WORKSHEET -1

13. How is cluster analysis calculated?

1) Cluster analysis is all finding observations which are similar data.

2) However, when group them together, they should be homogeneous within group.

3) Observation of the group should be quiet dissimilar to any object of the other group.

4) In cluster analysis if we consider shape, color, or some other then you can make group based on that.

14. How is cluster quality measured?

Ans :- If all the data objects in the cluster are highly similar then the cluster has high quality.

And the data is quiet distinct their similar categories

15. What is cluster analysis and its types?

Ans :- Cluster analysis is a data analysis technique, that data is occurring groups within a data set known as cluster.

Types of Cluster analysis :-

1) Hierarchical Clustering.

2) Density based Clustering.

3) Distribution based Clustering.

4) Centroid based Clustering.

WORKSHEET-2

11. What is data-warehouse?

A data warehousing is process for collecting and managing data from varied sources.

A data warehouse is typically used to connect and analyze business data from heterogeneous sources.

12) What is the difference between OLTP VS OLAP?

OLTP – (Online transaction process)

1) It is online transactional system. It manages modification.

2) It characterized by large number of short online transaction.

3) It is an data base online modifying system.

4) It is used traditional data base management system.

OLAP –(Online Analysis process)

1) It is an analysis and data retrieving process.

2) It is characterized by a large volume of data.

3) It is an database query management system.

4) It is customer orientated process.

13) What are the various characteristics of data-warehouse?

- Subject Oriented
- Integrated
- Time-Variant
- Non-Volatile

14) What is star-Schema?

A star-schema is a multi-dimensional data model used to organize data in a database,

So that it is easy to understand and analyze.

15) What do you mean SETL?

SETL is a very high-level programming language that based on the mathematical theory of sets.

WORKSHEET 3

10) What do you understand by the normal Distribution?

The normal distribution describes how the values of a variable are distributed.

It is most important distribution in statistics because it accurately describes the

Distribution of values for many natural phenomena.

11) How do you handle missing data? What imputation techniques do you recommend?

1] In missing data, only when we have huge dataset and from that data set some data is missing in that case delete that particular data, but only when we have huge dataset.

2] Separate model dataset – for example :- Data is arrange in the rows and columns so some data is missing in particular row from the present data in that column is input and missing data is output then we can handle it. In this way we handle missing data.

X	Y	Z	Output	so Y and Z is input and X is output
_	10	30	50	

Imputation techniques: - 1) in the list of data for any value is empty for any of the column remove that data from the list.

2) The values any data are different in a list and one is missing then add all values and divide total number of value we can get missing data.

12) What is A/B testing?

A/B testing is a multi variant testing. A method of comparing two versions of a web page or app against each other to determine which one performs better.

13) Is mean imputation of missing data acceptable practice?

Mean imputation is typically considered terrible practice since it ignores feature correlation.

14) What is linear regression in statistics?

It is a basic and commonly used type of predictive analysis.

In linear regression are dependent variable vs independent variable in between variables line of regression and that line is lie in data points

15) What are the various branches of statistics?

Data collection

Descriptive statistics

Inferential statistic.