Trade&Ahead

Project 7 – UT DSBA Program

08/27/2022

Contents

Executive Summary

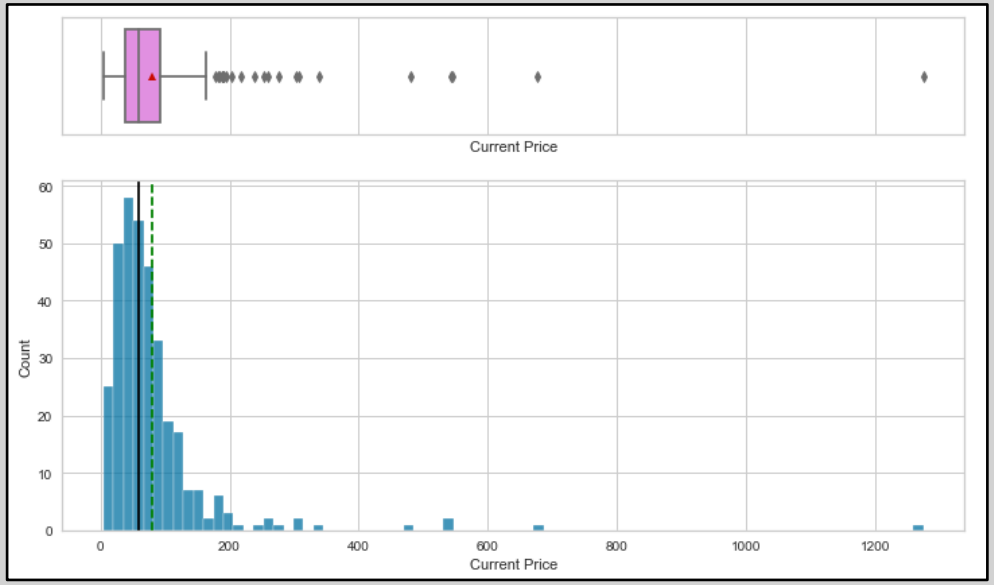
- Two models were created, one using K-Means Clustering and the other using Hierarchical Clustering.
    - The optimal number of clusters for each of the two models was 4.
    - The two models produced very similar results (see the appendix for more details).
- Each of the 4 clusters represents a group of securities. Trade&Ahead can consider assigning a specific investment strategy to each of these groups. For example, the clusters with the highest current price and price change (hierarchical cluster 2/k-means cluster 4) might be considered “high risk” and would be suitable for young investors. On the other hand, the clusters with the highest segment count (hierarchical cluster 4/k-means cluster 1) might be considered more “conservative” and would be suitable for older investors who have more to lose from the volatility of the market.
- Since the dataset was small (340 records), a more refined and accurate model could be created with a larger dataset. Trade&Ahead may want to revisit the exercise and create some additional models after it has the chance to collect some additional data. Perhaps, Trade&Ahead can set up a pipeline that feeds new data to and updates the models on a schedule.

Business Problem Overview and Solution Approach

- The stock market has consistently proven to be a good place to invest in and save for the future. There are a lot of compelling reasons to invest in stocks. It can help in fighting inflation, create wealth, and also provides some tax benefits. Good steady returns on investments over a long period of time can also grow a lot more than seems possible. Also, thanks to the power of compound interest, the earlier one starts investing, the larger the corpus one can have for retirement. Overall, investing in stocks can help meet life's financial aspirations.
- It is important to maintain a diversified portfolio when investing in stocks to maximize earnings under any market condition. Having a diversified portfolio tends to yield higher returns and face lower risk by tempering potential losses when the market is down. It is often easy to get lost in a sea of financial metrics to analyze while determining the worth of a stock, and doing the same for a multitude of stocks to identify the right picks for an individual can be a tedious task. By doing a cluster analysis, one can identify stocks that exhibit similar characteristics and ones that exhibit minimum correlation. This will help investors better analyze stocks across different market segments and help protect against risks that could make the portfolio vulnerable to losses.
- Trade&Ahead is a financial consultancy firm that provides its customers with personalized investment strategies. In order to help maximize returns for the customers, we will analyze the data, group the stocks based on certain attributes using multiple clustering techniques, and share insights about the characteristics of each grouping. This should help drive larger returns for the customer and will generate greater confidence in the company’s ability to assist in wealth creation.

Exploratory Data Analysis (Univariate)

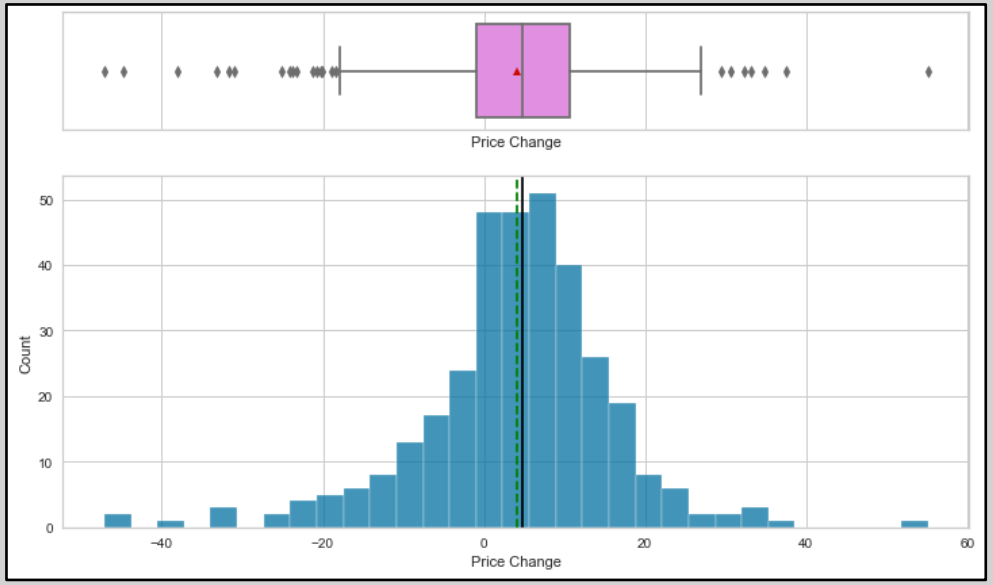
Distribution of Current Price



Observations

- The distribution of the Current Price follows a normal distribution for the most part but is slightly positively skewed.
- This indicates that the price for most stocks in the dataset fall between 0 and 200 while there are a few that go higher.
- There are not any stocks with a Current Price less than 0. This makes sense because a stock cannot have a negative price.

Distribution of Price Change

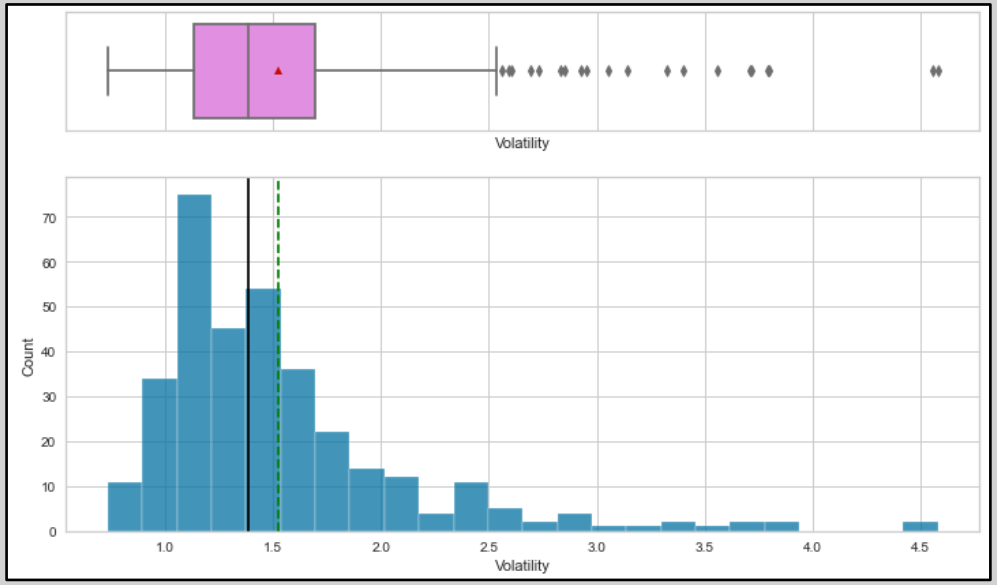


Observations

- The distribution of Price Change is approximately normal.
- The mean and median are both right in the center of the distribution.
- There are a handful of outliers on either side of the distribution.
- There are some negative values. This makes sense because a stock price can either rise (positive change) or fall (negative change).

Exploratory Data Analysis (Univariate)

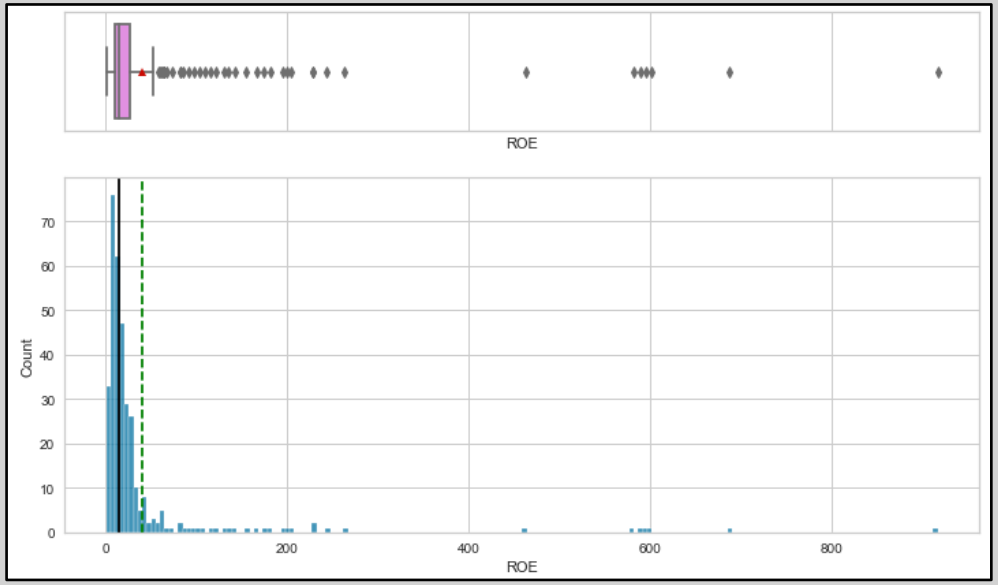
Distribution of Volatility



Observations

- The distribution of Volatility is positively skewed with most of the values falling between approximately 0.5 to 2.0.
- There are quite a few outliers on the right tail of the box chart visual, hence the positive skewness.
- There are not any negative values. This makes sense because we would not expect the percentage of price change of a stock (definition of volatility) to have a negative standard deviation.

Distribution of ROE

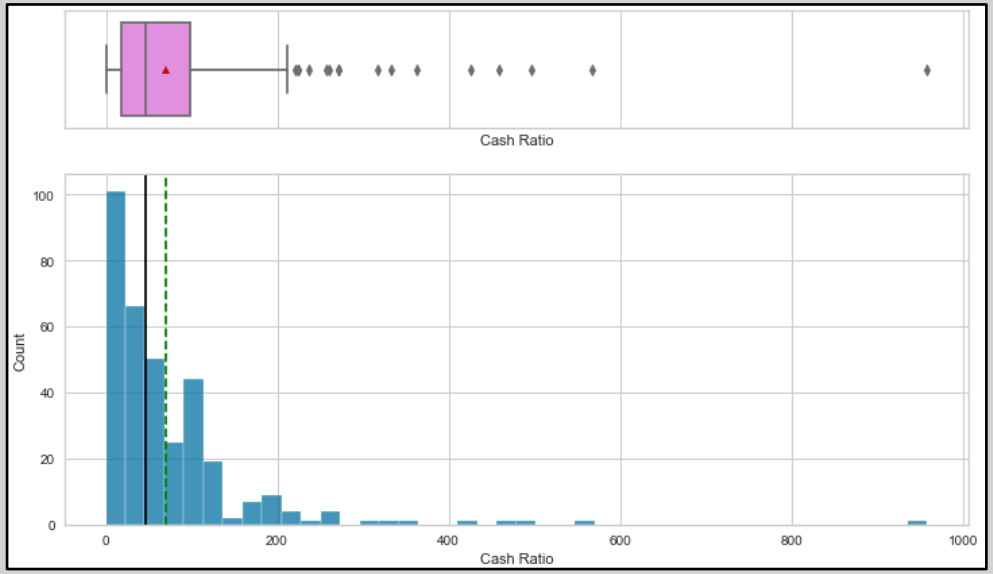


Observations

- The ROE is highly positively skewed as seen in both the box chart and histogram.
- Most values fall between 0 and 100, with many outliers beyond 100. The mean and median are both less than 100.
- There are no negative values which suggests that the stocks in the dataset have generally performed well enough to not lose money for the shareholders.

Exploratory Data Analysis (Univariate)

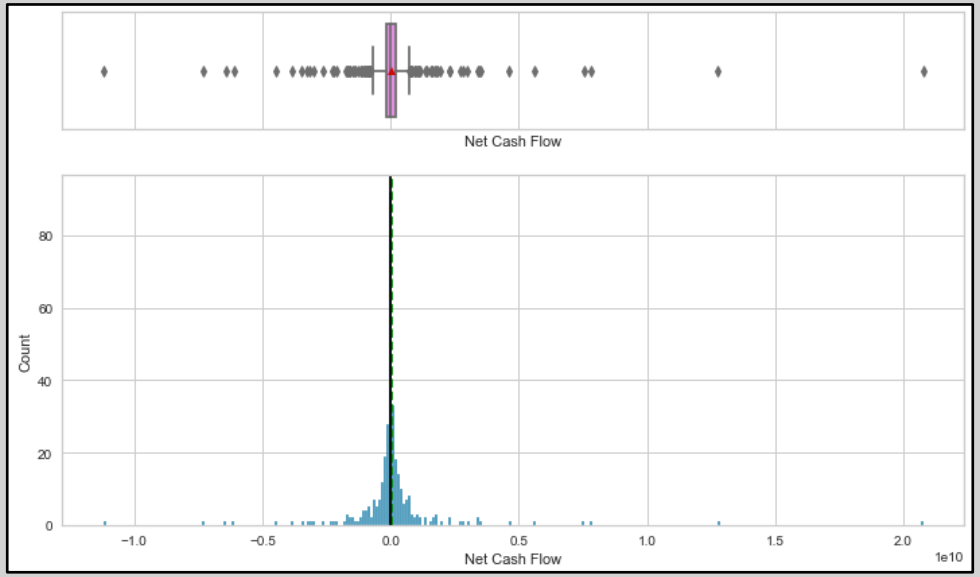
Distribution of Cash Ratio



Observations

- The distribution of the Current Ratio is positively skewed.
- Most of the data falls between 0 and 200 with the mean around 75 and median around 50.
- There are a handful of positive outliers in the boxplot, hence the positive skewness of the histogram.
- There are not any negative values which indicates that the companies seem to all be in a positive position relative to cash versus current liabilities.

Distribution of Net Cash Flow

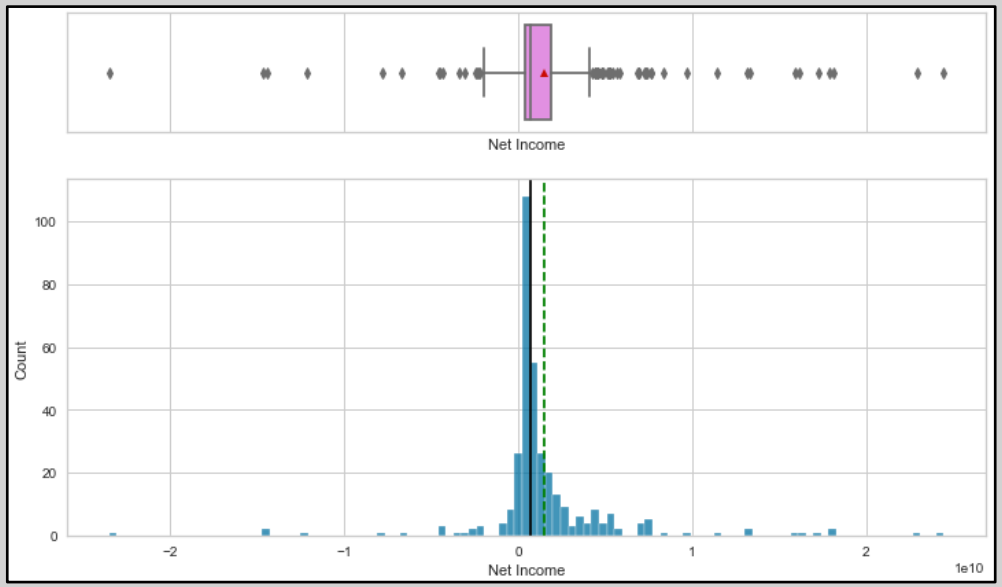


Observations

- The distribution of Net Cash Flow is approximately normal with many positive and negative outliers.
- Most of the data, including mean and median, is hovering around 0.0 (billion) which indicates most companies have a near net cash inflow and outflow.
- The presence of negative values here indicates that some companies have a higher cash outflow than inflow and vice versa.

Exploratory Data Analysis (Univariate)

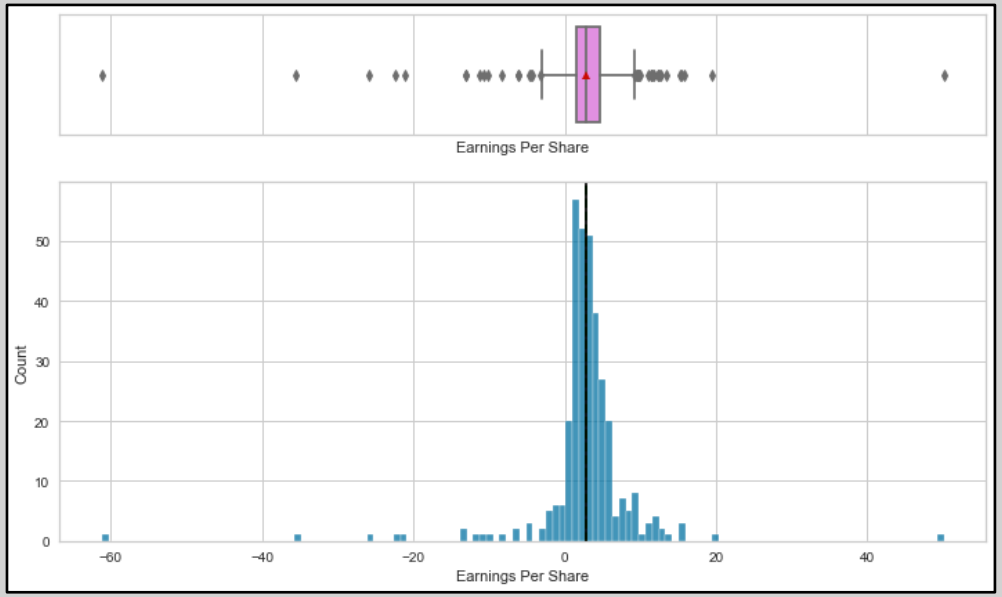
Distribution of Net Income



Observations

- The distribution of the Net Income is approximately normal with positive and negative outliers.
- Most of the data falls between -1 and 1 (billion), with median and mean hovering near 0.
- It makes sense that this data could be either positive or negative because a companies can operate at a profit (positive) or loss (negative) depending on expenses, interest, and taxes.

Distribution of Earnings Per Share

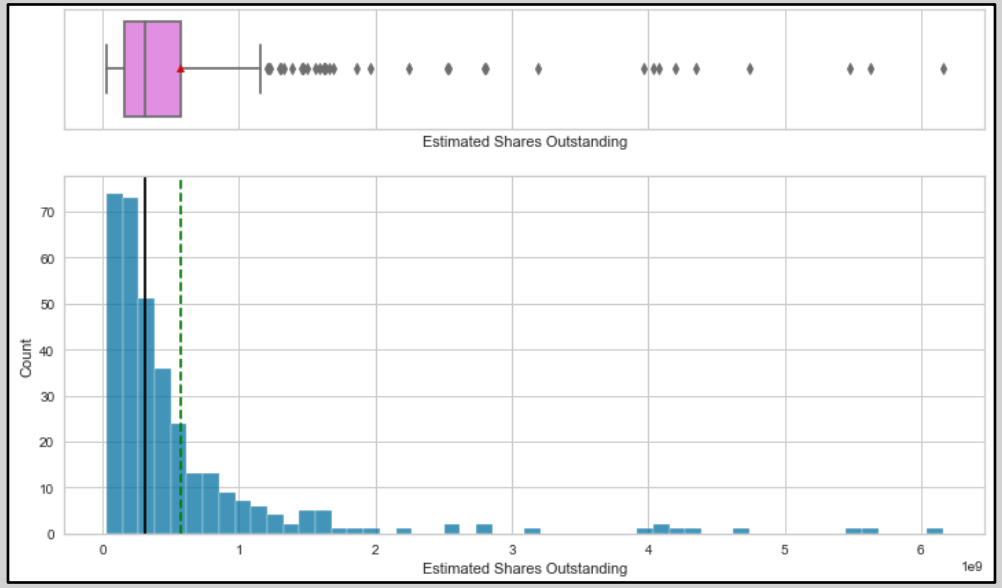


Observations

- The distribution of Earnings Per Share is approximately normal with positive and negative outliers.
- Most of the data falls between -10 and 10 with he median and mean around 3.
- The presence of negative values makes sense because a company's EPS will be negative if it is operating at a loss.

Exploratory Data Analysis (Univariate)

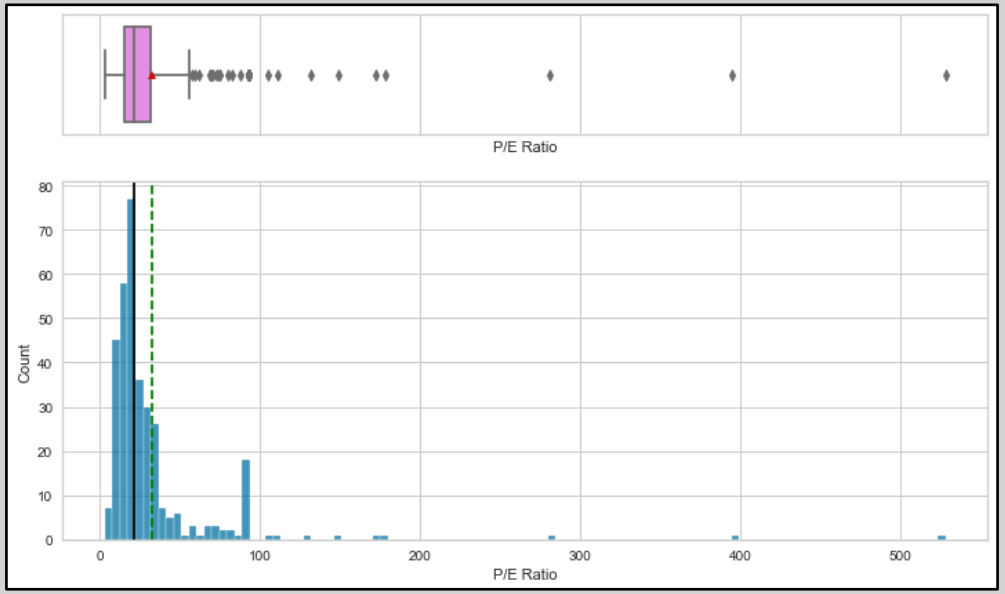
Distribution of Shares Outstanding



Observations

- The distribution of Shares Outstanding is highly positively skewed.
- Most values fall between 0 and 2 (hundred million).
- Mean and median are both between 0 and 1 (hundred million).
- The lack of negative values makes sense because a company cannot have negative shares of stock.

Distribution of P/E Ratio

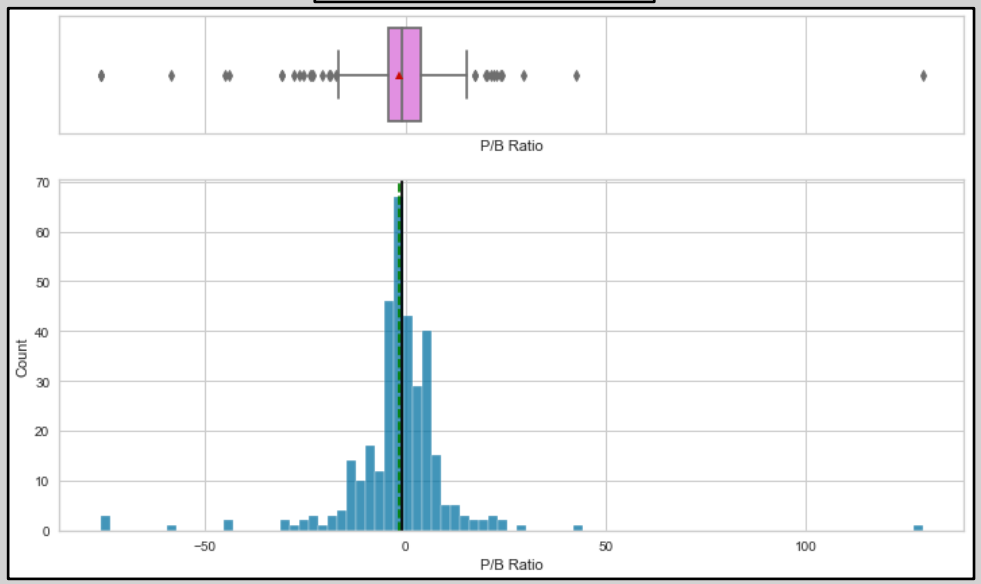


Observations

- The distribution of P/E Ratio is positively skewed.
- Mean and median both fall between 0 and 50.
- There are not any negative values which makes sense because there cannot be a negative stock price (half of the calculation) and most stock prices would have been high enough to offset a small negative EPS value (other half of the calculation).

Exploratory Data Analysis (Univariate)

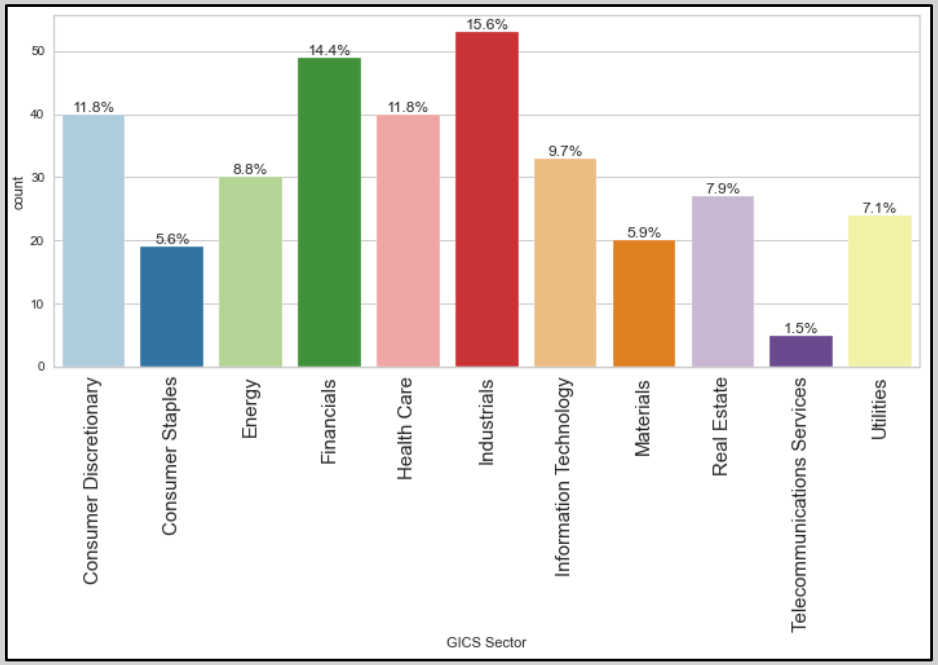
Distribution of P/B Ratio



Observations

- The distribution of the P/B Ratio is approximately normal with some positive and negative outliers.
- Most of the data falls between -50 and 50, and the mean and median are slightly less than 0.
- Negative values are caused by companies whose book value is negative. It makes sense that we would have some negatives in this dataset because there are some negative net income values.

Distribution of GICS Sector



Observations

- The largest sector is Industrials at 15.6% of the total and over 50 occurrences in the dataset.
- Financials is the second largest sector at 14.4% and just under 50 occurrences in the dataset.
- Consumer Discretionary and Health Care are tied for third largest sector at 11.8% and 40 occurrences in the dataset each.

Exploratory Data Analysis (Univariate)

GICS Sub Industry – Top 10 (custom visual)

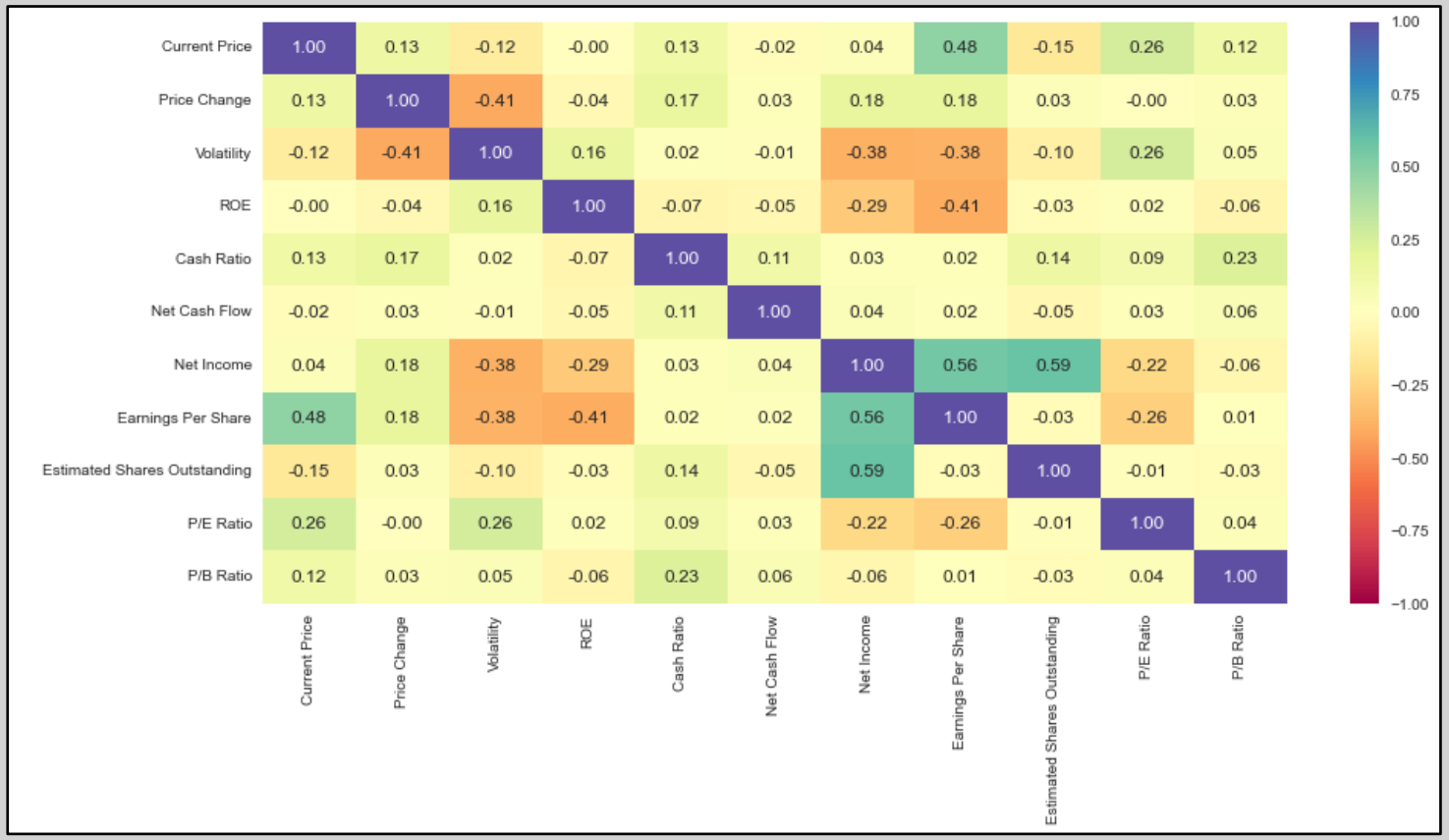
GICS Sector	GICS Sub Industry	Count	% of Total
Energy	Oil & Gas Exploration & Production	16	4.7%
Industrials	Industrial Conglomerates	14	4.1%
Real Estate	REITs	14	4.1%
Utilities	Electric Utilities	12	3.5%
Information Technology	Internet Software & Services	12	3.5%
Health Care	Health Care Equipment	11	3.2%
Utilities	MultiUtilities	11	3.2%
Financials	Banks	10	2.9%
Financials	Property & Casualty Insurance	8	2.4%
Health Care	Biotechnology	7	2.1%
Financials	Diversified Financial Services	7	2.1%

Observations

- Since there are too many Sub Industries to fit on a single graph and still make it visually appealing, we will analyze the top 10 instead (11 shown due to tie at 10th spot).
- The largest Sub Industry is Oil & Gas Exploration & Production with 16 occurrences and 4.7% of the total. This falls in the Energy sector.
- The second largest Sub Industry is a tie between Industrial Conglomerates (of Industrials sector) and REITs (or Real Estate sector) with 14 occurrences and 4.1% of the total.
- The third largest Sub Industry is a tie between Electric Utilities (of Utilities sector) and Internet Software & Services (of Information Technology sector) at 12 occurrences and 3.5% of the total.
- There are 6 different sectors represented in the top 10 Sub Industries: Energy (1), Industrials (1), Real Estate (1), Utilities (2), Information Technology (1), Health Care (2), Financials (3).

Exploratory Data Analysis (Multivariate)

Correlation Heatmap

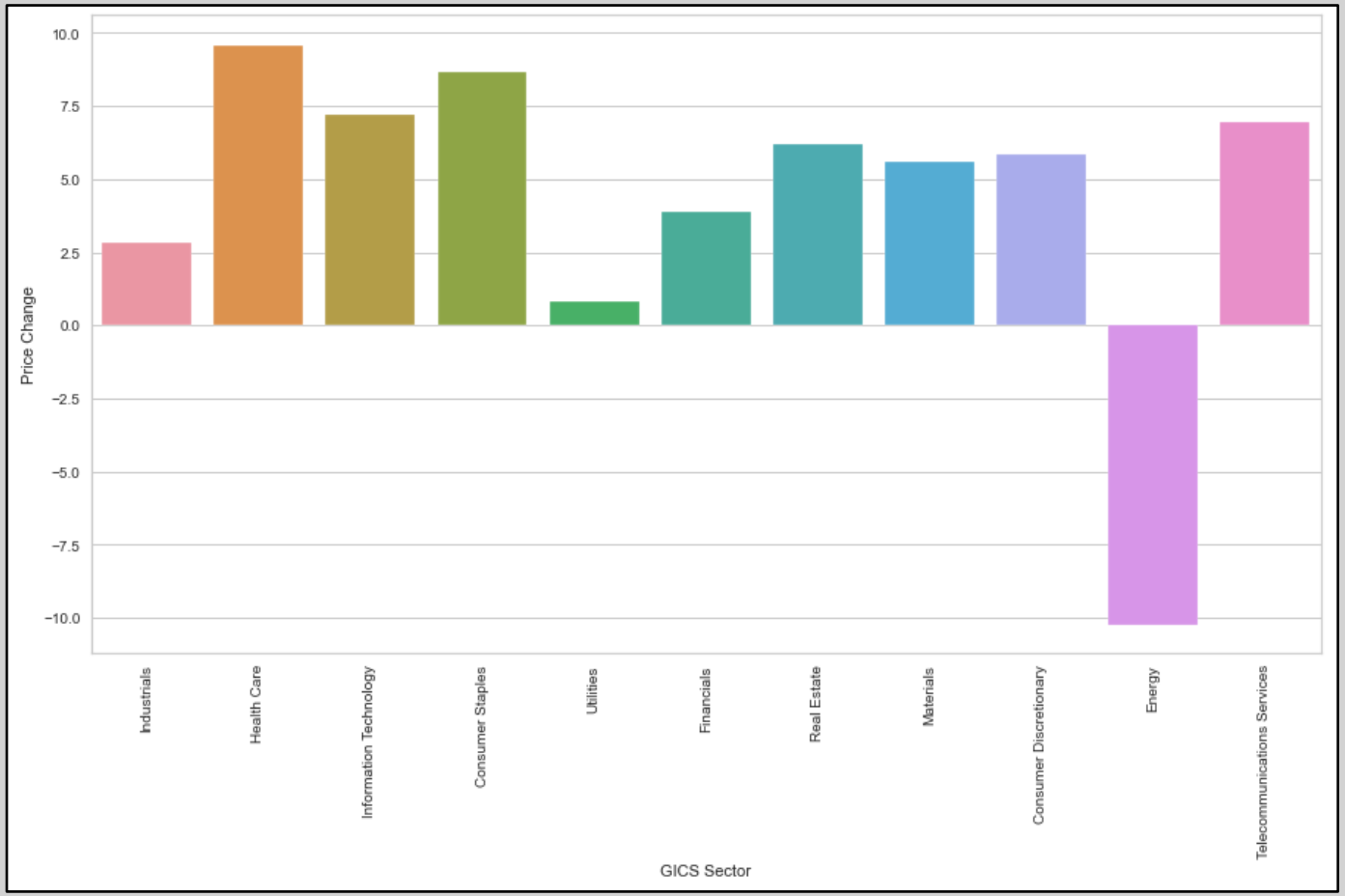


Observations

- The strongest positive correlation is 0.59 between Estimated Shares Outstanding and Net Income. This makes sense because a company with a higher Net Income is probably more likely to issue more shares of stock than one that is not performing quite as well.
- The next strongest positive correlation is 0.56 between Earnings Per Share and Net Income. This relationship makes sense because we would expect that a company's Net Income would ultimately drive things like profitability which impacts things like share price and earnings down the line.
- The strongest negative correlation is -0.41 between Volatility and Price Change.
- The second strongest negative correlation is -0.38 and occurs twice: between Volatility and Net Income & between Volatility and Earnings Per Share.

Exploratory Data Analysis (Multivariate)

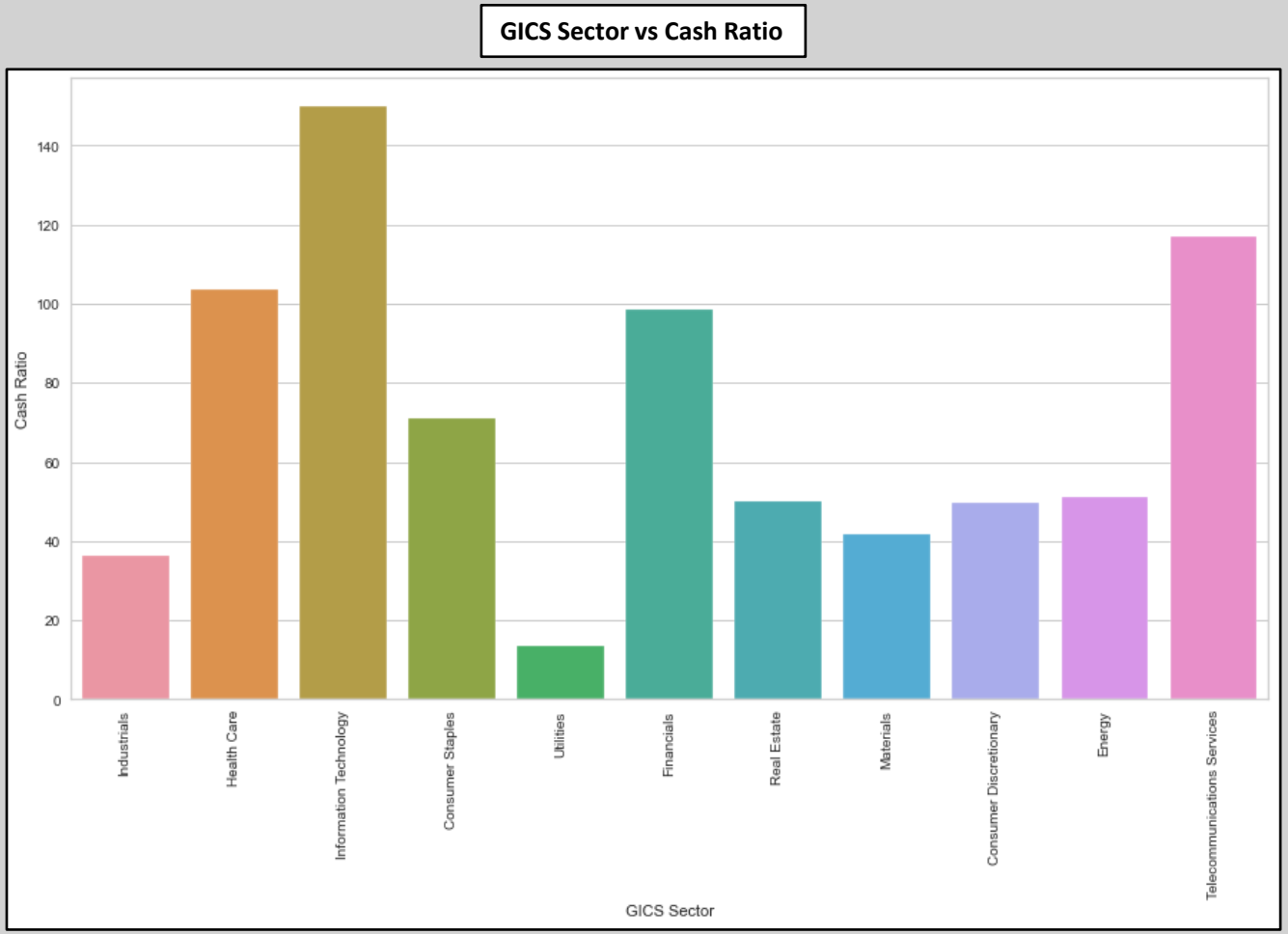
GICS Sector vs Price Change



Observations

- The largest Price Change among sectors occurred in the Energy sector with a change of over -10.
- The next largest change occurred in the Health Care sector at just under 10.
- Consumer Staples is the third largest change among sectors at 8 or 8.5.
- Utilities had the lowest change among sectors at around 1.
- Industrials had the second lowest change among sectors at just over 2.5.
- The median change was 4.82 and the mean change was 4.08.

Exploratory Data Analysis (Multivariate)

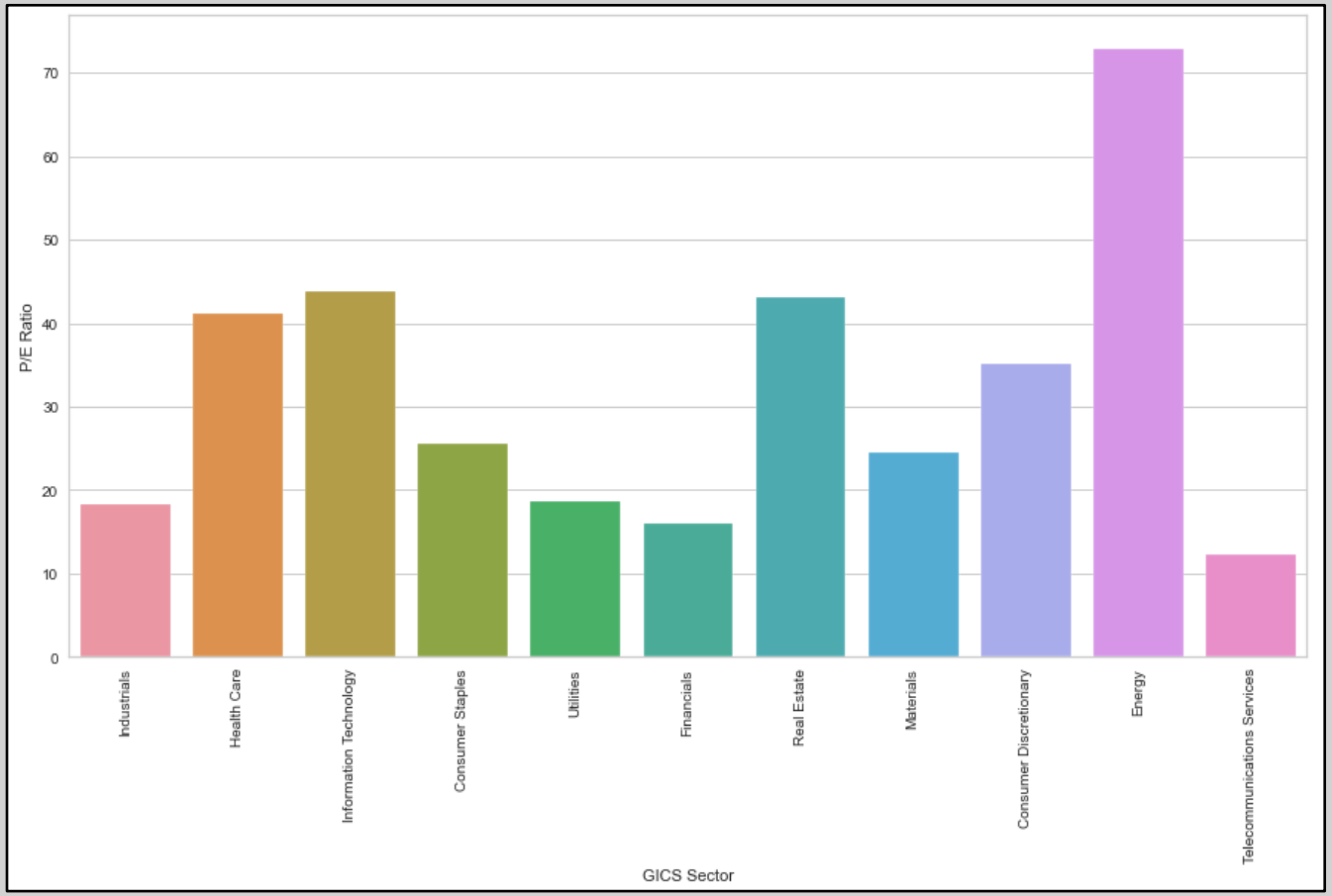


Observations

- The highest Cash Ratio belongs to the Information Technology sector at over 140.
- The second highest Cash Ratio belongs to the Telecommunications Services sector at just under 120.
- The third highest Cash Ratio belongs to the Health Care sector at just over 100.
- The lowest Cash Ratio belongs to the Utilities industry at just over 10.
- The second lowest Cash Ration belongs to the Industrials sector at under 40.

Exploratory Data Analysis (Multivariate)

GICS Sector vs P/E Ratio

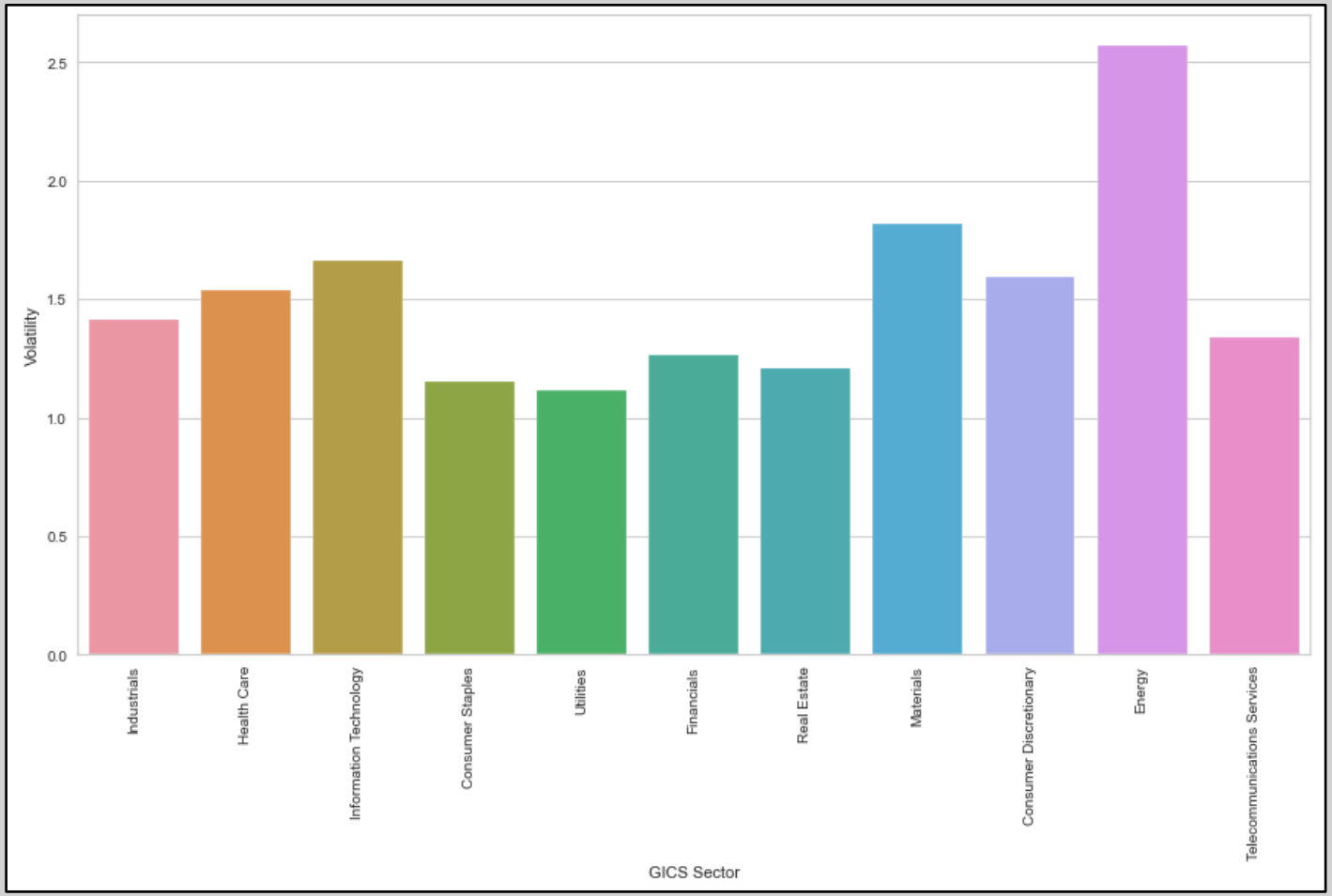


Observations

- The largest P/E Ratio belongs to the Energy sector at over 70.
- The second largest P/E Ratio is almost evenly split between Information Technology, Real Estate, and Health Care at the mid 40s.
- The lowest P/E Ratio belongs to the Telecommunications Services sector at just over 10.
- The second lowest P/E Ratio belongs to the Financials sector at around 15.

Exploratory Data Analysis (Multivariate)

GICS Sector vs Price Change



Observations

- The Energy sector has the highest volatility out of all sectors at over 2.5.
- The next four sectors with the highest volatility are Health Care, Information Technology, Materials, and Consumer Discretionary at between 1.5 and 2.0.
- The Utilities sector has the lowest volatility at just over 1.0.
- Consumer Staples is the second lowest volatile sector.
- All sectors, except for Energy, fall between 1.0 and 2.0 volatility.

Data Preprocessing

- The dataset does not have any duplicate values to treat.
- The dataset does not have any missing values to treat.
- Each of the numerical columns has outliers; however, we will not treat these as they are real world values.
- The dataset needed to be scaled prior to clustering.
    - The numeric columns were saved to a new dataframe, numeric_columns
    - Data was scaled using scaler.fit_transform
    - Finally, a new dataframe was created from the scaled data to use in the clustering steps.
- No additional feature engineering or data preprocessing was required.

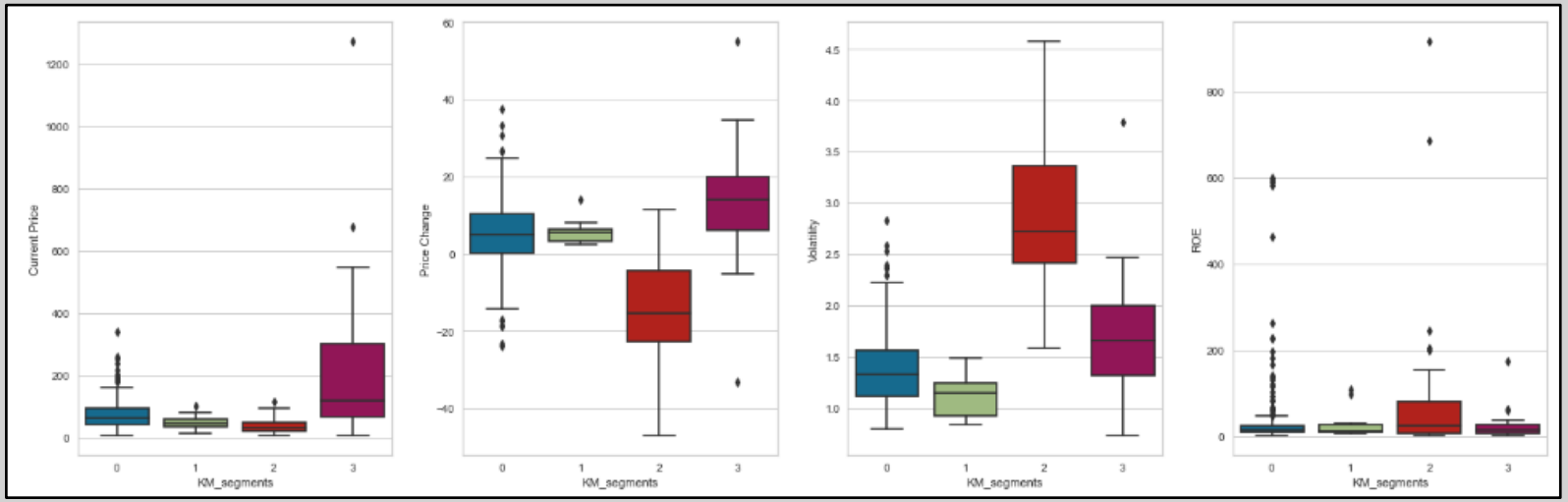
K-Means Clustering Summary

- The optimal number of clusters using K-Means was found to be 4.
- The final model was creating using the following steps:
  - kmeans = KMeans(n_clusters=4, random_state=1)
  - kmeans.fit(k_means_df) ; where k_means_df is the scaled subset of the original data
- Results of Cluster Profiling:
  - The first cluster (KM_segment 0) had the highest count of securities at 277.
  - A breakout of the distribution for each column is below and on the subsequent slides.

KM_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	count_in_each_segment
0	72.399112	5.066225	1.388319	34.620939	53.000000	-14046223.826715	1482212389.891697	3.621029	438533835.667184	23.843656	-3.358948	277
1	50.517273	5.747586	1.130399	31.090909	75.909091	-1072272727.272727	14833090909.090910	4.154545	4298826628.727273	14.803577	-4.552119	11
2	38.099260	-15.370329	2.910500	107.074074	50.037037	-159428481.481481	-3887457740.740741	-9.473704	480398572.845926	90.619220	1.342067	27
3	234.170932	13.400685	1.729989	25.600000	277.640000	1554926560.000000	1572611680.000000	6.045200	578316318.948800	74.960824	14.402452	25

K-Means Clustering Summary

Cluster Profiling - Distribution

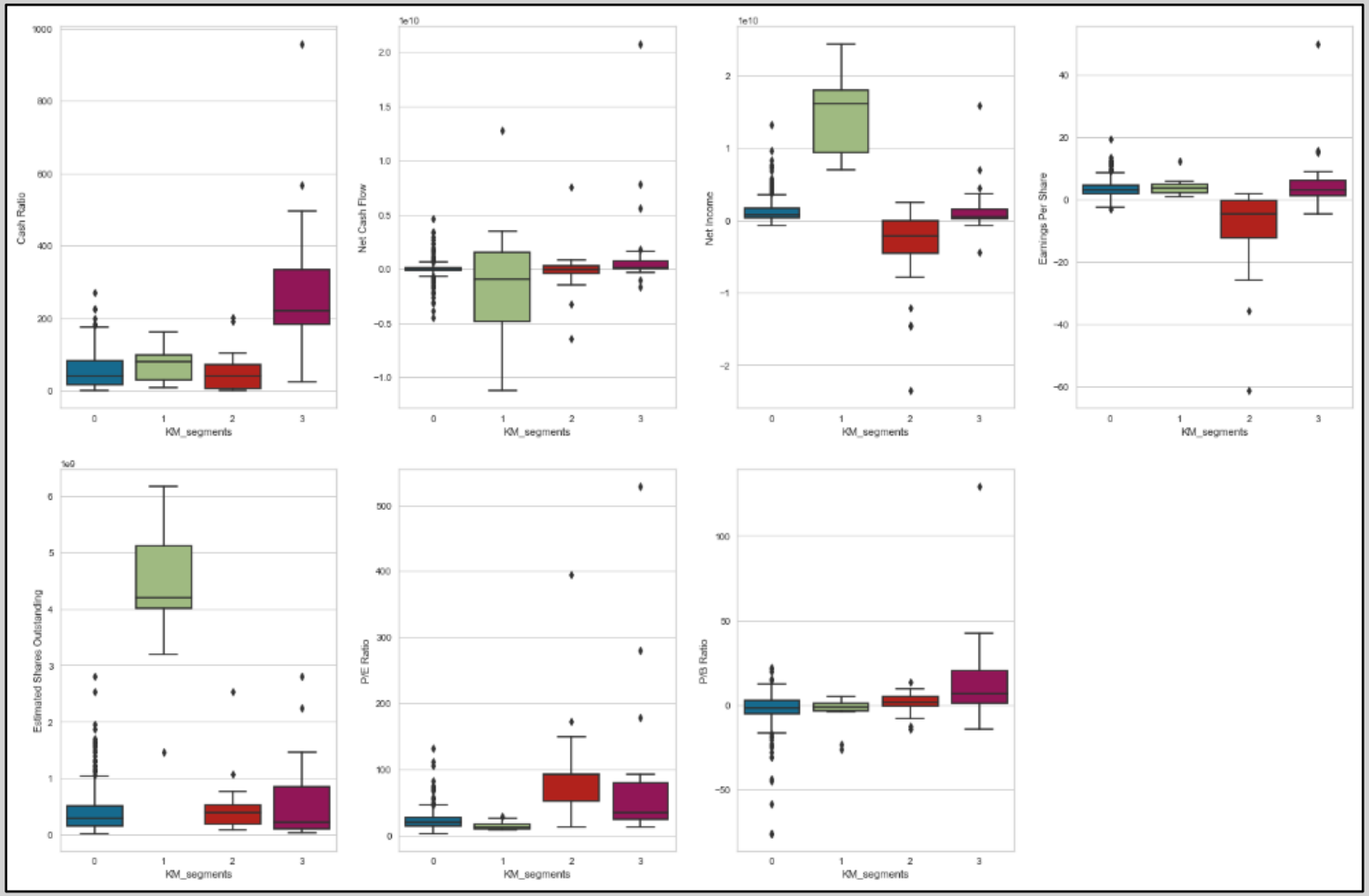


Count of GICS Sector by Cluster

KM_segments	GICS Sector	
0	Consumer Discretionary	33
	Consumer Staples	17
	Energy	6
	Financials	45
	Health Care	29
	Industrials	52
	Information Technology	24
	Materials	19
	Real Estate	26
	Telecommunications Services	2
	Utilities	24
1	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	3
	Health Care	2
	Information Technology	1
	Telecommunications Services	2
2	Energy	22
	Industrials	1
	Information Technology	3
	Materials	1
3	Consumer Discretionary	6
	Consumer Staples	1
	Energy	1
	Financials	1
	Health Care	9
	Information Technology	5
	Real Estate	1
	Telecommunications Services	1

K-Means Clustering Summary

Cluster Profiling - Distribution



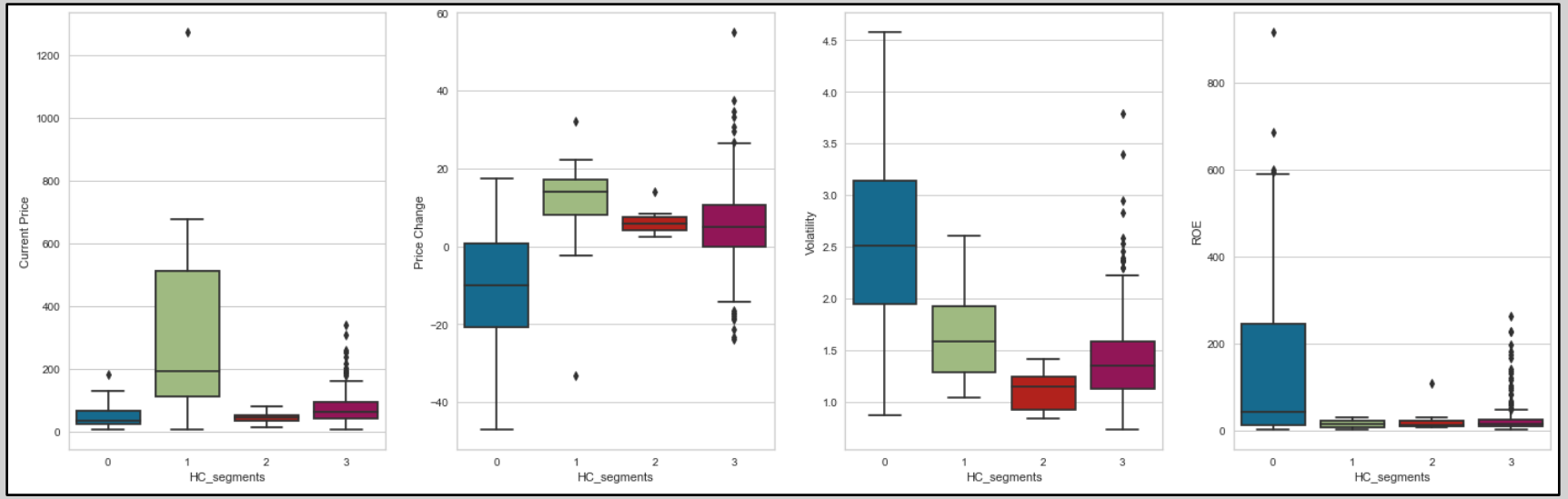
Hierarchical Clustering Summary

- The optimal number of clusters using Hierarchical Clustering was found to be 4.
- The final model was creating using the following steps:
  - HCmodel = AgglomerativeClustering(n_clusters=4, affinity='euclidean', linkage='ward')
  - HCmodel.fit(hc_df) ; where hc_df is the scaled subset of the original data
- Results of Cluster Profiling:
  - The fourth cluster (HC_segment 3) had the highest count of securities at 285.
  - A breakout of the distribution for each column is below and on the subsequent slides.

HC_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	count_in_each_segment
0	48.006208	-11.263107	2.590247	196.551724	40.275862	-495901724.137931	-3597244655.172414	-8.689655	486319827.294483	75.110924	-2.162622	29
1	326.198218	10.563242	1.642560	14.400000	309.466667	288850666.666667	864498533.333333	7.785333	544900261.301333	113.095334	19.142151	15
2	42.848182	6.270446	1.123547	22.727273	71.454545	558636363.636364	14631272727.272728	3.410000	4242572567.290909	15.242169	-4.924615	11
3	72.760400	5.213307	1.427078	25.603509	60.392982	79951512.280702	1538594322.807018	3.655351	446472132.228456	24.722670	-2.647194	285

Hierarchical Clustering Summary

Cluster Profiling - Distribution

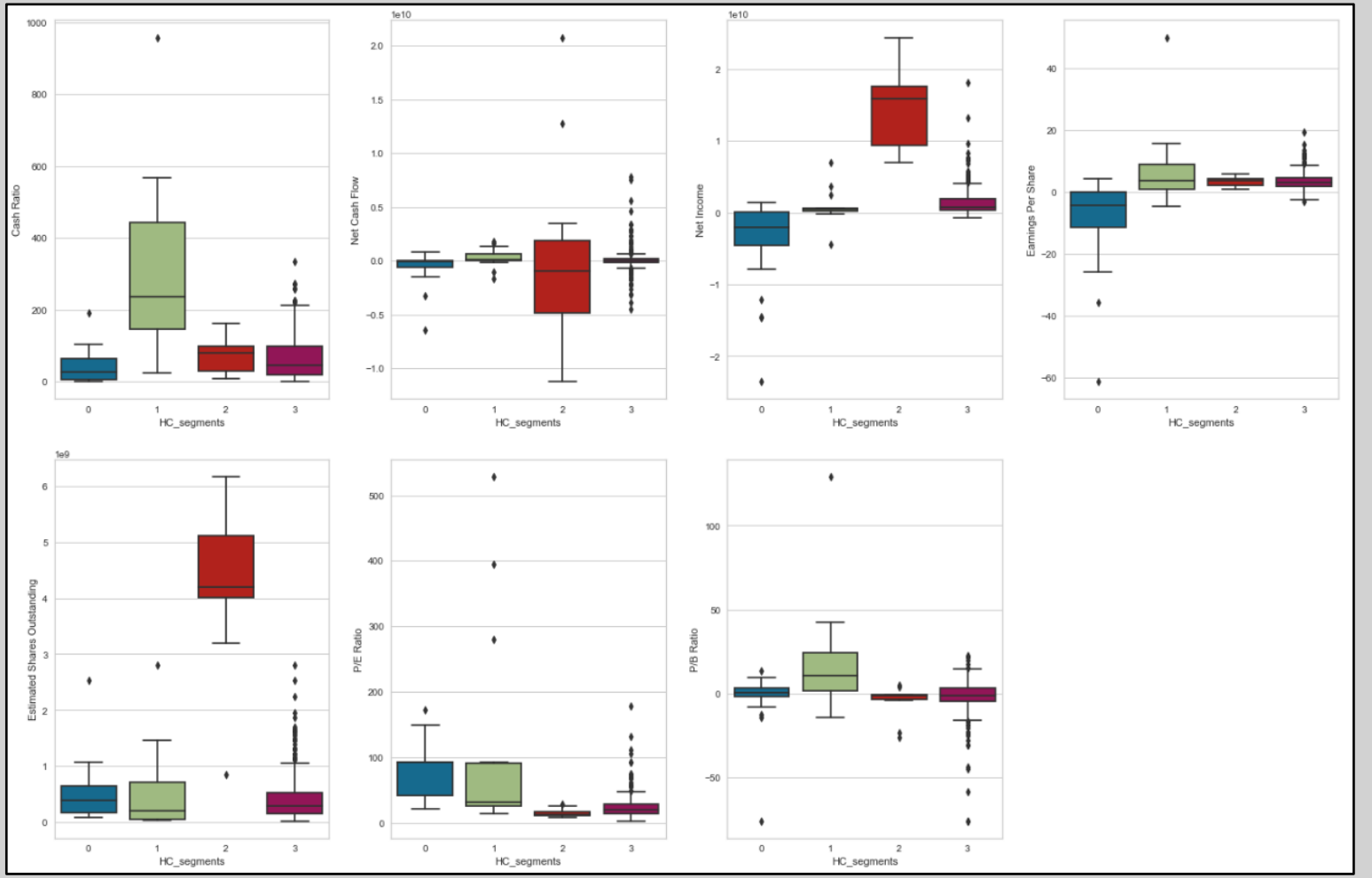


Count of GICS Sector by Cluster

HC_segments	GICS Sector	
0	Consumer Discretionary	1
	Consumer Staples	2
	Energy	22
	Financials	1
	Industrials	1
	Information Technology	1
	Materials	1
1	Consumer Discretionary	3
	Consumer Staples	1
	Health Care	5
	Information Technology	4
	Real Estate	1
	Telecommunications Services	1
2	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	4
	Health Care	1
	Information Technology	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

Hierarchical Clustering Summary

Cluster Profiling - Distribution



Appendix

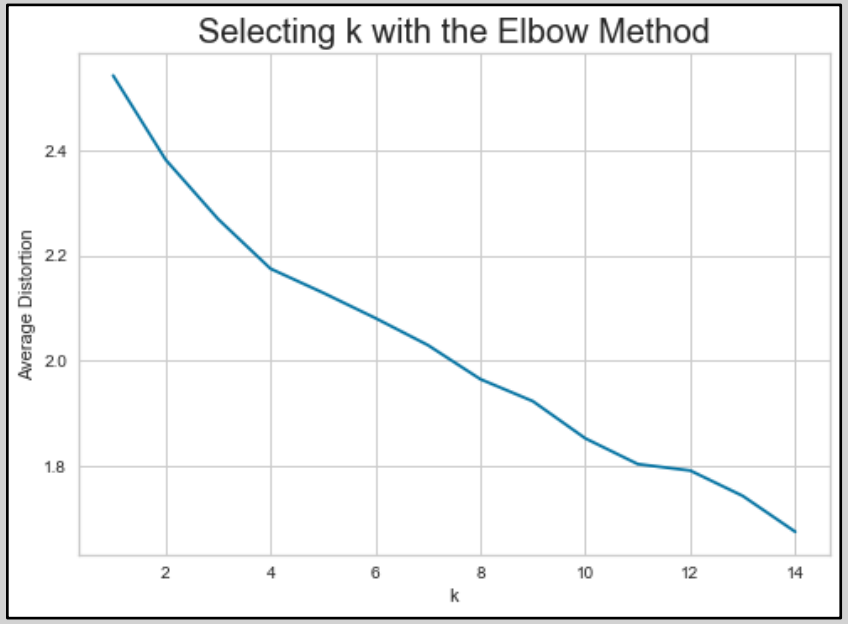
Data Dictionary and Shape

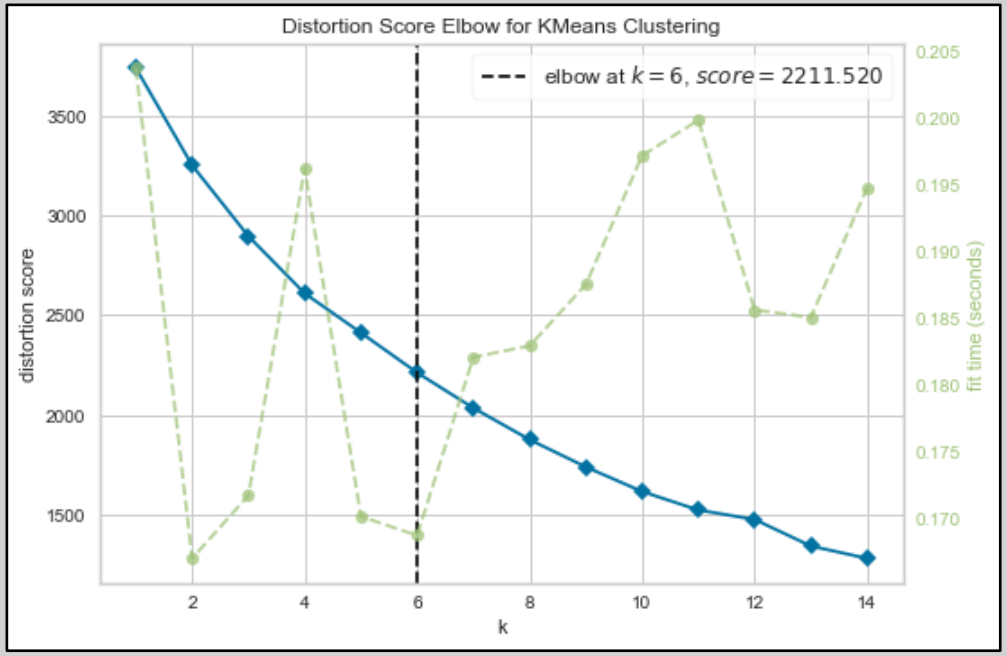
Column	Data Type	Description
Ticker Symbol	object	An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
Company	object	Name of the company
GICS Sector	object	The specific economic sector assigned to a company by the GICS that best defines its business operations
GICS Sub Industry	object	The specific sub-industry group assigned to a company by the GICS that best defines its business operations
Current Price	float64	Current stock price in dollars
Price Change	float64	Percentage change in the stock price in 13 weeks
Volatility	float64	Standard deviation of the stock price over the past 13 weeks
ROE	int64	A measure of financial performance calculated by dividing net income by shareholders' equity
Cash Ratio	int64	The ratio of a company's total reserves of cash and cash equivalents to its total current liabilities
Net Cash Flow	int64	The difference between a company's cash inflows and outflows (in dollars)
Net Income	int64	Revenues minus expenses, interest, and taxes (in dollars)
Earnings Per Share	float64	Company's net profit divided by the number of common shares it has outstanding (in dollars)
Estimated Shares Outstanding	float64	Company's stock currently held by all its shareholders
P/E Ratio	float64	Ratio of the company's current stock price to the earnings per share
P/B Ratio	float64	Ratio of the company's stock price per share by its book value per share

- Data types: float64(7), int64(4), object(4)
- Shape of the data is 340 rows by 15 columns.

K-Means Clustering – Elbow Method

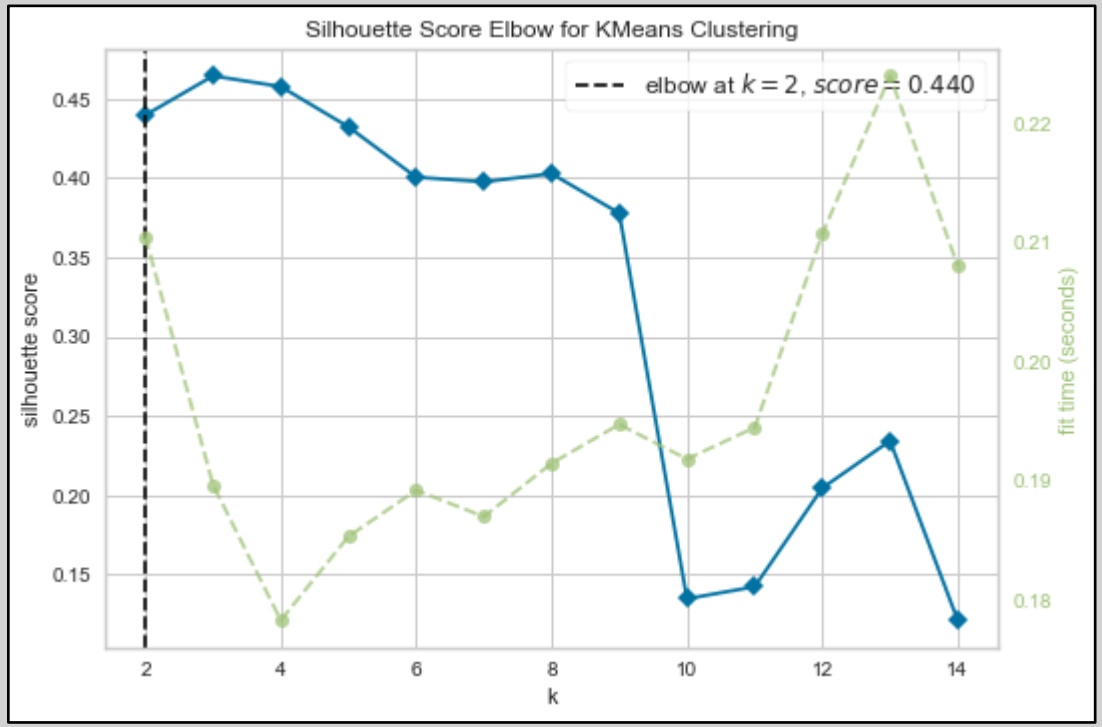
- Prior to clustering, the data was scaled using StandardScaler() and scaler.fit_transform()
- Next, we used the “elbow method” to analyze average distortion per each cluster (1 to 14 clusters)
- 4 clusters with average distortion of ~2.17 seems to be a sweet spot
- KElbowVisualizer(model, k=(1,15), timings=True) was also used to help determine how many clusters to use.
- We will do further analysis by checking the silhouette scores

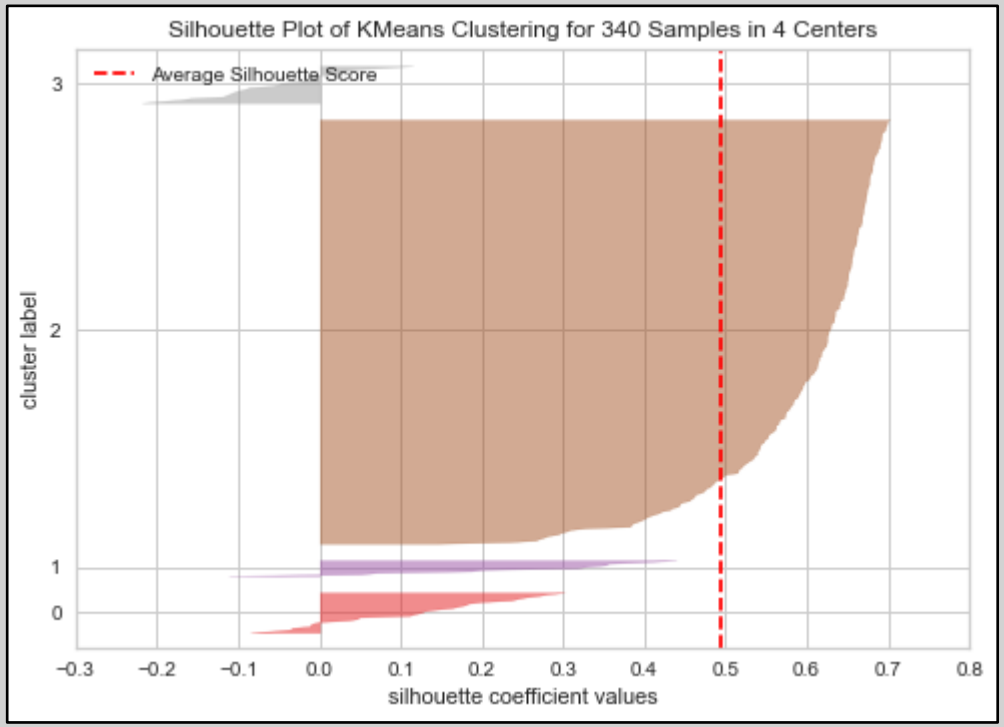




K-Means Clustering – Silhouette Scores

- 4 clusters provides the second highest silhouette score but at a significant time savings over the highest score (3 clusters).
- After running the SilhouetteVisualizer, 4 clusters was determined to be the optimal amount.
- A final model using 4 clusters was created and results with profiling analysis is included as part of the main slides.





Hierarchical Clustering – Cophenetic Correlation

- The highest cophenetic correlation found was 0.942, which was obtained using Euclidean distance and average linkage.
- The rest of the cophenetic correlation values are listed below.

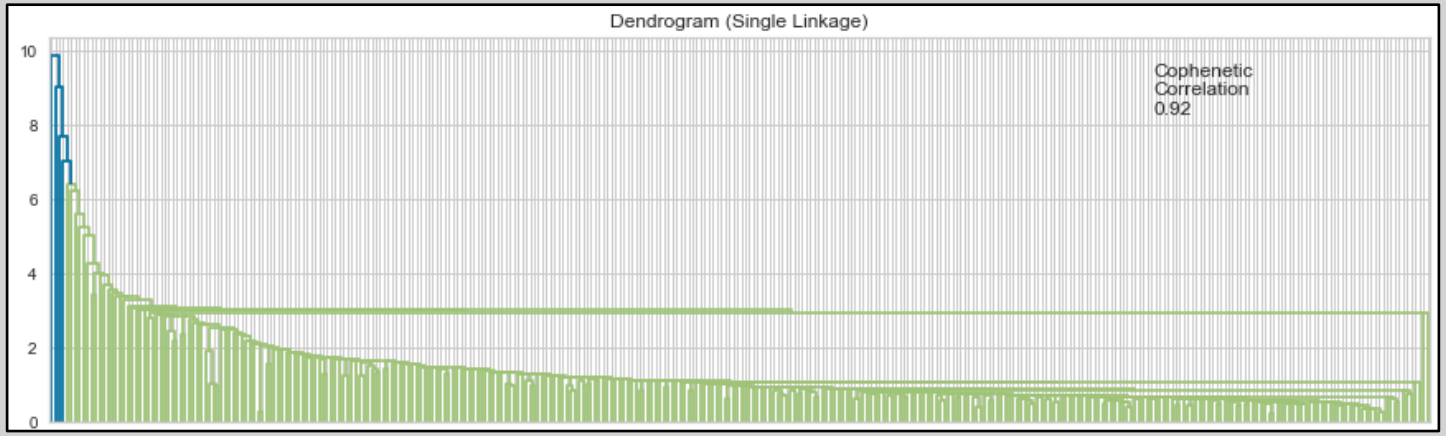
```
Cophenetic correlation for Euclidean distance and single linkage is 0.9232271494002922.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7873280186580672.
Cophenetic correlation for Euclidean distance and average linkage is 0.9422540609560814.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.8693784298129404.
Cophenetic correlation for Chebyshev distance and single linkage is 0.9062538164750717.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.598891419111242.
Cophenetic correlation for Chebyshev distance and average linkage is 0.9338265528030499.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.9127355892367.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.925919553052459.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.792530720285.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159737.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8708317490180426.
Cophenetic correlation for Cityblock distance and single linkage is 0.9334186366528574.
Cophenetic correlation for Cityblock distance and complete linkage is 0.7375328863205818.
Cophenetic correlation for Cityblock distance and average linkage is 0.9302145048594667.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.731045513520281.
```

- Next, we explored different linkage methods using Euclidean distance only. The highest found was still the average linkage at 0.942.
- The rest of the values are listed below.

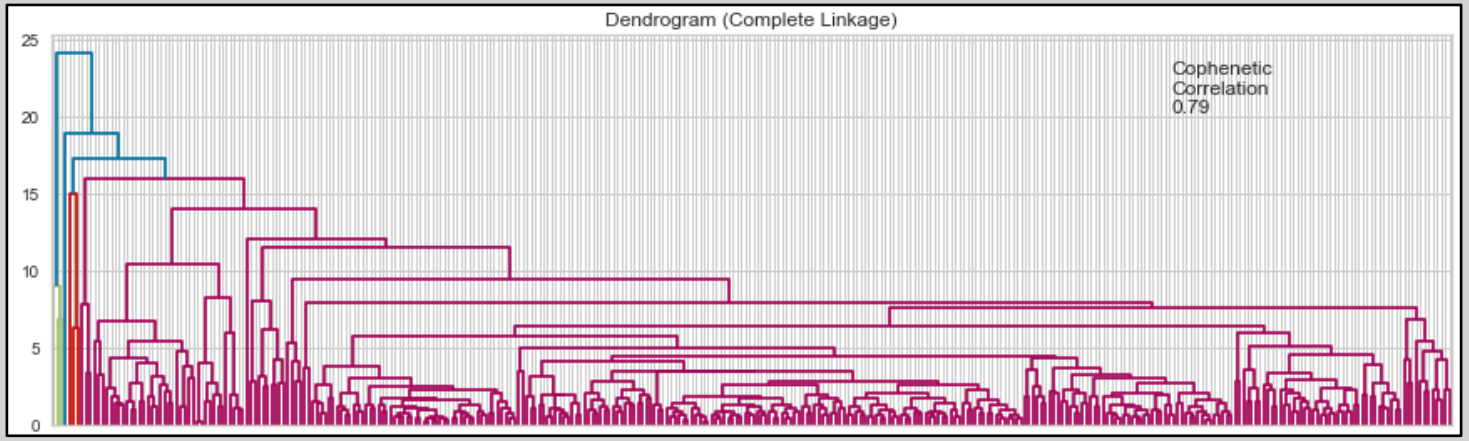
```
Cophenetic correlation for single linkage is 0.9232271494002922.
Cophenetic correlation for complete linkage is 0.7873280186580672.
Cophenetic correlation for average linkage is 0.9422540609560814.
Cophenetic correlation for centroid linkage is 0.9314012446828154.
Cophenetic correlation for ward linkage is 0.7101180299865353.
Cophenetic correlation for weighted linkage is 0.8693784298129404.
```

Hierarchical Clustering – Dendograms

- Single linkage has a high cophenetic correlation coefficient at 0.92, but the clustering is messy.

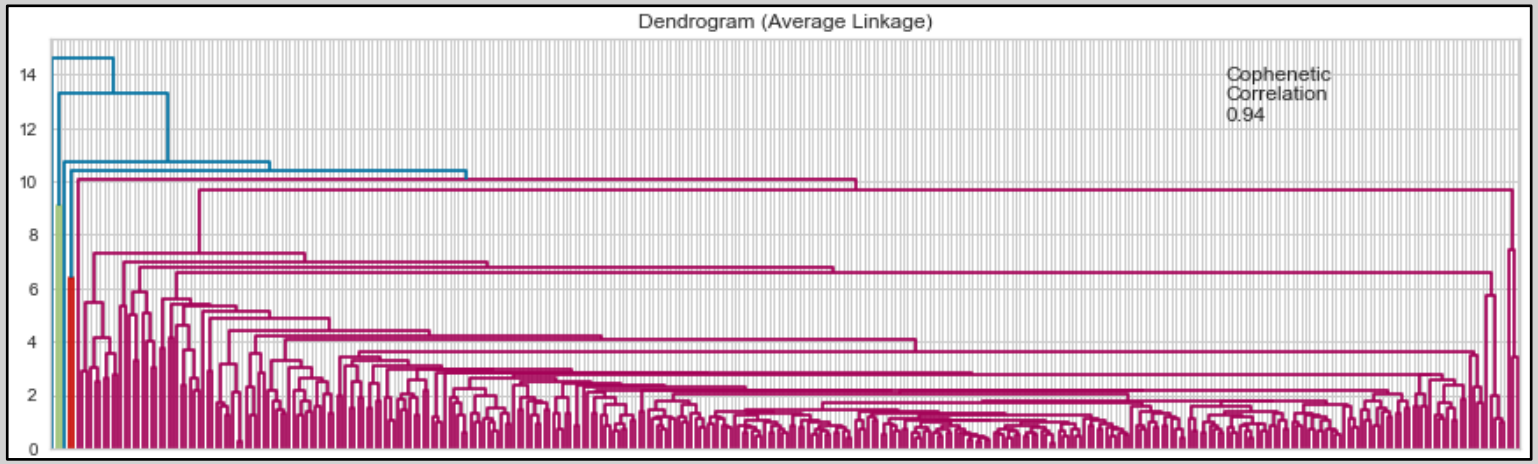


- Complete linkage has better clustering, but the correlation coefficient is lower than the single linkage at 0.79.

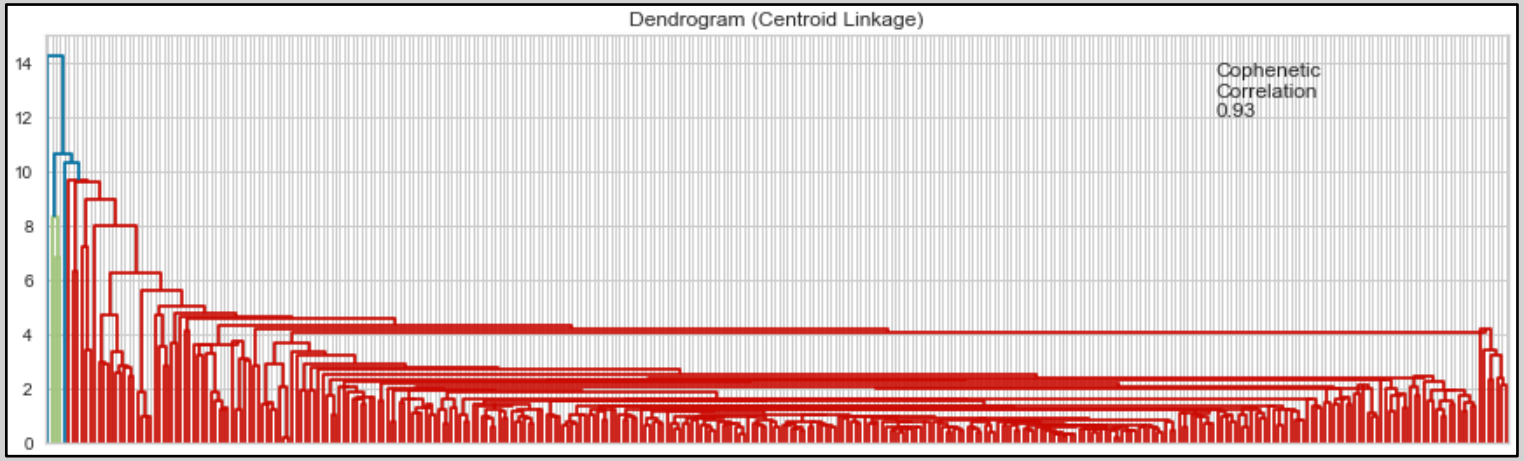


Hierarchical Clustering – Dendograms

- Average linkage has the highest cophenetic correlation coefficient at 0.94, but the clustering is not quite as good as complete linkage.

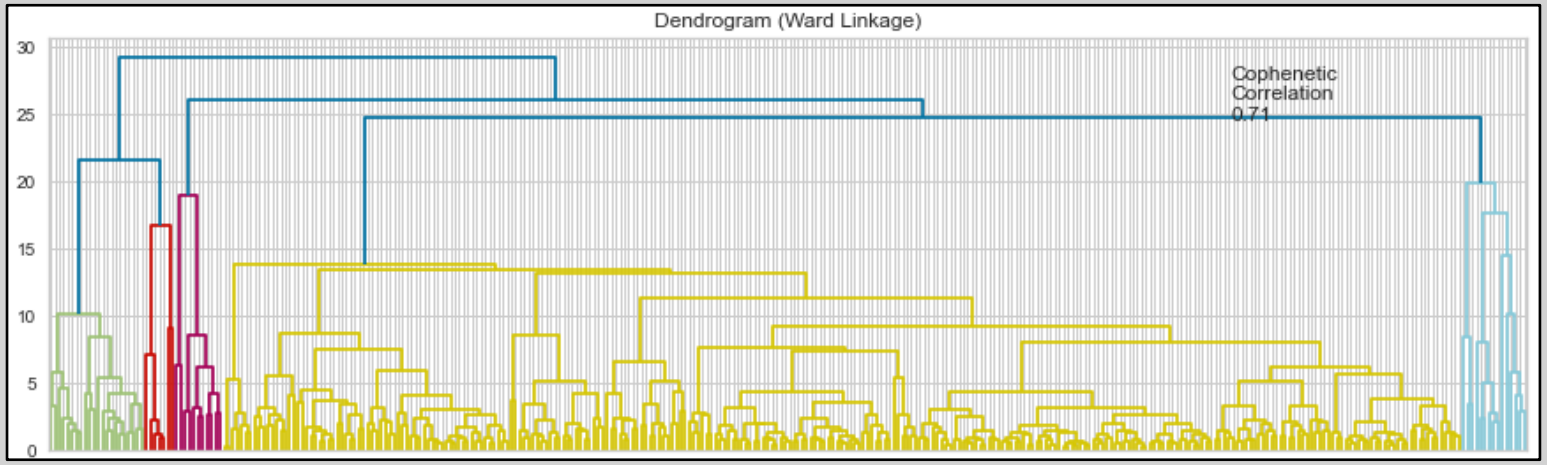


- Centroid linkage has the second highest cophenetic correlation at 0.93, but clustering still isn't as good as complete linkage.

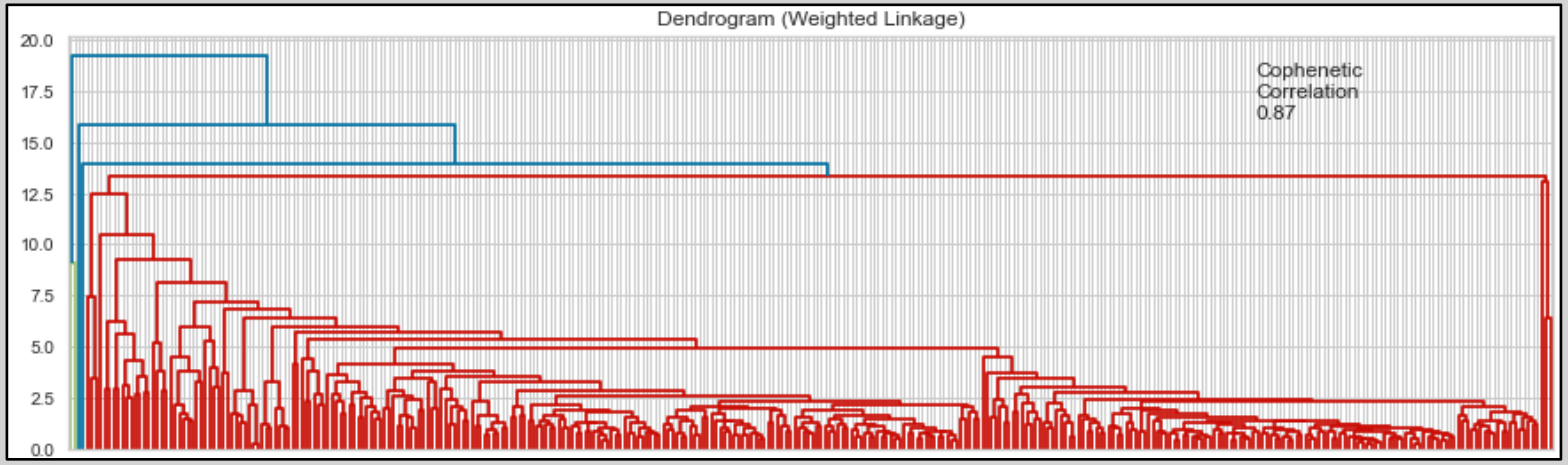


Hierarchical Clustering – Dendograms

- Ward linkage has the lowest cophenetic correlation, but the best clustering.



- Weighted linkage has middle of the road second highest cophenetic correlation at 0.87, but the clustering is not great.



Hierarchical Clustering – Observations

- Ward linkage has the lowest cophenetic correlation coefficient, but it has the best clustering. Even though its cophenetic correlation is the lowest of the group, 0.71 is still good so we will use ward linkage for the final model.
- The final model will be built using the AgglomerativeClustering method, using 4 clusters, with Euclidean distance and ward linkage.
- Results of the final model and cluster profiling are in included as part of the main slides.

K-Means vs Hierarchical Clustering

- The optimal number of clusters ended up being 4 for both clustering techniques.
- Hierarchical cluster 4 (HC_segment 3) ended up being the most like K-Means cluster 1 (KM_segment 0)
  - Each of these have the highest segment count and second highest current price.
  - Earnings per share is similar between these two – 3.66 for HC and 3.62 for KM.
- Hierarchical cluster 2 (HC_segment 1) ended up being the most like K-Means cluster 4 (KM_segment 3)
  - Each of these have the highest current price and highest price change.
  - Price change is similar between these two – 13.40 for HC and 10.56 for KM.
- Hierarchical cluster 1 (HC_segment 0) ended up being the most like K-Means cluster 3 (KM_segment 2)
  - Each of these have the highest Volatility and ROE.
- Hierarchical cluster 3 (HC_segment 2) ended up being the most like K-Means cluster 2 (KM_segment 1)
  - Each of these have the highest Net Income and Estimated Shares Outstanding.