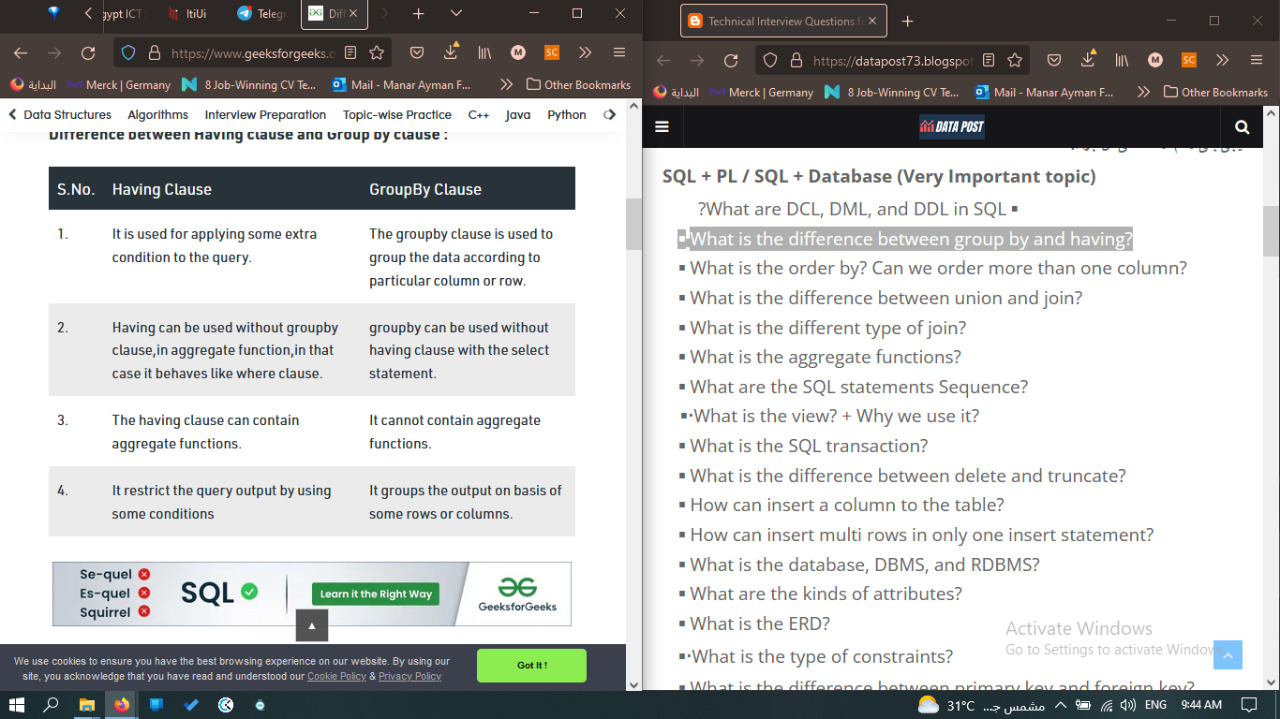
# SQL + PL / SQL + Database (Very Important topic)

1. What are DCL, DML, and DDL in SQL?

* ال DCL هى ال Data Controlled language و دى المسؤلة عن عملية ال Control للداتا و اشهر ال Syntax بتاعها هم ال Revoke / Grant
* ال DML دة هى ال Data Manipulation language و دى المسؤولة عن انها بتعمل بقى ال ال Functions على ال Data اللى معايا عشان أقدر اطلع منها insights
* ال DDL و هى ال Data Definition language و دى المسؤولة عن ال Create / Alter / Drop for tables .

1. What is the difference between group by and having?



1. What is the order by? Can we order more than one column?

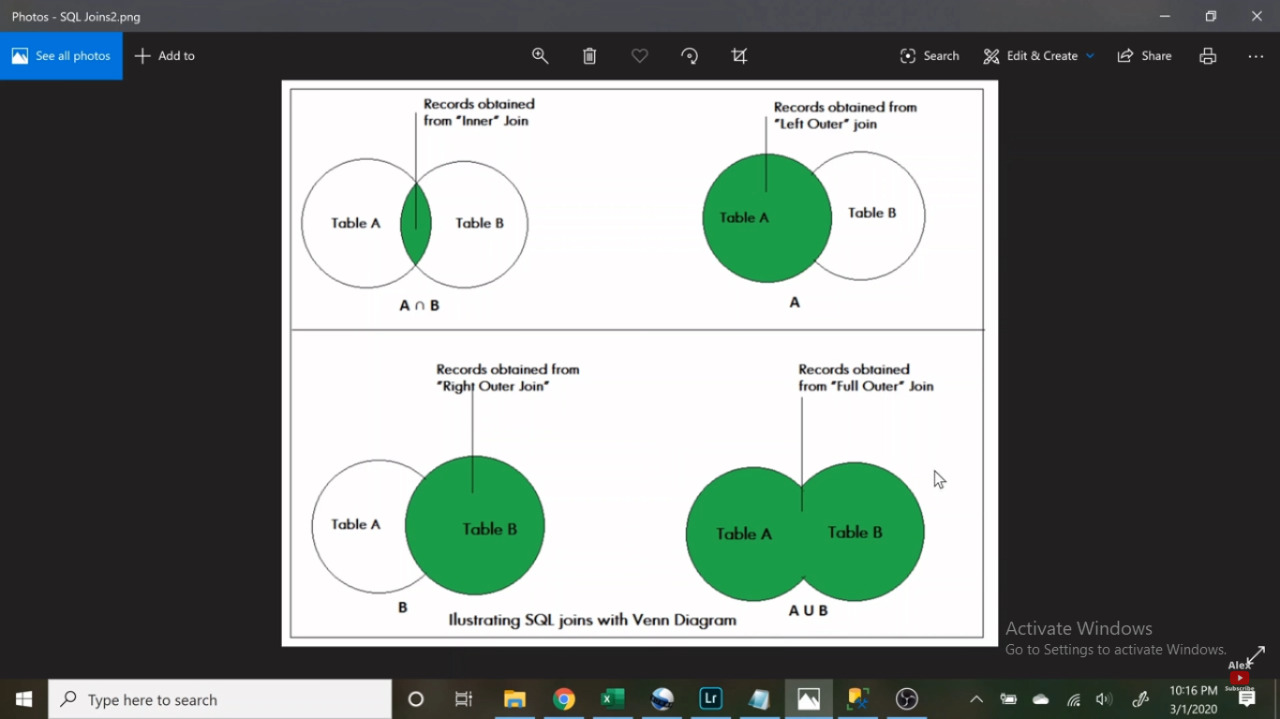
* هى عبارة عن كود انا بكتبه عشان ارتب القيم اللى فى جدول محدد بالنسبة ل Coulmn سواء كان الترتيب دة ترتيب Asc او Desc و عادى ممكن اعمل ترتيب بالنسبة لاكتر من Column و بيبقى عن طريق انى اكتب Order by Gender Asc , Age Desc

1. What is the difference between union and join?

* الاتنين بيجمعو القيم اللى هتظهر من جدولين لكن بيختلفو فى طريق ال merge اللى بيحصل بمعنى ان ال join بتجمع الجدولين horizontal جنب بعض على اساس ال entity و دة بيحصل فى أى جدولين بشرط ان يكون فى على الاقل 1 Column موجود فى الاتنين و على اساسه
* لكن فى ال union ال merge بيحصل على اساس ال Data Type لل ِAttribute و الناتج بيكون Vertically فممكن لو الجدولين ليهم نفس ال Enitiies هيبقى مظبوط لكن لو ال entities مختلفة و ال Data type واحد هيطلع جدول فعلا لكن المعنى مش هيبقى مظبوط

1. What is the different type of join?

Inner outer Join , Full Outer Join , Left / Right Outer Join



1. What is the aggregate functions?

In database management an aggregate function is a function where the values of multiple rows are grouped together as input on certain criteria to form a single value of more significant meaning.

Various Aggregate Functions: Count () /Sum() / Avg() / Min()/ Max()

1. What are the SQL statements Sequence?

What actually sets SQL Server apart from other programming languages is the way SQL Server processes its code. Generally, most programming languages process statement from top to bottom. By contrast, SQL Server processes them in a unique order which is known as Logical Query Processing Phase. These phases generate a series of virtual tables with each virtual table feeding into the next phase (virtual tables not viewable). These phases and their orders are given as follows:

* + FROM
  + ON
  + OUTER
  + WHERE
  + GROUP BY
  + CUBE | ROLLUP
  + HAVING
  + SELECT
  + DISTINCT
  + ORDER BY
  + TOP

فى اجابة تانية بتقول

1-getting data  
2-row filter  
3-grouping  
4-group filter  
5-return experssion  
6-order

1. What is the view? + Why we use it?

Views in SQL are kind of virtual tables. A view also has rows and columns as they are in a real table in the database. We can create a view by selecting fields from one or more tables present in the database. A View can either have all the rows of a table or specific rows based on certain condition.

CREATE VIEW view\_name AS

SELECT column1, column2.....

FROM table\_name

WHERE condition;

**view\_name**: Name for the View

**table\_name**: Name of the table

**condition**: Condition to select rows

1. What is the SQL transaction?

What are Transactions?

Transactions group a set of tasks into a single execution unit. Each transaction begins with a specific task and ends when all the tasks in the group successfully complete. If any of the tasks fail, the transaction fails. Therefore, a transaction has only two results: success or failure.

Incomplete steps result in the failure of the transaction. A database transaction, by definition, must be atomic, consistent, isolated and durable. These are popularly known as ACID properties.

Following commands are used to control transactions. It is important to note that these statements cannot be used while creating tables and are only used with the DML Commands such as – INSERT, UPDATE and DELETE.

1. BEGIN TRANSACTION: It indicates the start point of an explicit or local transaction.

2. SET TRANSACTION: Places a name on a transaction.

3. COMMIT: If everything is in order with all statements within a single transaction, all changes are recorded together in the database is called committed. The COMMIT command saves all the transactions to the database since the last COMMIT or ROLLBACK command.

BEGIN TRANSACTION transaction\_name ;

SET TRANSACTION [ READ WRITE | READ ONLY ];

COMMIT;

1. What is the difference between delete and truncate?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Delet | Drop | Truncate |
| Command | DML | DDL | DDL |
| Can Delete | 1 by 1 or the whole table by using where clause | The whole table + the Structure | The whole table |
| Speed | Slower than truncate |  | Faster than Delete |
| Syntax | Delete From Where; | Drop Table; | Truncate; |
| Rollback | مبيعملش auto commit فعادى نعمل rollback | بيعمل auto commit فمش ممكن نعمل rollback | بيعمل auto commit فمش ممكن نعمل rollback |

1. How can insert a column to the table?

**Insert into Table\_Name Values**

**( ….,……,…..);**

1. How can insert multi rows in only one insert statement?

Insert into Table\_Name Values

( ….,……,…..),

(….,…..,…..);

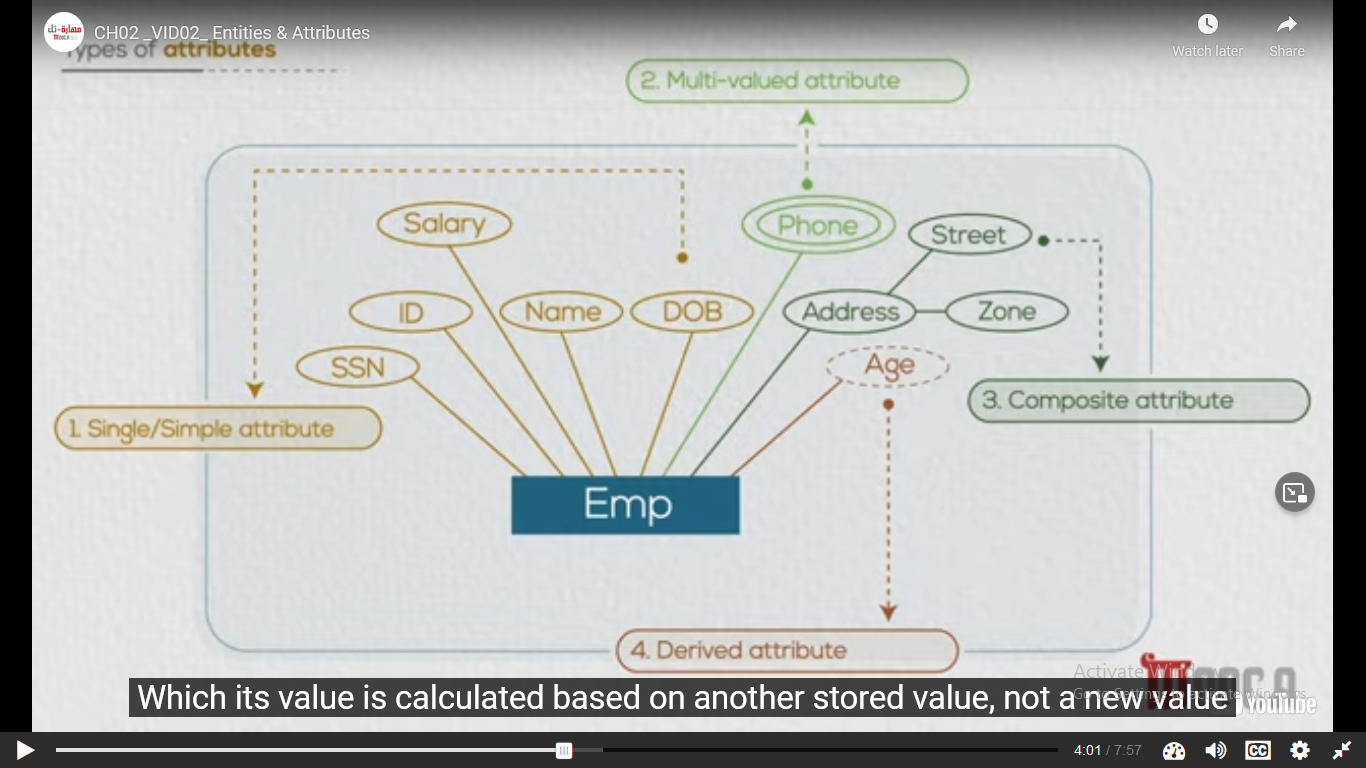
1. What is the database, DBMS, and RDBMS?

**Database Management System (DBMS)** is a software that is used to define, create and maintain a database and provides controlled access to the data.

**Relational Database Management System (RDBMS)** is an advanced version of a DBMS. 

|  |  |
| --- | --- |
| DBMS | RDBMS |
| DBMS stores data as file. | RDBMS stores data in tabular form. |
| Data elements need to access individually. | Multiple data elements can be accessed at the same time. |
| No relationship between data. | Data is stored in the form of tables which are related to each other. |
| Normalization is not present. | Normalization is present. |
| DBMS does not support distributed database. | RDBMS supports distributed database. |
| It stores data in either a navigational or hierarchical form. | It uses a tabular structure where the headers are the column names, and the rows contain corresponding values. |
| It deals with small quantity of data. | It deals with large amount of data. |
| Data redundancy is common in this model. | Keys and indexes do not allow Data redundancy. |
| It is used for small organization and deal with small data. | It is used to handle large amount of data. |
| It supports single user. | It supports multiple users. |
| Data fetching is slower for the large amount of data. | Data fetching is fast because of relational approach. |
| The data in a DBMS is subject to low security levels with regards to data manipulation. | There exists multiple levels of data security in a RDBMS. |
| Low software and hardware necessities. | Higher software and hardware necessities. |
| Examples: XML, Window Registry, etc. | Examples: MySQL, PostgreSQL, SQL Server, Oracle, Microsoft Access etc. |

1. What are the kinds of attributes?



Complex attribute :

Those attributes, which can be formed by the nesting of composite and multi-valued attributes, are called “Complex Attributes“. These attributes are rarely used in DBMS(DataBase Management System). That’s why they are not so popular.

1. What is the ERD?

ER Diagram is known as Entity-Relationship Diagram, it is used to analyze to structure of the Database. It shows relationships between entities and their attributes. An ER Model provides a means of communication.

1. What is the type of constraints?

Constraints are the rules that we can apply on the type of data in a table. That is, we can specify the limit on the type of data that can be stored in a particular column in a table using constraints.

The available constraints in SQL are: 

* **NOT NULL**: This constraint tells that we cannot store a null value in a column. That is, if a column is specified as NOT NULL then we will not be able to store null in this particular column any more.
* **UNIQUE**: This constraint when specified with a column, tells that all the values in the column must be unique. That is, the values in any row of a column must not be repeated.
* **PRIMARY KEY**: A primary key is a field which can uniquely identify each row in a table. And this constraint is used to specify a field in a table as primary key.
* **FOREIGN KEY**: A Foreign key is a field which can uniquely identify each row in a another table. And this constraint is used to specify a field as Foreign key.
* **CHECK**: This constraint helps to validate the values of a column to meet a particular condition. That is, it helps to ensure that the value stored in a column meets a specific condition.
* **DEFAULT**: This constraint specifies a default value for the column when no value is specified by the user.

CREATE TABLE sample\_table

(

column1 data\_type(size) constraint\_name,

column2 data\_type(size) constraint\_name,

column3 data\_type(size) constraint\_name,

....

);

1. What is the difference between primary key and foreign key?

|  |  |  |
| --- | --- | --- |
| S.NO. | PRIMARY KEY | FOREIGN KEY |
| 1 | A primary key is used to ensure data in the specific column is unique. | A foreign key is a column or group of columns in a relational database table that provides a link between data in two tables. |
| 2 | It uniquely identifies a record in the relational database table. | It refers to the field in a table which is the primary key of another table. |
| 3 | Only one primary key is allowed in a table. | Whereas more than one foreign key are allowed in a table. |
| 4 | It is a combination of UNIQUE and Not Null constraints. | It can contain duplicate values and a table in a relational database. |
| 5 | It does not allow NULL values. | It can also contain NULL values. |
| 6 | Its value cannot be deleted from the parent table. | Its value can be deleted from the child table. |
| 7 | It constraint can be implicitly defined on the temporary tables. | It constraint cannot be defined on the local or global temporary tables. |

1. What the difference is between delete and truncate?

Q11

1. What is delete set null and delete cascade?

This deletes the rows from the entire parent and child table as well. There are other permitted actions – no action, cascade, set null, set default.

* **On Delete No Action –**  
  It raises an error and rolls back the delete action on parent table.
* **On Delete Cascade –**  
  The cascade action deletes all the rows from the parent and child table.
* **On Delete Set Null –**  
  The rows from the parent and child table are set null only if the foreign key is nullable.
* **On Delete Set Default –**  
  The child table rows are set to default if the corresponding parent table rows are deleted only if the foreign key has default definitions.

1. What is the normalization and why are we making it?

Database normalization is the process of organizing the attributes of the database to reduce or eliminate **data redundancy (having the same data but at different places)** .

1. What are the types of normalization?

Diagram, text

Description automatically generated

1. What are the update anomalies?

**Insertion anomaly:** If a tuple is inserted in referencing relation and referencing attribute value is not present in referenced attribute, it will not allow inserting in referencing relation. For Example, If we try to insert a record in STUDENT\_COURSE with STUD\_NO =7, it will not allow.

**Deletion and Updation anomaly:** If a tuple is deleted or updated from referenced relation and referenced attribute value is used by referencing attribute in referencing relation, it will not allow deleting the tuple from referenced relation.

حلها اما اعمل و انا بمسح ال delete cascade or delete set null و فى ال inserion ممكن اغير ال design او ال relation بين الجدولين

1. What is the difference between SQL and PL/SQL?

|  |  |
| --- | --- |
| SQL | PL SQL |
| SQL Stands for Structured Query language in which single SQL statement is executed at a time. | PL SQL stands for Procedural Language Structured Query Language which is used as programming language in databases. |
| SQL simply used to perform DDLs (Data Definition Language) and DMLs with using single query at a time. | PL SQL is nothing but the block of code embed with SQL statements which is used to add business logic with using programming concepts. |
| SQL is called as declarative language which is used to define what needs to be done. | PL SQL is called as procedural language which is used to define how the things should be done. |
| SQL executes as a single statement. | PL SQL executes as a block. The code of PL SQL is always written in to the block. |
| The main use of SQL statement is to process and manipulate the data. The SQL statements are mainly used for reporting purpose. | PL SQL statements are used to perform the transactions to the databases. PL SQL is mainly used to build the application logic. |
| SQL directly interacts with database server | There is no interaction between database server. |
| User can add SQL in to PL SQL statements | User can not add PL SQL in to SQL statements |

1. What are the types of loops in PL/SQL?

* Basic Loop

Basic loop or simple loop is preferred in PL/SQL code when there is no surety about how many times the block of code is to be repeated. When we use the basic loop the code block will be executed at least once.

While using it, following two things must be considered:

* Simple loop always begins with the keyword LOOP and ends with a keyword END LOOP.
* A basic/simple loop can be terminated at any given point by using the exit statement or by specifying certain condition by using the statement exit when.

**Syntax:**

LOOP

sequence of statements

END LOOP;

* While Loop

t is an entry controlled loop which means that before entering in a while loop first the condition is tested, if the condition is TRUE the statement or a group of statements get executed and if the condition is FALSE the control will move out of the while loop.

Syntax:

WHILE <test\_condition> LOOP

<action>

END LOOP;

* For Loop

his loop is used when some statements in PL/SQL code block are to be repeated for a fixed number of times.

When we use the for loop we are supposed to define a counter variable which decides how many time the loop will be executed based on a starting and ending value provided at the beginning of the loop.

The for loop automatically increments the value of the counter variable by 1 at the end of each loop cycle.

The programmer need not have to write any instruction for incrementing or decrementing value.

Syntax:

FOR counter\_variable IN start\_value..end\_value LOOP

statement to be executed

END LOOP;

1. What are the cursors and what are the cursors types?

**Cursor** is a Temporary Memory or Temporary Work Station. It is Allocated by Database Server at the Time of Performing DML(Data Manipulation Language) operations on Table by User. Cursors are used to store Database Tables. There are 2 types of Cursors: Implicit Cursors, and Explicit Cursors. These are explained as following below.

1. **Implicit Cursors:**  
   Implicit Cursors are also known as Default Cursors of SQL SERVER. These Cursors are allocated by SQL SERVER when the user performs DML operations.
2. **Explicit Cursors :**  
   Explicit Cursors are Created by Users whenever the user requires them. Explicit Cursors are used for Fetching data from Table in Row-By-Row Manner.

**How to create Explicit Cursor:**

1. **Declare Cursor Object.**  
   **Syntax :** DECLARE cursor\_name CURSOR FOR SELECT \* FROM table\_name

DECLARE s1 CURSOR FOR SELECT \* FROM studDetails

1. What is the procedure?

A stored procedure is a prepared SQL code that you can save, so the code can be reused over and over again.

1. What is the difference between procedure and function?

|  |  |
| --- | --- |
| Function | Procedure |
| It computes the outcome based on the inputs. | It completes tasks in a specific order. |
| A function is a tool for calculating something based on input. | While a procedure is a collection of commands that are executed in a specific order. |
| A SELECT statement can include a function call. | A SELECT statement does not include a procedure call. |
| The resultant value or control is returned to the calling function or code. | It returns control to the code or calls the function but not a value. |
| In a function, DML commands like (Insert, Delete, and Update) are not allowed. | DML statements can be utilized in procedures. |
| We can call a function via the procedure. | A procedure, on the other hand, cannot be called via a function. |
| It has to do with a particular expression. | It is not required to deal with a facial expression. |

1. What are the triggers and what are the triggers types?

**Trigger** is a statement that a system executes automatically when there is any modification to the database. In a trigger, we first specify when the trigger is to be executed and then the action to be performed when the trigger executes. Triggers are used to specify certain integrity constraints and referential constraints that cannot be specified using the constraint mechanism of SQL.

**Example –**   
Suppose, we are adding a tuple to the ‘Donors’ table that is some person has donated blood. So, we can design a trigger that will automatically add the value of donated blood to the ‘Blood\_record’ table.

**Types of Triggers –**   
We can define 6 types of triggers for each table:

1. **AFTER INSERT** activated after data is inserted into the table.
2. **AFTER UPDATE:** activated after data in the table is modified.
3. **AFTER DELETE:** activated after data is deleted/removed from the table.
4. **BEFORE INSERT:** activated before data is inserted into the table.
5. **BEFORE UPDATE:** activated before data in the table is modified.
6. **BEFORE DELETE:** activated before data is deleted/removed from the table.

Examples showing implementation of Triggers:

**1. Write a trigger to ensure that no employee of age less than 25 can be inserted in the database.**

delimiter $$

CREATE TRIGGER Check\_age BEFORE INSERT ON employee

FOR EACH ROW

BEGIN

IF NEW.age < 25 THEN

SIGNAL SQLSTATE '45000'

SET MESSAGE\_TEXT = 'ERROR:

AGE MUST BE ATLEAST 25 YEARS!';

END IF;

END; $$

delimiter;

**Explanation:** Whenever we want to insert any tuple to table ’employee’, then before inserting this tuple to the table, trigger named ‘Check\_age’ will be executed. This trigger will check the age attribute. If it is greater then 25 then this tuple will be inserted into the tuple otherwise an error message will be printed stating “ERROR: AGE MUST BE ATLEAST 25 YEARS!”

1. Write SQL Statements

# Business Intelligence

1. What is Business Intelligence?

Business intelligence combines business analytics, data mining, [data visualization](https://www.tableau.com/learn/articles/data-visualization), data tools and infrastructure, and best practices to help organizations make more data-driven decisions. In practice, you know you’ve got [modern business intelligence](https://www2.deloitte.com/content/dam/Deloitte/tr/Documents/deloitte-analytics/Modern%20Business%20Intelligence.pdf) when you have a comprehensive view of your organization’s data and use that data to drive change, eliminate inefficiencies, and quickly adapt to market or supply changes. Modern BI solutions prioritize flexible self-service analysis, governed data on trusted platforms, empowered business users, and speed to insight.

The benefits of business intelligence technologies include:

* Improved decision-making.
* Identifying customer behaviors, trends, and patterns.
* Enhanced operational efficiency and competitive advantage.
* Reduced costs and increased revenues.

1. What are the steps in BI?

* Phase 1. Extract data – Connecting to the original data sources and retrieving it from them. Data sources may be internal (databases, CRM, ERP, CMS, tools like Google Analytics or Excel) or external (order confirmation from suppliers, reviews from social media sites, public dataset repositories, etc.).
* Phase 2. Transform data – Placing data in temporary storage known as a staging area. Formatting data according to specified requirements and standards to make it suitable for analysis.
* Phase 3. Load data – Moving standardized data to a final storage destination – database, data lake, or data warehouse. If necessary, creating data marts – subsets of a data warehouse to store information from each of the company units, HR or sales department, for instance.
* Phase 4. Visualization – Exploring data and presenting results with visuals (via a user interface).

Steps 1, 2, and 3 are combined into ETL (extract, transform, load) operations. The ETL process specifies how heterogeneous data is retrieved from disparate sources, transformed into a form suitable for analysis, and loaded into a single destination. We won’t spend much time explaining it here since we did that in a dedicated article about an ETL developer.

1. What are the tools we use in BI (for ETL, Analysis, and Visualization)?

Timeline

Description automatically generated

**Supported deployment scenarios.** Based on your requirements for data security, budget, and possible need to scale, define what type of deployment you need and check whether the solution supports it. If you’re not sure about it, look for a consultant.

**Scalability.** Is it easy to increase storage, computing load, and the number of users? Most providers highlight this capability in a product description. You can always contact a vendor for more details.

**Ease of use.** From clear and well-structured documentation and published video tutorials to intuitive interface with drag-and-drop feature for data management, a good tool must have that all.

**Number and types of supported data sources.** Figure whether the tool allows for sourcing data in batches, in real-time, or near-to-real time. What connection options (API or/and web connectors) are included?

**Data transformation.** Solutions always include basic features for data transformation. Unique features could make a difference for you.

**Visualization.** You may want to look at the number and types of supported visuals, the capability to customize them, to embed dashboards into other applications. If you deal with streaming data, learn whether a platform supports real-time data visualization.

**Pricing options and terms.** Evaluate pricing plans and their conditions (fixed, pay-as-you-go, or discounted). Is there a trial to test a tool?

**Customer support.** Look what channels (email, call, etc.) you can use to reach customer support specialists, their working hours. Review website generally includes questions about the quality of service and support.

# Data Warehouse

1. What is the data warehouse?

A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transactional systems, [relational databases](https://aws.amazon.com/relational-database/), and other sources, typically on a regular cadence. Business analysts, data engineers, data scientists, and decision makers access the data through [business intelligence (BI) tools](https://aws.amazon.com/quicksight/), SQL clients, and other analytics applications.

1. What are the characteristics of a data warehouse?

* **Subject-oriented –**  
  A data warehouse is always a subject oriented as it delivers information about a theme instead of organization’s current operations. It can be achieved on specific theme. That means the data warehousing process is proposed to handle with a specific theme which is more defined. These themes can be sales, distributions, marketing etc.  
    
  A data warehouse never put emphasis only current operations. Instead, it focuses on demonstrating and analysis of data to make various decision. It also delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.
* **Integrated –**  
  It is somewhere same as subject orientation which is made in a reliable format. Integration means founding a shared entity to scale the all similar data from the different databases. The data also required to be resided into various data warehouse in shared and generally granted manner.  
    
  A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database. In addition, it must have reliable naming conventions, format and codes. Integration of data warehouse benefits in effective analysis of data. Reliability in naming conventions, column scaling, encoding structure etc. should be confirmed. Integration of data warehouse handles various subject related warehouse.
* **Time-Variant –**  
  In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time-variance is that once data is stored in the data warehouse then it cannot be modified, alter, or updated.
* **Non-Volatile –**  
  As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted. It includes the mammoth quantity of data that is inserted into modification between the selected quantity on logical business. It evaluates the analysis within the technologies of warehouse.  
    
  In this, data is read-only and refreshed at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Two types of data operations done in the data warehouse are:
  1. Data Loading
  2. Data Access

1. What is the difference between the database and the data warehouse?

|  |  |
| --- | --- |
| Database System | Data Warehouse |
| It supports operational processes. | It supports analysis and performance reporting. |
| Capture and maintain the data. | Explore the data. |
| Current data. | Multiple years of history. |
| Data is balanced within the scope of this one system. | Data must be integrated and balanced from multiple system. |
| Data is updated when transaction occurs. | Data is updated on scheduled processes. |
| Data verification occurs when entry is done. | Data verification occurs after the fact. |
| 100 MB to GB. | 100 GB to TB. |
| ER based. | Star/Snowflake. |
| Application oriented. | Subject oriented. |
| Primitive and highly detailed. | Summarized and consolidated. |
| Flat relational. | Multidimensional. |

1. What is the difference between a data warehouse and big data?

|  |  |  |
| --- | --- | --- |
| S.No. | Big Data | Data Warehouse |
| 1. | Big data is the data which is in enormous form on which technologies can be applied. | Data warehouse is the collection of historical data from different operations in an enterprise. |
| 2. | Big data is a technology to store and manage large amount of data. | Data warehouse is an architecture used to organize the data. |
| 3. | It takes structured, non-structured or semi-structured data as an input. | It only takes structured data as an input. |
| 4. | Big data does processing by using distributed file system. | Data warehouse doesn’t use distributed file system for processing. |
| 5. | Big data doesn’t follow any SQL queries to fetch data from database. | In data warehouse we use SQL queries to fetch data from relational databases. |
| 6. | Apache Hadoop can be used to handle enormous amount of data. | Data warehouse cannot be used to handle enormous amount of data. |
| 7. | When new data is added, the changes in data are stored in the form of a file which is represented by a table. | When new data is added, the changes in data do not directly impact the data warehouse. |
| 8. | Big data doesn’t require efficient management techniques as compared to data warehouse. | Data warehouse requires more efficient management techniques as the data is collected from different departments of the enterprise. |

1. What is the difference between OLTP and OLAP?

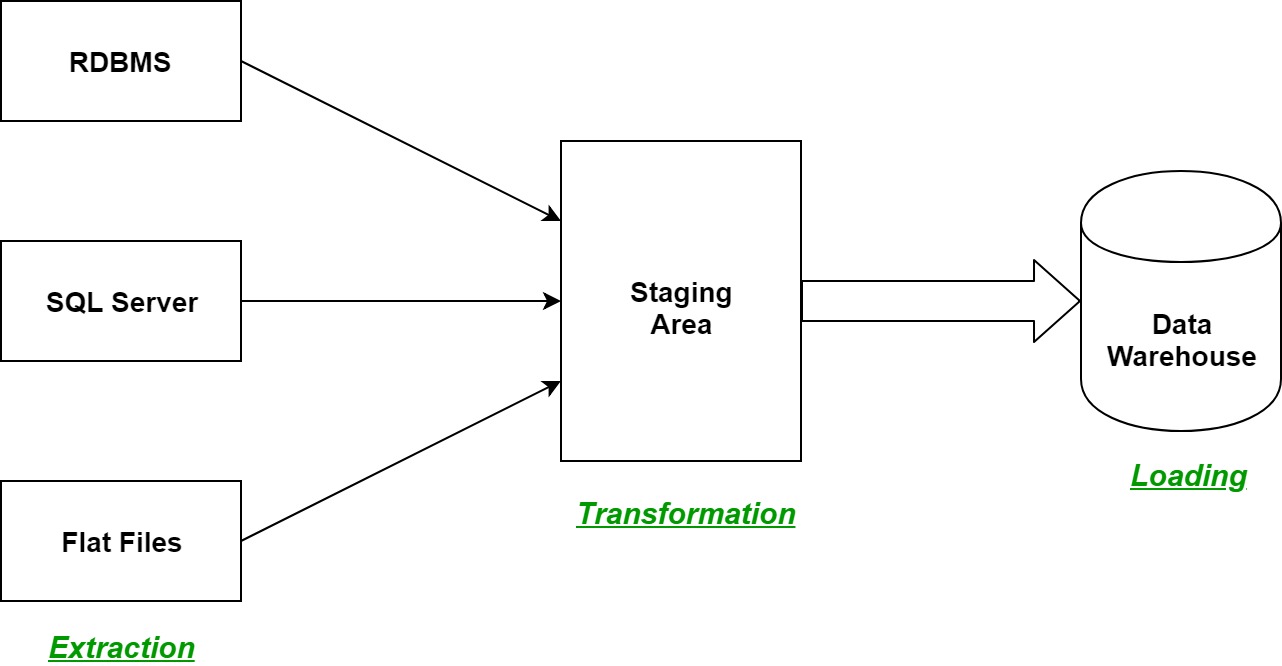
|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Category | OLAP (Online analytical processing) | OLTP (Online transaction processing) |
| 1. | **Definition** | It is well-known as an online database query management system. | It is well-known as an online database modifying system. |
| 2. | **Data source** | Consists of historical data from various Databases. In other words, different OLTP databases are used as data sources for OLAP. | Consists of only of operational current data. In other words, the original data source is OLTP and its transactions. |
| 3. | **Method used** | It makes use of a data warehouse. | It makes use of a standard database management system (DBMS). |
| 4. | **Application** | It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc. | It is application-oriented. Used for business tasks. |
| 5. | **Normalized** | In an OLAP database, tables are not normalized. | In an OLTP database, tables are normalized (3NF). |
| 6. | **Usage of data** | The data is used in planning, problem-solving, and decision-making. | The data is used to perform day-to-day fundamental operations. |
| 7. | **Task** | It reveals a snapshot of present business tasks. | It provides a multi-dimensional view of different business tasks. |
| 8. | **Purpose** | It serves the purpose to extract information for analysis and decision-making. | It serves the purpose to Insert, Update, and Delete information from the database. |
| 9. | **Volume of data** | A large amount of data is stored typically in TB, PB | The size of the data is relatively small as the historical data is archived. For ex MB, GB |
| 10. | **Queries** | Relatively slow as the amount of data involved is large. Queries may take hours. | Very Fast as the queries operate on 5% of the data. |
| 11. | **Update** | The OLAP database is not often updated. As a result, data integrity is unaffected. | The data integrity constraint must be maintained in an OLTP database. |
| 12. | **Backup and Recovery** | It only need backup from time to time as compared to OLTP. | Backup and recovery process is maintained rigorously |
| 13. | **Processing time** | The processing of complex queries can take a lengthy time. | It is comparatively fast in processing because of simple and straightforward queries. |
| 14. | **Types of users** | This data is generally managed by CEO, MD, GM. | This data is managed by clerks, managers. |
| 15. | **Operations** | Only read and rarely write operation. | Both read and write operations. |
| 16. | **Updates** | With lengthy, scheduled batch operations, data is refreshed on a regular basis. | The user initiates data updates, which are brief and quick. |
| 17. | **Nature of audience** | Process that is focused on the customer. | Process that is focused on the market. |
| 18. | **Database Design** | Design with a focus on the subject. | Design that is focused on the application. |
| 19. | **Productivity** | Improves the efficiency of business analysts. | Enhances the user’s productivity. |

1. What is Data Warehousing?

Q 1

1. What are the processes that can be done in the data warehouse?

ETL is a process in Data Warehousing and it stands for **Extract**, **Transform** and **Load**. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area, and then finally, loads it into the Data Warehouse system.



Let us understand each step of the ETL process in-depth:

1. **Extraction:**  
   The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML, and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process.
2. **Transformation:**  
   The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes/tasks:
   * Filtering – loading only certain attributes into the data warehouse.
   * Cleaning – filling up the NULL values with some default values, mapping U.S.A, United States, and America into USA, etc.
   * Joining – joining multiple attributes into one.
   * Splitting – splitting a single attribute into multiple attributes.
   * Sorting – sorting tuples on the basis of some attribute (generally key-attribute).
3. **Loading:**  
   The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading solely depends on the requirements and varies from system to system.

ETL process can also use the pipelining concept i.e. as soon as some data is extracted, it can transformed and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed. The block diagram of the pipelining of ETL process is shown below:

Diagram

Description automatically generated

**ETL Tools:** Most commonly used ETL tools are **Hevo**, Sybase, Oracle Warehouse builder, CloverETL, and MarkLogic.

**Data Warehouses:** Most commonly used Data Warehouses are **Snowflake**, Redshift, BigQuery, and Firebolt.

1. What is Data Modeling? + Types of Data Modeling?

**Data modeling (data modelling)** is the process of creating a data model for the data to be stored in a database. This data model is a conceptual representation of Data objects, the associations between different data objects, and the rules.

The **Data Model** is defined as an abstract model that organizes data description, data semantics, and consistency constraints of data. The data model emphasizes on what data is needed and how it should be organized instead of what operations will be performed on data. Data Model is like an architect’s building plan, which helps to build conceptual models and set a relationship between data items.

The two types of Data Modeling Techniques are

1. Entity Relationship (E-R) Model
2. UML (Unified Modelling Language)

**Types of Data Models**: There are mainly three different types of data models: conceptual data models, logical data models, and physical data models, and each one has a specific purpose. The data models are used to represent the data and how it is stored in the database and to set the relationship between data items.

1. **Conceptual Data Model:** This Data Model defines **WHAT** the system contains. This model is typically created by Business stakeholders and Data Architects. The purpose is to organize, scope and define business concepts and rules.
2. **Logical Data Model:** Defines **HOW** the system should be implemented regardless of the DBMS. This model is typically created by Data Architects and Business Analysts. The purpose is to developed technical map of rules and data structures.
3. **Physical Data Model**: This Data Model describes **HOW** the system will be implemented using a specific DBMS system. This model is typically created by DBA and developers. The purpose is actual implementation of the database.

Diagram

Description automatically generated

The snowflake design is the result of further expansion and normalized of the dimension table. In other words, a dimension table is said to be snowflaked if the low-cardinality attribute of the dimensions has been divided into separate normalized tables. These tables are then joined to the original dimension table with referential constraints (foreign key constrain).   
Generally, snowflaking is not recommended in the dimension table, as it hampers the understandability and performance of the dimension model as more tables would be required to be joined to satisfy the queries.

**Characteristics of snowflake schema:**

**Introduction:** The snowflake schema is a variant of the star schema. Here, the centralized fact table is connected to multiple dimensions. In the snowflake schema, dimensions are present in a normalized form in multiple related tables. The snowflake structure materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent tables. The snowflake effect affects only the dimension tables and does not affect the fact tables.

The dimension model of a snowflake under the following conditions:

* The snowflake schema uses small disk space.
* It is easy to implement dimension that is added to the schema.
* There are multiple tables, so performance is reduced.
* The dimension table consists of two or more sets of attributes that define information at different grains.
* The sets of attributes of the same dimension table are being populated by different source systems.

Diagram

Description automatically generated

**Advantages:** There are two main advantages of snowflake schema given below:

* It provides structured data which reduces the problem of data integrity.
* It uses small disk space because data are highly structured.

**Disadvantages:**

* Snowflaking reduces space consumed by dimension tables but compared with the entire data warehouse the saving is usually insignificant.
* Avoid snowflaking or normalization of a dimension table, unless required and appropriate.
* Do not snowflake hierarchies of one dimension table into separate tables. Hierarchies should belong to the dimension table only and should never be snowflakes.
* Multiple hierarchies that can belong to the same dimension have been designed at the lowest possible detail.

**Star schema** is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables. The star schema is a necessary cause of the snowflake schema. It is also efficient for handling basic queries.

It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star’s points. Below is an example to demonstrate the Star Schema:

In the above demonstration, SALES is a fact table having attributes i.e. (Product ID, Order ID, Customer ID, Employer ID, Total, Quantity, Discount) which references to the dimension tables. **Employee dimension table** contains the attributes: Emp ID, Emp Name, Title, Department and Region. *Product dimension table* contains the attributes: Product ID, Product Name, Product Category, Unit Price. *Customer dimension table* contains the attributes: Customer ID, Customer Name, Address, City, Zip. *Time dimension table* contains the attributes: Order ID, Order Date, Year, Quarter, Month.

**Model of Star Schema :**   
In Star Schema, Business process data, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact data. Sales price, sale quantity, distant, speed, weight, and weight measurements are few examples of fact data in star schema.   
Often, A Star Schema having multiple dimensions is termed as Centipede Schema. It is easy to handle a star schema which have dimensions of few attributes.

Diagram

Description automatically generated

**Advantages of Star Schema :**

1. **Simpler Queries –**  
   Join logic of star schema is quite cinch in comparison to other join logic which are needed to fetch data from a transactional schema that is highly normalized.
2. **Simplified Business Reporting Logic –**   
   In comparison to a transactional schema that is highly normalized, the star schema makes simpler common business reporting logic, such as of reporting and period-over-period.
3. **Feeding Cubes –**   
   Star schema is widely used by all OLAP systems to design OLAP cubes efficiently. In fact, major OLAP systems deliver a ROLAP mode of operation which can use a star schema as a source without designing a cube structure.

**Disadvantages of Star Schema –**

1. Data integrity is not enforced well since in a highly de-normalized schema state.
2. Not flexible in terms if analytical needs as a normalized data model.
3. Star schemas don’t reinforce many-to-many relationships within business entities – at least not frequently.
4. Can we update a record in a data warehouse?

Slowly Changing Dimensions in Data Warehouse is an important concept that is used to enable the historic aspect of data in an analytical system. As you know, the data warehouse is used to analyze historical data, it is essential to store the different states of data.

In data warehousing, we have fact and dimension tables to store the data. Dimensional tables are used to analyze the measures in the fact tables. In a data environment, data is initiated at operational databases and data will be extracted-transformed-loaded (ETL) to the data warehouse to suit the analytical environment.

1. What is data mart?

|  |  |  |
| --- | --- | --- |
| S.NO | Data Warehouse | Data Mart |
| 1. | **Data warehouse is a Centralised system.** | **While it is a decentralised system.** |
| 2. | **In data warehouse, lightly denormalization takes place.** | **While in Data mart, highly denormalization takes place.** |
| 3. | **Data warehouse is top-down model.** | **While it is a bottom-up model.** |
| 4. | **To built a warehouse is difficult.** | **While to build a mart is easy.** |
| 5. | **In data warehouse, Fact constellation schema is used.** | **While in this, Star schema and snowflake schema are used.** |
| 6. | **Data Warehouse is flexible.** | **While it is not flexible.** |
| 7. | **Data Warehouse is the data-oriented in nature.** | **While it is the project-oriented in nature.** |
| 8. | **Data Ware house has long life.** | **While data-mart has short life than warehouse.** |
| 9. | **In Data Warehouse, Data are contained in detail form.** | **While in this, data are contained in summarized form.** |
| 10. | **Data Warehouse is vast in size.** | **While data mart is smaller than warehouse.** |
| 11. | **It collects data from various data sources.** | **It generally stores data from a data warehouse.** |
| 12. | **Long time for processing the data because of large data.** | **Less time for processing the data because of handling only a small amount of data.** |
| 13. | **Complicated design process of creating schemas and views.** | **Easy design process of creating schemas and views.** |

1. What is Data Cube?

Grouping of data in a multidimensional matrix is called data cubes. In Dataware housing, we generally deal with various multidimensional data models as the data will be represented by multiple dimensions and multiple attributes. This multidimensional data is represented in the data cube as the cube represents a high-dimensional space. The Data cube pictorially shows how different attributes of data are arranged in the data model. Below is the diagram of a general data cube.

### **Data cube classification:**

The data cube can be classified into two categories:

* **Multidimensional data cube:** It basically helps in storing large amounts of data by making use of a multi-dimensional array. It increases its efficiency by keeping an index of each dimension. Thus, dimensional is able to retrieve data fast.
* **Relational data cube:** It basically helps in storing large amounts of data by making use of relational tables. Each relational table displays the dimensions of the data cube. It is slower compared to a Multidimensional Data Cube.

### Data cube operations:

Diagram

Description automatically generated

Data cube operations are used to manipulate data to meet the needs of users. These operations help to select particular data for the analysis purpose. There are mainly 5 operations listed below-

* **Roll-up**: operation and aggregate certain similar data attributes having the same dimension together. For example, if the data cube displays the daily income of a customer, we can use a roll-up operation to find the monthly income of his salary.
* **Drill-down**: this operation is the reverse of the roll-up operation. It allows us to take particular information and then subdivide it further for coarser granularity analysis. It zooms into more detail. For example- if India is an attribute of a country column and we wish to see villages in India, then the drill-down operation splits India into states, districts, towns, cities, villages and then displays the required information.
* **Slicing**: this operation filters the unnecessary portions. Suppose in a particular dimension, the user doesn’t need everything for analysis, rather a particular attribute. For example, country=”jamaica”, this will display only about jamaica and only display other countries present on the country list.
* **Dicing**: this operation does a multidimensional cutting, that not only cuts only one dimension but also can go to another dimension and cut a certain range of it. As a result, it looks more like a subcube out of the whole cube(as depicted in the figure). For example- the user wants to see the annual salary of Jharkhand state employees.
* **Pivot**: this operation is very important from a viewing point of view. It basically transforms the data cube in terms of view. It doesn’t change the data present in the data cube. For example, if the user is comparing year versus branch, using the pivot operation, the user can change the viewpoint and now compare branch versus item type.

### Advantages of data cubes:

* Helps in giving a summarised view of data.
* Data cubes store large data in a simple way.
* Data cube operation provides quick and better analysis,
* Improve performance of data.

1. What is the ETL?

Q 7

1. What is the difference between snowflake and star schema?

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Star Schema** | **Snowflake Schema** |
| Definition and Meaning | A star schema contains both dimension tables and fact tables in it. | A snowflake schema contains all three- dimension tables, fact tables, and sub-dimension tables. |
| Type of Model | It is a top-down model type. | It is a bottom-up model type. |
| Space Occupied | It makes use of more allotted space. | It makes use of less allotted space. |
| Time Taken for Queries | With the Star Schema, the process of execution of queries takes less time. | With the Snowflake Schema, the process of execution of queries takes more time. |
| Use of Normalization | The Star Schema does not make use of normalization. | The Snowflake Schema makes use of both Denormalization as well as Normalization. |
| Complexity of Design | The design of a Star Schema is very simple. | The designing of a Snowflake Schema is very complex. |
| Query Complexity | It is very low in the case of a Star Schema. | It is comparatively much higher in the case of a Snowflake Schema. |
| Complexity of Understanding | It is very easy to understand a Star Schema. | It is comparatively more difficult to understand a Snowflake Schema. |
| Total Number of Foreign Keys | The total number of foreign keys is less in the case of a Star Schema. | The total number of foreign keys is more in the case of a Snowflake Schema. |
| Data Redundancy | Data redundancy is comparatively higher in Star Schema. | Data redundancy is comparatively lower in Snowflake Schema. |

|  |  |  |
| --- | --- | --- |
| S.NO | Star Schema | Snowflake Schema |
| 1. | In star schema, The fact tables and the dimension tables are contained. | While in snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained. |
| 2. | Star schema is a top-down model. | While it is a bottom-up model. |
| 3. | Star schema uses more space. | While it uses less space. |
| 4. | It takes less time for the execution of queries. | While it takes more time than star schema for the execution of queries. |
| 5. | In star schema, Normalization is not used. | While in this, Both normalization and denormalization are used. |
| 6. | It’s design is very simple. | While it’s design is complex. |
| 7. | The query complexity of star schema is low. | While the query complexity of snowflake schema is higher than star schema. |
| 8. | It’s understanding is very simple. | While it’s understanding is difficult. |
| 9. | It has less number of foreign keys. | While it has more number of foreign keys. |
| 10. | It has high data redundancy. | While it has low data redundancy. |

1. What is the difference between 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. |

# Big Data

1. Why is the big data?

Big data is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML) projects, [predictive modeling](https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling) and other advanced analytics applications.

Companies [use big data in their systems](https://www.techtarget.com/searchbusinessanalytics/feature/8-big-data-use-cases-for-businesses-and-industry-examples) to improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profits. Businesses that use it effectively hold a potential competitive advantage over those that don't because they're able to make faster and more informed business decisions.

1. What is the big data? (V's of Big Data)

**1. Volume:**

* The name ‘Big Data’ itself is related to a size which is enormous.
* Volume is a huge amount of data.
* To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a ‘Big Data’. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
* Hence while dealing with Big Data it is necessary to consider a characteristic ‘Volume’.
* Example: In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 ExaBytes of data.

**2. Velocity:**

* Velocity refers to the high speed of accumulation of data.
* In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.
* There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
* Sampling data can help in dealing with the issue like ‘velocity’.
* Example: There are more than 3.5 billion searches per day are made on Google. Also, FaceBook users are increasing by 22%(Approx.) year by year.

**3. Variety:**

* It refers to nature of data that is structured, semi-structured and unstructured data.
* It also refers to heterogeneous sources.
* Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

**4. Veracity:**

* It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
* Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
* Example: Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

**5. Value:**

* After having the 4 V’s into account there comes one more V which stands for Value!. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
* Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V’s.

1. What are the data types?
   1. **Structured data**: This data is basically an organized data. It generally refers to data that has defined the length and format of data.
   2. **Semi- Structured data**: This data is basically a semi-organised data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.
   3. **Unstructured data**: This data basically refers to unorganized data. It generally refers to data that doesn’t fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can’t be stored in the form of rows and columns.
2. What is Data Lake?

A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semistructured, and unstructured data. It can store data in its native format and process any variety of it, ignoring size limits.

1. What is the difference between ETL & ELT?

|  |  |
| --- | --- |
| ELT | ETL |
| ELT tools do not require additional hardware | ETL tools require specific hardware with their own engines to perform transformations |
| Mostly Hadoop or NoSQL database to store data.Rarely RDBMS is used | RDBMS is used exclusively to store data |
| As all components are in one system, loading is done only once | As ETL uses staging area, extra time is required to load the data |
| Time to transform data is independent of the size of data | The system has to wait for large sizes of data. As the size of data increases, transformation time also increases |
| It is cost effective and available to all business using SaaS solution | Not cost effective for small and medium business |
| The data transformed is used by data scientists and advanced analysts | The data transformed is used by users reading report and SQL coders |
| Creates ad hoc views.Low cost for building and maintaining | Views are created based on multiple scripts.Deleting view means deleting data |
| Best for unstructured and non-relational data. Ideal for data lakes. Suited for very large amounts of data | Best for relational and structured data. Better for small to medium amounts of data |

1. What is the difference between Database and Big data?

|  |  |
| --- | --- |
| Tradational Data | Big Data |
| Traditional data is generated in enterprise level. | Big data is generated outside the enterprise level. |
| Its volume ranges from Gigabytes to Terabytes. | Its volume ranges from Petabytes to Zettabytes or Exabytes. |
| secondTraditional database system deals with structured data. | Big data system deals with structured, semi-structured,database, and unstructured data. |
| Traditional data is generated per hour or per day or more. | But big data is generated more frequently mainly per seconds. |
| Traditional data source is centralized and it is managed in centralized form. | Big data source is distributed and it is managed in distributed form. |
| Data integration is very easy. | Data integration is very difficult. |
| Normal system configuration is capable to process traditional data. | High system configuration is required to process big data. |
| The size of the data is very small. | The size is more than the traditional data size. |
| Traditional data base tools are required to perform any data base operation. | Special kind of data base tools are required to perform any databaseschema-based operation. |
| Normal functions can manipulate data. | Special kind of functions can manipulate data. |
| Its data model is strict schema based and it is static. | Its data model is a flat schema based and it is dynamic. |
| Traditional data is stable and inter relationship. | Big data is not stable and unknown relationship. |
| Traditional data is in manageable volume. | Big data is in huge volume which becomes unmanageable. |
| It is easy to manage and manipulate the data. | It is difficult to manage and manipulate the data. |
| Its data sources includes ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data etc. | Its data sources includes social media, device data, sensor data, video, images, audio etc. |

1. What are the tools in big data?

APACHE Hadoop / Cassandra / Qubole / Xplenty / ….

### 1. APACHE Hadoop

It’s a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel. It can process both structured and unstructured data from one server to multiple computers. [Hadoop](https://www.geeksforgeeks.org/hadoop-an-introduction/?ref=gcse) also offers **cross-platform** support for its users. Today, it is the best ***big data analytic tool*** and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

**Features of Apache Hadoop:**

* Free to use and offers an efficient storage solution for businesses.
* Offers quick access via HDFS (Hadoop Distributed File System).
* Highly flexible and can be easily implemented with MySQL, and JSON.
* Highly scalable as it can distribute a large amount of data in small segments.
* It works on small commodity hardware like JBOD or a bunch of disks.

### 2. Cassandra

[APACHE Cassandra](https://www.geeksforgeeks.org/introduction-to-apache-cassandra/?ref=gcse) is an open-source NoSQL distributed database that is used to fetch large amounts of data. It’s one of the **most popular tools for data analytics** and has been praised by many tech companies due to its high scalability and availability without compromising speed and performance. It is **capable of delivering thousands of operations every second** and can handle petabytes of resources with almost zero downtime. It was created by Facebook back in 2008 and was published publically.

**Features of APACHE Cassandra:**

* *Data Storage Flexibility:* It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
* *Data Distribution System:* Easy to distribute data with the help of replicating data on multiple data centers.
* *Fast Processing:* Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.
* *Fault-tolerance:* The moment, if any node fails, it will be replaced without any delay.

### 3. Qubole

It’s an open-source big data tool that helps in fetching data in a value of chain using ad-hoc analysis in machine learning. Qubole is a data lake platform that offers end-to-end service with reduced time and effort which are required in moving data pipelines. It is capable of configuring multi-cloud services such as AWS, Azure, and Google Cloud. Besides, it also helps in lowering the cost of cloud computing by 50%.

**Features of Qubole:**

* *Supports ETL process:* It allows companies to **migrate data from multiple sources in one place**.
* *Real-time Insight:* It monitors user’s systems and allows them to view real-time insights
* *Predictive Analysis:* Qubole offers predictive analysis so that companies can take actions accordingly for targeting more acquisitions.
* *Advanced Security System:* To protect users’ data in the cloud, Qubole uses an advanced security system and also ensures to protect any future breaches. Besides, it also allows encrypting cloud data from any potential threat.

### 4. Xplenty

It is a data analytic tool for building a data pipeline by using minimal codes in it. It offers a wide range of solutions for sales, marketing, and support. With the help of its interactive graphical interface, it provides solutions for [*ETL*](https://www.geeksforgeeks.org/etl-tools-overview/?ref=gcse)*, ELT*, etc. The best part of using Xplenty is its low investment in hardware & software and its offers support via **email, chat, telephonic and virtual meetings**. Xplenty is a platform to process data for analytics over the cloud and segregates all the data together.

**Features of Xplenty:**

* *Rest API:* A user can possibly do anything by implementing Rest API
* *Flexibility:* Data can be sent, and pulled to databases, warehouses, and salesforce.
* *Data Security:* It offers SSL/TSL encryption and the platform is capable of verifying algorithms and certificates regularly.
* *Deployment:* It offers integration apps for both cloud & in-house and supports deployment to integrate apps over the cloud.

### 5. Spark

[APACHE Spark](https://www.geeksforgeeks.org/overview-of-apache-spark/) is another framework that is used to process data and perform numerous tasks on a large scale.  It is also used to process data via multiple computers with the help of distributing tools. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is **capable of handling multi-petabytes of data** as well. Recently, Spark made a record of processing **100 terabytes of data in just 23 minutes** which broke the previous world record of **Hadoop (71 minutes)**. This is the reason why big tech giants are moving towards spark now and is highly suitable for ML and AI today.

**Features of APACHE Spark:**

* *Ease of use:* It allows users to run in their preferred language. (JAVA, Python, etc.)
* *Real-time Processing:* Spark can handle real-time streaming via Spark Streaming
* *Flexible:* It can run on, Mesos, Kubernetes, or the cloud.

### 6. Mongo DB

Came in limelight in 2010, is a free, open-source platform and a **document-oriented (NoSQL) database** that is used to store a high volume of data. It uses collections and documents for storage and its document consists of key-value pairs which are considered a basic unit of [Mongo DB](https://www.geeksforgeeks.org/mongodb-an-introduction/). It is so popular among developers due to its availability for multi-programming languages such as Python, Jscript, and Ruby.

**Features of Mongo DB:**

* *Written in C++:* It’s a schema-less DB and can hold varieties of documents inside.
* *Simplifies Stack:* With the help of mongo, a user can easily store files without any disturbance in the stack.
* *Master-Slave Replication*: It can write/read data from the master and can be called back for backup.

### 7. Apache Storm

A storm is a robust, user-friendly tool used for data analytics, especially in small companies. The best part about the storm is that it has no language barrier (programming) in it and can support any of them. It was designed to **handle a pool of large data in fault-tolerance and horizontally scalable methods**. When we talk about real-time data processing, Storm leads the chart because of its distributed real-time big data processing system, due to which today many tech giants are using APACHE Storm in their system. Some of the most notable names are Twitter, Zendesk, NaviSite, etc.

**Features of Storm:**

* *Data Processing:* Storm process the data even if the node gets disconnected
* *Highly Scalable:* It keeps the momentum of performance even if the load increases
* *Fast:* The speed of APACHE Storm is impeccable and can process up to **1 million messages of 100 bytes on a single node.**

### 8. SAS

Today it is one of the best tools for creating statistical modeling used by data analysts. By using [SAS](https://www.geeksforgeeks.org/introduction-to-sas-programming/), a data scientist can mine, manage, extract or update data in different variants from different sources. Statistical Analytical System or SAS allows a user to access the data in any format (SAS tables or Excel worksheets). Besides that it also offers a cloud platform for business analytics called **SAS Viya** and also to get a strong grip on AI & ML, they have introduced new tools and products.

**Features of SAS:**

* *Flexible Programming Language:* It offers easy-to-learn syntax and has also vast libraries which make it suitable for non-programmers
* *Vast Data Format:* It provides support for many programming languages which also include SQL and carries the ability to read data from any format.
* *Encryption:* It provides end-to-end security with a feature called **SAS/SECURE**.

### 9. Data Pine

Datapine is an analytical used for BI and was founded back in 2012 (Berlin, Germany). In a short period of time, it has gained much popularity in a number of countries and it’s mainly used for data extraction (for small-medium companies fetching data for close monitoring). With the help of its enhanced UI design, anyone can visit and check the data as per their requirement and offer in 4 different price brackets, starting from $249 per month. They do offer dashboards by functions, industry, and platform.

**Features of Datapine:**

* *Automation:* To cut down the manual chase, datapine offers a wide array of AI assistant and BI tools.
* *Predictive Tool:* datapine provides forecasting/predictive analytics by using historical and current data, it derives the future outcome.
* *Add on:* It also offers intuitive **widgets, visual analytics & discovery, ad hoc reporting**, etc.

### 10. Rapid Miner

It’s a fully automated visual workflow design tool used for data analytics. It’s a no-code platform and users aren’t required to code for segregating data. Today, it is being heavily used in many industries such as ed-tech, training, research, etc. Though it’s an open-source platform but has a limitation of adding **10000 data rows and a single logical processor**. With the help of Rapid Miner, one can easily deploy their ML models to the web or mobile (only when the user interface is ready to collect real-time figures).

**Features of Rapid Miner:**

* *Accessibility:* It allows users to access 40+ types of files (SAS, ARFF, etc.) via URL
* *Storage:* Users can access cloud storage facilities such as AWS and dropbox
* *Data validation:* Rapid miner enables the visual display of multiple results in history for better evaluation.

## Conclusion

[***Big data***](https://www.geeksforgeeks.org/what-is-big-data/) has been in limelight for the past few years and will continue to dominate the market in almost every sector for every market size. The demand for **big data** is booming at an enormous rate and ample tools are available in the market today, all you need is the right approach and choose the **best data analytic tool** as per the project’s requirement.

1. What are (Hadoop/Spark / Mapreduce / Hive / impala/ Impala/ Kafka/…)?

Tools in the big data.

و نكمل المعلومات من q 7

# Data Science + Machine Learning + Data Mining (Data Science Track)

1. What is data science?

[**Data Science**](https://www.geeksforgeeks.org/data-science-solving-linear-equations-2/) is an interdisciplinary field that focuses on extracting knowledge from data sets which are typically huge in amount. The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization. As such, it incorporates skills from computer science, mathematics, statics, information visualization, graphic, and business.

1. What is the difference between data scientists and data analysts?

Generally, we hear different designations about CS Engineers like Data Scientist, Data Analyst and Data Engineer. Let us discuss the differences between the above three roles.

* **Data Analyst** – The main focus of this person’s job would be on optimization of scenarios, say how an employee can improve the company’s product growth. Data Cleaning and organizing of raw data, analyzing and visualization of data to interpret the analysis and to present the technical analysis of data.  
  Skills needed for Data Analyst are R, Python, SQL, SAS, SAS Miner.
* **Data Scientist** – The predominant focus will be on the futuristic display of data. They provide both supervised and unsupervised learning of data, say classification and regression of data, Neural networks. The continuous regression analysis would be using machine learning techniques.  
  Skills needed for Data Scientist are R, Python, SQL, SAS, Pig, Apache Spark, Hadoop, Java, Perl.
* **Data Engineer** – Data Engineers concentrate more on optimization techniques and building of data in a proper manner. The main aim of a data engineer is continuously improving the data consumption. Mainly a data engineer works at the back end. Optimized machine learning algorithms were used for maintaining data and to make data to be available in most accurate manner.  
  Skills needed for Data Engineer are Pig, Hive, Hadoop, MapReduce techniques.

1. What is the data cleaning? How we clean the data?

**Data Cleaning:**   
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. 

* **(a). Missing Data:**   
  This situation arises when some data is missing in the data. It can be handled in various ways.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**   
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**   
  Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**   
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

1. What is Data Mining?

Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets. Given the evolution of [data warehousing](https://www.ibm.com/cloud/learn/data-warehouse) technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades, assisting companies by transforming their raw data into useful knowledge. However, despite the fact that that technology continuously evolves to handle data at a large-scale, leaders still face challenges with scalability and automation.

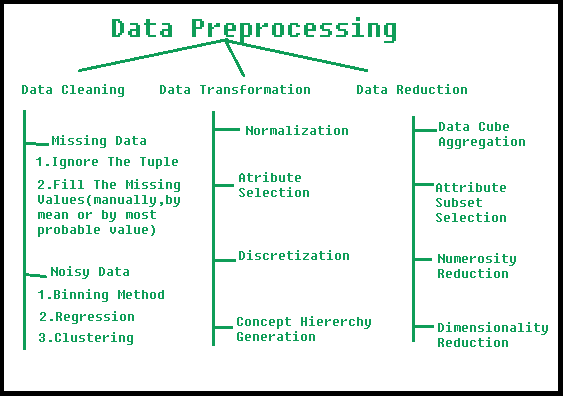
Data mining has improved organizational decision-making through insightful data analyses. The data mining techniques that underpin these analyses can be divided into two main purposes; they can either describe the target dataset or they can predict outcomes through the use of [machine learning](https://www.ibm.com/cloud/learn/machine-learning) algorithms. These methods are used to organize and filter data, surfacing the most interesting information, from fraud detection to user behaviors, bottlenecks, and even security breaches.

When combined with data analytics and visualization tools, like [Apache Spark](https://www.ibm.com/cloud/learn/apache-spark), delving into the world of data mining has never been easier and extracting relevant insights has never been faster. Advances within [artificial intelligence](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) only continue to expedite adoption across industries.

1. What are the real-life applications of data mining + ML?

* Marketing. Data mining is used to explore increasingly large databases and to improve market segmentation. By analysing the relationships between parameters such as customer age, gender, tastes, etc., it is possible to guess their behaviour in order to direct personalised loyalty campaigns. Data mining in marketing **also predicts which users are likely to unsubscribe from a service,** what interests them based on their searches, or what a mailing list should include to achieve a higher response rate.
* Retail. Supermarkets, for example, use joint purchasing patterns to identify product associations and decide how to place them in the aisles and on the shelves. **Data mining also detects which offers are most valued by customers** or increase sales at the checkout queue.
* Banking. Banks use data mining to better understand market risks. It is commonly applied to credit ratings and to intelligent anti-fraud systems to analyse transactions, card transactions, purchasing patterns and customer financial data. Data mining also **allows banks to learn more about our online preferences or habits** to optimise the return on their marketing campaigns, study the performance of sales channels or manage regulatory compliance obligations.
* Medicine. Data mining enables more accurate diagnostics. Having all of the patient's information, such as medical records, physical examinations, and treatment patterns, allows more effective treatments to be prescribed. It also **enables more effective, efficient and cost-effective management of health resources** by identifying risks, predicting illnesses in certain segments of the population or forecasting the length of hospital admission. Detecting fraud and irregularities, and strengthening ties with patients with an enhanced knowledge of their needs are also advantages of using data mining in medicine.
* Television and radio. There are networks that apply real time data mining to measure their online television (IPTV) and radio audiences. These systems collect and analyse, on the fly, anonymous information from channel views, broadcasts and programming. **Data mining allows networks to make personalised recommendations** to radio listeners and TV viewers, as well as get to know their interests and activities in real time and better understand their behaviour. Networks also gain valuable knowledge for their advertisers, who use this data to target their potential customers more accurately.

1. What is the Process of Data Mining/knowledge Discovery Process?



1. What is the Challenges of Data Mining?

Nowadays [Data Mining](https://www.geeksforgeeks.org/data-mining/) and knowledge discovery are evolving a crucial technology for business and researchers in many domains.Data Mining is developing into established and trusted discipline, many still pending challenges have to be solved.

Some of these challenges are given below.

1. **Security and Social Challenges:**  
   Decision-Making strategies are done through data collection-sharing, so it requires considerable security. Private information about individuals and sensitive information are collected for customers profiles, user behaviour pattern understanding. Illegal access to information and the confidential nature of information becoming an important issue.
2. **User Interface:**  
   The knowledge discovered is discovered using data mining tools is useful only if it is interesting and above all understandable by the user. From good visualization interpretation of data, mining results can be eased and helps better understand their requirements. To obtain good visualization many research is carried out for big data sets that display and manipulate mined knowledge.  
   **(i) Mining based on Level of Abstraction:** Data Mining process needs to be collaborative because it allows users to concentrate on pattern finding, presenting and optimizing requests for data mining based on returned results.  
   **(ii) Integration of Background Knowledge:** Previous information may be used to express discovered patterns to direct the exploration processes and to express discovered patterns.
3. **Mining Methodology Challenges:**  
   These challenges are related to data mining approaches and their limitations. Mining approaches that cause the problem are:

**(i)** Versatility of the mining approaches,

**(ii)** Diversity of data available,

**(iii)** Dimensionality of the domain,

**(iv)** Control and handling of noise in data, etc.

Different approaches may implement differently based upon data consideration. Some algorithms require noise-free data. Most data sets contain exceptions, invalid or incomplete information lead to complication in the analysis process and some cases compromise the precision of the results.

1. **Complex Data:**  
   Real-world data is heterogeneous and it could be multimedia data containing images, audio and video, complex data, temporal data, spatial data, time series, natural language text etc. It is difficult to handle these various kinds of data and extract the required information. New tools and methodologies are developing to extract relevant information.  
   **(i) Complex data types:** The database can include complex data elements, objects with graphical data, spatial data, and temporal data. Mining all these kinds of data is not practical to be done one device.  
   **(ii) Mining from Varied Sources:**The data is gathered from different sources on Network. The data source may be of different kinds depending on how they are stored such as structured, semi-structured or unstructured.
2. **Performance:**  
   The performance of the data mining system depends on the efficiency of algorithms and techniques are using. The algorithms and techniques designed are not up to the mark lead to affect the performance of the data mining process.  
   **(i) Efficiency and Scalability of the Algorithms:** The data mining algorithm must be efficient and scalable to extract information from huge amounts of data in the database.  
   **(ii) Improvement of Mining Algorithms:** Factors such as the enormous size of the database, the entire data flow and the difficulty of data mining approaches inspire the creation of parallel & distributed data mining algorithms.
3. What is Machine learning?

* Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.
* IBM has a rich [history](https://www.ibm.com/ibm/history/ibm100/us/en/icons/ibm700series/impacts/) with machine learning. One of its own, Arthur Samuel, is credited for coining the term, “machine learning” with his [research](https://hci.iwr.uni-heidelberg.de/system/files/private/downloads/636026949/report_frank_gabel.pdf) (PDF, 481 KB) (link resides outside IBM) around the game of checkers. Robert Nealey, the self-proclaimed checkers master, played the game on an IBM 7094 computer in 1962, and he lost to the computer. Compared to what can be done today, this feat seems trivial, but it’s considered a major milestone in the field of artificial intelligence.
* Over the last couple of decades, the technological advances in storage and processing power have enabled some innovative products based on machine learning, such as Netflix’s recommendation engine and self-driving cars.
* Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase. They will be required to help identify the most relevant business questions and the data to answer them.
* Machine learning algorithms are typically created using frameworks that accelerate solution development, such as TensorFlow and PyTorch.

1. What is deep learning?

* Deep learning is a subset of [machine learning](https://www.ibm.com/cloud/learn/machine-learning), which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to “learn” from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.
* Deep learning drives many [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) applications and services that improve automation, performing analytical and physical tasks without human intervention. Deep learning technology lies behind everyday products and services (such as digital assistants, voice-enabled TV remotes, and credit card fraud detection) as well as emerging technologies (such as self-driving cars).

1. What are the data mining tasks/algorithms?

[Data Mining](https://www.geeksforgeeks.org/data-mining/) functions are used to define the trends or correlations contained in data mining activities.

In comparison, data mining **activities** can be divided into 2 categories:

1. **Descriptive Data Mining:**  
   It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set.  
   For examples: count, average etc.
2. **Predictive Data Mining:**  
   It helps developers to provide unlabeled definitions of attributes. Based on previous tests, the software estimates the characteristics that are absent.  
   For example: Judging from the findings of a patient’s medical examinations that is he suffering from any particular disease.

**Data Mining Functionality:**

**1. Class/Concept Descriptions:**  
Classes or definitions can be correlated with results. In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts.  
These class or concept definitions are referred to as class/concept descriptions.

* **Data Characterization:**  
  This refers to the summary of general characteristics or features of the class that is under the study. For example. To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these type of data related to such products by running SQL queries.
* **Data Discrimination:**  
  It compares common features of class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts.

**2. Mining Frequent Patterns, Associations, and Correlations:**  
Frequent patterns are nothing but things that are found to be most common in the data.

There are different kinds of frequency that can be observed in the dataset.

* **Frequent item set:**  
  This applies to a number of items that can be seen together regularly for eg: milk and sugar.
* **Frequent Subsequence:**  
  This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.
* **Frequent Substructure:**  
  It refers to the different kinds of data structures such as trees and graphs that may be combined with the itemset or subsequence.

**Association Analysis:**  
The process involves uncovering the relationship between data and deciding the rules of the association. It is a way of discovering the relationship between various items. for example, it can be used to determine the sales of items that are frequently purchased together.

**Correlation Analysis:**  
Correlation is a mathematical technique that can show whether and how strongly the pairs of attributes are related to each other. For example, Highted people tend to have more weight.

1. What is the difference between Supervised and unsupervised? ( examples of algorithms for each )

**Supervised learning:** Supervised learning is the learning of the model where with input variable ( say, x) and an output variable (say, Y) and an algorithm to map the input to the output.  
That is, **Y = f(X)**

**Why supervised learning?**  
The basic aim is to approximate the mapping function(mentioned above) so well that when there is a new input data (x) then the corresponding output variable can be predicted.

It is called supervised learning because the process of an learning(from the training dataset) can be thought of as a teacher who is supervising the entire learning process. Thus, the “learning algorithm” iteratively makes predictions on the training data and is corrected by the “teacher”, and the learning stops when the algorithm achieves an acceptable level of performance(or the desired accuracy).

**Example of Supervised Learning**  
Suppose there is a basket which is filled with some fresh fruits, the task is to arrange the same type of fruits at one place.  
Also, suppose that the fruits are apple, banana, cherry, grape.

Suppose one already knows from their *previous work* (or experience) that, the shape of each and every fruit present in the basket so, it is easy for them to arrange the same type of fruits in one place.

Here, the previous work is called as **training data** in Data Mining terminology. So, it learns the things from the training data. This is because it has a response variable which says y that if some fruit has so and so features then it is grape, and similarly for each and every fruit.

This type of information is deciphered from the data that is used to train the model.  
This type of learning is called **Supervised Learning**.  
Such problems are listed under classical *Classification Tasks*.  
   
**Unsupervised Learning:** Unsupervised learning is where only the input data (say, X) is present and no corresponding output variable is there.

**Why Unsupervised Learning?**  
The main aim of Unsupervised learning is to model the distribution in the data in order to learn more about the data.

It is called so, because there is no correct answer and there is no such teacher(unlike supervised learning). Algorithms are left to their own devises to discover and present the interesting structure in the data.

**Example of Unsupervised Learning**  
Again, Suppose there is a basket and it is filled with some fresh fruits. The task is to arrange the same type of fruits at one place.

This time there is no information about those fruits beforehand, its the first time that the fruits are being seen or discovered

So how to group similar fruits without any prior knowledge about those.  
First, any physical characteristic of a particular fruit is selected. Suppose *color*.

Then the fruits are arranged on the basis of the color. The groups will be something as shown below:  
**RED COLOR GROUP**: apples & cherry fruits.  
**GREEN COLOR GROUP**: bananas & grapes.

So now, take another physical character say, *size*, so now the groups will be something like this.  
**RED COLOR** AND **BIG SIZE**: apple.  
**RED COLOR** AND **SMALL SIZE**: cherry fruits.  
**GREEN COLOR** AND **BIG SIZE**: bananas.  
**GREEN COLOR** AND **SMALL SIZE**: grapes.  
The job is done!  
Here, there is no need to know or learn anything beforehand. That means, no train data and no response variable. This type of learning is known as **Unsupervised Learning**.

**Difference b/w Supervised and Unsupervised Learning :**

|  |  |  |
| --- | --- | --- |
|  | SUPERVISED LEARNING | UNSUPERVISED LEARNING |
| Input Data | Uses Known and Labeled Data as input | Uses Unknown Data as input |
| Computational Complexity | Very Complex | Less Computational Complexity |
| Real Time | Uses off-line analysis | Uses Real Time Analysis of Data |
| Number of Classes | Number of Classes are known | Number of Classes are not known |
| Accuracy of Results | Accurate and Reliable Results | Moderate Accurate and Reliable Results |

1. What is the difference between Classification and Clustering?

In DBMS

**1. Clustering :**  
Database Clustering is the process of combining more than one servers or instances connecting to a single database. Sometimes one server may not be adequate to manage the amount of data or the number of requests, that is when a Data Cluster is needed.SQL is the language used to manage the database information. Clustering takes different forms, depending on how the data is stored and allocated resources.

**2. Classification of Database :**  
Database management systems can be classified based on several criteria, such as the data model, user numbers and database distribution etc as shown in the below figure.

Diagram

Description automatically generated

**Difference between Classification and Clustering in DBMS :**

|  |  |
| --- | --- |
| CLASSIFICATION | CLUSTERING |
| Its all about predicting the output when input data is given. | It is all about grouping data points together based on similarities among them and difference from others. |
| Labeled data is provided. | Unlabeled data provided. |
| This model function classifies the data into one of defined definite classes. | This function maps the data into one of the multiple clusters where the arrangement of data items is relies on the similarities between them. |
| In classification data are grouped by analyzing data objects whose class label is known. | Clustering analyzes data objects without knowing class label. |
| There is some prior knowledge of attributes of each classification. | There is no prior knowledge of attributes of data to form clusters. |
| It is done by classifying output based on the value of input data. | It is done by grouping only the input data because output is not predefined. |
| The number of class are known before classification as there is predefined output based input data. | The number of clusters are not known before clustering.These are identified after completion of clustering. |
| It is considered as the supervised learning because class labels are known before. | It is considered as unsupervised learning because their is no prior knowledge of class label. |

IN ML:

As you have read the articles about classification and clustering, here is the difference between them.

Both Classification and Clustering is used for the categorization of objects into one or more classes based on the features. They appear to be a similar process as the basic difference is minute. In the case of Classification, there are predefined labels assigned to each input instance according to their properties whereas in clustering those labels are missing.

Diagram

Description automatically generated

**Comparison between Classification and Clustering:**

| **Parameter** | **CLASSIFICATION** | **CLUSTERING** |
| --- | --- | --- |
| Type | used for supervised learning | used for unsupervised learning |
| Basic | process of classifying the input instances based on their corresponding class labels | grouping the instances based on their similarity without the help of class labels |
| Need | it has labels so there is need of training and testing dataset for verifying the model created | there is no need of training and testing dataset |
| Complexity | more complex as compared to clustering | less complex as compared to classification |
| Example Algorithms | Logistic regression, Naive Bayes classifier, Support vector machines, etc. | k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering algorithm, etc. |

**Differences between Classification and Clustering**

1. Classification is used for supervised learning whereas clustering is used for unsupervised learning.
2. The process of classifying the input instances based on their corresponding class labels is known as classification whereas grouping the instances based on their similarity without the help of class labels is known as clustering.
3. As Classification have labels so there is need of training and testing dataset for verifying the model created but there is no need for training and testing dataset in clustering.
4. Classification is more complex as compared to clustering as there are many levels in the classification phase whereas only grouping is done in clustering.
5. Classification examples are Logistic regression, Naive Bayes classifier, Support vector machines, etc. Whereas clustering examples are k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering algorithm, etc.
6. Examples of clustering algorithms

**Clustering Algorithms :**   
[K-means clustering algorithm](https://www.geeksforgeeks.org/k-means-clustering-introduction/) – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.

Chart, scatter chart

Description automatically generated

**Applications of Clustering in different fields**

* **Marketing:** It can be used to characterize & discover customer segments for marketing purposes.
* **Biology:** It can be used for classification among different species of plants and animals.
* **Libraries:** It is used in clustering different books on the basis of topics and information.
* **Insurance:** It is used to acknowledge the customers, their policies and identifying the frauds.

**City Planning:** It is used to make groups of houses and to study their values based on their geographical locations and other factors present.

**Earthquake studies:** By learning the earthquake-affected areas we can determine the dangerous zones.

1. Examples for classification algorithms

### Popular Classification Algorithms:

* [Logistic Regression](https://monkeylearn.com/blog/classification-algorithms/#logistic-regression)
* [Naive Bayes](https://monkeylearn.com/blog/classification-algorithms/#naive-bayes)
* [K-Nearest Neighbors](https://monkeylearn.com/blog/classification-algorithms/#knn)
* [Decision Tree](https://monkeylearn.com/blog/classification-algorithms/#decision-tree)
* [Support Vector Machines](https://monkeylearn.com/blog/classification-algorithms/#svm)

### Logistic Regression

Logistic regression is a calculation used to predict a binary outcome: either something happens, or does not. This can be exhibited as Yes/No, Pass/Fail, Alive/Dead, etc.

Independent variables are analyzed to determine the binary outcome with the results falling into one of two categories. The independent variables can be categorical or numeric, but the dependent variable is always categorical. Written like this:

**P(Y=1|X) or P(Y=0|X)**

It calculates the probability of dependent variable Y, given independent variable X.

This can be used to calculate the probability of a word having a positive or negative connotation (0, 1, or on a scale between). Or it can be used to determine the object contained in a photo (tree, flower, grass, etc.), with each object given a probability between 0 and 1.

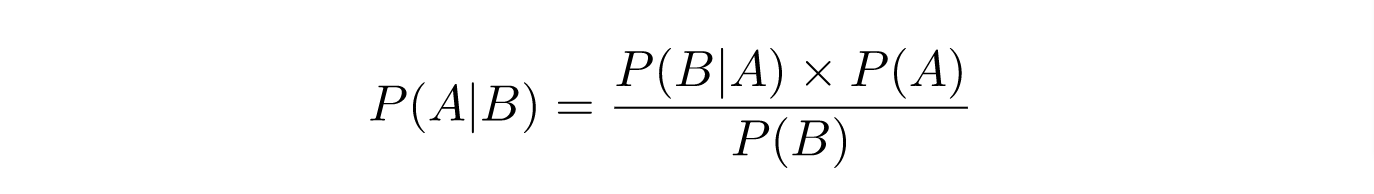
### Naive Bayes

[Naive Bayes](https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/) calculates the possibility of whether a data point belongs within a certain category or does not. In [text analysis](https://monkeylearn.com/text-analysis/), it can be used to categorize words or phrases as belonging to a preset “tag” (classification) or not. For example:

Table

Description automatically generated

To decide whether or not a phrase should be tagged as “sports,” you need to calculate:



Or… the probability of A, if B is true, is equal to the probability of B, if A is true, times the probability of A being true, divided by the probability of B being true.

### K-nearest Neighbors

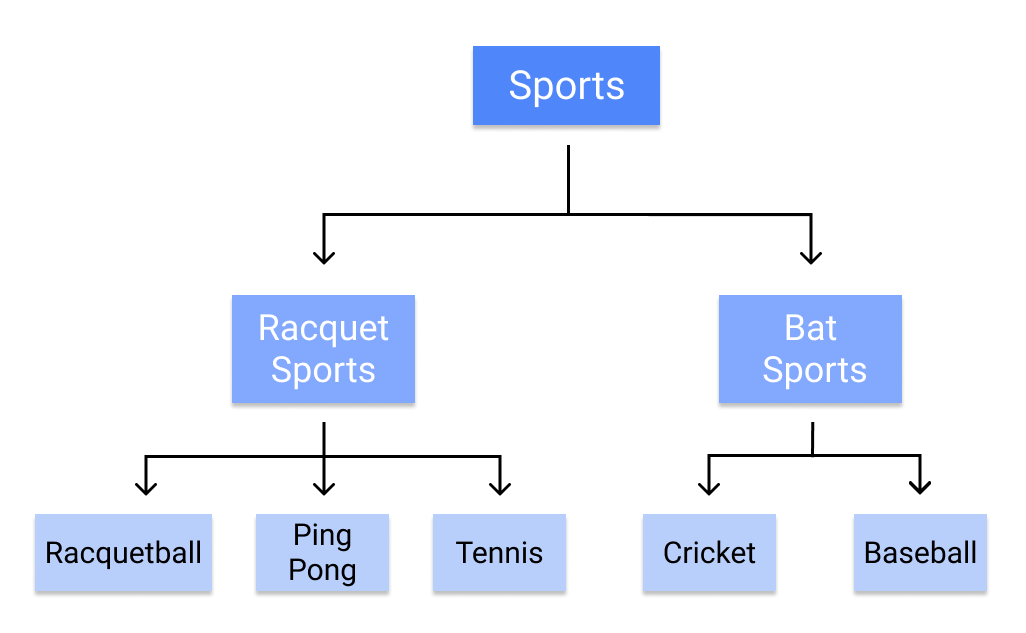
K-nearest neighbors (k-NN) is a pattern recognition algorithm that uses training datasets to find the k closest relatives in future examples.

When k-NN is used in classification, you calculate to place data within the category of its nearest neighbor. If k = 1, then it would be placed in the class nearest 1. K is classified by a plurality poll of its neighbors.

### Decision Tree

A decision tree is a supervised learning algorithm that is perfect for classification problems, as it’s able to order classes on a precise level. It works like a flow chart, separating data points into two similar categories at a time from the “tree trunk” to “branches,” to “leaves,” where the categories become more finitely similar. This creates categories within categories, allowing for organic classification with limited human supervision.

To continue with the sports example, this is how the decision tree works:



#### Random Forest

The random forest algorithm is an expansion of decision tree, in that you first construct a multitude of decision trees with training data, then fit your new data within one of the trees as a “random forest.”

It, essentially, averages your data to connect it to the nearest tree on the data scale. Random forest models are helpful as they remedy for the decision tree’s problem of “forcing” data points within a category unnecessarily.

### Support Vector Machines

A [support vector machine (SVM)](https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/) uses algorithms to train and classify data within degrees of polarity, taking it to a degree beyond X/Y prediction.

For a simple visual explanation, we’ll use two tags: red and blue, with two data features: X and Y, then train our classifier to output an X/Y coordinate as either red or blue.

Chart, scatter chart

Description automatically generated

The SVM then assigns a hyperplane that best separates the tags. In two dimensions this is simply a line. Anything on one side of the line is red and anything on the other side is blue. In sentiment analysis, for example, this would be positive and negative.

In order to maximize machine learning, the best hyperplane is the one with the largest distance between each tag:

Chart

Description automatically generated

However, as data sets become more complex, it may not be possible to draw a single line to classify the data into two camps:

Chart, scatter chart

Description automatically generated

Using SVM, the more complex the data, the more accurate the predictor will become. Imagine the above in three dimensions, with a Z-axis added, so it becomes a circle.

Mapped back to two dimensions with the best hyperplane, it looks like this:

Diagram, schematic

Description automatically generated

SVM allows for more accurate machine learning because it’s multidimensional.

1. What is association rule?

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

| **TID** | **Items** |
| --- | --- |
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

Before we start defining the rule, let us first see the basic definitions.

**Support Count() –** Frequency of occurrence of a itemset.

Here ({Milk, Bread, Diaper})=2

**Frequent Itemset –** An itemset whose support is greater than or equal to minsup threshold.

**Association Rule –** An implication expression of the form X -> Y, where X and Y are any 2 itemsets.

Example: {Milk, Diaper}->{Beer}

**Rule Evaluation Metrics –**

* **Support(s) –**  
  The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction.It is a measure of how frequently the collection of items occur together as a percentage of all transactions.
* **Support = (X+Y) total –**  
  It is interpreted as fraction of transactions that contain both X and Y.
* **Confidence(c) –**  
  It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
* **Conf(X=>Y) = Supp(XY) Supp(X) –**  
  It measures how often each item in Y appears in transactions that contains items in X also.
* **Lift(l) –**  
  The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the itemsets X and Y are independent of each other.The expected confidence is the confidence divided by the frequency of {Y}.
* **Lift(X=>Y) = Conf(X=>Y) Supp(Y) –**  
  Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected.Greater lift values indicate stronger association.

**Example –** From the above table, {Milk, Diaper}=>{Beer}

s= ({Milk, Diaper, Beer}) |T|

= 2/5

= 0.4

c= (Milk, Diaper, Beer) (Milk, Diaper)

= 2/3

= 0.67

l= Supp({Milk, Diaper, Beer}) Supp({Milk, Diaper})\*Supp({Beer})

= 0.4/(0.6\*0.6)

= 1.11

The Association rule is very useful in analyzing datasets. The data is collected using bar-code scanners in supermarkets. Such databases consists of a large number of transaction records which list all items bought by a customer on a single purchase. So the manager could know if certain groups of items are consistently purchased together and use this data for adjusting store layouts, cross-selling, promotions based on statistics.

1. How does this algorithm work (K-Mean, Regression, SVM, association rule, decision tree, KNN...)?

Q 14

1. What is recall and precision, F1?

[Information Systems](https://www.geeksforgeeks.org/types-of-information-system/) can be measured with two metrics: precision and recall. When a user decides to search for information on a topic, the total database and the results to be obtained can be divided into 4 categories:

1. Relevant and Retrieved
2. Relevant and Not Retrieved
3. Non-Relevant and Retrieved
4. Non-Relevant and Not Retrieved

Relevant items are those documents that help the user in answering his question.Non-Relevant items are items that don’t provide actually useful information. For each item there are two possibilities it can be retrieved or not retrieved by the user’s query. Precision is defined as the ratio of the number of relevant and retrieved documents(number of items retrieved that are actually useful to the user and match his search need) to the number of total retrieved documents from the query. Recall is defined as ratio of the number of retrieved and relevant documents(the number of items retrieved that are relevant to the user and match his needs) to the number of possible relevant documents(number of relevant documents in the database).Precision measures one aspect of information retrieval overhead for a user associated with a particular search. If a search has 85 percent precision then 15(100-85) percent of user effort is overhead reviewing non-relevant items. Recall measures to what extent a system processing a particular query is able to retrieve the relevant items the user is interested in seeing. Recall is a very useful concept but due to the denominator is non-calculable in operational systems. If the system is made known the total set of relevant items in the database, recall can be made calculable.

In this article, we will be looking at the approach to calculate F1 Score using the various packages and their various functionalities in the R language.

## **F1 Score**

The F-score or F-measure is a measure of a test’s accuracy. It is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

Graphical user interface, text, application, chat or text message

Description automatically generated

## Method 1: Using F1\_Score function from Mlmetrics package

Under this approach to calculate the f1 score, the user needs to install and import the Mlmetrics package in the current working R console and further, the user needs to call the F1\_Score() function from this package and pass it with the required parameter to get the F1 score of the predicted and the actual value and further in return this function will be returning the F1 score of the given actual and the predicted values.

**Syntax to install and import the Mlmetrics package in R language:**

install.package("MLmetrics")

library("MLmetrics")

**F1\_Score() function:** This function is used to calculate the F1 score.

**Syntax:** F1\_Score(y\_true, y\_pred, positive = NULL)

**Parameters:**

* y\_true: Ground truth (correct) 0-1 labels vector
* y\_pred: Predicted labels vector, as returned by a classifier
* positive: An optional character string for the factor level that corresponds to a “positive” result

**Example:** In this example, we are creating two vectors of 10 data points one with the actual values and another with the predicted values and with the help of the F1\_Score() function from the MLmetrics package we are calculating the f1 score in the R programming.

|  |
| --- |
| # Import Mlmetrics library  library(MLmetrics)    # Create Data  actual = c(1,2,28,1,5,6,7,8,9,10)  predicted = c(1,2,3,4,5,6,7,8,9,10)    # Calculate F!\_Score  F1\_Score(predicted,actual) |

**Output:**

[1] 0.6666667

## Method 2: Using confusionMatrix() function from caret package

In this approach to calculate the F1 score, the user needs to first install and import the caret package in the working R console, and then further the user needs to call the confusionMatrix() function and pass the required parameter into it. This will be returning the F1 score back to the user of the given data in the R language.

**Syntax to install and import the caret package in R language:**

install.package("caret")

library("caret")

**confusionMatrix() function:** Calculates a cross-tabulation of observed and predicted classes with associated statistics.

**Syntax:** confusionMatrix(data, reference, positive = NULL, dnn = c(“Prediction”, “Reference”), …)

**Parameters:**

* data: a factor of predicted classes
* reference: a factor of classes to be used as the true results
* positive: an optional character string for the factor level that corresponds to a “positive” result (if that makes sense for your data).
* dnn: a character vector of dimnames for the table
* …: options to be passed.

**Example:** In this example, we are two vectors, one with the actual data and another with the predicted data, and further, we are using the confusionMatrix() function to get the F1 score of the given data.

|  |
| --- |
| # Import caret library  library(caret)    # Create Data  actual <- factor(rep(c(1, 2),                       times=c(16, 24)))  predicted <- factor(rep(c(1, 2, 1, 2),                          times=c(12, 4, 7, 17)))    # create confusion matrix  confusionMatrix(predicted, actual,                  mode = "everything",                  positive="1") |

**Output:**

Confusion Matrix and Statistics

Reference

Prediction 1 2

1 12 7

2 4 17

Accuracy : 0.725

95% CI : (0.5611, 0.854)

No Information Rate : 0.6

P-Value [Acc > NIR] : 0.07095

Kappa : 0.4444

Mcnemar's Test P-Value : 0.54649

Sensitivity : 0.7500

Specificity : 0.7083

Pos Pred Value : 0.6316

Neg Pred Value : 0.8095

Precision : 0.6316

Recall : 0.7500

F1 : 0.6857

Prevalence : 0.4000

Detection Rate : 0.3000

Detection Prevalence : 0.4750

Balanced Accuracy : 0.7292

'Positive' Class : 1

## Method 3: Calculate F1 score of the model:

In this method to calculate the F1 score of the model, the user needs to first create the model regarding the given data then the user needs to calculate the confusion matric of that model, further the err\_metric() function with the confusion matrix pass as its parameter to the f1 score of the built model in the R programming language.

**Syntax:** err\_metric(cm)

**Where,** cm: confusion matrix

**Example:** In this example, we will be simply creating a model of logistic regression of the given data set and then using the err\_metrics() function to calculate the f1 score in the R programming language.

The [link](https://drive.google.com/file/d/1Y-5CAsPoAQ0C1RmMS-OtoLHNyKy259rF/view?usp=sharing) of the dataset.

|  |
| --- |
| library(caTools)  data = read.csv('Social\_Network\_Ads.csv')  data = data[3:5]  split = sample.split(data$Purchased, SplitRatio = 0.75)  train = subset (data, split == TRUE)  test = subset (data, split == FALSE)  train[-3] = scale(train[-3])  test[-3] = scale(test[-3])  classifier = glm(formula = Purchased ~ .,                   family = binomial,                   data = train)  prob\_pred = predict (classifier, type = 'response',                       newdata = test[-3])  y\_pred = ifelse (prob\_pred > 0.5, 1, 0)  cm = table (test[, 3], y\_pred > 0.5)  err\_metric(cm) |

**Output:**

[1] "Precision value of the model: 0.72"

[1] "Accuracy of the model: 0.77"

[1] "Recall value of the model: 0.12"

[1] "False Positive rate of the model: 0.12"

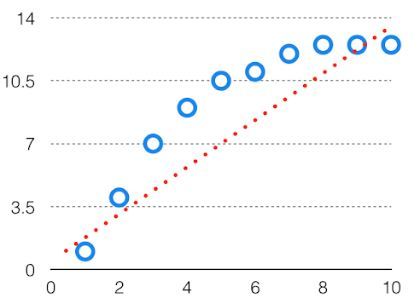
[1] "False Negative rate of the model: 0.42"

[1] "f1 score of the model: 0.21"

1. What is the bias, variance trade-off?

It is important to understand prediction errors (bias and variance) when it comes to accuracy in any machine learning algorithm. There is a tradeoff between a model’s ability to minimize bias and variance which is referred to as the best solution for selecting a value of **Regularization** constant. Proper understanding of these errors would help to avoid the overfitting and underfitting of a data set while training the algorithm.

**Bias**  
The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.  
By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set. Such fitting is known as **Underfitting of Data**. This happens when the hypothesis is too simple or linear in nature. Refer to the graph given below for an example of such a situation.

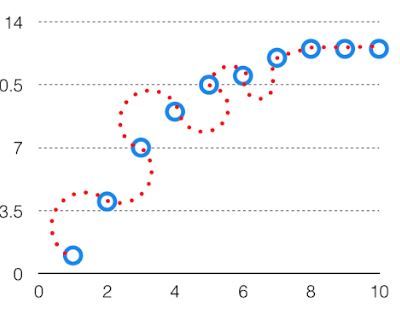


HighBias

In such a problem, a hypothesis looks like follows.  
Text

Description automatically generated with medium confidence  
**Variance**  
The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model. The model with high variance has a very complex fit to the training data and thus is not able to fit accurately on the data which it hasn’t seen before. As a result, such models perform very well on training data but has high error rates on test data.  
When a model is high on variance, it is then said to as **Overfitting of Data**. Overfitting is fitting the training set accurately via complex curve and high order hypothesis but is not the solution as the error with unseen data is high.  
While training a data model variance should be kept low.

The high variance data looks like follows.



High Variance

In such a problem, a hypothesis looks like follows.  
Calendar

Description automatically generated with low confidence  
**Bias Variance Tradeoff**

If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex ( hypothesis with high degree eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time. For the graph, the perfect tradeoff will be like.

Chart, scatter chart

Description automatically generated  
The best fit will be given by hypothesis on the tradeoff point.

The error to complexity graph to show trade-off is given as –  
A picture containing diagram

Description automatically generated  
This is referred to as the best point chosen for the training of the algorithm which gives low error in training as well as testing data.

1. What is the confusion matrix?

Classification is the process of categorizing a given set of data into classes.  
In Machine Learning(ML), you frame the problem, collect and clean the data, add some necessary feature variables(if any), train the model, measure its performance, improve it by using some cost function, and then it is ready to deploy.   
But how do we measure its performance? Is there any particular feature to look at?  
A trivial and broad answer would be to compare the actual values to the predicted values. But that does not solve the issue.   
Let us consider the famous MNIST dataset and try to analyze the problem.

|  |
| --- |
| # Importing the dataset.  from sklearn.datasets import fetch\_openml  mnist = fetch\_openml('mnist\_784', version=1)    # Creating independent and dependent variables.  X, y = mnist['data'], mnist['target']    # Splitting the data into training set and test set.  X\_train, X\_test, y\_train, y\_test = X[:60000], X[60000:], y[:60000], y[60000:]    """  The training set is already shuffled for us, which is good as this guarantees that all  cross-validation folds will be similar.  """    # Training a binary classifier.  y\_train\_5 = (y\_train == 5) # True for all 5s, False for all other digits.  y\_test\_5 = (y\_test == 5)    """  Building a dumb classifier that just classifies every single image in the “not-5” class.  """    from sklearn.model\_selection import cross\_val\_score  from sklearn.base import BaseEstimator  class Never5Classifier(BaseEstimator):      def fit(self, X, y=None):          pass      def predict(self, X):          return np.zeros((len(X), 1), dtype=bool)    never\_5\_clf = Never5Classifier()  cross\_val\_score(never\_5\_clf, X\_train, y\_train\_5, cv=3, scoring="accuracy") |

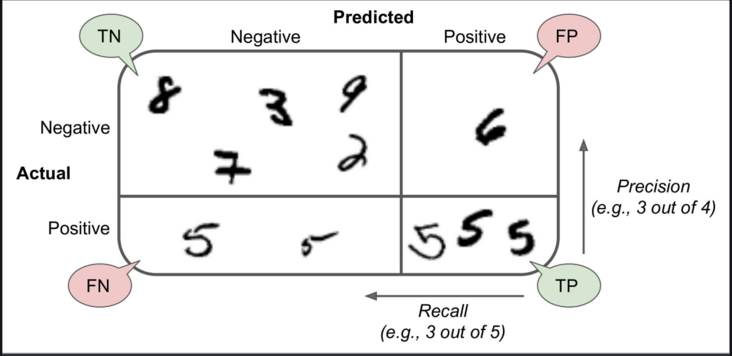
If you executed the same code on an IDE, you would get an array of accuracies each with above 90% accuracy! This is simply because only about 10% of the images are 5s, so if you always guess that an image is not a 5, you will be right about 90% of the time.  
This demonstrates why accuracy is generally not the preferred performance measure for classifiers, especially when you are dealing with skewed datasets (i.e., when some classes are much more frequent than others).

**Confusion Matrix**  
A much better way to evaluate the performance of a classifier is to look at the confusion matrix. The general idea is to count the number of times instances of class A are classified as class B. For example, to know the number of times the classifier confused images of 5s with 3s, you would look in the 5th row and 3rd column of the confusion matrix.

|  |
| --- |
| # Creating some predictions.  from sklearn.model\_selection import cross\_val\_predict  y\_train\_pred = cross\_val\_predict(sgd\_clf, X\_train, y\_train\_5, cv=3)    """  You could make predictions on the test set, but use the test set only at the very end of your project, once you have a classifier that you are ready to launch.  """    # Constructing the confusion matrix.  from sklearn.metrics import confusion\_matrix  confusion\_matrix(y\_train\_5, y\_train\_pred) |

Each row in a confusion matrix represents an actual class, while each column represents a predicted class. For more info about the confusion, matrix clicks [here.](https://github.com/Nitin1901/Confusion-Matrix)  
The confusion matrix gives you a lot of information, but sometimes you may prefer a more concise metric.

* **Precision**  
  precision = (TP) / (TP+FP)  
  TP is the number of true positives, and FP is the number of false positives.   
  A trivial way to have perfect precision is to make one single positive prediction and ensure it is correct (precision = 1/1 = 100%). This would not be very useful since the classifier would ignore all but one positive instance.
* **Recall**  
  recall = (TP) / (TP+FN)



|  |
| --- |
| # Finding precision and recall  from sklearn.metrics import precision\_score, recall\_score  precision\_score(y\_train\_5, y\_train\_pred)  recall\_score(y\_train\_5, y\_train\_pred) |

Now your 5-detector does not look as shiny as it did when you looked at its accuracy. When it claims an image represents a 5, it is correct only 72.9% (precision) of the time. Moreover, it only detects 75.6% (recall) of the 5s.   
It is often convenient to combine precision and recall into a single metric called the F1 score, in particular, if you need a simple way to compare two classifiers.   
The F1 score is the harmonic mean of precision and recall 

|  |
| --- |
| # To compute the F1 score, simply call the f1\_score() function:  from sklearn.metrics import f1\_score  f1\_score(y\_train\_5, y\_train\_pred) |

The F1 score favors classifiers that have similar precision and recall.   
This is not always what you want: in some contexts, you mostly care about precision, and in other contexts, you really care about the recall. For example, if you trained a classifier to detect videos that are safe for kids, you would probably prefer a classifier that rejects many good videos (low recall) but keeps only safe ones (high precision), rather than a classifier that has a much higher recall but lets a few terrible videos show up in your product (in such cases, you may even want to add a human pipeline to check the classifier’s video selection). On the other hand, suppose you train a classifier to detect shoplifters on surveillance images: it is probably fine if your classifier has only 30% precision as long as it has 99% recall (sure, the security guards will get a few false alerts, but almost all shoplifters will get caught).  
Unfortunately, you can’t have it both ways: increasing precision reduces recall and vice versa. This is called the *precision/recall tradeoff*.

1. What is the ROC Curve?

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

* True Positive Rate
* False Positive Rate

**True Positive Rate** (**TPR**) is a synonym for recall and is therefore defined as follows:

**False Positive Rate** (**FPR**) is defined as follows:

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

**Figure 4. TP vs. FP rate at different classification thresholds.**

To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient. Fortunately, there's an efficient, sorting-based algorithm that can provide this information for us, called AUC.

## AUC: Area Under the ROC Curve

**AUC** stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

**Figure 5. AUC (Area under the ROC Curve).**

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. For example, given the following examples, which are arranged from left to right in ascending order of logistic regression predictions:

**Figure 6. Predictions ranked in ascending order of logistic regression score.**

AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example.

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

AUC is desirable for the following two reasons:

* AUC is **scale-invariant**. It measures how well predictions are ranked, rather than their absolute values.
* AUC is **classification-threshold-invariant**. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

However, both these reasons come with caveats, which may limit the usefulness of AUC in certain use cases:

* **Scale invariance is not always desirable.** For example, sometimes we really do need well calibrated probability outputs, and AUC won’t tell us about that.
* **Classification-threshold invariance is not always desirable.** In cases where there are wide disparities in the cost of false negatives vs. false positives, it may be critical to minimize one type of classification error. For example, when doing email spam detection, you likely want to prioritize minimizing false positives (even if that results in a significant increase of false negatives). AUC isn't a useful metric for this type of optimization.

1. Explain cross-validation?

In machine learning, we couldn’t fit the model on the training data and can’t say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique.

**Cross-Validation**

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

**Methods of Cross Validation**

**Validation**  
In this method, we perform training on the 50% of the given data-set and rest 50% is used for the testing purpose. The major drawback of this method is that we perform training on the 50% of the dataset, it may possible that the remaining 50% of the data contains some important information which we are leaving while training our model i.e higher bias.

**LOOCV (Leave One Out Cross Validation)**  
In this method, we perform training on the whole data-set but leaves only one data-point of the available data-set and then iterates for each data-point. It has some advantages as well as disadvantages also.  
An advantage of using this method is that we make use of all data points and hence it is low bias.  
The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point. If the data point is an outlier it can lead to higher variation. Another drawback is it takes a lot of execution time as it iterates over ‘the number of data points’ times.

**K-Fold Cross Validation**  
In this method, we split the data-set into k number of subsets(known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

***Note:***

It is always suggested that the value of k should be 10 as the lower value

of k is takes towards validation and higher value of k leads to LOOCV method.

**Example**  
The diagram below shows an example of the training subsets and evaluation subsets generated in k-fold cross-validation. Here, we have total 25 instances. In first iteration we use the first 20 percent of data for evaluation, and the remaining 80 percent for training([1-5] testing and [5-25] training) while in the second iteration we use the second subset of 20 percent for evaluation, and the remaining three subsets of the data for training([5-10] testing and [1-5 and 10-25] training), and so on.

Chart

Description automatically generated

Total instances: 25

Value of k : 5

No. Iteration Training set observations Testing set observations

1 [ 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24] [0 1 2 3 4]

2 [ 0 1 2 3 4 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24] [5 6 7 8 9]

3 [ 0 1 2 3 4 5 6 7 8 9 15 16 17 18 19 20 21 22 23 24] [10 11 12 13 14]

4 [ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 20 21 22 23 24] [15 16 17 18 19]

5 [ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19] [20 21 22 23 24]

**Comparison of train/test split to cross-validation**

Advantages of train/test split:

1. This runs K times faster than Leave One Out cross-validation because K-fold cross-validation repeats the train/test split K-times.
2. Simpler to examine the detailed results of the testing process.

Advantages of cross-validation:

1. More accurate estimate of out-of-sample accuracy.
2. More “efficient” use of data as every observation is used for both training and testing.

Python code for k fold cross-validation.

|  |
| --- |
| # This code may not be run on GFG IDE  # as required packages are not found.    # importing cross-validation from sklearn package.  from sklearn import cross\_validation    # value of K is 10.  data = cross\_validation.KFold(len(train\_set), n\_folds=10, indices=False) |

1. What is the difference between a validation set and a test set?

The fundamental purpose for splitting the dataset is to assess how effective will the trained model be in generalizing to new data. This split can be achieved by using **train\_test\_split** function of [scikit-learn.](https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/)

## **Training Set**

This is the actual dataset from which a model trains .i.e. the model sees and learns from this data to predict the outcome or to make the right decisions. Most of the training data is collected from several resources and then preprocessed and organized to provide proper performance of the model. Type of training data hugely determines the ability of the model to generalize .i.e. the better the quality and diversity of training data, the better will be the performance of the model. This data is more than 60% of the total data available for the project.

**Example:**

|  |
| --- |
| # Importing numpy & scikit-learn  import numpy as np  from sklearn.model\_selection import train\_test\_split    # Making a dummy array to  # represent x,y for example  # Making a array for x ranging  # from 0-15 then reshaping it  # to form a matrix of shape 8x2  x = np.arange(16).reshape((8,2))    # y is just a list of 0-7 number  # representing target variable  y = range(8)    # Splitting dataset in 80-20 fashion .i.e.  # Testing set is 20% of total data  # Training set is 80% of total data  x\_train, x\_test, y\_train, y\_test = train\_test\_split(x,y,                                                      train\_size=0.8,                                                      random\_state=42)    # Training set  print("Training set x: ",x\_train)  print("Training set y: ",y\_train) |

**Output:**

Training set x: [[ 0 1]

[14 15]

[ 4 5]

[ 8 9]

[ 6 7]

[12 13]]

Training set y: [0, 7, 2, 4, 3, 6]

**Explanation:**

* Firstly we created a dummy matrix of 8×2 shape using NumPy library to represent input x. And a list of 0 to 7 integers representing our target variable y.
* Now in order to split our dataset into training and testing data, a function named **train\_test\_split** of sklearn library is used.
* Input data x with target variable y is passed as parameters to function which then divides the dataset into 2 parts on the size given in train\_size i.e. if train\_size=0.8 is given then the dataset will be divided in such an way that the training set will be 80% of given dataset and testing set will be 20% of given dataset.
* And as we specify random\_state to be a positive number, train\_test\_split function will randomly split data.

## **Testing Set**

This dataset is independent of the training set but has a somewhat similar type of probability distribution of classes and is used as a benchmark to evaluate the model, used only after the training of the model is complete. Testing set is usually a properly organized dataset having all kinds of data for scenarios that the model would probably be facing when used in the real world. Often the validation and testing set combined is used as a testing set which is not considered a good practice. If the accuracy of the model on training data is greater than that on testing data then the model is said to have overfitting. This data is approximately 20-25% of the total data available for the project.

**Example:**

|  |
| --- |
| # Importing numpy & scikit-learn  import numpy as np  from sklearn.model\_selection import train\_test\_split    # Making a dummy array to represent x,y for example  # Making a array for x ranging from 0-15 then  # reshaping it to form a matrix of shape 8x2  x = np.arange(16).reshape((8, 2))    # y is just a list of 0-7 number representing  # target variable  y = range(8)    # Splitting dataset in 80-20 fashion .i.e.  # Training set is 80% of total data  # Testing set is 20% of total data  x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y,                                                      test\_size=0.2,                                                      random\_state=42)    # Testing set  print("Testing set x: ", x\_test)  print("Testing set y: ", y\_test) |

**Output:**

Testing set x: [[ 2 3]

[10 11]]

Testing set y: [1, 5]

**Explanation:**

* To show how the train\_test\_split function works we first created a dummy matrix of 8×2 shape using NumPy library to represent input x. And a list of 0 to 7 integers representing our target variable y.
* Now in order to split our dataset into training and testing data, input data x with target variable y is passed as parameters to function which then divides the dataset into 2 parts on the size given in test\_size i.e. if test\_size=0.2 is given then the dataset will be divided in such an away that testing set will be 20% of given dataset and training set will be 80% of given dataset.
* And as we specify random\_state to be a positive number, train\_test\_split function will randomly split data.

## **Validation Set**

The validation set is used to fine-tune the hyperparameters of the model and is considered a part of the training of the model. The model only sees this data for evaluation but does not learn from this data, providing an objective unbiased evaluation of the model. Validation dataset can be utilized for regression as well by interrupting training of model when loss of validation dataset becomes greater than loss of training dataset .i.e. reducing bias and variance. This data is approximately 10-15% of the total data available for the project but this can change depending upon the number of hyperparameters .i.e. if model has quite many hyperparameters then using large validation set will give better results. Now, whenever the accuracy of model on validation data is greater than that on training data then the model is said to have generalized well.

**Example:**

|  |
| --- |
| # Importing numpy & scikit-learn  import numpy as np  from sklearn.model\_selection import train\_test\_split    # Making a dummy array to represent x,y for example  # Making a array for x ranging from 0-23 then reshaping it  # to form a matrix of shape 8x3  x = np.arange(24).reshape((8,3))    # y is just a list of 0-7 number representing  # target variable  y = range(8)    # Splitting dataset in 80-20 fashion .i.e.  # Training set is 80% of total data  # Combined set of testing & validation is  # 20% of total data  x\_train, x\_Combine, y\_train, y\_Combine = train\_test\_split(x,y,                                                train\_size=0.8,                                                random\_state=42)    # Splitting combined dataset in 50-50 fashion .i.e.  # Testing set is 50% of combined dataset  # Validation set is 50% of combined dataset  x\_val, x\_test, y\_val, y\_test = train\_test\_split(x\_Combine,                                                  y\_Combine,                                                  test\_size=0.5,                                                  random\_state=42)    # Training set  print("Training set x: ",x\_train)  print("Training set y: ",y\_train)  print("  ")    # Testing set  print("Testing set x: ",x\_test)  print("Testing set y: ",y\_test)  print("  ")    # Validation set  print("Validation set x: ",x\_val)  print("Validation set y: ",y\_val) |

**Output:**

Training set x: [[ 0 1 2]

[21 22 23]

[ 6 7 8]

[12 13 14]

[ 9 10 11]

[18 19 20]]

Training set y: [0, 7, 2, 4, 3, 6]

Testing set x: [[15 16 17]]

Testing set y: [5]

Validation set x: [[3 4 5]]

Validation set y: [1]

**Explanation:**

* So as to get the validation set, a dummy matrix of 8×3 shape is created using the NumPy library to represent input x. And a list of 0 to 7 integers representing our target variable y.
* Now it gets a bit tricky to divide dataset into 3 parts. To begin with, the dataset is divided into two parts, input data x with target variable y is passed as parameters to function which then divides the dataset into 2 parts on the size given in train\_size (from this we’ll get our training set) i.e. if train\_size=0.8 is given then the dataset will be divided in such a way that training set will be 80% of given dataset and another set will be 20% of given dataset.
* So now we have validation and testing combined set having 20% of the initially given dataset. This dataset is divided further to get validation set and testing set, output of above distribution is then passed as parameters to train\_test\_split again which then divides the combined dataset into 2 parts on the size given in test\_size .i.e. if test\_size=0.5 is given then the dataset will be divided in such a way that testing set and validation set will be 50% of the combined dataset.

1. How do you treat missing/outlier values?

### Missing Attributes

When the sample contains missing attributes, there is unfortunately no miracle cure! However, there are several possible approaches you can take.

For a given variable (for instance, last chapter’s example of Date of Birth), if the proportion of missing attributes is low, you can just forget about them and do nothing: leave the sample intact. You will then be working with a data set that has “holes,” like a Swiss cheese. Depending on the statistical process you plan to apply, this solution might be acceptable, or it might not.

#### Forget a Variable

However, if for this same variable, the proportion of missing attributes is way too high, you’d better just forget about it—provided that the variable is not too important to your analysis. This is the same as not including a column in a table, as we saw in the last chapter.

#### Forget Individuals

If the variable with the missing data is crucial to your analysis, it’s better to create a sub-sample, removing the individuals for whom this variable is missing. For example, if you are analyzing your bank statements by looking at the amount of money you spend/earn, the “transaction amount” variable will be crucial. If the transaction amount is unknown for some of the rows of your statement, it’s better to create a sub-sample that removes all of the offending rows.

However, this method is risky. You might find yourself with a number of individuals (a number of rows) so small that your analysis no longer meaningful. In addition, your sample might no longer be representative of the overall population. To find out why, go to the Take It Further section at the end of this chapter.

#### Guess

A more adventurous approach consists of filling in your holes with values you have guessed. This is pretty much the method for daredevils! :zorro: Of course, these values will not correspond to actual values, but some methods manage to create values that are not too far off. Guessing a missing attribute is referred to as Imputation.

For example, we can replace the missing attributes of the height variable with the average height of the individuals in our sample. In our example, to correct Hannah’s height (which we assume is erroneous), we would replace it with the average height of the other individuals in the sample, which is 1.52 m. This is known as **Mean Imputation**.

#### Guess Based on Other Variables

But we can do better! To replace a given variable, we can look at the variables around it. A number of methods apply this principle.

Imagine a new individual named Luke, born in 1991, whose height is unknown. Rather than assign him the mean of the entire sample (1.52 m), we can assign him the average height of people who are about his age. So let’s assign him the average height of people born between 1990 and 2000, or 1.49 m. Here, we looked at the value of the date\_of\_birth variable to come up with a value for the height variable.

Other methods also deduce the value of a variable by looking at other variables. These include Hot-deck, and methods based on linear regression.

Guessing (imputing) values changes your sample, because the imputed values are false. In particular, your calculations of variances and correlations will be false. You must therefore use this method only sparingly.

In all cases, you must specify which method you used for each of the analysis results you present. It’s a question of intellectual integrity ^^.

### Outliers

Hannah is 3.45 meters tall. You think that’s not so tall? You’re wrong. It’s very tall compared with the heights of other human beings.

But proceed cautiously, because an outlier value isn’t always necessarily false! Hannah might actually be 3.45 meters tall. Okay, that’s hard to believe—but it’s possible.

An outlier can be:

* An aberration: a value that’s obviously false
* An atypical value: a value that “deviates from the norm,” but is not necessarily false.

Ideally, outliers should be checked to determine whether or not they’re erroneous. For example, a thermometer in Canada in April might indicate 40°C, but this could be due to a defective temperature sensor, or it could be an actual value....(although it usually a little colder in Canada in the spring ;)).

So what should we do with outliers? If we are sure that the value is erroneous (input error or flawed sensor, for example) and we can’t find the actual value, it has to be deleted. If we are not sure whether it’s erroneous, we can choose between:

* Deleting the value. We then find ourselves with a missing attribute, to which we can impute a value, as we saw previously. But imputation isn’t mandatory.
* Keeping the value.

How to choose between these two options? It all depends on how you will be processing your data after that. Some methods are considered “robust,” meaning they are not destabilized by outliers. For example, we will see below that the mean is very sensitive to outliers, while the median is not. So if you want to find a mean, create a sub-sample in which you don’t include outliers. But if you also want to calculate a median, work with the original sample.

1. How do you prepare the data for the ML Model?

It is very rare that you get data exact in the form in which you want. [Data Preprocessing](https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/) is a crucial and very first step before building and deploying your Machine Learning Model. And while building a model it’s not the case that every time you will get clean and formatted data to work on. It is mandatory to clean and check the data before use. So, we use data preprocessing for these. Let’s check out some steps before building the model which we should perform.

1. Getting dataset
2. Importing libraries
3. Import dataset
4. Finding missing values
5. Encoding categorical data
6. Split data in training and testing set
7. Feature scaling

### **1. Getting Dataset**

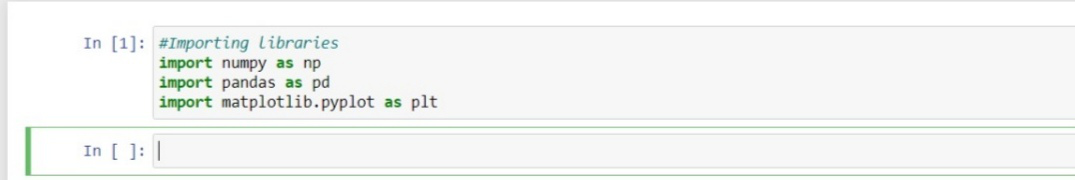
The very first thing we require is a dataset as Machine Learning completely works on a dataset. The collected data in a particular format is known as **DATASET.** It’s very necessary to understand your dataset well to work upon. Because the dataset may be in different formats for different purposes. And you need to know well about the dataset to build and analyze the model. To use the dataset in our code we basically use it in CSV format or xlsx(Excel) format.

### **2. Importing Libraries**

In order to perform the data preprocessing using Python, we need to import some of the predefined Python Libraries. To perform some particular task these libraries are very useful.

* [**Numpy**](https://www.geeksforgeeks.org/numpy-in-python-set-1-introduction/)**:** Numpy Python library is used to perform any kind of scientific and mathematical computation in the code.
* [**Pandas**](https://www.geeksforgeeks.org/python-pandas-dataframe/)**:** Pandas is the most famous and useful Python Library used for importing and managing the Dataset. Pandas is open-source data that provide high-performance data manipulation in python.
* [**Matplotlib**](https://www.geeksforgeeks.org/using-matplotlib-with-jupyter-notebook/)**:** Matplotlib is very important in order to visualize our results and have a better view of data. And with this library, we need to import a sub-package called [**Pyplot**](https://www.geeksforgeeks.org/pyplot-in-matplotlib/). This library is used to plot any kind of plot or chart.

|  |
| --- |
| import numpy as np  import pandas as pd  import matplotlib.pyplot as plt |

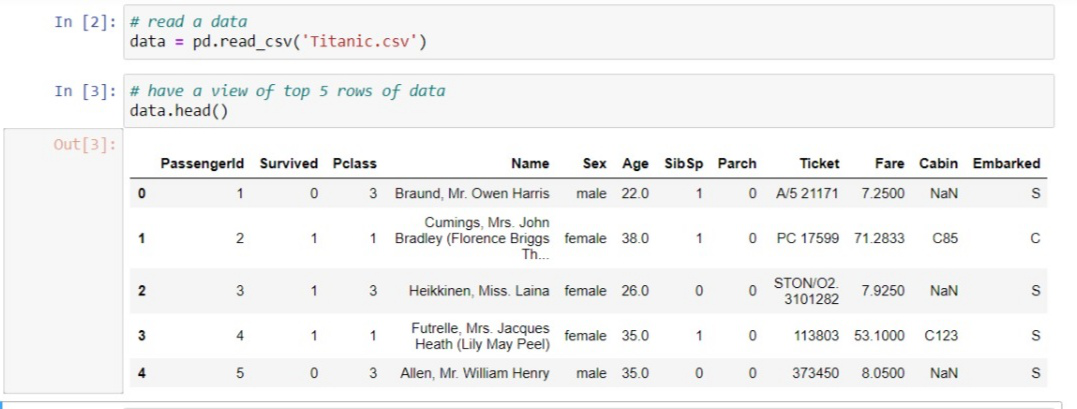


### **3. Importing Dataset**

Now, it’s time to import the dataset which we have collected for the Machine Learning Model. Before importing a dataset make sure to set the current directory as the working directory. Now to import the dataset use [**read\_csv()**](https://www.geeksforgeeks.org/python-read-csv-using-pandas-read_csv/) of the pandas’ library which is used to read a CSV file and perform various operations on it. There are various other options available to read the file in different formats like [**read\_excel()**](https://www.geeksforgeeks.org/working-with-excel-files-using-pandas/) for reading excel files.

|  |
| --- |
| data = pd.read\_csv('dataset.csv') |

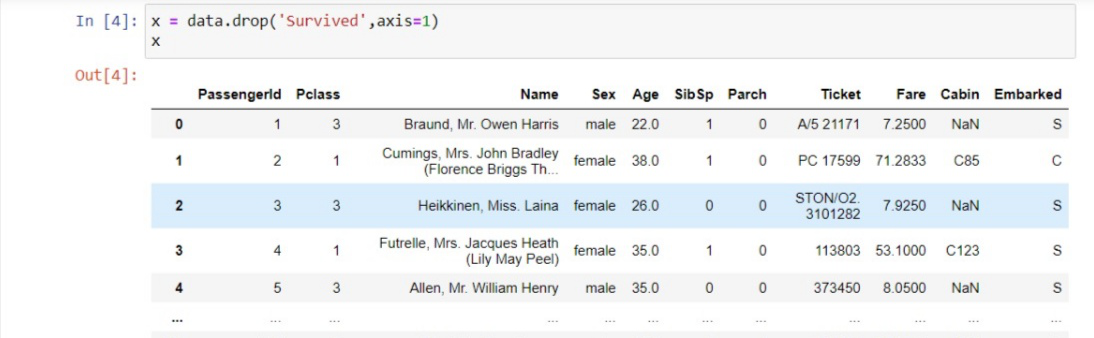
The dataset is the name of the variable which stores the loaded dataset as filed named as a dataset in CSV format. Here is an example with a random dataset known as the Titanic dataset which you can download [here](https://www.kaggle.com/c/titanic/data?select=train.csv).



**Extracting Dependent And Independent Variables:** With a filtered data set explored, you need to create a matrix of Independent variables and a vector of dependent variables. At first, you should decide and confirm that which columns or factors you are using as independent variables(also known as features) to train your model which affects your target variables. For example, If the target column is the last column of your dataset then you can easily create a matrix of your Independent variables as:

|  |
| --- |
| x = data.iloc[:,:-1].values |

Here, the first colon **(:)** represents that we want all the lines in our dataset. **:-1** means we want to take all the columns except the last one. And **.values** means we want to take all the values. And if the target variable is not at last as in our case (Survived) then we can drop that and take the all other columns in the train set. Example:



And to take the vector of the dependent variable which is simply the last column of the dataset you can write as:

|  |
| --- |
| y = data.iloc[:,-1].values |

Or use this method,

|  |
| --- |
| y = data['Survived'] |

### **4. Dealing with missing values**

Missing data can create huge problems to the results hence they are necessary to be removed from the dataset. Missing data is the most common problem and an important step involved in data cleaning. Missing values usually take the form of **NaN or NONE**. The cause of missing value is: sometimes most of the fields in columns are empty which needs to be filled by correct data and sometimes there is incorrect data or poor quality of data which adversely affects the outputs. There are several ways to deal with missing values and fill them:

* The first way to deal with NULL values: you can simply delete the rows or columns which are having NULL values. But most of the time this may lead to loss of information so this method is not so efficient.
* Another important method is to fill the data in place of NULL values. You can calculate the mean of a specific row and column and fill that in place of Null values. The method is very useful in which columns have numerical data such as age, salary, weight, year, etc.
* You can also decide to fill up missing values with whatever value comes directly after it in the same column.

The decisions depend on the type of data, and the results which you want to draw with the data. If missing values are less then it’s good to delete them and if more than 50% of values are missing then it needs to fill with correct data.



Here, the cabin column contains approximately 55% missing data. name and ticket column are not so efficient to predict the person’s survival rate so drop the column which is not significant in prediction otherwise they will harm your machine learning model accuracy.

### **5. Encoding Categorical Data**

Machine learning works with Mathematics and numbers(float or int type). However, it often happens that there are some columns in the dataset which have categorical values(string type) and we want to use them to train our model. In that case, we have to convert categories values to numeric types. For example, there is a sex column in a dataset containing values Male and Female which need to convert to numeric then we can convert it simply like we can assign 0 to Male and 1 to Female and train our model on it. It mostly happens with many columns such as the name of the State, Country, Color, etc. To convert values in numeric we can simply use SKLEARN**preprocessing** library which has the class name **LabelEncoder()**. For example, we take the state column and convert it to numeric values.

|  |
| --- |
| # CATEGORICAL DATA  from sklearn.preprocessing import LabelEncoder  le\_x = LabelEncoder()  x['State'] = le\_x.fit\_transform(x['State']) |

In the above code, we have imported **LabelEncoder** class of **preprocessing** library from sklearn. The class has successfully encoded the variables in digits. If there are more digits encoded such as 0,1,2 then the machine learning model will assume there is some relationship between this And by these, the outputs predicted by the model will be wrong and performance will be reduced. So, to avoid this problem we will use dummy variables.

**Dummy Variables:** Dummy variables are the variables which have value only 0 and 1. In our dataset, if we have 3 values as 0,1 & 2 then it will convert it into 0 and 1. For this, we will use [**OneHotEncoder()**](https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/) class of preprocessing library. it is used when categories are less in a column such as a gender column has values as male and female then it can be encoded as 0,1.

|  |
| --- |
| from sklearn.preprocessing import OneHotEncoder  onehot\_encoder = OneHotEncoder(categorical\_features = ['State'])  x = onehot\_encoder.fit\_transform(x).toarray()    # or  x = pd.get\_dummies(x['gender']) |

Table

Description automatically generated

### **6. Splitting dataset into training and testing set**

The very crucial step of data preprocessing to split the dataset into training and testing sets. By doing this step we can enhance the performance and accuracy of our model in a better and efficient way. If we train the model on the different datasets and test it with a completely different dataset from training then it will create difficulties for the model to estimate the creation between independent and dependent variables and the accuracy of the model will be very less. So, we always try to make a machine learning model that performs well with the training set as well with the testing set.

* **Training set:** Subset of the dataset to train the model, the outputs are known to us as well to model.
* **Testing set:** Subset of the dataset to test the model, which model predicts the output based on training given to the model.

For this, we use the **train\_test\_split** class of **model\_selection** library from the [sklearn](https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/).

|  |
| --- |
| 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) |

* **x\_train:** independent variables(features) for training set
* **y\_train:** features for testing data
* **x\_test:** Dependent variables training data
* **y\_test:** Independent variables testing data

In **train\_test\_split()** we passed 4 parameter in which two are arrays and the third one is test\_size which define the size of testing data. Here we have taken 20% data as testing so given as 0.2 and the remaining 80% will be given to the training set. The last parameter random\_state is used to set a seed for a random generator so that you always get the same result.

### **7. Feature Scaling**

The final step of the machine learning model is feature scaling. It’s a method to Standardize the training dataset in a specific range. In feature scaling, all the values are kept in the same range and on the same scale so that no variable dominates the other variable. For example, we have age and salary in the training dataset then they are not on the same scale as age is 31 and salary is 48000. As Machine learning model is based on Euclidean Distance and if we did not scale the variable it will cause a problem in results and performance.

**EUCLIDEAN DISTANCE BETWEEN A & B = ∑[(x2 – x1)² + (y2 – y1)²]**

For feature scaling, we use the **StandardScaler** class from the **preprocessing**library.

|  |
| --- |
| from sklearn.preprocessing import StandardScaler |

Now, we will create an object of StandardScaler class for independent variables. And then we will fit and transform the training dataset. For the testing dataset, we will directly apply the transform function because it has already done in the training dataset.

|  |
| --- |
| std\_x = StandardScaler()  x\_train = std\_x.fit\_transform(x\_train)  x\_test = std\_x.transform(x\_test) |

As the output, the variables are scaled between -1 to 1. After performing these steps fit the model, predict and deploy it, and then test the accuracy of the model. It will give exact results and the accuracy of the model will always be above 90 percent.

### **Combining All the Above Steps**

Now, in the end, we will combine all the steps to make the code more readable and understandable.

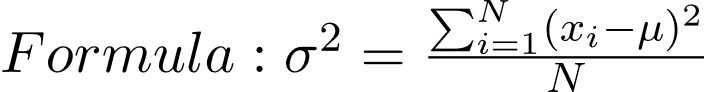
|  |
| --- |
| # Importing libraries  import numpy as np  import pandas as pd  import matplotlib.pyplot as plt    # import dataset  data = pd.read\_csv('datasets/Titanic.csv')    # extract dependent and independent features  x = data.drop('Survived', axis = 1)  y = data['Survived']    # Missing value Imputation  # drop the unrelaven features  x.drop(['Name', 'Ticket', 'Cabin'],         axis = 1, inplace = True)    # numeric value imputation with mean  x['Age'] = x['Age'].fillna(x['Age'].mean())    # categorical value imputation with mode(most frequent category)  x['Embarked'] = x['Embarked'].fillna(x['Embarked'].mode()[0])    # category encoding  x = pd.get\_dummies(x, columns = ['Sex', 'Embarked'],                     prefix = ['Sex', 'Embarked'],                     drop\_first = True)    # 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 = 0.2,                                                      random\_state = 0)    # feature scaling  from sklearn.preprocessing import StandardScaler  std\_x = StandardScaler()  x\_train = std\_x.fit\_transform(x\_train)  x\_test = std\_x.transform(x\_test) |

In the above code, we have covered the necessary steps required to prepare your data or for data preprocessing. But there are some of the areas and the datasets in which all the steps are necessary and in those cases, we can exclude them from our code.

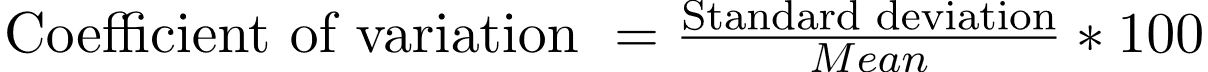
# Statistics (Data Science Track)

1. What is the difference between standard deviation and variance

**Variance** is the sum of squares of differences between all numbers and means.



**Standard Deviation** is square root of variance. It is a measure of the extent to which data varies from the mean.



1. What are (Mean, Median, and Mode)?

Mean is average of a given set of data. Let us consider below example

A Median is a middle value for a sorted data.

A mode is the most frequent value or item of the data set.

1. What is the difference between variance and standard deviation

Q1

1. What is the Box plot?

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is 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.

1. What are the types of skewed data?

**Right skew** **(also called positive skew**)**.** A right-skewed distribution is longer on the right side of its peak than on its left.

**Left skew** **(also called negative skew).** A left-skewed distribution is longer on the left side of its peak than on its right.

**Zero skew.** It is symmetrical and its left and right sides are mirror images.

Chart, line chart

Description automatically generated

1. What is the Z-score?

**Simply put, a z-score (also called a standard score) gives you an idea of how far from the** [**mean**](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/) **a data point is.** But more technically it’s a measure of how many [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) below or above the [population mean](https://www.statisticshowto.com/population-mean/) a [raw score](https://www.statisticshowto.com/raw-score/) is.

A z-score can be placed on a [**normal distribution**](https://www.statisticshowto.com/probability-and-statistics/normal-distributions/) curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the [mean](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/#mean) μ and also the population standard deviation σ.

Z-scores are a way to **compare results** to a “normal” population. Results from tests or surveys have thousands of possible results and units; those results can often seem meaningless. For example, knowing that someone’s weight is 150 pounds might be good information, but if you want to compare it to the “[average](https://www.statisticshowto.com/arithmetic-mean/)” person’s weight, looking at a vast table of data can be overwhelming (especially if some weights are recorded in kilograms). A z-score can tell you **where** that person’s weight is **compared to the average population’s** mean weight.

The **basic z score formula** for a [sample](https://www.statisticshowto.com/sample/) is:

**z = (x – μ) / σ**

## Z Score Formula: Standard Error of the Mean

When you have **multiple samples** and want to describe the standard deviation of those sample means ([the standard error](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-the-standard-error-of-a-sample/)), you would use this z score formula:  
**z = (x – μ) / (σ / √n)**

1. What is the P-value?

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical [hypothesis test](https://www.investopedia.com/terms/h/hypothesistesting.asp), assuming that the [null hypothesis](https://www.investopedia.com/terms/n/null_hypothesis.asp) is correct. The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

1. What is the Pearson correlation coefficient?

The **Pearson correlation coefficient (r)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

|  |  |  |  |
| --- | --- | --- | --- |
| Pearson correlation coefficient (r) | Correlation type | Interpretation | Example |
| Between 0 and 1 | Positive correlation | When one variable changes, the other variable changes in the **same direction**. | Baby length & weight:  The longer the baby, the heavier their weight. |
| 0 | No correlation | There is **no relationship** between the variables. | Car price & width of windshield wipers:  The price of a car is not related to the width of its windshield wipers. |
| Between 0 and –1 | Negative correlation | When one variable changes, the other variable changes in the **opposite direction**. | Elevation & air pressure:  The higher the elevation, the lower the air pressure. |

|  |  |  |
| --- | --- | --- |
| Pearson correlation coefficient (r) value | Strength | Direction |
| Greater than .5 | Strong | Positive |
| Between .3 and .5 | Moderate | Positive |
| Between 0 and .3 | Weak | Positive |
| 0 | None | None |
| Between 0 and –.3 | Weak | Negative |
| Between –.3 and –.5 | Moderate | Negative |
| Less than –.5 | Strong | Negative |

1. What is the A/B Testing?

[A/B testing](https://vwo.com/testing/ab-testing/), also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

1. What is the hypothesis testing

**Hypothesis testing** is a formal procedure for investigating our ideas about the world using [statistics](https://www.scribbr.com/?cat_ID=34372). It is most often used by scientists to test specific predictions, called [hypotheses](https://www.scribbr.com/methodology/hypothesis/), that arise from theories.

There are 5 main steps in hypothesis testing:

1. State your [research hypothesis](https://www.scribbr.com/methodology/hypothesis/) as a [null hypothesis and alternate hypothesis](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/) (Ho) and (Ha or H1).
2. [Collect data](https://www.scribbr.com/methodology/data-collection/) in a way designed to test the hypothesis.
3. Perform an appropriate [statistical test](https://www.scribbr.com/statistics/statistical-tests/).
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your [results](https://www.scribbr.com/dissertation/results/) and [discussion](https://www.scribbr.com/research-paper/discussion/) section.

# ممكن تسأل أسئلة مهمة جدا فى الأنترفيو وفى الغالب هتجاوب عليها بالانجليزى

1. شوفت وظايف التراك اللي علي السايت؟ وحابب تشتغل في اي وظيفة منهم؟ و عارف الوظيفة دي مفروض بتعمل ايه؟ وإيه ال Tools اللى بيستخدمها وبيعملها وإيه الفرق ما بينها وبين ال Jobs التانية اللى ممكن تتخرجيها من التراك وإيه البديل بتاعها (English)
2. هتضيف ايه للتراك ده ؟ أو هتضيفى أيه للمعهد ؟ أو المنحة دى ؟ (English)
3. ليه أقبلك فى التراك ده ؟ (English)
4. حضرت لإنترفيو إزاى ؟ (English)
5. أاتكلم عن مشروع التخرج و كان ايه دورك فيه ؟ (English)
6. شايف نفسك فين بعد خمس سنين ؟(English)
7. أحكى تحدى ليك وإزاى اتصرفت فيه (English)
8. ايه هي عيوبك ؟ + مميزاتك (English)

# Data Set Task (Optional)

ممكن تشوف Data Set (علي جهاز عندهم) لمدة 3 دقائق بكتير وممكن تكون عن سوبر ماركتت أو شركة أو بنك وبعدها بتقول إيه ال insights اللى طلعت بيها + بتشوف فكرت فيها إزاى وعملت إيه؟

للجزئية دى : ممكن تعرف إزاى تعمل Pivot table وتطلع منه Charts + ازاى تعرف ال duplication فى الداتا قد ايه؟ وهتعمل معاها إيه؟ + أزاى تعرف ال missing values وهتعمل معاها إيه؟ + انك تطلع ال Pearson r ما بين اتنين Columns + أنك تعمل Descriptive Statistics لداتا اللى عندك ... دى كلها أفكار ممكن تعملها المهم أنك لازم تكون عارف بتتعمل إزاى وإيه المشاكل اللى ممكن تقابلك وأنت بتنفذها؟ + أختار حاجة أو حاجتين تعملهم طبقا لداتا والوقت اللى هيكون متوفرين

ممكن يكون فى إمتحان Statistics للناس اللى مقدمين على تراك ال Data Science ، الامتحان بيكون مركز على الاسئلة ليها علاقة ب Ratio وال Probability زى مثلا سحب كورة او رمى زهرة وهكذا ، وبيكون إمتحان MCQ على الكمبيوتر تقدر تستخدم ورقة وقلم بس مش مسموح بإستخدام الالة الحاسبة مشكلته فى الوقت أنه ضيق جدا ومحتاج تحل أسئلة زى دى كتير جدا ، ده [موقع](https://www.examveda.com/mcq-question-on-arithmetic-ability/) فى اسئلة كتير كانت جاى منه السنة اللى فاتت (مش شرط السنة دى يجى منه عادى جدا ، كمان ممكن الأفكار فى الامتحان تتغير وده بيكون على حسب نظام كل دفعة)