


TRADE & AHEAD



Project 7 – Unsupervised Learning
Joshua Willis, PGP-DSBA

Agenda

- 
- ✓ BUSINESS PROBLEM & OBJECTIVE
 - ✓ DATA OVERVIEW
 - ✓ EDA RESULTS
 - ✓ DATA PREPROCESSING
 - ✓ OBSERVATIONS: K-MEANS CLUSTERING
 - ✓ OBSERVATIONS: HIERARCHICAL CLUSTERING
 - ✓ INSIGHTS & RECOMMENDATIONS

Business Problem & Objective



Business Problem:

The stock market is a great opportunity to invest and save for the future. It is important to maintain a diversified portfolio in order to maximize earnings in any market condition. It is often easy to get lost in the overwhelming number of financial metrics in order to determine the worth of a stock. By doing a cluster analysis, one can identify stocks that exhibit similar characteristics and ones that have minimal correlation. This helps investors better analyze stocks across different market segments and help protect against risks that could make the portfolio vulnerable to loss.

Objective:

Trade & Ahead is a financial consultancy firm who provide their customers with personalized investment strategies. Our data science team has been assigned the task of helping Trade & Ahead analyze the data and group the stocks based on the attributes provided and share insights about the characteristics of each group.

Data Overview



Data Dictionary:

Ticker Symbol: An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market

Company: Name of the company

GICS Sector: The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations

GICS Sub Industry: The specific sub-industry group assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations

Current Price: Current stock price in dollars

Price Change: Percentage change in the stock price in 13 weeks

Volatility: Standard deviation of the stock price over the past 13 weeks

ROE: A measure of financial performance calculated by dividing net income by shareholders' equity (shareholders' equity is equal to a company's assets minus its debt)

Cash Ratio: The ratio of a company's total reserves of cash and cash equivalents to its total current liabilities

Net Cash Flow: The difference between a company's cash inflows and outflows (in dollars)

Net Income: Revenues minus expenses, interest, and taxes (in dollars)

Earnings Per Share: Company's net profit divided by the number of common shares it has outstanding (in dollars)

Estimated Shares Outstanding: Company's stock currently held by all its shareholders

P/E Ratio: Ratio of the company's current stock price to the earnings per share

P/B Ratio: Ratio of the company's stock price per share by its book value per share (book value of a company is the net difference between that company's total assets and total liabilities)

Data Overview



Data Shape:

- 340 Rows & 15 Columns

Data Info:

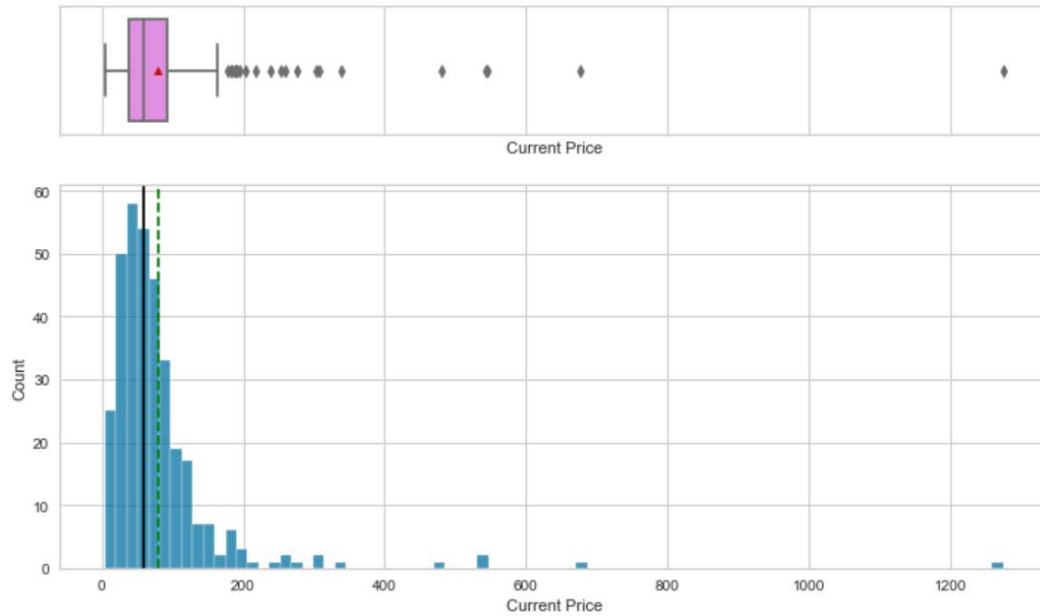
- There is object and numerical data

Data Sample:

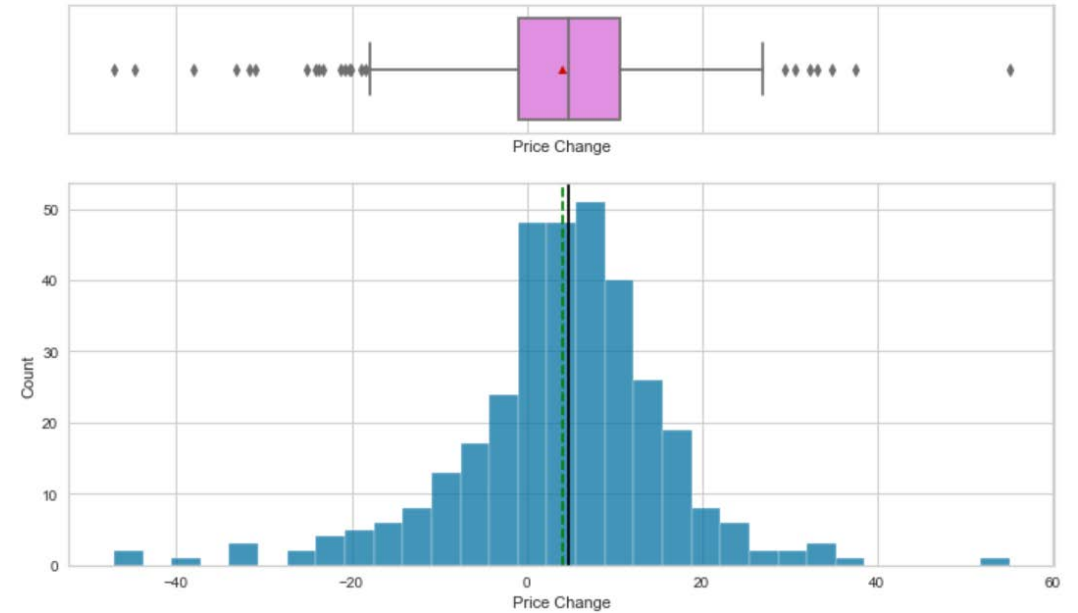
- Numerical data is inconsistent and will need to be treated

There are no duplicates or missing values

EDA Results

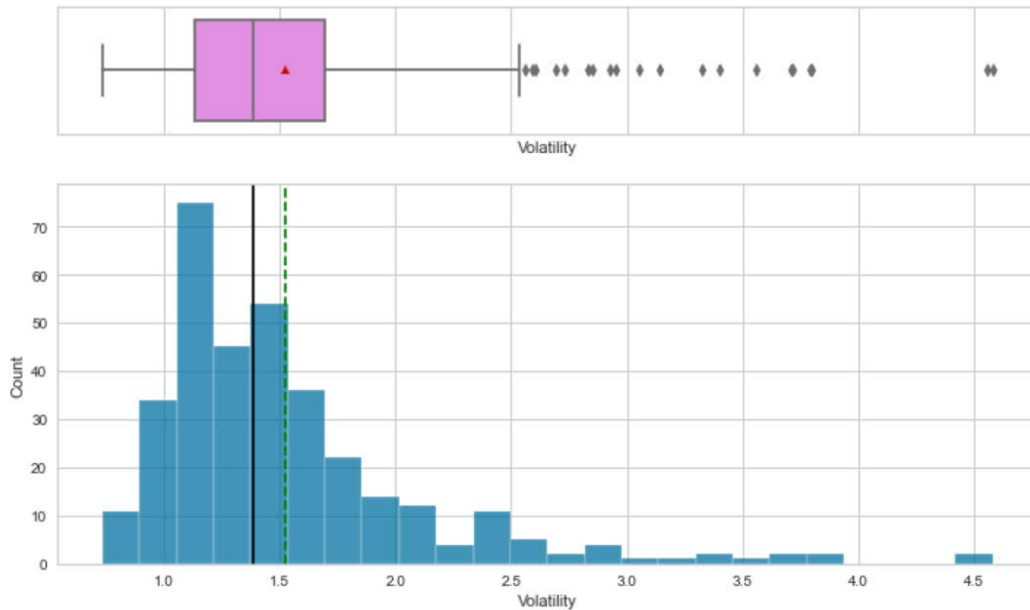


- *Current Price has outliers and is somewhat right skewed*
- *Average current price is \$80.86*

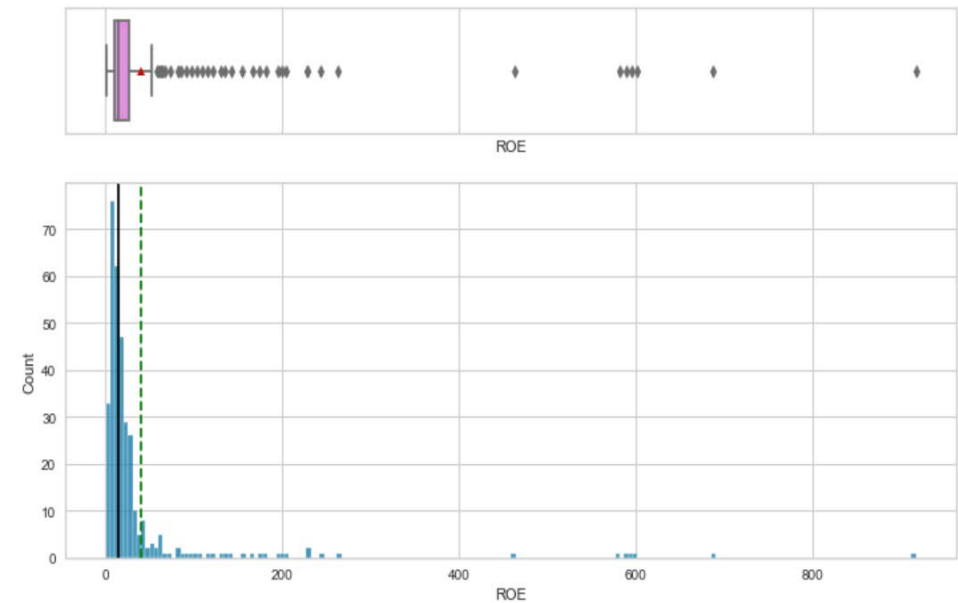


- *Price Change is normally distributed with some outliers*
- *Average price change is \$4.07*

EDA Results

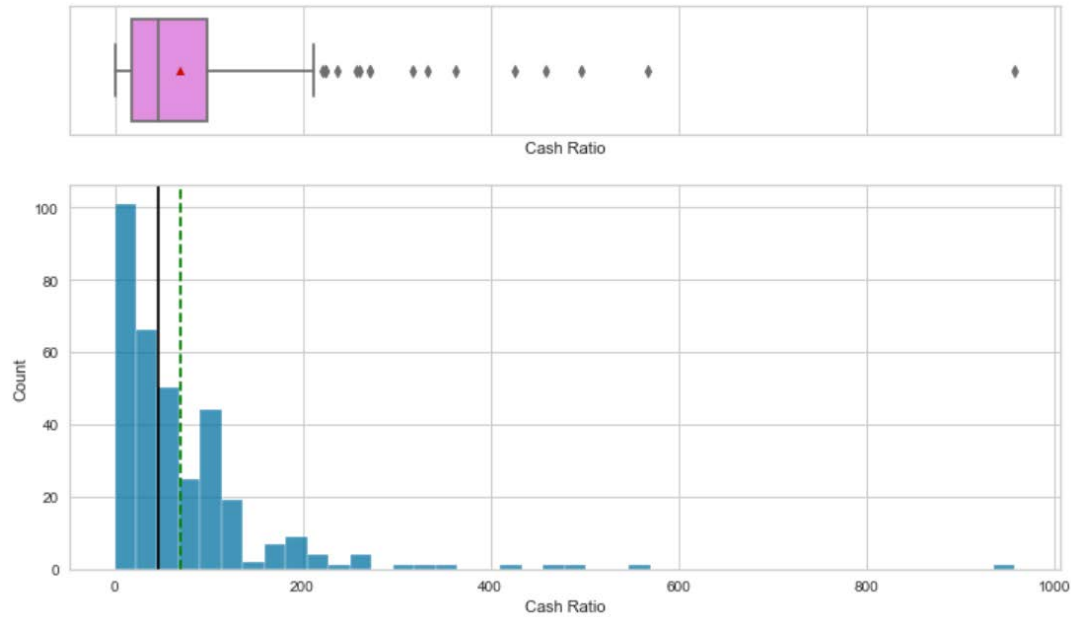


- *Volatility is right skewed with outliers*
- *Average Volatility is 1.52*

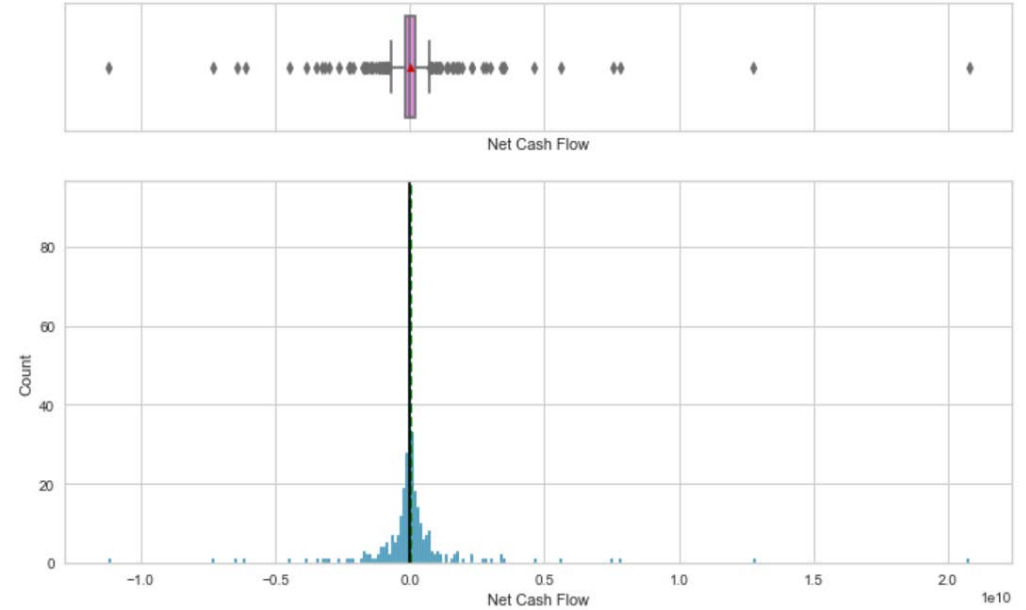


- *ROE is right skewed with a lot of outliers beyond 75% quartile*
- *Average ROE is \$39.59*

EDA Results

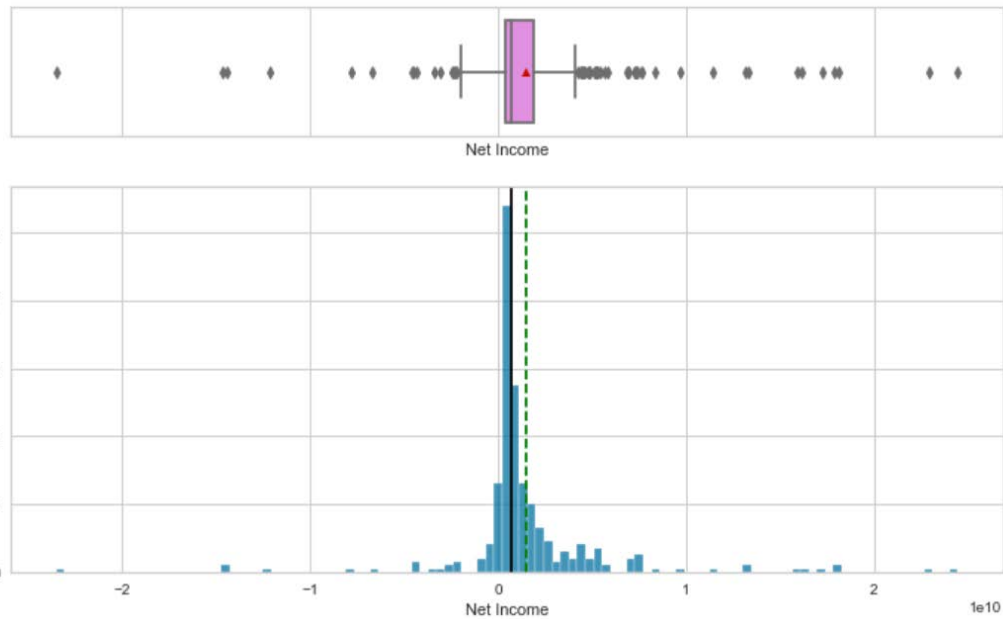


- *Cash Ratio is right skewed with outliers*
- *Average Cash Ratio is \$70.02*

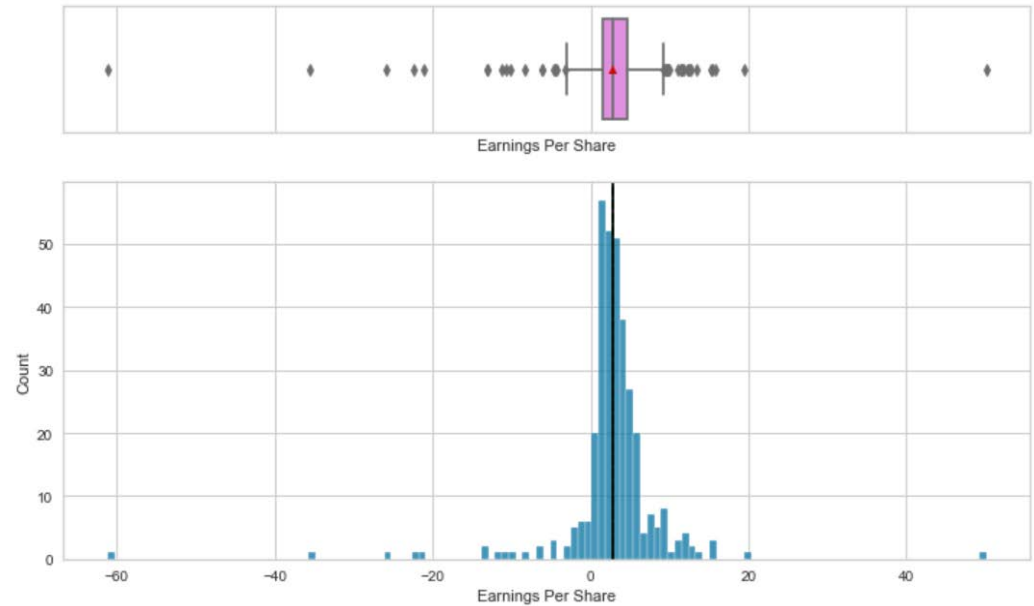


- *Net Cash Flow is normally distributed with outliers*
- *Average Net Cash Flow is \$55,537,620*

EDA Results



- *Net Income is normally distributed with outliers*
- *Average Net Income is \$1,494,384,602.94*

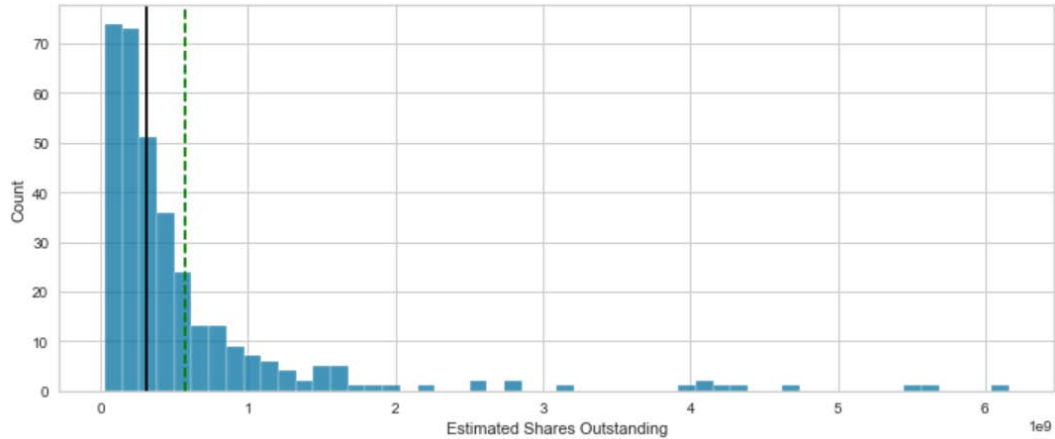


- *Earnings Per Share is somewhat normally distributed with outliers*
- *Average Earnings Per Share is \$2.77*

EDA Results



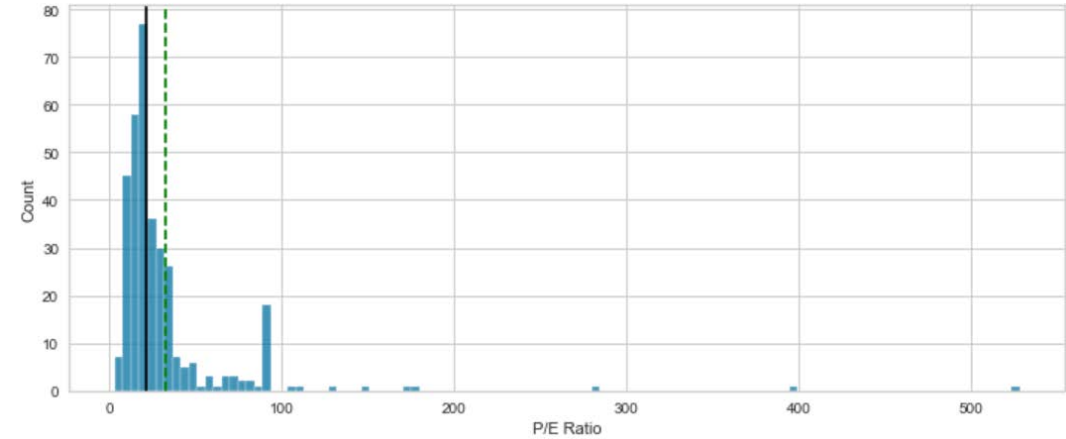
Estimated Shares Outstanding



- *Estimated Shares Outstanding is right skewed with outliers*
- *Average Estimated Shares Outstanding is 577,028,337*

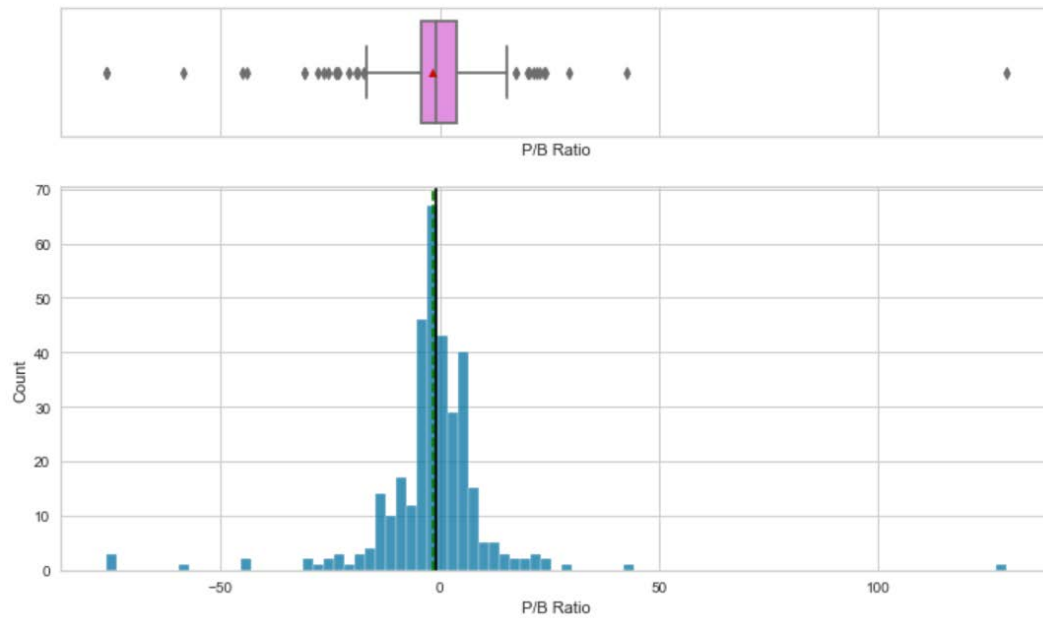


P/E Ratio

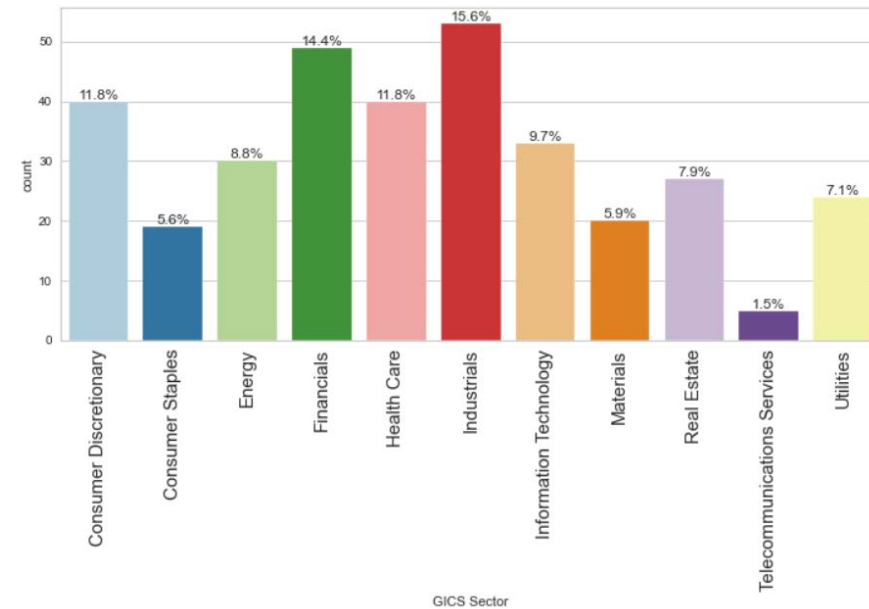


- *P/E ratio is right skewed with outliers*
- *Average P/E Ratio is \$32.61*

EDA Results

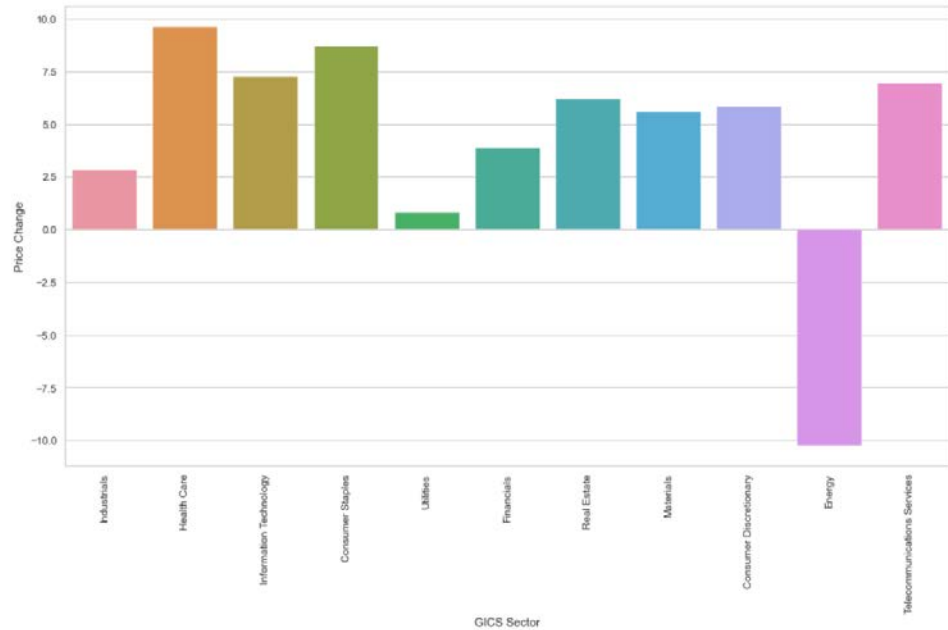


- *P/B Ratio is somewhat normally distributed with outliers*
- *Average P/B Ratio is -1.71*

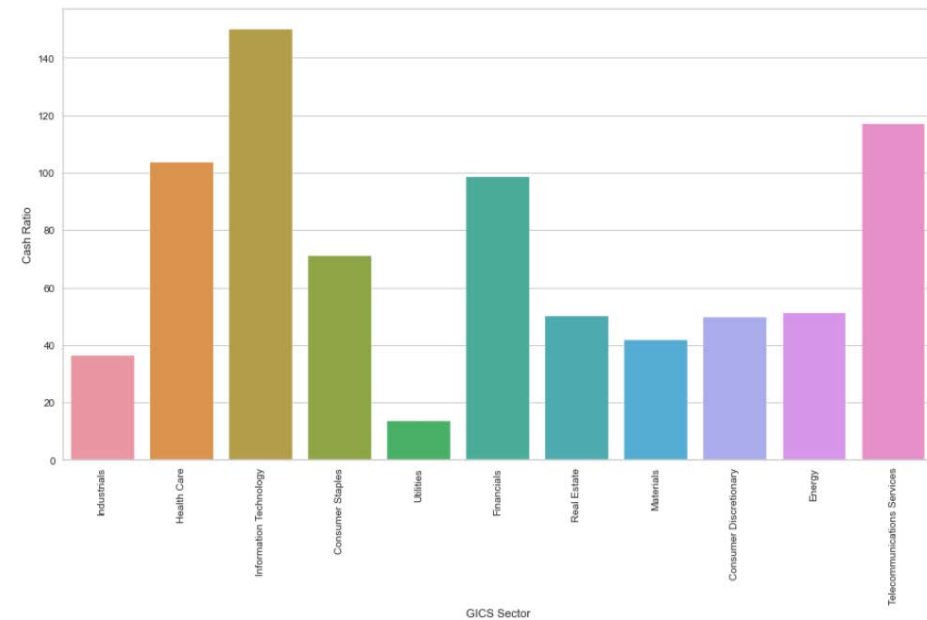


- *Top GICS Sectors for number of investments were Industrials, Financials, and Health Care / Consumer Discretionary*

EDA Results

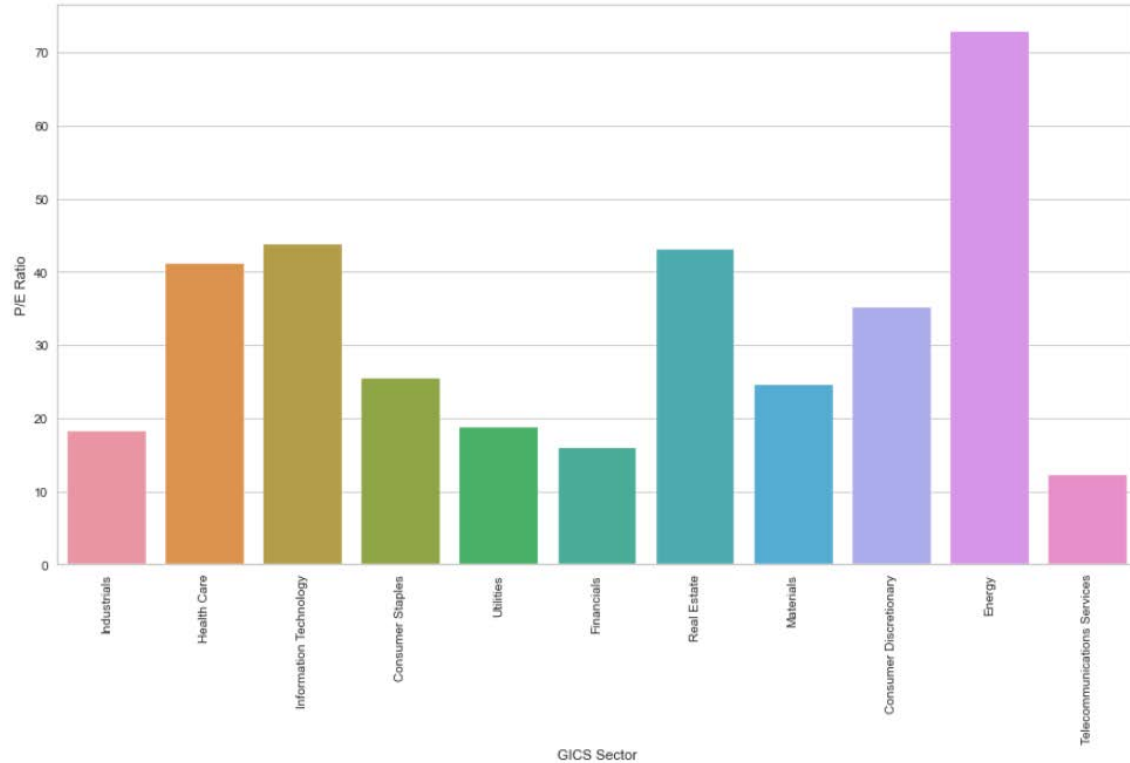


- Health Care, IT, Consumer Staples, and Telecommunication Services had the biggest price increases
- Energy had the biggest price decreases

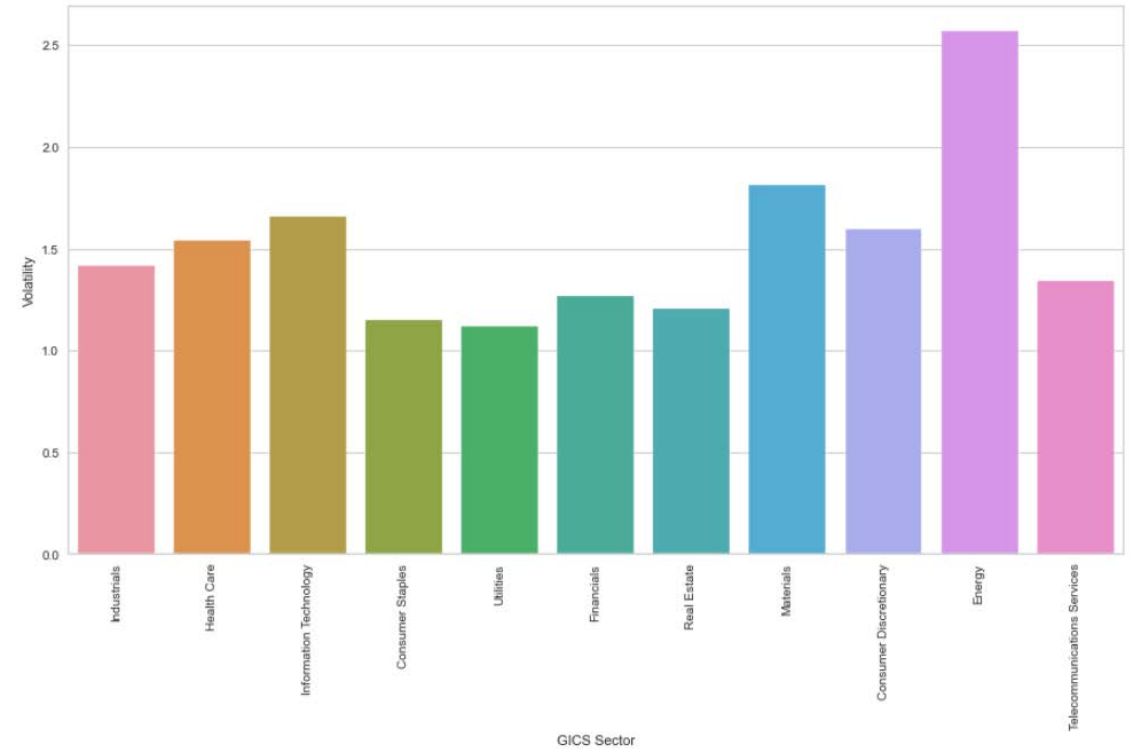


- IT, Telecommunication services, Health Care had the highest Cash Ratio
- Utilities had the lowest Cash Ratio

EDA Results

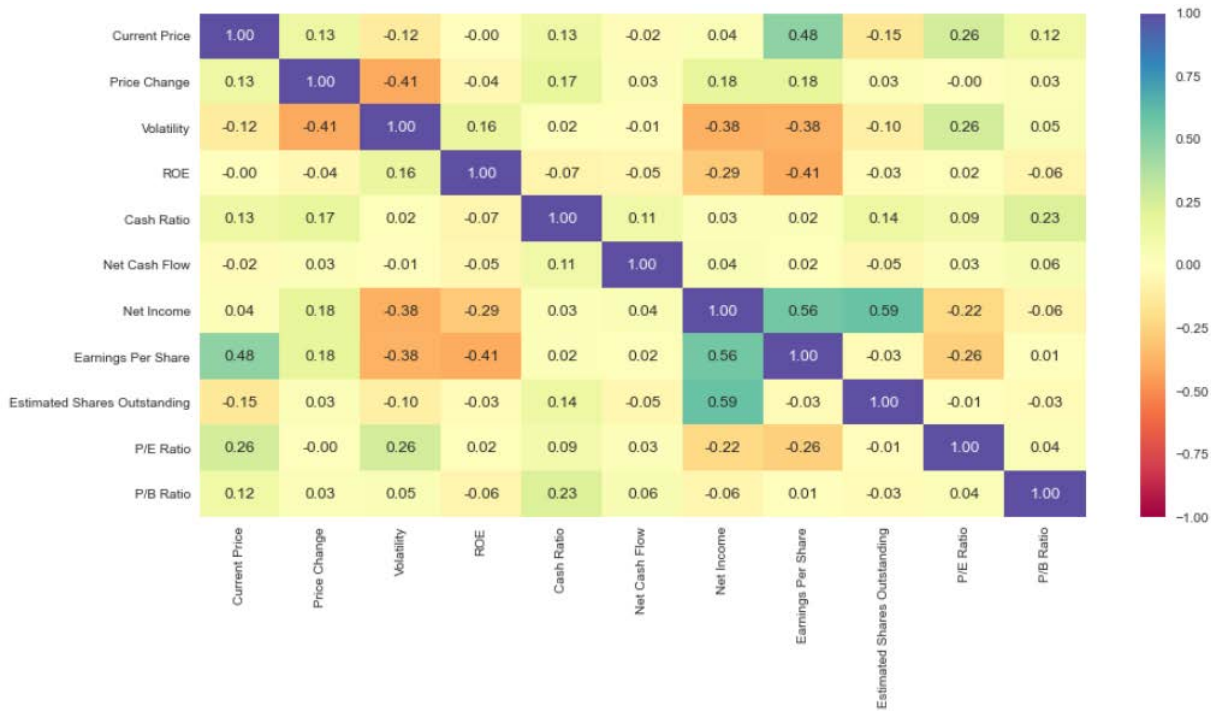


- *Energy has the highest P/E ratio and Telecommunication services has the lowest*



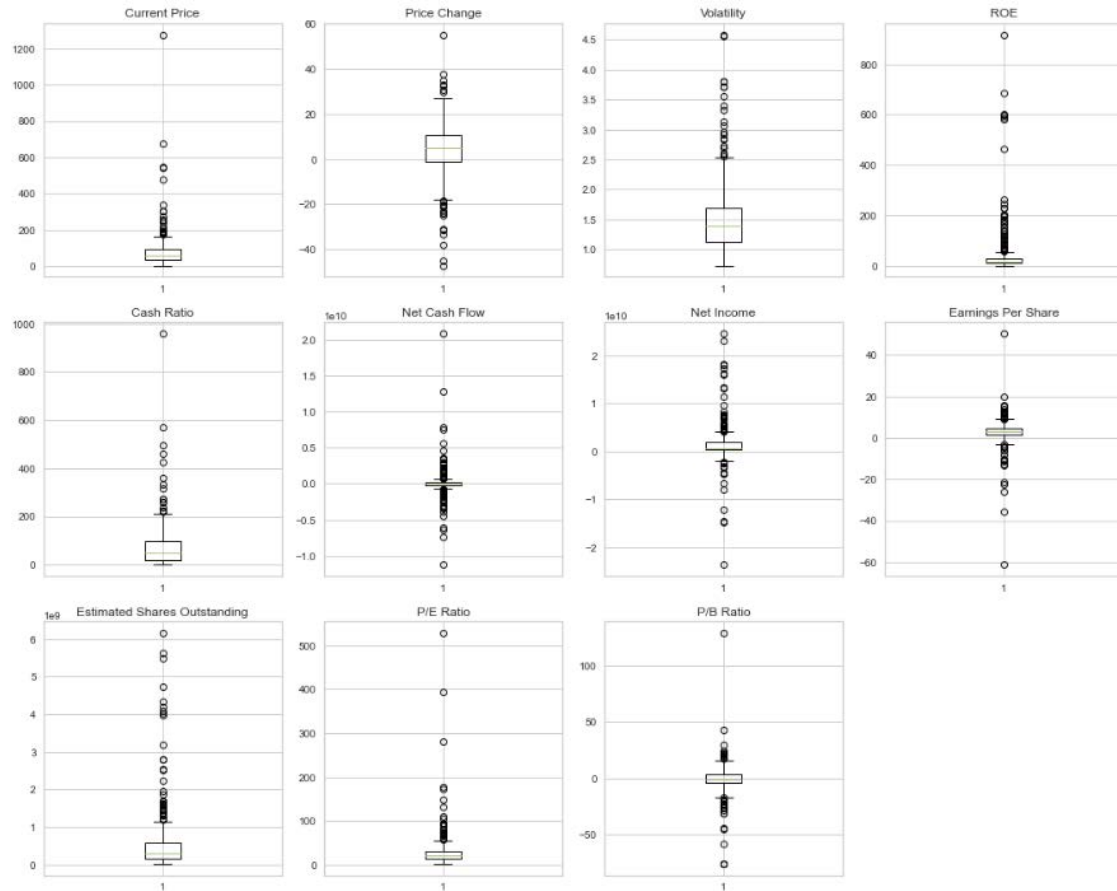
- *Energy is the most volatile stock*

EDA Results



- Net Income & Estimated Shares Outstanding have moderate correlation
- Net Income & Earnings Per Share have moderate correlation
- Current Price & Earnings Per Share have moderate correlation

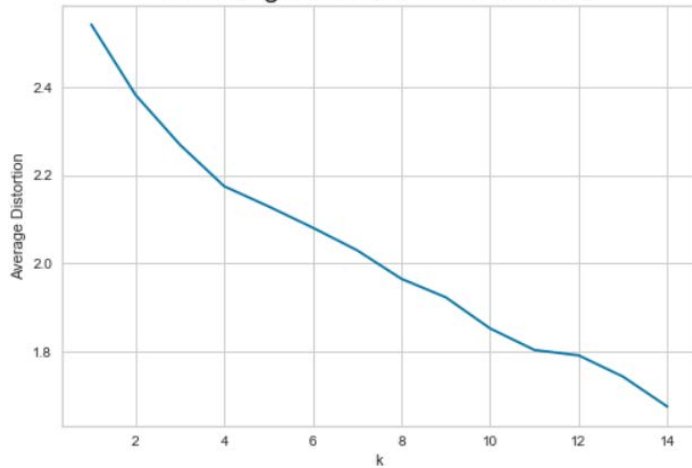
Data PreProcessing



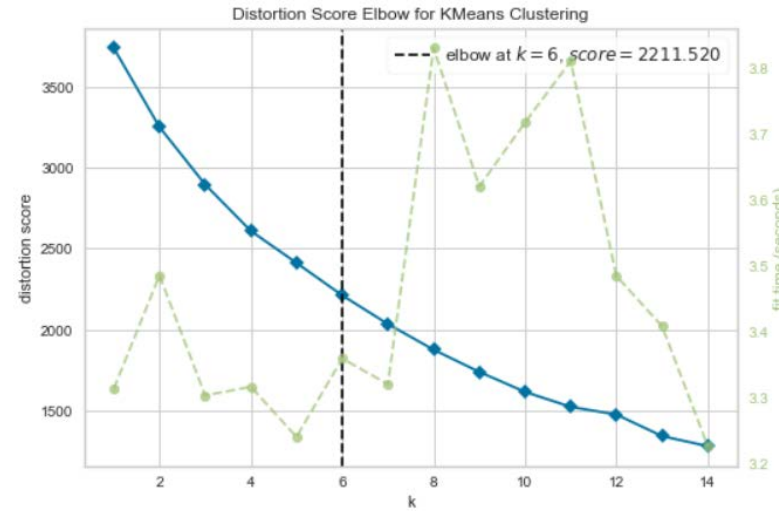
- All the numerical columns have outliers
- Outliers were kept and removed to see difference in cluster analysis
- Ultimately, outliers were kept because they didn't make much difference in cluster analysis
- Numerical columns were scaled to provide data consistency and prevent bias in cluster analysis

K-Means Clustering Analysis (Distortion)

Selecting k with the Elbow Method

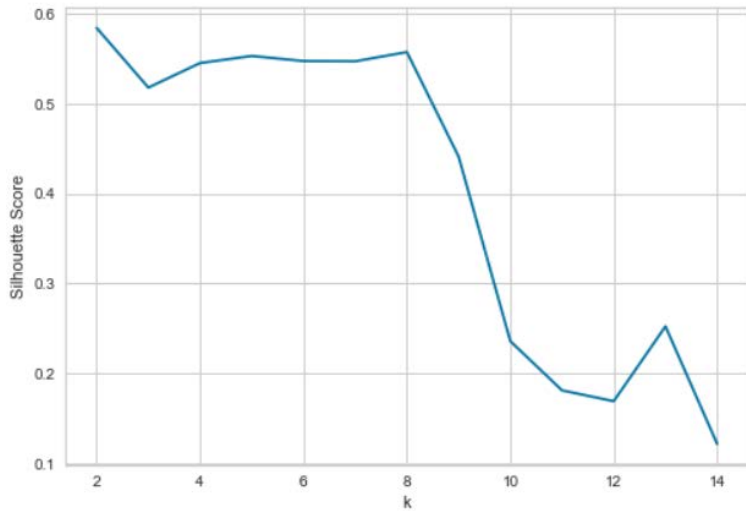


Number of Clusters: 1	Average Distortion: 2.5425069919221697
Number of Clusters: 2	Average Distortion: 2.382318498894466
Number of Clusters: 3	Average Distortion: 2.2692367155390745
Number of Clusters: 4	Average Distortion: 2.1745559827866363
Number of Clusters: 5	Average Distortion: 2.128799332840716
Number of Clusters: 6	Average Distortion: 2.080400099226289
Number of Clusters: 7	Average Distortion: 2.0289794220177395
Number of Clusters: 8	Average Distortion: 1.964144163389972
Number of Clusters: 9	Average Distortion: 1.9221492045198068
Number of Clusters: 10	Average Distortion: 1.8513913649973124
Number of Clusters: 11	Average Distortion: 1.8024134734578485
Number of Clusters: 12	Average Distortion: 1.7900931879652673
Number of Clusters: 13	Average Distortion: 1.7417609203336912
Number of Clusters: 14	Average Distortion: 1.673559857259703

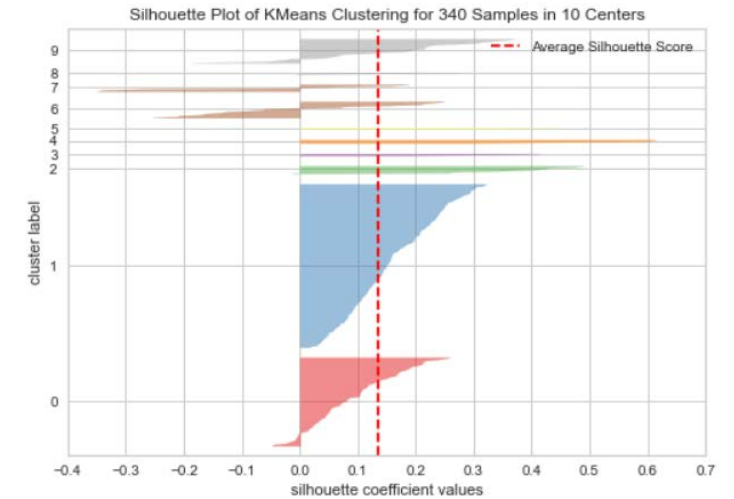
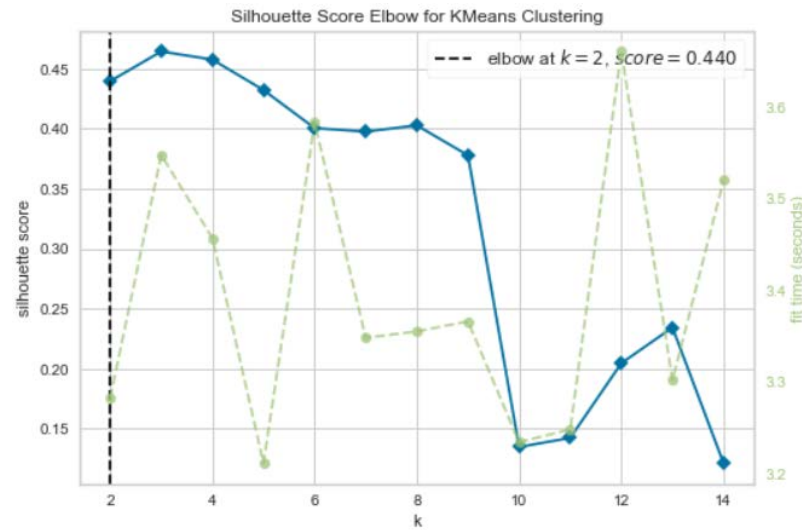


- Range of clusters established between 1 to 14
- Average distortion scores established for each cluster
- K-Means analysis conducted using Elbow Method
 - Elbows appear at k=4 and 11
- K-Elbow Visualizer shows at k=6

K-Means Clustering Analysis (Silhouette)



For `n_clusters = 2`, the silhouette score is 0.5841109092067904)
 For `n_clusters = 3`, the silhouette score is 0.5177021598641379)
 For `n_clusters = 4`, the silhouette score is 0.5450053730911077)
 For `n_clusters = 5`, the silhouette score is 0.5528840021919487)
 For `n_clusters = 6`, the silhouette score is 0.5472332275342476)
 For `n_clusters = 7`, the silhouette score is 0.5469598805975546)
 For `n_clusters = 8`, the silhouette score is 0.5572234186127173)
 For `n_clusters = 9`, the silhouette score is 0.4409247831651318)
 For `n_clusters = 10`, the silhouette score is 0.23553795483938178)
 For `n_clusters = 11`, the silhouette score is 0.18074249318178504)
 For `n_clusters = 12`, the silhouette score is 0.16873198005331133)
 For `n_clusters = 13`, the silhouette score is 0.25200178483797864)
 For `n_clusters = 14`, the silhouette score is 0.12118938181555283)



- Range of clusters established between 2 to 15
- Average silhouette scores established for each cluster
- K-Means analysis conducted using Elbow Method
 - Elbows appear at $k=10$ and 11
- K-Elbow Visualizer shows $k=2$
- Silhouette Visualizer shows at $k=10$

K-Means Clustering Analysis (Cluster Profiles)

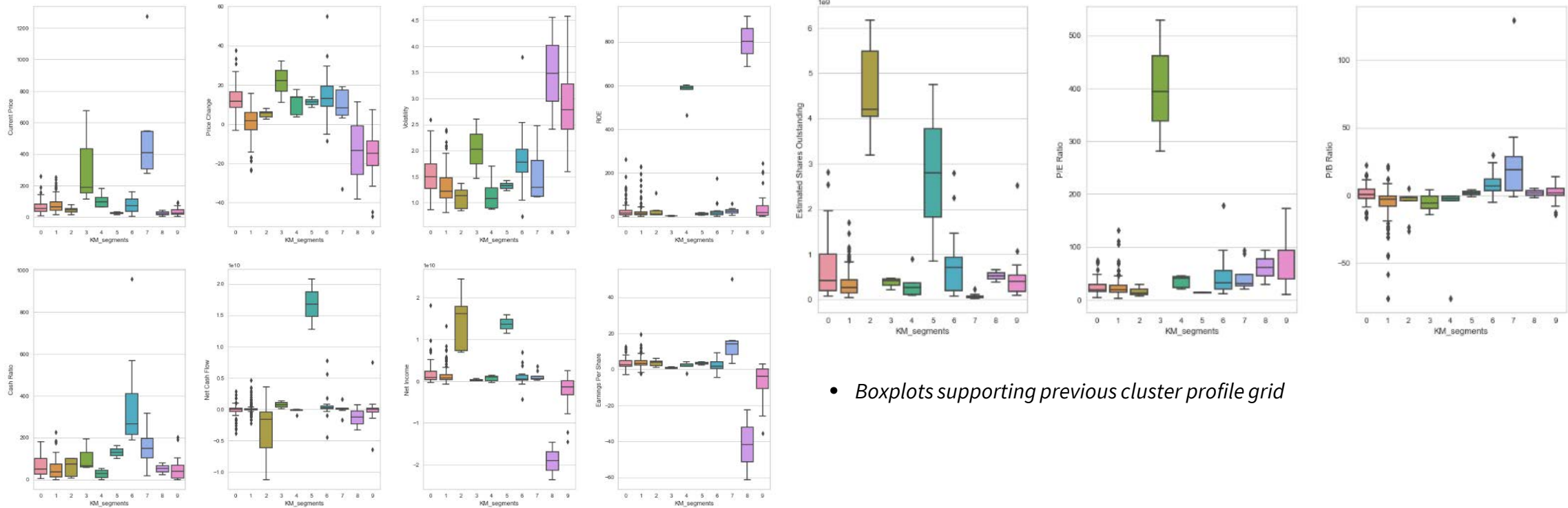
KM_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio
0	62.644030	12.720586	1.529654	29.223404	61.319149	-156258638.297872	1919175936.170213	3.399149	636937648.906064	23.345566
1	76.374133	0.834108	1.297704	23.023121	47.121387	115498173.410405	1390461699.421965	3.851069	337271215.505665	23.384698
2	46.672222	5.166566	1.079367	25.000000	58.333333	-3040666666.666667	1484844444.444445	3.435556	4564959946.222222	15.596051
3	327.006671	21.917380	2.029752	4.000000	106.000000	698240666.666667	287547000.000000	0.750000	366763235.300000	400.989188
4	108.304002	10.737770	1.165694	566.200000	26.600000	-278760000.000000	687180000.000000	1.548000	349607057.720000	34.898915
5	25.640000	11.237908	1.322355	12.500000	130.500000	16755500000.000000	13654000000.000000	3.295000	2791829362.100000	13.649696
6	75.775186	14.419381	1.854929	29.111111	338.555556	696745611.111111	935969944.444444	2.005000	792523728.361111	44.919121
7	508.534992	5.732177	1.504640	27.250000	150.875000	37895875.000000	1116994125.000000	15.965000	75654420.935000	43.727459
8	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608
9	35.263847	-16.175693	2.841300	49.769231	48.153846	-135215038.461538	-2525946153.846154	-6.514231	482428533.751538	77.817252

P/B Ratio count_in_each_segment

0.739498	94
-5.428802	173
-6.354193	9
-5.322376	3
-16.851358	5
1.508484	2
8.778016	18
29.581664	8
1.565141	2
1.618150	26

- **Cluster 0:**
✓ Moderate current price, moderate price change increase, moderate ROE, negative cash flow, and moderate EPS
- **Cluster 1:**
✓ Moderate current price, minimal price change increase, small ROE, low cash flow, and moderate EPS
- **Cluster 2:**
✓ Moderate current price, moderate price change increase, moderate ROE, negative cash flow, moderate EPS
- **Cluster 4:**
✓ Moderate to high price, moderate price increase, high ROE, negative cash flow, moderate EPS
- **Cluster 6:**
✓ Moderate current price, moderate price increase, moderate ROE, moderate cash flow, moderate EPS
- **Cluster 7:**
✓ Expensive current price, low to moderate price increase, moderate ROE, decent cash flow, high EPS
- **Cluster 9:**
✓ Low to moderate current price, steep price drop, moderate ROE, negative cash flow, negative EPS

K-Means Clustering Analysis (Cluster Profiles)



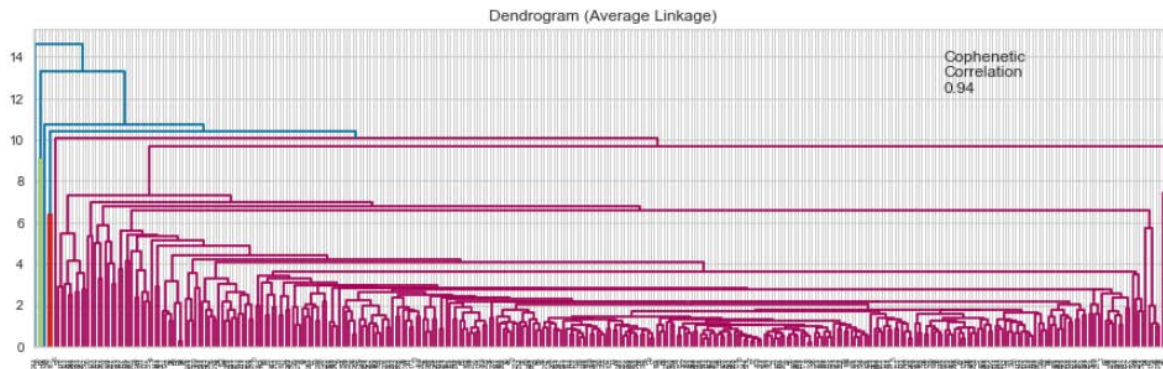
- *Boxplots supporting previous cluster profile grid*

Hierarchical Clustering Analysis

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 0.9259195530524591.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.7925307202850002.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159737.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8708317490180428.
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.

Highest cophenetic correlation is 0.9422540609560814, which is obtained with Euclidean distance and average linkage.

Linkage	Cophenetic Coefficient
4 ward	0.710118
1 complete	0.787328
5 weighted	0.869378
0 single	0.923227
3 centroid	0.931401
2 average	0.942254



- Highest cophenetic correlation is obtained using Euclidean distance and average linkage at 0.94
- Dendrogram shows $k=9$

Hierarchical Clustering Analysis (Cluster Profiles)

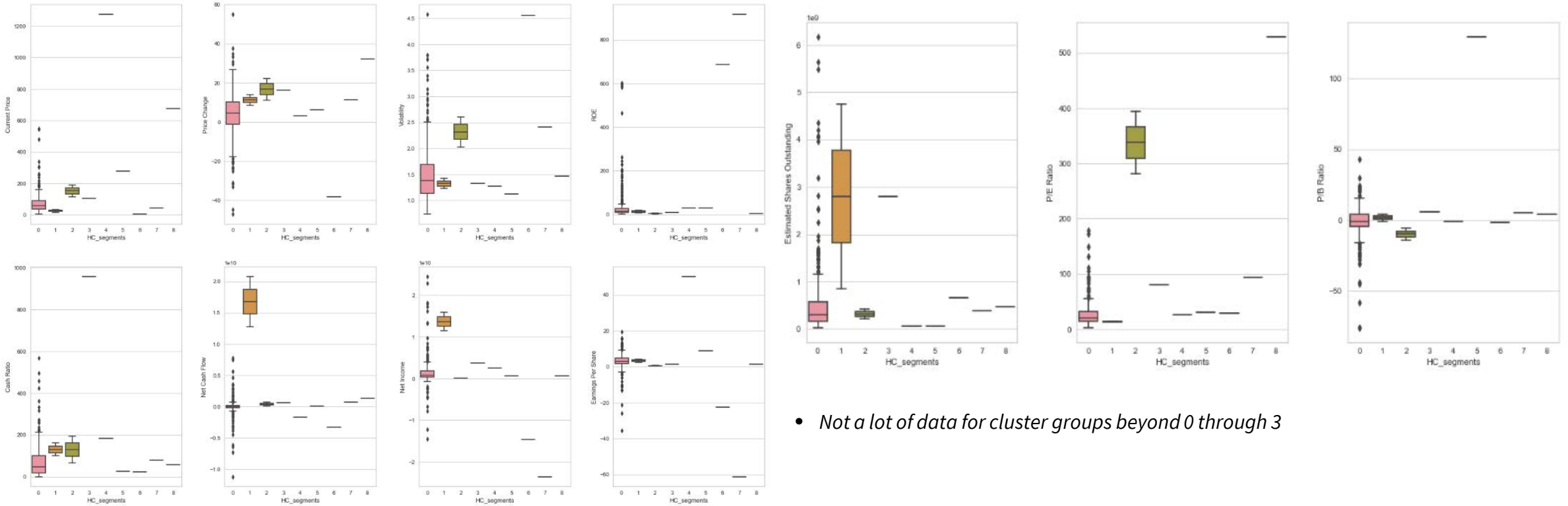
HC_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio
0	75.017416	3.937751	1.513415	35.621212	66.545455	-39846757.575758	1549443100.000000	2.904682	562266326.402576	29.091275
1	25.640000	11.237908	1.322355	12.500000	130.500000	16755500000.000000	13654000000.000000	3.295000	2791829362.100000	13.649696
2	152.564999	16.742017	2.314435	4.000000	130.000000	380861000.000000	133320500.000000	0.485000	317332352.950000	337.464244
3	104.660004	16.224320	1.320606	8.000000	958.000000	592000000.000000	3669000000.000000	1.310000	2800763359.000000	79.893133
4	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000	50.090000	50935516.070000	25.453183
5	276.570007	6.189286	1.116976	30.000000	25.000000	90885000.000000	596541000.000000	8.910000	66951851.850000	31.040405
6	4.500000	-38.101788	4.559815	687.000000	22.000000	-3283000000.000000	-14685000000.000000	-22.430000	654703522.100000	28.407929
7	44.470001	11.397804	2.405408	917.000000	80.000000	698000000.000000	-23528000000.000000	-61.200000	384444444.400000	93.089287
8	675.890015	32.268105	1.460386	4.000000	58.000000	1333000000.000000	596000000.000000	1.280000	465625000.000000	528.039074

P/B Ratio count_in_each_segment

-2.146308	330
1.508484	2
-9.935778	2
5.884467	1
-1.052429	1
129.064585	1
-1.840528	1
4.970809	1
3.904430	1

- Majority of data in Cluster 0
- Hierarchical Cluster Analysis didn't perform well

Hierarchical Clustering Analysis(Cluster Profiles)



- *Not a lot of data for cluster groups beyond 0 through 3*

Insights & Recommendations



Insights:

- K-Means Clustering & Hierarchical Clustering
 - ✓ K-Elbow Visualizer (Distortion) shows at k=6
 - ✓ K-Elbow Visualizer (Silhouette) shows at k=10
 - ✓ Hierarchical Dendrogram shows at k=9
- Hierarchical Clustering did not have good variation of data and had most of data in Cluster 0
- K-Means Clustering technique took longer to execute but provided better clusters
- K-Means (Silhouette) Visualizer recommended clusters at k=10; only 7 of them had enough data to be used
- Cluster 5 companies had low current price, moderate to high price increase, and good earnings per share
- Cluster 7 companies had an expensive current price, low to moderate price increase, and good earnings per share
- Cluster 8 companies had low price, considerable price decreases, and negative earnings per share
- Most of companies in Cluster 5 were in the Financial GICS Sector
- Most of companies in Cluster 7 were in the Health Care GICS Sector
- Most of the companies in Cluster 8 were in the Energy GICS Sector

Recommendations:

- Inform customers about Cluster 5 companies with low current price and moderate to high increases. This can help increase value of portfolio.
- Inform customers about Cluster 7 companies with expensive current price but have low to moderate price increases. This can help customers save money.
- Warn customers about Cluster 8 companies and the Energy GICS Sector; market dynamics has caused prices to decrease significantly.



THANK YOU

Josh Willis, PGP-DSBA