KLE Society's

KLE Technological University



**Exploratory Data Analysis**

**(22ECAC210)**

**Course Project Report on**

**INDIAN STOCK MARKET**

*Submitted in partial fulfilment of the requirement for the award of*

**Degree of Bachelor of Engineering**

**in**

**Computer Science and Engineering**

*Submitted By*

**Name SRN**

**Shivan Singh 01FE21BCS246**

**Somil Yadav**    **01FE21BCS192**

**Rajat Singh Jakhar**   **01FE21BCS129**

**Prasad Bharatesh Palled 01FE21BCS123**

*Submitted To*

**SCHOOL OF COMPUTER SCIENCE & ENGINEERING,**

KLE TECHNOLOGICAL UNIVERSITY

HUBLI – 580 031 (India).

**Abstract**

The purpose of this EDA project is to analyse the Indian stock market through historical data. By exploring stock prices, trading volumes, market indices, and sector performances, valuable insights into market behaviour and trends can be gained. This analysis aims to identify patterns, correlations, and indicators to assist investors, analysts, and researchers in making informed decisions. Through data visualization, outlier detection, distribution examination, and variable relationship exploration, a deeper understanding of the Indian stock market will be achieved, aiding stakeholders in formulating investment strategies and predicting market movements.

**Contents**

**Table of Content:** **Page No.**

1. Introduction 4
   1. Overview
   2. Importance of EDA in Data Analysis
   3. Objectives
2. Data Collection 5-7
   1. Data Preprocessing
3. Data Exploration 8-11
   1. Attributes
   2. Visualisation Stock Market Prices
   3. Visualisation Stock Market Prices of a day
   4. Visualisation of Moving Average of the companies
   5. Correlation Heatmap of the Companies
4. Data Cleaning 11-13
   1. Checking NULL Values
   2. Replacing NULL Values
   3. Data Transformation
   4. Outliers
5. Feature Selection 14
   1. Process of selecting relevant features
   2. Data Transformation
   3. The rationale behind the feature selection process
6. Data Analysis 15-20
   1. Model Selection
   2. Model Result and Analysis
   3. Model Prediction Graph Analysis
7. Insights and Findings 21
   1. Key findings
   2. Significant findings or correlations.
   3. Limitation or challenges encountered
8. Recommendations 22
   1. Recommendations based on the findings.
   2. Potential areas for further analysis or investigation.
   3. Potential impact of implementing the recommendations
9. Conclusions 23
   1. Highlights
   2. Value of EDA in understanding the dataset.

References

**1. Introduction**

**1.1 Overview**

This EDA project aims to analyse the Indian stock market through historical data, including stock prices, trading volumes, market indices, and sector performances. By utilizing data visualization, outlier detection, distribution examination, and variable relationship exploration, valuable insights into market behaviour and trends will be uncovered. The goal is to identify patterns, correlations, and indicators that can assist investors, analysts, and researchers in making informed decisions and formulating investment strategies. Ultimately, this analysis seeks to provide a deeper understanding of the Indian stock market, aiding stakeholders in predicting market movements and optimizing their investment approaches.

**1.2 Importance of Exploratory Data Analysis in Data Analysis**

Exploratory Data Analysis (EDA) is crucial in data analysis as it serves as a foundational step to understand, visualize, and summarize the main characteristics of the dataset. By utilizing EDA techniques, such as data visualization, distribution examination, and outlier detection, analysts can gain valuable insights, identify patterns, and understand the relationships between variables. EDA helps uncover hidden trends, potential correlations, and data quality issues, allowing for more informed decision-making and the formulation of appropriate analytical approaches. It acts as a powerful tool to guide further analysis, hypothesis testing, and model building, making it an essential step in any data analysis process.

**1.3 Objectives**

Objective 1:

* Understand dataset structure, variables, and data quality.
* Gain insights into the dataset's composition, key variables, and identify any missing or erroneous values.

Objective 2:

* Study the distribution of stocks value of the company, identify trends over time, and explore relationships with other variables.
* Analyse the company’s performance metrics in stock markets.

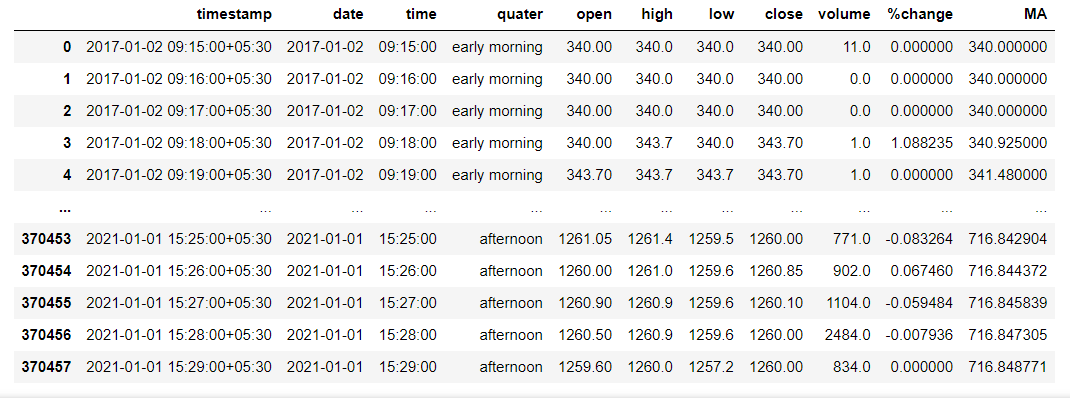
Objective 3:

* Visualize the company’s performance and compare them to each other.
* Predicting the best company to invest stocks.

**2. Data Collection**

* Data set which is being used in this project is Stock Market Data sets of Indian companies.
* The collection of datasets includes the data of 159 Indian Companies of size approximately 6.67 GB from the year 1st January 2017 to 1st January 2021 which comprises of stock market data for every minute.
* Data set is taken from Kaggle. Kaggle is a popular platform that hosts a wide range of datasets related to various topics, including the Indian stock market. These datasets can provide valuable insights for researchers, analysts, and investors interested in studying the dynamics of the Indian stock market.

Stock Data of Company AARTIIND

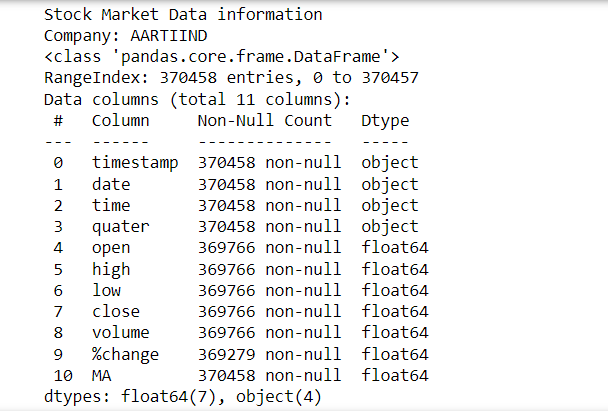


370458 rows × 11 columns

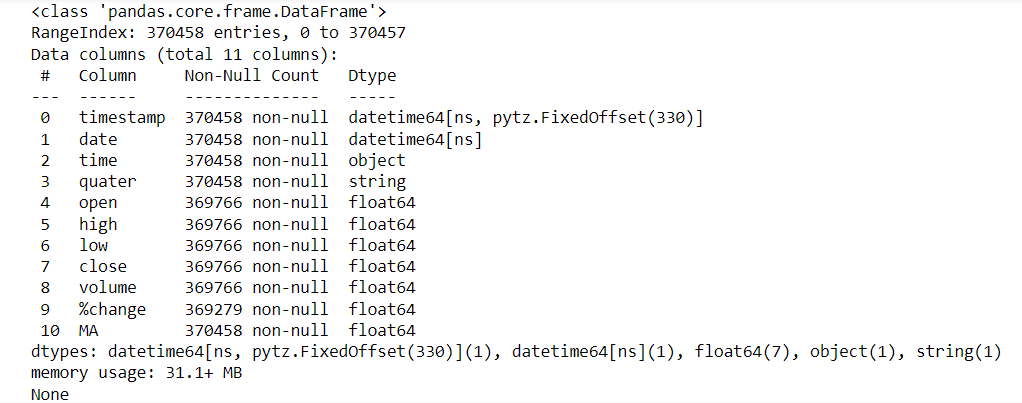
Table 1

Table 1: This Table shows Stock market dataset of the company AARTIIND with 11 Columns and 370458 rows

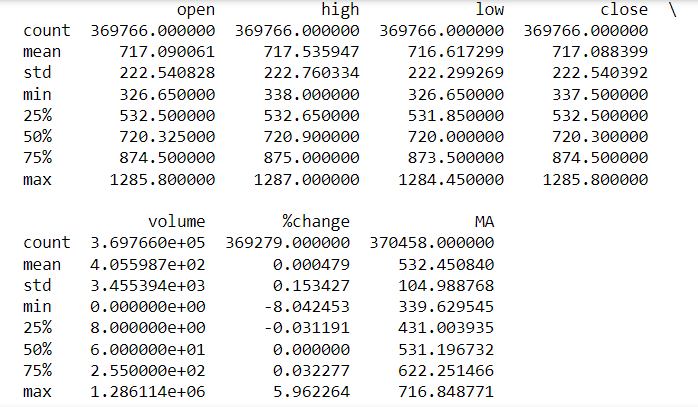
Stock Market Data Information for Company AARTIIND



Stock Market Data Information for Company ABCAPITAL



Stock Market Data Description for Company ABCAPITAL



**2.1 Data Preprocessing**

* The data preprocessing steps used in the analysis of the stock market data include data collection, data cleaning, data transformation, data exploration and visualization, and the prediction of stock market prices. These steps are essential to ensure that the data is reliable, consistent, and suitable for further analysis and modelling. By following a systematic preprocessing approach, valuable insights can be gained, and accurate predictions of stock market prices can be made.

**3. Data Exploration**

**3.1 Attributes**

|  |  |  |
| --- | --- | --- |
| At No. | Attribute Name | Attribute Description |
| 1 | Timestamp | Timestamp of the dataset, representing the exact time the data was recorded |
| 2 | Date | Date of the Stock Market Prices, indicating the specific trading date |
| 3 | Time | Time of the Stock Market, indicating the time within the trading day when the data was recorded |
| 4 | Quarter | Periods of the day (morning, afternoon, evening) during which the data was recorded |
| 5 | Open | The stock market price at which trading opened on the specified date |
| 6 | Close | The stock market price at which trading closed on the specified date |
| 7 | High | The highest stock market price recorded during the trading session |
| 8 | Low | The lowest stock market price recorded during the trading session |
| 9 | Volume | The total number of shares traded during the specific time frame |
| 10 | %Change | The percentage change in stock market prices from the previous trading day’s closing price |
| 11 | MA (Moving Average) | A technical indicator that calculates the average price of the stock over a specified period |

Table 2

Table 2: Table shows Attributes in the date and there description

**3.2 Visualisation Stock Market Prices**

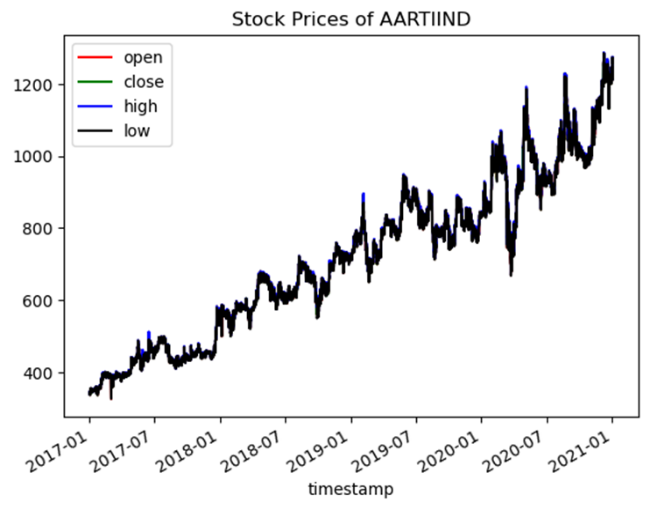
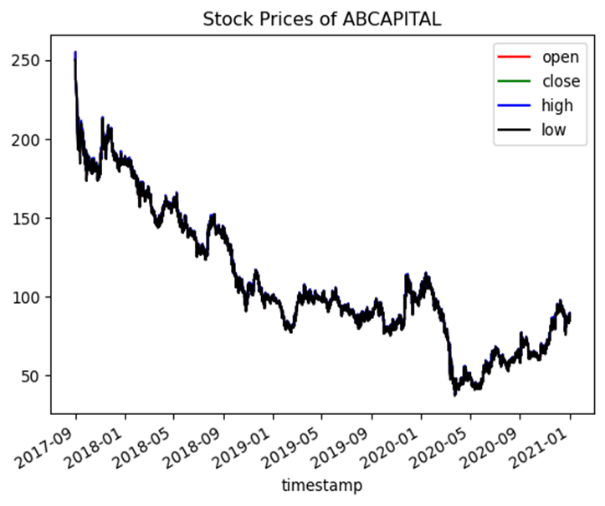
 

Figure 1a. Figure 1b.

Figure 1

Figure 1: Figure 1 shows graph of open, high, low, close with respect to date and time for all data.

Similarly, we created 157 more graphs of other companies with same process.

**3.3 Visualisation Stock Market Prices of a day**

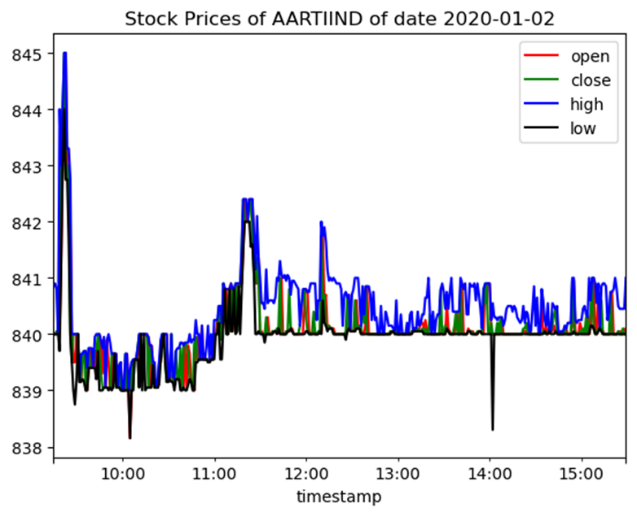
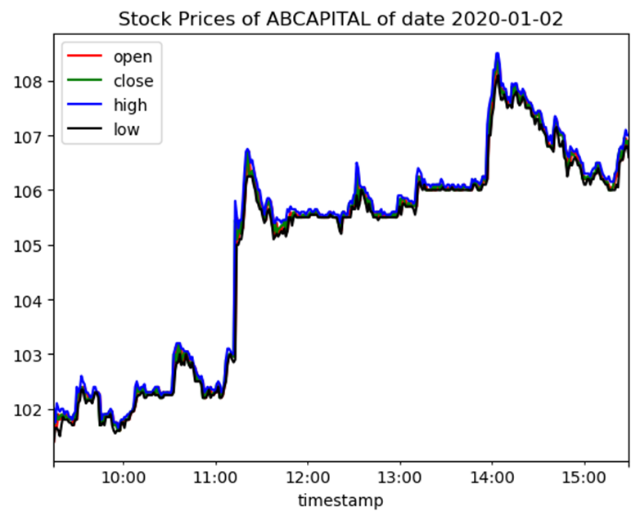
 

Figure 2a. Figure 2b.

Figure 2.

Figure 2: Figure 2 shows graph of open, high, low, close with respect to date 2020-01-02

Similarly, we created 157 more graphs of other companies with same process.

**3.4** **Visualisation of Moving Average of the company**

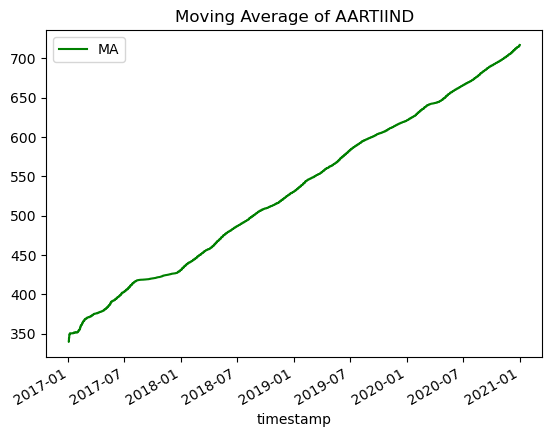
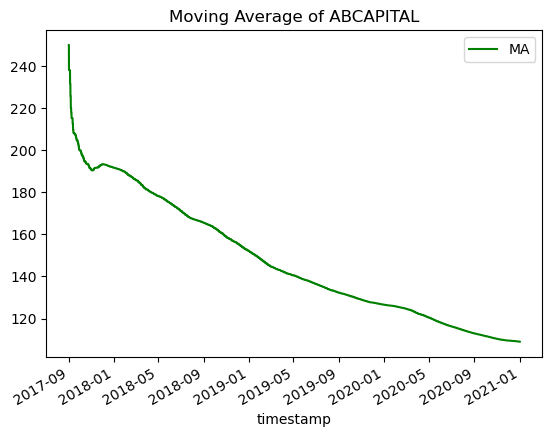
 

Figure 3a. Figure 3b.

Figure 3

Figure 3: Figure 3 shows graph of open, high, low, close with respect to date 2020-01-02

Similarly, we created 157 more graphs of other companies with same process.

**3.5** **Correlation Heatmap of the Company**

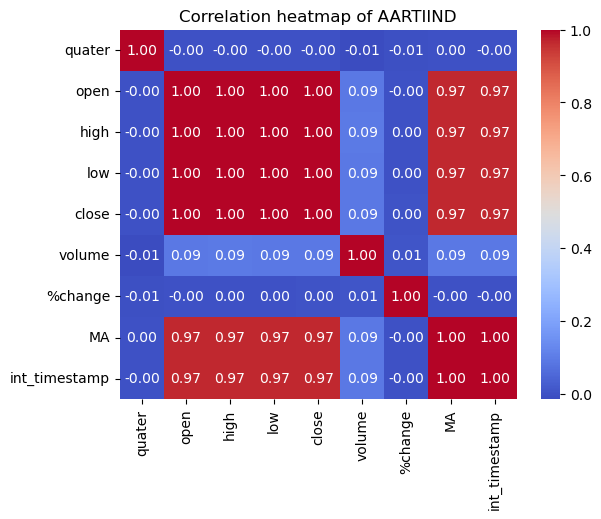


Figure 4a.

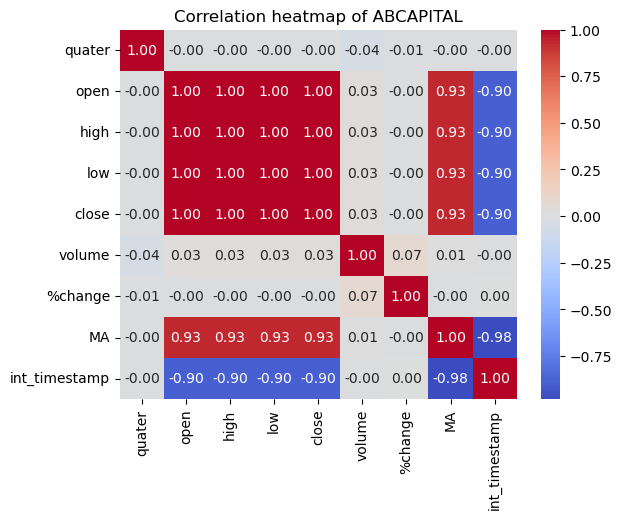


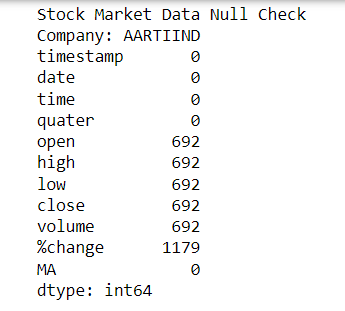
Figure 4b.

Figure 4: Figure 4 shows Heatmap of correlation between columns of different companies

Similarly, we created 157 more heatmaps of other companies with same process.

**4. Data Cleaning**

**4.1 Initial NULL check for Companies**



From Above Data, we can see data containing open, high, low, close, volume, %change NULL Values.

**NULL DATA OF COMPANY AARTIIND**

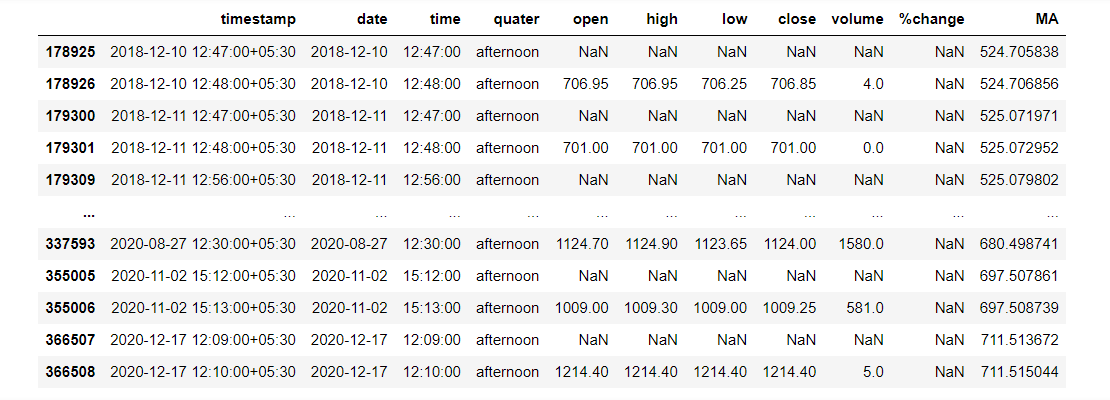


Table 3

Table 3: Table showing rows with containing NULL/NAN Values of columns AARTIIND

**4.2** **Replacing of NULL values**

* Replacing of NULL Values for attributes- open, close, high, low:

In stock market analysis, it’s not a good way to drop value or replace with mean in stock market for ex. If values like 20 NULL 20, then if mean is 40 it will make data inconsistent.

The best way to replace them is to replace them with previous non-NULL values.

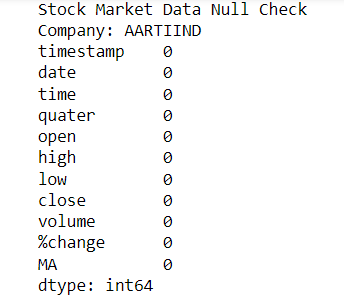
* Replacing NULL Values for attributes- volume, %change:

Volume: Assuming no one brought the stock for that minute. So, we are replacing values with 0.

%Change: We can recalculate them by formula given in stock market.

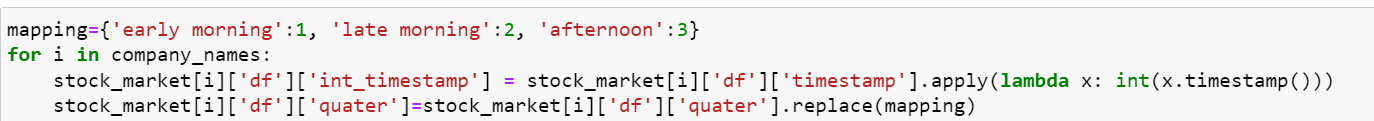
%Change=(curr\_close-prev\_close)/prev\_close\*100

**After Data Cleaning on Companies**



**4.3 Data Transformation**

In our data analysis, we performed data transformation to convert the timestamp data from datetime format to integer format. This conversion enables us to represent time-related information in a numerical form, which can be beneficial for certain types of analyses and modelling tasks. Also, we mapped Quarter to integer format so that it can be used in Models prediction



**4.4 Outliers**

In stock market data, inconsistencies are common, and every data point at any given moment can be valuable. Despite the presence of noise or outliers, it's crucial to handle them carefully without modifying the original data drastically, as it may adversely affect subsequent predictions and analyses. Proper data preprocessing techniques, such as outlier detection and handling, can be applied to preserve the integrity of the data while minimizing the impact on future predictions. Careful consideration of the data cleaning process is essential to ensure accurate and reliable results in any analysis or predictive modelling.

**5. Feature Selection**

**5.1** **Process of selecting relevant features**  
  
Correlation Analysis: We perform a correlation analysis using Pearson Correlation among the numerical attributes like stock price, trading volume. This analysis helps us identify which numerical attributes have a strong positive or negative correlation with the stock price, indicating their relevance for predicting stock movements.

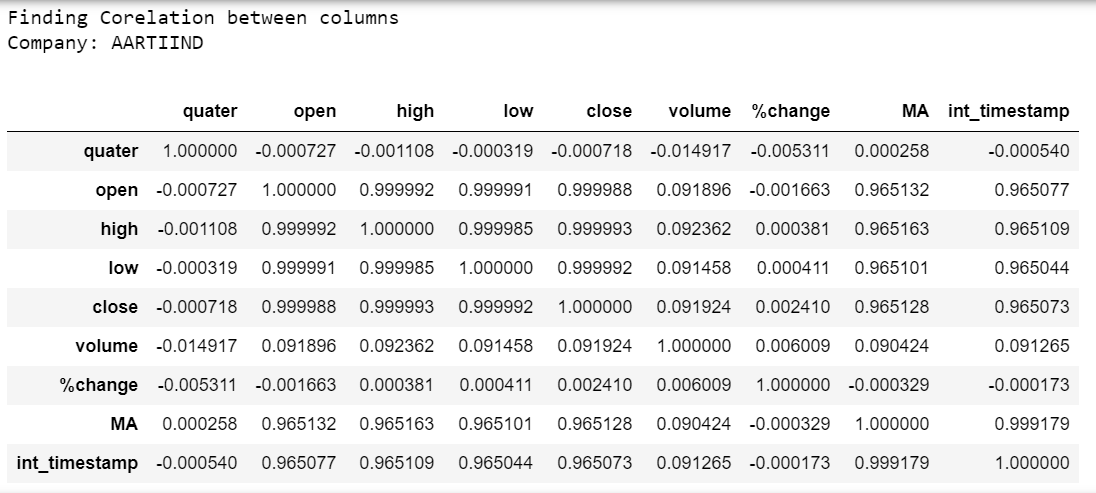


Table 4

Table 4: Table shows correlation values between the columns using Pearson coefficient

**5.2 Data Transformation**

* Conversion of Timestamp from datetime64 format to integer format
* Quarter has been Mapped from [‘early morning’, ‘late morning’, ‘afternoon’] to [1,2,3]

**5.3** **The rationale behind the feature selection process**

The rationale behind feature selection in stock market analysis and prediction revolves around improving model interpretability, generalization, stability, and performance. By selecting the most relevant features, the model becomes more focused, efficient, and better equipped to extract meaningful insights from complex market data, aiding investors and financial experts in making well-informed decisions.

**6. Data Analysis**

**6.1 Model Selection**

Based on real time scenario, as person would want open, high, low, close for a particular date and time. So, we will use Linear Regression to predict open, high low, close

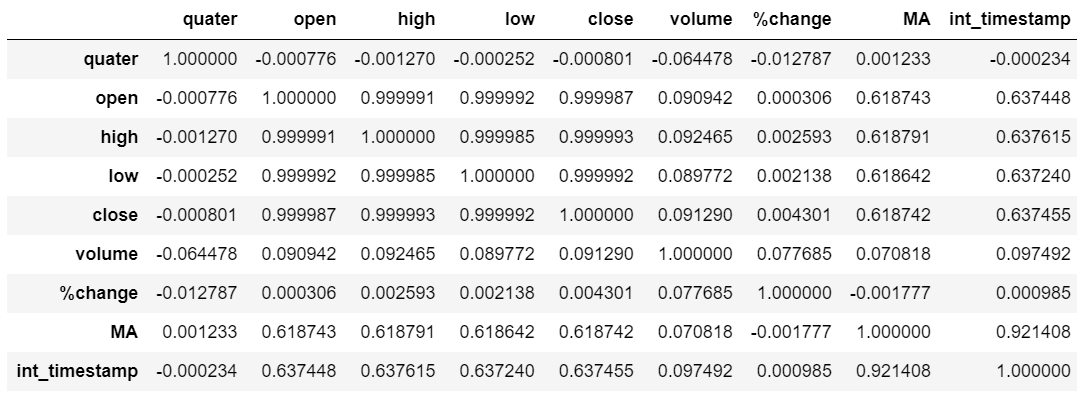


Table 5

Table 5: Table shows correlation values between the columns using Pearson coefficient

So, starting with timestamp we can predict MA as correlation between MA and timestamp is high as well as open and MA.

Now, using given Timestamp and predicted MA we will predict open, then using given Timestamp and predicted MA and predicted open, we will predict close.

Using same process, we will predict high and low.

**6.2** **Model Result and Analysis**

Mean Square Error for Linear Regressions Model.

COMPANY: GMRINFRA

Mean Square Error for MA: 0.681561275914831

Mean Square Error for open: 6.670817967219985

Mean Square Error for close: 0.0019042509203748925

Mean Square Error for high: 0.0006570024659466621

Mean Square Error for low: 0.0006728098888524553

COMPANY: BHEL

Mean Square Error for MA: 14.484769170526645

Mean Square Error for open: 62.28507206105021

Mean Square Error for close: 0.0073055527489031425

Mean Square Error for high: 0.0021371956684333336

Mean Square Error for low: 0.0019233462384935366

**Value Predictions for Date 02-01-2021**

Prediction for company AARTIIND

Date: 2021-01-02 15:30:00

Open: [1087.35883945]

High: [1088.18221237]

Low: [1086.46903305]

Close: [1087.35264896]

Prediction for company ABCAPITAL

Date: 2021-01-02 15:30:00

Open: [42.88245879]

High: [42.94530514]

Low: [42.82060939]

Close: [42.88301602]

Prediction for company ABFRL

Date: 2021-01-02 15:30:00

Open: [180.09527577]

High: [180.26532268]

Low: [179.92512748]

Close: [180.09572198]

Prediction for company ADANIENT

Date: 2021-01-02 15:30:00

Open: [247.0818159]

High: [247.32507447]

Low: [246.83692971]

Close: [247.08461724]

Prediction for company ADANIGAS

Date: 2021-01-02 15:30:00

Open: [220.73773653]

High: [220.96261004]

Low: [220.49964087]

Close: [220.74125142]

Prediction for company ADANIPORTS

Date: 2021-01-02 15:30:00

Open: [365.15035923]

High: [365.4164192]

Low: [364.87580837]

Close: [365.14924213]

Prediction for company AJANTPHARM

Date: 2021-01-02 15:30:00

Open: [1212.93081215]

High: [1213.94916528]

Low: [1211.89439218]

Close: [1212.930471]

Prediction for company AMARAJABAT

Date: 2021-01-02 15:30:00

Open: [667.90344798]

High: [668.41199445]

Low: [667.40041724]

Close: [667.90507997]

**MODEL PREDICTION FOR COMPANY TO INVEST ON DATE 2021-01-02 AND TIME 15:30:00**

Company We should invest on date 2021-01-02 and time 15:30:00 is SRF with profit of 0.02635884097253438

**MODEL PREDICTION FOR COMPANY TO INVEST ON DATE 2023-07-20**

Company We should invest on date 2023-07-20 is NESTLEIND with profit of 2.097678437919967

**6.3 Model Prediction Graph Analysis**

**Open Stock Prices of Companies: Prediction vs actual**

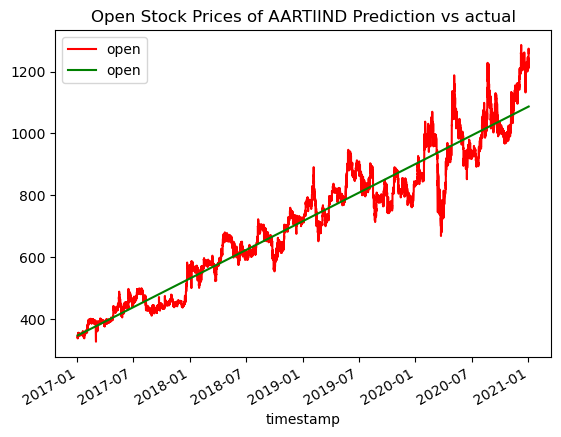


Figure 5a.

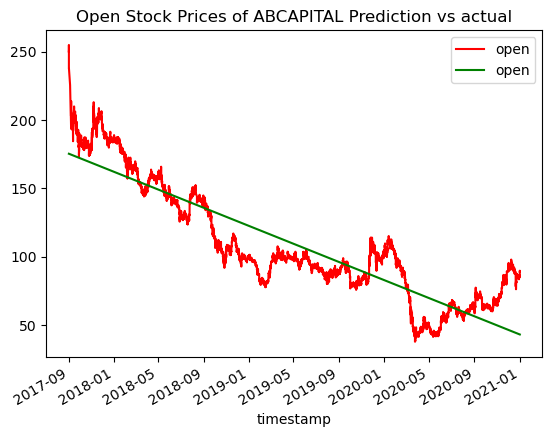


Figure 5b.

Figure 5: Figure shows Actual vs Prediction red shows Actual data and green shows prediction shows for open prices

**Close Stock prices of Companies: Prediction vs actual**

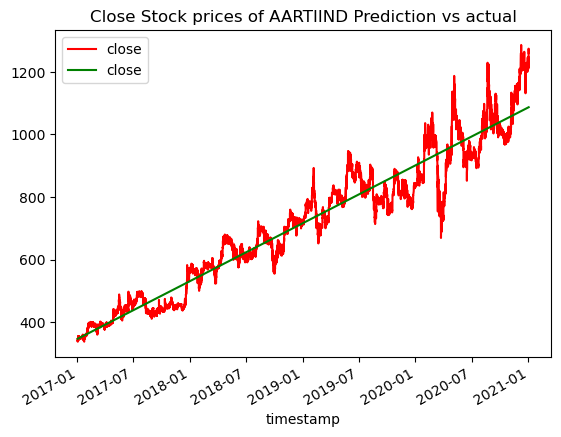


Figure 6a.

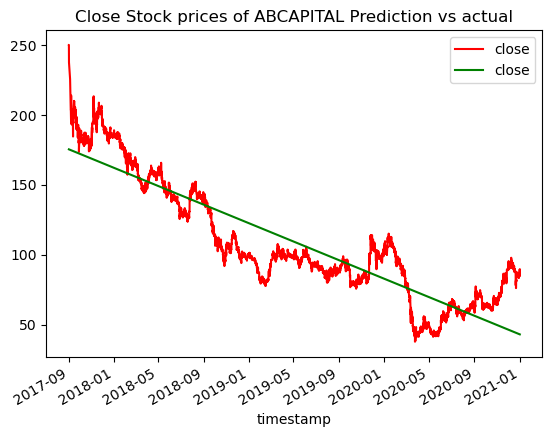


Figure 6b.

Figure 6: Figure shows Actual vs Prediction red shows Actual data and green shows prediction shows for open close

**High Stock prices of Companies: Prediction vs actual**

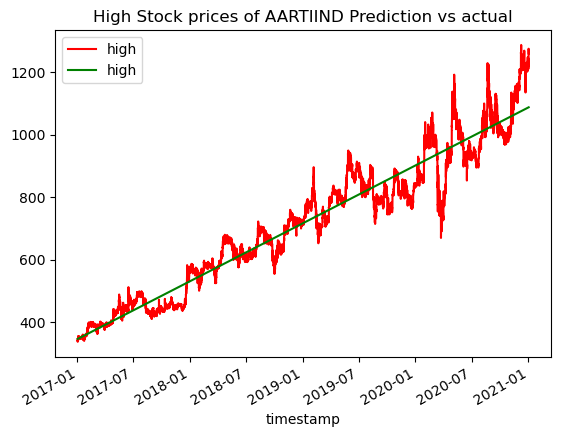


Figure 7a

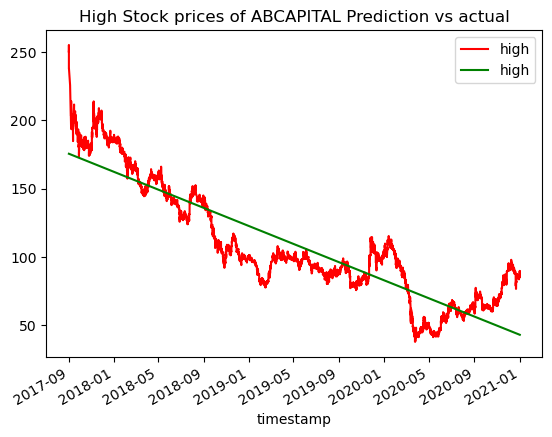


Figure 7b

Figure 7: Figure shows Actual vs Prediction red shows Actual data and green shows prediction shows for high prices

**Low Stock prices of Companies: Prediction vs actual**

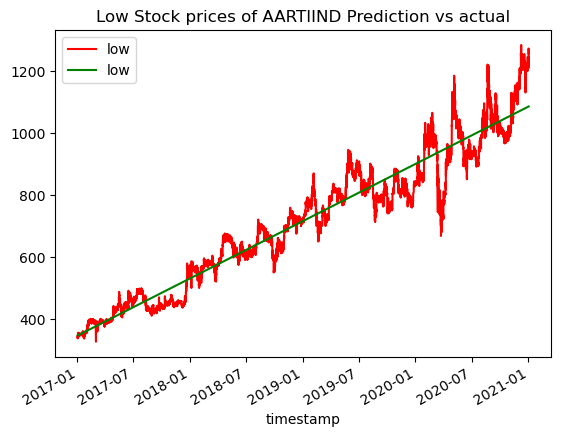


Figure 8a

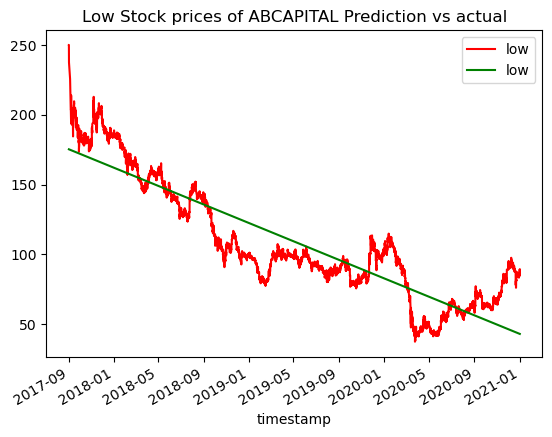


Figure 8b.

Figure 8: Figure shows Actual vs Prediction red shows Actual data and green shows prediction shows for low prices

Similarly, We created prediction graph for more 157 companies using same process.

**7. Insights and Findings**

**7.1 Key Findings**

* The dataset contains information about stock market of different Indian companies like timestamp, date, time, open, high, low etc….
* Missing values were handled for numerical features like open, high, low, close columns by replacing it with previous non null values. For Volume it is replaced by 0 and %change is recalculated.
* Timestamp is transformed in integer and Quarter is mapped to integer
* Visualizations were created to show open, high, low, close, moving average
* A correlation heatmap was used to show the relationships between numerical features. It helped identify any significant correlations between attributes.
* A linear regression model was trained to predict open, close, high, close value. The model's accuracy on the validation set was calculated to evaluate its performance.

**7.2 Significant findings or correlations.**

* The Graph of open, high, low, close are almost identical this means they are high correlated
* The Moving Average Shows slope on Company stock market whether rising and or falling

**7.3 Limitations or challenges encountered**

* Limited data exploration and analysis.
* Linear Regression is not the best model can be used in stock market
* Stock market prices are not linear, predicted value may not much useful for accurate time prediction.

**8. Recommendations**

**8.1** **Recommendations based on the findings.**

* We can conduct a more thorough data exploration to gain a deeper understanding of the dataset. This could involve visualizing distributions, identifying outliers, and analysing relationships between variables to reveal hidden insights.

**8.2** **Potential areas for further analysis or investigation.**

* Finding more Columns in which open, high, low, close depends
* Exploring more companies' data with real time data for more accurate prediction

**8.3 Potential impact of implementing the recommendations**

* Understanding the stock market previous prices, so Customer can know in which company stocks will give him max profit.
* Understanding the stock market previous prices, so Customer can know future values of stock market prices of different companies.

**9. Conclusion**

**9.1 Highlights**

* The challenge was to predict stock market values.
* Missing values were handled for numerical features like open, high, low, close columns by replacing it with previous non null values. For Volume it is replaced by 0 and %change is recalculated.
* A linear regression model was trained to predict open, close, high, close value. The model's accuracy on the validation set was calculated to evaluate its performance.
* Total 795 Linear model created with each company 5 linear model to predict best value using linear regressions

**9.2** **Value of EDA in understanding the dataset.**

* EDA helps data analysts and scientists to gain a comprehensive understanding of the dataset's structure, size, and content. It allows them to familiarize themselves with the data's features, types, and distributions.
* EDA can be used to validate assumptions required for various modelling techniques.
* EDA provides valuable insights that can inform decision-making processes.
* EDA guides data preprocessing steps such as data imputation, feature engineering, and scaling. These steps help prepare the data for modelling and improve the model's accuracy and performance.

**References**

* <https://www.kaggle.com/>
* <https://geeksforgeeks.org/>