Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical variables like season, holiday, and weather conditions significantly influence bike demand. For instance, demand is higher in summer and lower on holidays.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)
 It helps avoid the dummy variable trap, which causes multicollinearity by removing one
 category from each categorical variable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?   (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)
 Temperature (temp) has the highest correlation with bike demand, showing a strong
 positive relationship.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)
Checked for linearity using scatter plots, multicollinearity using VIF, normality using Q-Q plots, and homoscedasticity using residual plots.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)

Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression models the relationship between a dependent and independent variable(s) using the equation $Y = \beta\_0 + \beta\_1X + \epsilon$. It minimizes the sum of squared residuals to find the best-fitting line.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet consists of four datasets with identical statistical properties but different distributions, showing the importance of visualizing data rather than relying solely on summary statistics.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R measures the linear correlation between two variables, ranging from -1 (perfect negative) to 1 (perfect positive). A value near 0 indicates no correlation.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling adjusts feature values to a common scale to improve model performance. Normalization scales between 0 and 1, while standardization transforms data to have a mean of 0 and standard deviation of 1.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

  <Your answer for Question 10 goes here>
 Infinite VIF occurs when a predictor is a perfect linear combination of other variables, causing
 perfect multicollinearity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
 (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

  <Your answer for Question 11 goes here>
 A Q-Q plot compares the distribution of residuals to a normal distribution. It helps assess the
 normality assumption in linear regression.