UNDERSTANDING PORTFOLIO THEORY WITH DATA SCIENCE AND NATURAL LANGUAGE PROCESSING

ARUN S BHARADWAJ, SPRINGBOARD DATA SCIENCE INTENSIVE CAPSTONE PROJECT

INTRODUCTION

The goal of the Springboard Data Science Intensive capstone project is to understand portfolio theory using data science, natural language processing and Python programming. The intended audience for this project are hedge fund managers, academics and students of portfolio theory. The project is divided into two parts. The first part uses fundamental company data collected from StockPup.com to develop a relationship between stock price and predictor variables. The StockPup data contains 739 csv files with quarterly financial statement information from 1993 to 2017 for 739 companies listed in the US stock market. Each csv file was separately collected by downloading from the StockPup website. The data is arranged with each row containing fundamental information for a company during a specific quarterly time-period. The data has more than 40 columns that include revenue, earnings, stockholder`s equity, number of shares, stock price and other metrics commonly seen in company filings. The second part of the project tries to develop a classification algorithm for finding an investor`s investing philosophy using the letters they write to clients and written research reports. This part of the project will use natural language processing.

SIGNIFICANCE OF THE PROJECT

This project tries to answer some important questions in the realm of investment finance. Investors, hedge fund managers and academics have always been interested in understanding the factors driving stock price growth. Fundamental, technical and macroeconomic data are the factors considered to be driving stock prices. Fundamental data refers to financial statement information released by companies. Technical data refers to data collected from stock price movement and trading volume. Macroeconomic data refers to national statistics like GDP, unemployment rate, inflation etc. In contrast, there is another school of thought which hypothesizes that stock price movements are a random function and cannot be predicted. The first part of this project tries to find if fundamental data can be used to predict stock prices. The second part of the project builds a classification algorithm that can identify a market participant`s investing philosophy based on their language corpora. Investors are divided based on their investing philosophies. There are growth investors, value investors, quantitative investors, macro investors etc. Financial markets across the world have thousands of investors. By building a simple classification algorithm, it is possible to identify the different types of investors across financial markets that use the english language as their medium of communication. Further, this information can be used to identify and automate the returns generated by different types of investors. This area of research is interesting for hedge fund managers, academics and students of portfolio theory.

For the sake of simplicity, the capstone project only tries to classify investors into bearish and bullish. Bearish investors have negative views about stocks while bullish investors have positive views. Bearish investors sell short while bullish investors go long.

WHY IS THIS PROBLEM STATEMENT IMPORTANT

The two parts of this problem statement constitute, at a very high level, some of the most pressing questions in investment finance, with respect to stock market investing.

Some academics claim that stock prices are a random function that have only a minor relationship with other economic variables. In contrast, many market participants and other academics claim that stock price movements can be predicted using company specific and macroeconomic variables. If stock prices are a random function, investors will be better off buying index funds since stock selection will be a futile exercise. If stock prices are a function of some underlying variables, stock picking is likely to be more profitable than buying an index fund.

The investor classification algorithm is significant since it allows future researchers to understand the effect of investing philosophy on the returns generated by investors. An algorithm, compared to a manual classification, allows us to classify the thousands of investors operating in stock markets in an efficient manner. This classification can then be used to identify the aggregate returns generated by different types of investors.

FOR WHOM IS THIS PROBLEM STATEMENT IMPORTANT

This problem statement is important for academics, hedge funds and students of portfolio theory. Academics are always interested in understanding whether humans can generate returns better than the S&P 500 index. If humans are better, academics would further like to understand the behavioral theory behind this outperformance. Hedge funds are also interested in these topics since an understanding of stock price movement will help them outperform the index and generate better returns. Today`s students are tomorrow`s academics and hedge fund managers. So, they are interested in this field of study.

DATA LOADING

All 739 csv files are loaded into Python Jupyter notebook using the glob.glob function. One of the problems with each csv file is that the name of the company is seen only in the filename. So, when using glob.glob, while all the data inside these csv files get imported, we will never be able to know the name of the actual company. To ensure that the name of the company is seen in each row of the csv file of the respective company, we use the assign function in conjunction with the pd.read_csv. Inside the assign function, the symbol attribute is made equal to os.path.basename of the respective file name. After all files are imported, the quarter end column is assigned a data type of datetime to ensure that future operations using timeseries data are performed easily. The assign function inputs companyname_quarterly_financial_data.csv into each row. Since we only want company name, the _quarterly_financial_data.csv is removed from each row using str.replace function.

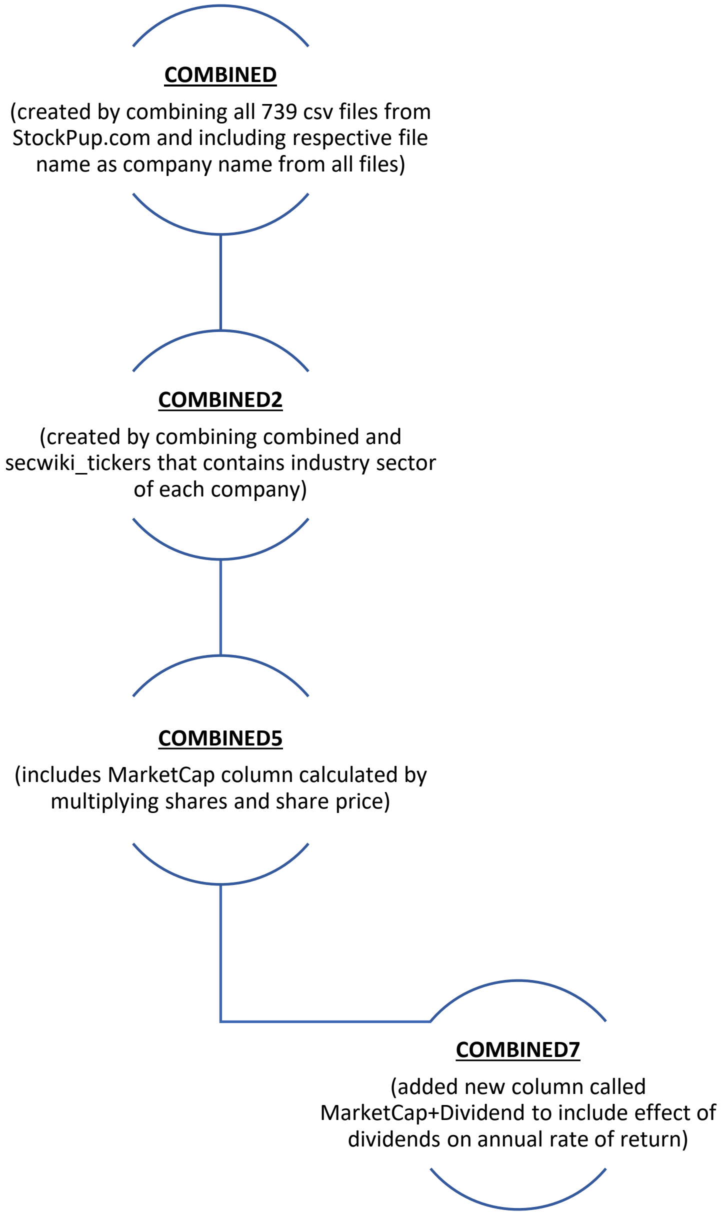
For the second part of the project, the nltk library is used to load the language corpora. Further, the correct encoding is used to import the text files.

DATA WRANGLING

In the first part of the project, since we also want information about the industry the respective company is operating in, the secwiki_tickers file is imported and included in the combined dataset. By running the info function on the combined dataset, it is seen that many columns are of data type, object. These columns are converted to numeric data type.

DATA STRUCTURE

This paragraph of the report allows a reader to understand and make sense of the Python code that has been written for this project. The link to the code is attached in the appendix. The variables in the code, which are mostly dataframes, are named based on their function in the project. For instance, main dataframes that are created by combining initial datasets are called combined. Data wrangling and feature extraction on the main combined dataframe creates combined 1, combined2 etc. From these dataframes, new dataframes are created, which are used in exploratory data analysis. These dataframes are named as eda1, eda1 etc. Below figure shows the description of some combined dataframes.



COMBINED

(created by combining all 739 csv files from StockPup.com and including respective file name as company name from all files)

COMBINED2

(created by combining combined and secwiki_tickers that contains industry sector of each company)

COMBINED5

(includes MarketCap column calculated by multiplying shares and share price)

COMBINED7

(added new column called MarketCap+Dividend to include effect of dividends on annual rate of return)

EXPLORATORY DATA ANALYSIS

MARKET CAPITALIZATION AND PRICE TO EARNINGS RATIO

The scale and complexity of the company fundamental data allows for a lot of exploratory analysis. Figure 1 shows the market capitalization of US markets from 1994 to present. The figure shows the current total market capitalization of all companies in the dataset to be $ 17.5 trillion. A simple internet search shows the current total market capitalization of US stock markets as $ 25 trillion. This shows that the StockPup dataset of 739 companies accounts for 70% of the total market capitalization of US markets.

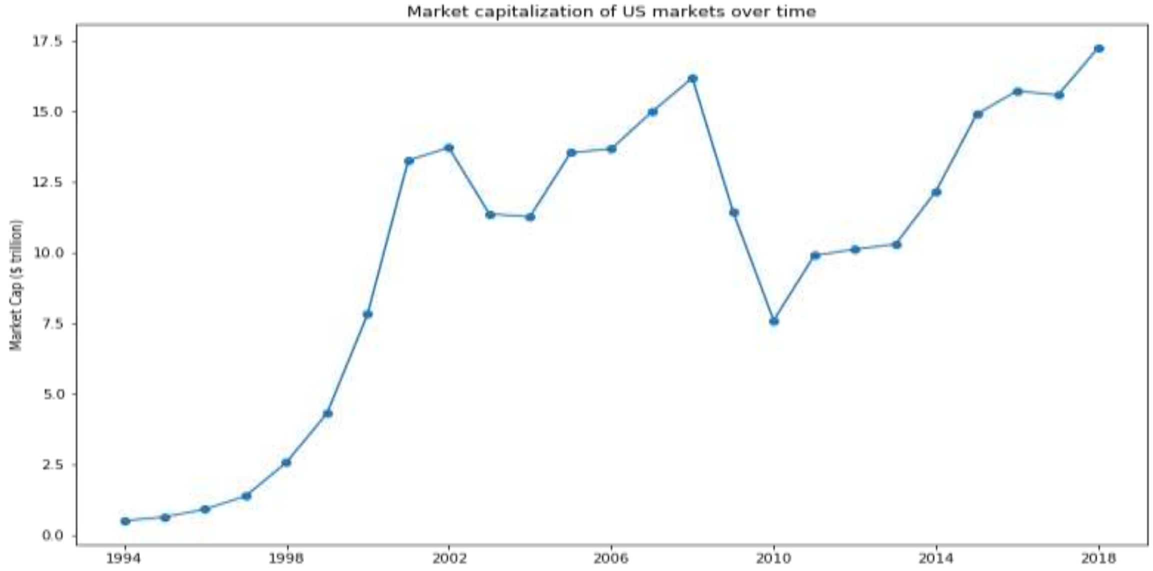


Figure 1

Figure 2 shows the number of different market capitalization companies over time. There are 4 different types of companies divided based on market capitalization: < $ 1 billion, $ 1-10 billion, $ 10-30 billion and > $ 30 billion. It is to be noted that a company`s sub-type may change from year to year, depending on its market capitalization. From the graph, it can be inferred that from 2010, the number of $ 1-10 billion market capitalization companies has halved while the number of $ 10-30 billion and > $ 30 billion companies have increased substantially.

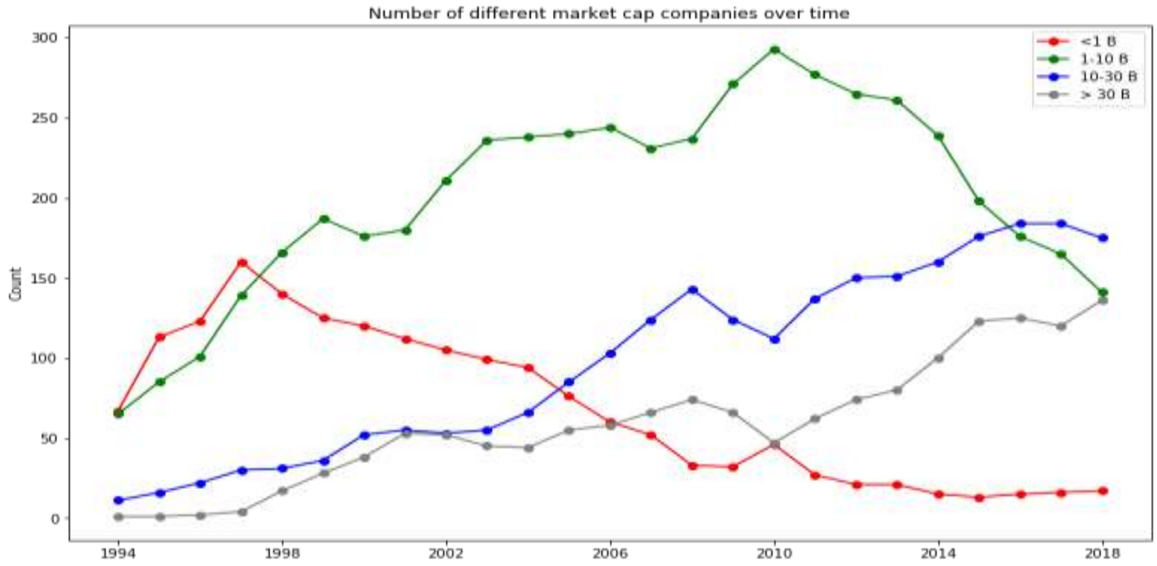


Figure 2

The P/E ratio stands for price to earnings ratio. Easiest way to interpret the P/E ratio is to think of it as the number of years needed to breakeven the investment in a company if the company expects to generate the same level of earnings into the future. So, a high P/E ratio implies a longer time to breakeven and hence a riskier investment. Figure 3 shows box plots of P/E ratios of different sized companies historically. It is seen that < $ 1 billion market capitalization companies have a slightly higher median P/E ratio compared to the other categories. The > $ 30 billion companies have lowest median P/E ratios of all sub-types. This makes intuitive sense since lower market capitalization companies have higher potential for appreciation compared to larger companies. Small companies have faster expected growth rates and hence higher P/E ratios.

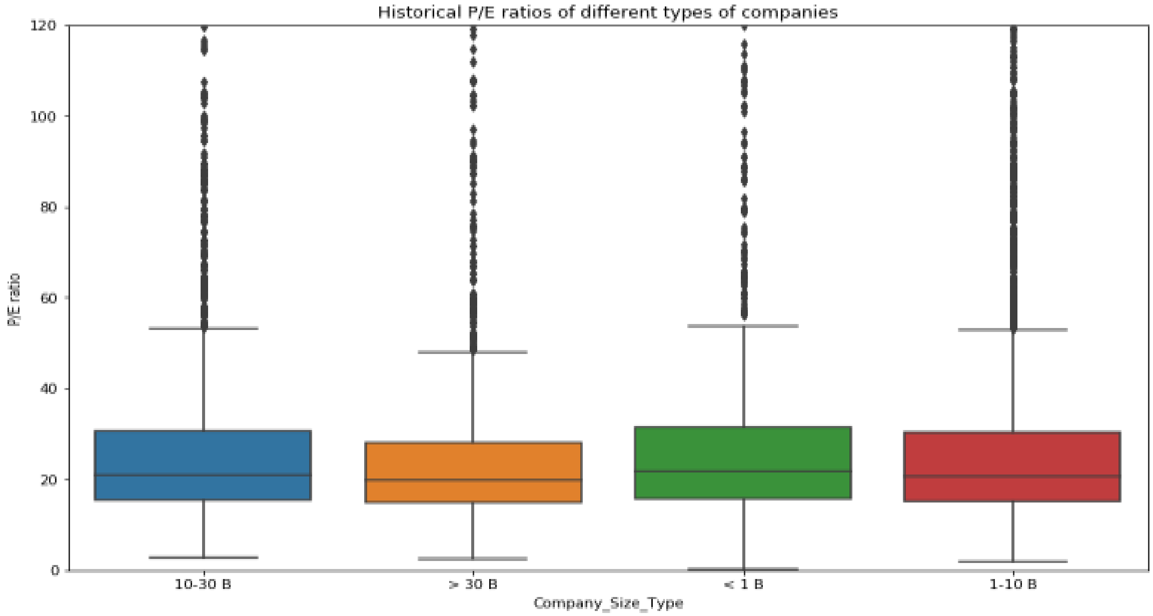


Figure 3

Figure 4 breaks down market capitalization of US markets into sub-types based on sectors. There are total of eight sectors: basic materials, consumer goods, financial, healthcare, industrial goods, services, technology and utilities. While six of these sectors have grown consistently over time, financials and technology have seen the largest increases in market capitalization at different points of time. From 1994 to 2008, financials saw the largest gains in market capitalization and was the largest sector based on market capitalization. The technology sector was mostly the second largest sector by market capitalization from the year 2000 and overtook financials in the year 2012 to become the largest sector by market capitalization. The utilities sector has seen the slowest growth in market capitalization over time.

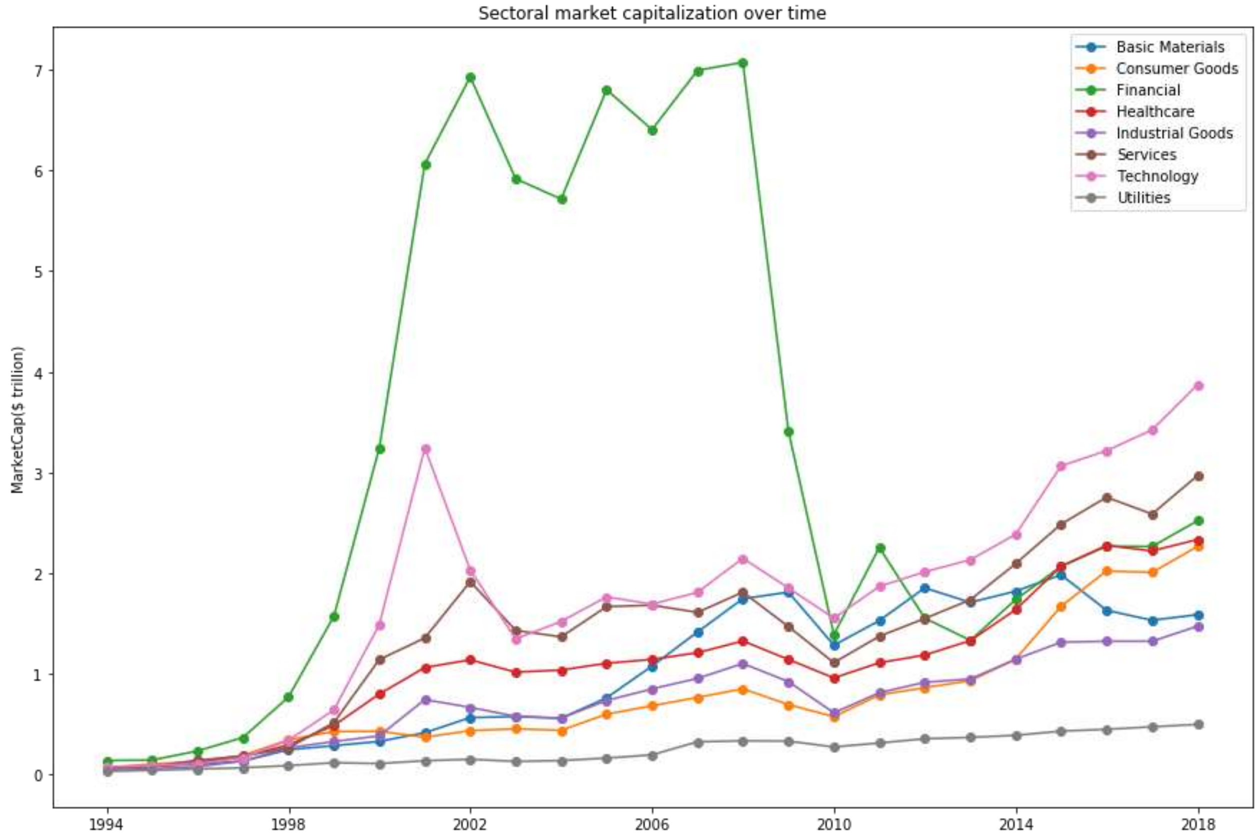


Figure 4

Figure 5 and figure 6 show the variation in sectoral P/E ratios over time. The sectoral P/E ratios were calculated differently compared to the P/E variation with respect to company size. The company size P/E ratio variation was plotted as a box plot in the previous page. The company size P/E ratio took individual P/E ratios of all companies at year end for all years. The sectoral P/E ratio for a given year was calculated by adding the market capitalization of all companies in that sector at year end and dividing that by the sum of earnings of all companies in that sector at year end.

Figure 5 and figure 6 have some large outlier values. To ensure that the reader discerns patterns in sectoral P/E ratios over time, the y-axis scale has been restricted. Negative P/E ratios occur when the whole sector makes a loss or has negative earnings. The areas where the plots extend upwards or downwards after end of the figure are outliers.

Figure 5 shows two large peak P/E ratios for the financials sector in the years 2002 and 2010. These were years that succeeded financial crises and these peaks were likely caused by large drops in earnings due to the crisis, which was not matched by a proportional drop in market capitalization. Figure 5 shows that the consumer goods sector has seen the lowest volatility in sectoral P/E ratios over time. Similarly, in Figure 6, the industrial goods sector has seen the lowest volatility in sectoral P/E ratios over time. Also, from Figure 6, it is seen that the sectoral P/E ratio of the technology sector was negative in the year 2002. This is after the dotcom bubble of 1999, when technology companies would have experienced losses or negative earnings.

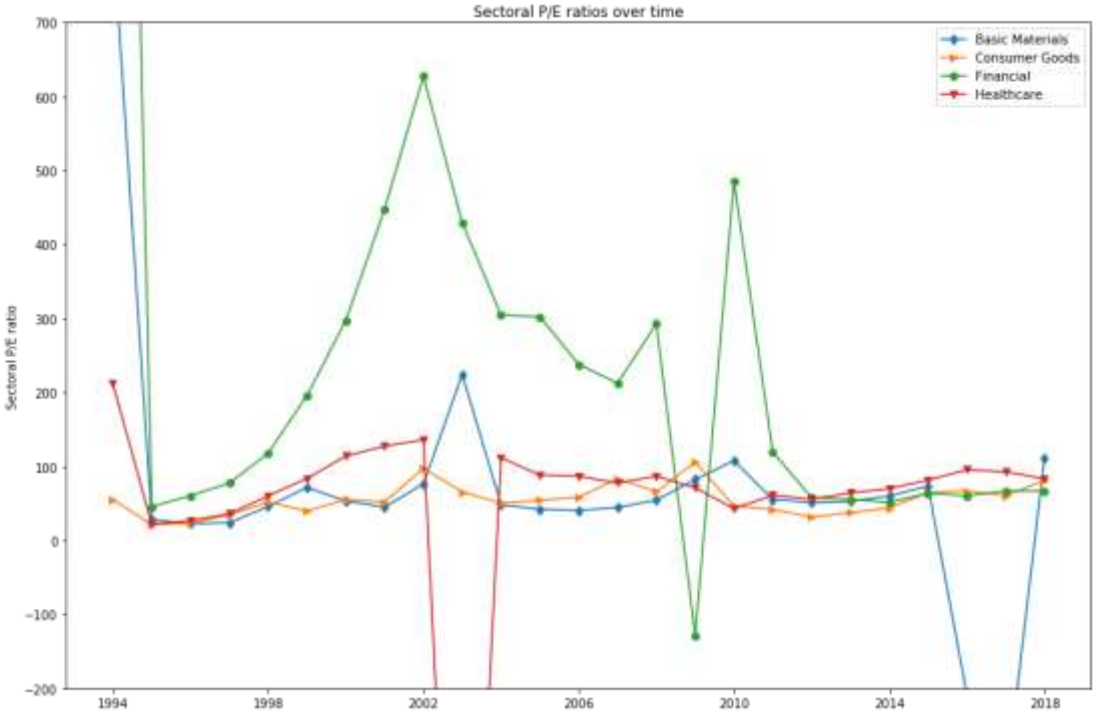


Figure 5

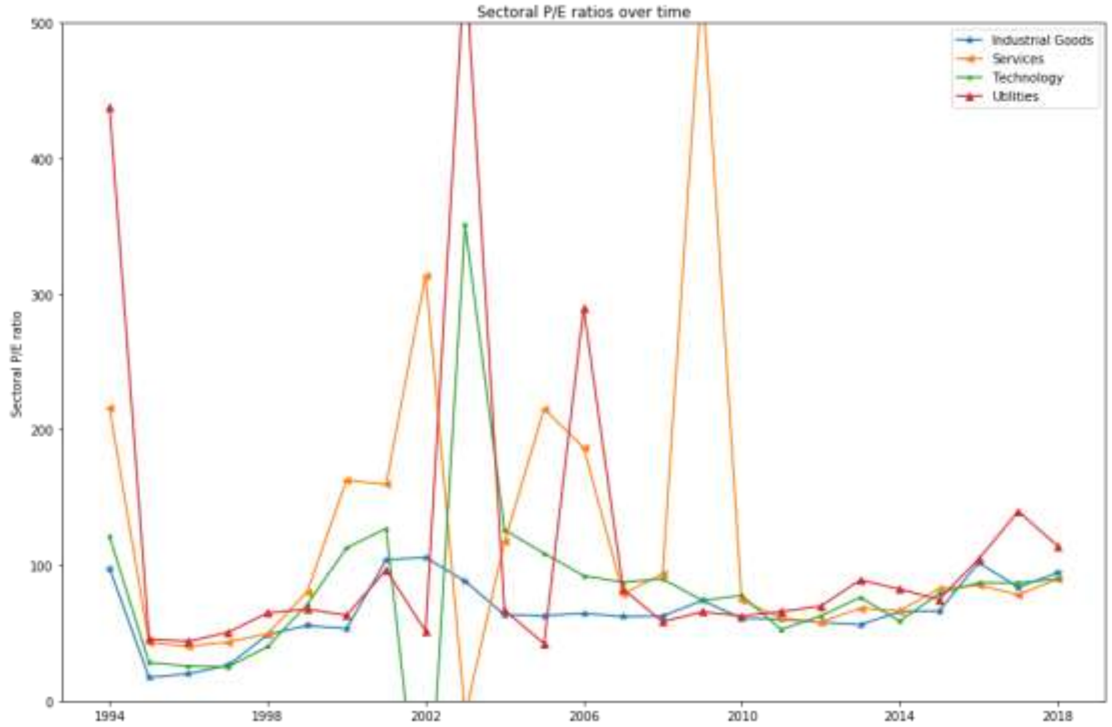


Figure 6

LONG TERM DEBT AND EARNINGS

Figure 7 shows the pattern of long term debt taken on by US companies over time. The graph shows a consistent increase in long term debt over time. It is improper to look at an absolute metric like long term debt since fast growing companies can afford to take on larger amounts of debt. So, figure 8 plots the ratio of earnings to long term debt for all US companies.

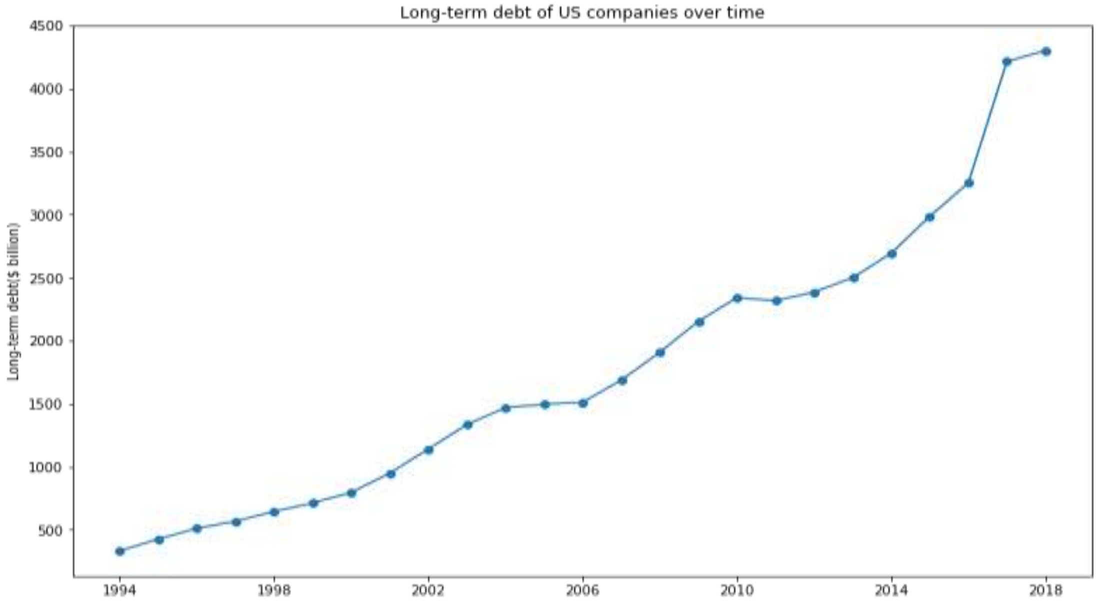


Figure 7

Figure 8 shows large dips in earnings to long term debt ratio after the 1999 tech bubble and after the 2008 financial crisis. This happened due to the sharp drops in earnings experienced by companies after these bubbles, while the amount of long term debt remained constant.

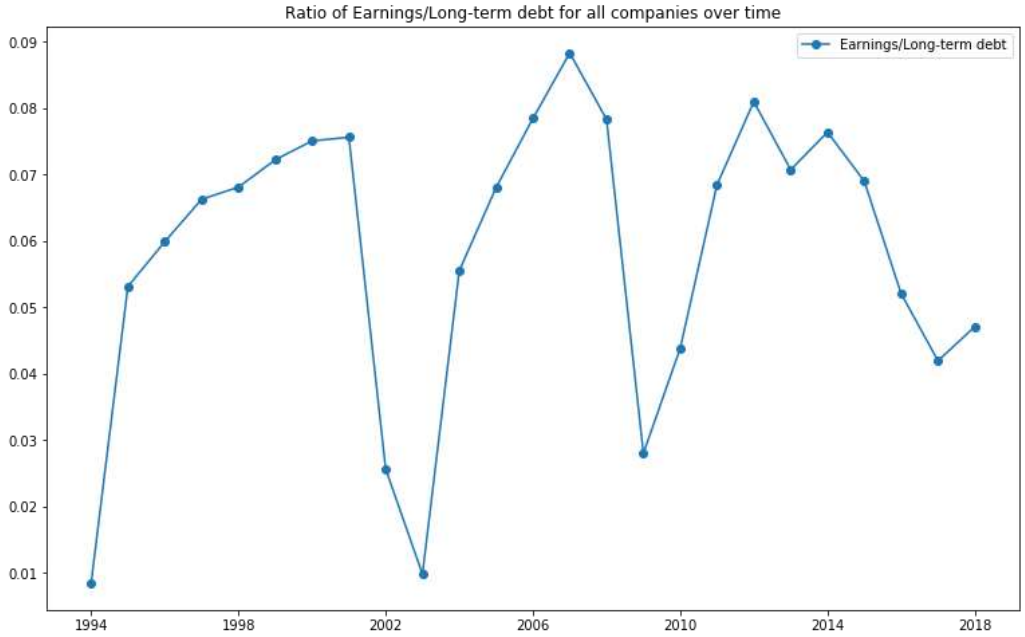


Figure 8

Figure 9 shows the long-term debt taken on by US companies, based on size. As expected, the bulk of long term debt has been taken on by companies that are > $ 30 billion market capitalization.

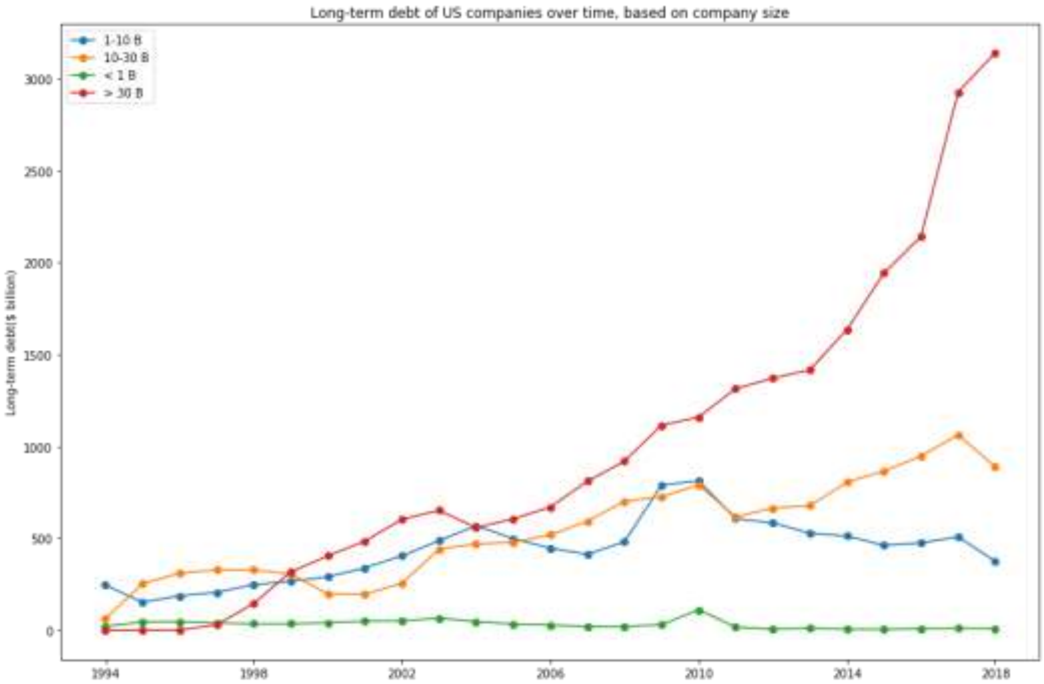


Figure 9

Figure 10 shows the financial sector to be the largest holder of long-term debt in 2017.

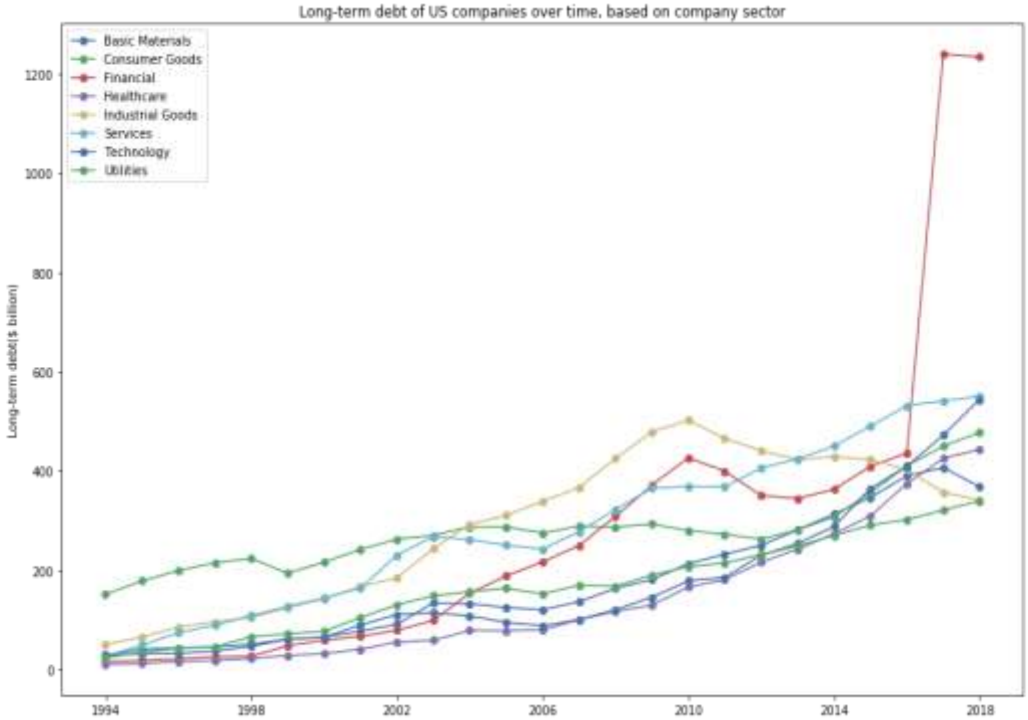


Figure 10

The ratio of earnings to long-term debt for different sized companies shows a very high degree of consistency over time, except for 2006 and 2014, when < $ 1 billion market capitalization companies saw large dips in their ratio.

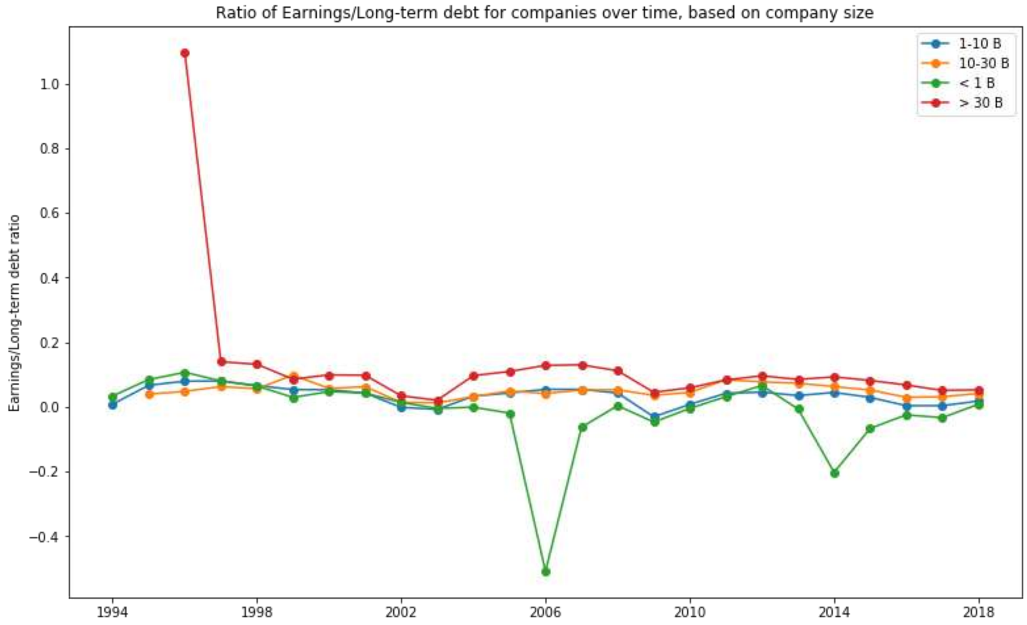


Figure 11

The ratio of earnings to long-term debt based on sectors is strikingly different from the previous graph. It shows a lot of volatility. Predictably, the ratio dips in the year 2002 for the technology sector and in 2009 for financials. In 2017, all sectors have earnings to long term debt ratio between 0 and 0.1.

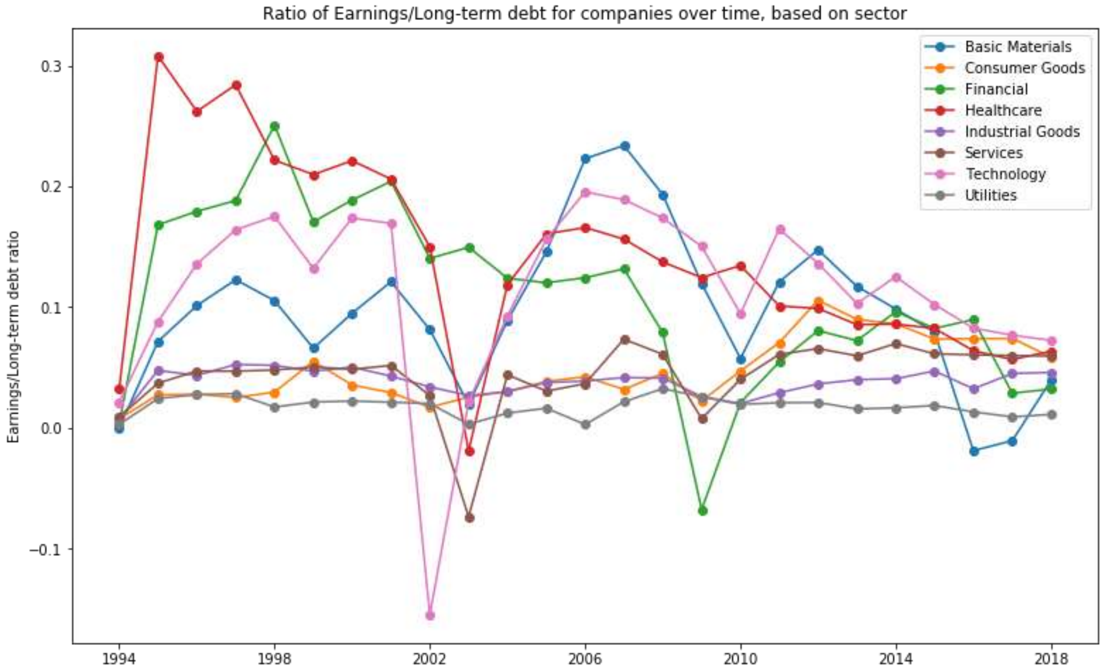


Figure 12

STOCK PRICE ANNUAL RATE OF RETURNS

Figure 13 shows the annual rate of return of all companies in the dataset. The annual rate of return is calculated using market capitalization and the cumulative dividends paid out for the year. Figure 14 gives some descriptive statistics about the annual rate of return.

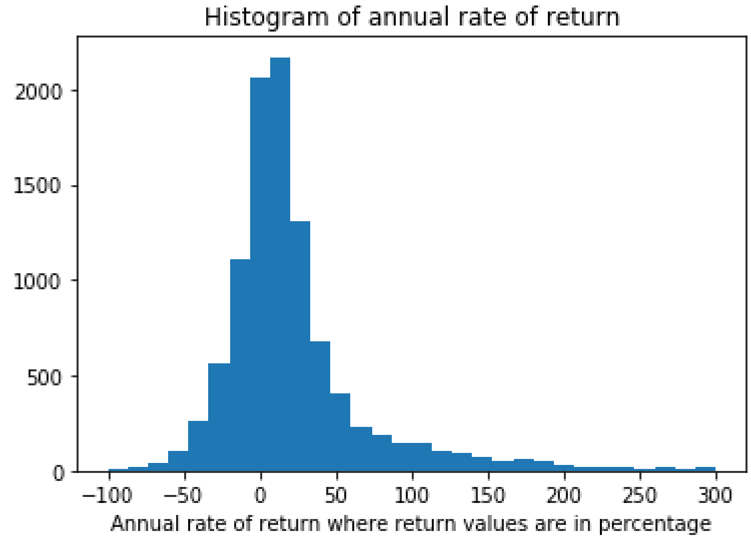


Figure 13

Mean	29.24%
Standard deviation	106%
Minimum	-99%
25%	-3.22%
Median	11.71%
75%	32.62%
Max	3926%

Figure 14

Figure 13 had its x-axis restricted to only include the range of -100% to 300%. The mean value of 29% is abnormally high since an average investor, investing in a randomly selected sample of companies, will never be able to achieve an annual return of 29%. This mean value is due to the presence of large positive outliers. The median value of 11% is more reasonable since an average investor will be able to achieve this annual return when investing in a randomly selected group of companies.

Figure 15, figure 17 and figure 19 are annual rate of return histograms split based on time periods. The three time periods that were used are 1993-2001, 2002-2009 and 2010-2017. These time periods were chosen since they are of nearly equal length and because they signify three different eras in the evolution of US stock markets.

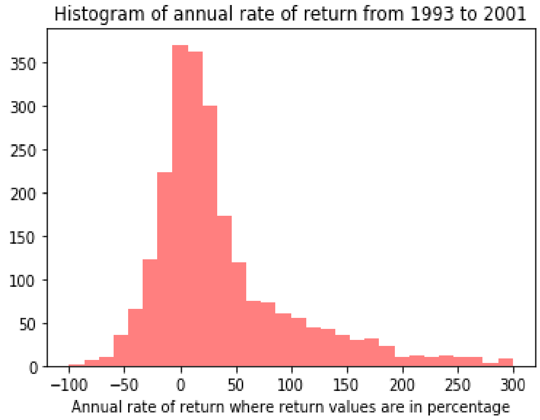


Figure 15

	1993-2001	2002-2009	2010-2017
Mean	60.44%	23.85%	13.46%
Standard deviation	184.79%	79.11%	37.51%
Minimum	-90.75%	-94.91%	-99.99%
25%	-0.74%	-8.61%	-0.82%
Median	20.52%	9.29%	8.87%
75%	65.47%	32.8%	20.62%
Maximum	3926%	2544%	858%

Figure 16

Figure 16 describes the findings of the three histograms. The mean of annual rate of return has been falling consistently over time. This also corresponds to a drop in standard deviation or market volatility. The median annual rate of return in the 90`s was substantially higher compared to the 2010-2017 time period. From the histogram, we see that the returns for the 1993-2001 time period has a short peak and the returns are heavily scattered around the mean of the distribution. For the 2002-2009 time period, the peak is taller and the scattering of returns are smaller compared to the earlier period. The 2010-2017 time period has the tallest peak and the lowest scattering of returns around the mean.

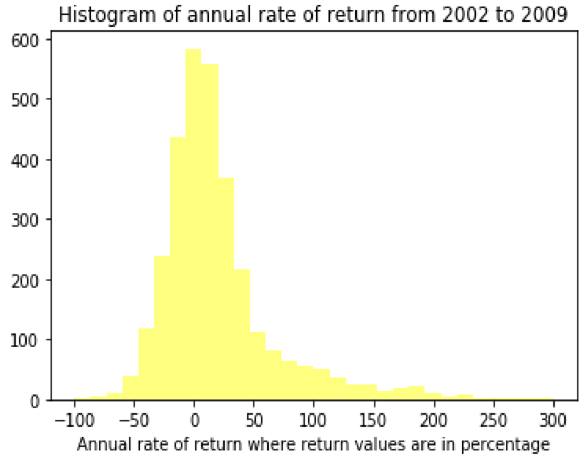


Figure 17

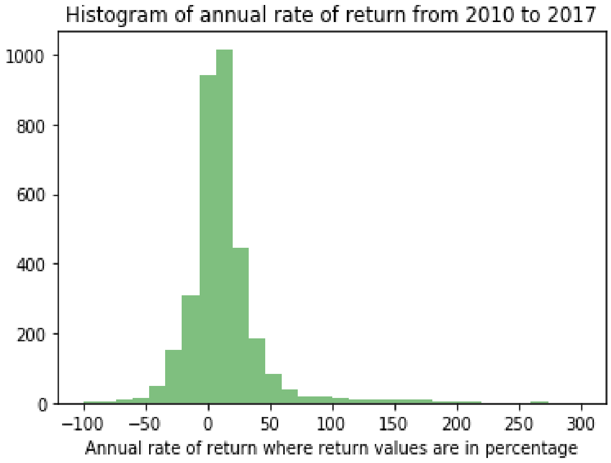
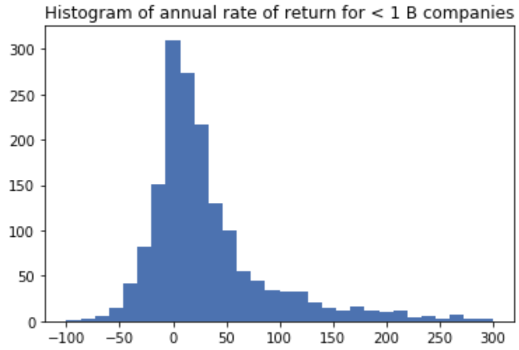


Figure 18

The annual rate of return for the whole time period was also calculated based on the size of companies. As expected, based on the below table, the < $ 1 billion market capitalization companies have the highest median and standard deviation of annual rate of return. While the highest median shows that small companies have the highest expected growth rates, the highest standard deviation or volatility shows that investors perceive investment in small companies as riskier than larger companies. The mean annual rate of return constantly decreases as the size of the company increases, except for > $ 30 billion market capitalization companies, which have a higher mean annual rate of return compared to $ 10-30 billion market capitalization companies. But the trend is maintained in the median annual rate of return metrics, which are inversely correlated with size of the companies.

	< $ 1 billion	**$ 1 – 10 billion**	**$ 10-30 billion**	**>$ 30 billion**
Mean	43.04%	18.98%	8.99%	11.66%
Standard deviation	108.71%	45.08%	23.16%	43.97%
Minimum	-90.03%	-76.36%	-55.7%	-82.57%
25%	-0.74%	-3.86%	-3.27%	-2.9%
Median	17.69%	10.51%	7.08%	6.18%
75%	50.25%	29.05%	17.94%	16.78%
Maximum	1988%	750%	352%	655%

Figure 19



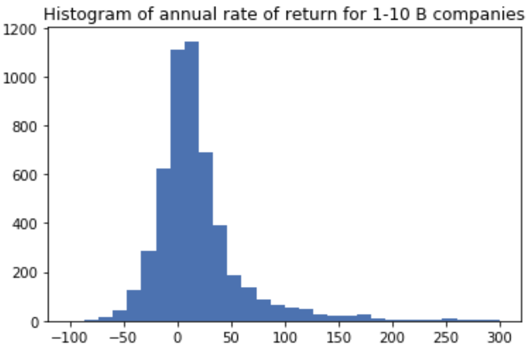
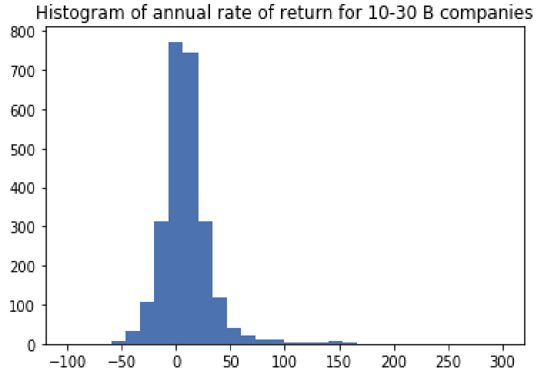


Figure 20

Figure21



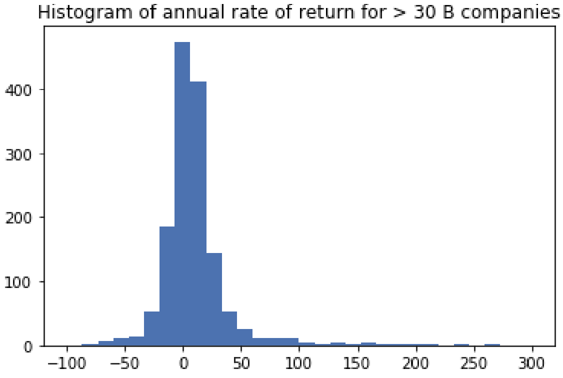


Figure 22

Figure 23

Figure 24 shows the summary statistics of annual rates of return for companies in different sectors. The utilities sector had the lowest median annual rate of return, while financials, healthcare and industrial goods had the highest median annual rate of returns. The highest volatility in annual rate of return was seen in the technology sector, followed by the services sector.

	Basic materials	**Consumer goods**	**Financials**	**Healthcare**	**Industrial goods**	**Services**	**Technology**	**Utilities**
Mean	23.62%	25.75%	22.36%	35.4%	22.88%	27.94%	47.11%	18.31%
Standard deviation	58.84%	67.6%	50.81%	99.96%	46.39%	102.9%	194.1%	86.44%
Minimum	-94.6%	-59.97%	-94.91%	-70.24%	-71.78%	-87.6%	-99.9%	-83.5%
25%	-5.39%	-1.69%	-1.24%	-1.21%	-1.23%	-5.62%	-7.5%	1.86%
Median	11.04%	11.08%	13.14%	13.82%	13.62%	11.35%	11.83%	9.58%
75%	33.85%	30.4%	28.74%	41.86%	32.15%	33.72%	41.54%	17.62%
Maximum	774%	1246%	533%	1988%	358%	2544%	3926%	1594%

Figure 24

ANNUAL RATE OF RETURNS VERSUS STANDARD DEVIATION OF RETURNS

Figure 25 shows that the mean annual rate of returns increases with increasing standard deviation of returns. This implies that stocks that have higher annual rate of returns also have higher standard deviation of returns. The curve is very similar to the Markowitz efficiency frontier. The Markowitz efficiency frontier is a right facing parabola drawn between standard deviation of stock returns and actual returns. According to Markowitz theorem, the points that fall below the vertex or the turning point of the parabola are inefficient. Also, points that are far off from the frontier are inefficient. Each point corresponds to a stock and the points below the vertex have higher risk for a lower return compared to the points above. This is true also for points away from the frontier. The points above the vertex are considered efficient compared to their peers below the vertex since they generate better returns for the same level of risk. The relationship between standard deviation of returns and rate of returns is quadratic and non-linear. Our plot confirms this.

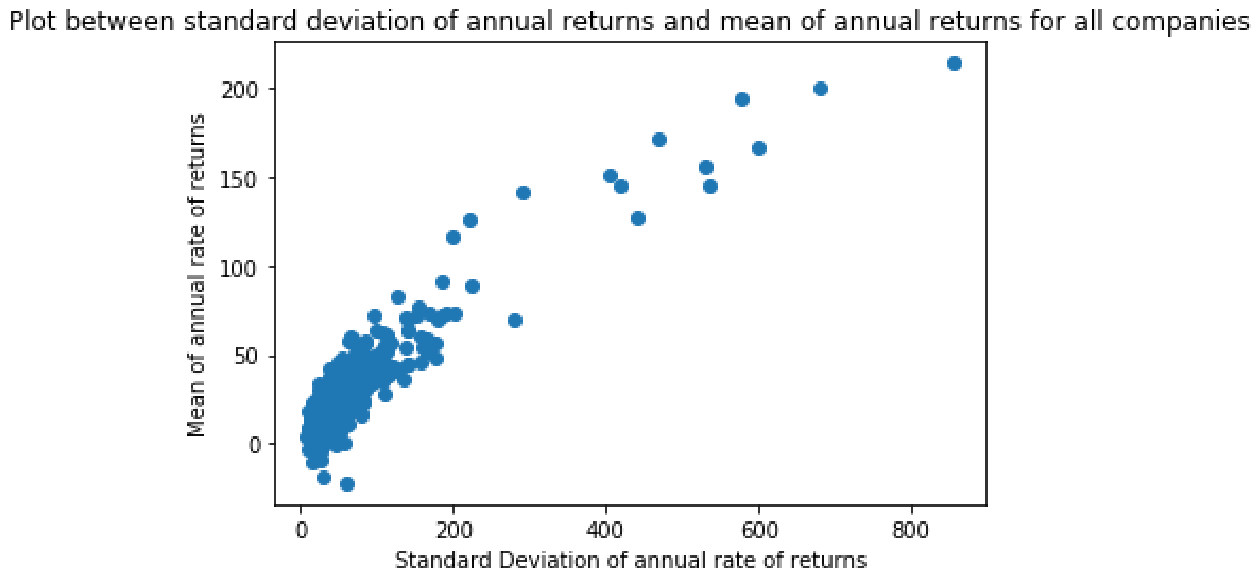
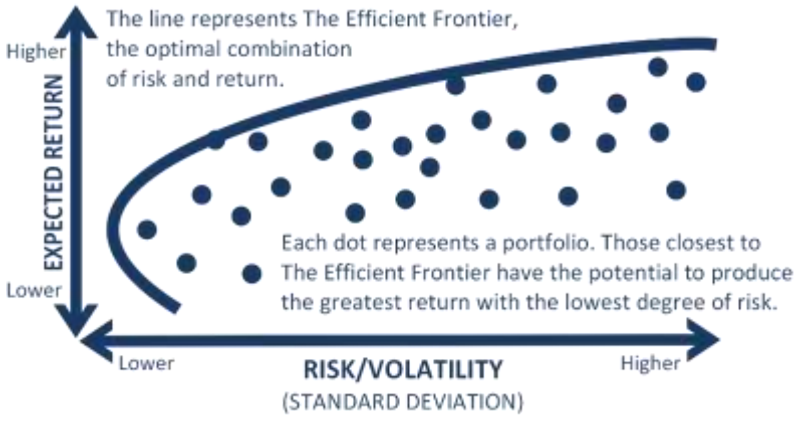


Figure 25


The line represents The Efficient Frontier, the optimal combination of risk and return.

Each dot represents a portfolio. Those closest to The Efficient Frontier have the potential to produce the greatest return with the lowest degree of risk.

Figure 26

Finally, our plot does not have any data points below the vertex. This means that all the 739 companies in our dataset are efficient, according to the Markowitz theorem. The very likely reason for this phenomenon is survivorship bias. The companies data downloaded from StockPup.com is probably composed mostly of successful companies whose share prices do well. An unbiased dataset composed of both successful and unsuccessful companies will likely produce a more complete version of the Markowitz frontier.