2018

**Created by - Sohong Chakraborty and Zeeshan Khan**



Stock Market Analysis During the Great Recession of 2007-09 using R

1. **Data set URL’s**

* <http://eoddata.com/products/historicaldata.aspx>

The reason for choosing one comprehensive dataset is that, the dataset is compact with all the required columns present to carry out a holistic analysis. Stocks have limited attributes associated with them. As a result, one comprehensive dataset proved enough to carry out this study.

The data sets contain data of all the companies listed in NASDAQ (1) and NYSE (2) from the last quarter of 2008 and ‘09 (i.e July to September). Following are the columns:

* **<ticker>** : This column contains information of tickers of various listed companies
* **<date>** : This column contains information about the date of transaction in the market
* **<open>** : This column contains information about the opening price of company stocks
* **<high>** : This column contains information about the intra-day high price
* **<close>** : This column contains information about the closing price of company stocks
* **<vol>** : This column contains data about the total number of stocks traded during the day

One more advantage of having this dataset is that stocks from both NASDAQ and NYSE have the same attributes, i.e the same columns. So, joining the two datasets – NASDAQ.xls and NYSE.xls never turned out to be an arduous job.

The reason we chose to carry out analysis of stocks from the recession period is that, studying the trend analysis of the stocks and their behaviour can help predict future recessions and enable us to take preventive measures. There are multiple factors which leads to a recession. So, careful predictive trend analysis can help solve that problem.

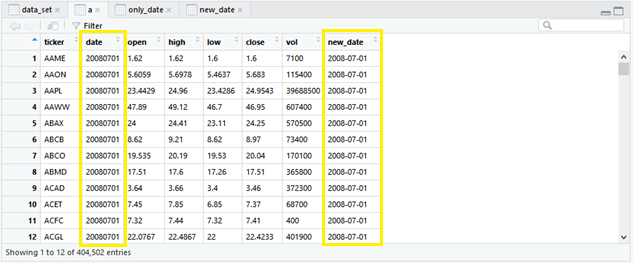
(1) NASDAQ - "National Association of Securities Dealers Automated Quotations"

(2) NYSE – “New York Stock Exchange”

1. **Data Set Cleaning**

**Removing duplicate column:**

**Before:**



**Code:**

> clean\_data <- a[,-2]

> View(clean\_data)

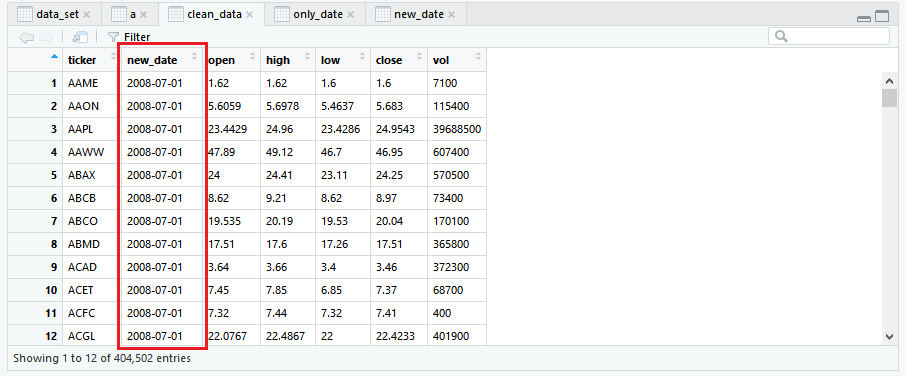
> colnames(clean\_data)

[1] "ticker" "open" "high" "low" "close" "vol" "new\_date"

> clean\_data <- clean\_data[c(1,7,2,3,4,5,6)]

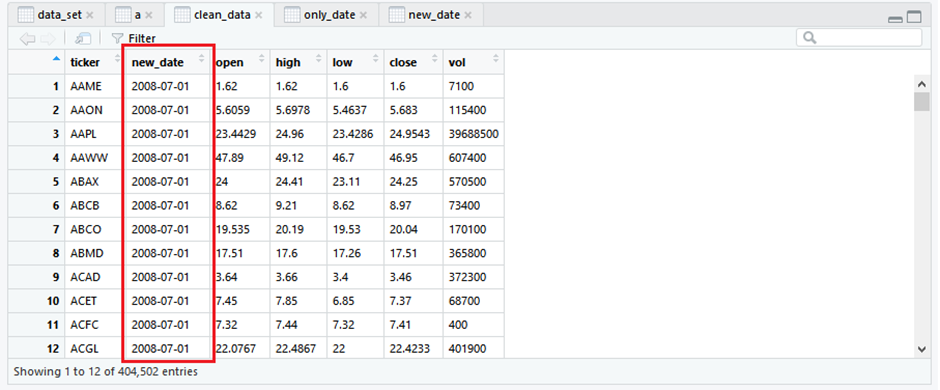
> View(clean\_data)

**After:**



**Splitting columns:**

**Before:**

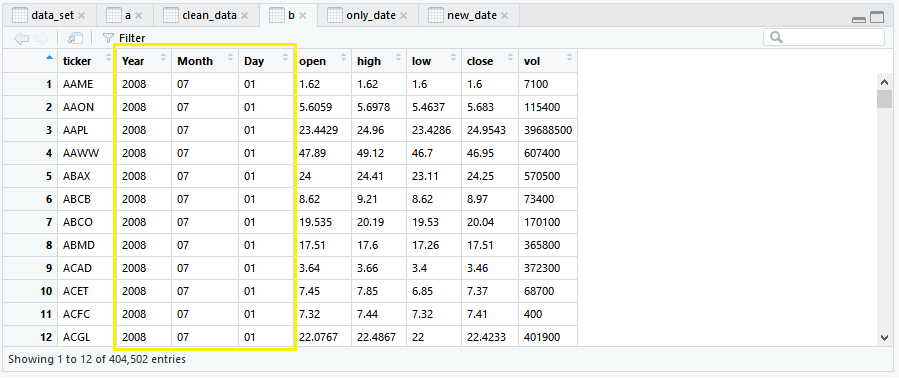


**Code:**

> b <- separate(clean\_data,new\_date,c("Year","Month","Day"), sep='-')

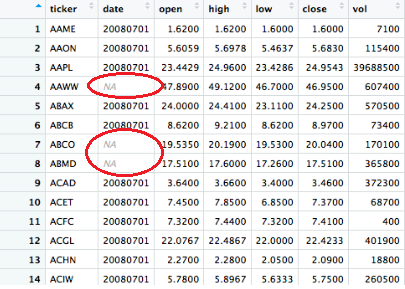
> View(b)

**After:**



**Removing null values:**

**Before:**



**Code:**

> i<- read.csv("cleaning.csv" , header = T)

> View(i)

> date <- i$date

> mean(date, na.rm=TRUE)

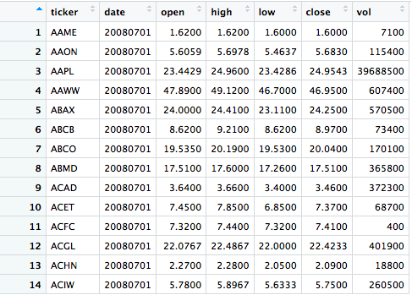
[1] 20080701

> date[is.na(date)]=mean(date, na.rm=TRUE)

> i$date<- date

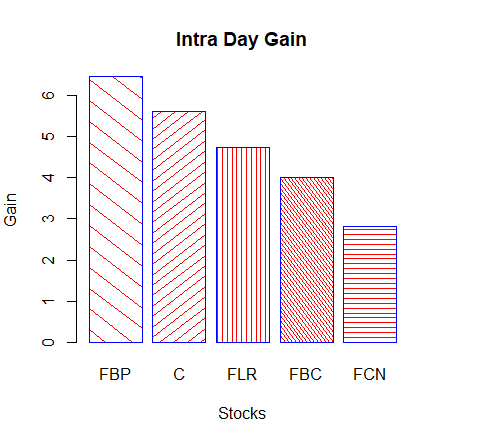
> View(i)

**After:**



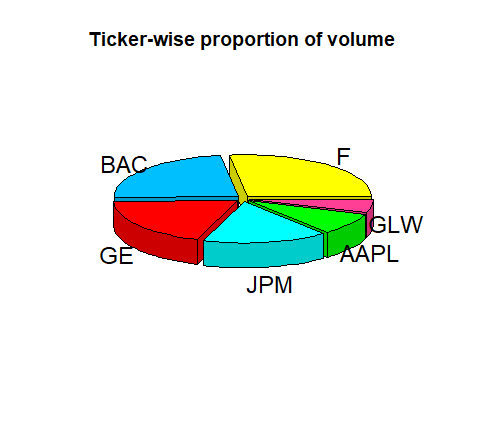
1. **Analysis & Visualizations**

**Which are the top 5 stocks with highest intra-day gain?**



Intraday trading as the name suggests refers to the trading system where one must square-off the trade on the same day. Squaring off the trade means that one must do the buy and sell or sell and buy transaction on the same day before the market closes. Intraday Trading is also referred to as Day trading by many traders. Intra-day gain means the difference between the opening value of the stock and the closing value. If the difference comes out to be positive, it is called intra-day gain, else, it is called intra-day loss. Out of the top 5 stocks which have seen a surge in their pricing, three of them are bank stocks. First BanCorp (FBP) is a publicly owned financial holding company located out of San Juan, Puerto Rico. Citigroup Inc. or Citi (C) is an American multinational investment bank and financial services company headquartered in New York City, New York. It saw a pretty hefty intra-day gain of $5.60 on 1st July 2008. Another financial institution which features in the top 5 is the Flagstar Bancorp, Inc. which operates as the holding company for Flagstar Bank, a federally chartered stock savings bank. From the above trend it quite clear that, during the volatile times of a recession the stocks of any financial institution are monitored closely as they are the ones to be first affected. While stocks of some financial institutions bear the grunt of panic selling, other may eventually come out as clear gainers. But, all of this depends upon how the reserve bank of any country acts during that time to stop the volatility in the market!

**Which are the top 6 stocks with highest proportion of volumes traded?**

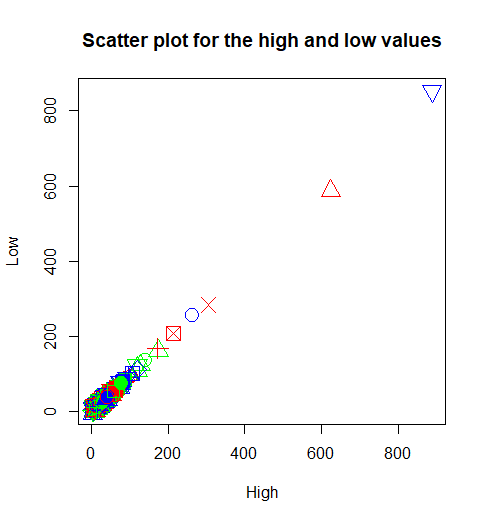


The above visualization shows the top 6 company stocks that have had the highest proportion of volume during the recession of 2008-09. It is clearly visible from the above depiction that Bank of America Corporation (BAC) leads the ranking with around 22 billion shares traded during the said time-period. This is normal in trend with a recession, during which people seem to lose faith in the banking institutions and undertake panic selling. Over the short term, the financial crisis affects the banking sector by causing banks to lose money on mortgage defaults, interbank lending to freeze, and credit to consumers and businesses tend to dry up. Thus, people tend to play safe and limit their losses by selling the already plunging stock values.

Another financial institution which features on the top 10 lists is J.P Morgan & co. This tends to show that the financial sector is among the worst to be affected. According to an excerpt, “*the financial meltdown in 2008 was not the result of ethereal and enigmatic forces, it was the result of fraud in the financial markets. That truth was initially hidden by politicians, regulators, banks, and the bought priesthood of mainstream economics. But despite all the lies, the day of reckoning has finally arrived for one of the world’s most powerful corporations, J.P Morgan. Of course, there won’t be jail time for executives – prison is for the little people. But, contrary to some Republican presidential candidates, corporations aren’t people. They are mindless engines of profit and the only acts of contrition the machines can output is in the form of paying money and JPMorgan has already agreed to pay $13 billion for crashing the economy in 2008 through fraud and greed*.” (3) This event bolsters the fact that people lose all faith in the financial institutions, who are regarded as culprit behind any economic meltdown. Even behemoths like J.P Morgan & co. who contribute greatly to the GDP of a nation like the US, falter due to their greed and hunger for power.

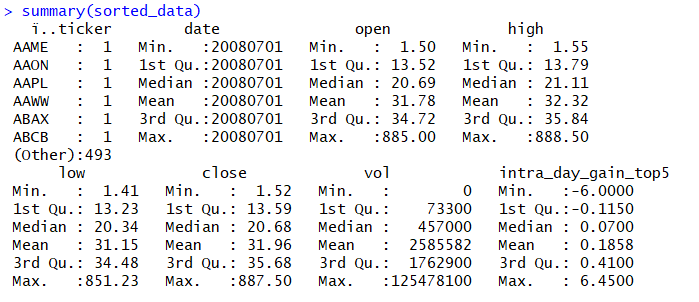
(3) <https://shadowproof.com/2013/10/21/jpmorgan-to-pay-13-billion-for-causing-2008-financial-crisis/>

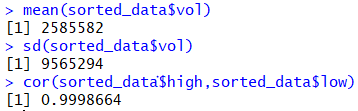
**What trend do you see between the High and Low values of the stocks?**



Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation. The closer the data points come when plotted to making a straight line, the higher the correlation between the two variables, or the stronger the relationship. If the data points make a straight line going from the origin out to high x- and y-values, then the variables are said to have a positive correlation. If the line goes from a high-value on the y-axis down to a high-value on the x-axis, the variables have a negativecorrelation. This scatter plot shows the variance of High with Low values of the stock. The curve seems to be fairly-linear, signifying that there is a strong positive correlation between the two attributes - High & Low. This is in tandem with the recession period where the market remains highly volatile and the high and low values of any stock vary correspondingly.

**Statistical Summary and Functions:**





It is clear from the statistical data the data is evenly spread out. The maximum and the minimum values of each column are far apart from each other, signifying a wide range. The mean is somewhere close to the middle, signifying an overall even distribution of data (even though some of the columns do have a lot of outliers which result in the huge variance of the data in the column).

The high mean of the volume of stocks traded defines that during the recession there is always high trading of data due to panic buying and selling of stocks. This is in synchronization with the traits of a recession period. The high standard deviation of the volume column suggest that the range of stocks exchanged is quite high. While some stocks may have seen negligible trading, some have been traded on magnanimous scale (especially the stocks of financial institutions). The high correlation between the high value and the low value of the stocks signify that there is a positive correlation between the two.

**Codes:**

**Data Cleaning:**

> setwd("C:/Users/Hp/Downloads")

> data\_set <- read.csv("NASDAQ and NYSE.csv", header = TRUE, sep = ',')

> View(data\_set)

> only\_date <- data\_set[,2]

> View(only\_date)

> new\_date <- as.Date(only\_date,"%Y%m%d"); new\_date

> View(new\_date)

> a <- within(data\_set,{new\_date <- as.Date(only\_date,"%Y%m%d")})

> View(a)

> clean\_data <- a[,-2]

> View(clean\_data)

> colnames(clean\_data)

[1] "ticker" "open" "high" "low" "close" "vol" "new\_date"

> clean\_data <- clean\_data[c(1,7,2,3,4,5,6)]

> View(clean\_data)

> install.packages("tidyr")

Installing package into ‘C:/Users/Hp/Documents/R/win-library/3.2’

(as ‘lib’ is unspecified)

also installing the dependencies ‘utf8’, ‘bindrcpp’, ‘glue’, ‘pillar’, ‘dplyr’, ‘Rcpp’, ‘rlang’, ‘tibble’, ‘tidyselect’

> library(tidyr)

Warning message:

package ‘tidyr’ was built under R version 3.2.5

> b <- separate(clean\_data,new\_date,c("Year","Month","Day"), sep='-')

> View(b)

> i<- read.csv("cleaning.csv" , header = T)

> View(i)

> date <- i$date

> mean(date, na.rm=TRUE)

[1] 20080701

> date[is.na(date)]=mean(date, na.rm=TRUE)

> i$date<- date

> View(i)

**First Visualization:**

> data <- read.csv("Book1.csv", header = TRUE, sep = ',')

> View(data)

> sorted\_data <- data[order(data$vol),]

> View(sorted\_data)

> sorted\_data <- data[order(-data$vol),]

> View(sorted\_data)

> install.packages("plyr")

Installing package into ‘C:/Users/Hp/Documents/R/win-library/3.2’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.2/plyr\_1.8.4.zip'

Content type 'application/zip' length 1121290 bytes (1.1 MB)

downloaded 1.1 MB

package ‘plyr’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Hp\AppData\Local\Temp\Rtmp0C7Ydv\downloaded\_packages

> library(plyr)

Warning message:

package ‘plyr’ was built under R version 3.2.5

> install.packages("plotrix")

Installing package into ‘C:/Users/Hp/Documents/R/win-library/3.2’

(as ‘lib’ is unspecified)

There is a binary version available but the source version is later:

binary source needs\_compilation

plotrix 3.6-4 3.7-1 FALSE

installing the source package ‘plotrix’

trying URL 'https://cran.rstudio.com/src/contrib/plotrix\_3.7-1.tar.gz'

Content type 'application/x-gzip' length 247706 bytes (241 KB)

downloaded 241 KB

\* installing \*source\* package 'plotrix' ...

\*\* package 'plotrix' successfully unpacked and MD5 sums checked

\*\* R

\*\* data

\*\* demo

\*\* inst

\*\* preparing package for lazy loading

\*\* help

\*\*\* installing help indices

\*\* building package indices

\*\* testing if installed package can be loaded

\*\*\* arch - i386

\*\*\* arch - x64

\* DONE (plotrix)

The downloaded source packages are in

‘C:\Users\Hp\AppData\Local\Temp\Rtmp0C7Ydv\downloaded\_packages’

> library(plotrix)

> new\_data <- sorted\_data[1:6,]

> View(new\_data)

> colors <- c("yellow", "deepskyblue", "red", "cyan", "green1", "violetred1")

> pie3D(new\_data$vol, labels = new\_data$ï..ticker, explode = 0.05, main = "Ticker-wise proportion of volume", col = colors)

**Second visualization:**

> f <- sorted\_data$close

> g <- sorted\_data$open

> highest\_intra\_day\_gain <- f – g

> sorted\_data <- sorted\_data[order(-sorted\_data$intra\_day\_gain\_top5),]

> View(sorted\_data)

> x <- sorted\_data[1:5,]

> View(x)

> barplot(x$intra\_day\_gain\_top5, main="Intra Day Gain", xlab="Stocks", ylab="Gain or Loss", names.arg=x$ï..ticker,col="red", border="blue",density=c(5,10,20,30,15) , angle=c(-45,45,90,-60,180))

**Third visualization:**

> s <- sorted\_data$high

> t <- sorted\_data$low

> scat <- plot(sorted\_data$high, sorted\_data$low, xlab = "High",ylab = "Low", col=c("red","green","blue"), pch=1:25, cex=2, main = "Scatter plot for the high and low values")