Statistics Worksheet 1

Q1) A → True

Q2) A → Central Limit Theorem

Q3) B → Modeling bounded count data

Q4) D → All options

Q5) C → Poisson Distribution

Q6) B → False

Q7) B → Hypothesis Testing

Q8) A → 0

Q9) C → Outliers cannot conform to regression relationship

Q10) Normal distribution also known as Gaussian Distribution resembles bell shape curve wherein the data is majorly distributed around the mean and is symmetric around the mean and tappers off as they move away from centre. In case of Normal Distribution the Mean = Median = Mode.

Q11) Missing data can be managed in multiple ways either by Deletion or Imputation

- If the data missing is 5 to 10% of the entire dataset it can be deleted using dropna.
- In case of continuous variable by eliminating the outliers and then replacing the null values with Mean.
- In case of categorical data the missing values can be replaced with Mode.

Q12) A/B testing also known as Split Test is one of the most prominent statistical tools to compare 2 versions of a variable/item to see which variable performs better. In this case some modification are made in A and then relabelled as B and then tested that which variable performs better.

Q13) Mean imputation can be a good option as it preserves the mean of the entire data in case there are no outliers because in presence of outliers imputing with Mean may lead to distort the overall mean. Imputing with mean will also help in preserving the entire data and hence will not lead to loss of data.

Q14) Linear Regression tries to draft a relationship between 2 variables by applying linear equation y=mx+c where m is intercept and c is constant. By using Linear Regression Model best fit line is generated which tries to cover maximum data points with least RMSE.

Q15) Branches of Statistics: Descriptive Statistics and Inferential Statistics.

SQL Worksheet 1

Q1) A & D → Create & Alter

Q2) A,B&C → Update, Delete, Select

Q3) B → Structured Query Language

Q4) B → Data Defination Language

Q5) A → Data Manipulation Language

Q6) C → Create Table A (B int,C float)

Q7) B → Alter Table A ADD COLUMN D Float

Q8) B → Alter Table A Drop Column D

Q9) B → Alter Table A Alter Column D int

Q10) C → Alter Table A Add Primary key B

Q11) Data Warehouse is a system that aggregates data (structured and semi-structured) from different sources for the purpose of analysis.

Q12) Primary difference between OLTP and OLAP is

- A) OLTP is optimized for processing massive no of transactions and designed for frontline workers whereas OLAP is optimized for smart decision making and used by DS and BA.
- B) OLTP uses traditional DBMS to accommodate large volume of real time transactions whereas OLAP supports complex queries of multiple data facts from current and historical data.
- C) OLTP requires frequent/concurrent backsups as the data is modified frequently whereas OLAP system backup is less frequent.
- D) OLTP response time is faster than OLAP

Q13) Characteristics of Data Warehouse are:

- Subject Oriented
- Time Variant
- Non Volatile
- Integrated

Q14) Star Schema is the elementary form of a dimensional model, in which data are organized into facts and dimensions. It is known as Star Schema because entity relationship diagram of this schema stimulates a star with points diverging from center, center consists of large facts and the diverging points are dimensions.

Q15) SETL is a high level programming language based on mathematical theory of sets

Machine Learning

Q1) D → 8

Q2) D → 1,2 and 4

Q3) D → Formulating the clustering problem

Q4) A → Euclidean Distance

Q5) B → Divisive Clustering

Q6) D → All answers are correct

Q7) A → Divide data points into groups

Q8) B → Unsupervised Learning

Q9) D → All

Q10) A → K Means Clustering Algorithm

Q11) D → All

Q12) A → Labelled Data

Q13) By using Internal and External evaluation.

Q14) If all the data objects in the cluster are highly similar then the cluster has high quality. Quality of cluster can be measured by using Dissimilarity/Similarity Metric.

Q15) Cluster Analysis is data mining technique which groups the data based on similar characteristics or attributes. Types of Clustering are mentioned below

- Hierarchical Based Clustering
- Density Based Clustering
- Centroid Bases Clustering
- Distribution Based Clustering