Section A: Data Wrangling (Questions 1-6)

1) The primary objective of data wrangling is

b) Data cleaning and transformation.

This process involves transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

2) The technique used to convert categorical data into numerical data is called encoding. This technique is used to transform non-numerical data into numerical data so that it can be used in data analysis. There are two commonly used encoding techniques: LabelEncoding and OneHotEncoding. LabelEncoding assigns a unique numerical value to each category in the data, while OneHotEncoding creates a binary column for each category in the data.

3) LabelEncoding and OneHotEncoding are techniques used to convert categorical data into numerical data. LabelEncoding assigns a unique numerical value to each category, while OneHotEncoding creates a binary column for each category and assigns a value of 1 or 0 to indicate whether the category is present or not. OneHotEncoding is preferred over LabelEncoding when there is no inherent order or hierarchy among the categories.

4) A commonly used method for detecting outliers in a dataset is the Z-score approach. Here are the steps involved:

1. Calculate the Z-score for each data point using the formula: $Z = (X - \mu) / \sigma$, where Z is the Z-score, X is the data point, μ is the mean, and σ is the standard deviation.

2. Define a threshold value, typically 3, to determine if a data point is an outlier.

3. Mark the data points whose absolute value of Z-score is greater than the threshold as outliers.

It is important to identify outliers for the following reasons:

i. Outliers can significantly impact statistical measures such as the mean and median, leading to biased results.

ii. Outliers may indicate errors in data collection or measurement, which can affect the validity and reliability of the analysis.

iii. Outliers can provide valuable insights and information about unusual or unexpected phenomena in the dataset.

iv. Outliers can affect the performance of machine learning algorithms, as they can introduce noise and distort the patterns in the data.

5) The Quantile Method is a robust technique for handling outliers in a dataset. Here's how outliers are typically handled using the Quantile Method:

i. Determine the lower and upper quantiles (Q1 and Q3) of the dataset.

ii. Calculate the interquartile range (IQR) by subtracting Q1 from Q3: IQR = Q3 - Q1.

iii. Define the lower and upper bounds for outliers as follows:

- Lower Bound: Q1 - 1.5 * IQR

- Upper Bound: Q3 + 1.5 * IQR

iv. Identify any data points that fall below the lower bound or above the upper bound as outliers.

v. Replace or remove the identified outliers based on the specific requirements of the analysis.

6) Handling outliers using the Quantile Method is beneficial because it is less sensitive to extreme values compared to other methods like Z-score. By focusing on the interquartile range, the Quantile Method provides a more robust measure of variability in the dataset and helps in identifying and addressing outliers effectively without being influenced by the mean and standard deviation.

Overall, the Quantile Method is a reliable approach for detecting and managing outliers in a dataset, ensuring that the data analysis is more accurate and reliable.A Box Plot, also known as a Box-and-Whisker Plot, is a powerful visualization tool used in data analysis to provide a graphical summary of the distribution of a dataset. Here are some key points on the significance of a Box Plot in data analysis and how it aids in identifying potential outliers:

i. Visualizing Data Distribution: A Box Plot displays the five-number summary of a dataset, including the minimum, first quartile (Q1), median (second quartile or Q2), third quartile (Q3), and maximum. This visual representation helps in understanding the central tendency, spread, and skewness of the data distribution.

ii. Identification of Outliers: Box Plots are effective in identifying potential outliers in a dataset. Outliers are data points that fall significantly below the lower whisker (Q1 - 1.5 * IQR) or above the upper whisker (Q3 + 1.5 * IQR) of the Box Plot. By visually inspecting the Box Plot, analysts can easily spot data points that lie beyond these whiskers, indicating potential outliers.

iii. Comparison of Data: Box Plots are useful for comparing the distribution of multiple datasets or groups. By plotting multiple Box Plots side by side, analysts can visually compare the central tendency, spread, and variability of different datasets, making it easier to identify patterns, differences, and outliers across groups.

iv. Robustness to Skewed Data: Box Plots are robust to outliers and skewed data distributions. Unlike measures such as the mean and standard deviation, which can be heavily influenced by outliers, the Box Plot's representation of the quartiles and median provides a more robust summary of the data distribution.

v. Communication of Results: Box Plots are effective tools for communicating data insights to a non-technical audience. The simplicity and intuitive nature of the Box Plot make it a valuable tool for presenting key statistical information in a clear and concise manner.

Section B: Regression Analysis (Questions 7-15)

7) When predicting a continuous target variable, the type of regression typically employed is Linear Regression. Linear Regression is a statistical method used to model the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables) by fitting a linear equation to the observed data points. The goal of Linear Regression is to find the best-fitting line that minimizes the sum of squared differences between the observed values and the values predicted by the model. This linear relationship allows for the prediction of continuous outcomes based on the input variables.

8) The two main types of regression are:

I. Simple Linear Regression:

Simple Linear Regression is a basic form of regression analysis that models the relationship between a single independent variable (predictor variable) and a continuous dependent variable (target variable).

i.The relationship between the variables is represented by a straight line (linear relationship) in a two-dimensional space.

ii.The equation of a simple linear regression model is represented as: $Y = \beta 0 + \beta 1 * X + \varepsilon$, where Y is the dependent variable, X is the independent variable, $\beta 0$ is the intercept, $\beta 1$ is the slope coefficient, and ε is the error term.

Iii. Simple Linear Regression is used when there is a linear relationship between the variables and when we want to predict the value of a continuous outcome based on a single predictor variable.

2. Multiple Linear Regression:

i.Multiple Linear Regression extends the concept of Simple Linear Regression by modeling the relationship between multiple independent variables and a continuous dependent variable.

ii.The relationship between the variables is represented by a linear equation in a multidimensional space.

iii. The equation of a multiple linear regression model is represented as: $Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \ldots + \beta n * Xn + \varepsilon$, where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, $\beta 0$ is the intercept, $\beta 1$, $\beta 2$, ..., βn are the coefficients of the independent variables, and ε is the error term.

iv. Multiple Linear Regression is used when there are multiple predictors influencing the outcome variable and when we want to understand how each predictor contributes to the prediction of the target variable.

9) Simple Linear Regression is used when there is a linear relationship between a single independent variable (predictor variable) and a continuous dependent variable (target variable). It is suitable for scenarios where we want to understand and predict the impact of a single predictor on the outcome variable. Here is an example scenario where Simple Linear Regression would be appropriate:

Example Scenario:

Suppose a company wants to analyze the relationship between the number of years of experience an employee has and their salary. The company believes that there is a linear relationship between the years of experience and the salary earned. In this case, Simple Linear Regression can be used to build a model that predicts an employee's salary based on the number of years of experience they have.

Key Points:

Dependent Variable (Y): Salary

Independent Variable (X): Years of Experience

Objective: To determine how the number of years of experience influences the salary of employees.

Model: The Simple Linear Regression model will estimate the slope and intercept of the line that best fits the relationship between years of experience and salary.

Interpretation: The coefficient of the independent variable (slope) will indicate the change in salary for each additional year of experience.

In this scenario, Simple Linear Regression would be used to quantify and understand the linear relationship between years of experience and salary, allowing the company to make predictions about salary based on an employee's years of experience.

10).In Multiple Linear Regression, multiple independent variables are involved. The model includes more than one predictor variable to predict the continuous dependent variable. The equation for Multiple Linear Regression involves multiple coefficients corresponding to each independent variable, along with an intercept term.

11).Polynomial Regression should be utilized when the relationship between the independent and dependent variables is non-linear. It is suitable when the data points do not follow a straight line pattern. Polynomial Regression can capture more complex relationships by introducing polynomial terms (e.g., quadratic, cubic) into the model. For example, in a scenario where the relationship between temperature and ice cream sales is curvilinear, Polynomial Regression would be preferable over Simple Linear Regression to capture the non-linear trend accurately.

12).In Polynomial Regression, a higher degree polynomial represents a more complex relationship between the independent and dependent variables. Increasing the degree of the polynomial allows the model to fit the data more closely, capturing intricate patterns and fluctuations. However, higher-degree polynomials can lead to overfitting, where the model fits the noise in the data rather than the underlying trend, resulting in reduced generalization to new data.

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

The key difference between Multiple Linear Regression and Polynomial Regression lies in the form of the relationship they can capture. Multiple Linear Regression models linear relationships between multiple independent variables and a dependent variable, while Polynomial Regression can capture non-linear relationships by introducing polynomial terms of the independent variables.

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

Multiple Linear Regression is most appropriate when there are multiple independent variables influencing the dependent variable, and the relationship between the variables is assumed to be linear. It is suitable for scenarios where the outcome is influenced by several predictors, and the goal is to understand how each predictor contributes to the overall prediction.

15). The primary goal of regression analysis is to understand and quantify the relationship between one or more independent variables and a dependent variable. Regression analysis aims to model the relationship, make predictions, and infer insights about how changes in the independent variables affect the dependent variable. It is used for prediction, forecasting, and understanding the underlying patterns in the data.