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 bike rental rates are probably going to be higher in the summer and the fall, are more noticeable in the months of September and October, are more so in the days of Saturday, Wednesday, and Thursday, and are more so in the year 2019. These conclusions may be drawn from the analysis of the categorical variables from the dataset. In addition, we discovered that vacations are the best times to rent bikes.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
-drop_first=True the extra column added during the formation of the dummy variable, True helps to avoid all repetition.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
-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? (2 marks)
-The top 3 features, including temperature, year, and holiday factors, have a major impact on demand for shared bikes.

General Subjective Questions

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

- An ML approach used for supervised learning is linear regression. Based on the provided independent variable, it aids in forecasting a dependent variable (goal).(s). A dependent variable and the other independent variables are typically connected linearly by the regression approach. Simple linear regression and multiple linear regression are the two forms 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 factors are used to forecast the numerical value of the target variable, this is known as multiple linear regression. Regression lines are linear graphs that depict the connection between dependent and independent variables. When both the dependent and independent variables are on the X-axis, there is a positive linear connection. It is a negative linear relationship, though, if the value of the dependent variable falls as the value of the independent variable rises on the X-axis.

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

-Four data sets make up Anscombe's quartet, which have essentially similar simple descriptive statistics but radically diverse distributions and visual appearances. There are eleven points in every dataset. The main goal of Anscombe's quartet is to emphasise the significance of visualising a group of data before starting an analysis process because statistics simply cannot accurately describe two datasets being compared.

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

-Correlation by Pearson To establish a linear relationship between two quantities, coefficient is used. The value of the coefficient, which can range from -1 to +1, indicates the strength between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

-Scaling is a pre-processing method used to standardise the independent feature variables in the dataset within a predetermined range.
The dataset may contain a number of features that range widely in high magnitudes and units. There will be some discrepancy in the units of all the characteristics included in the model if scaling is not done on this data, which results in erroneous modelling.
Normalization and standardisation are different in that standardisation replaces the values with their Z scores, whereas normalisation sets all the data points in a range between 0 and 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

-When the two independent variables are perfectly correlated, VIF has an infinite value. In this instance, the R-squared value is 1. Given that VIF equals 1/, this results in VIF infinity.(1-R2). According to this idea, multi-collinearity is an issue, and one of these variables must be eliminated in order to create a useful regression model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

-To assess whether a dataset in question follows a certain 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.