Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable? (3 marks)

The analysis of categorical variables indicates that bike rental rates tend to be higher during the summer and fall seasons, particularly in the months of September and October. Moreover, there is a notable increase in bike rentals on Saturdays, Wednesdays, and Thursdays, as well as in the year 2019. Additionally, the data suggests that bike rentals experience a surge on holidays.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Using **drop_first=True** in dummy variable creation is crucial to avoid the dummy variable trap. It prevents multicollinearity issues by excluding one of the dummy variables, preventing perfect predictability.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

In the pair-plot, we observed that temp Variable has the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

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

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

The key contributors significantly influencing the demand for shared bikes are the temperature, the year, and the holiday variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is like fitting a straight line to a set of points on a graph. Imagine you have data points that form a pattern, and you want to draw a line that best represents that pattern. Linear regression helps you find that line. The line helps you make predictions about one thing based on another. For example, if you know the temperature, you might predict how many bikes people will rent.

The algorithm looks for the best-fitting line by figuring out how much each input (like temperature or time) contributes to the output (like bike rentals). It aims to minimize the difference between its predictions and the actual outcomes. So, in simple terms, linear regression helps us understand and predict relationships between different things.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R? (3 marks)

Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling? (3 marks)

Scaling is a preprocessing step in machine learning that involves transforming the numeric features of a dataset to a common scale. It is performed to ensure that no single feature dominates the others, especially when working with algorithms sensitive to the scale of the input data, such as gradient-based methods.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

The value of VIF is infinite when there is a perfect correlation between the two independent variables. The Rsquared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)
Q-Q plots, short for quantile-quantile plots, serve the purpose of comparing the quantiles of a sample distribution against the quantiles of a theoretical distribution. These plots are useful for assessing whether a given dataset adheres to a specific distribution like normal, uniform, or exponential. By visually comparing the points on the Q-Q plot to a straight line (representing a theoretical distribution), one can identify the distributional characteristics of the dataset. Additionally, Q-Q plots aid in recognizing if the errors in the dataset exhibit a normal distribution or deviate from it.