**WORKSHEET 1 SQL**

Answer No. 1: A & D

Answer No. 2: A, B&C

Answer No. 3: B

Answer No. 4: B

Answer No. 5: A

Answer No. 6: C

Answer No. 7: B

Answer No. 8: B

Answer No. 9: B

Answer No. 10: A

Answer No. 11:  **Data Warehousing** 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 data warehouse is the core of the BI system which is built for data analysis and reporting.

It is a blend of both technologies and components which aids the strategic use of data. It is an electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

Answer No. 12: Difference between OLTP VS OLAP

|  |  |
| --- | --- |
| OLTP | OLAP |
| It is characterized by large numbers of short online transactions. | It is characterized by a large volume of data. |
| OLTP is an online database modifying system. | OLAP is an online database query management system. |
| OLTP uses traditional DBMS. | OLAP uses the data warehouse. |
| Data in the OLTP database is always detailed and organized | Data in OLAP process might not be organized. |
| It helps to control and run fundamental business tasks. | It helps with planning, problem-solving, and decision support. |
| OLTP is designed to have fast response time, low data redundancy and is normalized. | In OLAP a data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database |
| OLTP is market orientated process | It is customer-oriented process. |
| Users in this are the clerk, DBA and data base professionals. | Used by Data knowledge users like workers, managers, and CEO. |

Answer No. 13: Various characteristics of data-warehouse are as follows:

Subject-Oriented: - The warehouse organizes data around the essential subjects of the business rather than (customers and products) around applications such as inventory management or order processing.

Integrated: In Data Warehouse, integration means the establishment of a common unit of measure for all similar data from the dissimilar database. The data also needs to be stored in the Datawarehouse in common and universally acceptable manner. Data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. Moreover, it must keep consistent naming conventions, format, and coding.

Time-Variant: The time horizon for data warehouse is quite extensive compared with operational systems. The data collected in a data warehouse is recognized with a particular period and offers information from the historical point of view. It contains an element of time, explicitly or implicitly.

One such place where Datawarehouse data display time variance is in in the structure of the record key. Every primary key contained with the DW should have either implicitly or explicitly an element of time. Like the day, week month, etc.

Another aspect of time variance is that once data is inserted in the warehouse, it can't be updated or changed.

Non-volatile: The previous data is not erased when new data is entered in it. Data is read-only and periodically refreshed. This also helps to analyse historical data and understand what & when happened. It does not require transaction process, recovery and concurrency control mechanisms.

Answer No. 14: The star schema is the simplest form of a dimensional model used in business intelligence and data warehousing wherein data is arranged in dimensions and facts. In the star schema, there is a single fact table, which is usually expressed in the third normal form (3NF), and multiple de-normalized dimension tables connected to it, radiating out like the points of a star. The star schema has been optimized for querying large data sets and is generally used in data marts and warehouses in order to support OLAP cubes, ad hoc queries, analytic applications and business intelligence.  
  
The fact tables in a star schema usually have two columns: the first is for the foreign keys pointing to the dimension tables, and the second is for the measures that contain numeric facts, hence, the name fact table. The dimension tables are actually structuring that are usually composed of multiple hierarchies that categorize data.

Answer No. 15: **SETL** (SET Language) is a [very high-level programming language](https://en.wikipedia.org/wiki/Very_high-level_programming_language) based on the mathematical [theory of sets](https://en.wikipedia.org/wiki/Set_theory).

SETL provides two basic aggregate data types: *unordered sets*, and *sequences* (the latter also called *tuples*). The elements of sets and tuples can be of any arbitrary type, including sets and tuples themselves. Primitive operations in SETL include set membership, union, intersection, and power set construction, among others.

SETL provides quantified Boolean expressions constructed using the [universal](https://en.wikipedia.org/wiki/Universal_quantifier) and [existential quantifiers](https://en.wikipedia.org/wiki/Existential_quantifier) of [first-order predicate logic](https://en.wikipedia.org/wiki/First-order_predicate_logic).

**STATISTICS WORKSHEET -1**

Answer No. 1: A

Answer No. 2: A

Answer No. 3: B

Answer No. 4: D

Answer No. 5: C

Answer No. 6: B

Answer No. 7: B

Answer No. 8: A

Answer No. 9: C

Answer No. 10: Normal distribution, also known as the Gaussian distribution, is a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. Normal distribution will appear as a [bell curve](https://www.investopedia.com/terms/b/bell-curve.asp). The standard normal distribution has two parameters: the mean and the [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp). The normal distribution model is motivated by the [Central Limit Theorem](https://www.investopedia.com/terms/c/central_limit_theorem.asp).

Answer No. 11: The most common problems I have faced in Data Cleaning/Exploratory Analysis is handling the missing values. Firstly, understand that there is No good way to deal with missing data.

Ways to deal missing values in the dataset:

1. **Deleting Rows with missing values**: - Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of rows as null, then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.
2. **Impute missing values with Mean/Median**: - Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.
3. **Impute using most frequent for categorical variable**: - When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large, then it can be replaced with a new category.
4. **Impute using k-NN: -** The k nearest neighbours is an algorithm that is used for simple classification. The algorithm uses ‘**feature similarity**’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This can be very useful in making predictions about the missing values by finding the k’s closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood.
5. **Impute using miss forest**: - It is non-standardd, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.
6. **Impute using Deep Learning (Datawig)**: - This method works very well with categorical and non-numerical features. It is a library that learns Machine Learning models using Deep Neural Networks to impute missing values in a dataframe. It also supports both CPU and GPU for training.

In conclusion, there is no prefect way to compensate for the missing values in a dataset because each strategy can perform better for certain datasets and missing data types but may perform much worse on other types of datasets. There are set of rules to decide which strategy to use for particular types of missing values, but beyond that you should experiment and check which model works best for our dataset.

Answer No. 12: 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. A/B testing is one of the most prominent and widely used statistical tools. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. In which population refers to all the customers buying your product, while the sample are those number of customers that participated in the test.

Answer No. 13: Bad practice in general.

If just estimating means, mean imputation preserves the mean of the observed data.

Leads to an underestimate of the standard deviation.

Distorts relationships between variables by “pulling” estimates of the correlation toward zero.

Answer No. 14: Linear regression is the next step up after correlation*.* **Linear regression**quantifies the relationship between one or more predictor variables and one outcome variable. Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).  Linear regression is also known asmultiple regression*,* multivariate regression*,* ordinary least squares (OLS)*, and*regression*.* This post will show you examples of linear regression, including an example of simple linear regression and an example of multiple linear regression*.*

Answer No. 15: The two main branches of statistics are descriptive statistics and inferential statistics. Both are employed in scientific analysis of data.

Descriptive Statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis.

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**

Answer No. 1: B

Answer No. 2: D

Answer No. 3: D

Answer No. 4: A

Answer No. 5: B

Answer No. 6: D

Answer No. 7: D

Answer No. 8: B

Answer No. 9: A

Answer No. 10: A

Answer No. 11: D

Answer No. 12: A

Answer No. 13: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory [data mining](https://en.wikipedia.org/wiki/Data_mining), and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [image analysis](https://en.wikipedia.org/wiki/Image_analysis), [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics), [data compression](https://en.wikipedia.org/wiki/Data_compression), [computer graphics](https://en.wikipedia.org/wiki/Computer_graphics) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning).

Cluster analysis itself is not one specific [algorithm](https://en.wikipedia.org/wiki/Algorithm), but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small [distances](https://en.wikipedia.org/wiki/Distance_function) between cluster members, dense areas of the data space, intervals or particular [statistical distributions](https://en.wikipedia.org/wiki/Statistical_distribution). Clustering can therefore be formulated as a [multi-objective optimization](https://en.wikipedia.org/wiki/Multi-objective_optimization) problem. The appropriate clustering algorithm and parameter settings (including parameters such as the [distance function](https://en.wikipedia.org/wiki/Metric_(mathematics)) to use, a density threshold or the number of expected clusters) depend on the individual [data set](https://en.wikipedia.org/wiki/Data_set) and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of [knowledge discovery](https://en.wikipedia.org/wiki/Knowledge_discovery) or interactive multi-objective optimization that involves trial and failure.

Answer No. 14: Cluster quality is measured by two methods first is elbow method. The **elbow method** is a [heuristic](https://en.wikipedia.org/wiki/Heuristic) used in [determining the number of clusters in a data set](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set). The method consists of plotting the [explained variation](https://en.wikipedia.org/wiki/Explained_variation) as a function of the number of clusters, and picking the [elbow of the curve](https://en.wikipedia.org/wiki/Elbow_of_the_curve) as the number of clusters to use

**Silhouette Method**is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. It is also better than elbow method. The silhouette method computes silhouette coefficients of each point that measure how much a point is like its own cluster compared to other clusters. by providing a **succinct graphical representation** of how well each object has been classified.

Answer No. 15: 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.

Four basic types of cluster analysis used in data science. These types are Centroid clustering, Density clustering, Distribution clustering and Connectivity clustering.

In centroid clustering the algorithm will start by randomly selecting centroids to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all interstation’s, resulting in optimized clusters.

In Density clustering, it groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the denser the data points the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it.

In Distribution clustering, it identifies the probability that a point belongs to a cluster. Around each possible centroid the algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions the algorithm optimizes the characteristics of the distributions to best represent the data.

In connectivity clustering, the iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.