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?**
   The coefficients of the categorical variables (seasons, weather conditions, etc.) indicate their impact on the target variable (bike rental count). For example, season_2, season_3, and season_4 show substantial positive coefficients, indicating increased bike rentals in these respective seasons compared to the base season (probably season_1). Similarly, weathersit_2 and weathersit_3 indicate adverse weather conditions negatively affect bike rentals.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Using drop_first=True helps to prevent multicollinearity by removing the first category in each categorical variable. It reduces redundancy and avoids perfect multicollinearity that might affect the regression model's performance.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Among the numerical variables, temp exhibits the highest positive correlation with the target variable (cnt), suggesting a positive relationship between temperature and bike rentals.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Various assumptions of linear regression, like linearity, normality of residuals, homoscedasticity, and independence of residuals, are typically validated using diagnostic plots like residuals vs. fitted values plot, QQ plot, residual autocorrelation plot, etc., and statistical tests.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   temp: Positive coefficient indicates higher temperatures correlate with increased bike rentals.
   weathersit_3: Strong negative coefficient suggests severe weather conditions greatly reduce bike rentals.
   season_4: Positive coefficient indicates higher rentals during the fourth season compared to the base season.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable and one or more independent variables. Its aim is to establish a linear equation that predicts the value of the dependent variable based on the independent variable(s).

Understanding Linear Regression:

Model Representation:

In a simple linear regression with one independent variable, the model is represented as:

Y=Beta0 + Beta1 + X + e

Where,

Y represents the dependent variable (target).

X represents the independent variable (predictor).

Beta0 is the intercept (where the line intersects the y-axis).

Beta1 is the slop line

e represents the error term (the difference between predicted and actual values).

Objective:

The goal of linear regression is to find the best-fitting line that minimizes the sum of squared differences between predicted and actual values. This is achieved by estimating the coefficients that minimize the sum of squared residuals (also known as the "least squares" method).

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet comprises four datasets that have nearly identical statistical properties despite vastly different distributions. It emphasizes the importance of data visualization in understanding the data.

3. **What is Pearson's R?**

   Pearson's correlation coefficient measures the linear relationship between two variables, ranging from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 a perfect negative linear relationship, and 0 no linear relationship.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Scaling transforms numerical variables to a standard scale to ensure all features contribute equally to model fitting. Normalization and standardization are two common scaling techniques; normalization scales features between 0 and 1, while standardization rescales features to have mean 0 and variance 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

   Infinite VIF values often indicate multicollinearity issues in the dataset, where one predictor can be linearly predicted from the others, causing numerical instability in the regression model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?**

   The Q-Q plot (Quantile-Quantile plot) helps visualize if a dataset is approximately normally distributed. It plots the quantiles of the observed data against the quantiles of a theoretical normal distribution, aiding in assessing normality assumptions in linear regression.