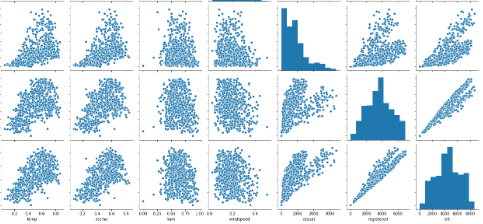
**Summary Report**

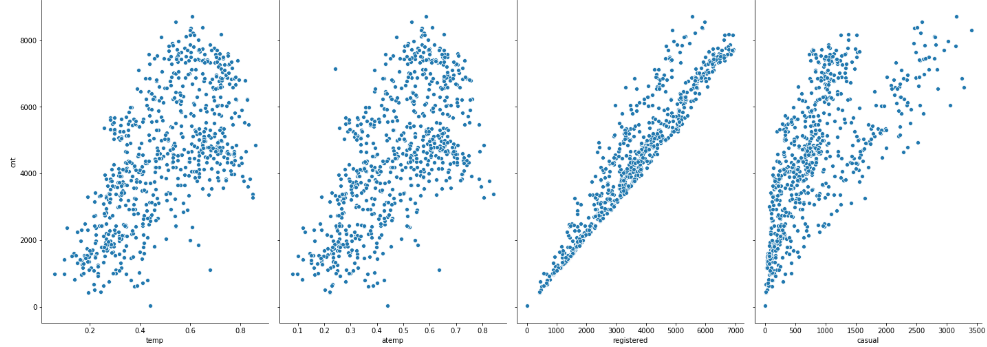
**Multi Linear Regression:**

Before we started with any model, we did explore the data to have a general idea on how the data looks like. We have visualized the data and I have notices there some sort of linear relationship between dependent and independent variables. I preferred to start with multi linear regression. The following charts shows the relationship that exists between variables.



**Splitting the data:**

We have used the sklearn packages to split the data into 2 haves for testing and training. Before splitting the data, we decided to drop some other variables that have weak relationship with the dependent variable. At the first stage of training, we have included only four variables as predictor and this can be seen clearly from the following graph.



**Fit and Prediction:**

After we fit the model, we tested the model to see the level of the accuracy. For the evaluation we have relied on 2 things: the mean square error and the R square. The performance on the

Prediction was very good based on the R square and the mean square error.

**Random Forest:**

I haven’t done any project with random forest before and this was a good chance for me to test this type of machine learning and see its performance.

**Data Preparation:**

As usual, we have started with the data preparation first. The exact steps for preparation of the data will depend on the model used and the data gathered, but some amount of data manipulation will be required for any machine learning application.

**Response Encoding:**

Instead of using one-hot-encoding for the dummy variables, I have decided to use the response encoding. I have reconstructed the dataset for the final stage and ready for the model fit.

Train Model:

After all the work of data preparation, creating and training the model is pretty simple [using Scikit-learn](http://scikit-learn.org/stable/tutorial/basic/tutorial.html). We import the random forest regression model from skicit-learn, instantiate the model, and fit (scikit-learn’s name for training) the model on the training data.

**Prediction:**

Getting accuracy of 83% This tells us that we actually do not need all the data we collected to make accurate predictions! If we were to continue using this model, we could only collect the two variables and achieve nearly the same performance.

**Conclusion:**

In comparison between the 2 models, the linear regression model has performed well and the accuracy was higher than the random forest.