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?

Ans s: From the analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season, are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019. Additionally we could discern that bike rental are higher on holidays.

2: Why is it important to use **drop_first=True** during dummy variable creation?

Ans: drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: The 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?

Ans: 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?

Ans : Ans: 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

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables. There are two types of linear regression- simple linear regression and multiple linear regression. Simple linear regression is used when a single independent variable is used to predict the value of the target variable. Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

3. What is Pearson's R?

Ans: 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?
   Ans: Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model. The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.