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)

Season, weathersit, season and month are good predictors of bike usage. There isn't any trend on weekdays except weekday vs. weekend which is partly captured in the holiday variable.

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

One of the fundamental assumptions of linear regression is the independent variables shouldn't be linear dependent on each other. Having drop_first = True breaks this dependency between the dummy variables.

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

Among the continuous variables temp has the highest correlation with the target variable

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

I plotted the residuals and ensured that the residuals are following a normal distribution with mean close to zero

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

Season, year and windspeed

General Subjective Questions

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

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Some of the assumptions of linear regression are:

- Normal distribution of error terms.
- Independence of error terms.
- Constant variance of error terms

The relation between the dependent and the independent variables is obtained such that the mean square error is minimum

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

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

It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations. It always lies between -1 and 1

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

Often independent variables are in different scale and the parameters are difficult to interpret when independent variables are of different scale. To bring them to the same scale we follow a process called scaling.

Normalized scaling is (X-X min)/(X max - X min)
Standardized scaling is (X-mu)/sigma

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

This means that the target variable is linearly dependent on some of the independent variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.