Unsupervised Learning: Trade&Ahead Project

By: Esra Mercan

Contents

Context

The stock market has consistently proven to be a good place to invest in and save for the future. There are a lot of compelling reasons to invest in stocks. It can help in fighting inflation, create wealth, and also provides some tax benefits. Good steady returns on investments over a long period of time can also grow a lot more than seems possible. Also, thanks to the power of compound interest, the earlier one starts investing, the larger the corpus one can have for retirement. Overall, investing in stocks can help meet life's financial aspirations.

It is important to maintain a diversified portfolio when investing in stocks in order to maximise earnings under any market condition. Having a diversified portfolio tends to yield higher returns and face lower risk by tempering potential losses when the market is down. It is often easy to get lost in a sea of financial metrics to analyze while determining the worth of a stock, and doing the same for a multitude of stocks to identify the right picks for an individual can be a tedious task. By doing a cluster analysis, one 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.

Objective

Trade&Ahead is a financial consultancy firm who provide their customers with personalized investment strategies. They have hired you 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. They have assigned you the tasks of analyzing the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

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)

Exploratory Data Analysis (EDA)

This is how our sample data looks:

	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
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

```
data.shape
```

```
(340, 15)
```

There are 340 rows and 15 column in the data set

```
# checking for duplicate
data.duplicated().sum()
```

```
0
```

There is no duplicate values.

```
df.isnull().sum() ## Complete the
```

```
Ticker Symbol                  0
Security                       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
```

There is no missing values in the data.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 340 entries, 0 to 339
Data columns (total 15 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Ticker Symbol                340 non-null    object
 1   Security                     340 non-null    object
 2   GICS Sector                  340 non-null    object
 3   GICS Sub Industry            340 non-null    object
 4   Current Price                340 non-null    float64
 5   Price Change                 340 non-null    float64
 6   Volatility                   340 non-null    float64
 7   ROE                          340 non-null    int64
 8   Cash Ratio                   340 non-null    int64
 9   Net Cash Flow                340 non-null    int64
 10  Net Income                   340 non-null    int64
 11  Earnings Per Share           340 non-null    float64
 12  Estimated Shares Outstanding 340 non-null    float64
 13  P/E Ratio                    340 non-null    float64
 14  P/B Ratio                    340 non-null    float64
dtypes: float64(7), int64(4), object(4)
memory usage: 40.0+ KB
```

There are 4 object, 4 integer and 7 float type in the data.

Exploratory Data Analysis (EDA)

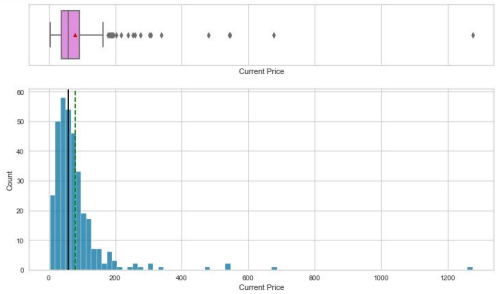
Let's check the statistical summary of the data.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ticker Symbol	340	340	GT	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Security	340	340	The Walt Disney Company	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sector	340	11	Industrials	53	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sub Industry	340	104	Oil & Gas Exploration & Production	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Current Price	340.0	NaN	NaN	NaN	80.862345	98.055086	4.5	38.555	59.705	92.880001	1274.949951
Price Change	340.0	NaN	NaN	NaN	4.078194	12.006338	-47.129693	-0.939484	4.819505	10.695493	55.051683
Volatility	340.0	NaN	NaN	NaN	1.525976	0.591798	0.733163	1.134878	1.385593	1.695549	4.580042
ROE	340.0	NaN	NaN	NaN	39.597059	96.547538	1.0	9.75	15.0	27.0	917.0
Cash Ratio	340.0	NaN	NaN	NaN	70.023529	90.421331	0.0	18.0	47.0	99.0	958.0
Net Cash Flow	340.0	NaN	NaN	NaN	55537620.588235	1946365312.175789	-11208000000.0	-193906500.0	2098000.0	169810750.0	20764000000.0
Net Income	340.0	NaN	NaN	NaN	1494384602.941176	3940150279.327937	-23528000000.0	352301250.0	707336000.0	1899000000.0	24442000000.0
Earnings Per Share	340.0	NaN	NaN	NaN	2.776662	6.587779	-61.2	1.5575	2.895	4.62	50.09
Estimated Shares Outstanding	340.0	NaN	NaN	NaN	577028337.754029	845849595.417695	27672156.86	158848216.1	309675137.8	573117457.325	6159292035.0
P/E Ratio	340.0	NaN	NaN	NaN	32.612563	44.348731	2.935451	15.044653	20.819876	31.764755	528.039074
P/B Ratio	340.0	NaN	NaN	NaN	-1.718249	13.966912	-76.119077	-4.352056	-1.06717	3.917066	129.064585

- We can see that there top Ticker Symbol is GT and top Security is The Walt Disney Company.The top GICS Sector is Industrials with frequency of 53 and the top GICS Sub Industry is Oil and Gas Exploration and Production.
- Net Cash Flow, Net Income and Estimated Shares Outstanding has a huge range.

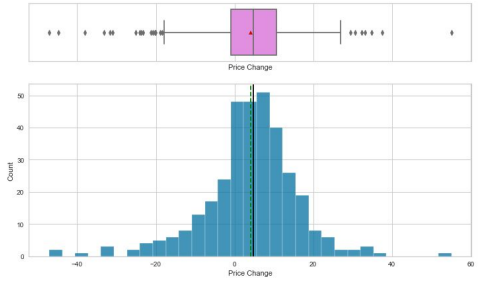
Exploratory Data Analysis (EDA)

Current Price



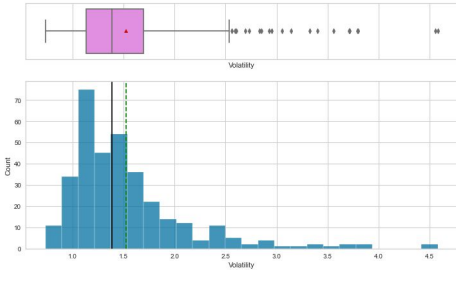
Current price data is right skewed, there are outliers.

Price Change



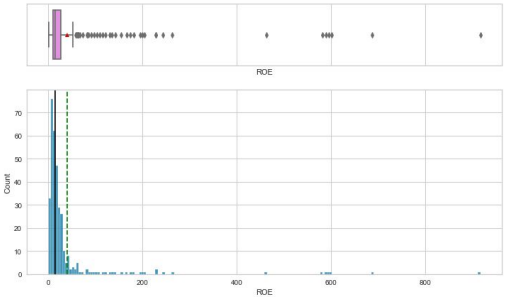
Price change data has close to a normal distributions. There are some outliers.

Volatility



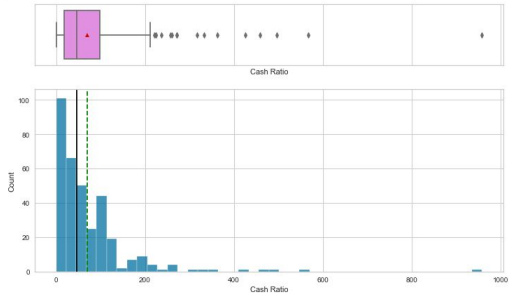
Volatility data is right skewed. There are outliers.

ROE



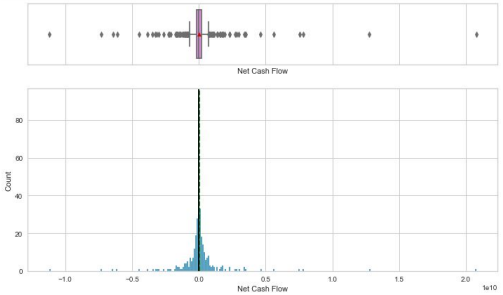
ROE data is also right skewed. There are outliers.

Cash Ratio



Cash Ratio is right skewed.

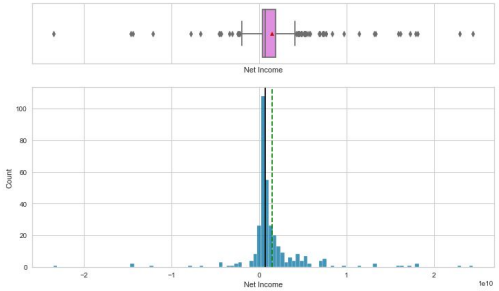
Net Cash Flow



Net Cash Flow data has close to a normal distributions. There are some outliers.

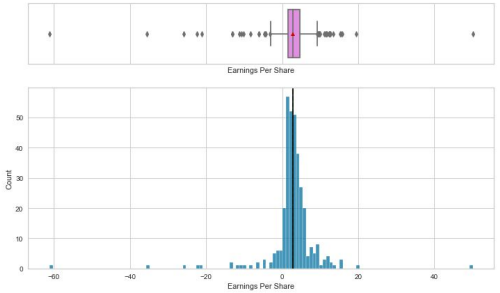
Exploratory Data Analysis (EDA) - continued

Net Income



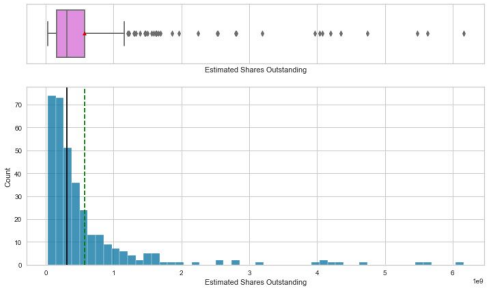
Net income data has close to a normal distributions. There are some outliers.

Earnings Per Share



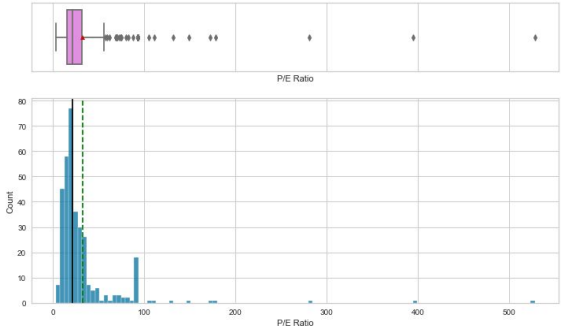
Earning Per Share data has a normal distributions. There are outliers.

Estimated Shares Outstanding



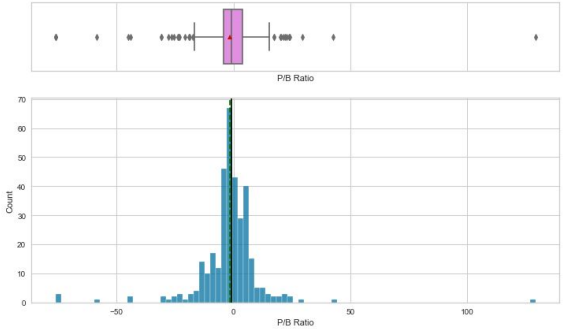
Estimated Shares Outstanding data is right skewed, there are outliers.

P/E Ratio



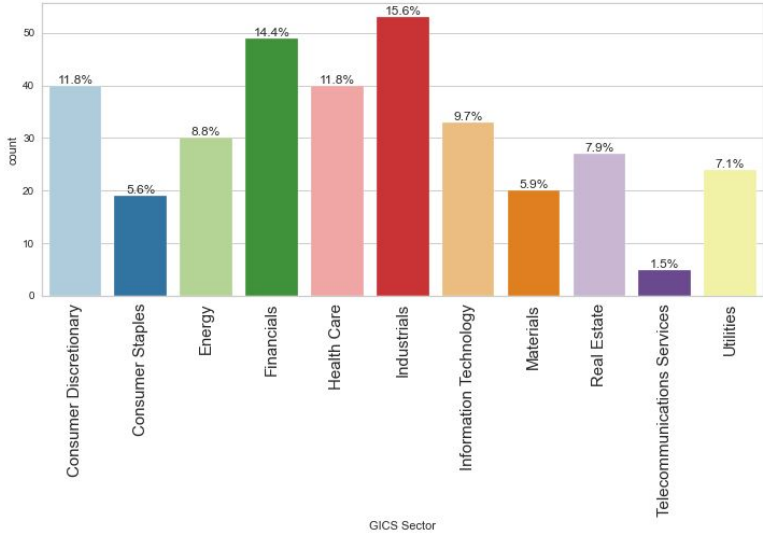
P/E Ratio data is right skewed, there are outliers.

P/B Ratio

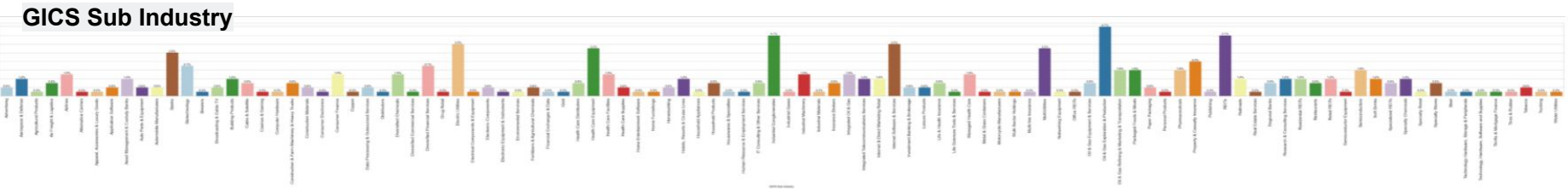


P/B Ratio data has close to a normal distributions. There are some outliers.

Exploratory Data Analysis (EDA) - continued



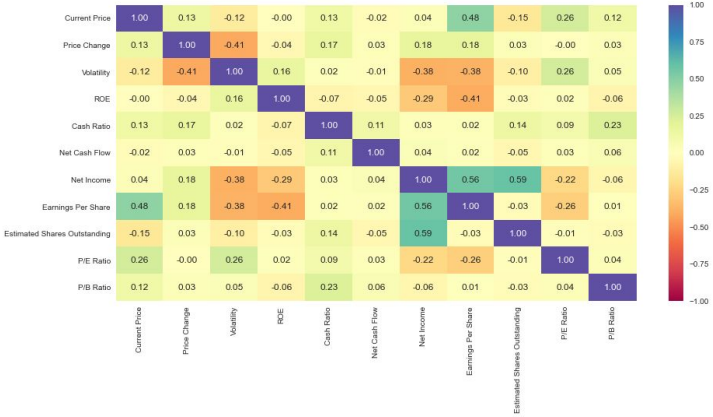
Industrials 15.6%, Financials 14.4%, Consumer Discretionary 11.8% and Health Care 11.8%.



- Oil Gas Exploration and Production has the highest(4.7%), Industrial Conglomerates (4.1%), REITs (4.1%), Internet Software & Services(3.5%), Electric Utilities(3.5%), and Health Care Equipment (3.2%) are the other leading GICS Sub Industry.

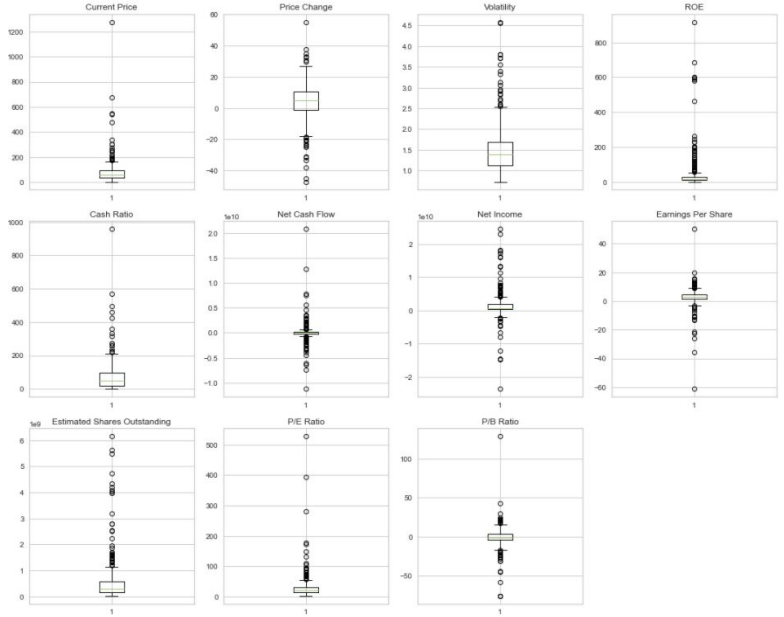
Exploratory Data Analysis (EDA)-Univariate Analysis

Bivariate Analysis



There is no high correlation between the different variables.
Net income and Estimated Shares Outstanding is positively correlated with 0.59.
Net Income and Earning Per Share is positively correlated with 0.56.
Earning Per Share and Current Price is positively correlated with 0.48.
There are some negatively correlated variables such as ROE and Earning Per Share -0.41 .

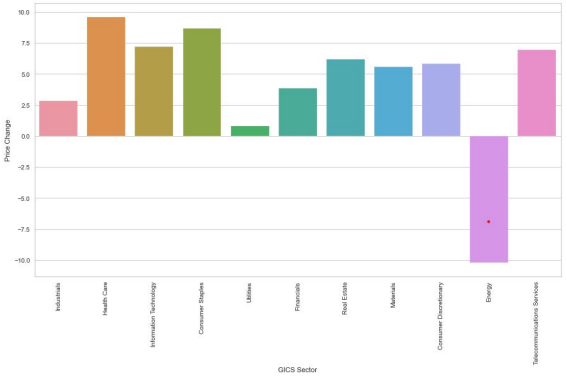
Outlier Check



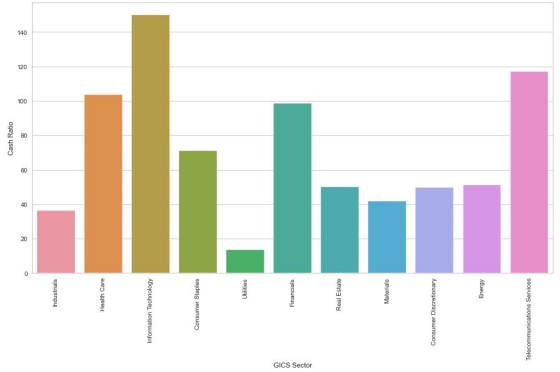
There are outliers ,but we are not treating the them since all are genuine values.

Exploratory Data Analysis (EDA) - continued

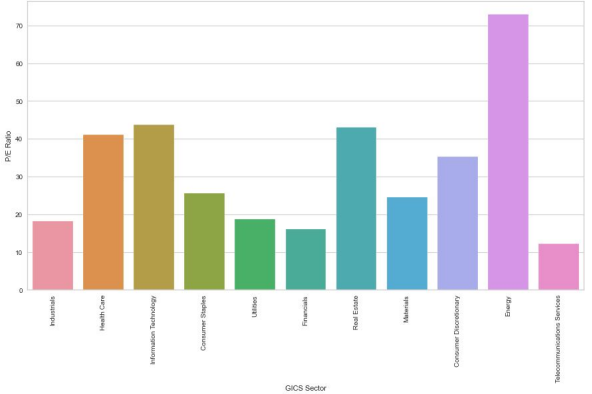
Let's check the stocks of which economic sector have seen the maximum price increase on average.



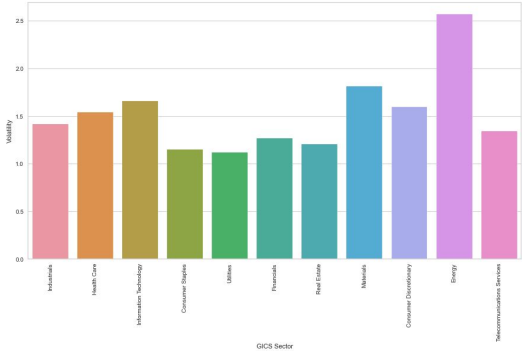
Cash ratio provides a measure of a company's ability to cover its short-term obligations using only cash and cash equivalents. Let's see how the average cash ratio varies across economic sectors.



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. Let's see how the P/E ratio varies, on average, across economic sectors.



Volatility accounts for the fluctuation in the stock price. A stock with high volatility will witness sharper price changes, making it a riskier investment. Let's see how volatility varies, on average, across economic sectors.



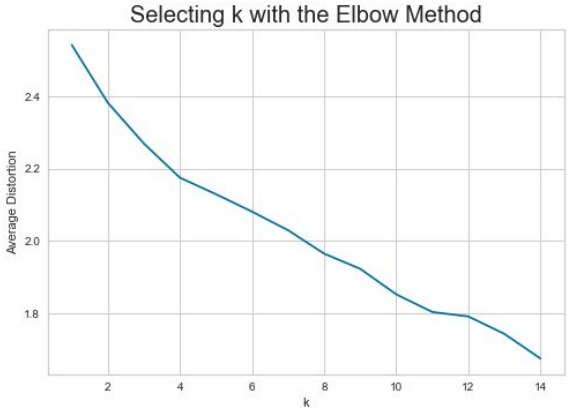
- Health care has highest positive price change following with Consumer Staples. Energy has the highest negative price change
- Energy, Materials and Information Technology has high Volatility which will witness sharper price changes. Investors have to be aware of high risks in this GICS Sectors.
- Cash ratio is high on Information Technology, Telecommunication, Health Care and Financials. This GICS Sectors will have an advantage on short-terms using only cash and cash equivalents.
- It is obvious that Energy Sector has the highest P/E Ratio. Information Technology, Real Estate and Health care are next highest P/E Ratio respectively. It means that they signify the amount of money an investor is willing to invest in a single share of a company per dollar of its earnings.

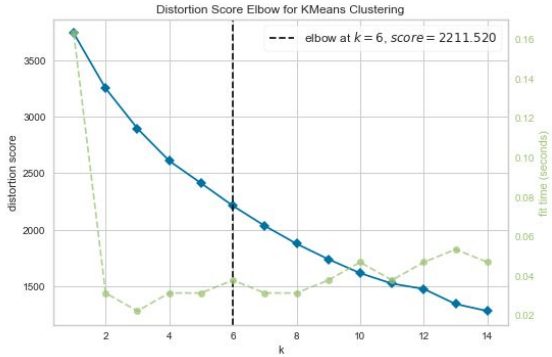
Scaling

```
scaler = StandardScaler()
subset = df[numeric_columns].copy()  ## Compl
subset_scaled = scaler.fit_transform(subset)
subset_scaled_df = pd.DataFrame(subset_scaled, columns=subset.columns)
```

K-means Clustering

```
Number of Clusters: 1   Average Distortion: 2.5425069919221697
Number of Clusters: 2   Average Distortion: 2.382318498894466
Number of Clusters: 3   Average Distortion: 2.2692367155390745
Number of Clusters: 4   Average Distortion: 2.1745559827866363
Number of Clusters: 5   Average Distortion: 2.128799332840716
Number of Clusters: 6   Average Distortion: 2.080400099226289
Number of Clusters: 7   Average Distortion: 2.0289794220177395
Number of Clusters: 8   Average Distortion: 1.964144163389972
Number of Clusters: 9   Average Distortion: 1.9221492045198068
Number of Clusters: 10  Average Distortion: 1.8513913649973124
Number of Clusters: 11  Average Distortion: 1.8024134734578485
Number of Clusters: 12  Average Distortion: 1.7900931879652673
Number of Clusters: 13  Average Distortion: 1.7417609203336912
Number of Clusters: 14  Average Distortion: 1.673559857259703
```

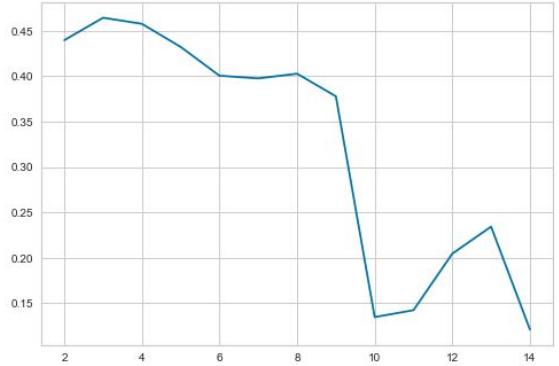


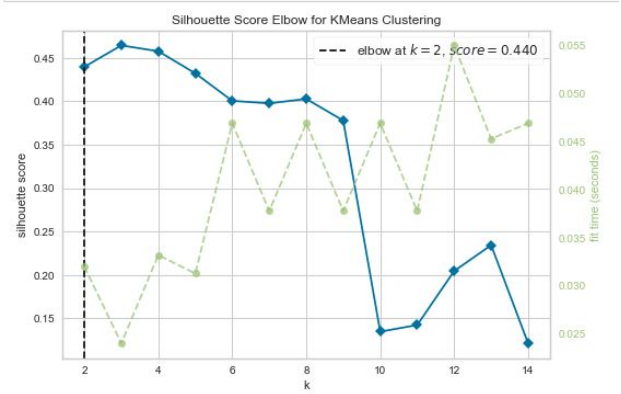


The appropriate value of k from the elbow curve seems to be 6.

Let's check the silhouette scores.

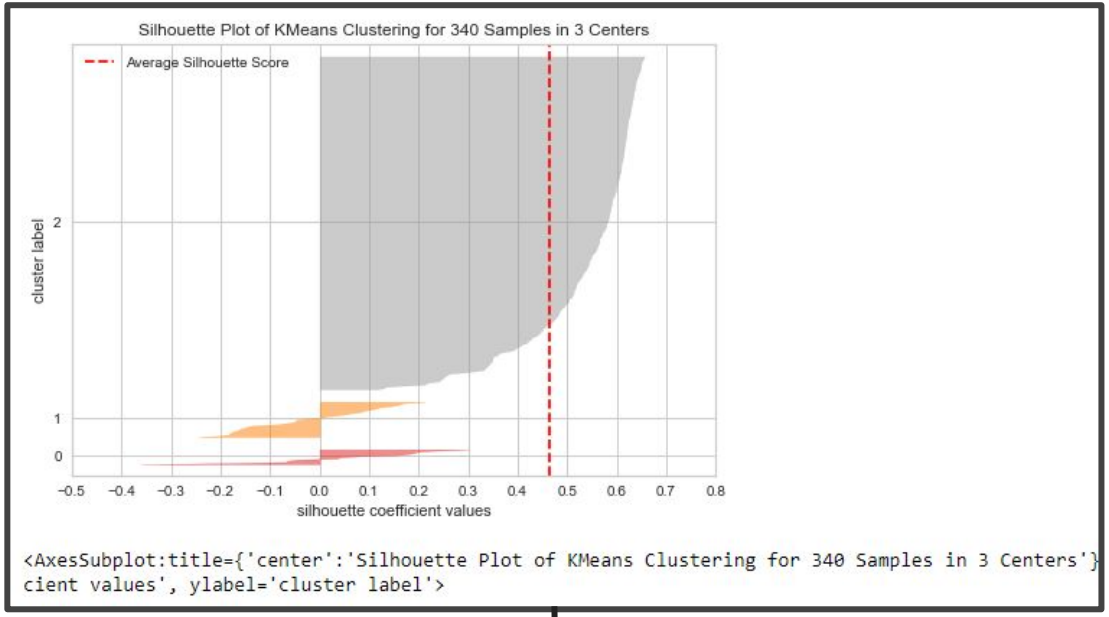
```
For n_clusters = 2, the silhouette score is 0.43969639509980457)
For n_clusters = 3, the silhouette score is 0.4644405674779404)
For n_clusters = 4, the silhouette score is 0.4577225970476733)
For n_clusters = 5, the silhouette score is 0.43228336443659804)
For n_clusters = 6, the silhouette score is 0.4005422737213617)
For n_clusters = 7, the silhouette score is 0.3976335364987305)
For n_clusters = 8, the silhouette score is 0.40278401969450467)
For n_clusters = 9, the silhouette score is 0.3778585981433699)
For n_clusters = 10, the silhouette score is 0.13458938329968687)
For n_clusters = 11, the silhouette score is 0.1421832155528444)
For n_clusters = 12, the silhouette score is 0.2044669621527429)
For n_clusters = 13, the silhouette score is 0.23424874810104204)
For n_clusters = 14, the silhouette score is 0.12102526472829901)
```





Silhouette score for 3 is higher than that for 6. So, we will choose 3 as value of k

Finding optimal number of clusters with silhouette coefficients




```
<AxesSubplot:title={'center':'Silhouette Plot of KMeans Clustering for 340 Samples in 3 Centers'}
cient values', ylabel='cluster label'>
```

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	52.142857	6.779993	1.175153	26.142857	140.142857	760285714.285714	13368785714.285715	3.769286	3838879870.871428	20.654832	-3.529270	14
1	64.183438	-10.557046	2.797776	96.531250	70.718750	159171125.000000	-3250005968.750000	-7.886875	526459323.057500	111.333230	1.783445	32
2	84.045331	5.542488	1.404255	34.040816	66.608844	10698350.340136	1445333183.673469	3.890051	427206184.715408	24.613743	-2.013147	294

We highlighted the highest values for each variables in KM segments.
Cluster 0 captured the highest values.

Companies in each cluster:

Cluster 2

```
In cluster 2, the following companies are present:
['American Airlines Group' 'AbbVie' 'Abbott Laboratories'
 'Adobe Systems Inc' 'Archer-Daniels-Midland Co' 'Alliance Data Systems'
 'Ameren Corp' 'American Electric Power' 'AFLAC Inc'
 'American International Group, Inc.' 'Apartment Investment & Mgmt'
 'Assurant Inc' 'Arthur J. Gallagher & Co.' 'Akamai Technologies Inc'
 'Albemarle Corp' 'Alaska Air Group Inc' 'Allstate Corp' 'Allegion'
 'Applied Materials Inc' 'AMETEK Inc' 'Affiliated Managers Group Inc'
 'Amgen Inc' 'Ameriprise Financial' 'American Tower Corp A'
 'AutoNation Inc' 'Anthem Inc.' 'Aon plc' 'Amphenol Corp' 'Arconic Inc'
 'Activision Blizzard' 'AvalonBay Communities, Inc.' 'Broadcom'
 'American Water Works Company Inc' 'American Express Co' 'Boeing Company'
 'Baxter International Inc.' 'BB&T Corporation' 'Bard (C.R.) Inc.'
 'BIOGEN IDEC Inc.' 'The Bank of New York Mellon Corp.' 'Ball Corp'
 'Bristol-Myers Squibb' 'Boston Scientific' 'Borgwarner'
 'Boston Properties' 'Caterpillar Inc.' 'Chubb Limited' 'CBRE Group'
 'Crown Castle International Corp.' 'Carnival Corp.' 'Celgene Corp.'
 'CF Industries Holdings Inc' 'Citizens Financial Group' 'Church & Dwight'
 'C. H. Robinson Worldwide' 'Charter Communications' 'CIGNA Corp.'
 'Cincinnati Financial' 'Colgate-Palmolive' 'Comerica Inc.'
 'CME Group Inc.' 'Chipotle Mexican Grill' 'Cummins Inc.' 'CMS Energy'
 'Centene Corporation' 'CenterPoint Energy' 'Capital One Financial'
 'The Cooper Companies' 'CSX Corp.' 'CenturyLink Inc'
 'Cognizant Technology Solutions' 'Citrix Systems' 'CVS Health'
 'Chevron Corp.' 'Dominion Resources' 'Delta Air Lines' 'Du Pont (E.I.)'
 'Deere & Co.' 'Discover Financial Services' 'Quest Diagnostics'
 'Danaher Corp.' 'The Walt Disney Company' 'Discovery Communications-A'
 'Discovery Communications-C' 'Delphi Automotive' 'Digital Realty Trust'
 'Dun & Bradstreet' 'Dover Corp.' 'Dr Pepper Snapple Group' 'Duke Energy'
 'DaVita Inc.' 'eBay Inc.' 'Ecolab Inc.' 'Consolidated Edison'
 'Equifax Inc.' "Edison Int'l" 'Eastman Chemical' 'Equinix'
 'Equity Residential' 'Eversource Energy' 'Essex Property Trust, Inc.'
 "E*Trade' 'Eaton Corporation' 'Entergy Corp.' 'Edwards Lifesciences'
 'Exelon Corp.' 'Expeditors Int'l' 'Expedia Inc.' 'Extra Space Storage'
 'Fastenal Co' 'Fortune Brands Home & Security' 'FirstEnergy Corp'
 'Fidelity National Information Services' 'Fiserv Inc' 'FLIR Systems'
 'Fluor Corp.' 'Flowserve Corporation' 'FMC Corporation'
 'Federal Realty Investment Trust' 'First Solar Inc'
 'Frontier Communications' 'General Dynamics'
 'General Growth Properties Inc.' 'Corning Inc.' 'General Motors'
 'Genuine Parts' 'Garmin Ltd.' 'Goodyear Tire & Rubber'
 'Grainger (W.W.) Inc.' 'Hasbro Inc.' 'Huntington Bancshares'
 'HCA Holdings' 'Welltower Inc.' 'HCP Inc.' 'Hartford Financial Svc.Gp.'
 'Harley-Davidson' "Honeywell Int'l Inc." 'HP Inc.' 'Hormel Foods Corp.'
 'Henry Schein' 'Host Hotels & Resorts' 'The Hershey Company'
 'Humana Inc.' 'International Business Machines' 'IDEXX Laboratories'
 'Intl Flavors & Fragrances' 'International Paper' 'Interpublic Group'
 'Iron Mountain Incorporated' 'Intuitive Surgical Inc.'
 'Illinois Tool Works' 'Invesco Ltd.' 'J. B. Hunt Transport Services'
 'Jacobs Engineering Group' 'Juniper Networks' 'Kimco Realty'
 'Kimberly-Clark' 'Kansas City Southern' 'Leggett & Platt' 'Lennar Corp.'
 'Laboratory Corp. of America Holding' 'LKQ Corporation'
 'L-3 Communications Holdings' 'Lilly (Eli) & Co.' 'Lockheed Martin Corp.'
 'Alliant Energy Corp' 'Leucadia National Corp.' 'Southwest Airlines'
 'Level 3 Communications' 'LyondellBasell' 'Mastercard Inc.'
 'Mid-America Apartments' 'Macerich' "Marriott Int'l." 'Masco Corp.'
 'Mattel Inc.' "McDonald's Corp." "Moody's Corp" 'Mondelez International'
 'MetLife Inc.' 'Mohawk Industries' 'Mead Johnson' 'McCormick & Co.'
 'Martin Marietta Materials' 'Marsh & McLennan' '3M Company'
 'Monster Beverage' 'Altria Group Inc' 'The Mosaic Company'
 'Marathon Petroleum' 'M&T Bank Corp.' 'Mettler Toledo' 'Mylan N.V.'
 'Navient' 'NASDAQ OMX Group' 'Nextera Energy'
 'Newmont Mining Corp. (Hldg. Co.)' 'Nielsen Holdings'
 'Norfolk Southern Corp.' 'Northern Trust Corp.' 'Nucor Corp.'
 'Newell Brands' 'Realty Income Corporation' 'Omnicom Group'
 "O'Reilly Automotive" "People's United Financial" 'Pitney-Bowes'
 'PACCAR Inc.' 'PG&E Corp.' 'Priceline.com Inc'
 'Public Serv. Enterprise Inc.' 'PepsiCo Inc.' 'Principal Financial Group'
 'Procter & Gamble' 'Progressive Corp.' 'Pulte Homes Inc.'
 'Philip Morris International' 'PNC Financial Services' 'Pentair Ltd.'
 'Pinnacle West Capital' 'PPG Industries' 'PPL Corp.'
 'Prudential Financial' 'Phillips 66' 'Praxair Inc.' 'PayPal'
 'Ryder System' 'Royal Caribbean Cruises Ltd' 'Regeneron'
 'Robert Half International' 'Roper Industries' 'Republic Services Inc'
 'SCANA Corp' 'Charles Schwab Corporation' 'Spectra Energy Corp.'
 'Sealed Air' 'Sherwin-Williams' 'SL Green Realty'
 'Scripps Networks Interactive Inc.' 'Southern Co.'
 'Simon Property Group Inc' 'S&P Global, Inc.' 'Stericycle Inc'
 'Sempra Energy' 'SunTrust Banks' 'State Street Corp.'
 'Skyworks Solutions' 'Synchrony Financial' 'Stryker Corp.'
 'Molson Coors Brewing Company' 'Tegna, Inc.' 'Torchmark Corp.'
 'Thermo Fisher Scientific' 'TripAdvisor' 'The Travelers Companies Inc.'
 'Tractor Supply Company' 'Tyson Foods' 'Tesoro Petroleum Co.'
 'Total System Services' 'Texas Instruments' 'Under Armour'
 'United Continental Holdings' 'UDR Inc' 'Universal Health Services, Inc.'
 'United Health Group Inc.' 'Unum Group' 'Union Pacific'
 'United Parcel Service' 'United Technologies' 'Varian Medical Systems'
 'Valero Energy' 'Vulcan Materials' 'Vornado Realty Trust'
 'Verisk Analytics' 'Verisign Inc.' 'Vertex Pharmaceuticals Inc'
 'Ventas Inc' 'Waters Corporation' 'Wec Energy Group Inc'
 'Whirlpool Corp.' 'Waste Management Inc.' 'Western Union Co'
 'Weyerhaeuser Corp.' 'Wyndham Worldwide' 'Xcel Energy Inc' 'XL Capital'
 'Dentsply Sirona' 'Xerox Corp.' 'Xylem Inc.' 'Yahoo Inc.'
 'Yum! Brands Inc' 'Zimmer Biomet Holdings' 'Zions Bancorp' 'Zoetis']
```

Cluster 1

```
In cluster 1, the following companies are present:
['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.' 'Southwestern Energy'
 'Teradata Corp.' 'Williams Cos.' 'Wynn Resorts Ltd' 'Cimarex Energy']
```

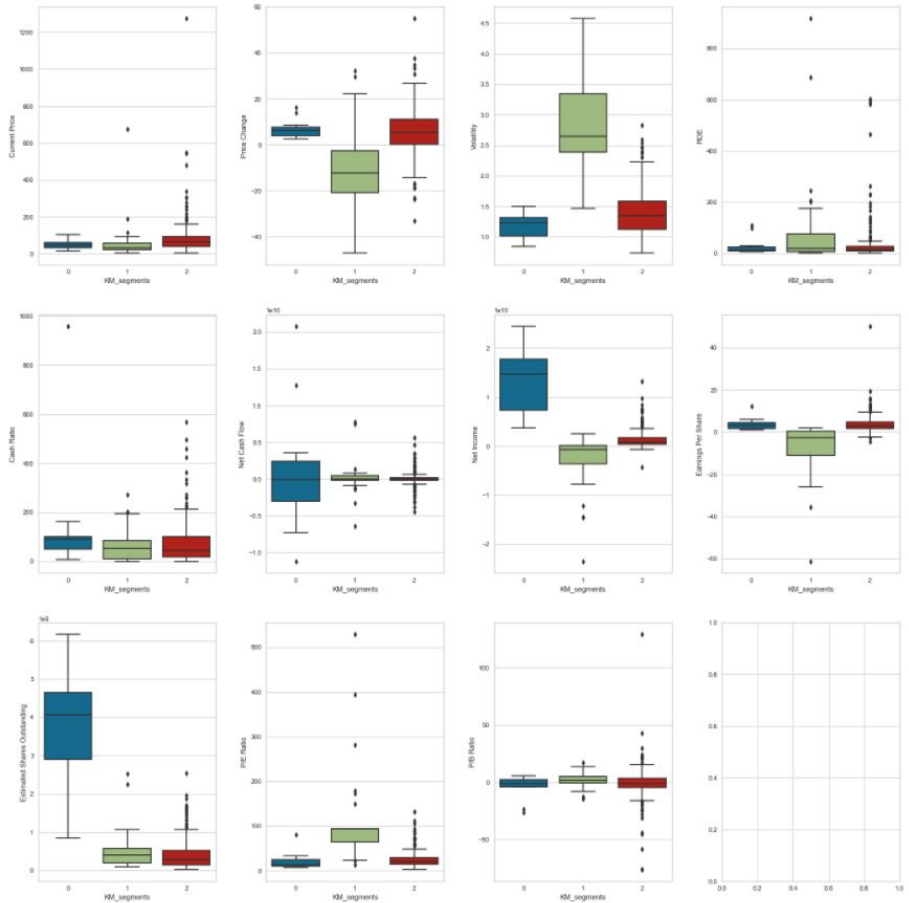
Cluster 0

```
In cluster 0, the following companies are present:
['Bank of America Corp' 'Citigroup Inc.' 'Ford Motor' 'Facebook'
 'Gilead Sciences' 'Intel Corp.' 'JPMorgan Chase & Co.'
 'Coca Cola Company' 'Merck & Co.' 'Pfizer Inc.' 'AT&T Inc'
 'Verizon Communications' 'Wells Fargo' 'Exxon Mobil Corp.']
```

KM_segments	GICS Sector	
0	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	4
	Health Care	3
	Information Technology	2
	Telecommunications Services	2
1	Consumer Discretionary	2
	Energy	23
	Health Care	1
	Industrials	1
	Information Technology	4
	Materials	1
2	Consumer Discretionary	37
	Consumer Staples	18
	Energy	6
	Financials	45
	Health Care	36
	Industrials	52
	Information Technology	27
	Materials	19
	Real Estate	27
	Telecommunications Services	3
	Utilities	24

Name: Security, dtype: int64

As we see on the left Cluster 2 holds most of the companies.



- Price Change, Cash Ratio, Net Income, Net Cash Flow, Estimated Share Outstanding is high in **Cluster 0.**
- Vololity, ROE, P/E Ratio and P/B Ratio is high in **Cluster 1.**
- Current Price and Earning Per Share is high in **Cluster 2.** It also has most of the count in each segment.

Hierarchical Clustering

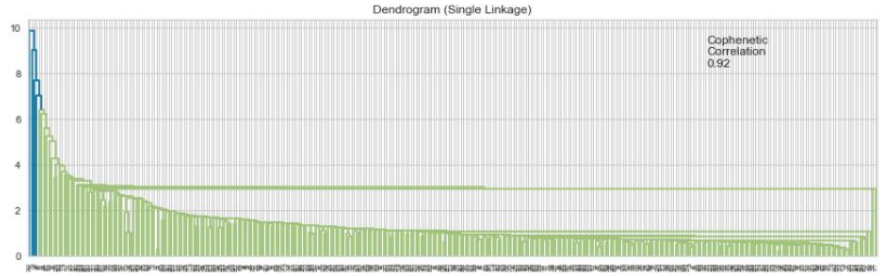
```
Cophenetic correlation for Euclidean distance and single linkage is 0.9232271494002922.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7873280186580672.
Cophenetic correlation for Euclidean distance and average linkage is 0.9422540609560814.
Cophenetic correlation for Euclidean 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 0.792530720285.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159737.
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.
```

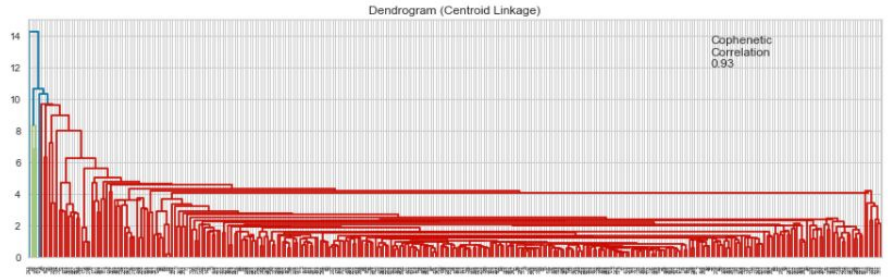
We will use Euclidean distance and average linkage for the Hierarchical Clustering.

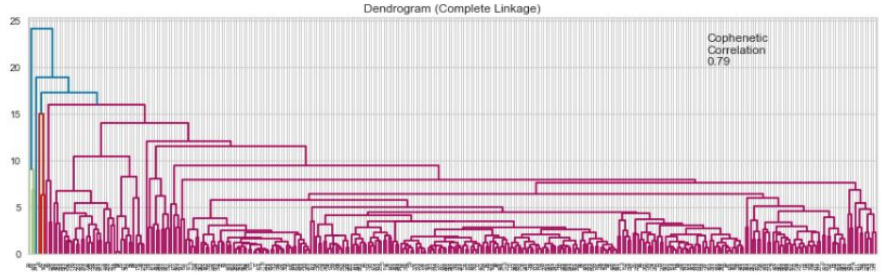
Exploring different linkage methods with Euclidean distance only.

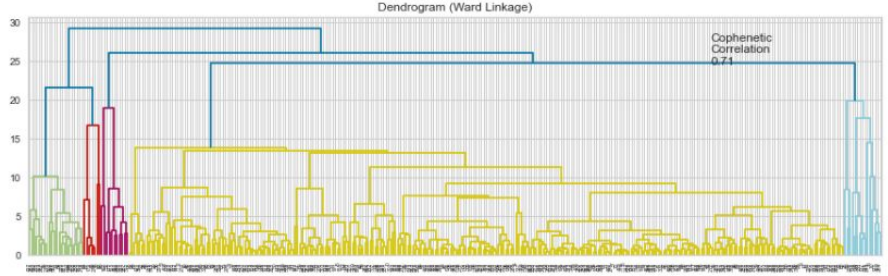
```
Cophenetic correlation for single linkage is 0.9232271494002922.
Cophenetic correlation for complete linkage is 0.7873280186580672.
Cophenetic correlation for average linkage is 0.9422540609560814.
Cophenetic correlation for centroid linkage is 0.9314012446828154.
Cophenetic correlation for ward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

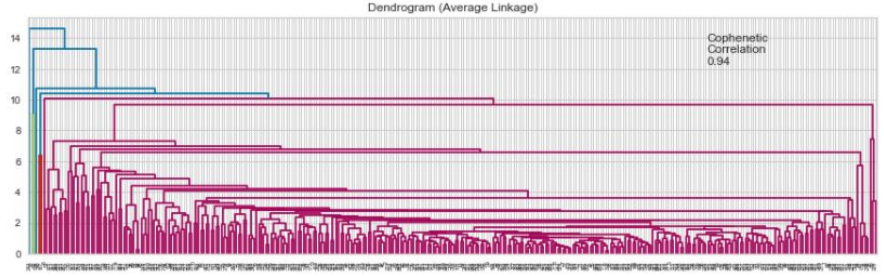
Let's view the dendrograms for the different linkage methods with Euclidean distance

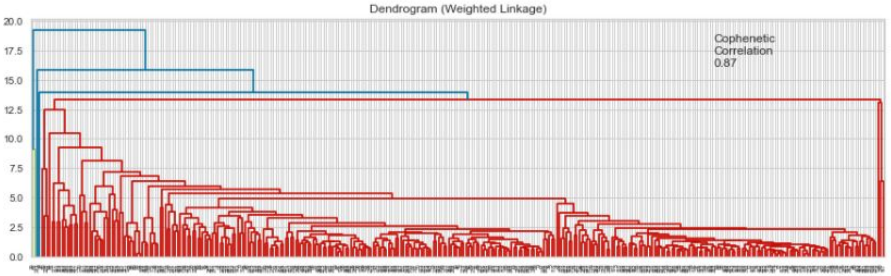












We will choose Average Linkage with 6 clusters.

Comparing cophenetic correlations for different linkage methods

	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

Cluster Profiling

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	77.287589	4.099730	1.518066	35.336336	66.900901	-33197321.321321	1538074666.666667	2.885270	560505037.293543	32.441706	-2.174921	333
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
5	276.570007	6.189286	1.116976	30.000000	25.000000	90885000.000000	596541000.000000	8.910000	66951851.850000	31.040405	129.064585	1

The Companies in each cluster

In cluster 0, the following companies are present:
['American Airlines Group' 'AbbVie' 'Abbott Laboratories'
'Adobe Systems Inc' 'Analog Devices, Inc.' 'Archer-Daniels-Midland Co'
'Ameren Corp' 'American Electric Power' 'AFLAC Inc'
'American International Group, Inc.' 'Apartment Investment & Mgmt'
'Assurant Inc' 'Arthur J. Gallagher & Co.' 'Akamai Technologies Inc'
'Albemarle Corp' 'Alaska Air Group Inc' 'Allstate Corp' 'Allegion'
'Alexion Pharmaceuticals' 'Applied Materials Inc' 'AMETEK Inc'
'Affiliated Managers Group Inc' 'Amgen Inc' 'Ameriprise Financial'
'American Tower Corp A' 'Amazon.com Inc' 'AutoNation Inc' 'Anthem Inc.'
'Aon plc' 'Anadarko Petroleum Corp' 'Amphenol Corp' 'Arconic Inc'
'Activision Blizzard' 'AvalonBay Communities, Inc.' 'Broadcom'
'American Water Works Company Inc' 'American Express Co' 'Boeing Company'
'Baxter International Inc.' 'BB&T Corporation' 'Bard (C.R.) Inc.'
'Baker Hughes Inc' 'BIOGEN IDEC Inc.' 'The Bank of New York Mellon Corp.'
'Ball Corp' 'Bristol-Myers Squibb' 'Boston Scientific' 'BorgWarner'
'Boston Properties' 'Citigroup Inc.' 'Caterpillar Inc.' 'Chubb Limited'
'CBRE Group' 'Crown Castle International Corp.' 'Carnival Corp.'
'Celgene Corp.' 'CF Industries Holdings Inc' 'Citizens Financial Group'
'Church & Dwight' 'C. H. Robinson Worldwide' 'Charter Communications'
'CIGNA Corp.' 'Cincinnati Financial' 'Colgate-Palmolive' 'Comerica Inc.'
'CME Group Inc.' 'Chipotle Mexican Grill' 'Cummins Inc.' 'CMS Energy'
'Centene Corporation' 'CenterPoint Energy' 'Capital One Financial'
'Cabot Oil & Gas' 'The Cooper Companies' 'CSX Corp.' 'CenturyLink Inc'
'Cognizant Technology Solutions' 'Citrix Systems' 'CVS Health'
'Chevron Corp.' 'Concho Resources' 'Dominion Resources' 'Delta Air Lines'
'Du Pont (E.I.)' 'Deere & Co.' 'Discover Financial Services'
'Quest Diagnostics' 'Danaher Corp.' 'The Walt Disney Company'
'Discovery Communications-A' 'Discovery Communications-C'
'Delphi Automotive' 'Digital Realty Trust' 'Dun & Bradstreet'
'Dover Corp.' 'Dr Pepper Snapple Group' 'Duke Energy' 'DaVita Inc.'
'Devon Energy Corp.' 'eBay Inc.' 'Ecolab Inc.' 'Consolidated Edison'
'Equifax Inc.' "Edison Int'l" 'Eastman Chemical' 'EOG Resources'
'Equinix' 'Equity Residential' 'EQT Corporation' 'Eversource Energy'
'Essex Property Trust, Inc.' 'E*Trade' 'Eaton Corporation'
'Entergy Corp.' 'Edwards Lifesciences' 'Exelon Corp.' "Expeditors Int'l"
'Expedia Inc.' 'Extra Space Storage' 'Ford Motor' 'Fastenal Co'
'Fortune Brands Home & Security' 'Freeport-McMoran Cp & Gld'
'FirstEnergy Corp' 'Fidelity National Information Services' 'Fiserv Inc'
'FLIR Systems' 'Fluor Corp.' 'Flowserve Corporation' 'FMC Corporation'
'Federal Realty Investment Trust' 'First Solar Inc'
'Frontier Communications' 'General Dynamics'
'General Growth Properties Inc.' 'Gilead Sciences' 'Corning Inc.'
'General Motors' 'Genuine Parts' 'Garmin Ltd.' 'Goodyear Tire & Rubber'
'Grainger (W.W.) Inc.' 'Halliburton Co.' 'Hasbro Inc.'
'Huntington Bancshares' 'HCA Holdings' 'Welltower Inc.' 'HCP Inc.'
'Hess Corporation' 'Hartford Financial Svc.Gp.' 'Harley-Davidson'
"Honeywell Int'l Inc." 'Hewlett Packard Enterprise' 'HP Inc.'
'Hormel Foods Corp.' 'Henry Schein' 'Host Hotels & Resorts'
'The Hershey Company' 'Humana Inc.' 'International Business Machines'
'IDEXX Laboratories' 'Intl Flavors & Fragrances' 'International Paper'
'Interpublic Group' 'Iron Mountain Incorporated'
'Intuitive Surgical Inc.' 'Illinois Tool Works' 'Invesco Ltd.'
'J. B. Hunt Transport Services' 'Jacobs Engineering Group'
'Juniper Networks' 'JPMorgan Chase & Co.' 'Kimco Realty' 'Kimberly-Clark'
'Kinder Morgan' 'Coca Cola Company' 'Kansas City Southern'
'Leggett & Platt' 'Lennar Corp.' 'Laboratory Corp. of America Holding'
'LKQ Corporation' 'L-3 Communications Holdings' 'Lilly (Eli) & Co.'
'Lockheed Martin Corp.' 'Alliant Energy Corp' 'Leucadia National Corp.'
'Southwest Airlines' 'Level 3 Communications' 'LyondellBasell'
'Mastercard Inc.' 'Mid-America Apartments' 'Macerich' "Marriott Int'l."
'Masco Corp.' 'Mattel Inc.' "McDonald's Corp." "Moody's Corp"
'Mondelez International' 'MetLife Inc.' 'Mohawk Industries'
'Mead Johnson' 'McCormick & Co.' 'Martin Marietta Materials'
'Marsh & McLennan' '3M Company' 'Monster Beverage' 'Altria Group Inc'
'The Mosaic Company' 'Marathon Petroleum' 'Merck & Co.'
'Marathon Oil Corp.' 'M&T Bank Corp.' 'Mettler Toledo' 'Murphy Oil'
'Mylan N.V.' 'Navient' 'Noble Energy Inc' 'NASDAQ OMX Group'
'NextEra Energy' 'Newmont Mining Corp. (Hldg. Co.)' 'Netflix Inc.'
'Newfield Exploration Co' 'Nielsen Holdings'
'National Oilwell Varco Inc.' 'Norfolk Southern Corp.'
'Northern Trust Corp.' 'Nucor Corp.' 'Newell Brands'
'Realty Income Corporation' 'ONEOK' 'Omnicom Group' "O'Reilly Automotive"
'Occidental Petroleum' "People's United Financial" 'Pitney-Bowes'
'PACCAR Inc.' 'PG&E Corp.' 'Public Serv. Enterprise Inc.' 'PepsiCo Inc.'
'Pfizer Inc.' 'Principal Financial Group' 'Procter & Gamble'
'Progressive Corp.' 'Pulte Homes Inc.' 'Philip Morris International'
'PNC Financial Services' 'Pentair Ltd.' 'Pinnacle West Capital'
'PPG Industries' 'PPL Corp.' 'Prudential Financial' 'Phillips 66'
'Quanta Services Inc.' 'Praxair Inc.' 'PayPal' 'Ryder System'
'Royal Caribbean Cruises Ltd' 'Regeneron' 'Robert Half International'
'Roper Industries' 'Range Resources Corp.' 'Republic Services Inc'
'SCANA Corp' 'Charles Schwab Corporation' 'Spectra Energy Corp.'
'Sealed Air' 'Sherwin-Williams' 'SL Green Realty'
'Scripps Networks Interactive Inc.' 'Southern Co.'
'Simon Property Group Inc' 'S&P Global, Inc.' 'Stericycle Inc'
'Sempra Energy' 'SunTrust Banks' 'State Street Corp.'
'Skyworks Solutions' 'Southwestern Energy' 'Synchrony Financial'
'Stryker Corp.' 'AT&T Inc' 'Molson Coors Brewing Company'
'Teradata Corp.' 'Tegna, Inc.' 'Torchmark Corp.'
'Thermo Fisher Scientific' 'TripAdvisor' 'The Travelers Companies In
'Tractor Supply Company' 'Tyson Foods' 'Tesoro Petroleum Co.'
'Total System Services' 'Texas Instruments' 'Under Armour'
'United Continental Holdings' 'UDR Inc' 'Universal Health Services,
'United Health Group Inc.' 'Unum Group' 'Union Pacific'
'United Parcel Service' 'United Technologies' 'Varian Medical Systems'
'Valero Energy' 'Vulcan Materials' 'Vornado Realty Trust'
'Verisk Analytics' 'Verisign Inc.' 'Vertex Pharmaceuticals Inc'
'Ventas Inc' 'Verizon Communications' 'Waters Corporation'
'Wec Energy Group Inc' 'Wells Fargo' 'Whirlpool Corp.'
'Waste Management Inc.' 'Williams Cos.' 'Western Union Co'
'Weyerhaeuser Corp.' 'Wyndham Worldwide' 'Wynn Resorts Ltd'
'Cimarex Energy' 'Xcel Energy Inc' 'XL Capital' 'Exxon Mobil Corp.'
'Dentsply Sirona' 'Xerox Corp.' 'Xylem Inc.' 'Yahoo Inc.'
'Yum! Brands Inc' 'Zimmer Biomet Holdings' 'Zions Bancorp' 'Zoetis']

In cluster 5, the following companies are present:
['Alliance Data Systems']

In cluster 2, the following companies are present:
['Apache Corporation' 'Chesapeake Energy']

In cluster 1, the following companies are present:
['Bank of America Corp' 'Intel Corp.']

In cluster 3, the following companies are present:
['Facebook']

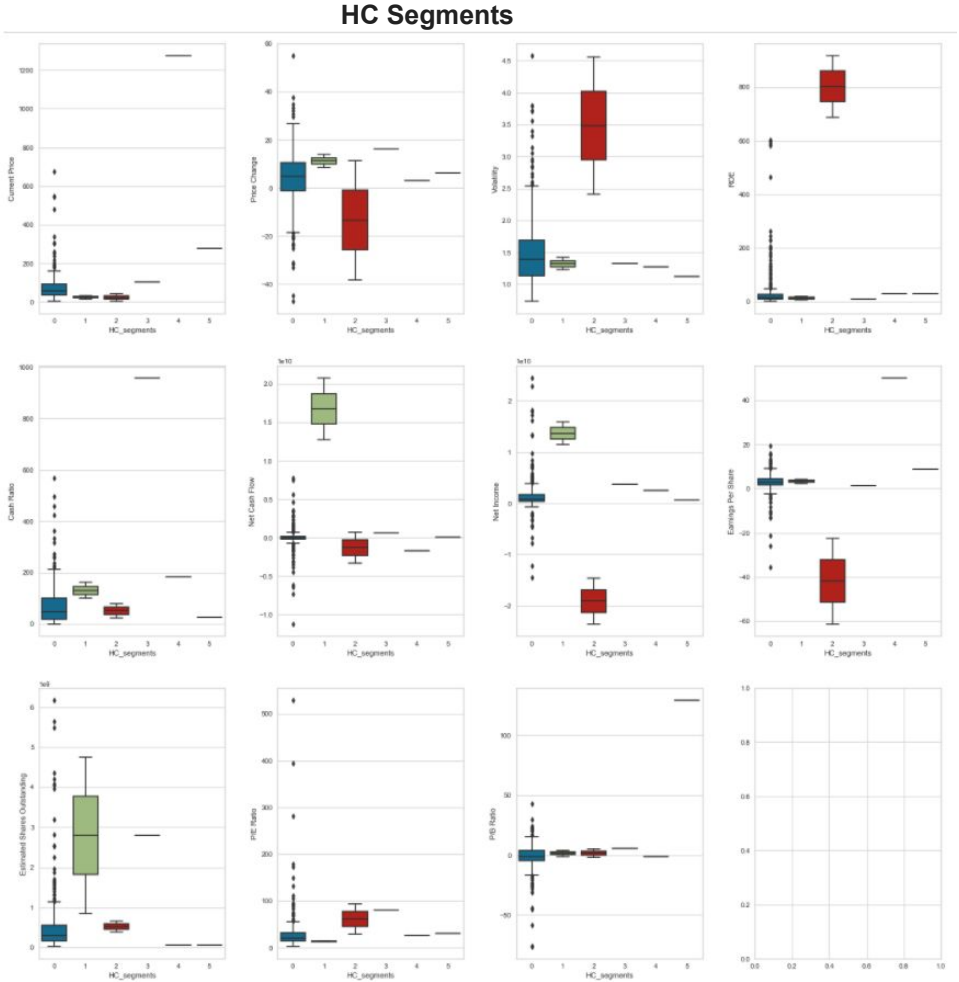
In cluster 4, the following companies are present:
['Priceline.com Inc']

HC Segments and GICS Sectors grouped by Security

HC_segments	GICS Sector	
0	Consumer Discretionary	39
	Consumer Staples	19
	Energy	28
	Financials	48
	Health Care	40
	Industrials	53
	Information Technology	30
	Materials	20
	Real Estate	27
	Telecommunications Services	5
	Utilities	24
1	Financials	1
	Information Technology	1
2	Energy	2
3	Information Technology	1
4	Consumer Discretionary	1
5	Information Technology	1

Name: Security, dtype: int64

- Cluster 0 has the highest count in each segment.
- Net Cash flow and Net Income has the highest value in Cluster 1.
- Volatility and ROE is highest in Cluster 2.
- Price Change, Cash Ratio, Estimated Share Outstanding, P/E Ratio is highest in Cluster 3.
- Current Price and Earning Per Share is highest in Cluster 4.
- P/B Ratio is the highest in Cluster 5.



K-means vs Hierarchical Clustering

Questions:

- Which clustering technique took less time for execution?
- Which clustering technique gave you more distinct clusters, or are they the same?
- How many observations are there in the similar clusters of both algorithms?
- How many clusters are obtained as the appropriate number of clusters from both algorithms?

Answers:

- K-Means took less time since I have to choose the k value.
- K-Means has more distinct clusters. Hierarchical Clustering has more clusters however there are 1 or 2 different Ticker Symbol.
- Hierarchical Clustering Cluster 2(2 Observations) and K-Means Cluster 1(32 Observations) is similar due to Volatility and ROE is highest in both. Current Price and Earning Per Share is highest in Cluster 4(Hierarchical Clustering 2 observations) and Cluster 2(K-Means 294 observations).
- K-Means algorithm has 3 clusters, and Hierarchical Clustering has 6 clusters.

Recommendations

- Health care has highest positive price change following with Consumer Staples. Energy has the highest negative price change. Customers should invest when their price go down to get more benefit. They will earn more in a long term.

- Energy, Materials and Information Technology has high Volatility which will witness sharper price changes. Investors have to be aware of high risks in this GICS Sectors and should not be investing on a single sector.

- Cash ratio is high on Information Technology, Telecommunication, Health Care and Financials. This GICS Sectors will have an advantage on short-terms using only cash and cash equivalents. This sectors will reduce the risk of losing money when a problem affects stock market.

- It is obvious that Energy Sector has the highest P/E Ratio. Information Technology, Real Estate and Health care are next highest P/E Ratio respectively. It means that they signify the amount of money an investor is willing to invest in a single share of a company per dollar of its earnings.