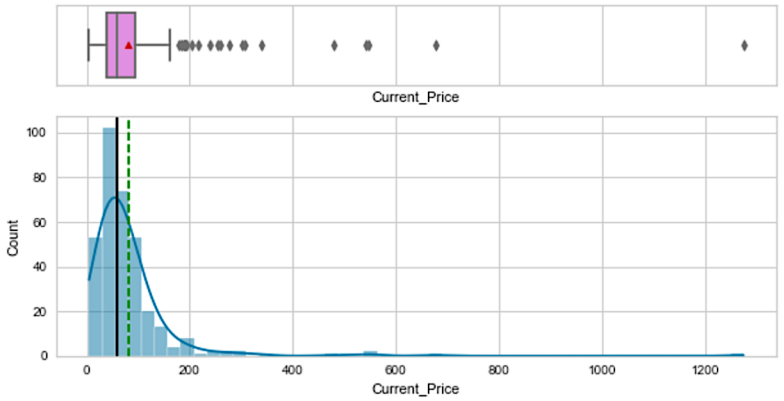
Machine Learning for Stock Clustering

- 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

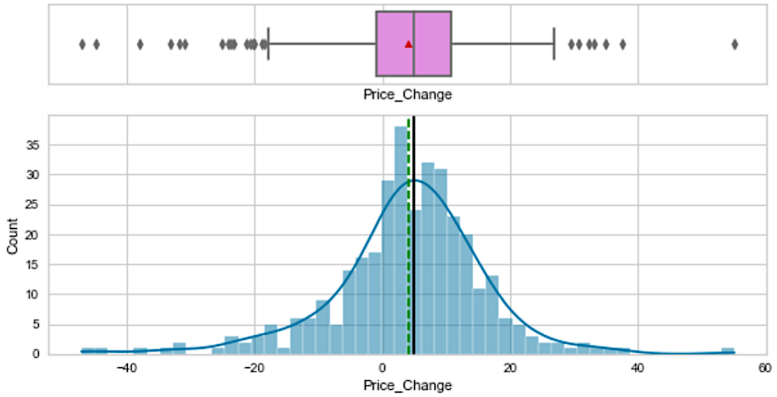
- Available data is of stocks listed under New York Stock Exchange with following financial indicators -

  - **Ticker Symbol:** abbreviation used to uniquely identify publicly traded shares
  - **Security**: name of the company
  - **GICS Sector:** specific economic sector assigned to a company by the Global Industry Classification Standard (GICS)
  - **GICS Sub Industry:** specific sub-industry group assigned to a company by the Global Industry Classification Standard (GICS)
  - **Current Price:** current stock price in dollars
  - **Price Change:** percentage change in the stock price in 13 weeks
  - **Net Cash Flow:** difference between a company's cash inflows and outflows (in dollars)
  - **Net Income:** revenues minus expenses, interest, and taxes (in dollars)
  - **Volatility:** standard deviation of the stock price over the past 13 weeks
  - **ROE:** net income by shareholders' equity (shareholders' equity is equal to a company's assets minus its debt)
  - **Cash Ratio:** ratio of a company's total reserves of cash and cash equivalents to its total current liabilities
  - **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)

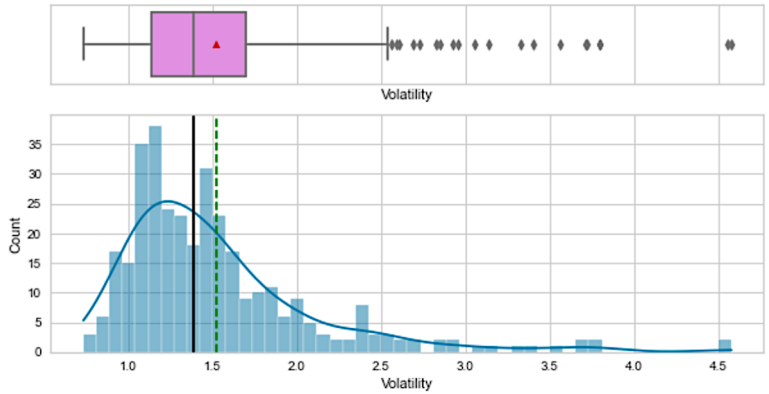
Data Overview – Univariate Analysis



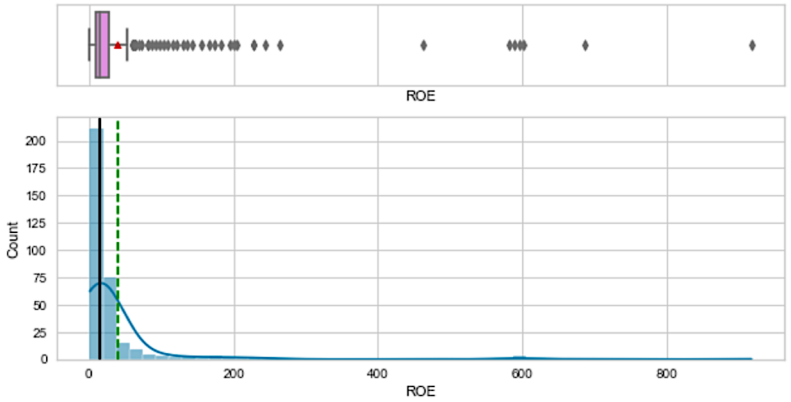
- Current_Price is right skewed with several positive outliers



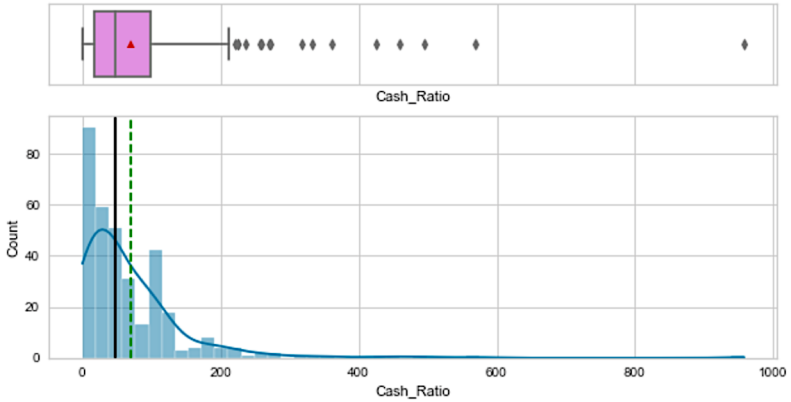
- Price_Change has a normal distribution with +ve & -ve outliers



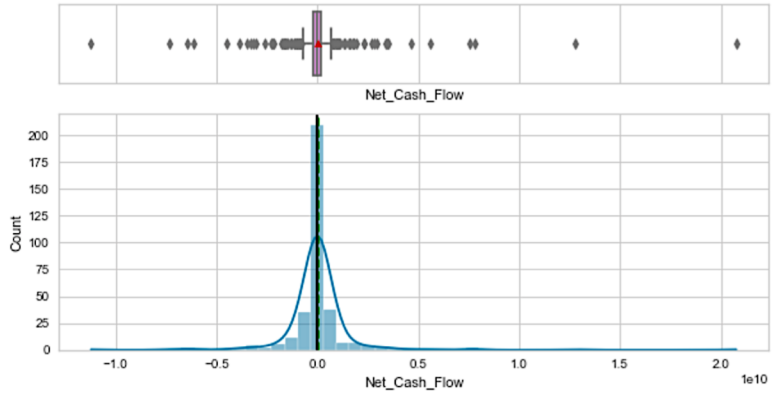
- Volatility is right skewed with some positive outliers



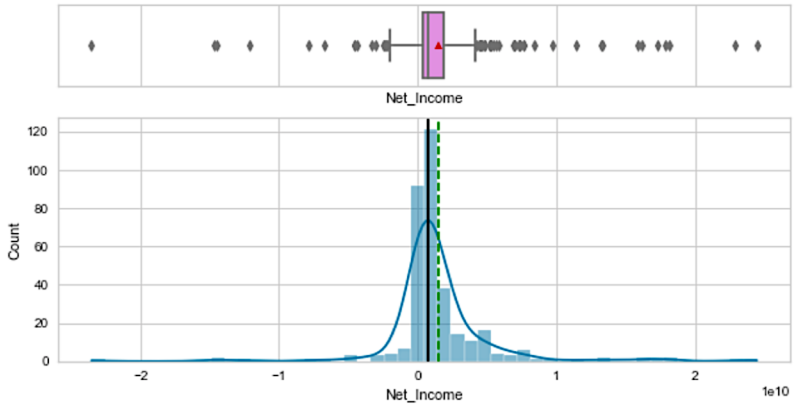
- ROE is right skewed with several +ve outliers



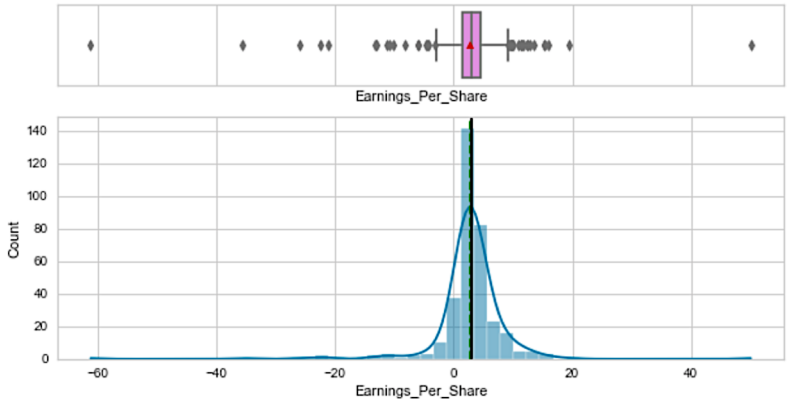
- Cash_Ratio is right skewed with +ve outliers



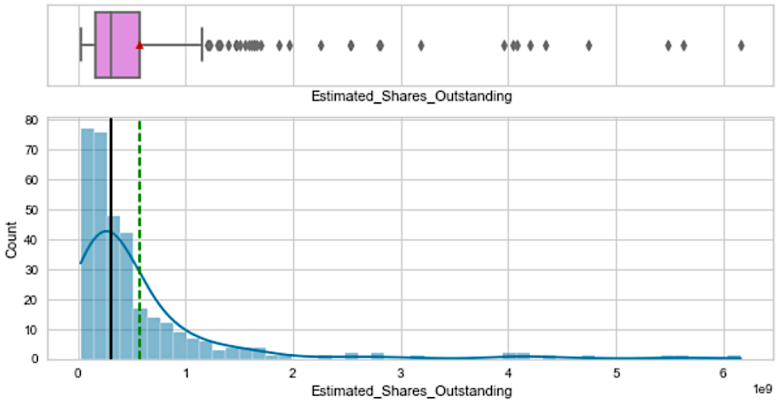
- Net_Cash_Flow has a normal distribution with +ve and –ve outliers



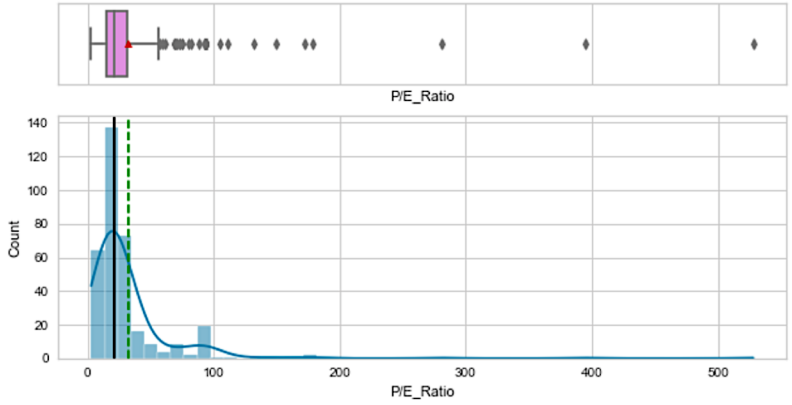
Net_Income has a normal distribution with some +ve and a few –ve outliers



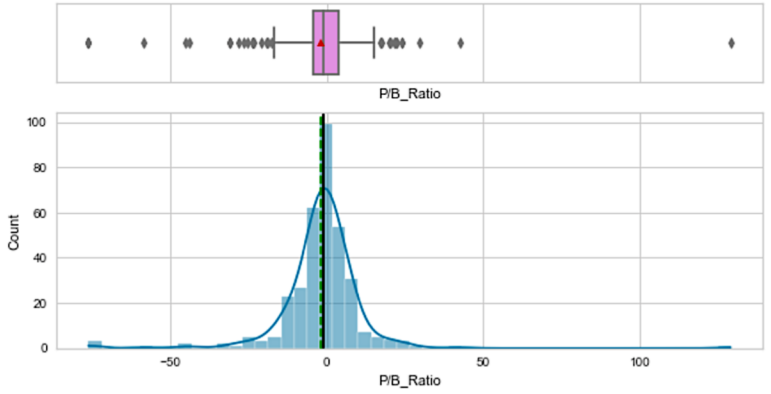
Earnings_Per_Share has a normal distribution with some +ve and –ve outliers



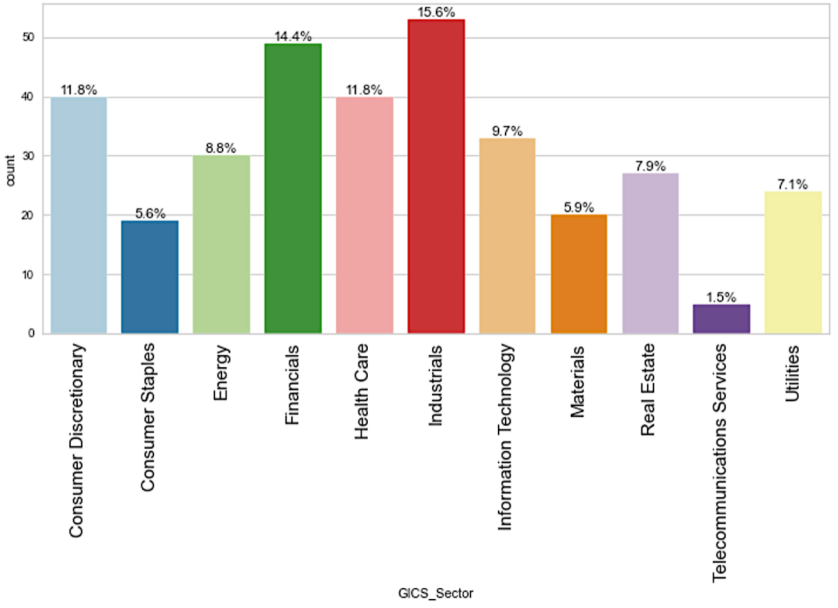
Estimated_Shares_Outstanding is right skewed with several +ve outliers



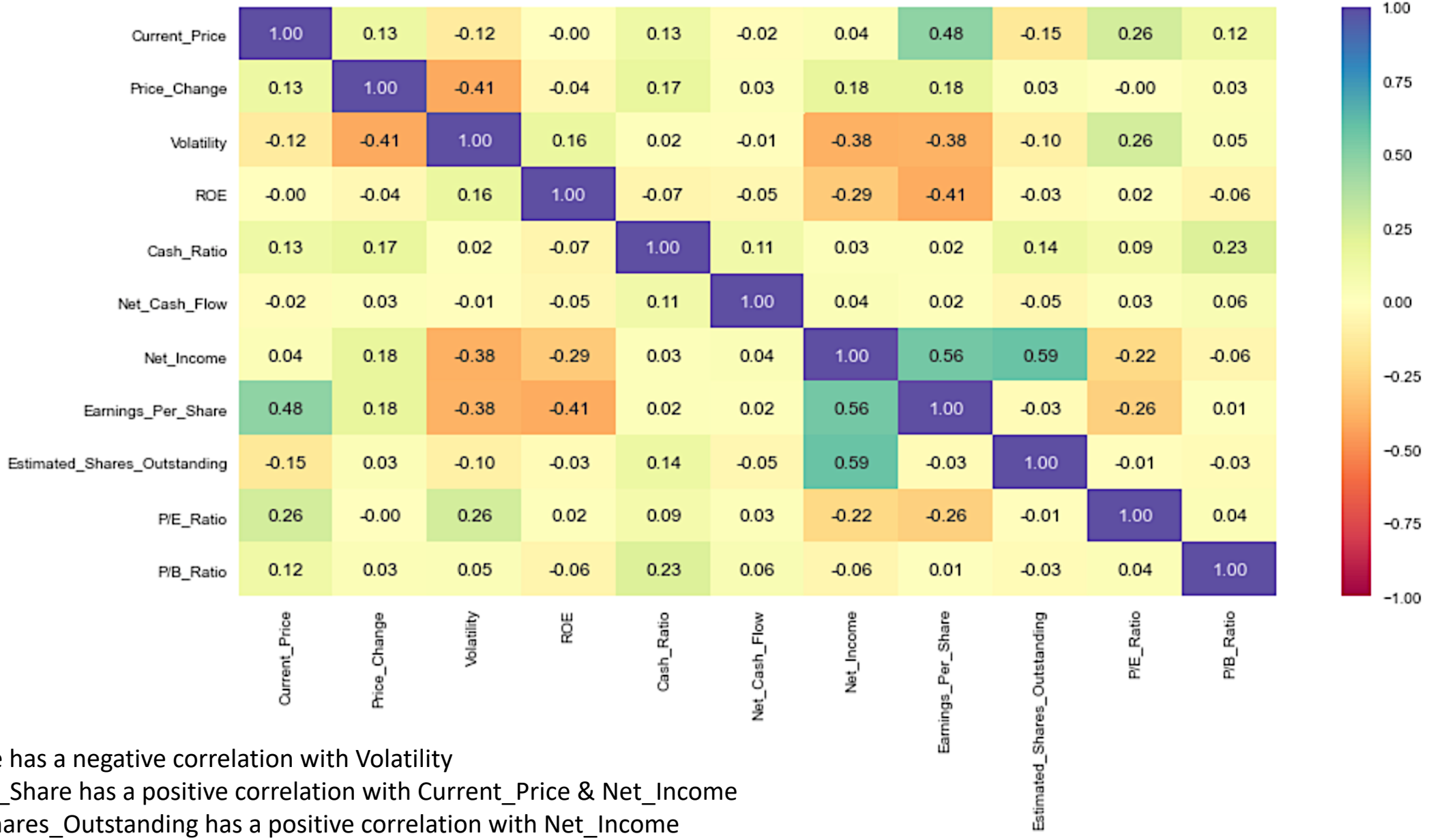
- P/E_Ratio is right skewed with some +ve outliers



- P/B_Ratio has a normal distribution with a few +ve and –ve outliers



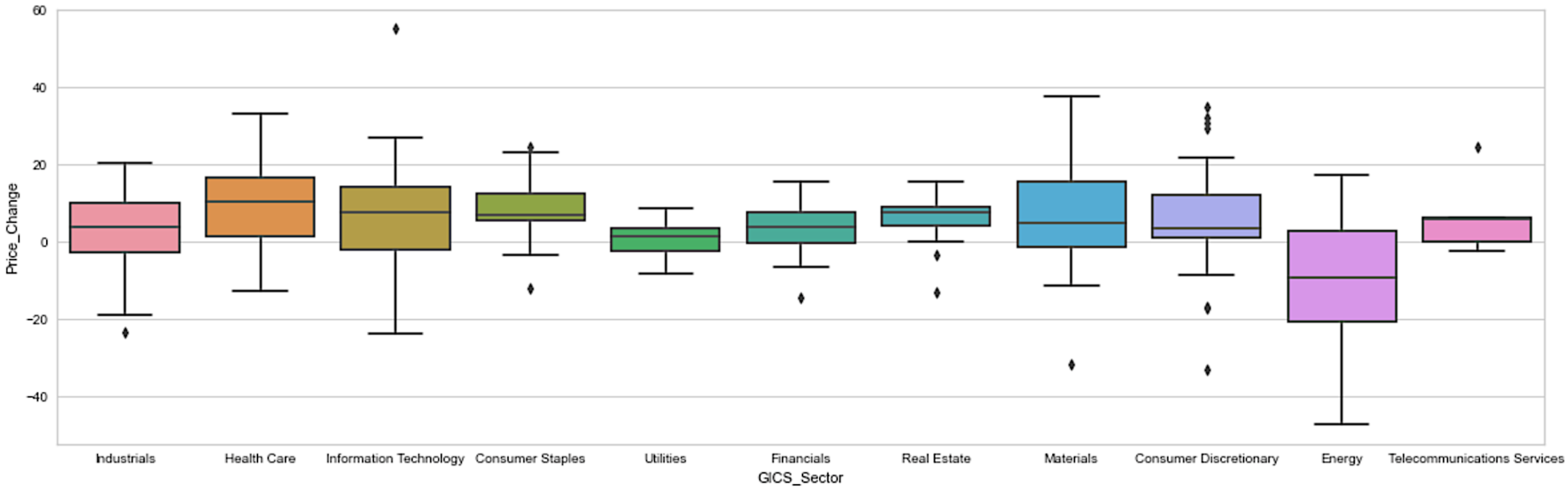
- Majority belong to “Industries” GICS and minority to “Telecommunication Services”



- 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

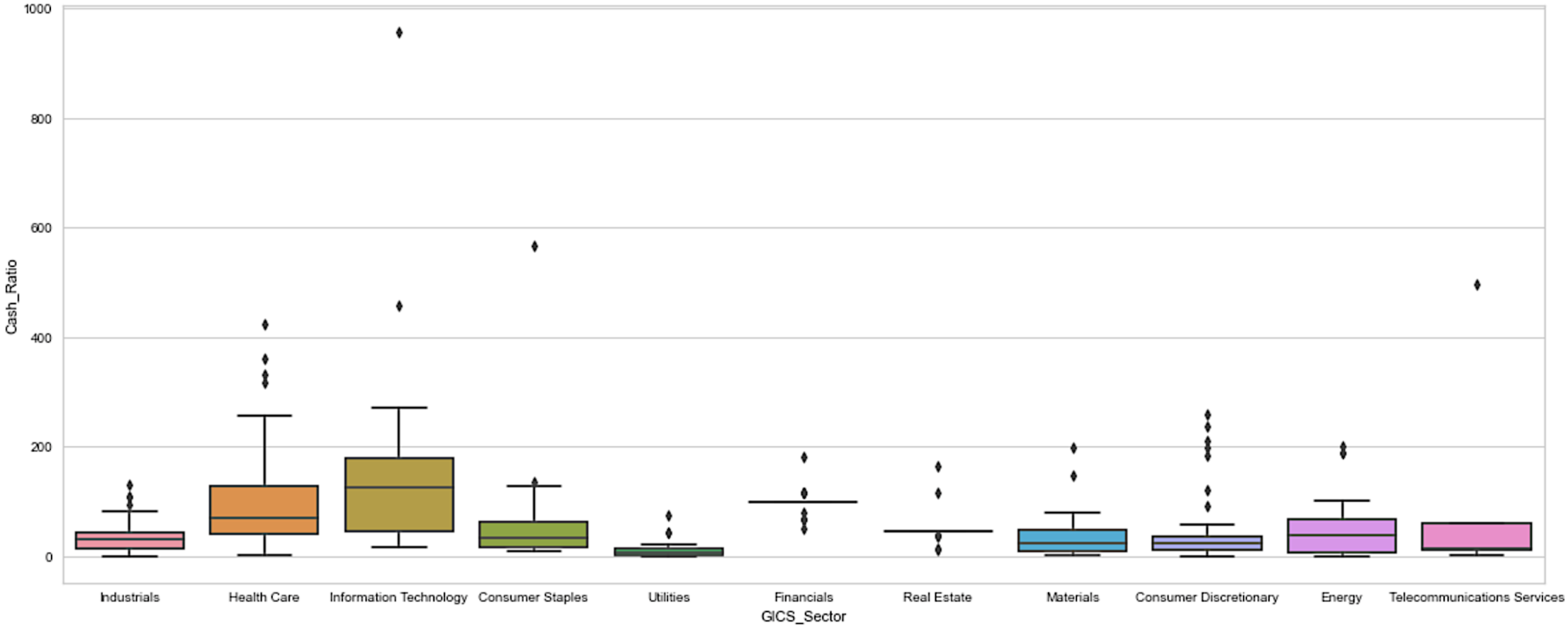
Data Overview – Bivariate Analysis

Price_Change vs. GICS_Sector



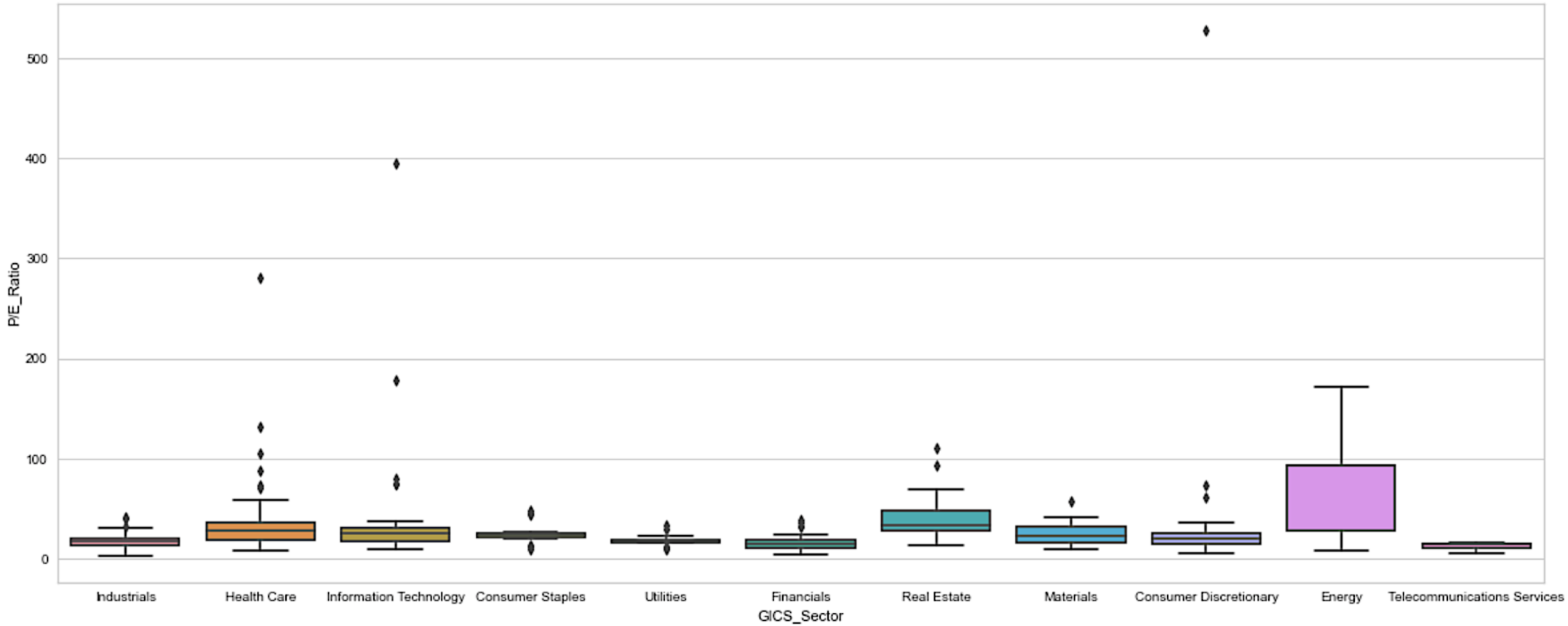
- Real_Estate has seen the minimum variation in Price_Change across different securities it encompasses while Energy GICS_Sector has seen the maximum variation in Price_Change across its securities
- Healthcare and Information Technology have maximum number of securities with a high positive Price_Change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

Cash_Ratio vs. GICS_Sector



- Real_Estate and Financials have the minimum Cash_Ratio variance across securities/companies in the respective GICS_sector while Informational Technology and Health Care have the maximum Cash_Ratio variance
- Informational_Sector and Financials also have high median Cash_Ratios in comparison to other GICS_Sectors; Informational_Technology has some of the highest positive outliers for Cash_Ratio

P/E_Ratio vs. GICS_Sector



- 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: All numerical attributes were transformed using standardscaler() to bring them to the same scale; with a mean of 0 and standard deviation of 1

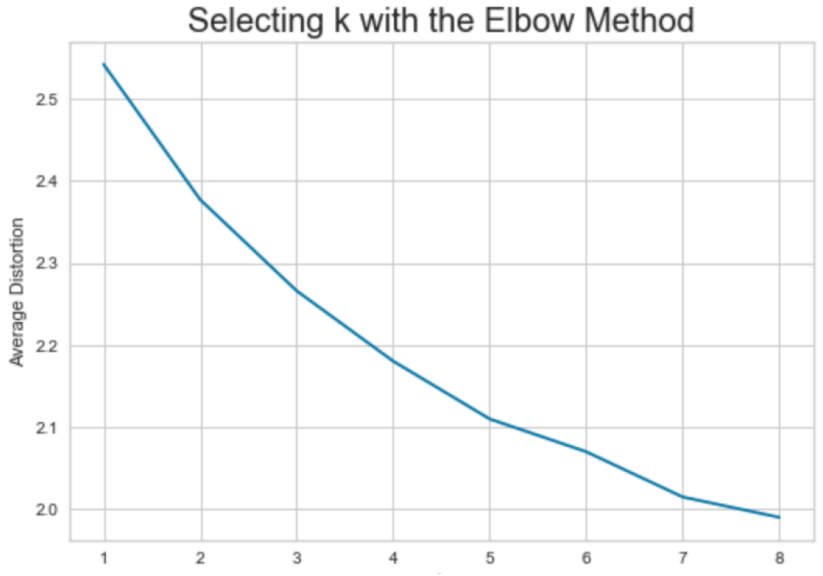
K-Means Clustering

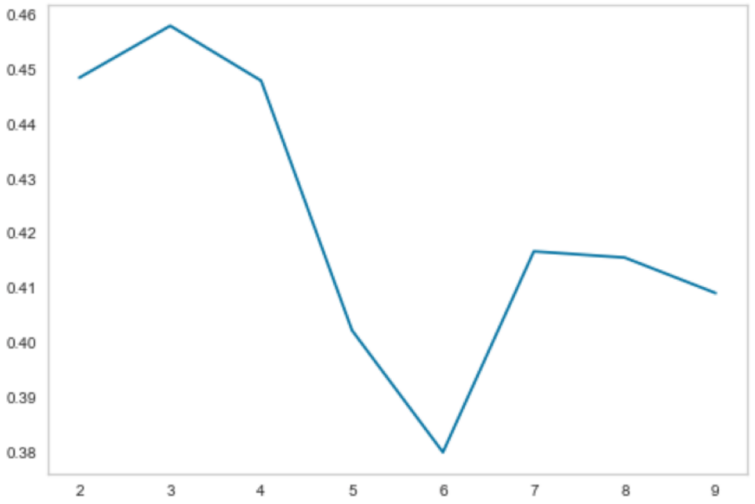
- **Cluster 0:** has 25 securities
high avg. Current_Price, high Cash_Ratio, high Earnings_Per_Share, high P/B_Ratio

- **Cluster 1:** 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 277 securities
Intermediate between clusters 0 & 1

- **Cluster 2:** has 11 securities ; similar to cluster 2; however has 10 times as high avg. Net_Income & Estimated_Shares_Outstanding



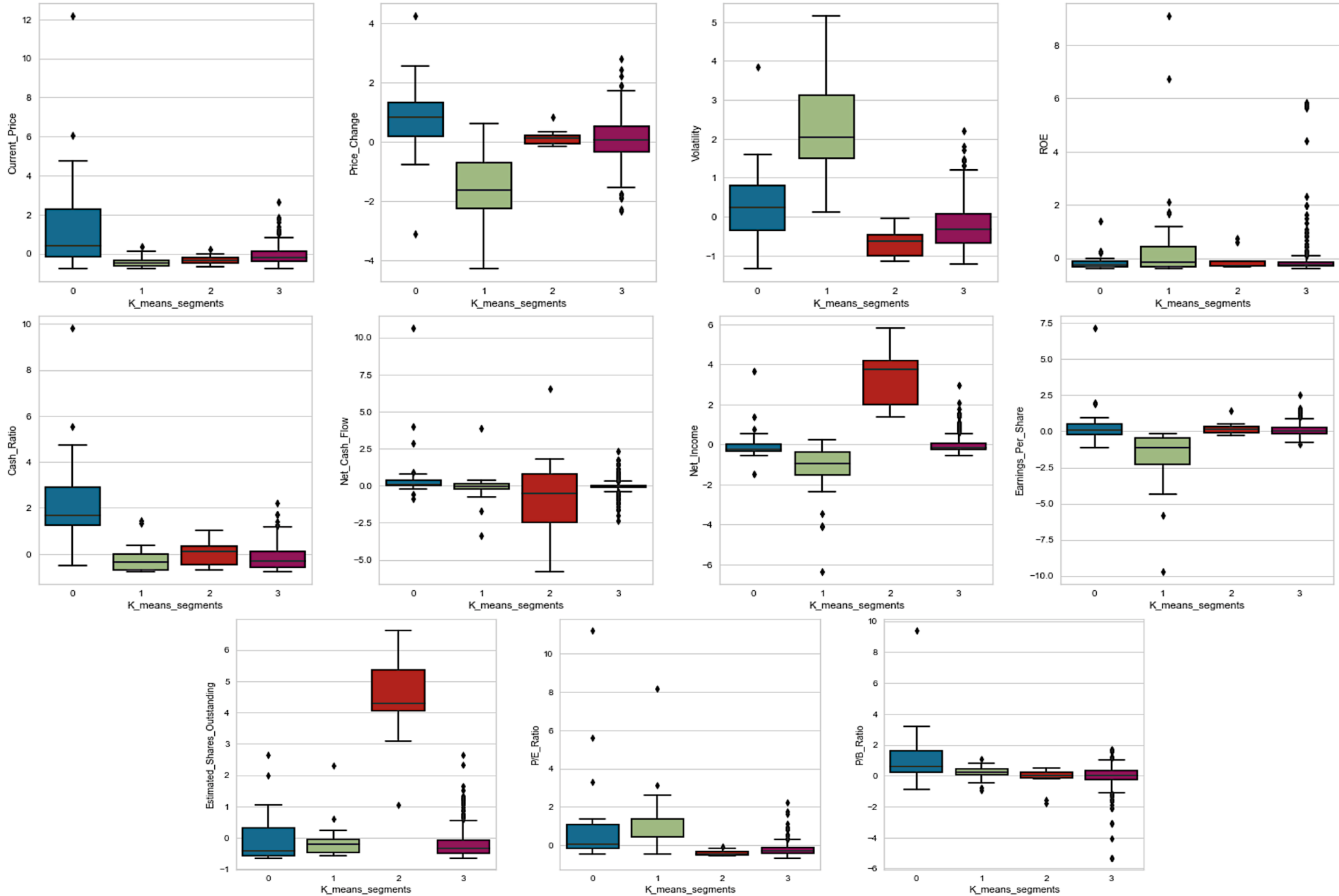


Silhouette score vs. K groupings

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

Groups	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_segments
0	234.170932	13.400685	1.729989	25.600000	277.640000	1.554927e+09	1.572612e+09	6.045200	5.783163e+08	74.960824	14.402452	25
1	38.099260	-15.370329	2.910500	107.074074	50.037037	-1.594285e+08	-3.887458e+09	-9.473704	4.803986e+08	90.619220	1.342067	27
2	50.517273	5.747586	1.130399	31.090909	75.909091	-1.072273e+09	1.483309e+10	4.154545	4.298827e+09	14.803577	-4.552119	11
3	72.399112	5.066225	1.388319	34.620939	53.000000	-1.404622e+07	1.482212e+09	3.621029	4.385338e+08	23.843656	-3.358948	277

Boxplot of numerical variables for each cluster



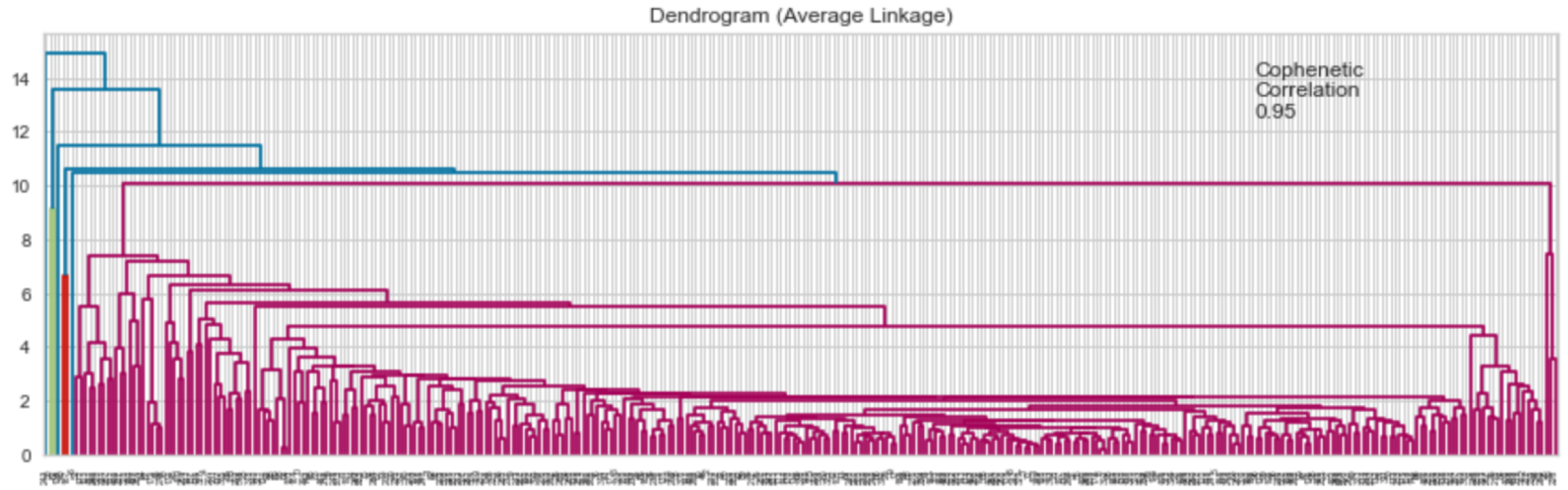
Clusters 3 and 2 are the safe clusters, with clusters 2 containing more exclusive securities. Clusters 0 and 1 are more riskier securities, former being high performing and later historically speaking low performing

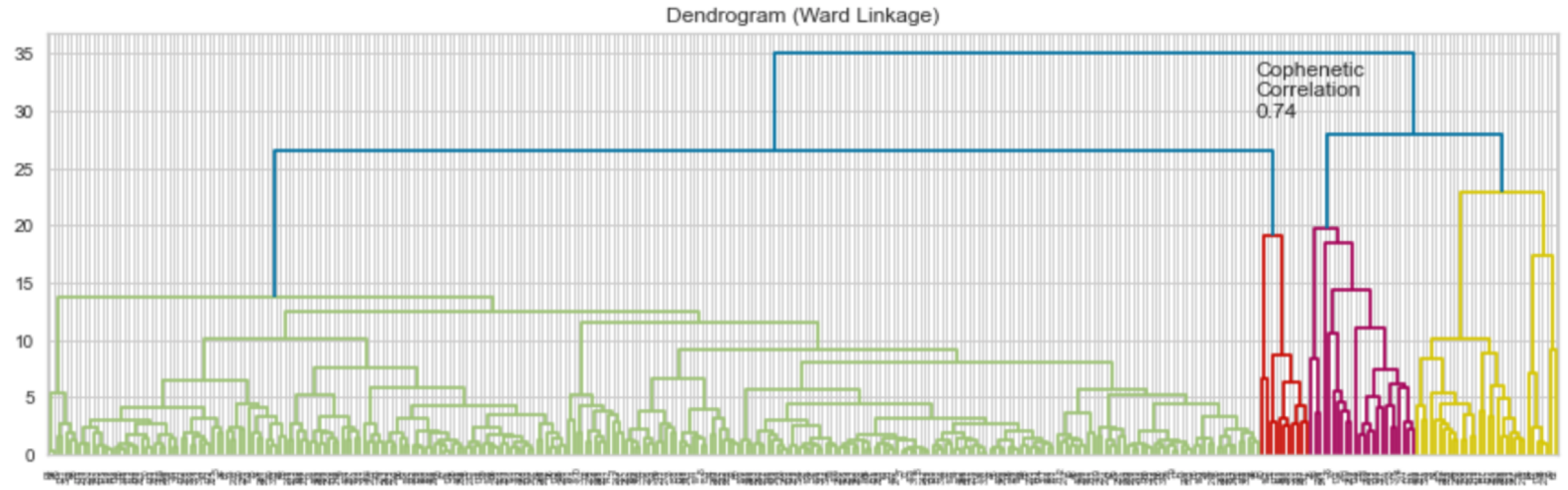
K_means_segments	0	1	2	3
GICS_Sector				
Consumer Discretionary	6	0	1	33
Consumer Staples	1	0	1	17
Energy	1	22	1	6
Financials	1	0	3	45
Health Care	9	0	2	29
Industrials	0	1	0	52
Information Technology	5	3	1	24
Materials	0	1	0	19
Real Estate	1	0	0	26
Telecommunications Services	1	0	2	2
Utilities	0	0	0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

- Among the securities in Cluster 0, majority are Health Care, followed by Consumer Discretionary & Informational Technology
- Cluster 1 is dominated by securities belonging to Energy division
- Cluster 2 is predominantly Financials
- Cluster 3 is diversified with Securities predominantly belonging to Industrials followed by Financials, Consumer Discretionary, Real Estate & Informational Technology

Hierarchical Clustering

- Cophenetic correlation was found to be highest for **Euclidian distance** in comparison to Chebyshev, Mahalanobis or Cityblock distance
- Cophenetic correlations were found to be highest for average or centroid linkages, but distinct clusters were observed only with ward linkage (providing enough variability among clusters)





Euclidean distance, Ward linkage, n=4 clusters were further chosen for hierarchical clustering

HC_Clusters	0	1	2	3
GICS_Sector				
Consumer Discretionary	1	32	6	1
Consumer Staples	2	15	1	1
Energy	23	6	0	1
Financials	1	44	1	3
Health Care	0	30	9	1
Industrials	2	51	0	0
Information Technology	2	24	7	0
Materials	1	19	0	0
Real Estate	0	26	1	0
Telecommunications Services	0	2	1	2
Utilities	0	24	0	0

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

- Cluster 0 of Hierarchical clustering
  - Similar to cluster 1 of K-Means clustering
- Cluster 1 of Hierarchical clustering
  - Similar to cluster 3 of K-Means clustering
- Cluster 2 of Hierarchical clustering
  - Similar to cluster 0 of K-Means clustering
- Cluster 3 of Hierarchical clustering
  - Similar to cluster 2 of K-Means clustering

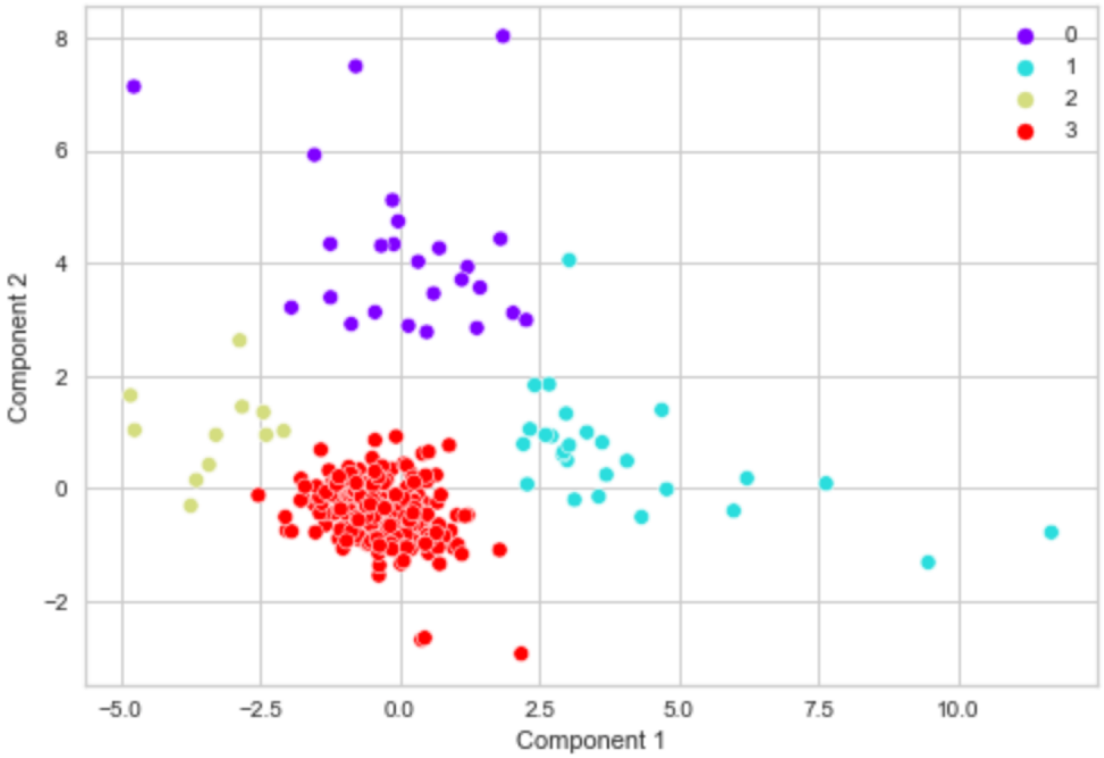
The industry segregation into clusters yielded similar results as well across both Hierarchical & K-means clustering

HC_Clusters	0	1	2	3
K_means_segments				
0	1	0	24	0
1	26	0	1	0
2	0	1	1	9
3	5	272	0	0

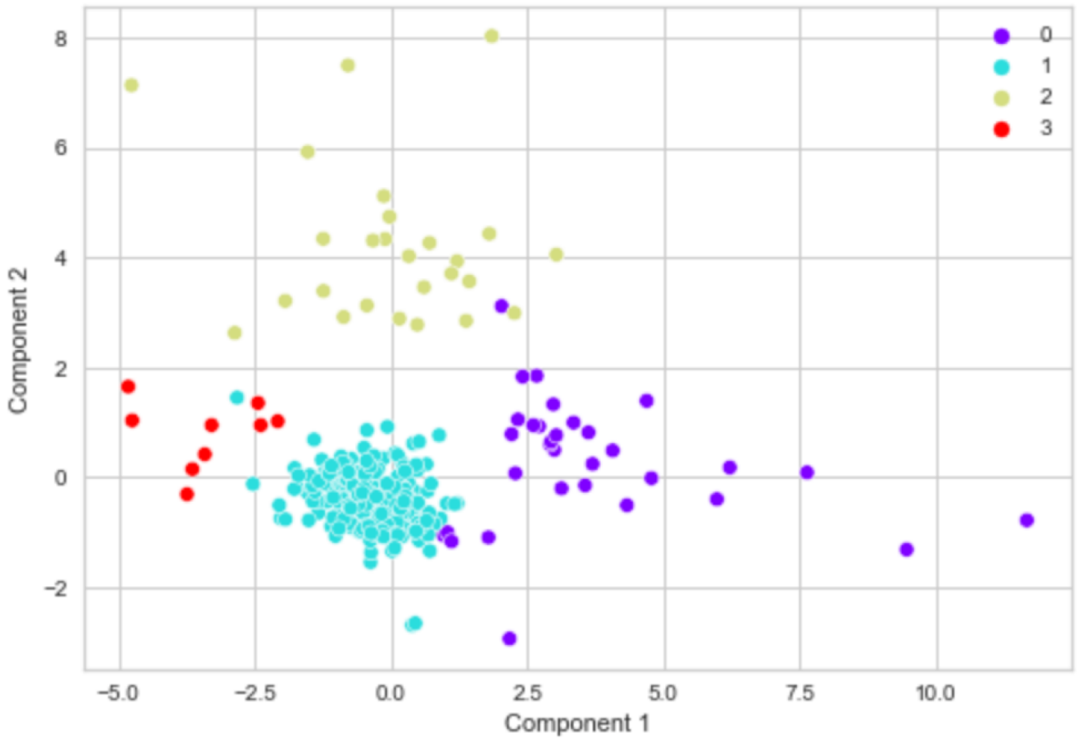
The 4 clusters as identified by K-Means and Hierarchical Clustering are able to group (26+272+24+9 i.e.,) 331 securities out of 340 securities in an identical fashion

Dimensionality Reduction using PCA for visualization

Scatterplot by cluster labels - K-Means Clustering



Scatterplot by cluster labels - Hierarchical Clustering



- The 4 clusters are well separated, with 1 cluster containing majority of data points and other clusters capturing the outliers
- While there are subtle differences in the plots obtained with K-Means clustering & Hierarchical clustering, broadly, there are more similarities than dissimilarities in the obtained clusters

Insights

Exploratory Data Analysis

- Current_Price of stocks, and Estimated_Shares_Outstanding is right skewed (with several positive outliers)
- Health Care and Financial sectors have seen high positive Price_Change in recent weeks, making them favourable
- Informational Technology and Financial sectors have high Cash_Ratio making them favourable
- Real Estate sector has seen minimum variation in Price_Change & minimum variation in Cash_Ratio making them a safer investment choice
- Energy sector has high variance in Price_Change, more volatility making them riskier. However, also has high P/E_Ratios (an investor is willing to invest more in a single share of a company in Energy sector per dollar value of its earnings)

Clustering Profiles

- Both clustering methods clustered 331 securities in a similar fashion with 9 securities being clustered differently. The industry segregation into clusters yielded similar results
- One cluster (25+securities) was identified as very aggressive (& high performing) belonging predominantly to Health Care followed by Consumer Discretionary and Information Technology sectors
- Another cluster (25+securities) was identified as very aggressive (but historically low performing) belonging predominantly to Energy sector
- Another cluster (~10securities) was identified as moderately aggressive (& high performing) belonging predominantly to Financials sector
- Finally, 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

Recommendations

- Securities were segregated into 4 different clusters identifying very aggressive (high & low performing), moderately aggressive (& high performing) and mildly aggressive options
    - This is important in an effort to split the stocks across investments that are diversified, enabling one to maximize earnings in any market condition

- Stock market is often volatile, and past indicators may not always indicate future trends. Dynamic clustering (as more data is added each day) & movement of stocks across cluster groups due to changing market conditions needs to be further analysed for making better predictions