

Section A: Data Wrangling

1. What is the primary objective of data wrangling?

The primary objective of data wrangling is:

b) Data cleaning and transformation

Explanation: Data wrangling involves preparing and organizing raw data for analysis, which includes tasks such as cleaning, transforming, and restructuring the data.

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

One common technique is **One-Hot Encoding**, where each category is represented by a binary (0 or 1) value in a separate column. This helps in numerical analysis by ensuring that the categorical values don't imply a false ordinal relationship.

3. How does LabelEncoding differ from OneHotEncoding?

LabelEncoding assigns a unique numerical label to each category, introducing ordinal relationships.

OneHotEncoding creates binary columns for each category, avoiding ordinal implications.

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

A commonly used method is the **Z-score method**, where data points significantly far from the mean are considered outliers. Identifying

outliers is crucial as they can distort statistical analyses and machine learning models, leading to inaccurate results.

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

The Quantile Method involves setting thresholds based on percentiles (e.g., 25th and 75th percentiles). Values beyond these thresholds are considered outliers and can be replaced or removed. This method is less sensitive to extreme values than Z-score.

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

A Box Plot visually represents the distribution of data, providing information about the median, quartiles, and potential outliers. It helps identify skewed data and outliers through visual inspection of the plot's "whiskers" and individual data points.

Section B: Regression Analysis

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

Linear Regression is employed when predicting a continuous target variable.

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

Linear Regression: Predicts a continuous target variable using a linear equation.

Logistic Regression: Predicts the probability of a binary outcome using a logistic function.

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

Use Simple Linear Regression when there is a linear relationship between a single independent variable and the dependent variable. Example: predicting house prices based on square footage.

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

Multi Linear Regression involves more than one independent variable.

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 and dependent variables is nonlinear. For example, predicting the height of a plant based on time might involve a curve rather than a straight line.

12. What does a higher degree polynomial represent in Polynomial Regression? How does it affect the model's complexity?

A higher degree polynomial represents a more complex curve. Higher-degree polynomials can fit the training data more closely, but they may also lead to overfitting, capturing noise in the data rather than the underlying pattern.

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

The key difference is in the nature of the relationship: Multi Linear Regression assumes a linear relationship between variables, while Polynomial Regression can capture nonlinear relationships through the use of polynomial functions.

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

Multi Linear Regression is most appropriate when there are multiple independent variables influencing a single dependent variable. For example, predicting a person's income based on education level, years of experience, and age.

15. What is the primary goal of regression analysis?

The primary goal of regression analysis is to understand the relationship between the independent variables and the dependent variable, make predictions, and quantify the strength and nature of those relationships.