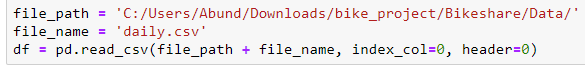
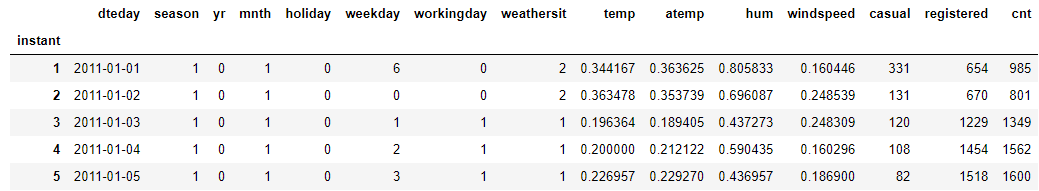
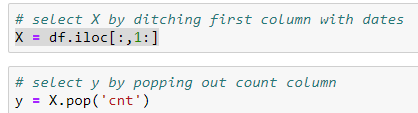
Decided to build my Linear Regression model based on the daily dataset instead of the hourly data. After reviewing the data it appears that the first column is an index column, first row are the headers. Imported my data as follows using the Pandas read\_csv() function defining the previous assertions.



From there I reviewed to see what the data looked like using the Pandas head() function in order to view the first 5 rows.



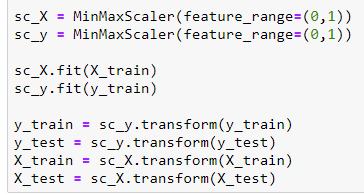
From this I determined that there were 15 columns in the data, including a date column at index 0. Since dates can’t be used in Multiple Linear Regression models I decided to drop this data from the selection, leaving 13 columns of predictors and the last column as my value to be predicted. But because the columns such as humidity or windspeed are very small values by comparison to the registered column my next step was to normalize the data.



Before normalizing, I split the data into 80% training data and 20% test data. I did this using the train\_test\_split() function from sklearn to randomly split into X and y, test and train variables. Because a linear regression model does not take into account a time series’ positioning I randomly did so using 123 as my seed to be able to recreate my results.



In order to normalize the data you take each column, find the min value, max value, and for each point subtract min and divide by max-min. Or, in order to make my life easier, I could use the MaxMinScaler function from the sklearn library. Used the training X’s and training y’s to scale the test data to a number between 0 and 1



Because all columns are numerical I decided to test them all out in an exhaustive feature search. In order to select the best features to use, instead of running individual analysis on 13 columns vs the y variable, I used the ExhaustiveFeatureSelector function from the mlxtend.feature\_selection library. This does an exhaustive search between all variables to find the variables with the highest correlation. The function returned the following output, which aligns with columns for workingday, casual, and registered as the highest predictors of count.

