FA-541

Applied Statistics with Application in Finance

PROJECT REPORT

ON

MARKET VOLATILITY PREDICTION USING S&P 500 INDEX USING S&P 500 Index.

Under the guidance of-

Prof. Zhenyu Cui

Submitted by-

Abhimanyu Sheth
Jayesh Kartik
Chandra Gopi Reddy
Dheeraja Katram, MS(FIN)

Contents

INTRODUCTION

S&P 500 Index-

S&P 500 is the joint venture of S&P Dow Jones Indices' registered trademark, often known as the Standard & Poor's 500. The 500 largest U.S. corporations make up this stock index, which is generally regarded as the most accurate gauge of the state of the market for American equities. From a different perspective, the S&P 500 index serves as a statistical gauge of the performance of the 500 largest stocks in the United States. The S&P 500 serves as a common standard against which the performance of a portfolio can be measured in this situation. Market capitalization determines the weighting of the S&P 500 index. This implies that the degree of influence a firm has over the performance of the index depends on its valuation. Not just 1/500th of the index is represented by each listed company. The S&P 500 index is more impacted by large corporations like Apple (NASDAQ: AAPL) and Amazon (NASDAQ: AMZN) than by comparably sized firms like Macy's (NYSE: M) and Harley-Davidson (NYSE: HOG).

One crucial issue is that, even though there are 500 significant enterprises, the valuations differ widely. The market capitalizations of several of the index's biggest companies exceed $1 trillion. The lowest S&P 500 companies, with market capitalization between $6 billion and $7 billion, are more than 200 times smaller than this. Throughout the trading day, the S&P 500 index's value swings continuously depending on performance-weighted market data for the underlying companies.

S&P 500 Dataset-

We used the head and tail function on a dataset of S&P 500 data from Yahoo Finance for the dates between 01/02/1990 and 10/10/2022. These are the columns that make up the S&P 500 data: Date, Open, High, Low, Close, Adj. Close, and Volume.

Date	Open	High	Low	Close	Adj Close	Volume
1990-01-02	353.399994	359.690002	351.980011	359.690002	359.690002	162070000
1990-01-03	359.690002	360.589996	357.890015	358.760010	358.760010	192330000
1990-01-04	358.760010	358.760010	352.890015	355.670013	355.670013	177000000
1990-01-05	355.670013	355.670013	351.350006	352.200012	352.200012	158530000
1990-01-08	352.200012	354.239990	350.540009	353.790009	353.790009	140110000
...	...	...	...	...	...	...
2022-10-03	3609.780029	3698.350098	3604.929932	3678.429932	3678.429932	4806680000
2022-10-04	3726.459961	3791.919922	3726.459961	3790.929932	3790.929932	5146580000
2022-10-05	3753.250000	3806.909912	3722.659912	3783.280029	3783.280029	4293180000
2022-10-06	3771.969971	3797.929932	3739.219971	3744.520020	3744.520020	4252100000
2022-10-07	3706.739990	3706.739990	3620.729980	3639.659912	3639.659912	4449660000

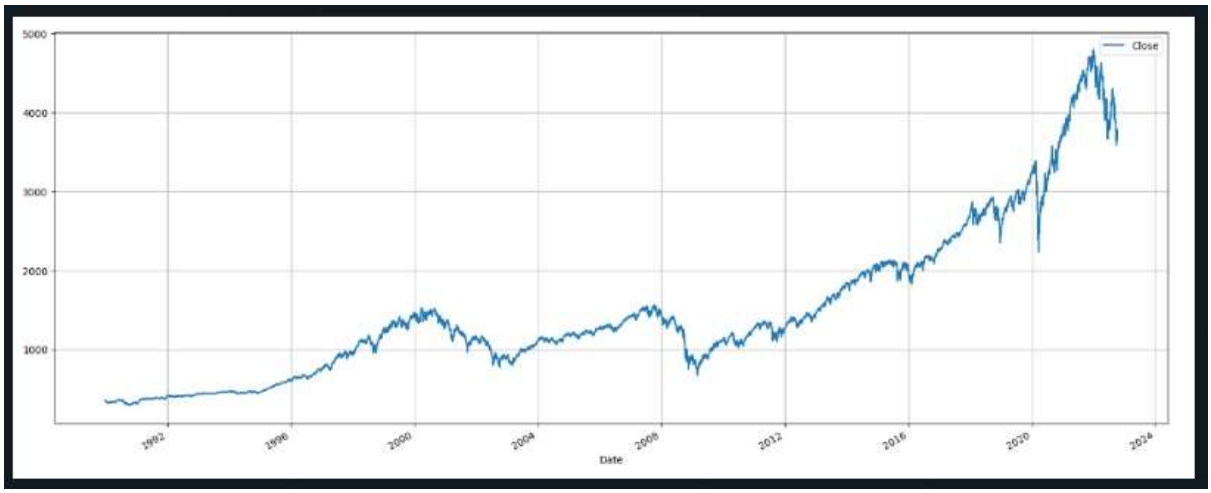
Summary Statistics of S&P 500 Index-

To have a broader understanding of the dataset and to decide which quantitative operations need to be performed on the dataset, we used the summary function to calculate the summary of the dataset.

```
count    8257.000000
mean     1486.147807
std       981.251312
min       295.459991
25%       876.770020
50%      1245.859985
75%      1922.030029
max      4796.560059
Name: Adj Close, dtype: float64
```

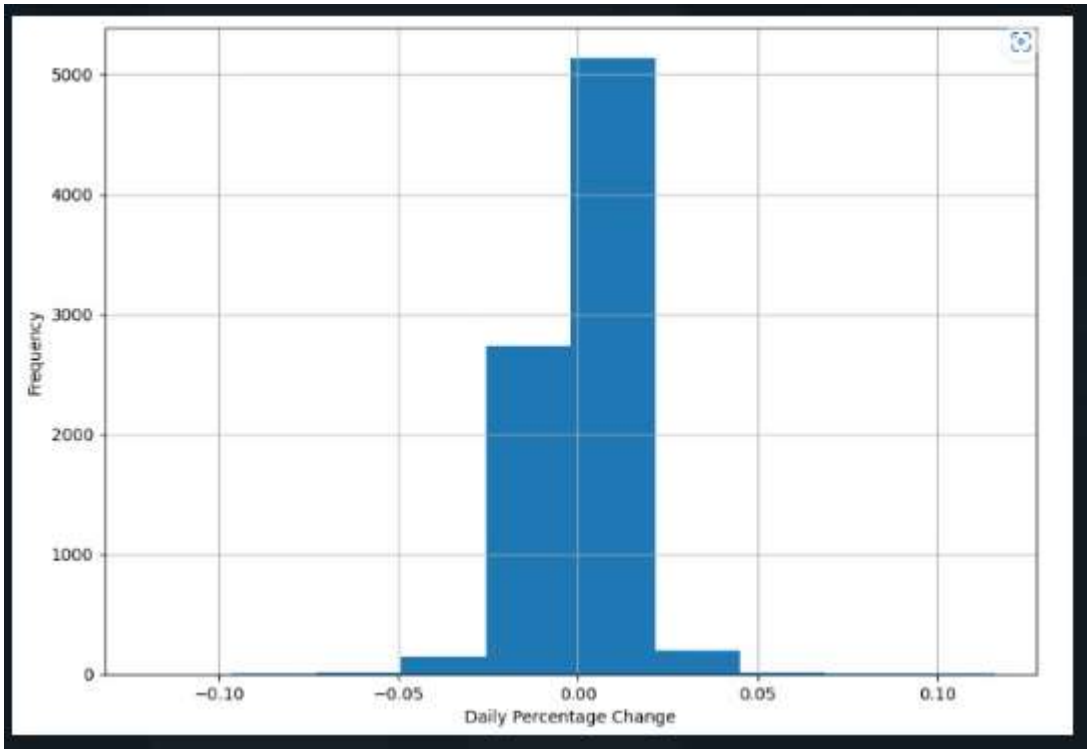
The close price movement of the S&P 500 over the last 22 years

As we know, the close price is the most widely used parameter by investors for investing in a stock. It gives the true price of the stock. This graph, it is shown how the close price of the S&P 500 data has changed over the period of 22 years. The increase in price also confirms the popularity and continuous investing by investors in the last 22 years.



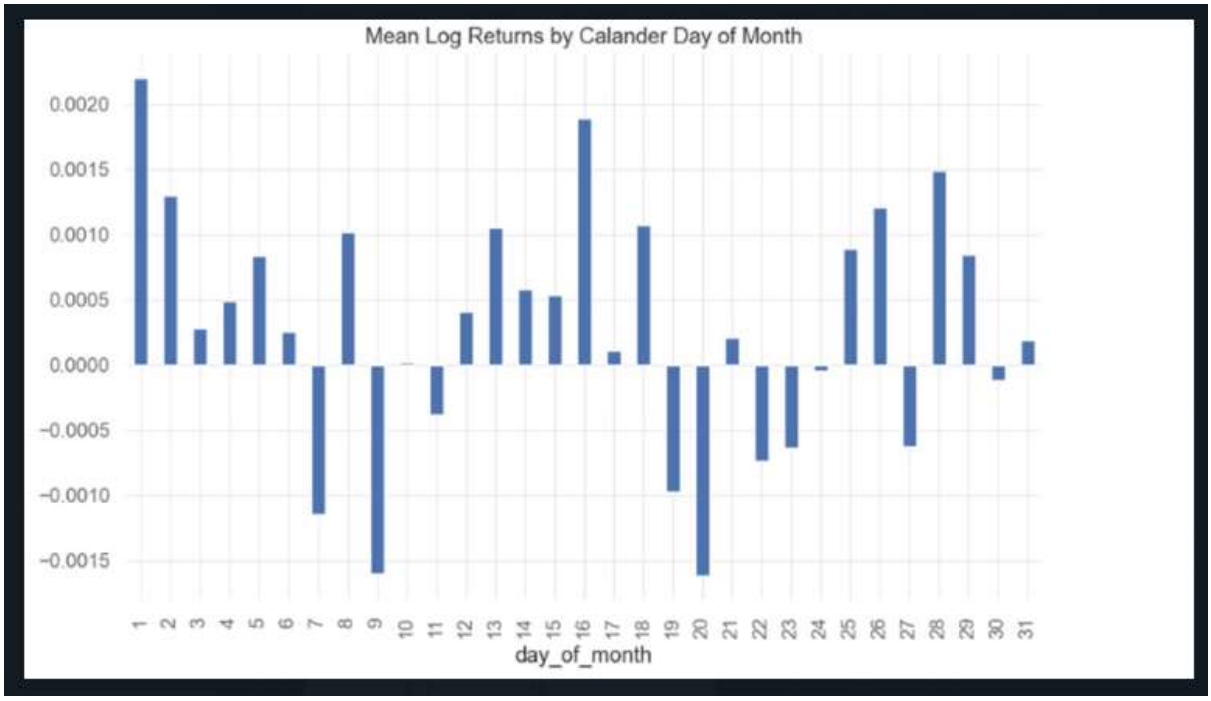
Daily percentage change in the S&P 500's closing price

As we all know, when buying stocks, investors most frequently utilize the adjusted close price. It displays the stock's actual price. This graph illustrates how, during a 22-year period, the adjusted close price of the S&P 500 data has changed. The price growth is further evidence of the popularity and ongoing investment of investors over the past 22 years.



Mean log closing price returns

The S&P 500 index's mean log return was determined, and the graph that follows shows how the index behaved on various days of the month. Positive returns demonstrate that the market acted in the investors' favor, while negative values demonstrate that the market did not act in the investors' favor because of the numerous socioeconomic activities taking place throughout the world.



Usage of S&P 500 Index

The reason the S&P 500 is regarded as being so valuable as a market and economic indicator is that it comprises a diverse basket of stocks without too many small or obscure businesses, and it contains the companies that individual investors own in the greatest numbers. Approximately 80% of the total value of the American stock market is made up of the 500 largest corporations.

VIX

The market's expectations for the relative strength of impending price swings of the S&P 500 Index are reflected in the Volatility Index (VIX), a live index (SPX). It produces a 30-day forecast of volatility because it is drawn from the pricing of SPX index options with close-in expiration dates. Volatility, or how quickly prices fluctuate, is frequently used to assess market emotion, particularly the level of anxiety among traders. This index is frequently referred to as "the VIX" despite being more well known by its ticker symbol. It is managed by CBOE Global Markets and was developed by the CBOE Options Exchange.

- Traders can price derivatives using VIX values or they can trade the VIX using a range of options and exchange-traded products.
- The VIX often goes up when stocks are down and down when they are up.

VIX Dataset-

Now it is essential to know the risk associated with the S&P 500 data. To know it, we have taken the VIX data of the same date range and following is a small screenshot of the VIX data.

```
vix.head()
```

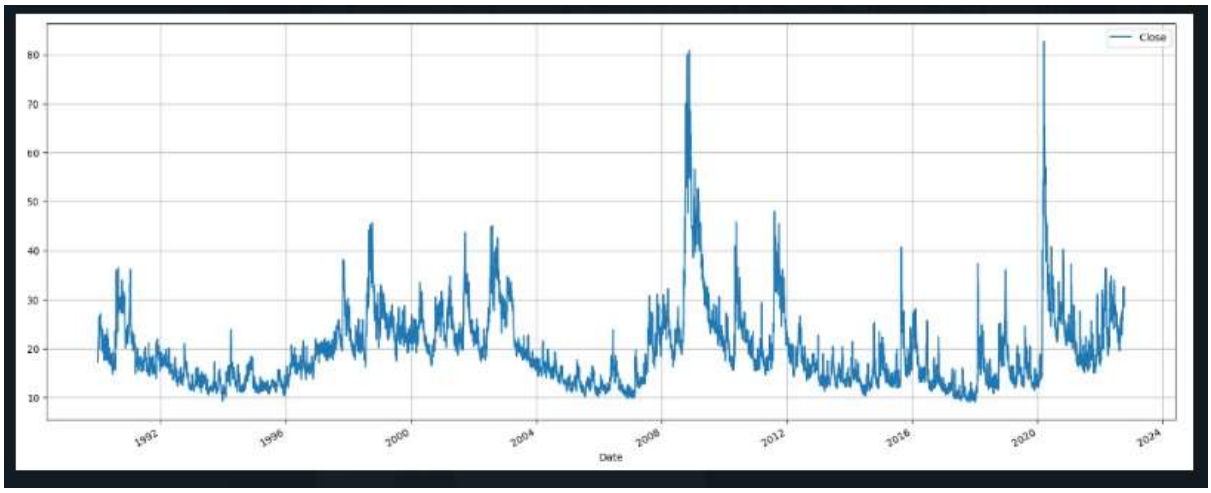
Date	Date	Open	High	Low	Close	Adj Close	Volume
1990-01-02	1990-01-02	17.240000	17.240000	17.240000	17.240000	17.240000	0
1990-01-03	1990-01-03	18.190001	18.190001	18.190001	18.190001	18.190001	0
1990-01-04	1990-01-04	19.219999	19.219999	19.219999	19.219999	19.219999	0
1990-01-05	1990-01-05	20.110001	20.110001	20.110001	20.110001	20.110001	0
1990-01-08	1990-01-08	20.260000	20.260000	20.260000	20.260000	20.260000	0

```
vix.tail()
```

Date	Date	Open	High	Low	Close	Adj Close	Volume
2022-10-04	2022-10-04	29.520000	29.620001	28.559999	29.070000	29.070000	0
2022-10-05	2022-10-05	29.360001	30.110001	28.500000	28.549999	28.549999	0
2022-10-06	2022-10-06	28.600000	30.740000	28.559999	30.520000	30.520000	0
2022-10-07	2022-10-07	30.370001	32.020000	29.879999	31.360001	31.360001	0
2022-10-10	2022-10-10	32.930000	33.990002	32.049999	32.450001	32.450001	0

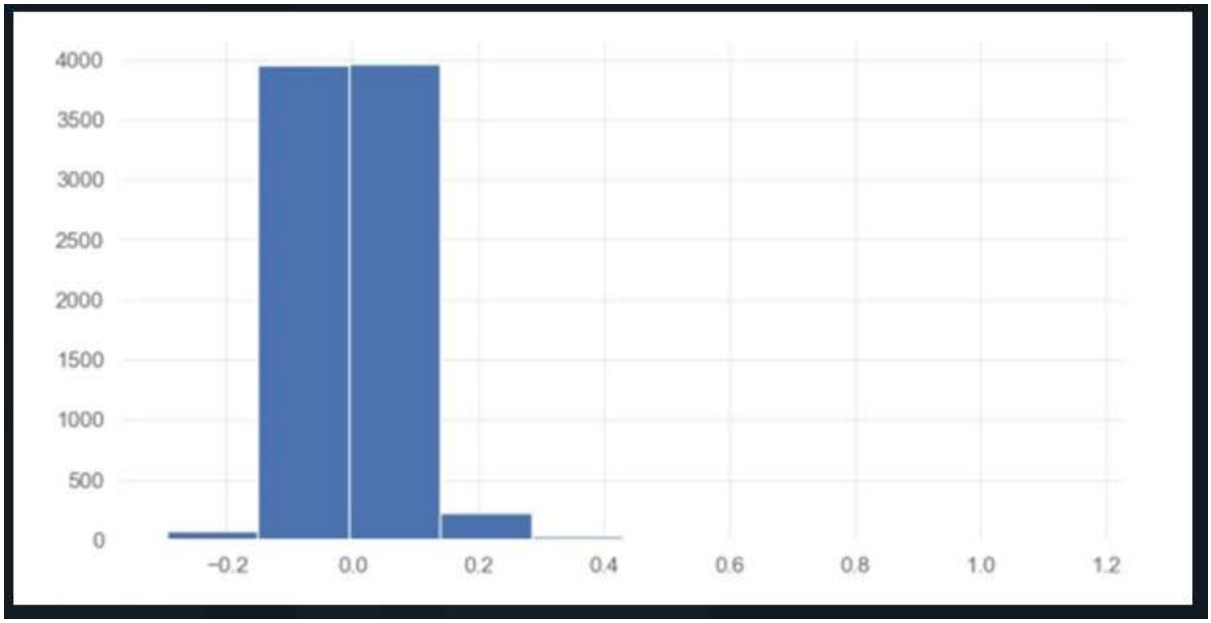
The movement of VIX data Close Price in the last 22 years-

It is ideally known that whenever the volatility increases in the market, the scenario of the market is not good. The following picture gives us proof of the fact that in 2008, because of the recession, volatility reached its peak. Similarly, the market was not ready for Covid-19 and the steep increase in the S&P 500 adjusted close price validates the bad scenario of the market. If we recall, everyone was trying to pull their money out of the market and the liquidity of the market was increasing thus making the market dry.



Daily Percentage change in VIX's Closing Price-

As we already saw that the prices of S&P stock changed daily so does the volatility associated with it. Following is a snapshot of the daily percentage change in the VIX data.

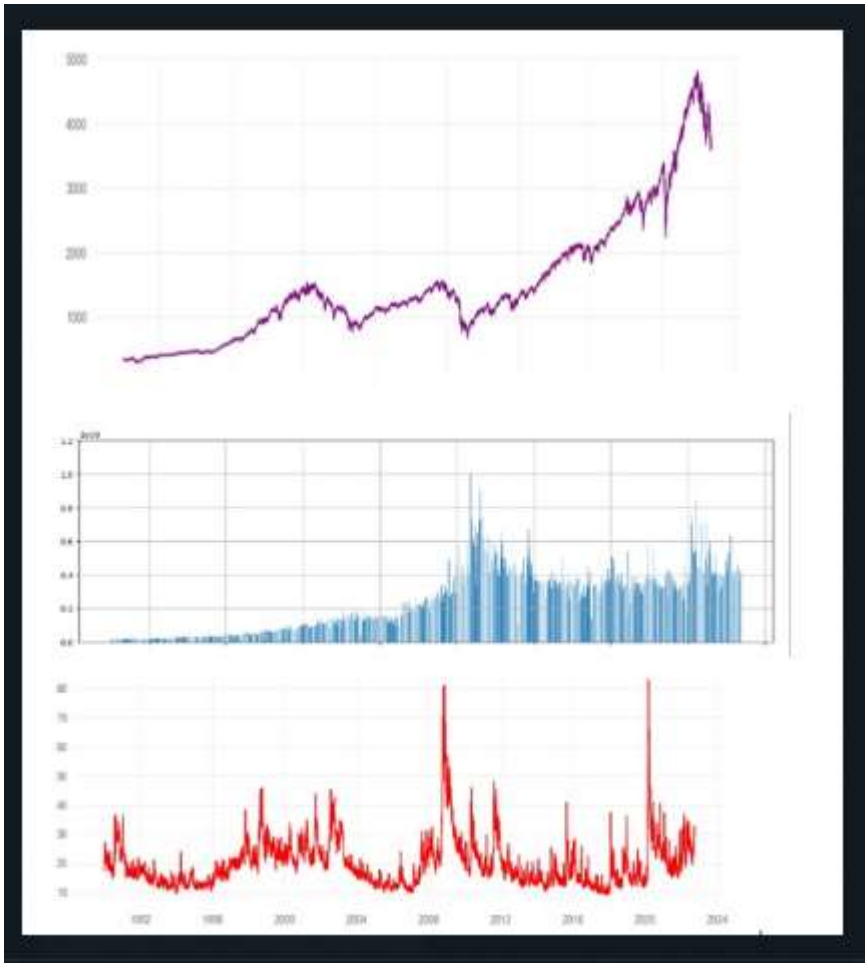


VIX vs. S&P 500 Price

Volatility value, investors' fear, and VIX values all move up when the market is falling. The reverse is true when the market advances—the index values, fear, and volatility decline. The price action of the S&P 500 and the VIX often shows inverse price action: when the S&P falls sharply, the VIX rises—and vice versa.

Comparison of the movement of S&P 500 and VIX-

This is a perfect picture to understand the behaviour of the S&P data with the VIX data. It clearly shows that whenever the volatility of the market has increased with time the prices has decreased hence validating the fact the market has tended towards getting dried up because of increasing liquidity. In the following picture, we can see that the volatility of the stock has increased drastically for 3/4 times in the last 22 years and hence getting the price to decline very swiftly.

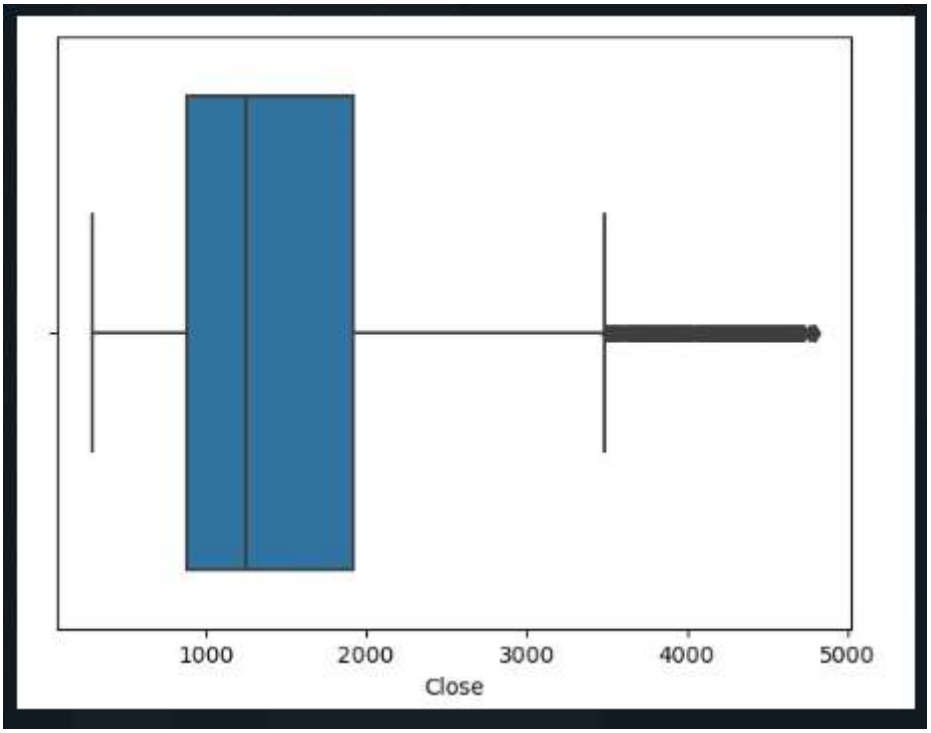


MANAGING AND DEALING WITH THE OUTLIERS

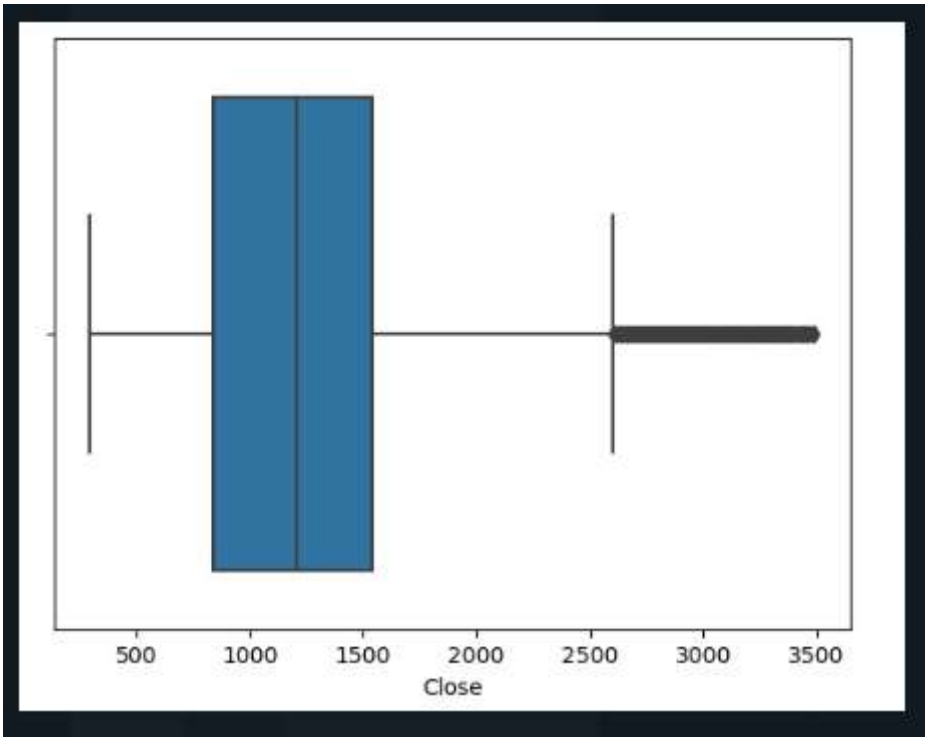
Using Inter Quartile Range and boxplot to remove the outliers of closing prices of S&P 500 index

As we already discussed in class, sometimes because of the presence of outliers in the data, we cannot get a true picture. Using the interquartile range concept, we have judged the suspected outliers to 1.5*IQR and thus plotted the boxplot to remove the outliers out of the research domain. The concept took into various parameters like largest observation, smallest observation, median and it is very important to take correct data to get the appropriate results.

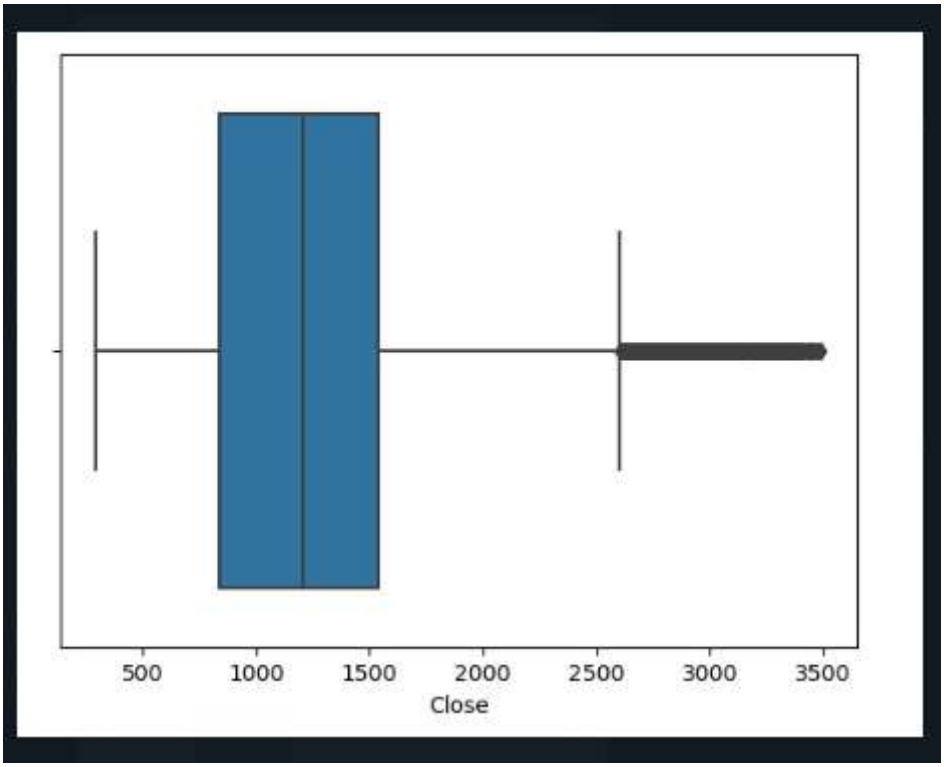
Following picture illustrates the Interquartile range of S&P and VIX data-



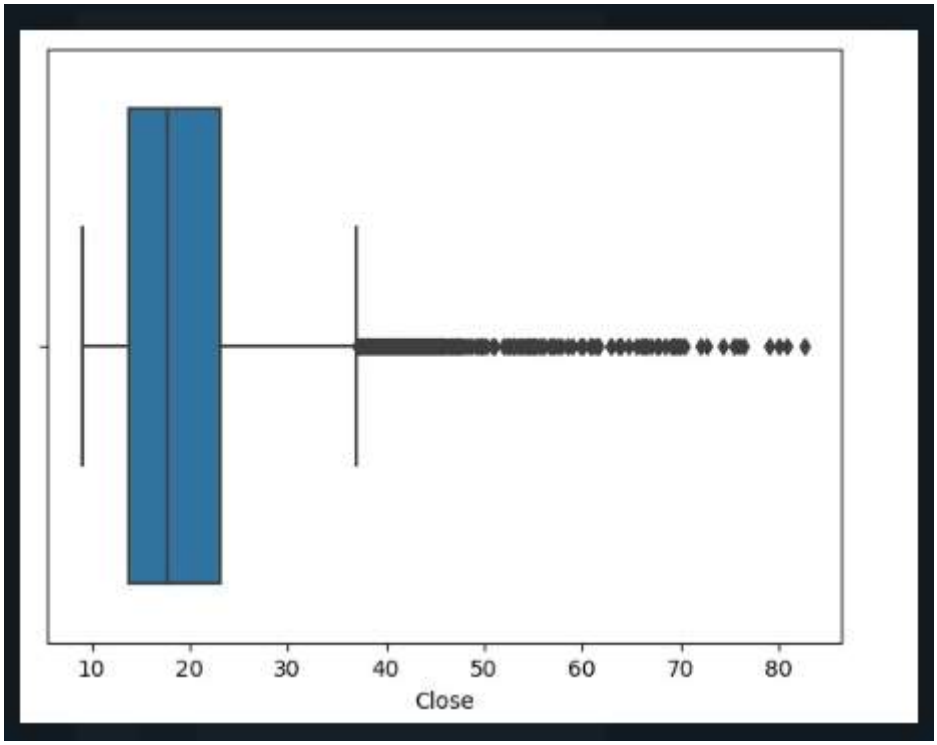
S&P DATA Closing Price before Removing the Outliers



S&P DATA Closing Price after Removing the Outliers

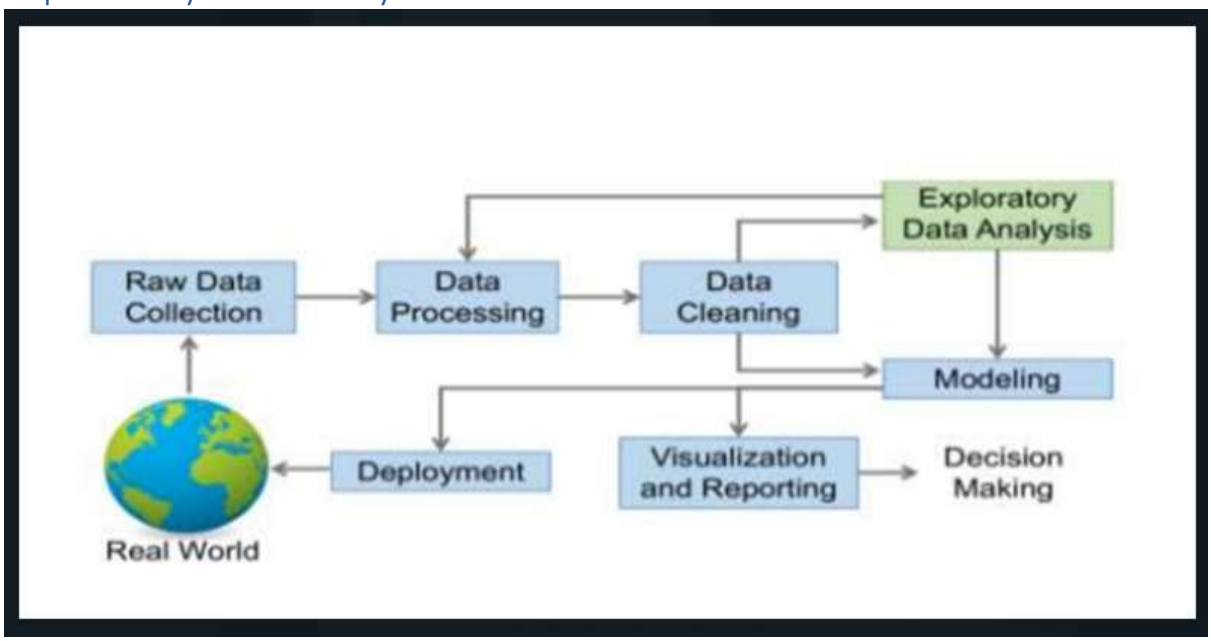


VIX DATA Closing Price before Removing the Outliers



VIX DATA Closing Price after Removing the Outliers

Exploratory Data Analysis



Data Cleaning

Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate, or non-relevant parts of the data and then restoring, remodelling, or removing the dirty or crude data. Data cleaning may be performed as batch processing through scripting or interactively with data wrangling tools. After cleaning, a dataset should be uniform with other related datasets in the operation. The discrepancies identified or eliminated may have been basically caused by user entry mistakes, corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores.

It is more important for any organization to have the right data as compared to a large data set. Data cleansing solutions can have several problems during the process of data scrubbing. The company needs to understand the various problems and figure out how to tackle them. Some of the key data cleaning problems and solutions include –

Data is never static-It is important that the data cleansing process arranges the data so that it is easily accessible to everyone who needs it. The warehouse should contain unified data and not in a scattered manner. The data warehouse must have a documented system that is helpful for the employees to easily access the data from different sources. Data cleaning also further helps to improve the data quality by removing inaccurate data as well as corrupt and duplicate entries.

Incorrect data may lead to bad decisions-While operating your business you rely on certain sources of data, based on which you make most of your business decisions. If the data has a lot of errors, the decisions you take may be incorrect and prove to be hazardous for your business. The way you collect data and how your data warehouse functions can easily have an impact on your productivity.

Incorrect data may affect district records-Complete client records are only possible when the names and addresses match. Names and addresses of the client can be poor sources of data. To avoid these mistakes, companies should provide external references which can verify the data, supplement data points, and correct any inconsistencies.

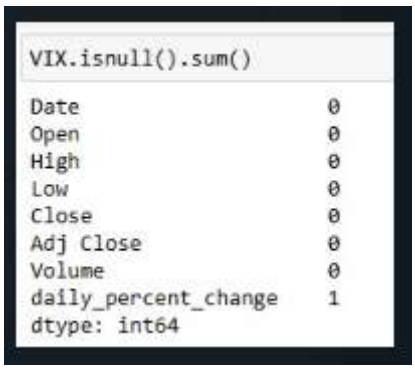
Developing a data cleaning framework in advance-Data cleansing can be a timeconsuming and expensive job for your company. Once the data is cleaned it needs to be stored in a secure location. The staff should keep a complete log of the entire process to ascertain which data went through which process. If a data scrubbing framework is not created in advance, the entire process can become repetitive.

Big data can bring in bigger problems-Big data needs regular cleansing to maintain its effectiveness. It requires complex computer data analysis of semi-structured or structured and voluminous data. Data cleansing helps in extracting information from such a big set of data and coming up with some data which can be used to make certain key business decisions.

We have cleaned the data for our further process and have removed null values to make sure that the modelling we are doing is appropriate.

```
spx.isnull().sum()

Open                  0
High                  0
Low                   0
Close                 0
Adj Close             0
Volume                0
daily_percent_change  1
log_return            1
day_of_month          0
Year                  0
dtype: int64
```

Counting the Null Values in Every Column and summing them up of S&P 500



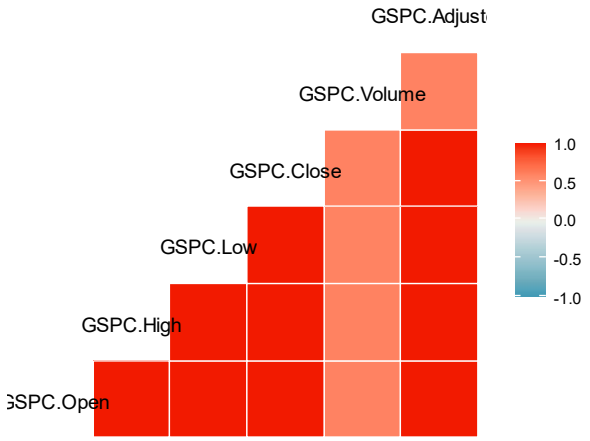
```
VIX.isnull().sum()

Date                  0
Open                  0
High                  0
Low                   0
Close                 0
Adj Close             0
Volume                0
daily_percent_change  1
dtype: int64
```

Counting the Null Values in Every Column and summing them up of VIX

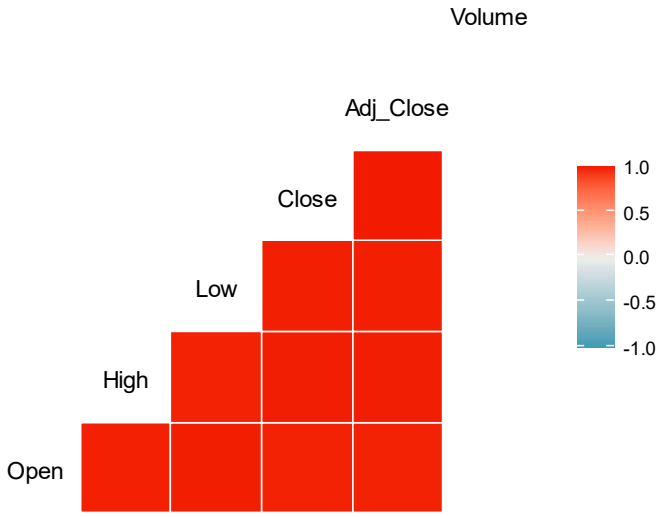
Correlation-

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

Possible values of the correlation coefficient range from -1 to +1, with -1 indicating a perfectly linear negative, i.e., inverse, correlation (sloping downward) and +1 indicating a perfectly linear positive correlation (sloping upward).



Spearman Correlation Matrix in S&P 500 data



Spearman Correlation Matrix in VIX data

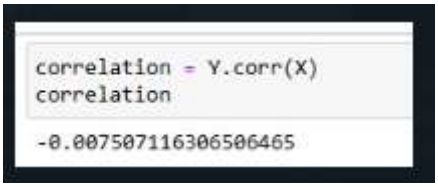
Correlation of SP500 Close and VIX Close Price

```
X = sp500['Close']
X
```

```
Date
1990-01-02     359.690002
1990-01-03     358.760010
1990-01-04     355.670013
1990-01-05     352.200012
1990-01-08     353.790009
                  ...
2022-10-04    3790.929932
2022-10-05    3783.280029
2022-10-06    3744.520020
2022-10-07    3639.659912
2022-10-10    3612.389893
Name: Close, Length: 8258, dtype: float64
```

```
Y = vix['Close']
Y
```

```
Date
1990-01-02    17.240000
1990-01-03    18.190001
1990-01-04    19.219999
1990-01-05    20.110001
1990-01-08    20.260000
                ...
2022-10-04    29.070000
2022-10-05    28.549999
2022-10-06    30.520000
2022-10-07    31.360001
2022-10-10    32.450001
Name: Close, Length: 8258, dtype: float64
```



```
correlation = Y.corr(X)
correlation
```

```
-0.007507116306506465
```

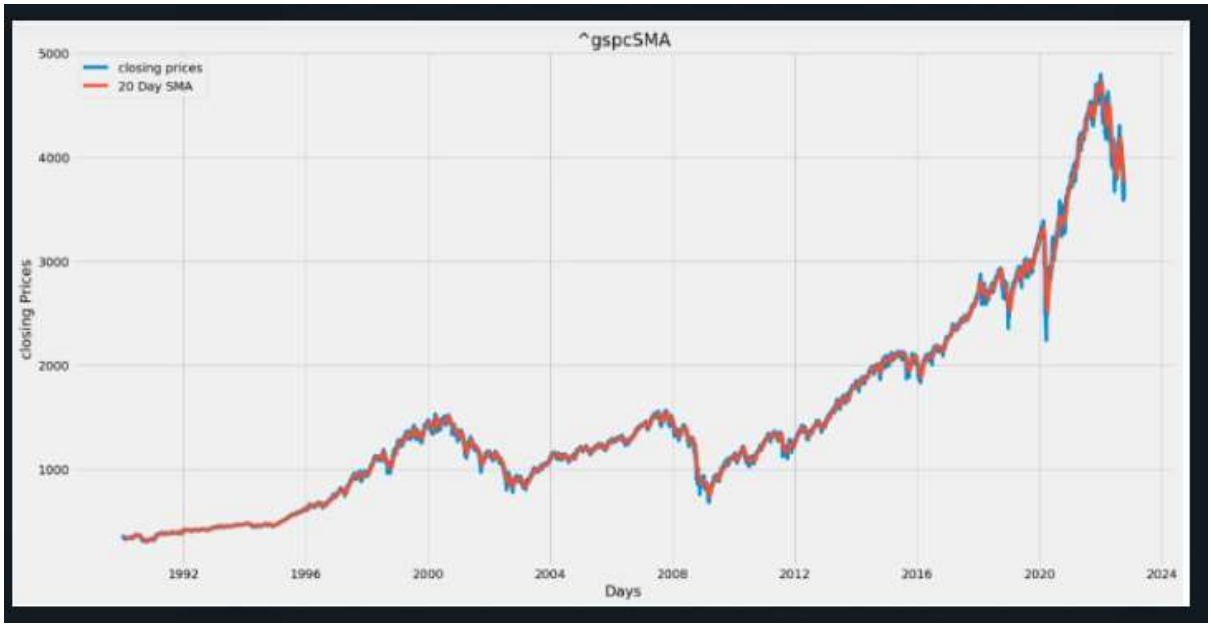
We can clearly see that there is a negative correlation between S&P 500 index and VIX which is correct as when there is a higher volatility in the market, people start pulling out the money and share prices decreases and when the VIX decreases the correction starts in the market.

Applying Bollinger Bands

Before jumping on to explore Bollinger Bands, it is essential to know Simple Moving Average (SMA). Simple Moving Average is nothing, but the average price of a stock given a specified period. Now, Bollinger Bands are trend lines plotted above and below the SMA of the given stock at a specific standard deviation level

Bollinger Bands are great to observe the volatility of a given stock over a period. The volatility of a stock is observed to be lower when the space or distance between the upper and lower band is less. Similarly, when the space or distance between the upper and lower band is more, the stock has a higher level of volatility. While observing the chart, you can observe a trend line named 'MIDDLE BB 20' which is nothing but SMA 20 of the Tesla stock. The formula to calculate both upper and lowers bands of stock are as follows:

- UPPER_BB = STOCK SMA + SMA STANDARD DEVIATION * 2.5
- LOWER_BB = STOCK SMA - SMA STANDARD DEVIATION * 2.5

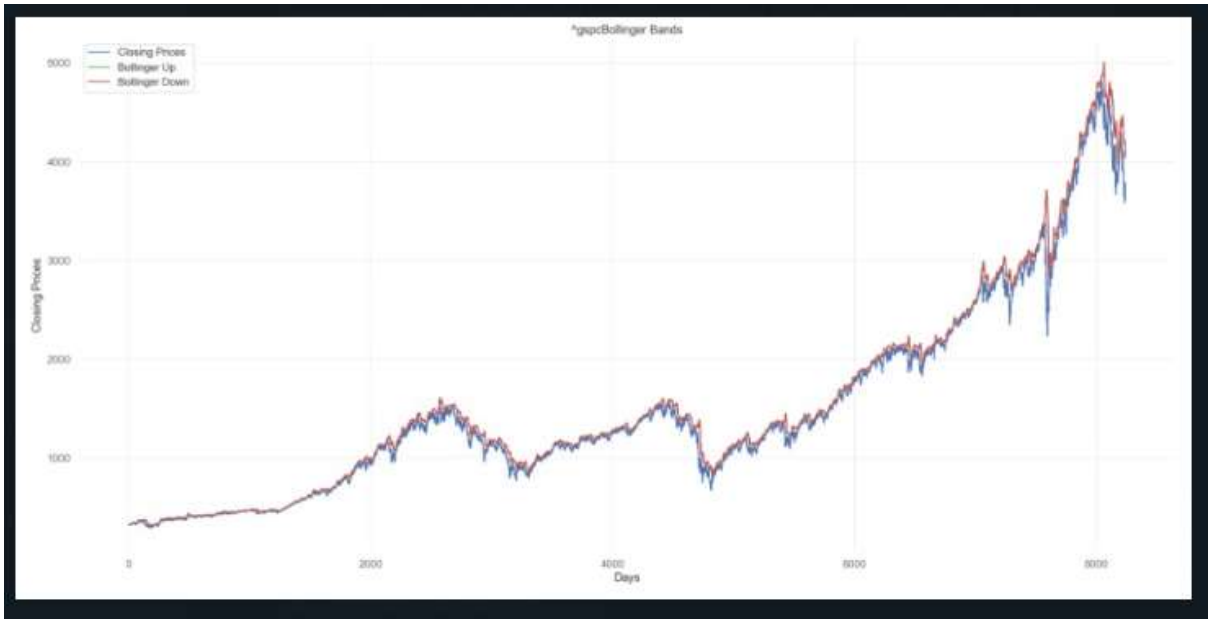


Bollinger Bands adapt dynamically to price expanding and contracting as volatility increases and decreases. Therefore, the bands naturally widen and narrow in sync with price action, creating a very accurate trending envelope.

Benefits of Bollinger bands:

- Bollinger Bands® help you identify sharp, short-term price movements and potential entry and exit points
- Bollinger Bands are visually easy to interpret
- They can be used both as a volatility indicator and a momentum oscillator
- Bollinger Bands can be applied on any underlying asset across any time frame
- The indicator generates signals that not only provide precise entry levels, but also specify stop loss and take-profit zones

APPLICATION OF BOLLINGER BAND



STRATEGY OF BOLLINGER BANDS

- Daily Time Frame

- 20MA Knowing the Trend
- BB length = 20, std = 2.5

"IF PREV_STOCK > PREV_LOWERBB & CUR_STOCK < CUR_LOWER_BB => BUY IF PREV_STOCK < PREV_UPPERBB & CUR_STOCK > CUR_UPPER_BB => SELL"

BACK TESTING

- Back testing involves applying a strategy or predictive model to historical data to determine its accuracy. It can be used to test and compare the viability of trading strategies so traders can employ and tweak successful strategies.
- It allows traders to test trading strategies without the need to risk capital.
- Common back testing measures include net profit/loss, return, risk-adjusted return, market exposure, and volatility.

How Back testing Works

- Analysts use back testing to test and compare various trading techniques without risking money. The theory is that if their strategy performed poorly in the past, it is unlikely to perform well in the future (and vice versa).
- The two main components looked at during testing are the overall profitability and the risk level taken.
- However, a back test will look at the performance of a strategy relative to many different factors. A successful back test will show traders a strategy that's proven to show positive results historically. While the market never moves the same, back testing relies on the assumption that stocks move in similar patterns as they did historically.

Implementation

- A back test is usually coded by a programmer running a simulation on the trading strategy. The simulation is run using historical data from stocks, bonds, and other financial instruments. The person facilitating the back test will assess the returns on the model across several different datasets.
- It is also essential that the model is tested across many different market conditions to assess performance objectively. Variables within the model are then tweaked for optimization against several different back testing measures.

Common Back testing Measures

- Net Profit/Loss
- Return: The total return of the portfolio over a given time frame
- Risk-Adjusted Return: The return of the portfolio adjusted for a level of risk
- Market Exposure: the degree of exposure to different segments of the market
- Volatility: The dispersion of returns on the portfolio

Benefits of Back testing-

- Do you know most of the traders in the market lose money?
- They lose money not because they lack understanding of the market. But simply because their trading decisions are not based on sound research and tested trading methods.
- They make decisions based on emotions, suggestions from friends and take excessive risks in the hope to get rich quickly. If they remove emotions and instincts from the trading and back test the ideas before trading, then the chance to trade profitability in the market is increased.

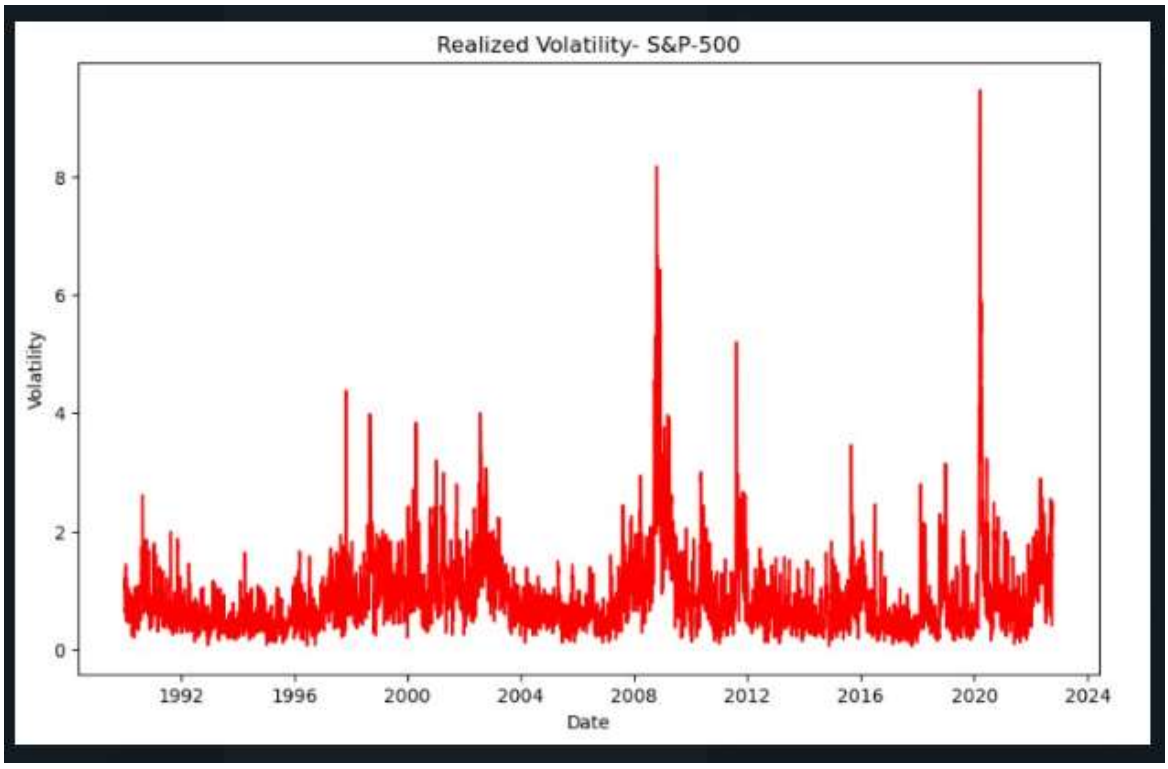
Result of Back Testing-

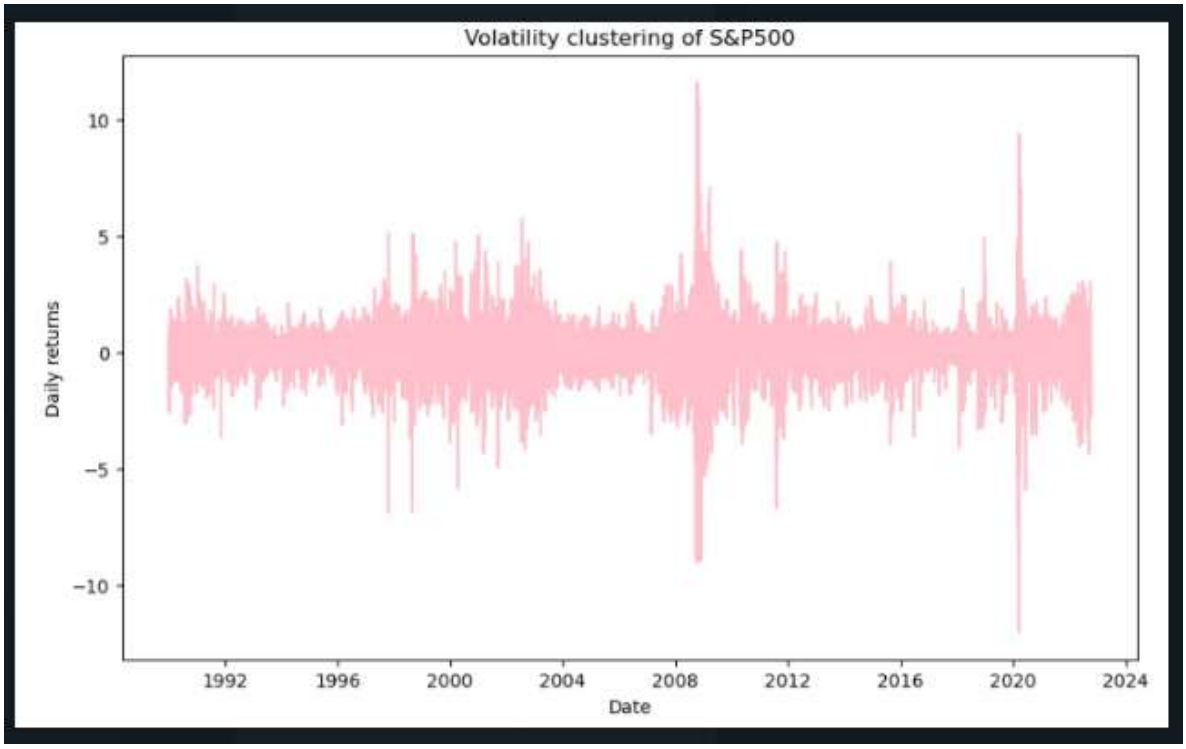
Link to the tear sheet-

https://drive.google.com/file/d/1iete-PKCP90kL7KK76TBJdQc-69T2VIg/view?usp=share_link

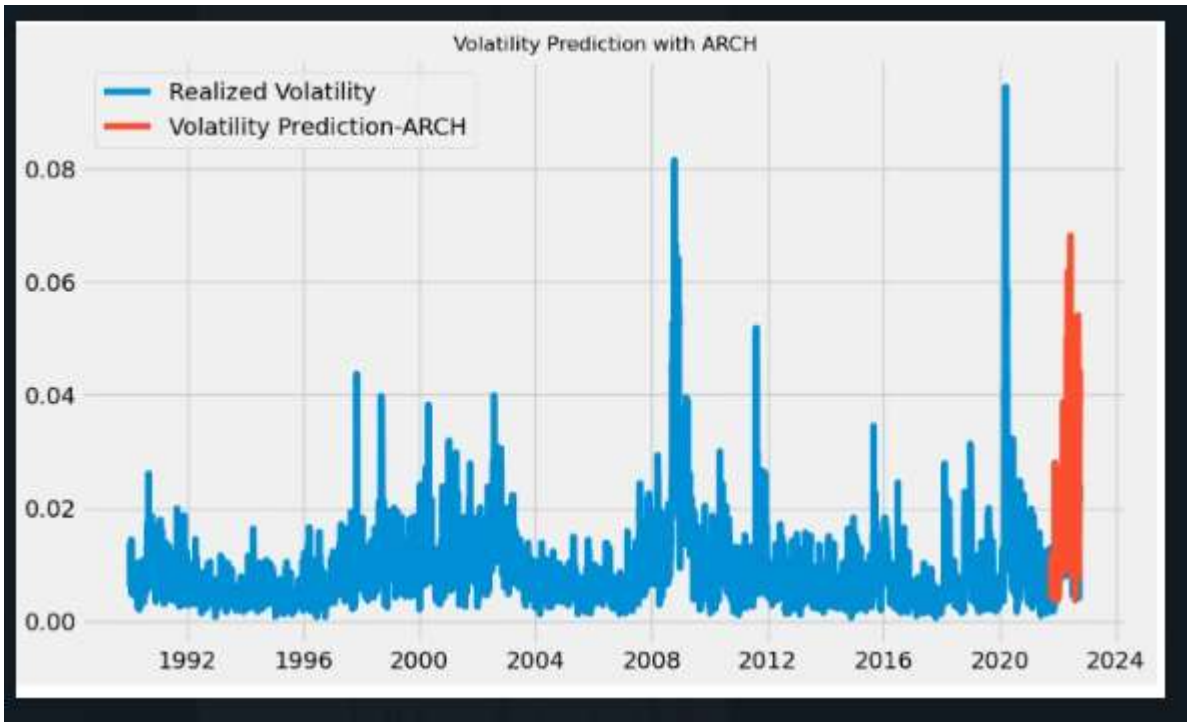
ARCH Model-

An ARCH (autoregressive conditionally heteroscedastic) model is a model for the variance of a time series. ARCH models are used to describe a changing, possibly volatile variance. Although an ARCH model could possibly be used to describe a gradually increasing variance over time, most often it is used in situations in which there may be short periods of increased variation. (Gradually increasing variance connected to a gradually increasing mean level might be better handled by transforming the variable.)





PREDICTION OF VOLATALITY USING ARCH MODEL-



Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them. We have used simple linear regression in our project and the same is discussed below.

Simple linear regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$Y = a + bx + \epsilon$

Where:

Y – Dependent variable
X – Independent (explanatory) variable
a – Intercept
b – Slope
ϵ – Residual (error)

Linear Model Assumptions

- Linear regression analysis is based on six fundamental assumptions:
- The dependent and independent variables show a linear relationship between the slope and the intercept.
- The independent variable is not random.
- The value of the residual (error) is zero.
- The value of the residual (error) is constant across all observations.
- The value of the residual (error) is not correlated across all observations.
- The residual (error) values follow the normal distribution.

Benefits of Linear regression

Linear regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You'll find that linear regression is used in everything from biological, behavioural, environmental, and social sciences to business. Linear regression models have become a proven way to predict the future scientifically and reliably. Because linear regression is a long-established statistical procedure, the properties of linear regression models are well understood and can be trained very quickly.

Results of Linear Regression

```
model <- lm(SPX$GSPC.Close ~ SPX$GSPC.Volume, data = SPX)
summary(model)

##
## Call:
## lm(formula = SPX$GSPC.Close ~ SPX$GSPC.Volume, data = SPX)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3496.1  -366.8  -121.8   286.0  3200.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.159e+02  1.408e+01   50.85   <2e-16 ***
## SPX$GSPC.Volume 3.212e-07  4.647e-09   69.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 781.2 on 8256 degrees of freedom
## Multiple R-squared:  0.3665, Adjusted R-squared:  0.3664
## F-statistic:  4776 on 1 and 8256 DF,  p-value: < 2.2e-16
```

Calculation of MSE and RMSE in our Model

```
summary(model)$r.squared

## [1] 0.3664659

#calculate MSE
mean(model$residuals^2)

## [1] 610200

#calculate RSME
sqrt(mean(model$residuals^2))

## [1] 781.153
```

ANOVA TESTING RESULTS-

```
                   Df    Sum Sq   Mean Sq F value Pr(>F)
SPX$GSPC.Volume     1 5.695e+09 5.695e+09   11203 <2e-16 ***
Residuals       10784 5.482e+09 5.083e+05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```