TRADE & AHEAD PROJECT

BY: DR. CYNTHIA OFFOHA

AGENDA

- EXECUTIVE SUMMARY
- BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH
- DATA OVERVIEW
- EDA: UNIVARIATE ANALYSIS
- EDA: BIVARIATE ANALYSIS
- DATA PREPROCESSING
- K-MEANS & HIERARCHICAL CLUSTERING
- RECOMMENDATIONS & CONCLUSIONS

EXECUTIVE SUMMARY

- Stock market is shown to be an effective way for investment and saving for the future
- There are various reasons for investing in stocks, which are:
  - It can assist in fighting inflation, creating wealth, and providing assistant in tax benefits
  - The steady returns on stock investments over long period of time can lead to tremendous growth in investment returns
  - Due to compound interest seen on stock investments at earlier stages has resulted greater outcomes on retirement
  - It can assist in meeting life's financial aspirations

- It is suggested to maintain diversified portfolio when investing in stocks in order to enhance earnings under any market condition
  - This tends to yield higher returns and resulting in lower risk by tempering potential losses when the market is down
- With the aid of cluster analysis, the identification of stocks with similar characteristics and ones with minimum correlations can be identified
  - This will assist investors to better investigate stocks across different segments and help protect risks that could make the portfolio vulnerable to losses

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

- Trade&Ahead is a financial consultancy firm who assist their customers with personalized investment strategies
  - They are in search of Data Scientist that will assist in data analysis of New York Stock Exchange
    - Data comprising stock prices and some financial indicators for few companies have been provided
    - The data scientist would be assigned the tasks of analyzing the data, grouping the stocks based on attributes provided, and sharing insights regarding the characteristics of each group

DATA OVERVIEW

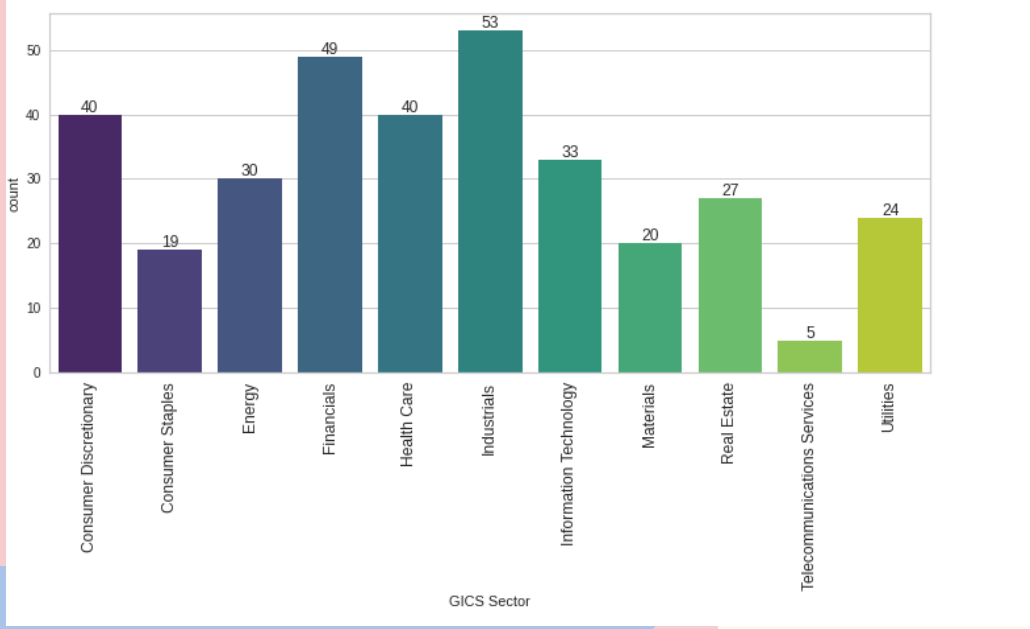
VARIABLE	DESCRIPTIONS
Ticker Symbol	An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
Company	Name of the company
GICS Sector	The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that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

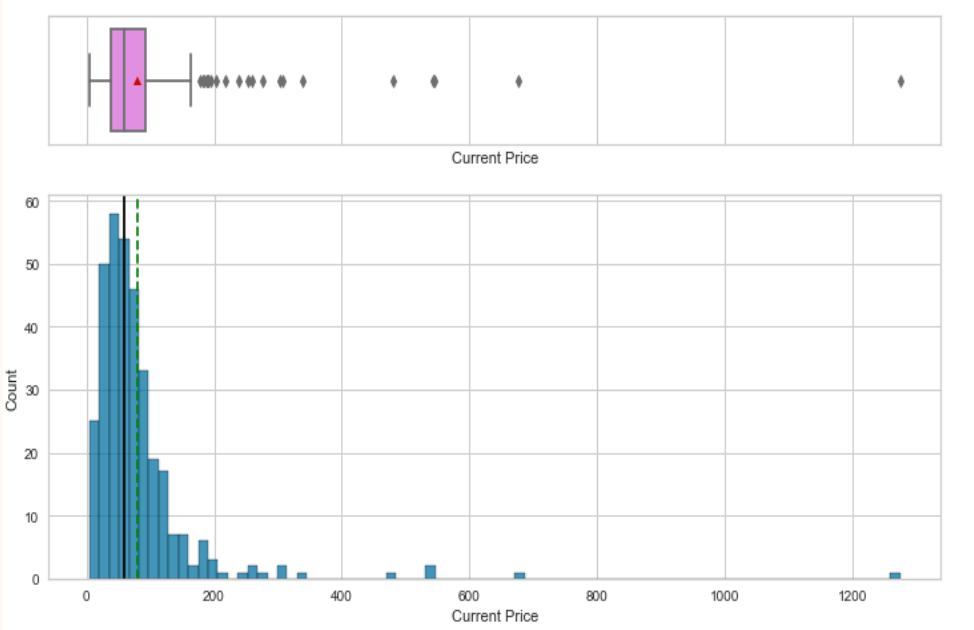
	Ticker Symbol	Security	GICS Sector	GICS Sub Industry	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share
0	AAL	American Airlines Group	Industrials	Airlines	42.35	10.00	1.69	135	51	-604000000	7610000000	11.39
1	ABBV	AbbVie	Health Care	Pharmaceuticals	59.24	8.34	2.20	130	77	51000000	5144000000	3.15
2	ABT	Abbott Laboratories	Health Care	Health Care Equipment	44.91	11.30	1.27	21	67	938000000	4423000000	2.94
3	ADBE	Adobe Systems Inc	Information Technology	Application Software	93.94	13.98	1.36	9	180	-240840000	629551000	1.26
4	ADI	Analog Devices, Inc.	Information Technology	Semiconductors	55.32	-1.83	1.70	14	272	315120000	696878000	0.31

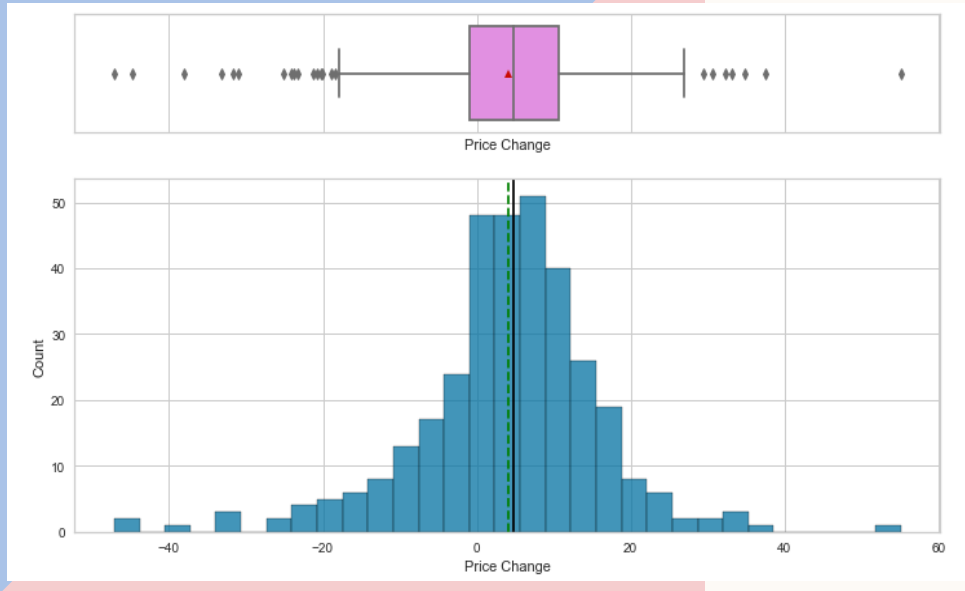
	count	mean	std	min	25%	50%	75%	max
Current Price	340.0	8.086234e+01	9.805509e+01	4.500000e+00	3.855500e+01	5.970500e+01	9.288000e+01	1.274950e+03
Price Change	340.0	4.078194e+00	1.200634e+01	-4.712969e+01	-9.394838e-01	4.819505e+00	1.069549e+01	5.505168e+01
Volatility	340.0	1.525976e+00	5.917984e-01	7.331632e-01	1.134878e+00	1.385593e+00	1.695549e+00	4.580042e+00
ROE	340.0	3.959706e+01	9.654754e+01	1.000000e+00	9.750000e+00	1.500000e+01	2.700000e+01	9.170000e+02
Cash Ratio	340.0	7.002353e+01	9.042133e+01	0.000000e+00	1.800000e+01	4.700000e+01	9.900000e+01	9.580000e+02
Net Cash Flow	340.0	5.553762e+07	1.946365e+09	-1.120800e+10	-1.939065e+08	2.098000e+06	1.698108e+08	2.076400e+10
Net Income	340.0	1.494385e+09	3.940150e+09	-2.352800e+10	3.523012e+08	7.073360e+08	1.899000e+09	2.444200e+10
Earnings Per Share	340.0	2.776662e+00	6.587779e+00	-6.120000e+01	1.557500e+00	2.895000e+00	4.620000e+00	5.009000e+01
Estimated Shares Outstanding	340.0	5.770283e+08	8.458496e+08	2.767216e+07	1.588482e+08	3.096751e+08	5.731175e+08	6.159292e+09
P/E Ratio	340.0	3.261256e+01	4.434873e+01	2.935451e+00	1.504465e+01	2.081988e+01	3.176476e+01	5.280391e+02
P/B Ratio	340.0	-1.718249e+00	1.396691e+01	-7.611908e+01	-4.352056e+00	-1.067170e+00	3.917066e+00	1.290646e+02

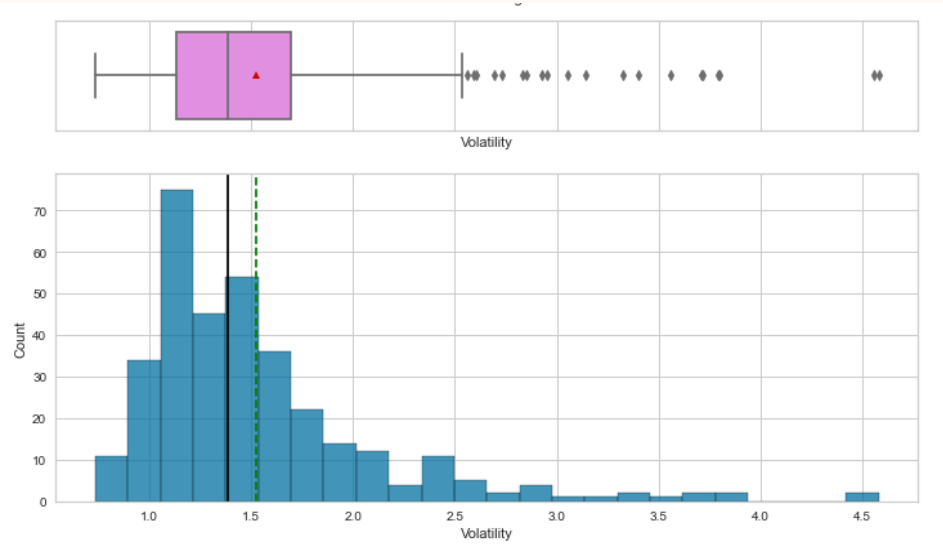
- Data set consist of 340 rows and 15 columns
  - Columns with dtype object should be dtype category to conserve memory
- No missing or duplicate value in dataset
- Variables --> 7 float, 4 integer, and 4 object
- It is shown that the current price, ROE, Cash ratio, Net Cash flow Net income, Estimated shares outstanding and P/E ratio, the average is greater than the median
  - Distribution is right skewed

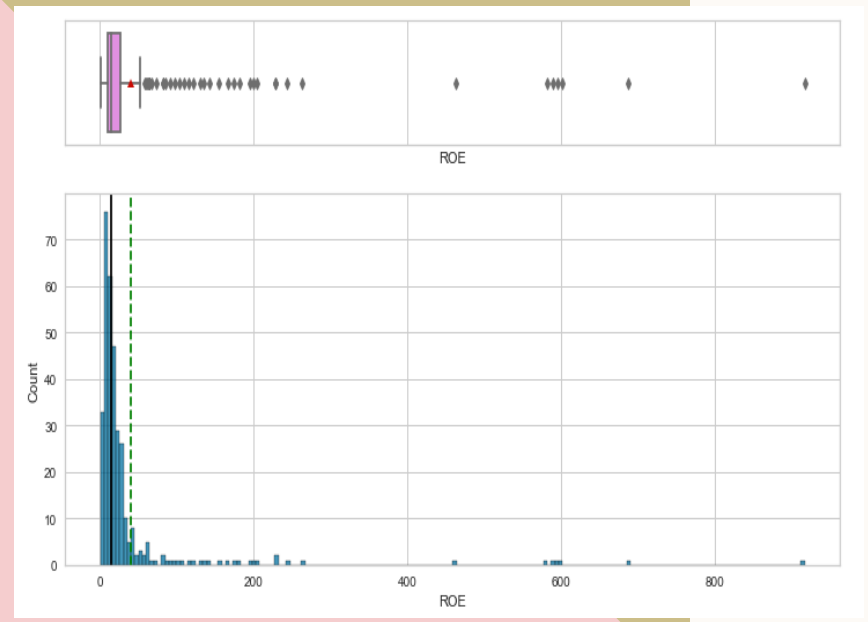
EDA: UNIVARIATE ANALYSIS

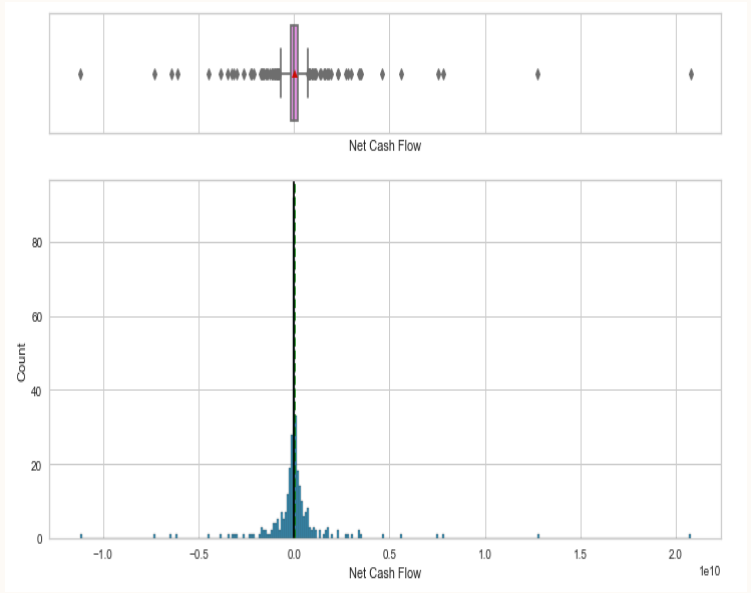


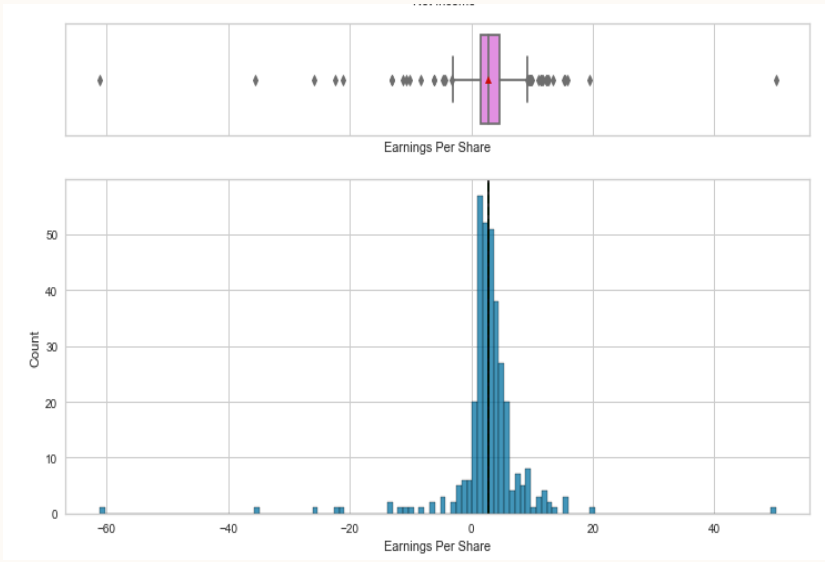


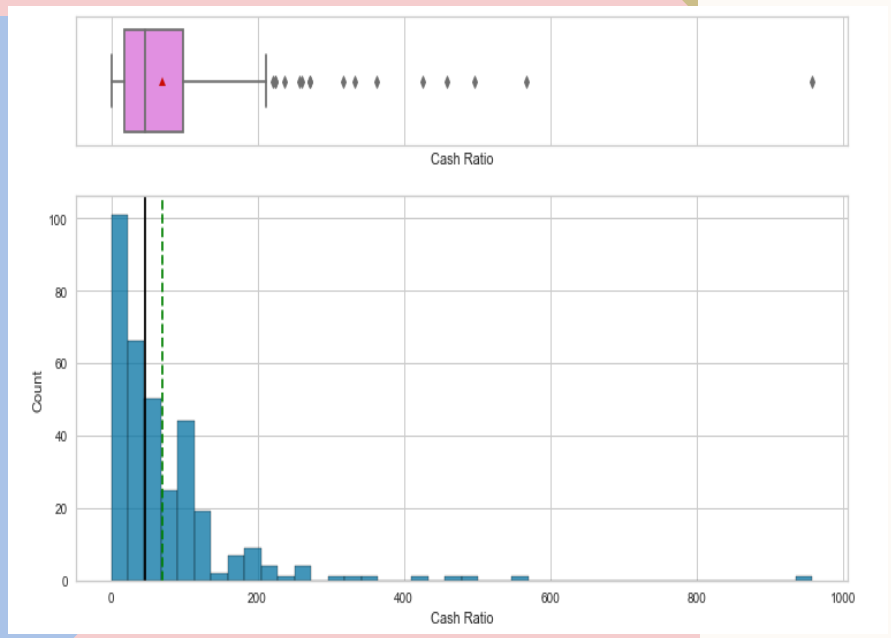


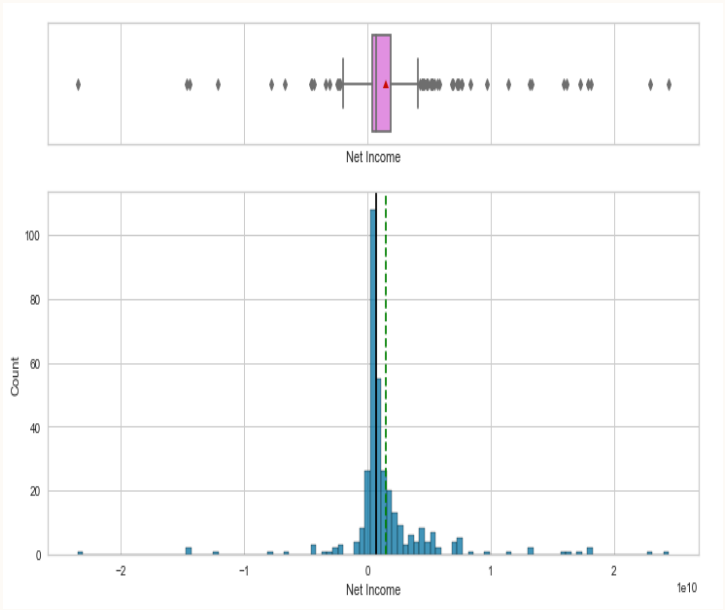


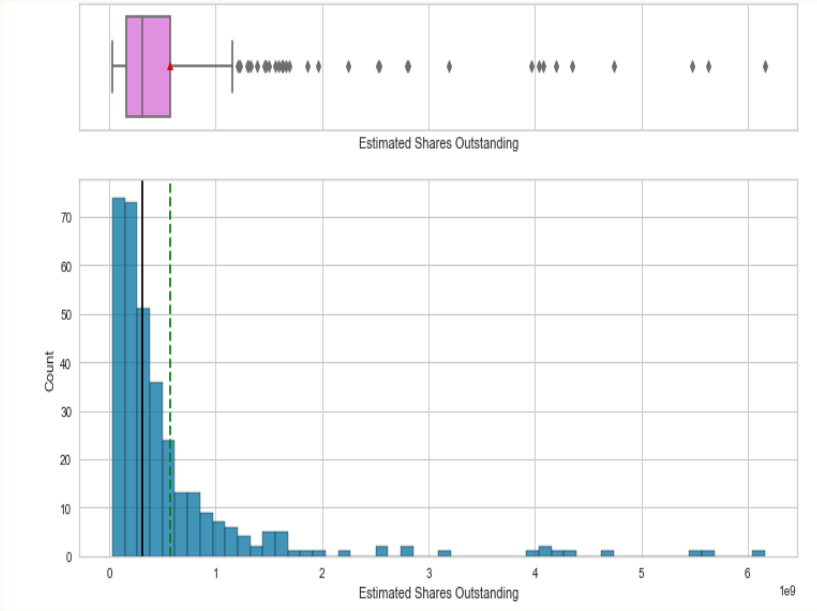

ROE
Count
ROE

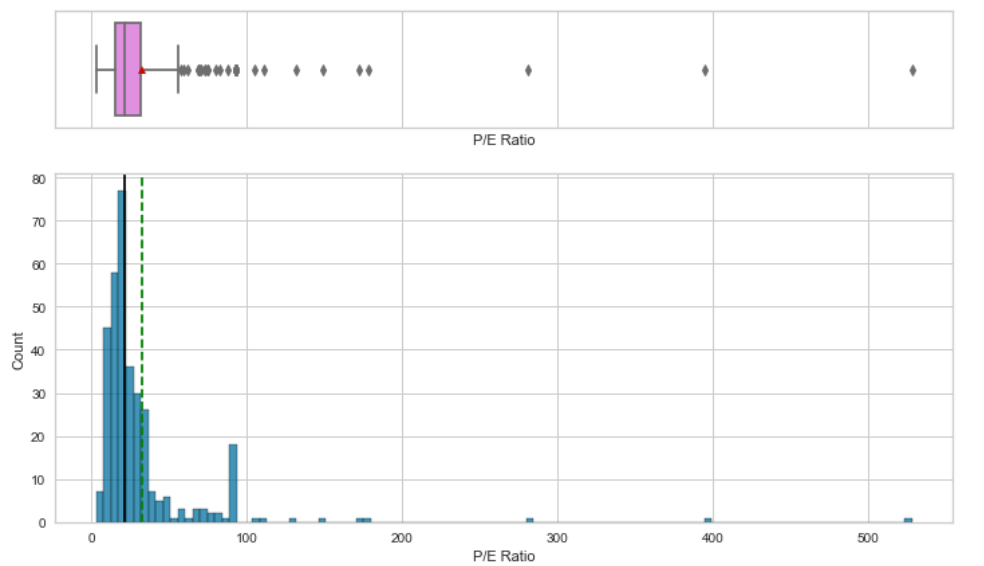

Net Cash Flow
Count
Net Cash Flow
1e10

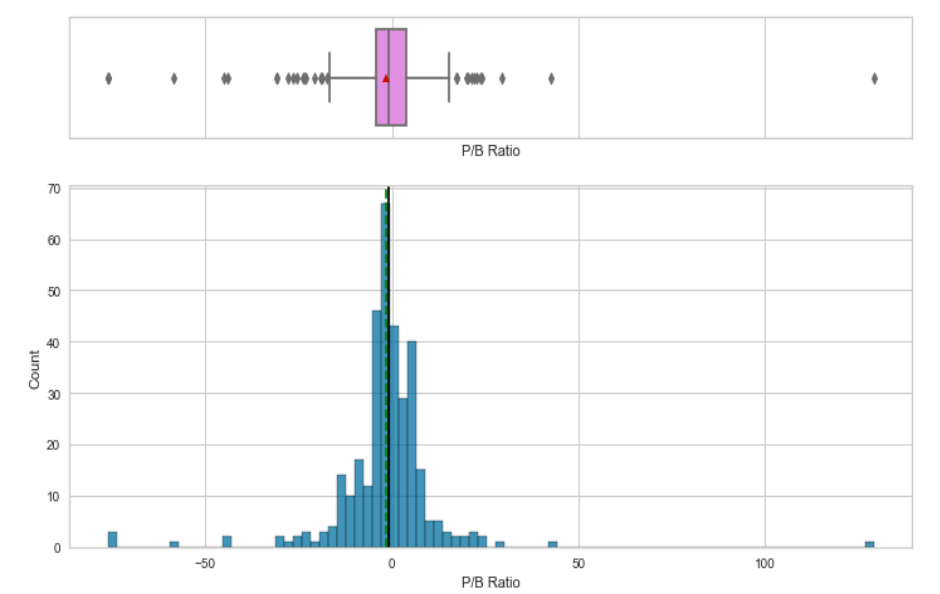

Earnings Per Share
Count
Earnings Per Share


Cash Ratio
Count
Cash Ratio


Net Income
Count
Net Income
1e10


Estimated Shares Outstanding
Count
Estimated Shares Outstanding
1e9





Observations:

Current price

- Distribution --> right skewed
  - 49 of the 340 stocks possessed twice the median value of all stocks
- No stock is listed at less than 0 dollars

- **Price change**
  - Distribution is biased towards lower volatilities and long tails do exist for (+) and (-) price changes
  - Most volatile stocks show as low as a 47% decrease to as high as a 55%
    - This increased over 13 weeks

Volatility

- Distribution of standard deviations is right skewed and not normalized

Cash Ratio / ROE

- Distributions are heavily right skewed and no stock is listed with either metric with a value of less than 0
  - For example, 24 stocks are listed with returns on equity of less than 5 and 25 stocks are listed with returns of over 100 percent

Net Income / EPS

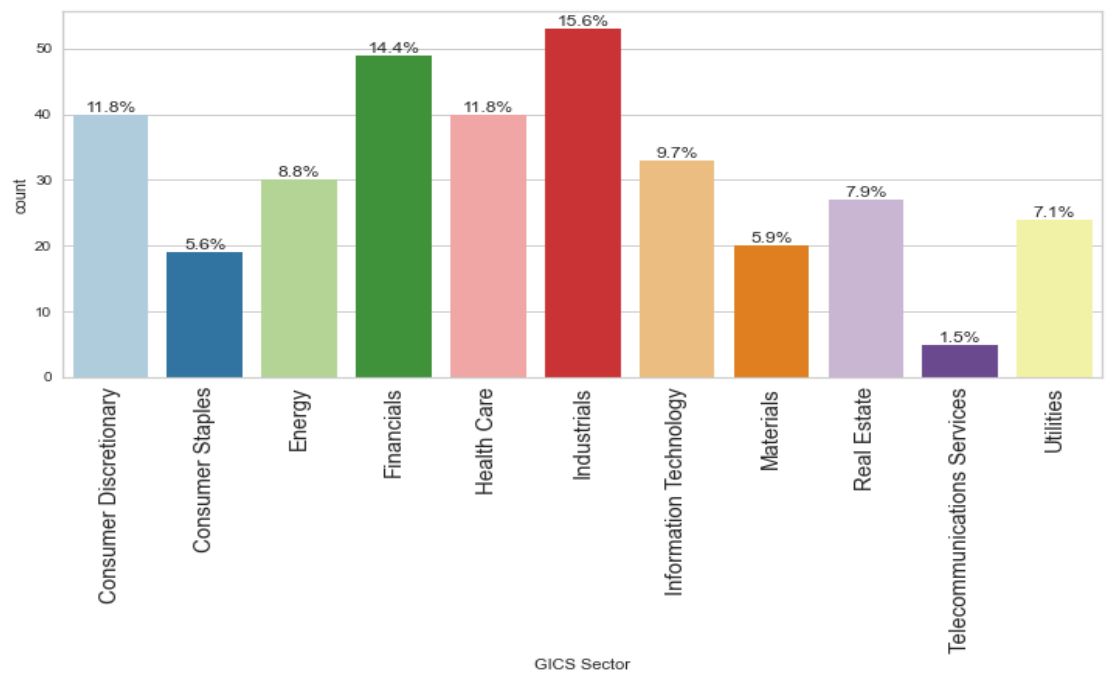
- Net income is shown to be right skewed with both long (+) and (-) tails
  - I.e., most companies generate meager profits, but some are failing and some are significantly successful
- 32 companies within the dataset indicates a net income of less than 0 dollars
- EPS, as a derivative of Net Income, demonstrated a similar distribution

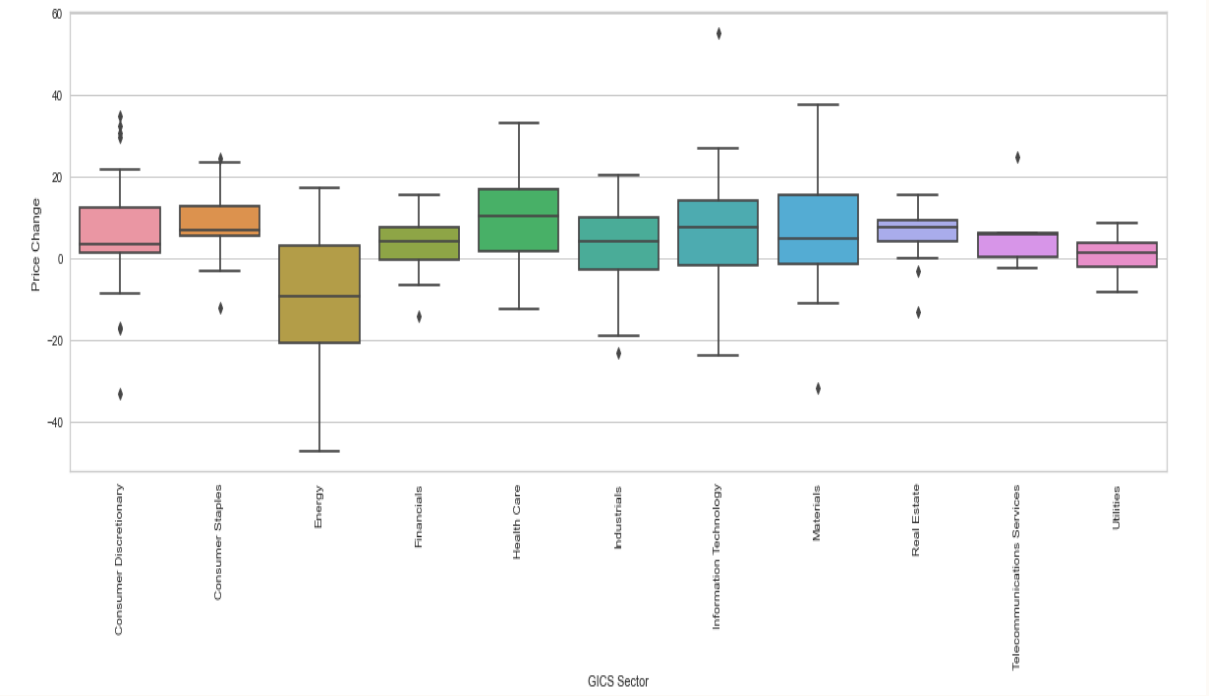
Estimated shares outstanding

- Distribution --> right skewed, but no stock has a value of outstanding shares that is unrealistic

P/E and P/B Ratio

- Distribution of P/E ratios --> right skewed
  - No stock shows a negative ratio, even though several stocks have a negative EPS and no stock consist price listed of less than 0
- Distribution for P/B ratios is mostly centered around 0 but with long positive and negative
  - 175 of the 340 total stocks are shown to below the 25th percentile and above the 75th percentile
  - 31 of the stocks are outliers

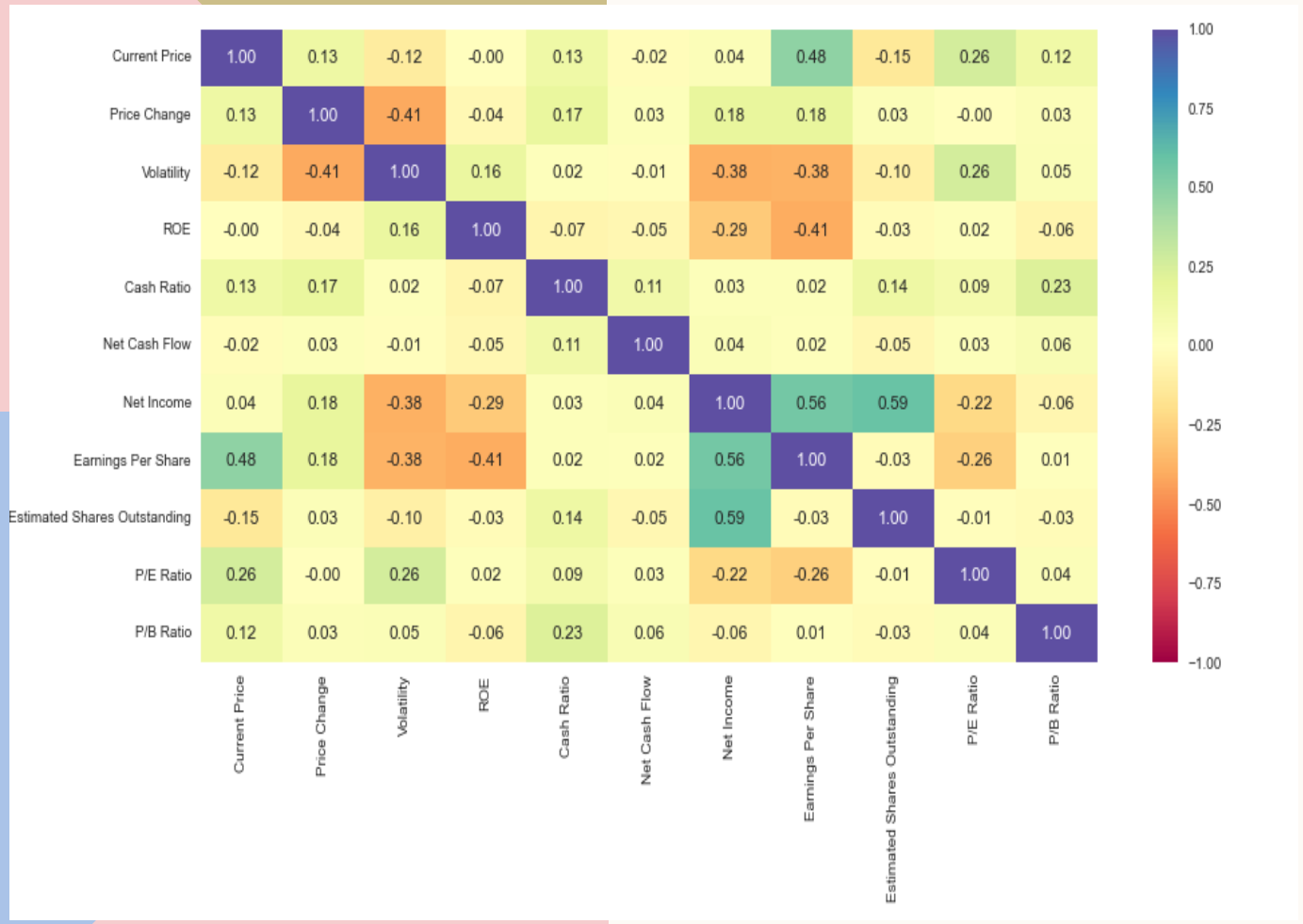




- GICS sector with most stock --> industries sector with 15.6%
- The top 5 --> Industries, Financials, Health care, Consumer Discretionary, and Information Technology
- Telecoms services --> least number of stocks

- Healthcare sector consist greatest price increase (+), closely followed by consumer discretionary GICS sector.
  - Sector with lowest (+) price change is telecoms services
  - Energy sector has showed a great amt. (-) price change
- Real estate consist of least variation in Price Change across different companies.
  - Energy GICS_Sector --> most variation

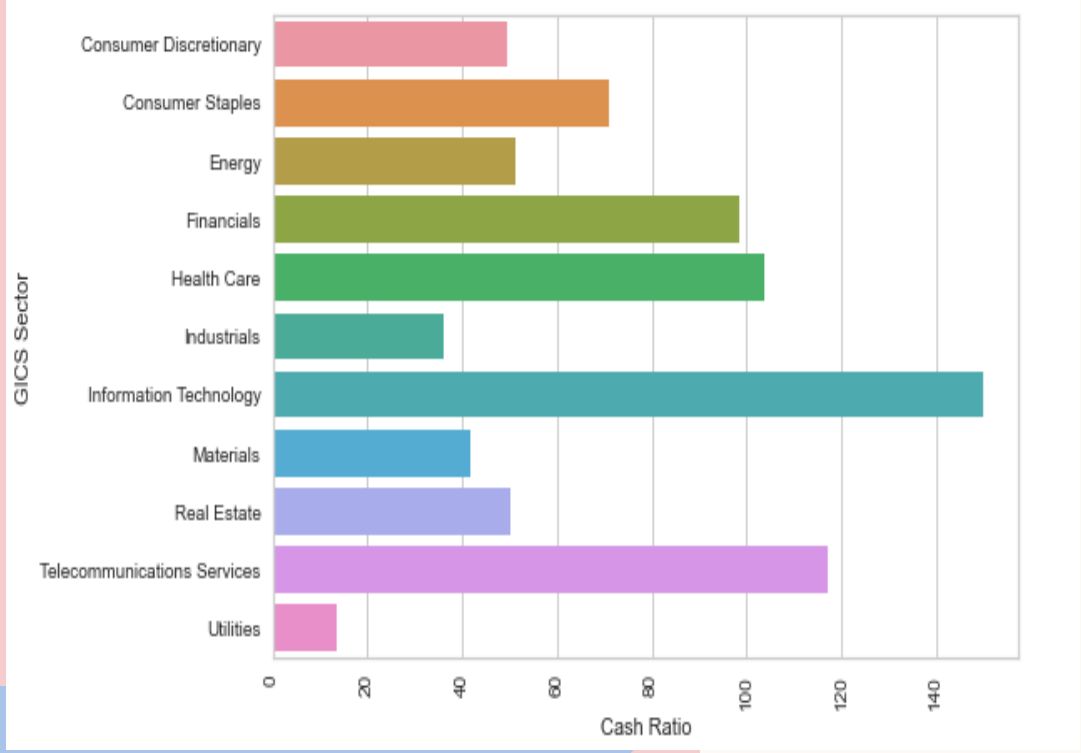
EDA: BIVARIATE ANALYSIS

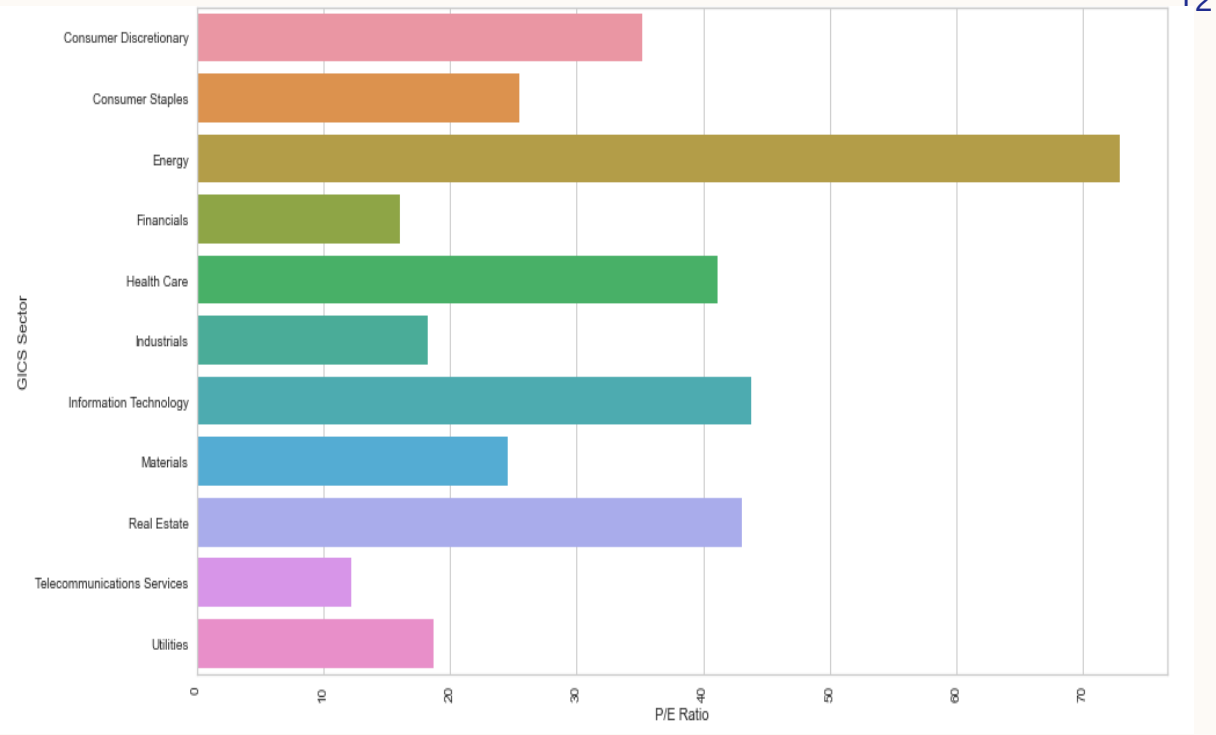


Observations:

Most variables are moderately correlated (+/- .40) in respect to one another

- Volatility is (-) correlated with price change
  - As stock becomes more volatile --> Dec. in price
- Net income is (-) correlated with volatility
  - As company generates higher net income, the price is likely less volatile
- Net income is (+) correlated with earnings per share (EPS) and estimated shares outstanding
- EPS is (+) correlated with current price
  - As a company's EPS rises --> Inc. prices
- EPS is (-) correlated with ROE
  - As company generates more equity for shareholders then equivalent amount of net income following periods will generate a lower return



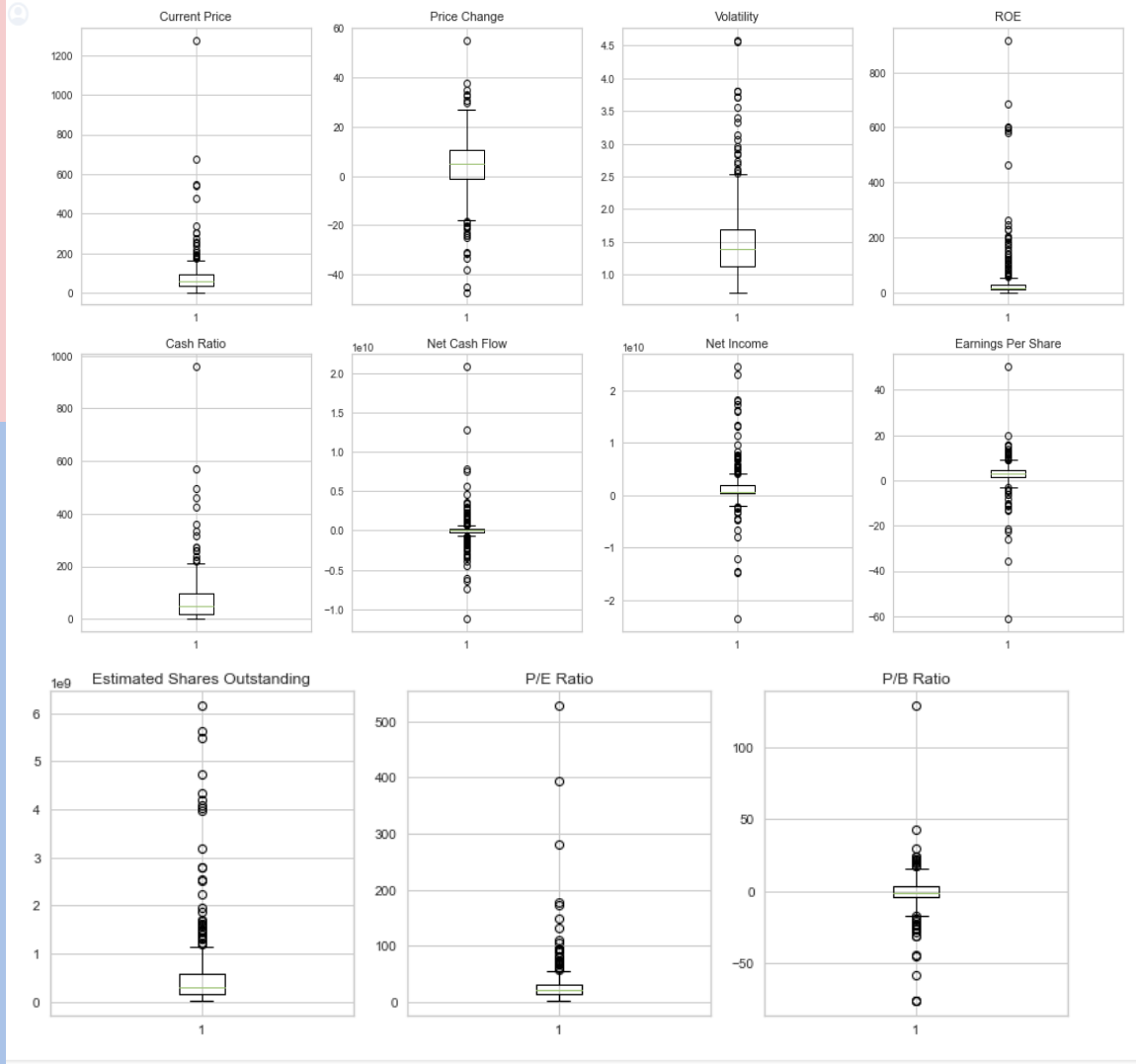


- The GICS sector w/greatest cash ratio on avg. is information technology then telecoms services
- Sector with the lowest cash ratio is utilities sector with cash ratio below 20

- GICS sector with highest P/E ratio is Energy sector while telecom services has the lowest ratio

DATA PREPROCESSING

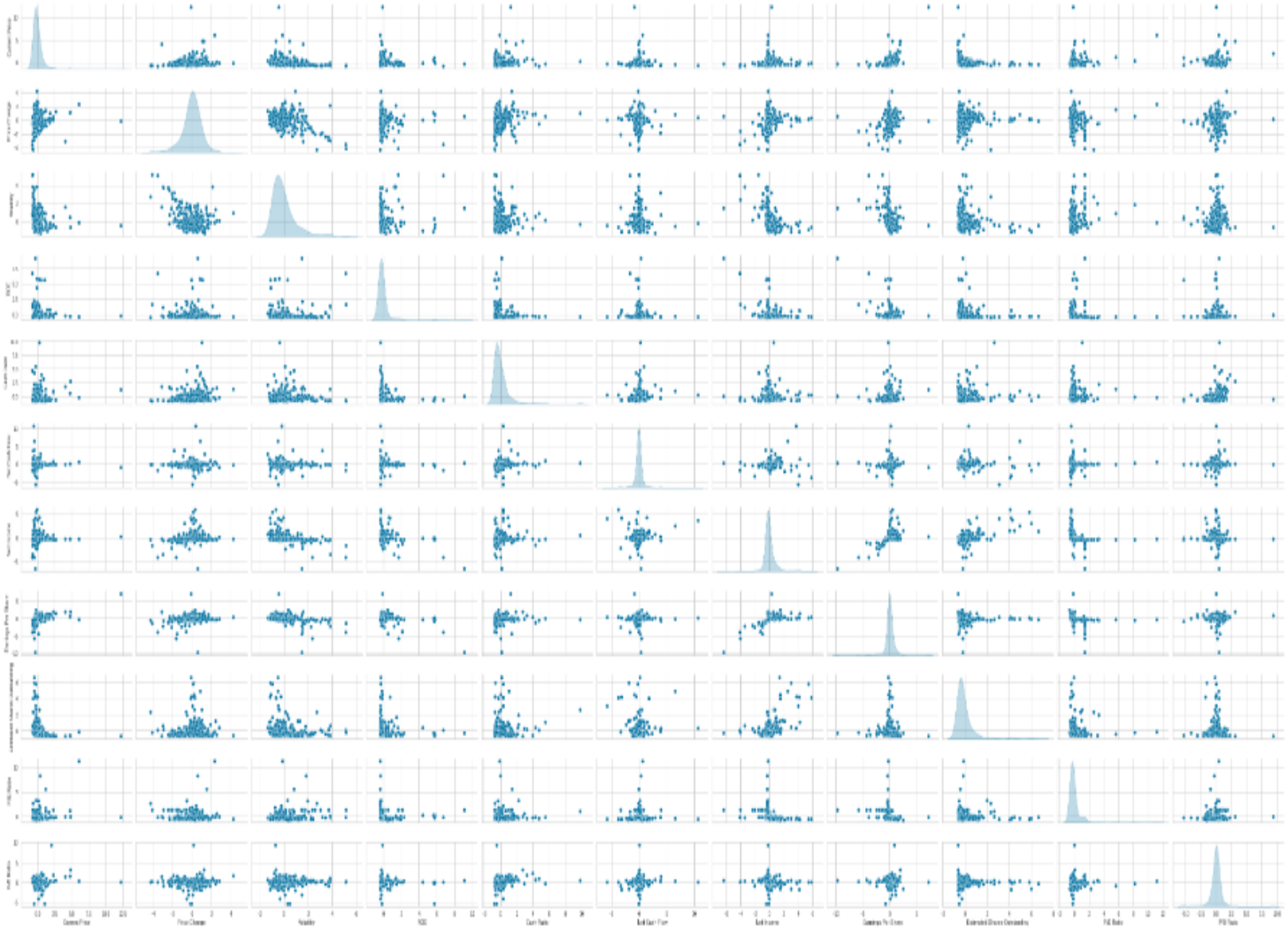
OUTLIERS



Observations:

- Variables in the dataset consist of outliers
- Variables such as current price, cash ratio, estimated shares outstanding, P/E ratio and ROE consist of outliers w/n (+) distribution,
  - Others consist of outliers w/n (+) & (-) distribution
  - Thus, outliers will not be treated

Scaled Dataframe

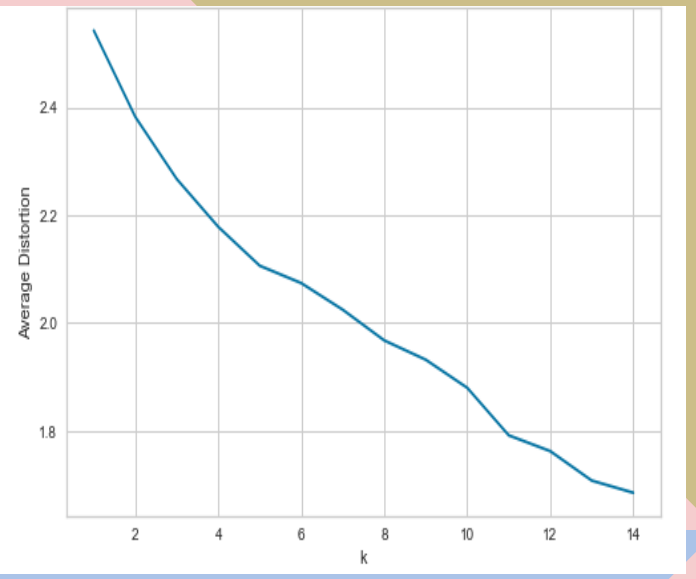


Observations:

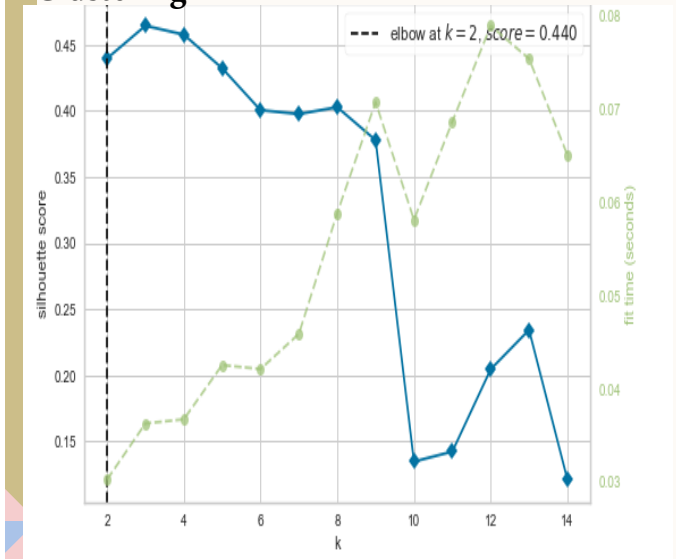
- Distribution and variation between within the variables were not affected
- All variables reside within same scale with mean of 0 and STD of 1

K-MEANS CLUSTERING

K with the Elbow Method

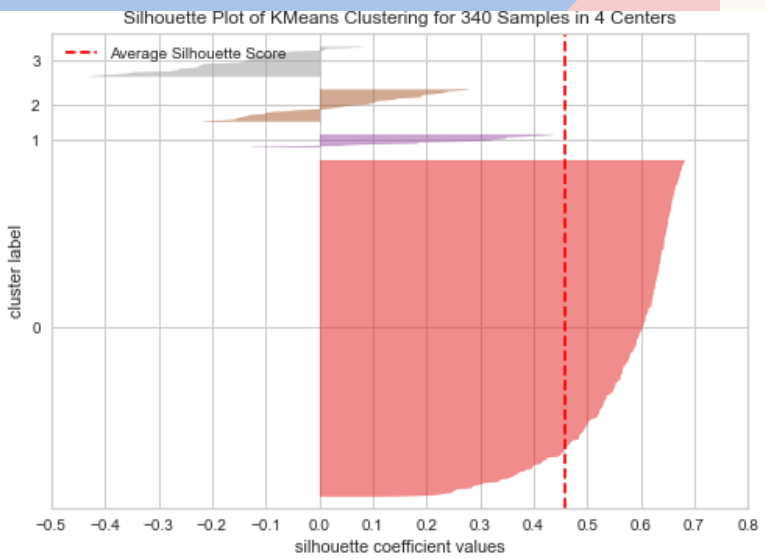


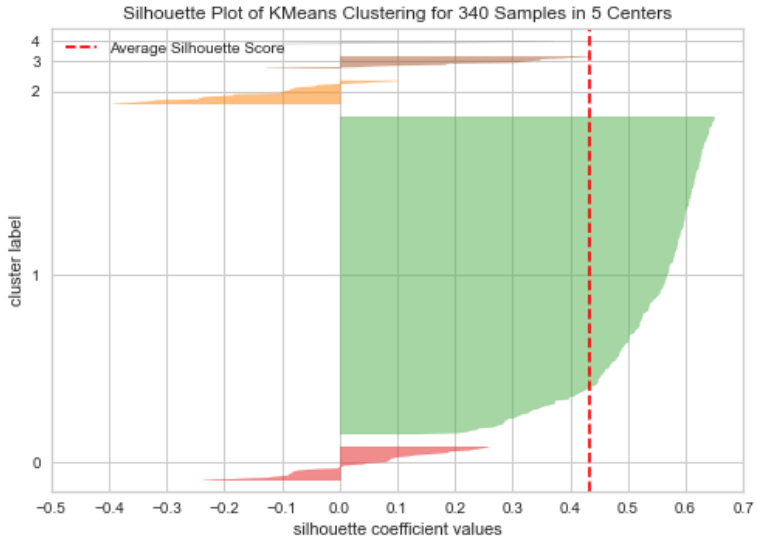
Silhouette Score Elbow for KMeans Clustering

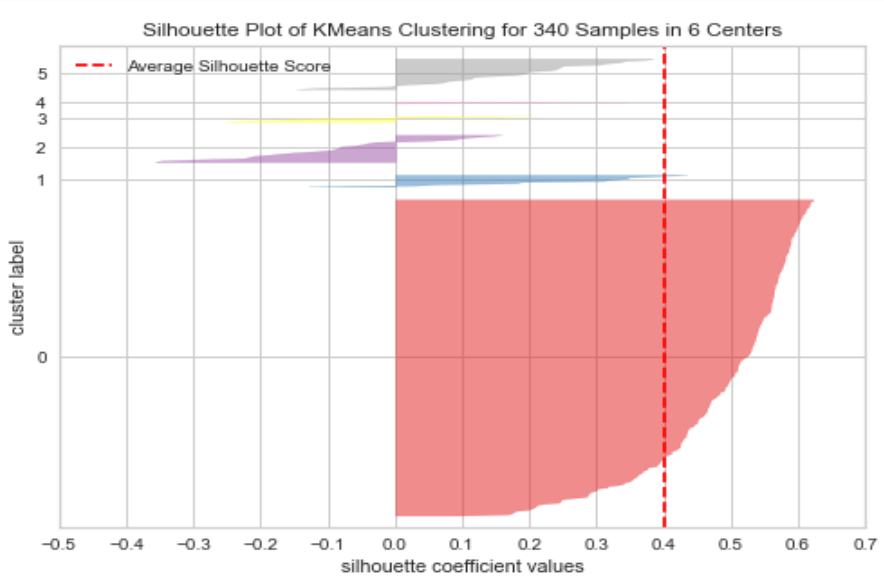


Observations:

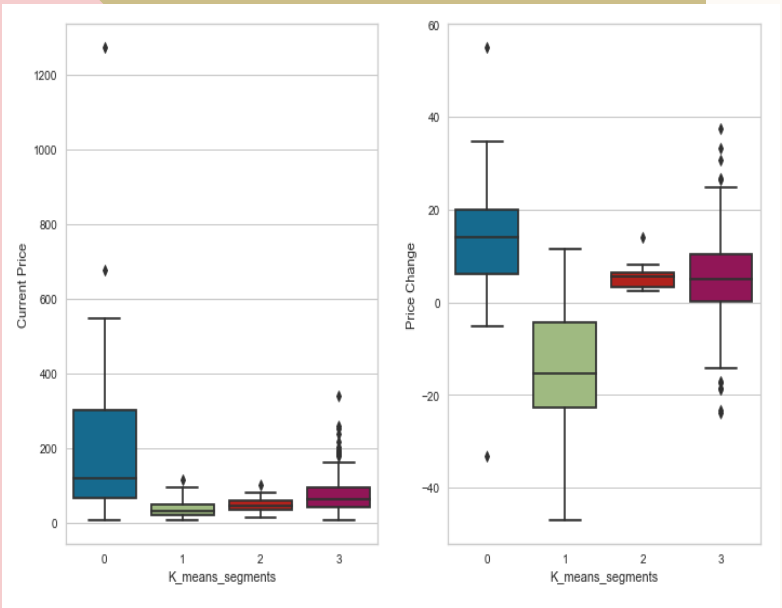
- Based on elbow plots and the distortions, #of cluster choices is best around 4, 5 and 6 clusters
- Best silhouette score is 3 but not too good in the elbow plot
- 5 and 6 --> didn't yield good silhouette scores
  - Of the 3 choices based on elbow plot (4,5 and 6) --> 4 shows best silhouette score, then is the best option
- Thus, based on Elbow and Silhouette plots -->4 shows the number of clusters with good performance

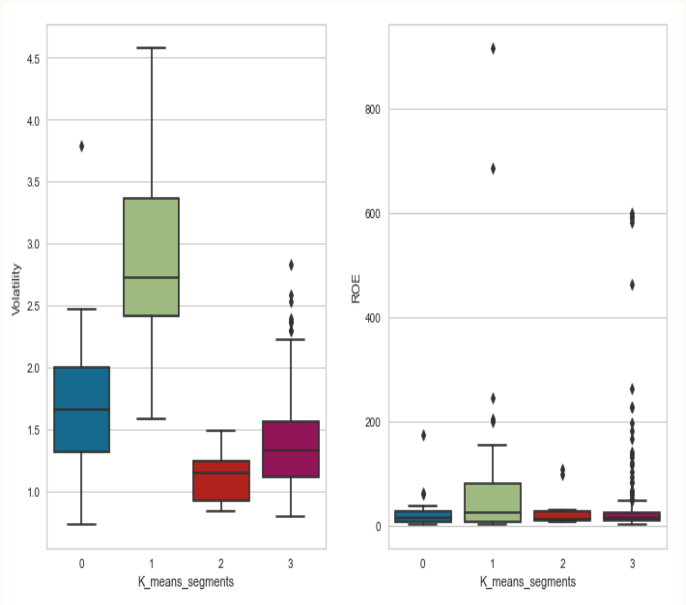


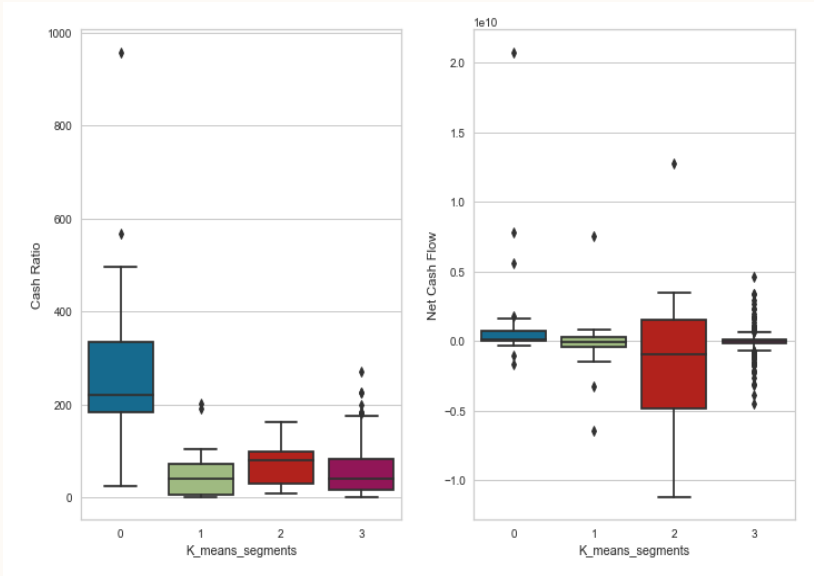


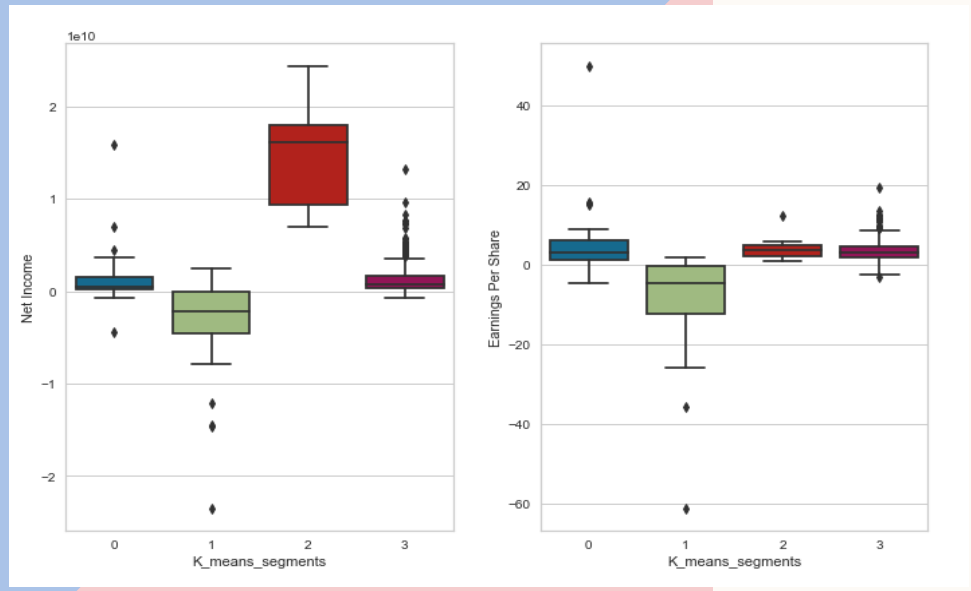


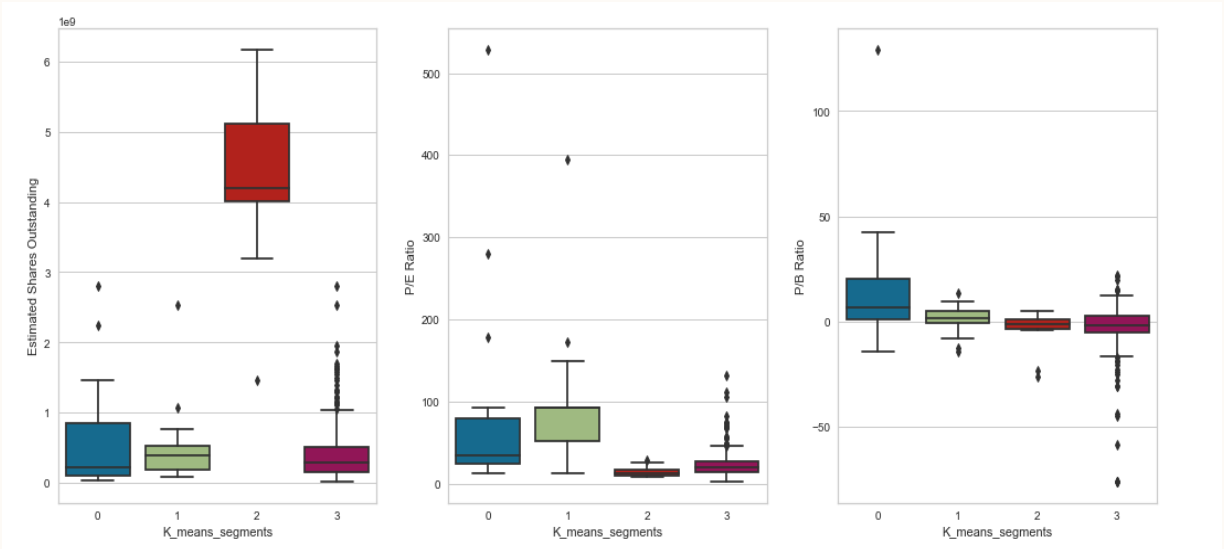
BOXPLOT OF NUMERICAL VARIABLES PER CLUSTER











OBSERVATIONS

- Cluster 0 has the highest current price and variance in current price
  - Lowest current price with little variance is cluster 1
- Cluster 0 has highest price change but cluster 1 has highest variance in price change
- Cluster 1 has highest volatility and most variance in volatility
- Cluster 1 has highest ROE and shows most variance in ROE
- Cluster 0 has highest cash ratio and shows the highest variance
  - In Net cash flow, cluster 2 --> highest variance
- Cluster 2 has highest net income and most variance of the 4 clusters
- Cluster 1 has lower earnings per share than others, and the most variance
  - Cluster 0 has the largest earnings per share not higher than clusters 2 and 3
- Cluster 2 consist of highest estimated shares outstanding
- Cluster 1 -->highest P/E ratio on average
  - Cluster 0 consist the largest variance
- Cluster 0 -->highest P/B ratio and variance while cluster 2 has lowest variance

K-MEANS CLUSTERING INSIGHTS

Cluster 0 - Large Market Capitalization / Dow Jones Industrial Average

- 11 stocks --> stocks within the Financials, Health Care, Information Technology (IT), and Consumer Discretionary sectors
- Companies' w/n this cluster consist of:
- Low volatility
  - Most of the companies with the highest outflows of cash
  - The highest net incomes
  - The highest number of shares outstanding

- **Cluster 1 - "Cash is King"**
- 13 stocks --> stocks within the Healthcare and IT sectors
- Companies within this cluster have:
  - Moderate volatility
  - Mostly profitable
  - Most of the highest cash ratios and cash inflows
- **Cluster 2 - S&P 500 / Diversification**
- 280 stocks (~84% of all stocks in the dataset) drawn from all sectors present in the dataset
- Companies within this cluster have:
  - Low P/E ratios
  - Most of the outliers on negative P/B ratios

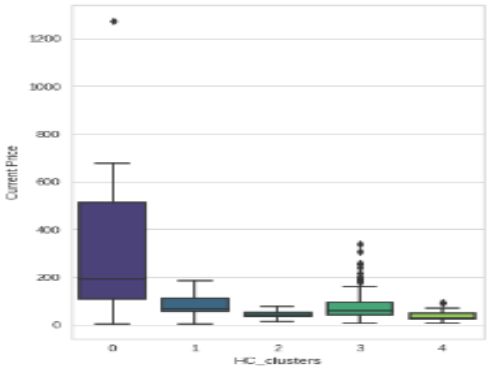
Cluster 3 - "Ride the Energy Rollercoaster" portfolio / Growth mindset

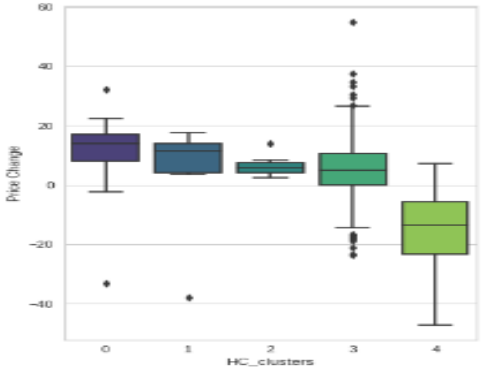
- 29 stocks, -->The Energy sector
- Companies within this cluster consist of:
  - Low stock prices and high ROE
  - High beta
  - Most of volatile stocks with outliers in price decreases
  - Mostly (-) net incomes and earnings per share

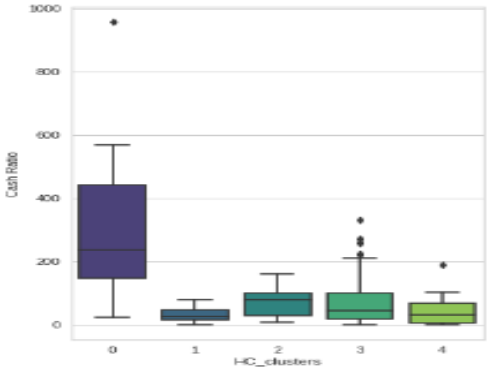
Cluster 4 - High Earnings for a High Price

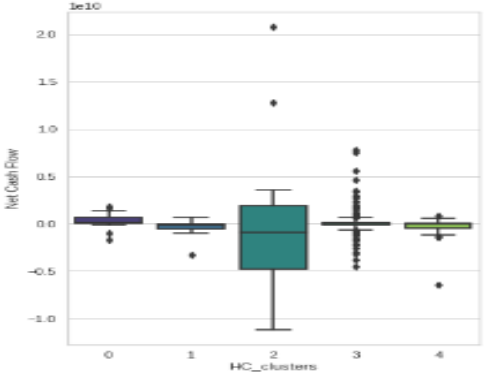
- 7 stocks --> Stocks from the Health Care and Consumer Discretionary sectors
- Companies' w/n this cluster consist of:
  - Stocks with the highest prices
  - Favorable cash ratios
  - Most favorable P/B ratios
  - Most of the highest earnings-per-share

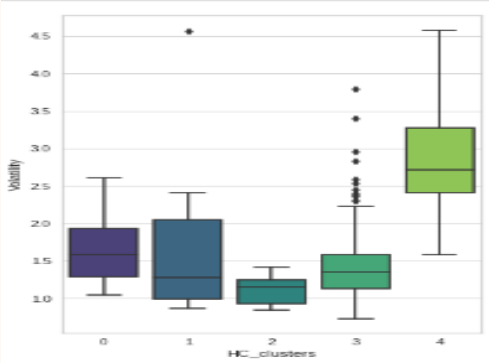
HIERARCHICAL CLUSTERING

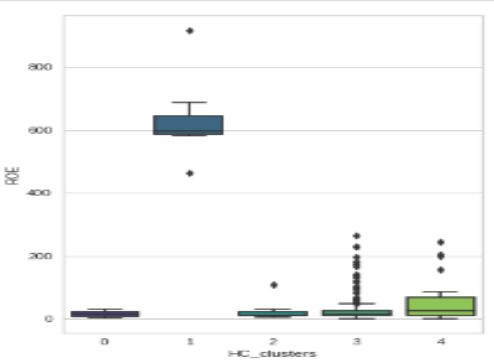


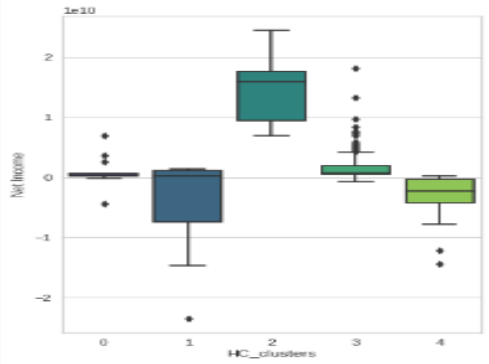


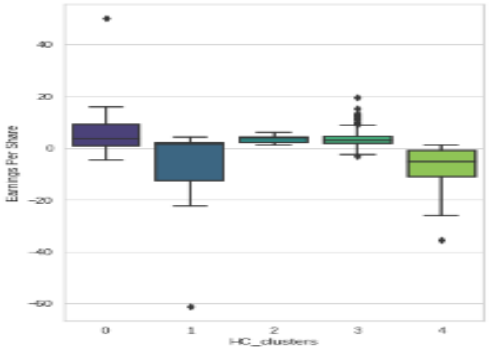


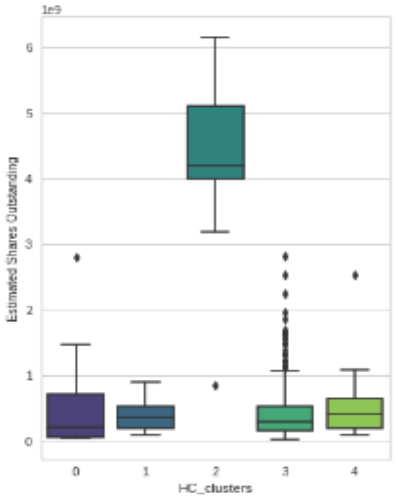


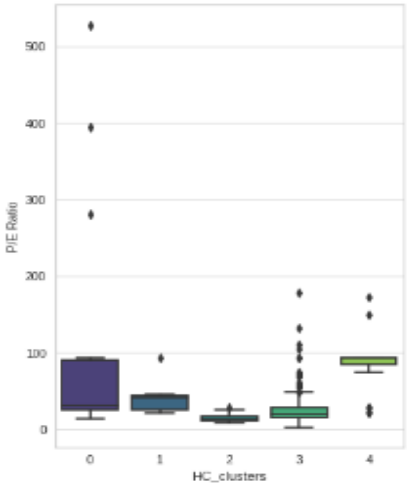


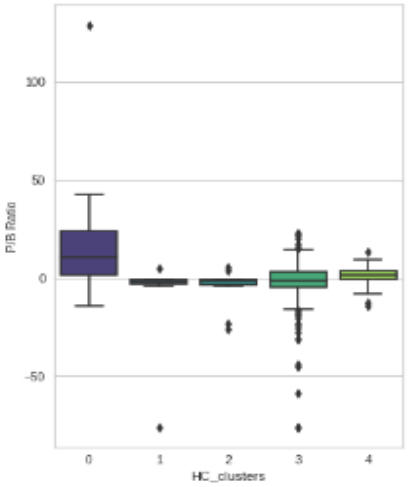












OBSERVATIONS

Cluster 0 - Growth for a Price

- 15 stocks --> within Health Care, Information Technology (IT), and Consumer Discretionary sectors
- Companies within this cluster consist of :
  - Most of stocks in highest prices
  - Significant outliers in price-to-equity ratio
  - Most favorable price-to-book (P/B) ratios
  - Most of highest cash ratios

Cluster 1 - Short-term Poor, Long-term Rich

- 7 stocks --> within the Consumer Staples and Energy sectors
- Companies within this cluster consist of :
  - Highest returns-on-equity
  - Lowest net incomes
  - Mostly negative earnings per share

Cluster 2- DJIA

- 11 stocks --> within the Financials and Telecommunications sectors
- Companies within this cluster have:
  - Companies with the highest inflows and outflows of cash
  - Highest net incomes
  - Highest number of shares outstanding

Cluster 3 - Diversification

- 285 stocks (~84% of all stocks in the dataset) drawn from all sectors present in the dataset
- Companies within this cluster consist of:
  - Most of outliers in price increases and some of the outliers in price decreases
  - Some of outliers in cash inflows and outflows
  - Most of the outliers in P/B ratio

Cluster 4 - Energy-specific portfolio

- 22 stocks, a vast majority of which are in the Energy sector
- Companies within this cluster consist of:
  - Most of the most volatile stocks, especially those with outliers in price decreases
  - Mostly (-) net incomes and earnings per share

K-MEANS VS HIERARCHICAL CLUSTERING

Which clustering technique took less time for execution?

- The clustering technique that took less time are KMeans model and Agglomerative Clustering model that fit the dataset within ~0.1s

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?

- Both algorithms yield identical clusters, with a single cluster of majority of the stocks and remaining four clusters consisting of 7-29 stocks

How many clusters are obtained as the appropriate number of clusters from both algorithms?

- For both algorithms, 5 clusters provided different clusters with enough observations in each to distinguish which "type" of stock is representative of the cluster

Differences or similarities in the cluster profiles from both the clustering techniques

- Both algorithms yield similar clusters based on the outliers within the 11 variables

CONCLUSIONS

- Trade&Ahead should identify the financial goals, risk tolerance, and investment behaviors of their clients
  - This is can be achieved via survey conduction in which the suitable cluster can be identified
- A lot of the clusters based on characteristics of the stocks are essentially substitutes for standard indexes, such as Dow Jones Industrial Average and S&P 500
- The energy company displayed a high volatility on stock prices which becomes a risk when investing
- Utilities sector demonstrated a low risk invest due to low volatility on stock prices
- IT, telecommunication services, and healthcare demonstrated the highest cash ratio based on the tendency to cover short term obligations with cash value
- Thus, the energy industry shows very high P/E ratio within the distribution compared to other companies

RECOMMENDATIONS

- Trade & Ahead could use these clusters as a starting point for further financial statement analysis to identify which individual stocks do not fit the "profile" of the cluster
- Further investigation through financial statement analysis to understand the best fits for the clients is recommended
- Trade & Ahead would need to understand and clarify the characteristics of its clients in terms of financial goals and position, risk appetite, and investment time/duration via surveys conduction
- Trade & Ahead should recommend to its clients to invest in clusters that consist of highly volatile securities for those with interest to make payoffs in shorter time or purchase stocks at cheaper price
- The firm can also recommend to clients that prefer investments with moderate price, risks and profits to invest in safe clusters such as cluster 3 (both k-means and cluster 0)
- Cluster 3 should be recommended to clients that prefer diversified portfolio
- For the clients looking for high yield and companies with good track record with long term profits should be advised to invest in cluster 2
- For those searching for investments with larger companies with strong brands and requires consistent and regular dividend payements should be recommended to utilized cluster 0 in k-means and cluster 1 in HC algorithms
- Moreover, Trade & Ahead, need to make sure that the clustering exercise is performed regularly and updated due to volatility of stocks and newly introduced securites

THANK YOU!!!