TradeAhead-Unsupervised Learning

Michail Mersinias

06/06/2023

Contents / Agenda

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

Executive Summary

- The four investment clusters, resulted from the both clustering methods on the Trade Ahead financial data of 340 companies, are the following:
    - **Unicorn Cluster:** Stocks have a high current price. They are overvalued (high P/E ratio and high P/B ratio), however they are reliable as they exhibit low volatility combined with a positive and very high net cash flow and high earnings per share. Finally, they have the highest price change, indicating their continuous rise.
    - **Opportunity Cluster:** Stocks have a very low current price and very high net income. They also have a growth potential as they are also currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit very low volatility combined with a high net cash flow and medium earnings per share. Finally, they have a positive and high price change and the highest estimated shares outstanding, indicating their rise.
    - **Normal Cluster:** Stocks have a low current price. They have a growth potential as they are currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit low volatility combined with medium earnings per share and positive price change. However, while the price change is positive, the remaining metrics are of low magnitude, indicating safety but only a low or moderate rise.
    - **Junk Cluster:** Stocks have a low current price but a very high volatility and ROE. They are currently overvalued (high P/E ratio), and the remaining metrics also indicate bad performance with significantly negative net income, earnings per share and price change. Plus, moderately negative net cash flow and cash ratio.

Executive Summary

- The two clustering methods on the Trade Ahead financial data of 340 companies, are the following:
  - **K-Means Clustering Model:** We chose 4 clusters, as indicated by the elbow method. The silhouette score for k=4 was the second highest but very close to the highest one (k=3). Thus, we chose 4 clusters for our K-Means final model.
  - **Hierarchical Clustering (Agglomerative) Model:** After evaluating the Cophenetic correlation scores, we chose the Euclidean distance as the distance component. Afterwards, for choosing the linkage component, we evaluated the the Cophenetic correlation scores again using the Euclidean distance. Although average linkage displayed the highest score, after a more careful examination of the dendrograms, we decided to select the ward linkage as our component, because it provided a more distinct clustering result. From the dendrogram, we also selected the number of clusters to be 4 (k=4) for our Hierarchical Clustering final model.
  - **Conclusion:** Both clustering methods provide the same number of distinct clusters (k=4), with approximately the same attributes for each cluster. If we wish to optimize for performance (greater diversification), we choose the **Hierarchical Clustering (Agglomerative) Model**, while if we want to optimize for speed, we choose the **K-Means Clustering Model** which is slightly faster.
  - **Recommendation:** In both cases, we recommend a portfolio containing a mix of stocks from the *Unicorn* and *Opportunity* clusters, plus selected individual stocks from *Normal* cluster and none from the *Junk* clusters.

Business Problem Overview and Solution Approach

- Problem Definition:
  - Trade Ahead is a financial consultancy firm who provide their customers with personalized investment strategies. They provide data comprising stock price and some financial indicators for companies listed under the New York Stock Exchange.
  - The objective is to analyze the data, group the stocks based on the attributes provided, and share insights about the characteristics of each group.
- Solution Approach and Methodology:
  - For EDA, both univariate and multivariate analysis will be performed to find insights.
  - For Data Preprocessing, missing value imputation and feature engineering will be performed.
  - For Model Building, we will use the following unsupervised learning models: K-Means Clustering and Hierarchical (Agglomerative) Clustering.

EDA Results: Data Overview

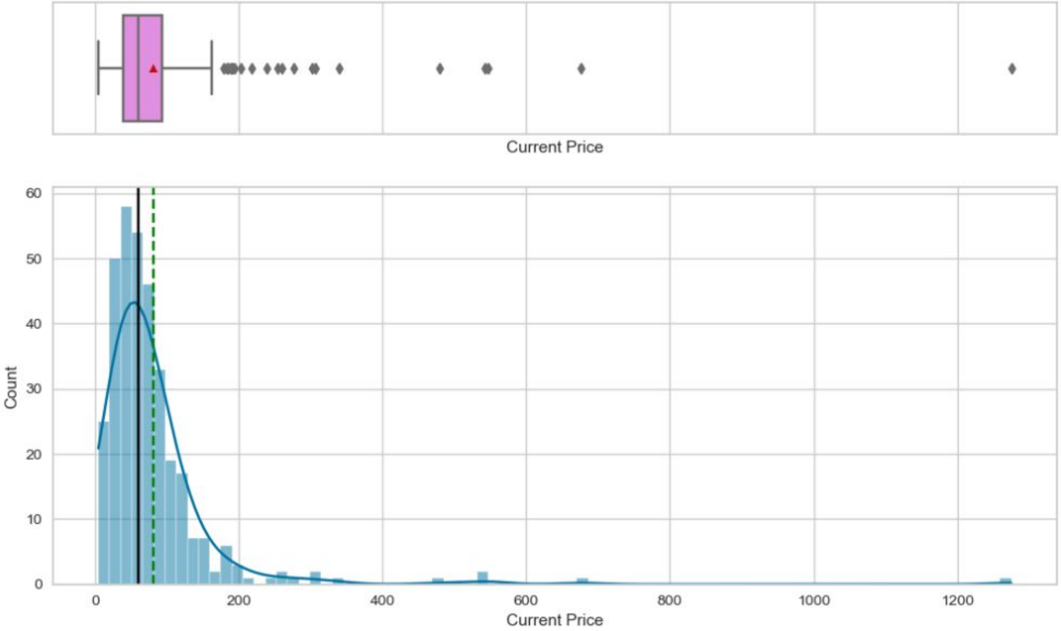
- The dataset is comprised of 340 rows and 15 columns.
- The columns are as follows: Ticker Symbol, Security, GICS Sector, GICS Sub Industry, Current Price, Price Change, Volatility, ROE, Cash Ratio, Net Cash Flow, Net Income, Earnings Per Share, Estimated Shares Outstanding, P/E Ratio, P/B Ratio.
- Each row represents a company listed in the New York Stock Exchange.
- There are no duplicate values in the dataset.
- There are no missing values in the dataset.

EDA Results: Univariate Analysis

- In this section, we perform univariate analysis on the data.
    - For each attribute, we perform a descriptive statistical analysis where only that attribute is involved as a variable.
    - We also analyze the corresponding data and present the distribution of the attribute, with confidence intervals to signify variance and statistical significance.
    - Finally, we write the conclusion based on both quantitative and qualitative observations.

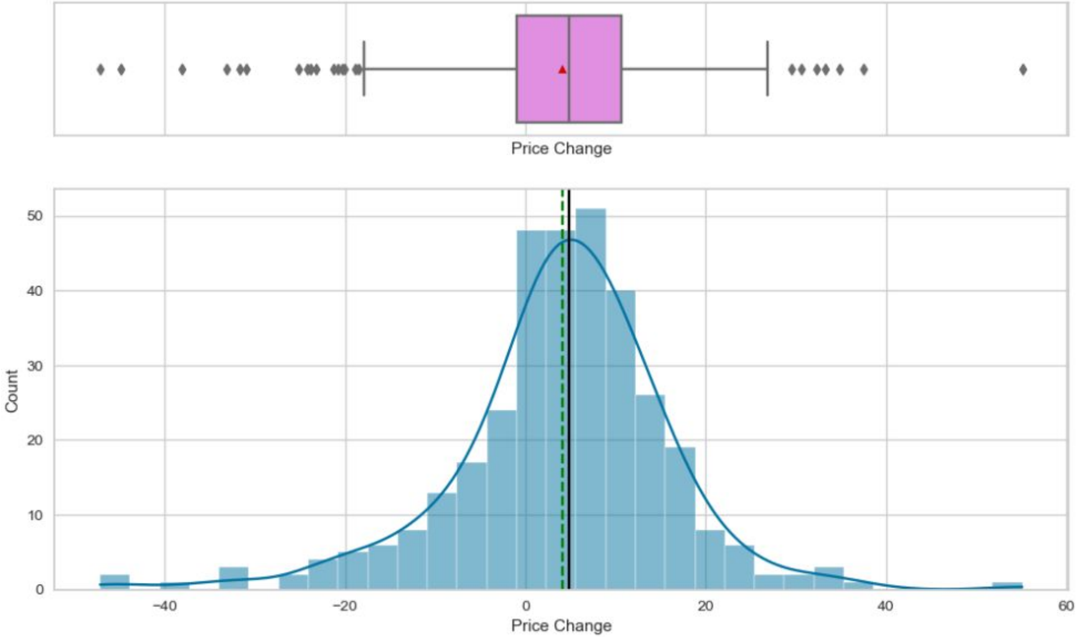
EDA Results: Univariate Analysis

- Analyzing the Current Price attribute, we report that the average value is 80.86 with a standard deviation of 98.05. A boxplot and a countplot are presented to depict the distribution more clearly.



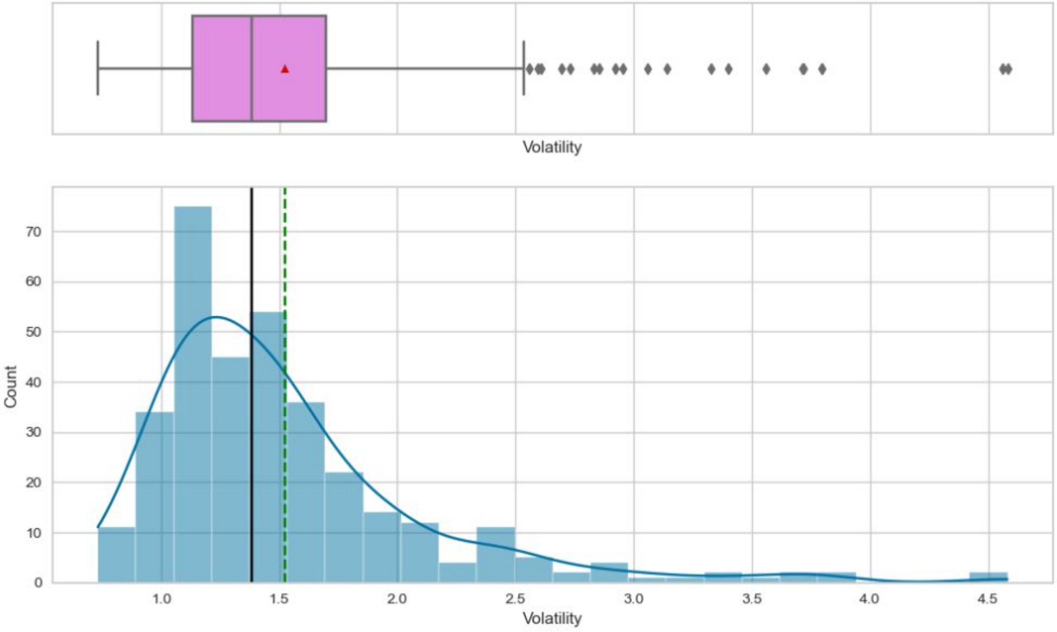
EDA Results: Univariate Analysis

- Analyzing the Price Change attribute, we report that the average value is 4.07 with a standard deviation of 12.00. A boxplot and a countplot are presented to depict the distribution more clearly.



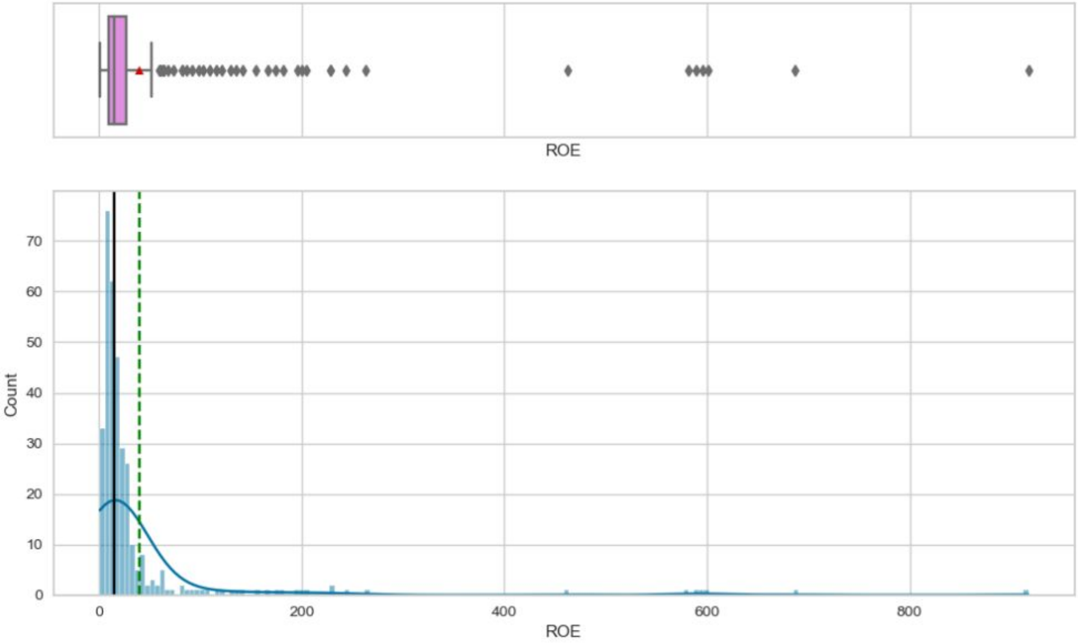
EDA Results: Univariate Analysis

- Analyzing the Volatility attribute, we report that the average value is 1.52 with a standard deviation of 0.59. A boxplot and a countplot are presented to depict the distribution more clearly.



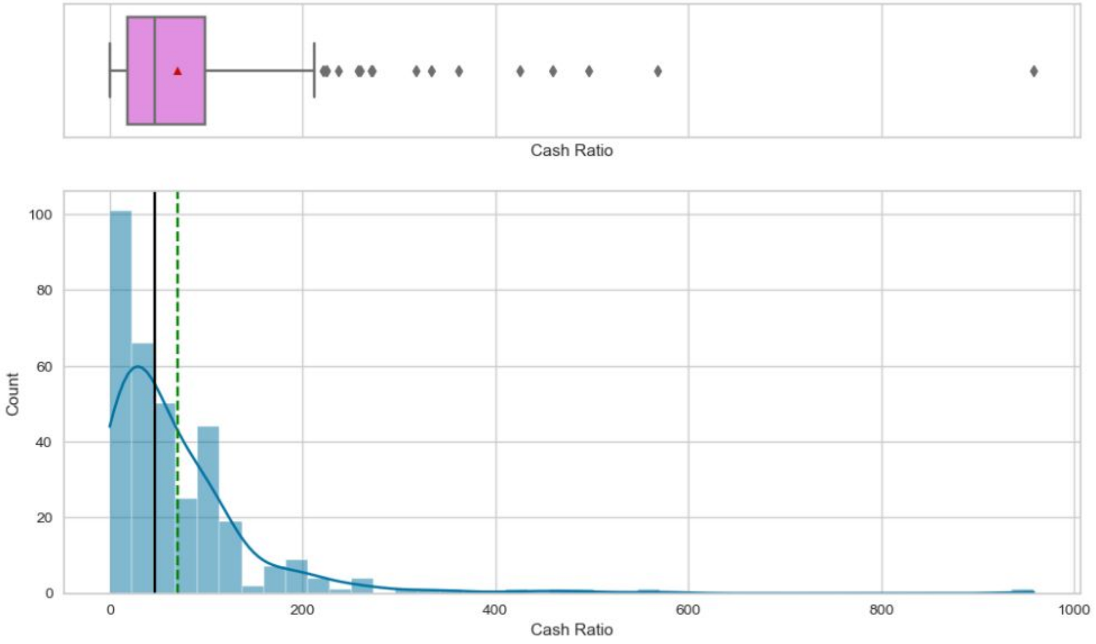
EDA Results: Univariate Analysis

- Analyzing the ROE attribute, we report that the average value is 39.59 with a standard deviation of 96.54. A boxplot and a countplot are presented to depict the distribution more clearly.



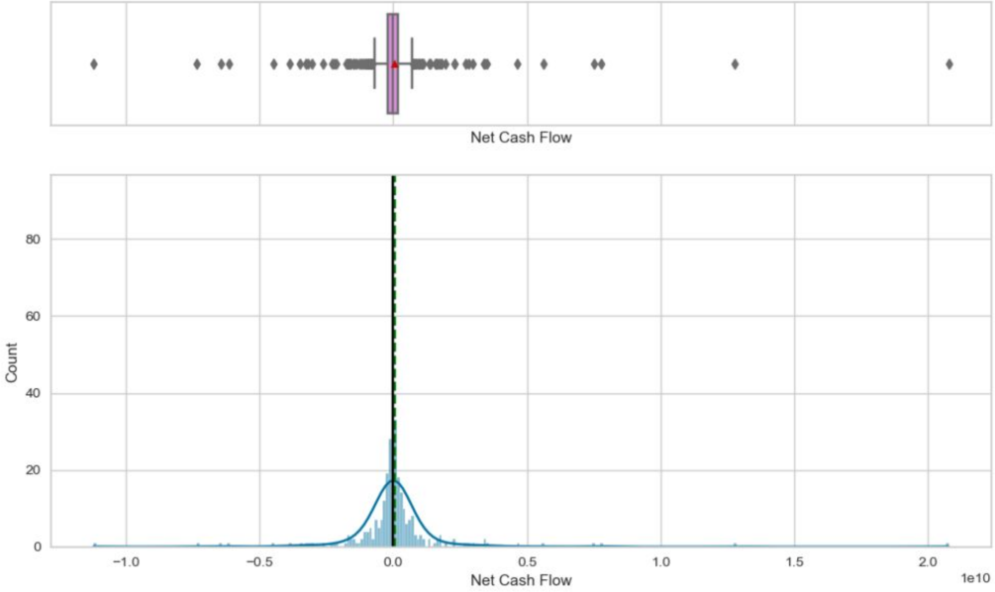
EDA Results: Univariate Analysis

- Analyzing the Cash Ratio attribute, we report that the average value is 70.02 with a standard deviation of 90.42. A boxplot and a countplot are presented to depict the distribution more clearly.



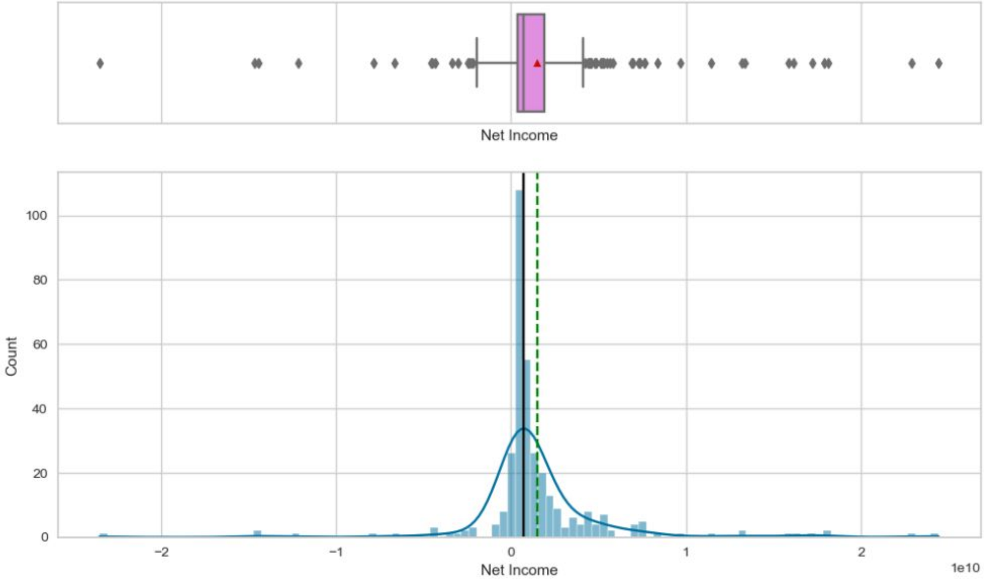
EDA Results: Univariate Analysis

- Analyzing the Net Cash Flow attribute, we report that the average value is 55537620 with a standard deviation of 1946365312. A boxplot and a countplot are presented to depict the distribution more clearly.



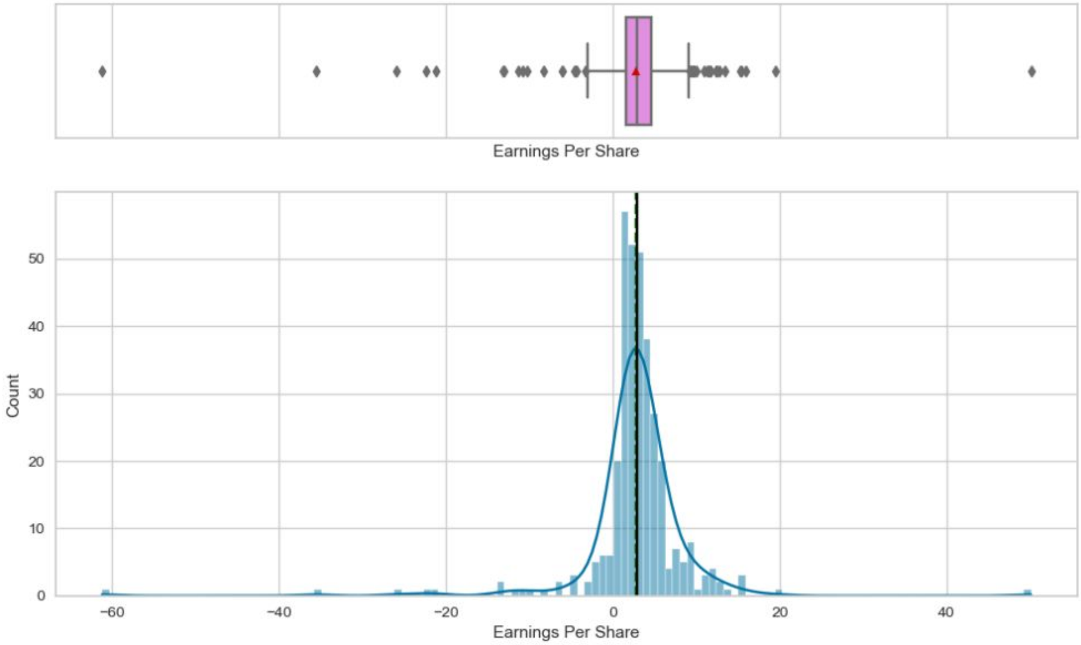
EDA Results: Univariate Analysis

- Analyzing the Net Income attribute, we report that the average value is 1494384602 with a standard deviation of 3940150279. A boxplot and a countplot are presented to depict the distribution more clearly.



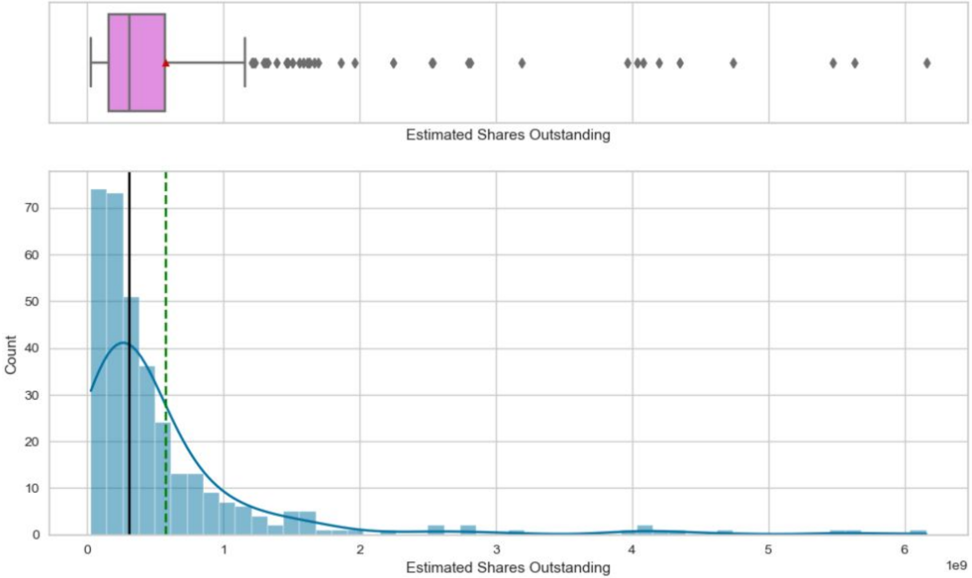
EDA Results: Univariate Analysis

- Analyzing the Earnings Per Share attribute, we report that the average value is 2.77 with a standard deviation of 6.58. A boxplot and a countplot are presented to depict the distribution more clearly.



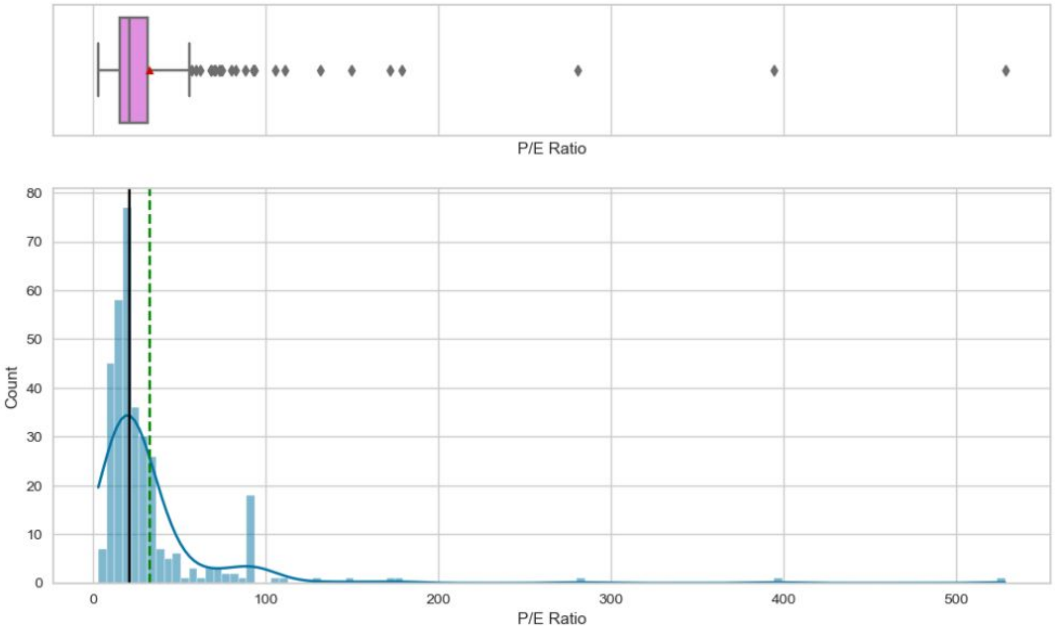
EDA Results: Univariate Analysis

- Analyzing the Estimated Shares Outstanding attribute, we report that the average value is 577028337 with a standard deviation of 845849595. A boxplot and a countplot are presented to depict the distribution more clearly.



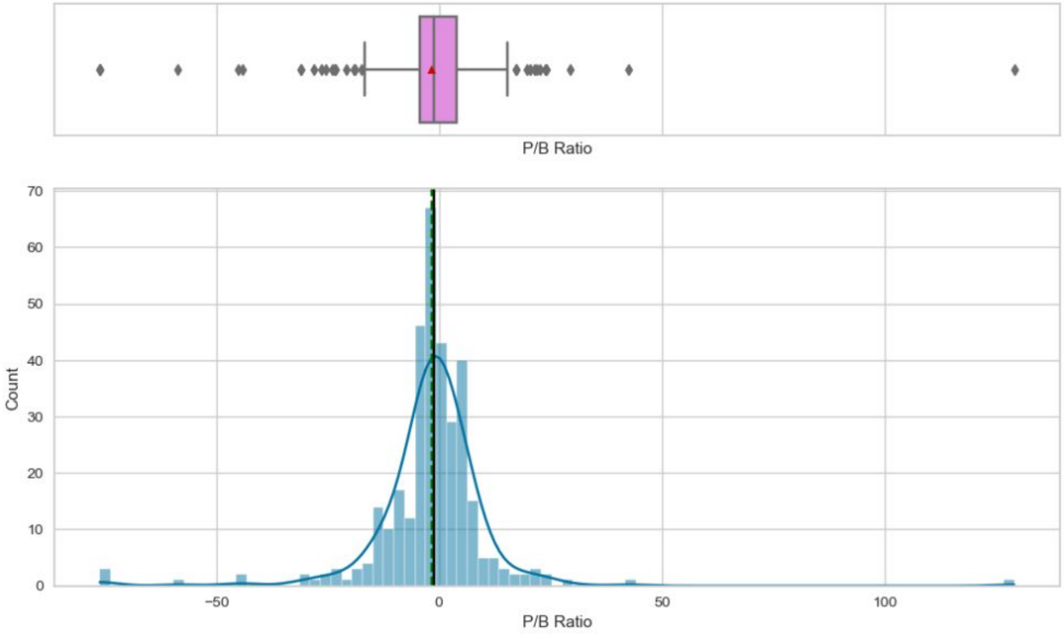
EDA Results: Univariate Analysis

- Analyzing the P/E Ratio attribute, we report that the average value is 32.61 with a standard deviation of 44.34. A boxplot and a countplot are presented to depict the distribution more clearly.



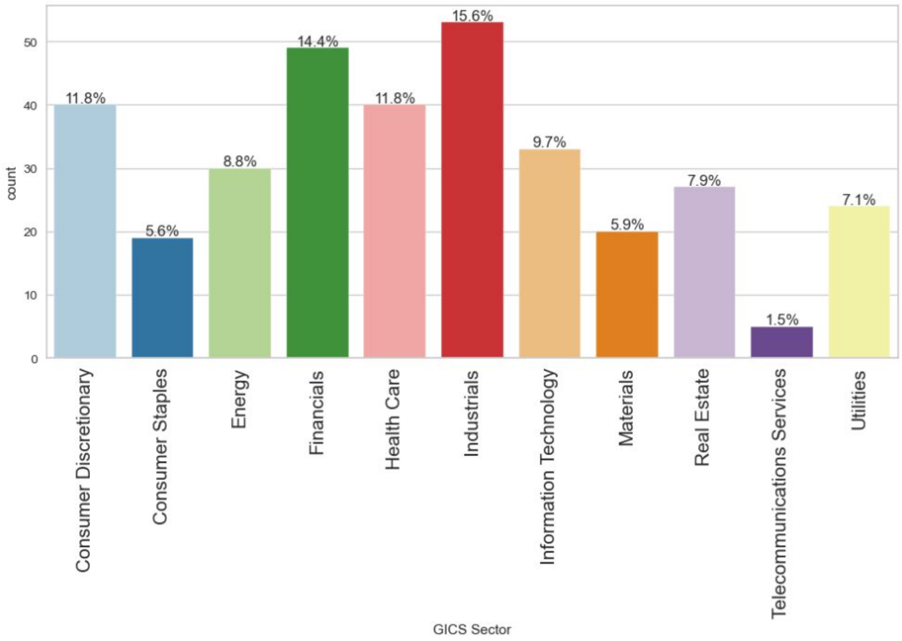
EDA Results: Univariate Analysis

- Analyzing the P/B Ratio attribute, we report that the average value is -1.71 with a standard deviation of 13.96. A boxplot and a countplot are presented to depict the distribution more clearly.



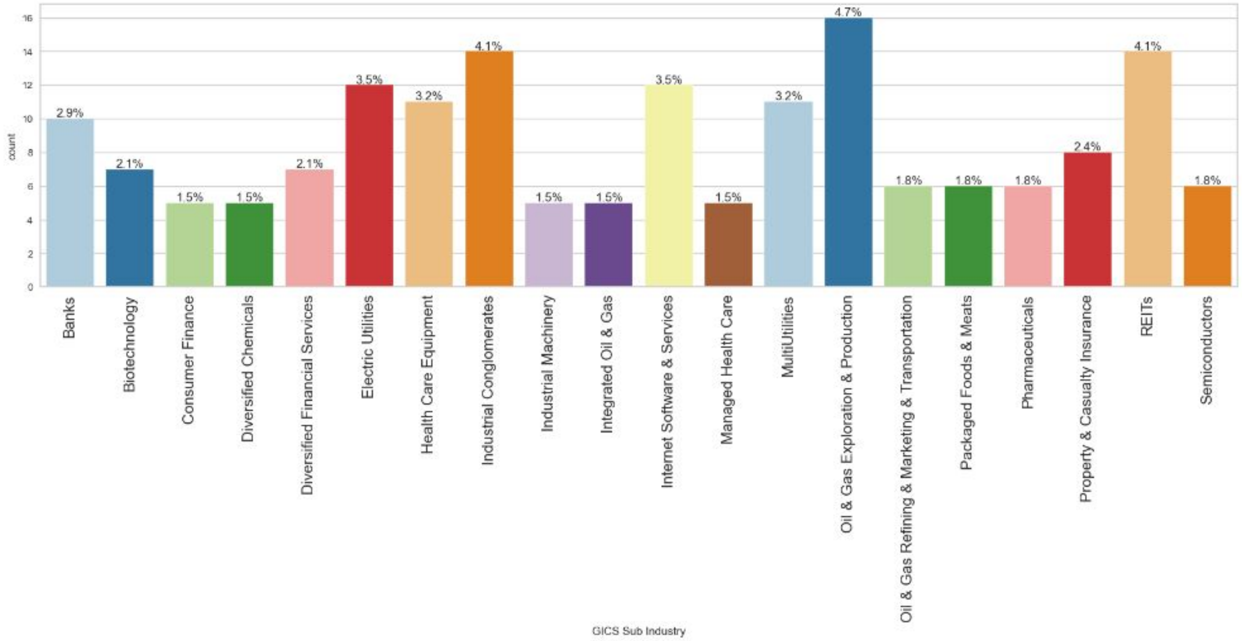
EDA Results: Univariate Analysis

- Analyzing the GICS sector attribute, we report that there is sufficient diversification between different sectors. A countplot is presented to depict the distribution more clearly.



EDA Results: Univariate Analysis

- Analyzing the GICS Sub Industry attribute, we report that there is sufficient diversification between different industries. A countplot (top 20 of 104) is presented to depict the distribution more clearly.

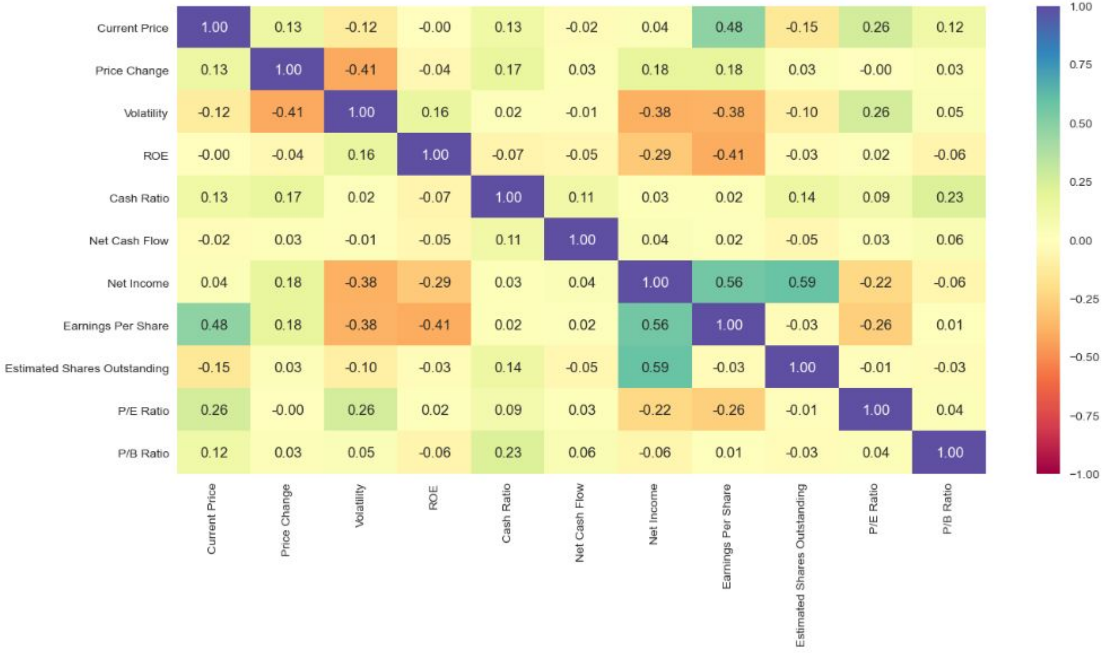


EDA Results: Multivariate Analysis

- In this section, we perform multivariate analysis on the data.
    - For selected attributes, we perform a descriptive statistical analysis where each attribute is jointly examined along with other attributes in order to evaluate their relationship and correlation.
    - We also analyze the corresponding data and present the joint distribution of the attributes, with confidence intervals to signify variance and statistical significance.
    - Finally, we write the conclusion based on both quantitative and qualitative observations.

EDA Results: Multivariate Analysis

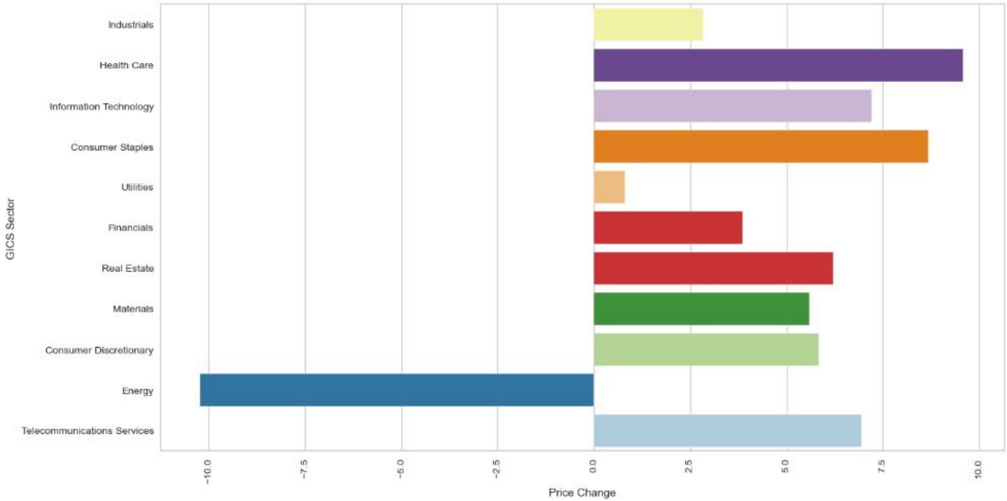
- Analyzing the correlation matrix, we can observe that negative volatility combined with positive cash ratio, net income and earnings per share correlates with the desired positive price change.



	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
Current Price	1.00	0.13	-0.12	-0.00	0.13	-0.02	0.04	0.48	-0.15	0.26	0.12
Price Change	0.13	1.00	-0.41	-0.04	0.17	0.03	0.18	0.18	0.03	-0.00	0.03
Volatility	-0.12	-0.41	1.00	0.16	0.02	-0.01	-0.38	-0.38	-0.10	0.26	0.05
ROE	-0.00	-0.04	0.16	1.00	-0.07	-0.05	-0.29	-0.41	-0.03	0.02	-0.06
Cash Ratio	0.13	0.17	0.02	-0.07	1.00	0.11	0.03	0.02	0.14	0.09	0.23
Net Cash Flow	-0.02	0.03	-0.01	-0.05	0.11	1.00	0.04	0.02	-0.05	0.03	0.06
Net Income	0.04	0.18	-0.38	-0.29	0.03	0.04	1.00	0.56	0.59	-0.22	-0.06
Earnings Per Share	0.48	0.18	-0.38	-0.41	0.02	0.02	0.56	1.00	-0.03	-0.26	0.01
Estimated Shares Outstanding	-0.15	0.03	-0.10	-0.03	0.14	-0.05	0.59	-0.03	1.00	-0.01	-0.03
P/E Ratio	0.26	-0.00	0.26	0.02	0.09	0.03	-0.22	-0.26	-0.01	1.00	0.04
P/B Ratio	0.12	0.03	0.05	-0.06	0.23	0.06	-0.06	0.01	-0.03	0.04	1.00

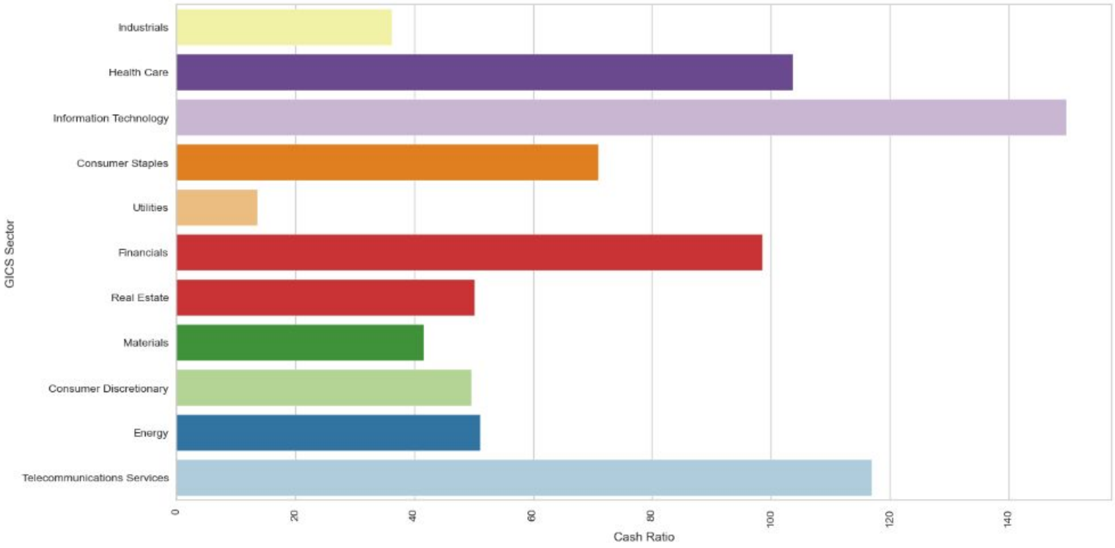
EDA Results: Multivariate Analysis

- Analyzing the relationship between the Price Change and GICS Sector attributes, we observe that the sectors with the highest price change increase are Health Care (~10), Consumers Staples (~8.5), Information Technology (~7.5) and Telecommunication Services (~7). The Energy sector is the only one with negative price change(-10). The remaining sectors had a moderate price change increase between 1 and 6.



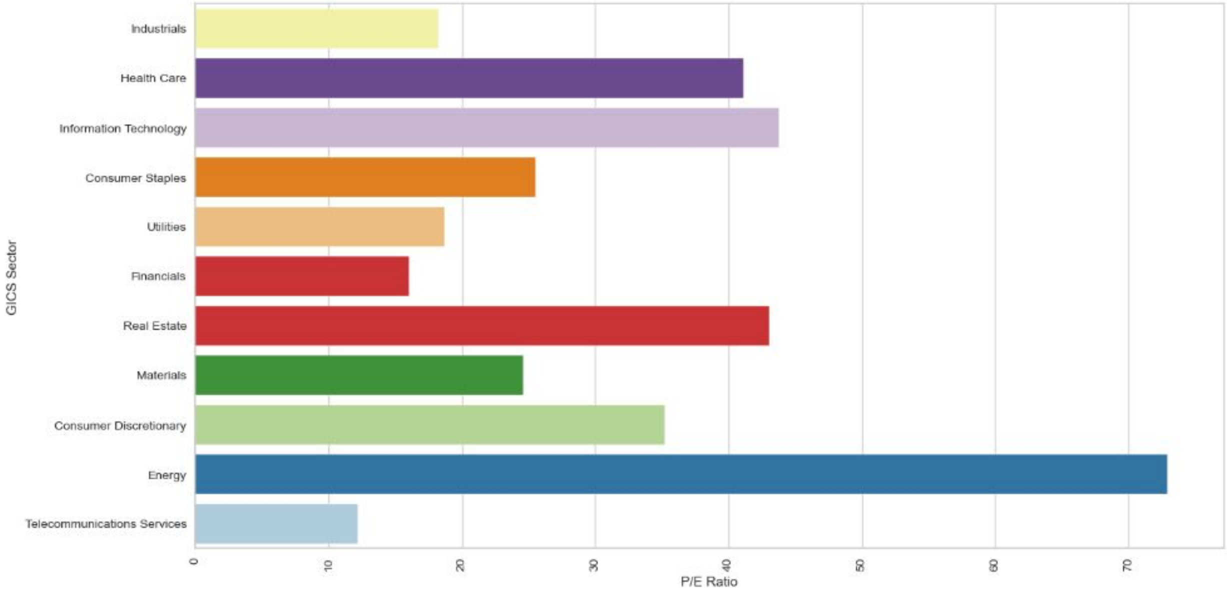
EDA Results: Multivariate Analysis

- Analyzing the relationship between the Cash Ratio and GICS Sector attributes, we observe that the Information Technology sector has the highest cash ratio (~150), followed by Telecommunication Services (~120), Healthcare (~100) and Financials (~100). The remaining sectors have significantly lower cash ratios.



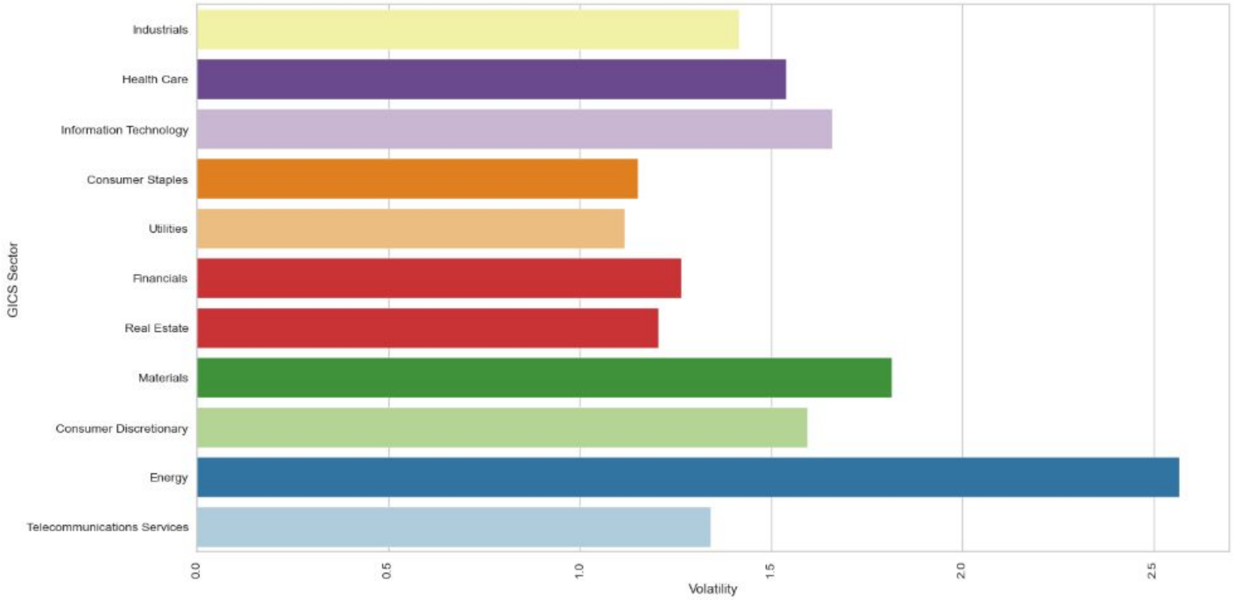
EDA Results: Multivariate Analysis

- Analyzing the relationship between the P/E Ratio and GICS Sector attributes, we observe that the Energy sector displays the highest P/E ratio (~70), followed by Information Technology, Real Estate and Health Care (~40). The remaining sectors have significantly lower P/E ratios.



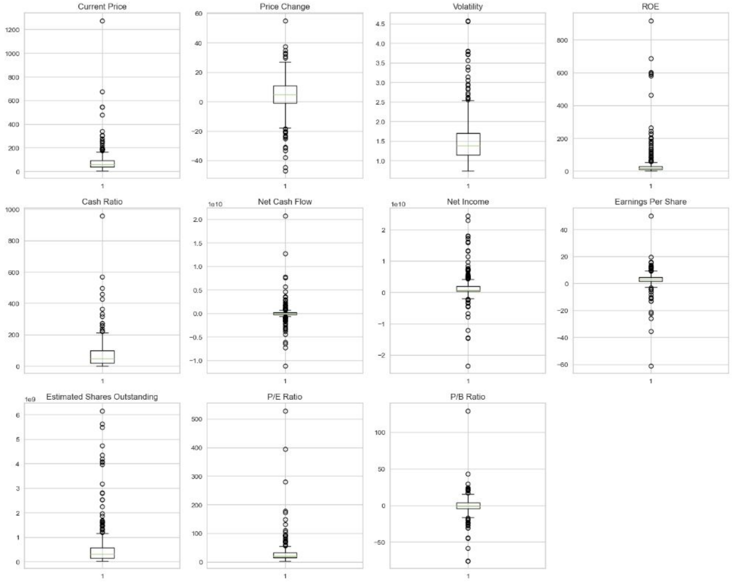
EDA Results: Multivariate Analysis

- Analyzing the relationship between the Volatility and GICS Sector attributes, we observe that the Energy sector displays a high volatility of 2.5, while the remaining sectors exhibit significantly lower volatility approximately between 1.0 and 1.5.



Data Preprocessing

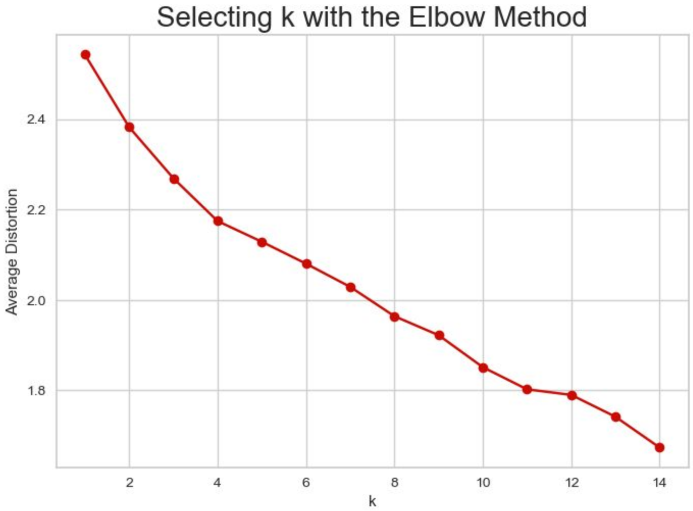
- Outlier Check: There is a significant number of outliers in the financial data of Trade Ahead.

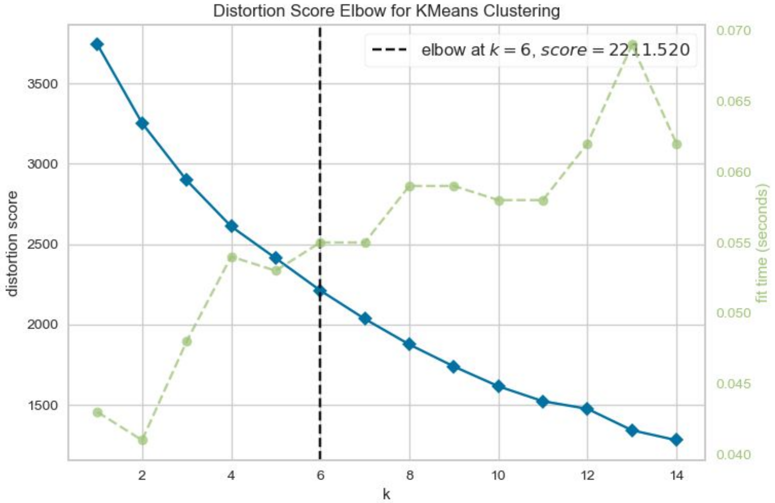


- Solution: We use StandardScaler to scale the data and minimize the negative effect outliers can have on the distribution attributes (mean, standard deviation) of each feature.

K-Means Clustering

- Using the Elbow Method, we select the best number of clusters (k) for K-Means Clustering.

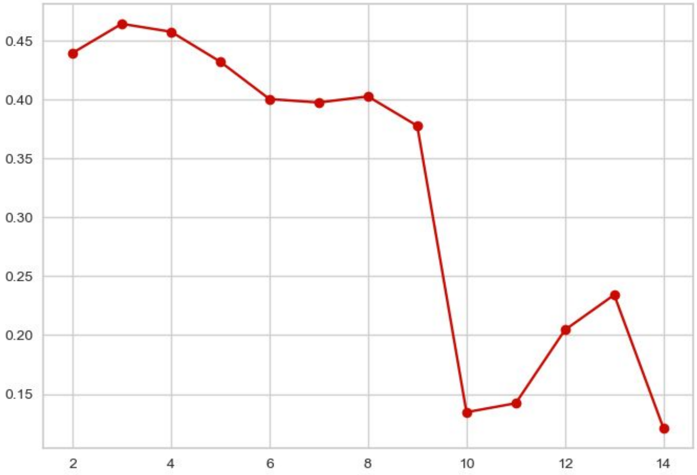


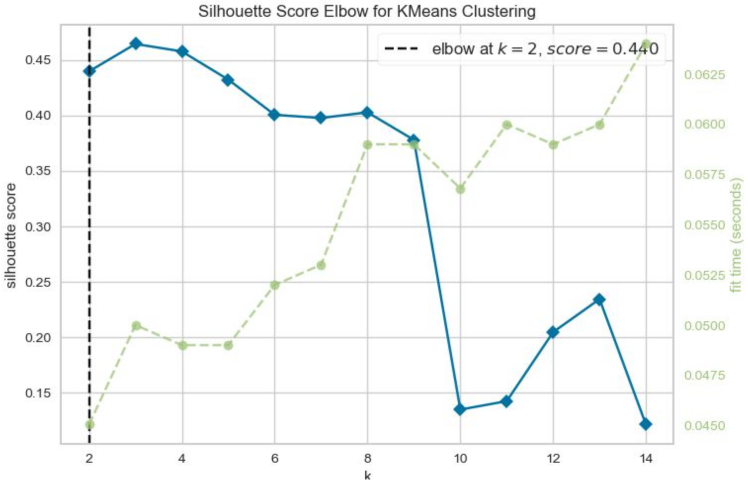


- From the Elbow Method, we select k=4 to be the number of clusters.

K-Means Clustering

- We also calculate the silhouette scores to verify the validity of our selection.

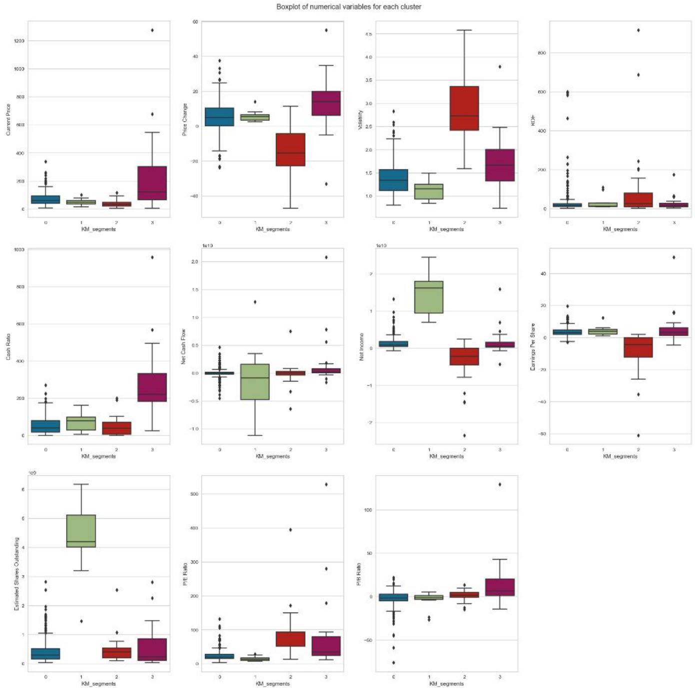




- We observe that the highest silhouette score is achieved for k=3, however we will keep our selection of k=4 which resulted from the Elbow Method, because both the silhouette scores and the running times are approximately similar, while selecting k=4 can lead to greater diversification.

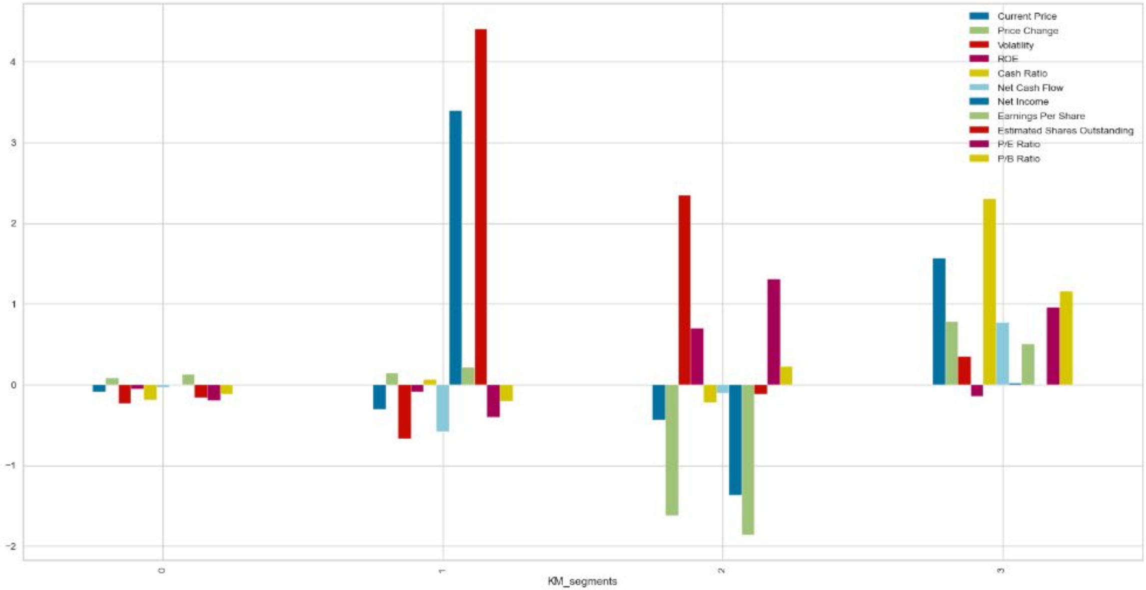
K-Means Clustering

- The result of the K-Means final model (k=4), for each cluster and attribute combination, is as it is depicted in the following boxplot:



K-Means Clustering

- The result of the K-Means final model (k=4), for each cluster and scaled attribute combination, is as it is depicted in the following barplot:



K-Means Clustering

- The first cluster (8 companies) is the "unicorn cluster" where companies whose stocks have a very high current price are located. These stocks are overvalued (high P/E ratio and high P/B ratio), however they are reliable as they exhibit low volatility combined with a positive net cash flow, a high cash ratio and high earnings per share. Finally, they have a positive and the highest price change, indicating their continuous rise.

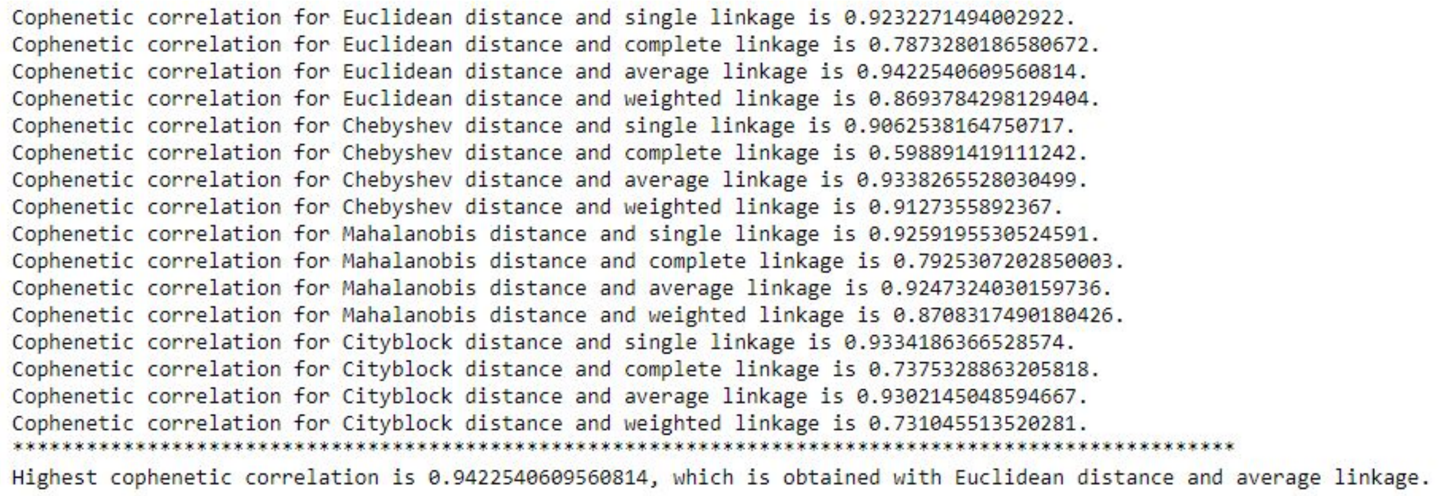
- The second cluster (290 companies) is the "normal cluster" where companies whose stocks have a low current price are located. These stocks have a growth potential as they are currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit low volatility combined with medium earnings per share and positive price change. However, while the price change is positive, the remaining metrics are of low magnitude.

- The third cluster (14 companies) is the "opportunity cluster" where companies whose stocks have a very low current price and very high net income are located. These stocks also have a growth potential as they are also currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit very low volatility combined with a high net cash flow, high cash ratio and medium earnings per share. Finally, they have a positive and high price change and the highest estimated shares outstanding, indicating their continuous rise.

- The fourth cluster (28 companies) is the "junk cluster" where companies whose stocks have a low current price but a very high volatility and ROE are located. These stocks are currently overvalued (high P/E ratio and high P/B ratio), and the remaining metrics also indicate bad performance with significantly negative net income, earnings per share and price change.

Hierarchical (Agglomerative) Clustering

- Calculating the Cophenetic Correlation scores, we examine all combinations of distance metrics and linkage methods.

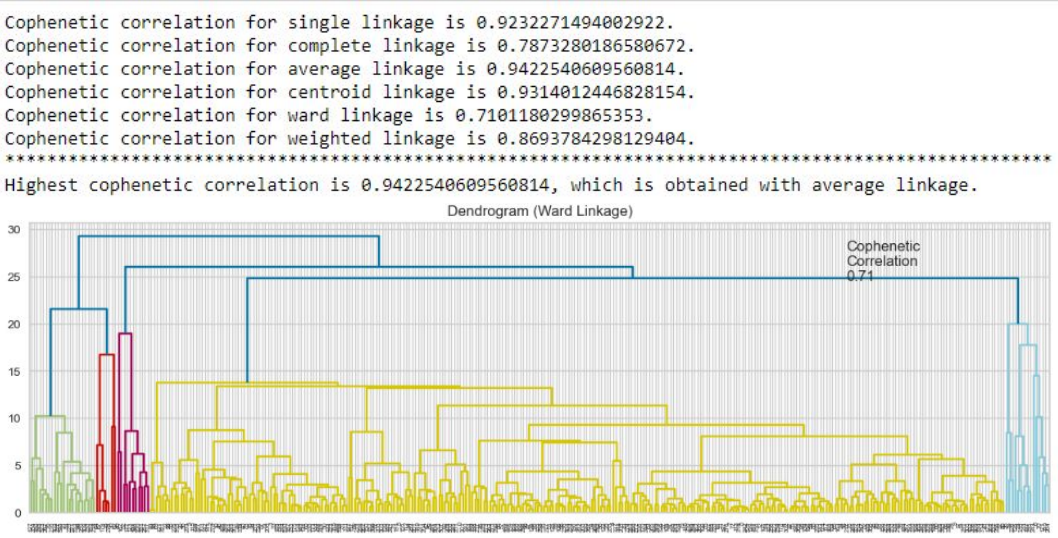


```
Cophenetic correlation for Euclidean distance and single linkage is 0.9232271494002922.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7873280186580672.
Cophenetic correlation for Euclidean distance and average linkage is 0.9422540609560814.
Cophenetic correlation for Euclidean 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 0.9259195530524591.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.7925307202850003.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159736.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8708317490180426.
Cophenetic correlation for Cityblock distance and single linkage is 0.9334186366528574.
Cophenetic correlation for Cityblock distance and complete linkage is 0.7375328863205818.
Cophenetic correlation for Cityblock distance and average linkage is 0.9302145048594667.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.731045513520281.
****************************************************************************************************
Highest cophenetic correlation is 0.9422540609560814, which is obtained with Euclidean distance and average linkage.
```

- Thsu, based on the resulting scores, we select the Euclidean distance as our distance metric.

Hierarchical (Agglomerative) Clustering

- We also calculate the Cophenetic Correlation scores for all possible linkage methods.

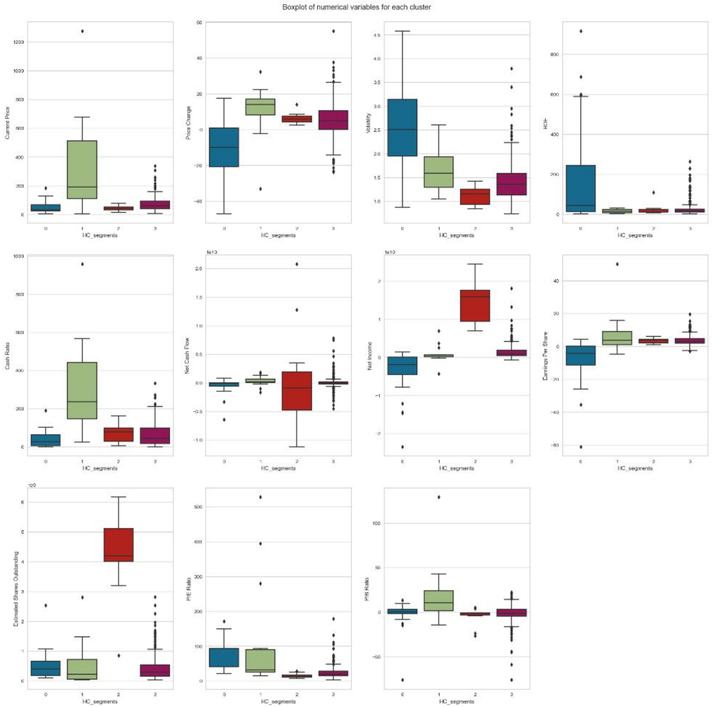


```
Cophenetic correlation for single linkage is 0.9232271494002922.
Cophenetic correlation for complete linkage is 0.7873280186580672.
Cophenetic correlation for average linkage is 0.9422540609560814.
Cophenetic correlation for centroid linkage is 0.9314012446828154.
Cophenetic correlation for ward linkage is 0.7101180299865353.
Cophenetic correlation for weighted linkage is 0.8693784298129404.
*****************************************************************************************
Highest cophenetic correlation is 0.9422540609560814, which is obtained with average linkage.
```

- Although the average linkage provides the highest Cophenetic Correlation score, from the dendrograms of each linkage method, we find the ward linkage method to be the most suitable as it provides the greatest diversification between the four clusters. Thus, we select ward linkage as our linkage method.

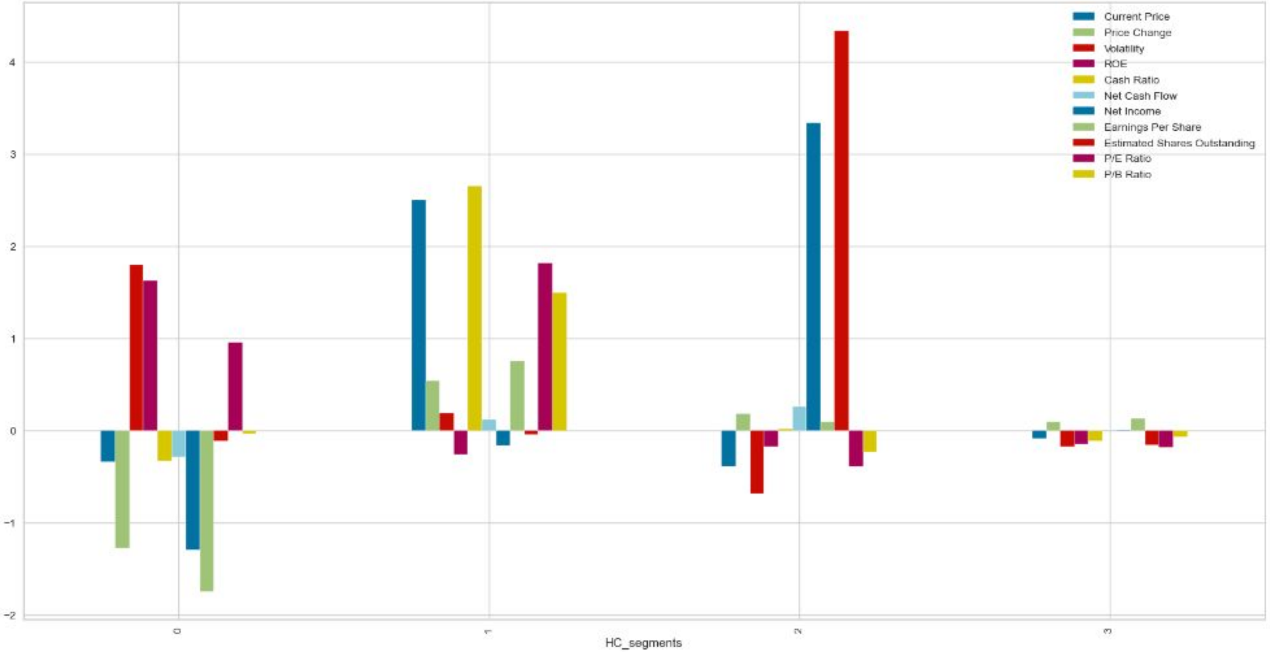
Hierarchical (Agglomerative) Clustering

- The result of the Hierarchical (Agglomerative) Clustering final model (k=4, affinity='euclidean', linkage='ward'), for each cluster and attribute combination, is as it is depicted in the following boxplot:



Hierarchical (Agglomerative) Clustering

- The result of the Hierarchical (Agglomerative) Clustering final model (k=4, affinity='euclidean', linkage='ward'), for each cluster and scaled attribute combination, is as it is depicted in the following barplot:



Hierarchical (Agglomerative) Clustering

- The first cluster (29 companies) is the "junk cluster" where companies whose stocks have a low current price but a very high volatility and ROE are located. These stocks are currently overvalued (high P/E ratio), and the remaining metrics also indicate bad performance with significantly negative net income, earnings per share and price change. Moreover, it displays moderately negative net cash flow and cash ratio.

- The second cluster (15 companies) is the "unicorn cluster" where companies whose stocks have a high current price are located. These stocks are overvalued (high P/E ratio and high P/B ratio), however they are reliable as they exhibit low volatility combined with a positive and very high net cash flow and high earnings per share. Finally, they have a positive and the highest price change, indicating their continuous rise.

- The third cluster (11 companies) is the "opportunity cluster" where companies whose stocks have a very low current price and very high net income are located. These stocks also have a growth potential as they are also currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit very low volatility combined with a high net cash flow and medium earnings per share. Finally, they have a positive and high price change and the highest estimated shares outstanding, indicating their continuous rise.

- The fourth cluster (285 companies) is the "normal cluster" where companies whose stocks have a low current price are located. These stocks have a growth potential as they are currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit low volatility combined with medium earnings per share and positive price change. However, while the price change is positive, the remaining metrics are of low magnitude.

Executive Summary

- The four investment clusters, resulted from the both clustering methods on the Trade Ahead financial data of 340 companies, are the following:
  - **Unicorn Cluster:** Stocks have a high current price. They are overvalued (high P/E ratio and high P/B ratio), however they are reliable as they exhibit low volatility combined with a positive and very high net cash flow and high earnings per share. Finally, they have the highest price change, indicating their continuous rise.
  - **Opportunity Cluster:** Stocks have a very low current price and very high net income. They also have a growth potential as they are also currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit very low volatility combined with a high net cash flow and medium earnings per share. Finally, they have a positive and high price change and the highest estimated shares outstanding, indicating their rise.
  - **Normal Cluster:** Stocks have a low current price. They have a growth potential as they are currently undervalued (low P/E ratio and low P/B ratio), and they are reliable as they exhibit low volatility combined with medium earnings per share and positive price change. However, while the price change is positive, the remaining metrics are of low magnitude, indicating safety but only a low or moderate rise.
  - **Junk Cluster:** Stocks have a low current price but a very high volatility and ROE. They are currently overvalued (high P/E ratio), and the remaining metrics also indicate bad performance with significantly negative net income, earnings per share and price change. Plus, moderately negative net cash flow and cash ratio.

Executive Summary

- The two clustering methods on the Trade Ahead financial data of 340 companies, are the following:
    - **K-Means Clustering Model:** We chose 4 clusters, as indicated by the elbow method. The silhouette score for k=4 was the second highest but very close to the highest one (k=3). Thus, we chose 4 clusters for our K-Means final model.
    - **Hierarchical Clustering (Agglomerative) Model:** After evaluating the Cophenetic correlation scores, we chose the Euclidean distance as the distance component. Afterwards, for choosing the linkage component, we evaluated the the Cophenetic correlation scores again using the Euclidean distance. Although average linkage displayed the highest score, after a more careful examination of the dendrograms, we decided to select the ward linkage as our component, because it provided a more distinct clustering result. From the dendrogram, we also selected the number of clusters to be 4 (k=4) for our Hierarchical Clustering final model.
    - **Conclusion:** Both clustering methods provide the same number of distinct clusters (k=4), with approximately the same attributes for each cluster. If we wish to optimize for performance (greater diversification), we choose the **Hierarchical Clustering (Agglomerative) Model**, while if we want to optimize for speed, we choose the **K-Means Clustering Model** which is slightly faster.
    - **Recommendation:** In both cases, we recommend a portfolio containing a mix of stocks from the *Unicorn* and *Opportunity* clusters, plus selected individual stocks from *Normal* cluster and none from the *Junk* clusters.

Thank you for your time!

Michail Mersinias

06/06/2023