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 - The examination of the data's categorical variables reveals that bike rental fees are expectedly higher during summer and fall, with a greater emphasis on September and October. Additionally, rentals are particularly common on Saturdays, Wednesdays, and Thursdays, and notably increased during the year 2019. Moreover, it became apparent that bicycle rentals are more expensive during holidays.

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Ans – By reducing the extra column added when the dummy variable was created, drop first=True helps to eliminate all redundancy.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans – By examining the VIF, the residual error distribution, and the linear relationship between the dependent variable and a feature variable, it was possible to validate the assumptions of linear regression.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans - The top 3 factors, including temperature, year, and holiday factors, have a significant impact on demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans - An ML algorithm used for supervised learning is linear regression. Based on the provided independent variable(s), it aids in forecasting a dependent variable (target). A dependent variable and the other independent variables are typically connected linearly by the regression technique. Simple linear regression and multiple linear regression are the two types of linear regression. When a single independent variable is used to predict the value of the target variable, simple linear regression is used. When several independent variables are used to forecast the numerical value of the target variable, this is known as multiple linear regression.

2. Explain the Anscombe’s quartet in detail.
   Ans - Four data sets with essentially identical simple descriptive statistics make
   up Anscombe's quartet.
   very different distributions, and when represented graphically, they look very different. Eleven are contained in each dataset.
   points. The main goal of Anscombe's quartet is to emphasize the value of carefully examining a collection of data.
   before starting the analysis process, as statistics merely do not provide an accurate picture comparison of two datasets as represented.

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.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans - When there is a perfect correlation between the two independent variables, VIF has an infinite value.
   In this instance, the Rsquared value is 1. Given that VIF is equal to 1/(1-R2), this results in VIF infinity.
   According to this idea, multi-collinearity is a problem, and one of these variables must be eliminated in order to define a useful regression model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   Ans - To determine whether a dataset in question follows a particular distribution, such as a normal, uniform, or exponential distribution, the quantile-quantile (Q-Q) plot is used to plot quantiles of a sample distribution with a theoretical distribution. It enables us to determine whether the distribution of two datasets is the same. It is also useful to determine whether or not the errors in the dataset are typical.