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 were significantly affecting the bike usage demand. Variables like year and workingday were directly proportional to the demand.

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)
It uses all values of the categorical column. It just adds another feature which can be reduced by adding drop_first=True. Less features will always help to reduce the complexity of the model. For example, instead of evaluating effects of all four seasons, we can have 3 season as features of the data.

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 seems to has the highest correlation.

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)
We performed residual analysis. I found that residuals were normally distributed. Also, they were centered around 0 value. R2_score was around 0.80 that explains around 80% variance in demand.

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)

Year, windspeed and weather situation (Light snow or rain)

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>

It is a supervised learning algorithm. Target variable will be available upfront. It is used in predicting target variable. Using linear regression algorithm, we try to find the effect of one dependent variable on the target while keeping other constants.

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 helps to emphasize the importance of visualization. It comprises 4 datasets that have nearly identical descriptive statistics. On plotting those datasets, they represented totally different probability 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 R is a measure of correlation. It measures linear correlation. Its value can be between -1 and 1. It measure the strength and direction of 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 is a preprocessing step in regression analysis that is performed on independent variables. Scaling helps to prevent one feature from dominating other features. Normalized scaling brings all data between 0 and 1. Standardized scaling replaces the values by their z score.

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>

If VIF is infinite, it means that variables are perfectly correlated. It can happen due to multicollinearity and high correlation between dependent variables.

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 is used to determine if dataset follows a particular distribution. It is also known as Quantile-Quantile plot. It is used to check assumptions and identify departures from expected distributions.