Clustering Model for Stocks

Shravani Ponde
PGDM in Business Analytics
REVA Academy of Corporate Excellence
REVA University
Bangalore, INdia
shravanip.ba03@reva.edu.in

Ratnakar Pandey
Mentor, Business Analytics
REVA Academy of Corporate Excellence
REVA University
Bangalore, India
ratnakarpandey20@gmail.com

Abstract

Investment risk is loss of capital or reduced returns due to market fluctuations and stock's performance. Technical and Fundamental Analysis of stocks can be used to identify stocks with low risk and better returns.

CANSLIM is one such techno-fundamental approach. CANSLIM stands for,

C - Current Quarterly Earnings

A - Annual Earnings Growth

N - New Product or Service, New Management, New High

S - Supply & Demand

L - Leader or Laggard

I - Institutional Sponsorship

M - Market Trend or Market Direction

In this paper the author would study the stocks listed in NASDAQ & NYSE. The stock performance data as well as the company financial data is for 3 years are collected from Yahoo Finance.

A company's stock gets a rating (0-7) basis the CANSLIM criteria, We use earnings per share thisquarter vs samequarter previous year, annual earnings per share, number of outstanding stocks, investor concentration, debt to equity ratio, market share, market Indices like S&P 500 to rate the stocks.

For analyzing stock performance, we calculate Risk/Reward ratio, 200 day moving average, 52 week change % etc. For day trading, we use the Open/Low/Close to calculate the Risk/Reward ratio. For position trading we use the Open/52 week low/ 52 week high to calculate the Risk/ Reward ratio. We then use Risk/ Reward Ratio to classify stocks into Low/Medium/High Risk Categories (Low > 0.5, Medium - 0.3 to 0.5, High <0.3).

We also use visualization techniques to study the stock performance:

Histogram to see the distribution of returns of a given stock

OHCL Graphs

Bollinger bands

Identify what properties constitute a stock into Low/Medium/High Risk Categories:

Study the relationship between risk categories and - Industry Type, Market Cap, Revenue, Beta, P/E Ratio.

This study will use clustering techniques to groups the stocks into Good, Average and Bad. This can be used to predict the behavior of stocks when they move from one cluster to another, or when there is a new stock in the market.

This analysis can help traders and investors identify better stocks, which can help minimize risk and maximize the returns for day as well as position trading.

Keywords—Clustering Technique, Investment Risk, CANSLIM, Stocks

I. INTRODUCTION

After the market debacles of 2000 and 2008, Investors now realize they must take charge and learn much more when they save and invest their hard earned money.

CANSLIM is a system for selecting stocks, created by Investor's Business Daily founder William J. O'Neil (2009). The CANSLIM strategy is based on his analysis of the 500 of the biggest stock market winners from 1953 to 1993. This methodology can help individual investors to pick the best 1-2% of stocks.CANSLIM suggests investors buy higher priced, better quality stocks rather than the lowest priced stocks. Focus on proven factors such as strong earnings and sales growth, price and volume action, and whether the company is the number one profit leader in its field with a superior new product.

Thus, individual investors could make use of this strategy without having to spend a substantial amount of time developing the market expertise.

The American Association of Individual Investors (AAII) 11 year independent study, done in real time, rated it top investment strategy in America is its based 100% on realistic historical studies of how the stock market has actually worked.

Each letter in the acronym stands for a key factor of the greatest winning stocks at their early developing stages, just before they made huge profits for their shareholders.

Key factors as per CANSLIM are Quarterly Earnings, Annual Earnings history, New Product or New Management, Percentage of stock owned by the management, with strong Institutional Sponsorship.

Most preliminary research regarding the effectiveness of the CANSLIM strategy involves using CANSLIM criteria to

select stocks from the S&P 500 and using back testing to compare the returns of this CANSLIM portfolio versus S&P 500 index. We use similar method in this paper. Back testing is used to compute the average gain of the portfolio of stocks.

We will also be using various clustering techniques like KMeans and Self Organising Maps to cluster the stocks based on the CANSLIM criteria.

II. CANSLIM

S&P 500 is an America stock market index which represents the large cap US Equity market. We have evaluated all the 505 stocks listed under S&P 500 as of October 15th 2018.

Fig1 shows the Bar Graph with the number of stocks per Global Industry Classification Standard (GICS) Sector. Industrials has the highest number of stocks listed in S&P 500 followed by Financials. Materials Sector had the least number of stock listed in S&P 500.

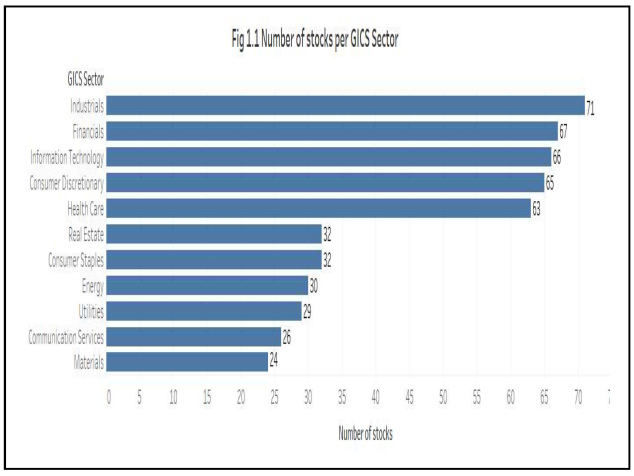


Fig 1 Bar Graph of number of stocks per GICS Sector.

A. C – Current Quarterly Earnings Growth

Stocks must show major percentage increase in their current quarterly earnings per share(EPS) when compared to the same quarter previous year (year over year). Quarter EPS Growth of at least 25% is considered as meeting this criteria.

Table I shows the sample calculation for two stocks. Here AAPL has Current Quarterly Earnings Growth of more than 25%. Hence, the Flag is set to 1. In the sample chosen, overall 10% (51 companies) met this criteria.

TABLE I

Ticker Symbol	C-Current Quarterly Earnings Growth			
	EPS Q2 2017	*EPS Q2 2018*	*% EPS Growth*	*Flag*
AAP	1.43	1.74	21.68%	0
AAPL	2.07	2.78	34.30%	1

B. A – Annual Earnings Growth

Stocks must show consistent annual earnings growth rate for the last three years.Annual EPS Growth of at least 25% for the past three years, is considered as meeting this criteria.

Table II shows Annual EPS Growth Percentage calculation based on EPS (Diluted) for the stock FB.

TABLE II

Ticker Symbol FB	Calculating Annual Earnings Growth			
	2014	*2015*	*2016*	*2017*
EPS Diluted	1.10	1.29	3.49	5.39
EPS Growth %		17.27%	170.54%	54.44%

Table III shows the Annual EPS Growth Percentage for five stocks. Here, FDX has consistent growth of more than 25% for the past three years. Hence, the flag is set to 1.

Only 3% (15 stocks) met this criteria. This criteria filtered out most of the stocks.

TABLE III

Ticker Symbol	A-Annual Earnings Growth			
	% EPS Growth 2015	*% EPS Growth 2016*	*% EPS Growth 2017*	*Flag*
FCX	-797.74%	73.44%	140.21%	0
FDX	78.52%	70.12%	51.46%	1

C. N – New Supply, New Services, New Management, New Price Highs

O' Neil states that it takes something new to cause a substantial increase in the price of a stock. It can be new product or service that sells rapidly or it can be change of management, new industry conditions or revolutionary technologies can have positive effect on stocks.

Since, it is time consuming to evaluate each stock for new supply, new service, new management, new price highs. We skip this parameter.

D. S – Supply and Demand

Demand of a stock should be greater than supply of it. Volume change of at least 50% when compared to the average trading volume over the last 50 days. Also, the top management of the company should own a reasonable percentage of stock and buys its own stock in the open market consistently. For simplification, stocks where the management (insiders) owns at least 1% of the stock is considered to be meeting this criteria.

Table IV shows the percentage of stock held by insiders for two sample stocks. Wherever, the percentage of stock held by insiders is more than 1%, the Flag is set to 1.

TABLE IV

Ticker Symbol	S-Supply and Demand	
	% held by investors	*Flag*
HD	0.12%	0
HES	18.24%	1

Overall, 61.6% of companies had less than 1% of stock owned by insiders. Only 38% (192 companies) met this criteria

E. L – Leader or Laggard

The top one or two companies that are the leaders in terms of quarterly and annual earnings growth, sales growth in the specific industry are considered as Leader. Since, we already evaluated the stocks for Quarterly and Annual Earnings growth, we skip this parameter as well.

F. I – Institutional Sponsorship

The biggest source of demand for the stock should be institutional investors like mutual funds, insurance companies etc. Meaning the stock should be owned by several institutional sponsors. A stock that has at least 20 institutional owners is considered to be meeting this criteria.

TABLE V

Ticker Symbol	I-Institutional Sponsorship	
	Number of Institutional investors	*Flag*
FIS	998	1
FISV	1141	1
FL	1	0

Only very few companies had the number of institutional investors less than 20. Almost 95% of companies met this criteria.

G. Market Direction

William J. O' Neil stresses on the fact that picking the right stocks which satisfy the first six criteria is not enough. It is important to evaluate the Market direction, and study the Market Indices for perfect timing of stocks. Since this parameter is not specific to stocks, we leave this parameter out.

H. CANSLIM Stocks

The Stock that met the four criteria listed above is considered as a CANSLIM stock.Overall, 5 companies met the CANSLIM criteria.

III. Reward Classification:

A. Reward/ Risk Ratio

For the sake of calculation, we are assuming that the stock was bought exactly one year ago on 19th Oct 2017. We are considering the closing price of the stock on 15th Oct 2018 as the selling price of the stock to calculate the Reward/Risk ratio. So, the difference of closing price of the stock for the past one year will be the Reward (Profit).

Generally speaking Risk is set to only 7-8% or at the most to 10% lesser than the price that the stock was bought at. This is also called stop loss. However, we are considering the price at which the stock was bought as the Risk.

Reward/Risk percentage is calculated for each stock following the Equation (1).

Equation(1)Reward/ Risk = (Closing Price of current year – Closing Price of previous year)/Closing Price of previous year.

TABLE VI

S. No	Reward Classification Table	
	Ratio	*Reward Classification*
1	<0.0	Loss
2	0.0 – 0.3	Low
3	0.3 – 0.6	Medium
4	>0.6	High

B. Variance

The Closing Price of the stock for each day (weekday) for the pas one year (19th Oct 201& – 15th Oct 2018) is considered for calculating Variance for each stock.

In Figure2, we have plotted the variance of all the stocks against the Reward Classification. We can see that the Reward Classification of High and Medium had more variance compared to Loss and Low. Also, the variance was more pronounced in third Quartile Q3 for the Reward Classification of High and Medium.

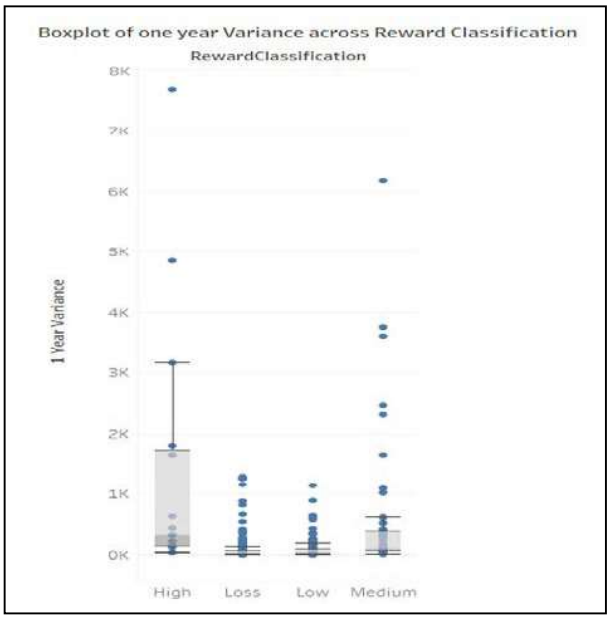


Fig 2 Boxplot of variance across GICS Sector

C. Reward of CANSLIM Stocks

We use back testing to compute the gains of stocks. We categorize the stock that as,met CANSLIM 100%, 75% and

<= 50%. Then using random selection we selected 5 stocks in each category and computed the net gain for each of these 15 stocks.

Table VII shows the comparison of average gain in each of these three buckets. CANSLIM stocks have a very high average gain.

TABLE VII

S. No	Average Gain Table	
	CANSLIM	*Average Gain*
1	100%	165.8
2	75%	9.32
3	<= 50%	3.18

Fig 3 shows the scatter plot of Price (closing price as on 15th Oct 2018) plotted against Variance. As price increases so does the variance with almost a linear relationship. Pearson correlation coefficient between Price and Variance is 0.86.

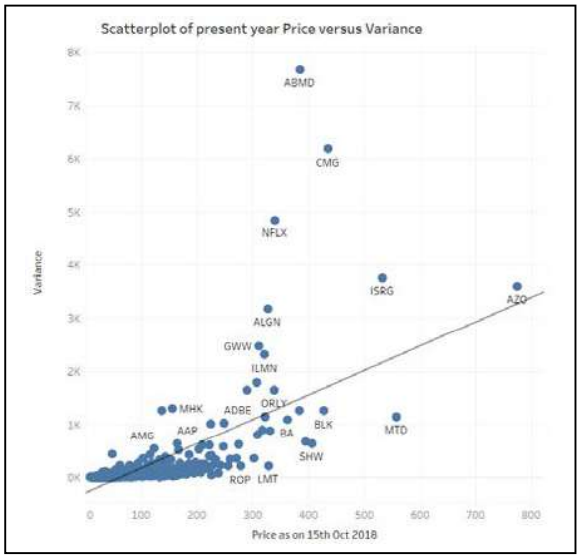


Fig 3Scatterplot of price versus variance

Fig 4shows the Boxplot of variance across various GICS Sector. Communication Services, Consumer Discretionary and Health Care Sectors have more outliers with higher values.

Fig 5 shows the performance of the 5 CANSLIM stocks against the S&P 500 Index for the past one year. We have used Close Price for this line chart.(S&P 500 Index Close Price is scaled).

D.Bollinger Band

Bollinger Band is a set of lines plotted two standard deviations(positively and negatively) away from the a simple moving average of the stock's closing price.

The upper and lower band are calculated two standard deviations away from the 20 day moving average. Because Standard deviation is a measure of volatility, when the markets are volatile, the bands widen and during less volatile periods, the bands contract.

Many traders believe that the closer the prices move to the upper band, the more over bought the market, and the closer the prices move to the lower band, the more oversold the market.

Fig 6 shows the Bollinger Band for the S&P 500 Index.

Fig 7 and Fig 8 show the Bollinger Bans for two of the CANSLIM stocks.

All charts for the exploratory analysis are built using Tableau Public 2018.2.

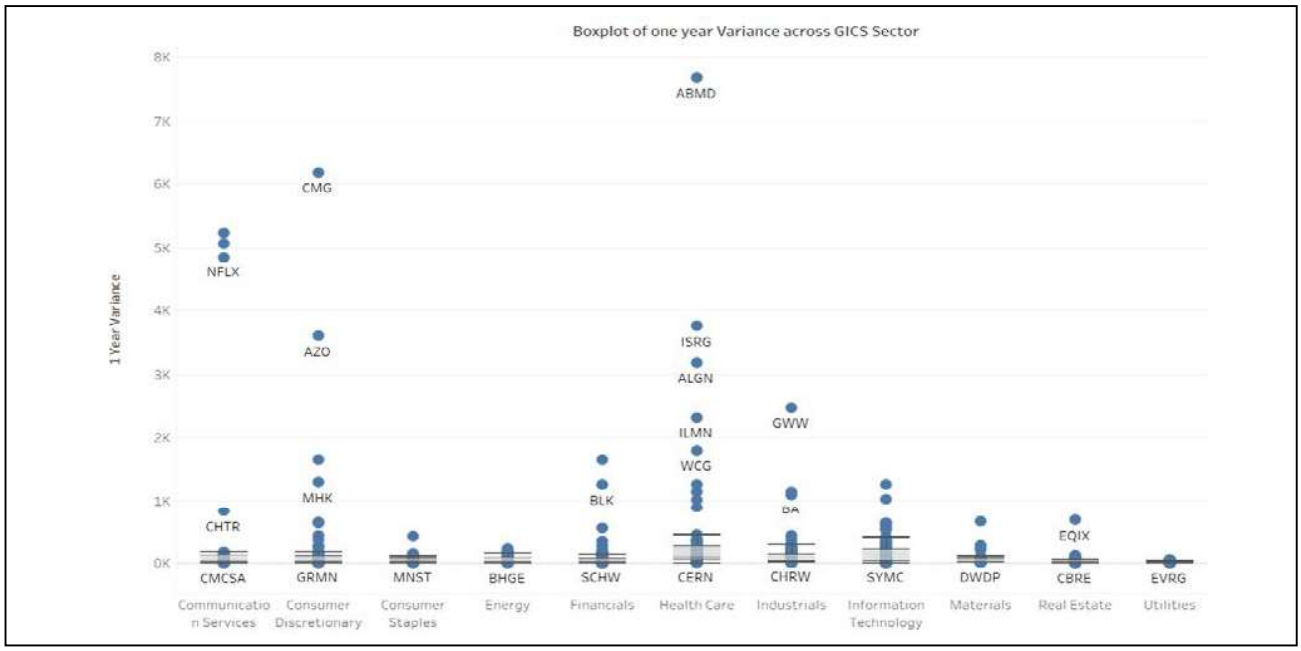


Fig 4 Boxplot of variance across GICS Sector

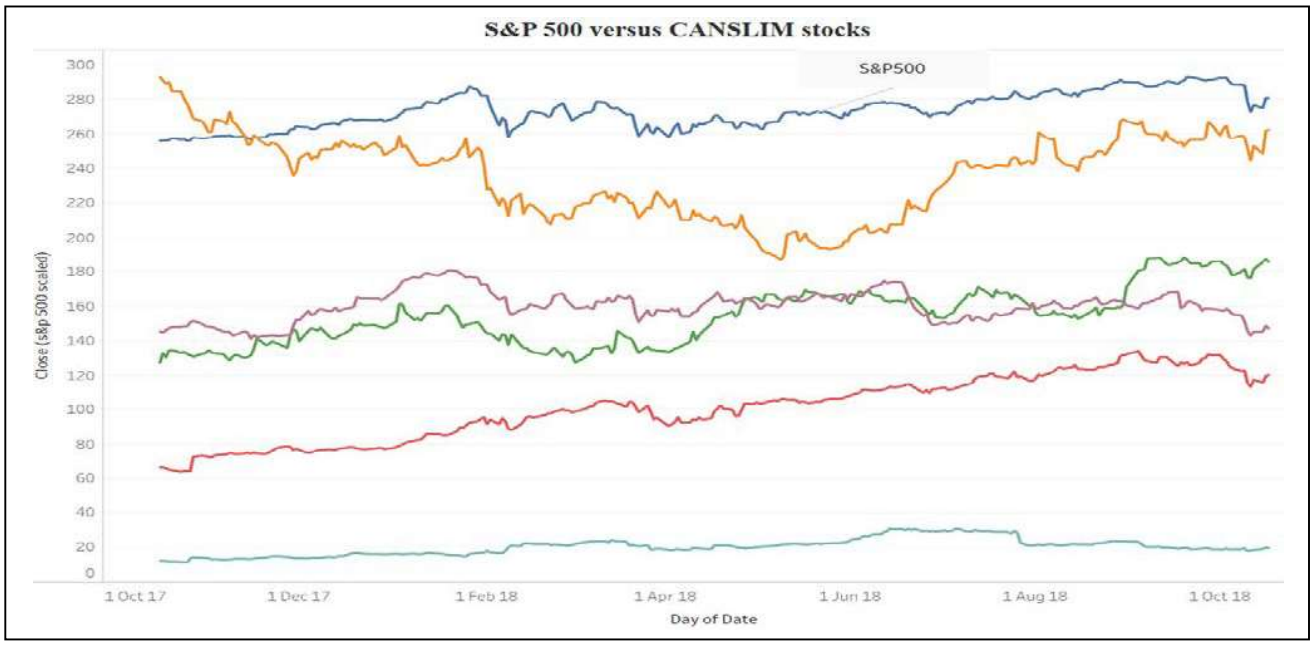


Fig 5 Line chart showing performance of S&P 500 Index v/s CANSLIM

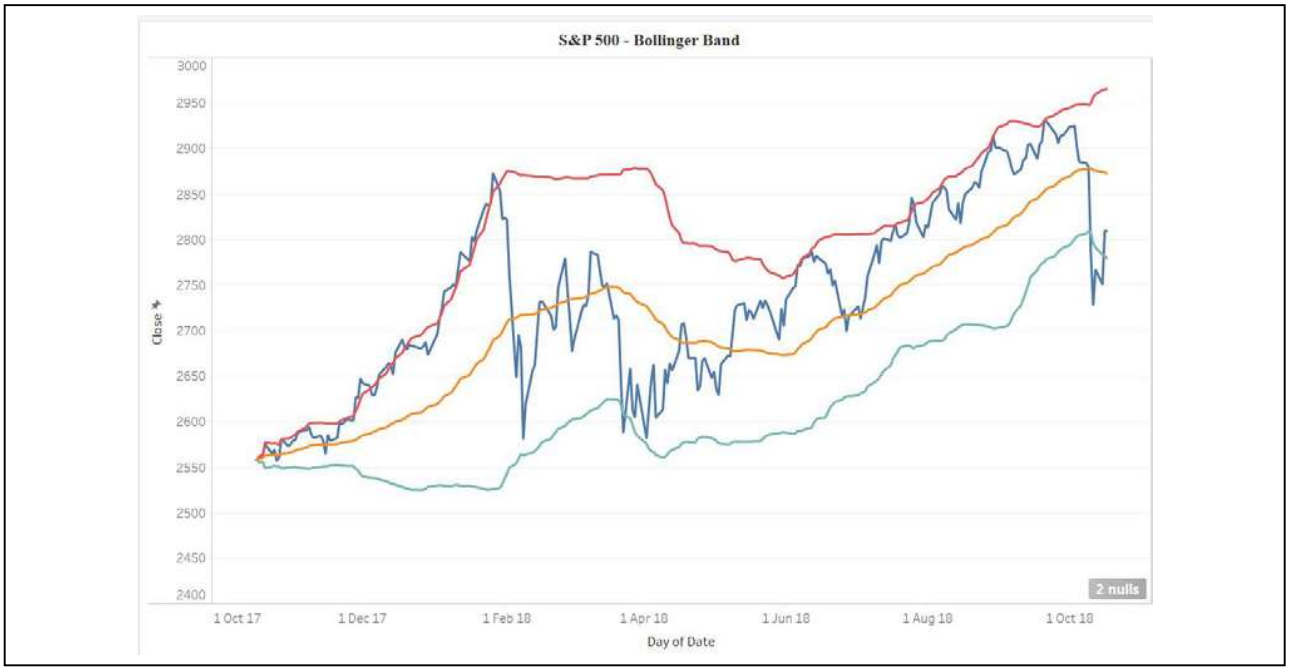


Fig 6Bollinger band showing the perfromance of S&p 500

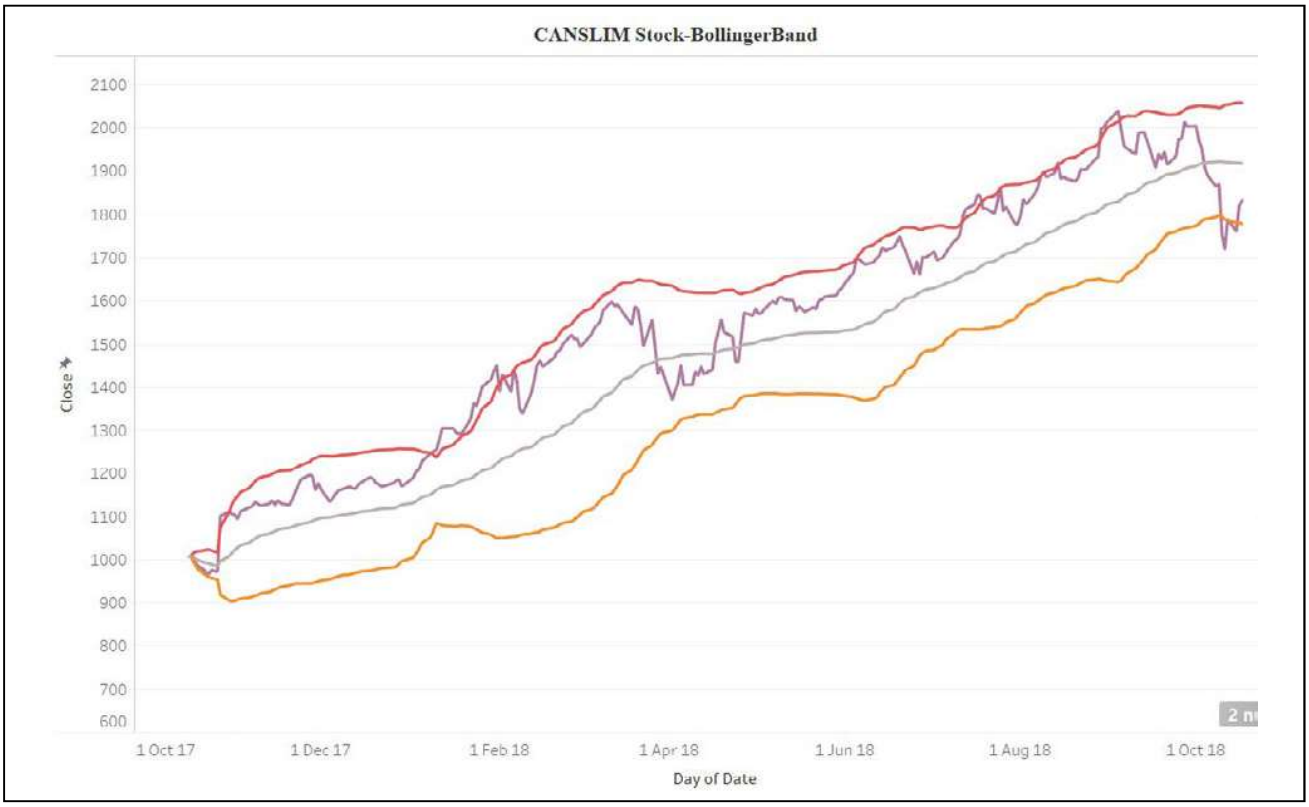
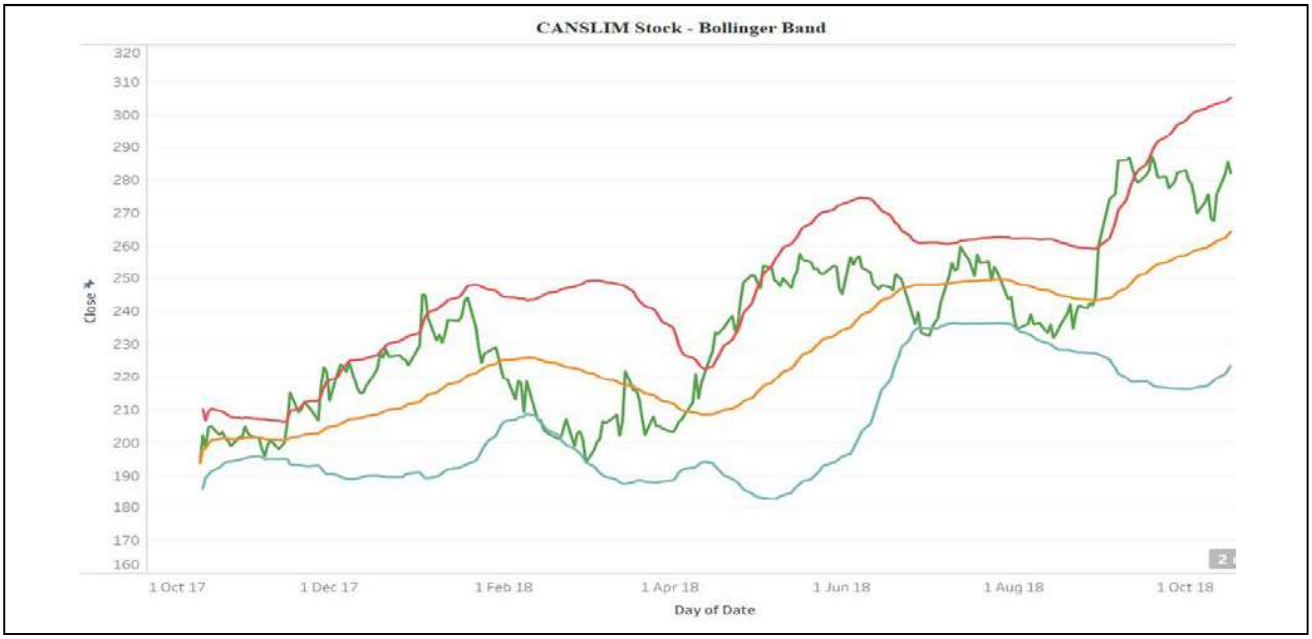


Fig 7 Bollinger band showing the performance of CANSLIM stock

Fig 8 Bollinger band showing the performance of CANSLIM stock



IV. Clustering Model

A. Data Collection

We have used the same data used for evaluating CANSLIM criteria to build the clustering model. Annual EPS 2015, Annual EPS 2016, Annual EPS 2017, Q3 2018 EPS, Q3 2017 EPS, Percentage held by insiders, Number of Institutional Investors.The stocks with the missing values are eliminated. All the numeric values are scaled before modelling.

B. K Means

K Means is an unsupervised learning algorithm that tries to cluster data based on similarity. It randomly assign the each observation to a cluster and finds the centroid of each cluster. Then the algorithm iterates till the within sum of squares cannot be reduced any further.

We have used K Means clustering algorithm to cluster the stocks into 10 clusters.

K Means Model output

Between ss/ total ss = 69.2%

CANSLIM stocks are members of cluster # 2.

C. Cluster Plot

Clusterplot uses the first two principal components to explain the data. Fig shows the cluster plot of the 10 clusters created using K means. The two components explain 38% variability in the data.

D. Scree Plot

The within sum of squares (WSS) is plotted against the number of cluster centres. Fig shows the Scree plot of optimum number of cluster. WSS is least when the number of clusters is between 10-15 (at the elbow) of the WSS curve.

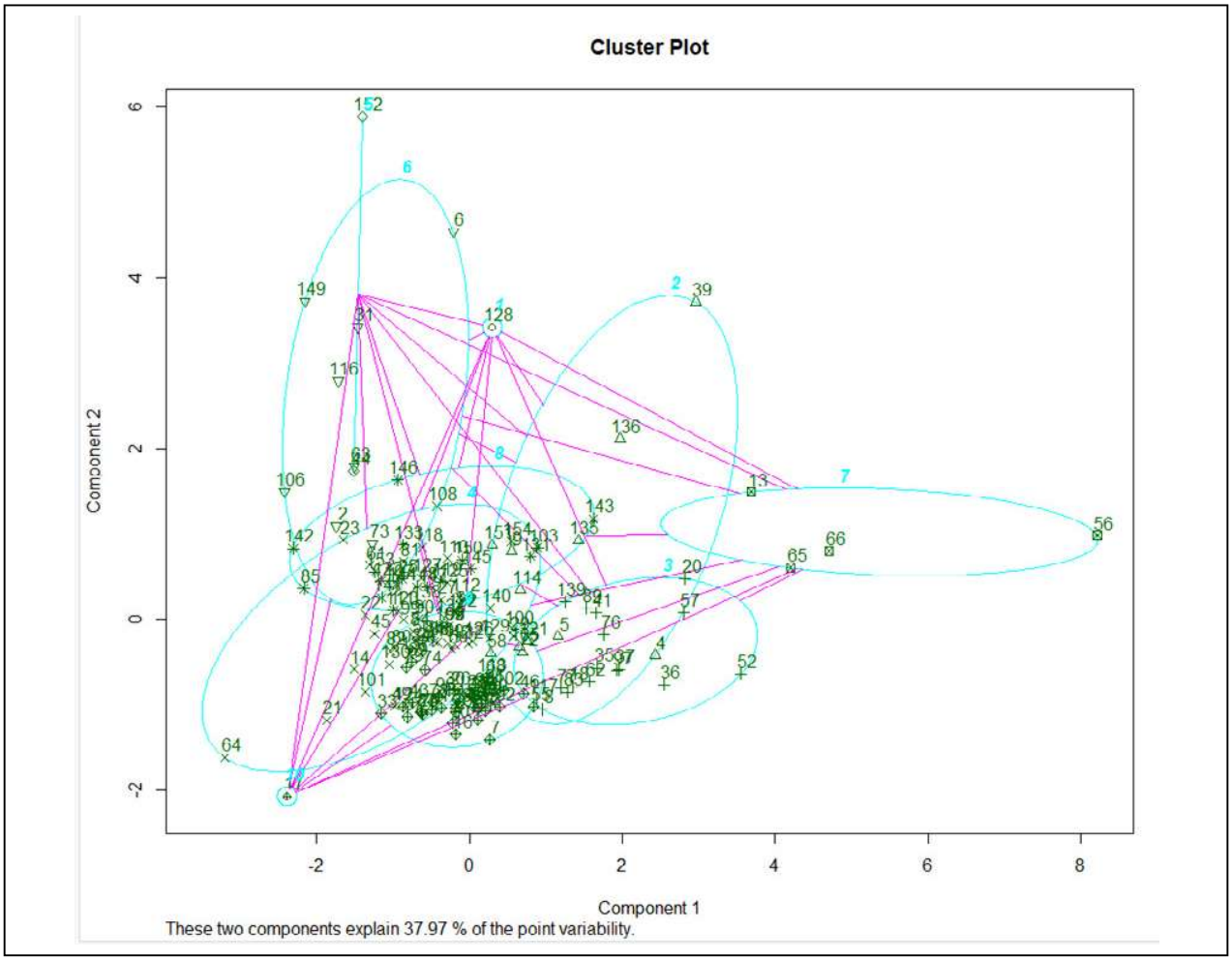


Fig 9Cluster plot of Stocks using K Means

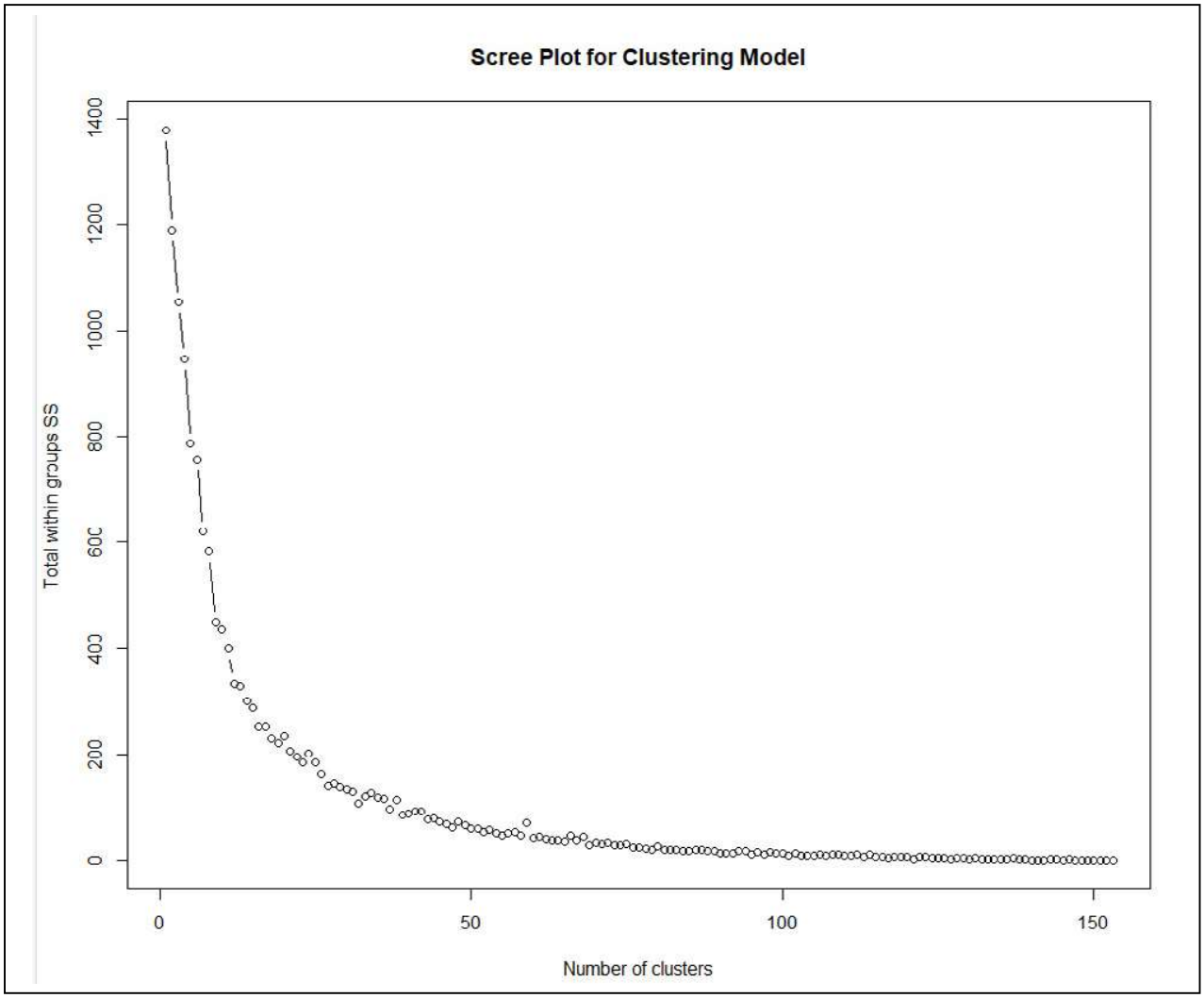


Fig 10 Screeplot showing optimum number of clusters

V. KOHONEN SELF ORGANISING MAPS(SOM)

A. SOM

SOMs were first described by Teuvo Kohonen in Finland in 1982. SOMs are an unsupervised data visualization technique that can be used to visualize high dimensional data sets. Typical SOM Visualizations are of 'heatmaps'. A Heatmap shows distribution of a variable across the SOM.

B. Counts Plot

Count plot gives us the count of how many samples are mapped to each node on the map. Fig 10 shows us the count of samples mapped to 5*5 nodes. The number of members per node is varying from 2 to 14. There are three nodes with 14 members and six nodes with up to four members.

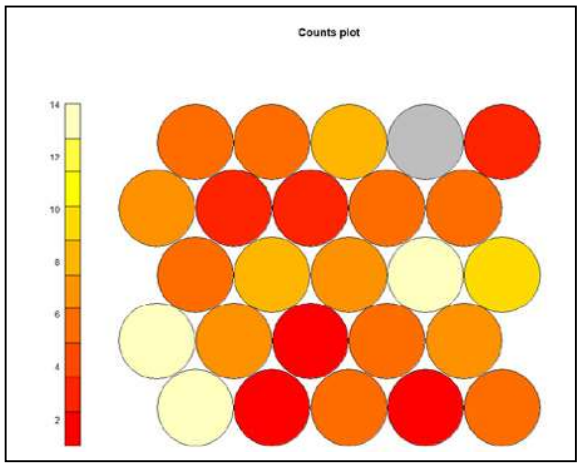


Fig 11 Counts plot of stocks using KohonenSOM

C Neighbour Distance

This visualization shows the distance between each node and its neighbours. Areas with large distance indicate nodes are much more dissimilar.

Fig 11 shows the distance between each node and its neighbours.

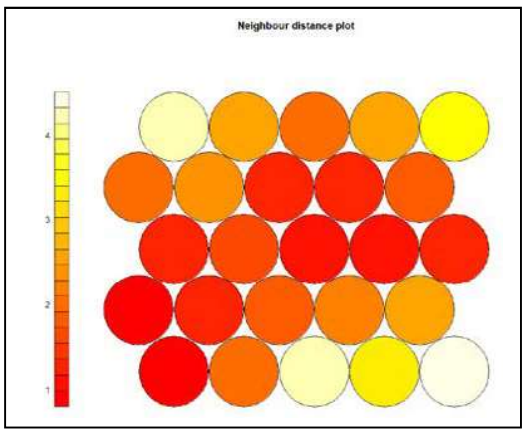


Fig 12Neighbour distance plot using Kohonen SOM

D. Weight Vectors

Weight Vectors (codes) are made up of normalised values of the original variables used to generate the SOM. Each node's weight vector is representative of the samples mapped to that node. The default visualization of the weight vector is a fan diagram, where individual fan represents the magnitude of each variable.

In Fig 12, Each node is representing the magnitude of each variable in the form of a fan. Nodes 2 and 3 in the first row have magnitude for all the eight parameters of the CANSLIM methodology. These (or one of these) could be the node representing the CANSLIM stocks.

Also, we can see that in most of nodes the stocks have satisfied the Percentage held by investors and Institutional ownership. Only in very few nodes EPS Quarterly Growth(yoy) and EPS 2015 Growth is visible.

C. Hierarchical Clustering

SOM can be clustered using Hierarchical clustering. Fig 13 shows the 5 clusters created using Hierarchical clustering.

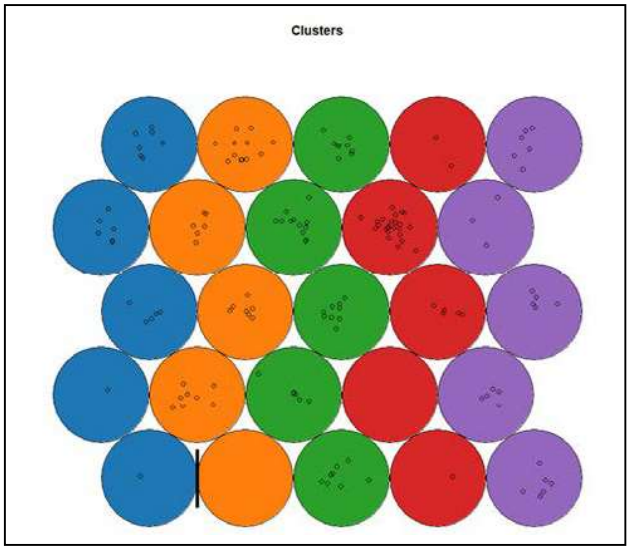


Fig 14 Hierarchical clustering in SOM

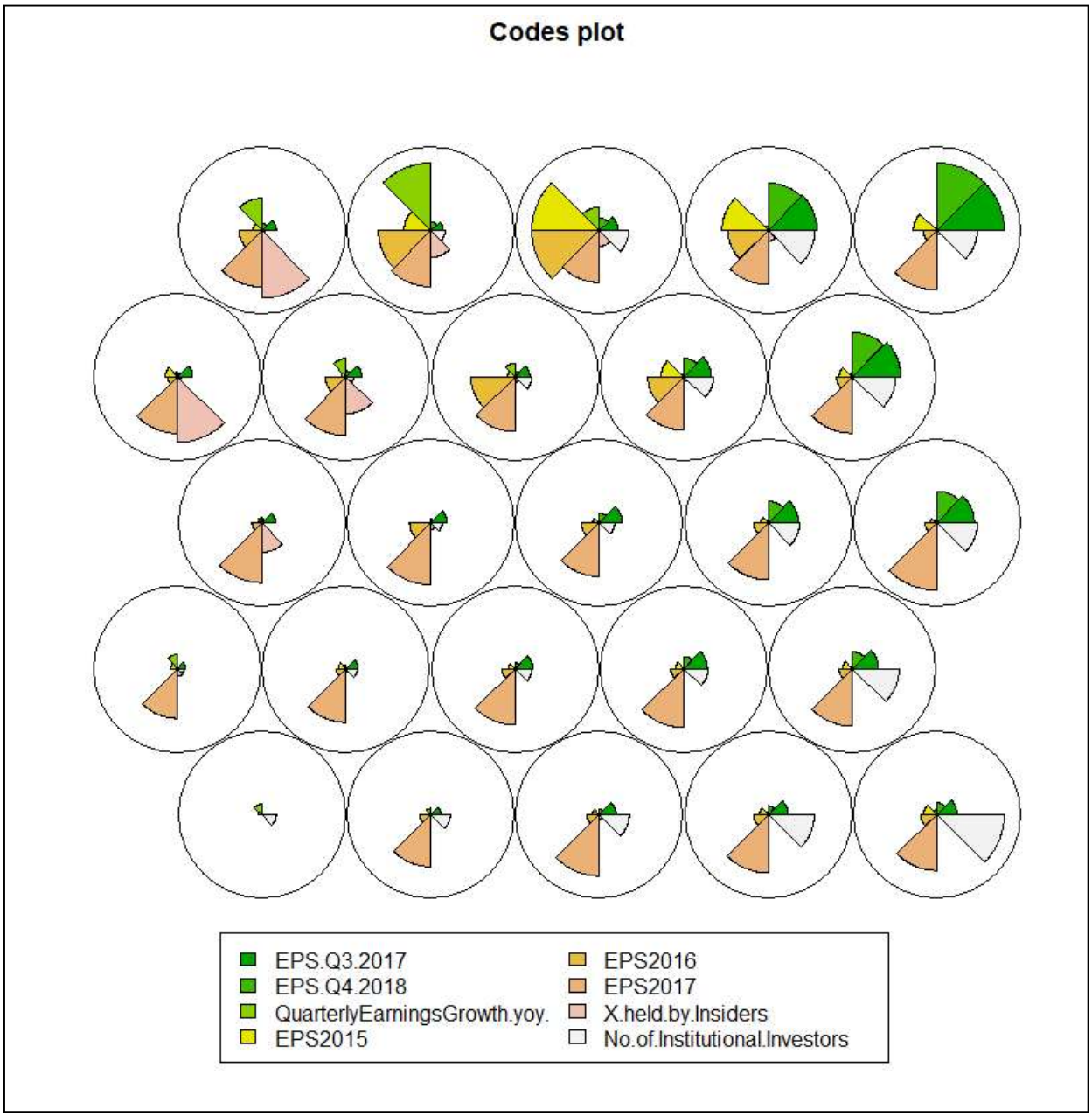


Fig 13Weight vector node plot using Kohonen SOM

VI. Conclusion

In this paper we collected the latest data for the 505 S&P 500 stocks. The stock were evaluated using four out of the seven CANSLIM criteria to identify the CANSLIM stocks. The variance and the reward of the stocks was studied.

1. The performance of the CANSLIM stocks was evaluated using the Average gain and Visualization techniques like Line charts and Bollinger Band
2. The performance of the CANSLIM stocks is found to be superior compared to other stocks that did not meet the CANSLIM criteria.
3. The stock parameters used for evaluating CANSLIM criteriawas then clustered using K-Means and Kohonen Self Organised Maps.

REFERENCES

[1] O'Neil, W. J., & O'Neil, W. J. (1988). How to make money in stocks (Vol. 10). New York: McGraw-Hill.

[2] O'Neil, W. (2009). How to make money in stocks: A winning system in good times and bad. McGraw-Hill Education.

[3] Lutey, Matthew, Michael Crum, and David Rayome. "OPBM II: An Interpretation of the CAN SLIM Investment Strategy." Journal of Accounting & Finance (2158-3625) 14, no. 5 (2014).

[4] https://www.r-bloggers.com/self-organising-maps-for-customer-segmentation-using-r/

[5] https://en.wikipedia.org/

[6] https://www.investopedia.com/

[7] https://in.finance.yahoo.com/

[8] https://www.marketwatch.come/