TRADE AND AHEAD

UNSUPERVISED LEARNING

01/29/2023

Contents / Agenda

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

EXECUTIVE SUMMARY

- Cluster analysis can help identify stocks exhibiting similar characteristics and ones exhibiting minimum correlation, thereby helping investors diversify and invest in stocks across different market segments, protecting against risks that could make the portfolio vulnerable to losses
- Both K-means clustering, and Hierarchical Clustering methods were performed on the datasets
- Out of 340 securities in our data set, both clustering methods clustered 270+ securities in a similar fashion with other securities being clustered differently. The industry segregation into clusters yielded similar results across both algorithms.
- Both the K-Means model and the Agglomerative Clustering model fit the dataset within ~0.1s
- Both algorithms yielded similar clusters based on the outliers within the 11 variables
- A major cluster (270+securities) was identified as mildly aggressive & safe investment option. This cluster is diversified with securities predominantly belonging to Industrials, followed by Financials, Consumer Discretionary, Real Estate, & Informational Technology sectors

- One cluster (25+securities) was high performing belonging predominantly to Health Care followed by Consumer Discretionary and Information Technology sectorsaaaaaaaaa

- Another cluster (25+securities) was identified as historically low performing belonging predominantly to Energy sector

- Another cluster (~10securities) was identified as moderate belonging to Health Care, Consumer Discretionary

- Another cluster (~10securities) was identified as moderately aggressive (& high performing) belonging predominantly to Financials sector

- Finally, a major cluster (270+securities) was identified as safe investment option. This cluster is diversified with securities predominantly belonging to Industrials, followed by Financials, Consumer Discretionary, Real Estate, & Informational Technology sectors

Business Problem Overview and Solution Approach

Trade & Ahead is a financial consultancy firm who provide their customers with personalized investment strategies as a Data Scientist and provided you with data comprising stock price and some financial indicators for a few companies listed under the New York Stock Exchange.

Business Problem

The tasks of analyzing the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

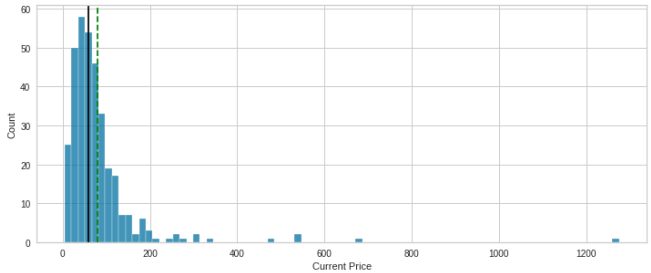
Solution Approach

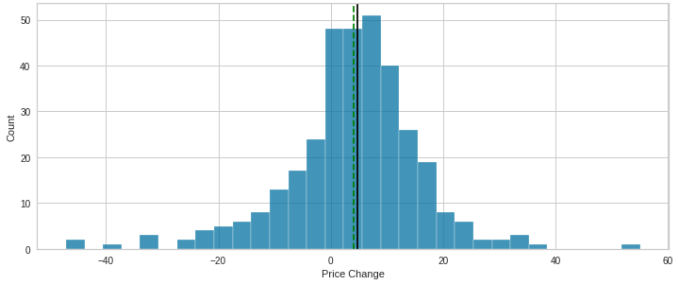
By using a cluster analysis, we can identify stocks that exhibit similar characteristics and ones which exhibit minimum correlation. This will help investors better analyze stocks across different market segments and help protect against risks that could make the portfolio vulnerable to losses.

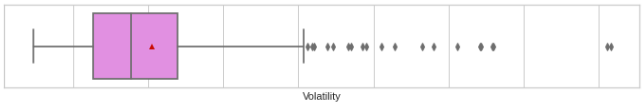
DATA OVERVIEW

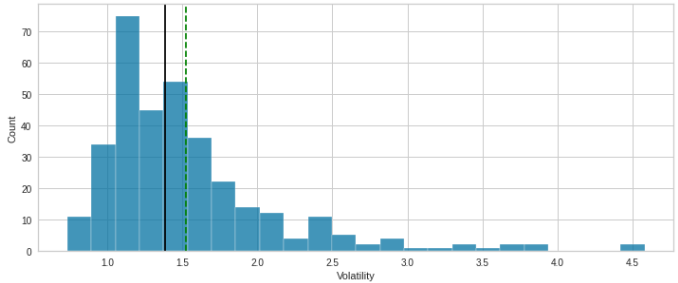
The dataset consist of 340 rows and 15 columns.
Ticker Symbol: An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
Company: Name of the company
GICS Sector: The specific economic sector assigned to a company by the Global Industry Classification 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)

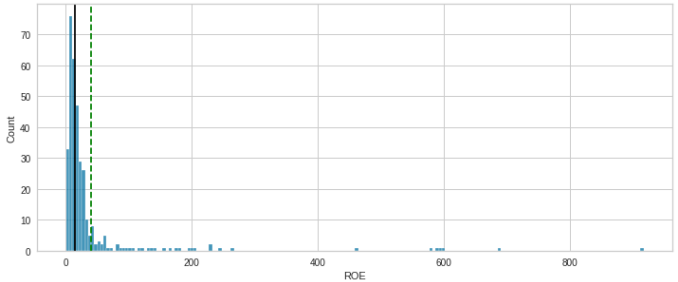
EDA Results



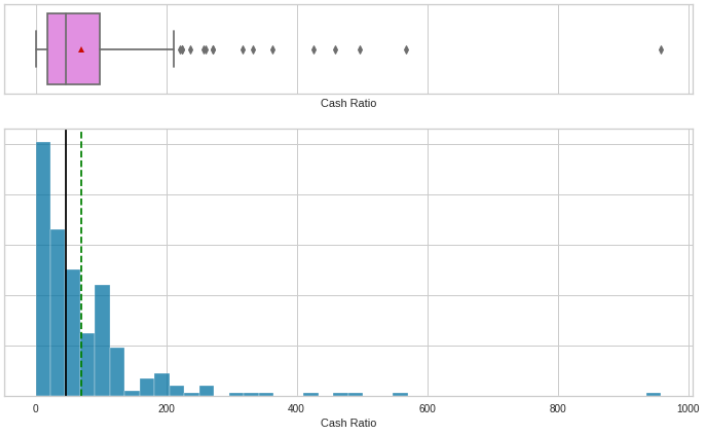


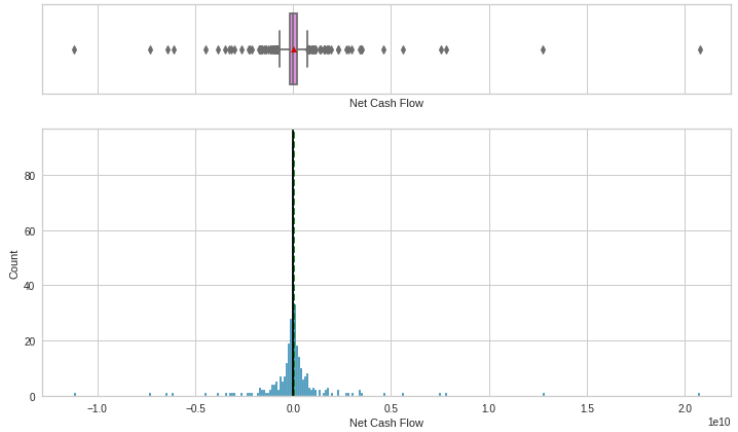


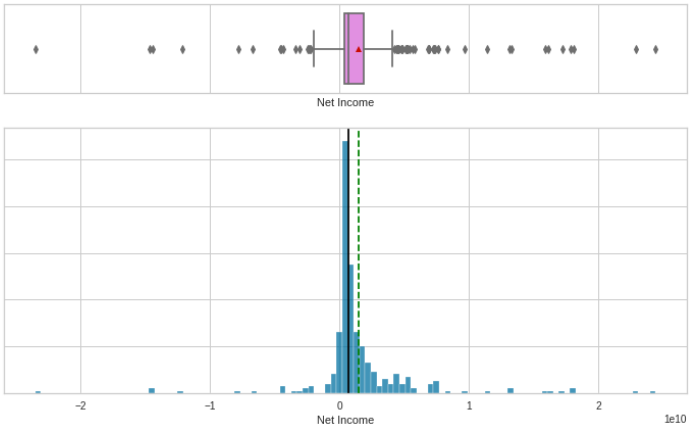


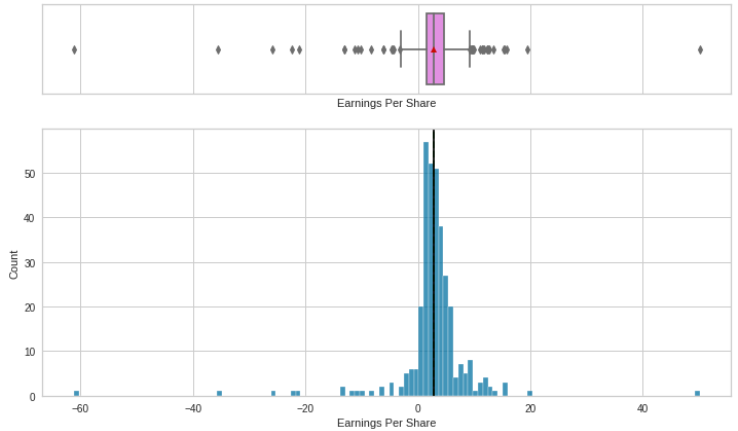


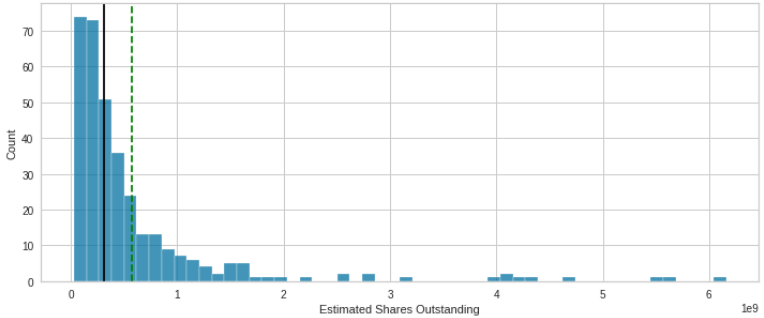
[ata background check]

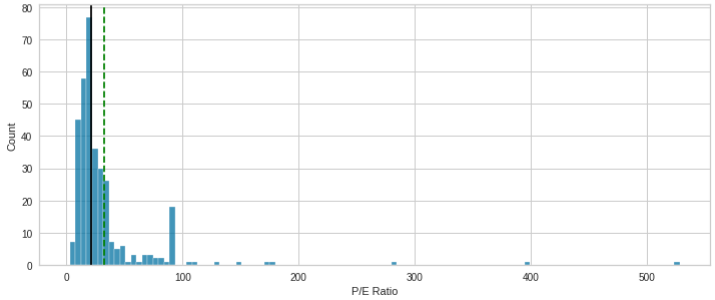


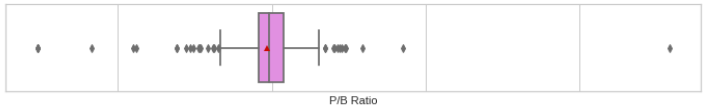


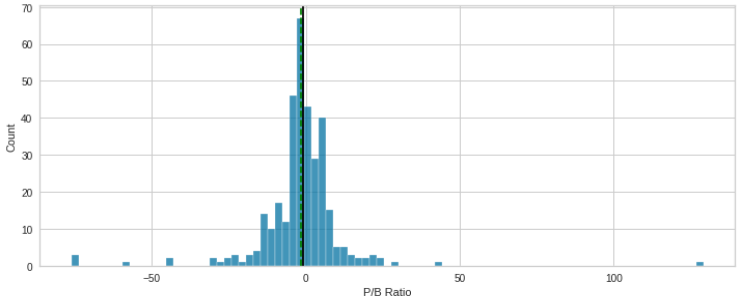


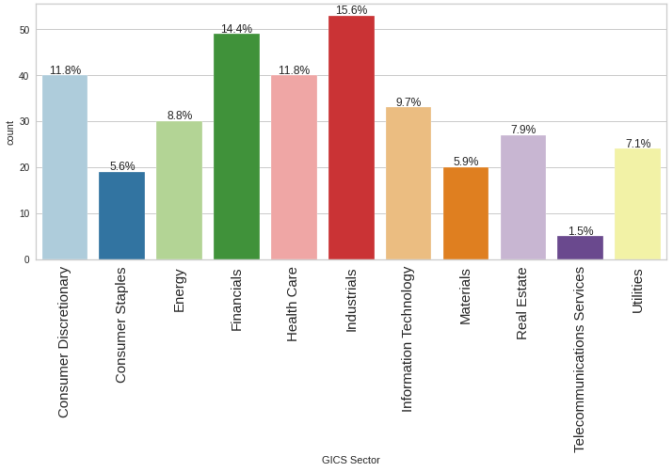










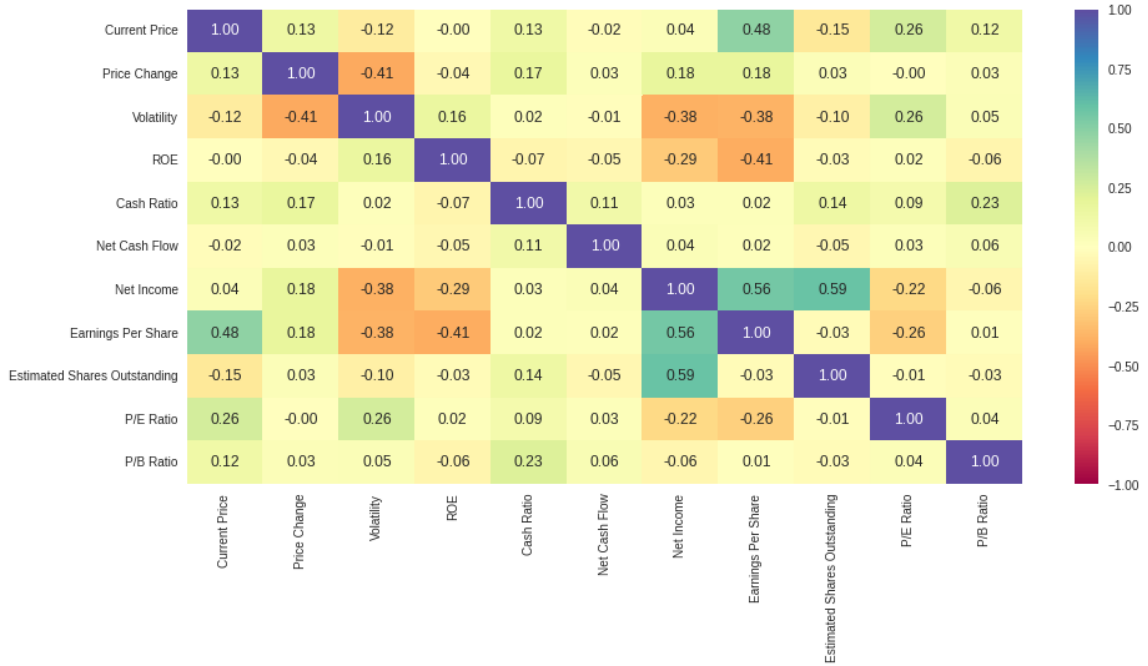


UNIVARIATE ANALYSIS INSIGHTS

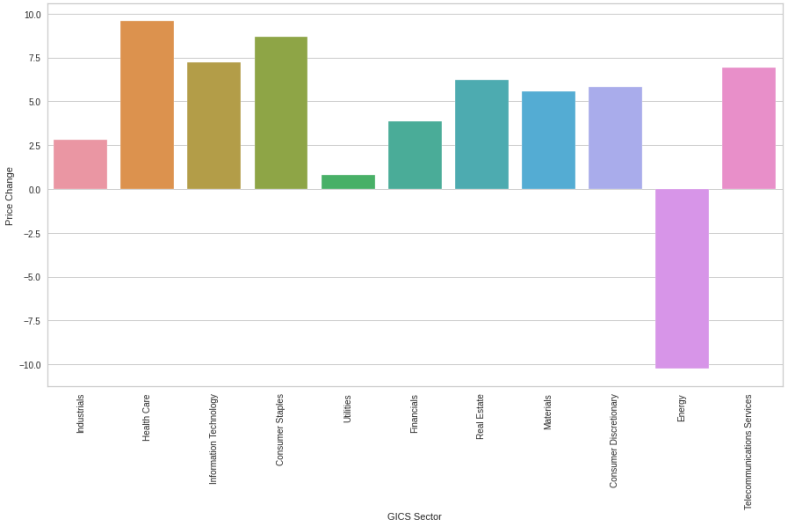
- **Current_price** is right skewed with positive outliers
- **Price_change** has normal distribution with +ve and –ve outliers
- **Volatility** is right skewed with some +ve outliers
- **ROE** is right skewed with some +ve outliers
- **Cah_ratio** is right skewed with +ve outliers
- **Net_Cash_Flow** has a normal distribution with +ve and –ve outliers
- **Net_Income** has a normal distribution with some +ve and a few –ve outliers
- **Earnings_Per_Share** has a normal distribution with some +ve and –ve outliers
- **Estimated_Shares_Outstanding** is right skewed with several +ve outliers
- **P/E_Ratio** is right skewed with some +ve outliers
- **P/B_Ratio** has a normal distribution with a few +ve and –ve outliers

- Current_Price of stocks, and Estimated_Shares_Outstanding across securities for all sectors is right skewed (with several positive outliers)
- Health Care and Financial sectors have seen some of the highest positive Price_Change in the last 13 weeks, making them favorable to investors
- Informational Technology and Financial sectors have some of the highest Cash_Ratios making them favorable more so than other sectors
- Real Estate sector has seen minimum variation in Price_Change & minimum variation in Cash_Ratio across securities it encompasses making them a safer investment choice for investors
- Energy sector has some of the highest variance in Price_Change across securities it encompasses, being more volatile and riskier for investors. However, this sector has securities with high P/E_Ratios.

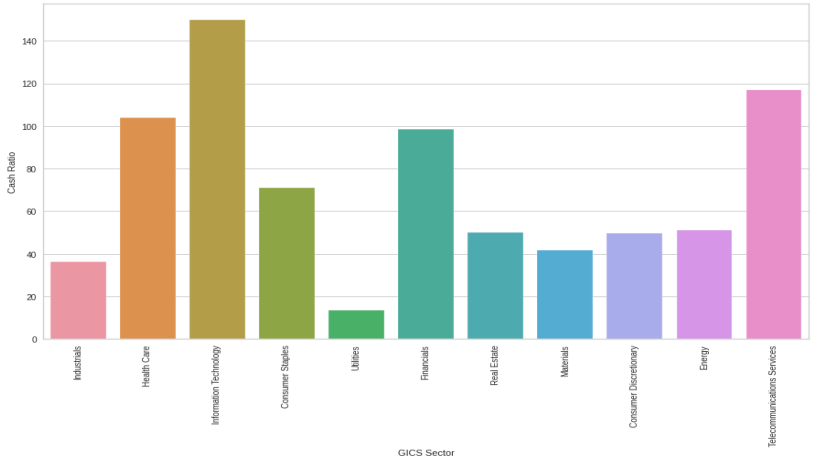
BI-VARIATE ANALYSIS



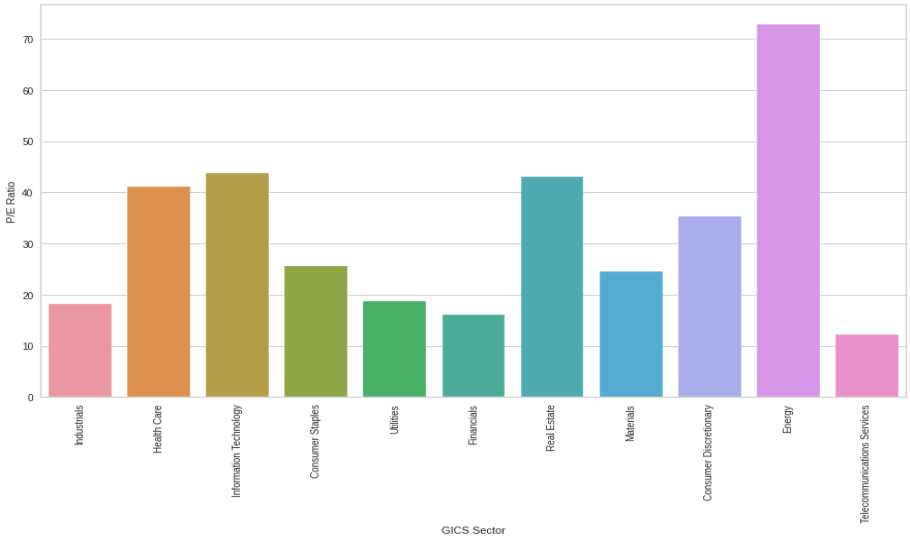
- Price_Change has a negative correlation with Volatility
- Earnings_Per_Share has a positive correlation with Current_Price & Net_Income
- Estimated_Shares_Outstanding has a positive correlation with Net_Income
- Earnings_Per_Share has a negative correlation with ROE and Volatility

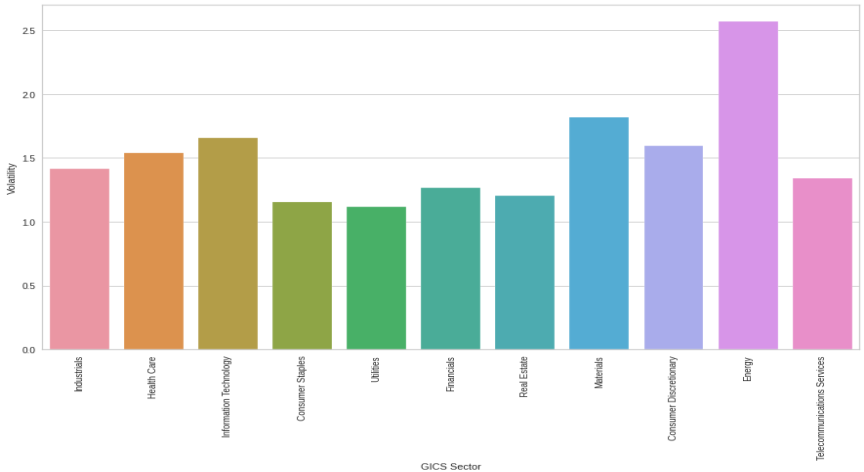


- Real_Estate has seen the minimum variation in Price_Change across different securities it encompasses while Energy GICS_Sector has seen the maximum variation in Price_Change across its securities
- Healthcare and Information Technology have maximum number of securities with a high positive Price_Change making them more favorable



•IT and Telecommunications sectors, both relatively newer and unregulated industries, are able to generate significantly higher average cash ratios than other sectors
•Utilities, a highly regulated industry, generates the lowest average cash ratios of all sectors

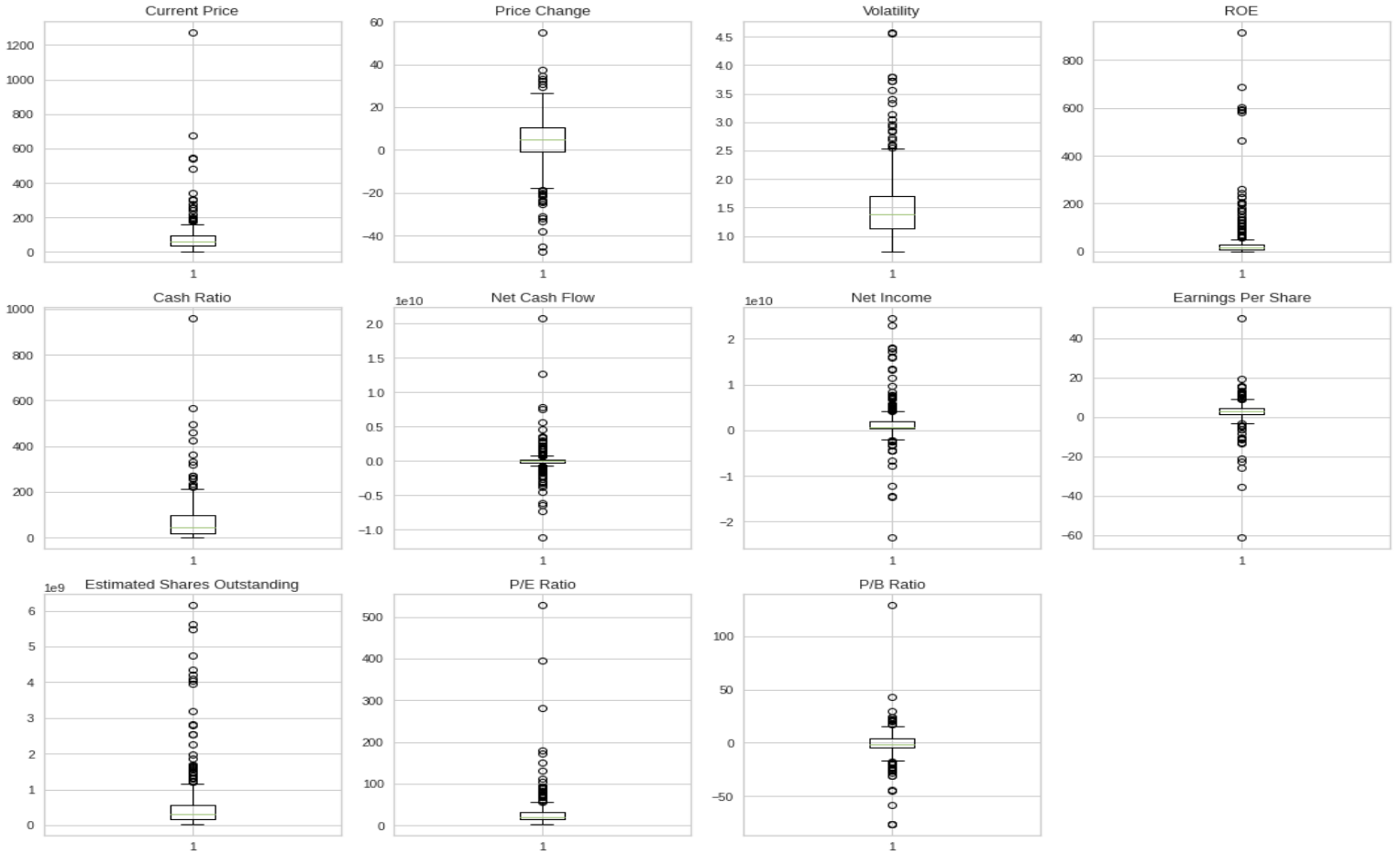


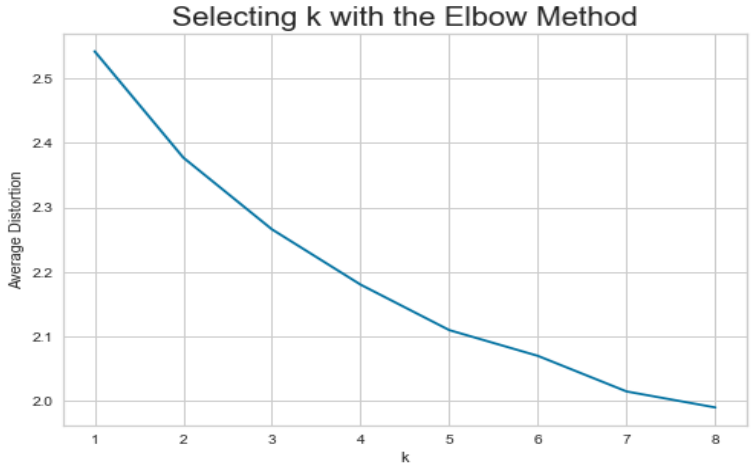


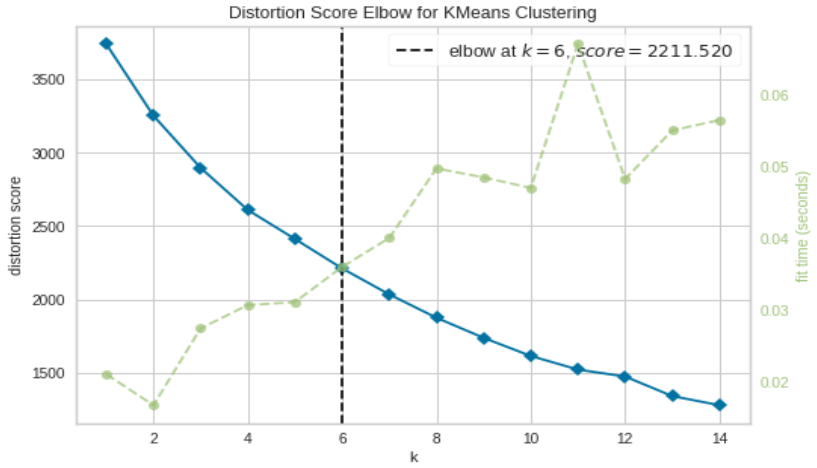
- Energy GICS_Sector has some of the highest variance in P/E ratios as well has some securities/companies with high P/E ratios.
- This indicates an investor is willing to invest more in a single share of a company in this sector per dollar of its earnings as opposed to securities/companies in other GICS_Sectors

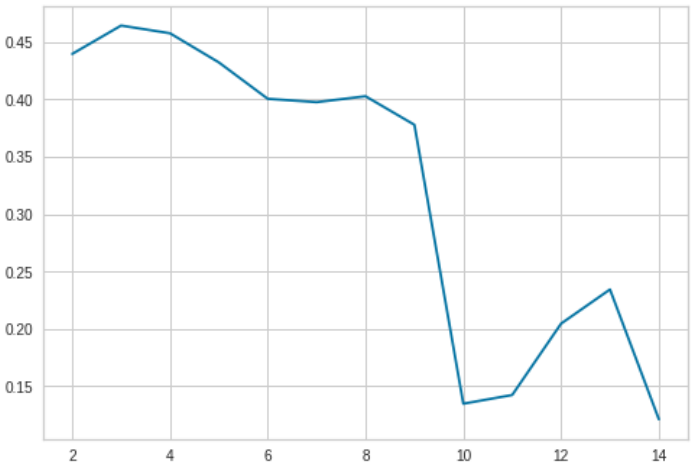
Data Preprocessing

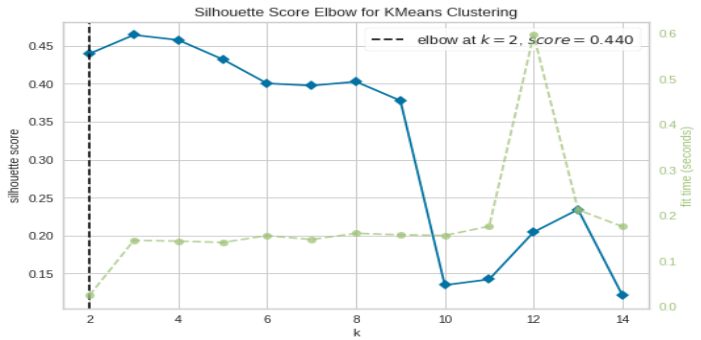
- After feature engineering (scalar transformation), the relationship between the attributes have been maintained.
- However, the attributes are now all on the same scale, with an average of 0, standard deviation of 1
- The data has no missing values, nor duplicate entries.
- Outliers have been identified, but not treated and they are assumed to be real data points & not anomalies in this context for modeling.

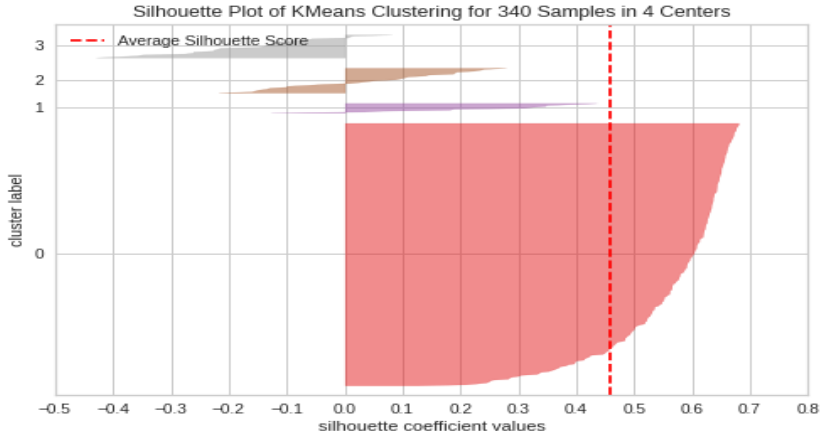












Selecting k=4 groups based on elbow method & silhouette score

K-Means Clustering

Cluster 0: : has 277 securities
Intermediate between clusters 1 & 2

- Cluster 1: has 11 securities has 10 times as high avg. Net_Income & Estimated_Shares_Outstanding

Cluster 2: has 27 securities low avg. Current_Price, negative avg.Price_Change, high volatility, low Cash_Ratio, low Net_Income, & low Earnings_Per_Share

Cluster 3 has 25 securities:
:high avg. Current_Price, high Cash_Ratio, high Earnings_Per_Share, high P/B_Ratio

KM_segments	current Price	Price change	Volatility	ROE	Cash Ratio	Net Cash flow	Net Income	Earning Per share	Est shares outstand	P/E	P/B	Count in each seg
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

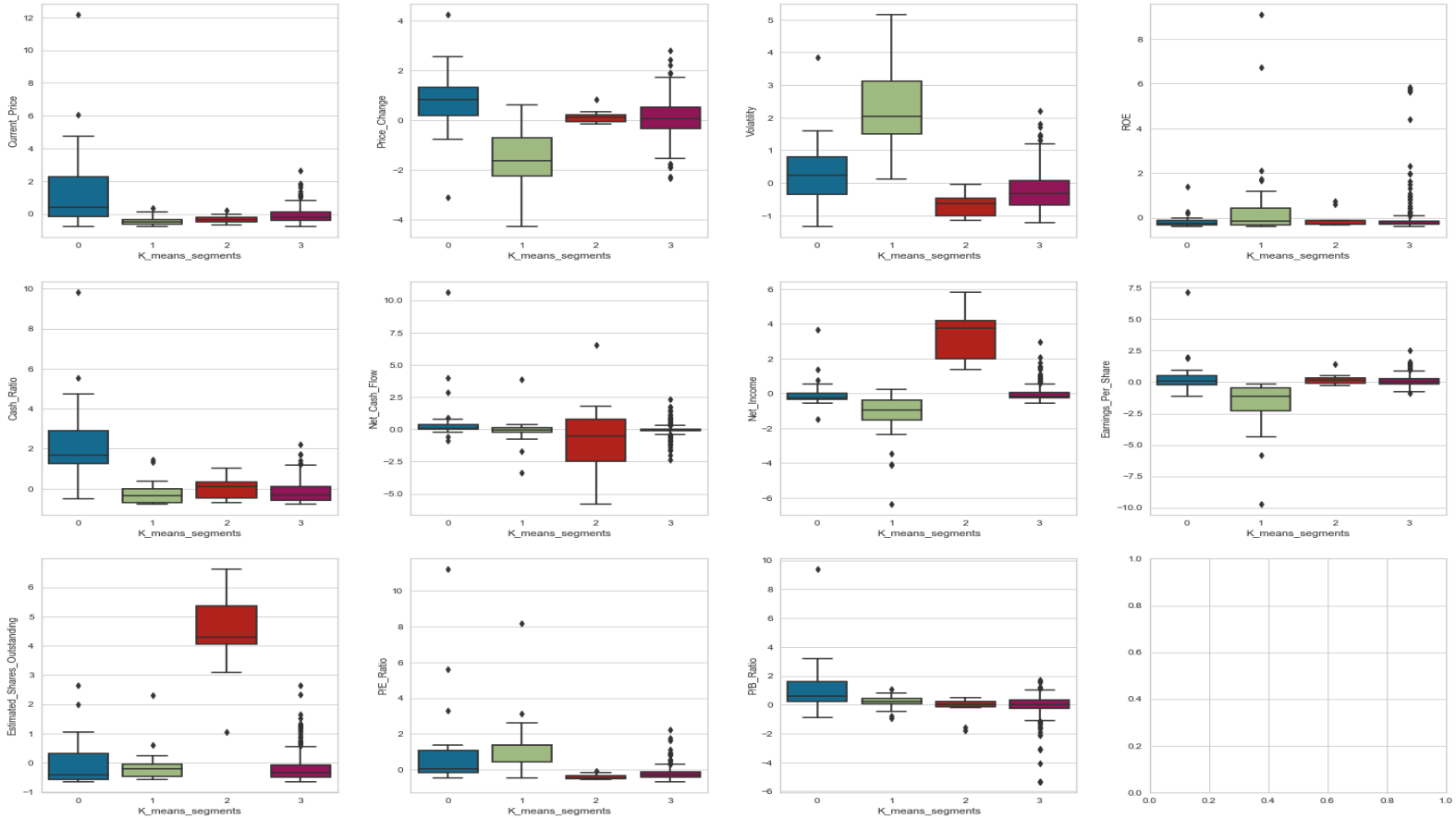
KM_segments	0	1	2	3
GICS Sector				
Consumer Discretionary	33	1	0	6
Consumer Staples	17	1	0	1
Energy	6	1	22	1
Financials	45	3	0	1
Health Care	29	2	0	9
Industrials	52	0	1	0
Information Technology	24	1	3	5
Materials	19	0	1	0
Real Estate	26	0	0	1
Telecommunications Services	2	2	0	1
Utilities	24	0	0	0

- Among the securities in Cluster 0, majority are Industrials, Financials, Consumer Discretionary ,Real Estate And IT
- Cluster 1 is dominated by Financial sector
- Cluster 2 is High securities in Energy
- Cluster 3 is high in Health Care sector
- Clusters 0 and 2 are the safe clusters, with clusters 2 containing more exclusive securities. Clusters 1 and 3 are more riskier securities, former being high performing and later historically speaking low performing

Great Learning

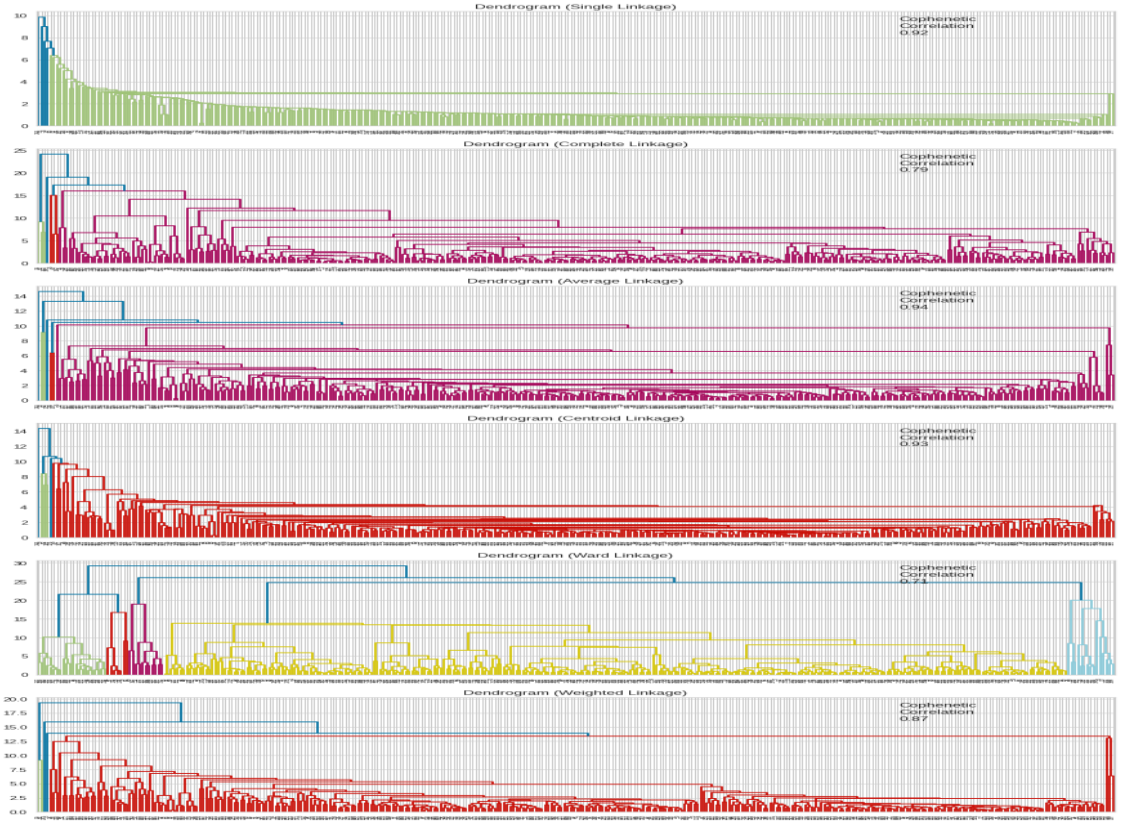
ER AHEAD

Boxplot of numerical variables for each cluster



HIERARCHIAL CLUSTERING

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



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

The dendrogram for Ward linkage appears to provide better clustering, with 4 appearing to be the appropriate number of clusters

CLUSTERING PROFILES IN HIERARCHIAL CLUSTERING

HC_segments	0	1	2	3
GICS Sector				
Consumer Discretionary	1	5	1	33
Consumer Staples	2	1	1	15
Energy	22	0	1	7
Financials	1	1	3	44
Health Care	0	8	1	31
Industrials	1	0	0	52
Information Technology	1	9	0	23
Materials	1	1	0	18
Real Estate	0	1	0	26
Telecommunications Services	0	1	2	2
Utilities	0	0	0	24

- Cluster 0:29 stocks,comprises maximum in Energy sector
- Cluster 1: it comprises of 27 stocks mostly in IT, HEALTH CARE
- Cluster 2:9 stocks, comprised mostly of stocks within the Financials and Telecommunications sectors
- Cluster 3:275 stocks (~84% of all stocks in the dataset) drawn from all sectors present in the dataset

CLUSTER PROFILING

HC_segments	current_price	Price_change	volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earning per Share	Est outstanding share	P/E	P/B	COUNT
0	48.006208	-11.263107	2.590247	196.551724	40.275862	-495901724.137931	-3597244655.172414	-8.689655	486319827.294483	75.110924	-2.162622	29
1	213.518640	15.252913	1.779861	22.333333	258.740741	1504052814.814815	1716529851.851852	5.177407	689838338.441482	78.441603	13.022590	27
2	46.672222	5.166566	1.079367	25.000000	58.333333	-3040666666.666667	14848444444.444445	3.435556	4564959946.222222	15.596051	-6.354193	9
3	72.421687	4.563230	1.403434	25.218182	55.014545	72801872.727273	1572467469.090909	3.728564	445003946.148764	24.188244	-2.966949	275

Companies within cluster 0 have:

a. The highest returns-on-equity
b. The lowest net incomes
c. Mostly negative earnings per share

Companies within this cluster 1 have:

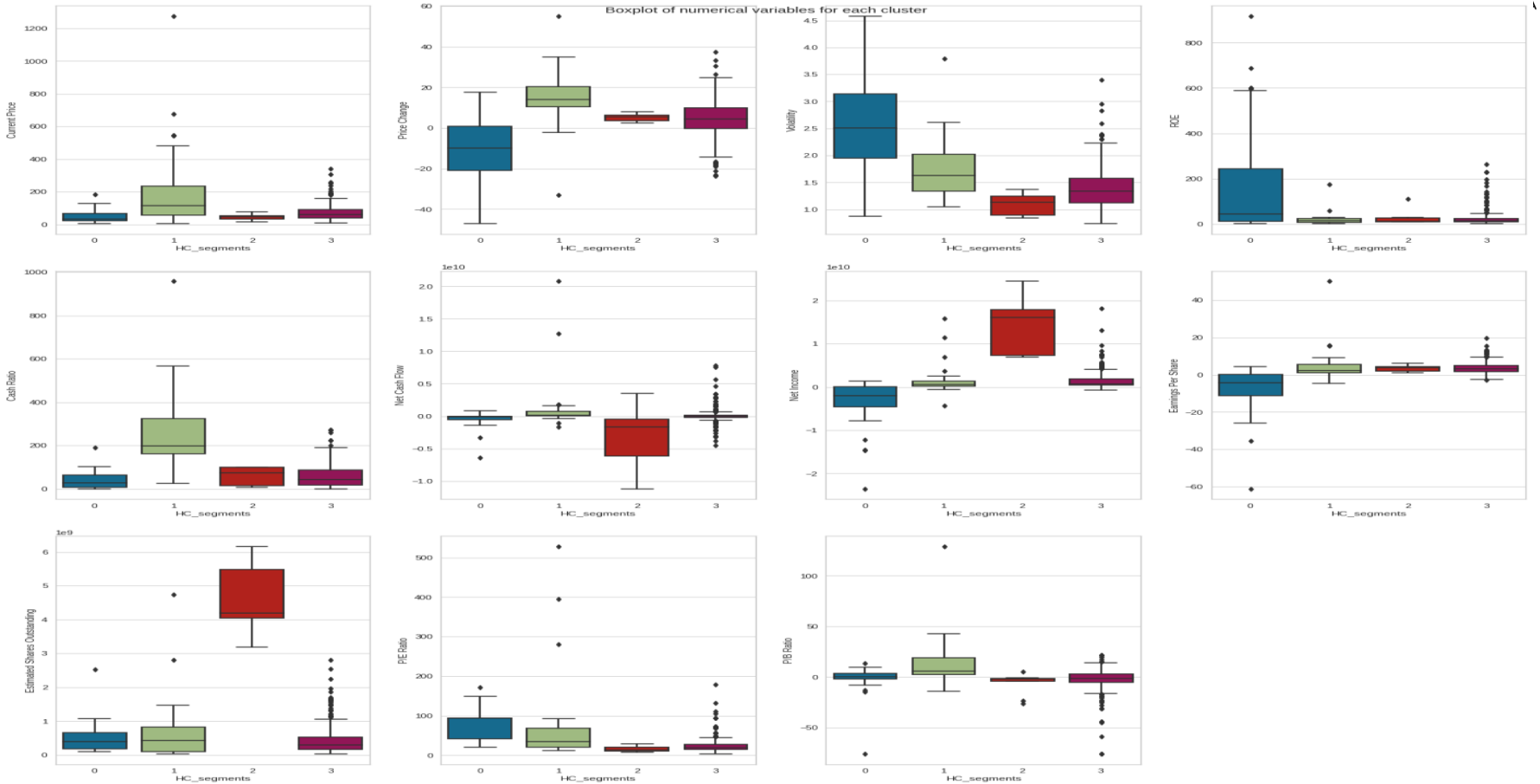
a. Most of stocks with the highest prices
b. Significant outliers in price-to-equity ratio
c. The most favorable price-to-book (P/B) ratios
d. Most of the highest cash ratios

Companies within cluster 2 have:

a. Most of the companies with the highest inflows and outflows of cash
b. The highest net incomes
c. The highest number of shares outstanding

Companies within cluster 3 have:

Most of outliers in price increases and some of the outliers in price decreases
Some of outliers in cash inflows and outflows
Most of the outliers in P/B ratio



K-Means vs Hierarchical Clustering

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

HC_segments	0	1	2	3
GICS Sector				
Consumer Discretionary	1	5	1	33
Consumer Staples	2	1	1	15
Energy	22	0	1	7
Financials	1	1	3	44
Health Care	0	8	1	31
Industrials	1	0	0	52
Information Technology	1	9	0	23
Materials	1	1	0	18
Real Estate	0	1	0	26
Telecommunications Services	0	1	2	2
Utilities	0	0	0	24

Although, minor differences here and there, groupings obtained with Hierarchical clustering using Eucledian distance & Ward linkage is similar to the one obtained using K-Means clustering!

Both algorithms give similar clusters, with a single cluster of a majority of the stocks and the remaining four clusters containing 7-29 stocks

Both algorithms yielded similar clusters based on the outliers within the 11 variables

BUSINESS INSIGHTS AND RECOMMENDATION

Securities were segregated into 4 different clusters identifying high & low performing, moderately performing stocks.

o This is important in an effort to split the stocks across investments that are diversified, enabling one to maximize earnings in any market condition .

However, it is important to keep in mind that stock market is often volatile, and past indicators may not always indicate future trends. Dynamic clustering & movement of stocks across cluster groups due to changing market conditions needs to be further analyzed for making better predictions.

Great Learning

Happy Learning !