Assignment

Subjective Questions

1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A

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

A Since it helps in reducing the extra column created during dummy variable creation.

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

A Temperature is highly correlated with count of bike

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

A After calculating predicted values those are compared with test sets. This is how one can validate LR.

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

A 1. Temperature 2. Weather situation 3.Year

General Subjective Questions

1 Explain the linear regression algorithm in detail

A 1.Data Pre-processing 2. Splitting Data into training and test data set 3. Fitting the Simple Linear Regression to the Training Set 4. Prediction of test set result 5. visualizing the Training set results 6. visualizing the Test set results:

2 Explain the Anscombe's quartet in detail.

A Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots

3 What is Pearson's R?

A Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

A Scaling is converting original data set into o to 1 form or some readable numbers which are not much deviated from other variables.
Scaling is performed to achieve correct results during LR analysis.
Normalized scaling will scale data into 0 to 1 range
Standardized scaling where values are centered about mean

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

A An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

6 What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution