WORKSHEET 1 SQL

1. A) Create , D) ALTER
2. A) Update, B) Delete
3. B) Structured Query Language
4. B) Data Definition Language
5. A) Data Manipulation Language
6. C) Create Table A (B int,C float)
7. B) Alter Table A ADD COLUMN D float
8. B) Alter Table A Drop Column D
9. B) Alter Table A Alter Column D int
10. B) Alter table (B primary key)
11. Data warehouse is a type of data management system where data is stored and used to perform queries,reporting and analysis the data whenever needed.
12. Online transaction processing(OLTP) is a system captures and maintains transaction data ina database and handles large number of small transactions wheras Online analytical processing(OLAP) is a system uses large amount of data for data mining, data analyzing and business intellengence projects and handles large volumes of data with complex queries.
13. There are four main characteristics of data warehouse - subject oriented, integrated, time-variant, non-volatile.
14. Star schema is the elementary form of a dimentional model in which data are organized into facts and dimentions. A fact is an event that is counted or measured where as dimensions includes reference data about the fact such as date.
15. SETL means set language . It is very high level programming language based on mathematical theory of sets.

STATISTICS WORKSHEET-1

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. a) Probability
8. a) 0
9. c) Outliers cannot conform to the regression relationship
10. Normal distribution is a distribution of data which is normally distributed. The normal distribution graph is bell shaped and there will be no outliners.
11. Firstly we should check if the missing data is related to any other data,if the data is not related to other data we can delete the missing data or we can replace with mean imputation technique.
12. A/B testing is a way of comparing two or more versions to determine which version perfoms better and also to understand which version is significant.
13. No the mean imputation of missing data is not accepatable practice, mean imputation does not preserve the the relationship between variables and it leads to underestimate of standard errors.
14. Linear regression is a basic regression method used for predictive analysis. The regression is to exmine does the set of predictor variables do a good job in predicting an outcome variable.

15. There are four braches in statistics they are methamatical or theoretical statistics, stastical methods, descriptive statistics and inferential statistics.

MACHINE LEARNING

1. b) 4
2. d) 1, 2 and 4
3. d) formulating the clustering problem
4. a) Euclidean distance
5. b) Divisive clustering
6. d) All answers are correct
7. b) Classify the data point into different classes
8. b) Unsupervised learning
9. a) K- Means clustering
10. a) K-means clustering algorithm
11. d) All of the above
12. a) Labeled data
13. Cluster analysis is calculated by calculating the distance, linking the clusters and choosing the clusters by selecting the right number of clusters.
14. Cluster quality is measured in various ways, if the clusters are highly similar than the cluster has high quality. We can measure the quality of clustering by using dissimilarity/similarity metric. The other ways are cluster completeness,ragbag and small cluster preservation.
15. Cluster analysis is a data mining technique whose goal is to group objects based on set of user selected data. The types of cluster analysis are hierarchical cluster analysis,centroid based clustering,disrtibustion based clustering and density based clustering.