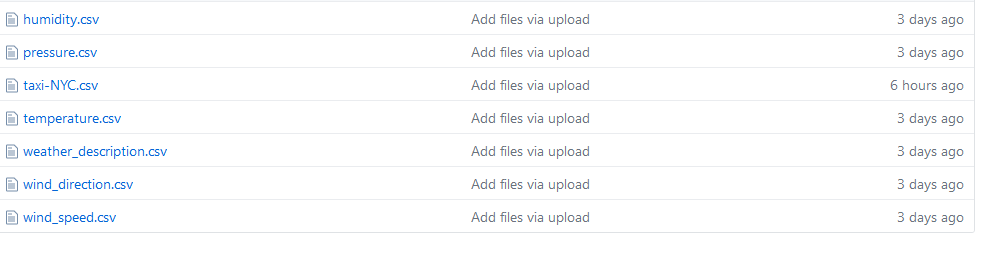
**Techincal Blog Assignment**

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.

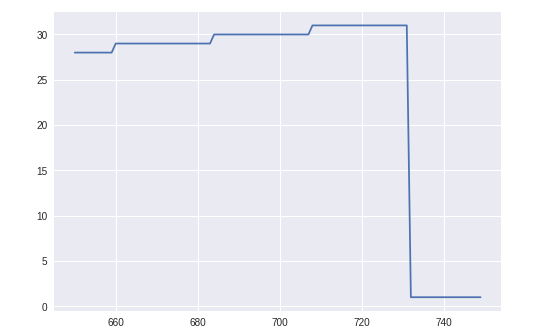
The datasets had information from different date and time.



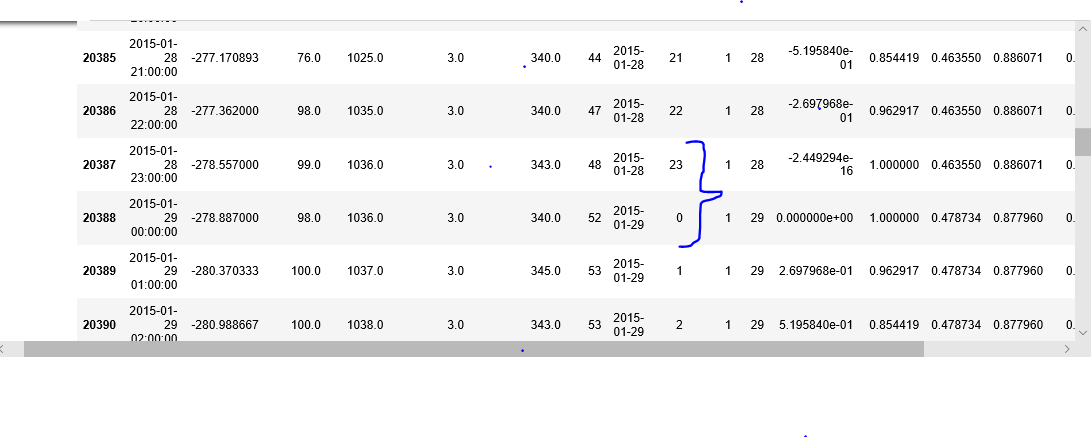
After cleaning, converting and aggregating data, I achieved to my processed database.

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 "20387" and 35 : 23-22=1

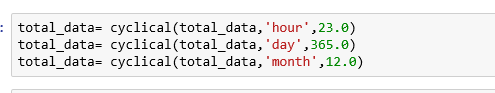
between record "20388" and 36 : 0-23=-23

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 between 0 and 23. We can do that using the following transformations:

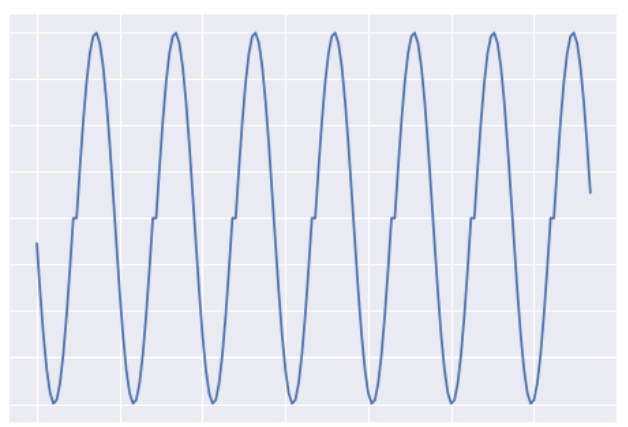
*xsin*=sin(2∗*π*∗*x*max(*x*))

*xcos*=cos(2∗*π*∗*x*max(*x*))

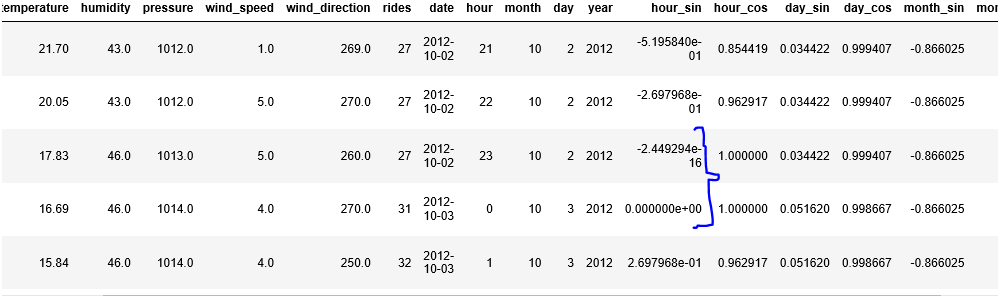
Let's do this for our hourly data by defining a function :



Why two dimensions, using sine and cosine you may ask? Let's have a look at just one dimension:

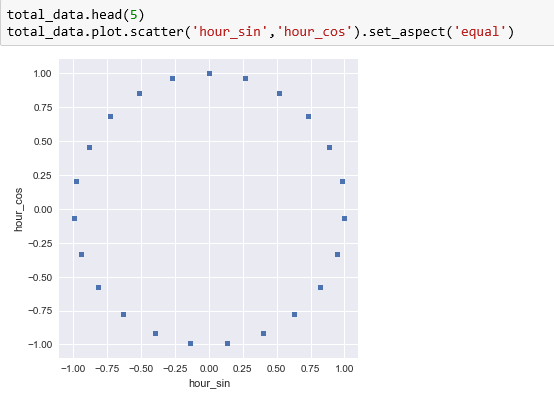


As expected, it is cyclical, based on the sine graph. Looking at the values around midnight again:



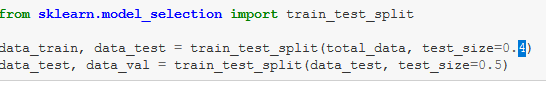
Great, it appears the absolute difference an hour\_sin before, at and after midnight is now the same! However, if we look at the plot of hour\_sin (following any flat line intersection with the graph), we can see there is a problem. If we consider just the one dimension, there are two records with exactly the same hour\_sin values, This is why we also need the cosine transformation, to separate these records from each other.

Indeed, if we plot both features together in two dimensions we get the following:

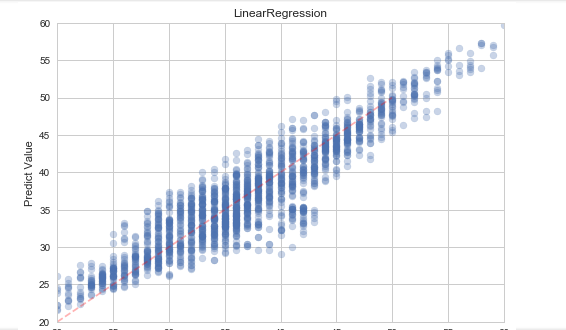


Now I can fit them in to the model. It suppose to have better results with encoded data so I will try my model with both.

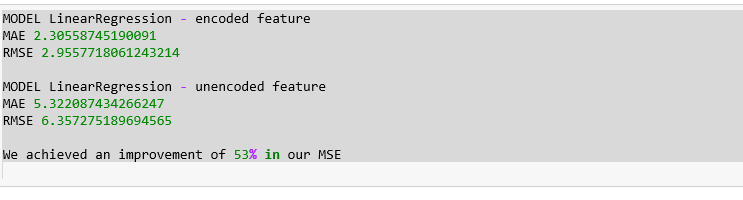
First of all this data set has been split to train and test.



I found the best result by linear regression as the best model.



After evaluation of each model I got the result below:



**Conclusion :**

It's important to encode features correctly for the specific machine learning algorithm being used. Other machine learning algorithms might be robust towards raw cyclical features, particularly tree-based approaches. However, deep neural networks stand to benefit from the encoding strategy discussed above, particularly in terms of the convergence speed of the network.