APAN5420 — HW 1

Megan Wilder

5/25/18

Contents

---

1 Load both ETF and Stock Data

```
#load packages
library(readr)  # for read_csv()
library(dplyr)  # for mutate()
library(tidyr)  # for unnest()
library(purrr)  # for map(), reduce()
library(stringr) # extract symbols

# find all file names ending in .txt
files <- dir(pattern = "*.txt")

#load ETF and Stock data and create symbol column with ticker from file name

DF = tibble(File = files) %>%
  extract(File, "Symbol", "([^.]+)", remove = FALSE) %>%
  mutate(Data = lapply(File, read_csv)) %>%
  unnest(Data) %>%
  select(-File)
```

2 Explore Data

```
#Summary
summary(DF)
```

```
##     Symbol              Date                 Open
##  Length:17453243    Min.   :1962-01-02   Min.   :0.000e+00
##  Class :character   1st Qu.:2008-05-22   1st Qu.:9.000e+00
##  Mode  :character   Median :2012-06-08   Median :1.800e+01
##                     Mean   :2010-11-12   Mean   :2.625e+04
##                     3rd Qu.:2015-06-23   3rd Qu.:3.400e+01
##                     Max.   :2017-11-10   Max.   :1.424e+09
##                                          NA's   :2868
##       High               Low                Close
```

```
##  Min.   :0.000e+00   Min.   :        -1   Min.   :0.000e+00
##  1st Qu.:9.000e+00   1st Qu.:         9   1st Qu.:9.000e+00
##  Median :1.900e+01   Median :        18   Median :1.800e+01
##  Mean   :2.697e+04   Mean   :     25363   Mean   :2.613e+04
##  3rd Qu.:3.400e+01   3rd Qu.:        33   3rd Qu.:3.400e+01
##  Max.   :1.442e+09   Max.   :1362117844   Max.   :1.438e+09
##  NA's   :3179        NA's   :3081         NA's   :4502
##      Volume             OpenInt
##  Min.   :0.000e+00   Min.   :0
##  1st Qu.:2.572e+04   1st Qu.:0
##  Median :1.574e+05   Median :0
##  Mean   :1.581e+06   Mean   :0
##  3rd Qu.:7.843e+05   3rd Qu.:0
##  Max.   :2.070e+09   Max.   :0
##  NA's   :2
```

```
#View NA's in DF
new_DF_open <- DF[is.na(DF$Open),]
new_DF_high <- DF[is.na(DF$High),]
new_DF_low <- DF[is.na(DF$Low),]
new_DF_close <- DF[is.na(DF$Close),]
new_DF_volume <- DF[is.na(DF$Volume),]

unique(new_DF_open$Symbol)
```

```
## [1] "aezs"  "brk-a" "clbs"  "rgse"  "tvix"  "uvxy"
```

```
unique(new_DF_high$Symbol)
```

```
## [1] "aezs"  "brk-a" "clbs"  "rgse"  "tvix"  "uvxy"
```

```
unique(new_DF_low$Symbol)
```

```
## [1] "aezs"  "brk-a" "clbs"  "rgse"  "tvix"  "uvxy"
```

```
unique(new_DF_close$Symbol)
```

```
## [1] "aezs"  "brk-a" "clbs"  "fbc"   "rgse"  "tvix"
```

```
unique(new_DF_volume$Symbol)
```

```
## [1] "bac"   "brk-b"
```

```
#tvix and uvxy, aezs, brk_a, brk_b, clbs, fbc, rgse and bac did not load properly

#remove tvix and uvxy, aezs, brk_a, brk_b, clbs, fbc, rgse and bac data and reload
DF = DF[!DF$Symbol == "tvix", ]
DF = DF[!DF$Symbol == "uvxy", ]
DF = DF[!DF$Symbol == "aezs", ]
DF = DF[!DF$Symbol == "brk-a", ]
DF = DF[!DF$Symbol == "brk-b", ]
DF = DF[!DF$Symbol == "clbs", ]
DF = DF[!DF$Symbol == "fbc", ]
DF = DF[!DF$Symbol == "rgse", ]
DF = DF[!DF$Symbol == "bac", ]

#reload tvix and uvxy, aezs, brk_a, brk_b, clbs, fbc, rgse and bac
tvix <- read.csv("tvix.us.txt")
```

```
uvxy <- read.csv("uvxy.us.txt")
aezs <- read.csv("aezs.us.txt")
brk_a <- read.csv("brk-a.us.txt")
brk_b <- read.csv("brk-b.us.txt")
clbs <- read.csv("clbs.us.txt")
fbc <- read.csv("fbc.us.txt")
rgse <- read.csv("rgse.us.txt")
bac <- read.csv("bac.us.txt")

#add symbol column
tvix$Symbol <- "tvix"
uvxy$Symbol <- "uvxy"
aezs$Symbol <- "aezs"
brk_a$Symbol <- "brk_a"
brk_b$Symbol <- "brk_b"
clbs$Symbol <- "clbs"
fbc$Symbol <- "fbc"
rgse$Symbol <- "rgse"
bac$Symbol <- "bac"

#attach tvix and uvxy, aezs, brk_a, brk_b, clbs, fbc, rgse and bac to DF_new
DF_new <- rbind(DF, aezs)
DF_new <- rbind(DF_new, brk_a)
DF_new <- rbind(DF_new, brk_b)
DF_new <- rbind(DF_new, clbs)
DF_new <- rbind(DF_new, fbc)
DF_new <- rbind(DF_new, rgse)
DF_new <- rbind(DF_new, bac)
DF_new <- rbind(DF_new, tvix)
DF_new <- rbind(DF_new, uvxy)

#check for NAs
summary(DF_new)
```

```
##     Symbol               Date                 Open
##  Length:17453243    Min.   :1962-01-02   Min.   :0.000e+00
##  Class :character   1st Qu.:2008-05-22   1st Qu.:9.000e+00
##  Mode  :character   Median :2012-06-08   Median :1.800e+01
##                     Mean   :2010-11-12   Mean   :2.625e+04
##                     3rd Qu.:2015-06-23   3rd Qu.:3.400e+01
##                     Max.   :2017-11-10   Max.   :1.424e+09
##       High                Low                 Close
##  Min.   :0.000e+00   Min.   :        -1   Min.   :0.000e+00
##  1st Qu.:9.000e+00   1st Qu.:         9   1st Qu.:9.000e+00
##  Median :1.900e+01   Median :        18   Median :1.800e+01
##  Mean   :2.697e+04   Mean   :     25362   Mean   :2.613e+04
##  3rd Qu.:3.400e+01   3rd Qu.:        33   3rd Qu.:3.400e+01
##  Max.   :1.442e+09   Max.   :1362117844   Max.   :1.438e+09
##      Volume             OpenInt
##  Min.   :0.000e+00   Min.   :0
##  1st Qu.:2.572e+04   1st Qu.:0
##  Median :1.574e+05   Median :0
##  Mean   :1.581e+06   Mean   :0
##  3rd Qu.:7.843e+05   3rd Qu.:0
```

```
##  Max.   :2.424e+09   Max.   :0
```

```
#View Class of each variable
sapply(DF_new, class)
```

```
##      Symbol        Date        Open        High         Low       Close
## "character"      "Date"   "numeric"   "numeric"   "numeric"   "numeric"
##      Volume     OpenInt
##   "numeric"   "integer"
```

```
#Find Stock with -1 Low
DF_new %>% filter(Low == -1)
```

```
## # A tibble: 1 x 8
##   Symbol       Date  Open    High   Low   Close Volume OpenInt
##    <chr>     <date> <dbl>   <dbl> <dbl>   <dbl>  <dbl>   <int>
## 1    hlg 2017-08-10 10.18 863.473    -1 863.473     10       0
```

```
#HLG on 2017-08-10, low should be 10.13, high should be 10.98,
#close should be 10.35 and Volume of 32,600

DF_new %>% filter(Symbol == "hlg") %>% filter(Date >= as.Date("2017-08-1") &
Date <= as.Date("2017-08-20"))
```

```
## # A tibble: 13 x 8
##    Symbol       Date    Open     High     Low   Close Volume OpenInt
##     <chr>     <date>   <dbl>    <dbl>   <dbl>   <dbl>  <dbl>   <int>
##  1    hlg 2017-08-01  9.1800   9.1800  8.9200   9.000   8654       0
##  2    hlg 2017-08-02  9.0000   9.1800  8.8100   9.000  18439       0
##  3    hlg 2017-08-04  8.5501   9.2100  8.5501   9.000  19563       0
##  4    hlg 2017-08-07  9.0000   9.5000  8.8200   9.420  31118       0
##  5    hlg 2017-08-08  9.0800   9.8000  9.0000   9.800  43329       0
##  6    hlg 2017-08-09  9.6600  10.4000  9.6600   9.900  25756       0
##  7    hlg 2017-08-10 10.1800 863.4730 -1.0000 863.473     10       0
##  8    hlg 2017-08-11 10.3510  11.5654 10.3510  11.260  36295       0
##  9    hlg 2017-08-14 13.5100  16.0000 11.6100  13.141  47581       0
## 10    hlg 2017-08-15 13.6500  15.8976 13.6500  15.300 130083       0
## 11    hlg 2017-08-16 16.8100  16.9600 15.4000  16.520 129494       0
## 12    hlg 2017-08-17 16.2700  16.2700 15.0500  15.720  72065       0
## 13    hlg 2017-08-18 15.8000  16.8000 15.8000  16.470  41977       0
```

```
#Replace incorrect HLG data
DF_new$Low[DF_new$Low == "-1"] <- "10.13"
DF_new[DF_new$High == "863.473" &
DF_new$Symbol == "hlg", 4] <- "10.98"
DF_new[DF_new$Close == "863.473" &
DF_new$Symbol == "hlg", 6] <- "10.35"
DF_new[DF_new$Volume == "10" & DF_new$Symbol == "hlg", 7] <- "32600"



#change from scientific notation
options(scipen = 999)

#check summary
summary(DF_new)
```

```
##     Symbol               Date                 Open
##  Length:17453243    Min.   :1962-01-02   Min.   :         0
##  Class :character   1st Qu.:2008-05-22   1st Qu.:         9
##  Mode  :character   Median :2012-06-08   Median :        18
##                     Mean   :2010-11-12   Mean   :     26249
##                     3rd Qu.:2015-06-23   3rd Qu.:        34
##                     Max.   :2017-11-10   Max.   :1423712891
##      High               Low                Close
##  Length:17453243    Length:17453243    Length:17453243
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Volume             OpenInt
##  Length:17453243    Min.   :0
##  Class :character   1st Qu.:0
##  Mode  :character   Median :0
##                     Mean   :0
##                     3rd Qu.:0
##                     Max.   :0
```

```
#Find Stock with 1423712891 Open
DF_new %>% filter(Open == 1423712891) #drys checked on yahoo finance and correct
```

```
## # A tibble: 1 x 8
##   Symbol       Date       Open          High           Low         Close
##    <chr>     <date>      <dbl>         <chr>         <chr>         <chr>
## 1   drys 2007-10-16 1423712891 1432606231.61 1285480853.13 1301071730.91
## # ... with 2 more variables: Volume <chr>, OpenInt <int>
```

```
DF_new %>% filter(Symbol == "drys") %>% filter(Date >= as.Date("2007-10-10") &
Date <= as.Date("2007-10-20"))
```

```
## # A tibble: 8 x 8
##   Symbol       Date       Open          High           Low         Close
##    <chr>     <date>      <dbl>         <chr>         <chr>         <chr>
## 1   drys 2007-10-10 1279222509 1339390221.52 1248040763.03 1294044859.89
## 2   drys 2007-10-11 1331155626 1349381562.73 1225862142.08 1242221549.12
## 3   drys 2007-10-12 1257263491 1338292293.28 1257153694.09 1337743294.16
## 4   drys 2007-10-15 1344989758 1405047674.21 1330826217.17  1391872265.1
## 5   drys 2007-10-16 1423712891 1432606231.61 1285480853.13 1301071730.91
## 6   drys 2007-10-17 1347515096 1354761571.56 1286249382.38 1337084561.35
## 7   drys 2007-10-18 1321383812 1388249062.09 1302828466.77 1385174797.87
## 8   drys 2007-10-19 1385064994 1392640866.74 1255396960.65 1261874897.72
## # ... with 2 more variables: Volume <chr>, OpenInt <int>
```

```
#Find Stock with 2423735131 Volume
DF_new %>% filter(Volume == 2423735131)
```

```
## # A tibble: 1 x 8
##   Symbol       Date   Open   High    Low  Close     Volume OpenInt
##    <chr>     <date>  <dbl>  <chr>  <chr>  <chr>      <chr>   <int>
## 1    bac 2012-03-07 7.4073 7.6065 7.3694 7.6065 2423735131       0
```

```
#View days around that date
DF_new %>% filter(Symbol == "bac") %>% filter(Date >= as.Date("2012-03-01") &
```

```
Date <= as.Date("2012-03-10"))
```

```
## # A tibble: 7 x 8
##   Symbol       Date   Open   High    Low  Close     Volume OpenInt
##    <chr>     <date>  <dbl>  <chr>  <chr>  <chr>      <chr>   <int>
## 1    bac 2012-03-01 7.6730 7.7488 7.6351 7.7012  208370245       0
## 2    bac 2012-03-02 7.6920 7.7871 7.6823 7.7109  151539730       0
## 3    bac 2012-03-05 7.6730 7.7109 7.5401 7.5591  206813942       0
## 4    bac 2012-03-06 7.3789 7.3883  7.265 7.3125  277322560       0
## 5    bac 2012-03-07 7.4073 7.6065 7.3694 7.6065 2423735131       0
## 6    bac 2012-03-08 7.6539 7.7012 7.5874 7.6445  169142238       0
## 7    bac 2012-03-09 7.7012 7.7677 7.5967 7.6351  205637681       0
```

```
#bac checked on yahoo finance and incorrect should be 328,331,900
#replace incorrect volume
DF_new[DF_new$Volume == "2423735131" &
DF_new$Symbol == "bac", 7] <- "328331900"

#Find Stock with 2304018600 Volume
DF_new %>% filter(Volume == 2304018600)  #brk_b
```

```
## # A tibble: 1 x 8
##   Symbol       Date  Open  High   Low Close     Volume OpenInt
##    <chr>     <date> <dbl> <chr> <chr> <chr>      <chr>   <int>
## 1  brk_b 2010-02-11  74.5  76.8 74.15 76.69 2304018600       0
```

```
#View days around that date
DF_new %>% filter(Symbol == "brk_b") %>% filter(Date >= as.Date("2010-02-07") &
Date <= as.Date("2010-02-15"))  #appears correct based on morningstar
```

```
## # A tibble: 5 x 8
##   Symbol       Date  Open  High   Low Close     Volume OpenInt
##    <chr>     <date> <dbl> <chr> <chr> <chr>      <chr>   <int>
## 1  brk_b 2010-02-08 73.99  74.5 72.93 74.23 1199250400       0
## 2  brk_b 2010-02-09 74.72 74.72  73.9 74.53 1356372800       0
## 3  brk_b 2010-02-10 74.54 74.59 74.25 74.42 1379239200       0
## 4  brk_b 2010-02-11 74.50  76.8 74.15 76.69 2304018600       0
## 5  brk_b 2010-02-12 77.15 77.86 74.57  76.9  316134200       0
```

```
# http://performance.morningstar.com/stock/performance-return.action?p=price_history_page&t=BRK.B&regio

#Change columns back to numeric
DF_new$Low <- as.numeric(DF_new$Low)
DF_new$High <- as.numeric(DF_new$High)
DF_new$Close <- as.numeric(DF_new$Close)
DF_new$Volume <- as.numeric(DF_new$Volume)

#check class
sapply(DF_new, class)
```

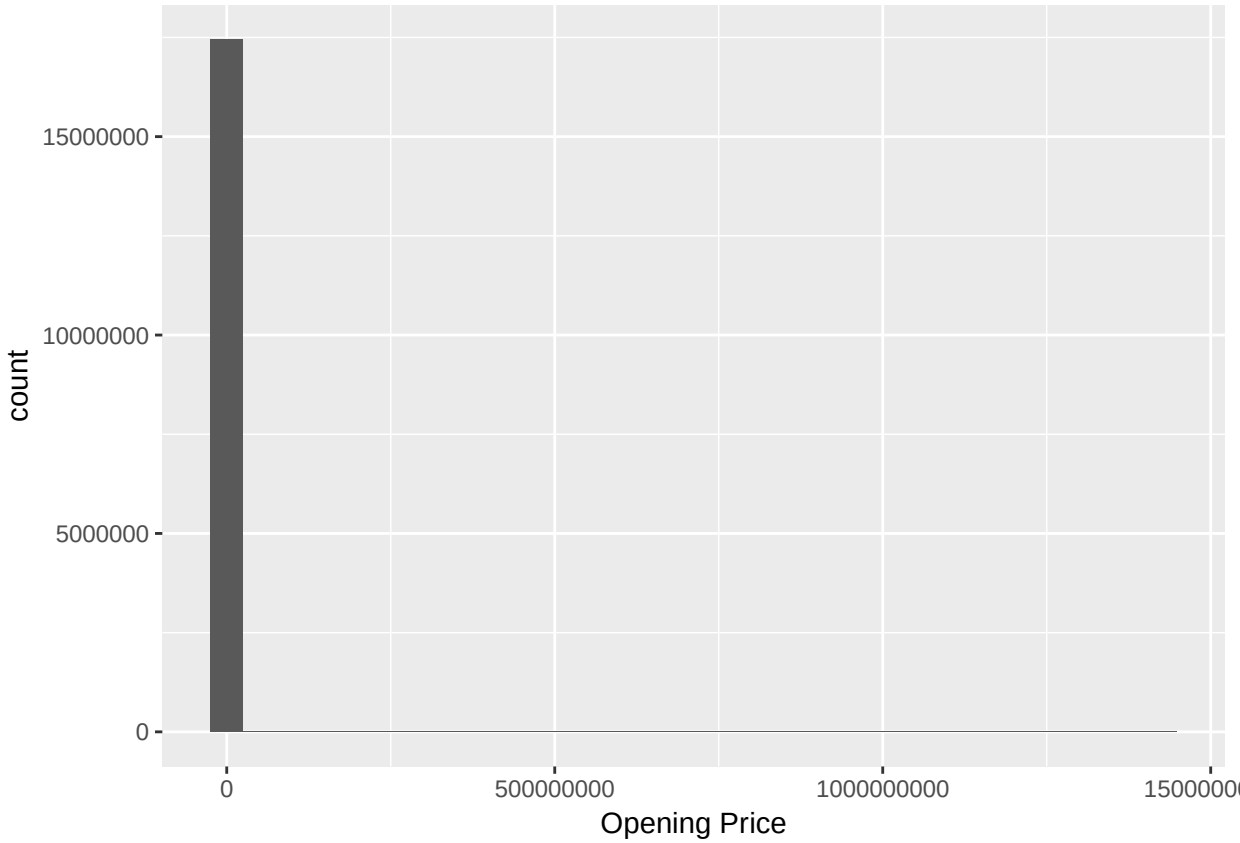
```
##      Symbol        Date        Open        High         Low       Close
## "character"      "Date"   "numeric"   "numeric"   "numeric"   "numeric"
##      Volume     OpenInt
##   "numeric"   "integer"
```

3 Feature Creation

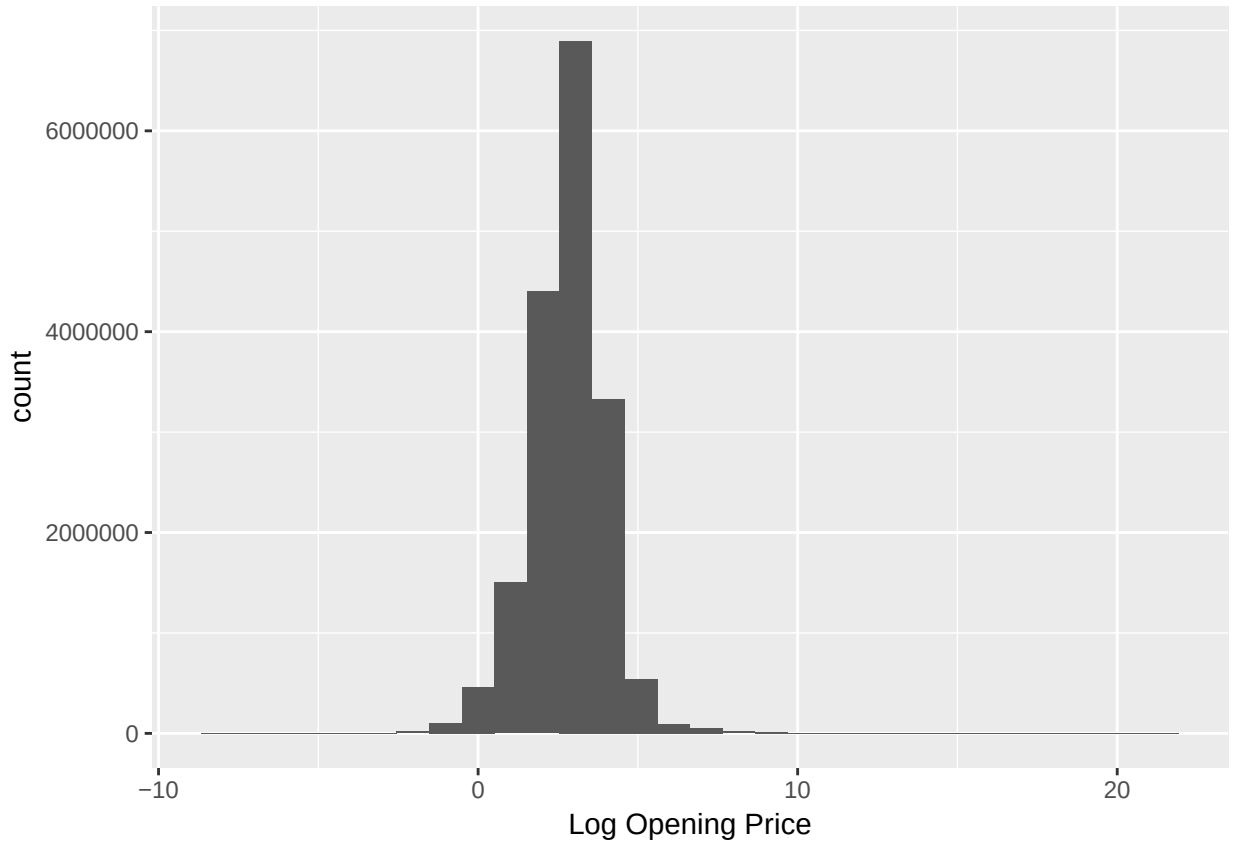
3.1 Log Prices and Volumes

Lead: I'm going to convert stock and ETF prices to log prices and volume to log volume.

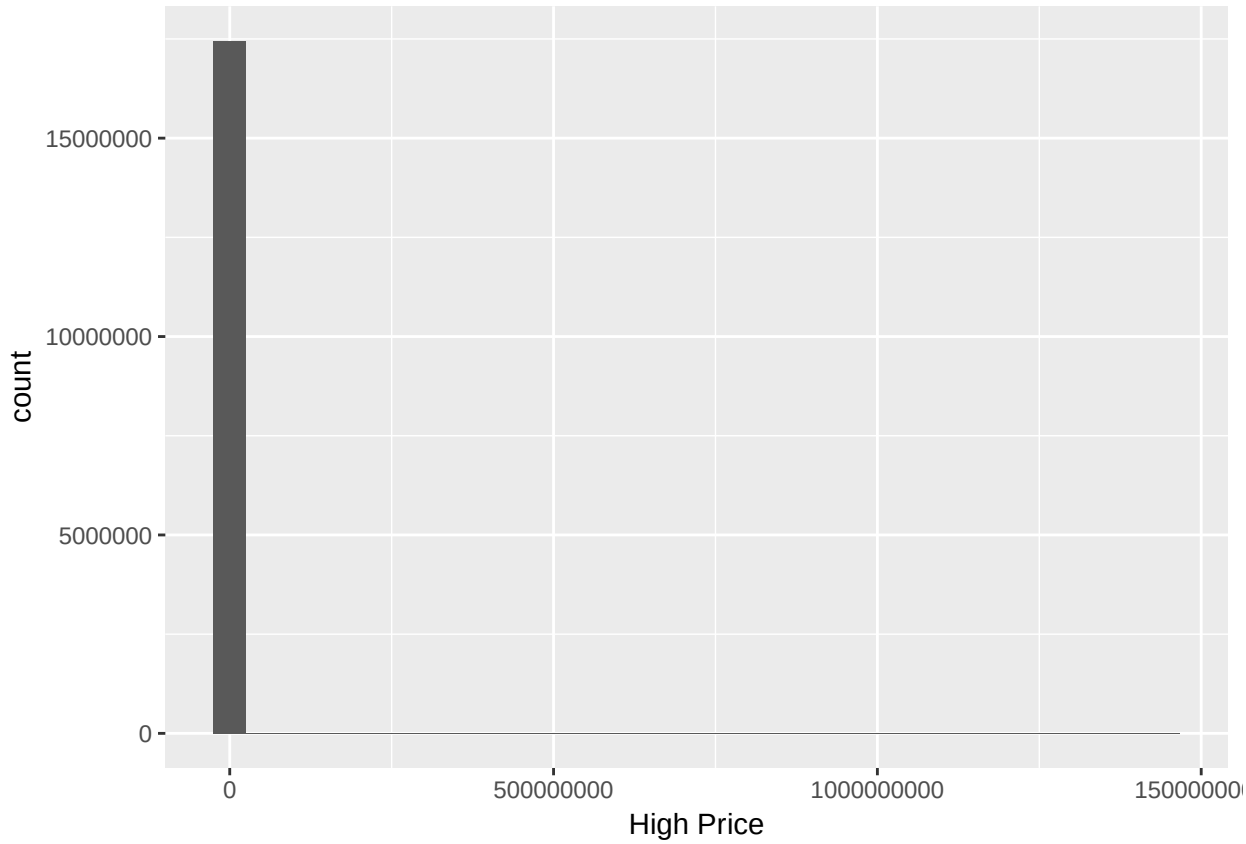
```
#Calculate log prices and volumes and create new columns
DF_new <- DF_new %>%
mutate(
Log_Open = log(Open),
Log_High = log(High),
Log_Low = log(Low),
Log_Close = log(Close),
Log_Volume = log(Volume)
) %>%
ungroup()

#graphically compare Arithmetic vs log prices/volumes
#load ggplot2
library(ggplot2)
#Opening Price Arithmetic
ggplot(data = DF_new, aes(DF_new$Open)) + geom_histogram() + xlab("Opening Price")
```



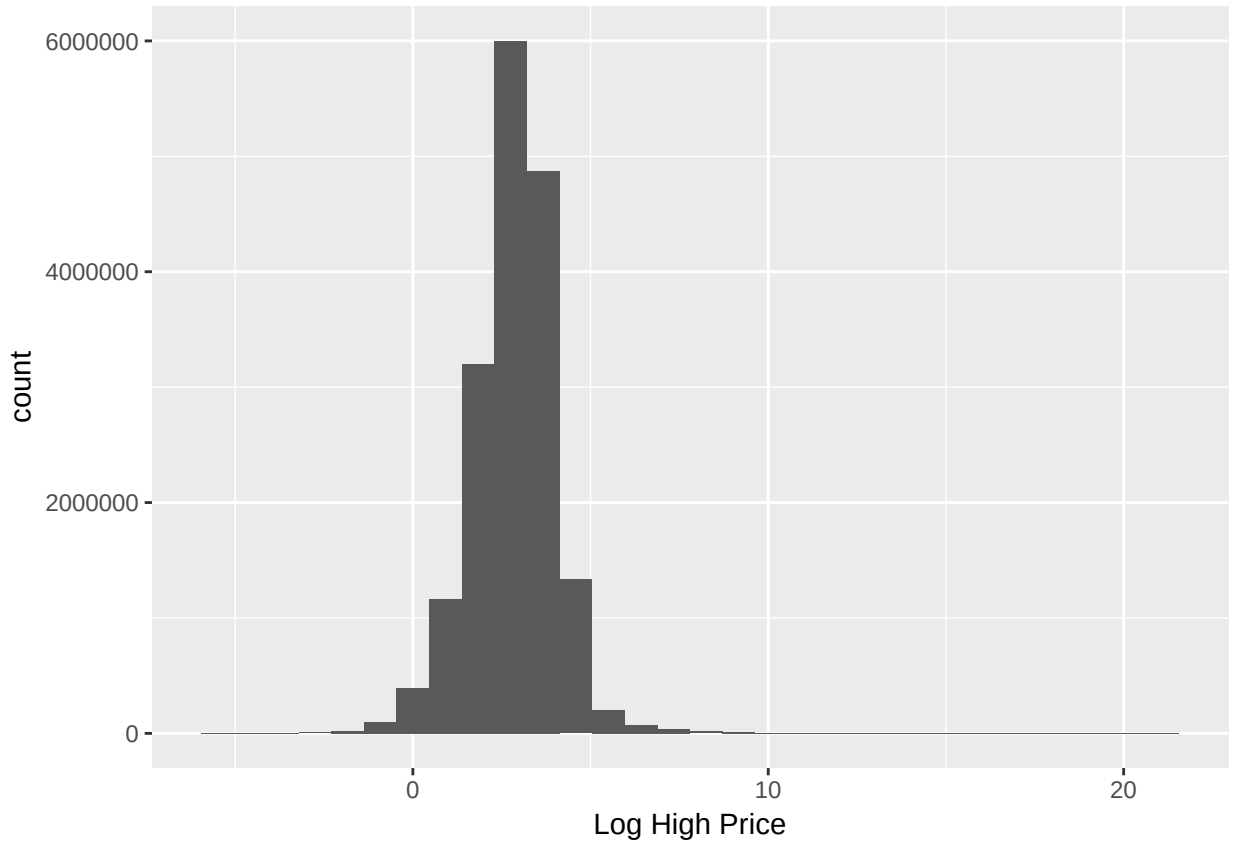

```
#Opening Price log
ggplot(data = DF_new, aes(DF_new$Log_Open)) + geom_histogram() + xlab("Log Opening Price")
```



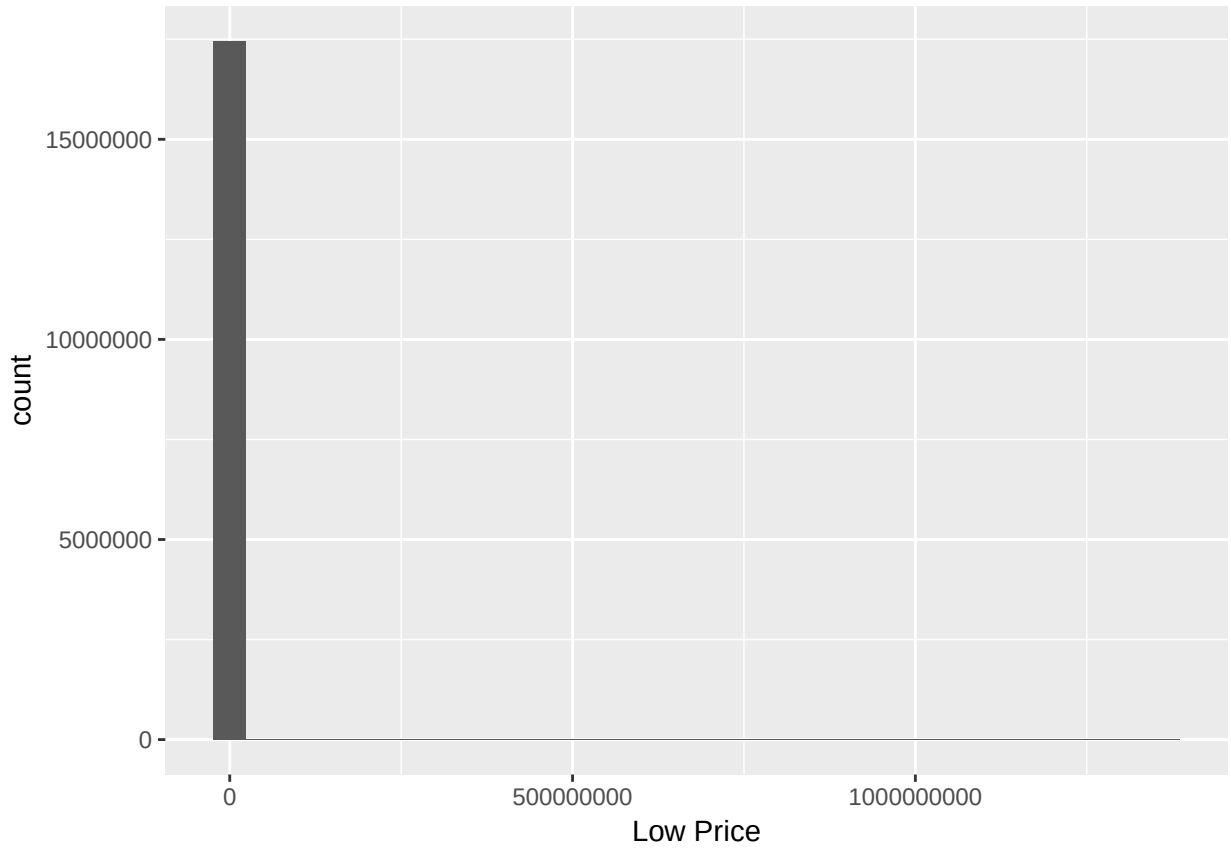

```
#High Price Arithmetic
ggplot(data = DF_new, aes(DF_new$High)) + geom_histogram() + xlab("High Price")
```



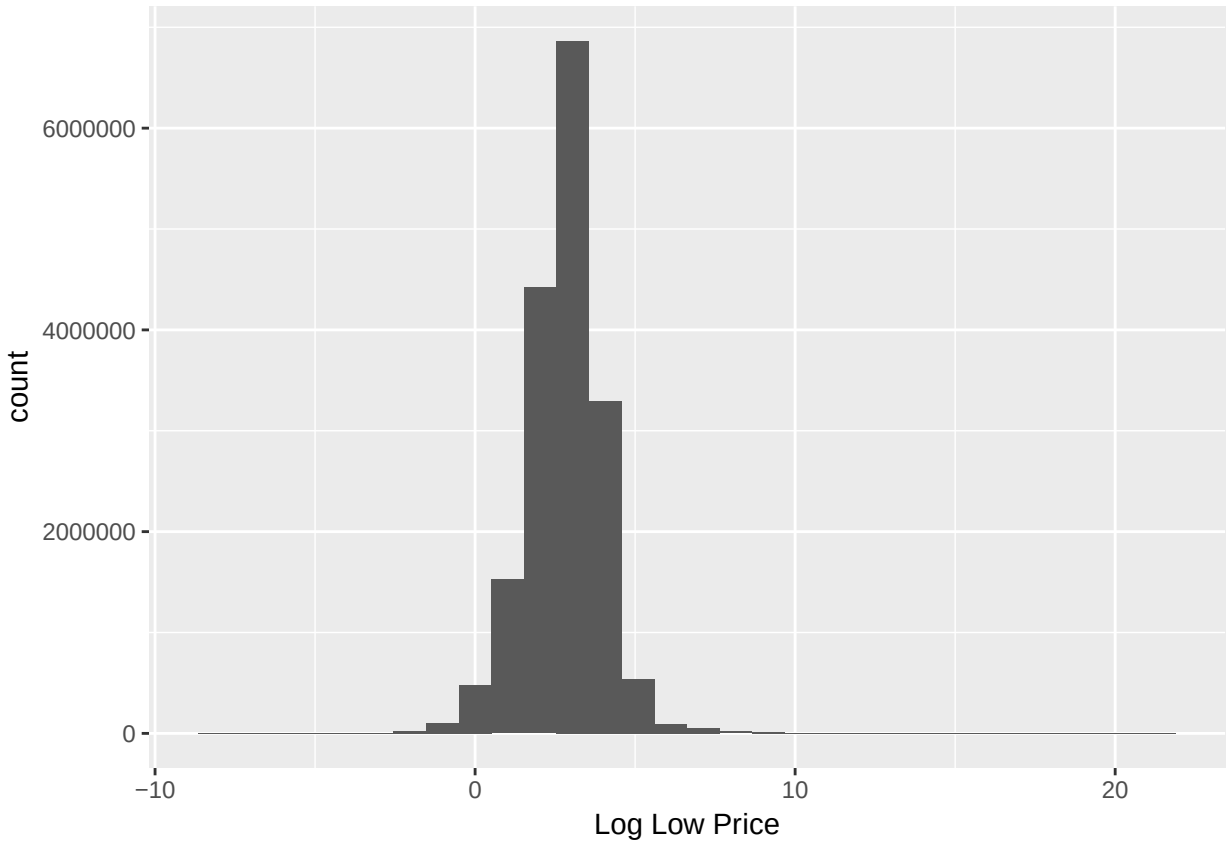

```
#High Price log
ggplot(data = DF_new, aes(DF_new$Log_High)) + geom_histogram() + xlab("Log High Price")
```



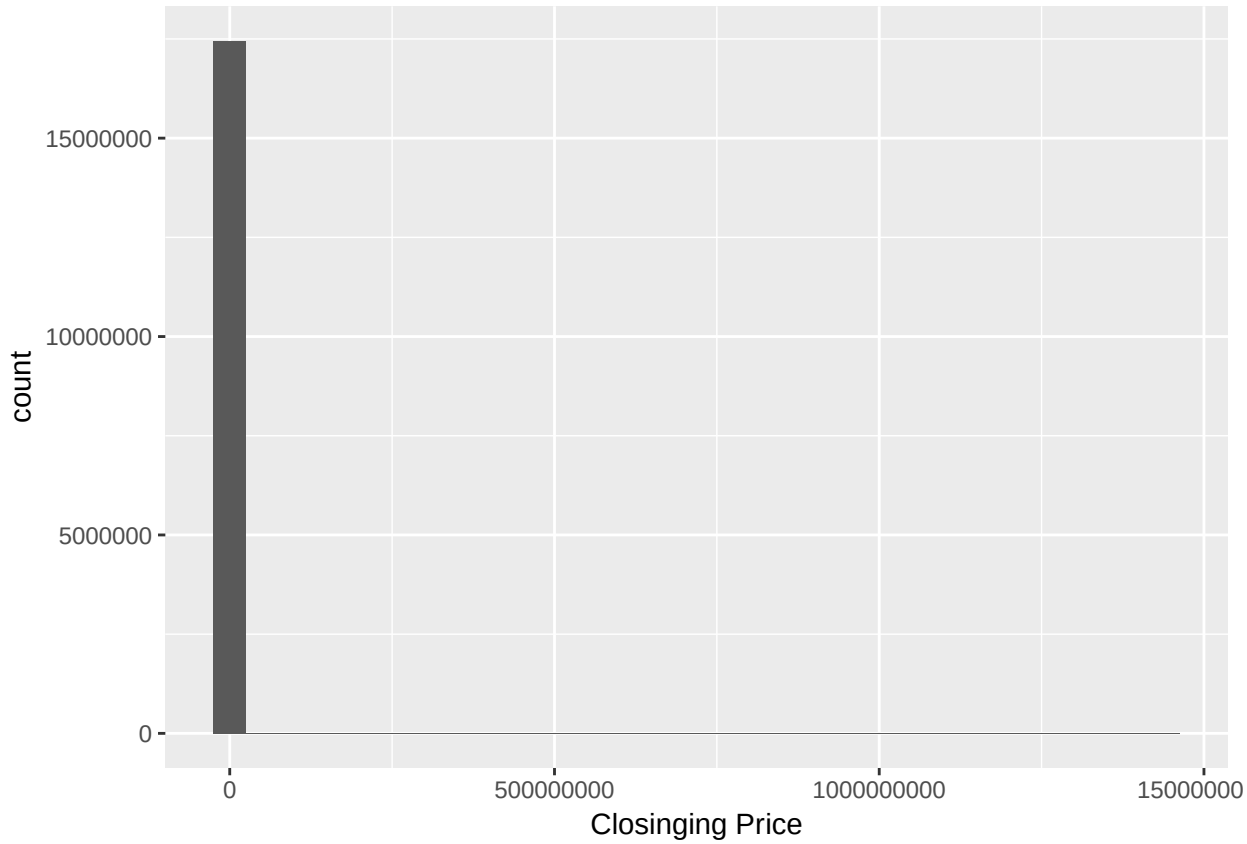

```
#Low Price Arithmetic
ggplot(data = DF_new, aes(DF_new$Low)) + geom_histogram() + xlab("Low Price")
```



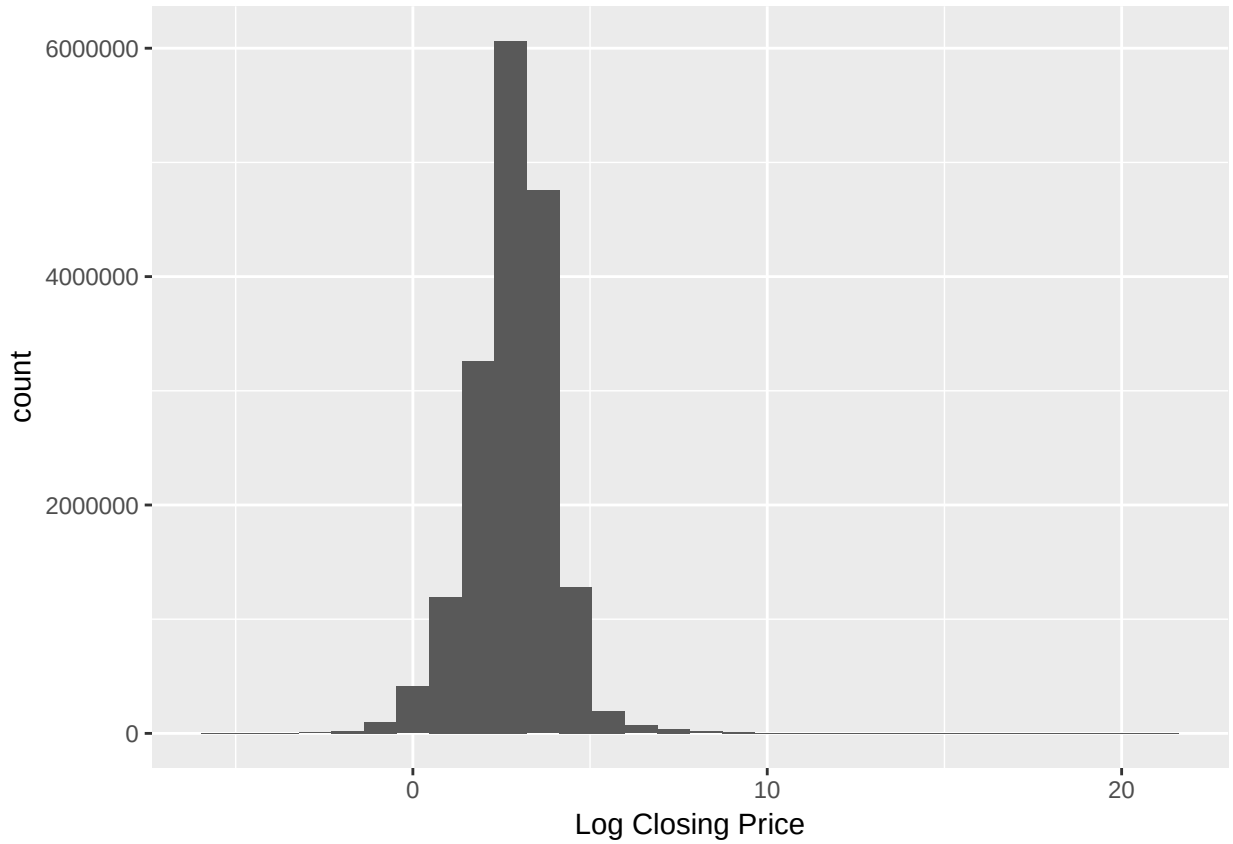

```
#Low Price log
ggplot(data = DF_new, aes(DF_new$Log_Low)) + geom_histogram() + xlab("Log Low Price")
```



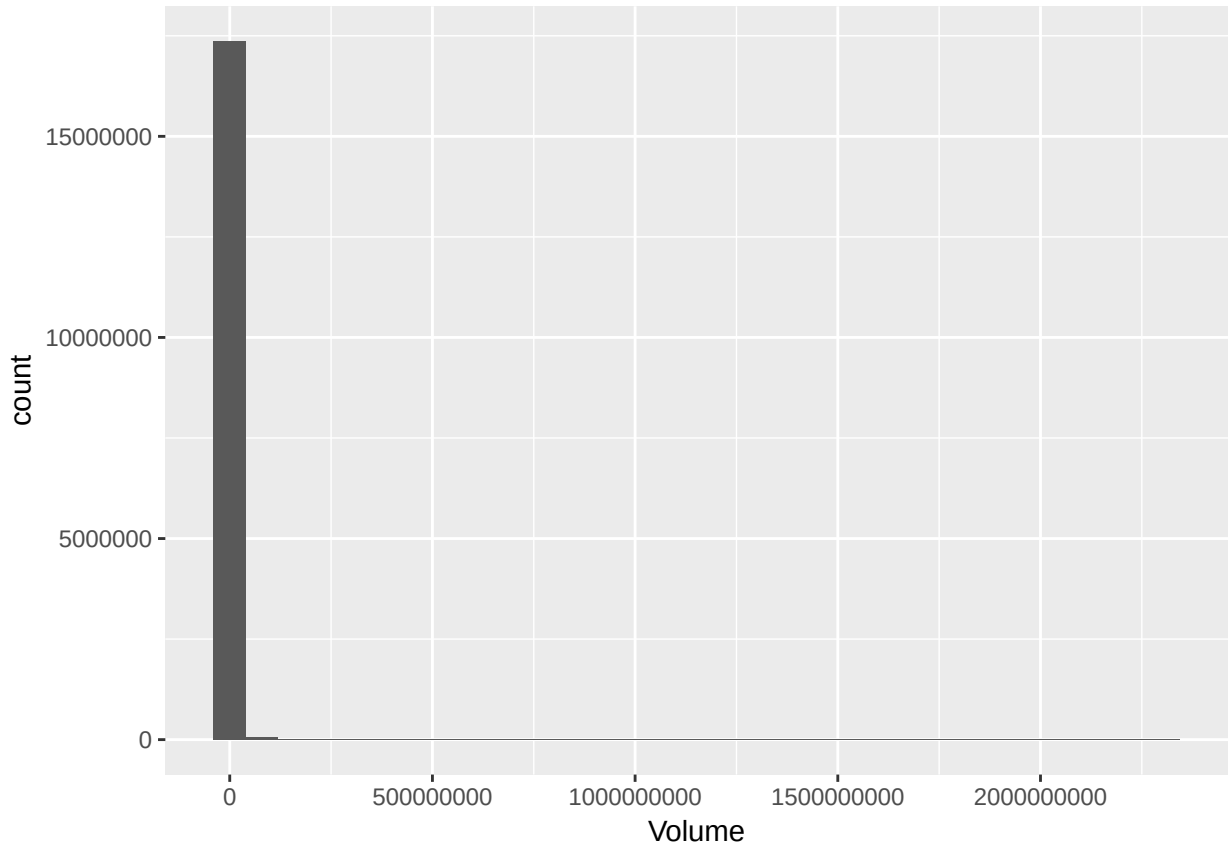

```
#Closing Price Arithmetic
ggplot(data = DF_new, aes(DF_new$Close)) + geom_histogram() + xlab("Closinging Price")
```



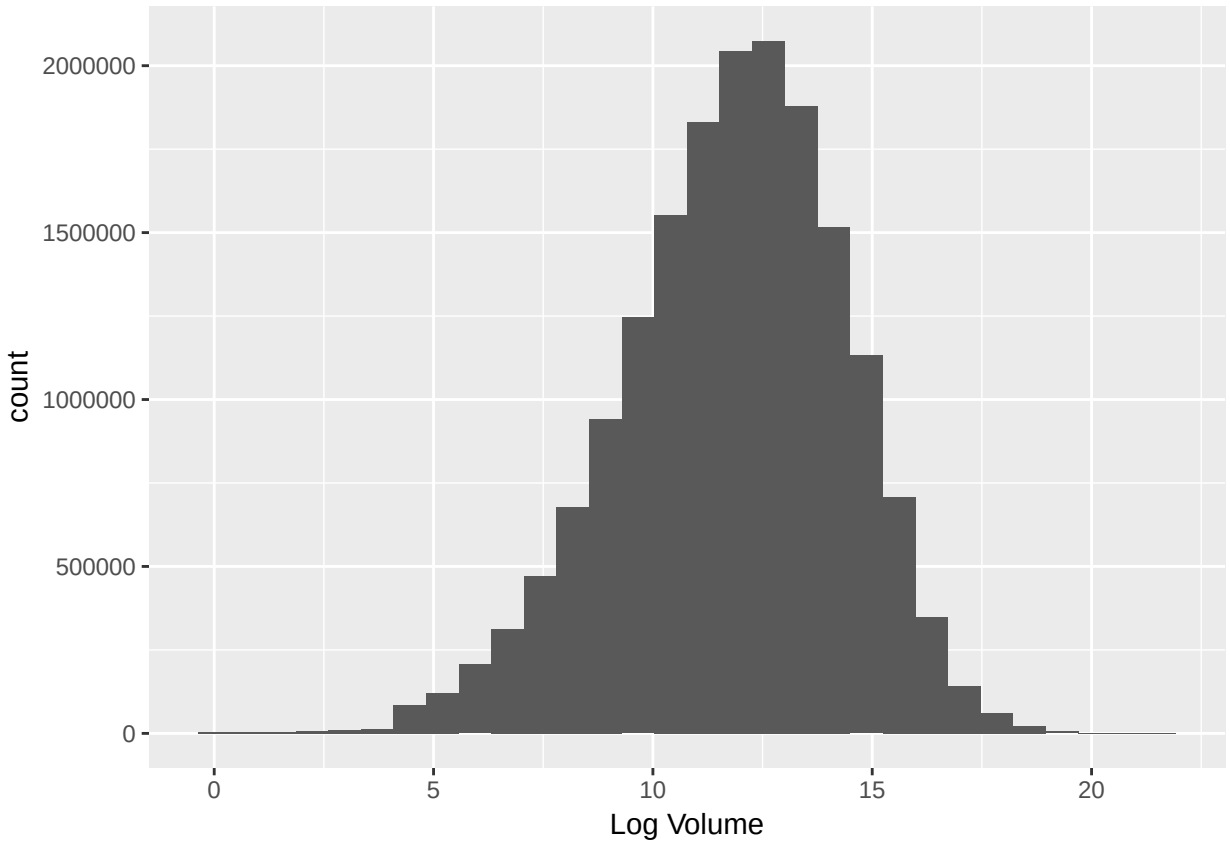

```
#Closing Price log
ggplot(data = DF_new, aes(DF_new$Log_Close)) + geom_histogram() + xlab("Log Closing Price")
```




```
#Volume Arithmetic
ggplot(data = DF_new, aes(DF_new$Volume)) + geom_histogram() + xlab("Volume")
```




```
#Volume log
ggplot(data = DF_new, aes(DF_new$Log_Volume)) + geom_histogram() + xlab("Log Volume")
```



Analysis: As you can see in the above graphs converting to log values reduces the wide range of data into a more manageable size. For example, with arithmetic scaling, all of the metrics were right skewed. For price, DRYS drove this due to the very high stock price. For volume, brk_b drove this due to high volume. Additionally, the log graphs, are log-normally distributed.

Conclusion: Converting price and volume to log scale enabled better visualization of the data. Additionally, by converting to log scale the metrics approximately follow a normal distribution enabling the use of many statistical tests that require the assumption of normality. (Source: http://www.automatedtrader.net/articles/strategies/158619/models-for-daily-and-intra_day-volume-prediction.)

3.2 Log Returns

Lead: I'm going to calculate the daily log returns of each stock and ETF.

```
#Calculate daily log return, add new column for variable
DF_new$Log_Daily_Return <- log(DF_new$Close / DF_new$Open)
#summary
summary(DF_new)
```

```
##     Symbol               Date                 Open
##  Length:17453243    Min.   :1962-01-02   Min.   :         0
##  Class :character   1st Qu.:2008-05-22   1st Qu.:         9
##  Mode  :character   Median :2012-06-08   Median :        18
##                     Mean   :2010-11-12   Mean   :     26249
##                     3rd Qu.:2015-06-23   3rd Qu.:        34
##                     Max.   :2017-11-10   Max.   :1423712891
##       High                 Low                 Close
##  Min.   :         0   Min.   :         0   Min.   :         0
```

```
##  1st Qu.:         9   1st Qu.:         9   1st Qu.:         9
##  Median :        19   Median :        18   Median :        18
##  Mean   :     26971   Mean   :     25362   Mean   :     26128
##  3rd Qu.:        34   3rd Qu.:        33   3rd Qu.:        34
##  Max.   :1442048636   Max.   :1362117844   Max.   :1437986240
##      Volume             OpenInt     Log_Open         Log_High
##  Min.   :         0   Min.   :0   Min.   :  -Inf   Min.   :-5.521
##  1st Qu.:     25717   1st Qu.:0   1st Qu.: 2.183   1st Qu.: 2.197
##  Median :    157428   Median :0   Median : 2.911   Median : 2.925
##  Mean   :   1581166   Mean   :0   Mean   :  -Inf   Mean   : 2.844
##  3rd Qu.:    784313   3rd Qu.:0   3rd Qu.: 3.517   3rd Qu.: 3.529
##  Max.   :2304018600   Max.   :0   Max.   :21.077   Max.   :21.089
##     Log_Low          Log_Close        Log_Volume    Log_Daily_Return
##  Min.   :  -Inf   Min.   :-5.599   Min.   : -Inf   Min.   :-9.692429
##  1st Qu.: 2.166   1st Qu.: 2.183   1st Qu.:10.15   1st Qu.:-0.008434
##  Median : 2.897   Median : 2.911   Median :11.97   Median : 0.000000
##  Mean   :  -Inf   Mean   : 2.829   Mean   : -Inf   Mean   :      Inf
##  3rd Qu.: 3.504   3rd Qu.: 3.517   3rd Qu.:13.57   3rd Qu.: 0.008178
##  Max.   :21.032   Max.   :21.087   Max.   :21.56   Max.   :      Inf
```

```
#find INF results for log return
DF_new %>% filter(Log_Daily_Return == "Inf") #Log Daily Return infinite when opening price = 0
```

```
## # A tibble: 34 x 14
##    Symbol       Date  Open    High   Low   Close Volume OpenInt Log_Open
##     <chr>     <date> <dbl>   <dbl> <dbl>   <dbl>  <dbl>   <int>    <dbl>
##  1   bcom 2011-03-17     0 19.4370     0 19.4370   1269       0     -Inf
##  2    blj 2008-02-07     0 12.1380     0 12.1380      0       0     -Inf
##  3    cwi 2007-01-10     0 29.5820     0 29.5820      0       0     -Inf
##  4    cwi 2007-01-11     0 29.7890     0 29.7890      0       0     -Inf
##  5    cwi 2007-01-12     0 30.1480     0 30.1480      0       0     -Inf
##  6    cwi 2007-01-16     0 30.2460     0 30.2460      0       0     -Inf
##  7    drh 2005-05-25     0  7.8113     0  7.8113      0       0     -Inf
##  8    efv 2005-10-10     0 41.1390     0 41.1390      0       0     -Inf
##  9    emi 2007-03-27     0  9.8480     0  9.8480      0       0     -Inf
## 10    emj 2008-02-05     0 10.4650     0 10.4650      0       0     -Inf
## # ... with 24 more rows, and 5 more variables: Log_High <dbl>,
## #   Log_Low <dbl>, Log_Close <dbl>, Log_Volume <dbl>,
## #   Log_Daily_Return <dbl>
```

```
#check to see if data is correct
#bcom
DF_new %>% filter(Symbol == "bcom") %>% filter(Date >= as.Date("2011-03-10") &
Date <= as.Date("2011-03-20"))
```

```
## # A tibble: 4 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <int>    <dbl>
## 1   bcom 2011-03-10 19.079 19.079 19.079 19.079    483       0 2.948588
## 2   bcom 2011-03-15 18.391 18.391 18.391 18.391    324       0 2.911861
## 3   bcom 2011-03-16 19.419 19.567 18.701 19.301   1938       0 2.966252
## 4   bcom 2011-03-17  0.000 19.437  0.000 19.437   1269       0     -Inf
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```r
#open and low price missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "bcom", 3] <- "19.301"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "bcom", 5] <- "19.301"

#blj
DF_new %>% filter(Symbol == "blj") %>% filter(Date >= as.Date("2008-02-02") &
Date <= as.Date("2008-02-10"))
```

```
## # A tibble: 4 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <dbl>   <int>    <dbl>
## 1    blj 2008-02-04 11.966 11.966 11.966 11.966   3514       0 2.482069
## 2    blj 2008-02-05 11.966 11.966 11.966 11.966   1464       0 2.482069
## 3    blj 2008-02-06 12.066 12.138 12.066 12.138   1464       0 2.490392
## 4    blj 2008-02-07      0 12.138      0 12.138      0       0     -Inf
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```r
#open, low volume missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "blj", 3] <- "12.138"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "blj", 5] <- "12.138"
DF_new[DF_new$Volume == "0" & DF_new$Symbol == "blj", 7] <- "1464"

#cwi
DF_new %>% filter(Symbol == "cwi") %>% filter(Date >= as.Date("2007-01-07") &
Date <= as.Date("2007-01-20"))
```

```
## # A tibble: 7 x 14
##   Symbol       Date  Open   High   Low  Close  Volume OpenInt Log_Open
##    <chr>     <date> <chr>  <dbl> <chr>  <dbl>   <chr>   <int>    <dbl>
## 1    cwi 2007-01-10     0 29.582     0 29.582       0       0     -Inf
## 2    cwi 2007-01-11     0 29.789     0 29.789       0       0     -Inf
## 3    cwi 2007-01-12     0 30.148     0 30.148       0       0     -Inf
## 4    cwi 2007-01-16     0 30.246     0 30.246       0       0     -Inf
## 5    cwi 2007-01-17 30.29 30.622 30.29 30.321 1146241       0 3.410818
## 6    cwi 2007-01-18 30.53 30.530 30.29 30.329  258379       0 3.418710
## 7    cwi 2007-01-19 30.29 30.554 30.29 30.547  252371       0 3.410818
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```r
#open, low volume missing, impute with prior days close price
#didn't start trading until 1/17/2007
#source:https://finance.yahoo.com/quote/CWI/history?period1=1167886800&period2=1169269200&interval=1d&f
#remove rows 2007-01-10 to 2007-01-16
DF_new = DF_new[!(
DF_new$Date >= as.Date("2007-01-10") &
DF_new$Date <= as.Date("2007-01-16") & DF_new$Symbol == "cwi"
), ]

#drh
#didn't start trading until 2005-05-25
#https://finance.yahoo.com/quote/DRH/history?period1=1116561600&period2=1117425600&interval=1d&filter=h
#remove row 2005-05-25
DF_new = DF_new[!(DF_new$Date == as.Date("2005-05-25") &
DF_new$Symbol == "drh"), ]
```

```r
#efv
DF_new %>% filter(Symbol == "efv") %>% filter(Date >= as.Date("2005-10-06") &
Date <= as.Date("2005-10-15"))
```

```
## # A tibble: 7 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    efv 2005-10-06 41.336 41.480 40.927 40.927  31611       0 3.721734
## 2    efv 2005-10-07   41.4 41.400 41.296 41.368  10792       0 3.723281
## 3    efv 2005-10-10      0 41.139      0 41.139      0       0     -Inf
## 4    efv 2005-10-11 41.233 41.439  41.21 41.233  19170       0 3.719239
## 5    efv 2005-10-12 41.288 41.336  40.84 41.014  18156       0 3.720572
## 6    efv 2005-10-13 40.533 40.808 40.416 40.808  19805       0 3.702116
## 7    efv 2005-10-14  40.84 41.139  40.76 41.123  13584       0 3.709662
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```r
#open, low volume missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "efv", 3] <- "41.368"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "efv", 5] <- "41.368"
DF_new[DF_new$Volume == "0" & DF_new$Symbol == "efv", 7] <- "10792"

#emi
DF_new %>% filter(Symbol == "emi") %>% filter(Date >= as.Date("2007-03-25") &
Date <= as.Date("2007-03-30"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date  Open  High   Low Close Volume OpenInt Log_Open
##    <chr>     <date> <chr> <dbl> <chr> <dbl>  <chr>   <int>    <dbl>
## 1    emi 2007-03-26 9.864 9.967 9.837 9.837  31695       0 2.288892
## 2    emi 2007-03-27     0 9.848     0 9.848      0       0     -Inf
## 3    emi 2007-03-28 9.857 9.864 9.857 9.864  14919       0 2.288182
## 4    emi 2007-03-29 9.864 9.880 9.864 9.880   7315       0 2.288892
## 5    emi 2007-03-30  9.88 9.933  9.88 9.933   4877       0 2.290513
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```r
#open, low volume missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "emi", 3] <- "9.837"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "emi", 5] <- "9.837"
DF_new[DF_new$Volume == "0" & DF_new$Symbol == "emi", 7] <- "31695"

#emj
DF_new %>% filter(Symbol == "emj") %>% filter(Date >= as.Date("2008-02-01") &
Date <= as.Date("2008-02-10"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    emj 2008-02-01 10.239 10.341 10.239 10.341   5028       0 2.326204
## 2    emj 2008-02-04 10.362 10.368 10.308 10.308   2872       0 2.338145
## 3    emj 2008-02-05      0 10.465      0 10.465      0       0     -Inf
## 4    emj 2008-02-07 10.434 10.434 10.434 10.434   1444       0 2.345070
## 5    emj 2008-02-08 10.434 10.440 10.434 10.440    870       0 2.345070
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
```

```
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low volume missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "emj", 3] <- "10.308"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "emj", 5] <- "10.308"
DF_new[DF_new$Volume == "0" & DF_new$Symbol == "emj", 7] <- "2872"

#fmo
DF_new %>% filter(Symbol == "fmo") %>% filter(Date >= as.Date("2007-11-01") &
Date <= as.Date("2007-11-07"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    fmo 2007-11-01 13.675 13.771 13.577 13.652 116517       0 2.615569
## 2    fmo 2007-11-02 13.663 13.771 13.598 13.618  72343       0 2.614691
## 3    fmo 2007-11-05      0 13.709      0 13.709  84047       0     -Inf
## 4    fmo 2007-11-06 13.756 13.916 13.657 13.663  65213       0 2.621475
## 5    fmo 2007-11-07 13.512 13.685 13.512 13.680  53389       0 2.603578
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "fmo", 3] <- "13.618"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "fmo", 5] <- "13.618"

#fud
DF_new %>% filter(Symbol == "fud") %>% filter(Date >= as.Date("2008-03-28") &
Date <= as.Date("2008-05-01"))
```

```
## # A tibble: 22 x 14
##    Symbol       Date  Open  High   Low Close Volume OpenInt Log_Open
##     <chr>     <date> <chr> <dbl> <chr> <dbl>  <chr>   <int>    <dbl>
##  1    fud 2008-04-01     0 25.00     0 25.00      0       0     -Inf
##  2    fud 2008-04-03 25.26 25.72 25.26 25.72  14700       0 3.229222
##  3    fud 2008-04-04 25.64 25.64 25.64 25.64    100       0 3.244154
##  4    fud 2008-04-07 26.02 26.12 25.88 25.88   6000       0 3.258865
##  5    fud 2008-04-08 25.98 25.98 25.98 25.98    300       0 3.257327
##  6    fud 2008-04-09 26.52 26.52  26.5 26.50   1500       0 3.277899
##  7    fud 2008-04-10 26.86 26.88 26.46 26.59   4400       0 3.290638
##  8    fud 2008-04-11 26.71 26.71 26.35 26.35  21200       0 3.285038
##  9    fud 2008-04-14 26.51 26.63 26.51 26.62   2069       0 3.277522
## 10    fud 2008-04-15 27.04 27.06 26.95 26.96   3280       0 3.297317
## # ... with 12 more rows, and 5 more variables: Log_High <dbl>,
## #   Log_Low <dbl>, Log_Close <dbl>, Log_Volume <dbl>,
## #   Log_Daily_Return <dbl>
```

```
#didn't start trading until 2008-04-03 source:
#https://finance.yahoo.com/quote/FUD/history?period1=1207022400&period2=1211688000&interval=1d&filter=h

#remove row 2008-04-01
DF_new = DF_new[!(DF_new$Date == as.Date("2008-04-01") &
DF_new$Symbol == "fud"), ]

#gbb
DF_new %>% filter(Symbol == "gbb") %>% filter(Date >= as.Date("2007-05-05") &
```

```
Date <= as.Date("2007-05-10"))
```

```
## # A tibble: 2 x 14
##   Symbol       Date  Open  High   Low Close Volume OpenInt Log_Open
##    <chr>     <date> <chr> <dbl> <chr> <dbl>  <chr>   <int>    <dbl>
## 1    gbb 2007-05-08     0 50.00     0  50.0      0       0     -Inf
## 2    gbb 2007-05-09 50.14 50.14  50.1  50.1    800       0 3.914819
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#remove row 2007-05-08
DF_new = DF_new[!(DF_new$Date == as.Date("2007-05-08") &
DF_new$Symbol == "gbb"), ]

#ite
DF_new %>% filter(Symbol == "ite") %>% filter(Date >= as.Date("2007-05-20") &
Date <= as.Date("2007-06-10"))
```

```
## # A tibble: 7 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    ite 2007-05-23      0 47.615      0 47.615      0       0     -Inf
## 2    ite 2007-05-24      0 47.608      0 47.608      0       0     -Inf
## 3    ite 2007-05-25      0 47.599      0 47.599      0       0     -Inf
## 4    ite 2007-05-29      0 47.564      0 47.564      0       0     -Inf
## 5    ite 2007-05-30 47.705 47.705 47.705 47.705    101       0 3.865036
## 6    ite 2007-06-01 47.502 47.502 47.502 47.502    101       0 3.860772
## 7    ite 2007-06-08  47.32 47.320  47.32 47.320    201       0 3.856933
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#remove rows 2007-01-10 to 2007-01-16
#didn't start trading
#https://finance.yahoo.com/quote/ITE/history?period1=1181448000&period2=1183176000&interval=1d&filter=h
DF_new = DF_new[!(
DF_new$Date >= as.Date("2007-05-23") &
DF_new$Date <= as.Date("2007-05-29") & DF_new$Symbol == "ite"
), ]

#itm
DF_new %>% filter(Symbol == "itm") %>% filter(Date >= as.Date("2009-04-15") &
Date <= as.Date("2009-04-25"))
```

```
## # A tibble: 8 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    itm 2009-04-15 17.496 17.600 17.462 17.540  36672       0 2.861972
## 2    itm 2009-04-16 17.531 17.660 17.531 17.660  25335       0 2.863971
## 3    itm 2009-04-17 17.675 17.691 17.608 17.608  11968       0 2.872151
## 4    itm 2009-04-20      0 17.744      0 17.717  13497       0     -Inf
## 5    itm 2009-04-21 17.767 17.825 17.767 17.825  17004       0 2.877343
## 6    itm 2009-04-22 17.799 17.832 17.675 17.825  55955       0 2.879142
## 7    itm 2009-04-23 17.842 17.866 17.774 17.816  28822       0 2.881555
## 8    itm 2009-04-24 17.784 17.792 17.744 17.774  25364       0 2.878299
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "itm", 3] <- "17.608"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "itm", 5] <- "17.608"

#mln
DF_new %>% filter(Symbol == "mln") %>% filter(Date >= as.Date("2009-04-15") &
Date <= as.Date("2009-04-30"))
```

```
## # A tibble: 12 x 14
##    Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##     <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
##  1    mln 2009-04-15 13.325 13.362 13.301 13.354   2809       0 2.589642
##  2    mln 2009-04-16 13.484 13.484 13.309 13.397  18041       0 2.601504
##  3    mln 2009-04-17 13.516 13.581 13.397 13.397  11606       0 2.603874
##  4    mln 2009-04-20      0 13.557      0 13.430 550823       0     -Inf
##  5    mln 2009-04-21  13.54 13.683 13.516 13.566   9397       0 2.605648
##  6    mln 2009-04-22 13.612 13.675 13.612 13.659   5092       0 2.610952
##  7    mln 2009-04-23 13.659 13.683 13.621 13.683   8277       0 2.614399
##  8    mln 2009-04-24 13.629 13.683 13.629 13.675   5197       0 2.612200
##  9    mln 2009-04-27 13.645 13.712 13.581 13.705  35350       0 2.613373
## 10    mln 2009-04-28 13.636 13.712 13.636 13.712  10321       0 2.612713
## 11    mln 2009-04-29 13.683 13.692 13.557 13.575  26159       0 2.616154
## 12    mln 2009-04-30 13.588 13.629 13.454 13.548  21151       0 2.609187
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "mln", 3] <- "13.397"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "mln", 5] <- "13.397"

#nom
DF_new %>% filter(Symbol == "nom") %>% filter(Date >= as.Date("2008-02-02") &
Date <= as.Date("2008-02-10"))
```

```
## # A tibble: 3 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    nom 2008-02-04 11.246 11.246 11.246 11.246    100       0 2.420013
## 2    nom 2008-02-05      0 11.246      0 11.246      0       0     -Inf
## 3    nom 2008-02-08 11.281 11.310 11.281 11.310   1223       0 2.423120
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low volume missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "nom", 3] <- "11.246"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "nom", 5] <- "11.246"
DF_new[DF_new$Volume == "0" & DF_new$Symbol == "nom", 7] <- "100"

#smb
DF_new %>% filter(Symbol == "smb") %>% filter(Date >= as.Date("2009-04-15") &
Date <= as.Date("2009-04-25"))
```

```
## # A tibble: 8 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    smb 2009-04-15 15.427 15.427  15.39 15.399   6239       0 2.736119
```

```
## 2    smb 2009-04-16 15.417 15.435 15.399 15.435   5350       0 2.735471
## 3    smb 2009-04-17 15.435 15.489 15.435 15.489  19768       0 2.736638
## 4    smb 2009-04-20      0 15.508      0 15.444  14831       0     -Inf
## 5    smb 2009-04-21 15.499 15.527 15.499 15.527  23145       0 2.740776
## 6    smb 2009-04-22 15.554 15.564 15.527 15.554  13169       0 2.744318
## 7    smb 2009-04-23 15.536 15.573 15.508 15.554  14018       0 2.743160
## 8    smb 2009-04-24 15.536 15.536 15.453 15.518  48165       0 2.743160
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low missing, impute with prior days close price
DF_new[DF_new$Open == "0" & DF_new$Symbol == "smb", 3] <- "15.489"
DF_new[DF_new$Low == "0" & DF_new$Symbol == "smb", 5] <- "15.489"

#spab
DF_new %>% filter(Symbol == "spab") %>% filter(Date >= as.Date("2007-05-20") &
Date <= as.Date("2007-05-30"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1   spab 2007-05-23      0 21.604      0 21.604      0       0     -Inf
## 2   spab 2007-05-24      0 21.604      0 21.604      0       0     -Inf
## 3   spab 2007-05-25      0 21.597      0 21.597      0       0     -Inf
## 4   spab 2007-05-29      0 21.575      0 21.575      0       0     -Inf
## 5   spab 2007-05-30 21.655 21.655 21.655 21.655   7032       0 3.075236
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#remove rows 2007-05-23 to 2007-05-29
DF_new = DF_new[!(
DF_new$Date >= as.Date("2007-05-23") &
DF_new$Date <= as.Date("2007-05-29") & DF_new$Symbol == "spab"
), ]

#sptl
DF_new %>% filter(Symbol == "sptl") %>% filter(Date >= as.Date("2007-05-20") &
Date <= as.Date("2007-05-30"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date  Open   High   Low  Close Volume OpenInt Log_Open
##    <chr>     <date> <chr>  <dbl> <chr>  <dbl>  <chr>   <int>    <dbl>
## 1   sptl 2007-05-23     0 20.990     0 20.990      0       0     -Inf
## 2   sptl 2007-05-24     0 20.998     0 20.998      0       0     -Inf
## 3   sptl 2007-05-25     0 20.990     0 20.990      0       0     -Inf
## 4   sptl 2007-05-29     0 21.060     0 21.060      0       0     -Inf
## 5   sptl 2007-05-30 21.06 21.060 21.06 21.060    200       0 3.047376
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#remove rows 2007-05-23 to 2007-05-29
DF_new = DF_new[!(
DF_new$Date >= as.Date("2007-05-23") &
DF_new$Date <= as.Date("2007-05-29") & DF_new$Symbol == "sptl"
), ]
```

```
#uci
DF_new %>% filter(Symbol == "uci") %>% filter(Date >= as.Date("2008-03-25") &
Date <= as.Date("2008-04-05"))
```

```
## # A tibble: 4 x 14
##   Symbol       Date  Open  High   Low Close Volume OpenInt Log_Open
##    <chr>     <date> <chr> <dbl> <chr> <dbl>  <chr>   <int>    <dbl>
## 1    uci 2008-04-01     0 25.00     0 25.00      0       0     -Inf
## 2    uci 2008-04-02 25.02 25.02 25.02 25.02    100       0 3.219676
## 3    uci 2008-04-03 25.36 25.54 25.36 25.43    800       0 3.233173
## 4    uci 2008-04-04 25.46 25.46 25.46 25.46    100       0 3.237109
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#didn't start trading until 2008-04-02
#https://finance.yahoo.com/quote/UCI/history?period1=1205553600&period2=1207195200&interval=1d&filter=h
#remove row 2008-04-01
DF_new = DF_new[!(DF_new$Date == as.Date("2008-04-01") &
DF_new$Symbol == "uci"), ]

#usv
DF_new %>% filter(Symbol == "usv") %>% filter(Date >= as.Date("2008-03-25") &
Date <= as.Date("2008-04-05"))
```

```
## # A tibble: 3 x 14
##   Symbol       Date  Open  High   Low Close Volume OpenInt Log_Open
##    <chr>     <date> <chr> <dbl> <chr> <dbl>  <chr>   <int>    <dbl>
## 1    usv 2008-04-01     0 25.00     0 25.00      0       0     -Inf
## 2    usv 2008-04-03 25.79 25.79 25.79 25.79    100       0 3.249987
## 3    usv 2008-04-04 25.91 25.91 25.91 25.91    100       0 3.254629
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#remove row 2008-04-01
DF_new = DF_new[!(DF_new$Date == as.Date("2008-04-01") &
DF_new$Symbol == "usv"), ]

#vxz
DF_new %>% filter(Symbol == "vxz") %>% filter(Date >= as.Date("2009-01-28") &
Date <= as.Date("2009-02-05"))
```

```
## # A tibble: 6 x 14
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>  <dbl>  <chr>   <int>    <dbl>
## 1    vxz 2009-01-29      0 400.00      0 400.00      0       0     -Inf
## 2    vxz 2009-01-30 400.44 414.00 399.32 413.56  18475       0 5.992564
## 3    vxz 2009-02-02    420 420.00 415.48 416.84  21049       0 6.040255
## 4    vxz 2009-02-03  416.2 416.84 413.32 413.68  40500       0 6.031166
## 5    vxz 2009-02-04    410 412.84 407.44 410.24  24075       0 6.016157
## 6    vxz 2009-02-05 412.32 413.20 406.56 406.72   8375       0 6.021800
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#remove row 2009-01-29
DF_new = DF_new[!(DF_new$Date == as.Date("2009-01-29") &
DF_new$Symbol == "vxz"), ]
```

```
#rwx
#didn't start trading until 12/19/2006
#source: http://www.kibot.com/Historical_Data/All_Stocks_And_ETFs_Historical_Intraday_Data.aspx
#remove rows 2006-12-15 to 2006-12-18
DF_new = DF_new[!(
DF_new$Date >= as.Date("2006-12-15") &
DF_new$Date <= as.Date("2006-12-18") & DF_new$Symbol == "rwx"
), ]

#summary
summary(DF_new)
```

```
##     Symbol               Date                Open
##  Length:17453219    Min.   :1962-01-02   Length:17453219
##  Class :character   1st Qu.:2008-05-22   Class :character
##  Mode  :character   Median :2012-06-08   Mode  :character
##                     Mean   :2010-11-12
##                     3rd Qu.:2015-06-23
##                     Max.   :2017-11-10
##       High                Low                Close
##  Min.   :         0   Length:17453219    Min.   :         0
##  1st Qu.:         9   Class :character   1st Qu.:         9
##  Median :        19   Mode  :character   Median :        18
##  Mean   :     26971                      Mean   :     26128
##  3rd Qu.:        34                      3rd Qu.:        34
##  Max.   :1442048636                      Max.   :1437986240
##     Volume             OpenInt     Log_Open          Log_High
##  Length:17453219    Min.   :0   Min.   :  -Inf   Min.   :-5.521
##  Class :character   1st Qu.:0   1st Qu.: 2.183   1st Qu.: 2.197
##  Mode  :character   Median :0   Median : 2.911   Median : 2.925
##                     Mean   :0   Mean   :  -Inf   Mean   : 2.844
##                     3rd Qu.:0   3rd Qu.: 3.517   3rd Qu.: 3.529
##                     Max.   :0   Max.   :21.077   Max.   :21.089
##     Log_Low          Log_Close        Log_Volume    Log_Daily_Return
##  Min.   :  -Inf   Min.   :-5.599   Min.   : -Inf   Min.   :-9.692429
##  1st Qu.: 2.166   1st Qu.: 2.183   1st Qu.:10.15   1st Qu.:-0.008434
##  Median : 2.897   Median : 2.911   Median :11.97   Median : 0.000000
##  Mean   :  -Inf   Mean   : 2.829   Mean   : -Inf   Mean   :      Inf
##  3rd Qu.: 3.504   3rd Qu.: 3.517   3rd Qu.:13.57   3rd Qu.: 0.008178
##  Max.   :21.032   Max.   :21.087   Max.   :21.56   Max.   :      Inf
```

```
#Check very large returns to make sure they are correct
#filter for log price changes greater than or less than 4%
DF_change <- DF_new %>% filter(!between(Log_Daily_Return, -4, 4))
View(DF_change)

#ako-a
#View days around that date
DF_new %>% filter(Symbol == "ako-a") %>% filter(Date >= as.Date("2010-08-25") &
Date <= as.Date("2010-09-04")) #incorrect open and low prices
```

```
## # A tibble: 8 x 14
##   Symbol       Date    Open   High     Low  Close Volume OpenInt  Log_Open
##    <chr>     <date>   <chr>  <dbl>   <chr>  <dbl>  <chr>   <int>     <dbl>
```

```
## 1  ako-a 2010-08-25  21.016 21.076  20.858 20.858   2383       0  3.045284
## 2  ako-a 2010-08-26  21.086 21.316  21.086 21.316   7368       0  3.048609
## 3  ako-a 2010-08-27    21.3 21.464  21.208 21.429   2493       0  3.058707
## 4  ako-a 2010-08-30  21.226 21.827  21.226 21.540   6677       0  3.055227
## 5  ako-a 2010-08-31   21.53 21.779  21.483 21.622   2383       0  3.069447
## 6  ako-a 2010-09-01 0.00924 22.021 0.00924 21.686   2954       0 -4.684213
## 7  ako-a 2010-09-02  21.686 21.733  21.393 21.669   4333       0  3.076667
## 8  ako-a 2010-09-03  21.889 21.889  21.242 21.272  22199       0  3.085984
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low wrong, impute with prior days close price
DF_new[DF_new$Open == "0.00924" &
DF_new$Symbol == "ako-a", 3] <- "21.622"
DF_new[DF_new$Low == "0.00924" &
DF_new$Symbol == "ako-a", 5] <- "21.622"

#avb
DF_new %>% filter(Symbol == "avb") %>% filter(Date >= as.Date("2010-09-12") &
Date <= as.Date("2010-09-19"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date   Open   High    Low   Close  Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <dbl>  <chr>   <dbl>   <chr>   <int>    <dbl>
## 1    avb 2010-09-13  87.83 88.766 87.204 88.7300 1170144       0 4.475403
## 2    avb 2010-09-14 88.175 89.674 87.547 89.1140 1073190       0 4.479323
## 3    avb 2010-09-15 88.989 90.951  88.89 90.8430 1125083       0 4.488513
## 4    avb 2010-09-16 90.616 90.843 89.844  0.8997  886562       0 4.506631
## 5    avb 2010-09-17  90.39 90.481 89.213 89.2840 1398547       0 4.504134
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#closing price wrong, change to 87 (source: yahoo finance)
DF_new[DF_new$Close == "0.8997" & DF_new$Symbol == "avb", 6] <- "87"

#bofil
DF_new %>% filter(Symbol == "bofil") %>% filter(Date >= as.Date("2016-03-07") &
Date <= as.Date("2016-03-12"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date      Open       High    Low  Close Volume OpenInt
##    <chr>     <date>     <chr>      <dbl>  <chr>  <chr>  <chr>   <int>
## 1  bofil 2016-03-07    22.497     22.507 22.379 22.497  21556       0
## 2  bofil 2016-03-08    22.507     22.516 22.315 22.516  42046       0
## 3  bofil 2016-03-09 182681.41 182681.410 22.342 22.487  30671       0
## 4  bofil 2016-03-10    22.762     23.036  22.47 22.561   7784       0
## 5  bofil 2016-03-11    22.552     22.607 22.516 22.589  28816       0
## # ... with 6 more variables: Log_Open <dbl>, Log_High <dbl>,
## #   Log_Low <dbl>, Log_Close <dbl>, Log_Volume <dbl>,
## #   Log_Daily_Return <dbl>
```

```
#open, high wrong, impute with prior days close price
DF_new[DF_new$Open == "182681.41" &
DF_new$Symbol == "bofil", 3] <- "22.516"
DF_new[DF_new$High == "182681.41" &
DF_new$Symbol == "bofil", 4] <- "22.516"
```

```r
#bxp
DF_new %>% filter(Symbol == "bxp") %>% filter(Date >= as.Date("2007-04-30") &
Date <= as.Date("2007-05-10"))
```

```
## # A tibble: 9 x 14
##   Symbol       Date   Open   High    Low  Close  Volume OpenInt   Log_Open
##    <chr>     <date>  <chr>  <chr>  <chr>  <chr>   <chr>   <int>      <dbl>
## 1    bxp 2007-04-30 89.164  90.69 89.164 89.226 1248300       0  4.4904774
## 2    bxp 2007-05-01 89.353 89.568 87.651 88.504 1356145       0  4.4925948
## 3    bxp 2007-05-02 88.168 89.155 87.499 0.8874 1034628       0  4.4792441
## 4    bxp 2007-05-03 89.065 89.335 87.613 89.011  792930       0  4.4893664
## 5    bxp 2007-05-04 89.206 89.226 87.445 87.773  748163       0  4.4909483
## 6    bxp 2007-05-07 0.8874 89.065 87.895 88.206  886684       0 -0.1194594
## 7    bxp 2007-05-08 88.134 88.299 87.436 87.436  727205       0  4.4788584
## 8    bxp 2007-05-09 87.606 89.497 87.541 89.441 1106182       0  4.4728495
## 9    bxp 2007-05-10 88.805 89.261 88.159 88.214 1337922       0  4.4864430
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```r
#close wrong, should be 87.88 (source: yahoo finance)
DF_new[DF_new$Close == "0.8874" &
DF_new$Symbol == "bxp", 6] <- "87.88"
#open wrong, should be 87.35 (source: yahoo finance)
DF_new[DF_new$Open == "0.8874" &
DF_new$Symbol == "bxp", 3] <- "87.35"

#cbmxw
DF_new %>% filter(Symbol == "cbmxw") %>% filter(Date >= as.Date("2016-05-28") &
Date <= as.Date("2016-06-3")) #correct
```

```
## # A tibble: 3 x 14
##   Symbol       Date  Open  High    Low  Close Volume OpenInt   Log_Open
##    <chr>     <date> <chr> <chr>  <chr>  <chr>  <chr>   <int>      <dbl>
## 1  cbmxw 2016-05-31 0.011  0.84  0.011   0.84  14793       0 -4.5098600
## 2  cbmxw 2016-06-01  0.84  1.12   0.84   1.12  35100       0 -0.1743534
## 3  cbmxw 2016-06-03  1.12  1.12 1.1127 1.1127   1907       0  0.1133287
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```r
#fud
DF_new %>% filter(Symbol == "fud") %>% filter(Date >= as.Date("2008-10-06") &
Date <= as.Date("2008-10-11"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date  Open  High   Low Close Volume OpenInt  Log_Open
##    <chr>     <date> <chr> <chr> <chr> <chr>  <chr>   <int>     <dbl>
## 1    fud 2008-10-06 20.38 20.45 19.16 19.16   2700       0  3.014554
## 2    fud 2008-10-07 19.85 19.85 18.11 18.16   3351       0  2.988204
## 3    fud 2008-10-08   0.3 17.65   0.3 17.39   4795       0 -1.203973
## 4    fud 2008-10-09 17.67 17.67 16.57 16.57   1483       0  2.871868
## 5    fud 2008-10-10  15.9 16.19 15.25 16.19    579       0  2.766319
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low wrong, fix source: yahoo finance
DF_new[DF_new$Open == "0.3" & DF_new$Symbol == "fud", 3] <- "17.65"
DF_new[DF_new$Low == "0.3" & DF_new$Symbol == "fud", 5] <- "17.39"

#gcbc
DF_new %>% filter(Symbol == "gcbc") %>% filter(Date >= as.Date("2016-04-25") &
Date <= as.Date("2016-04-30"))
```

```
## # A tibble: 5 x 14
##   Symbol       Date     Open     High    Low  Close Volume OpenInt
##    <chr>     <date>    <chr>    <chr>  <chr>  <chr>  <chr>   <int>
## 1   gcbc 2016-04-25   17.454   17.491  17.17 17.491   2481       0
## 2   gcbc 2016-04-26   17.491   17.997 17.491 17.997   1100       0
## 3   gcbc 2016-04-27 77739.67 77739.67 17.268 17.439  10481       0
## 4   gcbc 2016-04-28    17.18   17.491  17.18 17.491   1133       0
## 5   gcbc 2016-04-29   17.491   17.491 17.268 17.394    741       0
## # ... with 6 more variables: Log_Open <dbl>, Log_High <dbl>,
## #   Log_Low <dbl>, Log_Close <dbl>, Log_Volume <dbl>,
## #   Log_Daily_Return <dbl>
```

```
#open, high wrong source:yahoo finance
DF_new[DF_new$Open == "77739.67" &
DF_new$Symbol == "gcbc", 3] <- "17.81"
DF_new[DF_new$High == "77739.67" &
DF_new$Symbol == "gcbc", 4] <- "19"

#phii
DF_new %>% filter(Symbol == "phii") %>% filter(Date >= as.Date("2016-04-12") &
Date <= as.Date("2016-04-22"))
```

```
## # A tibble: 8 x 14
##   Symbol       Date     Open     High     Low   Close Volume OpenInt
##    <chr>     <date>    <chr>    <chr>   <chr>   <chr>  <chr>   <int>
## 1   phii 2016-04-12     19.5     19.5 18.5001 18.5001    750       0
## 2   phii 2016-04-13    18.85    18.85   18.68   18.85   2350       0
## 3   phii 2016-04-15  19.4999     19.5  19.014  19.014   3613       0
## 4   phii 2016-04-18 76800.02 76800.02   18.56   19.28   4108       0
## 5   phii 2016-04-19    18.87    19.75   18.87   19.61   5100       0
## 6   phii 2016-04-20    19.57   21.069   19.42 20.2101   2311       0
## 7   phii 2016-04-21     20.6  20.9081    20.6 20.9081    203       0
## 8   phii 2016-04-22    20.74    20.74   20.74   20.74    425       0
## # ... with 6 more variables: Log_Open <dbl>, Log_High <dbl>,
## #   Log_Low <dbl>, Log_Close <dbl>, Log_Volume <dbl>,
## #   Log_Daily_Return <dbl>
```

```
#open, high wrong source:yahoo finance
DF_new[DF_new$Open == "76800.02" &
DF_new$Symbol == "phii", 3] <- "18.90"
DF_new[DF_new$High == "76800.02" &
DF_new$Symbol == "phii", 4] <- "20.95"

#slg
DF_new %>% filter(Symbol == "slg") %>% filter(Date >= as.Date("2014-03-01") &
Date <= as.Date("2014-03-07"))
```

```
## # A tibble: 5 x 14
```

```
##   Symbol       Date   Open   High    Low  Close Volume OpenInt Log_Open
##    <chr>     <date>  <chr>  <chr>  <chr>  <chr>  <chr>   <int>    <dbl>
## 1    slg 2014-03-03 89.905  90.55 89.293 90.341 851028       0 4.498754
## 2    slg 2014-03-04 90.542 91.806 90.542 0.9161 731180       0 4.505814
## 3    slg 2014-03-05 91.363 91.874 90.441 91.079 526697       0 4.514841
## 4    slg 2014-03-06 91.125 91.354 90.459 90.806 404581       0 4.512232
## 5    slg 2014-03-07 90.897 91.024 89.502 90.122 521838       0 4.509727
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#closing price wrong  (source: yahoo finance)
DF_new[DF_new$Close == "0.9161" &
DF_new$Symbol == "slg", 6] <- "90.11"

#smbk
DF_new %>% filter(Symbol == "smbk") %>% filter(Date >= as.Date("2016-01-10") &
Date <= as.Date("2016-01-20"))
```

```
## # A tibble: 6 x 14
##   Symbol       Date      Open      High   Low Close Volume OpenInt
##    <chr>     <date>     <chr>     <chr> <chr> <chr>  <chr>   <int>
## 1   smbk 2016-01-11      15.5      15.5  15.1  15.5   8544       0
## 2   smbk 2016-01-12      15.5      15.5  15.5  15.5   1100       0
## 3   smbk 2016-01-13    15.485      15.5 15.35  15.4  27242       0
## 4   smbk 2016-01-14      15.4      15.4  15.4  15.4   5000       0
## 5   smbk 2016-01-15 199999.99 199999.99 15.01 15.55  35836       0
## 6   smbk 2016-01-20     15.45     15.45 15.35 15.35  15500       0
## # ... with 6 more variables: Log_Open <dbl>, Log_High <dbl>,
## #   Log_Low <dbl>, Log_Close <dbl>, Log_Volume <dbl>,
## #   Log_Daily_Return <dbl>
```

```
#open, high wrong source:yahoo finance
DF_new[DF_new$Open == "199999.99" &
DF_new$Symbol == "smbk", 3] <- "16.90"
DF_new[DF_new$High == "199999.99" &
DF_new$Symbol == "smbk", 4] <- "16.90"

#smed
DF_new %>% filter(Symbol == "smed") %>% filter(Date >= as.Date("2016-07-20") &
Date <= as.Date("2016-07-25"))
```

```
## # A tibble: 4 x 14
##   Symbol       Date   Open  High    Low Close Volume OpenInt  Log_Open
##    <chr>     <date>  <chr> <chr>  <chr> <chr>  <chr>   <int>     <dbl>
## 1   smed 2016-07-20   5.54  5.73   5.47  5.65  12980       0  1.711995
## 2   smed 2016-07-21   5.55  5.69   5.55  5.69  16888       0  1.713798
## 3   smed 2016-07-22 0.0096  5.65 0.0093  5.32  30714       0 -4.645992
## 4   smed 2016-07-25   5.33  5.51      5  5.04  91497       0  1.673351
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, low wrong, fix source: yahoo finance
DF_new[DF_new$Open == "0.0096" &
DF_new$Symbol == "smed", 3] <- "5.38"
DF_new[DF_new$Low == "0.0093" & DF_new$Symbol == "smed", 5] <- "5"
```

```
#tgen
DF_new %>% filter(Symbol == "tgen") %>% filter(Date >= as.Date("2016-03-15") &
Date <= as.Date("2016-03-25"))
```

```
## # A tibble: 7 x 14
##   Symbol       Date     Open     High    Low  Close Volume OpenInt
##    <chr>     <date>    <chr>    <chr>  <chr>  <chr>  <chr>   <int>
## 1   tgen 2016-03-15     4.01     4.09   4.01   4.09   1158       0
## 2   tgen 2016-03-17   4.0999      4.1 4.0999 4.0999   2534       0
## 3   tgen 2016-03-18      4.1     4.48    4.1   4.48    847       0
## 4   tgen 2016-03-21 80000.99 80000.99 4.4999   4.94  18532       0
## 5   tgen 2016-03-22     4.99     5.46   4.91   5.07  35808       0
## 6   tgen 2016-03-23      5.2     5.25   5.15   5.15  11802       0
## 7   tgen 2016-03-24     5.25    6.499      5   5.87  64452       0
## # ... with 6 more variables: Log_Open <dbl>, Log_High <dbl>,
## #   Log_Low <dbl>, Log_Close <dbl>, Log_Volume <dbl>,
## #   Log_Daily_Return <dbl>
```

```
#open, high wrong source:yahoo finance
DF_new[DF_new$Open == "80000.99" &
DF_new$Symbol == "tgen", 3] <- "3.85"
DF_new[DF_new$High == "80000.99" &
DF_new$Symbol == "tgen", 4] <- "3.86"

#wsr
DF_new %>% filter(Symbol == "wsr") %>% filter(Date >= as.Date("2010-09-28") &
Date <= as.Date("2010-10-15"))
```

```
## # A tibble: 14 x 14
##    Symbol       Date    Open    High    Low  Close Volume OpenInt Log_Open
##     <chr>     <date>   <chr>   <chr>  <chr>  <chr>  <chr>   <int>    <dbl>
##  1    wsr 2010-09-28  6.7445  6.8135 6.7445 6.8135  11897       0 1.908727
##  2    wsr 2010-09-29  6.7678  6.8019 6.7678 6.8019  36540       0 1.912176
##  3    wsr 2010-09-30   6.798  6.8152 6.7458 6.7865  12604       0 1.916628
##  4    wsr 2010-10-01 1133.16 1133.16 6.6818 6.7963  33756       0 7.032765
##  5    wsr 2010-10-04  6.7789  6.8188 6.7506 6.8076  20501       0 1.913815
##  6    wsr 2010-10-05  6.8248  6.8248 6.7445 6.8076 148105       0 1.920563
##  7    wsr 2010-10-06  6.7963  6.8538 6.7963 6.8475  45025       0 1.916378
##  8    wsr 2010-10-07  6.7789  6.8818 6.7789 6.8704  37921       0 1.913815
##  9    wsr 2010-10-08  6.8762  7.2246 6.8303 7.0699  67526       0 1.928066
## 10    wsr 2010-10-11  570.08  570.08 7.1222 7.2861 262436       0 6.345777
## 11    wsr 2010-10-12  7.2808  7.3276 7.0926 7.3042 198988       0 1.985241
## 12    wsr 2010-10-13  7.3385  7.3385 7.1721 7.3042 137544       0 1.993134
## 13    wsr 2010-10-14  7.3385  7.3563 7.2073 7.2684  89595       0 1.993134
## 14    wsr 2010-10-15  7.3563   7.424 7.2133 7.2808  67965       0 1.995557
## # ... with 5 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>
```

```
#open, high wrong source:yahoo finance
DF_new[DF_new$Open == "1133.16" &
DF_new$Symbol == "wsr", 3] <- "6.25"
DF_new[DF_new$High == "1133.16" &
DF_new$Symbol == "wsr", 4] <- "6.25"
DF_new[DF_new$Open == "570.08" &
DF_new$Symbol == "wsr", 3] <- "6.70"
```

```
DF_new[DF_new$High == "570.08" &
DF_new$Symbol == "wsr", 4] <- "6.70"


#Change columns back to numeric
DF_new$Low <- as.numeric(DF_new$Low)
DF_new$Open <- as.numeric(DF_new$Open)
DF_new$High <- as.numeric(DF_new$High)
DF_new$Close <- as.numeric(DF_new$Close)
DF_new$Volume <- as.numeric(DF_new$Volume)

#check class
sapply(DF_new, class)
```

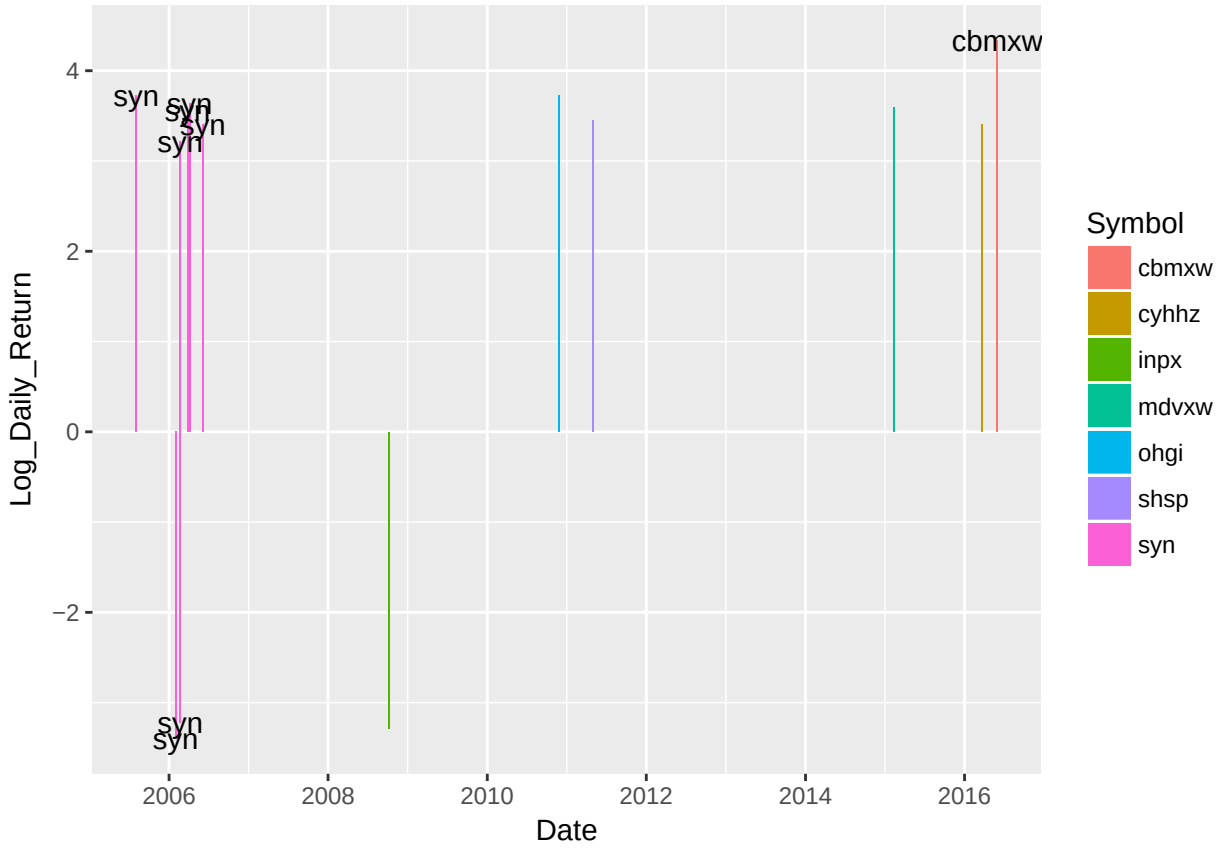
```
##           Symbol             Date             Open             High
##      "character"           "Date"        "numeric"        "numeric"
##              Low            Close           Volume          OpenInt
##        "numeric"        "numeric"        "numeric"        "integer"
##         Log_Open         Log_High          Log_Low        Log_Close
##        "numeric"        "numeric"        "numeric"        "numeric"
##       Log_Volume Log_Daily_Return
##        "numeric"        "numeric"
```

```
#rerun daily log return
DF_new$Log_Daily_Return <- log(DF_new$Close / DF_new$Open)

#summary, check to see if INF and large daily changes are gone
summary(DF_new)
```

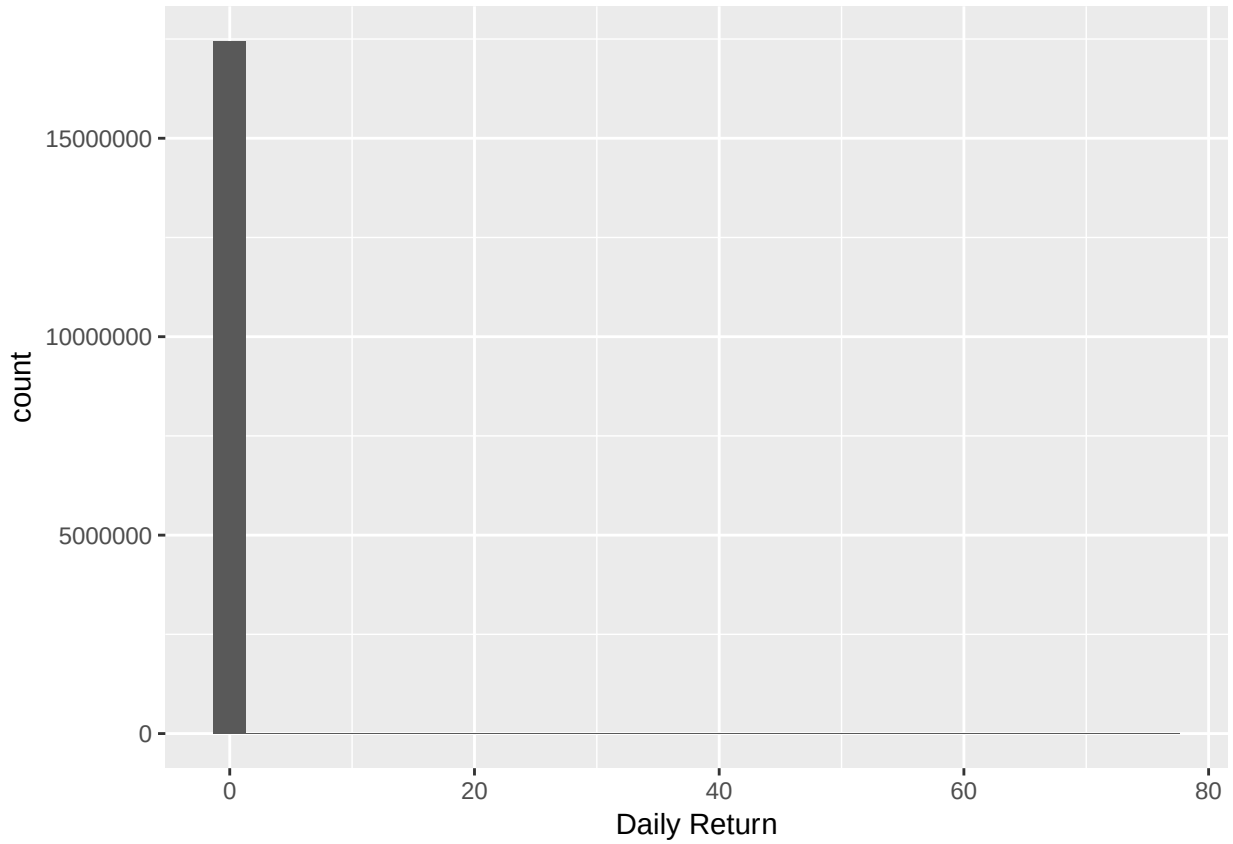
```
##     Symbol               Date                 Open
##  Length:17453219    Min.   :1962-01-02   Min.   :         0
##  Class :character   1st Qu.:2008-05-22   1st Qu.:         9
##  Mode  :character   Median :2012-06-08   Median :        18
##                     Mean   :2010-11-12   Mean   :     26249
##                     3rd Qu.:2015-06-23   3rd Qu.:        34
##                     Max.   :2017-11-10   Max.   :1423712891
##       High                 Low                 Close
##  Min.   :         0   Min.   :         0   Min.   :         0
##  1st Qu.:         9   1st Qu.:         9   1st Qu.:         9
##  Median :        19   Median :        18   Median :        18
##  Mean   :     26971   Mean   :     25362   Mean   :     26128
##  3rd Qu.:        34   3rd Qu.:        33   3rd Qu.:        34
##  Max.   :1442048636   Max.   :1362117844   Max.   :1437986240
##      Volume              OpenInt     Log_Open         Log_High
##  Min.   :         0   Min.   :0   Min.   :  -Inf   Min.   :-5.521
##  1st Qu.:     25717   1st Qu.:0   1st Qu.: 2.183   1st Qu.: 2.197
##  Median :    157428   Median :0   Median : 2.911   Median : 2.925
##  Mean   :   1581168   Mean   :0   Mean   :  -Inf   Mean   : 2.844
##  3rd Qu.:    784314   3rd Qu.:0   3rd Qu.: 3.517   3rd Qu.: 3.529
##  Max.   :2304018600   Max.   :0   Max.   :21.077   Max.   :21.089
##     Log_Low          Log_Close        Log_Volume    Log_Daily_Return
##  Min.   :  -Inf   Min.   :-5.599   Min.   : -Inf   Min.   :-3.401197
##  1st Qu.: 2.166   1st Qu.: 2.183   1st Qu.:10.15   1st Qu.:-0.008434
##  Median : 2.897   Median : 2.911   Median :11.97   Median : 0.000000
```

```
##  Mean   :  -Inf   Mean   : 2.829   Mean   : -Inf   Mean   :-0.000153
##  3rd Qu.: 3.504   3rd Qu.: 3.517   3rd Qu.:13.57   3rd Qu.: 0.008178
##  Max.   :21.032   Max.   :21.087   Max.   :21.56   Max.   : 4.335507
```

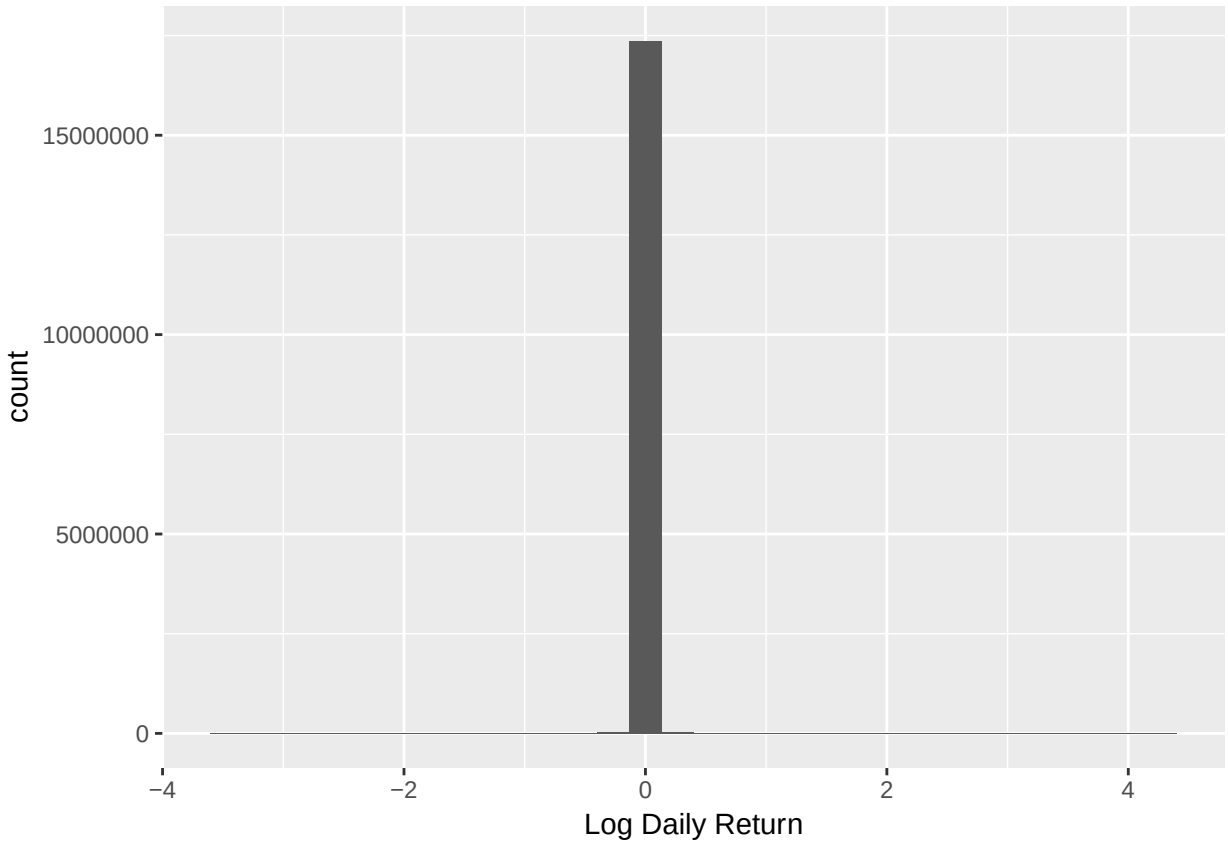
```r
#filter for log price changes greater than or less than 3%
DF_change2 <- DF_new %>% filter(!between(Log_Daily_Return, -3, 3))
#cbmxw only stock with log daily return/loss of greater than 4%
#due to time constraints I did not validate all of the stock and ETF prices, though I expect some of th

#graph
DF_change2$Symbol <- as.factor(DF_change2$Symbol)
ggplot(DF_change2,
aes(
x = Date,
y = Log_Daily_Return,
position = "stack",
fill = Symbol
)) +
geom_bar(stat = "identity", width = 5) + geom_text(data = subset(DF_change2, Symbol == 'cbmxw' |
Symbol == 'syn'),
aes(label = Symbol))
```




```r
#Daily return Arithmetic
#Calculate daily return arithmetic
DF_new$Daily_Return <- ((DF_new$Close / DF_new$Open) - 1)
#graph daily return
ggplot(data = DF_new, aes(DF_new$Daily_Return)) + geom_histogram() + xlab("Daily Return")
```




```
#graph Daily return log
ggplot(data = DF_new, aes(DF_new$Log_Daily_Return)) + geom_histogram() + xlab("Log Daily Return")
```



Analysis: Log return equaled infinite when opening price was $0. This occurred due to errors in the data, either the data was incorrect (i.e. wrong price) or the stock/ETF had not started trading as of that date (i.e. prior to IPO). When the data was incorrect, I fixed the data. When the stock had not started trading as of that date, I removed the row from the data frame. Additionally, the log return variable revealed incorrect data. I saw large changes in price. Upon further inspection, many of the data inputs were incorrect. I corrected the prices. This resulted in one stock showing a significant daily change, CBMXW. It had a log daily return of 4.335507% on 2016-05-31. As you can see in the above graph, SYN had the greatest frequency of large stock moves (7 days) and CBMXW had the greatest daily change in price.

Conclusion: By calculating this variable, I was able to discover errors in the data that required fixing. I was also able to see which stocks/ETF's had the greatest daily price change, and which stocks/ETF's had the largest volatility.
Additionally, the benefit of using log returns instead of absolute prices is normalization, "measuring all variables in a comparable metric, thus enabling evaluation of analytic relationships among two or more variables despite originating from price series of unequal values." (source: https://quantivity.wordpress.com/2011/02/21/why-log-returns/). Another benefit of log returns is, assuming normal distribution, adding period returns produces an end period return that is also normally distributed (source: https://www.youtube.com/watch?v=PtoUlt3V0CI). Additionally,graphing log returns enables a person to see big moves on a percentage basis not on an absolute dollar basis (source: https://www.usatoday.com/story/money/columnist/krantz/2013/08/25/linear-logarithmic-stock-charts/2657493/). Also, converting prices to a log scale is highly useful when charting stock prices, as a significant percentage move will always correspond to a significant visual change (source :https://finance.zacks.com/use-logarithmic-scale-stocks-8760.html)

3.3 Rolling moving average

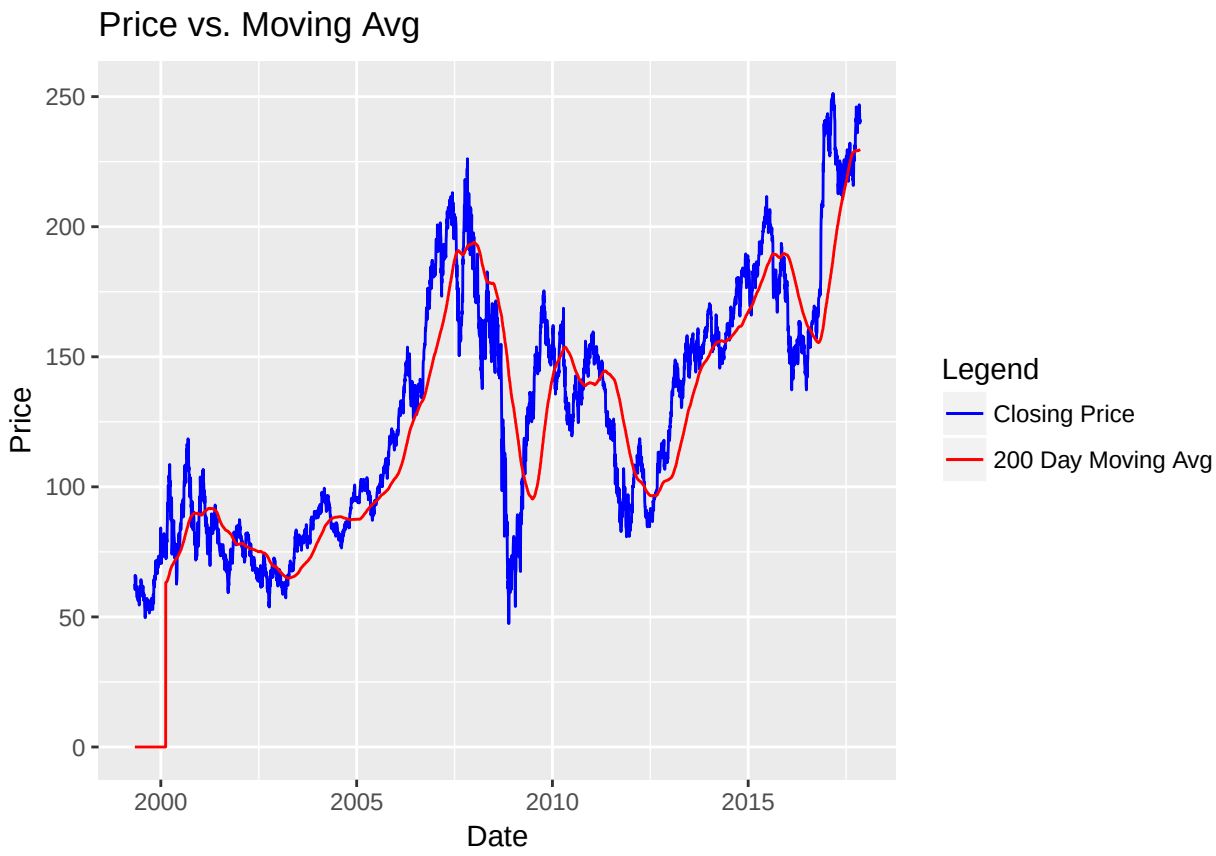
Lead: I'm going to calculate the 5 day moving average on the closing price of each stock and ETF.

```r
#load library
library(RcppRoll)
#caclulate 200 day moving average on closing price, create new column
DF_new <- DF_new %>%
group_by(Symbol) %>%
mutate(Roll_Avg_Price = roll_mean(
Close,
200,
na.rm = TRUE,
align = "right",
fill = 0
)) %>%
ungroup()

#summary
summary(DF_new)
```

```
##     Symbol               Date                 Open
##  Length:17453219    Min.   :1962-01-02   Min.   :         0
##  Class :character   1st Qu.:2008-05-22   1st Qu.:         9
##  Mode  :character   Median :2012-06-08   Median :        18
##                     Mean   :2010-11-12   Mean   :     26249
##                     3rd Qu.:2015-06-23   3rd Qu.:        34
##                     Max.   :2017-11-10   Max.   :1423712891
##
##       High                 Low                 Close
##  Min.   :         0   Min.   :         0   Min.   :         0
##  1st Qu.:         9   1st Qu.:         9   1st Qu.:         9
##  Median :        19   Median :        18   Median :        18
##  Mean   :     26971   Mean   :     25362   Mean   :     26128
##  3rd Qu.:        34   3rd Qu.:        33   3rd Qu.:        34
##  Max.   :1442048636   Max.   :1362117844   Max.   :1437986240
##
##      Volume              OpenInt     Log_Open         Log_High
##  Min.   :         0   Min.   :0   Min.   :  -Inf   Min.   :-5.521
##  1st Qu.:     25717   1st Qu.:0   1st Qu.: 2.183   1st Qu.: 2.197
##  Median :    157428   Median :0   Median : 2.911   Median : 2.925
##  Mean   :   1581168   Mean   :0   Mean   :  -Inf   Mean   : 2.844
##  3rd Qu.:    784314   3rd Qu.:0   3rd Qu.: 3.517   3rd Qu.: 3.529
##  Max.   :2304018600   Max.   :0   Max.   :21.077   Max.   :21.089
##
##     Log_Low          Log_Close        Log_Volume    Log_Daily_Return
##  Min.   :  -Inf   Min.   :-5.599   Min.   : -Inf   Min.   :-3.401197
##  1st Qu.: 2.166   1st Qu.: 2.183   1st Qu.:10.15   1st Qu.:-0.008434
##  Median : 2.897   Median : 2.911   Median :11.97   Median : 0.000000
##  Mean   :  -Inf   Mean   : 2.829   Mean   : -Inf   Mean   :-0.000153
##  3rd Qu.: 3.504   3rd Qu.: 3.517   3rd Qu.:13.57   3rd Qu.: 0.008178
##  Max.   :21.032   Max.   :21.087   Max.   :21.56   Max.   : 4.335507
##
##   Daily_Return      Roll_Avg_Price
##  Min.   :-0.96667   Min.   :        0
##  1st Qu.:-0.00840   1st Qu.:        7
##  Median : 0.00000   Median :       16
##  Mean   : 0.00028   Mean   :    24953
```

```
##  3rd Qu.: 0.00821   3rd Qu.:       31
##  Max.   :75.36364   Max.   :911390331
##                     NA's   :69829
```

```
#graph rolling average vs. closing price for a select stock, GS
ggplot() +
geom_line(data = subset(DF_new, Symbol == 'gs'),
aes(x = Date, y = Close, colour = "blue")) +
geom_line(data = subset(DF_new, Symbol == 'gs'),
aes(x = Date, y = Roll_Avg_Price, colour = "red")) +
labs(
title = "Price vs. Moving Avg",
x = "Date",
y = "Price",
color = "Legend"
) +
scale_color_manual(
labels = c("Closing Price", "200 Day Moving Avg"),
values = c("blue", "red")
)
```



Analysis: As you can see in the above graph, the 200 day moving average smooths out the data.
A moving average can also act as support or resistance. In a downtrend, a moving average may act as resistance; like a ceiling, the price hits the level and then starts to drop again. This can be seen in the GS graph during 2008.

Conclusion: The moving average is a common technical analysis tool used by traders. It is often used to identify trend direction and can also be used to generate potential buy and sell signals (source: http://www.onlinetradingconcepts.com/TechnicalAnalysis/MASimple.html).

3.4 Feature Creation: Day of Week

Lead: Add day of week to the data frame.

```
#add day of week to data frame
DF_new$Day <- weekdays(DF_new$Date)

#calculate return by day of week for each ETF and stocks
DF_new <- DF_new %>%
  group_by(Symbol, Day) %>%
  mutate(Weekday_Return = mean(Log_Daily_Return)) %>%
  ungroup()

#summary
summary(DF_new)
```

```
##     Symbol               Date                 Open
##  Length:17453219    Min.   :1962-01-02   Min.   :         0
##  Class :character   1st Qu.:2008-05-22   1st Qu.:         9
##  Mode  :character   Median :2012-06-08   Median :        18
##                     Mean   :2010-11-12   Mean   :     26249
##                     3rd Qu.:2015-06-23   3rd Qu.:        34
##                     Max.   :2017-11-10   Max.   :1423712891
##
##       High                 Low                 Close
##  Min.   :         0   Min.   :         0   Min.   :         0
##  1st Qu.:         9   1st Qu.:         9   1st Qu.:         9
##  Median :        19   Median :        18   Median :        18
##  Mean   :     26971   Mean   :     25362   Mean   :     26128
##  3rd Qu.:        34   3rd Qu.:        33   3rd Qu.:        34
##  Max.   :1442048636   Max.   :1362117844   Max.   :1437986240
##
##      Volume              OpenInt     Log_Open         Log_High
##  Min.   :         0   Min.   :0   Min.   :  -Inf   Min.   :-5.521
##  1st Qu.:     25717   1st Qu.:0   1st Qu.: 2.183   1st Qu.: 2.197
##  Median :    157428   Median :0   Median : 2.911   Median : 2.925
##  Mean   :   1581168   Mean   :0   Mean   :  -Inf   Mean   : 2.844
##  3rd Qu.:    784314   3rd Qu.:0   3rd Qu.: 3.517   3rd Qu.: 3.529
##  Max.   :2304018600   Max.   :0   Max.   :21.077   Max.   :21.089
##
##     Log_Low         Log_Close        Log_Volume    Log_Daily_Return
##  Min.   :  -Inf   Min.   :-5.599   Min.   : -Inf   Min.   :-3.401197
##  1st Qu.: 2.166   1st Qu.: 2.183   1st Qu.:10.15   1st Qu.:-0.008434
##  Median : 2.897   Median : 2.911   Median :11.97   Median : 0.000000
##  Mean   :  -Inf   Mean   : 2.829   Mean   : -Inf   Mean   :-0.000153
##  3rd Qu.: 3.504   3rd Qu.: 3.517   3rd Qu.:13.57   3rd Qu.: 0.008178
##  Max.   :21.032   Max.   :21.087   Max.   :21.56   Max.   : 4.335507
##
##   Daily_Return      Roll_Avg_Price          Day
##  Min.   :-0.96667   Min.   :        0   Length:17453219
##  1st Qu.:-0.00840   1st Qu.:        7   Class :character
##  Median : 0.00000   Median :       16   Mode  :character
##  Mean   : 0.00028   Mean   :    24953
##  3rd Qu.: 0.00821   3rd Qu.:       31
##  Max.   :75.36364   Max.   :911390331
```

```
##                     NA's   :69829
##  Weekday_Return
##  Min.   :-0.5877867
##  1st Qu.:-0.0006922
##  Median : 0.0000448
##  Mean   :-0.0001530
##  3rd Qu.: 0.0006762
##  Max.   : 0.4398429
##
```

```
#filter
DF_new %>% filter(Symbol == "gs") %>% filter(Date >= as.Date("1999-05-10") &
                                               Date <= as.Date("1999-05-14")) # Wednesday's achieved th
```

```
## # A tibble: 5 x 18
##   Symbol       Date   Open   High    Low  Close  Volume OpenInt Log_Open
##    <chr>     <date>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>   <int>    <dbl>
## 1     gs 1999-05-10 65.331 65.441 62.549 62.941 2839054       0 4.179467
## 2     gs 1999-05-11 62.329 63.389 61.885 62.888 2017182       0 4.132427
## 3     gs 1999-05-12 63.216 66.671 61.547 65.441 2915697       0 4.146557
## 4     gs 1999-05-13 65.222 67.000 64.555 65.166 1253094       0 4.177797
## 5     gs 1999-05-14 63.444 64.329 61.491 62.497 2256860       0 4.150158
## # ... with 9 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>,
## #   Daily_Return <dbl>, Roll_Avg_Price <dbl>, Day <chr>,
## #   Weekday_Return <dbl>
```

```
#calculate return by day of week for all ETF and stocks
DF_new <- DF_new %>%
  group_by(Day) %>%
  mutate(Weekday_Return_All = mean(Log_Daily_Return)) %>%
  ungroup()

#summary
summary(DF_new)
```

```
##     Symbol               Date                 Open
##  Length:17453219    Min.   :1962-01-02   Min.   :         0
##  Class :character   1st Qu.:2008-05-22   1st Qu.:         9
##  Mode  :character   Median :2012-06-08   Median :        18
##                     Mean   :2010-11-12   Mean   :     26249
##                     3rd Qu.:2015-06-23   3rd Qu.:        34
##                     Max.   :2017-11-10   Max.   :1423712891
##
##       High                 Low                 Close
##  Min.   :         0   Min.   :         0   Min.   :         0
##  1st Qu.:         9   1st Qu.:         9   1st Qu.:         9
##  Median :        19   Median :        18   Median :        18
##  Mean   :     26971   Mean   :     25362   Mean   :     26128
##  3rd Qu.:        34   3rd Qu.:        33   3rd Qu.:        34
##  Max.   :1442048636   Max.   :1362117844   Max.   :1437986240
##
##      Volume              OpenInt     Log_Open         Log_High
##  Min.   :         0   Min.   :0   Min.   :  -Inf   Min.   :-5.521
##  1st Qu.:     25717   1st Qu.:0   1st Qu.: 2.183   1st Qu.: 2.197
##  Median :    157428   Median :0   Median : 2.911   Median : 2.925
```

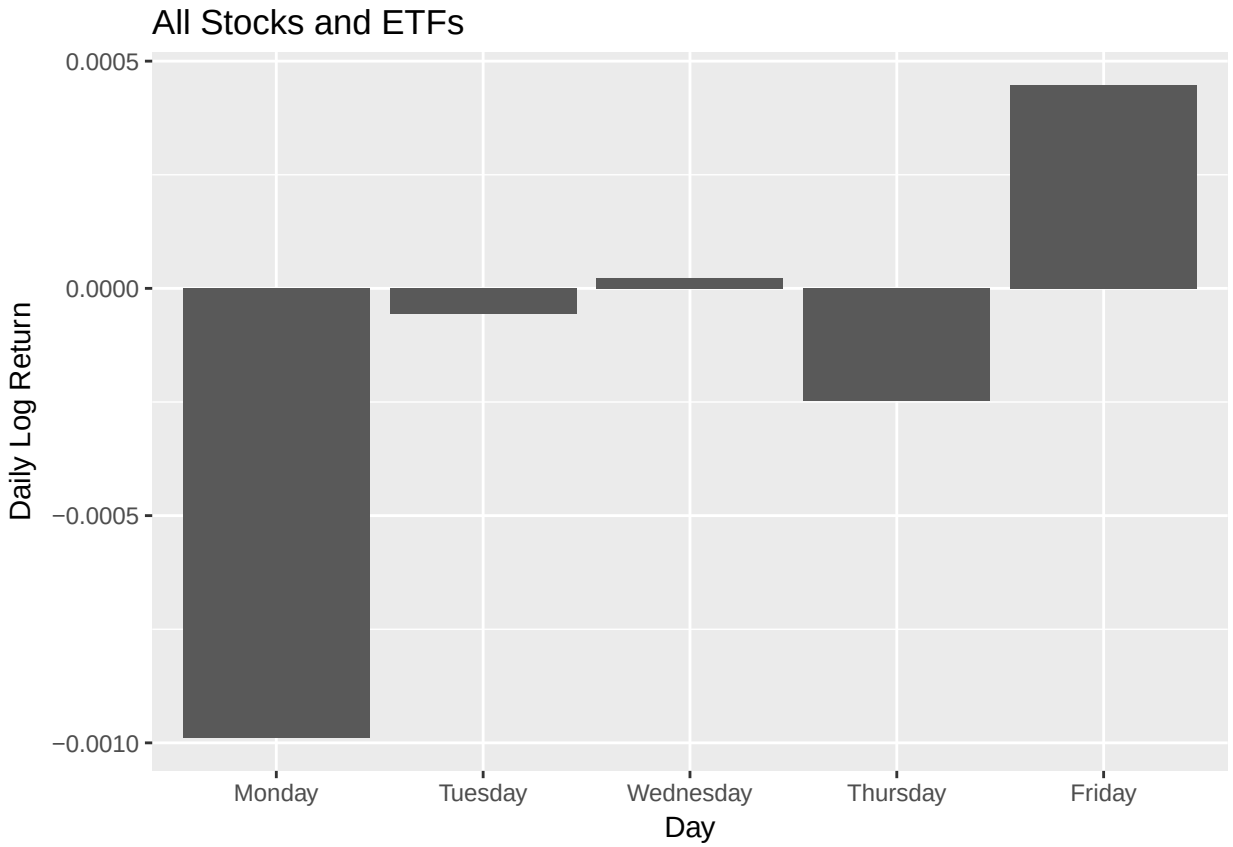
```
##  Mean   :   1581168   Mean   :0   Mean   :  -Inf   Mean   : 2.844
##  3rd Qu.:    784314   3rd Qu.:0   3rd Qu.: 3.517   3rd Qu.: 3.529
##  Max.   :2304018600   Max.   :0   Max.   :21.077   Max.   :21.089
##
##     Log_Low          Log_Close        Log_Volume    Log_Daily_Return
##  Min.   :  -Inf   Min.   :-5.599   Min.   : -Inf   Min.   :-3.401197
##  1st Qu.: 2.166   1st Qu.: 2.183   1st Qu.:10.15   1st Qu.:-0.008434
##  Median : 2.897   Median : 2.911   Median :11.97   Median : 0.000000
##  Mean   :  -Inf   Mean   : 2.829   Mean   : -Inf   Mean   :-0.000153
##  3rd Qu.: 3.504   3rd Qu.: 3.517   3rd Qu.:13.57   3rd Qu.: 0.008178
##  Max.   :21.032   Max.   :21.087   Max.   :21.56   Max.   : 4.335507
##
##   Daily_Return      Roll_Avg_Price          Day
##  Min.   :-0.96667   Min.   :        0   Length:17453219
##  1st Qu.:-0.00840   1st Qu.:        7   Class :character
##  Median : 0.00000   Median :       16   Mode  :character
##  Mean   : 0.00028   Mean   :    24953
##  3rd Qu.: 0.00821   3rd Qu.:       31
##  Max.   :75.36364   Max.   :911390331
##                     NA's   :69829
##  Weekday_Return       Weekday_Return_All
##  Min.   :-0.5877867   Min.   :-0.00098970
##  1st Qu.:-0.0006922   1st Qu.:-0.00024810
##  Median : 0.0000448   Median :-0.00005603
##  Mean   :-0.0001530   Mean   :-0.00015297
##  3rd Qu.: 0.0006762   3rd Qu.: 0.00002279
##  Max.   : 0.4398429   Max.   : 0.00044819
##
```

```
#filter
DF_new %>% filter(Symbol == "gs") %>% filter(Date >= as.Date("1999-05-10") &
                                               Date <= as.Date("1999-05-14")) # Friday's achieved the h
```

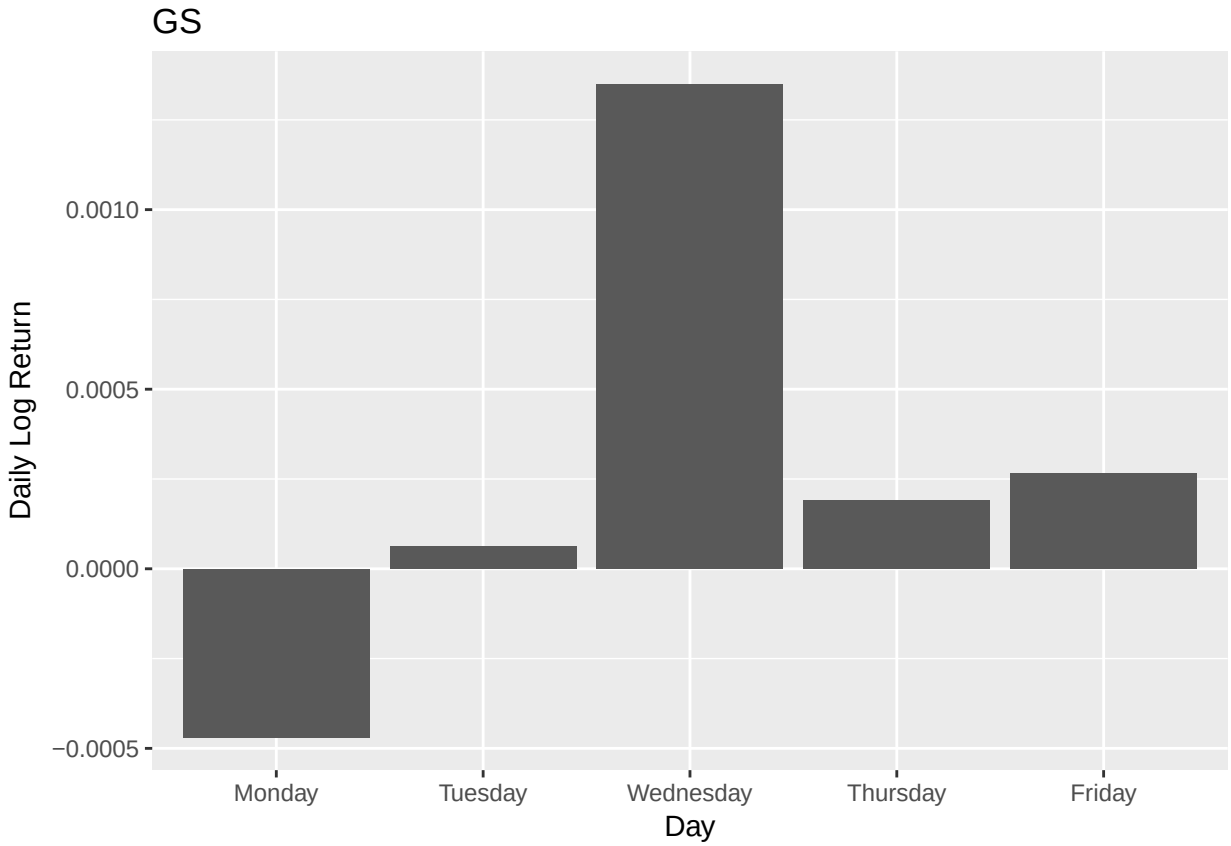
```
## # A tibble: 5 x 19
##   Symbol       Date   Open   High    Low  Close  Volume OpenInt Log_Open
##    <chr>     <date>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>   <int>    <dbl>
## 1     gs 1999-05-10 65.331 65.441 62.549 62.941 2839054       0 4.179467
## 2     gs 1999-05-11 62.329 63.389 61.885 62.888 2017182       0 4.132427
## 3     gs 1999-05-12 63.216 66.671 61.547 65.441 2915697       0 4.146557
## 4     gs 1999-05-13 65.222 67.000 64.555 65.166 1253094       0 4.177797
## 5     gs 1999-05-14 63.444 64.329 61.491 62.497 2256860       0 4.150158
## # ... with 10 more variables: Log_High <dbl>, Log_Low <dbl>,
## #   Log_Close <dbl>, Log_Volume <dbl>, Log_Daily_Return <dbl>,
## #   Daily_Return <dbl>, Roll_Avg_Price <dbl>, Day <chr>,
## #   Weekday_Return <dbl>, Weekday_Return_All <dbl>
```

```
#graph
#filter
DF_days = DF_new %>% filter(Symbol == "gs") %>% filter(Date >= as.Date("1999-05-10") &
                                                         Date <= as.Date("1999-05-14"))
#set factor levels
DF_days$Day <- as.factor(DF_days$Day)
DF_days$Day <-
  factor(DF_days$Day,
         levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))
```

```
#graph for all stocks and etfs combined
ggplot(DF_days, aes(Day, Weekday_Return_All)) +
  geom_col() + ylab("Daily Log Return") + ggtitle("All Stocks and ETFs")
```




```
#graph for GS
ggplot(DF_days, aes(Day, Weekday_Return)) +
  geom_col() + ylab("Daily Log Return") + ggtitle("GS")
```



Analysis: When looking at all stocks and ETF's combined, Friday's achieved the highest daily log return, Monday's performed worst. Looking at a specific stock, GS, Wednesday's performed best while Monday's again performed worst.

Conclusion: There seems to be a bias towards positive market performance on Friday's and a bias towards under performance on Monday's. The business insight for this is investors are better off buying on Monday's selling on Friday's. Additionally, this analysis can be done on specific stocks. For example, over GS's trading history, Wednesday's outperformed and Monday's performed worst.