Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: From the analysis of categorical variables, such as season, holiday, and weathersit, we can infer that these factors significantly influence bike demand. For example**, demand tends to be higher during warmer seasons (summer and spring) compared to colder ones (fall and winter).** Additionally, bike rentals are likely to increase on holidays as people engage in leisure activities. The weather condition also impacts demand; clear weather leads to higher rentals, whereas adverse weather conditions like rain or snow decrease demand. Analyzing these variables helps identify patterns and relationships with the target variable, enabling better demand forecasting.

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Answer: Using drop_first=True when creating dummy variables is important to avoid the "dummy variable trap," which occurs due to multicollinearity. This situation arises when one variable can be perfectly predicted from the others, leading to redundancy. By dropping the first category of a categorical variable, we reduce redundancy and maintain the integrity of the model. This approach allows us to interpret the coefficients meaningfully, as the remaining dummy variables represent changes relative to the dropped category.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Based on the pair-plot analysis among numerical variables, the variable **temp (temperature)** shows the highest correlation with the target variable cnt (total bike rentals). The scatterplot indicates a strong positive relationship, suggesting that as the temperature increases, bike rentals also tend to rise significantly.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: To validate the assumptions of Linear Regression after building the model, we conducted several checks:

i. Linearity: We examined scatterplots of predicted vs. actual values to ensure a linear relationship.

ii. Normality of Residuals: We created histograms and Q-Q plots of residuals to verify that they follow a normal distribution.

iii. Homoscedasticity: We plotted residuals against predicted values to check for constant variance, ensuring that the spread of residuals is consistent across all levels of predicted values.

iv. Independence: We ensured that the residuals were independent by observing the patterns in residual plots.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes? (2 marks)

Answer: Based on the final model, the top three features contributing significantly to explaining the demand for shared bikes are:

i. Temperature (temp): This variable has a strong positive correlation with bike rentals, indicating that higher temperatures lead to increased usage.

ii. Weather Situation (weathersit): Different weather conditions significantly affect bike demand, with clear weather leading to higher rentals.

iii. Season (season): The seasonality effect also plays a crucial role, as bike rentals vary significantly between different seasons, particularly peaking in summer.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The core idea is to find a linear equation that best predicts the target variable based on the features.

The general form of a linear regression equation is:

$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn+\epsilon$

where:

i. Y is the dependent variable.

ii. $\beta 0$ is the intercept.

iii. $\beta 1,\beta 2,...,\beta n$ are the coefficients of the independent variables $X1,X2,...,Xn$.

iv. ϵ represents the error term.

Steps in Linear Regression:

i. Data Preparation: Clean and preprocess the dataset to handle missing values and categorical variables.

ii. Model Training: Fit the model to the training data by minimizing

the residual sum of squares (RSS) between the observed and predicted values.

iii. Evaluation: Assess model performance using metrics such as R-squared, Mean Squared Error (MSE), and residual analysis.

iv. Assumption Checks: Validate the assumptions of linear regression, including linearity, independence, homoscedasticity, and normality of residuals.

Linear regression is widely used due to its simplicity and interpretability, making it a fundamental technique in statistical modeling and machine learning.

2. Explain Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line) yet appear very different when graphed. The quartet was created by statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to illustrate how statistical properties can be misleading.

The Four Datasets:

Dataset I: Shows a linear relationship with a positive slope.

Dataset II: Shows a nonlinear relationship with a parabolic shape.

Dataset III: Contains a single outlier that significantly affects the linear regression results.

Dataset IV: Shows a vertical distribution, indicating no correlation.

Key Takeaway:

Anscombe's quartet highlights that relying solely on summary statistics can lead to incorrect interpretations. Visualizations are crucial for understanding data distributions and relationships, as they can reveal

patterns, outliers, and anomalies that summary statistics may obscure.

3. What is Pearson's R? (3 marks)

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

i. 1 indicates a perfect positive linear correlation (as one variable increases, the other also increases).

ii. -1 indicates a perfect negative linear correlation (as one variable increases, the other decreases).

iii. 0 indicates no linear correlation between the variables.

Importance:

Pearson's R is widely used in statistics to quantify relationships between variables and is essential for hypothesis testing and building regression models. However, it only captures linear relationships and can be affected by outliers.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is the process of transforming the features in a dataset to a common scale without distorting differences in the ranges of values. This step is crucial in many machine learning algorithms, particularly those that rely on distances between data points, such as k-nearest neighbors and gradient descent.

Reasons for Scaling:

i. Improved Convergence: Algorithms that use gradient descent

converge faster when features are on a similar scale.

ii. Equal Weightage: Scaling ensures that features with larger ranges do not dominate the model's performance.

iii. Distance Metrics: In algorithms that compute distances, unscaled features can lead to biased results.

Types of Scaling:

i. **Normalization** (Min-Max Scaling): This technique rescales the feature to a fixed range, usually [0, 1]. It is calculated as:

$$X' = (X - \min(X)) / (\max(X) - \min(X))$$

This method is sensitive to outliers.

i. **Standardization** (Z-score Scaling): This technique rescales the feature to have a mean of 0 and a standard deviation of 1. It is calculated as:

$$X' = (X - \mu) / \sigma$$

where μ is the mean and σ is the standard deviation of the feature. Standardization is less sensitive to outliers and is often preferred when the dataset follows a normal distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The Variance Inflation Factor (VIF) quantifies how much the variance of an estimated regression coefficient increases when your predictors are correlated. A VIF value of 1 indicates no correlation, while values greater than 1 indicate some correlation. A VIF value can become infinite when there is perfect multicollinearity among the independent variables.

Reasons for Infinite VIF:

i. Perfect Multicollinearity: If one independent variable is a

perfect linear combination of one or more other variables, it causes the determinant of the correlation matrix to be zero, resulting in an infinite VIF.

ii. Duplicated Variables: Including the same variable multiple times in the model or having identical columns in the dataset can lead to infinite VIF values.

Implications:

Infinite VIF values suggest that the model may not be reliable due to redundancy among predictors. It is crucial to identify and address multicollinearity by removing or combining correlated features.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, typically the normal distribution. It compares the quantiles of the sample data against the quantiles of the theoretical distribution.

How It Works:

i. The data points are sorted and plotted against the expected quantiles from the normal distribution.

ii. If the points fall approximately along a straight line (typically the 45-degree line), the data is considered to follow the normal distribution.

Use and Importance in Linear Regression:

i. Normality Assumption: Linear regression assumes that the residuals (errors) are normally distributed. A Q-Q plot helps validate this assumption.

ii. Model Diagnostics: By visualizing the distribution of residuals, researchers can identify potential deviations from normality,

which could indicate issues with the model fit or the presence of outliers.

iii. Improving Model Performance: If the residuals are not normally distributed, transformations may be necessary to improve model performance and interpretability.