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?**
   Among the categorical variables in the dataset (season, yr, mnth, holiday, weekday, workingday, and weathersit), we can infer their effects on the dependent variable as follows:
   - Fall is the season with the highest number of active customers.
   - In 2019, the number of active customers is higher than in 2018.
   - June and July are the months with the highest activity levels observed throughout the year.
   - During holidays, the count of active customers is lower than on weekdays.
   - The weathersit condition "Light snow" corresponds to the lowest customer count observed.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   drop_first=True will remove the first column from the dummy variable's dataframe, which will help to reduce the correlation created among the dummy variables and prevent over parameterization.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   From the pair plot it is evident that "temp" and "atemp" numerical variables has the highest correlation with the target variable("cnt").

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   After building the linear regression model, I validated the assumptions through various checks. Firstly, I created a histogram to examine the distribution of the error terms, and it is indicative of a normal distribution. This suggests that the residuals follow a bell-shaped curve, aligning with the assumption of normality in linear regression. Additionally, I generated a scatter plot to inspect the relationship between the predicted values and the residuals. The plot exhibits a linear pattern, and the best-fit line further supports this observation. The linearity in the scatter plot aligns with the assumption that the relationship between the independent and dependent variables is adequately captured by the linear model.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   The top 3 features contributing significantly towards explaining the demand of the shared bikes are temp, yr(2019) - positive contribution, and weathersit(Light_snow) - negative contribution.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear regression is a supervised machine learning method used to predict the dependent variable by learning patterns from independent variables. It predicts the relationship between the dependent and independent variables by establishing a linear correlation, illustrating how an independent variable is linearly associated with the dependent variable.
   Initially, we comprehend and clean the data, followed by preprocessing for model creation. Subsequently, we split the data into training and testing datasets. The training dataset is utilized to train the model, while the testing dataset is employed for model evaluation. After training, we assess the model's performance on the test dataset, focusing on error terms to validate the model. Ideally, the error terms should follow a normal distribution.
   In conclusion, the model, once validated, can be employed to derive important insights or make predictions.

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

   Anscombe's quartet is a collection of 4 dataset having nearly identical statistical properties(mean, variance, correlation, and linear regression) but differ significantly when we represent them using scatter plot.
   The different dataset will show different representations when we graphically represent them.

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

   Pearson's correlation coefficient is denoted by R, and it is a measure of the strength of the linear relationship between two different variables. It takes values between -1 and +1, where -1 represents a negative association between the variables, 0 represents no association between the variables, and +1 represents a positive association between variables.

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

   Scaling is a method to normalize the range of the independent variables. Scaling brings all the variables within the same range.
   The difference between the normalized scaling and standardized scaling is that the normalized scaling brings the values of the variables between 0 and 1 whereas standardized scaling transforms.

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

   If the value of the VIF is infinite, it indicates perfect multicollinearity. This occurs when one variable can be exactly predicted from another variable. It could happen due to many reasons, such as a perfect linear relationship, duplicate variables, or perhaps due to data issues.

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

   The Q-Q plot is a graphical tool used to identify whether two different sets of data belong to the same dataset. It is useful when checking if the training and test datasets belong to the same distribution. The Q-Q plot helps maintain the integrity of the data, ensuring the consistency of the datasets. This, in turn, contributes to the model's robustness and sanity.