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 - Upon analyzing the categorical variables of the data, it was found that bicycle rental charges experience a predictable surge during summer and autumn, particularly in the months of September and October. Moreover, Saturdays, Wednesdays, and Thursdays are days when rentals are frequently observed, and there was a marked rise in rental activity during the year 2019. Additionally, it was revealed that the cost of renting bicycles increases during holiday periods.

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Ans - Drop first=True reduces the additional column generated by the creation of the dummy variable, leading to the elimination of any duplicity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans -The temperature variable exhibits the strongest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans - The assumptions of linear regression were confirmed by evaluating the VIF, the distribution of residual errors, and the correlation between the dependent and feature variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans - The demand for shared bikes is greatly influenced by the three most important factors, namely temperature, year and holiday factors.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans - Linear regression is an ML algorithm that is utilized for supervised learning. Predicting a target variable is facilitated by utilizing the given independent variable(s). The technique of regression often links a dependent variable and the remaining independent variables through linear means. The two categories of linear regression are simple linear regression and multiple linear regression. Simple linear regression is utilized when there is only one independent variable being employed to estimate the value of the dependent variable. The method of using numerous independent variables to predict the numerical value of the objective variable is called multiple linear regression.

2. Explain the Anscombe's quartet in detail.
   Ans - Anscombe's quartet comprises four sets of data that share the same basic descriptive statistics. The distributions are significantly diverse, and their graphical representation would demonstrate noticeable dissimilarities. Each dataset comprises eleven items. points. Anscombe's quartet is designed to highlight the importance of thoroughly scrutinizing a set of

data as its primary objective. Prior to commencing the analysis, it should be noted that relying solely on statistics may not prove to be a precise means of comparing two datasets.

3. What is Pearson's R?
   Ans - The Pearson correlation coefficient is utilized to determine a straight correlation between two variables. The coefficient, which can range anywhere from -1 to +1, offers insight into the level of correlation between two variables in terms of strength.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans - Scaling is a pre-processing method utilized in constructing machine learning models in order to standardize the independent feature variables within the dataset to a consistent range.
   The dataset may contain numerous characteristics that vary greatly in both magnitude and units. Incorrect modeling can occur if the data is not scaled because the units of the features in the model will not match.
   Normalization and standardization differ in their techniques of data transformation. Normalization scales all data points to a range of 0 to 1, while standardization replaces the values with their respective Z scores.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans - Variance inflation factor (VIF) attains an infinite value when the independent variables are absolutely correlated.
   The value of Required is 1 in this case. As VIF equals 1 divided by 1 minus R2, the outcome is an infinite VIF.
   Normalization and standardization differ in their techniques of data transformation. Normalization scales all data points to a range of 0 to 1, while standardization replaces the values with their respective Z scores.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans - The (Q-Q) plot is a visual tool that compares quantiles of a dataset to a theoretical distribution, such as normal, uniform, or exponential, in order to determine if the dataset follows that distribution. We can use it to ascertain if the distribution of two sets of data is identical. Assessing if the errors in the dataset are customary is also beneficial.