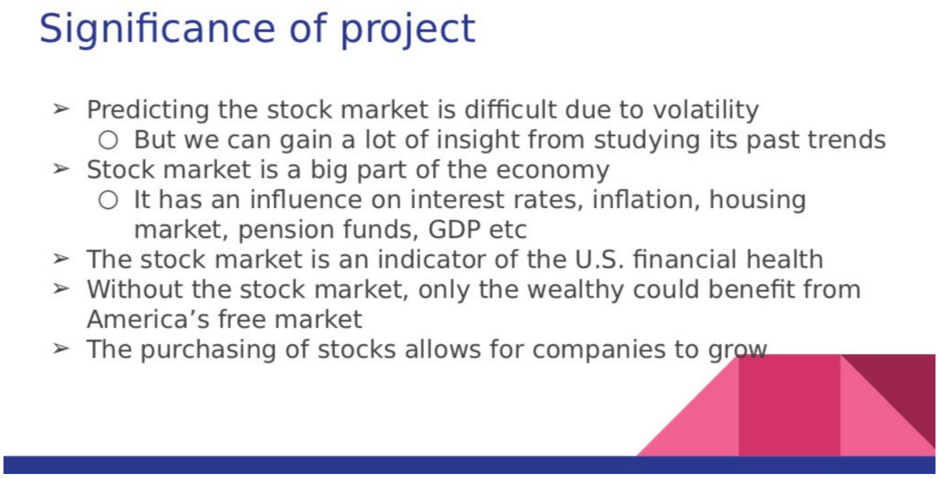
Final Project Report

By: Abraham Carreon, Abhishek Wagh, Pradeep Kumar

Our final project consisted of three existing datasets from Kaggle.com "https://www.kaggle.com/aaron7sun/stocknews/data." The name of each dataset is as follows, "Combined_News_DJIA.csv", "DJIA_table.csv", and "RedditNews.csv". These three datasets are gathered in one overall folder called "Daily News for Stock Market Predictions." The Reddit News dataset comes from historical news headlines from the reddit channel called "Reddit WorldNews Channel" from 2008-2016. Whereas, the stock dataset focuses on stocks from the Dow Jones Industrial Average from 2008-2016. The premise of our project revolves around attempting to identify various price trends within the Dow Jones. In addition, we also took into account news reports from Reddit due to the fact that the price of stocks can be indirectly affected by current events. We were able to implement three different softwares throughout our project such as Tableau, Power BI, and Watson analytics. We also developed graphs using R studio in the latter part of our project.

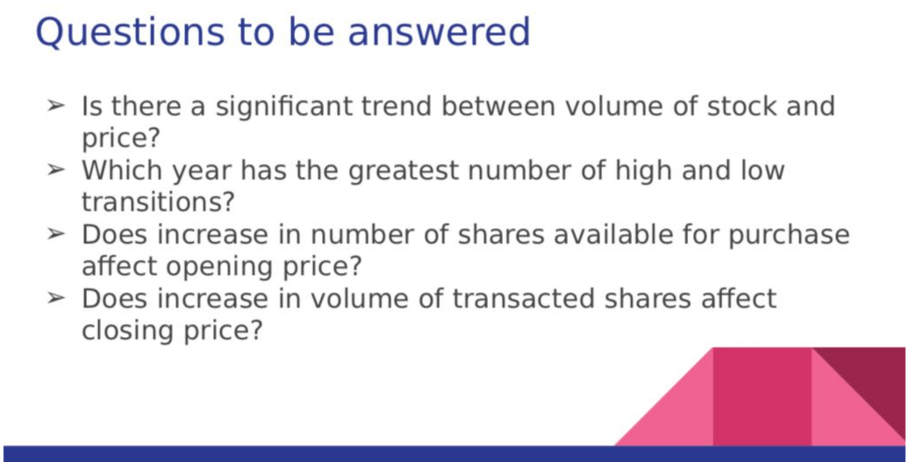
Our motive behind this project was based off of our self interest within the stock market. We decided it was only right to create a project of which we were passionate about because it made the process enjoyable. In addition, we were able to collect robust datasets of which made our analysis and implementation easier.



Significance of project

- Predicting the stock market is difficult due to volatility
  - But we can gain a lot of insight from studying its past trends
- Stock market is a big part of the economy
  - It has an influence on interest rates, inflation, housing market, pension funds, GDP etc
- The stock market is an indicator of the U.S. financial health
- Without the stock market, only the wealthy could benefit from America's free market
- The purchasing of stocks allows for companies to grow

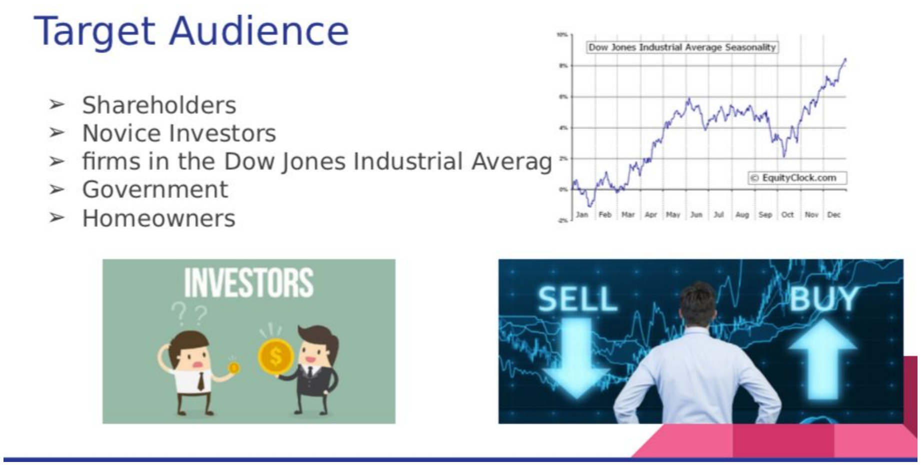
The significance behind our project was quite obvious from the beginning. We felt that our project was impactful for the sole fact that every American is affected by the performance of the stock market. For instance, the stock market's influence within the economy has a lasting effect in the long-run as well as the short-run. In the long run interest rates, inflation, and the housing market all can negatively or positively affect the general population, contingent on its performance. Whereas, the short-run doesn't necessarily have that much of an impact, however we are able to determine the health of the U.S. financial system based off of factors such as Gross domestic product and our national debt. All in all, we accepted the challenge of taking on this project knowing that predicting stock prices are difficult due to volatility. Volatility plays a role in stock prices due to the fact that it takes into account various different variables such as current events, stock volume, and fear to name a few.



Questions to be answered

- Is there a significant trend between volume of stock and price?
- Which year has the greatest number of high and low transitions?
- Does increase in number of shares available for purchase affect opening price?
- Does increase in volume of transacted shares affect closing price?

Before we started our project, we wanted it to be impactful enough to answer a few basic questions. As a result, we made a list of questions that we wanted to answer based on our results. Thus, were able to conclude that the significant trend between volume and price contains a trend line that steadily increases from 2008-2014 then drops off sharply in 2015-2016. In addition, we were also able to determine that 2009 contained the highest number of transactions, whereas 2014 contained the lowest. We were also able to determine in the following question that the number of shares available, increases the opening price of the stock. Lastly, an increase in volume will result in a higher closing price due to simply economics. Hence, an increase in supply results in an increase in price.



Target Audience

- Shareholders
- Novice Investors
- firms in the Dow Jones Industrial Averag
- Government
- Homeowners

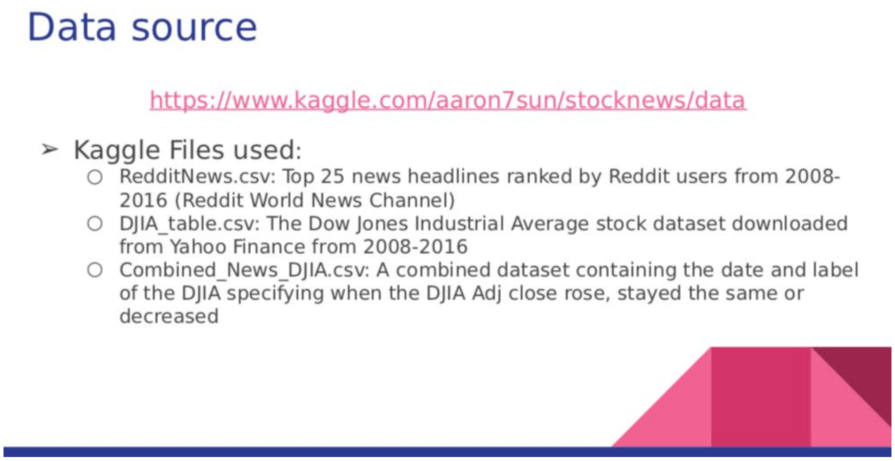
Dow Jones Industrial Average Seasonality
© EquityClock.com
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
INVESTORS
SELL
BUY

Our target audience consisted of groups of people that are not only directly affected by the stock market, but also indirectly affected as well. With that being said, shareholders as well as investors are directly affected by the stock market because the value in their portfolios are directly determined by stock price. On the other hand, companies are also measured by their given stock price. Thus, if Microsoft's stock for instance reached an all time high. We are able to conclude that Microsoft as a company is doing really well because their stock price reflects their companies performance. If we look at the government sector, they tend to regulate the stock market. A perfect example was apparent last month during the governmental corrections. The shutdown of the government resulted in chaos as the stock market crashed dramatically. Homeowners are more so indirectly affected by the stock market because as we saw in 2006-2009, the housing bubble also destroyed the stock market due to bad mortgages.

Tools Used

- ➢ Tableau Desktop
- ➢ Microsoft Power BI
- ➢ IBM Watson Analytics

We decided to approach our project with the utilization of three different softwares. With that being said, each of the three softwares had a range of advantages as well as disadvantages. As far as advantages go, Tableau had the most freedom out of the three softwares. To further elaborate, Tableau allowed me to create any graph possible using any variable of my choosing. This amount of freedom is unparalleled compared to the rest of the two softwares. The only disadvantage with Tableau in my opinion is that at times it seems quite difficult when importing certain files. Tableau loves csv and json files, however any other file type is a nightmare to import. On the other hand, Microsoft Power BI is a little more forgiving. One advantage of using Power BI is the fact that it is integrated with excel. Thus, our expertise in excel directly translated when using Power BI. This was evident by the ease of creating graphs in Power BI. Hence we did not run into many issues at all. Although, one disadvantage would have to be that modification was at times bothersome. This was due to the different amount of options that it offers. The AI behind it is quite impressive, but it leads to more complexity. We decided to use Watson Analytics due to its intelligence. To further elaborate, all you have to do is import a dataset and it will take care of the rest. It creates graphs on its own of which requires little to no effort. However, a disadvantage that this carries is the fact that you cannot make modifications. This obviously a huge disadvantage, but we felt that the software itself still had a lot to offer especially coming from a different viewpoint.

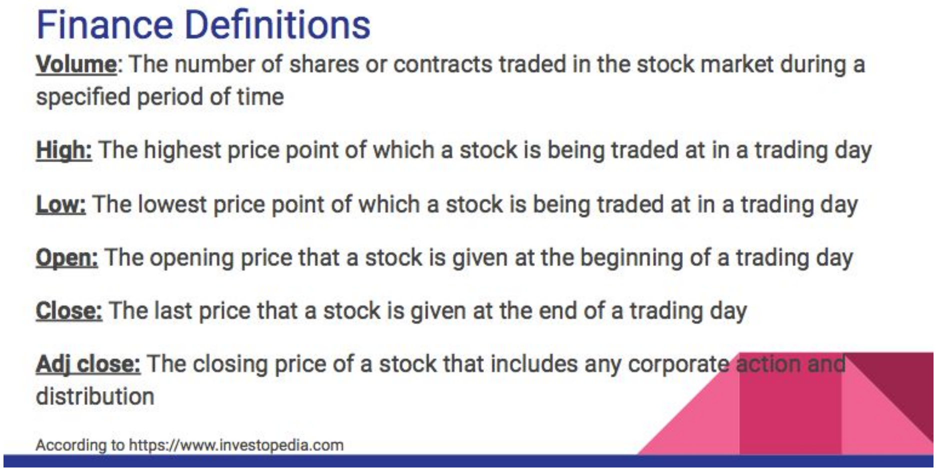


Data source

https://www.kaggle.com/aaron7sun/stocknews/data

➢ Kaggle Files used:
  ○ RedditNews.csv: Top 25 news headlines ranked by Reddit users from 2008-2016 (Reddit World News Channel)
  ○ DJIA_table.csv: The Dow Jones Industrial Average stock dataset downloaded from Yahoo Finance from 2008-2016
  ○ Combined_News_DJIA.csv: A combined dataset containing the date and label of the DJIA specifying when the DJIA Adj close rose, stayed the same or decreased

We preferred to use kaggle as our data source. Kaggle offers millions of different datasets and is quite reliable. As a result we knew that we could not go wrong by selecting it as our data source. We also preferred kaggle because it includes an in depth description of each data file. This is essential because most of the times people are able to find online data sets, however they do not include descriptions. This acts as a disadvantage because you can be misinformed about the dataset that you are trying to analyze.



Finance Definitions

Volume: The number of shares or contracts traded in the stock market during a specified period of time

High: The highest price point of which a stock is being traded at in a trading day

Low: The lowest price point of which a stock is being traded at in a trading day

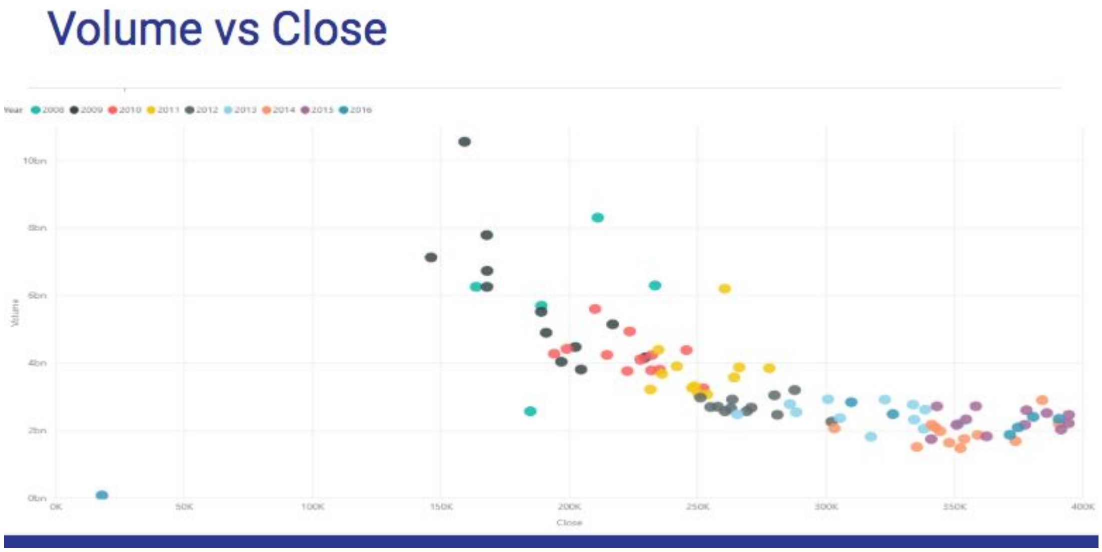
Open: The opening price that a stock is given at the beginning of a trading day

Close: The last price that a stock is given at the end of a trading day

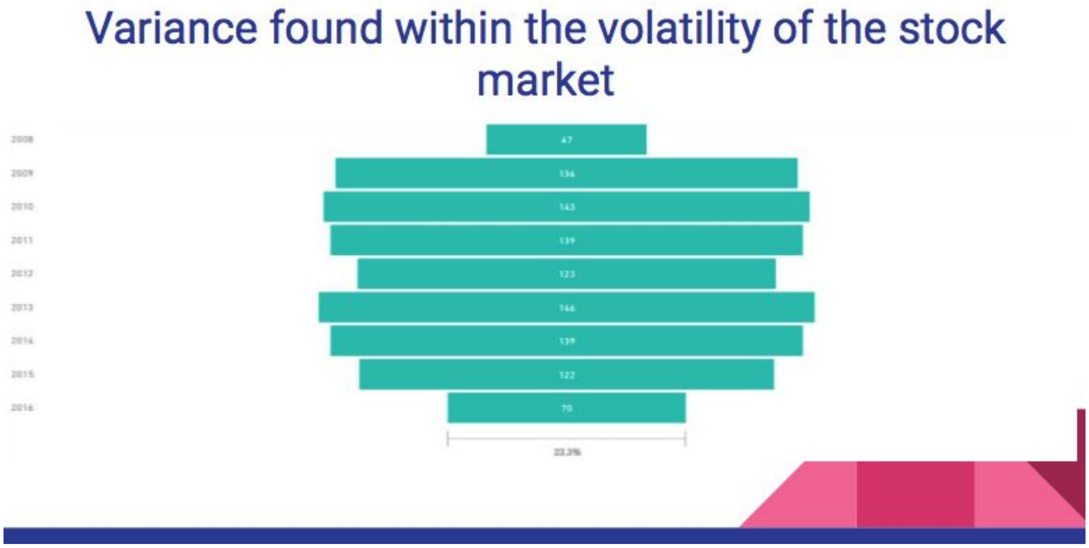
Adj close: The closing price of a stock that includes any corporate action and distribution

According to https://www.investopedia.com

We felt that it was necessary to include a list of financial definitions in our presentation due to the fact that not everyone is financially literate. Hence, this list was also provided in order for our audience to follow along during our presentation all while eliminating confusion. The most important definition that we wanted our viewers to understand was volume. Volume plays a vital role in stock price variation, and must be accounted for whenever one discusses stock price movement.

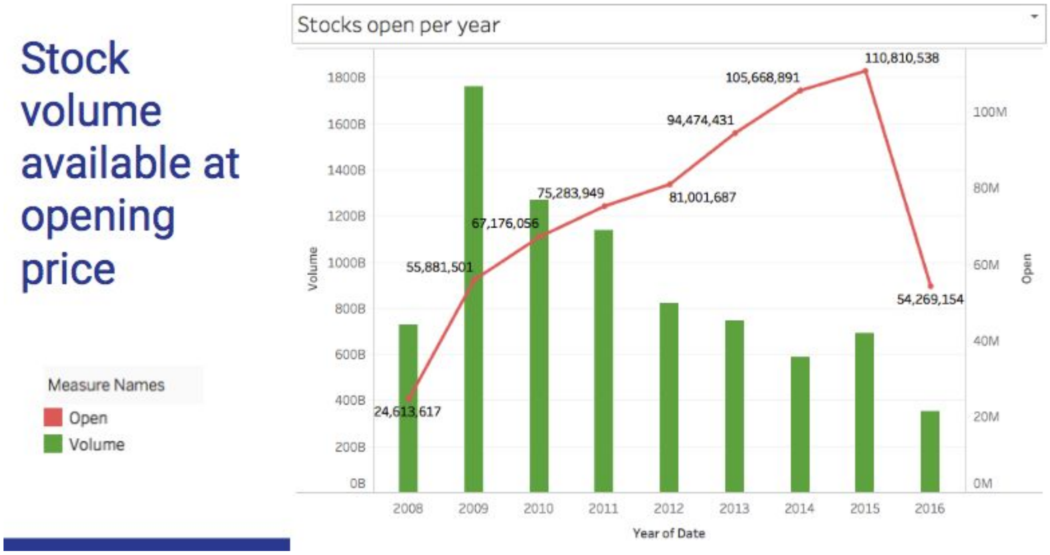


The figure was prepared using PowerBI. It shows that as the number of transactions increase in a unit time, the close price reduces. Thus, there is a negative correlation between volume and closing price. This is because the economy recovered in recent years from recession. High volumes and low closing means that there is a high interest in purchasing of stocks, however there is a lack of investors who are willing to purchase the stocks.



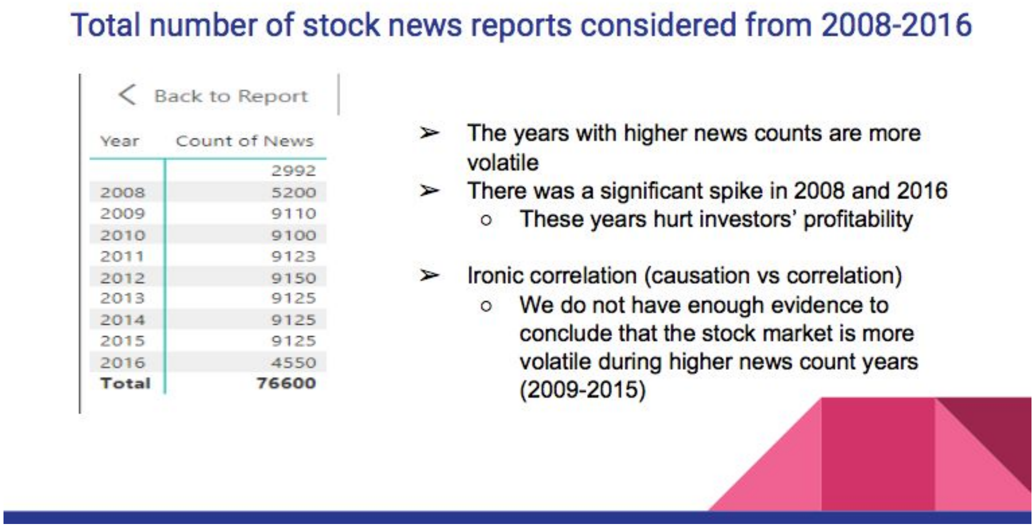
Variance found within the volatility of the stock market

The figure shows the degree of volatility in the stock market for every year from 2008 to 2016. Degree of volatility is important because it may be used to predict whether it may be the right time to buy stocks. The figure shows that 2008 and 2016 was when the stock market was the least volatile. This is because in 2008, there was a recession and the stock prices were always falling. In 2016, the federal bank increased interest rates due to falling oil and utility prices, thus the stock prices fell continuously. In 2013, the stock prices was most volatile because that was the year when the stock prices rose and fell the most number of times. The aggregate volatility in the stock market from 2008 to 2016 was 23.3%.



Stock volume available at opening price

The figure shows a strong negative correlation between opening price of stock and volume. In most years, as the volume of stock increases, the stock prices reduce. This is because the companies need to sell as many stocks as possible in order to get the money for investment. In 2009, the volume was relatively high because the companies were offering stocks at a lower price in the hope that the economy will bounce from recession. In 2016, the stock price was less even if the volume of stock sold reduced because Federal Reserve raised interest rates to lower energy and utility rates.



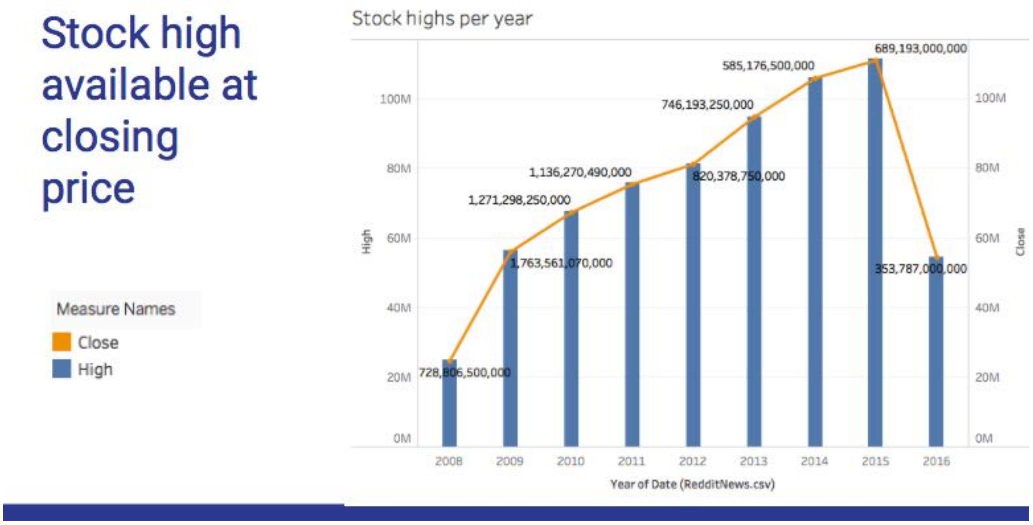
Total number of stock news reports considered from 2008-2016

< Back to Report

Year	Count of News
	2992
2008	5200
2009	9110
2010	9100
2011	9123
2012	9150
2013	9125
2014	9125
2015	9125
2016	4550
Total	**76600**

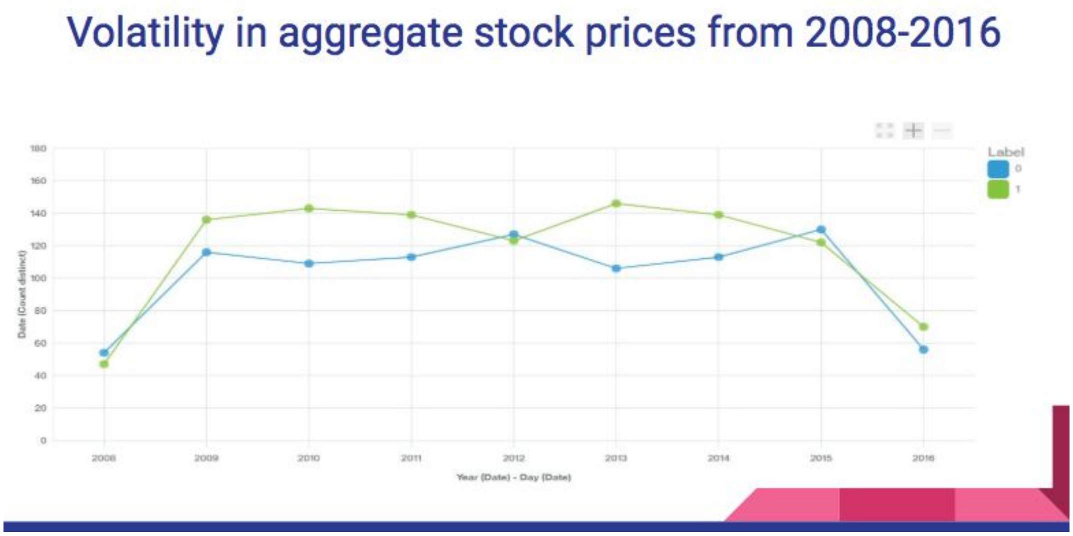
- ➢ The years with higher news counts are more volatile
- ➢ There was a significant spike in 2008 and 2016
  - ○ These years hurt investors' profitability
- ➢ Ironic correlation (causation vs correlation)
  - ○ We do not have enough evidence to conclude that the stock market is more volatile during higher news count years (2009-2015)

The years with more news reports contained stock prices that were more volatile compared to the years with less news reports. This is because in the years with more volatility in stock price, people tend to see the news more so as to decide when to buy the stocks. In years with lower volatility in stock price, people tend to see the news less since they already know that the stock price will reduce or increase continuously. Thus, in 2008 and 2016 the the count of news is less and also the stock price is less volatile.



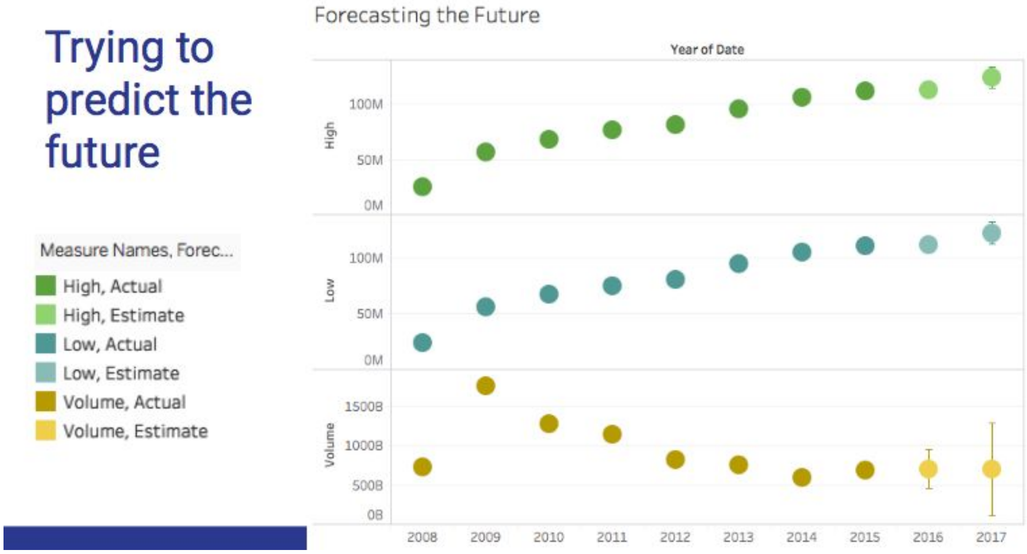
Stock high available at closing price

The graph in the figure shows that if an investor purchased stock in years from 2008 to 2010, most profit would have been made if the stocks purchased in those years were sold in 2015. You are able to grasp one of the most basic investing concepts from this graph alone which is the longer you hold on to a stock the more money you will make in the long run. This is a general rule of thumb of which is based on the economic history of the stock market. Thus, since the stock market is always expanding over time, you are more likely to be most profitable the longer you hold on to your stock versus selling away your stock in the short run.



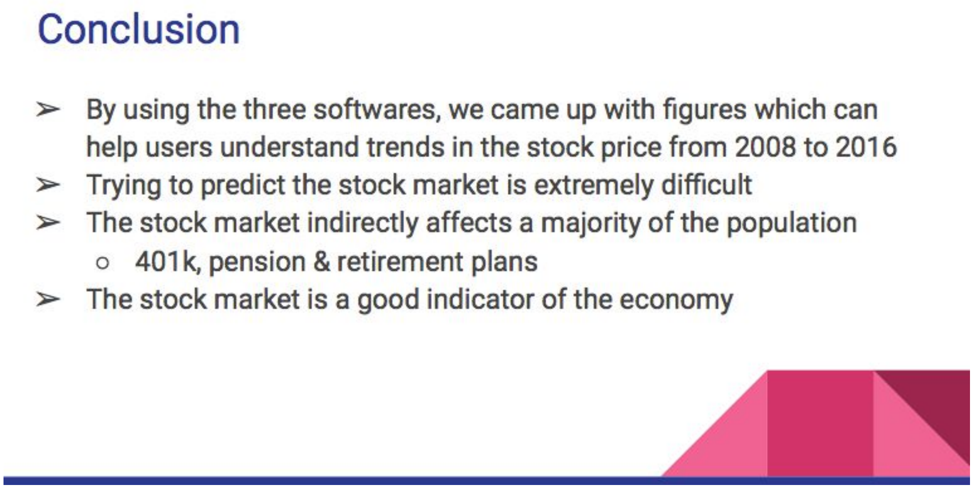
Volatility in aggregate stock prices from 2008-2016

This graph in the figure which has been prepared using IBM Watson Analytics can be compared to the volatility graph obtained using Microsoft PowerBI. This graph verifies that in 2008 and 2016, the stock prices were less volatile compared to other years. However in other years, the stock prices were volatile. In 2012, the stock prices were comparatively less volatile, which can be verified by the PowerBI graph. The period from 2009 to 2011 and 2013-2014 were risky time for investors to enter the stock market. The graphs show that investors can benefit if they buy the stock when the stock price is low during periods of volatility. The investor has to time when to buy the stock accurately to reap maximum benefit.



Trying to predict the future

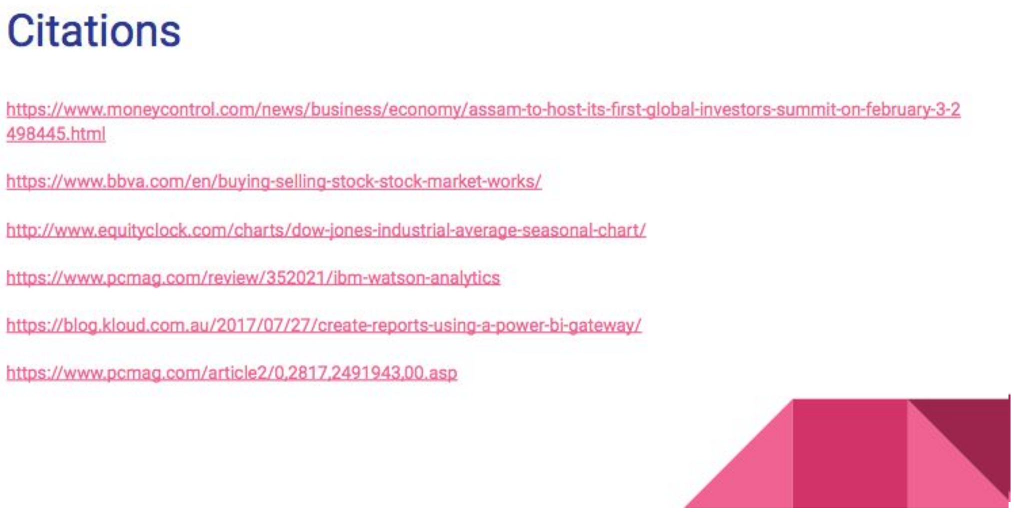
We decided to create a forecasting model through Tableau in order to attempt to predict the stock market within another year. It was not to our surprise that we were not able to predict the sevre dip in 2016. This is only proof that even sophisticated algorithms and machine learning cannot take into account volatility within the stock market. Although, the good thing about the stock market is that it is almost always guaranteed to rise in the long run. This is evident through the stock market's passed history. Hence, it has always been on the rise, and recovered from every recession.



Conclusion

- ➢ By using the three softwares, we came up with figures which can help users understand trends in the stock price from 2008 to 2016
- ➢ Trying to predict the stock market is extremely difficult
- ➢ The stock market indirectly affects a majority of the population
  - ○ 401k, pension & retirement plans
- ➢ The stock market is a good indicator of the economy

To wrap up, we took it upon ourselves to construct our project through the use of several different software programs, in order to gain the insight and perspective of different tools. We came to the conclusion that everyone should consider at some point in their lives attempt to invest into low risk investments that will benefit them in the long run. With a little bit of knowledge in the stock market, you can avoid common mistakes amongst investors such as trying to predict short run trends in the market. The average American does not take into account how much the stock market actually affects their financial health, however we wanted to bring awareness through our project. At the end of the day, choosing not to invest in the stock market is your ultimate choice. However, keeping an eye on it is adequate enough in order educate yourself as to how the economy is behaving.



Citations

https://www.moneycontrol.com/news/business/economy/assam-to-host-its-first-global-investors-summit-on-february-3-2498445.html

https://www.bbva.com/en/buying-selling-stock-stock-market-works/

http://www.equityclock.com/charts/dow-jones-industrial-average-seasonal-chart/

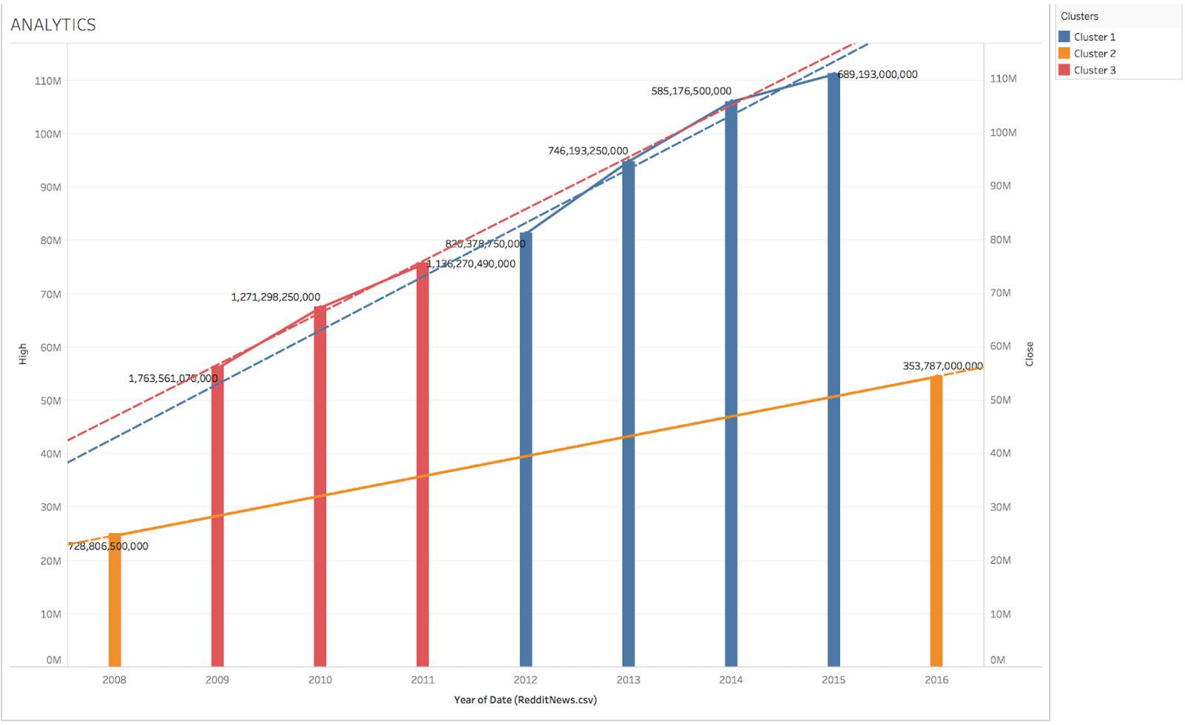
https://www.pcmag.com/review/352021/ibm-watson-analytics

https://blog.kloud.com.au/2017/07/27/create-reports-using-a-power-bi-gateway/

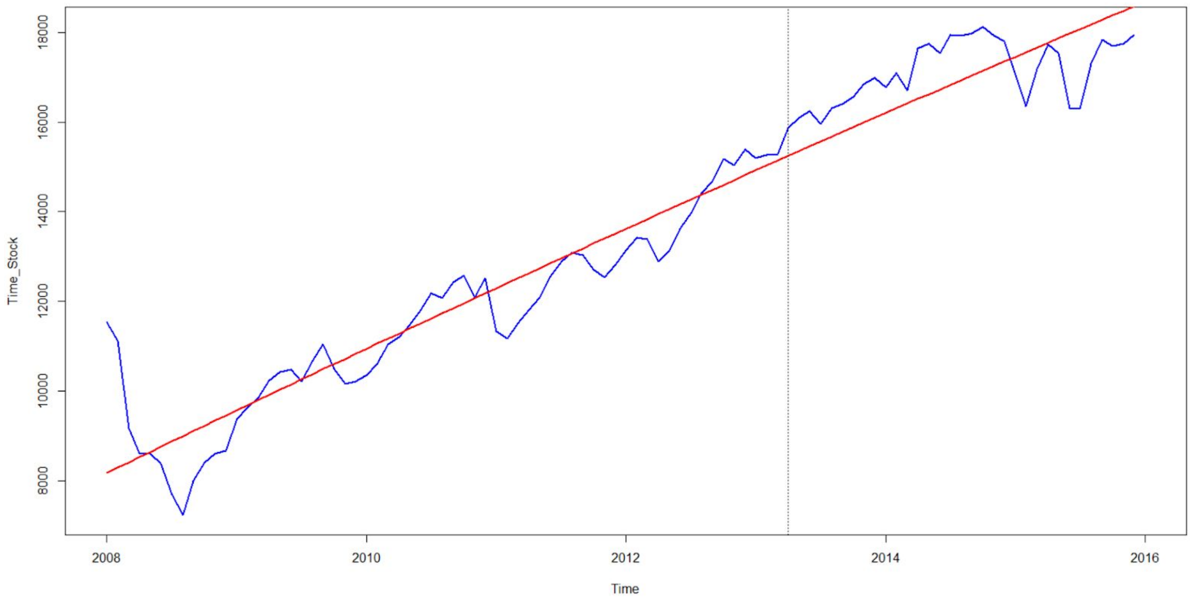
https://www.pcmag.com/article2/0,2817,2491943,00.asp

Extra:

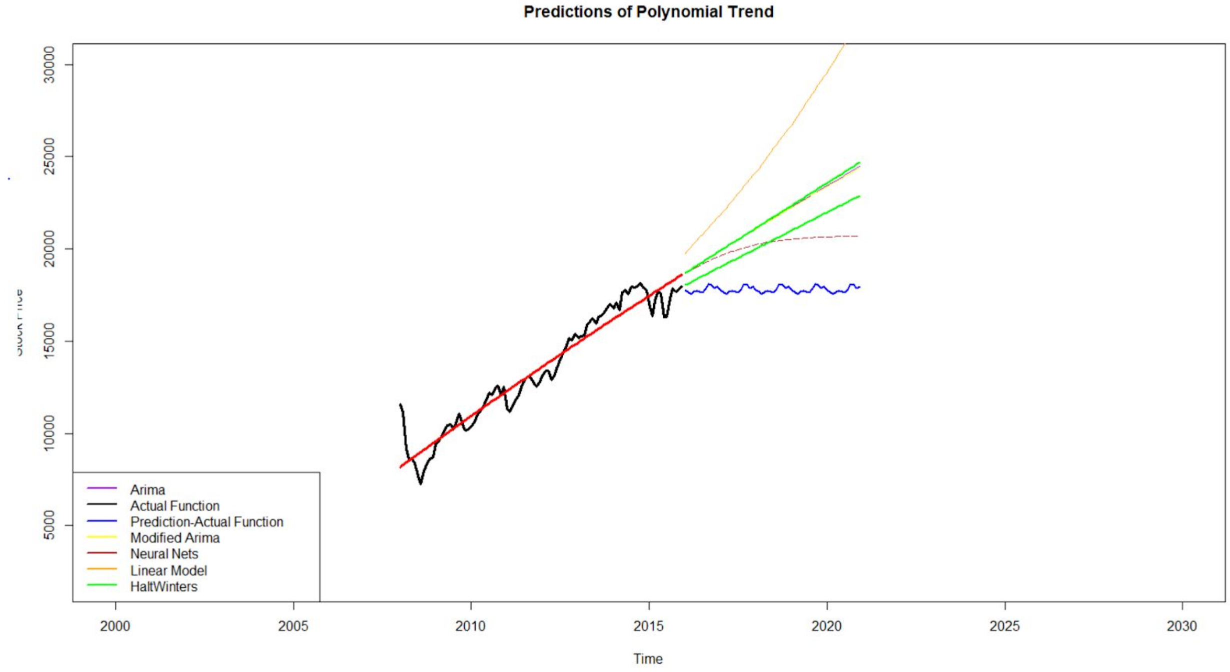
We took it upon ourselves as a group of three to make additional graphs as well as analyses with the use of R studio. We did not cover R in class as it is a quite technical programming language, however it is very useful. The following graphs were constructed after our in-class presentation was given, thus they were not in our original power point slide.



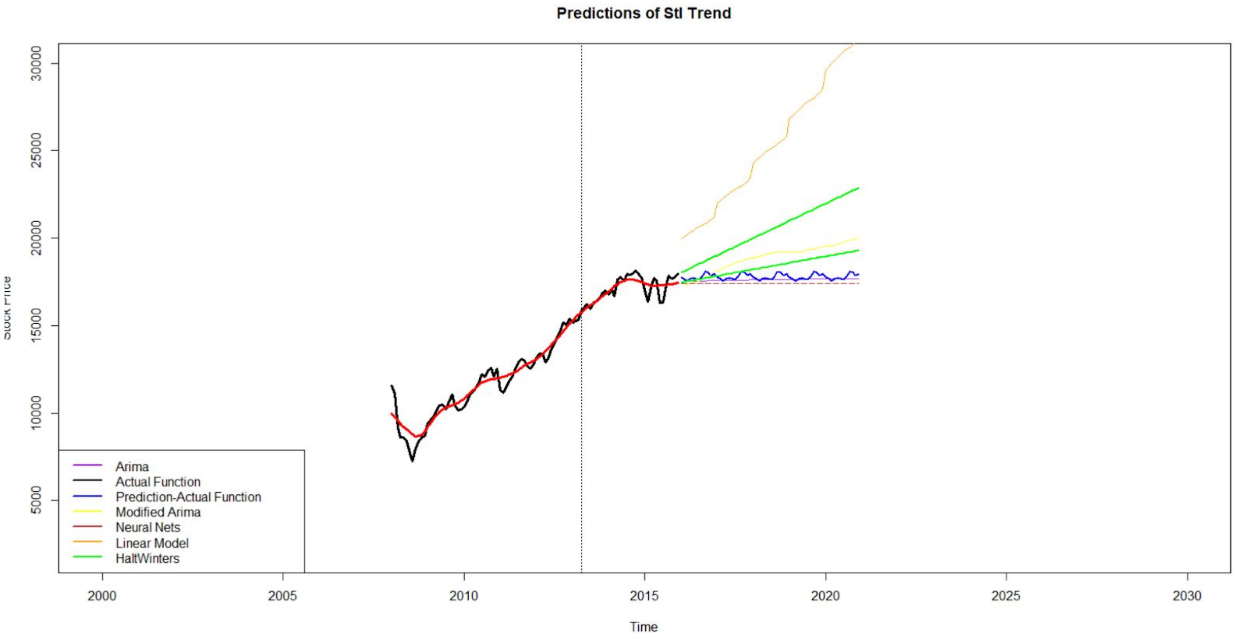
This graph was created through the use of Tableau, however it is still a very in depth analysis. With that being said, it took into account the stock price high as well as the stock price close. It is important to note, that the graph was broken up into three different clusters. The first cluster is made of years 2008 and 2016. Meanwhile, the second cluster is made up of years 2008-2011, while the last cluster is made up of years 2012-2015. Each cluster has its given trend line that foresees the average closing stock price date. We are able to conclude that the second cluster contains a higher degree of slope within its trend line meaning the average closing costs across the 2008-2011 span were at its highest point. This means the profitability across 2008-2011 were at its highest point assuming investors were cashing out during this period.

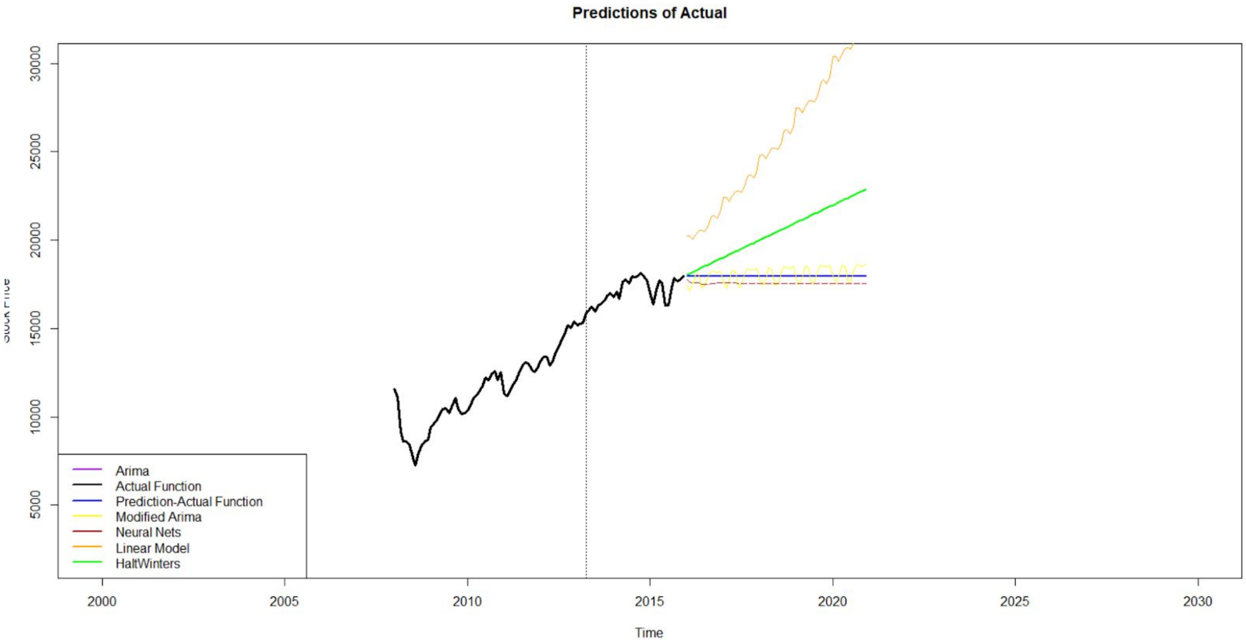


The following graph was coded through the use of R. It is a Time Series trend where we are able to see the positively upward trend. A perfect market should always consists of a positive upward trend because as history shows, the American economy has always been expanding since its existence of time. The trend line is essentially the average stock price across the span of 2008-2016. As you can see there are multiple places where the price beats the trend line. This is the most ideal time to sell any stock for profits. That is only true of course if we implement the “buy low sell high” investing strategy. On the other hand, if stock prices are below the trend line investors are better off holding onto their investments until it surpasses the trend line in order to prevent themselves from losing profits.



The following graph was also implemented through the use of R. This graph is a Stl Analysis with a trend line. In other words this graph contains the previous trend graph with an addition of a polynomial trend. The green lines within the graph contain a continuation of the time series trend. It is not a surprise to me that it is still upward trending because that is what we can assume. However, the important thing to note from the graph is the blue line that is staying constant. We can conclude from this forecast a stagnant market where the market is essentially paused. This is a rare instance in which we must not waste our time on because natural occurrences as well as volatility would not allow for this to happen. The outlying trend (orange line) contains exponential growth. Not only would this trend be profitable for everyone, but it is an extremely bullish trend in the market that would not continue for long. With that being said, the green lines contain the most predictable future if we were to forecast the market up until the year 2020.





The following two graphs are Forecast graphs using polynomial trends. These trends are similar to the previous graph, however they tend to be more accurate. The first graph contains a red trend that eliminates a majority of the peaks and troughs. This is necessary in order to get rid of the unnecessary volatility when trying to forecast the future. As previously explained, the green line is the line to pay most attention to because it captures the possible range of the trend when looking at the following years. However, a major difference between these two graphs and the previous one is the orange line. In the previous graphs it did not take into account volatility, whereas in the current two graphs it has. This is apparent by the amount of movement captured in the orange graph.