Assignment-based Subjective Questions

1. Analysis of Categorical Variables and Their Impact on the Dependent Variable (3 marks)

Upon examining the categorical variables, it became evident that some of them considerably influence the dependent variable, the count of bike rentals (cnt). For instance:
- Season: Different seasons demonstrate varying bike rental counts, with higher numbers in summer.
- Month: There is a clear variation in rentals across months, showing a seasonal trend.
- Weather Situation: Weather conditions greatly affect bike rentals, with favorable weather increasing rentals.
- Holiday: Rentals tend to be lower on holidays compared to regular workdays.
These findings indicate that both environmental and temporal factors significantly impact bike rental demand.

2. Importance of Using drop_first=True in Dummy Variable Creation (2 marks)

Setting drop_first=True when creating dummy variables is crucial to avoid multicollinearity. Without this setting, one category could be perfectly predicted by the others, resulting in redundant information. Dropping the first category prevents the dummy variable trap, ensuring the model remains reliable without perfect multicollinearity.

3. Numerical Variable with Highest Correlation with Target Variable (1 mark)

From the pair-plot analysis, it was noted that the variable 'atemp' (apparent temperature) has the highest correlation with the target variable, 'cnt' (count of bike rentals).

4. Validating Assumptions of Linear Regression After Model Building (3 marks)

After constructing the linear regression model, the following steps were taken to validate its assumptions:
- Linearity: Verified by plotting predicted vs. actual values to ensure a linear relationship.
- Homoscedasticity: Checked through a residual plot to confirm constant variance of residuals.
- Normality of Residuals: Assessed using a Q-Q plot and residual histogram to see if residuals follow a normal distribution.
- Multicollinearity: Examined using the Variance Inflation Factor (VIF) to ensure no high correlation between independent variables.

5. Top 3 Features Contributing to Bike Rental Demand (2 marks)

According to the final model, the top three features significantly contributing to the demand for shared bikes are:

- 'atemp' (apparent temperature)
- 'yr' (year, indicating trends over years)
- 'season_3' (indicating the third season)

General Subjective Questions

1. Detailed Explanation of the Linear Regression Algorithm (4 marks)

Linear regression is a technique for modeling the relationship between a dependent variable and one or more independent variables. It aims to find a linear equation of the form:
$$ y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_nx_n + \epsilon $$
where:
- $ y $ is the dependent variable.
- $ \beta_0 $ is the y-intercept.
- $ \beta_1, \beta_2, \ldots, \beta_n $ are coefficients of the independent variables $ x_1, x_2, \ldots, x_n $.
- $ \epsilon $ is the error term.

The algorithm involves:
1. Fitting the Model: Using the least squares method to minimize the sum of squared residuals (differences between observed and predicted values).
2. Evaluating the Model: Assessing performance using metrics like R-squared, adjusted R-squared, and p-values of coefficients.
3. Checking Assumptions: Validating assumptions of linearity, independence, homoscedasticity, normality of residuals, and absence of multicollinearity.

2. Explanation of Anscombe's Quartet (3 marks)

Anscombe's quartet consists of four datasets with nearly identical simple descriptive statistics but differing distributions and relationships. It underscores the importance of visualizing data:
1. First dataset: Exhibits a typical linear relationship.
2. Second dataset: Shows a parabolic relationship.
3. Third dataset: Demonstrates a linear relationship with an outlier.
4. Fourth dataset: Features a linear relationship with a single influential point.
The quartet illustrates that relying solely on summary statistics can be misleading and highlights the necessity of data visualization for accurate analysis.

3. Explanation of Pearson's R (3 marks)

Pearson's R, or the Pearson correlation coefficient, quantifies the linear relationship between two variables, ranging from -1 to 1:
- 1: Perfect positive linear correlation.

- -1: Perfect negative linear correlation.
- 0: No linear correlation.
It is calculated as the covariance of the variables divided by the product of their standard deviations.

4. Explanation of Scaling, Including Normalized and Standardized Scaling (3 marks)

Scaling adjusts data to a standard range or distribution to:
- Enhance the performance of algorithms sensitive to data scale.
- Ensure features equally contribute to model training.

Normalized Scaling: Adjusts data to a [0, 1] range, suitable for bounded data distributions.
Standardized Scaling: Transforms data to have a mean of 0 and standard deviation of 1, preserving the distribution shape and suitable for unbounded data distributions.

5. Explanation of Infinite VIF Values (3 marks)

VIF (Variance Inflation Factor) becomes infinite when perfect multicollinearity occurs, meaning one predictor is a perfect linear combination of others. This results in an undefined or infinite VIF, signaling redundancy in predictors and complications in model estimation.

6. Explanation and Importance of a Q-Q Plot in Linear Regression (3 marks)

A Q-Q (Quantile-Quantile) plot assesses whether a dataset follows a specific distribution, typically the normal distribution. It plots the data's quantiles against theoretical distribution quantiles. In linear regression, Q-Q plots are used to:
- Verify if residuals follow a normal distribution.
- Validate the normality assumption, essential for hypothesis testing and constructing confidence intervals.
Deviations from the diagonal line in a Q-Q plot indicate departures from normality.