Worksheet 1

1. A

2. D

3. D

4. A

5. B

6. D

7. B

8. B

9. A

10. A

11. D

12. D

13. Cluster analysis can be calculated by following 3 important steps those are calculating the distance, linking the clusters, choosing the solution by the appropriate number of clusters.

14. The quality of cluster analysis can be seen be a number of metrices. An efficient cluster has minimal intra cluster distance and maximum inter cluster distance.

15. Cluster analysis is an unsupervised machine learning. It helps in clustering data points into groups. There are three types of cluster analysis such as Kmeans , density based, distribution based.

Worksheet 2

1. A

2. A

3. B

4. B

5. A

6. C

7. B

8. D

9. B

10. A

11. A data warehouse is the main repository of date that can be used for analyzing the data. It is a process of collecting and managing the data from various source.

12. Online Transaction Processing captures and stores the real time data whereas Online analytical processing is a method of using queries to analyze historical data of OLTP.

13. Data Warehouse generally have 4 important characteristics such as subject oriented as it will have all details about the clients, integrated as it will integrate data from all sources, time variant and non-volatile.

14. A star schema is like database organizational structure that is used in data warehouse. It is like a large single fact table that store transactional or measured data.

15. It's a language for SET theory of mathematics.

Worksheet 3

1. A

2. A

3. C

4. A

5. C

6. B

7. B

8. A

9. C

10. A normal distribution is such type of distribution of which has mean = median = mode. It is a type of continuous probability distribution in which most data points cluster towards middle range. It has bell shaped curve.

11. Missing data can be handled by deleting the rows if it is less than 20 % else it can be replaced by median or mode.

12. AB testing is a process of splitting the data in different ways to check which split is giving more accurate value like cross validation in Python.

13. Mean imputation should not be used as it does not consider the correlations.

14. Linear regression used for predictive analysis, it is used to describe date and explain the relationship.

15. The 3 main branches of statistics are data collection, descriptive statistics and inferential statistics.