FE-582 Foundations of Financial Data Science

Fall 2019
Final Project Report

Stock analysis for 30 Dow Jones Stocks

Junzhe Wang
Ashish Negi
Rohnit Shetty

Introduction

The enormous availability of data in today's time enables investors, at any scale to make better investment decisions. Keeping such a theory in mind, our goal is to use historical price of 30 Dow Jones stocks to analyze the stock market and predict the performance of stocks.

Research Question

The main question we wish to find an answer is, if we can use the content of news analytics to predict stock price performance?

Dataset Description

We chose the dataset which provides free end of day data for all stocks currently in the Dow Jones Industrial Average, from 2015-01-02 to 2019-06-11. There are 30 csv files in the current version of the dataset. For each of the 30 components of the index, there is one CSV file named by the stock's symbol (e.g. AAPL for Apple). Each file provides historically adjusted market-wide data (daily, max. 5 years back).
List of stocks and symbols as per:
https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average

symbol date	AAPL	AXP	BA	CAT	CSCO	CVX	DIS	DWDP	GS	HD	...	
2015-01-02	101.1385	86.6383	115.7103	79.7573	24.0850	93.6102	88.4442	NaN	182.4928	93.5363	...	2
2015-01-05	98.2893	84.3470	114.9089	75.5472	23.6052	89.8685	87.1517	NaN	176.7949	91.5739	...	2
2015-01-06	98.2985	82.5495	113.5555	75.0611	23.5965	89.8269	86.6895	NaN	173.2185	91.2935	...	2
2015-01-07	99.6769	84.3524	115.3185	76.2243	23.8145	89.7521	87.5763	NaN	175.7999	94.4225	...	2
2015-01-08	103.5067	85.5481	117.3576	77.0055	23.9977	91.8059	88.4819	NaN	178.6066	96.5116	...	2

5 rows × 30 columns

Raw data sample - Head

symbol date	AAPL	AXP	BA	CAT	CSCO	CVX	DIS	DWDP	GS	HD	...	PFE	PG	TR
2019-06-05	182.54	118.12	348.75	123.12	54.75	117.65	135.94	NaN	188.44	196.69	...	42.48	106.73	149.025
2019-06-06	185.22	119.43	350.64	123.39	55.10	120.68	137.21	NaN	189.81	197.17	...	42.71	107.38	148.180
2019-06-07	190.15	121.11	353.70	124.46	55.93	121.48	138.04	NaN	189.81	197.30	...	42.92	108.77	149.530
2019-06-10	192.58	122.66	353.80	125.74	56.42	122.29	137.07	NaN	194.12	198.05	...	43.07	108.72	149.320
2019-06-11	194.81	123.23	349.33	127.28	57.11	121.17	135.08	NaN	194.73	198.01	...	42.67	109.38	149.190

5 rows × 30 columns

Raw data sample - Tail

Objectives and methods of Exploratory Data Analysis

The objectives of the EDA are as follows:

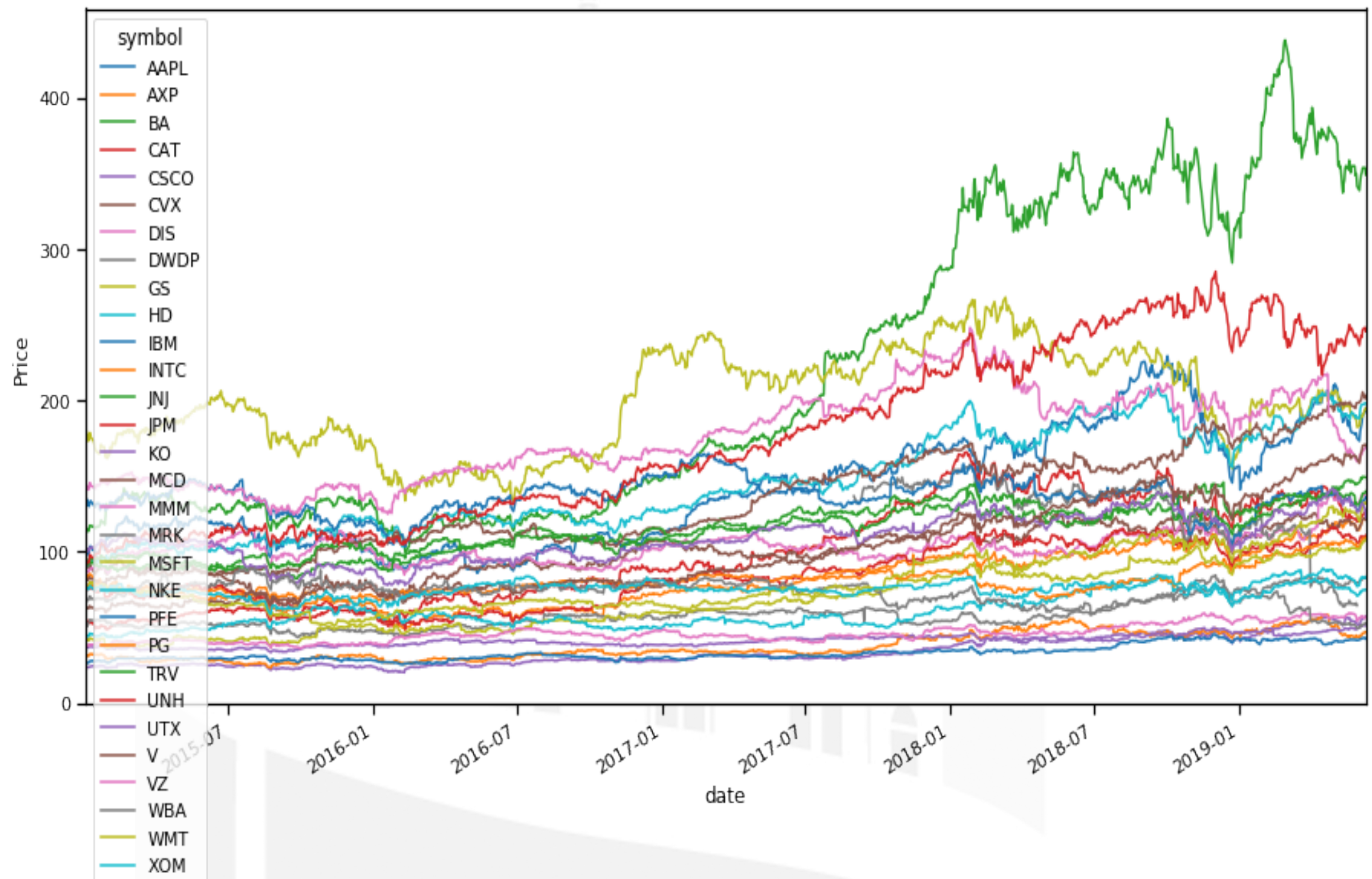
1. To get an overview of the data and distribution of the dataset, including data type, the size(shape) of the whole data frame, the highest price, the lowest price, the volume and the mean price by different window size in two months or 1 year(41 or 252 trading day).

2. Check for missing numerical values, outliers or other anomalies in the dataset. If there is any missing data, we can fill forward or remove the invalid data.

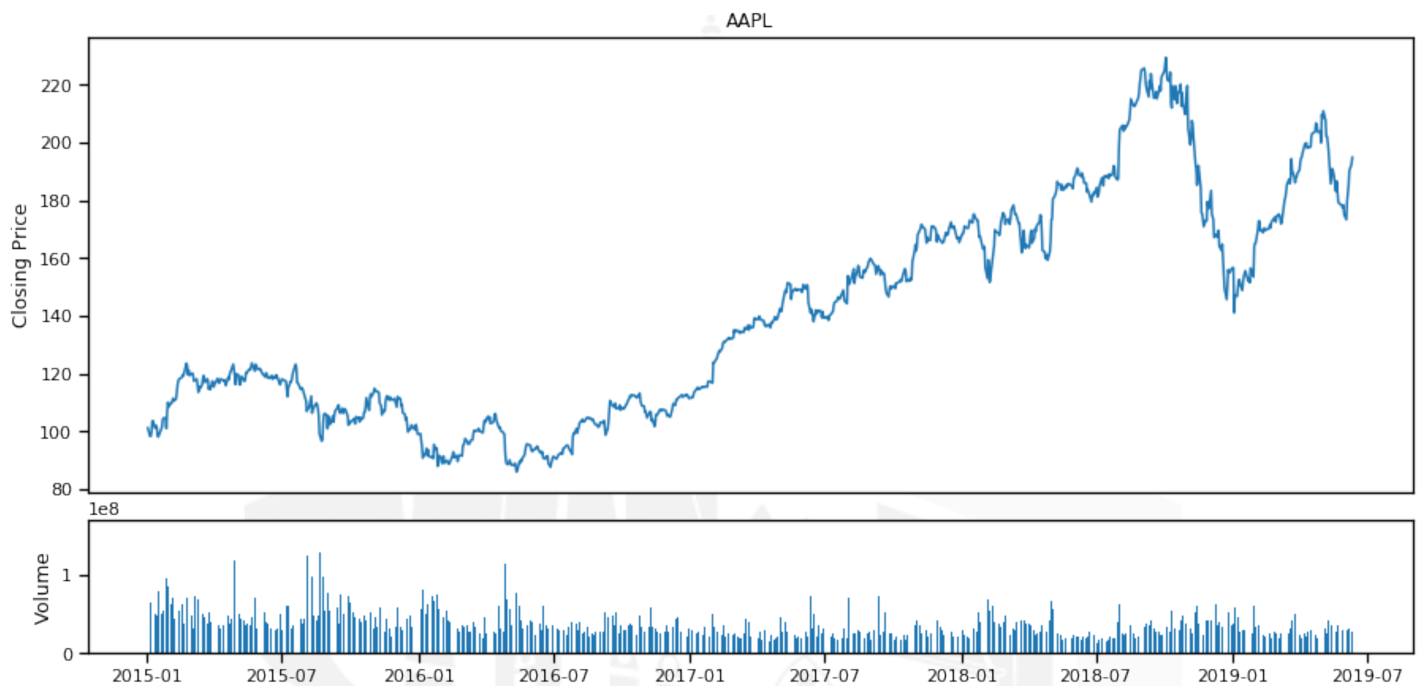
3. Discover patterns and relationships between variables in the dataset. We can solve this question by ranking the correlation, or visualizing the correlation by heap map.

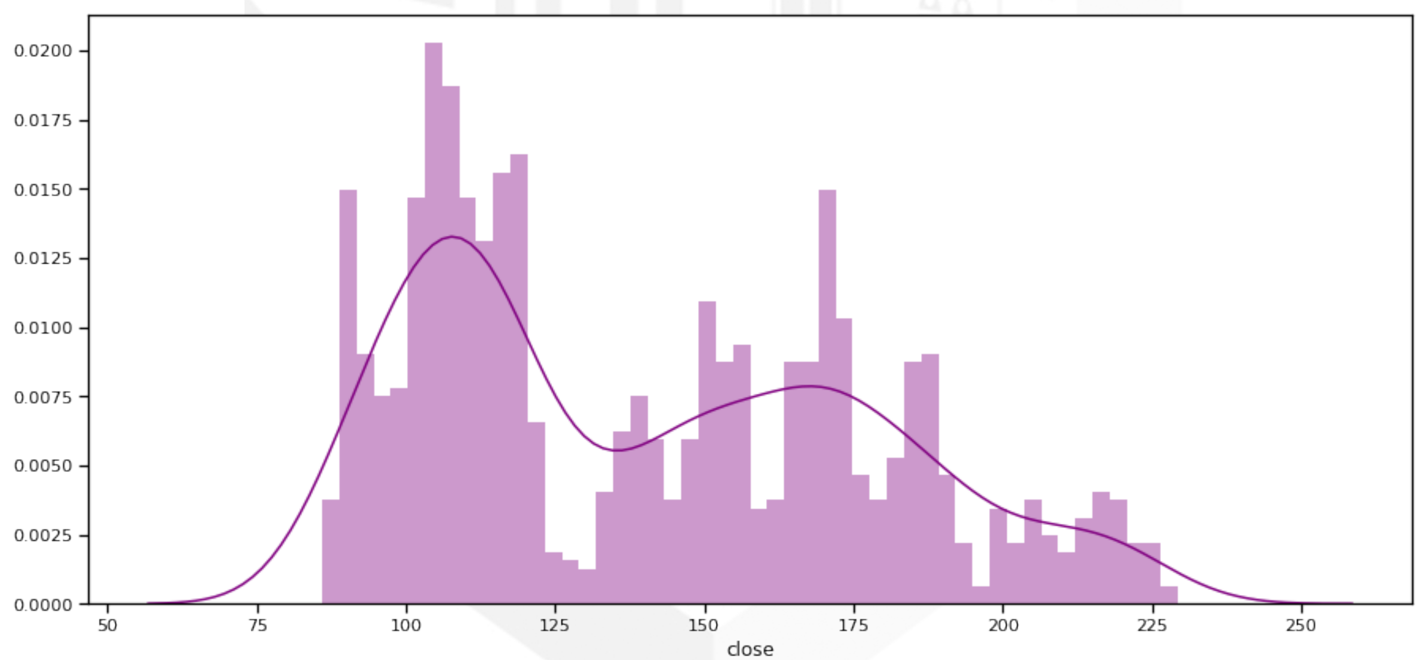
4. Check the underlying assumptions in the dataset. In this case, we hope to come up with a trading strategy and back test with the dataset.

Some results of Exploratory Data Analysis

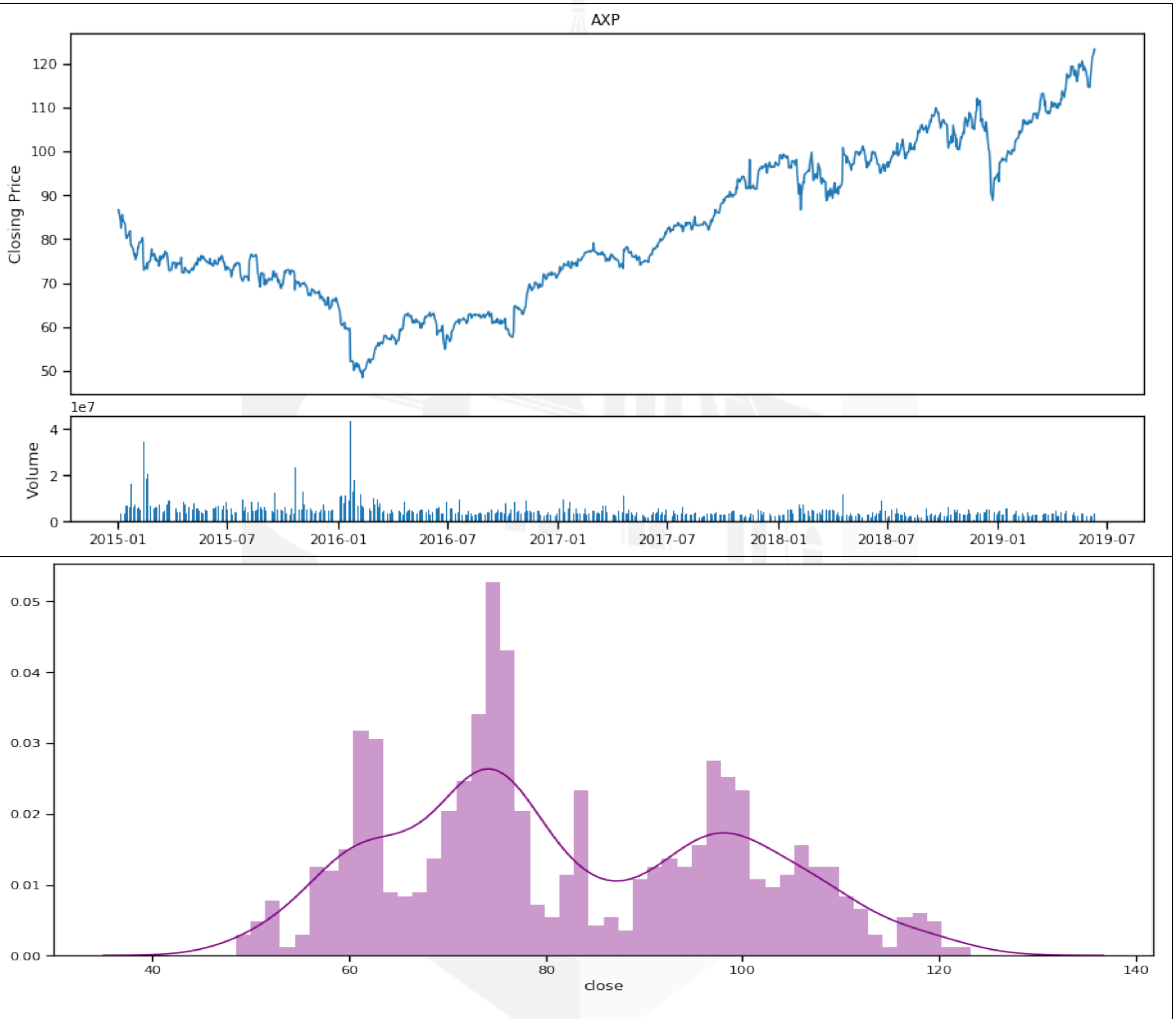


Trend of all 30 Dow Jones stocks

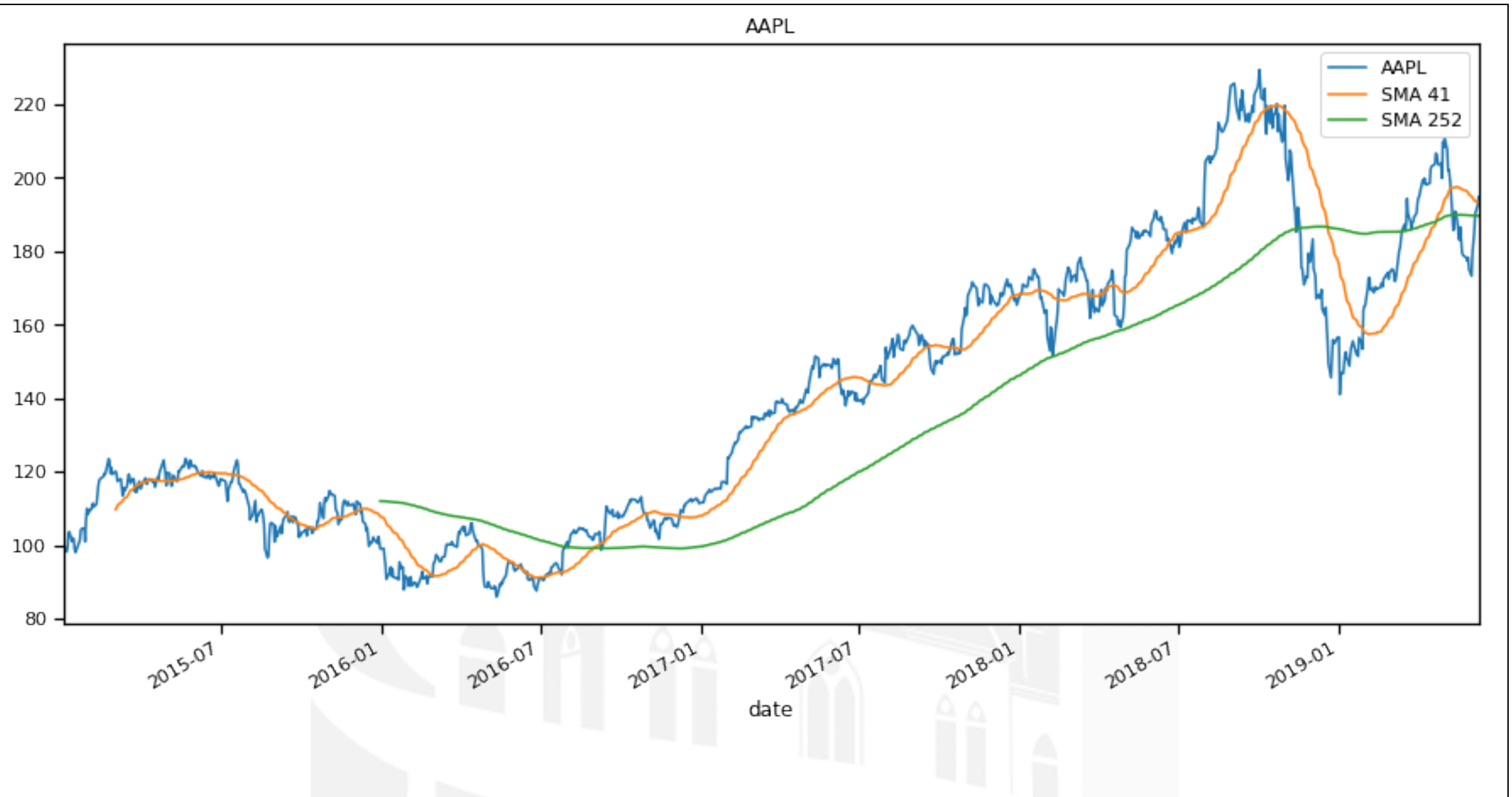




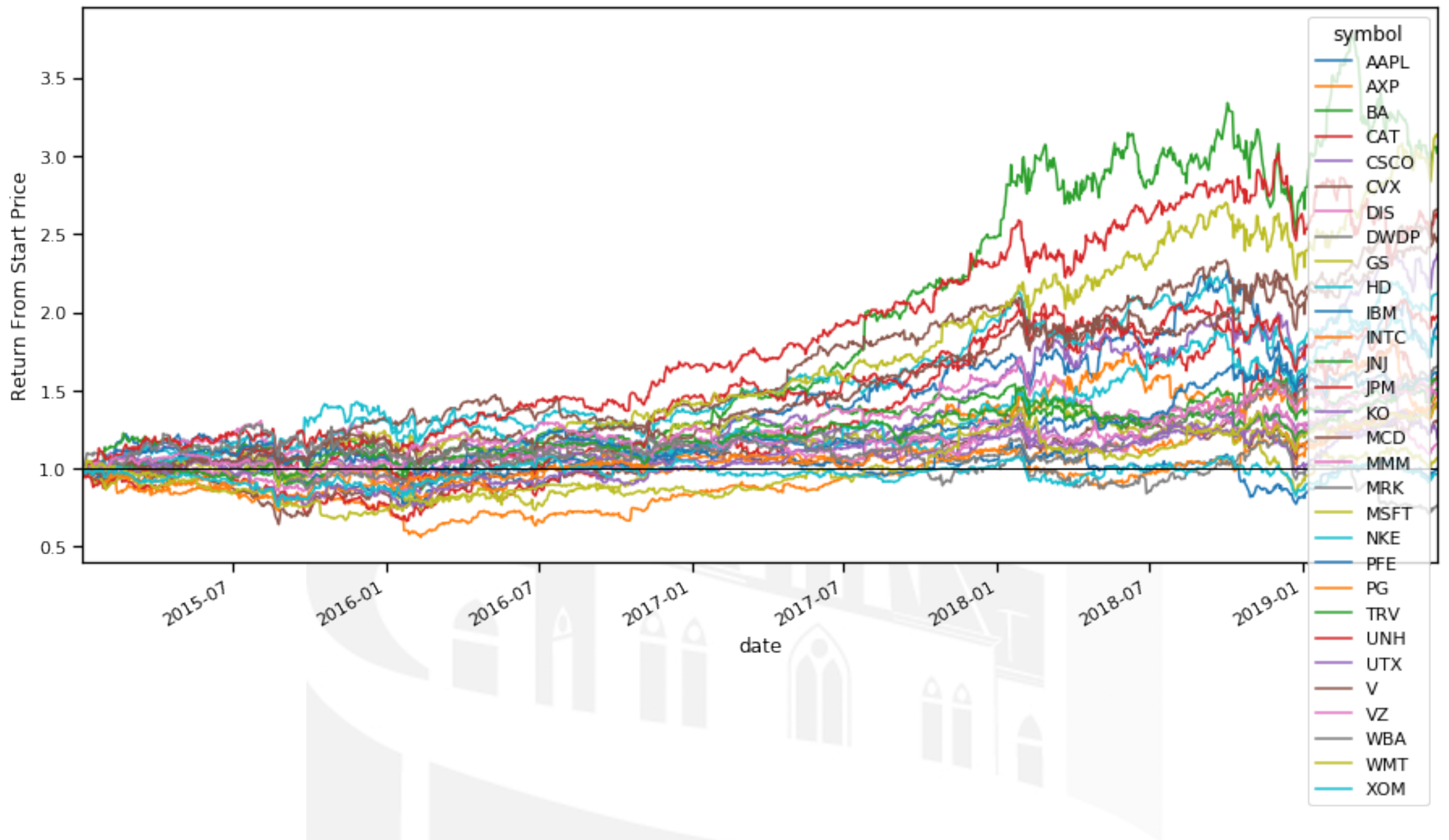
Close Price of AAPL



Close Price of AXP



42 and 252 days rolling mean of AAPL



Daily Returs of all 30 Dow Jones Stocks

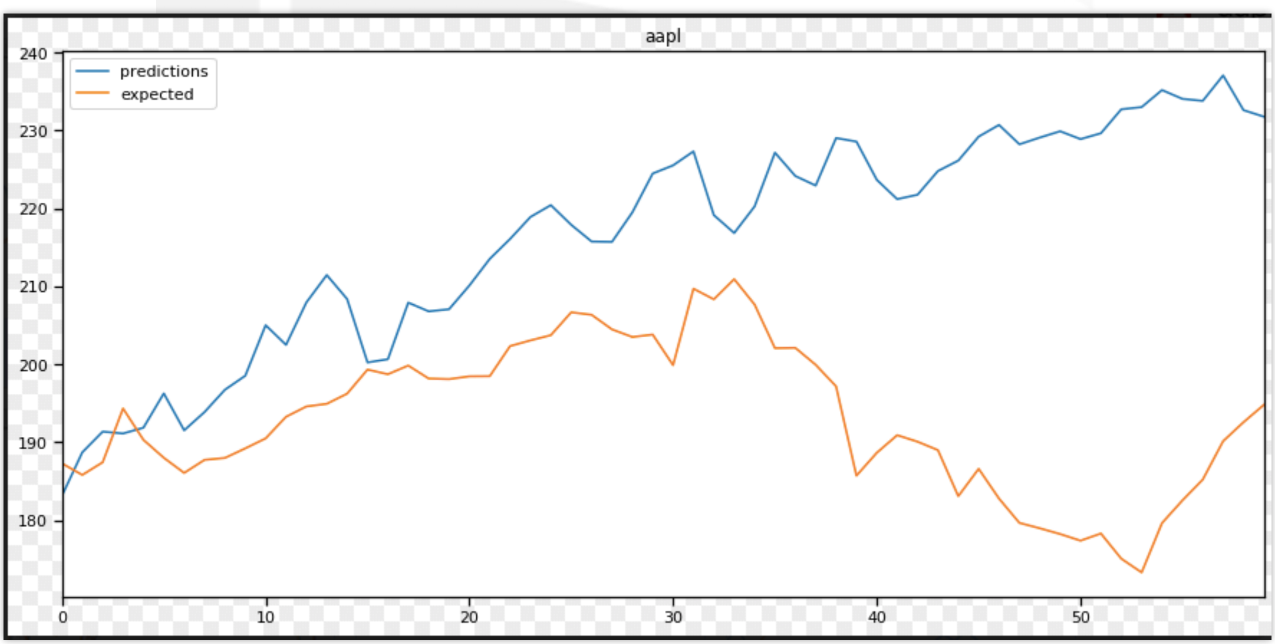
Algorithms and methods

Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables.
Considering the data, we have and the prediction we wish to do, starting the process with linear regression seems to be the best bet.
We know that, linear regression would not perform well, keeping up with the trends of the stock market, but constructing a good model will help us establish a well-formed foundation to carry out further predictions and achieve perfect models.

We split the data into a 66% training and 34% test set, before continuing with the algorithm.

Here's how linear regression performed for Apple.



Arima

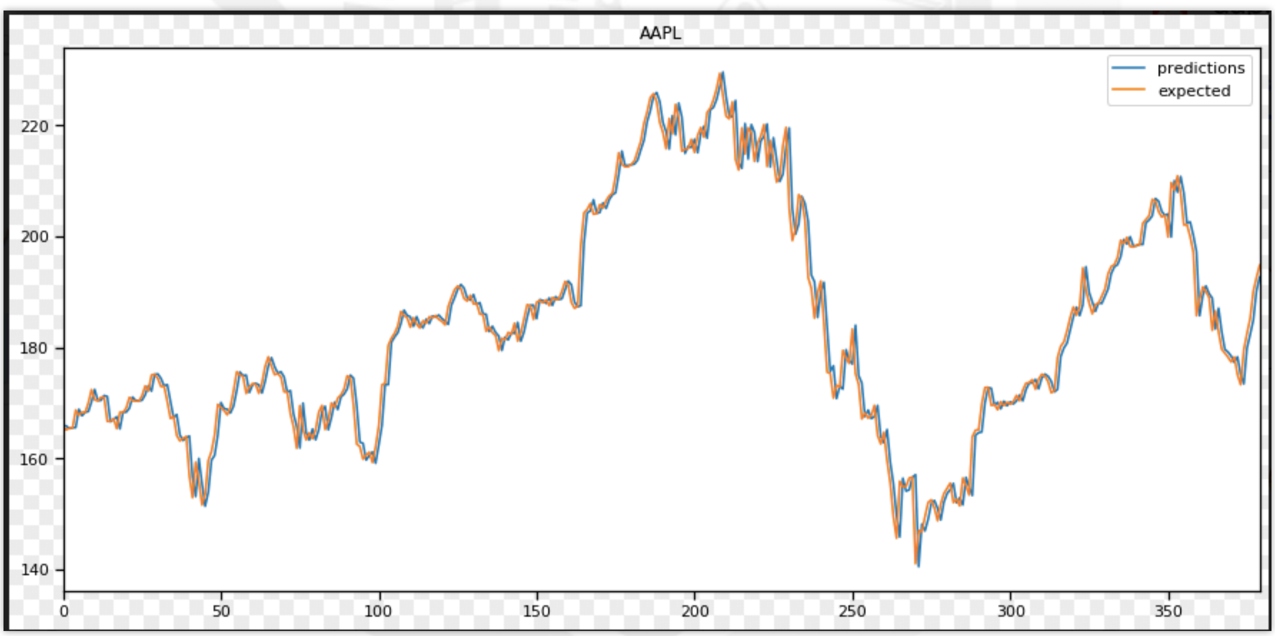
The obvious next bet would be to try Arima prediction. Using ARIMA model, you can forecast a time series using the series past values.

Since the Arima function doesn't demand a separate training and testing set. We gave it an initial training window of 66%

This means that the algorithm started working with a window of 66% of the data (around 737 values) from one column and predicted the 738th value. Next it considered the 738 values from the raw data set and predicted the next value. This continued till the dataset was eventually exhausted with all the values being tested for.

Arima turned out to be a very favorable model for our project.

Here's how Arima turned out for Apple,

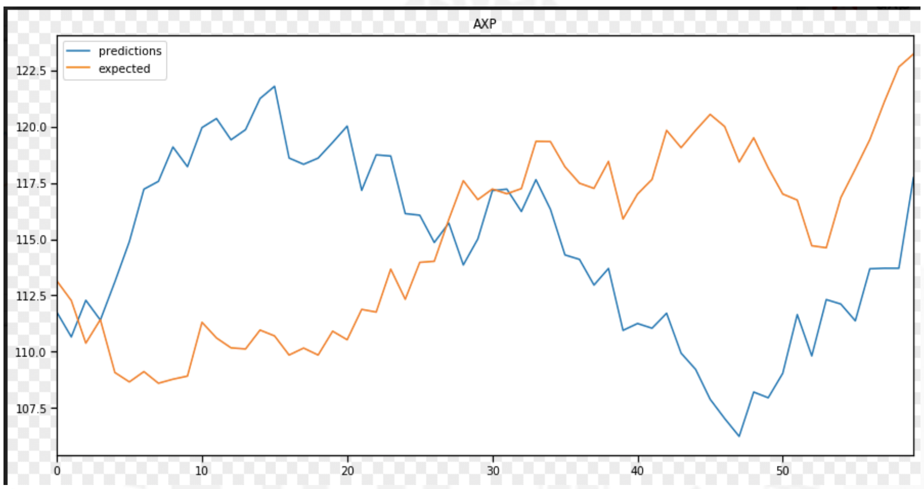



```
In [94]: df_aapl = arima_prediction(ts_aapl)

Test MSE: 10.900
```

Comparison between Linear Regression and ARIMA

American Express

Linear Regression

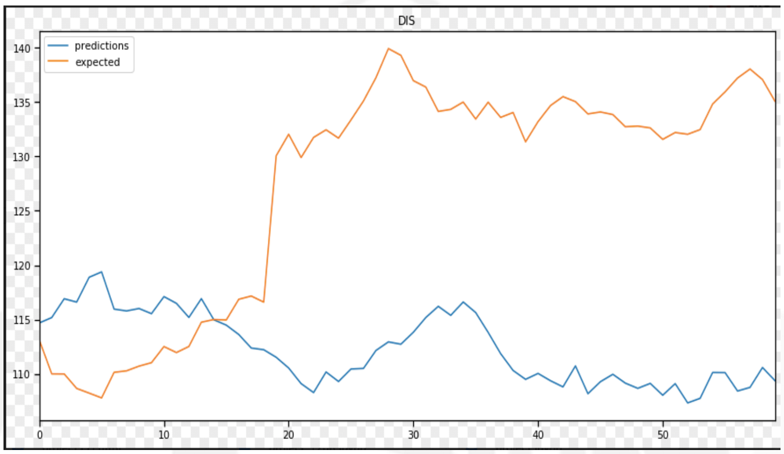


Arima

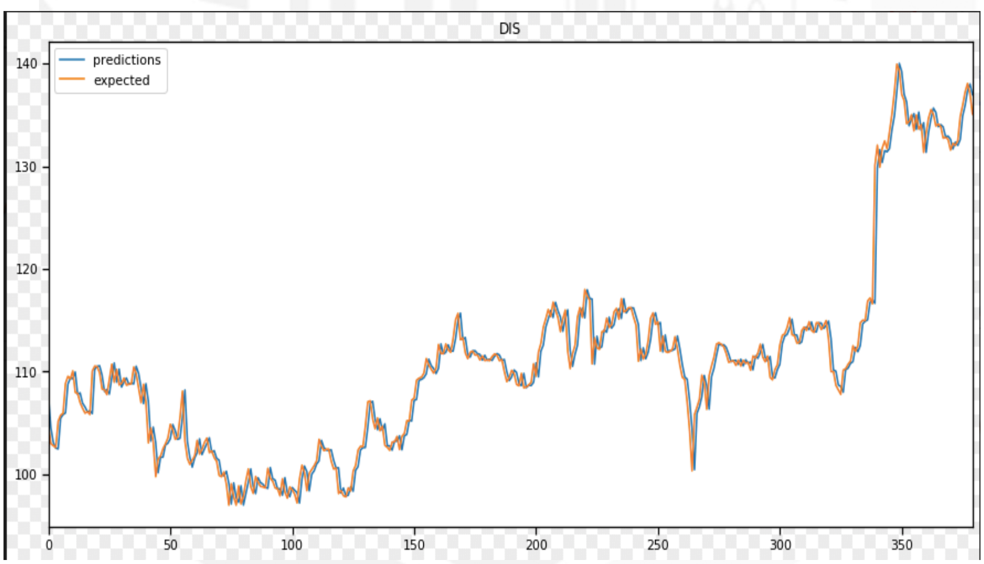

```
In [87]: df_axp = arima_prediction(ts_axp)

Test MSE: 1.866
```

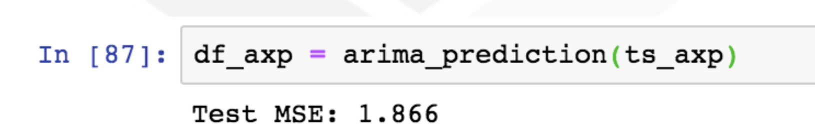
Walt Disney Company

Linear Regression



Arima





Conclusion

- Considering the fact that we have to deal with time-series data here and the trend can go up or down with time, choosing Arima is a better option for our dataset instead of models like linear regression
- We also have to understand here that the accuracy we achieved is so high because the time frame is only worth a variably short amount of time
- By analyzing decades worth of data there might be introduced certain volatility, change in variance
- For such situations, Arima falls short. To extend our accuracy we need to research working of more models

Future Plans

We researched a few models which can help us tackle issues related to volatility

- ARCH or Autoregressive Conditional Heteroskedasticity method provides a way to model a change in variance in a time series that is time dependent, such as increasing or decreasing volatility.
- GARCH or Generalized Autoregressive Conditional Heteroskedasticity allows the method to support changes in the time dependent volatility, such as increasing and decreasing volatility in the same series

References

- Dow Jones Industrial Average (https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average)
- Dow Jones Stock dataset (https://www.kaggle.com/timoboz/stock-data-dow-jones)
- Seasonal Adjustment (https://en.wikipedia.org/wiki/Seasonal_adjustment)
- Linear Regression Example (https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html)