Faculty of Computer& Informatics.
Department of Artificial Intelligence &
Data science
Department of Software Engineering

Analyzing Stock Market Data using Big Data System Techniques

Prepared by

Abdullah Khadem Aljame

Ali Hussain Alaswad

Supervised by

Dr.Eng.Mouhib Alnoukari

Eng.Anas Abdulaziz

Academic Year

2023-2024

Abstract

The stock market plays a pivotal role in the global economy, serving as a platform for buying and selling of financial securities, such as stocks, bonds, and commodities. An essential aspect of stock market operations involves the analysis of vast amounts of financial data to gain insights, make informed decisions, and predict market trends. This abstract discusses the significance of data analysis in the stock market, focusing on the Nifty 100 index as an overview and the benefits of employing big data systems for enhanced analysis.

The Benefits of Analyzing Stock Market Data:

Analyzing stock market data confers several advantages. Firstly, it enables investors and financial institutions to identify patterns, trends, and correlations within the market, assisting in making informed investment decisions. Additionally, data analysis aids in risk assessment and management, allowing for the mitigation of potential financial losses. Moreover, understanding market movements through data analysis is crucial for developing effective trading strategies and optimizing portfolio performance.

Overview of Nifty 100:

The Nifty 100 index represents a diversified portfolio of top 100 companies listed on the National Stock Exchange (NSE) of India. This index provides a comprehensive view of the Indian equity market, encompassing companies from various sectors, and is widely used by investors and fund managers as a benchmark for evaluating market performance and constructing investment portfolios.

The Benefits of Using Big Data Systems in Stock Market Analysis:

Employing big data systems in stock market analysis offers numerous advantages. Firstly, big data technologies facilitate the processing and analysis of vast volumes of financial data in real time, enabling rapid decision-making and the extraction of valuable insights. Furthermore, big data systems can uncover complex market patterns and anomalies that may not be discernible through traditional analytical approaches. Additionally, the integration of big data tools with machine learning and predictive analytics empowers market participants to develop sophisticated models for forecasting stock price movements and market behavior.

In conclusion, the stock market is intrinsically linked to the analysis of financial data, and leveraging big data systems further enhances the depth and precision of market analysis. The Nifty 100 index serves as a key barometer of the Indian equity market, while the utilization of big data technologies offers substantial potential for refining investment strategies, managing risk, and gaining a competitive edge in the dynamic landscape of the stock market.

Content list

Chapter 1: Project Summary

1.1. Project Summary:

1.1.1. Overview:

Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity at which it is created and collected, and the variety of the data points being covered.

1.1.2. Aim:

In this project, we are supposed to build a working environment for a big data system and perform data analysis using big data technology.

1.2. Project Goals:

- Build A Real Big Data System In SPU
- Analyzing Data using Big Data Techniques
- Make the environment ready to receive future big data projects
- Work on implementing financial analysis for financial companies

Chapter 2: Weekly Project Plan

Week 1 & 2: Preparing servers to start working

- Data center and Information about it
- Servers types and about the specification
- Chose ubuntu server as the operating system for Hadoop and config it.
  - steps:
    - install ubuntu server at each server
    - configure Server by passing the IP Address,

    Faced Some problem configuration:

    FIRST

    - The operating system did not recognize the server's network card

    SECOND

    - the SSH problem: the way to identify the server with the college laboratories and halls so that we communicate with the servers from the halls

- How to solve all these problems?

- Changing the operating system to Ubuntu desktop22.04

- configure the IP Address for each server

  Master server: spu@10.0.1.219

  Slave server: spu@10.0.1.220

-Now we can access the server from all halls and laboratories using ssh by using " ssh spu@10.0.1.219/220 "

Week 3 & 4: Hadoop environment &configuration files

1-Install jdk 8 for Hadoop nodes (Master and Slave)

2-Install Apache Hadoop 3.3.6 for each server and config it by edit the xml files

Steps:

- install java 8 on each server

-download hadoop 3.3.6 https://www-eu.apache.org/dist/hadoop... and extract

-copy hadoop-3.3.6.tar.gz folder to each slave server with scp and extract

-add hostname in /etc/hosts/

-move folder: hadoop-3.3.6 to /usr/local/hadoop/

-add PATH /etc/environment

JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

-create user: hadoop

-create ssh keygen from master server and copy to slave server

-Configuring the Hadoop Master

-Copy the configuration files to each of your Hadoop Nodes from your Hadoop Master.

-Format the HDFS file system on master server

-Now you can start HDFS

- open master server: http://192.168.56.101:9870

-config Yarn

Week 5 & 6: Find and extract data

Gather, Find and understand the data:

https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data

About data:

Overview:

This dataset contains historical daily prices for Nifty 100 stocks and indices currently trading on the Indian Stock Market.

About Nifty 50:

The NIFTY 50 index is a well-diversified 50 companies index reflecting overall market conditions. NIFTY 50 Index is computed using the free float market capitalization method.

The Nifty 100 index tracks the performance of the top 100 companies listed on the National Stock Exchange (NSE) of India. This broader index includes constituents of the Nifty 50 as well as an additional 50 companies, providing a more comprehensive view of the Indian stock market. The Nifty 100 index offers exposure to a diversified set of stocks across various sectors, making it a valuable benchmark for investors and fund managers. It serves as a barometer of the Indian equity market, allowing for a broader assessment of market performance and trends. The inclusion of additional companies in the Nifty 100 offers investors a more extensive representation of the Indian stock market compared to the Nifty 50, allowing for a broader perspective on market dynamics and diversification opportunities.

Data samples are of 1-minute intervals and the availability of data is from Jan 2015 to Feb 2022.

Content:

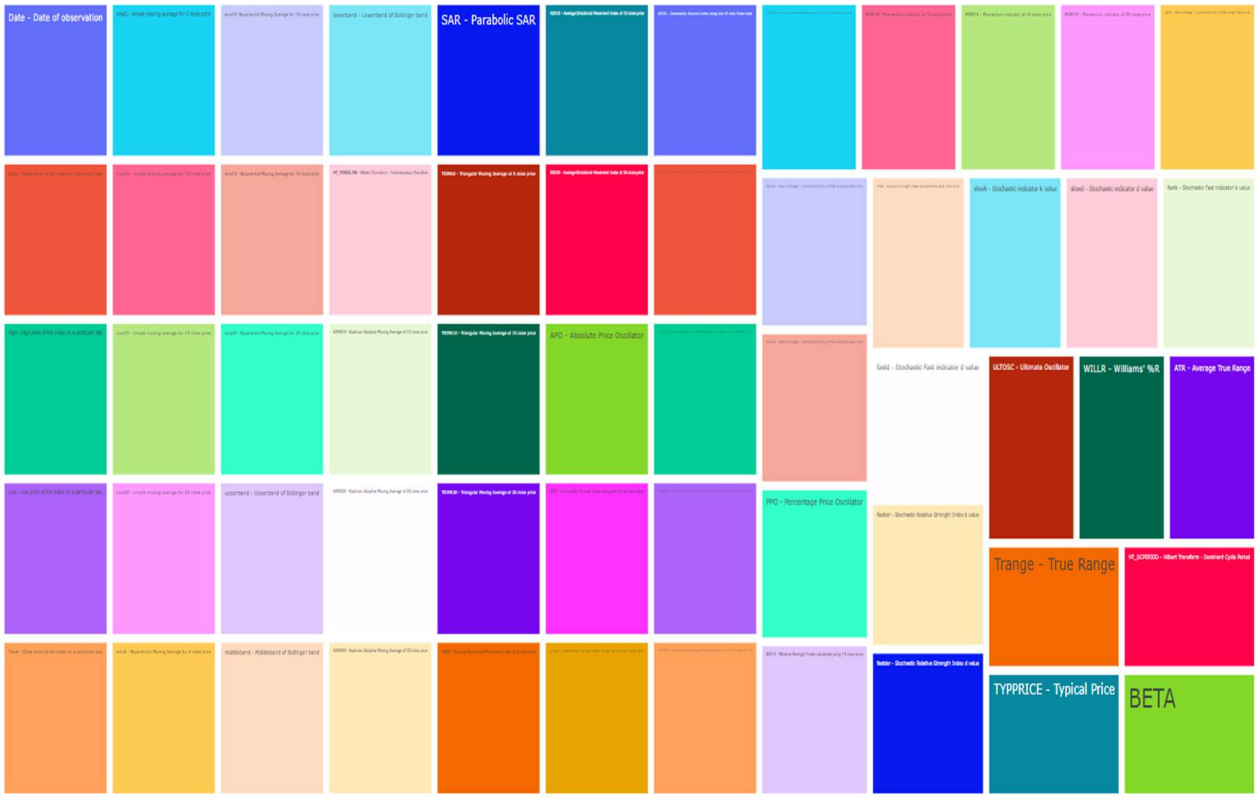
The whole dataset is around 66.2 GB (compressed here to 26 GB), and 100 stocks (Nifty 100 stocks) and 2 indices (Nifty 50 and Nifty Bank indices) are present in this dataset.

Details about each file:

OHLCV (Open, High, Low, Close, and Volume) data

55 Technical indicator values

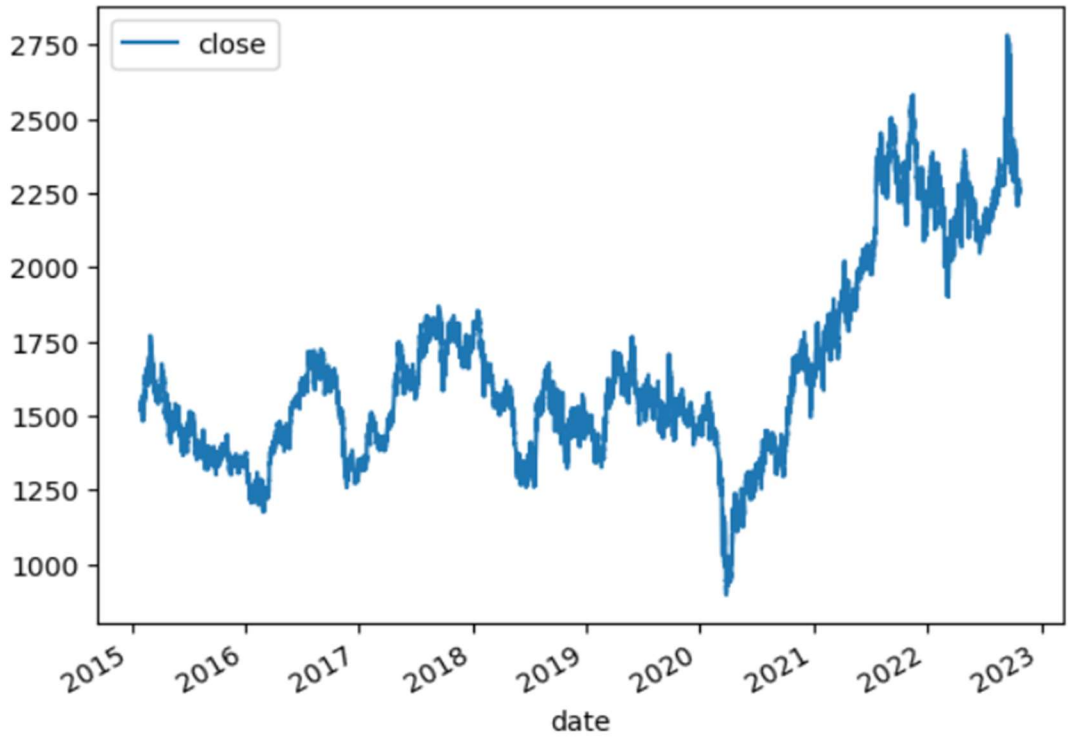
The columns contain different technical indicators related to financial market data. Below is a brief explanation of each:



Week 7 & 8: Descriptive and Predictive Analytics

Simple visualize:

```
In [4]: da=df[0:1000]
        import plotly.graph_objects as go
        fig = go.Figure(data=go.Candlestick(x=da.index,open=da['open'],high=da['high'],low=da['low'],close=da['close']))
        fig.show()
```



Descriptive Statistics:

Descriptive statistics are numbers that describe a larger set of numbers, which quickly allow people to understand some important features about the data.

Understanding descriptive statistics

Descriptive statistics are numbers that convey information about a larger or more complex set of numbers (also called a dataset). In business, descriptive statistics are the outputs from descriptive analytics. They allow someone to gain insight about the data without diving into the details and unique circumstances that each data point represents. In statistics, descriptive statistics are the standard summary statistics of a dataset when conducting a research project. It includes some information about how dispersed (spread out) the data is, what the average values look like, and how unbalanced the observations are. Descriptive statistics are bite-sized pieces of information that provide general insight about the larger dataset.

What is the main purpose of descriptive statistics?

The primary purpose of descriptive statistics is to convey information quickly. In business, the person receiving the information may not have the time or skills required to analyze data. That is one reason why data analysts take complex information and reduce it into something more digestible.

A management team can take the descriptive statistics into account as they consider changes in a company's strategy. These descriptive statistics allow managers to understand if the current plan is working and if course corrections are required.

Investors often use descriptive statistics to get a feel for a company's finances, performance, value, and growth potential. Researchers may use descriptive statistics to understand and communicate the details of the set of data they are using.

What are the types of descriptive statistics?

Basic Descriptive Statistics:

Calculate the mean, median, mode, minimum, and maximum values for each column. This will give you a general overview of the central tendency and range of values.

Variability:

Compute the standard deviation and variance for columns like Open, Close, High, Low, and the moving averages (sma5, sma10, etc.). This helps you understand the dispersion or volatility in the data.

Trend Indicators:

Analyze the trend indicators (sma and ema columns) to identify trends in the stock prices over different time frames. You can calculate the rate of change (ROC) for these indicators to measure the percentage change in the moving averages.

Bollinger Bands:

Explore the upperband, middleband, and lowerband columns to understand the volatility and potential reversal points in the stock prices.

Momentum Indicators:

Investigate momentum indicators (MOM and ROC columns) to identify periods of strong price movement and potential trend reversals.

Relative Strength Indicators:

Examine the RSI columns (RSI14 and RSI8) to identify overbought or oversold conditions in the market.

Stochastic Indicators:

Explore the stochastic indicators (slowk, slowd, fastk, fastd, fastksr, and fastdsr) to identify potential turning points in the market.

MACD (Moving Average Convergence/Divergence):

Analyze the MACD columns (macd510, macd520, etc.) to identify potential crossovers and divergences, which can be indicative of trend changes.

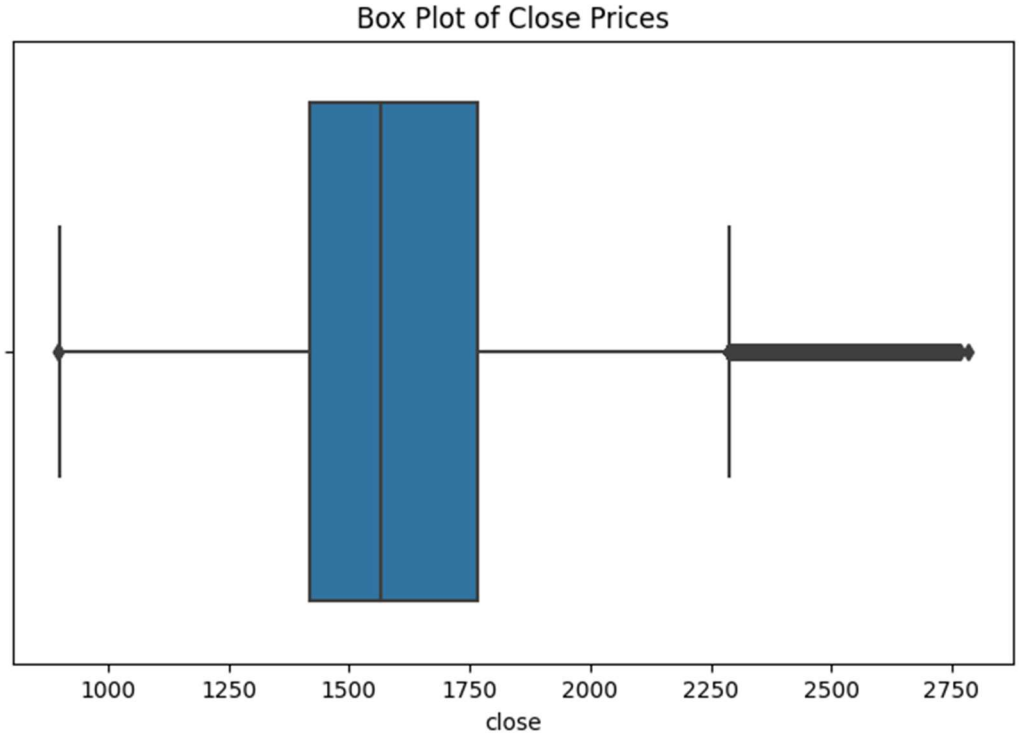
Other Technical Indicators:

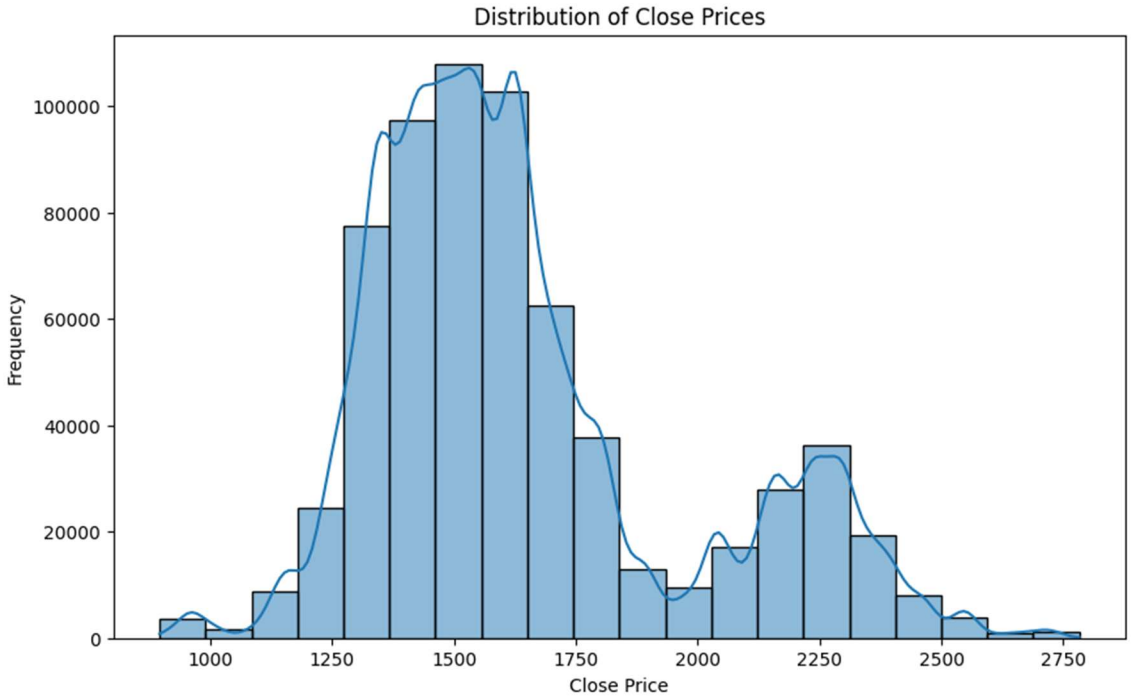
Examine the values of ATR, Trange, TYPPRICE, HT_DCPERIOD, BETA, and other columns to gain insights into market volatility, true range, typical prices, dominant cycle periods, and beta coefficients.

Correlation Analysis:

Conduct correlation analysis to understand the relationships between different columns. For example, you can explore how moving averages correlate with price trends or how momentum indicators relate to price changes.

Basic Descriptive Statistics: Calculate the mean, median, mode, minimum, and maximum values for each column. This will give you a general overview of the central tendency and range of values.





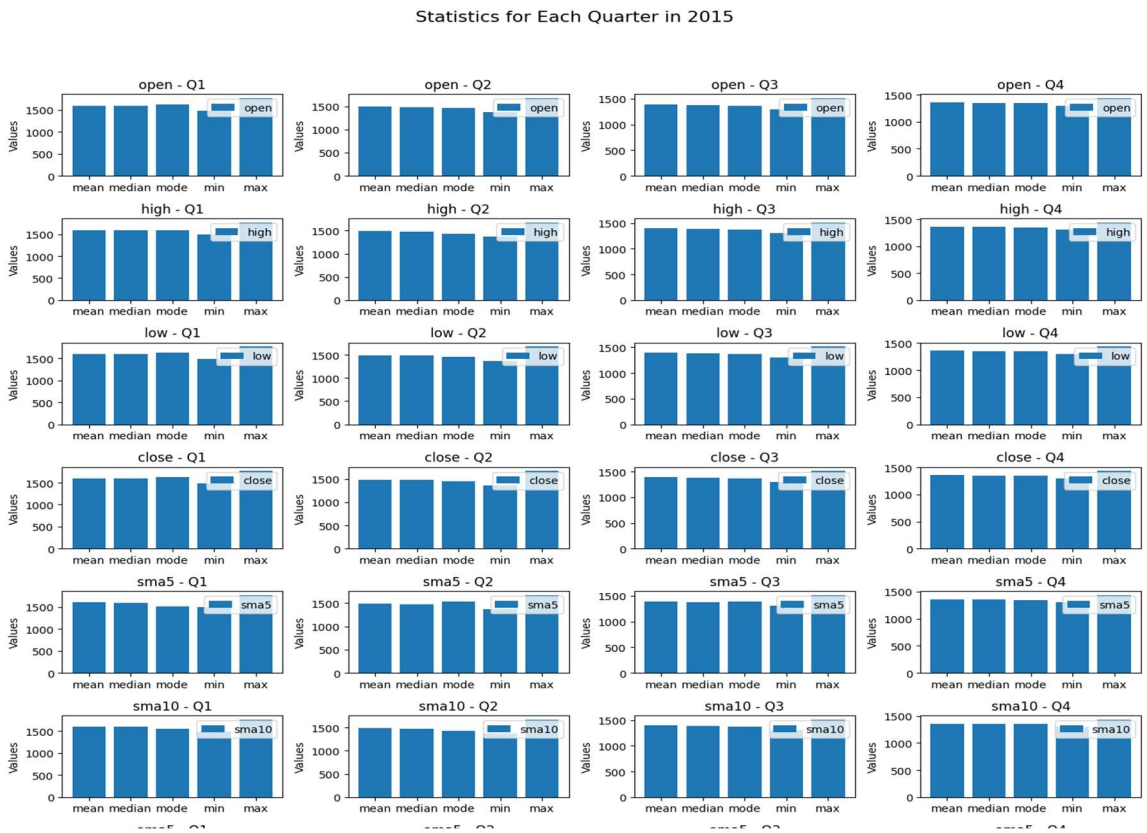
The basics Descriptive statistics for close price and Volume:

Year	Close Mean	Close Median	Close Min	Close Max	Close Std	Volume Mean	Volume Median	Volume Min	Volume Max	Volume Std
2015	1447.3036461974552	1416.7	1301.1	1770.0	100.22217765200737	804.4324338066124	352.0	0.0	825479.0	3668.380516361473
2016	1459.6305087843416	1453.0	1175.9	1724.8	157.5215953475246	725.2871359509492	316.0	0.0	210112.0	2145.5224628037136
2017	1622.8521777084056	1659.25	1314.6	1869.65	153.06642862332475	1055.7428141981732	449.0	0.0	500892.0	3661.282795594651
2018	1514.281260082452	1526.55	1260.9	1854.1	134.4344369283616	1445.7640795841548	636.0	0.0	930263.0	5972.3665454859965
2019	1530.8125605887278	1530.35	1326.7	1766.25	85.08687735471864	2108.802094525186	1083.0	0.0	324126.0	4588.097841293055
2020	1389.9351965125304	1403.0	896.5	1782.5	191.65594827392633	3484.402656685078	1851.5	0.0	586897.0	7400.378674141749
2021	2080.4188201513966	2045.3	1585.85	2581.45	255.6532407983641	1963.7455447189197	850.0	0.0	260083.0	4352.169306647161
2022	2230.5591151993617	2218.95	1901.1	2782.65	132.6511165454724	1768.718770484795	685.0	0.0	962709.0	5286.303846382271

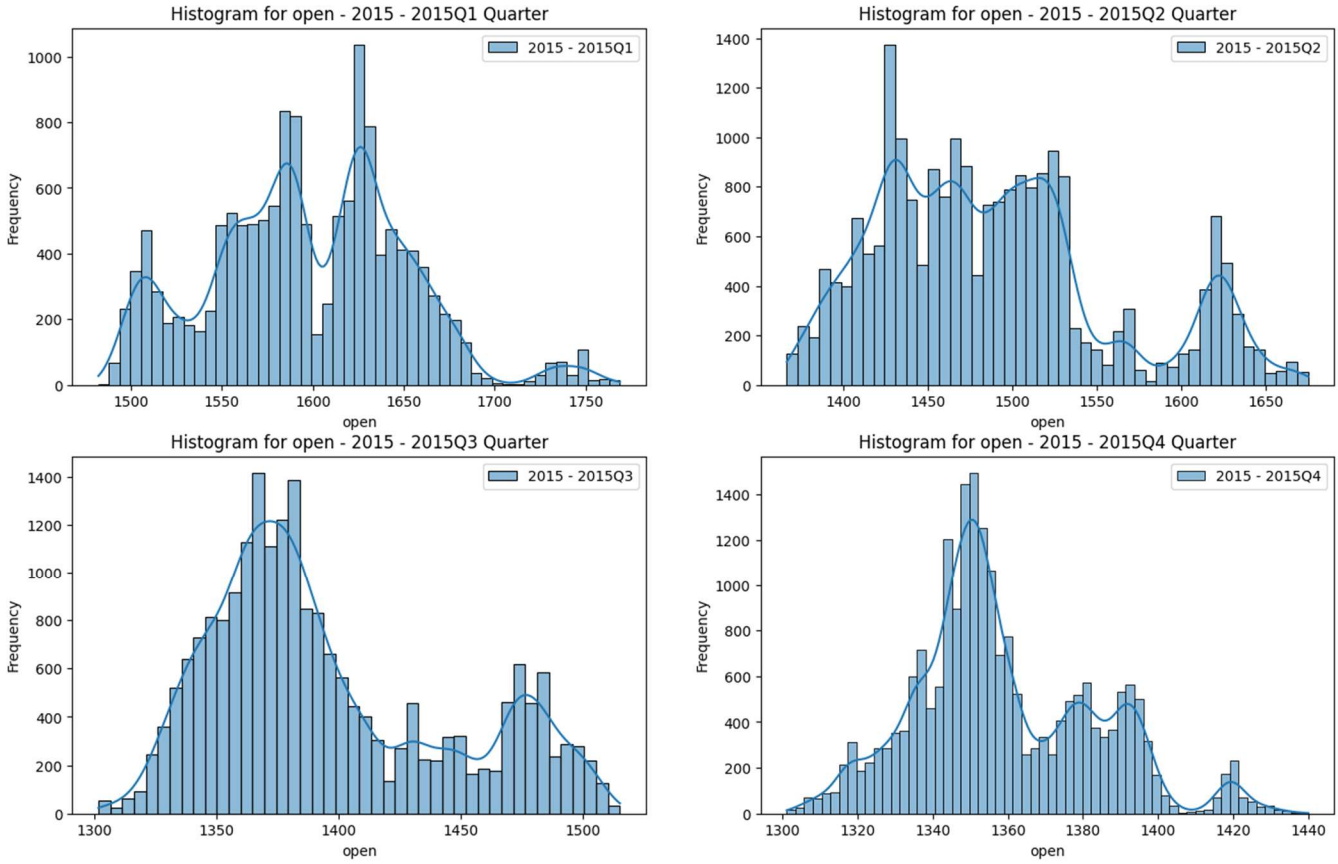
The best selling for each Quarter from 2015 to 2022:

Year	Best-Selling Quarter	volume
2015	2015Q2	18069994
2016	2016Q3	16687525
2017	2017Q3	29382788
2018	2018Q4	39534071
2019	2019Q3	51242118
2020	2020Q4	104609456
2021	2021Q1	61945983
2022	2022Q3	52985210

Basics Statistics for 2015:

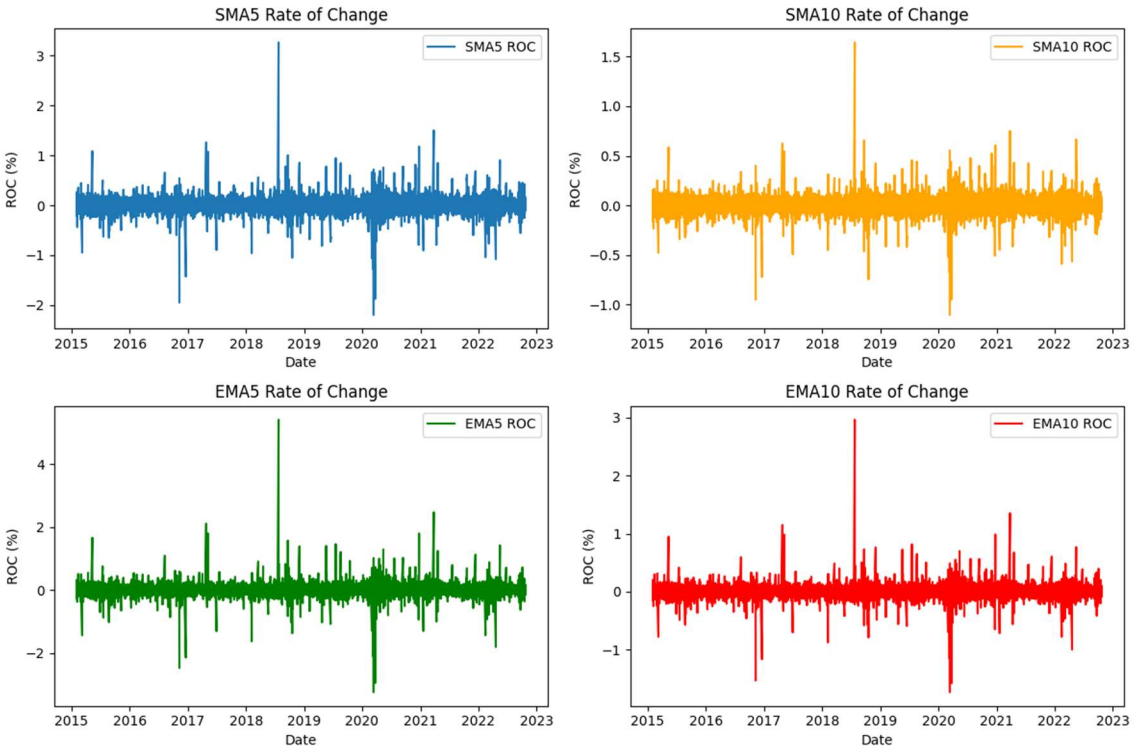


Distribution for each quarter in 2015:



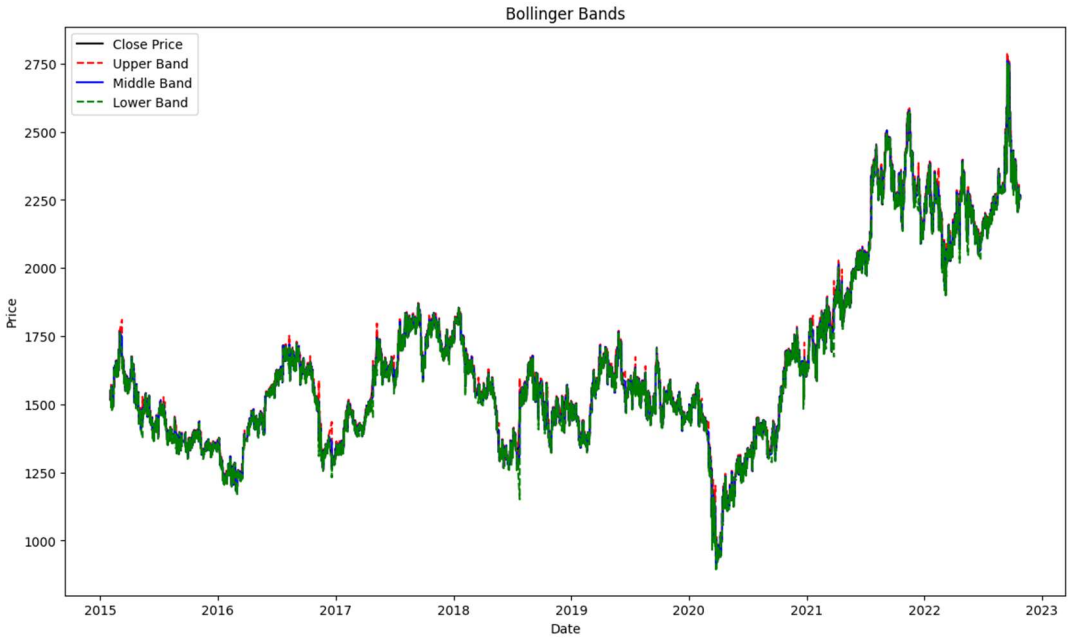
2- **Trend Indicators:**

Analyze the trend indicators (sma and ema columns) to identify trends in the stock prices over different time frames. You can calculate the rate of change (ROC) for these indicators to measure the percentage change in the moving averages.



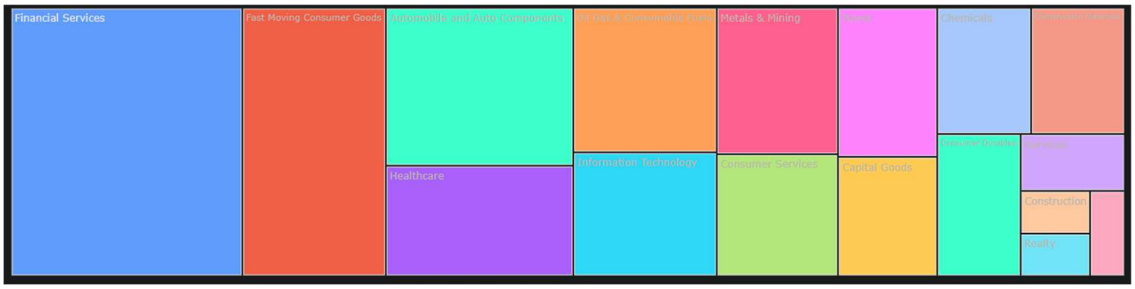
3- **Bollinger Bands:**

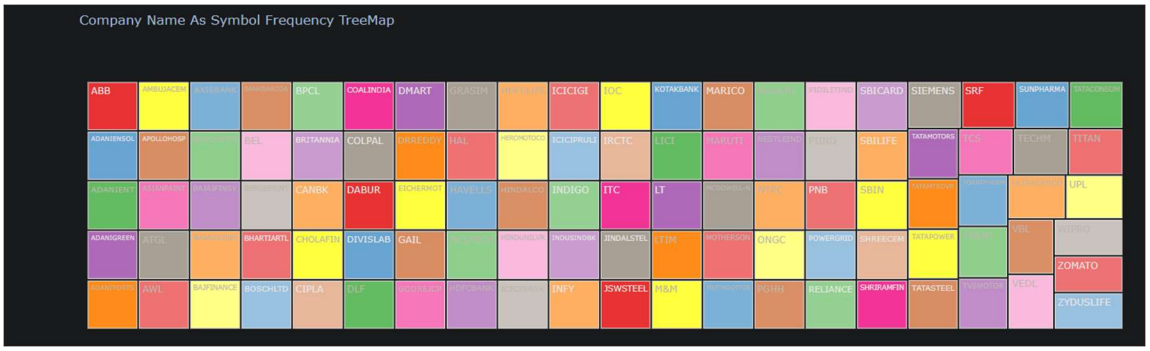
Explore the upperband, middleband, and lowerband columns to understand the volatility and potential reversal points in the stock prices



3-TreeMap:

create a TreeMap to visualize the frequency of industries in your dataset, you can use Plotly Express in Python:





4- **Variability:**

Compute the standard deviation and variance for columns like Open, Close, High, Low, and the moving averages (sma5, sma10, etc.). This helps you understand the dispersion or volatility in the data.