ASSESSMENT-2

1. What is the primary objective of data wrangling? • a) Data visualization • b) Data cleaning and transformation • c) Statistical analysis • d) Machine learning modeling

A)The primary objective of data wrangling is Data transformation and cleaning. Data wrangling involves the process of cleaning, structuring, and transforming raw data into a suitable format for analysis. This often includes tasks such as handling missing data, removing duplicates, converting data types, and restructuring data to make it more suitable for analysis.

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

One hot encoding is a technique that we use to represent categorical variables as numerical values in a machine learning model. One hot encoding creates new (binary) columns, indicating the presence of each possible value from the original data.

3. How does LabelEncoding differ from OneHotEncoding?

Label Encoding: Assign each categorical value an integer value based on alphabetical order.

One Hot Encoding: Create new variables that take on values 0 and 1 to represent the original categorical values.

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

A) Standard Deviation (σ): In statistics, standard deviation measures the spread of data around the mean. It captures how far away from the mean the data points are. For data that is normally distributed:

   a. Approximately 68.2% of the data lie within one standard deviation from the mean.
   b. Close to 95.4% lie within two standard deviations.
   c. About 99.7% lie within three standard deviations.

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

A) The Quantile Method (using the Interquartile Range or IQR) is a robust technique for identifying outliers in a dataset. Here's how it works:

Calculate the first quartile (Q1) and the third quartile (Q3).

Define the lower bound as Q1 - 1.5 × IQR and the upper bound as Q3 + 1.5 × IQR.

Any data point outside these bounds is considered an outlier

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

A) A Box Plot, also known as a box-and-whisker plot, is a powerful data visualization tool used in data analysis. Let's explore its significance and how it helps identify potential outliers:

1. Visual Summary of Data:
   - A Box Plot provides a concise summary of the distribution of numerical data.
   - It displays key statistical measures, including the median, quartiles, and potential outliers.
2. Components of a Box Plot:
   - Box: Represents the interquartile range (IQR), covering the middle 50% of the data.
   - Whiskers: Extend from the box to the minimum and maximum values within a certain range.
   - Outliers: Shown as individual points outside the whiskers.
3. Significance:
   - Identifying Outliers: Outliers are data points significantly different from the rest. Box plots help detect them visually.
   - Skewness: The position of the median within the box indicates skewness (symmetric, positively skewed, or negatively skewed).
   - Spread and Dispersion: The length of the whiskers shows the spread of data.
   - Comparison: Box plots allow easy comparison of multiple datasets side by side.

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

A)When predicting a **continuous target variable**, the most common type of regression to use is **Linear Regression**. Linear regression models the relationship between one or more independent variables and a continuous dependent variable. It estimates parameters by minimizing the sum of squared errors (SSE) and can handle both linear and curved relationships. Other advanced types of linear regression include polynomial regression (which models curvature) and interaction effects. Linear regression is a robust choice for continuous outcomes.

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

1. Linear Regression

Linear regression is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable.

2. Logistic Regression

Logistic regression is used to fit a regression model that describes the relationship between one or more predictor variables and a binary response variable**.**

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

Simple Linear Regression is used when we want to understand the relationship between one predictor variable (also called independent variable) and one response variable (also called dependent variable). Here are two real-life examples:

1. Advertising Spending and Revenue:
    - Scenario: A business wants to understand how their advertising spending affects revenue.
    - Variables:
        - Predictor Variable (X): Advertising spending (in dollars).
        - Response Variable (Y): Revenue (in dollars).

      Interpretation:

        - (\beta_0) represents the expected revenue when ad spending is zero.
        - (\beta_1) represents the average change in revenue for each additional dollar spent on advertising.
    - Decision: Depending on the value of (\beta_1), the company may adjust their ad spending.

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

A) In Multiple Linear Regression**,** typically two or more independent variables are involved. This allows us to model the relationship between these predictors and a single continuous response variable. For example, we might use multiple linear regression to analyze how both temperature and humidity affect crop yield.

11. When should Polynomial Regression be utilized?

A) Polynomial Regression is a technique used when the relationship between the predictor variable(s) and the response variable is nonlinear. Here are three ways to determine if you should use polynomial regression:

1. Scatterplot Inspection:
    - Create a scatterplot of the predictor variable and the response variable.
    - If the scatterplot shows a nonlinear pattern, consider using polynomial regression.

- Example: If hours studied vs. exam score exhibits a curved relationship, polynomial regression may be appropriate.

2. Fitted Values vs. Residuals Plot:
   - Fit a linear regression model and create a fitted values vs. residuals plot.
   - If residuals exhibit a clear nonlinear pattern (e.g., a "U" shape), consider polynomial regression.
   - Example: If residuals show a curved trend, polynomial regression might be better.
3. Adjusted R-Squared Comparison:
   - Fit both linear and polynomial regression models.
   - Calculate the adjusted R-squared for both models.
   - The model with the higher adjusted R-squared explains more variation in the response variable.
   - Example: If the adjusted R-squared is higher for the polynomial model, choose it

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

A) In Polynomial Regression, a higher degree polynomial represents a more complex relationship between the predictor variable(s) and the response variable. Let's explore this further:

Degree of the Polynomial:

a. The degree of the polynomial determines how many terms (powers) of the predictor variable(s) are included in the model.
b. For example, a quadratic polynomial has terms up to the second power , while a cubic polynomial includes terms up to the third power .

Effect on Complexity:

c. As the degree increases, the model becomes more flexible and can fit nonlinear patterns in the data.
d. However, higher-degree polynomials can also lead to overfitting:
   i. Overfitting: The model captures noise and random fluctuations in the training data, making it perform poorly on new, unseen data.
   ii. Complexity Trade-off: Increasing the degree balances better fit to the training data with the risk of overfitting.

Choosing the Degree:

e. Domain Knowledge: Choose the degree based on your understanding of the problem and the underlying relationship.
f. Cross-Validation: Use techniques like cross-validation to find the optimal degree that minimizes overfittin

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

1. Linear Regression (MLR):
    - Assumption: Assumes a linear relationship between the dependent variable and multiple independent variables.
    - Equation: Uses a straight line equation to represent the relationship.
    - Use Case: Suitable when the relationship is approximately linear.
    - Example: Predicting house prices based on square footage, number of bedrooms, and location.
2. Polynomial Regression:
    - Assumption: Does not assume a linear relationship; it models nonlinear patterns.
    - Equation: Uses a polynomial equation (e.g., quadratic or cubic) to represent the relationship.
    - Complexity: More complex due to fitting a curve rather than a straight line.
    - Example: Modeling the growth of a plant based on time (using a quadratic term).

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

Multiple Linear Regression (MLR) is most appropriate in the following scenarios:

1. Multiple Predictors:
    - When you have two or more independent variables (predictors) that potentially influence a single dependent variable (response).
    - Example: Predicting crop yield based on rainfall, temperature, and fertilizer usage.
2. Complex Relationships:
    - When the relationship between predictors and the response is not strictly linear.
    - MLR can capture combined effects and interactions among predictors.
    - Example: Modeling how both education level and work experience affect salary.
3. Business and Social Sciences:
    - MLR is widely used in fields like economics, marketing, social sciences, and finance.
    - Example: Analyzing factors affecting stock prices (e.g., interest rates, company performance, market indices).
4. Controlled Experiments:
    - When conducting controlled experiments with multiple manipulated variables.
    - Example: Testing the impact of advertising spend, pricing, and product quality on sales.

15)What is the primary goal of regression analysis?

A)The primary goal of regression analysis is to model and understand the relationship between a dependent variable (also called the response variable or outcome variable) and one or more independent variables (also called predictor variables or explanatory variables). By estimating the parameters of the regression model, we can:

- Predict the value of the dependent variable based on the independent variables.
- Identify which independent variables significantly influence the dependent variable.
- Assess the strength and direction of these relationships. Regression analysis helps us make informed decisions, predict outcomes, and gain insights into the underlying processes.