R Final Project

Adam Patterson

4/13/2020

Introduction

For our final project, I would like to explore data on the stock market obtained from kaggle at https://www.kaggle.com/sturrion/stock-market-data/data#. After the EDA is complete, I am interested in running a regression to better understand if we can predict stock price with the selected independent variables. This would be helpful in evaluating stock price as I could put the actual values into the model and get a stock price. From that point, I could then see if (according to my model) the stock is overvalued or undervalued. I will have to log some of the independent variables and I should learn how to lag variables in R to complete a proper model.

One question I am curious about is if sector influences stock price. To run this test, I will create 10 dummy variables for the 11 Sector variables. My hypothesis is Technology will have the largest intercept value. Another question that I want to investigate is how many companies in each sector are within the top 100 companies in cash ? Is there a correlation between cash and profitMargin? Is there a correlation between profitMargin and revenuePerEmployee? I plan on experimenting many tests between the variables and providing visualizations about company and sector rank.

The variables used in this analysis:

Company.Name = Name of Company being observed

Sector = Sector that the company belongs to

EBITDA = Earnings Before Interest,Tax, Depreciation and Amortization

cash = cash position of company

debt = total liabilities of the company

dividendYield = percent of stock price paid out in dividends

institutionPercent = percent of institutional ownership

priceToSales = price to sales ratio

priceToBook = price to book ratio

profitMargin = profit margin of the company

returnOnAssets = return on company assets - a measure of capital efficiency

revenuePerEmployee = reveneue realized per employee

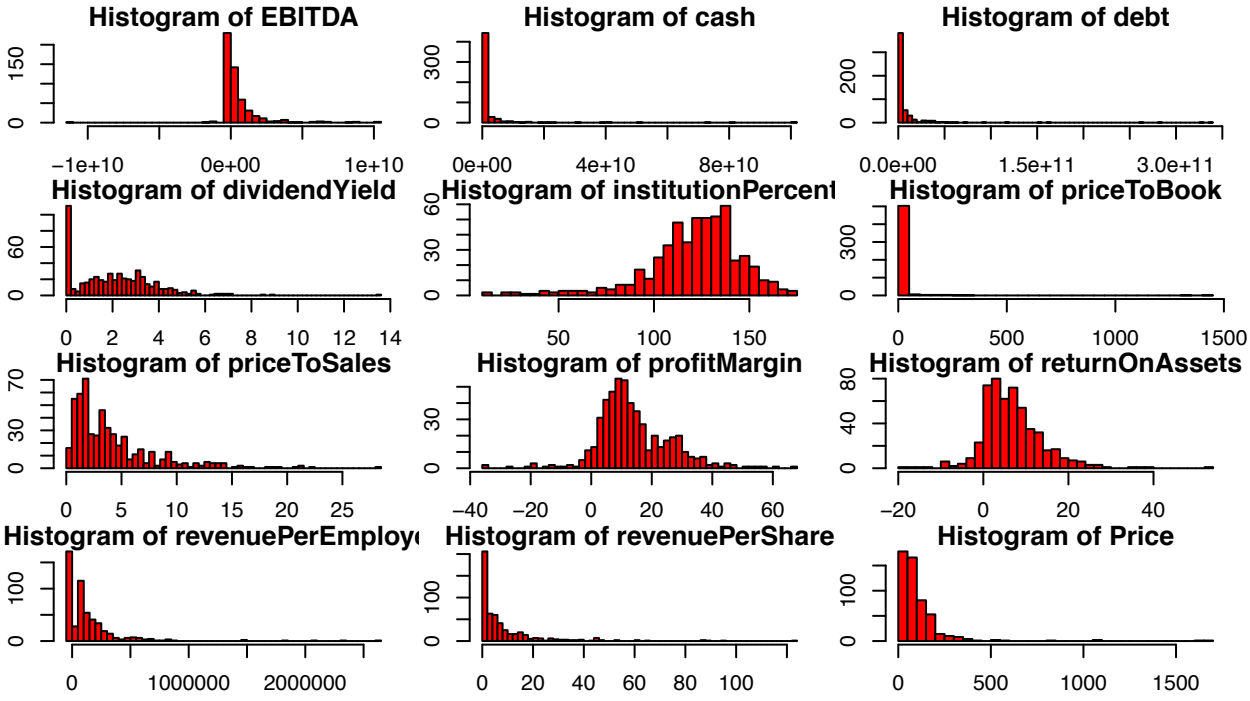
revenuePerShare = revenue realized per share of common stock

Price = price of the stock trading in the public market

Step 1 Variable Visualization

To start the EDA, we look at a histogram for each column variable.

```
setwd("~/Desktop/R_Final")
library(plyr)
data<-read.csv("stocks.csv", sep = ";")
stat<-data[,c(6,9,11,15,22,24,29,39,40,41,42,46,47,61)]
par(mfrow=c(5,3),mar=c(2,2,1,1))
for (j in 3:ncol(stat)) {
  hist(stat[,j], xlab=colnames(stat)[j],
       main=paste("Histogram of", colnames(stat)[j]),
       col="red", breaks=50)
}
```



We learn: 1) That priceToSales, revenuePerShare, revenuePerEmployee, dividendYield, cash, debt, price and insider percent are all right skewed 2) Most cash and debt values are closer to 0 3) The profitMargin, institutionPercent and returnOnAssets variables look rather symmetric and even normally distributed. 4) priceToBook values are all close to the lower bound of 0. We would expect this as priceToBook is just a ratio of the companys stock price over its book value per share

Step 2 Variable Correlation

We explore correlation coefficients for all of our column variables.

```
cor.stat<-cor(stat[,3:ncol(stat)])
round(cor.stat,3)
cor.stat[lower.tri(cor.stat,diag = TRUE)]=0
cor.stat
cor.stat.sorted<-sort(abs(cor.stat),decreasing = TRUE)
cor.stat.sorted[cor.stat.sorted>0]
Variable1<-rep(NA,15)
Variable2<-rep(NA,15)
Correlation<-rep(NA,15)
correlation.table<- data.frame(Variable1,Variable2,Correlation,row.names = 1:15)
for (i in 1:15) {
  which(abs(cor.stat)==cor.stat.sorted[i])
  var.big.cor <- arrayInd(which(abs(cor.stat)==cor.stat.sorted[i]), dim(cor.stat))
  correlation.table[i,1:2] <- colnames(cor.stat)[var.big.cor]
  correlation.table[i,3] <- cor.stat[var.big.cor]
}
correlation.table
```

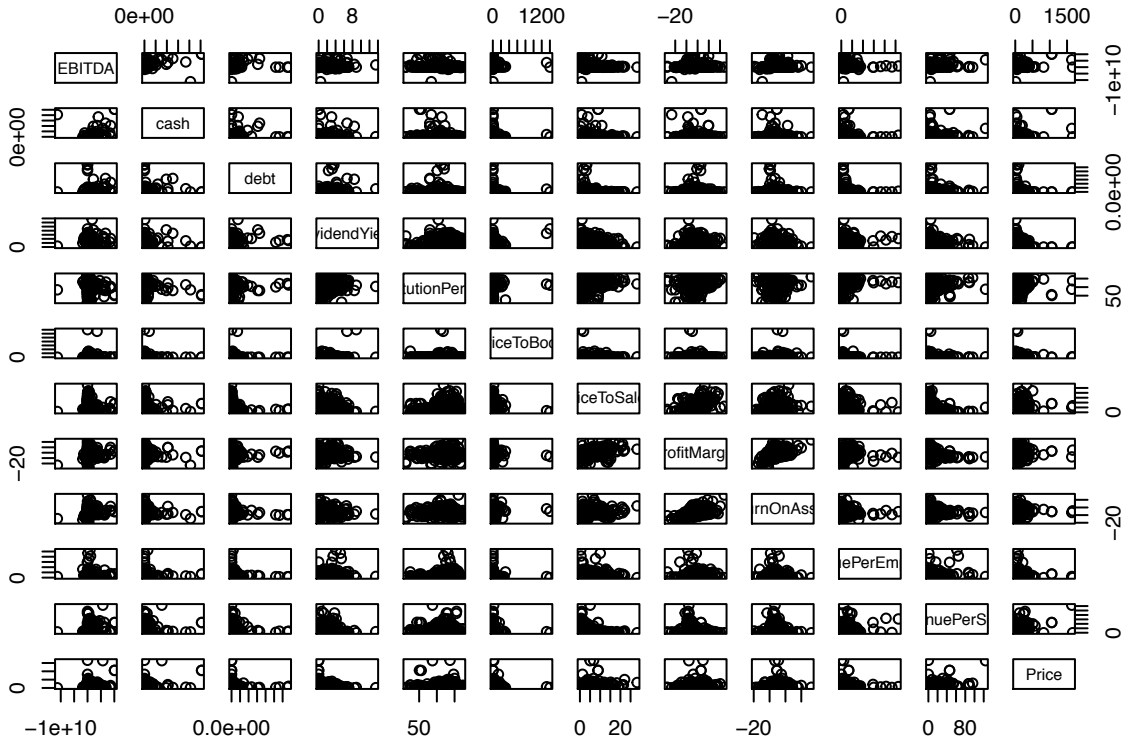
```
correlation.table
```

```
##             Variable1          Variable2 Correlation
## 1              EBITDA               cash   0.4317329
## 2     revenuePerShare              Price   0.4160784
## 3                cash              Price   0.3402518
## 4              EBITDA    revenuePerShare   0.3060142
## 5  revenuePerEmployee    revenuePerShare   0.2802446
## 6              EBITDA              Price   0.2787214
## 7       dividendYield              Price  -0.2672998
## 8        priceToSales              Price   0.2435039
## 9              EBITDA institutionPercent  -0.2166832
## 10               cash institutionPercent  -0.2131054
## 11       priceToSales    revenuePerShare  -0.2122174
## 12             EBITDA               debt   0.2116371
## 13               cash    revenuePerShare   0.2076943
## 14      dividendYield       priceToSales  -0.1918414
## 15               debt      dividendYield   0.1696685
```

Looking at the results, we notice EBITDA and cash has the highest correlation. revenuePerShare and Price are also highly correlated. Cash and Price have a high correlation as well. I am extremely surprised that there is not a more heavily correlation between DividendYield and Price as Dividend Yield is a direct function of stock price. Many variables have to be taken into account, such as dividend growth, however this correlation coefficient still seemed low to me. The negative correlation of cash and institutionalPercent ownership makes me almost think this dataset is fraudulent. I will have to think about reasons why this could be. That is 100% completely opposite of what my intution would tell me. After seeing such a high correlation between cash and EBITDA, I have decided to remove EBITA from my original model to potentially avoid multicollinearity.

Next, we can visualize the correlations between variables.

```
pairs(stat[,3:ncol(stat)])
```



Step 3 Model 1- Regression Model

In this step, we will perform a linear regression to see what variables have significance.

```
model_1<-lm(stat$Price~stat$cash+stat$debt+stat$dividendYield+stat$institutionPercent+stat$priceToSales+
summary(model_1)
```

```
##
## Call:
## lm(formula = stat$Price ~ stat$cash + stat$debt + stat$dividendYield +
##     stat$institutionPercent + stat$priceToSales + stat$profitMargin +
##     stat$returnOnAssets + stat$revenuePerShare + stat$revenuePerEmployee +
##     stat$priceToBook)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -332.29  -40.17   -9.28   23.19 1524.27
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.379e+01  2.588e+01   1.306    0.192
## stat$cash                3.898e-09  5.589e-10   6.975 9.53e-12 ***
## stat$debt               -2.629e-10  1.667e-10  -1.577    0.115
## stat$dividendYield      -1.268e+01  3.054e+00  -4.152 3.86e-05 ***
## stat$institutionPercent  5.004e-02  1.973e-01   0.254    0.800
```

```
## stat$priceToSales        9.445e+00  1.532e+00   6.163 1.45e-09 ***
## stat$profitMargin        6.911e-01  5.468e-01   1.264    0.207
## stat$returnOnAssets      1.192e+00  8.086e-01   1.474    0.141
## stat$revenuePerShare     4.748e+00  3.842e-01  12.359  < 2e-16 ***
## stat$revenuePerEmployee -8.551e-05  2.018e-05  -4.236 2.70e-05 ***
## stat$priceToBook         5.727e-02  5.575e-02   1.027    0.305
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.4 on 510 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4098, Adjusted R-squared:  0.3982
## F-statistic: 35.41 on 10 and 510 DF,  p-value: < 2.2e-16
```

Step 4 Model 2- Regression Model with Log

I have chosen two percentage variables to be run with a log( ) function.

```
model_2<-lm(stat$Price~stat$cash+stat$debt+stat$dividendYield+log(stat$institutionPercent)+stat$priceToS
```

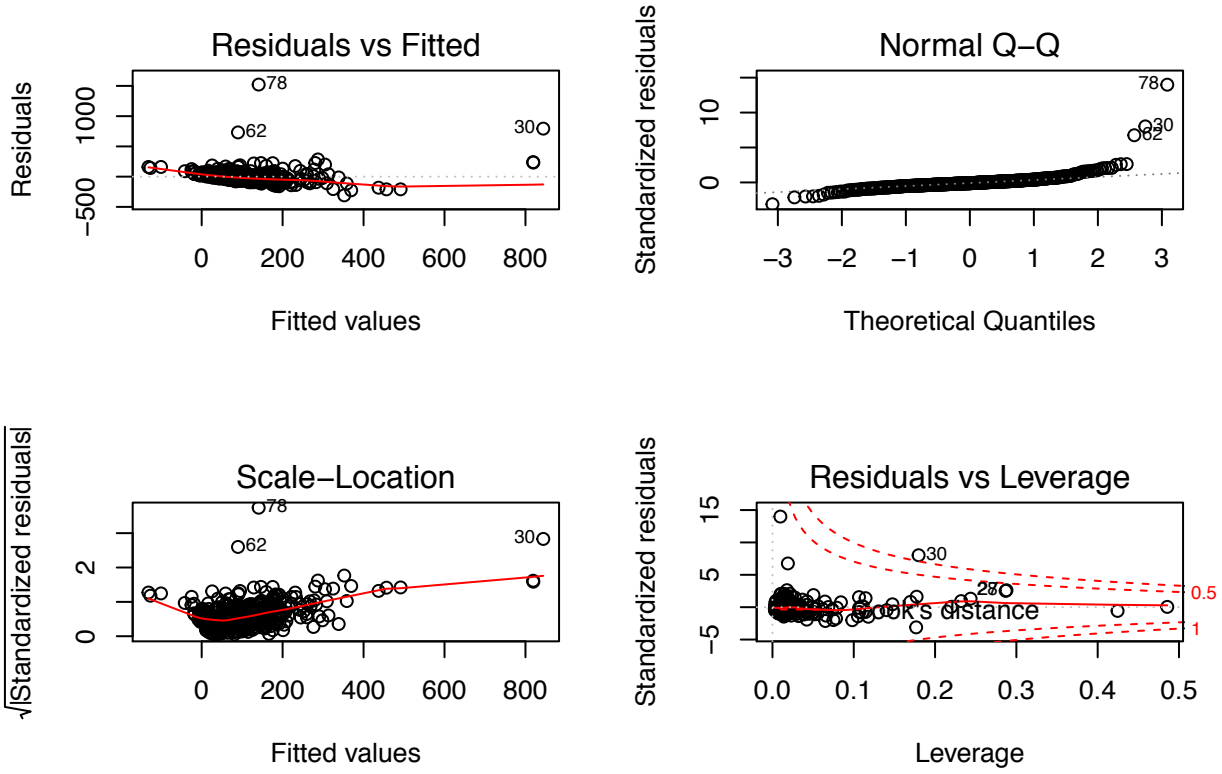
```
## Warning in log(stat$profitMargin): NaNs produced
```

```
summary(model_2)
```

```
##
## Call:
## lm(formula = stat$Price ~ stat$cash + stat$debt + stat$dividendYield +
##     log(stat$institutionPercent) + stat$priceToSales + log(stat$profitMargin) +
##     stat$returnOnAssets + stat$revenuePerShare + stat$revenuePerEmployee +
##     stat$priceToBook)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -309.12  -41.33  -11.59   22.52 1522.07
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.473e+01  9.382e+01  -0.264   0.7922
## stat$cash                    4.636e-09  6.134e-10   7.557 2.11e-13 ***
## stat$debt                   -3.216e-10  1.673e-10  -1.922   0.0551 .
## stat$dividendYield          -1.318e+01  3.208e+00  -4.107 4.71e-05 ***
## log(stat$institutionPercent)  8.043e+00  1.947e+01   0.413   0.6797
## stat$priceToSales            8.618e+00  1.589e+00   5.424 9.25e-08 ***
## log(stat$profitMargin)        1.708e+01  8.104e+00   2.108   0.0356 *
## stat$returnOnAssets          8.465e-01  8.229e-01   1.029   0.3041
## stat$revenuePerShare         5.027e+00  4.156e-01  12.095  < 2e-16 ***
## stat$revenuePerEmployee     -9.452e-05  2.140e-05  -4.418 1.23e-05 ***
## stat$priceToBook             5.971e-02  5.626e-02   1.061   0.2892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.3 on 479 degrees of freedom
```

```
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.4324, Adjusted R-squared:  0.4205
## F-statistic: 36.49 on 10 and 479 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model_2)
```



Looking over the results, Model 2 has a larger adjusted R square. The logged value of profitMargin became significant at the .05 level. This second model has a higher adjusted R squared than Model 1. Including regression visualization, we can see that a linear model appears to fit the data until a certain point. At that point, the residuals start to show a pattern moving away from 0. This gets really noticeable around the fitted value 350 and higher in the Residuals v Fitted scatterplot with abline. The deviation does not overwhelm the residuals from 0, so we will use a linear model in the Results section of this report.

Data Description

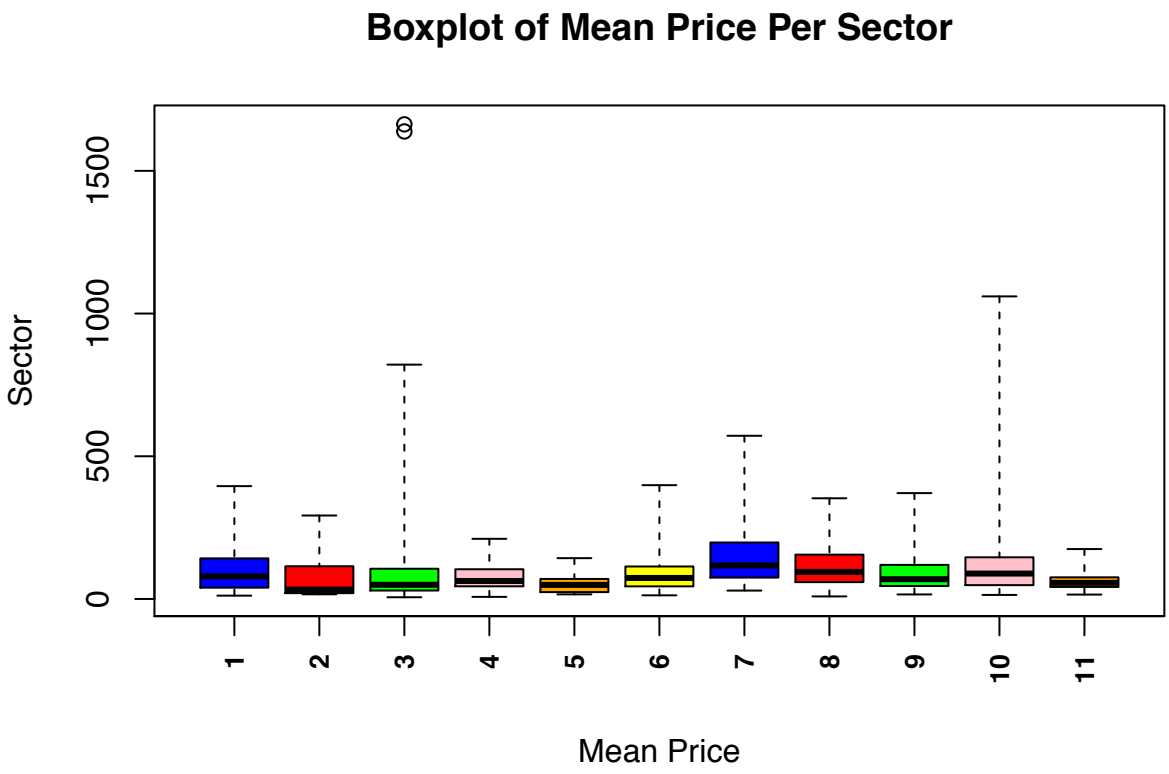
Observing the mean stock price by Sector

```
mean<-aggregate(stat$Price, list(stat$Sector),mean)
y<-order(mean$x, decreasing = TRUE)
industry.mean<-mean[y,]
names(industry.mean)[2]<-" Average Stock Price"
names(industry.mean)[1]<-" Sector"
industry.mean
```

```
##                    Sector  Average Stock Price
## 7              Healthcare            149.08039
## 3       Consumer Cyclical            125.73907
## 10             Technology            124.19043
## 8              Industrials            110.49247
## 1          Basic Materials             99.84675
## 9              Real Estate             97.45911
## 6      Financial Services             86.46932
## 2  Communication Services             76.65667
## 4      Consumer Defensive             75.76694
## 11               Utilities             62.42328
## 5                   Energy             53.12583
```

Looking through the results in the above table, I was really surprised that the Technology Sector was not the first. Adding to my surprise was how far Healthcare is ahead of all other industries. This is very intriguing to me. It is important to note that stock price is not a reflection of market cap, so this data alone does not tell us too much. I will run this exercise in the "Discussion" section with market cap instead of average stock price. That will give us a better guide into which Sectors have more access to capital in the public market. I did not want to include market cap in this my original model as it would not doubt be significant or correlated to stock price. That is an obvious relationship, unless one sector has companies constantly creating new shares, that is not a very interesting one to explore. In a longer report, or just for fun, I would explore this but not now. To see if we can extract any more information from the Sectors, I will create 10 dummy variables to replicate the 11 Sectors and run a linear regression of stock price regressed on the dummy variables. That model will be executed in "Results" section of this report. Below you will find a visualization of the above table.

```
boxplot(stat$Price~stat$Sector,col=c("blue","red","green","pink","orange","yellow","blue","red","green"
axis(1,seq(0,30,1),las=2,cex.axis=.8,font=2)
```



Finding out what Sectors are cash rich

```
x<-order(stat$cash, decreasing = TRUE)
stat1<-stat[x,]
cash<-stat1[,c("Company.Name","cash","Sector")]
top.cash<-cash[1:100,]
summary(top.cash$Sector)
```

```
##         Basic Materials Communication Services       Consumer Cyclical
##                       3                       4                      11
##      Consumer Defensive                  Energy      Financial Services
##                       7                       9                       6
##              Healthcare             Industrials             Real Estate
##                      19                      20                       0
##              Technology               Utilities
##                      20                       1
```

Although these results were not as surprising as the Sector mean results, I found the strength in the industrials suprising. This leads me to question if the Industrials are also the largest debt Sector in the top 100 debt rankings. This is because with higher fixed costs, I would not anticipate industrials to be cash rich on the technological level. I am noting the strength in the Healthcare industry in cash. I would be curious to see how they compare in debt also. First, I will just look at the top 20 cash rich companies and evaluate the Sector from that point. I am anticipating Technology to dominate this more narrow range of cash. After that, I will do the same exercise with debt rankings.

```
top1.cash<-top.cash[1:20,]
summary(top1.cash$Sector)
```

```
##         Basic Materials Communication Services      Consumer Cyclical
##                       1                       1                       3
##     Consumer Defensive                  Energy     Financial Services
##                       2                       1                       0
##              Healthcare             Industrials             Real Estate
##                       4                       3                       0
##              Technology               Utilities
##                       5                       0
```

As we can see in the above results, there is not a tremendous change in the ratios. The top 3 sectors remain relatively close and the laggards of Energy and Consumer Defensive are homogenously low. This is not really the result I thought/wanted to see. For the next exercise, I will include information about company debt in determing the cash position of a company. Although the three top Sectors of Healthcare, Technology, and Industrials are very well represented, it is important to note how many different sectors are represented here. Perhaps our economy is more diverse than I thought.

```
z<-order(stat$debt, decreasing = TRUE)
stat2<-stat[z,]
debt<-stat2[,c("Company.Name","debt","Sector")]
top.debt<-debt[1:100,]
summary(top.debt$Sector)
```

```
##         Basic Materials Communication Services      Consumer Cyclical
##                       3                       7                       9
##     Consumer Defensive                  Energy     Financial Services
##                       8                       8                      20
##              Healthcare             Industrials             Real Estate
##                      16                      15                       2
##              Technology               Utilities
##                       2                      10
```

Here, we find the major differences in industry. Technology companies are much more cash rich given their lower liabilites. As we can see, only 2 technology companies are on this list whereas 15 Industrial and 16 Healthcare comapnies made the cut. I am curious as to which Technology companies are so heavily indebted.

```
top.debt[which(top.debt$Sector == "Technology"),]
```

```
##           Company.Name       debt     Sector
## 116 Cisco Systems Inc. 3.9366e+10 Technology
## 253  Intel Corporation 2.6813e+10 Technology
```

Intel is not completely surprising, however Cisco surprised me a little as I know they had such a large cash position. Long live low interest rates, I guess. Next, lets try cash rich companies with no debt.

```
x1<-order(stat$cash, decreasing = TRUE)
stat3<-stat[x,]
cash1<-stat3[,c("Company.Name","cash","Sector","debt")]
top.cash1<-cash1[1:100,]
topcash.nodebt<-top.cash1[which(top.cash1$debt==0),]
topcash.nodebt
```

```
##                          Company.Name         cash             Sector debt
## 27                       Alphabet Inc. 101871000000         Technology    0
## 28                       Alphabet Inc. 101871000000         Technology    0
## 215          General Electric Company  82000000000        Industrials    0
## 191                      Facebook Inc.  41711000000         Technology    0
## 30                     Amazon.com Inc.  30986000000  Consumer Cyclical    0
## 100               Celgene Corporation  12042000000         Healthcare    0
## 349      Northrop Grumman Corporation  11225000000        Industrials    0
## 306           Mastercard Incorporated   7782000000 Financial Services    0
## 52              Applied Materials Inc.   7454000000         Technology    0
## 363               PayPal Holdings Inc.   5695000000 Financial Services    0
## 6               Activision Blizzard Inc   4713000000         Technology    0
## 487         Wellcare Health Plans Inc.   4668100000         Healthcare    0
## 24  Alliance Data Systems Corporation   4190000000 Financial Services    0
## 464         United Parcel Service Inc.   4069000000        Industrials    0
## 91               Campbell Soup Company   3906000000 Consumer Defensive    0
## 273              Juniper Networks Inc.   3032600000         Technology    0
## 335                       Netflix Inc.   2822795000  Consumer Cyclical    0
## 121                     CME Group Inc.   1993700000 Financial Services    0
## 261            Intuitive Surgical Inc.   1960600000         Healthcare    0
## 516                       NetEase Inc.   1922920961         Technology    0
## 435          T. Rowe Price Group Inc.   1902700000 Financial Services    0
## 119               Citrix Systems Inc.   1747646000         Technology    0
## 505               Zoetis Inc. Class A   1564000000         Healthcare    0
## 56                Arista Networks Inc.   1535555000         Technology    0
```

```
summary(topcash.nodebt$Sector)
```

```
##        Basic Materials Communication Services      Consumer Cyclical
##                      0                       0                       2
##     Consumer Defensive                  Energy     Financial Services
##                      1                       0                       5
##             Healthcare             Industrials             Real Estate
##                      4                       3                       0
##             Technology               Utilities
##                      9                       0
```

As we can see in the above results, Technology starts to widen the gap as a leader however the economy reamins relatively balanced. This is relieving to see. In theory, we are a very diverse economy and this just proves that with 6 sectors being represented. I noticed when I ranked companies by debt, Financial Service companies occupied the top several. This made me very curious as to who are the 5 Financial Service companies that have large cash and no debt.

```
topcash.nodebt[which(topcash.nodebt$Sector=="Financial Services"),]
```

```
##                          Company.Name       cash             Sector debt
## 306           Mastercard Incorporated 7782000000 Financial Services    0
## 363              PayPal Holdings Inc. 5695000000 Financial Services    0
## 24  Alliance Data Systems Corporation 4190000000 Financial Services    0
## 121                     CME Group Inc. 1993700000 Financial Services    0
## 435          T. Rowe Price Group Inc. 1902700000 Financial Services    0
```

I did not know they were including Fintech in Financial Services, in this case the results are not surprising. I thought they were actual banks and I wanted to see if Ally bank was listed. I thought they might have potentially been on the list due to their low fixed costs (relative to the banking industry). Ally is an online bank with no retail locations, that is what led me to believe they may have been cash rich. Paypal and Mastercard are on my top inflation hedges (along with Visa) so this information was not very surprising for me. Especially with the stimulus checks and massive recent money printing, I look for fintech companies to hedge potential inflation. It is worth noting that in recent years, it seems as if inflation has not risen in relation to the money supply. This could be due to Technology but I am preparing for an inflationary cycle after this most recent round of QE (hopefully not stagflation). I would have liked to see Visa in this group, however I know they have similar metrics.

```
grep("Ally",stat$Company.Name)
```

```
## integer(0)
```

Above, I tried to look up the observed data for Ally Financial and they are not in the database. I was curious as to their metrics given the above paragraph.

Description About Methods

The above Exploratory Data Analysis is very useful in providing visualization and descriptions of the data. In exploring two different models in the below Results section, both will be linear regression methods. While inspecting the results of the Actual vs Fitted results in Model 2 of the Introduction, I did not think that there was enough of a patterened deviation from 0 to use a non-linear model. Also in the results section is model 3, a regression model of stock price on Sector dummy variables. Model 3 will provide more statistical insight into the visualization (boxplot) that we observed in the Data Description section. Besides the heavy amount of data summarizations and exploration, the results section will consist of two linear regression analyses.

Results

Running Model 2 of the Introduction section so that we can further analyze and visualize results.

```
model_2<-lm(stat$Price~stat$cash+stat$debt+stat$dividendYield+log(stat$institutionPercent)+stat$priceToS
```

```
## Warning in log(stat$profitMargin): NaNs produced
```

```
summary(model_2)
```

```
##
## Call:
## lm(formula = stat$Price ~ stat$cash + stat$debt + stat$dividendYield +
##     log(stat$institutionPercent) + stat$priceToSales + log(stat$profitMargin) +
##     stat$returnOnAssets + stat$revenuePerShare + stat$revenuePerEmployee)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -311.89  -40.89  -11.27   22.69 1523.24
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2.188e+01  9.368e+01  -0.234   0.8154
## stat$cash                     4.656e-09  6.129e-10   7.596 1.60e-13 ***
## stat$debt                    -3.272e-10  1.671e-10  -1.958   0.0508 .
## stat$dividendYield           -1.238e+01  3.127e+00  -3.958 8.70e-05 ***
## log(stat$institutionPercent)  7.387e+00  1.943e+01   0.380   0.7040
## stat$priceToSales             8.705e+00  1.587e+00   5.486 6.67e-08 ***
## log(stat$profitMargin)        1.631e+01  8.076e+00   2.019   0.0440 *
## stat$returnOnAssets           1.001e+00  8.093e-01   1.236   0.2169
## stat$revenuePerShare          5.000e+00  4.139e-01  12.081  < 2e-16 ***
## stat$revenuePerEmployee      -9.504e-05  2.135e-05  -4.452 1.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.3 on 481 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.4304, Adjusted R-squared:  0.4197
## F-statistic: 40.38 on 9 and 481 DF,  p-value: < 2.2e-16
```

Reviewing the results of the linear regression above, we notice that five variables are statistically significant at the 0 level: cash, dividendYield, priceToSales, revenuePerShare, and revenuePerEmployee. One variable, log of profit margin, was significant at the .05 level thus the use of the logarithmic scale for our independent variable was valuable. Although the adjusted R squared is low at .41, it is useful to learn of significant independent variables. I plan on merging these significant independent variables with significant indpendent variables found in a similar project but with economic variables instead of business measures. In that project, I found interest rate, price of oil, price of gold, DXY, unemployment rate, CPI, and GDP to all be signiciant variables in estimating the S&P 500 Index Price.

Model 3- Running a regression after creating a dummy variable for 10 of the 11 Sectors of the Economy

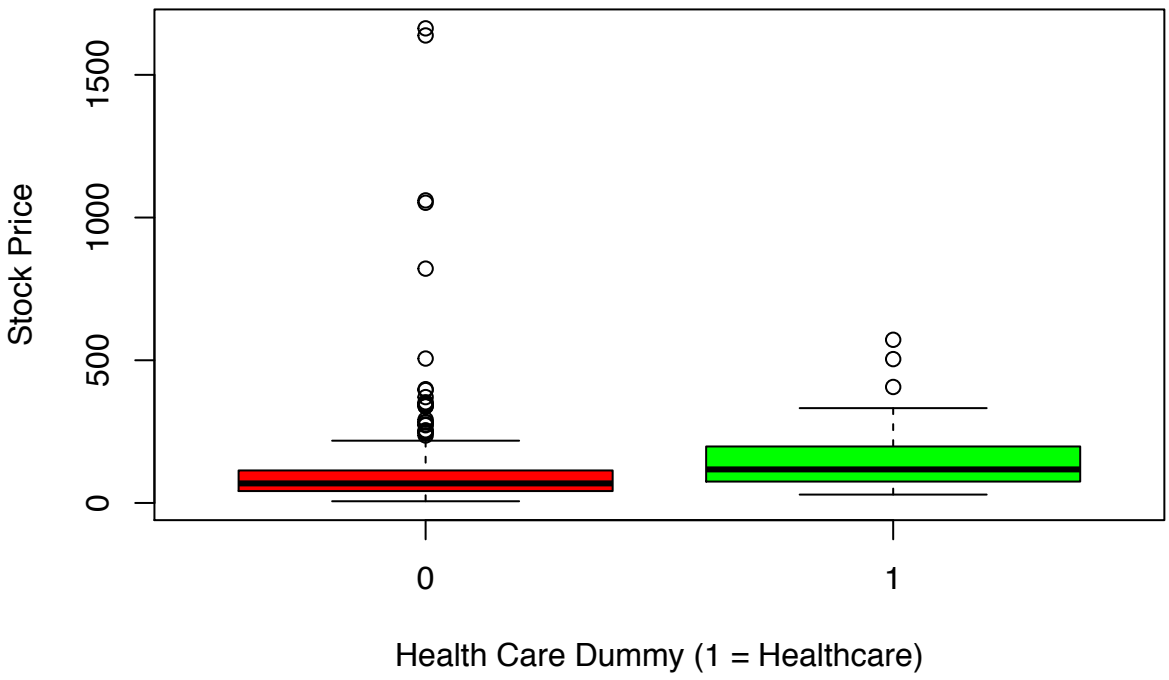
```
stat5<-stat
stat5$BasicMaterialsDummy<-revalue(stat5$Sector,c("Basic Materials"=1, "Communication Services"=0,"Consu
stat5$CommunicationServicesDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=
stat5$ConsumerCyclicalDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"Co
stat5$ConsumerDefensiveDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"C
stat5$EnergyDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"Consumer Cycl
stat5$FinancialServicesDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"C
stat5$HealthcareDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"Consumer
stat5$IndustrialsDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"Consume
stat5$RealEstateDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"Consumer
stat5$TechnologyDummy<-revalue(stat5$Sector,c("Basic Materials"=0, "Communication Services"=0,"Consumer
model_1<-lm(Price~BasicMaterialsDummy+CommunicationServicesDummy+ConsumerCyclicalDummy+ConsumerDefensive
summary(model_1)
```

```
##
## Call:
## lm(formula = Price ~ BasicMaterialsDummy + CommunicationServicesDummy +
##     ConsumerCyclicalDummy + ConsumerDefensiveDummy + EnergyDummy +
```

```
##     FinancialServicesDummy + HealthcareDummy + IndustrialsDummy +
##     RealEstateDummy + TechnologyDummy, data = stat5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.71  -61.24  -25.81   22.01 1537.18
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     99.847     31.185   3.202  0.00145 **
## BasicMaterialsDummy0           -37.423     40.537  -0.923  0.35634
## CommunicationServicesDummy1     14.233     47.871   0.297  0.76634
## ConsumerCyclicalDummy1          63.316     29.948   2.114  0.03498 *
## ConsumerDefensiveDummy1         13.344     34.800   0.383  0.70155
## EnergyDummy1                    -9.297     36.319  -0.256  0.79806
## FinancialServicesDummy1         24.046     30.613   0.785  0.43253
## HealthcareDummy1                86.657     31.219   2.776  0.00571 **
## IndustrialsDummy1               48.069     30.497   1.576  0.11560
## RealEstateDummy1                35.036     36.951   0.948  0.34349
## TechnologyDummy1                61.767     30.799   2.005  0.04544 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 139.5 on 512 degrees of freedom
## Multiple R-squared:  0.03767,    Adjusted R-squared:  0.01887
## F-statistic: 2.004 on 10 and 512 DF,  p-value: 0.0311
```

Looking at the above model results, we can see that the intercept (Utilities) has the most significance and the highest coefficient. We notice that Healthcare is also significant at the .01 level with a very large coefficient of 86.65. This regression actually supports the visualization earlier in this report. My hyopthesis was completely wrong. I underestimated the Healthcare sector and very much overestimated the Technology industry. Thus a company belonging to the Healthcare Sector can anticipate a stock price 86 dollars greather than if they were not in the Healthcare Sector. I noticed, although not significant, the negative coefficients for Basic Materials and Energy. We would not anticipate the adjusted R squared in this model to be high but we can most definitely identify sectors that are more favored on Wall Street. This in turn would result in an edge because the company would have more capital available for capital expenditures. Below is a visualization of the most significant dummy variable. You can notice a substantial increase in the mean but more importanty, the interquartile range is larger for the Healthcare Sector. This suggests that although the mean is higher, variation is also higher. Which is very interesting because we can see that the non-health care industries have many more outliers than the Healthcare Industry. This further emphasizes the large deviation in Healthcare Sector.

```
plot(stat5$HealthcareDummy,stat5$Price, xlab="Health Care Dummy (1 = Healthcare)",ylab="Stock Price", c
```



Discussion

We have found cash to be very important in estimating stock price. In this discussion section, I would like to discuss and explore what companies and Sectors could potentially grow their cash reserves in the future. My intuition tells me that profitMargin should be heavily associated. Recalling what we learned in the correlation table in Introduction Step #2, EBITDA and cash had the highest correlation out of all independent variables in the dataset. EBITDA could be somewhat used as predictor of cash but what about the other measures ? For curiousity, I decided to regress cash on the other indepdent variables used in this dataset.

```
model_5<-lm(stat$cash~stat$debt+stat$dividendYield+log(stat$institutionPercent)+stat$priceToSales+log(s
```

```
## Warning in log(stat$profitMargin): NaNs produced
```

```
summary(model_5)
```

```
##
## Call:
## lm(formula = stat$cash ~ stat$debt + stat$dividendYield + log(stat$institutionPercent) +
##     stat$priceToSales + log(stat$profitMargin) + stat$returnOnAssets +
##     stat$revenuePerShare + stat$revenuePerEmployee)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -1.413e+10 -2.263e+09 -6.726e+08  8.535e+08  8.765e+10
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.123e+10  6.816e+09   4.582 5.88e-06 ***
## stat$debt                     3.660e-02  1.231e-02   2.974  0.00309 **
## stat$dividendYield            2.344e+08  2.322e+08   1.010  0.31317
## log(stat$institutionPercent) -6.860e+09  1.410e+09  -4.864 1.56e-06 ***
## stat$priceToSales             3.636e+08  1.168e+08   3.114  0.00195 **
## log(stat$profitMargin)       -3.881e+07  6.002e+08  -0.065  0.94848
## stat$returnOnAssets          -9.520e+06  6.015e+07  -0.158  0.87430
## stat$revenuePerShare          1.656e+08  2.982e+07   5.554 4.63e-08 ***
## stat$revenuePerEmployee      -2.129e+02  1.587e+03  -0.134  0.89331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.122e+09 on 482 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.1376, Adjusted R-squared:  0.1232
## F-statistic:  9.61 on 8 and 482 DF,  p-value: 2.185e-12
```

The results are not good in terms of my hypothesis. Debt would have been a more obvious choice, but I am not surprised to see that the two significant variables were the other variables correlated with cash in our Introduction correlation table. I only took EBITDA out because it was the highest correlated. The adjusted R square is not very good in this model so there are much more efficient predictors of cash out there. profitMargin was not significant at any level and that is very disappointing to my hypothesis and intuition. I wonder if this has something to do with net cash of debt. I am curious to see if profitMargin would be correlated with net cash. To do this, I will create a new variable called "net.cash" and regress net.cash on the independent variables to see if that makes a difference with regard to profitMargin. After reviewing the results below, we notice that the adjusted R squared actually decreases while creating the the new "net.cash" variable (probably due to the large 0 or NA values included within the transformed net.cash variable). Again these are not any desired results we wish to see, and normally I would exclude them from a report but this project is about our ability to use the materials learned through R, which is why I include these exercises that had undesirable results.

```
stat6<-stat
stat6$net.cash<-(stat6$cash-stat6$debt)
model_4<-lm(stat6$cash~stat6$dividendYield+log(stat6$institutionPercent)+stat6$priceToSales+log(stat6$p
```

```
## Warning in log(stat6$profitMargin): NaNs produced
```

```
summary(model_4)
```

```
##
## Call:
## lm(formula = stat6$cash ~ stat6$dividendYield + log(stat6$institutionPercent) +
##     stat6$priceToSales + log(stat6$profitMargin) + stat6$returnOnAssets +
##     stat6$revenuePerShare + stat6$revenuePerEmployee)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.334e+10 -2.431e+09 -8.295e+08  7.294e+08  8.696e+10
```

```
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.331e+10  6.834e+09   4.874 1.48e-06 ***
## stat6$dividendYield            3.294e+08  2.318e+08   1.421  0.15598
## log(stat6$institutionPercent) -7.336e+09  1.412e+09  -5.194 3.04e-07 ***
## stat6$priceToSales             3.371e+08  1.174e+08   2.873  0.00425 **
## log(stat6$profitMargin)        1.858e+08  6.003e+08   0.309  0.75709
## stat6$returnOnAssets          -3.374e+07  6.007e+07  -0.562  0.57468
## stat6$revenuePerShare          1.721e+08  2.998e+07   5.740 1.67e-08 ***
## stat6$revenuePerEmployee      -3.706e+02  1.599e+03  -0.232  0.81677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.188e+09 on 483 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.1217, Adjusted R-squared:  0.109
## F-statistic: 9.563 on 7 and 483 DF,  p-value: 3.765e-11
```

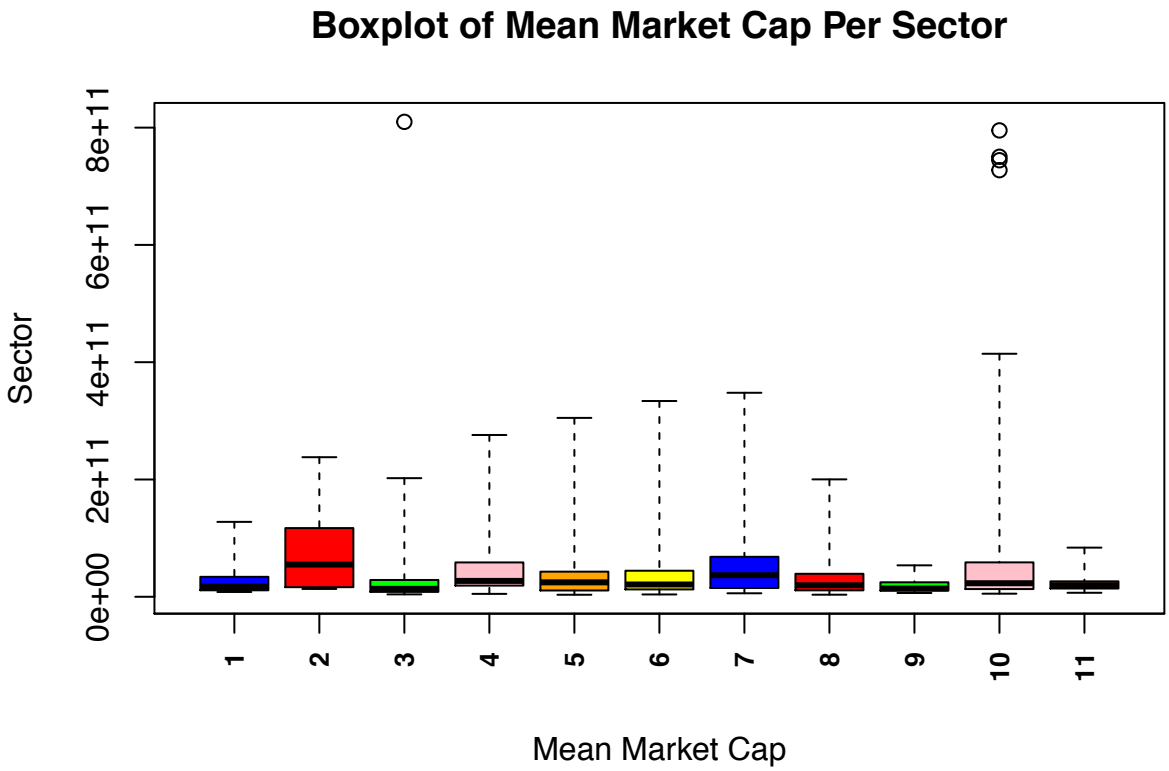
Continuting on from the Data Description section, I was curious as to the average market cap per industry. I knew that stock price was rather an inaccurate predictor because that data did not account for the number of shares.

```
stat8<-data[,c(6,9,11,15,22,24,29,32,39,40,41,42,46,47,61)]
mean1<-aggregate(stat8$marketcap, list(stat8$Sector),mean)
w<-order(mean1$x, decreasing = TRUE)
sector.mean<-mean1[w,]
names(sector.mean)[1]<-"Sector"
names(sector.mean)[2]<-"Mean Market Cap"
sector.mean
```

```
##                    Sector Mean Market Cap
## 10             Technology     85500968005
## 2  Communication Services     79219394774
## 7              Healthcare     55692873511
## 4       Consumer Defensive     53586827099
## 6       Financial Services     48580740241
## 5                   Energy     41015546291
## 3        Consumer Cyclical     37153713719
## 8              Industrials     31009306166
## 1          Basic Materials     26060594642
## 11               Utilities     24816590350
## 9              Real Estate     18972073102
```

The results are much more in line with what I hypothesized with Technology leading the way. Commmunication Services is a close second as former Technology companies, such as Alphabet and Facebook, have recently been reclassified from the Technology Sector to the Communication Services Sector. Reviewing the results further, the ranking of Energy is much more than I would have anticipated. Given the combination of recent events, the price of oil has plummeted. Although OPEC plus just reached a production cut on April 12, I believe the demand side is a larger story than the supply side. I am very worried about the overall health of our economy if oil stays low (around 20 dollars per barrel). Many of these energy companies will default on debt, causing ripple effects into the junk bond market which is not good. The bond market dwarfs the stock market in size so whenever the bond market could have a problem, economists must be alert. Below, we will visualize our above table results:

```
boxplot(stat8$marketcap~stat8$Sector,col=c("blue","red","green","pink","orange","yellow","blue","red","
axis(1,seq(0,30,1),las=2,cex.axis=.8,font=2)
```



We can see several outliers in the non-Healthcare Sectors, Amazon, Apple, Microsoft, and Alphabet. When I look at this visual, it really puts into perspective how dominant a few companies are in the S&P 500. The top 5 account for more than 20% of the total stock market value. It will be interesting to see where these companies go in the future. Will the age of Technology allow them to keep their competitive advantages?

Conclusion

Although this project was fun, and I got to brush up on some R skills that I have not used in a while, it was not statistically meaningful in that I would feel confident predicting price from any model in this report. My original goal of the project has failed, for now. Company data alone does not explain much of the variation in stock price. In previous projects, I have found that economic variables have much more importance. While running the models in this report, I tried lagging the price variable by 1 unit and did not get any difference. I would have to explore lagging more because my intution tells me that should have an effect. I did find interesting relations on the Sector level and it was fun to speculate ranking the companies by cash. I was very impressed with the profit margins of real estate companies such as Public Storage. Also, Twitter became a real company to me while doing this project. I was pleasantly surprised with their metrics and recent stock price. There are a few more companies I will continue to keep an eye on, Paypal for sure.