General Subjective Questions:

1. **Explanation of Linear Regression Algorithm:**

   Linear regression endeavors to elucidate the relationship between independent and dependent variables using a straight line, applicable solely to numerical variables. The algorithm follows these steps:

   - The dataset is partitioned into test and training data.
   - Training data is segregated into feature (independent) and target (dependent) datasets.
   - A linear model is fitted using the training dataset, often utilizing gradient descent algorithm to determine the coefficients of the best fit line by minimizing a cost function (e.g., residual sum of squares).
   - In the case of multiple features, the predicted variable is a hyperplane.
   - The predicted variable is then compared with the test data, and assumptions are evaluated.

2. **Explanation of Anscombe's Quartet:**

   Anscombe's quartet comprises four datasets sharing similar simple descriptive statistics but demonstrating distinct distributions when graphically visualized. These datasets challenge the reliance solely on simple statistics for data interpretation, emphasizing the importance of graphical analysis to discern subtle nuances in data patterns.

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

   Pearson's R quantifies the strength of association between two variables, calculated as the covariance divided by the product of their standard deviations. Its values range from +1 to -1, indicating perfect positive linear correlation, no correlation, and perfect negative linear correlation, respectively

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

   Scaling standardizes variables to a specific range, a preprocessing step in linear regression. Two common types are:

   - Normalized Scaling: Scales variables to a range of 0–1, suitable for non-Gaussian distributions.
   - Standardized Scaling: Centers variables around mean with a unit standard deviation, suitable for Gaussian distributions.

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

   VIF (Variance Inflation Factor) becomes infinite when the R square is 1, indicating perfect correlation between features. - $\mathrm{VIFi} = 1/1 - Ri\,2$

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

   A Q-Q plot compares two sets of quantiles, serving to check if they originate from the same distribution, aiding in visual data assessment.

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?

Analysis of categorical variables suggests several insights regarding their impact on the dependent variable (Count), including seasonal variations, trends over time, holiday effects,

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

Setting drop_first=True ensures that a variable with n levels can be represented by n-1 dummy variables, simplifying the representation while avoiding multicollinearity issues.

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 correlation coefficient of 0.63 with the target variable.

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

Assumptions of linear regression are validated post-model building by examining the distribution of residuals, which should ideally form a normal distribution with a mean of 0.

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

The top three features significantly contributing to bike demand in the final model are identified as 'Holiday', 'Temp', and 'Humidity'.