**STOCK MARKET PREDICTION**

* **Group Member**: Rishabh Vijay Agarwal, Wesley Stevens, Haoyang Shang, Xin Bai
* **Objective**

We will research how machine learning models and preprocessing methods perform at predicting stock prices. We will evaluate using a running error metric over a period of one year using the years previous for training. We would be working on the 30 stocks which are a part of the Dow Jones Industrial Average and identify the top performing stocks from it.

* **Data Set Description**

Overview / Description:

The dataset contains the stock market data for the 30 companies which are a part of Dow Jones Industrial Average (^DJI) ETF. The dataset contains the data of each of the ticker symbol from 8th February 2013 to 7th February 2018. The data set contains 7 columns which are Date, Open, High, Low, Close, Volume and Ticker.

The DJIA is a benchmark that tracks American stocks that are the leaders of the economy and are on the Nasdaq and NYSE. The DJIA covers 30 large-cap companies, which are subjectively picked by the editors of The Wall Street Journal.

The Dow Jones Industrial Average (DJIA) is a stock index of 30 blue-chip industrial and financial companies in the U.S. The index is used in the media as a barometer of the broader stock market and the economy as a whole.

Number of rows and cols:

There are 36620 rows in the dataset.

The data set contains 7 columns which are

1. Date: It ranges between 8th February 2013 to 7th February 2018. It had to be converted from string to datetime format.

2. Open: Open price of the stock on a particular day. (Price of stock at the beginning of the stock market at 9:30 am)

3. High: High price of the stock on a particular day.

4. Low: Low price of the stock on a particular day.

5. Close: Close price of the stock on a particular day. (Price of stock at market close 4:30 pm)

6. Volume: Volume of shares traded in that day.

7. Ticker: Symbol for the stock as traded on the New York Stock Exchange.

Sample predictors (does not need to be an exhaustive list):

For the Autoregressive Integrated Moving Average model, the predictors would be the lags of the Close price and its lagged forecast errors.

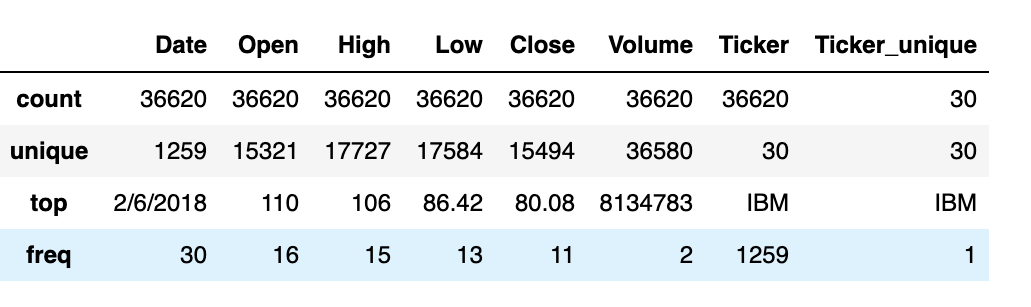
For the Random Forest Model, we would be using the Open, High, Low, Close and Volume attributes.

A link to the dataset:

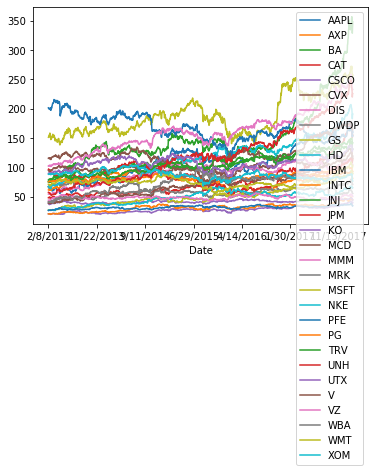
<https://drive.google.com/file/d/12AlXVGyI289zs7k-1Vuy6HEecTesFYSa/view?usp=sharing>

* **Preliminary Data Exploration**

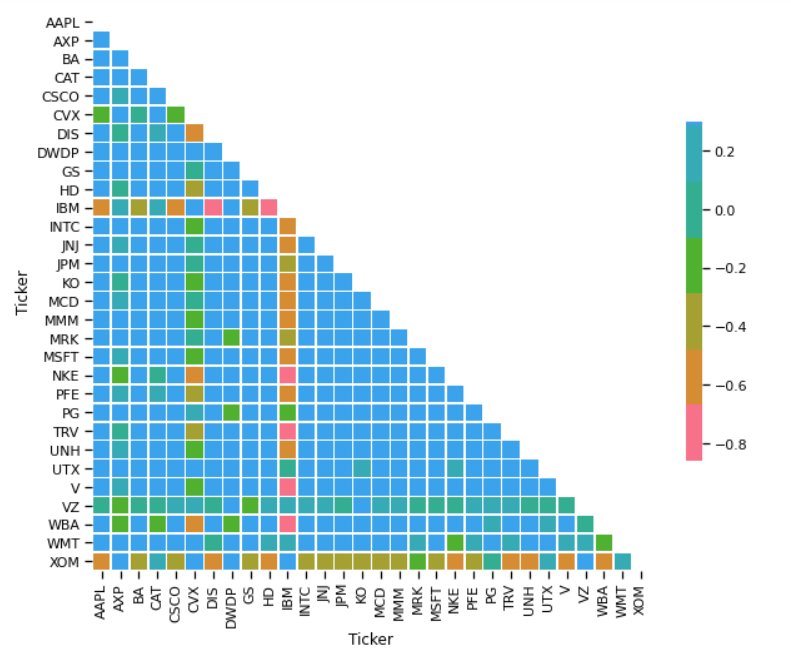
Summary statistics:

**

The following graph is a time series graph through 5 years from 2013-2018. It includes 30 stock names and pictures daily activity changes throughout the time period. It is interesting to see that BA has been the fastest growing stock based on increasing close price and it has reached the highest among others. Compared with BA, GS is also on the top but has smaller variance with a slower pace of growing. This graph provides information in order to predict the future behavior of those stocks and yield larger profit.

**

The following graph shows how volumes of different stocks change over time throughout the 5 years period from 2013-2018. It depicts volume change activities and each color represents one stock name. It is interesting to see that there are a couple of abnormal behaviors of volume for some specific stocks at certain time periods. Those facts allow us to further investigate why there were sudden increases in volume for certain stocks during those specific time periods.

**

This is a correlation plot of the close prices of stocks with one and another. The pair of stocks with the highest correlation are used in pairs trading. The stocks belonging to the same sector exhibit high correlation as change in their prices are affected by the same conditions.

* **Prediction**

We will try to predict the future stock price for each of the stocks which are a part of the Dow Jones Industrial Average(^DJI) ETF.

The initial plan is to implement a RF (Random Forest) Model and a ARIMA (Auto Regressive Integrated Moving Average) Model.

Compared to decision tree which sometimes could be overfitting due to the high variance, the RF could average multiple decision tree to reduce the variance to present a higher accuracy.

The idea of ARIMA is to learn patterns of the dataset that changed over time from the historical data, and to use these patterns to predict the future which is exactly what we are trying to do with the stock market dataset. ARIMA model is particularly used in pairs trading.

* **Inference**

We want to determine the minimum number of stock price points needed to give a good estimate. This will give us a lower bound that we can then use for estimation. We will also investigate how more datapoints prior to the desired prediction affects the outcome. For example, after N number of datapoints, do we see diminishing returns?

We want to know if smoothing methods give a more general prediction of how the stock will perform. If so, then we can more easily tell how well a stock is or will perform. For the smoothing methods, we will experiment with a rolling average window and an exponential rolling average window. We believe these to be suited to this dataset due to the volatile nature of stock prices. We can do this with a package called Flint.

* **Non Spark Packages**

1. Flint: enhance Spark’s functionality for time series analysis. Flint is an open source library and available via Maven and PyPI.