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? (3 marks)

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

Answer: Using this we reduce the number of columns by one for a categorical variable with n levels. This helps in reducing redundancy and increase the speed of calculation.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Temp and atemp variable have highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: Temp, yr and workingday are the top 3 features contributing to the demand of shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

2. Explain the Anscombe's quartet in detail. (3 marks)

3. What is Pearson's R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

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

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

Q-Q plots also known as quantile plots which stand for quantile-quantile plots. As the name suggests Q-Q plot is obtained by plotting quantiles of two probability distributions against each other and can be used to compare the probability distributions. Generally, the Q-Q plots are used to check if the distributions are gaussian distributions or not by comparing the existing distribution with the gaussian distribution.