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?
   Answer: Month and week of the day shows some interesting insights form the categorical variables
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Answer: drop_first will delete the first created coulmn after the dummy variable creation and it is necessary to drop because the value which has all the 0's will be considered as the droped variable
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? **Answer:** Temp (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? **Answer:** Using Error Terms (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? **Answer:** WorkingDay, Windspeed, fall (2 marks)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (3 marks)

   Answer: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and aims to minimize the mean squared error between predicted and actual values. The coefficients are typically estimated using Ordinary Least Squares (OLS) or gradient descent. Key evaluation metrics include R-squared and Root Mean Squared Error (RMSE). Linear regression is widely applicable in various fields for tasks like forecasting and trend analysis.
2. Explain the Anscombe's quartet in detail. (3 marks)
   Answer: Anscombe's quartet consists of four datasets that have identical statistical properties—such as mean, variance, and correlation—but differ significantly in their distributions and visual patterns. Each dataset consists of 11 pairs of (x,y)(x, y)(x,y) values, yet they reveal different underlying relationships when graphed. This illustrates the importance of data visualization, as relying solely on summary statistics can be misleading. The quartet highlights that similar statistical measures can arise from vastly different datasets. It serves as a reminder that understanding data requires both numerical analysis and graphical representation.
3. What is Pearson's R? (3 marks)
   Answer: Pearson's R, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables, ranging from -1 to +1. A value of +1 indicates a perfect positive correlation, -1 a perfect negative correlation, and 0 signifies no linear correlation. It is calculated using the covariance of the variables divided by the product of their standard deviations. While useful for identifying linear relationships, Pearson's R is sensitive to outliers and does not capture non-linear relationships. Therefore, it's important to complement it with data visualization for a complete analysis.
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Answer: Scaling is the process of transforming data to a common range or distribution, essential for improving the performance of machine learning algorithms. It helps ensure that features contribute equally and enhances convergence in optimization. **Normalized scaling**

(min-max scaling) rescales data to a fixed range, typically [0, 1], while **standardized scaling** (z-score scaling) centers the data around a mean of 0 with a standard deviation of 1. Normalization maintains the original distribution shape, whereas standardization can alter it. Choosing the appropriate scaling method depends on the data characteristics and the specific analysis or model being used.

5. You might have observed that sometimes the value of VIF is infinite.
   Why does this happen? (3 marks)
   Answer: A Variance Inflation Factor (VIF) value of infinity indicates perfect multicollinearity among independent variables, where one variable is a perfect linear combination of others. This results in a singular matrix that cannot be inverted, leading to undefined VIF calculations. Redundant or identical variables can also contribute to this issue. Overly complex models may capture relationships that cause multicollinearity. An infinite VIF suggests a need to simplify the model by removing or combining variables for better stability.
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Answer: A **Q-Q plot** (quantile-quantile plot) compares the distribution of a dataset against a theoretical distribution, typically the normal distribution. In linear regression, it is used to assess whether the residuals are normally distributed, an important assumption of the model. If the points follow a straight line, it indicates normality; deviations may suggest issues like skewness or outliers. The plot helps in diagnosing model fit and guiding necessary transformations to meet assumptions. Overall, Q-Q plots provide a clear visual tool for evaluating the distribution of residuals and improving model accuracy.