**ERICH LEONARDO RATZAT and**

**NIKOLAS LIPPMANN PARESCHI**

**ALPHA DESIGN WITH MACHINE LEARNING**

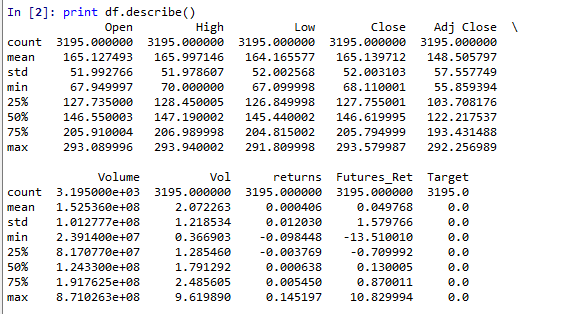
**MINI PROJECT I**

For this project, we followed the instructions below:

1. Identify a data set that you can use for the project (in Unit 5 and 6).

We decided to use the SP500 data from YahooFinance! From 2006 until 2018:





1. Use a subset of the data and load it into the WebSim environment.

We asked permission to use python instead of WebSim and it was accepted.

1. Find an off-the-shelf clustering algorithm.

Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

A clustering problem is where you want to discover the inherent groupings in the data, such as grouping financial market data into more volatile or less volatile days.

We decided to use the k-means algorithm and the goal is to put each data point into a group or cluster and do an analysis from there. It starts with the center data and iterates until minimize the distance between data points and the nearest centroid (the center of the group data).

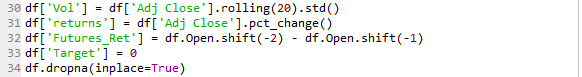
1. Use the WebSim python API to call the clustering algorithm.

As we said before we worked with python.

1. Discuss how this may be applied to your alpha design and in preparation for your project.

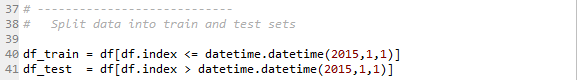
The idea is to work with SP500 data and try to divide it into high or low volatility groups. The idea is try to predict trading activity using this volatility group example.

The first step was to calculate the volume and the trading range for daily activity. We used a 20 day moving average for this:

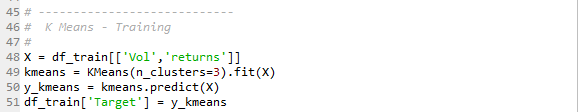


After that, we divided the data into training and testing groups. The goal here is to find “groups” of different volatility as low, medium and high. This in theory would give us some insights to build a strategy with an alpha in the stock market.

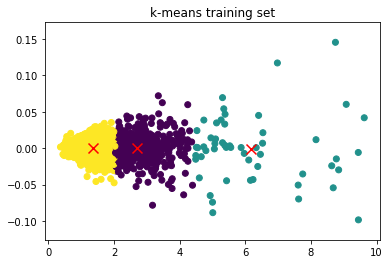
The k-means algo give us clusters for each data point.



The k-means training set is as follow:

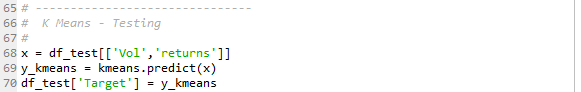


When we plotted the data we got this:

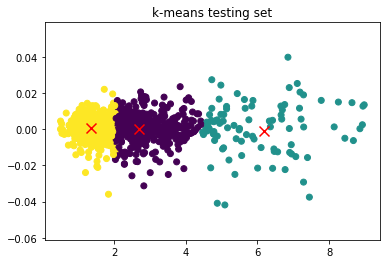


We can see 3 different “clusters”. The first two are more homogeneous. The x-axis is our volume variable and the y-axis is our return variable.

The next step is to test the algorithm again in a out of sample data, which is our testing data.

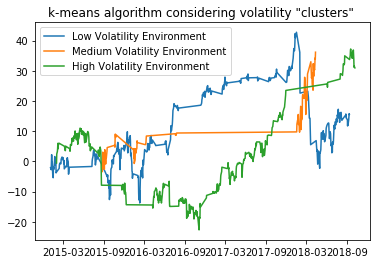


And the result is as follow:



Visually speaking our testing data seems to be less dispersed (cluster 3, the green one) when compared with the training data.

The final goal is to check if this strategy produced any good results for trading the Sp500. We would buy a contract based on today cluster. So, if today volatility was medium we would buy a contract. We can see the strategy performance based on the low, medium and high volatility as showed on the plot below:



The low volatility performed better until 2018 and then fell. As we thought in the beginning, the medium volatility strategy was smoother than the others and produced better results in the end.

Sources:

<http://www.pythonforfinance.net/2018/02/08/stock-clusters-using-k-means-algorithm-in-python/>

<https://www.quantstart.com/articles/k-means-clustering-of-daily-ohlc-bar-data>

<https://www.quantnews.com/k-means-clustering-creating-simple-trading-rule-smoother-returns/>