Project 7: unsupervised Learning

BY: LUKE HARRIS

Agenda

- Executive Summary
- EDA results
- Bivariate results
- K-Means Clustering
- Hierarchical Clustering

Executive Summary

The stock market combines the benefits of compound interest, tax incentives, and acting as a hedge against inflation to be one of the best places to store money for long periods of time. It is the most important tool to help Americans reach their financial goals later in life.

Diversification is one of the core fundamentals of a healthy stock portfolio. We will see in this presentation why putting all your eggs in one basket is not the correct way to navigate the market. While there may be greater reward in smaller, individual stocks, there is also great risk. A balanced portfolio takes risk, but mostly focuses on conservative picks like ETFs over individual stocks.

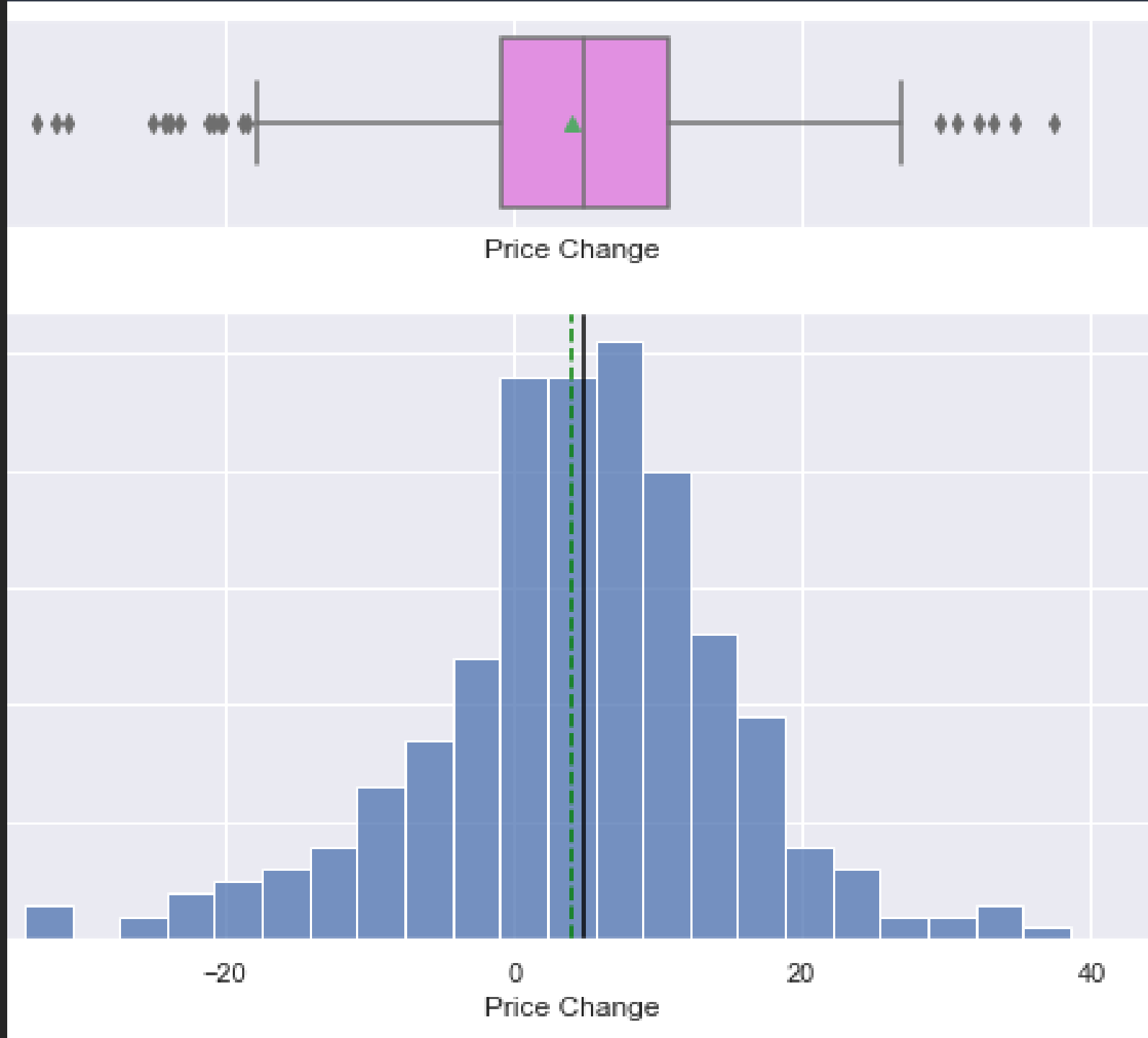
EDA Results

When looking at a stock it is important to look at how the company holds its money, where the cash is coming from, and what it is investing in. A company that stockpiles cash, invests it back into itself, and is conservative with taking on debt is usually a good place to invest in. Over the next few slides, I will explain a few graphs and show the important aspects to look at when deciding on a company to invest in.

Price Change

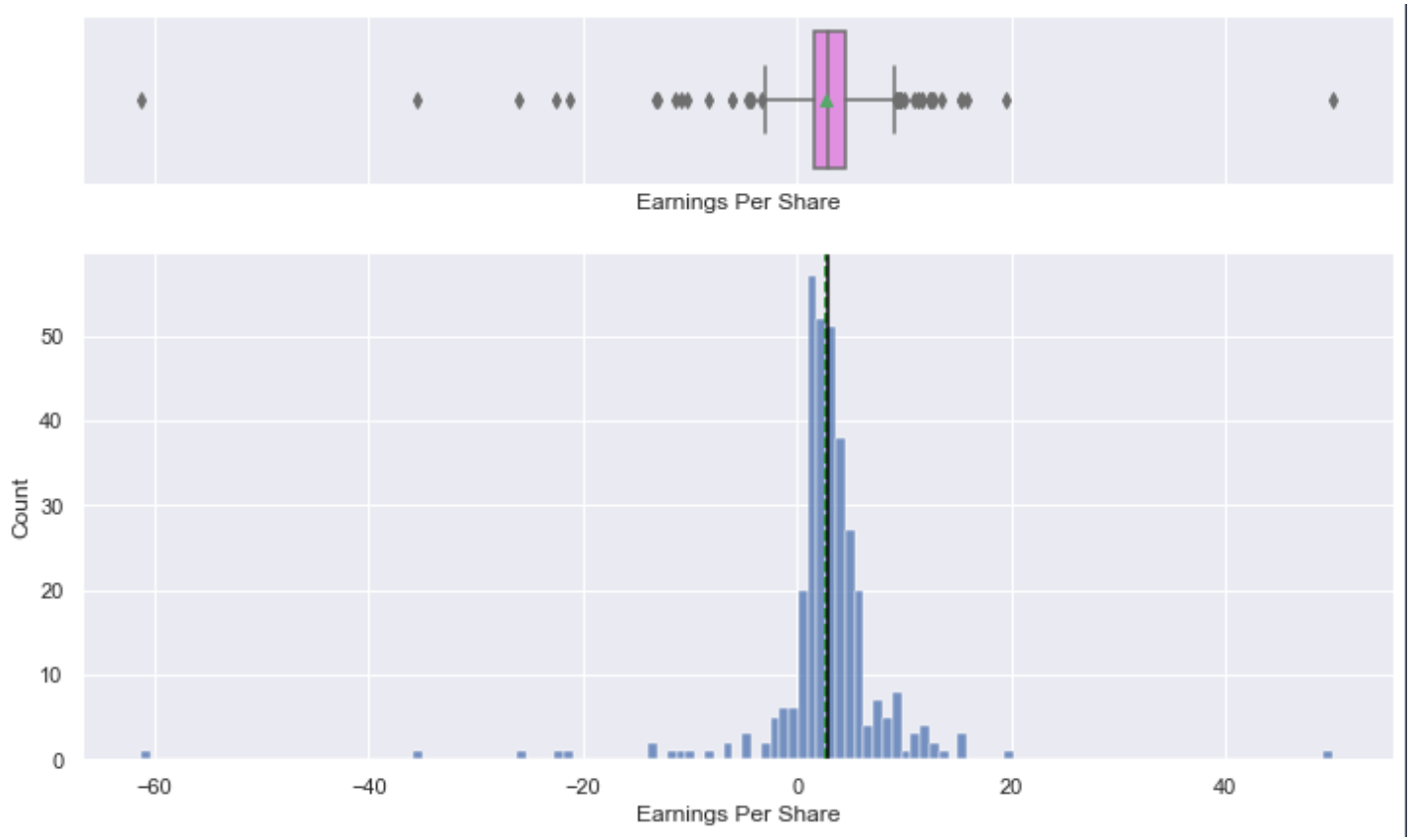
A GOOD RETURN ON INVESTMENT IN THE STOCK MARKET IS CONSIDERED 5-10%. AS SEEN HERE, MOST COMPANIES IN OUR DATASET FALL WITHIN THIS RANGE. LOOKING AT DIFFERENT TIME PERIODS AND FRAMES CAN CHANGE WHICH COMPANIES ARE STRONG.

RIGHT NOW, TECH HAS BEEN THE STRONGEST MARKET. BUT THROUGHOUT HISTORY THE STRONGEST SEGMENT AND WHERE THOSE COMPANIES ARE LOCATED HAS CHANGED. IT'S IMPORTANT TO BE AGILE AND CHANGE WITH THE TIMES IN THE STOCK MARKET



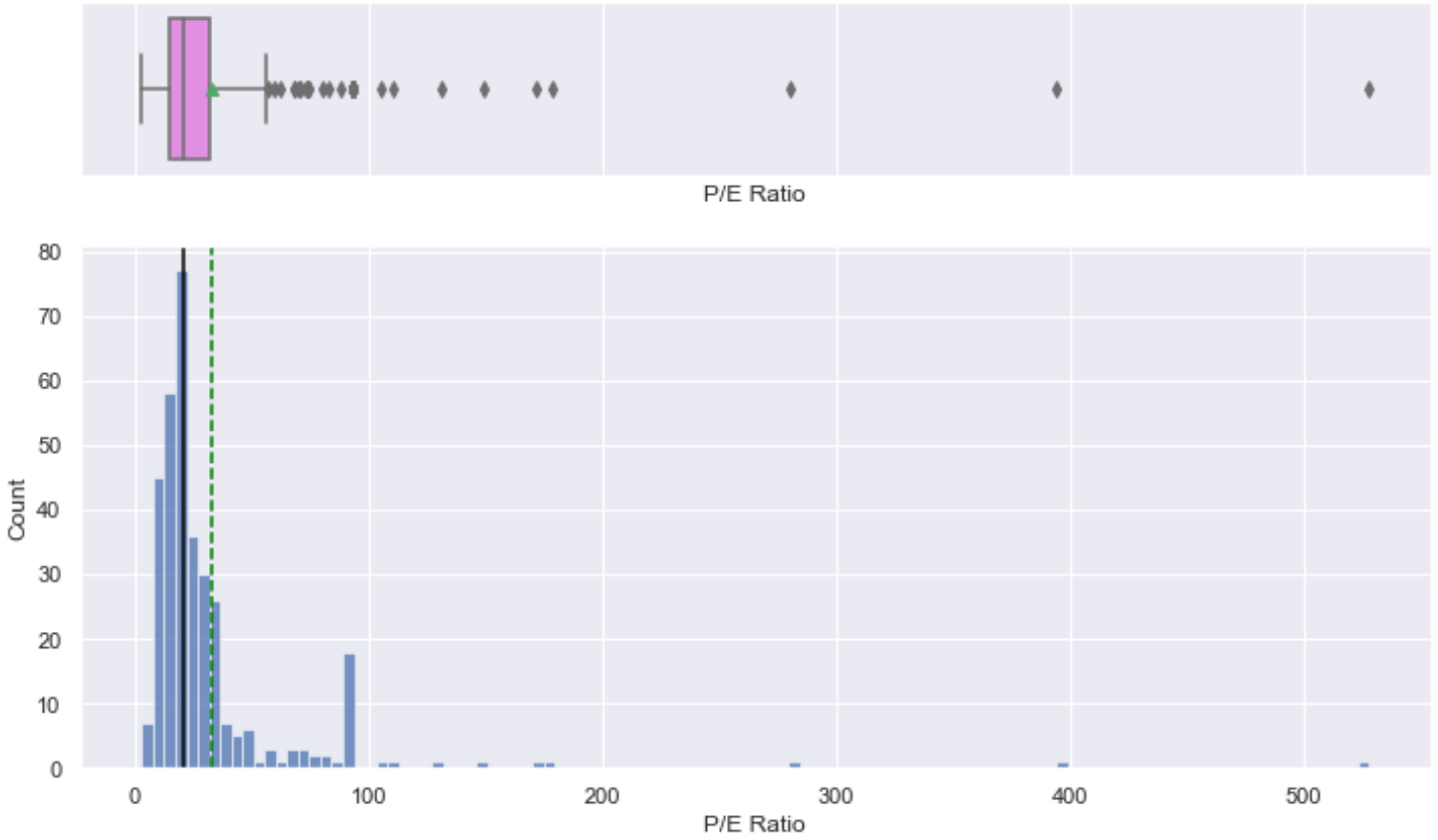
Earnings Per Share

Most individual people can't afford to buy hundreds or thousands of shares in large cap companies, so we need to look for the best bang for our buck. Which companies can give us the best earnings per each share we have? This is an important question and as we can see here, the higher the number the better. Most companies are slightly higher than 0 meaning that for each share we own, we are gaining a few dollars back, just by letting our money sit in the market!



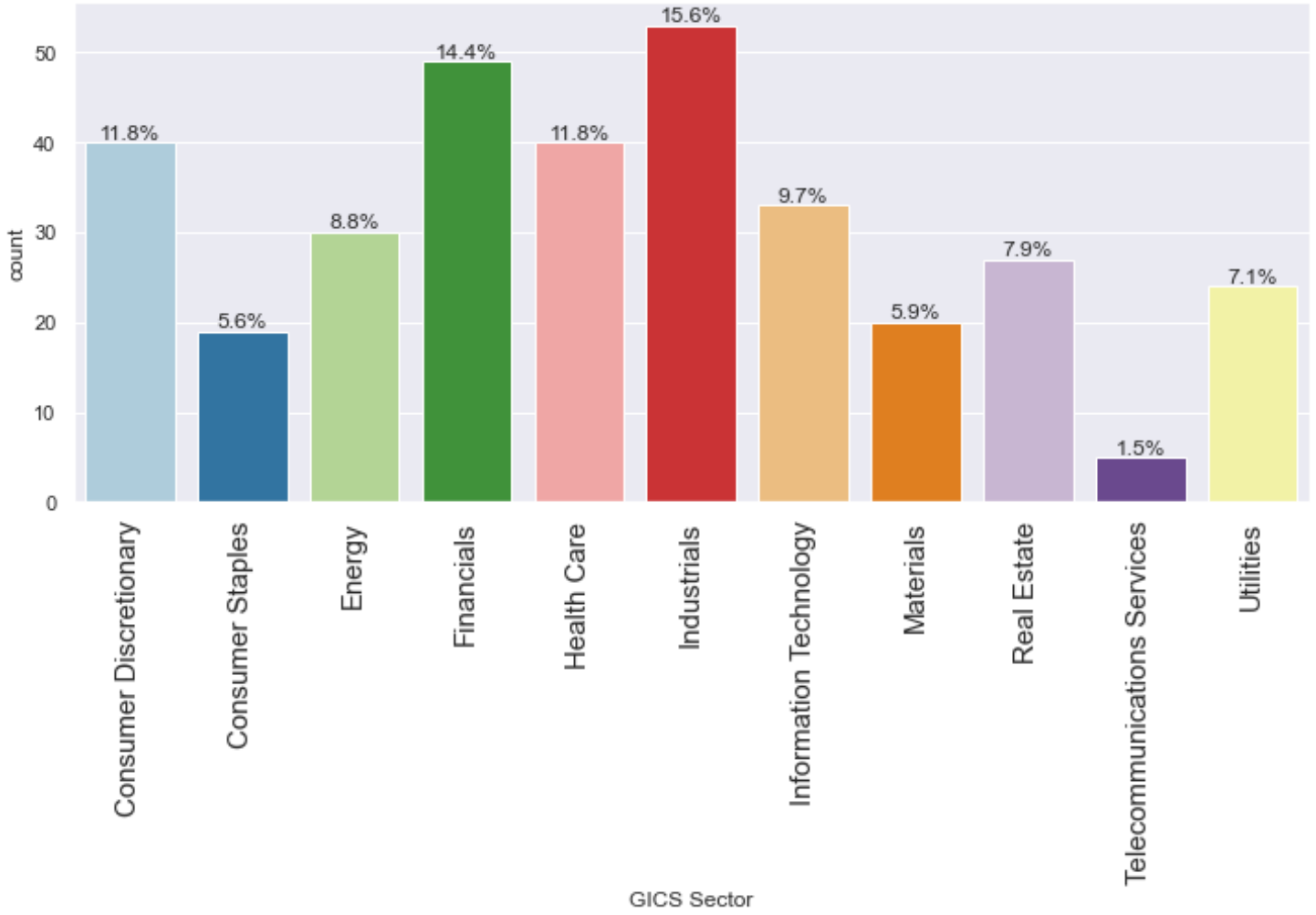
P/E Ratio

Finally, we have a very important number in the stock market, the P/E Ratio. This number tells us the company's current price, vs their earnings per share. The lower the P/E ratio, the better it is for the business and for potential investors. The way we can look at this number is seeing if the company is currently overvalued or undervalued. The more undervalued a company is the more enticing it becomes to invest it. There could be large potential gains for finding the diamond in the rough.



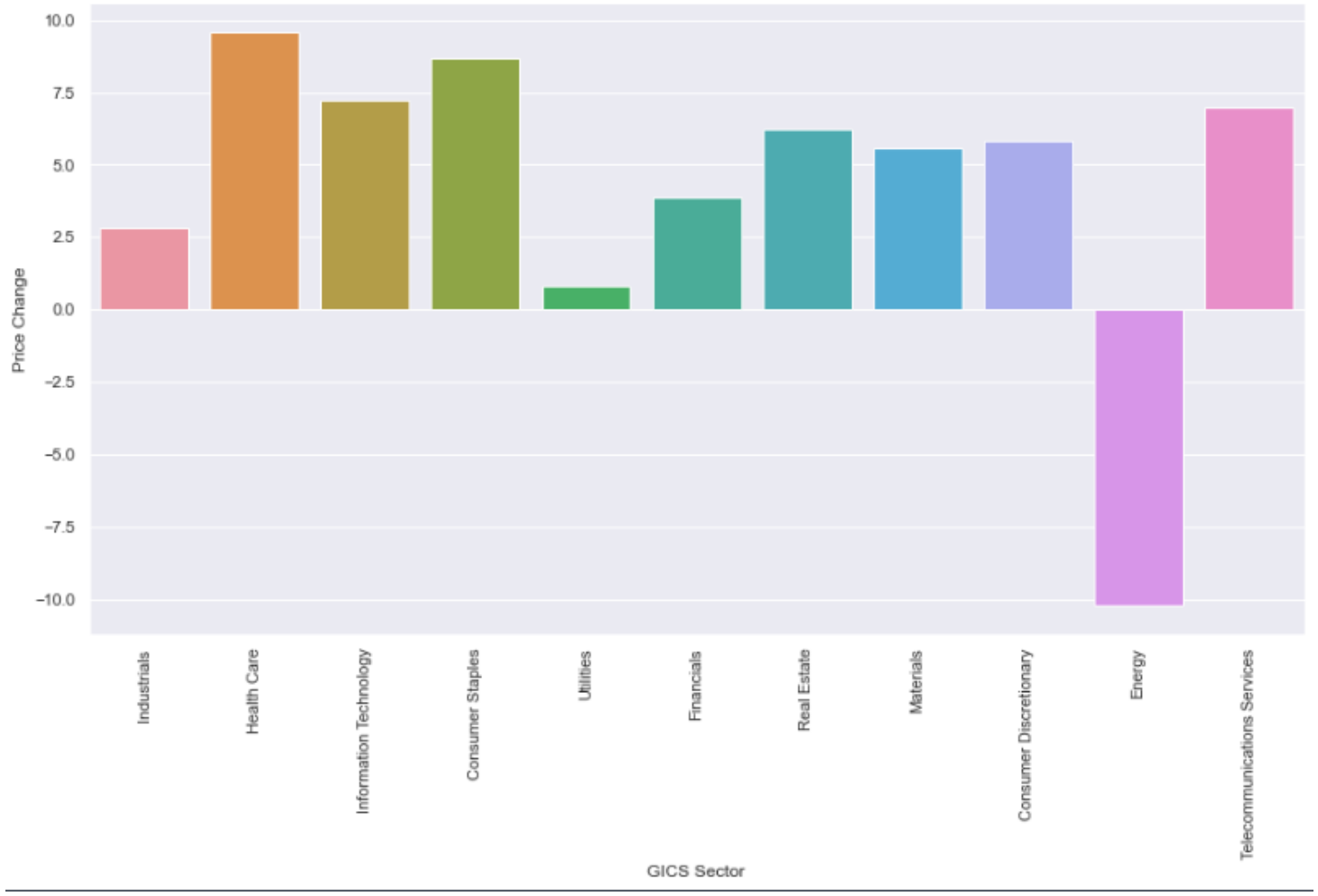
Bivariate Analysis

Before we get into the graphs that show the two variables against each other, let's look at the sectors we will be analyzing. As I stated before there are many different companies to invest in, and each of these companies can fall into a certain sector of the market.



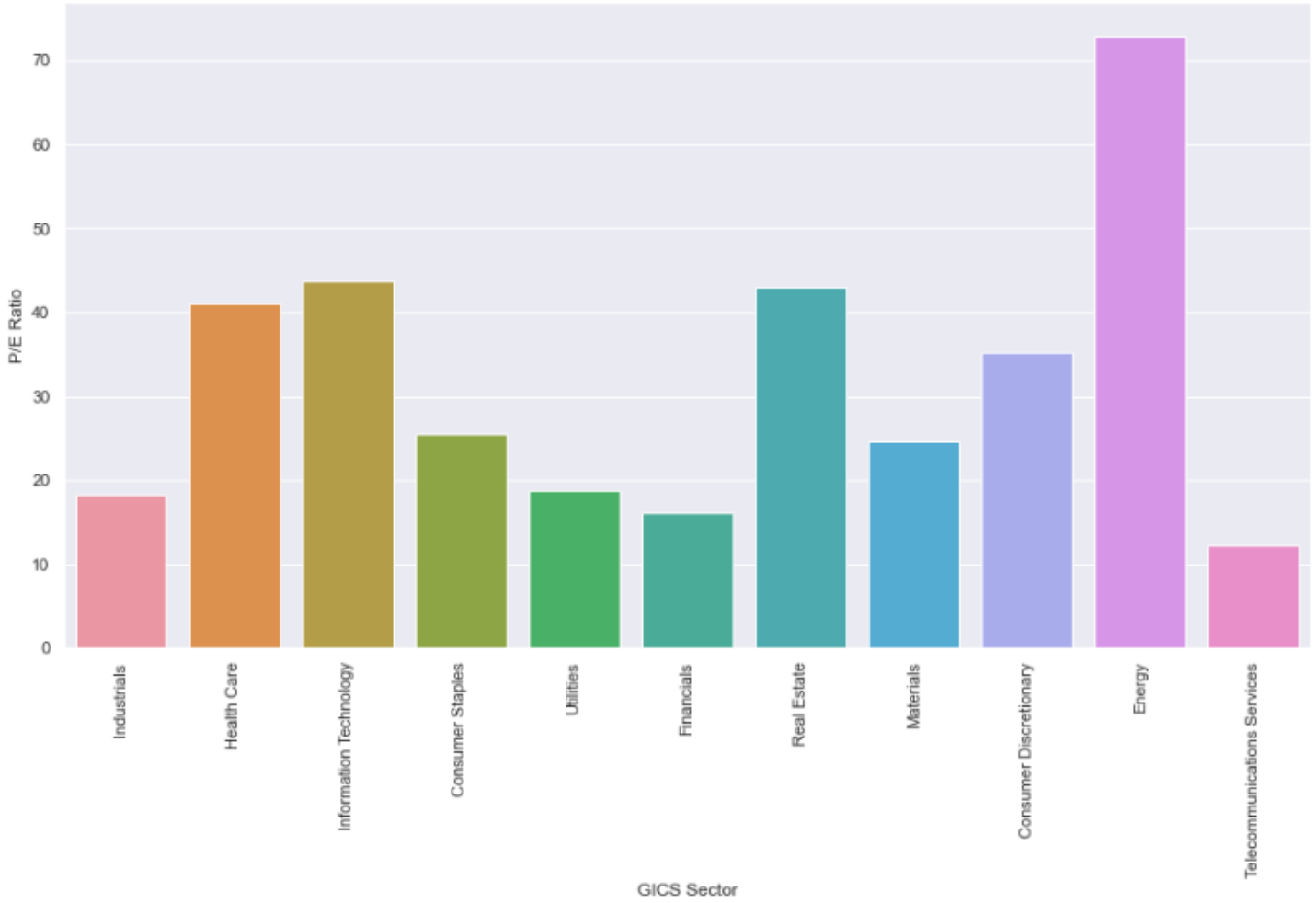
Price increase vs Sector

Here we can see that each, except one, has been performing above average. So, looking at this graph we would be inclined to put our money into health care, IT, consumer supplies. But one graph does not tell the full story.



P/E Ratio vs Sector

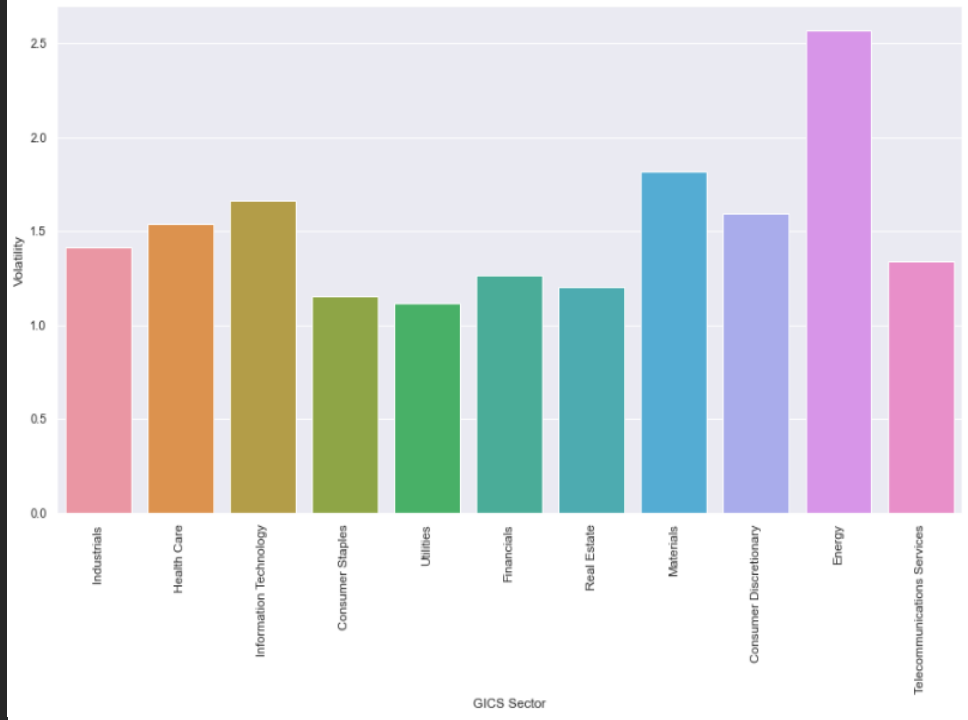
In the last graph, Energy looked very unnapealing, but here we see that when the market is down on a sector, it could present opportunity for undervalued companies. Energy is by far the most undervalued sector we find in our dataset which could potentially produce huge gains. We see that IT, and health care are also high up on undervalued, even if they also had the highest price increases, indicating that they are still safe bets in the today's market



Volatility of each Sector

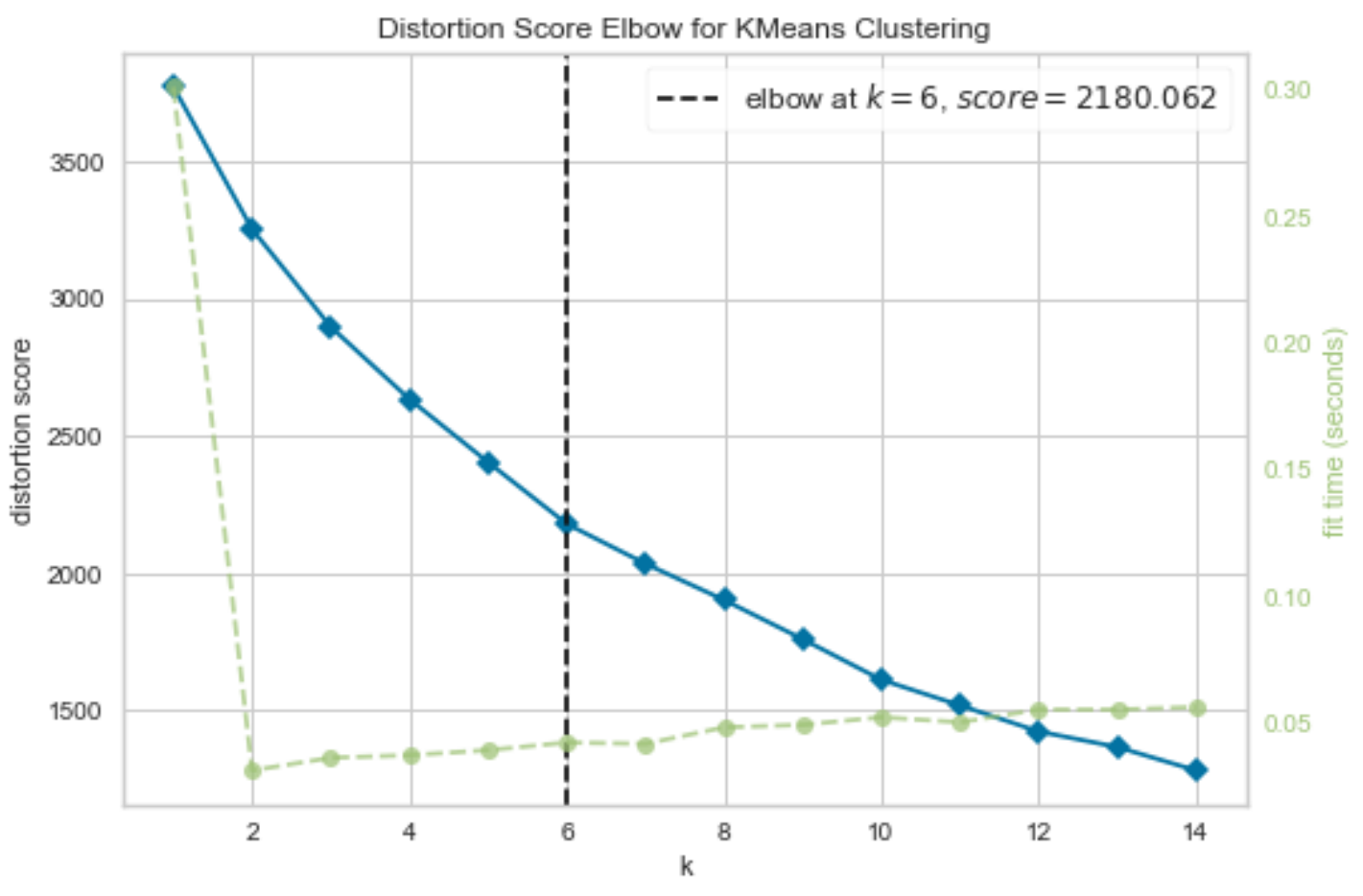
Volatility is a of an investors ability to stay calm in a storm. Some people are very conservative and want the most cautious and steady investment, guaranteeing the 5% increase per year. Other investors, especially recently, have had the “lambo or bust attitude”, looking for the most volatile stocks possible that could make or break their entire financial future.

The story that has been going on in this presentation has been focused on energy. To end the story, we see that they have had the worst price increases but are potentially highly undervalued. But this may be because many investors fear the volatility of this sector. It’s not good to not know how much the price will change from day to day. In the case of an emergency, you don’t want your investment to be down 10% because political reasons. Each person has their own risk tolerance and while there may be a lot of money to be made in the energy sector, it is much safer to stay in Health Care or IT sectors.



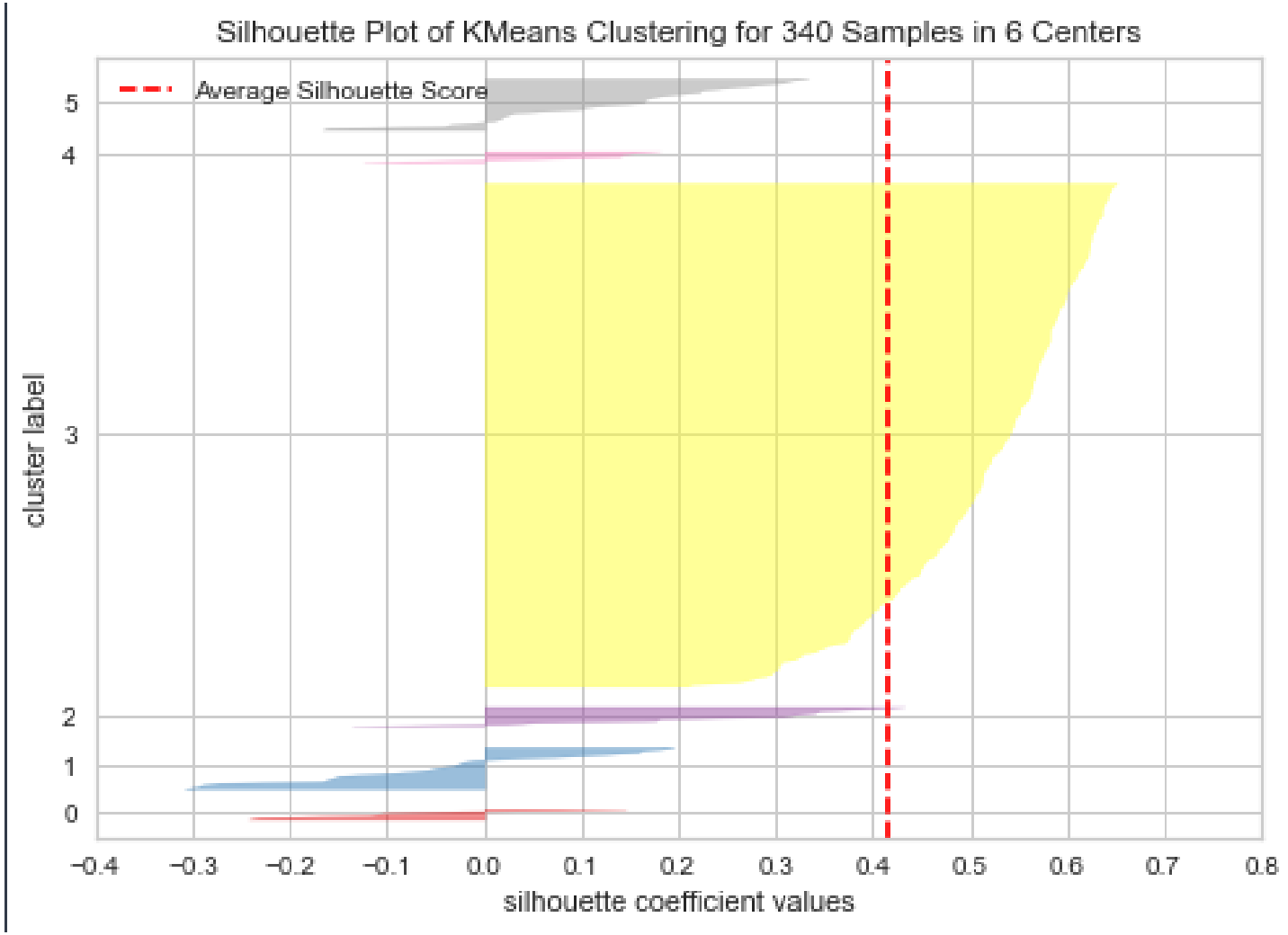
K-Means Clustering

K-Means clustering uses “centroids” to cluster together N number of observations belonging to the nearest centroid. When looking at this graph, we see that k = 6 is where the elbow is so that is what we will choose



Finding the Optimal Number of Clusters using Silhouettes

Using the Silhouette method, we can compute the coefficients of each point that measures how similar a cluster is to other clusters. Here is the results for the silhouette score when using 6 clusters as stated by our elbow point in the previous slide



HC_segments	GICS Sector	
0	Consumer Discretionary	2
	Consumer Staples	1
	Health Care	4
	Information Technology	3
	Real Estate	1
	Telecommunications Services	1
1	Consumer Discretionary	1
	Consumer Staples	2
	Energy	2
	Financials	1
	Industrials	1
2	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	3
	Health Care	1
	Telecommunications Services	2
3	Consumer Discretionary	35
	Consumer Staples	15
	Energy	7
	Financials	44
	Health Care	34
	Industrials	52
	Information Technology	27
	Materials	19
	Real Estate	26
	Telecommunications Services	2
	Utilities	24
4	Energy	20
	Information Technology	1
	Materials	1
5	Consumer Discretionary	1
	Health Care	1
	Information Technology	1
6	Financials	1
	Information Technology	1

Name: Security, dtype: int64

Hierarchical Clustering

This is the creation of creating clusters that have a predetermined ordering from top to bottom. The endpoint is a set of clusters that are distinct from other clusters. In this dataset, we look at the average linkage because it holds the highest cophenetic correlation. Here we can see how some of the companies in each sector were clustered together. As we see the largest cluster is the 4th (named 3), which has a lot of companies within similar se3ctors except energy and telecommunications, this is because mostly every other secotor was similar to one another except for those two, as seen earlier in the presentation

Each of the two techniques were very quick to execute on my machine, neither took too much processing power to complete. I K-Means was easier to read and understand on the graph, when dealing with this many datapoints, the hierarchical becomes bogged down with how large the tree grows to. They both were similar in how they clustered each sector when split into the same number of clusters. For K-means clustering, I used k=6 to determine the number of clusters, while hierarchical called for 7 separate clusters.

K-means vs Hierarchical

A Quick Thank You

I’m not sure who grades these, but I’d like to say thank you for taking the time and helping me grow in data science over the past 7 projects. This entire program has been very helpful for allowing me to break into a new industry and career. I hope to use this certification in the future and continue my endeavors in the data science world. I appreciate the time taken, especially by my mentor Shahin who has been amazing in all the lectures over the past weeks.