# DATA ANALYSIS & FINDINGS

## Data Screening

Data screening is necessary in ensuring that data of the variables are correctly entered, free from large missing values, outliers and to confirm that the distribution of the data for scale variables is normal. The variables in this project which were examined through data screening are: humidity, temperature, wind\_speed, wind\_direction and rides.

### Replacing Missing Values

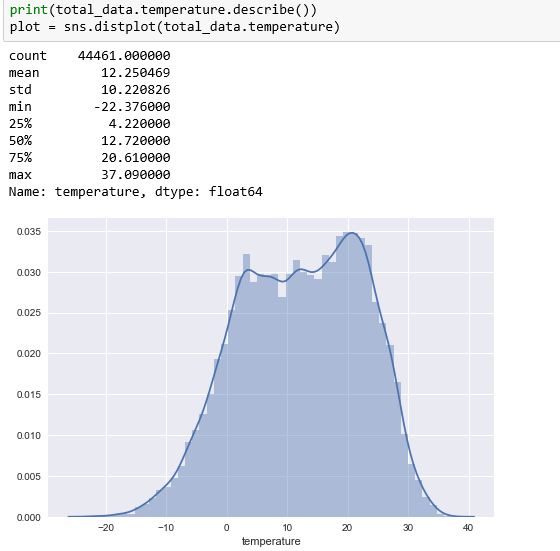
Missing data occurs when there is no information for one or more cases in relation to a variable. If that missing data up to 5% may not cause any serious problem in the interpretation of the findings. In this project as I have hourly data, I used previous data to fill the missing value.

### Assessment of the Data Normality

Normality is the main assumption of parametric tests such as Linear Regression.

Using powerful liberary in python I plot the skewness and kurtosis, and virtually observe the shape of the distribution. For this reason, in the present project, skewness and kurtosis were employed to assess normality of the data. Skewness values reflect the symmetry of the distribution score and a skewed variable means that the score is not at the centre of the distribution. On the other hand, kurtosis gives information about the "peakness" of the distribution which can be either too peaked

or too flat.



### K-Means Clustering

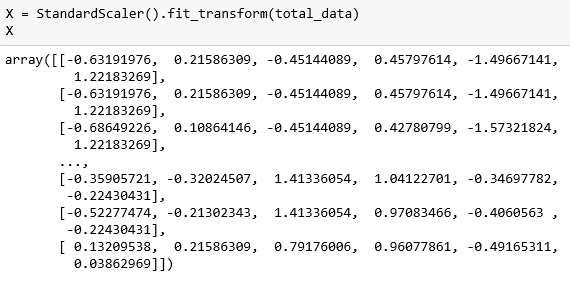
Clustering is a powerful tool which has been used in several forecasting works, such as time series forecasting, real time storm detection, flood forecasting and so on. In this project, K-means clustering algorithm is proposed. To find the result I use typical K-means clustering on the processed database that I created it in the past part. In this notebook, I perform k-means clustering in Python. I will use cluster analysis to generate a big picture model of the weather at NYC regarding the taxi. In this dataset, I used 12 clusters to present my dataset.

At first I select the feature of interest of clustering.

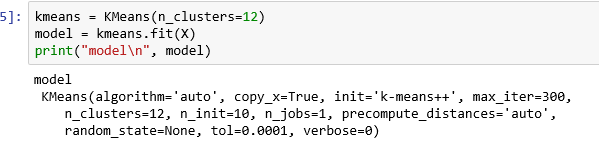


**Feature scaling**

To standardize the range of features of data I used StandardScaler from Sklearn.

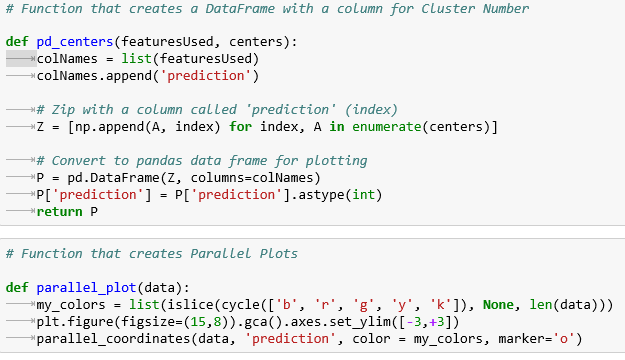


Here our data is ready to use in clustering models. I chose K-means model. Using Sklearn I did it simply. we can do it simply using the predefined function.



I achieved to 12 clusters determining the center of the cluster.

Finally, I create some utility functions which will be helpful in plotting graphs:



The visualized maps are available to study. For example we can plot cold and windy days if we chose low temperature and high wind\_speed .



**Conlusion :**

There is a great bound of weather analysis use cases that I just worked on a simple question of one of them. I learned that I can create a prediction from unsupervised data using clustering. When I don’t have any target as my test data I can use clustering to define the rules and prediction model. Here in cluster number 4 (blue line) I can see that I have low temperature with high taxi rides in condition that humidity and wind approximately have the same level. In cluster number I see that by increasing the temperature and the wind speed the number of the taxi rides goes down. We can study the similarity between the entity inside the clusters and the difference between 2 independent clusters and find the secret of the data.