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 and weather conditions significantly influence bike demand, with higher rentals during pleasant seasons like summer and fall, and clear weather. Severe weather and less favorable seasons, such as winter, reduce bike demand noticeably.

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)

Using drop_first=True prevents the dummy variable trap caused by multicollinearity among dummy variables. By dropping one category, the baseline, the remaining variables stay independent and avoid redundancy. This ensures the regression model's estimates are accurate and not distorted.

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)

The variable registered has the highest correlation with cnt, as registered users directly contribute to the total bike rentals.

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)

To validate the assumptions of Linear Regression, I checked scatterplots of residuals vs. predicted values.

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)

Based on the final model's coefficients, the top 3 features contributing significantly to bike demand are: 1. Year, 2. Season (2-4 once split), 3. Month. This is because of the growth from one year to the next coupled with significant seasonality in the bike sharing.

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 is a statistical method for modeling the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a linear equation to the data. It minimizes the sum of squared residuals to estimate coefficients (β) that best predict Y using the equation. Linear regression is widely used for predicting continuous outcomes, understanding relationships between variables, and assessing the impact of features on the target variable.

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)

Anscombe's quartet is a set of four datasets that have the same statistical properties with the same mean, variance, and correlation, but don't look the same when plotted. This shows that relying only on statistics can be misleading, as data with the same summary statistics can have completely different relationships. Visualization is crucial to spot patterns, outliers, and non-linear relationships. Anscombe's quartet reminds us to always plot our data during analysis to truly understand it.

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)

Pearson's R measures how strongly two variables are related and whether they move in the same direction. It ranges from -1 to 1 with 1 meaning a perfect positive relationship and -1 meaning a negative relationship, and 0 meaning no linear relationship. It's useful for understanding how two things are connected.

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)

Scaling ensures features contribute equally to model training. Normalization bounds data to a fixed range like [0, 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)

VIF can become infinite when there is perfect multicollinearity among the independent variables. This means that one variable is a perfect linear combination of one or more other variables. If two variables are perfectly correlated the model cannot determine their independent effects, leading to an infinite VIF.

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)

A Q-Q plot is a tool used to compare a dataset's distribution to a theoretical distribution like a normal distribution. A Q-Q plot can be used to evaluate the distribution of residuals to see if they are normal, which is expected in linear regression.
