WORKSHEET_1_SQL

1. Which of the following is/are DDL commands in SQL?

Create and Alter

2. Which of the following is/are DML commands in SQL?

Select and Delete

3. Full form of SQL is:

Structured Query Language

4. Full form of DDL is:

Data Definition Language

5. DML is:

Data Manipulation Language

6. Which of the following statements can be used to create a table with column B int type and C float type?

Create Table A (B int, C float)

7. Which of the following statements can be used to add a column D (float type) to the table A created above?

Alter Table A ADD COLUMN D float

8. Which of the following statements can be used to drop the column added in the above question?

Alter Table A Drop Column D

9. Which of the following statements can be used to change the data type (from float to int) of the column D of table A created in above questions?

Alter table A Column D float to int

10. Suppose we want to make Column B of Table A as primary key of the table. By which of the following statements we can do it?

Alter Table A Add Constraint Primary Key B

11. What is data-warehouse?

1. A data warehouse is a data management system, designed to enable and support business intelligence (BI) activities and analytics.

2. The large amount of data in data warehouses comes from different sources such as internal applications as marketing, Sales, and finance, customer-facing apps and external partner systems.

3. Data warehouse is the core of the BI system which is built for data analysis and reporting.

4. Data warehouses are designed to give a long-range view of data over time. They trade off transaction volume and instead specialize in data aggregation.

5. A Data Warehouse works as a central repository where information arrives from one or more data sources.

Data flows into a data warehouse from the transactional system and other relational databases.

6. Three main types of Data Warehouses (DWH) are:

Enterprise Data Warehouse (EDW)

Operational Data Store

Data Mart

7. Components of Data warehouse

1. Load manager

2. Warehouse Manager

3. Query Manager

8. Benefits of Data Warehouse

Subject-oriented: They can analyze data about a particular subject or functional area (such as sales)

Integrated: Data warehouses create consistency among different data types from disparate sources

Nonvolatile: Once data is in a data warehouse, it's stable and doesn't change.

Time-variant: Data warehouse analysis looks at change over time.

12. What is the difference between OLTP VS OLAP?

Functionality	OLTP : Manages transactions that modify data in databases.
	OLAP: Used for analytical and reporting purposes.
Source	OLTP: Real-time transactions of organizations.
	OLAP: Data is consolidated from various OLTP databases.
Storage format	OLTP: Tabular form in Relational Databases. OLAP: Multidimensional form in OLAP cubes.
Operation	OLTP: Read and write Read-only OLAP: Read-only
Response time	OLTP: Fast processing since queries are simple. OLAP: Slower than OLTP
Users	OLTP: Executives, Data scientists OLAP: Programmers, Database professionals

13. What are the various characteristics of data-warehouse?

Characteristics of data-warehouse:

1. Subject-oriented: It delivers information about a theme instead of organization's current operations. These themes can be sales, distributions, marketing etc.

It delivers an easy and precise demonstration around particular theme by eliminating data which is not required to make the decisions.

2. Integrated:

The way data is extracted and transformed is uniform, regardless of the original source.

Integration in Data Warehouse means establishing a standard unit of measurement from the different databases for all the similar data.

The data must get stored in a simple and universally acceptable manner within the Data Warehouse.

3. Time-Variant

Data is organized via time-periods (weekly, monthly, annually, etc.)

Time horizon for the data warehouse is quite extensive.

The data collected in a data warehouse is acknowledged over a given period and provides historical information.

4. Non-volatile

A data warehouse is not updated in real-time. It is periodically updated via the uploading of data, protecting it from the influence of momentary change. The data warehouse is non-volatile, meaning that prior data will not be erased when new data are entered into it.

Data is read-only, only updated regularly.

14. What is Star-Schema?

In data warehousing and business intelligence (BI), a star schema is the simplest form of a dimensional model, in which data is organized into facts and dimensions.

A fact is an event that is counted or measured, such as a sale or login.

A dimension contains reference information about the fact, such as date, product, or customer.

A star schema is diagramed by surrounding each fact with its associated dimensions.

Star schemas are optimized for querying large data sets and are used in data warehouses and data marts to support OLAP cubes,

Business intelligence and analytic applications, and ad hoc queries.

Characteristics of Star Schema:

It creates a DE-normalized database that can quickly provide query responses.

It provides a flexible design that can be changed easily or added to throughout the development cycle, and as the database grows.

It provides a parallel in design to how end-users typically think of and use the data.

It reduces the complexity of metadata for both developers and end-users.

15. What do you mean by SETL?

Set Theory as a Language

SETL is a high-level programming language that’s based on the mathematical theory of sets.

It was developed in the early 1970’s by mathematician Professor J. Schwartz.

SETL is an interpreted language with a syntax that is resembles C and in many cases similar to Perl.

In SETL every statement is terminated by a semicolon.

Variable names are case-insensitive and are automatically determined by their last assignment.

Statistic Worksheet_1

1. Bernoulli random variables take (only) the values 1 and 0.

True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Modeling bounded count data

4. Point out the correct statement.

All of the mentioned

5._____ random variables are used to model rates.

Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.

False

7. Which of the following testing is concerned with making decisions using data?

Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

0

9. Which of the following statement is incorrect with respect to outliers?

Outliers cannot conform to the regression relationship

10. What do you understand by the term 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.

Characteristics:

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.

In reality, most pricing distributions are not perfectly normal.

11. How do you handle missing data? What imputation techniques do you recommend?

5 Ways to Handle Missing Values:

1. Deleting Rows

This method commonly used to handle the null values.

Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values.

This method is advised only when there are enough samples in the data set.

2. Replacing With Mean/Median/Mode

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare.

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. But the loss of the data can be

Negated by this method which yields better results compared to removal of rows and columns.

3. Assigning an Unique Category

4. Predicting the Missing Values

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm.

This method may result in better accuracy, unless a missing value is expected to have a very high variance.

5. Using Algorithms Which Support Missing Values

KNN is a machine learning algorithm which works on the principle of distance measure.

This algorithm can be used when there are nulls present in the dataset.

While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values.

Another algorithm which can be used here is Random Forest.

This model produces a robust result because it works well on non-linear and the categorical data.

It adapts to the data structure taking into consideration of the high variance or the bias, producing better results on large datasets.

12. What is A/B testing?

A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

A/B testing allows individuals, teams, and companies to make careful changes to their user experiences while collecting data on the results.

This allows them to construct hypotheses, and to learn better why certain elements of their experiences impact user behavior.

13. Is mean imputation of missing data acceptable practice?

It can be a acceptable practice with certain limitations:

1. A better approach when the data size is small

2. It can prevent data loss which results in removal of the rows and columns.

3. Imputing the approximations adds variance and bias.

14. What is linear regression in statistics?

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables.

The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regressions.

1. Linear regression models are used to show or predict the relationship between two variables or factors.
2. The factor that is being predicted (the factor that the equation solves for) is called the dependent variable

15. What are the various branches of statistics?

Two branches:

Descriptive statistics and Inferential statistics

Descriptive 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 that are so easy to creep into the experiment.

Inferential Statistics

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.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other.

Machine Learning

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram?

4

2. In which of the following cases will K-Means clustering fail to give good results?

2 and 3

3. The most important part of is selecting the variables on which clustering is based.

Formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.

Euclidean distance

5. ----------is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

b) Divisive clustering

6. Which of the following is required by K-means clustering?

Number of clusters

7. The goal of clustering is to-

Classify the data point into different classes

8. Clustering is a-

Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?

K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis-

d) All of the above

12. For clustering, we do not require-

Unlabeled data

13. How is cluster analysis calculated?

Cluster analysis is an exploratory analysis that tries to identify structures within the data.

Cluster analysis is also called segmentation analysis or taxonomy analysis.

More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known.

Because it is exploratory, it does not make any distinction between dependent and independent variables.

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. What is cluster analysis and its types?

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

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: Its simply a division of the set of data objects into non-overlapping clusters such that each objects 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 objects belongs to every cluster with a membership weight that goes between 0: if it absolutely doesn't belong to cluster and

1: if it absolutely belongs to the cluster.

6. Complete clustering: It perform 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.