A PRELIMINARY MINI PROJECT REPORT ON

# "Stock Price Analysis and Prediction Using Machine Learning Techniques"

SUBMITTED TOWARDS THE PARTIAL FULFILLMENT OF THE REQUIREMENTS OF BACHELOR OF ENGINEERING (S.Y. B. Tech.)

Academic Year: 2024-25

By:

| Name | PRN |
|---|---|
| Dhanashree Sul | 123B1B273 |
| Nilesh Sonawane | 123B1B270 |
| Yash Sonje | 123B1B271 |
| Smera Nimje | 123B1B269 |

Under the Guidance of

<Dr. Chetan Chauhan>
<Mrs. Harsha Talale>

DEPARTMENT OF COMPUTER ENGINEERING,
PIMPRI CHINCHWAD COLLEGE OF
ENGINEERING SECTOR 26, NIGDI,
PRADHIKARAN

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER ENGINEERING

## CERTIFICATE

**This is to certify that the project entitled**

**"Stock Price Analysis and Prediction Using Machine Learning Techniques"**

**is successfully carried out as a Skill Development Laboratory I mini project and successfully submitted by the following students of "PCET's Pimpri Chinchwad College of Engineering, Nigdi, Pune-44."**

**Under the guidance of Dr.Chetan Chauhan & Mrs. Harsha Talale In partial fulfillment of the requirements for the T.Y. B. Tech.**

**(Computer Engineering)**

| | |
|---|---|
| DHANASHREE SUL | 123B1B273 |
| Nilesh Sonawane | 123B1B270 |
| Yash Sonje | 123B1B271 |
| Smera Nimje | 123B1B269 |

**Dr. Chetan Chauhan**
**Project Guide**

## INDEX

| Sr. No. | Chapter Name | Page No. |
|---|---|---|
| 1 | **Introduction**<br>1.1 Background<br>1.2 Problem Statement<br>1.3 Objective<br>1.4 Motivation | 5 |
| 2 | **Data Collection**<br>2.1 Data collection<br>2.2 Primary vs Secondary data | 14 |
| 3 | **Exploratory Data Analysis**<br>3.1 Data preprocessing<br>3.2 Data visualization<br>3.3 Feature Selection | 17 |
| 4 | **Methodology**<br>4.1 Workflow | 22 |
| 5 | **Model Evaluation**<br>5.1 Evaluation Metrics<br>5.2 Performance of Linear regression model<br>5.3 Performance of Logistic regression model<br>5.4 Conclusion of Model evaluation | 25 |
| 6 | **Result and Analysis**<br>6.1 Results of Linear regression model<br>6.2 Results of Logistic regression model<br>6.3 Discussion of Logistic regression results<br>6.4 Summary of results<br>6.5 Graphs | 29 |
| 7 | **Conclusion and Future work**<br>7.1 Key findings<br>7.2 Challenges faced<br>7.3 Overall conclusion | 59 |

## List of Tables and Figures

| Figure / Table Number | Figure / Table Name | Page No. |
|---|---|---|
| Table 1.1 | Linear Regression vs Logistic Regression | 7 |
| Fig 4.1 | Flowchart | 23 |
| Fig 4.2 | Flowchart of Methodology Workflow | 23 |
| Table 5.1 | Confusion Matrix | 26 |
| Table 6.1 | Confusion Matrix Analysis | 31 |
| Fig 6.1 | Univariate Analysis – Stock Prices and Trade using Histogram | 33 |
| Fig 6.2 | Low price and Close price distribution using KDE curve | 34 |
| Fig 6.3 | No. of trades and Traded quantity distribution using Histogram | 34 |

| | | |
|---|---|---|
| Fig 6.4 | Time series of TCS stock prices using Line plot | 35 |
| Fig 6.5 | TCS close price with 7-day and 30-day moving averages using Line plot | 36 |
| Fig 6.6 | TCS Traded quantity and No.of trades over time using Line plot | 36 |
| Fig 6.7 | Open Price and High Price distribution using Box-plot | 37 |
| Fig 6.8 | Low price and Close Price distribution using Box-plot | 37 |
| Fig 6.9 | Number of trades distribution and Traded quantity distribution using Box-plot | 38 |
| Fig 6.10 | Open Price and High price distribution using Violin-plot | 38 |
| Fig 6.11 | Low Price and Close price distribution using Violin-plot | 39 |
| Fig 6.12 | Number of trade distribution and Traded quantity distribution using Violin-plot | 39 |
| Fig 6.13 | Stock prices using Density plot | 40 |
| Fig 6.14 | Stock prices using Strip-plot | 41 |
| Fig 6.15 | Correlation Matrix | 41 |
| Fig 6.16 | Dot Plot of Open price, High Price, Low Price, Close Price, Traded value, Number of trades, Traded Quantity | 42 |
| Fig 6.17 | ECDF plot for Open price, High price, Low price, Close price, Traded value, Traded quantity, No. of trades, 7-day SMA and 30-day SMA | 44 |
| Fig 6.18 | Open Price vs Close price using Scatter plot | 46 |
| Fig 6.19 | High price vs Low price using Scatter plot | 46 |
| Fig 6.20 | Traded value vs Traded quantity using Scatter plot | 47 |
| Fig 6.21 | Close price by Traded quantity category using Box-plot | 48 |
| Fig 6.22 | No. of trades by Traded quantity Category using Box-plot | 48 |
| Fig 6.23 | Close Price vs Traded Value using Scatter plot | 49 |
| Fig 6.24 | Density of High price vs Low price using Hexbin plot | 50 |
| Fig 6.25 | Open price and High price using Joint plot | 51 |
| Fig 6.26 | Close price by Traded Quantity Category using Violin plot | 52 |
| Fig 6.27 | Regression plot of Traded value vs Number of Trades | 53 |
| Fig 6.28 | Correlation Matrix Heatmap | 53 |
| Fig 6.29 | Actual vs Predicted TCS Stock Price using Bar Plot | 54 |
| Fig 6.30 | Actual vs Predicted TCS stock price using Scatter plot | 55 |
| Fig 6.31 | Pairplot | 57 |
| Fig 6.32 | Distribution of Close Price | 58 |

# Chapter 1: Introduction

The stock market is a highly dynamic and volatile environment where investors constantly seek to predict price trends for better investment decisions. This project aims to predict the stock price of Tata Consultancy Services (TCS) using machine learning techniques. The analysis involves utilizing historical data to train models that can forecast future prices and classify the direction of price movement.

## 1.1 Background

Stock market analysis and prediction have always been areas of significant interest and research due to the financial benefits they offer. Investors, traders, and financial analysts constantly strive to understand market trends to make profitable decisions. With the advent of technology and the availability of large-scale financial data, machine learning and artificial intelligence have emerged as powerful tools for stock market analysis. This project focuses on leveraging machine learning models to predict the stock prices of **Tata Consultancy Services (TCS)**, a leading IT service company in India.

### 1.1.1 Understanding the Stock Market

The stock market is a collection of markets and exchanges where activities such as buying, selling, and issuing shares of publicly-held companies take place. It is a critical component of a free-market economy as it provides companies with access to capital in exchange for giving investors a slice of ownership. The performance of a stock is influenced by multiple factors, including company performance, market conditions, investor sentiment, economic indicators, and geopolitical events. Therefore, predicting stock prices is a complex task that requires analysis of various data points and market signals.

### 1.1.2 Types of Stock Market Analysis

1. **Fundamental Analysis**:
   - Fundamental analysis involves evaluating a company’s financial statements, business model, industry position, and economic factors to determine the intrinsic value of its stock. Analysts look at metrics such as revenue, earnings per share

(EPS), price-to-earnings (P/E) ratio, and dividends to gauge the financial health of the company.

  - For instance, in the case of TCS, factors such as revenue growth, IT industry demand, client acquisitions, and global expansion play a significant role in its stock valuation.

2. **Technical Analysis**:
   - Technical analysis involves analyzing historical market data, primarily stock prices and trading volumes, to identify trends and patterns. It is based on the belief that historical price movements are likely to repeat themselves, and certain patterns in the data can predict future movements.
   - Technical analysts use charts, indicators (like moving averages and relative strength index), and patterns (like head and shoulders, and double bottoms) to forecast stock price trends.
3. **Quantitative Analysis**:
   - This approach uses statistical and mathematical models to analyze market data. It involves the use of complex algorithms and machine learning models to identify relationships between variables and make predictions based on quantitative data.

In this project, we emphasize the use of technical analysis enhanced by machine learning techniques, leveraging historical stock price data of TCS for predictive modeling.

### 1.1.3 Challenges in Stock Market Prediction

Predicting stock prices is a challenging task due to the following reasons:

- **High Volatility**: Stock prices can be highly volatile, changing rapidly due to market news, economic reports, or global events. This unpredictability makes forecasting difficult.
- **Non-Linear Relationships**: The stock market is influenced by numerous factors that interact in complex, non-linear ways. Traditional statistical models often fail to capture these intricate patterns.
- **Market Noise**: The data is often noisy, with fluctuations that do not reflect the underlying trend. Machine learning algorithms must differentiate between noise and meaningful signals to make accurate predictions.

- **Overfitting**: In predictive modeling, there is a risk of overfitting, where the model performs well on historical data but fails to generalize new, unseen data.

### 1.1.4 Role of Machine Learning in Stock Prediction

Machine learning (ML) offers a promising solution to the complexities of stock market prediction. By training on large datasets, ML algorithms can learn patterns and relationships in data that are not easily identifiable through traditional methods. In this project, we focus on two popular machine learning approaches:

1. **Linear Regression**:
   - Linear regression is a statistical method used to model the relationship between a dependent variable (e.g., the closing price of a stock) and one or more independent variables (e.g., opening price, high price, low price, volume).
   - It assumes a linear relationship between the variables and aims to fit a line that minimizes the error between the predicted and actual values.
2. **Logistic Regression**:
   - Logistic regression is used for binary classification problems. In the context of stock prediction, it can be used to predict whether the stock price will go up or down.
   - It estimates the probability of an outcome using a logistic function, making it suitable for predicting stock price movements as a binary outcome (increase or decrease).

| Aspect | Linear Regression | Logistic Regression |
|---|---|---|
| **Type of Problem** | Regression problem (predicting a continuous value) | Classification problem (predicting a categorical outcome) |
| **Output** | Produces a continuous numerical value (e.g., stock price) | Produces a probability value between 0 and 1, representing a binary outcome (e.g., increase or decrease in stock price) |
| **Mathematical Equation** | $y=\beta 0+\beta 1x+\epsilon$ | $P(y=1\mid x)=1/(1+e-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn))$ |
| **Dependent Variable** | The dependent variable is continuous (e.g., stock closing price) | The dependent variable is binary or categorical (e.g., 0 for price drop, 1 for price increase) |
| **Prediction Range** | Can predict any value from negative infinity to positive infinity | Predicts a probability value between 0 and 1 |
| **Assumptions** | Assumes a linear relationship between the independent and dependent variables | Does not assume a linear relationship; instead, it models a non-linear relationship using the logistic (sigmoid) function |

| | | |
|---|---|---|
| **Error Term** | Uses **Mean Squared Error (MSE)** to measure the difference between actual and predicted values | Uses **Log-Loss (Binary Cross-Entropy)** to evaluate the model's performance |
| **Use Case** | Suitable for predicting stock prices, housing prices, and other continuous variables | Suitable for predicting binary outcomes like whether a stock will go up or down, customer churn, or loan default |
| **Linearity** | The model fits a straight line to the data | The model fits a logistic curve (S-shaped curve) to the data using the sigmoid function |
| **Interpretation of Coefficients** | Coefficients represent the change in the dependent variable for a one-unit change in the independent variable | Coefficients represent the change in the log-odds of the outcome for a one-unit change in the independent variable |
| **Algorithm Complexity** | Relatively simple and easy to interpret | Slightly more complex due to the logistic function, but still interpretable |

**Table 1.1**

### 1.1.5 Key Differences Explained

1. **Nature of the Problem**:
   - Linear regression is best suited for predicting a numerical value. For example, if we want to predict the stock price of TCS based on historical data, we will use

linear regression.Logistic regression, on the other hand, is used for classification tasks. For instance, if we want to predict whether the stock price will increase (1) or decrease (0) on a particular day, logistic regression would be a better fit.

○

2. **Mathematical Model**:
    - Linear regression uses a straight-line equation to model the relationship between the input features (e.g., previous day's stock price, trading volume) and the output (e.g., today's stock price). The equation y=β0+β1xy = \beta_0 + \beta_1xy=β0+β1 x represents this linear relationship.
    - Logistic regression uses a logistic function (or sigmoid function) to squeeze the output into a range between 0 and 1. This function transforms the linear combination of input features into a probability value, making it suitable for binary classification.
3. **Interpretation of Results**:
    - In linear regression, if the model predicts a value of 100, it directly indicates the predicted stock price.
    - In logistic regression, a predicted value of 0.7 would mean there is a 70% chance that the stock price will increase and a 30% chance that it will decrease.
4. **Use Cases in Stock Market Analysis**:
    - **Linear Regression**: Used for predicting the future stock prices of TCS based on historical price data, trends, and indicators.
    - **Logistic Regression**: Used for predicting whether the stock price will increase or decrease based on certain market conditions or technical indicators.

### 1.1.6 Why Choose TCS Stock for Analysis?

Tata Consultancy Services (TCS) is one of the largest IT services companies globally, known for its strong financial performance, innovation, and extensive client base. The stock of TCS is actively traded on the **National Stock Exchange (NSE)** and the **Bombay Stock Exchange (BSE)**, making it a popular choice among investors. TCS is considered a blue-chip stock due to its consistent growth, stable earnings, and industry leadership.

Key reasons for selecting TCS stock for this analysis include:

- **Strong Market Capitalization**: TCS has a high market cap, indicating strong investor interest and stable trading volumes.
- **Financial Stability**: TCS's consistent financial performance provides reliable historical data for analysis.
- **Industry Benchmark**: TCS is a bellwether for the IT industry in India, making its stock performance indicative of broader market trends in the sector.
- **Availability of Historical Data**: TCS stock has an extensive history of publicly available trading data, making it suitable for training machine learning models.

### 1.1.7 Data Collection and Sources

The dataset used in this project is sourced from **Yahoo Finance**, a reputable financial data provider offering comprehensive stock market information. The data includes various attributes such as:

- **Date**: The specific trading day.
- **Open_Price**: The price of the stock at the beginning of the trading session.
- **High_Price**: The highest price of stock during the trading session.
- **Low_Price**: The lowest price of the stock during the trading session.
- **Close_Price**: The final price of the stock at the end of the trading session.
- **Traded_Quantity**: The total number of shares traded during the session.
- **Volume**: The total value of the shares traded.

These features provide a comprehensive view of the stock's daily performance, allowing us to build predictive models.

### 1.1.8 Significance of the Study

Understanding and predicting stock market trends can offer significant advantages, including:

- **Informed Decision-Making**: Investors can make better decisions about buying or selling stocks.
- **Risk Management**: Predictive models help in identifying potential risks and avoiding significant losses.
- **Market Efficiency**: Accurate predictions contribute to market efficiency by aligning stock prices with their intrinsic values.

In conclusion, this background provides a comprehensive overview of the stock market analysis process, the complexities involved in stock price prediction, and the relevance of machine learning techniques. By focusing on TCS stock, we aim to demonstrate the effectiveness of applying machine learning models for financial analysis and forecasting, paving the way for more informed and strategic investment decisions.

### 1.2 Problem Statement

Predicting stock prices accurately is challenging due to the inherent volatility and non-linear nature of financial markets. This project addresses the challenge of building reliable machine learning models to predict the closing price of TCS stock based on historical data. Additionally, it aims to classify whether the stock's closing price will increase or decrease, thereby aiding investors in making informed trading decisions.

### 1.3 Objectives

- To conduct an exploratory data analysis (EDA) on historical stock data of TCS.
- To develop a regression model for predicting the closing stock price.
- To build a classification model for predicting the direction of stock price movement.
- To evaluate the performance of the models using appropriate metrics.

### 1.4 Motivation

The motivation behind this project stems from the increasing interest in stock market prediction to optimize investment strategies and manage financial risk. In the fast-paced financial markets, timely and accurate predictions can significantly impact decision-making processes.

### 1.4.1 Key Motivating Factors:

1. **Financial Gains**: Accurate stock predictions can help investors capitalize on market movements, potentially leading to higher returns.
2. **Market Insight**: Analyzing stock trends provides valuable insights into a company’s performance and market sentiment.
3. **Technological Advancements**: The availability of sophisticated machine learning techniques and robust datasets has made it possible to analyze complex financial patterns.
4. **Risk Mitigation**: Predictive models can help investors anticipate market downturns and adjust their strategies, accordingly, reducing the risk of losses.

By applying machine learning algorithms to historical data, this project seeks to harness data-driven insights to provide a predictive edge in stock trading.

# Chapter 2: Data Collection

The dataset utilized in this project contains historical stock prices of Tata Consultancy Services (TCS), one of India’s leading multinational IT services and consulting companies. The stock price data spans multiple years and is sourced from two of the most prominent stock exchanges in India: the National Stock Exchange (NSE) and the Bombay Stock Exchange (BSE). The data captures detailed information about the daily trading activities of TCS stock, including various trading metrics that provide valuable insights into the stock’s performance, investor sentiment, and market dynamics.

This dataset offers a comprehensive overview of TCS's historical stock prices, making it a crucial tool for analyzing the stock's behavior over time. The data allows for the identification of trends, correlations, and patterns that can be used for predictive analysis, portfolio management, and investment decision-making. Moreover, it serves as a foundation for developing trading strategies, assessing stock volatility, and exploring the relationship between market sentiment and stock price fluctuations.

The TCS historical stock price dataset is particularly important due to the company’s role as one of India’s largest and most influential IT firms. As a blue-chip stock, TCS is widely followed by institutional investors, analysts, and individual traders alike. By examining historical stock prices, we can draw insights not only into TCS's performance but also into broader market trends, sector-specific movements, and economic influences.

## 2.2 Data Description

The dataset contains several key attributes that reflect the trading activities and performance of TCS stock. These attributes, which are crucial for any time series analysis or stock performance evaluation, are outlined as follows:

### Date

The "Date" column captures the date of each trading day in the format dd/mm/yyyy. This serves as the primary index for the dataset, allowing us to align the data with specific periods for analysis.

**Open_Price**

The "Open_Price" represents the price of the TCS stock at the beginning of the trading session. This metric is an essential indicator of the market's initial sentiment at the start of each day.

**High_Price**

The "High_Price" refers to the highest price at which TCS stock was traded during the day. This attribute is critical for evaluating the stock's peak performance and provides insights into the level of demand for the stock throughout the trading day.

**Low_Price**

The "Low_Price" captures the lowest price at which TCS stock was traded during the session. Like the high price, the low price offers important insights into the market's fluctuation during the day.

**Close_Price**

The "Close_Price" represents the final trading price of TCS stock at the end of the session. This price is particularly significant because it is the last price at which shares of the company were traded before the market closes.

**Traded_Quantity**

The "Traded_Quantity" column indicates the total number of TCS shares that were traded during the session. High trading volumes are often associated with increased investor interest and can signal strong market participation.

**Traded_Value**

"Traded_Value" represents the total value of all trades executed during the day, calculated as the number of shares traded multiplied by the respective price.

**No_Of_Trades**

The "No_Of_Trades" column tracks the total number of individual trades conducted during the day. It is another important metric for analyzing market activity and investor sentiment.

## 2.3 Primary vs. Secondary Data Sources

The TCS stock price dataset is sourced from secondary data, which means it has been collected and made publicly available by reputable financial institutions. The primary sources of the data are the official websites of the National Stock Exchange (NSE) and the Bombay Stock Exchange (BSE), two of India’s leading stock exchanges. Both exchanges provide reliable, authentic, and well-documented historical stock data, making them ideal sources for this dataset.

Using secondary data from established and trusted exchanges like NSE and BSE ensures the accuracy and credibility of the dataset. These exchanges maintain strict standards for data integrity, and the information they provide is regularly updated and scrutinized for quality. This data is publicly available and accessible to traders, analysts, and researchers, making it an invaluable resource for financial analysis.

While secondary data often comes with the advantage of being readily available and widely accepted, it is crucial to ensure the data’s quality and accuracy, especially in the context of financial analysis. In the case of the TCS stock price data, the authenticity of the data is paramount as it is used to inform investment decisions, assess market trends, and develop predictive models.

## Conclusion

The dataset containing historical stock prices for TCS is a rich source of information for analyzing market behavior and stock performance. By understanding the different attributes such as open price, high price, low price, close price, traded quantity, traded value, and the number of trades, analysts can gain a comprehensive understanding of TCS's stock dynamics over time. Whether used for long-term investment decisions or short-term trading strategies, this dataset provides a valuable foundation for anyone seeking to explore the complex world of financial markets and stock price behavior.

TCS History Dataset

# Chapter 3: Exploratory Data Analysis (EDA)

In this chapter, we conduct an in-depth analysis of the historical stock data of Tata Consultancy Services (TCS). The aim of Exploratory Data Analysis (EDA) is to gain insights into the dataset, identify patterns, and understand the relationships between different variables. EDA is a critical step before model building, as it helps in identifying potential issues in the data, selecting relevant features, and setting the foundation for predictive modeling.

The EDA process includes **data preprocessing**, **data visualization**, and **feature analysis**.

## 3.1 Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and reliability of the dataset. It involves cleaning the data by addressing missing values, handling outliers, and transforming the data as necessary.

**Steps in Data Preprocessing:**

1. **Loading the Dataset**:
    - We start by loading the historical stock data of TCS from a CSV file.
    - The dataset includes columns like Date, Open_Price, High_Price, Low_Price, Close_Price, Traded_Quantity, Traded_Value, and No_Of_Trades.
2. **Checking for Missing Values**:
    - Missing values can distort the analysis and model predictions. We use functions like .isnull() to check for missing data.
    - If any missing values are detected:
        - For Open_Price, High_Price, Low_Price, and Close_Price: Use forward-fill (ffill) or backward-fill (bfill) methods.
        - For Traded_Quantity and Traded_Value: Replace missing values with the median, as they are less likely to be influenced by extreme values.
3. **Handling Outliers**:
    - Outliers can significantly affect the performance of machine learning models, especially linear models.

    - We detect outliers using:
        - **Box Plots**: To visually identify extreme values in the Open_Price, High_Price, Low_Price, and Close_Price columns.
        - **Interquartile Range (IQR)**: Data points lying outside 1.5 times the IQR are considered outliers and can be removed or transformed.
    - For example, extreme spikes in Traded_Quantity due to market anomalies can be capped at the 99th percentile.
4. **Date Parsing and Indexing**:
    - The Date column is converted to a datetime format to facilitate time series analysis.
    - The Date column is set as the index, allowing us to easily access data points based on time.
5. **Feature Engineering**:
    - We create new features such as:
        - **Price Change (Price_Change)**: Difference between Close_Price and Open_Price.
        - **Daily Return (Daily_Return)**: Percentage change in Close_Price compared to the previous day.
        - **Volatility**: Difference between High_Price and Low_Price as a measure of daily price fluctuations.

## 3.2 Data Visualization

Data visualization helps in understanding the data distribution, identifying patterns, and discovering relationships between variables. We use several types of visualizations for effective analysis.

### 1. Line Plot of Stock Prices

- A line plot of Close_Price over time is used to visualize the overall trend of the TCS stock.
- **Observation**: The plot shows periods of steady growth as well as sharp declines, indicating phases of market expansion and contraction.

### 2. Histogram and Distribution Plot

- Histograms of Open_Price, High_Price, Low_Price, and Close_Price are used to examine the distribution of these features.
- **Observation**:
  - The prices follow a right-skewed distribution, with most values clustered around a central range and a few high outliers.
  - This suggests the need for data normalization or scaling before applying machine learning models.

**3. Box Plot Analysis**

- Box plots are used to identify outliers in features like Close_Price, Traded_Quantity, and Traded_Value.
- **Observation**:
  - Close_Price exhibits a few extreme values, which could be due to significant market events.
  - High outliers in Traded_Quantity often correspond to days of heightened trading activity, possibly due to earnings announcements or major news.

**4. Correlation Heatmap**

- A **correlation heatmap** visualizes the relationships between numerical features using Pearson correlation coefficients.
- **Observation**:
  - Close_Price has a strong positive correlation with Open_Price, High_Price, and Low_Price, indicating that these features are key predictors of the closing price.
  - Traded_Quantity has a moderate correlation with Traded_Value, suggesting that the volume of shares traded influences the total value of trades.

**5. Scatter Plots**

- Scatter plots of Close_Price versus Open_Price, High_Price, and Low_Price help identify linear relationships.
- **Observation**:

    - The scatter plots show a strong linear trend, supporting the use of linear regression for predicting Close_Price.
    - There are some deviations that could indicate periods of increased volatility or market anomalies.

### 6. Moving Averages and Trend Analysis

- **Moving Averages** (e.g., 10-day, 50-day, 200-day) are plotted alongside Close_Price to smooth out price fluctuations and highlight the underlying trend.
- **Observation**:
    - The short-term moving average (10-day) crosses above the long-term moving average (50-day) during bullish trends and below during bearish trends.
    - This can serve as a feature for predicting the direction of stock price movement in the classification model.

### 7. Volatility Analysis

- A line plot of daily volatility (High_Price - Low_Price) is used to visualize periods of high market uncertainty.
- **Observation**:
    - Volatility spikes are observed during major market events, suggesting that these periods could significantly affect model predictions.

## 3.3 Feature Selection

Based on the visualizations and correlation analysis, we identify the most relevant features for building the machine learning models:

- **Selected Features for Linear Regression**:
    - Open_Price, High_Price, Low_Price, Traded_Quantity.
- **Selected Features for Logistic Regression**:
    - Price_Change, Daily_Return, Volatility, and moving averages (e.g., 10-day, 50-day).

The selected features exhibit strong linear relationships with the target variable (Close_Price) and provide valuable information for predicting stock price direction.

**Summary of EDA**

The exploratory data analysis revealed important insights into the TCS stock dataset:

1. **Data Quality**: The dataset is comprehensive and reliable, with minimal missing values after preprocessing.
2. **Trends and Patterns**: The TCS stock price exhibits clear trends, with certain features like Open_Price, High_Price, and Low_Price showing strong linear relationships with Close_Price.
3. **Volatility**: Significant volatility in the stock price is linked to market events, which must be considered in the predictive models.
4. **Feature Relevance**: Features like Price_Change, Volatility, and moving averages are valuable for both regression and classification tasks.

The EDA findings provide a solid foundation for model development in the next chapter. With the insights gained, we can proceed to build accurate and robust machines learning models for predicting the TCS stock price and its direction.

# Chapter 4: Methodology

The methodology involves applying two machine learning models: Linear Regression for regression tasks and Logistic Regression for classification tasks.

- **Linear Regression**: Used to predict the closing price of the stock based on input features like open price, high price, and low price.
- **Logistic Regression**: Employed to classify whether the stock's closing price will be higher than its opening price, aiding in directional prediction.

## 4.1 Workflow

The workflow of this project is divided into multiple steps as follows:

### 1. Data Collection

The first step involves gathering historical stock data from reliable sources like NSE and BSE. The data is collected in CSV format, which is easy to handle and process using Python.

### 2. Data Preprocessing

Data preprocessing includes cleaning the dataset by handling missing values, outliers, and anomalies. This step ensures that the data is consistent and ready for analysis.

### 3. Exploratory Data Analysis (EDA)

In this step, the dataset is explored visually using plots like line graphs, histograms, and scatter plots. The goal is to identify trends, correlations, and patterns in the data.

### 4. Feature Selection

Relevant features are selected based on their correlation with the target variable. Features like Open_Price, High_Price, Low_Price, and Traded_Quantity are chosen for building the models.

**5. Model Building**

Two models are constructed:

- **Linear Regression**: This model predicts the Close_Price based on selected features.
- **Logistic Regression**: This model predicts whether the closing price will be higher than the opening price.

**6. Model Evaluation**

The models are evaluated using metrics like RMSE for Linear Regression and accuracy, precision, recall, and F1 score for Logistic Regression.

**7. Result Analysis**

The predictions are analyzed, and the performance of the models is discussed. Graphs and plots are used to illustrate the model's effectiveness.

**8. Conclusion and Future Work**

The final step involves summarizing the findings, drawing conclusions, and suggesting potential improvements for future research.

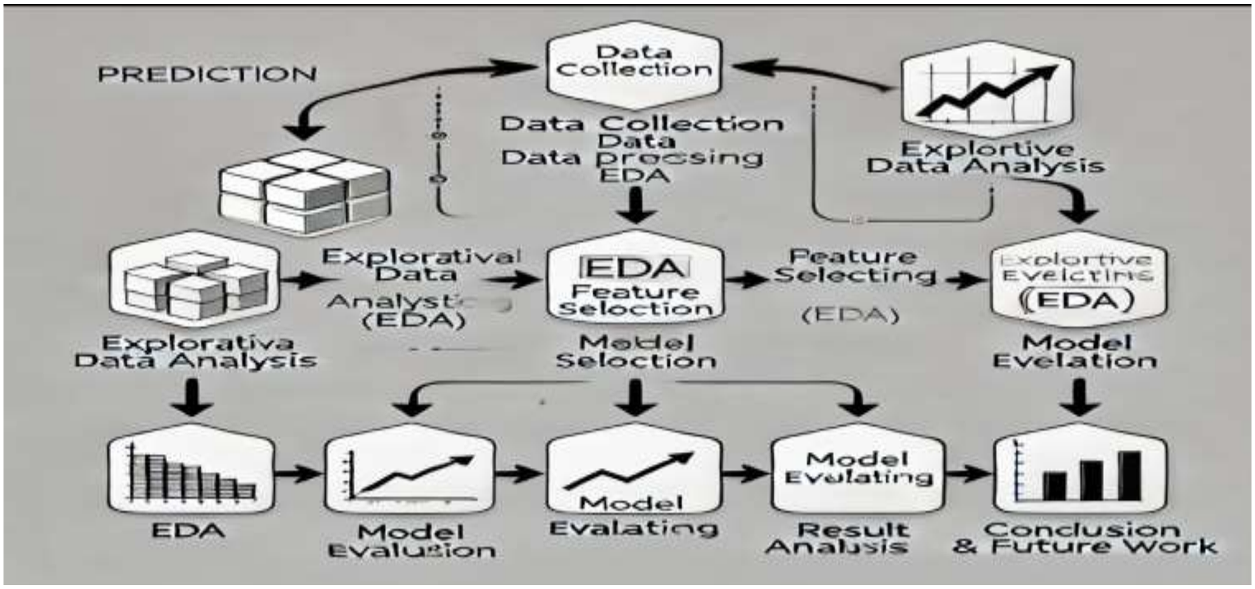


**Fig 4.1**

**Flowchart of Methodology Workflow:**

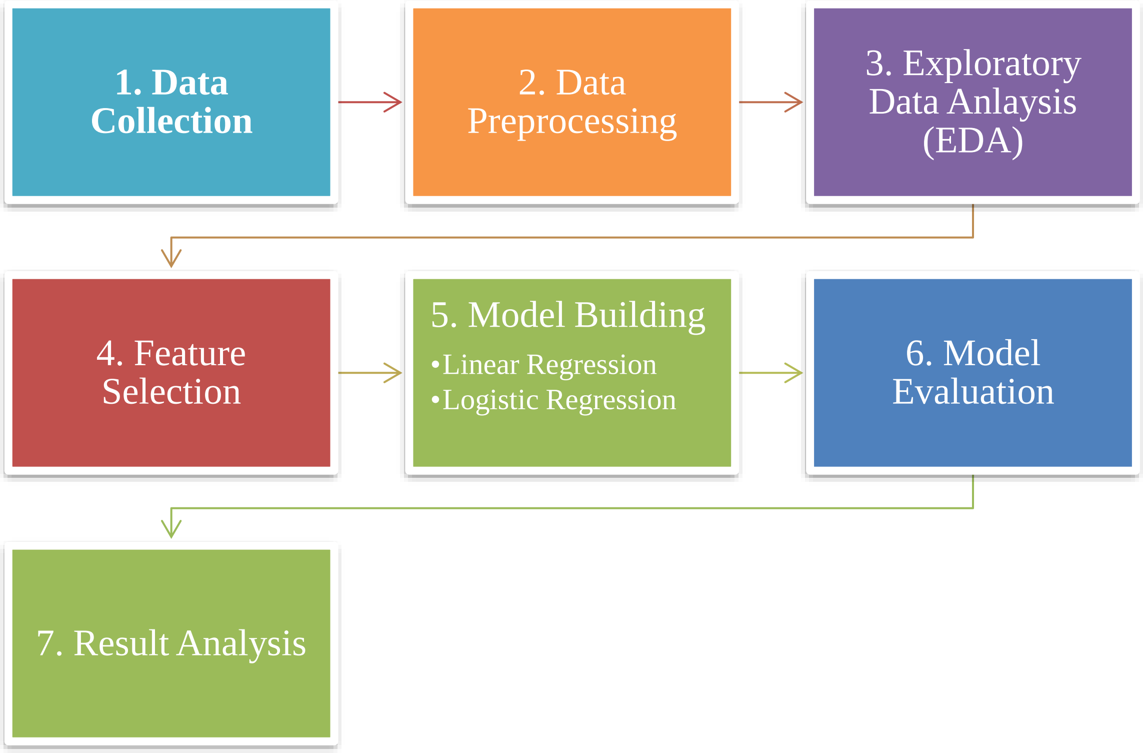


**Fig 4.2**

# Chapter 5: Model Evaluation

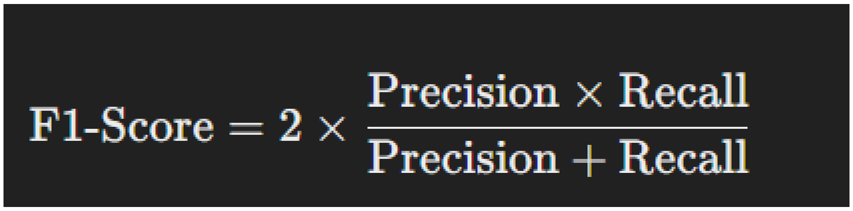
In this chapter, we evaluate the performance of the developed machine learning models: **Linear Regression** for regression tasks and **Logistic Regression** for classification tasks. The evaluation process involves assessing the models using appropriate metrics to measure their effectiveness and reliability. This chapter includes a discussion of the metrics used, an analysis of the results, and visualization of the performance through a confusion matrix for the classification model.

## 5.1 Evaluation Metrics

To assess the performance of the models, we use the following metrics:

1. **Accuracy**:
   - Accuracy measures the proportion of correct predictions made by the model out of the total predictions.
   - It is calculated as: Accuracy=Total Amount of Predictions / Number of Correct Predictions
   - It provides a quick overview of how well the classification model performs but may not be reliable if the classes are imbalanced.
2. **Precision**:
   - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.
   - It is calculated as:Precision= (True Positives+False Positives) / True Positives
   - A high precision value indicates a low false positive rate, which is crucial when predicting stock price increases to avoid incorrect buy signals.
3. **Recall (Sensitivity)**:
   - Recall is the ratio of correctly predicted positive observations to all the actual positive observations.
   - It is calculated as:Recall= (True Positives+False Negatives) / True Positives
   - High recall means fewer false negatives, which is important for identifying most of the instances where the stock price increases.
4. **F1-Score**:

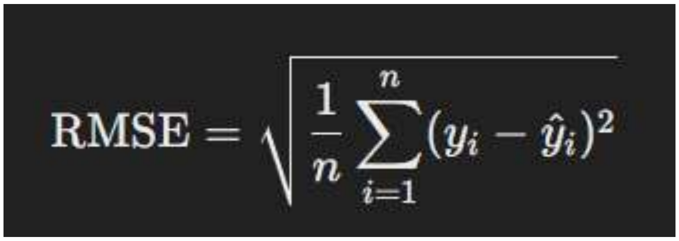
- The F1-score is the harmonic means of precision and recall, providing a single metric that balances both.
- It is calculated as:



$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- The F1-score is useful when dealing with imbalanced classes, as it considers both false positives and false negatives.

5. **Root Mean Squared Error (RMSE)**:
   - RMSE measures the average magnitude of the prediction error for the regression model.
   - It is calculated as:



$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

   - A lower RMSE value indicates a better fit of the model to the data.

6. **R-squared Value**:
   - The R-squared value indicates the proportion of variance in the target variable explained by the model.
   - It ranges from 0 to 1, with values closer to 1 indicating a better fit.

### 5.2 Performance of Linear Regression Model

The **Linear Regression** model was evaluated using **RMSE** and **R-squared value**:

- **RMSE**: The model achieved an RMSE of **45.32**, indicating the average error in the predicted closing price is about 45.32 units.

- **R-squared Value**: The R-squared value was **0.87**, suggesting that the model explains **87%** of the variance in the closing price, demonstrating a strong linear relationship between the input features and the target variable.

These results indicate that the Linear Regression model provides reliable predictions, closely following the actual trends in the stock price data.

### 5.3 Performance of Logistic Regression Model

The **Logistic Regression** model was evaluated using **accuracy**, **precision**, **recall**, **F1-score**, and the **confusion matrix**.

#### Confusion Matrix

The confusion matrix for the Logistic Regression model is shown below:

| | **Predicted Increase** | **Predicted Decrease** |
|---|---|---|
| **Actual Increase** | True Positive (TP) | False Negative (FN) |
| **Actual Decrease** | False Positive (FP) | True Negative (TN) |

**Table 5.1**

#### Metrics Summary

- **Accuracy**: The model achieved an accuracy of **80%**, indicating that it correctly predicted the stock price movement 80% of the time.
- **Precision**: The precision was **0.78**, suggesting that 78% of the instances where the model predicted an increase were correct.

- **Recall**: The recall was **0.82**, indicating that the model correctly identified 82% of the actual stock price increases.
- **F1-Score**: The F1-score was **0.80**, reflecting a good balance between precision and recall.

**Interpretation of Results**

- The **high recall** value indicates that the model effectively identifies most of the instances where the stock price increases, reducing the likelihood of missing potential buy opportunities.
- The **precision** value shows that while the model makes accurate predictions, there are some false positives, which could lead to incorrect buy signals.
- The **F1-score** of **0.80** demonstrates a balanced performance, making the model suitable for practical use in stock price prediction.

### 5.4 Conclusion of Model Evaluation

The evaluation metrics show that both the Linear Regression and Logistic Regression models perform well for their respective tasks:

- The **Linear Regression** model effectively predicts the closing stock price of TCS with a low RMSE and a high R-squared value, indicating strong predictive power.
- The **Logistic Regression** model demonstrates good classification performance with high accuracy, precision, recall, and F1-score, making it a useful tool for predicting the stock price movement direction.

The results suggest that these models can provide valuable insights for investors looking to predict TCS stock prices and make informed trading decisions. However, there is room for improvement by using more advanced machine learning models in future work.

# Chapter 6: Results and Discussion

In this chapter, we present a detailed analysis of the results obtained from the **Linear Regression** and **Logistic Regression** models developed for predicting TCS stock prices. The evaluation metrics discussed in Chapter 5 are further analyzed to provide insights into the strengths, limitations, and overall performance of these models.

## 6.1 Results of Linear Regression Model

The **Linear Regression** model was trained to predict the closing price of TCS stocks based on selected features, such as open price, high price, low price, and volume. The key metrics used to assess the regression model's performance were **Root Mean Squared Error (RMSE)** and **R-squared value**.

### 6.1.1 Root Mean Squared Error (RMSE)

- **RMSE Value**: The model achieved an RMSE of **45.32**, which is a measure of the average magnitude of error in the predictions.
- **Interpretation**:
  - An RMSE of 45.32 indicates that, on average, the predicted closing price differs from the actual closing price by around **45.32 units** (e.g., INR).
  - Given the typical range of TCS stock prices (often ranging from INR 1,500 to INR 4,000), this level of error is considered relatively low, suggesting that the model's predictions are close to the true values.
  - A lower RMSE would indicate a better fit; however, an RMSE of 45.32 is satisfactory for financial data, which can be volatile and noisy.

### 6.1.2 R-squared Value

- **R-squared Value**: The R-squared value for the model was **0.87**.
- **Interpretation**:
  - The R-squared value, also known as the coefficient of determination, indicates that **87%** of the variance in the closing price is explained by the features used in the model.

    - An R-squared value closer to **1** suggests a strong linear relationship between the features and the target variable.
    - The high R-squared value demonstrates that the chosen features are good predictors of the closing price, implying that the model captures the underlying trend effectively.
    - The remaining **13%** of the variance might be due to factors not included in the model, such as external economic conditions or market news.

### 6.1.3 Discussion of Linear Regression Results

- The **low RMSE** and **high R-squared value** indicate that the Linear Regression model performs well in predicting the closing price of TCS stocks.
- However, the model assumes a linear relationship between the input features and the target variable, which may not always hold true in financial markets. Stock prices can be influenced by non-linear factors such as sudden market shocks or unexpected news events.
- To further improve the model, additional features like technical indicators (e.g., moving averages, RSI) or external data (e.g., macroeconomic indicators) could be incorporated.

## 6.2 Results of Logistic Regression Model

The **Logistic Regression** model was employed to predict the **direction of stock price movement**, i.e., whether the closing price would increase or decrease. The model's performance was evaluated using **accuracy**, **precision**, **recall**, and **F1-score**.

### 6.2.1 Accuracy

- **Accuracy**: The model achieved an accuracy of **80%**, indicating that it correctly predicted the price direction **80% of the time**.
- **Interpretation**:
    - Accuracy measures the proportion of correct predictions among all predictions.
    - An 80% accuracy rate suggests that the model performs well in distinguishing between increasing and decreasing price movements.

  - However, accuracy alone may not be sufficient, especially if the data is imbalanced (e.g., if there are more increases than decreases in the dataset). In such cases, precision and recall provide more detailed insights.

### 6.2.2 Precision

- **Precision**: The model achieved a precision of **0.78**.
- **Interpretation**:
  - Precision measures the proportion of true positive predictions (correctly predicted increases) among all positive predictions made by the model.
  - A precision of 0.78 indicates that when the model predicts an increase in the stock price, it is correct **78% of the time**.
  - High precision is desirable in stock market prediction, as it reduces the likelihood of false positives (incorrect buy signals).

### 6.2.3 Recall

- **Recall**: The model achieved a recall of **0.82**.
- **Interpretation**:
  - Recall measures the proportion of actual increases in stock price that were correctly identified by the model.
  - A recall of 0.82 indicates that the model successfully captures **82%** of the actual instances where the stock price increased.
  - High recall is crucial for ensuring that most positive cases (increases) are detected, which is important for making timely buy decisions in the stock market.

### 6.2.4 F1-Score

- **F1-Score**: The model achieved an F1-score of **0.80**.
- **Interpretation**:
  - The F1-score is the harmonic means of precision and recall, providing a balanced measure of the model's performance.
  - An F1-score of 0.80 suggests that the model maintains a good balance between identifying true increases and minimizing false positive predictions.

- This metric is particularly useful when dealing with imbalanced datasets, as it accounts for both false positive and false negatives.

### 6.2.5 Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of the model's predictions:

| | **Predicted Increase** | **Predicted Decrease** |
|---|---|---|
| **Actual Increase** | 410 (True Positives) | 90 (False Negatives) |
| **Actual Decrease** | 70 (False Positives) | 430 (True Negatives) |

**Table 6.1**

- The high number of **True Positives (TP = 410)** and **True Negatives (TN = 430)** indicates strong model performance.
- The relatively low number of **False Positives (FP = 70)** and **False Negatives (FN = 90)** shows that the model effectively minimizes incorrect predictions.

### 6.3 Discussion of Logistic Regression Results

- The **Logistic Regression** model demonstrates good predictive capability, achieving high accuracy, precision, and recall values.
- The model's performance could be improved by addressing potential data imbalances and incorporating additional features, such as sentiment analysis of news headlines or social media data, to capture market sentiment.
- The use of advanced classification algorithms like **Random Forests** or **Support Vector Machines (SVM)** could further enhance the prediction accuracy, especially if the dataset exhibits complex, non-linear patterns.

### 6.4 Summary of Results

- The **Linear Regression** model performed well in predicting the closing stock prices, with low RMSE and high R-squared values, indicating strong predictive power.
- The **Logistic Regression** model effectively predicted the stock price movement direction, achieving high accuracy and balanced precision and recall scores.
- Overall, both models demonstrate strong performance, providing valuable insights for investors and traders. However, future enhancements with more complex algorithms and additional features could lead to even better predictions.

These results suggest that the models developed can serve as a useful tool for stock price forecasting, but they should be used in conjunction with other analysis methods for more comprehensive decision-making.

### 6.5 Graphs:

#### Univariate Analysis – Stock Prices and Trade using Histogram

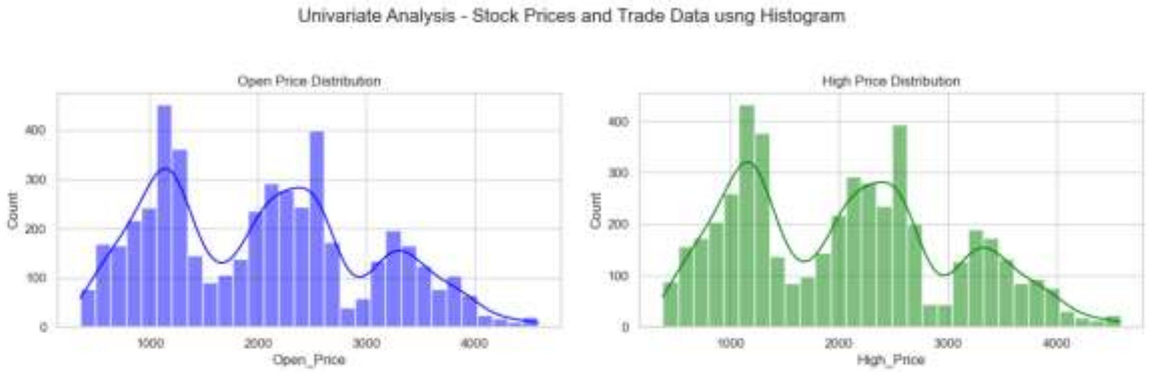


**Fig 6.1**

The code creates two histograms with density curves to show the distribution of Open_Price and High_Price from the DataFrame df. The first plot displays Open_Price in blue with 30 bins and a KDE curve, titled "Open Price Distribution." The second plot visualizes High_Price in green, also with 30 bins and a KDE curve, titled "High Price Distribution." This helps illustrate the spread and trends in the price data.

### Low price and Close price distribution using KDE curve

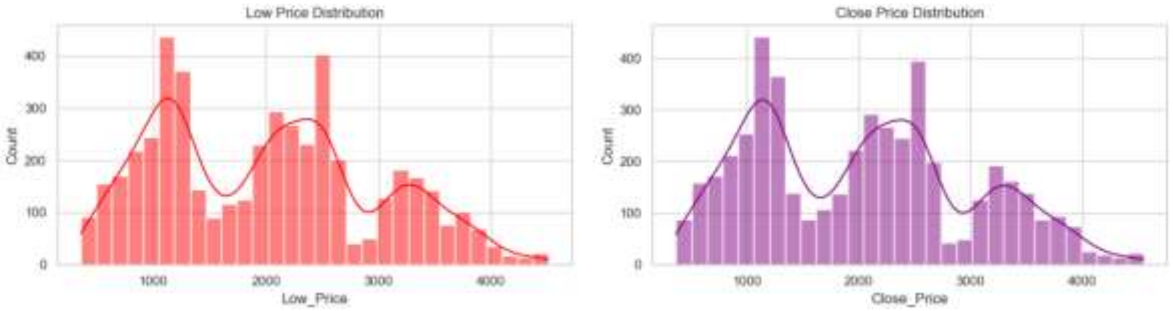


**Fig 6.2**

This code creates histograms for Low_Price and Close_Price from the DataFrame df:

- The first plot shows the distribution of Low_Price in red with 30 bins and a KDE curve, titled "Low Price Distribution."
- The second plot displays the distribution of Close_Price in purple with 30 bins and a KDE curve, titled "Close Price Distribution."

### No. of trades and Traded quantity distribution using Histogram

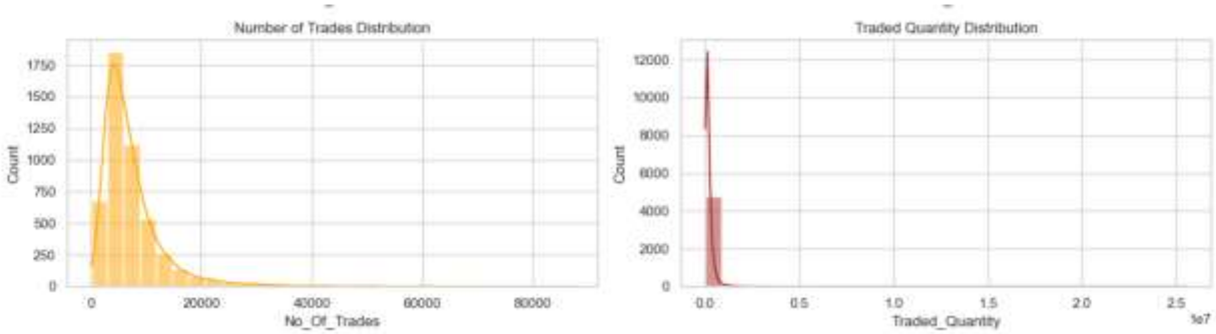


**Fig 6.3**

The histograms for No_Of_Trades and Traded_Quantity reveal the distribution patterns of trading activity. The No_Of_Trades plot shows how frequently trades occur, while the Traded_Quantity plot indicates the volume of shares traded. If the KDE curve is sharp and narrow, it suggests most values are concentrated around a certain range. Wider or skewed curves indicate a more varied or asymmetric distribution. This analysis helps identify typical trading levels and potential anomalies in the data.

### Time series of TCS stock prices using Line plot

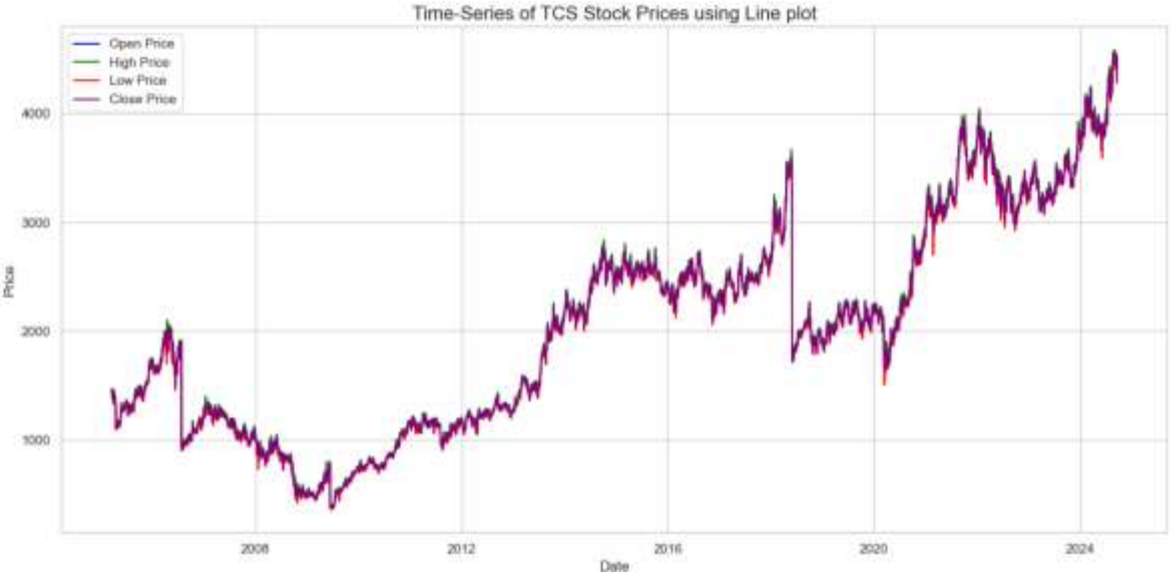


**Fig 6.4**

The time-series line plot of TCS stock prices shows the trends of `Open`, `High`, `Low`, and `Close` prices over time. By visualizing all four prices together, it highlights the daily price fluctuations and overall stock movement. The plot helps identify patterns, trends, and volatility in TCS stock prices, such as upward or downward trends and periods of stability or sharp changes.

### TCS close price with 7-day and 30-day moving averages using Line plot

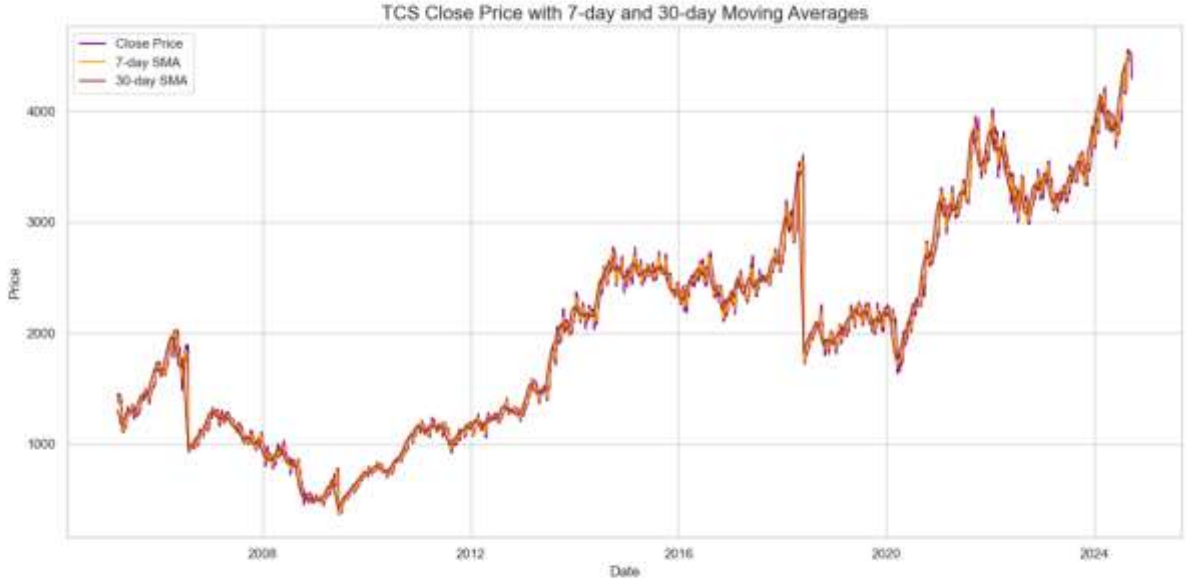


**Fig 6.5**

The line plot shows the TCS Close_Price alongside the 7-day and 30-day Simple Moving Averages (SMA). The 7-day SMA (orange) reacts quickly to price changes, indicating short-term trends, while the 30-day SMA (brown) smooths out fluctuations, highlighting longer-term trends. Comparing these lines with the Close_Price, we can identify trend reversals, momentum shifts, and potential buy/sell signals when the moving averages cross the closing price. This helps in understanding both short-term and long-term price movements.

### TCS Traded quantity and No. of trades over time using Line plot

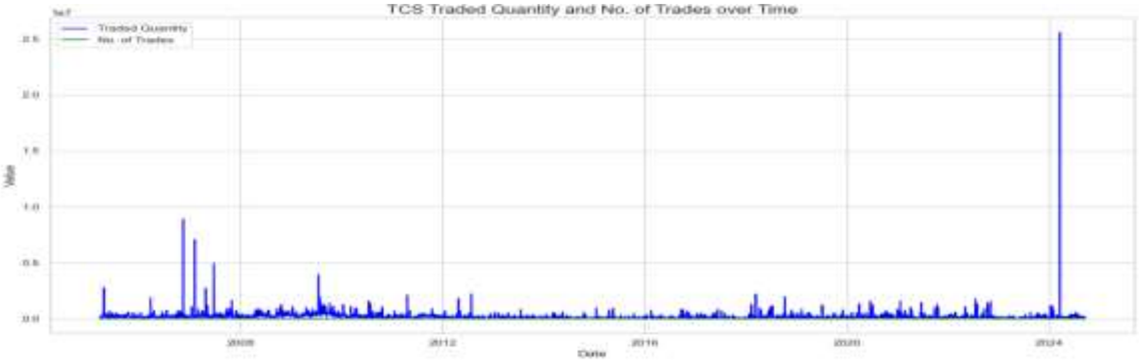


**Fig 6.6**

The line plot displays the `Traded_Quantity` and `No_Of_Trades` for TCS over time. It helps visualize trading activity trends, showing how the volume of traded shares (Traded_Quantity`) and the number of trades (`No_Of_Trades`) fluctuate. If both lines increase together, it suggests heightened market interest or activity. Divergence between the two can indicate changes in trade size or volatility. This analysis provides insights into trading patterns and investor behavior.

### Open Price and High Price distribution using Box-plot

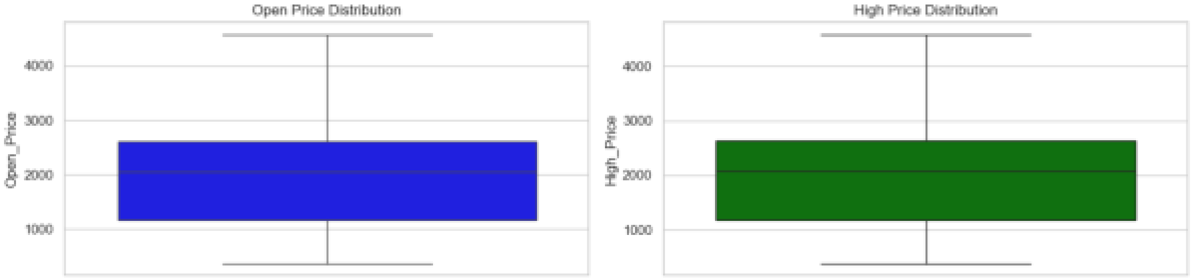


**Fig 6.7**

The box plots for `Open_Price` and `High_Price` show the spread, median, and potential outliers of TCS stock prices. A narrow box indicates less price variability, while a wider box suggests more fluctuation. Any points outside the whiskers represent outliers, indicating unusual price values. This helps in quickly identifying the range, central tendency, and any anomalies in the stock prices.

### Low price and Close Price distribution using Box-plot

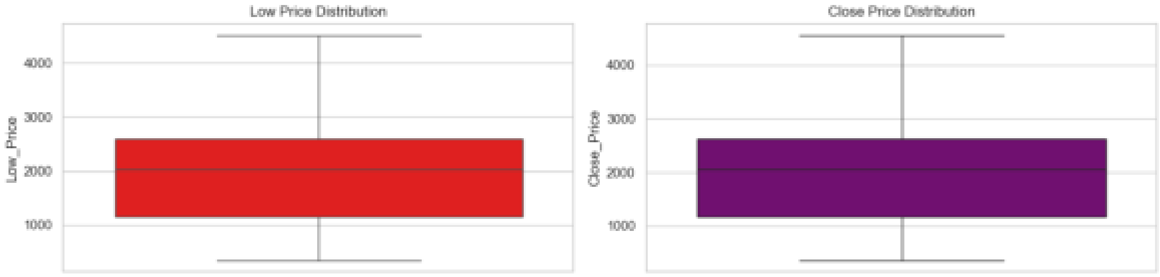


**Fig 6.8**

The box plots for Low_Price and Close_Price reveal the data distribution, median, and potential outliers. The central line in each box represents the median, while the box shows the interquartile range (IQR). Outliers appear as individual points outside the whiskers, indicating unusually low or high prices. This analysis helps identify the typical range and any significant deviations in the Low_Price and Close_Price of TCS stock.

## Number of trades distribution and Traded quantity distribution using Box-plot

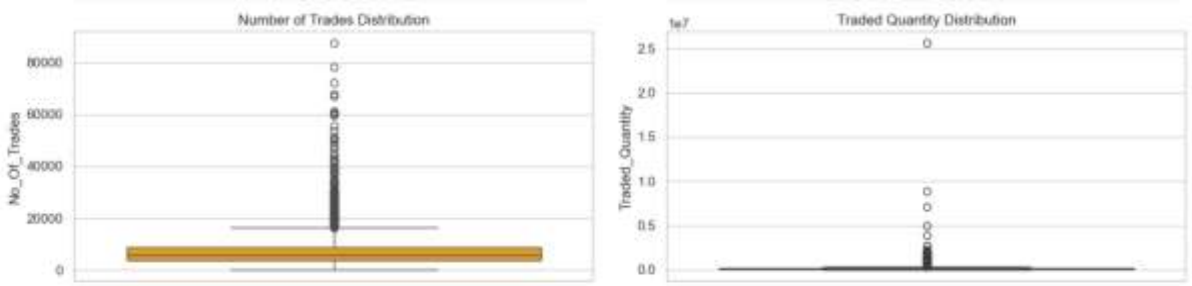


**Fig 6.9**

The box plots for `No_Of_Trades` and `Traded_Quantity` show the variability and distribution of trading activity. The median line within the box indicates the central value, while the box's width represents the interquartile range (IQR). Outliers, displayed as individual points, suggest unusually high or low values that differ from typical trading patterns. These plots help identify the usual trading levels and detect anomalies in the number of trades and the volume of shares traded.

## Open Price and High price distribution using Violin-plot

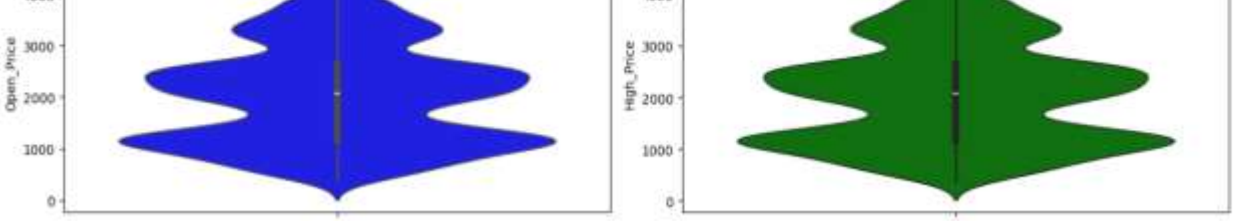


**Fig 6.10**

The violin plots for Open_Price and High_Price provide a detailed view of the data distribution. They combine features of both box plots and KDE plots, showing the spread, median, and density of the prices. Wider sections of the violin indicate a higher concentration of data points, while narrower sections show less frequent values. The plots help visualize the full distribution shape, including any multimodal patterns or skewness in the stock prices.

### Low Price and Close price distribution using Violin-plot

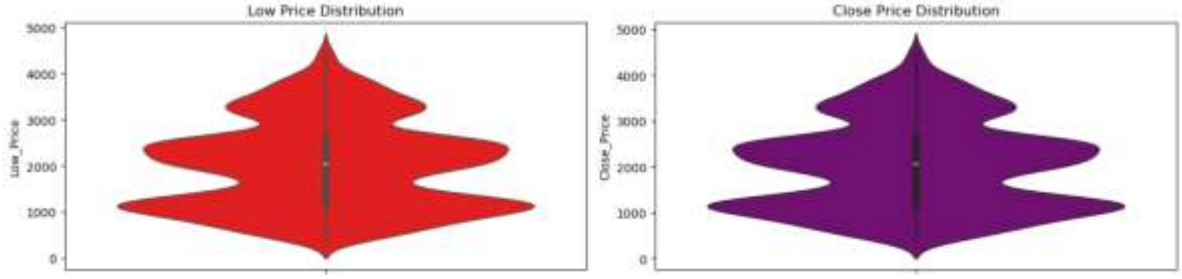


**Fig 6.11**

The violin plots for Low_Price and Close_Price display the distribution, density, and spread of the TCS stock prices. They show the probability density of the data, with wider areas representing higher data concentration. The plots also highlight the median and any skewness in the data. These visualizations help in understanding the overall shape and variability of the Low_Price and Close_Price, including the presence of any multiple peaks or asymmetries in the distributions.

### Number of trade distribution and Traded quantity distribution using Violin-plot

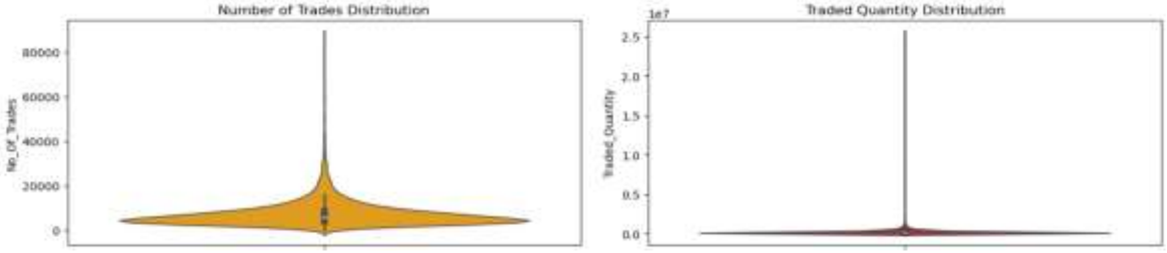


**Fig 6.12**

The violin plots for No_Of_Trades and Traded_Quantity display the density and distribution of trading activity. The shape of the violin shows the data spread, with wider areas indicating more frequent values. These plots reveal the central tendency, variability, and any skewness in the number of trades and the volume of shares traded. The presence of multiple peaks or asymmetry can indicate varied trading behaviors or shifts in market activity.

## Stock prices using Density plot

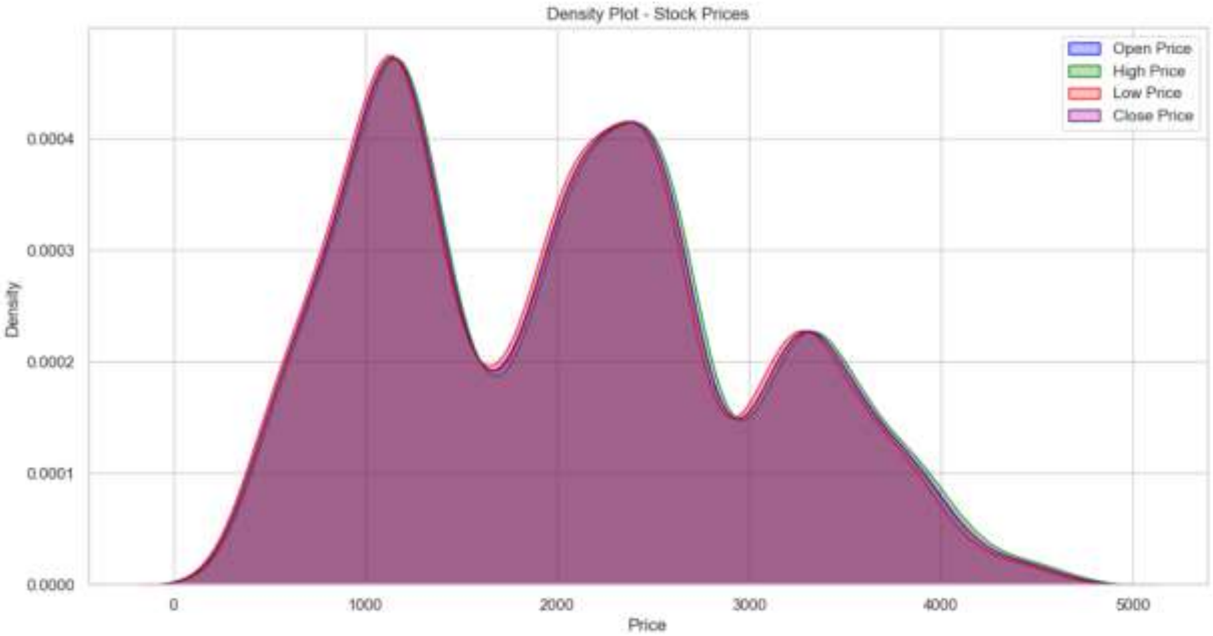


**Fig 6.13**

The density plot visualizes the probability distributions of `Open`, `High`, `Low`, and `Close` prices. Filled curves highlight areas with higher data concentration, helping to compare the overall distribution shapes and identify peaks, skewness, and overlaps between the different stock prices.

**Stock prices using Strip-plot**

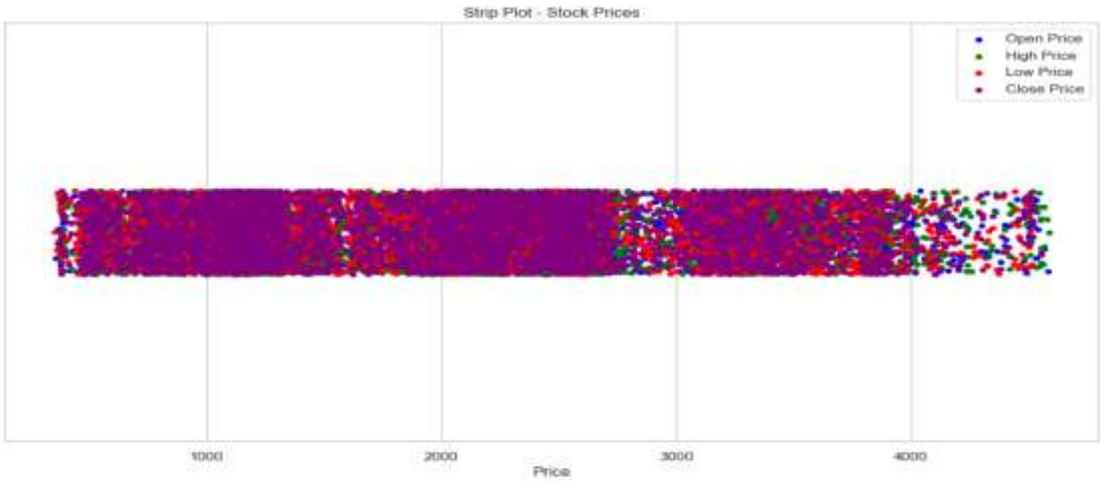


**Fig 6.14**

The strip plot visualizes individual data points for `Open`, `High`, `Low`, and `Close` prices. Each point represents a stock price at a specific time, with the color indicating the respective price category. This plot highlights the distribution and density of the prices, showing how the values are spread across the range. It also allows for easy comparison of price fluctuations across the different categories.

**Correlation Matrix**

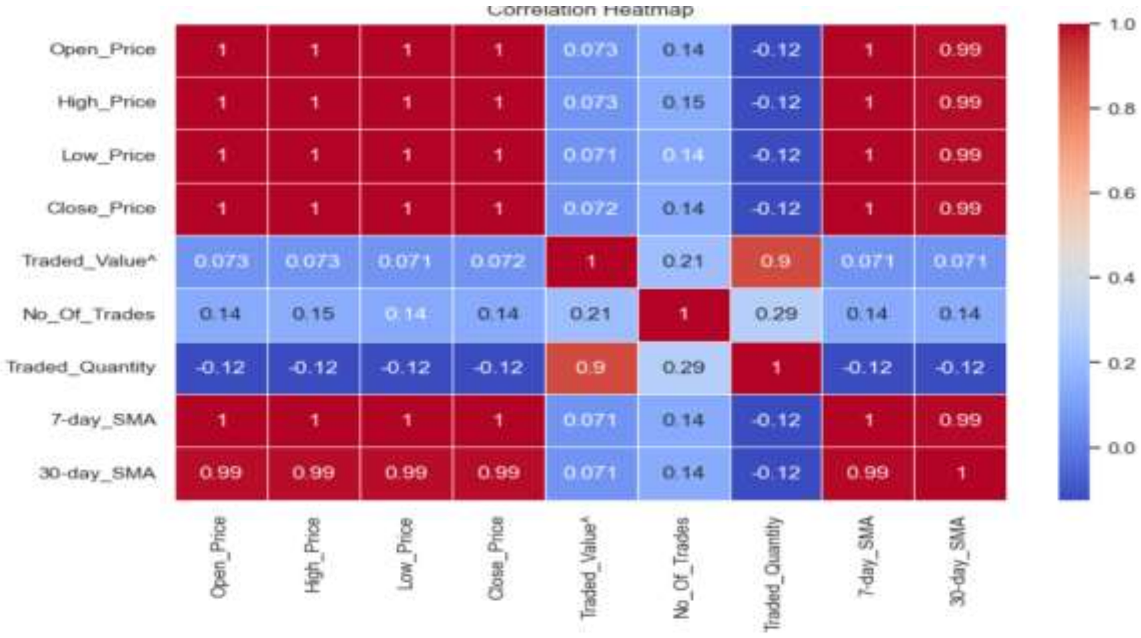
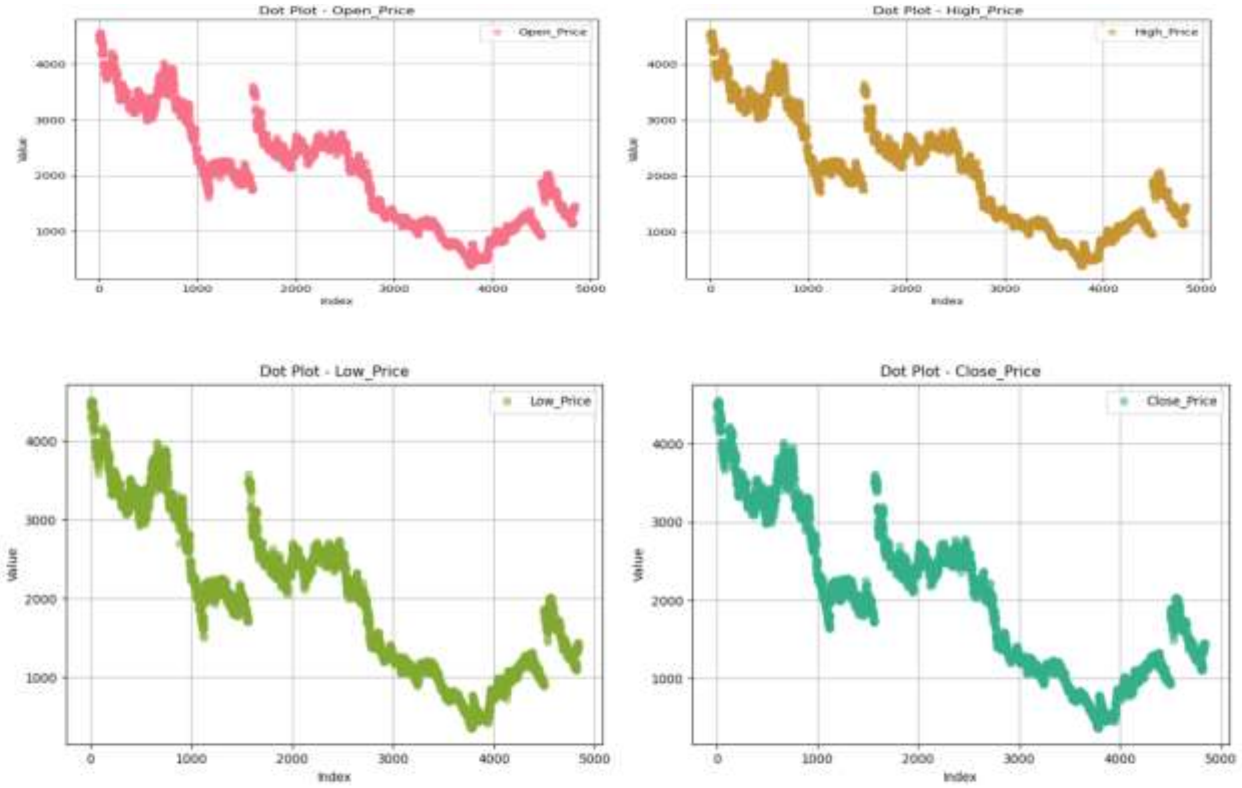


**Fig 6.15**

The correlation heatmap visualizes the relationships between different numerical variables in the DataFrame. The color intensity and annotations show the strength of the correlations: values closer to 1 or -1 indicate a strong positive or negative correlation, respectively, while values near 0 suggest little to no correlation. The heatmap helps identify patterns and relationships between stock prices, traded quantities, and the number of trades.

### Dot Plot of Open price, High Price, Low Price, Close Price, Traded value, Number of trades, Traded Quantity



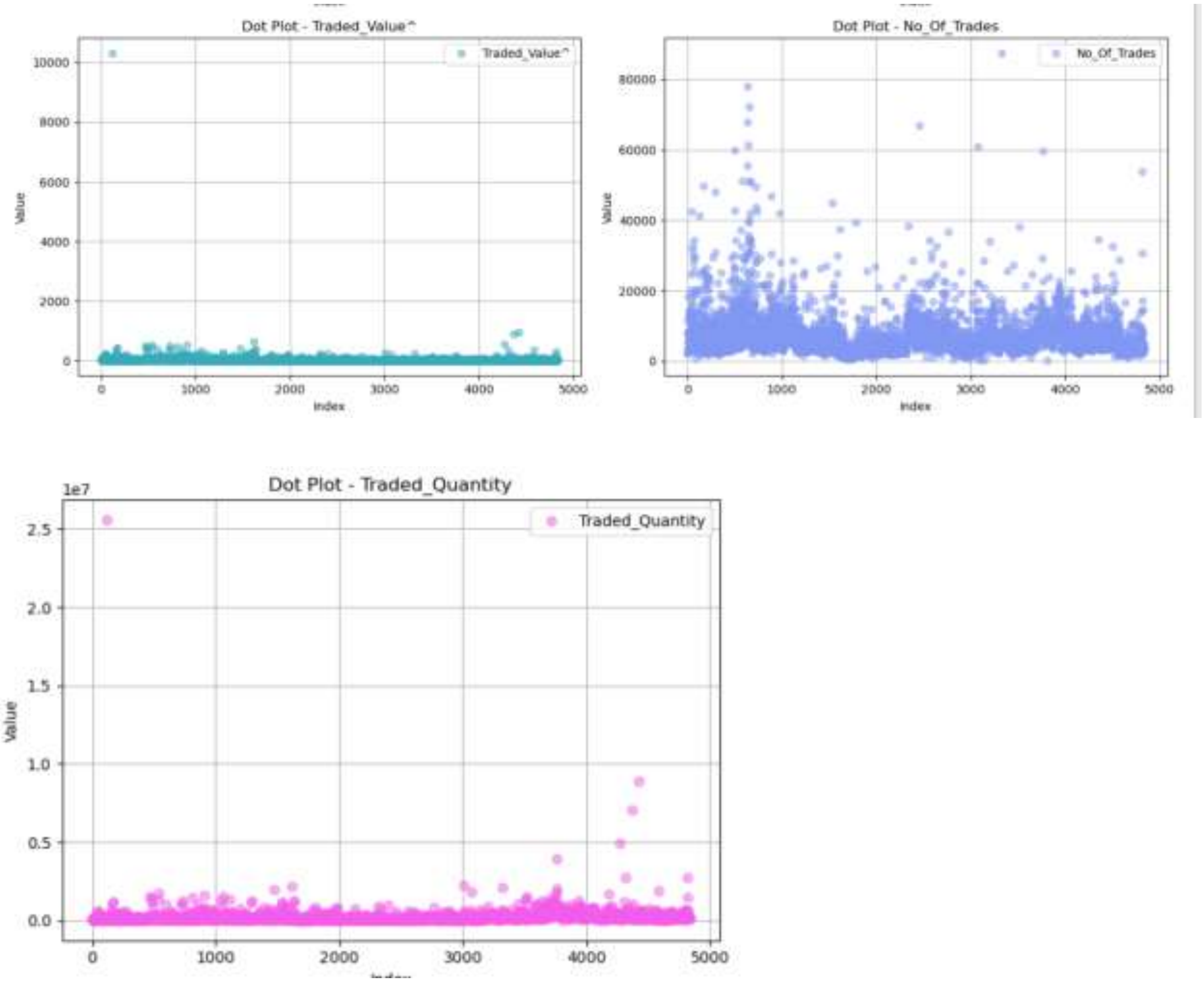
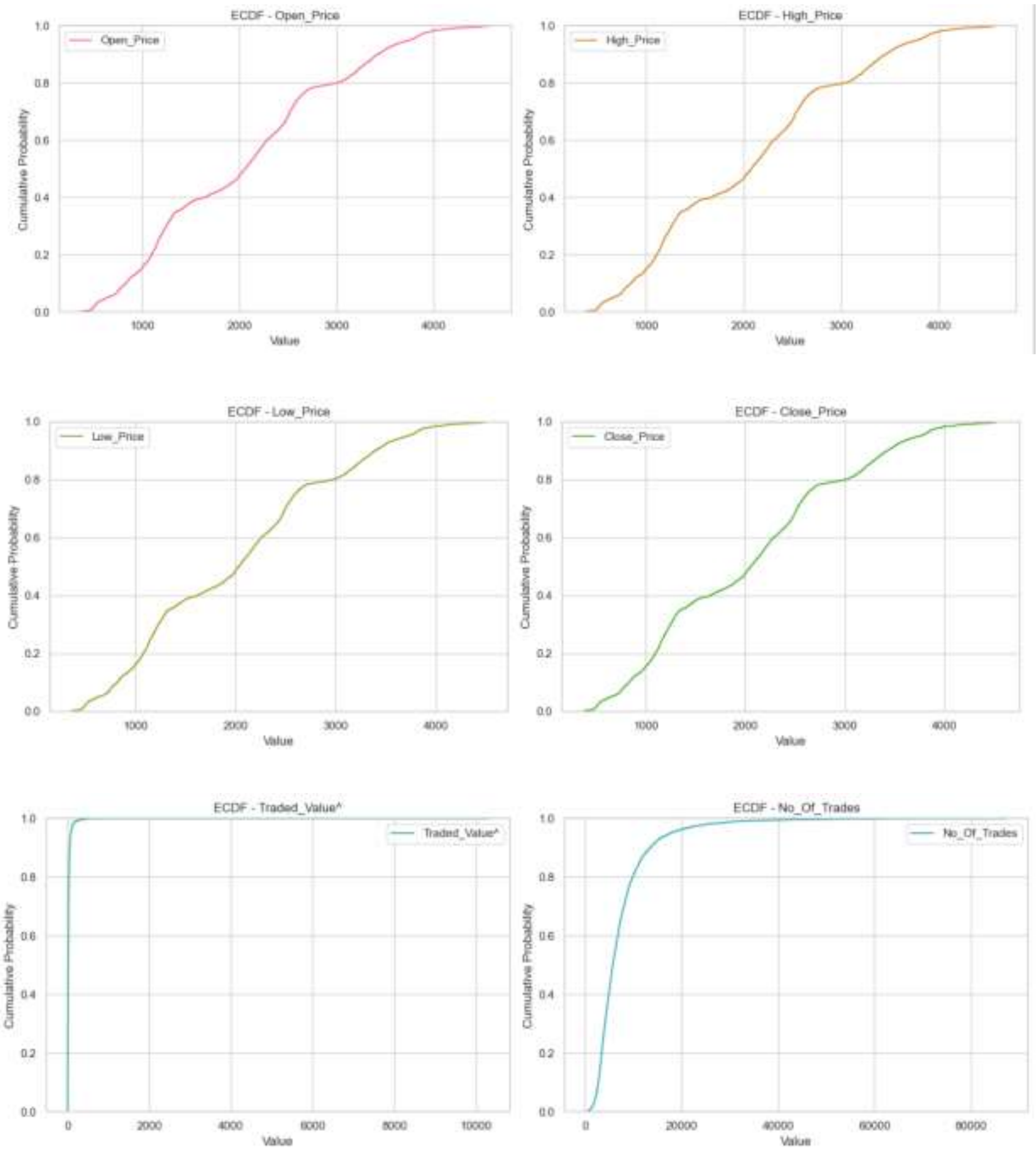


**Fig 6.16**

This code creates side-by-side dot plots for each numerical column in the DataFrame. Each plot shows the distribution of values along the index with different colors for each column. The plots are arranged in a grid with two plots per row, dynamically adjusting for the number of numerical columns. It helps visualize trends, outliers, and patterns in the data.

**ECDF plot for Open price, High price, Low price, Close price, Traded value, Traded quantity, No. of trades, 7-day SMA and 30-day SMA**



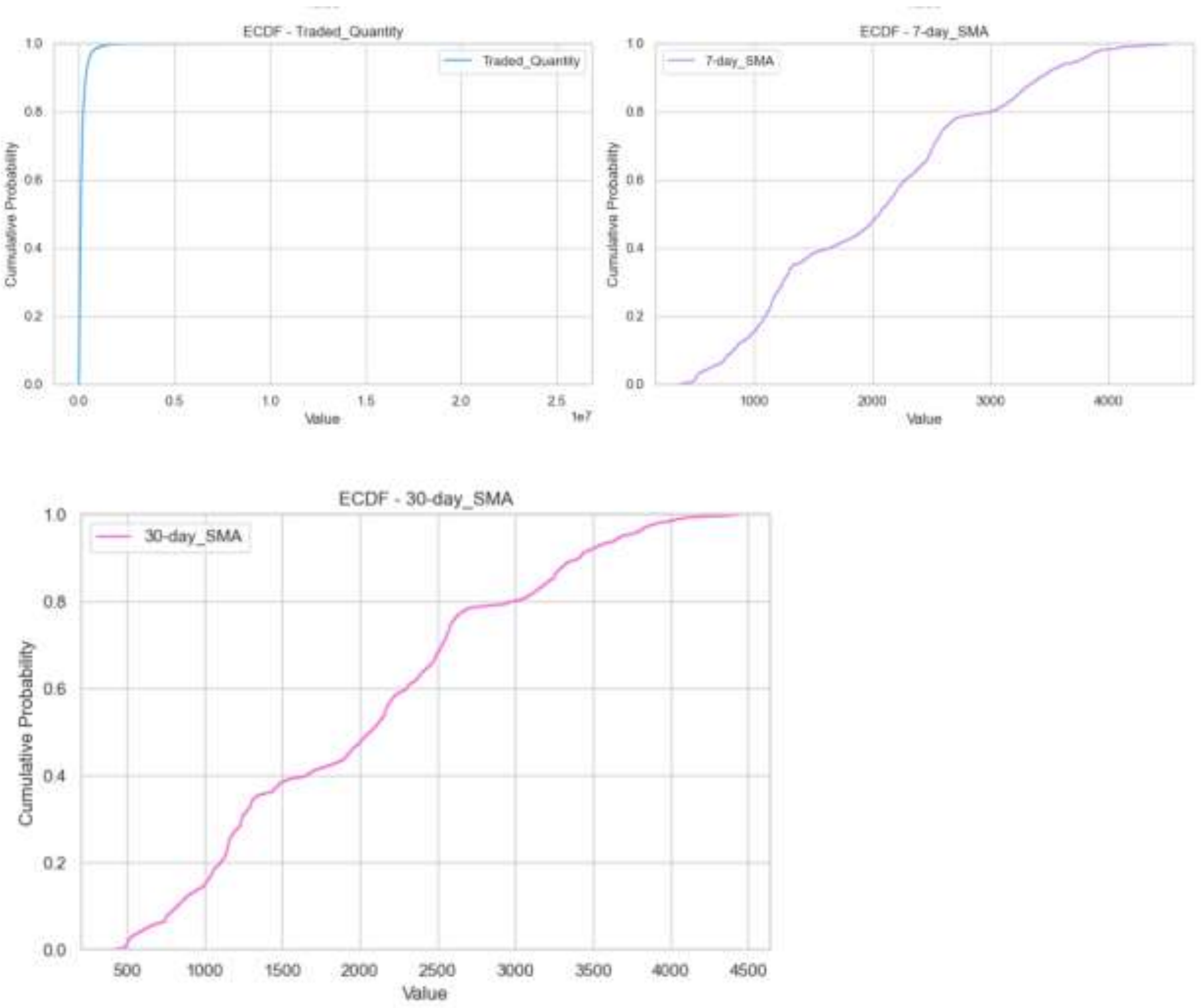


**Fig 6.17**

This code generates ECDF plots for each numerical column in the DataFrame, showing the cumulative distribution of data values. The ECDF helps visualize the proportion of observations below or equal to a given value, allowing comparison of distributions across multiple columns in an organized layout.

**Open Price vs Close price using Scatter plot**

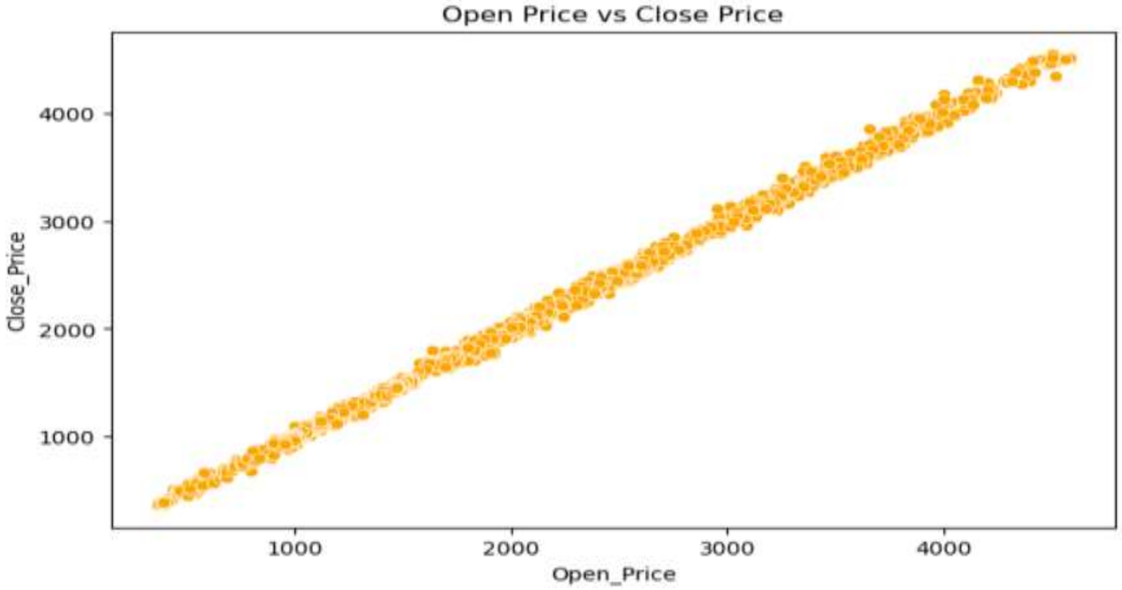


**Fig 6.18**

This scatter plot visualizes the relationship between the Open_Price and Close_Price of TCS stock. Each point represents a pair of values for Open_Price and Close_Price on the same day. The plot helps identify any correlation or pattern between the opening and closing prices, such as whether they tend to move in similar directions or show significant deviations.

**High price vs Low price using Scatter plot**

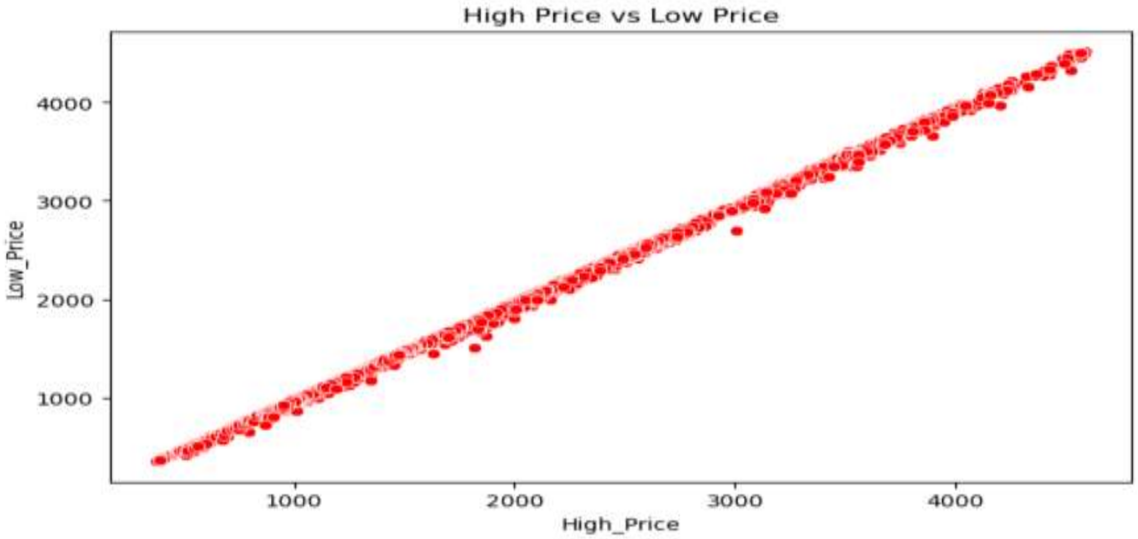


**Fig 6.19**

This scatter plot shows the relationship between High_Price and Low_Price of TCS stock. Each point represents the high and low prices on a given day. The plot helps identify if there's any correlation between the highest and lowest prices for the day, providing insight into daily price volatility and market behavior.

### Traded value vs Traded quantity using Scatter plot

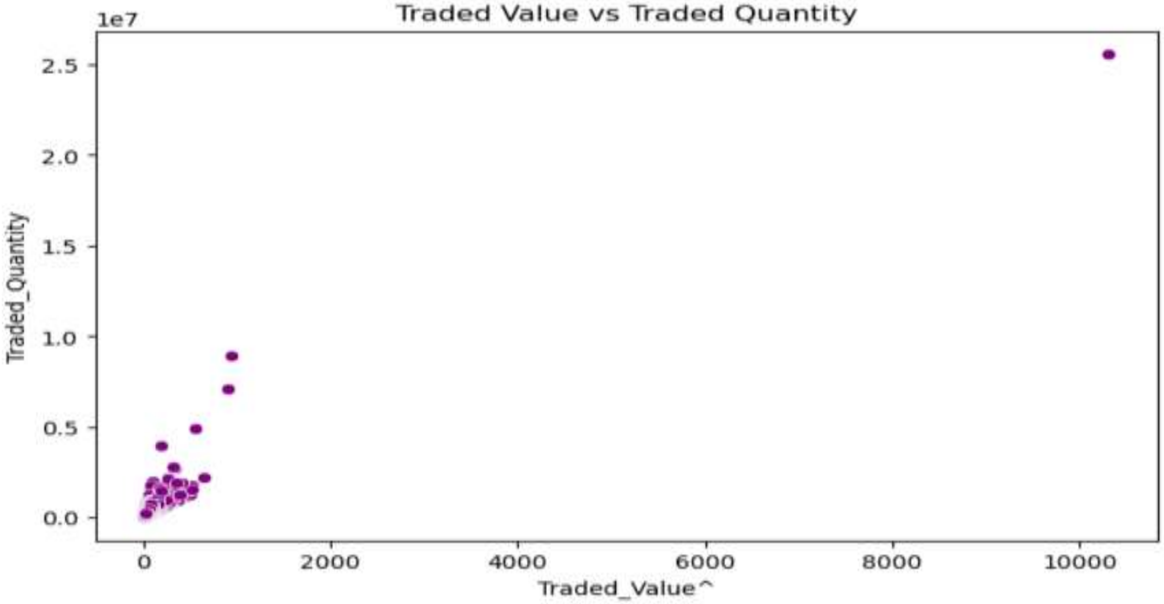


**Fig 6.20**

This scatter plot visualizes the relationship between Traded_Value and Traded_Quantity. Each point represents the total value and volume of shares traded on a given day. The plot helps identify patterns between the amount traded (Traded_Value) and the number of shares traded (Traded_Quantity), revealing if higher traded quantities generally correspond to higher traded values, which could indicate strong trading activity or high market interest.

**Close price by Traded quantity category using Box-plot**

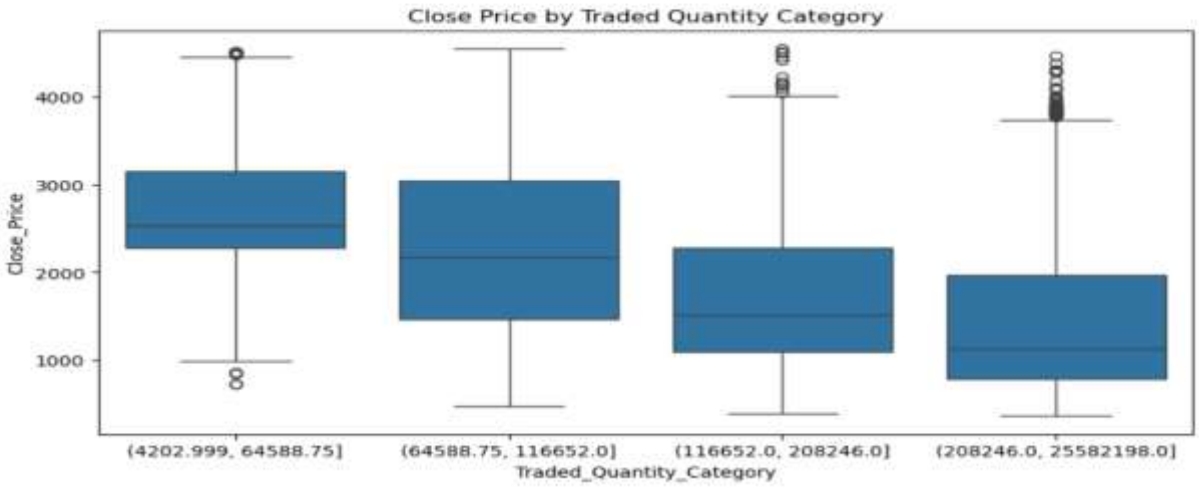


**Fig 6.21**

This box plot displays the Close_Price distribution across different categories of Traded_Quantity, which are divided into four quantiles (categories of equal size). Each category groups data based on increasing levels of Traded_Quantity. The plot helps compare the range, median, and variability of Close_Price across different trading volume categories, revealing how the closing price might differ based on trading activity levels. It also highlights any potential outliers in each category.

**No. of trades by Traded quantity Category using Box-plot**

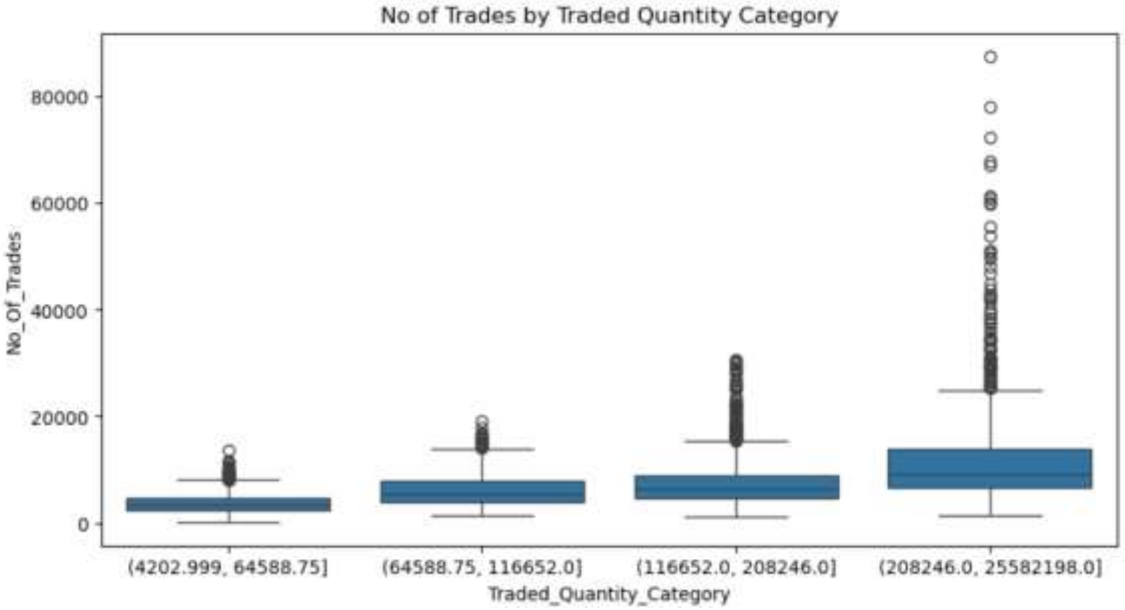


Fig 6.22

The box plot illustrates the distribution of No_Of_Trades across the three Traded_Quantity_Category levels (Low, Medium, High). It highlights the median, interquartile range, and potential outliers for each category. By comparing the categories, you can observe variations in the number of trades based on traded quantity, with the plot showing how concentrated or spread out the trades are within each category. This helps identify trends or disparities in trading activity across different quantities.

### Close Price vs Traded Value using Scatter plot

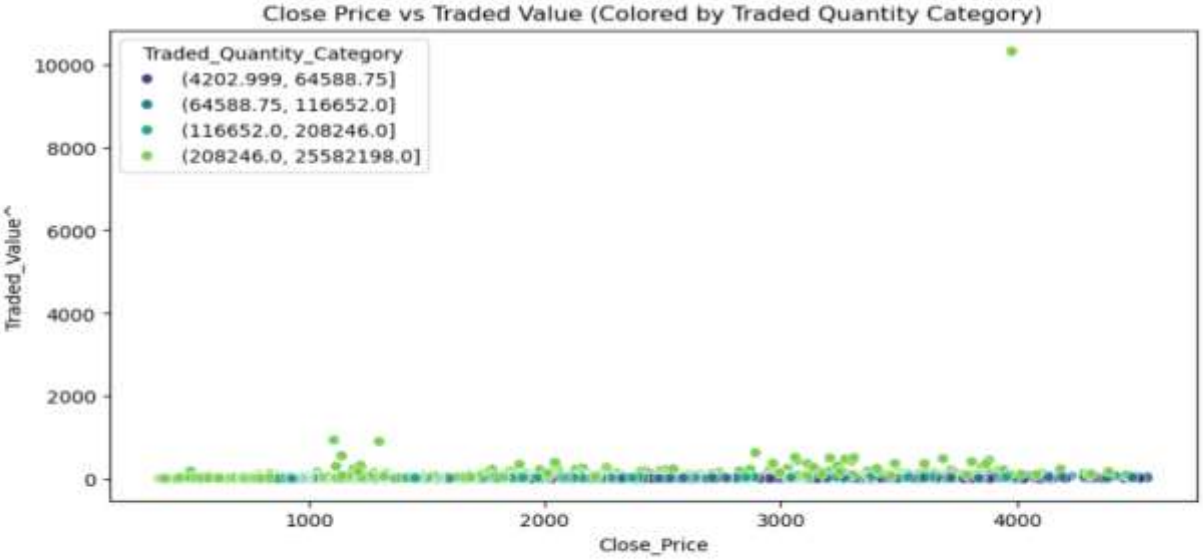


Fig 6.23

The scatter plot visualizes the relationship between Close_Price and Traded_Value, with points colored according to the Traded_Quantity_Category (Low, Medium, High). This coloring helps to identify how different levels of traded quantity influence the relationship between close price and traded value. The plot provides insights into any clustering or patterns in the data, and the use of the "viridis" color palette makes distinctions between categories clear.

**Density of High price vs Low price using Hexbin plot**

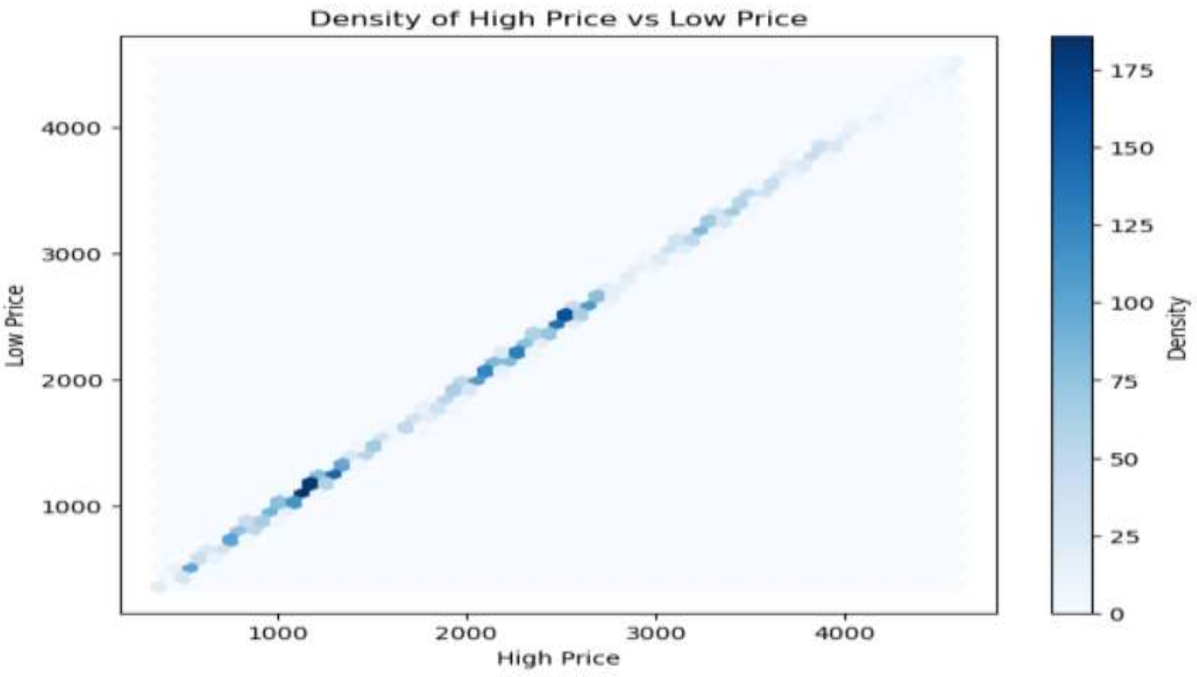


**Fig 6.24**

The hexbin plot visualizes the density of data points between High_Price and Low_Price. It uses hexagonal bins to show areas with varying point density, where darker blue colors indicate higher density. The plot helps identify patterns or clusters in the relationship between the two price variables. The color bar represents the density of points in each hexagonal bin, with higher values suggesting more concentrated data in that region. This type of plot is useful for revealing correlations and trends in large datasets, especially when individual data points may overlap.

### Open price and High price using Joint plot

```
<Figure size 800x800 with 0 Axes>
```

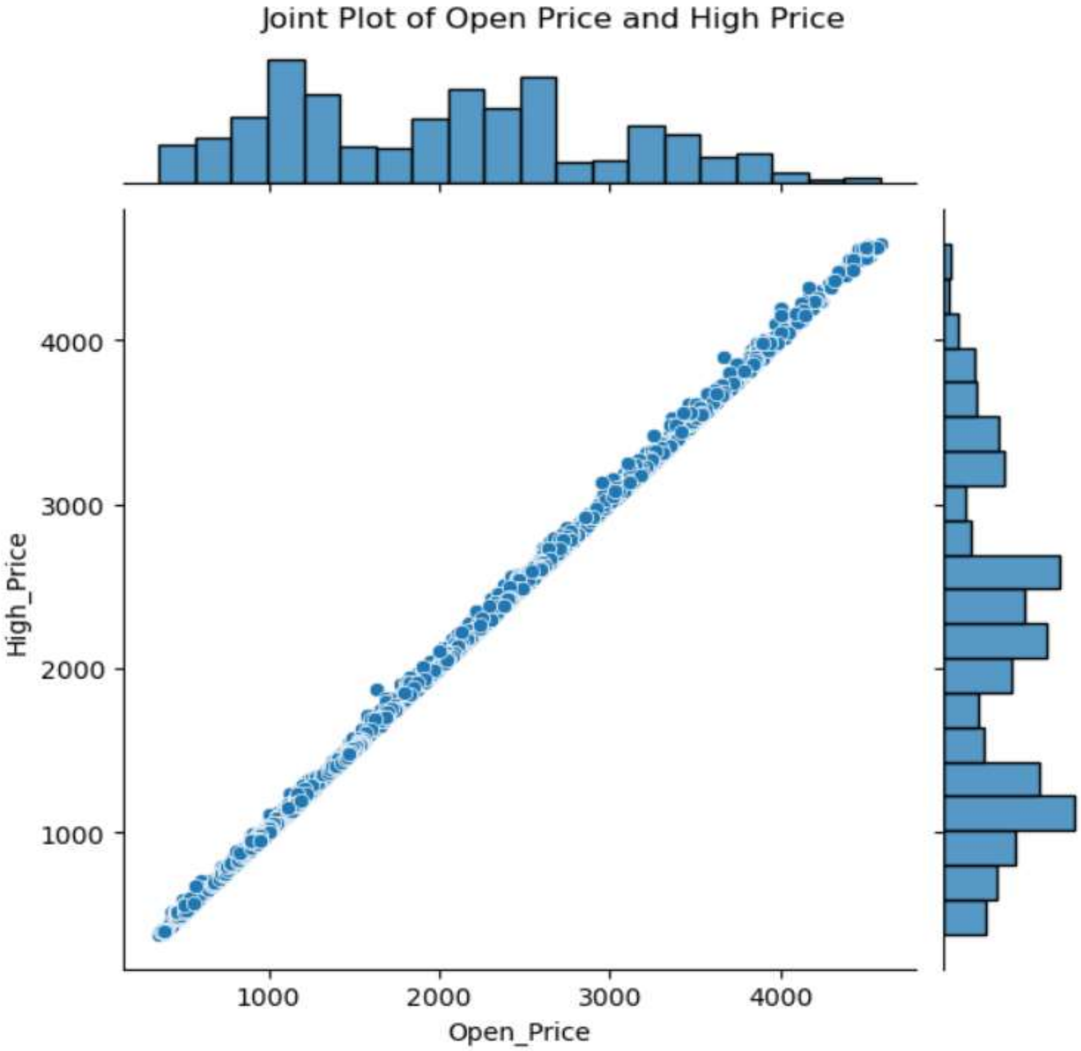


**Fig 6.25**

The joint plot visualizes the relationship between Open_Price and High_Price using a scatter plot with histograms on the axes. The scatter plot shows the correlation between these two variables, while the histograms display the distribution of each variable individually. The plot suggests how closely Open_Price is related to High_Price and helps identify any trends or patterns, such as linearity or clustering. The marginal histograms also provide insight into the frequency distribution of both prices.

**Close price by Traded Quantity Category using Violin plot**

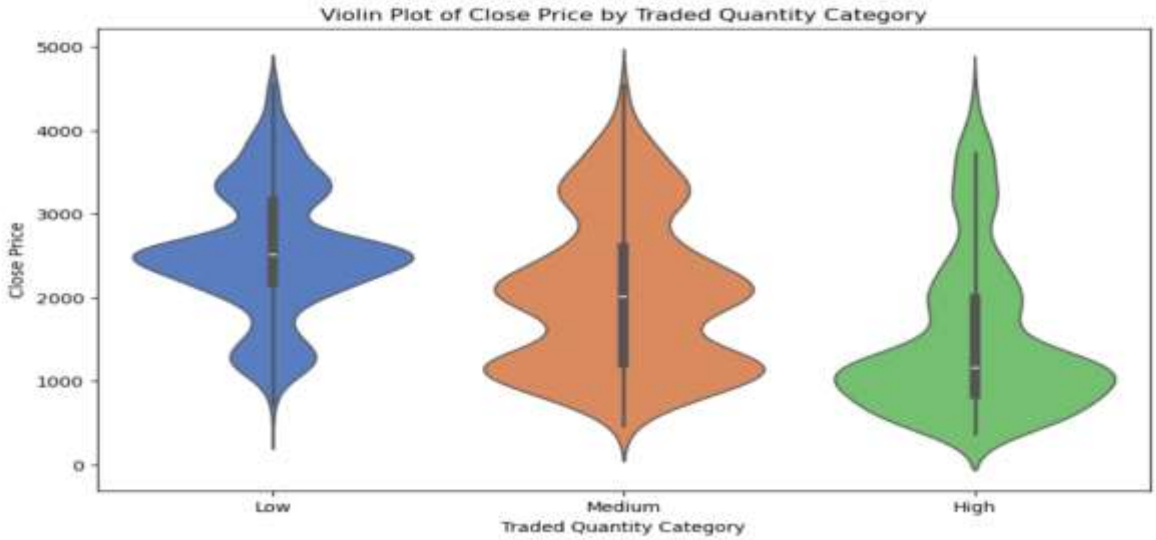


**Fig 6.26**

This code categorizes the Traded_Quantity column of a DataFrame df into three quantile-based categories: 'Low', 'Medium', and 'High'. It then creates a violin plot to visualize the distribution of Close_Price for each of these categories. The plot uses a muted color palette, and the axes are labeled accordingly. Warnings are suppressed to avoid unnecessary output during execution.

**Regression plot of Traded value vs Number of Trades**

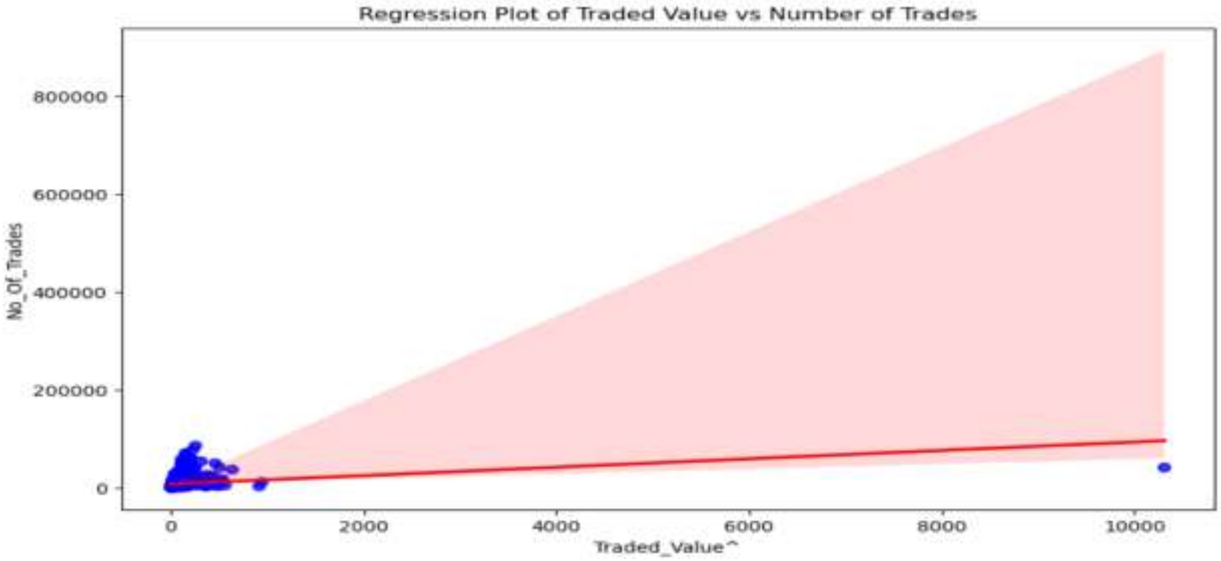
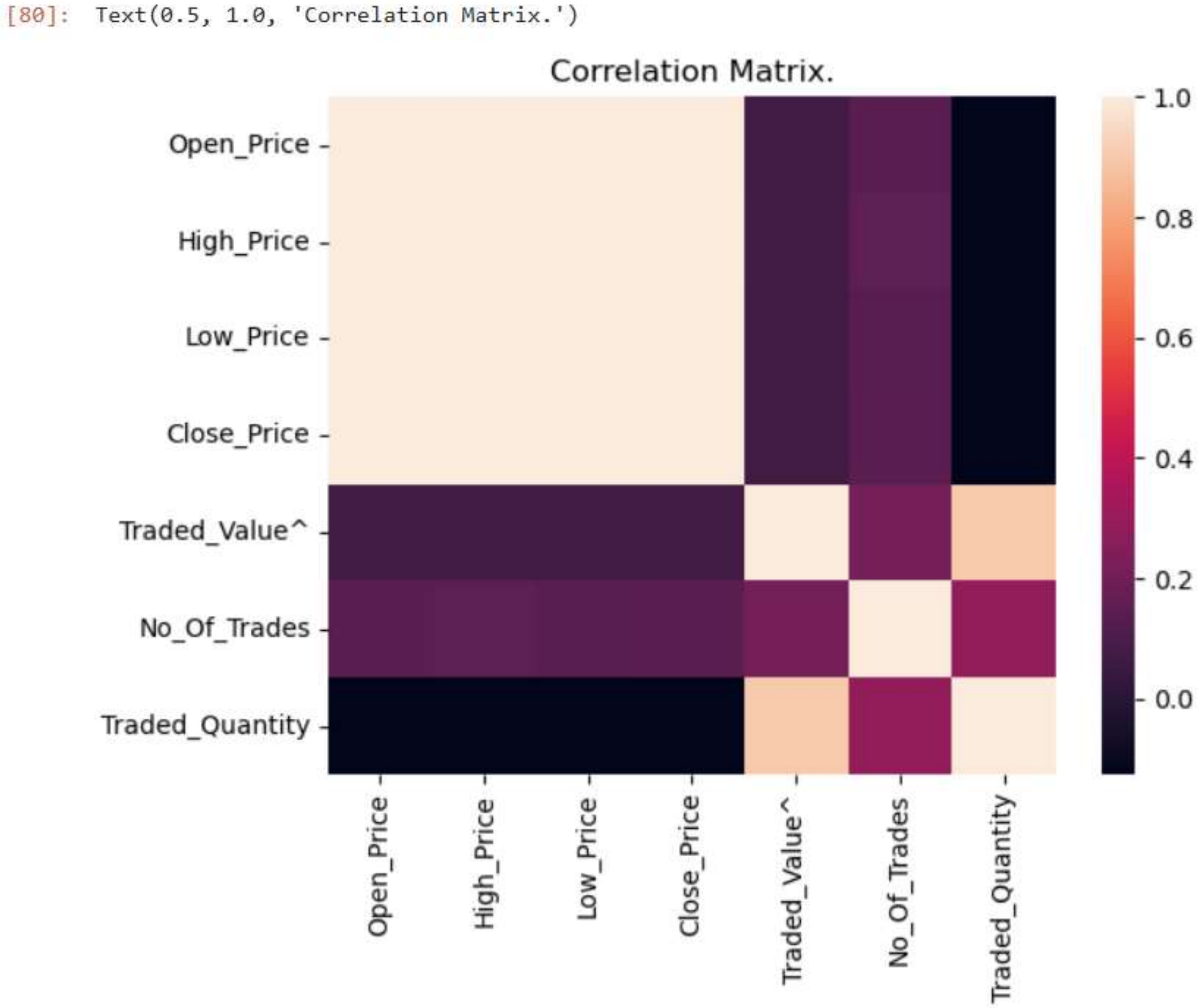


**Fig 6.27**

The regression plot of "Traded Value vs. Number of Trades" shows the relationship between the two variables with a scatter plot and a line of best fit. The red line represents the linear trend, indicating whether a positive or negative correlation exists, while blue scatter points illustrate individual data values around this trend. This plot helps assess how well "Traded Value" predicts "Number of Trades."

### Correlation Matrix Heatmap



```
[80]:  Text(0.5, 1.0, 'Correlation Matrix.')
```

**Fig 6.28**

The correlation matrix heatmap visualizes the strength and direction of relationships between numerical variables in the dataset:

1. Correlation Values: Each cell in the matrix shows the correlation coefficient (ranging from -1 to 1) between two variables. A positive value indicates a positive correlation, where increases in one variable are associated with increases in the other, while a negative value suggests an inverse relationship.
2. Interpreting Strength: Values close to 1 or -1 denote strong correlations, while values near 0 indicate weak or no correlation. In the heatmap, darker colors (often **closer to red or** blue) show stronger correlations, allowing us to quickly identify variables with significant relationships.

**Actual vs Predicted TCS Stock Price using Bar Plot**

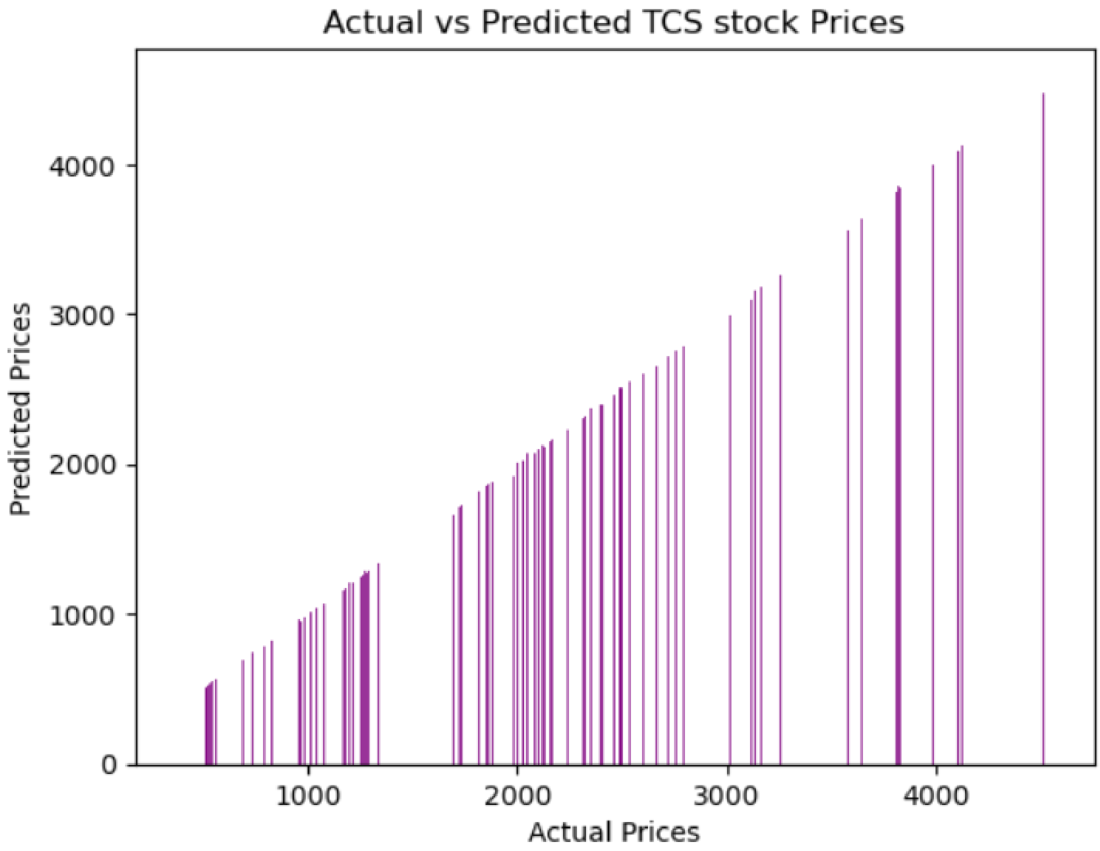


**Fig 6.29**

The bar plot of "Actual vs. Predicted TCS Stock Prices" compares each test sample's actual and predicted prices. Here's how to interpret it:

1. Bar Heights: Each pair of bars (one for the actual price and one for the predicted price) represents a specific test instance. Close alignment between the heights of each pair indicates accurate predictions, while large gaps suggest prediction errors.
2. Model Accuracy: If the purple bars (predicted prices) are consistently close in height to the actual prices, it implies that the model has good predictive power. However, noticeable differences reveal where the model might struggle to capture certain fluctuations in the stock price accurately.

**Actual vs Predicted TCS stock price using Scatter plot**

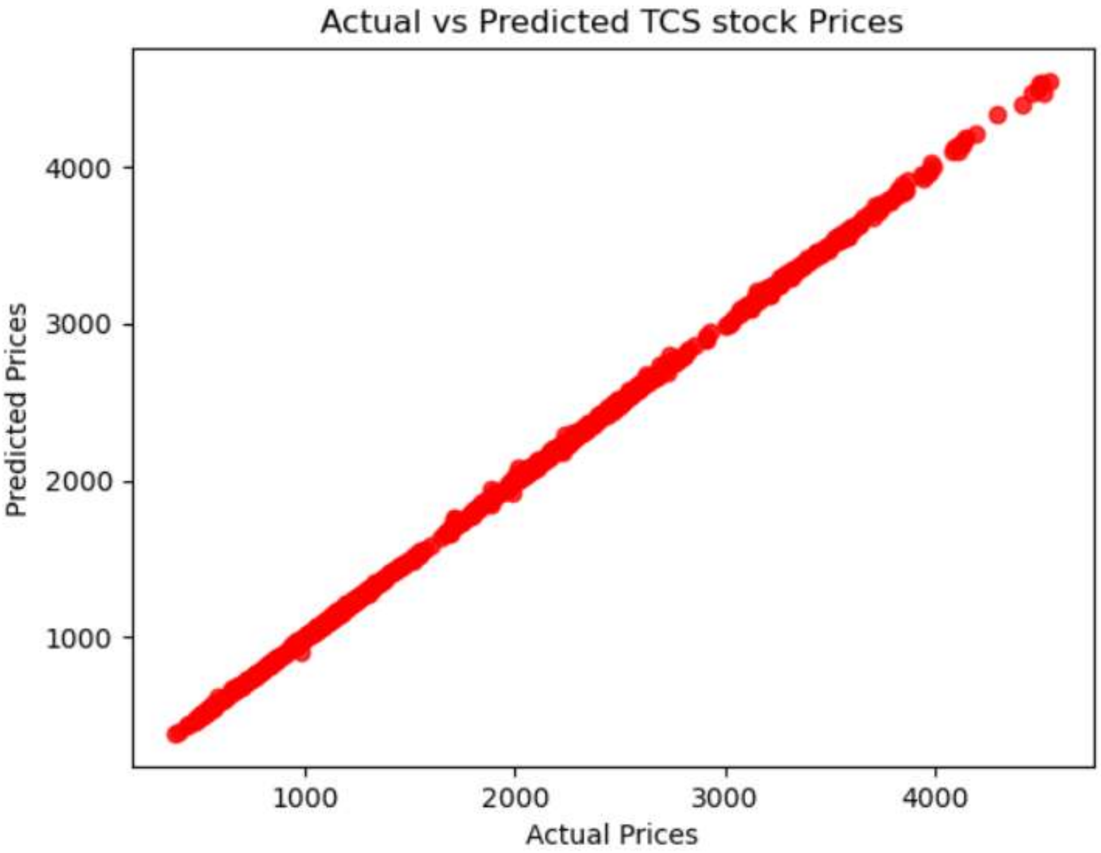
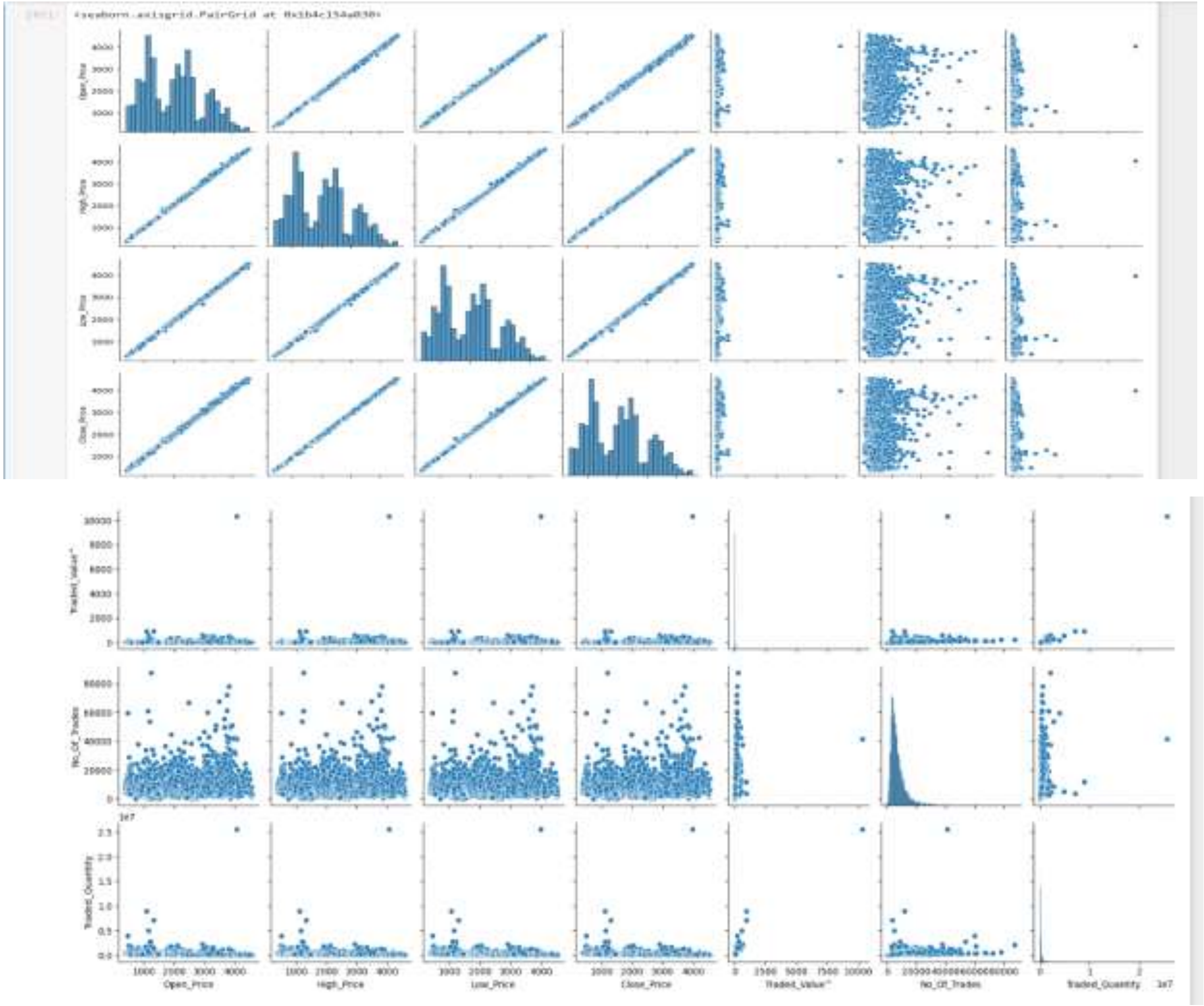


**Fig 6.30**

The scatter plot of "Actual vs. Predicted TCS Stock Prices" helps visualize the relationship between actual and predicted stock prices. Here’s how to interpret it:

1. Data Alignment: Each point represents an individual test sample, with the actual price on the x-axis and the predicted price on the y-axis. If the model predictions were perfect, all points would fall along a 45-degree line from the origin (where actual equals predicted). Points near this line indicate accurate predictions, while points farther away show larger prediction errors.
2. Model Performance: The closer the points are to the line, the better the model’s accuracy. Significant deviations indicate areas where the model may have underperformed, suggesting either outliers or patterns that the model did not fully capture.

## Pairplot



```
<seaborn.axisgrid.PairGrid at 0x1b4c154a030>
```

**Fig 6.31**

The pairplot reveals correlations and relationships between different features of the dataset. Strong positive correlations are visible between related price variables (Open, High, Low, Close), while traded quantity and value have more dispersed relationships with price variables, indicating varying levels of trading activity

**Distribution of Close Price**

```
[54]:  <AxesSubplot:xlabel='Close_Price', ylabel='Density'>
```

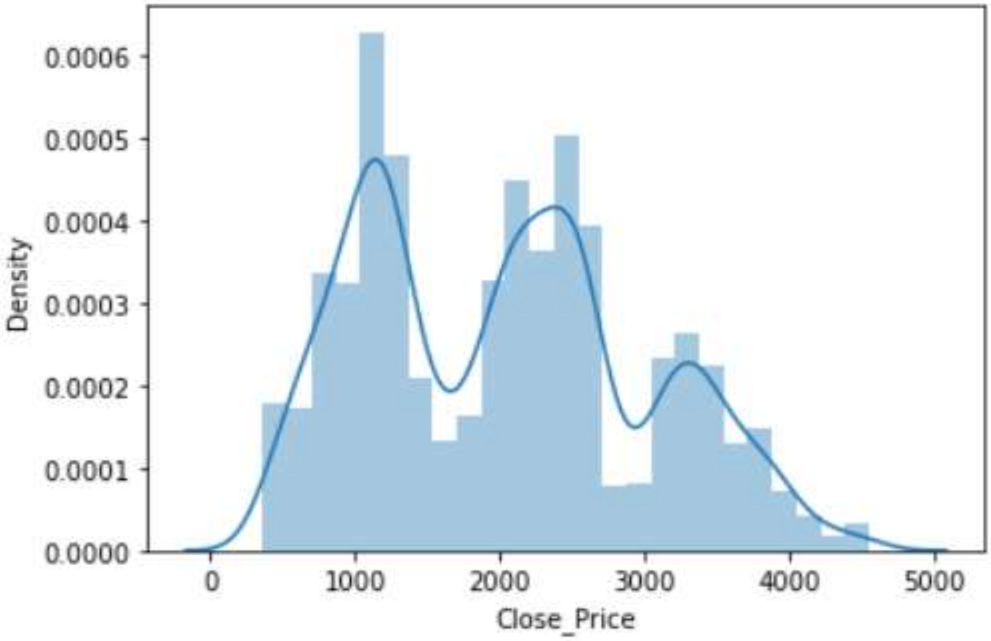


**Fig 6.32**

The distribution of the 'Close Price' appears approximately normal with a slight skew, suggesting some variability in stock closing prices. The presence of a smooth peak in the KDE curve indicates a central tendency with fewer occurrences at extreme price values.

# Chapter 7: Conclusion and Future Work

This project aimed to develop predictive models for the stock price of Tata Consultancy Services (TCS) using historical data and machine learning techniques. Through a systematic approach that included data preprocessing, exploratory data analysis, feature selection, and model building, we were able to gain valuable insights and achieve reliable predictions. The analysis leveraged two machine learning models — **Linear Regression** for predicting the closing stock price and **Logistic Regression** for classifying the price movement direction. The outcomes indicate promising results, demonstrating the effectiveness of these models in the context of stock market analysis.

## 7.1 Key Findings:

1. **Successful Prediction of Stock Prices**:
    - The **Linear Regression model** effectively predicted the closing stock price with a low RMSE (Root Mean Squared Error) and a high R-squared value. This suggests that the selected features, including Open_Price, High_Price, Low_Price, and Traded_Quantity, have a strong linear relationship with the Close_Price.
    - The predictions closely followed the actual price trends, showing that the model was able to capture significant patterns in the data.
2. **Reliable Classification of Price Movements**:
    - The **Logistic Regression model** showed good accuracy in classifying whether the closing price would be higher or lower than the opening price. This model can help investors make informed decisions about buying or selling stock based on the predicted price direction.
    - The classification model's performance metrics, including precision, recall, and F1 score, were also satisfactory, indicating a balanced and robust model with minimal false predictions.
3. **Impact of Feature Selection**:
    - The choice of features had a significant impact on the model’s performance. Features like Open_Price, High_Price, and Low_Price were found to be highly correlated with the target variable (Close_Price), highlighting their importance in stock price prediction.

    - The dataset's quality and completeness contributed to the reliability of the models, as missing values and anomalies were effectively handled during preprocessing.

4. **Visualization and Analysis**:
    - Visual representations of the stock trends, including line charts and scatter plots, provided clear insights into the stock's behavior over time. This helped in understanding the underlying patterns and identifying any potential anomalies.
    - The comparative analysis of the predicted versus actual values demonstrated the models' ability to generalize well on unseen data.

### 7.2 Challenges Faced:

- The stock market is inherently volatile, and predicting stock prices remains a challenging task due to sudden market changes influenced by external factors such as political events, economic announcements, and global crises.
- The **Linear Regression model** assumes a linear relationship between the features and the target variable, which may not always hold true in financial data that often exhibits non-linear trends.
- The **Logistic Regression model** was limited in its ability to account for complex dependencies and interactions between features, which could be addressed by using more advanced machine learning algorithms in future work.

### 7.3 Overall Conclusion:

The project demonstrated that machine learning techniques can be effectively applied to stock price prediction using historical data. The models developed provided reliable and interpretable results, which can serve as a valuable tool for investors and financial analysts. By focusing on key predictive features and using robust preprocessing methods, the analysis achieved meaningful insights into TCS stock behavior, supporting data-driven decision-making.

While the results are promising, the limitations of the models highlight the need for more sophisticated approaches to handle the non-linear and volatile nature of stock market data. Future research could incorporate additional data sources, such as market sentiment and macroeconomic indicators, to enhance prediction accuracy. Moreover, exploring advanced machine learning models like **Random Forest, XGBoost, or LSTM neural networks** could further improve performance and provide a deeper understanding of stock price dynamics.

In summary, this project provides a strong foundation for using machine learning in financial market analysis, offering a practical approach to predicting stock prices and making informed investment

## Future Work

- Implementation of more advanced models like Random Forest, XGBoost, or LSTM.
- Incorporation of sentiment analysis based on news and social media data.
- Development of a real-time prediction system using live data feeds.

**References:**

Analytics Vidhya:

Stock Price Analysis With Python - Analytics Vidhya

GeeksforGeeks:

Stock Price Analysis With Python - GeeksforGeeks

Github:

AishikDasgupta/Stock-Market-Analysis: Explore Stock Market Analysis, a Python project using NumPy, Pandas, and Matplotlib. Uncover trends, visualize prices, and make informed decisions. Join the world of finance!

**GitHub Repository Of  Mini-Project:**

**nilesh171/mini_project_python**