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)

Potential inferences about how each might affect the demand for shared bikes :-
1. Summer and Fall might see higher demand due to favorable weather for biking.
2. Clear weather would likely encourage more biking, leading to higher demand.

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 when creating dummy variables is important for avoiding multicollinearity in your regression models. When you create dummy variables for a categorical feature with n unique categories, you'll end up with n columns (dummy variables) if you don't drop any.

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 highest correlation with the target variable is of “temp” variable.

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)

The relationship between each predictor and the target variable is linear.
The residuals (errors) should be independent of each other.
The variance of the errors should remain constant across all levels of the predicted values.
The residuals should be normally distributed, particularly for small samples, as this affects the validity of confidence intervals and p-values.
The independent variables should not be highly correlated with each other, as this can inflate the variance of coefficient estimates and lead to unstable estimates.

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)

The top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows :-

1. temp
2. yr
3. season_spring

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 foundational algorithm in machine learning and statistics, primarily used for predicting a continuous target variable based on one or more independent variables (features). The algorithm works by fitting a linear equation to the observed data and finding the best-fitting line that minimizes the difference between the predicted and actual values.

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 is a set of four small datasets that were created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. Each dataset in the quartet has identical summary statistics—mean, variance, correlation coefficient, and linear regression line—yet they are very different in their distribution and relationships when plotted.

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, also known as the Pearson correlation coefficient or simply correlation coefficient, is a measure of the linear relationship between two continuous variables. It was developed by Karl Pearson and is one of the most widely used correlation metrics in statistics and data analysis.

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>

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>

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, or Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. It is widely used in statistics and data analysis to visually check the normality assumption, which is crucial in many statistical methods, including linear regression.