CASE STUDY, GUESSTIMATES, SQL, PYTHON, R, MS EXCEL, AND STATISTICS

CRACK YOUR NEXT DATA SCIENCE INTERVIEW 300+ QUESTIONS

BY THE DATA MONK

DATA SCIENCE MASTERBOOK

Crack your next Data Science Interview

For the last few weeks we have been working extensively with Data Scientists to find out the type of questions they are asked in a Data Scientist interview. Those who are aware of The Data Monk (www.thedatamonk.com) already know that we have a pool of 100s of Data Scientists who have provided us with valuable input on How to get into the best Data Science companies.

Our colleagues are working in Amazon, Google, ZS, Mu Sigma, Microsoft, Book My Show, etc.

We are publishing this book with the help of their experience and expertise. For any information or feedback visit at www.thedatamonk.com or mail us at contact@thedatamonk.com

List of The Data Monk books on Amazon

1. The Monk who knew Linear Regression (Python): Understand, Learn and Crack Data Science Interview
2. 100 Python Questions to crack Data Science/Analyst Interview
3. Complete Linear Regression and ARIMA Forecasting project using R
4. 100 Hadoop Questions to crack data science interview: Hadoop Cheat Sheet
5. 100 Questions to Crack Data Science Interview
6. 100 Puzzles and Case Studies To Crack Data Science Interview
7. 100 Questions To Crack Big Data Interview
8. 100 Questions to Learn R in 6 Hours
9. Complete Analytical Project before Data Science interview
10. 112 Questions To Crack Business Analyst Interview Using SQL
11. 100 Questions To Crack Business Analyst Interview
12. A to Z of Machine Learning in 6 hours
13. In 2 Hours Create your first Azure ML in 23 Steps
14. How to Start A Career in Business Analysis
15. Web Analytics - The Way we do it
16. Write better SQL queries + SQL interview Questions
17. How To Start a Career in Data Science
18. Top Interview Questions And All About Adobe Analytics
19. Business Analyst and MBA Aspirant's Complete Guide to Case Study - Case Study Cheat sheet
20. What do they ask in top Data Science Interviews?
21. What do they ask in top Data Science Interviews Part 2 ?
22. 100 Questions to Master Forecasting in R
23. Python Master Book

A typical interview for a Data Scientist role will have the following rounds:-

1. Telephonic Basic information round
2. Aptitude round
3. SQL and Excel
4. Guesstimate and Case Study
5. Project discussion
6. Statistics or Language(R/Python round)
7. Human Resource round

This does not imply that every company will have all these rounds. In general there are 4-5 rounds in any recruitment drive.

P.S. – More than 80% of the questions given below were asked in recruitment drive of companies like Amazon, BookMyShow, Accenture, Cognizant, Sapient, Deloitte, OYO Rooms, Flipkart, Myntra, etc.
We have tried to give ample number of examples to help you getting the vibe of these interviews

1.Telephonic Information Round

As soon as your resume is selected by the company or a third party recruitment firm. You will get a call to see if you are available for the telephonic round.
This round is mostly taken by the HR team of the recruiter. The common questions asked in this interview are:

1. Where do you work at?
2. He/She will give some information about the company and the requirement of the company
3. Are you comfortable with the tools and technologies that we use? Like SQL, R/Python, etc.
4. Why do you want to leave the current company?
5. What is your current CTC?
6. What is your expected CTC?
7. What is your notice period?

There is very little scope of asking anything other than these questions. The call will end with a description of the next round and the mode of the next round(Telephonic/Face-to-Face)

Bonus Tip – Do ask the HR about the complete recruitment process and what should you work on for the interview. They might drop you a sample assignment. Shot in the dark !!

2. Aptitude Test – Elimination Round

Now a days, more than 50% of the companies wants to eliminate the major chunk of candidates with the aptitude test. We have personally seen people getting eliminated after scoring more than 70% marks in the test. So be aware of it.

The aptitude test is either a written test or a HackerRank test.

Following are the questions asked in Tredence Incorporation and Bounce. The difficulty level of the questions are mostly easy to moderate.

You can easily find these questions on various websites, but there is a plethora of questions and you need to know the right type of questions which are asked in these rounds. Whether it's an online test or a centre test, you need to have good speed to solve as many questions as possible. So practice these questions before an aptitude round.

Time – 30 minutes
Questions – 25

1.
T is older than E.
C is older than T.
E is older than C.

If first two statements are true, then the third statement is?

a. True
b. False
c. Uncertain

Ans.
B i.e False as E is the youngest of the three.

2. Two cards are drawn from a pack of 52 cards, What is the probability that both of the cards are being Kings?

Ans.
$n(S) = 52C_2 = (52*51)/(2*1) = 1326$
E = Event of getting 2 kings out of 4
$n(E) = 4C_2 = 6$

$P(E) = n(E)/n(S) = 1/221$

3. 7,10,8,11,9,12, Next Term ?

Ans.
In the first pattern, 3 is added, in the second,2 is subtracted. So, the answer is 10.

4. How many 5 letter words can be formed using BIHAR?

Ans.
5! = 120

5. Two unbiased coins are tossed, What is the probability of getting at most one head?

Ans.
S = {HH,HT,TH,TT}
E = {HH,HT,TH}
P(E) = n(E)/n(S) = ¾

6. Find Cost Price when Selling Price is Rs. 40.60 and profit is 16%.

Ans.
C.P. = Rs.(100/116)*40.60 = Rs. 35

7. A bag contains 4 white, 5 red and 6 blue balls. Three balls are drawn at random from the bag. The probability that all of them are red is:

n(S) = number of ways of drawing 3 balls out of 15
n(S) = 15C3 = (15 * 14 * 13)/(3 * 2 * 1) = 455
n(E) = event of getting all the 3 red balls
n(E) = 5C3 = 5C2 = (5 * 4)/(2 * 1) = 10
P(E) = n(E)/n(S) = 10/455 = 2/91

8. A can finish a work in 18 days and B can do the same work in half the time take by A. Then, working together, what part of the same work they can finish in a day?

A's 1 day's work = 1/18
B's 1 day's work = 1/9
(A + B)'s 1 day's work = (1/18) + (1/9) = 1/6

9. In how many ways can a cricket eleven be chosen out of a batch of 15 players?

Ans.

$1365 = 15C_{11}$

10. A takes twice as much time as B or thrice as much time to finish a piece of work. Working together, they can finish the work in 2 days. B can do the work alone in:

suppose A, B and C take x, x/2 and x/3 hrs respectively to finish the work
then, (1/x) + (2/x) + (3/x) = ½ .i.e. 6/x = ½
x = 12 hrs, B = x/2 = 12/2 = 6 hrs

11. If x=y=2z and xyz=256 then what is the value of x
(a)12
(b)8
(c)16
(d)6

Ans. (b)

12. Pipe A can fill in 20 minutes and Pipe B in 30 mins and Pipe C can empty the same in 40 mins. If all of them work together, find the time taken to fill the tank
(a) 17 1/7 mins
(b) 20 mins
(c) 8 mins
(d) none of these

Ans. (a)

13. (51+52+53+……+100) is equal to:
(a) 2525
(b) 2975
(c) 3225
(d) 3775

Use of formula sum of n natural no.
=((n*(n+1))/2) [51+52+53+54+.....+100]
=(sum(1 to 100)) - (sum(1 to 50))
=((100*101)/2) - ((50*51)/2)
=5050-1275 =3775
Answer is D

14. If A is thrice as fast as B and together can do a work in 21 days. In how many days A alone can do the work?
a. 36
b. 42
c. 28
d. 54

15. How to calculate Mode, Median, and Mean from a given number

Numbers – 2,3,6,6,6,9,11

Mean = 44/7 = 6.28
Median = 6
Mode = 6

16. Find the average of all the numbers between 6 and 34 which are divisible by 5
a. 18
b. 20
c. 24
d. 30

Multiples of 5 between 6 and 34 are 10,15,20,25,30
Average = (10+15+20+25+30)/5 =5(10+30)/2*5 = 40/2 =20

17. A man can row a boat at 10 km/hr in still water and the speed of the stream is 8 km/hr. What is the time taken to row a distance of 90 km down the stream?
a) 8hrs
b)5 hrs
c) 15 hrs
d) 20 hrs

Speed in down stream = 10 +8 = 18
Time taken to cover 90 km down stream = 90/18 = 5 hrs.

3. SQL and Excel

This is one of the most important round. You have to get the best feedback out of this round(be it a telephonic round or a face-to-face round). We have did extensive research and figures out the areas from which questions are asked in the interview. Following are the questions which will definitely help you in understanding the type of questions asked in this round.

First we will deal with some very basic questions. We will be using one table and there will be 30 questions from these tables. The questions will start with the most basic ones, you can skip these 30 questions if you think you are good with SQL.

Below is the employee(emp) table

EmpNo	EName	Job	MGR	HireDate	Sal	Comm	DeptNo
1234	Amit	Waiter	8382	19-Oct-18	50000	500	50
5678	Ashish	Analyst	8635	2-Nov-18	60000	200	51

Below is the department(dept) table

DeptNo	Dname	Loc
50	Service	Delhi
51	Account	Mumbai

18. Show all the data from emp table

SELECT * FROM emp;

19. Show all the data from Dept table

SELECT * FROM dept;

19. Display distinct jobs from Dept table

SELECT DISTINCT(job) FROM dept;

20. Number of employees

SELECT COUNT(*) FROM emp;

21. List the employee in the ascending order of salary

SELECT * from emp ORDER BY Sal;

22. Show the employee information of the Managers

SELECT * from emp WHERE EmpNo in (SELECT MGR FROM emp);

23. List of employees who were hired before 2018.

SELECT * FROM emp WHERE HireDate < ’01-Jan-2018′;

24. List the detail of employees along with the annual salary, order it on the annual salary

```
SELECT *, sal*12 as Annual_Income
FROM emp
ORDER BY Annual_Income;
```

25. Display number of months of experience of all the Managers

```
SELECT *, months_between(sysdate,HireDate) as Exp
FROM emp
WHERE EmpNo IN (SELECT MGR FROM emp);
```

26. Display the name of the employees with Commission(Comm) less than Salary (Sal)

```
SELECT EName
FROM emp
WHERE Comm < Sal;
```

27. Display the name of the employee with Daily income more than 200

```
SELECT EName
FROM emp
WHERE (sal/30)>200;
```

28. Show information of all the Waiters

```
SELECT *
FROM emp
WHERE Job = 'Waiter';
```

29. Show all the employee who joined on 01-Aug-2018, 4-Aug-2018, 29-Oct-2018 in descending order of Hire Date

```
SELECT *
FROM emp
WHERE HireDate IN ('01-Aug-2018','04-Aug-2018','29-OCt-2018')
ORDER BY HireDate DESC;
```

30. List the employees who joined in 2018

```
SELECT *
FROM emp
WHERE HireDate BETWEEN (’01-Jan-2018′) AND (’31-Dec-2018′);
```

31. Employees with Annual Salary between 600000 and 1000000

```
SELECT *
FROM emp
WHERE Sal*12 BETWEEN 600000 AND 1000000;
```

32. List the employees with name starting with N and containing 5 alphabets

```
SELECT *
FROM emp
WHERE EName LIKE ‘N____’;
```

Or

```
SELECT *
FROM emp
WHERE EName LIKE ‘N%’ AND len(EName) = 5;
```

33. List the employee with the third alphabet in their name as K

```
SELECT *
FROM emp
WHERE Upper(EName) LIKE '__K%';
```

34. Show the name of the employees who joined in August month of any year.

```
SELECT *
FROM emp
WHERE to_char(HireDate,'mon')='Aug'
```

35. Show the employee details of those who were hired in the 90s

```
SELECT *
FROM emp
WHERE to_char(HireDate, 'yy') LIKE '9_';
```

36. Show the employee who were not hired in the month of October.

```
SELECT *
FROM emp
WHERE to_char(HireDate,'MON') NOT IN '('Oct');
```

37. List the total information of the employees along with DName and Location of people working under ‘Accounts’

```
SELECT *
FROM emp e
INNER JOIN dept d ON (e.DeptNo = d.DeptNo)
WHERE d.DName = ‘Account’
```

38. List all the employees with more than 10 years of experience as of now

```
SELECT *
FROM emp
WHERE TIMESTAMPDIFF(MONTH, HireDate, sysdate)
```

39. List the detail of all the employees whose salary is less than that of Aman

```
SELECT *
FROM emp
WHERE sal > (SELECT sal FROM emp WHERE EName = ‘Aman’);
```

40. Show the name of those employees who are senior to their own Manager.

```
SELECT *
FROM emp w, emp m
WHERE w.MGR = m.EmpNo and w.HireDate < m.HireDate
```

Or

```
SELECT *
FROM emp w, emp m
WHERE w.EmpNo = m.MGR and w.HireDate < m.HireDate
```

41. Show the employees who are senior to Aman

```
SELECT *
FROM emp
WHERE HireDate < (SELECT HireDate FROM emp WHERE EName =
'Aman')
```


42. Show the employees who are senior to Aman and are working in Delhi or Bangalore

```
SELECT *
FROM emp e, dept d
WHERE UPPER(d.loc) IN ('DELHI,'BANGALORE') AND e.DeptNo =
d.DeptNo
AND e.HireDate < (SELECT e.HireDate FROM emp e WHERE EName =
'Aman');
```

43. Show the employees with the same job as Aman or Amit.

```
SELECT *
FROM emp
WHERE job in (SELECT job from emp WHERE EName IN
('Aman','Amit');
```

44. Find the highest salary of any employee

```
SELECT MAX(Sal)
FROM emp;
```

45. Find the detail of the employee with the minimum pay

```
SELECT *
FROM emp
WHERE Salary = (SELECT MIN(Salary) FROM emp);
```

46. Show the detail of the recently hired employee working in Delhi

Try it yourself

SQL Tricky interview Questions

The above questions were to make sure that you are good with the basics. The below questions are asked mostly in the interviews

46. Explain the difference between DENSE_RANK and RANK function
Try it yourself

46. What is the difference between NVL and NVL2?
Ans. In SQL, NVL() converts a null value to an actual value. Data types that can be used are date, character and number. Data type must match with each other i.e. expr1 and expr2 must of same data type.

NVL (expr1, expr2)
expr1 is the source value or expression that may contain a null.
expr2 is the target value for converting the null.

NVL2(expr1, expr2, expr3) : The NVL2 function examines the first expression. If the first expression is not null, then the NVL2 function returns the second expression. If the first expression is null, then the third expression is returned i.e. If expr1 is not null, NVL2 returns expr2. If expr1 is null, NVL2 returns expr3. The argument expr1 can have any data type.

NVL2 (expr1, expr2, expr3)
expr1 is the source value or expression that may contain null
expr2 is the value returned if expr1 is not null
expr3 is the value returned if expr2 is null

47. What is the function of COALESCE()?

The COALESCE() function examines the first expression, if the first expression is not null, it returns that expression; Otherwise, it does a COALESCE of the remaining expressions.

The advantage of the COALESCE() function over the NVL() function is that the COALESCE function can take multiple alternate values. In simple words COALESCE() function returns the first non-null expression in the list.

48. How to find count of duplicate rows?

```
Select rollno, count (rollno) from Student
Group by rollno
Having count (rollno)>1
Order by count (rollno) desc;
```

49. How to find Third highest salary in Employee table using self-join?

```
Select * from Employee a Where 3 = (Select Count (distinct Salary) from
Employee where a.salary<=b.salary;
```

50. How to calculate number of rows in table without using count function?

```
SELECT table_name, num_rows
FROM user_tables
WHERE table_name='Employee';
```

51. What is the use of FETCH command?

The FETCH argument is used to return a set of number of rows. FETCH can't be used itself, it is used in conjuction with OFFSET.

```
SELECT column_name(s)
FROM table_name
ORDER BY column_name
OFFSET rows_to_skip
FETCH NEXT number_of_rows ROWS ONLY;
```

52. Write a query to find maximum salary of each department in an organization

```
SELECT Department_Name, Max(Salary)
FROM Department_Table
GROUP BY Department_Name
```

53. What is wrong with the following query?
SELECT Id, Year(PaymentDate) as PaymentYear
FROM Bill_Table
WHERE PaymentYear > 2018;

Though the variable PaymentYear has already been defined in the first line of the query, but this is not the correct logical process order. The correct query will be

SELECT Id, Year(PaymentDate) as PaymentYear
FROM Bill_Table
WHERE Year(PaymentDate) > 2018;

54. What is the order of execution in a query?

The order of query goes like this:-

FROM – Choose and join tables to get the raw data
WHERE – First filtering condition
GROUP BY – Aggregates the base data
HAVING – Apply condition on the base data
SELECT – Return the final data
ORDER BY – Sort the final data
LIMIT – Apply limit to the returned data

55. What is ROW_NUMBER() function?

It assigns a unique id to each row returned from the query ,even if the ids are the same. Sample query:-

```
SELECT emp.*,
row_number() over (order by salary DESC) Row_Number
from Employee emp;
```

Employee Name	Salary	Row_Number
Amit	7000	1
Bhargav	6000	2
Chirag	6000	3
Dinesh	5000	4
Esha	3000	5
Farhan	3000	6

Even when the salary is the same for Bhargav and Chirag, they have a different Row_Number, this means that the function row_number just gives a number to every row

56. What is RANK() function?

RANK() function is used to give a rank and not a row number to the data set. The basic difference between RANK() and ROW_NUMBER is that Rank will give equal number/rank to the data points with same value. In the above case, RANK() will give a value of 2 to both Bhargav and Chirag and thus will rank Dinesh as 4. Similarly, it will give rank 5 to both Esha and Farhan.

```
SELECT emp.*,
RANK() over (order by salary DESC) Ranking
from Employee emp;
```

57. What is NTILE() function?

NTILE() function distributes the rows in an ordered partition into a specific number of groups. These groups are numbered. For example, NTILE(5) will divide a result set of 10 records into 5 groups with 2 record per group. If the number of records is not divided equally in the given group, the function will set more record to the starting groups and less to the following groups.

```
SELECT emp.*,
NTILE(3) over (order by salary DESC) as GeneratedRank
from Employee emp
```

This will divide the complete data set in 3 groups from top. So the GeneratedRank will be 1 for Amit and Bhargav, 2 for Chirag and Dinesh: 3 for Esha and Farhan

58. What is DENSE_RANK() ?

This gives the rank of each row within a result set partition, with no gaps in the ranking values. Basically there is no gap, so if the top 2 employees have the same salary then they will get the same rank i.e. 1 , much like the RANK() function. But, the third person will get a rank of 2 in DENSE_RANK as there is no gap in ranking where as the third person will get a rank of 3 when we use RANK() function. Syntax below:-

```
SELECT emp.*,
DENSE_RANK() OVER (order by salary DESC) DenseRank
from Employee emp;
```

59. Write a query to get employees name starting with vowels.

```
SELECT EmpID,EmpName
FROM Employee
where EmpName like '[aeiou]%'
```

60. Write a query to get employee name starting and ending with vowels.

```
SELECT EmpID,EmpName
FROM Employee
where EmpName like '[aeiou]%[aeiou]'
```

61. What are the different types of statements supported in SQL?

There are three types of statements in SQL:-
a. DDL – Data Definition Language
b. DML – Data Manipulation Language
c. DCL – Data Control Language

62. What is DDL?

It is used to define the database structure such as tables. It includes 3 commands:-
a. Create – Create is for creating tables

```
CREATE TABLE table_name (
    column1 datatype,
    column2 datatype,
    column3 datatype,
   ….
);
```

b. Alter – Alter table is used to modifying the existing table object in the database.

```
ALTER TABLE table_name
ADD column_name datatype
```

c. Drop – If you drop a table, all the rows in the table is deleted and the table structure is removed from the database. Once the table is dropped, you can't get it back

63. What is DML?

Data Manipulation Language is used to manipulate the data in records. Commonly used DML commands are Update, Insert and Delete. Sometimes SELECT command is also referred as a Data Manipulation Language.

64. What is DCL?

Data Control Language is used to control the privileges by granting and revoking database access permission

65. What is the difference between DELETE and TRUNCATE?

66. What are the important SQL aggregate functions?
a. AVG()
b. COUNT()
c. MAX()
d. MIN()
e. SUM()

67. What is Normalization? How many Normalization forms are there?

Normalization is used to organize the data in such manner that data redundancy will never occur in the database and avoid insert, update and delete anomalies.
There are 5 forms of Normalization
First Normal Form (1NF): It removes all duplicate columns from the table. Creates a table for related data and identifies unique column values
Second Normal Form (2NF): Follows 1NF and creates and places data subsets in an individual table and defines the relationship between tables using a primary key
Third Normal Form (3NF): Follows 2NF and removes those columns which are not related through primary key
Fourth Normal Form (4NF): Follows 3NF and do not define multi-valued dependencies. 4NF also known as BCNF

68. How to fetch only common records between two tables?

SELECT * FROM Employee
INTERSECT
SELECT * FROM Employee1

69. Full Outer Join is a combination of which of the following:-
a. Left Outer and Left Inner Join
b. Left Outer and Right Inner Join
c. Left Outer and Right Outer Join
d. Left Outer and Right Outer Join

A. Full Outer Join is a combination of Left Outer and Right Outer Join in SQL

70. Right Outer Join is similar to:-
a. Right Inner Join
b. Left Inner Join
c. Left Outer Join
d. Right Outer Join

A. Right Outer Join is similar to Left Outer Join in SQL

71. What is the use of OFFSET command?
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.
OFFSET can only be used with ORDER BY clause. It cannot be used on its own.
OFFSET value must be greater than or equal to zero. It cannot be negative, else return error.

```
SELECT column_name(s)
FROM table_name
WHERE condition
ORDER BY column_name
OFFSET rows_to_skip ROWS;
```


72. Difference between DELETE and DROP command.
DROP is a Data Definition Language (DDL) command which removes the named elements of the schema like relations, domains or constraints and you can also remove an entire schema using DROP command.

```
DROP SCHEMA schema_name RESTRICT;

DROP Table table_name CASCADE;
```

DELETE is a Data Manipulation Language (DDL) command and used when you want to remove some or all the tuples from a relation. If WHERE clause is used along with the DELETE command it removes only those tuples which satisfy the WHERE clause condition but if WHERE clause is missing from the DELETE statement then by default all the tuples present in relation are removed.

```
DELETE FROM relation_name
WHERE condition;
```

There were some output related questions where a table was given, mostly

on group by, order by, top, etc. command

ADVANCE EXCEL

73. What are the ways to create a dynamic range?
Creating a Table
Using OFFSET and COUNTA Functions

74. What is the order of operations that Excel uses while evaluating formulas?
PEMDAS Rule
Parenthesis
Exponentiation
Multiplication/Division
Addition
Subtraction

75. Difference between FUNCTION and FORMULA in excel?
FORMULA – is a statement which written by the user for calculations. Ex: =1+2+3
FUNCTION – is a built-in formula by Excel. Ex: =SUM(1+2+3)
Q4)How will you find duplicate values in a column?
To highlight duplicate values – use Conditional Formatting
To get a number of duplicate values – use COUNTIF function.

76.What is the use of Slicer and Timeline in Excel?
Slicer – is used to filter the Table, Pivot Table data visually.
Timeline – is used to filter the dates interactively by Year, Month, Quarter and Day.
Q6)What is Recommended Pivot Tables and Recommended Charts?
Recommended Pivot Table – Based on the raw data, Excel will recommend some pivot table automatically.
Recommended Charts – Same like as above, Excel will recommend charts (Ex: Column chart, Bar Chart, etc…) based on the data.

77.What is the Name Manager in Excel?
Names which we give for a cell/Range, Table will be managed by the Name Manager.

78. What is COUNT and COUNTA?
COUNT – Counts the number of cells which contains only numbers except for blank cells.
COUNTA – Counts the number of cells which contains alpha-numeric except blank cells.
Q9)Is it possible to make a single Pivot Table for multiple data sources?
Yes, it is possible. Using Pivot Table Data Modeling technique.

79.VLOOKUP Vs INDEX-MATCH
VLOOKUP – Using VLOOKUP, we can retrieve the data from left to right in the range/Table.
INDEX-MATCH – Using a combination of INDEX and MATCH, we can retrieve the data from left to right/right to left in a range/table.

80. How would you get the data from different data sources?
Data > Get External Data section > Choose your data source

81. What is the use of Option Explicit in VBA?
Option Explicit will force the user to declare variables. If the user uses undeclared variables, an error occurs while compiling the code.

82. What is Excel object Model?
Application – Workbooks – Worksheets – Range

83. What is the default data type in VBA?
A variant is the default data type in VBA.

84. How will you fix/lock the cell/range reference?
Using $ symbol.
Ex: $A $1 – Here Locked Column A and row 1

85. Difference between Function and Subroutine in VBA?
The function will return a value whereas Subroutine may or may not return a value. The function can be called in the procedure. We can create custom functions using FUNCTION like as built-in functions.

86. Does VBA support OOPs concepts?
No, it will not. VBA is Object based programming language, not Object Oriented Programming language.

87. Difference between ThisWorkbook and ActiveWorkbook in VBA?
ThisWorkbook – is the workbook where the VBA code is written.
ActiveWorkbook – is the workbook which is in Active state at present.

88. How will you debug codes in VBA?
Step by step execution – F8
Breakpoints – F9
Using Debug.Print
Immediate Window
Watch Window

89. Explain – ADO, ODBC, and OLEDB
ADO –ActiveX Data Objects is a data access framework and is useful to get the data from the databases.
ODBC – Open Database Connectivity is useful to get the data from the external database.
OLEDB – Object Linking and Embedding, Database.

90. What are the different types of errors that you can encounter in Excel?
#N/A
#DIV/0!
#VALUE!
#REF!
#NAME
#NUM

91. What are volatile functions in Excel?
Volatile functions recalculate the formula, again and again, so Excel workbook performance will be slow down. Volatile functions recalculate the formulas when any changes happen in the worksheet.

Ex: NOW (), RAND ()

92. What is the difference between Report and Dashboard?
Dashboards: Dashboard is a visual display of the data and these are dynamic and live, so data is being updated in real time and visuals can show changes from minute to minute.

Reports: Reports are not live and we use historical data to make reports. sometimes Reports are included with visuals such as Table, Graphs and Charts, Text, Numbers or anything.

93. What are the structured references in Excel?
Structured references – Instead of using cell references, we can use an Excel table name or the Column name for reference.

94. Name some of the Excel formats which an Excel file can be saved?
.XLS
.XLSX
.XLSM (Macro-enabled workbook)
.XLSB (Binary format)
.CSV (Comma Separated Values)

95. How will you pass arguments to VBA Function?
In 2 ways we can pass arguments to VBA Functions
ByVal
ByRef

96. What are the modules available in VBE?
Code Module: Default module to write procedures.
User form: Helps to develop GUI (Graphical User Interface) applications.
Class Module: Allows to create new objects.

97. What is Collection in VBA?
The Collection object contains a set of related group of items as a single object.

98. What is the difference between an Array and Collection?
Collections and Arrays are used to group variables. In Arrays, before using to start adding elements we normally set the size. But in Collection, we will not set the size, because we don't know the number of elements in advance.

99. What are the modules available in Excel VBA?
Sheet Module
Thisworkbook Module
Code Module
Userforms
Class Modules

100. How to run faster your VBA macro?
Below are the some tips –

Turn off screenupdating
Declare the variables and avoid "Variant" data type.
Disable events
Using the WITH statement.
Avaoding the Select statement
Select Case instaed of If Then

101. What are the error handling techniques in Excel VBA?
Item
On Error Goto 0
On Error Resume Next
On Error Goto [Label]
Err Object
Err. Number
Err. Description
Err. Source
Err. Raise
Error Function
Error Statement

102. What are the ways in which a variable can be declared in the VBScript language?
Implicit Declaration: When variables are used without declaration is called Implicit declaration.

Explicit Declaration: Declaring variables before using is called Explicit Declaration.

103. Name some of the operators available in VBA?
Arithmetic Operators, Comparison Operators, Logical Operators etc.

104. What are the types of arrays available in VBA?
There are 2 types of arrays available in VBA.
Single Dimensional Array: Single Dimensional array is used more often in the VBA. An array uses only one index.
Multi-Dimensional Array: If an array has more than 1 dimension is called Multi-Dimensional Array.

105. Which object is used to work with Databases in the VBA?
Connection Objects are used to provide a connection between Excel and Databases with the help of ADODB Objects. So, we can interact with the database and can use the SQL queries to fetch the data from the database.

ADO stands for, ActiveX Data Objects.

106. Why VBA is not Object Oriented Programming language?
VBA does not support all the OOPs concepts (VBA will support Polymorphism and Encapsulation, not supports Inheritance). Hence, VBA is called Object Based Programming Language.

107. What is Conditional Formatting in Excel?
Conditional Formatting is used to format cells/Range based on a condition/Conditions.

Ex: Highlighting a cell based on cell Value

108. What is a Slicer?
Slicer is used to filter the Table, Pivot Table data. Instead of using Filters section in a Pivot Table, we can use Slicer.

109. What is a Goal Seek in Excel?
Goal Seek – is used to achieve your goal by changing the dependent value.

Ex: If you have taken a personal loan, and if you can able to pay the EMI of 6K instead of 10K, how many months do you need to close your personal loan?

110. What is Scenario Manager in Excel?
Scenario Manager – Excel can be the tool of choice when you have multiple variables, and you want to see the effect on the final result when these variables change.

111. What is a UDF in VBA?
UDF stands for User Defined Function, and these are custom functions. Using VBA, you can create your own functions and those can be used in Excel worksheets as normal built-in Functions.

112. What are the ways to run a macro?
Assigning a macro to a shape.
Assigning a macro to a button
Run a macro from the ribbon
Run a macro using a keyboard short cut key.

113. How do you remove duplicate spaces from a cell?
Using TRIM () function, we can delete duplicate spaces and gives unique/single space between words.

114. What is Data Validation in Excel?
Data Validation – is used to validate the Data in a cell/Range. In Data Validation, we have criteria such as List, Whole Number etc. And have custom criteria option where we can give function/formula.

115. How will you find the number of duplicate values in a range?
There might be different ways to find the duplicate values from a range. One of that is, using COUNTIF function we can find duplicate values.

116. Which function will you use instead of VLOOKUP?
Instead of VLOOKUP, we can use INDEX and MATCH function. Limits of VLOOKUP is we cannot fetch the data from the left side of the Lookup range. Using INDEX-MATCH, we can fetch the data any ways.

117. How would you add a new column to an existing pivot table for calculations?
Using Calculated Field

118. Can you name some Text Functions?
CONCATENATE( ) – used to join several text strings to one string
TEXT( ) – Converting a value into text formatting
PROPER ( ) – Arranging the characters in proper way.
LEFT ( ) – Returns the specified number of characters from the starting character.

Guesstimate

119.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*0.7*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*0.2*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*0.65*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*0.15*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.

120. 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 : 3/4th = 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*45=180 Mn
16–22 yr = 23 Mn, on an average individual own 4 t shirts -> 4*23 = 92 Mn
23–50 yr = 65 Mn, on an average individual own 3 t shirts -> 3*65 = 195 Mn
50 - 80 yr = 18 Mn, on an average individual have 2 t shirts -> 2*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*45=90 Mn
16–22 yr = 23 Mn, on an average individual own 4 t shirts -> 4*23 = 92 Mn
23–30 yr = 15 Mn, on an average individual own 3 t shirts -> 3*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 Mn ~ 100 Mn per year

Number of t-shirts sold in India per day(From ecommerce portal) = 100 * 10^6/365 ~ 27,000

121. 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 Mn -> 1.2 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+.6= 2.4 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 year (in India)
Number of sold per day in Bangalore = 0.9 Mn/ 365*5 ~ 500 laptops

122. What are the number of smartphones 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

123. 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

OYO Rooms Case Study

124. What are the KPIs for OYO rooms?

- The main KPIs for any online room booking company could be:-

a. Online Rating
b. Occupancy %
c. Average daily Rate
d. Revenue per available room
e. Customer Satisfaction
f. Advertising ROI

125. How do calculate the average daily rate?

- If I were to calculate, it should be equal to Total Revenue Per Room/Total rooms occupied

126. What are the types of marketing scheme?

- Banners on websites
- Video ads on youtube
- Pamphlet distribution
- Hoarding
- Email/SMS

127. Any idea about Average Length of Stay?

- The ALOS metrics makes it easy to identify the length of stay of guests at your hotel. This is calculated by dividing the occupied rooms by a number of bookings. It is said that a higher number means an improved profit as less labor is required. On the other hand, a lower ALOS results in reduced profit. The concept is that if a guest stays for a long period of time then it requires less labor. Whereas if several guests book rooms for one-nights for the same period of time then it requires more labor.

128. What is Market Penetration Index?

- To stay ahead of the competition you need to know how your hotel is performing in the local market. The MPI metrics can be used as a tool to compare your hotel's market share with your competitors. It helps you to know how many guests are choosing your hotel as compared to other hotels in your location. It can be calculated by dividing your hotel's occupancy by market occupancy and multiplying by 100. If the result is more than 100 that means you have a very good hold on the market. Else if it is less than 100 then it indicates your hotel isn't performing well and losing a lot of bookings to your competitors.

129. 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

130. The client of our company is The Minnesota Mining Manufacturing Company of United States whose main office is in Minnesota, United States. The company manufactures products such as car décor products, medical products, adhesives, electronic products and dental products etc. On a global level, the company is a thriving one with its employee population of over 84000 and product types of over 55000 and business in over 60 countries. One of their major investments is in Brazil, which is the manufacturing of a particular kind of steel that is only produced by two other companies in Brazil. Throughout the world, Steel has an amazing market capture and increasing demand. Now, the company has hired BCG to frame a plan for the progress of this business only after acquiring a proper knowledge of the market trend. What would you do about it?

Possible answer:
The candidate would begin by discussing the market dynamics in Brazil as well as globally on which he/she has to base the suggestion. Furthermore, an idea is to be framed up about the cost, market, value, customers, transportation facility and price if the steel is to be exported. Also, Brazil has some taxes on foreign goods export which would only add up to the price. Since the local market is more profitable than the international trade, it is advisable to try out the products first in the local market of Brazil since there is a chance of price war.

R Language

131. How can you combine 2 vectors?
Ans.) Vectors can be combined from 2 to 1 by using the c() function
Example.
> first <- c(1,2,3,4)
> second <- (“a”, “b”, “c”)
> third <- c(first, second)
> print(third)
[1] “1” “2” “3” “4” “a” “b” “c”

132. What is the difference between Matrix and an array ?
Ans.) Matrix can have only 2 dimensions where as an array can have as many dimensions as you want. Matrix is defined with the help of data, number of rows, number of columns and whether the elements are to be put in row wise or column wise.
In array you need to give the dimension of the array. An array can be of any number of dimensions and each dimension is a matrix. For example a 3x3x2 array represents 2 matrices each of dimension 3x3.

133. What is the difference between a matrix and a dataframe?
Ans.)A dataframe can contain vectors with different inputs and a matrix cannot. (You can have a dataframe of characters, integers, and even other dataframes, but you can't do that with a matrix.A matrix must be all the same type.)
So, the data frame can have different vector of character, numbers, logical, etc. and it is still cool. But, for matrix you need only one type of data type. Phewww !!

134. Explain general format of Matrices in R?

Ans.) General format is

>Temp_matrix< – matrix (vector, nrow=r ,ncol=c , byrow=FALSE, dimnames = list (char_vector_ rowname, char_vector_colnames))

135. Define repeat.

Ans.) Repeat loop executes a sequence of statement multiple times. It don't put the condition at the same place where we put the keyword repeat.

Example

```
> name <- c("Pappu", "John")
> temp <- 5
> repeat {
print(name)
temp <- temp+2

if(temp > 11) {
break
}
}
```


So, this will return the name vector 4 times. First it prints the name and then increase the temp to 7 and so on.

136. Define while.
In the while loop the condition is tested and the control goes into the body only when the condition is true

Example

```
> name <- c("Pappu", "John")
> temp <- 5
> repeat (temp<11) {
print(name)
temp <- temp+2
}
```

The name will be printed 4 times

137. Define the for loop.
Ans.) The for loop are not limited to integers. You can pass character vectors, logical vectors, lists or expressions.
Example.

```
> x<- LETTERS[1:2]
for ( i in x) {
print(i)
}
[1] "A"
[2] "B"
```

138. What is the syntax of a function in R?
Ans.)

Name_of_function<- function(argument_1,argument_2,..)
{
function body
}
Argument is the place holder , whenever a function is invoked, it pass a value to the argument. Arguments are optional

139. What is the use of sort() function? How to use the function to sort in descending order?
Ans.) Elements in a vector can be sorted using the function sort()
Example.

> temp <- c(3,5,2,6,7,1)
>sort_temp<- sort(temp)
> print(sort_temp)
[1] 1 2 3 5 6 7
>rev_sort<- sort(temp, decreasing = TRUE)
[1] 7,6,5,3,2,1

This function also works with the words

140. What is the use of subset() function and sample() function in R ?
Ans.) In R, subset() functions help you to select variables and observations while through sample() function you can choose a random sample of size n from a dataset.

**Suppose there is adata.frame
data_frame_example<- data.frame(a = c(10, 20, 30), b = c(40, 50, 60), c(70, 80, 90))...

141. Determine the output of the following function f(2).

```
b <- 4
f <- function(a)
{
b <- 3
b^3 + g(a)
}
g <- function(a)
{
a*b
}
```


Ans.) The global variable b has a value 4. The function f has an argument 2 and the function's body has the local variable b with the value 3. So function f(2) will return 3^3 + g(2) and g(2) will give the value 2*4 = 8 where 4 is the value of b.
Thus, the answer is 35

142. How to set and get the working directory in R ?
Ans.)setwd() and getwd() functions are used to set a working directory and get the working directory for R.
setwd() is used to direct R to perform any action in that location and to directly import objects from there itself.
getwd() is used to see which is the current working directory of R

143. Get all the data of the person having maximum salary.
Ans.)

```
max_salary_person<- subset(data, salary == max(salary))
print(max_salary_person)
```

144. Now create an output file which will have data of all the people who joined TCS in 2016 with salary more than 300000
Ans.)

```
temp <- subset(data, company==”TCS” &  salary > 300000 &as.Date(DOJ) >as.Date(“2016-01-01”))
write.csv(temp,”output.csv”,row.names = FALSE)
new_temp<- read.csv(“output.csv”)
print(new_temp)
```

145. How to combine multiple vectors in one data frame?
Ans.)
Example.
a <- c(1,2,3,4)
b<- c("Amit","Sumit","Gaurav")
c<- c("TCS","CTS","Musigma")
df<- cbind(a,b,c)
print(df)

146. What is the function of merge() function?
Ans.) We can merge two data frames by using the **merge()** function. The data frames must have same column names on which the merging happens.
Example.
df1<- data.frame(id<- c(1:6), name <- c(rep("Amit",3), rep("Sumit",3))
df2<- data.frame(id<-c(7,8,9), name<- c(rep("Nitin",2), rep("Paplu",1))

*outer join
merge(x=df1, y=df2, by = "id", all = TRUE)
This all = TRUE will give you the outer join, so the new data set will have all the value from both the data frame merged on the id

147. How to ideally use the read.csv() function?

Ans.) You must be wondering that it's very easy to use a csv file by putting the name inside the read.csv() function. But, in most of the cases we also need to put some extra conditions in order to get things right and less frustrating for us.
Use the below syntax for the use.

my_data<- read.csv("filename.csv", stringsAsfactors = FALSE, strip.white=TRUE, na.strings=c("NA",""))

stringsAsFactors = FALSE tells R to keep character variables as they are rather than convert to factors.
strip.white = TRUE removes spaces at the start and end of character elements. R treats "game" and " game" differently, which is not usually desired.
na.strings = c("NA","") tells R that in addition to the usual NA, empty strings in columns of character data are also to be treated as missing

148. What is lapply() function in R?

Ans.)lapply() function is used *when you want to apply a function to each element of a list in turn and get a list back.*
Example.

```
x<- list(a=1, b=1:3, c=10:100)
lapply(x,FUN=length)
$a
[1] 1
$b
[1] 3
$c
[1] 91
```

You can use other functions like max, min, sum, etc.

149. What is sapply() function in R?

Ans.)sapply() function is used *when you want to apply a function to each element of a list in turn, but you want a* ***vector*** *back, rather than a list. Vector* is useful sometimes because it will get you a set of values and you can easily perform an operation on it.

Example.

```
x <-list(a =1, b =1:3, c =10:100)
#Compare with above; a named vector, not a list
sapply(x, FUN = length)
a  b  c
1391

sapply(x, FUN = sum)
a    b    c
165005
```


150. How to make scatterplot in R?

Ans.) Scatterplot is a graph which shows many points plotted in the Cartesian plane. Each point holds 2 values which are present on the x and y axis. The simple scatterplot is plotted using plot() function.

The syntax for scatterplot is:-

```
plot(x, y ,main, xlab, ylab, xlim, ylim, axes)
```

Where

x is the data set whose values are the horizontal coordinates
y is the data set whose values are the vertical coordinates
main is the tile in the graph
xlab and ylab is the label in the horizontal and vertical axis
xlim and ylim are the limits of values of x and y used in the plotting
axes indicates whether both axis should be there on the plot

```
plot(x =input$wt,y=input$mpg,
xlab="Weight",
```

```
ylab="Milage",
xlim= c(2.5,5),
ylim= c(15,30),
main="Weight vsMilage"
)
```

151. How to write a countdown function in R?

Ans.)

```
timer<- function(time)
{
print(time)
while(time!=0)
  {
Sys.sleep(1)
time<- time - 1
print(time)
  }
}
countdown(5)
```

[1] 5
[2] 4
[3] 3
[4] 2
[5] 1

152. Is array a matrix or matrix an array?

Ans.) Every matrix can be an array but every array need not be a matrix. A matrix cannot have more than 2 dimensions, whereas an array can be multi dimensional.

Python

153. WAP to show the use of if-elif-else

```
if salary>20000:
        print("Good Salary")
elif salary<20000
        print("Average Salary")
else
        print("Salary is 20000")
```

154. WAP to create a dictionary and then iterate over it and print

```
lucky_number = {'Amit':4,'Rahul':6,'Nihar':8}
for name,number in lucky_number.items():
        print(name+'prefers'+str(number))
```

155. WAP to access all the keys in the dictionary.

```
lucky_number = {'Amit':4,'Rahul':6,'Nihar':8}
for name in lucky_number.keys():
        print(name)
```

156. How to read a file and store the lines in your variable.

```
filename = 'abc.txt'
with open(filename) as file_object:
        lines = file_object.readlines()
for line in lines:
        print(line)
```

157. Exceptions helps us to be prepared for an error which might occur in the program. WAP to showcase how an exception works.

```
x = “What is your age?”
inp = input(x)

try:
        inp = int(inp)
except ValueError:
        print(“Sorry, Please Try again latter”)
else:
        print(“That’s a beautiful age “)
```

158. Try the following operations using List

TDM = [‘The’,’Data’,’Monk]

a. Print the last object of the list
print(TDM[-1])

b. Change the last element to Monkey
TDM[-1] = ‘Monkey

c. Remove Monkey from the list
del TDM[-1]

d. GET Monk back to the list
TDM.append(‘Monk’)

159. Print all the Prime numbers less than 20

```
i = 2
while(i < 20):
   j = 2
   while(j <= (i/j)):
      if not(i%j):
        break
      j = j + 1
   if (j > i/j) :
    print (i," is a prime number")
   i = i + 1
```

160. Write a function to print the square of all numbers from 0 to 11

```
sq = [x**2 for x in range(10)]
print(sq)
```

```
In [60]: sq = [x**2 for x in range(10)]
         print(sq)

         [0, 1, 4, 9, 16, 25, 36, 49, 64, 81]
```

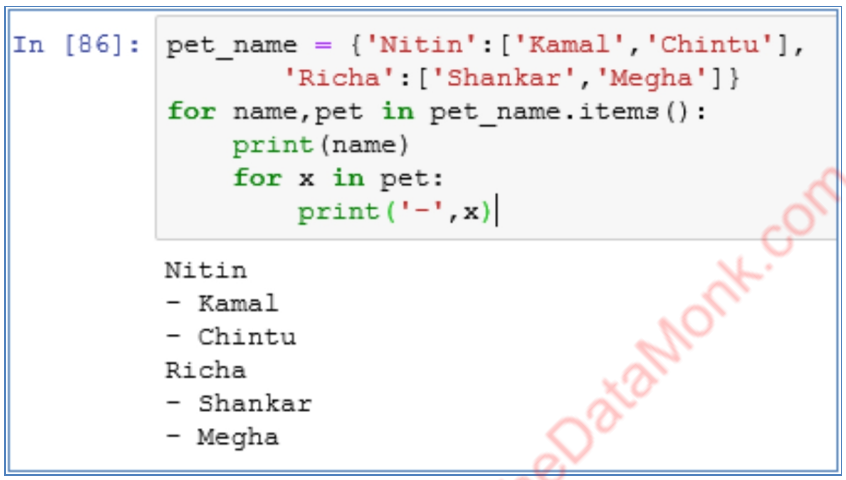
161. How does a mutable list works?

```
list_example = ['Amit','Sumit','Rahul']
print(list_example)
list_example[1] = 'Kamal'
print(list_example)
```

```
['Amit', 'Sumit', 'Rahul']
['Amit', 'Kamal', 'Rahul']
```

162. Write a code to put a list into dictionary

```
pet_name = {'Nitin':['Kamal','Chintu'],
      'Richa':['Shankar','Megha']}
for name,pet in pet_name.items():
   print(name)
   for x in pet:
      print('-',x)
```



163. How to pass a list to a function?

```
def game_name(name):
for x in game_name
      print(x)

example = [‘Cricket’,’Football’,’TT’]
game_name(example)
```

164.When you don't know how many arguments will be passed to a function, then you need to pass a variable number of arguments. Show by an example.

```
def pizza(size, *toppings):
  print("\nMaking a " + size + " pizza.")
  print("Toppings:")
  for topping in toppings:
    print("- " + topping)

# Make three pizzas with different toppings.

make_pizza('small', 'pepperoni')
make_pizza('large', 'bacon bits', 'pineapple')
make_pizza('medium', 'mushrooms', 'peppers', 'onions', 'extra cheese')
```


165. How to split a dataset into train and test in python?

Splitting a dataset into train and test is one of the initial stage of most of the machine learning models. Following is how you can split dataset in python:-

```
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test =
train_test_split(dataframe_name,target_variable,test_size=0.3,
random_state=42)
```

dataframe_name = the complete dataset as a panda dataframe
target_variable = the name of the target variable
test_size = 0.3 denotes 70-30 split of the dataset in train and test
random_state = 42, Look for the explanation in the next question

166.What will be the output of the print(str*3) if str=”TheDataMonk”?
It will print TheDataMonk three times

```
str='TheDataMonk'
print(str*3)
```

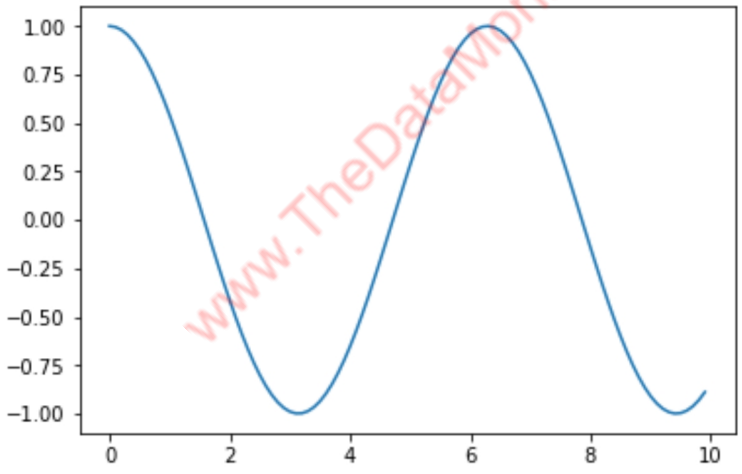
TheDataMonkTheDataMonkTheDataMonk

167.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()

```
In [145]: x = [x*0.1 for x in range(100)]
          y = cos(x)
          plt.plot(x,y)
          plt.show()
```

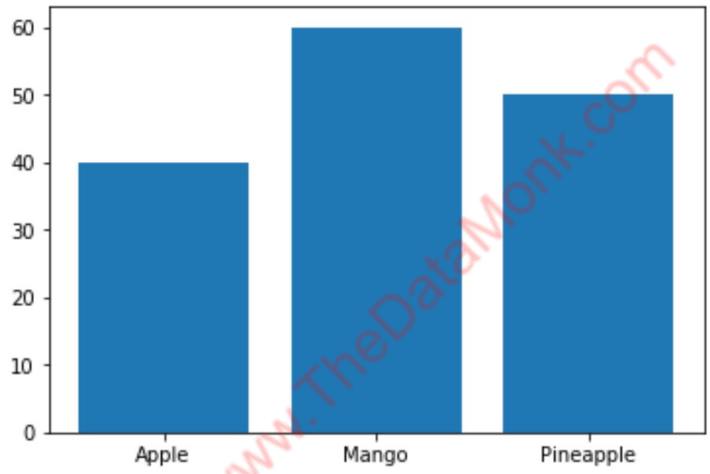


168. Plot a bar chart

import matplotlib.pyplot as plt
a = ['Apple','Mango','Pineapple']
b = [40,60,50]
plt.bar(a,b)

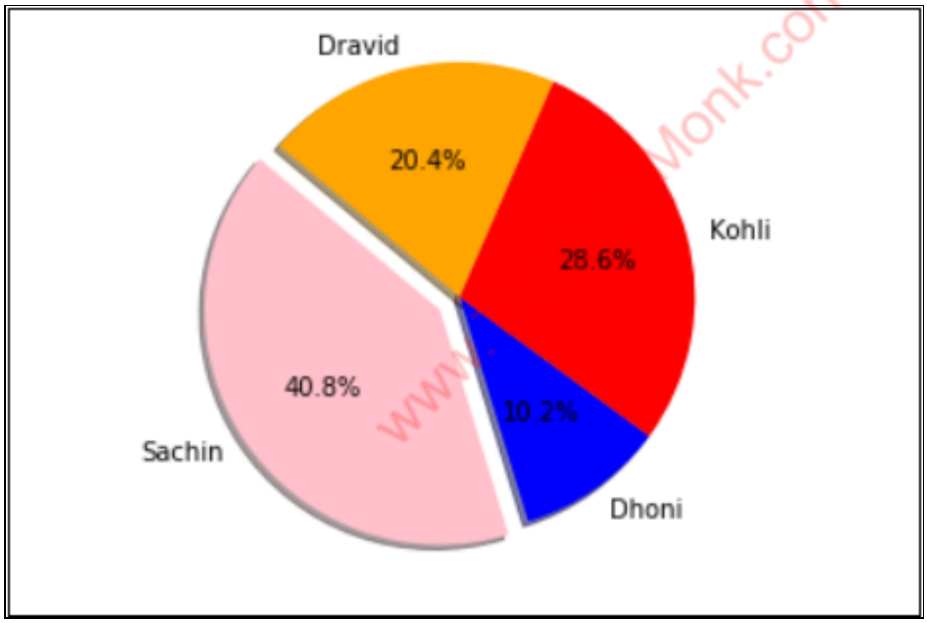
```
In [172]: a = ['Apple','Mango','Pineapple']
          b = [40,60,50]
          plt.bar(a,b)
```

```
Out[172]: <BarContainer object of 3 artists>
```



169. 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='%1.1f%
%',shadow=True,startangle=140)
plt.axis('equal')
plt.show()
```



Project discussion - 1
I had this project on forecasting number of tickets for our client. So, the questions were mostly with respect to predictive modeling.

170. What was the project?
- The project was to predict the number of tickers which will be coming in the future months

171. What language did you use?
- R for creating the model and Microsoft PowerBI for visualization

172.What algorithms did you use?
- Linear Regression, ARIMA, and ARIMAX

173. What are the three components of time series data?
- Trend, Seasonality, and Cyclicity

174. What is cyclicity?
- When a particular behavior is repeated time and again. Generally we talk about a span of 3 to 10-12 years.

175. Give an example of cyclicity?
- The recession which happened in the past, and it is supposed to happen in the coming few years based on cyclicity

176. What is the full form of ARIMA?
- ARIMA stands for Auto-Regressive Integrated Moving Average.

177. What is auto regression?
Auto regression 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. It is a very simple idea that can result in accurate forecasts on a range of time series problems

178. What was the impact of the project?
- The project was intended to reduce the number of employee who are solving tickets. The prediction will help them in estimating the number of employees needed to solve the tickets in the upcoming quarter

179. What is the advantage of ARIMAX?

- Basically ARIMAX gives you an added advantage over Linear Regression and AIRMA model. ARIMAX merges both the concepts i.e. it takes the time series forecasting from ARIMA and then uses other variables to check if these variables have any added advantage over the already predicted value of ARIMA. So, we can safely assume that ARIMAX is the sum of ARIMA and Linear Regression.

180. What were the problems that you faced in the project?

- There were different levels at which we were predicting the volume of tickets. For the top most level we were using Linear Regression and for the lower levels we were using ARIMA and ARIMAX, so the main pain point was the Normalization of number because the numbers predicted by Linear Regression need not match the summation of all the sub level prediction from ARIMA and ARIMAX. So, we normalized the ARIMA and ARIMAX prediction with the Linear Regression.

181. What could have made the result better?
- We could have done more feature engineering to include in the Linear Regression. So, there is a scope to better the already 96% accuracy

There were few more questions on the basics of Statistics, you can easily find the answers to these questions on internet, the reason behind not answering these 4 questions is that Amazon does not allow us to share knowledge which is easily available on the internet:-
182. What is normal distribution?
183. How to calculate correlation?
184. Formula of Variance and Standard Deviation.
185. What is chi-square test?

Miscellaneous Questions

In the above sections we dealt with all the possible domain from which an interviewer can ask you questions. The Miscellaneous Question section will have mixed questions so that you can switch between sections in quick time. Solve or learn each concept

186. How would you construct a feed to show relevant content for a site that involves user interactions with items?

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).

187. 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.)
-Any 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

188. 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 a travel map
-Based on the number of past pickups
-Account for periodicity (seasonal, monthly, weekly, daily, hourly)
-Special events (concerts, festivals, etc.) from tweets

189.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.

190. Give an example of Normal Distribution from daily life.
Height of all the employees on this floor or in this office

191. How do you think TVF makes a profit? Did moving to it's own website advantageous to TVF?

Approach

Honestly, I did not expect such a topic in a case study. I took some 4-5 minutes to shape my idea. Following are the points on which we discussed:-
1. TVF has some 10Million subscriber on Youtube, and it release it's video on Youtube after a week of it's original release on the TVF website. These videos give it a good amount of money to keep the show running
2. The main reason for TVF to move to it's own website was to create an ecosystem comparable to Netflix so that people buy subscription to watch the show.
3. Netflix charges some $9 for subscription, TVF could be planning to launch it's series exclusively to any of these and can get some part of the subscription. Even a dollar per person can get them close to 10Million dollars
4. The estimated revenue of a Youtube channel with 10 Million subscriber is ~500,000 dollars per year.
5. Apart from these, a major chunk of the production cost is taken care by the sponsor of the show. For example Tiago in Trippling, Kingfisher in Pitchers, etc. So the production cost is next to zero for the episodes
6. TVF is also going for it's own website and raising funding to acquire customers and drive them to their website

It's hard to get a $10 subscription, but even a basic subscription or tie-up with some other production can get them a handful of money.

192. The mean of a distribution is 20 and the standard deviation is 5. What is the value of the coefficient of variation?

Variation
= (Standard Deviation/Mean)*100
= (5/20)*100
= 25%

193. When the mean is less than mode and median, then what type of distribution is it?

Negatively Skewed

194. Which of the following describe the middle part of a group of numbers?

a. Measure of Variability
b. Measure of Central Tendency
c. Measure of Association
d. Measure of Shape

Measure of Central Tendency

195. According to the empirical rule, approximately what percent of the data should lie within μ±2σ?

95% of the data should lie between μ±2σ

196. The sum of the deviations about the mean is always:
a. Range
b. Zero
c. Total Deviation
d. Positive

Zero

197. The middle value of an ordered array of numbers is the
a. Mode
b. Mean
c. Median
d. Standard Deviation

Median

198. Height of employees is a :-
a. Continuous value
b. Qualitative value
c. Discrete value
d. None of these

Continuous Value

199. Which of these is a measure of dispersion:-
a. Mean
b. Median
c. Quartile
d. Standard Deviation

Standard deviation is a measure of dispersion

200. Variance of a dataset is 144, what is the Standard Deviation?

Standard deviation is square root of Variance, so the Standard deviation will be 12

201. Which of these is a qualitative data:-
a. Weight of family members
b. Salary
c. Feedback of 100 customers about your website
d. Number of burgers sold in India

Feedback of 100 customers about your website, rest all are discrete

202. Which of these is/are measure of central tendency?
a. Median
b. Mean
c. Mode
d. Mid range
e. Mid hinge

All of these are measures of central tendency

203. What divides a data set in a group of 10 parts?
a. Deciles
b. Percentile
c. Quartile
d. Standard Deviation

Deciles divide the complete dataset in a group of 10 parts

204. What is Mid range?
Arithmetic mean of the maximum and minimum values of a dataset is called mid-range

205. What is Mid hinge?
Arithmetic mean of the two quartiles is called mid hinge

206. What is Inter Quartile Range?
a. 0-50th percentile
b. 25-50th percentile
c. 25-75th percentile
d. 50-100th percentile

25-75th percentile is called IQR i.e. Inter Quartile Range

207. What is a cap in a box-plot?
A upper cap contains the values which falls between 75th percentile and 75th Percentile+1.5*IQR. Similarly lower cap contains the values which falls between 25th Percentile and 25th Percentile-1.5*IQR

208. What values are termed as an outlier in a box plot?
Any value which is more than upper cap and less than lower cap will fall under the definition of outlier

209. Which of the following divides the group of data into four subgroups?
a. Median
b. Mean
c. Quartile
d. Percentile

Quartile divides a group of dataset into four subgroups

210. A chance variation in an observational process is
a. Dispersion
b. Measurement error
c. Random error
d. Instrument error

Random error

211. Given X1=6,X2=5,X3=4,X4=3, then ∑4i=1 Xi equals?
18

212. Consider the following data set and create a linear regression model in Python
X = 3,4,5,6,7
Y = 15,22,25,33,40

```
from sklearn import linear_model
import numpy as np
xl = [3,4,5,6,7]
x = np.asarray(xl).reshape(-1, 1)
y = [15,22,25,33,40]
lm = linear_model.LinearRegression()
lm.fit(x, y)
print(lm.intercept_)
print(lm.coef_[0])
```


213. A train 120 meters long is running with a speed of 60 km/hr. In what time will it pass a boy who is running at 6 km/hr in the direction opposite to that in which the train is going?

Speed of train relative to boy= `(60 + 6)` km/hr = 66 km/hr
`= [66 xx 5/18]` m/sec = `[55/3]` m/sec.
Time taken to pass the boy=` [120 xx 3/55]` sec = 6.54 seconds

214. What is the value of c , If 8 is 4% of a, and 4 is 8% of b. c equals b/a?

Let be the 4% of a is 4a/100.
Since this equals 8, we have 4a/100=8.
Solving for a yields a=8×`(100/4)`=200.
Also, 8% of b equals 8b/100, and this equals 4.
Hence, we have `(8/100)`×b=4.
Solving for b yields b = 50.
Now, c=`b/a`=`50/200`=`1/4.`

215. P and Q take part in 100 m race. P runs at 6kmph. P gives Q a start of 8 m and still beats him by 8 seconds. The speed of Q is:
`speed =( 6**5/18)m/sec) = (30/18) m/sec`
Time taken by P to cover 100 m = `(100 ** 18/30)`m/sec =60 sec
Time taken by Q to cover 92 m = (60 + 8) = 68 sec.
Q's speed =`("Distance"/"Time"**18/5) kmph =(92/68 ** 18/5)` kmph = 4.86kmph.

216. Find the speed of the train, if a train 142 m long passes a pole in 6 seconds.
Speed = `[142/6]` m/sec
= `[23.6 ** 18/5]` km/hr
= 84.9 km/hr

217. Sum of present ages of A, B and C is 92 years. If 4 years ago, the ratio of their ages were 1:2:3 respectively, find A's present age.
Sum of present ages of A, B and C is = 92 years
Therefore , Sum of their ages 4 years ago = 92 – (4 * 3)= 80 years.
4 years ago ratio of the ages of A , B and C was = 1:2:3
Therefore, A's age four years ago = `1/6` * 80 = 13.3 years.
So , A's present age =13.3 + 4 = 17.3 years

218. A guy bought 10 pencils for Rs. 50 and sold them for Rs. 60.What is his gain in terms of percentage?
Gain%"=("Gain"/"C.P")*100=20%

219. Common Data Quality Issues
Missing Values
Noise in the Data Set
Outliers
Mixture of Different Languages (like English and Chinese)
Range Constraints

220. What is Imbalanced Data Set and how to handle them? Name Few Examples.
Fraud detection
Disease screening
Imbalanced Data Set means that the population of one class is extremely large than the other
(Eg: Fraud – 99% and Non-Fraud – 1%)
Imbalanced dataset can be handled by either oversampling, under sampling and penalized Machine Learning Algorithm.

221.If you are dealing with 10M Data, then will you go for Machine learning (or) Deep learning Algorithm?
Machine learning algorithm suits well for small data and it might take huge amount of time to train for large data.
Whereas Deep learning algorithm takes less amount of data to train due to the help of GPU(Parallel Processing).

222. Examples of Supervised learning algorithm?
Linear Regression and Logistic Regression
Decision Trees and Random Forest
SVM
Naïve Bayes
XGBoost

223. In Logistic Regression, if you want to know the best features in your dataset then what you would do?
Apply step function, which calculates the AIC for different permutation and combination of features and provides the best features for the dataset.

224. What is Feature Engineering? Explain with Example?
Feature engineering is the process of using domain knowledge of the data to create features for machine learning algorithm to work
-Adding more columns (or) removing columns from the existing column
-Outlier Detection
-Normalization etc

225. How to select the important features in the given data set?
In Logistic Regression, we can use step() which gives AIC score of set of features
In Decision Tree, We can use information gain(which internally uses entropy)
In Random Forest, We can use varImpPlot

226. When does multicollinearity problem occur and how to handle it?
It exists when 2 or more predictors are highly correlated with each other.

Example: In the Data Set if you have grades of 2nd PUC and marks of 2nd PUC, Then both gives the same trend to capture, which might internally hamper the speed and time.so we need to check if the multi collinearity exists by using VIF(variance Inflation Factor).

Note: if the Variance Inflation Factor is more than 4, then multi collinearity problem exists.

227. What is Variance inflation Factors (VIF)?
Measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

228. Examples of Parametric machine learning algorithm and non-parametric machine learning algorithm
Parametric machine learning algorithm– Linear Regression, Logistic Regression
Non-Parametric machine learning algorithm – Decision Trees, SVM, Neural Network

229. What are parametric and non-parametric machine learning algorithm? And their importance
Algorithm which does not make strong assumptions are non-parametric algorithm and they are free to learn from training data. Algorithm that makes strong assumptions are parametric and it involves

select the form for the function and
learn the coefficients for the function from training data.

230. When does linear and logistic regression performs better, generally?
It works better when we remove the attributes which are unrelated to the output variable and highly co-related variable to each other.

231. Why you call naïve bayes as "naïve" ?
Reason: It assumes that the input variable is independent, but in real world it is unrealistic, since all the features would be dependent on each other.

232. Give some example for false positive, false negative, true positive, true negative
False Positive – A cancer screening test comes back positive, but you don't have cancer
False Negative – A cancer screening test comes back negative, but you have cancer
True Positive – A Cancer Screening test comes back positive, and you have cancer
True Negative – A Cancer Screening test comes back negative, and you don't have cancer

233. What is Sensitivity and Specificity?
Sensitivity means "proportion of actual positives that are correctly classified" in other words "True Positive"

Specificity means "proportion of actual negatives that are correctly classified" "True Negative"

234. When to use Logistic Regression and when to use Linear Regression?
If you are dealing with a classification problem like (Yes/No, Fraud/Non Fraud, Sports/Music/Dance) then use Logistic Regression.

If you are dealing with continuous/discrete values, then go for Linear Regression.

235. What are the different imputation algorithm available?
Imputation algorithm means "replacing the Blank values by some values)

Mean imputation

Median Imputation
MICE
miss forest
Amelia

236. What is AIC(Akaike Information Criteria)
The analogous metric of adjusted R^2 in logistic regression is AIC.

AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.

237. Suppose you have 10 samples, where 8 are positive and 2 are negative, how to calculate Entropy (important to know)
E(S) = 8/10log(8/10) – 2/10log(2/10)

Note: Log is à base 2

238. What is perceptron in Machine Leaning?
In Machine Learning. Perceptron is an algorithm for supervised classification of the input into one of several possible non-binary outputs

239. How to ensure we are not over fitting the model?
Keep the attributes/Columns which are really important
Use K-Fold cross validation techniques
Make use of drop-put incase of neural network
How the root node is predicted in Decision Tree Algorithm?
Mathematical Formula “Entropy” is utilized for predicting the root node of the tree.

240. What are the different Backend Process available in Keras?
TensorFlow
Theano
CNTK
Q126. Name Few Deep Learning Algorithm
TensorFlow
Theano

Lasagne
mxnet
blocks
Keras
CNTK
TFLearn

241. How to split the data with equal set of classes in both training and testing data?
Using Stratified Shuffle package

242. What do you mean by Ensemble Model? When to use?
Ensemble Model is a combination of Different Models to predict correctly and with good accuracy.

Ensemble learning is used when you build component classifiers that are more accurate and independent from each other.

243. When will you use SVM and when to use Random Forest?
SVM can be used if the data is outlier free whereas Naïve Bayes can be used even if it has outliers (since it has built in package to take care).
SVM suits best for Text Classification Model and Random Forest suits for Binomial/Multinomial Classification Problem.
Random Forest takes care of over fitting problem with the help of tree pruning

244. Applications of Machine Learning?
Self Driving Cars
Image Classification
Text Classification
Search Engine
Banking, Healthcare Domain

245. If you are given with a use case – 'Predict whether the transaction is fraud (or) not fraud", which algorithm would you choose
Logistic Regression

246. If you are given with a use case – 'Predict the house price range in the coming years", which algorithm would you choose
Linear Regression

247. What is the output of 3*32?**

27
The order of precedence is ** then *. Thus 3**2 = 9 and then 9*3 = 27.

248. How to convert a tuple into a list?

```
tup = ('the','Data','Monk')
list_example = list(tup)
print(list_example)
```

['the','Data','Monk']

249. Though you are comfortable with string, but do try to answer the output of the following basic operations on string.

```
str="TheDataMonk"
print (str)
print (str*2)
print (str[2:5])
print (str[3:])
print (str + ".com")
print ("www."+str+".com")
```

```
TheDataMonk
TheDataMonkTheDataMonk
eDa
DataMonk
TheDataMonk.com
www.TheDataMonk.com
```

250. Calculate IQR

Step 1: Put the numbers in order.
1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
Step 2: Find the median.
1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
Step 3: Place parentheses around the numbers above and below the median. Not necessary statistically, but it makes Q1 and Q3 easier to spot.
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).
Step 4: Find Q1 and Q3
Think of Q1 as a median in the lower half of the data and think of Q3 as a median for the upper half of data.
(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27). Q1 = 5 and Q3 = 18.
Step 5: Subtract Q1 from Q3 to find the interquartile range.
18 – 5 = 13.

251. Calculate Standard Deviation
You survey households in your area to find the average rent they are paying. Find the standard deviation from the following data:
$1550, $1700, $900, $850, $1000, $950.

Step 1: Find the mean:
($1550 + $1700 + $900 + $850 + $1000 + $950)/6 = $1158.33

Step 2: Subtract the mean from each value. This gives you the differences:
$1550 – $1158.33 = $391.67
$1700 – $1158.33 = $541.67
$900 – $1158.33 = -$258.33
$850 – $1158.33 = -$308.33
$1000 – $1158.33 = $158.33
$950 – $1158.33 = $208.33

Step 3: Square the differences you found in Step 3:
$391.672 = 153405.3889
$541.672 = 293406.3889
-$258.332 = 66734.3889
-$308.332 = 95067.3889
$158.332 = 25068.3889
$208.332 = 43401.3889

Step 4: Add up all of the squares you found in Step 3 and divide by 5 (which is 6 – 1):
(153405.3889 + 293406.3889 + 66734.3889 + 95067.3889 + 25068.3889 + 43401.3889) / 5 = 135416.66668

Step 5: Find the square root of the number you found in Step 4 (the variance):
√135416.66668 = 367.99

The standard deviation is 367.99.

252. Which of the following Measures of averages are affected by the outliers in the data set?
a. Median
b. Mode
c. Mean
d. Geometric Mean

A. Mean is affected badly by the outliers. It's said that if a Billionaire walks into a cheap bar, the average crowd becomes millionaire

253. Which of the following is not possible to compute for the following data set?
Data set – 23,43,223,54,64,0,1,2
a. Median
b. Mean
c. Mode
d. Standard Deviation
e. Geometric Mean
f. Harmonic Mean

A. Whenever there is a data point 0, then you cannot compute the Harmonic mean

254. Can a geometric mean be negative?
Like zero, it is impossible to calculate Geometric Mean with negative numbers. However, there are several work-around for this problem, all of which require that the negative values be converted or transformed to a meaningful positive equivalent value.

255. 2nd Quartile = 5th Decile = 50th Percentile = ?
a. Harmonic Mean
b. Mode
c. Median
d. Mean

A. Median is the 50^{th} percentile which is equal to 5^{th} decile, each decile denotes 10 percentile

256. A distribution having two mode is called a :
a. Tri-modal
b. Uni-modal
c. Multi-modal
d. Bi-modal

A. Bi-modal

257. Numerical methods and graphical methods are specialized procedures used in
a. social statistics
b. business statistics
c. descriptive statistics
d. education statistics

A. Descriptive Statistics

258. Which measure of average can have more than one value?
a. Mode
b. Mean
c. Median
d. Harmonic Mean

A. A series can have more than one values for mode. Example – 2,3,4,5,5,5,4,4,6,7,7 Here both 5 and 4 are mode

259. The measure of Dispersion can never be?
a. Positive
b. Negative
c. Zero
d. One

A. The measure of dispersion can never be negative. It is never negative since every term in the variance sum is squared and therefore either positive or zero. It has squared units.

260. The expected value or _____ of a random variable is the center of its distribution.
a) mode
b) median
c) mean
d) bayesian inference

A. Mean is generally taken as the center of distribution. So, the answer is c

261. Point out the wrong statement:
a) A percentile is simply a quantile with expressed as a percent
b) There are two types of random variable
c) R cannot approximate quantiles for you for common distributions
d) None of the Mentioned

A. Discrete and continuous are the two types of random variable. A percentile is definitely a quantile. R can approximate quantiles for common distribution

262. The sum of the percent frequencies for all classes will always equal

A. one
B. the number of classes
C. the number of items in the study
D. 100

Ans. The sum of the percent frequencies will always be 100

263. What is the use of correlation?
Correlation can tell you something about the relationship between variables. It is used to understand:
1. Whether the relationship is positive or negative

2. The strength of the relationship.

264. What is p-value and give an example?
In statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If the p-value is less than 0.05 or 0.01, corresponding respectively to a 5% or 1% chance of rejecting the null hypothesis when it is true (Type I error).

265. Display the depart numbers with more than three employees in each dept.

```
SELECT deptno, count(deptno)
FROM emp
GROUP BY deptno
HAVING count(*)>3;
```

266. In a single toss of 2 fair (evenly-weighted) six-sided dice, find the probability that their sum will be at most .
When you roll two dice, you have 6 possibilities for each roll (6 sides). This is 36 total combinations.
Let's list the combinations that result in sums greater than 9.
(4,6) (6,4) (5,5) (6,5) (5,6) (6,6)
That's 6 out of the 36 total possibilities. Therefore, the remaining 30/36 possibilities fulfill the less than or equal to 9 requirement. Simplifying by a factor of 6, that's 5/6 chance.

267. Let's say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?

Shorter. Regression to the mean

268. Let's say you're building the recommended music engine at Spotify to recommend peoplemusic based on past listening history. How would you approach this problem?
Collaborative Filtering

269. What is R2? What are some other metrics that could be better than R2 and why?
Goodness of fit measure. Variance explained by the regression / total variance
Remember, the more predictors you add the higher R^2 becomes. Hence use adjusted R^2 which adjusts for the degrees of freedom or train error metrics

271. Is more data always better?
Statistically, It depends on the quality of your data, for example, if your data is biased, just getting more data won't help. It depends on your model. If your model suffers from high bias, getting more data won't improve your test results beyond a point. You'd need to add more features, etc.
Practically, Also there's a tradeoff between having more data and the additional storage, computational power, memory it requires. Hence, always think about the cost of having more data.

272. You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?
Multi collinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.
Leave the model as is, despite multi collinearity. The presence of multi collinearity doesn't affect the efficiency of extrapolating the fitted model to new data provided that the predictor variables follow the same pattern of multi collinearity in the new data as in the data on which the regression model is based.
principal component regression

273. What are the tests which are performed on data sets?
There are many tests, few are:-
-A/B Test
-Student's T Test
-Chi Square Test
-Fisher's Exact Test
-Mann-Whitney Test

Human Resource Round

Though these questions are asked only at the end of the interview and most probably you have already made it to the final list. But, whenever you have time, give these questions a shot.

1. Why are you interested in working for our organization/company?
Be True in this question. Do ample research before the interview.

2. Where do you see yourself in 5 years?
Keep it concrete, like a higher management position in this company.

3. Are you willing to relocate?
Should be Yes

4. Are you willing to travel?
Yes only if you willing to travel

5. Why should we hire you?
Quick Learner, Ability to lead a project, etc.

6. How do you handle pressure?
Give a few examples of your ability to handle pressure situation.

7. What are your career goals?
No MBA or higher studies. Keep the goals align to the present position which you are applying for.

8. How are you planning to achieve them?
Support the answer with the learning which you are thinking to achieve.

9. What was the last book you've read for entertainment?
If you are not a reader, just say it directly. Else be true with the book's name.

10. What do you like to do in your spare time?
Anything will do, ranging from solving Sudoku to Table tennis or writing books.

11. What are your strengths?
Be true with your strength, this is the time to show your leadership quality or a tech guy or a communicator.

12. What are your weaknesses?
A big no to the below answers:
-My honesty is my weakness
-My hard working nature is my weakness
Prepare something where you are weak but make sure you present it in such a way that it looks like you want to get better on it

13. What do you know about our organization/company?
The company knowledge will guide you here

14. Why do you want to leave your current job?
Be true but don't be blunt. Don't bash the manager or colleagues.

15. What are your salary expectations?
Be blunt with the expectation but make sure you check glassdoor to know how much they offer to the same position

16. Tell me about a difficult situation you have faced, and how you dealt with it.
Nail it

17. What relevant experience do you have?
Make sure that you only refer the “relevant experience”

18. What motivates you to do a good job?
Success is what motivates me to do a good job. Knowing the fact that my hard work and perseverance will help me achieve greater professional success is what keeps me going. I feel that aligning the company's vision and values with my own is one way to achieve that

19. Is there anyone you just could not work with?
People who procrastinate their work

20. What makes you think you are qualified for this position?
I have thoroughly gone through the job description and the profile of the company. It seems to be aligned with my expertise and interest, so I will be able to work in this role with efficiency

21. Tell me about a project you initiated?
Take any example and explain it. Focus on the impact it created for the team/project

22. What interests you about this job?
Same as Q.20

23. What can you contribute to this company?

24. With which other organizations/companies are you interviewing?
Be true if you think your interview has been decent.

25. What characteristics are most important in a good manager?
Ability to take decisions in tricky situations(quote an example here), ability to hold the team as one, etc.

26. What challenges are you looking for in a position?
New tools, understanding the project from scratch, etc.

27. What are your core beliefs and values?

28. What gets you up in the morning?
If it's a week day then my work motivates me because if I am able to do it efficiently then I will have a lot of time for myself and will award my body with ample sleep. If it's a weekend then daily shores motivates me because again if I am able to do it efficiently then I will have a lot of time for myself and will award my body with ample sleep :P

29. What questions do you have for me?
Ask questions related to the team you will going to, the tools they use so that you can learn it in the coming days, you can basically ask anything.

The point of creating this book is to make you aware of the pattern and difficulty level of the interviews. I would highly recommend you to go through the Interview Questions book(link and name of the book given above).

Mail us at contact@thedatamonk.com for any help.

Hopefully you will crack your next Data Science interview