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 categorical variables especially the **seasons** have a high correlation with the count of bikes rented, during fall and summer seasons there is high correlation. Then the **weather**, especially clear weather, had high correlation with the dependent variable.

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

It would eliminate a dummy column which can be inferred from the other dummy columns, for instance furnished and unfurnished can be inferred with unfurnished taking 0 or 1 as the value.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temperature (temp & atemp), Registered and casual had the most correlation with total count of rentals (cnt)

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

I performed manual feature selection with the help of P Value and VIF calculations, I was able to eliminate the features that caused multicollinearity and had low correlation, Then I performed RFE to get a more narrowed down set of features, just to be sure if my assumptions on feature selection were correct

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

a. **mnth**
b. **Workingday**
c. **windspeed**

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Linear regression is the technique through which a linear equation is formulated to predict a variable based on its correlation with other variables, this linear equation forms the model for the linear regression.

2. Explain the Anscombe's quartet in detail.
Graphs which are visually different but the mean and standard deviation but has totally different regression lines

3. What is Pearson's R?

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

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

This is due to the fact that multiple columns in the dataset has collinearity so they should be removed to eliminate the infinite VIF

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