Indian Institute of Technology, Kanpur
Computer Science and Engineering

CS685 Data Mining Project Report - Group 25

Stock Data Extraction, Analysis and Prediction

Author:
Abhishek Reddy(21111062)
Mail Id: tareddy21@iitk.ac.in

Deepak Kumar(21111023)
Mail Id: dkumar21@iitk.ac.in

Mandar Dhake(21111405)
Mail Id: mkdhake21@iitk.ac.in

Yash Patil(21111410)
Mail Id: yashpp21@iitk.ac.in

Supervisor:
Prof. Arnab Bhattacharya

Year 2021-22

Acknowledgement

We would like to extend our sincere gratitude to our Project Guide, Prof. Arnab Bhattacharya for his advice, guidance, patience and timely help during the project. The freedom he gave us, to work with anything at any time, encouraged us to try out new things and helped to work more efficiently. It is a great honor for us to have been a part of his course. The successful completion of the work has been only possible due to his excellent guidance, meticulous observation and critical analysis.

Abhishek Reddy(21111062)
Deepak Kumar(21111023)
Mandar Dhake(21111405)
Yash Patil(21111410)

Contents

Abstract

Stocks is the most general topic. From teenagers to elderly everyone looks out for different opportunities to invest in stocks. Its always the one of the topics in discussions. Stocks have entangled people as if it controls their fate. Stocks is the best way of investment and most of stocks are freely available for people to buy/sell as they please. Now a days, handling stocks have became so easy, without any broker just through some online applications. This easy accessibility has increased the number of investors by nearly 200%. This leads to the question what are these stocks and whats so interesting about it?

Stocks are a type of security that gives stockholders a share of ownership in a company. The company shares some of its share in open market which are exchanged in stock market among the investors. There are two types of companies public and private. A public company is one which can share its shares to the general public on the stock exchange (share market) eg: - Reliance, ICICI Bank , Yes Bank. A private company is one which holds its shares to a few, big money investors eg: Paytm, Dell. Stock and share are nearly same the only difference between two is that, a 'stock' can refer to any arbitrary company, but the word 'share' is used when we are referring to a specific company. In India NSE (National Stock Exchange) and BSE (Bombay Stock Exchange) are basically the 'markets' where one can buy and sell shares of a company. Unlike the conventional markets, these markets are electronic and not physical. All transactions take place electronically. But how the stock prices change.

The stock prices are simply controlled by buyers-seller trends. If buyers are willing to pay a higher price for a stock than its current price and seller sells this stock at that high price then stock price increases. Similarly, if buyers are lowering the buying price for a stock than its current price and seller are selling these stock at that lower price then stock price decreases. So if a investors want to make most profit out of stock market. He must but stocks when it reaches valley point and sell the stocks when it reaches peak point. But that's the gamble of stock market no one can predict exact time or value at which stocks reach these extreme points. The one who understands these patterns may even earn 15-20% profits, but those who doesn't may remain in loss forever. As The market runs electronically the number of investors have increased tremendously, which lead to too much volatility in the market. There are 7400 companies in Bombay Stock Exchange and 5400 companies in National Stock exchange. Keeping track of trends in most of these companies manually to earn some descent profit got harder and harder.

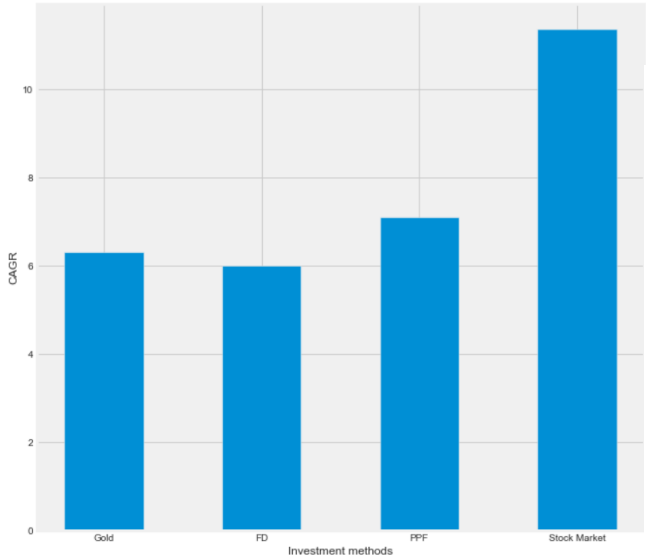
Our system considers this problem faced by investors and provides a user convenient platform that extracts the live data, analyses it for investors, predicts the best possible solutions for user. Hence, users have to just pick the right information from the system and they are welcome to earn great profits in stock market.

Chapter 1

1 Introduction and Broad Aims of the Project

1.1 Introduction and Motivation

The common options available for investment for common people are gold, Fixed deposits, PPF(Public provident fund), Real Estate(but it is illiquid form of investment), Stock Market. These all are the best options for investment and it is always preferred to invest in more than one way. As Figure 1a suggests stocks are on another level it has the highest returns. CAGR(Compound Annual Growth Rate) is calculated using data of a decade. Nifty50 is the market index which represents the top 50 companies listed on the NSE(National Stock Exchange). We can see that in last decade stock market returns out performs all other investment strategy gives an average annual returns more that 11% without taking any high risk as its the return of nifty50 which is index of market.



(a) CAGR of Gold, FD, PPF, Stock Market in this decade

(b) CAGR of nifty50 from 2013 to 2021

Figure 1: Compound Annual Growth Rate of different investment methods

Stock market is the most volatile market. With high returns it also has high risks. In 2020 we saw the big fall in Indian stock market due to coronavirus, lockdown and Yes Bank Crisis, recession was at its peak. But as everything was eventually under control the market started gradually rising. As the consumption rate of common man decreased, he started moving his money in market. And eventually we saw a great rise in market in 2021. The stock market performed better than expected. But, now as the spread of coronavirus has melted down and people are back to routine. As we see the economic boom, it affected the market. The regular goods and services consumption of people has increased, this overheating economy could easily suck money out of the stock markets [1].

The market rise of 2021 attracted many investors, but this rise is expected to decrease due to the current economic rise. This motivated us to create an application that provides a proper analysis and prediction of the current stock market to keep the repelled investors in loop and do not let them loose their interest in Indian stock market investment.

1.2 Broad aims of the project

The main aim of this project is to thoroughly analyse and predict the Indian stock market so that users can invest effectively. The project attains the aim by providing solutions to following:

1. Scrape the data of overall performance of all company stocks in national stock exchange.
2. Scrape the daily performance of selective companies for analysis
3. Classify the companies into small cap, middle cap and large cap categories.
4. Find the top six companies that have higher returns and analyse their trends.
5. Analysis of closing price and volume of stock over time.
6. Analysis of Simple Moving Averages of various stocks
7. Analysis of daily returns of various stocks.
8. Analysis of the relation between volume and daily returns of stock.
9. Analysis of correlation of daily returns between different stocks.
10. Risk analysis of various stocks.
11. Analysing various trends of each stock.
12. Trade calls using bollinger bands.
13. Beta calculation using regression.
14. Diversification analysis using clustering.
15. Trade call prediction using classification models.
16. Prediction of closing prices using LSTM.

2 Data Extraction and Preparation

The use case required the availability of live data, which can be analysed thoroughly to produce results. But extracting the data requires large processing power as we need to continuously run the extractor in background. Hence, more the companies, more is the processing power i.e multiple cores required to get and analyse the data. Therefore, we decided to work with some selective companies which would be best investing for users. And gradually as project progresses we can gradually increase the number of companies.

To select the best companies, in first phase we extracted the data of 5 year performance of all national companies, analysed them and classified them into three categories named large cap, medium cap, small cap. Furthermore, we analysed the top 6 companies from large cap categories. These were the top companies we used to perform final analysis. In second phase we extracted the data of last 2 years for performing analysis on these six companies.

2.1 The performance data of all companies

In the first phase we extracted the data for company classification [2]. For this we used website named https://www.tickertape.in/screener. This website was best reference because it had briefly evaluated the performance of most of the national company stocks. The parameters we gathered from the site to evaluate the performance are as follows:

- Stock name : Names of the company's stock.
- Sub-sector : Sub-Sector of the company's stock. Depends on product or service they provide.
- Market Cap : Market Cap Price of the company's stock i.e market value of company stock.
- Closing Price : The close price from the day company's stock last traded
- PE ratio : Close price divided by the earnings per share, excluding extraordinary items of the company's stock.
- Five Year Growth : Compounded annual growth rate of the company's stock price over previous 5 years.
- Alpha : Excess return of a company's stock relative to its benchmark, calculated using 104 weekly price close points of the stock.
- Beta : Measure of a company's stock price volatility relative to the market.
- Five Year Historical Revenue Growth : The annual compounded growth rate of revenue over the last 5 years of a company's stock.

These were the best sophisticated parameters we were able to find to evaluate overall performance of stock.

To extract the data we used selenium web-driver. It is the most efficient web-driver and easy to use. Where we can select the fields by simply clicking on the tags. The steps we followed are as follows

1. Create a new web-driver connection and open the site "https://www.tickertape.in/screener" in chrome using it's web-driver tool.

```
[4]: driver1 = webdriver.Chrome(executable_path = "chromedriver.exe")
     driver1.get("https://www.tickertape.in/screener")
     driver1.set_window_size(720,1280)
```

Figure 2: Initialize the web-driver

2. Emulate clicks to add filters to the site so that we can get various required attributes for our data.

```
#CLICKING EDIT FILTERS
driver1.execute_script(
    "arguments[0].click();", WebDriverWait(driver1, 20)
        .until(EC.element_to_be_clickable(
            (By.XPATH,"/html/body/div/div[2]/header/div/div[1]/div[1]/div[3]/div[2]/div/span")
        ))
)
```

Figure 3: Example of code to add the filter to scrape the data

3. Get the number of stocks provided by the site.
4. Define the lists to store extracted data.
5. Scrape the data from the site into previously defined lists with each list containing the corresponding attribute.

```
names.append(
    driver1.find_element_by_xpath(
        "/html/body/div/div[3]/div[1]/div/div[2]/section/div[2]/section[2]/div[3]/span["+str(i)+"]/span[2]"
    ).text
)
```

Figure 4: Example of code to add the scraped data into respective lists

6. Replace the null values, "-", or " " with Nan which can be handled easily in python

The extracted data is used in section 4.1 and 4.2

2.2 The daily performance of different companies

In second phase we needed data of daily performance of company stocks [3]. The website we referred is https://finance.yahoo.com/. The data for different national companies is readily available on the site. The parameters we gathered from the site are as follows:

- Date: The day of accounting
- High: The highest price the stock reached in whole day.
- Low: The lowest price the stock reached in whole day.
- Open: The opening price of the stock.

- Close: The closing price of stock.
- Volume: Volume measures the number of stocks traded throughout the day
- Adj Close: The adjusted closing price is a stock's closing price to reflect the stock's value after accounting for any corporate actions.

These are perfect parameters to evaluate the daily performance of a stock. The procedure followed to extract the data is same as before. But here we had to extract data of more than one company for different time periods for different types of analysis. Hence we created a modular fuction which could extract data using company name and starting date of extraction. The steps followed to scrape the data is as follows:

1. Input the comapany date and starting date and set current date as end date.
2. Create the driver with url to extract the data for given period
3. Emulate the clicks to scrape the data from web page
4. Scrape the data into a dataframe for all the days the data is available
5. Store it in csv.

As the perfect function is ready to extract data of any company in any time period, we extracted data as needed. Following are the different extractions performed on the website.

- The data for three years of top six companies which is analysed in section 4.2 and 3.2
- The data for three years of top 20 companies which are analysed in section 4.5
- The data for ten years of Mindtree which is used for prediction in section 5.2
- The data for eleven years of nifty50 which is used for beta analysis in section 4.4.4 and 4.5.1

3 Feature Selection and Extraction

3.1 Feature selection on dataset of performance of all companies

The dataset has many features, out of which the necessary features are to be selected. To evaluate the overall performance of company the useful features are Market Cap, Five Year Growth Rate, Alpha and Beta. The other parameters are also relevant but have least effect on stock performance. Hence we decided to evaluate the performance based on these given parameters. And Sub-sector parameter is used to draw some inferences like how different parameters like covid affected different types of industries.

3.2 Feature Extraction for three year dataset of top six companies

The features extracted are well selected and all the features are required for analysis, but also we some features can be extracted from given features, like derived features. The extracted features are given below:

3.2.1 Simple Moving Average(SMA)

Simple Moving Average(SMA) [4] is an arithmetic moving average calculated by adding recent prices and then dividing that figure by the number of time periods in the calculation average. Short-term averages respond quickly to changes in the price of the underlying security, while long-term averages are slower to react.

$$SMA = \frac{A_1 + A_2 + A_3 + ... + A_n}{n}$$

where, A_n = Price of a asset at a period n, n = Number of total periods.

Here for calculating SMA we have taken 2 intervals such as per day, 20 days, 35 days.

3.2.2 Daily returns

The Daily returns of the stock is rate of change of Adj Close. Formula for percent change can be given as:

$$Daily\ Returns = \frac{Current\ Adj\ Close - Previous\ Adj\ Close}{Previous\ Adj\ Close}$$

It may be negative or positive representing rise or fall in stock value respectively.

3.2.3 Daily Trends

As a part of feature extraction, for each company stocks we analysed the daily trends. This mainly depended on daily returns on stock. Each daily return signifies a certain trend in market, these trends are given below:

1. Slight or No change for 'Daily Return' in between -0.5 and 0.5
2. Slight positive for 'Daily Return' in between 0.5 and 1
3. Slight negative for 'Daily Return' in between -0.5 and -1
4. Positive for 'Daily Return' in between 1 and 3
5. Negative for 'Daily Return' in between -1 and -3
6. Among top gainers for 'Daily Return' in between 3 and 7
7. Among top losers for 'Daily Return' in between -3 and -7
8. Bull run for 'Daily Return' >7
9. Bear drop for 'Daily Return' <-7

This signifies the daily progress of stock

3.2.4 Volume Weighted Average Price (VWAP)

The volume-weighted average price (VWAP) [5] is a trading benchmark used by traders that gives the average price a security has traded at throughout the day, based on both volume and price. Large institutional buyers and mutual funds use the VWAP ratio to help move into or out of stocks. Therefore, when possible, institutions will try to buy below the VWAP, or sell above it. VWAP is calculated by adding up the dollars traded for every transaction (price multiplied by the number of shares traded) and then dividing by the total shares traded.

$$VWAP = \frac{\sum Price * Volume}{\sum Volume}$$

4 Data analysis

We referred [6] for analysis research. The data analysis is done in different phases as follows:

4.1 Classify all companies based on their capitalization

The company stocks are classified in three classes large capitalization, medium capitalization, small capitalization. We classified stocks in these categories because each class has its specific distinction. But most varying parameter is volatility. The small cap stocks are too volatile and it decreases as we move towards large cap stocks. As small cap stocks are so volatile analysing them would not be an appropriate choice as one day changes are enough to contradict the whole analysis. Hence we decided to initially consider the large cap stocks as our main priority. Also, it would be fruitful to study such stocks. Hence we decided to start with large cap stocks.

The steps for classification are as follows:

1. Import stock data
2. Filter out rows with invalid market cap
3. Sort the companies based on market cap in descending order
4. select top 200 companies as large cap companies
5. Select top 200 companies with market cap less than 19000 in medium cap companies
6. Select top 200 companies with market cap less than 4000 in small cap companies

4.2 Top six companies as benchmark for analysis

As discussed large cap companies are fruitful to analyse. So, we decided to select top six companies to start the analysis. But in analysis we had too many parameters to find these companies. Based on observation the features like alpha, beta, five year company growth were selected. The five year company growth is the best parameter to analyse any company because company performance itself is best indicator of any stock, hence it is the highest priority parameter. Then the alpha which signifies how better the stock returns are, so it had the second highest priority. And last beta that speaks of volatility, which as discussed have great effect on stock, it had the third priority. Other parameters like sub-sector, closing price or PE ratio etc. were irrelevant with respect to calculating company stock performance. Hence the top six company stocks were deduced

1. Import the data of large cap company stocks
2. Filter out rows with invalid entries i.e if alpha or beta are not between 0 and 1
3. Select alpha, beta and 5 year growth as final attributes
4. Filter the companies First by beta parameter, then alpha and then by five year company growth.
5. Select the top six companies as the result.

The final six company stocks that performed best in last five years are as follows:

- LTI
- MINDTREE
- NAUKRI
- ASTRAL
- DEEPAKNTR
- PERSISTENT

4.3 Analysis on the top six company stocks

For analysing these company stocks we extracted the data as mentioned in 2.2. Each company data had 7 features as discussed. The different analysis is performed on the company stocks as discussed below

4.3.1 Change in price of the stock overtime

The graph plots the trends in closing price of stocks as follows:

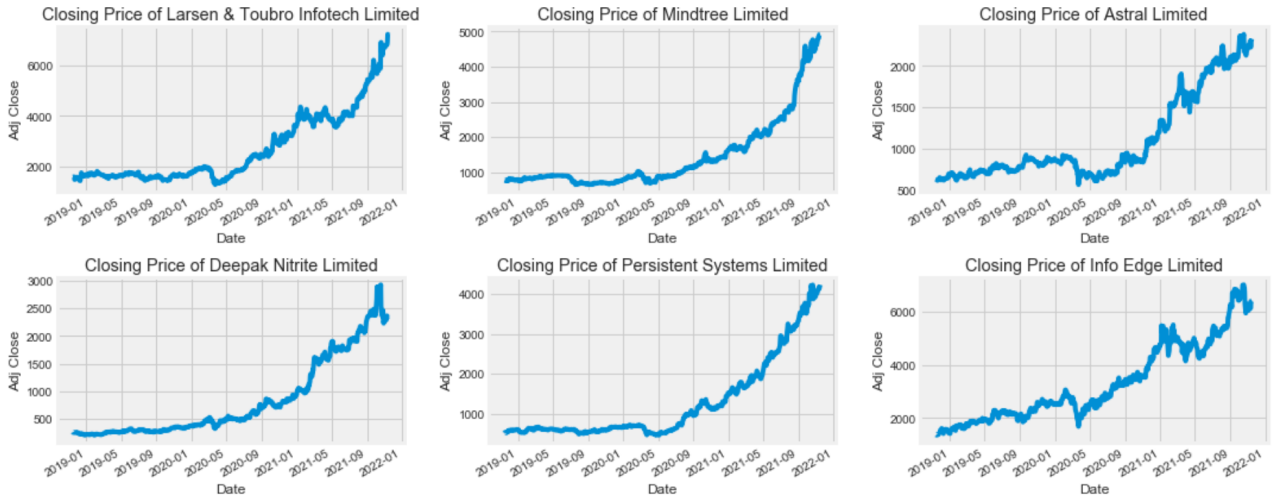


Figure 5: Stock's Closing price over time.

From the graph we can infer that in last three years the stocks had an exponential growth because large cap companies are least volatile and gives us good returns.

4.3.2 Stock's CAGR

The compound annual growth rate (CAGR) [7] is the rate of return (RoR) that would be required for an investment to grow from its beginning balance to its ending balance, assuming the profits were reinvested at the end of each period of the investment's life span. It is calculated as follows:

$$CAGR = \left(\frac{EndingValue}{BeginningValue}\right)^{(1/n)} - 1 * 100$$

The graph representing CAGR for each stock is plotted as follows:

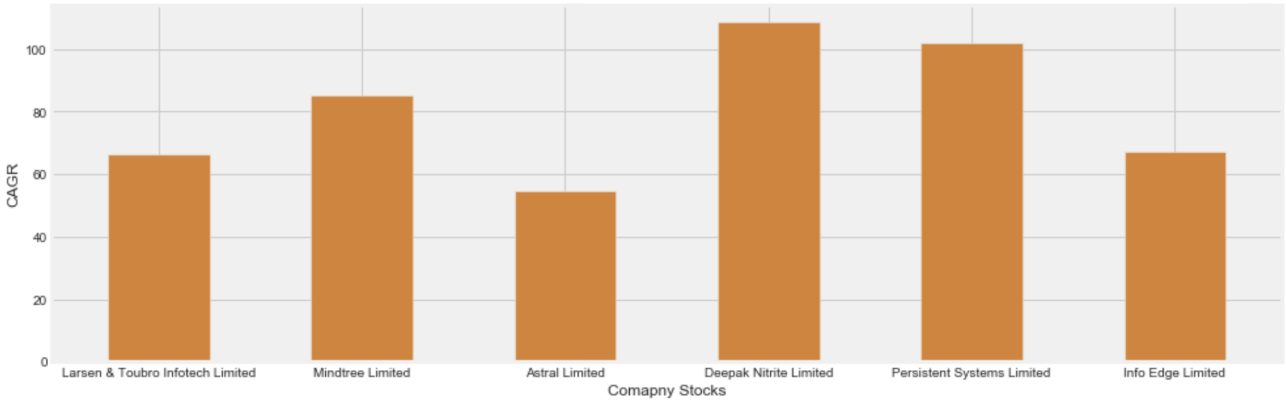


Figure 6: Stock's CAGR

Observations:

1. Deepak nitrite gives more that 100% average annual returns in these three years. It is a chemical company and we all know that covid had a great impact on humanity last year and demand of chemical(life saving chemicals) increased rapidly and this is well reflected in the stock price of this company.

2. L&T, Mindtree, Persitent, Indo edge are IT companies and in covid time IT industry was on high demand as this industry won't affect much from the lockdown and people are working from home hence we can see their returns are high .

3. Astral Limited is pipe(related to construction) company, we can observe that in 2020 it's stock price values decreased. Due to the lockdown construction was stopped throughout the nation but in 2021 as lockdown was removed and as new and pending construction was needed to be done we can see a high growth in its stock price and it keeps on increasing.

4.3.3 Stock's Volume Change

Every transaction that takes place between a buyer and a seller of a security contributes to the total volume count of that security. One transaction occurs whenever a buyer agrees to purchase what a seller is offering for sale at a certain price. If only five transactions occur in a day, the volume for that day is set at five. The graph representing volume trends is plotted as follows:

Figure 7: Stock's Volume change over time

We can see that Deepak Nitrite has high volume after mid 2020 because of covid people start investing in chemical industries as they are the most demanding during pandemic.

4.3.4 Stock's Simple Moving Average

The SMA features are extraction is mentioned in feature extraction part. Here is a demonstration to draw inferences. The graph representing the SMA trends is plotted as follows:

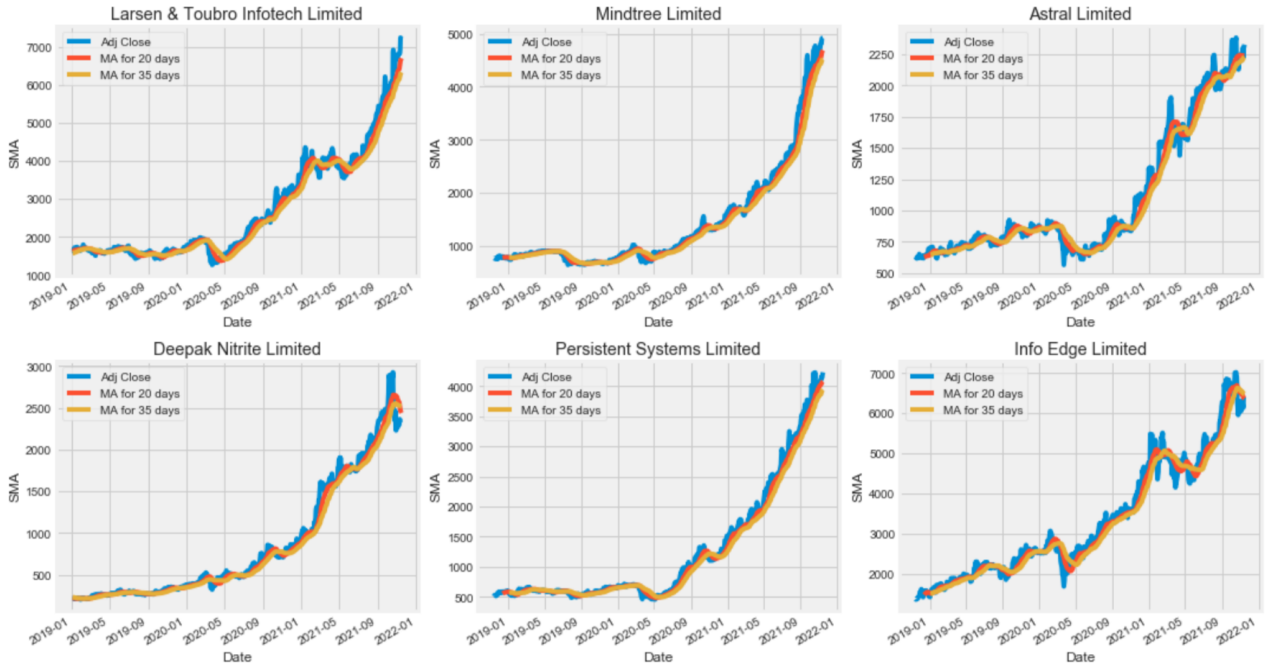


Figure 8: Stock's SMA over different interval

Trade Calls - Using Simple Moving Averages

- When price crosses up and over the moving average, *it's signal to buy.* When price crosses down and under the moving average *it's signal to sell.*
- Two popular trading patterns that use simple moving averages include the death cross and a golden cross. A death cross occurs when the 50-day SMA crosses below the 200-day SMA. This is considered a bearish signal, that further losses are in store. The golden cross occurs when a short-term SMA breaks above a long-term SMA. Reinforced by high trading volumes, this can signal further gains are in store.
- Call should be buy whenever the smaller moving average (20) crosses over longer moving average (35) AND the call should be sell whenever smaller moving average crosses under longer moving average. One of the most widely used technical indicators.

4.3.5 Stock's Daily Returns

The features that we extracted before had a feature called daily returns which shows the rate at which closing price changes everyday. The graph representing the trends in daily returns is plotted as follows:

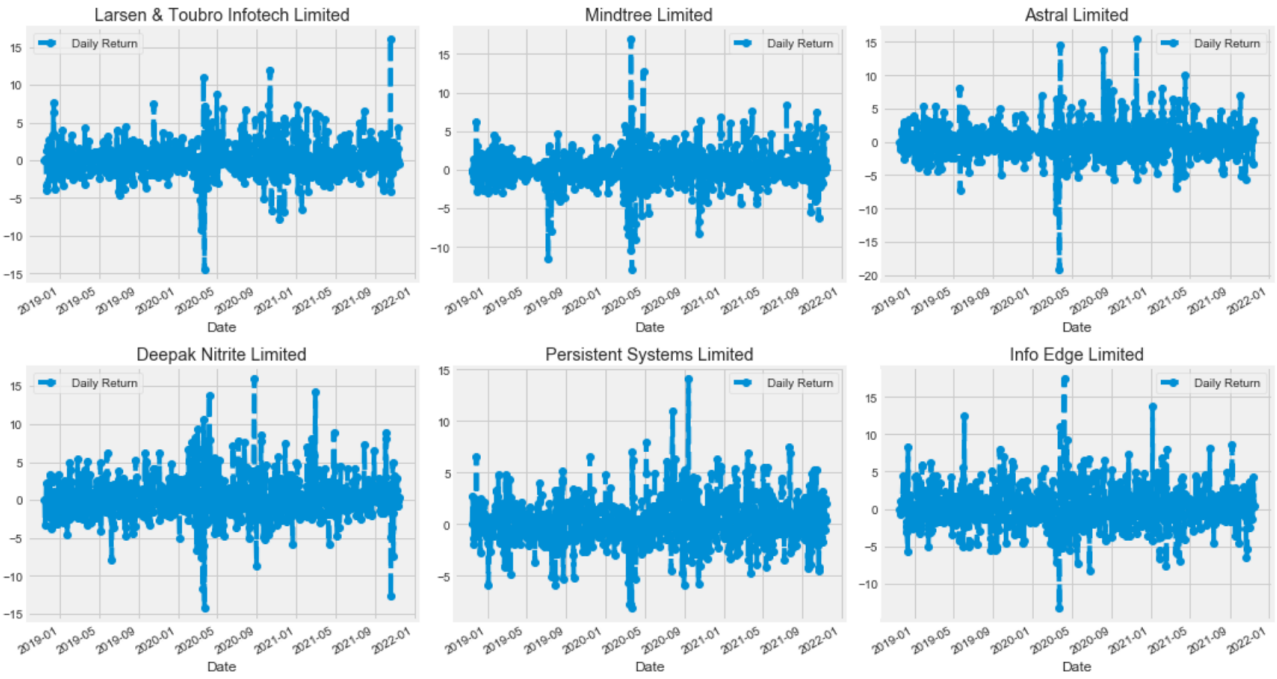


Figure 9: Daily returns of stocks over time

The histogram representing the frequency of different daily returns is plotted as follows:

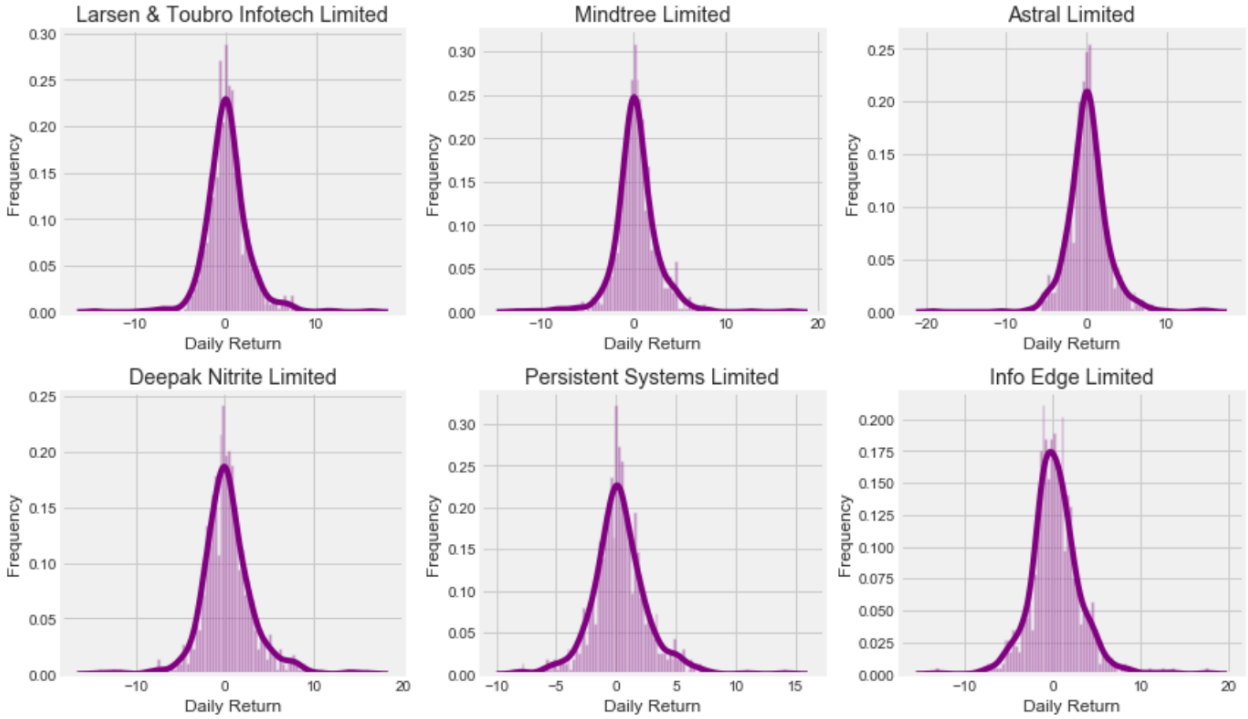


Figure 10: The frequency analysis of stocks daily returns

These graph shows that daily returns of these companies are in a fixed range not too high not too low. Hence we not need to check their performance too frequently we can check in a quarter or six month and then forget about them for next quarter. So these are best form of investment for those who don't have much market knowledge. AS we discussed large cap stocks are less volatile

4.3.6 Correlation between different stocks closing prices

The daily closing prices are compared with each other. To show the correlation among these stocks python jointplot, heatmap is used. The jointplot [8] draws a plot of two variables with bivariate and univariate graphs. If two stocks are perfectly (and positively) correlated with each other a linear relationship between its daily return values should occur. The heatmaps [9] are the graphical representation of values depicted using various shades. For plotting the correlation the daily percent change in closing price for each company is calculated. The graph representing the correlation between different stocks closing prices is plotted as follows:

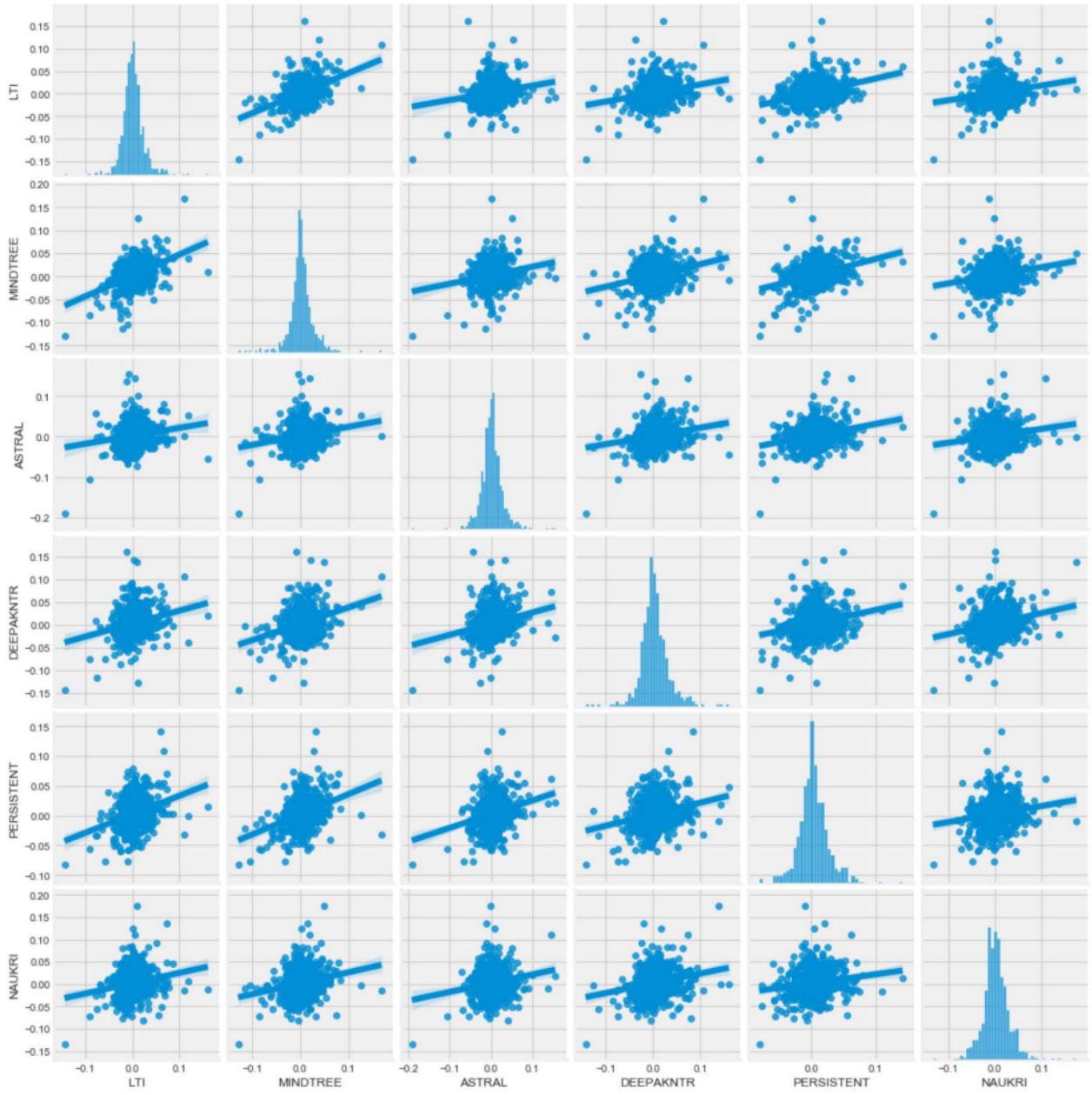


Figure 11: Jointmap representing correlation between stocks

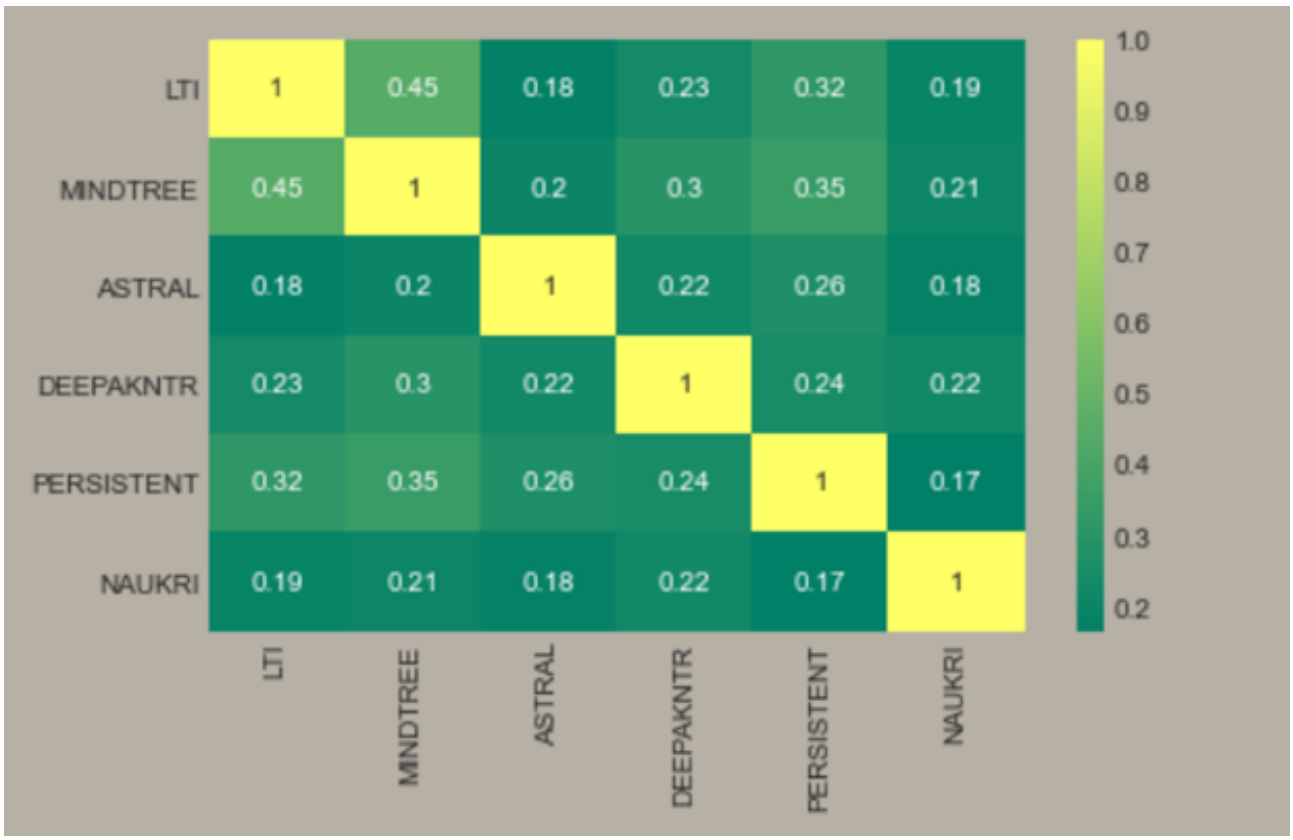


Figure 12: Heatmap of correlation between daily returns

The common conclusions can be derived from this plot such as:

- There is no positive correlation between any two companies, i.e. stock of one company do not affect / do not have relationship with other company.
- Investor can freely invest in all company because change in one stock does not affect other.Even all IT companies don't have correlation.
- All histogram plots look like marginal distribution while all scatter plot look like joint distribution between two different company.

4.3.7 Monthly highest percentage of top gains

The trends are classified, hence using these trends we found which month had highest number of days with top gains. For this we filtered out the days with trend 'Among top gainers', calculated the density of the trend in each month for each company stocks.

These graphs provide the good perspective of the months with top gains:

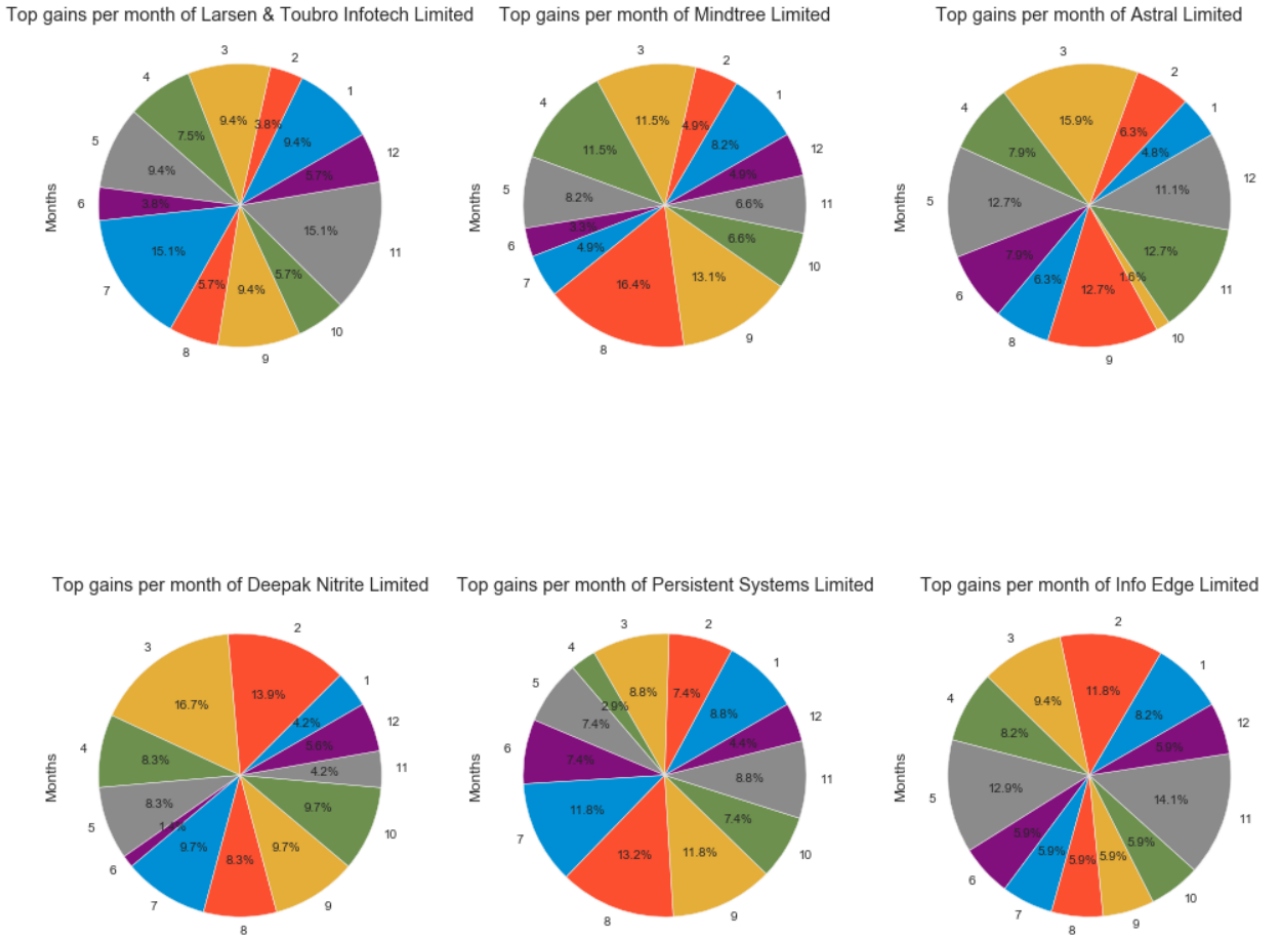


Figure 13: Monthly percentage of highest gain days

For all companies we see that stable gain days per month are nearly same hence, no such months are seen to have specifically high return months.

4.3.8 Stock Risk Analysis

There are many ways we can quantify risk, one of the most basic ways using the information we've gathered on daily percentage returns is by comparing the expected return with the standard deviation of the daily returns. Volatility is the change in variance in the returns of a stock over a specific period of time. In most cases, the higher the volatility, the riskier the security. Volatility is often measured as either the standard deviation or variance between returns from that same stock or market index. Volatile assets are often considered riskier than less volatile assets because the price is expected to be less predictable. The graph representing the risk based on daily average and standard deviation is plotted as follows:

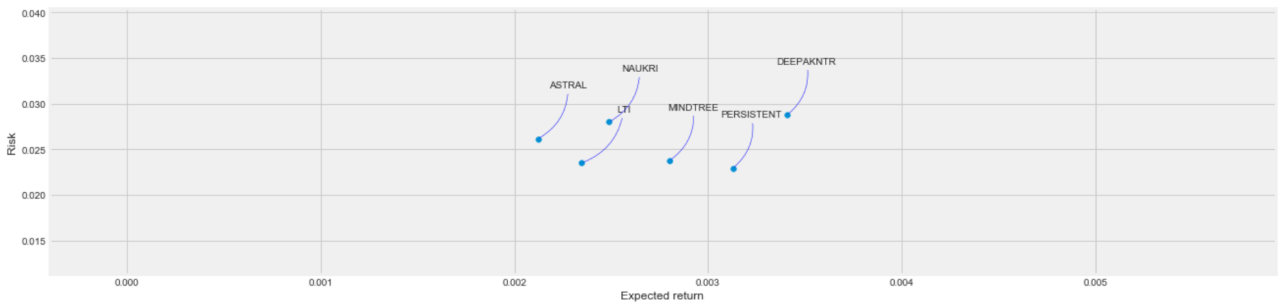


Figure 14: Simple risk analysis based on mean and standard deviation of daily returns

Risk analysis using Monte Carlo Simulation

The Monte Carlo [10] model makes it possible for researchers from all different kinds of professions to run multiple trials, and thus to define all the potential outcomes of an event or a decision. Monte Carlo analysis is a kind of multivariate modeling technique.Research analysts use them to forecast investment outcomes, to understand the possibilities surrounding their investment exposures, and to better mitigate their risks. The model uses the 4 parameters start price, number of days, mean, standard deviation. It just calculates a simple propagation from starting point and deviates by using mean and standard deviation for each iteration of day. As everything is random we can create more than one propagation. The graph representation of the risk analysis using Monte Carlo simulation is plotted as follows:

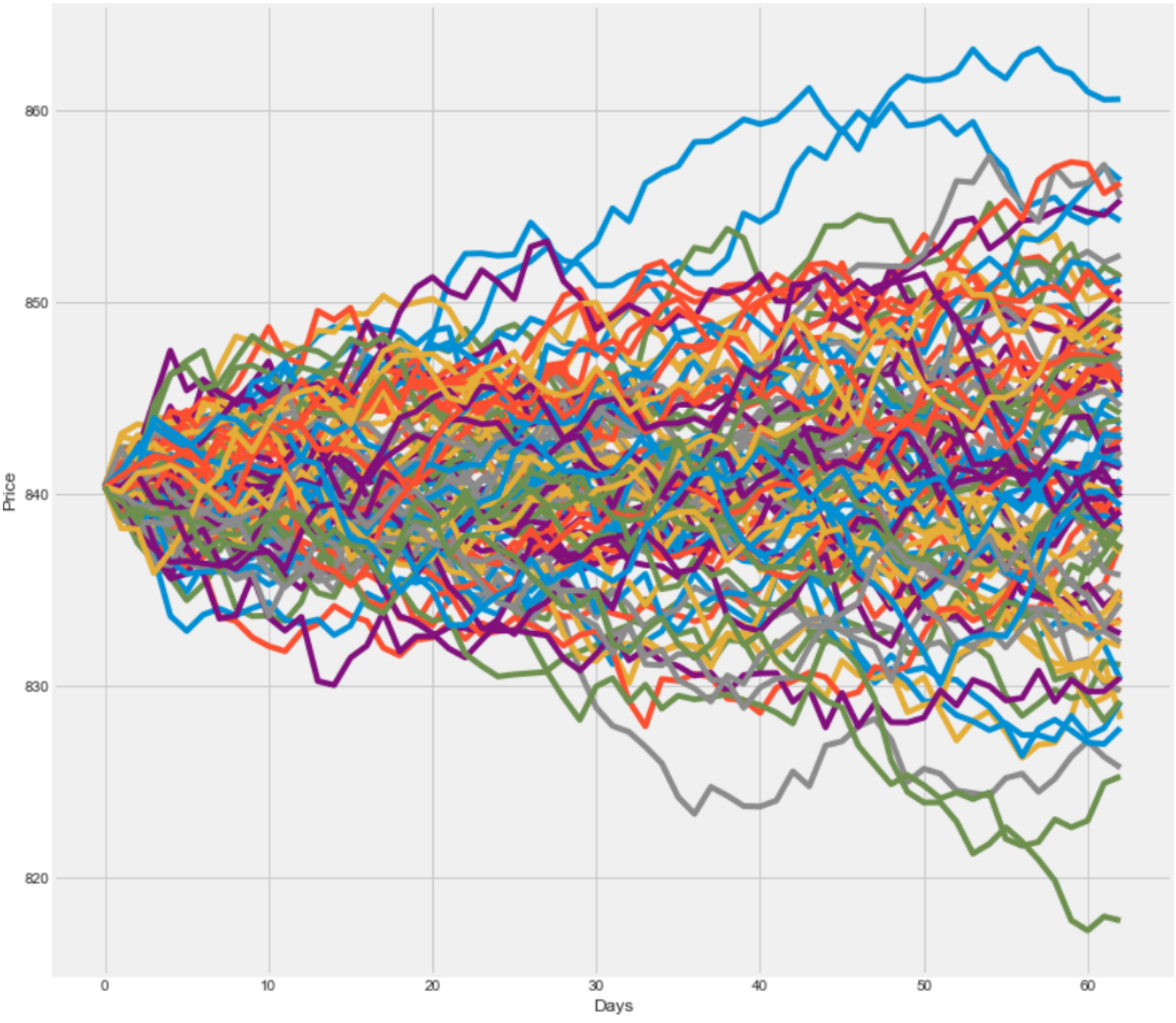


Figure 15: Risk analysis using Monte Carlo Simulations

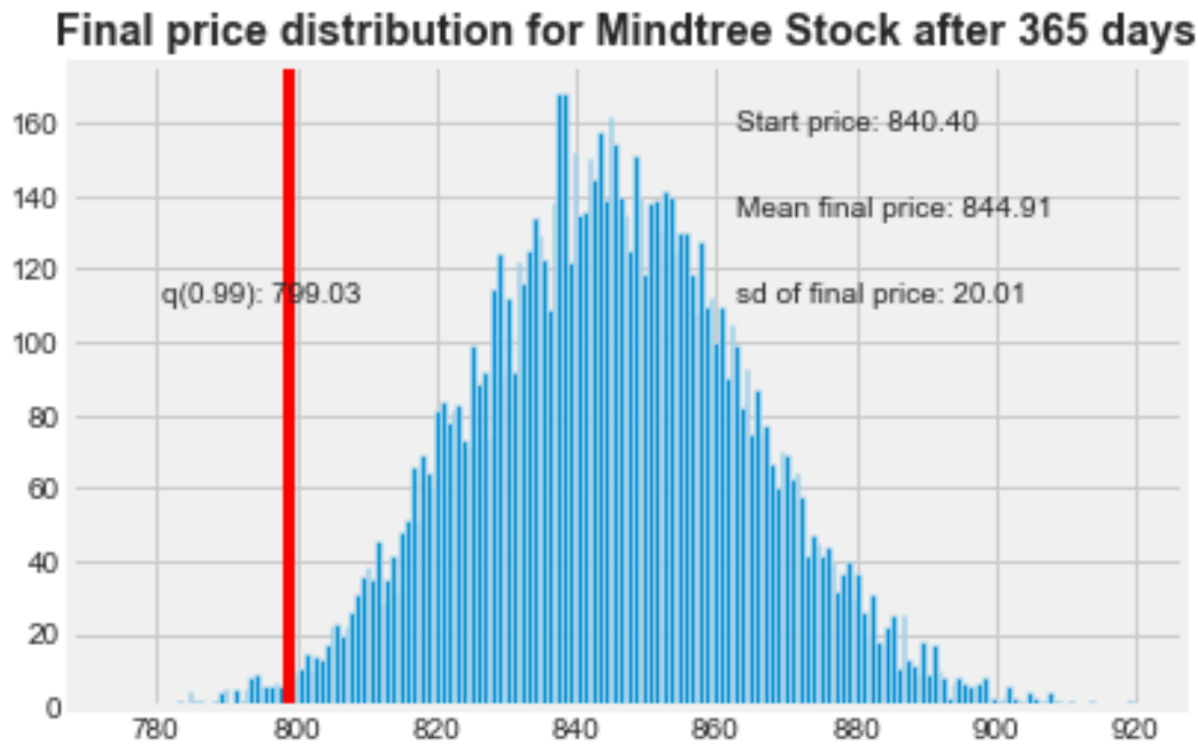


Figure 16: The Monte Carlo Simulation for 10,000 simulations with 1 percent impirical quantile

Generally frequencies of different outcomes generated by this simulation will form a normal distribution. The most likely return is in the middle of the curve, meaning there is an equal chance that the actual return will be higher or lower than that value. The probability that the actual return will be within one standard deviation of the most probable ("expected") rate is 68%, while the probability that it will be within two standard deviations is 95%, and that it will be within three standard deviations 99.7%.

4.3.9 The mean and median values of the Volume for each of the types of Trend in each stock

The mean and median are good indicators to relate to a certain trend. So, for every trend we can analyse a single value of volume and derive an inference. The table of output data is as follows:

Trends	LTI	Mindtree	Astral	Deepakntr	Persistent	Naukri
	Mean values					
Bear drop	260896	3210047	223317	2124859	227162	896982
Among top losers	347016	1760062	334973	991652	231041	587458
Negative	225728	1110874	199335	649567	137851	386071
Slight negative	151184	872712	153447	700013	122855	350021
Slight or No change	162065	783567	171927	605166	131622	308571
Slight positive	176112	1039725	200020	669820	157750	378974
Positive	271146	1309640	235884	839462	186547	452817
Among top gainers	607192	2247740	426655	1399417	351183	623125
Bull run	956896	4501073	883490	2940115	1080472	1407651

	Median values					
Bear drop	207466	1359731	235073	1893352	132424	746283
Among top losers	270725	1296888	240319	726172	150914	417878
Negative	180422	838557	145883	508522	121239	291505
Slight negative	110503	692667	98256	588393	97116	263363
Slight or No change	138844	646973	114024	444663	102235	280117
Slight positive	149117	903715	152446	433563	116569	254955
Positive	203297	1069039	156339	692815	153308	394123
Among top gainers	399879	1941360	282923	1170778	290743	491334
Bull run	659006	4412318	911670	2517684	774051	915556

Table 1: The mean and median values of volume for different trends

From the table we infer that volume increases as difference between closing price increases same may not be true for small cap stocks

4.3.10 The relationship between volume and daily return

Compared the daily returns and volume using simple plot. The graph representing the rends in volume and daily returns for past 100 days is as follows:

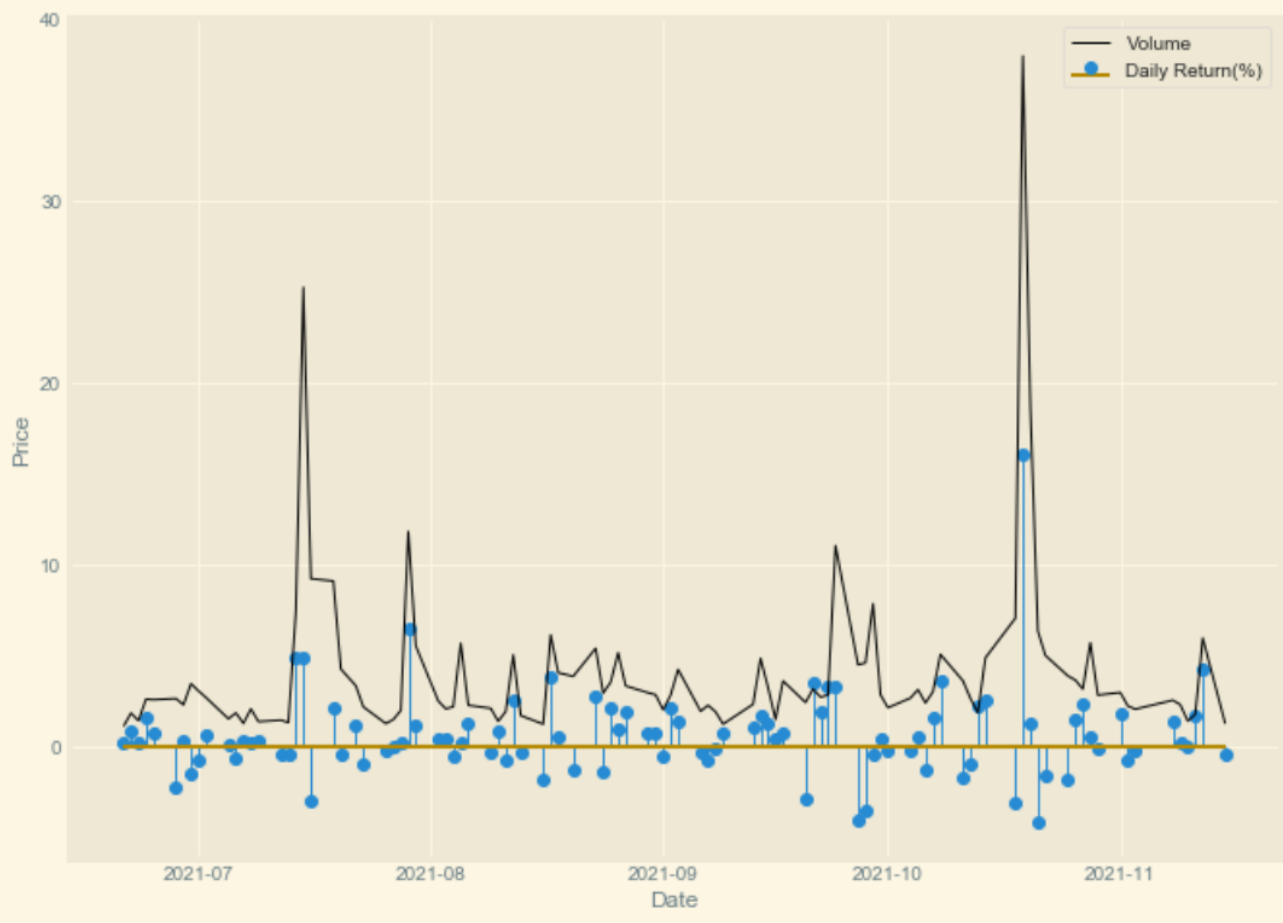


Figure 17: The relation of volume and daily returns over time

- As inferred from previous analysis. we can infer the same that both are interrelated for large cap stocks. When Volume suddenly increase, It means Daily return either increased or decreased which affect Stock price directly.
- When some news came up about company, it highly affect the stock price of the company. When there is good news about Company, People tend to buy the stock and stock price start to increase. Which means, at end of day Volume is large. It means Positive change happen in stock than previous day. When there is bad news about Company, People tend to sell the stock and stock price start to decrease. Which means, at end of day Volume is large. It means negative change happen in stock than previous day.

4.4 Analysis for LTI

For the analysing a stock individually we extracted more data for LTI company stocks and plotted some graphs to draw inferences.

4.4.1 The trends percentile

Each stock trends are classified into 9 categories. For this stock the percentage of times each trend occurred is plotted. The pie chart representing the trend percentile is plotted as follows:

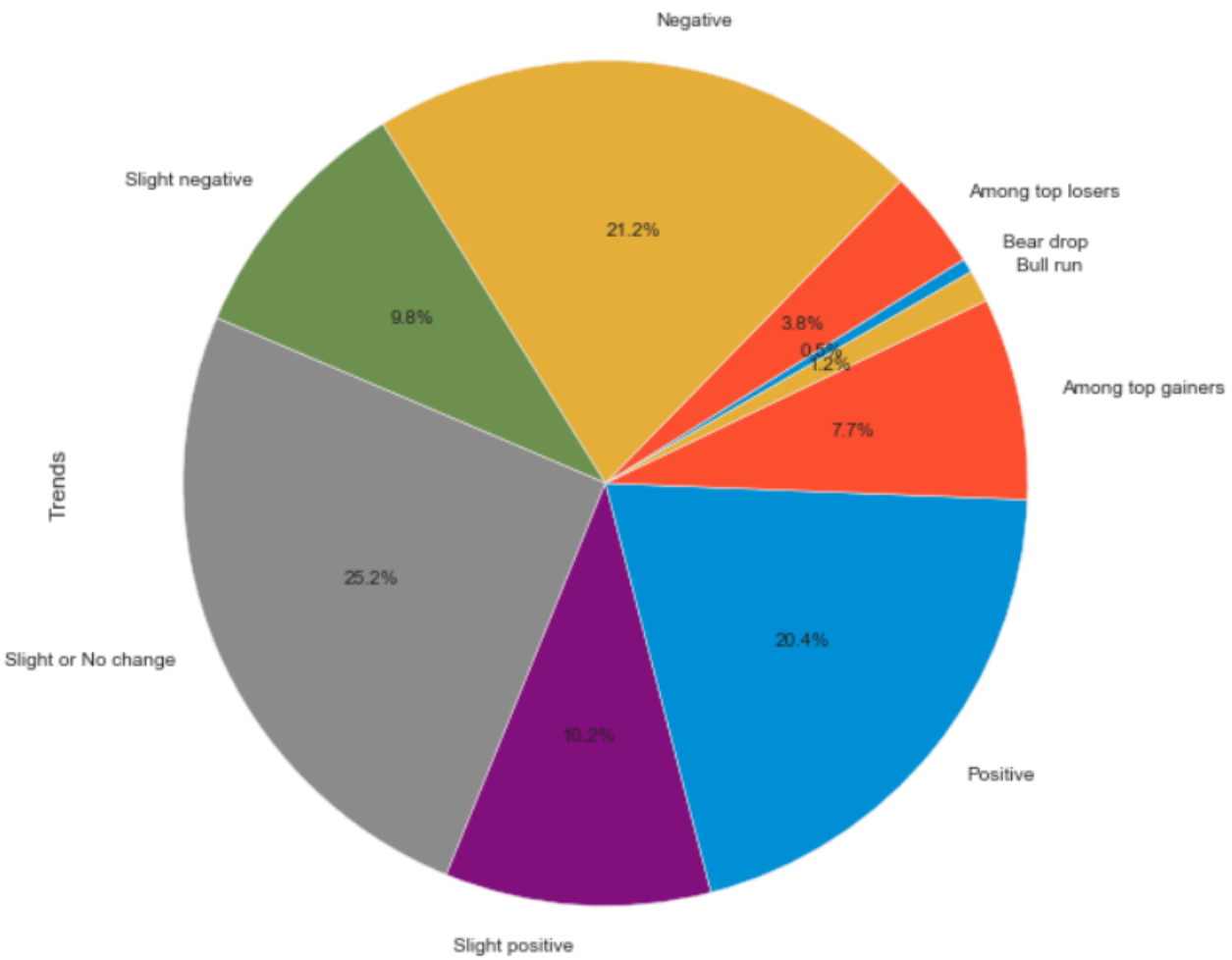


Figure 18: The trend percentile distribution

LTI has very small bull run and bear drop means it doesn't gives too high and too low returns it mostly fluctuate between positive, negative, or no change. This is exactly what expected from a large cap company.

4.4.2 The mean and median for volumes of different trends

The trends are compared to volume to infer if volume has any effect on trends. The bar graph representing the mean and median trends for each trend is plotted as follows:

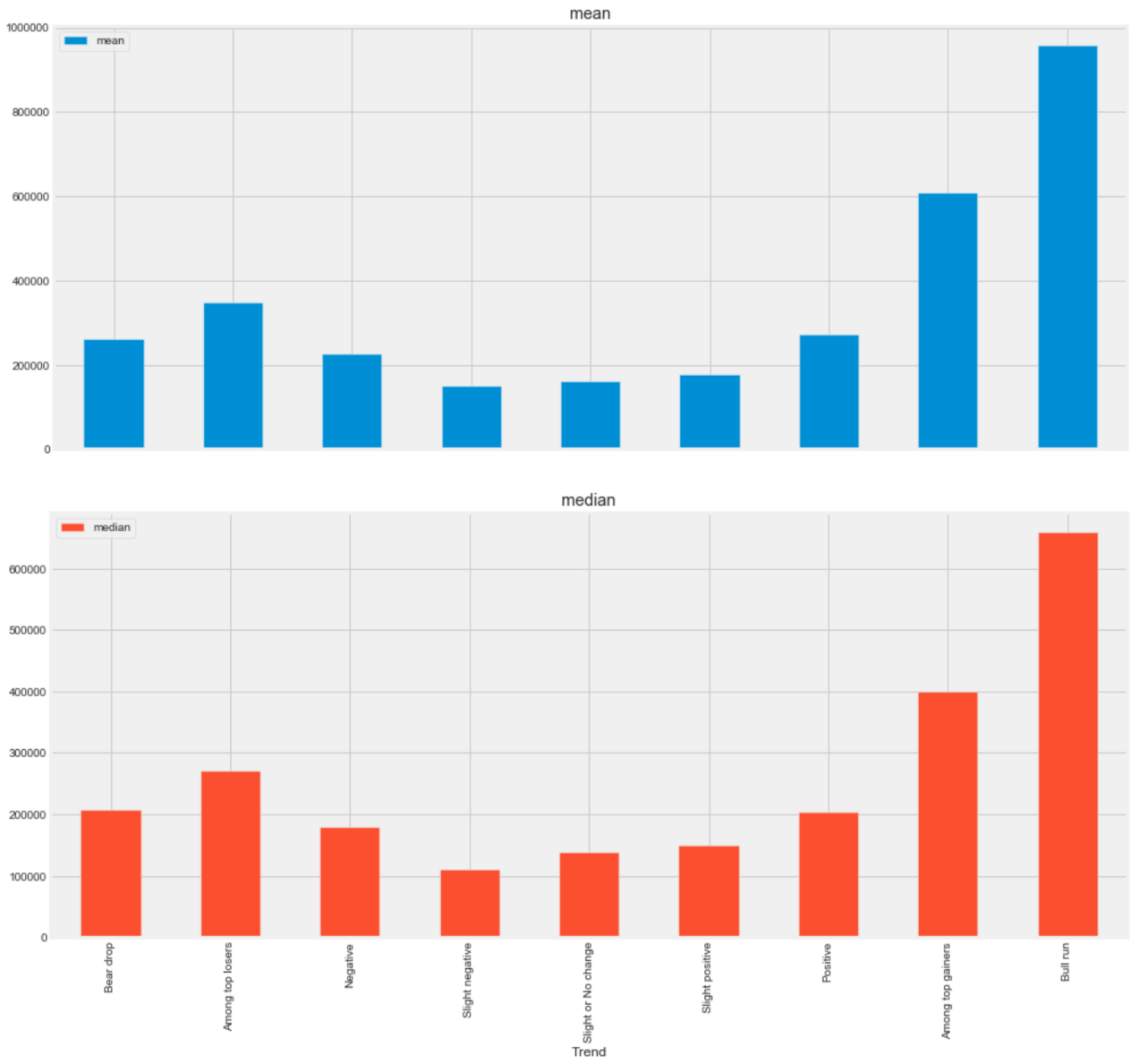


Figure 19: The mean and median for volumes over trends

The expected graph is that, values must increase as we move left/right from center i.e. as changes increases volume must increases. This is noticed on right side which is stock increase side and not on left side. This happened because we saw tremendous increase in stock and the drops in stock were nominal hence more unpredictable.

4.4.3 Trade Calls - Using Bollinger Bands

A Bollinger Band [11] is a technical analysis tool defined by a set of trend lines plotted two standard deviations (positively and negatively) away from a simple moving average (SMA) of a

stocks price, but which can be adjusted to user preferences. There are three lines that compose Bollinger Bands: A simple moving average (middle band) and an upper and lower band. A 14-day moving average would average out the closing prices for the first 14 days as the first data point. Bollinger bands are extremely reliable , with a 95% accuracy at 2 standard deviations. The graph representing Bollinger bands for different intervals are plotted as follows:

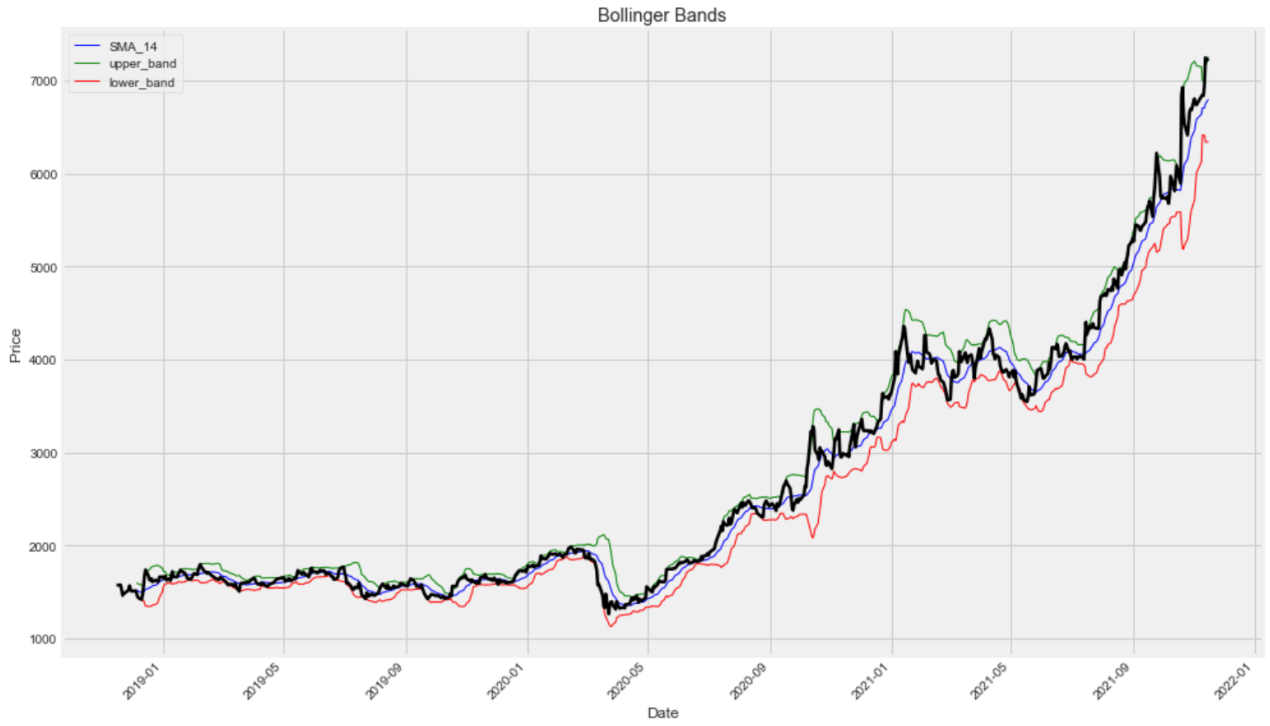


Figure 20: Bollinger bands

Many traders believe the closer the prices move to the upper band, the more overbought the market, and the closer the prices move to the lower band, the more oversold the market. We can observe that for most part closing price is between upper and lower bands means there is no issue of overbought and oversold the market.

4.4.4 Beta Calculation using regression

The Beta [12] of an asset is a measure of the sensitivity of its returns relative to a market benchmark.How sensitive/insensitive is the returns of an asset to the overall market returns. What happens when the market jumps, does the returns of the asset jump accordingly or jump somehow? For answering these questions we compared the beta values of LTI and nifty50 stocks. As discussed nifty decides a relative market status based on top 50 stocks. The graph representing beta trends for LTI and Nifty is plotted as follows:

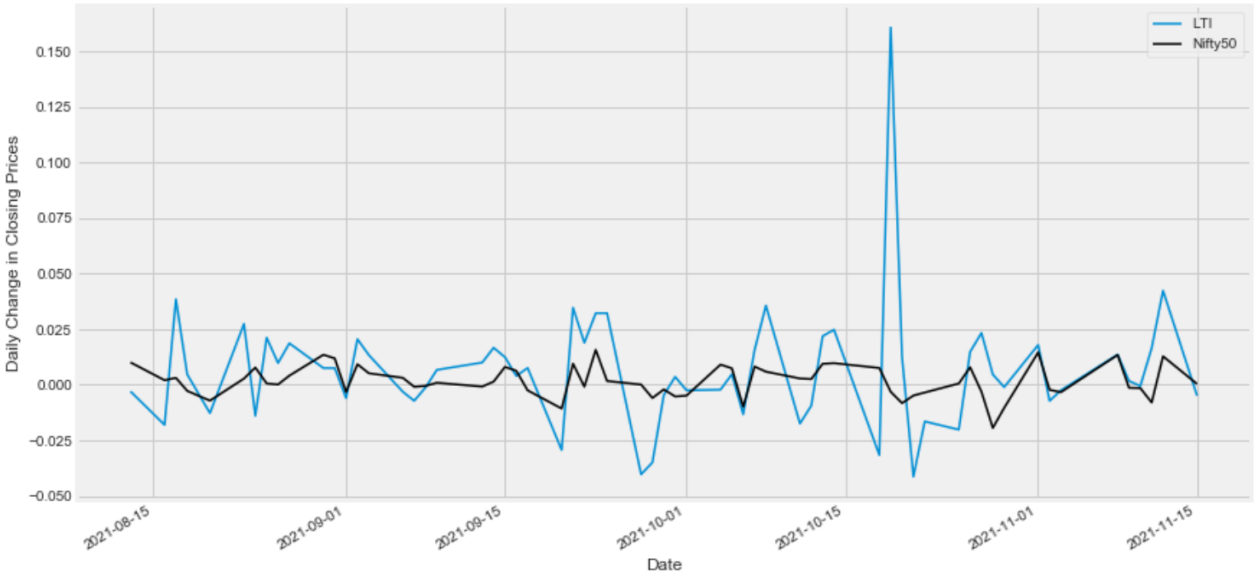


Figure 21: Relation of beta values between LTI and nifty

The graph infers that stock beta value almost tracks the same path as market beta value. For the plot we calculated the regression coefficient among both of them to capture similarity. It comes out to be 0.83. It means that it is less volatile with respect to market that we can also see in the above graph which is good for long term investment.

4.5 Analysis of 20 company's stock data

Here we compared many companies at a time to consider this performance as performance of large cap stocks in general. And find if they are similar or dissimilar in different respects.

4.5.1 Beta Calculation using regression

As we calculated regression coefficient for LTI and nifty, in similar way nifty and these top 20 companies are compared. The result is plotted in following table:

Stock Name	Regression coefficient
LTI	0.83
MINDTREE	1.51
ASTRAL	0.89
DEEPAKNTR	1.97
PERSISTENT	1.18
NAUKRI	1.38
VENKEYS	0.77
TATASTLLP	1.23
JKTYRE	1.38
RAYMOND	0.67
TATAMETALI	0.9
TCS	0.9
COGNIZANT	-0.1
INFOSYS	0.21
HCL	0.99
WIPRO	1.31

ONGC	1.3
RAJESHEXPORTS	0.25
ICICI	1.02
HDFC	0.81

Table 2: Regression coefficient for different stocks in comparison to nifty

The maximum value observed is 1.97 for DEEPAKNTR and minimum value observed is -0.105 for COGNIZANT. A negative beta correlation would mean an investment that moves in the opposite direction from the stock market. When the market rises, then a negative-beta investment generally falls. When the market falls, then the negative-beta investment will tend to rise.

Inferences from the Beta Values and Regression results:

- Beta measures the volatility of a stock compared with the volatility of the market as a whole.
- A high beta means the stock price will move faster than a stock with low beta. High beta means high volatility, but also the possibility of high returns.
- β=0: indicates no correlation with NIFTY or some chosen Index/Benchmark.
- β=1: shows a stock has equally sensitive as the market.
- $\beta >$ 1: indicates a stock that's more volatile/unstable than NIFTY.
- $\beta <$ 1: shows less sensitive than NIFTY 1.45 is 45% more sensitive than NIFTY
- We can observe that companies like Raymond, Venkeys are very less volatile.
- Companies like Deepakntr,Mindtree are relatively more volatile.
- It shows DEEPAKNTR stock is very volatile. It very nice option for Intraday trading but not for long term trading. Due to high volatility we can get a very big return on investment

4.5.2 Diversification analysis using Clustering

Clustering [13] is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. In financial Markets, Cluster analysis is a technique used to group sets of objects that share similar characteristics. It is common in statistics, but investors will use the approach to build a diversified portfolio. Stocks that exhibit high correlations in returns fall into one basket, those slightly less correlated in another, and so on, until each stock is placed into a category.

The elbow curve for selecting the optimal k for kmeans clustering is plotted as follows:

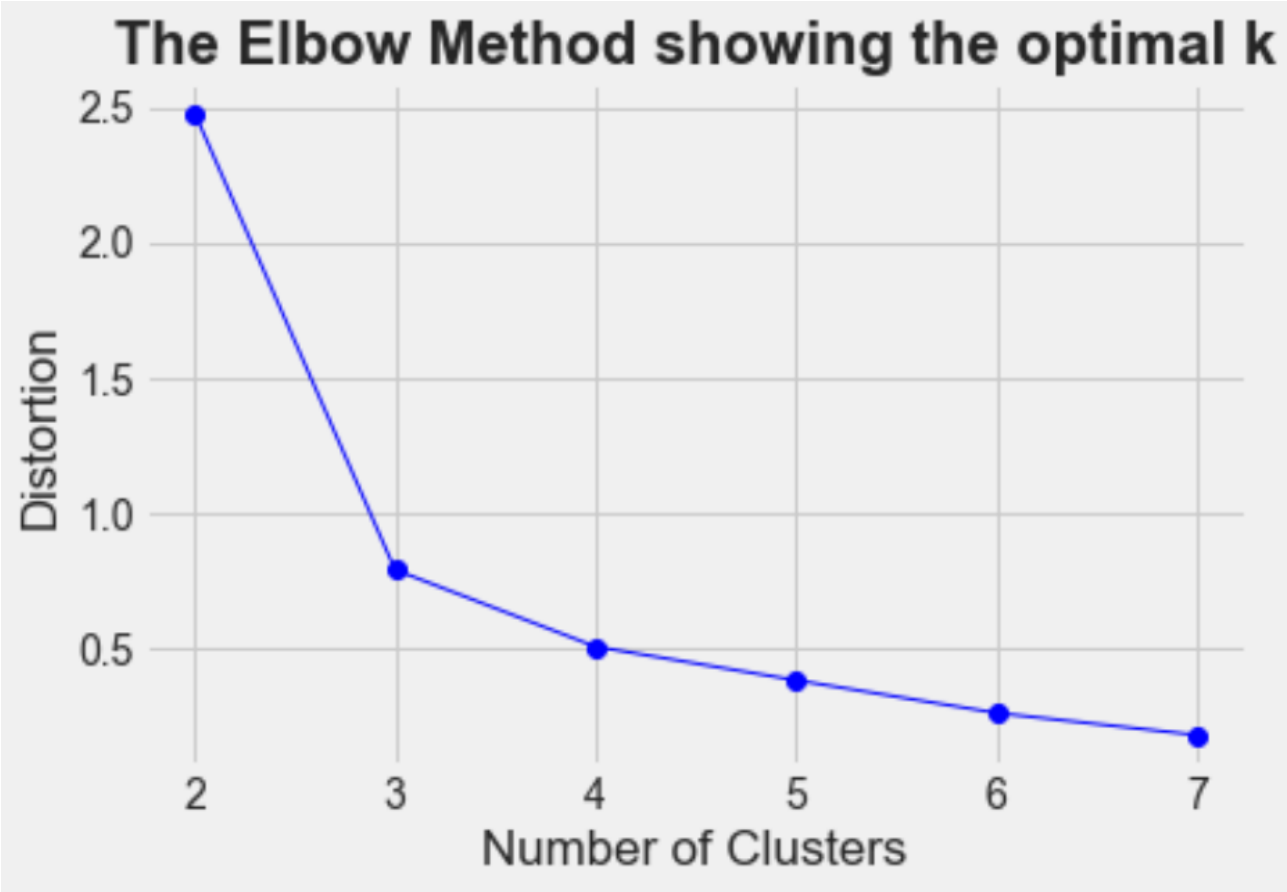


Figure 22: Elbow graph

The Number of clusters From above graph, it looks like 3 or 4 would be better number of cluster. Let's, plot scatter plot for 3 and 4 cluster. And decide. The graph representing the clustering with k=3,4 is plotted as follows:

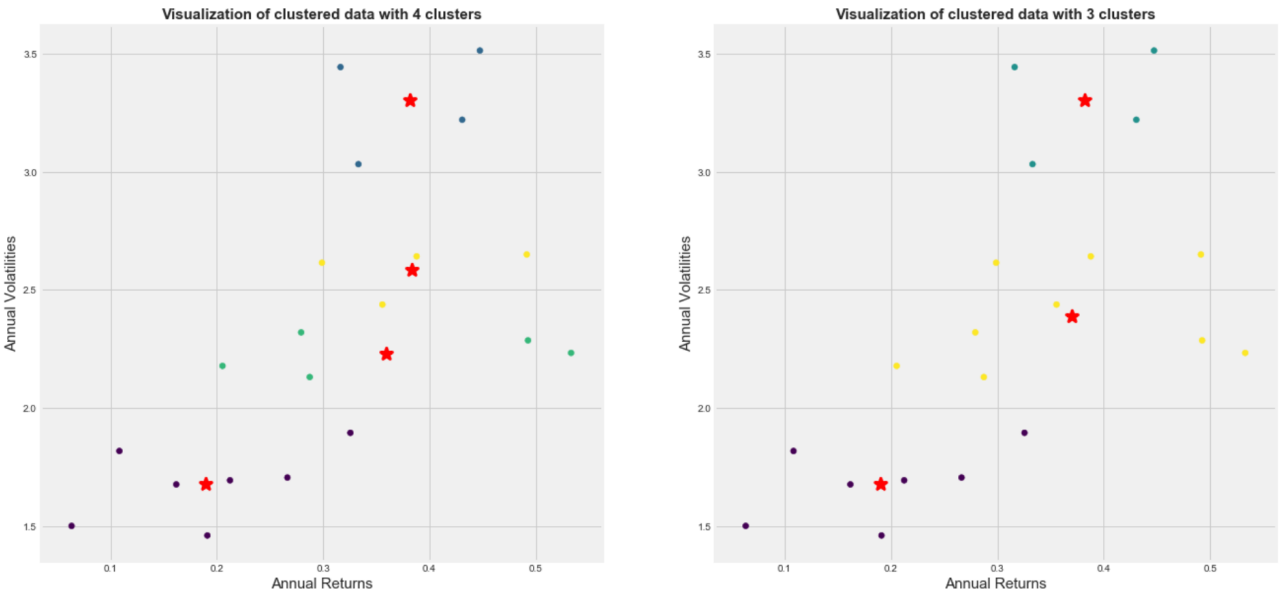


Figure 23: The cluster graphs for k=3,4

3 clusters seperation looks better than 4 clusters,so we go for three clusters. The Final visualization of clustering with stock labels is as follows:

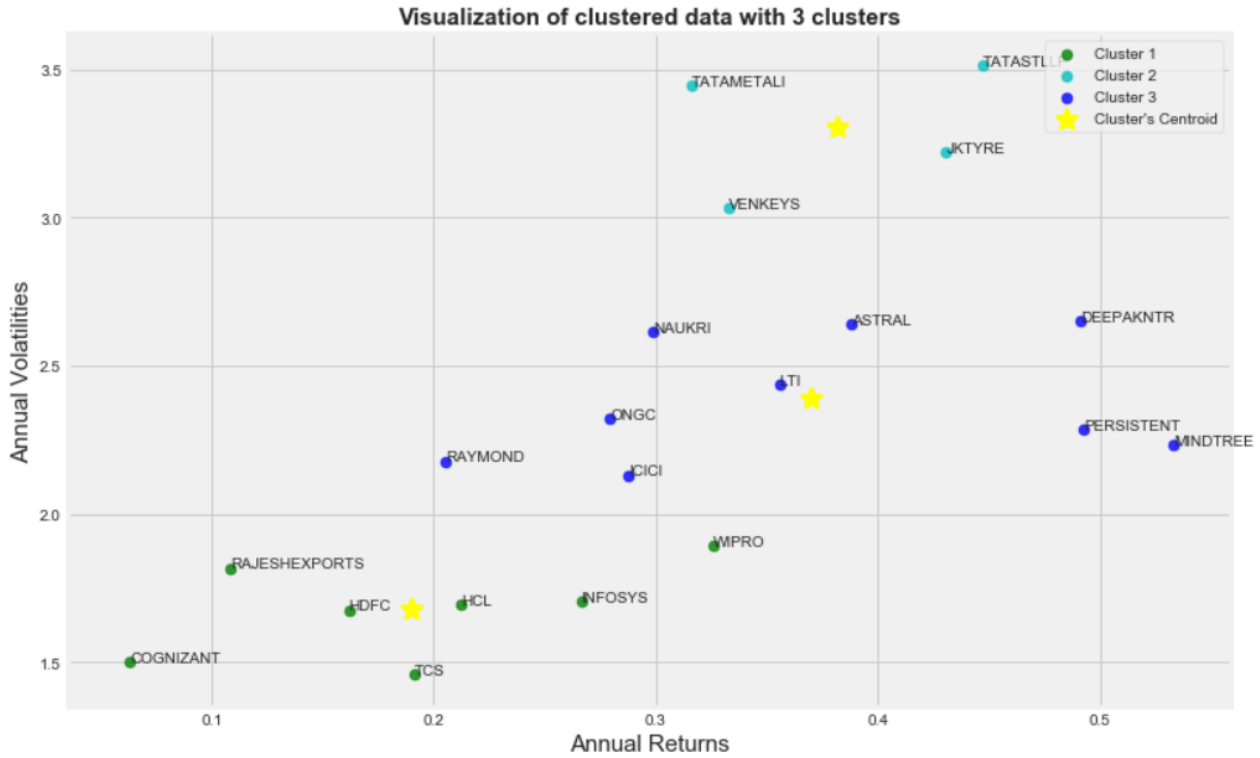


Figure 24: The visulization of clusters with stock names

We can observe that cluster1 companies are less volatile but also gives us less returns. Cluster2 companies are relatively high returns that cluster1 and also not so high risky/volatile.(These companies we found above as best companies to invest as they gives high returns and also less risky). Cluster3 companies gives moderate returns but high risky. Someone according to their risk taking capacity can invest in these different cluster companies. But best approach will be to split the investment amount and invest in all these cluster according to their risk taking capacity and expected returns.

5 Stock Prediction

For the users the most important part is prediction. Like when should one invest how the company may perform etc. For this purpose we train different models on the data of different companies and find the suitable model for prediction. And present the users with useful predictions

5.1 Trade Call Prediction using Classification

For classification we created a new category called 'Call'. The category specifies what should have been the ideal call the user should have taken to achieve highest interest over the stocks. The different categories used to classify are as follows:

- 'Buy' if the stock price is below the lower Bollinger band
- 'Hold Buy/ Liquidate Short' if the stock price is between the lower and middle Bollinger band
- 'Hold Short/ Liquidate Buy' if the stock price is between the middle and upper Bollinger band
- 'Short' if the stock price is above the upper Bollinger band

Now we will train different classification model with the 3 bollinger columns and the stock price as inputs and 'Calls' as output. The steps are as follows:

1. Classify the available data into given categories using Bollinger bands.
2. Split the data into trining and testing sets.
3. Select the different models suitable for classification.
4. Train each model on training set and test it using testing dataset.
5. Compare the accuracies and select the best model for prediction

The accuracy of different algorithms is tabulated in following table:

Classifier	Train Accuracy	Test Accuracy
Naive Bayes	0.48	0.55
Logistic Regression	0.96	0.97
SVM	0.46	0.45
KNN	0.93	0.81
Random Forest	0.58	0.55

Table 3: The accuracy of different models for trade call classification

Logistic regression gives the best accuracy, hence we use this classification model.

5.2 Predicting the closing price stock price of mindtree limited

For stock prediction Long short-term memory (LSTM) [14] is a best fit. Long short-term memory is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feed forward neural networks, LSTM has feedback connections LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.

The steps for prediction are as follows:

1. Read the data
2. Scale the data using MinMaxScaler
3. Split the data in training and testing sets
4. Build the LSTM model
5. Compile and train the model
6. Test the accuracy of model using testing set

The graph representing the predicted values in comparison to test values is plotted as follows:

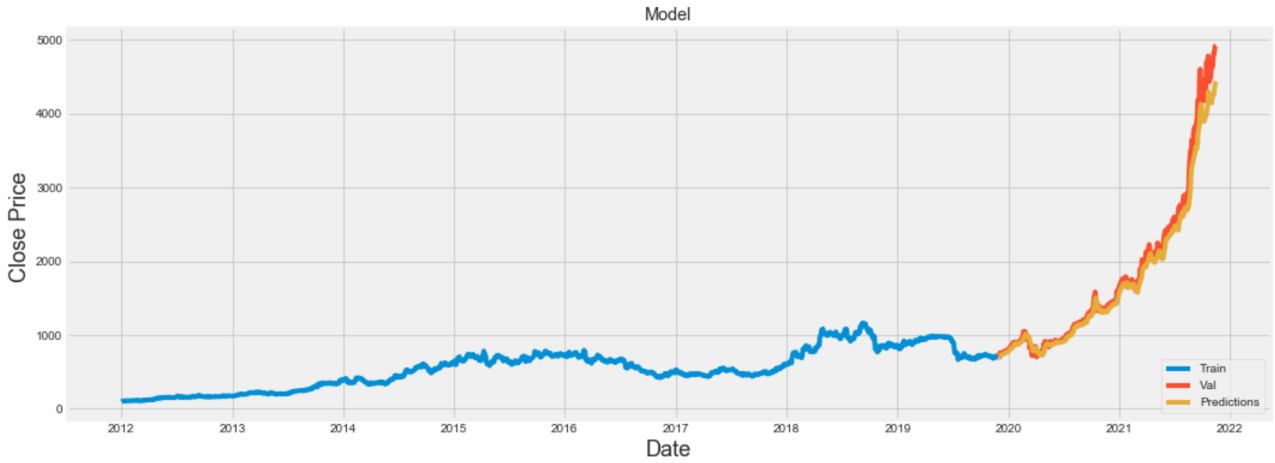


Figure 25: The closing price prediction using LSTM

After testing the model the Root Mean Square Error(RMSE) comes out to be 164.41. This is very low comparing to high closing price values.

6 Results

Many inferences are drawn from the analysis and prediction, and are mentioned after the explanation of each part. This section summarises all the result in brief. The results are limited to certain top companies. The derived results are as follows:

- **Data extraction**: The data is extracted perfectly from websites efficiently without any errors. But as always it has null or invalid values which are handled effectively.
- **Stock Classification**: Stocks are classified into large cap, middle cap and small cap classes for better analysis.
- **Find the top companies**: Sorted the companies based on different parameters like beta, alpha, five year growth and got top companies.
- **Compared closing price and volume**: The exponential growth is noticed in closing price of all companies. The volume of companies had different trends based on their subsector, eg: We can see that Deepak Nitrite has high volume after mid 2020 because of covid people start investing in chemical industries as they are the most demanding during pandemic.
- **Analysing Simple Moving Average for different intervals**: This analysis can be used to make different trade calls. It is used to decide trade calls i.e when price is greater than SMA sell stock and when it is less buy it. And as observed in analysis the stock was always greater hence anytime was best to sell the stock.
- **Compare daily returns of stocks**: The frequency of least change in daily returns was always high which denotes cap large cap stocks are less volatile. Hence, no need to keep a continuous look at large cap stocks
- **Correlation between stocks**: There is no positive correlation between any two companies, i.e. stock of one company do not affect / do not have relationship with other company. Investor can freely invest in all company because change in one stock does not affect other.Even all IT companies don't have correlation.
- **Monthly percentage of top gains**: For all companies we see that stable gain days per month are nearly same hence, no such months are seen to have specifically high return months.
- **Risk Analysis**: Risk analysis is done by two methods simple and monte carlo. The outcomes generated by this simulation will form a normal distribution. The most likely return is in the middle of the curve, meaning there is an equal chance that the actual return will be higher or lower than that value. The probability that the actual return will be within one standard deviation of the most probable ("expected") rate is68%, while the probability that it will be within two standard deviations is 95%, and that it will be within three standard deviations 99.7%.
- **The relationship between volume and daily return**: Both are interrelated for large cap stocks. When Volume suddenly increase, it means daily return either increased or decreased which affect stock price directly.
- **LTI stock analysis**: LTI has very small bull run and bear drop means it doesn't gives too high and too low returns it mostly fluctuate between positive, negative, or no change. This is exactly what expected from a large cap company. The mean and median values

for the stocks are least when slight changes in volume are noticed. Similarly , volume increases as change increases , but change may be negative or positive. The volume on positive side as stock saw a large rise in market.

- **Trade calls using bollinger bands**: Many traders believe 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. We can observe that for most part closing price is between upper and lower bands means there is no issue of overbought and oversold the market.

- **The relation between top companies and nifty using regression coefficient**: The maximum value observed is 1.97 for DEEPAKNTR and minimum value observed is-0.105 for COGNIZANT. A negative beta correlation would mean an investment that moves in the opposite direction from the stock market. When the market rises, then a negative-beta investment generally falls. When the market falls, then the negative-beta investment will tend to rise. We can observe that companies like Raymond, Venkeys are very less volatile. Companies like Deepakntr,Mindtree are relatively more volatile. It shows DEEPAKNTR stock is very volatile. It very nice option for Intraday trading but not for long term trading. Due to high volatility we can get a very big return on investment.

- **Diversification analysis using Clustering**: Cluster1 companies like HCL, HDFC, ICS, Cognizant etc. are less volatile but also gives us less returns. Cluster2 companies like Venkeys, TataSTL, KTYRE etc. are relatively high returns that cluster1 and also not so high risky/volatile.(These companies we found above as best companies to invest as they gives high returns and also less risky). Cluster3 companies like ICICI, Raymond, LTI, Persistant etc. gives moderate returns but high risky. Someone according to their risk taking capacity can invest in these different cluster companies. But best approach will be to split the investment amount and invest in all these cluster according to their risk taking capacity and expected returns.

- **Trade call prediction using classification**: Logistic regression gives the best accuracy, hence we use this classification model. This model predicts perfectly whether to hold, buy or sell the stocks.

- **Predicting the closing price stock price using LSTM**. After testing the model the Root Mean Square Error(RMSE) comes out to be 164.41. Thisis very low comparing to high closing price values. Hence can be used for predicting closing price for any required day.

7 Conclusion and Future work

The stock market is a trending topic and current market is on boom. And hence many companies have taken the advantage by creating online applications which can work as brokers, have made tremendous profits out of it. But everyone is focused on certain analysis and predictions which may not be sufficient enough for people to judge the company background and invest into the companies. The people had put the blind faith in these systems and invest with the front these applications present. So, we built the system to overcome these drawbacks.

The system extracts the live data to keep the system upto date with perfect latency, so that analysis provides live performance.The system focuses on thoroughly analysing a certain company using many different kinds of methods. Instead of opting for quantitative analysis i.e analysing each and every stock in market, we opted for qualitative analysis i.e select limited but best companies to invest. We used more than one alternative for same analysis to ensure that result is not contradictory. Using the analysis user can easily study the company background and setup a perfect foreground before investing. Furthermore, the company performance is compared to market performance so we can refer to market to infer the company performance. Additionally, the risk of investment is analysed to tell how the steps taken here onward would result. Ten thousand simulations are ran to predict the different risks a stock may present. The companies are classified in different clusters to see similarities and dissimilarities among the different companies.

The system also provides the predictions for different company stocks. The most asked question about stock is what should be my trade call, should I hold the onto the stock, buy more stocks or sell the stocks. Our system answers to all these questions with high confidence using classification methods. Additionally, The system predicts the closing price for any input of date with best accuracy using LSTM model. It is the best model for stock prediction.Hence concluding the report, the system live up to its expectation. It extracts, analyses and predicts the stock market data efficiently and effectively.

For future work the current models can be integrated to perform all operations on any kind of company stock and create a dynamic framework to present the work in user friendly manner.

References

[1] 2021 Market Status Article. https://economictimes.indiatimes.com/markets/stocks/news/3-reasons-why-2021-could-be-a-mirror-image-of-2020-for-dalal-street/articleshow/81324438.cms?from=mdr, 2021. [Online].

[2] The overall performance data of national stocks. https://www.tickertape.in/screener, 2021. [Online].

[3] The daily performace data of different companies. https://finance.yahoo.com/, 2021. [Online].

[4] Simple Moving Average. https://www.investopedia.com/terms/s/sma.asp, 2021. [Online].

[5] Volume Weighted-Averge Price. https://www.investopedia.com/terms/v/vwap.asp, 2021. [Online].

[6] Articles related to stocks. https://www.investopedia.com/stocks-4427785. [Online].

[7] Compound Annual Growth rate. https://www.investopedia.com/terms/c/cagr.asp, 2021. [Online].

[8] Joint Plot. https://www.geeksforgeeks.org/python-seaborn-jointplot-method/, 2021. [Online].

[9] Heatmap. https://www.tutorialspoint.com/python_data_science/python_heat_maps.htm, 2021. [Online].

[10] Risk analysis by Monte Carlo. https://www.investopedia.com/articles/financial-theory/08/monte-carlo-multivariate-model.asp, 2021. [Online].

[11] Bollinger Bands. https://www.investopedia.com/terms/b/bollingerbands.asp, 2021. [Online].

[12] Regression Coefficient. https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/, 2021. [Online].

[13] The portfolio diversification. https://www.investopedia.com/articles/03/072303.asp, 2021. [Online].

[14] The stock prediction using LSTM. https://www.analyticsvidhya.com/blog/2021/05/stock-price-prediction-and-forecasting-using-stacked-lstm/, 2021. [Online].