1. b) Data cleaning and transformation

2. Technique to Convert Categorical Data into Numerical Data:

The technique commonly used to convert categorical data into numerical data is called Encoding. Encoding is essential because many machine learning algorithms and statistical models require numerical input.

One common method is One-Hot Encoding, which creates binary columns for each category and represents the presence or absence of the category with a 1 or 0. Another method is Label Encoding, where each category is assigned a unique numerical label.

3. Difference Between Label Encoding and One-Hot Encoding:

Label Encoding: It assigns a unique integer to each category. The order of the assigned numbers may imply an ordinal relationship, which might not be suitable for some algorithms.

One-Hot Encoding: It creates binary columns for each category, eliminating the ordinal relationship issue. Each category is represented by a binary column, and only one of them is "hot" (1) for each row.

4. Method for Detecting Outliers:

A commonly used method for detecting outliers is the Interquartile Range (IQR) method. The IQR is the range between the first quartile (25th percentile) and the third quartile (75th percentile) of the data. Outliers are identified as values beyond a certain distance from the quartiles.

5. Handling Outliers Using the Quantile Method:

The Quantile Method involves setting a threshold based on quantiles (percentiles) of the data. Values beyond a certain quantile range are considered outliers and can be either removed or transformed. This method helps in ensuring that extreme values don't unduly influence statistical analysis or machine learning models.

6. Significance of a Box Plot in Data Analysis:

A Box Plot (Box-and-Whisker Plot) is a graphical representation that displays the distribution of a dataset. It includes the median, quartiles, and potential outliers. It aids in identifying the following:

Center and Spread: The box represents the interquartile range, giving an indication of data spread.

Skewness: The length of the whiskers helps identify the skewness of the distribution.

Outliers: Outliers can be visually identified as points beyond the whiskers.

The Box Plot is a powerful tool for summarizing and visually presenting the key characteristics of a dataset, making it easier to identify patterns, outliers, and potential issues.

7. Type of Regression for Predicting a Continuous Target Variable:

The type of regression employed when predicting a continuous target variable is Linear Regression.

8. Two Main Types of Regression:

Linear Regression: It models the relationship between the dependent variable and one or more independent variables by fitting a linear equation to the observed data.

Logistic Regression: Despite its name, logistic regression is used for binary classification problems. It models the probability that the dependent variable belongs to a particular category.

9. When to Use Simple Linear Regression:

Simple Linear Regression is used when there is a linear relationship between the independent and dependent variables. It is suitable when you have only one independent variable. Example scenario: predicting a student's exam score based on the number of hours they studied.

10. Number of Independent Variables in Multi Linear Regression:

Multi Linear Regression involves more than one independent variable. It can include two or more independent variables.

11. When to Use Polynomial Regression:

Polynomial Regression is utilized when the relationship between the independent and dependent variables is nonlinear. It is preferable over Simple Linear Regression when the data exhibits a curved or non-linear pattern.

12. Higher Degree Polynomial in Polynomial Regression:

A higher degree polynomial in Polynomial Regression represents a more complex model. It allows the algorithm to fit the data more closely but may also lead to overfitting. Higher degree polynomials introduce more flexibility, capturing intricate patterns in the data.

13. Key Difference Between Multi Linear Regression and Polynomial Regression:

The key difference lies in the relationship between the independent variables:

Multi Linear Regression: Involves multiple independent variables with a linear relationship to the dependent variable.

Polynomial Regression: Involves a single independent variable raised to various powers or multiple independent variables with nonlinear relationships.

14. Scenario for Using Multi Linear Regression:

Multi Linear Regression is appropriate when there are multiple independent variables, and the relationship with the dependent variable is assumed to be linear. For example, predicting house prices based on features like square footage, number of bedrooms, and location.

15. Primary Goal of Regression Analysis:

The primary goal of regression analysis is to model and understand the relationship between the dependent variable and one or more independent variables, making predictions or inferences about the dependent variable based on the independent variables. It aims to quantify the strength and nature of these relationships.