WORKSHEET 1 SQL

1. A,D
2. A,B
3. B
4. B
5. A
6. C
7. B
8. B
9. B
10. C

11. A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

    A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making. Over time, it builds a historical record that can be invaluable to data scientists and business analysts. Because of these capabilities, a data warehouse can be considered an organization's "single source of truth."

    A typical data warehouse often includes the following elements:

    - A relational database to store and manage data
    - An extraction, loading, and transformation (ELT) solution for preparing the data for analysis
    - Statistical analysis, reporting, and data mining capabilities
    - Client analysis tools for visualizing and presenting data to business users
    - Other, more sophisticated analytical applications that generate actionable information by applying data science and artificial intelligence (AI) algorithms

12. The basic difference between OLTP and OLAP is that OLTP is an online database modifying system, whereas, OLAP is an online database query answering system

OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)
Consists only operational current data.	Consists of historical data from various Databases.
It is application oriented. Used for business tasks.	It is subject oriented. Used for Data Mining, Analytics, Decision making, etc.

The data is used to perform day to day fundamental operations.	The data is used in planning, problem solving and decision making
Simpler queries.	Complex queries.
Tables in OLTP database are normalized (3NF).	Tables in OLAP database are not normalized.
The processing time of a transaction is comparatively less in OLTP.	The processing time of a transaction is comparatively more in OLAP.

13. Characteristics of Data Warehouse are:

- 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.

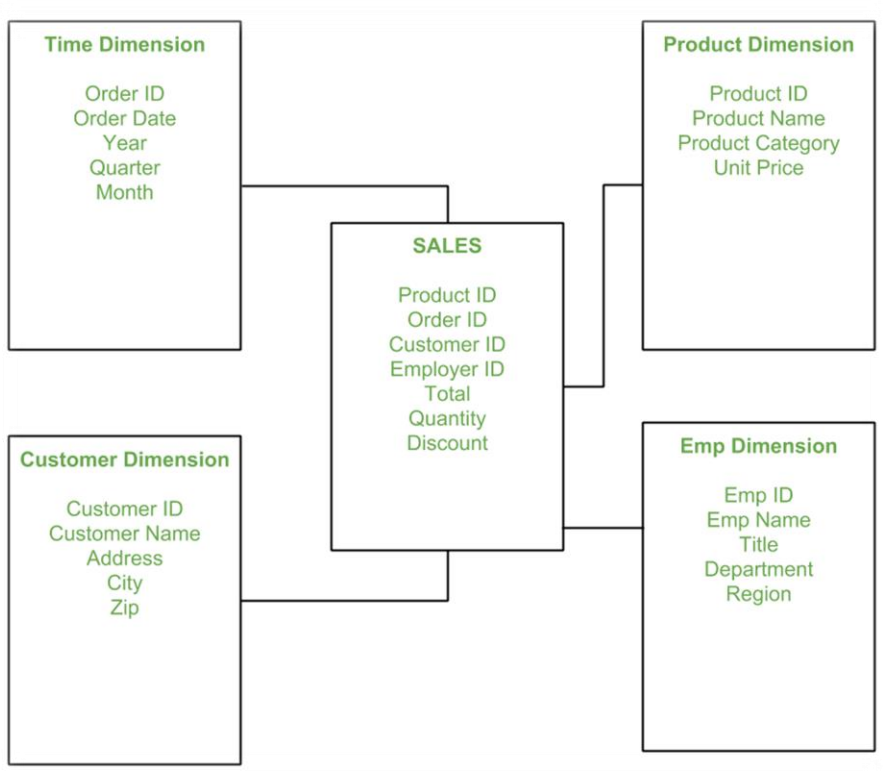
In this, data is read-only and refreshed at particular intervals. This is beneficial in analysing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Two types of data operations done in the data warehouse are:

a) Data Loading
b) Data Access

14.

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 case 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.

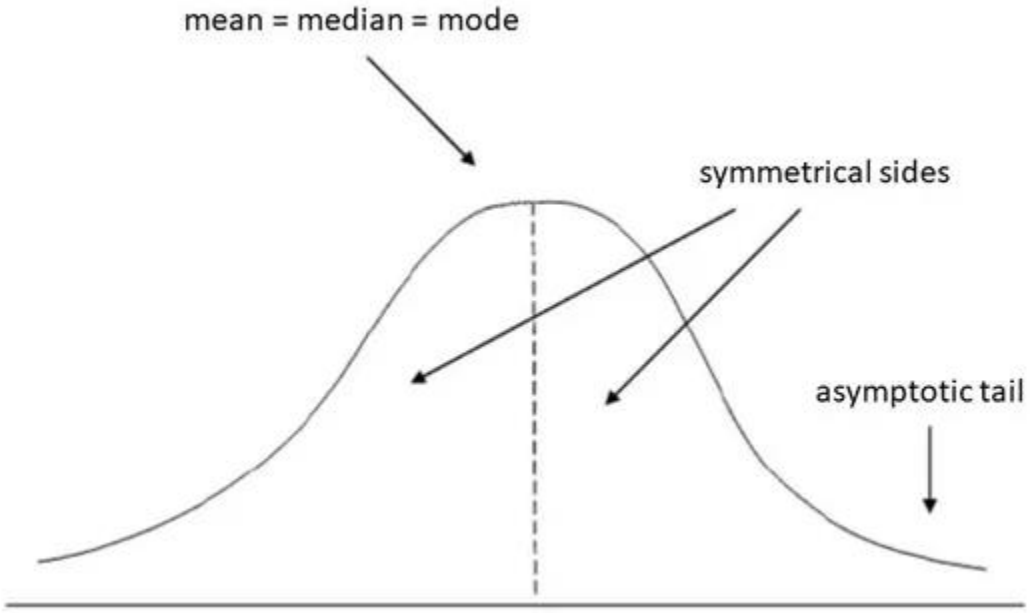
15. SETL (SET Language) is a very high-level programming language based on the mathematical theory of sets.
    SETL provides two basic aggregate data types: unordered sets, and sequences (the latter also called tuples). The elements of sets and tuples can be of any arbitrary type, including sets and tuples themselves. Maps are provided as sets of pairs (i.e., tuples of length 2) and can have arbitrary domain and range types. Primitive operations in SETL include set membership, union, intersection, and power set construction, among others.

    SETL provides quantified boolean expressions constructed using the universal and existential quantifiers of first-order predicate logic.

    SETL provides several iterators to produce a variety of loops over aggregate data structures.

WORKSHEET 1 STATISTICS

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. A
10. The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.
    The area under the normal distribution curve represents probability and the total area under the curve sums to one.
    Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).
    For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.



    The normal distribution is often called the bell curve because the graph of its probability density looks like a bell. It is also known as called Gaussian distribution, after the German mathematician Carl Gauss who first described it.

11. Handling missing data is one of the most important part of data cleaning. Because our model results will be based on how will we treat the missing data.

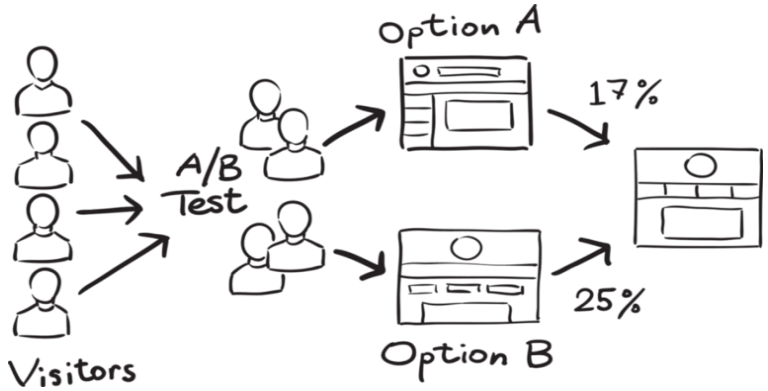
If 70% or more data is missing, it's better to drop that columns as imputing values to missing data will make the model more biased.
Techniques of dealing with missing values depends on the type of column we are dealing with, size of the dataset, percentage of missing values, co-relation of that column with other columns and many other factors. So, it can be concluded that there is no hard and fast rule to impute missing values. Depending on the situation, problem statement and some of the above mentioned factors correct method of imputing the missing values should be used. Some of the methods of dealing with missing values is:

- Dropping the column(If the cost of losing data is very less. As data is costly, this method is not recommended).
- Imputing with a constant value or with zero.
- Imputing with mean/median/mode.
- Using ffill/bfill methods.

12. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

    For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

    In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



    It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.
13. Even though mean imputation of missing values is a very easy process, it causes more problems while building the model. So it's better to keep mean imputation as the last option. Some of the problems of mean imputation may be:

a) It ignores feature co-relation:

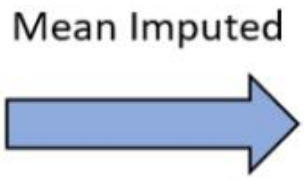
Let's understand this with a small example:

The following table have 3 variables: Age, Gender and Fitness Score. It shows a Fitness Score results (0–10) performed by people of different age and gender.

	Age	Gender	Fitness_Score
0	20	M	8
1	25	F	7
2	30	M	7
3	35	M	7
4	36	F	6
5	42	F	5
6	49	M	6
7	50	F	4
8	55	M	4
9	60	F	5
10	66	M	4
11	70	F	3
12	75	M	3
13	78	F	2

Now let's assume that some of the data in Fitness Score is actually missing, so that after using a mean imputation we can compare results using both tables.

	Age	Gender	Fitness_Score
0	20	M	NaN
1	25	F	7.0
2	30	M	NaN
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	NaN
12	75	M	3.0
13	78	F	NaN



	Age	Gender	Fitness_Score
0	20	M	5.1
1	25	F	7.0
2	30	M	5.1
3	35	M	7.0
4	36	F	6.0
5	42	F	5.0
6	49	M	6.0
7	50	F	4.0
8	55	M	4.0
9	60	F	5.0
10	66	M	4.0
11	70	F	5.1
12	75	M	3.0
13	78	F	5.1

Imputed values don't really make sense — in fact, they can have a negative effect on accuracy when training our ML model. For example, 78 year old women now has a Fitness Score of 5.1, which is typical for people aged between 42 and 60 years old. Mean imputation doesn't take into account a fact that Fitness Score is correlated to Age and Gender features. It only inserts 5.1, a mean of the Fitness Score, while ignoring potential feature correlations.

b) Mean reduces the variance of data:

Based on the previous example, variance of the real Fitness Score and of their mean imputed equivalent will differ. Figure below presents the variance of those two cases:

	Variance
Real Data	3.302198
Missing Data	1.300000

As we can see, the variance was reduced (that big change is because the dataset is very small) after using the Mean Imputation. Going deeper into mathematics, a smaller variance leads to the narrower confidence interval in the probability distribution. This leads to nothing else than introducing a bias to our model.

14. Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.
Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.
The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$y = B0 + B1*x$

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model ($0 * x = 0$). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

15. Statistics have majorly categorized into two types:

- Descriptive statistics
- Inferential statistics

Descriptive Statistics
In this type of statistics, the data is summarized through the given observations. The summarization is one from a sample of population using parameters such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.
Descriptive statistics are also categorized into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics
This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analyzed and summarized then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.

Worksheet-Machine Learning

1. B
2. D
3. D
4. A
5. B
6. D
7. A
8. B
9. D
10. A
11. D
12. A
13. Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Cluster Analysis has been used in marketing for various purposes. Segmentation of consumers in cluster analysis is used on the basis of benefits sought from the purchase of the product. It can be used to identify homogeneous groups of buyers.

Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

The variables on which the cluster analysis is to be done should be selected by keeping past research in mind. It should also be selected by theory, the hypotheses being tested, and the judgment of the researcher. An appropriate measure of distance or similarity should be selected; the most commonly used measure is the Euclidean distance or its square.

Clustering procedures in cluster analysis may be hierarchical, non-hierarchical, or a two-step procedure. A hierarchical procedure in cluster analysis is characterized by the development of a tree like structure. A hierarchical procedure can be agglomerative or divisive. Agglomerative methods in cluster analysis consist of linkage methods, variance methods, and centroid methods. Linkage methods in cluster analysis are comprised of single linkage, complete linkage, and average linkage.

The non-hierarchical methods in cluster analysis are frequently referred to as K means clustering. The two-step procedure can automatically determine the optimal number of clusters by comparing the values of model choice criteria across different clustering solutions. The choice of clustering procedure and the choice of distance measure are interrelated. The relative sizes of clusters in cluster analysis should be meaningful. The clusters should be interpreted in terms of cluster centroids.

There are certain concepts and statistics associated with cluster analysis:

- Agglomeration schedule in cluster analysis gives information on the objects or cases being combined at each stage of the hierarchical clustering process.
- Cluster Centroid is the mean value of a variable for all the cases or objects in a particular cluster.
- A dendrogram is a graphical device for displaying cluster results.
- Distances between cluster centers in cluster analysis indicate how separated the individual pairs of clusters are. The clusters that are widely separated are distinct and therefore desirable.
- Similarity/distance coefficient matrix in cluster analysis is a lower triangle matrix containing pairwise distances between objects or cases.

14. Cluster validation is used to design the procedure of evaluating the goodness of clustering algorithm results. Generally, clustering validation statistics can be categorized into 3 classes:

Internal cluster validation: which uses the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. It can be also used for estimating the number of clusters and the appropriate clustering algorithm without any external data.

External cluster validation: which consists in comparing the results of a cluster analysis to an externally known result, such as externally provided class labels. It measures the extent to which cluster labels match externally supplied class labels. Since we know the "true" cluster number in advance, this approach is mainly used for selecting the right clustering algorithm for a specific data set.

Relative cluster validation: which evaluates the clustering structure by varying different parameter values for the same algorithm (e.g.,: varying the number of clusters k). It's generally used for determining the optimal number of clusters.

Internal validation measures reflect often the compactness, the connectedness and the separation of the cluster partitions.

- **Compactness or cluster cohesion:** Measures how close are the objects within the same cluster. A lower within-cluster variation is an indicator of a good compactness (i.e., a good clustering). The different indices for evaluating the compactness of clusters are base on distance measures such as the cluster-wise within average/median distances between observations.

- **Separation:** Measures how well-separated a cluster is from other clusters. The indices used as separation measures include:
  - distances between cluster centers
  - the pairwise minimum distances between objects in different clusters

- **Connectivity:** corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between 0 and infinity and should be minimized.

Generally most of the indices used for internal clustering validation combine compactness and separation measures as follow:
Index=(α×Separation) / (β×Compactness)
Where α and β are weights.

The two commonly used indices for assessing the goodness of clustering: **the silhouette width and the Dunn index.** These internal measure can be used also to determine the optimal number of clusters in the data.

Silhouette coefficient :

The silhouette analysis measures how well an observation is clustered and it estimates the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.

For each observation i, the silhouette width si is calculated as follows:

a. For each observation i, calculate the average dissimilarity ai between i and all other points of the cluster to which i belongs.
b. For all other clusters C, to which i does not belong, calculate the average dissimilarity d(i,C) of i to all observations of C. The smallest of these d(i,C) is defined as bi=minCd(i,C).

The value of bi can be seen as the dissimilarity between i and its “neighbor” cluster, i.e., the nearest one to which it does not belong.

c. Finally the silhouette width of the observation i is defined by the formula: $Si=(bi-ai)/max(ai,bi)$.

Silhouette width can be interpreted as follow:

1. Observations with a large Si (almost 1) are very well clustered.
2. A small Si (around 0) means that the observation lies between two clusters.
3. Observations with a negative Si are probably placed in the wrong cluster.

Dunn index :

The Dunn index is another internal clustering validation measure which can be computed as follow:

- For each cluster, compute the distance between each of the objects in the cluster and the objects in the other clusters
- Use the minimum of this pairwise distance as the inter-cluster separation (min.separation)
- For each cluster, compute the distance between the objects in the same cluster.
- Use the maximal intra-cluster distance (i.e maximum diameter) as the intra-cluster compactness
- Calculate the Dunn index (D) as follow:

  D=min.separation/max.diameter

If the data set contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large. Thus, Dunn index should be maximized.

15. Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables. There are many clustering methods. But some of the most important types of clustering analysis are:

    - **Connectivity models:** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.

- **Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.

- **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.

- **Density Models:** These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.