Final Project

Rohit Kumar & Channing Vernon

11/12/2019

Installing required packages and Libraries

```r
library(tidyverse)
library(dplyr)
```

Loading Data + Cleaning

```r
## Goals:
# Cleaning up data, making it easier to read.
# **ALso need to filter out any NA's.**

data <- read.csv("all_stocks_5yr.csv")

summary(data)
```

```
##          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)
```

```
## '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
```

Ommiting NA rows vs replacing them with 0

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

# Omitting rows with NA values in order to make calculations error free, vs setting them to 0.

summary(SP_data)
```

```
##       Name                date             open              low
##  A      :  1259   2017-12-05:   505   Min.   :   1.62   Min.   :   1.50
##  AAL    :  1259   2017-12-06:   505   1st Qu.:  40.22   1st Qu.:  39.83
##  AAP    :  1259   2017-12-07:   505   Median :  62.59   Median :  62.02
##  AAPL   :  1259   2017-12-08:   505   Mean   :  83.02   Mean   :  82.26
##  ABBV   :  1259   2017-12-11:   505   3rd Qu.:  94.37   3rd Qu.:  93.54
##  ABC    :  1259   2017-12-12:   505   Max.   :2044.00   Max.   :2035.11
##  (Other):611475   (Other)   :615999
##       high             close             volume
##  Min.   :   1.69   Min.   :   1.59   Min.   :      101
##  1st Qu.:  40.62   1st Qu.:  40.24   1st Qu.:  1070351
##  Median :  63.15   Median :  62.62   Median :  2082165
##  Mean   :  83.78   Mean   :  83.04   Mean   :  4321892
##  3rd Qu.:  95.18   3rd Qu.:  94.41   3rd Qu.:  4284550
##  Max.   :2067.99   Max.   :2049.00   Max.   :618237630
##
```

Changing date column from a factro to date

```r
# Making the "date" column usable.

SP_data$Date <- as.Date(SP_data$date, format = "%Y-%m-%d")
SP_data$date <- NULL
# Changing the date column format into a new variable, and deleting the previous one

summary(SP_data)
```

```
##       Name             open              low               high
##  A      :  1259   Min.   :   1.62   Min.   :   1.50   Min.   :   1.69
##  AAL    :  1259   1st Qu.:  40.22   1st Qu.:  39.83   1st Qu.:  40.62
##  AAP    :  1259   Median :  62.59   Median :  62.02   Median :  63.15
##  AAPL   :  1259   Mean   :  83.02   Mean   :  82.26   Mean   :  83.78
##  ABBV   :  1259   3rd Qu.:  94.37   3rd Qu.:  93.54   3rd Qu.:  95.18
##  ABC    :  1259   Max.   :2044.00   Max.   :2035.11   Max.   :2067.99
##  (Other):611475
##      close             volume               Date
##  Min.   :   1.59   Min.   :      101   Min.   :2013-02-08
##  1st Qu.:  40.24   1st Qu.:  1070351   1st Qu.:2014-05-20
##  Median :  62.62   Median :  2082165   Median :2015-08-21
```

```
##  Mean   :  83.04   Mean   :  4321892   Mean   :2015-08-18
##  3rd Qu.:  94.41   3rd Qu.:  4284550   3rd Qu.:2016-11-15
##  Max.   :2049.00   Max.   :618237630   Max.   :2018-02-07
##
```

```r
## Goals
# Figuring out how to categorize the data into 500 indivivual companies, manually or via
# automation/funtion.

SP_data %>%
  group_by(Name) %>%
  select(Name) %>%
  unique()
#summary(SP_data)
```

#Top S&P 500 Stocks

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

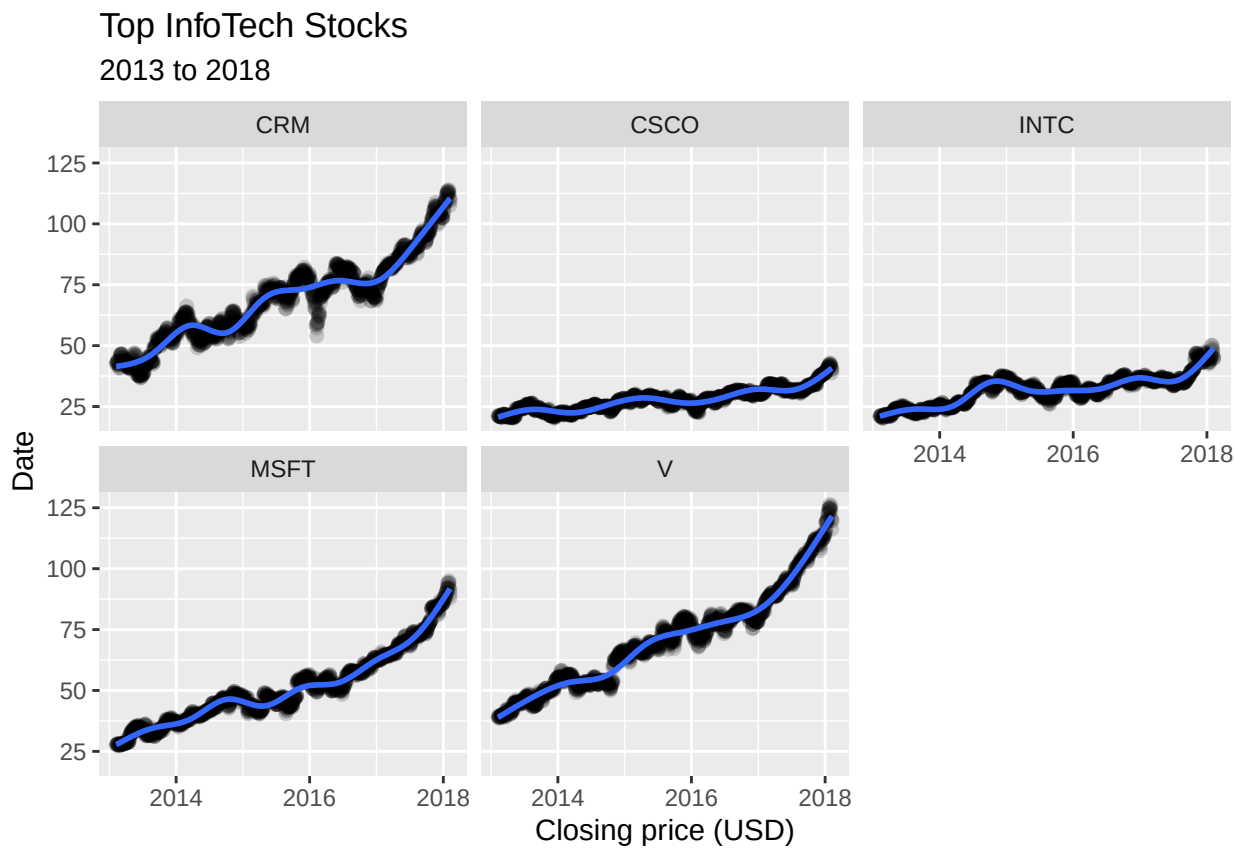
Information Technology - 19.9%

- CRM
- MFST
- CSCO
- V
- INTC

`https://us.spindices.com/indices/equity/sp-500-information-technology-sector`

```
SP_InfoTech <- ggplot(SP_data %>% filter(Name %in% c("CRM", "MSFT", "CSCO", "V", "INTC")), aes(x = Date

SP_InfoTech
```



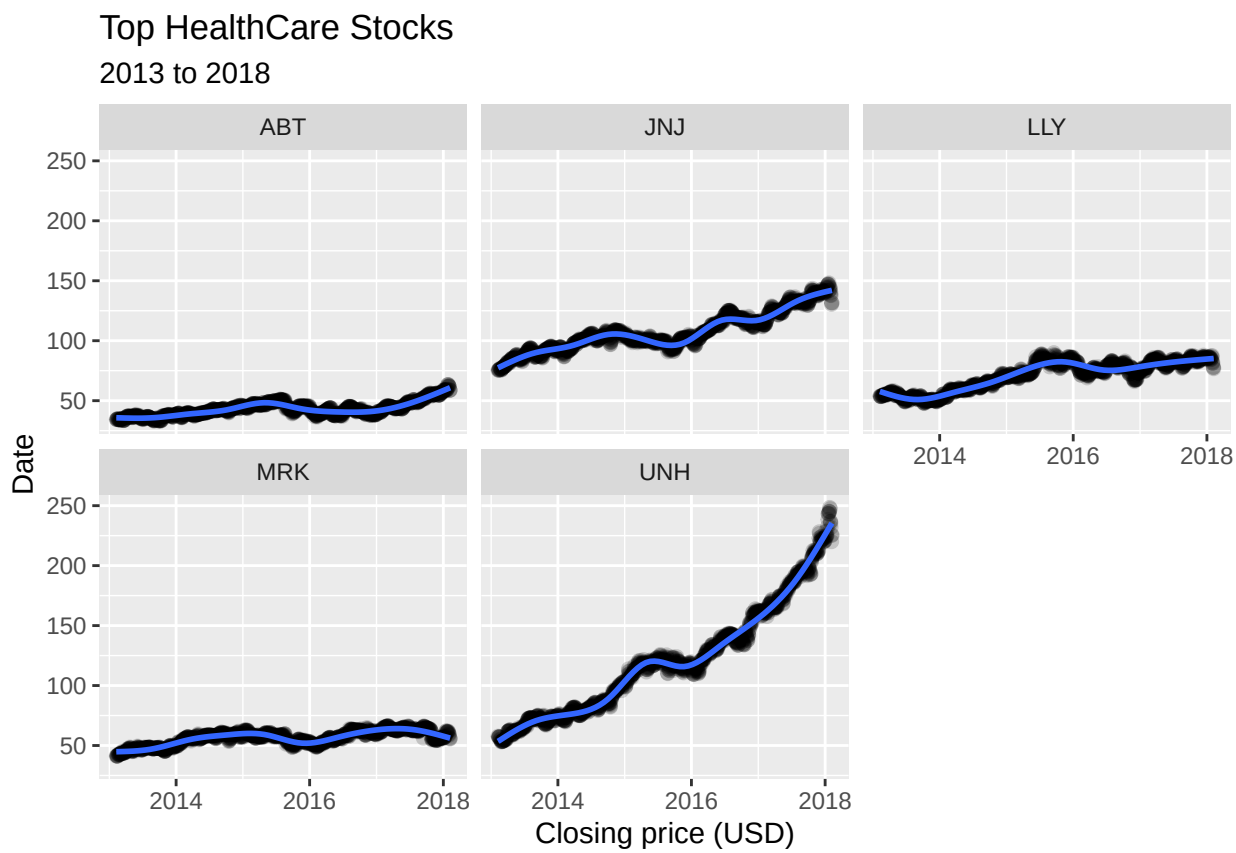
Health Care - 15.8%

- JNJ
- UNH
- MRK
- ABT
- LLY

https://us.spindices.com/indices/equity/sp-500-health-care-sector

```
SP_HealthCare <- ggplot(SP_data %>% filter(Name %in% c("JNJ", "UNH", "MRK", "ABT", "LLY")), aes(x = Dat

SP_HealthCare
```



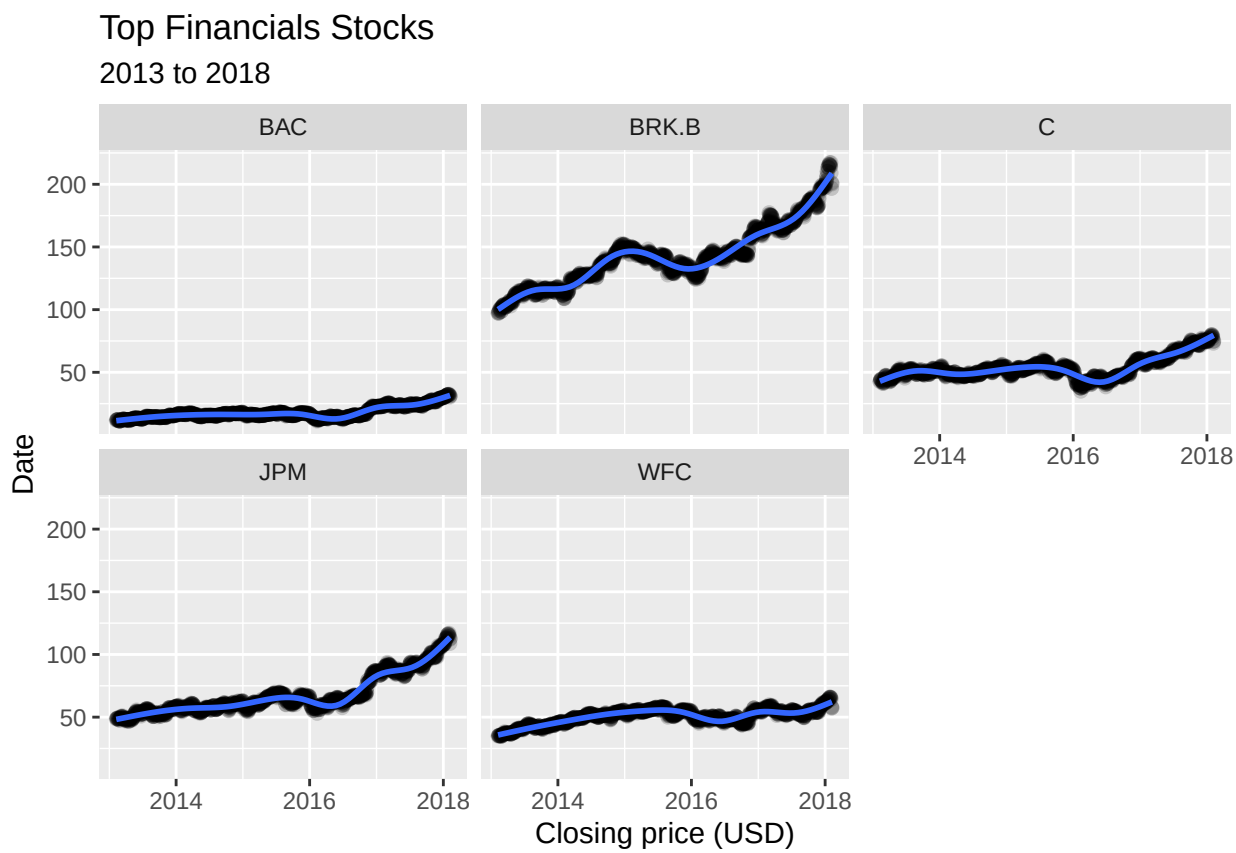
Financials - 13.7%

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

```
https://us.spindices.com/indices/equity/sp-500-financials-sector
```

```
SP_Financials <- ggplot(SP_data %>% filter(Name %in% c("BRK.B", "JPM", "BAC", "WFC", "C")), aes(x = Dat

SP_Financials
```

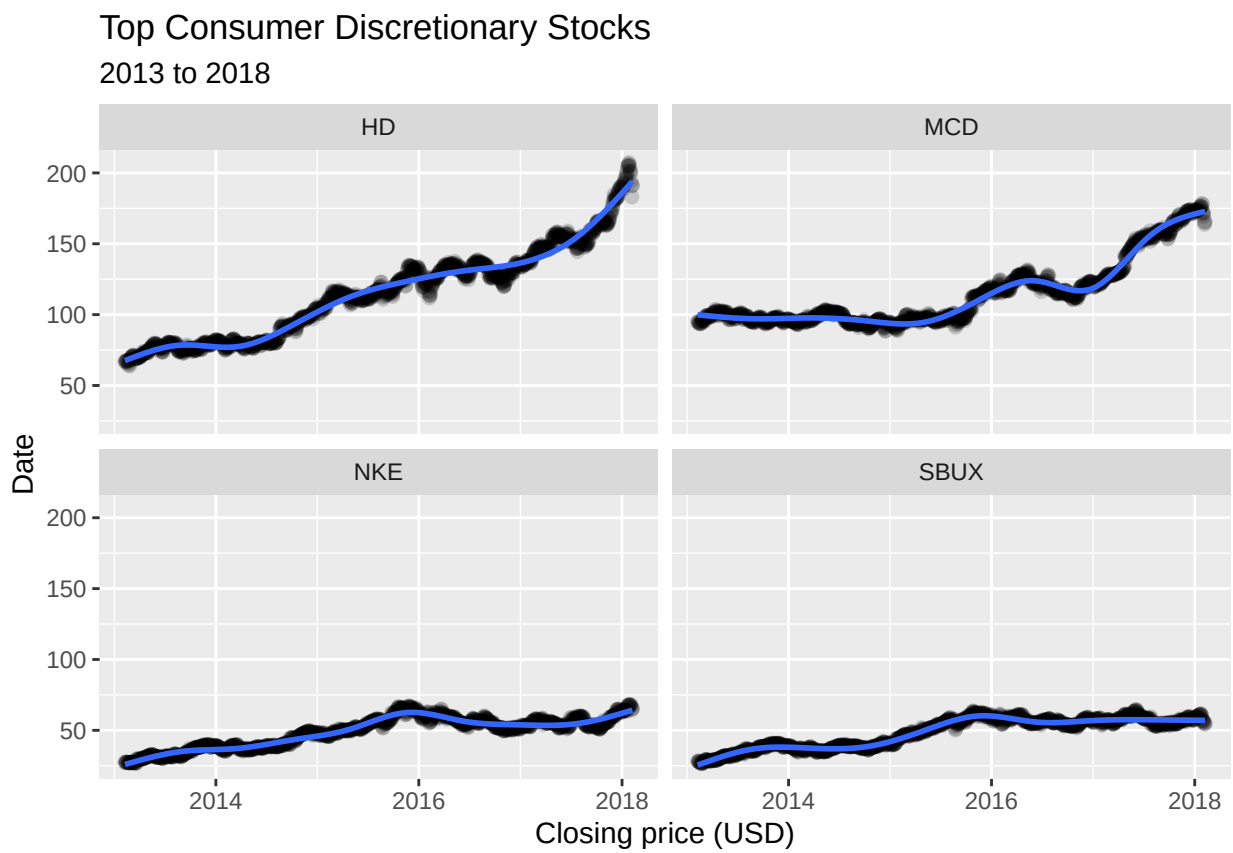


Consumer Discretionary - 9.9% * AMZN * HD * MCD * NKE * SBUX

https://us.spindices.com/indices/equity/sp-500-consumer-discretionary-sector

```
SP_ConsumerDisc <- ggplot(SP_data %>% filter(Name %in% c("HD", "MCD", "NKE", "SBUX")), aes(x = Date, y =

SP_ConsumerDisc
```




```
"AMZN"
```

```
## [1] "AMZN"
```

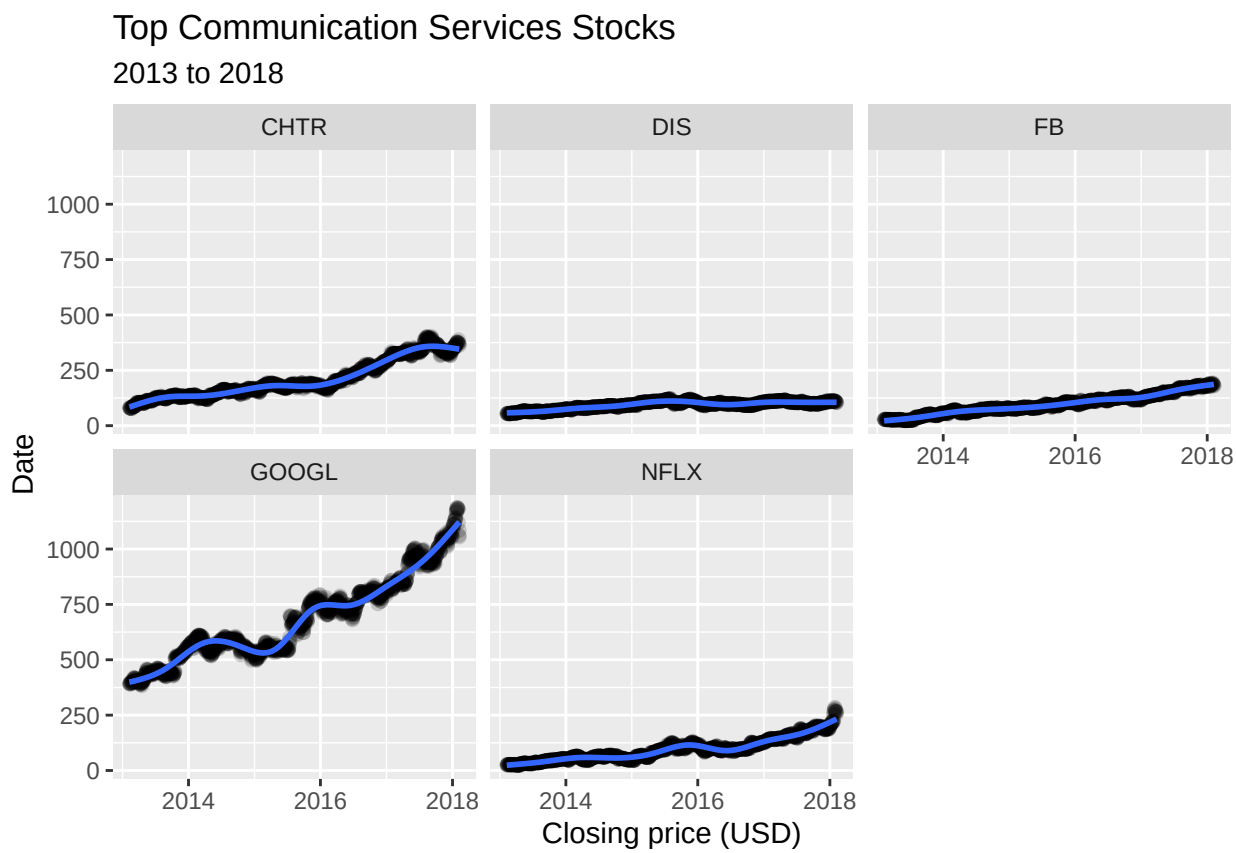
Communications Services - 9.9%

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

```
https://us.spdrs.com/en/etf/the-communication-services-select-sector-spdr-fund-XLC
```

```
SP_CommServices <- ggplot(SP_data %>% filter(Name %in% c("FB", "GOOGL", "CHTR", "DIS", "NFLX")), aes(x =

SP_CommServices
```



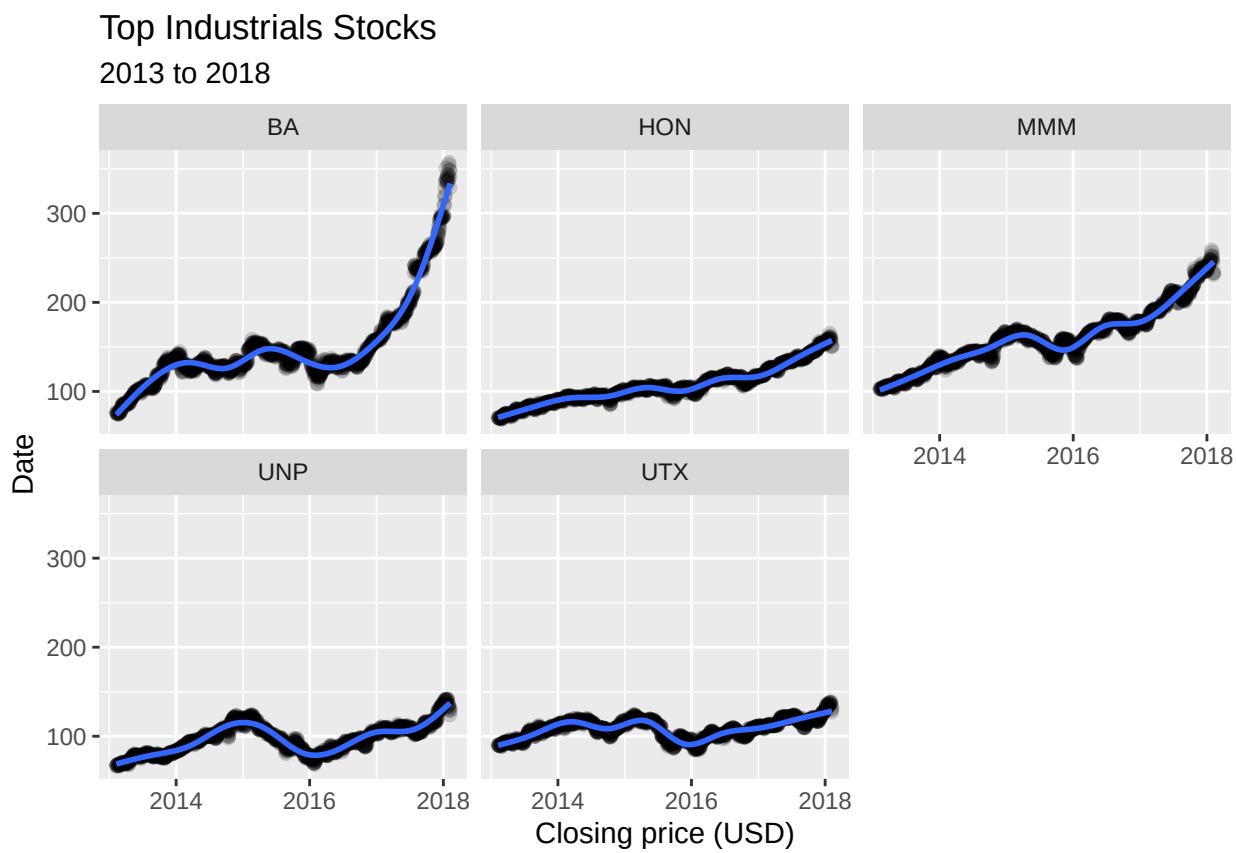
Industrials - 9.4%

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

```
https://us.spindices.com/indices/equity/sp-500-industrials-sector
```

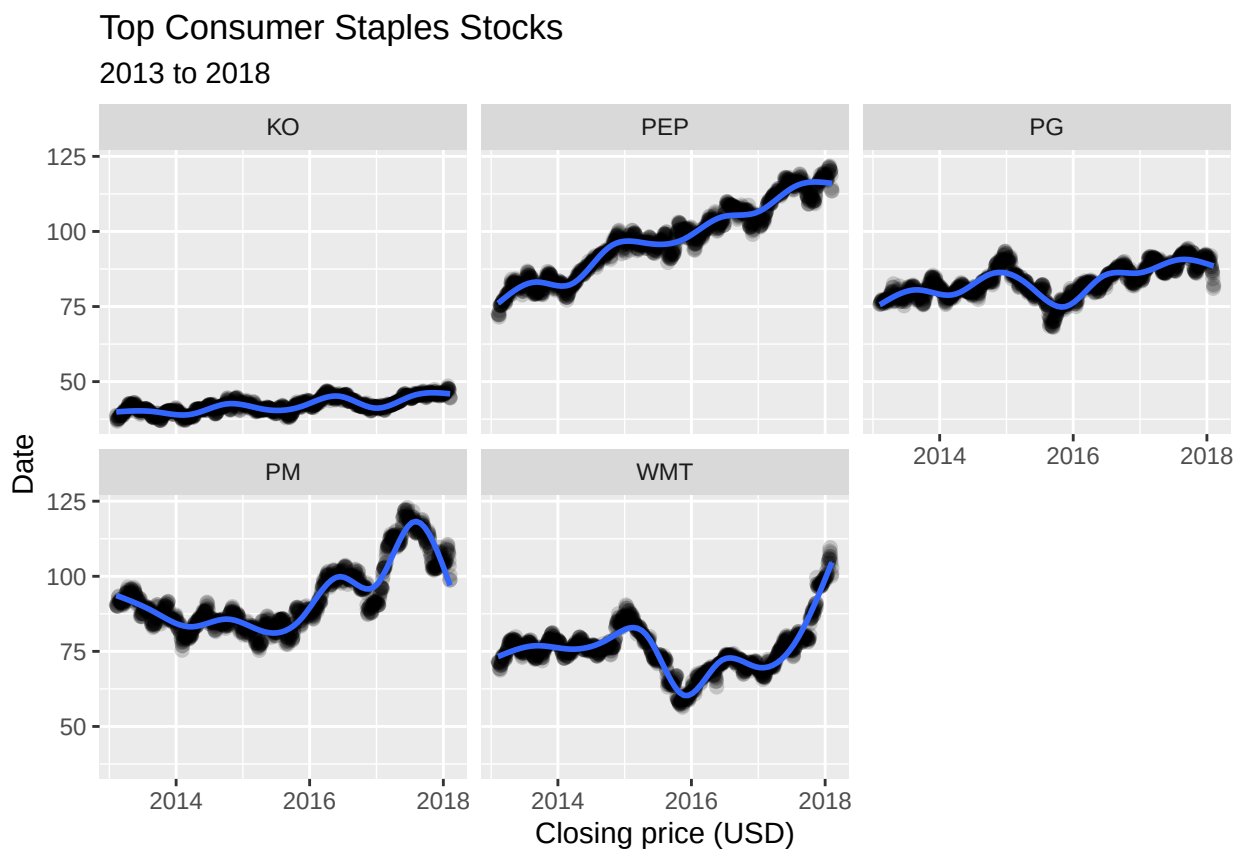
```
SP_Industrials <- ggplot(SP_data %>% filter(Name %in% c("BA", "HON", "UNP", "UTX", "MMM")), aes(x = Dat

SP_Industrials
```




```
SP_ConsumerStaples <- ggplot(SP_data %>% filter(Name %in% c("PM", "PEP", "KO", "PG", "WMT")), aes(x = Da

SP_ConsumerStaples
```



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

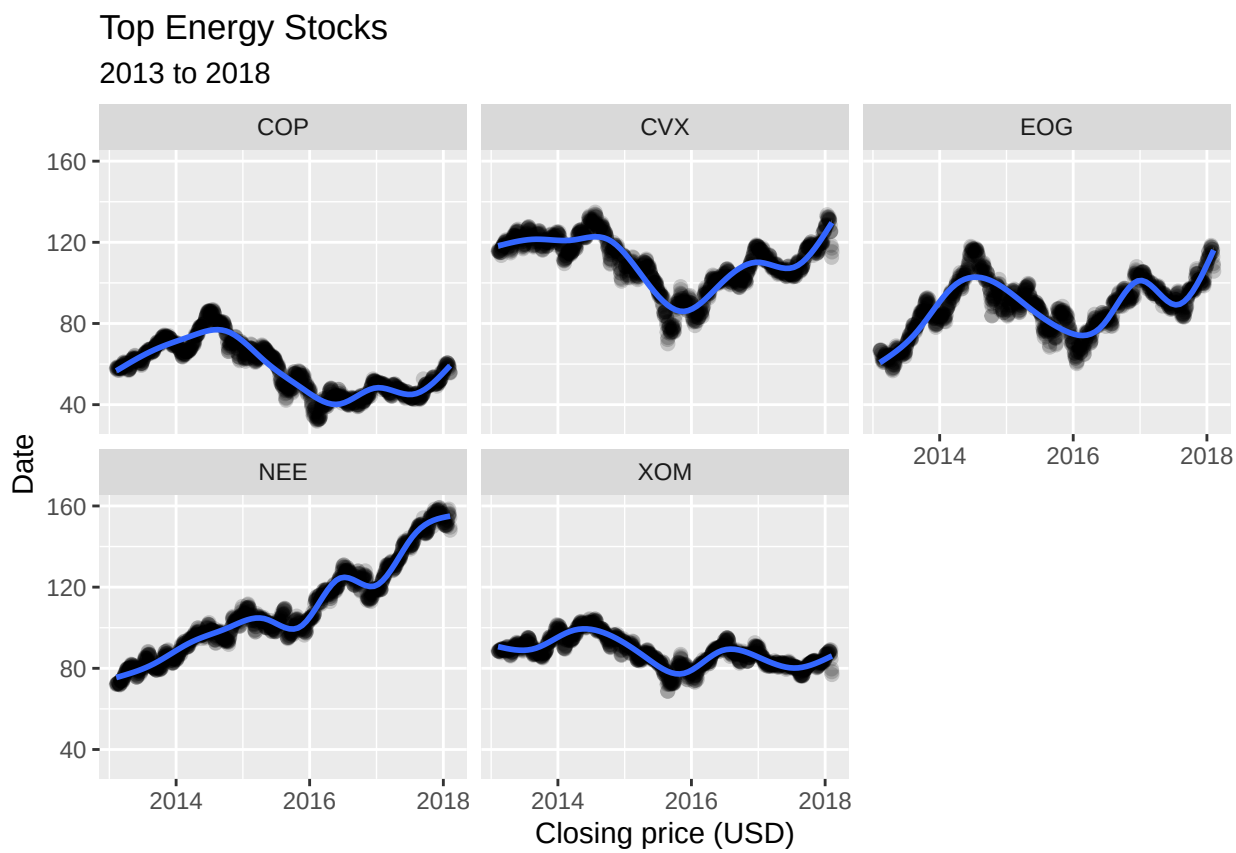
`https://www.fool.com/investing/the-10-biggest-energy-stocks.aspx`

&

`https://www.fool.com/investing/2018/09/06/the-5-biggest-oil-stocks-in-the-us.aspx`

```
SP_Energy <- ggplot(SP_data %>% filter(Name %in% c("EOG", "COP", "NEE", "XOM", "CVX")), aes(x = Date, y

SP_Energy
```



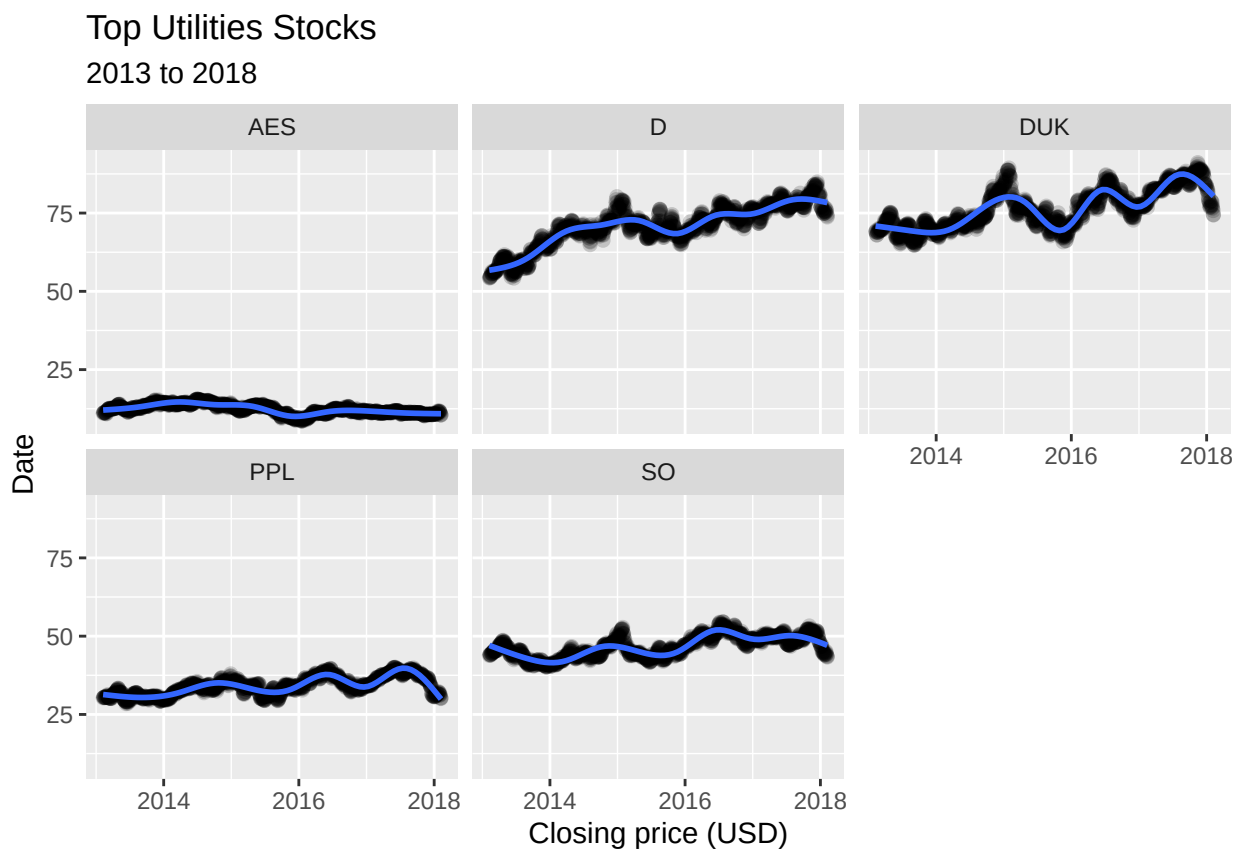
Utilities - 3.1%

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

```
https://www.barrons.com/articles/you-may-be-surprised-which-stock-market-sector-has-been-the-hottest-in-
```

```
SP_Utilities <- ggplot(SP_data %>% filter(Name %in% c("PPL", "D", "DUK", "SO", "AES")), aes(x = Date, y

SP_Utilities
```



Real Estate - 2.9%

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

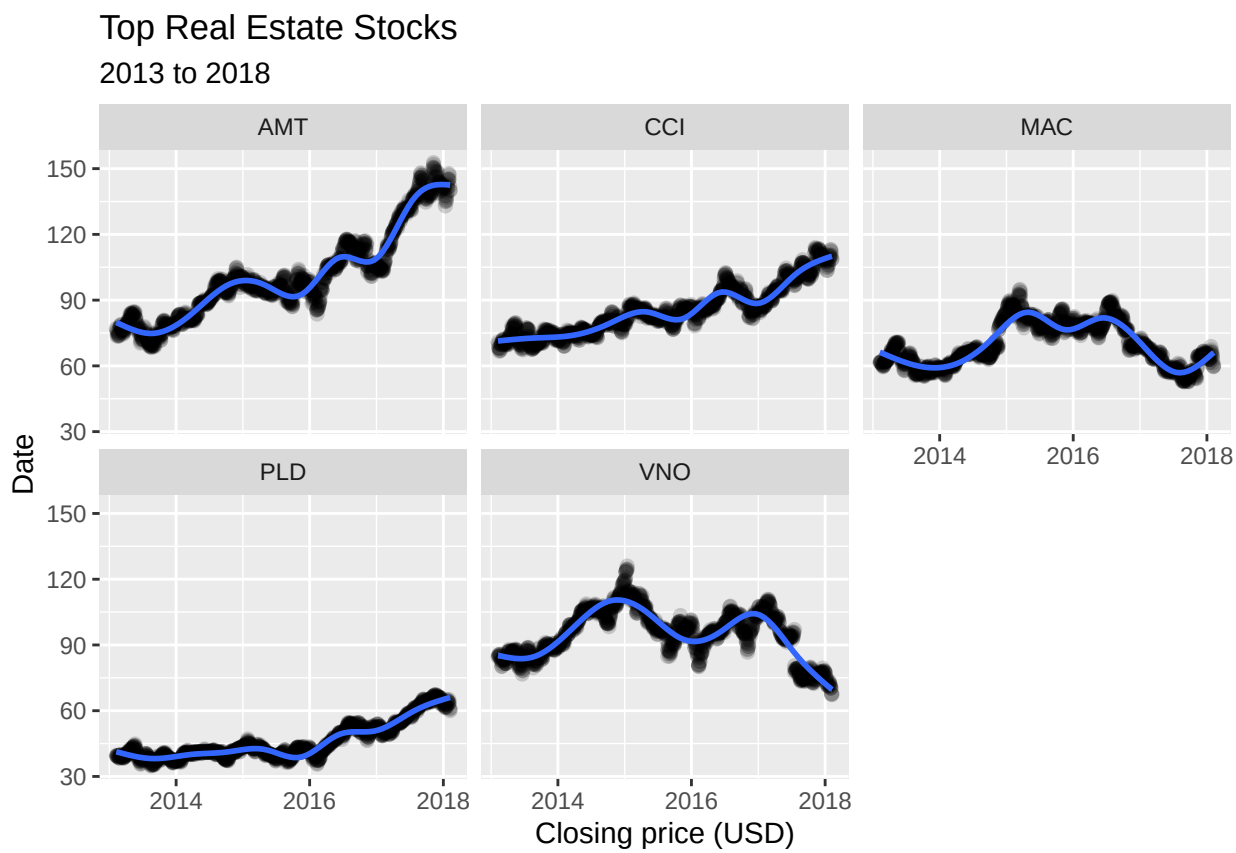
`https://www.fool.com/investing/2019/01/13/the-3-best-real-estate-stocks-to-buy-in-2019.aspx`

&

`https://www.investopedia.com/top-real-estate-stocks-4582128`

```
SP_RealEstate <- ggplot(SP_data %>% filter(Name %in% c("CCI", "PLD", "AMT", "VNO", "MAC")), aes(x = Dat

SP_RealEstate
```



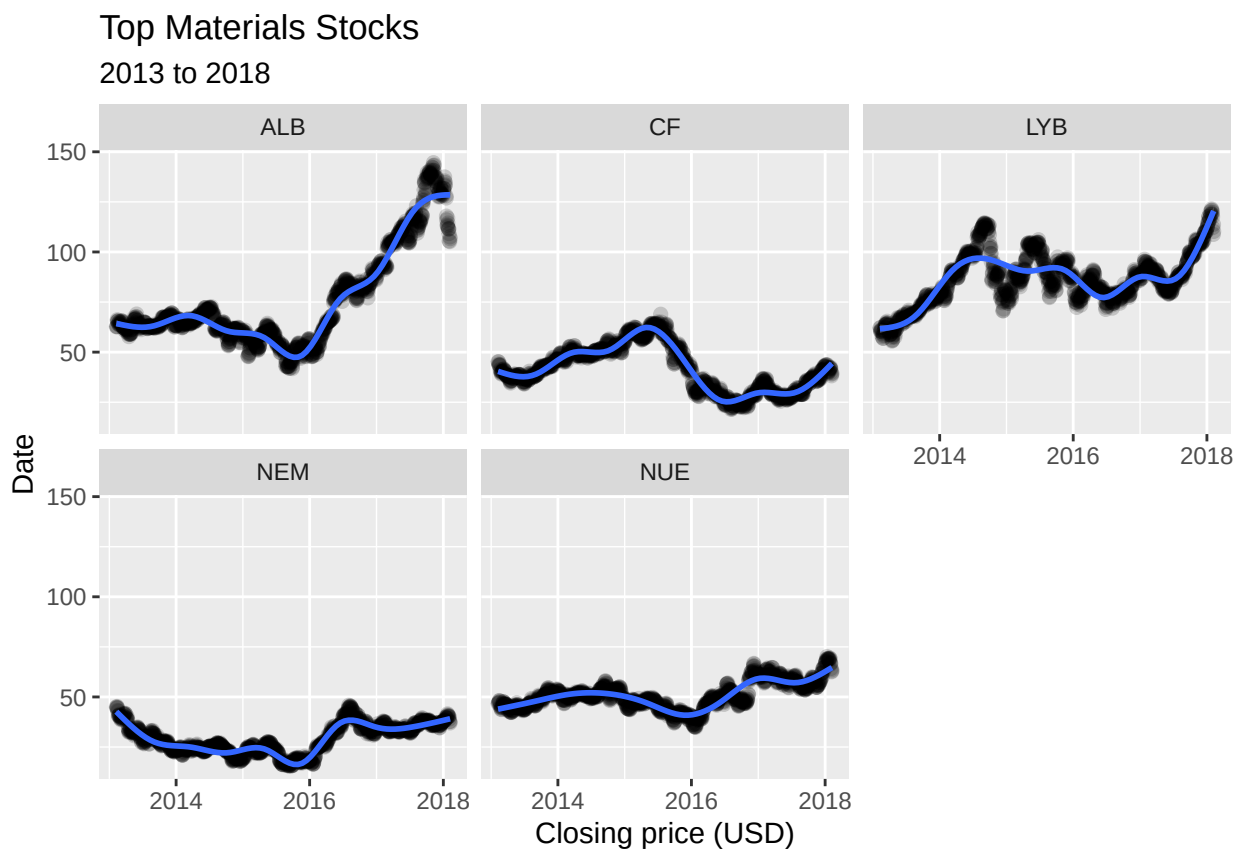
Materials - 2.6%

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

```
https://www.investopedia.com/top-materials-stocks-4582152
```

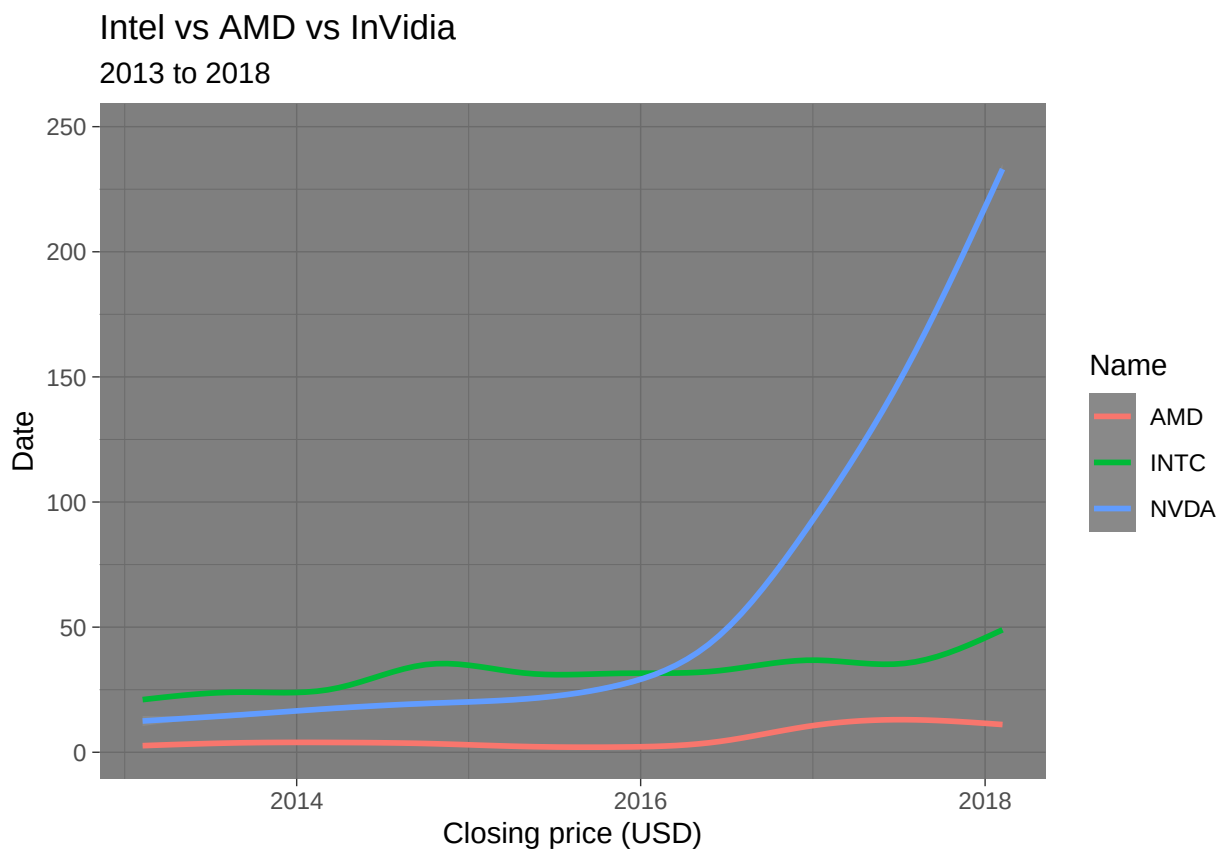
```
SP_Materials <- ggplot(SP_data %>% filter(Name %in% c("NUE", "LYB", "ALB", "NEM", "CF")), aes(x = Date,

SP_Materials
```



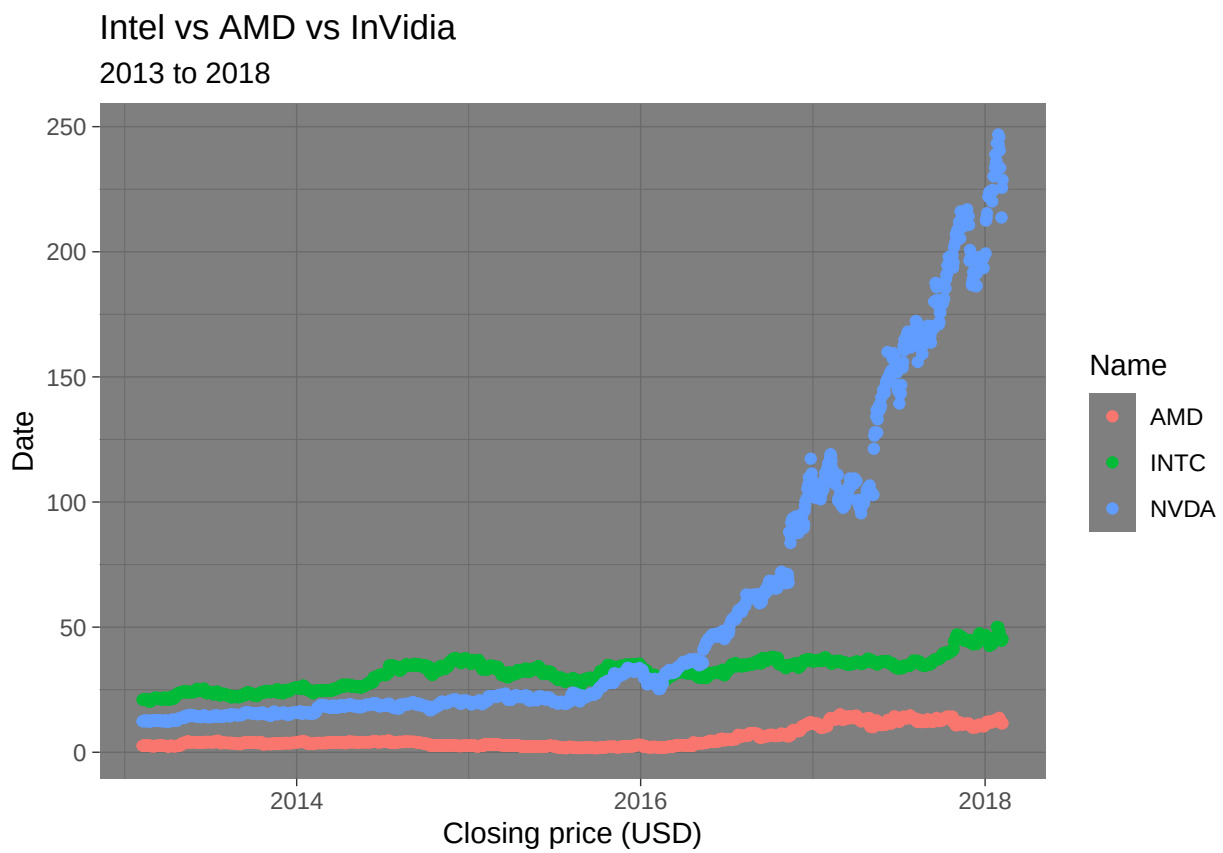

```
AMD_NVDA_INTC <- ggplot(filter(SP_data, Name %in% c("INTC", "AMD","NVDA")), aes(x = Date, y = close, col
                        subtitle = "2013 to 2018",
                        y = "Date",
                        x = "Closing price (USD)") + theme_dark()

AMD_NVDA_INTC
```




```
AMD_NVDA_INTC_ <- ggplot(filter(SP_data, Name %in% c("INTC", "AMD","NVDA")), aes(x = Date, y = close, co
                        subtitle = "2013 to 2018",
                        y = "Date",
                        x = "Closing price (USD)") + theme_dark()

AMD_NVDA_INTC_
```



Questions

- 1. Which stock has the highest gains from 2013- 2018?
- 2. Which has the lowest?
- 3. Finding the p value with a t test and the null hypothesis.
- 4. Finding a way to implement the stocks with a confidence interval.
- 5. Compare the top companies.
- 6. Events that affected the stock market from 2013 - 2018.
- 7. Most popular stock?
- 8. Least popular stock?