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)
   Ans: categorical variables such as season and its impact on the bike sharing is driven by parameters such as temperature and humidity. Hence from analysis it can be concluded that the categorical variables and dependent variables have a high correlation.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Ans: Not sure.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans: Temperature
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Ans: by calculating RMSE which is acceptable and in good range.

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

General Subjective Questions

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

Ans:

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables.

The case of one explanatory variable is called simple linear regression. For more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

The equation for linear regression is y = mx+c, where m is the slope of the line and c is intercept. Where y and X are dependent and independent variables.

Linear regression is a powerful statistical method to find the relationship between variables. But it's only limited to linear relationships.

Linear regression produces the high predictive accuracy for linear relationship whereas its little sensitive to outliers and only looks at the mean of the dependent variable.

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

   Ans: Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.

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

   Ans: Pearson's R is a measure of linear correlation between two sets of data.

   It is the covariance of two variables, divided by the product of their standard deviations.Thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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

   Ans: Scaling in statistics is a linear transformation of the form f(x)=ax+b.

   Normalizing means applying a transformation so that the transformed data is normally distributed. Standardizing means subtracting the mean and dividing by the standard deviation.

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

   Ans: Infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

   Ans: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. It is a graphical technique for determining if two data sets come from populations with a common distribution.

   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.