|  |  |
| --- | --- |
| BREAKING DOWN “new york stock exchange - NYSE”  SUBMITTED TO: DR. SHILPA BALAN | Abstract  Can trading be fully automated? It is said that 30% of traffic on stocks is already generated by machines. Using R, we will try to fetch some interesting insights based on historic data about NYSE like ROI, chance of bankruptcy, under-valued and/or over-valued companies.  Manali Joshi  CIS: 5270 – Business Intelligence |

* **Objective of the Study:**

Can trading be fully automated? It is said that 30% of traffic on stocks is already generated by machines. Using R, we will try to fetch some interesting insights based on historic data about NYSE like ROI, chance of bankruptcy, under-valued and/or over-valued companies. This dataset also contains columns like Effect of Exchange Rate, Earning Before and After Tax, Fixed Assets, Gross Margin & Gross Profit, Investments, Liabilities, Long-Term Debts and Investments, Net Income, Opening and Closing of Stock Prices, Volumes, High, Low. All the previously listed columns have numeric values, so it’s very favorable for doing statistical analysis using R. We can predict the best company or listed security to invest into, forecasting of stock prices of companies, comparison of expected profit and actual profit, net income increase or decrease. This analysis will help investors or traders with a detailed view and prediction about New York Stock Exchange.

* **DataSet URL:** <https://www.kaggle.com/dgawlik/nyse>

**DataSet Format:** CSV

* **About the DataSet and Project Topic:**

The New York Stock Exchange is an American stock exchange, it is abbreviated as NYSE

and, sometimes called “The Big Board”

It is by far the world’s largest stock exchange by market capitalization of its listed

companies at US$ 19.3 trillion as of June 2016. [1] [2]

The NYSE is still by far the most important equity market in the world. With a market cap

of about $21 trillion, the NYSE about three times larger than NASDAQ (National

Association of Securities Dealers Automated Quotations), and the two US exchanges

together have a larger market cap than the next ten exchanges combined, which includes

Japan Stock Exchange, London Stock Exchange, Hong Kong Stock Exchange & Shanghai

Stock Exchange. [3]

The New York Stock Exchange Dataset consists of following files:

• Prices.csv: raw, as-is daily prices. Most of data spans from 2010 to the end 2016, for companies new on stock market date range is shorter. There have been approx. 140 stock splits in that time, this set doesn't account for that.

• Prices-Split-Adjusted.csv: same as prices, but there have been added adjustments for splits.

• Securities.csv: general description of each company with division on sectors

• Fundamentals.csv: metrics extracted from annual SEC 10K fillings (2012-2016), should be enough to derive most of popular fundamental indicators.

Prices were fetched from Yahoo Finance, fundamentals are from Nasdaq Financials, extended by some fields from EDGAR SEC databases.

**Questions:**

Q1: Maximum number of highs for the 5 IT Companies.

Q2: Comparison of Profit vs Investment.

Q3: Comparison of actual profit vs expected profit.

Q4: Calculate Return on Investment.

**References:**

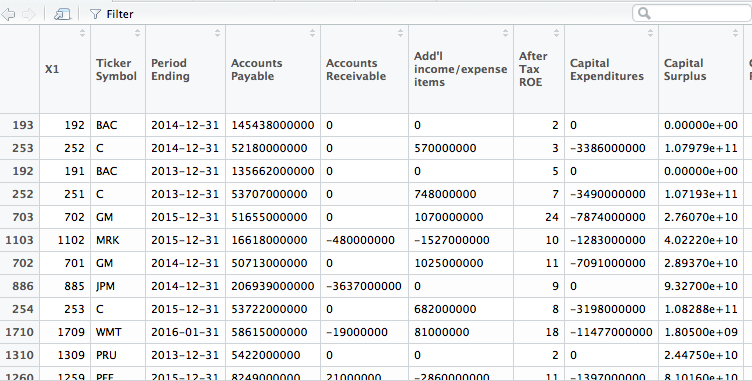
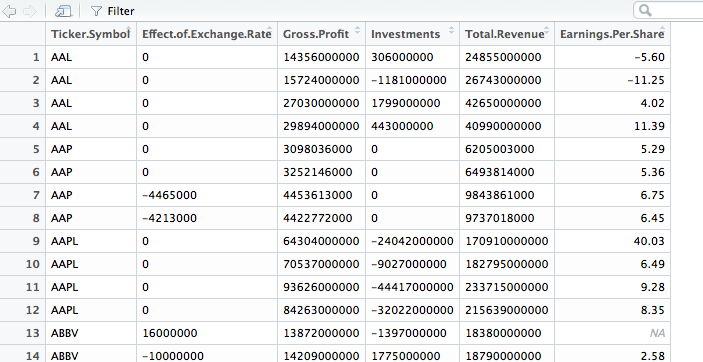
[1] NYSE Group Shares Outstanding and Market Capitalization of Companies Listed,2016

<http://www.nyxdata.com/nysedata/asp/factbook/viewer_edition.asp?mode=tables&key=333&category=5>

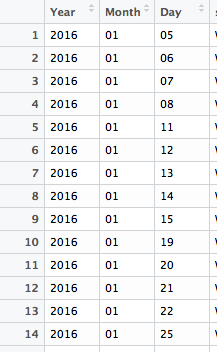
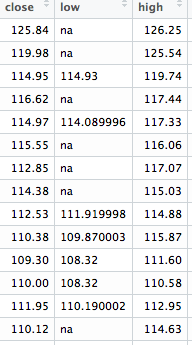
[2] <http://finance.zacks.com/new-york-stock-exchange-largest-stock-market-world-5426.html>

[3] <http://www.businessinsider.com/global-stock-market-capitalization-chart-2014-11?IR=T>

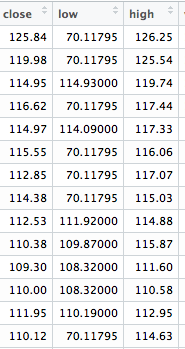
**Data Cleaning:**

* A). Removing Columns
* Before
* 
* After
* 

In data cleaning part-A I’ve removed columns like Period Ending, Account Payable, Account Receivable, Capital Expenditures, etc. which are not useful further for analysis or visualization.

* B). Splitting Data
* Before
* 
* After
* 
* Data cleaning part-B includes splitting of data, i.e. Date Column is spited into 3 columns Year, Month & Day.
* C). Replacing Null Values
* Before
* 

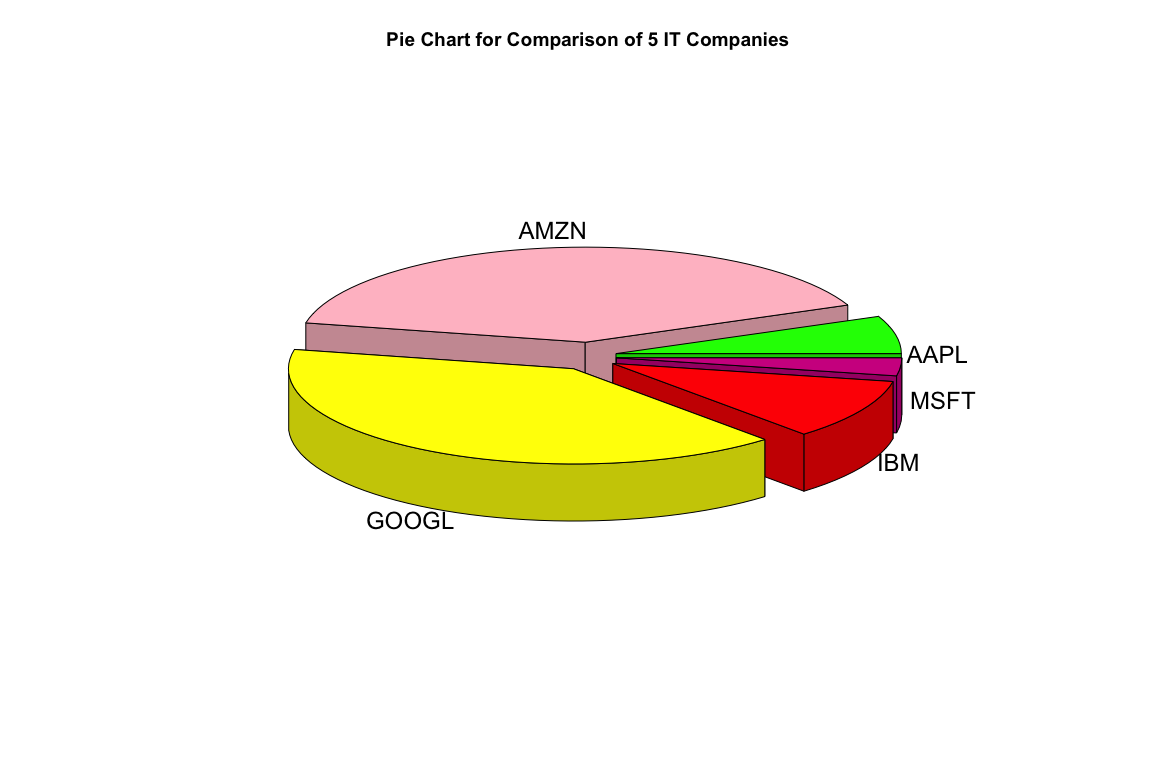
After



Data cleaning part-C I’ve replaced null values with mean values for the same column.

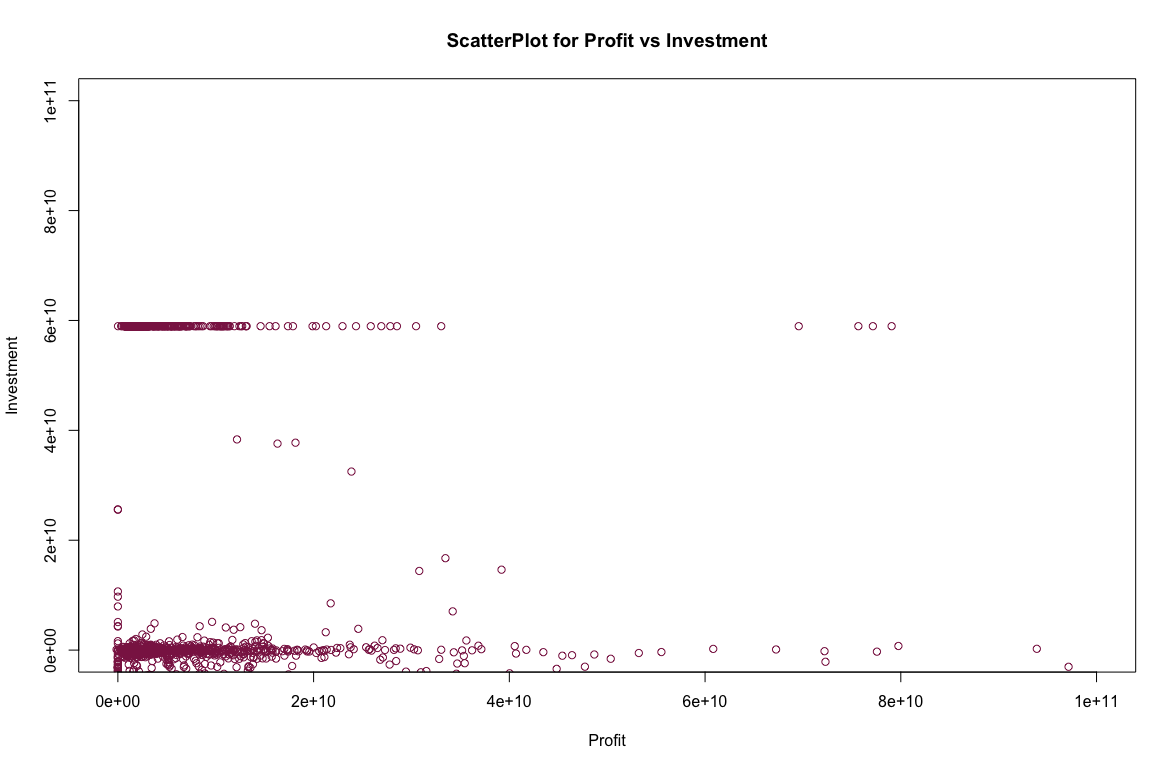
**Visualization:**

1). Pie Chart showing maximum number of highs for the 5 IT Companies.



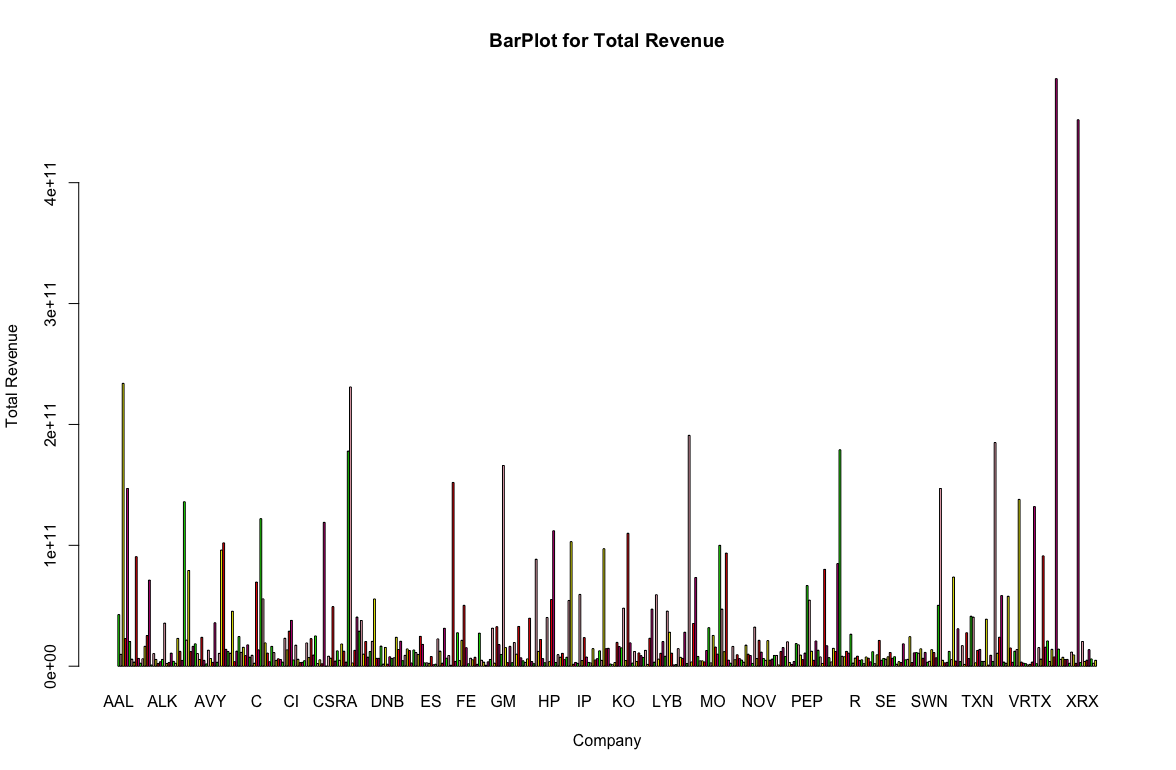
The above given pie-chart shows maximum number of highs of the 5 IT Companies which are Amazon, Google, IBM, Apple & Microsoft. As the market opens everyday it hits the maximum, which is the maximum price at which the shares have been sold. Here with using the maximum function I’ve aggregated the highest throughout the year for each company. We can see that Amazon has the maximum high value amount for their shares, followed by google on second, then IBM on third, Apple after that & Microsoft at the last among them.

2). Scatter Plot for comparing Profit vs Investment



Scatter Plots are best to depict the relationships between two attributes. This Scatter Plot shows comparison of total money invested by company vs profit gained by company. For any company, Return on Investment is one of the crucial factors. Hence, this graph shows how much benefit, i.e. profit the company has gained. From the Scatter Plot, we can conclude that more the investment, more the profit. On the contrast, we can also say that companies which has invested less also got good profit.

3). Comparison of Total Revenue



Above given Bar Plot shows total revenue for the companies in NYSE and NASDAQ. We have total of 448 companies like Apple, Google, Microsoft, Cisco, VMware, IBM. The maximum revenue is more than 4e+11. As there were lots of data for each company’s total revenue, we extracted the highest revenue for the span of 4 years, i.e. 2012-2016 and represented it using Bar Plot. The maximum revenue was found to exist for the ticker symbol “WMT” which is of Wall-mart.

**Code Used in the Project:**

**Data Cleaning-A**

> data1<-read.csv("fundamentals.csv", header = TRUE)

> View(data1)

> keepcols<-c(2,21,26,31,75,78)

> data4<-data1[,keepcols]

> View(data4)

**Data Cleaning-B**

>data5<-read.csv("prices-split-adjusted.csv", header = TRUE)

> View(data5)

> install.packages("tidyr")

> library("tidyr")

> data6<-separate(data5,date,c("Year","Month","Day"),sep="-")

> View(data6)

**Data Cleaning-C**

> data7<-read.csv("prices.csv", header = TRUE)

> low<-data7$low

> mean(low,na.rm=TRUE)

[1] 70.11795

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

> View(data7)

**Visualization-1:**

>install.packages("plyr")

>library("plyr")

>install.packages("plotrix")

>library("plotrix")

>colors1<-c("Green","Pink","Yellow","Red","VioletRed")

>data9<-read.csv("Top5.csv", header = TRUE)

>View(data9)

>data11<-aggregate(data9$high, by=list(data9$symbol), FUN=max)

>View(data11)

>pie3D(data11$x, labels = data11$Group.1, explode = 0.1, main = "Pie Chart for Comparison of 5 IT Companies", col = colors1)

**Visualization-2:**

> data1<-read.csv("fundamentals.csv", header = TRUE)

> View(data1)

> keepcols<-c(2,21,26,31,75,78)

> data4<-data1[,keepcols]

> View(data4)

>plot(data4$Gross.Profit ,data4$Investments, xlim=c(0,100000000000), ylim=c(0,100000000000),xlab = "Profit", ylab="Investment", main="ScatterPlot for Profit vs Investment",col= "VioletRed4")

**Visualization-3:**

> data1<-read.csv("fundamentals.csv", header = TRUE)

> View(data1)

> keepcols<-c(2,21,26,31,75,78)

> data4<-data1[,keepcols]

> View(data4)

>data00<-aggregate(data4$Total.Revenue, by=list(data4$Ticker.Symbol), FUN=max)

> View(data00)

> barplot(data00$x,main = "BarPlot for Total Revenue",names.arg = data00$Group.1,xlab = "Company",ylab = "Total Revenue",col = colors1)

**Script: (Used in Visualization 2)**

data9<-read.csv("Top5.csv", header = TRUE)

data11<-aggregate(data9$high, by=list(data9$symbol), FUN=max)

barplot(data11$x,main = "Highest of the companies",names.arg = data11$Group.1,xlab = "Companies",ylab = "High",col = colors1)

>source("Bar.R")

**Function: (Function answers the Question-4: Calculate Return on Investment)**

ROI<- function ()

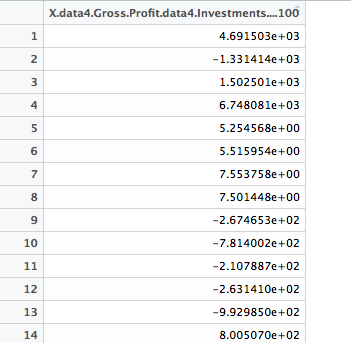
{

df<-data.frame((data4$Gross.Profit/data4$Investments)\*100) a

View(df)

}

>source("ROI.R")



**Statistical Function:**

1). Calculate Mean (Used in Data Cleanig-C)

> data7<-read.csv("prices.csv", header = TRUE)

> low<-data7$low

> mean(low,na.rm=TRUE)

[1] 70.11795

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

> View(data7)

2). SD:

> sddf<-data.frame(sd(data12$x))

> View(sddf)

