Statistics in the Stock Market

Camden Jones

2023-11-27

Data

At the link below you will find the data to be used for this paper. The data consists of the 500 companies in the S&P500 index as of 2014. For each company, the name, sector, and price per share are listed, along with a litany of other financial metrics. Today, we will be concerned only with the market cap, the Price to Earnings Ratio, and the EBITDA (Earnings Before Interest Tax Depreciation and Amortization).

Click Here For the Data

```r
urlremote <- "https://www.dropbox.com"
dbpath <- "/scl/fi/vaobsyplry4dmknx1x0pi/constituents-financials.csv"
misc <- "?rlkey=xie7rysshi63eaoi1pnizbl2k&dl=1"
index <- paste0(urlremote, dbpath, misc) %>%
    read.csv()
```

Question

Stocks are often described as following random paths or reacting to hidden catalysts that only the richest and most powerful are privy to know about. Burton Malkiel's "A Random Walk Down Wall Street" is a prime example, as Malkiel sets forth a bounty of data supporting the thesis that broad index fund investing is superior to individual stock picking. Why should one submit oneself to the random walk of individual stocks when one can just enjoy the rising tide of the US economy which inherently lifts all boats (stocks)? I would argue that though index investing is passive on the investor's end, there is an even greater amount of activity happening on the index end, as companies go bankrupt or slowly die off and leave the index, and as new startups burst into a meteoric rise thanks to disruption. This amount of change and activity is far greater than an individual investor's 30 stock portfolio with possibly a few slow growing companies and dividend aristocrats. So why are people so bad at picking individual stocks with the immense power of the human brain, all while the S&P500 cranks out consistent 10-11% annual gains for 100 years just by slapping the largest companies on a list? We will examine the relationship between a company's earnings (EBITDA) and its market cap. Market cap is the number of outstanding shares multiplied by the share price; it essentially gives a rough estimate of the "market value" of any given business as evaluated by Mr. Market. We will also take a look at how Price to Earnings ratios are distributed to see if there is any pattern to the valuation of companies. So let's take a look and see how much of a random walk any individual stock had taken to get to where it was in 2014.

The stock market, while consisting of thousands of companies listed across multiple exchanges, is usually evaluated on performance by examing one of a few select indexes that contain a specific group of companies. The S&P500 is one of those indexes, and it contains the 500 largest companies listed on US exchanges by market capitalization.The S&P500 itself is often used as a sort of wind vane for the US economy and stock market overall. For this reason and that of manageability, we will be working with the approximately 500 companies in the index for our analysis. The index undergoes quarterly updates, so that is the reason for the extra companies on the list.

Descriptive Statistics

First, a breakdown of the descriptive statistics for our three aforementioned metrics: Market Cap, EBITDA, and Price Earnings. We will clean the original data of any entries containing “NA” so that statistics can be calculated for the whole set.

Market Cap

```r
cleanindex <- na.omit(index)
summary(cleanindex$Market.Cap)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 2.626e+09 1.275e+10 2.143e+10 4.938e+10 4.499e+10 8.095e+11
```

```r
sd(cleanindex$Market.Cap)
```

```
## [1] 90691616350
```

Below, the raw numbers in scientific notation have been converted into cleaner figures in ($) billions.

Min.: 2.626

1st Q: 12.75

Median: 21.43

Mean: 49.38

3rd Q: 44.99

Max: 809.5

Standard Deviation: 90.69

Just this summary paints a vivid picture of the data. The vast majority (75%) of companies are below $45 billion in market cap. However, the mean, a supposed measure of central tendency, is actually greater than $45 billion! This indicates some significant influence from outliers on the high end of the data. A maximum of $809 billion corroborates this suspicion when all other statistics fall under $50 billion. Additionally, with a whopping $90 billion standard deviation (a spread which is larger than double the range of the first 3 quartiles), we almost know for certain multiple outliers exist.

EBITDA

```r
summary(cleanindex$EBITDA)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
## -5.067e+09  7.709e+08  1.613e+09  3.576e+09  3.658e+09  7.939e+10
```

```r
sd(cleanindex$EBITDA)
```

```
## [1] 6880742484
```

Because of the variance in EBITDA figures, below the statistics are listed with more clarity.

Min: -5.067 billion

1st Q: 770.9 million

Median: 1.613 billion

Mean: 3.576 billion

3rd Q: 3.658 billion

Max: 79.39 billion

Standard Deviation: 6.88 billion

With some small differences, this set of summary statistics paints a similar picture to those of market caps. For one, EBITDA can go negative (a loss), while the value of a company (market cap) cannot be negative. This introduces a wider range on the lower end of the data. However, we still see a large concentration of the data on the lower end of the range (75% of the data falls below $4 billion), but the maximum extends way out to almost $80 billion. Here the mean appears ever so slightly less influenced by outliers, as it fails to eclipse the 3rd quartile, but it is still more than double the median. Once again the standard deviation at almost $7 billion spans a large range of the data, but this time it does not fully span the first 3 quartiles since the minimum reaches far into negative territory. These similarities and differences indicate multiple outliers, but likely fewer than in the market cap set. What does this mean? Fewer companies are enjoying outsize earnings than are enjoying outsize market cap. This means there are likely some outliers in the way that companies are valued (Price Earnings ratios).

Price Earnings Ratio

```
summary(cleanindex$Price.Earnings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -251.53   15.35   19.45   24.65   25.77  520.15
```

```
sd(cleanindex$Price.Earnings)
```

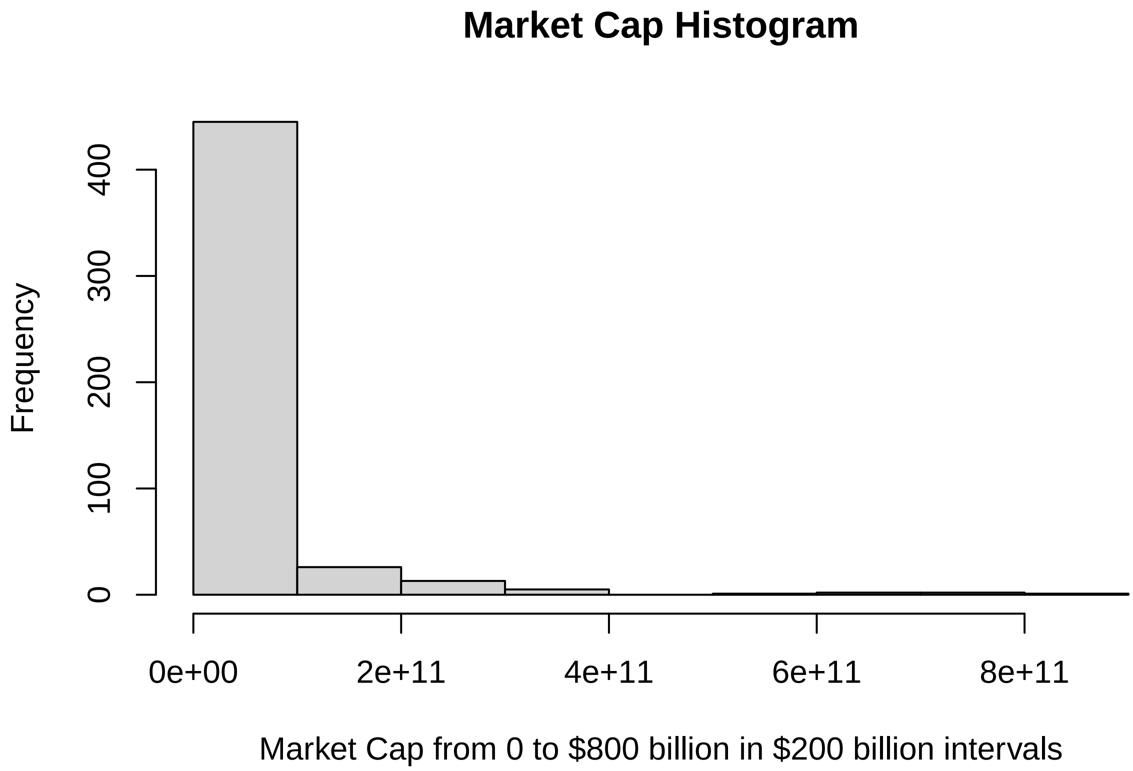
```
## [1] 40.76708
```

Since these numbers are already ratios of price per share to earnings per share, there is no need for reformatting of the figures. An overall picture of the data is starting to form. All three data sets yield summary statistics which indicate significant spread, multiple outliers, and a concentration of data around the lower end of the range.

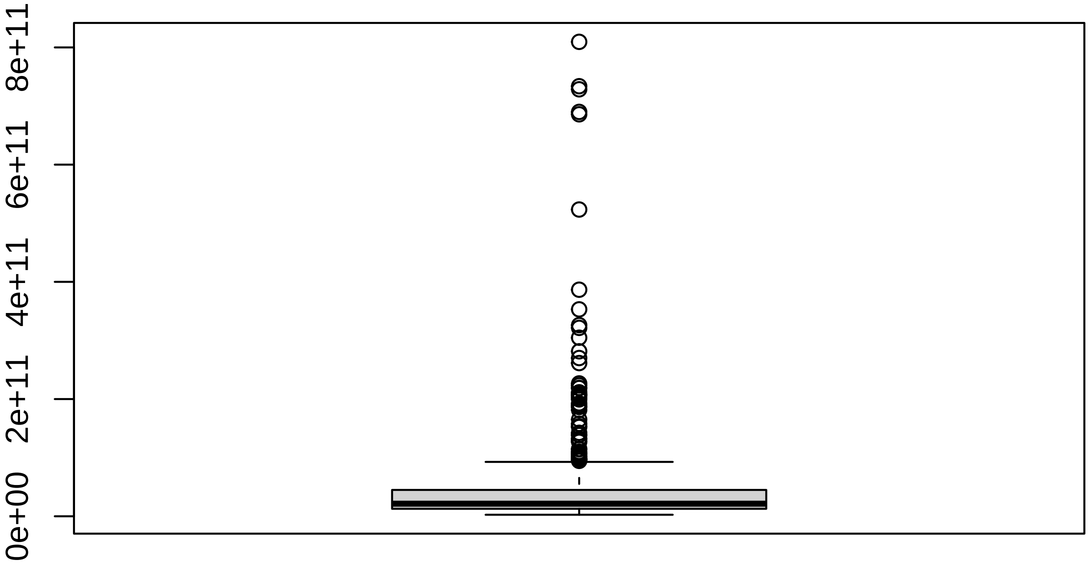
Visualization

Market Cap

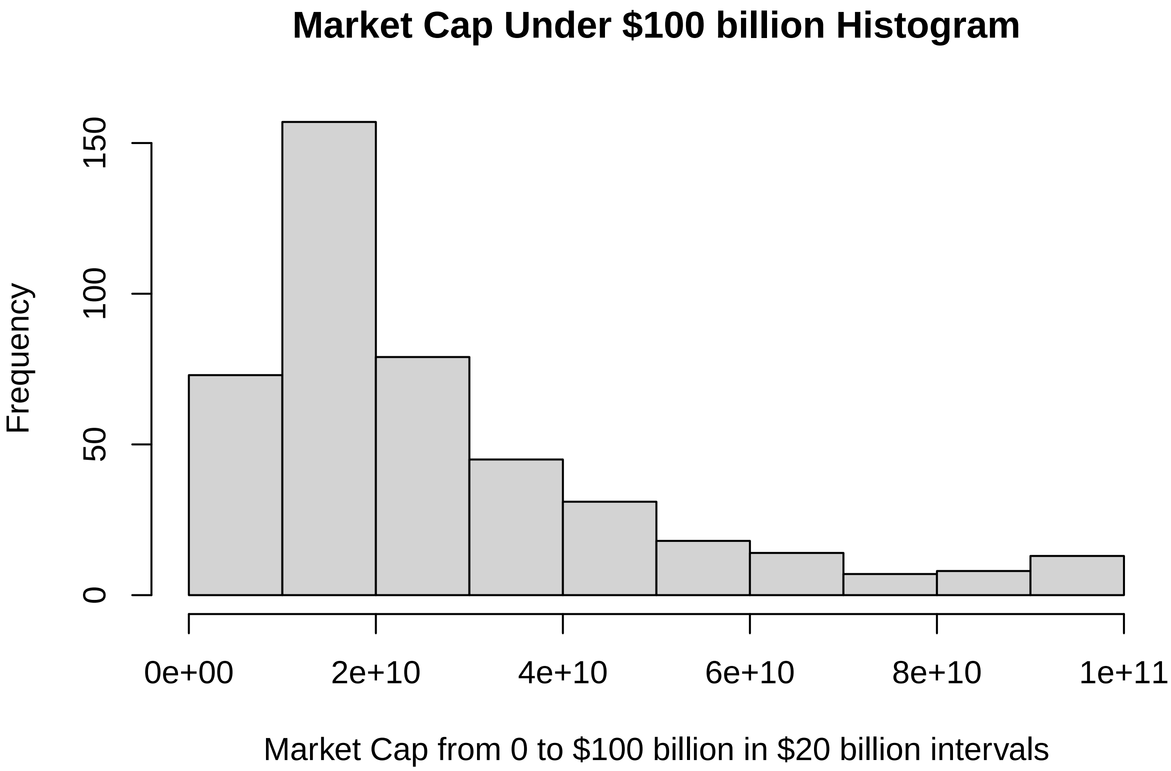
```
hist(cleanindex$Market.Cap, main = "Market Cap Histogram", xlab = "Market Cap from 0 to $800 billion in
```



```
boxplot(cleanindex$Market.Cap)
```



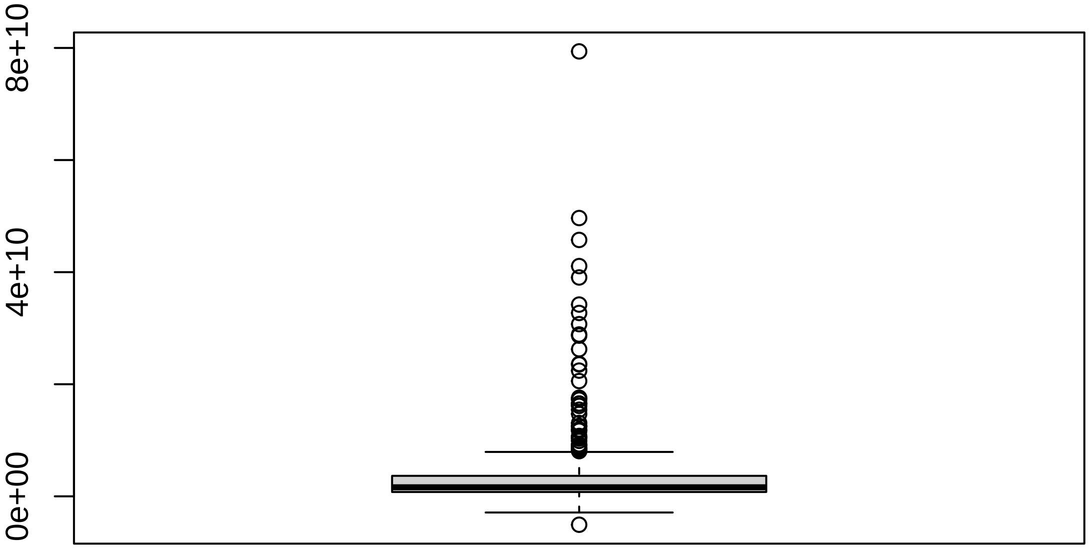
```
subhundred <- filter(cleanindex, cleanindex$Market.Cap < 1e+11)
hist(subhundred$Market.Cap, main = "Market Cap Under $100 billion Histogram",
    xlab = "Market Cap from 0 to $100 billion in $20 billion intervals")
```



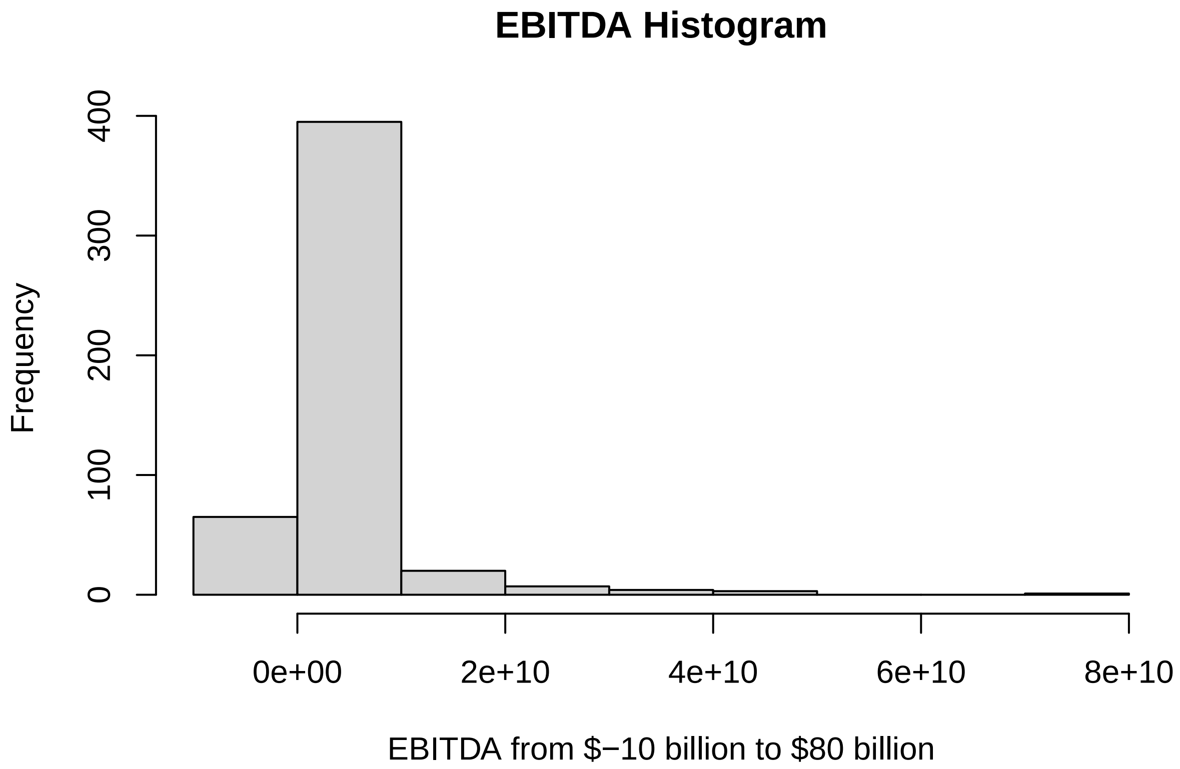
Above we have a histogram of the market caps which indicates an extreme skew to the right and confirms a significant concentration of the data on the lower end of the range. We see that the vast majority of the market caps fall between $0 and $100 billion. So we filter the data for those companies only, and this zoomed in histogram presents a right-hand skew in even the concentrated part of the data. Additionally, it presents even stronger concentration on the lower end of the range. In fact, just 50 companies were removed from the data set when the sub-$100 billion market cap constraint was introduced. The bulk of the data lies in the second histogram, where a more precise skew is taking place. The boxplot confirms these findings and reveals a plethora of outliers on the high end of the data, with none on the low end.

EBITDA

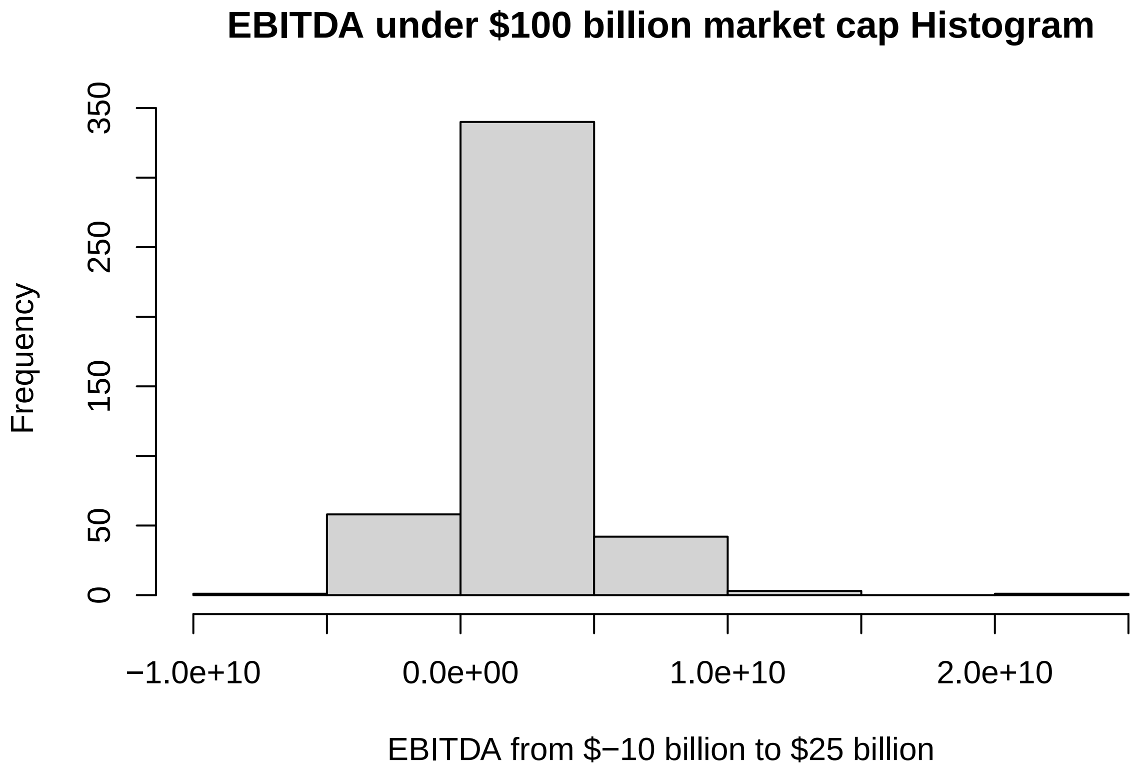
```r
boxplot(cleanindex$EBITDA)
```



```
hist(cleanindex$EBITDA, main = "EBITDA Histogram", xlab = "EBITDA from $-10 billion to $80 billion")
```



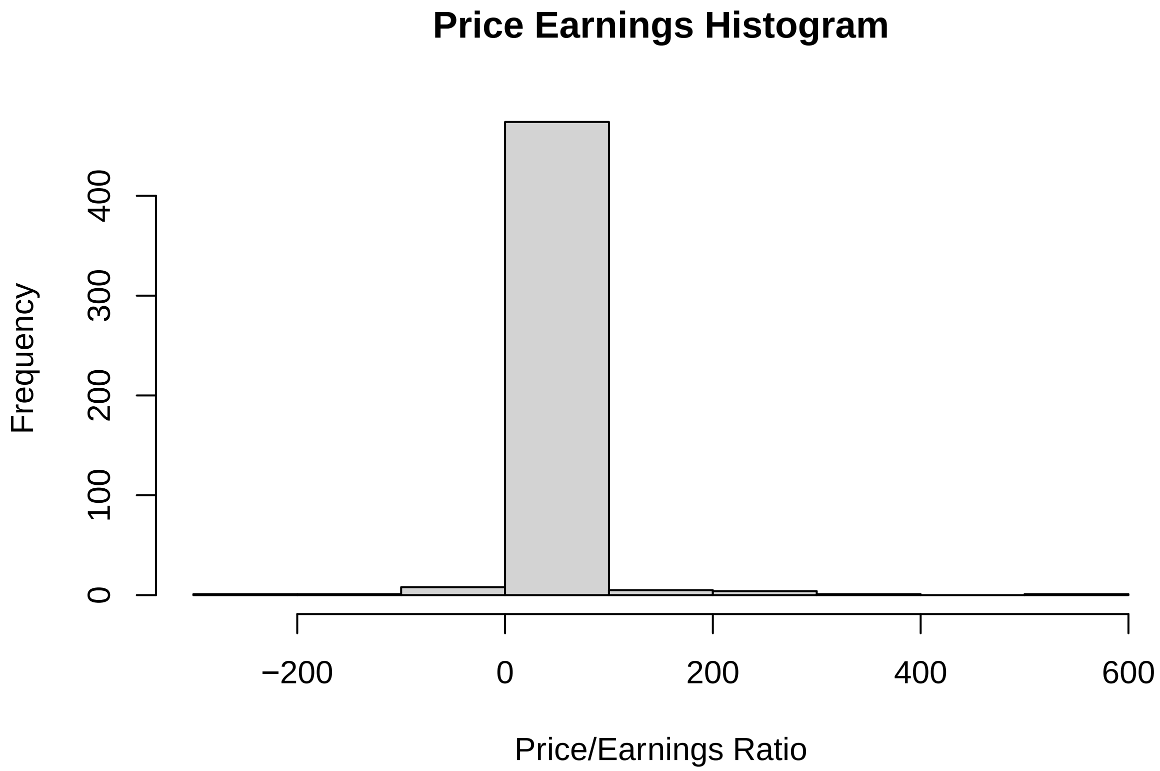
```
hist(subhundred$EBITDA, main = "EBITDA under $100 billion market cap Histogram",
    xlab = "EBITDA from $-10 billion to $25 billion")
```



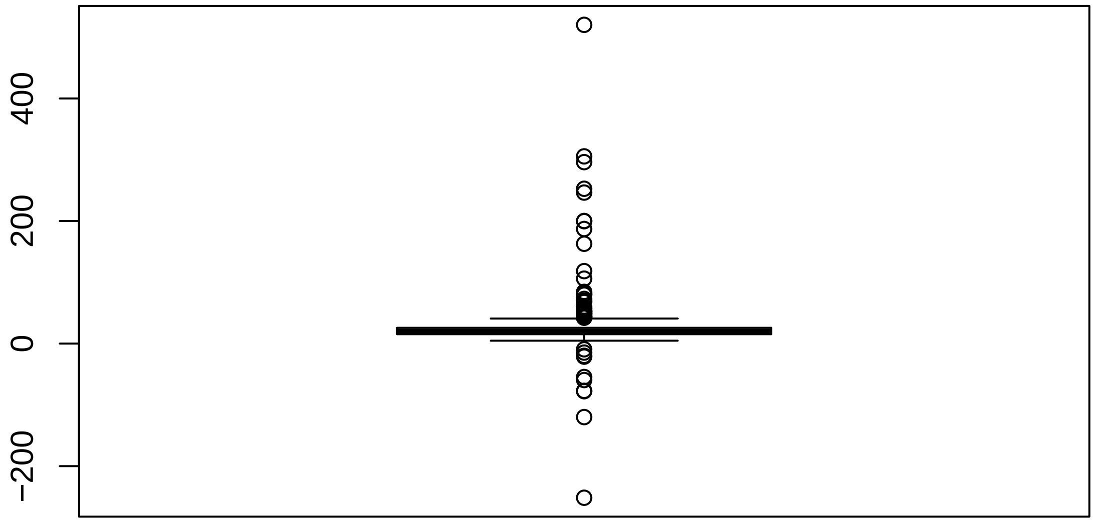
Here we examine the histogram of EBITDA and again notice extreme skew to the right and concentration in the low end of the data. When we limit the data to just companies under $100 billion in market cap, we start to notice an approximately normal shape (unimodal, symmetric, and centered around the low end of the range where we saw concentration in the first histogram). The right skew in the lower end of the data appears to be a result of market caps being limited to positive values while many companies have negative earnings (EBITDA). Would investors value companies in negative dollars if they could? With a loss in the hundreds of millions, probably. The boxplot again confirms many outliers on the high end of the data, with only one on the low end.

Price Earnings Ratio

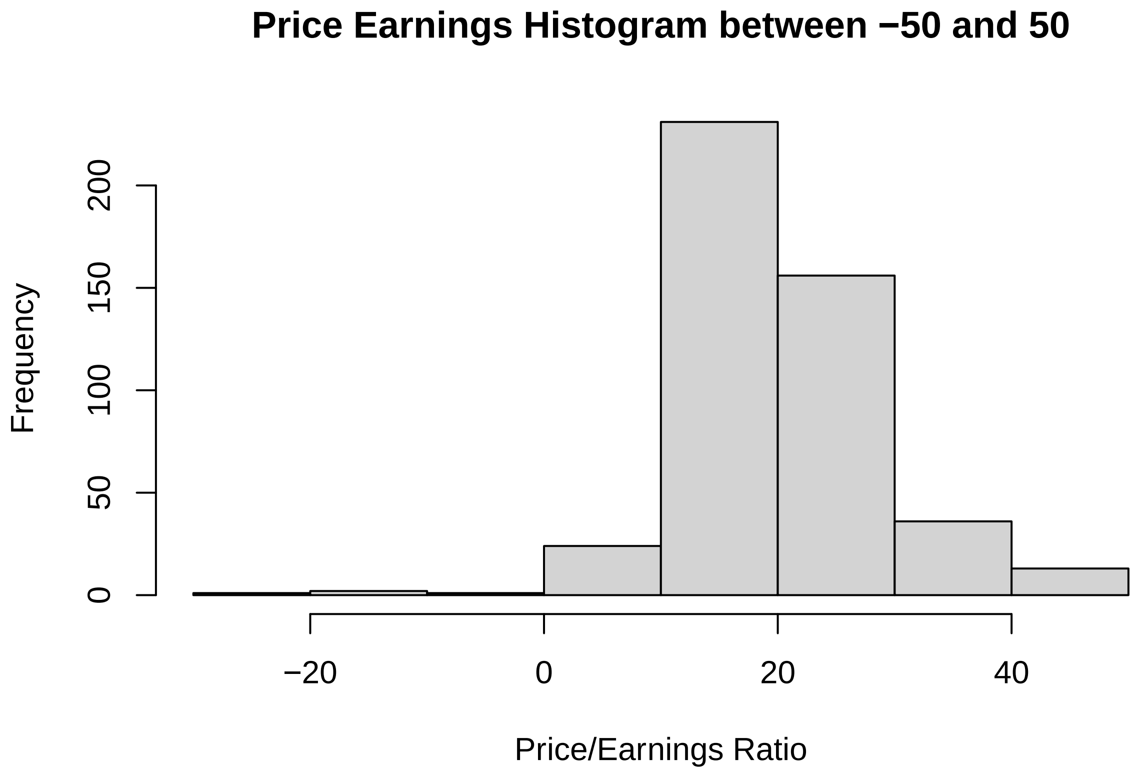
```r
hist(cleanindex$Price.Earnings, main = "Price Earnings Histogram",
    xlab = "Price/Earnings Ratio")
```



```r
boxplot(cleanindex$Price.Earnings)
```



```
step <- filter(cleanindex, cleanindex$Price.Earnings > -50)
neg50plus50 <- filter(step, step$Price.Earnings < 50)
hist(neg50plus50$Price.Earnings, main = "Price Earnings Histogram between -50 and 50",
    xlab = "Price/Earnings Ratio")
```



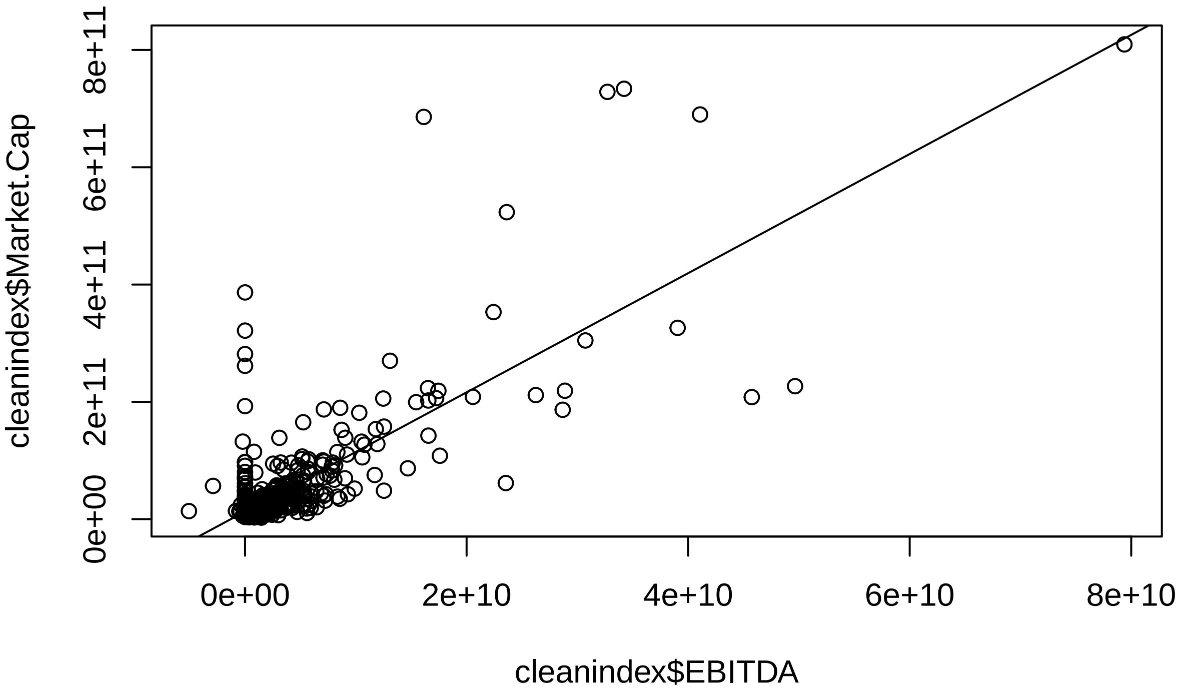
Finally, the histogram of Price Earnings Ratios appears unimodal and approximately symmetric with perhaps a slight skew to the right. Zooming in on the data to only include companies with ratios in between -50 and 50, we see a confirmation of an approximately normal distribution with a slightly larger righthand tail. Once again, this slight skew appears to be the result of negative ratios working differently than positive ones. The larger a company's loss (i.e. more negative EBITDA), then the closer to 0 the negative Price Earnings Ratio is despite actually being more "expensive" in terms of valuation. When examining the boxplot, outliers exist on both ends of the data, with this boxplot appearing more symmetrical than those of the other variables.

Regression

The final part of our analysis will include linear regression to examine the relationships of the variables.

Market Cap vs. EBITDA

```
plot(cleanindex$EBITDA, cleanindex$Market.Cap)
abline(lm(cleanindex$Market.Cap ~ cleanindex$EBITDA))
```



```
model.mcap_ebitda <- lm(cleanindex$Market.Cap ~ cleanindex$EBITDA)
summary(model.mcap_ebitda)
```

```
##
## Call:
## lm(formula = cleanindex$Market.Cap ~ cleanindex$EBITDA)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.907e+11 -1.543e+10 -9.434e+09  3.216e+09  5.090e+11
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.305e+10  2.931e+09   4.451 1.06e-05 ***
## cleanindex$EBITDA 1.016e+01  3.782e-01  26.857  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.784e+10 on 493 degrees of freedom
## Multiple R-squared:  0.594,  Adjusted R-squared:  0.5932
## F-statistic: 721.3 on 1 and 493 DF,  p-value: < 2.2e-16
```

```
cor(cleanindex$Market.Cap, cleanindex$EBITDA)
```

```
## [1] 0.7707216
```

First we examine the relationship between Market Cap and EBITDA. The scatterplot and line of best fit together indicate a positive linear relationship. Modeling the relationship and examining the summary reveals a p-value of 2e-16 for EBITDA, indicating a strong likelihood that the beta of EBITDA is non-zero, meaning EBITDA meaningfully contributes to the prediction of market cap. The correlation coefficient of 0.77 indicates a strong, direct correlation between market cap and EBITDA. Additionally, the adjusted R-squared value of .5932 means that approximately 59% of the variation in market cap can be accounted for by the variation in EBITDA. As we would intuitively expect, an increase in EBITDA (earnings) tends to correspond with an increase in market cap. This generally means an increase in share price and corresponding profits for investors in the form of dividends or capital gains. Next, we will incorporate the Price/Earnings Ratios into the regression model.

Market Cap vs. EBITDA & Price Earnings Ratio

```
model.mcap_pe_ebitda <- lm(cleanindex$Market.Cap ~ cleanindex$EBITDA +
    cleanindex$Price.Earnings)
summary(model.mcap_pe_ebitda)
```

```
##
## Call:
## lm(formula = cleanindex$Market.Cap ~ cleanindex$EBITDA + cleanindex$Price.Earnings)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.870e+11 -1.485e+10 -9.233e+09  3.963e+09  4.293e+11
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.815e+09  3.257e+09   1.785   0.0749 .
## cleanindex$EBITDA         1.016e+01  3.704e-01  27.421  < 2e-16 ***
## cleanindex$Price.Earnings 2.935e+08  6.252e+07   4.695 3.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.665e+10 on 492 degrees of freedom
## Multiple R-squared:  0.6114, Adjusted R-squared:  0.6098
## F-statistic: 387.1 on 2 and 492 DF,  p-value: < 2.2e-16
```

Upon examination of the second model, we note that p-values for both EBITDA and Price/Earnings Ratios are well below .01, indicating that both variables have nonzero betas, and hence that both variables significantly contribute to the prediction of market cap. Taking the square root of the multiple R-squared value gives us the correlation coefficient with a value of .7819. Both this value and the Adjusted R-Squared of .6098 have slightly improved upon the first model, corroborating the idea that Price/Earnings ratios supplement EBITDA well when predicting market cap.

Conclusions

The analysis of the distributions of the three variables reveals that EBITDA and market cap data are skewed to the right with a concentration of data in the low end of the range and many outliers on the high end. This indicates that most companies have similarly low earnings and market values, with a few companies massively outperforming in terms of both earnings and value. On the other hand, price/earnings ratios appeared to be unimodal and distributed approximately symmetrically. This indicates that companies, while valued in a range of ways, tend to have valuations around a central point, with some deviation likely based on industry, profitability, and future growth prospects. When trying to predict market cap (or market value), the more salient variation appears to come from actual earnings (EBITDA). Combining EBITDA and price/earnings ratios yields the best predictive model for market cap. This intuitively makes sense; when you factor in how much a company is making along with what multiple the market values those earnings at, you get a pretty good predictor for the value of that company.

So why do we always seem to get it wrong? Why does a nearly fully automated index tend to outperform the smartest animal on the planet? Maybe we are not as objective and truth-seeking as we like to think. Stocks tell the stories of businesses and their interactions with the public eye of the markets. Humans like to hear and tell stories, and we often mix in our own biases and opinions, skewing the truth ever so slightly. As this skew filters down through multiple people via news media and word of mouth, perhaps the story is too distorted to make an accurate evaluation of the stock. The distortion can even show up in the stock price, as each investor contributes part of the story through buying and selling. We act with our emotions and intuition, rather than being purely logical; this leads to poor decision-making in investing, and can result in mispriced businesses and hand-selected portfolios that can't even outperform a simple sorted list (S&P500). Our analysis indicates that selecting individual stocks isn't so hard by virtue of the activity itself or "mysterious market forces", but rather because of our irrational nature. Additionally, the distributions we examined reveal that only a minority of companies will enjoy a large portion of the earnings and market cap valuations, meaning there isn't a lot of room for irrationality when selecting individual stocks for a portfolio.

Most likely, utilization of other relevant financial statistics such as revenue, price/sales ratios, dividend yields, free cash flow yields, and more could yield an even better model for predicting a stock's performance. Jim Simons' Renaissance Capital has arguably cracked the code to the market, generating an astounding 62% annualized return from its founding in 1988 through 2021. Without closely-kept secrets and limited investment into the fund, Simons' self-described mathematical approach would likely become the new norm for investing, forever changing the way companies are valued, and heavily diluting the amazing returns across all investors. The code is nowhere near cracking in this paper, but perhaps a more mathematical approach to investing could benefit everyone. This analysis is a good start.