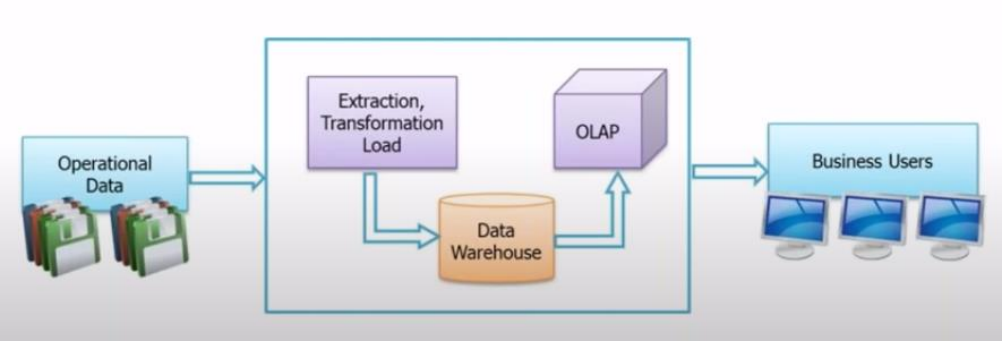
WORKSHEET 1 SQL

1. Create and Alter
2. Update, Delete and Select
3. Structured Query Language
4. Data Definition Language
5. Data Manipulation Language
6. Create Table A (B int, C float)
7. Alter Table A ADD COLUMN D float
8. Alter Table A Drop Column D
9. Alter Table A Column D float to int
10. None of them

11. Data-warehouse
    - Data-warehouse is a central location where consolidated data from multiple locations (databases) are stored
    - It functions on the basis of OLAP
    - Data-warehouse is maintained separately from an organization's operational database
    - End users access it when ever any information is needed



The Data collected from various sources and stored in various data bases cannot be directly visualized.

So the data, first need to be integrated and then processed before visualization takes place.

Advantages of Data-warehouse:

Strategic questions can be answered by studying trends

Data warehousing is faster and more accurate

12. OLTP (Online Transactional Processing)
    - It is used to manage very large number of online transactions (running business).
    - Contains current data
    - Based on entity relationship model
    - Provided primitive and highly detailed data
    - Used for writing into the database
    - Database size ranges from 100mb to 1gb
    - Fast provides high performance
    - Number of records accessed in tens
    - Ex: All bank transaction made by a customer

    OLAP (Online Analytical Processing)

    - It is used for data analysis
    - Contains historical data
    - Based on Star, Snowflake and Fact constellation schema
    - Provides summarized and consolidated data
    - Used for reading data from data-warehouse
    - Data-warehouse size ranges from 100gb to 1tb
    - Highly flexible but not fast
    - Number of records accessed in millions
    - Ex: Bank transactions made by a customer at a particular time

13. Characteristics of Data-warehouse
    - Subject Oriented

      Data is categorized and stored by business subject rather than by applications. Data-warehouse concentrate on emphasized modeling and analyzing decision-making data.

    - Integrated

      Data on a given subject is collected from various sources and stored in a single place. The data must also get stored in a universally acceptable manner within the data-warehouse. It must also keep the naming conventions, format and coding consistent.

    - Time variant

      The data collected in a data warehouse is acknowledged over a given period and provides historical information. It contains a temporal element, either explicitly or implicitly.

- Non-volatile

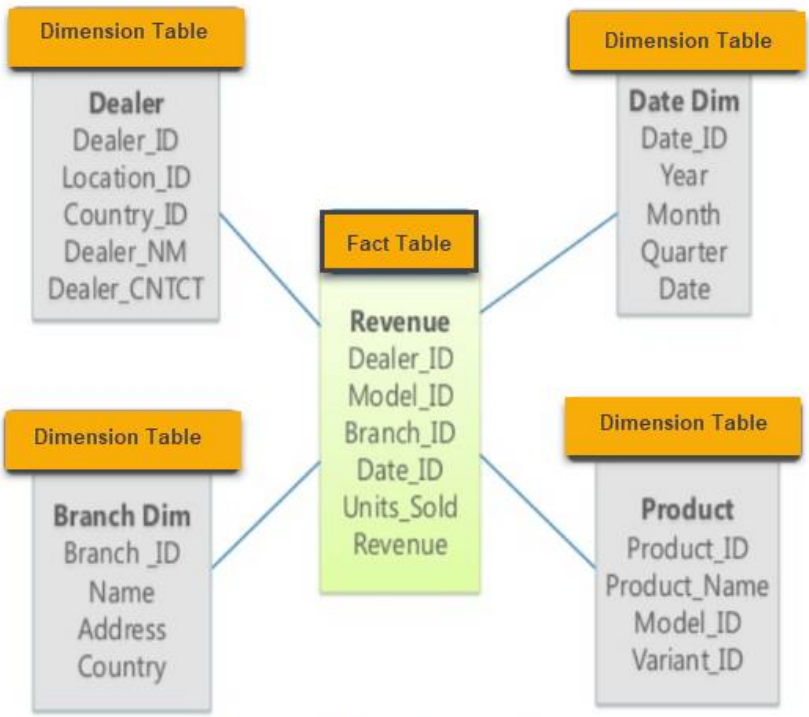
  Data will not be erased when new data are entered into it. Typically, the data in the data-warehouse is not updated or deleted. It also assists in analyzing historical data and understanding what and when it happened.

14. Star Schema
    - Star schema is a relational database schema for representing multidimensional data.
    - It is known as star schema as its structure resembles a star
    - The center of the schema consists of large Fact table and it point towards dimension tables

    Advantages of Star Schema

    - It is very simple to implement
    - Simple form of queries is used
    - It optimizes the data base navigation

    Eg: In the following star schema example, the fact table is at the center which contains keys to every dimension table like Dealer_ID, Model_ID, Date_ID, Product_ID, Branch_ID and other attributes like Units_Sold and Revenue



Example of Star Schema Diagram

Machine Learning Worksheet

1. 4
2. 1,2 and 4
3. formulating the clustering problem
4. Euclidean distance
5. Divisive clustering
6. All of the answers are correct
7. Divide the data points into groups
8. Unsupervised learning
9. K-means clustering
10. K-means clustering algorithm
11. All of the above
12. Labeled data

13. Cluster Analysis

Cluster analysis groups the similar data in same group. 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

Greater the similarity within the group and greater difference between the group, more distinct the clustering

Types of clustering methods

- Partitioning method

  Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k<=n. It means it will classify the data into k groups
  For a given number of partitions (says k), the partitioning method will create an initial partitioning.
  Then it uses the iterative location technique to improve the partitioning by moving objects from one group to other.

- Hierarchical method

  This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here-
  Agglomerative Approach
  Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

This approach is also known as top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, ie, once a merging or splitting is done, it can never be undone

- Density-based method

  This method is based on the notion of the density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, ie, for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points

- Grid-based method

  In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

  Advantages

  The major advantage of this method is fast processing time.
  It is dependent only on the number of cells in each dimension in the quantized space.

- Model-based method

  In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

  This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

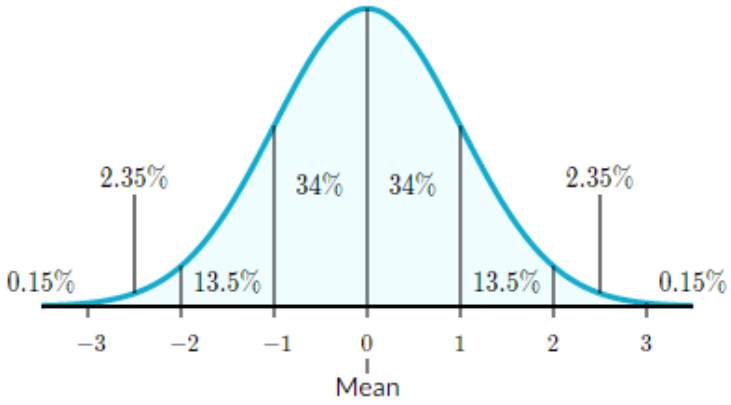
Statistic Worksheet 1

1. True
2. Central Limit Theorem
3. Modeling bounded count data
4. All of the mentioned
5. Poisson
6. False
7. Hypothesis
8. 0
9. Outliers cannot perform 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.

Normal distribution has the following features.



- Symmetric bell shape
- Mean and median are equal: both located at the center of the distribution.
- 68% of the data falls within +/- 1 standard deviation of the mean
- 95% of the data falls within +/- 2 standard deviation of the mean
- 99.7% of the data falls within +/- 3 standard deviation of the mean

11. Handling missing data

In pandas I use isnull() to find columns of data with missing or corrupted data. If I find missing data in dataset, I will use these imputation techniques.

- Drop those rows or columns. If not, many rows contain missing data, dropping those rows doesn't bias the data. And it's a reasonable thing to do.
- replace them the with another value such as 0, mean, median, mode. Replacing missing value with 0 if the values are relatively closer to that or leaving other

than nan value. Replacing missing values with mean value from the rest of the column (columns, not rows!). median may be better choice if outliers are present.

- Create a separate model to handle missing value

12. A/B tests is used widely in industry to make decisions. In the simplest form there are two variants control A and control B. Typically the control group uses the existing features while treatment group uses new features.
    A/B testing allows tech companies to evaluate a feature with a subset of users to infer how it may be received

13. No its not acceptable practice. The main problems of mean Imputation are
    - Mean imputation does not preserve the relationships among variables. If estimating means and if the data are missing completely at random, mean imputation will not bias the parameter estimate
    - If estimating means mean imputation preserves the mean of the observed data will leads to underestimate the standard deviation

14. Linear Regression

    Linear regression expresses the relationship between one or more predictor variables(s) and one outcome variable. Linear regression is commonly used for predictive analysis and modeling. 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 = dependent variable score, c = constant, b= regression coefficient and x = score on the independent variable.

    <u>Types of Linear Regression</u>

    1. Simple linear regression
       1 dependent variable, 1 independent variable
    2. Multiple linear regression
       1 dependent variable, 2+ independent variable
    3. Logistic regression
       1 dependent variable, 2+ independent variable(s)
    4. Ordinal regression
       1 dependent variable(ordinal), 1+ independent variable(s)
    5. Multinomial regression
       1 dependent variable(nominal), 1+ independent variable(s)
    6. Discriminant analysis
       1 dependent variable(nominal), 1+ independent variable(s)

15. Branches of Statistics

- Descriptive Statistics

  Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs, and diagrams and finding averages and other measures which would describe the data.

  Eg: Industrial statistics, population statistics, trade statistics etc. Business makes use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

- Inferential Statistics

  Inferential statistics deals with techniques used for the analysis of data, making estimates and drawing, conclusions from limited information obtained through sampling and testing the reliability of the estimates.

  Eg: Suppose we want to have an idea about the percentage of illiterate population of our country. We take sample from population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enable us to make inferences about the population proportion. This study belongs to inferential statistics.