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)

Based on the analysis, categorical variables like season, weathersit, and holiday have a noticeable impact on the demand for shared bikes. For instance, demand tends to be higher during certain seasons (such as summer and fall) compared to others, likely due to favorable weather. Similarly, clear weather conditions (weathersit) are associated with higher demand, while misty or rainy conditions see reduced demand. Additionally, on holidays, demand might vary as fewer people commute to work, impacting rentals.

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` helps prevent multicollinearity by removing one level of each categorical variable, as it can be inferred from the remaining categories. This makes the model more stable and prevents redundancy in the feature set, ensuring that dummy variables do not introduce linear dependencies.

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 `temp` (temperature) has the highest positive correlation with the target variable `cnt` (total rental count), suggesting that warmer temperatures are associated with increased demand for 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)

After building the model, I validated linear regression assumptions by:

- Checking residuals for normality using a Q-Q plot.
- Analyzing residual plots for homoscedasticity to ensure constant variance.
- Using the Variance Inflation Factor (VIF) to check for multicollinearity.
- Reviewing the distribution of residuals to confirm they are approximately normal.

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 to bike demand are `temp` (temperature),

`season_Fall` (indicating fall season), and `yr` (with increased demand in 2019 compared to 2018).

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 supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation. The algorithm aims to minimize the sum of squared residuals, which represent the difference between observed and predicted values. The line of best fit is determined by finding the optimal values of coefficients using methods like Ordinary Least Squares (OLS), which minimizes the mean squared error. This model assumes linearity, independence, homoscedasticity, and normality of residuals.

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 that have nearly identical statistical properties, such as mean, variance, and correlation, but appear very different when graphed. This illustrates the importance of visualizing data before analysis, as summary statistics alone can be misleading. Anscombe's quartet highlights the risk of relying solely on numerical summaries and encourages the use of graphical representations to detect patterns, outliers, or relationships.

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 Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's R is useful for quantifying the strength and direction of a linear relationship between two variables.

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 the range of features for improved model performance. It is particularly important for algorithms sensitive to the scale of data. Normalization (min-max scaling) scales data within a specified range, usually 0 to 1, whereas standardization scales data to have a mean of 0 and standard deviation of 1. Standardization preserves outliers, while normalization confines data to a defined range.

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>

An infinite VIF indicates perfect multicollinearity, meaning one variable is a perfect linear combination of others. This typically happens if there are duplicate variables or if dummy variable encoding hasn't excluded a reference category, creating redundancy in the predictors.

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 (quantile-quantile plot) compares the distribution of residuals to a normal distribution. Points close to the line indicate normally distributed residuals, which is an assumption of linear regression. Deviations suggest non-normality, indicating potential issues in the model, such as skewness or outliers, that may affect validity.
