Joshua Lugo
Daniel Muniz

Stock Analysis by Industry

Abstract

We look to examine trends in stock prices sorted by industry, to perform various analyses. We find that volatility greatly increased following the market crash in 2009, and while most stocks eventually recovered to their pre-crash values, some did not. For out changepoint analysis, we were able to find meaningful comparisons between and within various industries.

Introduction

The stock market is one of the most lucrative investment opportunities in the world. An entire multi-billion dollar industry is built around trying to predict fluctuations and movements of stocks. While it is a difficult, oftentimes nebulous task, being able to predict even a 1% change can lead to a big payoff.

Our goal is to analyze indices of company stocks from specific industries to see if we can find underlying trends specific to those industries. The industries we'll look at are Tech, Energy, Finance, Capital Goods, Healthcare, Customer Services, Consumer Non-Durables, and Public Utilities . We'll attempt to make these analyses by plotting using various tools and techniques, including Moving Averages, Candlestick charts, Market Return, and Change-Point graphs. We will look to examine volatility, especially after the market crash in 2009. We also will examine changepoints, and hope to find that any anomalous dips and rises will correspond with notable events in the U.S. and abroad.

Related Work

In order to help us explore the data, we found an article, lecture, and tutorial describing stock analysis and based our analysis from the different methods we learned. Karlijn Willems's article was our starting point where we initially started exploring the stock data with Yahoo! Finance and ultimately modeled our volatility analysis as described in the tutorial.

Curtis Miller's piece is from a lecture and describes many different areas for stock analysis. While we did initially explore the data through candlestick plots as he described, we ultimately decided to focus on other methods in order to analyze trends between industries. Curtis' lecture showed us how to simply calculate the returns from a stock for a given start date and explained why we should compare our stocks to SPY.

William Koehrsen's article introduced us to Stocker, "a Python class-based tool for stock analysis and prediction" that he had made himself. We used Stocker for the majority of our

analysis, but we modified the code heavily to tailor it to our specific needs. We write about this in greater detail in the Methodology section.

Wayne Taylor's article on change-point analysis provides an excellent overview on the pros and cons of the analysis, which was a key part of our experiment.

Experiments/Methodology

Our experiments were run in Python 3.6, using Jupyter notebook through Anaconda. We pulled almost all our of data through the "quandl" and "yahoo_finance" packages. Little pre-processing of the data was needed since stock data is both well documented and easily accessible.

Experiment 1: Volatility Analysis

Data Exploration

Initially we pulled our data from Yahoo! Finance through the pandas-datareader package. Because the Yahoo! developers were still working on a fix for query data from their API, we would then also have to import fix_yahoo_finance. While it did work most of the time, occasionally we would get the following error when trying to retrieve the stock data: `ValueError: zero-size array to reduction operation maximum which has no identity`. We then found a permanent fix for our project by using pandas_datareader.data to pull SPY's data (a fund which attempts to imitate the S&P 500 stock index), and quandl to pull everything else. Quandl pulls a variety of time series data for stocks consisting of the price the stock opened on a given day, adjusted close price, and volume, to just name a few.

Data Preprocessing

The preprocessing was very limited as the data from quandl is very reliable. We were mainly interested in the Adj. Close, a stock's closing price that's been amended due to any corporate actions occurring before the market's next opening, so we used Stocker to create DataFrames consisting of the dates and the Adj. Close for five major stocks in each industry. We define major as the being part of the top ten largest market caps for a given industry, as described by NASDAQ.

Implementation

As previously mentioned, while there were many benefits to using Stocker, we needed to modify that code and create a methods file to automate the retrieval of the stock data and the generation of most of the plots. The only inputs that were necessary were the names of the stock indices. The methods.py file includes some functions that weren't used in the final iteration of the analysis but every method was used during the initial data exploration. As our understanding of both the material and techniques available to us grew, we kept adding to methods.py. While we didn't keep a change_log, we did keep our initial attempts to explore the

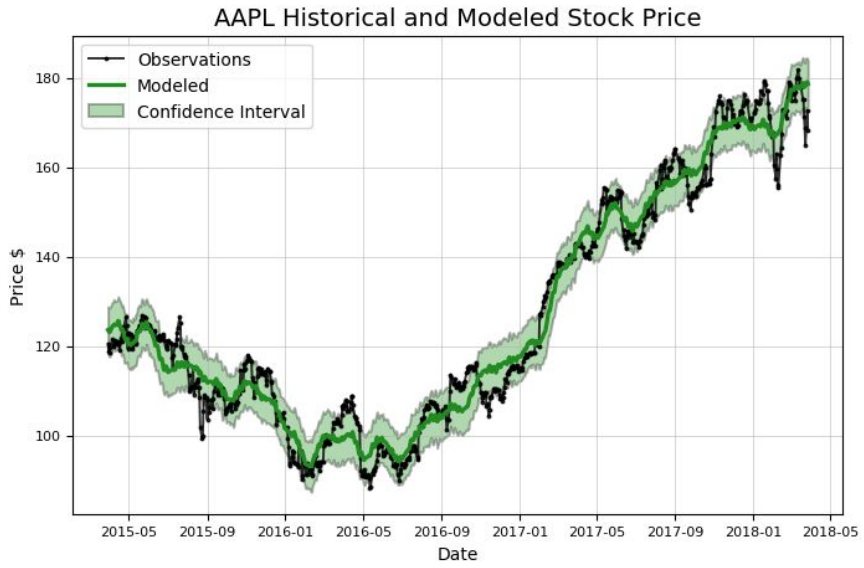
data in our github (these attempts still run properly because we kept the relevant methods in methods.py).

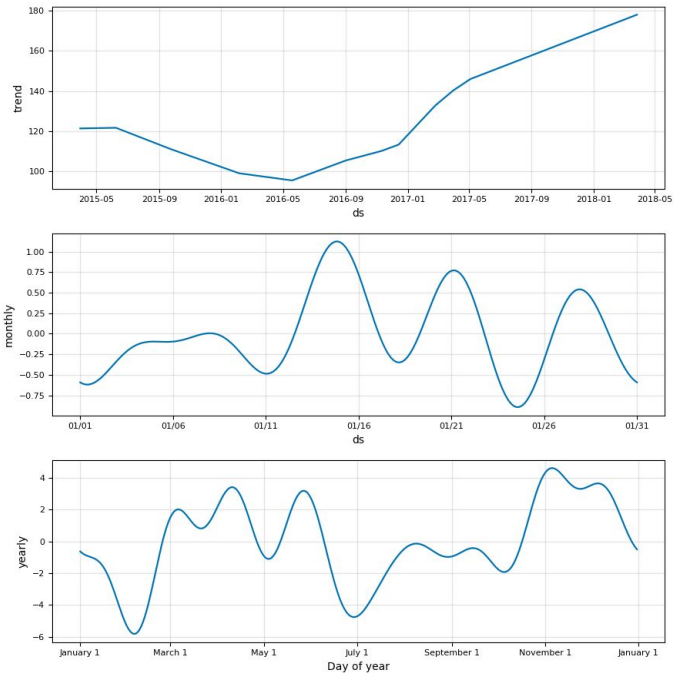
Another issue with Stocker is that it only lets you look at one stock at a time, which would complicate our analysis of 40 stocks across eight different industries. This is another reason why we modified parts of the stocker.py and created a methods.py; to house the different functions we would applying, automate the process for different industries, and to be able to compare multiple stocks in the same plot.

The first industry we explored was Technology, so I'll describe our initial data exploration using Tech stocks as an example. We first plotted each stock's Adj. Close price for five stocks in each industry, but in order to better compare them, we needed to calculate the returns for each stock. We calculated the returns by looking at the Stock Price for a given start date for our stock and dividing the Stock Price for each day afterwards start date.

As you can see, by looking at the graphs, while Google's (GOOGL) has a larger Adj. Close than the others, it's actually Apple (AAPL) that has had larger returns since 2007. While it may be a bit hard to see, Facebook's (FB) stock data actually starts in 2012. This was one issue that we ran into but fixed by creating a check which would look to see if there was no data for our given start date, then it would set the start date for that stock to the first date with information.

We then used Stocker's create_prophet_model() function to create a model for each stock; this is based on the fbprophet package. Stocker allowed us to visualize the general trend for each stock, as well as monthly or yearly trends. For Apple, they trending upwards but had a dip around May 2016. Perhaps this is something that we should look into? If we look at the yearly trends, then we can see how there are dips in February and July but a local maximum in November. Perhaps Apple tends to release products in November?

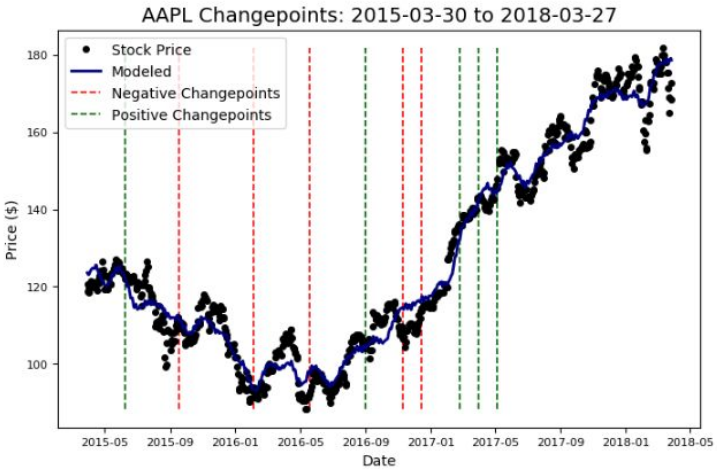




We can supplement these initial trend observations with Stocker's changepoint_date_analysis method which serves to illustrate the dates with the greatest rate of change. Identifying these points could help us not only predict swings in the future, but also gives us an opportunity to explore the news related to the company around that given date. While Stocker does automatically plot the 10 largest changepoints, it only prints out the exact dates for five of them. For our purposes, we need all ten, so we again modified the stocker.py code to allow us to see the exact dates for all ten depicted change points. Here is the example for Apple.

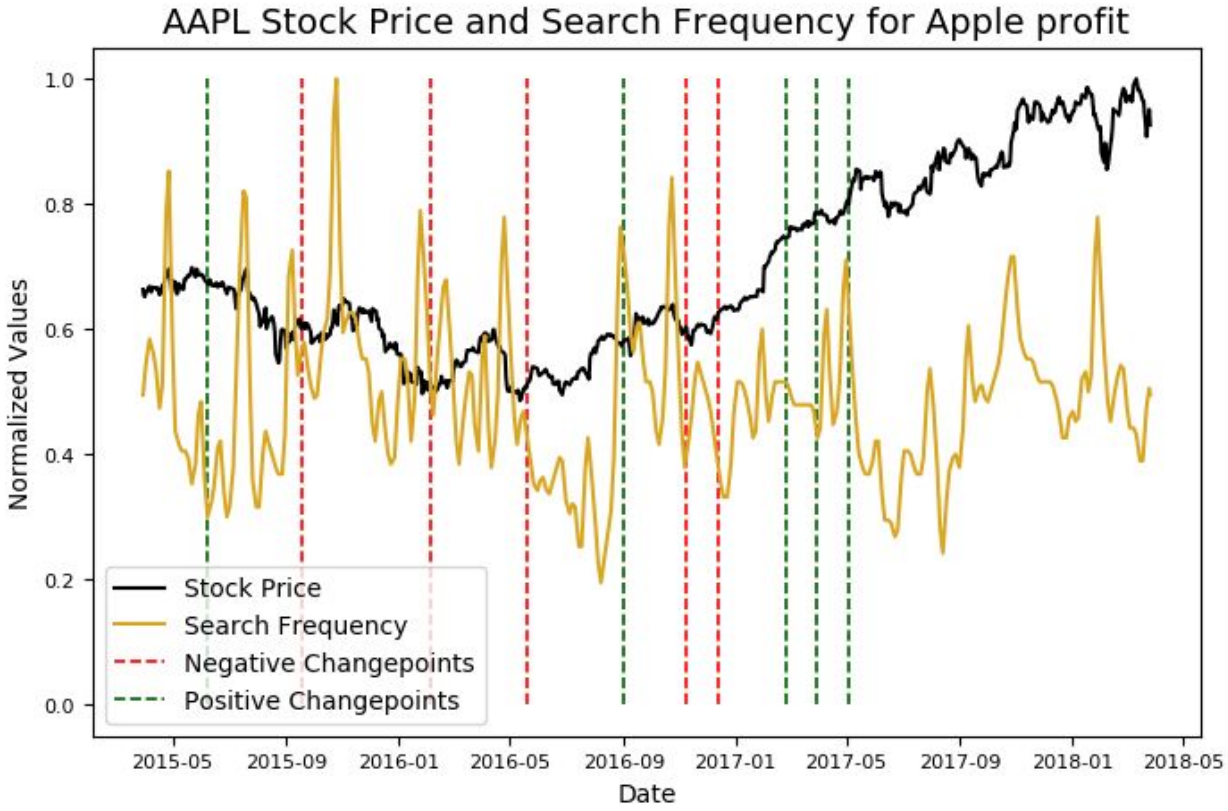
```
Changepoints sorted by slope rate of change (2nd derivative):

          Date  Adj. Close     delta
48  2015-06-08  122.392291 -0.622993
120 2015-09-18  109.140225  0.058418
216 2016-02-05   91.326278  0.331918
288 2016-05-19   92.060491  0.772369
361 2016-09-01  104.867483 -0.157710
409 2016-11-09  109.510470  0.163036
433 2016-12-14  113.767235  1.063407
481 2017-02-24  135.553035 -0.337295
505 2017-03-30  142.764147 -0.272297
529 2017-05-04  145.343086 -0.425737
```



We decided that for the models and the changepoint analysis, each graph should just consist of data from one stock, else it would be too noisy and we wouldn't be able to draw meaningful insights.

An additional function of Stocker's changepoint analysis is that it lets you look up key search terms as well (using pytrends) and plots normalized values for the Stock Price and Search frequency. However, Koehrsen's code gives an error as documented here. We eventually created a fix within stocker.py, but had already manually searched up for events around each change point, so we decided to not include it. Here's an example of what it looks like when we got it to run:

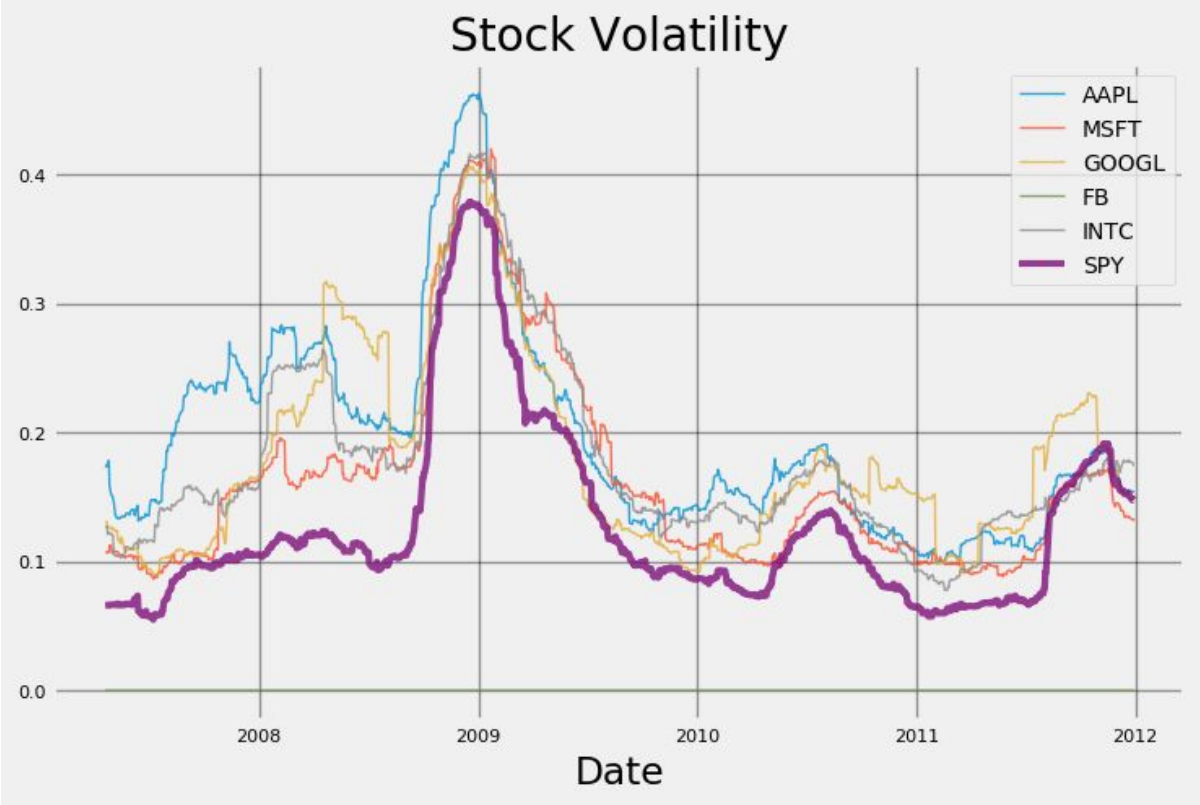


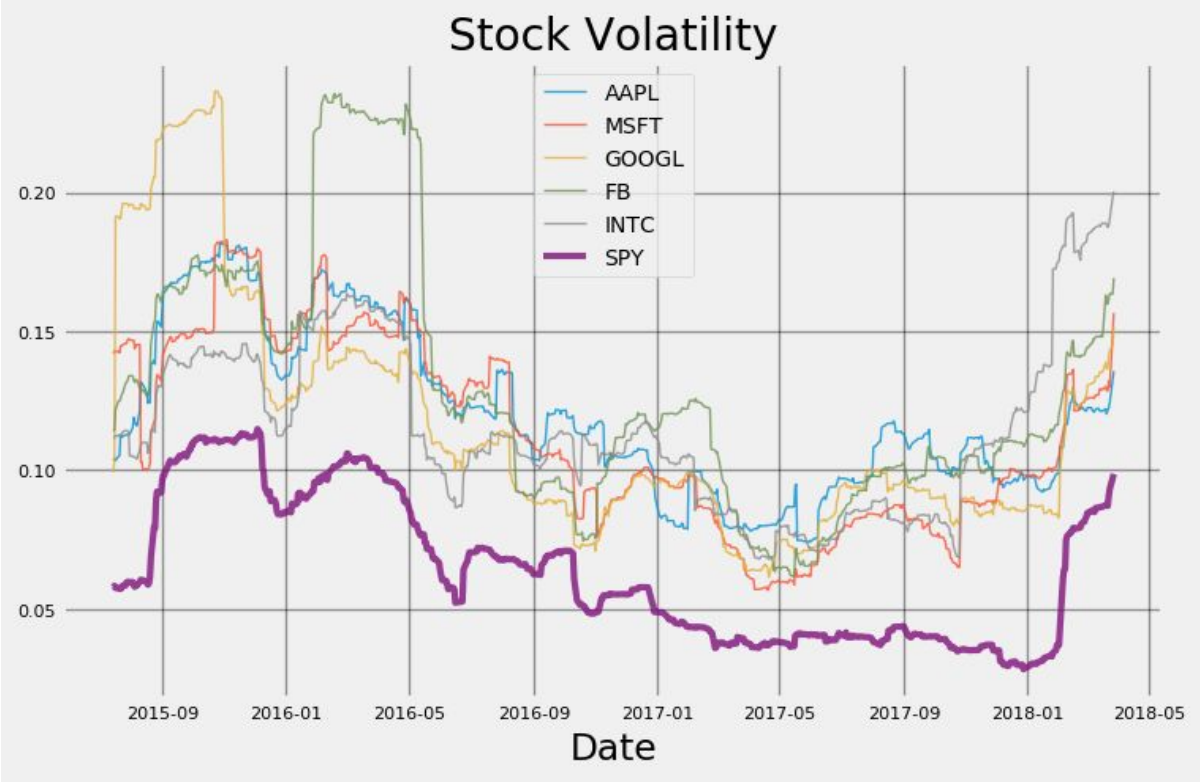
We had also done some initial candlestick graphs, but found that volatility graphs would be more effective for comparing trends, both around the stock market crash and in the years afterwards. These graphs become more powerful by plotting multiple stocks together, so we chose to do one per industry for two time intervals. The first time interval was to capture the data before and after the crash. The second interval was to capture the volatility during the same time frame as the changepoint analysis.

Results

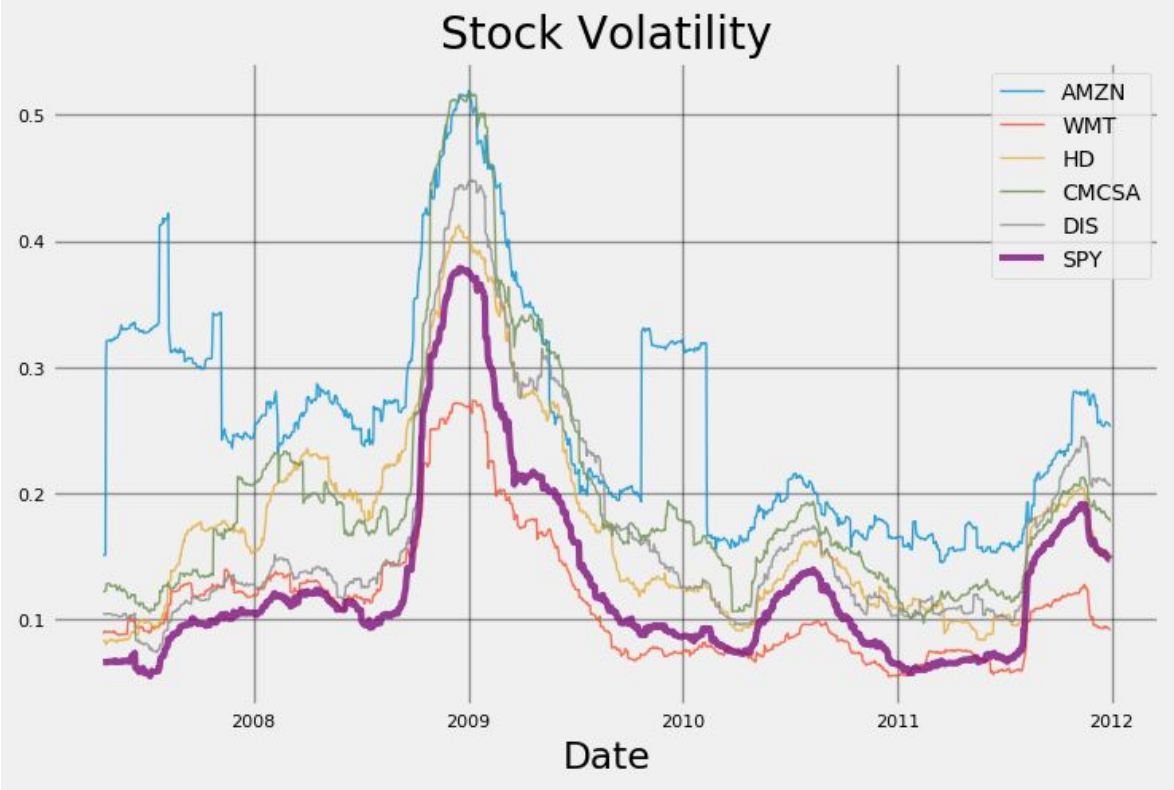
Below I'll include the volatility graphs for each industry and for the two time frames described above.

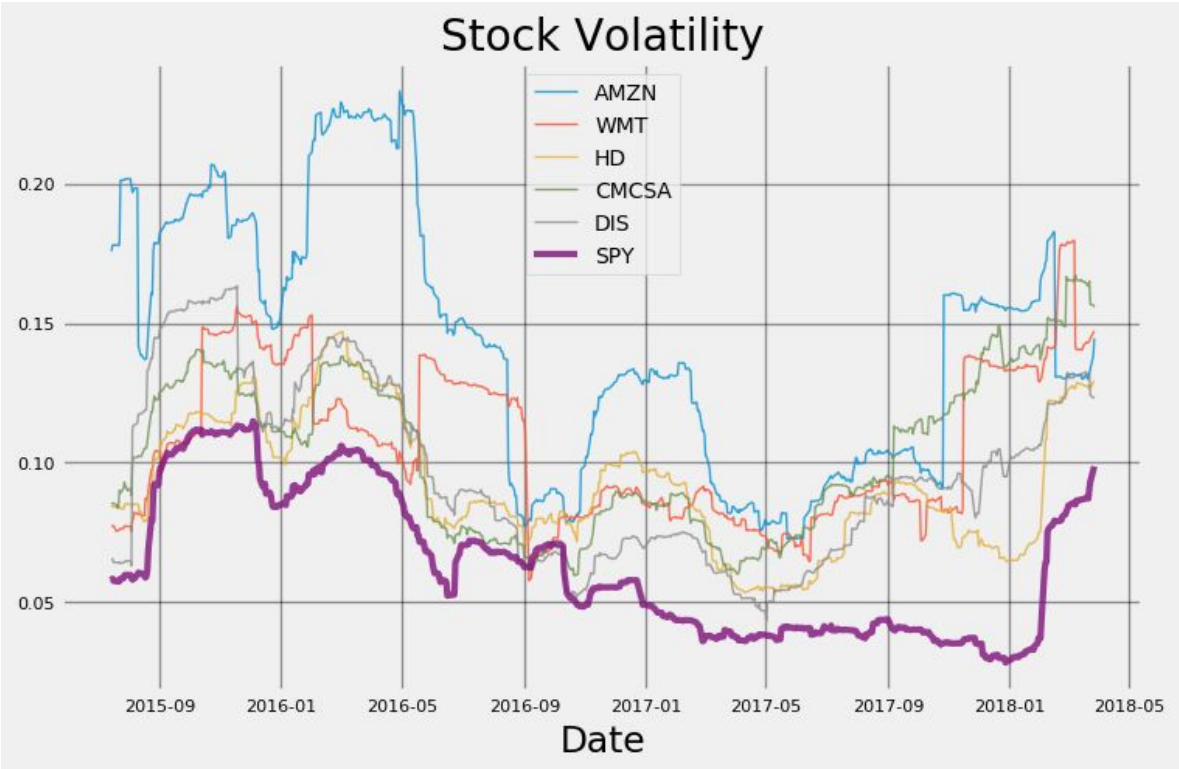
Technology


Stock Volatility
AAPL
MSFT
GOOGL
FB
INTC
SPY
0.4
0.3
0.2
0.1
0.0
2008
2009
2010
2011
2012
Date

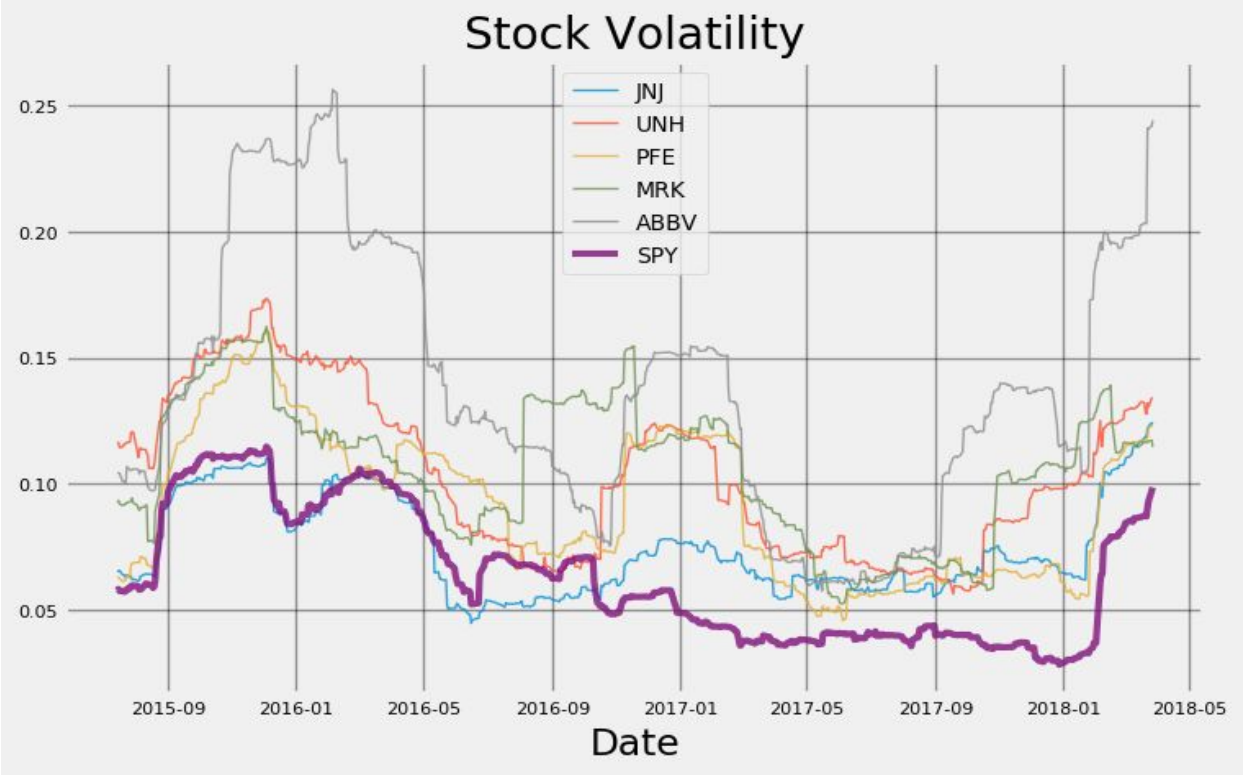

Stock Volatility
AAPL
MSFT
GOOGL
FB
INTC
SPY
0.20
0.15
0.10
0.05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
Date

Customer Services



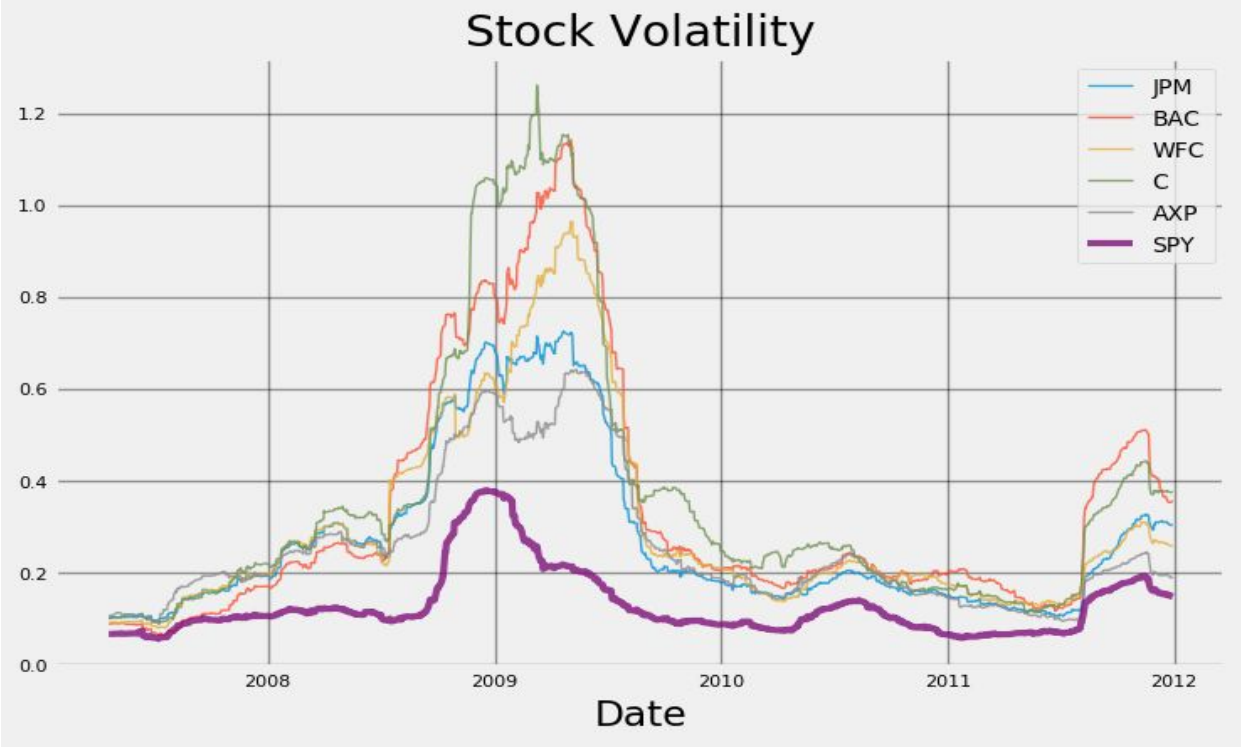


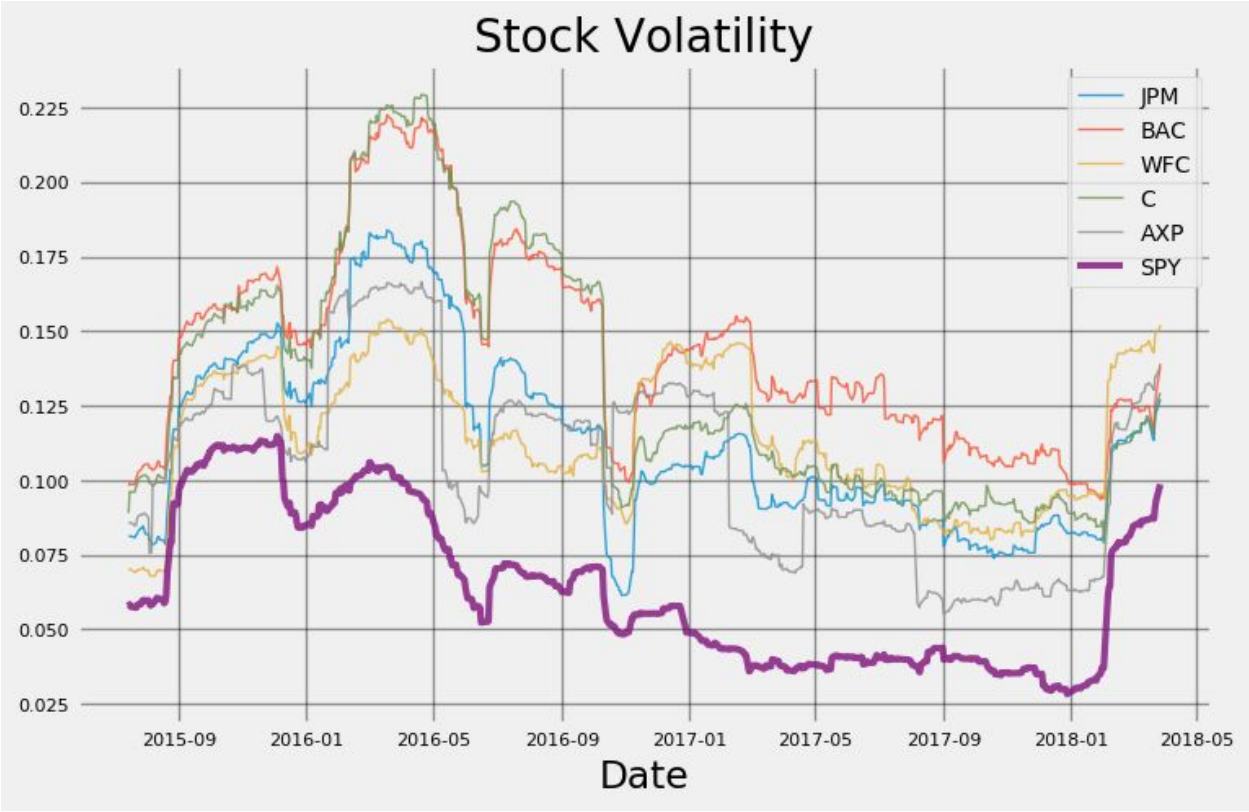
Healthcare


Stock Volatility
JNJ
UNH
PFE
MRK
ABBV
SPY
0.25
0.20
0.15
0.10
0.05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
Date

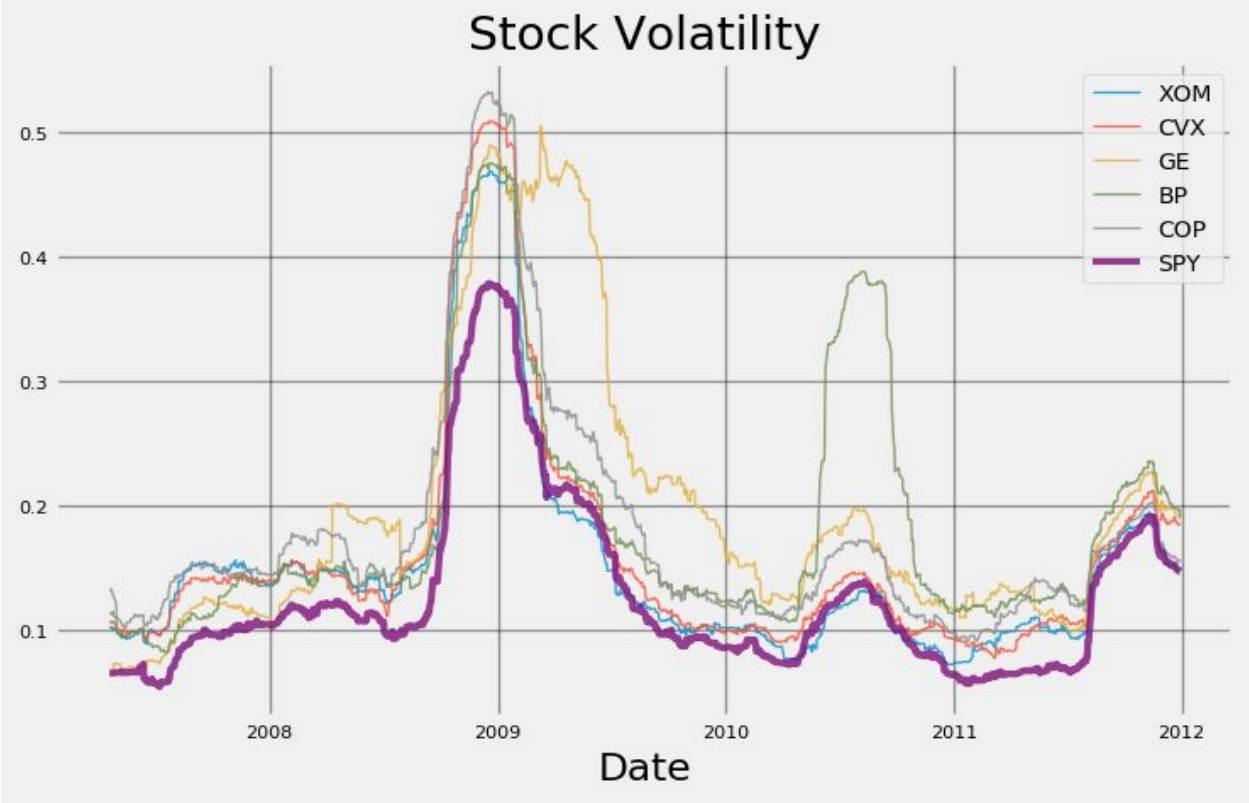

Stock Volatility
JNJ
UNH
PFE
MRK
ABBV
SPY
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2008
2009
2010
2011
2012
Date

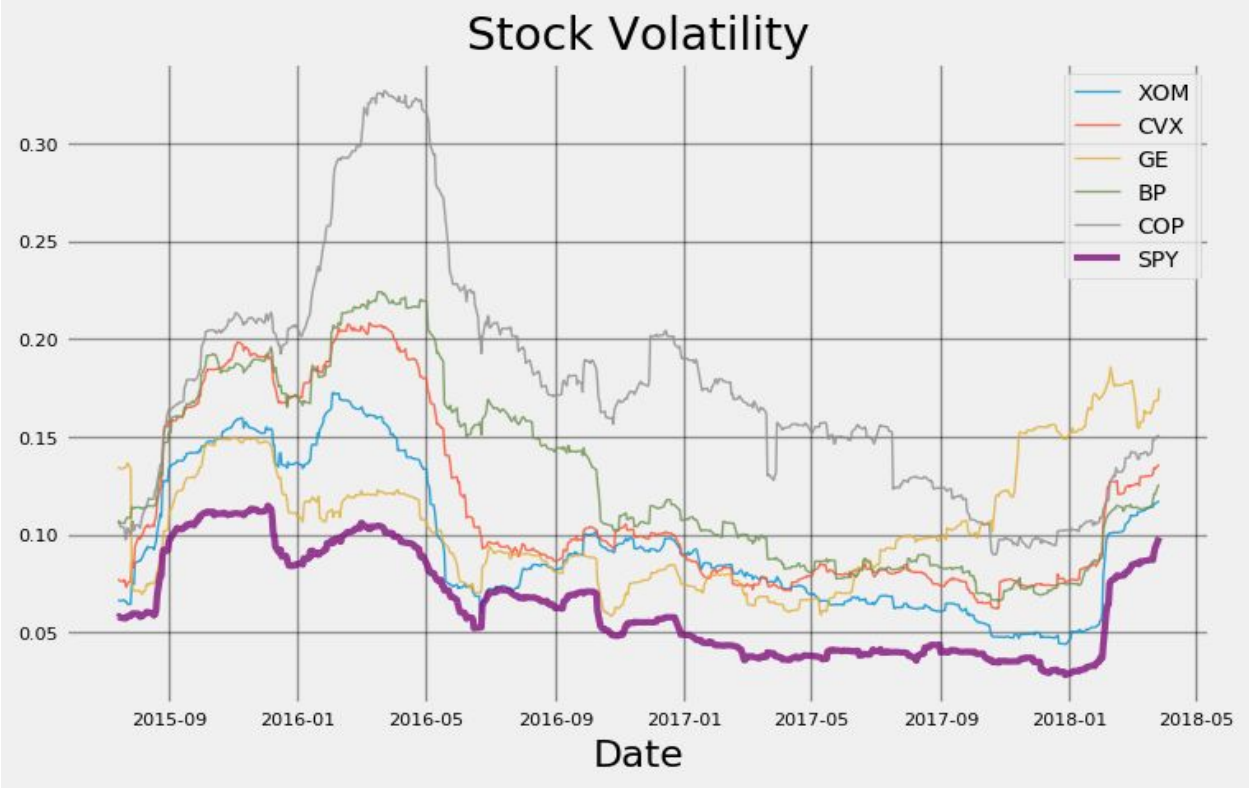
Finance


Stock Volatility
JPM
BAC
WFC
C
AXP
SPY
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2008
2009
2010
2011
2012
Date

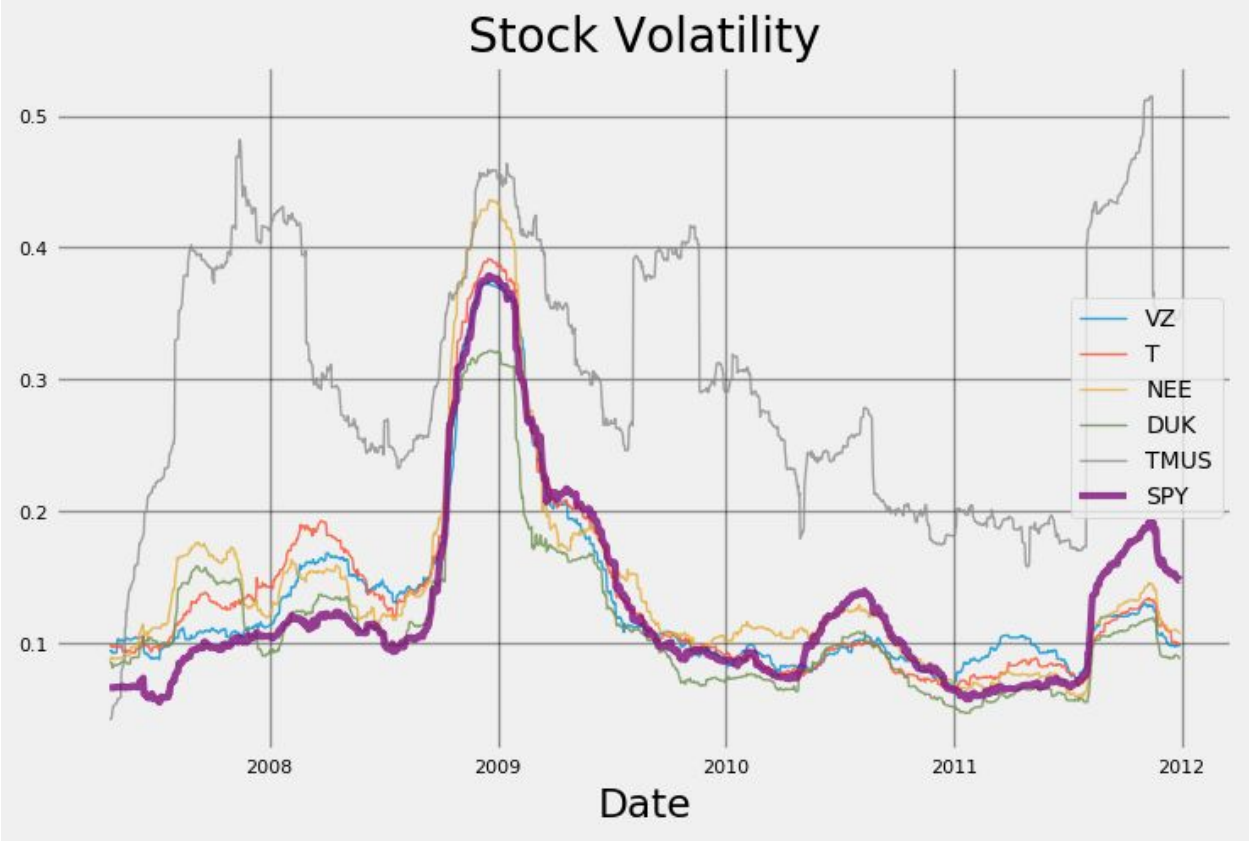

Stock Volatility
JPM
BAC
WFC
C
AXP
SPY
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
Date

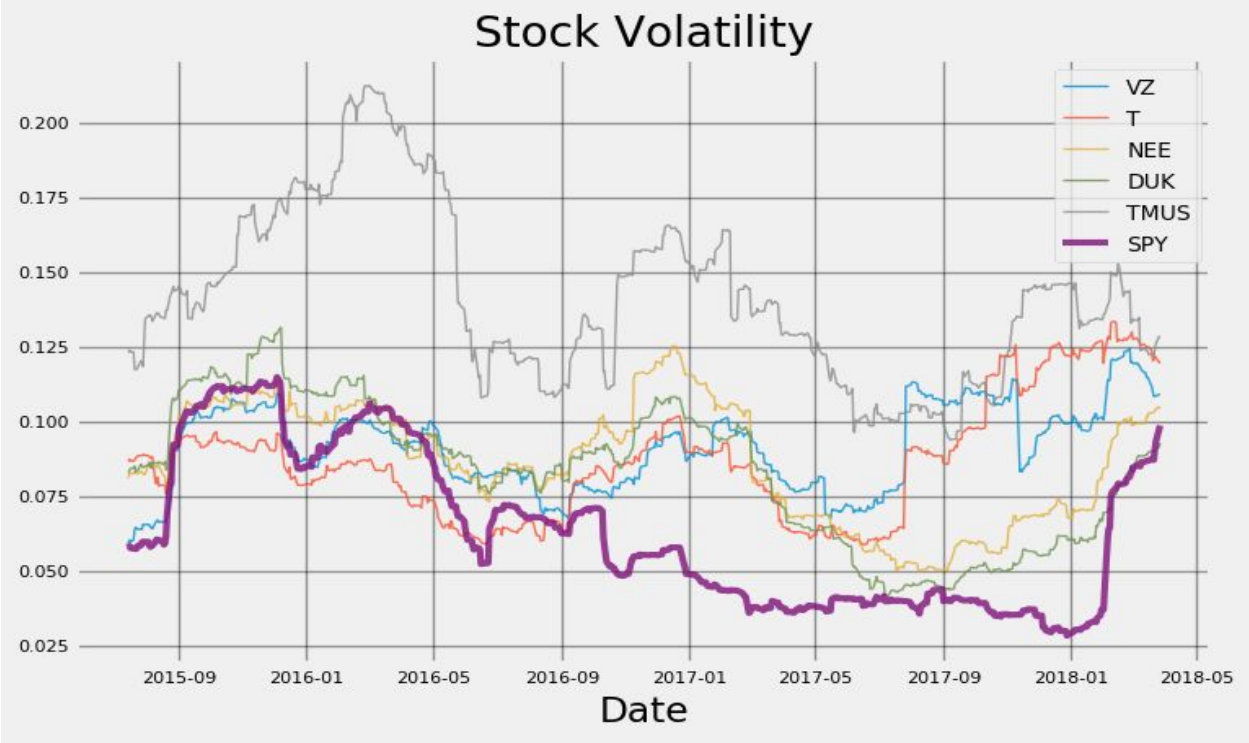
Energy


Stock Volatility
XOM
CVX
GE
BP
COP
SPY
0.5
0.4
0.3
0.2
0.1
2008
2009
2010
2011
2012
Date

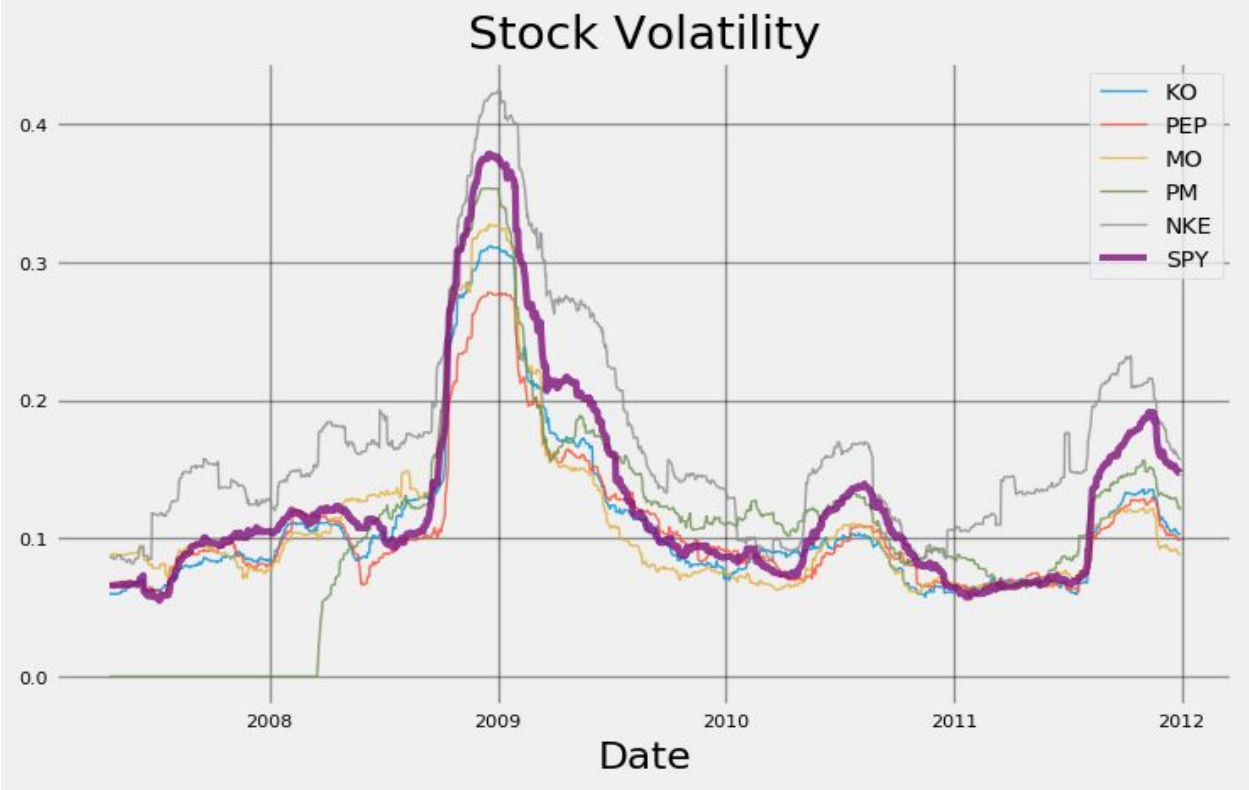

Stock Volatility
XOM
CVX
GE
BP
COP
SPY
0.30
0.25
0.20
0.15
0.10
0.05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01
2018-05
Date

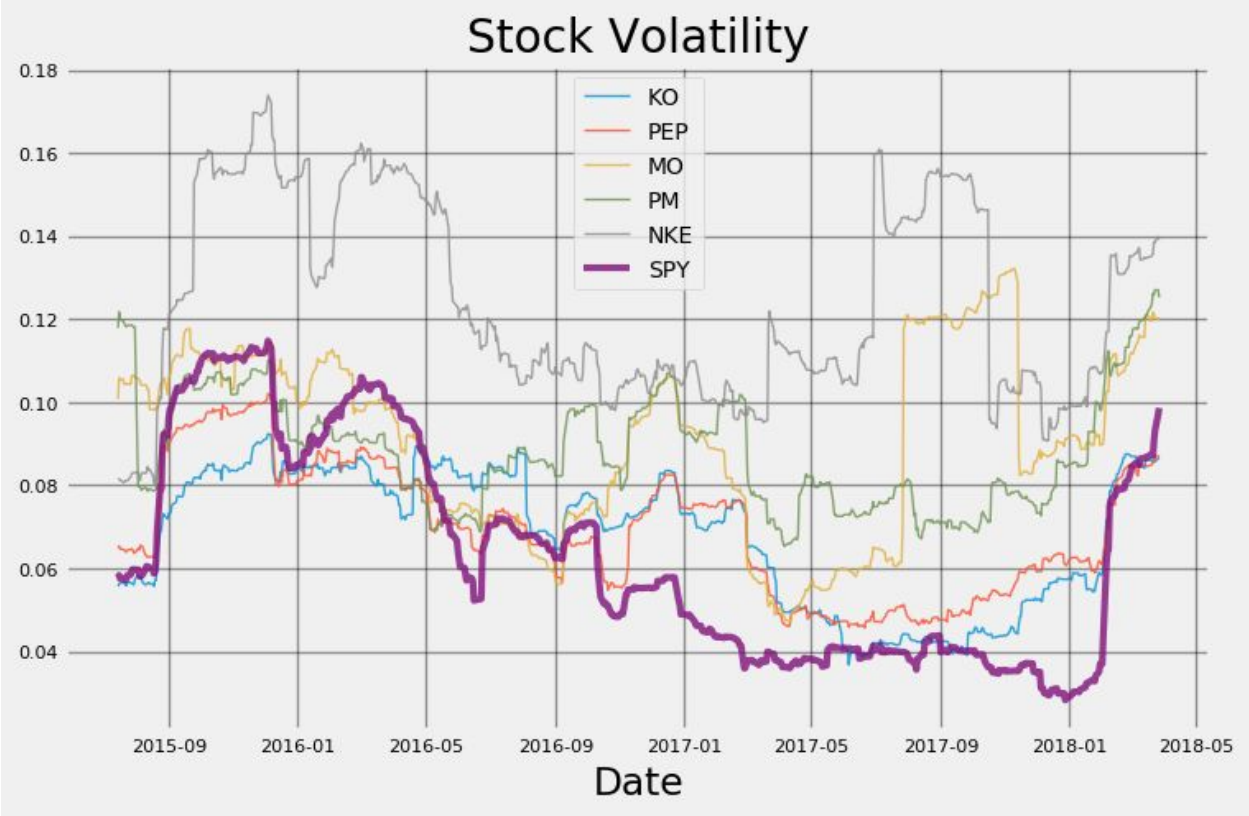
Public Utilities



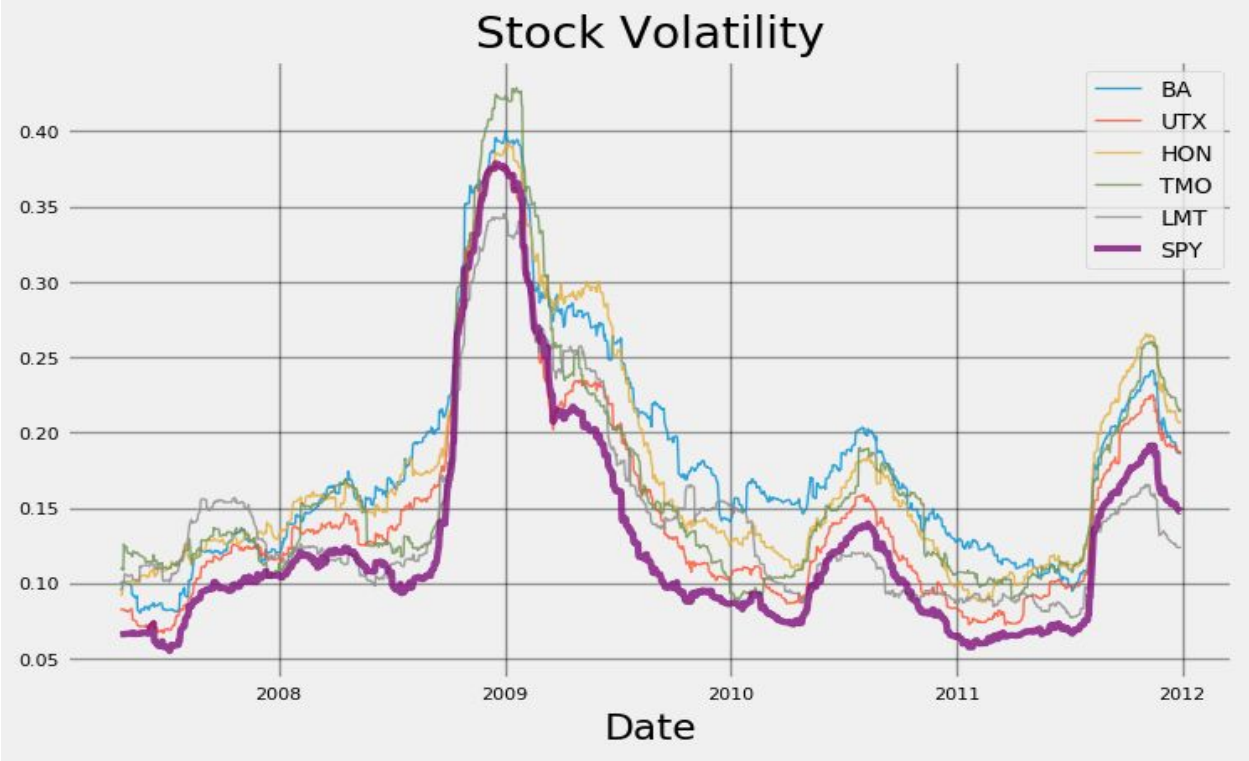


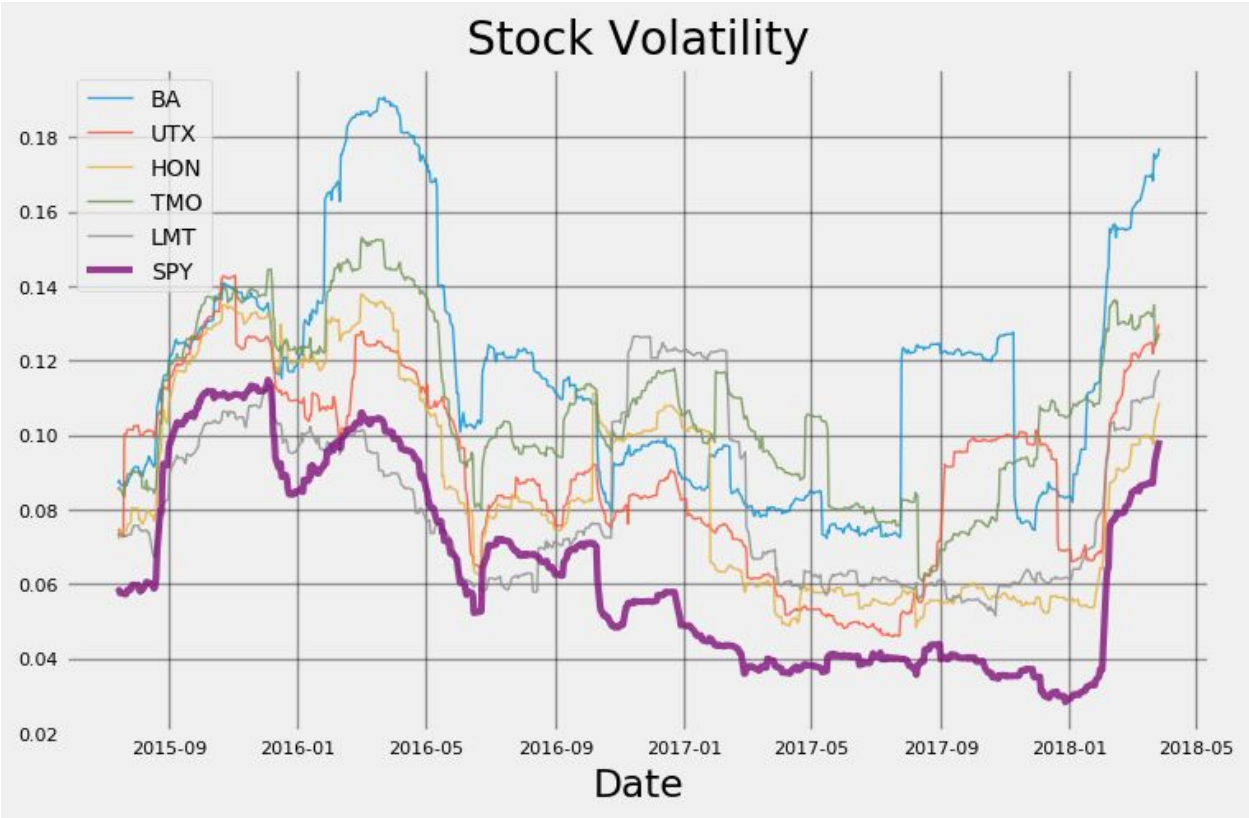
Consumer Non-Durables





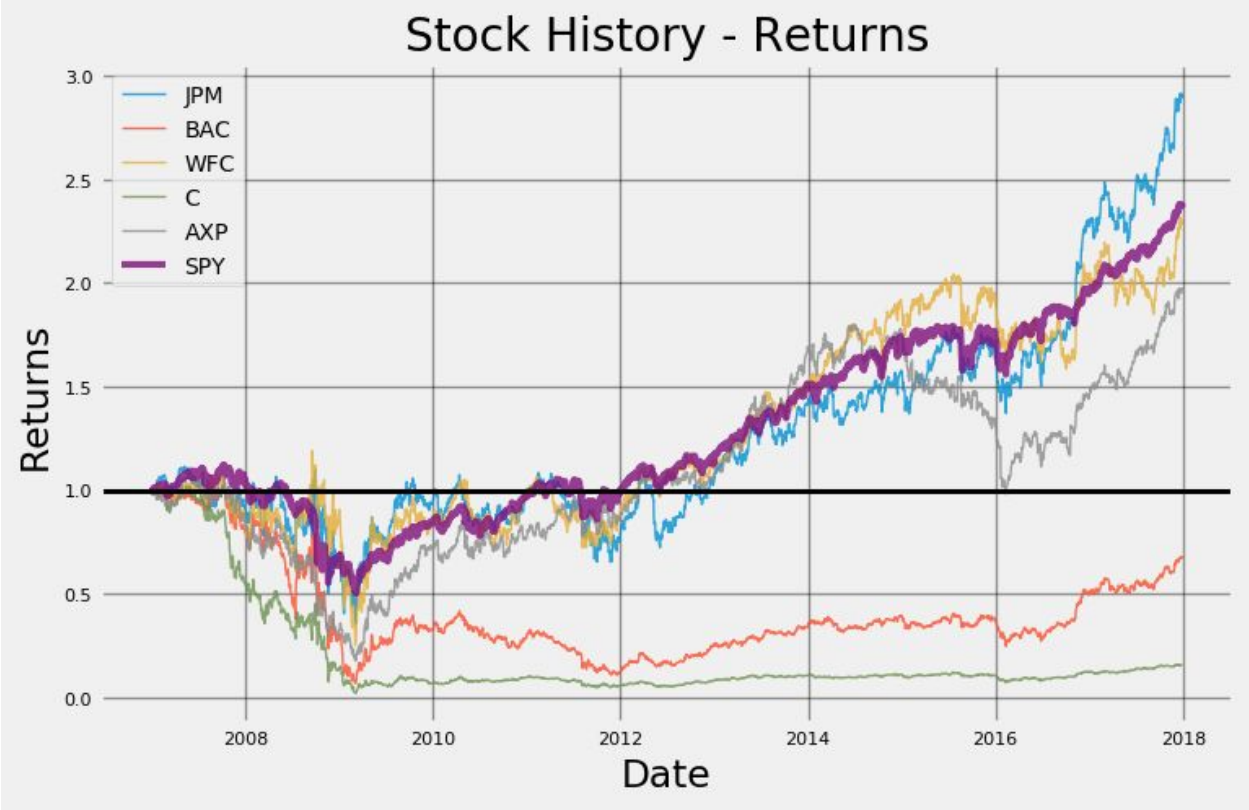
Capital Goods





Conclusion

The aim of this analysis was to observe trends between stocks of different industries. The SPY ticker has been added as benchmark. This is because SPY mimics the S&P 500, so essentially it becomes an indicator for the market. Without even having to do a google search, it becomes apparent when the crash occurred and how it affected every industry. Without fail, every industry's volatility shot up in 2009 due to the stock market crash. We can also see this by observing plots of the returns for each industry, as there is a dip in 2009. While we decided to focus on the volatility for each industry, it is interesting to show the returns graph for the financial industry. Some stocks have yet to recover from the crash, namely Bank of America and Citibank.



Looking back to the volatility graphs, these numbers decreased as time went on, which is more apparent in the second date range. Note that the volatilities of some industries were larger than others. Technology, Consumer Non-Durables, and Capital Goods all had relatively low volatilities while Finance and (surprisingly) Healthcare had larger volatilities. Using SPY as our benchmark for safe stocks that you can rely on in the years to come, investing in the industries of Technology, Consumer Non-Durables, and Capital goods seem would all be sound investments. An improvement to better solidify this claim would be to look at the over fifty stocks for each industry and compare them in the same manner as illustrated above. We could also

look at the average volatility of each industry over time and plot them together, in order better illustrate this point.

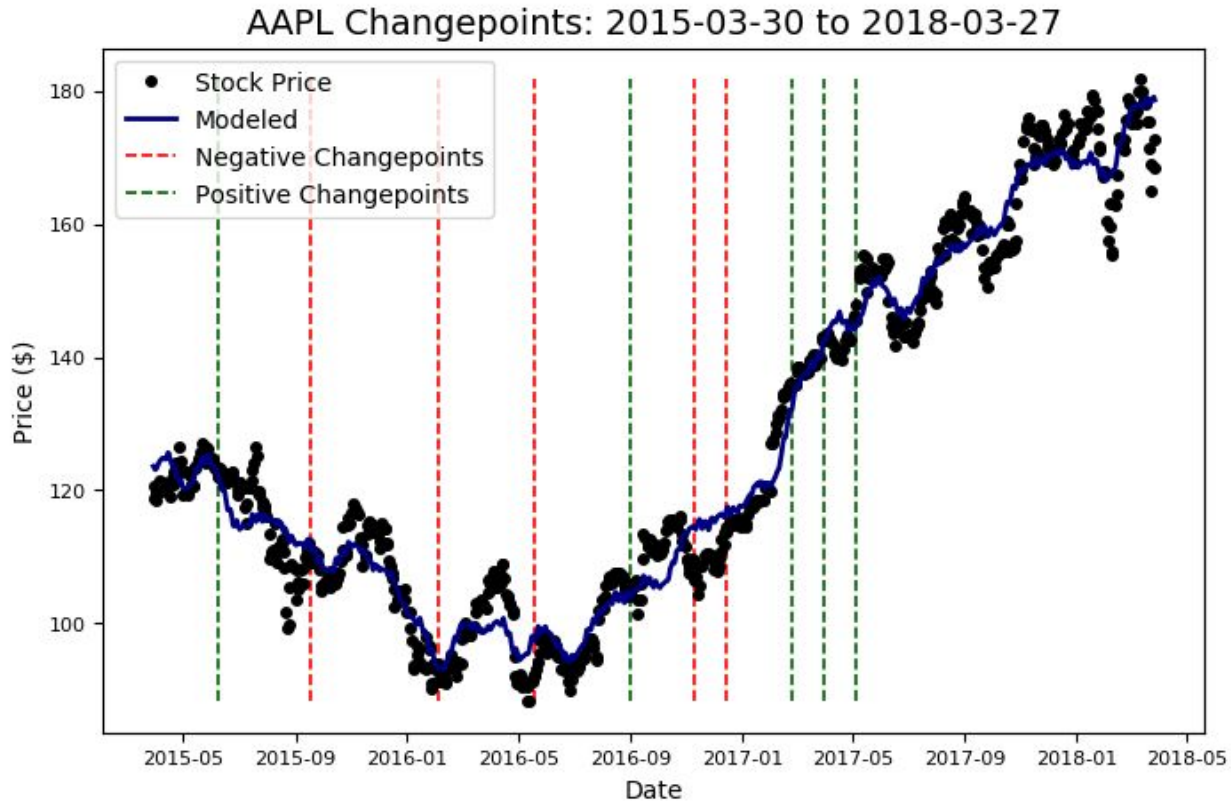
Experiment 2: Changepoint Analysis using Stocker

Methodology

Changepoints are points in time-series data that hint at potential trend shifts in the graph in the near future. They are found by looking at the concavity of every point in the graph. The larger the numerical value of the concavity, the stronger the potential shift will be. A positive concavity is correlated with a negative shift in the graph, and vice versa.

We analyzed companies from 8 different industries based on the changepoints of their stock prices. The industries we chose to look at are as follows: Tech, Finance, Energy, Healthcare, Capital Goods, Consumer Services, Consumer Non-Durable, and Public Utilities.

Our set-up is as follows. From each industry, we chose 5 companies listed among the top 10 in market share. We plotted their stock prices using stocker, and analyzed the graphs to find the 10 largest change-points from March 30th, 2015 up until August 31, 2017, sorted by the exact date and whether the changepoint was positive or negative. We listed all the dates for a given industry, and took note of any dates that were common changepoints for at least 3 out of the 5 companies. We will call these dates “shared changepoints.”



This is a graph of the stock price of Apple, as well as a simple linear model, and the 10 largest changepoints based off that model. Positive changepoints are marked by dashed green vertical lines, while negative changepoints are marked by dashed red vertical lines. Here is the list of the 10 changepoints from the previous graph, sorted by their exact date.

	Date	Adj. Close	delta
48	2015-06-08	122.392291	-0.622993
120	2015-09-18	109.140225	0.058418
216	2016-02-05	91.326278	0.331918
288	2016-05-19	92.060491	0.772369
361	2016-09-01	104.867483	-0.157710
409	2016-11-09	109.510470	0.163036
433	2016-12-14	113.767235	1.063407
481	2017-02-24	135.553035	-0.337295
505	2017-03-30	142.764147	-0.272297
529	2017-05-04	145.343086	-0.425737

For every shared changepoint, we first looked to see if it showed an entirely positive trend shift, an entirely negative trend shift, or a mixed trend shift. We then looked to see what notable news events happened on that particular date. The archive of the Wall Street Journal was very helpful in this regard. We looked for specific keywords related to the industry, as well as general keywords “market”, “stocks”, “growth”, and “economy”.

We analyze our results to see what kind of events cause shared changepoints to happen, and which direction, positive or negative, the changepoint points to. We will also compare across industries to see how many shared changepoints each industry has, whether any of these dates are shared among industries, and how do the trends compare for similar dates and events.

We used a modified version of “Stocker” for this analysis. We modified the code to allow us to load multiple companies at once. We also implemented QoL changes, fixing the title of each graph and sorting the data for easier processing.

Results

First, a summary analysis. The Finance and Non-Durables industries had the most shared changepoints with 10, while Capital Goods had the least with 5. The most common date for changepoints was February 8, 2016, which was a changepoint for every industry in our analysis, and a down trend for 7 out of 8 industries. Finance had the most positive trend shared changepoints with 7, Tech had the most negative with 4, and Healthcare had the most mixed with 3.

Industry Analysis

- Tech
    - The tech industry was unique in that company stocks moved in reaction to other companies in the industry
    - The election of Donald Trump was a big negative, accounting for 2 changepoints right around his election
    - The tech industry's movements seemed to be independent of any other in our analysis
- Finance
    - The finance industry was entirely in sync, always trending either positive or negative, and never mixed
    - Seemed to be entirely dependent on foreign markets and currency valuation, and was not affected at all by politics
    - Movement seems to be correlated with Energy
- Energy
    - Energy companies had a high number of shared changepoints with 9, suggesting that they're heavily connected with each other
    - The price of oil was heavily correlated with stock price and trends, accounting for 4 out of 9 changepoints
    - As such, foreign markets have an affect on the industry when oil is involved
- Capital Goods
    - Capital Goods had the least shared changepoints with 5, suggesting that the companies' stocks trend independently of one another
    - This may be due to the fact that this is a very broad and diverse industry tag
- Healthcare
    - The Healthcare industry had 8 shared changepoints, which suggests some level of connection within the industry
    - Very dependent on politics, with no global market influence
- Customer Services
    - This industry had 7 shared changepoints
    - This industry was unique in that it seemed to move entirely against a single company, Amazon. When Amazon trended one way, oftentimes the entire industry trended the opposite. Even for changepoints with only 2 companies, it was usually Amazon trending one way with another company trending the opposite way.
- Non-Durables
    - Like Finance, Non-Durables had 10 shared changepoints. With only 1 mixed changepoint, it seems that this industry also moves in sync.
    - The election of President Trump hurt the industry
    - Events in the U.K. (Brexit, the following election) favored the industry
- Public Utilities
    - The industry had 8 shared changepoints.

- It is unique in that it seemed to move counter to other industries on the same dates (February 8, 2016 and December 14, 2016).

Conclusion

In conclusion we feel that we were able to lay down the groundwork for a potential model to predict potential trends in stocks based on industry. This model would look at top news stories and events each day, and use specific criteria, tuned to each industry, to filter for news that could potentially be the signal for a changepoint.

We understand that there are some limitations in our experiment. First, we could only pull the data for companies based in the U.S. While the majority of companies are based in the U.S., we missed out on some we'd have liked to look at. Second, there are more industries that can be examined still. Third, while we believe picking 5 among the top 10 companies in market share per industry was enough to generate sufficient data for analysis, there's a lot of room to expand on this, by increasing the number of companies and/or widening the range of companies chosen.

Choosing to define a shared changepoint as having 3 out of 5 common companies also merits some debate. We chose this because of the previous limitation of only 5 companies per industry, but perhaps a higher percentage is better to allow for more robust changepoints. Ultimately, it's a trade-off; a higher percentage allows for stronger conclusions to be made, but potentially misses out on interesting trends that might affect a (large) part of the industry.

Something we found interesting is that the stock trends hinted at by changepoints were oftentimes very localized, and were quickly overtaken by a larger trend affecting the entire market. Accounting for this in future analysis, and building a predictive model, will prove to be a difficult task.