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)

'dteday' is the only categorical variable and it is was dropped because it was redundant. Therefore there is no effect of categorical variables on the dataset

---

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)

If there are 'n' levels for a categorical variable, 'n-1' levels can be enough to infer information. Hence drop_first=True is important

---

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 'registered' has the highest correlation with the target variable

---

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)

Checking for normal distribution of the error terms/residuals, and dropping of highly correlated variables to show they are independent of each other

---

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)

Constant, Casual and Yr

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>

The linear regression algorithm is the total sum of the product of variables and their weights/coefficients, with the aim of regression being to obtain the least sum of squared errors

---

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 share identical summary statistics, yet look very different when graphed. It serves as a reminder of the importance of visualizing data, rather than depending 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)

<Your answer for Question 8 goes here>

Pearson's R coefficient gives a measure of the linear relationship between two variables, with negative correlation indicating inverse relationship and positive correlation indicating directly proportional 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>

Scaling is done to bring values of all variables within the same range, thereby making it easy for the model to learn the patterns. Standardised Scaling converts the features with mean centered at 0 and unit standard deviation. Normalised scaling converts the values in the range of 0 to 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>

VIF value being infinity indicates that there is perfect correlation between the predictor variables. This may happen when two variables identical in nature, which can make it redundant.

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, or quantile-quantile plot, is a graphical tool used to compare the quantiles of two distributions. In this plot, the quantiles from one distribution are plotted against the quantiles from another. When both sets of quantiles come from the same distribution, the points on the plot will align roughly along a straight line