STATISTICS WORKSHEET-2

1. What represent a population parameter?
   Mean

2. What will be the median of following set of scores(18,6,12,10,14)
   12

3. What is standard deviation?
   All of the above

4. The intervals should be _____ in a grouped frequency distribution

5. What is the goal of descriptive statistics?
   All of the above

6. A set of data organized in a participant by variables format is called
   Data Set

7. In multiple regressions, _____ dependent variables are used
   2 or More

8. Which of the following is used when you want to visually examine the relationship between 2 quantitative variables?
   Scatter Plot

9. Two or more group's means are compared by using
   Analysis of Variance

10. _____is a raw score which has been transformed into standard deviation units?
    Z-score

11. _____is the value calculated when you want the arithmetic average?

Mean

12. Find the mean of these set of number (4, 6, 7, 9, 2000000)?
    7

13._____ is a measure of central tendency that takes into account the magnitude of scores?
    Median

14._____ focuses on describing or explaining data whereas _____involves going beyond immediate data and making inferences?
    Descriptive and inferences

15. What is the formula for range?
    H-L

WORKSHEET 2 SQL

1. Which of the following constraint requires that there should not be duplicate entries?
    a. Unique

2. Which of the following constraint allows null values in a column?
    a. None of them

3. Which of the following statements are true regarding Primary Key?
   Each entry in the primary key uniquely identifies each entry or row in the table

4. Which of the following statements are true regarding Unique Key?
   Multiple columns can make a single unique key together

5. Which of the following is/are example of referential constraint?

Foreign Key

6. How many foreign keys are there in the Supplier table?
2

7. The type of relationship between Supplier table and Product table is:
One to one

8. The type of relationship between Order table and Headquarter table is:

9. Which of the following is a foreign key in Delivery table?
Delivery id

10. The number of foreign keys in order details is:.
1

11. The type of relationship between Order Detail table and Product table is:
One to many

12. DDL statements perform operation on which of the following database objects?
Table

13. Which of the following statement is used to enter rows in a table?
Insert in to

Q14 and Q15 have one or more correct answer. Choose all the correct option to answer your question.

14. Which of the following is/are entity constraints in SQL?
Unique
Primary Key
Null

15. Which of the following statements is an example of semantic Constraint?
A blood group can contain one of the following values - A, B, AB and O.
Two or more donors can have same blood group

Machine Learning

1. Movie Recommendation systems are an example of:

Clustering

2. Sentiment Analysis is an example of:

Regression

Classification

Reinforcement

3.Can decision trees be used for performing clustering?

True.

4.Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

Capping and flooring of variables

5.What is the minimum no. of variables/ features required to perform clustering?

1

6.For two runs of K-Mean clustering is it expected to get same clustering results?

No

7.Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

Yes

8.Which of the following can act as possible termination conditions in K-Means?

All of the above

9. Which of the following can act as possible termination conditions in K-Means?

10.Which of the following algorithms is most sensitive to outliers?

K-means clustering algorithm

11.How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

All of the above

12. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

All of the above

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

13. Is K sensitive to outliers?

The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers.

Instead of using the mean point as the center of a cluster, K-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points,

14. Why is K means better?

K-Means for Clustering is one of the popular algorithms for this approach. Where K means the number of clustering and means implies the statistics mean a problem. It is used to calculate code-vectors (the centroids of different clusters).

K-means has been around since the 1970s and fares better than other clustering algorithms like density-based, expectation-maximization. It is one of the most robust methods, especially for image segmentation and image annotation projects.

According to some users, K-means is very simple and easy to implement. However, it is unlikely to be the state-of-the-art, but for straightforward clustering, it is also a part of a larger data-processing pipeline, K-means is a reasonable default choice, at least until you figure out that the clustering step is your bottleneck in terms of overall performance.

K-means is used to learn feature representations for images (use k-means to cluster small patches of pixels from natural images, then represent images in the basis of cluster centres; repeat this several times to form a “deep” network of feature representations) gives image classification results that are competitive with much more complex / intimidating deep neural network models. In fact, a lot of k-means applications are now done using support vector machines.

15. Is K means a deterministic algorithm?

The basic k-means clustering is based on a non-deterministic algorithm.

This means that running the algorithm several times on the same data, could give different results.

The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids.

The key idea of the algorithm is to select data points which belong to dense regions and which are adequately separated in feature space as the initial centroids.