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)

Season: Impact is significant. Higher ridership is observed in fall than spring.
Weathersit: Impact is significant. People prefer cloudy skies to Snowy weather
Year: Higher ridership was observed in 2019 compared to 2018
Holiday: Has a negative impact on ridership
WorkingDay: Doesn't show much of an impact

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 avoid multicollinearity in linear regression.

When all dummy variables for a categorical feature are included in the regression model, it creates a situation of perfect multicollinearity, where one variable can be perfectly predicted from the others

---

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 & atemp (0.63)

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)

Normality of errors using Q-Q plot
No Multicollinearity by model summary checking VIF

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)

Regression Equation:

y = 0.26 + 0.23 * yr + -0.08 * holiday + 0.62 * temp + -0.26 * hum + -0.21 * windspeed + 0.07 * season_2 + 0.14 * season_4 + -0.19 * weathersit_3

3 significant features are: temp, humidity 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>

Linear regression predicts a dependent variable (y) by using one ore more independent variables (X) by fitting a straight line (or plane in multiple dimensions). The algorithm works by minimizing the sum of squared errors (difference between actual and predicted y) using techniques like OLS (Ordinary Least Squares). It assumes:

a) A linear relation between X and y
b) Errors are normally distributed, independent and have a constant variance.

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 datasets that have nearly identical simple descriptive statistics (such as mean, variance, correlation) but differ significantly in their distributions and relationships.

It is used to demonstrate that Summary Statistics like mean, variance, correlation can be misleading without visualizing the data.

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, also known as the Pearson Correlation Coefficient, is a statistical measure that describes the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 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)

Scaling is the process of adjusting the range of feature values in data to a common scale. This is carried out to prevent variables with larger ranges from dominating the model as well as improve model performance (Gradient Descent time).

Normalization adjusts data to a specific range (typically between 0 and 1), while standardization adjusts to have a mean of 0 and unit variance.

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>

It happens when there's perfect multicollinearity between two or more independent variables. This means that one variable can be perfectly predicted from the others.

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)

Q-Q plot is a graphical tool to assess if a dataset follows a normal distribution. If the data points form a straight line, it suggests that the data follows a normal distribution.

Q-Q plot helps visually to check if the residuals are normally distributed.
