Assignment-Based Subjective Questions:

1. Inference from Categorical Variables Analysis:

Ans: Categorical variables in the dataset include `season`, `yr`, `mnth`, `holiday`, `weekday`, and `weathersit`. From the analysis: - Fall, especially in September, sees the maximum active customers. - The year 2019 observed more sales than 2018. - Holidays negatively impact active counts. - Heavy rain results in no users, while partly cloudy/clear sky sees the maximum count.

2. Importance of `drop_first=True` in Dummy Variable Creation:

Ans: Using `drop_first=True` during dummy variable creation is essential to avoid correlation between dummy variables, preventing redundancy in the analysis.

3. Numerical Variable with Highest Correlation:

Ans: 'atemp' and 'temp' exhibit the highest correlation with the target variable, as observed in the pair plot.

4. Validation of Linear Regression Assumptions:

Ans: The assumption that error terms correspond to a normal curve is validated through a histogram analysis, ensuring the model's robustness.

5. Top 3 Features Contributing to Bike Demand:

Ans: The top three features significantly contributing to bike demand are 'Temp,' 'Year' (positively influencing), and snowy/rainy weather (negatively influencing), based on their coefficients.

General Subjective Questions:

6. Explanation of Linear Regression Algorithm:

Ans: Linear regression is an interpolation technique predicting correlations between variables. After data exploration and cleaning, the dataset is split into training and testing sets. Features are selected, collinearity is checked, and the model is iteratively refined, considering R-values and p-values. The model is then tested against the assumptions, providing insights and predictions.

7. Explanation of Anscombe's Quartet:

Ans: Anscombe's Quartet consists of four datasets with identical descriptive statistics but different characteristics. It highlights that a regression model can be misled by distinct datasets that appear similar upon training.

8. Pearson's R:

Ans: Pearson's correlation coefficient, or Pearson's R, measures the strength of correlation between two variables, ranging from -1 to +1. The values denote the linear correlation between variables, with +1 indicating a perfect positive correlation and -1 indicating a perfect negative correlation.

9. Scaling and Differences Between Normalized and Standardized Scaling:

Ans: Scaling is essential for model functionality, ensuring variable coefficients are within an appropriate range. Normalized scaling focuses on Gaussian distribution without a preset range, often used in neural networks. Standardized scaling compresses variable values into a specific range, aiding the model.

10. VIF and Infinite Values:

Ans: Infinite VIF values occur when there is perfect correlation between the dependent and independent variables, resulting in an R-squared value close to 1.

11. Q-Q Plot and Its Importance in Linear Regression:

Ans: Q-Q plot assesses if data comes from the same statistical distribution, crucial in linear regression with separate testing and training datasets. It ensures both datasets share the same background for model integrity.