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)

The analysis shows that bike rentals are likely higher in summer and fall, especially in September and October. Rentals are more frequent on Saturdays, Wednesdays, and Thursdays, with higher demand in 2019. Additionally, bike rentals tend to increase on holidays.

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)

drop_first=True removes one dummy variable to avoid extra columns and prevent redundancy in the dataset.

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)

temp and atemp are highly correlated.

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)

Validated linear regression assumptions by checking VIF (multicollinearity), residual error distribution, and linear relationship between the target and feature variables.

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)

temp, yr and season

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 used to predict continuous outcomes based on one or more independent variables. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables.

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 data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.

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 Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +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>

Scaling is a pre-processing technique used in machine learning to standardize independent features within a fixed range. Datasets often have features with different magnitudes and units, which can affect the model's performance. Scaling helps avoid this issue by bringing all features to a similar scale.

1. Normalization scales data between 0 and 1.
2. Standardization transforms data into Z-scores with a mean of 0 and standard deviation of 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>

When two independent variables are perfectly correlated, the R-squared value becomes 1, making the VIF (Variance Inflation Factor) infinite using the formula:

VIF = 1 / (1 – R^2)

This indicates multicollinearity, meaning the variables carry duplicate information. To fix this, one of the correlated variables should be removed to build a better regression model.

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 quantiles of a sample dataset with a theoretical distribution (like normal, uniform, or exponential).

It helps check:

- If the dataset follows a **specific distribution**.
- If two datasets have the same type of distribution.
- Whether the errors in the dataset are normally distributed.
