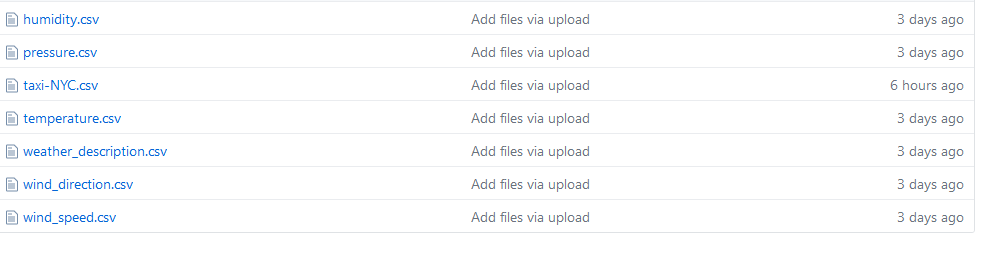
**Techincal Blog Assignment**

Goal: Finding a model to predict the number of the taxi rides given time date and temperature.

Tools:

* Python
* Random Forest of predictive regression modeling.
* Multiple relatinal data bases.

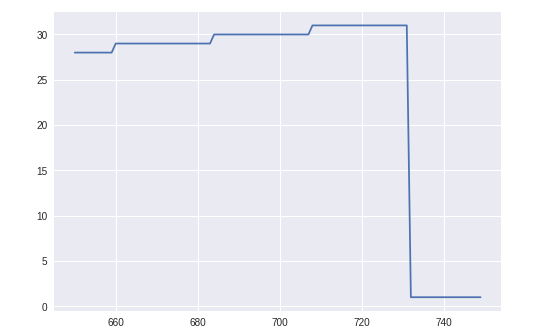
The datasets had information from different date and time.



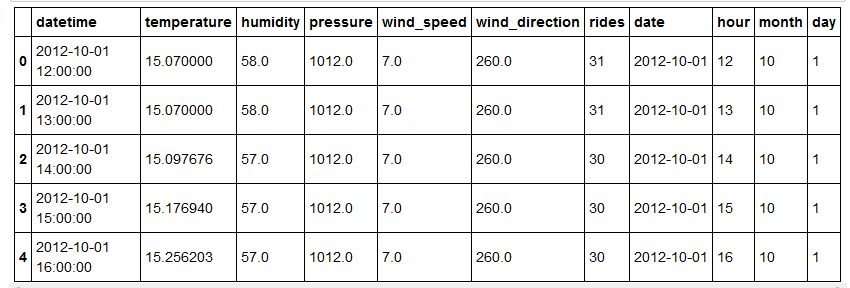
After cleaning, converting and aggregating data, I achieved to my processed database.

So I chose data related to 2015 to build my RandomForest model.

Here I faced the first problem. To use machine learning I need to have cyclical data.For example Time. I chose one day to have a look at the time around midnight.



when the time goes from 23 to 0 we have a jump, so there is a problem about presenting syclical data to machine learning algorithem.

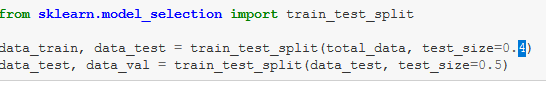


between record 34 and 35 : 23-22=1 between record 35 and 36 : 0-23=-23 so we need to change the encoding of the feature. A common method for encoding cyclical data is to transform the data into two dimensions using a sine and consine transformation.

we need to see a cycle betwwn 0 and 23.

From this set, I used 80% of the data for training and 20% for

test.



The analysis above shows that our model has more than 99% accuracy. At first, it seems fantastic! We made the best model! But there’s a problem…