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: From the analysis of categorical variables, it can be inferred that variables like `season`, `weathersit`, and `yr` significantly impact the demand for shared bikes. For instance, demand tends to be higher during summer and fall, as indicated by the `season` variable. Similarly, the `weathersit` variable shows that clear weather leads to higher rentals compared to misty or snowy conditions. The `yr` variable indicates growing popularity, as bike rentals increased in 2019 compared to 2018. (Do not edit)

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: Using `drop_first=True` during dummy variable creation avoids the dummy variable trap. It removes one dummy variable from each set of encoded variables to ensure no multicollinearity is introduced into the model. (Do not edit)

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: Based on the pair-plot, `temp` (temperature) has the highest correlation with the target variable `cnt`, indicating that higher temperatures are associated with increased bike rentals. (Do not edit)

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: After building the model, the assumptions of Linear Regression were validated by: 1) Checking residual normality using

histograms and Q-Q plots, 2) Ensuring homoscedasticity by plotting residuals versus predicted values, and 3) Evaluating multicollinearity using the Variance Inflation Factor (VIF). (Do not edit)

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: The top 3 features contributing significantly to the demand for shared bikes are: 1) Temperature (`temp`), 2) Feeling Temperature (`atemp`), and 3) Year (`yr`), as they show strong correlations and have high coefficients in the model. (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)

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, fitting a line (y = mx + c) to minimize the sum of squared residuals. The method uses least squares estimation to find the best fit and is widely used for prediction.

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 collection of four datasets that have nearly identical statistical properties (mean, variance, correlation, etc.) but differ significantly in their distributions and visual patterns. It emphasizes the importance of visualizing data to avoid

misleading conclusions based 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)

Pearson's R is a measure of the linear correlation between two variables, ranging from -1 to 1. A value closer to 1 indicates a strong positive linear relationship, while -1 indicates a strong negative linear relationship. It helps quantify the strength and direction of the relationship.

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 adjusts feature values to a specific range to ensure equal weighting in machine learning models. Normalization scales data between 0 and 1, while standardization transforms it to have a mean of 0 and a standard deviation of 1. Scaling is crucial for algorithms sensitive to feature magnitudes.

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)

The value of VIF can be infinite when there is perfect multicollinearity among the independent variables, meaning one variable is a perfect linear combination of others. This indicates redundancy and hampers model performance.

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 (Quantile-Quantile plot) is a graphical tool to compare the distribution of residuals with a normal distribution. In linear regression, it helps assess whether the residuals follow a normal distribution, an important assumption for model validity.
