Uber Stock Prices 2019-2021

By: Matthew Cooper

Dataset Details:

Dataset source: https://www.kaggle.com/varpit94/uber-stock-data

Column Names:

Date - Y/M/D

Open Price - Price from the first transaction of a trading day

Close Price - Price from the last transaction of a trading day

High Price - Maximum price in a trading day

Low Price - Minimum price in a trading day

Adjusted Close Price - Closing price adjusted to reflect the value after accounting for any corporate actions

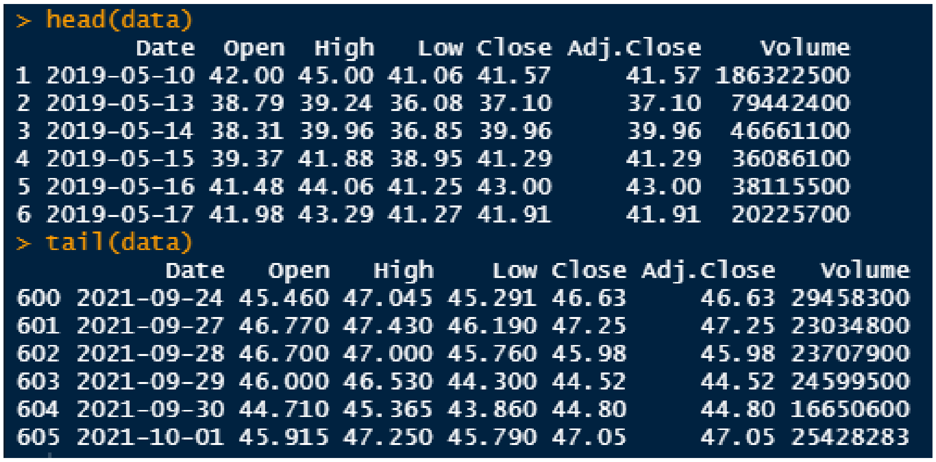
Volume - Number of units traded in a day

Number of data rows: 605

The Project:

```
# Set the seed (even if not really needed, it's always good to do)
set.seed(951)

#Load the dataset
data <- read.csv("C: /Datasets/UBER_Stock_dataset/UBER.csv")
head(data)
tail(data)
```

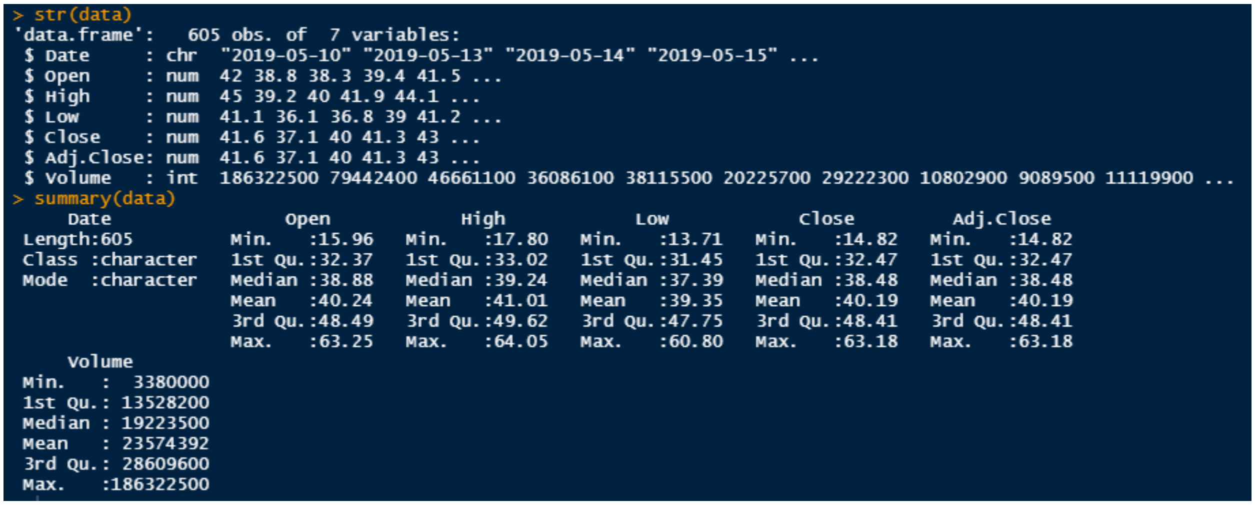
So, first load the dataset and print the head and tail to see what the data looks like.



```
> head(data)
        Date  Open  High   Low Close Adj.Close    Volume
1 2019-05-10 42.00 45.00 41.06 41.57     41.57 186322500
2 2019-05-13 38.79 39.24 36.08 37.10     37.10  79442400
3 2019-05-14 38.31 39.96 36.85 39.96     39.96  46661100
4 2019-05-15 39.37 41.88 38.95 41.29     41.29  36086100
5 2019-05-16 41.48 44.06 41.25 43.00     43.00  38115500
6 2019-05-17 41.98 43.29 41.27 41.91     41.91  20225700
> tail(data)
          Date   Open   High    Low Close Adj.Close   Volume
600 2021-09-24 45.460 47.045 45.291 46.63     46.63 29458300
601 2021-09-27 46.770 47.430 46.190 47.25     47.25 23034800
602 2021-09-28 46.700 47.000 45.760 45.98     45.98 23707900
603 2021-09-29 46.000 46.530 44.300 44.52     44.52 24599500
604 2021-09-30 44.710 45.365 43.860 44.80     44.80 16650600
605 2021-10-01 45.915 47.250 45.790 47.05     47.05 25428283
```

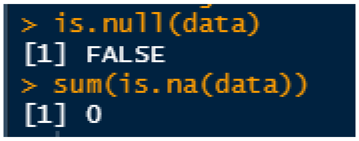
```
#-----Doing some checking and fixing

#Gaining some basic insight
str(data)
summary(data)

#Convert  Date column from "chr" to "date"
data$Date <- as.Date(data$Date, format="%Y-%m-%d" )
#Checkking for nulls and missing data
is.null(data)
sum(is.na(data))
```

First print out some descriptive info about the variables.



```
> str(data)
'data.frame':   605 obs. of  7 variables:
 $ Date     : chr  "2019-05-10" "2019-05-13" "2019-05-14" "2019-05-15" ...
 $ Open     : num  42 38.8 38.3 39.4 41.5 ...
 $ High     : num  45 39.2 40 41.9 44.1 ...
 $ Low      : num  41.1 36.1 36.8 39 41.2 ...
 $ Close    : num  41.6 37.1 40 41.3 43 ...
 $ Adj.Close: num  41.6 37.1 40 41.3 43 ...
 $ Volume   : int  186322500 79442400 46661100 36086100 38115500 20225700 29222300 10802900 9089500 11119900 ...
> summary(data)
     Date                Open            High            Low            Close          Adj.Close    
 Length:605         Min.   :15.96   Min.   :17.80   Min.   :13.71   Min.   :14.82   Min.   :14.82  
 Class :character   1st Qu.:32.37   1st Qu.:33.02   1st Qu.:31.45   1st Qu.:32.47   1st Qu.:32.47  
 Mode  :character   Median :38.88   Median :39.24   Median :37.39   Median :38.48   Median :38.48  
                    Mean   :40.24   Mean   :41.01   Mean   :39.35   Mean   :40.19   Mean   :40.19  
                    3rd Qu.:48.49   3rd Qu.:49.62   3rd Qu.:47.75   3rd Qu.:48.41   3rd Qu.:48.41  
                    Max.   :63.25   Max.   :64.05   Max.   :60.80   Max.   :63.18   Max.   :63.18  
     Volume         
 Min.   :  3380000  
 1st Qu.: 13528200  
 Median : 19223500  
 Mean   : 23574392  
 3rd Qu.: 28609600  
 Max.   :186322500  
```

There is a problem with the "Date" column being typed incorrectly, so it is re-typed to the correct 'date' type. Now check for missing data.

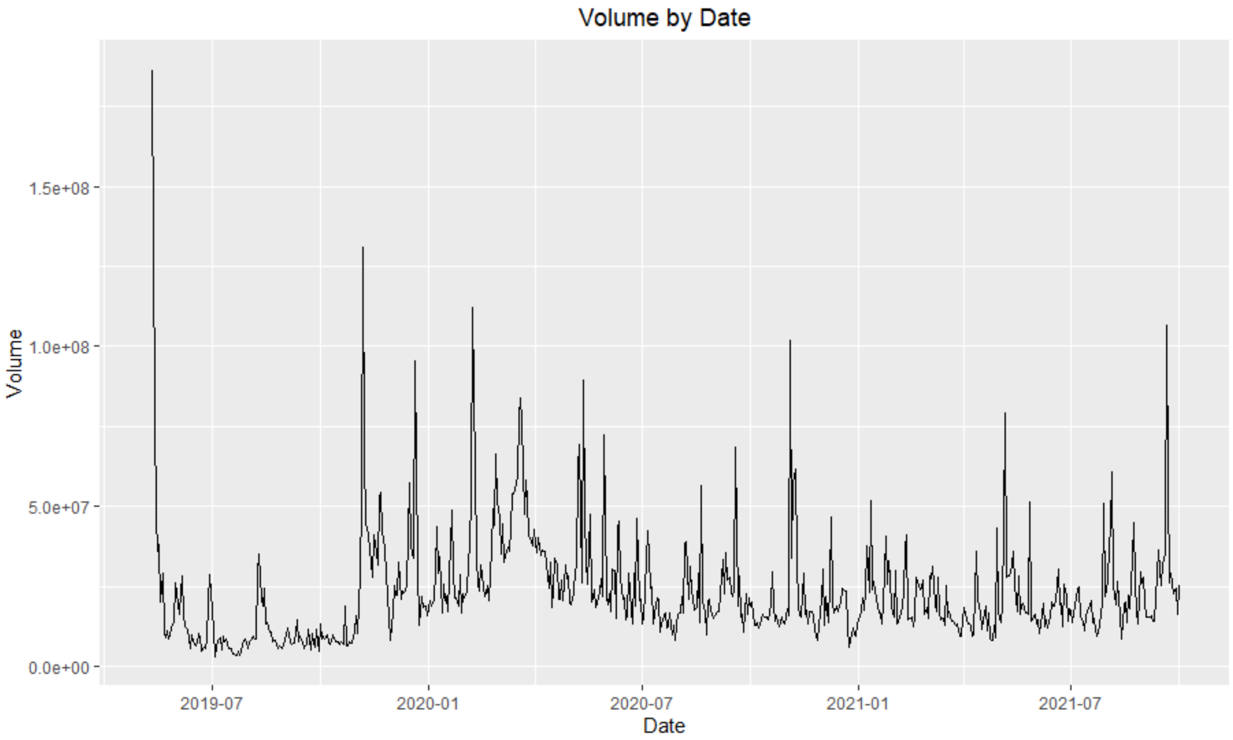


```
> is.null(data)
[1] FALSE
> sum(is.na(data))
[1] 0
```

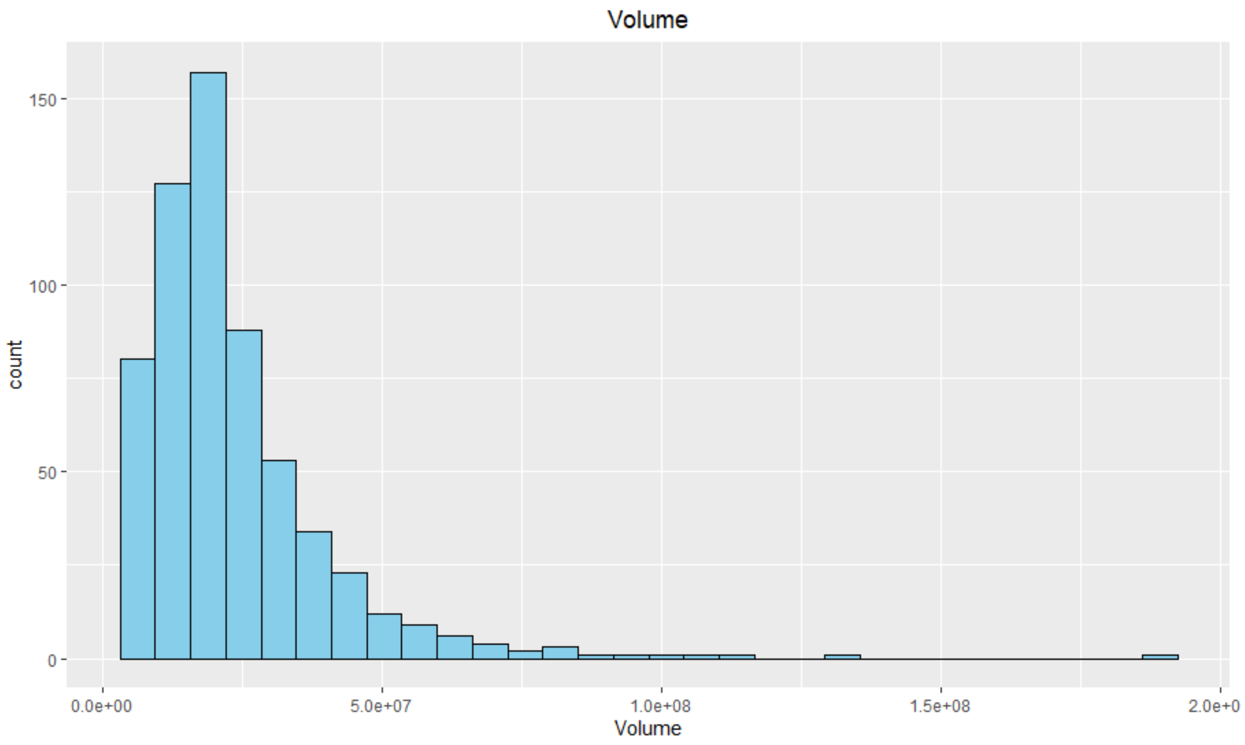
```
#-----Some Exploratory Visualizations

library(ggplot2)
```

```
#Vis of Volume column by date
ggplot(data=data, aes(x=Date)) +
        geom_line(aes(y=Volume), color="Black") +
        ggtitle("Volume by Date") +
        theme(plot.title = element_text(hjust = 0.5))

#Histogram of Volume
ggplot(data, aes(x=Volume)) +
        geom_histogram(color="Black", fill="Sky Blue") +
        ggtitle("Volume") +
        theme(plot.title = element_text(hjust = 0.5))
```

The first thing chosen to look at is the "Volume" variable. The full date range is used. I was curious to see how those numbers looked over time. We see some very high number of trades, but most of the volume of trades is steady.

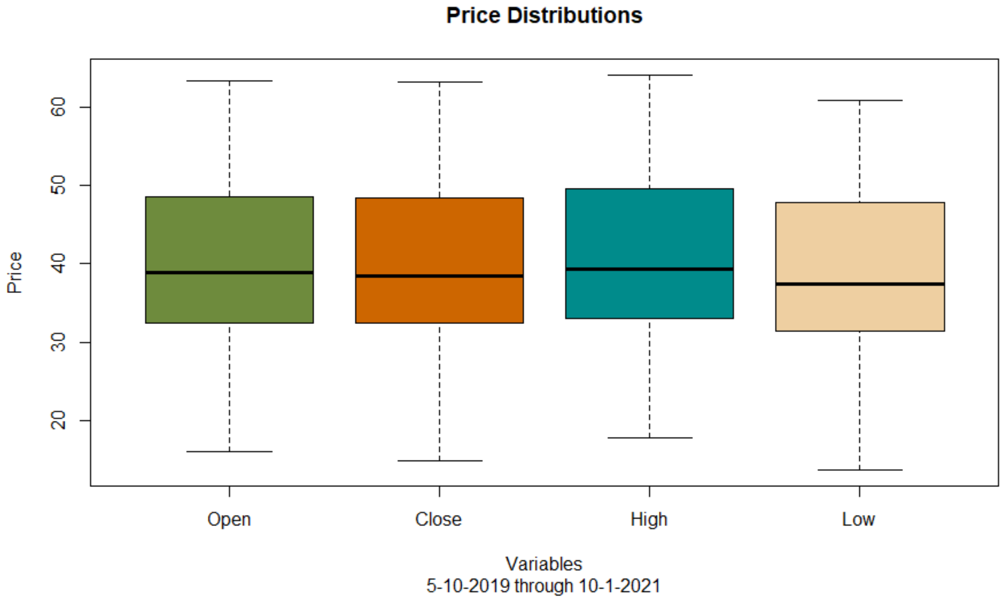


The next item of interest is looking at a histogram of "Volume". This graph confirms my previous statement that there are occasional high points but the majority of the trade volume falls in between 0 - 50,000,000

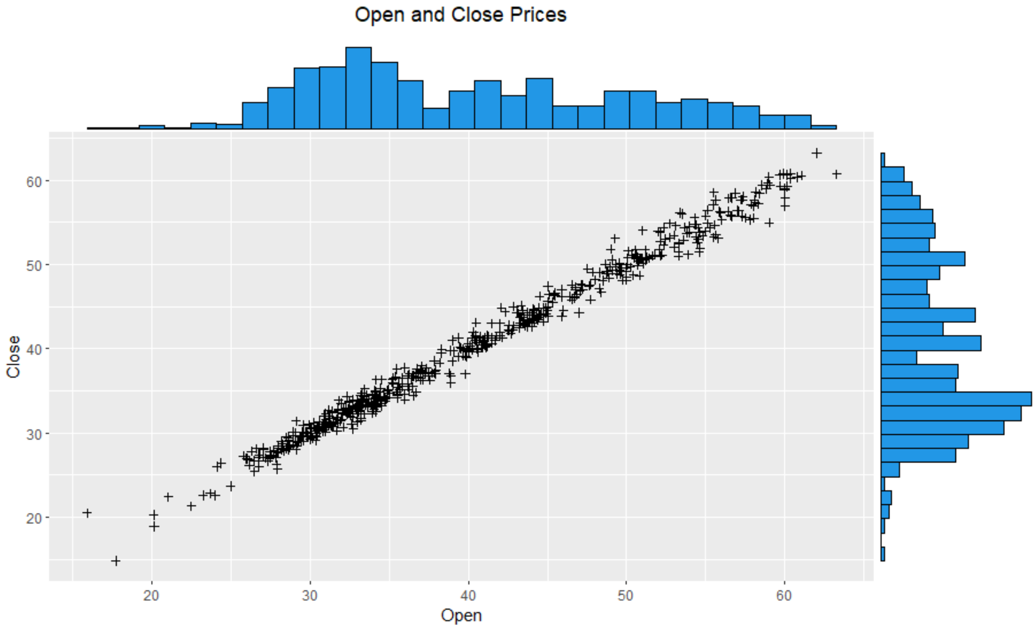


```
#Boxplot of all prices
boxplot(data[, c("Open", "Close", "High", "Low")],
    main="Price Distributions", sub="5-10-2019 through 10-1-2021",
    ylab ="Price", xlab ="Variables",
    col=c("darkolivegreen4", "darkorange3", "cyan4",
    "navajowhite2"))

#Scatterplot + Histograms of Open and Close Prices
spOC<-ggplot(data =data, aes(x=Open, y=Close)) +
                  geom_point(shape=3, color="Black" ) +
                  ggtitle("Open and Close Prices") +
                  theme(plot.title=element_text(hjust =0.5))
                  library(ggExtra)
                  ggMarginal(spOC, type="histogram" , fill=4)
```

Now the interest is in the "Price" variables: Open, Close, High & Low. We see in the box plot a very interesting pattern. The spread of each variable is very similar to the others, and doesn't seem to be much variation. Since "Open" and "Close" are so similar, let's do a scatterplot.

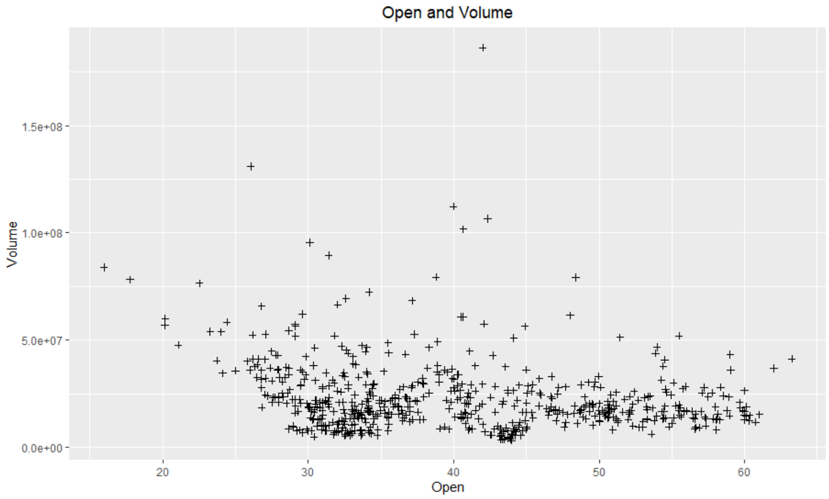


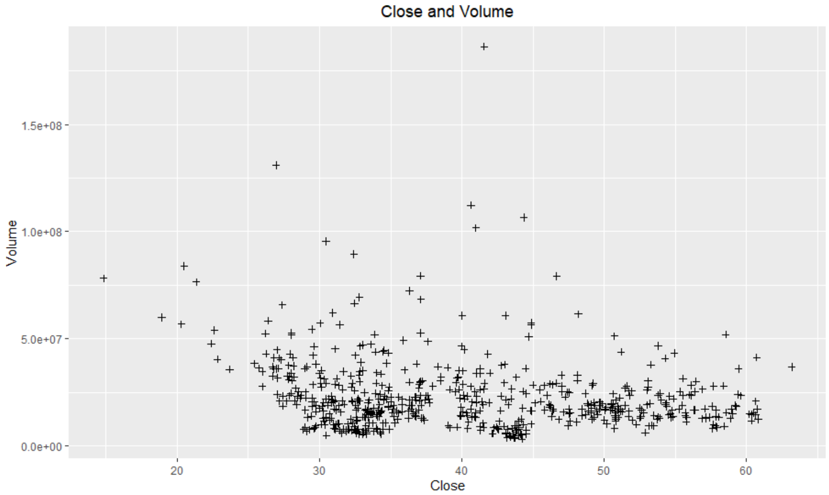
This is even more interesting. We see a positive linear relationship between the "Open" and "Close" variables. With the helpful addition of the respective histograms, we see how the bulk of the stock price occurs between $26 - $36. It is also interesting to see how tight the data point are from $25 - $52. We also see at the upper and lower ends a little spread occurring.



```
#Scatterplot of Open and Volume
ggplot(data=data, aes(x=Open, y=Volume)) +
        geom_point(shape=3, color="Black") +
        ggtitle("Open and Volume") +
        theme(plot.title = element_text(hjust = 0.5))

#Scatterplot of Close and Volume
ggplot(data=data, aes(x=Close, y=Volume)) +
        geom_point(shape=3, color="Black") +
        ggtitle("Close and Volume") +
        theme(plot.title = element_text(hjust = 0.5))
```

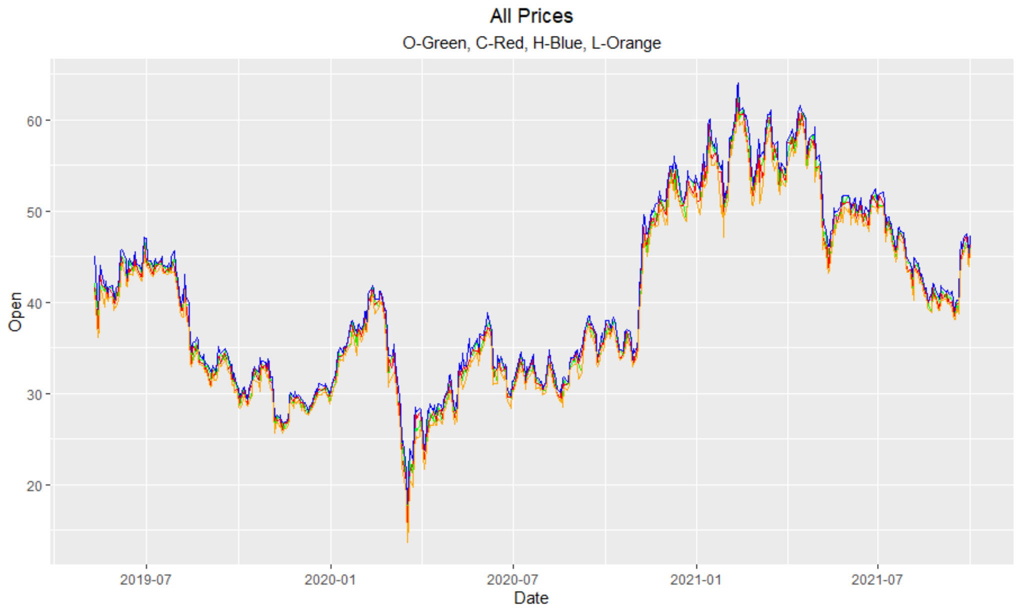
I was curious if there was any relationship between the "Open" and "Close" prices vs "Volume". The scatterplots below of each price show that there is no relationship. This would mean that the price most likely has no effect on the number of trades occurring in a day.



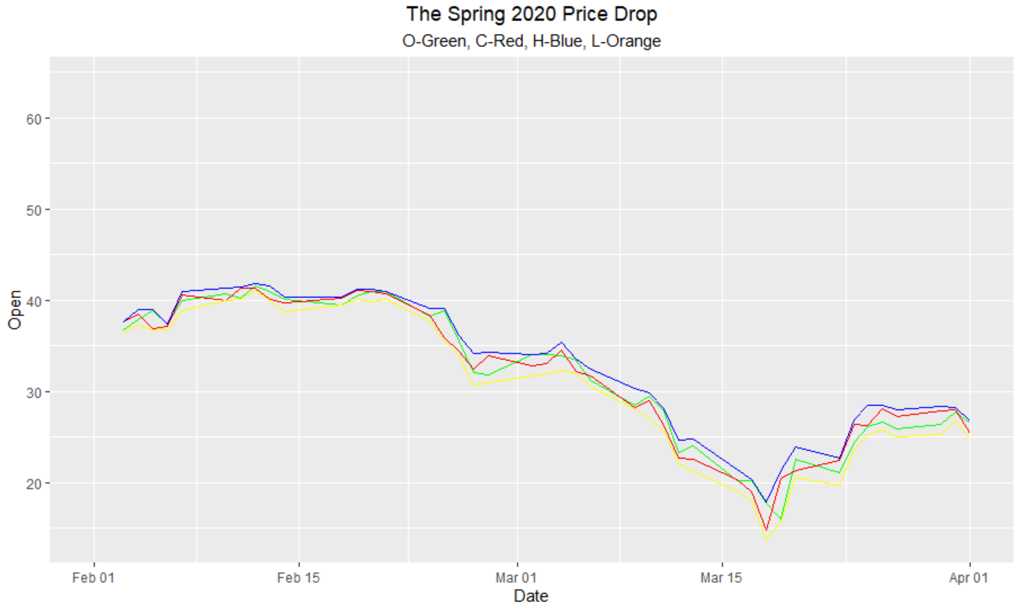


```
#Line chart of Open, Close, High, and Low prices
ggplot(data=data, aes(x=Date)) +
        geom_line(aes(y=Open), color="Green") +
        geom_line(aes(y=Close), color="Red") +
        geom_line(aes(y=High), color="Blue") +
        geom_line(aes(y=Low), color="Orange") +
        ggtitle("All Prices", subtitle="O-Green, C-Red, H-Blue, L-
Orange") +
        theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5))

#Line chart of O, C, H, and L focusing on the severe price drop in
2020
ggplot(data=data, aes(x=Date)) +
        geom_line(aes(y=Open), color="Green") +
        geom_line(aes(y=Close), color="Red") +
        geom_line(aes(y=High), color="Blue") +
        geom_line(aes(y=Low), color="Yellow") +
        scale_x_date(limits = as.Date(c("2020-02-01","2020-04-01"))) +
        ggtitle("The Spring 2020 Price Drop", subtitle="O-Green, C-
Red, H-Blue, L-Orange") +
        theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5))

#Line chart of O, C, H, and L focusing on the severe price rise in
late 2020
ggplot(data=data, aes(x=Date)) +
        geom_line(aes(y=Open), color="Green") +
        geom_line(aes(y=Close), color="Red") +
        geom_line(aes(y=High), color="Blue") +
        geom_line(aes(y=Low), color="Yellow") +
        scale_x_date(limits = as.Date(c("2020-10-25","2020-11-15"))) +
        ggtitle("The Winter 2020 Price Rise", subtitle="O-Green, C-
Red, H-Blue, L-Orange") +
        theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
element_text(hjust = 0.5))
```

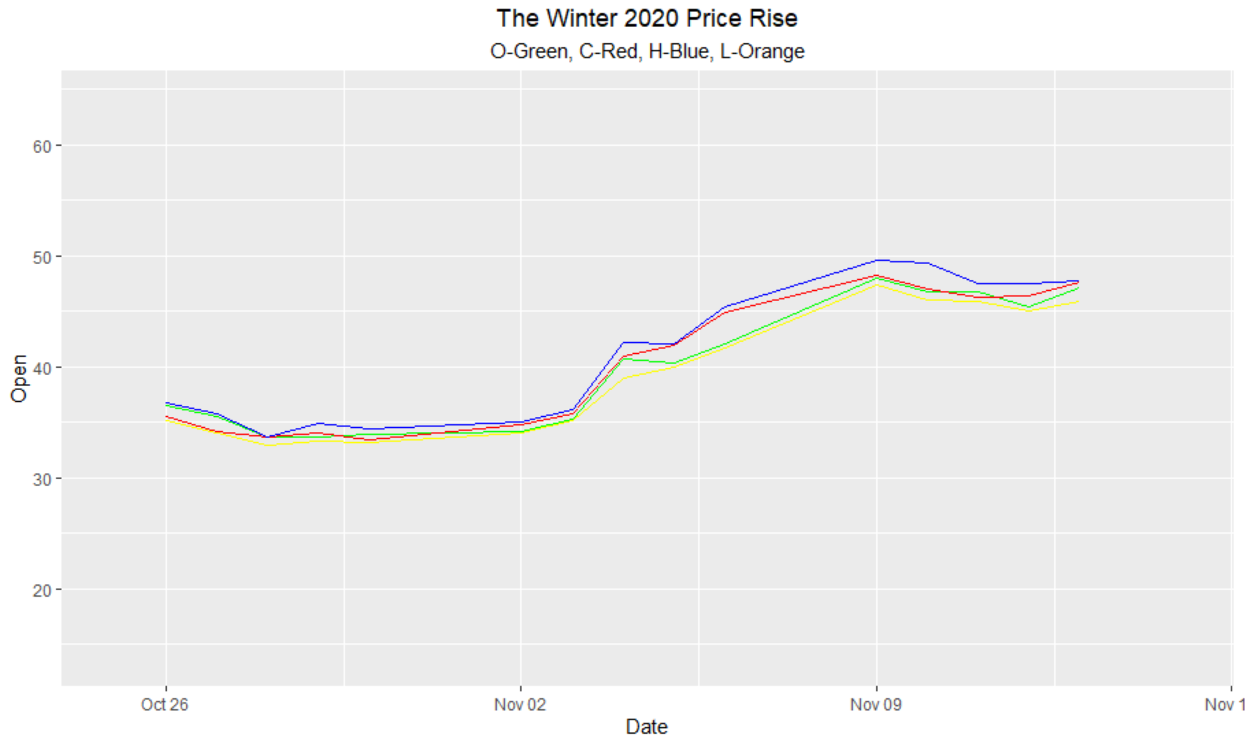
Let's continue looking at the "Price" variables: Open, Close, High & Low. Here is a line graph for the entire date range. We see a few interesting things. There is not much steadiness in the stock price over this time range. We also see some severe drops and rises at certain points. Keep in mind that this data occurs during the COVID-19 pandemic.



Let's take a closer look at the sharp drop in stock price that occurred in early 2020. What originally looked like a sudden severe drop actually happened over a period of about 1.5 months--from late February to about March 18.



Now, let's look at the sudden sharp rise that happened in late 2020. Like before, it doesn't happen as quickly as it seemed. In this case it occurred over a period of about two (2) weeks.



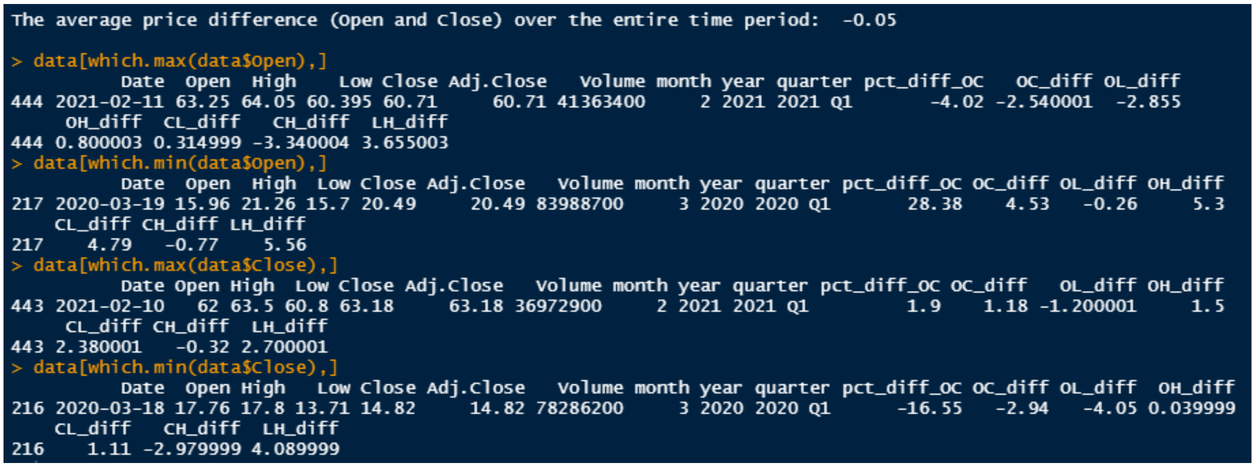
The Winter 2020 Price Rise

O-Green, C-Red, H-Blue, L-Orange

```
#-------Making some calculations and adding some more columns

#Pulling Month/Year/Quarter to separate columns
library(lubridate)
data$month<-month(data$Date) data$year<-year(data$Date)
#data$quarter<-as.numeric(substr(quarters(as.Date(data$Date)), 2, 2))
#data$quarter<-quarters(as.Date(data$Date))
library(zoo)
data$quarter=as.yearqtr(data$Date, format = "%Y-%m-%d")

#Creates volume by quarter in separate dataframe
aggregated_data=aggregate(data$Volume,by=list(data$quarter),FUN=sum)

#Calculating percent change in Open & Close
data$pct_diff_OC<-((data$Close-data$Open)/(abs(data$Open))*100)
data$pct_diff_OC<-as.numeric(format(round(data$pct_diff_OC, 2), nsmall
= 2))

#Making calculated columns of the price differences
```

```
data$OC_diff <- data$Close - data$Open
data$OL_diff <- data$Low - data$Open
data$OH_diff <- data$High - data$Open
data$CL_diff <- data$Close - data$Low
data$CH_diff <- data$Close - data$High
data$LH_diff <- data$High - data$Low

#Print Ave, max, and min of the Open and Close prices--Just to see
cat("\nThe average price difference (Open and Close) over the entire
time period: ",
as.character(round(mean(data$OC_diff), 2)), "\n", "\n")
data[which.max(data$Open),]
data[which.min(data$Open),]
data[which.max(data$Close),]
data[which.min(data$Close),]
```

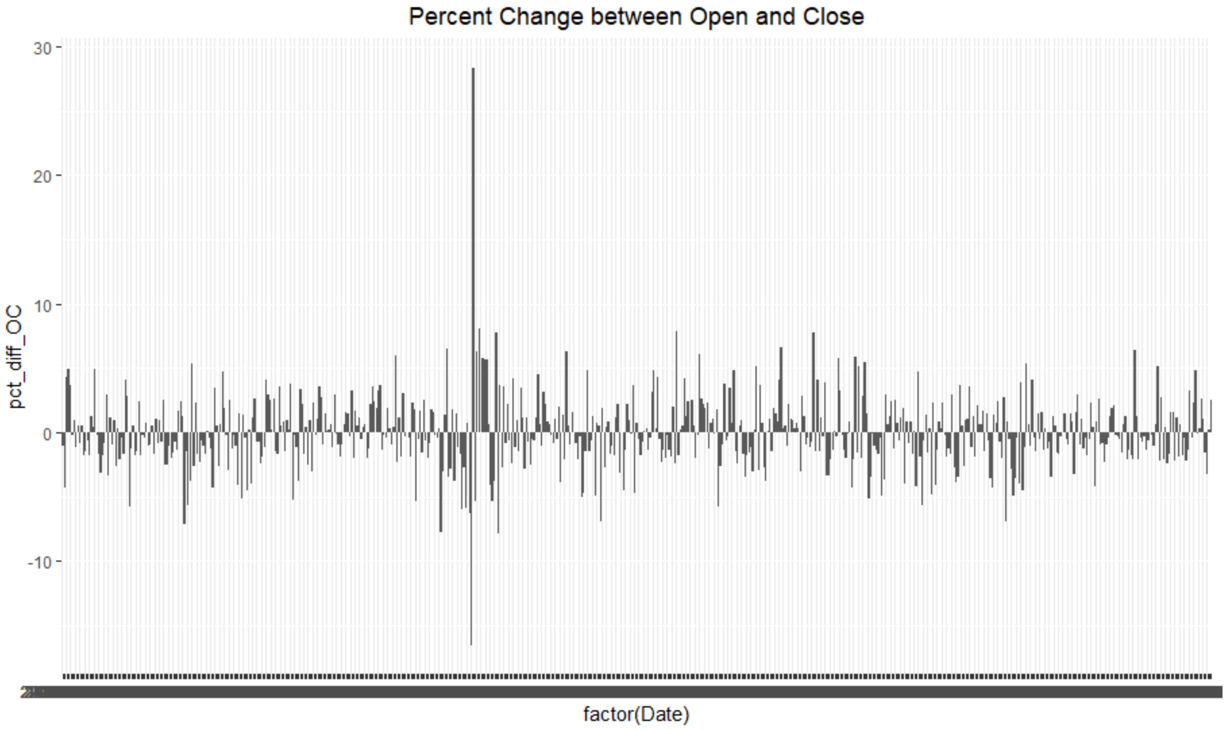
In this code block, I wanted to expand the data and see what can be gained from it. Year, month and quarter were pulled from the data to give finer control over how to display the data. The "Volume" variable was then aggregated by quarter and placed into a new dataframe. The daily percent change was calculated for "Open" and "Close" and price differences were calculated for all price variables. The final code block prints a couple of things I was curious about: Average Open/Close price difference, the MAX Open price, the MIN Open price, the Max Close price, and the MIN Close price. The method I used to get MIN/MAX returned the entire rows of data, which allows us to see the other variables for comparison.



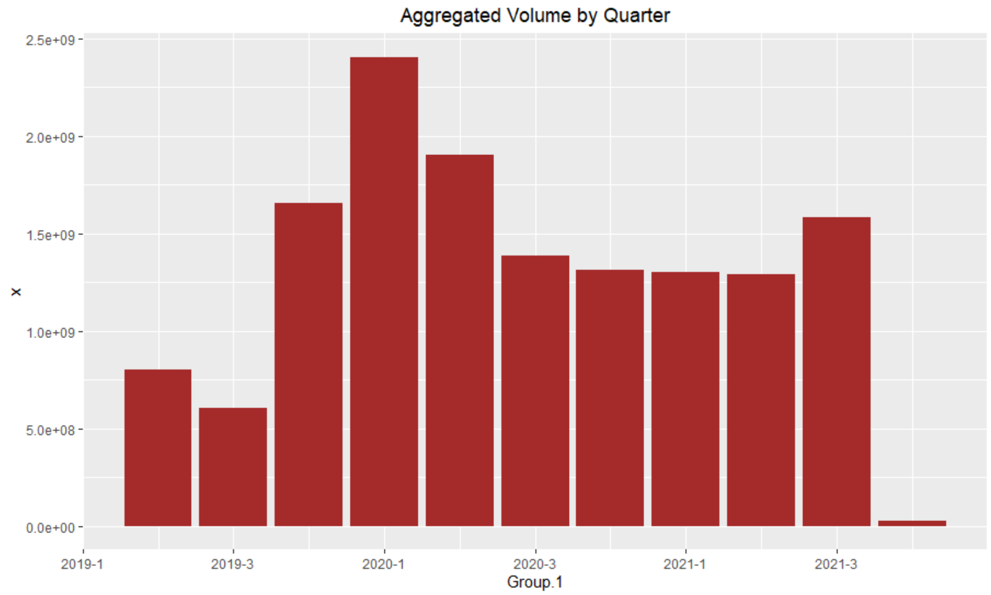
```
The average price difference (Open and Close) over the entire time period:  -0.05

> data[which.max(data$Open),]
          Date  Open  High    Low Close Adj.Close   Volume month year quarter pct_diff_OC   OC_diff OL_diff
444 2021-02-11 63.25 64.05 60.395 60.71     60.71 41363400     2 2021 2021 Q1       -4.02 -2.540001  -2.855
     OH_diff  CL_diff   CH_diff  LH_diff
444 0.800003 0.314999 -3.340004 3.655003
> data[which.min(data$Open),]
          Date  Open  High  Low Close Adj.Close   Volume month year quarter pct_diff_OC OC_diff OL_diff OH_diff
217 2020-03-19 15.96 21.26 15.7 20.49     20.49 83988700     3 2020 2020 Q1       28.38    4.53   -0.26     5.3
    CL_diff CH_diff LH_diff
217    4.79   -0.77    5.56
> data[which.max(data$Close),]
          Date Open High  Low Close Adj.Close   Volume month year quarter pct_diff_OC OC_diff   OL_diff OH_diff
443 2021-02-10   62 63.5 60.8 63.18     63.18 36972900     2 2021 2021 Q1         1.9    1.18 -1.200001     1.5
     CL_diff CH_diff  LH_diff
443 2.380001   -0.32 2.700001
> data[which.min(data$Close),]
          Date  Open High   Low Close Adj.Close   Volume month year quarter pct_diff_OC OC_diff OL_diff  OH_diff
216 2020-03-18 17.76 17.8 13.71 14.82     14.82 78286200     3 2020 2020 Q1      -16.55   -2.94   -4.05 0.039999
    CL_diff   CH_diff  LH_diff
216    1.11 -2.979999 4.089999
```

```
#---------More Visualizations

#Graph the Percent change over entire date range
ggplot(data=data) +
        geom_col(aes(x=factor(Date), y=pct_diff_OC)) +
        ggtitle("Percent Change between Open and Close") +
        theme(plot.title = element_text(hjust = 0.5))

#Column chart of aggregated volume by quarter
ggplot(data=aggregated_data) +
        geom_col(aes(x=Group.1, y=x), fill="Brown") +
        ggtitle("Aggregated Volume by Quarter") +
        theme(plot.title = element_text(hjust = 0.5))

#Making a boxplot of all difference calculations
diff <- data[, c("OC_diff", "OL_diff" , "OH_diff" , "CL_diff" ,
        "CH_diff" , "LH_diff" )]
        boxplot(diff, main="Differences Between Prices" , sub="5-10-
2019 through 10-1-2021" ,
        col=c("darkolivegreen4", "darkorange3" , "cyan4" ,
        "indianred3" , "navajowhite2" , "palevioletred2" ),
        ylab="Price" , xlab="Differences" ,
        names=c("Open & Close", "Open & Low" , "Open & High" ,
        "Close & Low" , "Close & High" , "Low & High" ))
```

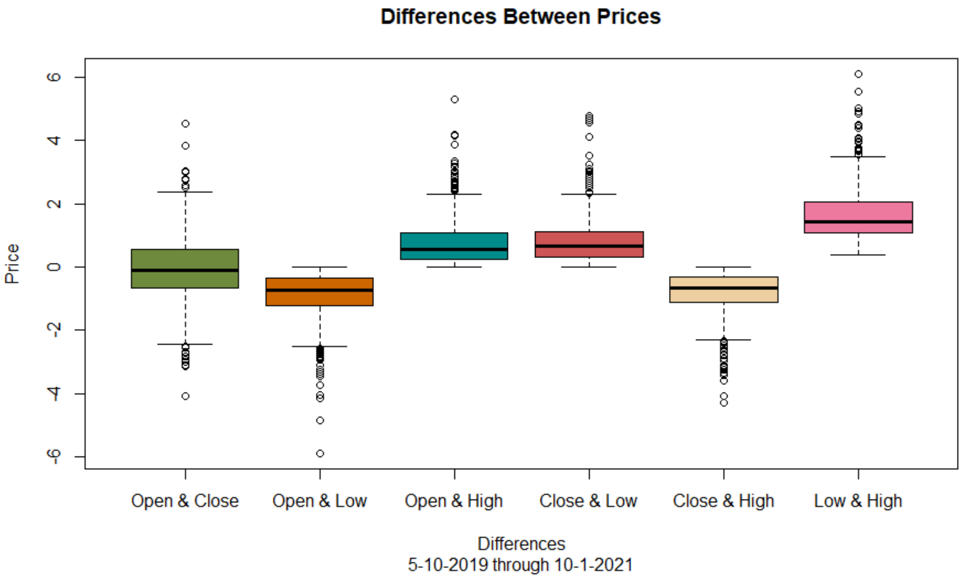
Now that those calculations are complete, let's graph them. We'll begin with the percent change of the Open and Close prices. Now, this graph looks very busy, but it is interesting to see the positive and negative changes between Open and Close prices.



Next is a graph of the aggregated Volume by quarter data. One item of potential interest is that we see an increase in trades entering the fourth (4th) quarter of 2019, first (1st) and second (2nd) quarters of 2020. We see what could be a repeat in the third (3rd) quarter of 2021, but the dataset ends early. It would be interesting to see if that trade volume increase occurs every year.



The final graph of this project is a boxplot of the differences in prices. An interesting item is how narrow the range is for all of the price differences. If we recall the scatterplot of "Open" and "Close", the data was mostly tightly packed. This boxplot reinforces that point. It also leads to the idea that the price of the stock generally doesn't stray to far from the "Open" price, regardless is the price is rising or falling.


Differences Between Prices
Price
Open & Close
Open & Low
Open & High
Close & Low
Close & High
Low & High
Differences
5-10-2019 through 10-1-2021