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)

My analysis of the categorical variables reveals that bike rental rates tend to peak during the summer and fall seasons, with especially high demand in September and October. Rentals are also notably higher on Saturdays, Wednesdays, and Thursdays, and they saw an raise in 2019. Additionally, holidays show a clear increase in bike rentals.

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)

Attribute drop_first=True removes the extra column created during dummy variable creation, helping to avoid redundancy and simplifying 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)

The registered, 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 the p-value, VIF, error distribution of residuals and linear relationship between the dependent variable and a 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)

The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, the year and the holiday variables.

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 is a supervised learning algorithm used to predict a target variable based on one or more input variables. It works by finding a linear relationship between the target variable and the input variables. There are two types of linear regression:

1. Simple Linear Regression: Uses one independent variable to predict the target variable.
2. Multiple Linear Regression: Uses multiple independent variables to predict the target variable.

The line that shows this relationship is called the regression line. If the target variable (on the Y-axis) increases as the independent variable (on the X-axis) increases, it's a positive relationship. If the target variable decreases as the independent variable increases, it's a negative relationship.

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 includes four datasets that have nearly identical basic statistics but look very different when graphed. Each dataset has eleven points, and they highlight how important it is to visualize data before analyzing it, as statistics alone may not capture the true differences between datasets.

Example: There is a clothing store in 3 major cities, Hyderabad, Bangalore, Chennai. On excel the sales are approximately same numbers but, mapping the numbers against marketing the sales show linear, exponential and no relationship in each city.

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 Correlation Coefficient measures the strength of the linear relationship between two variables, with values ranging from -1 to +1. A value close to +1 indicates a strong positive relationship, while a value close to -1 indicates a strong negative relationship.

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 a preprocessing step in machine learning used to standardize feature variables to a fixed range. Datasets often have features with different magnitudes and units, which can cause issues if not scaled, leading to inaccurate modeling.

The main difference between normalization and standardization is that normalization adjusts values to a range between 0 and 1, while standardization transforms values to their Z-scores,

reflecting how far each value is from the mean in terms of standard deviations.

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)

When two independent variables are perfectly correlated, the Variance Inflation Factor (VIF) becomes infinite because the R-squared value is 1, making VIF = 1/(1 – R2) undefined. This indicates multicollinearity, meaning one of these variables should be removed to build an effective 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 compare the quantiles of a sample distribution to a theoretical distribution, like normal, uniform, or exponential. This helps us see if the dataset follows a specific distribution and whether two datasets share a similar distribution. Q-Q plots are also useful for checking if the errors in a dataset are normally distributed.
