Trade&Ahead

Investment Portfolio Analysis _ Unsupervised Learning

Julie Kistler _ 2/29/24

Contents / Agenda

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

Executive Summary

Navigating the stock market requires a thoughtful strategy to maximize returns and minimize risks. A diversified portfolio is key to this approach, balancing potential gains and safeguarding against market downturns. The challenge lies in sifting through the myriad of financial metrics to identify stocks that align with an investor's profile and investment goals.

Trade&Ahead has hired us to analyze stock data from the New York Stock Exchange and develop personalized investment strategies for their clients. The task involved employing data analysis to cluster stocks based on various financial indicators and providing insights into the defining characteristics of each group.

Insights and Recommendations:

- **K-Means Clustering**: It is recommended that Trade&Ahead utilize the K-Means clustering algorithm to classify stocks into distinct groups, providing a structured approach to stock categorization.
- **Investor Profile Analysis**: A crucial step is to thoroughly understand each client's financial aspirations, risk tolerance, and investment habits. This understanding will inform the alignment of clients with stock clusters that best match their individual investment criteria, offering a customized potential investment portfolio.
- **Strategic Stock Selection**: The clusters formed can act as a foundation for in-depth financial statement analysis, with a focus on identifying outlier stocks that do not conform to the general trends of their cluster. This allows for the identification of unique investment opportunities and risks.
- **Tailored Investment Plans**: For clients whose strategies involve selecting individual stocks, the insights from the clustering process can be used to pinpoint stocks likely to outperform their peers or those at risk of underperformance, ensuring a tailored and strategic investment approach.

Overall, the integration of K-Means clustering in Trade&Ahead's analytical toolkit is expected to enhance the firm's ability to craft bespoke investment portfolios that are both responsive to client needs and resilient to market fluctuations.

Business Problem Overview and Solution Approach

The core problem is to assist Trade&Ahead to making informed decisions by providing personalized investment strategies. The challenge lies in navigating through complex financial data to identify stocks that not only promise good returns but also align with Trade&Ahead risk tolerance and investment goals. The goal is to create a diversified portfolio that can withstand market volatility, optimize returns, and mitigate risks.

The solution approach is to leveraging data science and machine learning techniques, particularly clustering, to analyze and group stocks, We will use two types of clustering algorithms - K-Means and Hierarchical clustering methods to derive insights from the clusters.

This approach focuses on leveraging data-driven insights to minimize risk and maximize returns for investors. By systematically analyzing and grouping stocks.

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 Cont…

Column	Dtype
Ticker Symbol	object
Company	object
GICS Sector	object
GICS Sub Industry	object
Current Price	float64
Price Change	float64
Volatility	float64
ROE	int64
Cash Ratio	int64
Net Cash Flow	int64
Net Income	int64
Earnings Per Share	float64
Estimated Shares Outstanding	float64
P/E Ratio	float64
P/B Ratio	float64

Rows	Columns
340	15

- There are no duplicate values
- There are no missing values

- **4 object data types** (Ticker Symbol, Company, GICS Sector, GICS Sub Industry)
- **4 integer data types** (ROE, Cash Ratio. Net Cash Flow, Net Income)
- **7 float data type** (Current Price, Price Change, Volatility, Earnings Per Share, Estimated Shared Outstanding, P/E Ratio, P/B Ratio)

EDA Results

Statistical Summary

- There are 340 unique stock symbols
- There are 11 unique GICS sectors with industrials as the most frequent (53)
- There are 104 GICS sub industry with Oil &Gas Exploration & Production as the most frequent (16)

~~~~~~~~~~~~~~~~~~~~~

Current Price

- Highly skewed right
- No stocks are listed for less than zero
- Average price appears to be approx. $81.00

Price Change

- Distribution appears to favor lower prices
- There are outliers on both ends
- Over the past 13 weeks the most volatile stocks showed an approximate price increase of 55% and a decrease of 47%

Link to Appendix slide on statistical summary of data
~~~~~~~~~~~~~~~~~~~~~

EDA Results _ Univariate Analysis Cont…

Volatility

- Standard deviation is not normal and skewed right

ROE

- Highly skewed right
- No stock is listed below $0
- There are some outliers

Cash Ratio

- Highly skewed right
- No stocks have a cash ratio of less than $0
- The average cash ratio is approximately $70.00

Net Cash Flow

- Distribution appears normal
- There are outliers on both ends

Net Income

- Skewed right
- There are outliers on both ends
- As expected some companies are reflecting a positive net income while others are reflecting a negative net income

Link to Appendix slide to EDA Results

EDA Results _ Univariate Analysis Cont…

Earnings Per share

- Skewed right
- There are outliers on both ends
- It appears most companies are reflecting positive earnings per share

Estimated Shares Outstanding

- Highly skewed right
- There are outliers on the positive tail
- No values stand out as highly out of normal range

P/E Ratio

- Highly skewed right
- There are outliers on the positive tail
- No stocks appear to have a negative P/E ratio

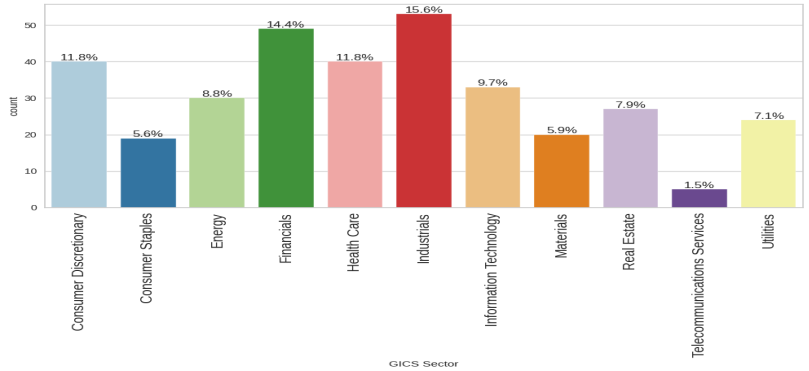
P/B Ratio

- Skewed slightly right
- There are large outliers on the both ends
- The distribution is centered around zero (0)

Link to Appendix slide to EDA Results

EDA Results _ Univariate Analysis Cont…

GIS Sector



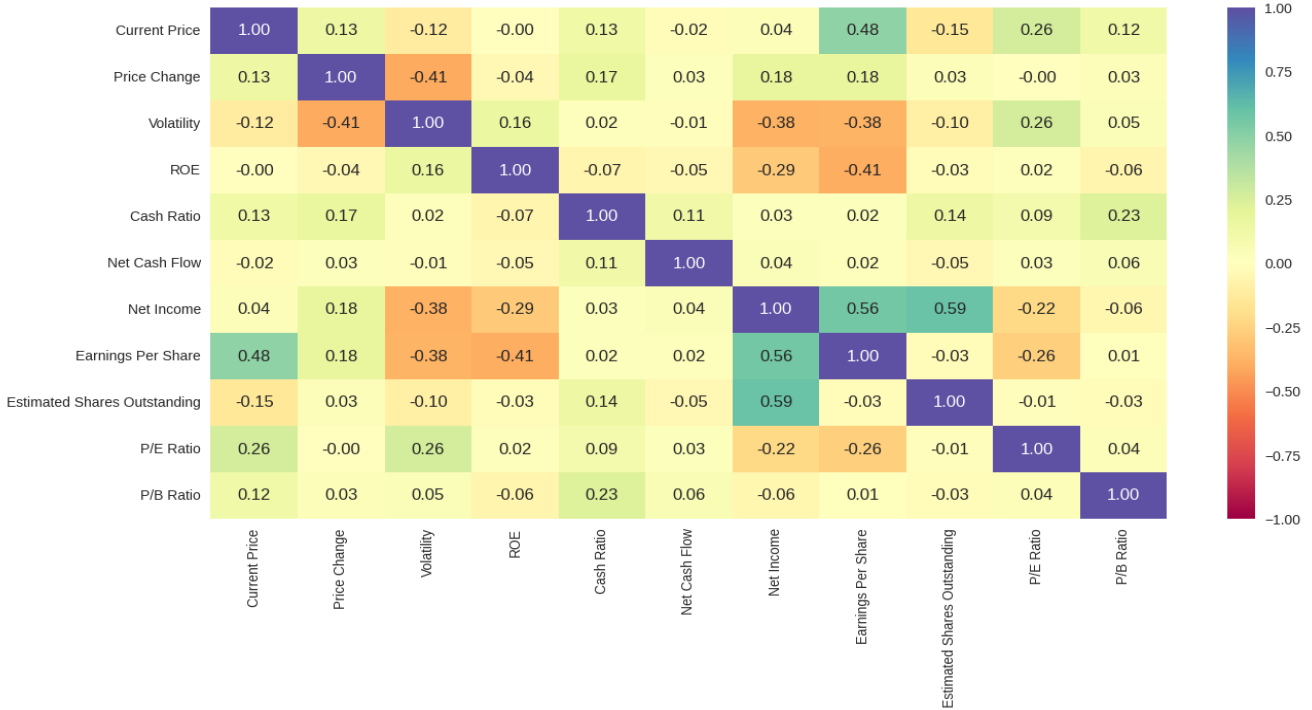
- There are 11 unique GICS sectors
- Industrials is the largest sector reflecting 15.6% of the data and Information Services the lowest reflecting 1.5% of the data

GIS Sub Industry

- There are 104 Unique GICS sub industry categories
- The most dominant sub industry is Oil & Gas Exploration & Production at 4.7%
- Approximately 36% of the data falls within 12 of the GICS sub industries

Link to Appendix slide to EDA Analysis

EDA Results _ Bivariate Analysis



- Net Income has a strong positive correlation to Earnings Per Share and Estimated Share Outstanding
- Earning Per Share and Current Price have a positive correlation
- Volatility is negatively correlated to Net Income, Earnings Per Share, and Price Change
- ROE is negatively correlated with Earnings Per Share

Link to Appendix slide to Bivariate Analysis

EDA Results _ Bivariate Analysis Cont…

Stocks with maximum price increase on average

- Health Care and Customer Staples sectors appear to have the highest price increases
- Energy sector has a significant negative price decrease on average
- Utilities demonstrates a minimal price increase

Average Cash Ratio Across Economic Sectors

- Information Technology has the greatest Cash Ratio across the economic sectors
- Utilities is showing the least average cash ratio across economic sectors

P/E ratio across economic sectors

- Energy sector has the highest P/E ratio on average – significantly higher than the others
- Telecommunications services has the lowest P/E ratio on average

Volatility across economic sectors

- The energy sector has the the highest volatility across the sectors.

Link to Appendix slide to Bivariate Analysis

Data Preprocessing

- There are zero (0) duplicate values
- There are zero (0) missing values
- There was an incorrect header in the CSV File – had to change ”Security” to “Company” to be consistent with the data dictionary
- There are quite a few outliers in the data- these all seem to appear in the dataset -- we will not treat them as they provide valuable information
- The data was scaled before clustering

Link to Appendix slide to Data Preprocessing

K-Means Clustering Summary (Recommended)

It appears the optimal number or clusters is Four (4)

Cluster 0

- 277 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, Telecommunications Services, Utilities)
- Companies in this cluster have:
  - * Moderate to higher price changes
  - * ROE is moderate to high
  - * Current pricing are relatively moderate
  - * This cluster does not have a real significant variable to stands out

Cluster 1

- 11 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Telecommunications Services)
- Companies in this cluster have:
  - * Higher than normal net income
  - * Estimated Shares Outstanding is very high
  - * This cluster does have two significant variable to stands out – Net Income and Estimated Shares Outstanding. Much higher that the other clusters

Cluster 2

- 27 Stocks in the following sectors (Energy, Industrials, Information Technology, Materials)
- Companies in this cluster have:
  - * Higher P/E Ratios
  - * Low Earning per Share
  - * Lower net income
  - * High Volatility
  - * This cluster does have one significant variable to stands out -- Volatility. Much higher that the other clusters

Cluster 3

- 25 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Information Technology, Real Estate, Telecommunications Services)
- Companies in this cluster have:
  - * Higher stock prices
  - * Higher cash ratios
  - * Higher price changes
  - * This cluster does have one significant variable to stands out – Price Change. Much higher that the other clusters

Link to Appendix slide on K-Means Clustering

Hierarchical Clustering Summary

- It appears the optimal number or clusters using Hierarchical Clustering is Five (5)

Cluster 0
- 344 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, Telecommunications Services, Utilities)
- Companies in this cluster have:
  * Moderate Price Changes
  * ROE is moderate
  * Cash Ratios are relatively moderate
  * This cluster does not have a real significant variable to stands out

Cluster 1
- 2 Stocks in the following sectors (Financials& Information Technology)
- Companies in this cluster have:
  * Higher than normal Net Cash Flow
  * Higher P/E Ratios
  * Higher Net Income
  * Higher Cash Ratios
  * This cluster does have two significant variable to stands out – Net Income and Net Cash Flow. Higher than the other clusters

Cluster 2
- 2 Stocks in the following sector (Energy)
- Companies in this cluster have:
  * Moderate to Low Price Changes
  * Very Low Earnings Per Share
  * Lower net income
  * High Volatility
  * This cluster does have two significant variable to stands out – Earning Per Share much lower than others and Net Cash Flow. Higher than the other clusters

Cluster 3
- 1 Stock in the following sector (Information Technology)
- Company in this cluster have:
  * Higher current prices
  * Higher cash ratios
  * Higher price changes
  * Higher Estimated Shares Outstanding
  * This cluster does have one significant variable to stands out – Cash Ratio. Much higher that the other clusters

Cluster 4
- 1 stock in the following sector (Consumer Discretionary)
- Company in this cluster have:
  * Very high Current Price
  * Lower Volatility
  * Higher Cash Ratio
  * Very high Earning Per Share
  * This cluster does have two significant variables that stand out – Current Price and Earnings Per Share higher than the other clusters

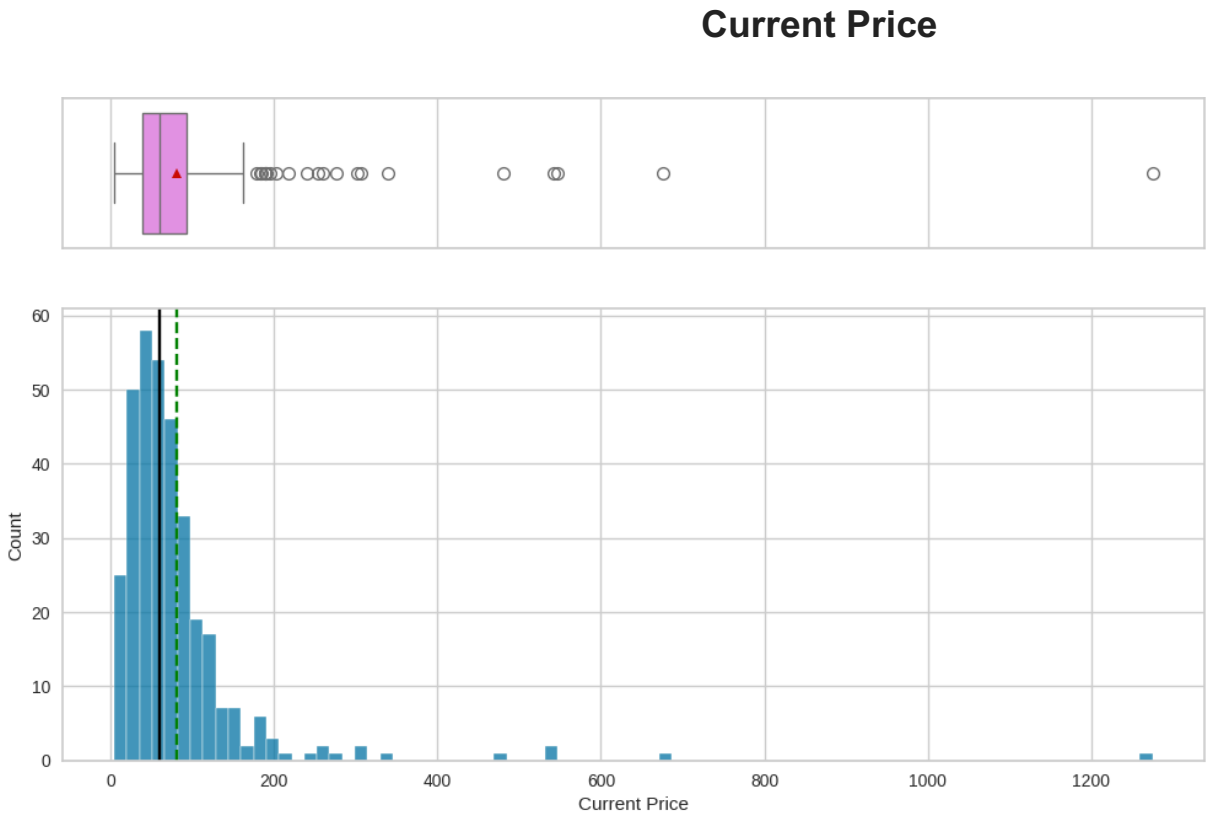
Link to Appendix slide on Hierarchical Clustering

APPENDIX

Statistical Summary of the Data

	Ticker Symbol	Company	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

EDA _ Univariate Analysis

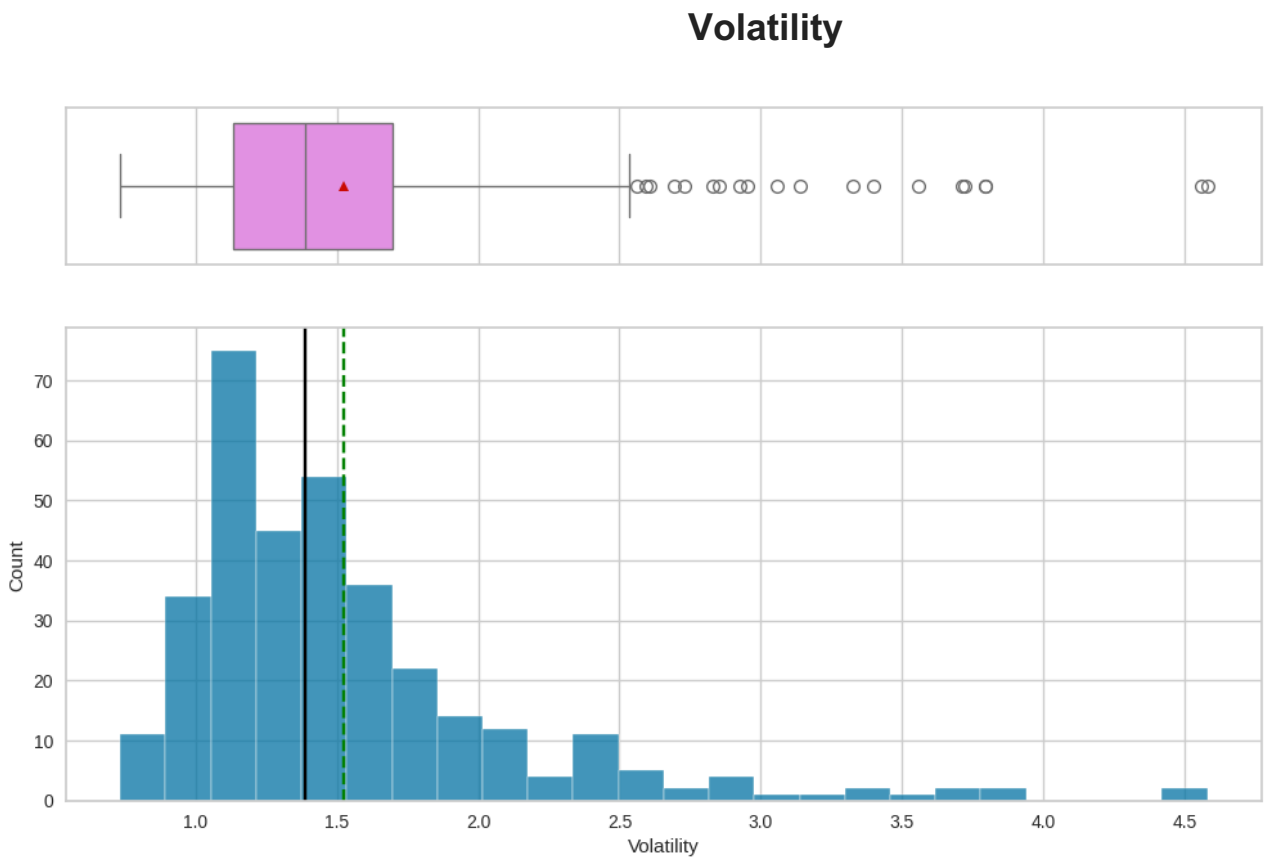


- Highly skewed right
- No stocks are listed for less than zero
- Average price appears to be approx. $80.00

EDA _ Univariate Analysis, Cont...

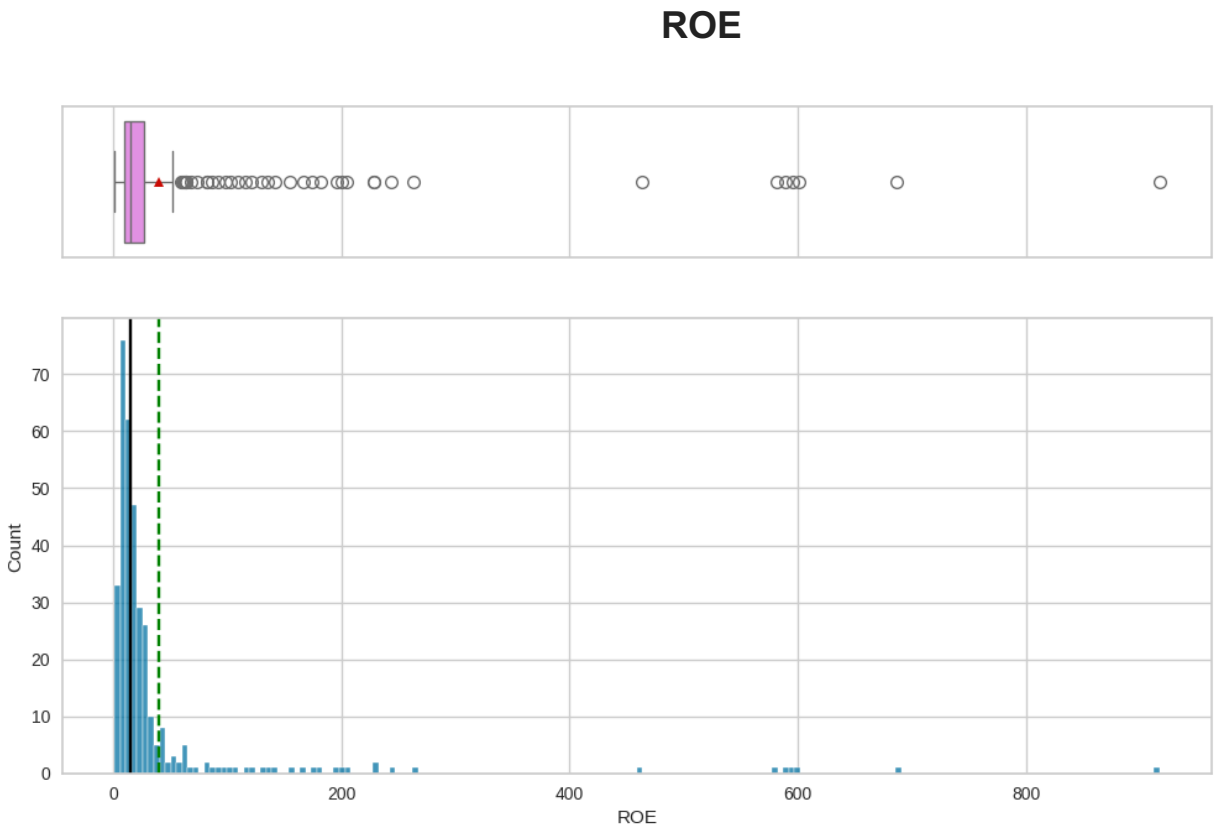
- Distribution appears to favor lower prices
- There are outliers on both ends
- Over the past 13 weeks the most volatile stocks showed an approximate price increase of 55% and a decrease of 47%

EDA _ Univariate Analysis, Cont...



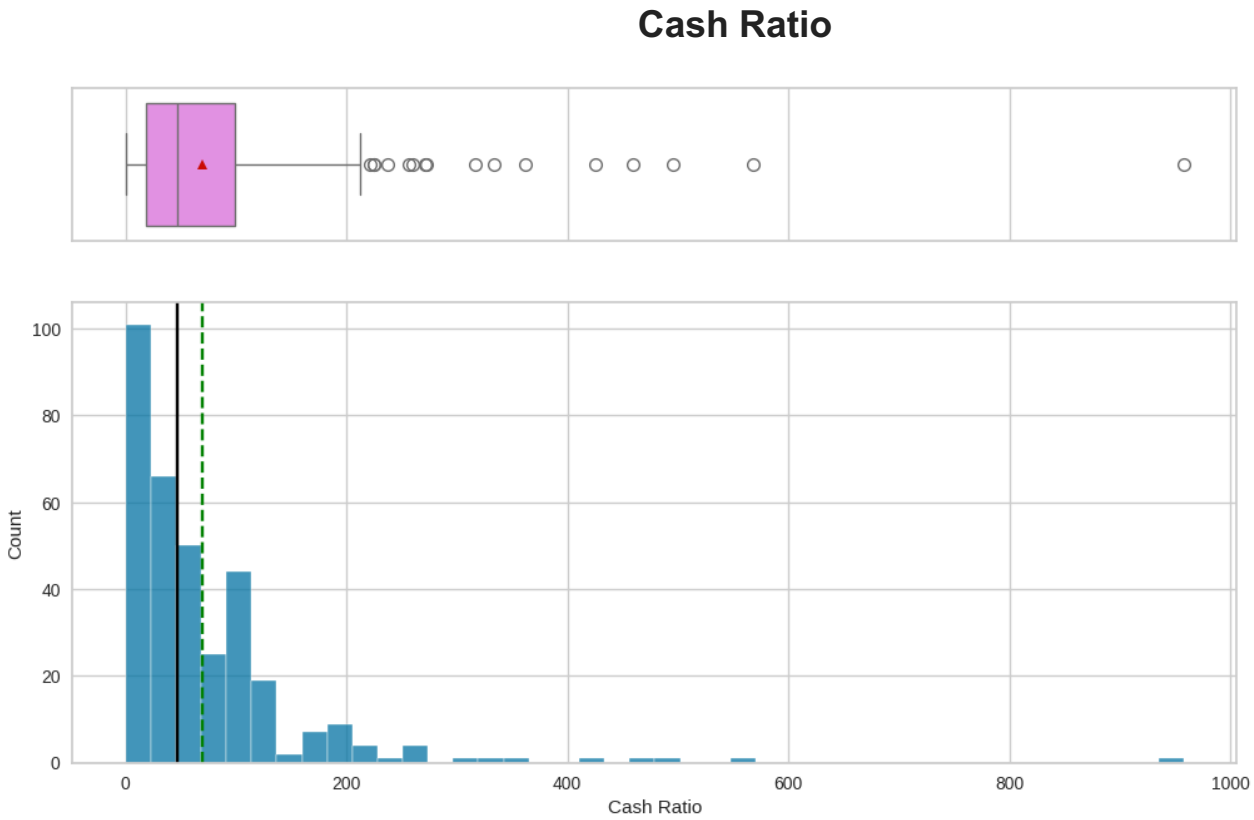
- Standard deviation is not normal and skewed right

EDA _ Univariate Analysis, Cont...



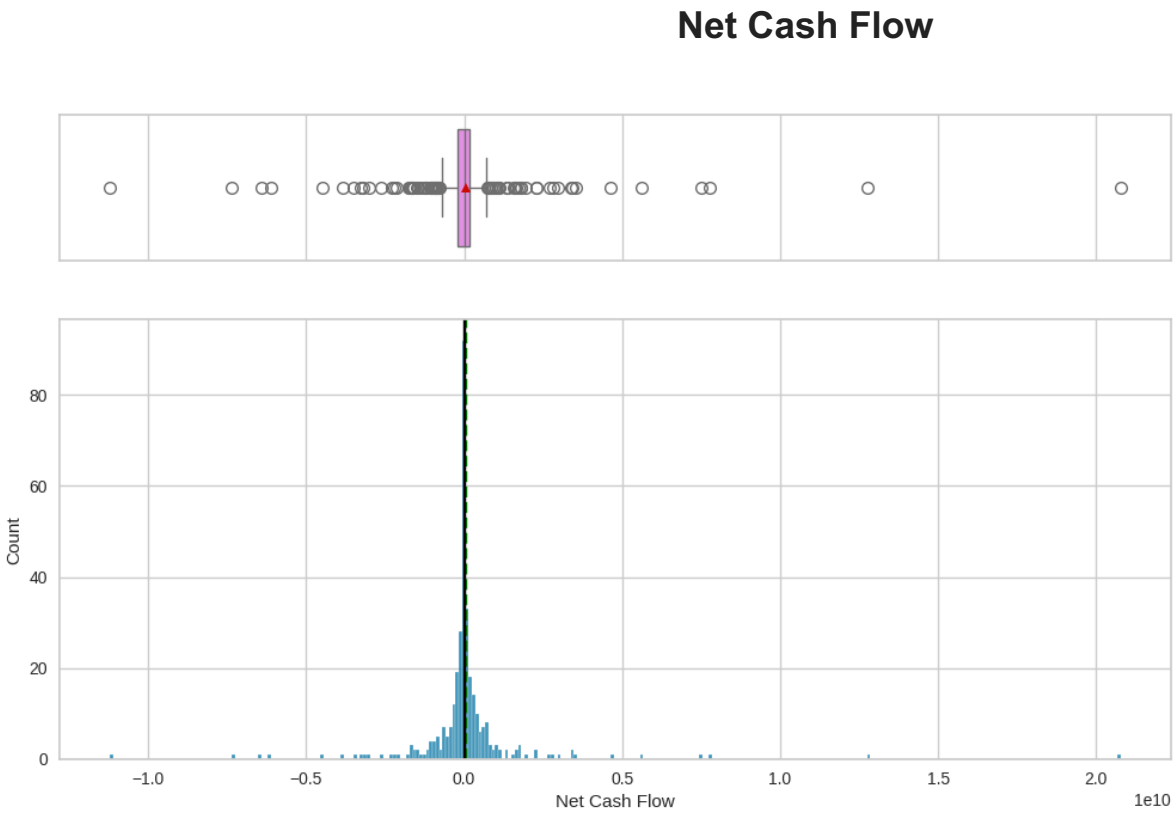
- Highly skewed right
- No stock is listed below $0
- There are some outliers

EDA _ Univariate Analysis, Cont...



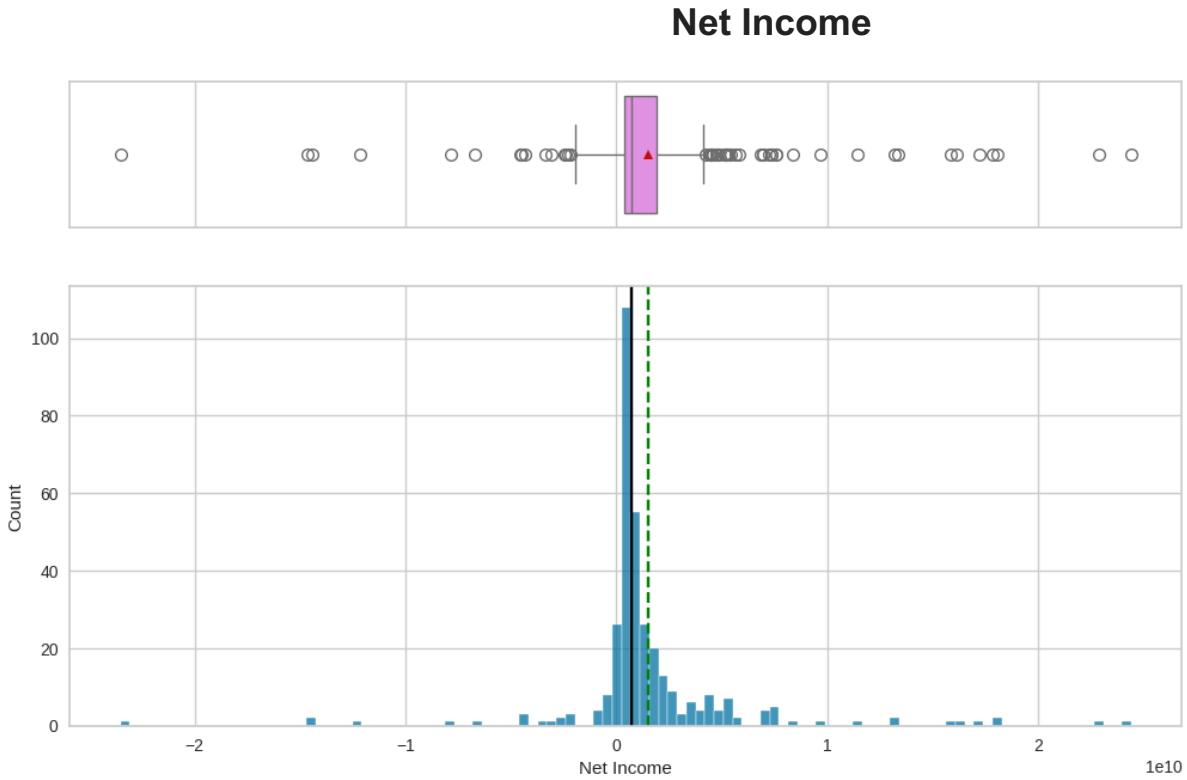
- Highly Skewed Right
- No stocks have a cash ratio of less than $0
- The average cash ratio is approximately $70.00

EDA _ Univariate Analysis, Cont...



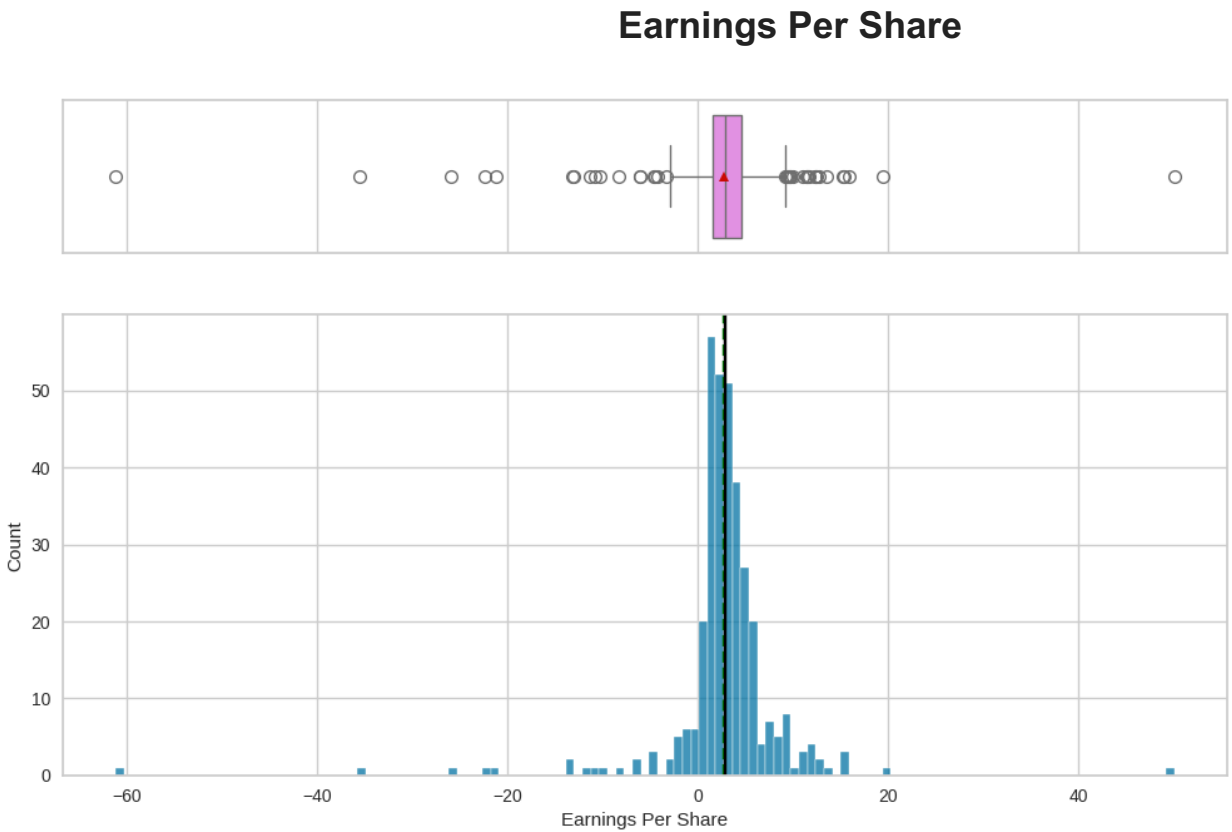
- Distribution appears normal
- There are outliers on both ends

EDA _ Univariate Analysis, Cont...



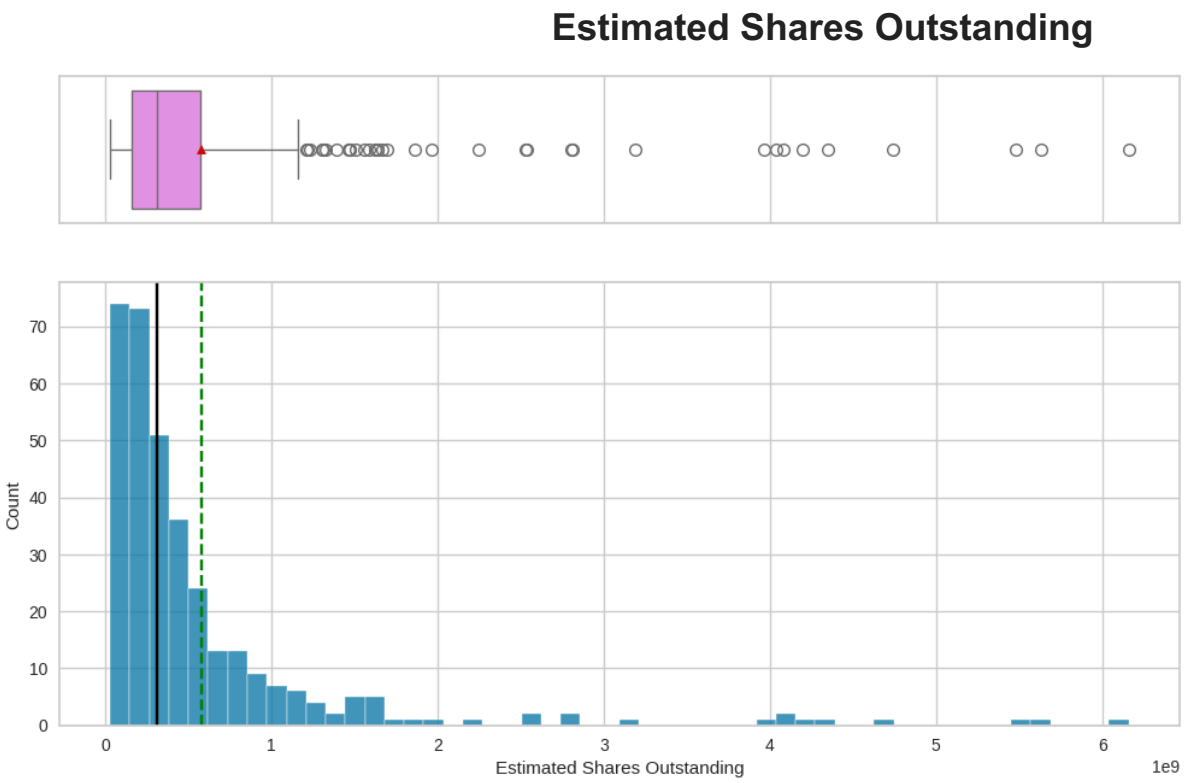
- Skewed right
- There are outliers on both ends
- As expected some companies are reflecting a positive net income while others are reflecting a negative net income

EDA _ Univariate Analysis, Cont...



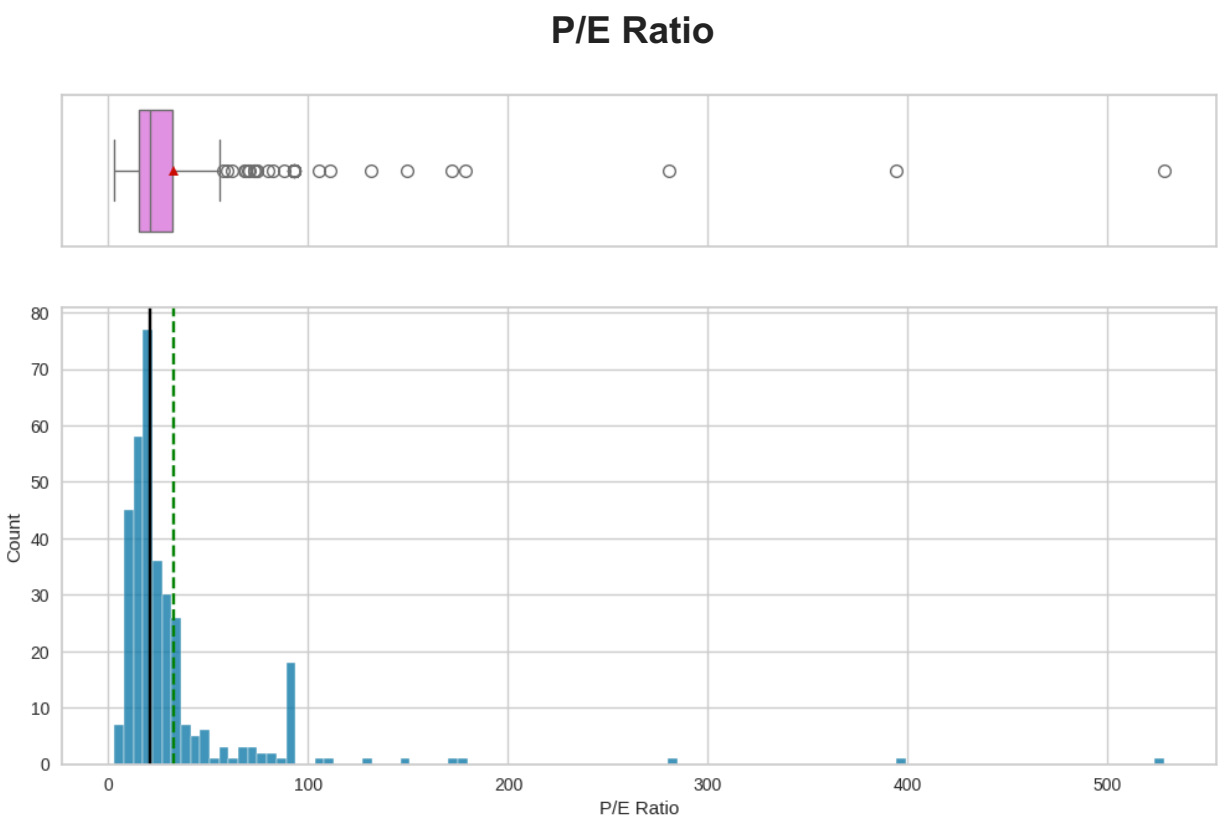
- Skewed right
- There are outliers on both ends
- It appears most companies are reflecting positive earnings per share

EDA _ Univariate Analysis, Cont...



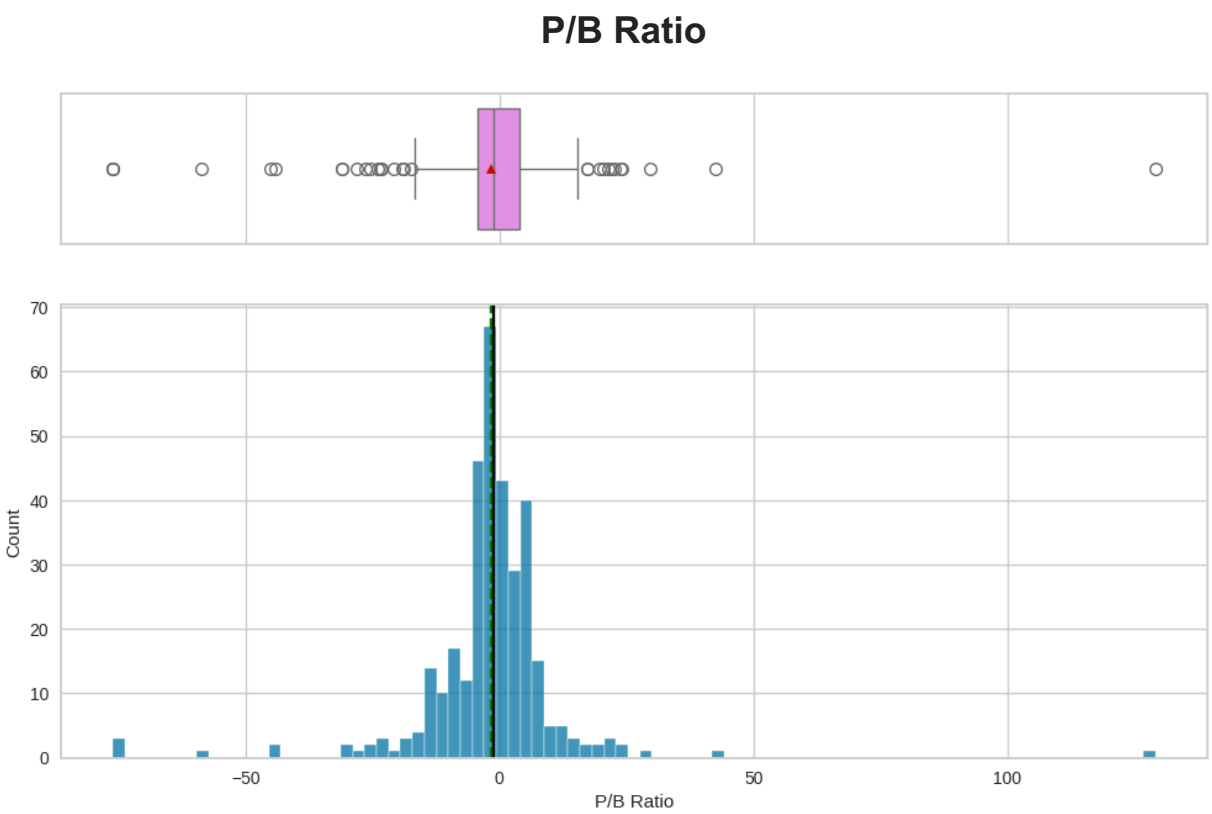
- Highly skewed right
- There are outliers on the positive tail
- No values stand out as highly out of normal range

EDA _ Univariate Analysis, Cont...



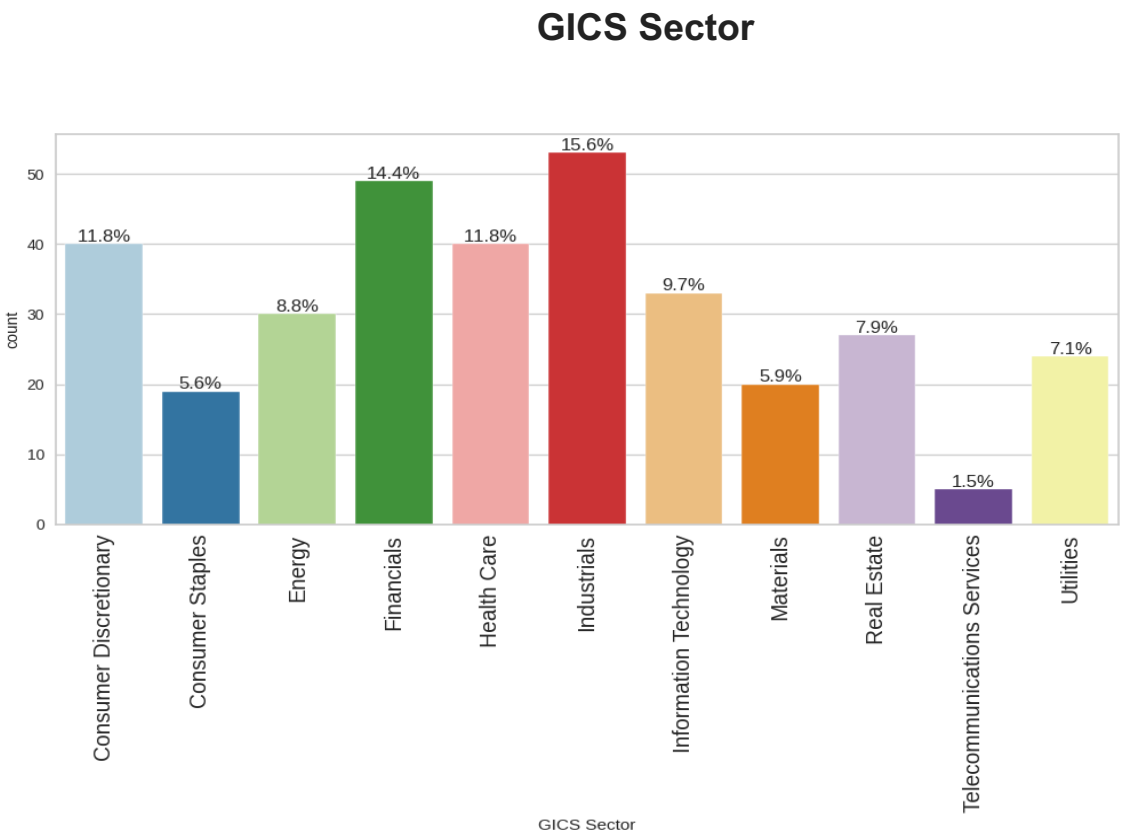
- Highly skewed right
- There are outliers on the positive tail
- No stocks appear to have a negative P/E ratio

EDA _ Univariate Analysis, Cont...



- Skewed slightly right
- There are large outliers on the both ends
- The distribution is centered around zero (0)

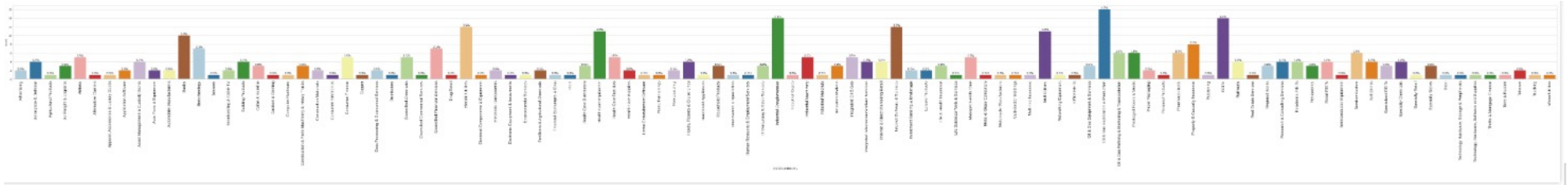
EDA _ Univariate Analysis, Cont...



- There are 11 unique GICS Sectors
- Industrials is the largest sector reflecting 15.6% of the data and Information Services the lowest reflecting 1.5% of the data

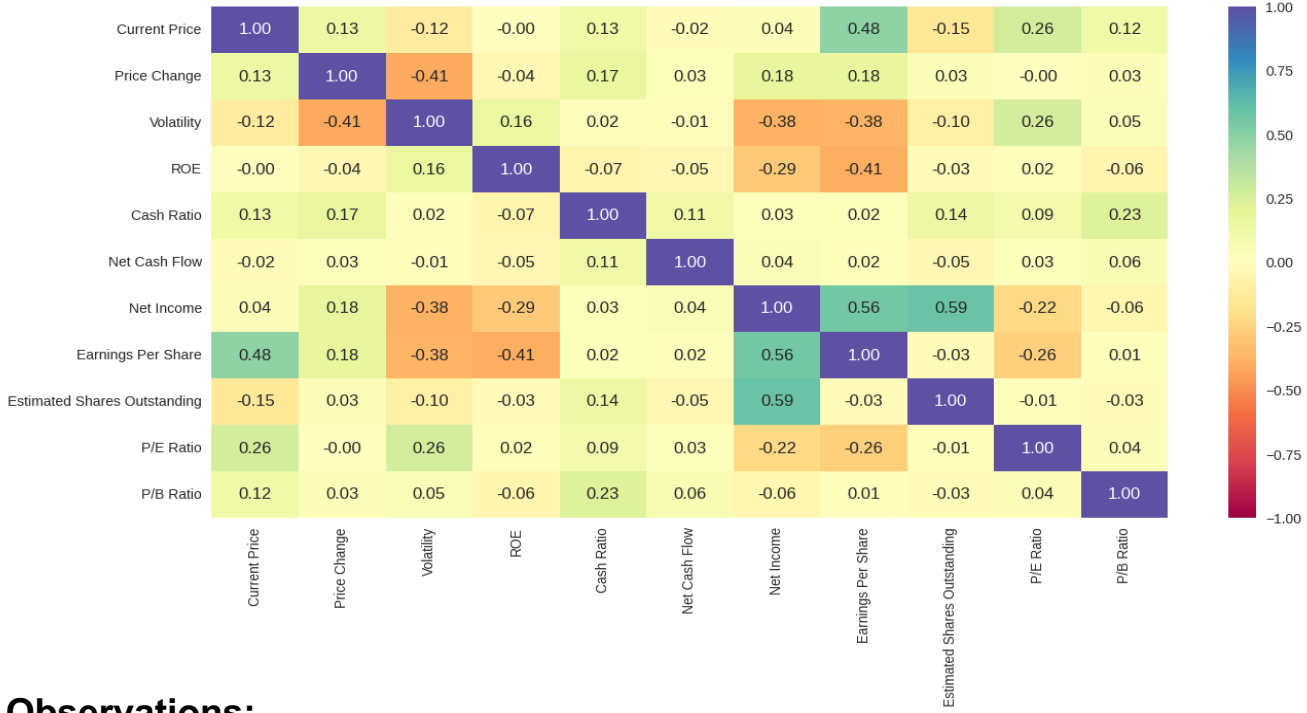
EDA _ Univariate Analysis, Cont...

GICS Sub Industry



- There are 104 Unique GICS Sub Industry categories
- The most dominant sub industry is Oil & Gas Exploration & Production at 4.7%
- Approximately 36% of the data falls within 12 of the GICS Sub Industries

EDA _ Bivariate Analysis

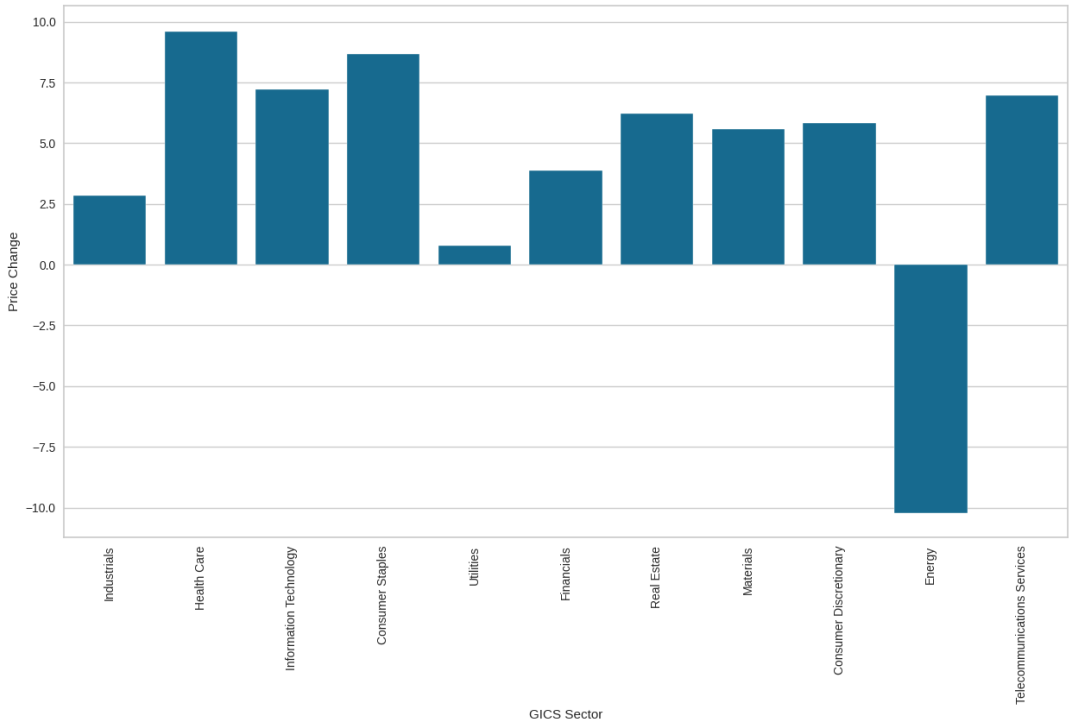


Observations:

- Net Income has a strong positive correlation to Earnings Per Share and Estimated Share Outstanding
- Earning Per Share and Current Price have a positive correlation
- Volatility is negatively correlated to Net Income, Earnings Per Share, and Price Change
- ROE is negatively correlated with Earnings Per Share

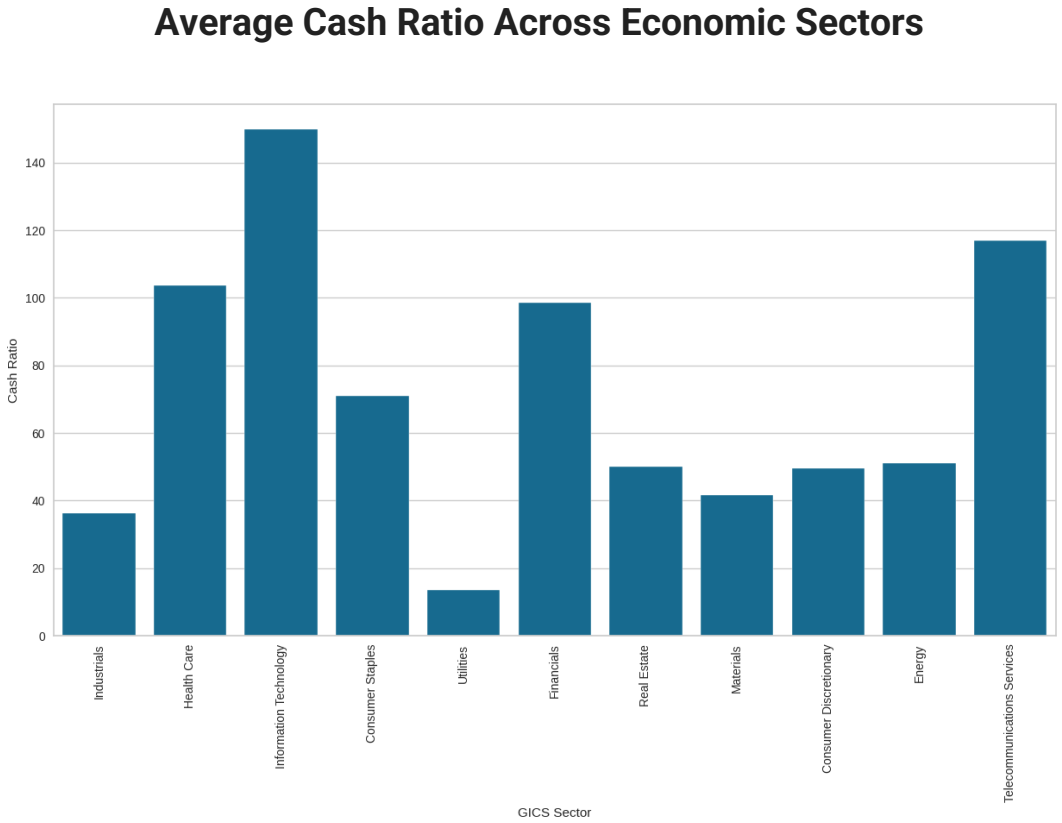
EDA _ Bivariate Analysis, Cont...

Stocks of which economic sector have seen the maximum price increase on average



- Health Care and Customer Staples sectors appear to have the highest price increases.
- Energy sector has a significant negative price decrease on average.
- Utilities demonstrates a minimal price increase

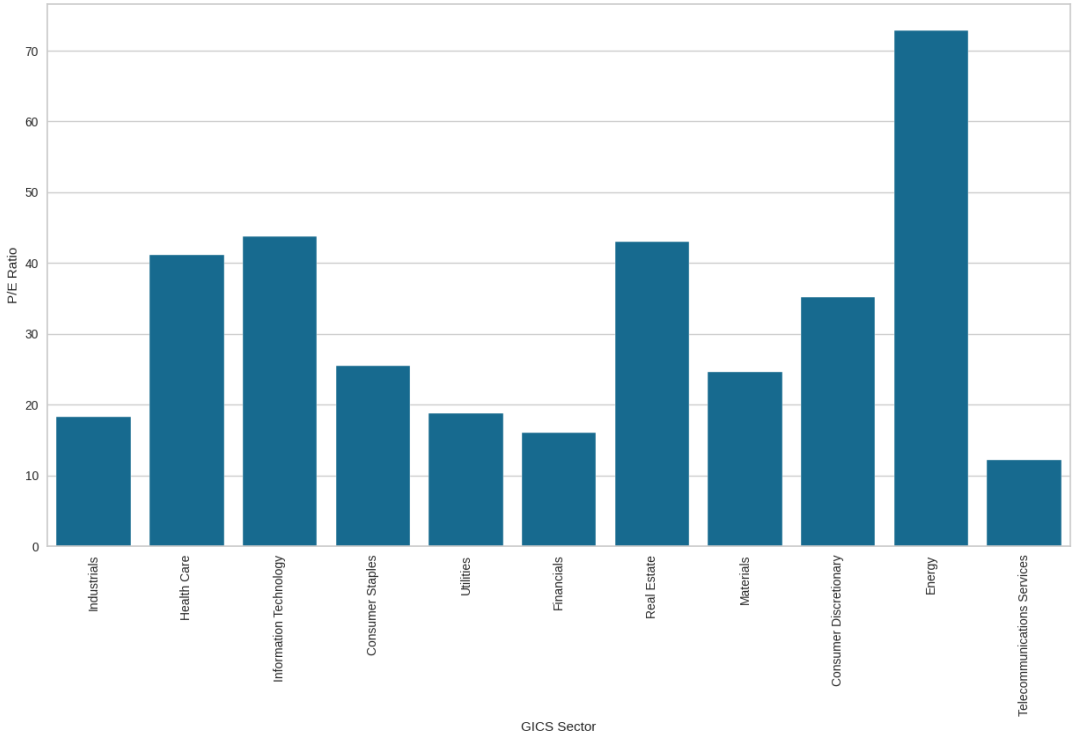
EDA _ Bivariate Analysis, Cont...



- Information Technology has the greatest Cash Ratio across the economic sectors
- Utilities is showing the least average cash ratio across economic sectors

EDA _ Bivariate Analysis, Cont...

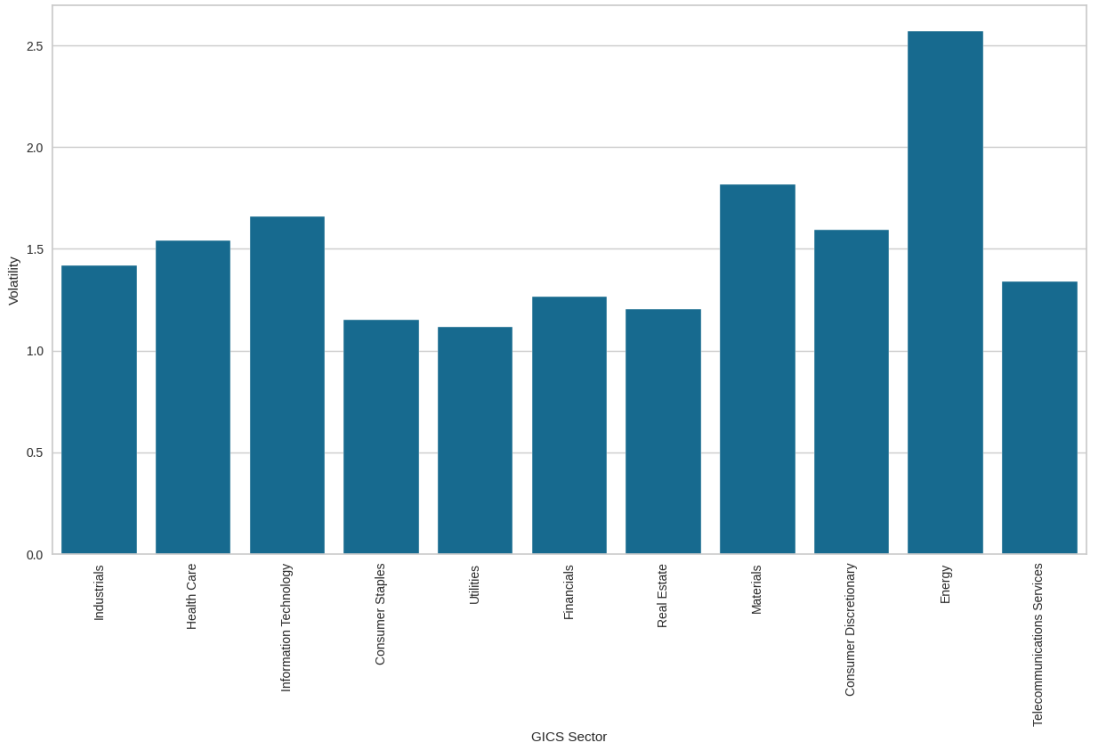
P/E ratio varies, on average, across economic sectors



- Energy sector has the highest P/E ratio on average – significantly higher than the others
- Telecommunications services has the lowest P/E ratio on average

EDA _ Bivariate Analysis, Cont...

View of how volatility varies, on average, across economic sectors



The energy sector has the the highest volatility across the sectors.

Data Processing

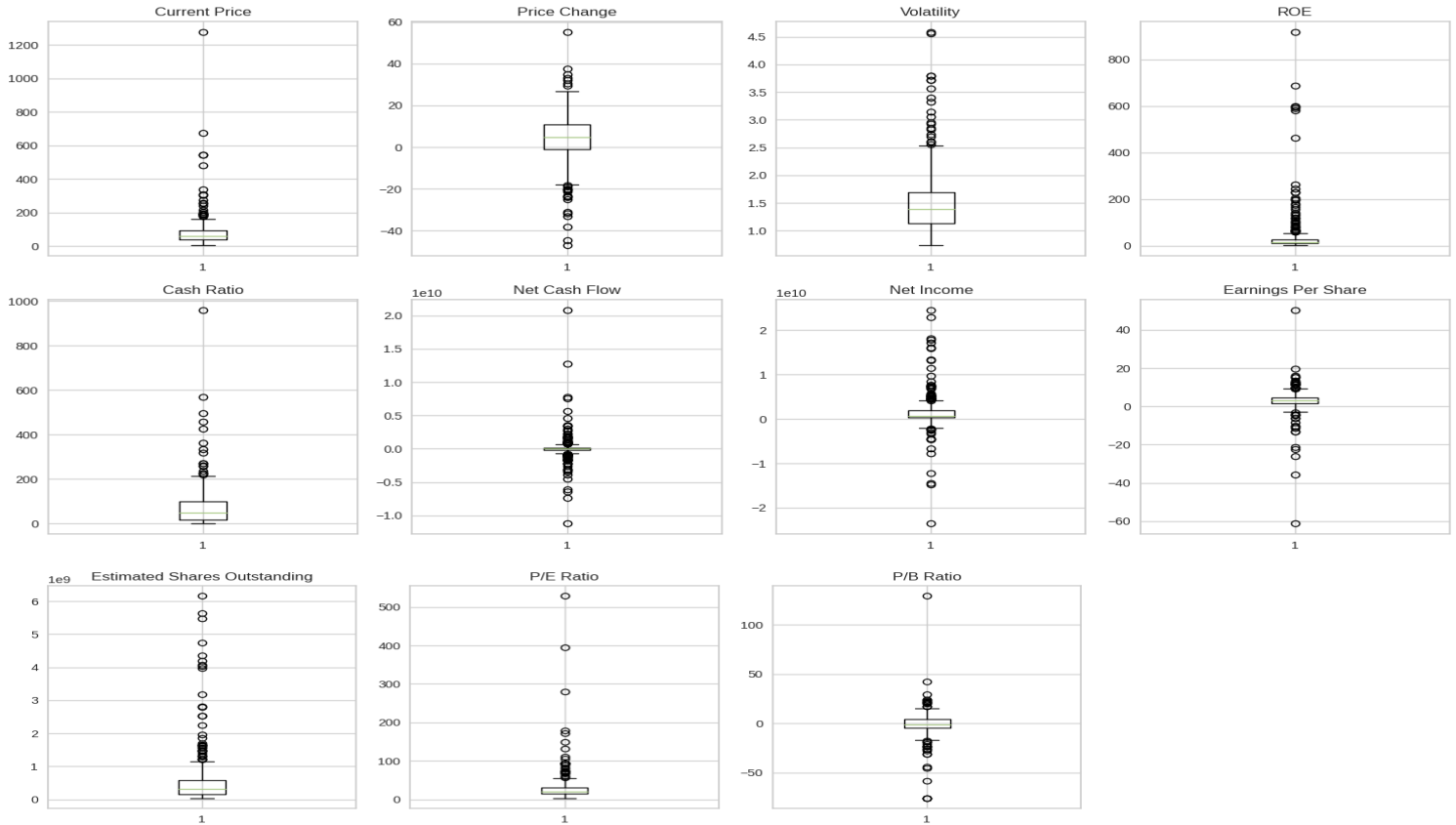
- There are zero (0) duplicate values
- There are zero (0) missing values

```
Ticker Symbol                0
Company                      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
```

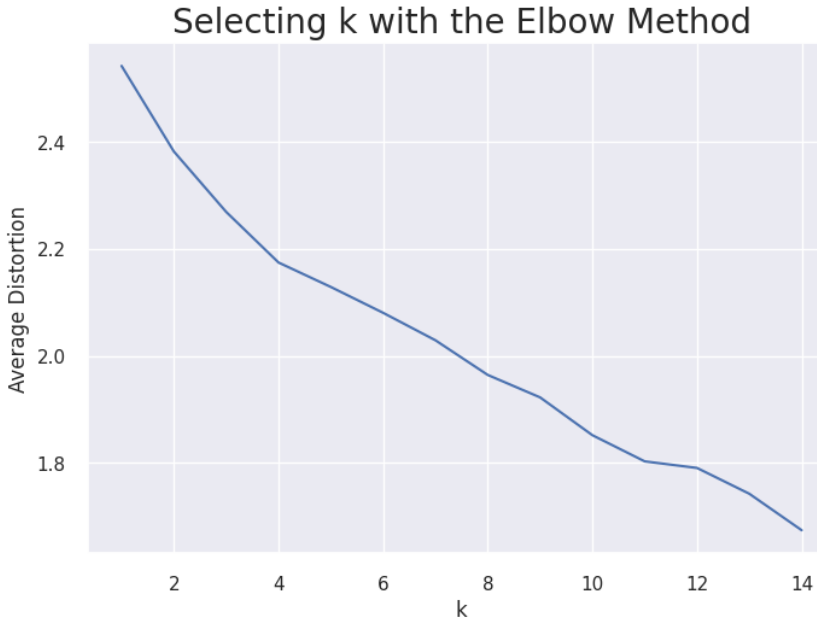
- The data was scaled before clustering

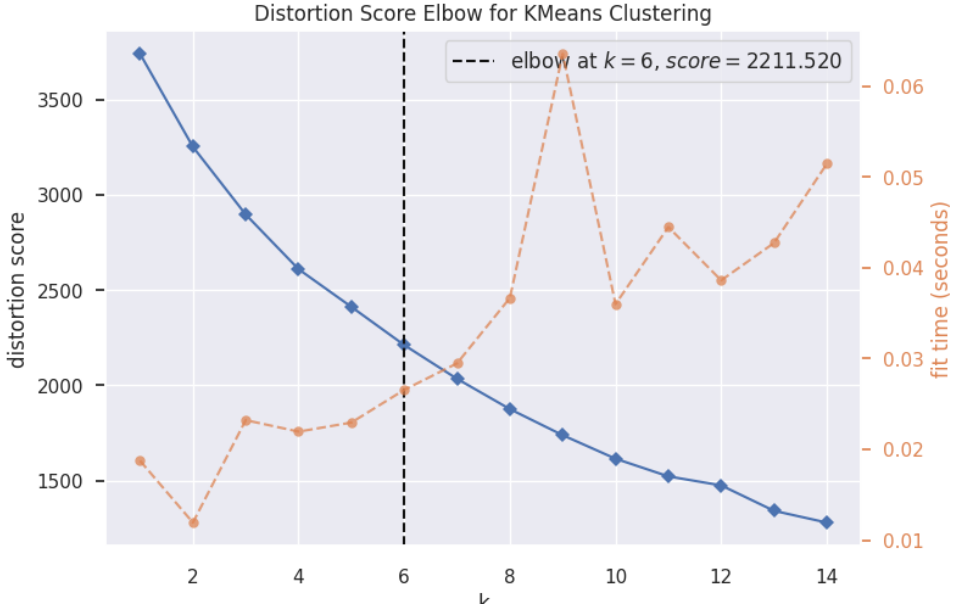
Data Processing, Cont…

Outlier Check



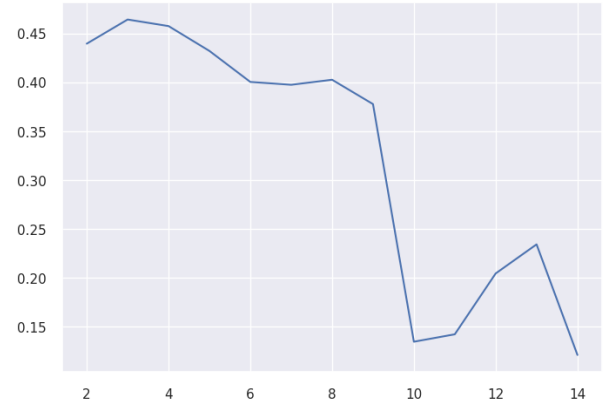
K-Means Clustering Technique, Elbow...

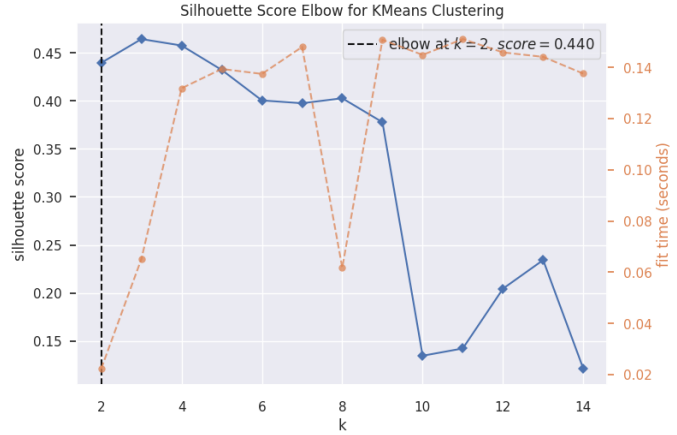


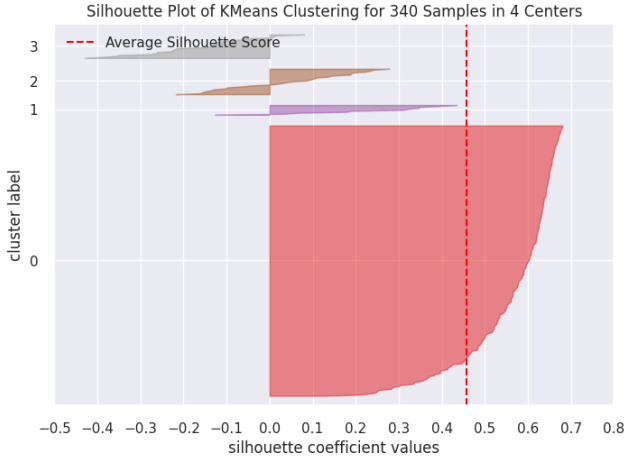


It appears the k score for the elbow method is between 6-7

K-Means Clustering Technique, Silhouette







It appears the score for the silhouette method is 4

K-Means Clustering Technique, Final Model

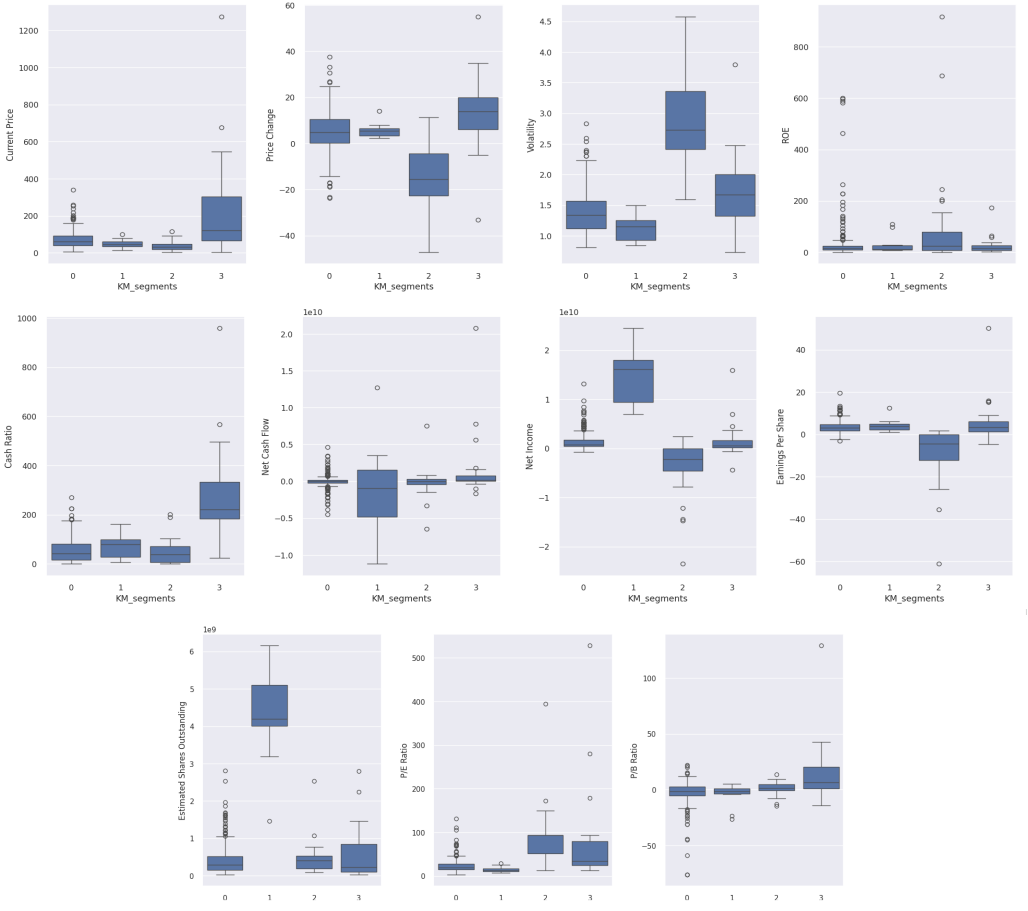
I ran clusters of 4, 5, 6 and found 4 the have the most balanced distributions multiple segments in each cluster

```
KMeans
KMeans(n_clusters=4, random_state=1)
```

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

Cluster 0: Total cluster count of 277 stocks in 11 Segments -- Industrials is the top segment (52)
Cluster 1: Total cluster count of 11 stocks in 7 Segments – Financials is the top segment (3)
Cluster 2: Total cluster count of 27 stocks in 4 Segments – Energy is the top segment (22)
Cluster 3: Total cluster count of 25 stocks in 8 Segments – Health Care is the top segment (9)

K-Means Clustering Technique, Final Model, Cont...



K-Means Clustering Technique, Final Model, Cont...

Cluster 0
- Current Price moderate to high
- Price Change is moderate to high
- Volatility is moderate
- ROE is moderate to high
- Cash Ratio is moderate
- Net Cash Flow is low
- Net Income is moderate
- Earning Per Share is moderate
- Estimated Shares Outstanding is moderate
- P/E Ratio is moderate
- P/B Ratio is moderate to low

Cluster 1
- Current Price moderate
- Price Change is moderate
- Volatility is moderate
- ROE is moderate
- Cash Ratio is moderate
- Net Cash Flow is moderate to low
- Net Income is high to very high
- Earning Per Share is moderate to low
- Estimated Shares Outstanding is very high
- P/E Ratio is moderate
- P/B Ratio is moderate to low

Cluster 2
- Current Price moderate
- Price Change is low to very low
- Volatility is high
- ROE is moderate to high
- Cash Ratio is moderate
- Net Cash Flow is moderate to low
- Net Income is moderate to low
- Earning Per Share is moderate to low
- Estimated Shares Outstanding moderate
- P/E Ratio is moderate to high
- P/B Ratio is moderate

Cluster 3
- Current Price is high to very high
- Price Change is high to very high
- Volatility is moderate
- ROE is moderate
- Cash Ratio is high to very high
- Net Cash Flow is moderate
- Net Income is moderate
- Earning Per Share is moderate
- Estimated Shares Outstanding is moderate
- P/E Ratio is moderate to high
- P/B Ratio is moderate to high

K-Means Clustering Technique, Final Model, Cont...

Cluster 0

- 277 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, Telecommunications Services, Utilities)
- Companies in this cluster have:
  - * Moderate to higher price changes
  - * ROE is moderate to high
  - * Current pricing are relatively moderate
  - * This cluster does not have a real significant variable to stands out

Cluster 1

- 11 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Telecommunications Services)
- Companies in this cluster have:
  - * Higher than normal net income
  - * Estimated Shares Outstanding is very high
  - * This cluster does have two significant variable to stands out – Net Income and Estimated Shares Outstanding. Much higher that the other clusters

Cluster 2

- 27 Stocks in the following sectors (Energy, Industrials, Information Technology, Materials)
- Companies in this cluster have:
  - * Higher P/E Ratios
  - * Low Earning per Share
  - * Lower net income
  - * High Volatility
  - * This cluster does have one significant variable to stands out -- Volatility. Much higher that the other clusters

Cluster 3

- 25 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Information Technology, Real Estate, Telecommunications Services)
- Companies in this cluster have:
  - * Higher stock prices
  - * Higher cash ratios
  - * Higher price changes
  - * This cluster does have one significant variable to stands out – Price Change. Much higher that the other clusters

Hierarchical Clustering Technique

- The following Hierarchical Clustering methods were used:

*. Distance Metrics
  - Euclidean
  - Chebyshev
  - Mahalanobis
  - Cityblock

* Linkage Methods
  - Single
  - Complete
  - Average
  - Weighted

- Observations using different linkage methods:

```
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 0.9422540609560814, which is obtained with Euclidean distance and average linkage.*

Hierarchical Clustering Technique, Cont

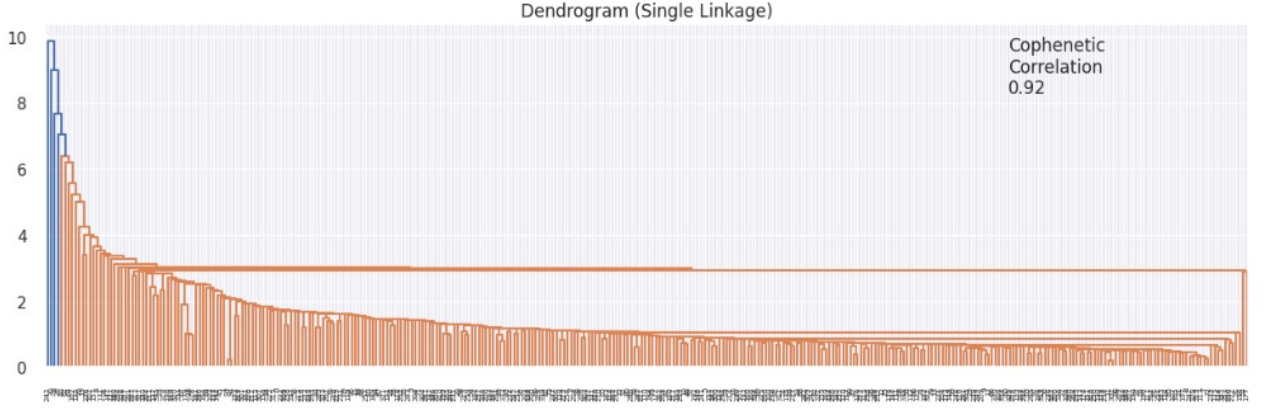
- Observations from Cophenetic correlation for different combinations of distance and metrics

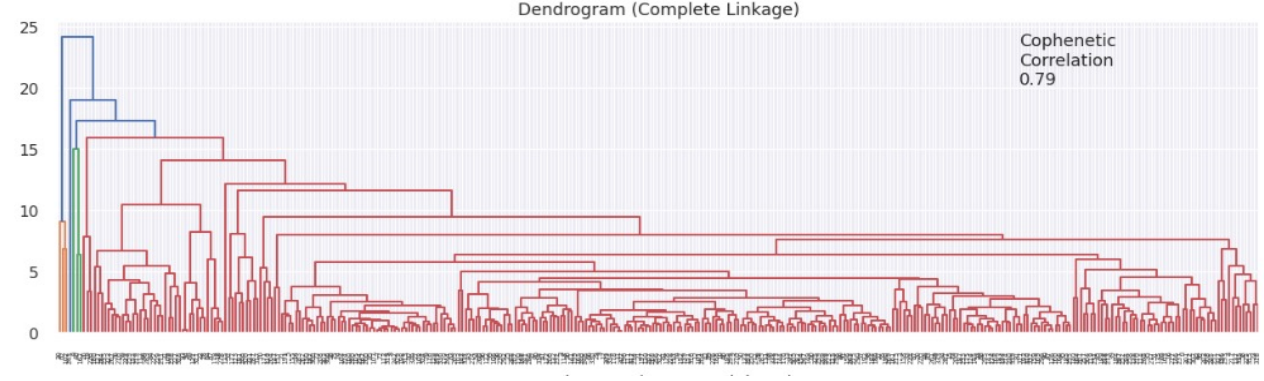
```
Cophenetic correlation for single linkage is 0.9232271494002922.
Cophenetic correlation for complete linkage is 0.7873280186580672.
Cophenetic correlation for average linkage is 0.9422540609560814.
Cophenetic correlation for centroid linkage is 0.9314012446828154.
Cophenetic correlation for ward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

* Highest cophenetic correlation is 0.9422540609560814, which is obtained with average linkage *

Hierarchical Clustering Technique, Cont...

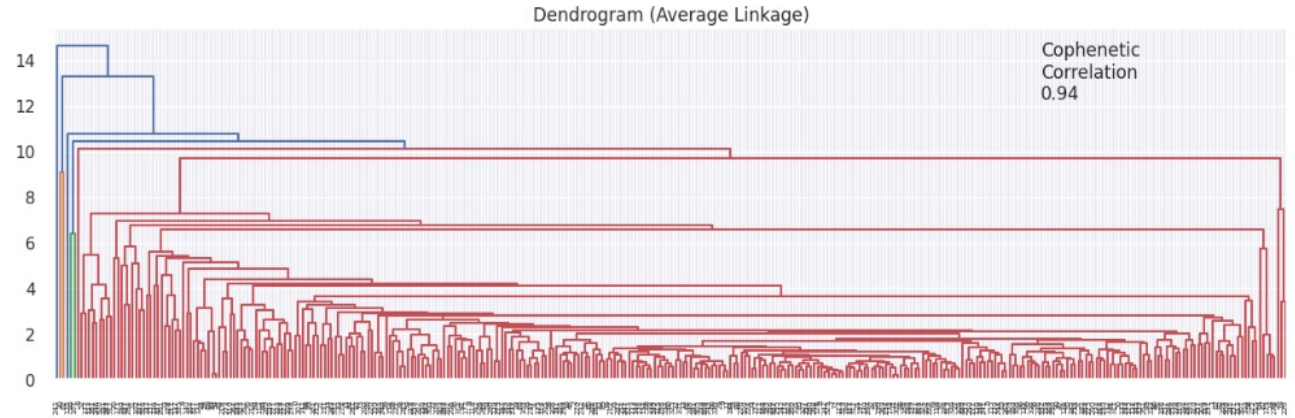
Dendrograms for linkage methods used and their observations

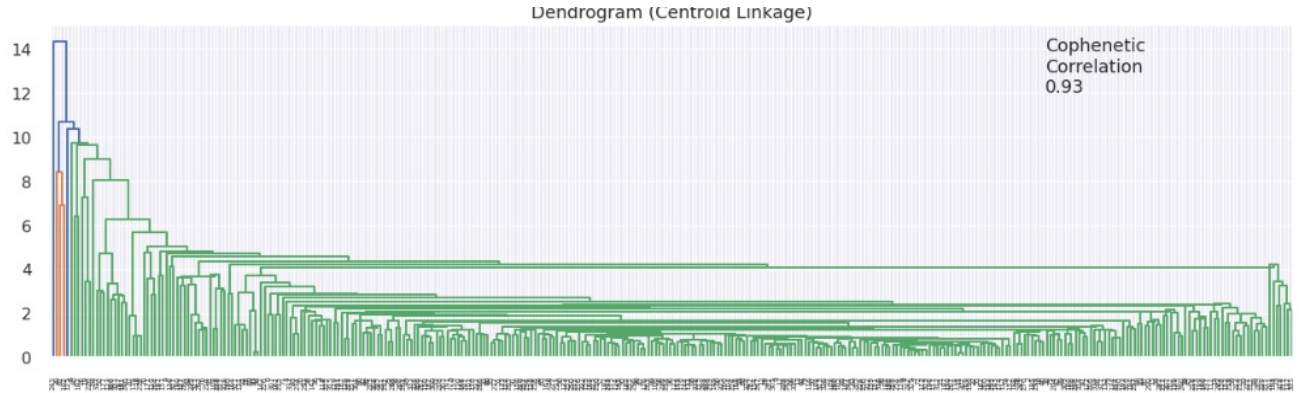




Hierarchical Clustering Technique, Cont...

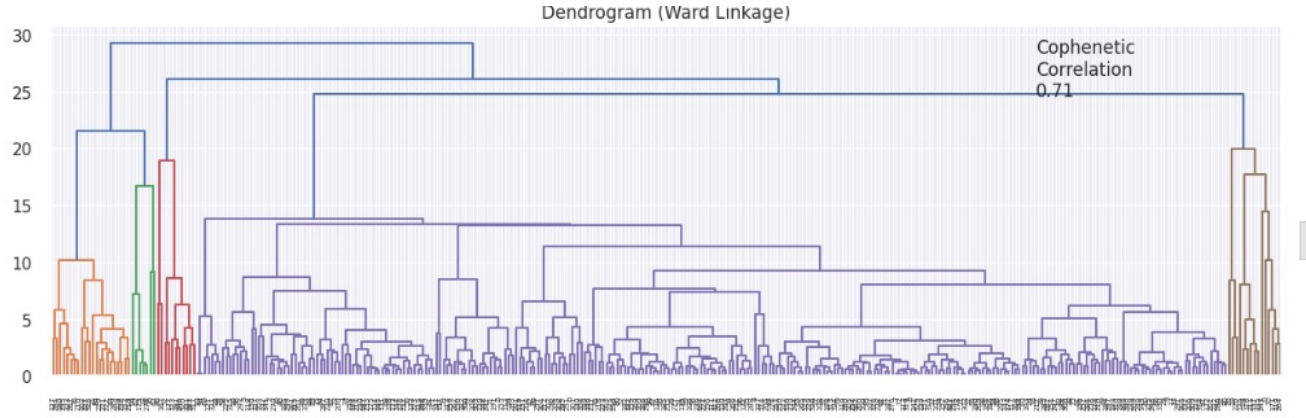
Dendrograms for linkage methods used and their observations

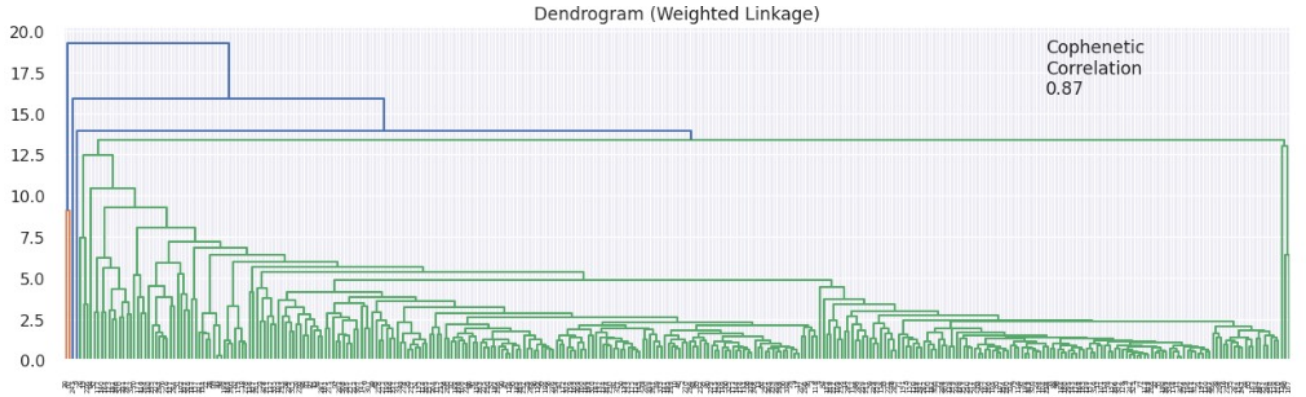




Hierarchical Clustering Technique, Cont...

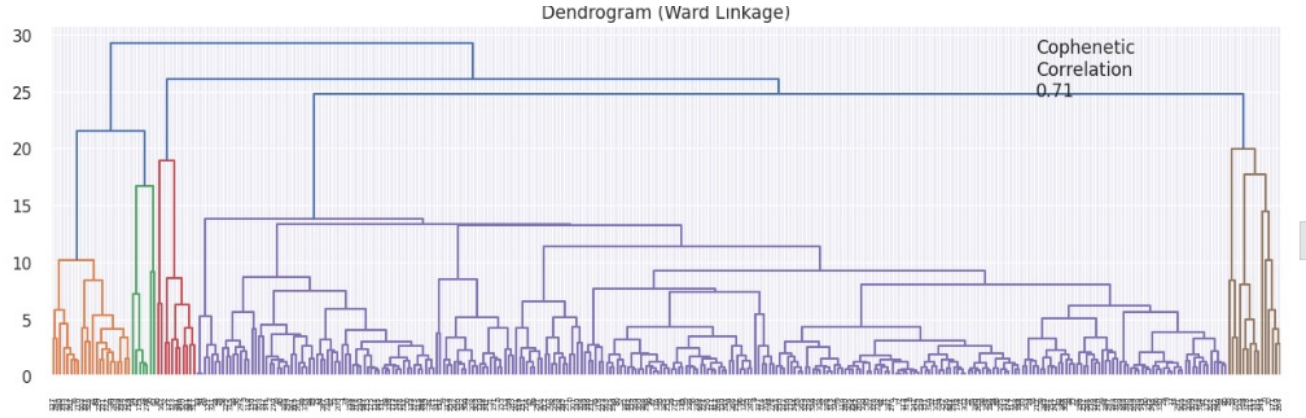
Dendrograms for linkage methods used and their observations

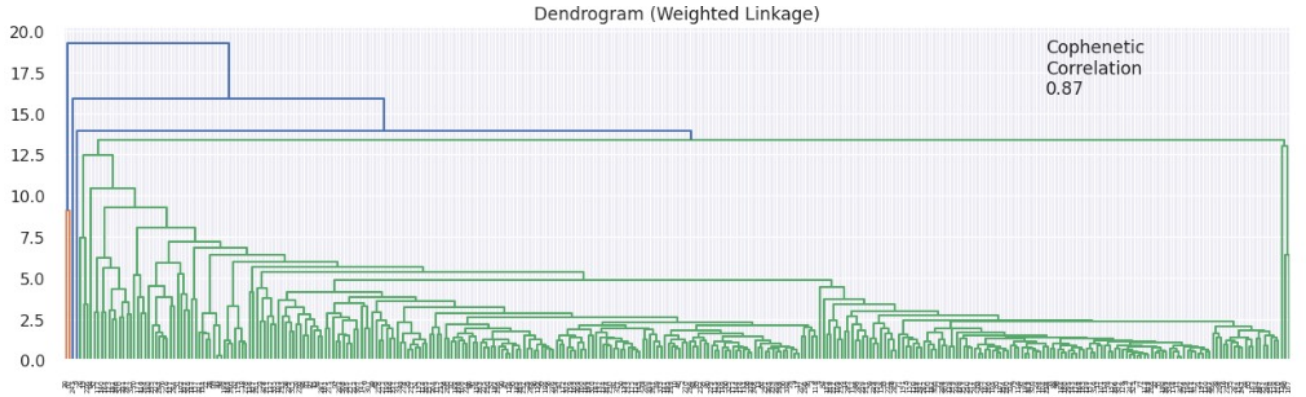




Hierarchical Clustering Technique, Cont...

Dendrograms for linkage methods used and their observations





Hierarchical Clustering Technique, Cont...

- Dendrogram Cophenetic Coefficient

	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 0.942254, which is obtained with average linkage **

Hierarchical Clustering Technique, Cont...

- Cluster Profiling -- It appears 5 is the right number of clusters

```
AgglomerativeClustering
AgglomerativeClustering(affinity='euclidean', linkage='average', n_clusters=5)
```

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	77.884243	4.105986	1.516865	35.320359	66.775449	-32825817.365269	1535255703.592814	2.903308	559027333.145509	32.437511	-1.781988	334
1	25.640000	11.237908	1.322355	12.500000	130.500000	16755500000.000000	13654000000.000000	3.295000	2791829362.100000	13.649696	1.508484	2
2	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1.565141	2
3	104.660004	16.224320	1.320606	8.000000	958.000000	592000000.000000	3669000000.000000	1.310000	2800763359.000000	79.893133	5.884467	1
4	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000	50.090000	50935516.070000	25.453183	-1.052429	1

Cluster 0: Total cluster count of 334 stocks in 11 Segments -- Industrials is the top segment (53)
Cluster 1: Total cluster count of 2 stocks in 2 Segments – Financials and Information Technology
Cluster 2: Total cluster count of 2 stocks in 1 Segment – Energy
Cluster 3: Total cluster count of 1 stock in 1 Segment – Information Technology
Cluster 4: Total cluster count of 1 stock in 1 Segment – Consumer Discretionary

Hierarchical Clustering Technique, Cont...

Cluster 0
- Current Price is moderate to high
- Price Change is moderate to high
- Volatility is moderate
- ROE is moderate
- Cash Ratio is moderate to high
- Net Cash Flow is moderate
- Net Income is moderate
- Earning Per Share is moderate
- Estimated Shares Outstanding is moderate
- P/E Ratio is moderate
- P/B Ratio is moderate to low

Cluster 1
- Current Price is moderate
- Price Change is high
- Volatility is moderate
- ROE is moderate
- Cash Ratio is high
- Net Cash Flow is high
- Net Income is high to very high
- Earning Per Share is moderate
- Estimated Shares Outstanding is high to very high
- P/E Ratio is moderate
- P/B Ratio is moderate to low

Cluster 2
- Current Price is moderate
- Price Change is low to very low
- Volatility is high
- ROE is very high
- Cash Ratio is moderate
- Net Cash Flow is moderate to low
- Net Income is low to very low
- Earning Per Share is low to very low
- Estimated Shares Outstanding moderate
- P/E Ratio is moderate to high
- P/B Ratio is moderate

Cluster 3
- Current Price is high
- Price Change is high to very high
- Volatility is high
- ROE is moderate
- Cash Ratio is very high
- Net Cash Flow is moderate
- Net Income is moderate to high
- Earning Per Share is moderate
- Estimated Shares Outstanding is high
- P/E Ratio is moderate
- P/B Ratio is moderate

Cluster 4
- Current Price is very high
- Price Change is moderate
- Volatility is moderate
- ROE is moderate
- Cash Ratio is high
- Net Cash Flow is moderate to low
- Net Income is moderate
- Earning Per Share is very high
- Estimated Shares Outstanding is moderate
- P/E Ratio is moderate
- P/B Ratio is moderate

Hierarchical Clustering Technique, Cont…

Cluster 0
- 344 Stocks in the following sectors (Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, Telecommunications Services, Utilities)
- Companies in this cluster have:
  * Moderate Price Changes
  * ROE is moderate
  * Cash Ratios are relatively moderate
  * This cluster does not have a real significant variable to stands out

Cluster 1
- 2 Stocks in the following sectors (Financials& Information Technology)
- Companies in this cluster have:
  * Higher than normal Net Cash Flow
  * Higher P/E Ratios
  * Higher Net Income
  * Higher Cash Ratios
  * This cluster does have two significant variable to stands out – Net Income and Net Cash Flow. Higher than the other clusters

Cluster 2
- 2 Stocks in the following sector (Energy)
- Companies in this cluster have:
  * Moderate to Low Price Changes
  * Very Low Earnings Per Share
  * Lower net income
  * High Volatility
  * This cluster does have two significant variable to stands out – Earning Per Share much lower than others and Net Cash Flow. Higher than the other clusters

Cluster 3
- 1 Stock in the following sector (Information Technology)
- Company in this cluster have:
  * Higher current prices
  * Higher cash ratios
  * Higher price changes
  * Higher Estimated Shares Outstanding
  * This cluster does have one significant variable to stands out – Cash Ratio. Much higher that the other clusters

Cluster 4
- 1 stock in the following sector (Consumer Discretionary)
- Company in this cluster have:
  * Very high Current Price
  * Lower Volatility
  * Higher Cash Ratio
  * Very high Earning Per Share
  * This cluster does have two significant variables that stand out – Current Price and Earnings Per Share higher than the other clusters

K-Means vs Hierarchical Clustering

- Which clustering technique took less time for execution?
  - K-Means Clustering was much faster to execute

- Which clustering technique gave you more distinct clusters, or are they the same?
  - K-Means Clustering provided much more distinct clusters

- How many observations are there in the similar clusters of both algorithms?
  - Cluster 0 as the most similar for both: K-Means had 277 stocks/11 segments and Hierarchical clustering had 334 stocks /11 segments

- How many clusters are obtained as the appropriate number of clusters from both algorithms?
  - K-Means had 4 clusters and Hierarchical Clustering had 5 clusters