balls can be packed up to 70%, hence the total number of balls is 700.

746. How many cups of tea were consumed in Delhi in a month?

Ans. We will assume that fewer people will consume tea during the weekend since these are not working days. The next number to consider is the population.

There are 20 million people in the city and let us assume that 20% of youngsters do not consume tea. Out of the rest, 30% consume tea on a daily basis, 20% consume tea occasionally and 10% do not consume tea. Let us say that daily drinkers could be having three cups of tea in a day and occasional drinkers consume tea twice a week.

Then, the total number of cups of tea consumed will be:

Daily drinkers – $3 \times 0.2 \times 7 = 4.2$

Occasional drinkers – $1 \times 0.2 \times 1 = 0.2$

Non-drinkers= 0

Total= Daily + Occasionally + Non-drinkers = 4.4 cups in a day

Per month = $4 \times 4.4 \times 1.4 \text{ crore} = 24.64 \text{ crore cups}$.

747. How much paint will be required for painting a 20 m x 20 m wall?

Ans. Let us estimate the amount of paint required for every square meter. Now, we will find the area to be painted.

The wall to be painted will have the main area as $20 \text{ m} \times 20 \text{ m}$ which is 400 square meters.

Let us assume that the depth is 1 mm.

We will also consider that the oil in half of the paint has dried after a few hours of the paint application on the wall.

Let us consider the width of the paint to be considered as 2 mm.

Thus, the volume to be painted is 400 square meters x 0.002 meters = 0.8 meters cube of paint is required.

748. How many refrigerators are sold in India every year?

Ans. First of all, clarify whether we will consider domestically produced refrigerators or both domestically and internationally manufactured refrigerators.

Suppose we are considering both, then we will exclude segments based on a few factors.

Consider the population of India and now, divide it by the average number of members in Indian Households i.e. 4 members per household.

Now, further segment the population into urban (tier 1), suburban (tier 2) and rural (tier 3).

Classify these tiers as per availability of electricity (1.3 million Indians do not have access to electricity).

Exclude the number of people below the poverty line.

Your approximation should also include the annual demand for new refrigerators and replacements.

Consider the average life of a refrigerator (10 years) and the annual projected growth rate of refrigerators in your calculation.

749. How many cups of coffee do Americans drink in New York City per month?

Let's begin by establishing that people drink fewer cups of coffee on the weekends because they don't need the caffeine boost for their jobs. Next, we need to look up the population of NYC, which is about 9 million people.

Let's now say that 20 percent of these people are children, and you don't want to caffeinate children! So of the remaining number, 30 percent drink coffee every day, 20 percent drink coffee occasionally, and 10 percent drink tea instead.

Now, we assume that daily coffee drinkers could be drinking three cups of coffee a day, and the occasional coffee drinkers are happy with just two cups a week. Here's the formula breakdown:

Daily coffee drinkers: $3 \times 0.2 \times 7 = 4.2$

Occasional coffee drinkers: $1 \times 0.2 \times 1 = 0.2$

Tea drinkers: 0

Total: Daily + Occasionally + Tea drinkers = 4.4 cups in a day

Per month = $4 \times 4.4 \times 7.2$ million = 126,720,000 cups of coffee per month.

Chapter 6 Linear Regression

750. What is Linear Regression?

Linear Regression is the most commonly used supervised machine learning algorithm. Linear regression is used to discover a linear relationship between one dependent variable(Y) and one or more independent variables(X). Linear regression is also known as Ordinary Least-Squares(OLS). Linear regression is used to predict the future scores of the dependent variable(Y) based on the measured score of the independent variable(X) when the dependent variable(Y) is continuous such as salary, age, sales, product price, etc.

Example - If you want to predict the sales of biryani from a particular biryani outlet then your independent variables will be something like day of the week, date of the month, time of the day (people order less biryani in the morning), etc.

And your dependent variable will be number of plates of biryani per hour

751. Why is linear regression important?

The importance of Linear Regression is that it is one of the easiest to understand machine learning algorithm that can help business owners to grow by understanding the data they have the factor which help their business to grow, the factor which is contributing to the growth of the business and the other factors which is not at all helping the business to grow. So after understanding the data knowing which data is more significantly contributing to the growth and which is not contributing to the growth they can manipulate or change the data for maximum profit.

752. What are some common places where Linear Regression is used?

Linear Regression is there in almost all the prediction models. Have you ever watched a cricket match?

In every match you can observe the telecast of 'projected score', it is nothing but a very simple projection of linear regression.

The linear regression algorithm is also used for:

- Predicting the sales of Company.
- Predicting the revenue from different advertisements on the basis of the placement of advertisements
- For Insurance prediction.

753.What is the Equation of Linear Regression?

Linear Regression has a very basic formula and you can easily understand it.

Dependent variable is something which you want to predict or extrapolate.

Independent variable is something which helps you predict the dependent variable.

If you ever get confused, remember that a dependent variable is dependent on n numbers of independent variables.

Score of Indian cricket team is dependent on the performance of Dhoni, Sachin, Kohli, and 8 other players

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where, Y : Dependent Variable X : Independent Variable β_0 : Y Intercept

β_1 : Slope Coefficient ϵ : Error Term or Residual

754. What are Independent(X) and Dependent variables(Y)? Explanation of these terms:

1.Dependent Variable(Y)

In a dataset we can have one or more independent variables and only one dependent variable.

A dependent variable is a variable that is dependent on the other variable it means that it can be changed by the other variables

Example of Dependent Variable:

In a dataset, we have a variable as sales which is nothing but the overall sales of a company etc. So this sales variable is dependent on many other factors or variables like advertisement by TV, RADIO, NEWSPAPER . So we can say that the sales variable is a dependent variable.

2.Independent Variable(X)

An Independent variable is a variable that is not dependent on the other variable; it means that it cannot be changed by the other variables.

OR

The variable that is controlled throughout the experiment but is not affected by other variables is called an independent variable.

Example of Independent Variable:

If the dependent variable is sales then there are many factors or variables that will definitely affect the dependent variable such as advertisements by TV, RADIO, NEWSPAPER hence all these variables can be identified as independent variables.

755. What is β_0 OR Y Intercept?

Y intercept is nothing but the point where the function cuts or intersects the Y-axis, when the value of X = 0.

756. What is β_1 OR Slope coefficient?

Usually the slope coefficient refers to the coefficient of an independent variable(X) in a regression equation. It tells the amount of change in dependent variable(Y) that can be expected to result from a unit increase in independent variable(X).

757. What is ϵ OR Error Term OR Residual?

ϵ is nothing but the distance between the regression line and the data point.

758. What is correlation? (V.V.I.)

Correlation measures the relative strength of linear relationship between two variables Independent(X) and Dependent Variable(Y)

Correlation ranges from -1 to 1. The closer to -1, the stronger is the negative linear relationship. The closer to 1, the stronger is the positive linear relationship. The closer to 0, the weaker is the positive linear relationship.

759. What is Positive Correlation?

ANS:

Income(X) Savings(Y)

10	5
20	10
30	15
40	50
50	25

The above table shows that there are 2 columns in it Income(X) which is the Independent Variable and Savings(Y) which is the Dependent Variable. As we can see in the above table as the Income(X) increases the Savings(Y) increases this is called Positive Correlation.

760. What is Negative Correlation?

ANS:

Age(X) Eggs(Y)

2	50
4	40
6	30
8	20
10	10

In the above table we have Age of Chicken and the number of eggs lays by the chicken. As we can see in the above table as the age increases the number of Eggs lays by the chicken decreases this is called Negative Correlation.

761. What is a Simple linear regression?

ANS: Simple linear regression or Simple regression is a supervised machine learning algorithm. Simple regression as the name suggest it has only two variables in which one of them is dependent variable(Y) and other one is the independent variable(X). Simple linear regression is a technique used to discover a linear relationship between one dependent variable(Y) and one independent variable(X).

General form of Simple linear regression:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Where,

Y : Dependent Variable

X : Independent Variable

β_0 : Y Intercept

β_1 : Slope Coefficient

ϵ : Error Term or Residual

762. What is a Multiple linear regression?

ANS: Multiple linear regression or Multiple regression is a supervised machine learning algorithm. Multiple regression as the name suggest contains multiple variable in which we have only one dependent variable(Y) but multiple independent variables(X). Multiple linear regression is a technique used to discover a linear relationship between the dependent variable(Y) and independent variables(X).

General form of Simple linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_x X_x + \epsilon$$

Where,

Y : Dependent Variable

X : Independent Variable

β_0 : Y Intercept

β_1 : Slope Coefficient

ϵ : Error Term or Residual

763. What are the advantages of Linear Regression?

ANS: The advantages of Linear Regression are as follow:

- The Linear regression is easy to understand and simple.
- Easy to interpret the output.
- Linear Regression is less complex compared to other Machine learning algorithm.
- When we have a regression problem the first choice of every individual is Linear Regression.
- Linear Regression tends towards overfitting but can be reduced by applying or implementing regularization L1 and L2.

764. What are the disadvantages of Linear Regression?

ANS: The disadvantages of Linear Regression are as follow:

- The major disadvantage of Linear Regression is the assumptions of Linear Regression in many real-life scenario the assumptions are not met so in this case it is very difficult to produce a useful result.
- Underfitting occurs in Linear Regression when the model fails to fit the data properly.

765. What is Overfitting?

ANS: Overfitting is a scenario in which the model tries to fit the training data very closely but fails to fit the testing data. Overfitting occurs when the model learns each and every detail in the training data and the noise in the training data. The problem which occurs is that we try to pass the new data to the model to predict it gives a negative result. Overfitting also occurs if the model is too complex.

766. How to deal with overfitting in Linear Regression?

ANS:

- Training the model with more data.
- Cross-Validation.
- Regularization
- Data Augmentation.
- Feature Selection.
- Reducing the model complexity.

767. What is Underfitting?

ANS: Underfitting is a scenario in which the model is not able to fit the training data and the results of the testing data is also poor. Underfitting occurs when the model is not complex enough to perfectly fit the training data.

Q18. How to deal with underfitting in Linear Regression? ANS:

- Increasing the size and the number of features in the machine learning model.
- Increasing the complexity of the model.
- Get more training data.

768. What is regularisation?

ANS: Regularization is a technique which is used to solve the problem of overfitting in Linear Regression. Regularization technique is used to reduce the magnitude of the features by keeping the same number of features. Regularization works by adding a penalty term to the complex model.

769. What is Feature Selection?

ANS: Feature selection is technique which helps us choose only that feature or variable which significantly helps in contributing to the accuracy of the model. The feature which we select using feature selection technique will highly contribute to the performance of the model. In feature selection technique most of the insignificant variable or the variable which are not contributing to the accuracy of the model are eliminated which also helps in the computation cost of the model and in model performance.

Feature Selection is important because:

- Improves Accuracy
- Reduces Overfitting
- Reduces training time

770. What are the type of feature selection?

ANS: The type of feature selection are:

- Feature Importance
- Univariate Selection
- Correlation Matrix with Heatmap
- Manual Feature Selection.

771. What are the assumptions of Linear Regression?

ANS: The assumptions of linear regression are as follow:

Assumption 1: There should be no outliers.

Assumption 2: Assumption of Linearity.

Assumption 3: Assumption of Normality.

Assumption 4: Assumption of Multicollinearity.

Assumption 5: Assumption of Independence.

772. When to drop an outlier and when not to drop an outlier?

Drop an outlier: We can drop an outlier when we know that it is wrongly entered that is a data entry error. For example if we have an outlier in the age column where age is 150 which is far from the normal range and which is impossible so in that case we can remove or eliminate that is outlier. We can drop an outlier if we are having a lot of data and a very small sample of data can be dropped.

· Do not drop an outlier: When the data is critical we should not drop an outlier if we do so the results may change it will affect the accuracy of the model. We should not drop an outlier when there are lot of outlier maybe something interesting will be there in the data.

773. What is the assumption of Linearity?

ANS. Assumption of Linearity says that there should be a linear relationship between the independent and the dependent variable.

774. What is the assumption of multicollinearity?

ANS. Assumption of multicollinearity says that there should be no multicollinearity which means that the independent variable should not be highly correlated with each other.

Multicollinearity is one of the assumptions of Linear regression. Multicollinearity is a scenario in which the predictor variables or the independent variables are somehow highly correlated with each other.

775. What are the functions used to check multicollinearity in python?

ANS: The function used to check multicollinearity are:

- corr() function
- sns.heatmap() function
- variance_inflation_factor()

776. What is corr() function?

ANS: corr() is a very useful function to check the multicollinearity in the dataset. corr() gives the correlation of the independent variable. corr() returns a correlation matrix. The values in the correlation matrix are known as correlation coefficients.

SYNTAX:

```
corr_df = X.corr(method = 'pearson')
print(corr_df)
print()
```

777. What is vif?

ANS: The Variance inflation factor is used to identify the correlation between the independent variables. If the vif value is 1 it means there is no correlation, if vif is between 1 to 5 there is a correlation and if the vif value is greater than 5 then it is highly correlated and we can eliminate that variable.

Example:

Calculating vif using variance_inflation_factor() function: SYNTAX:

```
from statsmodels.stats.outliers_influence import variance_inflation_factors as vif
vif_df = pd.DataFrame()
vif_df["features"] = X.columns
vif_df["VIF Factor"] = [vif(X.values, i) for i in range(X.shape[1])]
vif_df.round(2)
```

778. How to deal with the problem of multicollinearity?

ANS: To deal with the problem of multicollinearity:

- We can remove some of the independent variables which are highly related with each other with the help of heatmap.
- We can use the variance inflation factor to identify the correlation between the independent variables. If the vif value is 1 it means there is no correlation, if vif is between 1 to 5 there is a correlation and if the vif value is greater than 5 then it is highly correlated and we can eliminate that variable.
- We can use Principle Component Analysis to eliminate the unwanted or irrelevant variables.

779.What is R square?

ANS. R square is also described as the coefficient of determination. R square is used to determine the strength of correlation between the independent and the dependent variable. In simple terms R square lets us know how accurate our regression model is when compared to average. R square ranges between 0 to 1 higher the number the better is the accuracy or prediction of the model. If our R square is greater than 70% which is 0.7 indicated a good fit model.

780. What is Adjusted R square?

ANS. The Adjusted R square is a modified version of the R square. Adding more independent variables will result in an increased value of R square irrespective of whether the new independent variable is significant or not. But in the case of Adjusted R square if the new independent variable added is insignificant the adjusted r square has the capability to decrease therefore resulting in a better, more reliable, and accurate evaluation.

781. Difference between R2 and Adjusted R2?

ANS. Adding more independent variables will result in an increased value of R. This is the disadvantage of R square adding more independent variable irrespective of whether the new independent variable is significant or not the value of R square increases. But in the case of Adjusted R square if the new independent variable added is insignificant the adjusted r square has the capability to decrease therefore resulting in a better, more reliable, and accurate evaluation.

782. What is RMSE?

ANS. RMSE stands for ROOT MEAN SQUARE ERROR is a standard way to measure the error rate of the model. RMSE is a standard deviation of residuals or errors. Residuals or Errors are a measure of how far the data points are from the regression line. RMSE is a value that should be closer to 0.

783. Difference between Correlation and Regression?

ANS. Correlation measures the relative strength of linear relationship between two variables Independent(X) and Dependent Variable(Y). Correlation ranges from -1 to 1. The closer to -1, the stronger is the negative linear relationship. The closer to 1, the stronger is the positive linear relationship. The closer to 0, the weaker is the positive linear relationship.

Regression is nothing but to describe the relationship between two variables and how change in one variable affects the other variable. Regression is described by the best fit line. It is used for model building and prediction.

784. Why do we split our data into training and testing?

ANS. Splitting the data into a training data and testing data is a very important step for model evaluation. We divide our data into two sets which are training and testing most of the data is used for the training purpose and a small sample of the whole data is used for testing.

The whole process works in the following manners:

- First we train the model on the training data which we have created by splitting the original data set.
- After training the model we do our prediction on the testing data and store the predicted value in variable.
- The last step is to compare the predicted data and the testing data that we already know this is how we try to evaluate the model and calculate the accuracy of the model.

SYNTAX:

```
from sklearn.model_selection import train_test_split  
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size =0.2, random_state = 10)  
Basically we divide our data in X_train, Y_train which are the training data. X_test,Y_test are the testing data.
```

We only have to pass the data which is X and Y and the size of the test data in our case we have pass 0.2 which is nothing but 20% of the whole data.

That means our training data is 0.8 which is 80 % of the whole data.

Now since you are done with all the basic information about Linear Regression and you have a fair bit of idea about why we use Linear Regression, so let's now build our model

785. Steps for performing Linear Regression in Python.

ANS: Performing Linear Regression in Python:

1. Create a dataframe properly --> pd.read_csv(), pd.read_excel()
2. Assumption 1-There should be no outliers in the data --> pd.boxplot()
3. Assumption 2-Assumption of Linearity --> pairplot()
4. Create X and Y
5. Assumption

- 3-Assumption of Normality of Y --> distplot(), log()
6. Handle the skewness in the X --> skew(), log1p()
7. Assumption no 4-There should be no multicollinearity -->corr(), heatmap(), vif()
8. Splitting the data --> train_test_split(), manual splitting
9. Build the model:
- Create the model object --> obj=LinearRegression()
 - Train the model --> obj.fit(X_train,Y_train)
 - Predict using the model --> Y_pred=obj.predict(X_test)
- 10.Evaluating the model:
- Rsquare
 - Adjusted Rsquare
 - RMSE (ROOT MEAN SQUARE ERROR)
- 11.Tuning the model --> Manual feature selection, pvalues, Ridge Regression, Lasso Regression, Applying Feature engineering, PCA principle component analysis.

786. How does multicollinearity affect the linear regression?

Ans Multicollinearity occurs when some of the independent variables are highly correlated (positively or negatively) with each other. This multicollinearity causes a problem as it is against the basic assumption of linear regression. The presence of multicollinearity does not affect the predictive capability of the model. So, if you just want predictions, the presence of multicollinearity does not affect your output. However, if you want to draw some insights from the model and apply them in, let's say, some business model, it may cause problems.

One of the major problems caused by multicollinearity is that it leads to incorrect interpretations and provides wrong insights. The coefficients of linear regression suggest the mean change in the target value if a feature is changed by one unit. So, if multicollinearity exists, this does not hold true as changing one feature will lead to changes in the correlated variable and consequent changes in the target variable. This leads to wrong insights and can produce hazardous results for a business.

A highly effective way of dealing with multicollinearity is the use of VIF (Variance Inflation Factor). Higher the value of VIF for a feature, more linearly correlated is that feature. Simply remove the feature with very high VIF value and re-train the model on the remaining dataset.

787. Can you name a possible method of improving the accuracy of a linear regression model?

You can do so in many ways. One of the most common ways is 'The Outlier Treatment.'

Outliers have great significance in linear regression because regression is very sensitive to outliers. Therefore, it becomes critical to treat outliers with appropriate values. It can also prove useful if you replace the values with mean, median, mode or percentile depending on the distribution.

788. What are outliers? How do you detect and treat them?

An outlier is an observation point distant from other observations. It might be due to a variance in the measurement. It can also indicate an experimental error. Under such circumstances, you need to exclude the same from the data set. If you do not detect and treat them, they can cause problems in statistical analysis.

789. How do you interpret a Q-Q plot in a linear regression model?

As the name suggests, the Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words, you plot quantiles against quantiles.

Whenever you interpret a Q-Q plot, you should concentrate on the ' $y = x$ ' line. You also call it the 45-degree line in statistics. It entails that each of your distributions has the same quantiles. In case you witness a deviation from this line, one of the distributions could be skewed when compared to the other.

790. What is the importance of the F-test in a linear model?

The F-test is a crucial one in the sense that it tests the goodness of the model. When you reiterate the model to improve the accuracy with the changes, the F-test proves its utility in understanding the effect of the overall regression.

791. What are the disadvantages of the linear regression model?

One of the most significant demerits of the linear model is that it is sensitive and dependent on the outliers. It can affect the overall result. Another notable demerit of the linear model is overfitting. Similarly, underfitting is also a significant disadvantage of the linear model.

792. What is the curse of dimensionality? Can you give an example?

When you analyze and organize data in high-dimensional spaces (usually in thousands), various situations can arise that usually do not do so when you analyze data in low-dimensional settings (3-dimensional physical space). The curse of dimensionality refers to such phenomena.

Here is an example.

All kids love to eat chocolates. Now, you bring a truckload of chocolates in front of the kid. These chocolates come in different colors, shapes, tastes, and price. Consider the following scenario.

The kid has to choose one chocolate from the truck depending on the following factors.

Only taste – There are usually four tastes, sweet, salty, sour, and bitter. Hence, the child will have to try out only four chocolates before choosing one to its liking.

Taste and Color – Assume there are only four colors. Hence, the child will now have to taste a minimum of 16 (4×4) before making the right choice.

Taste, color, and shape – Let us assume that there are five shapes. Therefore, the child will now have to eat a minimum of 80 chocolates ($4 \times 4 \times 5$).

793. Compare Linear Regression and Decision Tree

Linear regression is used to predict continuous outputs where there is a linear relationship between the features of the dataset and the output variable.

Decision trees work by splitting the dataset, in a tree-like structure, into smaller and smaller subsets and make predictions based on which subset the new example falls into.

Linear regression is used for regression problems where it predicts something with infinite possible answers such as the price of a house.

Decision trees can be used to predict both regression and classification problems.

Linear regression is prone to underfitting the data. Switching to polynomial regression will sometimes help in countering underfitting.

Decision trees are prone to overfit the data. Pruning helps with the overfitting problem.

794. Name a disadvantage of R-squared and explain how would you address it?

R-squared (R^2) is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

R-squared takes values between 0 and 1, with 0 indicating that the proposed model does not improve prediction over the mean model and 1 indicating the perfect prediction. However, one drawback of R-squared is that its values can increase if we add predictors to the regression model, leading to a possible overfitting.

To address this issue, we can use Adjusted R-squared: a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance, and it decreases when a predictor improves the model by less than expected.

794. Does correlation imply causation? Why or why not?

No, while correlation is popularly used to provide information on the extent and direction of the linear relationship between two variables and can be used to determine whether a variable can be used to predict another, a high correlation does not imply causation.

For instance, you might find a correlation between umbrella malfunctions and a carpenter's income. As you can imagine, it is unlikely that there is a direct relation between the two, except

that people tend to open up their umbrellas during the rainy season and that wooden doors swell due to high humidity. In this case, there is a hidden cause, rain, that causes both the phenomena as mentioned above and consequently the high correlation between them.

795. Is linear regression suitable for time series analysis?

While linear regression can be used for time series analysis and generally yield workable results, the performance is not particularly remarkable. The two main factors for this are :

Time series generally have seasonal or periodic trends (such as peak seasons or even peak hours), which might be treated as outliers in linear regression and hence not appropriately accounted for.

Future prediction is a generally sought-after use case in time series analysis, which will require extrapolation and rarely results in good predictions.

ARIMA, ARCH, and LSTM are widely used and better performing algorithms for time series analysis.

796. How do you determine if a linear regression model is a good fit for your data?

To determine if a linear regression model is a good fit for the data, you can examine the residual plots, which should show no obvious patterns, and the distribution of the residuals, which should be approximately normal. You can also use statistical tests, such as the F-test or t-test, to determine if the model is significant and the coefficients are statistically significant. Another useful metric is the R-squared value, which measures the proportion of variance in the dependent variable that is explained by the independent variables.

797. What are some common techniques for dealing with multicollinearity in multiple linear regression models?

Multicollinearity occurs when the independent variables in a multiple linear regression model are highly correlated with each other. This can lead to unreliable coefficient estimates and reduce the interpretability of the model. Some common techniques for dealing with multicollinearity include:

- Removing one or more of the highly correlated independent variables
- Combining the highly correlated independent variables into a single variable
- Using regularization techniques, such as Ridge or Lasso regression, which shrink the coefficient estimates towards zero
- Collecting more data or using a larger sample size to reduce the effect of multicollinearity.

798. How do you check for heteroscedasticity in a linear regression model, and what are some techniques for addressing it?

Heteroscedasticity occurs when the variance of the residuals is not constant across all values of the independent variables. To check for heteroscedasticity, you can plot the residuals against the predicted values and look for a cone-like or fan-like shape. Statistical tests, such as the Breusch-Pagan or White test, can also be used to confirm the presence of heteroscedasticity.

Some common techniques for addressing heteroscedasticity include:

Transforming the dependent or independent variables

Using weighted least squares regression, which gives more weight to observations with smaller variances

Using robust regression techniques, which are less sensitive to outliers and non-normality.

799. How do you evaluate the performance of a linear regression model, and what metrics do you use?

To evaluate the performance of a linear regression model, you can use metrics such as:

Mean Squared Error (MSE): measures the average squared difference between the predicted and actual values of the dependent variable

Root Mean Squared Error (RMSE): takes the square root of the MSE to make the metric more interpretable in the same units as the dependent variable

Mean Absolute Error (MAE): measures the average absolute difference between the predicted and actual values of the dependent variable

R-squared: measures the proportion of variance in the dependent variable that is explained by the independent variables.

800. How do you handle outliers and influential points in a linear regression model?

Outliers and influential points can have a significant impact on the results of a linear regression model. Some common techniques for handling outliers and influential points include:

Removing them from the dataset if they are caused by errors or measurement issues

Transforming the dependent or independent variables if they are skewing the results

Using robust regression techniques, which are less sensitive to outliers and non-normality

Using diagnostic plots, such as Cook's distance or leverage plots, to identify influential points and remove them from the dataset or reweight them in the analysis.

Chapter 7 - Logistic Regression

801

What is Classification?

Ans: - In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data. Examples of classification problems include: Given an data, classify if it will rain or not. Given a data, classify it as new or not

802. Why is logistic regression called regression if it does the job of classification?

Ans: - It is called ‘Logistic Regression’ because its underlying technique is quite the same as Linear Regression. The term “Logistic” is taken from the logit function that is used in this method of classification.

803. What is Logistic Regression?

Ans: - Logistic Regression is one of the basic and popular algorithms to solve a classification problem.

It is mainly used in situations where there is a binary classification needed.

804. What is the similarity between linear regression and logistic regression?

Ans: - With linear regression you're looking for the k_i parameters:

$$h = k_0 + \sum k_i \cdot X_i = K_t \cdot X$$

With logistic regression you've the same aim but the equation is:

$$h = g(K_t \cdot X)$$

Where g is the sigmoid function:

$$g(w) = 1 / (1 + e^{-w})$$

So:

$$h = 1 / (1 + e^{-K_t \cdot X})$$

and you need to fit K to your data.

Assuming a binary classification problem, the output h is the estimated probability that the example x is a positive match in the classification task:

$$P(Y = 1) = 1 / (1 + e^{-K_t \cdot X})$$

When the probability is greater than 0.5 then we can predict "a match".

The probability is greater than 0.5 when:

$$g(w) > 0.5$$

and this is true when:

$$w = K_t \cdot X \geq 0$$

The hyperplane:

$$Kt \cdot X = 0$$

is the decision boundary.

Logistic regression is a generalized linear model using the same basic formula of linear regression but it is regressing for the probability of a categorical outcome.

805. Explain the mechanism of Logistic Regression

Ans: -

- Unlike actual regression, logistic regression does not try to predict the value of a numeric variable given a set of inputs.
- Instead, the output is a probability that the given input point belongs to a certain class.

For simplicity, lets assume that we have only two classes, and the probability in Q is

- P_+ -> the probability that a certain data point belongs to the '+' class.
- $P_- = 1 - P_+$.

Thus, the output of Logistic Regression always lies in [0, 1].

806. What are the applications of Logistic Regression?

Ans: -

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, it is used to predict mortality in injured patients, to predict the risk of developing a given disease. based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.).

Another example might be to predict whether a voter will vote which party. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It can also be used in weather forecasting, i.e. whether it will rain or not based on the weather condition variables.

Other example will be of the credit card issue. The Credit Card Fraud Detection problem is of significant importance to the banking industry because banks each year spend hundreds of millions of dollars due to fraud. When a credit card transaction happens, the bank makes a note of several factors. For instance, the date of the transaction, amount, place, type of purchase, etc.

Based on these factors, they develop a Logistic Regression model of whether or not the transaction is a fraud. For instance, if the amount is too high and the bank knows that the concerned person never makes purchases that high, they may label it as a fraud.

The logistic regression can also be used in marketing domain. Every day, when you browse your Facebook newsfeed, the powerful algorithms running behind the scene predict whether or not you would be interested in certain content (which could be, for instance, an advertisement). Such algorithms can be

viewed as complex variations of Logistic Regression algorithms where the Q to be answered is simple – will the user like this particular advertisement in his/her news feed?

Another example will be in the medical domain. A Logistic Regression classifier may be used to identify whether a tumour is malignant or if it is benign.

Several medical imaging techniques are used to extract various features of tumours. For instance, the size of the tumour, the affected body area, etc.

These features are then fed to a Logistic Regression classifier to identify if the tumour is malignant or if it is benign.

807. What are the differences between Logistic Regression and Linear Regression?

Linear Regression	Logistic Regression
In linear regression, the outcome (dependent variable) is continuous.	<ul style="list-style-type: none">Binary classification;is used when the response variable is categorical in nature. E.g. yes/no, true/false, red/green
The data is modelled using a straight line.	The probability of some obtained event is represented as a linear function of a combination of predictor variables.
Linear relationship between dependent and independent variables is required	Linear relationship between dependent and independent variables is NOT required
In the linear regression, the independent variable can be correlated with each other.	The variable must not be correlated with each other

808. What is a Sigmoid Function?

Ans:-

In order to map predicted values to probabilities, we use the sigmoid function.

The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.

Formula:- $S(z) = 1/(1+e^{-z})$

Code:-

```
def sigmoid(z):
```

```
return 1.0 / (1 + np.exp(-z))
```

809. What is the difference between Sigmoid function and SoftMax function?

SoftMax Function	Sigmoid Function
Used for multi-classification in logistic regression model.	Used for binary classification in logistic regression model.
The probabilities sum will be 1	The probabilities sum need not be 1.
Used in the different layers of neural networks.	Used as activation function while building neural networks.
The high value will have the higher probability than other values	The high value will have the high probability but not the higher probability.

810. In a nutshell, Explain the advantages and disadvantages of Logistic Regression?

Ans:-

Advantages:-

1. Highly interpretable, Outputs well-calibrated predicted probabilities.
2. Model training and prediction are fast.
3. Can perform well with a small number of observations.

Disadvantages:-

1. Presumes a linear relationship between the features and the log-odds of the response.
2. Is it not possible to apply a logistic regression algorithm on a larger Classification problem?

811. What are the assumptions of Logistic Regression?

Ans:-

- The Response Variable should be Binary in nature.
- The Observations are Independent
- There is No Multicollinearity Among Explanatory Variables.
- There are No Extreme Outliers
- There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable
- The Sample Size is Sufficiently Large.

812. Why does the response variable in the data should be binary in nature when using the logistic regression algorithm?

Ans:- Logistic Regression can work on multivariate variables but it will not give a precise accuracy since works and classifies on the sigmoid function and that sigmoid function is used for only binary nature variables. Hence Logistic Regression is prefeed in the data which is binary in nature.

813. What is Hypothesis test?

Ans:-

In logistic regression, hypotheses are of interest:

Null hypothesis :- Null hypothesis which is when all the coefficients in the regression equation take the value zero.

Alternate hypothesis :- Alternate hypothesis is that the model currently under consideration is accurate and differs significantly from the null of zero, i.e. gives significantly better than the chance or random prediction level of the null hypothesis.

814. What is Log Transformation?

Ans: -

The log transformation is, arguably, the most popular among the different types of transformations used to transform skewed data to approximately conform to normality. Log transformations and sq. root transformations moved skewed distributions closer to normality. So what we are about to do is common. This log transformation of the p values to a log distribution enables us to create a link with the normal regression equation. The log distribution (or logistic transformation of p) is also called the logit of p or $\text{logit}(p)$.

In logistic regression, a logistic transformation of the odds (referred to as logit) serves as the depending variable:

$$\text{Log}(odds) = \text{logit}(P) = \ln(P/1-P)$$

If we take the above dependent variable and add a regression equation for the independent variables, we get a logistic regression:

$$\text{logit}(p) = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots$$

As in least-squares regression, the relationship between the $\text{logit}(P)$ and X is assumed to be linear.

815. What is the general workflow of Logistic Regression Algorithm?

Ans:-

The general workflow is:

- 1) get a dataset
- 2) train a classifier
- 3) make a prediction using such classifier

816. What are the libraries required for implementing Logistic Regression?

Ans:-

```

import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
import warnings
warnings.filterwarnings("ignore")
pd.set_option("display.max_columns",None)

```

will be working on an Employee Dataset in which we have to predict whether the salary of the employee will be higher than 50k or less than 50k.

817 How to import the dataset into the python Environment?

Ans:-

Code:-

```

adult_df = pd.read_csv('adult_data.csv',header = None, delimiter=' *, *')
adult_df.head()

```

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States <=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States <=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States <=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States <=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba <=50K

818. The dataset does not has headers, how to define the headers on the dataset?

Ans: -

This can be done with columns functions:-

```

adult_df.columns = ['age', 'workclass', 'fnlwgt', 'education',
'education_num','marital_status','occupation', 'relationship','race', 'sex',
'capital_gain', 'capital_loss',
'hours_per_week', 'native_country', 'income']
adult_df.head()

```

819. How can you revert the original data frame in case any failure in your data analysis?

Ans: -

```

#CREATE A COPY OF THE DATAFRAME
adult_df_rev=pd.DataFrame.copy(adult_df)

```

820. How to drop some of the variables in the dataset?

Ans:-

Code:-

```
adult_df_rev = adult_df_rev.drop(['education','fnlwgt'], axis=1)
```

821. How to have look at the dataset?

Ans: - We can have a look at the dataset using head() and tail() functions

Code:- `adult_df_rev.head()`

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership
0	1077501	1296599	5000.0	5000.0	4975.0	36 months	10.65	162.87	B	B2	NaN	10+ years	RENT
1	1077430	1314167	2500.0	2500.0	2500.0	60 months	15.27	59.83	C	C4	Ryder	< 1 year	RENT
2	1077175	1313524	2400.0	2400.0	2400.0	36 months	15.96	84.33	C	C5	NaN	10+ years	RENT
3	1076863	1277178	10000.0	10000.0	10000.0	36 months	13.49	339.31	C	C1	AIR RESOURCES BOARD	10+ years	RENT
4	1075358	1311748	3000.0	3000.0	3000.0	60 months	12.69	67.79	B	B5	University Medical Group	1 year	RENT

822. How to check the null values in the dataset?

Ans:- The null values in the dataset can be checked by using `isnull()` function.

Code:-

```
df_rev.isnull().sum()
```

loan_amnt	0
funded_amnt	0
funded_amnt_inv	0
term	0
int_rate	0
installment	0
grade	0
sub_grade	0
emp_length	43061
home_ownership	0
annual_inc	0
verification_status	0
issue_d	0
purpose	0
zip_code	0
dti	0
delinq_2yrs	0
earliest_cr_line	0
inq_last_6mths	0
open_acc	0
pub_rec	0
revol_bal	0
revol_util	446

823. How to impute the missing values?

Ans:-

1. By using the mean

Code:-

```
colname2=['revol_util','collections_12_mths_ex_med',
'total_rev_hi_lim']
for x in colname2[:]:
    data[x].fillna(data[x].mean(),inplace=True)

data.isnull().sum()
```

Data.shape

2. By using info from other variables

```
emp_avg_income = data.groupby('emp_length').annual_inc.agg('mean')
```

```
def impute_emp_length(cols):
    emp_length = cols[0]
    annual_inc = cols[1]

    if pd.isnull(emp_length):
        if annual_inc < 70800:
            return '< 1 year'
```

```

elif annual_inc in range(70801,72000):
    return '1 year'
elif annual_inc in range(72000,72800):
    return '2 years'
elif annual_inc in range(72800,73600):
    return '3 years'
elif annual_inc in range(73600,74000):
    return '4 years'
elif annual_inc in range(74000,74500):
    return '5 years'
elif annual_inc in range(74500,74600):
    return '6 years'
elif annual_inc in range(74600,74700):
    return '7 years'
elif annual_inc in range(74700,74800):
    return '8 years'
elif annual_inc in range(74800,75900):
    return '9 years'
else:
    return '10+ years'
else:
    return emp_length

data['emp_length'] =
data[['emp_length','annual_inc']].apply(impute_emp_length, axis=1)

```

824. How to know the count of occurrences of the variables?

Ans: -

Code: - df.workclass.value_counts()

825. How to check for Outliers ?

Ans:- Outliers can be detected by the following ways: -

1. Extreme Value Analysis by Box Plot
2. Visualizing the data

826. How to convert the categorical data into numerical data?

Ans: - The different ways by Which you can convert the categorical variables into numerical ones are: -

1. Label Encoder
2. Manually Mapping
3. Dummy Variables
4. One hot label Encoding

827. On what basis should we decide that outliers should be eliminated or not ??

Ans:- If the quantity of outlier is less then we should eliminate them since the logistic regression doesn't allow outliers and if the quantity of outliers is more then we should keep them as it is and let the algorithm handle it.

828. How to check for outliers in the data?

1.By Using the Boxplot range

Code:-

```
df.boxplot(column='age')
plt.show()
```

2.Create a for loop that will calculate the IQR

```
#for value in colname:
```

```
    q1 = df['age'].quantile(0.25) #first quartile value
```

```
    q3 = df['age'].quantile(0.75) #third quartile value
```

```
    iqr = q3-q1 #Interquartile range
```

```
    low = q1-1.5*iqr #acceptable range
```

```
    high = q3+1.5*iqr #acceptable range
```

```
df_include = df.loc[(df['age'] >= low) & (df['age'] <= high)]
```

```
df_exclude = df.loc[(df['age'] < low) | (df['age'] > high)]
```

3.Finding the mean of the acceptable range.

```
age_mean=int(df_include.age.mean())
```

```
print(age_mean)
```

```
df_exclude.age=age_mean
```

4.Getting back the original shape of the dataframe.

```
df_rev=pd.concat([df_include,adult_df_exclude]) #concatenating both dfs to
get
```

```
#the original shape
```

```
df_rev.shape
```

5.The capping approach

```
df_exclude.loc[df_exclude["age"] < low, "age"] = low
```

```
df_exclude.loc[df_exclude["age"] > high, "age"] = high
```

829. Explain Label Encoder

Ans:- One hot encoding is used to encode the categorical column. It replaces a categorical column with its labels and fills values either 0 or 1. For example,

you can see the “color” column, there are 3 categories such as red, yellow, and green. 3 categories labeled with binary values.

Code:-

```
colname1=['grade','term','sub_grade','emp_length','home_ownership','verification_status',
'purpose','zip_code','earliest_cr_line','last_pymnt_d',
'next_pymnt_d','last_credit_pull_d','application_type','initial_list_status']
```

```
data.head()
from sklearn import preprocessing
le={}
for x in colname1:
    le[x]=preprocessing.LabelEncoder()
for x in colname1:
    data[x]=le[x].fit_transform(data[x])
data.head()
```

Before transforming

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length
0	1077501	1296599	5000.0	5000.0	4975.0	36 months	10.65	162.87	B	B2	NaN	10+ years
1	1077430	1314167	2500.0	2500.0	2500.0	60 months	15.27	59.83	C	C4	Ryder	< 1 year
2	1077175	1313524	2400.0	2400.0	2400.0	36 months	15.96	84.33	C	C5	NaN	10+ years
3	1076863	1277178	10000.0	10000.0	10000.0	36 months	13.49	339.31	C	C1	AIR RESOURCES BOARD	10+ years

After transforming

	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_length	home_ownership	annual_inc	verification_status
0	5000.0	5000.0	4975.0	0	10.65	162.87	1	6	1	5	24000.0	2
1	2500.0	2500.0	2500.0	1	15.27	59.83	2	13	10	5	30000.0	1
2	2400.0	2400.0	2400.0	0	15.96	84.33	2	14	1	5	12252.0	0
3	10000.0	10000.0	10000.0	0	13.49	339.31	2	10	1	5	49200.0	1
4	3000.0	3000.0	3000.0	1	12.69	67.79	1	9	0	5	80000.0	1

830. Explain Manual Mapping

Ans: - Manual mapping is a technique where we individually take one by one element and assign them a value. This is done where you need to convert specific values and the number of these values in less.

Syntax:-

Example : -

```
df["column_name"] = df.column_name.map({Desired Value : Actual Value })
```

831. What are Dummy Variables?

Ans:- A Dummy variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. Its requires less computational power compared to other techniques. However the coding length is more compared to other techniques.

Syntax:-

```
import pandas as pd
raw_data = {'first_name': ['Saurabh', 'Amit', 'Mansi', 'Pranjali', 'Ankita'],
'last_name': ['Parab', 'Parab', 'Rane', 'Gawde', 'Lokande'],
'sex': ['male', 'male', 'female', 'female', 'female']}
df = pd.DataFrame(raw_data, columns = ['first_name', 'last_name', 'sex'])
Df
```

	first_name	last_name	sex
0	Saurabh	Parab	male
1	Amit	Parab	male
2	Mansi	Rane	female
3	Pranjali	Gawde	female
4	Ankita	Lokande	female

```
pd.get_dummies(df, columns=['sex'])
```

	first_name	last_name	sex_female	sex_male
0	Saurabh	Parab	0	1
1	Amit	Parab	0	1
2	Mansi	Rane	1	0
3	Pranjali	Gawde	1	0
4	Ankita	Lokande	1	0

832. Explain One Hot Label Encoding

Ans:- It is a process that converts categorical data to integers or a vector of ones and zeros. The length of vector is determined by number of expected classes or categories. Each element in the vector represents a class. Therefore, a one is used to indicate which class it is and everything else will be zero.

Code:-

```
from sklearn.preprocessing import OneHotEncoder
type_one_hot = OneHotEncoder(sparse=False).fit_transform(
train_new.array.to_numpy().reshape(-1,1))
```

If we have categorical data that we think may be important, we want to be able to use this in the model. This is because regression algorithms and classification algorithms won't be able to process it. This is when one-hot

encoding is useful.

So we will use the most convenient way to transform the categorical variables

833. Explain the steps of Label Encoding?

Ans: -

1. Create a user defined function which will loop through the entire dataset and will return all the categorical variables into the list.

Code:-

```
colname1=['grade','term','sub_grade','emp_length','home_ownership','verification_status',
'purpose','zip_code','earliest_cr_line','last_pymnt_d',
'next_pymnt_d','last_credit_pull_d','application_type','initial_list_status']
data.head()
from sklearn import preprocessing

le={}
```

2. Use the Label Encoder Function.

```
for x in colname1:
    le[x]=preprocessing.LabelEncoder()
for x in colname1:
    data[x]=le[x].fit_transform(data[x])
data.head()
```

	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_length	home_ownership	annual_inc	verification_status
0	5000.0	5000.0	4975.0	0	10.65	162.87	1	6	1	5	24000.0	2
1	2500.0	2500.0	2500.0	1	15.27	59.83	2	13	10	5	30000.0	1
2	2400.0	2400.0	2400.0	0	15.96	84.33	2	14	1	5	12252.0	0
3	10000.0	10000.0	10000.0	0	13.49	339.31	2	10	1	5	49200.0	1
4	3000.0	3000.0	3000.0	1	12.69	67.79	1	9	0	5	80000.0	1

834. What is Scaling?

Ans:- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Machine learning is like making a mixed fruit juice. If we want to get the

best-mixed juice, we need to mix all fruit not by their size but based on their right proportion. We just need to remember apple and strawberry are not the same unless we make them similar in some context to compare

their attribute. Similarly, in many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant. The two major techniques for Feature Scaling are:

- Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].
- Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

835. Explain the syntax of Scaling and which scaling technique we will be using in this algorithm?

Ans:- We will be using the standard scaler in this algorithm. Standard Scaler assumes a normal distribution for data within each feature. The scaling makes the distribution centred around 0, with a standard deviation of 1 and the mean removed.

Formula:-

$$x(i) - \text{mean}(x)$$

$$\frac{x(i) - \text{mean}(x)}{\text{std}(x)}$$

Where sd is the standard deviation of x.

Syntax:-

```
from sklearn.preprocessing import StandardScaler  
scaler = StandardScaler()  
scaler.fit(X)  
X=scaler.transform(X)  
print(X)
```

836. How will we decide how to train the model and how to test the model on the data which is available to us?

Ans:- In Machine Learning, we split the data into 2 parts, training and testing parts.

We train the model on training data and compare its results with the test Data.

837. What is the threshold for splitting the data?

Ans: - Usually we follow the threshold of 70:30 of the data i.e., 70 % of the data to the training and 30% of the data. It depends on the situation whether you need more data for your model if it's not giving you the

accuracy.

Syntax: -

```
from sklearn.model_selection import train_test_split  
X_, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3,  
random_state=101)
```

838. What do you mean by feature splitting?

Ans: - A feature splitting is a technique to generate a few other features from the existing one to improve the model performance. for example, splitting names into first and last names.

For Example :- Deriving the profit variable from selling price and cost price variable .

839. What do you mean by feature selection?

Ans:- Feature selection means the process of selecting the independent variables and the dependant variables for your model.

Syntax:-

```
Independent variables alias name = dataframe_name.values[column  
names]  
dependent variable alias name = dataframe_name.values[column name]
```

Code:-

```
X=adult_df_rev.values[:,0:-1]  
Y=adult_df_rev.values[:, -1]
```

840. What is loc and iloc in python and what is the difference between them?

Ans:- The main distinction between the two methods is:-

- loc gets rows (and/or columns) with particular labels.
- iloc gets rows (and/or columns) at integer locations.

For Example:-

ILOC:-

```
X=adult_df_rev.values[:,0:-1]  
Y=adult_df_rev.values[:, -1]
```

LOC:-

```
X= adult_df_rev.values['age','post',gender',etc]  
Y=adult_df_rev.values['income']
```

841. What is the code for building the Logistic Regression Model?

Ans:-

```
from sklearn.linear_model import LogisticRegression  
#create a model
```

```

classifier=LogisticRegression()
#build train the model
classifier.fit(X_train,Y_train)
#predict using the model you created

Y_pred=classifier.predict(X_test)
#we are using this for comparison
#print(list(zip(Y_test,Y_pred)))

print(classifier.coef_)
print(classifier.intercept_)

```

842. Can we create a custom function of Confusion Matrix so that we can picturized it beautifully?

Ans:-

```

import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
from sklearn.utils.multiclass import unique_labels
import itertools
def plot_confusion_matrix(cm, classes,
                         normalize=False,
                         title='Confusion Matrix',
                         cmap=plt.cm.Greens):
    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
    print("Normalized Confusion Matrix")
    else:
        print("Confusion Matrix")
    print(cm)

    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=35)
    plt.yticks(tick_marks, classes)

    fmt = '.2f' if normalize else 'd'
    thresh = cm.max() / 2.

    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[0])):
        plt.text(j, i, format(cm[i, j], fmt),
                 horizontalalignment='center',
                 color='white' if cm[i, j] > thresh else 'black')

```

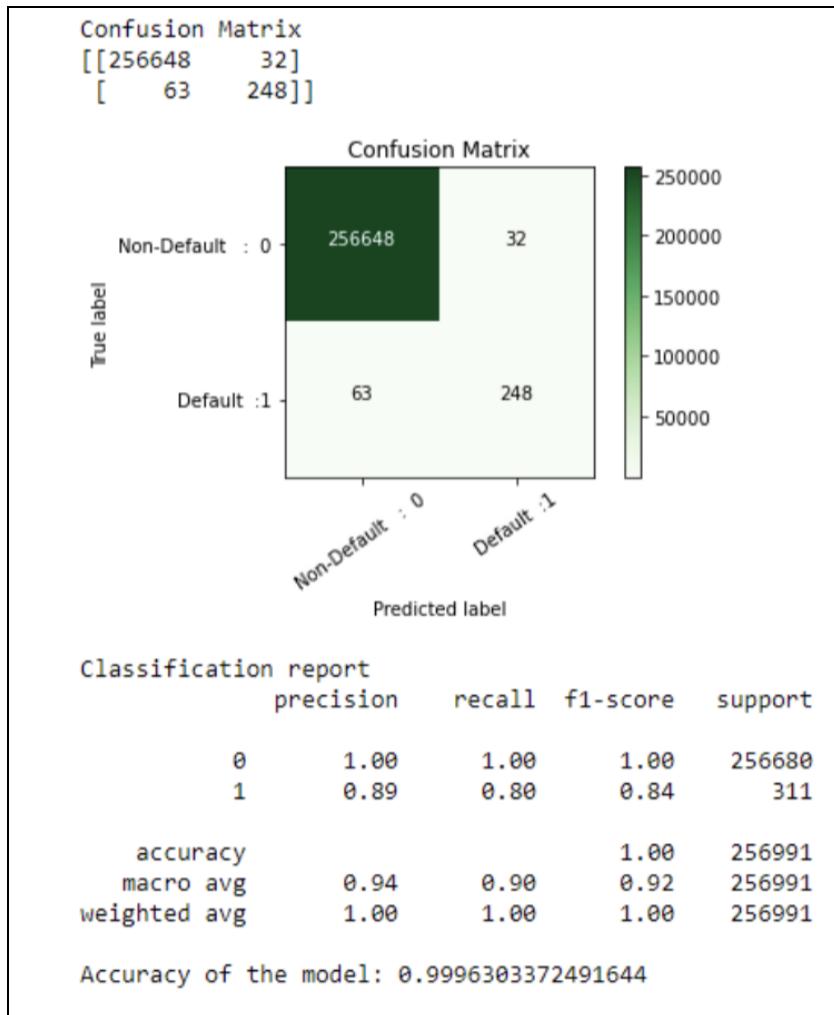
```
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.tight_layout()
```

843. How to check whether our model is performing well or not?

Ans:- Sklearn provides various functions for this purpose like accuracy ,confusion Matrix,classification report etc,

Code:-

```
from sklearn.metrics import confusion_matrix, accuracy_score,
classification_report
cfm = confusion_matrix(Y_test,Y_pred)
print(cfm)
print("CLASSIFICATION MATRIX:")
print(classification_report(Y_test,Y_pred))
acc = accuracy_score(Y_test,Y_pred)
print("ACCURACY OF THE MODEL:",acc)
```



844. What is Accuracy of model?

Accuracy is the quintessential classification metric. It is pretty easy to understand. And easily suited for binary as well as a multiclass classification problem.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

Accuracy is the proportion of true results among the total number of cases examined.

When to use?

Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or No class imbalance.

So here we have an accuracy of 82 which Is pretty good.

845. What is Precision of model?

Ans:-

Precision means what proportion of predicted Positives is truly Positive?

Precision = $(TP)/(TP+FP)$

In the asteroid prediction problem, we never predicted a true positive.

And thus precision=0

When to use?

Precision is a valid choice of evaluation metric when we want to be very sure of our prediction. For example: If we are building a system to predict if we should decrease the credit limit on a particular account, we want to be very sure about our prediction or it may result in customer dissatisfaction.

846. What is Recall Factor ?

Ans:- Recall Factor means proportion of actual Positives is correctly classified?

Recall = $(TP)/(TP+FN)$

In the asteroid prediction problem, we never predicted a true positive.

And thus recall is also equal to 0.

When to use?

Recall is a valid choice of evaluation metric when we want to capture as many positives as possible. For example: If we are building a system to predict if a person has cancer or not, we want to capture the disease even if we are not very sure.

Here we have an recall factor of 0.95 for 0 and 0.45 for 1 meaning our algorithm is performing well for 0 and not 1

847. Does Logistic Regression provide any feature in these condition where the recall factor is not satisfying?

Ans:- Logistic Regression provides the proba function in such cases

Proba function adjusts the threshold and the caps the values so that the recall factor improves resulting in better accuracy.

```
#Store the predicted probabilities
y_pred_prob=classifier.predict_proba(X_test)
print(y_pred_prob)
y_pred_class=[]
for value in y_pred_prob[:,1]:
if value > 0.46:
y_pred_class.append(1)
else:
y_pred_class.append(0)
print(y_pred_class)
TYPE 1 ERROR TYPE 2 ERROR
for a in np.arange(0.4,0.61,0.01):
predict_mine = np.where(y_pred_prob[:,1] > a, 1, 0)
```

```

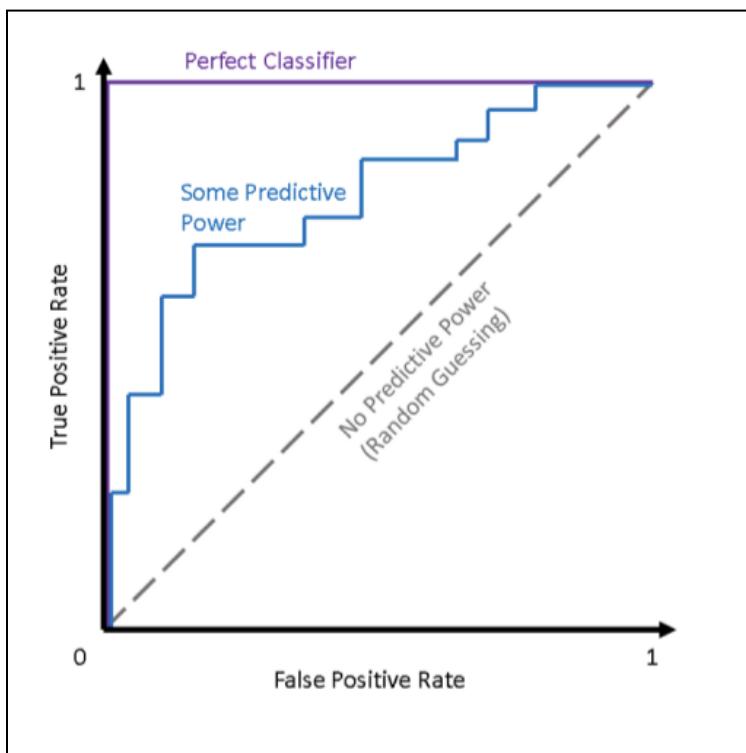
cfm=confusion_matrix(Y_test, predict_mine)
total_err=cfm[0,1]+cfm[1,0]
print("Errors at threshold ", a, ":" ,total_err, " , type 2 error :",
      cfm[1,0]," , type 1 error:", cfm[0,1])

```

848. What is the difference between SVM and Logistic Regression?

SVM	Logistic Regression
<ul style="list-style-type: none"> SVM tries to finds the “best” margin (distance between the line and the support vectors) that separates the classes and this reduces the risk of error on the data 	<ul style="list-style-type: none"> Logistic regression does not, instead it can have different decision boundaries with different weights that are near the optimal point.
<ul style="list-style-type: none"> SVM works well with unstructured and semi-structured data like text and images 	<ul style="list-style-type: none"> Logistic regression works with already identified independent variables.
<ul style="list-style-type: none"> SVM is based on geometrical properties of the data 	<ul style="list-style-type: none"> Logistic regression is based on statistical approaches.
<ul style="list-style-type: none"> The risk of overfitting is less in SVM 	<ul style="list-style-type: none"> Logistic regression is vulnerable to overfitting.

849. What is ROC?



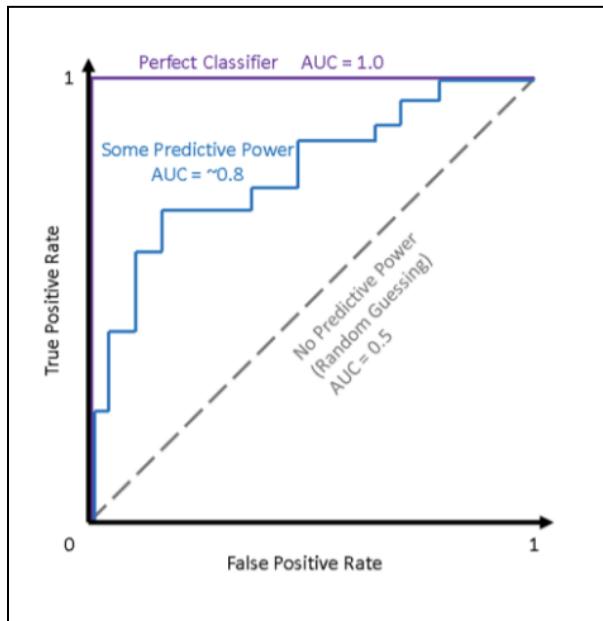
The ROC curve is produced by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds. For example, in logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class. Normally in logistic regression, if an observation is predicted to be positive at > 0.5 probability, it is labelled as positive. ROC curves help us visualize how these choices affect classifier performance.

One advantage presented by ROC curves is that they aid us in finding a classification threshold that suits our specific problem. For example, if we were evaluating an email spam classifier, we would want the false positive rate to be really, really low. We wouldn't want someone to lose an important email to the spam filter just because our algorithm was too aggressive. We would probably even allow a fair amount of actual spam emails (true positives) through the filter just to make sure that no important emails were lost.

850. What is AUC?

Ans:-

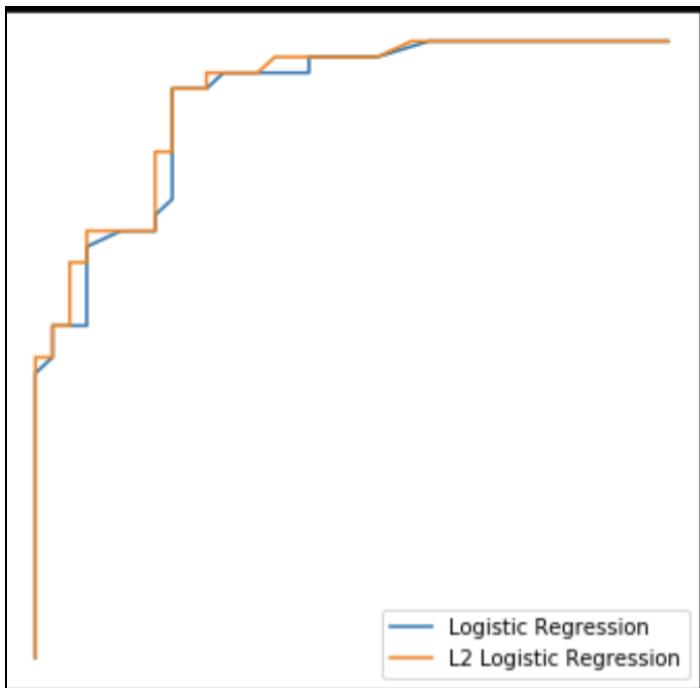
While it is useful to visualize a classifier's ROC curve, in many cases we can boil this information down to a single metric — the AUC. AUC stands for area under the (ROC) curve. Generally, the higher the AUC score, the better a classifier performs for the given task.



851. What is the code for plotting ROC curve?

Ans:-

```
# Plot ROC curves
fig, ax = plt.subplots(figsize=(6,6))
ax.plot(lr_fp_rates, lr_tp_rates, label='Logistic Regression')
ax.plot(l2_fp_rates, l2_tp_rates, label='L2 Logistic Regression')
ax.set_xlabel('False Positive Rate')
ax.set_ylabel('True Positive Rate')
ax.legend();
```



852. How to calculate the AUC scores?

The sklearn library has an `auc()` function, which I'll make use of here to calculate the AUC scores for both versions of the classifier. `auc()` takes in the true positive and false positive rates we previously calculated it and returns the AUC score to you.

Code:-

```
# Get AUC scores
from sklearn.metrics import auc
print(f'Logistic Regression (No reg.) AUC {auc(lr_fp_rates, lr_tp_rates)}')
print(f'Logistic Regression (L2 reg.) AUC {auc(l2_fp_rates, l2_tp_rates)}')
```

OP:-

```
Logistic Regression (No reg.) AUC 0.902979902979903
Logistic Regression (L2 reg.) AUC 0.9116424116424116
```

853. What is SGD Stochastic Gradient Descent Classifier?

Ans:- Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as Support Vector Machines and Logistic Regression.

The advantages of Stochastic Gradient Descent are:

- 1.Efficiency.
- 2.Ease of implementation (lots of opportunities for code tuning).

The disadvantages of Stochastic Gradient Descent include:

1. SGD requires a number of hyperparameters such as the regularization parameter and the number of iterations.
2. SGD is sensitive to feature scaling.

854. What is the difference between (SGD)Stochastic Gradient Descent Classifier and Logistic Regression?

SGD is a optimization method, while Logistic Regression (LR) is a machine learning algorithm/model. You can think of that a machine learning model defines a loss function, and the optimization method minimizes/maximizes it. Some machine learning libraries could make users confused about the two concepts. For instance, in scikit-learn there is a model called SGD Classifier which might mislead some user to think that SGD is a classifier. But no, that's a linear classifier optimized by the SGD.

Chapter 8 - NLP

855. What is NLP?

NLP stands for Natural Language Processing and it is a branch of data science that consists of systematic processes for analyzing, understanding, and deriving information from the text data in a smart and efficient manner.

856. What are the uses of NLP?

Natural Language Processing is useful in various domains like Chat bots, Extracting insights from feedback and surveys, text-classification, etc.

857. What are the different algorithms in NLP?

NLP is used to analyze text, allowing machines to understand how humans speak.

This human-computer interaction enables real-world applications like

- a. automatic text summarization
- b. sentiment analysis
- c. topic extraction
- d. named entity recognition
- e. parts-of-speech tagging
- f. relationship extraction
- g. stemming, and more.

NLP is commonly used for text mining, machine translation, and automated question answering.

858. What problems can NLP solve?

NLP can solve many problems like, automatic summarization, machine translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation etc.

859. What is Regular Expression?

A regular expression (sometimes called a rational expression) is a sequence of characters that define a search pattern, mainly for use in pattern matching with strings, or string matching, i.e. “find and replace”-like operations.

Regular expressions are a generalized way to match patterns with sequences of characters.

6. What are the different applications of Regular Expression in Data Science?

- a. Search engines like Google, Yahoo, etc. Google search engine understands that you are a tech guy so it shows you results related to you.
- b. Social websites feed like the Facebook news feed. The news feed algorithm understands your interests using natural language processing and shows you related Ads and posts more likely than other posts.
- c. Speech engines like Apple Siri.

d. Spam filters like Google spam filters. It's not just about the usual spam filtering, now spam filters understand what's inside the email content and see if it's a spam or not.

860. What are the packages in Python to help in Regular Expression

The package which we commonly use for regular expression is re. We can import the package using following command

```
Import re
```

What is match function?

```
import re  
re.match('ni','nitin')  
Match='ni'
```

9. What are the common patterns used in regular expression?

\w+ -> word

\d -> digit

\s -> space

* -> wildcard

+ or * -> greedy match

\S -> anti space i.e. it matches anything which is not a space

[A-Z] – matches all the character in the range of capital A and capital Z

10. What are the important functions to use in Regular Expression?

findall() – It finds all the patterns in a string

search() - It search for a pattern

match() – It matches an entire string or a sub string

split() – It splits a string in Regular Expression. It returns a list object

861.What is the difference between match and search function?

Match tries to match the string from beginning whereas search matches it wherever it finds the pattern. The below example will help you understand better

```
import re  
print(re.match('kam', 'kamal'))  
print(re.match('kam', 'nitin kamal'))  
print(re.search('kam','kamal'))  
print(re.search('kam','nitin kamal'))  
<re.Match object; span=(0, 3), match='kam'>  
None  
<re.Match object; span=(0, 3), match='kam'>  
<re.Match object; span=(6, 9), match='kam'>
```

862. Guess the output of the following

```
import re  
re.split('\s','The Data Monk is cool')
```

```
[‘The’, ‘Data’, ‘Monk’, ‘is’, ‘cool’]
```

863. Work in finding the output of the following

```
regx = r"\w+"
strx = "This isn't my pen"
re.findall(regx,strx)
['This', 'isn', 't', 'my', 'pen']
```

864. How to write a regular expression to match some specific set of characters in a string?

```
special_char = r"[?/}{;:]"
The above Regular Expression will take all the characters between []
```

865. Write a regular expression to split a paragraph every time it finds an exclamation mark

```
import re
exclamation = r"!"
strr = "Data Science comprises of innumerable topics! The aim of this 100 Days series is to get you started assuming ! that you have no prior! knowledge of any of these topics. "
excla = re.split(exclamation,strr)
print(excla)
['Data Science comprises of innumerable topics', ' The aim of this 100 Days series is to get you started assuming ', ' that you have no prior', ' knowledge of any of these topics. ']
```

Get all the words starting with capital letter

```
capital = r"[A-Z]\w+"
print(re.findall(capital,strr))
['Data', 'Science', 'The', 'Days']
```

867. Find the output of the following code?

```
digit = "12 34 98"
find_digit = r"\d+"
print(re.findall(find_digit,digit))
['12', '34', '98']
```

868. What is tokenization?

Tokenization is one of the most important part of NLP. It simply means to break down the string into smaller chunks. It breaks the paragraph into words, sentences, etc.

869. What is NLTK?

NLTK stands for Natural Language Toolkit Library and it is a package in Python which is very commonly used for tokenization.

```
from nltk.tokenize import word_tokenize  
word_tokenize("This is awesome!")  
['This', 'is', 'awesome', '!']
```

870. What are the important nltk tokenizer?

sent_tokenize – Tokenize a sentence

tweet_tokenize – This one is exclusively for tweets which can come handy if you are trying to do sentiment analysis by looking at a particular hashtag or tweets

regexp_tokenize – tokenize a string or document based on a regular expression pattern

871. What is the use of the function set() ?

The data type set is a collection. It contains an unordered collection of unique and immutable objects. So when you extract a set of words from a novel, then it will get you the distinct words from the complete novel. It is a very important function and it will continue to come in the book as you go ahead.

Tokenize the paragraph given below in sentence.

```
para = "This is the story about Piyush,29, Senior Data Scientist at Imagine Incorporation and myself, Pihu,24, Junior Data Scientist at the same organization. This is about the journey of Piyush once he retired from his job, after being unsatisfied with the way his career was moving ahead. Be with Piyush and Pihu to understand Data Science and Machine Learning."
```

```
import nltk.tokenize import sent_tokenize  
import nltk.tokenize import word_tokenize  
para = "This is the story about Piyush,29, Senior Data Scientist at Imagine Incorporation and myself, Pihu,24, Junior Data Scientist at the same organization. This is about the journey of Piyush once he retired from his job, after being unsatisfied with the way his career was moving ahead. Be with Piyush and Pihu to understand Data Science and Machine Learning."  
sent = sent_tokenize(para)  
print(sent)  
['This is the story about Piyush,29, Senior Data Scientist at Imagine Incorporation and myself, Pihu,24, Junior Data Scientist at the same organization.', 'This is about the journey of Piyush once he retired from his job, after being unsatisfied with the way his career was moving ahead.', 'Be with Piyush and Pihu to understand Data Science and Machine Learning.']}
```

Now get all the words from the above paragraph

```
word = word_tokenize(para)
['This', 'is', 'the', 'story', 'about', 'Piyush,29', ',', 'Senior', 'Data',
'Scientist', 'at', 'Imagine', 'Incorporation', 'and', 'myself', ',', 'Pihu,24',
',', 'Junior', 'Data', 'Scientist', 'at', 'the', 'same', 'organization', '',
'This', 'is', 'about', 'the', 'journey', 'of', 'Piyush', 'once', 'he', 'retired',
'from', 'his', 'job', ',', 'after', 'being', 'unsatisfied', 'with', 'the', 'way',
'his', 'career', 'was', 'moving', 'ahead', ',', 'Be', 'with', 'Piyush', 'and',
'Pihu', 'to', 'understand', 'Data', 'Science', 'and', 'Machine', 'Learning', '.']
```

872. Now get the unique words from the above paragraph

```
word=set(word_tokenize(para))
print(word)
{'retired', 'ahead', 'the', 'about', 'with', 'Piyush,29', 'Senior', 'Piyush',
'being', 'Science', 'was', 'Imagine', 'at', 'journey', 'way', 'same', 'and',
'Pihu', 'Pihu,24', 'Learning', 'from', 'story', 'he', 'Be', 'Machine', 'once',
'to', 'unsatisfied', 'Junior', 'of', 'career', 'Data', 'moving', 'is',
'understand', ',', 'myself', 'after', 'job', ',', 'Incorporation', 'Scientist',
'organization', 'This', 'his'}
```

873. What is the use of .start() and .end() function?

Basically .start() and .end() helps you find the starting and ending index of a search. Below is an example:

```
x = re.search("Piyush",para)
print(x.start(),x.end())
```

24 30

874. What is the OR method?

OR method, as the name suggests is used to give condition to the regular expression. See the example below:-

```
x = r"\d+ | \w+"
```

The above regex will get you all the words and numbers, but it will ignore other characters like punctuation, ampersand, etc.

875. What are the advance tokenization techniques?

Take for example [A-Za-z]+, this will get you all the alphabets regardless of upper or lowercase Alphabets

876. How to write a regex to match spaces or commas?

(/s+|,) – The /s+ will get you one or more spaces, and the pipe will mark an OR operator to take the comma into consideration

877. How to include special characters in a regex?

If you have any experience with regular expression or SQL queries, then this syntax will look familiar. You need to give a backward slash before any special character like below

(\,\.\.\?) – This will consider comma, full stop and question mark in the text

878. What is the difference between (a-z) and [A-Z]?

This is a very important concept, when you specify (a-z), it will only match the string “a-z”. But when you specify [A-Z] then it covers all the alphabet between upper case A and Z

879. Once again go through the difference between search() and match() function.

Search() will find your desired regex expression anywhere in the string, but the match always looks from the beginning of the string. If a match() function hits a comma or something, then it will stop the operation then and there itself. Be very particular on selecting a function out of these.

880. What is topic modeling?

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body.

881. What is bag-of-words?

Bag-of-words is a process to identify topics in a text. It basically counts the frequency of the token in a text. Example below to help you understand the simple concept of bag-of-words

para = “The game of cricket is complicated. Cricket is more complicated than Football”

The – 1

game – 1

of-1

cricket-1

is-2

complicated-2

Cricket – 1

than – 1

Football – 1

As you can see, the word cricket is counted two times as bag-of-words is case sensitive.

882. How to counter the case sensitive nature of bag-of-words?

It's a logical question, just convert every word in lower or upper case and then count the words. Look for question 35 to convert every word in lower case using loop.

883. What is counter?

A counter is a container that keeps count of number of times equivalent values are added. It looks similar to dictionary in Python. Counter supports three forms of initialization. Its constructor can be called with a sequence of items, a dictionary containing keys and counts, or using keyword arguments mapping string names to counts.

884. How to import Counter in Python?

Counter is present in the Collection package, you can use it directly by importing it like below:
from collections import Counter

885. Use the same paragraph used above and print the top 3 most common words

The code is self explanatory and is given below:

```
word2 = word_tokenize(para)
lower_case = [t.lower() for t in word2]
bag_of_words = Counter(lower_case)
print(bag_of_words.most_common(3))
[('the', 4), ('.', 4), ('data', 3)]
```

886. What is text preprocessing?

text pre processing is a complete process to make the text ready for analysis by removing stop words, common punctuations, spelling mistakes, etc. Before any analysis you are suppose to process the text.

887. What are the commonly used methods of text preprocessing?

Converting the complete text in either lower or upper case

Tokenization

Lemmatization/Stemming

Removing stop words

888. How to tokenize only words from a paragraph while ignoring the numbers and other special character?

```
x = "Here is your text. Your 1 text is here"
from nltk.corpus import stopwords
only_alphabet = [w for w in word_tokenize(x.lower())
if w.isalpha()]
print(only_alphabet)
w.isalpha() function will check if the word has only text in it and will remove the numbers
Output
['here', 'is', 'your', 'text', 'your', 'text', 'is', 'here']
```

889. What are stop words?

Stop words are common occurring words in a text which have high frequency but less importance.
Words like the, are, is, also, he, she, etc. are some of the examples of English stop words.

890. How to remove stop words from my text?

```
from nltk.corpus import stopwords
para = "Your text here. Here is your text"
tokens = [w for w in word_tokenize(para.lower())
if w.isalpha()]
stoppy = [t for t in tokens
if t not in stopwords.words('english')]
```

891. What is Lemmatization?

Lemmatization is a technique to keep words in its base form or dictionary form of the word.
Example will help you understand better
The lemma of better will be good.
The word “walk” is the base form of the word “Walking”

892. Give an example of Lemmatization in Python

```
x = "running"
import nltk
nltk.download('wordnet')
lem.lemmatize(x,"v")
Output
'Run'
```

893. How to lemmatize the texts in your paragraph?

Use the module WordNetLemmatizer from nltk.stem
from nltk.stem import WordNetLemmatizer
lower_tokens = word_tokenize(para)

```
lower_case = [t.lower() for t in lower_tokens]
only_alphabet = [t for t in lower_case if t.isalpha()]
without_stops = [x for x in only_alphabet if x not in stopwords.words("English")]
lemma = WordNetLemmatizer()
lemmatized = [lemma.lemmatize(t) for t in without_stops]
```

894. What is gensim?

Gensim is a very popular open-source NLP library. It is used to perform complex tasks like:-

- a. Building document or word vectors
- b. Topic identification

895. What is a word vector?

Word vector is a representation of words which helps us in observing relationships between words

and documents. Based on how the words are used in text, the word vector help us to get meaning and context of the words. Example, the word vector will connect Bangalore to Karnataka and Patna to Bihar where Bangalore and Patna are capital of the Indian state Karnataka and Bihar.

These are multi-dimensional mathematical representation of words created using deep learning method. They give us insight into relationships between words in a corpus.

896. What is LDA?

LDA is used for topic analysis and modeling. It is used to extract the main topics from a dataset. LDA stands for Latent Dirichlet Allocation. Topic Modelling is the task of using unsupervised learning to extract the main topics (represented as a set of words) that occur in a collection of Documents.

897. What is gensim corpus?

Gensim corpus converts the tokens in bag or words. It gives result in a list of (token id, token reference). The gensim dictionary can be updated and reused

898. What is stemming?

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP). Stemming is also a part of queries and Internet search engines.

899. Give an example of stemming in Python

```
from nltk.stem.porter import PorterStemmer
```

```

stem = PorterStemmer()
x = "running"
stem.stem(x)
Output
'run'

```

900. What is tf-idf?

term frequency and inverse document frequency. It is to remove the most common words other than stop words which are there in a particular document, so this is document specific.

$$w_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right)$$

$w_{i,j}$ = tf-idf weight for token i in document j

$tf_{i,j}$ = number of occurrences of token i in document j

df_i = number of documents that contain token i

N = total number of documents

The weight will be low in two cases:-

- a. When the term frequency is low i.e. number of occurrence of a word is low
- b. When N is equal to df_i , then the log will be close to zero

So, using (b), if a word occurs in all the document, then the log value will be low

If the word “abacus” is present 5 times in a document containing 100 words. The corpus has 200 documents, with 20 documents mentioning the word “abacus”.

The formula for tf-idf will be :-

$$(5/100) * \log(200/20)$$

901. How to create a tf-idf model using gensim?

```

from gensim.models.tfidfmodel import TfidfModel
tfidf = TfidfModel(corpus)
tf_idf_weights = tfidf[doc]
# Sort the weights from highest to lowest: sorted_tfidf_weights
sorted_tfidf_weights = sorted(tfidf_weights, key=lambda w: w[1], reverse=True)

```

```
# Print the top 5 weighted words
for term_id, weight in sorted_tfidf_weights[:5]:
    print(dictionary.get(term_id), weight)
```

902. What is Named Entity Recognition?

It is a process of identifying important named entity texts in a document. Ex. organization, dashboard names, work of arts, etc.

It is present in the ne_chunk_sents() function in nltk package. It can be used as below:-
chunk_Sent = nltk.ne_chunk_sents(Part_Of_Speech_sentence_token, binary = True)

903. What is POS?

Part of Speech tag in Natural Language Processing is used to tag a word according to its use in the sentence. It tags the word as a part of speech.

It is present as pos_tag() in nltk package. You can feed the tokenized word in a loop to get the POS tag for each word like below:-

```
pos = [nltk.pos_tag(x) for x in tokenized_word_variable]
```

904. What is the difference between lemmatization and stemming?

Lemmatization gets to the base of the word whereas stemming just chops the tail of the word to get the base form. Below example will serve you better:

See is the lemma of saw, but if you try to get the stem of saw, then it will return 's' as the stem.
See is the lemma of seeing, stemming seeing will get you see.

905. What is spacy package?

Spacy is a very efficient package present in Python which helps in easy pipeline creation and finding entities in tweets and chat messages.

906. How to initiate the English module in spacy?

```
import spacy
x = spacy.load('en',tagger=False,parser=False,matcher=False)
```

907. Why should one prefer spacy over nltk for named entity recognition?

Spacy provides some extra categories, other than the one provided by nltk. These categories are:-

NORP

Cardinal

money

Work of art

Language
Event
So, you can try spacy for NER according to your need

908. What are the different packages which uses word vectors?

Spacy and gensim are the two packages which we have covered so far that uses word vectors.

909. What if your text is in various different languages? Which package can help you in Named

Entity Recognition for most of the largely spoken languages?

Polygot is one of the package which supports more than 100 languages and uses word vector for Named Entity Recognition

910. What is supervised learning?

Supervised learning is a form of Machine Learning where your model is trained by looking at a given output for all the inputs. The model is trained on this input-output combination and then the learning of the model is tested on the test dataset. Linear Regression and Classification are two examples of supervised learning.

911. How can you use Supervised Learning in NLP?

Suppose you have a chat data and looking at the keyword you have specified the sentiment of the customer. Now you have got a set of data which have complete chat and the sentiment associated with the chat. Now you can use supervised learning to train the data on this dataset and then use it while there is alive chat to identify the ongoing sentiment of the customer.

912. What is Naïve-Bayes model?

Naive Bayes classifiers are linear classifiers that are known for being simple yet very efficient. The probabilistic model of naive Bayes classifiers is based on Bayes' theorem, and the adjective naive comes from the assumption that the features in a dataset are mutually independent.

913. What is the flow of creating a Naïve Bayes model?

```
from sklearn import metrics
from sklearn.naive_bayes import MultinomialNB
# Instantiate a Multinomial Naive Bayes classifier: nb_classifier
nb_classifier = MultinomialNB()
# Fit the classifier to the training data
nb_classifier.fit(count_train,y_train)
# Create the predicted tags: pred
```

```

pred = nb_classifier.predict(count_test)
# Calculate the accuracy score: score
score = metrics.accuracy_score(y_test,pred)
print(score)
# Calculate the confusion matrix: cm
cm = metrics.confusion_matrix(y_test,pred,labels=['FAKE','REAL'])
print(cm)

```

Let's take some sample text and try to implement basic algorithms first

914. What is TF-IDF?

TF-IDF stands for term frequency and Inverse Term Frequency which first takes the frequency of the words in the document and then looks for most relevant terms. IDF will filter out most of the high frequency words like preposition, etc. and will keep the less occurring important terms

915. What is POS?

POS stands for Parts of Speech tagging and it is used to tag the words in your document according to Parts of Speech. So, noun, pronoun, verb, etc. will be tagged accordingly and then you can filter what you need from the dataset. If I am just looking for names of people mentioned in the comment box then I will look for mainly Nouns. This is a basic but very important algorithm to work with.

916. Take an example to take a sentence and break it into tokens i.e. each word

```

text = "The Data Monk will help you learn and understand Data Science"
tokens = word_tokenize(text)
print(tokens)
['The', 'Data', 'Monk', 'will', 'help', 'you', 'learn', 'and', 'understand',
'Data', 'Science']

```

917.. Take the same sentence and get the POS tags

```

from nltk import word_tokenize, pos_tag
text = "The Data Monk will help you learn and understand Data Science"
tokens = word_tokenize(text)
print(pos_tag(tokens))
[('The', 'DT'), ('Data', 'NNP'), ('Monk', 'NNP'), ('will', 'MD'), ('help', 'VB'),
('you', 'PRP'), ('learn', 'VB'), ('and', 'CC'), ('understand', 'VB'), ('Data',
'NNP'), ('Science', 'NN')]

```

918. Take the following line and break it into tokens and tag POS using function

```
data = "The Data Monk was started in Bangalore in 2018. Till now it has more than 30 books on
```

```

Data Science on Amazon"
data = "The Data Monk was started in Bangalore in 2018. Till now it has more than 30 books on
Data Science on Amazon"
#Tokenize the words and apply POS
def token_POS(token):
    token = nltk.word_tokenize(token)
    token = nltk.pos_tag(token)
    return token
token = token_POS(data)
Token

```

Output

```

[('The', 'DT'),
 ('Data', 'NNP'),
 ('Monk', 'NNP'),
 ('was', 'VBD'),
 ('started', 'VBN'),
 ('in', 'IN'),
 ('Bangalore', 'NNP'),
 ('in', 'IN'),
 ('2018', 'CD'),
 ('.', '.'),
 ('Till', 'VB'),
 ('now', 'RB'),
 ('it', 'PRP'),
```

919. What is NER?

NER stands for Named Entity Recognition and the work of this algorithm is to extract specific chunk of data from your text data. Suppose you want to get all the Nouns from the dataset . It is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, medical codes. Etc.

920. What are some of the common tags in POS. You need to know the meaning of the tags to use it in your regular expression

DT – Detreminer

FW – Foreign word

JJ – Adjective

JJR – Comparative Adjective

NN – Singular Noun
NNS – Plural Noun
RB – Adverb
RBS – Superlative Adverb
VB – Verb
You can get the complete list on internet

921. Implement NER on the tokenized and POS tagged sentence used above.

```
nltk.download('maxent_ne_chunker')
nltk.download('words')
ne_chunked_sents = nltk.ne_chunk(token)
named_entities = []
for tagged_tree in ne_chunked_sents:
    if hasattr(tagged_tree, 'label'):
        entity_name = ''.join(c[0] for c in tagged_tree.leaves()) #
        entity_type = tagged_tree.label() # get NE category
        named_entities.append((entity_name, entity_type))
print(named_entities)
[('Data Monk', 'ORGANIZATION'), ('Bangalore', 'GPE'), ('Data Science', 'PERSON'),
 ('Amazon', 'ORGANIZATION')]
```

Code Explanation

`nltk.download` will import `maxent_ne_chunker` which is used to break the sentence into named entity chunks and `nltk.download('words')` will download the dictionary. We already have a variable `token` which contains POS tagged tokens. `nltk.ne_chunk(token)` will tag the tokens to Named entity chunks.

`function hasattr()` is used to check if an object has the given named attribute and return true if present, else false.

`.leaves()` function is used to get the leaves of the node and `label()` will get you the NER label

922. What are n-grams?

A combination of N words together are called N-Grams. N grams ($N > 1$) are generally more informative as compared to words (Unigrams) as features. Also, bigrams ($N = 2$) are considered as the most important features of all the others. The following code generates bigram of a text.

923. Create a 3-gram of the sentence below

“The Data Monk was started in Bangalore in 2018”

```
def ngrams(text, n):
    token = text.split()
    final = []
```

```

for i in range(len(token)-n+1):
    final.append(token[i:i+n])
return final
ngrams("The Data Monk was started in Bangalore in 2018",3)

```

```

In [63]: def ngrams(text, n):
    token = text.split()
    final = []
    for i in range(len(token)-n+1):
        final.append(token[i:i+n])
    return final
ngrams ("The Data Monk was started in Bangalore in 2018",3)

Out[63]: [['The', 'Data', 'Monk'],
           ['Data', 'Monk', 'was'],
           ['Monk', 'was', 'started'],
           ['was', 'started', 'in'],
           ['started', 'in', 'Bangalore'],
           ['in', 'Bangalore', 'in'],
           ['Bangalore', 'in', '2018']]

```

924. What is the right order for a text classification model components?

- Text cleaning
- Text annotation
- Text to predictors
- Gradient descent
- Model tuning

925. What is CountVectorizer?

CountVectorizer is a class from sklearn.feature_extraction.text. It converts a selection of text documents to a matrix of token counts.

Let's take up a project and try to solve it using NLP. Here we will only create the dataset and will apply Random forest and NLP to train our dataset to identify the sentiment of a review
 Objective of the project is to predict the correct tag i.e. whether people liked the food or not using NLP and Random Forest.

926. How to create a dataset? What to write in it?

Open an excel file and save it as Reviews (in the csv format). Now make two columns in the sheet like the one given below

Review	Liked
This restaurant is awesome	1
Food not good	0
Ambience was wow	1
The menu is good	1
Base was not good	0
Very bad	0
Wasted all the food	0
Delicious	1
Great atmosphere	1
Not impressed with the food	0
Nice	1
Bad taste	0
Great presentation	1
Lovely flavor	1
Polite staff	1
Bad management	0

Basically you can write the review of anything like Movies, food, restaurant, etc. Just make sure to keep the format like this. Thus your dataset is ready.

927. What all packages do I need to import for this project?

It's always good to start with importing all the necessary packages which you might use in the Project

```
import re
import pandas as pd
import numpy as np
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
We will discuss each of these as tackle the problem
```

928. How to import a csv file in Python?

Importing csv file in python requires importing pandas library and using read_csv function
review = pd.read_csv('C://Users//User//Downloads//Restaurant_Reviews.csv')

929. Let's view the top and bottom 5 lines of the file to make sure we are good to go with the Analysis

Use the commands given below

review.head() and review.tail()

```
In [43]: review.tail()  
  
Out[43]:  
      Review  Liked  
11    Bad taste     0  
12  Great presentation     1  
13   Lovely flavor     1  
14   Polite staff     1  
15  Bad management     0
```

930. Now we will clean the dataset. Will start with removing numbers and punctuations.

Write

a regular expression for removing special characters and numbers review is the name of the data set and Review is the name of the column

```
final = []  
for i in range(0,16):  
x = re.sub('[^a-zA-Z]', ' ',review['Review'][i] )
```

931. Now we want to stem the words. Do you remember the definition of stemming?

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP). Stemming is also a part of queries and Internet search engines.

```
final = []  
for i in range(0,16):  
x = re.sub('[^a-zA-Z]', ' ',review['Review'][i] )
```

```
x = x.lower()
x = x.split()
port = PorterStemmer()
x = [port.stem(words) for words in x
if not words in set(stopwords.words('english'))]
```

932. What does the above snippet do?

port = PorterStemmer() allocates the stemming function to the variable port
port.stem(words) for words in x - It takes all the words individually. Also remove the words which are stopwords.

The above loop will get all the non stop words and stem the words

933. Create the final dataset with only stemmed words.

```
final = []
for i in range(0,16):
    x = re.sub('[^a-zA-Z]', ' ',review['Review'][i] )
    x = x.lower()
    x = x.split()
    port = PorterStemmer()
    x = [port.stem(words) for words in x
    if not words in set(stopwords.words('english'))]
    x = ' '.join(x)
    final.append(x)
```

Let's see how the final dataset looks like after removing the stop words and stemming the text

```
In [47]: final

Out[47]: ['restaur awesom',
 'food good',
 'ambienc wow',
 'menu good',
 'base good',
 'bad',
 'wast food',
 'delici',
 'great atmospher',
 'impress food',
 'nice',
 'bad tast',
 'great present',
 'love flavor',
 'polit staff',
 'bad manag']
```

934. How to use the CountVectorizer() function? Explain using an example

```
from sklearn.feature_extraction.text import CountVectorizer
corpus = ['The Data Monk helps in providing resource to the users',
'It is useful for people making a career in Data Science',
'You can also take the 100 days Challenge of TDM']
counter = CountVectorizer()
X = counter.fit_transform(corpus)
print(counter.get_feature_names())
print(X.toarray())
get_feature_name() will take all the words from the above dataset and will arrange it in an
alphabetical order
fit_transform() will transform each line of the dataset as compared to the result of
get_feature_name()
toArray will change the datatype to Array
Lets understand the output
```

```
['100', 'also', 'can', 'career', 'challenge', 'data', 'days', 'for', 'helps',
'in', 'is', 'it', 'making', 'monk', 'of', 'people', 'providing', 'resource',
'science', 'take', 'tdm', 'the', 'to', 'useful', 'users', 'you']

[[0 0 0 0 0 1 0 0 1 1 0 0 0 1 0 0 1 1 0 0 0 2 1 0 1 0]
[0 0 0 1 0 1 0 1 0 1 1 1 0 0 1 0 0 1 0 0 0 0 1 0 0]
[1 1 1 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 1]]
```

The first output is the 26 unique words from the 3 lines of document arranged in alphabetical Order. The next three contains the presence of the above words in the document. 0 present in the 1,2,3, and 4th place of the first row suggests that the words 100, also, can, and career are not present in the first line of the input.

Similarly 2 present on the 22nd position shows that the word “the” is present twice in the first row of input.

The first row of input is “The Data Monk helps in providing resource to the users”

935. Now let's apply CountVectorizer on our dataset

```
from sklearn.feature_extraction.text import CountVectorizer  
cv = CountVectorizer(max_features = 1000)  
X = cv.fit_transform(final).toarray()  
max_feature = 1500 will make sure that at max 1000 words are put into the master array. In  
case you are planning to apply this on a huge dataset, then do increase the max_feature  
component.  
X will have the same array of occurrence across all the features as we have seen in above  
example
```

```
print(X)  
[[0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]  
[0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  
[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1]  
[0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0]  
[0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  
[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  
[0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  
[0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1]  
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  
[0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0]  
[0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

936. How to separate the dependent variable?

As we know we want to see whether the review was positive or not. So the dependent variable here is the second column and we have put the value of the second column in a different variable i.e. y

```
from sklearn.feature_extraction.text import CountVectorizer  
cv = CountVectorizer(max_features = 1500)  
X = cv.fit_transform(final).toarray()  
y = review.iloc[:,1].values
```

So, X has the array containing array of occurrence of different words across all the words and y has the binary value where 1 denotes like and 0 denotes did not like

937. Now we need to split the complete data set into train and test

```
from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)  
You already know about X and y, the test_size will divide the train and test dataset in 75:25 ratio  
respectively  
Now you will have to train the model on X_train and y_train
```

938. Random forest is one of the best model to work on supervised learning. By the way, what is Random forest?

Before we start with explaining a forest, we need to know what is a tree? Random forest is made of decision trees. To illustrate the concept, we'll use an everyday example: predicting the tomorrow's maximum temperature for our city. To keep things straight, I'll use Seattle, Washington, but feel free to pick your own city.

In order to answer the single max temperature question, we actually need to work through an entire series of queries. We start by forming an initial reasonable range given our domain knowledge, which for this problem might be 30–70 degrees (Fahrenheit) if we do not know the time of year before we begin. Gradually, through a set of questions and answers we reduce this range until we are confident enough to make a single prediction.

Since temperature is highly dependent on time of year, a decent place to start would be: what is the season? In this case, the season is winter, and so we can limit the prediction range to 30–50 degrees because we have an idea of what the general max temperatures are in the Pacific Northwest during the winter. This first question was a great choice because it has already cut our range in half. If we had asked something non-relevant, such as the day of the week, then we could not have reduced the extent of predictions at all and we would be back where we started. Nonetheless, this single question isn't quite enough to narrow down our estimate so we need to find out more information.

A good follow-up question is:

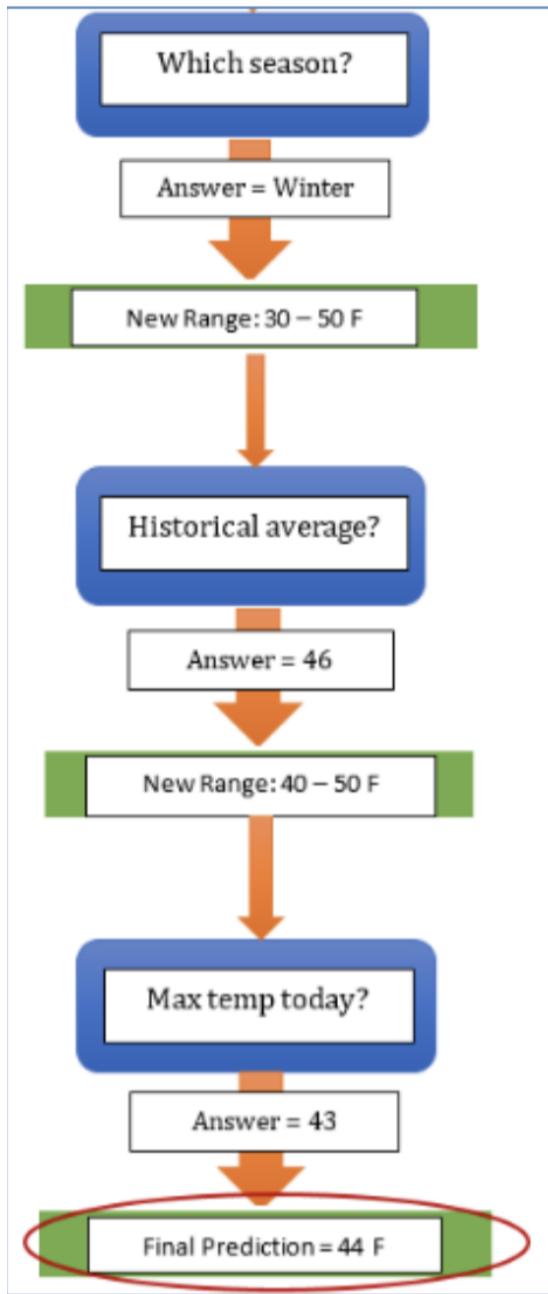
what is the historical average max temperature on this day?

For Seattle on December 27, the answer is 46 degrees.

This allows us to further restrict our range of consideration to 40–50 degrees. Again, this was a high-value question because it greatly reduced the scope of our estimate.

We need to have similar questions and once we put everything in a flow we will get a decision Tree.

So, to arrive at an estimate, we used a series of questions, with each question narrowing our possible values until we were confident enough to make a single prediction. We repeat this decision process over and over again in our daily lives with only the questions and answers changing.



939. What is Random Forest?

Every person comes to the problem with different background knowledge and may interpret the exact same answer to a question entirely differently. In technical terms, the predictions have variance because they will be widely spread around the right answer. Now, what if we take predictions from hundreds or thousands of individuals, some of which are high and some of which are low, and decided to average them together? Well, congratulations, we have created a random forest! The fundamental idea behind a random forest is to combine many decision trees into a single model.

Every person comes to the problem with different background knowledge and may interpret the exact same answer to a question entirely differently. In technical terms, the predictions have variance because they will be widely spread around the right answer. Now, what if we take predictions from hundreds or thousands of individuals, some of which are high and some of which are low, and decided to average them together? Well, congratulations, we have created a random forest! The fundamental idea behind a random forest is to combine many decision trees into a single model.

940. Let's create our Random forest model here

```
model = RandomForestClassifier(n_estimators = 10,  
criterion = 'entropy')  
model.fit(X_train, y_train)
```

941. Define n_estimator

n_estimator is basically the number of trees you want to create in your forest. Try to vary the number of trees in this forest.

In general, the more trees you use the better get the results. However, the improvement decreases as the number of trees increases, i.e. at a certain point the benefit in prediction performance from learning more trees will be lower than the cost in computation time for learning these additional trees.

Random forests are ensemble methods, and you average over many trees. Similarly, if you want to estimate an average of a real-valued random variable (e.g. the average height of a citizen in your country) you can take a sample. The expected variance will decrease as the square root of the sample size, and at a certain point the cost of collecting a larger sample will be higher than the benefit in accuracy obtained from such larger sample.

942. Define criterion. Why did you use entropy and not gini?

Gini is intended for continuous attributes and Entropy is for attributes that occur in classes.

Gini is to minimize misclassification

Entropy is for exploratory analysis

Entropy is a little slower to compute

943. What is model.fit()?

model.fit() helps you in create your model. The two parameters are that of training dataset i.e. X_train and y_train. It will take the values or the output of the reviews and will create a lot of decision trees to fit the output on the basis of input. These rules will be applied to your testing dataset to get the results

944. Let's predict the output for the testing dataset

```
y_pred = model.predict(X_test)
```

You have just created the model on X_train and y_train. Now you need to predict the output for X_test. We already have the output for these, but we want our model to predict the answer so that we can match the answers or output

945. Now let's check the confusion matrix to see how many of our outputs were correct

```
from sklearn.metrics import confusion_matrix  
cm = confusion_matrix(y_test, y_pred)
```

946. Lastly, what is confusion matrix and how to know the accuracy of the model?

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

Let's take example of a confusion matrix

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

So, our rows contains real values for a binary classifier and the columns have our predicted values. 50 and 100 shows that the predicted and actual values were correctly identified. 10 and 5 shows that the predicted values were not correct. Explore precision, recall, etc.

As far as accuracy is concerned, the formula is simple = $(50+100)/(50+10+5+100)$
i.e. total correct prediction divided by all the prediction.

Our model had very less dataset. The confusion matrix resulted in the following

```
In [51]: from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)
|
cm
```

Out[51]: array([[1, 0],
 [0, 3]], dtype=int64)

Complete code

```
import re
import pandas as pd
import numpy as np
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
review = pd.read_csv('C://Users//User//Downloads//Restaurant_Reviews.csv')
review.tail()
final = []

for i in range(0,16):
    x = re.sub('[^a-zA-Z]', ' ', review['Review'][i] )
    x = x.lower()
    x = x.split()
    port = PorterStemmer()
    x = [port.stem(words) for words in x]
    if not words in set(stopwords.words('english'))]
    x = ' '.join(x)
    final.append(x)
cv = CountVectorizer(max_features = 1500)
X = cv.fit_transform(final).toarray()
y = review.iloc[:, 1].values
print(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)
from sklearn.ensemble import RandomForestClassifier
```

```
model = RandomForestClassifier(n_estimators = 501,  
criterion = 'entropy')  
model.fit(X_train, y_train)  
y_pred = model.predict(X_test)  
y_pred  
cm = confusion_matrix(y_test, y_pred)  
cm
```

947. What is sub() method?

The re.sub() function in the re module can be used to replace substrings.
The syntax for re.sub() is re.sub(pattern,repl,string).
That will replace the matches in string with repl.

948. Convert all the text into lower case and split the words

```
final = []  
for i in range(0,16):  
    x = re.sub('[^a-zA-Z]', ' ',review['Review'][i] )  
    x = x.lower()  
    x = x.split()
```

Decision Tree and Random Forest

901.

Q2. What are the different types of terminologies of Decision Tree?

Ans. The terminologies related to decision tree are as follows:

Root Node: Root node represents the entire population of a small part of population that further gets divided into two or more homogeneous sets.

Splitting: Splitting is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node is divided or splitted into further nodes then that node is called as Decision node.

Leaf Node or Terminal Node: Node that do not split further is known as Leaf Node or Terminal Node.

Pruning: When we remove a sub-node of a decision node that process is known as pruning. Pruning is the opposite process of splitting.

Branch or Sub-Tree: A subset or sub section of a entire tree is called as branch or sub-tree.

Parent Node and Child Node: A node which is divided into sub-nodes is known as Parent Node and the sub-nodes are the child node of the parent node.

Q3. Why the Decision Trees are useful?

Ans. The decision trees are useful because:

Decision tree is simple and easy to use, understand and explain.

Decision tree can do feature scaling.

Decision tree can provide strategic answers.

Decision tree required very little effort to prepare.

Decision tree can be used for both classification and regression problems.

Q4. What are the types of decision tree?

Ans. Decision tree algorithm can be used to solve both the classification problems as well as the regression problems.

Categorical Variable Decision tree: This type of decision tree is used to solve the classification problems where the dependent variable is categorical in nature.

Continuous Variable Decision tree: This type of decision tree is used to solve the regression problems where the dependent variable is

continuous or numerical in nature.

Q5. What is a Categorical Variable Decision Tree?

ANS. A categorical variable decision tree is nothing but the tree that is used to solve the classification problems where the dependent variable is categorical in nature. We can also say that when the Y is categorical in nature the tree is called as decision tree classification. X can be numeric or categorical.

Q6. What is a Continuous Variable Decision Tree?

Ans. A continuous variable decision tree is nothing but the tree that is used to solve the regression problems where the dependent variable is continuous or numerical in nature. We can also say that when the Y is continuous in nature the tree is called as decision tree regression. X can be numeric or categorical.

Q7. What are the assumptions while creating a decision tree?

Ans. The assumptions while creating a decision tree are:

At the start the whole training dataset is considered as the root.

Features values are preferred to be categorical in nature and if not then they are converted to discrete values before building the model.

All the records are disturbed recursively on the basis of the attribute values.

Ordering of the attributes as the root node or the internal node is done by using some statistical method.

Q8. What kind of problems are decision tree most suitable?

Ans. The decision tree are most suitable for solving:

Where we have structured data or tabular data.

Where the outputs are discrete in nature.

Where the explanation of the decisions are required.

Where the training set contains error.

Where the training data has missing values or null values.

Q9. What is Attribute Selection Measures?

Ans. Suppose we have a large dataset containing N numbers of attributes or features so then deciding which attribute to place at the root node and which attribute to place at the other internal nodes or at the sub-nodes is a difficult step. We cannot randomly select any attribute to be the root node. If we select random attribute it will give us poor results and a very low accuracy.

So to solve this issue of attribute selection we can use some criteria like:

Entropy

Information Gain

Gini Index

Gain Ratio

Reduction in Variance

Chi square

All these criteria will help us calculate the values for every attribute. The values will be sorted and the attributes will be placed according to the following order. The attribute or the feature with high value will be placed at the root node when the criterion used is Information gain.

When we use information gain as the criterion we assume that the attributes are categorical and when we use gini index as the criterion we assume that the attributes are continuous.

Q10. What is Entropy?

Ans. Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

The formula of Entropy:

$$\text{Entropy}(S) = -P(\text{yes})\log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{no})$$

Where,

S = Total number of samples.

P(yes) = Probability of yes.

P(no) = Probability of no.

Q11. What is Information Gain?

Ans. Information gain is a statistical property which is used measure how well a given attribute differentiate or separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute or the feature that returns the highest information gain and the smallest entropy. Attribute having the highest information gain is split first. According to the value of the Information we split the node and build the decision tree.

Information gain is nothing but it calculates the entropy before split and average entropy after splitting the dataset based on the attribute values.

The formula of Information Gain:

$$\text{Information Gain}(A,B) = \text{Entropy}(A) - \text{Entropy}(A,B)$$

Q12. On what basis Information Gain selects an attribute?

Ans. First the Information Gain is calculated for all the variable and selects the variable or attribute which has the highest information gain value.

Q13. What is Gini Index?

Ans. We can understand the Gini index as a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions and easy to implement whereas information gain favors smaller partitions with distinct values.

The formula for Gini Index:

$$\text{Gini} = 1 - \sum_j P_j^2$$

Gini Index works with the categorical target variable “Yes” or “No”, “True” or “False”. It performs only Binary splits.

Steps to Calculate Gini index for a split

1. Calculate Gini for sub-nodes, using the above formula for success(p) and failure(q) ($p^2 + q^2$).

2. Calculate the Gini index for split using the weighted Gini score of each node of that split.

CART (Classification and Regression Tree) uses the Gini index method to create split points.

Q14. What is Reduction in Variance?

Ans. Reduction in Variance is one of the technique which is used for Attribute Selection Measures(ASM) for decision tree. Reduction in variance is used when we have continuous target variable or when we are dealing with continuous variable decision tree. This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population.

The formula of variance:

$$\text{Variance} = \sum(X - (\text{sample mean}))^2 / N$$

Q3. Why the Decision Trees are useful?

Ans. The decision trees are useful because:

Decision tree is simple and easy to use, understand and explain.

Decision tree can do feature scaling.

Decision tree can provide strategic answers.

Decision tree required very little effort to prepare.

Decision tree can be used for both classification and regression problems.

Q4. What are the types of decision tree?

Ans. Decision tree algorithm can be used to solve both the classification problems as well as the regression problems.

Categorical Variable Decision tree: This type of decision tree is used to solve the classification problems where the dependent variable is categorical in nature.

Continuous Variable Decision tree: This type of decision tree is used to solve the regression problems where the dependent variable is continuous or numerical in nature.

Q5. What is a Categorical Variable Decision Tree?

ANS. A categorical variable decision tree is nothing but the tree that is used to solve the classification problems where the dependent variable is categorical in

nature. We can also say that when the Y is categorical in nature the tree is called as decision tree classification. X can be numeric or categorical.

Q6. What is a Continuous Variable Decision Tree?

Ans. A continuous variable decision tree is nothing but the tree that is used to solve the regression problems where the dependent variable is continuous or numerical in nature. We can also say that when the Y is continuous in nature the tree is called as decision tree regression. X can be numeric or categorical.

Q7. What are the assumptions while creating a decision tree?

Ans. The assumptions while creating a decision tree are:

At the start the whole training dataset is considered as the root.

Features values are preferred to be categorical in nature and if not then they are converted to discrete values before building the model.

All the records are disturbed recursively on the basis of the attribute values.

Ordering of the attributes as the root node or the internal node is done by using some statistical method.

Q8. What kind of problems are decision tree most suitable?

Ans. The decision tree are most suitable for solving:

Where we have structured data or tabular data.

Where the outputs are discrete in nature.

Where the explanation of the decisions are required.

Where the training set contains error.

Where the training data has missing values or null values.

Q9. What is Attribute Selection Measures?

Ans. Suppose we have a large dataset containing N numbers of attributes or features so then deciding which attribute to place at the root node and which attribute to place at the other internal nodes or at the sub-nodes is a difficult step. We cannot randomly select any attribute to be the root node. If we select random attribute it will give us poor results and a very low accuracy.

So to solve this issue of attribute selection we can use some criteria like:

Entropy

Information Gain

Gini Index

Gain Ratio

Reduction in Variance

Chi square

All these criteria will help us calculate the values for every attribute. The values will be sorted and the attributes will be placed according to the following order.

The attribute or the feature with high value will be placed at the root node when

the criterion used is Information gain.

When we use information gain as the criterion we assume that the attributes are categorical and when we use gini index as the criterion we assume that the attributes are continuous.

Q10. What is Entropy?

Ans. Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

The formula of Entropy:

$$\text{Entropy}(s) = -P(\text{yes})\log_2 P(\text{yes}) - P(\text{no})\log_2 P(\text{no})$$

Where,

S = Total number of samples.

P(yes) = Probability of yes.

P(no) = Probability of no.

Q11. What is Information Gain?

Ans. Information gain is a statistical property which is used measure how well a given attribute differentiate or separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute or the feature that returns the highest information gain and the smallest entropy. Attribute having the highest information gain is split first. According to the value of the Information we split the node and build the decision tree.

Information gain is nothing but it calculates the entropy before split and average entropy after splitting the dataset based on the attribute values.

The formula of Information Gain:

$$\text{Information Gain}(A,B) = \text{Entropy}(A) - \text{Entropy}(A,B)$$

Q12. On what basis Information Gain selects an attribute?

Ans. First the Information Gain is calculated for all the variable and selects the variable or attribute which has the highest information gain value.

Q13. What is Gini Index?

Ans. We can understand the Gini index as a cost function used to evaluate splits in the dataset. It is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions and easy to implement whereas information gain favors smaller partitions with distinct values.

The formula for Gini Index:

$$\text{Gini} = 1 - \sum_j P_j^2$$

2

Gini Index works with the categorical target variable "Yes" or "No", "True" or "False". It performs only Binary splits.

Steps to Calculate Gini index for a split

1. Calculate Gini for sub-nodes, using the above formula for success(p) and failure(q) (p_2+q_2).

2. Calculate the Gini index for split using the weighted Gini score of each node of that split.

CART (Classification and Regression Tree) uses the Gini index method to create split points.

Q14. What is Reduction in Variance?

Ans. Reduction in Variance is one of the technique which is used for Attribute Selection Measures(ASM) for decision tree. Reduction in variance is used when we have continuous target variable or when we are dealing with continuous variable decision tree. This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population.

The formula of variance:

$$\text{Variance} = \sum(X - (\text{sample mean}))^2 / N$$

Q15. What do you mean by Pruning Decision Trees?

Ans. When we remove a sub-node of a decision node that process is known as pruning. Pruning is the opposite process of splitting. The splitting of the tree results in fully grown trees until the stopping criteria are reached. The fully grown tree is likely to overfit the data resulting in poor accuracy on unseen data. Pruning is a process where we trim off the branches of the tree remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets: training data set A and validation data set B. Prepare the decision tree using the segregated training data set A. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set B.

Q16. What are the different types of pruning?

Ans. There are two types of pruning:

Pre-Pruning: Pre-Pruning is nothing but a technique that refers to early stopping of the growth of the decision tree. The Pre-Pruning is also known as the hyperparameter technique which takes place before the training of the model. The hyperparameters which we use for Pre-Pruning are `max_depth`, `min_samples_leaf`, `min_samples_split`. Pre-Pruning is also used to handle the overfitting problem in the decision tree.

Post-Pruning: Post-Pruning is nothing but a technique that allows the decision tree to grow fully and after it is grown fully it then removes the tree branches. The technique which is used for post-pruning is known as cost complexity pruning(ccp). `ccp_alpha` is the parameter to achieve Post_Pruning. Post-Pruning is also used to handle the problem of overfitting in the Decision Tree.

Q17. Which is better Linear model or Tree-Based-Model?

Ans. Basically, it depends that what kind of data we are working upon.

If we are working upon a dataset in which the relationship between the dependent variable and independent variable is very high in such kind of data the linear regression will perform very well but the tree-based model will perform poorly.

Suppose we are working upon a dataset where there is high non-linearity and complex relation between the dependent and the independent variable

in this case the tree-based model will perform very well and the linear regression will fail.

Suppose we are working upon a dataset where we want to build a model which can be easily explained to the people in such cases decision tree model will always perform better as compared to linear models. A decision tree model can also be used to explain the linear regression models

Q18. What is hyperparameter tuning in Machine Learning?

Ans. A machine learning consists of many parameters that tries to learn from the data. If we are using this method and not getting a very good accuracy then we can try another method which is known as Hyperparameters.

Hyperparameter tuning is also known as Hyperparameter Optimization.

Hyperparameter tuning is a technique of determining the best combination of hyperparameters that allows the machine learning model to maximize the model performance. The only way of getting the best accuracy of the model can be done by hyperparameter tuning. The hyperparameters are set before the actual training starts.

Q19. How to choose hyperparameters?

Ans. Choosing a correct hyperparameters is a very difficult task.

Manual Hyperparameter Tuning: In this process different types of hyperparameter are fixed manually. This can be done by selecting the hyperparameter randomly or with some kind of understanding of the model. This type of process can be very tedious and time consuming when the number of hyperparameter is large.

Automated Hyperparameter Tuning: Two of the most popular hyperparameter technique are Random Search and Grid Search. In this technique we create a dictionary of all the possible hyperparameters. In Random Search each iteration tries a random combination of hyperparameter and returns the best combination of the hyperparameter that can give us the best accuracy. In Grid Search each and every combination tries to find a hyperparameter that best suits the model and returns the best combination of the hyperparameter that can give us the best accuracy.

Q20. What are the various hyperparameters used in Decision Tree for tuning?

Ans. The hyperparameter used for tuning Decision Tree are as follows:

Criterion : {"gini", "entropy"}, default="gini"

Criterion indicates which type of splitting we want to perform. By default it is gini. But we can choose with gini or entropy where gini can be used for gini impurity and entropy can be used for information gain.

max_depth = int

Maximum depth the depth of the tree by default it is none and none means that the tree will expand all the leaves are pure or cannot expand.

random_state = int

Controls the randomness of the estimator.

min_samples_leaf = int or float

By default it is 1. The minimum number of samples required to be at the leaf node.

splitter : {"best", "random"}

In splitter we have two options either to choose best or random. By default it is best. Best is used to choose the best split and random to choose the best random split.

Q21. What are the Advantages of Decision Tree?

Ans. The advantages of Decision Tree are:

Decision Tree do not require much effort in data cleaning and data pre-processing as compared to other algorithm. Decision tree can

automatically handle missing values.

Clear and easy to interpret and visualize.

Decision trees are very simple and easy to understand as we check the diagram of a decision tree it is like a simple if-else statements.

One of the major advantages of decision tree is that it can be used for both classification and regression problems.

Decision tree can work upon both the continuous and categorical variable.

Decision tree is robust to outliers and it can handle outliers automatically.

The training time required by the decision tree is less as compared to the random forest as the decision tree creates only one tree and random forest creates number of trees.

Q22. What are the Disadvantages of Decision Tree?

Ans. The Disadvantages of Decision Tree are:

One of the major drawbacks of decision tree is overfitting.

Decision tree is unstable as adding new data points can lead to

regeneration of the whole tree which can be time consuming.
Decision tree is highly affected by noise which can make the decision tree unstable and in the end decision tree will do wrong predictions.
Decision tree does not perform well if we are predicting the outcome of the dependent variable which is continuous in nature.

Q23. What is Overfitting?

Ans. Overfitting is a scenario in which the model tries to fit the training data very closely but fails to fit the testing data. Overfitting occurs when the model learns each and every detail in the training data and the noise in the training data. The problem which occurs is that we try to pass the new data to the model to predict it gives a negative result. Overfitting also occurs if the model is too complex.

Q24. How to solve the problem of overfitting in Decision Tree?

Ans. There are various methods to tackle the problem of overfitting:

Pre-pruning: Pre-Pruning is a technique that is applied before training the model.

Post-Pruning: Post-Pruning is a technique that is applied after the decision tree is fully grown or after the training of the model.

Random Forest: If the decision tree model is overfitted we can use a random forest classifier as it uses more than one decision tree to overcome the problem of overfitting in the Decision Tree.

Cross-Validation: Cross Validation is a method that is used to train the model using the sub-set of the data and then evaluate it on the remaining data set. Cross-Validation can be used to solve the problem of overfitting in the Decision Tree.

Q25. What is Pre-Pruning and Post-Pruning?

Ans. Pre-Pruning and Post-Pruning are the two types of pruning techniques that are used to solve the problem of overfitting in decision tree:

Pre-Pruning: Pre-Pruning is nothing but a technique that refers to early stopping of the growth of the decision tree. The Pre-Pruning is also known as the hyperparameter technique which takes place before the training of the model. The hyperparameters which we use for Pre-Pruning are max_depth, min_samples_leaf, min_samples_split. Pre-Pruning is also used to handle the overfitting problem in the decision tree.

Post-Pruning: Post-Pruning is nothing but a technique that allows the decision tree to grow fully and after it is grown fully it then removes the tree branches. The technique which is used for post-pruning is known as cost complexity pruning(ccp). ccp_alpha is the parameter to achieve Post_Pruning. Post-Pruning is also used to handle the problem of overfitting in the Decision Tree.

Q26. What is Random Forest?

Ans. Random Forest is a Machine Learning Algorithm that comes under supervised learning. Random Forest can be used to solve both the classification problems and the regression problems. Random Forest is a type of Ensemble Learning model. Basically, Ensemble Learning is a technique which tries to combine multiple classifiers to solve the complex problem and improve the performance of the model. Random Forest uses multiple decision trees on various subsets of the training data and takes the average of all the decision tree classifiers. By combining all the decision tree classifiers random forest tries to improve the accuracy of the model. The greater the number of decision trees in the random forest the higher the chance of accuracy and random forest also tries to handle the problem of overfitting in the Decision Tree.

Q27. What is cross-validation?

Ans. Cross-validation is a method that is used to train the model using the subset of the data and then evaluate it on the remaining data set. Cross-Validation

can be used to solve the problem of overfitting in the Decision Tree.

One of the most famous cross-validation and mostly used cross-validation technique is K-fold cross validation. To perform K-fold cross validation first we divide our data k parts which is also known as k folds then we train our data on

all the k parts and leave one part($k-1$) for the testing purpose. Example if our $k = 10$ then we perform training on the 9 subsets and testing on the remaining one subset. As the value of k is equal to 10 we will iterate it 10 times keeping one subset reserved for testing.

Q28. What are the applications of Decision Tree?

Ans. The applications of Decision Tree are as follows:

Decision Tree Algorithm can be used in marketing which can help the businesses to expand and earn profit.

In understanding customers behaviours and releasing various offers which suits the customer behaviour.

In Diagnosis of Diseases and alignments decision tree can help the doctors in identifying the condition of the patients.

Decision trees can be used in detection of frauds by identifying the fraudulent behaviour. This can help company in saving a lot of resources.

Q29. What are the ways of evaluation of the Continuous Variable Decision tree?

Ans. The ways of Evaluation of the continuous variable decision tree are:

- a. R square
- b. Adjusted R square

c. RMSE (ROOT MEAN SQUARE ERROR)

Q30. What is R square?

ANS. R square is also described as the coefficient of determination. R square is used to determine the strength of correlation between the independent and the dependent variable. In simple terms R square lets us know how accurate our regression model is when compared to average. R square ranges between 0 to 1 higher the number the better is the accuracy or prediction of the model. If our R square is greater than 70% which is 0.7 indicated a good fit model.

Q31. What is Adjusted R square?

ANS. The Adjusted R square is a modified version of the R square. Adding more independent variables will result in an increased value of R square irrespective of whether the new independent variable is significant or not. But in the case of Adjusted R square if the new independent variable added is insignificant the adjusted r square has the capability to decrease therefore resulting in a better, more reliable, and accurate evaluation.

Q32. Difference between R²

and Adjusted R²?

ANS. Adding more independent variables will result in an increased value of R. This is the disadvantage of R square adding more independent variable irrespective of whether the new independent variable is significant or not the value of R square increases. But in the case of Adjusted R square if the new independent variable added is insignificant the adjusted r square has the capability to decrease therefore resulting in a better, more reliable, and accurate evaluation.

Q33. What is RMSE?

ANS. RMSE stands for ROOT MEAN SQUARE ERROR is a standard way to measure the error rate of the model. RMSE is a standard deviation of residuals or errors. Residuals or Errors are a measure of how far the data points are from the regression line. RMSE is a value that should be closer to 0.

Q34. What are the ways of evaluation of the Categorical Variable Decision tree?

Ans. The ways to evaluate the categorical variable decision tree are:

Confusion Matrix

Accuracy Score

Classification Report

Q35. What is Confusion Matrix?

Ans. Basically Confusion Matrix is used to evaluate the performance classification problems. If we have two classes in the target or dependent variable then the matrix will be 2 X 2.

In this 2 X 2 confusion matrix we have four values True Positive, True Negative, False Positive and False Negative this four values denotes that how many observation got correctly predicted, how many observation got misplaced or were not predicted correctly.

So with the help of the confusion matrix can understand how well the model is performing.

Important Terms of Confusion matrix are:

True Positive: The cases which were predicted 1 and actually it was 1.

True Negative: The cases which were predicted 0 and actually it was 0.

False Positive: The cases which were predicted 1 and actually it was 0.

False Negative: The cases which were predicted 0 and actually it was 1.

Q36. What is Accuracy Score?

Ans. Accuracy Score is nothing but the ratio or the calculation of the total number of correct predictions upon the total number of observations. It gives us a value which is known as the accuracy score which is used to evaluate the performance of the model.

Accuracy = Number of correct Predictions / Total number of observations

In terms of confusion matrix it is denoted by

Accuracy score = True Positive + True Negative / True Positive + True Negative + False Positive + False Negative

Q37. What is classification report?

Ans. As the name classification signifies it is a report for the classification models which helps us to evaluate the performance of the model. The classification report gives us the precision, recall, support, f1 score and the accuracy of the model.

Precision : Accuracy of the positive class.

Precision = True Positive / True Positive + False Positive

Recall : Total number of positive class that were correctly classified.

Recall = True Positive / True Positive + False Negative

F1_score is the harmonic mean of the precision and the recall.

Support is the number of observation of each classes.

Accuracy Score is nothing but the ratio or the calculation of the total number of correct predictions upon the total number of observations. It gives us a value which is known as the accuracy score which is used to evaluate the performance of the model.

Accuracy score = True Positive + True Negative / True Positive + True Negative + False Positive + False Negative

Q38. Steps for performing Decision Tree in Python.

Ans. Performing Decision Tree in Python:

1. Create a data frame properly --> pd.read_csv(), pd.read_excel()
2. Performing Exploratory Data Analysis(EDA) to better understand the data.
3. After performing Exploratory data analysis if there are some missing values then we need to compute if it is a numeric value then compute it by mean and if the value is categorical then compute it by mode.
4. After computing the missing values we need to convert the categorical variable into numerical using preprocessing.LabelEncoder().
5. Creating X and Y
6. Applying Standard Scaler if required.
7. Splitting the data --> train_test_split(), manual splitting
8. Importing the DecisionTreeClassifier model from sklearn.tree:
9. Build the model:
10. Create the model object --> modeldt = DecisionTreeClassifier()
11. Train the model --> modeldt.fit(X_train,Y_train)
12. Predict using the model --> Y_pred=modeldt.predict(X_test)
13. Evaluating the model using confusion_matrix, accuracy_score, classification_report when the problem is classification
14. Evaluating the model using R square, Adjusted R square and RMSE when the problem is regression.
15. Tuning the model --> Manual feature selection, Hyperparameter tuning.

Implementation of Decision tree

39. Description of the dataset: The dataset which we are using contains the information of the individuals like age, sex, education, marital status, capital gain, capital loss, income. Basically the last variable which is income is the dependent variable and all the other variables are independent so what we want to predict is that which individual income is >50k and which individual income is <=50k. It is a classification problem. So we will be using DecisionTreeClassifier() to solve this problem.

What are the libraries do you need to

Chapter 9 - Decision Tree and Random Forest

949. What is a decision tree, and how does it work?

A decision tree is a supervised machine-learning model that is used for both classification and regression tasks. It works by recursively partitioning the training data into subsets based on the values of the features. The goal is to create a tree that predicts the target variable as accurately as possible. The top node in the tree is the root node, which represents the entire dataset. The root node is then split into child nodes, which represent the subsets created by the split. This process is repeated recursively until a stopping criterion is met.

950. What are the advantages and disadvantages of using decision trees as a machine learning model?

Advantages:

Decision trees are easy to understand and interpret, making them useful for explaining the model to others.

Decision trees can handle both categorical and numerical data.

Decision trees are fast and efficient to train and can handle large datasets.

Decision trees can capture non-linear relationships between the features and the target variable.

Disadvantages:

Decision trees are prone to overfitting, which can lead to poor generalization performance.

Decision trees are sensitive to small changes in the training data, which can result in different trees being generated for the same dataset.

Decision trees can be biased towards features that have more levels or categories.

Decision trees can be unstable, meaning that a small change in the data can result in a completely different tree.

951. What are the different types of decision trees, and how do they differ from each other?

The two main types of decision trees are classification trees and regression trees. Classification trees are used for categorical target variables, while regression trees are used for continuous target variables. Another type of decision tree is the binary decision tree, which only has two possible outcomes at each node.

952. What is entropy, and how is it used in decision tree algorithms?

Entropy is a measure of the impurity or randomness of a set of examples. In the context of decision trees, entropy is used to measure the impurity of a node. A node with a low entropy has a high degree of homogeneity, meaning that most of the examples belong to the same class. A node with a high entropy has a low degree of homogeneity, meaning that the examples are distributed across multiple classes.

953. What is the difference between the Gini impurity and entropy as measures of impurity in decision trees?

Gini impurity and entropy are both measures of impurity in decision trees, but they differ in how they measure impurity. Gini impurity measures the probability of misclassifying an example in a given node, while entropy measures the average information content of the examples in a given node. In practice, the choice between Gini impurity and entropy as the impurity measure does not have a significant impact on the performance of the decision tree model.

954. How are decision trees used in regression problems, and what is the difference between regression trees and classification trees?

In regression problems, decision trees are used to predict a continuous target variable. The tree is constructed in a similar way to a classification tree, but instead of predicting a class at each leaf node, the tree predicts a continuous value. Regression trees are different from classification trees in that they predict a continuous value instead of a categorical value.

955. What is overfitting, and how can it be addressed in decision tree models?

Overfitting is a common problem in decision trees where the model is too complex and fits the noise in the training data rather than the underlying patterns. This can lead to poor performance on new, unseen data. Overfitting can be addressed in decision tree models by using techniques such as pruning, early stopping, and limiting the depth of the tree.

956. How are missing values handled in decision tree algorithms?

In decision tree algorithms, missing values can be handled by assigning probabilities to each possible value of the missing variable based on the observed values in the same node. The probabilities are then used to calculate the expected information gain for each variable as usual.

957. What is pruning, and how is it used to prevent overfitting in decision tree models?

Pruning is a technique used to prevent overfitting in decision tree models by removing branches that do not improve the performance of the model. Pruning can be done by removing individual branches, or by collapsing entire subtrees into single nodes

958. What are some common algorithms used to build decision trees, and how do they differ from each other?

Some common algorithms used to build decision trees include ID3, C4.5, CART, and CHAID. These algorithms differ in their splitting criteria, handling of missing data, and other

implementation details.

959. How can decision trees be used in combination with other machine learning models, such as random forests or gradient boosting machines?

Decision trees can be used in combination with other machine learning models by using them as base learners in ensemble methods such as random forests, gradient boosting machines, and AdaBoost. This can improve the performance and robustness of the model.

960. How can the performance of a decision tree model be evaluated, and what metrics are commonly used?

The performance of a decision tree model can be evaluated using metrics such as accuracy, precision, recall, F1 score, and ROC curves. Cross-validation can also be used to estimate the performance of the model on new data

961. What is the bias-variance tradeoff in machine learning, and how does it apply to decision trees?

The bias-variance tradeoff is a fundamental concept in machine learning that refers to the tradeoff between underfitting and overfitting. In the context of decision trees, increasing the depth of the tree can reduce bias (i.e., improve the fit to the training data) but also increase variance (i.e., make the model more sensitive to noise in the data).

962. How can decision trees be used for feature selection, and what advantages does this approach offer?

Decision trees can be used for feature selection by measuring the importance of each feature in the tree, and then selecting the most important features for use in other models. This approach can reduce the dimensionality of the data and improve the interpretability of the model.

963. What are some limitations of decision trees, and when might they not be the best choice of model for a given problem?

Some limitations of decision trees include their sensitivity to small changes in the data, their tendency to overfit to noisy or irrelevant features, and their inability to capture complex interactions between variables. Decision trees may also be less effective on high-dimensional data or data with highly correlated features. In some cases, other machine learning models such as neural networks or support vector machines may be a better choice.

964. List down some popular algorithms used for deriving Decision Trees along with their attribute selection measures.

Some of the popular algorithms used for constructing decision trees are:

1. ID3 (Iterative Dichotomiser): Uses Information Gain as attribute selection measure.
2. C4.5 (Successor of ID3): Uses Gain Ratio as attribute selection measure.
3. CART (Classification and Regression Trees) – Uses Gini Index as attribute selection measure.

965. Explain the CART Algorithm for Decision Trees.

The CART stands for Classification and Regression Trees is a greedy algorithm that greedily searches for an optimum split at the top level, then repeats the same process at each of the subsequent levels.

Moreover, it does verify whether the split will lead to the lowest impurity or not as well as the solution provided by the greedy algorithm is not guaranteed to be optimal, it often produces a solution that's reasonably good since finding the optimal Tree is an NP-Complete problem that requires exponential time complexity.

As a result, it makes the problem intractable even for small training sets. This is why we must go for a “reasonably good” solution instead of an optimal solution

966. List down the attribute selection measures used by the ID3 algorithm to construct a Decision Tree.

The most widely used algorithm for building a Decision Tree is called ID3. ID3 uses Entropy and Information Gain as attribute selection measures to construct a Decision Tree.

1. Entropy: A Decision Tree is built top-down from a root node and involves the partitioning of data into homogeneous subsets. To check the homogeneity of a sample, ID3 uses entropy. Therefore, entropy is zero when the sample is completely homogeneous, and entropy of one when the sample is equally divided between different classes.
2. Information Gain: Information Gain is based on the decrease in entropy after splitting a dataset based on an attribute. The meaning of constructing a Decision Tree is all about finding the attributes having the highest information gain.

967. Briefly explain the properties of Gini Impurity.

Let X (discrete random variable) takes values y_+ and y_- (two classes). Now, let's consider the different cases:

Case- 1: When 100% observations belong to y_+ . Then, the Gini impurity of the system would be: –

Decision Trees Questions gini impurity

Case- 2: When 50% observations belong to y_+ . Then, the Gini impurity of the system would be:

—

Decision Trees Questions gini impurity example

Case- 3: When 0% observations belong to y_+ . Then, the Gini impurity of the system would be:

968. Briefly explain the properties of Gini Impurity.

Gini impurity is a measure of the impurity or homogeneity of a set of labels in a binary classification problem. It measures the probability of misclassifying a randomly chosen element in the dataset if it were randomly labeled according to the distribution of labels in the set.

Some of the properties of Gini impurity are:

It ranges from 0 to 0.5, with 0 indicating perfect homogeneity (all elements have the same label) and 0.5 indicating perfect heterogeneity (an equal number of elements in each class).
It is computationally efficient to calculate and is less sensitive to outliers compared to entropy.
It tends to be biased towards larger partitions and is not recommended for problems where the size of the partitions is not proportional to their homogeneity.
It can be used for both binary and multiclass classification problems, but it is less informative than other measures in the case of multiclass problems.

969. Measurement of the performance of decision tree algorithm.

The performance of a decision tree algorithm can be measured using various metrics. Some of the commonly used metrics include:

Accuracy: This is the most basic metric used to measure the performance of a decision tree algorithm. It simply measures the percentage of correctly classified instances.

Precision: Precision measures the percentage of true positive instances among all instances that were classified as positive.

Recall: Recall measures the percentage of true positive instances that were correctly classified as positive among all instances that are actually positive.

F1 score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of both metrics.

Area under the ROC curve (AUC): AUC is a metric that measures the ability of a decision tree algorithm to distinguish between positive and negative instances. It is the area under the receiver operating characteristic curve.

Confusion matrix: A confusion matrix is a matrix that shows the actual and predicted class labels for each instance. It is useful for understanding the types of errors made by the algorithm.

These metrics can be used to evaluate the performance of the decision tree algorithm on a validation set or in cross-validation. By comparing the performance of different decision tree algorithms, or the same algorithm with different hyperparameters or features, we can determine the best approach for a given problem.

970. What is F1 Score?

The F1 score is a measure of a model's accuracy that considers both precision and recall. It is the harmonic mean of precision and recall, where precision is the proportion of true positives among all positive predictions and recall is the proportion of true positives among all actual positives.

971. What is Recall?

Recall, also known as sensitivity or true positive rate, is a measure of a model's ability to identify all positive instances correctly. It is the proportion of true positives among all actual positives.

972. What AUC?

AUC, or the area under the ROC (Receiver Operating Characteristic) curve, is a measure of a model's ability to distinguish between positive and negative instances. The ROC curve plots the true positive rate against the false positive rate at various classification thresholds, and the AUC represents the area under this curve. A model with an AUC of 1.0 is perfect, while a model with an AUC of 0.5 performs no better than random chance.

973. What is a confusion matrix?

A confusion matrix is a table that summarizes the performance of a classification model by comparing the actual labels of a set of instances to the predicted labels produced by the model. The table typically has four cells: true positive, false positive, false negative, and true negative. From the confusion matrix, various metrics such as accuracy, precision, recall, F1 score, and AUC can be calculated.

974. Can decision trees handle categorical variables? If so, how are they treated?

Yes, decision trees can handle categorical variables. Categorical variables are usually converted into dummy variables (also known as one-hot encoding), which are binary variables representing the presence or absence of a particular category.

975. How does the depth of a decision tree affect its performance and complexity?

The depth of a decision tree can affect its performance and complexity in several ways. A deeper tree may fit the training data better, but it may also overfit the data and perform poorly on new, unseen data. A shallower tree may be simpler and easier to interpret, but it may also underfit the data and not capture all the relevant patterns in the data.

976. How can decision trees be used to perform feature selection, and what advantages does this approach offer over other feature selection methods?

Decision trees can be used for feature selection by evaluating the importance of each feature in the tree-building process. The importance of a feature can be measured by the decrease in impurity (e.g., Gini impurity or entropy) that results from splitting on that feature. Features with higher importance are considered more informative and can be used for feature selection.

One advantage of using decision trees for feature selection is that they can handle both continuous and categorical features. Decision trees can also capture complex nonlinear relationships between features, which may not be captured by linear methods such as correlation or regression.

977. What is the difference between a greedy algorithm and an exhaustive search algorithm for building decision trees?

A greedy algorithm for building decision trees makes locally optimal choices at each node of the tree, without considering the global optimal solution. In other words, it chooses the feature and split point that maximally reduces impurity at each node, without looking ahead to future nodes.

An exhaustive search algorithm, on the other hand, considers all possible splits of all possible features at each node of the tree, and chooses the best split. This approach is computationally expensive and may lead to overfitting, especially in high-dimensional feature spaces.

978. Can decision trees be used for time-series forecasting or sequence prediction problems? If so, how are they adapted?

Yes, decision trees can be adapted for time-series forecasting or sequence prediction problems. One common approach is to use sliding windows to transform the time-series data into a tabular format, where each row represents a window of time-series data, and each column represents a feature.

Another approach is to use decision trees in an ensemble method, such as random forests or gradient boosting, which can handle time-series data more effectively by incorporating information from multiple trees or iterations.

979. What is Pre-Pruning and Post-Pruning?

Ans. Pre-Pruning and Post-Pruning are the two types of pruning techniques that are used to solve the problem of overfitting in decision tree:

Pre-Pruning: Pre-Pruning is nothing but a technique that refers to early stopping of the growth of the decision tree. The Pre-Pruning is also known as the hyperparameter technique which takes place before the training of the model. The hyperparameters which we use for Pre-Pruning are max_depth, min_samples_leaf, min_samples_split. Pre-Pruning is also used to handle the overfitting problem in the decision tree.

Pruning: Post-Pruning is nothing but a technique that allows the decision tree to grow fully and after it is grown fully it then removes the tree branches. The technique which is used for post-pruning is known as cost complexity pruning(ccp). ccp_alpha is the parameter to achieve Post_Pruning. Post-Pruning is also used to handle the problem of overfitting in the Decision Tree.

980. How to solve the problem of overfitting in Decision Tree?

Ans. There are various methods to tackle the problem of overfitting:

Pre-pruning: Pre-Pruning is a technique that is applied before training the model.

Post-Pruning: Post-Pruning is a technique that is applied after the decision tree is fully grown or after the training of the model.

Random Forest: If the decision tree model is overfitted we can use a random forest classifier as it uses more than one decision tree to overcome the problem of overfitting in the Decision Tree.

Cross-Validation: Cross Validation is a method that is used to train the model using the sub-set of the data and then evaluate it on the remaining data set. Cross-Validation can be used to solve the problem of overfitting in the Decision Tree.

981. What is Pre-Pruning and Post-Pruning?

Ans. Pre-Pruning and Post-Pruning are the two types of pruning techniques that are used to solve the problem of overfitting in decision tree:

Pre-Pruning: Pre-Pruning is nothing but a technique that refers to early stopping of the growth of the decision tree. The Pre-Pruning is also known as the hyperparameter technique which takes place before the training of the model. The hyperparameters which we use for Pre-Pruning are `max_depth`, `min_samples_leaf`, `min_samples_split`. Pre-Pruning is also used to handle the overfitting problem in the decision tree.

Post Pruning: Post-Pruning is nothing but a technique that allows the decision tree to grow fully and after it is grown fully it then removes the tree branches. The technique which is used for post-pruning is known as cost complexity pruning(ccp). `ccp_alpha` is the parameter to achieve Post_Pruning. Post-Pruning is also used to handle the problem of overfitting in the Decision Tree.

982. What are the ways of evaluation of the Continuous Variable Decision Tree?

Ans. The ways of Evaluation of the continuous variable decision tree are:

- a. R square
- b. Adjusted R square
- c. RMSE (ROOT MEAN SQUARE ERROR)

983. What is Accuracy Score?

Ans. Accuracy Score is nothing but the ratio or the calculation of the total number of correct predictions upon the total number of observations. It gives us a value which is known as the accuracy score which is used to evaluate the performance of the model.

Accuracy = Number of correct Predictions / Total number of observations

In terms of confusion matrix it is denoted by

Accuracy score = True Positive + True Negative / True Positive + True Negative + False Positive + False Negative

984. What is classification report?

Ans. As the name classification signifies it is a report for the classification models which helps us to evaluate the performance of the model. The classification report gives us the precision, recall, support, f1 score and the accuracy of the model.

Precision : Accuracy of the positive class.

Precision = True Positive / True Positive + False Positive

Recall : Total number of positive class that were correctly classified.

Recall = True Positive / True Positive + False Negative

985. F1_score is the harmonic mean of the precision and the recall.

Support is the number of observation of each classes.

Accuracy Score is nothing but the ratio or the calculation of the total number of correct predictions upon the total number of observations. It gives us a value which is known as the accuracy score which is used to evaluate the performance of the model.

Accuracy score = True Positive + True Negative / True Positive + True Negative + False Positive + False Negative

986. What are the Advantages of Decision Tree?

Ans. The advantages of Decision Tree are:

Decision Tree do not require much effort in data cleaning and data pre-processing as compared to other algorithm. Decision tree can

automatically handle missing values.

Clear and easy to interpret and visualize.

Decision trees are very simple and easy to understand as we check the diagram of a decision tree it is like a simple if-else statements.

One of the major advantages of decision tree is that it can be used for both classification and regression problems.

Decision tree can work upon both the continuous and categorical variable.

Decision tree is robust to outliers and it can handle outliers automatically.

The training time required by the decision tree is less as compared to the random forest as the decision tree creates only one tree and random forest creates number of trees.

987. What are the Disadvantages of Decision Tree?

Ans. The Disadvantages of Decision Tree are:

One of the major drawbacks of decision tree is overfitting.

Decision tree is unstable as adding new data points can lead to regeneration of the whole tree which can be time consuming.

Decision tree is highly affected by noise which can make the decision tree unstable and in the end decision tree will do wrong predictions.

Decision tree does not perform well if we are predicting the outcome of the dependent variable which is continuous in nature.

988. Which is better Linear model or Tree-Based-Model?

Ans. Basically, it depends that what kind of data we are working upon. If we are working upon a dataset in which the relationship between the

dependent variable and independent variable is very high in such kind of data the linear regression will perform very well but the tree-based model will perform poorly.

Suppose we are working upon a dataset where there is high non-linearity and complex relation between the dependent and the independent variable

in this case the tree-based model will perform very well and the linear regression will fail.

Suppose we are working upon a dataset where we want to build a model which can be easily explained to the people in such cases decision tree model will always perform better as compared to linear models. A decision tree model can also be used to explain the linear regression models

989. What is hyperparameter tuning in Machine Learning?

Ans. A machine learning consists of many parameters that tries to learn from the data. If we are using this method and not getting a very good accuracy then we can try another method which is known as Hyperparameters.

Hyperparameter tuning is also known as Hyperparameter Optimization.

Hyperparameter tuning is a technique of determining the best combination of hyperparameters that allows the machine learning model to maximize the model performance. The only way of getting the best accuracy of the model can be done by hyperparameter tuning. The hyperparameters are set before the actual training starts.

990. What are the various hyperparameters used in Decision Tree for tuning?

Ans. The hyperparameter used for tuning Decision Tree are as follows:

`Criterion : {"gini", "entropy"}, default="gini"`

Criterion indicates which type of splitting we want to perform. By default it is gini. But we can choose with gini or entropy where gini can be used for gini impurity and entropy can be used for information gain.

`max_depth = int`

Maximum depth the depth of the tree by default it is none and none means that the tree will expand all the leaves are pure or cannot expand.

`random_state = int`

Controls the randomness of the estimator.

`min_samples_leaf = int or float`

By default it is 1. The minimum number of samples required to be at the leaf node.

`splitter : {"best", "random"}`

In splitter we have two options either to choose best or random. By default it is best. Best is used to choose the best split and random to choose the best random split.

991. Steps for performing Decision Tree in Python.

Ans. Performing Decision Tree in Python:

1. Create a data frame properly --> pd.read_csv(), pd.read_excel()
2. Performing Exploratory Data Analysis(EDA) to better understand the data.
3. After performing Exploratory data analysis if there are some missing values then we need to compute if it is a numeric value then compute it by mean and if the value is categorical then compute it by mode.
4. After computing the missing values we need to convert the categorical variable into numerical using preprocessing.LabelEncoder().
5. Creating X and Y
6. Applying Standard Scaler if required.
7. Splitting the data --> train_test_split(), manual splitting
8. Importing the DecisionTreeClassifier model from sklearn.tree:
9. Build the model:
10. Create the model object --> modeldt = DecisionTreeClassifier()
11. Train the model --> modeldt.fit(X_train,Y_train)
12. Predict using the model --> Y_pred=modeldt.predict(X_test)
13. Evaluating the model using confusion_matrix, accuracy_score, classification_report when the problem is classification
14. Evaluating the model using R square, Adjusted R square and RMSE when the problem is regression.
15. Tuning the model --> Manual feature selection, Hyperparameter tuning.

Random Forest

992. What is a random forest?

A random forest is an ensemble learning method that is used for both classification and regression tasks. It consists of a large number of decision trees, which are created by randomly selecting subsets of the features and samples. The final output is determined by aggregating the predictions of all the decision trees.

993. What is the difference between a decision tree and a random forest?

A decision tree is a single tree-based model that is used for both classification and regression tasks. It is built by recursively splitting the data into smaller and smaller subsets based on the most discriminative features. On the other hand, a random forest is an ensemble of decision trees, where each tree is built using a random subset of the features and samples.

994. How does a random forest work?

A random forest works by creating multiple decision trees and then combining their outputs to make a final prediction. Each decision tree is built on a random subset of the features and samples, which reduces overfitting and increases the generalization of the model. The final prediction is made by aggregating the outputs of all the decision trees.

995. What are the advantages of using a random forest algorithm?

The advantages of using a random forest algorithm are:

- It can handle both regression and classification tasks.
- It can work with a large number of input features, even if some of them are irrelevant.
- It is less prone to overfitting, which means it can generalize well to new data.
- It is relatively easy to use and requires minimal tuning of hyperparameters.

996. What are some of the parameters that can be tuned in a random forest algorithm?

Some of the parameters that can be tuned in a random forest algorithm are:

- The number of decision trees in the forest.
- The maximum depth of each decision tree.
- The minimum number of samples required to split an internal node.
- The minimum number of samples required to be at a leaf node.
- The number of features to consider when looking for the best split.
- The criterion used to evaluate the quality of a split (e.g., Gini impurity or information gain).

997. What is bagging in the context of a random forest?

Bagging stands for bootstrap aggregating and is a technique used in ensemble learning methods like random forests. It involves generating multiple random subsets of the training data and using them to train multiple models. In the context of a random forest, bagging is used to create multiple decision trees by training each tree on a random subset of the features and samples.

998. How does the random forest algorithm handle missing data?

The random forest algorithm can handle missing data by using two approaches:

Imputing the missing data: One approach is to impute the missing values with a default value such as the mean or median of the feature. Alternatively, a more sophisticated method like k-nearest neighbors can be used to impute the missing values.

Ignoring the missing data: Another approach is to simply ignore the missing data and train the model on the available data. This can be done by assigning a special category or value to represent the missing data.

999. What are some of the techniques used to evaluate the performance of a random forest model?

Some of the techniques used to evaluate the performance of a random forest model are:

Cross-validation: A common technique used to evaluate the performance of a random forest model is k-fold cross-validation, where the data is split into k subsets and the model is trained and tested on each subset.

Out-of-bag error: Since each decision tree in a random forest is built on a different subset of the data, the remaining samples can be used to calculate the out-of-bag error, which can be a good estimate of the generalization error.

Feature importance: Feature importance is a metric that measures the relative importance of each feature in a random forest model. This can be used to identify the most important features and to eliminate irrelevant or redundant features.

1000. What is the feature importance in a random forest model?

Feature importance is a metric that measures the relative importance of each feature in a random forest model. It is calculated by averaging the importance of each feature over all the decision trees in the forest. The importance of each feature is typically calculated using metrics such as the Gini importance or the mean decrease impurity, which measure the contribution of each feature to the reduction in impurity during the training process.

Feature importance can be used to identify the most important features in the data and to eliminate irrelevant or redundant features, which can lead to a more efficient and accurate model.

1001. What are the limitations of the random forest algorithm?

Some of the limitations of the random forest algorithm are:

It can be difficult to interpret the output of a random forest, especially if the number of decision trees is large.

It can be computationally expensive, especially if the number of input features is large.

It can overfit if the number of decision trees is too large or if the hyperparameters are not well-tuned.

It may not perform well on data with a large number of classes or imbalanced classes.

It may not work well with high-dimensional data or data with complex interactions between features.

1002. Does Random Forest need Pruning? Why or why not?

Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances.

Random Forest usually does not require pruning because it will not over-fit like a single decision tree. This happens due to the fact that the trees are bootstrapped and that multiple random trees use random features so the individual trees are strong without being correlated with each other.

1003.. Explain how the Random Forests give output for Classification, and Regression problems?

Classification: The output of the Random Forest is the one selected by the most trees.

Regression: The output of the Random Forest is the mean or average prediction of the individual trees.

1004. How is a Random Forest related to Decision Trees?

Random forest is an ensemble learning method that works by constructing a multitude of decision trees. A random forest can be constructed for both classification and regression tasks. Random forest outperforms decision trees, and it also does not have the habit of overfitting the data as decision trees do.

A decision tree trained on a specific dataset will become very deep and cause overfitting. To create a random forest, decision trees can be trained on different subsets of the training dataset, and then the different decision trees can be averaged with the goal of decreasing the variance.

1005. How would you find the optimal size of the Bootstrapped Dataset?

Due to the observations being sampled with replacements, even if the size of the bootstrapped dataset is different, the datasets will be different.

Due to this, the full size of the training data can be used.

1006. What are Ensemble Methods?

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

Random Forest is a type of ensemble method.

The number of component classifier in an ensemble has a great impact on the accuracy of the prediction, although there is a law of diminishing results in ensemble construction.

1007. Explain the advantages of using Random Forest

Random Forest is very versatile and can be used in both regression and classification tasks. It can also handle all binary, categorical, and numerical features.

The process is parallelizable where the process can be split to run in different machines.

It performs better in high dimensionality since the work is on subsets of data.

The training speed is faster than decision trees because they are working only on a subset of features. Even if there are hundreds of features the training speed will be significantly faster.

The Random Forest is good at balancing errors for class population unbalanced data sets.

It has low bias, but moderate variance because when all the trees are averaged in random forest, all the variances are also averaged so it has low bias but a moderate variance.

1008. How does Random Forest handle missing values?

The Random Forest methods encourage two ways of handling missing values:

Drop data points with missing values. This is not recommended due to the fact that all the available data points is not used.

Fill in the missing values with the median (for numerical values) or mode (for categorical values). This method will brush too broad a stroke for datasets with many gaps and significant structure.

There are other methods of filling in missing values such as calculating the similarity between the missing features, and the missing values estimated by weighting.

1009. How is it possible to perform Unsupervised Learning with Random Forest?

As part of their construction, random forest predictors naturally lead to a dissimilarity measure among the observations. One can also define a random forest dissimilarity measure between unlabeled data:

the idea is to construct a random forest predictor that distinguishes the observed data from suitably generated synthetic data.

Many unsupervised learning methods require the inclusion of an input dissimilarity measure among the observations. Hence, if a dissimilarity matrix can be produced using Random Forest, unsupervised learning can be successfully implemented. The patterns found in the process will be used to make clusters.

1010. How would you improve the performance of Random Forest?

Some things to try to improve the performance of Random Forest are:

Using a higher quality dataset and feature engineering. Using too many features and data are not good for the model so sometimes it is important to perform some feature reduction too.

Tuning the hyperparameters of the algorithm.

Trying different algorithms.

1011. What are proximities in Random Forests?

Proximity is the closeness or nearness between pairs of cases.

Proximities are calculated for each pair of cases/observations/sample points. If two cases occupy the same terminal node through one tree, their proximity is increased by one. At the end of the run of all trees, the proximities are normalized by dividing by the number of trees.

Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

1012. What does Random refer to in Random Forest?

Random forest is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Hence, random forest is Random in the following ways:

Each tree is trained on a random subset of features, which ensures low correlation among decision trees.

Each tree in the forest is trained in 2/3-rd of the total training data and data points are drawn at random from the original dataset.

1013. What is Entropy?

The basic definition of entropy is a measure of disorder. The equation of entropy is as follows

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

where,

p is the probability of frequentist probability of the element i in the data.

In machine learning models, the goal is to decrease uncertainties. Thus, the models should have less entropy.

The reduction of entropy is defined as information gain. It is shown below:

$$IG(Y, X) = E(Y) - E(Y|X)$$

Information gain is just the subtraction of entropy of Y given X from the entropy of just Y . $E(Y|X)$ corresponds to the information of Y that we already know. So, $E(Y|X)$ is not new information for the model.

1014. Why Random Forest models are considered not interpretable?

Decision trees can be easily converted into rules which increase human interpretability of the results and explain why a decision was made.

For Random Forest the general recommendation is to use as many trees as possible. In most cases, with hundreds of trees, you wouldn't be able to understand why did they collectively made the decision that they made.

1015. Why is the training efficiency of Random Forest better than Bagging?

The difference between Random Forest and Bagging is the fact that for Random Forest only a subset of features out of all are selected in random and the best split feature from the subset is used to split each node in a tree.

In bagging all the features are considered in splitting the node.

Due to the fact that bagging considers all the features, the training efficiency of random forest is better.

1016. Implement Random Forest in Python

```
# Import required libraries
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.datasets import load_iris

# Load iris dataset
iris = load_iris()

# Split the dataset into training and testing data
X_train, X_test, y_train, y_test = train_test_split(iris.data,
iris.target, test_size=0.2, random_state=42)

# Create a random forest classifier with 100 trees
rf = RandomForestClassifier(n_estimators=100, random_state=42)

# Train the model using the training data
rf.fit(X_train, y_train)

# Make predictions on the test data
y_pred = rf.predict(X_test)

# Calculate the accuracy of the model
accuracy = accuracy_score(y_test, y_pred)

# Print the accuracy
print("Accuracy:", accuracy)
```

In this example, we first load the iris dataset and split it into training and testing data using the `train_test_split` function. We then create a `RandomForestClassifier` object with 100 trees and train the model using the training data. Finally, we make predictions on the test data and calculate the accuracy of the model using the `accuracy_score` function.

1017. What is `n_estimators` in Random forest?

`n_estimators` is a hyperparameter in the random forest algorithm that represents the number of decision trees to be used in the ensemble. Each decision tree in the forest is trained on a different subset of the training data, using a different subset of the features. The predictions of all the decision trees are then combined to make the final prediction.

Increasing the number of decision trees in the forest can improve the performance of the model by reducing the variance, but it can also increase the computational cost and the risk of overfitting. The optimal value of `n_estimators` depends on the size and complexity of the

dataset, as well as the other hyperparameters of the model, such as `max_depth`, `min_samples_split`, and `max_features`.

In practice, a common value for `n_estimators` is 100, but this can vary depending on the problem at hand. It is often a good idea to experiment with different values of `n_estimators` and other hyperparameters using techniques such as cross-validation to find the optimal values for a given problem.

1018. What is `random_state` in random forest?

`random_state` is a hyperparameter in the random forest algorithm (as well as other machine learning algorithms) that is used to set the random seed for the random number generator. This is important because random forest models can behave differently each time they are trained on the same data due to the randomness involved in the process, such as the random selection of features and samples for each decision tree.

Setting the `random_state` parameter to a fixed value ensures that the model will behave in a consistent way each time it is trained, which can make it easier to reproduce the results and compare different models. For example, if you set `random_state` to 42 and train a random forest model on a dataset, the model will always use the same random seed and produce the same results, as long as the dataset and other hyperparameters are kept constant.

It is important to note that the choice of the value for `random_state` does not affect the performance of the model, but rather only the consistency of the results. If `random_state` is not set, it defaults to `None`, which means that a random seed will be used every time the model is trained.

1019. What are the hyperparameters in Random Forest?

Hyperparameters are adjustable parameters that determine the behavior of the random forest model during training. Some of the key hyperparameters in the random forest algorithm are:

`n_estimators`: The number of decision trees in the forest.

`max_depth`: The maximum depth of each decision tree.

`min_samples_split`: The minimum number of samples required to split an internal node.

`min_samples_leaf`: The minimum number of samples required to be at a leaf node.

`max_features`: The maximum number of features to consider when making a split.

`criterion`: The function used to measure the quality of a split.

`bootstrap`: Whether or not to use bootstrap samples when building the decision trees.

`random_state`: The seed used by the random number generator.

The optimal values of these hyperparameters depend on the specific dataset and problem being solved. In practice, it is common to use techniques such as grid search and random search to tune the hyperparameters and find the optimal values for a given problem.

It is also worth noting that the number of decision trees (`n_estimators`) is the most important hyperparameter in a random forest model and increasing it often leads to better performance, up to a certain point. The other hyperparameters tend to have less impact on the performance, but can still have an effect, especially when the dataset is small or noisy.

1020. What is `max_depth` in Random Forest ?

`max_depth` is a hyperparameter in the random forest algorithm that determines the maximum depth of each decision tree in the ensemble. It controls the complexity of the trees and can help prevent overfitting. If the `max_depth` is too large, the model may overfit the training data, while if it is too small, the model may not capture enough of the complexity in the data.

1021. What is `min_samples_split` in Random Forest?

`min_samples_split` is a hyperparameter in the random forest algorithm that determines the minimum number of samples required to split an internal node in a decision tree. It controls the trade-off between model complexity and underfitting. If `min_samples_split` is too small, the decision tree may overfit the data and capture noise, while if it is too large, the tree may not capture enough of the complexity in the data.

1022. What is `min_samples_leaf` in Random Forest?

`min_samples_leaf` is a hyperparameter in the random forest algorithm that determines the minimum number of samples required to be at a leaf node in a decision tree. It also controls the trade-off between model complexity and underfitting. If `min_samples_leaf` is too small, the decision tree may overfit the data and capture noise, while if it is too large, the tree may not capture enough of the complexity in the data.

1023. What is `max_features` in Random Forest?

`max_features` is a hyperparameter in the random forest algorithm that determines the maximum number of features to consider when making a split in a decision tree. It controls the randomness of the model and can help prevent overfitting. If `max_features` is too large, the model may capture noise in the data, while if it is too small, the model may not capture enough of the complexity in the data.

1024. What is the criterion in Random Forest?

criterion is a hyperparameter in the random forest algorithm that determines the function used to measure the quality of a split in a decision tree. The two most common criteria are "gini" and "entropy", which measure the impurity of the classes in the nodes. The choice of criterion can have an impact on the performance of the model, but the effect is typically small.

1025. What is bootstrap in Random Forest?

bootstrap is a hyperparameter in the random forest algorithm that determines whether or not to use bootstrap samples when building the decision trees in the ensemble. If bootstrap is set to True (the default), each tree is built on a random subset of the training data, with replacement. This introduces randomness into the model and can help prevent overfitting. If bootstrap is set to False, each tree is built on the entire training data, which can make the model more stable but also less diverse.

1026. Implement Random Forest in R.

```
# Load the randomForest package
library(randomForest)

# Load the iris dataset
data(iris)

# Split the data into training and test sets
set.seed(123)
train_index <- sample(nrow(iris), nrow(iris)*0.7)
train_data <- iris[train_index, ]
test_data <- iris[-train_index, ]

# Train a random forest model
rf_model <- randomForest(Species ~ ., data = train_data, ntree = 500)

# Make predictions on the test data
predictions <- predict(rf_model, test_data)

# Evaluate the model performance
table(predictions, test_data$Species)
```

In this example, we first load the randomForest package and the iris dataset. We then split the data into training and test sets, using 70% of the data for training and the remaining 30% for testing.

We then train a random forest model using the randomForest function. The first argument to this function is the formula that specifies the target variable (Species) and the predictor variables (all

other columns in the data). The second argument is the training data, and the ntree parameter specifies the number of trees in the forest (500 in this case).

We then use the predict function to make predictions on the test data using the trained model. Finally, we evaluate the model performance by comparing the predicted values to the true values using the table function.

1027. Working of Random Forest Classifier?

A Random Forest Classifier is an ensemble model that combines multiple decision trees to improve the overall accuracy and robustness of the model. Here's how a Random Forest Classifier works:

Data preparation: The first step is to prepare the data for the model. This typically involves cleaning and preprocessing the data, such as handling missing values and converting categorical variables to numerical ones.

Random sampling: A random sample of the data is selected for each tree in the forest. This process is known as bootstrapping, and it involves sampling the data with replacement. This means that some data points may be included multiple times, while others may not be included at all.

Building decision trees: A decision tree is built using the sampled data. Each node in the tree represents a feature, and the branches represent the possible values of that feature. The tree is built recursively by selecting the best feature to split the data based on some criterion, such as information gain or Gini impurity.

Repeat steps 2 and 3: Steps 2 and 3 are repeated to create multiple decision trees. The number of trees is specified by the n_estimators hyperparameter.

Predictions: To make a prediction for a new data point, the Random Forest Classifier combines the predictions of all the trees in the forest. Each tree makes a prediction, and the class with the most votes is chosen as the final prediction.

Hyperparameter tuning: The Random Forest Classifier has several hyperparameters that can be tuned to improve its performance. These include n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features.

The main advantage of a Random Forest Classifier is that it is more accurate and less prone to overfitting than a single decision tree. This is because it averages the predictions of multiple trees, which reduces the impact of any individual tree's errors. Additionally, the random sampling and feature selection process helps to decorrelate the trees and improve the robustness of the model.

1028. Working of Random Forest Regresor

A Random Forest Regressor is similar to a Random Forest Classifier, except that it is used for regression problems instead of classification problems. The steps for building a Random Forest Regressor are the same as for a Random Forest Classifier, except that the prediction for a new data point is the average of the predictions of all the trees in the forest. The output of the model is a continuous numerical value, rather than a discrete class label.

1029. Grid Search in Hyperparameter tuning?

Grid Search is a technique for hyperparameter tuning that involves evaluating a model's performance for a range of hyperparameter values. The hyperparameters are specified in a grid, and the model is trained and evaluated for every combination of hyperparameters in the grid. The combination of hyperparameters that results in the best performance is selected as the optimal set of hyperparameters. Grid Search is a brute-force approach that can be computationally expensive, but it is guaranteed to find the best set of hyperparameters within the specified grid.

1030. Random Search in hyperparameter tuning?

Random Search is an alternative technique for hyperparameter tuning that involves randomly sampling hyperparameters from a distribution. The distribution can be uniform or non-uniform, and it can be defined for each hyperparameter separately. The model is trained and evaluated for a fixed number of randomly sampled hyperparameter combinations, and the combination that results in the best performance is selected as the optimal set of hyperparameters. Random Search is more efficient than Grid Search in high-dimensional hyperparameter spaces, as it does not evaluate every possible combination of hyperparameters. However, it is not guaranteed to find the best set of hyperparameters, as it relies on random sampling.

1031. What is the impact of correlated features on a Random Forest model, and how can this issue be addressed?

Correlated features can have a negative impact on the performance of a Random Forest model. In the presence of correlated features, the trees in the forest tend to be similar to each other, which can lead to over-representation of certain features in the model and an overall reduction in the diversity of the forest. This can result in lower accuracy and higher variance of the model.

To address this issue, several techniques can be used. One approach is to perform feature selection to remove one or more of the correlated features before training the Random Forest model. Another approach is to use techniques like Principal Component Analysis (PCA) or Independent Component Analysis (ICA) to transform the features into a set of uncorrelated components before training the model. Finally, feature bagging can be used to create different subsets of the original feature set and train a separate Random Forest model on each subset, which can help to reduce the correlation among the trees.

1032. Can a Random Forest model suffer from overfitting? If so, how can overfitting be avoided or mitigated?

Yes, a Random Forest model can suffer from overfitting if the model is too complex or if the trees in the forest are allowed to grow too deep. When a Random Forest model overfits, the model becomes too specialized to the training data and fails to generalize well to new data.

To avoid or mitigate overfitting in a Random Forest model, several techniques can be used. One approach is to limit the depth of the trees in the forest, which can help to reduce the complexity of the model and prevent overfitting. Another approach is to increase the minimum number of samples required to split a node or to create a leaf, which can also help to reduce the complexity of the model. Additionally, feature bagging and tree bagging can be used to increase the diversity of the forest and reduce the variance of the model, which can also help to prevent overfitting. Finally, cross-validation can be used to evaluate the performance of the model on new data and select the optimal set of hyperparameters.

1033. How can the importance of individual trees in a Random Forest model be measured and used to improve the overall model?

The importance of individual trees in a Random Forest model can be measured by examining the decrease in impurity or information gain associated with each split point in each tree. This information can be aggregated across all trees in the forest to obtain a measure of feature importance. The feature importance can be used to improve the overall model by identifying the most important features and removing less important features, or by selecting a smaller subset of features to use for training the model.

1034. How does the out-of-bag (OOB) error estimate work in Random Forest, and what are its limitations?

The out-of-bag (OOB) error estimate in Random Forest is an estimate of the generalization error of the model that is computed during training. The OOB error estimate works by using a subset of the training data that is not included in the bootstrap sample for each tree to evaluate the performance of that tree. The OOB error estimate has the advantage of not requiring a separate validation set, and it can be used to select the optimal set of hyperparameters or to compare the performance of different models. However, the OOB error estimate can have higher variance than a cross-validation estimate, and it may not be as reliable for small datasets.

1035. What is the difference between a feature selection technique and a feature importance measure in the context of Random Forest?

Feature selection techniques aim to identify a subset of the original feature set that is most relevant to the prediction task, and they are often used to reduce the dimensionality of the feature space. In contrast, feature importance measures aim to assign a score to each feature based on its contribution to the accuracy of the model. Feature importance measures can be used to gain insight into the underlying factors that drive the prediction task, and they can be used to identify the most

important features for further analysis.

1036. How can the performance of a Random Forest model be improved by reducing the correlation among the trees?

The performance of a Random Forest model can be improved by reducing the correlation among the trees. This can be achieved through techniques like feature bagging, where a random subset of features is used for each tree, or tree bagging, where a random subset of the training data is used to train each tree. Additionally, the correlation among the trees can be reduced by using a subset of the available trees for prediction or by averaging the predictions of multiple models trained on different subsets of the data.

1037. How can imbalanced class distributions affect the performance of a Random Forest model, and what techniques can be used to address this issue?

Imbalanced class distributions can affect the performance of a Random Forest model, as the model may be biased towards the majority class. Techniques like oversampling, undersampling, or class weighting can be used to address this issue. Oversampling involves duplicating instances of the minority class to balance the class distribution, while undersampling involves removing instances of the majority class. Class weighting involves assigning a higher weight to instances of the minority class to increase their influence on the model.

1038. How can the computational efficiency of a Random Forest model be improved for large datasets or high-dimensional feature spaces?

The computational efficiency of a Random Forest model can be improved for large datasets or high-dimensional feature spaces by using techniques like parallelization or distributed computing. Parallelization involves splitting the data across multiple processors or cores to speed up the training process, while distributed computing involves splitting the data across multiple machines. Additionally, techniques like dimensionality reduction or feature selection can be used to reduce the dimensionality of the feature space and speed up the training process.

1039. How can the interpretability of a Random Forest model be improved, and what techniques can be used to extract insights from the model?

The interpretability of a Random Forest model can be improved by using techniques like feature importance measures, partial dependence plots, or permutation feature importance. Feature importance measures can be used to identify the most important features in the model, while partial dependence plots can be used to visualize the relationship between individual features and the predicted outcome. Permutation feature importance involves randomly permuting the values of a feature and observing the effect on the model performance, which can be used to estimate the importance of that feature.

1040. What are some of the limitations and drawbacks of Random Forest, and when might it not be the best choice of model for a given problem?

Some limitations and drawbacks of Random Forest include the tendency to overfit if the trees in the forest are too deep or if the model is too complex, the inability to extrapolate beyond the range of the training data, the difficulty of handling missing data, and the lack of interpretability of the model at the individual

Chapter 10 - K-means and KNN

K-Means Algorithm and implementation in Python

1041.What is K-means?

ANS. K-means clustering is an unsupervised machine learning algorithm. K-means is the most common type of clustering technique. The process of k-means clustering is very simple it tries to classify the given dataset into number of clusters(groups) given by the k. Suppose we take k = 3 in this case 3 clusters are formed. It tries to find the pattern, similarity to create the clusters.

Example of k-means:

Suppose I have a data set of a customers where the columns in the data are Annual income and spending scores of the individuals. We have to find that depending on the annual income and spending scores what kind of group does the individual falls for example if the individual annual income is low and spending score is high then he will be grouped as careless. Let's take another example if the individual annual score is high and the spending score is precise not that low not that high then he will be place in the careful and so on. So basically we are creating different subgroups based on the similarity of the data.

1042. What is clustering?

ANS. Clustering is an Exploratory data analysis technique. It is used to discover subgroups from a dataset based on similar pattern. The data points in each subgroups must be very similar to each other and the data points in the other subgroups must be different. This groups are formed based on the Euclidean distance formula. Clustering can be used in market segmentation where we try to find the customers with similar characteristics.

1043. How does k-means clustering works?

ANS. The k-means clustering works in the following ways:

1. First we have to select the number of clusters by specifying k.
2. Selecting random points or centroids.
3. Assigning each data points to the nearest centroids.
4. Calculating the sum of the squared distance between data points and all centroids and assigning the new clusters to each data points.
5. Now repeat the third step which is nothing but assigning the data points to the nearest clusters.
6. In case any reassignment happens repeat the step forth or else done.
7. Our model is ready.

1044. What does k represents in k-means clustering?

ANS. k in k-means clustering represents number of clusters.

1045. What is a cluster?

ANS. A cluster is nothing but the collection of data points aggregated together because of some similarities.

1046. What is a centroid?

ANS. A centroid is a point that represents the centre of the cluster.

1047. What does k-means term denotes?

ANS. k in k-means denotes the number of centroids or number of clusters we want to create from the data. Means in k-means denotes averaging of the data.

1048. How to find the optimal value of k in k-means?

ANS. To find the optimal(minimal) value of k in k means the most popular method is known as elbow method.

To find the optimal value we use elbow method on the K-means clustering algorithm using a for loop on a range of value from 1 to 11.

First we perform k-means clustering on all the values from the range 1 to 11. For each k value we calculate the Within-Cluster Sum of Squared Error.

After this we plot a line graph for the value 1 to 11 which is nothing but the number of clusters against wsse(Within-Cluster Sum of Squared Error).

The point at which the line graph suddenly falls(elbow point) that is the optimal value of k in k-means clustering or the optimal number of cluster.

1049. What is Elbow method?

The "elbow method" is a common technique used in machine learning and data science to determine the optimal number of clusters for a given dataset. It involves plotting the within-cluster sum of squared errors (WCSS) against the number of clusters and selecting the number of clusters where the WCSS starts to level off and forms an "elbow".

The idea behind the elbow method is that as the number of clusters increases, the WCSS will decrease. However, at some point, the marginal decrease in WCSS will start to level off and become much smaller. This is the point where the WCSS forms an "elbow" in the plot, and the number of clusters corresponding to the "elbow" is considered to be the optimal number of clusters.

The elbow method is a simple and intuitive way to determine the optimal number of clusters, but it is not always reliable and may not always produce the optimal solution. Other techniques, such as the silhouette score or the gap statistic, may also be used to determine the optimal number of clusters. Ultimately, the choice of technique depends on the specifics of the dataset and the goals of the clustering analysis.

1050. What is the Within-Cluster Sum of Squared Error?

The within-cluster sum of squared errors (WCSS) is a measure of the total variability of the data points within a cluster. It is commonly used as a metric for evaluating the quality of clustering solutions. The WCSS is calculated as the sum of the squared distances between each data point in a cluster and the cluster centroid. The cluster centroid is the mean of all the data points in the cluster.

The WCSS is often used to compare different clustering solutions and determine the optimal number of clusters. The idea is that as the number of clusters increases, the WCSS will decrease. However, there is often a trade-off between the number of clusters and the WCSS. Too few clusters can result in large WCSS values because the data points are not being grouped effectively, while too many clusters can result in small WCSS values but may not capture the underlying structure of the data.

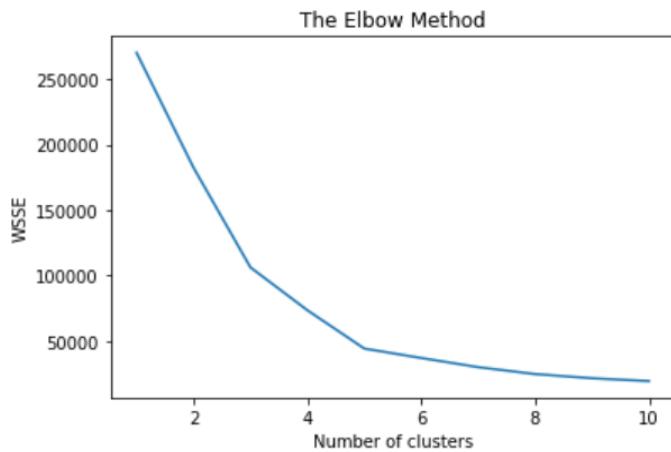
The goal of clustering is to find a solution that balances the WCSS and the number of clusters. A good clustering solution should have a small WCSS and a reasonably low number of clusters. The WCSS can be used to determine the optimal number of clusters through techniques such as the elbow method or the silhouette score.

1051. Implement WCSS Error in Python

CODE FOR ELBOW METHOD:

```
# Using the elbow method to find the optimal number of clusters
from sklearn.cluster import KMeans
wsse = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, random_state = 10)
    kmeans.fit(X)
    wsse.append(kmeans.inertia_)
plt.plot(range(1, 11), wsse)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WSSE')
plt.show()
```

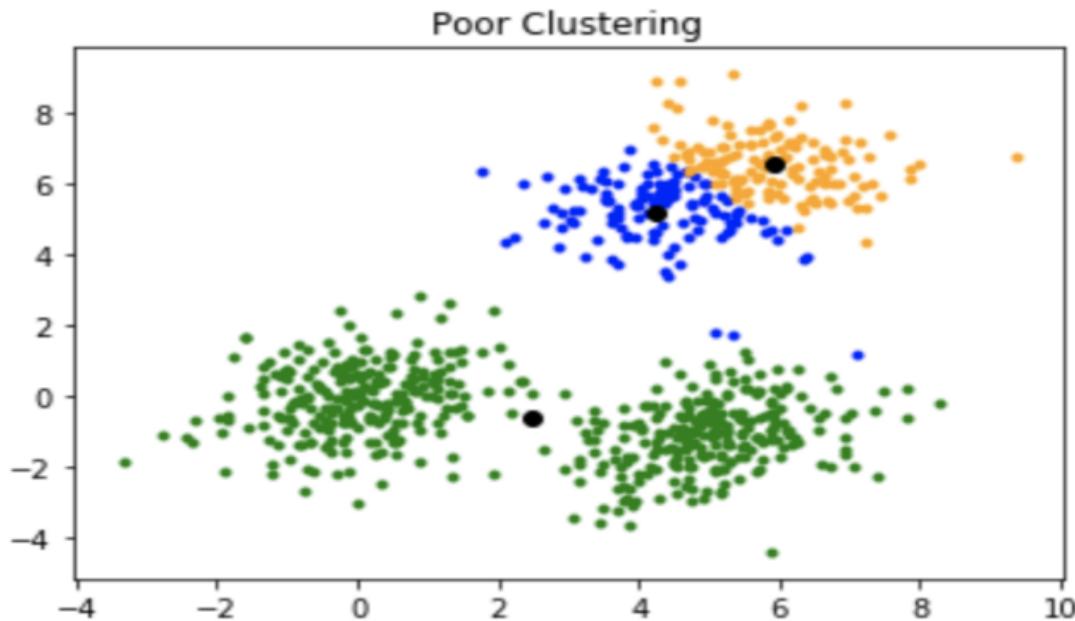
OUTPUT:



1052. What is the drawback of K-means clustering algorithm?

ANS. The drawback of K-means clustering algorithm is that it is sensitive for initialization of the centroids. Suppose if a centroid is initialized to a very far data point it might end up with no other data points associated with it or it might consider more than one cluster with a single centroid and vice versa more than one centroids might be initialized into the common cluster this results in poor clustering. Poor initialization of centroids might give us poor results in clustering.

Example:



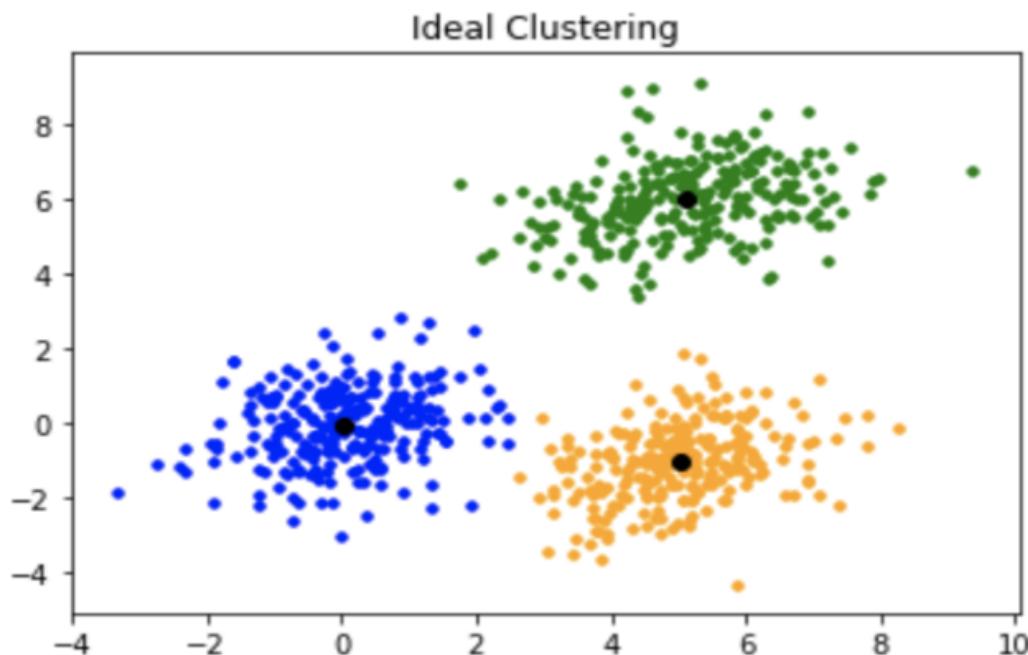
1053. What are the different ways to solve the problem of initialization sensitivity in k-means algorithm?

ANS. There are two ways to solve the problem of initialization sensitivity in k-means algorithm:

- Repeat k-means: In this case the algorithm repeats itself again and again initializing the centroids thus creating the clusters with small intracluster distance and large intercluster distance.
- K-means++: K-means++ algorithm uses a smart initialization process that deals with the problem of initialization sensitivity in k-means algorithm.

1054. What is K-means++ algorithm?

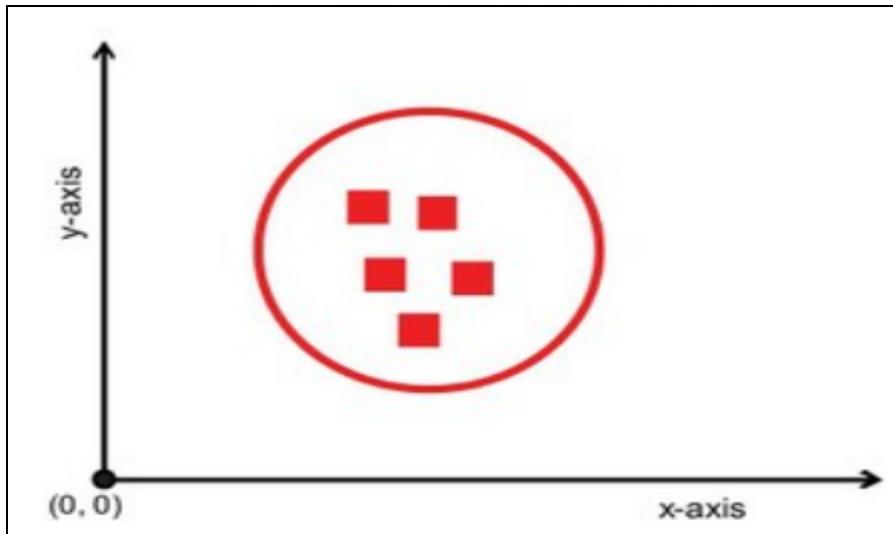
ANS. To deal with the drawback of k-means algorithm which is the initialization problem of centroids we use k-means++. K-means++ algorithm does a smart initialization of the centroids and tries to improve the quality of the clusters. Only the initialization problem is different from the standard k-means algorithm other than that everything else is the same. To solve the above problem we used k-means++ algorithm which results the following output



1055. What do you mean by intracluster distance?

ANS. Intracluster distance is nothing but the distance between the two data points or the members of the same cluster. This gives us the idea that how well the distance measures are able to bring the items together. The intracluster distance between the members of the cluster should be small as compared to the intercluster distance. The intracluster distance should be as small as possible so that it is able to bring similar data points together.

Example:



1056. What do you mean by intercluster distance?

ANS. Intercluster distance is nothing but the distance between the two data points or the members of the different clusters. The intercluster distance between the members of the cluster should be big as compared to the intracluster distance. The distance should be maximum so that it can distinguish that the two points belong to different clusters.

1057. Difference between K-means and K-means++ algorithm?

ANS. The K-means and the K-means++ algorithm are clustering techniques that comes under Unsupervised Learning. K-means++ algorithm is used to overcome the drawback of the k-means algorithm. The K-means++ algorithm gives a more intelligent initialization of centroids by which the cluster takes place and therefore it improves the nature of clusters. Besides the initialization, there is no other differences and they are almost the same.

1058. Difference between Classification and Clustering?

ANS. Difference between classification and clustering are as follow:

- Classification is used for Supervised Learning Algorithm and Clustering is used for Unsupervised Learning Algorithm.
- Classification have labelled data associated with it whereas Clustering is associated with unlabelled data.
- Classification is a process where the inputs are classified based on their corresponding labels and Clustering is a process where grouping is done on the basis of similarity.
- In Classification we have labels so in this case there is need of training dataset and testing dataset for evaluating the accuracy of the model created whereas in case of clustering there is no need of training dataset and testing dataset.

- Classification is much more complex as compared to the clustering because there are many levels in classification technique and in clustering only grouping is done.
- Examples of classification are Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, K-nearest neighbours, etc.
- Examples of clustering are k-means clustering algorithm, Hierarchical Clustering, etc.

1059. What are the advantages of k-means clustering algorithm?

ANS. The advantages of k-means clustering algorithm are:

- It is easy to understand and easy to implement.
- It works on unlabelled data.
- Working with large number of variables k-means can be computationally faster.
- Guaranteed convergence.
- Works better with spherical clusters.
- Easy to interpret and flexible.

1060. What are the Disadvantages of k-means clustering algorithm?

ANS. The disadvantages of k-means clustering algorithm are:

- We are suppose to choose the value of k manually which is nothing but the total number of clusters.
- K-mean does not work well when the clusters are of different size and different density. To get a better result the cluster must be spherical and equally sized.
- In K-means data should be numerical if not then some pre-processing of the data is necessary.
- K-means algorithm cannot handle outliers and noisy data.
- We cannot pass a very huge dataset if we do so the results may be poor or the computer may crash.
- If there are two data points which are overlapping then it is unable to differentiate that there are two clusters.

1061. Difference between KNN and K-means algorithm?

ANS. Sometime we get confused by the K in both the algorithm so the differences are as follow:

- KNN is a supervised learning algorithm whereas K-means is an unsupervised learning algorithm.
- KNN need labelled data to train, test and evaluate whereas K-means need unlabelled data there is no need of training and testing.
- K in KNN indicates number of nearest neighbours and k in K-means indicated total number of clusters or groups.
- KNN can be used for both classification and regression whereas K-means is used for clustering.

- There are many differences between KNN and K-means but there is one similarity that both the algorithm works on distance metrics.

1062. What are the different types of distance metrics used in K-means algorithm?

ANS. The different types of distance metrics in K-means algorithm are:

- Euclidean: The Euclidean distance determines the distance between two points. If we have a point A and point B the Euclidean distance is an ordinary straight line. It is the distance between the two points in Euclidean space.
- Manhattan: The Manhattan distance is nothing but the (Manhattan distance between two points (a_1,b_1) and (a_2,b_2) is $|a_1 - a_2| + |b_1 - b_2|$) simple sum of the distance between two points measured along axes at right angles.

1063. What are the different types of clustering?

ANS. The different types of clustering are:

- Hierarchical Clustering: Hierarchical clustering is a technique that uses a tree-like structure. Hierarchical clustering is an unsupervised clustering algorithm that creates clusters that have predominant ordering from top to bottom.
- Partitioning Clustering: Partitioning Clustering is a clustering technique that classifies the data into several parts as denoted by the k. Suppose if k=2 the two clusters will be created k_1, k_2 . The objects in the clusters will be different from each cluster but the objects within each cluster will be similar.

1064. What are the applications of K-means Clustering Algorithm?

ANS. The applications of K-means Clustering Algorithm are:

- Academic performance of the students: K-means helps in categorizing the students into different grades like A1, B1, C1 based on the marks obtained by the students.
- Search engines: K-means helps the search engines like when a search is performed the results are grouped therefore the search engines uses clustering techniques.
- Customer Segmentation: To better understand the customer in the market the owner uses customer segmentation to understand which customer they should target with the help of clustering technique and to understand the customer behaviour.

1065. Steps for performing K-means Clustering Algorithm in Python.

ANS. The steps for performing K-means clustering in Python are as follows:

Step 1:

Select k data points as the initial cluster centers.(Randomly)

Step 2:

Find the euclidean distance of each data point towards each cluster centres.

Step 3:

Assign each data point to the nearest cluster.

Step 4:

Recompute the new cluster centres by taking mean of the data points belonging to that cluster.

Step 5:

Repeat step 2 to 4.

Step 6:

Stop the process when zero convergence is reached.

End result:

You get the data points clustered into k clusters.

Implementation of K-means Clustering Algorithm with Python:

Description of the dataset: The data set which we are using for the k-means clustering is a customers dataset in which we have columns like CustomerId, Gender, Age, Annual Income and Spending score. For K-means we are only using the two column named Annual Income and Spending score. We have to find that depending on the annual income and spending scores what kind of group does the individual falls for example if the individual annual income is low and spending score is high then he will be grouped as careless. Let's take another example if the individual annual score is high and the spending score is precise not that low not that high then he will be place in the careful and so on. So basically we are creating different subgroups based on the similarity they have.

1066. What are the libraries to apply K-Mean cluster?

1067. How to import dataset as Data Frame?

1068. How to check the missing values in the dataset?

1069. Plot using scatter plot in Python

1070. How to save the result in the excel file?

- Importing the required Libraries:

```
import numpy as np  
import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt
```

- Importing the dataset as DataFrame:

```
df = pd.read_csv(r'Mall_Customers.csv')
```

- Checking the dataset using the head() function:

```
df.head()
```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

- Checking the size of the dataset using shape function:
`df.shape`

```
df.shape
```

```
(200, 5)
```

- Checking for missing values in the dataset using isnull() function:
df.isnull().sum()

```
df.isnull().sum()
```

```
CustomerID      0  
Gender          0  
Age             0  
Annual Income (k$)  0  
Spending Score (1-100)  0  
dtype: int64
```

- Converting the data frame in array because array are much faster then data frame in building models the columns which we are passing are Annual Income (k\$) and Spending Score(1-100) because these are the two columns which we will be using for k-means clustering:
`X = df.values[:, [3,4]]`
- Printing the array created above:
`print(X)`

```
print(X)
```

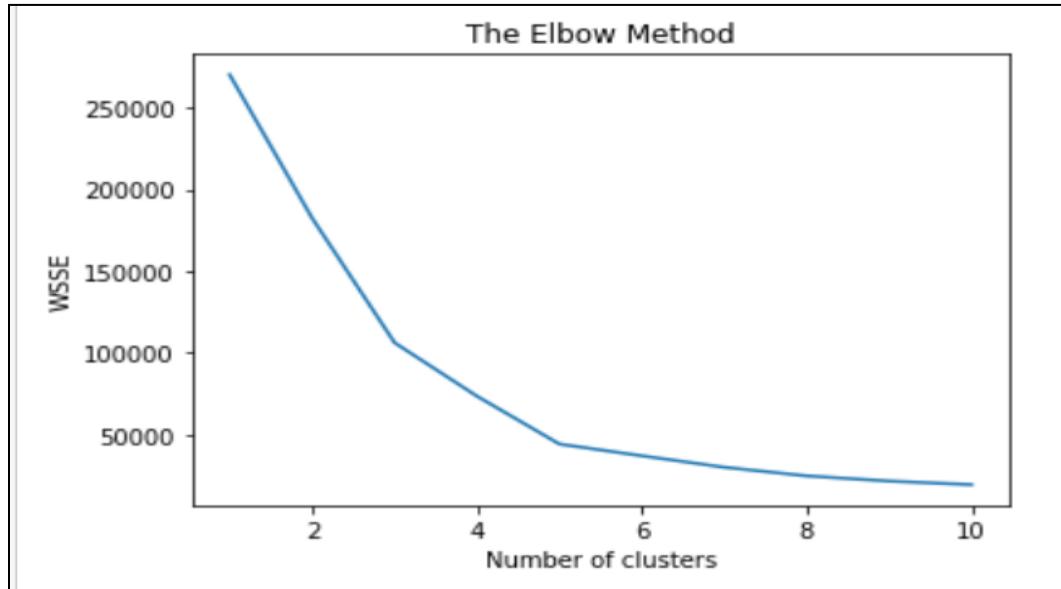
```
[[15 39]  
[15 81]  
[16 6]  
[16 77]  
[17 40]  
[17 76]  
[18 6]  
[18 94]]
```

- Using the elbow method to find the optimal number of clusters:
`from sklearn.cluster import KMeans`

```

wsse = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, random_state = 10)
    kmeans.fit(X)
    wsse.append(kmeans.inertia_)
plt.plot(range(1, 11), wsse)
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WSSE')
plt.show()

```



The point at which the line graph suddenly falls(elbow point) that is the optimal value of k in k-means clustering or the optimal number of cluster.

In this case the optimal number of cluster is k = 5.

- Printing within-cluster sum of squared error(wsse)
print(wsse)

```

print(wsse)
[269981.2800000014, 182440.30762987016, 106348.37306211119, 73679.78903948837, 44448.45544793369, 37265.86520484345, 30273.394
312070028, 25007.38394731206, 21826.936303231643, 19669.71099830122]

```

- Fitting k-means to the dataset:
kmeans = KMeans(n_clusters=5, random_state=10)
Y_pred = kmeans.fit_predict(X)
- Printing the Y_pred:
print(Y_pred)

```
print(Y_pred)
```

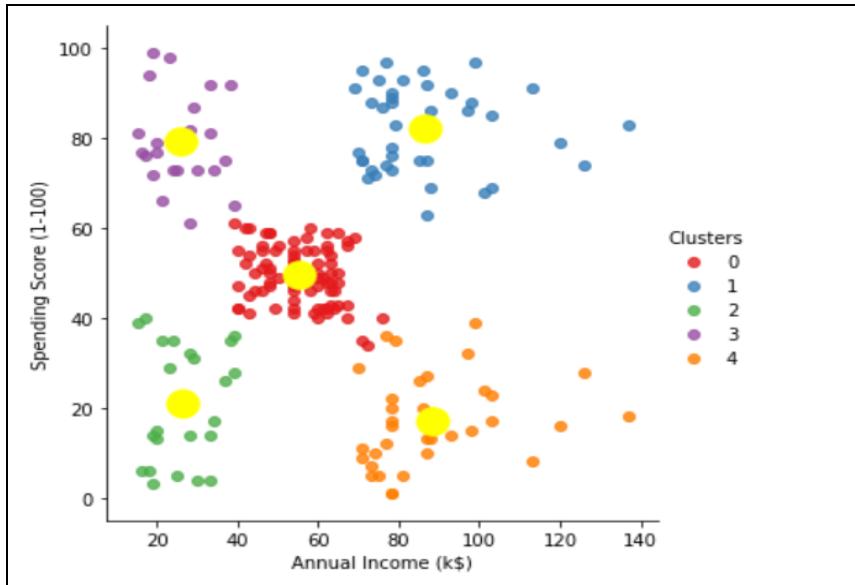
- Now we are creating a new column name Clusters which is the value of Y_pred and storing it in the original dataframe:

```
df['Clusters']=Y_pred  
df.head()
```

CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Clusters
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

- Plotting the clusters:

```
import seaborn as sns
sns.lmplot( data=df, x='Annual Income (k$)', y='Spending Score (1-100)',
fit_reg=False, # No regression line
hue='Clusters',palette="Set1")
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
s = 300, c = 'yellow')
plt.show()
```



We have created a scatter plot on the model kmeans where we have taken n_clusters = 5 so 5 cluster are plotted in the above scatter plot.

- Manually mapping or assigning the names of the clusters created:

```
df['Clusters']=df.Clusters.map({0:"Careless",1:"Standard",2:"Target",3:"Not-Sensible",4:"Careful"})
```

Now we have manually assigned the name if the clusters as the clusters have properties like the cluster 0 which is named careless because the income is low but the spending score is high.

```
df.head()
```

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Clusters
0	1	Male	19	15	39	Target
1	2	Male	21	15	81	Not-Sensible
2	3	Female	20	16	6	Target
3	4	Female	23	16	77	Not-Sensible
4	5	Female	31	17	40	Target

- If we want to check for a particular category of clusters we can check that in the following ways:

```
new_df=df[df["Clusters"]=="Careless"]
new_df
```

new_df						
CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Clusters	
43	44	Female	31	39	61	Careless
46	47	Female	50	40	55	Careless
47	48	Female	27	40	47	Careless
48	49	Female	29	40	42	Careless
49	50	Female	31	40	42	Careless
...
121	122	Female	38	67	40	Careless
122	123	Female	40	69	58	Careless
126	127	Male	43	71	35	Careless
132	133	Female	25	72	34	Careless
142	143	Female	28	76	40	Careless

81 rows × 6 columns

- We can save the result in the excel file in the following ways:
`new_df.to_excel("CarelessCustomers.xlsx",index=False)`

KNN

1071. What is the KNN Algorithm?

Ans:- KNN(K-nearest neighbours) is a supervised learning and non-parametric algorithm that can be used to solve both classification and regression problem statements. It uses data in which there is a target column present i.e, labelled data to model a function to produce an output for the unseen data. It uses the euclidean distance formula to compute the distance between the data points for classification or prediction. The main objective of this algorithm is that similar data points must be close to each other so it uses the distance to calculate the similar points that are close to each other.

1072. Why is KNN a non-parametric Algorithm?

Ans:- The term “non-parametric” refers to not making any assumptions on the underlying data distribution. These methods do not have any fixed numbers of parameters in the model.

Similarly in KNN, the model parameters grow with the training data by considering each training case as a parameter of the model. So, KNN is a non-parametric algorithm.

1073. What is “K” in the KNN Algorithm?

Ans:- K represents the number of nearest neighbours you want to select to predict the class of a given item, which is coming as an unseen dataset for the model.

1074. How does the KNN algorithm make the predictions on the unseen dataset?

Ans:- The following operations have happened during each iteration of the algorithm. For each of the unseen or test data point, the kNN classifier must:

- Step-1: Calculate the distances of test point to all points in the training set and store them
- Step-2: Sort the calculated distances in increasing order
- Step-3: Store the K nearest points from our training dataset
- Step-4: Calculate the proportions of each class
- Step-5: Assign the class with the highest proportion

1075. Is Feature Scaling required for the KNN Algorithm? Explain with proper justification.

Ans:- Yes, feature scaling is required to get the better performance of the KNN algorithm. For Example, Imagine a dataset having n number of instances and N number of features. There is one feature having values ranging between 0 and 1. Meanwhile, there is also a feature that varies from -999 to 999. When these values are substituted in the formula of Euclidean Distance, this will affect the performance by giving higher weightage to variables having a higher magnitude.

1076. What is space and time complexity of the KNN Algorithm? Ans:-

Time complexity:

The distance calculation step requires quadratic time complexity, and the sorting of the calculated distances requires an $O(N \log N)$ time. Together, we can say that the process is an $O(N^3 \log N)$ process, which is a monstrously long process.

Space complexity:

Since it stores all the pairwise distances and is sorted in memory on a machine, memory is also the problem. Usually, local machines will crash, if we have very large datasets.

1077. Can the KNN algorithm be used for regression problem statements? Ans:-Yes, KNN can be used for regression problem statements. In other words, the KNN algorithm can be applied when the dependent variable is continuous. For regression problem statements, the predicted value is given by the average of the values of its k nearest neighbours.**1078. Why is the KNN Algorithm known as Lazy Learner?**

Ans:- When the KNN algorithm gets the training data, it does not learn and make a model, it just stores the data. Instead of finding any discriminative function with the help of the training data, it follows instance-based learning and also uses the training data when it actually needs to do some prediction on the unseen datasets.

As a result, KNN does not immediately learn a model rather delays the learning thereby being referred to as Lazy Learner.

1079. Why is it recommended not to use the KNN Algorithm for large datasets?

Ans:-

The Problem in processing the data:

KNN works well with smaller datasets because it is a lazy learner. It needs to store all the data and then make a decision only at run time. It includes the computation of distances for a given point with all other points. So if the dataset is large, there will be a lot of processing which may adversely impact the performance of the algorithm.

Sensitive to noise:

Another thing in the context of large datasets is that there is more likely a chance of noise in the dataset which adversely affects the performance of the KNN algorithm since the KNN algorithm is sensitive to the noise present in the dataset.

1080. How to handle categorical variables in the KNN Algorithm?

Ans:- To handle the categorical variables we have to create dummy variables out of a categorical variable and include them instead of the original categorical variable. Unlike regression, create k dummies instead of (k-1).

For example, a categorical variable named “Degree” has 5 unique levels or categories. So we will create 5 dummy variables. Each dummy variable has 1 against its degree and else 0.

1081. How to choose the optimal value of K in the KNN Algorithm?

Ans:- There is no straightforward method to find the optimal value of K in the KNN algorithm. You have to play around with different values to choose which value of K should be optimal for my problem statement. Choosing the right value of K is done through a process known as Hyperparameter Tuning. The optimum value of K for KNN is highly dependent on the data itself. In different scenarios, the optimum K may vary. It is more or less a hit and trial method. There is no one proper method of finding the K value in the KNN algorithm. No method is the rule of thumb but you should try the following suggestions:

1. Square Root Method: Take the square root of the number of samples in the training dataset and assign it to the K value.
2. Cross-Validation Method: We should also take the help of cross-validation to find out the optimal value of K in KNN. Start with the minimum value of k i.e, K=1, and run cross-validation, measure the accuracy, and keep repeating till the results become consistent.

As the value of K increases, the error usually goes down after each one-step increase in K, then stabilizes, and then raises again. Finally, pick the optimum K at the beginning of the stable zone. This technique is also known as the Elbow Method.

3. Domain Knowledge: Sometimes with the help of domain knowledge for a particular use case we are able to find the optimum value of K (K should be an odd number). I would therefore suggest trying a mix of all the above points to reach any conclusion.

1082. How can you relate KNN Algorithm to the Bias-Variance tradeoff? Ans:- Problem with having too small K:

The major concern associated with small values of K lies behind the fact that the smaller value causes noise to have a higher influence on the result which will also lead to a large variance in the predictions.

Problem with having too large K:

The larger the value of K, the higher is the accuracy. If K is too large, then our model is under-fitted. As a result, the error will go up again. So, to prevent your model from under-fitting it should retain the generalization capabilities otherwise there are fair chances that your model may perform well in the training data but drastically fail in the real data. The computational expense of the algorithm also increases if we choose the k very large.

So, choosing k to a large value may lead to a model with a large bias(error).

The effects of k values on the bias and variance is explained below : As the value of k increases, the bias will be increases

As the value of k decreases, the variance will increases

With the increasing value of K, the boundary becomes smoother

So, there is a tradeoff between overfitting and underfitting and you have to maintain a balance while choosing the value of K in KNN. Therefore, K should not be too small or too large.

1083. Which algorithm can be used for value imputation in both categorical and continuous categories of data?

Ans:- KNN is the only algorithm that can be used for the imputation of both categorical and continuous variables. It can be used as one of many techniques when it comes to handling missing values.

To impute a new sample, we determine the samples in the training set “nearest” to the new sample and averages the nearby points to impute. A Scikit learn library of Python provides a quick and convenient way to use this technique.

1084. Explain the statement- “The KNN algorithm does more computation on test time rather than train time”.

Ans:- The above-given statement is absolutely true.

The basic idea behind the kNN algorithm is to determine a k-long list of samples that are close to a sample that we want to classify. Therefore, the training phase is basically storing a training set, whereas during the prediction stage the algorithm looks for k-neighbours using that stored data. Moreover, KNN does not learn anything from the training dataset as well.

1085. What are the things which should be kept in our mind while choosing the value of k in the KNN Algorithm?

Ans:- If K is small, then results might not be reliable because the noise will have a higher influence on the result. If K is large, then there will be a lot of processing to be done which may adversely impact the performance of the algorithm

So, the following things must be considered while choosing the value of K:

- K should be the square root of n (number of data points in the training dataset).
- K should be chosen as the odd so that there are no ties. If the square root is even, then add or subtract 1 to it.

1086. What are the advantages of the KNN Algorithm?

Some of the advantages of the KNN algorithm are as follows:

1. No Training Period: It does not learn anything during the training period since it does not find any discriminative function with the help of the training data. In simple words, actually, there is no training period for the KNN algorithm. It stores the training dataset and learns from it only when we use the algorithm for making the real-time predictions on the test dataset. As a result, the KNN algorithm is much faster than other algorithms which require training. For Example, SupportVector Machines(SVMs), Linear Regression, etc.

Moreover, since the KNN algorithm does not require any training before making predictions as a result new data can be added seamlessly without impacting the accuracy of the algorithm.

2. Easy to implement and understand: To implement the KNN algorithm, we need only two parameters i.e. the value of K and the distance metric(e.g. Euclidean or Manhattan, etc.). Since both the parameters are easily interpretable therefore they are easy to understand.

1087. What are the disadvantages of the KNN Algorithm?

Ans:- Some of the disadvantages of the KNN algorithm are as follows:

1. Does not work well with large datasets: In large datasets, the cost of calculating the distance between the new point and each existing point is huge which decreases the performance of the algorithm.

2. Does not work well with high dimensions: KNN algorithms generally do not work well with high dimensional data since, with the increasing number of dimensions, it becomes difficult to calculate the distance for each dimension.

3. Need feature scaling: We need to do feature scaling (standardization and normalization) on the dataset before feeding it to the KNN algorithm otherwise it may generate wrong predictions.

4. Sensitive to Noise and Outliers: KNN is highly sensitive to the noise present in the dataset and requires manual imputation of the missing values along with outliers removal.

1088. Is it possible to use the KNN algorithm for Image processing?

Ans:- Yes, KNN can be used for image processing by converting a 3- dimensional image into a single-dimensional vector and then using it as the input to the KNN algorithm.

1089. What are the real-life applications of KNN Algorithms?

Ans:- The various real-life applications of the KNN Algorithm includes:

1. KNN allows the calculation of the credit rating. By collecting the financial characteristics vs. comparing people having similar financial features to a database we can calculate the same. Moreover, the very nature of a credit rating where people who have similar financial details would be given similar credit ratings also plays an important role. Hence the existing database can then be used to predict a new customer's credit rating, without having to perform all the calculations.

2. In political science: KNN can also be used to predict whether a potential voter "will vote" or "will not vote", or to "vote Democrat" or "vote Republican" in an election. Apart from the above- mentioned use cases, KNN algorithms are also used for handwriting detection (like OCR), Image recognition, and video recognition.
3. Forecasting stock market: Predict the price of a stock, on the basis of company performance measures and economic data.
4. Currencyexchangerate
5. Bank bankruptcies
6. Understanding and managing financial risk
7. Trading futures
8. Creditrating
9. Loan management
10. Bank customer profiling
11. Money laundering analyses

1090. Explain the Difference between K-Means and KNN

Ans:-

K- means

It is an Unsupervised learning technique

It is used for Clustering

KNN

It is a Supervised learning technique

It is used mostly for Classification, and sometimes even for Regression

<p>'K' in K-Means is the number of clusters the algorithm is trying to identify/learn from the data. The clusters are often unknown since this is used with Unsupervised learning.</p>	<p>'K' in KNN is the number of nearest neighbours used to classify or (predict in case of continuous variable/regression) a test sample</p>
<p>It is typically used for scenarios like understanding the population demomgraphics, market segmentation, social media trends, anomaly detection, etc. where the clusters are unknown to begin with.</p>	<p>It is used for classification and regression of known data where usually the target attribute/variable is known before hand.</p>

In training phase of K-Means, K observations are arbitrarily selected (known as centroids). Each point in the vector space is assigned to a cluster represented by nearest (euclidean distance) centroid. Once the clusters are formed, for each cluster the centroid is updated to the mean of all cluster members. And the cluster formation restarts with new centroids. This repeats until the centroids themselves become mean of clusters, i.e., when updating centroids to mean doesn't change them. The prediction of a test observation is done based on nearest centroid.

K-NN doesn't have a training phase as such. But the prediction of a test observation is done based on the K- Nearest (often euclidean distance) Neighbours (observations) based on weighted averages/votes.

1091. Show an implementation of KNN Algorithm

Ans:-

Code(1):-

```
from mlxtend.plotting import plot_decision_regions import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt import seaborn as sns  
sns.set()  
import warnings warnings.filterwarnings('ignore') %matplotlib inline
```

Basic Data Science and ML Pipeline

1. Obtaining our data
2. Scrubbing / Cleaning our data
3. Exploring / Visualizing our data will allow us to find patterns and trends
4. Modelling our data will give us our predictive power as a wizard
5. Interpreting our data

Code(2):-

#Loading the dataset

```
diabetes_data = pd.read_csv(r'C:\Users\Saurabh\Downloads\dia\diabetes.csv')  
diabetes_data.head() #Print the first 5 rows of the dataframe.
```

OP:-

Code(3):- diabetes.describe()

The Question creeping out of this summary

Can minimum value of below listed columns be zero (0)?

On these columns, a value of zero does not make sense and thus indicates missing value.

Following columns or variables have an invalid zero value:

1. Glucose
2. BloodPressure
3. SkinThickness
4. Insulin
5. BMI

It is better to replace zeros with nan since after that counting them would be easier and zeros need to be replaced with suitable values

Code(4):-

```
diabetes_data_copy = diabetes_data.copy(deep = True)  
diabetes_data_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BM I']] =  
diabetes_data_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BM I']].replace(0,np.NaN)  
  
## showing the count of Nans print(diabetes_data_copy.isnull().sum()) OP:-
```

To fill these Nan values the data distribution needs to be understood

Code(5):-

```
p = diabetes_data.hist(figsize = (20,20)) OP:-
```

Aiming to impute nan values for the columns in accordance with their distribution.

Code(6):-

```
diabetes_data_copy['Glucose'].fillna(diabetes_data_copy['Glucose'].mean(), inplace = True)
```

```
diabetes_data_copy['BloodPressure'].fillna(diabetes_data_copy['BloodPressure'].mean(), inplace = True)
```

```
diabetes_data_copy['SkinThickness'].fillna(diabetes_data_copy['SkinThickness'].median(), inplace = True)
```

```
diabetes_data_copy['Insulin'].fillna(diabetes_data_copy['Insulin'].median(), inplace = True)
```

```
diabetes_data_copy['BMI'].fillna(diabetes_data_copy['BMI'].median(), inplace = True)
```

Plotting after Nan removal

Code(7):-

```
p = diabetes_data.hist(figsize = (20,20))
```

Skewness

A left-skewed distribution has a long left tail. Left-skewed distributions are also called negatively-skewed distributions. That's because there is a long tail in the negative direction on the number line. The mean is also to the left of the peak.

A right-skewed distribution has a long right tail. Right-skewed distributions are also called positive-skew distributions. That's because there is a long tail in the positive direction on the number line. The mean is also to the right of the peak.

Code(8):-

```
## null count analysis import missingno as msno p=msno.bar(diabetes_data)
```

OP:-

Code(9):-

```
## checking the balance of the data by plotting the count of outcomes by their value
```

```
color_wheel = {1: "#0392cf", 2: "#7bc043"}
```

```
colors = diabetes_data["Outcome"].map(lambda x: color_wheel.get(x + 1))
print(diabetes_data.Outcome.value_counts())
p=diabetes_data.Outcome.value_counts().plot(kind="bar")
```

OP:-

The above graph shows that the data is biased towards datapoints having outcome value as 0

where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

Scatter matrix of uncleaned data.

Code(9):-

```
from pandas.tools.plotting import scatter_matrix  
p=scatter_matrix(diabetes_data,figsize=(25, 25))
```

OP:-

The pairs plot builds on two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship (or lack thereof) between two variables.

Heatmap for clean data

Code(10):- plt.figure(figsize=(12,10)) # on this line I just set the size of figure to 12 by 10.

```
p=sns.heatmap(diabetes_data_copy.corr(), annot=True,cmap ='RdYIGn') OP:-
```

Scaling the data

Code(11):-

```
from sklearn.preprocessing import StandardScaler
```

```
sc_X = StandardScaler()
```

```
X=pd.DataFrame(sc_X.fit_transform(diabetes_data_copy.drop(["Outcome"], axis = 1)),  
columns=['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI',  
'DiabetesPedigreeFunction', 'Age'])
```

OP:-

Why Scaling the data for KNN?

it is always advisable to bring all the features to the same scale for applying distance based algorithms like KNN.

Let's see an example of distance calculation using two features whose magnitudes/ranges vary greatly.

Euclidean Distance = $[(100000 - 80000)^2 + (30 - 25)^2]^{(1/2)}$

We can imagine how the feature with greater range will overshadow or diminish the smaller feature completely and this will impact the performance of all distance based model as it will give higher weightage to variables which have higher magnitude.

Test Train Split and Cross Validation methods

Train Test Split : To have unknown datapoints to test the data rather than testing with the same points with which the model was trained. This helps capture the model performance much better.

Code(11):-

```
#importing train_test_split  
  
from sklearn.model_selection import train_test_split  
  
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=1/3,random_state=42, stratify=y)
```

Code(12):-

```
from sklearn.neighbors import KNeighborsClassifier test_scores = []  
train_scores = []  
for i in range(1,15):  
  
    knn = KNeighborsClassifier(i) knn.fit(X_train,y_train)  
    train_scores.append(knn.score(X_train,y_train)) test_scores.append(knn.score(X_test,y_test))
```

Code(13):-

```
max_train_score = max(train_scores)
```

```
train_scores_ind = [i for i, v in enumerate(train_scores) if v == max_train_score]
```

```
print('Max train score {} % and k = {}'.format(max_train_score*100,list(map(lambda x: x+1, train_scores_ind))))
```

Code(14):-

```
max_test_score = max(test_scores)
```

```
test_scores_ind = [i for i, v in enumerate(test_scores) if v == max_test_score]
```

```
print('Max test score {} % and k = {}'.format(max_test_score*100,list(map(lambda x: x+1, test_scores_ind))))
```

Result Visualisation

Code(15):-

```
plt.figure(figsize=(12,5))
```

```
p = sns.lineplot(range(1,15),train_scores,marker='*',label='Train Score') p =  
sns.lineplot(range(1,15),test_scores,marker='o',label='Test Score')
```

OP:-

The best result is captured at k = 11 hence 11 is used for the final model.

Code(16):-

```
#Setup a knn classifier with k neighbors knn = KNeighborsClassifier(11)
```

```
knn.fit(X_train,y_train) knn.score(X_test,y_test)
```

OP:- 0.765625

Model Performance Analysis

1. Confusion Matrix

Code(17):-

```
#import confusion_matrix  
  
from sklearn.metrics import confusion_matrix  
  
#let us get the predictions using the classifier we had fit above y_pred = knn.predict(X_test)
```

```
confusion_matrix(y_test,y_pred)
```

```
pd.crosstab(y_test, y_pred, rownames=['True'], colnames=['Predicted'], margins=True)
```

Code(18):-

```
y_pred = knn.predict(X_test)
```

```
from sklearn import metrics
```

```
cnf_matrix = metrics.confusion_matrix(y_test,y_pred)
```

```
p = sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YIGnBu" ,fmt='g')
```

```
plt.title('Confusion matrix', y=1.1) plt.ylabel('Actual label') plt.xlabel('Predicted label')
```

OP:-

2.ROC-AUC Curve

ROC (Receiver Operating Characteristic) Curve tells us about how good the model can distinguish between two things (e.g If a patient has a disease or no). Better models can accurately distinguish between the two. Whereas, a poor model will have difficulties in distinguishing between the two

Code(19):-

```
from sklearn.metrics import roc_curve  
  
y_pred_proba = knn.predict_proba(X_test)[:,1]  
  
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba) plt.plot([0,1],[0,1],'k--')  
  
plt.plot(fpr,tpr, label='Knn')  
  
plt.xlabel('fpr')  
  
plt.ylabel('tpr')  
  
plt.title('Knn(n_neighbors=11) ROC curve') plt.show()
```

OP:-

3.Hyper Parameter optimization

Grid search is an approach to hyperparameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

Let's consider the following example:

Suppose, a machine learning model X takes hyperparameters a1, a2 and a3. In grid searching, you first define the range of values for each of the hyperparameters a1, a2 and a3. You can think of this as an array of values for each of the hyperparameters. Now the grid search technique will construct many versions of X with all the possible combinations of hyperparameter (a1, a2 and a3) values that you defined in the first place. This range of hyperparameter values is referred to as the grid.

Suppose, you defined the grid as: a1 = [0,1,2,3,4,5] a2 = [10,20,30,40,5,60] a3 = [105,105,110,115,120,125]

Note that, the array of values of that you are defining for the hyperparameters has to be legitimate in a sense that you cannot supply Floating type values to the array if the hyperparameter only takes Integer values.

Now, grid search will begin its process of constructing several versions of X with the grid that you just defined. It will start with the combination of [0,10,105], and it will end with [5,60,125]. It will go through all the intermediate combinations between these two which makes grid search computationally very expensive.

Chapter 11 - Power BI

Chapter - 11

Power BI

Q1. What are the parts of Microsoft Self-Service BI solution? Q2 What is Self-Service Solution?

Q3. What is Power BI?

Q4. What is Power BI Desktop?

Q5. What data sources can Power BI connect to?

Q6. What are Building Blocks in Power BI?

Q7. What are the different types of filters in Power BI Reports?

Q8. What are content packs in Power BI?

Q9. What is DAX?

Q9. What is DAX?

Q11. How is the FILTER function used?

Q12. What is DAX?

Q9. What is DAX?

Q14. What are some benefits of using Variables in DAX ?

Q15. How would you create trailing X month metrics via DAX against a non-standard calendar?

Q16. What is DAX?

Q17. What is Power Pivot?

Q18. What is Power Pivot Data Model?

Q19. What is xVelocity in-memory analytics engine used in Power Pivot?

Q20. What are some of differences in data modeling between Power BI Desktop and Power Pivot for Excel? Q21. Can we have more than one active relationship between two tables in data model of power pivot?

Q22. What is Power Query?

Q23. What are the data destinations for Power Queries?

Q24. What is query folding in Power Query?

Q25. What are some common Power Query/Editor Transforms?

Q26. Can SQL and Power Query/Query Editor be used together?

Q28. What are query parameters and Power BI templates?

Q29. Which language is used in Power Query?

Q30. Why do we need Power Query when Power Pivot can import data from mostly used sources?

Q31. What is Power Map?

Q32. What are the primary requirement for a table to be used in Power Map?

Q33. What are the data sources for Power Map?

Q34. What is Power View?

Q35. What is Power BI Designer?

Q36. Can we refresh our Power BI reports once uploaded to cloud (Share point or Powebi.com)?

- Q37. What are the different types of refreshing data for our published reports?
- Q38. Is Power BI available on-premises?
- Q39. What is data management gateway and Power BI personal gateway?
- Q40. What is Power BI Q&A?
- 41). What are some ways that Excel experience can be leveraged with Power BI?
- Q42. What is a calculated column in Power BI and why would you use them?
- Q43. How is data security implemented in Power BI ?
- Q44. What are many-to-many relationships and how can they be addressed in Power BI ?
- Q45. Why might you have a table in the model without any relationships to other tables?
- 46). What is the Power BI Publisher for Excel?
- Q47. What are the differences between a Power BI Dataset, a Report, and a Dashboard?
- Q48. What are the three Edit Interactions options of a visual tile in Power BI Desktop?
- Q49. What are some of the differences in report authoring capabilities between using a live or direct query connection such as to an Analysis Services model, relative to working with a data model local to the Power BI Desktop file?
- Q50. How does SSRS integrate with Power BI?
- Q51. What is the general business need for Power BI?
- Q52. Which data sources are often associated with Power BI?
- Q53. Can you please tell me what the distinctive channels in the Power BI reports.
- Q54. Can you please tell me what some of the key competitive differences are between Power BI and Power BI Pro?
- Q55. Do you know how much Power BI costs?
- Q56. Can you please tell me about the Power BI desktop application?
- Q57. Can you please tell me what DAX is?
- Q58. What are some of the core parts of Microsoft self service business intelligence solutions?
- Q59. Can you please tell me what the self-service business intelligence offering is?
- Q60. What are the building blocks when it comes to Power BI?
- Q61. What are the content packs when it comes to Power BI?
- Q62. Can you please tell me how you use the FILTER function?
- Q63. What are the various names of Power BI?
- Q64. Name some of the filters that you can utilize with the Power BI Reports.
- Q65. Can you tell me how well you understand the Power BI designer?
- Q66. What are the main differences between the Power BI gateway and Data management gateways?
- Q67. Can you please tell me what the selection pane is inside of Power BI?
- Q68. What is your opinion of the dynamic filtering option within Power BI?
- Q69. What data sets are generally used for creating a visual dashboard within the streaming data tiles?
- Q70. What would you consider to be the normal amount of table capacity for gathering/importing data in Power BI?
- Q71. Can you tell me what some of the advantages are of utilizing the variables within DAX?

Q72. Can you tell me what Power Pivot is?

Q73. Tell me what some of the key contrasts in the display of data between Power BI Desktop and Power Pivot for Excel?

Q74. Can you tell me what the Power Pivot Data Model is?

Q75. When is the x-Velocity examination tool utilized for Power Pivot?

Q76. Can you tell me if you would have access of one dynamic connection between two tables in the data model in the intensity rotate process?

Q77. Can you please describe to me what Power Query is ?

Q78. What are data goals when it comes to Power Queries?

Q79. Can you tell me what question collapsing within Power Query is?

Q80. Can you tell me if Power BI available for on-prem (or on premises)?

Q81. Does Power BI support general mobile devices like Android and iOS?

Q82. The Power BI Desktop's Software License Terms show that one may install and use one copy of the software on the premises. True or false?

Q83. Can you tell me what the query folding tool is in Power Query?

Q84. What are some of the most commonly used Power Query Data Transforms?

Q85. Can the SQL and Power Query Editor be used together?

Q86. What are the main query parameters and Power BI Templates you have access to?

Q87. What is the main language, which is used in Power Query?

Q88. Can you tell me why you would need Power Query when Power Pivot can import the data for you?

Q89. Can you tell me what the Power Map is within Power BI?

Q90. What are some of the main requirements for a table to be used within Power Map for Power BI?

Q91. What are the data hotspots when it comes to Power Map?

Q92. Is it possible to invigorate the Power BI reports once they are uploaded to the cloud?

Q93. What are some of the unique data invigorations used for generating the distributed reports within Power BI? Q94. Explain the Power BI Designer.

Q95. Is there any process for refreshing Power BI reports once uploaded to the cloud?

Q96. What is the major difference between Power BI personal Gateway and Data Management Gateway?

Q97. What is the use of split function?

Q98. Name all the platforms for which the Power BI app is available.

Q99. Differentiate between older and newer Power BI.

Q100. Is it possible in the power pivot data model to have more than one active relationship between two tables?

Q101. What is the purpose of the 'Get Data' icon in Power BI?

Q102. What is Row-level Security?

Q103. What are the general data shaping techniques?

Q104. Which data sets can be used to create dashboards with streaming data tiles? Q106.

What could be the difference between Distinct() and Values() in DAX? Q107. State the advantages of the Direct query method.

Q108. What is What if the parameter in power bi?

Q109. What is the incremental refresh?

Q110. What are the three main tabs in Reports development Window?

1100. What are the parts of Microsoft self-service business intelligence solution?

Microsoft has two parts for Self-Service BI

- a. Excel BI Toolkit
- b. Power BI

It Allows users to create an interactive report by importing data from different sources and model data according to report requirement.

It is The online solution that enables you to share the interactive reports and queries that you have created using the Excel BI Toolkit.

1101. What is self-service business intelligence?

Self-Service Business Intelligence (SSBI)

- SSBI is an approach to data analytics that enables business users to filter, segment, and analyze their data, without the in-depth technical knowledge in statistical analysis, business intelligence (BI).
- SSBI has made it easier for end users to access their data and create various visuals to get better business insights.
- Anybody who has a basic understanding of the data can create reports to build intuitive and shareable dashboards.

1102. What is Power BI?

Power BI is a cloud-based data sharing environment. Once you have developed reports using Power Query, Power Pivot and Power View, you can share your insights with your colleagues. This is where Power BI enters the equation. Power BI, which technically is an aspect of SharePoint online, lets you load Excel workbooks into the cloud and share them with a chosen group of co-workers. Not only that, but your colleagues can interact with your reports to apply filters and slicers to highlight data. They are completed by Power BI, a simple way of sharing your analysis and insights from the Microsoft cloud.

Power BI features allow you to:

- Share presentations and queries with your colleagues.
- Update your Excel file from data sources that can be on-site or in the cloud.
- Display the output on multiple devices. This includes PCs, tablets, and HTML 5-enabled mobile devices that use the Power BI app.
- Query your data using natural language processing (or Q&A, as it is known).

1103. What is Power BI Desktop?

Power BI Desktop is a free desktop application that can be installed right on your own computer. Power BI Desktop works cohesively with the Power BI service by providing advanced data exploration, shaping, modeling, and creating report with highly interactive visualizations. You can save your work to a file or publish your data and reports right to your Power BI site to share with others.

1104. What data sources can Power BI connect to?

The list of data sources for Power BI is extensive, but it can be grouped into the following:

- Files: Data can be imported from Excel (.xlsx, xlsm), Power BI Desktop files (.pbix) and Comma Separated Value (.csv).
- Content Packs: It is a collection of related documents or files that are stored as a group. In Power BI, there are two types of content packs, firstly those from services providers like Google Analytics, Marketo or Salesforce and secondly those created and shared by other users in your organization.
- Connectors to databases and other datasets such as Azure SQL, Database and SQL, Server Analysis Services tabular data, etc.

1105. What are Building Blocks in Power BI?

The following are the Building Blocks (or) key components of Power BI:

1. Visualizations: Visualization is a visual representation of data.

Example: Pie Chart, Line Graph, Side by Side Bar Charts, Graphical Presentation of the source data on top of Geographical Map, Tree Map, etc.

2. Datasets: Dataset is a collection of data that Power BI uses to create its visualizations.

Example: Excel sheets, Oracle or SQL server tables.

3. Reports: Report is a collection of visualizations that appear together on one or more pages.

Example: Sales by Country, State, City Report, Logistic Performance report, Profit by Products report etc.

4. Dashboards: Dashboard is single layer presentation of multiple visualizations, i.e we can integrate one or more visualizations into one page layer.

Example: Sales dashboard can have pie charts, geographical maps and bar charts.

5. Tiles: Tile is a single visualization in a report or on a dashboard.

Example: Pie Chart in Dashboard or Report.

1106. What are the different types of filters in Power BI Reports?

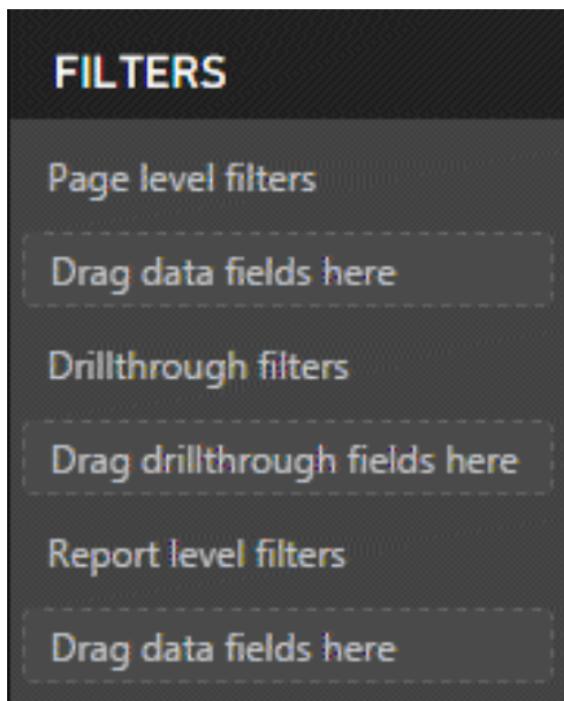
Power BI provides variety of option to filter report, data and visualization. The following are the list of Filter types.

- Visual-level Filters: These filters work on only an individual visualization, reducing the amount of data that the visualization can see. Moreover, visual-level filters can filter both data and calculations.

- Page-level Filters: These filters work at the report-page level. Different pages in the same report can have different page-level filters.
- Report-level Filters: These filters work on the entire report, filtering all pages and visualizations included in the report.

We know that Power BI visual have interactions feature, which makes filtering a report a breeze. Visual interactions are useful, but they come with some limitations:

- The filter is not saved as part of the report. Whenever you open a report, you can begin to play with visual filters but there is no way to store the filter in the saved report.
- The filter is always visible. Sometimes you want a filter for the entire report, but you do not want any visual indication of the filter being applied.



1107. What are content packs in Power BI?

Content packs for services are pre-built solutions for popular services as part of the Power BI experience. A subscriber to a supported service, can quickly connect to their account from Power BI to see their data through live dashboards and interactive reports that have been pre-built for them. Microsoft has released content packs for popular services such as Salesforce.com, Marketo, Adobe Analytics, Azure Mobile Engagement, CircuitID, comScore Digital Analytix, Quickbooks Online, SQL Sentry and tyGraph. Organizational content packs provide users, BI professionals, and system integrator the tools to build their own content packs to share purpose-built dashboards, reports, and datasets within their organization.

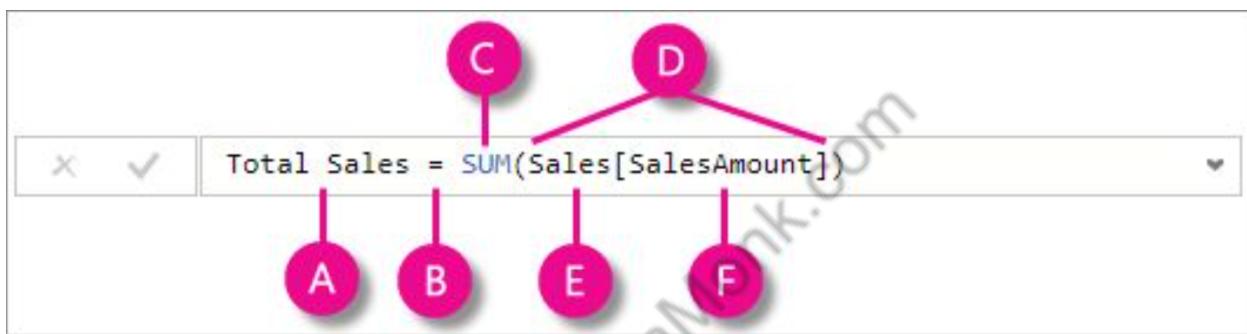
1108. What is DAX?

To do basic calculation and data analysis on data in power pivot, we use Data Analysis Expression (DAX). It is formula language used to compute calculated column and calculated field.

- DAX works on column values.
- DAX can not modify or insert data.
- We can create calculated column and measures with DAX but we can not calculate rows using DAX.

Sample DAX formula syntax:

For the measure named Total Sales, calculate (=) the SUM of values in the [SalesAmount] column in the Sales table.



A- Measure Name

B- = – indicate beginning of formula C- DAX Function

D- Parenthesis for Sum Function

E- Referenced Table

F- Referenced column name

1109. What are some of the DAX functions?

Below are some of the most commonly used DAX function:

- SUM, MIN, MAX, AVG, COUNTROWS, DISTINCTCOUNT
- IF, AND, OR, SWITCH
- ISBLANK, ISFILTERED, ISCROSSFILTERED
- VALUES, ALL, FILTER, CALCULATE,
- UNION, INTERSECT, EXCEPT, NATURALINNERJOIN, NATURALLEFTEROUTERJOIN, SUMMARIZECOLUMNS, ISEMPTY,
- VAR (Variables)
- GEOMEAN, MEDIAN, DATEDIFF

1110. How is the FILTER function used?

The FILTER function returns a table with a filter condition applied for each of its source table rows. The FILTER function is rarely used in isolation, it's generally used as a parameter to other functions such as CALCULATE.

- FILTER is an iterator and thus can negatively impact performance over large source tables.
- Complex filtering logic can be applied such as referencing a measure in a filter expression.
- FILTER(MyTable,[SalesMetric] > 500)

1111. What are the functions and limitations of DAX?

These are the only functions that allow you modify filter context of measures or tables.

- Add to existing filter context of queries.
- Override filter context from queries.
- Remove existing filter context from queries.

Limitations:

- Filter parameters can only operate on a single column at a time.
- Filter parameters cannot reference a metric.

1112. What is SUMMARIZE() and SUMMARIZECOLUMNS() DAX?

SUMMARIZE()

- Main group by function in SSAS.
- Recommended practice is to specify table and group by columns but not metrics. You can use ADDCOLUMNS function.

SUMMARIZECOLUMNS

- New group by function for SSAS and Power BI Desktop; more efficient.
- Specify group by columns, table, and expressions.

1113. What are some benefits of using Variables in DAX ?

Below are some of the benefits:

- By declaring and evaluating a variable, the variable can be reused multiple times in a DAX expression, thus avoiding additional queries of the source database.
- Variables can make DAX expressions more intuitive/logical to interpret.
- Variables are only scoped to their measure or query, they cannot be shared among measures, queries or be defined at the model level.

1114. How would you create trailing X month metrics via DAX against a non- standard calendar?

The solution will involve:

1. CALCULATE function to control (take over) filter context of measures.
2. ALL to remove existing filters on the date dimension.
3. FILTER to identify which rows of the date dimension to use.

Alternatively, CONTAINS may be used:

- CALCULATE(FILTER(ALL('DATE'),.....))

1115. What are the different BI add-in to Excel ?

Below are the most important BI add-in to Excel:

- Power Query: It helps in finding, editing and loading external data.
- Power Pivot: Its mainly used for data modeling and analysis.
- Power View: It is used to design visual and interactively reports.
- Power Map: It helps to display insights on 3D Map.

1116. What is Power Pivot?

Power Pivot is an add-in for Microsoft Excel 2010 that enables you to import millions of rows of data from multiple data sources into a single Excel workbook. It lets you create relationships between heterogeneous data, create calculated columns and measures using formulas, build PivotTables and PivotCharts. You can then further analyze the data so that you can make timely business decisions without requiring IT assistance.

1117. What is Power Pivot Data Model?

It is a model that is made up of data types, tables, columns, and table relations. These data tables are typically constructed for holding data for a business entity.

1118. What is xVelocity in-memory analytics engine used in Power Pivot?

The main engine behind power pivot is the xVelocity in-memory analytics engine. It can handle large amount of data because it stores data in columnar databases, and in memory analytics which results in faster processing of data as it loads all data to RAM memory.

1119. What are some of differences in data modeling between Power BI Desktop and Power Pivot for Excel?

Here are some of the differences:

- Power BI Desktop supports bi-directional cross filtering relationships, security, calculated tables, and Direct Query options.

- Power Pivot for Excel has single direction (one to many) relationships, calculated columns only, and supports import mode only. Security roles cannot be defined in Power Pivot for Excel.

1120. Can we have more than one active relationship between two tables in data model of power pivot?

No, we cannot have more than one active relationship between two tables. However, can have more than one relationship between two tables but there will be only one active relationship and many inactive relationship. The dotted lines are inactive and continuous line are active.

1121. What is Power Query?

Power query is a ETL Tool used to shape, clean and transform data using intuitive interfaces without having to use coding. It helps the user to:

- Import Data from wide range of sources from files, databases, big data, social media data, etc.
- Join and append data from multiple data sources.
- Shape data as per requirement by removing and adding data.

1122. What are the data destinations for Power Queries?

There are two destinations for output we get from power query:

- Load to a table in a worksheet.
- Load to the Excel Data Model.

1123. What is query folding in Power Query?

Query folding is when steps defined in Power Query/Query Editor are translated into SQL and executed by the source database rather than the client machine. It's important for processing performance and scalability, given limited resources on the client machine.

1124. What are some common Power Query/Editor Transforms?

Changing Data Types, Filtering Rows, Choosing/Removing Columns, Grouping, Splitting a column into multiple columns, Adding new Columns ,etc.

1125. Can SQL and Power Query/Query Editor be used together?

Yes, a SQL statement can be defined as the source of a Power Query/M function for additional processing/logic. This would be a good practice to ensure that an efficient database query is passed to the source and avoid unnecessary processing and complexity by the client machine and M function.

1126. What are query parameters and Power BI templates?

Query parameters can be used to provide users of a local Power BI Desktop report with a prompt, to specify the values they're interested in.

- The parameter selection can then be used by the query and calculations.
- PBIX files can be exported as Templates (PBIT files).
- Templates contain everything in the PBIX except the data itself.

Parameters and templates can make it possible to share/email smaller template files and limit the amount of data loaded into the local PBIX files, improving processing time and experience.

1127. Which language is used in Power Query?

A new programming language is used in power query called M-Code. It is easy to use and similar to other languages. M-code is case sensitive language.

1128. Why do we need Power Query when Power Pivot can import data from mostly used sources?

Power Query is a self-service ETL (Extract, Transform, Load) tool which runs as an Excel add-in. It allows users to pull data from various sources, manipulate said data into a form that suits their needs and load it into Excel. It is most optimum to use Power Query over Power Pivot as it lets you not only load the data but also manipulate it as per the users needs while loading.

1129. What is Power Map?

Power Map is an Excel add-in that provides you with a powerful set of tools to help you visualize and gain insight into large sets of data that have a geo-coded component. It can help you produce 3D visualizations by plotting upto a million data points in the form of column, heat, and bubble maps on top of a Bing map. If the data is time stamped, it can also produce interactive views that display, how the data changes over space and time.

1130. What are the primary requirement for a table to be used in Power Map?

For a data to be consumed in power map there should be location data like:

- Latitude/Longitude pair
- Street, City, Country/Region, Zip Code/Postal Code, and State/Province, which can be geolocated by Bing

The primary requirement for the table is that it contains unique rows. It must also contain location data, which can be in the form of a Latitude/Longitude pair, although this is not a requirement. You can use address fields instead, such as Street, City, Country/Region, Zip Code/Postal Code, and State/Province, which can be geolocated by Bing.

1131. What are the data sources for Power Map?

The data can either be present in Excel or could be present externally. To prepare your data, make sure all of the data is in Excel table format, where each row represents a unique record.

Your column headings or row headings should contain text instead of actual data, so that Power Map will interpret it correctly when it plots the geographic coordinates. Using meaningful labels also makes value and category fields available to you when you design your tour in the Power Map Tour Editor pane.

To use a table structure which more accurately represents time and geography inside Power Map, include all of the data in the table rows and use descriptive text labels in the column headings, like this:

Year	UFO sightings	City
2006	43	Portland
2006	45	Seattle
2007	34	Portland
2007	23	Seattle

In case you wish to load your data from an external source:

1. In Excel, click Data > the connection you want in the Get External Data group.
2. Follow the steps in the wizard that starts.
3. On the last step of the wizard, make sure Add this data to the Data Model is checked.

1132. What is Power View?

Ans: Power View is a data visualization technology that lets you create interactive charts, graphs, maps, and other visuals which bring your data to life. Power View is available in Excel, SharePoint, SQL Server, and Power BI.

The following pages provide details about different visualizations available in Power View:

- Charts
- Line charts
- Pie charts
- Maps
- Tiles
- Cards
- Images
- Tables
- Power View
- Multiples Visualizations

- Bubble and scatter charts
 - Key performance indicators (KPIs)
1133. What is Power BI Designer?

Ans: It is a stand alone application where we can make Power BI reports and then upload it to Powerbi.com, it does not require Excel. Actually, it is a combination of Power Query, Power Pivot, and Power View.

1134. Can we refresh our Power BI reports once uploaded to cloud (Share point or Powebi.com)?

Ans: Yes we can refresh our reports through Data Management gateway(for sharepoint), and Power BI Personal gateway(for Powerbi.com)

1135. What are the different types of refreshing data for our published reports?

Ans: There are four main types of refresh in Power BI. Package refresh, model or data refresh, tile refresh and visual container refresh.

- Package refresh

This synchronizes your Power BI Desktop, or Excel, file between the Power BI service and OneDrive, or SharePoint Online. However, this does not pull data from the original data source. The dataset in Power BI will only be updated with what is in the file within OneDrive, or SharePoint Online.

- Model/data refresh

It refers to refreshing the dataset, within the Power BI service, with data from the original data source. This is done by either using scheduled refresh, or refresh now. This requires a gateway for on-premises data sources.

- Tile refresh

Tile refresh updates the cache for tile visuals, on the dashboard, once data changes. This happens about every fifteen minutes. You can also force a tile refresh by selecting the ellipsis (...) in the upper right of a dashboard and selecting Refresh dashboard tiles.

- Visual container refresh

Refreshing the visual container updates the cached report visuals, within a report, once the data changes.

1136. Is Power BI available on-premises?

No, Power BI is not available as a private, internal cloud service. However, with Power BI and Power BI Desktop, you can securely connect to your own on-premises data sources. With the On-premises Data Gateway, you can connect live to your on-premises SQL Server Analysis Services, and other data sources. You can also scheduled refresh with a centralized gateway. If a gateway is not available, you can refresh data from on-premises data sources using the Power BI Gateway – Personal.

1137. What is data management gateway and Power BI personal gateway?

Gateway acts a bridge between on-premises data sources and Azure cloud services.

Personal Gateway:

- Import Only, Power BI Service Only, No central monitoring/managing.
- Can only be used by one person (personal); can't allow others to use this gateway.

On-Premises Gateway:

- Import and Direct Query supported.
- Multiple users of the gateway for developing content.
- Central monitoring and control.

1138. What is Power BI Q&A?

Power BI Q&A is a natural language tool which helps in querying your data and get the results you need from it. You do this by typing into a dialog box on your Dashboard, which the engine instantaneously generates an answer similar to Power View. Q&A interprets your questions and shows you a restated query of what it is looking from your data. Q&A was developed by Server and Tools, Microsoft Research and the Bing teams to give you a complete feeling of truly exploring your data.

1139. What are some ways that Excel experience can be leveraged with Power BI?

Below are some of the ways through which we can leverage Power BI:

- The Power BI Publisher for Excel:
 - Can be used to pin Excel items (charts, ranges, pivot tables) to Power BI Service.
 - Can be used to connect to datasets and reports stored in Power BI Service.
- Excel workbooks can be uploaded to Power BI and viewed in the browser like Excel Services.
- Excel reports in the Power BI service can be shared via Content Packs like other reports.
- Excel workbooks (model and tables) can be exported to service for PBI report creation.
- Excel workbook Power Pivot models can be imported to Power BI Desktop models.

1140. What is a calculated column in Power BI and why would you use them?

Calculated Columns are DAX expressions that are computed during the model's processing/refresh process for each row of the given column and can be used like any other column in the model.

Calculated columns are not compressed and thus consume more memory and result in reduced query performance. They can also reduce processing/refresh performance if applied on large fact tables and can make a model more difficult to maintain/support given that the calculated column is not present in the source system.

1141. Why might you have a table in the model without any relationships to other tables?

There are mainly 2 reasons why we would have tables without relations in our model:

- A disconnected table might be used to present the user with parameter values to be exposed and selected in slicers (e.g. growth assumption.)
 - DAX metrics could retrieve this selection and use it with other calculations/metrics.
 - A disconnected table may also be used as a placeholder for metrics in the user interface.
 - It may not contain any rows of data and its columns could be hidden but all metrics are visible.

1142. What is the Power BI Publisher for Excel?

You can use Power BI publisher for Excel to pin ranges, pivot tables and charts to Power BI.

- The user can manage the tiles – refresh them, remove them, in Excel.
- Pinned items must be removed from the dashboard in the service (removing in Excel only deletes the connection).
- The Power BI Publisher for Excel can also be used to connect from Excel to datasets that are hosted in the Power BI Service.
- An Excel pivot table is generated with a connection (ODC file) to the data in Azure.

1143. What are the differences between a Power BI Dataset, a Report, and a Dashboard?

Dataset: The source used to create reports and visuals/tiles.

- A data model (local to PBIX or XLSX) or model in an Analysis Services Server
- Data could be inside of model (imported) or a Direct Query connection to a source.

Report: An individual Power BI Desktop file (PBIX) containing one or more report pages.

- Built for deep, interactive analysis experience for a given dataset (filters, formatting).
- Each Report is connected to atleast one dataset
- Each page containing one or more visuals or tiles.

Dashboard: a collection of visuals or tiles from different reports and, optionally, a pinned.

- Built to aggregate primary visuals and metrics from multiple datasets.

1144. What are the three Edit Interactions options of a visual tile in Power BI Desktop?

The 3 edit interaction options are Filter, Highlight, and None.

Filter: It completely filter a visual/tile based on the filter selection of another visual/tile.

Highlight: It highlight only the related elements on the visual/tile, gray out the non-related items.

None: It ignore the filter selection from another tile/visual.

1145. What are some of the differences in report authoring capabilities between using a live or direct query connection such as to an Analysis Services model, relative to working with a data model local to the Power BI Desktop file?

With a data model local to the PBIX file (or Power Pivot workbook), the author has full control over the queries, the modeling/relationships, the metadata and the metrics.

With a live connection to an Analysis Services database (cube) the user cannot create new metrics, import new data, change the formatting of the metrics, etc – the user can only use the visualization, analytics, and formatting available on the report canvas.

With a direct query model in Power BI to SQL Server, for example, the author has access to the same features (and limitations) available to SSAS Direct Query mode.

- Only one data source (one database on one server) may be used, certain DAX functions are not optimized, and the user cannot use Query Editor functions that cannot be translated into SQL statements.

1146. How does SSRS integrate with Power BI?

Below are some of the way through which SSRS can be integrated with Power BI:

- Certain SSRS Report items such as charts can be pinned to Power BI dashboards.
- Clicking the tile in Power BI dashboards will bring the user to the SSRS report.
- A subscription is created to keep the dashboard tile refreshed.
- Power BI reports will soon be able to be published to SSRS portal

1147. What is the general business need for Power BI?

Power BI is a powerful business analytics service provided by Microsoft that helps businesses to analyze data and share insights. The general business need for Power BI is to gain better visibility into business operations and make informed decisions based on data-driven insights. Some specific reasons why businesses use Power BI include:

Data visualization: Power BI allows businesses to create interactive and visually appealing reports and dashboards that enable users to quickly and easily understand complex data.

Data exploration: Power BI provides businesses with powerful tools for exploring and analyzing data in real-time, which can help identify trends and patterns that might otherwise go unnoticed.

Data collaboration: Power BI enables businesses to share data and insights with others in the organization, making it easier to collaborate and make decisions based on a common understanding of the data.

Data integration: Power BI can integrate with a wide range of data sources, including cloud-based and on-premises data, allowing businesses to access all their data in one place and gain a comprehensive view of their operations.

Business intelligence: Power BI provides businesses with advanced analytics capabilities, such as predictive modeling and machine learning, that can help them identify opportunities and make more informed business decisions.

Overall, the general business need for Power BI is to help organizations make sense of their data and turn it into actionable insights that can drive business growth and success.

1148. Which data sources are often associated with Power BI?

You can take the data and create robust reporting very easily. This assists in attracting new clients toward servicing and monitoring the customers already present. It also becomes possible to track information and set goals. Therefore, completely building an extraction, transformation and loading solution ultimately assists the management so they are able to make better decisions in the process. The return on investment when it comes to Power BI is also very high. Lastly, it makes some of the unwanted data into information, which can be utilized progressively.

There are a number of associated software tools that make Power BI more powerful. If you had to show some of the highlights of Power BI, it would be:

- Records: data ingestion from excel spreadsheets as well as shared Power BI desktop documents (.pbix) and comma separated value operations or (.csv) files
- Content packs: This is an accumulation of related reports or records. There are two types of substance packs. Connectors to databases and datasets like Azure SQL and SQL, Server Analysis Services.

1149. Can you please tell me what the distinctive channels in the Power BI reports.

- Visual level filters: these channels can chip away at the particular perception that decreasing the amount of data, which can be viewed. The visual level channels may also channel computations.
- Page level filters: these channels operate at the report level. They may have more diverse page level data.
- Report level type of filters: these are channels, which chip at the entire report and sift all the pages and representations that are included within the report. The realization is Power BI visual has certain operations, which make the act of sifting that much better. Visual associations are also valuable though they come with a number of constraints. You should be able to open a report created by another user and begin to play with the visual channels. Although there is no way of storing those channels within the spared report itself. The channel is also noticeable. Sometimes you may need a channel for the report in a variety of ways. However, there is no need for a visual sign to be indicated on the channel which is being connected.

Power BI provides an assortment of choice by which to channel a report, data and/or perception.

1150. Can you please tell me what some of the key competitive differences are between Power BI and Power BI Pro?

Power BI gives you a number of key features to assist with exploring data in a multitude of ways. Power BI Pro provides some of the better features within Power BI along with the additional type of features including the storage capacity, and scheduling data refresh more frequent than normal, live data sources with interactivity and much more.

1151. Do you know how much Power BI costs?

Microsoft offers several pricing options for Power BI depending on the needs of the organization. Here are the pricing options as of my knowledge cutoff date:

Power BI Free: This option is completely free, and users can create and share interactive reports and dashboards with others. However, it has limited functionality and data capacity.

Power BI Pro: This option costs \$9.99 per user per month and provides additional features such as more data capacity, collaboration tools, and more advanced sharing options.

Power BI Premium: This option starts at \$20,000 per year and provides organizations with dedicated cloud resources for their data, as well as additional features such as advanced analytics, higher data capacity, and the ability to share reports and dashboards with non-Pro users.

Power BI Embedded: This option is designed for developers who want to integrate Power BI into their own applications and pricing varies based on usage.

It's worth noting that Microsoft frequently updates its pricing and licensing options, so it's always a good idea to check their website for the latest information.

1152. Can you please tell me about the Power BI desktop application?

For the Power BI Pro package, there is a 60-day free trial period. After that, it is \$9.99 per month or there is an annual fee as well which you can opt for.

The Power BI desktop application is a simple way to access Power BI via your Mac or PC. It works pretty cohesively with the web application, giving all the advanced data exploration features you might see in the web app. Data modeling and shaping are included, well as report creation with highly interactive and beautiful visualizations. It is possible to save your work to a number of export types and then publish your data and reporting to the Power BI site to share with team members.

1153. Can you please tell me what DAX is?

- It works on the column values
 - DAX cannot modify or insert information or data
- In order to conduct basic math or calculations and data analysis when it comes to power pivot, there is a need to use Data Analysis Expression or DAX.
- It is possible to calculate calculated column and measures but it is not possible to calculate rows with the use of DAX

1154. What are some of the core parts of Microsoft self service business intelligence solutions?

Microsoft has two parts for self-service business intelligence solution:

- Excel BI toolkit: this allows the customers to initiate an interactive report through the importation of data from different sources and modeling the data according to the requirements of the report.
- It is an online solution, which allows one to share the interactive reports and queries, which have been created with the use of Excel BI Toolkit.

1155. Can you please tell me what the self-service business intelligence offering is?

- Self-service business intelligence or SSBI refers to the approach of data analytics which allows the users or colleagues to filter, analyze and segment data without an in-depth technical knowledge of business intelligence and/or statistical analysis.
- SSBI has made it pretty easy for the general end users of company or organization to access information from it and then create data visuals in order to get key business insights.
- Anyone that has a basic understanding of the data should be able to create reports in order to come up with intuitive and sharable types of dashboards.

1156. What are the building blocks when it comes to Power BI?

- Visualizations; this is a visual representation of the data which can be line graphs, pie charts, side bar charts, graphical representations of the source data and tree maps.
- Datasets: a dataset refers to a collection of data, which the Power BI software may utilize in order to come up with their visualizations.
- Dashboards: the dashboard is a type of single layer presentation interface that provides a variety of visualization modes for you. If you the user integrates one or more of the visualizations into one of the page layers, it becomes a pretty fancy and powerful display. For example: a sales dashboard may have pie charts, geographical maps as well as, bar charts.
- Tiles: A tile refers to a single visualization made within a report. An example in this case would be a pie chart in dashboard or report.

1157. What are the content packs when it comes to Power BI?

Content packs are essentially prebuilt pieces of logic for a number of your services as part of the Power BI platform. A customer to one of the supported content pack services should be able

to connect their account from Power BI admin panel in order to see the data. Microsoft is one of the more popular providers of this. They've released content packs for Marketo, Salesforce.com, Adobe Analytics, Mobile Engagement, Azure and many more. The organizational content packs allow the users, system integrators (programmers) and BI professionals with the tools for building their content packs in order to share purpose built dashboards, datasets and reports within their organization.

1158. Can you please tell me how you use the FILTER function?

- For example: FILTER(MyTable,[SalesSheet] > 1000)

Ans. The FILTER function returns a table with a set of conditions applied for each of the source table rows.

- FILTER is the function that refers to an "iterator" so it can negatively effect the returned data set.

1159. What are the various names of Power BI?

- Excel BI Toolkit: In the heart of Microsoft's Power BI would be the Excel BI Toolkit. This is what provides it with power and boost. This comprises of Excel and a number of add-ins, which allow you to shape, create and project the data as analyzed and consider some of the components. There are four main elements, which would be: Power View, Power Pivot, Power Map and Power Query.
- Power View: this mainly analyses and represents the data in a means that is interactive as a data visualization manner with the use of Power View.
- Power Query: this investigates; change public and internal information as well as access to the information sources.
- Power Pivot: this works as an information modeling for the in-memory analytics.
- Power Map: Assist in bringing information to bear with visualization approaches, which are interactive.

1160. Name some of the filters that you can utilize with the Power BI Reports.

- Visual filters
- Drill filters
- Page filters
- Report filters

1161. Can you tell me how well you understand the Power BI designer?

This is a solo application that is associated with Power BI. And yes, it is entirely possible to make Power BI reports and then transfer them to PowerBI.com pretty easily.

1162. What are the main differences between the Power BI gateway and Data management gateways?

The data management gateway is a component which gathers data continuously and exposes tables and relevant data to the user. In the case of Power BI gateway, this is software, which considers the on-premise Network. This data is not generally stored in the cloud for security purposes. It can be utilized for a single data or different data sources. The data is usually stored within an encrypted gateway cloud service.

1163. Can you please tell me what the selection pane is inside of Power BI?

It can generally be used for controlling the tab order between the visuals on the data viz page. You can actually combine two or more of the visual pages into one single visual group if you'd like. This can be used to select data in a single visual for highlighting purposes and drill down a bit further from there.

1164. What is your opinion of the dynamic filtering option within Power BI?

Dynamic filtering is a useful feature in Power BI that allows users to interactively filter data based on the selections made in other visuals or slicers. This means that when a user selects a value in one visual or slicer, other visuals on the same report page will automatically update to reflect that selection.

Dynamic filtering is especially useful in creating interactive dashboards and reports, where users can explore data by making selections and seeing the results in real-time. It can also help users save time by avoiding the need to manually update multiple visuals with the same filter.

Overall, the dynamic filtering option within Power BI is a powerful tool that can help users gain insights into their data more efficiently and effectively.

1165. What data sets are generally used for creating a visual dashboard within the streaming data tiles?

Streaming data tiles in Power BI are designed to show real-time data, such as data from sensors, IoT devices, or other streaming sources. The types of data sets that are generally used for creating a visual dashboard within the streaming data tiles will depend on the specific application and the data sources being used. Here are some examples:

Sensor data: Streaming data tiles can be used to visualize real-time data from sensors, such as temperature, humidity, or pressure sensors. This can be useful in industries such as manufacturing, agriculture, or energy management.

IoT data: Streaming data tiles can also be used to monitor and visualize data from IoT devices, such as smart home devices, wearables, or vehicles. This can be useful in industries such as healthcare, transportation, or consumer electronics.

Web analytics data: Streaming data tiles can also be used to monitor and visualize real-time web analytics data, such as website traffic, clickstream data, or social media analytics. This can be useful in industries such as marketing, e-commerce, or digital media.

Financial data: Streaming data tiles can be used to monitor and visualize real-time financial data, such as stock prices, currency exchange rates, or trading volumes. This can be useful in industries such as finance, investment management, or trading.

Overall, the types of data sets that are used for creating a visual dashboard within the streaming data tiles will depend on the specific use case and the types of streaming data sources that are available.

1166. What would you consider to be the normal amount of table capacity for gathering of importing data in Power BI?

The "Summarize ()" function

Fundamental group work in produced within SSAS

New gathering by work for SSAS and Power BI Desktop; more productive Indicate bunch by different segments, table and articulations.

The prescribed practice would be to determine table and gathering by sections not metrics. You may utilize ADDCOLUMNS work. SUMMARIZECOLUMNS

1167. Can you tell me what some of the advantages are of utilizing the variables within DAX?

- Through assessing a variable, the variable can be reused in different circumstances or instances within the DAX articulation. Because of this, the manner stays away from the extra inquiries of the particular source database.
- Factors may make the DAX articulations instinctive so they can be deciphered.
- Factors are just perused to their measure or question as they cannot share among the measures, inquiries or characterize at the model level.

1168. Can you tell me what Power Pivot is?

Power Pivot is included as part of the software for Microsoft excel which allows you to import several lines of data from the different data sources into a solitary Excel exercise manual. This generally provides the opportunity to make connections between data and make figures based on that. It can also be used to measure the utilizing equations as well as the assembly of Pivot Tables and Pivot Charts. Additionally, it is possible to break down the data so that a team member can settle on the convenient alternatives without the need for support.

1169. Tell me what some of the key contrasts in the display of data between Power BI Desktop and Power Pivot for Excel?

- Power BI Desktop underpins the multidirectional connections, figures/tables, security and Direct Query options you've setup.
- Power Pivot for Excel also has single bearing connections, ascertained segments and underpins the import mode, as it would be. The parts related to security cannot be characterized in Power Pivot for Excel.

1170. Can you tell me what the Power Pivot Data Model is?

This is a data model which is comprised of data write tables as well as table relationships. The data tables usually are there for holding data together when it comes to business query that is being made.

1171. When is the x-Velocity examination tool utilized for Power Pivot?

This is a core part of the engine that is behind "power rotate", or what would be the x- Velocity in memory examination engine. It can assist with a big data measurement because it stores data in a column database and then in the memory examination, which brings a much faster preparation of data as it places most of the query power to your machines RAM memory. It is an efficient way to analyze large data.

1172. Can you tell me if you would have access of one dynamic connection between two tables in the data model in the intensity rotate process?

It is generally not possible to have access to one dynamic connection between two tables. However, it usually is possible to have access of one connection between two tables. There is usually just one single dynamic relationship available.

1173. Can you please describe to me what Power Query is ?

Power Query is the ETL tool used to shape, clean and alter the data via an interface. It allows us the users to import data from an extensive variety of sources. You can shape the data according to necessity through evacuation and inclusion of the data.

1174. What are data goals when it comes to Power Queries?

- Loading to a table within a worksheet
- Loading to the Excel Data Model

1175. Can you tell me what question collapsing within Power Query is?

In Power Query, question collapsing refers to a feature that allows you to collapse a table to show only the unique values in one or more columns. This is useful when you have a large table

with repeated values and you want to quickly see the unique values and their corresponding counts.

To collapse a table in Power Query, you can follow these steps:

Select the column or columns that you want to collapse.

Go to the "Transform" tab in the ribbon and click on the "Group By" button.

In the "Group By" dialog box, select the column or columns that you want to group by and choose the aggregation function you want to apply to the other columns. For example, you might choose to count the number of rows for each unique value in the grouping column.

Click on the "OK" button to apply the grouping and collapse the table.

Once the table is collapsed, you can further refine the results by applying filters or sorting the data. The collapsed table can be loaded into a new worksheet or merged with other tables in your data model.

1176. Can you tell me if Power BI available for on-prem (or on premises)?

- SQL Server mobile reports on the iPad
- SQL Server mobile reports on the iPhone.

It would be described as two goals for a yield that are obtained from the question you asking the data.

Query collapsing is the process that happens when steps characterized in Power Query/Query Editor are converted into SQL or MySQL queries and then executed. It is important for the preparing of execution and versatility given the restricted assets on the customer machine.

Ans. No, it is not available as a private server setup. This is usually called "on-prem". However, with Power BI, you do have a secure connection to your on-prem data sources if you'd like to set that up. With the on-premises Data Gateway, you should be able to connect your on-premises SQL Server and some of the other data sources you have available. You may also want to schedule refreshes with a centralized type of gateway. In the event that a gateway is not there then you can refresh the data from the on-premises data sources with the use of Power BI Gateway Personal. It is also possible to view the on-premises SQL server mobile reports with the Power BI iOS application:

1177. Does Power BI support general mobile devices like Android and iOS?

Power BI does, in fact, have native applications for Android phones, iOS devices and some of the Windows 10 devices. You can find the application in many of the branded app stores.

1178. The Power BI Desktop's Software License Terms show that one may install and use one copy of the software on the premises. True or false?

The question is whether this means you are limited to one copy of Power BI Desktop for the entire company. Power BI desktop's use rights do not limit the individual to one copy of Power BI Desktop for the entire firm. Each of the individual users at the firm is able to install and use one copy on their premises.

1179. Can you tell me what the query folding tool is in Power Query?

Query folding is the process of going through the steps defined in Power Query which translated SQL executed by the source database. This is opposed to the users local machine. It is important for processing performance as well as, scalability considering the limited resources when it comes to the client machine.

1180. What are some of the most commonly used Power Query Data Transforms?

Power Query provides a wide range of data transformation options to clean, reshape, and combine data from various sources. Some of the most commonly used Power Query data transforms include:

Filtering rows based on conditions

Removing duplicates

Splitting columns into multiple columns

Merging or appending tables

Pivoting or unpivoting columns

Grouping and summarizing data

Adding calculated columns

Renaming columns

Changing data types

Filling or replacing missing values

Extracting text using regular expressions

Combining multiple queries using joins or unions

These transforms can be accessed and applied through the Power Query Editor, which provides a visual interface to create, modify, and manage data transforms.

1181 . Can the SQL and Power Query Editor be used together?

Absolutely, usually a SQL statement can be defined as the source of the Power Query function. That would be one of the better practices to make certain that an efficient type of database query is passed to the source in order to avoid the unnecessary type of complexity or processing by the client machine and the M function.

1182. What are the main query parameters and Power BI Templates you have access to?

The Query parameters are parameters which can be used to provide the user of the local Power BI desktop report via a prompt that is geared to be specified for the values to access.

Parameters and templates as well can make it possible to share or email smaller template files and limit the amount of data loaded within the local PBIX files, thus improving the time for processing and experience.

1183. What is the main language, which is used in Power Query?

The main language used in Power Query is called the "M" language. M is a functional programming language that is used to define data transforms and perform data manipulations in Power Query. It is a powerful language that supports a wide range of data types, operators, and functions to manipulate data, including text, numbers, dates, and lists. M is designed to be both expressive and concise, allowing users to perform complex data transformations with ease. In addition, Power Query also supports a range of other programming languages, including SQL, C#, and Python, to extend the functionality and perform custom data processing tasks.

1184. Can you tell me why you would need Power Query when Power Pivot can do the same?

Power Query is apparently a self-service type of ETL tool that runs as an excel add-in. It allows the end users to pull their data from different sources and manipulate it into whatever forms they wish, which would best suit its needs and then load it into Excel or some other type of data export source. Its usually better to use Power Query because it allows you to load the data much easier and also manipulate it according to the needs of the user.

1185. Can you tell me what the Power Map is within Power BI?

The Power Map refers to a Microsoft Excel plugin that provides a set of functions which usually assists in the dat visualizations and data insight from large data sets. It can also assist in the production of 3D visualizations if you choose. It can plot up to a million data points in the form of a heatmap, into columns and also bubble maps if you choose. If the data is time coded, then it may also give some interactive views, which illustrate the manner the data is changing over the course of both space and time.

1186. What are some of the main requirements for a table to be used within Power Map for Power BI?

For the data to be consumed in power map there ought to be location data such as the:

- Latitude and longitude pair
- Street, country, region, zip, postal code and state, which may be geo-located with the assistance of Bing. It also has to have data that can be in the form of a latitude/ longitude pair though this is not necessarily one of the requirement. It is possible to use some of the address fields instead like the country, city, street/ region, zip code. These can be geo- located with Bing.

1187. What are the data hotspots when it comes to Power Map?

The data can either be available in Excel or it can be available remotely. For simplicity of setting up your data make sure the majority of your data is inside an Excel table where each of the columns is setup to a unique record. The segment headings or the column headings ought to have a message as opposed to genuine data with the objective that Power Map is there to decipher it in an accurate manner especially when it plots the geographic directions. The utilization of some of the important names makes the classification fields accessible in the Power Map Tour Editor Sheet. In order to utilize table structures, both time and geology within the Power Map have to incorporate the better part of the data within the table lines. In order to do this you might have to:

- You wish to stack the data from an outside source.
- Take the means or results within the wizard that begins.
- In Excel click on the menu data > the association you need within the Get external data gathering.
- On the last advance of the wizard, make certain Add this data to the Data Model is checked.

1188. Is it possible to invigorate the Power BI reports once they are uploaded to the cloud?

Yes, absolutely. It is definitely possible to revive the reports through a Data Management gateway when it comes to SharePoint and Power BI Personal gateway.

1189. What are some of the unique data invigorations used for generating the distributed reports within Power BI?

There are four main types of invigorating steps when it comes to Power BI. Bundle invigorate can model or revive your data.

- Visual compartment - This type of invigorating is a visual compartment refresh stored in a report type within a report once the data starts to change. In order to find out if the data invigorate worked and view how to actualize the data, you can check on the accompanying connections.
- Tile Invigorate - Tile invigorate refreshes the stores for the tile visuals on the dashboard once the data begins to change. This happens at regular type intervals. You may also constrain a tile for invigorating through the option of the ellipsis (which is three dots like this ...)on the upper right of the dashboard and choosing Refresh Dashboard Tiles.
- Bundle Invigorate – This would then synchronize the Power BI Desktop or Excel document between the Power BI benefit and OneDrive, or SharePoint Online. This does not pull the data from the first data source though. When the data comes to Power BI it can refresh it once it is inside the OneDrive record.
- Display/data revive – This refers to the type of revising to the dataset inside the Power BI benefit with data from the initial data source. This would either be finished by utilizing booked revive or invigorate now. That usually requires an entry point for the for the on- premises data sources, in the form of access tokens or login credentials.

1190. Explain the Power BI Designer.

Power BI Designer, a standalone app that is used to create reports in Power BI and to upload it to Powerbi.com. It is a combination of Power View, Power Pivot, and Power Query.

1191. Is there any process for refreshing Power BI reports once uploaded to the cloud?

Of course, Power BI reports can be refreshed with Data Management Gateway and Power BI Personal Gateway.

1192. What is the major difference between Power BI personal Gateway and Data Management Gateway?

Power BI Personal Gateway is used for reports that are deployed in Powerbi.com. Data management, on the other hand, is an app installed the gateway on source data machines to deploy reports on Sharepoint and schedule to refresh automatically.

1193. What is the use of split function?

The split function is used for splitting the string database on the given delimiter.

1194. Name all the platforms for which the Power BI app is available.

Power BI app is available for:

- Android
- iPhone and iPad
- Windows tablets and Windows Desktops
- Coming for Windows phone soon

1195. Differentiate between older and newer Power BI.

There is a new design tool that is used in the new Power BI called Power BI Desktop. It is a standalone designer, including Power Pivot, Power View and Power Query in the back end. Whereas, Older Power BI consists of excel add-ins. In the newer Power BI version, there are several graphs available including treemap, line area chart, waterfall, combo chart, etc.

1196. Is it possible in the power pivot data model to have more than one active relationship between two tables?

No, it is not possible. There cannot be more than one active relationship in the power pivot data model between two tables. It is possible to have only one active and many inactive relationships.

1197. What is the purpose of the 'Get Data' icon in Power BI?

When users in Power BI click on the icon “Get Data”, a drop-down menu appears displaying all data sources from which data can be ingested. Data can directly get ingests from any source such as files in Excel, XML, PDF, JSON, CSV, and SharePoint folder databases and formats such as SQL, SQL Server Analysis Services, IBM, Access, Oracle, MySQL, and much more.

1198. What is Row-level Security?

Row-level security restricts the data that users view and access, based on filters. To configure row-level security, users can define rules and roles within Power BI Desktop and publish them to Power BI Service. Also, the `username()` function can be used to restrict data in the table to the current user.

However, to enable row-level security, a Power BI Pro subscription account is essential, and Excel sheets can be used when converted to the .pbix file format.

1199. What are the general data shaping techniques?

The common data shaping techniques are:

- Removing Columns and Rows
- Adding Indexes
- Applying for a Sort Order

1200. Which data sets can be used to create dashboards with streaming data tiles?

- Streaming datasets
- Hybrid Datasets

1201. What are the KPIs in Power BI?

KPIs are Key Performance Indicators, which evaluates the organization’s performance in distinct areas by evaluating measurable goals and values. A KPI has a measure or base value that is evaluated against target values. It includes a comparison of the performance with the target. The KPI also helps you evaluate the analysis performances with their graphical representation. Thus, KPIs will show whether your goals have met or not.

1202. What could be the difference between Distinct() and Values() in DAX?

The `Distinct()` and `Values()` are the same. The only difference between them is that the function values don’t calculate null values whereas `distinct ()` calculates even the null values.

1203. State the advantages of the Direct query method.

The advantages of Direct query method are listed as follows:

User can build huge data sets data visualizations using the Direct Query Method, but Power BI desktop supports data visualizations on smaller sets alone. There is no limit to the dataset for direct query method and a 1GB dataset limit is not applicable in this method.

1204. What is What if the parameter in power BI?

If you want to put a scenario and based on that if you wanted to see the visuals, the best is What if parameter. It helps you to forecast data and perform advanced analytics. For example, if you have set up the product discount from the what-if parameter from 1 to 10. And user can change the values and see the changes in profit, sales, revenue, margin etc. that helps in detailed analysis.

1205. What is the incremental refresh?

Incremental refresh is used to refresh the newly added data to avoid truncating and loading data.

1206. What are the three main tabs in Reports development Window?

The major tabs in Reports development Window are as follows:

- Relationship tab
- Data Modeling Tab
- Report Tab

1206. How is data security implemented in Power BI ?

Power BI can apply Row Level Security roles to models.

- A DAX expression is applied on a table filtering its rows at query time.
- Dynamic security involves the use of USERNAME functions in security role definitions.
- Typically a table is created in the model that relates users to specific dimensions and a role.

1207. What are many-to-many relationships and how can they be addressed in Power BI ?

Many to Many relationships involve a bridge or junction table reflecting the combinations of two dimensions (e.g. doctors and patients). Either all possible combinations or those combinations that have occurred.

- Bi-Directional Crossfiltering relationships can be used in PBIX.
- CROSSFILTER function can be used in Power Pivot for Excel.
- DAX can be used per metric to check and optionally modify the filter context.

Chapter 12 - Forecasting

Forecasting is a method of predicting the future looking at the historic trend. You can think of forecasting as a scientific approach to understand the previous values and then to predict on the same line.

A simple example - Suppose you have a burger outlet i.e. Burger Singh. Noe the owner of the outlet knows that on an average he needs to prepare 500 burgers on the weekdays and 700 on the weekends. But recently he has observed that somedays the burgers are getting sold out at 8 pm itself whereas somedays he is facing a loss due to over preparation.

How would an analyst try to solve this problem?

To start with he will first try to look at the average and moving average. Then he might like to check if the trend is having any seasonality or not, he might also be interested in understanding the correlation between any event happening in the city vs the sold burgers. Don't worry, we will deal with each term in detail in the upcoming section.

But, just understand the underlying principles behind forecasting. As usual we will try to clear out concepts with the help of questions. So, let's get started.

1208. What is a time series?

A times series is a type of data where in all the data points are in a sequence that is dependent on time. That is, the data is in the formed over some intervals of time. For example, stock market price, or weather forecast.

Basically, a timeseries is any two column and multiple rows dataset which contains any time metrics in one column like hour, day, date, month, year, etc. and the second column will have the metrics that you want to forecast, like number of burgers, number of sales, etc.

1209. What is time series analysis?

Time series analysis is the analysis of time series data to find out if there are any existing trends, seasonality or any other meaningful characteristics to the data.

1210. How is time series analysis performed?

Time series analysis is performed using different statistical and mathematical methods to find out any relations between one or more variables with time to perform future predictions.

1211. What is a trend in time series?

A trend in time series refers to the increase or decrease of the values present in the data, that is, there is a positive or a negative slope present in general over a long duration of time.

1212. What is a seasonality in time series?

Seasonality in time series refers to a specific time interval over which the data experiences the same changes every time. Which means that a particular pattern is observed in the time series at some regular time periods. Example: A stock price that experiences a rapid increase in its price every year in the month of June.

1213. What is stationarity in time series?

Stationarity in time series refers to a time series where the mean and variance of the data do not change over time. Note that this does not mean the data itself does not change, it doesn't even mean that the data changes itself in a uniform way. It simply means that the change in data is such that the statistical indicators remain the same.

1214. What is time-series forecasting, and how is it used in data analysis?

Time-series forecasting is a statistical technique used to predict future values of a variable based on its past values. It is used in data analysis to make informed decisions and plans based on trends, patterns, and seasonality in the data.

1215. Can you explain the differences between time-series forecasting and other types of data analysis techniques?

Time-series forecasting is different from other types of data analysis techniques, such as regression analysis and classification analysis, in that it specifically focuses on analyzing patterns and trends in time-ordered data. Unlike regression analysis, which seeks to establish a causal relationship between two variables, time-series forecasting is concerned with predicting future values of a variable based solely on its past values.

1216. What is autoregression, and how is it used in time-series forecasting?

Autoregression is a time-series forecasting technique that uses a regression model with one or more lagged values of the variable being predicted as the input. It is used to model the relationship between an observation and a number of lagged observations.

1217. How can you evaluate the accuracy of a time-series forecasting model?

There are several ways to evaluate the accuracy of a time-series forecasting model, including

Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Another approach is to use forecast error measures, such as the Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE).

1218. Can you explain the differences between moving average and exponential smoothing methods for time-series forecasting?

Moving average and exponential smoothing are two methods used for time-series forecasting. Moving average takes the average of a certain number of the most recent observations as the prediction for the next period. Exponential smoothing is similar to moving average, but it gives more weight to the most recent observations, making it more responsive to recent changes in the data.

1219. How would you use ARIMA models for time-series forecasting in Python?

To use an ARIMA model for time-series forecasting in Python, you can use the statsmodels library. ARIMA stands for Autoregressive Integrated Moving Average and is a type of time-series model that uses three parameters: p, d, and q. p represents the order of the autoregressive model, d represents the degree of differencing, and q represents the order of the moving average model.

1220. Can you explain how to use Prophet library for time-series forecasting in Python?

Prophet is a time-series forecasting library developed by Facebook that uses an additive model to predict future values. To use Prophet for time-series forecasting in Python, you can install the library and follow the documentation to fit the model to the data and make predictions.

1221. How can you handle missing values in a time-series data set?

To handle missing values in a time-series data set, you can use interpolation techniques, such as forward or backward filling, or impute the missing values using statistical methods, such as mean or median imputation.

1222. Can you explain the differences between additive and multiplicative seasonality in time-series forecasting?

Additive seasonality refers to a seasonal effect that is constant across the range of the data, while multiplicative seasonality refers to a seasonal effect that increases or decreases in proportion to the level of the data. The choice between additive and multiplicative seasonality depends on the nature of the time-series data and can be determined through visual inspection or statistical tests.

1223. How can you use machine learning algorithms like random forests and neural networks for time-series forecasting?

Machine learning algorithms, such as random forests and neural networks, can be used for time-series forecasting by training the algorithm on historical data and using it to make predictions on new data. These algorithms can handle complex relationships between variables and capture nonlinear patterns in the data.

1224. Can you explain how to use cross-validation to evaluate time-series forecasting models?

Cross-validation is a technique used to evaluate the performance of a time-series forecasting model by testing it on a subset of the data that was not used to train the model. This can help to identify whether the model is overfitting or underfitting the data.

1225. How would you deal with non-stationarity in time-series data?

Non-stationarity refers to a time-series data set where the mean, variance, or autocorrelation structure changes over time. To deal with non-stationarity, you can use techniques such as differencing, which involves taking the difference between consecutive values to remove the trend, or seasonal differencing, which involves taking the difference between values at the same point in the seasonal cycle to remove seasonality.

1226. Can you explain how to use state space models for time-series forecasting in Python?

State space models are a flexible framework for modeling time-series data that allows for complex patterns to be represented. The basic idea is that the underlying state of a system evolves over time according to some set of equations, and the observed data is generated from this underlying state. The goal of state space modeling is to estimate the unobserved state of the system, given the observed data.

In Python, state space modeling can be done using the statsmodels library, which provides a range of models including the Kalman filter and its extensions. The basic process for using state space models in Python involves specifying the model, fitting it to the data, and then using the model to make predictions.

1227. How can you use time-series forecasting to predict demand for a product or service?

Time-series forecasting can be used to predict demand for a product or service by analyzing historical sales data and using it to make predictions about future sales. There are a variety of methods that can be used for this type of forecasting, including ARIMA models, exponential smoothing, and machine learning algorithms.

To use time-series forecasting to predict demand, you would start by collecting historical sales data, which can be broken down by various factors such as time period, product or service category, region, or customer demographic. This data is then used to build a forecasting model, which can be calibrated using different methods depending on the specific requirements of the business.

Once the model has been trained on historical data, it can be used to make predictions about future sales. These predictions can then be used to inform business decisions such as inventory management, marketing strategies, and resource allocation.

1228. How can you use time-series forecasting to identify trends and patterns in financial data?

Time-series forecasting can be used to identify trends and patterns in financial data by analyzing historical data and making predictions about future trends. This can be useful for a variety of purposes, including forecasting stock prices, predicting economic indicators, and identifying opportunities for investment.

To use time-series forecasting for financial data analysis, you would start by collecting historical data on the financial metric of interest, such as stock prices, interest rates, or GDP. This data is then used to build a forecasting model, which can be calibrated using a range of methods such as ARIMA, exponential smoothing, or machine learning algorithms.

Once the model has been trained on historical data, it can be used to make predictions about future trends in the financial metric. These predictions can then be used to inform investment decisions, such as which stocks to buy or sell, or to forecast economic indicators such as inflation or unemployment.

1229. What is an ARIMA model, and how is it used for time-series forecasting?

An ARIMA (Autoregressive Integrated Moving Average) model is a statistical method used for time-series forecasting. It uses the past values of a time-series to predict future values. ARIMA models are based on the assumption that the time-series is stationary, which means that the statistical properties of the time-series (such as the mean and variance) remain constant over time. The model is composed of three components: the autoregressive (AR) component, the integrated (I) component, and the moving average (MA) component. The AR component models the dependence of the current value on the past values, the MA component models the dependence of the current value on the past errors, and the I component models the non-stationarity of the time-series.

1230. Can you explain the differences between AR, MA, and ARMA models, and when would you use each?

The autoregressive (AR) model is a model that uses the past values of the time-series to predict

future values. The moving average (MA) model uses past errors to predict future values. The ARMA (Autoregressive Moving Average) model combines both the AR and MA components. The AR model is used when the time-series is auto-correlated, meaning that the values are dependent on past values. The MA model is used when the errors in the time-series are auto-correlated. The ARMA model is used when both the time-series and the errors are auto-correlated.

1231. How would you determine the order of an ARIMA model for a given time-series data set?

The order of an ARIMA model is determined by analyzing the autocorrelation and partial autocorrelation plots of the time-series. The autocorrelation plot (ACF) shows the correlation between the current value and past values of the time-series. The partial autocorrelation plot (PACF) shows the correlation between the current value and past values, while removing the influence of the intermediate values. By analyzing these plots, we can determine the order of the AR, MA, and I components of the ARIMA model.

1232. Can you explain the differences between stationary and non-stationary time-series data, and how does it impact the use of ARIMA models?

A stationary time-series is one in which the statistical properties such as the mean and variance remain constant over time. Non-stationary time-series are those where the statistical properties change over time. ARIMA models require the time-series to be stationary, and if the time-series is non-stationary, then the ARIMA model can result in incorrect predictions. To use ARIMA models on non-stationary time-series, we can first make the time-series stationary by taking differences of the time-series or using other methods.

1233. How can you use ACF and PACF plots to select the parameters for an ARIMA model?

The autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots can be used to select the parameters for an ARIMA model. The ACF plot shows the correlation between the current value and past values of the time-series. The PACF plot shows the correlation between the current value and past values, while removing the influence of the intermediate values. By analyzing these plots, we can determine the order of the AR, MA, and I components of the ARIMA model.

1234. Can you explain the Box-Jenkins method for building ARIMA models?

The Box-Jenkins method is a three-step process for building ARIMA models: model identification, parameter estimation, and model checking. In the identification step, the time-series data is examined to determine the order of the ARIMA model, including the number of autoregressive (AR) terms, the number of differencing (I) terms, and the number of moving average (MA) terms. In the estimation step, the model parameters are estimated using

maximum likelihood estimation. In the model checking step, the fitted model is evaluated for goodness of fit and modified if necessary.

1235. How would you handle seasonality in a time-series data set when using ARIMA models?

To handle seasonality in a time-series data set when using ARIMA models, a seasonal ARIMA (SARIMA) model can be used. SARIMA models extend ARIMA models to include seasonal components, such as seasonality in weekly, monthly, or quarterly data. These models include additional seasonal parameters, such as the seasonal period and the seasonal differencing parameter.

1236. Can you explain how to fit an ARIMA model in Python using the Statsmodels library?

To fit an ARIMA model in Python using the Statsmodels library, you can use the ARIMA class. First, you need to identify the order of the model using ACF and PACF plots or by performing a grid search over different parameter combinations. Then, you can create an instance of the ARIMA class and fit the model to the data using the fit method.

1237. How would you evaluate the accuracy of an ARIMA model, and what metrics would you use?

To evaluate the accuracy of an ARIMA model, you can use metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). These metrics can be calculated by comparing the predicted values to the actual values of the time-series data.

1238. How can you use ARIMA models for anomaly detection in time-series data?

ARIMA models can be used for anomaly detection in time-series data by identifying data points that deviate significantly from the model's predictions. This can be done by setting a threshold for the prediction errors or using statistical tests, such as the Granger causality test, to detect significant changes in the data.

1239. Can you explain how to use SARIMA models to incorporate seasonal effects in ARIMA modeling?

SARIMA models can be used to incorporate seasonal effects in ARIMA modeling by including seasonal differences and seasonal AR and MA terms in the model. The seasonal order of the model is denoted by $(p, d, q)(P, D, Q)m$, where p , d , and q are the non-seasonal parameters, P , D , and Q are the seasonal parameters, and m is the seasonal period.

1240. How would you use ARIMA models for forecasting multiple time-series data sets simultaneously?

To use ARIMA models for forecasting multiple time-series data sets simultaneously, a vector autoregression (VAR) model can be used. VAR models extend ARIMA models to include multiple time-series data sets by modeling the relationships between them.

1241. Can you explain the differences between ARIMA models and exponential smoothing methods for time-series forecasting?

The main difference between ARIMA models and exponential smoothing methods is that ARIMA models rely on the historical values of the time-series data to make predictions, while exponential smoothing methods use a weighted average of the past values. ARIMA models are more flexible and can handle a wider range of time-series patterns, while exponential smoothing methods are simpler and more computationally efficient.

1242. How would you use ARIMA models for forecasting in the presence of missing data in a time-series data set?

To use ARIMA models for forecasting in the presence of missing data in a time-series data set, the missing values can be imputed using interpolation techniques or using a state space model that can handle missing values. Another option is to use a model that can handle irregularly spaced time-series data, such as a Bayesian structural time series model.

1243. Can you explain the limitations of ARIMA models and when it may not be appropriate to use them for time-series forecasting?

The limitations of ARIMA models include the assumption of linearity, stationarity, and independence of the errors. Nonlinear relationships and non-stationary time-series data may require more complex models. Additionally, the accuracy of ARIMA models may be limited by the amount and quality of the historical data available, and the model may not perform well in the presence of outliers or sudden changes in the data.

ARIMA Model

1244. What are the key components of an ARIMA model?

The key components of an ARIMA model are:

Autoregressive (AR) component: captures the dependence of the current value on its past values.

Integrated (I) component: deals with the non-stationarity of the time series by differencing the series.

Moving Average (MA) component: captures the dependence of the current value on its past errors.

Order: represents the number of terms in each component.

1245. What are the assumptions underlying an ARIMA model?

The assumptions underlying an ARIMA model are:

The time series is stationary or can be made stationary by differencing.

The errors are normally distributed and have constant variance.

The errors are independent and identically distributed (iid).

1246. How do you determine the appropriate order of differencing for an ARIMA model?

The appropriate order of differencing for an ARIMA model can be determined by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the time series. If the ACF shows a slow decay or a sinusoidal pattern and the PACF shows a significant spike at lag 1, then the time series may need to be differenced. The appropriate order of differencing can be determined by differencing the time series and analyzing the ACF and PACF plots of the differenced series.

1247. How do you determine the appropriate order of autoregression and moving average terms for an ARIMA model?

The appropriate order of autoregression and moving average terms for an ARIMA model can be determined by analyzing the ACF and PACF plots of the time series. If the ACF shows a significant spike at lag k and the PACF shows a slow decay, then the time series may need an autoregressive term of order k. If the PACF shows a significant spike at lag k and the ACF shows a slow decay, then the time series may need a moving average term of order k. The appropriate order of autoregression and moving average terms can be determined by iteratively adding and removing terms and analyzing the ACF and PACF plots of the residuals.

1248. What is the difference between an ARIMA model and an ARMA model?

ARIMA and ARMA models are both used for time series forecasting, but the main difference between them is that ARIMA models include an additional "I" (integrated) component to deal with non-stationarity, while ARMA models do not. Therefore, ARMA models are suitable only for stationary time series, while ARIMA models can handle both stationary and non-stationary time series.

1249. What is the difference between a stationary and non-stationary time series?

A stationary time series has constant statistical properties over time, such as constant mean, variance, and autocorrelation structure. A non-stationary time series, on the other hand, has statistical properties that change over time, such as a trend or seasonal variation.

1250. How do you test for stationarity in a time series?

There are several statistical tests that can be used to test for stationarity in a time series, such as the Augmented Dickey-Fuller (ADF) test, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, and the Phillips-Perron (PP) test. These tests check whether the time series has a unit root, which indicates non-stationarity, or not. Another way to check for stationarity is to visually inspect the time series plot and its autocorrelation and partial autocorrelation functions.

1251. What are some common techniques for transforming a non-stationary time series into a stationary one?

There are several common techniques for transforming a non-stationary time series into a stationary one, such as:

Differencing: taking the difference between consecutive observations to remove a trend or seasonal pattern.

Detrending: fitting a trend model to the time series and subtracting it from the original series.

Seasonal differencing: taking the difference between observations at the same season across different years to remove seasonal patterns.

Transformation: applying a mathematical transformation, such as a logarithmic or square root transformation, to stabilize the variance.

1252. How do you evaluate the performance of an ARIMA model?

The performance of an ARIMA model can be evaluated using several metrics, such as:

Mean absolute error (MAE): measures the average absolute difference between the predicted and actual values.

Root mean squared error (RMSE): measures the square root of the average squared difference between the predicted and actual values.

Mean absolute percentage error (MAPE): measures the average percentage difference between the predicted and actual values.

Symmetric mean absolute percentage error (SMAPE): measures the percentage difference between the predicted and actual values, taking the average of the absolute difference and the actual value.

1253. How do you forecast future values using an ARIMA model?

To forecast future values using an ARIMA model, you need to first estimate the model parameters using historical data. Once the model is estimated, you can use it to predict future values by feeding in the most recent observations and iteratively forecasting one step ahead. This process can be repeated for any desired number of time steps.

1254. What are some limitations of using an ARIMA model for time series analysis?

Some limitations of using an ARIMA model for time series analysis include:

- It assumes that the past patterns and relationships between variables will continue into the future, which may not always hold true.
- It requires a stationary time series, which may be difficult to achieve in practice.
- It may not capture all the complexities of the time series, such as non-linear relationships and sudden changes in the underlying patterns.
- It may not work well for long-term forecasting, as the model tends to lose accuracy as the forecast horizon increases.

1255. What are the conditions for a time series to be stationary?

The three conditions that are required for a time series to be labelled as stationary are:

1. The mean of the time series is constant, that is the mean of the data does not change over time.
2. The variance of the time series is constant, that is the variance of the data does not change over time.
3. There is no seasonality observed in the time series.

1256. What is white noise?

White noise in time series is data that cannot be predicted at all with data points that have no correlation between any two data points whatsoever. The conditions for a time series to be defined as data points are:

1. The mean of the time series is constant and is zero throughout the series.
2. The variance of the time series is constant, that is the variance of the data does not change over time.
3. There is no seasonality observed in the time series, that means there is no correlation between any two data points.

1257. Is white noise a type of stationary time series?

Yes, white noise is a type of stationary time series with the mean always equal to zero. All white noise time series are stationary but all stationary time series aren't white noise.

1258. What is correlation?

Correlation as the word suggests is basically the correlation between two variables, that is, it tells us about the relationship or the dependence of one variable with another. For example, if on the increase in value of one variable the other one increases as well, that will be a positive correlation. Similarly, if for the increase in value of one variable the other decreases, that will be a negative correlation.

1259. How is correlation calculated?

A11. There are many different formulas for calculating correlation, but the most used one is the “Pearson Correlation Coefficient”. It is calculated as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

Where,

r is the correlation coefficient

x is the value of the x variable in data

x bar is the mean of the values of x in data

y is the value of the y variable in data

y bar is the mean of the values of y in data

Here,

If $r = 1$, that means that there is a perfect positive correlation between x and y.

If $r = -1$, that means that there is a perfect negative correlation between x and y.

If $r = 0$, that means that there is absolutely no correlation between x and y.

1260. What does ACF stand for?

ACF stands for “Autocorrelation Function”.

1261. What is Autocorrelation Function?

Autocorrelation is the correlation between a time series with a lagged version of the same time series. Let's say we have a time series T and denote the value of the time series at time interval 'i' as T_i . So in autocorrelation we take the correlation of the data point T_i with the points T_{i-1} , T_{i-2} , T_{i-3} and so on for every 'i' present in the dataset.

The correlation we are talking about here is the normal Pearson correlation where we calculate the correlation using the formula given in the previous question for every pair $T_i - T_{i-1}$, $T_i - T_{i-2}$ and so on.

1262. What does the Autocorrelation function signify?

Autocorrelation shows the direct and indirect dependence of the observed variable on the lagged observation. For example, if we are considering the autocorrelation between two variables T_i and T_{i-2} , autocorrelation will consider the direct correlation of the variable T_{i-2} and T_i , but also the indirect correlation that T_{i-2} has on T_{i-1} , which will in turn have a correlation with T_i .

1263. What does PACF stand for?

PACF stands for “Partial Autocorrelation Function”.

1264. What is Partial Autocorrelation Function?

Partial autocorrelation function also calculates the correlation between the data points in a time series, but in a different way. As we saw in the previous questions, Autocorrelation for the variables T_i and T_{i-2} includes the direct correlation of T_i with T_{i-2} and also the indirect relation of T_{i-2} with T_i through T_{i-1} . Generally speaking, for any two variables T_i and T_{i-k} , autocorrelation will include all the intermediate correlations of T_{i-k} through T_i .

But in the partial autocorrelation function, we only consider the direct correlation of T_i with T_{i-k} , eliminating all the other in between intermediate indirect correlations.

1265. How is Partial Autocorrelation calculated?

A17. To calculate Partial Autocorrelation, we need to use a different and more complex method. Let's say we need the partial correlation between the variables T_i and T_{i-2} . We will take a regression function as follows:

$$T_i = x_2 * T_{i-2} + x_1 * T_{i-1} + \epsilon$$

And the coefficient ‘ x_2 ’ is what will be the PACF for T_{i-2} . Similarly, for finding out the PACF for T_i and T_{i-k} , we will take a regression function of $k+1$ terms and the coefficient of T_{i-k} will be the value of our PACF function.

1266. What does the Partial Autocorrelation Function signify?

The Partial Autocorrelation Function tells us about the correlation between the two variables only, eliminating all in between correlations. It signifies only the direct dependence of the variables and nothing else.

1267. What is a persistence model in time series?

The persistence model in time series is the simplest model that can be used to perform predictions on a sequence data.

1268. How does a persistence model work?

A persistence model has a very simple workflow. It simply uses the last observation to predict the next value. That is, it uses the value of the previous time step to predict the value of the next time step. So for any time step 't', a persistence model will use the value at time step 't-1' to predict the value at time step 't+1'.

1269. Why is a persistence model used?

A persistence model is usually used to set a baseline performance that can be compared to other models that we will use for the same problem statement.

1270. What is regression?

Regression is a method or a function that is used to find a relationship between two or more variables given from a dataset. Usually, you would have a dependent variable that is to be predicted using some independent variables.

For example, predicting the price of a house given its area, number of rooms, how old it is etc.

1271. What are different types of regression?

The 5 basic types of regression are:

1. Linear regression
2. Logistic regression
3. Ridge regression
4. Lasso regression
5. Polynomial regression

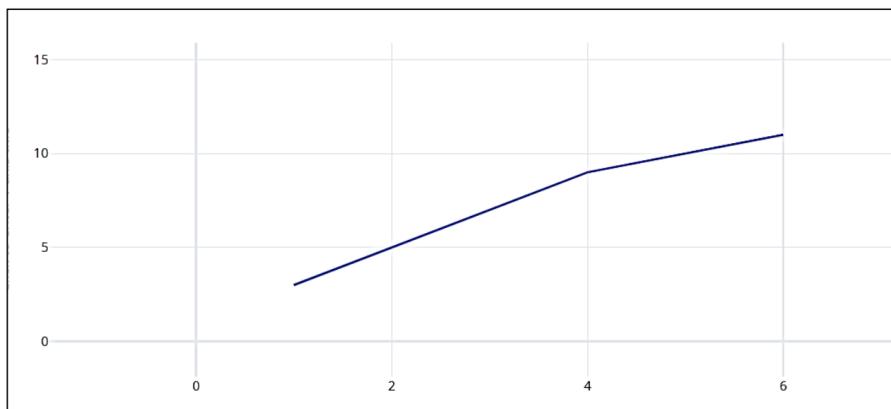
1272. How does regression work?

To understand how regression works, let us take a simple problem statement where we have to predict the price of a house given the area it covers using linear regression.

Let's denote the price of the house by 'y' and the area it covers by 'x'.
Let's say we have the following data:

Price of house (y)	Area it covers (x)
3	1
5	2
7	3
9	4
11	6

If we were to plot the given data on a graph, it would look something like this:



As you can see in the graph above, the data points form almost a straight line when connected together. Now to apply linear regression, let's take a function like this:

$$y = \alpha^*x + \beta$$

As you can see, the equation above is that of a straight line with some slope ' α ' and intercept ' β '. Our objective is to find the correct values for the parameters α and β such that the line that it will form will be closest to the straight line represented by our data.

The way we do that is by training the model to fit to our given dataset by minimizing a loss function that will minimize the error between our line and the line represented by the dataset.

1273. What is Auto-Regression?

Auto regression is a kind of regression used in time series forecasting where you use a regression model to predict the value of a future time step using the values of previous time steps as inputs.

That means in this case, our dependent variable is the current time step T_i , and our inputs or our independent variables will be the values of the previous time steps T_{i-1} , T_{i-2} , T_{i-3} and so on. This is what the equation will look like:

$$T_i = \alpha_1 * T_{i-1} + \alpha_2 * T_{i-2} + \alpha_3 * T_{i-3} + \dots$$

After training the above equation to fit on the dataset, we will obtain the optimal values for the parameters α_1 , α_2 , α_3 and so on, and we can use that model to perform time series forecasting then.

1274. How is ACF and PACF used in Auto regression in time series?

The auto correlation function or the partial auto correlation function can be used in Auto regression to choose which previous time steps should be used as inputs for the regression function. Sometimes, by using all the previous time steps as input can result in the model overfitting the data, that means that model won't generalize or make good predictions over other datasets than the training dataset.

So, we can use ACF and PACF to choose which time steps have the highest correlation with the current time step. That means we have to only consider the values that actually contribute to the value of the current time step, and eliminate the ones because the value of the current time step does not depend on those values.

1275. What will happen if you don't use ACF or PACF to filter the parameters to be used in the auto regression model?

These are the advantages of using ACF or PACF to select the input parameters for your model:

1. Only relevant time steps are used, values that don't have any contributing factor towards predictions are eliminated.
2. Model size is reduced due to decrease in number of parameters.
3. Overfitting is prevented and the model can generalize much easily on other datasets.

1276. What is a moving average?

A moving average is simply the average of a dataset observed over regular successive time intervals. Suppose for a time series with a 100 data points, you could have a moving average over 5 data points, that is first you will take the average of the first 5 points 1-5, then the average of the points 2-6, then 3-7 and so on and plot all of these averages right up to 96-10. This will be known as the moving average. You can have the intervals of any length.

1277. What is the use of using moving average on a time series?

By using moving average on a time series, you can smooth out the time series graph up to a certain extent. Using moving averages also helps to eliminate random small sized fluctuations that will help any model that you train to fit the data much better.

1278. What is an exponential moving average (EMA)?

Exponential moving average is a type of moving average where the recent datapoints are given more importance than the older datapoints. It uses the exponential function to give more weight to the recent time steps.

1279. What is a moving average model in time series forecasting?

A moving average model is a model that is kind of like a regression model, but instead of using the values of the previous time steps as it is, a moving average model instead uses the average of the dataset and the error values from the previous time steps predictions to forecast the values at the next time step.

Let us understand this better with an example. Let's say you have to use a moving average model to perform predictions for weather forecast. Let's say that the average temperature you have observed to be is 25 degrees Celsius.

$$T_i = \mu + \alpha * \epsilon_{i-1} + \epsilon_i$$
$$T'_i = \mu + \alpha * \epsilon_{i-1}$$

Where

T_i is the weather at the current time step

T'_i is the predicted value of the current time step

μ is the average value of the weather (in this case = 25)

α is the weight parameter (which is let's say = 0.5 in this case)

ϵ_{i-1} is the error value of the previous time step (= $T_i - T_{i-1}$)

So, let's say at the first time step you predict the value 25 and the actual value is 23. So, the error at this step will be:

$$\epsilon_1 = T_1 - T_0$$

$$\epsilon_1 = -2$$

$$\epsilon_1 = T_1 - T_0$$

$$\epsilon_1 = -2$$

So for the prediction at second value, by using the equation

$$T'2 = \mu + \alpha * \epsilon_1$$

We get:

$$T'2 = 25 + 0.5 * -2 = 24$$

Now let's say at the second time step, the actual value was 28. So for the third time step:

$$\epsilon_2 = T_2 - T_1$$

$$\epsilon_1 = 4$$

$$T'2 = 25 + 0.5 * 4 = 27$$

And this way, we can keep performing predictions on the time series. This is how a basic moving average model works.

Note: The above example was of a basic moving average model of order one. You can have higher order moving average models that consider the errors of more time steps as follows:

$$T'i = \mu + \alpha_1 * \epsilon_{i-1} + \alpha_2 * \epsilon_{i-2} + \alpha_3 * \epsilon_{i-3} + \dots$$

1280. What does 'ARMA' stand for?

ARMA stands for 'Auto Regressive Moving Average'.

1281. What is an Auto Regressive Moving Average (ARMA) model?

An Auto Regressive Moving Average or ARMA model is simply the combination of the Auto Regressive (AR) model and the Moving Average (MA) model.

1282. What is the notation used for an ARMA model?

ARMA models are usually written as follows:

ARMA(p,q)

Where 'p' is the order of the Auto Regression model and 'q' is the order of the Moving Average model.

1283. How does an ARMA model work?

An ARMA model can be formed simply by concatenating the equations of the Auto Regressive model and the Moving Average model. So, for an ARMA (1,1) model:

Auto Regressive model (AR):

$$T_i = \alpha_1 * T_{i-1} + K$$

Moving Average model (MA):

$$T'_i = \mu + \alpha * \epsilon_{i-1}$$

Auto Regressive Moving Average model (ARMA):

$$T_i = \alpha_1 * T_{i-1} + \beta_1 * \epsilon_{i-1} + K$$

For ARMA models of higher orders, the terms in the equation will be adjusted accordingly.

For example, for an ARMA(2,3) model, this will be what the equation will look like:

$$T_i = \alpha_1 * T_{i-1} + \alpha_2 * T_{i-2} + \beta_1 * \epsilon_{i-1} + \beta_2 * \epsilon_{i-2} + \beta_3 * \epsilon_{i-3} + K$$

1284. What does ARIMA stand for?

ARIMA stands for "Auto Regressive Integrated Moving Average".

1285. What is the notation used for ARIMA models?

ARIMA models are usually written in the following way:

$$\text{ARIMA}(p,d,q)$$

Where

'p' is the order of the Auto Regression model

'q' is the order of the Moving Average model

'd' is the order of differencing

1286. When is an ARIMA model usually used in time series forecasting?

An ARIMA model is usually used in time series forecasting when the given time series is not stationary. What that means is that the given time series does not have either a constant mean or a constant standard deviation. When a given time series isn't stationary, we can't use ARMA models as they don't perform well. In this case, we can use the ARIMA model.

1287. How does an ARIMA model work?

As the name suggests, the ARIMA model consists of both the Auto Regressive model and the Moving average model, much like the ARMA models. Let's understand what the Integrated part means.

As discussed in the previous question, we use ARIMA when the time series isn't stationary. The working of the ARIMA model is very similar to the ARMA model. In fact, the equation is exactly the same as the ARMA model. But in ARIMA, instead of using the exact value of the variable as it is, we take the differences between the values as the variable to be predicted.

Let us take the example of weather forecast. Let's say you have the following data:

Day (x)	Temperature (t)
1	25
2	28
3	24
4	27
5	26
6	31
7	28

Now for implementing an ARMA model where $d=1$, it means the difference is to be taken once. So, we will generate a new series where we take the differences between two consecutive values. Let's call this series D where

$$d_i = t_{i+1} - t_i$$

x	D
1	3
2	4
3	3
4	1
5	5
6	-3

Now we will implement an ARMA model where instead of the value 't' we will be using the variable 'd':

$$d_i = \alpha_1 * d_{i-1} + \beta_1 * \epsilon_{i-1} + K$$

And for predicting the value of 't', we will simply use the equation mentioned before to transform the value back:

$$d_i = t_{i+1} - t_i$$

Therefore,

$$t_i = d_{i-1} + t_{i-1}$$

Note, you can substitute t_{i-1} with $d_{i-2} + t_{i-2}$

Therefore,

$$t_i = d_{i-1} + d_{i-2} + t_{i-2}$$

If you keep on substituting, you will end up with the following equation:

$$t_i = \sum_{i=1}^K d_{K-i} + t_K$$

Note: In case of change in values of 'p' and 'q', the equation will change accordingly as it did in ARMA.

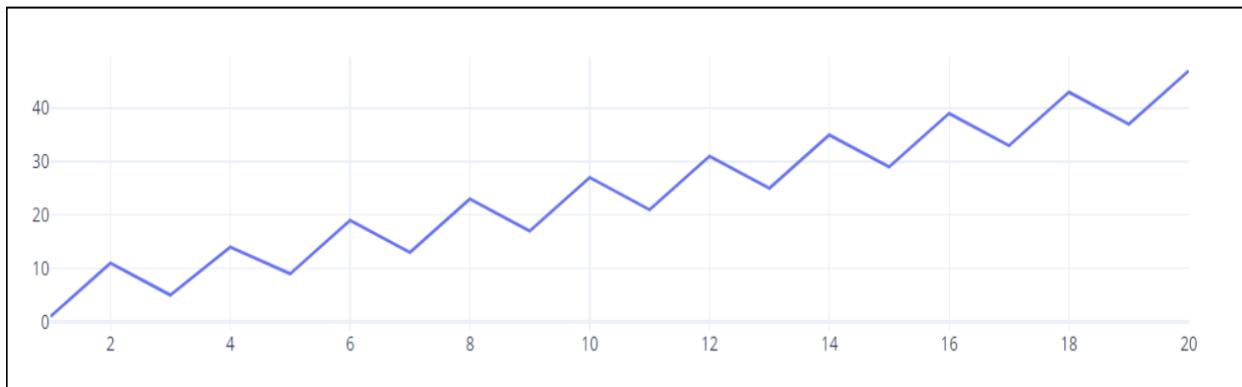
In the case of change in value in 'd', we will simply take the difference that many times. For example, if $d = 2$, we will create a new series that has the differences of the series 'D', let's say the new series is called 'E'.

If $d = 3$, we will create another series 'F' that will have the differences of successive values of the series 'E'.

And accordingly, we will keep applying reverse transforms on every level to get the value of the original variable 't'. Usually, one or at most two differences are enough to get a decent performance on the time series.

1288. What is the use of using the differences in ARIMA model?

The use of using differences in ARIMA is that it helps to make the time series stationary somewhat. Let us understand this better with an example. Let's say you have a time series which when plotted looks like the graph shown below:

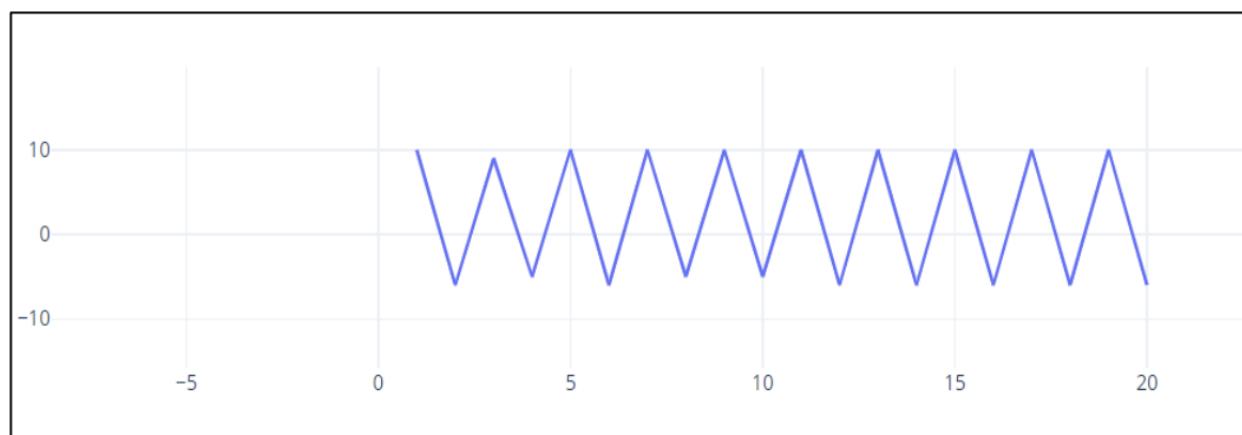


As you can see, the graph above has a pretty constant standard deviation. It does not show any seasonality too. But it definitely does not have a constant mean.

The mean seems to be gradually increasing with time.

Clearly, the graph isn't stationary. But one thing that is visible is that the graph seems to be linear in nature, that is the mean of the graph seems to be increasing linearly at a constant rate slowly. What that means basically that the mean seems to have a constant gradient.

So, when we take the differences of every two consecutive values and plot them, this is what it might look like:



As you can see, this series has a constant mean as well as a constant variance. This is a stationary series. We can now implement ARMA models for this series which will perform very well.

This is why we use differences in ARIMA model so that the model works on non-stationary series.

1289. What are some variations of the ARIMA model?

Some variations of the ARIMA model are the “SARIMA” model and the “ARIMAX” model.

1290. What does ‘SARIMA’ stand for?

SARIMA stands for “Seasonal Auto Regressive Integrated Moving Average”.

1291. When is the “SARIMA” model used in time series forecasting?

The SARIMA model is used in time series forecasting when you would usually use an ARIMA model but the time series given has some seasonality to it. In this case you would use the SARIMA model instead as it works better on seasonal time series data.

1292. What is the notation used for SARIMA?

This is how we usually denote SARIMA models: SARIMA (p,d,q) (P,D,Q)m

Where

‘p’ is the order of the Auto Regression model

‘q’ is the order of the Moving Average model

‘d’ is the order of differencing

‘P’ is the seasonal order of the Auto Regression model

‘Q’ is the seasonal order of the Moving Average model

‘D’ is the seasonal order of differencing

‘m’ is the seasonality factor

1293. What does the seasonality factor ‘m’ in the SARIMA model signify?

The seasonality factor ‘m’ in the SARIMA model is basically the number of time steps required for the seasonality to repeat. It simply means how many more time steps at a current time step for the time series to show the same cycle again.

1294. How does a SARIMA model work?

The SARIMA model works exactly like the ARIMA model, just along with the normal orders, the model also considers the lagged observations to factor in seasonality.

For example, if you have a time series where the seasonal effect occurs every year, we will set $m = 12$ for 12 months so that the SARIMA model at every time step will use the observations from 12 months before to perform forecasting.

1295. What does ARIMAX stand for?

ARIMAX stands for “Auto Regressive Integrated Moving Average with Explanatory Variable”.

1296. What is ARIMAX model ?

ARIMAX (Autoregressive Integrated Moving Average with eXogenous variables) is a type of time series model that combines the ARIMA model with additional exogenous variables that are believed to influence the time series.

In an ARIMAX model, the time series is modeled as a linear combination of its own past values, past errors, and the values of exogenous variables. The ARIMA component models the time series itself, while the exogenous variables are included as additional predictors in the model.

The order of the ARIMA component (i.e., the autoregressive, integrated, and moving average terms) is chosen based on the characteristics of the time series itself, while the selection of the exogenous variables is based on domain knowledge or statistical techniques such as feature selection or regularization.

ARIMAX models can be useful for forecasting time series that are influenced by known external factors, such as weather patterns, economic indicators, or marketing campaigns. They can also be used to estimate the impact of these factors on the time series, which can be useful for decision making and planning.

1297. What are two types of time series classified according to the number of variables?

The two types of time series classified according to the number of variables are:

1. Univariate time series: In a univariate time series, there is only one variable that is observed that changes with time.
2. Multivariate time series: In a multivariate time series, there are more than one variables observed that are changing with time.

1298. When is an ARIMAX model used in time series forecasting?

The ARIMAX model is used in time series forecasting when the time series given is not univariate but multivariate. Which means that it has more than one variables that might factor in the on deciding the value of the dependent variable.

1299. How does an ARIMAX model work?

An ARIMAX model works the same way as an ARIMA model, just with the addition of another variable in the equation called the Exogenous variable:

$$d_i = \alpha_1 * d_{i-1} + \beta_1 * e_{i-1} + K + \gamma * X$$

Where 'X' is the new Exogenous variable.

X could be anything like a technical indicator or a statistic value that could in any way influence the value to be predicted.

Python code

1300. Libraries for ARIMAX model in python?

```
import pandas as pd  
import statsmodels.api as sm
```

1301. Load and create an ARIMAX model

```
data = pd.read_csv('data.csv', index_col='date', parse_dates=True)  
model = sm.tsa.ARIMA(data, order=(p, d, q))
```

where p, d, and q are the autoregressive, integrated, and moving average orders, respectively.

1302. Fit the model and generate prediction

```
results = model.fit()  
forecast = results.predict(start=start_date, end=end_date, exog=exog_data)
```

where start_date and end_date are the start and end dates of the forecast, and exog_data is a DataFrame of exogenous variables (if any).

Evaluate the performance of the model using appropriate metrics such as MAE, RMSE, or MAPE.

Forecasting Model Evaluation

1303. What is MAE?

MAE (Mean Absolute Error) is a metric that measures the average absolute difference between the predicted and actual values. It is calculated by taking the absolute difference between the predicted and actual values, then averaging those differences over the entire dataset.

1304. What is RMSE?

RMSE (Root Mean Squared Error) is a metric that measures the square root of the average squared difference between the predicted and actual values. It is calculated by taking the square of the difference between the predicted and actual values, averaging those squared differences over the entire dataset, and then taking the square root of that average.

1305. What is MAPE?

MAPE (Mean Absolute Percentage Error) is a metric that measures the average percentage difference between the predicted and actual values. It is calculated by taking the absolute difference between the predicted and actual values, dividing that difference by the actual value, then averaging those percentage differences over the entire dataset and multiplying by 100.

1306. What is SMAPE?

SMAPE (Symmetric Mean Absolute Percentage Error) is a metric that measures the percentage difference between the predicted and actual values, taking the average of the absolute difference and the actual value. It is calculated by taking the absolute difference between the predicted and actual values, adding them together, then dividing by the sum of the predicted and actual values. The result is multiplied by 100 to express the error as a percentage. The use of the absolute difference and the actual value in the denominator makes this metric symmetric around zero, which can be useful in some applications.

1307. Difference between ARIMA, ARIMAX and SARIMA model?

The main difference between ARIMA, SARIMA, and ARIMAX models lies in the way they handle seasonal patterns and external variables.

ARIMA: This model is used to model and forecast non-seasonal time series data. It uses autoregressive, integrated, and moving average terms to model the past behavior of the time series and generate future forecasts.

SARIMA: This model extends the ARIMA model to include seasonality in the time series data. It adds seasonal autoregressive, seasonal integrated, and seasonal moving average terms to the ARIMA model to capture the seasonality in the data.

ARIMAX: This model is an extension of the ARIMA model that includes external variables or predictors in addition to the time series data. These external variables can be used to capture the impact of other factors on the time series, such as the effect of weather patterns on sales data.

In summary, the differences between these models can be summarized as follows:

ARIMA models are used for non-seasonal time series data.

SARIMA models are used for time series data with seasonal patterns.

ARIMAX models are used for time series data with external variables that are believed to impact the time series.

It's important to note that the selection of the appropriate model depends on the characteristics of the data being analyzed and the specific requirements of the analysis. It may be necessary to try multiple models and compare their performance to determine the best model for a given application.

1308. Application of ARIMA model?

ARIMA models can be applied in a variety of fields and industries where time series data is present. Some common applications of ARIMA models include:

Finance: ARIMA models can be used to model and forecast stock prices, exchange rates, and other financial time series.

Economics: ARIMA models can be used to analyze and forecast economic indicators such as GDP, inflation, and unemployment rates.

Sales and Marketing: ARIMA models can be used to forecast sales and demand for products, as well as to analyze the effectiveness of marketing campaigns.

Energy: ARIMA models can be used to forecast energy demand and consumption, as well as to analyze trends in energy production and consumption.

Engineering: ARIMA models can be used to analyze and forecast system behavior in engineering applications such as manufacturing, transportation, and logistics.

Healthcare: ARIMA models can be used to forecast patient demand and hospital occupancy rates, as well as to analyze trends in disease incidence and mortality rates.

Social sciences: ARIMA models can be used to analyze and forecast social phenomena such as crime rates, voting patterns, and migration flows.

Overall, ARIMA models can be a valuable tool for understanding and forecasting time series data in a wide range of applications.

Advance Forecasting questions

1309. What is the difference between qualitative and quantitative forecasting methods, and what are some common techniques used in each?

Qualitative forecasting methods involve expert opinions, judgment, and intuition, while quantitative forecasting methods use mathematical models and statistical analysis. Common qualitative techniques include market research, panel consensus, and Delphi method, while common quantitative techniques include time series methods (such as ARIMA and exponential smoothing) and causal methods (such as regression and econometric models).

1310. How do you evaluate the accuracy and performance of a forecasting model, and what are some common metrics used to do so?

Forecast accuracy and performance can be evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), and others. These metrics measure the difference between actual and predicted values and can help identify the best forecasting model for a given data set.

1311. What are some common challenges and limitations associated with time series forecasting, and how do you address them?

Common challenges and limitations associated with time series forecasting include missing data, outliers, seasonality, trend, and non-stationarity. These issues can affect the accuracy and reliability of forecasting models, and they can be addressed by using appropriate techniques, such as imputation for missing data, outlier detection and removal, seasonal adjustment, differencing for trend, and stationarity testing.

1312. How do you handle missing data or outliers in a time series data set, and how do they affect your forecasting results?

Missing data and outliers can be handled using imputation and outlier detection techniques. The choice of technique depends on the nature and severity of the problem and the specific characteristics of the data set. Missing data can be imputed using interpolation, regression, or time series methods, while outliers can be identified and removed using statistical tests or visual inspection.

1313. How do you handle seasonality in time series data, and what are some common techniques used to remove or adjust for it?

Seasonality in time series data can be handled using seasonal adjustment techniques, such as seasonal decomposition or seasonal ARIMA models. These methods can help remove the periodic component of the data and improve the accuracy of forecasting models.

1314. What is the difference between stationary and non-stationary time series data, and how do you handle each when building a forecasting model?

Stationary time series data has a constant mean and variance over time, while non-stationary data has a changing mean, variance, or both. Stationary data is easier to model and forecast, and it can be achieved by using appropriate techniques, such as differencing or detrending. Non-stationary data can be handled by transforming it into a stationary form or by using appropriate models that can handle non-stationarity.

1315. How do you select the appropriate forecasting model for a given data set, and what factors should be considered in making this decision?

The appropriate forecasting model for a given data set depends on several factors, such as the nature of the data, the level of complexity desired, the availability of historical data, and the specific problem being addressed. Some common techniques used in model selection include time series analysis, exploratory data analysis, cross-validation, and goodness-of-fit tests.

1316. What is the difference between univariate and multivariate time series analysis, and when would you choose to use one over the other?

Univariate time series analysis involves modeling and forecasting a single time series variable, while multivariate analysis involves modeling and forecasting multiple time series variables simultaneously. Univariate analysis is suitable for simple problems with one dependent variable, while multivariate analysis is suitable for more complex problems with multiple dependent variables that are related.

1317. How do you handle trend and drift in time series data, and what are some common techniques used to remove or adjust for them?

Trend and drift in time series data can be handled by using differencing or detrending techniques. Differencing involves taking the difference between consecutive observations, while detrending involves fitting a regression line to the data and subtracting it from the original time series. These techniques can help remove the non-stationary component of the data and improve the accuracy of forecasting models.

1318. What are some common software tools used for time series forecasting, and how do you choose the appropriate tool for a given application?

Common software tools used for time series forecasting include R, Python, SAS, MATLAB, and Excel. The appropriate tool for a given application depends on the level of complexity desired, the availability of historical data, the specific problem being addressed, and the user's familiarity with the software. Factors such as the cost, support, and ease of use should also be considered when choosing a software tool.

Chapter 13 - Data Preprocessing and Statistics

1319. What is Data??

Data is a representation of facts stored in digital form. Data may be the clean or not. This representation of the facts may or may not be valid and accurate.

1320. What is the difference between data and information?

The terms information and data are often used interchangeably. There is a difference between these two. Data can be any sequence of values, numbers, text, picture, files and so on. All of these things do not necessarily have to be informative to a consumer of that data. In most cases, data needs to be processed and put into context to make it informative for the consumer.

1321. Why is data important?

Data is information stored in digital form. Information has always been important. When information was stored exclusively in analogue form, the information storage capacity of human mankind was extremely limited. By storing information as data in digital form, we have decoupled the growth of information from these limitations. Thanks to computers, hard drives, our smartphones and other technological innovations we are able to create, store and process data at staggering levels. Today, data growth is following an exponential path. Data is the oil of the 21st century. Most importantly, data is often regarded as the fuel of the 21st century. Much like oil and electricity have powered innovations and economies in the past, data will be the (not so natural) resource that fuels these in the present and future.

1322. What is Data Pre-processing?

Data pre-processing is a process of preparing and transforming the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

1323. Why data pre-processing is so important?

Mistakes, redundancies, missing values, and inconsistencies all effect the quality of the dataset, we need to fix all those issues for a more accurate outcome. That is why data pre processing is very important. Imagine you are training a Machine Learning algorithm to deal with the share market with a faulty dataset. Chances are that the system will develop biases and deviations that will produce a poor user experience. Thus, before using the data for the purpose you want, you need it to be clean as possible.

1324. What is raw data?

Raw data refers to any data that hasn't gone through any processing, either manually or through automated computer software. Raw data may be gathered from various processes, manual data entry or IT resources. Raw data is also known as source data or primary data.

1325. What are the types of raw data?

Ans: -

1. Missing data: Missing data often appears when there's a problem in the collection phase, such as a glitch that caused a system's downtime, mistakes in data entry, or issues with biometrics use, among others. This is common in pretty much in any data available
2. Noisy data: This group encompasses outliers that you can find in the data set but that is just meaningless information. Here you can see noise made of human mistakes, rare exceptions, mislabels, and other issues during data gathering.
3. Inconsistent data: Duplicates in different formats, mistakes in codes of names, or the absence of data constraints often lead to inconsistent data, that introduces deviations that you have to deal with before analysis.

1326. What is types of data?

Data can be categorized into three types: -

1. Structured Data
2. Unstructured Data
3. Semi- Structured data

Structured Data: - Data with a high degree of organization, typically stored in a spreadsheet-like manner. Think of a spreadsheet or data in a tabular format. Data is structured in a spreadsheet-like manner Within that table, entries have the same format and a predefined length and follow the same order. It is easily machine-readable and can therefore be analysed without major pre processing of the data. It is commonly said that around 20% of the world's data is structured.

E.g.

- Excel spreadsheets
- Comma-separated value file (.csv)
- Relational database tables

Semi-structured Data: - Data with some degree of organization. Think of a TXT file with text that has some structure (headers, paragraphs, etc.)

E.g.: -

- Hypertext Markup Language (HTML) files
- JavaScript Object Notation (JSON) files
- Extensible Markup Language (XML) files

Data is stored in files that have some degree of organization and structure. Tags or other markers separate elements and enforce hierarchies, but the size of elements can vary and their order is not important. Needs some pre-processing before it can be analysed by a computer. Has gained importance with the emergence of the World Wide Web

Unstructured Data: - Data with no predefined organizational form and no specific format. Essentially anything that is not structured or semi-structured data (which is a lot). Data that can take any form and thus be stored as any kind of file (formless). Within that file, there is no structure of content. Typically needs major pre-processing before it can be analysed by a computer, but often easily consumable for humans (e.g., pictures, videos, plain texts. Most of the data that is created today is unstructured.

E.g.

- Images such as .jpeg or .png files
- Videos such as .mp4 or m4a files
- Sound files such as .mp3 or .wav files
- Plain text files
- Word files
- PDF files

1327. What are the 5 Major Steps in Data Pre-processing?

Following are the 5 major steps in Data Pre-processing

1. Import the libraries.

Importing the libraries which you'll need to work with mainly pandas ,NumPy ,seaborn
2. Import the data-set

Importing the dataset which you want to do your work

3. Check out the missing values.

Analysis and imputing the missing values

4. Transforming the Categorical Values.

Converting the Categorical values into numerical ones

5. Splitting the data-set into Training and Test Set.

Cutting the dataset into train and test so we'll train the dataset on train and test its results

on test.

6. Feature Scaling.

Scaling the features so every feature has the same level of importance for the model.

1328. What type of dataset does the machine learning algorithms work?

A machine learning model completely works on data. Each dataset is different from another dataset. To use the dataset in our code, we usually put it into a CSV file. However, sometimes, we may also need to use an HTML, Json or xlsx file.

1329. What are Libraries and why do we need them?

In order to perform data pre-processing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs.

1330. What are the popular libraries in Python we use for data pre-processing?

Pandas: The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library.

import pandas as pd

Matplotlib: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot.

Import matplotlib.pyplot as plt

Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python.

Import numpy as np

Here we use alias words for efficient coding.

1331. How to set the working directory in Spyder?

To set a working directory in Spyder IDE:

- Save your Python file in the directory which contains dataset.
- Click on to the File explorer option in Spyder IDE, and select the required directory.
- Execute the file.

1332. How to import the dataset which we want to work on?

We use `read_csv()` function of pandas library, which is used to read a csv file and performs various operations on it. Using this function, we can read a csv file locally as well as through an URL.

Syntax: -

```
df= pd.read_csv('Dataset.csv')
```

Here, df is a name of the variable to store our dataset, and inside the function, we have passed the name of our dataset.

1333. How to pass the address of the dataset on the function?

Ans: - We can the pass the address of the dataset by using the `read_csv()`

function Syntax: - `df = pd.read_csv(r"Address of the dataset")`

1334. What are the additional parameters of the `read_csv()` function

? `read_csv()` has multiple paramters:

- `Index_col` : It is used to set which columns to be used as the index of the dataframe. The default value is `None`, and pandas will add a new column start from 0 to specify the index column. It can be set as a column name or column index, which will be used as the index column.
Syntax :- `df = pd.read_csv("Fllename", index_col = 0)`

- `Header` : Header parameter is used to specify you have the names of columns in the first row in the file and if you don't you will have to specify `header=None`.
Syntax: `df = pd.read_csv("Filename",header= None)`

- `Sep` : The sep parameter is used to specify by which element are the columns separated so that the pandas library treats the data that way
Syntax:- `df = pd.read_csv("filename" ,sep = ",")`

- `Skiprows` : The skiprows parameter of the `read_csv()` function is used to the rows from csv at specified indices in the list
Syntax:- `df = pd.read_csv("filename" , , skiprows=[0,2,5])`

1335. How to have a brief look at the data ?

Ans:-

- Info() function :You can use the info() function to have a brief look at the data
Syntax:- df.info()
- Head() function :The head() Function is used to look at the first 5 rows of the dataset. Syntax: df.head()
- Tail() function :The tail() function is used to look at the first 5 rows of the dataset. Syntax: df.tail()

1336. Which commands will give you a descriptive look at the data?

Ans:-

- Describe() function: describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values.
- Shape() Function: The shape attribute returns a tuple of the number of rows and the number of columns in the DataFrame.
Syntax:- df.shape

1337. What is missing data and why it is a big problem?

Missing data presents various problems.

- The absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false.
- The lost data can cause bias in the estimation of parameters.
- It can reduce the representativeness of the samples.

1338. When should be consider deleting the missing data?

If data is missing for more than 60% of the observations, it may be wise to discard it if the variable is insignificant.

1339. How can we impute the categorical missing values?

Ans:-

1. Ignore the observation and let the algorithm handle it
2. Replace by most frequent value.

1340. How can we impute the continuous missing variable?

Ans:-

1. Ignore the observation and let the algorithm handle it.
2. Replace by the mean

1341. How to find out the missing numbers present in the data?

The isnull() function is used to show where missing values are present in the data.

The output is in Boolean format.

1342. How to find out the total number of missing values in the data? Ans:-

The isnull().sum() is used to return the total number of missing values in the data.

Syntax:- df.isnull().sum()

OP:-

TV	0
radio	0
newspaper	0
sales	0
dtype:	int64

1343. What are Outliers??

Outliers are considered to be extreme values. They are defined as samples that are significantly different from the remaining data. Those are points that lie outside the overall pattern of the distribution. Statistical measures such as mean, variance, and correlation are very susceptible to outliers.

Example: - Suppose you are handling a dataset of cricketers of the Indian cricket team. In the variable of total runs, you will encounter that Sachin Tendulkar and Virat Kohli would be considered as an outlier since their total runs scored is much higher than other cricketers, this doesn't mean that the data entered in their field is wrong, it's just that they are better than other cricketers that much.

1344. How should you handle Outliers in the dataset?

If you encounter outliers in the data, you should look at the quantity of the outliers. If the quantity is high then you should consider leaving them as they are and let the algorithm handle it. If the quantity is less then consider them imputing them.

1345. How can Outliers occur in the dataset?

Outliers can occur in the dataset due to one of the following reasons: -

1. Genuine extreme high and low values in the dataset
2. Introduced due to human or mechanical error
3. Introduced by replacing missing values.

1346. How to Detect Outliers??

Outliers can be detected by the following ways: -

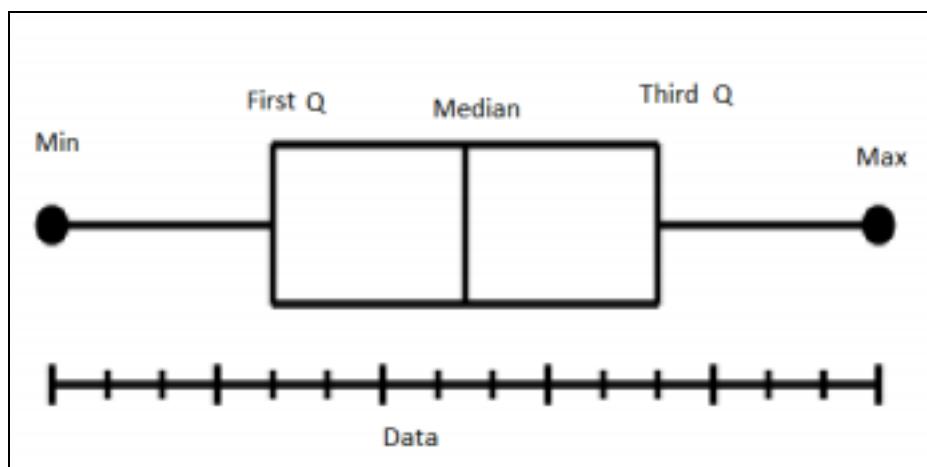
1. Extreme Value Analysis by Box Plot
2. Visualizing the data

1347. What is BoxPlot?

A box and whisker plot (box plot) summarizes the data in 5 numbers.

The five-numbers are the minimum, first quartile, median, third quartile, and maximum.

In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.



The "interquartile range", abbreviated "IQR", is just the width of the box in the box-and-whisker plot. That is, $IQR = Q3 - Q1$. The IQR can be used as a measure of how

spread-out the values are.

Outliers are the values which fall outside these margins.

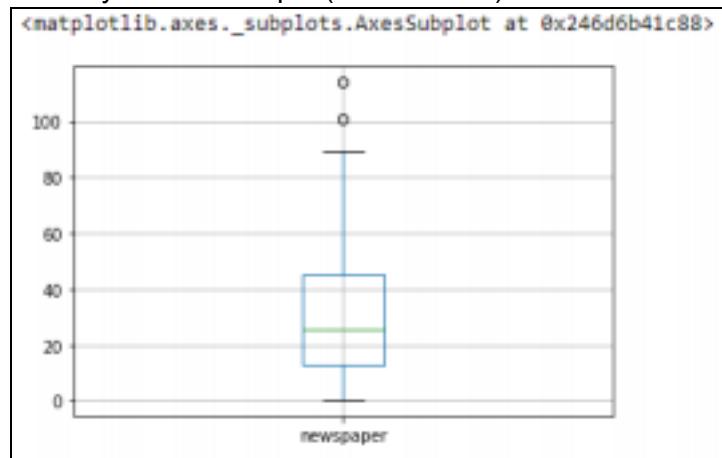
1348. How to Treat Outliers?

Ans:-

- 1.Mean/Median or random Imputation
- 2.Trimming
- 3.Discretization

1349. What is the syntax of the boxplot for visualizing the outliers?

Ans:- Syntax:- df.boxplot(column='TV')

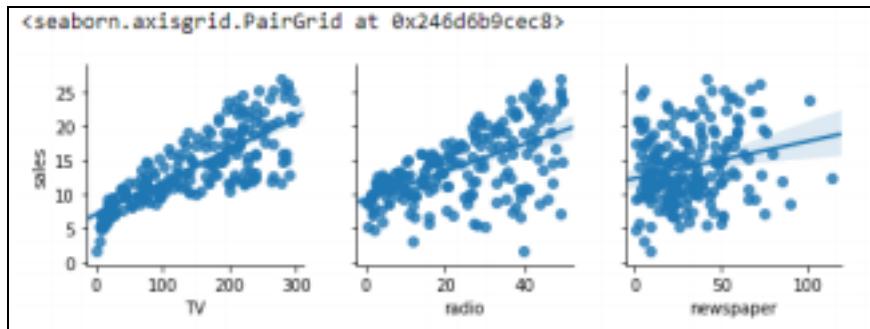


1349. How to generate a pairplot ?

Ans:- Syntax:-

```
sns.pairplot(df,x_vars = ['TV','radio','newspaper'],y_vars = 'sales',kind = 'reg')
```

OP:-



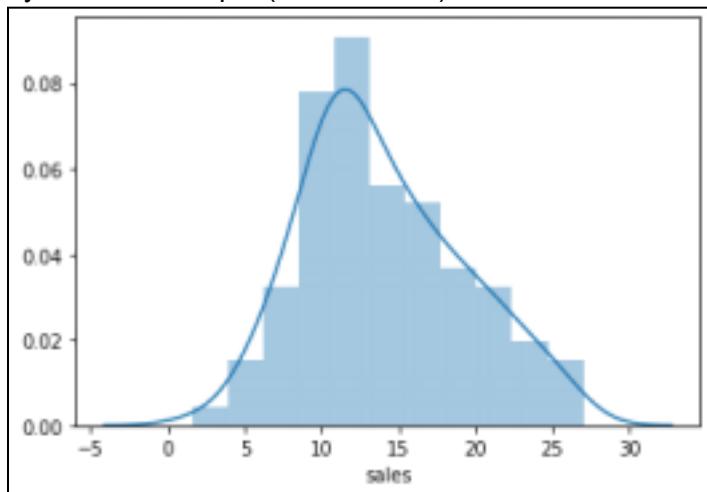
1350. What does the 'kind' parameter stand for?

The kind parameter is used to show the regression line if needed.

1351. What is Histogram and how to generate the histogram of the variables?

Ans:- A histogram is an approximate representation of the distribution of numerical data.

Syntax:- `sns.distplot(Y,hist = True)`



1352. What is Normal Distribution?

Normal distribution also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In normal distribution $\text{mean} = \text{mode} = \text{median}$. In graph form, normal distribution will appear as a bell curve.

1353. Why do we need Log Transformation?

Logarithmic transformation is a convenient means of transforming a highly skewed variable into a more normalized dataset. When modelling variables with non-linear relationships, the chances of producing errors may also be skewed negatively.

1354. What is Skewness?

Skewness is a measure of the asymmetry of a distribution. This value can be positive or negative.

- A negative skew indicates that the tail is on the left side of the distribution, which extends towards more negative values.
- A positive skew indicates that the tail is on the right side of the distribution, which extends towards more positive values.
- A value of zero indicates that there is no skewness in the distribution at all, meaning the distribution is perfectly symmetrical.

1355. What is Kurtosis?

Kurtosis is a measure of whether or not a distribution is heavy-tailed or light-tailed relative to a normal distribution.

- The kurtosis of a normal distribution is 3.
- If a given distribution has a kurtosis less than 3, it is said to be platykurtic, which means it tends to produce fewer and less extreme outliers than the normal distribution.
- If a given distribution has a kurtosis greater than 3, it is said to be leptokurtic, which means it tends to produce more outliers than the normal distribution.

1356. How can we convert the variables data type ?

In machine learning there might be situations when you need to convert the data type of the variable to your liking and hence there are function for that.

Example :-

```
pincode = 400070
```

```
mystring = str(pincode) # '400070'
```

Now the value 400070 is not integer but string

In this scenario the pin code variable is not a continuous variable ,it has a categorical essence to it and hence it should be treated as such ..

1357. What is corelation?

Correlation is usually defined as a measure of the linear relationship between two quantitative variables.

1358. How are covariance and correlation different from one another?

Ans: - Covariance measures how two variables are related to each other and how one would vary with respect to changes in the other variable. If the value is positive, it means there is a direct relationship between the variables and one would increase or decrease with an increase or decrease in the base variable respectively, given that all other conditions remain constant.

Correlation quantifies the relationship between two random variables and has only three specific values, i.e., 1, 0, and -1.

1 denotes a positive relationship, -1 denotes a negative relationship, and 0 denotes that the two variables are independent of each other.

1359. What is Multicollinearity and how we find it in Python?

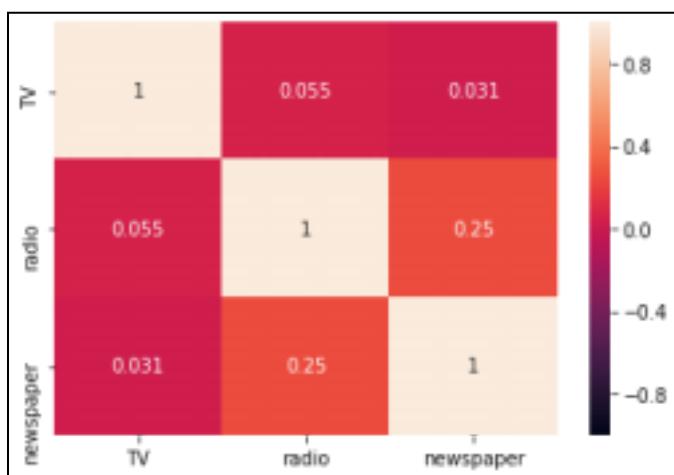
Ans: - Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

Syntax: -

```
a = sns.heatmap(corr_df,vmax = 1.0, vmin = -1.0, annot = True)
```

```
b, t = a.get_ylim()
```

```
a.set_ylim(b+0.5, t-0.5)
```



If any variables show high score, then that feature should be eliminated.

1360. What is VIF?

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. In general, a VIF above 5 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above and it depends on the situation.

1361. What is a confusion matrix and why do you need it?

Confusion matrix is a table that is frequently used to illustrate the performance of a classification model i.e., classifier on a set of test data for which the true values are well-known. It allows us to visualize the performance of an algorithm/model. It allows us to easily identify the confusion between different classes. It is used as a performance measure of a model/algorithm. It is summary of predictions on a classification model.

1362. How do we check the normality of a data set or a feature?

There is a list of Normality checks, they are as follow:

- Shapiro-Wilk W Test
- Anderson-Darling Test
- Martinez-Iglewicz Test
- Kolmogorov-Smirnov Test
- D'Agostino Skewness Test

1363. What is the idea behind Splitting the data?

In Machine Learning, we split the data into 2 parts, training and testing parts.

We train the model on training data and compare its results with the test data.

1364. What is the threshold for splitting the data?

Usually we follow the threshold of 70:30 of the data i.e., 70 % of the data to the training and 30% of the data. It depends on the situation whether you need more data for your model if its not giving you the accuracy.

Syntax: -

```
from sklearn.model_selection import train_test_split  
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=101)
```

1365. What is Scaling?

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Machine learning is like making a mixed fruit juice. If we want to get the best-mixed juice, we need to mix all fruit not by their size but based on their right proportion. We just need to remember apple and strawberry are not the same unless we make them similar in some context to compare their attribute. Similarly, in many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant.

The two major techniques for Feature Scaling are:

- Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].
- Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

1366. What are the different types of Scalers available?

Ans:-

- 1) Min Max Scaler
- 2) Standard Scaler
- 3) Max Abs Scaler
- 4) Robust Scaler
- 5) Quantile Transformer Scaler
- 6) Power Transformer Scaler
- 7) Unit Vector Scaler

1367. Explain in Min Max Scaler

An alternative approach to Z-score normalization (or standardization) is the so-called Min-Max scaling (often also simply called "normalization" - a common cause for ambiguities).

In this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

A Min-Max scaling is typically done via the following equation:

$$X_{sc} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

1368. Explain Max abs Scaler

Scale each feature by its maximum absolute value.

This estimator scales and translates each feature individually such that the maximal absolute value of each feature in the training set will be 1.0. It does not shift/centre the data, and thus does not destroy any sparsity.

Attributes:

- `scale_`
Per feature relative scaling of the data.
New in version 0.17: `scale_` attribute
- `max_abs_`
Per feature maximum absolute value.
- `n_samples_seen_`
The number of samples processed by the estimator. Will be reset on new calls to fit, but increments across `partial_fit` calls.

Syntax: - class `node_preprocessing.MaxAbsScaler`

1369. Explain Robust Scaler

Ans: - Robust Scaler algorithms scale features that are robust to outliers. It uses the interquartile range. The median and scales of the data are removed by this scaling algorithm according to the quantile range.

It, thus, follows the following formula:

$$\frac{X(i) - Q1(x)}{Q3(x) - Q1(x)}$$

Where Q1 is the 1st quartile, and Q3 is the third quartile.

EG:-

```
data = [[0,5],[2,13],[-3,7],[1,-4],[6,0]]
from sklearn.preprocessing import RobustScaler
rs = RobustScaler().fit(data)
print(rs.transform(data))
```

OP:-

```
[-0.5 0. ]  
[ 0.5 1.14285714]  
[-2. 0.28571429]  
[ 0. -1.28571429]  
[ 2.5 -0.71428571]]
```

1370. Explain Standard Scaler

Ans:- Standard Scaler assumes a normal distribution for data within each feature. The scaling makes the distribution centred around 0, with a standard deviation of 1 and the mean removed.

Formula:-

$$\frac{x(i) - \text{mean}(x)}{\text{sd}(x)}$$

Where sd is the standard deviation of x.

Syntax:-

```
from sklearn.preprocessing import StandardScaler  
ss = StandardScaler().fit(data)  
print(ss.transform(data))
```

1371. Why do we need to convert the categorical variables into numerical ones?

The machines we develop only understand categorical data, they only understand numerical data that is why we need to convert every categorical variable into numerical one.

1372. What to do when you have a variable with no missing values but has no variance?

In such situations where the variable has passed the other parameters into conducted into the model but has no variance. Then you should consider removing that variable because that variable is not contributing anything useful to the model.

1373. If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. Bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use polling technique to combine all the predicted outcomes of the model.

1374. What do you mean by feature engineering?

Features are the core characteristics of any prediction that impact the results. Feature engineering is the process of creating a new feature, transforming a feature, and encoding a feature. Sometimes we also use the domain knowledge to generate new features.

For e.g. Using the selling price variable and the cost price variable to calculate the profit. It prepares the data that easily input to the model and improves model performance.

1375. What do you mean by feature splitting?

A feature splitting is a technique to generate a few other features from the existing one to improve the model performance. for example, splitting names into first and last names.

1376. How do you select the important features in your data?

We can select the important features using random forest, or remove redundant features using recursive feature elimination. Let's all the categories of such methods.

1. Filter Methods: Pearson Correlation, Chi-Square, Anova, Information gain, and LDA.
2. Wrapper Methods: Recursive feature elimination.
3. Embedded Methods: Ridge and Lasso Regression

1377. What are the different ways by Which you can convert the categorical variables into numerical ones.?

Ans: -

1. Label Encoder
2. Manually Mapping
3. Dummy Variables
4. One hot label Encoding

1378. Explain Label Encoder

One hot encoding is used to encode the categorical column. It replaces a categorical column with its labels and fills values either 0 or 1. For example, you can see the “color” column, there are 3 categories such as red, yellow, and green. 3 categories labeled with binary values.

Syntax:-

```
from sklearn import preprocessing
le=preprocessing.LabelEncoder()
for x in colname:
    adult_df_rev[x]=le.fit_transform(adult_df_rev[x])
```

Before transforming

	age	workclass	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss
0	39	State-gov	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0
1	50	Self-emp-not-inc	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0
2	38	Private	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0
3	53	Private	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0
4	28	Private	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0

After transforming

	age	workclass	education_num	marital_status	occupation	relationship	race	sex	capital_gain
0	39	6	13	4	0	1	4	1	2174
1	50	5	13	2	3	0	4	1	0
2	38	3	9	0	5	1	4	1	0
3	53	3	7	2	5	0	2	1	0
4	28	3	13	2	9	5	2	0	0

1379. Explain Manual Mapping

Ans:- Manual mapping is a technique where we individually take one by one element and assign them a value. This is done where you need to convert specific values and the number of these values in less.

Example : -

```
df["Clusters"] = df.Clusters.map({0:"Careless",1:"Standard",2:"Target",3:"Sensible",4:"Careful"})
```

CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Clusters
0	1	Male	19	15	39 Sensible
1	2	Male	21	15	81 Careless
2	3	Female	20	16	6 Sensible
3	4	Female	23	16	77 Careless
4	5	Female	31	17	40 Sensible

1380. What are Dummy Variables?

Ans:- A Dummy variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. Its requires less computational power compared to other techniques. However the coding length is more compared to other techniques.

Syntax:-

```
import pandas as pd
raw_data = {'first_name': ['Saurabh', 'Amit', 'Mansi', 'Pranjali',
'Ankita'],
'last_name': ['Parab', 'Parab', 'Rane', 'Gawde', 'Lokande'],
'sex': ['male', 'male', 'female', 'female', 'female']}
df = pd.DataFrame(raw_data, columns = ['first_name', 'last_name', 'sex'])
```

	first_name	last_name	sex
0	Saurabh	Parab	male
1	Amit	Parab	male
2	Mansi	Rane	female
3	Pranjali	Gawde	female
4	Ankita	Lokande	female

```
pd.get_dummies(df, columns=['sex'])
```

	first_name	last_name	sex_female	sex_male
0	Saurabh	Parab	0	1
1	Amit	Parab	0	1
2	Mansi	Rane	1	0
3	Pranjali	Gawde	1	0
4	Ankita	Lokande	1	0

So after creating these variables the parent variable is of no use and hence must be eliminated.

1381. Explain One Hot Label Encoding

Ans:- It is a process that converts categorical data to integers or a vector of ones and zeros. The length of vector is determined by number of expected classes or categories. Each element in the vector represents a class. Therefore, a one is used to indicate which class it is and everything else will be zero.

Code:-

```
from sklearn.preprocessing import OneHotEncoder
type_one_hot = OneHotEncoder(sparse=False).fit_transform(
train_new.array.to_numpy().reshape(-1,1))
```

If we have categorical data that we think may be important, we want to be able to use this in the model. This is because regression algorithms and classification algorithms won't be able to process it. This is when one-hot encoding is useful.

1382. What are the Different Types of Feature Selection Techniques?

Ans:- Its is not possible that all the variables will be useful to the model and hence in machine learning you have to apply some feature selection techniques to make your model the best. Using all the features to the model reduces the overall accuracy of a classifier. The goal of feature selection in machine learning is to find the best set of features to build useful models.

The techniques for feature selection in machine learning can be broadly classified into the following categories:

- Filter methods
- Wrapper methods
- Embedded methods
- Hybrid methods

1383. What is a filter method and what are the types of filter method?

Filter Methods: - Filter methods use the properties of the features measured via univariate statistics. These methods are faster and less computationally expensive than wrapper methods. When dealing with high-dimensional data, it is computationally cheaper to use filter methods.

Some of the filter method Techniques: -

1. Information Gain: - Information gain calculates the reduction in entropy from the transformation of a dataset. It can be used for feature selection by evaluating the Information gain of each variable in the context of the target variable.
2. Chi-square Test: - The Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with the best Chi-square scores.
3. Correlation coefficient: - Correlation is a measure of the linear relationship of 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variables should be correlated with the target but should be uncorrelated among themselves. If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only really needs one of them, as the second one does not add additional information.

1384. What is a wrapper method and what are the types of a wrapper method?

Wrapper Methods: - Wrappers require some method to search the space of all possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset. The feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given data.

Some of the Wrapper method techniques are: -

1. Forward Feature Selection: - This is an iterative method wherein we start with the best performing variable against the target. Next, we select another variable that gives the best performance in combination with the first selected variable. This process continues until the target is achieved.
2. Recursive Feature Elimination :- Recursive feature elimination (RFE) selects features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a `coef_` attribute or through a `feature_importances_` attribute. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the data set until the desired number of features to select is eventually reached.

Statistics

1385. What is Statistics?

Statistics is a mathematical discipline which deals with summarizing and analyzing data sets, getting a suitable visual representation of a dataset for better understanding, drawing conclusions out of it and meeting specific demands as required.

1386. What is the importance of Statistics in Machine Learning?

We use Machine Learning to computationally analyze data and make predictions. However, many Machine Learning Algorithms and important concepts of data analysis derive its basic concepts from that of Statistics. So, to deeply understand Machine Learning Algorithms and data analysis concepts, Statistics might serve as a pre-requisite for the Beginners.

1387. How can Statistics be classified?

Statistics can be majorly classified into two broad classes:

- 1) Descriptive Statistics – It mostly deals with the summarization and visualization part of statistics, or to basically analyze and organize the data. Bar plots, Pie charts, Histograms and other exploratory data analysis techniques come into play here.
- 2) Inferential Statistics – It basically deals with a sample within a population and makes inferences or conclusions on the basis of that sample for the whole population. It is used in z-tests, t-tests and chi-square tests (which we'll see in detail, further in the article). Let's see what this actually means through an Example,

Say you have a dataset of height, weight and other physical parameters of boys in a school, in total there are 100 boys and you need to create a common activity schedule for them, that should consist of their everyday activities, exercises and diet plan.

So, this is where Inferential statistics come into play, you might want to select a sample (data of 15-20 boys), analyze them, get your conclusions and results and apply them to the whole population (100 boys). This process is more feasible and saves a lot of time and memory.

1388. What are the measures of Central Tendency?

The following are the measures of Central Tendency:

- 1) Mean
- 2) Median
- 3) Mode

1389. What is mean of a sample?

Mean or Average of a sample is the ratio of the summation of the values and the total number of values, i.e.,

$$\text{MEAN OR AVERAGE} = \frac{\text{SUM OF ALL THE VALUES}}{\text{TOTAL NUMBER OF VALUES}}$$

Or,

If there are 'n' observations in the sample, then,

MEAN OR AVERAGE = (

$$\sum h(i)$$

n

i=1

n

),

where $h(i)$ is the value of i th observation.

*Mean of a population means the ratio of sum of all the values in the

population and the total number of values*

1390. What is the median of a sample?

Median of a sample is basically a value that divides the whole set of data in higher and lower values.

Few terms one might want to look at other than the measures of Central Tendency when dealing with data science and machine learning algorithms,

1. Standard Deviation – It discusses about how scattered each data value is with respect to the mean.

Mathematically,

$$\text{Standard Deviation} = \sqrt{\frac{X(i) - \text{Mean}(X)}{N-1}}$$

where, $X(i)$ is the value of i th observation and 'N' is the total number of Observations.

2. Variance – It is the square of the Standard Deviation.

$$\boxed{\text{Variance} = (\text{Standard Deviation})^2}$$

3. Covariance – It is used to measure the relation or variability between two variables.

$$\boxed{\text{Covariance between X and Y} = \frac{(X - \text{Mean}(X))(Y - \text{Mean}(Y))}{N-1}}$$

1391. What steps should be followed in order to calculate the median of a Sample?

There is a different way of calculation of median if the number of observations is even and if the number of observations is odd. Let's see how the calculation follows:

- If number of observations(n) is odd,

- 1) Arrange the data values in ascending order
- 2) The $((n+1)/2)$ th value in the arranged data set is the median of the data.

- If the number of observations(n) is even,

- 1) Arrange the data values in ascending order
- 2) Take the middle values of the data set, i.e., $(n/2)$ th and $((n/2)+1)$ th
- 3) Take the mean of both the values and there you get your median of the dataset.

1392. What is the Mode of a sample?

Mode is the most frequently occurring value of the sample.

1393. What is train-test split?

Before applying a machine learning model, it is always better to test the model and be sure that it has a considerable accuracy and can be used further, this is what we do in train-test split.

We distribute the sample dataset in two groups, that are, the training set and the testing set,

- The training set is used to observe the data and make a model out of it, or the dataset from which the model learns.
- The testing set is further used to make predictions from the input variables and comparing the predicted values with those of the prior values present in the dataset. Let us see this through an example,

Say we have a dataset with 1000 input and output values, so we split it into two parts with training set having 800 value pairs while the testing set having 200 value pairs, i.e., 80% data is training set and 20% data is testing set (the ratio is not fixed, can be altered as per requirement, but the training set is usually bigger than the testing set) However, except the training and testing set, there is a Validation Set too, which determines the error in the training set. It is created as an individual set in big datasets but in smaller ones it is a subset of the training data set.

The training and testing datasets are split on a random basis, i.e., each value pair is placed in the training or testing set on a complete random basis

1394. How can we implement train test split in our python notebooks?

One should follow the following steps after clean and processing the whole data set in order to implement train test split:

Step 1) Distribute your dataset into independent and dependent variables. There are various ways to distribute the dataset, one of the most convenient ways according to me is using the 'iloc' function, as shown below:

```
In [ ]: X=data.iloc[:, :-1]
Y=data[:, -1]
```

Step 2) Once the data is distributed in the independent and dependent variables, one can further split it into training and testing data sets using the below mentioned lines of code,

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=0 )
print('Shape of X_train is ', X_train.shape)
print('Shape of X_test is ', X_test.shape)
print('Shape of Y_train is ', Y_train.shape)
print('Shape of Y_test is ', Y_test.shape)
```

Here, the 'train_test_split' library is imported from the 'sklearn' module (sklearn.model_selection)
• Four new variables namely, X_train, X_test, Y_train and Y_test are introduced and the independent and dependent variables are split into their training and testing datasets.

This is how a dataset can be split into its training and testing sets and further model can be created.

1395. How can one analyze a Regression Model?

One can analyze the credibility of a Regression Model by looking at various loss functions.

1396. What is a loss function?

- A loss function is a function at which a model can look at to achieve more precision and accuracy.
- A higher value of the loss function means less credibility of the model.
- A loss function is always preferred to have as minimum value as possible.

1397. What are the various types of loss function in order to analyze a regression model?

Two major types of loss function to analyze a regression model are:

1. MAE – Mean Absolute Error
2. RMSE or MSE – Root Mean Squared Error, or, Mean Squared Error

1398. What is MAE?

MAE stands for Mean Absolute error, i.e., it is the mean of the absolute error for each observation.

$$\text{MAE} = \frac{1}{n} \sum_{i=0}^n (h(i) - h(p))$$

n = Total number of observations

h(i) = Original Value

h(p) = Predicted Value

1399. How can you implement MAE in your python notebook and analyze the regression model?

One can implement MAE for analyzing the model by using the following steps:

Step 1) Define a function with arguments being the testing dependent dataset and the predicted values. Return the MAE value.

*Remember to import the NumPy library in order to perform the mentioned

mathematical calculations*

```
In [ ]: def mae(y_test , y_pred):  
    return (1/len(y_test))*np.sum(np.abs(y_test - y_pred))
```

Step 2) Fit the model and insert the required arguments in the function to get the MAE value.

```
In [ ]: model = LinearRegression()  
model.fit(Xtrain , Ytrain)  
print("mae : " , mae(Ytest , Ypred))
```

1400. What is RMSE and MSE?

RMSE and MSE stands for Root Mean Square Error and Mean Square Error respectively. RMSE means the root over mean of squared differences between the original values and the predictions while MSE simply implies the mean of squared differences between the original and predicted values.

$$\text{MSE} = \frac{(h(i)-h(p))^2}{n} \text{ and,}$$

$$\text{RMSE} = \sqrt{\frac{(h(i)-h(p))^2}{n}} \text{ where,}$$

n = Total number of observations

h(i) = Original Values

h(o) = Predicted Values

1401. How can you implement MSE and RMSE in your python notebooks and analyze the regression model?

One can find out to RMSE value by following the given series of steps:

Step 1) Define a function to calculate the RMSE score. In order to get the MSE value, one can ask to return from the function (in this case we have returned the RMSE value)

Remember to import the NumPy library in order to perform the given mathematical Calculations

```
In [ ]: def rmse_score(y_test , y_pred):
    value = (1/len(y_test))*np.sum((y_test - y_pred)**2)
    return np.sqrt(value)
```

Step 2) Fit the model and pass the required arguments in order to get the RMSE score and further you can evaluate your model on the basis of it.

```
In [ ]: model = LinearRegression()
model.fit(Xtrain , Ytrain)
print("rmse_score : " , rmse_score(Ytest , Ypred))
```

1402. What is the relation between the two main loss functions, RMSE and MAE?

For same set of error RMSE usually has a greater value as compared to MAE. However, if the difference between the original and predicted data points is equal for all the observations then MAE and RMSE yield the same value. (Can be verified with the help of Basic Mathematics)

1403. How can one analyze a classification model?

One can analyze the credibility of a classification model on the basis of various parameters:

1. Accuracy (sometimes used individually to evaluate the model)
2. Confusion Matrix
 - Precision
 - Recall
 - Accuracy (same value as without implementing confusion matrix)
 - F1 score

1404. What do you mean by accuracy of a classification model?

The accuracy score or accuracy of a classification model indicates how accurately the model is able to classify the input variables. A good accuracy score is good for model building or implementation, but it is not the only thing to take into consideration. There can be times when a false or disastrous model might give a good accuracy score.

Let's see how a model calculates accuracy with the help of a small example,

Original value	Predicted Value	True/False
0	0	True
1	1	True
1	0	False
0	1	False
1	1	True

Here, we have two columns of original and predicted values and a column of true and false.

- The third column returns true if the predicted value matches the original value and false if it doesn't match.
- We notice that out of five three values turn out to correctly predicted or 'True' while the other two are incorrectly predicted or 'False'.
- The accuracy will be the ratio of number of values correctly predicted and the total number of values.
- So, here in this case, the model has an accuracy of 3/5 or 0.6.

1405. What is a confusion Matrix?

A confusion matrix is another way to analyze a classification model. The confusion matrix however analyzes the model in a bit more detail than accuracy score. There are four attributes in the confusion matrix that helps it to analyze a model. Those attributes are as follows:

1. True Positive
2. False Negative
3. True Negative
4. False Positive

1406. What is True Positive?

True positive is an attribute of the confusion matrix that basically talks about the number of observations that are actually true and have been classified as true.

1407. What is False Positive?

False Positive talks about the values which are actually true but have been predicted false by the model. This is an error in detection; hence, False Positive values are often referred to as type 1 error.

1408. What is True Negative?

True Negative is yet another attribute that talks about the values that are actually false and have also been predicted false by the Model.

1409. What is False Negative?

False negative is another attribute that talks about the values that are actually true but have been predicted false by the Model. This is an error; hence, False Negative values are often referred to as type 2 error.

The aim in Confusion Matrix should always be to reduce type 1 and type 2 errors, that are False Positive and False Negative Values respectively

1410. What is the relation between accuracy and other confusion matrix attributes?

Accuracy in terms of Confusion Matrix attributes can be written as,

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} .$$

1411. What is precision?

Precision is another term related to the Confusion Matrix. It tells us how precisely our model is making predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Through the formula, we can see that precision is actually the ratio of true values that are predicted true and the total number of true predicted.

1412. What is the desired value of precision and why?

Desired Value of Precision is '1' and it will be when the number of false positives will be zero. Precision works towards decreasing the number of False Positive values as much as possible.

1413. What is recall?

It is also a term related to the Confusion Matrix and it basically checks the number of True predicted out of the actual number of True Values present in the sample. The formula for recall goes like,

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Recall is often referred to as ‘Sensitivity’.

1414. What is the desired value of recall and why?

The value of Recall should be as close to ‘1’ as possible. Recall equals one will imply that the model has very successfully detected all the True Values and there are no False Negatives. Recall works towards lowering the False Negatives value as much as possible.

1415. What do you mean by f1 score?

f1 score is the harmonic mean of Precision and Recall.

$$\text{F1 score} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

1416. What is the need of f1 score for evaluating a model and what should be its desired value?

Even after having various model evaluating measures like Precision and Recall, the need of another measure f1 score is justified to have an optimized model which ensures less number of False Negative Values as well as less number of False Positive Values. Less False Negative values is ensured by a good recall value while less False Positive Value is ensured by a good precision value. However, both of these things are taken into account by f1 score. A higher f1 score ensures a better model.

In order to have a good f1 score, one should have a high precision and recall value

1417. Explain Confusion Matrix and its attributes with the help of an example.

Let's take a scenario in which we are provided with 1000 laptops and we have to predict the number of laptops that are faulty or whose battery is damaged. Below the Confusion Matrix shows us the result of actual and predicted condition,

1000 Samples	Predicted Faulty	Predicted not faulty
Actual Faulted	100 = TN	50 = FP
No Fault	150 = FN	700 = TP

Here,

- TN stands for True Negative,
- FP for False Positive
- FN for False Negative and,
- TP for True Positive.

We can derive from this illustration and confusion Matrix that,

- there are 100 laptops that are faulty and are predicted faulty (True Negative),
- 50 laptops are predicted not faulty but are actually faulty (False Positive)
- 150 laptops actually have no fault but are predicted faulty (False Negative)
- And finally, 700 have no fault and are predicted not faulty (True Positive)

Also, let's calculate the accuracy, precision, recall and f1 score in the same case:

- Accuracy = $(700+100)/(700+100+50+150) = 800/1000 = 0.8 = 80\%$
- Precision = $700/(700+50) = 700/750 = 0.933$
- Recall = $700/(700+150) = 700/850 = 0.824$
- F1 score = $2*((0.933*0.824)/(0.933+0.824)) = 0.875$

With 80% accuracy and the given values of precision, recall and f1 score, we can very confidently say that our model is good and can be implemented further.

1418. How can we implement the confusion matrix along with all its attributes in python notebook?

A) One can follow the given series of steps in order to implement Confusion matrix along with all its attributes,

Step 1) Import the confusion matrix along with all its required attributes. Introduce a variable and pass confusion matrix with the testing set of dependent variable and the predictions made.

Note that f1-score is not imported as a predefined library, we'll see how to calculate the f1-score further in this question

```
In [69]: from sklearn.metrics import confusion_matrix,accuracy_score, precision_score,\  
recall_score,roc_curve,roc_auc_score  
  
matrix = confusion_matrix(Ytest , predictions)  
print(matrix)|  
[[1549  55]  
 [ 302  94]]
```

Step 2) Calculate accuracy by using a function and passing the required arguments, i.e., the testing set of dependent variable and the predictions made. One can simply use the 'accuracy_score' library for the same.

One can observe how the accuracy function works, it is the mean of zeros (if predictions not correct) and ones (if predictions are correct), similar to the way we have discussed accuracy in one of the above questions

```
In [70]: def Accuracy(Truths , Predictions):  
    return np.mean(Truths == Predictions)  
  
print("Validation Accuracy : " , Accuracy(Ytest , predictions))|  
Validation Accuracy :  0.8215
```

Step 3) Calculate Precision by simply passing the arguments as shown below.

```
In [71]: print("Precision : " , precision_score(Ytest , predictions))|  
Precision :  0.6308724832214765
```

Step 4) Calculate Recall by simply passing the arguments as shown below.

```
In [72]: print("Recall : " , recall_score(Ytest , predictions))|  
Recall :  0.23737373737373738
```

The formulae of Precision and Recall are discussed in the above questions; one can also create a function and calculate precision and recall through it, similar to the way we calculated accuracy

Step 5) We can calculate the f1-score by defining a library, passing precision and recall as its arguments and calculating the harmonic mean of precision and recall, as shown below.

Remember to import the NumPy library in order to perform mathematical calculations

```
In [ ]: def f1_score(precision , recall):
         return 2 * (precision * recall)/(precision + recall)
```

1419. What do you mean by threshold and optimizing the threshold value?

Threshold value is a value we set in our classification model according to which we determine our True and False values. It is very important to have a optimized Threshold Value in order to have a significant accuracy of the model with less type 1 and type 2 errors.

1420. What is ROC Curve?

ROC stands for Receiver Operator Characteristics Curve. We must have a very optimized threshold value for our threshold to have a model with good accuracy, precision and recall score.

ROC curve helps us to get that optimized threshold value, ROC curve has,

- True Positive Rate as its Y-axis.

*True Positive Rate is the same as Recall or Sensitivity Value, i.e., the ratio of positive values that are predicted as positive (True Positive) and the sum of actual number of Positive value (True Positive + False Negative) *

- False Positive Rate as its X-axis

*False Positive Rate is the ratio of false positives and the sum of false positives and true negatives, i.e., it talks about the percentage of sample that is actually false but predicted true (False Positives) from the sample that is actually false (False Positive + True Negatives) *

1421. What steps should be followed in order to create a ROC Curve?

The following steps should be followed in order to create a ROC Curve:

1. Start with a very less threshold value.
2. Create a confusion matrix according to this threshold.
3. Calculate True Positive Rate and False Positive Rate.
4. Plot it on a X-Y plane.
5. Increase the value of the Threshold by a bit.
6. Create another confusion matrix and plot on the X-Y plane.
7. Follow these set of steps until you reach the maximum Threshold.
8. Connect all the points in the X-Y plane to obtain your ROC Curve.

To implement ROC Curve in python notebook we can follow the given series of steps:

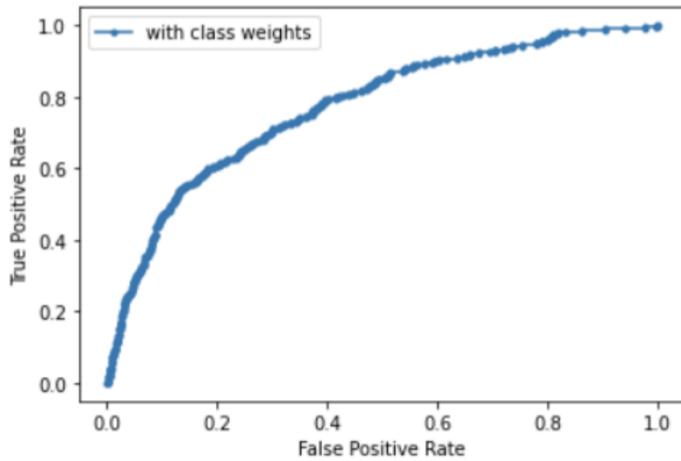
Step 1) Import roc_curve (and auc_score if required) from sklearn module (sklearn.metrics) as shown below,

```
In [69]: from sklearn.metrics import confusion_matrix,accuracy_score, precision_score,\  
recall_score,roc_curve,roc_auc_score  
  
matrix = confusion_matrix(Ytest , predictions)  
print(matrix)  
[[1549  55]  
 [ 302  94]]
```

Step 2) 'roc_curve' returns Fpr (False Positive Rate), Tpr (True Positive Rate) and threshold.
We can plot a ROC Curve as shown below.

Remember to import the matplotlib library in order to carry out the plotting functions

```
In [83]: Fpr, Tpr, _ = roc_curve(Ytest, pred_probs)  
  
plt.plot(Fpr, Tpr, marker='.', label='with class weights')  
plt.xlabel('False Positive Rate')  
plt.ylabel('True Positive Rate')  
plt.legend()  
plt.show()
```



1422. What is AUC?

- AUC stands for Area Under the Curve.
 - It basically determines the area under the ROC Curve.
 - AUC helps us to choose between models with different hyperparameters.
 - A Model with higher value of AUC (Area Under the Curve) is preferred.
 - Higher AUC implies the model has more True Positive Rate as compared to another model.
- Let us see how to implement or calculate the AUC in python notebook.

```
In [84]: Auc = roc_auc_score(Ytest, pred_probs)
print(" Area Under Curve : " , Auc)

Area Under Curve :  0.7736104813723268
```

We can simply calculate the AUC using the above library after importing it from sklearn module (sklearn.metrics). You can refer the above question in case you want to know how to import the roc_auc_score.

1423. What do you mean by a classification model?

A Classification Model is a model which classifies data into positive or negative (Binary Classification). It sets a threshold value and according to it classifies datasets into positives or negatives. However, python has very wide range of applications and there are times when data can be classified in more than two classes, i.e., no Binary Classification.

1424. What do you mean by a regression model?

A regression model is used to predict continuous values, i.e., it predicts values by observing the trend of previous input and output pairs. Regression model has applications in stock-market trading, bank insurance policy and many more professional sectors.

1425. What are parameters and hyperparameters?

- Parameters are the values that are set during the training phase. They have no impact on the efficiency of the model.
- Hyperparameters are values that are set before the training of the data. They impact the efficiency of the model and with different hyperparameters the model might lead to different results.

1426. What is Hypothesis Testing?

Hypothesis Testing is another Statistical measure that is used statistically infer about the probability of an event occurring or not. Hypothesis Testing has two forms:

1. Null Hypothesis (H_0)
2. Alternate Hypothesis (H_a)

1427. What are Null and Alternate Hypothesis?

- Null Hypothesis is the initial prediction about the occurrence of an event, the prediction made by Null Hypothesis is not based on any experimental data or historical records.
- Alternate Hypothesis on the other hand, makes prediction on the basis

on some historical records or experimental data. If Null Hypothesis doesn't hold true on the basis of evidences found, then we take Alternate Hypothesis into account.

1428. What are the steps involved in order to perform Hypothesis Testing?

Following are the steps one should follow in order to make Hypothesis Testing:

1. Make an initial Prediction or Statement (Null Hypothesis)
2. Find Evidences or records in accordance with the given condition or Hypothesis or Statement made.
3. Check if you are able to gather enough evidences to verify the hypothesis that you made (Null Hypothesis)
4. If Null Hypothesis is not verified by the evidences found then take Alternate Hypothesis into Account.

1429. Explain or illustrate hypothesis testing with the help of an example.

Let's say we have a football match going on, Now we have to make a statement regarding the striker of a team, that if he will be able to score a goal or not. So, let's follow the following steps,

1. We will assume that 'there is equal chance of scoring and not scoring a goal by the striker', i.e., the probability of the striker scoring a goal is 0.5, as either he will score a goal or he won't. (Null Hypothesis)
2. Now, we will try to collect data or records or evidences in accordance with the striker.
3. Say, we find out that the striker has scored eight goals out of his previous ten passes.
4. The evidence found nowhere justifies our Null Hypothesis so we will jump to Alternate Hypothesis and make the final statement that 'There is a greater chance or probability of scoring a goal by the striker'. Or saying that the Alternate Hypothesis holds true.

1430. What do you understand by p-value?

p-value basically puts up a label or a criterion up to which we can consider our 'Null Hypothesis'. If the original or observed experimental results or evidences are below the p-value then, we must reject our 'Null Hypothesis' and go with the 'Alternate Hypothesis'.

1431. What are statistical tests?

Statistical tests help us to determine whether the Hypothesis selected by us is true or not.

Statistical Tests are classified in two types:

1. Parametric tests
2. Non-parametric tests

1432. What are the various types of parametric tests and when do we use them?

A) There are various types of parametric tests and are used as per different requirements and features to be taken into consideration, the tests are as stated below:

Features Considered	Tests Applied
One Categorical	One sample proportion test
Two Categorical	Chi-square test
Continuous Variable ($n < 30$) or one continuous and one categorical (categorical variable having only two categories)	T-test
Continuous variable ($n > 30$) or one continuous and one categorical (categorical variable having only two categories)	Z-test
Two Continuous Variable	Correlation (further with the help of t-test)
One Continuous Variable and one or more categorical variable (with categorical variables having more than two categories)	Anova Test

1433. What is the need of all the parametric tests after applying Hypothesis Testing?

Hypothesis testing works on the principle of assumption and probability, say we assume null hypothesis, fail to collect evidences in favor of null hypothesis and then go with the alternate hypothesis or claim the alternate hypothesis to be true. Now, we go forward with alternate hypothesis because we were not able to verify alternate hypothesis, looking closely, we realize that we didn't had enough evidences at the time or the evidences don't direct towards the alternate hypothesis as well and there is a significant chance of the alternate hypothesis being false. This is where all the parametric tests come into play and try to make sure that a correct hypothesis is considered. If the tests lead to a p-value less than that of the original p-value set before the test (0.05 – standard), we reject the null hypothesis and accept the alternate hypothesis and vice versa if p value is greater than the original pre-set p-value (0.05).

1434. What is Ensemble Model?

Ensemble Model as the name suggests creates a model out of a number of models and does classification accordingly. Let's see how an ensemble model works through a series of steps,

1. m samples are created out of a dataset (say of n observations; $n > m$),
 2. A model is made out of each sample, i.e., m models are created.
 3. Each model is asked to predict or classify a particular requisite value.
 4. The value is further classified on a voting basis. (Say a majority of the m models classify the value as 'True' then, the value is indeed classified as 'True')
- So, this is a basic functioning of an Ensemble Model. However, there are various classifications and different types of ensemble models which we'll further see.

1435. What are the various types of ensemble techniques?

Ensemble Techniques are classified as:

1. Bagging (Bootstrap Aggregation)

- Random Forest

2. Boosting

- ADAboost
- Gradient Boosting
- XgBoost

1436. What do you mean by AIC and BIC?

- AIC stands for Akaike Information Criteria and BIC stands for Bayesian Information Criteria.
- They are used to select a better model; when a user is struck with a bunch of models, he/she can compare the AIC and BIC values of the models and can go with the model having the least AIC/BIC value.

1437. On what basis are AIC and BIC values calculated?

AIC and BIC values are calculated on the basis of various model parameters. The model parameters taken into account for the calculation of AIC and BIC are as follows:

1. Log likelihood (l)
2. Number of parameters (k)
3. Number of samples used for fitting (n)

1438. How does the above-mentioned model parameters affect the model?

The model parameters affect the model in the following ways:

1. Log Likelihood (l) – It basically talks about how well the model is fitting the data, a higher Log likelihood value is obviously preferred as the better the fit, better the model is. However, this is not the only model parameter, there are other parameters as well to make sure one doesn't encounter Overfitting.
2. Number of Parameters (k) – A lower value of 'k' is preferred as greater the number of parameters taken into account, greater the complexity or the memory and time consumption will be, that we surely don't want.
3. Number of samples used for fitting (n) – This model parameter is taken into account only while calculating BIC. A smaller number of samples used for fitting is preferred.

1439. How are AIC and BIC values calculated and what is the desired value for the selection of a better model?

A) AIC or Akaike Information Criteria is calculated as,

$$AIC = 2k - 2l$$

A value close to zero is preferred in case of AIC for the selection of a better model. A value close to zero will try to ensure an optimized value for 'k' and 'l', thus trying to make sure that the model is not taking into consideration a large number of parameters and also has a considerable fit to the data.

BIC or Bayesian Information Criteria is calculated as,

$$BIC = k * \ln(n) - 2l$$

For BIC as well, a value close to zero is preferred for the selection of a better model. Just like AIC it will also try to ensure a considerable fit model for the data without taking in account many parameters. It will also take into consideration the number of samples used to fit the data and will try to keep its value smaller, as preferred.

1440. What do you understand by Correlation?

Correlation between two rows having numerical values can be said as the relative relation or trend between those two rows. There are two components of correlation, namely,

1. Magnitude – It talks about the level of correlation between two numerical fields.

2. Sign (+ or -) – It talks about the way the two fields are correlated, i.e., Positive or Regular Correlation and Negative Correlation.

1441. What are the different types of Correlation?

- A) There are basically three types of correlation:
1. Pearson's Correlation
 2. Kendall Rank Correlation
 3. Spearman Rank Correlation

1442. What is Multicollinearity?

Multicollinearity refers to a condition when two columns are highly correlated, this might lead to a condition where the model created might be a bit biased due to the repetition of a particular feature. It is mostly observed in Linear Regression models and it is better to remove multicollinearity.

1443. What are the methods to remove Multicollinearity?

Multicollinearity is dealt in different ways under different scenarios,

1. If there are less observations in a dataset, so multicollinearity might affect the model a lot, because one can't afford repetition in a small dataset and the model created will surely be biased further adversely affecting the predictions to be made. In this case, the best thing to do is to remove one of the columns, surely a part of data will be lost, but the model created will be much more precise and accurate.
2. If we are dealing with a bigger dataset involving several features, then deleting one of the columns causing multicollinearity should not be the first priority as it might lead to loss of data. Here, multicollinearity can be resolved using Ridge and Lasso Regression.

1444. What is the statistical significance of degree of freedom?

In statistics, by degree of freedom we mean the number of values that are free to vary. Say we have a set of 20 integers and are supposed to calculate the mean of those integers, So, in this case degree of freedom will equal to one less than the number of independent variables, i.e., $20-1=19$. So, in total we will have 19 Degrees of Freedom for the system.

1445. What does one mean by Degree of Freedom in Machine Learning?

- A) In Machine learning, degree of freedom implies the number of parameters that are estimated from the data.

Let's take the example of linear regression in order to understand this more efficiently,

So, in a linear regression model we have to estimate the slope and the intercept in order to determine the line. So, we have 2 degrees of freedom in case of a linear regression model.

Chapter 14 - MS Excel

MS Excel is one of the most widely used software for data entry, data retrieval and data analysis. It is used by most of the large MNCs for tackling with humongous datasets or to maintain important records. Excel has its uses ranging from simple data collection to complex data handling and analyzing. With growing Modernization, Excel continues to remain an important Tool for Big Business and firms and a requisite for any job seekers in the Data Analysis industry.

1446. Explain MS Excel in brief.

Microsoft Excel is a spreadsheet or a computer application that allows the storage of data in the form of a table. Excel was developed by Microsoft and can be used on various operating systems such as Windows, macOS, IOS and Android.

Some of the important features of MS Excel are:

Availability of Graphing tools

Built-in functions such as SUM, DATE, COUNTIF, etc

Allows data analysis through tables, charts, filters, etc

The availability of Visual Basic for Application (VBA)

Flexible workbook and worksheet operations

Allows easy data validation

1447. What is a spreadsheet?

Spreadsheets are a collection of cells that help you manage the data. A single workbook may have more than one worksheet. You can see all the sheets at the bottom of the window, along with the names that you have given them.

1448. What is a cell address in Excel?

A cell address is used to identify a particular cell on a worksheet. It is denoted by a combination of the respective column letter and a row number.

1449. How to freeze columns in Excel?

Freeze panes keep the rows and columns visible while scrolling through a worksheet. To freeze panes, select the View tab and go to Freeze Panes. If you are looking to freeze the first two columns of a dataset, select the 3rd column, and click 'Freeze Panes'. A thick grey border indicates this.

1450. What is formula in Excel?

Formula - The formula is like an equation in Excel, the user types in that. It can be any type of calculation depending on the user's choice.

Manually typing out a formula every time you need to perform a calculation, consumes more time.

Ex: = A1+A2+A3

1451. What is function in Excel?

Function - a function in Excel is a predefined calculation which is in-built in Excel. Performing calculations becomes more comfortable and faster while working with functions.

Ex: = SUM(A1:A3)

1452. Mention the order of operations used in Excel while evaluating formulas.

The order of operations in Excel is referred to as PEDMAS. Shown below is the order of precedence while performing an Excel operation.

Parentheses

Exponentiation

Division/Multiplication

Addition

Subtraction

As seen above, first, the data in the parentheses is operated, followed by the exponentiation operation. After that, it can be either the division or multiplication operations. The result is then added and finally subtracted to give the final result.

1453. What is count function?

It counts the number of cells that contain numeric values only. Cells that have string values, special characters, and blank cells will not be counted.

1454. What is counta function?

It counts the number of cells that contain any form of content. Cells that have string values, special characters, and numeric values will be counted. However, a blank cell will not be counted.

1455. Can you protect workbooks in Excel?

Yes, workbooks can be protected. Excel provides three options for this:

Passwords can be set to open Workbooks

You can protect sheets from being added, deleted, hidden or unhidden

Protecting window sizes or positions from being changed

1456. Does each cell have unique address?

Yes, each cell has a unique address depends on the row and column value of the cell.

1457. How can you add cells, rows or columns in Excel?

If you want to add a cell, row or column in Excel, right click the cell you want to add to and after that select insert from the cell menu. The insert menu makes you able to add a cell, a column or a row and to shift the cells affected by the additional cell right or down.

1458. How would you format a cell? What are the options?

A cell can be formatted by using the format cells options. There are 6 format cells options:

Number

Alignment

Font

Border

Fill

Protection

1459. What is the use of comment? How to add comments to a cell?

Comments are used for a lot of reasons:

Comments are used to clarify the purpose of the cells.

Comments are used to clarify a formula used in the cell.

Comments are used to leave notes for others users about a cell.

To add a comment: Right click the cell and choose insert comment from the cell menu. Type your comment.

1460. What does the red triangle indicate at the top right hand corner of the cell?

The red triangle at the top right hand corner of a cell indicates that there is a comment linked to the particular cell. If you put your cursor on it, it will show the comment.

1461. How would you add comments to a cell?

To add a comment to a cell, you right click the cell and choose insert comment from the cell menu. Type your comment in the comment area provided. A red triangle at the top right hand corner of a cell indicates that there is a comment linked to that particular cell. To remove a comment from a cell, right click the cell and then select delete comment from the cell menu.

1462. How can you use the INDEX and MATCH functions together in a formula in Excel?

The INDEX and MATCH functions can be used together in a formula in Excel to perform a two-dimensional lookup, where you want to search for a value in one table based on a value in another table. The INDEX function returns the value of a cell within a specified range, while the MATCH function returns the position of a value within a range. By combining these two functions, you can create a formula that searches for a value in one table and returns a corresponding value from another table.

1463. What can you use instead of VLOOKUP to get same result?

Ans-Using a combination INDEX and MATCH, we can perform the same operations as VLOOKUP.

Following syntax using INDEX and MATCH together: =INDEX(range, MATCH(lookup_value, lookup_range, match_type))

1464. What is the difference between relative and absolute cell references in MS Excel?

Relative cell references are references to cells that adjust based on the location of the formula. For example, if you have a formula in cell A1 that references cell B1, and you copy that formula to cell A2, the reference to cell B1 will change to B2. Absolute cell references, on the other hand, always reference the same cell, even if the formula is copied to another location. To create an absolute cell reference, use the \$ symbol before the row and/or column reference (e.g. \$A\$1).

1465. How to check two columns are equal in excel?

Ans- To check if two columns are equal in Excel, you can use the IF function and the equality operator (=). For example, if you want to compare the values in columns A and B, you could use the following formula in cell C1:

=IF(A1=B1,"Equal","Not Equal")

And then copy the formula down to the rest of the cells in column C. This formula will compare the values in cells A1 and B1, and if they are equal, it will return "Equal" in cell C1. If the values are not equal, it will return "Not Equal".

1466. How can you use VLOOKUP in MS Excel to search for a specific value in a table?

Ans-The VLOOKUP function in MS Excel is used to search for a specific value in the first column of a table, and return a corresponding value from another column.

The syntax for the VLOOKUP function is =VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup]), where lookup_value is the value you want to search for, table_array is the range of cells that contains the table, col_index_num is the column number that contains the value you want to return, and range_lookup is an optional argument that determines whether to find an exact match (TRUE or omit) or an approximate match (FALSE).

1467. How would you use the SUMIF function in MS Excel to sum values based on a specific condition?

The SUMIF function in MS Excel allows you to sum values based on a specific condition. The syntax for the SUMIF function is =SUMIF(range, criteria, sum_range), where range is the range of cells to evaluate, criteria is the condition to be met, and sum_range is the range of cells to be summed if the criteria are met.

For example, if you want to sum the values in column A that match the word ""apple"" in column B, you could use the following formula: =SUMIF(B1:B10, ""apple"", A1:A10)

1468. Which are the different workbook protection types in Excel?

There are three ways to protect a workbook in Excel:

Password protection for opening a workbook
Protection for adding, deleting, hiding and unhiding sheets
Protection from changing size or position of windows.

1469. How is a Formula different from a Function in Excel?

Answer: Formula- Manually typing out a formula every time you need to perform a calculation, consumes more time.

Ex: = A1+A2+A3

Function-performing calculations becomes more comfortable and faster while working with functions.

Ex: = SUM(A1:A3)

1470. How will you write the formula for the following? -

Multiply the value in cell A1 by 10, add the result by 5, and divide it by 2.

We have to follow the PEDMAS Precedence. That is $((A1*10)+5)/2$.

1471. How do you create a hyperlink in Excel?

The shortcut used is Ctrl+K.

1482. How can we merge multiple cells text strings in a cell?

To merge text strings present in multiple cells into one cell, use the CONCATENATE().

1483. What is the syntax of Vlookup?

VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

1484. Which function is used to determine the day of the week for a date?

WEEKDAY () returns the day of the week for a particular date counting from Sunday.

Example: Let date at A1 be 12/30/2016

WEEKDAY(A1,1) =>6

1485. What is the shortcut to add a filter to a table?

The shortcut to add a filter to a table is Ctrl+Shift+L.

1486. Write a VBA function to calculate the area of a rectangle.

Function Area(Length As Double, Optional Width As Variant)

If IsMissing(Width) Then

Area = Length * Length

Else

Area = Length * Width

End If

End Function

1487. Calculate your age in years from the current date.

Use the YEARFRAC() or DATEDIF() function to return the number of whole days between start_date and end_date

1488. How many report formats are available in Excel and what are their names?

1. Compact
2. Report
3. Tabula

1489. How would you calculate the number of occurrences of a specific text in a range of data in Excel?

The formula to use would be `SUM(IF(range = ""specific text"", 1, 0))`.

1489. How would you find the second largest value in a range of data in Excel?

The formula to use would be `LARGE(range, 2)`.

1490. How would you calculate the product of the largest 3 values in a range of data in Excel?

The formula to use would be `PRODUCT(LARGE(range, {1,2,3}))`.

1491. How would you calculate the average of only positive numbers in a range of data in Excel?

The formula to use would be `AVERAGEIF(range, ">0")`.

1492. How would you create a drop-down list in Excel?

To create a drop-down list in Excel, you need to go to the ""Data"" tab, click on ""Data Validation,"" and choose ""List"" as the validation criteria. Then, you can specify the range of cells that contain the list items.

1493. How would you calculate the difference between the largest and smallest values in a range of data in Excel?

The formula to use would be `MAX(range) - MIN(range)`.

1494. . What is the difference between absolute and relative cell references?

Absolute cell references are cell addresses that stay the same when copied or filled to other cells, while relative cell references change when they are copied or filled. Absolute cell references are denoted by the "\$" symbol before the row and/or column reference.

1495. How would you calculate the number of days between two dates in Excel?

The formula to use would be DATEDIF(start_date, end_date, "unit"),

where "unit" can be "d" for days, "m" for months, or "y" for years.

1496. How would you calculate the percent change between two values in Excel?

The formula to use would be (new_value - old_value) / old_value.

1497. How would you calculate the correlation coefficient between two sets of data in Excel?

The formula to use would be CORREL(data1, data2), where "data1" and "data2" are two sets of data.

1498. How can we fix #N/A error when using VLOOKUP function?

Ways to solve #N/A error -

- 1) by referencing it to correct columns
- 2) Using Index/ Match functions instead of VLOOKUP

1499. What are other alternate functions which can be used in place of VLOOKUP?

Index/ Match functions and XLOOKUP()

1500. I want output as Hello/Hola in one cell and both the Hello, Hola are entered in 2 different cells now how can I achieve my desired output?

Ans- by using concatenate()

1501. How can one find outliers using Ms- Excel?

3 ways to find outliers in Ms-Excel are -

- 1) Using sort and filter feature. Finding the largest and smallest values.

2) Can use conditional formatting.

3) Finding it through IQR method.

1502. We have limited number of cells present in one sheet in excel but I want to work on more dataset on the same sheet using more rows and columns and do the required analyses then in this situation how can I do it?

By using Power Pivot we can import large data sets, from different sources, build relationship with them and perform calculations.

1503. What all quality checks a data analyst needs to do and how can it be done in Ms-Excel?

Following are the quality checks a data analyst needs to do -

- 1) Look for proper data types in each column
- 2) Remove extra spaces by using Trim()
- 3) Remove duplicate values by using remove duplicate option
- 4) Finding inconsistent categorical values. Using find and replace feature for it.
- 5) Null values & Outliers - either remove them, or replace them with closest value.

1504. In a column, values are entered as name, city_name. For my analyses I want to have both the values in different columns. How can it be done?

By using Text to columns option we can separate the values in different cells.

1505. What is the date format followed in Ms-Excel?

Ans - mm/dd/yyyy

1506. How can I fix date issues in excel something like 03/12/2022 where 12 should be my month and 03 as date of that day?

Using Text to column feature and setting it into DMY format i.e. date/month/year.

1507. I want to know count, sum and maximum of 5 different product categories of a brand. Then how can we do it quickly?

Ans - Using Pivot Table

1507. How can I filter data in Microsoft Excel?

To filter data in Microsoft Excel, first select the column that you want to filter. Then, go to the ""Data"" tab and click on the ""Filter"" option in the ""Sort & Filter"" group. This will add drop-down arrows to each header in the selected column, allowing you to filter the data based on specific criteria. You can also use advanced filtering options to filter based on complex criteria.

1508. How to use the IF function in MS Excel?

The syntax for the IF function in MS Excel is =IF(condition, value_if_true, value_if_false).

1509. How can I apply conditional formatting in a range of cells?

To apply conditional formatting in Microsoft Excel, first select the range of cells that you want to apply the formatting to. Then, go to the ""Home"" tab and click on the ""Conditional Formatting"" option in the ""Styles"" group. From there, you can select the type of conditional formatting you want to apply, such as highlighting cells that meet a certain criteria or displaying data bars.

1510. How can you create a dynamic named range in Microsoft Excel?

A dynamic named range in Microsoft Excel allows you to create a range of cells that automatically expands or contracts based on the data in your sheet. To create a dynamic named range, you can use a formula such as ""=OFFSET(A1,0,0,COUNT(A:A),1)"" which will create a named range that starts at cell A1 and extends down to the last non-empty cell in column A

1511. How can you use the Solver tool in Microsoft Excel to solve optimization problems?

The Solver tool in Microsoft Excel is a powerful tool that allows you to solve optimization problems by finding the best possible solution based on a set of constraints and objective function. For example, you can use Solver to find the combination of inputs that maximize profit or minimize cost, subject to constraints on the inputs.

1512. How can I format a cell as a percentage in Microsoft Excel?

To format a cell as a percentage in Microsoft Excel, first select the cell or cells that you want to format. Then, right-click on the selected cells and choose ""Format Cells." In the ""Format Cells"" dialog box, select ""Percentage"" from the ""Category"" list and choose the number of decimal places you want to display.

1513. How can I sort data in Microsoft Excel?

To sort data in Microsoft Excel, first select the data that you want to sort. Then, go to the ""Data"" tab and click on the ""Sort"" option in the ""Sort & Filter"" group. In the ""Sort"" dialog box, you can choose the sort order (ascending or descending) and select the column that you want to sort by.

1514. How can you use the VBA programming language in Microsoft Excel to automate tasks and create custom solutions?

The Visual Basic for Applications (VBA) programming language in Microsoft Excel allows you to automate tasks and create custom solutions. With VBA, you can create macros that automate repetitive tasks, create custom functions, and manipulate data in powerful ways. You can access the VBA editor by pressing ALT + F11 in Excel.

1515. How can you use the data analysis tools in Microsoft Excel, such as regression analysis, to perform statistical analysis?

Microsoft Excel provides a variety of data analysis tools that allow you to perform statistical analysis on your data, such as regression analysis. Regression analysis allows you to model the relationship between two or more variables and make predictions based on that relationship. To perform regression analysis in Excel, you can use the Analysis ToolPak add-in, which provides a variety of statistical functions, including regression analysis.

1516. How can you use the INDEX and MATCH functions in Microsoft Excel to perform advanced lookups?

The INDEX and MATCH functions in Microsoft Excel can be used together to perform advanced lookups, such as looking up a value in a table based on multiple criteria. For example, you can use the INDEX function to retrieve a value from a specific cell in a table and the MATCH function to find the row in the table that meets your criteria

1517. is it possible to hide or show the ribbon?

You can hide or show (minimize or maximize) the ribbon by pressing CNTRL F1.

	Math	chemistry	physics
amreez	82	65	99
raja	85	78	98
kiran	88	87	75

**1518. Apply vlookup formula to find the marks of "amreez" in Math subject given the column
is C and cell number is 6.**

You can assume the column labels and row numbers for the above table.

Answer:

=VLOOKUP(C6,J5:M8,2,0)

1519. . How to remove duplicates in the excel sheet?

Data->Data Tools->Remove Duplicates

There are many ways to do it , and the above mentioned is one of the way.

**1520. How to sum values of marks only for ""RAMESH"" in the name column?
You can Assume the range of cells**

Ans: =SUMIF(A2:A5,""=RAMESH"",B2:B5)

1521. How to extract First Name from a full name?

Ans: =MID (A2,1,FIND(" " ",A2)-1)

1522. How can you restrict someone from copying a cell from your worksheet?

Ans:

1. First, choose the data you want to protect.
2. Hit Ctrl + Shift + F. The Format Cells tab appears. Go to the Protection tab. Check Locked and click OK.
3. Next, go to the Review tab and select Protect Sheet. Enter the password to protect the sheet.

1523. How will you write the formula for the following? - Multiply the value in cell A1 by 10, add the result by 5, and divide it by 2.

Ans:

To write a formula for the above-stated question, we have to follow the PEDMAS Precedence. The correct answer is $((A1*10)+5)/2$.

Answers such as $=A1*10+5/2$ and $=(A1*10)+5/2$ are not correct. We must put parentheses brackets after a particular operation.

1524. How to highlight those cells where total sales > \$5000

step 1: Select Conditional Formatting from the home tab and under Highlight Cells Rules, choose Greater Than option.

step 2: Provide the condition ,here in this case 500 and choose the color for the cells to be highlighted.

1525.. Count the number of blank cells in column ""SALES"" from cells A5:A20?

A: $=COUNTBLANK(A5:A20)$ " "Q) Give an example of ""AND"" function

Ans:

$=AND(A35>200,A35<400)$ "

1526. What are left, right, fill and distributed alignments?

Left /Right alignment align the text to left and right most of the cell.Fill as the name suggests, fill the cell with same text repetitively.Distributed, spread the text across the width of the cell.

1527. How is VLOOKUP different from the LOOKUP function?

VLOOKUP lets the user look for a value in the left-most column of a table. It then returns the value in a left-to-right way. It is not very easy to use as compared to the LOOKUP function. LOOKUP function enables the user to look for data in a row/column. It returns the value in another row/column. It is easier and can also be used to replace the VLOOKUP function.

1528. Calculate your age in years from the current date.

Use the YEARFRAC() or DATEDIF() function to return the number of whole days between start_date and end_date

1529. Write a VBA function to calculate the area of a rectangle.

Function Area(Length As Double, Optional Width As Variant)

If IsMissing(Width) Then

 Area = Length * Length

Else

 Area = Length * Width

End If

End Function

1530. Write a VBA code to create a bar chart with the given data.

Consider the below data that has two features. You can use the lines of code below to create a bar chart. Once you have run the above VBA code lines, below is the bar chart you will get.

1531. How do you create a pivot chart in Excel?

To create a pivot chart, first, we need to create a pivot table. Go to the Insert tab next and select the ,Pivot Chart,option. Choose a suitable chart to represent your pivot table data."

1532. How would you work around the VLOOKUP limitation of working only in left to right direction?

In order to avoid error due to VLOOKUP's limitation, i would make sure that the location of the information i wish to seek is always to the left of the value we look for.

1533. - Which function will you use to count values containing a specific range containing data of any type?

COUNTA will return the count of values containing specific data of any type.

1534. When and how does a pivot table gets Refreshed?

A pivot table does not get refreshed automatically. it need to be refresed manually whenever there are any changes made to the dataset.

To refresh a pivot table we must click on the pivot table first to activate ""pivot table tools"" in the ribbon. then click analyse section, in which click refresh button and select refresh.

1535. How would you find the top 10 products sold during a year by revenue if you are given Product ID, Product Name, Unit Price, Unit Sold and Billing Date ?

To find top 10 products sold by revenue we will create a Revenue column by Multiplying Unit Price and Unit sold columns.

Then we use a advanced filter TOP N in filters on Revenue column. We make sure the filter the filter is applied to the whole sheet and not just the column. this will give us out Top 10 products sold by revenue.

1536. A Vlookup formula contains \$A\$3 in its Vlookup value, what pupose does \$ play here?

The \$ sign present in the formula indicates that it is an absolute cell refrence. Which means the creater of the formula wants ensure cell lookup value does not accidentally change while working.

1537. Can you protect your data in Excel? If so, how?

There are two ways in which we can do this.

First, Setting passwords to open the workbook.

Second, removing or hiding sheets.

1538. How do you apply a single format to all the sheets present in a workbook?

To apply the same format to all the sheets of a workbook, we follow the given steps:

Right-click on any sheet present in that workbook.

Then, click on the Select All Sheets option.

Format any of the sheets and you will see that the format has been applied to all the other sheets as well.

1539. There are multiple blank cells present in your dataset. How to ensure they do not interfere in your analysis?

To ensure my data is clean and ready for analysis, I would filter my data column wise to select any blank values present in the dataset. Then I would click on delete button in the ribbon and select delete rows. This ensures that any row with a blank value has been deleted. I will repeat the process for all data column in the dataset.

Second method to delete blank cells is,

click on find, select special, select blanks in the open dilog box, all blank cells in the sheet are selected.

Then, click on delete in the ribbon, select delete rows which deletes all rows that contains blanks."

1540. Is It Possible To Use Multiple Data Sources To Render PivotTables?

Yes, data can be imported from a variety of sources by accessing the Data tab and clicking Get External Data > From Other Sources. Excel worksheet data, data feeds, text files, and other such data formats can be imported, but we'll need to create relationships between the imported tables and those in your worksheet before using them to create a pivot table.

1541. Numerology Sum of the Digits aka Sum the Digits till the result is a single digit

Answer- example, $78 = 7 + 8 = 15 = 1 + 5 = 6$

The formula to achieve the same is

=MOD(A1-1,9)+1

1542. Get File Name through Formula

=CELL("filename",\$A\$1)

1543. Number of times a character appears in a string

Answer - Suppose you want to count the number of times, character ,"" appears in a string
=LEN(A1)-LEN(SUBSTITUTE(LOWER(A1),"a",""))

1544. Calculate Age from Given Birthday

Answer- =DATEDIF(A1,TODAY(),"y")&" Years "&DATEDIF(A1,TODAY(),"ym")&" Months
"&DATEDIF(A1,TODAY(),"md")&" Days"

1545. Number of Days in a Month

Suppose, you have been given a date say 15-Nov-21 and you have to determine how many days this particular month contains.

The formula which you need to use in the above case would be

=DAY(EOMONTH(A1,0))

1546. Reverse a String

Answer- Suppose cell A1:=""qwerty"" and you want to reverse it
=TEXTJOIN(,,MID(A1,LEN(A1)-SEQUENCE(LEN(A1))+1,1))

=TEXTJOIN(,,MID(A1,LEN(A1)-ROW(INDIRECT("'"1:""&LEN(A1)))+1,1))

1547. Convert English Alphabets to Numbers

Answer- There may be scenarios where you need to convert alphabets a, b to y, z to 1, 2 to 25, 26 (Or A, B to Y, Z to 1, 2 to 25, 26) You can use one of the following formulas to do it.

=CODE(LOWER(A1))-96

=CODE(UPPER(A1))-64" "Extract First Name from Full Name

Answer- =LEFT(A1,FIND(" " ",A1&" " ")-1)

1548. Roman Representation of Numbers

Answer- Use ROMAN function.

Hence ROMAN(56) will give LVI.

ROMAN works only for numbers 1 to 3999.

1548. Extract Domain Name from an E Mail ID

Answer- If you want to retrieve domain name which in above example is gmail.com, use following

formula ,

=REPLACE(A1,1,SEARCH("@",A1)+1,"")

1549. Write the function to solve this problem: Divide the value of cell H3 by square root of value of cell A2 and multiply this entire term by 5.

A1) = (H3/(SQRT(A2))*5)

1550. What is the order of precedence while executing operations in MS Excel?

PEDMAS: Parentheses, Exponential, Multiplication, Addition, Subtraction

1551. What is the result of INT(H2) where the data in H2 cell = 115.89?

Ans - 115

1552. Calculate your current age using MS Excel functions when the present date is provided.

Present date: 2023-02-10(CELL:A2), DOB: 2001-06-26(CELL: A3), Current Age =
DATEDIF(A3,A2,""y""") = 21
" "]]]

1553. What will be the result of the function *ch if the data in various cells are ""March, Chair, Chess""?

Chair, Chess

1554. Which function in Excel rounds a given number up, to the nearest multiple of significance?

CEILING

1554. Write a function to find random numbers between 5 and 90.

=RAND()*(90-5)+5

1555. Which function should be entered to calculate all products marked with ""YES"" in the expired column and ""Refrigerator"" in the store column?

=COUNTIFS(B5:B9,""YES"",C5:C9,""Refrigerator"")

1556. How can you update the values of formula cells if Auto Calculate mode of Excel is disabled?

By pressing F9

1557. Which of the following formulas will Excel not calculate?

- A. =SUM(Sales)-A3
- B. =SUM(A1:A5)*.5
- C. =SUM(A1:A5)/(10-10)
- D. =SUM(A1:A5)-10

Ans10) A

Chapter 15 - Big Data Technologies

What is Big Data, and where does it come from? How does it work?

Big Data refers to extensive and often complicated data sets so huge that they're beyond the capacity of managing with conventional software tools. Big Data comprises unstructured and structured data sets such as videos, photos, audio, websites, and multimedia content.

1558. What is PostgreSQL?

PostgreSQL is an enterprise-level, versatile, resilient, open-source, object-relational database management system that supports variable workloads and concurrent users. The international developer community has constantly backed it. PostgreSQL has achieved significant appeal among developers because of its fault-tolerant characteristics.

It's a very reliable database management system, with more than two decades of community work to thank for its high levels of resiliency, integrity, and accuracy. Many online, mobile, geospatial, and analytics applications utilise PostgreSQL as their primary data storage or data warehouse.

1559. What is Debezium used for?

The primary use of Debezium is to enable applications to respond almost immediately whenever data in databases change. Applications can do anything with the insert, update, and delete events. They might use the events to know when to remove entries from a cache. They might update search indexes with the data.

1560. What are CI and CD?

CI and CD stand for continuous integration and continuous delivery/continuous deployment. In very simple terms, CI is a modern software development practice in which incremental code changes are made frequently and reliably.

1561. What is Hadoop used for?

Apache Hadoop is an open-source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

1562. What is Data Warehouse?

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

1563. Difference between data lake and data warehouse.

Data lakes and data warehouses are both widely used for storing big data, but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose.

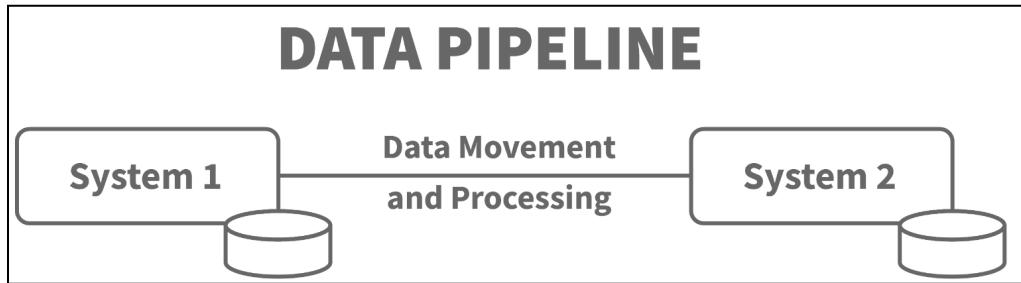
1564. What is orchestration?

IT departments must maintain many servers and apps, but doing it manually isn't scalable. The more complicated an IT system is, the more difficult it is to keep track of all the moving elements. As the requirement to combine numerous automated jobs and their configurations across groups of systems or machines grows, so does the demand to combine multiple automated tasks and their configurations across groups of systems or machines. This is where orchestration comes in handy.

The automated configuration, management, and coordination of computer systems, applications, and services are known as orchestration. IT can manage complicated processes and workflows more easily with orchestration. There are many container orchestration platforms available such as Kubernetes and OpenShift.

1565. What do you mean by data pipeline?

A data pipeline is a system for transporting data from one location (the source) to another (the destination) (such as a data warehouse). Data is converted and optimized along the journey, and it eventually reaches a state that can be evaluated and used to produce business insights. The procedures involved in aggregating, organizing, and transporting data are referred to as a data pipeline. Many of the manual tasks needed in processing and improving continuous data loads are automated by modern data pipelines.



1566. What is the difference between Hive and Presto?

Hive is optimized for query throughput, while Presto is optimized for latency. Presto has a limitation on the maximum amount of memory that each task in a query can store, so if a query requires a large amount of memory, the query simply fails.

1567. Define HDFS

HDFS stands for Hadoop Distributed File System. The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters.

With HDFS, data is written on the server once, and read and reused numerous times after that. HDFS has a primary NameNode, which keeps track of where file data is kept in the cluster.

1568. Difference between Apache Spark and Hadoop?

Hadoop reads and writes files to HDFS, Spark processes data in RAM using a concept known as an RDD, Resilient Distributed Dataset. Spark can run either in stand-alone mode, with a Hadoop cluster serving as the data source, or in conjunction with Mesos

1569.What are the 5 types of database?

They are namely:

- Hierarchical databases.
- Network databases.
- Object-oriented databases.
- Relational databases.
- NoSQL databases.

1570. Does Presto stores data?

Presto is an open source, distributed SQL query engine designed for fast, interactive queries on data in HDFS, and others. Unlike Hadoop/HDFS, it does not have its own storage system.

1571. What is Hive used for?

Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data.

1572. What is the difference between Hive and Presto?

1573. What is the importance of Hadoop?

Hadoop is a beneficial technology for data analysts. There are many essential features in Hadoop which make it so important and user-friendly.

The system is able to store and process enormous amounts of data at an extremely fast rate. A semi-structured, structured and unstructured data set can differ depending on how the data is structured.

Enhances operational decision-making and batch workloads for historical analysis by supporting real-time analytics.

Data can be stored by organisations, and it can be filtered for specific analytical uses by processors as needed.

A large number of nodes can be added to Hadoop as it is scalable, so organisations will be able to pick up more data.

A protection mechanism prevents applications and data processing from being harmed by hardware failures. Nodes that are down are automatically redirected to other nodes, allowing applications to run without interruption.

1574. What are the components of Hadoop?

Following are the main component of Hadoop:-

- Hadoop HDFS - Storage unit of Hadoop
- Hadoop Map Reduce - Processing unit of Hadoop
- Hadoop YARN - Yet Another Resource Negotiator - It is a resource management unit of Hadoop

Master - Namenode (Contains metadata)

Slave - Datanode(Contains data)

YARN and MApReduce divides the resources and assign tasks to the unit

1575. What are the three layers of Big Data Architecture?

The three layers of Big Data Architecture are:-

A. Persistence Layer

Example - HDFS, Amazon Web Services, MongoDB, NoSQL,RDBMS

B. Data Analysis Layer

Example - Data Receiving -> Data Preprocessing -> Data Integration -> Data Preparation -> Data Analytics Module -> Result Publishing module

C. Governance Layer

Ingestion Module, preprocessing module, result module, archival module and auditing module

1576. What is the use of Debezium?

Apache Debezium is used to monitor the change in data in any of the table or databases. Applications can do anything with the insert, update, and delete events. They might use the events to know when to remove entries from a cache. They might update search indexes with the data.

1577. What is the difference between Kafka Connect and Debezium?

Ans. Debezium platform has a vast set of CDC connectors, while Kafka Connect comprises various JDBC connectors to interact with external or downstream applications. However, Debezium's CDC connectors can only be used as a source connector that captures real-time event change records from external database systems.

1578. What is CDC?

Change data capture (CDC) refers to the process of identifying and capturing changes made to data in a database and then delivering those changes in real-time to a downstream process or system.

1579. How is Hadoop and Big Data related?

If we talk about Big Data, we do talk about Hadoop as well. So, this is one of the most critical questions from an interview perspective. That you might surely face. Hadoop is an open-source framework for saving, processing, and interpreting complex, disorganized data sets for obtaining insights and knowledge. So, that is how Hadoop and Big Data are related to each other.

1580. How Map Reduce works?

The responsible layer of Hadoop for data processing is MapReduce. It puts a request for processing of structured and unstructured data which is already stored in HDFS. It is liable for the parallel processing of a high volume of data by distributing data into detached tasks. There

are two stages of processing: Map and Reduce. In simple terms, Map is a stage where data blocks are read and made available to the executors (computers /nodes /containers) for processing. Reduce is a stage where all processed data is collected and collated.

1581. How to deploy a Big Data Model? Mention the key steps involved.

Deploying a model into a Big Data Platform involves mainly three key steps they are,

Data ingestion

Data Storage

Data Processing

** NoSQL databases (aka "not only SQL") are non-tabular databases and store data differently than relational tables. NoSQL databases come in a variety of types based on their data model. The main types are document, key-value, wide-column, and graph.

1582. What is the use of Cassandra?

Apache Cassandra is an open source, distributed and decentralized/distributed storage system (database), for managing very large amounts of structured data spread out across the world. It provides highly available service with no single point of failure. It is scalable, fault-tolerant, and consistent.

1583. Is Cassandra a SQL database?

Cassandra is a NoSQL distributed database. By design, NoSQL databases are lightweight, open-source, non-relational, and largely distributed. Counted among their strengths are horizontal scalability, distributed architectures, and a flexible approach to schema definition.

1584. Is MongoDB better than Cassandra?

Key Factors That Drive the Apache Cassandra Versus MongoDB Decision. Scalability and Speed: Cassandra can be preferred if high scalability with faster writing speed is the main requirement. Data availability: MongoDB is a good choice if consistency is a priority and availability can be compromised.

1585. S3 vs HDFS

S3 and cloud storage provide elasticity, with an order of magnitude better availability and durability and 2X better performance, at 10X lower cost than traditional HDFS data storage clusters. Hadoop and HDFS commoditized big data storage by making it cheap to store and distribute a large amount of data

1586. Is S3 slower than HDFS?

S3 is slower to work with than HDFS, even on virtual clusters running on Amazon EC2.

1587.. What are the different layers of Data platform architecture?

Data Ingestion Layer

Data Storage Layer

Data Processing Layer

User Interface Layer

Data Pipeline Layer

1588. What is meant by data cube?

A data cube is a data structure that, contrary to tables and spreadsheets, can store data in more than 2 dimensions. They are mainly used for fast retrieval of aggregated data. The key elements of a data cube are dimensions, attributes, facts and measures.

1589. What is the difference between OLAP and OLTP?

OLTP and OLAP: The two terms look similar but refer to different kinds of systems. Online transaction processing (OLTP) captures, stores, and processes data from transactions in real time. Online analytical processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems.

1590. Which is better OLAP or OLTP?

An OLAP system is designed to process large amounts of data quickly, allowing users to analyze multiple data dimensions in tandem. Teams can use this data for decision-making and problem-solving. In contrast, OLTP systems are designed to handle large volumes of transactional data involving multiple users.

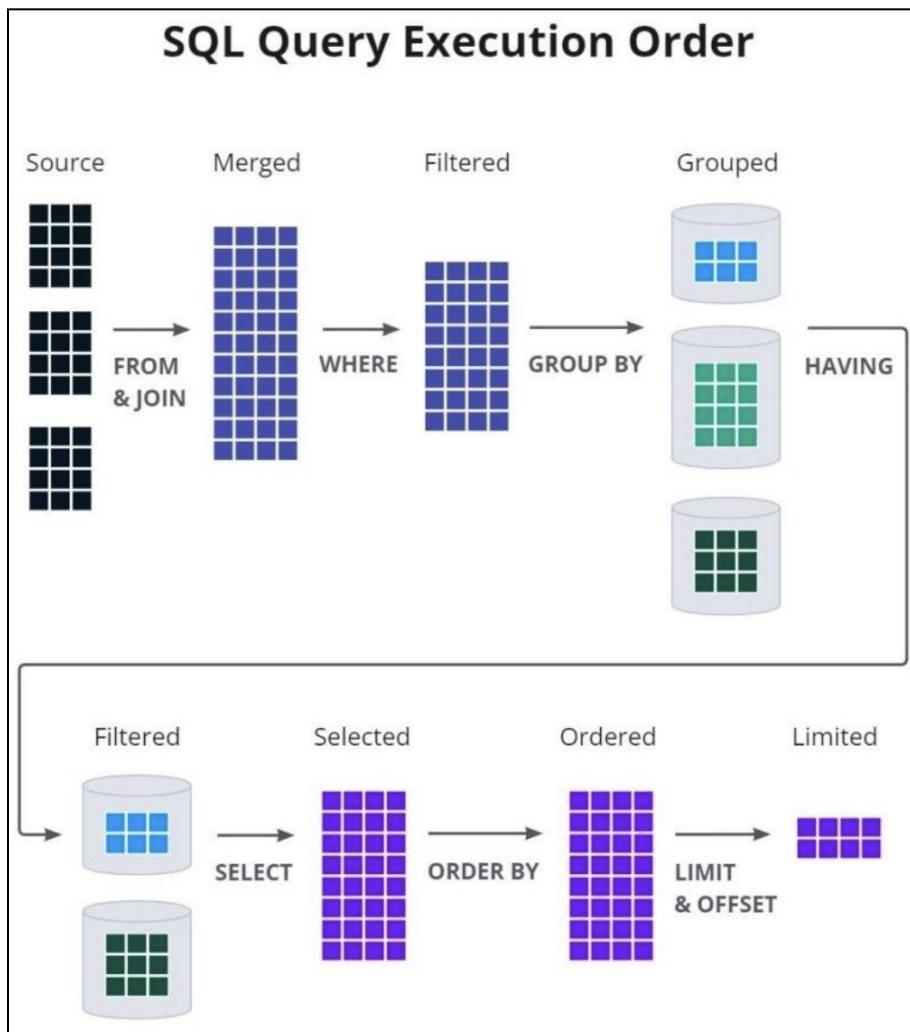
1591. Is Hadoop OLTP or OLAP?

Hadoop is an OLAP. Hadoop is neither OLAP nor OLTP. All above are true statements. Since we use Hadoop to process the analysis of big data & this is done by batch wise on historical data which is loaded in the HDFS (Hadoop distributed file system)

1592. What is the ACID in database?

ACID is an acronym that stands for atomicity, consistency, isolation, and durability. Together, these ACID properties ensure that a set of database operations (grouped together in a transaction) leave the database in a valid state even in the event of unexpected errors.

1593. SQL order of execution



1594. What is a Zookeeper? What are the benefits of using a zookeeper?

Hadoop's most remarkable technique for addressing big data challenges is its capability to divide and conquer with Zookeeper. After the problem has been divided, the conquering relies on employing distributed and parallel processing methods across the Hadoop cluster.

The interactive tools cannot provide the insights or timeliness needed to make business judgments for big data problems. In those cases, you need to build distributed applications to solve those big data problems. Zookeeper is Hadoop's way of coordinating all the elements of these distributed applications.

Zookeeper as technology is simple, but its features are powerful. Arguably, it would be difficult, if not impossible, to create resilient, fault-tolerant distributed Hadoop applications without it.

1595. What are the benefits of Zookeeper?

Benefits of using a Zookeeper are:

- Simple distributed coordination process: The coordination process among all nodes in Zookeeper is straightforward.
- Synchronization: Mutual exclusion and co-operation among server processes.
- Ordered Messages: Zookeeper tracks with a number by denoting its order with the stamping of each update; with the help of all this, messages are ordered here.
- Serialization: Encode the data according to specific rules. Ensure your application runs consistently.
- Reliability: The zookeeper is very reliable. In case of an update, it keeps all the data until forwarded.
- Atomicity: Data transfer either succeeds or fails, but no transaction is partial.

1596. What is the default replication factor in HDFS?

By default, the replication factor is 3. There are no two copies that will be on the same data node. Usually, the first two copies will be on the same rack, and the third copy will be off the shelf. It is advised to set the replication factor to at least three so that one copy is always safe, even if something happens to the rack.

We can set the default replication factor of the file system and each file and directory exclusively. We can lower the replication factor for files that are not essential, and critical files should have a high replication factor.

1597. What are the features of Apache Sqoop?

- Robust: It is extremely robust and easy to use. In addition, it has community support and contribution.
- Full Load: Loading a table in Sqoop can be done in one command. Multiple tables can also be loaded in the same process.
- Incremental Load: Incremental load functionality is also supported. Whenever the table is updated, with the help of Sqoop, it can be loaded in parts too.
- Parallel import/export: Importing and exporting of data is done by the YARN framework. It also provides fault tolerance too.
- Import results of SQL query: It allows us to import the output from the SQL query into the Hadoop Distributed File System.

1598. What is partitioning in Hive?

In general partitioning in Hive is a logical division of tables into related columns such as date, city, and department based on the values of partitioned columns. Then these partitions are

subdivided into buckets so that they provide extra structure to the data that may be used for more efficient querying

1599. Explain the steps of big data project life cycle

The Big Data Analytics Life cycle is divided into nine phases, named as :

Business Case/Problem Definition

Data Identification

Data Acquisition and filtration

Data Extraction

Data Munging(Validation and Cleaning)

Data Aggregation & Representation(Storage)

Exploratory Data Analysis

Data Visualization(Preparation for Modeling and Assessment)

Utilization of analysis results.

1600. What Are Big Data Infrastructure Solutions?

There are several big data architecture solutions today, many of which function across multiple distribution platforms.

Some of these solutions include the following:

Hadoop: Hadoop is an open-source big data framework written in Java and maintained by the Apache Foundation. Hadoop is actually a series of components that include the HDFS storage layer, the MapReduce analytics engine, the YARN HA cluster and application orchestration management system, and the Spark framework for machine learning apps. As an open-source application, Hadoop is a popular, cost-effective solution for engineers and admins who want a well-maintained project. It is, however, open-source, subject to a lack of central support and maintenance from a third party.

NoSQL: Known as “Not Only SQL,” NoSQL provides a distributed database framework for big data architectures that need low-latency retrieval for structured and semistructured data. NoSQL is actually an umbrella term for this type of technology, and as such, there are several implementations (Apache Cassandra, Oracle NoSQL, etc.) maintained by different enterprises and organizations. It also often works in tandem with other solutions like Hadoop.

Massively Parallel Processing: A solution for massive big data appliances, MPP often powers higher-end systems that require extremely large parallel processing applications across thousands of individual processors. Like NoSQL, MPP is a name for a set of technologies offered by specialist providers that have subsequently been absorbed by larger tech companies like IBM, HP, and EMC.

Cloud Computing: Big data infrastructures aren't always cloud-based. In fact, tech like Hadoop doesn't require a cloud environment, as it was designed to work across distributed hardware. However, cloud technology brings some of the key aspects of big data to the table (orchestration, HA storage, etc.) without requiring a company to implement them itself. That being said, these tech platforms aren't mutually exclusive. It's quite common, for example, to see big data cloud systems running Hadoop as part of their overall architecture.

1601. Different file formats to store and process input data using Apache Hadoop

CSV Files

CSV files are an ideal fit for exchanging data between hadoop and external systems. It is advisable not to use header and footer lines when using CSV files.

JSON Files

Every JSON File has its own record. JSON stores both data and schema together in a record and also enables complete schema evolution and splitability. However, JSON files do not support block-level compression.

Avro Fliles

This kind of file format is best suited for long term storage with Schema. Avro files store metadata with data and also let you specify independent schema for reading the files.

Parquet Files

A columnar file format that supports block level compression and is optimized for query performance as it allows selection of 10 or less columns from from 50+ columns records.

1602. What is zookeeper?

ZooKeeper is a distributed coordination service for distributed systems. It provides a simple interface to a centralized coordination service, which allows distributed applications to perform common services such as configuration management, naming, and group services in a reliable and easy-to-use manner. It is often used in distributed systems, such as Apache Kafka and Apache Hadoop, to manage and coordinate distributed processes.

1603. What is sqoop?

Sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. It can be used to import data from external structured sources into the Hadoop Distributed File System (HDFS) or related systems like Apache Hive or Apache HBase, and also to export data from Hadoop to external structured sources. Sqoop uses a connector-based architecture to support many different databases,

including MySQL, Oracle, and PostgreSQL. It also supports incremental imports and exports, and can be used to import and export individual tables or entire databases.

1604. What is cdc in data engineering?

In the context of data engineering, CDC stands for "Change Data Capture." Change Data Capture is a technique used in data integration and replication processes to track changes to source data systems and replicate those changes to one or more target systems.

CDC technology captures changes made to a source database, such as inserts, updates, and deletes, and records them in a log or journal. This log is then used to identify and extract only the changed data, which can be sent to one or more target systems for processing, analysis, or storage.

CDC is a powerful tool for data engineers and data analysts, as it allows them to keep data in sync across different systems, without having to manually perform data migrations or full data extracts. This can help to improve the accuracy, timeliness, and efficiency of data integration processes, and make it easier to analyze and use data across different parts of an organization.

1605. What is Hudi and what is it used for?

Hudi (Hadoop Upserts Deletes and Incrementals) is an open-source data management framework developed by Apache Software Foundation. Hudi is designed to simplify data engineering processes for big data workloads, with a focus on managing large-scale data sets that require frequent updates and real-time analysis.

Hudi enables data engineers and developers to perform CRUD (Create, Read, Update, and Delete) operations on large data sets in real-time, while also providing built-in support for data versioning, incremental processing, and ACID (Atomicity, Consistency, Isolation, and Durability) transactions.

Hudi achieves this by using a combination of techniques, including columnar file formats, data partitioning, and change data capture (CDC) to enable efficient and scalable data updates. Hudi is integrated with Apache Spark, which enables Hudi to leverage Spark's distributed computing capabilities for large-scale data processing.

Hudi is widely used in various industries, such as e-commerce, finance, and healthcare, where real-time processing of large-scale data sets is critical for making timely and accurate business decisions. Hudi provides a way to manage data at scale with ease, while also reducing the complexity and cost of data management in big data environments.

1606. What is CRUD in Data Engineering?

CRUD stands for Create, Read, Update, and Delete. It is a set of basic operations used in data engineering and database management to manipulate data stored in a database.

Create refers to the operation of adding new data records to a database, typically by inserting a new row or document.

Read refers to the operation of retrieving data records from a database, typically by querying the database for specific records that match certain criteria.

Update refers to the operation of modifying existing data records in a database, typically by changing the values of one or more fields in an existing row or document.

Delete refers to the operation of removing data records from a database, typically by deleting a row or document.

CRUD operations are the most basic building blocks of database management and are used extensively in data engineering processes, such as data integration, data processing, and data analysis. By enabling developers and data engineers to perform these basic operations, databases provide a flexible and scalable way to manage and manipulate large data sets.

1607. What is the difference between delete, upsert and incremental pipeline?

Upsert, delete, and incremental pipeline are all techniques used in data engineering for managing changes to data stored in a database or data warehouse. Here are the differences between these techniques:

Upsert: Upsert is a combination of update and insert operations. It is used to update an existing record in the database if it exists, or insert a new record if it doesn't exist. Upsert is useful for scenarios where data may be updated frequently and there may be a mix of new and existing records that need to be updated.

Delete: Delete is a data operation used to remove a record or set of records from a database. Deleting data is often necessary to maintain the integrity and accuracy of the data in a database.

Incremental pipeline: Incremental pipeline is a technique used to process only the changes or updates that have been made to a dataset since the last processing run, instead of processing the entire dataset again. This is useful for datasets that change frequently, as it reduces processing time and computational resources required to update the data.

In summary, upsert is a way to handle both inserts and updates in a single operation, delete is used to remove data, and incremental pipeline is a technique used to process only the changes

that have occurred since the last processing run. These techniques are often used in combination with each other to manage and maintain data integrity and accuracy in data engineering pipelines.

1608. How is data partitioned in Hadoop, and why is this important for performance?

In Hadoop, data is partitioned into blocks and stored across a cluster of machines in a distributed file system, such as Hadoop Distributed File System (HDFS). The data is divided into smaller chunks to be processed in parallel on multiple nodes. This process of dividing data into smaller pieces and distributing them across multiple nodes is called data partitioning.

Data partitioning is important for performance because it allows for parallel processing, which can significantly speed up data processing tasks. By dividing the data into smaller chunks and processing them in parallel, it reduces the time needed to process the data compared to traditional sequential processing methods.

1609. What is MapReduce, and how is it used in Hadoop?

MapReduce is a programming model used to process large volumes of data in a distributed computing environment, like Hadoop. It consists of two main tasks - Map and Reduce - that are executed in parallel across a large number of nodes.

In MapReduce, the Map task processes input data and converts it into key-value pairs. The Reduce task then processes the key-value pairs, aggregates the values, and produces output data. MapReduce is designed to efficiently process large amounts of data in a distributed computing environment by dividing the data into smaller chunks, processing them in parallel, and combining the results.

MapReduce is used in Hadoop to process large datasets and is a key component of Hadoop's distributed processing capabilities.

1610. What is Spark, and how does it differ from Hadoop? What are some common use cases for Spark?

Apache Spark is a distributed computing framework that is used to process large datasets across a cluster of machines. It is often used as an alternative to Hadoop's MapReduce for processing big data. Spark offers a more general-purpose, faster, and more flexible way to process data than MapReduce.

One of the key differences between Spark and Hadoop is the way they handle data processing. Spark is designed to keep data in memory whenever possible, whereas Hadoop's MapReduce writes intermediate results to disk, which can slow down processing times.

Some common use cases for Spark include real-time data processing, machine learning, and graph processing. It is particularly well-suited for iterative processing tasks, such as those involved in training machine learning models.

1611. What is a distributed file system, and how does it differ from a traditional file system?

A distributed file system is a file system that is distributed across multiple machines in a network. It allows multiple users to access the same files and data from different locations simultaneously. Distributed file systems are designed to be fault-tolerant and scalable, which makes them ideal for handling large amounts of data.

In contrast, traditional file systems are stored on a single machine and can only be accessed by one user at a time. Traditional file systems are not designed to handle large amounts of data or to be fault-tolerant.

1612. What is the CAP theorem, and how does it relate to distributed systems?

The CAP theorem, also known as Brewer's theorem, is a concept in distributed systems that states that it is impossible to achieve all three of the following properties simultaneously: consistency, availability, and partition tolerance.

Consistency refers to the idea that all nodes in a distributed system should see the same data at the same time. Availability means that a distributed system should always be available to respond to requests, even if some nodes in the system fail. Partition tolerance means that a distributed system should be able to continue operating even if network connections between nodes fail or are slow.

The CAP theorem states that a distributed system can only achieve two of these three properties at the same time. For example, a system can be consistent and partition tolerant, but not always available.

1613. What is NoSQL, and how does it differ from traditional relational databases?

NoSQL, or "not only SQL," refers to a class of database management systems that do not use the traditional tabular relations used in relational databases. NoSQL databases are designed to handle large amounts of unstructured or semi-structured data, such as social media data or sensor data. Unlike relational databases, which use a schema to define the structure of the data, NoSQL databases are schema-less, meaning they can handle data that is highly variable in structure. NoSQL databases can also be highly scalable, with the ability to handle large volumes of data and high levels of concurrent users. Some common types of NoSQL databases include document-oriented databases, key-value stores, and graph databases.

1614. What is a data warehouse, and how is it used for business intelligence?

A data warehouse is a large, centralized repository of data that is used for business intelligence (BI). The data in a data warehouse is typically extracted from multiple, heterogeneous sources, such as transactional databases, CRM systems, and marketing automation platforms. Once the data is loaded into the data warehouse, it is transformed and organized to support reporting and analysis. Data warehouses are designed to handle large volumes of data and complex queries, and they are optimized for read performance.

1615. How do you ensure the security and privacy of big data, especially when dealing with sensitive information?

There are several ways to ensure the security and privacy of big data, including implementing encryption and access controls, conducting regular security audits, and ensuring compliance with data protection regulations such as GDPR or HIPAA. Additionally, data masking and anonymization techniques can be used to hide sensitive data while still allowing for analysis.

1616. What are some of the most common techniques used to extract insights from big data?

Some of the most common techniques used to extract insights from big data include data mining, machine learning, natural language processing, and sentiment analysis. These techniques can be used to identify patterns, trends, and anomalies in large data sets and to generate predictive models that can be used to make data-driven decisions.

1617. How do you validate the accuracy and completeness of big data sets?

Validating the accuracy and completeness of big data sets can be challenging due to their size and complexity. One approach is to use data profiling techniques to identify any anomalies or missing data. Another approach is to use statistical methods to measure the distribution of data values and to identify outliers. Additionally, data visualization techniques can be used to quickly identify any issues with the data.

1618. What are some of the most common tools and technologies used in big data analytics?

Some of the most common tools and technologies used in big data analytics include Hadoop, Spark, Hive, Pig, Cassandra, and MongoDB. These tools provide the infrastructure and processing capabilities required for managing and analyzing large data sets.

1619. How do you handle data quality issues in big data sets?

Handling data quality issues in big data sets requires a systematic approach. This can include

data profiling to identify any issues with the data, data cleansing to remove any duplicate or inconsistent data, and data enrichment to supplement the data with additional information. Additionally, data governance policies can be put in place to ensure that data quality is maintained over time.

1620. What are some of the most common use cases for big data processing, and how do they provide value to organizations?

Common use cases for big data processing include fraud detection, customer segmentation, supply chain optimization, sentiment analysis, and predictive maintenance. These use cases provide value to organizations by allowing them to make data-driven decisions that can improve efficiency, reduce costs, and enhance the customer experience. For example, fraud detection can help financial institutions to identify and prevent fraudulent activity, while customer segmentation can help retailers to better understand their customers and tailor their marketing efforts accordingly.

1621. In what all modes can Hadoop be run?

Hadoop can be run in three different modes - local or standalone mode, pseudo-distributed mode, and fully-distributed mode. In local mode, both the Hadoop Distributed File System (HDFS) and the Hadoop processing engine run on a single machine. In pseudo-distributed mode, the Hadoop components are distributed across a single machine, simulating a multi-node cluster. In fully-distributed mode, the Hadoop components are distributed across multiple machines, forming a true multi-node cluster.

1622. What are the real-time industry applications of Hadoop?

Some of the real-time industry applications of Hadoop include fraud detection, sentiment analysis, customer segmentation, log processing, recommendation systems, and supply chain optimization. Hadoop is also used in industries such as finance, healthcare, retail, and telecommunications to manage and analyze large data sets.

1623. What is HBase?

HBase is a NoSQL, column-oriented database that runs on top of Hadoop. It is designed to handle large amounts of structured data and provide low-latency access to that data. HBase is commonly used in big data applications for storing and retrieving large amounts of data quickly and efficiently.

1624. What is a Combiner?

A Combiner is a function in Hadoop that performs an intermediate data reduction step during the map-reduce process. The output of the map function is typically a large amount of data that needs to be shuffled and sorted before being passed on to the reduce function. The Combiner

function is run on the output of the map function before it is shuffled and sorted, in order to reduce the amount of data that needs to be processed by the reduce function. This helps to improve the performance of the map-reduce process by reducing the amount of network traffic and disk I/O required.

1625. Name some of the important tools useful for Big Data analytics.

Ans. It is one of the most commonly asked big data interview questions.

The important Big Data analytics tools are –

NodeXL
KNIME
Tableau
Solver
OpenRefine
Rattle GUI
Qlikview

1626. What are the five ‘V’s of Big Data?

Ans. It is one of the most popular big data interview questions.

The five ‘V’s of Big data are –

Value – Value refers to the worth of the data being extracted.

Variety (Data in Many forms) – Variety explains different types of data, including text, audio, videos, photos, and PDFs, etc.

Veracity (Data in Doubt) – Veracity talks about the quality or trustworthiness and accuracy of the processed data.

Velocity (Data in Motion) – This refers to the speed at which the data is being generated, collected, and analyzed.

Volume (Data at Rest) – Volume represents the volume or amount of data. Social media, mobile phones, cars, credit cards, photos, and videos majorly contribute to the volumes of data.

1627. What are HDFS and YARN? What are their respective components?

Ans. HDFS or Hadoop Distributed File System runs on commodity hardware and is highly fault-tolerant. HDFS provides file permissions and authentication and is suitable for distributed

storage and processing. It is composed of three elements, including NameNode, DataNode, and Secondary NameNode.

YARN is an integral part of Hadoop 2.0 and is an abbreviation for Yet Another Resource Negotiator. It is a resource management layer of Hadoop and allows different data processing engines like graph processing, interactive processing, stream processing, and batch processing to run and process data stored in HDFS. ResourceManager and NodeManager are the two main components of YARN.

Also Read>> [Top Online IT Courses for IT Professionals](#)

1628. What is FSCK?

Ans. FSCK or File System Check is a command used by HDFS. It checks if any file is corrupt, has its replica, or if there are some missing blocks for a file. FSCK generates a summary report, which lists the overall health of the file system.

1629. What is the goal of A/B Testing?

Ans. A/B testing is a comparative study, where two or more variants of a page are presented before random users and their feedback is statistically analyzed to check which variation performs better.

1630. What is a Distributed Cache?

Ans. Distributed Cache is a dedicated service of the Hadoop MapReduce framework, which is used to cache the files whenever required by the applications. This can cache read-only text files, archives, jar files, among others, which can be accessed and read later on each data node where map/reduce tasks are running.

It is among the most commonly asked big data interview questions and you must read about Distributed Cache in detail.

1631. Explain the steps to deploy a Big Data solution.

Ans. Followings steps are followed to deploy a Big Data Solution –

Data Ingestion: It is the first step in the big data solution deployment. The data is extracted from various resources such as Salesforce, SAP, MySQL and other log files, documents, etc. Data ingestion can be performed through batch jobs or real-time streaming.

Data Storage: The extracted data is either stored in HDFS or NoSQL database (i.e. HBase). HDFS storage is used for sequential access while HBase is used for random read/write access.

Data Processing: This is the final step in big data solution deployment. Data is processed through a processing framework such as Pig, Spark, MapReduce, etc.

1632. What is Hive Metastore?

Answer: Hive metastore is a database that stores metadata about your Hive tables (eg. Table name, column names and types, table location, storage handler being used, number of buckets in the table, sorting columns if any, partition columns if any, etc.).

When you create a table, this metastore gets updated with the information related to the new table which gets queried when you issue queries on that table.

Hive is a central repository of hive metadata. it has 2 parts of services and data. by default, it uses derby DB in local disk. it is referred to as embedded metastore configuration. It tends to the limitation that only one session can be served at any given point of time.

1633. What kind of Dataware house application is suitable?

Answer: Hive is not a full database. The design constraints and limitations of Hadoop and HDFS impose limits on what Hive can do.

Hive is most suited for data warehouse applications, where

- 1) Relatively static data is analyzed,
- 2) Fast response times are not required, and
- 3) When the data is not changing rapidly.

Hive doesn't provide crucial features required for OLTP, Online Transaction Processing. It's closer to being an OLAP tool, Online Analytic Processing. So, Hive is best suited for data warehouse applications, where a large data set is maintained and mined for insights, reports, etc.

1634. It's true that HDFS is to be used for applications that have large data sets. Why is it not the correct tool to use when there are many small files?

Answer: In most cases, HDFS is not considered as an essential tool for handling bits and pieces of data spread across different small-sized files. The reason behind this is "Namenode" happens to be a very costly and high-performing system. The space allocated to "Namenode" should be used for essential metadata that's generated for a single file only, instead of numerous small files. While handling large quantities of data attributed to a single file, "Namenode" occupies lesser space and therefore gives off optimized performance. With this in view, HDFS should be used for supporting large data files rather than multiple files with small data.

1635. What are the main distinctions between NAS and HDFS?

Answer: HDFS needs a cluster of machines for its operations, while NAS runs on just a single machine. Because of this, data redundancy becomes a common feature in HDFS. As the replication protocol is different in the case of NAS, the probability of the occurrence of redundant data is much less.

Data is stored on dedicated hardware in NAS. On the other hand, the local drives of the machines in the cluster are used for saving data blocks in HDFS.

Unlike HDFS, Hadoop MapReduce has no role in the processing of NAS data. This is because computation is not moved to data in NAS jobs, and the resultant data files are stored without the same.

1636. Is it possible to create multiple tables in the hive for the same data?

Hive creates a schema and appends on top of an existing data file. One can have multiple schemas for one data file, the schema would be saved in hive's metastore and data will not be parsed read or serialized to disk in a given schema. When s/he will try to retrieve data schema will be used. Let's say if my file has 5 columns (Id, Name, Class, Section, Course) we can have multiple schemas by choosing any number of the column.

1637. What is Hadoop Big Data Testing?

Big Data means a vast collection of structured and unstructured data, which is very expansive & is complicated to process by conventional database and software techniques. In many organizations, the volume of data is enormous, and it moves too fast in modern days and exceeds the current processing capacity. Compilation of databases that are not being processed by conventional computing techniques, efficiently. Testing involves specialized tools, frameworks, and methods to handle these massive amounts of datasets. Examination of Big data is meant to the creation of data and its storage, retrieving of data and analysis them which is significant regarding its volume and variety of speed.

1638. What do we test in Hadoop Big Data?

In the case of processing of the significant amount of data, performance, and functional testing is the primary key to performance. Testing is a validation of the data processing capability of the project and not the examination of the typical software features.

1639. How do we validate Big Data?

In Hadoop, engineers authenticate the processing of quantum of data used by the Hadoop cluster with supportive elements. Testing of Big data needs asks for extremely skilled professionals, as the handling is swift. Processing is three types namely Batch, Real-Time, & Interactive.

1640. How is data quality being tested?

Along with processing capability, the quality of data is an essential factor while testing big data. Before testing, it is obligatory to ensure the data quality, which will be part of the examination of the database. It involves the inspection of various properties like conformity, perfection, repetition, reliability, validity, completeness of data, etc.

1641. What do you understand by Data Staging?

The initial step in the validation, which engages in process verification. Data from a different source like social media, RDBMS, etc. are validated, so that accurate uploaded data to the system. We should then compare the data source with the uploaded data into HDFS to ensure that both of them match. Lastly, we should validate that the correct data has been pulled, and uploaded into a specific HDFS. There are many tools available, e.g., Talend, Datameer, which are mostly used for validation of data staging.

1642. What is Architecture Testing?

This pattern of testing is to process a vast amount of data extremely resources intensive. That is why testing the architecture is vital for the success of any project on Big Data. A faulty planned system will lead to degradation of the performance, and the whole system might not meet the desired expectations of the organization. At least, failover and performance test services need a proper performance in any Hadoop environment.

1643. What is Performance Testing?

Performance testing consists of testing of the duration to complete the job, utilization of memory, the throughput of data, and parallel system metrics. Any failover test services aim to confirm that data is processed seamlessly in any case of data node failure. Performance Testing of Big Data primarily consists of two functions. First, is Data ingestion whereas the second is Data Processing

1644. What is Data ingestion?

The developer validates how fast the system is consuming the data from different sources. Testing involves the identification process of multiple messages that are being processed by a queue within a specific frame of time. It also consists of how fast the data gets into a particular data store, e.g., the rate of insertion into the Cassandra & Mongo database.

1645. What is Data Processing in Hadoop Big data testing?

It involves validating the rate at which map-reduce tasks are performed. It also consists of data testing, which can be processed in separation when the primary store is full of data sets. E.g., Map-Reduce tasks running on a specific HDFS.

1646. What are the tools applied in these scenarios of testing?

Cluster	Tools
NoSQL:	Cassandra, CouchDB, DatabasesMongoDB, Redis, HBase. ZooKeeper
MapReduce:	Hadoop, Pig, Hive, Cascading, Kafka, Oozie, S4, Flume, MapR
Storage:	HDFS, S3
Servers:	Elastic, EC2, Heroku
Processing	Mechanical Turk, R, Yahoo! BigSheets.

1647. What is Query Surge?

Query Surge is one of the solutions for Big Data testing. It ensures the quality of data quality and the shared data testing method that detects bad data while testing and provides an excellent view of the health of data. It makes sure that the data extracted from the sources stay intact on the target by examining and pinpointing the differences in the Big Data wherever necessary.

1648. What Benefits do Query Surge provides?

Query Surge helps us to automate the efforts made by us manually in the testing of Big Data. It offers to test across diverse platforms available like Hadoop, Teradata, MongoDB, Oracle, Microsoft, IBM, Cloudera, Amazon, HortonWorks, MapR, DataStax, and other Hadoop vendors like Excel, flat files, XML, etc.

Enhancing Testing speeds by more than thousands of times while at the same time offering the coverage of entire data.

Delivering Continuously – Query Surge integrates DevOps solution for almost all Build, QA software for management, ETL.

It also provides automated reports by email with dashboards stating the health of data.

Providing excellent Return on the Investments (ROI), as high as 1,500%

1649. What is Query Surge's architecture?

Query Surge Architecture consists of the following components:

Tomcat - The Query Surge Application Server

The Query Surge Database (MySQL)

Query Surge Agents – At least one has to be deployed

Query Surge Execution API, which is optional.

1650. Can you explain the difference between batch processing and stream processing, and when would you use one over the other?

Batch processing involves processing data in large volumes and in batches, while stream processing involves processing data in real-time as it arrives. Batch processing is typically used when dealing with large volumes of data that can be processed in batches, such as nightly processing of transactional data or log files. Stream processing, on the other hand, is used when immediate action is required, such as in real-time analytics or fraud detection.

1651. How do you measure the performance of a big data processing system, and what metrics are commonly used for this purpose?

The performance of a big data processing system can be measured using various metrics, including throughput, latency, and resource utilization. Throughput measures the rate at which data is processed, while latency measures the time it takes for a system to respond to a request or an event. Resource utilization measures the efficiency with which system resources such as CPU, memory, and disk are used. Other metrics can include the number of records processed per unit of time, the rate of errors or failures, and the system's scalability and fault tolerance.

1652. Can you describe the Lambda Architecture, and how it can be used to build a scalable and fault-tolerant big data processing system?

The Lambda Architecture is a data processing architecture designed to handle both batch and stream processing by combining the strengths of both approaches. It involves three layers: a batch layer, a speed layer, and a serving layer. The batch layer processes large volumes of data in a batch-oriented fashion and stores the results in a master dataset, while the speed layer processes data in real-time and stores the results in a stream. The serving layer merges the results from the batch and speed layers to provide a unified view of the data. The Lambda Architecture can be used to build a scalable and fault-tolerant big data processing system by handling both batch and real-time data processing needs.

1653. What are some of the common data integration challenges faced in big data projects, and how can they be addressed?

Common data integration challenges in big data projects include dealing with data quality

issues, data heterogeneity, and data volume. These challenges can be addressed by using data integration tools and technologies, such as ETL (extract, transform, load) tools, data mapping tools, and data quality tools. Other approaches include data profiling, data cleansing, and data standardization.

1654. Can you explain how data sharding works, and how it can be used to improve the performance of a distributed data processing system?

Data sharding is a technique used to horizontally partition data across multiple machines or nodes in a distributed data processing system. Each node is responsible for storing and processing a subset of the data, which improves system performance and scalability. Data sharding can be used to improve performance in distributed data processing systems by reducing the amount of data that needs to be processed by each node.

1655. What are the key considerations when designing a data model for a NoSQL database, and how does this differ from designing a data model for a relational database?

When designing a data model for a NoSQL database, key considerations include denormalization, data duplication, and the use of partition keys. NoSQL databases are often designed to be highly scalable, which requires a different approach to data modeling than relational databases. This can include using denormalized data models, where redundant data is stored to avoid costly joins, and using partition keys to distribute data across multiple nodes. In contrast, relational databases often require a more normalized data model to avoid data redundancy and ensure data consistency.

1656. Can you describe some of the common approaches for data governance in big data projects, and how they can be used to ensure compliance with data privacy and security regulations?

Common approaches for data governance in big data projects include data classification, data lineage, and access controls. Data classification involves categorizing data based on its sensitivity and importance, while data lineage tracks the origin and movement of data throughout the data processing pipeline. Access controls involve defining who can access data and what they can do with it. These approaches can be used to ensure compliance with data privacy and security regulations.

1657. How can you optimize the performance of a big data processing system, and what techniques are commonly used for this purpose?

Techniques for optimizing the performance of a big data processing system can include tuning the system's hardware and software settings, optimizing data partitioning, and using caching and in-memory computing. Other approaches include using data compression, data indexing,

and query optimization. Machine learning and artificial intelligence techniques can also be used to optimize performance by automating and improving data processing tasks.

Chapter 16 - Amazon Web Services

1658. Explain what is AWS ?

Amazon Web Services (AWS) is a cloud computing platform and infrastructure provided by Amazon.com. It provides a suite of cloud-based computing services, including storage, networking, databases, and computing power, that can be used to build, run, and manage applications and services on the cloud.

AWS is designed to provide scalable, reliable, and cost-effective computing resources on demand, allowing users to quickly and easily provision the resources they need to support their applications. With AWS, users can pay only for the services they use, and they can scale their resources up or down as needed to accommodate changes in demand.

AWS offers a wide range of services, including compute services (such as EC2, Elastic Container Service, and Lambda), storage services (such as S3 and Glacier), database services (such as RDS and DynamoDB), and networking services (such as VPC and Route 53). These services can be used together or separately, and they can be combined with other tools and technologies to build complex, high-performance applications and systems.

AWS is widely used by businesses of all sizes, from startups to large enterprises, and it is widely regarded as one of the most comprehensive and versatile cloud computing platforms available.

1659. Mention what the key components of AWS are?

The key components of Amazon Web Services (AWS) are:

Compute: This includes services such as Amazon Elastic Compute Cloud (EC2), Amazon Elastic Container Service (ECS), and AWS Lambda, which provide scalable computing resources to run applications and services.

Storage: This includes services such as Amazon Simple Storage Service (S3) and Amazon Glacier, which provide scalable and durable storage solutions for data and files.

Databases: This includes services such as Amazon Relational Database Service (RDS) and Amazon DynamoDB, which provide managed relational databases and NoSQL databases, respectively.

Networking: This includes services such as Amazon Virtual Private Cloud (VPC) and Amazon Route 53, which provide network infrastructure and domain name system (DNS) services.

Analytics: This includes services such as Amazon Redshift, Amazon Kinesis, and Amazon QuickSight, which provide data warehousing, real-time data processing, and business intelligence capabilities.

Security and Compliance: This includes services such as AWS Identity and Access Management (IAM), Amazon CloudWatch, and Amazon GuardDuty, which provide security and compliance services for applications and data.

Management and Governance: This includes services such as AWS CloudFormation, AWS CloudTrail, and AWS Organizations, which provide management and governance tools for cloud infrastructure and services.

Application Services: This includes services such as Amazon Simple Queue Service (SQS), Amazon Simple Notification Service (SNS), and Amazon Simple Email Service (SES), which provide application services for messaging, notifications, and email.

These components can be used together or separately to build, run, and manage a wide range of applications and services in the cloud.

1660. How can you send a request to Amazon S3?

You can send requests to Amazon S3 (Simple Storage Service) using the Amazon S3 API. The S3 API provides a set of RESTful (Representational State Transfer) web services interfaces that you can use to access S3. You can use these APIs to perform a wide range of operations on S3, such as creating and managing buckets, uploading and downloading objects, and managing permissions.

To send a request to S3, you need to make an HTTP request to the appropriate endpoint for the API operation you want to perform. The endpoint for an S3 API operation is constructed using the following format:

`https://<bucket-name>.s3.<region>.amazonaws.com/<object-key>`

Where <bucket-name> is the name of the S3 bucket, <region> is the AWS region where the bucket is located, and <object-key> is the key for the object you want to access.

You can make requests to the S3 API using any HTTP client that supports making HTTP requests, including command-line tools like curl, programming libraries for your programming language, or HTTP clients built into your operating system.

It's important to note that you will need to authenticate your requests to S3 using AWS Identity and Access Management (IAM) or by providing your AWS access key ID and secret access key in your requests. This is to ensure that only authorized users can access your data in S3.

1661. Explain how the buffer is used in Amazon web services(AWS)?

A buffer is a temporary holding area for data that is being sent from one component to another. In Amazon Web Services (AWS), a buffer is used to help manage the flow of incoming and outgoing data, ensuring that the rate of data flow is consistent and does not overwhelm the receiving component.

One common use of buffers in AWS is to absorb spikes in incoming data traffic and prevent the loss of data due to an overwhelmed system. For example, when receiving data from an Amazon Kinesis stream, a buffer can be used to temporarily store incoming data before it is processed, ensuring that the data processing component can keep up with the incoming data rate.

Another use of buffers in AWS is to manage the flow of outgoing data, ensuring that data is sent at a consistent rate and preventing the overloading of downstream components. For example, when sending data to Amazon S3, a buffer can be used to temporarily store outgoing data before it is transmitted, ensuring that the data transmission rate is consistent and does not overwhelm the S3 service.

Buffers can also be used in AWS to store data that is being processed, such as the data being processed by an AWS Lambda function, so that the processing component can continue to receive new data while it is processing the stored data.

Overall, buffers play a critical role in ensuring the smooth flow of data within AWS, providing a way to manage incoming and outgoing data traffic, absorb spikes in traffic, and prevent data loss due to overwhelmed systems.

1662. What is AWS Lambda?

AWS Lambda is a serverless computing platform provided by Amazon Web Services (AWS). It allows you to run code without provisioning or managing servers. With AWS Lambda, you can build and run applications and services without thinking about infrastructure.

Lambda allows you to upload your code and create a function, which can then be triggered by events such as changes to data in an Amazon S3 bucket or a new item in a DynamoDB table. The service automatically scales your application in response to incoming requests, so you don't have to worry about capacity planning.

Lambda supports multiple programming languages, including Java, Python, Node.js, and more, so you can use the language that's most familiar to you. The service provides a flexible and secure environment for running your code, with built-in security features such as automatic encryption of data at rest and in transit, and the ability to configure resource-based permissions for your functions.

AWS Lambda is often used for building microservices, real-time stream processing, and serverless applications, as well as for running background tasks that are triggered by events. With its automatic scaling, low cost, and ease of use, AWS Lambda has become a popular choice for building scalable and cost-effective applications in the cloud.

1663. Name the AWS service that exists only to redundantly cache data and images?

The AWS service that exists specifically to redundantly cache data and images is Amazon CloudFront.

CloudFront is a content delivery network (CDN) service that speeds up the delivery of your static and dynamic web content, such as HTML, CSS, JavaScript, images, and videos. CloudFront works by caching your content in edge locations around the world, so when a user requests content, it's delivered from the nearest edge location, providing low latency and high transfer speeds.

With CloudFront, you can reduce the load on your origin servers and improve the performance of your website or application by offloading the serving of static content to the edge locations. The service also integrates with other AWS services, such as Amazon S3 and Amazon EC2, so you can easily distribute content from your origin servers to the edge locations.

In summary, CloudFront is designed to redundantly cache data and images, so you can deliver content to your users quickly and efficiently, improving the performance of your website or application and reducing the load on your origin servers.

1664. What are the different types of Load Balancers in AWS services?

There are three types of load balancers in Amazon Web Services (AWS) services:

Classic Load Balancer (CLB): This is the original load balancing service in AWS, designed for simple load balancing of traffic across multiple EC2 instances. CLB supports TCP and SSL protocols, and provides basic features such as round-robin distribution of incoming traffic and health checks.

Application Load Balancer (ALB): This type of load balancer is designed for applications that require advanced routing and traffic management capabilities. ALB supports HTTP and HTTPS protocols, and provides features such as request routing based on the content of the request, support for multiple domains and paths, and the ability to redirect traffic based on rules.

Network Load Balancer (NLB): This type of load balancer is designed for high performance and extreme reliability, making it well-suited for applications that require consistent performance even under heavy traffic loads. NLB supports TCP and UDP protocols, and provides features such as extremely low latency and high throughput, and the ability to handle millions of requests per second.

Each of the three load balancer types has its own strengths and weaknesses, and the best choice for your application will depend on the specific requirements of your application and your network architecture. To help you choose the right load balancer for your application, AWS provides a load balancing comparison tool that allows you to compare the features and benefits of each type of load balancer.

1665. Name some of the DB engines which can be used in AWS RDS?

Amazon Web Services (AWS) Relational Database Service (RDS) supports several database engines, including:

Amazon Aurora: A high-performance relational database engine that is fully compatible with MySQL and PostgreSQL.

MariaDB: An open-source relational database engine that is a fork of the popular MySQL database.

Microsoft SQL Server: A relational database engine developed by Microsoft.

MySQL: An open-source relational database engine that is widely used for web-based applications.

Oracle Database: A commercial relational database engine developed by Oracle Corporation.

PostgreSQL: An open-source relational database engine that is well-suited for applications that require advanced data processing and analysis capabilities.

By using AWS RDS, you can easily set up, operate, and scale a relational database in the cloud, without the need for manual database administration. With RDS, you can choose the database engine that best fits your application's requirements, and enjoy the benefits of automatic database management, high availability, and data backup and recovery.

1666. What are the important features of Amazon cloud search?

Amazon CloudSearch is a managed search and analytics service that makes it easy to set up, manage, and scale a search solution for your website or application. Some of the key features of Amazon CloudSearch include:

Customizable Search Relevance: CloudSearch allows you to fine-tune the relevance of search results by adjusting the weight of specific fields, adjusting the search syntax, and using synonyms to match similar terms.

Multi-Language Support: CloudSearch supports multiple languages, including English, Spanish, French, German, Italian, Portuguese, Chinese, and Japanese, and allows you to provide search results in different languages based on the user's locale.

Automatic Scaling: CloudSearch automatically scales to handle spikes in search traffic, so you don't have to worry about capacity planning or managing infrastructure.

Security: CloudSearch is built on the secure AWS infrastructure and provides various security features, including encryption of data at rest and in transit, as well as integration with AWS Identity and Access Management (IAM) for fine-grained access control.

High Availability: CloudSearch is designed to provide high availability and automatic failover, so you can ensure that your search solution is always available to your users.

Easy Integration: CloudSearch integrates with a variety of AWS services, such as Amazon S3, Amazon DynamoDB, and Amazon CloudFront, so you can easily set up a search solution for your application.

Rich Search Features: CloudSearch provides a variety of advanced search features, such as faceted search, geospatial search, and synonym support, to help you deliver a rich search experience to your users.

These are just a few of the important features of Amazon CloudSearch. By using CloudSearch, you can quickly and easily build a highly scalable and customizable search solution for your application, without having to worry about the underlying infrastructure.

1667. What is the role of AWS CloudTrail?

AWS CloudTrail is a service provided by Amazon Web Services (AWS) that enables governance, compliance, operational auditing, and risk auditing of an AWS account. It provides a history of AWS API calls made on an account and delivers that information in the form of log files to an Amazon S3 bucket.

The main role of AWS CloudTrail is to provide an audit trail of all activity within an AWS account. This includes all API calls made to AWS services, including the identity of the user who made the call, the time the call was made, the source IP address of the request, and other relevant details.

By using CloudTrail, users can perform a variety of tasks, including:

Compliance: CloudTrail logs can help users demonstrate compliance with industry regulations and standards, such as HIPAA and PCI DSS.

Security: CloudTrail can help users identify and respond to security threats by monitoring changes made to their resources and accounts.

Troubleshooting: CloudTrail logs can be used to troubleshoot operational issues by providing detailed information about changes made to resources and accounts.

Governance: CloudTrail logs can be used to monitor and enforce governance policies across an AWS account, ensuring that resources are being used in accordance with established guidelines.

In summary, AWS CloudTrail plays an important role in providing visibility into the activity occurring within an AWS account, which helps users ensure compliance, improve security, troubleshoot issues, and enforce governance policies.

1668. Explain how Amazon SNS can be used for push messaging and event-driven architectures?

Amazon Simple Notification Service (Amazon SNS) is a fully managed pub/sub messaging service that enables you to send messages between decoupled microservices, distributed systems, and serverless applications. With Amazon SNS, you can publish messages to a large number of subscribers, and you don't have to worry about the underlying infrastructure. Additionally, Amazon SNS integrates with other AWS services, such as Amazon SQS, Amazon Lambda, and Amazon S3, to provide a complete solution for building event-driven architectures.

1669. How can you configure an Amazon S3 bucket to serve static content with custom domain names and SSL certificates?

To configure an Amazon S3 bucket to serve static content with custom domain names and SSL certificates, you need to use Amazon CloudFront, which is a content delivery network (CDN) service that integrates with S3. To set up CloudFront, you need to create a distribution, configure it to use your S3 bucket as the origin, and associate it with your custom domain name. To use SSL certificates, you need to use the AWS Certificate Manager to request and manage SSL certificates for your custom domain names. Once you have set up CloudFront and SSL certificates, you can use your custom domain names and SSL certificates to serve static content securely from your S3 bucket.

1670. How can you migrate a complex, multi-tier application to AWS with minimal downtime?

To migrate a complex, multi-tier application to AWS with minimal downtime, you can use a combination of AWS migration tools and techniques. To begin, you need to assess your application architecture and identify the dependencies between your components. Then, you can use AWS Migration Hub to track your migration progress and AWS Application Discovery

Service to discover and document the components of your application. After the assessment, you can use AWS Server Migration Service (SMS) to automate the migration of your virtual machines to Amazon EC2, and use Amazon RDS or Amazon DocumentDB to migrate your databases. To minimize downtime during the migration, you can use Amazon Route 53 and AWS Elastic Load Balancer to route traffic to the appropriate service based on the domain name and URL. Additionally, you can use Amazon CloudFormation or AWS Elastic Beanstalk to automate the deployment and management of your application components.

1671. how Amazon Kinesis can be used to process real-time streaming data?

Amazon Kinesis is a fully managed service that enables you to process real-time streaming data at scale. You can use Kinesis to ingest, buffer, and process real-time data streams, such as log files, application telemetry, and social media feeds. Kinesis allows you to process data in real-time and respond to changing business needs, and enables you to handle large amounts of data, such as terabytes of data per hour. Additionally, Kinesis integrates with other AWS services, such as Amazon S3, Amazon Redshift, and Amazon Machine Learning, to provide a complete solution for processing and analyzing real-time data.

1672. What is the difference between Amazon S3 Intelligent-Tiering and Amazon S3 Standard-IA?

Amazon S3 Intelligent-Tiering is a storage class that automatically moves objects between two access tiers (frequent and infrequent access) based on changing access patterns, while Amazon S3 Standard-IA is a storage class that provides infrequent access to data, with lower storage costs compared to S3 Standard. S3 Intelligent Tiering provides automatic cost optimization and performance optimization, while S3 Standard-IA provides a predictable and lower-cost option for infrequent access to data.

1673. A shipping company has to process requests received from different sources in batches at a periodic intervals in order to receipt of requests.

The company needs a service that can help him manage a queue and process information at an interval of their choice. Which AWS service can they use for managing the requests received?

ANS- SQS

1674. What is SQS?

Amazon Simple Queue Service (SQS) is a fully managed message queuing service that enables you to decouple the components of a cloud application. With SQS, you can transmit

any volume of data, at any level of throughput, without losing messages or requiring other services to be available.

SQS provides a highly scalable, durable, and available hosted queue for storing messages as they travel between applications. Applications can use SQS to transmit messages between components within the same application, or between different applications.

SQS is particularly useful for decoupling applications that need to communicate asynchronously. For example, you could use SQS to transmit messages from a web application to a background job processing service, or from a mobile application to a serverless function that processes images.

SQS supports two types of queues: standard queues and FIFO (first-in, first-out) queues. Standard queues provide a simple way to transmit messages between applications, and they guarantee that messages will be delivered at least once, although they do not guarantee the order of delivery. FIFO queues provide strict ordering and exactly-once processing, which is useful for applications that require a guarantee of message order and delivery.

SQS integrates with other AWS services, including Amazon S3, Amazon DynamoDB, Amazon SNS, and AWS Lambda, so you can use SQS to build complex, scalable, and highly available applications that leverage the full range of AWS services.

1675. The sales team of a company needs to see its sales data in form of visual dashboards.

The company should on rare occasions be able to edit the data if needed as well.

Which AWS service can help them achieve both requirements?

ANS- Amazon Quicksight

1676. What is Amazon Quicksight?

Amazon QuickSight is a fast, cloud-powered business intelligence (BI) service that makes it easy to build visualizations, perform ad hoc analysis, and quickly get business insights from data. QuickSight provides a user-friendly interface that allows you to create and share interactive dashboards, reports, and visualizations without the need for specialized skills in data analysis or visualization.

QuickSight can connect to a wide range of data sources, including Amazon S3, Amazon RDS, Amazon Redshift, and more. You can also use QuickSight with other AWS services, such as Amazon Athena, Amazon EMR, and Amazon Glue, to create a comprehensive data analytics environment.

Once you have connected to your data sources, you can use QuickSight to perform complex data analysis and create interactive dashboards, reports, and visualizations that are easy to understand and share. QuickSight provides a wide range of built-in visualizations and allows you to customize your visualizations using custom calculations, colors, and formatting.

QuickSight also provides advanced features for data analysis, such as machine learning-powered insights, predictive analytics, and real-time data streaming, so you can gain deeper insights into your data and make more informed decisions.

With QuickSight, you can quickly get started with business intelligence, regardless of your data size or complexity, and benefit from the scalability, reliability, and security of the AWS cloud.

1677. A company wants to track the usage of AWS services and resources by its departments in real time? What service can help them monitor the usage?

Ans- Cloudwatch

1678. What is Cloudwatch?

Amazon CloudWatch is a monitoring and management service provided by Amazon Web Services (AWS) for cloud resources and applications. It allows users to collect and track metrics, collect and monitor log files, set alarms, and automatically react to changes in their AWS resources.

CloudWatch collects and processes raw data from AWS resources, such as EC2 instances, RDS databases, and ELB load balancers, as well as custom metrics generated by applications and services running on AWS or on-premises. CloudWatch aggregates and summarizes this data into meaningful metrics and provides a real-time view of the performance and health of applications and resources.

CloudWatch provides several features that enable users to effectively monitor their resources, including:

Metrics: CloudWatch collects metrics from AWS resources and custom applications and services, allowing users to monitor resource utilization, performance, and availability.

Logs: CloudWatch allows users to collect and monitor log files from applications and resources running on AWS, providing real-time insights into application and system behavior.

Alarms: CloudWatch alarms enable users to set thresholds on metrics and receive notifications when those thresholds are breached, allowing them to proactively respond to issues before they become critical.

Dashboards: CloudWatch dashboards provide a customizable view of metrics, logs, and alarms, enabling users to monitor the health and performance of their resources at a glance.

In summary, Amazon CloudWatch is a powerful monitoring and management service that provides real-time visibility into the health and performance of AWS resources and applications, enabling users to monitor, troubleshoot, and optimize their cloud environments.

1679. A company wants to deploy its new application without using a virtual machine how can it do it?

They can use Amazon Fargate which does not need a server for running application

1680. What is Amazon's Frigate?

Amazon Frigate is an open source edge computing platform designed for running machine learning (ML) models and detecting events in video streams in real-time. It is designed to run on low-powered devices and can be used to analyze video feeds from a variety of sources, including security cameras, drones, and other devices.

Frigate uses a combination of object detection, motion detection, and facial recognition algorithms to analyze video streams in real-time. The platform is highly customizable and allows users to define their own detection rules, actions, and notifications. Frigate also supports integration with other AWS services, such as Amazon S3 and Amazon Rekognition.

Some of the key features of Amazon Frigate include:

Real-time video processing: Frigate can process video streams in real-time, enabling users to detect events and take actions immediately.

Customizable detection rules: Frigate allows users to define their own detection rules and customize the actions and notifications triggered by those rules.

Low-power computing: Frigate is designed to run on low-powered devices, making it a cost-effective solution for edge computing applications.

Open source: Frigate is an open source project, which means that it is free to use and can be customized and extended by users.

In summary, Amazon Frigate is an open source edge computing platform designed for running machine learning models and detecting events in video streams in real-time. It is highly customizable and can be used to analyze video feeds from a variety of sources, making it a versatile solution for edge computing applications.

1681. How can we host only one function which can execute task on dedicated event triggers?

AWS lambda can be used to deploy a single function as a program which can be triggered on events.

1682. What is the AWS service used for delivering push notifications to mobile devices?

Ans - SNS

1683. What is the AWS service used for executing, managing, and scaling background jobs?

Ans - Lambda.

1684. What is the AWS service used for storing and processing large amounts of data?

Ans - Redshift

1685. What is the AWS service used for disaster recovery and backing up data?

Ans - Backup

1686. What is the AWS service used for centralized management of AWS resources?

Ans - CloudFormation

1687. How many S3 buckets can be created?

Ans - Up to 100 buckets can be created by default.

1688. What is the maximum limit of elastic IPs anyone can produce?

Ans - A maximum of five elastic IP addresses can be generated per location and AWS account.

1689. Mention 2 major issues that are not supported by the AWS support?

Ans - Code development, Debugging custom software

1690. With specified private IP addresses, can an Amazon Elastic Compute Cloud (EC2) instance be launched? If so, which Amazon service makes it possible?

Ans-Yes. Utilizing VPC makes it possible (Virtual Private Cloud).

1691. Will your standby RDS be launched in the same availability zone as your primary?

Ans - No, standby instances are launched in different availability zones than the primary, resulting in physically separate infrastructures. This is because the entire purpose of standby instances is to prevent infrastructure failure. As a result, if the primary instance fails, the backup instance will assist in recovering all of the data.

1692.. Your business prefers to use its email address and domain to send and receive compliance emails. What service do you recommend to implement it easily and budget-friendly?

Ans-This can be accomplished by using Amazon Simple Email Service (Amazon SES), a cloud-based email-sending service.

1693. Give one instance where you would prefer Provisioned IOPS over Standard RDS storage?

Ans. Provisioned IOPS can be preferred over Standard RDS storage when we have batch-oriented workloads.

1694. Can you change the Private IP Address of an EC2 instance while it is running or in a stopped state?

Ans-No, a Private IP Address of an EC2 instance cannot be changed. When an EC2 instance is launched, a private IP Address is assigned to that instance at the boot time. This private IP Address is attached to the instance for its entire lifetime and can never be changed.

1695. How many Subnets can you have per VPC?

Ans-You can have 200 Subnets per VPC

1696. How can you send a request to Amazon S3?

Ans-Amazon S3 is a REST Service, and you can send a request by using the REST API or the AWS SDK wrapper libraries that wrap the underlying Amazon S3 REST API.

1697.. You plan to design an application by encrypting all the data in an Amazon Redshift cluster. How will you encrypt the data at rest?

Ans - Using the AWS KMS Default Customer master key

1698. An organization decides to build an Amazon Redshift cluster to host sensitive data

in their shared services VPC. What control does the organization implement for networks accessing the cluster?

Ans - Defining a cluster security group for the cluster allowing access from the allowed networks.

1699. An application saves the logs to an S3 bucket. A user needs to keep the logs for one month for troubleshooting purposes and then clear the logs. What action will enable this?

Ans - Configuring lifecycle configuration rules on the S3 bucket.

1700. A website experiences inconsistent traffic, and the database cannot keep up with the write requests during peak traffic times. What AWS Service helps to decouple the web application from the database?

Ans - Amazon SQS

1701. A solution architect is designing a new web application on AWS. To make the application very popular, the architect focuses on software development and new features without managing or provisioning instances. Which solution is best suited for that?

Ans - AWS Lambda and Amazon API Gateway

1702. What are the different types of instances?

Ans - General purpose, Computer Optimized, Storage Optimized

1703. The types of AMI provided by AWS are:

Ans - Instance store-backed and EBS backed

1704. Amazon Web Services supports which Type II Audits?

Ans - SAS70

1705. Amazon S3 is which type of storage service?

Ans - Object

1706. How many buckets can you create in AWS by default?

Ans - 100 buckets can you create in AWS by default.

1707. How would you design a cost-effective solution for processing large amounts of data in real-time using AWS technologies?

Amazon Kinesis Streams: This service can ingest large amounts of streaming data, process it in real time, and then route it to other AWS services for storage and analysis.

Amazon Lambda: This service can be used to process the data from Kinesis Streams and perform any necessary transformations or calculations.

Amazon S3: This service can be used for long-term storage of processed data.

Amazon Glue: This service can be used to create a catalog of the data stored in S3 for easier querying and analysis using tools like Amazon Athena or Amazon Redshift.

1708. How would you design an architecture to support a multi-region, highly available and disaster recovery setup for a web application?

Store data in a multi-region Amazon RDS database, with automatic failover to the secondary region in case of a disaster.

Deploy the web application instances in multiple AWS regions, using Amazon EC2 Auto Scaling groups behind an Amazon ELB Application Load Balancer for automatic scaling and failover.

Use Amazon Route 53 for routing traffic to the nearest healthy region, and for automatic failover to another region in case of an outage.

Store static assets and files in Amazon S3, with versioning enabled and cross-region replication configured to ensure data durability and accessibility in case of a disaster.

1709. How would you design a cost-efficient solution for archiving infrequently accessed data for compliance purposes, taking into consideration retrieval times and cost?

Store the data in Amazon S3 Glacier, which is a low-cost, durable, and secure storage solution specifically designed for archiving data.

Use Amazon S3 Lifecycle policies to automatically transition data to S3 Glacier after a certain period of time.

use Amazon S3 Select to retrieve only the necessary data, rather than restoring the entire archive, which can reduce retrieval times and costs.

Use Amazon S3 Transfer Acceleration for faster data uploads to S3.

Monitor costs using the AWS Cost Explorer service and adjust the solution as needed to maintain cost efficiency.

1710. What would you do if your AWS S3 bucket suddenly became public and sensitive data was being downloaded by unauthorized users?

Immediately stop the download by revoking public access to the S3 bucket.

Investigate the cause of the public access, such as an incorrectly configured bucket policy or an unauthorized IAM user.

Secure the sensitive data by either deleting or properly securing it within the S3 bucket.

Implement measures to prevent future security incidents, such as regularly monitoring access logs and implementing multi-factor authentication.

Document the incident and steps taken for compliance and auditing purposes.

1711. How would you design a solution for migrating an on-premises application to AWS?

Create a Virtual Private Cloud (VPC): A VPC is a logically-isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define.

Control network access with security groups: Security groups act as firewalls that control inbound and outbound network traffic to AWS resources.

Implement network segmentation: Use subnets to segment your network into smaller, isolated sections. This can help limit the blast radius of a security breach.

Use Network Access Control Lists (ACLs): ACLs are another layer of security for controlling network traffic. They provide stateful inspection for inbound and outbound traffic.

Enable monitoring and auditing: Use Amazon CloudWatch and AWS CloudTrail to monitor network activity and log all AWS API calls for auditing purposes.

1712. How would you implement a solution to process big data in real-time using AWS technologies, such as Kinesis, Lambda, and S3?

Store the big data in Amazon S3: You can use S3 to store your big data in a highly durable and scalable manner.

1713. How can you configure an Amazon S3 bucket to serve static content with custom domain names and SSL certificates?

To configure an Amazon S3 bucket to serve static content with custom domain names and SSL certificates, you need to use Amazon CloudFront, which is a content delivery network (CDN) service that integrates with S3. To set up CloudFront, you need to create a distribution, configure it to use your S3 bucket as the origin, and associate it with your custom domain name. To use SSL certificates, you need to use the AWS Certificate Manager to request and manage SSL certificates for your custom domain names. Once you have set up CloudFront and SSL certificates, you can use your custom domain names and SSL certificates to serve static content securely from your S3 bucket.

1714. How can you migrate a complex, multi-tier application to AWS with minimal downtime?

To migrate a complex, multi-tier application to AWS with minimal downtime, you can use a combination of AWS migration tools and techniques. To begin, you need to assess your

application architecture and identify the dependencies between your components. Then, you can use AWS Migration Hub to track your migration progress and AWS Application Discovery Service to discover and document the components of your application. After the assessment, you can use AWS Server Migration Service (SMS) to automate the migration of your virtual machines to Amazon EC2, and use Amazon RDS or Amazon DocumentDB to migrate your databases. To minimize downtime during the migration, you can use Amazon Route 53 and AWS Elastic Load Balancer to route traffic to the appropriate service based on the domain name and URL. Additionally, you can use Amazon CloudFormation or AWS Elastic Beanstalk to automate the deployment and management of your application components.

1715. How Amazon Kinesis can be used to process real-time streaming data?

Amazon Kinesis is a fully managed service that enables you to process real-time streaming data at scale. You can use Kinesis to ingest, buffer, and process real-time data streams, such as log files, application telemetry, and social media feeds. Kinesis allows you to process data in real-time and respond to changing business needs, and enables you to handle large amounts of data, such as terabytes of data per hour. Additionally, Kinesis integrates with other AWS services, such as Amazon S3, Amazon Redshift, and Amazon Machine Learning, to provide a complete solution for processing and analyzing real-time data.

1716. How disaster recovery can be implemented for AWS infrastructure?

There are several strategies for implementing disaster recovery for AWS infrastructure, including using Amazon S3 and Amazon EC2 to create a disaster recovery plan, using Amazon S3 Cross-Region Replication to replicate data between regions, and using Amazon Route 53 to redirect traffic in the event of an outage. Additionally, by implementing disaster recovery best practices, such as regularly backing up your data, testing your disaster recovery plan, and having a clear communication plan in place.

1717. What is the difference between Amazon S3 Intelligent-Tiering and Amazon S3 Standard-IA?

Amazon S3 Intelligent-Tiering is a storage class that automatically moves objects between two access tiers (frequent and infrequent access) based on changing access patterns, while Amazon S3 Standard-IA is a storage class that provides infrequent access to data, with lower storage costs compared to S3 Standard. S3 Intelligent Tiering provides automatic cost optimization and performance optimization, while S3 Standard-IA provides a predictable and lower-cost option for infrequent access to data.

Create an Amazon Kinesis Data Stream: You can use Kinesis Data Streams to collect, process, and analyze real-time, streaming data at a massive scale.

Create an Amazon Lambda function: You can use a Lambda function to process the data from the Kinesis Data Stream and store the processed data in a data store of your choice.

Connect the Kinesis Data Stream to the Lambda function: You can use the Kinesis Data Streams API to connect the Kinesis Data Stream to the Lambda function, so that the function is triggered each time a new batch of data is added to the stream.

Configure the Lambda function to process the data: You can use the code of your choice to process the data from the Kinesis Data Stream. For example, you can use Python or Java to write a function that filters and transforms the data, and stores it in a database, such as Amazon RDS or Amazon DynamoDB.

1718. How would you implement an auto-scaling solution for an application that experiences sudden spikes in traffic?

Create an Amazon CloudWatch Alarm: You can use CloudWatch Alarm to monitor specific metric, such as CPU utilization or network traffic, and trigger an action when the metric crosses a specified threshold.

Create an Amazon Auto Scaling group: You can use an Auto Scaling group to automatically increase or decrease the number of EC2 instances based on the demand.

Create an Amazon EC2 Launch Configuration: You can use a launch configuration to specify the instance type, image ID, key pair, and other parameters for the instances.

Attach the CloudWatch Alarm to the Auto Scaling group: You can attach the CloudWatch Alarm to the Auto Scaling group, so that it triggers a scale-out event when the metric crosses the threshold.

Create a scaling policy: You can create a scaling policy that defines how the Auto Scaling group should scale when a CloudWatch Alarm is triggered. For example, you can specify that the group should add 2 instances when the CPU utilization crosses 70%.

1719. How do you differentiate between terminating and stopping an instance from each other?

When you terminate an instance in the Amazon Elastic Compute Cloud (EC2), the instance is permanently deleted and its associated resources (such as Elastic IP addresses, attached volumes, etc.) are released. You cannot start or recover a terminated instance.

Stopping an instance, on the other hand, is a way to temporarily shut down an instance and preserve its state. When you stop an instance, its associated resources (such as the instance's IP address) remain allocated to your account. You can start a stopped instance at any time, and the instance will return to its previously running state.

In summary, terminating an instance is a permanent action that results in the loss of data and resources, while stopping an instance is a temporary action that allows you to preserve the instance's state and data for later use.

1720. How can you spot the instant difference between on-demand instances and reserved instances?

The primary difference between on-demand instances and reserved instances in Amazon EC2 is the cost and the payment options.

On-demand instances are the traditional way of using EC2, where you pay for each hour that the instance is running. On-demand instances provide the flexibility to scale capacity up or down as needed, without upfront commitments. They are ideal for applications with unpredictable workloads or for development and testing purposes.

Reserved instances, on the other hand, are a cost-effective option for applications with steady-state or predictable usage. With reserved instances, you can make a low, one-time, upfront payment for a portion or all of the instance's capacity, and then pay a lower hourly rate for the instance usage. By making a commitment to use an instance for a longer term, you can save up to 75% compared to on-demand pricing.

To spot the difference between on-demand instances and reserved instances, you can look at the billing options and pricing information in the EC2 console, or in the AWS Cost Explorer tool. On-demand instances will be billed at the on-demand rate, while reserved instances will show the upfront payment and the discounted hourly rate.

1721. Do you think one elastic IP address is enough for all instances of running?

Whether one Elastic IP address is enough for all instances depends on your specific use case and network architecture.

An Elastic IP address is a static IPv4 address that can be assigned to and unassigned from an EC2 instance. By default, every EC2 instance is assigned a dynamic IPv4 address, which can change if the instance is stopped and restarted or if the instance fails and is replaced.

If you need to maintain a persistent public IPv4 address for your instances, you can use an Elastic IP address. For example, if you have a single instance running a web server, one Elastic IP address may be sufficient. However, if you have multiple instances running multiple services, you may need more than one Elastic IP address, so that each instance can be reached through a different public IP address.

In general, it's best to plan your network architecture based on your specific requirements and use cases, and to consider factors such as scalability, security, and manageability when determining the number of Elastic IP addresses that you need.

1722. What are the differences between Amazon RDS, Dynamo DB, and Redshift?

Amazon RDS, DynamoDB, and Redshift are all managed databases provided by Amazon Web Services (AWS), but they have different design goals and capabilities.

Amazon RDS is a managed relational database service that makes it easy to set up, operate, and scale a relational database in the cloud. It supports popular database engines such as Amazon Aurora, MySQL, MariaDB, Microsoft SQL Server, Oracle, and PostgreSQL. RDS provides features such as automatic backups, point-in-time recovery, and read replicas for high availability and performance.

DynamoDB is a managed NoSQL database service that provides fast and predictable performance with seamless scalability. It is designed for high-scale, high-performance, and low-latency applications, such as mobile, gaming, and IoT. DynamoDB supports key-value and document data models, and provides features such as on-demand and provisioned capacity, global tables for multi-region access, and Streams for real-time data processing.

Amazon Redshift is a fast, fully managed, petabyte-scale data warehousing service that makes it simple and cost-effective to analyze big data using SQL and your existing business intelligence tools. Redshift is optimized for complex, large-scale data analytics and business intelligence workloads, and provides features such as columnar storage, data compression, and parallel query processing.

In summary, RDS provides a managed relational database service, DynamoDB provides a managed NoSQL database service, and Redshift provides a managed data warehousing service. The choice between these databases depends on the specific requirements and use cases of your application.

1723. What are the important components of IAM?

IAM (Identity and Access Management) is a service in Amazon Web Services (AWS) that provides centralized control over AWS resources. The following are the important components of IAM:

Users: An IAM user represents a person or service that interacts with AWS resources. You can create and manage IAM users in your AWS account.

Groups: An IAM group is a collection of IAM users, and you can use groups to specify permissions for multiple users at once.

Roles: An IAM role is an AWS identity with specific permissions. You can use roles to delegate access to AWS resources to AWS services, applications, or users.

Policies: An IAM policy is a document that defines one or more permissions. A policy can be attached to a user, group, or role to grant or deny access to AWS resources.

Access Keys: Access keys are a pair of access key ID and secret access key that are used to make programmatic requests to AWS services.

Multi-Factor Authentication (MFA): MFA is a security feature that requires users to provide two forms of authentication, such as a password and a time-based one-time password (TOTP) generated by a virtual MFA device.

Federation: IAM federation enables you to grant permissions to users in another AWS account or an external identity provider, such as Microsoft Active Directory, through a single set of AWS credentials.

These are the important components of IAM that you can use to secure your AWS resources and control access to them. By properly using these components, you can implement a comprehensive security strategy and meet your compliance requirements.

1724. What is Lambda@Edge in AWS?

Lambda@Edge is a feature of Amazon Web Services (AWS) that lets you run serverless code in response to specific CloudFront events, such as a request for a content object or an error response.

With Lambda@Edge, you can write Lambda functions that are triggered by CloudFront events and run close to the user, allowing you to perform real-time processing of content, improve the user experience, and customize the behavior of your application.

Lambda@Edge functions are automatically replicated to multiple AWS locations worldwide, so they run in the location closest to the user and respond to events in real time, with low latency. This allows you to perform tasks such as image resizing, content personalization, security verification, and error handling, directly at the edge.

Lambda@Edge is a powerful and flexible tool for building and deploying serverless applications that deliver high performance and reliability to your users. By using Lambda@Edge in conjunction with CloudFront, you can take advantage of the scale and reliability of the cloud to create fast, dynamic, and highly customized applications.

1725. What are the security mechanisms available in Amazon S3?

Amazon S3 (Simple Storage Service) provides several mechanisms to ensure the security of your data in the cloud:

Access Control Lists (ACLs): ACLs are used to grant or deny access to your S3 objects and buckets. You can use ACLs to specify the AWS accounts or groups that are allowed to access your S3 resources and the type of access they are granted (e.g. read or write).

IAM policies: IAM policies are used to specify the permissions for AWS users and roles. You can use IAM policies to grant or deny access to your S3 resources based on the identity of the user or role.

S3 Bucket Policies: S3 bucket policies are used to specify the permissions for access to a specific S3 bucket. You can use bucket policies to grant or deny access to your S3 resources based on the requester's identity, the request source, and other conditions.

S3 Object Lock: S3 Object Lock is a feature that allows you to protect your S3 objects from being deleted or overwritten for a specified period of time or indefinitely.

S3 Inventory: S3 Inventory provides reports about the objects and metadata stored in your S3 buckets. You can use S3 Inventory to verify the integrity and status of your data over time.

S3 Transfer Acceleration: S3 Transfer Acceleration is a feature that uses Amazon CloudFront's globally distributed edge locations to accelerate transfers of large files over the public internet to S3.

Server-side Encryption: S3 supports server-side encryption, which encrypts your data at rest in the bucket. You can choose to use Amazon S3 managed encryption keys, or bring your own encryption keys for even more security.

Client-side Encryption: S3 also supports client-side encryption, which allows you to encrypt your data before uploading it to the bucket. This provides an additional layer of security to ensure that your data is protected at all times.

These are the main security mechanisms available in Amazon S3 to help you secure your data and protect it from unauthorized access or tampering. By using these mechanisms in combination, you can implement a comprehensive security strategy for your S3 data.

1727. If an organization splits its workload between public cloud and private servers, what do you call this approach?

The approach of splitting workload between public cloud and private servers is often referred to as a "hybrid cloud" approach.

A hybrid cloud is a type of cloud computing environment that combines elements of public and private clouds to deliver the benefits of both. In a hybrid cloud, some workloads are run on public cloud infrastructure for its scalability, cost-effectiveness, and ease of use, while others are run on private servers for security and compliance reasons.

The hybrid cloud approach allows organizations to take advantage of the benefits of both public and private clouds, while minimizing the risks and limitations associated with each. For example, critical applications and sensitive data can be kept in the private cloud for enhanced security

and control, while less critical applications can be run on the public cloud for cost savings and scalability.

The hybrid cloud approach is becoming increasingly popular among organizations that require the flexibility, scalability, and security of both public and private clouds to meet their evolving business needs.

1728. In simple terms, explain the difference between vertical and horizontal scaling.

Vertical scaling and horizontal scaling refer to two different approaches for increasing the capacity of a system.

Vertical scaling, also known as "scaling up", involves adding more resources to a single server, such as adding more memory, disk space, or processing power. This increases the capacity of the system and allows it to handle more workload.

Horizontal scaling, also known as "scaling out", involves adding more servers to a system, distributing the workload across multiple servers. This increases the capacity of the system by adding more resources and also enhances its fault tolerance and scalability.

In summary, vertical scaling adds more resources to a single server to increase its capacity, while horizontal scaling adds more servers to the system to increase its capacity and enhance its scalability and fault tolerance. Both approaches have their own advantages and disadvantages, and the choice between them depends on the specific requirements of the system and the workload it needs to handle.

1729. What is an Elastic IP Address?

An Elastic IP address is a static IPv4 address that is assigned to your AWS account. It is designed to allow you to associate the same IP address with an EC2 instance, even if the instance is stopped or terminated and then restarted. This makes it easier to manage IP-based access controls and maintain connectivity to the Internet even if you need to restart your instances.

With an Elastic IP address, you can mask the failure of an instance or software by quickly remapping the address to another instance in your account. This eliminates the need to change the DNS record for your domain or update your application's configuration files.

Elastic IP addresses are associated with your AWS account, not a specific instance, so you can reuse the same IP address for different instances in the same region. They are also associated with your AWS account, not a specific region, so you can move them between regions if needed.

Overall, Elastic IP addresses provide a convenient and flexible way to manage IP addresses in your AWS environment and help ensure high availability and resiliency for your applications.

1730. Can you explain the difference between Amazon S3 and Amazon EC2?

Amazon S3 and Amazon EC2 are both cloud computing services provided by Amazon Web Services (AWS), but they serve different purposes and are designed for different use cases.

Amazon S3 is an object storage service that provides a scalable, high-speed, and low-cost data storage infrastructure. It allows you to store and retrieve any amount of data, at any time, from anywhere on the web. S3 is ideal for storing and serving large amounts of unstructured data, such as images, videos, backups, and big data analytics results.

Amazon EC2, on the other hand, is a compute service that provides scalable computing capacity in the cloud. It allows you to launch virtual machines (known as EC2 instances) that can run your applications, host your websites, and store your data. EC2 instances can be resized and scaled up or down as needed, providing you with the ability to handle sudden spikes in traffic or changing processing requirements.

In summary, Amazon S3 is an object storage service designed for storing and serving large amounts of unstructured data, while Amazon EC2 is a compute service designed for running and scaling applications and virtual machines in the cloud.

1731. How do you handle disaster recovery in AWS?

Disaster recovery in AWS can be handled through a combination of different services and best practices, including:

Backup and Recovery: Regularly backing up your data and applications to Amazon S3, Amazon Glacier, or other AWS storage services is a critical component of disaster recovery planning. This helps ensure that your data is protected and can be quickly restored in the event of a disaster.

Amazon EC2 Auto Scaling and Amazon Elastic Load Balancer: By using Amazon EC2 Auto Scaling and Amazon Elastic Load Balancer, you can automatically provision and scale EC2 instances to handle changing traffic patterns, which can help ensure high availability and minimize the impact of a disaster.

Amazon Route 53: Amazon Route 53 is a scalable and highly available Domain Name System (DNS) service that can help route traffic to healthy resources in the event of a disaster.

Amazon VPC: Amazon Virtual Private Cloud (VPC) allows you to launch EC2 instances and other AWS resources into a virtual network that is isolated from the public Internet. By using

VPC, you can increase the security and control of your AWS resources and help ensure that they are protected in the event of a disaster.

Amazon CloudWatch and Amazon CloudTrail: Amazon CloudWatch and Amazon CloudTrail provide monitoring and logging capabilities that can help you detect and respond to potential disasters in real-time.

Multi-Region Deployment: Deploying your resources across multiple AWS regions can help ensure that your data and applications are protected and available even in the event of a regional disaster.

Business Continuity Planning: In addition to technical measures, a comprehensive disaster recovery plan should also include business continuity planning, which involves assessing the impact of a disaster on your business operations and defining processes for restoring business operations as quickly as possible.

1732. How would you optimize an Amazon RDS database for improved performance?

There are several steps you can take to optimize the performance of an Amazon RDS database:

Proper instance sizing: Choose an appropriate instance type based on your database's workload and ensure that you have enough CPU, memory, and I/O capacity to handle your database's demands.

Read replicas: Consider using read replicas to offload read-intensive workloads from the primary database instance. Read replicas are copies of the primary database instance that are automatically updated with changes to the primary database.

Indexing: Make sure you have appropriate indexes in place to speed up query performance. Over-indexing can lead to decreased write performance, so it's important to strike a balance.

Caching: Use caching to store frequently accessed data in memory, reducing the need to retrieve it from disk. Amazon RDS supports caching using the Memcached or Redis engines.

Query optimization: Monitor your database's slow query log and optimize queries that are taking too long to execute.

Database tuning: Regularly monitor key database performance metrics such as CPU utilization, memory usage, and I/O performance, and make adjustments as needed.

Maintenance: Perform regular database maintenance tasks, such as vacuuming and optimizing tables, to help ensure that your database is running at peak performance.

Storage optimization: Consider using Provisioned IOPS storage if your database requires high I/O performance.

These are just a few of the many steps you can take to optimize an Amazon RDS database for improved performance. The specific optimizations you'll need to make will depend on the specifics of your database and its workload.

1733. What is the difference between an Amazon VPC and a public subnet?

An Amazon Virtual Private Cloud (Amazon VPC) is a virtual network dedicated to your AWS account. It enables you to launch AWS resources into a virtual network that you've defined. With Amazon VPC, you can define subnets, route tables, network gateways, security settings, and more.

A public subnet is a subnet within an Amazon VPC that has direct access to the internet. This is achieved by associating the public subnet with a route table that has a route to an Internet Gateway.

In contrast, a private subnet is a subnet within an Amazon VPC that does not have direct access to the internet. Instead, you can use a Network Address Translation (NAT) gateway or a VPN connection to allow communication between instances in a private subnet and the internet.

In summary, an Amazon VPC is a virtual network that you can use to launch AWS resources, while a public subnet is a type of subnet within an Amazon VPC that has direct access to the internet.

1734. Can you explain the security measures available in Amazon S3?

Amazon S3 provides several security measures to help protect your data:

Access Control: You can use Amazon S3 Access Control Lists (ACLs) and bucket policies to control access to your S3 objects and buckets.

Encryption: You can encrypt data in transit to and from Amazon S3 using SSL/TLS, and you can encrypt data at rest using server-side encryption with Amazon S3-managed keys (SSE-S3) or customer-managed keys (SSE-C).

Versioning: You can use versioning to preserve, retrieve, and restore versions of objects in your Amazon S3 bucket. This helps you protect against both unintended user actions and data loss due to events such as infrastructure failures.

Authentication: You can use AWS Identity and Access Management (IAM) to control access to your Amazon S3 resources and AWS Multi-Factor Authentication (MFA) to add an extra layer of security for critical actions.

Logging: Amazon S3 server access logs can provide detailed records for the requests that are made to a bucket. These logs can be useful for security and access auditing.

Amazon S3 Transfer Acceleration: This feature can speed up large data transfers over the public internet to Amazon S3 by using Amazon CloudFront's globally distributed edge locations.

S3 Inventory: S3 Inventory provides reports about your objects and their metadata, making it easier to manage your data in Amazon S3. With S3 Inventory, you can list objects and metadata, and also analyze object access patterns.

These security measures can help protect your data stored in Amazon S3 and ensure that it is only accessible by authorized users. It's important to understand how to use these security features to meet your specific security requirements.

1735. How do you migrate an on-premise database to Amazon RDS?

There are several steps to migrating an on-premise database to Amazon Relational Database Service (Amazon RDS):

Prepare the source database: Make sure your on-premise database is up-to-date, backed up, and has no corruption issues.

Set up Amazon RDS: Create an Amazon RDS instance and choose the appropriate database engine (such as MySQL, MariaDB, Microsoft SQL Server, Oracle, or PostgreSQL).

Configure network and security: Set up Amazon Virtual Private Cloud (Amazon VPC) and configure the security group rules to allow access to the Amazon RDS instance from your on-premise environment.

Migrate data: There are several methods to migrate data to Amazon RDS, including using AWS Database Migration Service (AWS DMS), using a database-specific migration tool (such as mysqldump for MySQL), or using a combination of the two.

Test the migration: After the data has been migrated, test the Amazon RDS instance to ensure that everything is working as expected.

Cut over to Amazon RDS: After testing is complete, switch over to the Amazon RDS instance as the primary source for your database.

It is important to plan the migration carefully and thoroughly test the Amazon RDS instance before cutting over to it as the primary database. This will help ensure that the migration is smooth and that there are no unexpected issues or downtime.

1736. How would you troubleshoot a high network latency issue in Amazon EC2?

High network latency in Amazon Elastic Compute Cloud (Amazon EC2) can impact the performance of your applications and services. Here are some steps to troubleshoot high network latency in Amazon EC2:

Monitor EC2 instances: Use Amazon CloudWatch to monitor the network performance of your EC2 instances. Look for spikes in network latency and track the source of the issue.

Check EC2 instance types: Ensure that the EC2 instance type you are using is appropriate for your workload and network requirements. Some instance types may have limited network performance, so you may need to switch to a different instance type.

Check network configurations: Ensure that the network configuration for your EC2 instances is set up correctly and is optimized for performance. For example, you can increase the number of network interfaces, configure multiple IP addresses, or use Amazon Elastic Network Interfaces (ENIs).

Check the VPC and subnet configurations: Ensure that your VPC and subnet configurations are optimized for network performance. For example, you can increase the size of your subnets, or use Amazon VPC peering to reduce network latency.

Monitor network traffic: Use Amazon CloudWatch to monitor network traffic and identify any bottlenecks. You can also use Amazon VPC Flow Logs to view network traffic at the subnet or VPC level.

Check for network congestion: Ensure that there is no network congestion in your VPC or subnets. Network congestion can cause high latency and reduced performance.

Check for packet loss: Check for any packet loss in your network. Packet loss can cause high latency and reduced performance.

By following these steps, you can identify the cause of high network latency in your Amazon EC2 environment and take steps to resolve the issue. It's important to monitor your network performance regularly to ensure that it remains optimal and that any issues are quickly identified and resolved.

1737. Can you explain the benefits and limitations of Amazon Lambda?

Amazon Lambda is a serverless computing service that allows you to run code without provisioning or managing servers. It is a highly scalable and cost-effective platform that can be

used for a variety of use cases, including back-end web applications, data processing, and mobile app backends.

Benefits of Amazon Lambda:

No server management: With Lambda, there is no need to manage servers, as all server management is handled by Amazon Web Services (AWS). This can save you time and resources and allow you to focus on writing and deploying your code.

Automatic scaling: Lambda automatically scales your applications in response to incoming traffic, ensuring that your applications have the capacity to handle any load.

Cost-effective: With its pay-per-use pricing model, Lambda can be more cost-effective than traditional server-based computing, as you only pay for the compute time that you consume.

Easy integration: Lambda can easily integrate with other AWS services, such as Amazon S3, Amazon DynamoDB, and Amazon Kinesis, making it a flexible and versatile platform.

High availability: Lambda is designed to be highly available, with multiple instances of your code running in different Availability Zones.

Limitations of Amazon Lambda:

Cold starts: Cold starts can occur when a new instance of your Lambda function is created after a period of inactivity. This can result in slower function execution times, as the new instance needs to be created and initialized.

Resource limitations: Lambda functions have resource limitations, including a maximum execution time and memory allocation. This can limit the types of applications that can be built on Lambda.

Network latency: Network latency can be an issue for applications that require low latency, as Lambda functions run in a remote environment.

Limited local storage: Lambda functions have limited local storage, so it's important to store any data that needs to persist in an external storage service, such as Amazon S3.

Despite its limitations, Amazon Lambda remains a popular and powerful platform for building serverless applications. Its benefits, including automatic scaling, cost-effectiveness, and easy integration, make it a valuable tool for many organizations.

1738. How do you handle storage management in Amazon S3?

Storage management in Amazon S3 involves the creation, organization, and maintenance of S3

buckets and objects to meet the needs of your application or organization. Here are some best practices for handling storage management in Amazon S3:

Use versioning: Versioning allows you to keep multiple versions of an object in the same S3 bucket. This can be useful for retaining previous versions of files, or for rolling back changes if necessary.

Use lifecycle policies: Lifecycle policies can be used to automatically transition objects to different storage classes over time, such as moving objects from the S3 Standard storage class to the S3 Infrequent Access storage class. This can help you reduce costs and improve storage efficiency.

Use access control lists (ACLs) and bucket policies: ACLs and bucket policies can be used to control access to S3 buckets and objects, allowing you to restrict access to specific users or groups.

Use Amazon S3 Transfer Acceleration: Transfer Acceleration can be used to speed up transfers of large objects over long distances by utilizing Amazon CloudFront's globally distributed edge locations.

Use Amazon S3 Inventory: S3 Inventory provides a report of your S3 objects and metadata, making it easier to manage your storage and identify objects that can be archived or deleted.

Use Amazon S3 object tagging: S3 object tagging allows you to add metadata to your objects, making it easier to categorize and manage your data.

Use Amazon S3 object storage classes: S3 provides several object storage classes, including S3 Standard, S3 Infrequent Access, and S3 One Zone. Choose the storage class that is appropriate for the access patterns of your data.

By following these best practices, you can effectively manage your storage in Amazon S3, ensuring that your data is organized, secure, and cost-effective.

1739. Can you explain the difference between Amazon Route 53 and Amazon CloudFront?

Amazon Route 53 and Amazon CloudFront are two different services offered by Amazon Web Services (AWS).

Amazon Route 53 is a highly available and scalable Domain Name System (DNS) web service. It translates domain names into IP addresses and helps users access web applications and services. It provides a variety of routing types, including simple routing, latency-based routing, geolocation-based routing, and weighted routing, to give you fine-grained control over how your users are directed to your application.

Amazon CloudFront is a content delivery network (CDN) service that speeds up the delivery of static and dynamic web content, such as HTML pages, images, videos, and APIs. CloudFront works by caching content at edge locations around the world, so that when a user requests content, they are served it from the nearest edge location, reducing latency and improving performance. CloudFront integrates with a variety of AWS services, including Amazon S3, Amazon EC2, and Amazon Route 53, to make it easy to distribute content globally.

In summary, Amazon Route 53 provides a highly available and scalable DNS service, while Amazon CloudFront provides a fast and scalable content delivery network service. They complement each other in that Route 53 helps direct users to your application, while CloudFront helps speed up the delivery of your application's content.

1740. Suppose you have an application that serves content over a network to millions of users and is running on an EC2 instance. The application has become increasingly popular, and the EC2 instance is now overutilized. How would you design a scalable solution to handle this increased demand?

To handle the increased demand and scale the application, you can consider using an Elastic Load Balancer (ELB) to distribute traffic across multiple EC2 instances. You can also use auto scaling to automatically adjust the number of instances in response to changes in demand. Additionally, you can use Amazon CloudFront to cache content closer to users and reduce the load on the application. You can also consider using Amazon RDS to offload the database load from the EC2 instances.

1741. How would you design a highly available and disaster-recovery solution for a multi-tier web application on AWS?

To design a highly available and disaster-recovery solution for a multi-tier web application on AWS, you can use multiple availability zones (AZs) to ensure that your application is available even if one AZ becomes unavailable. You can use ELB to distribute traffic across instances in different AZs and use auto scaling to maintain the desired capacity. You can use Amazon RDS to provide a highly available and scalable database solution. You can also implement backup and restore procedures for your data, and use Amazon S3 for storing backups and other important data.

1742. What strategies would you use to secure sensitive data stored in AWS?

To secure sensitive data stored in AWS, you can use encryption at rest and in transit. You can use AWS Key Management Service (KMS) to manage encryption keys, and use AWS Certificate Manager to generate and manage SSL/TLS certificates. You can also implement network security measures such as security groups and network ACLs, and use AWS Identity and Access Management (IAM) to control access to your resources.

1743. How would you implement a cost-effective disaster recovery solution for a large number of EC2 instances in different regions?

To implement a cost-effective disaster recovery solution for a large number of EC2 instances in different regions, you can use AWS Disaster Recovery (DR) services such as AWS Backup, AWS CloudFormation, and AWS Site-to-Site VPN. You can use AWS Backup to automate backups and restore processes, and AWS CloudFormation to automate the creation and configuration of your resources. You can also use AWS Site-to-Site VPN to connect your on-premises network to your resources in different regions.

1744. Can you explain the difference between EC2 placement groups and EC2 auto-scaling groups and when you would use each?

EC2 placement groups are used to place instances in a low-latency, high-bandwidth network, while EC2 auto-scaling groups are used to automatically adjust the number of instances based on demand. Placement groups are typically used for applications that require high network performance, such as HPC and Big Data applications. Auto-scaling groups are typically used for web applications and other applications that have variable traffic patterns.

1745. How would you design a scalable and cost-effective storage solution for a big data analytics application on AWS?

To design a scalable and cost-effective storage solution for a big data analytics application on AWS, you can use Amazon S3 for storing data and use Amazon EMR for processing and analyzing data. You can use Amazon Glue for data transformation and data cataloging. You can also use Amazon Redshift for data warehousing, and Amazon DynamoDB or Amazon Aurora for NoSQL and relational database needs

1746. Can you explain how to optimize network performance for an application running on multiple EC2 instances in different availability zones?

To optimize network performance for an application running on multiple EC2 instances in different availability zones, you can consider using Amazon VPC peering to connect the instances in different AZs. You can also use Amazon CloudFront to cache content closer to users and reduce the load on the instances. You can use Amazon Route 53 to load balance traffic across instances in different AZs.

1747. How would you set up a multi-tier application in AWS using EC2, RDS, and S3, and ensure high availability and reliability?

To set up a multi-tier application in AWS using EC2, RDS, and S3 and ensure high availability and reliability, you can use multiple availability zones to ensure that your application is available even if one AZ becomes unavailable. You can use ELB to distribute traffic across instances in different AZs, and use auto scaling to maintain the desired capacity. You can use Amazon RDS to provide a highly available and scalable database solution, and use Amazon S3 for storing and serving static assets.

1748. Can you describe how you would implement a serverless architecture for a web application on AWS using Lambda, API Gateway, and DynamoDB?

To implement a serverless architecture for a web application on AWS using Lambda, API Gateway, and DynamoDB, you can use Lambda to handle the application logic, API Gateway

1749. How would you design a solution to monitor and alert on critical issues in an AWS infrastructure?

To monitor and alert on critical issues in an AWS infrastructure, you could use a combination of AWS CloudWatch and AWS CloudTrail.

AWS CloudWatch can be used to collect and track metrics, collect and monitor log files, and set alarms. You can use CloudWatch to set up custom metrics to monitor the performance of specific applications or resources. You can also use CloudWatch Logs to monitor logs for specific keywords or phrases, and set up alarms to notify you if those keywords are found.

AWS CloudTrail can be used to record all API calls made to AWS services, and store the logs in an S3 bucket. This provides a record of all changes made to your infrastructure, which can be used for security analysis, resource change tracking, and compliance auditing.

To set up monitoring and alerting for critical issues in an AWS infrastructure, you can follow these steps:

Set up CloudWatch Metrics and Alarms: Use CloudWatch to collect and track metrics for your resources, and set up alarms to notify you when specific thresholds are reached.

Monitor logs using CloudWatch Logs: Use CloudWatch Logs to monitor logs for specific keywords or phrases, and set up alarms to notify you when those keywords are found.

Use CloudTrail to monitor API calls: Use CloudTrail to record all API calls made to AWS services, and set up alerts to notify you of any suspicious or unauthorized activity.

Set up Notifications: Use Amazon SNS or other notification services to send alerts and notifications to your team or stakeholders when critical issues are detected.

Regularly review and refine your monitoring and alerting strategy: Continuously monitor and review your alerts and metrics, and adjust your strategy as needed to ensure that your infrastructure is running smoothly and efficiently.

1750. What is the difference between an Amazon ELB and an Amazon ALB?

The main difference between Amazon ELB (Elastic Load Balancer) and Amazon ALB (Application Load Balancer) is that ELB is a classic load balancer that works at Layer 4 (TCP/UDP), while ALB is an application load balancer that works at Layer 7 (HTTP/HTTPS). This means that ALB can route traffic based on content, whereas ELB cannot.

1751. What is Amazon SNS and how do you use it?

Amazon SNS (Simple Notification Service) is a fully managed messaging service that enables you to send notifications to a variety of endpoints, including email, SMS, mobile push, and Amazon SQS (Simple Queue Service). You can use SNS to send notifications when certain events occur in your AWS environment, such as a new instance being launched or a resource becoming unavailable.

1752. What is the difference between an Amazon VPC and an Amazon VPN?

Amazon VPC (Virtual Private Cloud) is a logically isolated virtual network that you can use to launch AWS resources. Amazon VPN (Virtual Private Network) is a way to securely connect your on-premises network to your VPC. VPC is used to create a private, isolated network in the cloud, while VPN is used to securely connect that network to your on-premises network.

1753. How do you monitor AWS costs and optimize resources for cost savings?

To monitor AWS costs and optimize resources for cost savings, you can use AWS Cost Explorer and AWS Trusted Advisor. Cost Explorer allows you to visualize and understand your AWS costs and usage, while Trusted Advisor provides recommendations to optimize your AWS resources for cost savings.

1754. What are the different types of load balancers in AWS and when would you use each type?

There are three types of load balancers in AWS: Application Load Balancers (ALBs), Network Load Balancers (NLBs), and Classic Load Balancers (CLBs). ALBs are used for HTTP/HTTPS traffic, NLBs are used for TCP/UDP traffic, and CLBs are used for both HTTP/HTTPS and TCP/UDP traffic. You would use each type of load balancer based on the type of traffic you need to balance and the specific requirements of your application.

1755. What is AWS CloudFormation and how do you use it?

AWS CloudFormation is a service that enables you to create and manage AWS resources using templates. You can use CloudFormation to automate the creation and configuration of your AWS resources and ensure that they are created consistently and according to best practices.

1756. What is AWS Identity and Access Management (IAM) and how do you use it?

AWS Identity and Access Management (IAM) is a service that enables you to manage access to AWS resources. You can use IAM to create and manage users, groups, and roles, and to control access to AWS resources using policies. IAM enables you to grant permissions to users and applications to access your AWS resources securely.

1757. How do you ensure high availability and fault tolerance in your AWS architecture?

To ensure high availability and fault tolerance in your AWS architecture, you can use strategies such as deploying resources across multiple availability zones (AZs), using auto scaling to ensure that your application can handle increased load, and using AWS managed services that are designed for high availability, such as Amazon RDS (Relational Database Service).

1758. What are some of the key considerations for designing a scalable and highly available architecture on AWS?

When designing a scalable and highly available architecture on AWS, some key considerations include designing for elasticity and scalability, using multiple availability zones to increase availability, using managed services to reduce the operational burden, using autoscaling to handle increased load, and designing for failure by implementing redundancy and fault tolerance.

Chapter 17 - Hadoop

1. What is Hadoop used for?

Apache Hadoop is an open-source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

2. What are the differences between RDBMS and Hadoop?

RDBMS vs Hadoop

Name	RDBMS	Hadoop
Data volume	RDBMS cannot store and process a large amount of data	Hadoop works better for large amounts of data. It can easily store and process a large amount of data compared to RDBMS.
Throughput	RDBMS fails to achieve a high Throughput	Hadoop achieves high Throughput
Data variety	Schema of the data is known in RDBMS and it always depends on the structured data.	It stores any kind of data. Whether it could be structured, unstructured, or semi-structured.
Data processing	RDBMS supports OLTP(Online Transactional Processing)	Hadoop supports OLAP(Online Analytical Processing)
Read/Write Speed	Reads are fast in RDBMS because the schema of the data is already known.	Writes are fast in Hadoop because no schema validation happens during HDFS write.
Schema on reading Vs Write	RDBMS follows schema on write policy	Hadoop follows the schema on reading policy
Cost	RDBMS is a licensed software	Hadoop is a free and open-source framework

3. What is Hadoop and list its components?

Hadoop is an open-source framework used for storing large data sets and runs applications across clusters of commodity hardware.

It offers extensive storage for any type of data and can handle endless parallel tasks.

3. What are the core components of Hadoop?

Core components of Hadoop:

Storage unit– HDFS (DataNode, NameNode)

Processing framework– YARN (NodeManager, ResourceManager)

4. What is YARN and explain its components?

Yet Another Resource Negotiator (YARN) is one of the core components of Hadoop and is responsible for managing resources for the various applications operating in a Hadoop cluster, and also schedules tasks on different cluster nodes.

5. What are the components of YARN?

YARN components:

Resource Manager - It runs on a master daemon and controls the resource allocation in the cluster.

Node Manager - It runs on a slave daemon and is responsible for the execution of tasks for each single Data Node.

Application Master - It maintains the user job lifecycle and resource requirements of individual applications. It operates along with the Node Manager and controls the execution of tasks.

Container - It is a combination of resources such as Network, HDD, RAM, CPU, etc., on a single node.

6. What are the features of HDFS?

Supports storage of very large datasets

Write once read many access model

Streaming data access

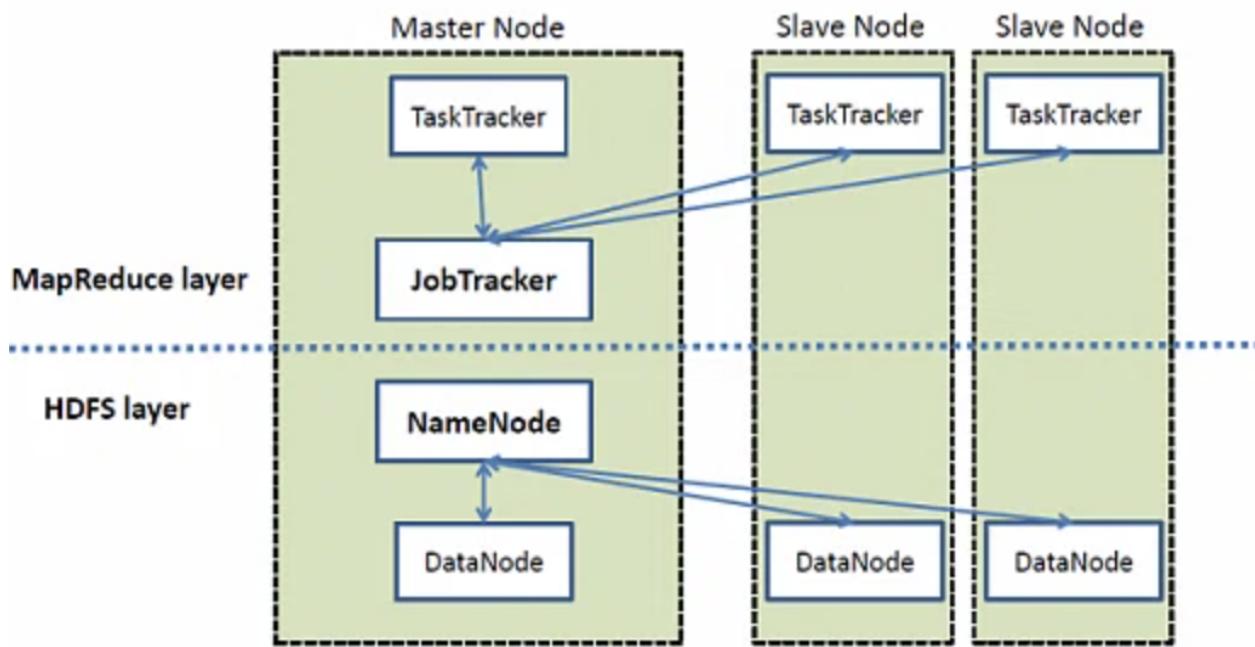
Replication using commodity hardware

HDFS is highly Fault Tolerant

Distributed Storage

7. Give the high-level architecture of Hadoop.

High Level Architecture of Hadoop



7. What is Namenode?

NameNode is the master service that hosts metadata in disk and RAM. It holds information about the various DataNodes, their location, the size of each block, etc.

8. What is Datanode?

DataNodes hold the actual data blocks and send block reports to the NameNode every 10 seconds. The DataNode stores and retrieves the blocks when the NameNode asks. It reads and writes the client's request and performs block creation, deletion, and replication based on instructions from the NameNode.

Data that is written to HDFS is split into blocks, depending on its size. The blocks are randomly distributed across the nodes. With the auto-replication feature, these blocks are auto-replicated across multiple machines with the condition that no two identical blocks can sit on the same machine.

As soon as the cluster comes up, the DataNodes start sending their heartbeats to the NameNodes every three seconds. The NameNode stores this information; in other words, it starts building metadata in RAM, which contains information about the DataNodes available in

the beginning. This metadata is maintained in RAM, as well as in the disk.

9. What are the different Hadoop configuration files?

The different Hadoop configuration files include:

hadoop-env.sh

mapred-site.xml

core-site.xml

yarn-site.xml

hdfs-site.xml

Master and Slaves

10. What are the three modes in which Hadoop can run?

The three modes in which Hadoop can run are :

Standalone mode: This is the default mode. It uses the local FileSystem and a single Java process to run the Hadoop services.

Pseudo-distributed mode: This uses a single-node Hadoop deployment to execute all Hadoop services.

Fully-distributed mode: This uses separate nodes to run Hadoop master and slave services.

11. What are the differences between regular FileSystem and HDFS?

Regular FileSystem: In regular FileSystem, data is maintained in a single system. If the machine crashes, data recovery is challenging due to low fault tolerance. Seek time is more and hence it takes more time to process the data.

HDFS: Data is distributed and maintained on multiple systems. If a DataNode crashes, data can still be recovered from other nodes in the cluster. Time taken to read data is comparatively more, as there is local data read to the disc and coordination of data from multiple systems.

12. Why is HDFS fault-tolerant?

HDFS is fault-tolerant because it replicates data on different DataNodes. By default, a block of data is replicated on three DataNodes. The data blocks are stored in different DataNodes. If one node crashes, the data can still be retrieved from other DataNodes.

13. If you have an input file of 350 MB, how many input splits would HDFS create and what would be the size of each input split?

By default, each block in HDFS is divided into 128 MB. The size of all the blocks, except the last block, will be 128 MB. For an input file of 350 MB, there are three input splits in total. The size of each split is 128 MB, 128MB, and 94 MB.

14. What would happen if you store too many small files in a cluster on HDFS?

Storing several small files on HDFS generates a lot of metadata files. To store these metadata in the RAM is a challenge as each file, block, or directory takes 150 bytes for metadata. Thus, the cumulative size of all the metadata will be too large.

15. Who takes care of replication consistency in a Hadoop cluster and what do under/over replicated blocks mean?

In a cluster, it is always the NameNode that takes care of the replication consistency. The fsck command provides information regarding the over and under-replicated block.

16. What are under-replicated blocks?

Under-replicated blocks are the blocks that do not meet their target replication for the files they belong to. HDFS will automatically create new replicas of under-replicated blocks until they meet the target replication.

Consider a cluster with three nodes and replication set to three. At any point, if one of the NameNodes crashes, the blocks would be under-replicated. It means that there was a replication factor set, but there are not enough replicas as per the replication factor. If the NameNode does not get information about the replicas, it will wait for a limited amount of time and then start the re-replication of missing blocks from the available nodes.

17. What are over-replicated blocks?

Over-replicated blocks are the blocks that exceed their target replication for the files they belong to. Usually, over-replication is not a problem, and HDFS will automatically delete excess replicas.

Consider a case of three nodes running with the replication of three, and one of the nodes goes down due to a network failure. Within a few minutes, the NameNode re-replicates the data, and then the failed node is back with its set of blocks. This is an over-replication situation, and the NameNode will delete a set of blocks from one of the nodes.

18. What is recordreader?

Recordreader communicates with the InputSplit and converts the data into key-value pairs suitable for the mapper to read.

19. What is Combiner?

This is an optional phase; it is like a mini reducer. The combiner receives data from the map tasks, works on it, and then passes its output to the reducer phase.

20. What is a partitioner?

The partitioner decides how many reduced tasks would be used to summarize the data. It also confirms how outputs from combiners are sent to the reducer, and controls the partitioning of keys of the intermediate map outputs.

21. Why is MapReduce slower in processing data in comparison to other processing frameworks?

This is quite a common question in Hadoop interviews; let us understand why MapReduce is slower in comparison to the other processing frameworks:

MapReduce is slower because:

It is batch-oriented when it comes to processing data. Here, no matter what, you would have to provide the mapper and reducer functions to work on data.

During processing, whenever the mapper function delivers an output, it will be written to HDFS and the underlying disks. This data will be shuffled and sorted, and then be picked up for the reducing phase. The entire process of writing data to HDFS and retrieving it from HDFS makes MapReduce a lengthier process.

In addition to the above reasons, MapReduce also uses Java language, which is difficult to program as it has multiple lines of code.

22. Is it possible to change the number of mappers to be created in a MapReduce job?

By default, you cannot change the number of mappers, because it is equal to the number of input splits. However, there are different ways in which you can either set a property or customize the code to change the number of mappers.

For example, if you have a 1GB file that is split into eight blocks (of 128MB each), there will only be only eight mappers running on the cluster. However, there are different ways in which you can either set a property or customize the code to change the number of mappers.

23. Explain spilling in Map Reduce

Spilling is a process of copying the data from the memory buffer to disk when the buffer usage reaches a specific threshold size. This happens when there is not enough memory to fit all of the mapper output. By default, a background thread starts spilling the content from memory to disk after 80 percent of the buffer size is filled.

For a 100 MB size buffer, the spilling will start after the content of the buffer reaches a size of 80 MB.

24. Can we write the output of MapReduce in different formats?

Yes. Hadoop supports various input and output File formats, such as:

TextOutputFormat - This is the default output format and it writes records as lines of text.

SequenceFileOutputFormat - This is used to write sequence files when the output files need to be fed into another MapReduce job as input files.

MapFileOutputFormat - This is used to write the output as map files.

SequenceFileAsBinaryOutputFormat - This is another variant of SequenceFileInputFormat. It writes keys and values to a sequence file in binary format.

DBOutputFormat - This is used for writing to relational databases and HBase. This format also sends the reduce output to a SQL table.

25. What is the default replication factor?

The replication factor means the minimum number of times the file will replicate(copy) across the cluster.

The default replication factor is 3

26. Name two messages that NameNode gets from DataNode?

There are two messages which NameNode gets from DataNode.

They are

1) Block report and 2) Heartbeat.

27. Explain Hadoop distributed file system

Hadoop works with scalable distributed file systems like S3, HFTP FS, FS, and HDFS. Hadoop Distributed File System is made on the Google File System. This file system is designed in a way that it can easily run on a large cluster of the computer system.

28. What is Heartbeat in Hadoop?

In Hadoop, NameNode and DataNode communicate with each other. Heartbeat is the signal sent by DataNode to NameNode on a regular basis to show its presence.

29. How to define the distance between two nodes in Hadoop?

The distance is equal to the sum of the distance to the closest nodes. The method `getDistance()` is used to calculate the distance between two nodes.

30. What is a partition in Hive and why is partitioning required in Hive

Partition is a process for grouping similar types of data together based on columns or partition keys. Each table can have one or more partition keys to identify a particular partition.

Partitioning provides granularity in a Hive table. It reduces the query latency by scanning only relevant partitioned data instead of the entire data set. We can partition the transaction data for a bank based on month — January, February, etc. Any operation regarding a particular month, say February, will only have to scan the February partition, rather than the entire table data.

31. Why does Hive not store metadata information in HDFS?

We know that Hive's data is stored in HDFS. However, the metadata is either stored locally or it is stored in RDBMS. The metadata is not stored in HDFS, because HDFS read/write operations are time-consuming. As such, Hive stores metadata information in the metastore using RDBMS instead of HDFS. This allows us to achieve low latency and is faster.

32. What are the main configuration parameters in a “MapReduce” program?

The main configuration parameters which users need to specify in “MapReduce” framework are:

Job's input locations in the distributed file system
Job's output location in the distributed file system
The input format of data
The output format of data
Class containing the map function
Class containing the reduced function
JAR file containing the mapper, reducer, and driver classes

33. State the reason why we can't perform "aggregation" (addition) in the mapper? Why do we need the "reducer" for this?

This answer includes many points, so we will go through them sequentially.

We cannot perform "aggregation" (addition) in the mapper because sorting does not occur in the "mapper" function. Sorting occurs only on the reducer side and without sorting aggregation cannot be done.

During "aggregation", we need the output of all the mapper functions which may not be possible to collect in the map phase as mappers may be running on a different machine where the data blocks are stored.

And lastly, if we try to aggregate data at the mapper, it requires communication between all mapper functions which may be running on different machines. So, it will consume high network bandwidth and can cause network bottlenecking.

34. What is the purpose of "RecordReader" in Hadoop?

The "InputSplit" defines a slice of work but does not describe how to access it. The "RecordReader" class loads the data from its source and converts it into (key, value) pairs suitable for reading by the "Mapper" task. The "RecordReader" instance is defined by the "Input Format".

35. Explain "Distributed Cache" in a "MapReduce Framework".

Distributed Cache can be explained as, a facility provided by the MapReduce framework to cache files needed by applications. Once you have cached a file for your job, the Hadoop framework will make it available on each and every data node where you map/reduce tasks running. Then you can access the cache file as a local file in your Mapper or Reducer job.

36. How do "reducers" communicate with each other?

This is a tricky question. The "MapReduce" programming model does not allow "reducers" to communicate with each other. "Reducers" run in isolation.

37. What does a "MapReduce Partitioner" do?

A "MapReduce Partitioner" makes sure that all the values of a single key go to the same "reducer", thus allowing even distribution of the map output over the "reducers". It redirects the

“mapper” output to the “reducer” by determining which “reducer” is responsible for the particular key.

38. What role do RecordReader, Combiner, and Partitioner play in a MapReduce operation?

RecordReader

This communicates with the InputSplit and converts the data into key-value pairs suitable for the mapper to read.

Combiner

This is an optional phase; it is like a mini reducer. The combiner receives data from the map tasks, works on it, and then passes its output to the reducer phase.

Partitioner

The partitioner decides how many reduced tasks would be used to summarize the data. It also confirms how outputs from combiners are sent to the reducer, and controls the partitioning of keys of the intermediate map outputs.

39. Name some Hadoop-specific data types that are used in a MapReduce program.

This is an important question, as you would need to know the different data types if you are getting into the field of Big Data.

For every data type in Java, you have an equivalent in Hadoop. Therefore, the following are some Hadoop-specific data types that you could use in your MapReduce program:

IntWritable
FloatWritable
LongWritable
DoubleWritable
BooleanWritable
ArrayWritable
MapWritable
ObjectWritable

40. Can we write the output of MapReduce in different formats?

Yes. Hadoop supports various input and output File formats, such as:

TextOutputFormat - This is the default output format and it writes records as lines of text.

SequenceFileOutputFormat - This is used to write sequence files when the output files need to be fed into another MapReduce job as input files.

MapFileOutputFormat - This is used to write the output as map files.

SequenceFileAsBinaryOutputFormat - This is another variant of SequenceFileInputFormat. It writes keys and values to a sequence file in binary format.

DBOutputFormat - This is used for writing to relational databases and HBase. This format also sends the reduce output to a SQL table.

41. Can we have more than one ResourceManager in a YARN-based cluster?

Yes, Hadoop v2 allows us to have more than one ResourceManager. You can have a high availability YARN cluster where you can have an active ResourceManager and a standby ResourceManager, where the ZooKeeper handles the coordination.

There can only be one active ResourceManager at a time. If an active ResourceManager fails, then the standby ResourceManager comes to the rescue.

42. What are the different schedulers available in YARN?

The different schedulers available in YARN are:

FIFO scheduler - This places applications in a queue and runs them in the order of submission (first in, first out). It is not desirable, as a long-running application might block the small running applications

Capacity scheduler - A separate dedicated queue allows the small job to start as soon as it is submitted. The large job finishes later compared to using the FIFO scheduler

Fair scheduler - There is no need to reserve a set amount of capacity since it will dynamically balance resources between all the running jobs

43. In a cluster of 10 DataNodes, each having 16 GB RAM and 10 cores, what would be the total processing capacity of the cluster?

Every node in a Hadoop cluster will have one or multiple processes running, which would need RAM. The machine itself, which has a Linux file system, would have its own processes that need a specific amount of RAM usage. Therefore, if you have 10 DataNodes, you need to allocate at least 20 to 30 percent towards the overheads, Cloudera-based services, etc. You could have 11 or 12 GB and six or seven cores available on every machine for processing. Multiply that by 10, and that's your processing capacity.

44. What happens if requested memory or CPU cores go beyond the size of container allocation?

If an application starts demanding more memory or more CPU cores that cannot fit into a container allocation, your application will fail. This happens because the requested memory is more than the maximum container size.

Now that you have learned about HDFS, MapReduce, and YARN, let us move to the next section. We'll go over questions about Hive, Pig, HBase, and Sqoop.

45. Write a query to insert a new column(new_col INT) into a hive table (h_table) at a position before an existing column (x_col).

The following query will insert a new column:

```
ALTER TABLE h_table  
CHANGE COLUMN new_col INT  
  
BEFORE x_col
```

46. What is “SerDe” in “Hive”?

Apache Hive is a data warehouse system built on top of Hadoop and is used for analyzing structured and semi-structured data developed by Facebook. Hive abstracts the complexity of Hadoop MapReduce.

The “SerDe” interface allows you to instruct “Hive” about how a record should be processed. A “SerDe” is a combination of a “Serializer” and a “Deserializer”. “Hive” uses “SerDe” (and “FileFormat”) to read and write the table’s row.

To know more about Apache Hive, you can go through this Hive tutorial blog.

47. Can the default “Hive Metastore” be used by multiple users (processes) at the same time? “Derby database” is the default “Hive Metastore”. Multiple users (processes) cannot access it at the same time. It is mainly used to perform unit tests.

48. What is the default location where “Hive” stores table data?

The default location where Hive stores table data is inside HDFS in /user/hive/warehouse.

49. What happens when a data node fails?

If a data node fails the job tracker and name node will detect the failure. After that, all tasks are re-scheduled on the failed node and then name node will replicate the user data to another node.

50. What is Hadoop Streaming?

Hadoop streaming is a utility which allows you to create and run map/reduce job. It is a generic API that allows programs written in any languages to be used as Hadoop mapper.

51. What are the network requirements for using Hadoop?

Following are the network requirement for using Hadoop:

Password-less SSH connection.

Secure Shell (SSH) for launching server processes.

52. What do you know by storage and compute node?

Storage node: Storage Node is the machine or computer where your file system resides to store the processing data.

Compute Node: Compute Node is a machine or computer where your actual business logic will be executed.

53. Is it possible to provide multiple inputs to Hadoop? If yes, explain.

Yes, It is possible. The input format class provides methods to insert multiple directories as input to a Hadoop job.

54. What is the relation between job and task in Hadoop?

In Hadoop, A job is divided into multiple small parts known as the task.

55. What is the difference between Input Split and HDFS Block?

The Logical division of data is called Input Split and physical division of data is called HDFS Block.

56. What is the difference between Input Split and HDFS Block?

The Logical division of data is called Input Split and physical division of data is called HDFS Block.

57. What is the difference between Hadoop and other data processing tools?

Hadoop facilitates you to increase or decrease the number of mappers without worrying about the volume of data to be processed.

58. What commands are used to see all jobs running in the Hadoop cluster and kill a job in LINUX?

Hadoop job - list

Hadoop job - kill jobID

59. Mention what daemons run on a master node and slave nodes?

Daemons run on Master node is “NameNode”

Daemons run on each Slave nodes are “Task Tracker” and “Data”

60. Explain what is storage and compute nodes?

The storage node is the machine or computer where your file system resides to store the processing data

The compute node is the computer or machine where your actual business logic will be executed.

61. Mention what is the use of Context Object?

The Context Object enables the mapper to interact with the rest of the Hadoop

system. It includes configuration data for the job, as well as interfaces which allow it to emit output.

62. Mention what is the next step after Mapper or MapTask?

The next step after Mapper or MapTask is that the output of the Mapper are sorted, and partitions will be created for the output.

63. Mention what is the number of default partitioner in Hadoop?

In Hadoop, the default partitioner is a “Hash” Partitioner.

64. Explain how is data partitioned before it is sent to the reducer if no custom partitioner is defined in Hadoop?

If no custom partitioner is defined in Hadoop, then a default partitioner computes a hash value for the key and assigns the partition based on the result.

65. Explain what happens when Hadoop spawned 50 tasks for a job and one of the task failed?
It will restart the task again on some other TaskTracker if the task fails more than the defined limit.

66. Mention what is the best way to copy files between HDFS clusters?

The best way to copy files between HDFS clusters is by using multiple nodes and the distcp command, so the workload is shared.

67. Mention what job does the conf class do?

Job conf class separate different jobs running on the same cluster. It does the job level settings such as declaring a job in a real environment.

68. Mention what is the Hadoop MapReduce APIs contract for a key and value class?

For a key and value class, there are two Hadoop MapReduce APIs contract

The value must be defining the org.apache.hadoop.io.Writable interface

The key must be defining the org.apache.hadoop.io.WritableComparable interface

69. Mention what does the text input format do?

The text input format will create a line object that is an hexadecimal number. The value is considered as a whole line text while the key is considered as a line object. The mapper will receive the value as ‘text’ parameter while key as ‘longwriteable’ parameter.

70. Mention how many InputSplits is made by a Hadoop Framework?

Hadoop will make 5 splits

1 split for 64K files

2 split for 65mb files

2 splits for 127mb files

71. Mention what is distributed cache in Hadoop?

Distributed cache in Hadoop is a facility provided by MapReduce framework. At the time of execution of the job, it is used to cache file. The Framework copies the necessary files to the slave node before the execution of any task at that node.

72. Explain how does Hadoop Classpath plays a vital role in stopping or starting in Hadoop daemons?

Classpath will consist of a list of directories containing jar files to stop or start daemons.

73. What is a block and block scanner in HDFS?

Block - The minimum amount of data that can be read or written is generally referred to as a "block" in HDFS. The de

size of a block in HDFS is 64MB.

Block Scanner - Block Scanner tracks the list of blocks present on a DataNode and verifies them to find any kind of checksum errors. Block Scanners use a throttling mechanism to reserve disk bandwidth on the datanode.

74. Checkpoint Node-

Checkpoint Node keeps track of the latest checkpoint in a directory that has same structure as that of NameNode's directory. Checkpoint node creates checkpoints for the namespace at regular intervals by downloading the edits and fsimage file from the NameNode and merging it locally. The new image is then again updated back to the active NameNode.

75. What is the port number for NameNode, Task Tracker and Job Tracker?

NameNode 50070

Job Tracker 50030

Task Tracker 50060

76. Explain about the process of inter cluster data copying.

HDFS provides a distributed data copying facility through the DistCP from source to destination. If this data copying is within the hadoop cluster then it is referred to as inter cluster data copying. DistCP requires both source and destination to have a compatible or same version of hadoop.

77. Explain what happens if during the PUT operation, HDFS block is assigned a replication factor 1 instead of the default value 3.

Replication factor is a property of HDFS that can be set accordingly for the entire cluster to adjust the number of times the blocks are to be replicated to ensure high data availability. For every block that is stored in HDFS, the cluster will have n-1 duplicated blocks. So, if the replication factor during the PUT operation is set to 1 instead of the default value 3, then it will have a single copy of data. Under these circumstances when the replication factor is set to 1 ,if the DataNode crashes under any circumstances, then only single copy of the data would be lost.

78. What happens to a NameNode that has no data?

There does not exist any NameNode without data. If it is a NameNode then it should have some sort of data in it.

79. What happens when a user submits a Hadoop job when the NameNode is down- does the job get in to hold or does it fail.

The Hadoop job fails when the NameNode is down.

80. What happens when a user submits a Hadoop job when the Job Tracker is down- does the job get in to hold or does it fail.

The Hadoop job fails when the Job Tracker is down.

81. Whenever a client submits a hadoop job, who receives it?

NameNode receives the Hadoop job which then looks for the data requested by the client and provides the block information. JobTracker takes care of resource allocation of the hadoop job to ensure timely completion.

82. Shuffle Phase-Once the first map tasks are completed, the nodes continue to perform several other map tasks and also exchange the intermediate outputs with the reducers as required. This process of moving the intermediate outputs of map tasks to the reducer is referred to as Shuffling.

83. What is speculative execution in Hadoop?

If a node appears to be running slow, the master node can redundantly execute another instance of the same task and first output will be taken .this process is called as Speculative execution.

84. What are Problems with small files and HDFS?

If a node appears to be running slow, the master node can redundantly execute another instance of the same task and first output will be taken .this process is called as Speculative execution.

85. Can Hadoop handle streaming data?

Yes, through Technologies like Apache Kafka, Apache Flume, and Apache Spark it is possible to do large-scale streaming.

86. How to make a large cluster smaller by taking out some of the nodes?

Hadoop offers the decommission feature to retire a set of existing data-nodes. The nodes to be retired should be included into the exclude file, and the exclude file name should be specified as a configuration parameter `dfs.hosts.exclude`.

The decommission process can be terminated at any time by editing the configuration or the exclude files and repeating the `-refreshNodes` command

87. Does the name-node stay in safe mode till all under-replicated files are fully replicated?

No. During safe mode replication of blocks is prohibited. The name-node awaits when all or majority of data-nodes report their blocks.

88. What does "file could only be replicated to 0 nodes, instead of 1" mean?

The namenode does not have any available DataNodes.

89. What is the communication channel between client and namenode/datanode?

The mode of communication is SSH

90. Are job tracker and task trackers present in separate machines?

Yes, job tracker and task tracker are present in different machines. The reason is job tracker is a single point of failure for the Hadoop MapReduce service. If it goes down, all running jobs are halted.

91. Explain about the replication and multiplexing selectors in Flume.

Channel Selectors are used to handle multiple channels. Based on the Flume header value, an event can be written just to a single channel or to multiple channels. If a channel selector is not specified to the source then by default it is the Replicating selector. Using the replicating selector, the same event is written to all the channels in the source's channels list. Multiplexing channel selector is used when the application has to send different events to different channels.

92. What is the process to change the files at arbitrary locations in HDFS?

HDFS does not support modifications at arbitrary offsets in the file or multiple writers but files are written by a single writer in append only format i.e. writes to a file in HDFS are always made at the end of the file.

93. What happens to a NameNode that has no data?

There does not exist any NameNode without data. If it is a NameNode then it should have some sort of data in it.

94. What happens when a user submits a Hadoop job when the NameNode is down- does the job get in to hold or does it fail.

The Hadoop job fails when the NameNode is down.

95. What happens when a user submits a Hadoop job when the Job Tracker is down- does the job get in to hold or does it fail.

The Hadoop job fails when the Job Tracker is down.

96. Whenever a client submits a hadoop job, who receives it?

NameNode receives the Hadoop job which then looks for the data requested by the client and provides the block information. JobTracker takes care of resource allocation of the hadoop job to ensure timely completion.

97. Mention some differences between HDFS high availability and HDFS federation.

HFS High Availability	HDFS Federation
There are two NameNodes, one active and one on standby.	In the HDFS Federation, the NameNodes are not related to each other.
If the primary NameNode goes down, the standby will take its place using the most recent metadata that it has.	Even if one NameNode goes down, there is no effect on the other NameNodes.
At a given time, only the active NameNode will be running, while the standby NameNode remains idle and only updates its metadata to stay up to date.	There is a pool of metadata which is shared by all the NameNodes. In addition, each NameNode will have its own metadata.
This requires two separate machines. One machine is configured to be the primary NameNode and the other as the Standby NameNode.	There is no limit on the number of machines that can be configured as the NameNodes.

98. What are side data distribution techniques in Hadoop?

The extra read only data required by a hadoop job to process the main dataset is referred to as side data. Hadoop has two side data distribution techniques -

i) Using the job configuration - This technique should not be used for transferring more than few kilobytes of data as it can pressurize the memory usage of hadoop daemons,particularly if your system is running several hadoop jobs.

ii) Distributed Cache - Rather than serializing side data using the job configuration, it is suggested to distribute data using hadoop's distributed cache mechanism.

99. What are the main components of a Hadoop Application?

Hadoop applications have a wide range of technologies that provide a great advantage in solving complex business problems.

Core components of a Hadoop application are-

- 1) Hadoop Common
- 2) HDFS
- 3) Hadoop MapReduce
- 4) YARN

Data Access Components are - Pig and Hive

Data Storage Component is - HBase

Data Integration Components are - Apache Flume, Sqoop, Chukwa

Data Management and Monitoring Components are - Ambari, Oozie and Zookeeper.

Data Serialization Components are - Thrift and Avro

Data Intelligence Components are - Apache Mahout and Drill.

100. What is the best hardware configuration to run Hadoop?

The best configuration for executing Hadoop jobs is dual core machines or dual processors with 4GB or 8GB RAM that use ECC memory. Hadoop highly benefits from using ECC memory though it is not low - end. ECC memory is recommended for running Hadoop because most of the Hadoop users have experienced various checksum errors by using non ECC memory. However, the hardware configuration also depends on the workflow requirements and can change accordingly.

Chapter - 18 - Data Pipeline

1.What is Data Engineering?

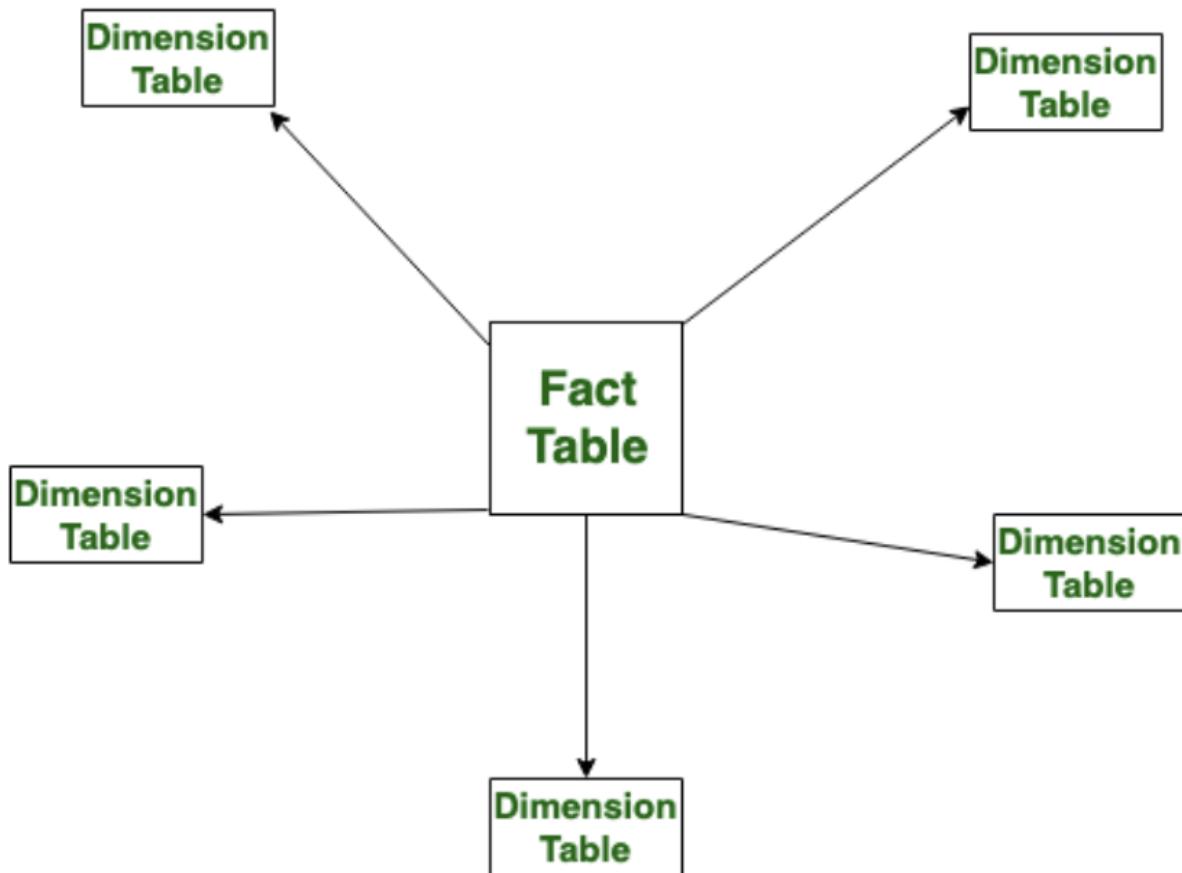
The application of data collecting and analysis is the emphasis of data engineering. The information gathered from numerous sources is merely raw information. Data engineering helps in the transformation of unusable data into useful information. It is the process of transforming, cleansing, profiling, and aggregating huge data sets in a nutshell.

2. What is Data Modeling?

Data Modeling is the act of creating a visual representation of an entire information system or parts of it in order to express linkages between data points and structures. The purpose is to show the many types of data that are used and stored in the system, as well as the relationships between them, how the data can be classified and arranged, and its formats and features. Data can be modeled according to the needs and requirements at various degrees of abstraction. The process begins with stakeholders and end-users providing information about business requirements. These business rules are then converted into data structures, which are used to create a concrete database design.

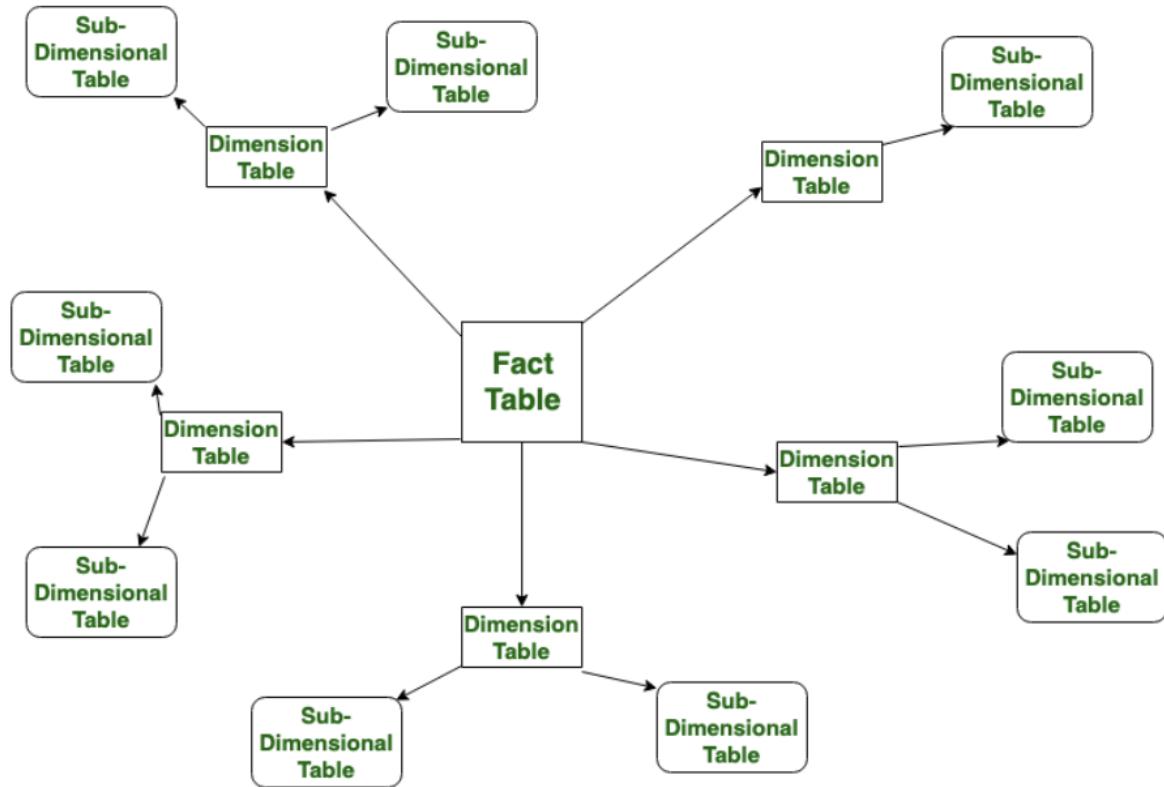
3. What is Star schema?

Star schema is the type of multidimensional model which is used for data warehouse. In star schema, The fact tables and the dimension tables are contained. In this schema fewer foreign-key join is used. This schema forms a star with fact table and dimension tables.



4. What is snowflake schema?

Snowflake Schema is also the type of multidimensional model which is used for data warehouse. In snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained. This schema forms a snowflake with fact tables, dimension tables as well as sub-dimension tables.



5. What is fact table?

A reality or fact table's record could be a combination of attributes from totally different dimension tables. The Fact Table or Reality Table helps the user to investigate the business dimensions that helps him in call taking to enhance his business.

6. What is dimension table?

Dimension Tables facilitate the reality table or fact table to gather dimensions on that the measures needs to be taken.

7. What is aggregate fact table?

- Aggregate fact tables are a special kind of fact tables in a data warehouse which contains new metrics which are been derived from one or more aggregate functions (COUNT, AVERAGE, MIN, MAX, etc.) or from some specialized functions whose outputs are totally derived from a grouping of base data.

- Aggregates are basically summarization of the fact related data which are been used as a purpose to improve the performance.

- These new metrics, called as "aggregate facts" or "summary statistics" are been stored and maintained in database of the data warehouse in special fact table at the grain of the aggregation.

- In similar way, the corresponding dimensions are been rolled up and compressed to match the new grain of the fact.

- These specialized tables are been used as an substitutions whenever possible for returning user queries. The reason is the speed.
- Querying a neat aggregate table is much faster and uses less of the disk I/O than the base, atomic fact table, especially when the dimensions are large as well.
- If you want to amaze your users then start adding the aggregates.
- Even you can use this technique in your operational systems as well, giving boost to the foundational reports.

8. What is the difference between a fact and dimension table?

S.NO	Fact Table	Dimension Table
1.	Fact table contains the measuring of the attributes of a dimension table.	Dimension table contains the attributes on that truth table calculates the metric.
2.	In fact table, There is less attributes than dimension table.	While in dimension table, There is more attributes than fact table.
3.	In fact table, There is more records than dimension table.	While in dimension table, There is less records than fact table.
4.	Fact table forms a vertical table.	While dimension table forms a horizontal table.
5.	The attribute format of fact table is in numerical format and text format.	While the attribute format of dimension table is in text format.
6.	It comes after dimension table.	While it comes before fact table.
7.	The number of fact table is less than dimension table in a schema.	While the number of dimension is more than fact table in a schema.
8.	It is used for analysis purpose and decision making.	While the main task of dimension table is to store the information about a business and its process.

9. What is the difference between a data engineer and a data scientist?

Data science is a broad topic of research. It focuses on extracting data from extremely huge datasets (sometimes it is known as "big data"). Data scientists can operate in a variety of fields, including industry, government, and applied sciences. All data scientists have the same goal: to analyze data and derive insights from it that are relevant to their field of work.

A data engineer's job is to develop or integrate many components of complex systems, taking into account the information needed, the company's goals, and the end requirements. This

necessitates the creation of extremely complicated data pipelines. These data pipelines, like oil pipelines, take raw, unstructured data from a variety of sources. They then channel them into a single database (or larger structure) for storage.

10. What are the differences between structured and unstructured data?

On the basis of	Structured	Unstructured
Storage	Structured data is stored in DBMS.	It is stored in unmanaged file structures.
Flexibility	It is less flexible as it is dependent on the schema.	It is more flexible.
Scalability	Not easy to scale.	Easy to scale.
Performance	Since we can perform a structured query, the performance is high.	The performance of unstructured data is low.
Analysis factor	Easy to analyze.	Hard to analyze.

11. What are the different types of data pipelines?

There are many different types of data pipelines, but some of the most common are Extract-Transform-Load (ETL) pipelines and Extract-Load-Transform (ELT) pipelines. ETL pipelines extract data from one or more sources, transform it into a format that can be loaded into a destination system, and then load it into that system. ELT pipelines extract data from one or more sources and load it into a destination system, where it is then transformed into the desired format.

12. Can you explain what ETL is? How does it differ from ELT?

ETL stands for Extract, Transform, Load. It is a process in which data is extracted from a source, transformed into a format that can be loaded into a destination, and then loaded into that destination. ELT stands for Extract, Load, Transform. It is a process in which data is extracted from a source, loaded into a destination, and then transformed into a format that can be used by that destination.

13. How do you identify a bottleneck in a data pipeline?

A bottleneck in a data pipeline is typically identified by a decrease in throughput or an increase in latency. If you are seeing either of these issues, it is likely that there is a bottleneck somewhere in the pipeline. To further narrow down the location of the bottleneck, you can look at individual components of the pipeline to see where the slowdown is occurring.

14. How do you analyze tables in ETL?

You can validate the structures of system objects using the ANALYZE statement. This statement generates statistics, which are then used by a cost-based optimizer to determine the most effective strategy for data retrieval. ESTIMATE, DELETE, and COMPUTER are some additional operations.

15. What SQL commands allow you to validate data completion?

You can validate the completeness of the data using the intersect and minus statements. When you run source minus target and target minus source, the minus query returns a value indicating rows that don't match. A duplicate row is present when the count intersects is less than the source count, and the minus query returns the value.

16. What role does impact analysis play in an ETL system?

Impact analysis analyzes the metadata relating to an object to determine what is impacted by a change in its structure or content. A data warehouse's proper loading can be affected by changing data-staging objects in processes. You must conduct an impact analysis before modifying a table once it has been created in the staging area.

17. What Are the Various Phases of Data Mining?

A phase in data mining is a logical procedure for sifting through vast amounts of data to identify crucial data.

Exploration: The exploration phase aims to identify significant variables and determine their characteristics.

Pattern Identification: In this phase, the main task is looking for patterns and selecting the best prediction.

Deployment stage: This stage can only be attained once a reliable, highly predictive pattern is identified in stage 2.

18. Explain the use of ETL in data migration projects.

The usage of ETL tools is popular in data migration projects. For instance, if a company previously managed its data in Oracle 10g and now wants to switch to a SQL Server cloud database, it will need to be moved from Source to Target. ETL tools are quite beneficial when performing this kind of data migration. ETL code writing will take a lot of the user's time. As a result, the ETL tools are helpful because they make coding easier than P-SQL or T-SQL. Therefore, ETL is a beneficial process for projects involving data migration.

19. What performs better, joining data first, then filtering it, or filtering data first, then joining it with other sources?

Filtering data and then joining it with other data sources is better.

Filtering unnecessary data as early as possible in the process is an excellent technique to enhance the efficiency of the ETL process. It cuts down on time necessary for data transfer, I/O, and/or memory processing.

The general idea is to minimize the number of processed rows and avoid altering data that is never utilized.

20. Explain how ETL and OLAP (Online Analytical Processing) tools differ.

ETL tools: ETL tools allow you to extract, transform, and load the data in the data warehouse or data mart. To apply business logic, several transformations are necessary before data is loaded into the target table.

OLAP (Online Analytical Processing) tools: OLAP tools help generate reports from data marts and warehouses for business analysis. It transfers data from the target tables into the OLAP repository and performs the necessary modifications to create a report.

21. Briefly explain ETL mapping sheets.

ETL mapping sheets usually offer complete details about a source and a destination table, including each column and how to look them up in reference tables. ETL testers may have to generate big queries with several joins at any stage of the testing process to check data. When using ETL mapping sheets, writing data verification queries is much easier.

22. Why is it important to automate your data pipelines?

Automating your data pipelines is important for a number of reasons. First, it can help to ensure that your data is consistently formatted and of high quality, as human error can be eliminated from the equation. Second, automating your data pipelines can help to improve efficiency and speed, as data can be processed and moved more quickly without the need for manual intervention. Finally, automating your data pipelines can help to improve security, as sensitive data can be better protected when it is not being handled by humans.

23. What's the difference between batch and real-time processing in data pipelines?

Batch processing is the process of collecting data over a period of time and then processing it all at once. Real-time processing is the process of collecting and processing data as it comes in, in near-real-time.

24. What are some challenges faced when building data pipelines?

There are a few challenges that are commonly faced when building data pipelines. One challenge is ensuring that data is consistently formatted as it moves through the pipeline. Another challenge is dealing with data that is incomplete or has errors. Finally, it can be difficult to monitor and optimize the performance of data pipelines.

25. When should you use a stream processor as opposed to a database?

A stream processor is best used when you need to process data in real time as it is coming in, such as for monitoring or logging purposes. A database is better suited for storing data for later retrieval and analysis.

26. What are some differences between Apache Storm, Spark Streaming, Flink, Samza, and Apex?

There are a few key differences between these data processing frameworks. Apache Storm is designed for real-time processing of streaming data, while Apache Spark Streaming is a more general-purpose framework that can be used for both batch and streaming data processing. Flink is another general-purpose framework, but with a focus on streaming data processing and event-based applications. Samza is a framework specifically for stream processing of data from Apache Kafka. Apex is a framework designed for low-latency, high-throughput processing of streaming data.

27. What is Kafka Streams? Why would you use it over other streaming systems like Spark or Storm?

Kafka Streams is a stream processing library that is built on top of Apache Kafka. It allows you to easily build streaming applications that process data from Kafka topics. Kafka Streams has several advantages over other streaming systems, including its simple API, its ability to handle out-of-order data, and its built-in support for fault tolerance.

28. Can you explain what Lambda Architecture is? Why is it so popular?

Lambda Architecture is a data processing architecture that is designed to handle massive quantities of data by using a combination of batch processing and real-time streaming. It is popular because it is able to provide low-latency results while still being able to process a large amount of data.

29. What are the main components of an enterprise-grade data pipeline?

The main components of an enterprise-grade data pipeline are a data ingestion system, a data processing system, a data storage system, and a data analysis system. The data ingestion system is responsible for collecting data from various sources and delivering it to the data processing system. The data processing system then cleans and transforms the data before storing it in the data storage system. The data analysis system then uses the data in the data storage system to generate insights and reports.

30. What are the advantages of using a solid data pipeline architecture?

A solid data pipeline architecture can help to ensure that data is processed efficiently and effectively. It can also help to ensure that data is accurate and consistent, and that it is easy to track and monitor the data pipeline.

31. How can you create a robust data pipeline that can handle high volumes of traffic with minimal latency?

There are a few key things to keep in mind when building a data pipeline that needs to handle high volumes of traffic. First, you need to make sure that your data pipeline is scalable so that it can easily handle increased traffic. Second, you need to optimize your data pipeline for performance so that it can minimize latency. Finally, you need to make sure that your data pipeline is fault-tolerant so that it can continue to operate even if there are errors or failures.

32. What is a data warehouse? How does it work?

A data warehouse is a database that is used to store data for reporting and analysis. Data warehouses are usually designed to hold data from multiple sources, and they are often used to track historical data. Data warehouses typically use a star schema, which means that data is organized into tables that are connected by relationships.

33. What are some advantages of using a data lake instead of a data warehouse?

A data lake can be less expensive to maintain than a data warehouse because it does not require the same level of data cleansing and organization. A data lake can also be more flexible, allowing for a wider variety of data types to be stored and accessed.

34. What are the strongest arguments for using machine learning in data pipelines?

The strongest arguments for using machine learning in data pipelines are its ability to automate complex processes and its ability to improve the accuracy of predictions. Machine learning can automate the process of feature engineering, which is often required in order to get the most out of data. Additionally, machine learning can help to improve the accuracy of predictions made by data pipelines by learning from past data and making adjustments accordingly.

35. Why is Kafka a great option for building scalable data pipelines?

Kafka is a great option for building scalable data pipelines because it is a high-performance, distributed streaming platform. Kafka can handle extremely large volumes of data very quickly, and it is able to do so while maintaining low latency. This makes it an ideal tool for building data pipelines that need to be able to handle large amounts of data in real-time.

36. How does Amazon Kinesis compare to Apache Kafka?

Both Amazon Kinesis and Apache Kafka are streaming data platforms that can be used to process and analyze large amounts of data in real time. However, there are some key differences between the two. For one, Kafka is an open source project, while Kinesis is a proprietary Amazon service. Kafka also has a more robust set of features and can handle higher throughputs than Kinesis. Finally, Kafka is more difficult to set up and manage than Kinesis.

37. Does Spark provide better performance than MapReduce? If yes, then how?

Spark does provide better performance than MapReduce in a few key ways. First, Spark is able to run computations in-memory, which can be a significant speed boost. Additionally, Spark uses a more efficient shuffle implementation than MapReduce, which can lead to better performance on certain types of workloads. Finally, Spark can also perform multiple computations in parallel, whereas MapReduce is limited to running one computation at a time.

38. Is it possible to build a data pipeline with multiple sources and sinks? If yes, then how?

Yes, it is possible to build a data pipeline with multiple sources and sinks. This can be done by creating a separate pipeline for each source and sink, and then connecting the pipelines together.

39. What are the four Vs of Big Data?

The four characteristics or four Vs of Big data are:

Volume
Veracity
Velocity
Variety

40. What is the Replication factor?

The replication factor is the number of times the Hadoop framework replicates each Data Block. Fault tolerance is provided by replicating the block. The replication factor is set to 3 by default, however, it can be modified to 2 (less than 3) or raised to meet your needs (more than 3.)

41. What do you mean by data pipeline?

A data pipeline is a system for transporting data from one location (the source) to another (the destination) (such as a data warehouse). Data is converted and optimized along the journey, and it eventually reaches a state that can be evaluated and used to produce business insights. The procedures involved in aggregating, organizing, and transporting data are referred to as a data pipeline. Many of the manual tasks needed in processing and improving continuous data loads are automated by modern data pipelines.

DATA PIPELINE



42. What are different data validation approaches?

The process of confirming the accuracy and quality of data is known as data validation. It is implemented by incorporating various checks into a system or report to ensure that input and stored data are logically consistent. Common types of data validation approaches are

43. Data type check: It confirms that the data entered is of the correct data type.

Code check: A code check verifies that a field is chosen from a legitimate list of options or that it corresponds to specific formatting constraints. Checking a postal code against a list of valid codes, for example, makes it easier to verify if it is valid.

Range check: It ensures that input falls in a predefined range.

Format check: Many data types follow a predefined format. Format check confirms that. For example, a date has formats like DD-MM-YY or MM-DD-YY.

Consistency check: It confirms that the data entered is logically correct.

Uniqueness check: It ensures that the same data is not entered multiple times.

44. Data pipeline vs ETL

There's frequently confusion about a pipeline and ETL. So the first order of business is to clear that up. Simply stated, ETL is just a type of data pipeline, that includes three major steps

Extract – getting/ingesting data from original, disparate source systems.

Transform – moving data in temporary storage known as a staging area. Transforming data to ensure it meets agreed formats for further uses, such as analysis.

Load – loading reformatted data to the final storage destination.

This is a common but not the only approach to moving data. For example, not every pipeline has a transformation stage. You simply don't need it if source and target systems support the same data format. We'll discuss ETL and other types of data pipelines later on.

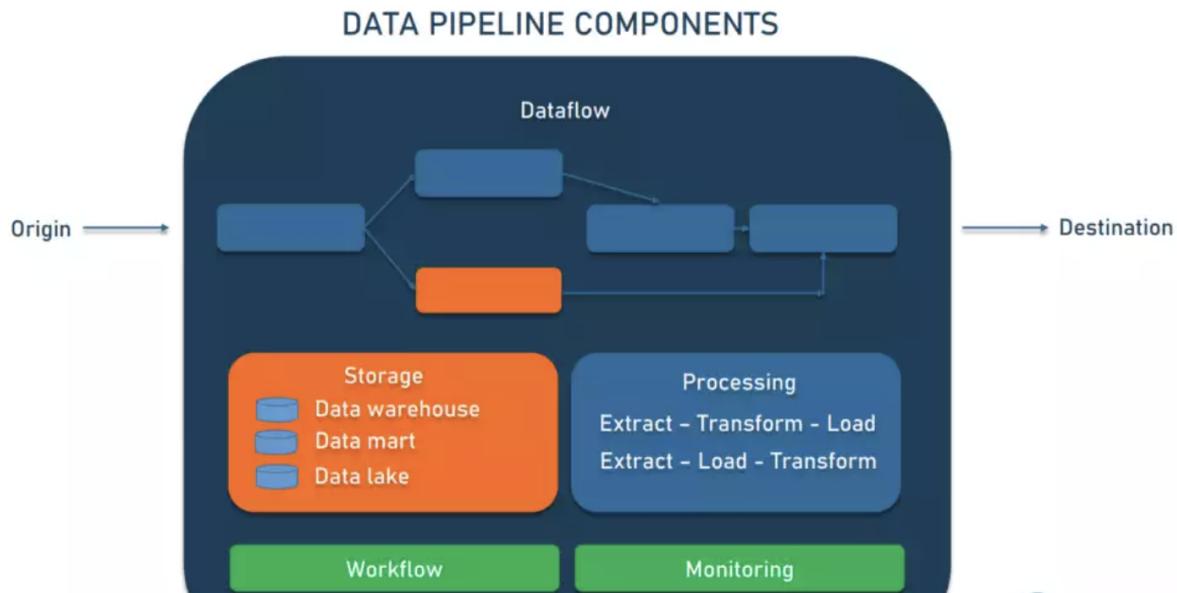
45. When do you need a data pipeline?

Reliable infrastructure for consolidating and managing data helps organizations power their analytical tools and support daily operations. Having a data pipeline is necessary if you plan to use data for different purposes, with at least one of them requiring data integration — for example, processing and storing transaction data and conducting a sales trend analysis for the whole quarter.

To carry out the analysis, you will have to pull data from a number of sources (i.e., a transaction system, CRM, a website analytics tool) to access it from single storage and prepare it for the analysis. So, a data pipeline allows for solving “origin-destination” problems, especially with large amounts of data.

Also, the more use cases, the more forms data can be stored in, and the more ways it can be processed, transmitted, and used.

46. Draw the Data Pipeline architecture



47. Explain Origin, Destination, Dataflow and storage

Origin. Origin is the point of data entry in a data pipeline. Data sources (transaction processing application, IoT devices, social media, APIs, or any public datasets) and storage systems (data warehouse, data lake, or data lakehouse) of a company's reporting and analytical data environment can be an origin.

Destination. The final point to which data is transferred is called a destination. Destination depends on a use case: Data can be sourced to power data visualization and analytical tools or moved to storage like a data lake or a data warehouse. We'll get back to the types of storage a bit later.

Dataflow. That's the movement of data from origin to destination, including the changes it undergoes along the way as well as the data stores it goes through.

Storage. Storage refers to systems where data is preserved at different stages as it moves through the pipeline. Data storage choices depend on various factors, for example, the volume of data, frequency and volume of queries to a storage system, uses of data, etc. (think of the online bookstore example).

48. Explain processing, workflow and monitoring in a data pipeline.

Processing. Processing includes activities and steps for ingesting data from sources, storing it, transforming it, and delivering it to a destination. While data processing is related to the dataflow, it focuses on how to implement this movement. For instance, one can ingest data by

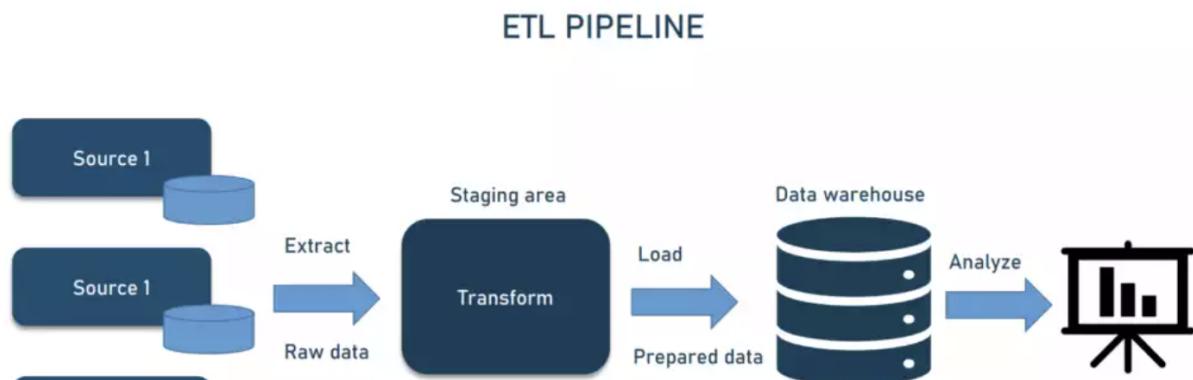
extracting it from source systems, copying it from one database to another one (database replication), or by streaming data. We mention just three options, but there are more of them.

Workflow. The workflow defines a sequence of processes (tasks) and their dependence on each other in a data pipeline. Knowing several concepts – jobs, upstream, and downstream – would help you here. A job is a unit of work that performs a specified task – what is being done to data in this case. Upstream means a source from which data enters a pipeline, while downstream means a destination it goes to. Data, like water, flows down the data pipeline. Also, upstream jobs are the ones that must be successfully executed before the next ones – downstream – can begin.

Monitoring. The goal of monitoring is to check how the data pipeline and its stages are working: Whether it maintains efficiency with growing data load, whether data remains accurate and consistent as it goes through processing stages, or whether no information is lost along the way.

49. ETL data pipeline

As we said before, ETL is the most common data pipeline architecture, one that has been a standard for decades. It extracts raw data from disparate sources, transforms it into a single pre-defined format, and loads it into a target system — typically, an enterprise data warehouse or data mart.



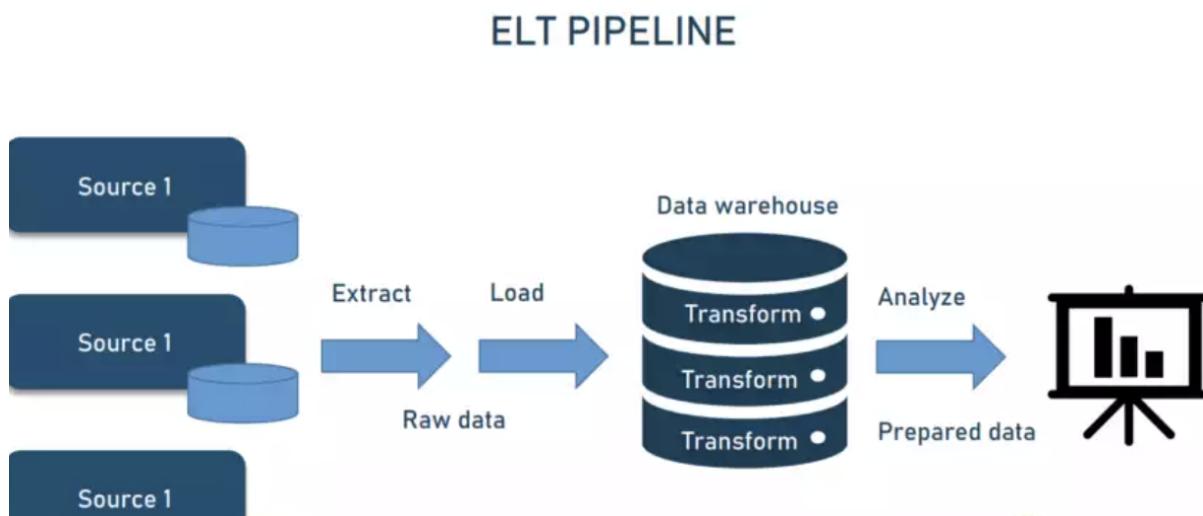
Typical use cases for ETL pipelines include

data migration from legacy systems to a data warehouse,
pulling user data from multiple touchpoints to have all information on customers in one place (usually, in the CRM system),
consolidating high volumes of data from different types of internal and external sources to provide a holistic view of business operations, and
joining disparate datasets to enable deeper analytics.

The key downside of the ETL architecture is that you have to rebuild your data pipeline each time business rules (and requirements for data formats) change. To address this problem, another approach to data pipeline architecture — ELT — appeared.

50. ELT data pipeline

Intuition correctly suggests that ELT differs from ETL in the sequence of steps: loading happens before the transformation. This seemingly minor shift changes a lot. Instead of converting huge amounts of raw data, you first move it directly into a data warehouse or data lake. Then, you can process and structure your data as needed — at any moment, fully or partially, once or numerous times.



ELT architecture comes in handy when

you're not sure what you're going to do with data and how exactly you want to transform it; the speed of data ingestion plays a key role; and huge amounts of data are involved.

Yet, ELT is still a less mature technology than ETL which creates problems in terms of available tools and talent pool.

Chapter 19 - Terminal Command

To excel in the field of Analytics, you need to be thorough with the Terminal and Hadoop Command, and most of the Big Data infrastructure is set up on Hadoop. Let's check out the most used 50 terminal commands. Do explore other Hadoop and terminal commands as per your use

1. Hadoop fs -ls
It lists the files in the Hadoop home directory
2. Hadoop fs -ls -R/
It lists the files in a recurring order
3. Hadoop fs -ls/
It lists the files in the Hadoop root directory
4. Hadoop fs -ls -t -r
It lists the files in reverse order, sorted by time
5. Hadoop fs -ls -s
It lists the files in descending order of size
6. Hadoop fs -ls /user |grep booking
It searches the files with the named booking
7. Hadoop fs -tail /user/thedatamonk/restaurant.txt
It will display the last 10 rows of the file restaurant.txt
8. Hadoop fs -rmdir /user/thedatamonk/direct
It will remove directories
9. Hadoop fs -rm -R /user/thedatamonk/direct
It will remove empty and non-empty directories as well
10. Hadoop fs -cp <complete file location 1> <complete directory location>
It will move a file from one location to another directory.
It copies a file from one location to another
11. Hadoop fs -mv <complete file location 1> <complete folder location>
It cuts and pastes the file from one location to another directory
12. Hadoop fs -copyFromLocal <Local folder location> <HDFS folder location>

It copies files from the local to the HDFS location. Remember, the keyword copyFromLocal is case-sensitive.

13. Hadoop fs -put <Local folder location> <HDFS folder location>
Another command to copy files from local to HDFS location. All the words are keywords.
14. Hadoop fs -copyToLocal <HDFS data location> <Local folder location>
It copies from HDFS to Local or desktop. Remember, the keyword copyToLocal is case-sensitive.
15. Hadoop fs -get <HDFS data location> <Local folder location>
Another command to copy from HDFS to Local or desktop.
16. Hadoop fs -df -h <location>
The above command gives the free space in the disk.
-h converts the space into bytes
17. Hadoop fs -du -h <location>
The above command gets the used space in the disk
18. Hadoop fs -touchz
To create an empty file on the file system
19. Hadoop fs -cat
It copies files to stdout
20. pwd
Display path of the current working directory
21. ls
Display the directory content
22. Mkdir <Directory Name>
Create a new directory with the given name. Example - mkdir alpha
It will create a directory name as alpha
23. Cat <File>
Output the content of the file. This is a very important command where the output of the content is displayed on the terminal
24. Head <File>
Output the top 10 lines of the file

25. Clear

To clear the command line window

26. Rm <File>

To remove the file

27. Rm -r <directory>

To delete the complete directory

Rm -f <File>

To force delete the directory

28. Mv <Old File> <New File>

Rename old file to new file

29. Mv <File> <Directory>

Move file from to a directory, overwriting an existing file(if present)

30. Cp <File> <Directory>

Copy file to a directory, overwriting an existing file (if present)

31. Cp -r <Directory 1 > <Directory 2>

Copy content of directory 1 to Directory 2

32. Find <directory> -name "<File>"

Search for all the names of the file present in the mentioned directory

33. Grep "<text>" <file>

This is also one of the important commands that helps you find all the files containing text in the directory

34. Ssh <username> @ <host>

Establish an ssh connection to host with user. This command is used to create connection with jump box and hadoop gateway

35. Kill <pid>

Quit process with ID <pid>

36. <cmd> > <file>

Direct the output of cmd in a file

37. <cmd> >> <file>

Append the output of cmd to file

38. `<cmd1> | <cmd2>`

Direct the output of cmd1 to cmd2

39. `Touch <file>`

Update file access and modification time

40. `Curl -O`

41. `cat` — command used to view the data from the file in HDFS

```
hadoop fs -cat <HDFS file path with file name>
```

42. `mv` — this command is used to move a file from one location to HDFS to another location in HDFS.

```
hadoop fs -mv <Source HDFS path> <Destination HDFS path>
```

43. `cp` — this command is used to copy a file from one location to HDFS to another location within HDFS only.

```
hadoop fs -cp <Source HDFS path> <Destination HDFS path>
```

44. `expunge` — this command is used to make the trash empty.

```
hadoop fs -expunge
```

45. `chown` — we should use this command when we want to change the user of a file or directory in HDFS.

```
hadoop fs -chown <HDFS file path>
```

46. `chgrp` — we should use this command when we want to change the group of a file or directory in HDFS.

```
hadoop fs -chgrp <HDFS file path>
```

47. `setrep` — this command is used to change the replication factor of a file in HDFS.

```
hadoop fs -setrep <Replication Factor> <HDFS file path>
```

48. `du` — this command is used to check the amount of disk usage of the file or directory.

```
hadoop fs -du <HDFS file path>
```

49. fsck — this command is used to check the health of the files present in the HDFS file system.

```
hadoop fsck <HDFS file path>
```

50. text — this is a simple command, used to print the data of an HDFS file on the console.

```
hadoop fs -text <HDFS file path>
```

51. appendToFile — this command is used to merge two files from the local file system to one file in the HDFS file.

```
hadoop fs -appendToFile <Local file path1> <Local file path2> <HDFS file path>
```

52. getmerge — this command is used to merge the contents of a directory from HDFS to a file in the local file system.

```
hadoop fs -getmerge <HDFS directory> <Local file path>
```

Chapter 20 - Regular Expression

1759. What is the most common field or application where regular expressions are used?

One of the most common fields that make use of regular expressions is software development. Some examples of how regular expressions are used in software development include:

Text Parsing: Regular expressions are often used to extract information from log files, configuration files, or other text-based data.

Input Validation: Regular expressions are commonly used to validate user input, such as email addresses, phone numbers, or passwords.

Search and Replace Operations: Regular expressions can be used to search for specific patterns within a text and replace them with new values.

Web Development: Regular expressions are frequently used in web development, such as in the implementation of URL routing, form validation, and input parsing.

Overall, regular expressions are a widely used tool in software development because they allow developers to efficiently perform complex text manipulation and validation tasks.

1760. How can you use backreferences in a regular expression to match a string that contains a specific pattern that is repeated multiple times, with each repetition having a unique variation?

Backreferences in Regular Expressions can be used to match a specific pattern that is repeated multiple times, with each repetition having a unique variation. Here's a simple explanation on how it works:

Capture the unique variation of the pattern using a capturing group. A capturing group is specified using parentheses in a regular expression.

Reference the capturing group in the rest of the expression using a backreference. A backreference is specified using the backslash "\(\)" followed by the group number.

For example, consider a string that contains repeating words, each word separated by a comma. To match a string that contains two words that are the same, you can use the following expression: "\(\w+\),\1".

The expression ""(\w+)" captures the first word in the string and \1 references the capturing group in the rest of the expression. This expression matches a string that contains two words that are the same, separated by a comma.

This technique can be used to match a specific pattern that is repeated multiple times, with each repetition having a unique variation, by capturing the unique variation in a capturing group and referencing the capturing group in the rest of the expression using a backreference.

1761. How would you use a regular expression to match and extract all the phone numbers from a large block of text, where the phone numbers can be in different formats explain with code?

Here is the code for above :

```
import re

# the regular expression pattern to match the phone numbers
pattern = re.compile(r'\b\d{3}[-.]?\d{3}[-.]?\d{4}\b')

# the text to search for phone numbers
text = "Phone numbers: 123-456-7890, (123) 456-7890, 123.456.7890, 1234567890"

# find all the matches and store them in a list
matches = re.findall(pattern, text)

# print the list of matches
print(matches)
```

This regular expression uses the \b word boundary, which matches a position that is between a word character (i.e. alphanumeric) and a non-word character. The \d matches a digit, and the {3} specifies that exactly three digits must be matched. The square brackets [-.] match either a hyphen or a dot. The ? following the square brackets makes the hyphen or dot optional. This allows the regular expression to match phone numbers with or without hyphens or dots. The {4} specifies that exactly four digits must be matched after the hyphen or dot. The \b at the end of the pattern matches another word boundary, which ensures that the phone number match does not include any extra characters before or after the match.

1762. How would you use a regular expression to match and extract all the URLs from a large block of text?

One way to use a regular expression to match and extract URLs from a large block of text is to use a pattern that matches the common structure of a URL, which includes a scheme (e.g. http, ftp, or https), a hostname or IP address, and a path.

This pattern uses the (http | ftp | https) group to match the scheme, [a-zA-Z0-9-]+ to match the hostname or IP address, and \.[a-zA-Z0-9-]+ to match the path. The + following the square brackets specifies that one or more characters in the specified range (i.e. a-z, A-Z, 0-9, or -) must be matched.

1763:

What the mean of different symbols like ^ \$ in Regular Expression?

ANSWER:

^a Search all string which start with "a".

a\$ Search all string which end with "a".

[a-zA-Z0-9]{2,5} there should be 2-5 string which can be a-z and 0-9 characters.

[a-zA-Z]{2,5} there should be 2-5 string which are a-z characters.

[0-9]{2,5} there should be 2-5 string which are 0-9 characters.

1764. How do you Improve the performance of regular expressions?

These are five regular expression techniques that can dramatically reduce processing time:

- 1) Character classes
- 2) Possessive quantifiers (and atomic groups)
- 3) Lazy quantifiers
- 4) Anchors and boundaries
- 5) Optimizing regex order

We can improve the performance of regular expressions in following ways mentioned:

1. When you require parentheses but not capture, use non-capturing groups.
2. Do a quick check before attempting a match, if the regex is very complex, e.g.
Does an email address contain '@'?
3. Present the most likely option(s) first, e.g.
light green|dark green|brown|yellow|green|pink leaf
4. Minimize the amount of looping
\d\d\d\d\d is faster than \d{6}
aaaaaa+ is faster than a{6,}
- 5 . Avoid obvious backtracking, e.g.
Mr|Ms|Mrs should be M(?:rs?)
Good night|Good morning should be Good (?:night|morning)

1765. How do you match starting and ending characters in a string using a regular expression?

To match the starting and ending characters in a string using a regular expression, you can use the '^' (caret) and '\$' (dollar sign) symbols, respectively.

The '^' symbol matches the position at the start of the string.

The '\$' symbol matches the position at the end of the string.

1766. How do you match a specific number of occurrences of a pattern?

To match a specific number of occurrences of a pattern in a regular expression, you can use quantifiers. There are several quantifiers available:

'{n}' matches exactly 'n' occurrences of the preceding pattern. For example, 'a{3}' matches the string ""aaa"".

'{n,}' matches n or more occurrences of the preceding pattern. For example, 'a{3,}' matches the string ""aaaaaaaa"".

'{m,n}' matches at least 'm' and at most 'n' occurrences of the preceding pattern. For example, 'a{2,4}' matches the strings ""aa"", ""aaa""", or ""aaaa""".

For example, to match a string that has exactly three digits, you could use the pattern:
\d{3}

1767. How do you match overlapping patterns in a string?

To match overlapping patterns in a string using regular expressions, you need to use a positive lookahead assertion. A positive lookahead '(?=...)' asserts that the pattern inside the parentheses should be present, but it should not consume any characters. In other words, it only matches the pattern without including it in the final match result.

For example, to match overlapping occurrences of the word ""cat"" in the string ""catcatcat""", you could use the pattern:

(?=cat)cat

This pattern matches all three occurrences of the word ""cat"", even though they overlap. The positive lookahead assertion ensures that each occurrence of ""cat"" is only matched once, even though it appears multiple times in the string.

1768. How do you perform a negative lookahead or look behind match in a regular expression?

Lookbehind has the same effect, but works backwards. It tells the regex engine to temporarily step backwards in the string, to check if the text inside the lookbehind can be matched there. (?<!a)b matches a ,Äub,Äù that is not preceded by an ,Äúa,Äù, using negative lookbehind.

Negative Lookahead: To match a pattern only if it is not followed by another pattern, you can use a negative lookahead assertion. For example, to match ""dog"" only if it is not followed by ""s"", you can use the pattern

dog(?!s)

Negative Lookbehind: To match a pattern only if it is not preceded by another pattern, you can use a negative lookbehind assertion. For example, to match ""dog"" only if it is not preceded by ""the""", you can use the pattern:

(?<!the)dog

1769. How do you match a pattern that is separated by a specific character or string?

To match a pattern that is separated by a specific character or string, you can use a separator pattern, such as a dot '.', with a character class '[]' or a negated character class '[^]'.

For example, to match words separated by a comma in a string, you can use the pattern:
\\w+\\b(,\\w+\\b)*

This pattern matches one or more words, separated by commas, surrounded by word boundaries. The word boundaries ensure that the pattern only matches complete words and not substrings of words. The '*' at the end of the pattern makes the separator optional, so it will also match strings with only one word.

If you want to match words separated by any character except a comma, you can use a negated character class:

\\w+\\b[^,]*

This pattern matches one or more words, separated by any character except a comma, surrounded by word boundaries.

1770. How do you match a pattern that is optionally present in a string?

To match a pattern that is optionally present in a string, you can use the '?' (question mark) quantifier. The '?' quantifier matches zero or one occurrence of the preceding pattern.

For example, to match a string that has an optional ""www."" prefix, you could use the pattern: www\.(.+)?

This pattern matches any string that starts with ""www."" followed by any number of characters. The parentheses around '.+' make the pattern inside optional, so the pattern will also match strings that only contain ""www."" . The '.+' matches any character (including spaces) one or more times.

1771. How do you extract a specific portion of a matched pattern?

To extract a specific portion of a matched pattern, you can use capturing parentheses '(...)' . Capturing parentheses create a capture group that can be referenced in the final match result.

For example, to extract the domain name from an email address, you can use the pattern:
\b([a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,})\b

This pattern matches an email address surrounded by word boundaries. The capturing parentheses around '([a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,})' create a capture group that matches the entire email address. The final match result will contain the email address in the first capture group.

Different regular expression implementations may have different methods for accessing the capture groups in the final match result. In some programming languages, you can access the capture groups by indexing the match result object, while in others you may need to use a separate function or method to extract the capture groups.

1772. What is the simple regular expression for a two-digit decimal?

\d+(d{1,2})?

" "Write a rex command to match the IP address like 120.456.348.670.

^(?:[0-9]{1,3}\.){3}[0-9]{1,3}\$

This pattern uses the following elements:

- 1 '^' matches the beginning of the string
2. '(?:[0-9]{1,3}\.){3}' matches three sets of 1 to 3 digits followed by a dot. The ?: in the parentheses makes the capture non-capturing, meaning that the match will not be stored in a separate capture group.
3. `'[0-9]{1,3}\$` matches 1 to 3 digits
4. '\$' matches the end of the string

This pattern will match IP addresses like 120.456.348.670 but not addresses like 120.456.348.670.1 or 12.456.348.67"

1773.

Write a regex to split String by a new line.

```
String lines[] = string.split("\r?\n");
```

1774. What is Zero width assertions in Regular expressions?

These are used for matching the position rather than text in the line.

\A - Matching the starting of the line.

\Z - Matching the end of the line.

1774.

How to replace all non-alphanumeric characters with empty strings?

```
replaceAll("[^A-Za-z0-9]", "");
```

1775. How can you match a string that does not contain a specific pattern in a regular expression?

If you want to match a string that doesn't have a specific pattern, you can use a negative look-ahead assertion. This is a special type of expression that you put inside your regular expression. It doesn't actually match anything in the string you're searching through. Instead, it checks to make sure that a certain pattern isn't present.

You write a negative lookahead assertion like this: `(?!pattern)`. The ""pattern"" part is replaced with the pattern you don't want to match. The expression `(?!pattern)` asserts that the pattern shouldn't be there, but the rest of the regular expression can still match if other parts match.

For example, if you want to find a string that doesn't have the word ""pattern"" in it, you can use this expression: `^(?!pattern).)*$`. The `.` part matches any characters in the string, and the `^` and `$` symbols make sure that the string starts and ends with the expression you've written.

1776. Can you explain how backreferences work in regular expressions?

Backreferences allow you to match patterns based on previous matches in the expression. They are represented by a backslash `()` followed by a number.

1777. Mention one way how can regular expressions be optimized for extracting a substring from non-consistent format of string?

One can avoid using nested subqueries, as nested subqueries can result in a large number of database calls, making the query slow and inefficient.

1778. How can you identify slow or inefficient regular expressions in SQL?

By testing regular expressions with sample data, it can help you identify slow or inefficient patterns (or) query execution plans can provide information about the performance of individual components of a query, including the cost of executing regular expressions.

1779. Why are REGEXPs' are considered to computationally expensive?

Regular expressions can be computationally expensive due to the complex nature of pattern matching, the use of backreferences, nested patterns, and process very large datasets.

1780. How do you use the tilde operator (~) or equivalent operator in SQL to match regular expressions against text data?

The tilde operator `(~)` is used in SQL to match regular expressions against text data. It can be used in a WHERE clause in a SQL query to filter rows based on a regular expression pattern. Tilde operator `(~)` is limited to PostgreSQL, for oracle we use `REGEXP_LIKE`.

1781. What are the factors that affect the performance of regular expressions in SQL?

Complexity of the pattern, choice of the regular expression engine, length of the input string.

1782. Can you explain the difference between greedy and non-greedy matching in regular expressions?

Greedy regular expressions will match as much of the input string as possible, while in contrast, non-greedy matching attempts to match the smallest possible portion of the input string. This can be achieved by using the ""?"" character immediately after a quantifier.

1783. Can non-greedy matching in regular expressions reduce the computational expense of the SQL query?

Non-greedy matching in regular expressions may reduce computational expense in some cases, but it depends on the specific use case and the complexity of the regular expression. When a regular expression is applied to a large input string, the engine will attempt to match the entire string using the pattern defined in the regular expression. In some cases, this can result in a large number of calculations and increased computational expense.

1784. What is the simple regular expression for a two-digit decimal or a decimal with a precision of 2 ?

\d+(\.\d{1,2})?

1785. How to search for all text files from a system ?

*.txt

Go to search and type *.txt which means select all the files which has extension .txt

1786. Can we give regular expression range in descending order like [30-5] ?

No , as out of range expression error occurs.

1787. What is the best way to match the term TheDataMonk at the start of a line?

^TheDataMonk\b

1787. Write simple regex expression to match the date pattern (like 23.08.2023, 07/04/2016, 12-08-2002) ?

(\d{2}).(\d{2}).(\d{4})

1788. Try naming more than one animal whose name consists of 3 letters and the middle letter is the vowel a in regex form ?

[bcr]at or (b|c|r)at

1788. Write regex expression for 10 digit mobile number starting with 8 or 9 ?

[8-9][0-9]{9}

1789. Write a regex expression with first character in uppercase and rest all characters in lower case with only one digit allowed in between ?

[A-Z][a-z]*[0-9][a-z]*

1790. Write a simple regular expression to capture the content of each line, without the extra whitespace ?

^\s*(.*\s*)\$

1791. Regex expression for trying to extract the protocol, host and port from any kind of general URL ?

(\w+://([\w\-\.\.]+)(:(\d+))?)

Python

1792. Write a Python program to check that a string contains only a certain set of characters (in this case a-z, A-Z and 0-9).

```
s1 = ""asjdJASDJA123asd123"""
s2 = ""&%576%&%7^$%#$%#)"""

def char_check(test_string):
    pattern = re.compile(r"""\b[a-zA-Z0-9]+\b""")
    check_str = pattern.search(test_string)
    return not bool(check_str)
```

1793. Write a Python program that matches a string that has an a followed by zero or more b's.

```
def custom_string_match(test_string):
    pattern = re.compile(r"""\b[a(b*)]+\b""")
    if pattern.search(test_string):
        return "match found"
    else:
        return "match not found"
```

1794. Write a Python program that matches a string that has an a followed by one or more b's.

```
def custom_string_match1(test_string):
    pattern = re.compile(r"""\^a(b+)""")
    if pattern.search(test_string):
        return ""match found"""
    else:
        return ""match not found"""


```

1795. Write a Python program to remove leading zeros from an IP address.

```
s1 = """192.09.08.0200"""
def custom_string_sub(ip_add):
    pattern = re.compile(r"""\.[0]*""")
    return pattern.sub("".", ip_add)


```

1796. Write a Python program to search the numbers (0-9) of length between 1 to 3 in a given string.

```
def custom_string_match14(test_string):
    pattern = re.compile(r"""\[0-9\]\{1\}|\[0-9\]\{2\}|\[0-9\]\{3\}""")
    if pattern.search(test_string):
        return ""match found"""
    else:
        return ""match not found"""


```

1797. Write a Python program to search some literals strings in a string.

```
def custom_string_match15(test_string, input_string, *string_pattern):
    pattern = [i for i in string_pattern]
    for i in pattern:
        if re.search(input_string ,test_string):
            return ""match found"""
        else:
            return ""match not found"""
    s1 = """I AM EATING FOOD NOW"""


```

1798. Write a Python program to search a literals string in a string and also find the

location within the original string where the pattern occurs.

```
def custom_string_match16(test_string, input_string, *string_pattern):
    pattern = [i for i in string_pattern]
    for i in pattern:
        match = re.search(input_string ,test_string)
        if match:
            start_position = match.start()
            end_position = match.end()
            return "" "match found at position ({}) to ({})" ".format(start_position, end_position)
        else:
            return "" "match not found"
s1 = "" "I AM EATING FOOD NOW""
```

1799. Write a Python program to find the substrings within a string.

```
def custom_string_match17(test_string, substring):
    for match in re.findall(substring, test_string):
        if match:
            return "" "match found""
```

1800. Write a Python program to replace whitespaces with an underscore and vice versa.

```
def replace_custom(test_string):
    if "" "" in test_string:
        return re.sub(" " " ", "_ ", test_string)
    else:
        return re.sub("_ ", " " " ", test_string)
```

1801. Write a Python program to match if two words from a list of words starting with letter 'P'.

```
def letter_checker(test_string):
    for word in test_string:
        match = re.match("^(P\w+)\W(P\w+)", word)
        if match:
            print(match.groups())
    l_word = ["Python PHP", "Java JavaScript", "c c++"]
```

1802. What are the most commonly used metacharacters in regular expressions?

- . (dot): matches any single character except newline
- ^ (caret): matches the start of a line or string

\$ (dollar sign): matches the end of a line or string
(asterisk): matches zero or more occurrences of the preceding character
(plus sign): matches one or more occurrences of the preceding character
? (question mark): matches zero or one occurrence of the preceding character
[] (square brackets): matches any single character from the enclosed set of characters
() (parentheses): groups characters together to create a subpattern
| (pipe): matches either the pattern on the left or the pattern on the right

1803. How do you match a specific character or sequence of characters using regular expressions?

To match a specific character or sequence of characters using regular expressions, you can simply include the characters in the pattern. For example, the pattern "hello" matches the string "hello".

1804. How do you match a range of characters using regular expressions?

To match a range of characters using regular expressions, you can use square brackets and a range of characters or character classes. For example, the pattern [a-z] matches any lowercase letter from a to z, and the pattern [0-9] matches any digit.

1805. How do you match repeating characters using regular expressions?

To match repeating characters using regular expressions, you can use the * or + metacharacters. The * matches zero or more occurrences of the preceding character, while the + matches one or more occurrences of the preceding character. For example, the pattern "ab*" matches "a", "ab", "abb", "abbb", and so on.

1806. How do you match multiple patterns using regular expressions?

To match multiple patterns using regular expressions, you can use the | (pipe) metacharacter to specify alternation. For example, the pattern "hello|world" matches either "hello" or "world".

1807. How do you match substrings using regular expressions?

To match substrings using regular expressions, you can use capture groups, which are defined by enclosing the pattern in parentheses. For example, the pattern "hello (world)" matches the string "hello world", and captures the substring "world".

1808. What is the difference between greedy and non-greedy matching in regular expressions?

Greedy matching and non-greedy matching are different ways of matching patterns in regular expressions. Greedy matching matches the longest possible substring that matches the pattern, while non-greedy matching matches the shortest possible substring that matches the pattern. Greedy matching is the default behavior in most regular expression engines, but you can make a quantifier non-greedy by adding a ? after it.

1809. How do you perform case-insensitive matching using regular expressions?

To perform case-insensitive matching using regular expressions, you can use the (?i) flag at the beginning of the pattern. For example, the pattern "(?i)hello" matches "hello", "Hello", "HELLO", and so on.

1810. How do you capture and extract parts of a string using regular expressions?

To capture and extract parts of a string using regular expressions, you can use capture groups, which are defined by enclosing the pattern in parentheses. The captured groups can be accessed using backreferences. For example, the pattern "(\\d{3})-(\\d{2})-(\\d{4})" captures a social security number in the format "123-45-6789", and the groups can be accessed using \1, \2, and \3.

1811. How do you validate and parse data using regular expressions, such as email addresses or phone numbers?

To validate and parse data using regular expressions, you can create a pattern that matches the valid format of the data. For example, the pattern "`^([a-zA-Z0-9_.+-]+)@([a-zA-Z0-9-]+\.[a-zA-Z0-9-]+\.)+$`" matches a valid email address.

1812. How do you use regular expressions with programming languages such as Python, Java, or JavaScript?

Regular expressions can be used with programming languages such as Python, Java, or JavaScript by calling the appropriate regular expression functions or methods provided by the language. These functions or methods typically take a regular expression pattern and a string to match

Chapter 21 - CLTV

1813. What is CLTV and why is it important for businesses?

CLTV (Customer Lifetime Value) is a measure of the total value that a customer is expected to generate for a business over the entire lifetime of their relationship with the company. It is an estimate of the total amount of money that a customer will spend on the company's products or services, minus the costs of acquiring and retaining the customer.

CLTV is important for businesses because it provides a more comprehensive view of customer value than other metrics like customer acquisition cost (CAC) or average order value (AOV). By understanding CLTV, companies can make informed decisions about how to allocate their resources, such as sales and marketing budget, and how to prioritize their customer engagement and retention initiatives. CLTV can also be used to measure the return on investment (ROI) of these initiatives, and to forecast future revenue and growth potential. In short, CLTV is an important tool for optimizing customer engagement and maximizing the value of customer relationships.

1814. How do you calculate CLTV, and what factors do you need to consider?

CLTV is typically calculated as the product of the average value of a customer's purchase, the average number of purchases per year, and the average customer lifespan. The formula for CLTV can be expressed as follows:

$$\text{CLTV} = \text{Average Value of a Purchase} \times \text{Average Number of Purchases per Year} \times \text{Average Customer Lifespan}$$

In order to calculate CLTV, you need to consider the following factors:

Average value of a purchase: This is the average amount of money that a customer spends on a single transaction.

Average number of purchases per year: This is the average number of transactions that a customer makes per year.

Average customer lifespan: This is the average length of time that a customer remains active with a company.

In order to estimate these values, you will need to collect data on customer behavior and purchase history, and you may need to make assumptions about future customer behavior based on past trends and patterns.

It is also important to note that CLTV is an estimate and not an exact value, and it will change over time as customer behavior and business conditions change. As such, companies should

regularly review and update their CLTV calculations to ensure that they have an accurate and up-to-date understanding of customer value.

1814. What is the difference between CLTV and customer acquisition cost (CAC)?

CLTV (Customer Lifetime Value) and customer acquisition cost (CAC) are both important metrics for businesses, but they measure different things and are used for different purposes.

CLTV is a measure of the total value that a customer is expected to generate for a business over the entire lifetime of their relationship with the company. CLTV helps companies understand the long-term value of customer relationships, and can be used to prioritize customer engagement and retention initiatives.

Customer acquisition cost (CAC), on the other hand, is a measure of the cost of acquiring a new customer. CAC is calculated as the total cost of sales and marketing efforts divided by the number of new customers acquired. CAC is important because it provides a measure of the cost efficiency of customer acquisition efforts, and helps companies understand the trade-off between customer acquisition and customer retention.

In summary, CLTV provides a measure of the long-term value of customer relationships, while CAC provides a measure of the cost efficiency of customer acquisition efforts. Both metrics are important for businesses, and they should be used together to make informed decisions about customer engagement and resource allocation.

1815. How do you measure and track CLTV over time, and what tools or techniques do you use?

Collect data: Collect data on customer behavior, purchase history, and customer demographic information. This data can be used to calculate the average value of a customer's purchase, the average number of purchases per year, and the average customer lifespan, which are all factors that contribute to CLTV.

Calculate CLTV: Use the formula for CLTV, which is: $CLTV = \text{Average Value of a Purchase} \times \text{Average Number of Purchases per Year} \times \text{Average Customer Lifespan}$.

Monitor CLTV over time: Regularly monitor CLTV to track changes in customer behavior and to ensure that the CLTV calculation remains accurate over time.

Use customer relationship management (CRM) software: CRM software can be used to automate the process of collecting, storing, and analyzing customer data, and can help to provide real-time insights into CLTV.

Analyze trends and patterns: Analyze trends and patterns in CLTV over time, and use this information to inform customer engagement and retention initiatives.

Use predictive analytics: Predictive analytics tools can be used to analyze customer data and predict future customer behavior, which can be used to estimate future CLTV.

In summary, CLTV can be measured and tracked over time by collecting and analyzing customer data, using customer relationship management software, and using predictive analytics tools. These tools and techniques can help companies understand the long-term value of customer relationships and make informed decisions about customer engagement and retention.

1816. How do you use CLTV to inform your marketing and customer retention strategies?
CLTV (Customer Lifetime Value) provides valuable insights into the long-term value of customer relationships, and can be used to inform marketing and customer retention strategies in the following ways:

Customer Segmentation: CLTV can be used to segment customers into different groups based on their expected lifetime value, and to prioritize customer engagement and retention efforts accordingly. For example, high-CLTV customers can be targeted with personalized marketing campaigns, while lower-CLTV customers may receive more generic marketing messages.

Customer Retention: CLTV helps companies understand the value of customer relationships, and can be used to prioritize customer retention initiatives. For example, companies may offer special incentives to high-CLTV customers to encourage them to remain loyal, or invest in customer service programs to improve the customer experience and reduce customer churn.

Marketing Spend: CLTV can be used to inform decisions about marketing spend, by providing a measure of the long-term value of customer relationships. For example, companies may choose to allocate more marketing resources to acquiring high-CLTV customers, or to customer retention initiatives aimed at high-CLTV customers.

Lifetime Value-Based Pricing: CLTV can be used to inform pricing strategies, by setting prices based on the expected lifetime value of a customer. For example, companies may charge a premium for products or services that are expected to generate high CLTV, or offer discounts to customers who are expected to generate lower CLTV.

In summary, CLTV provides valuable insights into the long-term value of customer relationships, and can be used to inform marketing and customer retention strategies by guiding customer segmentation, customer retention, marketing spend, and lifetime value-based pricing.

1817. How do you use CLTV to prioritize and allocate resources, such as sales and marketing budget?

CLTV (Customer Lifetime Value) can be used to prioritize and allocate resources, such as sales and marketing budget, in the following ways:

Targeting High-CLTV Customers: CLTV provides a measure of the expected value of a customer relationship, which can be used to prioritize sales and marketing efforts. For example, companies may choose to allocate more resources to acquiring and retaining high-CLTV customers, as they are likely to generate more revenue over the long-term.

Budget Allocation: CLTV can be used to inform decisions about budget allocation between customer acquisition and customer retention initiatives. For example, companies may choose to allocate more budget to acquiring high-CLTV customers, or to customer retention initiatives aimed at high-CLTV customers.

Channel Optimization: CLTV can be used to optimize marketing spend across different channels. For example, companies may choose to allocate more budget to channels that are most effective at acquiring high-CLTV customers, or to channels that are most effective at retaining high-CLTV customers.

Customer Segmentation: CLTV can be used to segment customers into different groups based on their expected lifetime value, and to prioritize customer engagement and retention efforts accordingly. For example, high-CLTV customers can be targeted with personalized marketing campaigns, while lower-CLTV customers may receive more generic marketing messages.

In summary, CLTV provides valuable insights into the long-term value of customer relationships, and can be used to prioritize and allocate resources by guiding targeting, budget allocation, channel optimization, and customer segmentation.

1818. How do you use CLTV to evaluate the impact of customer churn, upsells, and cross-sells on the business?

CLTV, or customer lifetime value, is a metric that helps businesses evaluate the economic value a customer will bring to the company over their lifetime. It takes into account factors such as customer acquisition costs, average purchase value, and the average number of purchases made by a customer over a given time period. By using CLTV, businesses can better understand the impact of customer churn, upsells, and cross-sells on their overall business performance.

To use CLTV to evaluate the impact of customer churn, you would calculate the average amount of money a customer spends with your company over a set period of time, and then subtract the cost of acquiring that customer. If the CLTV is positive, it means that the customer is generating more revenue than it costs to acquire them, and if it's negative, the opposite is true.

The impact of customer churn on CLTV can be significant, as it represents the loss of revenue from customers who no longer make purchases from your company. To mitigate the impact of churn, businesses can focus on retaining customers through loyalty programs, customer service initiatives, or targeted marketing campaigns.

Upselling and cross-selling, on the other hand, can have a positive impact on CLTV. By upselling customers to higher-priced products or services, businesses can increase the average purchase value, and by cross-selling complementary products, they can increase the frequency of purchases. Both of these strategies can lead to an increase in the overall CLTV for a customer.

In conclusion, CLTV is a valuable metric for businesses to evaluate the impact of customer churn, upsells, and cross-sells on their overall performance. By taking into account the cost of acquiring customers, the average purchase value, and the average number of purchases made, CLTV provides a clear picture of the financial value a customer brings to a business over their lifetime.

1819. How do you use CLTV to identify and target high-value customers, and what are the best practices for doing so?

CLTV can be used to identify and target high-value customers by calculating the expected economic value a customer will bring to a business over their lifetime. Customers with a higher CLTV are generally more profitable and have a greater potential to drive long-term revenue growth.

Here are some best practices for using CLTV to identify and target high-value customers:

Regularly calculate CLTV: In order to effectively use CLTV, it's important to regularly calculate the metric for each customer in your customer base. This allows you to track changes in customer value over time and identify patterns and trends.

Segment customers based on CLTV: By segmenting customers based on their CLTV, you can better understand the different types of customers you have and tailor your marketing and customer service efforts accordingly.

Personalize communication: High-value customers are often more engaged and responsive to personalized communication. By using CLTV to identify these customers, you can create tailored marketing campaigns and offers that will be more likely to drive additional sales.

Invest in customer retention: High-value customers are more valuable in the long-term, so it's important to invest in retention efforts to keep them coming back. This can include loyalty programs, targeted customer service initiatives, or personalized promotions.

Monitor changes in CLTV: Regularly monitoring changes in CLTV allows you to identify potential issues early on and take proactive steps to address them. For example, if you notice that a high-value customer's CLTV is declining, you can take action to improve their experience and retain their business.

Analyze customer behavior: Understanding the behavior of your high-value customers can provide valuable insights into what drives their purchasing decisions. This information can be used to develop targeted marketing campaigns, optimize your product offerings, and improve your overall customer experience.

By using CLTV to identify and target high-value customers, businesses can focus their efforts on the customers that are most likely to drive long-term revenue growth and profitability. This can help improve overall business performance and drive long-term success.

1820. How do you use CLTV to forecast future revenue and growth potential, and what challenges do you face in doing so?

CLTV can be used to forecast future revenue and growth potential by estimating the economic value a customer will bring to a business over their lifetime. By calculating the average purchase value, average purchase frequency, and customer lifespan, businesses can project the revenue that a customer will generate in the future. This information can then be used to forecast overall revenue and growth potential for the business.

Here are some of the challenges faced when using CLTV to forecast future revenue and growth potential:

Data accuracy: CLTV calculations rely on accurate data, so it's important to ensure that the data used in the calculation is accurate and up-to-date. This can be a challenge, particularly if the business has a large customer base or if customer data is spread across multiple systems.

Customer churn: CLTV calculations assume that customers will continue to make purchases at the same rate over their lifetime, but customer churn can impact the accuracy of these projections. Businesses need to factor in churn rates when making CLTV-based forecasts.

Changing customer behavior: Customer behavior can change over time, which can impact the accuracy of CLTV projections. Businesses need to monitor changes in customer behavior and make adjustments to their CLTV projections accordingly.

Uncertainty in customer lifespan: CLTV projections are based on an estimated customer lifespan, which can be difficult to predict with certainty. This can lead to uncertainty in CLTV projections, particularly if customers are relatively new.

Dynamic market conditions: Changes in market conditions, such as changes in competition or economic conditions, can impact the accuracy of CLTV projections. Businesses need to factor in these dynamic conditions when making CLTV-based forecasts.

In conclusion, while CLTV can be a useful tool for forecasting future revenue and growth potential, it's important to understand and manage the challenges that come with using this metric. By regularly monitoring and adjusting CLTV projections, businesses can improve the accuracy of their forecasts and make more informed decisions about their future growth and revenue potential.

1821. How do you use CLTV to measure the return on investment (ROI) of your customer engagement and retention initiatives?

CLTV can be used to measure the return on investment (ROI) of customer engagement and retention initiatives by comparing the expected economic value a customer will bring to a business over their lifetime with the cost of retaining them.

Here's how you can use CLTV to measure the ROI of customer engagement and retention initiatives:

Calculate CLTV: Start by calculating the CLTV of your customer base, taking into account the average purchase value, average purchase frequency, and customer lifespan. This will give you a baseline for the expected economic value of each customer.

Determine engagement and retention costs: Next, determine the cost of your customer engagement and retention initiatives, including the cost of customer service, loyalty programs, targeted marketing campaigns, and any other initiatives designed to retain customers.

Compare CLTV and engagement/retention costs: Compare the CLTV of your customer base with the cost of retaining them. If the CLTV of a customer exceeds the cost of retaining them, then the engagement and retention initiatives are considered to be cost-effective and generating a positive ROI.

Monitor changes in CLTV: Regularly monitor changes in CLTV and engagement and retention costs over time. This will help you track the effectiveness of your initiatives and make changes as needed to optimize your ROI.

By using CLTV to measure the ROI of customer engagement and retention initiatives, businesses can make more informed decisions about their customer retention strategies and allocate resources more effectively. This can help improve customer satisfaction, increase customer loyalty, and drive long-term revenue growth and profitability.

1822. How do you use CLTV to measure the impact of customer experience and customer satisfaction on the business?

CLTV can be used to measure the impact of customer experience and customer satisfaction on the business by analyzing the relationship between customer experience, customer satisfaction, and customer lifetime value.

Here's how you can use CLTV to measure the impact of customer experience and customer satisfaction on your business:

Measure customer satisfaction: Start by measuring customer satisfaction levels through customer surveys, focus groups, or other methods. This will provide a baseline for understanding customer experience and satisfaction levels.

Analyze CLTV by customer satisfaction: Next, segment your customer base by satisfaction level and calculate the CLTV of each segment. This will help you determine if customers who are more satisfied with their experiences have a higher CLTV compared to those who are less satisfied.

Monitor changes in customer experience and satisfaction: Regularly monitor changes in customer experience and satisfaction levels, and compare this with changes in CLTV. This will help you understand the relationship between customer experience, customer satisfaction, and CLTV, and identify areas where you can improve the customer experience to increase customer satisfaction and CLTV.

Implement changes: Based on the findings from your analysis, implement changes to improve the customer experience and increase customer satisfaction. Monitor changes in CLTV over time to assess the impact of these changes on customer lifetime value.

By using CLTV to measure the impact of customer experience and customer satisfaction on the business, businesses can make data-driven decisions about their customer engagement and retention strategies and prioritize investments that will improve the customer experience and drive long-term revenue growth.

1823. How do you use CLTV to evaluate the effectiveness of your customer loyalty and retention programs, and what metrics do you use to do so?

CLTV can be used to evaluate the effectiveness of customer loyalty and retention programs by comparing the expected economic value a customer will bring to a business over their lifetime with the cost of retaining them through these programs.

Here's how you can use CLTV to evaluate the effectiveness of customer loyalty and retention programs:

Calculate CLTV: Start by calculating the CLTV of your customer base, taking into account the average purchase value, average purchase frequency, and customer lifespan. This will give you a baseline for the expected economic value of each customer.

Determine program costs: Next, determine the cost of your customer loyalty and retention programs, including the cost of customer service, loyalty programs, targeted marketing campaigns, and any other initiatives designed to retain customers.

Compare CLTV and program costs: Compare the CLTV of your customer base with the cost of retaining them through your loyalty and retention programs. If the CLTV of a customer exceeds the cost of retaining them, then the programs are considered to be cost-effective and generating a positive return on investment (ROI).

Monitor changes in CLTV: Regularly monitor changes in CLTV over time and compare this with changes in program costs. This will help you track the effectiveness of your loyalty and retention programs and identify areas where you can optimize the programs to increase ROI.

To evaluate the effectiveness of customer loyalty and retention programs, it's important to use metrics such as customer retention rate, customer lifetime value, and customer acquisition cost to track the success of these programs. These metrics can help you understand the impact of your loyalty and retention programs on customer engagement, customer satisfaction, and long-term revenue growth.

By using CLTV to evaluate the effectiveness of customer loyalty and retention programs, businesses can make data-driven decisions about their customer engagement and retention strategies and allocate resources more effectively to drive long-term revenue growth and profitability.

1824. What is the formula for calculating CLTV

CLTV is calculated as follows:

$$\text{CLTV} = ((T * \text{AOV}) * \text{AGM}) / \text{ALT}$$

where

T is - average monthly transactions

AVPT = average value per transaction

ALT = Average customer lifespan

AGM = Average gross margin

One more formula for CLTV

$$\text{CLV} = (\text{Average Sales} * \text{Purchase Frequency}) / \text{Churn} * \text{Profit Margin}$$

Where,

Average Sales = (TotalSales) / (Total no. of orders)

Purchase Frequency = (Total no. of orders) / (Total unique customers)

Retention rate = (Total no. of orders greater than 1)/(Total unique customers)

Churn =1- Retention rate

Profit Margin = Based on business context

1825. How to Maximize Your CLTV?

Delivering more value than your competitors is the key to maximizing customer lifetime value. To do this, you need to focus on providing customers with a unique experience that they can't find elsewhere. Additionally, you must provide timely and accurate information, stay current with changing trends, and offer a variety of products and services. Finally, keep your prices reasonable so that customers feel they are getting a good deal.

1826. What are the Benefits of Maximizing Your CLTV?

There are many benefits to maximizing customer lifetime value (CLTV). First, increased CLTV leads to increased revenue and profitability. Second, CLTV is a key metric for competitive differentiation. Third, CLTV can help you attract and retain high-quality customers. Finally, CLTV is an important factor in employee retention and motivation.

1827. Applications of CLV

Boosting Retention and Loyalty

Forecasting Demand and Sales

Segregating Customers

1828. Strategies to Increase CLV

Up-sells and Cross-sells

Product Education

Add a Personal Touch

Add Sticky Features to the Product

Reassess your Value Proposition

Win them Over with your Customer Support

1829. Types of Customer Lifetime Value

There are various ways to calculate CLV, and you can divide them into these two types:

Historical Customer Lifetime Value:

Historical CLV calculates a customer's lifetime value depending on what the customer has spent with a business. There are two ways of calculating historical CLV: using ARPU and using cohort analysis.

Predictive Customer Lifetime Value:

The predictive CLV model is more complicated than its historical counterpart. Predictive CLV uses customers' historical behavior and the predicted retention to estimate future customer lifetime and revenue by applying artificial intelligence and machine learning.

This model, by nature, takes into account customer trends and behaviors throughout the lifetime of customers.

1830. How to Improve Your Purchase Frequency?

Maybe your customers spend a lot, and you, are making a great margin, but they just don't order very often. Try a few of these campaigns to maximize profit by getting your customers to pick up the pace:

Communicate dynamically. Using historical customer data collected with a true unified Single Customer View, you can automatically send emails that arrive at the ideal time for each individual customer. (Make sure those emails are personalized!).

Find out if you are communicating through the right channels does this particular customer respond better to email or SMS? Make sure each customer receives your message through their preferred channel.

Segment your customers by their customer lifecycle stage, then re-engage with customers who haven't purchased in a while, and are in danger of churning.

Use push notifications and banners to highlight time-sensitive deals based on their browsing history.

Use banners that trigger when a customer enters and exits the site. Use these banners to recommend personalized products or sales.

1831. How to Improve Your Customer Lifetime Value (CLTV)

Having trouble getting your customers to increase their spending? Try these campaigns, which focus on providing incentives to increase average order value:

Add personalized product recommendations to your website. The recommended products should be based on the ideal price point for each individual customer, thereby maximizing revenue.

Send personalized newsletter campaigns with dynamic product recommendations optimized for price.

Trigger product recommendations, based on what they have added to their shopping cart, directly on the site.

Send an email campaign with product recommendations based on what they have added to their shopping cart.

Create product bundles that offer a discount for making a larger purchase. Bundle products that can be used together, and recommend the bundle directly on the site, or through email, based on the user's browsing and shopping cart history.

Create a customer loyalty program, incentivizing spending by adding loyalty points that customers can use for discounts and freebies.

1832. How do you calculate Customer Lifetime Value (CLTV) in a B2B scenario where the customer acquisition cost is high and the customer relationship spans over several years?

To calculate CLTV in a B2B scenario, you can start by estimating the average revenue per customer per year, subtracting the average cost of serving that customer, and then discounting the result by an appropriate discount rate. The sum of these discounted values represents the CLTV.

1833. How do you take into account customer churn when calculating CLTV?

Customer churn can be taken into account by adjusting the estimated revenue and cost of serving the customer based on the likelihood of the customer churning.

1834. Is it fair to use CLTV at aggregate level or individual level? Why?

CLTV at aggregate level will give an idea of the health of business by one representative value for all the customers. But if there is an outlier or an extreme value, using mean CLTV can be problematic. It might give an over or underestimated CLTV for individual customers.

Customized CLTV for each individual is more efficient and useful. It is better for differentiating between every customer and then use the data in the best way to gain profits and decision making with regards to each customer.

1835. Is CLTV historic or predictive?

At a given point of time, it is desired to have the expected value from a customer, predicted because looking back in time won't be same for different customers given their different time period of acquisition to be able to better predict their behaviour.

1836. How is CLTV helpful for sales management decisions?

CLTV helps startegising sales. It helps in evaluating the factors that influence sales and hence help in acquiring and retaining value customers. For example it may help the sales team to analyse the purchasing pattern, timing pattern and type of product for the customer to be able to speed up the sales and reap out the best. It also helps in better allocation of resources. It helps in long term business growth and improve marketing efforts for customers of high value." "Why customer segmentation necessary in CLTV?

Customer segmentation becomes necessary to segregate High value customers from low value customers to use different communication for better results and judicious use of resources.

1837. What are the methods to improve CLTV?

There are several ways to increase average CLTV and generate more revenue. Put more emphasis on value and quality of the product or service. The customer service should be high end to beat the competition in the market as it might prevent customers from switching brands. Multiple and Omni support channel should be provided to the customers. Meaningful relationships with the high value customers should be made with periodic feedback collection and acting on them. Pain points should be detected alongwith exploring the solution. Personalised treatment goes a long way as it builds within the customers more trust towards the provider. Give incentives to loyal customers. Valuing customers will increase CLTV.

1838. What are the components of CLTV?

CLTV has 3 components, loyalty over time, sales over time and retention cost. " "How can CLTV become useful for product and service decisions?

CLTV can help in weighing the options while creating and developing products and services that are more client oriented and focus on fulfilling their needs for example by studying their pattern and frequency of choosing a product or service.

1839. What is the difference between cost to serve and customer lifetime value ?

When a firm invests in marketing and strategies to acquire and retain customers, it also has to spend money to maintain that customer base. This is called cost to serve. Whereas, customer lifetime value is the revenue that the company makes due to his customers. The comparison

between the two serves as an important metric to decide if the customers are helping the business or not. Cost to serve when less than CLTV, it serves in favour.

1840. What is a cohort model in CLTV?

it is the grouping of customers into different cohorts based on the transaction date, etc., And calculate average revenue per cohort. This method gives CLTV for each cohort. It is time consuming and can help in strategising moves to be administered on customers who are alike hence saving resources.

1841. Techniques to Raise CLV?

Here are some general tactics to raise your CLV:

1. Cross-selling and upselling

Cross-selling entails promoting more related products in addition to core products, whilst upselling refers to offering additional features or product extensions to current clients. Both can raise CLV since they deepen and broaden the consumer relationship. Additionally, selling to an existing client is less expensive than doing so to a new one.

2. Bring Your Own Style

Delivering a fantastic customer experience will be made possible by personalizations and an intuitive user interface. Chargebee, for instance, enables personalised bills to meet the demands of our customers. When you wish to upsell or cross-sell, personalization and knowing your customers will be helpful. The average client lifetime is increased by small gestures like customised in-app messages that make the user feel cherished.

3. Enhance the Product with Sticky Features

Enhance your product with features that make it challenging for customers to leave. For example, a personalised dashboard that the client can create using their data to increase the overall effectiveness of their firm. That kind of thing encourages long-term growth because it makes the user reliant on your product and improves their quality of life. This increases average customer lifespan and encourages client loyalty.

4.Examine your Value Proposition again.

You must follow their development if you want to realize your devoted clients on a deep level. To make sure that your product satisfies the customer's increasing wants, it would be beneficial if you frequently evaluated your value proposition. You might even think about changing your pricing structure or packaging, or bundling or unbundling some features. This calls for some

experimenting, which can only be successful if you are well-aware of your target audience and have statistics to support it.

1842. How much of the CLV has already been seen compared to the expected future revenue/profit?

We must next determine if this is a future forecast made using CLV as of the present or if it is based on data from a different time period. Using solely future-predicted revenue carries the risk of making it impossible to compare the future-predicted revenue of a highly engaged current customer to that of a prior cohort of clients. You can be comparing apples and oranges, for instance, if you compare an active, low-spend customer to an inactive, high-spend customer from a year ago. It makes more sense to calculate expected lifetime value from a variable reference, such as the start of a customer's association with your company.

1843. Is CLV historical, predictive, or both?

Make sure the CLV is forecast and not merely historic if CLV is being delivered at the individual level. Relying on historical profit or revenue has the drawback that you run the risk of misclassifying current clients. If you recently gained a customer, for instance, you cannot compare that customer to one you got a year ago unless you can forecast the behaviour of the current customer over the course of a year.

1844. Can the CLV take into consideration the different client tenures?

Consider calculating CLV for each customer today while projecting that value out one year. You could factor in each customer's complete history when calculating CLV. Given this, comparing consumers who joined five years ago to those who joined two years ago might not make sense. This is due to the fact that the first customer's total time consideration spans six years (five years of history and one year of prediction), but the second customer's spans only three years (two years of history and one year of prediction).

1845. Is it reasonable to include or exclude extremely elderly customers?

The time period you should use to compute lifetime value should also be taken into account. Imagine being a business as old as Coca-Cola. Should you include users who first began eating the beverage in 1920 in your model? Are those clients still important? I've used an extreme example to illustrate my thesis, but what about a business like Apple? Should the CLV model take into account a consumer who joined in order to purchase the original Mac? Although there is no definitive answer to this, we normally advise taking into account a relevant business window that spans at least a few years. This won't put the business and customers at a disadvantage, but it won't help the model comprehend seasonality either.

1846. When to use the simple CLV formula?

The straightforward CLV formula is suitable for usage when:

- * Because customer retention rates are so low (about 50% or less), the majority of client profit contribution is realised in the first few years.
- * Customers' profit contributions are essentially constant over time, meaning that the average customer profitability has not changed significantly over time.
- * The model's data inputs mostly depend on assumptions rather than past consumer data.
- * Instead of a precise measurement, only a rough estimate of client lifetime value is required. The company is unwilling or uneasy about applying discount rates.
- * Customers are extremely profitable, and they quickly recoup their initial investment.

1847. How does profit depend on customer lifetime value?

Marketing mix modelling and marketing return on investment are two marketing analytics indicators that are closely related to customer lifetime value (ROI). Utilizing statistics, marketing mix modelling examines the effects that various combinations of marketing strategies have on sales. Businesses can assess the effectiveness of their spending by looking at the marketing ROI. It is critical to understand how CLV (together with CSAT and CES) influences these areas when allocating resources, conducting gap analyses, and comprehending the behaviour of current and potential customers.

1848. Difference between LTV AND CLTV?

Ans. Customer lifetime value is abbreviated as LTV and CLTV, and both terms fundamentally mean the same thing. Marketers frequently use the terms interchangeably because there is no recognised distinction between them in the industry. In terms of specificity, some people distinguish between CLTV and LTV, with CLTV denoting the value of a specific client over the course of their association with a business and LTV denoting the average customer lifetime value of all current customers. However, without taking into account company-specific variables, this slight degree of differentiation generates essentially the same metric across all businesses, which is why the phrases are regarded to be synonymous.

1849. If the projected CLTV value fluctuates frequently, how can you follow those changes?

Ans. The effect of marketing and other factors on CLTV has been the subject of substantial academic investigation. Anand Bodapati, a professor at the University of California, Los

Angeles, has extensively studied this subject. He contends that depending on outside factors, predictive lifetime value will alter. Therefore, rather than computing projected lifetime value only once for each customer, you could ask your data team to do so continuously. It will be crucial to track the effects of your marketing, price, and new introductions over time on the CLTV figure.

1850. What type of people are most profitable through CLTV analysis?

Since not every consumer is created equal, not every customer will have the same CLL. Do more than just figure out your portfolio's average customer lifetime value; figure out CLTV for each individual customer. Then, aggregate customers based on their shared accounts (e.g., company size, industry, executive sponsor job title) to obtain an average. Your most profitable customers' habits can be found by calculating the average CLTV of each type of customer. For instance, "our CLTV is highest when a consumer technology company's IT director engages us to address a technical difficulty." These patterns might assist you in adjusting your sales and marketing strategies to target prospects and buyer personas that are comparable to your most lucrative clients.

1851. When does the profit margin normally start to fall?

To determine the exact moment at which margins normally begin to drop, plot the client lifetime value across time. You can come up with ideas to increase profitability at this point if your typical client lifespan is three years but your margin declines after two years. If you want to introduce your clients to higher-value services, such as the ones you recognised as your most lucrative projects, this might be the ideal point in the lifecycle. A gap in your service offerings may also become apparent, and you will need to make an investment in innovation to provide services that continue to meet the needs of your customers (and increase your CLTV).

1852. Difference between LTV AND CLTV?

Ans. Customer lifetime value is abbreviated as LTV and CLTV, and both terms fundamentally mean the same thing. Marketers frequently use the terms interchangeably because there is no recognised distinction between them in the industry. In terms of specificity, some people distinguish between CLTV and LTV, with CLTV denoting the value of a specific client over the course of their association with a business and LTV denoting the average customer lifetime value of all current customers. However, without taking into account company-specific variables, this slight degree of differentiation generates essentially the same metric across all businesses, which is why the phrases are regarded to be synonymous.

1853. Is the CLV at the aggregate level or individual level? If it is aggregate, is it the mean or median?

CLV can be calculated at both the aggregate level and the individual level.

At the aggregate level, CLV is calculated as the average or aggregate lifetime value of a group of customers. In this case, the CLV calculation takes into account the average purchase value, average purchase frequency, and average customer lifespan for the entire customer base. The aggregate CLV is usually calculated as the mean of the individual CLV calculations for each customer in the customer base.

At the individual level, CLV is calculated for each customer based on their specific purchase behavior, purchase frequency, and lifespan as a customer. This calculation provides a more precise estimate of the lifetime value of each individual customer.

Whether the mean or median is used depends on the data and the business context. If the data is relatively symmetrical, the mean can be a good representation of the aggregate CLV. However, if the data is heavily skewed, the median may be a more appropriate measure of central tendency as it is not affected by outliers.

In summary, it is important to understand that CLV can be calculated at both the aggregate and individual level and the choice of mean or median depends on the data and the business context.

1854. How much of CLV is already observed vs predicted future revenue/profit?

The calculation of customer lifetime value (CLV) is a combination of observed and predicted future revenue or profit.

The observed component of CLV is based on the historical data and reflects the revenue or profit a customer has already generated for the business. This component of CLV is based on actual purchase behavior and can be easily calculated using a customer's transaction history.

The predicted component of CLV is based on projections of future revenue or profit from a customer. This component of CLV takes into account factors such as expected future purchase behavior, expected customer lifespan, and any other factors that may impact future revenue or profit from a customer.

In general, the observed component of CLV represents a smaller portion of the total CLV calculation, while the predicted component represents a larger portion. The exact balance between the observed and predicted components of CLV will depend on the business context, the data available, and the level of confidence in the projections.

It is important to note that the accuracy of the predicted component of CLV is directly tied to the accuracy of the projections used in the calculation. The more accurate the projections, the more accurate the predicted component of CLV will be. In order to ensure the accuracy of the predicted component, it's important to use reliable data and advanced analytics techniques such as machine learning and predictive modeling.

1855. If the CLV is predictive, how far into the future does that prediction extend?

The prediction of customer lifetime value (CLV) usually extends into the future for the expected lifespan of the customer. The exact length of the prediction will depend on the business context and the data available.

In some cases, the expected customer lifespan may be a few years, while in other cases it may be several decades. The prediction horizon for CLV should be long enough to provide a meaningful estimate of the future revenue or profit that a customer will generate for the business, but not so long that the projections become unreliable or impractical.

It is also important to note that the accuracy of the CLV prediction decreases the further into the future it extends. This is because it is more difficult to predict customer behavior and other factors that may impact future revenue or profit with a high degree of accuracy over a longer period of time.

In general, it's a good practice to regularly review and update the CLV predictions to ensure that they remain accurate and relevant. This can involve using more recent data, adjusting projections based on changes in the business or market environment, and applying advanced analytics techniques such as machine learning and predictive modeling.

1856. Can the CLV account for varying tenure of customers?

Yes, the calculation of customer lifetime value (CLV) can account for varying tenure of customers. CLV is calculated as the present value of the future profit or revenue that a customer is expected to generate for a business over their lifetime. The expected lifespan of a customer is a key factor in the calculation of CLV, and different customers may have different expected lifespans.

In order to account for varying tenure of customers, the calculation of CLV should take into account the expected lifespan of each individual customer. This can be based on historical data or can be estimated using predictive analytics techniques such as machine learning or survival analysis.

By accounting for varying tenure of customers, the CLV calculation provides a more accurate picture of the value of each individual customer to the business. This information can be used to prioritize customer engagement and retention initiatives, target high-value customers, and allocate marketing budgets more effectively.

It is important to note that the accuracy of the CLV calculation will depend on the quality of the data used and the accuracy of the projections used to estimate the expected customer lifespan.

Regularly reviewing and updating the CLV calculation can help to ensure that it remains accurate and relevant.

1857. Does the predictive CLV number change often and, if so, how can you track changes to CLV?

The predictive customer lifetime value (CLV) number can change often, depending on several factors, including changes in customer behavior, market conditions, and the business environment.

To track changes in CLV, businesses can periodically update their CLV calculations using the latest available data and adjust their projections accordingly. This can involve using more recent data to update the expected customer lifespan, updating the customer purchase frequency and average transaction value, and adjusting projections based on changes in the business or market environment.

In addition, businesses can use advanced analytics techniques such as machine learning or predictive modeling to help track changes in CLV over time. These techniques can automatically identify changes in customer behavior and other factors that impact CLV and help businesses to make more accurate projections.

It is important to regularly review and update the CLV calculation in order to ensure that it remains accurate and relevant. This can help businesses to make more informed decisions about customer engagement and retention initiatives, target high-value customers more effectively, and allocate marketing budgets more efficiently.

1858. What do you think about the current maturity level around the Customer Lifetime Value metric?

As a language model AI, I don't have personal opinions, but I can tell you that the maturity level around the customer lifetime value (CLV) metric varies across different industries and organizations.

In general, the use of CLV as a key performance metric has become more widespread in recent years, as businesses recognize the importance of understanding the long-term value of their customers. Many organizations are using CLV to guide customer engagement and retention initiatives, prioritize marketing efforts, and make data-driven decisions about pricing and product development.

However, there is still a significant opportunity for organizations to improve their use of CLV. Some challenges that organizations face when using CLV include access to high-quality data,

difficulty in making accurate predictions, and limitations in the availability of advanced analytics tools and techniques.

In addition, while many organizations are using CLV as a high-level metric to track overall performance, there is still room for improvement in using CLV at a more granular level to inform customer segmentation and targeting strategies, optimize customer journeys, and drive business outcomes.

Overall, the maturity level around CLV is growing, but there is still much potential for further development and innovation in this area.

1859. What are the biggest challenges in using the Customer Lifetime Value (CLV) metric to design business strategy?

The customer lifetime value (CLV) metric can provide valuable insights into the long-term value of customers to a business, but there are several challenges that organizations face when using CLV to design business strategy. Some of the biggest challenges include:

Data quality and availability: CLV calculations rely on accurate and complete data, including customer demographic information, transaction history, and other relevant data. If data is incomplete, inaccurate, or outdated, the resulting CLV calculations may be unreliable.

Predictive accuracy: CLV calculations are based on projections about future customer behavior and revenue, which can be difficult to make accurately. Even small inaccuracies in the projections can have a significant impact on the overall CLV calculation.

Complexity of calculation: CLV calculations can be complex and require significant data processing and analysis, which can be time-consuming and resource-intensive.

Customer heterogeneity: CLV calculations can be challenging when customer behavior and preferences vary widely across different segments of the customer base. This can make it difficult to make accurate projections about future customer behavior and revenue.

Dynamic market and customer behavior: The market environment and customer behavior can change rapidly, making it difficult to make accurate projections about the future. CLV calculations need to be regularly reviewed and updated to ensure that they remain relevant and accurate.

Despite these challenges, CLV remains a valuable tool for organizations that are looking to understand the long-term value of their customers and make data-driven decisions about customer engagement and retention initiatives. By leveraging advanced analytics techniques and investing in the development of high-quality data systems, organizations can overcome these challenges and effectively use CLV to guide business strategy.

1860. What kind of external data can be leveraged to improve the confidence in CLV metric values?

There are several external data sources that can be leveraged to improve the confidence in customer lifetime value (CLV) metric values:

Demographic data: Demographic data, such as age, income, education level, and occupation, can be used to segment the customer base and understand the characteristics of different customer groups. This information can then be used to make more accurate projections about future customer behavior and revenue.

Customer behavior data: Customer behavior data, such as purchase history, product usage patterns, and customer engagement data, can be used to better understand the underlying drivers of customer value. This information can be used to create more accurate CLV projections and improve the accuracy of customer segmentation and targeting strategies.

Market and economic data: Market and economic data, such as consumer sentiment, inflation, and consumer spending patterns, can be used to understand the broader market context in which the business operates. This information can help inform CLV projections and help organizations understand how external factors may impact customer behavior and revenue over time.

Competitor data: Competitor data, such as market share, product offerings, and pricing strategies, can be used to understand the competitive landscape and how it may impact customer behavior and revenue. This information can be used to inform CLV projections and inform customer engagement and retention initiatives.

Technographic data: Technographic data, such as technology usage patterns and adoption rates, can be used to understand how customers are using technology and how it may impact customer behavior and revenue.

By leveraging external data sources, organizations can gain a more comprehensive understanding of customer behavior, market dynamics, and the factors that drive customer value, which can help improve the confidence in CLV metric values. However, it is important to ensure that the data sources used are relevant, high-quality, and up-to-date, as any inaccuracies in the data may impact the accuracy of the CLV calculations.

1861. In order to calculate Customer Lifetime Value (CLV), one needs to have the correct attribution model. In today's world of several consumer touch-points and multi-platform online experience, what are your recommendations to develop a reliable attribution model?

Assume that without a certain marketing investment gross profit is \$100,000, whereas with the investment it is \$200,000 (where the investment of \$50,000 has not been considered in the gross profit calculation). Compute the ROI Of the marketing investments?

The return on investment (ROI) of a marketing investment can be calculated using the following formula:

$$\text{ROI} = (\text{Gain from Investment} - \text{Cost of Investment}) / \text{Cost of Investment}$$

In this case, the gain from the investment is \$200,000 (the increase in gross profit) and the cost of the investment is \$50,000. So, the ROI can be calculated as follows:

$$\text{ROI} = (\$200,000 - \$50,000) / \$50,000 = 4$$

This means that the marketing investment has generated a return of 4 times its cost, or 400% ROI. This indicates that the marketing investment has been highly successful in increasing the gross profit of the business and has been a good investment decision.

Chapter 22 - Shell Scripting

1862. This is the shell script:

- (A) A file with a list of commands
- (B) A file with special symbols
- (C) Group of instructions
- (D) Group of operations

Ans:-(D)

1863. How do you add a new user to your Linux system?

- (A) Using useradd
- (B) Using adduser
- (C) Using Linuxconf
- (D) the entire list

Ans:- (D)

1864. Which shell is the most common and best to use?

- A. Korn shell
- B. C shell
- C. Bourne shell
- D. Bash Shell

Ans:- (D)

1865. With ____ first line of every shell script begins.

- (A). &
- (B). #
- (C). \$
- (D). !

Ans:- (B)

1866. What command allows the creation of environment variables?

- (A) Export
- (B) Set
- (C) Read and Export
- (D) None of the Above

Ans:- (A)

1867. The hash command:-

- (A)Manages an internal hash table
- (B)Finds and remembers the full path name of the specified command
- (C)Shows the number of hits and used command names
- (D)Performs all of the aforementioned functions.

Ans:- (D)

1868. Which option of the kill command sends the given signal name to the specified process?

- a) -l
- b) -n
- c) -s
- d) -a

Ans:- ©

1869. Which command identifies the resource of a command?

- a) type
- b) typeset
- c) select
- d) source

Ans:- (A)

1870. Which command prints the accumulated user and system times for processes run from the shell?

- a) time
- b) times
- c) both time and times
- d) none of the mentioned

Ans:- (B)

1871. Which command executes ""command,"" a built-in shell command, with the specified argument?

- (A)built-in
- (B)caller
- (C) No command is available for this purpose
- (D) None of the aforementioned are true

Ans:- (A)

1872. What does the following command do?

```
$ echo *
```

Answer: This command will print a list of all the files and directories in the current directory, separated by spaces.

1873. What is the difference between ""&&"" and "" ; ""in shell scripts?

The "" && "" operator only executes the second command if the first command returns a success exit status (0). The "" ; ""operator executes the second command regardless of the exit status of the first command.

1874. What is the difference between a shell script and an executable file?

A shell script is a text file that contains shell commands. An executable file is a binary file that has been compiled or linked and can be run directly by the operating system.

1875. How do you redirect the error output of a command to a file?

You can redirect the error output of a command to a file using the following syntax:

```
$ command 2> file.txt
```

1876. What does the following command do?

```
$ command > file.txt 2>&1
```

This command redirects both the standard output and standard error of the 'command' to the file 'file.txt.'

1877. What does the following command do?

```
$ command &> file.txt
```

This command is equivalent to 'command > file.txt 2>&1' and redirects both the standard output and standard error of the 'command' to the file 'file.txt.'

1878. What does the following command do?

```
$ command < file.txt
```

This command redirects the contents of 'file.txt' as the standard input to the 'command'.

1879. What does the following command do?

```
$ command << EOF
```

The '`<<`' operator is used for here-documents, which allows you to pass multiple lines of input to a command as standard input. The input will continue until a line containing only the specified delimiter (EOF in this case) is encountered.

1880. What does the following command do?

```
$ command > /dev/null 2>&1
```

This command redirects both the standard output and standard error of the ' command to `/dev/null`', effectively discarding them.

1881. What is the difference between a single and double square bracket in shell scripts?

The single square bracket (`' ['` and `'] '`) is used for test commands, while the double square bracket (`' [['and ']]`) is used for enhanced test commands. The double square bracket provides additional features and improved performance over the single square bracket.

1882. What is a Shell Scripting?

Shell Scripting is an open-source computer program designed to be run by the Unix/Linux shell. Shell Scripting is a program to write a series of commands for the shell to execute. It can combine lengthy and repetitive sequences of commands into a single and simple script that can be stored and executed anytime which, reduces programming efforts.

1883. What is the difference between `$*` and `$@`?

`$@` treats each quoted arguments as separate arguments but `$*` will consider the entire set of positional parameters as a single string.

Use sed command to replace the content of the file (emulate tac command)

Eg:

```
if cat fille  
ABCD  
EFGH  
Then O/p should be
```

```
EFGH  
ABCD
```

```
sed '1! G; h;$!d' file1
```

Here G command appends to the pattern space, h command copies pattern buffer to hold buffer and d command deletes the current pattern space.

1884. I have a file called 'mypaper.txt' or 'mypaper.tex' or some such somewhere, but I don't remember where I put it. How do I find it?

Solution:

```
$ find ~ -name ""mypaper*""
```

1885. I have a file containing a list of words. I want to sort it in reverse order and print it.

Solution:

```
$ cat myfile.txt | sort -r | lpr
```

1886. In a typical UNIX environment how many kernels and shells are available?

In a typical UNIX environment, only one kernel and many shells are available.

1887. What does the. (dot) indicate at the beginning of a file name and how should it be listed?

A file name that begins with a. (dot) is called as a hidden file. Whenever we try to list the files it will list all the files except hidden files..

1888. Shell programs are stored in which file?

Shell programs are stored in a file called sh.

1889. How to check whether a link is a hard one or a soft link?

We can use -h and -L operators of the test command to check whether a link is hard or soft (symbolic link).

-h file //true if the file is a symbolic link
-L file //true if the file is a symbolic link

1889. Write the difference between \$* and \$@

Unlike \$@, where each quoted argument is treated as a separate argument, \$* treats each positional parameter as a single argument.

1890. What are different types of variables mostly used in shell scripting?

Shell scripts usually have two types of variables:

System-defined variables: Also called environment variables, these are special built-in variables in the Linux kernel for each shell. They are normally defined in capital letters by the OS (Linux) and are standard variables.

Example: `SHELL`

It is a Unix Defined or System Variable, which specifies the default working shell.

User-defined variables: These variables are created and defined by users in order to store, access, read, and manipulate data. In general, they are defined in lowercase letters. The Echo command allows you to view them.

Example: `$ a=10`

In this case, the user has defined the variable `a`, and assigned it the value 10.

1891. Explain the term positional parameters.

In a shell program, positional parameters specify arguments that are used to launch the current process. A special set of variables is usually maintained by the shell for storing positional parameter values. Bash is an example of a shell that uses positional parameters. The bash shell can be used on Linux, BSD, macOS X, and Windows 10.

1892. How can you redirect the standard output and standard error to separate files in shell scripting?

You can redirect standard output and standard error to separate files using the `>` and `2>` operators, respectively. For example, command `> output.txt 2> error.txt` will redirect standard output to `output.txt` and standard error to `error.txt`

1893. What is the difference between `&&` and `||` in shell scripting?

The `&&` operator is used to execute the command following it only if the command preceding it is successful (returns a zero exit status). The `||` operator is used to execute the command following it only if the command preceding it is unsuccessful (returns a non-zero exit status).

1894. Write the command that is used to execute a shell file.

Firstly, use the `chmod` command to set execute permission on your script as shown below:

```
chmod +x script-name-here.sh
```

Secondly, run or execute your script as follows:

```
./script-name-here.sh
```

Alternatively, you can execute shell script by:

```
sh script-name-here.sh
```

1895. Write the difference between \$* and \$@

Unlike \$@, where each quoted argument is treated as a separate argument, \$* treats each positional parameter as a single argument.

1896. What is the best way to debug the shell script/program problems?

Following are some common methods of debugging a script:

The shell script can be modified to include debug statements to display the information that can be useful in identifying the problem.

By setting -x in the script, we can enable debugging.

1897. What is the difference between single quotes (' ') and double quotes ("") in shell scripting?

Ans) Single quotes (' ') preserve the literal value of each character within the quotes. Double quotes ("") allow for the interpretation of variables and command-substitution.

1898. How do you run a shell script in the background?

A shell script can be run in the background by appending an & symbol to the end of the command to execute the script.

1899. How to check whether a link is a hard one or a soft link?

Ans) We can use -h and -L operators of the test command to check whether a link is hard or soft (symbolic link).

-h file //true if the file is a symbolic link

-L file //true if the file is a symbolic link

One can also use:

```
readlink FILE; echo $? // This returns 1 if it's a hard link and 0 if it's a symbolic link.
```

1900. What is the difference between [[\$string == ""efg*]] and [[\$string == efg*]] ?**

Former, checks if string begins with efg. The later, checks if string is efg.

1901. What command needs to be used to take the backup?

The tar command is used to take the backup. It stands for tape archive. The command is mainly used to save and restore files to and from an archive medium like tape.

1902. What can be used to modify File permissions?

using umask.

1903. What is the alternative command available to echo and what does it do?

tput is an alternative command to echo. Using this, we can control the way in which the output is displayed on the screen.

1904. How to find out the number of arguments passed to the script?

```
echo $ #
```

1905. How many fields are present in a crontab file and what does each field specify?

The crontab file has six fields.

The first five fields contain information on when to execute the command and they are as follows:

minute(0-59)

hour(0-23)

day(1-31)

month(1-12)

day of the week(0-6, Sunday = 0).

The sixth field contains the command to be executed.

1906. What is the significance of the Shebang line in Shell Scripting?

The Shebang line is present at the top of the script,e.g. #!/bin/sh. It simply provides information regarding the location where the engine is placed. The engine is the one that executes the script.

1907. What are the two files of crontab command?

The two files of crontab command are:

crontab.allow which decides the users need to be permitted for using the crontab command.
crontab.deny which decides the users need to be prevented from using the crontab command.

How to redirect both standard output and standard error to the same location?

Ans- The two methods are :

```
2>&1(# ls /usr/share/doc > out.txt 2>&1 )
&>(# ls /usr/share/doc &> out.txt )"
```

1908. What is a shell and what is shell scripting?

A shell is a command-line interface that allows users to interact with the operating system by executing commands. Shell scripting refers to writing scripts or programs that automate tasks and perform operations using the shell's commands and utilities.

1909. What is the difference between a shell and a terminal?

A shell is a command-line interface, while a terminal is a window that displays the shell's output and provides an interface for input.

1910. What are some common shell environments, such as bash, zsh, or csh?

Bash (Bourne-Again SHell) is the most common shell environment on Linux and macOS systems. Other popular shell environments include Zsh (Z shell), Csh (C shell), and Fish (Friendly Interactive SHell).

1911. How do you navigate the file system and manage files and directories in shell?

To navigate the file system in shell, you can use commands like cd to change directories, ls to list files and directories, mkdir to create directories, rm to remove files and directories, and mv to move or rename files and directories.

1912. How do you execute and manage processes in shell?

To execute and manage processes in shell, you can use commands like ps to display information about running processes, kill to terminate processes, and jobs to manage background processes.

1913. How do you manipulate and process strings, numbers, and arrays in shell?

To manipulate and process strings, numbers, and arrays in shell, you can use built-in utilities like awk, sed, cut, and grep.

1914. How do you use variables, conditionals, loops, and functions in shell scripts?

Shell scripts can use variables to store values, conditionals (like if statements) to make decisions based on values, loops (like for and while loops) to perform repetitive tasks, and functions to group and reuse code.

1915. How do you perform input/output redirection and pipeline processing in shell?

Input/output redirection allows you to redirect input or output from one command or file to another. Pipeline processing allows you to combine multiple commands by piping the output of one command into the input of another.

1916. How do you use regular expressions in shell scripts?

Regular expressions can be used in shell scripts with utilities like grep and sed to search and manipulate text.

1917. How do you handle errors and exceptions in shell scripts?

You can handle errors and exceptions in shell scripts using commands like set -e to exit the script if a command fails, and using if statements to check for specific error conditions.

1918. How do you debug and troubleshoot shell scripts?

You can debug and troubleshoot shell scripts using commands like echo to display output, set -x to enable debugging mode, and using tools like shellcheck to check for syntax errors.

1919. How do you use shell scripts to automate tasks and run commands on multiple systems?

Shell scripts can be used to automate tasks like backups, system monitoring, and software deployments. They can also be used to run commands on multiple systems using tools like ssh or Ansible.

1920. How do you integrate shell scripts with other tools, such as sed, awk, or grep?

Shell scripts can be integrated with other tools like sed, awk, and grep by piping the output of one command into the input of another, or by using variables to store and manipulate data.

1921. What are the best practices for writing and testing shell scripts?

Use descriptive and meaningful variable and function names to improve readability and maintainability.

Comment your code to explain what it does, how it works, and any assumptions or limitations.

Use consistent indentation and formatting to make the code easier to read.

Use error handling and validation to catch and handle errors and exceptions.

Use shellcheck or other linting tools to check for syntax errors, best practices, and potential bugs.

Write modular and reusable code by creating functions and separating concerns.

Use version control to track changes, collaborate with others, and revert to previous versions if needed.

Write portable code that works across different shell environments and versions.

Test your code thoroughly, including edge cases and boundary conditions.

Optimize the performance of your shell scripts by avoiding unnecessary commands, reducing the number of subprocesses, and using built-in shell functions and utilities.

1922. How do you optimize their performance?

To optimize the performance of your shell scripts, you can follow these tips:

Minimize the use of external commands and subprocesses, which can slow down the script.

Use built-in shell functions and utilities whenever possible, as they are usually faster than external commands.

Use efficient data structures and algorithms, such as arrays and hashes, to manipulate data.

Avoid unnecessary string manipulation and processing, as it can be slow.

Use pipelines and redirection to efficiently process data and minimize subprocesses.

Avoid unnecessary IO operations, such as reading or writing to files, as they can slow down the script.

Use caching and memoization to avoid recomputing results that have already been computed.

Use parallelization and concurrency to speed up computationally intensive tasks.

Chapter - 23 Machine Learning Topics

1923. What is the importance of data cleaning?

Data needs to be transformed in a format which machines can make sense of and we can get the best out of the ML models. For example, when building a Linear model it is necessary to perform one-hot encoding to convert the Categorical variables into features, so that the ML model can make sense out of it.

Also, other techniques like treating missing values, dimensionality reduction etc. help to get our data in a better shape for applying ML algorithms. So, yes, Data Cleaning is important.

1924. What are the basic checks for data cleaning?

- 1) Missing values per column
- 2) Presence of outliers and the reason for their existence.
- 3) Treatment for missing values.
- 4) Converting datatypes of certain features.
- 5) Merging and restructuring datasets.
- 6) Performing one-hot encoding if required.

Code wise

Basic checks:

```
df.head()  
df.shape()  
df.describe()  
df.info()  
df.isnull().sum()
```

1925. What is the difference between forecasting & prediction.

Forecasting and Prediction are two different things.

You always forecast weather but never predict the weather.

Forecasting is nothing but an extrapolation of the past. You have some historic data, and you plotted it on the coordinate and extrapolated it for the future. This is forecasting

Prediction is subjective, it refers to an upcoming event that might or might not happen in the future.

Remember – Anyone can do a prediction.

Ex. Sachin will hit a century tomorrow. You predict which might be based on a deep analysis or could just be a simple shot in the dark.

Therefore all the forecasting is predictions but not all the predictions are forecasting.

1926. Explain Missing Value Treatment by mean, mode, median, and KNN Imputation

Missing Value treatment is no doubt one of the most important parts of the whole process of building a model. Why?

Because we can't afford to eliminate rows wherever there is a missing value in any of the columns. We need to tackle it in the best possible way. There are multiple ways to deal with missing values, and these are my top four methods:-

1. Mean – When do you take an average of a column? There is a saying which goes like this, "When a Billionaire walks in a small bar, everyone becomes a millionaire"

So, avoid using Mean as a missing value treatment technique when the range is too high.

Suppose there are 10,000 employees with a salary of Rs.40,000 each and there are 100 employees with a salary of Rs. 1,00,000 each. In this case you can consider using the mean for missing value treatment.

But, if there are 10 employees with 8 employees earning Rs.40,000 and one of them earning Rs. 10,00,00. Now, here you should avoid using mean for missing value treatment. You can use mode !!

2. Median – Median is the middle term when you write the terms in ascending or descending order. Think of one example where you can use this? The answer is at the bottom of the article

3. Mode – Mode is the maximum occurring number. As we discussed in point one, we can use Mode where there is a high chance of repetition.

4. KNN Imputation – This is the best way to solve a missing value, here n number of similar neighbors are searched. The similarity of two attributes is determined using a distance function.

5. What do you mean when I say “The model has high accuracy in Training dataset but low in testing dataset?”

When the model is performing well on the training set, but poorly on the test set , then this is the classic case of overfitting. In machine learning overfitting means that the data is so well fitted to the training set that it has also learnt noise. Such type of model is not generalized and will give very low accuracy on the test data.

The best way to make sure model is not overfitting is to split the training data into training and validation set and checking the model performance over the validation set.

1927. Define Loss function.

Let's say you are on the top of a hill and need to climb down. How do you decide where to walk towards?

Here's what I would do:

1. Look around to see all the possible paths
2. Reject the ones going up. This is because these paths would actually cost me more energy and make my task even more difficult
3. Finally, take the path that I think has the most slope downhill

This intuition that I just judged my decisions against? This is exactly what a loss function provides.

A loss function maps decisions to their associated costs.

Deciding to go up the slope will cost us energy and time. Deciding to go down will benefit us. Therefore, it has a negative cost.

In supervised machine learning algorithms, we want to minimize the error for each training example during the learning process. This is done using some optimization strategies like gradient descent. And this error comes from the loss function.

1928. What are the metrics to measure the performance of your Linear Regression model?

There are many metrics in the Data science community to measure the performance of a Machine learning Problem:

Confusion Matrix

F1 Score

Gain and Lift Charts

Kolmogorov Smirnov Chart

AUC – ROC

Log Loss

Gini Coefficient

Concordant – Discordant Ratio

Root Mean Squared Error

Cross Validation (Not a metric though!)

R Squared/Adjusted R squared

Mostly we use the R Squared/ Adjusted R squared but that also depends upon the problem statement and requirements

-> Let us first understand what is R-squared:

R-squared or R² explains the degree to which your input variables explain the variation of your output / predicted variable. So, if R-square is 0.8, it means 80% of the variation in the output variable is explained by the input variables. So, in simple terms, higher the R squared, the more variation is explained by your input variables and hence better is your model.

However, the problem with R-squared is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables. This is where “Adjusted R square” comes to help. Adjusted R-square penalizes you for adding variables which do not improve your existing model.

Hence, if you are building Linear regression on multiple variable, it is always suggested that you use Adjusted R-squared to judge goodness of model. In case you only have one input variable, R-square and Adjusted R squared would be exactly same.

Typically, the more non-significant variables you add into the model, the gap in R-squared and Adjusted R-squared increases.

1929. What is the need to remove multicollinearity?

It's very important to reduce the multicollinearity as it can significantly reduce the model performance and we may not know it. It can also reduce features which will result in less complex model and also the overhead to store these features will be less.

Understanding the multicollinearity Conceptually :

Imagine you went to watch a rock band concert . There are two singers , a drummer , a key board player , and two guitarists. You can easily differentiate between the voice of a Singers as one is male and the other female but you can face trouble in determining who is playing the better Guitar.

Both guitarists are playing on the same tone ,same pitch and at the same speed. if you could remove one of them then it wouldn't be a problem since both are almost same.

The benefit of removing one guitarist is cost cutting and fewer members in the team. In machine learning , it is fewer features for training which leads to less complex model.

Here both the guitarists are collinear. If one plays the guitar slowly then another guitarist also plays the guitar slowly. If one plays the guitar faster then the other also plays faster

1930. Define ROC in layman terms.

To Understand ROC first I will try to explain where it is used and why it is used.

As you all know about the classification problems which helps us in identifying the bank fraudulent transactions and helps in diagnosing diseases.

Many classification algorithms like Logistic Regressor uses probability to distribute samples into classes and in most of the cases we take the threshold value by default 0.5 , which means that the algorithm classifies a sample as positive if the probability of that sample being positive is

above 0.5(50%) and classifies a sample as negative if the probability of that sample being positive is less than 0.5(50%)

This threshold that we have taken may not be best case in case of many situations like in case of detecting a disease it may be wise to choose a lower probability threshold to prevent any chance of the disease going misclassified .Thus the classification of critical data demands a more custom threshold which meets certain requirements. This is where the Receiver operating characteristics comes into picture it illustrates the diagnostic ability of a binary classifier.

In layman's terms ,the ROC curve visualises the effect of a chosen probability threshold on the classification efficiency. It helps analyse how the efficiency of Binary classification changes with the values of Probability threshold.

1931. What is the best fit line in Linear Regression?

In linear regression, the best fit line is the line that minimizes the sum of the squared differences between the predicted values (i.e. the values on the line) and the actual values (i.e. the data points). This line is also known as the "line of best fit" or the "regression line." The line is represented by an equation of the form $y = mx + b$, where m is the slope of the line and b is the y-intercept. The slope and y-intercept are chosen such that the line is the best approximation of the relationship between the independent variable (x) and the dependent variable (y) in the data.

1932. When do we use Linear and when do we use Logistic regression?

Linear regression is used to model the relationship between a continuous dependent variable and one or more independent variables. It is used to predict a continuous outcome, such as the price of a house or the temperature of a gas.

Logistic regression, on the other hand, is used when the dependent variable is binary, i.e. it can take only two values, such as "success" or "failure". It is used to predict a probability of an event occurring, such as the probability of a customer buying a product or the probability of a patient having a disease. Logistic regression is a type of generalized linear model and it uses sigmoid function to predict the probability of a binary outcome.

In short, Linear regression is used for predicting continuous variables and Logistic regression is used for predicting binary outcomes.

1933. How is Ridge Regression different from Linear Regression?

Ridge regression is a variation of linear regression, but with a key difference: it adds a "shrinkage" term to the cost function that is being minimized. This shrinkage term, also known

as L2 regularization, is a way of preventing overfitting by penalizing large coefficients in the model.

In Ridge Regression, the cost function is the same as linear regression with an added "L2 regularization" term. It is defined as:

$$J(w) = \text{Sum}(y_i - (w^T x_i))^2 + \lambda * (w^2)$$

Where:

$J(w)$ is the cost function

y_i is the target variable

$w^T x_i$ is the predicted value

λ is the regularization parameter

w^2 is the square of the coefficients

This regularization term causes the optimization algorithm to shrink the coefficients of the model towards zero, but not all the way to zero. It reduces the magnitude of the coefficients, which can help to reduce overfitting.

On the other hand, Linear regression has no such regularization term and it aims to minimize the sum of the squared differences between the predicted values and the actual values, resulting in larger coefficient values.

In summary, Ridge Regression is a variation of linear regression that adds a regularization term to the cost function to prevent overfitting by shrinking the coefficients towards zero, while linear regression aims to minimize the sum of squared differences between the predicted values and actual values without any regularization term.

1934. Explain precision in the simplest terms.

I will always find it confusing to remember these formulations. I am going to tell you one way through which you might not forget in your entire life. I read this method on some quora post.

Imagine that, your girlfriend gave you a birthday surprise every year in the last 10 years. (Sorry, I didn't intend to depress you if you don't have one. Even I don't have one)

However, one day, your girlfriend asks you:

'Sweetie, do you remember all the birthday surprises from me?'

This simple question makes your life in danger.

To extend your life, you need to recall all 10 surprising events from your memory.

So, recall is the ratio of a number of events you can correctly recall to a number of all correct events.

If you can recall all 10 events correctly, then, your recall ratio is 1.0 (100%). If you can recall 7 events correctly, your recall ratio is 0.7 (70%).

Now, it's easier to map the word recall to real-life usage of that word.

However, you might be wrong in some answers.

For example, you answer 15 times, 10 events are correct and 5 events are wrong. This means you can recall all events but it's not so precise.

So, precision is the ratio of a number of events you can correctly recall to a number of all events you recall (mix of correct and wrong recalls). In other words, it is how precise your recall.

From the previous example (10 real events, 15 answers: 10 correct answers, 5 wrong answers), you get 100% recall but your precision is only 66.67% (10 / 15).

Yes, you can guess what I'm going to say next. If a machine-learning algorithm is good at recall, it doesn't mean that algorithm is good at precision. That's why we also need an F1 score which is the (harmonic) mean of recall and precision to evaluate an algorithm.

1935. Explain recall in simple terms.

Consider a example of fire alarm prediction.

A model for fire alarm will generally predict if the alarm will ring or not given a certain set of features.

The model will output a confusion matrix which will consists of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

True Positive – The model predicted the alarm rings and it actually rings

True Negative – The model predicted the alarm did not ring and it actually did not ring.

False Positive – The model predicted the alarm rings but it actually did not ring.

False Negative – The model predicted the alarm does not ring but it actually rings.

Recall = TP/P —————> When it is actually positive, how often does it predict positive.

Recall is also known as sensitivity.

1936. How does the value of R squared and adjusted R Squared error change when you add new variable to your model?

For every predictor you add in the model, the R^2 value goes on increasing but it is not necessary that the predictor you added plays a significant role in increasing the accuracy of the model. The adjusted R^2 adjusts for this artificial increase in the accuracy of the model and its value increases only if the predictor actually contributes to determine the target variable.

R^2 assumes that every single variable explains the variation in the dependent variable. The adjusted R^2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

1937. How to select the number of trees in a random forest?

One of the techniques is to use GridSearchCV() in scikit-Learn where you will have to tune the n_estimators parameter to find the correct no of trees. But, it is also necessary to pass in the adequate no of trees to the list of n_estimators.

Example – n_estimators = [10,30,100]

Typical values are 10, 30 or 100.

Passing lesser no of trees will not actually give you the benefits of the Random forest method as you loose on the benefit of creating large no of trees and averaging their output.

Also, creating large no of trees more than that are required will increase the training time and beyond a certain limit, you will not get substantial benefits in terms of accuracy.

1938. How to choose k in k-means?

I am familiar with two ways to find K values in K-means.

1) Elbow Method – It is based on the concept that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.

We compute the cluster using k-means for different k values by varying k from 1 to say 10 clusters. For each K, we calculate the total WSS (within-cluster sum of square distance). Then we plot the curve of WSS with respect to the K value. The location of the lowest WSS (knee point in a curve) is considered for the optimal value of K (respective K value).

You can see the sample plot of the elbow method in the above image uploaded by swapnil007.

2) The average silhouette method – it measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

We compute cluster using k-means algorithm for different k values vary from 1 to n. For each k value, we calculate the average silhouette of observations. Plot the curve of average silhouette

score with their respective k values. The location where the average silhouette is maximum is considered as optimal value of K.

At the End, varying K value depends on the problem statement, you are solving. Each problem has different requirements based on business requirements. Knowing the domain of the problem, can help to decide the range of K when you are handling large dataset.

1939. What are the different use cases where machine learning algorithms can be used?

Machine learning algorithms can be used in a wide variety of applications and domains. Some common use cases include:

Image and video analysis: Machine learning algorithms can be used to analyze images and videos for object recognition, facial recognition, and video surveillance.

Natural language processing: Machine learning algorithms can be used to process and understand human language, including tasks such as language translation, sentiment analysis, and text summarization.

Predictive modeling: Machine learning algorithms can be used to make predictions about future events or outcomes, such as stock prices, weather patterns, or customer behavior.

Recommender systems: Machine learning algorithms can be used to build personalized recommendations for users, such as product recommendations on e-commerce websites or movie recommendations on streaming platforms.

Robotics: Machine learning algorithms can be used to control robots and other autonomous systems, such as self-driving cars and drones.

Healthcare: Machine learning algorithms can be used to analyze medical data and assist with tasks such as medical image analysis, drug discovery, and patient diagnosis.

Fraud detection: Machine learning algorithms can be used to detect fraudulent behavior, such as credit card fraud or insurance fraud.

Financial services: Machine learning algorithms can be used in financial services such as risk management, portfolio optimization and algorithmic trading.

These are just a few examples of the many different use cases for machine learning algorithms. The potential for machine learning is vast and is being explored in many other fields.

1940. Differentiate between KNN and K-means.

Don't get mislead by 'k' in their names. You should know that the fundamental difference

between both these algorithms is, k-means is unsupervised in nature and kNN is supervised in nature. K-means is a clustering algorithm. kNN is a classification (or regression) algorithm. k-means algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels. kNN algorithm tries to classify an unlabeled observation based on its k (can be any number) surrounding neighbors. It is also known as a lazy learner because it involves minimal training of the model. Hence, it doesn't use training data to make generalization on an unseen data set.

Which approaches are used to evaluate the prediction accuracy of a logistics regression model? There are several approaches that can be used to evaluate the prediction accuracy of a logistic regression model. Some of the most common include:

Confusion Matrix: A confusion matrix is a table that is used to define the performance of a classification algorithm. It shows the number of true positives, true negatives, false positives, and false negatives. From these values, various metrics such as accuracy, precision, recall, and f1-score can be calculated.

Receiver Operating Characteristic (ROC) Curve: An ROC curve is a graphical representation of the performance of a binary classification model. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at different threshold settings. The area under the ROC curve (AUC) can be used to measure the overall performance of the model.

Log-loss: Log-loss, also known as cross-entropy loss, is a measure of the difference between predicted probabilities and actual outcomes. It is commonly used for probabilistic classification problems and it is a popular choice when dealing with imbalanced datasets.

Brier Score: The Brier score is a measure of the mean squared error between predicted probabilities and actual outcomes. It is a popular choice for probabilistic classification problems.

Precision-recall curve: Precision-recall curve is a plot of precision vs recall of a model. It is mostly used when dealing with imbalanced datasets.

Lift Curve: Lift curve is a plot of the ratio of true positive rate to the number of observations over the total population. It is commonly used in marketing, sales, and customer service to measure the effectiveness of a campaign.

1941. What is the meaning of “Overfitting”?

Overfitting occurs when our model becomes complex and it tries to learn the noise and the extreme details hence failing to generalise the new data.

To prevent overfitting we can use L1 or L2 regularization, dropout and early stopping

L1 regularization adds a penalty to minimize the absolute value of weights while L2 regularization adds penalty to minimize the square of weights.

If we are using a neural network to train, we can use dropout where certain number of neurons will be deactivated while training the model.

We can also use early stopping where the training of the model is stopped after a certain point so that our model does not become more complex and it can generalize the test data.

We can also reduce the complexity of the model by using less hidden layers, less number of neurons in the hidden layers, using simple model like linear regression instead of random forest so as to prevent overfitting

1942. How to do dimension reduction?

You are given a train data set having 1000 rows and 1 Million columns. The data set is based on a classification problem. You are asked to reduce the dimension of this data so that model computation time can become manageable. What will be your suggestion?

(You are free to make practical assumptions.)

Ans.

There are various ways to reduce dimensions:

1. Use L1 or lasso regression where the non-important parameters will be eliminated
2. Use Principal Component analysis
3. Use t-SNE(t-distributed stochastic neighbour embedding)

1943. Explain Confusion Matrix Machine Learning

Confusion Matrix is a table which is used for summarizing the performance of a classification algorithm ..

True positive: he is not murdered in reality and left him free

True negative: he is done murdered in reality ,but leave him free

False negative: he is not done any murdered in reality ,you took an action

False Negative: he is done murdered in reality ,but didnt take any action ,But actually action required

1944. What is Auto Regression?

Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step.

Let us understand it simply, it is just like predicting something about a product, based on its earlier progression and sale history. We will tabulate the structure of that product in last week,

last month or the last year. There will be some similarities that would be repeating after some regular intervals. If we could capture that pattern, we could generate a much precise assumption for the same.

1945. How to handle an imbalanced dataset?

There are several techniques to handle imbalanced dataset:

We can use random undersampling where the number of instance of majority class is deleted.

We can also use random oversampling where the number of instance of minority class is duplicated

We can use SMOTE (Synthetic Minority Oversampling Technique) to add instance of minority class. SMOTE uses nearest neighbours of the minority class to create synthetic data.

1946. Given that you have wifi data in your office, how would you determine which rooms and areas are underutilized and overutilized?

There are 3 popular ways of feature selection in ML models:

1. Filter method

a. Pearson's correlation

b. ANOVA

c. LDA

d. Chi-Square

2. Wrapper method

a. Forward selection

b. Backward selection

c. Recursive feature elimination

3. Embedded Methods

a. Lasso regression

b. Ridge regression

1947. What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

If the model is used for classification task, the metrics we should use are

Accuracy

Precision

Recall

F1 score

G1 score

Binary cross entropy loss

If the model is used for regression task, the metrics we should use are

RMSE

MSE

L1 loss

L2 loss

AUC score

1948. How can we implement different word2vec methods?

Word2Vec is a method to construct such an embedding. It can be obtained using two methods (both involving Neural Networks): Skip Gram and Common Bag Of Words (CBOW). Skip Gram works well with a small amount of data and is found to represent rare words well. On the other hand, CBOW is faster and has better representations for more frequent words.

1949. What is VIF?

VIF (Variance Inflation Factor) is a measure that quantifies the degree of multicollinearity in a multiple linear regression model. Multicollinearity refers to the situation where two or more predictor variables are highly correlated with each other. This can cause problems when interpreting the coefficients of the model and can lead to unstable and unreliable estimates.

VIF is calculated for each predictor variable by taking the ratio of the variance of the predictor variable when it is included in the model, to the variance of the predictor variable when it is not included in the model. A VIF of 1 indicates that there is no multicollinearity, while a VIF greater than 1 indicates that there is multicollinearity. The higher the VIF, the more correlated the predictor variable is with the other predictor variables in the model.

The VIF value can be used as a rule of thumb, where values over 5 or 10 are a cause of concern and indicates that the variable may be removed from the model. Additionally, it is important to note that it is not only the VIF value that should be considered, but also the correlation matrix, partial regression plots and the context of the problem when dealing with multicollinearity.

1950. How do linear and logistic regression differ in their error minimization techniques?

Linear regression and logistic regression differ in their error minimization techniques in two main ways:

The type of error function: Linear regression uses the Mean Squared Error (MSE) function as the error function, which measures the average of the squared differences between the predicted values and the actual values. Logistic regression, on the other hand, uses the

log-likelihood function as the error function, which measures the probability of the observed outcomes given the model parameters.

The optimization method: Linear regression uses the method of ordinary least squares (OLS) to minimize the MSE function and find the optimal values of the model parameters. OLS is a computational method that finds the values of the parameters that minimize the sum of the squared differences between the predicted values and the actual values. Logistic regression, on the other hand, uses the maximum likelihood estimation (MLE) method to minimize the log-likelihood function and find the optimal values of the model parameters. MLE finds the values of the parameters that maximize the probability of the observed outcomes given the model parameters.

In summary, Linear regression uses the MSE error function and OLS optimization method to minimize the error and find the optimal values of the model parameters. Logistic regression uses the log-likelihood error function and MLE optimization method to minimize the error and find the optimal values of the model parameters.

1951. What is topic modeling?

Topic modeling deals with mining large amounts of text data to identify topics in them and to identify hidden patterns in the data. It is an unsupervised approach. For example, if there is an article which has recurring occurrences of words like “Farm”, “Produce”, “Crop”, “Soil”, “water” , then we can infer that the article is related to “Farming”, which is our topic.

Topic modeling is a technique used in natural language processing and text mining to automatically identify and extract the latent topics or themes from a large collection of unstructured text data. The goal of topic modeling is to discover the underlying structure and patterns in the text data, and to organize the text into a set of coherent topics that can be used for further analysis and understanding.

Topic modeling algorithms typically use statistical methods such as Latent Dirichlet Allocation (LDA) or Latent Semantic Analysis (LSA) to identify the topics in the text. These algorithms work by identifying the most frequent words and phrases in the text, and then grouping them into clusters or topics based on their similarity. The result is a set of topics, each represented by a set of words or phrases that are highly indicative of that topic.

Topic modeling can be used for a wide range of applications, such as text summarization, document classification, text analytics, and information retrieval. It can also be used in combination with other text mining techniques such as sentiment analysis, named entity recognition, and text summarization to gain more insights from unstructured text data.

1952. What is pruning in case of decision trees?

Pruning in decision trees refers to the process of removing branches or sub-trees from the decision tree in order to simplify it and improve its generalization performance. The idea behind pruning is that a simpler model with fewer branches is less likely to overfit the training data, and therefore will perform better on unseen data.

There are two main types of pruning methods:

Reduced Error Pruning: it starts from the leaves and works its way up the tree. At each node, it checks if the accuracy of the tree is improved by removing the subtree rooted at that node. If so, the subtree is removed.

Cost Complexity Pruning: it starts from the root and works its way down the tree. At each node, it checks if the cost of the tree is improved by removing the subtree rooted at that node. The cost is the sum of the misclassification errors and the complexity of the tree. The complexity of the tree is represented by a parameter called complexity parameter, lambda (λ).

A good pruning method will result in a simpler tree with a good generalization performance. It is important to note that pruning is a trade-off between simplicity and accuracy, as pruning can result in loss of information and lead to underfitting if it is done aggressively.

1953. You are given a data set based on the cancer detection results. You've built a classification model and achieved an accuracy of 95%.

Are you satisfied by the performance? If not, suggest a way to improve it.

We have to check the dataset first. If the dataset is balanced, the model with 95% accuracy is a good model. If it is not balanced, then recall or F1-score has to be chosen as the performance metrics instead of accuracy.

1954. What are the assumptions required for linear regression?

Linear regression is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables. There are several assumptions that must be met in order for the results of a linear regression analysis to be valid:

Linearity: The relationship between the independent and dependent variables is linear.

Independence: The observations are independent of each other.

Homoscedasticity: The variance of the errors is constant across the different levels of the independent variable.

Normality: The errors are normally distributed.

No multicollinearity: The independent variables are not highly correlated with each other.

No Autocorrelation: There is no correlation between the residuals of the model.

No omitted variable bias: all the variables that are relevant to the response variable are included in the model.

It is important to note that if any of these assumptions are violated, the results of the linear regression analysis may be unreliable or invalid. Therefore, it is important to check these assumptions before interpreting the results.

1955. How to improve accuracy of linear regression model? Can you name a possible method of improving the accuracy of a linear regression model?

There are several ways to improve the accuracy of a linear regression model:

Feature selection: Identifying and selecting the most relevant features that have a strong correlation with the target variable can improve the accuracy of the model. Techniques such as correlation analysis, mutual information and feature importance can be used to identify the most informative features.

Feature engineering: Creating new features by combining or transforming existing features can also improve the accuracy of the model.

Model selection: Choosing the appropriate linear regression model based on the assumptions of the data and the problem can improve the accuracy of the model. For example, Ridge and Lasso regression are good options when there is multicollinearity among the independent variables.

Hyperparameter tuning: Hyperparameter tuning is the process of selecting the best set of hyperparameters for a model. Hyperparameter tuning can be done using techniques such as grid search or random search.

Regularization: Regularization is a technique that can be used to avoid overfitting and improve the generalization of the model. Ridge and Lasso regression are examples of regularized linear regression models.

Cross-validation: Cross-validation is a technique used to evaluate a model by dividing the data into training and testing sets. This can be used to identify overfitting and to estimate the generalization performance of a model.

Handling Outliers: Identifying and handling outliers can improve the accuracy of the model. Techniques such as winsorizing, capping, and removing outliers can be used.

Ensemble Methods: Ensemble methods are a way to combine the predictions of multiple models in order to improve the performance of the final model. Techniques such as bagging, boosting, and stacking can be used to improve the accuracy of the model.

1956. Explain about the box cox transformation in regression models.

Dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. In such Scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box-Cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape.

1957. How can we use the Naive Bayes classifier for categorical features?

Naive Bayes is a type of supervised learning algorithm which comes under the Bayesian Classification It uses probability for doing its predictive analysis .

The categorical Naive Bayes classifier is suitable for classification with discrete features that are categorically distributed. The categories of each feature are drawn from a categorical distribution.

For numerical columns, we can use Mean , Standard Deviation and Normal Distribution formula for calculating Likelihood

1958. When can parallelism make your algorithms run faster?

Parallelism can make an algorithm run faster in situations where the algorithm can be divided into smaller, independent tasks that can be executed simultaneously. By dividing the work across multiple processors or cores, the time required to complete the overall task can be reduced.

Here are some examples of when parallelism can be used to improve performance:

Large-scale data processing: When working with large datasets, parallelism can be used to divide the data into smaller chunks and process them simultaneously, reducing the overall processing time.

Monte Carlo simulations: These simulations are based on random sampling, and the samples can be generated independently and in parallel, which can reduce the overall runtime.

Machine Learning and Deep Learning: Training models such as Random Forest, Neural Networks, and Gradient Boosting require a lot of computational power, parallelism can be used to speed up the training process by distributing the computation across multiple cores or even multiple machines.

Image and video processing: Parallelism can be used to speed up image and video processing tasks such as image enhancement, object recognition, and video compression by dividing the work across multiple processors.

Optimization algorithms: Optimization algorithms such as genetic algorithms, particle swarm optimization and simulated annealing can be parallelized to speed up the optimization process.

1959. Suppose you are building a spam classifier. Which among the following two metrics would you consider while implementing the same? Precision or Recall ?

For implementing a spam classifier we need to reduce the false negatives and hence we would consider recall over precision

1960. How to cluster unsupervised data where all the attributes and its values are categorical?

Machine Learning model doesnot understand categories or letters. It only truly understand numbers. To group the categorical data, they are encoded. Mostly label encoding is done and after label encoding one-hot encoding is needed so that the data importance wont dissappear. From one-hot encoding, the dimensions of data gets increases so we need to reduce dimension afterwards.

1961. Is it beneficial to perform dimensionality reduction before fitting an SVM?

Usage of simple normalization techniques such as feature scaling and mean normalization can often result in good accuracy rather than using PCA with SVM.

Eventhough PCA can help improve the discriminative power of classifiers, this doesn't go well with SVMs since their kernel computation is not feature wise.

We could look at the distribution of Eigen values for the covariance matrix of our data, and see if they get very small. In general SVM's are pretty robust in the cases where your data spans of the full feature dimension. The reason for this is that the SVM operates at the sample level (the kernel is computed between samples) and not at the feature level

1962. What is the need of feature selection?

variables from a larger set of features that are most relevant and informative for a specific task or problem. It is an important step in the data pre-processing phase of machine learning and is used to improve the performance and interpretability of a model. There are several reasons why feature selection is important:

Reducing dimensionality: A large number of features can lead to high dimensionality, which can make the model more complex and harder to interpret. Feature selection can be used to reduce the dimensionality of the data and make the model more interpretable.

Improving model performance: Feature selection can improve the performance of a model by removing irrelevant, redundant or noisy features that do not contribute to the model's accuracy or can even negatively impact it. By selecting only the most relevant features, the model can make better predictions and generalize better to new data.

Reducing overfitting: Overfitting occurs when a model is too complex and is able to fit the noise in the data, which can lead to poor generalization performance. Feature selection can be used to reduce overfitting by removing irrelevant features that contribute to the complexity of the model.

Saving computational resources: Feature selection can be used to reduce the number of features and thus the computational cost of training and evaluating a model, which can be especially important for large datasets or complex models.

Improving interpretability: By reducing the number of features, it becomes easier to understand the relationship between the features and the target variable, and the model's predictions can be interpreted more easily.

Overall, feature selection is an important step in the data pre-processing phase of machine learning, it can be used to improve the performance, reduce overfitting, and make the model more interpretable and less computationally expensive.

1963. When are the linear regression lines perpendicular?

Linear regression lines are perpendicular when the slope of one line is the negative reciprocal of the other line's slope.

The slope of a line in a 2-dimensional space is defined as the ratio of the change in the y-coordinate to the change in the x-coordinate, and it is represented by the letter "m" in the equation of a line: $y = mx + b$.

Two lines are perpendicular when their slopes are negative reciprocals of each other, meaning that one line's slope is the opposite of the other's slope, and the product of their slopes is -1.

For example, if the slope of line A is 3, the slope of line B would be -1/3 in order to be perpendicular.

Alternatively, two lines are perpendicular if the angle between them is 90 degrees.

It is important to note that in linear regression, the model tries to fit a line that minimizes the sum of the residuals, and it's not guaranteed that the line will be perpendicular to the x-axis or any other line.

1964. Is rotation necessary in PCA?

It's not necessary, it's a form of optimization or ease of computation.

If you do not rotate the components, the inherent attributes of the metric informational points of the underlying co-ordinate system, will sustain as they are.

The fundamental part of the alignment and the “reconstruction” of the Matrixes - inherently means to “adjust” a piece of the larger set of metric points.

Everything you see on your monitor is 2-dimensional in space, 1-dimensional in time, if there is depth, then there is more information to see. If you don't rotate your POV, to see if there is 3-dimensional space to display, then you won't know if there is hidden information.

1965. Prior to building any kind of model, why do we need to complete the feature selection step?

Feature Selection is the most important part of modelling in Machine Learning, if we skip it then there might be chances of the following:

1. There are some columns in the data that are irrelevant and if we keep them in our dataset they might act as a noise to the data.
2. If we do not do feature selection then the data might have multicollinearity which will effect the modelling score afterwards.
3. If we select the most relevant features then our model will learn more efficiently.

1966. Name and describe three different kernel functions and in what situation you would use each.

Perform dimension reduction.

SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

Polynomial kernel- It is popular in image processing

Gaussian kernel -It is general purpose kernel. used when there is no prior knowledge of data

Sigmoid kernel -we can use it as a proxy for neural network

1967. What to do if one of my columns with integer value is having more than 30% missing values?

30% missing values can be handled by some form of imputation instead of completely deleting the entire column. There are various ways to handle missing values like filling up with mean, median and mode or k-nn imputation. Though, which method to choose is a different discussion as we need to first inspect as to why the missing values are missing in the first case.

Missing values can take many forms like Missing completely at random(MCAR), Missing at random(MAR), Missing Not at Random(MNAR) and each category needs to be dealt in its own way. As far as, it is the about the question of 30% missing values, they should be handled and replaced with one of the above methods.

1968. You are given a train data set having 2000 rows and 1.2 Million columns. The data set is based on a classification problem. You are asked to reduce the dimension of this data so that model computation time can become manageable. What will be your suggestion?

(You are free to make practical assumptions.)

Dimensionality reduction is the way to manage a dataset with a large number of variables.

Two classes in dimensionality reduction:

1. Feature Elimination
2. Feature Extraction

Feature Elimination eliminates the variables which are considered unimportant for the analysis. Though it doesn't give any information for the targeted analysis, we may lose some important data associated with the dropped variables

Feature Extraction creates new independent variables where each new variable is a combination of original variables. Since we have all the original variables combined in each new independent variable, we can drop some of the new variables that are not so important.

PCA (Principal Component Analysis) comes under Feature Extraction.
PCA helps in dimensional reduction.

1969. Which one would likely perform better- Linear Regression or Random Forest Regression? Why?

Linear regression works well when there is linearity in the data. It determines the best fit line using this data. Hence the error is less.

But if the data is non-linear and have multiple features, random forest works well as it combines several decision trees to fit the data.

1970. Do gradient descent methods at all times converge to a similar point?

No gradient descent does not always converge to local minima. It depends where we start (initialize) the weights. If we start near a local minima, there are high chances that it will get trapped in that minima failing to achieve global minima. Hence random initialization of weight is carried out multiple times to achieve global minima.

1971. How to make the model free from underfitting?

To prevent underfitting in a model, you can try these techniques:

Increase model complexity: Use a more complex model with more parameters to capture more information from the data.

Collect more data: Having more data helps the model to better understand the relationship between the features and target variable.

Feature engineering: Create new features or transform existing ones to capture more information.

Regularization: Use regularization techniques like L1, L2, dropout, etc. to prevent overfitting by limiting the model's complexity.

Cross-validation: Use cross-validation to evaluate the model's performance and prevent overfitting.

Early stopping: Stop training the model when the validation error stops decreasing to prevent overfitting on the training data

1972. How would you evaluate a logistic regression model?

Here are some ways to evaluate a logistic regression model:

Confusion matrix: This helps to calculate the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

Accuracy: It measures the overall accuracy of the model. It can be calculated as $(TP + TN) / (TP + TN + FP + FN)$.

Precision: It measures the model's ability to correctly predict positive outcomes. It can be calculated as $TP / (TP + FP)$.

Recall (Sensitivity, Hit Rate): It measures the model's ability to find all positive cases. It can be calculated as $TP / (TP + FN)$.

F1 Score: It is the harmonic mean of precision and recall. It provides a balance between precision and recall.

ROC curve: It plots the relationship between True Positive Rate and False Positive Rate and calculates the AUC (Area Under the Curve) to evaluate the model's performance.

Log loss: It measures the performance of the classifier by penalizing the false classifications.

1973. Do you think 50 small decision trees are better than a large one?

It depends on the context and the type of problem.

Bagging (Bootstrap Aggregation): If the aim is to reduce overfitting, then using multiple small decision trees (e.g. 50) via bagging can improve the model's performance by reducing variance.

Boosting: If the aim is to improve the model's accuracy, then using a large single decision tree may be better than many small ones, but boosting techniques can still be used to improve accuracy further.

Ultimately, the best approach is to experiment with different models and evaluate their performance on the specific dataset and problem using appropriate evaluation metrics.

1974. What metrics we should use to evaluate a binary classification model?

For evaluating a binary classification model, some commonly used metrics are:

Accuracy: The ratio of correctly predicted positive observations to the total observations.

Precision: The ratio of correctly predicted positive observations to the total predicted positive observations.

Recall (Sensitivity or True Positive Rate): The ratio of correctly predicted positive observations to the all positive observations in the dataset.

F1 Score: The weighted average of precision and recall.

Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC): A measure of a classifier's predictive ability.

Confusion Matrix: A table that is used to evaluate the performance of a binary classification model, it gives the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

It is important to choose the right evaluation metrics based on the problem and the data, as different metrics can provide different insights into the performance of a model.

1975. What are your favorite use cases of machine learning models?

Some common use cases for machine learning models are:

Image classification: Classifying images into different categories based on the features and patterns in the images.

Natural Language Processing (NLP): Tasks such as sentiment analysis, text classification, machine translation, etc.

Recommender Systems: Recommending products, items, or services to users based on their preferences, interests, and behavior patterns.

Fraud Detection: Detecting fraudulent activities in financial transactions, insurance claims, etc.

Predictive Maintenance: Predicting equipment failures and maintenance needs in industrial settings, to optimize maintenance schedules and reduce downtime.

Sales forecasting: Predicting future sales based on historical sales data and other relevant factors.

Customer Segmentation: Segmenting customers into different groups based on their behavior, demographics, and other characteristics, to better understand and target each group.

Anomaly Detection: Detecting unusual patterns or deviations in data, to identify outliers and anomalies, such as cyber-attacks, equipment failures, etc.

These are just a few examples, machine learning models have a wide range of applications in various domains and industries.

1976. If through training all the features in the dataset, an accuracy of 100% is obtained but with the validation set, the accuracy score is 75%. What should be looked out for?

If the model has achieved an accuracy of 100% on the training set, but only 75% on the validation set, it may be overfitting. Overfitting occurs when a model learns the training data too well and becomes too complex, resulting in poor generalization performance on unseen data.

To avoid overfitting, you can try the following techniques:

Regularization: This adds a penalty term to the loss function to prevent the model from becoming too complex and overfitting the training data.

Simplifying the model: This involves reducing the number of features, parameters, or hidden layers in the model to make it simpler and reduce its ability to overfit.

Cross-validation: This involves dividing the data into multiple folds, training the model on multiple combinations of training and validation sets, and averaging the results to obtain a more robust estimate of the model's performance.

Early stopping: This involves monitoring the model's performance on a validation set and stopping the training process when the performance stops improving or starts to degrade.

It's important to find the right balance between underfitting and overfitting, as both can impact the performance of the model and its ability to generalize to new data.

1977. What are the three components of time series data?

The three components of time series data are:

Trend: The underlying direction or pattern of the data over time, which can be upward, downward, or horizontal.

Seasonality: Regular patterns in the data that occur at predictable intervals, such as daily, weekly, or yearly patterns.

Cyclical patterns: Irregular patterns in the data that occur over a longer period than seasonality, such as business cycles or economic cycles.

1978. What should be the value of a good variable which we should include in our model?

The value of a good variable to include in a machine learning model depends on several factors, including:

Relevance: The variable should be relevant to the problem being solved and have a meaningful impact on the target variable.

Correlation: The variable should be correlated with the target variable, indicating that it is likely to have a strong influence on the target variable.

Independence: The variable should be independent of other variables in the model, as including highly correlated variables can lead to multicollinearity, making it difficult to interpret the impact of individual variables on the target variable.

Parsimony: The model should include only a limited number of variables, to reduce complexity and prevent overfitting.

Data Quality: The variable should have high quality and complete data, as missing or unreliable data can negatively impact the performance of the model.

It is important to consider these factors when selecting variables for a machine learning model and to perform feature engineering, such as feature selection and dimensionality reduction, to improve the performance and interpretability of the model

1979. What will you do if removing missing values from a dataset causes bias?

If removing missing values from a dataset causes bias, then it is generally not recommended to simply remove them. Bias can significantly impact the results of your analysis and make it difficult to draw valid conclusions. Instead, there are several alternatives that can be used to address missing data in a way that minimizes bias:

Imputation: This involves filling in missing values with estimates based on the available data. There are several imputation techniques, including mean imputation, median imputation, and multiple imputation, among others.

Weighting: This involves adjusting the weight of each observation in the analysis to account for missing data.

Using a different analysis method: Some analysis methods, such as multiple imputation and weighting, are specifically designed to handle missing data and can minimize bias.

Model-based imputation: This involves using statistical models to estimate missing values based on the relationships between variables in the data.

Ultimately, the choice of method will depend on the specifics of your data and the questions you are trying to answer. It is important to carefully consider the potential sources of bias and to choose a method that minimizes this bias while preserving the information in the data to the greatest extent possible.

1980. What is bag-of-words?

Bag-of-Words is a commonly used technique in Natural Language Processing (NLP) for representing text data as numerical vectors. The basic idea behind bag-of-words is to convert a text document into a numerical representation that can be used as input for machine learning algorithms.

In bag-of-words, a text document is represented as a vector of word frequencies, where each dimension in the vector corresponds to a unique word in the vocabulary and the value in each dimension indicates the frequency of that word in the document. The vocabulary is created by collecting all the unique words in the text corpus, and then a document vector is created by counting the number of times each word in the vocabulary appears in the document.

This representation discards information about the grammar and structure of the sentences, but it captures the word frequencies and the context in which the words appear. Bag-of-words is a simple and efficient representation that is widely used in NLP applications, such as text classification, sentiment analysis, and topic modeling.

1981. How to identify the important variable for my Linear regression model ?

There are several methods to identify important variables for a linear regression model, including:

Correlation analysis: Calculate the correlation coefficient between each independent variable and the dependent variable, and identify variables with a high correlation.

Stepwise regression: Use a stepwise regression procedure to select variables that contribute the most to the model's predictive power.

Regularization techniques: Lasso or Ridge regularization can be used to shrink the coefficients of variables with low importance.

Feature importance: Use techniques such as permutation importance or partial dependence plots to identify variables that have the most impact on the model's predictions.

Domain expertise: Consult with subject matter experts who can provide insight into which variables are likely to be the most important in predicting the dependent variable.

1982. What is the different regression algorithm?Explain the trend in forecasting.

Regression algorithms are used to model the relationship between a dependent variable and one or more independent variables. There are several types of regression algorithms, including:

Linear Regression: It models the relationship between the dependent variable and one or more independent variables by fitting a linear equation to the data. It is used for predicting continuous variables.

Logistic Regression: It models the relationship between the dependent variable and one or more independent variables by fitting a logistic function to the data. It is used for predicting binary or categorical variables.

Polynomial Regression: It models the relationship between the dependent variable and one or more independent variables by fitting a polynomial equation to the data. It is used when the relationship between the dependent variable and independent variables is nonlinear.

Ridge Regression: It is a regularization technique that adds a penalty term to the linear regression equation to avoid overfitting. It is used when there are many variables in the model.

Lasso Regression: It is also a regularization technique that adds a penalty term to the linear regression equation, but it shrinks the coefficients of the less important variables to zero. It is used for feature selection.

Trend forecasting is the process of predicting the future behavior of a particular variable or trend. In forecasting, time series models such as ARIMA (Autoregressive Integrated Moving Average) and Exponential Smoothing are commonly used to identify and forecast trends in data. These models use historical data to make predictions and can be used to forecast future values of a particular variable. The accuracy of trend forecasting can be improved by using multiple models and techniques, combining forecasts, and updating the model with new data.

1983. How can you handle an imbalanced dataset?

An imbalanced dataset is a dataset in which the classes are not represented equally. For example, in a binary classification problem, if the positive class represents only 5% of the data and the negative class represents 95% of the data, then the dataset is imbalanced. Handling imbalanced datasets is important because it can lead to biased models that have poor predictive power for the minority class. There are several techniques to handle imbalanced datasets, including:

Resampling: This involves either oversampling the minority class or undersampling the majority class. Oversampling can be done by duplicating samples from the minority class, while undersampling involves randomly removing samples from the majority class. However, oversampling can lead to overfitting and undersampling can result in loss of information.

Synthetic Minority Over-sampling Technique (SMOTE): It generates synthetic samples for the minority class by interpolating between existing samples. This can help to balance the classes without overfitting.

Cost-Sensitive Learning: This involves assigning different costs to misclassifying the minority and majority classes. This can help to improve the predictive power of the model for the minority class.

Ensemble Techniques: Techniques like bagging and boosting can be used to combine multiple models to create a more accurate and balanced model.

Anomaly Detection: Identify and remove the outliers, which are likely to be misclassified by the model.

The choice of method depends on the specific problem, the size of the dataset, and the resources available. It is important to choose the right method for the specific problem to ensure that the model is accurate and balanced.

1984. What are the steps for wrangling and cleaning data before applying machine learning algorithms?

Data wrangling and cleaning are essential steps in preparing data for machine learning. The following are the general steps for wrangling and cleaning data before applying machine learning algorithms:

Data Collection: Collect data from various sources and combine them into a single dataset.

Data Exploration: Explore the dataset to get an understanding of the data, its structure, and its quality. This involves checking for missing values, data types, and outliers.

Data Cleaning: Clean the dataset by handling missing values, removing duplicates, and correcting errors. This step is important because machine learning algorithms cannot handle missing values or incorrect data.

Feature Selection: Select relevant features or variables from the dataset. This is important because having irrelevant features can lead to overfitting, which reduces the model's predictive power.

Feature Engineering: Create new features or transform existing ones to improve the model's performance. This can involve scaling, encoding categorical variables, or creating new variables.

Data Sampling: Balance the dataset by sampling techniques such as over-sampling, under-sampling, or Synthetic Minority Over-sampling Technique (SMOTE).

Data Splitting: Split the dataset into training and test datasets. The training dataset is used to train the model, while the test dataset is used to evaluate its performance.

Data Normalization: Scale the data to the same range to improve the performance of the model.

Dimensionality Reduction: Reduce the dimensionality of the dataset using techniques such as Principal Component Analysis (PCA).

Data Visualization: Visualize the data to identify patterns and trends in the data that can be used to improve the model's performance.

These steps are iterative and require continuous refinement until the data is ready for the machine learning algorithm. The quality of the data is crucial for the accuracy and reliability of the machine learning model.

1985. What happens to our linear regression model if the column z in the data is a sum of columns x and y and some random noise?

If column z in the data is a sum of columns x and y and some random noise, then there will be perfect multicollinearity among the independent variables, x and y. This means that the values of x and y can be used to perfectly predict the value of z, making it impossible for the linear regression model to estimate the coefficients of x and y independently.

In this scenario, the linear regression model will fail to provide reliable estimates of the coefficients of x and y, as their values are not uniquely identifiable. This is because any change in x or y will lead to a corresponding change in z, making it impossible to determine the individual effect of x or y on z.

To handle multicollinearity, one possible solution is to remove one of the variables, x or y, from the model. Alternatively, one could use techniques such as Principal Component Analysis (PCA) or Ridge Regression to handle the multicollinearity and obtain reliable estimates of the coefficients.

Overall, it is important to detect and handle multicollinearity in the data, as it can lead to unreliable and misleading results in the linear regression model.

1986. Time series regression model got higher accuracy than decision tree model. Can this happen? Why?

Yes, it is possible for a time series regression model to have higher accuracy than a decision tree model. The reason is that the accuracy of a model depends on the quality and characteristics of the data and the algorithm used to model it. In some cases, time series data may have a strong trend or seasonal pattern that can be better captured by a time series regression model, whereas decision tree models are better suited for capturing non-linear relationships between features.

Additionally, the performance of a model depends on the specific parameters and hyperparameters used, the size of the dataset, the quality of the features, and the specific problem being solved. It is possible that the time series regression model may have been better suited for the specific problem, had better feature engineering, or was trained on a larger and higher quality dataset than the decision tree model.

Therefore, it is important to evaluate and compare different models using multiple metrics, such as accuracy, precision, recall, F1 score, AUC-ROC, and mean absolute error, among others, to determine which model performs better on a specific problem.

1987. Is it better to spend five days developing a 90-percent accurate solution or 10 days for 100-percent accuracy?

The decision to spend five days developing a 90-percent accurate solution or 10 days for 100-percent accuracy depends on the specific problem being solved and the trade-off between time, cost, and accuracy.

In some cases, a 90-percent accurate solution may be good enough to solve the problem, and the additional time and resources required to achieve 100-percent accuracy may not be worth it. For example, if the problem is to classify emails as spam or non-spam, a 90-percent accurate solution may be good enough to catch most spam emails while allowing legitimate emails to pass through, and spending additional time and resources to achieve 100-percent accuracy may not be worth it.

On the other hand, in some cases, a 100-percent accurate solution may be necessary to solve the problem. For example, in a medical diagnosis system, a false negative could be life-threatening, and a 100-percent accurate solution may be necessary to ensure patient safety.

Therefore, the decision of whether to spend five days developing a 90-percent accurate solution or 10 days for 100-percent accuracy should be made based on the specific problem being solved, the cost and resources required to achieve the desired accuracy, and the trade-off between time, cost, and accuracy.

1988. While working on model, what among them is more important- Model Accuracy or Model Performance?

While developing a model, both model accuracy and model performance are important metrics, but the relative importance of each metric depends on the specific problem being solved.

Model accuracy measures how well the model predicts the outcomes of new data based on the training data. It is an important metric, but it may not always reflect the performance of the model in real-world scenarios. For example, a model with high accuracy may perform poorly if it is slow to make predictions or has high memory requirements.

Model performance, on the other hand, measures how well the model performs in real-world scenarios, taking into account factors such as speed, scalability, interpretability, and reliability. This metric is important because it reflects the practical utility of the model in solving real-world problems.

Therefore, the relative importance of model accuracy and model performance depends on the specific problem being solved and the trade-off between accuracy and performance. In some cases, such as in medical diagnosis or fraud detection, accuracy may be more important than performance. In other cases, such as in real-time prediction systems, performance may be more important than accuracy.

Ultimately, it is important to evaluate the model using multiple metrics and to strike a balance between accuracy and performance that is appropriate for the specific problem being solved.

1989. Is it necessary to perform resampling in your dataset? How would you initiate with this process?

Whether or not to perform resampling in a dataset depends on the specific problem being solved and the characteristics of the data. Resampling can be useful in cases where the dataset is imbalanced, meaning that the number of samples in one class is significantly greater than the other class.

Resampling is a technique used to balance the distribution of classes in the dataset by either oversampling the minority class or undersampling the majority class. Oversampling involves increasing the number of samples in the minority class, while undersampling involves decreasing the number of samples in the majority class.

To initiate the resampling process, you can start by analyzing the class distribution of the dataset to determine if it is imbalanced. If there is a significant class imbalance, you can then decide on the resampling technique to use based on the specific problem being solved and the characteristics of the data.

Some popular resampling techniques include random oversampling, random undersampling, and Synthetic Minority Over-sampling Technique (SMOTE), among others. The choice of resampling technique depends on the specific problem being solved and the characteristics of the data.

It is important to note that resampling can introduce bias into the dataset, and it is important to evaluate the performance of the model on a separate test set to ensure that the model generalizes well to new data. Therefore, it is important to use caution and evaluate the effectiveness of the resampling technique in improving the performance of the model.

1990. Give an example of outlier values and how can they be treated?

Outlier values are extreme values that are significantly different from the other values in a dataset. Outliers can be caused by measurement errors, data entry errors, or rare events that are not representative of the underlying data distribution.

For example, consider a dataset of salaries in a company, and let's say that the majority of employees earn salaries in the range of \$50,000 to \$100,000. However, there is one employee who earns a salary of \$1,000,000. This value is significantly different from the other values in the dataset and can be considered an outlier.

There are several techniques to treat outlier values, including:

Winsorization: Winsorization involves replacing extreme values with less extreme values. For example, the highest and lowest 1% of values in the dataset can be replaced with the 99th and 1st percentiles, respectively.

Trimmed Mean: In this method, a certain percentage of the extreme values are removed from the dataset, and the mean is calculated using the remaining values.

Z-score method: The z-score method involves calculating the z-score of each value in the dataset and removing values that are above a certain threshold.

Interquartile range (IQR) method: This method involves calculating the IQR of the dataset and removing values that are above or below a certain threshold.

Using algorithms that are robust to outliers: There are several algorithms, such as Random Forest, that are less affected by the presence of outliers.

The choice of outlier treatment method depends on the specific problem being solved and the characteristics of the data. It is important to carefully evaluate the impact of outlier treatment on the performance of the model and to choose a method that does not introduce bias into the dataset.

1991. Do having more outlier values a good thing or a bad thing?

Having more outlier values is generally considered a bad thing as it can indicate a data quality issue, skew the distribution of the data, and negatively impact statistical analyses and machine learning models. However, the presence of outlier values can also provide valuable insights into the underlying data and help identify important patterns or anomalies. It is important to carefully evaluate the reasons for outlier values and determine whether they should be excluded or retained in the analysis.

1992. What is tokenization and lemmatization in NLP?

Tokenization and lemmatization are two common text preprocessing techniques used in Natural Language Processing (NLP).

Tokenization refers to the process of breaking up a large piece of text into smaller chunks, called tokens. In NLP, tokens are usually words or phrases. The tokenization process is

important because it enables the computer to analyze and understand the text at a more granular level. For example, a sentence can be tokenized into a list of words or phrases, which can then be used to calculate word frequency or perform other analyses.

Lemmatization, on the other hand, is the process of reducing a word to its base form or lemma. For example, the word "cats" would be reduced to "cat", and the word "running" would be reduced to "run". This is useful because it helps to reduce the number of unique words that a computer needs to process, making it easier to perform analyses and identify patterns in the text. Additionally, lemmatization can also help with tasks like sentiment analysis, where it is important to understand the meaning of individual words in context.

1993. How much data will you allocate for your training, validation and test sets?

The amount of data to allocate for training, validation, and test sets can depend on various factors such as the size of the dataset, the complexity of the problem, and the amount of available computational resources.

A common approach is to use a 60/20/20 split, where 60% of the data is used for training, 20% is used for validation, and 20% is used for testing. However, in some cases, it may be necessary to use a larger amount of data for training, especially if the dataset is large or the problem is complex. In other cases, a smaller amount of data may be sufficient for training, especially if the problem is simpler.

It is also important to note that the data should be randomly sampled and the splitting should be done in a way that ensures the sets are representative of the overall dataset. Additionally, it is important to keep the test set separate from the training and validation sets until the final evaluation to avoid overfitting.

1994. What is p-value?

P-value is a statistical measure that is used to determine whether a result or an observation is statistically significant or due to chance. In simpler terms, the p-value is the probability of obtaining a result as extreme or more extreme than what was observed, assuming that there is no real effect or difference.

For example, if a researcher conducts an experiment and obtains a p-value of 0.05, this means that there is a 5% chance that the result they obtained was due to chance, and a 95% chance that the result is statistically significant and there is a real effect.

Generally, a p-value of less than 0.05 (or 5%) is considered statistically significant, meaning that the result is unlikely to be due to chance, while a p-value greater than 0.05 is not statistically significant and the result may be due to chance. However, it is important to note that p-value is

just one measure of statistical significance and should be interpreted along with other measures and context.

1995. Is the mean imputation of missing data acceptable practice? Why or why not?

Mean imputation is a common technique used to handle missing data, where the missing values in a dataset are replaced by the mean value of the observed data. However, whether mean imputation is an acceptable practice depends on the specific context and the potential impact on the analysis.

One of the main concerns with mean imputation is that it can lead to biased estimates, as the imputed values do not reflect the actual values that are missing. Additionally, mean imputation can reduce the variability in the data and distort the relationships between variables.

There are also some cases where mean imputation may be inappropriate or unsuitable. For example, if there are a large number of missing values, mean imputation may not accurately reflect the true values in the dataset. Additionally, mean imputation may not be appropriate if the missing data are not missing at random, as it can introduce bias into the analysis.

Overall, mean imputation may be an acceptable practice in some contexts, but it is important to carefully consider the potential impact on the analysis and whether there are other techniques that may be more appropriate for handling missing data. Other techniques include using regression models, multiple imputation, or deleting cases with missing data (although this may lead to a loss of power and reduced representativeness of the data).

1996. What is multivariate normality in assumptions of Linear Regression?

Multivariate normality is an assumption in linear regression that refers to the normality of the distribution of the error term (also called the "residuals") across all levels of the predictor variables. In other words, it assumes that the residuals are normally distributed, not just at one level of the predictors but across all levels.

In linear regression, the residuals represent the difference between the actual values of the dependent variable and the predicted values based on the independent variables. The assumption of multivariate normality means that these residuals should be normally distributed with a mean of zero and constant variance, regardless of the values of the predictor variables.

This assumption is important because violations of multivariate normality can lead to biased estimates and inaccurate conclusions about the relationship between the independent and dependent variables. Non-normal residuals can also affect the results of other statistical tests, such as hypothesis tests and confidence intervals.

There are several techniques that can be used to assess the assumption of multivariate normality, including visual inspection of the residuals using normal probability plots and statistical tests such as the Shapiro-Wilk test. If the assumption is violated, there are several options to address it, such as transforming the data, removing outliers, or using non-parametric methods.

1997. If two predictors are highly correlated, what is the effect on the coefficients in the logistic regression?

When two predictors are highly correlated in logistic regression, it can lead to issues with multicollinearity, which occurs when two or more predictors are highly correlated with each other. In such cases, the effect of each predictor on the outcome variable may become difficult to estimate separately, and the coefficients may become unstable or difficult to interpret.

In particular, high correlation between two predictors can lead to a situation where the effect of each predictor on the outcome variable is difficult to distinguish from the other predictor, and this can result in large standard errors and unstable coefficients. As a result, it may become difficult to interpret the contribution of each predictor to the model, and the coefficients may be inconsistent or even have the opposite sign than expected.

To address the issue of multicollinearity, it is often recommended to remove one of the correlated predictors from the model, or to combine them into a single variable. Alternatively, regularization techniques such as ridge regression or lasso regression can be used to reduce the impact of multicollinearity on the coefficients. These techniques can help to stabilize the coefficients and reduce the impact of multicollinearity on the model.

1998. While working on a data set, how do you select important feature?

When working with a dataset, selecting important features is an important step in building a predictive model that is accurate and interpretable. Here are some common techniques for feature selection:

Correlation analysis: Correlation analysis is a quick and easy way to identify which features are strongly correlated with the target variable. Features with high correlation can be good candidates for inclusion in the model.

Univariate feature selection: Univariate feature selection involves selecting features based on their individual performance in a statistical test, such as ANOVA or t-tests. Features that have high statistical significance are selected for the model.

Recursive feature elimination: Recursive feature elimination involves training a model on all features, and then removing the least important feature and training the model again. This process is repeated until a desired number of features is reached.

Regularization: Regularization methods, such as ridge regression and LASSO, can be used to penalize the coefficients of features that are not important, effectively shrinking their impact on the model.

Feature importance from tree-based models: Tree-based models, such as decision trees and random forests, provide a measure of feature importance that can be used to select important features for the model.

It is important to note that selecting features is not a one-size-fits-all approach, and the selection method may vary depending on the specific data and problem. It is often helpful to try multiple methods and compare the results to find the best set of features for the model.

1999. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

If the distribution of the test data is significantly different from the distribution of the training data, there can be several issues with the predictive model. Some of the main issues are:

Overfitting or underfitting: If the distribution of the test data is significantly different from the training data, the model may overfit or underfit the training data. Overfitting occurs when the model learns the noise in the training data and fails to generalize to new data, while underfitting occurs when the model is too simple and cannot capture the complex patterns in the data.

Poor performance: If the test data has a different distribution from the training data, the model may perform poorly on the test data. This can lead to low accuracy, precision, recall, or other performance metrics, which can affect the utility of the model in real-world applications.

Biased estimates: If the test data has a different distribution from the training data, the estimates of the model parameters may be biased. This can lead to inaccurate predictions and affect the interpretation of the results.

Unforeseen problems: If the test data has a different distribution from the training data, the model may not capture all the important features of the data, and may miss important patterns or relationships that are present in the test data. This can lead to unforeseen problems in real-world applications, such as incorrect predictions or decisions.

To address these issues, it is important to ensure that the training and test data come from the same distribution, or to use techniques such as domain adaptation or transfer learning to adapt the model to the new data distribution. It is also important to evaluate the performance of the model on the test data and compare it to the performance on the training data, to ensure that the model is generalizing well to new data.

2000. What is the use of stringAsFactors set as False?

In R, when reading data from a file or creating a data frame, the `stringAsFactors` argument can be set to either `TRUE` or `FALSE` to control how character vectors are converted to factors. By default, `stringAsFactors` is set to `TRUE`, which means that R will automatically convert character vectors to factors, with each unique value in the vector becoming a level in the factor.

Setting `stringAsFactors` to `FALSE` can be useful in several situations, including:

Memory optimization: Factors can take up a significant amount of memory, especially for large datasets with many levels. By setting `stringAsFactors` to `FALSE`, you can save memory by keeping character vectors as they are, rather than converting them to factors.

Control over factor creation: When `stringAsFactors` is set to `TRUE`, R automatically creates factors based on the unique values in the character vectors. This can sometimes lead to unexpected results, such as factors with too many or too few levels. By setting `stringAsFactors` to `FALSE`, you can have more control over how factors are created, and can use additional functions such as `factor()` to create factors with specific levels.

Speed optimization: Converting character vectors to factors can be a time-consuming process, especially for large datasets. By setting `stringAsFactors` to `FALSE`, you can save time during data preparation and analysis.

It is important to note that setting `stringAsFactors` to `FALSE` can also have some drawbacks, such as making it more difficult to use certain functions that require factors, such as `lm()` for linear regression. Therefore, it is important to consider the specific needs of your analysis when deciding whether to set `stringAsFactors` to `FALSE` or `TRUE`.

2001. State one situation where the set-based solution is advantageous over the cursor-based solution.

In general, a set-based solution is advantageous over a cursor-based solution when working with large datasets, particularly in situations where performance is a critical factor. One specific situation where a set-based solution may be advantageous over a cursor-based solution is when performing aggregations on large datasets.

Aggregations involve performing some kind of computation on a set of values, such as calculating the sum, average, or count of a group of values. When using a cursor-based solution to perform an aggregation, the cursor needs to loop through each row of the dataset and perform the computation on each row individually. This can be slow and resource-intensive, particularly for large datasets.

In contrast, a set-based solution involves performing the computation on the entire dataset as a set, without the need to loop through each row individually. This can be much faster and more efficient, particularly for large datasets, and can lead to significant performance improvements.

For example, in SQL, using the GROUP BY clause to perform aggregations is a set-based solution, while using a cursor to loop through the dataset and perform the aggregation row by row is a cursor-based solution. In general, the GROUP BY clause is much faster and more efficient for performing aggregations on large datasets.

2002. What is Multicollinearity ?

Multicollinearity is a phenomenon that occurs when two or more predictor variables in a multiple regression model are highly correlated with each other. This can be problematic because it can make it difficult to determine the individual effect of each predictor variable on the response variable.

Multicollinearity can lead to unreliable and unstable estimates of the regression coefficients, making it difficult to interpret the results of the regression analysis. In some cases, multicollinearity can even lead to counterintuitive results, such as coefficients with the wrong sign or with very large standard errors.

There are several ways to detect multicollinearity in a regression analysis, including examining correlation matrices and variance inflation factors (VIFs) for the predictor variables. If multicollinearity is detected, there are several possible solutions, such as removing one of the correlated predictor variables, combining the correlated variables into a single variable, or using regularization techniques such as ridge regression or lasso regression.

Overall, multicollinearity is an important issue to consider in multiple regression analysis, and it is important to address it appropriately in order to ensure accurate and reliable results.

2003. We know that one hot encoding increases the dimensionality of a dataset, but label encoding doesn't. How?

One-hot encoding and label encoding are two different ways of converting categorical data into numerical data so that it can be used in machine learning algorithms. One-hot encoding increases the dimensionality of the dataset, while label encoding does not.

In one-hot encoding, each category in a categorical variable is converted into a new binary variable, where a value of 1 indicates that the observation belongs to that category, and 0 otherwise. This results in a new binary variable for each category, which increases the dimensionality of the dataset. For example, if we have a categorical variable with three

categories, one-hot encoding will result in three new binary variables, each representing one of the three categories.

In label encoding, each category is assigned a unique numerical value. This does not increase the dimensionality of the dataset because each categorical variable is replaced by a single numerical variable. For example, if we have a categorical variable with three categories, label encoding will assign each category a unique integer value, such as 1, 2, and 3.

However, it's important to note that label encoding has some limitations. Specifically, assigning numerical values to categories implies an order or rank to the categories, which may not be appropriate in all cases. For example, if we use label encoding to convert the categorical variable "red," "green," and "blue" into numerical values of 1, 2, and 3, this implies an order to the categories (red < green < blue), which may not be appropriate in all contexts.

Overall, the choice of encoding method depends on the specific requirements of the machine learning problem and the nature of the categorical variables being used.

2004. What are the various aspects of a Machine Learning process?

A typical machine learning process consists of several key aspects:

Data collection: This involves collecting and gathering relevant data that will be used to train and test the machine learning model.

Data preparation: Once the data is collected, it needs to be cleaned and preprocessed to ensure that it is accurate and suitable for training the machine learning model. This may involve tasks such as data cleaning, feature selection, and data transformation.

Data exploration and analysis: This involves exploring and analyzing the data to identify patterns and trends, and to gain a better understanding of the relationships between the variables.

Model selection and training: This involves selecting an appropriate machine learning algorithm and training it on the prepared data. The selection of the algorithm may depend on the specific problem being addressed, the nature of the data, and other factors such as the complexity of the model and the required accuracy.

Model evaluation: Once the model is trained, it needs to be evaluated to determine its performance and to identify any potential issues or areas for improvement. This may involve techniques such as cross-validation, which involves testing the model on data that was not used in training.

Model tuning: Based on the evaluation results, the model may need to be tuned or modified to improve its performance. This may involve adjusting the model parameters or changing the model architecture.

Deployment: Finally, once the model is trained and evaluated, it needs to be deployed in a real-world setting, where it can be used to make predictions or decisions based on new data.

Overall, the machine learning process is iterative, and each step is closely interconnected with the others. A successful machine learning project requires careful attention to each of these aspects to ensure that the final model is accurate, reliable, and useful.

2005. How would you perform feature selection on the dataset?

Feature selection is an important step in machine learning to identify the most relevant features in a dataset that will be used to train a model. Here are some common approaches to feature selection:

Correlation analysis: Correlation analysis involves calculating the correlation coefficient between each feature and the target variable. Features with low correlation coefficients can be removed from the dataset.

Feature importance ranking: This involves using a machine learning algorithm to rank the importance of each feature in the dataset. Popular algorithms for this purpose include decision trees and random forests.

Recursive feature elimination: This approach involves recursively removing features from the dataset and training a model with the remaining features. The process is repeated until a specified number of features is left or until a performance threshold is reached.

L1 regularization: L1 regularization is a type of regularization that can be applied to some machine learning algorithms to penalize the use of unnecessary features. This results in sparse models with fewer features.

Univariate feature selection: Univariate feature selection involves testing each feature individually and selecting the ones that have the strongest relationship with the target variable.

The choice of feature selection method depends on the specific problem being addressed and the nature of the data. It's important to balance the complexity of the model with its performance, and to avoid overfitting by selecting only the most relevant features. It's also important to evaluate the model's performance after feature selection to ensure that it has not been significantly impacted by the removal of features.

2006. What is the difference between machine learning and deep learning?

Machine learning and deep learning are both subfields of artificial intelligence that involve training models to make predictions or decisions based on data. However, there are some important differences between the two:

Model complexity: Machine learning models are typically less complex than deep learning models. Machine learning models often rely on hand-engineered features, while deep learning models use hierarchical representations learned from the data.

Data requirements: Deep learning models require large amounts of data to be trained effectively, while machine learning models may be trained on smaller datasets.

Computation requirements: Deep learning models require more computational power than machine learning models, which can make them more difficult and expensive to train.

Model interpretability: Machine learning models are often more interpretable than deep learning models, as they rely on a more transparent decision-making process. Deep learning models, on the other hand, can be more opaque and difficult to understand.

Applications: Machine learning is often used for more traditional, structured data, such as tabular data or time series data. Deep learning, on the other hand, is often used for more complex and unstructured data, such as images, audio, or natural language.

Both machine learning and deep learning have their strengths and weaknesses, and the choice of which approach to use depends on the specific problem being addressed, the nature of the data, and the available computational resources.

2007. What is CNN?

CNN stands for Convolutional Neural Network, which is a type of deep neural network commonly used in image and video recognition tasks.

The key characteristic of a CNN is the use of convolutional layers, which apply a set of learnable filters to the input image to create a set of feature maps. These feature maps capture the presence of specific visual features, such as edges, corners, and blobs, at different locations in the image. The convolutional layers are typically followed by pooling layers, which reduce the spatial dimension of the feature maps while preserving their key features.

CNNs also typically include fully connected layers, which take the flattened output of the convolutional and pooling layers and use it to classify the input image into one or more categories.

One of the main advantages of CNNs is their ability to automatically learn hierarchical representations of the input image, with each layer of the network capturing increasingly

abstract features. This allows CNNs to achieve high accuracy in tasks such as image recognition, object detection, and image segmentation.

CNNs have been used in a wide range of applications, from self-driving cars to medical image analysis, and continue to be an active area of research in the field of deep learning.

2008. Is it better to have too many false negatives or too many false positives?

The answer to this question depends on the specific problem being addressed and the costs associated with false positives and false negatives.

In some situations, such as in medical diagnosis or disease detection, false negatives may be more harmful than false positives. A false negative result means that a person who actually has the disease is incorrectly classified as not having the disease, which could lead to delayed or incorrect treatment and potentially negative health outcomes. In this case, minimizing false negatives may be more important than minimizing false positives.

On the other hand, in some situations such as spam email filtering, false positives may be more harmful than false negatives. A false positive result means that a legitimate email is incorrectly classified as spam, which could cause the recipient to miss important messages. In this case, minimizing false positives may be more important than minimizing false negatives.

Ultimately, the optimal balance between false positives and false negatives depends on the specific application and the costs associated with each type of error. In some cases, it may be necessary to weigh the costs of different types of errors and choose a threshold that balances the trade-offs between false positives and false negatives.

2009. Executing a binary classification tree algorithm is a simple task. But, how does a tree splitting take place?

A binary classification tree algorithm starts with a single node, which represents the entire dataset. The tree is then recursively split into smaller subsets, with each split dividing the data into two branches based on a specific feature or combination of features.

The process of splitting the tree involves selecting the best feature to split on at each node. The goal is to find a feature that results in the most effective separation of the data into the two branches, so that each branch contains as many samples as possible of a single class. There are various methods for selecting the best feature, including information gain, Gini impurity, and chi-squared.

Once the best feature is selected, the node is split into two child nodes based on a threshold value. Samples with feature values below the threshold are assigned to the left child node, and samples with feature values above the threshold are assigned to the right child node. This

process is repeated recursively for each child node until a stopping criterion is met, such as a maximum depth or a minimum number of samples per node.

The result of this process is a binary tree structure, with each leaf node representing a final decision or prediction. When a new sample is presented to the tree, it follows the branches of the tree based on its feature values until it reaches a leaf node, which provides the final classification or prediction for that sample.

2010. How to read PACF graph?

The PACF (partial autocorrelation function) graph is a plot that shows the correlation between a time series and its lagged values, while controlling for the influence of other lags.

To read a PACF graph, you can follow these steps:

Look for the point on the graph where the PACF value first crosses the significance level (usually shown as a dotted line at 0.05 or 0.01).

Identify the lag value corresponding to that point, which represents the number of lagged values that are significantly correlated with the time series, after controlling for the influence of other lags.

Look for any significant spikes or patterns in the PACF values at other lag values, which can indicate additional significant correlations.

Compare the PACF graph to the ACF (autocorrelation function) graph, which shows the correlation between the time series and its lagged values without controlling for other lags, to gain a fuller understanding of the autocorrelation structure of the time series.

In summary, the PACF graph helps identify the number of significant lags that should be included in a time series model, and provides insight into the autocorrelation structure of the time series.

2011. What cross-validation technique would you use on a time series dataset?

When working with a time series dataset, a common cross-validation technique is called "rolling window" or "walk-forward" cross-validation.

In this technique, the time series data is split into a series of consecutive and non-overlapping windows, with each window containing a fixed number of time steps. The model is then trained on the data from the first window, and its performance is evaluated on the data from the next window. The process is repeated for each subsequent window until the end of the time series.

This approach ensures that the model is trained on data that precedes the test data, mimicking the actual use case scenario of predicting future values based on past data.

Another variant of this technique is "expanding window" cross-validation, where the training window expands with each iteration and includes all past data up to the current test data point.

Overall, the choice of cross-validation technique depends on the nature of the time series data and the specific problem at hand, but rolling window cross-validation is a good starting point for most time series applications.

2012. When modifying an algorithm, how do you know that your changes are an improvement over not doing anything?

When modifying an algorithm, it's important to evaluate the impact of the changes to determine whether they are an improvement over not doing anything. There are several ways to do this:

Simulation and experimentation: One way to evaluate the impact of changes to an algorithm is to simulate the algorithm with the original version and the modified version using a range of input data, and compare the results. By running experiments on the same set of data using the original and modified versions, you can compare the performance of the two algorithms and determine whether the changes improve the algorithm's performance.

Cross-validation: Another way to evaluate the impact of changes to an algorithm is to use cross-validation, a technique that involves splitting the data into training and test sets, and comparing the performance of the original and modified algorithms on the test set. By comparing the performance of the two algorithms on a set of data that was not used in training, you can get a better sense of how the changes impact the algorithm's performance in the real world.

Performance metrics: It's also important to define performance metrics that can be used to quantify the impact of the changes. For example, if you're modifying a classification algorithm, you might use metrics such as accuracy, precision, and recall to evaluate the impact of the changes on the algorithm's performance. By defining specific metrics and comparing them between the original and modified versions of the algorithm, you can determine whether the changes are an improvement or not.

Overall, evaluating the impact of changes to an algorithm requires careful experimentation and analysis, and it's important to use a range of techniques and metrics to get a clear sense of the algorithm's performance.

2013. How do you select the appropriate algorithm for a given problem?

Selecting the appropriate algorithm for a given problem depends on the type of problem and the nature of the data. The problem may fall into either supervised or unsupervised learning, and

the nature of the data may dictate which algorithm is best suited for it. For example, linear regression is a good fit for continuous numerical data, while decision trees are better suited for categorical data. It's essential to evaluate the accuracy, efficiency, and scalability of different algorithms before selecting one.

2014. What is overfitting in machine learning and how can you avoid it?

Overfitting occurs when a machine learning model becomes too complex and starts to fit the noise in the data instead of the underlying pattern. This results in a model that performs well on the training data but poorly on the testing data. To avoid overfitting, we can use techniques such as regularization, early stopping, and cross-validation.

2015. What is cross-validation, and why is it important?

Cross-validation is a technique used to evaluate the performance of a machine learning model by splitting the data into multiple subsets. It involves training the model on one subset and testing it on another. This process is repeated multiple times, with different subsets being used for training and testing. The results are then averaged to give an overall estimate of the model's performance. Cross-validation is important because it helps to prevent overfitting and provides a more accurate estimate of a model's performance.

2016. How do you handle missing data in a dataset?

There are several ways to handle missing data in a dataset, such as imputation, deletion, and prediction. Imputation involves filling in missing values with a specific number or estimate. Deletion involves removing the rows or columns containing missing values. Prediction involves using machine learning algorithms to predict the missing values based on the available data.

2017. What is the difference between bias and variance in machine learning models?

Bias refers to the error that arises from incorrect assumptions in the learning algorithm, resulting in the model being unable to capture the underlying pattern in the data. Variance, on the other hand, refers to the error that arises from the model being too complex, resulting in it fitting the noise in the data instead of the underlying pattern.

2018. Can you explain the bias-variance tradeoff in machine learning?

The bias-variance tradeoff is a concept in machine learning that refers to the tradeoff between the model's ability to fit the training data accurately (low bias) and its ability to generalize to new data (low variance). A model with high bias will underfit the data, while a model with high variance will overfit the data. Therefore, it's important to strike a balance between bias and variance when training a machine learning model.

2019. What is regularization, and why is it important?

Regularization is a technique used to prevent overfitting in machine learning models by adding a penalty term to the loss function. The penalty term discourages the model from becoming too complex and fitting the noise in the data. Regularization is important because it helps to improve the model's ability to generalize to new data.

2020. How do you evaluate the performance of a machine learning model?

There are several ways to evaluate the performance of a machine learning model, such as accuracy, precision, recall, F1 score, ROC curve, and AUC. The choice of evaluation metric depends on the nature of the problem and the type of data.

2021.What are some common techniques used for feature selection in machine learning?

Some common techniques used for feature selection in machine learning are filter methods, wrapper methods, and embedded methods. Filter methods involve selecting features based on statistical measures such as correlation or mutual

2022. Can you cite some examples where both false positive and false negatives are equally important?

Examples where both false positives and false negatives are equally important include medical diagnosis tests for serious diseases, security screening at airports, and credit card fraud detection. In all these scenarios, a false positive can cause inconvenience, while a false negative can have serious consequences, so both types of errors need to be minimized

2023. How do you come up with an algorithm that will predict what the user needs after they type only a few letters?

An algorithm that predicts what the user needs after they type only a few letters can be developed using a combination of natural language processing (NLP) and machine learning techniques. The algorithm can be trained on a large corpus of text data to recognize patterns in the input and make predictions based on those patterns. It can also incorporate user feedback to improve its accuracy over time.

2024. If you're attempting to predict a customer's gender, and you only have 100 data points, what problems could arise?

When attempting to predict a customer's gender with only 100 data points, the main problem that could arise is overfitting. With a small dataset, the model may not be able to capture the true underlying patterns and instead memorize the noise in the data, leading to poor

performance on new data. Another problem is that the sample may not be representative of the entire population, leading to biases in the model's predictions.

2025. Give examples where a false negative is more important than a false positive, and vice versa.

A false negative is more important than a false positive in medical diagnosis tests for serious diseases, where a missed diagnosis can have life-threatening consequences. A false positive is more important than a false negative in spam email classification, where incorrectly labeling a legitimate email as spam can cause inconvenience but not as much harm as missing an important email.

2026. If the model isn't perfect, how would you like to select the threshold so that the model outputs 1 or 0 for label?

The threshold for the model's output can be selected based on the cost of each type of error. If false negatives are more costly, then the threshold can be set lower to minimize the number of false negatives, even if it increases the number of false positives. Conversely, if false positives are more costly, then the threshold can be set higher to minimize the number of false positives, even if it increases the number of false negatives.

2027. List out the difference between linear and logistic regression

The main differences between linear and logistic regression are:

Linear regression is used for predicting continuous numerical values, while logistic regression is used for predicting binary categorical values.

Linear regression uses a linear equation to model the relationship between the input variables and the output variable, while logistic regression uses a logistic function to model the probability of the output variable given the input variables.

Linear regression assumes that the errors are normally distributed and have constant variance, while logistic regression assumes that the errors are binomially distributed and have constant variance.

Linear regression can have negative predicted values, while logistic regression always produces values between 0 and 1.

2028. You have built a multiple regression model.

Your model R^2 isn't as good as you wanted. For improvement, you remove the intercept term, and your model R^2 becomes 0.8 from 0.3. Is it possible? How?

It is not possible for removing the intercept term in a multiple regression model to increase the R^2 value from 0.3 to 0.8. Removing the intercept term can reduce the bias of the model, but it

cannot increase the goodness of fit of the model. This may be due to overfitting or other issues with the original model.

2029. Why is mean square error a bad measure of model performance? What would you suggest instead?

Mean square error (MSE) is a bad measure of model performance because it is sensitive to outliers and can lead to biased results. Instead, one can use other measures such as mean absolute error (MAE), root mean square error (RMSE), or R^2 .

2030. Can we use the logistic regression algorithm for a regression problem?

No, logistic regression is not suitable for a regression problem as it is a classification algorithm. Logistic regression predicts the probability of an event occurring (binary classification) and outputs a value between 0 and 1, whereas regression predicts a continuous output value.

2031. Why is “Naive Bayes” naive?

Naive Bayes is considered naive because it makes the strong assumption of independence between the input features, which is often not true in real-world problems. Despite this simplification, Naive Bayes can still perform well in practice and is widely used in text classification and spam filtering.

2032. Give some problems or scenarios where map-reduce concept works well and where it doesn't work.

Map-reduce works well in problems where the data can be divided into smaller chunks and processed independently, such as counting word frequencies in a large text corpus or analyzing log files. It may not work well in problems where the data dependencies are complex and require a global view of the data, such as in some types of graph algorithms or numerical simulations.

2033. When it comes to Evaluation of Linear Regression which Evaluation Metrics would be the best when we have redundant Variables in our dataset?

When dealing with redundant variables in a linear regression model, it is important to use evaluation metrics that penalize model complexity, such as adjusted R^2 or Akaike Information Criterion (AIC).

2034. What is maximum likelihood estimation? Could there be any case where it doesn't exist?

Maximum likelihood estimation is a method for estimating the parameters of a statistical model by maximizing the likelihood function of the observed data. It is widely used in various statistical

models such as linear regression and logistic regression. However, maximum likelihood estimation may not exist if the likelihood function is not well-behaved, such as in some non-convex optimization problems.

2035. How do predict "y" at time t+1?

To predict "y" at time t+1, one can use a time-series forecasting model such as ARIMA, exponential smoothing, or recurrent neural networks (RNNs). These models use past observations to make predictions about future values.

2036. What is the role of trial and error in data analysis?

Trial and error can play a role in data analysis by allowing the data scientist to test various hypotheses and models until a satisfactory result is obtained. However, it is important to balance this approach with sound statistical and scientific principles to avoid overfitting and false conclusions.

2037. You notice a system uses a lot of triggers to enforce foreign key constraints, and the triggers are error-prone and difficult to debug. What changes can you recommend to reduce the use of triggers?

To reduce the use of triggers in a system that enforces foreign key constraints, one can consider using other methods such as check constraints or referential integrity constraints. Additionally, one can improve the data quality by enforcing data integrity rules at the application layer and using data validation techniques.

2038. What is the difference between Boosting and Bagging?

Boosting and bagging are both ensemble methods used in machine learning. The main difference is that boosting trains weak learners sequentially, with each model attempting to correct the errors of the previous one, while bagging trains weak learners independently in parallel. Boosting can lead to lower bias and higher variance, while bagging can reduce the variance but not necessarily the bias.

2039. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

If the distribution of the test data is significantly different from the distribution of the training data, the model may perform poorly on the test data and generalization may be limited. This can occur when the training data is not representative of the population or when the distribution of the data changes.

2040. Treating a categorical variable as a continuous variable would result in a better predictive model? How?

Treating a categorical variable as a continuous variable would not necessarily result in a better predictive model. This is because categorical variables are inherently different from continuous variables and do not follow the same mathematical properties. For example, a variable with categories "red," "green," and "blue" does not have a natural ordering or numerical representation, and treating it as a continuous variable would not make sense.

2041. Explain Long Short Term Memory Algorithm in brief Question With code snippet in R or Python

Long Short Term Memory (LSTM) is a type of recurrent neural network that is commonly used in natural language processing and time series prediction. Here's an example code snippet for LSTM in Python using the Keras library:

```
from keras.models import Sequential
from keras.layers import LSTM, Dense

model = Sequential()
model.add(LSTM(units=64, input_shape=(timesteps, features)))
model.add(Dense(units=1))
model.compile(loss='mean_squared_error', optimizer='adam')
model.fit(X_train, y_train, epochs=10, batch_size=32)
```

In this example, timesteps and features represent the number of time steps and features in the input data, respectively. The LSTM layer is used to learn patterns in the input sequence, and the Dense layer is used to make a prediction. The model is trained using mean squared error loss and the Adam optimizer.

2042. What is the main difference between K-Means and K-Means++ algorithm?

K-Means is a clustering algorithm that partitions data into K clusters based on their distance to K centroids. K-Means++ is an improvement over K-Means that selects initial centroids in a way that increases the chances of finding a better clustering.

In K-Means++, the first centroid is chosen at random from the data points, and each subsequent centroid is chosen with a probability proportional to its squared distance to the nearest centroid that has already been chosen. This ensures that the initial centroids are well spread out and reduces the chances of getting stuck in a local minimum.

2043. What is the difference between cyclicity and seasonality?

Cyclicity refers to a pattern that repeats at regular intervals, such as a weekly or yearly pattern. Seasonality refers to a pattern that repeats at fixed intervals but not necessarily at regular

intervals, such as a holiday season or a sales season.

2044. What is the acceptable value range for p,d and q in ARIMA?

The values of p, d, and q in ARIMA models depend on the properties of the time series being modeled. In general, p represents the order of the autoregressive component, d represents the order of differencing, and q represents the order of the moving average component.

There is no fixed acceptable value range for p, d, and q, and they need to be chosen based on the characteristics of the time series being modeled. One common approach is to use methods such as autocorrelation plots and partial autocorrelation plots to determine the values of p, d, and q that provide the best fit to the data.

2045. What is Homoscedasticity in assumptions of Linear Regression?

Homoscedasticity is an assumption of linear regression that states that the variance of the error terms is constant across all levels of the predictor variables. In other words, the spread of the residuals should be similar for all values of the independent variables.

If the assumption of homoscedasticity is violated, it can lead to biased and inefficient estimates of the regression coefficients, and the standard errors and confidence intervals may be incorrect.

2046. If you're attempting to predict a customer's gender, and you only have 100 data points, what problems could arise?

With only 100 data points, the sample size may not be large enough to capture the variability of the population. As a result, the model may not be accurate or reliable. In addition, if the sample is not representative of the population, the model may be biased and may not generalize well to new data.

2047. Differentiate between classification and regression in Machine Learning.

Classification and regression are two types of machine learning tasks. Classification involves predicting a categorical or discrete label, while regression involves predicting a continuous or numerical value. For example, predicting whether an email is spam or not is a classification task, while predicting the price of a house is a regression task.

2048. What are the types of Machine Learning?

There are three types of machine learning:

Supervised learning: The model is trained using labeled data and is used to make predictions on new, unseen data. Examples include classification and regression.

Unsupervised learning: The model is trained using unlabeled data and is used to find patterns and relationships in the data. Examples include clustering and dimensionality reduction.

Reinforcement learning: The model learns from feedback in an environment to maximize a reward. Examples include game playing and robotics.

2049. Why learning rate should be less in gradient descent?

In gradient descent, the learning rate controls the step size taken in the direction of the gradient. If the learning rate is too high, the algorithm may fail to converge or may oscillate around the optimal solution. If the learning rate is too low, the algorithm may take too long to converge or may get stuck in local optima. Therefore, it is recommended to use a smaller learning rate to ensure convergence while avoiding overshooting the optimal solution.

2050. How to identify given data is structured or unstructured?

Structured data refers to data that is organized in a predefined format, such as spreadsheets or databases, where each variable has a specific meaning and is stored in a separate column.

Unstructured data, on the other hand, refers to data that has no predefined format or organization, such as text, images, or audio.

2051. How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression?

To validate a model that predicts a quantitative outcome variable using multiple regression, one can use techniques such as cross-validation, where the data is split into training and validation sets, and the model is trained on the training set and evaluated on the validation set. Other techniques include assessing the goodness of fit of the model using metrics such as R-squared or root mean squared error, and checking for violations of the assumptions of the regression model, such as linearity and homoscedasticity.

2052. How would you optimize a web crawler to run much faster, extract better information, and better summarize data to produce cleaner databases?

To optimize a web crawler, one can use techniques such as parallel processing, caching, and load balancing. Parallel processing can speed up the crawling process by splitting the workload

across multiple threads or processes. Caching can reduce the number of requests made to the server by storing previously crawled pages in memory or on disk. Load balancing can distribute the crawling workload across multiple machines to increase throughput and reduce the load on individual machines.

2053. For tuning hyperparameters of your machine learning model, what will be the ideal seed?

While working on a data set, how do you select important variables?

The ideal seed for tuning hyperparameters of a machine learning model depends on the specific problem and the algorithm being used. In general, it is recommended to try multiple seeds and choose the one that gives the best performance on a validation set.

To select important variables in a data set, one can use techniques such as feature selection or feature importance. Feature selection involves selecting a subset of the most relevant features to improve model performance and reduce overfitting. Feature importance involves ranking the importance of each feature based on their contribution to the model's predictions, using techniques such as permutation importance or SHAP values. Other techniques include correlation analysis and principal component analysis.

2054. We know that one hot encoding increases the dimensionality of a dataset, but label encoding doesn't. How?

One-hot encoding increases the dimensionality of a dataset because it creates a binary variable for each category in a categorical variable. Label encoding, on the other hand, assigns a unique integer to each category, which does not increase the dimensionality of the dataset.

2055. What are Loss Function and Cost Functions? Explain the key Difference Between them?

When calculating loss we consider only a single data point, then we use the term loss function.

Whereas, when calculating the sum of error for multiple data then we use the cost function.
There is no major difference.

In other words, the loss function is to capture the difference between the actual and predicted values for a single record whereas cost functions aggregate the difference for the entire training dataset.

The Most commonly used loss functions are Mean-squared error and Hinge loss.

Mean-Squared Error(MSE): In simple words, we can say how our model predicted values against the actual values.

$$\text{MSE} = \sqrt{(\text{predicted value} - \text{actual value})^2}$$

Hinge loss: It is used to train the machine learning classifier, which is

$$L(y) = \max(0, 1 - yy)$$

Where $y = -1$ or 1 indicating two classes and y represents the output form of the classifier. The most common cost function represents the total cost as the sum of the fixed costs and the variable costs in the equation $y = mx + b$

2056. What is Ensemble learning?

Ensemble learning is a method that combines multiple machine learning models to create more powerful models.

There are many reasons for a model to be different. Few reasons are:

Different Population

Different Hypothesis

Different modeling techniques

When working with the model's training and testing data, we will experience an error. This error might be bias, variance, and irreducible error.

Now the model should always have a balance between bias and variance, which we call a bias-variance trade-off.

This ensemble learning is a way to perform this trade-off.

There are many ensemble techniques available but when aggregating multiple models there are two general methods:

Bagging, a native method: take the training set and generate new training sets off of it.

Boosting, a more elegant method: similar to bagging, boosting is used to optimize the best weighting scheme for a training set.

2057. How do you make sure which Machine Learning Algorithm to use?

It completely depends on the dataset we have. If the data is discrete we use SVM. If the dataset is continuous we use linear regression.

So there is no specific way that lets us know which ML algorithm to use, it all depends on the exploratory data analysis (EDA).

EDA is like “interviewing” the dataset; As part of our interview we do the following:

Classify our variables as continuous, categorical, and so forth.

Summarize our variables using descriptive statistics.

Visualize our variables using charts.

Based on the above observations select one best-fit algorithm for a particular dataset.

2058. How to Handle Outlier Values?

An Outlier is an observation in the dataset that is far away from other observations in the dataset. Tools used to discover outliers are

Box plot

Z-score

Scatter plot, etc.

Typically, we need to follow three simple strategies to handle outliers:

We can drop them.

We can mark them as outliers and include them as a feature.

Likewise, we can transform the feature to reduce the effect of the outlier.

2059. What are Recommender Systems?

A recommendation engine is a system used to predict users' interests and recommend products that are quite likely interesting for them.

Data required for recommender systems stems from explicit user ratings after watching a film or listening to a song, from implicit search engine queries and purchase histories, or from other knowledge about the users/items themselves.

2060. How do check the Normality of a dataset?

Visually, we can use plots. A few of the normality checks are as follows:

Shapiro-Wilk Test

Anderson-Darling Test

Martinez-Iglewicz Test

Kolmogorov-Smirnov Test

D'Agostino Skewness Test

27. Can logistic regression use for more than 2 classes?

No, by default logistic regression is a binary classifier, so it cannot be applied to more than 2 classes. However, it can be extended for solving multi-class classification problems (multinomial logistic regression)

2061. Explain Correlation and Covariance?

Correlation is used for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related. Examples like, income and expenditure, demand and supply, etc.

Covariance is a simple way to measure the correlation between two variables. The problem with covariance is that they are hard to compare without normalization.

2062. What are Parametric and Non-Parametric Models?

Parametric models will have limited parameters and to predict new data, you only need to know the parameter of the model.

Non-Parametric models have no limits in taking a number of parameters, allowing for more flexibility and to predict new data. You need to know the state of the data and model parameters.

2063. What is Reinforcement Learning?

Reinforcement learning is different from the other types of learning like supervised and unsupervised. In reinforcement learning, we are given neither data nor labels. Our learning is based on the rewards given to the agent by the environment.

2064. Difference Between Sigmoid and Softmax functions?

The sigmoid function is used for binary classification. The probabilities sum needs to be 1. Whereas, Softmax function is used for multi-classification. The probabilities sum will be 1.

2065. How should outlier values be handled?

An observation in the dataset that is pretty far from the others in the dataset is known as an outlier. The following tools can be used to discover outliers:

Box plot

Z-score

Scatter plot, etc.

Usually, three simple strategies can be followed to handle outliers:

Drop them.

Mark them as outliers and then include them as a feature.

Similarly, the feature can be transformed to decrease the effect of the outlier.

2066. Explain How a System Can Play a Game of Chess Using Reinforcement Learning.

Reinforcement learning has an environment and an agent. The agent performs some actions to achieve a specific goal. Every time the agent performs a task that is taking it towards the goal, it is rewarded. And, every time it takes a step that goes against that goal or in the reverse direction, it is penalized.

Earlier, chess programs had to determine the best moves after much research on numerous factors. Building a machine designed to play such games would require many rules to be specified.

With reinforced learning, we don't have to deal with this problem as the learning agent learns by playing the game. It will make a move (decision), check if it's the right move (feedback), and keep the outcomes in memory for the next step it takes (learning). There is a reward for every correct decision the system takes and punishment for the wrong one.

2067. When Will You Use Classification over Regression?

Classification is used when your target is categorical, while regression is used when your target variable is continuous. Both classification and regression belong to the category of supervised machine learning algorithms.

Examples of classification problems include:

Predicting yes or no

Estimating gender

Breed of an animal

Type of color

Examples of regression problems include:

Estimating sales and price of a product

Predicting the score of a team

Predicting the amount of rainfall

2068. How Do You Design an Email Spam Filter?

Building a spam filter involves the following process:

The email spam filter will be fed with thousands of emails

Each of these emails already has a label: 'spam' or 'not spam.'

The supervised machine learning algorithm will then determine which type of emails are being marked as spam based on spam words like the lottery, free offer, no money, full refund, etc. The next time an email is about to hit your inbox, the spam filter will use statistical analysis and algorithms like Decision Trees and SVM to determine how likely the email is spam. If the likelihood is high, it will label it as spam, and the email won't hit your inbox. Based on the accuracy of each model, we will use the algorithm with the highest accuracy after testing all the models.

2069. What is the Trade-off Between Bias and Variance?

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, variance, and a bit of irreducible error due to noise in the underlying dataset.

Necessarily, if you make the model more complex and add more variables, you'll lose bias but gain variance. To get the optimally-reduced amount of error, you'll have to trade off bias and variance. Neither high bias nor high variance is desired.

High bias and low variance algorithms train models that are consistent, but inaccurate on average.

High variance and low bias algorithms train models that are accurate but inconsistent.

2070. Briefly Explain Logistic Regression.

Logistic regression is a classification algorithm used to predict a binary outcome for a given set of independent variables.

The output of logistic regression is either a 0 or 1 with a threshold value of generally 0.5. Any value above 0.5 is considered as 1, and any point below 0.5 is considered as 0.

2071. What is the difference between Lasso and Ridge regression?

Lasso(also known as L1) and Ridge(also known as L2) regression are two popular regularization techniques that are used to avoid overfitting of data. These methods are used to penalize the coefficients to find the optimum solution and reduce complexity. The Lasso regression works by penalizing the sum of the absolute values of the coefficients. In Ridge or L2 regression, the penalty function is determined by the sum of the squares of the coefficients.

2072. What is the Bayes' Theorem? Why do we use it?

Bayes' Theorem is how we find a probability when we know other probabilities. In other words, it provides the posterior probability of a prior knowledge event. This theorem is a principled way of calculating conditional probabilities.

In ML, Bayes' theorem is used in a probability framework that fits a model to a training dataset and for building classification predictive modeling problems (i.e. Naive Bayes, Bayes Optimal Classifier).

2073. Explain difference between Type I and Type II error.

A Type I error is a false positive (claiming something has happened when it hasn't), and a Type II error is a false negative (claiming nothing has happened when it actually has).

2074. What is the difference between a discriminative and a generative model?

A discriminative model learns distinctions between different categories of data. A generative model learns categories of data. Discriminative models generally perform better on classification tasks.

2075. What are parametric models? Provide an example.

Parametric models have a finite number of parameters. You only need to know the parameters of the model to make a data prediction. Common examples are as follows: linear SVMs, linear regression, and logistic regression.

Non-parametric models have an unbounded number of parameters to offer flexibility. For data predictions, you need the parameters of the model and the state of the observed data. Common examples are as follows: k-nearest neighbors, decision trees, and topic models.

2076. Explain LDA for unsupervised learning.

Latent Dirichlet Allocation (LDA) is a common method for topic modeling. It is a generative model for representing documents as a combination of topics, each with their own probability distribution.

LDA aims to project the features of higher dimensional space onto a lower-dimensional space. This helps to avoid the curse of dimensionality.

2077. Explain how you would develop a data pipeline.

Data pipelines enable us to take a data science model and automate or scale it. A common data pipeline tool is Apache Airflow, and Google Cloud, Azure, and AWS are used to host them.

For a question like this, you want to explain the required steps and discuss real experience you have building data pipelines.

The basic steps are as follows for a Google Cloud host:

- Sign into Google Cloud Platform
- Create a compute instance
- Pull tutorial contents from GitHub
- Use AirFlow for an overview of the pipeline
- Use Docker to set up virtual hosts
- Develop a Docker container
- Open Airflow UI and run the ML pipeline
- Run the deployed web app

2078. How do you fix high variance in a model?

If the model has low variance and high bias, we use a bagging algorithm, which divides a data set into subsets using randomized sampling. We use those samples to generate a set of models with a single learning algorithm.

Additionally, we can use the regularization technique, in which higher model coefficients are penalized to lower the complexity overall.

2079. What are hyperparameters? How do they differ from model parameters?

A model parameter is a variable that is internal to the model. The value of a parameter is estimated from training data.

A hyperparameter is a variable that is external to the model. The value cannot be estimated from data, and they are commonly used to estimate model parameters.

- You are working on a dataset. How do you select important variables?
- Remove correlated variables before selecting important variables
- Use Random Forest and a plot variable importance chart
- Use Lasso Regression
- Use linear regression to select variables based on p values
- Use Forward Selection, Stepwise Selection, and Backward Selection

2080. What is the default method for splitting in decision trees?

The default method is the Gini Index, which is the measure of impurity of a particular node. Essentially, it calculates the probability of a specific feature that is classified incorrectly. When the elements are linked by a single class, we call this “pure”.

You could also use Random Forest, but the Gini Index is preferred because it isn't computationally intensive and doesn't involve logarithm functions.

2081. You are told that your regression model is suffering from multicollinearity. How do verify this is true and build a better model?

You should create a correlation matrix to identify and remove variables with a correlation above 75%. Keep in mind that our threshold here is subjective.

You could also calculate VIF (variance inflation factor) to check for the presence of multicollinearity. A VIF value greater than or equal to 4 suggests that there is no multicollinearity. A value less than or equal to 10 tells us there are serious multicollinearity issues.

You can't just remove variables, so you should use a penalized regression model or add random noise in the correlated variables, but this approach is less ideal.

Chap