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?

The analysis of the categorical variables shows some interesting results. In terms of seasons, bike rentals are shown to be higher in the summer season and in the fall season. In terms of days of the week, Saturday, Wednesday and Thursdays show the highest rental rates. In terms of months, the highest bike rentals are shown in September and October.

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

When the dummy variables are added to the dataset extra columns are added. In order to make the analysis more efficient drop_first=True is applied to remove the unnecessary extra columns.

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

Temperature variable.

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

The distribution of the error was checked (normal distribution), the VIF values were checked and a linear relationship was observed.

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

The top 3 features are Season, Windspeed and Weather Variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

In machine learning supervised learning is performed with the application of a Linear Regression algorithm. By predicting the value of a target variable with independent variables the value of the ML model is apparent. We need a relationship between inputs and outputs. Linear regression helps identify that relationship. We do this on one variable under Simple Linear Regression, whereas we add multiple variables in the Multiple Linear Regression model. The result will either show a positive relationship or a negative relationship.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four data sets with almost identical simple descriptive statistics but with different distributions and look different when displayed on a graph. Anscombe's quartet is a means of visualizing the dataset before proceeding with the analysis

3. What is Pearson's R?

Pearson's Correlation Coefficient measures the linear relationship between two sets of data. The ratio of the covariance of two variables and the product of their standard deviations is Pearson's R. 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?

Standardization is where the mean is subtracted from the value and is then divided by the standard deviation. Where as Normalization is the process of dividing a vector by its length to transform the data range into a value between 0 and 1.

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

VIF is infinite when the correlation between two variables is perfect. This suggests multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Q-Q plot, or the quantile-quantile plots are used to determine if two data sets show similar distributions and to determine if the error distribution is normal.