Trade&Ahead Investment Consulting

PGP-DSBA _ Trade&Ahead Project

February 29, 2024

Contents / Agenda

- Executive Summary
- Business Problem Overview
- Solution Approach
- EDA Results
- Data Preprocessing
- K-Means Clustering
- Hierarchical Clustering
- Appendix

Executive Summary

- Considering execution time, relative clustering performance measured, and number of clusters, K-Means 4 cluster option appears to be a preferable choice to the clusters obtained using hierarchical clustering
- Investment in KM_Segment 1 appears to be a safe bet considering the associated low to moderate current price of shares and high earnings per share; however, the benefits should be weighed against the relatively low to moderate ROE and Cash Ratio
- KM_Segment 2 also has a low to moderate price and thus relatively affordable shares. However, compared to earnings, the stocks might be considered expensive due to the high P/E ratio. Moreover, the low Cash Ratio and Earnings Per Share might be further reason to be careful about investment in this cluster
- KM_Segment 3 is characterized by high Current Price, Price Change, and P/B ratio. These characteristics render investment in this cluster expensive and risky. Though the Cash Ratio is high, the related benefit should be measured against the low ROE that characterizes the cluster.
- KM_Segment 3 is a low-risk, low-return cluster since all the reported metrics are low to moderate. It would be advisable for an investor to make minimal investment in this sector, if interested; however, the investment should be spread across several companies to further take advantage of the low-risk that characterizes the cluster

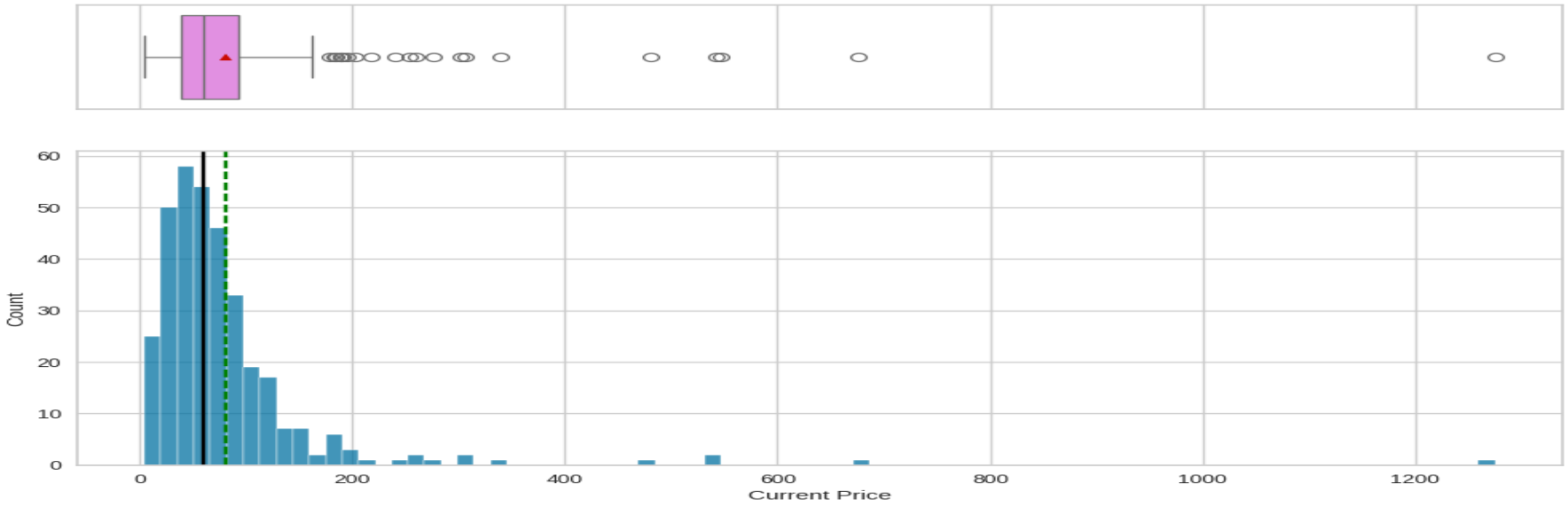
Business Problem Overview

- For several years now, investing in the stock market has demonstrably proven to be a good investment option
- Portfolio diversification is a critical consideration when contemplating or analyzing the purchase of stocks; on average, diversification reduces investment risk
- Several financial metrics characterize securities in the stock market and influence the profitability of investments
- I have been hired by Trade&Ahead, a financial consultancy firm, to help analyze some collected data on stock price and related financial metrics for a number of companies listed under the New York Stock Exchange
- The analysis comprises grouping the stocks using available attributes, providing insights on group characteristics and, thus, helping Trade&Ahead provide personalized investment guidance to its clients

Solution Approach

- Steps taken to accomplish the task:
  - ❖ General Analysis of the data to uncover information such as shape, data types, and statistical summary
  - ❖ Exploratory Data Analysis to explore the distributions of individual attributes as well as interaction between their respective variables
  - ❖ Data Preprocessing, specifically scaling using the Standard Scaler, in preparation for clustering using various cluster algorithms
  - ❖ Use of K-means and Hierarchical clustering to group the stocks into several number of clusters and evaluating the clusters using the Elbow Plot and Silhouette Scores for K-means and Dendrograms and Cophenetic Correlation for Hierarchical clustering
  - ❖ Finally, exploration of chosen clustering options and uncovering of insights such as the sizes and constitution of the clusters and the performance of the provided metrics across the clusters

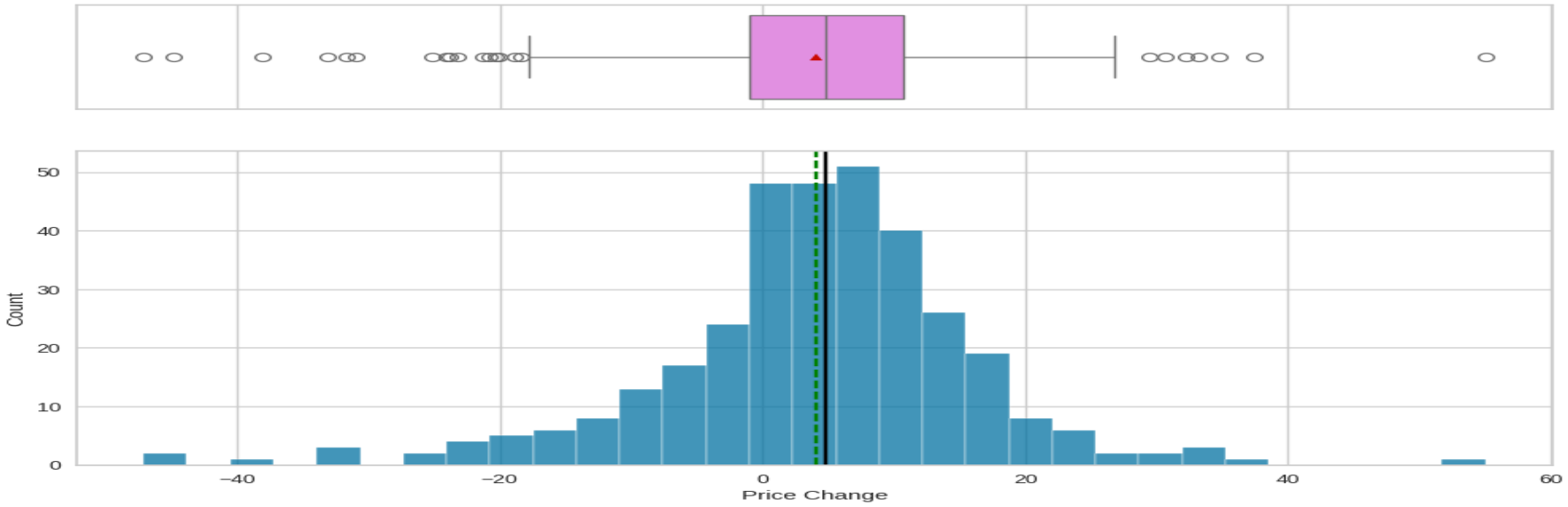
EDA Results: Univariate Analysis: Current Price



- We have confirmation that the distribution of Current Price is right-skewed with several upper outliers
- The mean, about 80, is 20 greater than the median

Link to Appendix slide on data background check

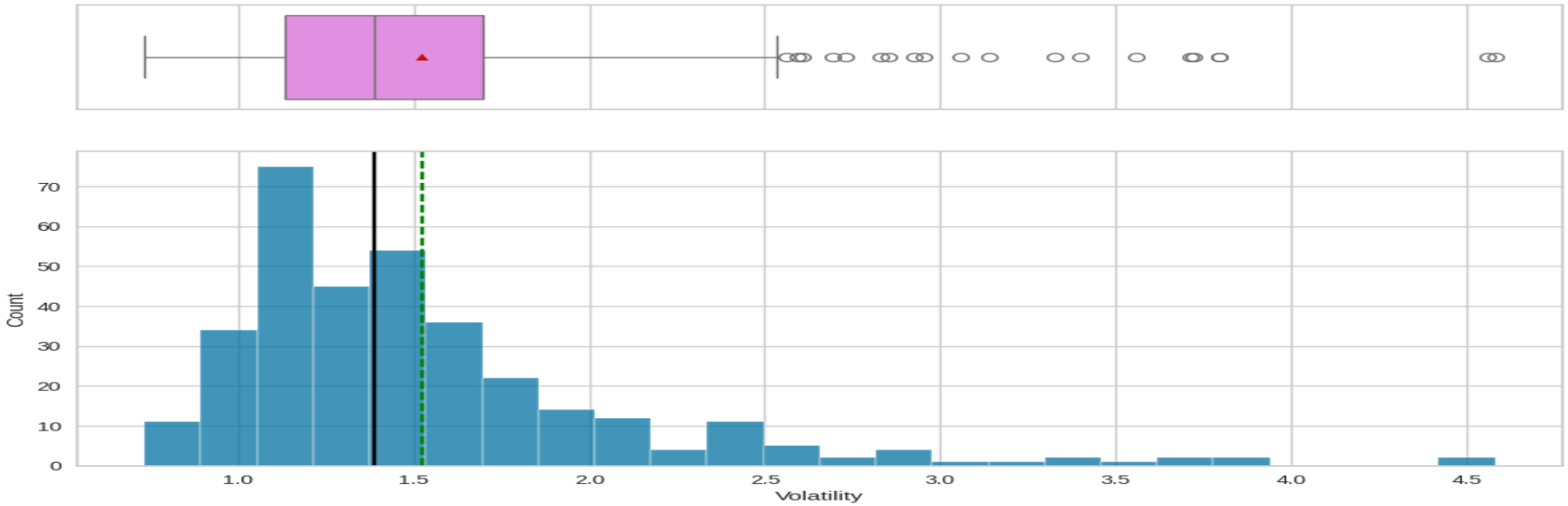
EDA Results: Univariate Analysis: Price Change



- Price Change is left-skewed but slightly symmetrical with several outliers on both ends of the distribution
- The mean, 4, is 1 lower than the median

Link to Appendix slide on data background check

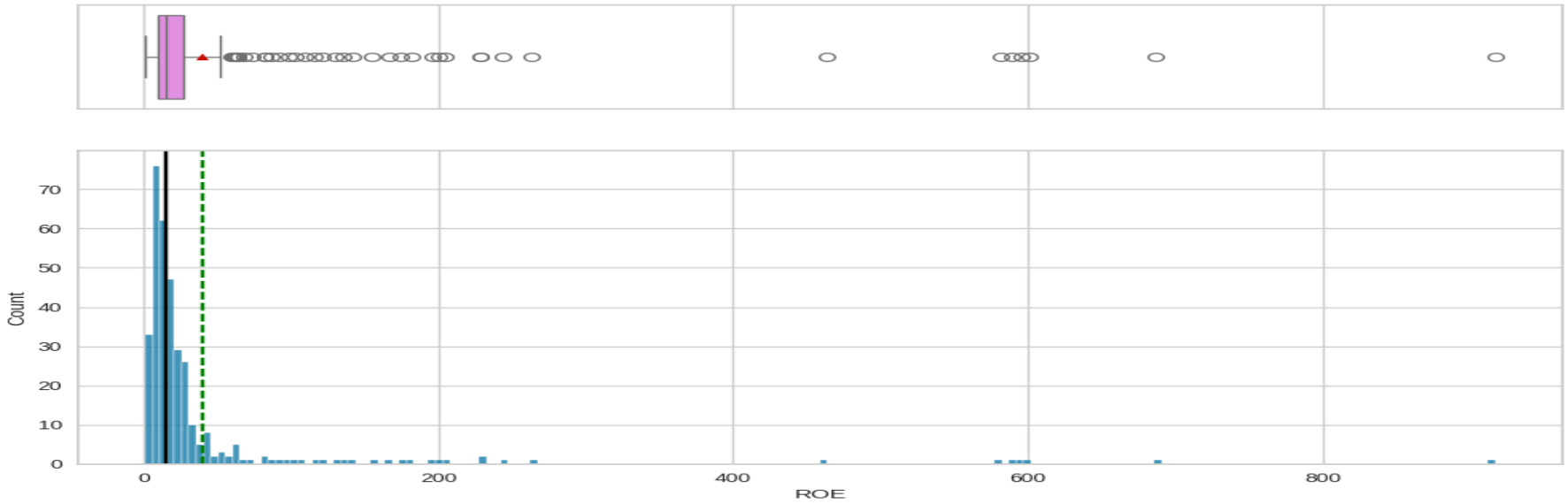
EDA Results: Univariate Analysis: Volatility



- Volatility is heavily right-skewed with several upper outliers
- The mean, about 1.5, is approximately 0.16 greater than the median

Link to Appendix slide on data background check

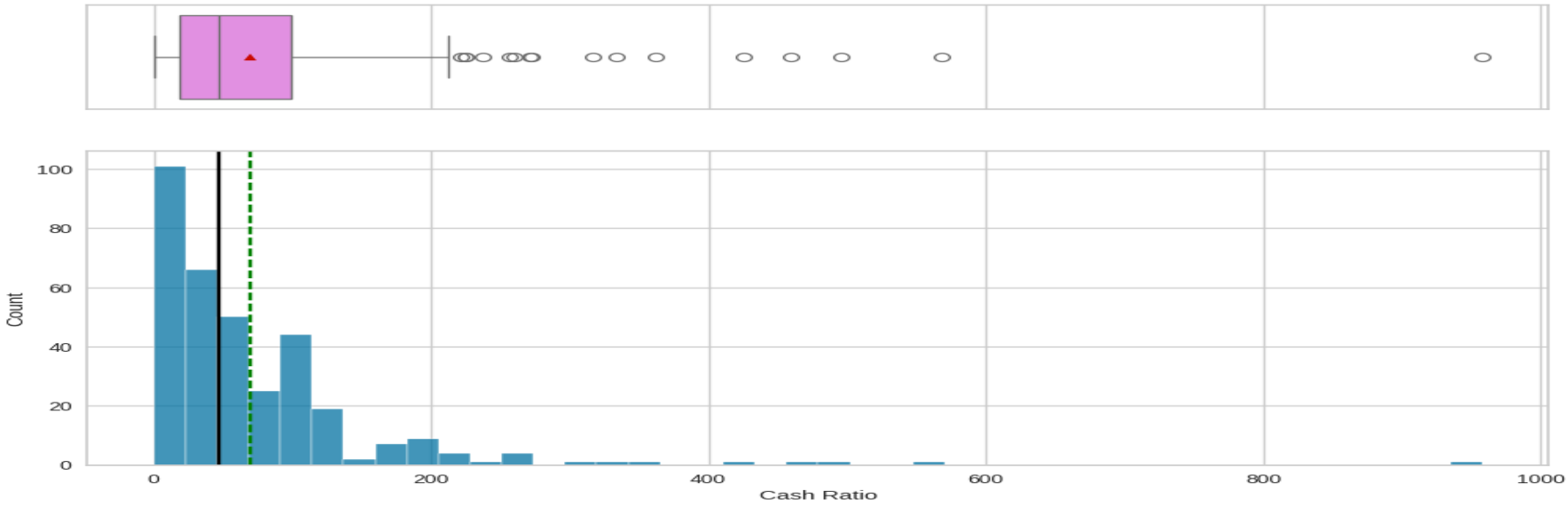
EDA Results: Univariate Analysis: ROE



- The ROE is heavily right-skewed with several upper outliers
- The mean, 40, is 25 greater than the median

Link to Appendix slide on data background check

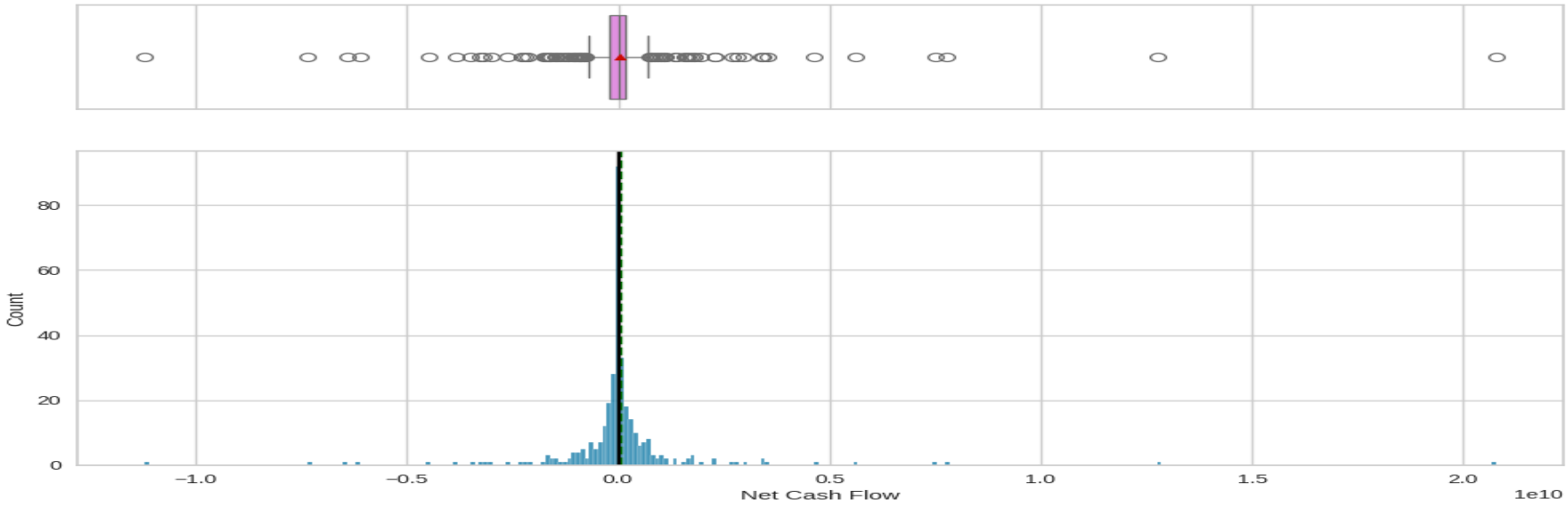
EDA Results: Univariate Analysis: Cash Ratio



- The cash ratio is right-skewed with several upper outliers
- The mean, 70, is 23 greater than the median

Link to Appendix slide on data background check

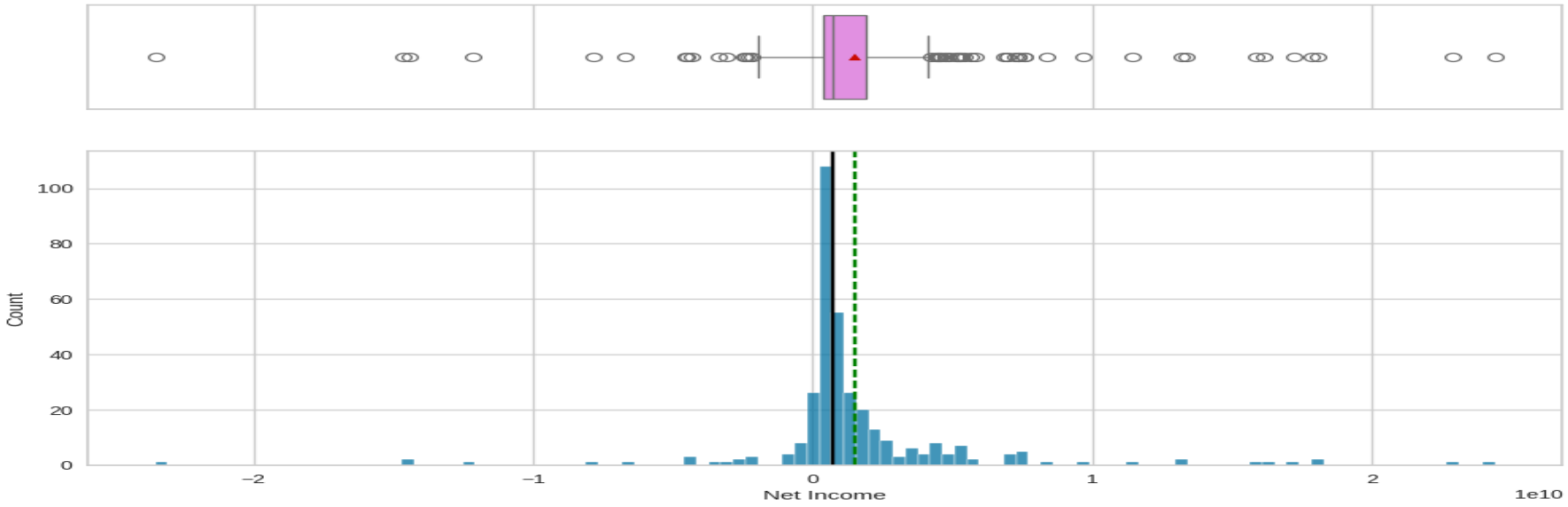
EDA Results: Univariate Analysis: Net Cash Flow



- Though the mean of the Net Cash Flow is about 56 M and the median about 2 M, the distribution is approximately normal with several outliers on both sides

Link to Appendix slide on data background check

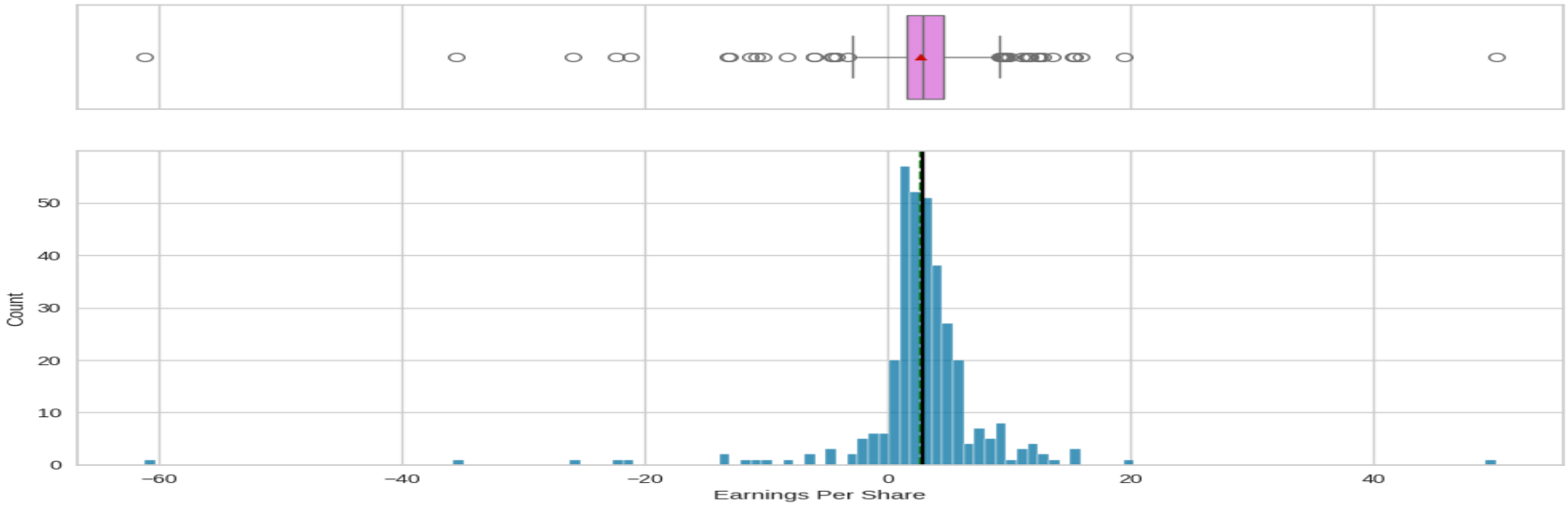
EDA Results: Univariate Analysis: Net Income



- The distribution of Net Income displays some level of symmetry but is right-skewed and has several outliers on both sides of the distribution
- The mean, about 1.5 B is approximately twice the median, about 700 M

Link to Appendix slide on data background check

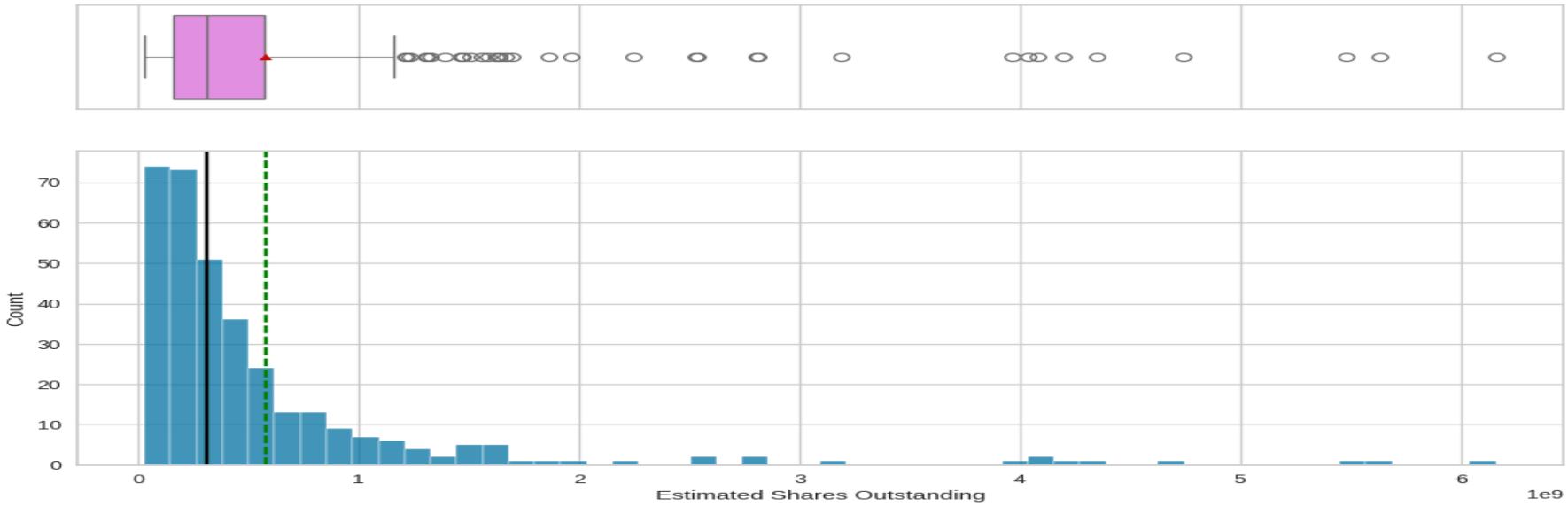
EDA Results: Univariate Analysis: Earnings Per Share



- The Earnings Per Share is approximately normal with several outliers on both sides of the distribution
- The mean, 2.8, is just about 0.1 less than the median

Link to Appendix slide on data background check

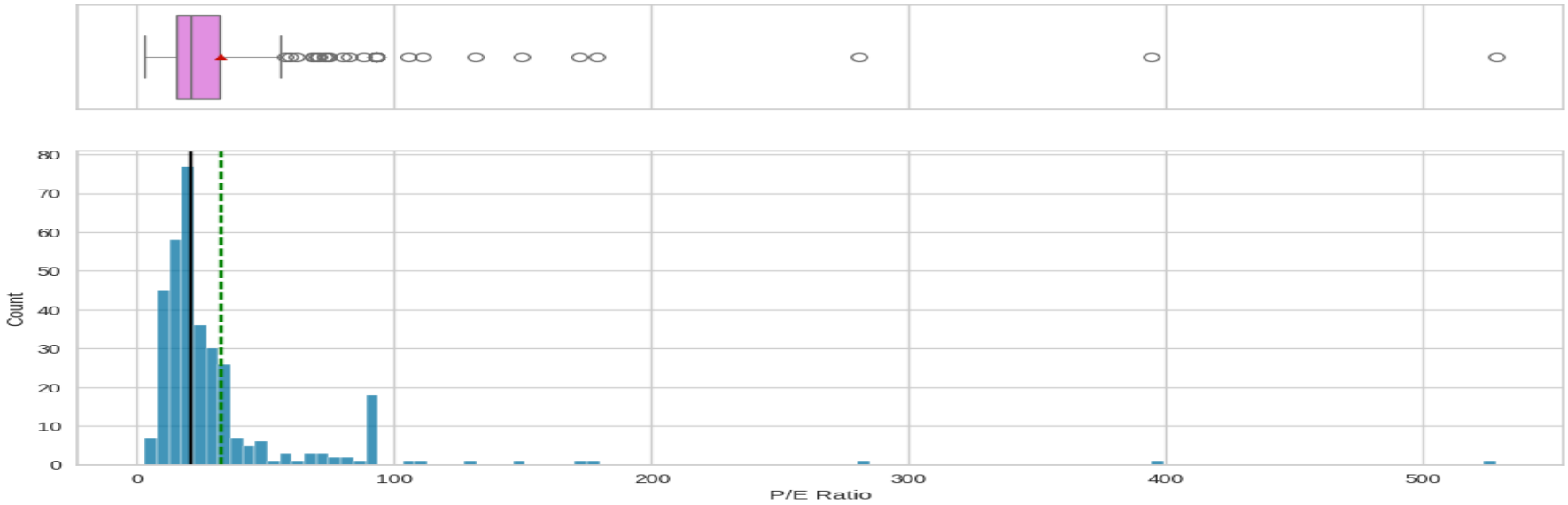
EDA Results: Univariate Analysis: Estimated Shares Outstanding



- The Estimated Shares Outstanding is right-skewed with several upper outliers
- The mean, about 580 M, is greater than the median, about 310 M

Link to Appendix slide on data background check

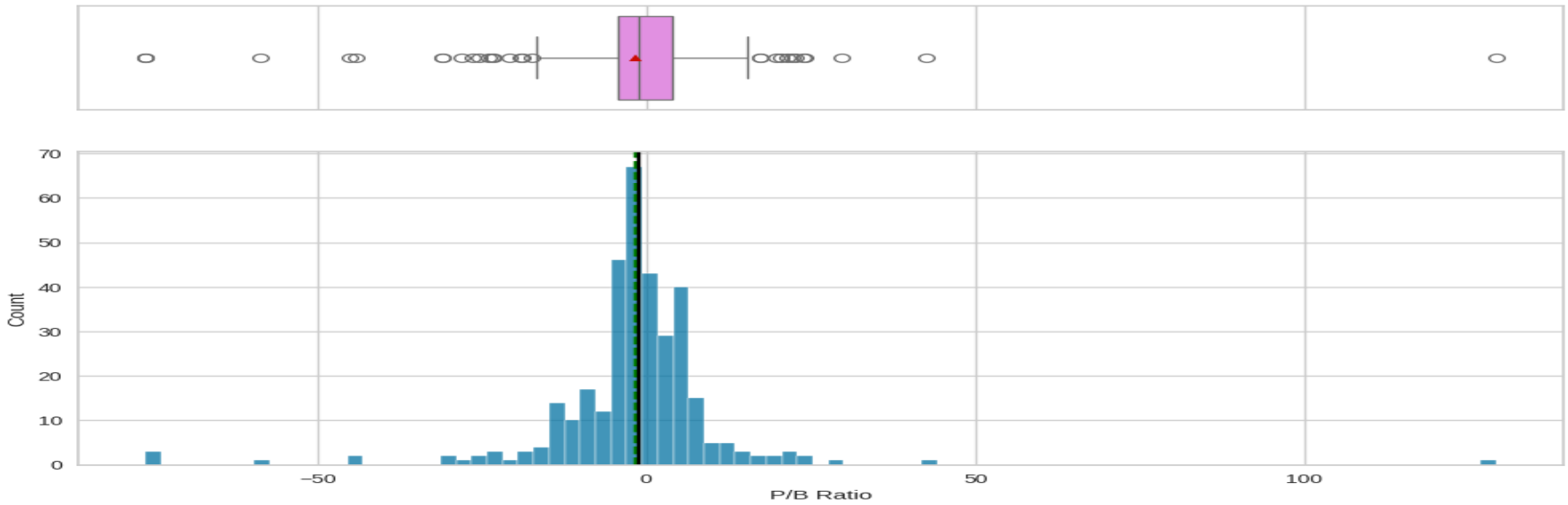
EDA Results: Univariate Analysis: P/E Ratio



- The P/E Ratio is heavily right-skewed with several upper outliers
- Peak observed between 80 and 100
- The mean, about 33, is approximately 12 greater than the median

Link to Appendix slide on data background check

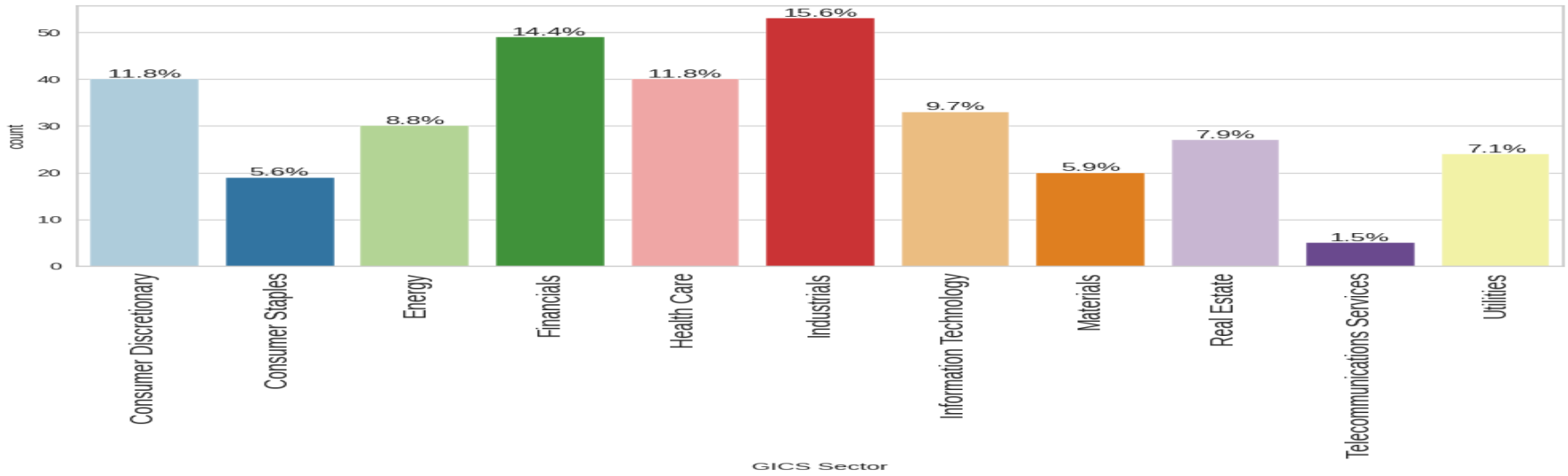
EDA Results: Univariate Analysis: P/B Ratio



- The P/B Ratio is approximately normally distributed with several outliers on both sides of the distribution
- The mean, about -2, is 1 less than the median

Link to Appendix slide on data background check

EDA Results: Univariate Analysis: GICS Sector

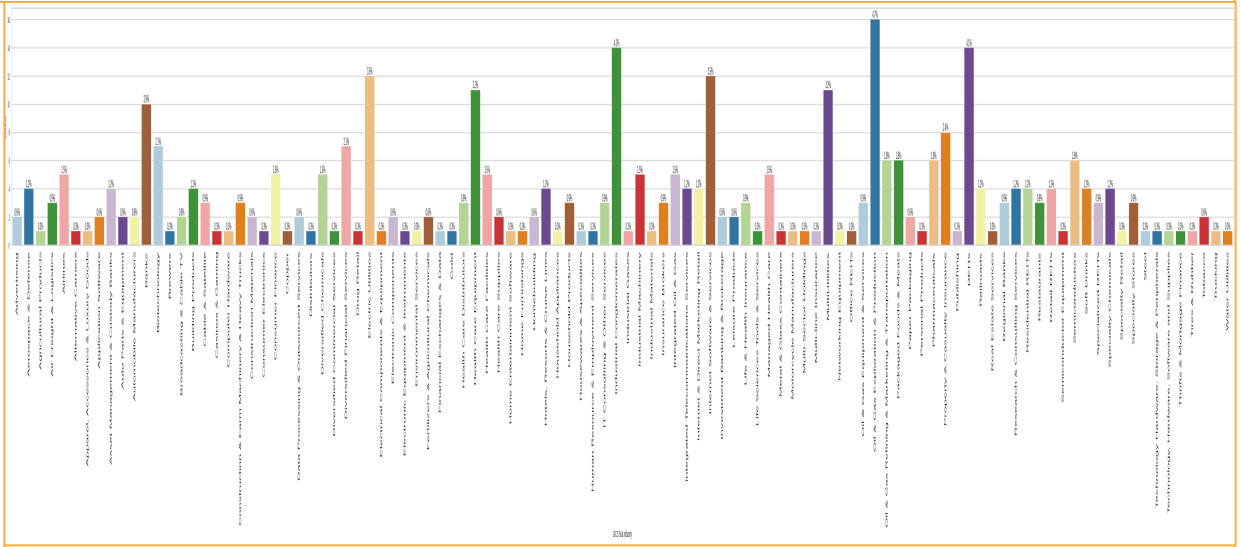


- Industrials, at 15.6%, constitute the most popular GICS Sector, followed but Financials (14.4%), Consumer Discretionary and Health Care (11.8% each)
- Telecommunications Services constitute the least popular sector (1.5%)

Link to Appendix slide on data background check

EDA Results: Univariate Analysis: GICS Sub Industry

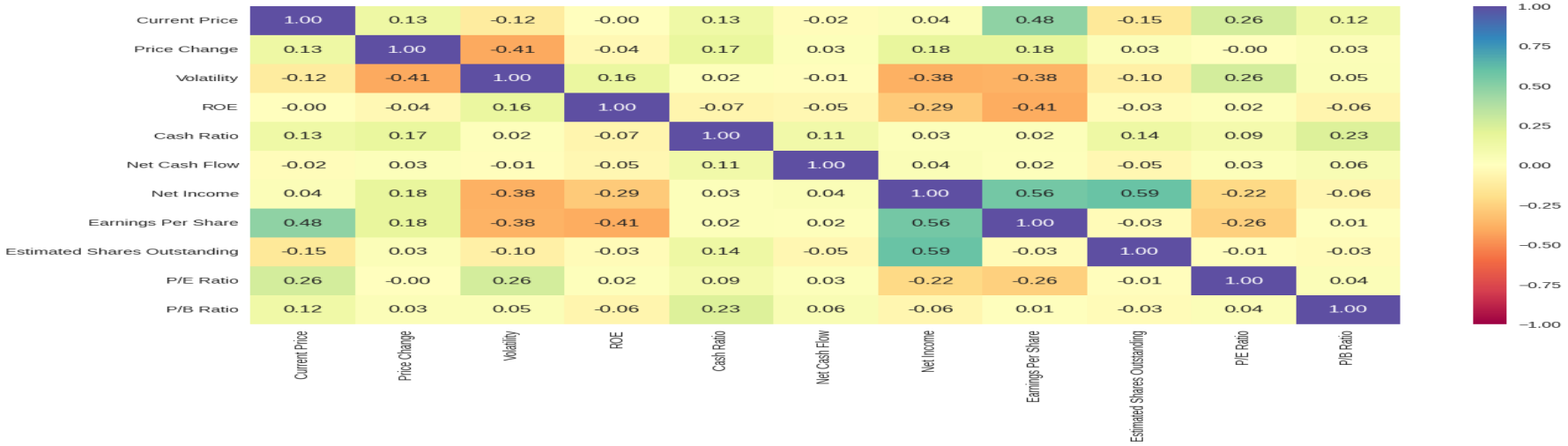
```
Oil & Gas Exploration & Production                 0.047059
REITs                                              0.041176
Industrial Conglomerates                           0.041176
Electric Utilities                                 0.035294
Health Care Equipment                              0.033333
Multi-Utilities                                    0.033333
Banks                                              0.029412
Property & Casualty Insurance                      0.023529
Diversified Financial Services                     0.020588
Pharmaceuticals                                    0.017647
Packaged Foods & Meats                             0.017647
Oil & Gas Refining & Marketing & Transportation    0.017647
Semiconductors                                     0.017647
Diversified Chemicals                              0.014706
Consumer Finance                                   0.014706
Integrated Oil & Gas                               0.014706
Industrial Machinery                               0.014706
Managed Health Care                                0.014706
Airlines                                           0.014706
Health Care Facilities                             0.014706
Internet & Direct Marketing Retail                 0.011765
Residential REITs                                  0.011765
Soft Drinks                                        0.011765
Research & Consulting Services                     0.011765
Asset Management & Custody Banks                   0.011765
Specialty Chemicals                                0.011765
Aerospace & Defense                                0.011765
Integrated Telecommunications Services             0.011765
Building Products                                  0.011765
Hotels, Resorts & Cruise Lines                     0.011765
Restaurants                                        0.008824
Cable & Satellite                                  0.008824
Electric Utilities & Statistics                    0.008824
Household Products                                 0.008824
Regional Banks                                     0.008824
IT Consulting & Other Services                     0.008824
Construction & Farm Machinery & Heavy Trucks       0.008824
Life & Health Insurance                            0.008824
Health Care Distributors                           0.008824
Gas Utilities                                      0.008824
Oil & Gas Equipment & Services                     0.008824
Insurance Brokers                                  0.008824
Specialty Stores                                   0.008824
Fertilizers & Agricultural Chemicals               0.005882
Tobacco                                            0.005882
Leisure Products                                   0.005882
Advertising                                        0.005882
Data Processing & Outsourced Services              0.005882
Paper Packaging                                    0.005882
Construction Materials                             0.005882
Application Software                               0.005882
Specialty Chemicals                                0.005882
Automobile Manufacturers                           0.005882
Investment Banking & Brokerage                     0.005882
Broadcasting & Cable TV                            0.005882
Auto Parts & Equipment                             0.005882
Health Care Supplies                               0.005882
Electronic Components                              0.005882
Gold                                               0.002941
Environmental Services                             0.002941
Home Furnishings                                   0.002941
Household Appliances                               0.002941
Apparel, Accessories & Luxury Goods                0.002941
Specialty Retail                                   0.002941
Life Sciences Tools & Services                     0.002941
Publishing                                         0.002941
Human Resource & Employment Services               0.002941
Steel                                              0.002941
Housewares & Specialties                           0.002941
Thrifts & Mortgage Finance                         0.002941
Brewers                                            0.002941
Technology Hardware, Software and Supplies         0.002941
Personal Products                                  0.002941
Industrial Gases                                   0.002941
Office REITs                                       0.002941
Multi-Sector Holdings                              0.002941
Alternative Carriers                               0.002941
Computer Hardware                                  0.002941
Distributors                                       0.002941
Agricultural Products                              0.002941
Metal & Glass Containers                           0.002941
Financial Exchanges & Data                         0.002941
Water Utilities                                    0.002941
Home Entertainment Software                        0.002941
Drug Retail                                        0.002941
Electrical Components & Equipment                  0.002941
Semiconductor Equipment                            0.002941
Multi-Line Insurance                               0.002941
Copper                                             0.002941
Electronic Equipment & Instruments                 0.002941
Diversified Commercial Services                    0.002941
Consumer Electronics                               0.002941
Tires & Rubber                                     0.002941
Industrial Materials                               0.002941
Motorcycle Manufacturers                           0.002941
Technology Hardware, Storage & Peripherals         0.002941
Real Estate Services                               0.002941
Trucking                                           0.002941
Networking Equipment                               0.002941
Casinos & Gaming                                   0.002941
Name: GICS Sub Industry, dtype: float64
```



- Oil & Gas Exploration & Production (4.7%) is the most popular GICS Sub Industry followed by REITs and Industrial Conglomerates (4.1% each)
- 42 GICS Sub Industries occupy the lowest rank of popularity (0.3% each)

Link to Appendix slide on data background check

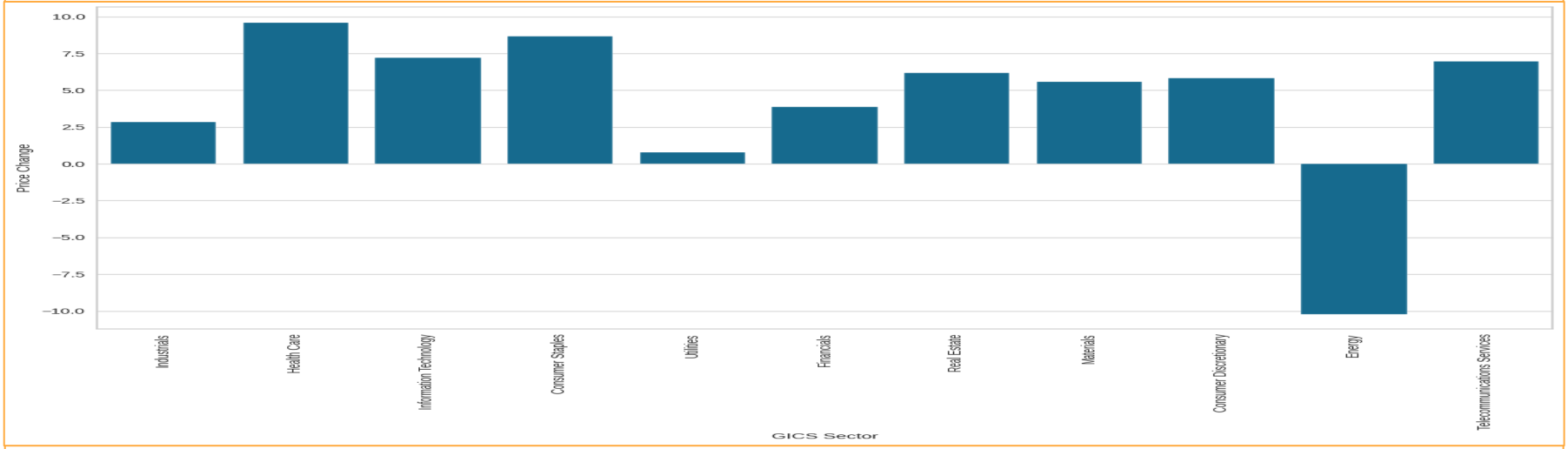
EDA Results: Bivariate Analysis: Correlation Check



- A fair amount of positive correlation is observed between Net Income on the one hand and Earnings Per Share and Estimated Shares Outstanding on the other; a little less between Earnings Per Share and Current Price
- Weaker negative correlations are observed between Volatility on the one hand and Net Income and Earnings Per Share on the other, between Price Change and Volatility, and between ROE and Earnings Per Share

Link to Appendix slide on data background check

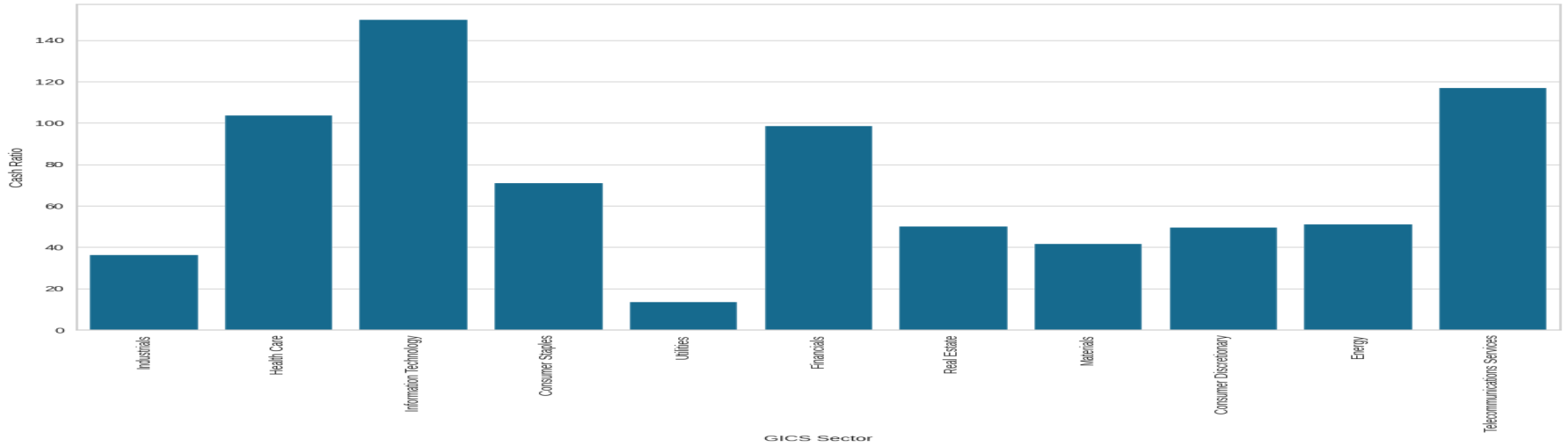
EDA Results: Bivariate Analysis: Price Change vs GICS Sector



- Health Care registered the highest percentage price increase in 13 weeks followed by Consumer Staples and then Information Technology
- Energy is the only GICS Sector that registered a percentage price decrease

Link to Appendix slide on data background check

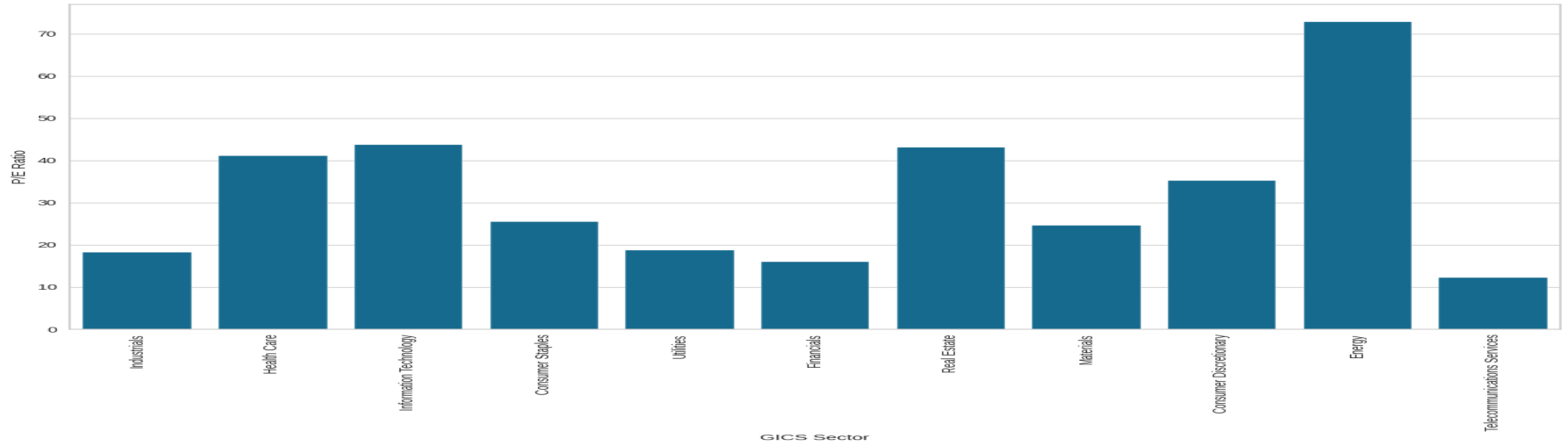
EDA Results: Bivariate Analysis: Cash Ratio vs GICS Sector



- The highest cash ratio was registered for Information technology (between 140 and 160) followed by Telecommunications Services (between 110 and 120), Health Care (between 100 and 110), and Financials (between 90 and 100)
- The lowest cash ratio was registered for Utilities (about 10) followed by Industrials (between 30 and 40)

Link to Appendix slide on data background check

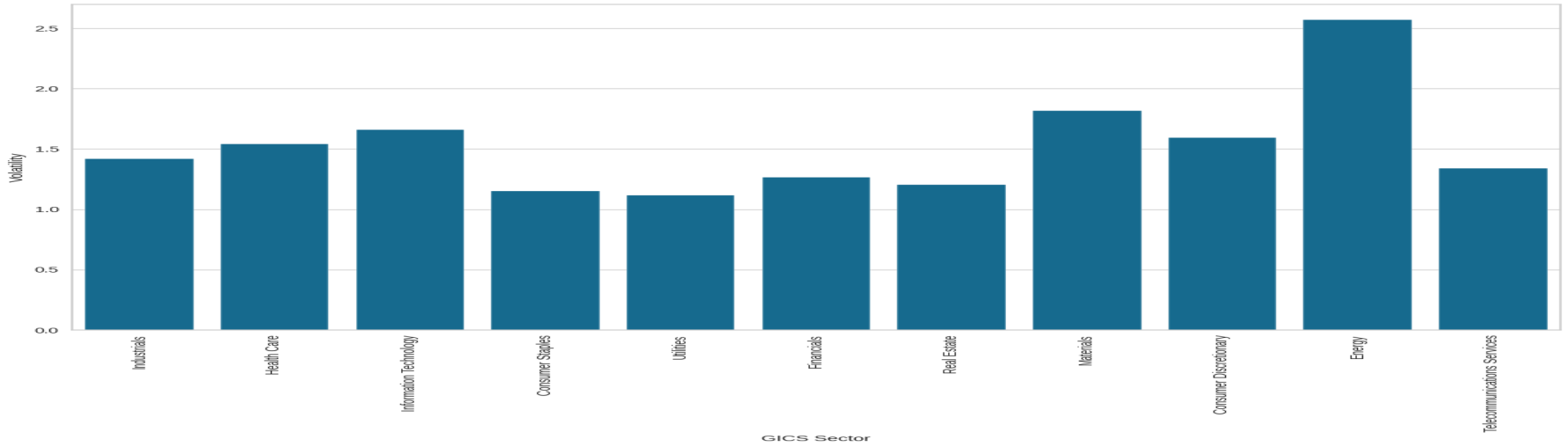
EDA Results: Bivariate Analysis: P/E Ratio vs GICS Sector



- Energy, the only GICS Sector that registered a percentage price decrease, has, by far, the highest P/E Ratio (over 70) followed by Real Estate, Information Technology, and Health Care (each between 40 and 45)
- Telecommunications Services registered the lowest P/E Ratio (between 10 and 15)

Link to Appendix slide on data background check

EDA Results: Bivariate Analysis: Volatility vs GICS Sector

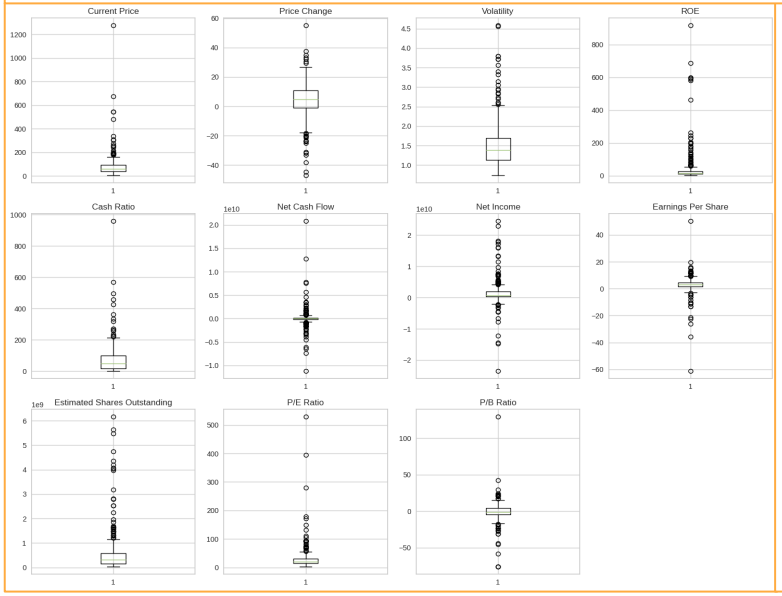


- Once again, Energy enjoys the top position with a volatility rating of over 2.5 followed by Materials (between 1.75 and 2.0) and Information Technology, Consumer Discretionary, and Health Care (all between 1.5 and 1.75)
- The lowest ranks are occupied by Utilities, Consumer Staples, and Real Estate (each between 1.0 and 1.25) followed by Financials (about 1.25) and Telecommunications Services (between 1.25 and 1.5)

Link to Appendix slide on data background check

Data Preprocessing

	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
0	-0.393341	0.493950	0.272749	0.989601	-0.210698	-0.339355	1.554415	1.309399	0.107863	-0.652487	-0.506653
1	-0.220837	0.355439	1.137045	0.937737	0.077269	-0.002335	0.927628	0.056755	1.250274	-0.311769	-0.504205
2	-0.367195	0.602479	-0.427007	-0.192905	-0.033488	0.454058	0.744371	0.024831	1.098021	-0.391502	0.094941
3	0.133567	0.825696	-0.284802	-0.317379	1.218059	-0.152497	-0.219816	-0.230563	-0.091622	0.947148	0.424333
4	-0.260874	-0.492636	0.296470	-0.265515	2.237018	0.133564	-0.202703	-0.374982	1.978399	3.293307	0.199196




```
Ticker Symbol                0
Security                     0
GICS Sector                  0
GICS Sub Industry            0
Current Price                0
Price Change                 0
Volatility                   0
ROE                          0
Cash Ratio                   0
Net Cash Flow                0
Net Income                   0
Earnings Per Share           0
Estimated Shares Outstanding 0
P/E Ratio                    0
P/B Ratio                    0
dtype: int64
```

- The data contains no missing values
- No duplicates
- All the numerical columns contain outliers
- The scaled dataset contains comparable values across columns, ready for effective clustering while avoiding large value columns from skewing the clustering models

K-Means Clustering Summary: Cluster Profiling (1/2)

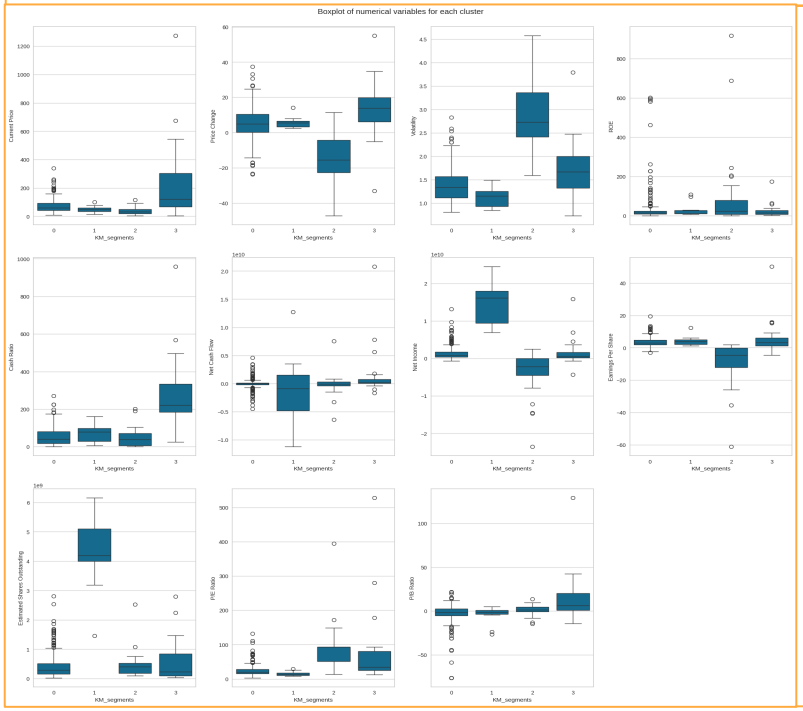
KM_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	count_in_each_segment
0	72.399112	5.066225	1.388319	34.620939	53.000000	-14046223.826715	148221289.891697	3.621029	438533835.667184	23.843656	-3.358948	257
1	50.517273	5.747586	1.130399	31.090909	75.909091	-107227272.727272	[illegible]	4.154545	[illegible]	14.803577	-4.552119	11
2	38.099260	-15.370329	2.591550	107.074074	50.037037	-159428481.481481	-3887457740.740741	-9.473704	480398572.845926	90.815556	1.342067	27
3	234.177032	19.400685	1.729989	25.600000	277.640000	1554235560.000000	1572611680.000000	8.042320	57831631 8.948800	74.960824	14.403462	25

```
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
Name: Security, dtype: int64
```

- KM_Segment 0 has the greatest number of companies whereas Cluster 1 has the least
- KM_Segment 1 has the greatest average Net Income and Estimated Shares Outstanding
- KM_Segment 2 has the greatest Volatility, ROE, and P/E Ratio
- KM_Segment 3 has the greatest Current Price, Price Change, Cash Ratio, Net Cash Flow, Earnings Per Share, and P/B Ratio
- It appears KM_Segment 0 comprises of the majority of companies with relatively low financial metrics while the other 3 KM Segments constitute, each, of relatively few companies with high financial performance, demonstrated via diverse metrics
- KM_Segment 0 has the largest number of GICS Sectors (11) followed by KM_Segment 3 (8) and KM_Segment 1 (7); KM_Segment 2 contains the least number of GICS Sectors (4)
- Telecommunications Services are spread across KM Segments 0, 1, and 3
- Consumer Discretionary is most represented in KM_Segment 0 and least in KM_Segment 2 where is has no representation at all, same with Financials, Consumer Staples, and a couple other sectors
- Utilities are only found in KM_Segment 0
- KM Segments 1, 2, and 3 contain some well known big players such as Coca Cola, Pfizer, Ford Motor, Yahoo, Facebok, and some Oil And Gaz Companies

Link to Appendix slide on K-Means Clustering

K-Means Clustering Summary: Cluster Profiling (2/2)



- KM Segments 1, 2, and 3 contain some well known big players such as Coca Cola, Pfizer, Ford Motor, Yahoo, Facebok, and some Oil And Gaz Companies
- The boxplots confirm the information obtained from the profile highlight
- In addition, signification variation is observed in Current Price, Cash Ratio, P/E Ratio, and P/B Ratio for KM_Segment 3, Price Change, Volatility, ROE, and Earnings Per Share for KM_Segment 2, and Net Cash Flow, Net Income, and Estimated Shares Outstanding for KM_Segment 1
- KM_Segment 0
  - Has 277 companies
  - Low Current Price, Volatility, and Cash Ratio
  - Moderate to high Price Change
  - Low ROE
  - Moderate Net Cash Flow, Net Income, and Earnings Per Share
  - Low Estimated Shares Outstanding, P/E Ratio, and P/B Ratio
- KM_Segment 1
  - 11 companies
  - Low to moderate Current Price
  - Moderate Price Change
  - Low to moderate Current Price, ROE, and Cash Ratio
  - Low Volatility, P/E Ratio, and P/B Ratio
  - Low but varied Net Cash Flow
  - High and varied Net Income
  - Relatively High Earnings Per Share
  - High and varied Estimated Shares Outstanding
- KM_Segment 2
  - 27 companies
  - Low Current Price, Cash Ratio, Net Income, and Estimated Shares Outstanding
  - Low and varied Price Change and Earnings Per Share
  - High and varied Volatility
  - Relatively high and varied ROE
  - Low to moderate Current Price
  - Medium Net Cash Flow and P/B Ratio
  - High P/E Ratio
- KM_Segment 3
  - 25 companies
  - High and varied Current Price, Price Change, Cash Ratio, and P/B Ratio
  - Medium Volatility, Net Income, and Earnings Per Share
  - Low ROE
  - High Net Cash Flow
  - Low and varied Estimated Shares Outstanding
  - Medium and varied P/E Ratio

Link to Appendix slide on K-Means Clustering

Hierarchical Clustering Summary: Cluster Profiling (1/2)

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	42.848182	6.270446	1.123547	22.727273	71.454545	558636363.636364	[redacted]	3.410000	[redacted]	15.242169	-4.924615	11
1	84.355716	3.854981	1.827670	[redacted]	33.571429	-568400000.000000	-496815742.857142	-10.841429	398169036.442857	42.284541	-11.589502	7
2	325.996105	7.724708	1.545762	17.000000	[redacted]	186503166.666667	100873646.666667	[redacted]	589434517.801667	41.121871	[redacted]	12
3	72.760400	5.213307	1.427078	25.603509	60.392982	79951512.280702	1538594322.807018	3.655351	446472132.228456	24.722670	-2.647194	[redacted]
4	36.440455	-16.073408	[redacted]	57.500000	42.409091	-472834090.909091	-3161045227.272727	-8.005000	514367806.201818	85.555682	0.836839	22
5	[redacted]	[redacted]	2.029752	4.000000	106.000000	[redacted]	287547000.000000	0.750000	366783235.300000	[redacted]	-5.322376	3

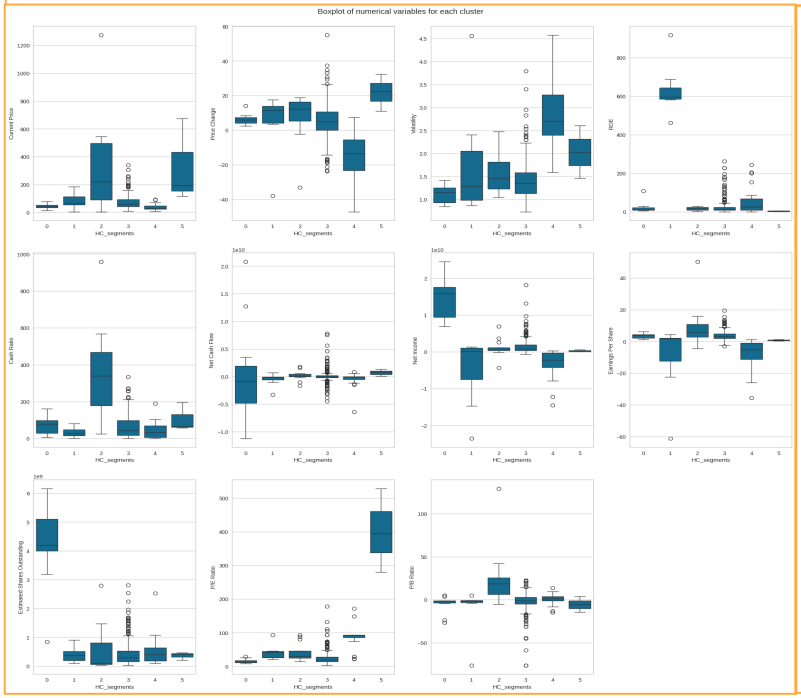
HC_segments	GICS Sector	
0	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	4
	Health Care	1
	Information Technology	1
	Telecommunications Services	2
1	Consumer Discretionary	1
	Consumer Staples	2
	Energy	2
	Financials	1
	Industrials	1
2	Consumer Discretionary	2
	Consumer Staples	1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

Name: Security, dtype: int64

- HC_segment 0 registered the highest Net Income and Estimated Shares Outstanding
- HC_segment 1 registered the highest ROE
- HC_segment 2 registered the highest Cash Ratio, Earnings Per Share, and P/B Ratio
- HC_segment 3 contains the greatest number of companies (285) but is not outstanding in any of the financial metrics - the large majority of average companies
- HC_segment 4 registered the highest Volatility
- HC_segment 5 registered the highest Current Price, Price Change, Net Cash Flow, and P/E Ratio but contains the lowest number of companies (3) - clearly a segment of elite companies
- HC_Segment 3 contains the greatest number of GICS Sectors
- Telecommunications Services, Information Technology, Consumer Discretionary, and a few other GICS Sectors are spread oveer several HC Segments
- Real Estate is found only in HC Segments 2 and 3
- Materials are only found in HC Segments 3 and 4
- Utilities are only found in HC Segments 3

Link to Appendix slide on Hierarchical Clustering

Hierarchical Clustering Summary: Cluster Profiling (2/2)



- HC_Segment 0
  - Has 11 companies
  - Low Current Price, Volatility, ROE, and P/E Ratio
  - Medium Price Change, Earnings Per Share, and P/B Ratio
  - Low to medium Cash Ratio
  - Low and varied Net Cash Flow
  - High and varied Net Income and Estimated Shares Outstanding
- HC_Segment 1
  - Has 7 companies
  - Medium Current Price, Price Change, Net Cash Flow, Estimated Shares Outstanding, P/E Ratio, and P/B Ratio
  - Low Volatility and Cash Ratio
  - High ROE
  - Medium and varied Net Income and Earnings Per Share
- HC_Segment 2
  - Has 12 companies
  - High and varied Current Price, Cash Ratio, and P/B Ratio
  - Medium Price Change, Volatility, Net Cash Flow, Net Income, and P/E Ratio
  - Low ROE
  - High Earnings Per Share
  - Low and varied Estimated Shares Outstanding
- HC_Segment 3
  - Has 285 companies
  - Low to medium Current Price and Estimated Shares Outstanding
  - Medium Price Change, Volatility, Cash Ratio, Net Cash Flow, Net Income, Earnings Per Share, and P/B Ratio
  - Low ROE and P/E Ratio
- HC_Segment 4
  - Has 22 companies
  - Low Current Price, Cash Ratio, Net Income, and Earnings Per Share
  - Low and varied Price Change
  - High and varied Volatility
  - Medium and varied ROE
  - Medium Net Cash Flow, Estimated Shares Outstanding, P/E Ratio, and P/B Ratio
- HC_Segment 5
  - Has 3 companies
  - Medium to high Current Price and Volatility
  - High Price Change
  - Medium Net Income, Cash Ratio, and Earnings Per Share
  - Low ROE and P/B Ratio
  - Relatively high Net Cash Flow
  - Low to medium Estimated Shares Outstanding
  - High and varied P/E Ratio

Link to Appendix slide on Hierarchical Clustering

APPENDIX

Data Background and Contents: Data Types and Sample display

	Ticker Symbol	Security	GICS Sector	GICS Sub Industry	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
102	DVN	Devon Energy Corp.	Energy	Oil & Gas Exploration & Production	32.000000	-15.478079	2.923698	205	70	830000000	-14454000000	-35.55	4.065823e+08	93.089287	1.785616
125	FB	Facebook	Information Technology	Internet Software & Services	104.660004	16.224320	1.320606	8	958	592000000	3669000000	1.31	2.800763e+09	79.893133	5.884467
11	AIV	Apartment Investment & Mgmt	Real Estate	REITs	40.029999	7.578608	1.163334	15	47	21818000	248710000	1.52	1.636250e+08	26.335526	-1.269332
248	PG	Procter & Gamble	Consumer Staples	Personal Products	79.410004	10.660538	0.806056	17	129	160383000	636056000	3.28	4.913916e+08	24.070121	-2.256747
238	OXY	Occidental Petroleum	Energy	Oil & Gas Exploration & Production	67.610001	0.865287	1.589520	32	64	-588000000	-7829000000	-10.23	7.652981e+08	93.089287	3.345102
336	YUM	Yum! Brands Inc	Consumer Discretionary	Restaurants	52.516175	-8.698917	1.478877	142	27	159000000	1293000000	2.97	4.353535e+08	17.682214	-3.838260
112	EQT	EQT Corporation	Energy	Oil & Gas Exploration & Production	52.130001	-21.253771	2.364883	2	201	523803000	85171000	0.56	1.520911e+08	93.089287	9.567952
147	HAL	Halliburton Co.	Energy	Oil & Gas Equipment & Services	34.040001	-5.101751	1.966062	4	189	7786000000	-671000000	-0.79	8.493671e+08	93.089287	17.345857
89	DFS	Discover Financial Services	Financials	Consumer Finance	53.619999	3.653584	1.159897	20	99	2288000000	2297000000	5.14	4.468872e+08	10.431906	-0.375934
173	IVZ	Invesco Ltd.	Financials	Asset Management & Custody Banks	33.480000	7.067477	1.580839	12	67	412000000	968100000	2.26	4.283628e+08	14.814159	4.218620

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 340 entries, 0 to 339
Data columns (total 15 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Ticker Symbol               340 non-null    object
 1   Security                    340 non-null    object
 2   GICS Sector                 340 non-null    object
 3   GICS Sub Industry           340 non-null    object
 4   Current Price               340 non-null    float64
 5   Price Change                340 non-null    float64
 6   Volatility                  340 non-null    float64
 7   ROE                         340 non-null    int64
 8   Cash Ratio                  340 non-null    int64
 9   Net Cash Flow               340 non-null    int64
 10  Net Income                  340 non-null    int64
 11  Earnings Per Share          340 non-null    float64
 12  Estimated Shares Outstanding  340 non-null    float64
 13  P/E Ratio                   340 non-null    float64
 14  P/B Ratio                   340 non-null    float64
dtypes: float64(7), int64(4), object(4)
memory usage: 40.0+ KB
```

- There are 340 rows and 15 columns
- The data appears to have 4 categorical columns and 11 numerical columns
- Negative price changes indicate decrease in stock price over the previous 13 weeks
- Negative net cash flow, net income, and earnings per share indicate, respectively, net cash outflow, total expenses, interest, and taxes greater than revenues, net loss by the respective companies
- Negative P/B ratios indicate that the companies have more liabilities than assets
- We have confirmation that the data contains 4 categorical columns and 11 numerical (7 float and 4 int) columns
- No obvious missing values

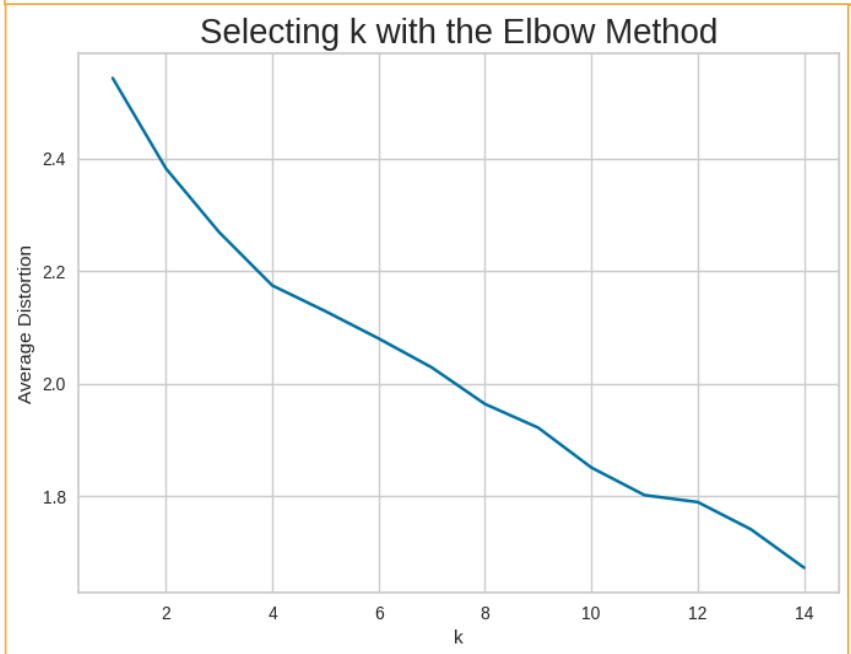
Data Background and Contents: Statistical Summary

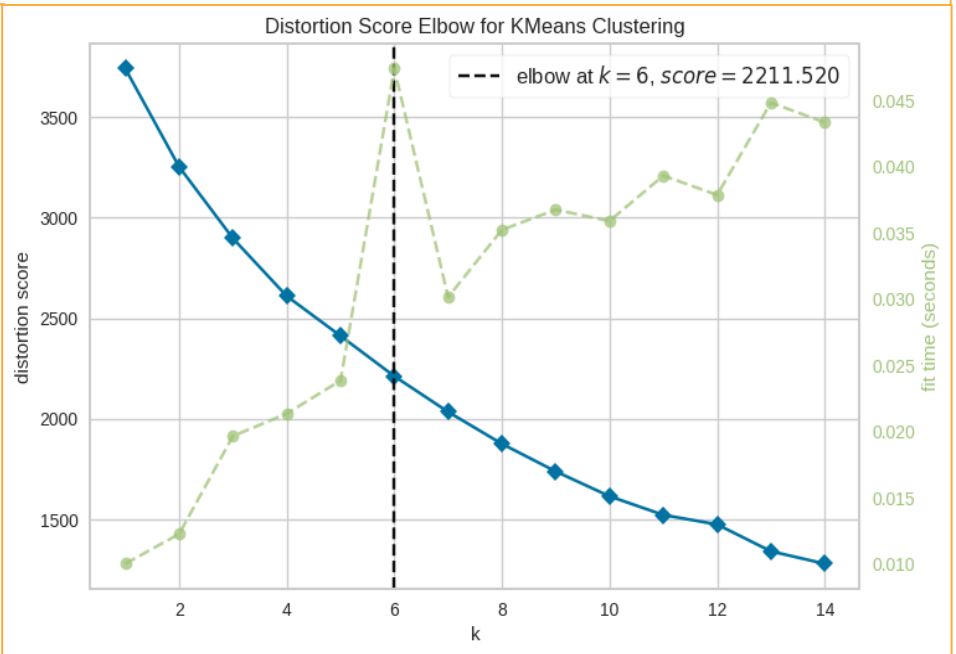
	count	mean	std	min	25%	50%	75%	max
Current Price	340.0	8.086234e+01	9.805509e+01	4.500000e+00	3.855500e+01	5.970500e+01	9.288000e+01	1.274950e+03
Price Change	340.0	4.078194e+00	1.200634e+01	-4.712969e+01	-9.394838e-01	4.819505e+00	1.069549e+01	5.505168e+01
Volatility	340.0	1.525976e+00	5.917984e-01	7.331632e-01	1.134878e+00	1.385593e+00	1.695549e+00	4.580042e+00
ROE	340.0	3.959706e+01	9.654754e+01	1.000000e+00	9.750000e+00	1.500000e+01	2.700000e+01	9.170000e+02
Cash Ratio	340.0	7.002353e+01	9.042133e+01	0.000000e+00	1.800000e+01	4.700000e+01	9.900000e+01	9.580000e+02
Net Cash Flow	340.0	5.553762e+07	1.946365e+09	-1.120800e+10	-1.939065e+08	2.098000e+06	1.698108e+08	2.076400e+10
Net Income	340.0	1.494385e+09	3.940150e+09	-2.352800e+10	3.523012e+08	7.073360e+08	1.899000e+09	2.444200e+10
Earnings Per Share	340.0	2.776662e+00	6.587779e+00	-6.120000e+01	1.557500e+00	2.895000e+00	4.620000e+00	5.009000e+01
Estimated Shares Outstanding	340.0	5.770283e+08	8.458496e+08	2.767216e+07	1.588482e+08	3.096751e+08	5.731175e+08	6.159292e+09
P/E Ratio	340.0	3.261256e+01	4.434873e+01	2.935451e+00	1.504465e+01	2.081988e+01	3.176476e+01	5.280391e+02
P/B Ratio	340.0	-1.718249e+00	1.396691e+01	-7.611908e+01	-4.352056e+00	-1.067170e+00	3.917066e+00	1.290646e+02

	count	unique	top	freq
Ticker Symbol	340	340	AAL	1
Security	340	340	American Airlines Group	1
GICS Sector	340	11	Industrials	53
GICS Sub Industry	340	104	Oil & Gas Exploration & Production	16

- Only Price Change, Net Cash Flow, and Net Income, Earnings Per Share, and P/B Ratio contain negative values; these are acceptable scenarios
- The means of Current Price, Volatility, ROE, Cash Ratio, Net Cash Flow, Net Income, Estimated Shares Outstanding, and P/E Ratio are greater than their respective medians, indicating that these variables might be right-skewed
- The means of Price Change, Earnings Per Share, and P/B Ratio are less than their respective medians, indicating that these variables might be left-skewed
- Comparing the minimum values, first quartile, median, third quartile, and maximum values, we expect the data to contain some outliers
- Ticker Symbol and Security each have 340 unique values while GICS Sub Industry and GICS Sector respectively have 104 and 11 distinct values
- The mode of GICS Sub Industry is "Oil & Gas Exploration & Production" whereas that of GICS Sector is "Industrials"

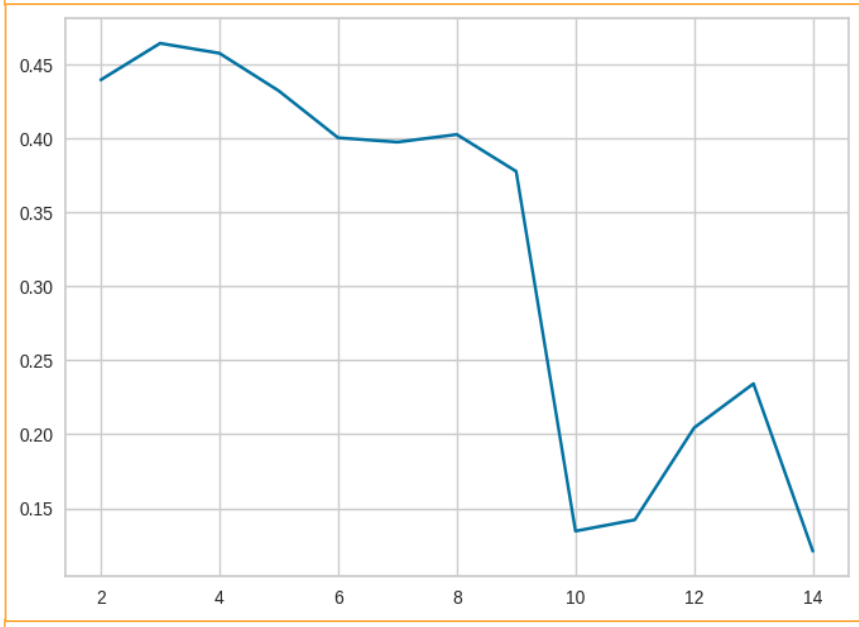
K-Means Clustering Technique: Checking Elbow Plot

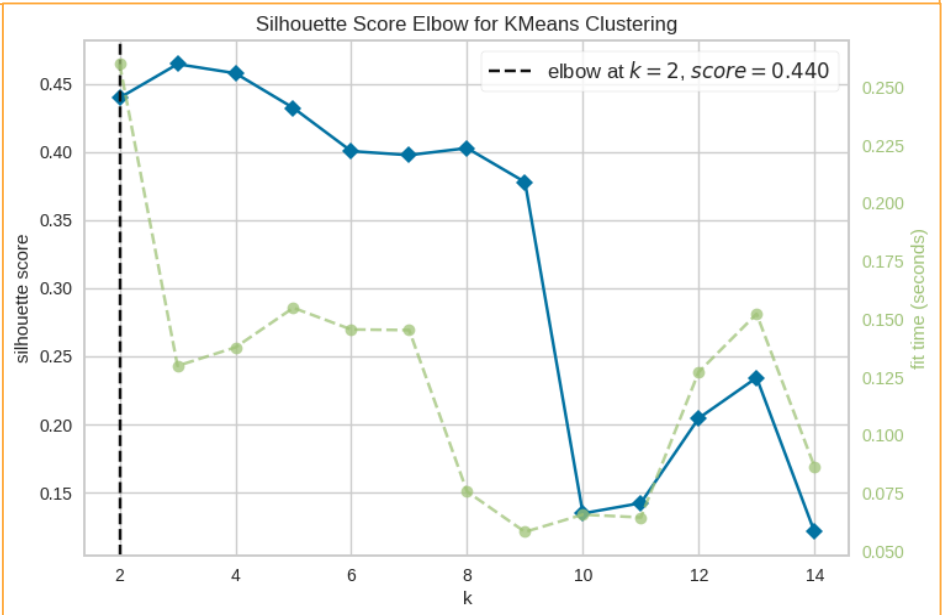




- From the elbow curve, 2, 4, and 11 appear to be good candidates for k, the number of clusters
- The KElbowVisualizer of the yellowbrick.cluster library suggests an elbow at k=6 with a fit time of about 50 ms

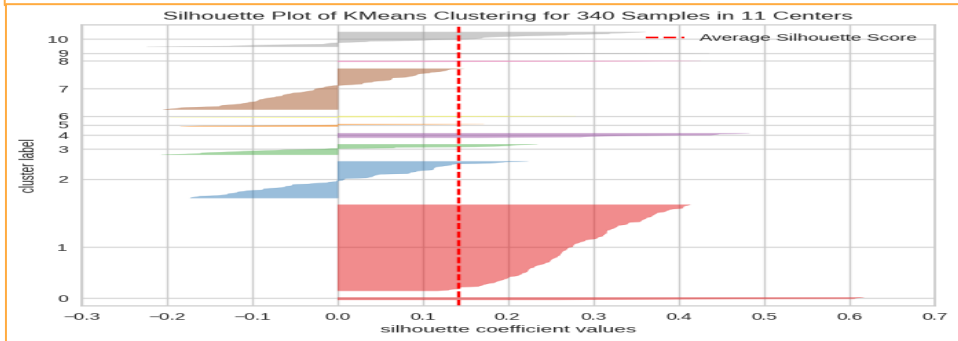
K-Means Clustering Technique: Silhouette Scores (1/2)

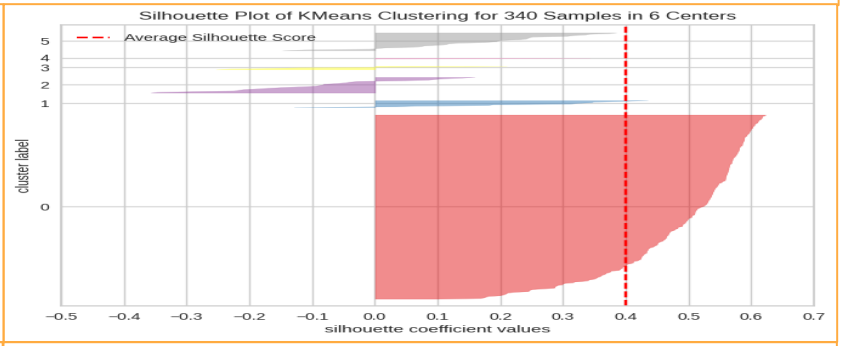


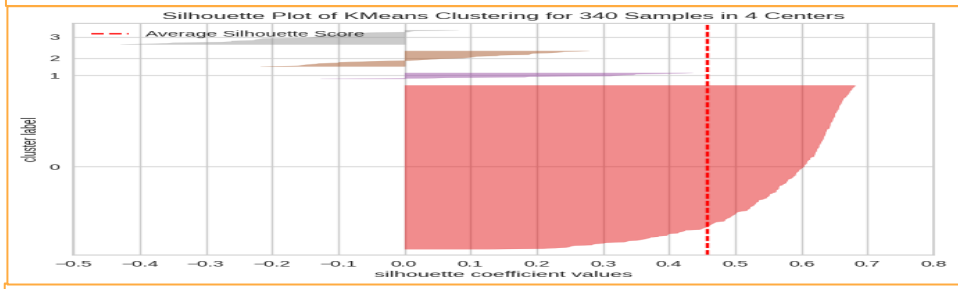


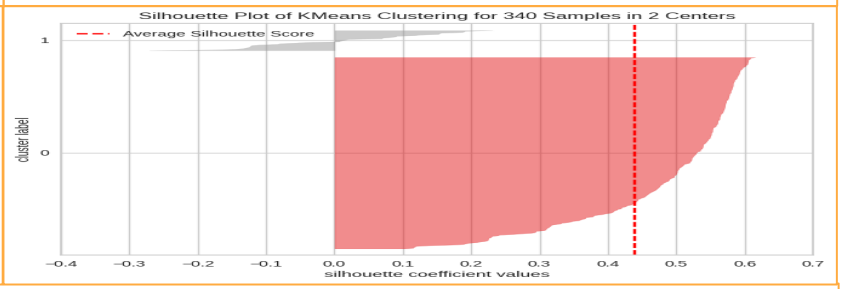
- The greatest silhouette score is registered for 3 clusters followed by 4 then 2 clusters
- Using the silhouette metric, the KElbowVisualizer of the yellowbrick.cluster library suggests an elbow at k=2 with a score of 0.44 and a fit time of about 260 ms

K-Means Clustering Technique: Silhouette Scores (2/2)









- We will be choosing 4 as the appropriate number of clusters since it has a relatively high silhouette score (about 0.46) and there is a knick in the elbow curve at k=4

Hierarchical Clustering Technique: Cophenetic Correlation (1/2)

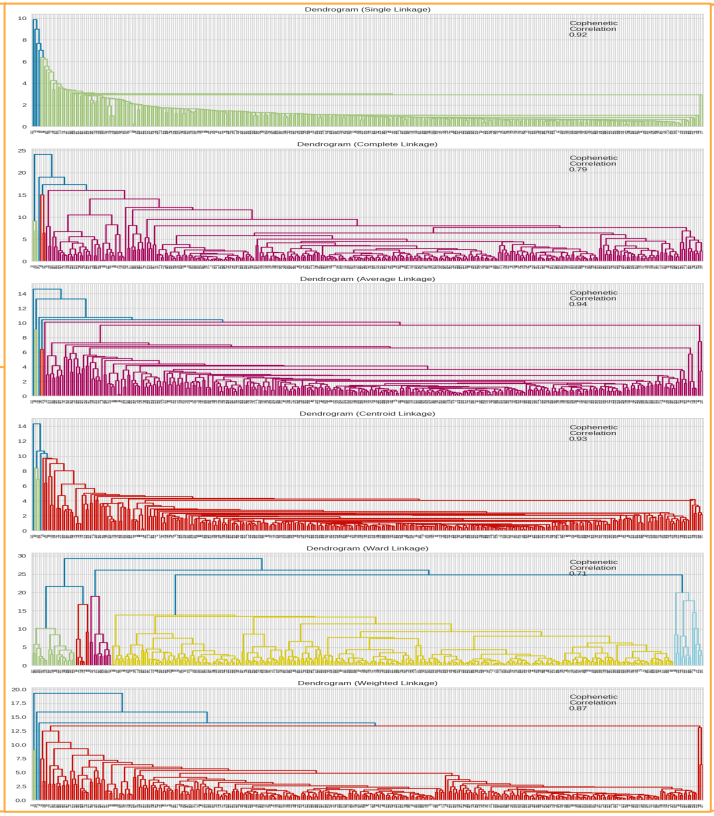
```
Cophenetic correlation for Euclidean distance and single linkage is 0.9232271494002922.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7873280186580672.
Cophenetic correlation for Euclidean distance and average linkage is 0.9422540609560814.
Cophenetic correlation for Euclidean 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 0.7925307202850002.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159736.
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 0.94, which is obtained with Euclidean distance and average linkage

Hierarchical Clustering Technique: Cophenetic Correlation (2/2): Focus on Euclidean Distance

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 0.7101188299865353.
Cophenetic correlation for weighted linkage is 0.8693784298129404.

	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 with Euclidean distance is 0.94, which is obtained with average linkage
- The cophenetic correlation with Euclidean distance is highest for average linkage (0.94).
- With Euclidean Distance:
  - 6 appears to be the appropriate number of clusters from the dendrogram for average linkage (At a distance of 10)
  - Average linkage has the highest cophenetic coefficient (0.94) whereas ward linkage has the least cophenetic coefficient (0.71)
  - We will however move ahead with ward linkage since it has the most distinct and separated clusters; from the dendrogram, 6 appears to be the appropriate number of clusters (just below a distance of 20)

K-Means vs Hierarchical Clustering

- It took about 31s to display the dendrograms of the different linkage methods with the Euclidean distance whereas fitting of K-Means for different values of k was accomplished in milliseconds
- As concerns the clustering performance of each technique:
  - Of the 13 values of K evaluated, the chosen number of clusters, 4, registered the second highest silhouette score (about 0.46)
  - Of the various linkage vs distance combinations evaluated for the hierarchical clustering, the chosen combination, ward linkage – Euclidean distance, registered a relatively low cophenetic score (0.71) but was chosen because, from the dendrogram, its clusters appeared to be the most distinct and separated; number of clusters chosen: 6
- Certain observations on corresponding similar clusters across both techniques:
  - The smallest HC Segment (5) contains 11 companies whereas the smallest KM Segment (1) has 3 companies. They appear to perform best on different metrics: HC_Segment 5 performs best on Price Change, Cash Flow, and P/E Ratio whereas KM_Segment 1 on Net Income, Earnings Per Share, and Estimated Shares Outstanding
  - The largest HC Segment (3) contains 285 companies whereas the largest KM Segment (0) contains 277 companies. Both have low to medium performance on the various metrics
  - Intermediate size clusters obtained using each technique display varying levels of performance on the provided metrics

Great Learning

Happy Learning !