## SMART INTERNZ-APSCHE

## AI/ML Training

ASSIGNMENT : Data Wrangling and Regression Analysis

## SECTION-A: DATA WRANGLING.

1. What is the primary objective of data wrangling?

 a) Data visualization

 b) Data cleaning and transformation

 c) Statistical analysis

 d) Machine learning modeling

b) Data cleaning and transformation

2. Explain the technique used to convert categorical data into numerical data. How

does it help in data analysis.

The technique used to convert categorical data into numerical data is called "one-hot encoding." One-hot encoding transforms categorical variables into a format that can be provided to machine learning algorithms to improve predictive performance.

Here's how it works:

1. \*Identification of Categorical Variables\*: First, you identify which columns in your dataset contain categorical variables. These variables represent categories, such as color, gender, or country.

2. \*Creation of Dummy Variables\*: For each categorical variable, you create a set of binary dummy variables. Each dummy variable represents one category and takes the value 1 if the observation belongs to that category, and 0 otherwise.

3. \*Transformation\*: You replace each categorical variable with its corresponding set of dummy variables. For example, if you have a categorical variable "color" with categories "red," "blue," and "green," you create three dummy variables: "color\_red," "color\_blue," and "color\_green." Each dummy variable takes the value 1 if the observation's color matches the category and 0 otherwise.

One-hot encoding helps in data analysis in several ways:

- \*Compatibility with Machine Learning Algorithms\*: Many machine learning algorithms require numerical input. By converting categorical variables into numerical form, you can use a broader range of algorithms for your analysis.

- \*Preservation of Information\*: One-hot encoding preserves the information contained in categorical variables without imposing an arbitrary ordinal relationship. This is important because categorical variables may not have a natural ordering.

- \*Avoidance of Bias\*: One-hot encoding prevents bias that might occur if categorical variables were encoded using arbitrary numerical values. For example, assigning numerical values like 1, 2, and 3 to categories may imply an order or magnitude that doesn't exist in the data.

- \*Increased Flexibility\*: One-hot encoding allows for more flexibility in the analysis, as it enables algorithms to learn different weights for each category independently.

Overall, one-hot encoding is a valuable technique for converting categorical data into a format suitable for analysis by machine learning algorithms, enabling better model performance and more accurate insights from the data.

3. How does LabelEncoding differ from OneHotEncoding.

Label Encoding and One-Hot Encoding are both techniques used to convert categorical data into numerical form, but they differ in how they represent categorical variables.

\*Label Encoding:\*

- Label Encoding assigns a unique integer to each category in the categorical variable.

- It is often used when the categorical variable has an inherent ordinal relationship, meaning the categories can be ordered or ranked.

- For example, if you have a categorical variable "Size" with categories "Small," "Medium," and "Large," Label Encoding might assign the values 0, 1, and 2 respectively.

- The problem with Label Encoding is that it may introduce unintended ordinal relationships between categories, which can be misleading to machine learning algorithms.

\*One-Hot Encoding:\*

- One-Hot Encoding creates a binary dummy variable for each category in the categorical variable.

- Each dummy variable represents one category and takes the value 1 if the observation belongs to that category, and 0 otherwise.

- It is useful when there is no inherent order or ranking among the categories, or when such an order should not be imposed on the data.

- For example, if you have a categorical variable "Color" with categories "Red," "Blue," and "Green," One-Hot Encoding would create three binary variables: "Color\_Red," "Color\_Blue," and "Color\_Green."

- One-Hot Encoding preserves the information of each category without introducing any ordinal relationship between them.

In summary, Label Encoding assigns a unique integer to each category, potentially introducing an ordinal relationship, while One-Hot Encoding creates binary dummy variables, preserving the categorical nature of the data without imposing any order. Which one to use depends on the nature of the categorical variable and the requirements of the analysis or machine learning algorithm.

4. Describe a commonly used method for detecting outliers in a dataset. Why is it

important to identify outliers?

One commonly used method for detecting outliers in a dataset is the "Interquartile Range (IQR)" method.

Here's how it works:

1. \*Calculate the Quartiles\*: First, calculate the first quartile (Q1) and the third quartile (Q3) of the dataset.

2. \*Calculate the Interquartile Range (IQR)\*: The IQR is the difference between the third quartile (Q3) and the first quartile (Q1), i.e., IQR = Q3 - Q1.

3. \*Identify Outliers\*: Any data point that falls below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR is considered an outlier.

4. \*Optional Step\*: Some implementations use a more aggressive threshold, like Q1 - 3 \* IQR and Q3 + 3 \* IQR, to identify extreme outliers.

It's important to identify outliers because:

1. \*Data Quality Assurance\*: Outliers may indicate errors in data collection, data entry, or processing. Identifying and investigating outliers can help ensure the quality and integrity of the dataset.

2. \*Preservation of Statistical Validity\*: Outliers can skew statistical analyses and machine learning models, leading to misleading results or inaccurate predictions. Removing or handling outliers appropriately helps preserve the statistical validity of the analysis.

3. \*Insight Generation\*: Outliers may represent rare or unusual events, phenomena, or behaviors in the data. Understanding these outliers can provide valuable insights into the underlying processes or patterns in the dataset.

4. \*Model Performance\*: Outliers can adversely affect the performance of machine learning models by influencing parameter estimation and model fitting. Detecting and properly handling outliers can improve the performance and generalization ability of the models.

In summary, identifying outliers using methods like the Interquartile Range (IQR) method is crucial for maintaining data quality, ensuring the validity of statistical analyses, gaining insights from the data, and improving the performance of machine learning models.

5. Explain how outliers are handled using the Quantile Method.

The Quantile Method, also known as the Interquartile Range (IQR) method, is a technique commonly used to handle outliers in a dataset. Here's how outliers are handled using the Quantile Method:

1. \*Calculate Quartiles\*:

- First, calculate the first quartile (Q1) and the third quartile (Q3) of the dataset.

2. \*Compute Interquartile Range (IQR)\*:

- The interquartile range (IQR) is calculated as the difference between the third quartile (Q3) and the first quartile (Q1), i.e., IQR = Q3 - Q1.

3. \*Identify Outliers\*:

- Any data point that falls below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR is considered an outlier.

- Optionally, a more aggressive threshold can be used, such as Q1 - 3 \* IQR and Q3 + 3 \* IQR, to identify extreme outliers.

4. \*Handle Outliers\*:

- Once outliers are identified, there are several ways to handle them:

- \*Removal\*: Outliers can be removed from the dataset. However, this approach should be used cautiously, as removing too many outliers can lead to loss of information and bias in the data.

- \*Transformation\*: Outliers can be transformed using mathematical functions such as log transformation, square root transformation, or Winsorization. Transformation can help reduce the impact of outliers without removing them entirely.

- \*Imputation\*: Outliers can be replaced with more typical values. This could involve replacing outliers with the median, mean, or some other statistically derived value.

- \*Model-based Methods\*: Outliers can be handled using robust statistical models that are less sensitive to outliers, such as robust regression or robust clustering algorithms.

5. \*Evaluate Impact\*:

- After handling outliers, it's essential to evaluate the impact on the dataset and the subsequent analysis or modeling tasks. This includes assessing changes in statistical summaries, model performance, and the interpretation of results.

Handling outliers using the Quantile Method helps ensure the integrity and reliability of the dataset for analysis or modeling tasks by identifying and mitigating the influence of extreme values that may skew results or affect model performance.

6. Discuss the significance of a Box Plot in data analysis. How does it aid in

identifying potential outliers?

A Box Plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It provides a visual summary of key statistical measures, including the median, quartiles, and potential outliers. Here's how a Box Plot aids in data analysis and helps identify potential outliers:

1. \*Visualization of Data Distribution\*:

- A Box Plot displays the distribution of the dataset, including the central tendency and spread of the data. It consists of a box, which represents the interquartile range (IQR) containing the middle 50% of the data, with a line inside representing the median. Whiskers extend from the box to indicate the range of the data, excluding potential outliers.

2. \*Identification of Central Tendency\*:

- The median line in the box represents the central tendency of the data. It indicates the value that divides the dataset into two equal halves.

3. \*Detection of Potential Outliers\*:

- Box Plots visually highlight potential outliers as individual data points that fall outside the "whiskers" of the plot, which extend to a specified distance from the quartiles. Outliers are represented as individual points beyond the "whiskers" and are plotted separately from the main box.

4. \*Comparison between Groups\*:

- Box Plots are particularly useful for comparing the distributions of different groups or categories within a dataset. By displaying multiple Box Plots side by side, analysts can compare the central tendency and spread of each group and identify any differences or similarities.

5. \*Robustness to Skewed Data\*:

- Unlike some other visualizations, such as histograms, Box Plots are robust to skewed distributions and outliers. They provide a clear representation of the data's central tendency and spread, even in the presence of extreme values.

6. \*Assessment of Symmetry and Skewness\*:

- Box Plots can also provide insights into the symmetry and skewness of the dataset. The length of the box and the position of the median line within the box indicate whether the distribution is symmetric or skewed.

Overall, Box Plots are valuable tools in data analysis because they offer a concise summary of the dataset's distribution and aid in the identification of potential outliers. By visually representing key statistical measures and outliers, Box Plots help analysts understand the underlying structure of the data and make informed decisions about subsequent analysis or modeling tasks.

## Section B: Regression Analysis

7. What type of regression is employed when predicting a continuous target

Variable?

When predicting a continuous target variable, linear regression is commonly employed. Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). It assumes that the relationship between the variables is linear, meaning that changes in the independent variables are associated with constant changes in the dependent variable.

In linear regression, the relationship between the independent variables \(X\) and the dependent variable \(y\) is represented by the equation of a straight line:

\[ y = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + ... + \beta\_n X\_n + \epsilon \]

Where:

- \( y \) is the dependent variable (target),

- \( X\_1, X\_2, ..., X\_n \) are the independent variables (features),

- \( \beta\_0, \beta\_1, \beta\_2, ..., \beta\_n \) are the coefficients (parameters) that represent the slope and intercept of the line,

- \( \epsilon \) is the error term representing the difference between the observed and predicted values.

The goal of linear regression is to estimate the coefficients \( \beta\_0, \beta\_1, \beta\_2, ..., \beta\_n \) that minimize the sum of squared differences between the observed and predicted values of the dependent variable. Once the coefficients are estimated, the linear regression model can be used to predict the value of the dependent variable for new observations based on their values of the independent variables.

Linear regression is widely used in various fields, including economics, finance, healthcare, and social sciences, for tasks such as predicting sales, estimating the impact of interventions, and modeling relationships between variables.

8. Identify and explain the two main types of regression?

The two main types of regression are:

1. \*Linear Regression\*:

- Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features).

- It assumes that the relationship between the variables is linear, meaning that changes in the independent variables are associated with constant changes in the dependent variable.

- The relationship between the independent variables \(X\) and the dependent variable \(y\) is represented by the equation of a straight line.

- Linear regression is commonly used for predicting continuous target variables.

- Example applications include predicting house prices based on features such as size, number of bedrooms, and location, or predicting sales based on advertising expenditure.

2. \*Logistic Regression\*:

- Logistic regression is a statistical method used for binary classification tasks, where the target variable has only two possible outcomes (e.g., yes/no, 1/0, true/false).

- Despite its name, logistic regression is a classification algorithm rather than a regression algorithm.

- Logistic regression models the probability that an observation belongs to a particular class using the logistic function (sigmoid function).

- Unlike linear regression, logistic regression does not assume a linear relationship between the independent variables and the log-odds of the target variable.

- Logistic regression is widely used in various fields, including healthcare (e.g., predicting the likelihood of disease occurrence), finance (e.g., credit risk assessment), and marketing (e.g., predicting customer churn).

In summary, linear regression is used for predicting continuous target variables by modeling a linear relationship between the variables, while logistic regression is used for binary classification tasks by modeling the probability of class membership. Both types of regression are powerful and widely used techniques in statistics and machine learning for different types of predictive modeling tasks.

9. When would you use Simple Linear Regression? Provide an example scenario.

Simple Linear Regression is used when you want to model the relationship between two variables: one independent variable (predictor) and one dependent variable (response), assuming a linear relationship between them. Here's an example scenario where you would use Simple Linear Regression:

\*Example Scenario:\*

You are working as a data analyst for a retail company, and you are tasked with analyzing the relationship between advertising expenditure and sales revenue. Your goal is to determine whether there is a linear relationship between the amount spent on advertising and the resulting sales revenue.

\*Variables:\*

- \*Independent Variable (X)\*: Advertising Expenditure - The amount of money spent on advertising campaigns.

- \*Dependent Variable (y)\*: Sales Revenue - The total revenue generated from product sales.

\*Objective:\*

To build a Simple Linear Regression model to predict sales revenue based on advertising expenditure.

\*Steps:\*

1. \*Data Collection\*: Gather data on advertising expenditure and sales revenue over a period of time. Each data point should represent a specific time period (e.g., month or quarter).

2. \*Data Exploration\*: Perform exploratory data analysis to understand the distribution and relationship between advertising expenditure and sales revenue. This may involve visualizations such as scatter plots and correlation analysis.

3. \*Model Building\*:

- Define the independent variable \(X\) (advertising expenditure) and dependent variable \(y\) (sales revenue).

- Split the dataset into training and testing sets.

- Fit a Simple Linear Regression model to the training data, estimating the coefficients (slope and intercept) of the linear relationship between advertising expenditure and sales revenue.

4. \*Model Evaluation\*:

- Evaluate the performance of the model using the testing set. Metrics such as mean squared error (MSE) or \(R^2\) score can be used to assess the goodness of fit.

- Interpret the coefficients of the model to understand the strength and direction of the relationship between advertising expenditure and sales revenue.

5. \*Prediction\*:

- Use the trained Simple Linear Regression model to make predictions on new data. Given a specific advertising expenditure, the model can predict the corresponding sales revenue.

6. \*Insights and Recommendations\*:

- Provide insights to stakeholders based on the results of the analysis. For example, if the model indicates a positive relationship between advertising expenditure and sales revenue, you may recommend increasing advertising budgets to boost sales.

In summary, Simple Linear Regression is used when you want to understand and quantify the relationship between two variables, such as advertising expenditure and sales revenue in this example scenario. It helps in making data-driven decisions and optimizing business strategies based on the insights gained from the analysis.

10. In Multi Linear Regression, how many independent variables are typically

involved?

In Multi Linear Regression, multiple independent variables are involved. The term "multi" in Multi Linear Regression refers to the presence of more than one independent variable.

In Multi Linear Regression, the relationship between the dependent variable and multiple independent variables is modeled using a linear equation:

\[ y = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + ... + \beta\_n X\_n + \epsilon \]

Where:

- \( y \) is the dependent variable,

- \( X\_1, X\_2, ..., X\_n \) are the independent variables,

- \( \beta\_0, \beta\_1, \beta\_2, ..., \beta\_n \) are the coefficients (parameters) that represent the slope and intercept of the linear relationship,

- \( \epsilon \) is the error term representing the difference between the observed and predicted values.

The number of independent variables \( n \) can vary depending on the specific problem and the dataset. It can range from just a few independent variables to several dozen or more, depending on the complexity of the relationships being modeled and the availability of data.

Multi Linear Regression is a powerful tool for analyzing the relationships between multiple independent variables and a dependent variable, and it is commonly used in various fields such as economics, finance, engineering, and social sciences for predictive modeling and inference.

11. When should Polynomial Regression be utilized? Provide a scenario where

Polynomial Regression would be preferable over Simple Linear Regression?

Polynomial Regression should be utilized when the relationship between the independent variable(s) and the dependent variable is not linear but can be better represented by a polynomial function. This is particularly useful when the relationship exhibits curvature or non-linearity.

A scenario where Polynomial Regression would be preferable over Simple Linear Regression is when analyzing phenomena where the relationship between variables is not strictly linear and cannot be adequately captured by a straight line. Here's an example scenario:

\*Scenario:\*

You are analyzing the relationship between a person's age and their annual income. Initially, you consider using Simple Linear Regression to model this relationship, assuming that income increases linearly with age. However, upon visualizing the data, you notice that the relationship does not appear to be strictly linear. Instead, it seems to follow a curved pattern, where income initially increases with age but then starts to level off or decrease at older ages.

\*Reasoning:\*

In this scenario, Polynomial Regression would be preferable over Simple Linear Regression because it can capture the curvature or non-linearity in the relationship between age and income. By fitting a polynomial function to the data, Polynomial Regression can better represent the underlying patterns and variations in the data, allowing for more accurate predictions and insights.

\*Steps:\*

1. \*Data Collection\*: Gather data on individuals' ages and their corresponding annual incomes.

2. \*Data Visualization\*: Visualize the relationship between age and income using scatter plots or other visualization techniques.

3. \*Modeling\*: Fit Polynomial Regression models of different degrees (e.g., quadratic, cubic) to the data and compare their performance.

4. \*Model Evaluation\*: Evaluate the performance of the Polynomial Regression models using metrics such as mean squared error (MSE) or \(R^2\) score.

5. \*Interpretation\*: Interpret the coefficients of the best-fitting Polynomial Regression model to understand the relationship between age and income and make predictions based on the model.

In summary, Polynomial Regression should be utilized when the relationship between variables is non-linear or exhibits curvature, as it can provide a more flexible and accurate model compared to Simple Linear Regression.

12.What does a higher degree polynomial represent in Polynomial Regression? How

does it affect the model's complexity?

In Polynomial Regression, the degree of the polynomial represents the highest power of the independent variable(s) in the polynomial equation. A higher degree polynomial includes additional terms with higher powers of the independent variable(s), resulting in a more flexible and complex model.

Here's what a higher degree polynomial represents in Polynomial Regression:

1. \*Flexibility in Capturing Patterns\*:

- Higher degree polynomials can capture more complex patterns and variations in the data compared to lower degree polynomials or simple linear models.

- For example, while a linear model (degree 1 polynomial) can only capture linear relationships, a quadratic model (degree 2 polynomial) can capture both linear and curved relationships, and higher degree polynomials can capture even more intricate patterns.

2. \*Increased Complexity\*:

- As the degree of the polynomial increases, the complexity of the model also increases.

- Higher degree polynomials introduce more parameters (coefficients) into the model, leading to increased model complexity.

- The model becomes more flexible and can fit the training data more closely, but it also becomes more susceptible to overfitting, where it captures noise or random fluctuations in the training data instead of the underlying patterns.

3. \*Risk of Overfitting\*:

- While higher degree polynomials can better fit the training data, they may not generalize well to new, unseen data.

- Overfitting occurs when the model captures noise or idiosyncrasies in the training data, leading to poor performance on test data.

- In Polynomial Regression, higher degree polynomials are more prone to overfitting, especially when the dataset is small or noisy.

4. \*Bias-Variance Tradeoff\*:

- Increasing the degree of the polynomial reduces bias (the error due to underfitting) but increases variance (the error due to overfitting).

- Finding the right balance between bias and variance is crucial for building a robust Polynomial Regression model.

- Techniques such as cross-validation and regularization can be used to manage model complexity and prevent overfitting in Polynomial Regression.

In summary, a higher degree polynomial in Polynomial Regression represents increased flexibility in capturing complex patterns in the data, but it also leads to higher model complexity and a greater risk of overfitting. Careful selection of the polynomial degree is essential to strike the right balance between model complexity and generalization performance.

13. Highlight the key difference between Multi Linear Regression and Polynomial

Regression.

The key difference between Multi Linear Regression and Polynomial Regression lies in the nature of the relationships they model:

1. \*Multi Linear Regression\*:

- Multi Linear Regression models the relationship between a dependent variable (target) and multiple independent variables (features).

- It assumes a linear relationship between the independent variables and the dependent variable.

- The equation for Multi Linear Regression is:

\[ y = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + ... + \beta\_n X\_n + \epsilon \]

- \( y \) is the dependent variable,

- \( X\_1, X\_2, ..., X\_n \) are the independent variables,

- \( \beta\_0, \beta\_1, \beta\_2, ..., \beta\_n \) are the coefficients (parameters) that represent the slope and intercept of the linear relationship,

- \( \epsilon \) is the error term representing the difference between the observed and predicted values.

- Multi Linear Regression is suitable for modeling linear relationships between multiple independent variables and a dependent variable.

2. \*Polynomial Regression\*:

- Polynomial Regression models the relationship between a dependent variable and one or more independent variables using polynomial functions.

- It allows for non-linear relationships between the independent variables and the dependent variable by including higher order terms (e.g., squared terms, cubic terms) in the regression equation.

- The equation for Polynomial Regression is:

\[ y = \beta\_0 + \beta\_1 X + \beta\_2 X^2 + ... + \beta\_d X^d + \epsilon \]

- In addition to the linear term \( X \), Polynomial Regression includes higher order terms \( X^2, X^3, ..., X^d \) up to a specified degree \( d \).

- Polynomial Regression can capture non-linear patterns and relationships in the data that cannot be represented by Multi Linear Regression.

- The choice of the polynomial degree determines the flexibility and complexity of the model.

In summary, while Multi Linear Regression models linear relationships between multiple independent variables and a dependent variable, Polynomial Regression allows for non-linear relationships by including higher order terms of the independent variable(s) in the regression equation. Polynomial Regression is a more flexible modeling approach suitable for capturing non-linear patterns in the data.

14. Explain the scenario in which Multi Linear Regression is the most appropriate

regression technique.

Multi Linear Regression is the most appropriate regression technique in scenarios where you have a single dependent variable (target) and multiple independent variables (features) that you believe collectively influence the dependent variable. Here's a scenario where Multi Linear Regression is commonly used:

\*Scenario:\*

You are a real estate analyst tasked with predicting house prices based on various factors such as size, number of bedrooms, location, and age of the house. Your goal is to build a model that can accurately predict house prices using these independent variables.

\*Variables:\*

- \*Dependent Variable (y)\*: House Price - The target variable representing the price of a house.

- \*Independent Variables (X)\*:

- Size: The size of the house in square feet.

- Number of Bedrooms: The number of bedrooms in the house.

- Location: The geographical location of the house (e.g., city, neighborhood).

- Age of the House: The age of the house in years.

\*Objective:\*

To build a Multi Linear Regression model to predict house prices based on the independent variables.

\*Steps:\*

1. \*Data Collection\*: Gather data on house prices and the independent variables (size, number of bedrooms, location, age) for a sample of houses.

2. \*Data Exploration\*: Explore the relationship between the independent variables and house prices using descriptive statistics and visualizations.

3. \*Feature Engineering\*: Preprocess the data, handle missing values, encode categorical variables, and scale/normalize the features if necessary.

4. \*Model Building\*:

- Define the independent variables \(X\) and dependent variable \(y\).

- Split the dataset into training and testing sets.

- Fit a Multi Linear Regression model to the training data, estimating the coefficients (slopes) for each independent variable.

5. \*Model Evaluation\*:

- Evaluate the performance of the model using metrics such as mean squared error (MSE), \(R^2\) score, and residual analysis.

- Validate the model's performance on the testing set to ensure it generalizes well to new data.

6. \*Prediction\*:

- Use the trained Multi Linear Regression model to make predictions on new data. Given the values of the independent variables (size, number of bedrooms, location, age), the model can predict the corresponding house price.

7. \*Interpretation\*:

- Interpret the coefficients of the model to understand the strength and direction of the relationships between the independent variables and house prices.

8. \*Insights and Recommendations\*:

- Provide insights to stakeholders based on the results of the analysis. For example, identify which features have the most significant impact on house prices and provide recommendations for pricing strategies or investment decisions.

In summary, Multi Linear Regression is the most appropriate regression technique in scenarios where you have multiple independent variables and want to model the relationship between these variables and a single dependent variable. It is widely used in various fields, including finance, economics, healthcare, and real estate, for predictive modeling and inference tasks.

15. What is the primary goal of regression analysis?

The primary goal of regression analysis is to understand and model the relationship between one or more independent variables (predictors) and a dependent variable (response).

In other words, regression analysis aims to:

1. \*Predict the Dependent Variable\*: Regression models are used to predict the value of the dependent variable based on the values of the independent variables. This allows for making informed predictions or estimates about the outcome variable given certain input variables.

2. \*Understand the Relationship\*: Regression analysis helps in understanding how changes in the independent variables are associated with changes in the dependent variable. It quantifies the strength and direction of the relationship between variables, allowing for interpretation of the underlying patterns and trends in the data.

3. \*Inferential Purposes\*: Regression analysis is also used for making inferences about the population parameters based on sample data. It helps in testing hypotheses about the relationships between variables and drawing conclusions about the population from which the sample was drawn.

Overall, the primary goal of regression analysis is to provide insights into the relationship between variables, facilitate prediction, and aid in decision-making processes across various fields such as economics, finance, healthcare, social sciences, and engineering.

Submitted By

Nagisetti Yamini

20HU1A4226