SQL WORKSHEET 1 SOLUTIONS

1. DDL commands : a) Create and d) Alter
2. DML commands: a) Update , b) Delete and 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. b) Alter Table A Alter Column D int
10. d)) None of them (Correct answer : Alter Table A Add Primary Key (B))
11. **Data Warehouse** : It is an e-storage of large amounts of information by a business organization. It is a system which has historical data, designed to run query and used for reporting and data analysis on the data derived from business transactional processes.

12. OLTP VS OLAP :

	OLTP	**OLAP**
Characteristics	Handles a large number of small transactions	Handles large volumes of data with complex queries
Query types	Simple standardized queries	Complex queries
Operations	Based on INSERT, UPDATE, DELETE commands	Based on SELECT commands to aggregate data for reporting
Response time	Milliseconds	Seconds, minutes, or hours depending on the amount of data to process
Design	Industry-specific, such as retail, manufacturing, or banking	Subject-specific, such as sales, inventory, or marketing
Source	Transactions	Aggregated data from transactions

Purpose	Control and run essential business operations in real time	Plan, solve problems, support decisions, discover hidden insights
Data updates	Short, fast updates initiated by user	Data periodically refreshed with scheduled, long-running batch jobs
Space requirements	Generally small if historical data is archived	Generally large due to aggregating large datasets
Backup and recovery	Regular backups required to ensure business continuity and meet legal and governance requirements	Lost data can be reloaded from OLTP database as needed in lieu of regular backups
Productivity	Increases productivity of end users	Increases productivity of business managers, data analysts, and executives
Data view	Lists day-to-day business transactions	Multi-dimensional view of enterprise data
User examples	Customer-facing personnel, clerks, online shoppers	Knowledge workers such as data analysts, business analysts, and executives
Database design	Normalized databases for efficiency	Denormalized databases for analysis

13. **Characteristics of Data Warehouse** :

a) Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

b) Integrated: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

c) Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

d) Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

14. The **star schema** is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts.[1] The star schema consists of one or more fact tables referencing any number of dimension tables. The star schema gets its name from the physical model's[3] resemblance to a star shape with a fact table at its center and the dimension tables surrounding it representing the star's points.

15. In data management, **semantic warehousing** is a methodology of digitized text data using similar functions to Data warehousing (DW), such as ETL(Extract, transform, load), ODS(Operational data store), and MODEL. Key value operation is less useful for the digitized text. Semantic warehousing is different from DW in that semantic information base from text(semantic) data.

Semantic warehousing is different from search engine in that semantic information base from text data is stored in the database.(DBMS)

SETL builds on Semantic Web (SW) standards and tools and supports developers by offering a number of powerful modules, classes, and methods for (dimensional and semantic) DW constructs and tasks. Thus it supports semantic data sources in addition to traditional data sources, semantic integration, and creating or publishing a semantic (multidimensional) DW in terms of a knowledge base. A comprehensive experimental evaluation comparing SETL to a solution made with traditional tools (requiring much more hand-coding) on a concrete use case, shows that SETL provides better programmer productivity, knowledge base quality, and performance.

STATISTICS WORKSHEET I SOLUTIONS

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship
10. Normal Distribution: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.
    - A normal distribution is the proper term for a probability bell curve.
    - In a normal distribution the mean is zero and the standard deviation is 1.
    - It has zero skew and a kurtosis of 3.
    - Normal distributions are symmetrical, but not all symmetrical distributions are normal.
11. Handle missing values by following methods:
    a) Delete rows or columns : If there are more than 70-75% null values in a column or row, we can delete them. This method is advised only when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

    b) Replace with Mean/Median/Mode : This strategy can be applied on a feature which has numeric data and for categorical data mode can be used. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set.

    c) Labelling the data : A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values.

    d) Predicting the null values : Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. For eg : We can use linear regression to predict numerical values. We can use KNN to find the k- nearest cluster and also can use random forest which works well on non-linear and categorical data.

Any of the above imputation methods can be used to fill in the missing values, based on the scenarios as each dataset is unique.

12. **A/B Testing :** A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

13. Yes, mean imputation of missing data is acceptable practise. It can be applied on numerical data.

14. **Linear Regression :** In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

15. The two main branches of statistics are : Descriptive and Inferential Statistics.

   a) **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 avoiding biases that are so easy to creep into the experiment.

   b) **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 SOLUTIONS

1. b) 4
2. d) 1, 2 and 4
3. d) formulating the clustering problem
4. a) Euclidean distance
5. c) Agglomerative clustering
6. d) All answers are correct
7. a) Divide the data points into groups
8. b) Unsupervised learning
9. d) All of the above
10. a) K-means clustering algorithm
11. d) All of the above
12. a) Labeled data
13. Cluster Analysis Calculation :

    Step 1: Choose the number of clusters *k :* The first step in k-means is to pick the number of clusters, k.

    Step 2: Select k random points from the data as centroids : Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:

    Step 3: Assign all the points to the closest cluster centroid : Once we have initialized the centroids, we assign each point to the closest cluster centroid:

    Step 4: Recompute the centroids of newly formed clusters :Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters:

    Step 5: Repeat steps 3 and 4 :We then repeat steps 3 and 4:

14. Measuring clustering quality is an important issue just because clustering is an unsupervised measure. We want to evaluate the goodness of clustering results by either some internal or external measures. Unfortunately there's no commonly recognized best suitable measure in practice. But there are three categories of measures we call external measures, internal measures and relative measures.

For **external measures**, we can consider them supervised. That means we may have some prior or expert knowledge. For example, some ground truth. Then we can compare the clustering results against the prior or expert specified knowledge, using certain clustering quality measures.

Then the second type of measure is called **internal measure**, which is unsupervised. That means the criteria derived from the data itself. In that case, we will evaluate the goodness of clustering by considering how well the clusters are separated and how compact the clusters are. For example, we can use silhouette coefficient.

The third one is a **relative measure**. That means we can directly compare different class rings using those obtained via different parameter settings for the same algorithm. For example, for the same algorithm, we can use a different number of clusters.

15. Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevance 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.
    a) It groups the similar data in the same group.
    b) The goal of this procedure is that the objects in a group are similar to one another and are different from the objects in other groups.
    c) Greater the similarity within a group and greater difference between the groups, more distinct the clustering.
    d) Cluster analysis provides a potential relationship and constructs systematic structure in a large number of variables and observations.

Types:

1. Hierarchical clustering: Also known as 'nesting clustering' as it also clusters to exist within bigger clusters to form a tree.
2. Partition clustering: It's simply a division of the set of data objects into non-overlapping clusters such that each object is in exactly one subset.
3. Exclusive Clustering: They assign each value to a single cluster.
4. Overlapping Clustering: It is used to reflect the fact that an object can simultaneously belong to more than one group.
5. Fuzzy clustering: Every object belongs to every cluster with a membership weight that goes between 0:if it absolutely doesn't belong to the cluster and 1:if it absolutely belongs to the cluster.
6. Complete clustering: It performs a hierarchical clustering using a set of dissimilarities on 'n' objects that are being clustered. They tend to find compact clusters of an approximately equal diameter.