Final Project

Rohit Kumar & Channing Vernon

11/12/2019

Loading Data + Cleaning

```
## 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
```

```
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 ...
```

```
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

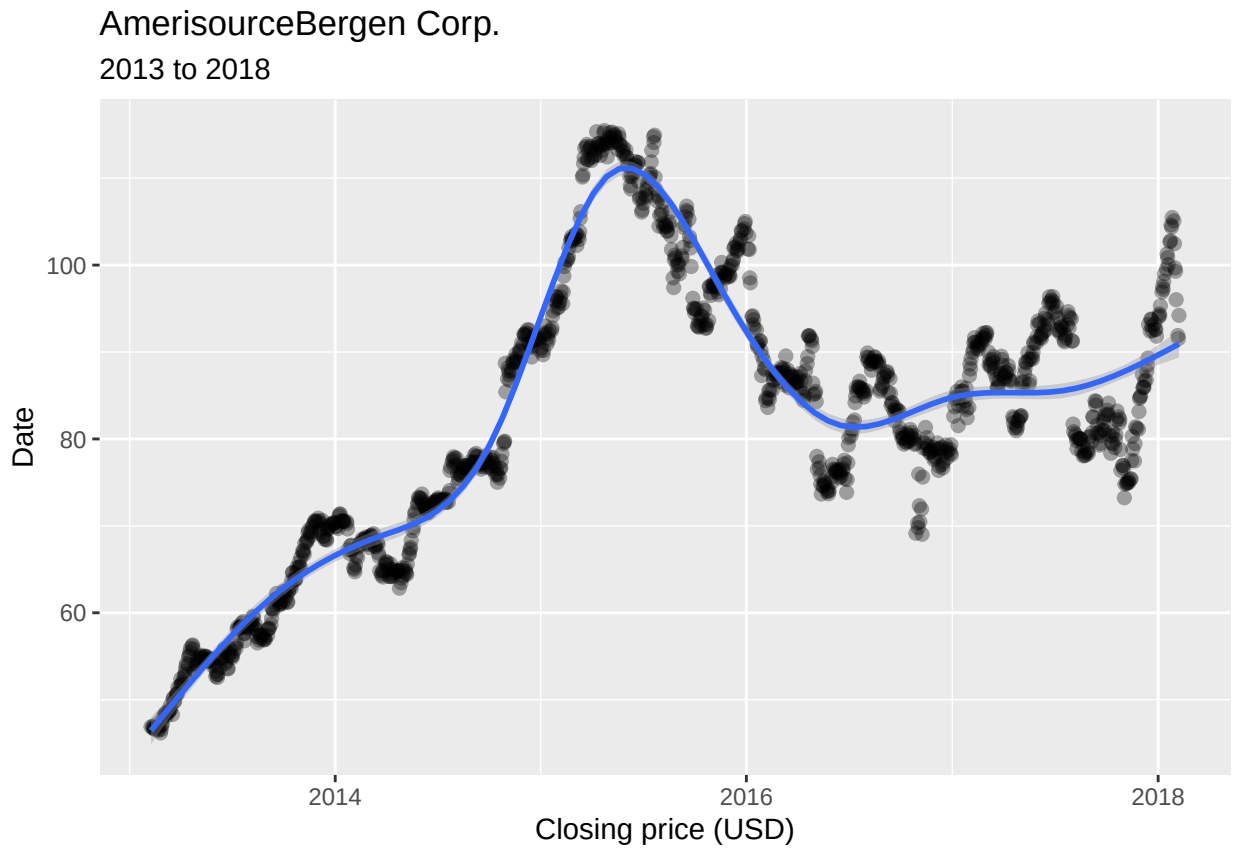
```r
# Making the "date" column usable.

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  
##
```

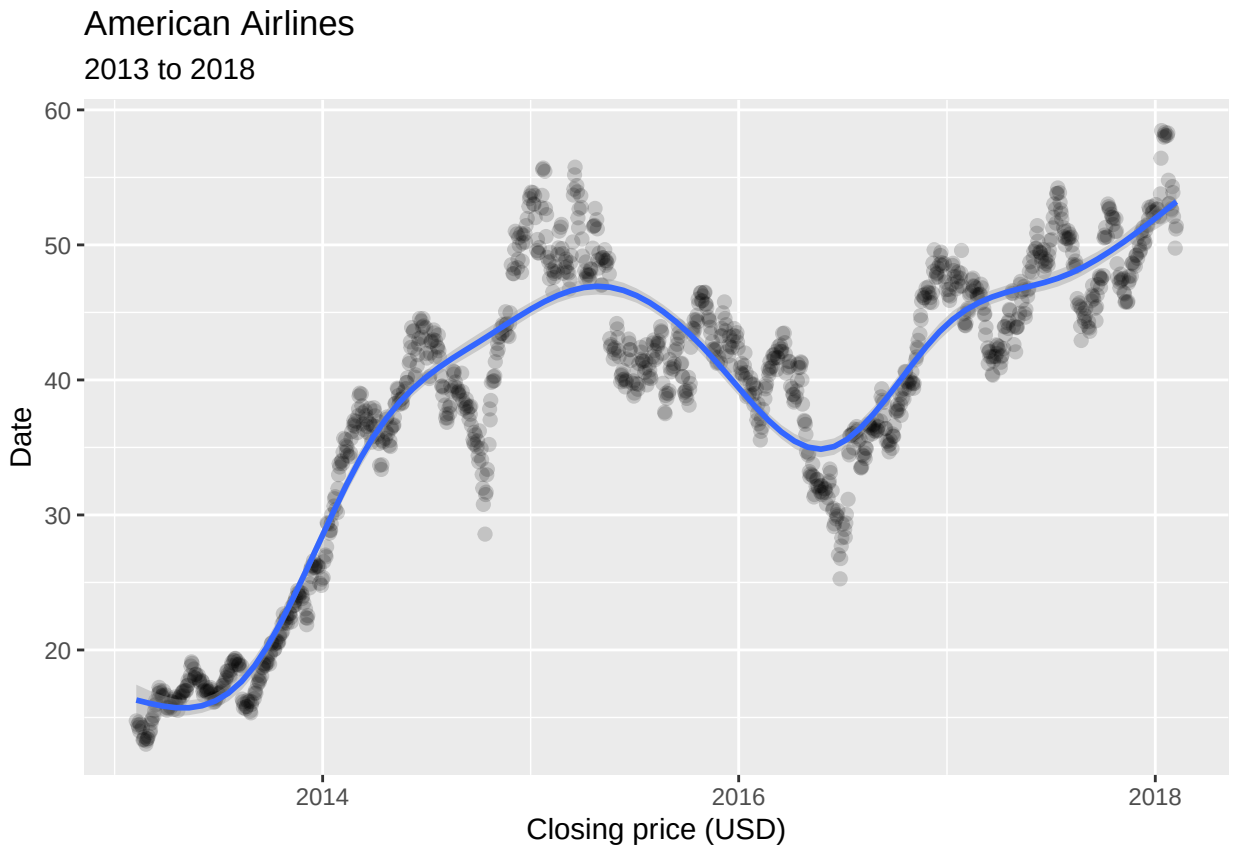
```r
## Goals
# Figuring out how to categorize the data into 500 indivivual companies, manually or via
# automation/funtion.

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

```
## # A tibble: 505 x 1
## # Groups:   Name [505]
##    Name
##    <fct>
##  1 AAL
##  2 AAPL
##  3 AAP
##  4 ABBV
##  5 ABC
##  6 ABT
##  7 ACN
##  8 ADBE
##  9 ADI
## 10 ADM
## # ... with 495 more rows
```

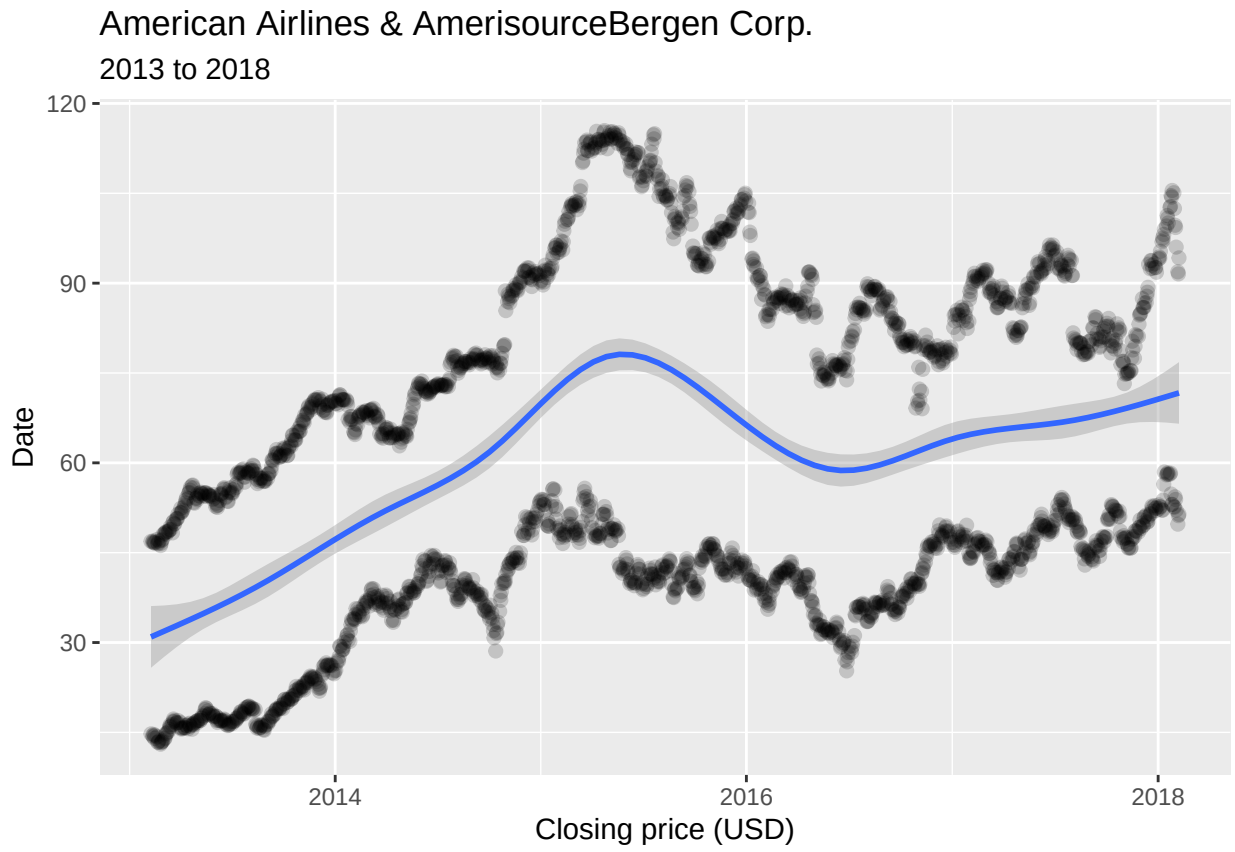
```r
# Listing all of the unique companies included in data set.

ABC <- ggplot(filter(SP_data, Name %in% c("ABC")), aes(x = Date, y = close)) + geom_point(alpha = (1/3)
                          subtitle = "2013 to 2018",
                          y = "Date",
                          x = "Closing price (USD)")

ABC
```




```
AAL <- ggplot(filter(SP_data, Name %in% c("AAL")), aes(x = Date, y = close)) + geom_point(alpha = (1/6)
                          subtitle = "2013 to 2018",
                          y = "Date",
                          x = "Closing price (USD)")

AAL
```




```
AAL_ABC <- ggplot(filter(SP_data, Name %in% c("AAL","ABC")), aes(x = Date, y = close)) + geom_point(alp
                          subtitle = "2013 to 2018",
                          y = "Date",
                          x = "Closing price (USD)")

AAL_ABC
```




```
#summary(SP_data)
```