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?

Ans: Categorical variables in the dataset include `season`, `yr`, `mnth`, `holiday`, `weekday`, and `weathersit`. The inference about the effect on the dependent variable would be as below.

- 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. Why is it important to use drop_first=True during 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. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

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

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

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

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

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. Explain the linear regression algorithm in detail.

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. Explain the Anscombe's quartet in detail.

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. What is 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. What is scaling? Why is scaling performed? What is the difference between normalized scaling 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. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

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. What is a Q-Q plot? Explain the use and importance of a Q-Q plot 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.