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)

1) Summer and fall might have higher bike rentals than spring and winter.
2) Clear weather conditions will likely lead to higher rentals than rainy or misty conditions.
3) Weekends or holidays might have different rental patterns than weekdays.
4) Working days might have higher rentals compared to non-working days (or vice versa).

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)
To remove the main column to avoid redundancy.

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 column

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)
Checked the R2_score. Also, plotted a scatter plot with line with y_test and y_pred.

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, windspeed, workingday

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>

In simple terms linear regression assumes that the relationship between the dependent variable (and the independent variable is linear and therefore can be represented by a straight line. The best fit line passes through the most number of actual points but not all of them. That is why there are errors. However, the error is minimized by ordinary least square method. There are few assumptions based on which the linear regression model operates:

1) The observations should be independent of each other.
2) No multi-colinearity among the attributes.
3) The error variance should always be constant.
4) The errors follow a normal distribution.

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 set of four different datasets that have nearly identical summary statistics (mean, variance, correlation, regression line) but exhibit completely different distributions when visualized. This highlights the importance of data visualization. That is why we always cannot depend upon the statistics summary. We always need to check the visualization for pattern or trend recognition.

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>

Karl Pearson's Coefficient is a statistical measure that determines the strength of relationship in between a direct and indirect variable along with its direction ; either positive or negative.

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>

When two or more columns in a dataset have records with far different ranges its difficult for the machine to compare them correctly. This is because the machine might give a weighted priority to the bigger numbers while analyzing and during training. That is why all the columns need to be brought to similar ranges like suppose in between 0 and 1 for example. This will help reducing overestimating by the machine. In case of normalized scaling which is also called min-max scaling the value ranges between (0,1) or (-1 , 1) whereas standardized scaling works in such a way that all the records would have 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>

VIF would become infinite if R^2 is 1. This can happen if one attribute is the linear combination of other features. Practically due to redundant data in the dataset may lead to this situation.

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>

Used to check for normal distribution of data. Help in determining one-tailed tests. Helps identify potential outliers.
