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)

Model was done to analyse how Bike Rent is influenced by host of categorical variables, among that significantly

Season is likely a significant influencer, with rentals peaking in fall and summer and dropping in winter.

A positive coefficient for yr is indicating an increasing trend in bike rentals over time that is increase in usage from 2018 to 2019.

Month is likely a significant predictor, with a cyclical pattern tied to weather and seasons.
Weather is likely a strong influencer, with rentals decreasing as weather conditions worsen.

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: Using drop_first=True during dummy variable creation, it is used to avoid multicollinearity, reduce redundancy and reduce the number of variables

---

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 (no. of registered users has the highest positive correlation.
The variable temp also shows a significant positive correlation.
Humidity has a negative correlation, indicating that higher humidity is associated with fewer bike rentals.

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)

Relationship between the independent variables and the dependent variable should be linear. Used Scatter plots to verify. The residuals (errors) should be independent. Independent variables should not be too highly correlated with each other. Variance Inflation Factor >10 confirmed the same

---

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)

The top 3 features contributing significantly towards explaining the demand of the shared bikes are following

- Temp
- Atemp
- 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)
Linear regression model is used here to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.
The core idea is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of the squared differences (residuals) between the observed values and the values predicted by the model.

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)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics yet exhibit very different distributions and relationships when graphed.
Four Data Sets

Dataset I: This dataset shows a strong linear relationship between the independent variable and the dependent variable
Dataset II: This dataset also has a linear relationship, but it is influenced by an outlier.
Dataset III: In this dataset, there is no linear relationship; instead, it follows a quadratic pattern.
Dataset IV: Similar to Dataset II, this dataset has an outlier, but in this case, it is positioned in such a way that it creates a non-linear relationship.
Anscombe's Quartet reminds us in statistics that understanding data requires more than just calculations; it necessitates careful exploration through visualization to uncover true relationships and avoid misleading conclusions.

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)

Pearson's R, a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted as r and ranges from -1 to +1.

Value Interpretation:

r=+1: Perfect positive linear correlation. As one variable increases, the other variable also increases.

r=−1: Perfect negative linear correlation. As one variable increases, the other variable decreases.

r=0: No linear correlation. Changes in one variable do not predict changes in the other.

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 ensures that each feature contributes equally to the calculations, preventing features with larger ranges from dominating the model.

Why is Scaling Performed?

Improves Model Performance, Prevents Bias, Enhances Interpretability and Facilitates Comparisons between different features and their effects on the target variable.

There are two common types of scaling: Normalization(Min-Max) and Standardization.

Min-Max scaler brings all variables values range from -1 till 1

Standardization for adopting maximum values and so it accounts outliers as well

---

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)

Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF value greater than 10 is often considered indicative of high multicollinearity, while a VIF value of infinity suggests an extreme case.

Reasons for Infinite VIF

Perfect Multicollinearity:

Infinite VIF occurs when one independent variable is a perfect linear combination of one or more other independent variables. For example, if you have two variables where one variable can be expressed as a multiple of another (e.g., X2 = 2 * X1), this leads to perfect multicollinearity.

Redundant Variables:

Including redundant variables in the model that do not provide additional information can also cause infinite VIF. For instance, if you include both temperature in Celsius and temperature in Fahrenheit in the same model, they are perfectly correlated.

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 plots are tools used in linear regression analysis for validating the normality assumption of residuals i.e., fitting commonly the normal distribution and diagnosing potential issues with model fit, and guiding necessary transformations to improve model performance. They provide a clear visual representation that complements numerical statistics, enhancing overall interpretability and reliability in statistical modeling.
