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 Categorical variable like WeatherSit, Season, Month from the data set will have impact on the Dependent variable,

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

The drop first is important to ensure the categorical column is dropped , as the get dummies function will create. New column based on the categorical values

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Temp has the highest correlations

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

looking at the coef values we can define the positive or negative impact on the final result

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1.aTemp

2.Season_Winter

3.Month_September

General Subjective Questions

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

The linear regression algorithm, is a mathematical equation that describes the relationship between two variables over a period of time and the change that can occur. With linear regression there is a mathematical model that can be used $Y = C + b0x + b1X + b2X$ the relationship

between variables used as inputs and the computed output. Generally, linear regressions emphasis the upward .the best-fit line is obtained by minimising a quantity called Residual Sum of Squares (RSS), this is the best time to be introduced to what is known as the cost function.

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

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

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

and standardized scaling? (3 marks)

1. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   The VIF value can be infinite as data sets with no relationships are analyzed together to depict a regression but end up being correlated to one another. When a large correlation is determined between the unrelated data sets, the VIF value is seen as infinite.
2. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   A Quantile-Quantile, or Q-Q plot, is a plot that graphically depicts the relationship between collected data and sample data from various distribution types. For example, in a Q-Q plot, the collected data and sample data from a prior distribution, which could be binomial, are broken down into equivalent groups and mapped into the Q-Q plot. The intersection of each data point where the sections have been created is plotted on the graph as data points. A linear line is then drawn onto the plot to show if the data points best fit this type of distribution. If the data points do not fall on a linear line, another sample distribution with a different data set should be selected to identify the type of distribution that best fits the collected data. A Q-Q plot is necessary for linear regression because the relationship between the values from the collected data set and the sample distribution set is used to determine if this is the proper distribution for the collected data. A linear regression shows the relationship between an x and a y value in a consistent rate and straight line. With this determination of data points aligned on a straight line, the Q-Q plot can be used to see if there is a consistent pattern among the distribution with additional data if the collected data set grows. Additionally, a straight line graphically is more straightforward in identifying if data points fit on the line for the selected distribution.