TRADE&AHEAD CONSULTING FIRM

Personalized Financial Investment Strategies
Unsupervised Learning: Module 7 - May 2024

AGENDA

❑ Executive Summary

❑ Objectives & Approach

❑ Data Definitions & Details

❑ EDA Results

❑ Data Preprocessing

❑ K-Means Clustering Summary & Techniques

❑ Hierarchical Clustering Summary & Techniques

❑ K-means vs. Hierarchical Clustering Questions

❑ Actionable Insights & Recommendations

Executive Summary

Investing in the stock market is one way to help meet your financial expectations. Getting a return on investment overtime with a diversified portfolio has proven to be the best combination. Trade&Ahead prides itself in analyzing stocks that maximize earnings under all market conditions by suggesting a blend of stocks that tend to yield a higher return and are lower risk of financial losses when the market is down.

Individuals tend to get overwhelmed by all of the financial metrics involved with trying to diversify their portfolios. Trade&Ahead is here to offer analysis and a sound mind decision for our clients to help protect against undo risks and losses in the future.

Objectives

1. Perform cluster analysis that can identify stocks that show similar characteristics and ones that exhibit minimum correlation

2. Analyze the given data comprised of stock price and some financial indicators for companies listed in the NYSE, and group them based on data definitions provided.

3. Share insights about the characteristics of each group with the client.

Approach

Our Trade&Ahead analysts will clean and verify the given data to make sure everything is in order for analysis. EDA analysis will provide a better picture of the information in an easy to understand manor. K-Means & Hierarchical Cluster Techniques will be used on the data and a summary of each technique individually will be provided as well as a summary of K-Means versus Hierarchical, to get the best possible result for client portfolios.

Data Definitions:

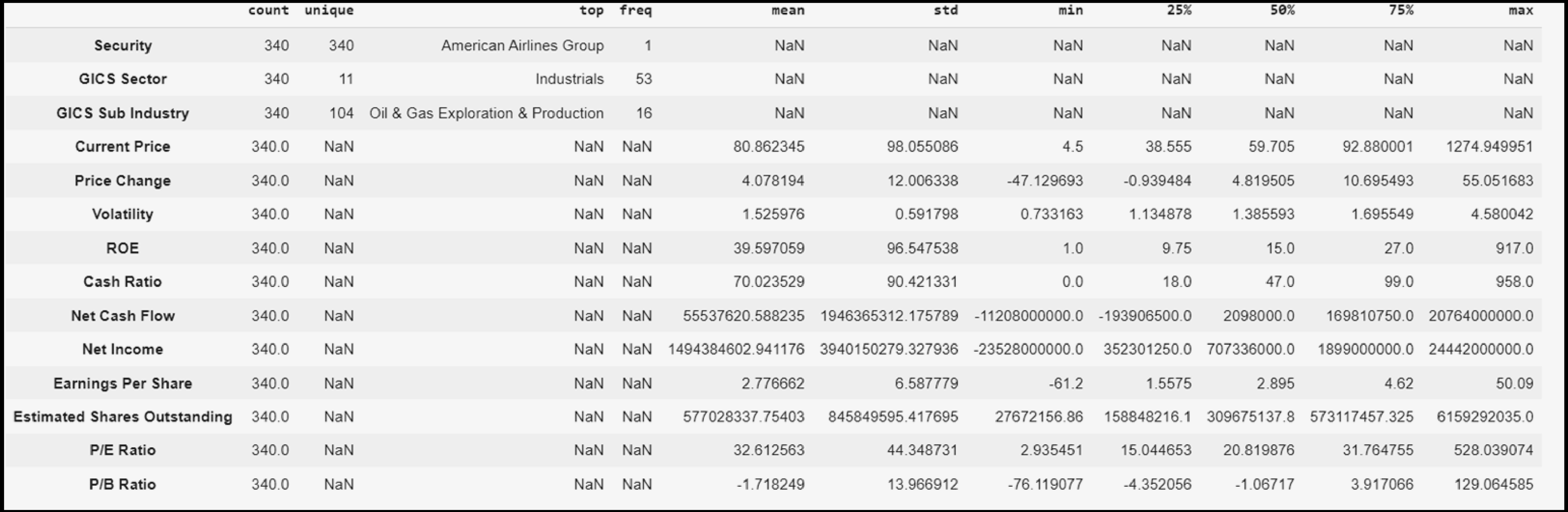
- **Ticker Symbol:** An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
- **Security:** Name of the company
- **GICS Sector:** The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that 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 is 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)

Dataset Details:

- Size: 340 Rows & 15 Columns
- Criteria: Analyze the data - Group by attributes - Share Characteristics & Insights
- EDA Analysis: Univariate & Bivariate
- Missing or Duplicate Values: No Missing or Duplicate Values within the Dataset
- Outlier Treatment/Feature Scaling: There were Outliers, but I did not treat them, the Histogram showed almost normal distribution
- K-means Clustering: Multiple Elbow Plots - Distortion Score, Silhouette Scores, & Silhouette Plot - Cluster Profiling
- Hierarchical Clustering: Cophenetic Correlation - Linkage Methods with Euclidean Distance - Dendograms - sklearn Model - Cluster Profiling
- K-means vs. Hierarchical Comparison: Clustering Techniques - Algorithms - Execution - Differences & Similarities

EDA Results

- Security - GICS Sector - GICS Sub Industry are Categorical, all other columns are Numeric
- The Statistical Summary of the data shows there will be many Outliers within the Numeric Columns
- The numeric variables will be scaled before Clustering



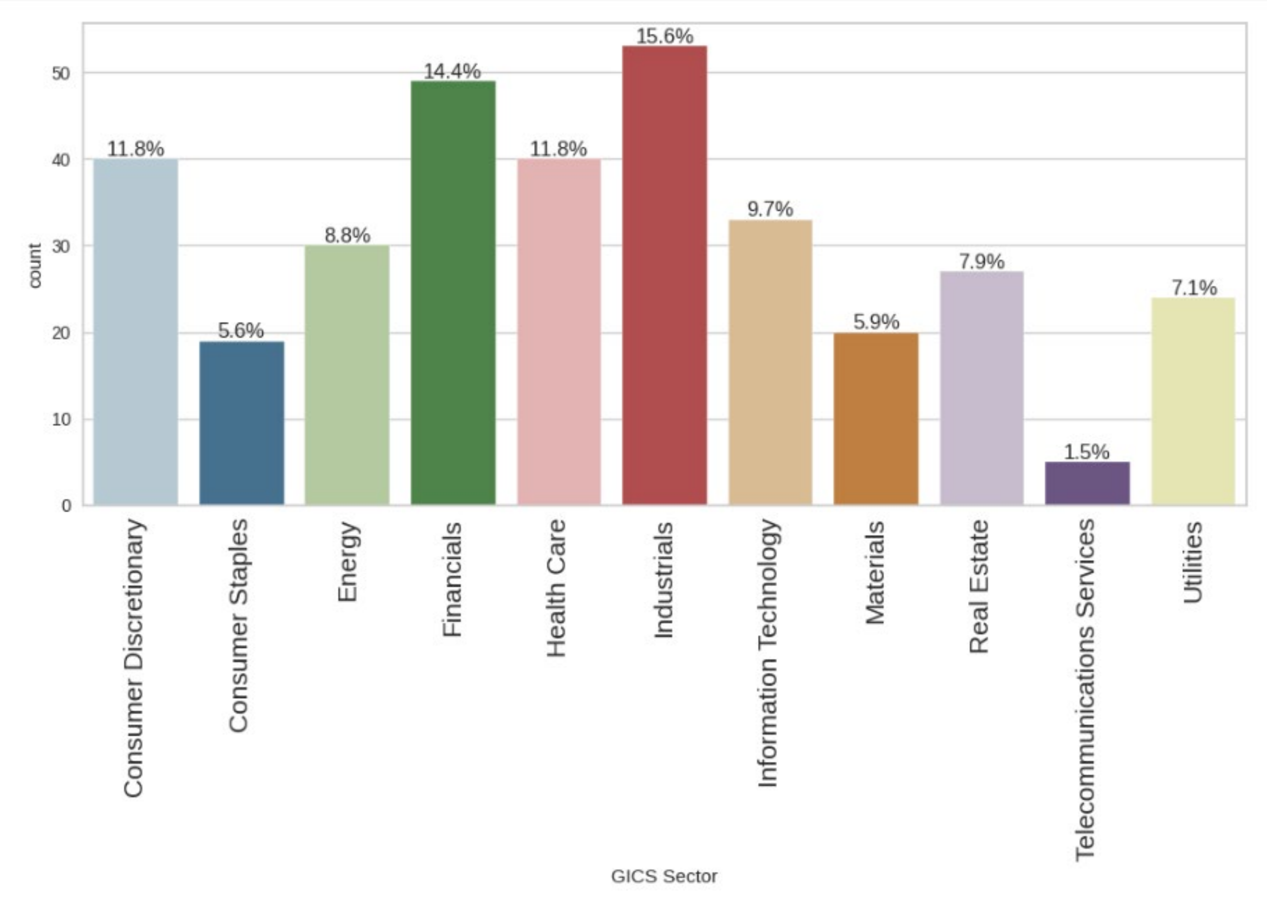
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Security	340	340	American Airlines Group	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sector	340	11	Industrials	53	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sub Industry	340	104	Oil & Gas Exploration & Production	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Current Price	340.0	NaN	NaN	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	3940150279.327936	-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	577028337.75403	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

EDA Results

UNIVARIATE ANALYSIS:

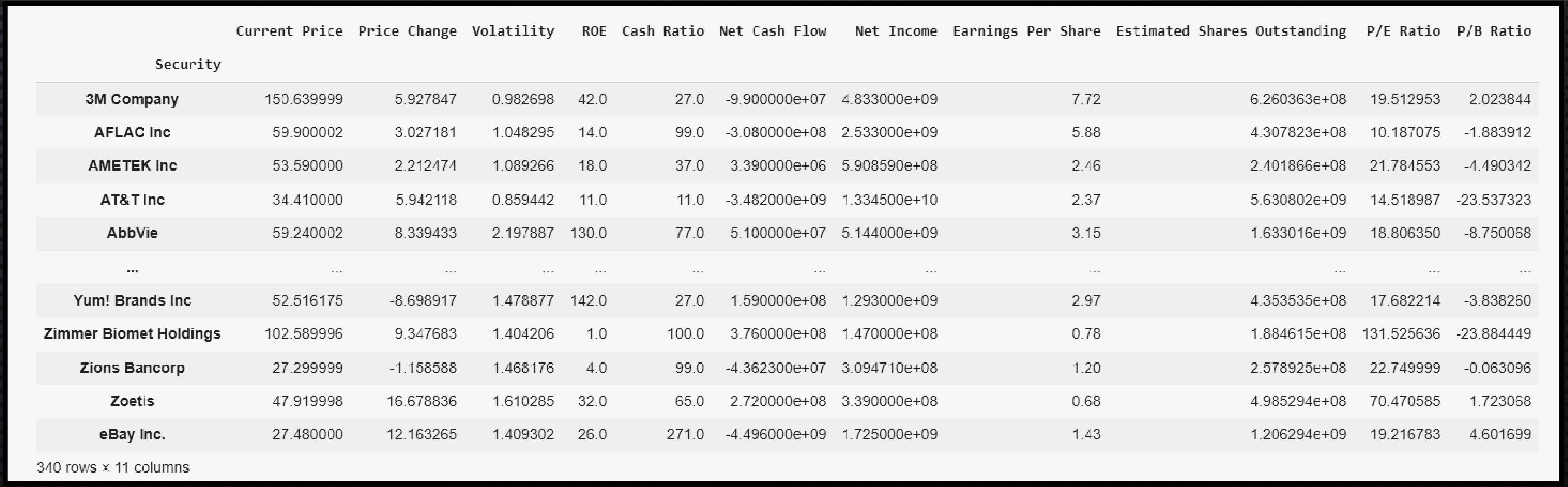
Histogram_Boxplots were completed for all Numeric Variables within the Dataset. Outliers were present in each Feature Variable

- Barplots for two of the Categorical Features [GICS Sector & GICS Sub Sector]
- GICS Sector - Specific Economic Sector assigned to a company that defines its Business Operations
- On the Right: the GICS Sector Barplot
- The highest percentages of businesses belong to these Economic Sectors
  - Industrials - Financials - Health Care - Consumer Discretionary



EDA Results

- Columns with Categorical Features were assessed to avoid problems when comparing Numerical Variables
- GICS Sector & GICS Sub Industry Columns were temporarily Dropped
- Security (Company Name) was grouped by Median Value



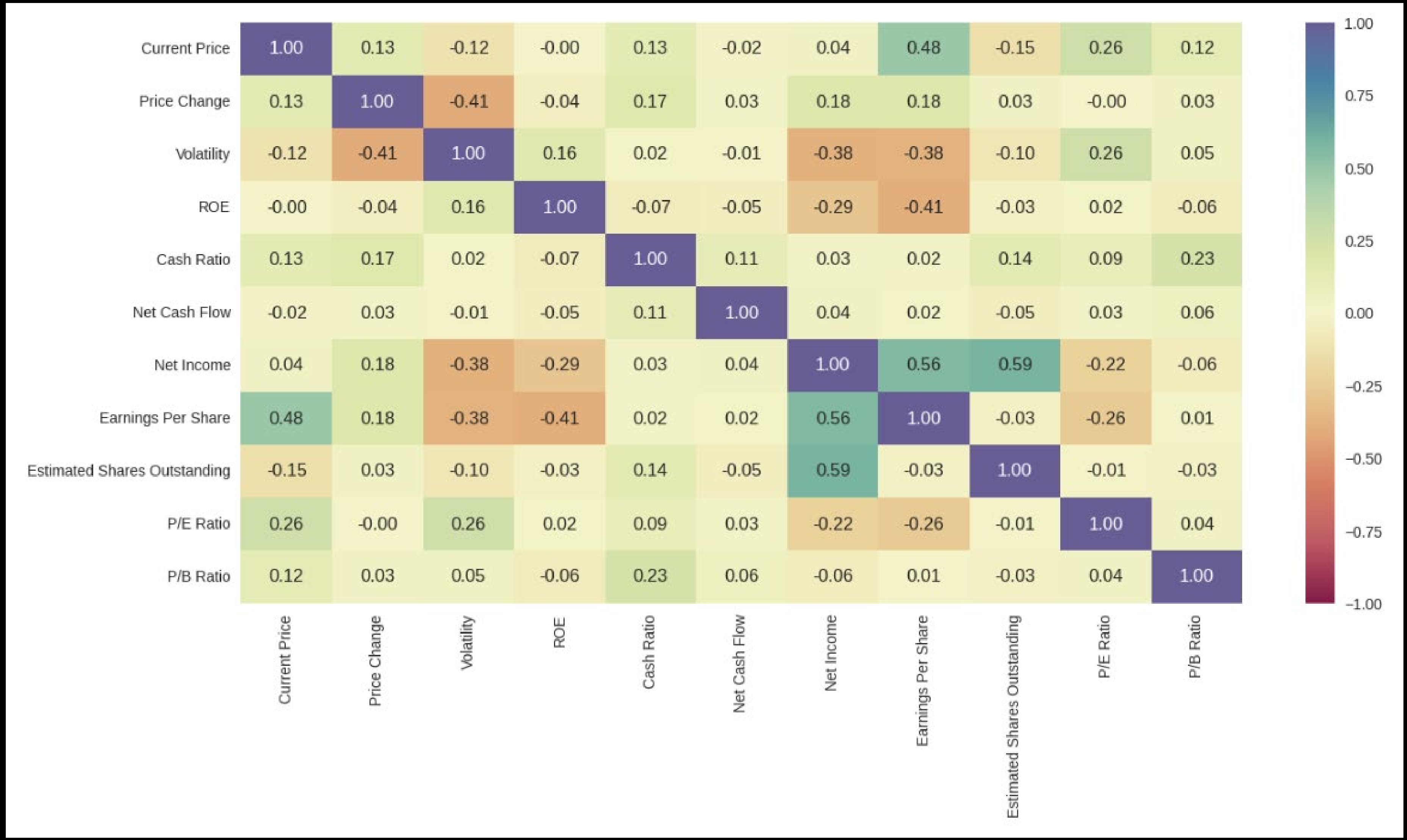
	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio
Security											
3M Company	150.639999	5.927847	0.982698	42.0	27.0	-9.900000e+07	4.833000e+09	7.72	6.260363e+08	19.512953	2.023844
AFLAC Inc	59.900002	3.027181	1.048295	14.0	99.0	-3.080000e+08	2.533000e+09	5.88	4.307823e+08	10.187075	-1.883912
AMETEK Inc	53.590000	2.212474	1.089266	18.0	37.0	3.390000e+06	5.908590e+08	2.46	2.401866e+08	21.784553	-4.490342
AT&T Inc	34.410000	5.942118	0.859442	11.0	11.0	-3.482000e+09	1.334500e+10	2.37	5.630802e+09	14.518987	-23.537323
AbbVie	59.240002	8.339433	2.197887	130.0	77.0	5.100000e+07	5.144000e+09	3.15	1.633016e+09	18.806350	-8.750068
...	...	...	...	...	...	...	...	...	...	...	...
Yum! Brands Inc	52.516175	-8.698917	1.478877	142.0	27.0	1.590000e+08	1.293000e+09	2.97	4.353535e+08	17.682214	-3.838260
Zimmer Biomet Holdings	102.589996	9.347683	1.404206	1.0	100.0	3.760000e+08	1.470000e+08	0.78	1.884615e+08	131.525636	-23.884449
Zions Bancorp	27.299999	-1.158588	1.468176	4.0	99.0	-4.362300e+07	3.094710e+08	1.20	2.578925e+08	22.749999	-0.063096
Zoetis	47.919998	16.678836	1.610285	32.0	65.0	2.720000e+08	3.390000e+08	0.68	4.985294e+08	70.470585	1.723068
eBay Inc.	27.480000	12.163265	1.409302	26.0	271.0	-4.496000e+09	1.725000e+09	1.43	1.206294e+09	19.216783	4.601699

340 rows × 11 columns

EDA Analysis

BIVARIATE ANALYSIS:

Heatmap was created for Correlation Check of Numeric Features of the Dataset



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

Heatmap - Correlation Comparison

- POSITIVE CORRELATIONS:

  Highest: Net Income & Estimated Shares Outstanding
  Net Income & Earnings Per Share
  Current Price & Earnings Per Share

- NEGATIVE CORRELATIONS:

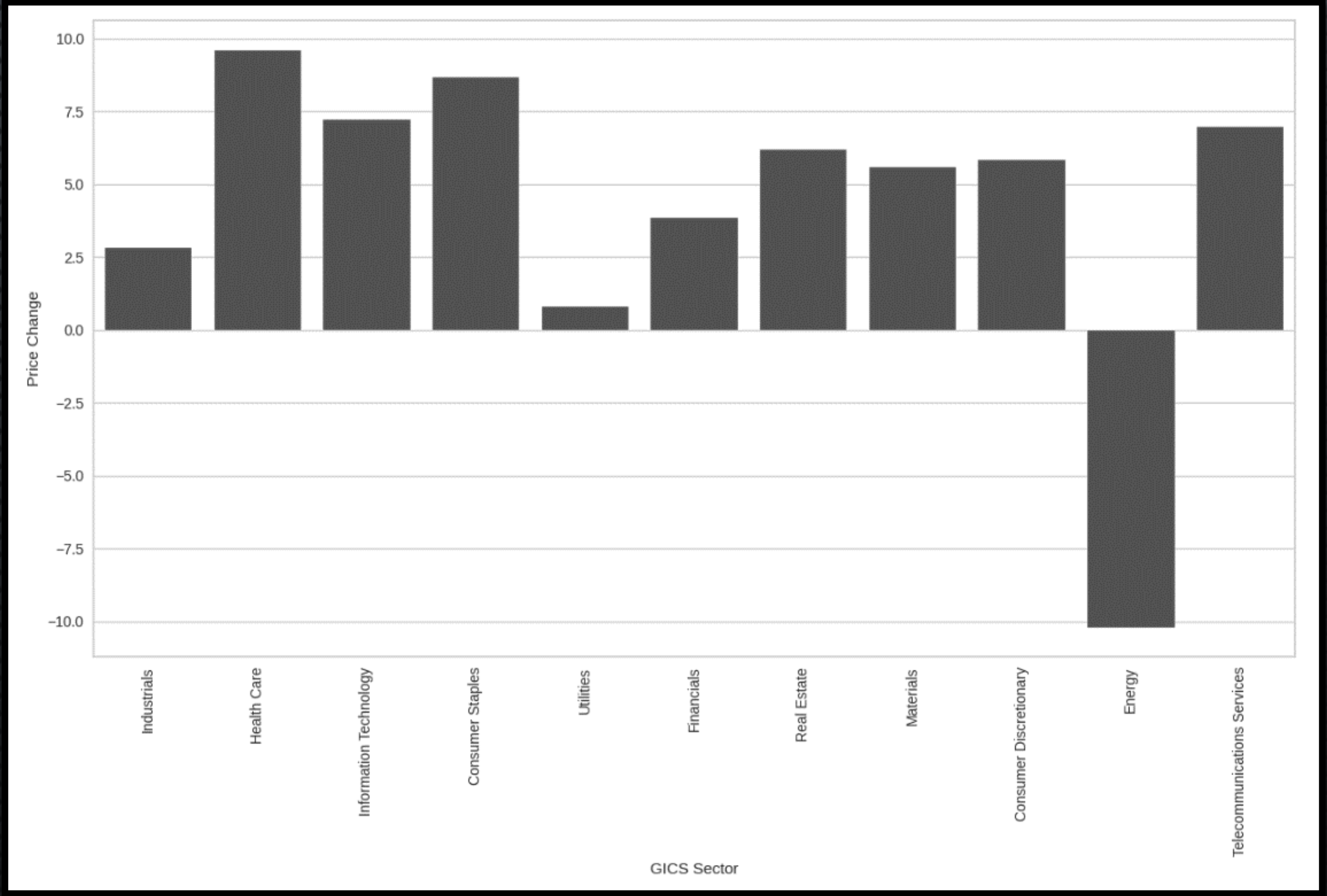
  Highest: Volatility & Price Change
  Volatility & Net Income
  Net Income & ROE

- Barplot comparisons will be made on specific features in relation to the Heatmap

Correlation Comparison

Maximum Price Increase on Average:

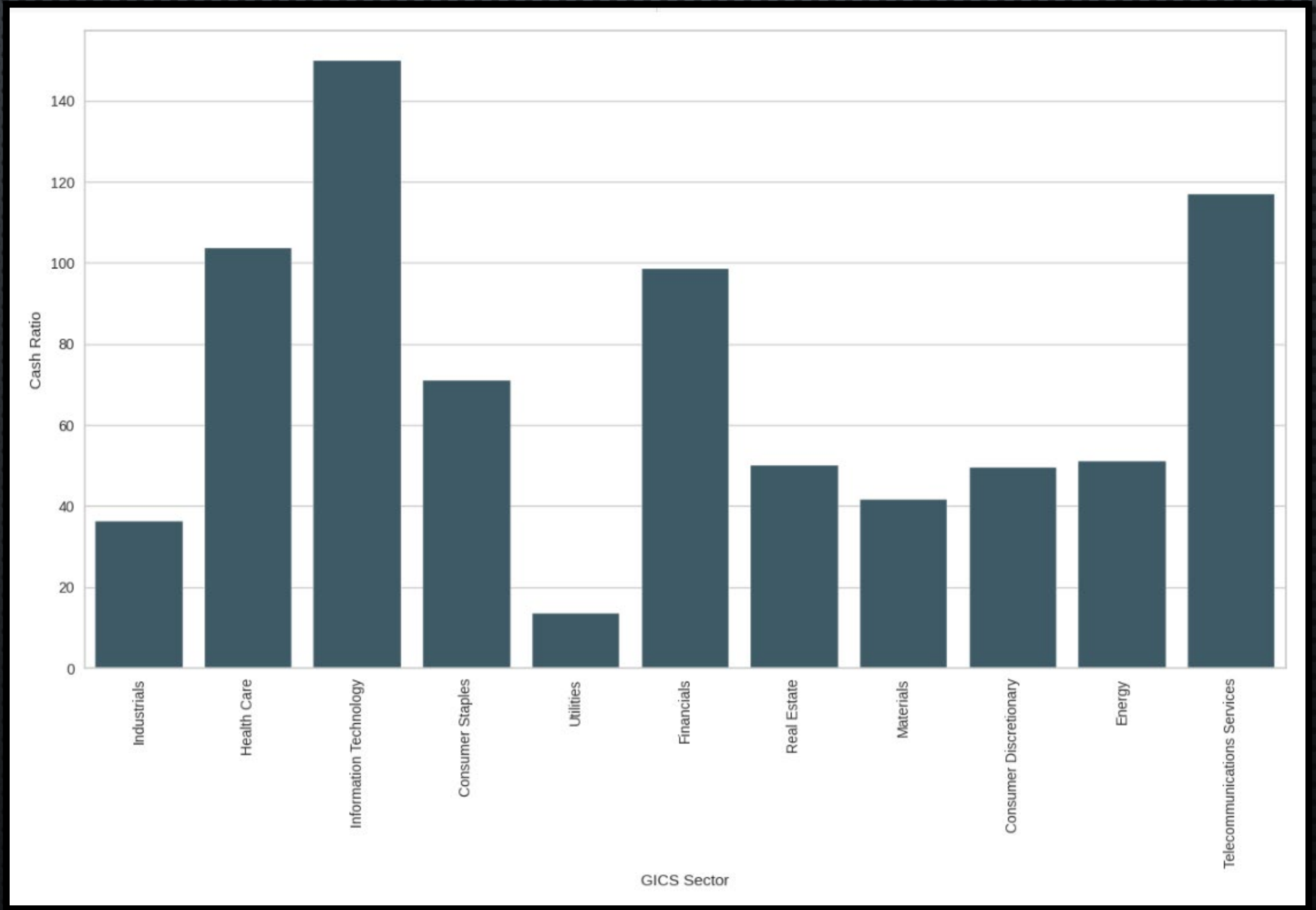
Health Care - Consumer Staples - IT - Communication Services



Correlation Comparison

Cash Ratio for short-term obligations - Average:

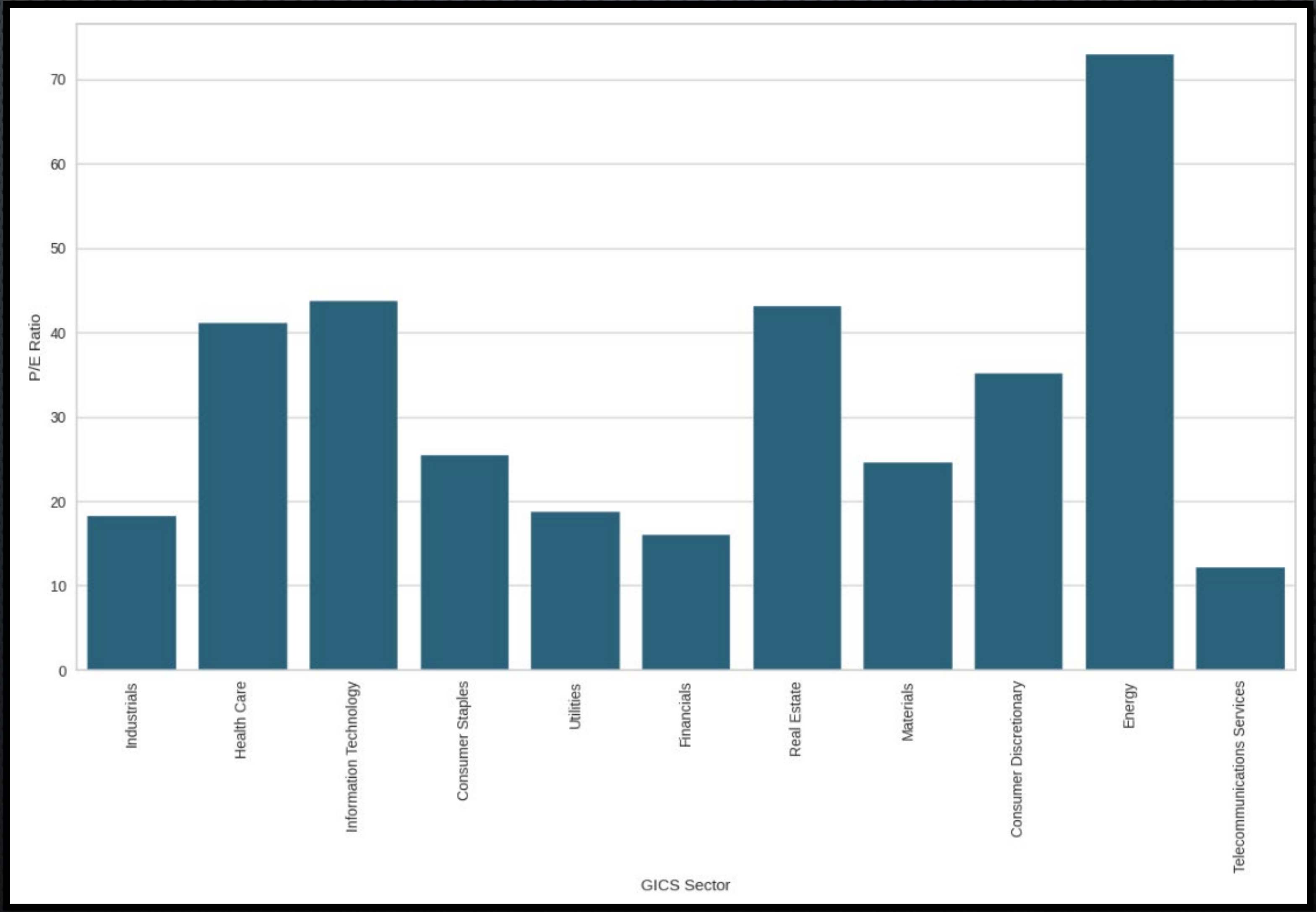
IT - Telecommunications Services - Health Care - Financials



Correlation Comparison

P/E Ratio (relative value of shares) Investments Average:

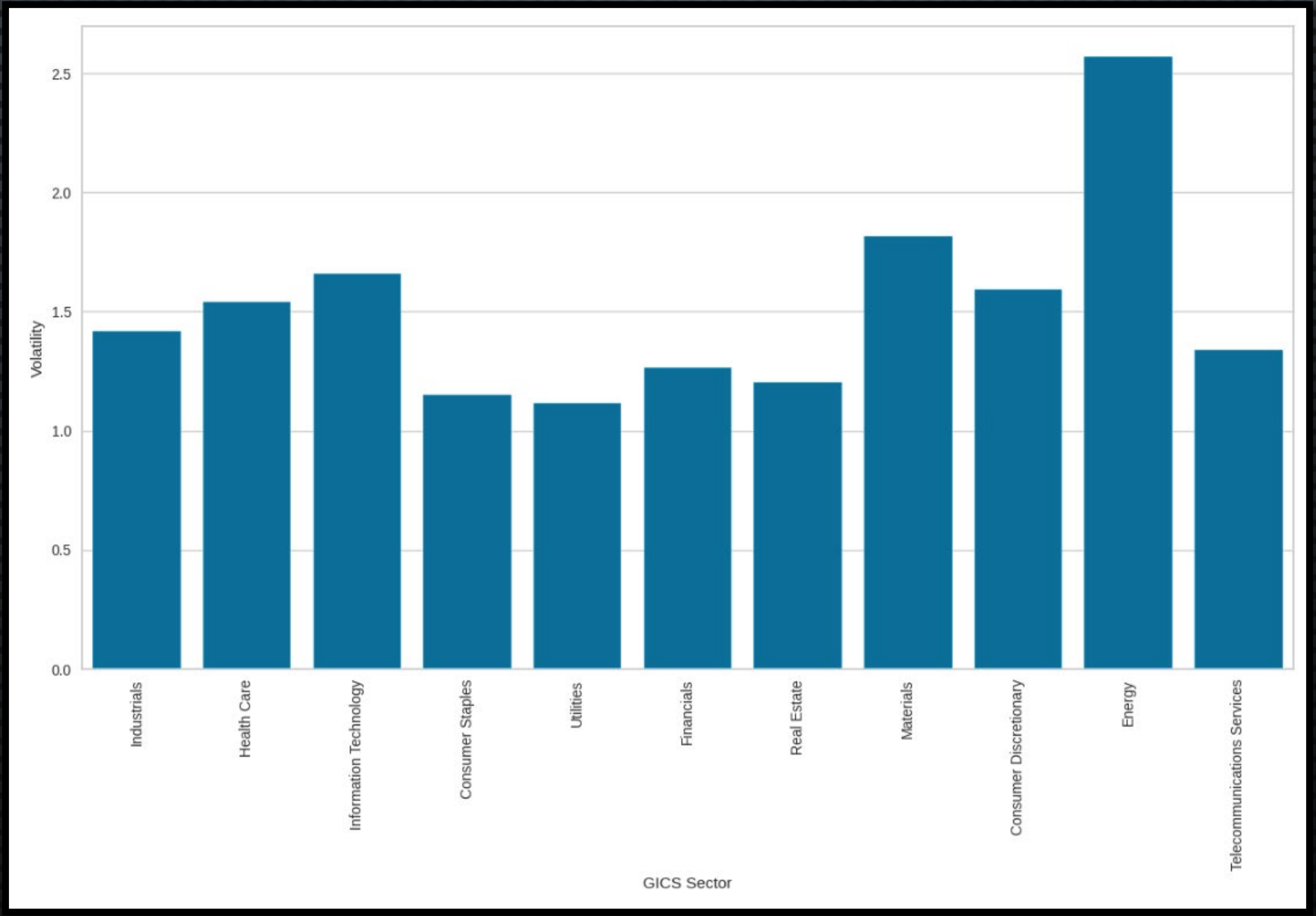
Energy - IT - Real Estate - Health Care - Consumer Discretionary



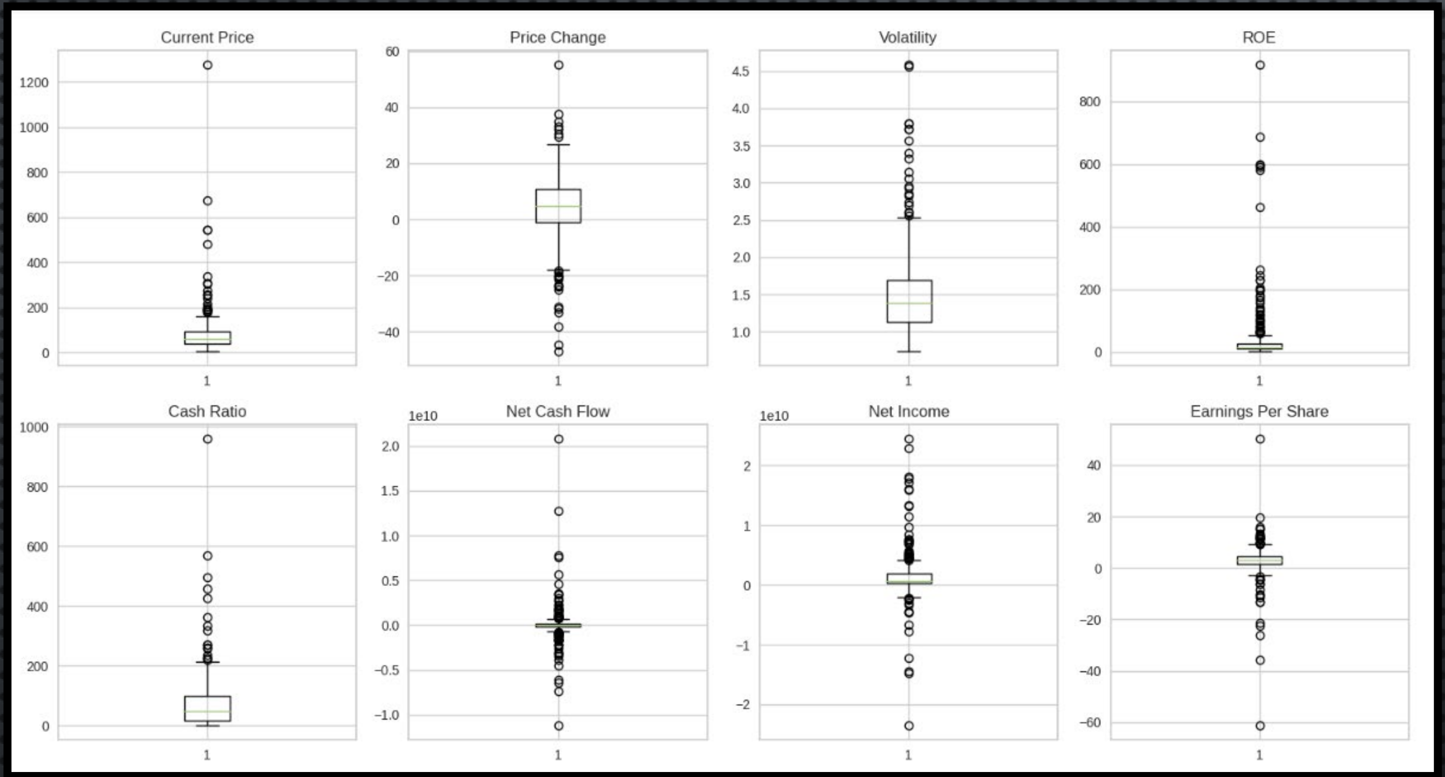
Correlation Comparison

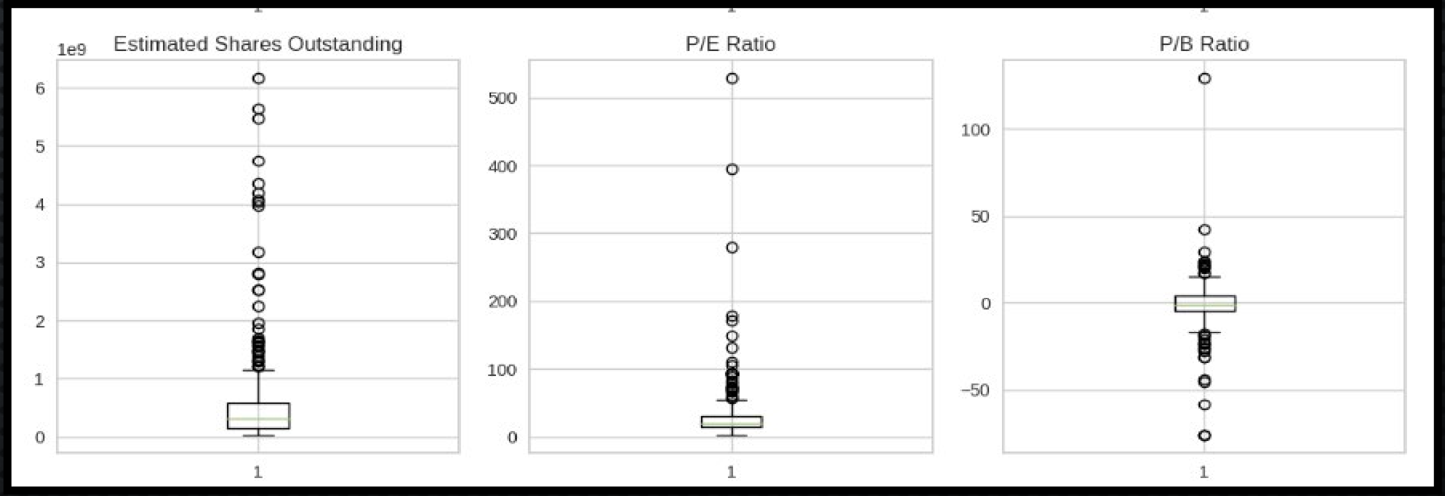
Volatility/Price Fluctuation = Risky investment Average across sectors:

Energy - Materials - IT - Consumer Discretionary - Health Care - Industrials



Data Preprocessing





Outlier Check

- All Numeric Features have Outliers within the Data
- Outliers were not treated or imputed
- No valuable information was lost
- Scaling of the Numerical Features will be the next step

Data Preprocessing

SCALING THE DATA - DISTANCE MEASURES

- To ensure that no single column of data highly influences Distance Measures we used Standard Scaler before performing K-means Clustering

- Different attributes within the dataset are represented on the same scale. Algorithmic Distance Calculations are improved after scaling

- Cluster Analysis and Distance Measures go hand in hand. Take into consideration whether attributes are independent of each other OR if the influence one another. Outliers also play a role in the dimensions of the clusters.

- Distance Measures: Euclidean Distance was used for K-means Clustering

- Euclidean Distance represents the shortest distance between two vectors

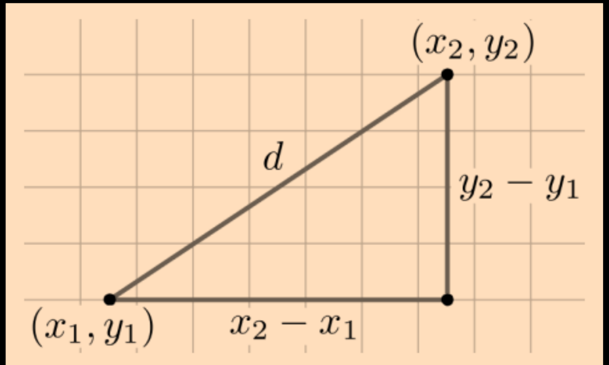
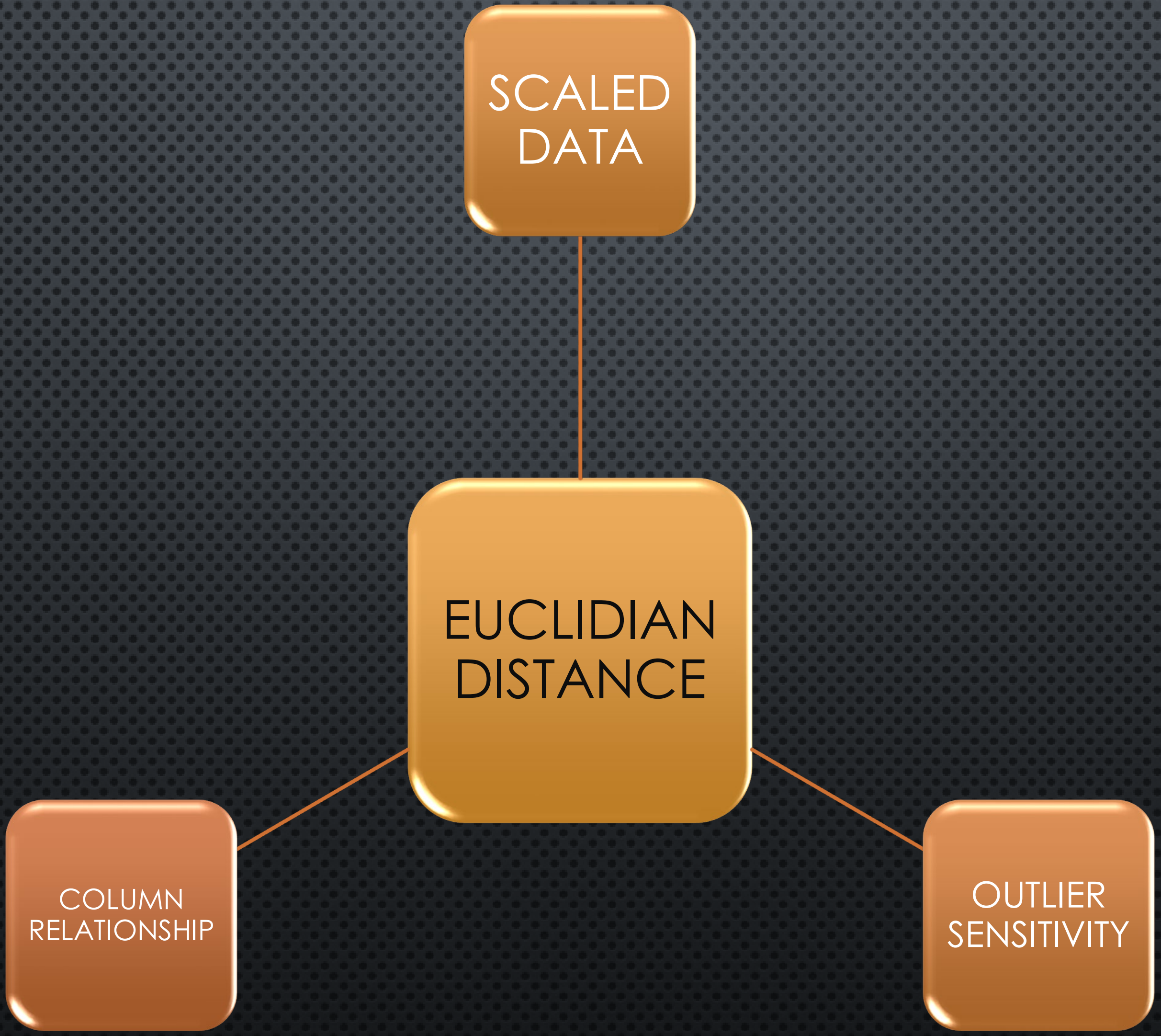


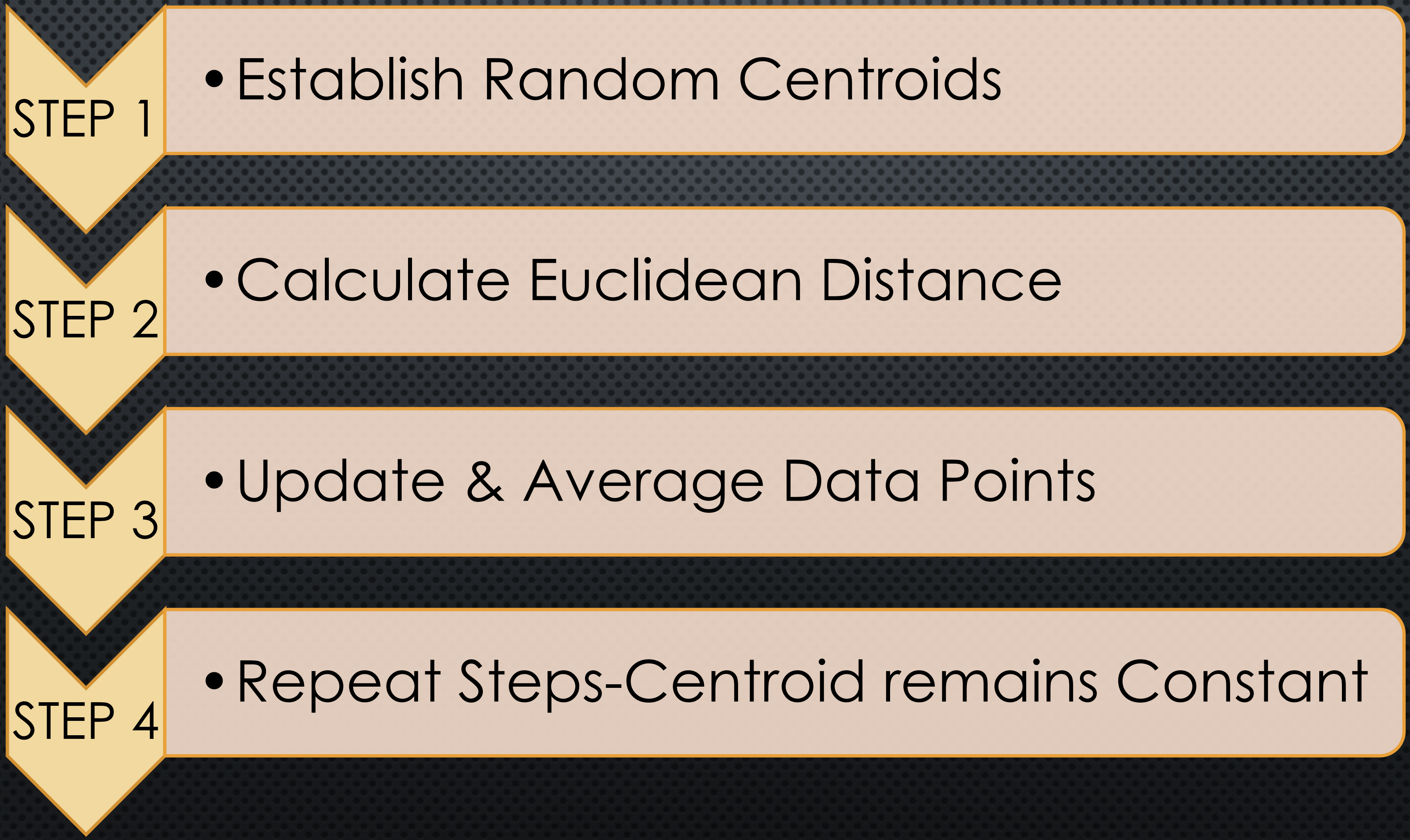
Figure 1: Two-dimensional Euclidean distance

Euclidean Distance Features



K-Means Clustering Summary

Centroid-Based Algorithm with the objective of finding the K-Number of clusters/groups



STEP 1

- Establish Random Centroids

STEP 2

- Calculate Euclidean Distance

STEP 3

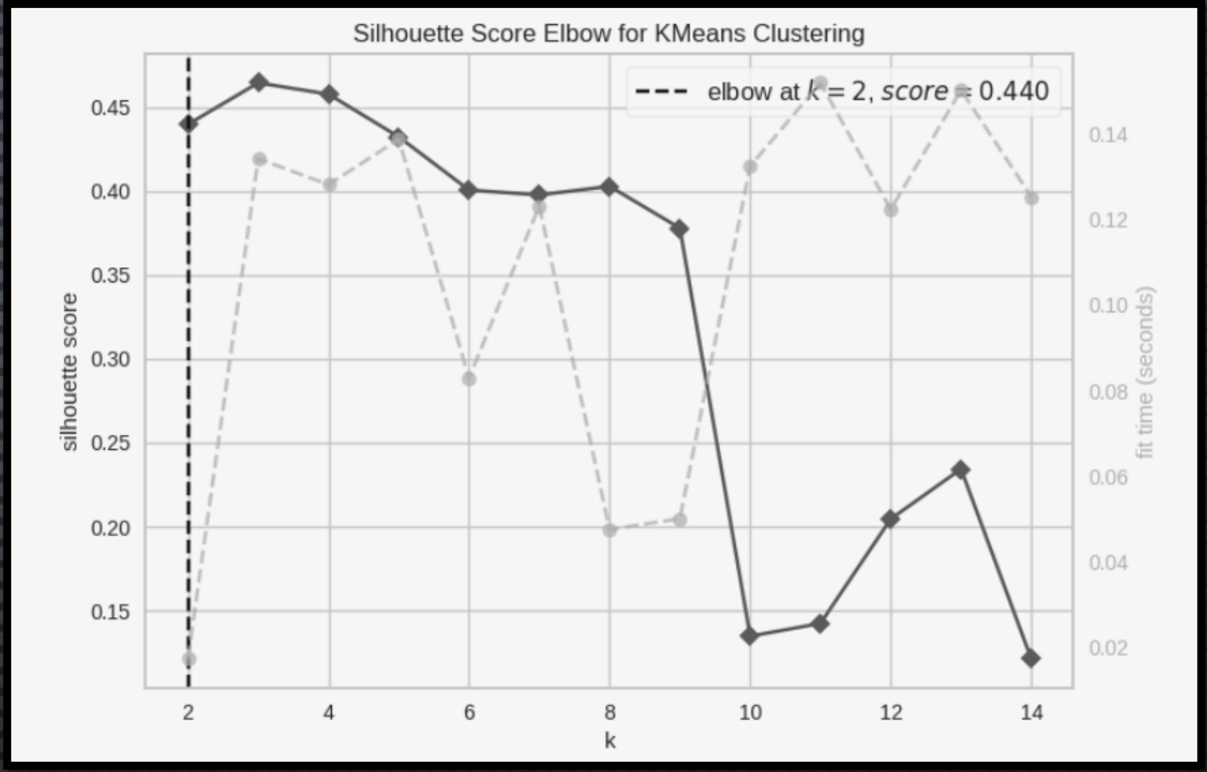
- Update & Average Data Points

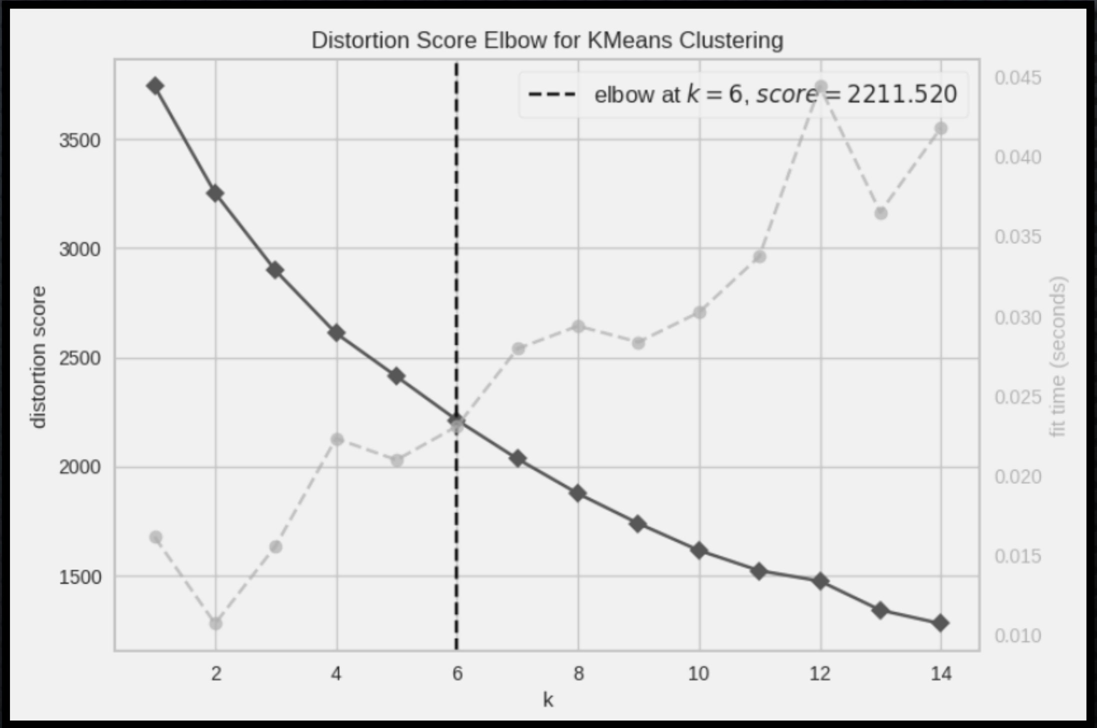
STEP 4

- Repeat Steps-Centroid remains Constant

K-Means Clustering

- Elbow Method - Optimal number of Clusters
- Plotting Cost Function against ranging values of K
- Distortion steepest decline defines the Elbow Point & the best number of clusters
- Here the optimal value is 6 & Distortion Score of 0.023



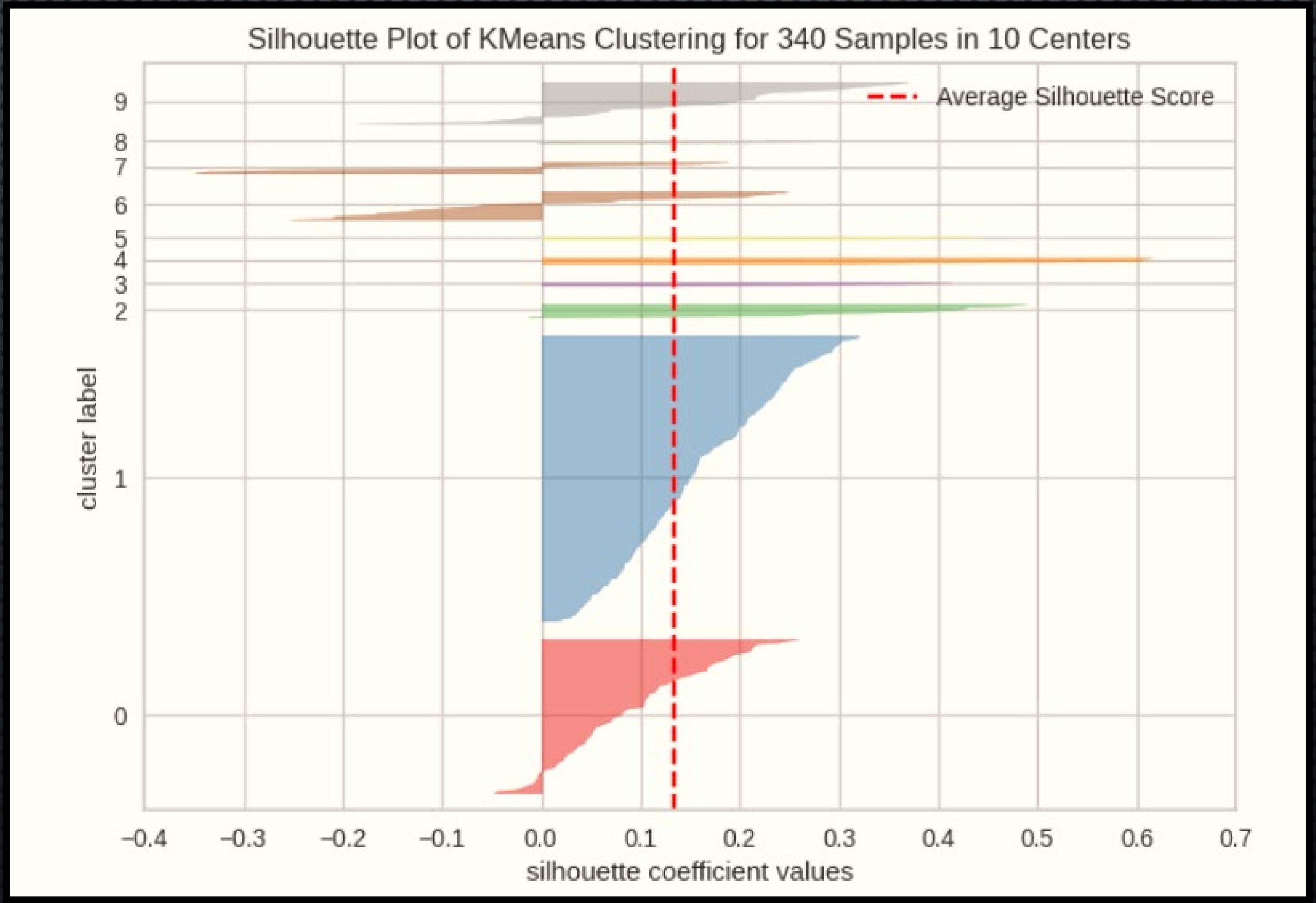


- Silhouette Score:
  - Average inter-cluster distance
  - Average cluster distance
  - 1 = Best | 0 = Not Good | -1 = Bad
- Here Silhouette Elbow is at 2 with Score of 0.440

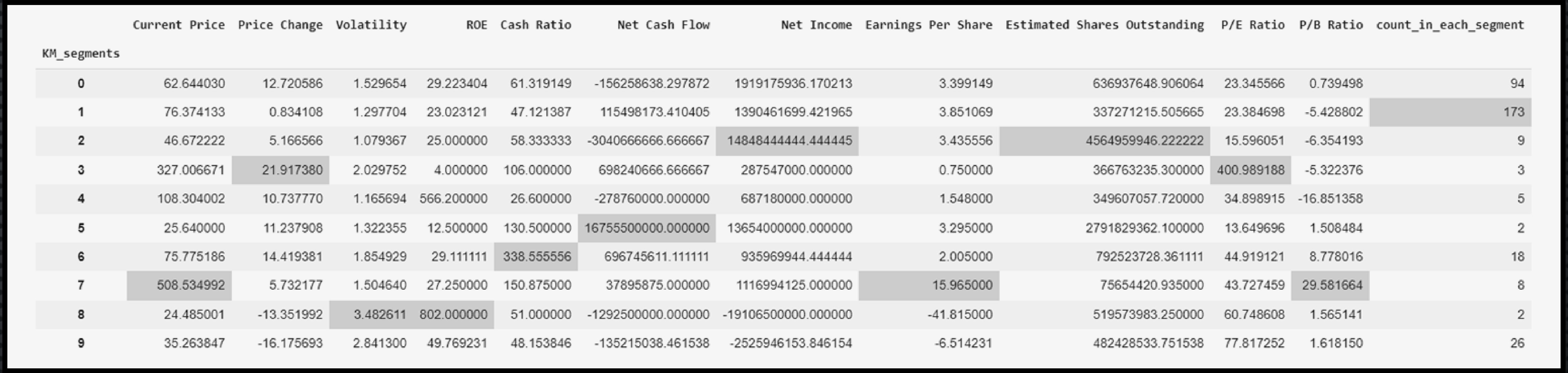
K-Means Clustering

SILHOUETTE PLOT:

We used 10 as the appropriate number of clusters. A sharp deviation was shown at 10 in the Elbow Method



K-means Final Model - Cluster Profiles



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	62.644030	12.720586	1.529654	29.223404	61.319149	-156258638.297872	1919175936.170213	3.399149	636937648.906064	23.345566	0.739498	94
1	76.374133	0.834108	1.297704	23.023121	47.121387	115498173.410405	1390461699.421965	3.851069	337271215.505665	23.384698	-5.428802	173
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	327.006671	21.917380	2.029752	4.000000	106.000000	698240666.666667	287547000.000000	0.750000	366763235.300000	400.989188	-5.322376	3
4	108.304002	10.737770	1.165694	566.200000	26.600000	-278760000.000000	687180000.000000	1.548000	349607057.720000	34.898915	-16.851358	5
5	25.640000	11.237908	1.322355	12.500000	130.500000	16755500000.000000	13654000000.000000	3.295000	2791829362.100000	13.649696	1.508484	2
6	75.775186	14.419381	1.854929	29.111111	338.555556	696745611.111111	935969944.444444	2.005000	792523728.361111	44.919121	8.778016	18
7	508.534992	5.732177	1.504640	27.250000	150.875000	37895875.000000	1116994125.000000	15.965000	75654420.935000	43.727459	29.581664	8
8	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1.565141	2
9	35.263847	-16.175693	2.841300	49.769231	48.153846	-135215038.461538	-2525946153.846154	-6.514231	482428533.751538	77.817252	1.618150	26

- Number of Clusters chosen is 10

- Create a new Copy of the Original Dataset

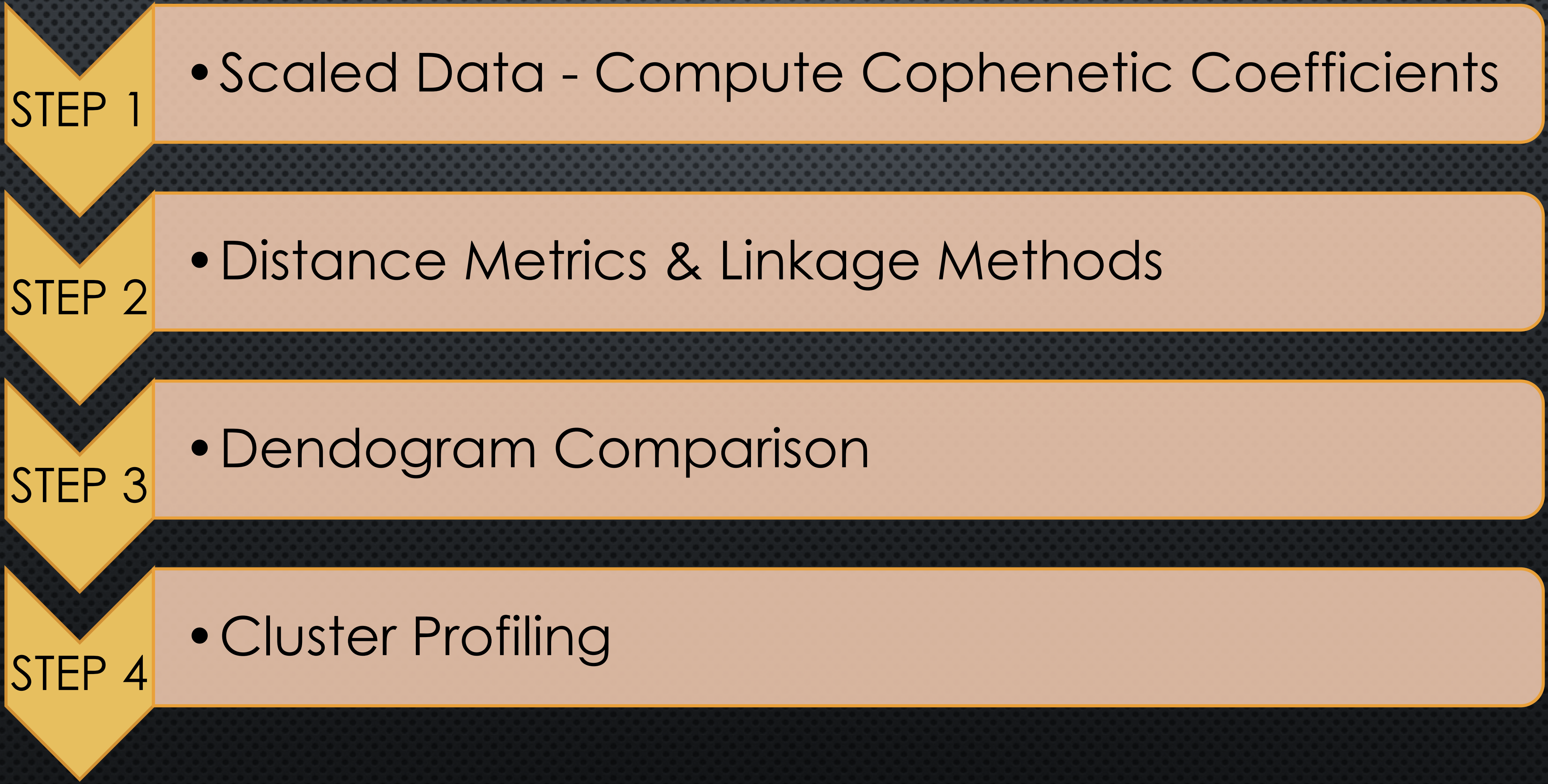
- Add the K-means Labels (KM_segments) to the Original & Scaled Dataset

- Cluster #7 only had Eight Companies Represented - Showed the highest Current Price - Earnings Per Share & P/B Ratio

- Industries Represented: 4 Health Care, 1 IT, 1 Real Estate, 2 Customer Discretionary

Hierarchical Clustering Summary

Density -Based Clustering: nearby points join to form clusters. Dendograms determine how many clusters should be formed



STEP 1

- Scaled Data - Compute Cophenetic Coefficients

STEP 2

- Distance Metrics & Linkage Methods

STEP 3

- Dendogram Comparison

STEP 4

- Cluster Profiling

Hierarchical Clustering Techniques

- Cluster Formation:

  1. Divisive - Single cluster and divide into multiple clusters
  2. Agglomerative - Multiple clusters and bringing the together to form one

- We used the Agglomerative Formation

  - Calculate distance between new closer clusters increases probability of being in the same cluster

  - Process repeated until One Cluster contains all Sub-Clusters

- Distance Metrics:

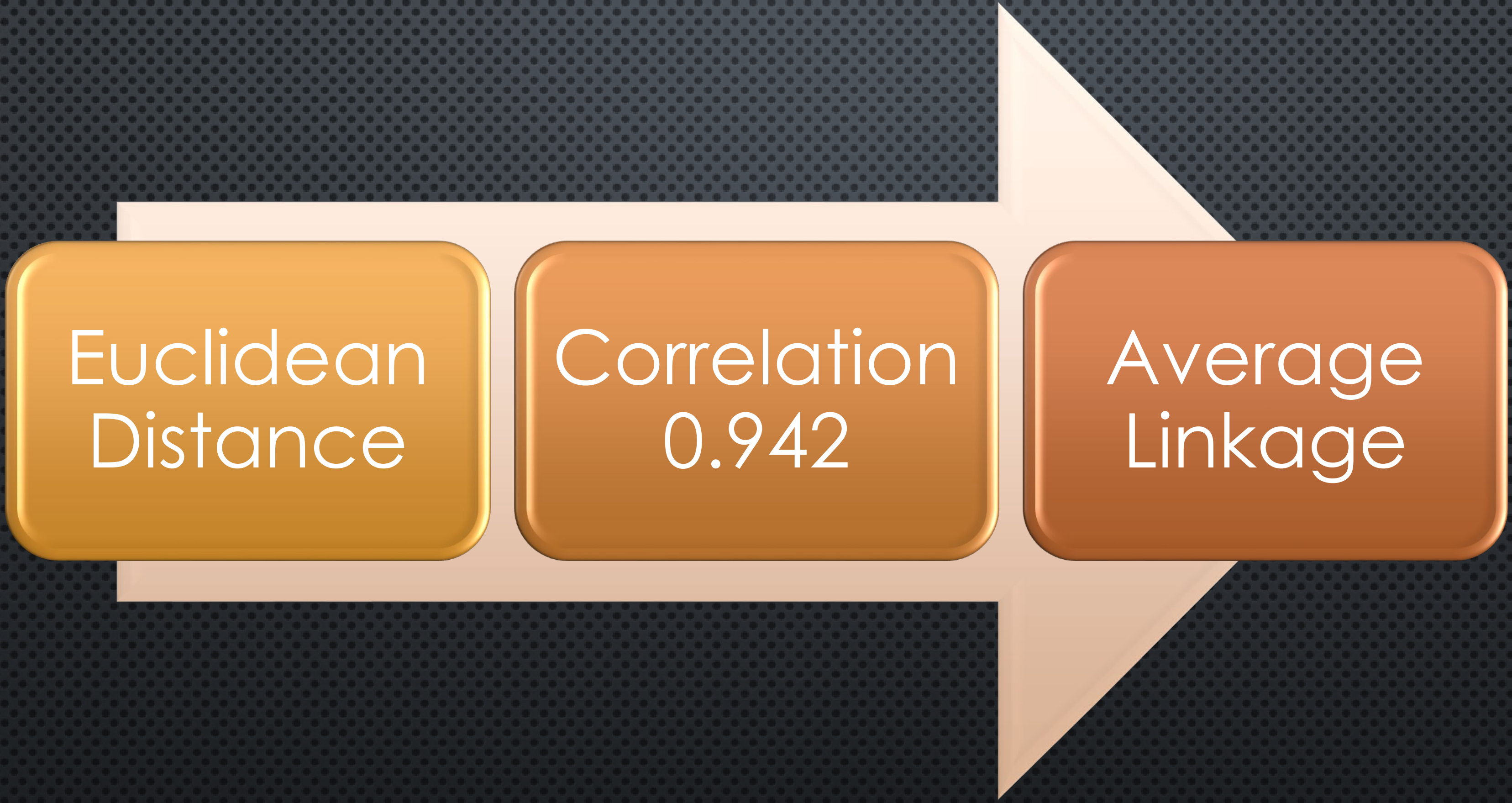
  Euclidean, Chebyshev, Mahalanobis, Cityblock

- Linkage Methods:

  Single, Complete, Average, Centroid, Ward, Weighted

- We used the Euclidean Distance Metric

Hierarchical Clustering



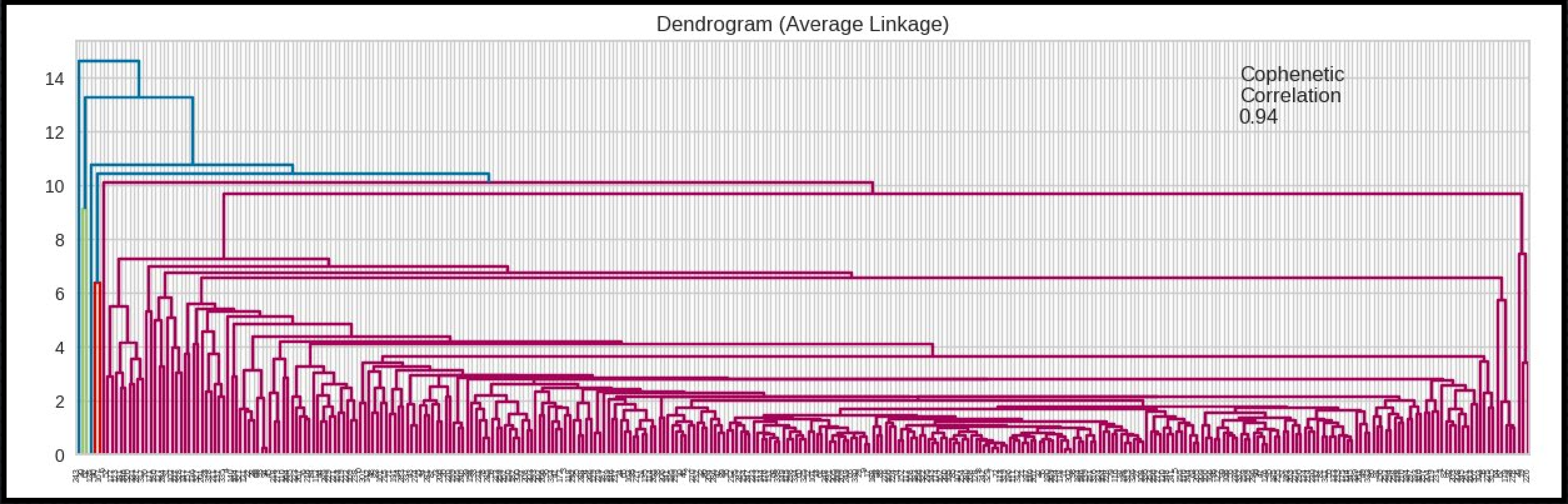
Euclidean Distance

Correlation 0.942

Average Linkage

Hierarchical Clustering

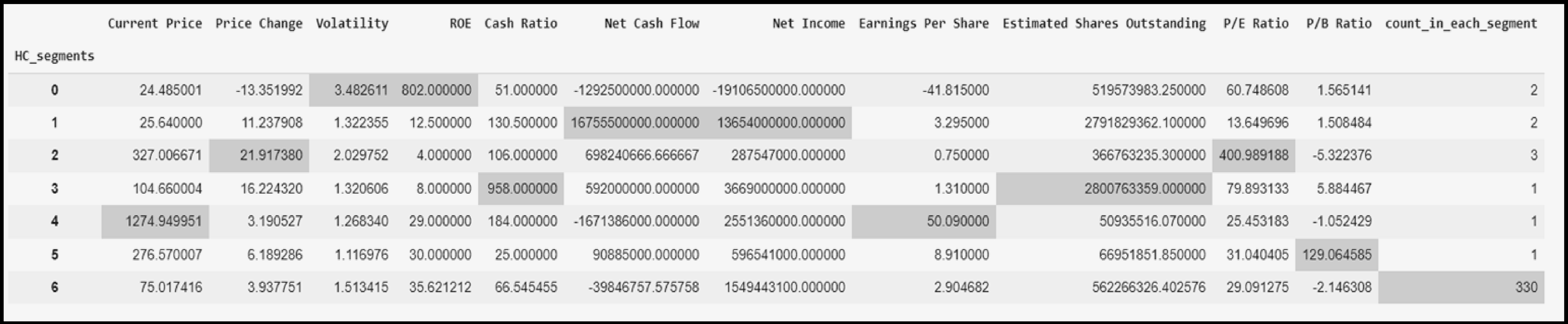
DENDOGRAM - Average Linkage - Correlation 0.94



Hierarchical Model - sklearn & Cluster Profiles

- Agglomerative: 7 Clusters - Euclidean Distance - Average Linkage

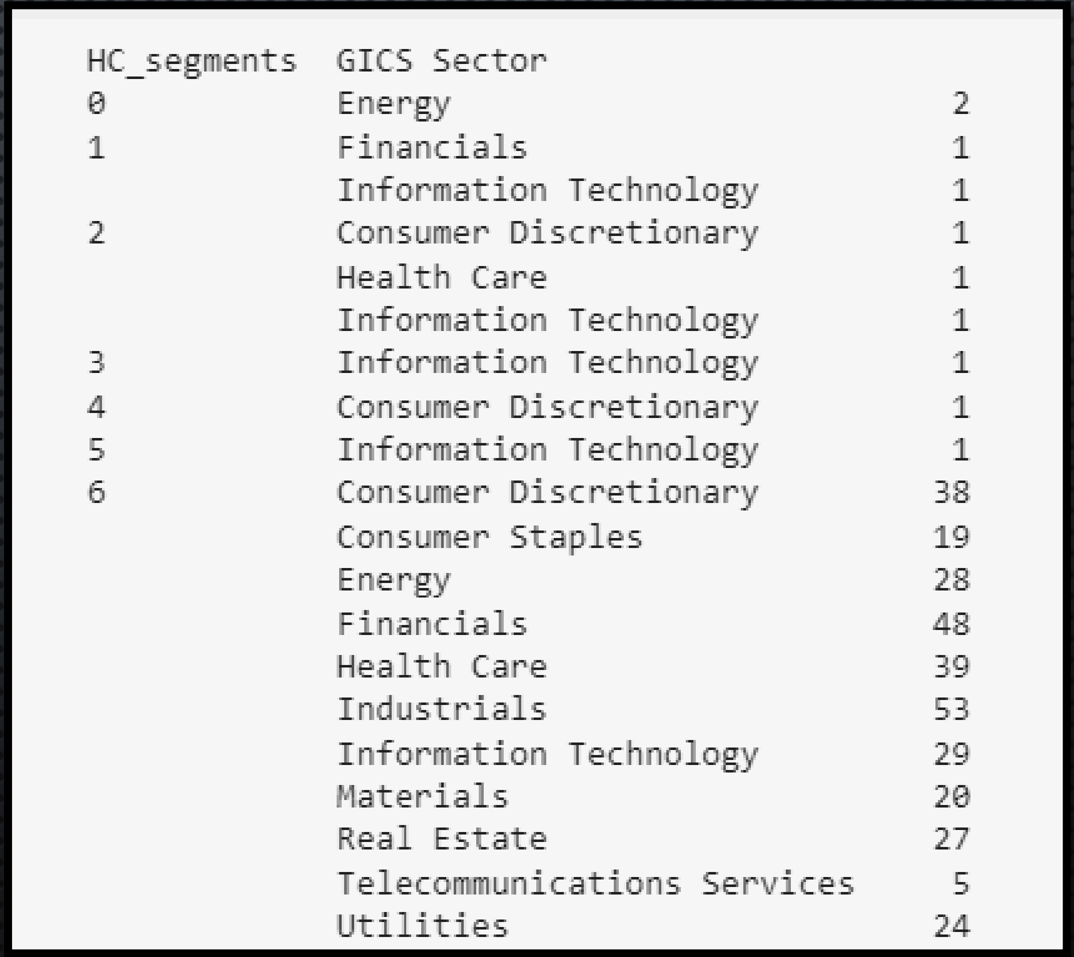
- Segment #6 has 300 Companies
  Current Price is on the low end at $75
  Price Change is stable in comparison to other segments
  Volatility is Average



	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
HC_segments												
0	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1.565141	2
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	327.006671	21.917380	2.029752	4.000000	106.000000	698240666.666667	287547000.000000	0.750000	366763235.300000	400.989188	-5.322376	3
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
6	75.017416	3.937751	1.513415	35.621212	66.545455	-39846757.575758	1549443100.000000	2.904682	562266326.402576	29.091275	-2.146308	330

Hierarchical Model - GICS SECTORS

The goal of getting to a Single Cluster was 97% successful using 7 Clusters



```
HC_segments  GICS Sector
0            Energy                          2
1            Financials                      1
             Information Technology          1
2            Consumer Discretionary          1
             Health Care                     1
             Information Technology          1
3            Information Technology          1
4            Consumer Discretionary          1
5            Information Technology          1
6            Consumer Discretionary         38
             Consumer Staples               19
             Energy                         28
             Financials                     48
             Health Care                    39
             Industrials                    53
             Information Technology         29
             Materials                      20
             Real Estate                    27
             Telecommunications Services     5
             Utilities                      24
```

K-means vs Hierarchical Comparison

1. Which clustering technique took less execution time?

   Hierarchical Clustering took less time, only because there were less steps within the process

2. Which technique gave more distinct clusters?

   K-means offered more defined clusters that offered more cluster profiles that showed separation of the clusters

3. How many observations are there in the similar clusters of both algorithms?

   In both K-means and Hierarchical the cluster with the highest count in a segment had no other highlighted areas of comparison

4. How many clusters were obtained as the appropriate number from both algorithms?

   K-means Clustering - 10 was the appropriate number of Clusters
   Hierarchical Clustering - 7 was the number of Clusters

Insights:

- Cluster #7 - Companies with a high Current Price, Earnings per Share & P/B Ratio, and low Volatility
- Cluster #0 - Showed average Current Price, Price Change, and Earnings per Share, but all other factors were below average
- Cluster #9 - 26 companies had no attributes that stood out. Net income was negative for the group and also Price Change

Recommendations:

- K-means Clustering lead to better attribute comparison of features in this particular situation
- Segment #7 gave the most measure of attributes among all other segments
- K-means Clustering required more steps than Hierarchical , but would not be considered time consuming or expensive for the company
- The Elbow Method within K-means gives direction and options to select the optimal K number of clusters on the scaled data