Project Final Report
Berkshire Hathaway Inc.

FA541-A: Applied Statistics with
Applications to Finance,

Fall 2021 Semester
December 16, 2021

Abstract

The purpose of this paper is to determine the performance of the Berkshire Hathaway stock (BRK-A) and to see how it perform with the comparison of S&P 500 index ETF (SPY). This paper uses data and analysis methods to answer the question: What influenced Berkshire Hathaway stock prices and What is the best fit model using the Multiple Regression Analysis? The data was used to compare the BRK-A performance to the S&P 500 using methods such as return comparison, looking at different trading strategies, and clustering.

Company Background

Berkshire Hathaway Inc. is an American multinational conglomerate holding company own by Warren Buffett and it's headquartered is in Omaha, Nebraska, United States. Berkshire Hathaway traces its roots to a textile manufacturing company established by Oliver Chace in 1839 as the Valley Falls Company in Valley Falls, Rhode Island. In 1965, Warren Buffet took full control of the company. As of 2021, Berkshire Hathaway Inc lands on No. 6 in the ranking of the Fortune 500 companies. As of 2020, The Revenue of the company is 245.5bn USD.

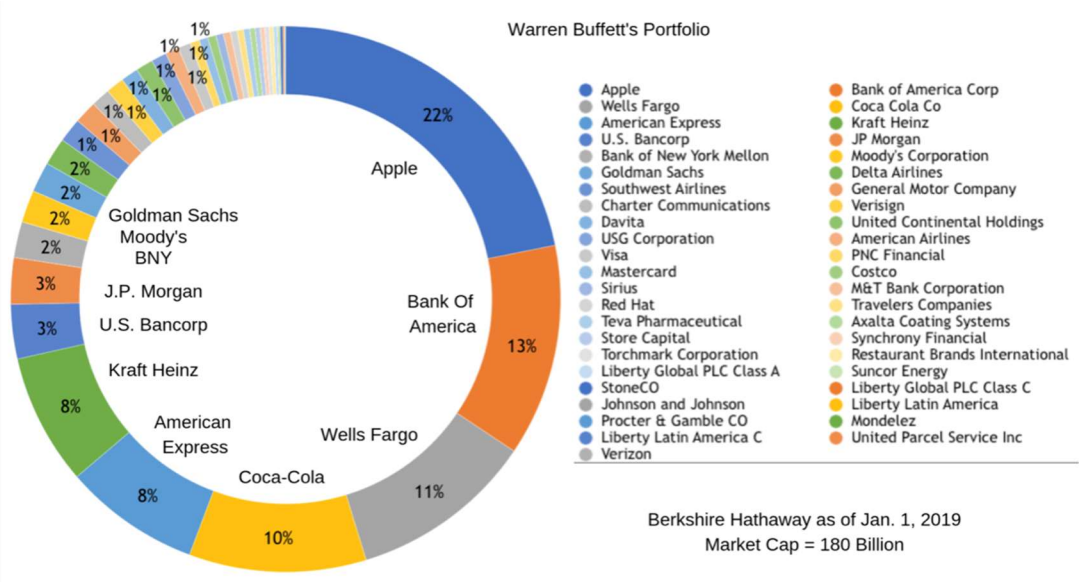

Figure 1: BRK-A Holdings

Above figure shows the holdings of the Berkshire Hathaway. Company has some of the biggest equites in their portfolio which are APPL, BAC, KO, AXP, AMZN, etc. The reason behind company's success is their investment strategies. Because of, company stock grown form $19 to $418,042 in nearly 50 years.

Berkshire Hathaway also have parent companies in their portfolio like, Berkshire Hathaway HomeService of America, Berkshire Hathaway Direct Insurance Company, Berkshire Hathaway Automotive, Berkshire Hathaway Energy Company.

To understand company's performance for long period of time, there are the way to understand it. How company did in financial crash time, if company's stock didn't tank hugely then company's performance is good. Warren buffet proves that 3 times in 2001,2008 and 2020 crash. When all the companies are tanking that time Berkshire Hathaway didn't take hugely, they were mange to stay with the market. In 2008 crash, they helps the other companies to survive too. they funded big banks like Bank of America, J. P. Morgan chase and Morgan Stanley.

Data and Methods

In order to achieve the goals outlined for the project, I put together a dataset using excel and up-to-date stock information from Yahoo Finance. Using a time period of 20 years beginning in 2001 to 2021, created a data frame consisting of stock price data and percent change. Also used the parameters like MarketCap, P/E Ratio, P/B Ratio. While if I only looked at performance in the past year, the selections may look different, since years of performance are being taken into account, stock can be analyze more deeply.

The datasets created can be seen in the figures below. The data was taken from Yahoo Finance as the monthly close prices for the BRK-A stock, as well as a dataset containing the close prices of the SPY which were used as comparisons. These were used to perform the majority of the data analysis within this project.

Date	Open	High	Low	Close	Adj Close	Volume
1/2/01	71500	74600	71500	72400	72400	1230
1/3/01	72400	73000	70000	70000	70000	680
1/4/01	69900	70600	68300	69000	69000	650
1/5/01	68600	68600	67500	67800	67800	270
1/8/01	67600	68600	66200	67700	67700	550
1/9/01	67900	69300	66500	66800	66800	390
1/10/01	66800	67300	64700	67200	67200	770
1/11/01	67500	68000	64900	65700	65700	740
1/12/01	65700	66500	64800	66200	66200	430
1/16/01	66000	68000	66000	67800	67800	290

Figure 2: Dataset for the BRK-A

Date	MarketCap	PeRatio	PbRatio
3/31/00	86,519,270,500	47.667	1.545
6/30/00	81,639,548,520	43.282	1.421
9/30/00	97,800,938,560	51.81	1.706
12/31/00	108,257,053,890	47.651	1.825
3/31/01	99,890,544,188	29.954	1.619
6/30/01	105,897,355,900	33.837	1.815
9/30/01	106,889,207,780	32.756	1.82
12/31/01	115,563,956,550	145.106	1.994
3/31/02	109,018,978,812	136.468	1.881

Figure 3: Dataset with additional parameters of BRK-A

Berkshire Hathaway has a lot of equities in its portfolio. To understand how these equities could influence the BRK-A stock price, also used the additional dataset for the 5 main holdings of Berkshire Hathaway which are, APPL, KO, BAC, AXP, and MCO. To compare BRK-A with the

S&P 500 index, I used the SPY ETF dataset. The dataset of these 5 companies and SPY ETF are shown in figure below.

Date	Open	High	Low	Close	Adj Close	Volume
1/2/01	132	132.15625	127.5625	128.8125	86.990189	8737500
1/3/01	128.3125	136	127.65625	135	91.16877	19431600
1/4/01	134.9375	135.46875	133	133.546875	90.187454	9219000
1/5/01	133.46875	133.625	129.1875	129.1875	87.243454	12911400
1/8/01	129.875	130.1875	127.6875	130.1875	87.918762	6625300
1/9/01	131.046875	131.5	129.421875	129.84375	87.68663	5702400
1/10/01	129	132.125	128.8125	132.125	89.227219	8746100
1/11/01	131.09375	133.484375	131.09375	132.25	89.311638	7245100
1/12/01	132.6875	133.71875	131.28125	132	89.142815	7244000
1/16/01	132	133.1875	131.515625	132.84375	89.712601	8542200

Figure 4: SPY ETF dataset

Date	BRK	Apple	BOA	Amex	KO	Moodys
1/2/15	223600	27.3325	17.9	93.019997	42.139999	95.889999
1/5/15	220980	26.5625	17.379999	90.559998	42.139999	94.010002
1/6/15	220450	26.565001	16.860001	88.629997	42.459999	93.019997
1/7/15	223480	26.9375	16.940001	90.300003	42.990002	94.440002
1/8/15	226680	27.9725	17.290001	91.580002	43.509998	95.360001
1/9/15	224675	28.002501	16.98	90.419998	43.029999	93.82
1/12/15	222424	27.3125	16.68	89.5	42.639999	93.010002
1/13/15	223000	27.555	16.450001	89.230003	42.630001	94.489998
1/14/15	221878	27.450001	16.040001	87.07	42.560001	93.739998
1/15/15	221510	26.705	15.2	85.879997	42.380001	92.669998
1/16/15	223615	26.497499	15.38	86.040001	42.529999	94.129997

Figure 5: BRK-A main holdings dataset

Data cleaning methods were applied to ensure that their datasets were in the correct, most efficient format for usage. This included ensuring the data was of the correct class, headings were in a format that was easy to reference throughout the code, and the dates and timelines were correct

and consistent amongst the datasets. From here, used preliminary data analysis techniques such as displaying summary statistics and plotting the changes in close price of the securities in their portfolio over the 20 year timeline. These methods are helpful to get a quick look at the behavior of the selected stocks over time and verify that their performance grew over the time period of interest. Additional analysis was done using investment and regression techniques. Regression analysis was used to plot log returns and compare the BRK-A to the SPY. Finally, used clustering methods to identify the groups of similar objects in the datasets to further compare the selected stock (BRK-A) against the popular indices. Each of these mentioned methods was important in obtaining results which helped to come to their conclusions within this report.

Results

U used the methods described in the previous section to carry out an analysis on the datasets and to answer their research question. In this section, the results obtained from the analysis will be presented and discussed.

The first step taken to analyze the data after loading the datasets in R was plotting the relationship between close price and the date over the last 10 years to get a better idea of BRK-A stock performance. Also used the stock's P/E ratio data over the last 20 years to analyze more about stock nature. The graph of the close price to the date and P/E ratio to date are shown in below figure.

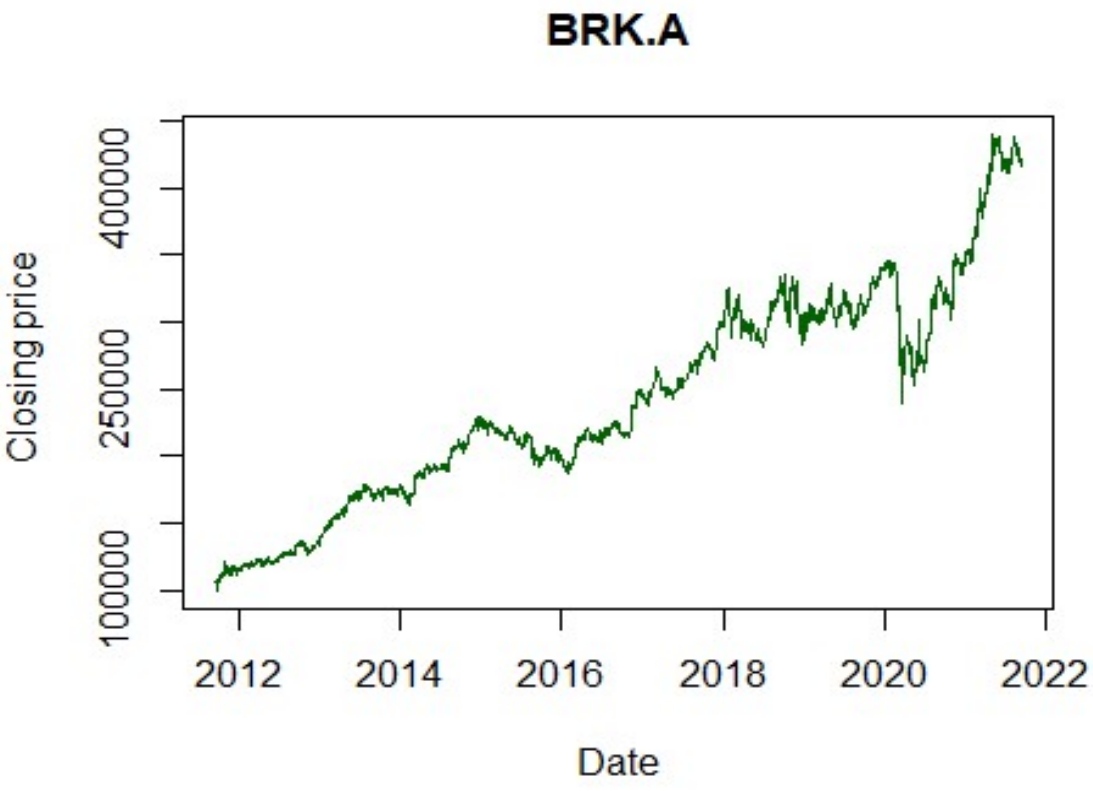

Figure 6: Performance of BRK.A over the 10 year period

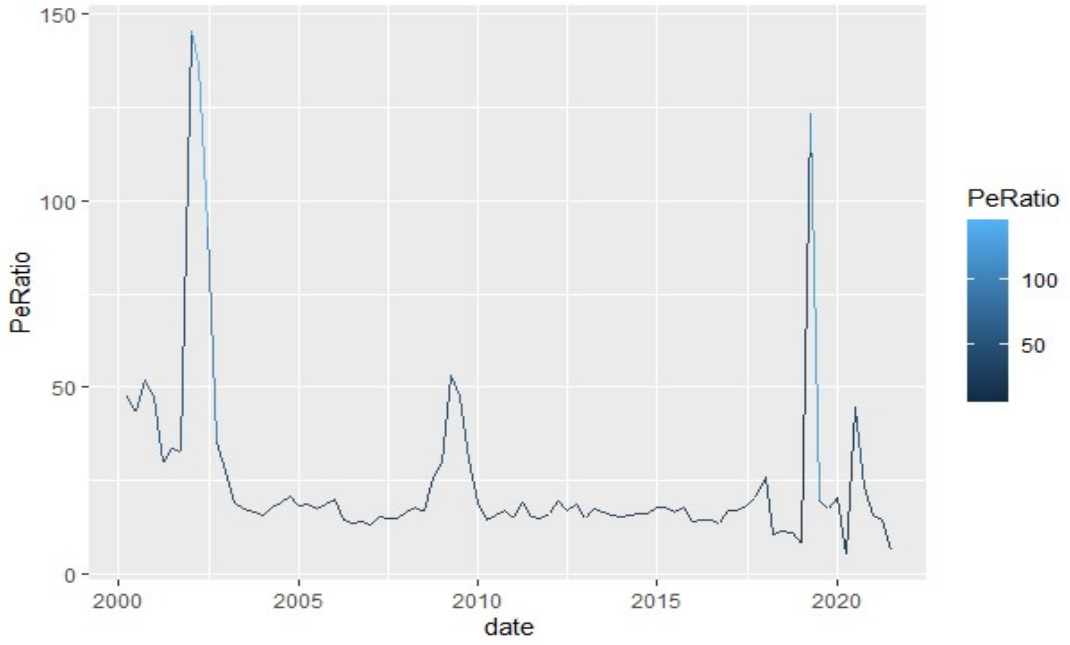

Figure 7: BRK-A P/E ratio graph over the last 20 years

From the above graphs, we can see that Berkshire Hathaway have grown in the last decade. The fall can be seen between the years 2020-2021 which is synonymous with the beginning of the COVID-19 pandemic.

Next, I looked at the summary statistics of the Berkshire Hathaway datasets to show a direct comparison of the distributions of the data. The summary statistics of the stock prices are shown below.

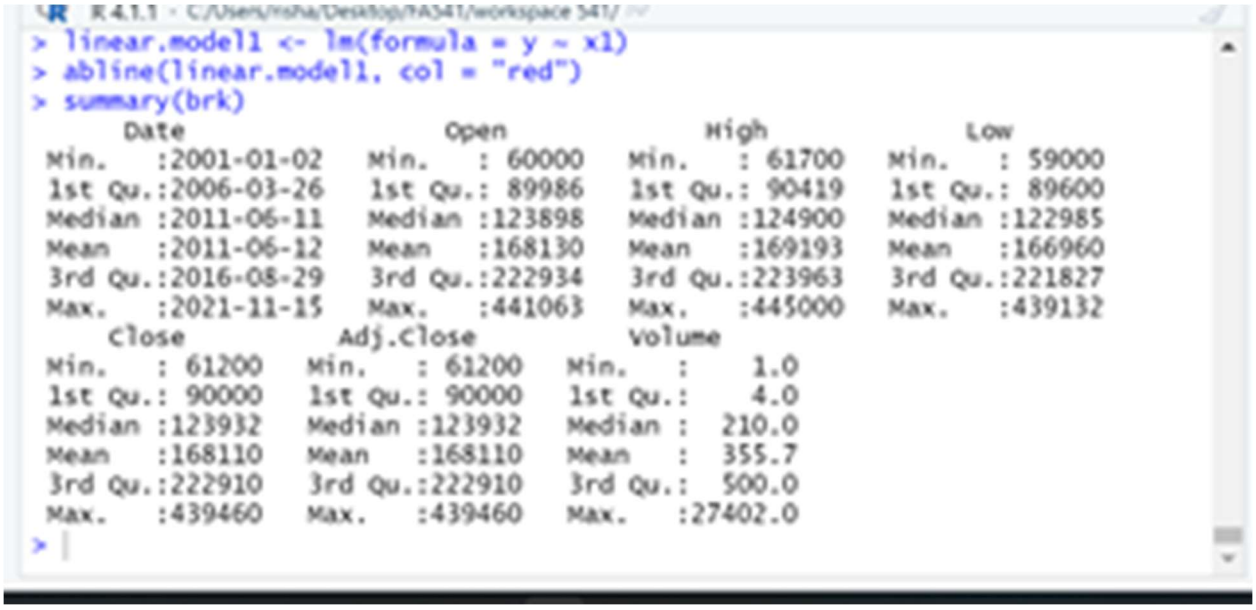

Figure 8: Summary of stock price

Used the linear regression to analyze more about the stock nature. I want to analyze that, if Berkshire Hathaway's main holdings stock prices changed then how it will affect the BRK-A stock price. Used the scatter plot for the analysis and results are shown below.

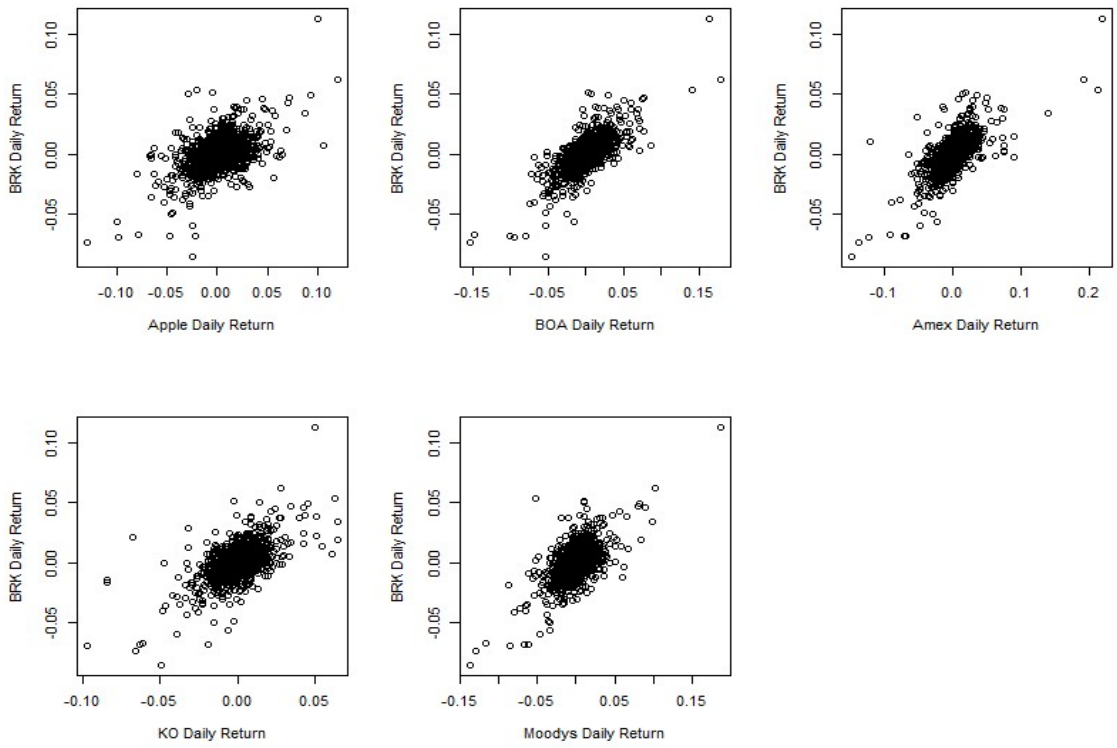

Figure 9: Scatter plot for Berkshire Hathaway main holding to the BRK-A stock

The above scatterplots are daily returns of each stock plotted against daily returns of Berkshire Hathaway. The list of stocks include Apple, Bank of America, American Express, Coca-Cola and Moody's Corporation. The summary of the linear regression are shown below.

```
> summary(multiple.model)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)

Residuals:
      Min        1Q    Median        3Q       Max
-0.038267 -0.003972 -0.000169  0.003386  0.047938

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.084e-05  1.703e-04  -0.122    0.903
x1           6.570e-02  1.138e-02   5.773 9.22e-09 ***
x2           2.210e-01  1.274e-02  17.353  < 2e-16 ***
x3           1.085e-01  1.338e-02   8.112 9.37e-16 ***
x4           2.025e-01  1.767e-02  11.461  < 2e-16 ***
x5           1.182e-01  1.395e-02   8.478  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007065 on 1724 degrees of freedom
Multiple R-squared:  0.663,     Adjusted R-squared:  0.6621
F-statistic: 678.5 on 5 and 1724 DF,  p-value: < 2.2e-16
```

Figure 10: Linear Regression summary for BRK-A holdings

In the above figure X1: Apple Daily Return, X2: Bank Of America Daily Return, X3: Amex Daily Return, X4: Coca Cola Daily Return, X5: Moody's Daily Return.

Performed Multiple linear regression, where the Adjusted R-squared for the multiple regression is significantly higher than all the single linear regression models. Higher the better and should be closer to 1.

Like linear regression, all the p-values are significant here as well. We can consider a model to be statistically significant only when the p-Values are less that the pre-determined statistical significance level, which is ideally 0.05. Also performed ANOVA test for the more analysis.

```
> anova(multiple.model, test="Chisq")
Analysis of Variance Table

Response: y
           Df   Sum Sq  Mean Sq  F value    Pr(>F)
x1          1 0.063134 0.063134 1264.899 < 2.2e-16 ***
x2          1 0.083096 0.083096 1664.832 < 2.2e-16 ***
x3          1 0.009880 0.009880  197.939 < 2.2e-16 ***
x4          1 0.009619 0.009619  192.715 < 2.2e-16 ***
x5          1 0.003588 0.003588   71.879 < 2.2e-16 ***
Residuals 1724 0.086049 0.000050
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: ANOVA test summary

Form the above figure we can say that the larger the F value, the more likely it is that the variation caused by the independent variable is real and not due to chance. Further analysis, plotted the correlation of each of the individual stock which are the main holding for the Berkshire Hathaway. The outputted correlation graph between all stock prices in the data set are shown below.

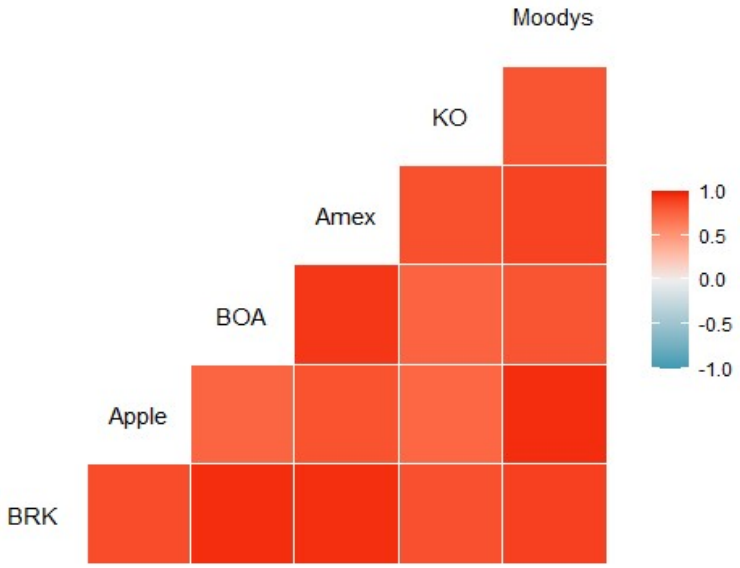

Figure 12: Correlation between 6 stocks prices

From the above correlation plot we can say that all stocks are highly correlated. Because of that, we can say that if stocks price change which are in the BRK-A portfolio then it would affect the BRK-A stock price. For the further analysis I used the BRK-A volume data to see how it affect the stock prices.

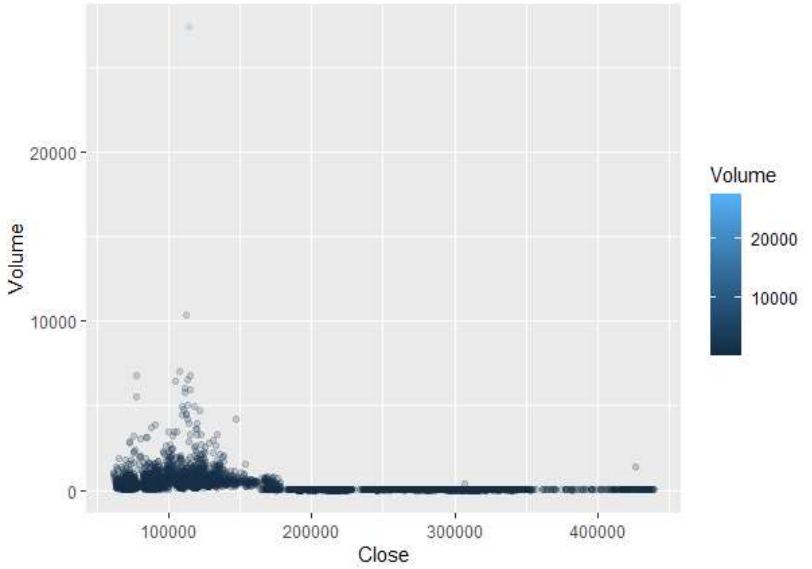

Figure 13: BRK-A Volume analysis over last 20 years

Above graph shows that, BRK-A volume used to change early in 2000's but after that it remain constant when stock price cross the $200,000 mark. In the volume analysis I used the 2008 financial crash period to understand how much volume had change during the crash time.

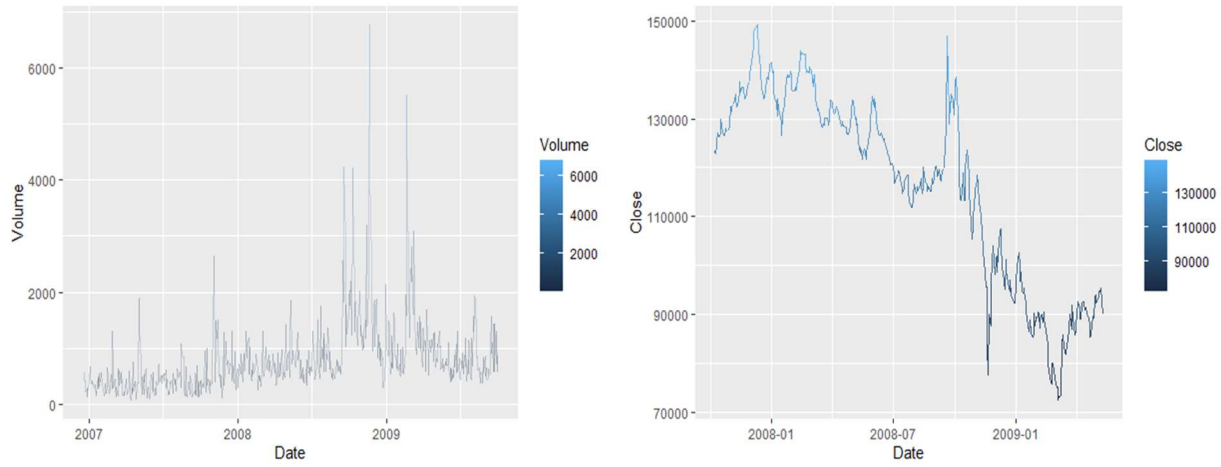

Figure 14: BRK-A Volume analysis during 2008 financial crash

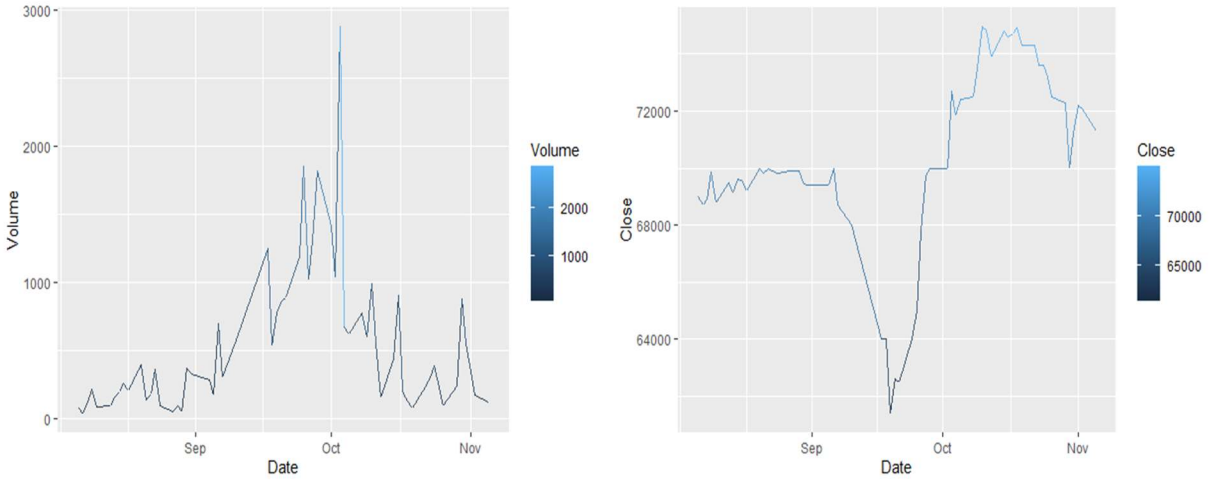

Figure 15: BRK-A Volume analysis during 2001 financial crash

Form the above figures we can say that there were lots of changes in volume during both crash time. This is natural, it happened with all the companies but not like this. Other companies changes are more high than the Berkshire Hathaway. BRK-A able to remain in good position during the crash of 2001 and 2008. Also used the MACD analysis which shown below.

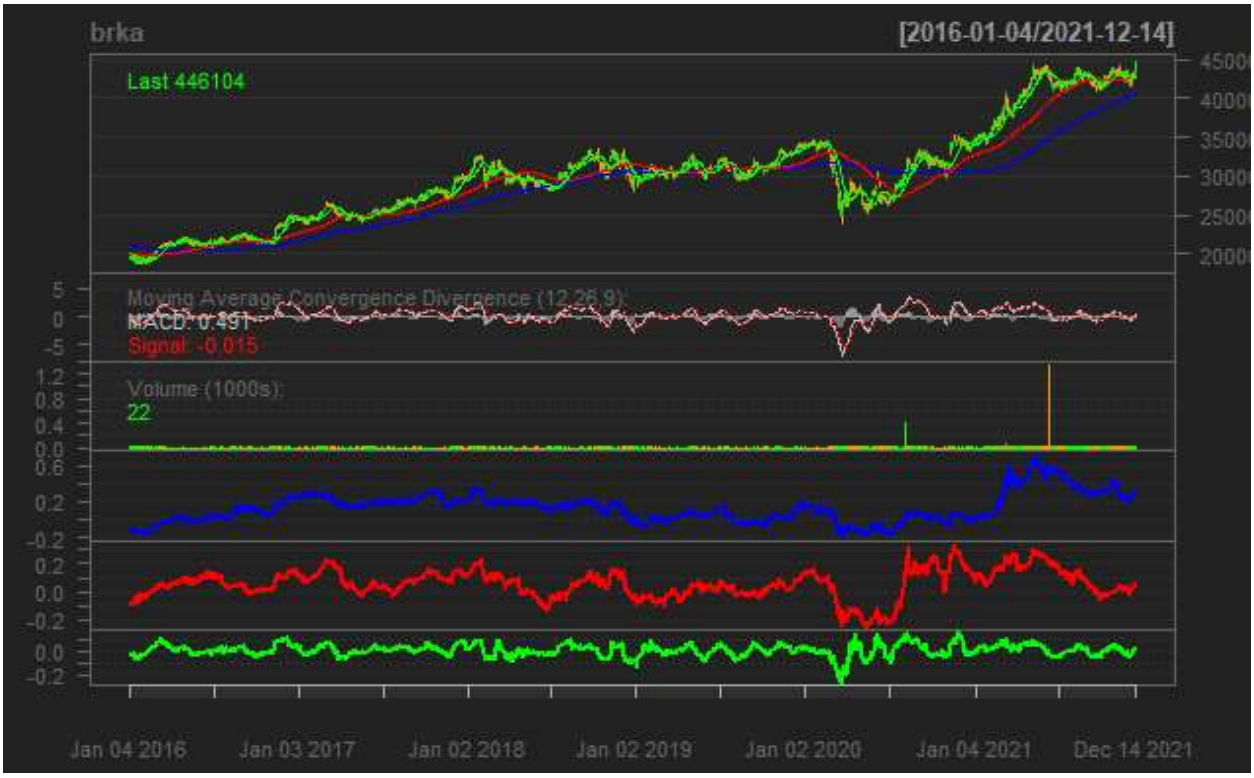

Figure 16: MACD graph for BRK-A

MACD, short for moving average convergence/divergence, is a trading indicator used in technical analysis of stock prices, created by Gerald Appel in the late 1970s. It is designed to reveal changes in the strength, direction, momentum, and duration of a trend in a stock's price. Used this trading tool to see the changes in BRK-A stock price over last 4 years period. Other trading tool used is Bollinger bands. The results are

Figure 17: Bollinger Bands for BRK-A

Bollinger Bands are a trading tool used to determine entry and exit points for a trade. The bands are often used to determine overbought and oversold conditions. Using only the bands to trade is a risky strategy since the indicator focuses on price and volatility, while ignoring a lot of other relevant information. This band represents 2020 crash.

To better understand the nature of the stock and how well it is doing, I compared the stock to the benchmark funds like SPY. The graph shown in Figures 18 below display the close price of the BRK-A stock against the SPY close price over last 20 years period.

Figure 18: Close price of BRK-A stock against SPY close price

Looking at the distributions of the close prices of BRK-A stock, it can be seen that there is a strong correlation between BRK-A stock closing prices and the SPY closing prices. As aforementioned, this result is expected due to the nature of the closing prices over time.

As of today, we can easily say that if we would've invested in the BRK-A stock then money would 300-400% by now. From this statement we can say that the time is most important part in the investing. BRK-A did this job very well, that bought the right stock at the right time. To understand this concept, I went back in time when BRK-A bought the stocks and get the data of that stocks. I selected the 4 parameters which stock analysis is highly depended on and which are P/E ratio, P/B ratio, Price/Sales and Ent. Value/Revenue.

Stock	P/E	Price/Sales	P/B Ratio	Ent. Value/Rev.
APPL	12.66	2.86	4.77	2.91
KO	17.06	4.37	4.86	4.64
BAC	5.71	0.66	0.28	0.45
AXP	6.51	1.82	4.4	2.54
KHC	20.37	2.85	1.5	16.48
MCO	14.58	3.55	1.32	14.53
VZ	15.2	1.07	3.73	1.49
GM	7.41	0.32	1.52	0.25
AMZN	58.14	3.5	16.28	3.34
V	14.93	9.53	2.16	5.81

Figure 19: Companies' Specifications when BRK-A invested

From the above data we can understand that, BRK-A bought this stocks because of the company's outlook, performance from over the period and future plans.

Used Clustering analysis as the last analysis for this paper. The purpose of clustering analysis is to identify patterns in the data and create groups according to those patterns. In this step of clustering analysis we should be able to check what features usually appear together and see what characterizes a group. After done with first two steps the next part is Clustering validation which can be done by using silhouette width. The Last part is to interpret the results and show them in a visualization.

The clustering method symmary are shown in below figure. Used the 4 cluster to understand the performance of the stock.

```
  cluster size ave.sil.width
1       1    3          0.66
2       2    2          0.54
3       3    1          0.00
4       4    4          0.39
> clustering
K-means clustering with 4 clusters of sizes 3, 2, 1, 4

Cluster means:
        P.E Price.Sales        PB Ent..Value.Rev. cluster
1  6.543333   0.9333333  2.066667          1.0800       1
2 17.475000   3.2000000  1.410000         15.5050       2
3 58.140000   3.5000000 16.280000          3.3400       3
4 14.962500   4.4575000  3.880000          3.7125       4

Clustering vector:
12 13 14 15 16 17 18 19 20 21
 4  4  1  1  2  2  4  1  3  4

Within cluster sum of squares by cluster:
[1] 14.83660 18.92450  0.00000 65.10022
 (between_SS / total_SS =  96.2 %)
```

Figure 20: Clustering Summary

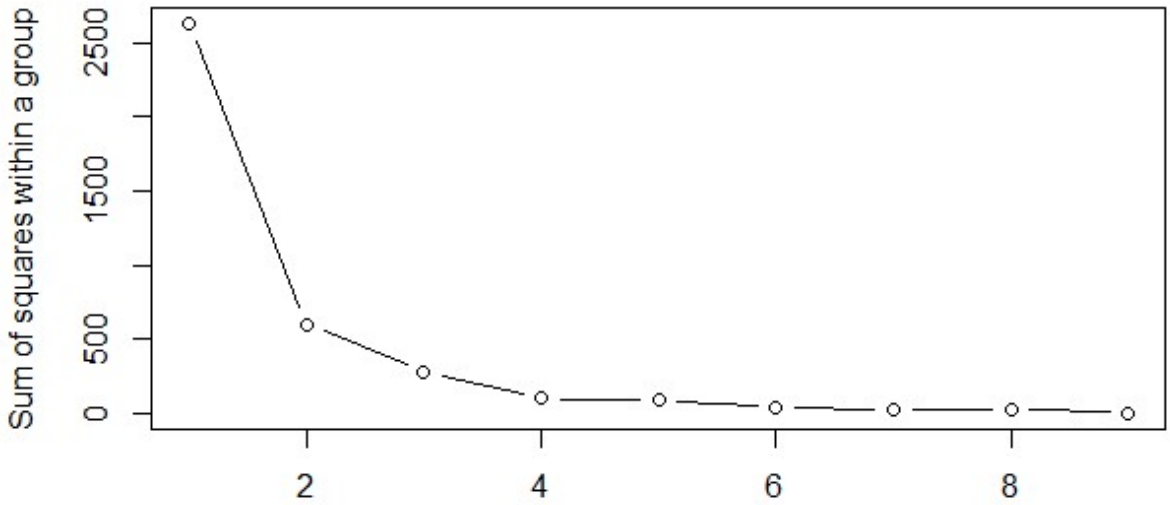

Figure 21: Sum of squares vs. Number of groups

Analyzing the above graph from right to left, we can see that the distance between the sum of squares within a group increases. That proves that there is reduction in the clustering. And the below graph shows the final result of the clustering.

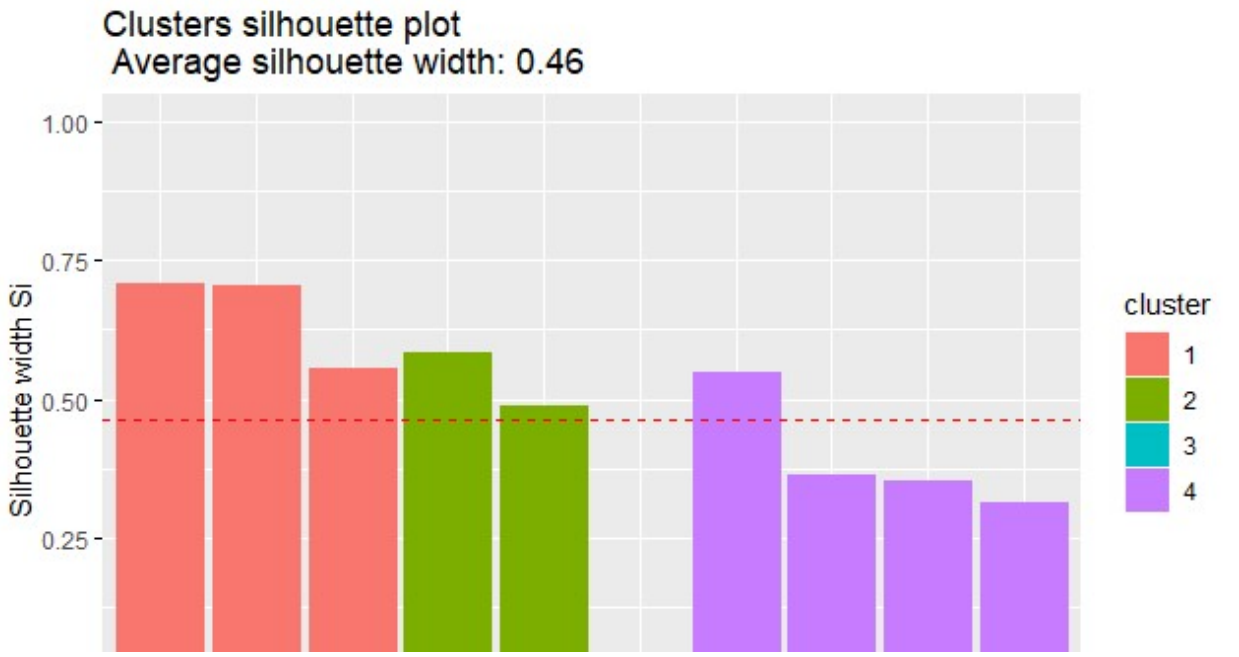

Figure 22: Cluster Silhouette plot

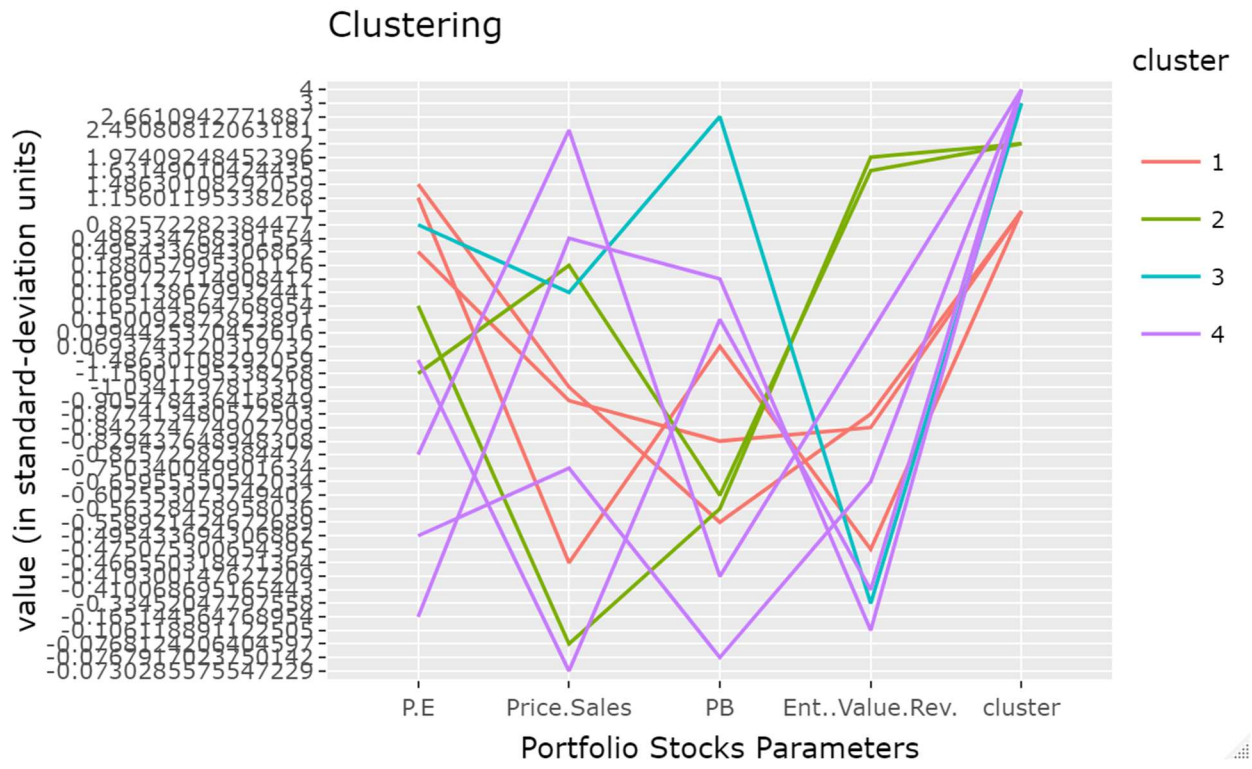

Figure 23: Clustering Analysis with P/E ratio, P/B ratio, price/sales, and Ent.Value/Rev

The above graph is for the clustering analysis using the parameters of Forward P/E ratio, Price/sales and price/book. The purpose of this chart is to determine the patterns in the data. The low values for the three ratios are as expected since we are analyzing the top performing stocks in different sector.

Conclusion

As of December 2021, **Berkshire Hathaway** has a market cap of **$657.56 Billion**. This makes Berkshire Hathaway the world's **9th** most valuable company by market cap. Our research suggests that the stock is still undervalued because it trades at a PE ratio of 7.96 and its book value of $317,000. Our analysis encourages that it's a **STRONG BUY** for the long term. 5 Major holdings are causing 66% of variations in the stock returns of Berkshire. The correlation coefficient for daily stock returns between Berkshire and rest of the holdings show that it's highest for Bank of America and lowest for Apple. It's a strong stock, it doesn't succumb to the high market volatility at several instances. The stock price recovers very quickly after any crisis and it posts a V-shaped recovery at every instance, which in turn means that investors have placed a lot of faith in the company. As BRK-A is basically a holding company that buys a stake in good performing businesses when the market undervalues them, hence it follows the concept of Value Investing. Through K means clustering we conclude that Berkshire Hathaway acquires most of the stocks which have a "low Enterprise Value to Price ratio", meaning that the companies are undervalued at the time of investing as compared to the others.

References

1. https://finance.yahoo.com/
2. https://www.berkshirehathaway.com

R Code-

```
options(scipen = 999)


library(ggplot2)

library(quantmod)

library(xts)

library(rvest)

library(tidyverse)

library(PerformanceAnalytics)

library(corrplot)

library(GGally)

library(cluster)

library(factoextra)

library(cluster.datasets)

library(tidyverse)

library(gridExtra)

library(factoextra)

library(plotly)


brk<-read.csv("BRK-A.csv")

brk<-data.frame(brk)

head(brk)


#Exploratory Analysis

#Converting Date into appropriate format


head(brk$Date)


strDates <- c(brk$Date)

brk$Date <- as.Date(strDates, "%m/%d/%Y")

brk$Date
```

```
a<-read.csv("annual.csv")

strDate <- c(a$Date)

date<- as.Date(strDate, "%m/%d/%Y")

a$MarketCap<- as.numeric(gsub("[^[:digit:]]","",a$MarketCap))


spx <- read.csv("SPY.csv")

strDates <- c(spx$Date)

spx$Date <- as.Date(strDates)


data <- read.csv("Closing prices.csv")

head(data)


summary(brk)


# RETURN SCREENSHOT


brka <- getSymbols("BRK-A",from="2000-12-30",auto.assign = F)

spy <- getSymbols("SPY",from="2000-12-31",auto.assign = F)

lineChart(spy$SPY.Open, line.type = 'h', theme = 'white', TA = NULL)

lineChart(brka$`BRK-A.Adjusted`, line.type = 'h', theme = 'white', TA = NULL)


#Market cap , PE ratio, PB Ratios


ggplot(data = a, mapping = aes(x = date, y = MarketCap)) +

  geom_point(alpha = 0.9, aes(color = MarketCap))


ggplot(data = a, mapping = aes(x = date, y = PeRatio)) +

  geom_line(alpha = 0.9, aes(color = PeRatio))


ggplot(data = a, mapping = aes(x = date, y = PbRatio)) +

  geom_line(alpha = 0.9, aes(color = PbRatio))



ret <- read.csv("returns.csv")

head(ret)
```

```
g <- ggplot(ret, aes(x=Year))
g <- g + geom_line(aes(y=brka), colour="blue")
g <- g + geom_line(aes(y=spy), colour="darkgreen")
g



# corrplot of holdings and BRK_A [STRONG CORRELATION]


ggcorr(data)
#Performing linear and multiple regression


#Data Description -
library(quantmod)
library(car)
library(MASS)
library(ggplot2)
data <- read.csv("dailyreturns.csv")
y <- data$BRK
x1 <- data$Apple
x2 <- data$BOA
x3 <- data$Amex
x4 <- data$KO
x5 <- data$Moodys


#scatterplots
par(mfrow=c(2,2))
plot(y ~ x1, xlab = "Apple Daily Return", ylab = "BRK Daily Return")
plot(y ~ x2, xlab = "BOA Daily Return", ylab = "BRK Daily Return")
plot(y ~ x3, xlab = "Amex Daily Return", ylab = "BRK Daily Return")
plot(y ~ x4, xlab = "KO Daily Return", ylab = "BRK Daily Return")
plot(y ~ x5, xlab = "Moodys Daily Return", ylab = "BRK Daily Return")



#Linear regression
linear.model1 <- lm(formula = y ~ x1)
abline(linear.model1, col = "red")
```

```
print(linear.model1)
summary(linear.model1)
linear.model2 <- lm(formula = y ~ x2)
abline(linear.model2, col = "red")
print(linear.model2)
summary(linear.model2)
linear.model3 <- lm(formula = y ~ x3)
abline(linear.model3, col = "red")
print(linear.model3)
summary(linear.model3)
linear.model4 <- lm(formula = y ~ x4)
abline(linear.model4, col = "red")
print(linear.model4)
summary(linear.model4)
linear.model5 <- lm(formula = y ~ x5)
abline(linear.model5, col = "red")
print(linear.model5)
summary(linear.model5)

#Multiple regression analysis
multiple.model <- lm(formula = y ~ x1+x2+x3+x4+x5)
abline(multiple.model, col = "red")
print(multiple.model)
summary(multiple.model)

#annova test
anova(multiple.model, test="Chisq")

#ANOVA helps you find out whether the differences between groups of data are statistically significant.

#Model Diagnostic Comparison
par(mfrow=c(2,2))
plot(multiple.model)

#finding the best model
full.model.formula <- y ~ x1+x2+x3+x4+x5
```

```
full.model <- lm(y ~ x1+x2+x3+x4+x5, data = data)
null.model <- lm(y ~ 1, data = data)
step(null.model, full.model.formula, direction = "both")

# From the multiple regression performed, we see that the model has p value which is most significant.
#We can consider a linear model to be statistically significant only when both these p-Values are less
#that the pre-determined statistical significance level, which is ideally 0.05
#Residual standard error - 10310 on 1725 degrees of freedom
#Multiple R squared : Higher the better and should be close to 1
#F-Statistic: Higher the better and Standard error closer to zero the better



#VOULME VISUALIZATION

ggplot(data = brk, mapping = aes(x = Close, y = Volume)) +
  geom_point(alpha = 0.2, aes(color = Volume))

#CURRENT VOLUME
ggplot(data = brk[5200:5250,], mapping = aes(x = Date, y = Volume)) +
  geom_point(alpha = 0.2, aes(color = Volume))

#DURING 2008 CRISIS
#High volume due and high volatility.
ggplot(data = brk[1500:2200,], mapping = aes(x = Date, y = Volume)) +
  geom_line(alpha = 0.2, aes(color = Volume))

# Date vs Close during 2008
#Stock halved

ggplot(data =brk[1700:2100,] , mapping = aes(x = Date, y = Close)) +
  geom_line(alpha = 0.9, aes(color = Close))

# Date vs Close during 9/11/2001

#High volume due and high volatility.
ggplot(data = brk[150:210,], mapping = aes(x = Date, y = Volume)) +
```

```
  geom_line(alpha = 0.9, aes(color = Volume))


# V shaped recovery
ggplot(data = brk[150:210,], mapping = aes(x = Date, y = Close)) +
  geom_line(alpha = 0.9, aes(color = Close))



chartSeries(brka,
            type = c("auto", "matchsticks"),
            subset = '2016-01::',
            show.grid = TRUE,
            major.ticks='auto', minor.ticks=TRUE,
            multi.col = FALSE,
            TA=c(addMACD(),addVo(),addSMA(n=250,col = 'blue'),addSMA(n=100,col = 'red'),addSMA(n=20,col = 'green'),
                 addROC(n=250,col = 'blue'),addROC(n=100,col = 'red'),addROC(n=20,col = 'green'))) # rate of change

#Measuring crisis using bollinger bands

chartSeries(brka, theme="white",
            TA="addVo();addBBands();addCCI()", subset = '2001-08::2001-12')

chartSeries(brka, theme="white",
            TA="addVo();addBBands();addCCI()", subset = '2007-01::2009-05')

chartSeries(brka, theme="white",
            TA="addVo();addBBands();addCCI()", subset = '2020-01::2020-06')




#K Means Clustering
data <- read.csv("clustering.csv")
# Let's remove the column with the stock names, so it won't be used in the clustering

input<-data[1:10,2:5]
input
```

```
# A function to check the best K value
wssplot <- function(data, nc, seed=123){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of groups",
       ylab="Sum of squares within a group")}

wssplot(input, nc =9)

set.seed(123)
clustering <- kmeans(input, centers = 4, nstart = 20)
clustering

sil <- silhouette(clustering$cluster, dist(input))
fviz_silhouette(sil)

input$cluster <- as.factor(clustering$cluster)

p <- ggparcoord(data = input, groupColumn = "cluster", scale = "std") +
  labs(x = "Portfolio Stocks Parameters", y = "value (in standard-deviation units)", title = "Clustering")
ggplotly(p)
```