## ASSIGNMENT-01

Q1. What are different types of Analytics?

Ans: There are four type of analytics

* Descriptive analysis.
* Diagnostic analysis.
* Predictive analysis.
* Perspective analysis.

Q2. Explain any four.

Ans:

1. Descriptive Analysis:

* Descriptive analysis involves summarizing and describing the main characteristics of dataset.
* In gerneralized word this says, “What’s happening in your business or past situation?”
* Descriptive statistic consist of three basic categories of measures : measure of central tendency(mean,mode,median) ,measure of variability(Variance,Standard Deviation,Range) and frequency distribution(Skewness ,Kurtosis).
* Uses Charts,graphs and dashboard in a visually gigestible format
* Example:

Retail Store Example: A retail chain generates monthly reports showing total sales, top-selling products, and busiest store locations.

* Use Case:

Retail Sales Monitoring: Retail companies use descriptive analytics to monitor daily, weekly, and monthly sales. For instance, a store may notice increased sales of winter clothing during cold months. This allows for better inventory management and marketing campaigns that align with seasonal trends.

1. Diagnostic Analysis:

* In short we say that it means “What is it happening?”
* This helps in understanding the root of the problem
* Breaks down data into finer categories, such as by region or customer segment, for more detailed insights.
* Correlation and Pattern Detection: Finds relationships between variables to explain why certain outcomes occurred.
* Identifies irregularities or deviations from expected patterns.
* Example:

Telecom Example: A telecom company investigates a rise in customer churn and finds that customers in certain regions experience frequent service outages, which leads to dissatisfaction and higher churn rates.

* Use Case:

Customer Churn Analysis: A subscription service notices an increase in cancellations. Diagnostic analytics reveals that the churn is highest among users who experienced poor customer support. By identifying this, the company can improve its support services to reduce churn.

1. Predictive Analytics

* Answering question like “What may happens next?”
* Uses historical data ,machine learning and other techniques to predicts future events or outcome.
* Risk Assessment: Estimates the probability of specific events, such as customer churn or equipment failure.
* Example:

Finance Example: A bank uses predictive analytics to predict the likelihood of loan defaults by analyzing past credit scores, transaction histories, and financial behavior.

* + Use Case:

Supply Chain Forecasting: A manufacturing company uses predictive analytics to forecast demand for products during peak seasons. This allows them to adjust their production schedules and inventory levels to meet future demand without overproducing.

1. Perspective Analysis :
   * Uses data,mathematical algorithm and business rules to suggest the best course of action for specific scenario.
   * Answring question like “What should I do ?”
   * Perscriptive analysis is used to recommend a couse of action based on the predictions.It helps organization to make data-driven decisions related to production,operations and services.It uses historical data and optimization algorithms.
   * Simulates different decisions to predict their outcomes and determine the best course of action.
   * Example:Logistics Example: A logistics company uses prescriptive analytics to recommend the most efficient delivery routes based on traffic patterns, fuel costs, and weather conditions.
   * Use Case:Inventory Management Optimization: A retail company uses prescriptive analytics to decide how much stock to hold in each store. By analyzing demand patterns, seasonal trends, and warehouse capacity, it recommends an optimal inventory level for each location, reducing overstock and stockouts.

**Industry-Specific Use Cases:**

Healthcare:

* Descriptive: A hospital generates reports on patient admissions and treatments to understand patient demographics and common illnesses.
* Diagnostic: Doctors investigate why certain patients experience longer recovery times after surgery and find it is linked to underlying health conditions.
* Predictive: The hospital predicts which patients are at higher risk of developing complications after surgery using historical patient data.
* Prescriptive: Recommends personalized treatment plans for high-risk patients based on their medical history and predictive models.

Manufacturing:

* Descriptive: A manufacturing company monitors production efficiency and output levels across different shifts.
* Diagnostic: The company investigates why production dropped in a particular month and discovers it was due to a machine malfunction.
* Predictive: Predicts when equipment is likely to fail based on usage data and past performance, enabling predictive maintenance.
* Prescriptive: Recommends the best maintenance schedule to minimize downtime while maximizing production output.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Key Features** | **Example** | **Use Case** |
| **Descriptive Analytics** | Aggregates historical data, visualizes trends, reports on performance. | A retailer summarizes last month’s sales to identify best-selling products. | Retailers use it for sales monitoring and trend analysis. |
| **Diagnostic Analytics** | Investigates causes of past events, identifies correlations, root cause analysis. | A telecom company analyzes customer churn and finds a service issue as the root cause. | Customer churn analysis to find out why customers are leaving. |
| **Predictive Analytics** | Forecasts future outcomes, uses models and machine learning for risk assessment. | A bank predicts which customers are most likely to default on loans. | Supply chain forecasting to predict future product demand. |
| **Prescriptive Analytics** | Provides recommendations and actions based on optimization and scenario analysis. | A logistics company uses it to recommend optimal shipping routes. | Inventory optimization to suggest the best stock levels for each location. |

Q3. What are different stages in Analytics?

Ans:

1.Buisness Issue - Define the question

2.Data Collection – Gathering raw data from various sources (databases,APIs,sensors,etc).

3.Data Cleaning – Removing errors,duplicates and handling missing or inconsistent data to ensure quality.

4.Analyzing Data - Using statistical methods, data visualization, or machine learning models to uncover trends, patterns, and insights in the cleaned data.

5.Visualization - After you’ve interpreted the results and drawn meaningful insights from them, the next step is to create data visualizations. Data visualization involves using several tools. Let's explore two popular tools that most data analysts use. e.g. Tableau ,PowerBi.

Q4. What are different data types and explain any 5?

Ans .Data types refer to the classification of data based on the kind of values they represents. Mainly types of data are broadly categorized into quantitative and qualitative ,along with some subcategories.

1.Quantitative Data

* Also known as Numerical Data
* This type of data represents quantities and can be measured in numbers.It is often use for statistical analysis.It can be futher classifies into two subtypes:
  + - * 1. Discrete Data :
* Represents countable values and is often whole numbers.
* Examples:

The number of products sold in a store.

The number of students in a classroom.

The number of defects in a production batch.

* + - * 1. Continues Data :
* Represents measurable values that can take any value within a range and often includes decimals.
* Examples:

Height and weight of individuals (e.g., 5.7 feet, 68.5 kg).

Temperature readings (e.g., 36.5°C, 75.3°F).

Time taken to complete a task (e.g., 4.25 hours).

* + 1. Qualitative Data
* Also known as categorical data
* This type of data represents characteristics,descriptions,or labels and it cannot be measured in numbers

a. Nominal Data

* Description: Represents categories with no inherent order or ranking.
* Examples**:**
  + Gender (Male, Female, Non-binary).
  + Types of fruits (Apple, Banana, Orange).
  + Eye color (Brown, Blue, Green).

b. Ordinal Data

* Represents categories that have a meaningful order or ranking, but the intervals between them are not equal.
* Examples:
  + Customer satisfaction ratings (Very Satisfied, Satisfied, Neutral, Dissatisfied).
  + Educational levels (High school, Bachelor's, Master's, PhD).
  + Severity of pain (Mild, Moderate, Severe).

c. Binary Data

* Description: A subtype of categorical data where there are only two possible outcomes or categories.
* Examples:
  + Yes/No responses (e.g., Did you purchase the product? Yes/No).
  + True/False statements (e.g., Is the user subscribed? True/False).
  + Pass/Fail results (e.g., Test results: Pass/Fail).

d. Time Series Data

* Data points collected or recorded at specific time intervals.
* Examples:
  + Daily stock prices of a company over a year.
  + Monthly sales data for a retail chain.
  + Temperature readings taken every hour.

e. Text Data (Unstructured Data)

* Data that doesn’t fit into structured formats and is often unorganized, making it harder to analyze directly.
* Examples:
  + Customer reviews on a website.
  + Social media posts or tweets.
  + Emails and text messages.

Q5. Difference between Data Munging / Data Cleaning / Data Mining.

|  |  |  |  |
| --- | --- | --- | --- |
| **Aspect** | **Data Munging (Data Wrangling)** | **Data Cleaning** | **Data Mining** |
| **Definition** | Process of transforming and mapping raw data into a usable format. | Process of fixing errors, removing inconsistencies, and ensuring data quality. | Process of analyzing large datasets to extract patterns, trends, and insights. |
| **Primary Goal** | Prepare raw data for analysis by reformatting and restructuring it. | Improve data quality by correcting, filling, or removing incorrect data. | Extract useful insights, relationships, and patterns from data. |
| **Stage in Data Workflow** | Performed early in the data analysis pipeline, after data collection. | Immediately follows data collection and before analysis. | Performed after data has been cleaned and prepared. |
| **Key Tasks** | - Merging datasets | - Removing duplicates | - Pattern recognition |
|  | - Converting data types | - Handling missing data | - Clustering |
|  | - Formatting | - Correcting errors | - Classification |
|  | - Mapping and transforming variables. | - Normalizing data formats. | - Predictive modeling |
| **Focus** | Transform raw data into a structured and usable format for analysis. | Correct data errors and ensure consistency for accurate analysis. | Discover hidden patterns, trends, and actionable insights. |
| **Input Data Type** | Mostly raw and unstructured or semi-structured data. | Raw or semi-cleaned data with potential errors or inconsistencies. | Cleaned and structured data that’s ready for analysis. |
| **Techniques Involved** | - Data merging | - De-duplication | - Machine learning algorithms (e.g., decision trees, clustering) |
|  | - Pivoting tables | - Handling missing values (imputation, deletion) | - Statistical models (e.g., regression, classification). |
|  | - Data reshaping | - Outlier removal. |  |
|  | - Aggregation and filtering. |  |  |
| **Tools Used** | - Python (Pandas, NumPy) | - Python (Pandas) | - Machine learning frameworks (e.g., Scikit-learn, TensorFlow) |
|  | - SQL | - Data quality tools | - Statistical tools (R, SAS, Python). |
|  | - Excel | - Excel |  |
|  | - R. | - SQL. |  |
| **Output** | Cleaned and well-structured data, ready for further cleaning or analysis. | High-quality, consistent, and error-free data, ready for analysis. | Actionable insights, predictions, and identified patterns. |
| **Challenges** | - Handling multiple data formats | - Dealing with large amounts of missing or corrupt data | - Correct model selection |
|  | - Integration of disparate sources | - Identifying hidden errors. | - Identifying meaningful patterns from complex datasets. |
|  | - Restructuring complex data. |  |  |
| **Use Case Example** | Combining multiple data sources (e.g., sales data and marketing data) into one structured dataset. | Correcting patient medical records by removing duplicates and fixing inconsistent entries. | Analyzing customer behavior data to predict which customers are likely to churn. |
| **When Used** | Before the cleaning process or when preparing data for analysis. | Directly after data collection or before performing analysis. | After the data is cleaned and prepared for analysis. |
| **Skills Required** | - Data manipulation | - Attention to detail | - Statistical and analytical skills |
|  | - Basic programming | - Knowledge of statistical imputation techniques | - Machine learning algorithms |
|  | - SQL | - Experience with data cleaning tools. | - Strong understanding of patterns and trends. |
|  | - Domain knowledge. |  |  |
| **Examples** | - Converting sales data from text to numeric format. | - Removing duplicate entries in customer datasets. | - Identifying purchasing patterns through clustering. |
|  | - Merging different customer data sources (CRM, website). | - Filling missing age values with the median age. | - Predicting customer churn with classification models. |

Q6. What are two major data collection methods explain with example.

Ans: There are two major methods for data collection as following

Primary Data Collection : Primary data collection is the process of collecting information directly from users.This type of data collection is usually done through surveys ,experiments and interviews.

e.g . A marketing company wants to understand customer preferences for a new product line .It conducts survey by directing asking customers about their product preference.The company h=gathers primary data by directly interacting with its target audience ,ensuring the data is highly relevant to their product development.

Secondary Data Collection : secondary data collection is the process of collecting data or information from other sources, such as public records or database ,Government database ,Commercial Database ,etc.

e.g : A business analyst wants to analyze trends in the retail sector over the last decade. Instead of collecting new data, they use secondary data from government databases and industry reports to track sales growth, consumer spending, and market share trends.

Q7. What do you mean by Inferential statistics and when would you be using it?

Ans:

* Inferential statistics allow researchers to make generalizations and draw conclusions about a larger population based on a sample.
* Drawing conclusion from data
* **Population vs. Sample:** Population refers to the entire group of individuals or items that you are studying, while a sample is a smaller subset of the population.
* **Example:**

**Business Decision-Making:**  
A retail company wants to know if a new marketing campaign has led to increased sales. They use a sample of sales data from a few stores to infer whether the campaign is effective across all stores. Using hypothesis testing, they check whether the observed sales increase in the sample is statistically significant and not just due to random chance.

Q8. What is SRS technique or method in inferential statics and why is it essential.

Ans :SRS is a method where each member of a population has an equal chance of being selected for the sample. This ensures that the sample is unbiased and representative of the population, allowing accurate inferences.

Why is SRS Essential?

1. Unbiased Sample: Ensures equal chances for all, reducing sampling bias.
2. Generalization: Allows researchers to generalize findings from the sample to the larger population.
3. Accurate Estimates: Provides reliable estimates of population parameters (mean, proportion, etc.).
4. Supports Hypothesis Testing: Many inferential techniques assume data comes from a random sample.

Example:

A researcher randomly selects 100 employees from a company of 1,000 to estimate average employee satisfaction. Since the sample is random, it can reliably reflect the views of the entire workforce.

Q9. What do you understand by 1st Business Moment. What methods are used to identify the tendency.

Ans: The 1st Business Moment refers to the average of a dataset, which is measure of central tendency .It represents the central point or typical value around which the data tends to cluster.The mean used to summarize the general level of data pointd in a dataset.

Methods:

1.Mean :The arithmetic average of all data points. It is calculated by summing all values and dividing by the total number of data points.  
Mean=∑x/n

2. Median:The middle value in a sorted dataset. If the dataset has an odd number of observations, the median is the middle value. If it's even, the median is the average of the two middle values.

3. Mode:The value that occurs most frequently in the dataset.

Q10. Which summarization method we reliable for Numerical and Categorical data?

|  |  |  |
| --- | --- | --- |
| **Data Type** | **Summarization Methods** | **When to Use** |
| **Numerical Data** | - Mean | Symmetric data without outliers |
|  | - Median | Skewed data or data with outliers |
|  | - Mode | Most common value in the dataset |
|  | - Range, Variance, Standard Deviation | Measure of spread or dispersion |
| **Categorical Data** | - Mode | To find the most frequent category |
|  | - Frequency Distribution | To see how often each category appears |

Q11. What is your instinct to use median over mean explain with example.

Ans :

It is best to use the median when the distribution is either skewed or there are outliers present. When a distribution is skewed, the median does a better job of describing the center of the distribution than the mean.

e.g :

* There is training institute for all long jump athletes , so in this institute a coach has been assigned so he will be training all the athletes into the institute for a duration of one month they will be deciding up weather we should keep this coach or not .
* using mean and median committee decides. Consider that there are five athletes into the coaching institute so there one month training has been completely, improvement produced by the five athletes with the help of coach.
* Player 1 improved performance by 0.15 , Player 2 improved performance by 0.11, Player 3 improved performance by 0.06, Player 4 improved performance by 0.12, Player 5 improved performance by -0.56
* Lets committee will be decide or not wheather the coach should kept or not.
* Lets us calculate mean of improvement length of athletes
  + Mean=0.44-0.56/5 = -0.12/5 =-0.024

Result=Athlete performance down by one month of training give by the coach

* Median = we set data in ascending order

-0.56 0.06 0.11 0.12 0.15

Median is odd no =0.11

Result = Athete performance is good within one month training hence committee kept coach

* Player 5 is outlier that damages data.

Q12. What do you mean by outlier/extreme values.

Ans :

* Values in our data which are unusual for a given context are called outliers.
* E.g :: 4 employees in one organization 3 employee’s salary less than one employee then the that one employee is outlier ,but that one employee head of the department and other employees are entry level then the we can not say “the outlier is not always bad data point or data”
* Using Box plot and IQR we handle outlier

Q13 Write down the formulas for Mean, Median, Mode. Also use an example to showcase.

Ans:

* Mean (Average):

Formula: Mean=∑XiN\text{Mean} = \frac{\sum X\_i}{N}Mean=N∑Xi​​ Where:

* + XiX\_iXi​ = each individual data point
  + NNN = total number of data points

Example: Let's calculate the mean for the dataset: 5,7,9,10,125, 7, 9, 10, 125,7,9,10,12 Mean=5+7+9+10+125=435=8.6\text{Mean} = \frac{5 + 7 + 9 + 10 + 12}{5} = \frac{43}{5} = 8.6Mean=55+7+9+10+12​=543​=8.6 So, the mean is 8.6.

* Median:

Formula: To calculate the median, the data points are first arranged in ascending order. The median is:

* + If NNN (the number of data points) is odd, the median is the middle value.
  + If NNN is even, the median is the average of the two middle values.

1. Example (Odd Number of Data Points): For the dataset: 5,7,9,10,125, 7, 9, 10, 125,7,9,10,12 The number of data points, N=5N = 5N=5 (odd), so the median is the middle value:  
   Median = 9

Example (Even Number of Data Points): For the dataset: 5,7,9,105, 7, 9, 105,7,9,10 The number of data points, N=4N = 4N=4 (even), so the median is the average of the two middle values:

Median=7+92=162=8\text{Median} = \frac{7 + 9}{2} = \frac{16}{2} = 8Median=27+9​=216​=8

So, the median is 8.

2) Example Median for Even Data Points:

In the same dataset:

4,8,15,16,23,424, 8, 15, 16, 23, 424,8,15,16,23,42

1. Sort the Data:
   * The data is already in ascending order: 4,8,15,16,23,424, 8, 15, 16, 23, 424,8,15,16,23,42
2. Find the Two Middle Values:
   * In this case, N=6N = 6N=6, so the middle two values are the 3rd and 4th values, which are 15 and 16.
3. Average the Middle Two Values:

Median=15+162=15.5\text{Median} = \frac{15 + 16}{2} = 15.5Median=215+16​=15.5

Thus, the median is 15.5.

* Mode:

Formula: The mode is the value that occurs most frequently in the dataset.

Example: For the dataset: 5,7,7,9,10,125, 7, 7, 9, 10, 125,7,7,9,10,12 The value 7 appears twice, while all other values appear once.  
So, the mode is 7.