TRADE & AHEAD PROJECT

COURSE TITLE: UNSUPERVISED LEARNING

DATE: AUGUST 15TH, 2022

CONTENTS

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

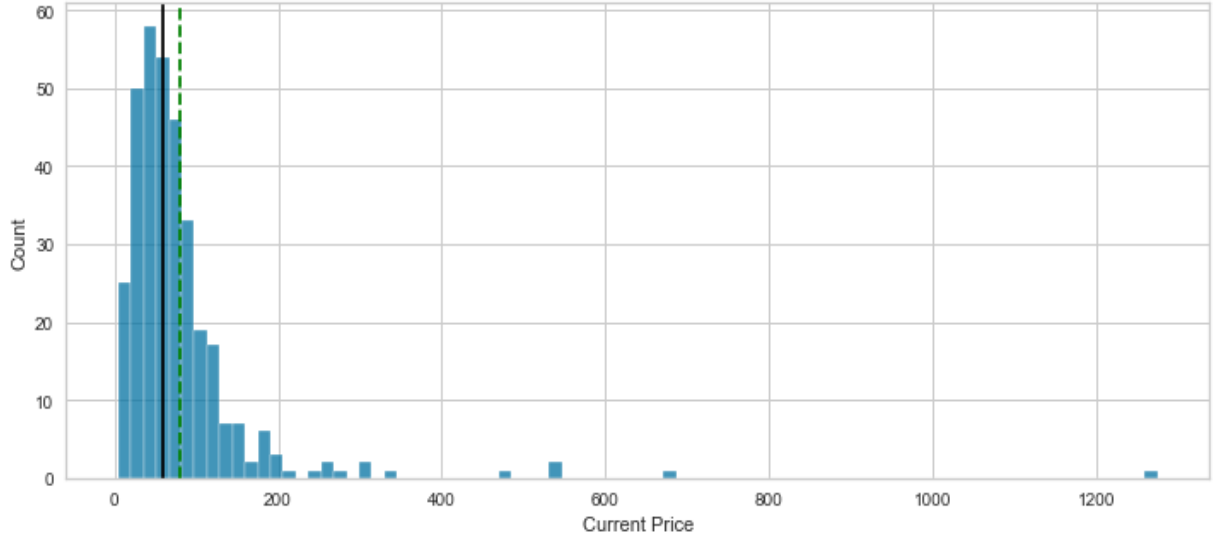
- The stock market has consistently proven to be a good place to invest in and save for the future
- There are a lot of compelling reasons to invest in stocks; It can help in fighting inflation, create wealth, and provide some tax benefits. Also, the power of compound interest, the earlier one starts investing, the larger one can have for retirement.
- Having a diversified portfolio tends to yield higher returns and face lower risk by tempering potential losses when the market is down.
- Getting lost in a sea of financial metrics to 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, identifying 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.
- To analyzing the data, group the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

DATA OVERVIEW

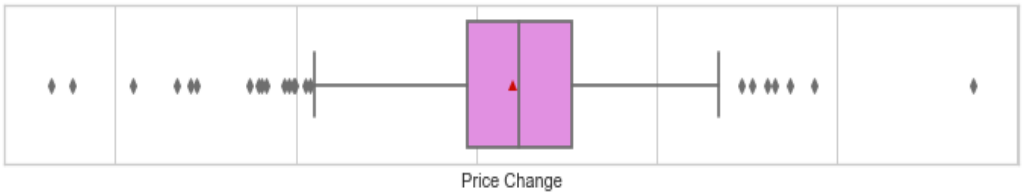
- The data set consist of 340 rows and 15 columns
- we have nine (4) object type, 7 float and 4 integer
- there are no missing or duplicated value in the dataset
- The average current price is 81
- the average price change is 4
- The average volatility rate 1.5
- The average ROE is 39
- the average cash ratio is 70
- The average net cash flow is 55,537,520
- The average net income is 1,494,384,602
- The average earnings per share is 2.7
- The average estimated shares is 577,028,337
- The average P/E Ratio is 32
- The average P/B Ratio is -1.7

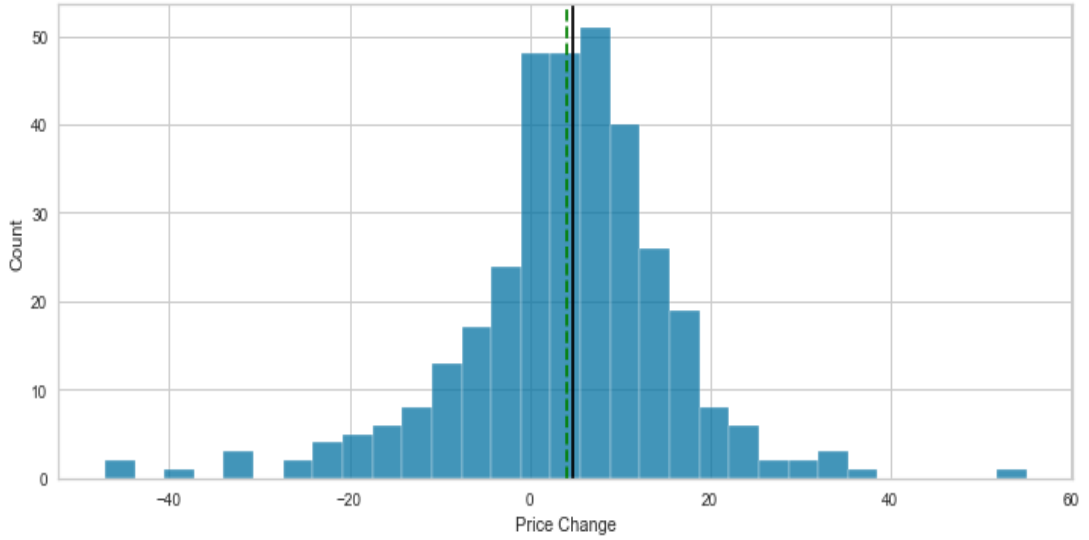
EXPLORATORY DATA ANALYSIS(EDA)

- The distribution for current price is rightly skewed
- The boxplot shows that there are a lot of upper outliers to the right for this variable.
- The average current price is higher than the median for current price indicating the distribution is skewed to the right



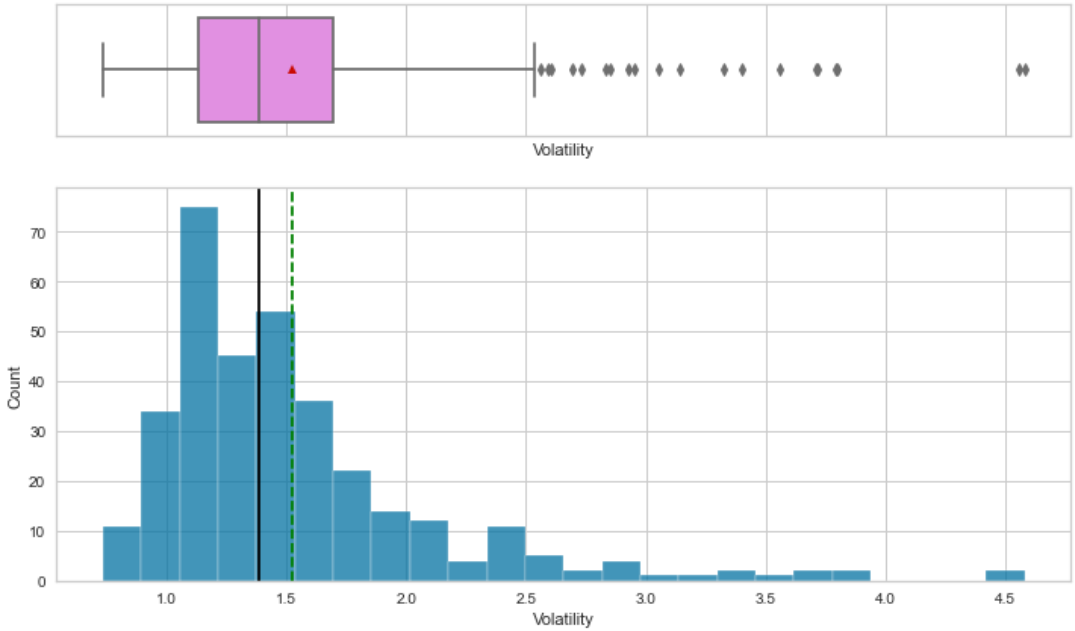
- The average price change is almost the same with the median indicating the distribution is nearly symmetrical
- There are outliers on both sides for this distribution
- The average price change is almost the same as the median price change



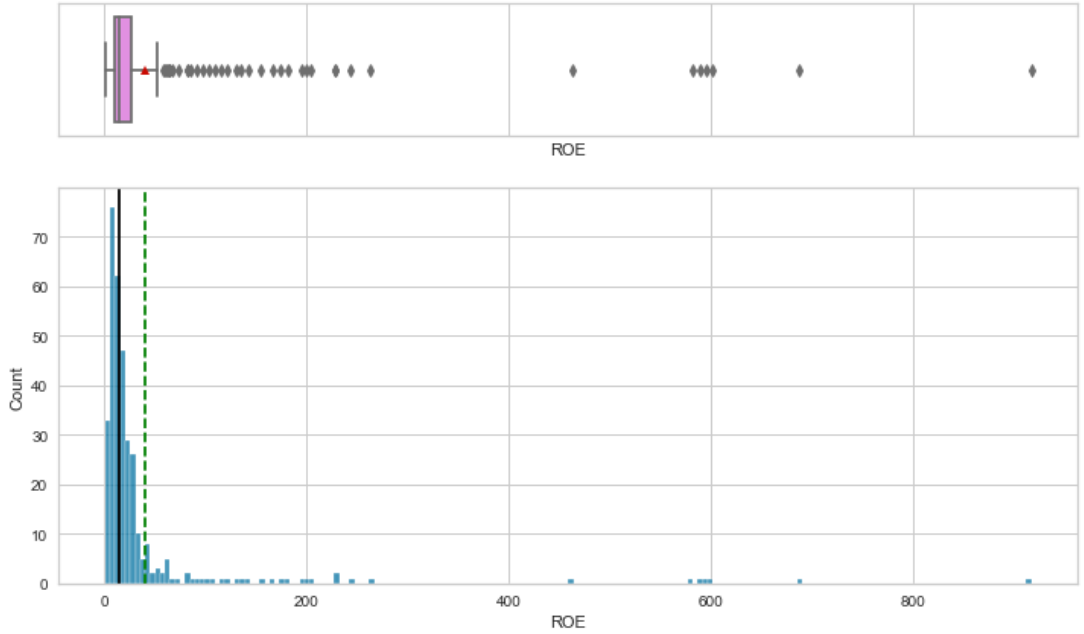


EXPLORATORY DATA ANALYSIS(EDA)

- The distribution for volatility is rightly skewed
- The boxplot shows that there are a lot of upper outliers to the right for this variable.
- There is a significant difference in the average and median volatility

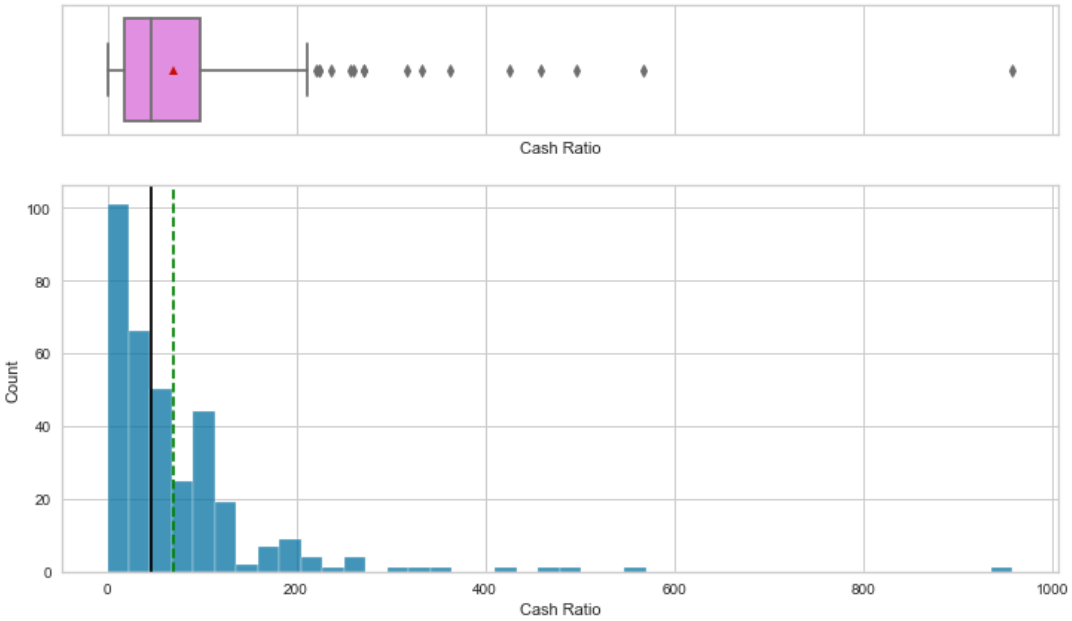


- The distribution for ROE is highly skewed to the right
- The boxplot shows that there are a lot of upper outliers to the right for this variable.
- The average ROE is higher than the median ROE

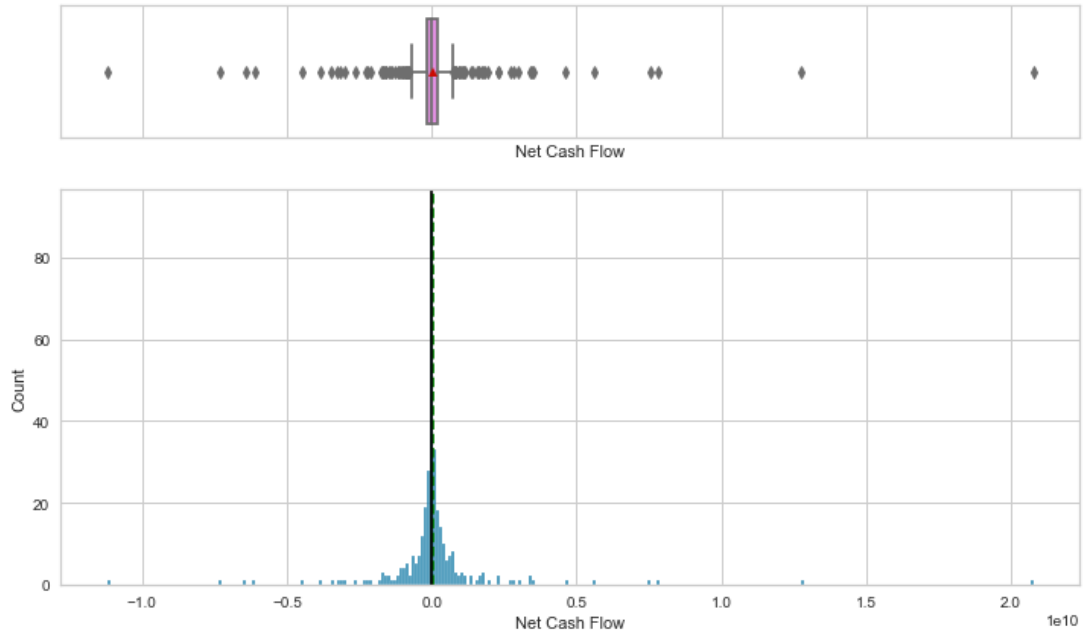


EXPLORATORY DATA ANALYSIS(EDA)

- The distribution for cash ratio is rightly skewed
- The boxplot shows that there are a lot of upper outliers to the right for this variable.
- There is a significant difference in the average and median for cash ratio

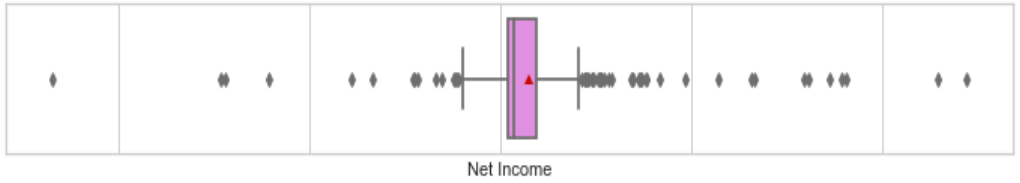


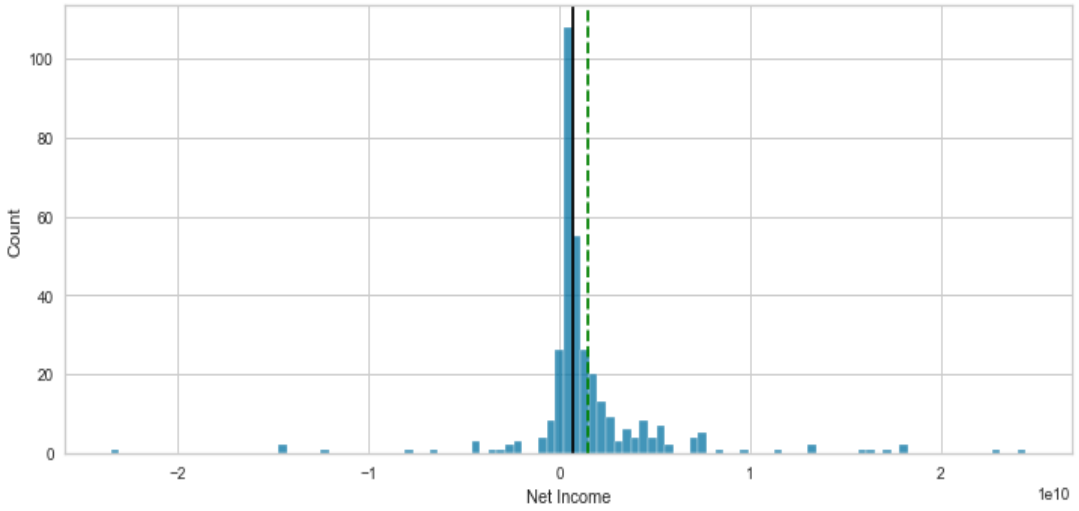
- The average net cash flow is almost the same with the median indicating the distribution is nearly symmetrical
- There are outliers on both sides for this distribution



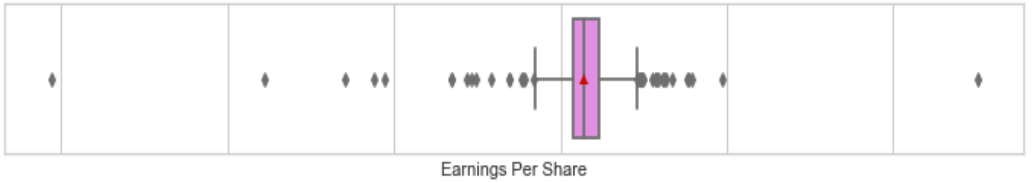
EXPLORATORY DATA ANALYSIS(EDA)

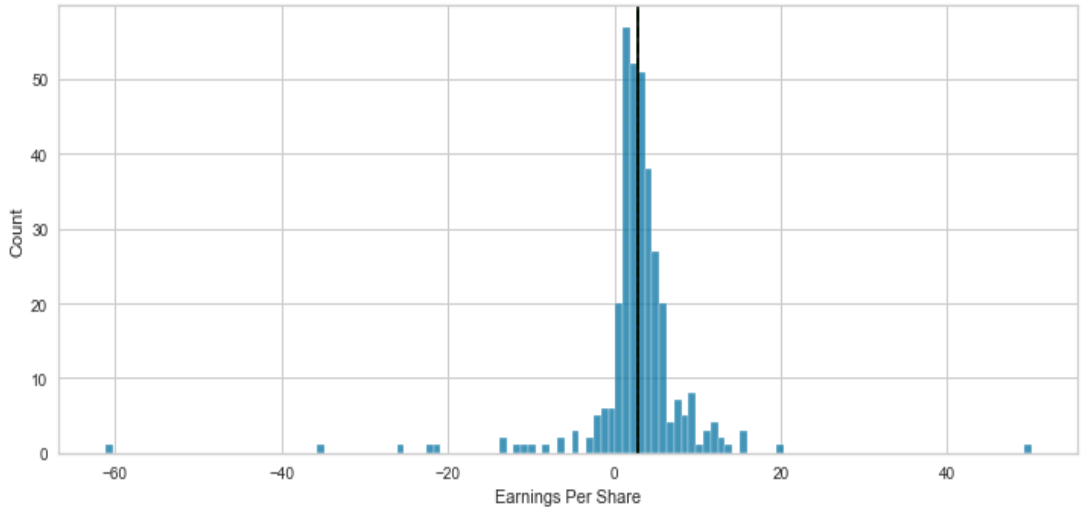
- The average net income is almost the same with the median indicating the distribution is almost symmetrical
- There are outliers on both sides for this distribution





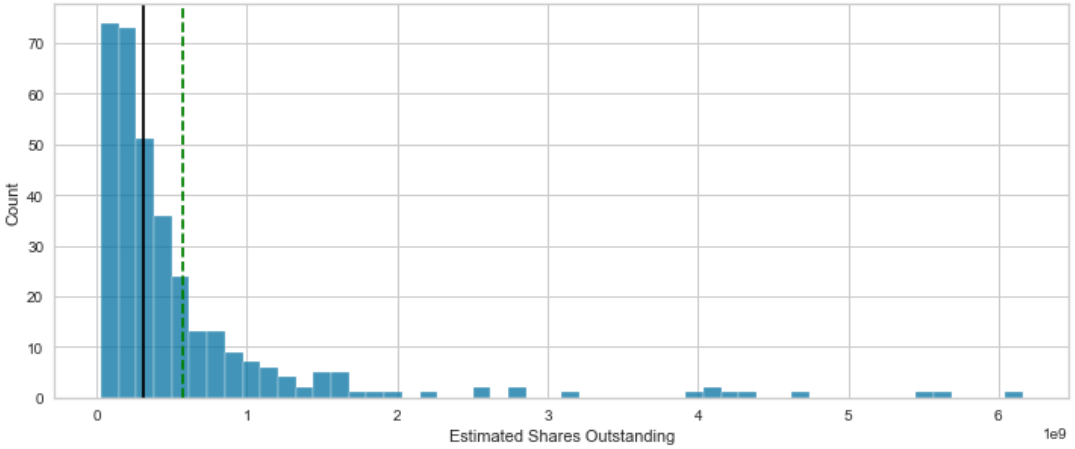
- The average earnings per share is the same with the median indicating the distribution is symmetrical
- There are outliers on both sides for this distribution



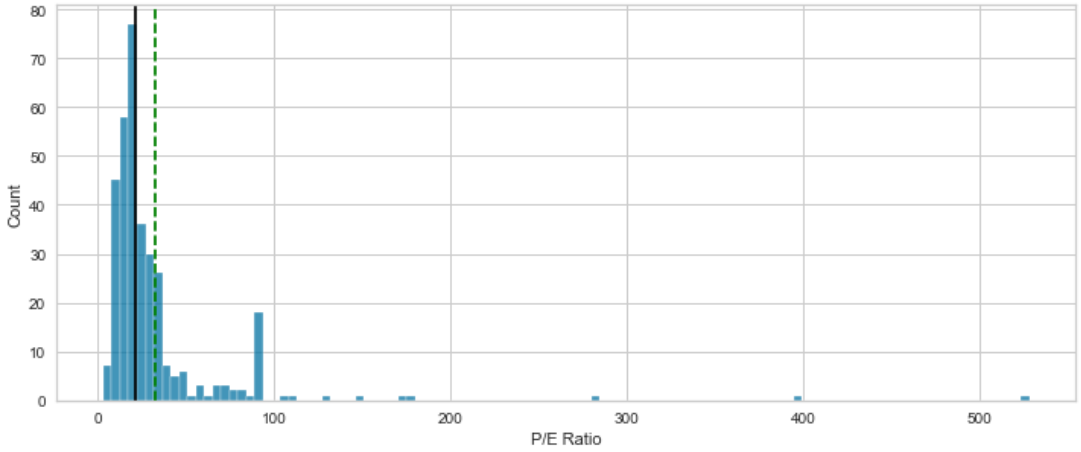


EXPLORATORY DATA ANALYSIS(EDA)

- The distribution for estimated shares outstanding is rightly skewed
- The boxplot shows that there are a lot of upper outliers to the right for this variable.
- There is a significant difference in the average and median for estimated shares outstanding

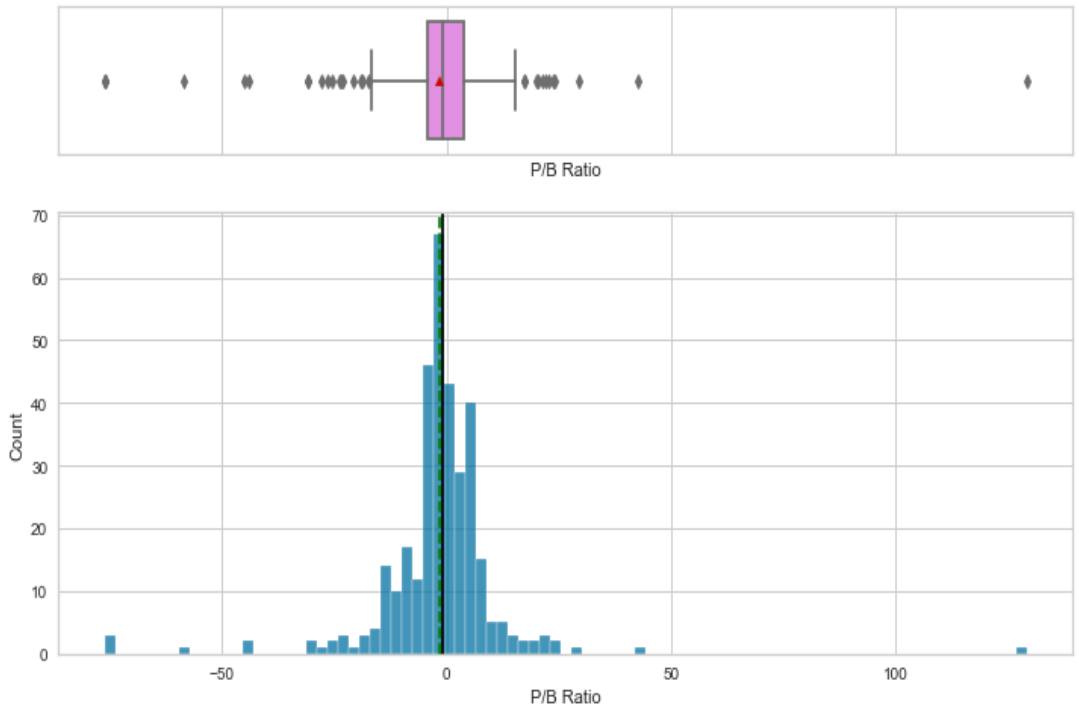


- The distribution for P/E Ratio is rightly skewed
- The boxplot shows that there are a lot of upper outliers to the right for this variable.
- There is a significant difference in the average and median for P/E Ratio

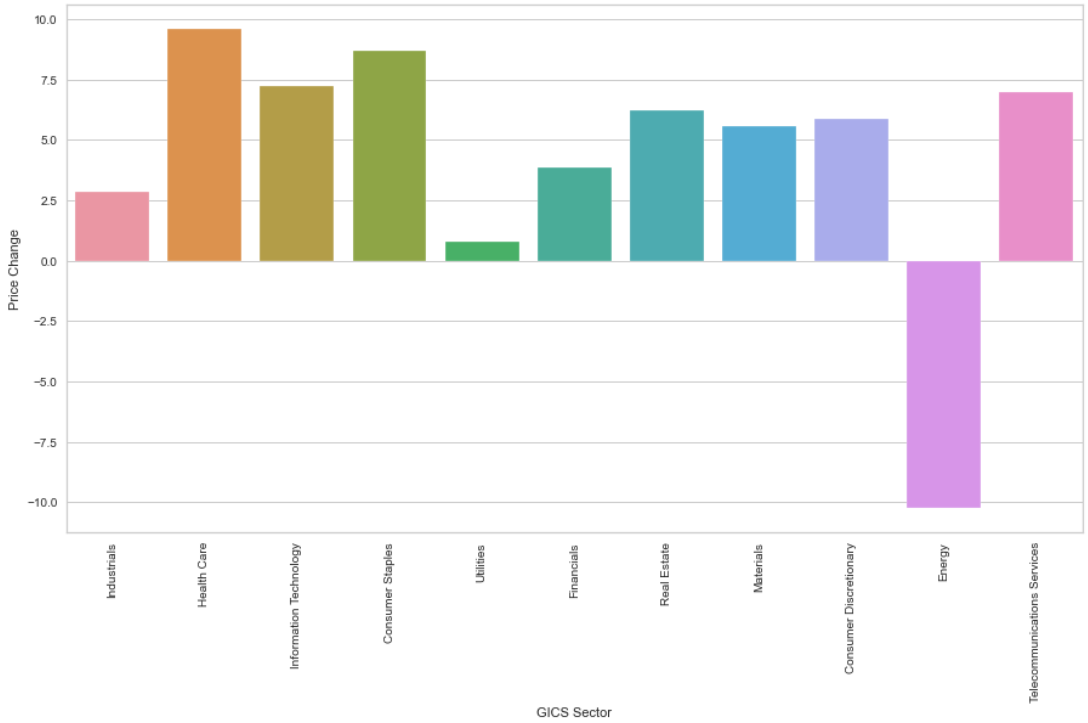


EXPLORATORY DATA ANALYSIS(EDA)

- The average P/B Ratio is almost the same with the median indicating the distribution is nearly symmetrical
- There are outliers on both sides for this distribution

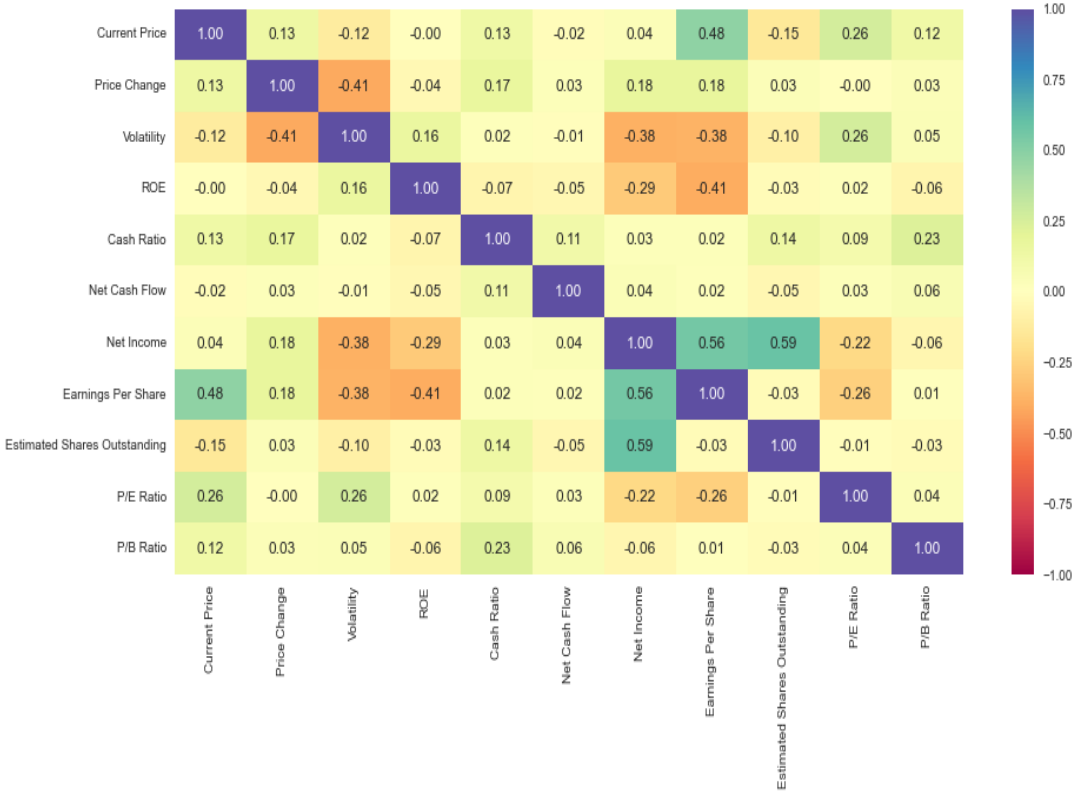


- Healthcare, consumer staples followed by information technology has seen the maximum stock price increase
- Energy has the least stock price change in 13 weeks

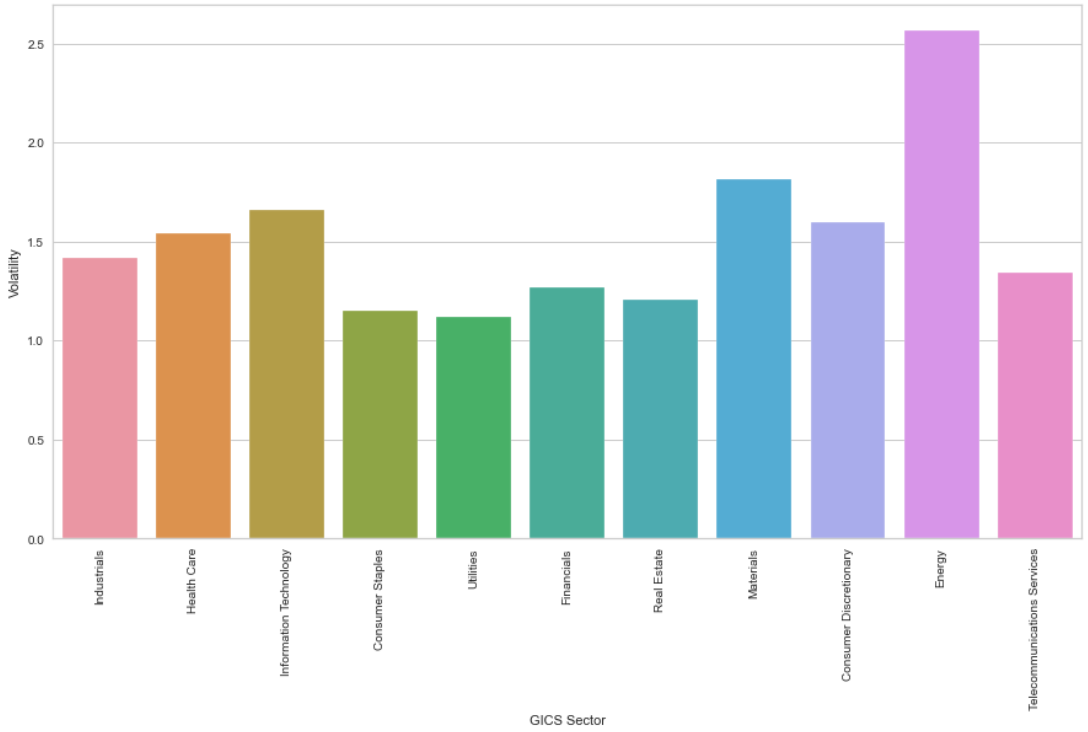


EXPLORATORY DATA ANALYSIS(EDA)

- The heat map distribution shows there is a high correlation between net income and earnings per share and estimated shares outstanding
- The distribution also shows little correlation between price change and current price
- There is no correlation between volatility and price change but shows little correlation with P/E Ratio
- The current price also indicates a high correlation with earnings per share

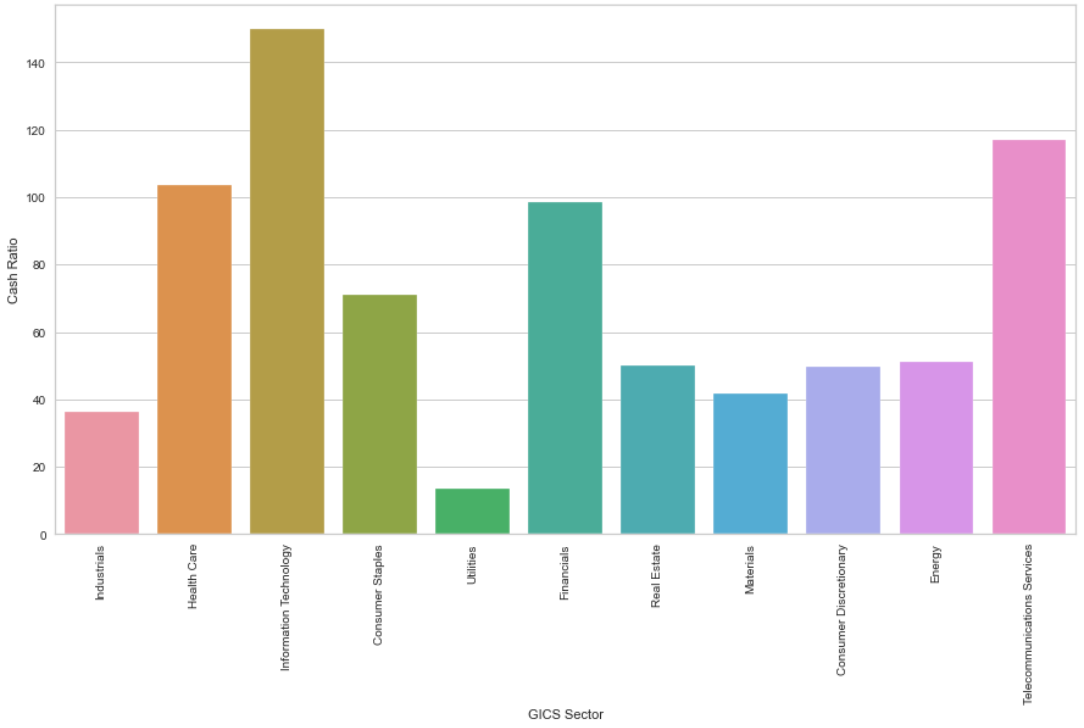


- The energy industry shows a high volatility on stock prices which becomes risker to invest followed by material industry
- The utilities sector presents a low risk invest as a result of low volatility on stock prices

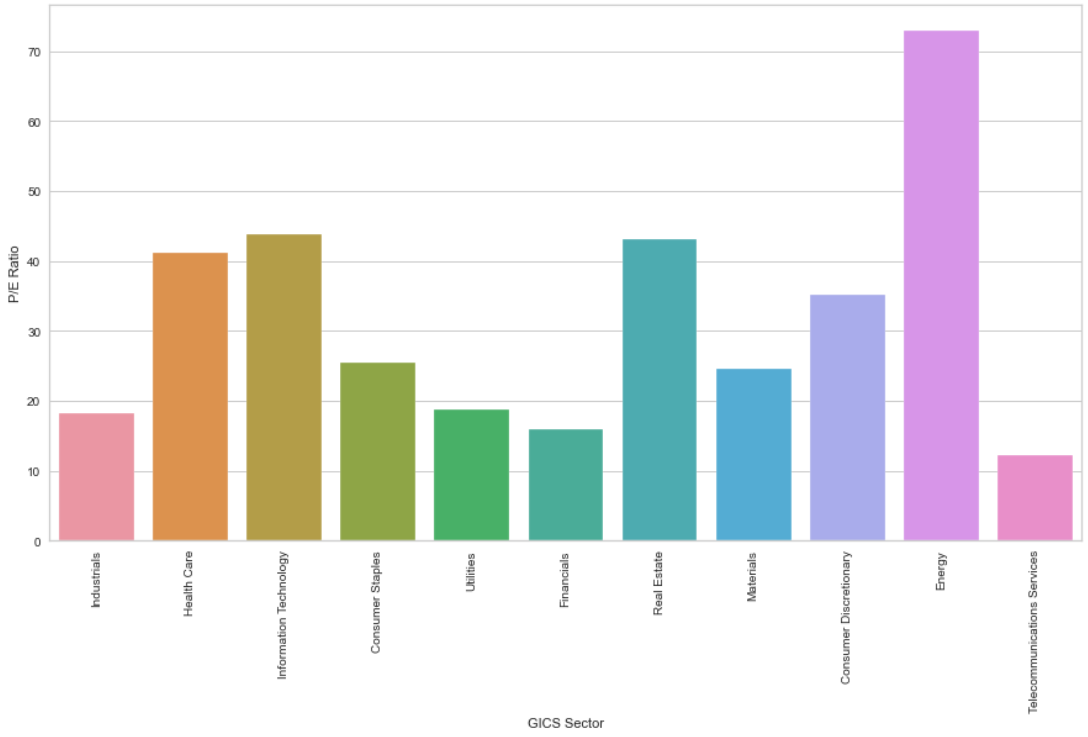


EXPLORATORY DATA ANALYSIS(EDA)

- Information technology, telecommunication services and health care has shown the highest cash ratio accordingly which gives the ability to cover short term obligations with cash equivalents



- The energy industry shows a very high P/E ratio in this distribution compared to the other industry
- There is a tie between information technology and real estate while telecommunication services displays the lowest P/E ratio



DATA PREPROCESSING

- There are no duplicates or missing value in the data set
- Outliers was found in the dataset
- The data preparation will be used in analyzing data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

K-MEANS CLUSTERING

- The appropriate value for k is 4 or 6
- The silhouette for 4 is higher than 6, so we choose 4 as value of k
- Cluster 0
  - This cluster indicates a relatively and medium prices in the stock which means it's a good or poor investment
- Cluster 1
  - This cluster has a low Volatility rate, P/B Ratio, and P/E Ratio which shows a very low risk investment
  - An indication of a high net cash flow, cash ratio and a net income of 14,833,090,909 shows a good investment sector
  - The net cash flow of -1,072,272,727 and cash ratio of 50 indicates low survival rate on the future of the company
- Cluster 2
  - This cluster has a high Volatility rate and P/E Ratio indicates a high risk investment
  - An indication of a low net cash flow, cash ratio and a negative net income of -3,887,457,740 shows a poor investment sector
  - This cluster has the lowest estimated outstanding shares which shows poor value of company
  - This cluster shows the lowest earnings per share of -9 indicating poor profitability
- Cluster 3
  - This cluster shows the highest stock price change in 13 weeks which can be caused by an increase or decrease in earnings
  - There is also a high net cash flow, cash ratio and net income of 1,572,611,680 which means the frequent price change is from an increase in profits
  - This cluster has a relatively low Volatility rate which serves a low risk investment
  - The estimated shares outstanding held by its shareholders is relatively high which is great for the market capitalization and value
  - A high earning per share of 6 and ROE of 26 shows a better profitability
  - A P/B ratio of 14 and P/E ratio of 74 which means expected high future earnings and the stocks can be overvalued

HIERARCHY CLUSTERING

- The highest cophenetic correlation which is obtained from Euclidean distance and ward linkage and average linkage is the same
- The dendrogram with Ward linkage gave us separate and distinct clusters.
- 4 would be the appropriate number of the clusters from the dendrogram with Ward linkage method.
- **Cluster 0**
  - This cluster has a high Volatility rate and relatively high P/E Ratio indicates a high risk investment
  - This cluster indicates low profitability by taking a look the lowest stock price change, cash ratio, net cash flow, a negative net income and earnings per share
- Cluster 1
  - This cluster shows the highest stock price change in 13 weeks which can be caused by an increase or decrease in earnings
  - There is also a high net cash flow, and cash ratio which means the frequent price change is from an increase in profits
  - This cluster has a relatively low Volatility rate which serves a low risk investment
  - The estimated shares outstanding held by its shareholders is relatively high which is great for the market capitalization and value
  - A high earning per share of 7 shows a better profitability
  - A P/B ratio of 14 and P/E ratio of 74 which means expected high future earnings and the stocks can be overvalued
  - However, this cluster shows a low ROE and net income which doesn't really project low profits can be as a result of high price change or company restructuring
- Cluster 2
  - This cluster has a high Volatility rate and P/E Ratio indicates a high risk investment
  - An indication of a high net cash flow, cash ratio and net income shows a great investment sector
  - This cluster has the highest estimated outstanding shares which shows great value of company
  - This cluster shows a relatively low earnings per share and P/B Ratio which might indicates funds are not used efficiently
- Cluster 3
  - This clusters consists of low and medium prices/ values which can indicate medium risk invest

BUSINESS INSIGHTS

The following are the insights the data displayed

- The energy industry shows a high volatility on stock prices which becomes risker to invest followed by material industry
- The utilities sector presents a low risk invest as a result of low volatility on stock prices
- Information technology, telecommunication services and health care has shown the highest cash ratio accordingly which gives the ability to cover short term obligations with cash equivalents
- The energy industry shows a very high P/E ratio in this distribution compared to the other industry
- In K-means clustering,
  - cluster 3 has the best attributes when It comes to investing,
  - cluster 1 and two indices poor investment sector,
  - while cluster 0 sights an average for investors
- In hierarchical clustering,
  - cluster 1 comes in as the best option for investing,
  - cluster 2 shows a great investment sector but comes with a high risk,
  - cluster 0 shows a bad investor and
  - cluster 3 comes in with average