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?

One technique is Label Encoding, where categories are assigned numerical labels. It helps in incorporating categorical data into machine learning models, as they often require numerical inputs.

3How does LabelEncoding differ from OneHotEncoding?

LabelEncoding converts each category into a unique integer, while OneHotEncoding creates binary columns for each category. OneHotEncoding avoids the ordinal implication of LabelEncoding and is preferred when categories are not ordered.

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

Z-score method is commonly used for outlier detection. It's important to identify outliers as they can significantly impact statistical analyses and machine learning models, leading to skewed results.

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

The Quantile Method involves setting thresholds based on the data distribution. Values outside the defined quantiles are considered outliers and can be removed or transformed to improve the overall dataset quality.

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, displaying quartiles and potential outliers. It aids in identifying outliers by highlighting values beyond the whiskers, providing a clear visualization of the data's spread.

Section B: Regression Analysis (Questions 7-15)

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.

The two main types of regression are Simple Linear Regression (one independent variable) and Multiple Linear Regression (more than one independent variable).

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

Simple Linear Regression is used when there is a linear relationship between the independent and dependent variables. For example, predicting a student's score based on the number of hours studied.

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

Multiple Linear Regression involves multiple independent variables. Multiple Linear Regression involves multiple independent variables.

11.When should Polynomial Regression be utilized? Provide a scenario where Polynomial Regression would be preferable over Simple Linear Regression.

Polynomial Regression is utilized when the relationship between the independent and dependent variables is non-linear. For instance, predicting house prices based on both square footage and the square of footage.

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

A higher degree polynomial in Polynomial Regression represents a more complex curve. It can capture intricate patterns in the data but may lead to overfitting, making the model less generalizable.

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

Multi Linear Regression involves multiple linear relationships, while Polynomial Regression captures non-linear relationships by introducing polynomial terms.

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

Multi Linear Regression is appropriate when there are multiple independent variables influencing the dependent variable. For example, predicting a person's income based on education, experience, and location.

15.What is the primary goal of regression analysis?

The primary goal of regression analysis is to understand and quantify the relationship between independent and dependent variables, enabling predictions and insights based on the observed data.