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)

Based on the analysis of categorical variables from the dataset, it can be concluded that bike rental rates are generally higher during the summer and fall seasons, especially in September and October. Moreover, rentals are more common on Saturdays, Wednesdays, and Thursdays, as well as in the year 2019. Additionally, there is a noticeable increase in bike rentals on holidays

---

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 reduce the extra column created during the dummy variable creation

---

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 variable 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)

Validated the assumptions of linear regression by checking VIF, error distribution of residuals and linear relationship between the dependent variable and feature variable.

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)

temperature, year, and holiday variables are the top three features that significantly contribute to the demand for shared bikes

---

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 is a machine learning algorithm used for supervised learning. It predicts a dependent variable (target) based on given independent variable(s). This regression technique aims to establish a linear relationship between the dependent variable and the independent variables. There are two types of linear regression: simple linear regression and multiple linear regression.

Simple Linear Regression: This is used when a single independent variable is employed to predict the value of the target variable.
Multiple Linear Regression: This is used when multiple independent variables are utilized to predict the numerical value of the target variable.
The linear line that represents the relationship between the dependent and independent variables is known as the regression line.

Positive Linear Relationship: This occurs when the value of the dependent variable on the Y-axis increases as the value of the independent variable on the X-axis increases.
Negative Linear Relationship: This occurs when the value of the dependent variable decreases as the value of the independent variable on the X-axis increases.

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 have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression lines, but exhibit very different distributions and appear distinctly different when visualized graphically. Each dataset consists of eleven points.

The primary purpose of Anscombe’s quartet is to highlight the importance of graphically examining data before performing statistical analyses. It demonstrates that relying solely on summary statistics can be misleading, as different datasets can share the same statistical properties but have very different underlying structures and patterns.

Here are the key points about Anscombe's quartet:

Identical Descriptive Statistics: Despite having nearly identical statistical measures (mean, variance, correlation, etc.), the datasets are fundamentally different.
Graphical Representation: Visualizing the data reveals the differences in distribution and patterns that are not apparent from the statistics alone.

Illustrative Purpose: It serves as a powerful reminder to analysts and statisticians to always visualize their data to avoid misinterpretation and to gain a better understanding of the data's characteristics

---

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 is widely used in various fields such as finance, economics, psychology, and the social sciences to measure and interpret the strength and direction of linear relationships between variables. It is an essential tool for identifying correlations and making predictions based on the linear association between variables. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +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)

<Your answer for Question 9 goes here>

Scaling is a pre-processing technique used in building machine learning models to standardize the independent feature variables within a fixed range.

Datasets often contain features with varying magnitudes and units. Without scaling, this disparity can lead to incorrect modeling due to mismatched units among the features involved in the model.

The key difference between normalization and standardization is that normalization adjusts all data points to fall within a range between 0 and 1, while standardization transforms the values into their Z scores.

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>

When there is a perfect correlation between two independent variables, the Variance Inflation Factor (VIF) becomes infinite. In this scenario, the R-squared value is 1, leading to an infinite VIF since VIF is calculated as $1/(1-R^2)$. This indicates a problem of multicollinearity, suggesting that one of these variables should be removed to create a functional regression model

---

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>

Quantile-Quantile (Q-Q) plots are used to compare the quantiles of a sample distribution with those of a theoretical distribution to assess whether the sample follows a specific distribution, such as normal, uniform, or exponential. Q-Q plots help determine if two datasets follow the same type of distribution and can also be used to check if the errors in a dataset are normally distributed.