A glimpse into the ups & downs of the S&P 500

Rohit Kumar & Channing Vernon

11/12/2019

Introduction

Have you wondered about how the stock market can have a positive or negative impact on our society as a whole? The stock market signifies the change in unemployment, interest rates, and economic growth. With that being said Rohit and Channing realize similarities and compare events that occur around the world from years 2013 - 2018 and how they impact stock prices and the market as a whole. We also look at how the market rises and falls with respect to itself, and if there is is a sole, major driving power that determines a change for most of the stock exchange in the S & P 500 index.

In order to address the problems at hand, we choose to use a data set found on Kaggle. Here we looked at a couple of situations that might have impacted the S&P 500 index, using summary statistics to show light on the general picture of what we are working with and then moving on to more robust analysis via hypothesis testing using t-tests.

To approach the problem we first use simple yet powerful graphing techniques to show the spread of the data. This will easily show the changes that have occurred over the 5-year span. With the help of text accompanying these graphs we can further emphasize what is truly going on here.

Installing packages required

```
library(tidyverse)
library(dplyr)
library(broom)
## We first use the "tidyverse" package which contains a variety of packages that help us show data thr
```

Loading Data + Cleaning

The data set was found from the following link on Kaggle: Source

The data was collected by Dr. Cam Nugent, Ph.D. Candidate, Bioinformatics at University of Guelph, Guelph, Ontario, Canada. He collected the data via a script that acquired `.csv` files of specified stock tickers which all together form the S&P 500 index. The script which was used can be found in this GitHub repository. The data was last updated/collected in February 2018. The script simply compiled data from the past 5 years at that point and compartmentalized the individual stocks with their respective observations. One peculiarity of the data was that as the S&P 500 index changed over time, some stocks were removed from the set in order to show a more true picture. This results in missing values throughout the data set which need to be fixed.

We simply downloaded the data from Kaggle into a new folder and imported the data into RStudio straight from there. Then we ran a couple of summary statistics to see exactly what we are working with. We notice a lot of `NA's` included in some variables, and that the date is not formatted correctly. In the str table we can notice that the date variable is a factor variable instead of a date.

```r
data <- read.csv("all_stocks_5yr.csv") # Loading the dataset

summary(data) # Showing basic summary statistics
```

```
##          date             open              high              low         
##  2017-12-05:   505   Min.   :   1.62   Min.   :   1.69   Min.   :   1.50  
##  2017-12-06:   505   1st Qu.:  40.22   1st Qu.:  40.62   1st Qu.:  39.83  
##  2017-12-07:   505   Median :  62.59   Median :  63.15   Median :  62.02  
##  2017-12-08:   505   Mean   :  83.02   Mean   :  83.78   Mean   :  82.26  
##  2017-12-11:   505   3rd Qu.:  94.37   3rd Qu.:  95.18   3rd Qu.:  93.54  
##  2017-12-12:   505   Max.   :2044.00   Max.   :2067.99   Max.   :2035.11  
##  (Other)   :616010   NA's   :11        NA's   :8         NA's   :8        
##      close              volume             Name       
##  Min.   :   1.59   Min.   :        0   A      :  1259  
##  1st Qu.:  40.24   1st Qu.:  1070320   AAL    :  1259  
##  Median :  62.62   Median :  2082094   AAP    :  1259  
##  Mean   :  83.04   Mean   :  4321823   AAPL   :  1259  
##  3rd Qu.:  94.41   3rd Qu.:  4284509   ABBV   :  1259  
##  Max.   :2049.00   Max.   :618237630   ABC    :  1259  
##                                        (Other):611486
```

```r
str(data) # Showing more information about the variables
```

```
## 'data.frame':    619040 obs. of  7 variables:
##  $ date  : Factor w/ 1259 levels "2013-02-08","2013-02-11",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ open  : num  15.1 14.9 14.4 14.3 14.9 ...
##  $ high  : num  15.1 15 14.5 14.9 15 ...
##  $ low   : num  14.6 14.3 14.1 14.2 13.2 ...
##  $ close : num  14.8 14.5 14.3 14.7 14 ...
##  $ volume: int  8407500 8882000 8126000 10259500 31879900 15628000 11354400 14725200 11922100 607140
##  $ Name  : Factor w/ 505 levels "A","AAL","AAP",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```r
SP_data <- data %>%
  select(Name,date,open,low,high,close, volume)
# Changing the order of the columns for aesthetic purposes
```

Ommiting `NA` rows vs replacing them with `0`

We choose to omit rows with `NA` values in order to make calculations error-free, vs setting them to 0.

```r
SP_data <- na.omit(SP_data) # Omitting all rows with NA values

## SP_data[is.na(SP_data)] <- 0

# Example of what setting NA values to 0 looks like

summary(SP_data)
```

Changing date column from a factor to date

Next, we make a new `Date` column and format the date column into Y-M-D, making it a more workable and usable column. After we have created this new `Date` column we delete the previous one.

```r
SP_data$Date <- as.Date(SP_data$date, format = "%Y-%m-%d")
# Selecting date variable and changing it into the date format

SP_data$date <- NULL # Deleting original date column

SP_data %>%
  select(Name, close, volume, Date) %>% # Selecting 3 useful variables and a name column
  head(5) # Returning the first 5 rows
```

```
##   Name close   volume       Date
## 1  AAL 14.75  8407500 2013-02-08
## 2  AAL 14.46  8882000 2013-02-11
## 3  AAL 14.27  8126000 2013-02-12
## 4  AAL 14.66 10259500 2013-02-13
## 5  AAL 13.99 31879900 2013-02-14
```

Above we can see the first 5 rows of the data set showing the three main variables we will be looking closely at; close (price), volume (stocks bought/sold) and the date.

- `close`
  - The closing price will provide us with exactly where the respective stock's value lies at the end of the day.
- `volume`
  - The volume represents how much of the stock changed hands, via buying and selling.
- `Date`
  - The date will show us exactly what point in time the stock was reacting to change(s).

Top S&P 500 Stocks

The S&P 500 index is composed of 10 different sectors in different industries, with Information Technology sitting with the highest perccentage. The top 5 best performing stocks among the respective sectors are compared below with the source linked at the bottom.

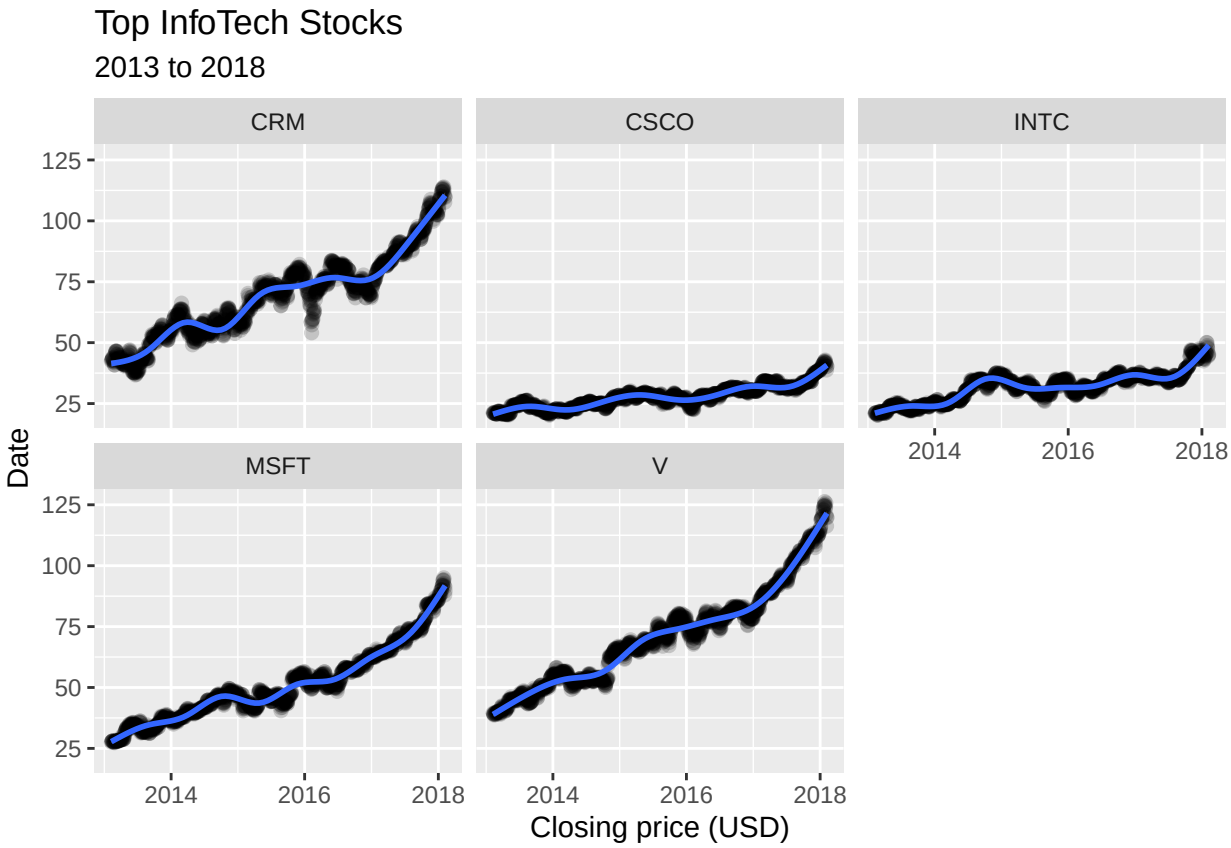
- S&P500 Sectors
- Information Technology
- Health Care
- Financials
- Communications Services
- Industrials
- Consumer Staples
- Energy
- Utilities
- Real Estate
- Materials

We take a quick look at these sectors in more detail, with percentage of sector from the S&P 500 written alongside

Information Technology - 19.9%

- CRM
- MFST
- CSCO
- V
- INTC

Source

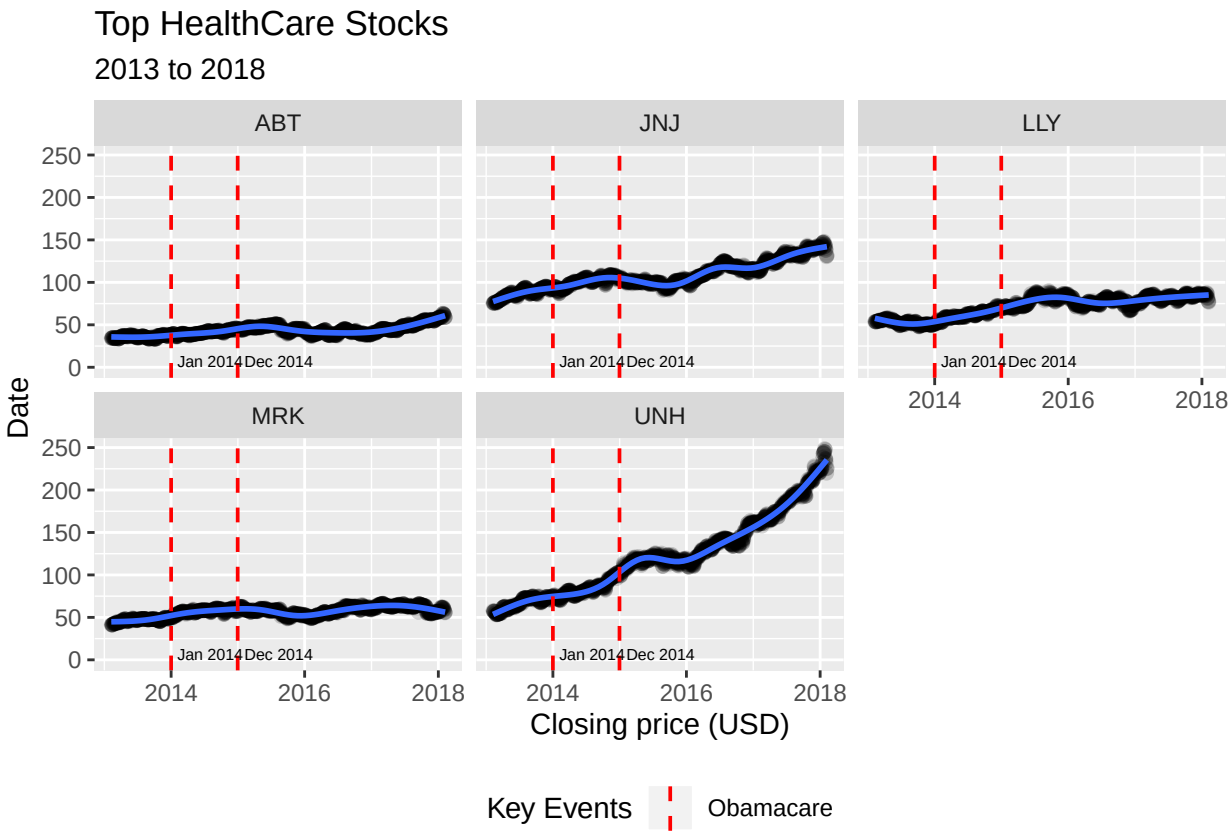


The top information technology stocks showed overall positive gains, with Salesforce (CRM) showing interesting dips with a respectable recovery. Cisco and Intel showed slow but sure improvement over the 5-year span. Visa and Microsoft, on the other hand, showed very steady positive growth. Compared together these top stocks seem to have a mind of their own when it comes to rising in gains or lowering.

Health Care - 15.8%

- JNJ
- UNH
- MRK
- ABT
- LLY

Source

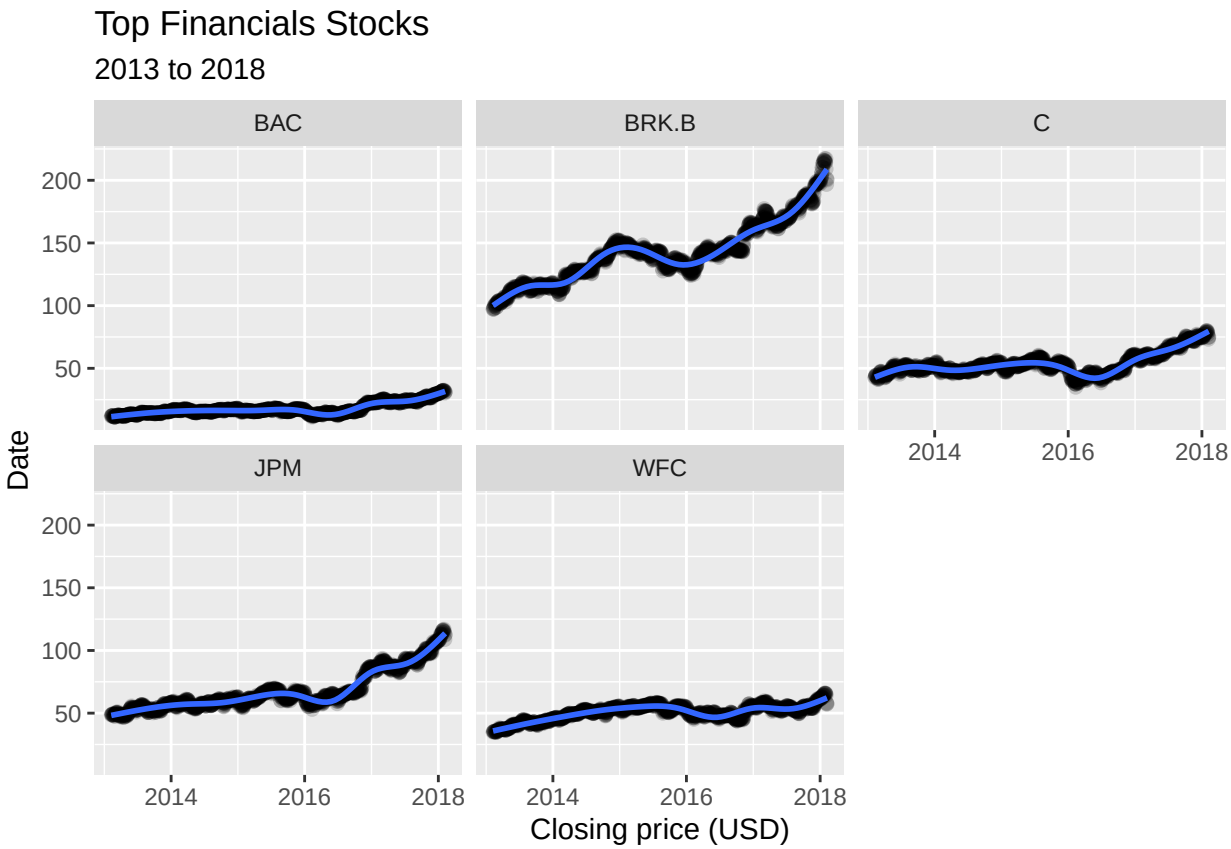


Although the top stocks in health care show near break-even gains, it is interesting to note UnitedHealth Group Inc's amazing performance. In 2014, former President Obama had signed to add $20 million in healthcare through his "Obamacare". However, by looking at the stocks within the marked range, there seems to be little to no effect in the year 2014 to the outside influence of "Obamacare".

Financials - 13.7%

- BRK.B
- JPM
- BAC
- WFC
- C

Source

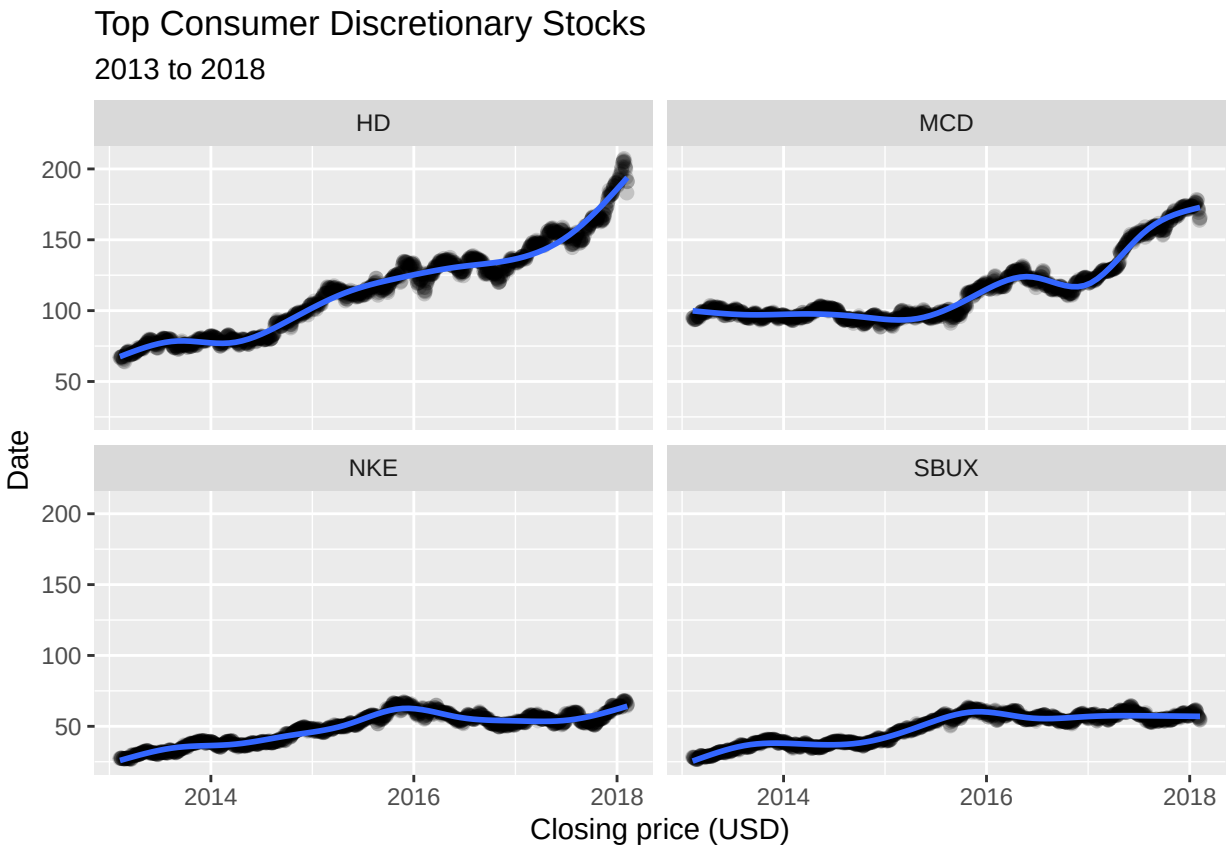


Among the top financials stocks, Warren Buffet's Berkshire Hathaway Inc. has seen some great gains. The rest of the leading financials stocks come in at very low gains compared to BRK.B. These stocks seem to have low gains except for BRK.B, but do so with no relation to one another.

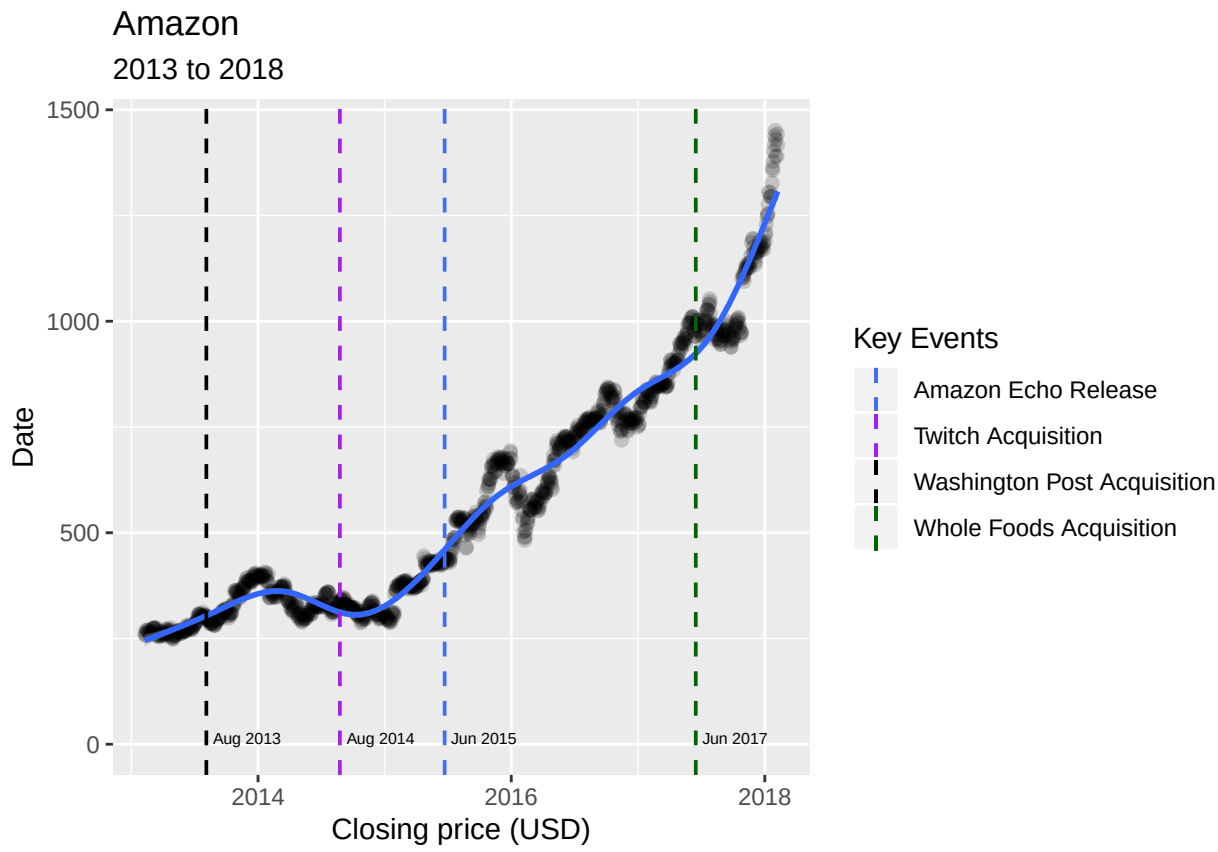
Consumer Discretionary - 9.9%

- AMZN
- HD
- MCD
- NKE
- SBUX

Source



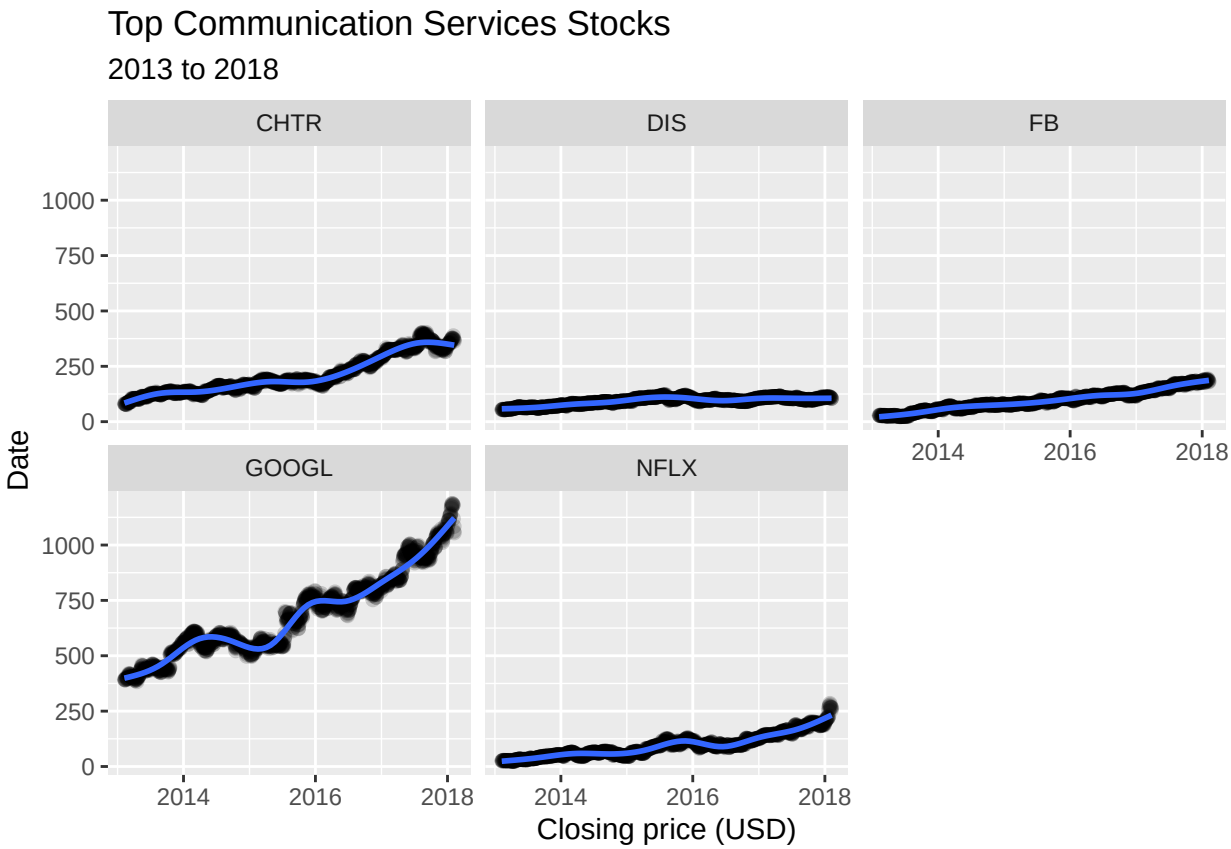
After taking out Amazon (below) out of the picture, for solely blowing every other stock out. Amazon displays extremely explosive growth and continues to do so, which, when comparing to the other throws off the data tremendously. Home depot and Mcdonalds show decent gains, whereas Nike and Starbucks seem to struggle during these years. All in all, there doesn't seem to be a clear force driving all of these stocks up or down.



Communications Services - 9.9%

- Facebook
- Alphabet Inc. Class A and Class C (GOOG & GOOGL)
- Charter Communications Inc Class A
- Walt Disney
- NFLX, AT&T, Verizon, comcast, activision

Source

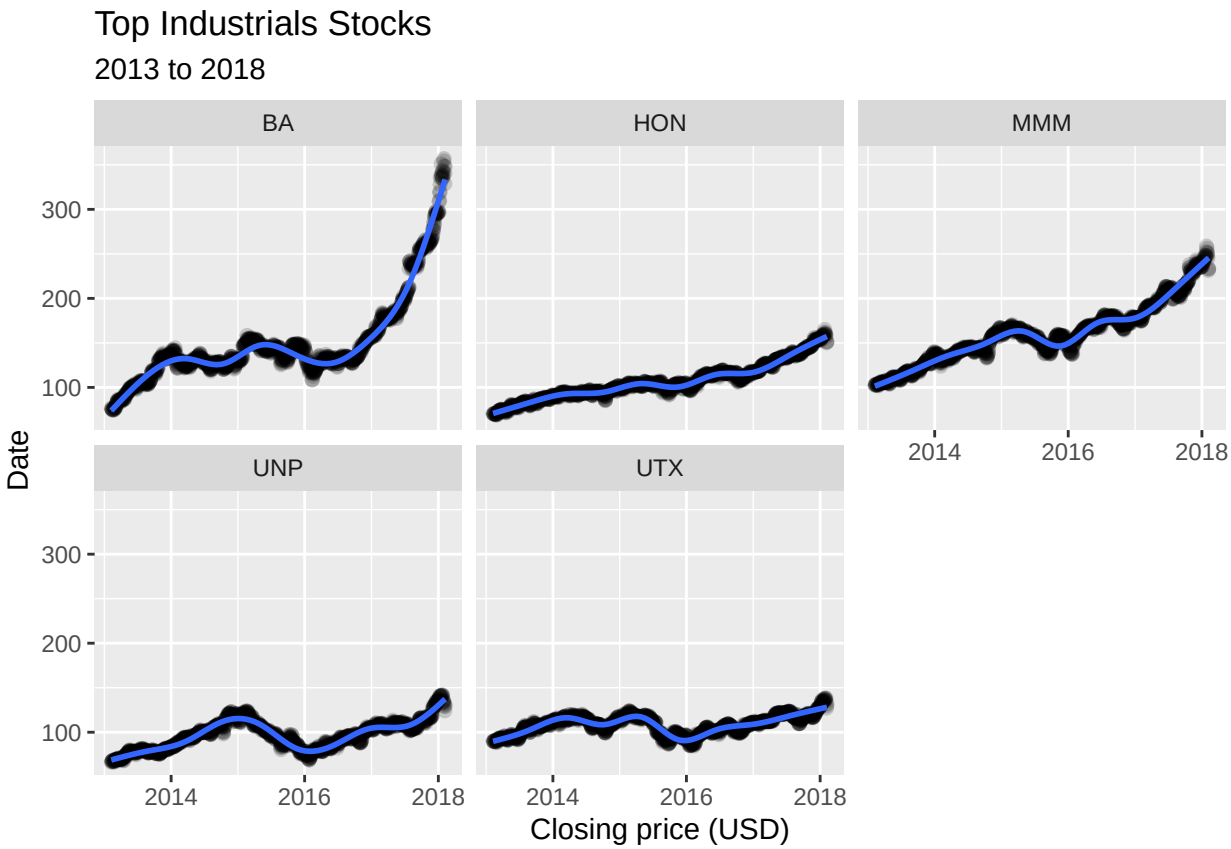


The above graphs show the dominance Google has over its competitors in the telecommunications sector of the S&P 500. With every other stock making minimal gains Google shows skyrocketing growth, out performing expected earnings. This can especially be seen around mid-2015 where Google seems to just take off. Source Once again there doesn't seem to be any major influence from outside that runs synonymous with all the stocks.

Industrials - 9.4%

- BA
- HON
- UNP
- UTX
- MMM, LMT, GE

Source

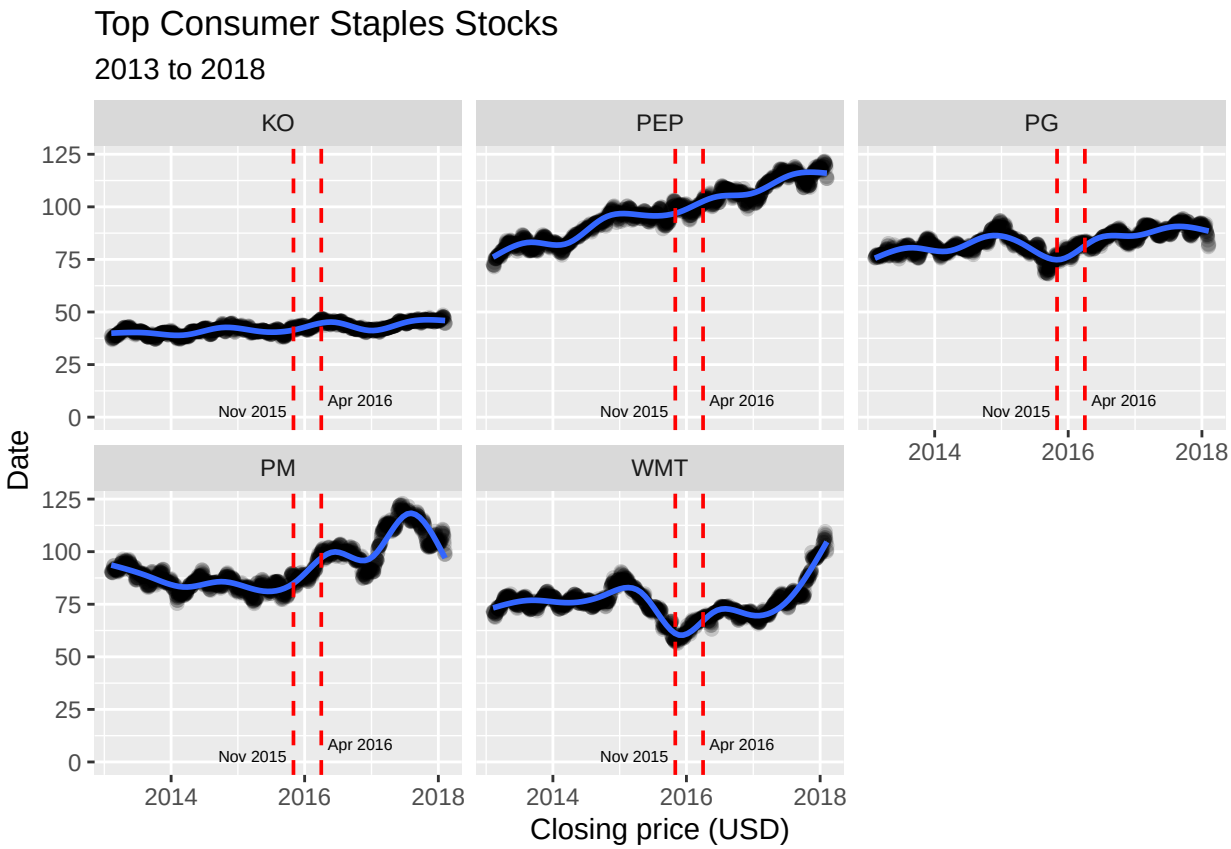


Once again, we can see that most of the stocks at the top of the industrials sector have average growth, there are two outliers almost on opposite spectrums. Boeing Co has seen unimaginable growth from 2017 and up, where if we glance at the Union Pacific Corporation there stock dipped horribly but recovered. With these volatile movements in mind compared to the average growth of the other stocks we can see a sort of pattern developing within all of the data. There isn't much outside influence that we can see driving the stock market, rather internal successes and failures within companies that may cause change to the stock.

Consumer Staples - 7.4%

- PM
- PEP
- KO
- PG
- WMT

SOURCE



The top consumer staples stocks paint a completely different picture. The gains are all over the place and the stocks dip and rise as if on random cue. However, through all of the confusion, it slowly becomes clear that around the 2016 year mark, the time proved to be a turning point, good or bad. Around this time one can see noticeable changes in the direction of all the stocks. This may indicate an outside force affecting the market.

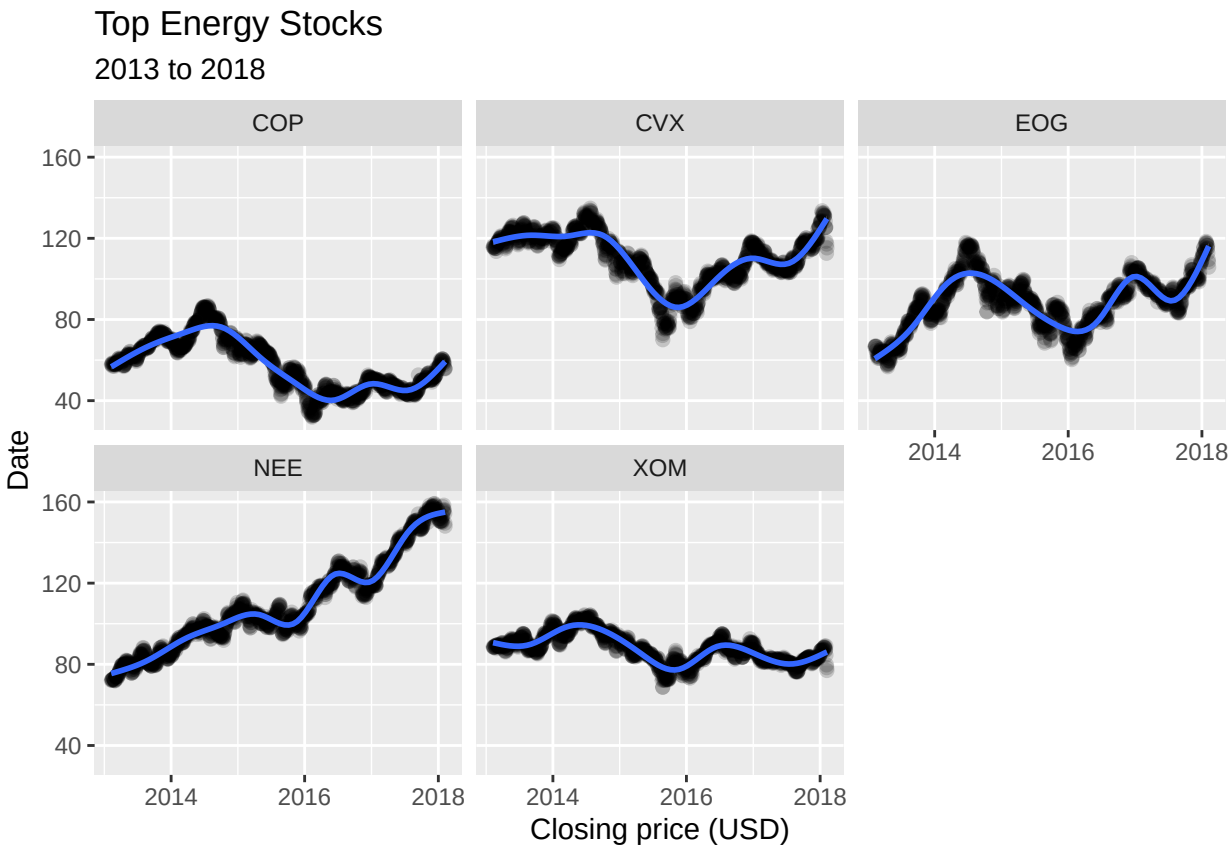
Energy - 5.4%

- XOM
- RDS.A (shell) *Not in the s&p 500 b/c it is not considered in the US market*
- CVX
- TOT
- BP, PTR, NEE

SOURCE 1

&

SOURCE 2

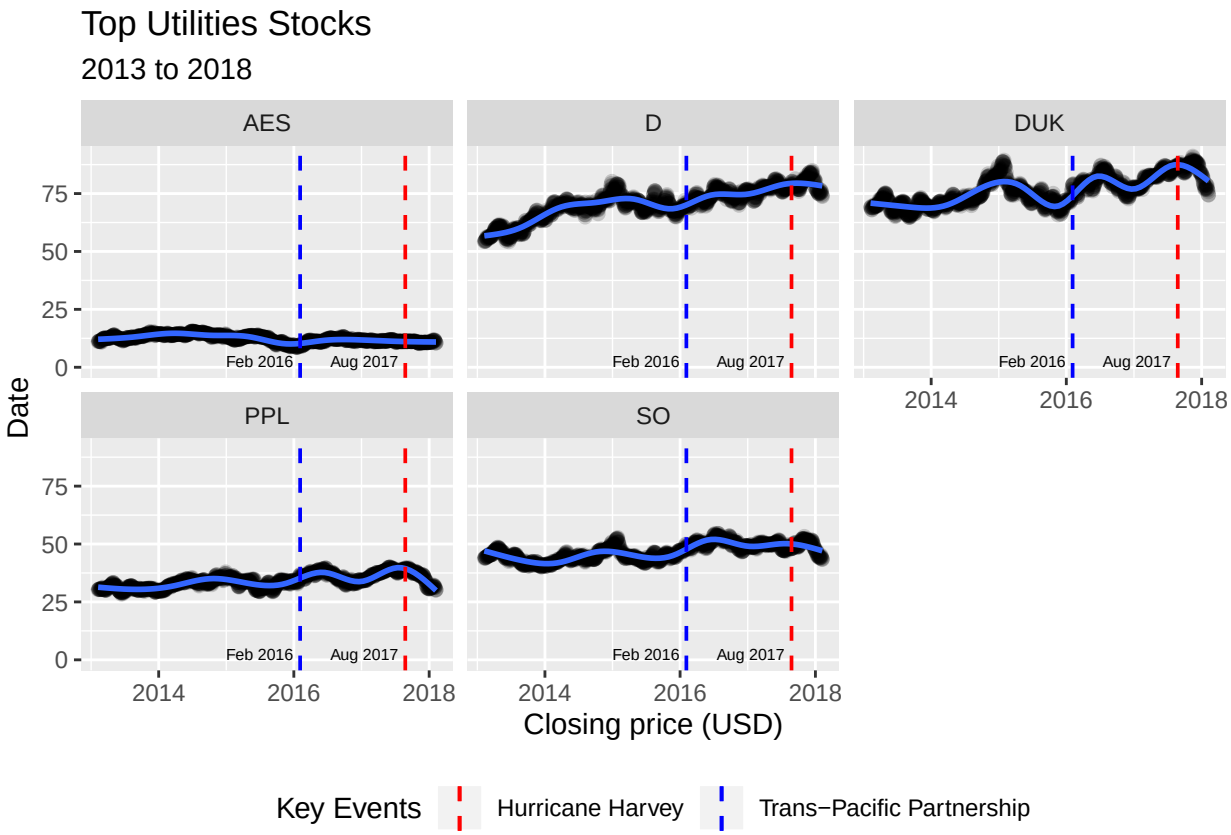


Here we can see the energy sector acts quite erratically. Most of these top companies deal with the production and exploration of hydrocarbons. Looking at these statistics really tells us nothing concrete, each company seems to come out positive or just about even in the long run. There is no way in telling if any outside entity influenced these stocks.

Utilities - 3.1%

- PPL
- D
- SO
- DUK
- CNNP, AES, EIX, FE, ED

SOURCE



Surprisingly, utilities stocks seem unfazed from anything happening around them. They show minimal but positive growth amidst many world events happening all the time. From hurricane Harvey in August of 2017 to the Trans-Pacific partnership in February of 2016, not much can sway these stocks from their course.

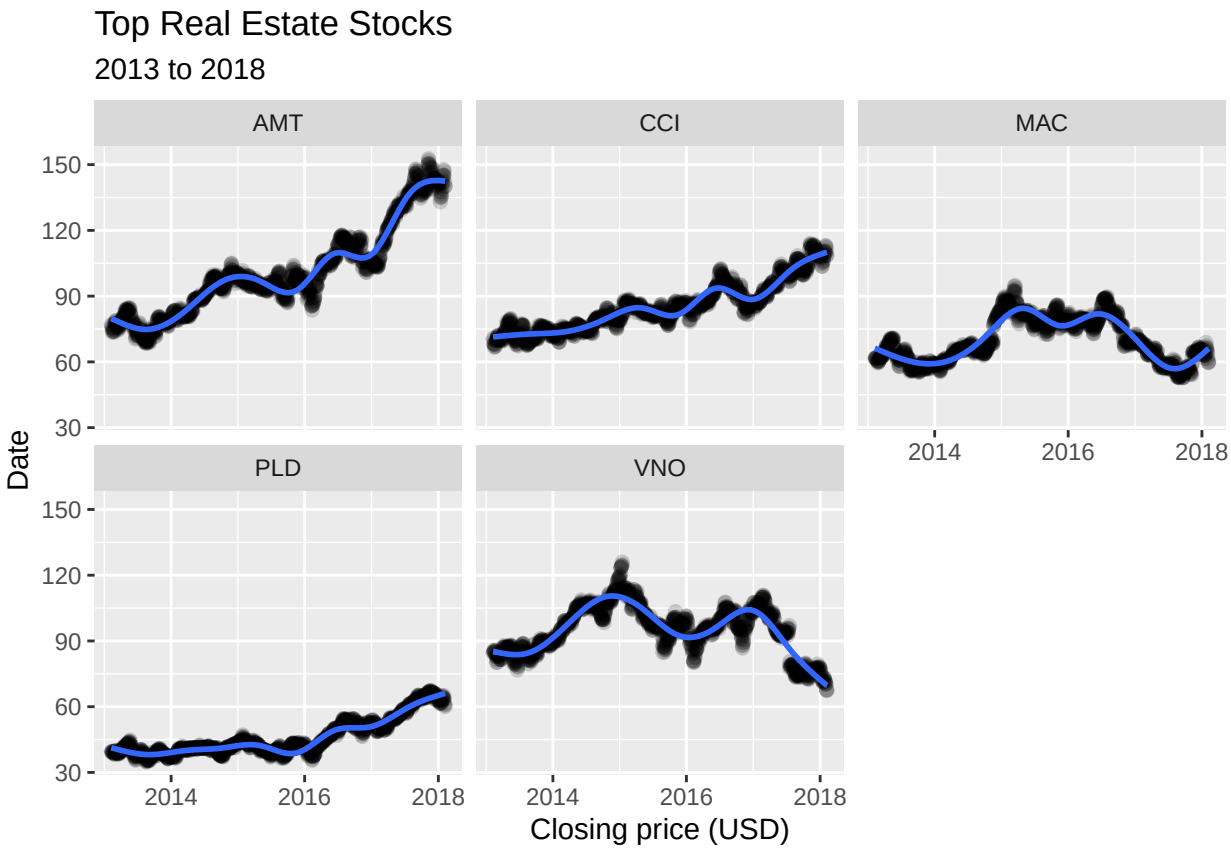
Real Estate - 2.9%

- ENSG
- ROIC
- AMT
- VNO
- HST, CBRE, WELL, PLD, CCI

SOURCE 1

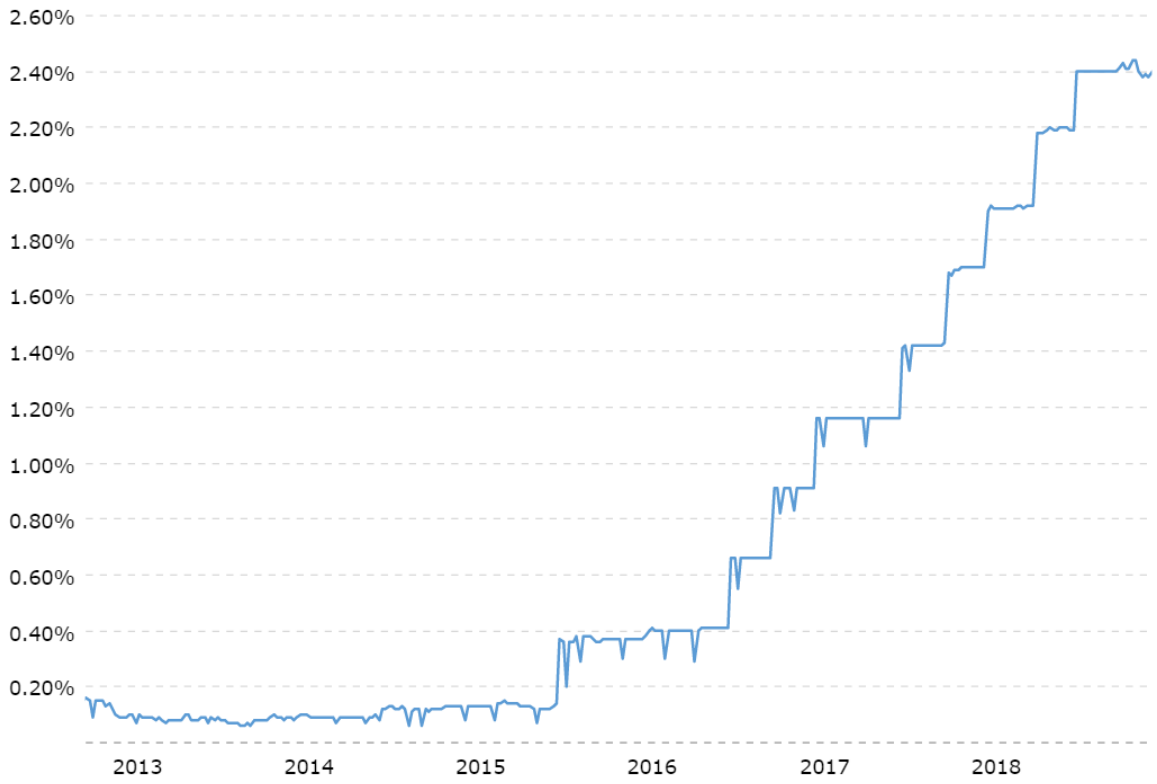
&

SOURCE 2



With interest rates on the rise, as seen below, we can see somewhat of an increase in gains, but there are still stocks that are lagging behind and even going negative. Once, again we cannot say for certain how these outside events are impacting the course of the market. We can look at the historical interest rates from 2013-2018 as a reference and compare them to how the stocks reacted throughout this time period.

Historical Interest rates 2013-2018

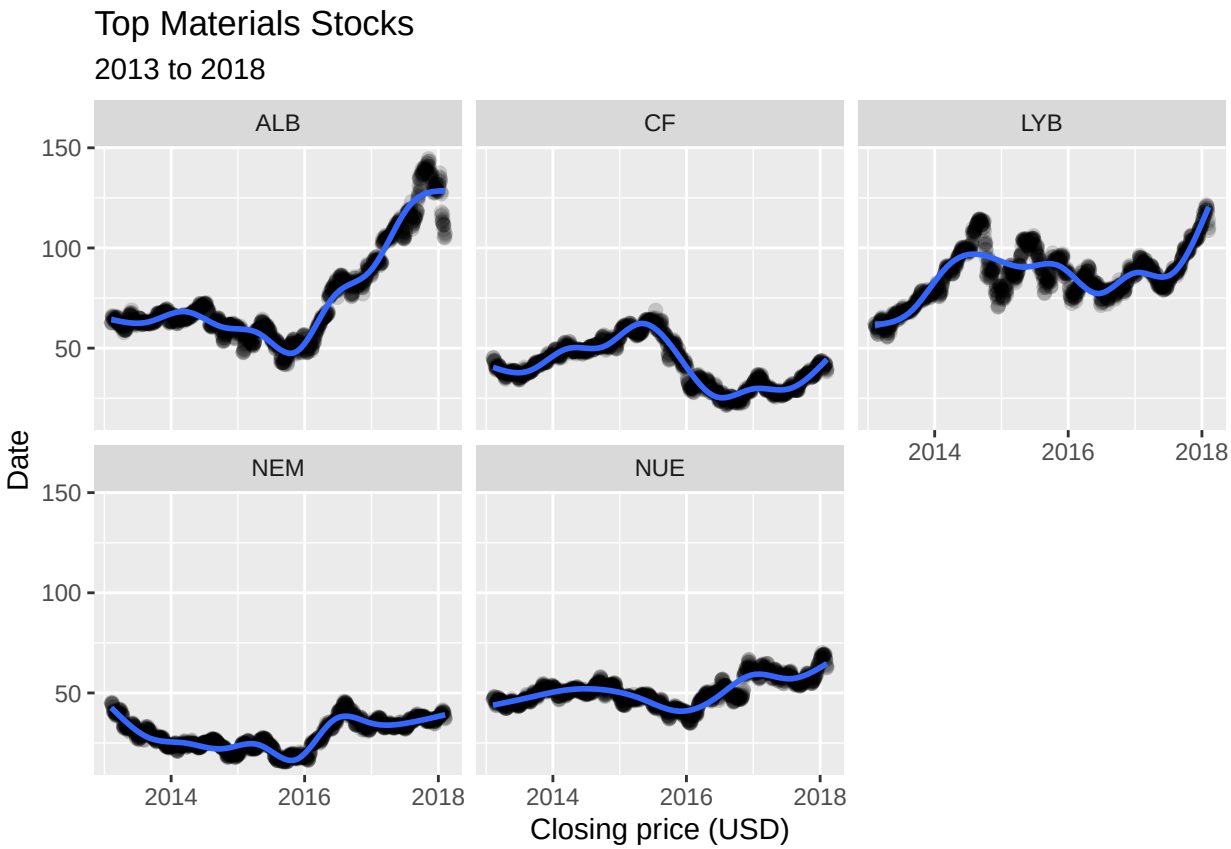


Source

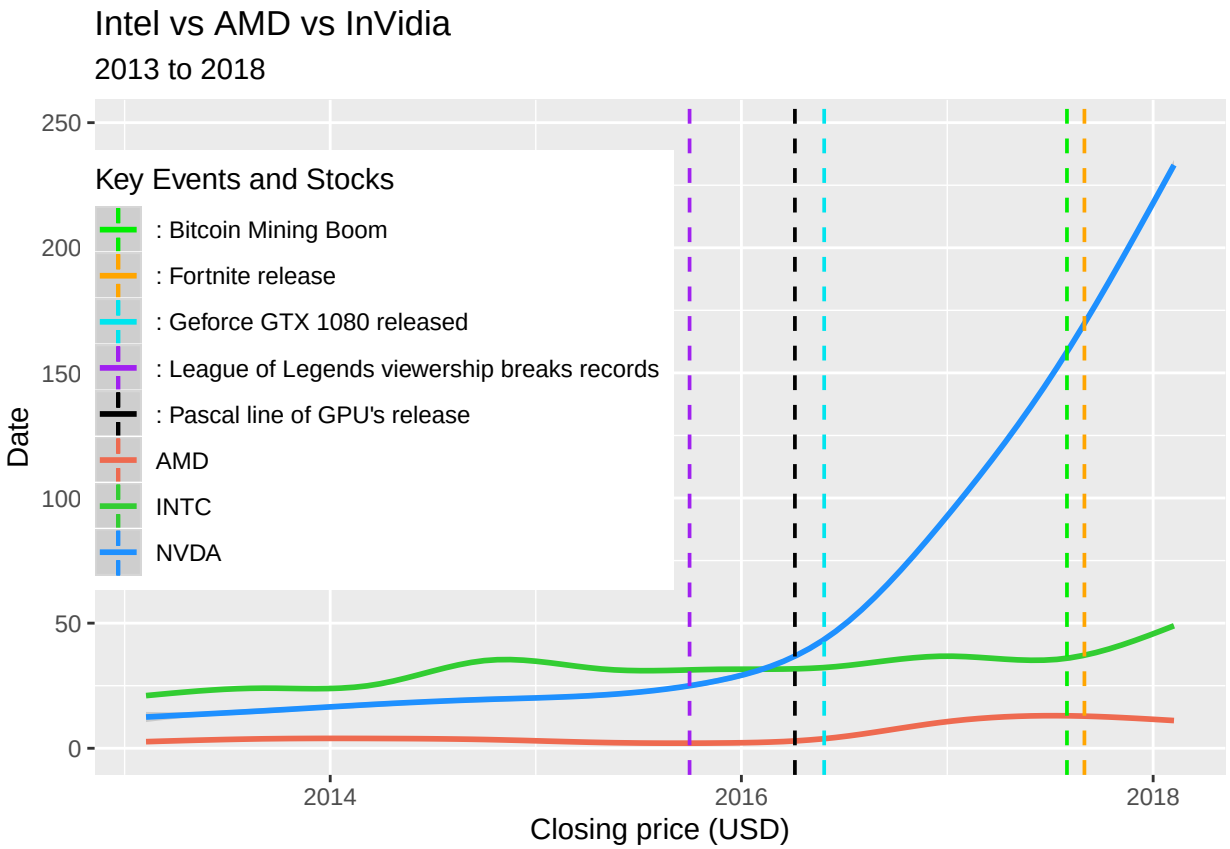
Materials - 2.6%

- NUE
- LYB
- ALB
- NEM
- CF, FMC,

SOURCE



Here, the top materials stocks shown the same sort of trends as seen previously, which was no trend at all! Although the stocks do seem to be positive as growth goes, there is still much to question about outside forces interacting with the future of the sector.



This graph shows that in the year 2016 Nvidia's stock tripled by years end. Compared to the other similar stocks in its sector, Nvidia exponentially increased its value. This is mainly because of the release of its Pascal line of graphics cards at the beginning of the pc gaming revolution. SOURCE This also shows that even though a stock has explosive growth in a single sector, accompanying stocks do not rise or drop as a result.

Delving Deeper into the Data

Interesting Q&A

- 1. Which stock has the highest gains from 2013 - 2018?

```
## # A tibble: 1 x 2
## # Groups:   Name [1]
##   Name  Gains
##   <fct> <dbl>
## 1 PCLN  1182.
```

Booking Holdings Inc

- 2. Which has the lowest gains from 2013 - 2018?

```
## # A tibble: 1 x 2
```

```
## # Groups:   Name [1]
##   Name  Gains
##   <fct> <dbl>
## 1 RL    -69.9
```

Ralph Lauren Corp

- 3. Most popular stock of 2017?

```
## # A tibble: 1 x 2
## # Groups:   Name [1]
##   Name  Total_Volume
##   <fct>        <dbl>
## 1 BAC    19779231722
```

Bank of America Corp

- 4 . Least popular stock of 2017?

```
## # A tibble: 1 x 2
## # Groups:   Name [1]
##   Name  Total_Volume
##   <fct>        <dbl>
## 1 APTV      36306643
```

Aptiv PLC

Hypothesis Test

In order to answer the problem at hand a bit more precisely, we set up a little test. Did stocks in the energy sector rise in the summer of '17 due to the higher demand for power in order to cool down? Solely on the fact that it was hotter than other months did this affect the stock gains?

```
## # A tibble: 13 x 2
## # Groups:   Name [13]
##    Name  mean_close
##    <fct>      <dbl>
##  1 NEE        151.
##  2 CXO        111.
##  3 CVX        108.
##  4 XEC         99.7
##  5 EOG         85.0
##  6 XOM         76.3
##  7 COP         43.7
##  8 HAL         39.0
##  9 HES         38.9
## 10 APA         38.8
## 11 DVN         31.4
## 12 COG         25.6
## 13 KMI         19.3
```

```
## # A tibble: 13 x 2
## # Groups:   Name [13]
##    Name  mean_close
##    <fct>      <dbl>
##  1 NEE        134.
##  2 CXO        127.
##  3 XEC        117.
##  4 CVX        107.
##  5 EOG         92.5
##  6 XOM         81.6
##  7 HES         48.8
##  8 APA         48.6
##  9 COP         47.9
## 10 HAL         45.9
## 11 DVN         39.5
## 12 COG         23.2
## 13 KMI         20.6
```

By refrencing the two tables one can see there is a slight increase in the means of the stocks during the summer months.

Using T.Tests

```
spring <- data.frame(stock = c("EOG", "COP", "NEE", "XOM", "CVX", "APA", "DVN", "XEC", "CXO", "COG", "HA

summer <- data.frame(stock = c("EOG", "COP", "NEE", "XOM", "CVX", "APA", "DVN", "XEC", "CXO", "COG", "HA

mean_spring = 71.7208

mean_summer = 66.6738

# creating dataframes for both the control and experimental groups of data
```

```
#(summer$mean_close)

t.test(as.numeric(summer$mean_close), mu = 71.7208, alt = "less", paired = FALSE, conf.level = 0.95)

t.test(as.numeric(spring$mean_close), mu = 71.7208, alt = "less", paired = FALSE, conf.level = 0.95)

# running the t.tests
```

The T-test above shows that the true mean of Stocks from the Energy sector during summer months is actually less than the average stock prices in off months

Final Assessment

Here, we tried to compare events that occur around the world from the years 2013 - 2018 and how they impacted stock prices. We also looked at how the market rises and falls with respect to itself, and if there is was sole, major driving power that determines a change for most of the stock exchange in the S & P 500 index. We began by first finding a reputable data set with the time frame we had in mind. Once found, on Kaggle.com, we loaded said data set and after a quick glimpse, we cleaned it up to be able to use for our needs. First, we went through each sector, organized by the top-performing stocks and compared their

trajectories for the 5-year span. Then using a timeline from outside events happening around the world, we compared those to the timelines of the stocks. With little to go on, we then turned to find that successes within companies were contributing to noticeable changes. In conclusion, we found that for the most part, the stocks that make up the S&P 500 index rose and fell loosely between 2013-2018 depending on economic conditions surrounding the market. Most of these stocks showed dramatic change based on catalysts from within the respective companies. We can look at Nvidia for example, only due to their new graphics cards and timing were they so successful. On the other hand, the health care sector didn't even budge even after former President Obama signed the "Obamacare" health bill making health care mandatory. Also, as we looked throughout the different sectors, by comparing the top-performing stocks there was no true relation of each stock with each other. The stocks rose and fell independently with respect to each other. Also, our t-test results proved to be the contrast of what we had thought going into testing the hypothesis concerning seasonal changes impacting stocks in the Energy sector. At the end of our analysis, we came to the end result that there is more than meets the eye to understanding how stocks move up and down on the totem pole. This Segways into one of the limitations of our findings. We feel that if there was more time at hand, as well as access to knowledge, we could use some sort of machine learning algorithm to figure out this phenomenon of the stock market. Although this project ultimately did not prove anything significant, we believe this is a strong preface to a future solution to explaining and accurately predicting/forecasting future S&P stock models.