




Trade and Ahead investment strategies

Contents / Agenda

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



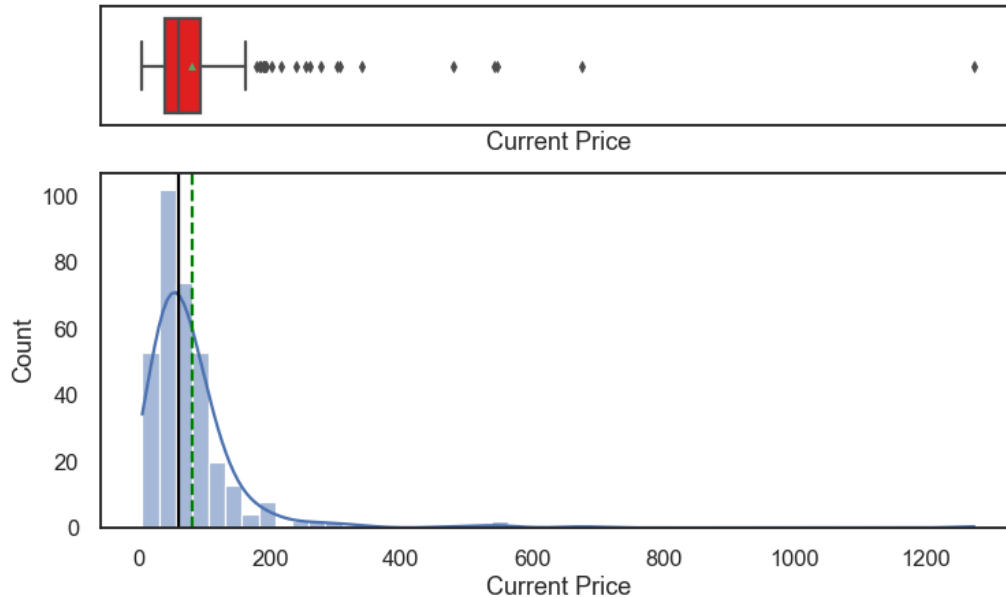
Business Problem Overview and Solution Approach

- Analyzing stocks across different market segments and Helping investors to protect against risks that could make the portfolio vulnerable to loss. Cluster analysis identifies stocks that exhibit similar characteristics.
- The objective of this project is to analyze the data, group the stocks based on the attributes provided, and sharing insights about the characteristics of each group.
- The approach: Applying clustering algorithms K_means and Hierarchical clustering. Choosing the best model and analyzing the results.

EDA Results

Distribution of stock prices:

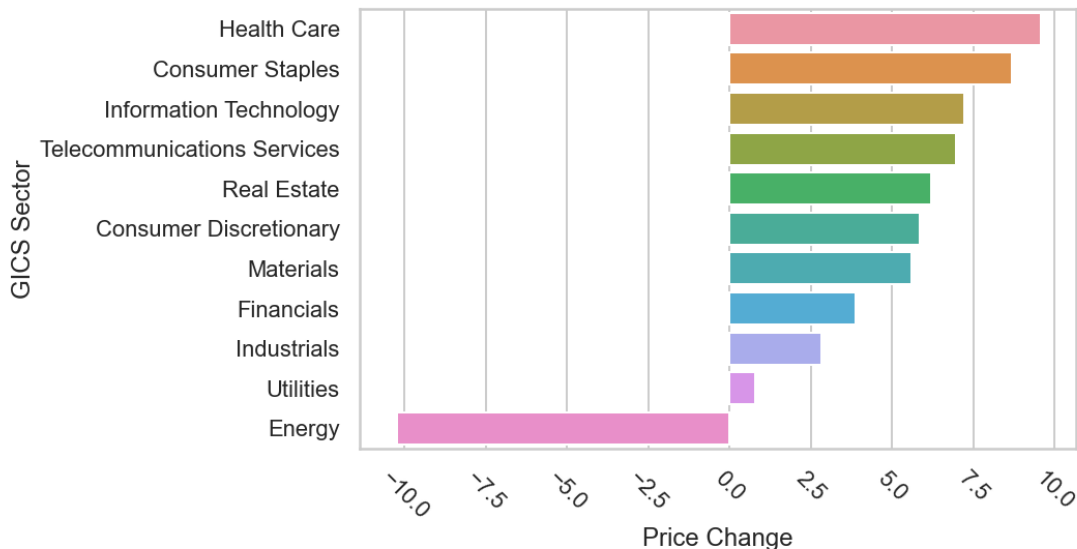
Current price skewed to right because of existing higher values. The mean is 80 and the median is 59 .



[Link to Appendix slide on data background check](#)

EDA Results

Economic sector have seen the maximum price increase on average:

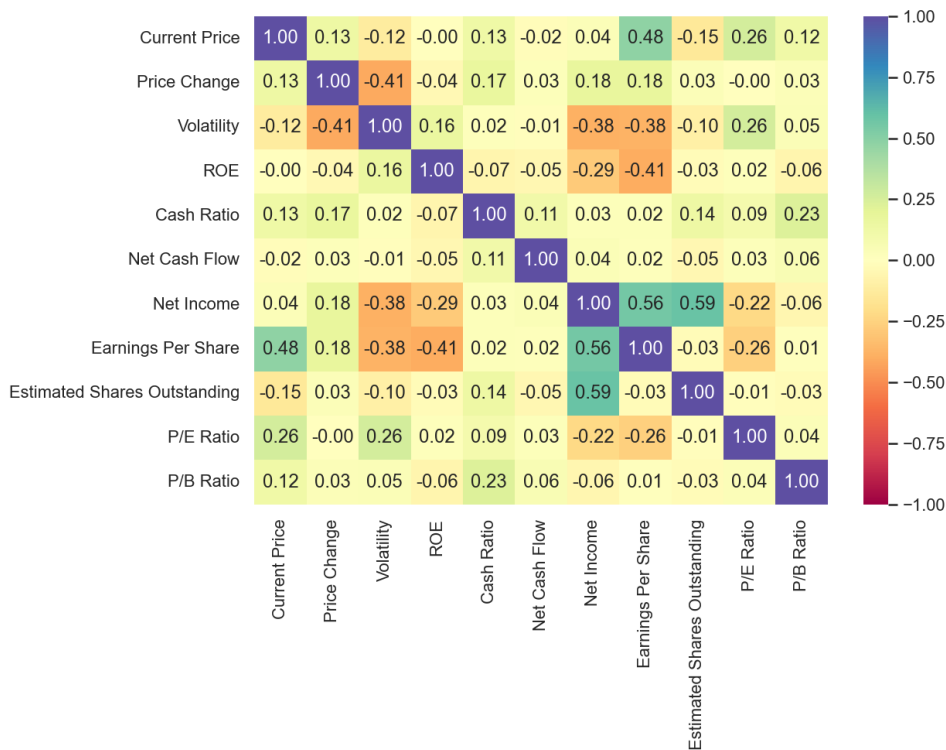


Health care has the highest positive price change following by consumer staples and information technology. While energy shows the highest negative price change

[Link to Appendix slide on data background check](#)

EDA Results

How are the different variables correlated with each other?

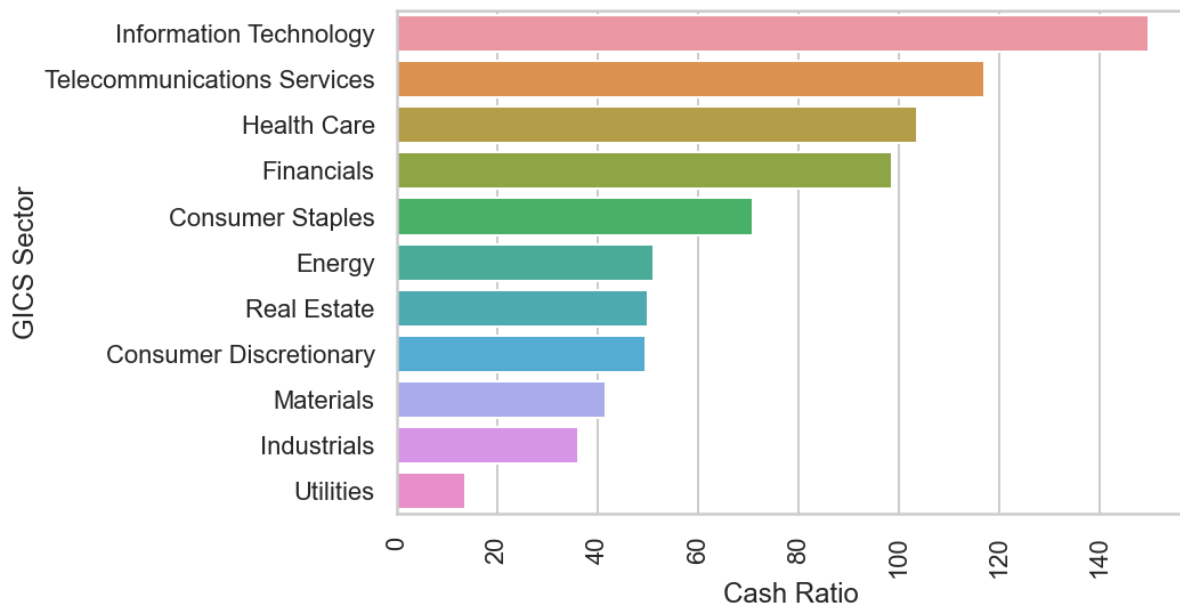


There is no high correlation between the numeric features. The highest is 0.59 for net income with estimated share outstanding

[Link to Appendix slide on data background check](#)

EDA Results

1. Cash ratio provides a measure of a company's ability to cover its short-term obligations using only cash and cash equivalents.



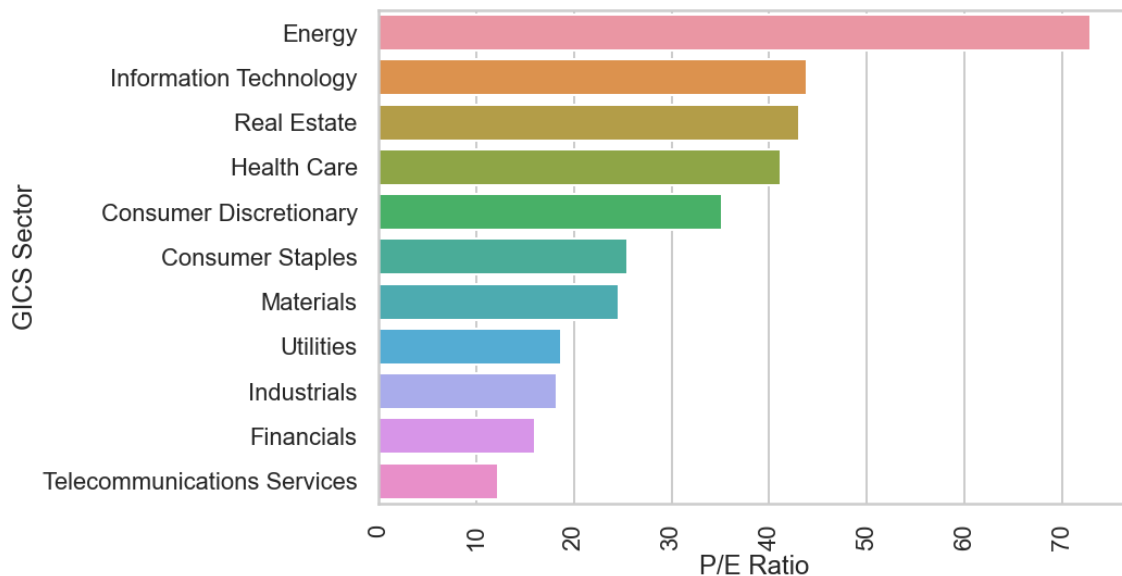
How does the average cash ratio vary across economic sectors?

Cash Ratio of information Technology, and Telecommunication services are the highest values following by Health Care and financials. While Utilities are the least Cash Ratio.

EDA Results

How does the P/E ratio vary, on average, across economic sectors?

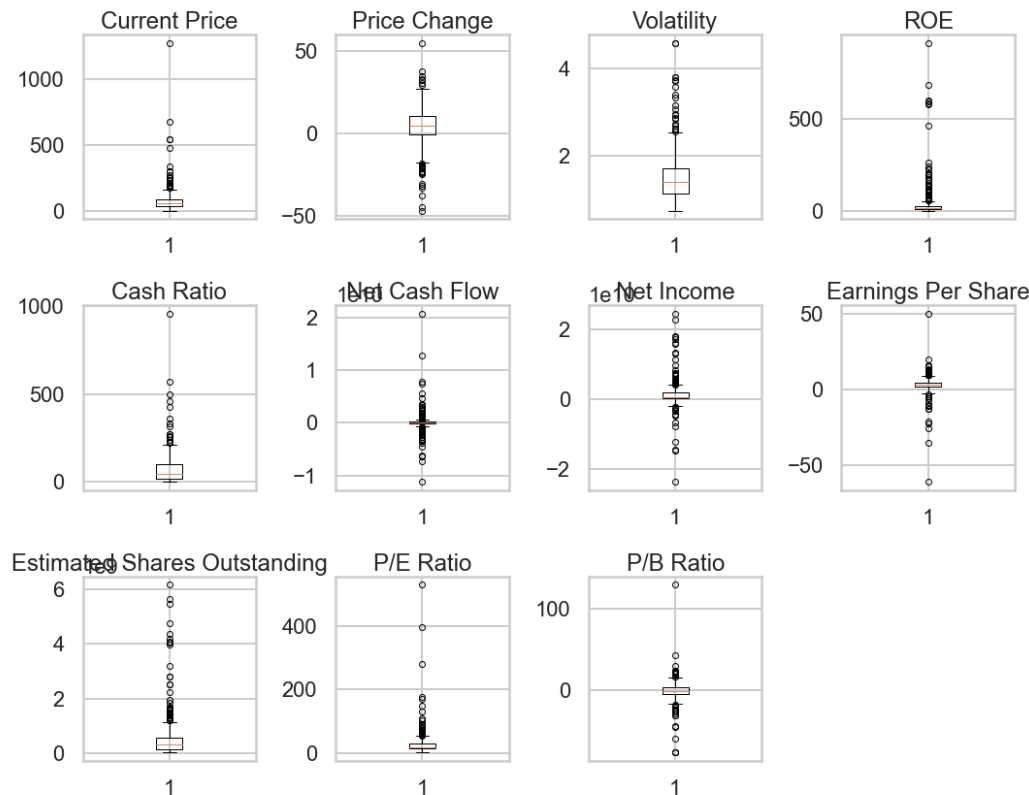
P/E ratios can help determine the relative value of a company's shares as they signify the amount of money an investor is willing to invest in a single share of a company per dollar of its earnings.



Energy has the highest P/E Ratio while the Telecommunications Services has the lowest P/E Ratio.

[Link to Appendix slide on data background check](#)

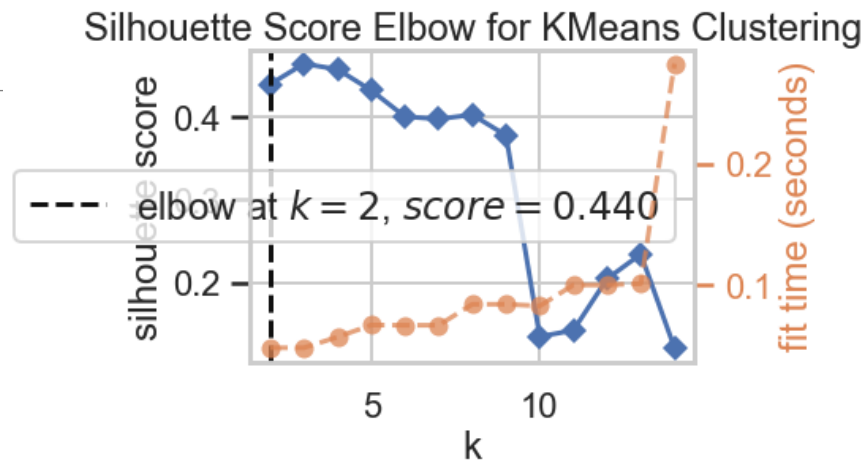
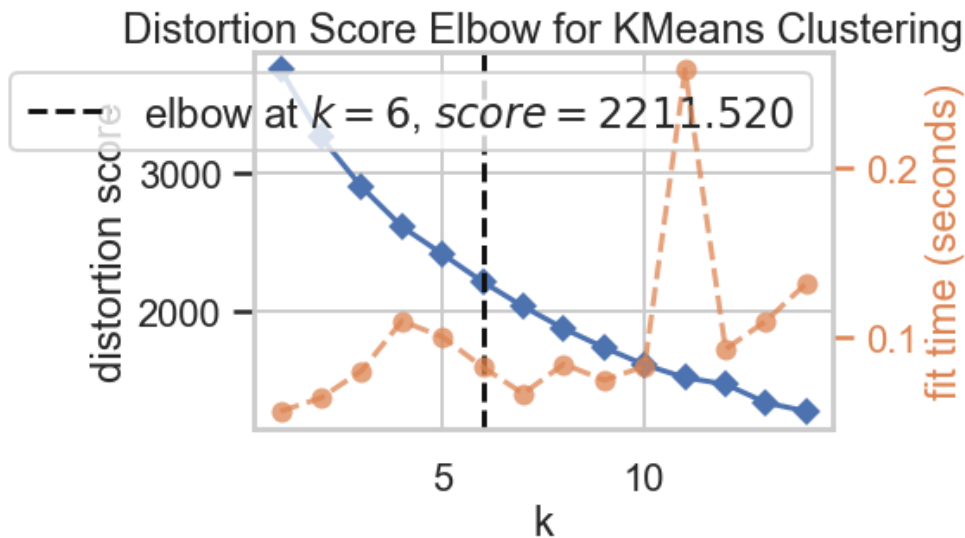
Data Preprocessing



- Duplicate value check
 - No duplicated values
- Missing value treatment
 - No missing values
- Outlier check (treatment if needed)
 - No need to outlier treatment
- Data preparation for modeling
 - Scaling the data using (z_score) StandardScaler from sklearn.

K-Means Clustering Summary

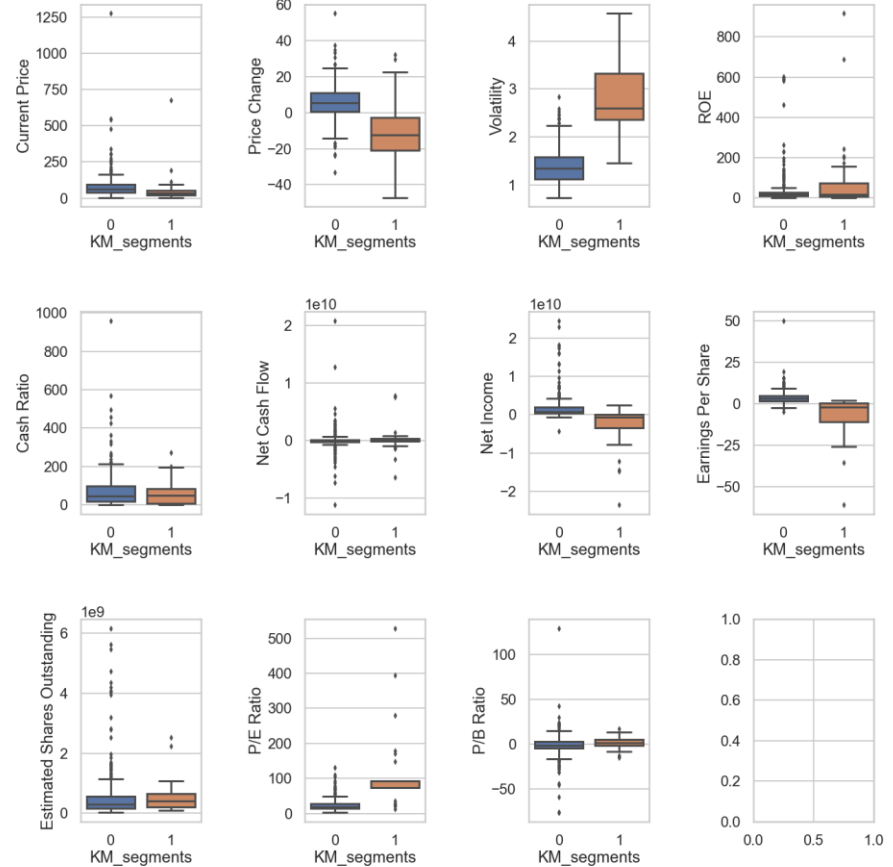
- Optimal Number of clusters using K-Means is 2 with Silhouette score 0.44 and during 0.02 s.



Number of clusters using Elbow method is 6 with score 2211 and during 0.05 s.

K-Means Clustering Summary

- Cluster Profiling
- Two distinguishable clusters.
- Net Cash Flow and Cash Ratio features median of the clusters are very similar. Which could result in overlaying in the clustering.



K-Means Clustering Summary

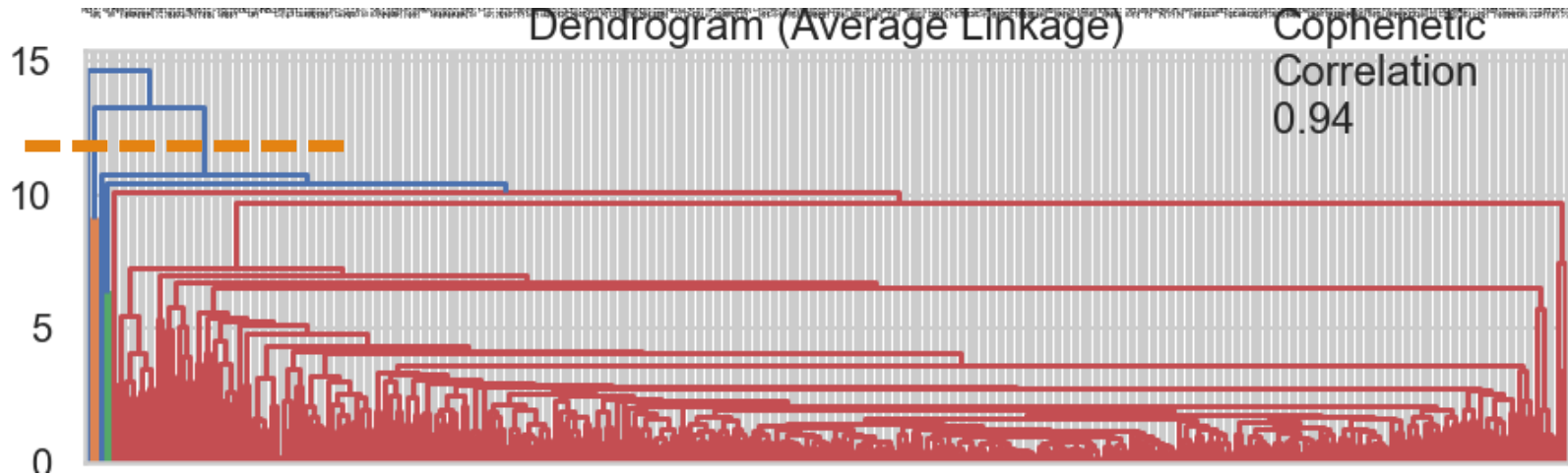
KM_segments (mean)	Cluster 0	Cluster 1
Current Price	82.786278	62.963940
Price Change	5.649217	-10.537087
Volatility	1.391766	2.774534
ROE	33.781759	93.696970
Cash Ratio	70.159609	68.757576
Net Cash Flow	44922850.162866	154287151.515152
Net Income	1993143179.153095	-3145581545.454545
Earnings Per Share	3.896270	-7.639091
Estimated Shares Outstanding	581977441.138534	530986678.995152
P/E Ratio	24.244484	110.461063
P/B Ratio	-2.080438	1.651207
count_in_each_segment	307	33

K-Means Clustering Summary

KM_segments/count sub-sector	Cluster 0	Cluster1
GICS Sector		
Consumer Discretionary	38	2
Consumer Staples	19	nan
Energy	6	24
Financials	49	nan
Health Care	39	1
Industrials	52	1
Information Technology	29	4
Materials	19	1
Real Estate	27	nan
Telecommunications Services	5	nan
Utilities	24	nan

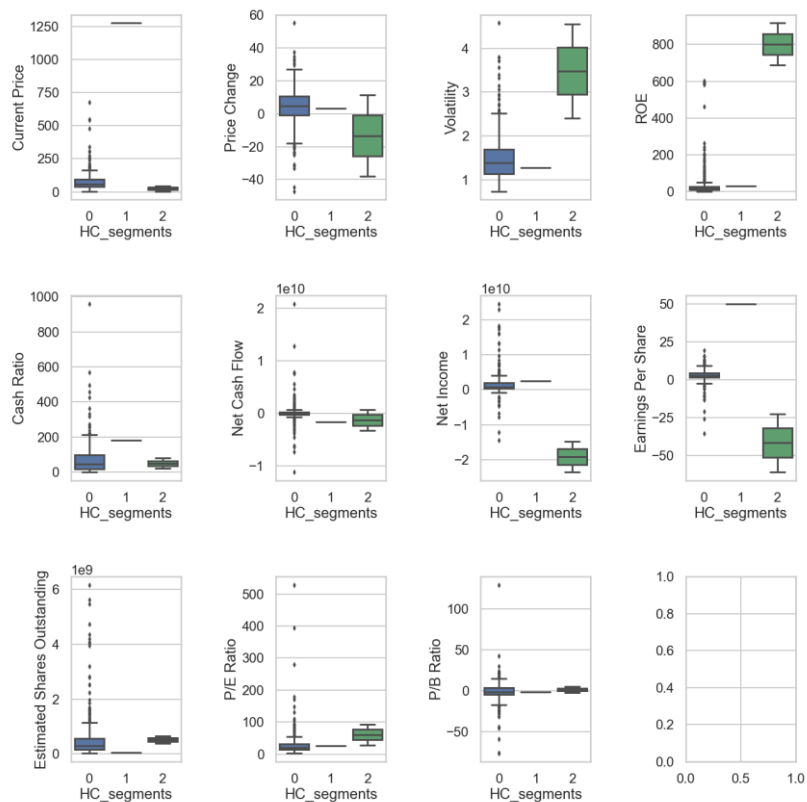
Hierarchical Clustering Summary

- Optimal Number of clusters using Hierarchical Clustering is 3 with Cophenetic Correlation 94%.



Hierarchical Clustering Summary

- The second cluster is only one value
- The differences between the first and third cluster is clear.



HC_segments (mean)	Cluster 0	Cluster 1	Cluster 2
Current Price	77.653642	1274.949951	24.485001
Price Change	4.184271	3.190527	-13.351992
Volatility	1.515129	1.268340	3.482611
ROE	35.103858	29	802
Cash Ratio	69.798220	184	51
Net Cash Flow	68662246.290801	-1671386000	-1292500000
Net Income	1613508620.178041	2551360000	-19106500000
Earnings Per Share	2.900905	50.090000	-41.815000
Estimated Shares Outstanding	578930419.447478	50935516.070000	519573983.250000
P/E Ratio	32.466828	25.453183	60.748608
P/B Ratio	-1.739711	-1.052429	1.565141
count_in_each_segment	337	1	2

Hierarchical Clustering Summary

Hierarchical Clustering Summary

HC_segments/count sub-sector	Cluster 0	Cluster 1	Cluster 2
GICS Sector			
Consumer Discretionary	39.0	1.0	NaN
Consumer Staples	19.0	NaN	NaN
Energy	28.0	NaN	2.0
Financials	49.0	NaN	NaN
Health Care	40.0	NaN	NaN
Industrials	53.0	NaN	NaN
Information Technology	33.0	NaN	NaN
Materials	20.0	NaN	NaN
Real Estate	27.0	NaN	NaN
Telecommunications Services	5.0	NaN	NaN
Utilities	24.0	NaN	NaN

Three clusters without distinguish features, it is not a accepted clustering model.

Executive Summary



The K_mean took less time than Hierarchical clustering.



K_mean clustering technique gives more distinct clusters while Hierarchical Clustering gave three non distinctual clusters therefore we choose K_mean results of clustering.



Observations in the similar clusters of both algorithms.

	K_mean (2 Clusters)	Hierarchical (3 Clusters)
Cluster 0	303	337
Cluster 1	37	1
Cluster 2		2

Executive Summary

Energy sector is a cluster while the rest of sectors belong to another cluster.

HC_segments/count sub-sector	Hierarchy Cluster 0	Hierarchy Cluster 1	Hierarchy Cluster 2	K_mean Cluster 0	K_mean Cluster1
GICS Sector					
Consumer Discretionary	39.0	1.0	NaN	38	2
Consumer Staples	19.0	NaN	NaN	19	Nan
Energy	28.0	NaN	2.0	6	24
Financials	49.0	NaN	NaN	49	Nan
Health Care	40.0	NaN	NaN	39	1
Industrials	53.0	NaN	NaN	52	1
Information Technology	33.0	NaN	NaN	29	4
Materials	20.0	NaN	NaN	19	1
Real Estate	27.0	NaN	NaN	27	Nan
Telecommunications Services	5.0	NaN	NaN	5	Nan
Utilities	24.0	NaN	NaN	24	Nan

Executive Summary

- For investors who invests in Energy will be recommend for them to invest in some of the companies in cluster 1 which include beside Energy companies such as Information Technology companies:
['Analog Devices, Inc.', 'Alexion Pharmaceuticals', 'Amazon.com Inc', 'Apache Corporation', 'Anadarko Petroleum Corp', 'Baker Hughes Inc', 'Chesapeake Energy', 'Cabot Oil & Gas', 'Concho Resources', 'Devon Energy Corp.', 'EOG Resources', 'EQT Corporation', 'Freeport-McMoran Cp & Gld', 'Halliburton Co.', 'Hess Corporation', 'Hewlett Packard Enterprise', 'Kinder Morgan', 'Marathon Oil Corp.', 'Murphy Oil', 'Noble Energy Inc', 'Netflix Inc.', 'Newfield Exploration Co', 'National Oilwell Varco Inc.', 'ONEOK', 'Occidental Petroleum', 'Quanta Services Inc.', 'Range Resources Corp.', 'Spectra Energy Corp.', 'Southwestern Energy', 'Teradata Corp.', 'Williams Cos.', 'Wynn Resorts Ltd', 'Cimarex Energy']
- We recommend to tracking the Price Change, Volatility, and P/E ratio because these features can differentiate between different clusters. (Cluster 1 shows negative Price Change and high Volatility while Cluster 0 shows positive Price Change and higher Volatility)
- Building diversified portfolio based on the two clusters which will help in decrease the risk of losing.

A solid orange vertical bar on the left side of the slide.

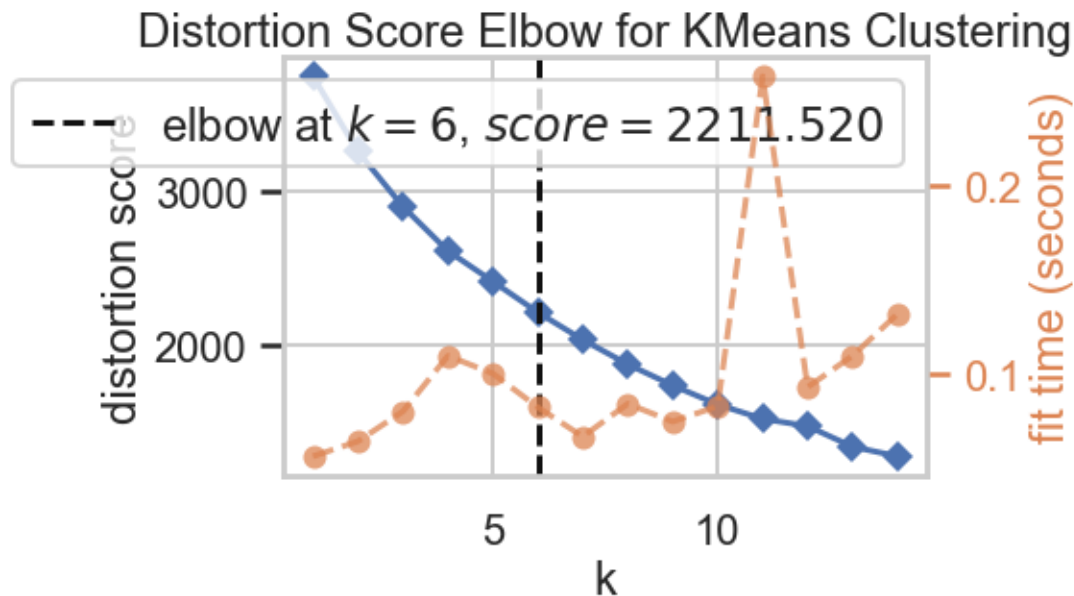
APPENDIX

Data Background and Contents

- 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)

K-Means Clustering Technique

- Observations using Elbow Curve along with visuals

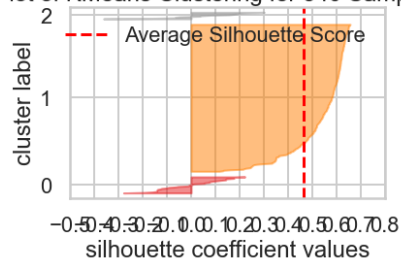


Elbow method shows
6 clusters for 0.02
seconds

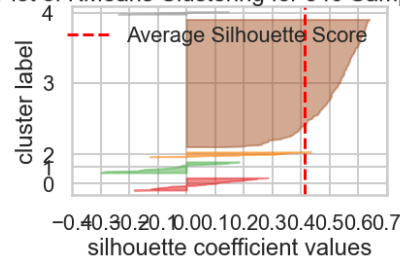
K-Means Clustering Technique

- Observations from Silhouette scores for different number of clusters

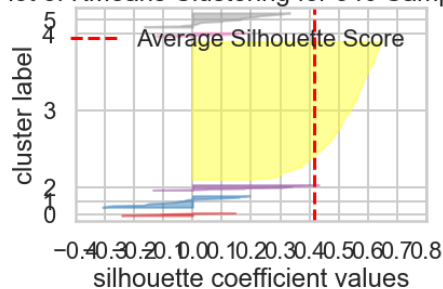
Silhouette Plot of KMeans Clustering for 340 Samples in 3 Centers



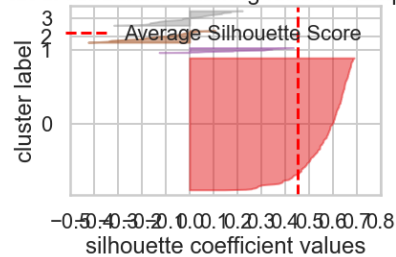
Silhouette Plot of KMeans Clustering for 340 Samples in 5 Centers



Silhouette Plot of KMeans Clustering for 340 Samples in 6 Centers



Silhouette Plot of KMeans Clustering for 340 Samples in 4 Centers



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

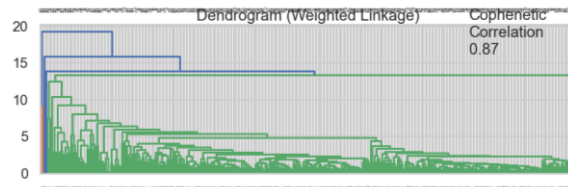
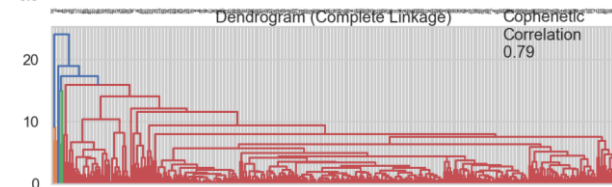
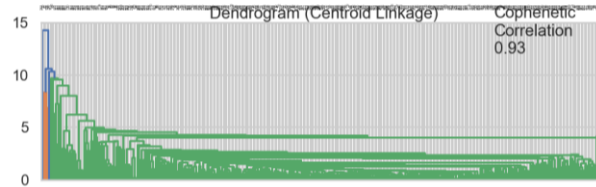
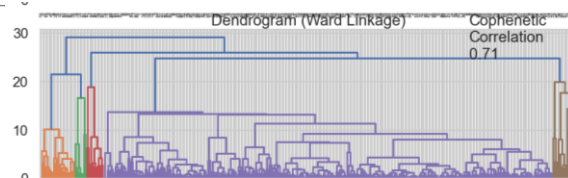
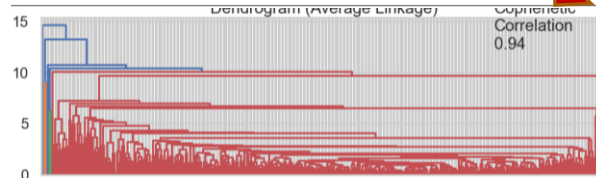
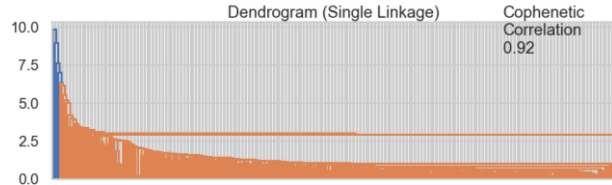
Hierarchical Clustering Technique

- Observations using different linkage methods

Hierarchical Clustering Technique

Winner

- Dendrograms for linkage methods used and their observations



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

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.

HC_segments/count sub-sector	Cluster 0	Cluster 1	Cluster 2	Cluster 0	Cluster1
GICS Sector					
Consumer Discretionary	39.0	1.0	NaN	38	2
Consumer Staples	19.0	NaN	NaN	19	Nan
Energy				6	24
Financials	49.0	NaN	NaN	49	Nan
Health Care	40.0	NaN	NaN	39	1
Industrials	53.0	NaN	NaN	52	1
Information Technology	33.0	NaN	NaN	29	4
Materials	20.0	NaN	NaN	19	1
Real Estate	27.0	NaN	NaN	27	Nan
Telecommunications Services	5.0	NaN	NaN	5	Nan
Utilities	24.0	NaN	NaN	24	Nan

K-Means vs Hierarchical Clustering

- Comparison of clusters obtained from K-Means and Hierarchical Clustering on various parameters
