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 significantly influenced bike demand. Demand was higher during summer and fall (season), on working days (workingday), and in 2019 (yr), reflecting seasonal trends and growing popularity. Weather conditions (weathersit) also played a role, with clear weather seeing the highest demand and reduced rentals in rain.

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 avoids multicollinearity by dropping one dummy variable from each categorical feature, as the dropped category can be inferred from the others. This ensures the regression model can function without redundant information.

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)

From the pair-plot analysis, temp (temperature) has the highest positive correlation with the target variable cnt (bike demand). Warmer temperatures are strongly associated with increased 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)

Residual Analysis: Examined the residuals' distribution for normality and checked for homoscedasticity.
Multicollinearity and Linearity

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)

yr (Year): Higher demand in 2019 compared to 2018.
workingday: Increased demand on working days.
windspeed: Negative impact, with lower demand on days with high wind speed.

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 method to predict a continuous value by finding a straight line that best fits the data points. It shows how one or more factors (independent variables) affect the outcome (dependent variable). The algorithm adjusts the line to minimize the difference between the predicted and actual values, making it useful for understanding trends and making forecasts.

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 collection of four datasets that demonstrate the importance of visualizing data when analyzing statistical properties. All four datasets have nearly identical summary statistics (mean, variance, correlation, regression line), yet their distributions and patterns differ significantly.

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, or the **Pearson Correlation Coefficient** is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It is denoted by r and ranges from -1 to +1.

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>

Scalling is a method to bring all the numeric columns of different scale to a same scale. This prevents larger values from dominating smaller ones in models like k-NN or SVM.

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>

VIF becomes infinite when there is perfect multicollinearity, meaning one feature is an exact linear combination of other features. This happens when redundant or highly correlated features exist in the dataset. Infinite VIF indicates the model cannot uniquely estimate coefficients. Removing one of the correlated features resolves this issue.

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>

Q-Q plot compares the distribution of residuals (errors) from a model to a theoretical normal distribution. In linear regression it checks whether the residuals follow a normal distribution. If the points on the Q-Q plot lie along a straight line, the residuals are normally distributed. This is important for validating model assumptions and ensuring accurate predictions and inferences.
