**Section A: Data Wrangling (Questions 1-6)**

**1. What is the primary objective of data wrangling?**

**a) Data visualization**

**b) Data cleaning and transformation**

**c) Statistical analysis**

**d) Machine learning modeling**

**Ans.** (b)Data cleaning and transformation

The primary objective of data wrangling, also known as data munging or data preparation, is to clean, transform, and prepare data into a suitable form.

**2. Explain the technique used to convert categorical data into numerical data. How does it help in data analysis?**

**Ans.** The technique used to convert categorical data into numerical data is called encoding. There are several methods of encoding categorical data, including Label Encoding, One-Hot Encoding, and Ordinal Encoding.

**🡪Label Encoding**: Label Encoding assigns a unique numerical value to each category in a categorical variable. For example, if we have categories like "red," "blue," and "green," they might be encoded as 0, 1, and 2, respectively. Label Encoding is often used for ordinal categorical variables where there is an inherent order or ranking among categories.

**🡪One-Hot Encoding**: One-Hot Encoding creates binary columns for each category in a categorical variable. Each column represents a category, and if an observation belongs to that category, the corresponding column value is set to 1; otherwise, it is set to 0. One-Hot Encoding is suitable for nominal categorical variables where there is no inherent order among categories.

**🡪Ordinal Encoding**: Ordinal Encoding is similar to Label Encoding but considers the ordinal relationship among categories. For example, if we have categories like "low," "medium," and "high," they might be encoded as 0, 1, and 2, respectively, indicating the order from low to high.

Encoding categorical data into numerical data is a crucial step in data analysis, enabling us to leverage a wide range of analytical techniques and algorithms for extracting insights and making informed decisions from the data.

**3. How does LabelEncoding differ from OneHotEncoding?**

**Ans.** Label Encoding assigns numerical labels to categories in a sequential manner, assuming an ordinal relationship among categories. It is suitable for ordinal categorical variables.

One-Hot Encoding creates binary features for each category, representing them as separate columns with binary values. It is suitable for nominal categorical variables and ensures that categorical variables are not interpreted as ordinal.

**4. Describe a commonly used method for detecting outliers in a dataset. Why is it important to identify outliers?**

**Ans.** Method for detecting outliers in a dataset is the Interquartile Range (IQR) method. The IQR method is based on the concept of quartiles, which divide a dataset into four equal parts.

Importance to Identify outliers:

🡪 Outliers can be indicative of data entry errors, measurement errors, or anomalies in the data collection process. Identifying and addressing outliers improves data quality and accuracy.

🡪 Outliers can distort visualizations such as histograms, box plots, and scatter plots, making it difficult to interpret and analyse the data effectively. Identifying outliers helps create more meaningful and accurate visualizations.

🡪 Outliers can negatively impact the performance of machine learning models by introducing noise, affecting model assumptions, and reducing predictive accuracy. Removing or handling outliers appropriately can improve model performance.

**5. Explain how outliers are handled using the Quartile Method.**

**Ans.** Steps to handle outliers using Quartile Method:

1. Calculating quartiles Q1(first quartile—25th percentile) and Q3(third quartile—75th percentile)
2. Computing inter quartile range: IQR=Q3-Q1
3. Find

Lower bound = Q1-k\*IQR

Upper bound = Q3+k\*IQR

Where k is 1.5 but it can be adjusted on requirement.

1. Values above upper bound and below lower bound are considered as outliers.
2. Handle outliers by removing them, transform data or impute values.
3. **Discuss the significance of a Box Plot in data analysis. How does it aid in identifying potential outliers?**

**Ans.** A Box Plot, also known as a Box-and-Whisker Plot, is a powerful visualization tool used in data analysis to understand the distribution, variability, and potential outliers in a dataset.

Significance of a Box Plot in data analysis and how it aids in identifying potential outliers:

🡪 he shape of the Box Plot can indicate whether the data is symmetrically distributed (box evenly centred), skewed to the left (longer whisker on the left), or skewed to the right (longer whisker on the right).

🡪 One of the main purposes of a Box Plot is to identify potential outliers in the dataset. Outliers are data points that fall significantly outside the range of the bulk of the data, beyond the whiskers of the plot. By visually inspecting the Box Plot, analysts can easily identify data points that may be outliers.

🡪 Box Plots are useful for comparing the distribution and variability of data across different groups or categories. Multiple Box Plots can be displayed side by side to compare distributions visually.

**Section B: Regression Analysis (Questions 7-15)**

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

**variable?**

**Ans.** When predicting a continuous target variable (also known as the dependent variable) in a regression analysis, the type of regression commonly employed is **linear regression**. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables (also known as predictor variables or features). It assumes a linear relationship between the independent variables and the dependent variable, which means that the relationship can be represented by a straight line.

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

**Ans.** There are two main types of linear regression:

**🡪Simple Linear Regression**: In simple linear regression, there is only one independent variable used to predict the dependent variable. The relationship between the independent variable *X* and the dependent variable *Y*.

**🡪Multiple Linear Regression**: In multiple linear regression, there are multiple independent variables used to predict the dependent variable. The relationship between multiple independent variables *X*1​,*X*2​,…,*Xn*​ and the dependent variable.

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

**Ans.** Linear regression is widely used for predicting continuous outcomes such as sales revenue, house prices, stock prices, temperature, and other numerical variables.

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

**involved?**

**Ans.** In multiple linear regression, there can be multiple independent variables involved, as the name suggests. The term "multiple" refers to the fact that there is more than one independent variable used to predict the dependent variable.

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

**Polynomial Regression would be preferable over Simple Linear Regression.**

**Ans.** Polynomial regression should be utilized when the relationship between the independent variable(s) and the dependent variable is not linear but shows a more complex, curved pattern.

**Scenario**: Predicting Sales Revenue Based on Advertising Spending

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

**does it affect the model's complexity?**

**Ans.**  In polynomial regression, the degree of the polynomial represents the highest power of the independent variable(s) included in the regression model.

A higher degree polynomial in polynomial regression provides increased flexibility to capture complex relationships in the data but also increases the risk of overfitting and model complexity. Careful consideration and validation are necessary to choose the optimal degree of the polynomial for a balanced and accurate predictive model.

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

**Regression.**

**Ans.** The key difference is that multiple linear regression is limited to modeling linear relationships between variables, while polynomial regression can model non-linear relationships by including higher-order polynomial terms in the regression equation. Polynomial regression provides more flexibility to capture complex patterns and curvature in the data compared to multiple linear regression. However, increasing the degree of the polynomial also increases the model's complexity and the risk of overfitting.

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

**regression technique.**

**Ans.**  Multiple linear regression is most appropriate in scenarios where there are multiple independent variables (predictor variables) that are believed to have a linear relationship with the dependent variable (target variable).

**Scenario: Predicting Housing Prices**

By using multiple linear regression, the real estate agency can analyze and quantify the relationships between various factors and housing prices, identify the most influential variables, and make more informed decisions regarding pricing, marketing strategies, and property investments.

**15. What is the primary goal of regression analysis?**

**Ans.** The primary goal of regression analysis is to understand and quantify the relationship between a dependent variable (also known as the outcome variable or target variable) and one or more independent variables (also known as predictor variables or features).