PROJECT REPORT ON

ANALYSIS OF TOP 100 STOCKS OF US MARKET

(BY WEIGHTED ALPHA AND PRICE VOLUME)

COURSE: Collecting Storing and Retrieving Data (DA 5020)

22 APRIL 2017

SUPERVISED BY: PROF. KATHLEEN DURANT

COMPILED BY:

YASH MEHTA - 001671624

SUMIT GUPTA - 001218158

Acknowledgement

We wish to express our sincere gratitude to Prof. Kathleen Durant for providing us an opportunity to undertake this project and guide us till its completion. Her class lectures, the various assignments and the lecture notes posted on blackboard were very helpful in realizing the objectives of the course and implementing them on completing this project. The valuable feedback received after the project proposal presentation from the professor and fellow students on discussion board was very helpful in incorporating those suggestions in our work. The results achieved have been a success primarily owing to the feedback received at the earlier stages.

We would also like to thank our TA(s) Mr. Jesse and Mr. Japan who have supported us throughout this coursework with their valuable inputs through learning sessions as well as clearing our doubts in TA hours. They have also helped us by providing valuable feedback on weekly assignments and on fundamentals which helped us in successfully executing this project.

Table of Contents

Abstract 4

Data Collection 5

Data Cleaning and Storing 10

Data Retrieval 12

Data Analysis: 13

- Analysis of Price Volume Category 14
- Analysis of Weighted Alpha Category 16
- Weighted alpha vs Price volume 18

Learning Outcomes 21

Future Scope 22

Abstract

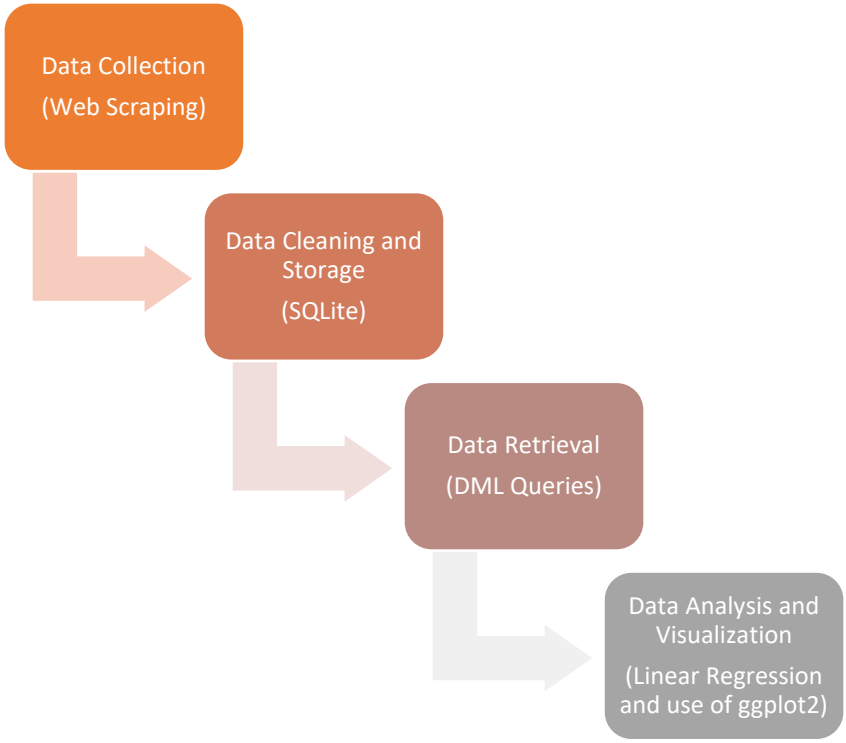
Stock values represent the price of a company's shares, which, is often used as an indication of the overall strength and health of a company. In general, if a company's share price has continued to climb over time, the company and its management are considered to be doing a good job.

Stock values are generally measured on two important parameters: **weighted alpha and price volume**. Stocks ranked based on weighted alpha measure how much a stock value has changed in a period of one year. Stocks ranked by price volume are a measure of the last closing price times volume, divided by 1000.

www.barchart.com is the leading provider of market data solutions for individuals and businesses. Analysis of top 100 stocks based on weighted alpha and price volume will provide us an insight into the relationship between various company fundamentals and their interactions which influence the stock values. An effort will be made to analyze the difference between the stocks from two different categories.

Through this project, our main objective is to extensively work on three main aspects of this course viz. Collecting, Storing and Retrieving Data by using a real-world data set. To establish relationship between the various company fundamentals (dependent and independent) we intend to make use of the statistical method of multiple linear regression. This statistical analysis at the end will help to establish relationship between the company fundamentals and their effect on stock values.

Here is a pictorial representation of workflow of this project:

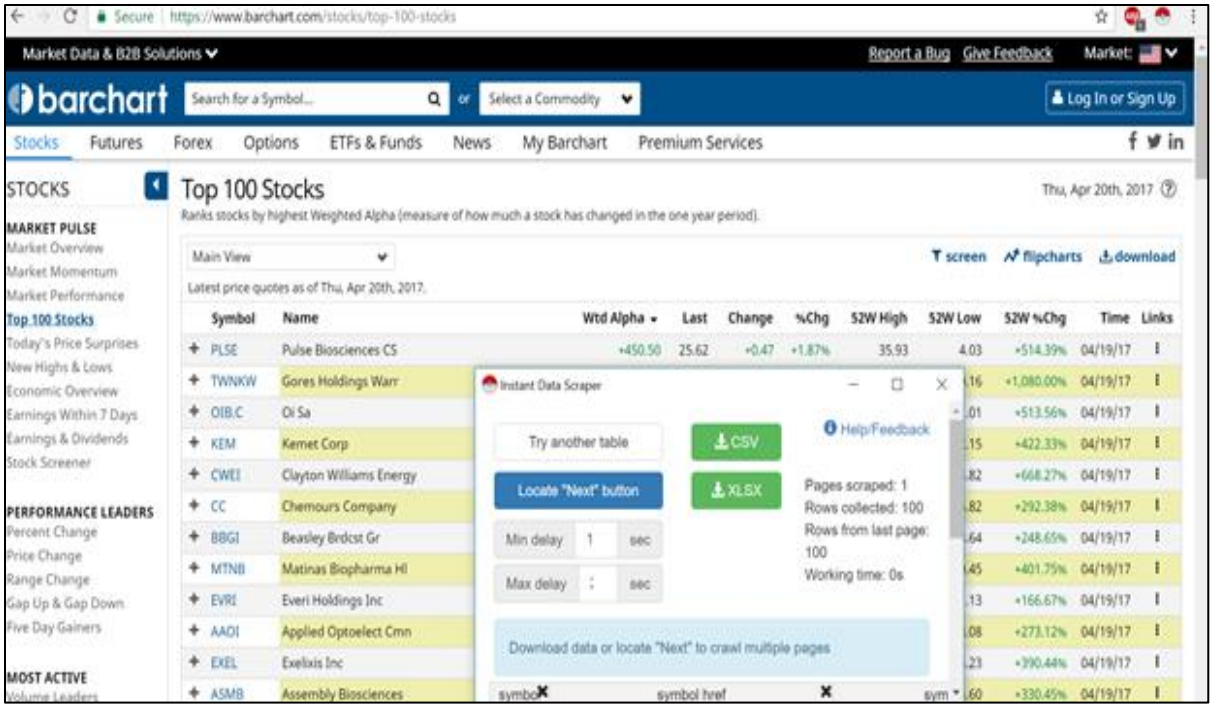


Data Collection

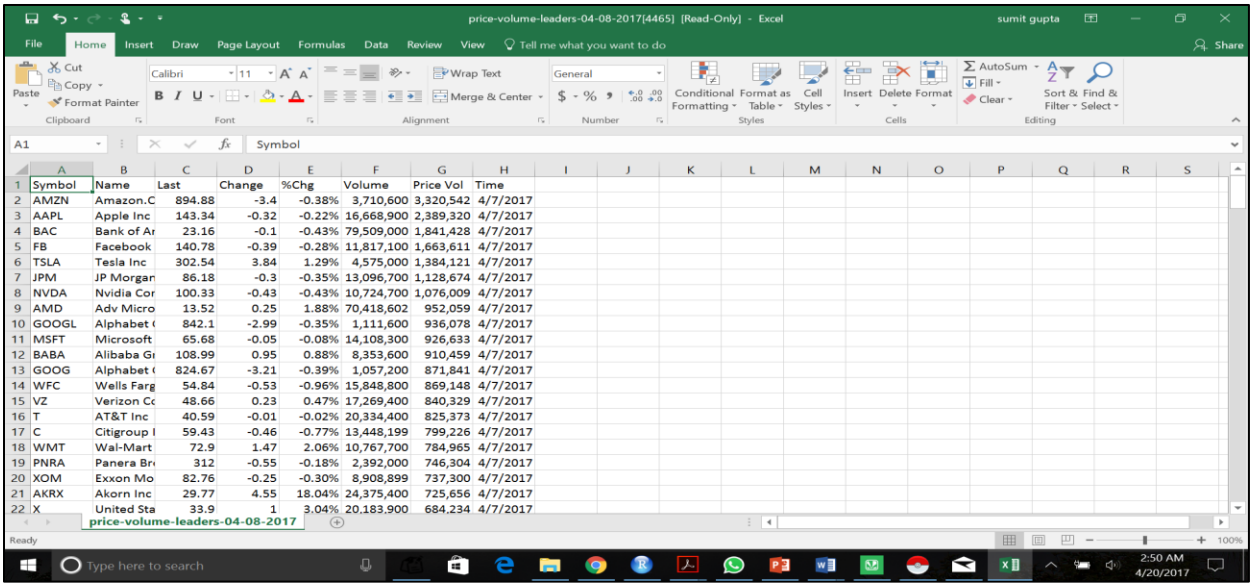
The stocks data and related company details have been web scraped in three different ways for different data values:

1. **List of top 100 stocks by weighted alpha and top 200 stocks by price volumes:** Collecting data of top 100 stocks from www.barchart.com (both weighted alpha and price volumes) using a web scraper i.e. Instant Data Scraper. Both the scraped files were saved in .csv format in the R working directory. These files were then read into the R-code.

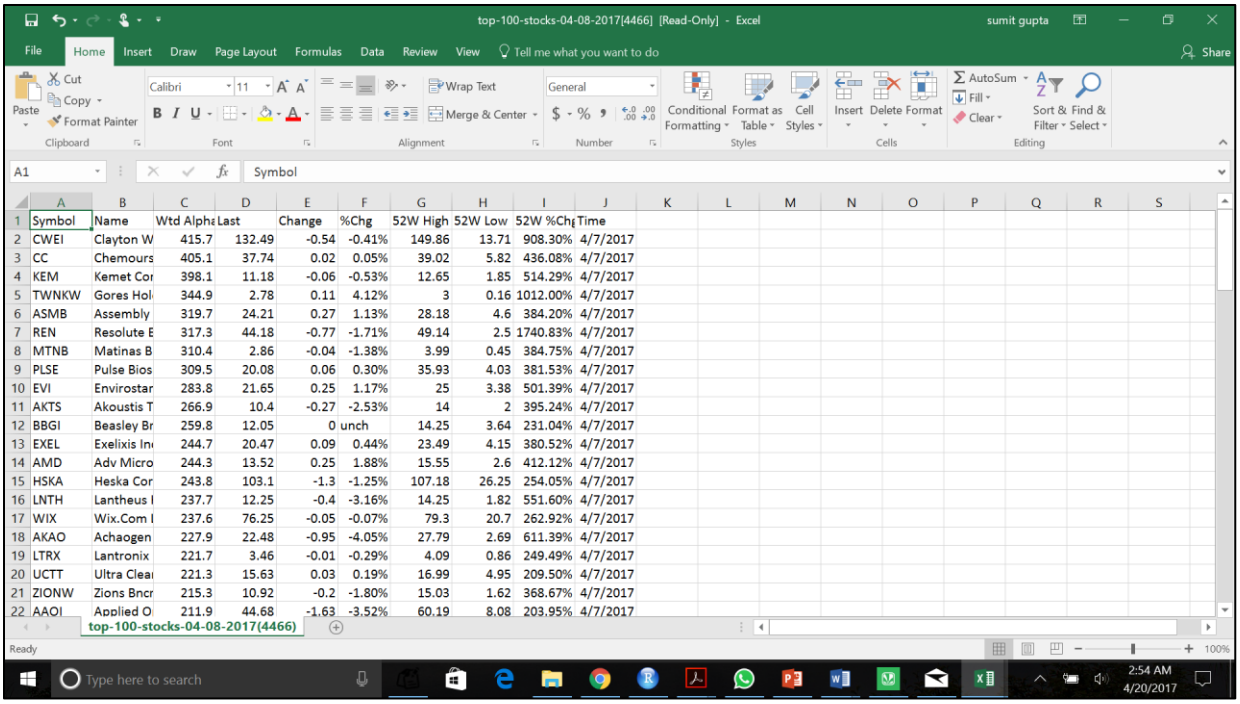
Below is the snapshot for the use of the web scraper for scraping data on the website using Instant data scraper:



Below is the snapshot of the scraped .csv file with top 200 stocks ranked based on price volume leaders:



Below is the snapshot of the scraped data in .csv file of top 100 stocks based on weighted alpha values:



Symbol	Name	Wtd Alpha	Last	Change	%Chg	52W High	52W Low	52W %Chg	Time
CWEI	Clayton W	415.7	132.49	-0.54	-0.41%	149.86	13.71	908.30%	4/7/2017
CC	Chemours	405.1	37.74	0.02	0.05%	39.02	5.82	436.08%	4/7/2017
KEM	Kemet Cor	398.1	11.18	-0.06	-0.53%	12.65	1.85	514.29%	4/7/2017
TWNKW	Gores Hol	344.9	2.78	0.11	4.12%	3	0.16	1012.00%	4/7/2017
ASMB	Assembly	319.7	24.21	0.27	1.13%	28.18	4.6	384.20%	4/7/2017
REN	Resolute E	317.3	44.18	-0.77	-1.71%	49.14	2.5	1740.83%	4/7/2017
MTNB	Matinas B	310.4	2.86	-0.04	-1.38%	3.99	0.45	384.75%	4/7/2017
PLSE	Pulse Bios	309.5	20.08	0.06	0.30%	35.93	4.03	381.53%	4/7/2017
EVI	Envirostar	283.8	21.65	0.25	1.17%	25	3.38	501.39%	4/7/2017
AKTS	Akoustis T	266.9	10.4	-0.27	-2.53%	14	2	395.24%	4/7/2017
BBGI	Beasley Br	259.8	12.05	0	unch	14.25	3.64	231.04%	4/7/2017
EXEL	Exelixis In	244.7	20.47	0.09	0.44%	23.49	4.15	380.52%	4/7/2017
AMD	Adv Micro	244.3	13.52	0.25	1.88%	15.55	2.6	412.12%	4/7/2017
HSKA	Heska Cor	243.8	103.1	-1.3	-1.25%	107.18	26.25	254.05%	4/7/2017
LNTH	Lantheus I	237.7	12.25	-0.4	-3.16%	14.25	1.82	551.60%	4/7/2017
WIX	Wix.Com I	237.6	76.25	-0.05	-0.07%	79.3	20.7	262.92%	4/7/2017
AKAO	Achaogen	227.9	22.48	-0.95	-4.05%	27.79	2.69	611.39%	4/7/2017
LTRX	Lantronix	221.7	3.46	-0.01	-0.29%	4.09	0.86	249.49%	4/7/2017
UCTT	Ultra Clea	221.3	15.63	0.03	0.19%	16.99	4.95	209.50%	4/7/2017
ZIONW	Zions Bncr	215.3	10.92	-0.2	-1.80%	15.03	1.62	368.67%	4/7/2017
AAOI	Applied O	211.9	44.68	-1.63	-3.52%	60.19	8.08	203.95%	4/7/2017

2. **Past 52 weeks lowest and highest price:** We constructed a function which was used to scrape the 52 week highest and lowest stock values of all the companies that have been extracted in last step. The values were stored in a data frame named Company_52_week_price. Attached below is the R-code snapshot of the same. (R-package used: **Rvest**)

```
## Get 52 week highest and lowest values for each
fetch_price<- function(company_tag)
{
  |
  webURL<- gsub(" ","",paste("https://www.barchart.com/stocks/quotes/",company_tag))
  webdata<- read_html(webURL)
  Period_table<- html_table(html_nodes(webdata, "table")[[1]],fill = T )[3,]
  pos<- gregexpr(pattern=' ',Period_table[1,2])
  fifty_two_week_low<- as.numeric(substring(Period_table[1,2],1,pos[[1]][1]-1))
  pos<- gregexpr(pattern=' ',Period_table[1,4])
  fifty_two_week_high<- as.numeric(substring(Period_table[1,4],1,pos[[1]][1]-1))
  df<- c(fifty_two_week_low,fifty_two_week_high)
  return(df)
}

for(i in 1:nrow(Company_fifty_two_week_price))
{
  fifty_two_week_price<- fetch_price(Company_fifty_two_week_price[i,1])
  Company_fifty_two_week_price$Fifty_Two_Week_Low[i]<- fifty_two_week_price[1]
  Company_fifty_two_week_price$Fifty_Two_Week_high[i]<- fifty_two_week_price[2]
  Company_fifty_two_week_price$fifty_Two_Week_Change[i]<-((fifty_two_week_price[2]-fifty_two_week_price[1])/fifty_two_week_price[1])

}
```

3. **Company fundamentals, Growth value, Per share info and ratios:** We developed a different function which was used to scrape data from 300 web pages and extract following information about the company and store them in respective data frames:

Fundamentals	**Ratios**	**Per share info**	**Growth**
Market Capitalization	Price per earnings	Most recent earnings	One year return
Shares Outstanding	Price per earnings Fwd	Next earnings date	Three-year return
Annual Sales	Price per earnings growth	Earnings per share	Five-year return
Annual Net Income	Return on equity	EPS growth vs previous quarter	Five-year revenue growth
Thirty-six-month beta	Return on Assets	EPS growth vs previous year	Five-year earnings growth
Percentage Insider shareholder	Profit Margin	Annual dividend rate	Five-year dividend growth
Percentage Institutional Shareholder	Debt per Equity	Annual dividend yield	
	Price per sales	Dividend payout ratio	

	Price per cash flow		
	Price per book		
	Book value per share		
	Interest Coverage		

Various validation checks were built in the code to make sure that the correct data is scraped and inserted into the expected data frame. Below is the R-code snapshot of the same:

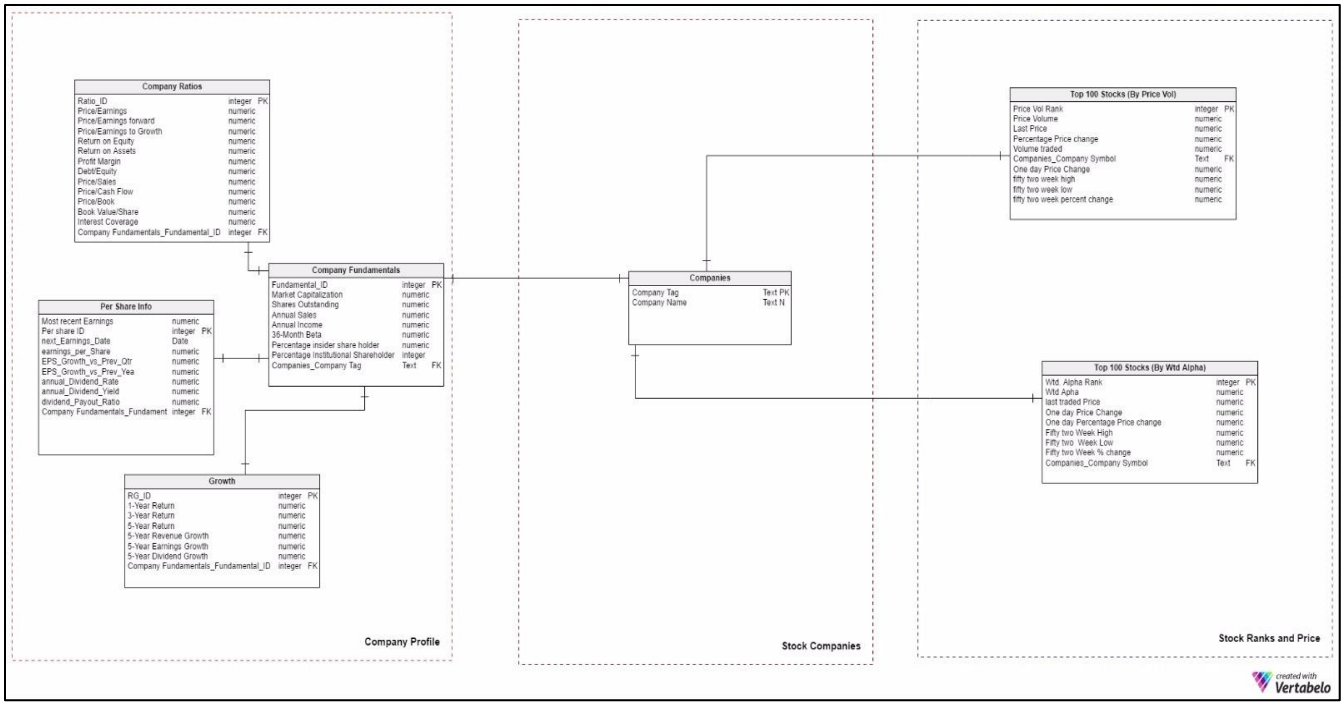
```
## Scapping data for fundamentals and others
for(i in 1:nrow(Company_fifty_two_week_price))
{
  company<-Company_fifty_two_week_price[i,1]
  company<-as.character(company)
  all_tables<- scraperFunc(company[1])
  if(length(all_tables)>0)
  {
    if(nrow(all_tables[[1]][2])==7) ## Checking if the correct table is being scrapped
    {
      Fundamentals[i,]<- cbind(as.character(company[1]),t(all_tables[[1]][2]))
    }
  }
  if(length(all_tables)>1)
  {
    if(nrow(all_tables[[2]][2])==6) ## Checking if the correct table is being scrapped
    {
      growth[i,]<- cbind(company,t(all_tables[[2]][2]))
    }
  }
  if(length(all_tables)>2)
  {
    if(nrow(all_tables[[3]][2])==12) ## Checking if the correct table is being scrapped
    {
      per_Share_Info[i,]<- cbind(company,t(all_tables[[3]][2]))
    }
  }
  if(length(all_tables)>3)
  {
    if(nrow(all_tables[[4]][2])==12) ## Checking if the correct table is being scrapped
    {
      ratios[i,]<- cbind(company,t(all_tables[[4]][2]))
    }
  }
```

Data Cleaning and Storing

The scraped data was then cleaned of any percentage signs, dollar signs etc. and stored in a relational database SQLite. RSQLite package was used for the same.

A relational database was chosen to store the data since the information being scraped was related to the companies and we could establish a good entity relationship between various entities. Further we chose to use SQLite because it is an all-inclusive server-less database system in a single file. So, there is no need of setting a database server and everything is included within the R package itself. It is very helpful in handling big chunks of data or analyzing subsets of data which can be retrieved via a SQLite query.

Before storing the data, we developed the below Entity-Relationship diagram using Vertabelo (an online platform to construct ER diagrams). We executed 7 SQL queries (DDL Statements) after establishing the SQLite server connection in R.



Below are the sample queries used for creating the entities for above E-R diagram in our database named ‘StocksProjectDB’.

```
sql_5<-"CREATE TABLE per_Share_Info(
per_Share_id INTEGER PRIMARY KEY AUTOINCREMENT,
most_Recent_Earnings NUMBER,
next_Earnings_Date DATE,
earnings_per_Share NUMBER,
EPS_Growth_vs_Prev_Qtr NUMBER,
EPS_Growth_vs_Prev_Year NUMBER,
annual_Dividend_Rate NUMBER,
annual_Dividend_Yield NUMBER,
divident_Payoiut_Ratio NUMBER,
fundamentals INTEGER,
FOREIGN KEY (fundamentals) REFERENCES Fundamentals(fundamentals_id))"

sql_6<- "CREATE TABLE growth(
growth_id INTEGER PRIMARY KEY AUTOINCREMENT,
one_year_return NUMBER,
three_year_return NUMBER,
five_year_return NUMBER,
five_year_revenue_growth NUMBER,
five_year_earnings_growth NUMBER,
five_year_dividend_growth NUMBER,
fundamentals INTEGER,
FOREIGN KEY (fundamentals) REFERENCES Fundamentals(fundamentals_id))"

# Create an SQLite database using above queries
con <- dbConnect(SQLite(), dbname = "StocksProjectDB2.sqlite")
dbListTables(con)
dbSendQuery(conn=con,sql_1)
dbSendQuery(conn=con,sql_2)
dbSendQuery(conn=con,sql_3)
dbSendQuery(conn=con,sql_4)
dbSendQuery(conn=con,sql_5)
```

We carried out data cleaning where we removed the unwanted characters from the data record such as $ sign, ‘M’ and ‘K’ characters representing Millions and thousand figures, % sign and some periods found in the data. Some unmapped records were found due to data inconsistency and they were dropped from the database.

Below is the sample R code used for cleaning the data from fundamentals table:

```
#Fundamentals table
Fundamentals<- cbind.data.frame(1:nrow(Fundamentals),Fundamentals)
colnames(Fundamentals)<- c("Fundamentals_id",colnames(Fundamentals[2:ncol(Fundamentals)]))
colnames(Fundamentals)<-c("fundamentals_id","company",
                          "market_Capitalization",
                          "shares_Outstanding",
                          "annual_Sales",
                          "annual_Net_Income",
                          "thirtysix_Month_Beta",
                          "percentage_Insider_Shareholder",
                          "percentage_Institutional_Shareholder")
# Column wise data cleaning
Fundamentals$market_Capitalization<- as.numeric(gsub(",","",as.character(Fundamentals$market_Capitalization)))
Fundamentals$shares_Outstanding<-as.numeric(gsub(",","",as.character(Fundamentals$shares_Outstanding)))

Fundamentals$annual_Sales<-gsub(",","",as.character(Fundamentals$annual_Sales))
Fundamentals$annual_Sales<-gsub(" K","000",as.character(Fundamentals$annual_Sales))
Fundamentals$annual_Sales<-as.numeric(gsub(" M","000000",as.character(Fundamentals$annual_Sales)))

Fundamentals$annual_Net_Income<-gsub(",","",as.character(Fundamentals$annual_Net_Income))
Fundamentals$annual_Net_Income<-gsub(" K","000",as.character(Fundamentals$annual_Net_Income))
Fundamentals$annual_Net_Income<-as.numeric(gsub(" M","000000",as.character(Fundamentals$annual_Net_Income)))

Fundamentals$thirtysix_Month_Beta<-as.numeric(Fundamentals$thirtysix_Month_Beta)
Fundamentals$percentage_Insider_Shareholder<- as.numeric(gsub("%","",as.character(Fundamentals$percentage_Insider_Shareholder)))/100
Fundamentals$percentage_Institutional_Shareholder<- as.numeric(gsub("%","",as.character(Fundamentals$percentage_Institutional_Shareholder)))/100

#Ratios table
tmp <- sapply(1:nrow(Fundamentals), function(fundamentals_id) {
  aa <- Fundamentals[fundamentals_id,]
  idx = which(ratios$company == aa$company)
  ratios[idx, "fundamentals"] <<- fundamentals_id
  return(NULL)
})
```

The R file contains similar codes for all the tables (refer R code file for code details.)

Data Retrieval

As mentioned in previous sections, we used a relational database to store the data using a well-defined entity-relationship. We then made use of SQL queries to retrieve the data. The SQL (SELECT) queries, mainly made use of JOIN and WHERE functions to extract required columns from various data tables defined earlier.

Below are the sample queries used for retrieving data from fundamental and ratios table and were stored in data frames defined locally.

```
##Data Retrival Queries
Companies_price_vol_Fundamental<- dbGetQuery(con, "SELECT
                                                   last_traded_Price,
                                                   (fifty_Two_Week_High+fifty_Two_Week_Low)/2 as Avg_Price,
                                                   market_Capitalization,
                                                   shares_Outstanding,
                                                   annual_Sales,
                                                   annual_Net_Income,
                                                   thirtysix_Month_Beta,
                                                   percentage_Insider_Shareholder,
                                                   percentage_Institutional_Shareholder
                                                   FROM stocks_By_Price_Vol, companies,Fundamentals
                                                   WHERE stocks_By_Price_Vol.company=companies.company_Tag
                                                   AND Fundamentals.company=companies.company_Tag")

Companies_price_vol_Ratios<- data.frame(dbGetQuery(con,"SELECT company_Tag,
                                        last_traded_Price,
                                        (fifty_Two_Week_High+fifty_Two_Week_Low)/2 as Avg_Price,
                                        price_per_Earnings,
                                        price_per_Earnings_Fwd,
                                        price_per_Earnings_Growth,
                                        return_on_Equity,
                                        return_on_Assets,
                                        profit_Margin,
                                        debt_per_Equity,
                                        price_per_sales,
                                        price_per_Cash_Flow,
                                        price_per_Book,
                                        bookValue_per_Share,
                                        interest_Coverage
                                        FROM ((stocks_By_Price_Vol JOIN companies ON stocks_By_Price_Vol.company=companies.company
                                        JOIN Fundamentals ON Fundamentals.company=companies.company_Tag)
                                        JOIN ratios ON ratios.fundamentals= Fundamentals.fundamentals_id"))
```

The R code contains close to 10 queries to retrieve required columns from various data tables which were used for analysis. (Please refer R code for full set of queries)

Data Analysis:

As mentioned in the above sections, the stocks values were collected in two different categories i.e**. Weighted alpha and Price volume**. Using the data, we have tried to find out how the stock prices vary in the above two categories with regards to factors such as Company Fundamentals, Growth parameters, return ratios etc. as mentioned in the E-R diagram above. To analyze this variation, we used Fifty-two weeks highest and lowest price values of various stocks to calculate the average stock value and further we employed linear multiple regression techniques.

Multiple Linear Regression attempts to model the relationship between two or more independent variables and a dependent variable by fitting a linear equation to observe variation in data. Here,

every value of the independent variable x is associated with a value of the dependent variable y. A general multiple linear regression can be represented mathematically as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Several models were generated manually using different parameter as independent variables and the one with the best R-Squared value was selected as the best model. Further to identify the most influential factors which play a vital role in variation of stock values, we employed forward section method using Forward AIC in R.

Analysis of Price Volume Category

For the stocks from price volume category, we generated four different models. The first three model tries to explain variation in average stock price based on columns (as independent variables) of fundamental, ratios and growth tables whereas the fourth model used columns from all the three tables. The highest R- squared value from the above four models was obtained for the fourth model (R squared ~ 52%). Below is the summary of 4th model (please refer the R-code for all the summary of other models).

```
Call:
lm(formula = Avg_Price ~ ., data = Data_Price_vol_ALL)

Residuals:
    Min      1Q  Median      3Q     Max
-235.42  -39.63  -12.88   26.24  822.74

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.184e+02  6.448e+01  -1.837 0.068005 .
one_year_return                 -2.192e+01  3.302e+01  -0.664 0.507771
three_year_return               -5.272e+00  2.678e+01  -0.197 0.844152
five_year_return                 1.988e+00  9.034e+00   0.220 0.826097
five_year_revenue_growth         9.386e-02  1.294e-01   0.725 0.469203
five_year_earnings_growth       -1.727e+00  4.346e+00  -0.397 0.691587
price_per_Earnings              -1.386e-02  4.213e-01  -0.033 0.973789
price_per_Earnings_Fwd          -2.303e-03  7.837e-02  -0.029 0.976592
return_on_Equity                -6.360e-01  2.801e-01  -2.270 0.024463 *
return_on_Assets                 8.724e+00  1.937e+00   4.504 1.24e-05 ***
profit_Margin                   -1.167e+00  6.841e-01  -1.706 0.089789 .
debt_per_Equity                 -6.395e-01  3.010e+00  -0.212 0.832033
price_per_sales                  1.152e+00  3.726e+00   0.309 0.757542
price_per_Cash_Flow              5.727e-02  1.651e-01   0.347 0.729049
price_per_Book                   3.169e+00  1.297e+00   2.444 0.015558 *
bookValue_per_Share              2.192e+00  2.688e-01   8.154 7.79e-14 ***
interest_Coverage               -3.571e-01  3.023e-01  -1.181 0.239103
market_Capitalization            4.493e-07  1.201e-07   3.740 0.000253 ***
shares_Outstanding              -2.325e-05  7.940e-06  -2.928 0.003885 **
annual_Sales                    -4.446e-09  7.224e-09  -0.615 0.539070
annual_Net_Income                1.492e-08  7.819e-09   1.909 0.058023 .
thirtysix_Month_Beta             7.033e+00  2.008e+01   0.350 0.726639
percentage_Insider_Shareholder   8.686e+01  1.113e+02   0.781 0.436145
percentage_Institutional_Shareholder  1.125e+02  7.074e+01   1.590 0.113614
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 115.8 on 168 degrees of freedom
Multiple R-squared:  0.5239,    Adjusted R-squared:  0.4587
F-statistic: 8.038 on 23 and 168 DF,  p-value: < 2.2e-16
```

To eliminate the less important variables we refined the full model by using forward AIC method. The Forward AIC model gave the relationship between average price and the following variables governing most of the variations:

Avg_Price ~ bookValue_per_Share

+ return_on_Assets

+ shares_Outstanding

+ market_Capitalization

+ interest_Coverage

+ profit_Margin

+ price_per_sales

+ thirtysix_Month_Beta

+ five_year_earnings_growth

+ percentage_Insider_Shareholder

+ percentage_Institutional_Shareholder

Analysis of Weighted Alpha Category

For the stocks from weighted alpha category, we again generated four different models. The first three model tries to explain variation in average stock price based on columns (as independent variables) of fundamental, ratios and growth tables whereas the fourth model used columns from all the three tables. The highest R- squared value from the above four models was again obtained for the fourth model. (R squared ~ 56.98%).

```
Call:
lm(formula = Avg_Price ~ ., data = Data_wtd_alpha_ALL)

Residuals:
    Min      1Q  Median      3Q     Max
-28.359  -7.674  -0.784   4.408  89.526

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.354e+00  8.346e+00   0.402 0.689472
one_year_return                  1.069e+00  1.289e+00   0.829 0.410783
three_year_return                1.634e+00  2.395e+00   0.682 0.498040
five_year_return                -3.459e-01  9.711e-01  -0.356 0.723162
five_year_revenue_growth        -5.111e-01  8.406e-01  -0.608 0.545857
five_year_earnings_growth        9.616e-02  1.243e-01   0.774 0.442514
price_per_Earnings              -1.514e+01  7.322e+00  -2.068 0.043603 *
price_per_Earnings_Fwd          -1.058e-01  4.512e-02  -2.345 0.022885 *
return_on_Equity                -1.573e-02  1.668e-02  -0.943 0.349894
return_on_Assets                 9.945e-02  1.016e-01   0.979 0.332308
profit_Margin                   -4.715e-04  5.510e-04  -0.856 0.396137
debt_per_Equity                  2.000e-03  5.942e-01   0.003 0.997328
price_per_sales                 -1.994e-03  2.833e-03  -0.704 0.484719
price_per_Cash_Flow              3.426e-01  1.363e-01   2.514 0.015083 *
price_per_Book                   6.237e-02  1.232e-01   0.506 0.614707
bookValue_per_Share             -5.839e-02  1.598e-01  -0.365 0.716368
interest_Coverage                2.422e-01  1.353e-01   1.790 0.079318 .
market_Capitalization            4.932e-06  1.351e-06   3.650 0.000609 ***
shares_Outstanding              -1.585e-04  4.333e-05  -3.658 0.000594 ***
annual_Sales                     5.068e-09  4.408e-09   1.150 0.255526
annual_Net_Income               -8.419e-08  3.468e-08  -2.428 0.018686 *
thirtysix_Month_Beta            -2.236e+00  2.664e+00  -0.840 0.404981
percentage_Insider_Shareholder   1.129e+00  1.487e+01   0.076 0.939803
percentage_Institutional_Shareholder  2.263e+01  1.042e+01   2.172 0.034427 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.54 on 52 degrees of freedom
Multiple R-squared:  0.5698,     Adjusted R-squared:  0.3795
F-statistic: 2.995 on 23 and 52 DF,  p-value: 0.0005308
```

(please refer the R-code for all the summary of other models).

Further, to eliminate the less important variables we again refined the values by using forward AIC method. The forward AIC model gave the relationship between average price and the following variables governing most of the variations:

Avg_Price ~ percentage_Institutional_Shareholder

+ annual_Net_Income

+ price_per_Cash_Flow

+ three_year_return

+ profit_Margin

+ price_per_Earnings_Fwd

For the above two analysis, we concluded:

1. The Percentage Institutional Shareholder and Profit Margin are the common factors across both the categories and one should consider these parameters while investing in the stock market. Profit margin had negative slope and Percentage institutional shareholders has positive slope in the linear relationship model.

2. The variation in stock value does not on depend only on fundamentals, growth or the ratios table but rather on the combination of all three.

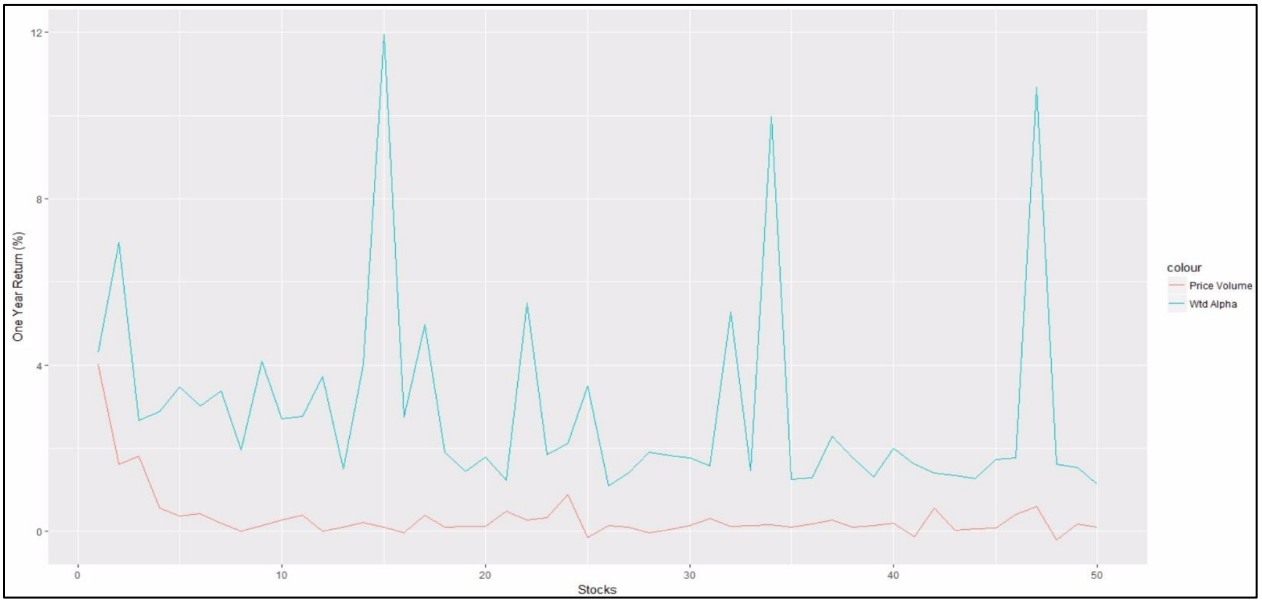
**Note: The models defined above are based on the values obtained on 20th April 2017 and provide only a fair idea of stock value dependency. One might find addition of new variables or deletion of the existing variable depending on the change in the data values.

Weighted alpha vs Price volume

To compare the two categories and decide regarding which category of stocks (weighted alpha or price volume) to choose from, we tried to compare the first 50 stocks from both the categories. We have used ggplot2 to carry out following comparisons:

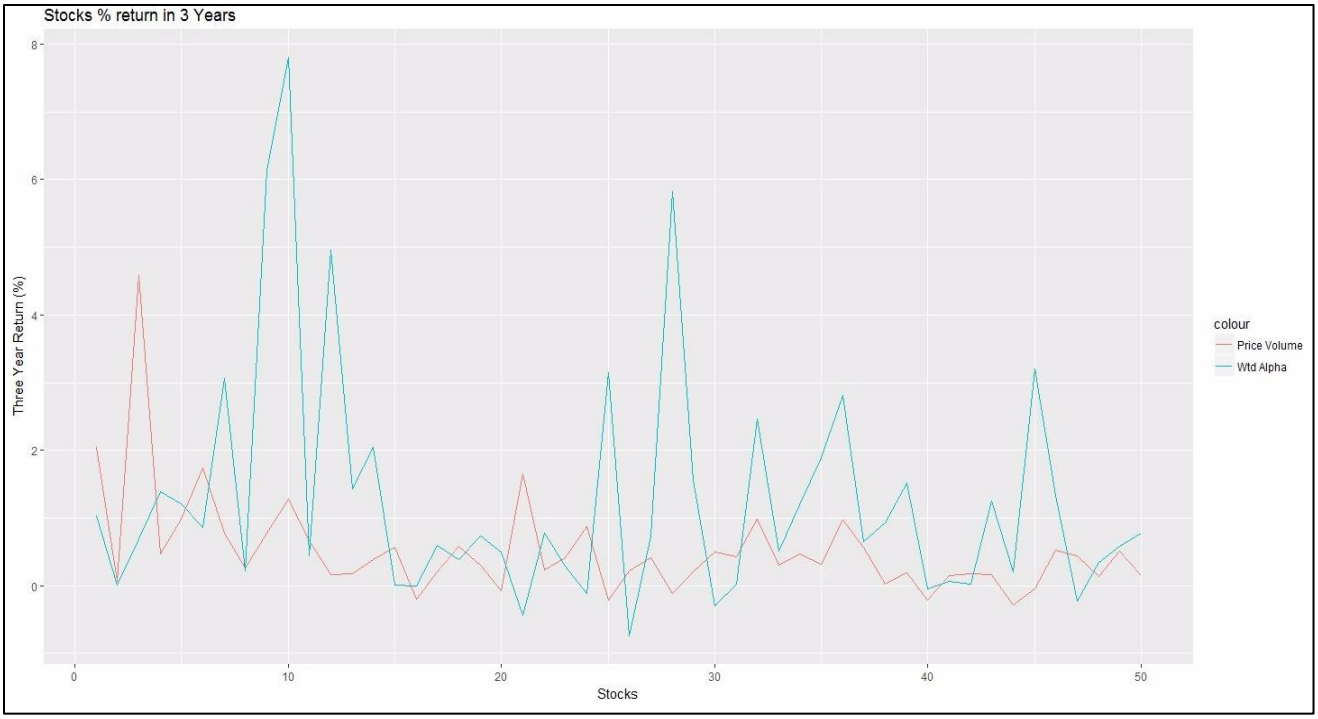
1. One year return on investments: The below curve provides comparison of percentage return after one year of investment from first fifty stock from both the categories. It has been observed that stocks in weighted alpha category seem to have higher return value compared to those in the price volume. However, the variation in return value is much higher for those stocks in weighted alpha category and hence one needs to be very careful while investing based on this analysis.

The blue line represents weighted alpha whereas the red line represents price volume.

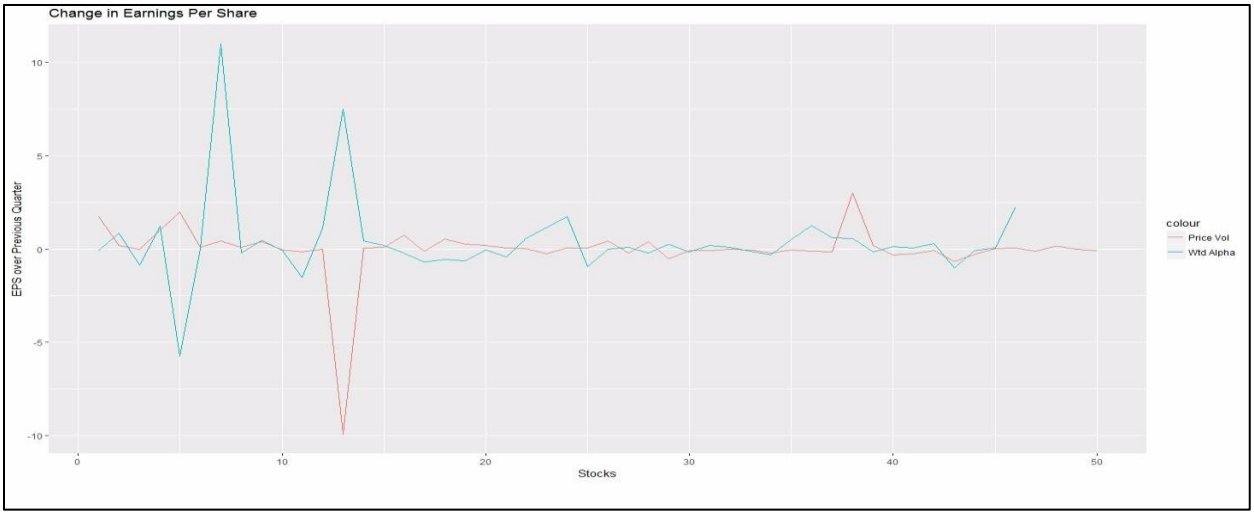


2. Three-year return on investment: To further find out the better categories to invest, we observed the return value after three years of investment. The below curve provides the comparison on percent return after three years. It was observed that after 3 years, the curves of price volume and weighted alpha interact with each other. However, huge spikes can be observed in the stocks for weighted alpha category. Overall stocks in weighted alpha category still provide higher return compared to the price volume category.

Below is the snapshot for this analysis with blue line for weighted alpha and red line for price volume.



3.Growth in earnings per share (EPS): To further compare the two categories, we plotted the growth or change in earning per share over previous quarter. Earnings per share serves as an indicator for a company's profitability. Below is the curve with blue line representing stocks from weighted alpha and red for price volume. We can observe that the change in earnings per share is almost similar across the two categories with a few outliers highlighted by the spikes. Thus, we found that there is not an appreciable difference in EPS values across the two categories.



Through this analysis, we can say that top 50 stocks of weighted alpha category have higher probability of providing better returns on investments compared to the top 50 stocks of price volume category. Also, the change in earnings per stock remains same across both the categories.

Learning Outcomes

The major aim of the project to collect, store and retrieve data was successfully achieved. The project uses various web scraping tools and techniques to scrape data from more than 300 webpages. Further, since the data collected was not in the right format and consisted of several null and redundant values, we performed intensive data cleaning and dropped the unwanted records.

SQLite, a relational database system, was used to store the values as per the pre-defined entity-relationship diagram. Further, both DDL and DML queries were employed to store and retrieve the data from the database. We made extensive use of SQL JOINs to retrieve the columns required for the analysis.

For the analysis, we successfully implemented linear regression to identify the best possible model which can explain the variation in average stock prices across the two categories of data. We identified the common parameters across the two categories of data that play a vital role in variation of stock prices. Also, to compare the two categories of data we made use of data visualization package “ggplot2” and plotted curves to identify the better category of stocks for one to make investments.

Future Scope

As per the analysis, we could explain only 50-60% of variation in stock prices across the two categories. Since, stock markets are dynamic and are subject to numerous factors, it is very difficult to identify variation in stock prices through linear regression models. One can look out for Non-linear regression models, ARIMA models, Time Series Analysis etc. to better predict the stock prices.

References:

- www.barchart.com
- www.wikipedia.com
- A book on Collecting, Storing and Retrieving Data by Yatish Jain, Martin Schedlbauer and Kathleen Durant
- www.vertabelo.com