MACHINE LEARNING WORKSHEET-1

Q1. b) 4
Q2. d) 1, 2 and 4
Q3. d) formulating the clustering problem
Q4. a) Euclidean distance
Q5. b) Divisive clustering
Q6. b) Number of clusters
Q7. a) Divide the data points into groups
Q8. b) Unsupervised learning
Q9. d) All of the above
Q10. a) K-means clustering algorithm
Q11. d) All of the above
Q12. a) Labeled data
Q13. How is cluster analysis calculated?

Ans. Steps to calculate cluster analysis are:-

Step 1 – Hypothesis building : This is the most crucial step of the whole exercise. Try to identify all possible variables that can help segment the portfolio regardless of its availability. Lets try to come up with a list for this example.

a. Customer balance with bank X

b. Number of transaction done in last 1/3/6/12 months

c. Balance change in last 1/3/6/12 months

d. Demographics of the customer

e. Customer total balance with all US banks

The list is just for illustrative purpose. In real scenario this list will be much longer.

Step 2 – Initial shortlist of variable : Once we have all possible variable, start selecting variable as per the data availability. Lets say, for the current example we have only data for Customer balance with bank X and Customer total balance with all US banks (total balance)

Step 3 – Visualize the data : It is very important to know the population spread across the selected variable before starting any analysis. For the current scenario, the exercise becomes simpler as the number of selected variables is only 2. Plot a scatter plot between total balance and Bank X balance (origin taken as mean of both the variables). This visualization helps me to identify clusters which I can expect after the final analysis. Here, we can see there are four clear clusters in four quadrants. We can expect the same result in the final solution.

Step 4 – Data cleaning: Cluster analysis is very sensitive to outliers. It is very important to clean data on all variables taken into consideration. There are two industry standard ways to do this exercise :-

1. Remove the outliers : (Not recommended in case the total data-points are low in number) We remove the data-points beyond mean +/- 3*standard deviation.

2. Capping and flouring of variables : (Recommended approach) We cap and flour all data-points at 1 and 99 percentile. Lets use the second approach for this case.

Step 4 – Variable clustering : This step is performed to cluster variables capturing similar attributes in data. And choosing only one variable from each variable cluster will not drop the separation drastically compared to considering all variables. Remember, the idea is to take minimum number of variables to justify the separation to make the analysis easier and less time consuming. You can simply use Proc VARCLUS to generate these clusters.

Step 5 – Clustering : We can use any of the two technique discussed in the article depending on the number of observation. k-means is used for a bigger samples. Run a proc fastclus with k=4 (which is apparent from the visualization). As we can see, the algorithm found 4 clusters which were already apparent in the visualization. In most business cases the number of variables will be much larger and such visualization won't be possible and hence

Step 6 – Convergence of clusters : A good cluster analysis has all clusters with population between 5-30% of the overall base. Say, my total number of customer for bank X is 10000. The minimum and maximum size of any cluster should be 500 and 3000. If any of the cluster is beyond the limit than repeat the procedure with additional number of variables. We will discuss in detail about other convergence criterion in the next article.

Step 7 – Profiling of the clusters : After validating the convergence of cluster analysis, we need to identify behaviuor of each cluster. Lets say we map age and income to each of the four clusters and get following results.

Now is the time to build story around each cluster. Lets take any two cluster and analyze. Cluster 1 : (High Potential Low balance customer) These customers do have high balance in aggregate but low balance with bank X. Hence, they are high potential customer with low current balance. Also the average salary is on a higher side which validates our hypothesis of customer being high potential.

Cluster 3 : (High Potential high balance customers) Even though the salary and total balance in aggregate is on a lower side, we see a lower average age. This indicates that the customer has a high potential to increase their balance with bank X.

Q14. How is cluster quality measured?

Ans. Three important factors by which clustering can be evaluated are:- *1. Clustering tendency 2. Number of clusters (**k**) 3. Clustering quality*

1. Clustering tendency:- Before evaluating the clustering performance, making sure that data set we are working has clustering tendency and does not contain uniformly distributed points is very important. If the data does not contain clustering tendency, then clusters identified by any state of the art clustering algorithms may be irrelevant. Non-uniform distribution of points in data set becomes important in clustering. To solve this, Hopkins test, a statistical test for spatial randomness of a variable, can be used to measure the probability of data points generated by uniform data distribution.

Null Hypothesis *(Ho): Data points are generated by uniform distribution (implying no meaningful clusters)*
Alternate Hypothesis *(Ha): Data points are generated by random data points (presence of clusters)*
If H>0.5, null hypothesis can be rejected and it is very much likely that data contains clusters. If H is more close to 0, then data set doesn't have clustering tendency.

2. Number of Optimal Clusters (k):-Some of the clustering algorithms like K-means, require number of clusters, *k*, as clustering parameter. Getting the optimal number of clusters is very significant in the analysis. If *k* is too high, each point will broadly start representing a cluster and if *k* is too low, then data points are incorrectly clustered. Finding the optimal number of clusters leads to granularity in clustering. There is no definitive answer for finding right number of cluster as it depends upon (a) Distribution shape (b) scale in the data set (c) clustering resolution required by user. Although finding number of clusters is a very subjective problem. There are two major approaches to find optimal number of clusters: (1)Domain knowledge (2)Data driven approach
Domain knowledge - Domain knowledge might give some prior knowledge on finding number of clusters. For example, in case of clustering iris data set, if we have the prior knowledge of species, then k = 3. Domain knowledge driven k value gives more relevant insights. Data driven approach - If the domain knowledge is not available, mathematical methods help in finding out right number of clusters.

Empirical Method:- A simple empirical method of finding number of clusters is Square root of N/2 where N is total number of data points, so that each cluster contains square root of 2 * N

Elbow method:- Within-cluster variance is a measure of compactness of the cluster. Lower the value of within cluster variance, higher the compactness of cluster formed.

Sum of within-cluster variance, *W*, is calculated for clustering analyses done with different values of k. *W* is a cumulative measure how good the points are clustered in the analysis. Plotting the *k* values and their corresponding sum of within-cluster variance helps in finding the number of clusters. Plot shows that number of optimal clusters = 4. Initially, Error measure (within-cluster variance) decreases with increase in cluster number. After a particular point, k=4 , Error measure starts flattening. Cluster number corresponding to that particular point, k=4, should be considered as optimal number of clusters.

Statistical approach:- Gap statistic is a powerful statistical method to find the optimal number of clusters, *k*. Similar to Elbow method, sum of within-cluster (intra-cluster) variance is

calculated for different values of k. Then Random data points from reference null distribution are generated and Sum of within-cluster variance is calculated for the clustering done for different values of k.

In Simpler words, Sum-of-within-Cluster variance of original data set for different values of k to Sum-of-within-cluster variance of reference data set (null reference data set of uniform distribution) of corresponding values of k is compared to find the ideal *k* value where 'deviation' or 'Gap' between two is highest. As Gap statistic quantifies this deviation, More the Gap statistic means more the deviation. Cluster number with maximum Gap statistic value corresponds to optimal number of cluster.

3.Clustering quality:- Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. Ideal clustering is characterised by minimal intra cluster distance and maximal inter cluster distance. There are majorly two types of measures to assess the clustering performance.

(i) *Extrinsic Measures* which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) *Intrinsic Measures* that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

Q15. What is cluster analysis and its types?

Ans. Cluster analysis can be a powerful data-mining tool for any organisation that needs to identify discrete groups of customers, sales transactions, or other types of behaviours and things. For example, insurance providers use cluster analysis to detect fraudulent claims, and in banks it is used for credit scoring. The objective of cluster analysis is to find similar groups of subjects, where "similarity" between each pair of subjects means some global measure over the whole set of characteristics. In this article we discuss various methods of clustering and the key role that distance plays as measures of the proximity of pairs of points.

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters. There are four basic types of cluster analysis used in data science. These types are Centroid Clustering, Density Clustering Distribution Clustering, and Connectivity Clustering.

1. Centroid Clustering:-

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you're a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.

The algorithm will start by randomly selecting centroids (cluster centres) 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 iterations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters. In this example, the data set will be segmented into customers who are own dogs and cats.

2. Density Clustering:-

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense 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. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

3. Distribution Clustering:-

Distribution clustering 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.

These maps look a lot like targets at an archery range. In the event that a data point hits the bulls eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bulls eye represents lessening percentage or certainty. Distribution clustering is a great technique to assign outliers to clusters, whereas density clustering will not assign an outlier to a cluster.

4. Connectivity Clustering:-

Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. 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.

SQL WORKSHEET-1

Q1. A) Create, D) ALTER
Q2. A) Update, B) Delete
Q3. B) Structured Query Language
Q4. B) Data Definition Language
Q5. A) Data Manipulation Language
Q6. C) Create Table A (B int, C float)
Q7. B) Alter Table A ADD COLUMN D float
Q8. B) Alter Table A Drop Column D
Q9. B) Alter Table A Alter Column D int
Q10. A) Alter Table A Add Constraint Primary Key B
Q11. What is data-warehouse?

Ans. A Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyse 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 technologies and components which aids the strategic use of data. It is 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. Data warehouse system is also known by the following name:

- Decision Support System (DSS)
- Executive Information System
- Management Information System
- Business Intelligence Solution
- Analytic Application
- Data Warehouse

How Datawarehouse works?

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. Data may be:

1. Structured
2. Semi-structured
3. Unstructured data

The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. A

data warehouse merges information coming from different sources into one comprehensive database.

Types of Data Warehouse

Three main types of Data Warehouses (DWH) are:

1. Enterprise Data Warehouse (EDW): Enterprise Data Warehouse (EDW) is a centralized warehouse. It provides decision support service across the enterprise. It offers a unified approach for organizing and representing data. It also provide the ability to classify data according to the subject and give access according to those divisions.

2. Operational Data Store: Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.

3. Data Mart: A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

Data Warehouse Tools

There are many Data Warehousing tools are available in the market. Here, are some most prominent one:

1. MarkLogic: MarkLogic is useful data warehousing solution that makes data integration easier and faster using an array of enterprise features. This tool helps to perform very complex search operations. It can query different types of data like documents, relationships, and metadata.

https://www.marklogic.com/product/getting-started/

2. Oracle: Oracle is the industry-leading database. It offers a wide range of choice of data warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.

https://www.oracle.com/index.html

3. Amazon RedShift**:** Amazon Redshift is Data warehouse tool. It is a simple and cost-effective tool to analyse all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data, using the technique of query optimization.

https://aws.amazon.com/redshift/?nc2=h_m1

Q12. What is the difference between OLTP VS OLAP?

Ans. What is OLAP?

Online Analytical Processing, a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time. The primary objective is data analysis and not data processing**.**

What is OLTP?

Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization. The primary objective is data processing and not data analysis.

KEY DIFFERENCE between OLTP and OLAP:

- Online Analytical Processing (OLAP) is a category of software tools that analyze data stored in a database whereas Online transaction processing (OLTP) supports transaction-oriented applications in a 3-tier architecture.
- OLAP creates a single platform for all type of business analysis needs which includes planning, budgeting, forecasting, and analysis while OLTP is useful to administer day to day transactions of an organization.
- OLAP is characterized by a large volume of data while OLTP is characterized by large numbers of short online transactions.
- In OLAP, data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database whereas OLTP uses traditional DBMS.

Parameters	OLTP	OLAP
Process	It is an online transactional system. It manages database modification.	OLAP is an online analysis and data retrieving pr
Characteristic	It is characterized by large numbers of short online transactions.	It is characterized by a large volume of data.
Functionality	OLTP is an online database modifying system.	OLAP is an online database query management system.
Method	OLTP uses traditional DBMS.	OLAP uses the data warehouse.
Query	Insert, Update, and Delete information from the database.	Mostly select operations
Table	Tables in OLTP database are normalized.	Tables in OLAP database are not normalized.
Source	OLTP and its transactions are the sources of data.	Different OLTP databases become the source of for OLAP.
Data Integrity	OLTP database must maintain data integrity constraint.	OLAP database does not get frequently modified Hence, data integrity is not an issue.
Response time	It's response time is in millisecond.	Response time in seconds to minutes.
Data quality	The data in the OLTP database is always detailed and organized.	The data in OLAP process might not be organized
Usefulness	It helps to control and run fundamental business tasks.	It helps with planning, problem-solving, and deci support.

Parameters	OLTP	OLAP
Operation	Allow read/write operations.	Only read and rarely write.
Audience	It is a market orientated process.	It is a customer orientated process.
Query Type	Queries in this process are standardized and simple.	Complex queries involving aggregations.
Back-up	Complete backup of the data combined with incremental backups.	OLAP only need a backup from time to time. Bac not important compared to OLTP
Design	DB design is application oriented. Example: Database design changes with industry like Retail, Airline, Banking, etc.	DB design is subject oriented. Example: Database design changes with subjects like sales, marketin purchasing, etc.
User type	It is used by Data critical users like clerk, DBA & Data Base professionals.	Used by Data knowledge users like workers, managers, and CEO.
Purpose	Designed for real time business operations.	Designed for analysis of business measures by category and attributes.
Performance metric	Transaction throughput is the performance metric	Query throughput is the performance metric.
Number of users	This kind of Database users allows thousands of users.	This kind of Database allows only hundreds of us
Productivity	It helps to Increase user's self-service and productivity	Help to Increase productivity of the business ana
Challenge	Data Warehouses historically have been a development project which may prove costly to build.	An OLAP cube is not an open SQL server data warehouse. Therefore, technical knowledge and experience is essential to manage the OLAP serv
Process	It provides fast result for daily used data.	It ensures that response to the query is quicker consistently.
Characteristic	It is easy to create and maintain.	It lets the user create a view with the help of a spreadsheet.
Style	OLTP is designed to have fast response time, low data redundancy and is normalized.	A data warehouse is created uniquely so that it c integrate different data sources for building a consolidated database

Q13. What are the various characteristics of data-warehouse?

Ans. General stages of Data Warehouse

Earlier, organizations started relatively simple use of data warehousing. However, over time, more sophisticated use of data warehousing begun. The following are general stages of use of the data warehouse (DWH):

1. Offline Operational Database: In this stage, data is just copied from an operational system to another server. In this way, loading, processing, and reporting of the copied data do not impact the operational system's performance.

2. Offline Data Warehouse: Data in the Data warehouse is regularly updated from the Operational Database. The data in Data warehouse is mapped and transformed to meet the Data warehouse objectives.

3. Real time Data Warehouse: In this stage, Data warehouses are updated whenever any transaction takes place in operational database. For example, Airline or railway booking system.

4. Integrated Data Warehouse**:** In this stage, Data Warehouses are updated continuously when the operational system performs a transaction. The Data warehouse then generates transactions which are passed back to the operational system.

Components of Data warehouse:

Four components of Data Warehouses are:

1. Load manager: Load manager is also called the front component. It performs with all the operations associated with the extraction and load of data into the warehouse. These operations include transformations to prepare the data for entering into the Data warehouse.

2. Warehouse Manager: Warehouse manager performs operations associated with the management of the data in the warehouse. It performs operations like analysis of data to ensure consistency, creation of indexes and views, generation of denormalisation and aggregations, transformation and merging of source data and archiving and baking-up data.

3. Query Manager: Query manager is also known as backend component. It performs all the operation operations related to the management of user queries. The operations of this Data warehouse components are direct queries to the appropriate tables for scheduling the execution of queries.

4. End-user access tools: This is categorized into five different groups like:-1. Data Reporting 2. Query Tools 3. Application development tools 4. EIS tools, 5. OLAP tools and data mining tools.

Who needs Data warehouse?

DWH (Data warehouse) is needed for all types of users like:

- Decision makers who rely on mass amount of data
- Users who use customized, complex processes to obtain information from multiple data sources.
- It is also used by the people who want simple technology to access the data
- It also essential for those people who want a systematic approach for making decisions.
- If the user wants fast performance on a huge amount of data which is a necessity for reports, grids or charts, then Data warehouse proves useful.
- Data warehouse is a first step If you want to discover 'hidden patterns' of data-flows and groupings.

Steps to Implement Data Warehouse

The best way to address the business risk associated with a Data warehouse implementation is to employ a three-prong strategy as below

1. Enterprise strategy: Here we identify technical including current architecture and tools. We also identify facts, dimensions, and attributes. Data mapping and transformation is also passed.
2. Phased delivery: Data warehouse implementation should be phased based on subject areas. Related business entities like booking and billing should be first implemented and then integrated with each other.
3. Iterative Prototyping: Rather than a big bang approach to implementation, the Data warehouse should be developed and tested iteratively.

Here, are key steps in Data warehouse implementation along with its deliverables.

Step	Tasks	Deliverables
1	Need to define project scope	Scope Definition
2	Need to determine business needs	Logical Data Model
3	Define Operational Datastore requirements	Operational Data Store Model
4	Acquire or develop Extraction tools	Extract tools and Software
5	Define Data Warehouse Data requirements	Transition Data Model
6	Document missing data	To Do Project List
7	Maps Operational Data Store to Data Warehouse	D/W Data Integration Map
8	Develop Data Warehouse Database design	D/W Database Design
9	Extract Data from Operational Data Store	Integrated D/W Data Extracts
10	Load Data Warehouse	Initial Data Load

Advantages of Data Warehouse (DWH):

- Data warehouse allows business users to quickly access critical data from some sources all in one place.
- Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
- Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
- Data warehouse helps to reduce total turnaround time for analysis and reporting.
- Restructuring and Integration make it easier for the user to use for reporting and analysis.
- Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
- Data warehouse stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.

Disadvantages of Data Warehouse:

- Not an ideal option for unstructured data.
- Creation and Implementation of Data Warehouse is surely time confusing affair.
- Data Warehouse can be outdated relatively quickly
- Difficult to make changes in data types and ranges, data source schema, indexes, and queries.
- The data warehouse may seem easy, but actually, it is too complex for the average users.
- Despite best efforts at project management, data warehousing project scope will always increase.
- Sometime warehouse users will develop different business rules.
- Organisations need to spend lots of their resources for training and Implementation purpose.

Q14. What is Star-Schema??

Ans. 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 diagrammed by surrounding each fact with its associated dimensions. The resulting diagram resembles a star. 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. 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.

Within the data warehouse or data mart, a dimension table is associated with a fact table by using a foreign key relationship. The dimension table has a single primary key that uniquely identifies each member record (row). The fact table contains the primary key of each associated dimension table as a foreign key. Combined, these foreign keys form a multi-part composite primary key that uniquely identifies each member record in the fact table. The fact table also contains one or more numeric measures. For example, a simple Sales fact with millions of individual clothing sale records might contain a Product Key, Promotion Key, Customer Key, and Date Key, along with Units Sold and Revenue measures. The Product dimension would hold reference information such as product name, description, size, and colour. The Promotion dimension would hold information such as promotion name and price. The Customer dimension would hold information such as first and last name, birth date, gender, address, etc. The Date dimension would include calendar date, week of year, month, quarter, year, etc. This simple Sales fact will easily support queries such as "total revenue for all clothing products sold during the first quarter of the 2010" or "count of female customers who purchased 5 or more dresses in December 2009".

The star schema supports rapid aggregations (such as count, sum, and average) of many fact records, and these aggregations can be easily filtered and grouped ("sliced & diced") by the dimensions. A star schema may be partially normalized (snow flaked), with related information stored in multiple related dimension tables, to support specific data warehousing needs.

Online analytical processing (OLAP) databases (data warehouses and data marts) use a denormalized star schema, with different but related information stored in one dimension table, to optimize queries against large data sets. A star schema may be partially normalized, with related information stored in multiple related dimension tables, to support specific data warehousing needs. In contrast, an online transaction processing (OLTP) database uses a normalized schema, with different but related information stored in separate, related tables to ensure transaction integrity and optimize processing of individual transactions.

Q15. What do you mean by SETL?

Ans. Short for *Set Theory as a Language* (or Set 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.

STATISTICS WORKSHEET-1

Q1. a) True
Q2. a) Central Limit Theorem
Q3. b) Modelling bounded count data
Q4. d) All of the mentioned
Q5. c) Poisson
Q6. b) False
Q7. b) Hypothesis
Q8. a) 0
Q9. c) Outliers cannot conform to the regression relationship
Q10. What do you understand by the term Normal Distribution?

Ans. 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.

KEY NOTES:-

- 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.

Understanding Normal Distribution

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans. An analysis is only as good as its data, and every researcher has struggled with dubious results because of missing data. In this article, I will cover three ways to deal with missing data.

Types of Missing Data

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

- Missing Completely At Random (MCAR): When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.
- Missing At Random (MAR): The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.
- Not Missing At Random (NMAR): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

Common Methods

1. Mean or Median Imputation:- When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

2. Multivariate Imputation by Chained Equations (MICE):- MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, Bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.
Secondly, as mentioned above, the algorithm treats different variables differently. So, all categorical variables should be treated as factor variables before implementing MICE.
Then you can implement the algorithm using the MICE library in R
You can also ignore some variables as predictors or skip a variable from being imputed using the MICE library in R. Additionally, the library also allows you to set a method of imputation discussed above depending upon the nature of the variable.

3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.
One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data. For example, if the distribution of race/ethnicity for non-missing data is similar to the distribution of race/ethnicity for missing data, overfitting is not likely to throw off results. However, if the two distributions differ, the accuracy of imputations will suffer.
The MICE library in R also allows imputations by random forest by setting the method to “rf”. The authors of the MICE library have provided an example on how to implement the random forest method here.
To sum up data imputations is tricky and should be done with care. It is important to understand the nature of the data that is missing when deciding which algorithm to use for imputations. While using the above algorithms, predictor variables should be set up carefully to avoid confusion in the methods implemented during imputation. Finally, you can test the quality of your imputations by normalized root mean square error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables.

Q12. What is A/B testing?

Ans. 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. For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the

customers buying your product, while the sample refers to the number of customers that participated in the test. Our objective here is to check which newsletter brings higher traffic on the website i.e. the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

Q13. Is mean imputation of missing data acceptable practice?

Ans. Mean imputation: So simple. And yet, so dangerous. Perhaps that's a bit dramatic, but mean imputation (also called mean substitution) really ought to be a last resort. It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power. But that doesn't make it a good solution, and it may not help you find relationships with strong parameter estimates. Even if they exist in the population.

On the other hand, there are many alternatives to mean imputation that provide much more accurate estimates and standard errors, so there really is no excuse to use it. This post is the first explaining the many reasons not to use mean imputation (and to be fair, its advantages). First, a definition: mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable.

1: Mean imputation does not preserve the relationships among variables:-True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That's a good thing. Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too.

This is the original logic involved in mean imputation. If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

2: Mean Imputation Leads to An Underestimate of Standard Errors:- A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low. In other words, yes, you get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data.

Ultimately, because your standard errors are too low, so are your p-values. Now you're making Type I errors without realizing it.

Q14. What is linear regression in statistics?

Ans. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable,

and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form ***Y = a + bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is, ***b*** and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

Linear regression is the next step up after correlation. It is used when we want to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable (or sometimes, the outcome variable). The variable we are using to predict the other variable's value is called the independent variable (or sometimes, the predictor variable). For example, you could use linear regression to understand whether exam performance can be predicted based on revision time; whether cigarette consumption can be predicted based on smoking duration; and so forth. If you have two or more independent variables, rather than just one, you need to use multiple regression.

Q15. What are the various branches of statistics?

Ans. Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. The mathematical theories behind statistics rely heavily on differential and integral calculus, linear algebra, and probability theory.

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific **analysis** of data and both are equally important for the student of statistics.

The two major areas of statistics are known as **descriptive** statistics, which describes the properties of sample and population data, and **inferential** statistics, which uses those properties to test hypotheses and draw conclusions.

KEY NOTES:-

- Statistics is the study and manipulation of data, including ways to gather, review, analyze, and draw conclusions from data.
- The two major areas of statistics are descriptive and inferential statistics.
- Statistics can be used to make better-informed business and investing decisions.

Understanding Statistics:-

- Statistics are used in virtually all scientific disciplines such as the physical and social sciences, as well as in business, the humanities, government, and manufacturing. Statistics is fundamentally a branch of applied mathematics that developed from the application of mathematical tools including calculus and linear algebra to probability theory.
- In practice, statistics is the idea we can learn about the properties of large sets of objects or events (a population) by studying the characteristics of a smaller number of similar objects or events (a sample). Because in many cases gathering comprehensive data about an entire population is too costly, difficult, or flat out impossible, statistics start with a sample that can conveniently or affordably be observed.
- Two types of statistical methods are used in analysing data: descriptive statistics and inferential statistics. Statisticians measure and gather data about the individuals or elements of a sample, then analyze this data to generate descriptive statistics. They can then use these observed characteristics of the sample data, which are properly called "statistics," to make inferences or educated guesses about the unmeasured (or unmeasured) characteristics of the broader population, known as the parameters.

Descriptive Statistics:- Descriptive statistics mostly focus on the central tendency, variability, and distribution of sample data. Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as mean, median, and mode. Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as range, variance, and standard deviation.

- The distribution refers to the overall "shape" of the data, which can be depicted on a chart such as a histogram or dot plot, and includes properties such as the probability distribution function, skewness, and kurtosis. Descriptive statistics can also describe differences between observed characteristics of the elements of a data set. Descriptive statistics help us understand the collective properties of the elements of a data sample and form the basis for testing hypotheses and making predictions using inferential statistics.

Inferential Statistics:- Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population from the characteristics of a sample and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution of the sample data statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.

- Inferential statistics are used to make generalizations about large groups, such as estimating average demand for a product by surveying a sample of consumers' buying habits, or to attempt to predict future events, such as projecting the future return of a security or asset class based on returns in a sample period.

- Regression analysis is a common method of statistical inference that attempts to determine the strength and character of the relationship (or correlation) between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). The output of a regression model can be analyzed for statistical significance, which refers to the claim that a result from findings generated by testing or experimentation is not likely to have occurred randomly or by chance but are instead likely to be attributable to a specific cause elucidated by the data. Having statistical significance is important for academic disciplines or practitioners that rely heavily on analyzing data and research.

The difference between descriptive and inferential statistics are

Descriptive statistics are used to describe or summarize the characteristics of a sample or data set, such as a variable's mean, standard deviation, or frequency. Inferential statistics, in contrast, employs any number of techniques to relate variables in a data set to one another, for example using correlation or regression analysis. These can then be used to estimate forecasts or infer causality.