**WORKSHEET 1 SQL**

1. (a).create and (d).alter

2. (a). update, (b) delete, (c) select

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. (D) Alter table A Column D float to int

10.(A) Alter Table A add constraint primary key B

11. Data-Warehouse:

A DataWarehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources.

The decision support database (Data Warehouse) is maintained separately from the organization's operational database. However, the data warehouse is not a product but an environment. It is an architectural construct of an information system which provides users with current and historical decision support information which is difficult to access or present in the traditional operational data store.

12. Difference b/w OLTP vs OLAP:

OLTP and OLAP both are the online processing systems. OLTP is a transactional processing while OLAP is an analytical processing system. OLTP is a system that manages transaction-oriented applications on the internet for example, ATM. OLAP is an online system that reports to multidimensional analytical queries like financial reporting, forecasting, etc.

### **Definition of OLTP :**

OLTP is an **Online Transaction Processing system**. The main focus of OLTP system is to record the current **Update, Insertion and Deletion** while transaction. The OLTP queries are **simpler** and **short** and hence require**less time in processing**, and also requires **less space**.

### **Definition of OLAP:**

OLAP is an **Online Analytical Processing system**. OLAP database stores historical data that has been inputted by OLTP. It allows a user to view different summaries of multi-dimensional data. Using OLAP, you can extract information from a large database and analyze it for decision making.

13. characteristics of data-warehouse:

* Some data is denormalized for simplification and to improve performance.
* Large amounts of historical data are used.
* Queries often retrieve large amounts of data.
* Both planned and ad hoc queries are common.
* The data load is controlled.

14. Star Schema :

In computing, the starschema is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts. The starschema consists of one or more fact tables referencing any number of dimension tables.

The star schema separates business process data into facts, which hold the measurable, quantitative data about a business, and dimensions which are descriptive attributes related to fact data. Examples of fact data include sales price, sale quantity, and time, distance, speed and weight measurements. Related dimension attribute examples include product models, product colors, product sizes, geographic locations, and salesperson names.

15.**SETL :**

SETL is an initiative to deploy a multi-asset, multi-currency institutional payment and settlements infrastructure based on blockchain technology. The SETL system will enable market participants to move cash and assets directly between each other, facilitating the immediate and final settlement of market transactions.

**STATISTICS WORKSHEET-1**

1. a). true

2. a). central limit theorem.

3. a). Modelling bounded count data

4. d). All of mentioned

5. c). Poisson distribution

6. b). False

7. b). Hypothesis

8. a). 0

9. c) Outliers cannot conform to the regression relationship

10. Normal Distribution:

The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean.

Common Continuous Probability Distribution is stated for Normal Distribution in statistics for advanced data evaluation and interpretation  
  
  
Data utility - Random variable ∈ R  
  
classification -  
Normal Distribution -  
1.) Probability Distribution Function [ P.D.F. ]  
2.) Cumulative Distribution Function [ C.D.F. ]  
  
STANDARD CASE -  
  
STANDARD NORMAL DISTRIBUTION FUNCTION  
WITH σ = 1 and μ = 0  
  
Stated as PDF  
  
INFLECTION POINTS - ±1

11. How to Handle missing data :

1. Use the 'mean' from each column. Filling the NaN values with the mean along each column.
2. Use the 'most frequent' value from each column. Now let's consider a new DataFrame, the one with categorical features.
3. Use 'interpolation' in each column.
4. Use other methods like K-Nearest Neighbor.

And some of the imputation techniques are as follows:

1. Listwise or case deletion. ...
2. Mean substitution. ...
3. Regression imputation. ...
4. Maximum likelihood. ...
5. Multiple imputation.

12. A/B testing :

Like any type of scientific testing, A/B testing is basically statisticalhypothesistesting, or, in other words, statistical inference It is an analytical method for making decisions that estimates population parameters based on sample statistics.

The population refers to all the visitors coming to your website (or specific group of pages), while the sample refers to the number of visitors that participated in the test.

In statistics your hypothesis breaks down into:

* Null hypothesis
* Alternative hypothesis

The null hypothesis states the default position to be tested or the situation as it is (assumed to be) now, i.e. the status quo.

The alternative hypothesis challenges the status quo (the null hypothesis) and is basically a hypothesis that the researcher (you) believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

13. Is mean imputation of missing data acceptable practice?

It is a non-standard, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missingdata. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.

14. Linear Regression :

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15. Various branches of statistics :

The two main branches of statistics are [descriptive statistics](https://explorable.com/descriptive-statistics) and [inferential statistics](https://explorable.com/inferential-statistics). Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics :deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid [biases](https://explorable.com/research-bias) that are so easy to creep into the [experiment](https://explorable.com/conducting-an-experiment).

Inferential Statistics: as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

**MACHINE LEARNING WORKSHEET**

1. d). 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. d). all the above

8. b). unsupervised learning

9. a), K Means clustering.

10. a). K Means clustering

11. c). data points with non-convex shapes.

12. a). labeled data

13. how is cluster analysis calculated?

The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.

14. how is cluster quality measured?

To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

15. cluster analysis and its types:

Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

Types : 1. Centroid clustering

2. Density clustering

3. Distribution clustering

4. Connectivity clustering.