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 Weather_3 being 3: light rain/snow have a high negative Coeff impact on dependent variable

Categorical Variables like holiday Whether the day is a holiday have a high negative Coeff impact on the dependent variable.

Categorical Variables like weekday_3 have a high positive Coeff impact on the dependent variable.

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 in dummy creation helps to avoid the dummy variable trap, a condition of perfect multicollinearity where one category can be linearly predicted by others. This improves model stability and interpretation.

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 numerical variable with the highest positive correlation with the target variable. In bike-sharing data, “atemp” and “temp” from both will be using the same as it’s a derivative of one another have a strong positive correlation with demand.

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)

Ensure the relationship between predictors and target is linear.

Check if residuals follow a normal distribution.

Residuals have constant variance across levels of the predictor variables.

Evaluate VIF values to ensure no high multicollinearity among predictors.

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)

Temperature (atemp), year (yr), and casual users (casual).

The temperature plays an vital role in the demand of the bikes, the year as the demand of the bike will increase, and the causal users

---

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 finds a line that best fits the data by minimizing the sum of squared differences between predicted and actual values.
The equation is Y = beta0 + beta1X1 + beta2X2 + ... + betanXn + ε, where: beta0 is the intercept, beta1, beta2,... betaN are coefficients, and ε represents error terms.
The model is solved using the least squares method, which minimizes the residual sum of squares (RSS) to optimize parameter 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 comprises four datasets with nearly identical statistical properties (mean, variance, correlation), yet each has distinct visual characteristics when plotted. It demonstrates the importance of visual data analysis beyond summary statistics, as similar metrics can represent vastly different data distributions.

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 correlation coefficient, measures the linear relationship between two variables. R ranges from -1 to +1:
R = +1 indicates a perfect positive correlation,
R = -1 a perfect negative correlation,
and R = 0 no linear 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)

<Your answer for Question 9 goes here>
Adjusts the data range of features to improve model performance and convergence. It is particularly helpful for algorithms sensitive to feature magnitudes:

Standardization centers data around mean 0 and scales to unit variance.
Normalization scales data to a range, typically [0, 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)

<Your answer for Question 10 goes here>

A VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity—when one predictor variable can be exactly predicted by a combination of others. This signals the need to remove or combine multicollinear features.

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 (Quantile-Quantile plot) compares residuals to a normal distribution to assess if residuals are normally distributed, which is a key assumption in linear regression. Deviations from the diagonal line suggest deviations from normality, impacting inference reliability.
