**SQL**

11. What is data-warehouse?

A data warehouse is a large collection of business data used to help an organization make decisions. . The large amount of data in data warehouses comes from different places such as internal applications such as marketing, sales, and finance; customer-facing apps; and external partner systems, among others. Data warehouses store data from multiple databases, which makes it easier to analyze the data and find meaningful inferences with the dataset provided.

12. What is the difference between OLTP VS OLAP?

|  |  |
| --- | --- |
| **OLTP**(Online transaction processing ) | **OLAP**(Online Analytical Processing) |
| * Administers day to day transaction of an organization. | * Consists of a type of software tools that are used for data analysis for business decisions. |
| * The data is used to perform day to day fundamental operations. | * The data is used in planning, problem solving and decision making. |
| * It reveals a snapshot of present business tasks. | * It provides a multi-dimensional view of different business tasks. |
| * Both read and write operations. | * Only read and rarely write operation. |
| * ATM center is an OLTP application. | * Eg: Netflix movie recommendation system. |
|  |  |

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

* The primary task that a data warehouse is given is mostly around the modeling of data and then analyzing it for different decision making processes that might affect the day to day working of the company as well as shape the long term plans.
* Whenever any new data points are stored in the data warehouse, the previous data is not removed or affected in any way. This property of a data warehouse makes it non-volatile.
* Within a data warehouse, there are multiple sources of data which leads to a distinct set and types of databases. But a data warehouse makes sure that for measuring the data, it maintains a constant unit of measurement.

14. What is Star-Schema??

Star Schema in data warehouse, in which the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The Star Schema data model is the simplest type of Data Warehouse schema.

15. What do you mean by SETL?

SETL is a very-high level language with dynamic typing and dynamic data structures, based on the mathematical notion of set. The very name SETL is an abbreviation of ‘SET Language’. The set-oriented paradigm is based on the assumption that sets are as essential constructions in programming as they are in mathematics.

**STATISTICS**

10. What do you understand by the term Normal Distribution?

Normal Distribution is a bell-shaped frequency distribution curve which helps describe all the possible values a random variable can take within a given range with most of the distribution area is in the middle and few are in the tails, at the extremes.

* The mean, median, and mode of this distribution are all equal.
* Half of the values are to the left of the center and the other half to the right.
* Value under the standard curve is always one.

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

There are various methods to handle missing data, the below are some methods:

* Deleting Rows: When the missing value is a lot in a column more than 80%, we delete the particular column .This method is advised only when there are enough samples in the data set

## Replacing With Mean/Median/Mode: While dealing with continuous data, we should consider mean or median and if the data has a categorical value we need to replace it with mode.

It is best to replace it with median, as median considers the middle value and it won’t be much affected by the outliers present in the dataset whereas when imputing mean it will be affected by the outliers present.

12. What is A/B testing?

An AB testing is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a significant relationship or not between the two.

13. Is mean imputation of missing data acceptable practice?

We can replace missing data with mean only when the data is symmetrical which suggests that there will be no outliers present in the dataset as outlier’s data points will have significant impact on the mean. So it is only acceptable if the data is normally distributed with no skewness.

14. What is linear regression in statistics?

Linear regression is a linear approach to modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). The case of one explanatory variable is called simple linear regression.

X is the independent variable and y is the dependent variable, with the change in x, y is also affected.

Linear regression consists of finding the best-fitting straight line through the points

15. What are the 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).

* [Descriptive statistics](https://explorable.com/descriptive-statistics) deals with the presentation and collection of data. This is usually the first part of a statistical analysis. Descriptive statistics are used in use to do various kinds of [analysis](https://statanalytica.com/data-analysis-assignment-help) on different studies.

Descriptive statistics have three parts:

* The distribution concerns the frequency of each value.
* The [central tendency](https://www.scribbr.com/statistics/central-tendency/)concerns the averages of the values.
* The [variability](https://www.scribbr.com/statistics/variability/)or dispersion concerns how spread out the values are.
* The inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population. Inference statistics often speak in terms of [probability](https://statanalytica.com/probability-assignment-help) by using descriptive statistics. Besides, these techniques are used primarily by a statistician for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

Types of Inferential Statistics:

* One sample test of difference/One sample hypothesis test
* Confidence Interval
* Contingency Tables and Chi Square Statistic
* T-test or Anova
* Pearson Correlation
* Bi-variate Regression
* Multi-variate Regression

13. How is cluster analysis calculated?

1: Select the number of clusters you want to identify in your data. This is the “K” in K-means clustering. In case, we will select K=3. That is to say, we want to identify 3 clusters.

2: Randomly select 3 distinct data points.

3: Measure the distance between the first point and the three initial clusters.

4: Assign the first point to the nearest cluster. In case, the nearest cluster is the blue cluster. Thus, we will continue the same process for the next points. Once all the points are in clusters, we will move on to the next step.

5: Calculate the mean of the cluster. Then again, measure and cluster using the mean values. If the clustering did not change at all during the last iterations, then the process is complete

14. How is cluster quality measured?

Below are some methods to measure the quality of clusters formed.

**Cluster homogeneity.** This requires that the more pure the clusters in a clustering are, the better the clustering.

**Cluster completeness.** This is the counterpart of cluster homogeneity. Cluster completeness requires that for a clustering, if any two objects belong to the same category according to ground truth, then they should be assigned to the same cluster

**Rag bag.**  The rag bag criterion states that putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag.

**Small cluster preservation.** The small cluster preservation criterion states that splitting a small category into pieces is more harmful than splitting a large category into pieces.

15. What are cluster analysis and its types?

Clustering is a method of unsupervised learning. It is a common technique for statistical data analysis .This method draws references from datasets as a part of input data without labeled responses.

Hierarchical clustering: This technique includes the process of forming clusters that have a definite ordering from top to bottom. For instance, all books are arranged in bookshelves, are organized in a hierarchy or alphabetical order. Furthermore, there are two types of hierarchical clustering, Divisive and Agglomerative

* Divisive method: In this technique, all the observation is allotted to a single cluster and then divided to 2 least similar clusters.  Finally, we continue several times on each cluster until there is one cluster for each observance.
* Agglomerative method: In this technique, each observation is allotted to its own cluster and then the similarity is computed. This analogy is compared between each of these clusters. Finally, join the two most similar clusters and repeat this until there is only a single cluster left.

In centroid -based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to *k*, k-means clustering gives a formal definition as an optimization problem. In these models, the no. of clusters required at the end has to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.

Distribution models: These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from over fitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.

Density Models: These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assigns the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.

Constraint-based Method: In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.