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?

Observation:

Fall season has high rentals from season plot

More Bikes rented in 2019 from year box plot

Rental high on normal working day from workday plot

Saturday has highest from weekday plot

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

Reduces correlation among dummy variables.

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

atemp

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

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

Atemp

temp

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear Regression is an machine learning algorithm used based learning by supervision. It performs the task of predicting a dependent variable which is called as target based on some given independent variables. The regression technique finds the linear relationship between the 2 variables.

2. Explain the Anscombe's quartet in detail.

   It consists of four data sets that have nearly identical simple descriptive statistics yet have very different distributions, appear different when graphed

3. What is Pearson's R?

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It’s a step of data pre processing which is applied to independent variables to normalize the data within a particular range. Its speeds up calculations in algorithm.

Collected data sets contain features vaying high in magnitude, units and range. To solve this we bring all values to the same scale.

Standardized scaling brings all data into a standard normal distribution which has a mean zero and standard deviation. Sklearn.preprocessing.scale helps to implement standardization in python.

Normalized scaling brings all of the data in the range of 0 and 1.

Sklearn.preprocessing.MinMaxScaler helps in implement normalization in python.

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

IF there is perfect correlation then it is infinity. In this case R2 is 1. In such cases we have to drop one of the variables from the dataset which is causing this.

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

Q-Q plots are plots of 2 quantiles against each other. A quantile is a fraction where some values fall below the quantile. The purpose is to find out if 2 data set of data is emerging from same distribution.

It provides a graphical view of how properties such as location scale are similar or different in 2 distributions.