Solution- WORKSHEET 1 SQL

1. A , D
2. A, B, C
3. B
4. B
5. A
6. C
7. B
8. B
9. B
10. C
11. Dataware house stores structured ,meaningful data which can be read easily, analyzed and helpful in making business decisions.

12. **OLAP** refers to **Online Analytical processing,** which implies category of software tools which are used to analyze data for business decisions. OLAP systems allow users to analyze data from multiple databases at one time.
    OLTP refers to **Online Transaction Processing**, which supports transaction-oriented applications in a 3-tier architecture.
    OLAP is characterized by a large volume of data while **OLTP** is characterized by large numbers of short online transactions.

13. Below are the four main characteristics of data warehouse:
    Non – Volatile: Data present in data warehouse is permanent i.e data is not deleted on insertion of ne data. Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Data warehouse supports two type of operations : DATA ACCESS AND DATA LOADING
    Time_variant: Data in data warehouse is maintained via different intervals of time such as weekly , monthly or yearly etc. The data resided in data warehouse is predictable with a specific interval of time
    Integrated: 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.
    Subject-oriented –A data warehouse is always subject oriented as it delivers information about a 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.

14. A star schema is a concept where data is organized into **Facts and Dimensions.** Facts refers to an event that is counted or measured example Sale, Inventory etc. Dimensions includes reference data about the fact like amount, time ,location,date or cutomer etc.

15. **ETL** refers to **Extract , Transform** and **Load** process. Where data ins extracted from different source systems , transformed into the data worthy to be stored in data warehouse and then finally loading the transformed data to the data warehouse. The loading can be done in two ways i.e Refresh where the complete data is rewritten into the data warehouse OR Update where only the fresh data given by source system is stored in data warehouse without deleting the existing data.
The whole ETL process is usually carried out with the help of ETL tools for example: **Informatica** which is a widely used ETL tool in IT industry.

Solution- STATISTICS WORKSHEET-1

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C
10. **Normal distribution** is a probability function which describes how the values of a variable is distributed. This distribution is symmetric with most of the values gathering around the central peak and probabilities for values other than peak goes on flattening the curve in both the directions. The shape of this distribution is always bell shaped.
11. Missing data should be handled after looking problem under consideration and the variable for which we are going to fix the missing values.
In python we have a libraries like **fancyimpute, impyute** using which missing data can be handled in more appropriate way. Other than python there are some commonly know algorithms used in real life datasets like **:**
MICE: In this first missing values are replaced by taking mean and the positions of missing data points are noted. Then Linear Regression algorithm is performed

where the variables which contained missing values becomes independent variable and all other features becomes independent variables. The missing values are then predicted using number of iterations (usually 10). This way a more accurate results can be obtained than simple mean method of replacement.

KNN: This algo gives n most similar neighbors of the missing data point and then the mean (in continuous variable) or mode(in categorical variable) of these neighbors are taken to replace null values

12. A/**B testing** is basically statistical hypothesis testing, or, in other words, statistical inference. In this decisions / inferences are made for the population by studying a sample of that population.
13. Mean imputation is not considered best practice because it ignores the feature correlation with other variables in the dataset and just perform mathematical mean of the variable which may not follow the correlation ,which non missing values have, with other variables. Another fact is in case the dataset is small then replacement with mean reduces the variance and hence reducing the confidence interval.
14. Linear Regression is a technique to model the relationship between dependent and independent variable using a linear equation.The parameters of this linear equation are estimated from the data under consideration. In case of only one independent variable this technique is known as simple linear regression and in case of multiple independent variables this is known as multiple linear regression.
15. There are two main branches of statistics :
    Descriptive staistics:This deals with collecting ,summarizing and presenting the data.
    Inferential statistics: This makes use of statistical analysis performed during descriptive statistics to draw inferences. Inferences made on the population by studying and analyzing the sample from the population.

Solution- MACHINE LEARNING

1. B
2. D
3. D
4. A
5. C
6. D

7. B
8. B
9. A
10. A
11. D
12. A
13. There are three main factors by which clustering analysis can be calculated:
    Clustering tendency: Before evaluating the clustering performance, make sure that data set we are working has clustering tendency and does not contain uniformly distributed points is very important. If so then any algorithm will not give correct results.
    Number of Clusters: Some of the clustering algorithms like K-means, require number of clusters, *k*, as clustering parameter. Getting the optimal number of clusters is very significant in the analysis. Getting right k can be done using data visualization, knowledge of domain or data driven methods like elbow method.
    Clustering quality: Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. (Detailed explanation is in answer 14.)

14. Cluster quality measurement can be done using quality metrices. Ideal clustering is characterized by minimal intra cluster distance and maximum inter cluster distance. There are majorly two types of measures to measure quality of cluster:
    Extrinsic Measures: which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.
    Intrinsic Measures :that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.
15. Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related.
    There are majorly 4 types of clustering analysis in data science:
    Centroid clustering: In this clustering is done using the centroids which equals to the number of clusters we want to have in our cluster. At first a any random points are considered as clusters and then the distance of each data point is measured from these centroids and each data point is assigned to a cluster of minimal distance. Then the mean (can be mode, median) of all data points is taken in each cluster to have new centroid and the process is repeated till we find no further moving clusters.
    Density clustering: Density clustering groups data points by how densely populated they are and not by taking the distance as in Centroid clustering. To group closely

related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are.

Distribution Clustering: Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid , the algorithm defines the density distributions for each cluster, giving the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data.

Connectivity clustering: Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster.