A Data analysis and Machine Learning Project

This project focuses on Data Analysis of Microsoft stocks from the Years 2015-2021, the project further expands its scope by applying machine learning Techniques to predict features.

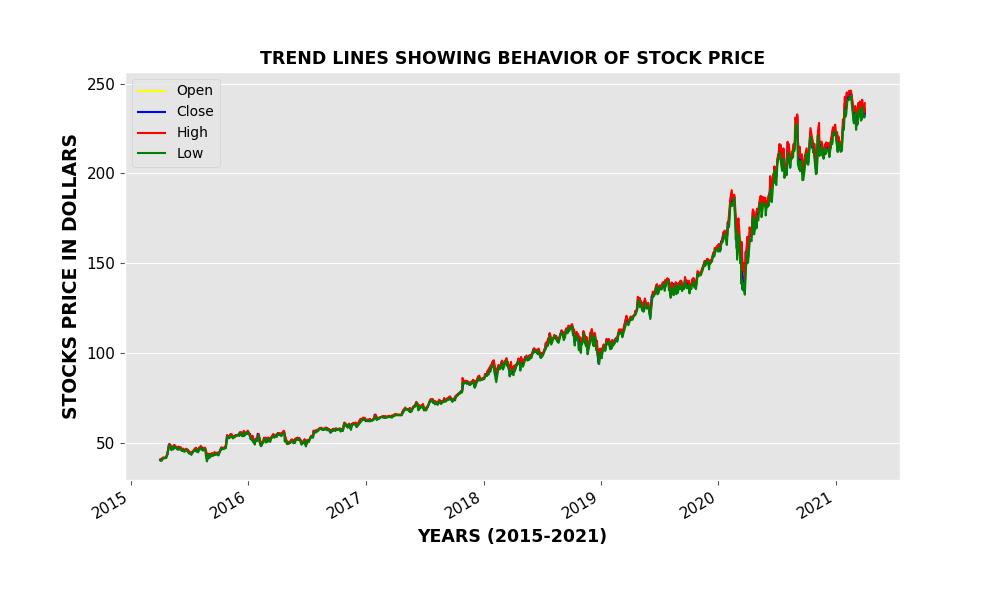
An interactive project, below are the steps taken for the project;

* Data Collection
* Exploratory Data Analysis
* Feature Engineering
* Model Building and Testing

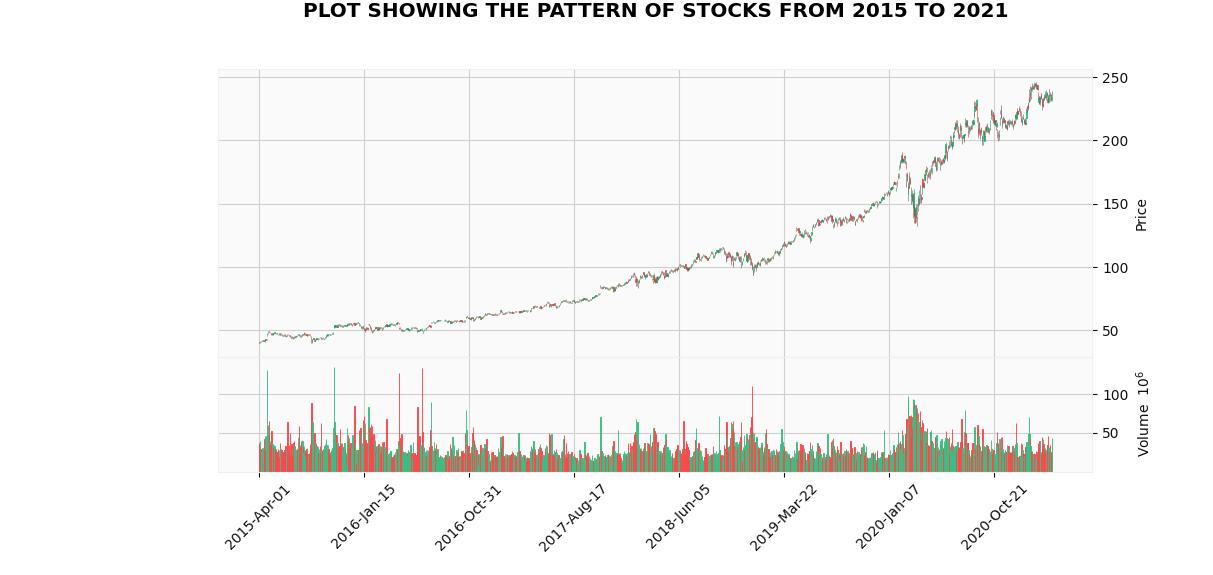
A Microsoft Stocks Data Set downloaded from www.kaggle.com;

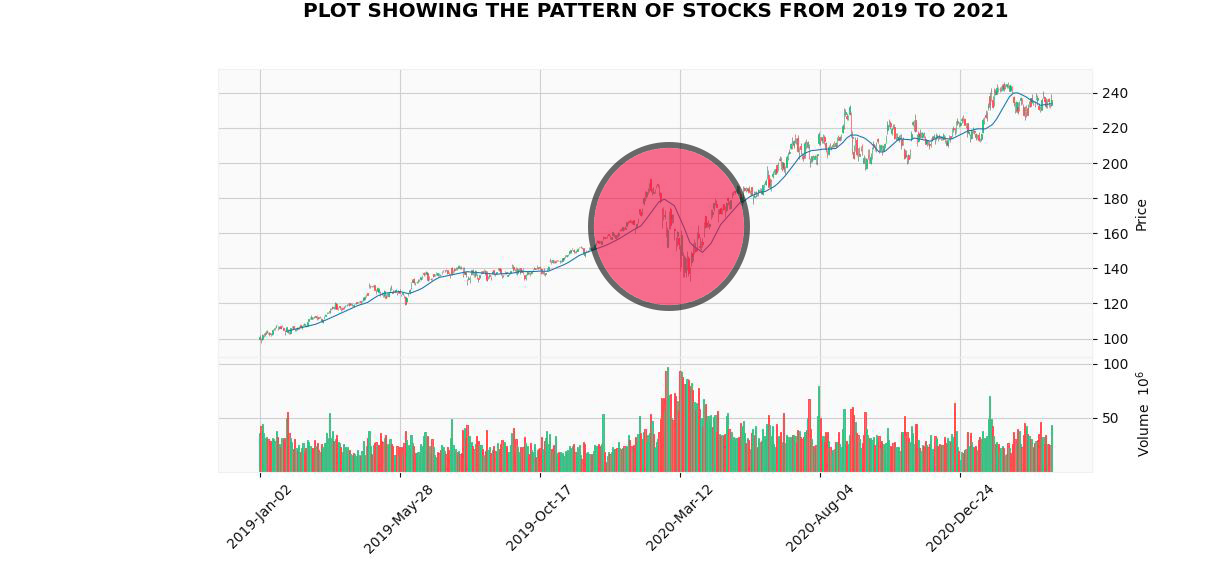
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Volume |
| 04-01-15 16:00 | 40.6 | 40.76 | 40.31 | 40.72 | 36865322 |
| 04-02-15 16:00 | 40.66 | 40.74 | 40.12 | 40.29 | 37487476 |
| 04-06-15 16:00 | 40.34 | 41.78 | 40.18 | 41.55 | 39223692 |
| 04-07-15 16:00 | 41.61 | 41.91 | 41.31 | 41.53 | 28809375 |
| 04-08-15 16:00 | 41.48 | 41.69 | 41.04 | 41.42 | 24753438 |
| 04-09-15 16:00 | 41.25 | 41.62 | 41.25 | 41.48 | 25723861 |
| 04-10-15 16:00 | 41.63 | 41.95 | 41.41 | 41.72 | 28022002 |

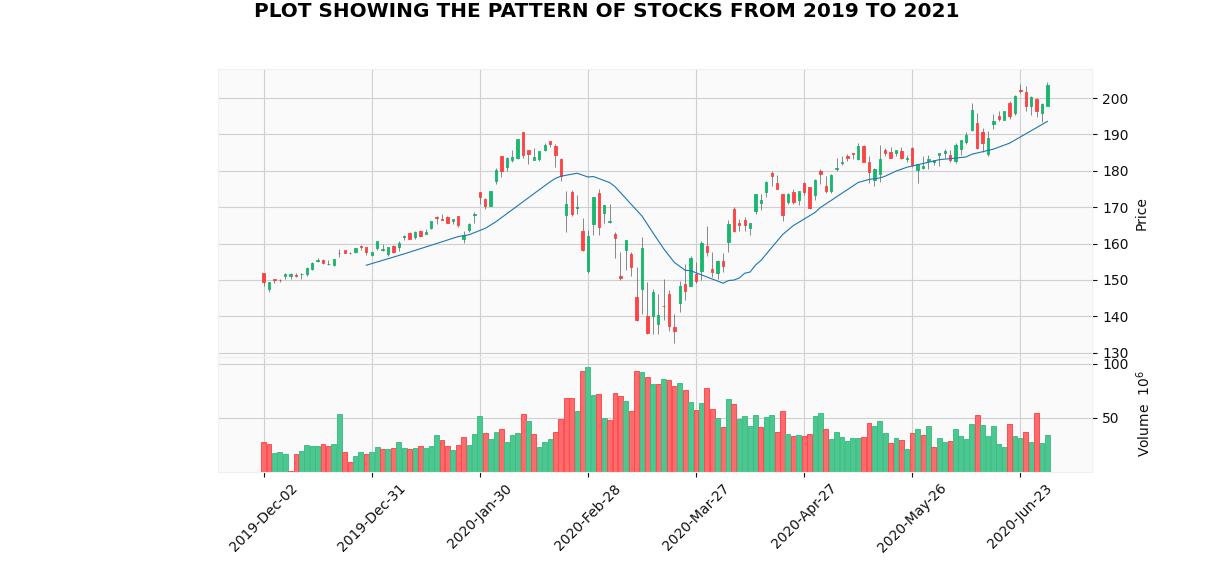
The above table shows a brief and randomly selected portion of the Data Set, originally, the Data Set has 1000+ entries dated from year 2015 to year 2021.

The exploratory Data Analysis aspect of this project involves finding lost insights like how stock was purchased aggressively after the Stock Price dropped drastically, the amount of times the market closed at higher prices than it opened.

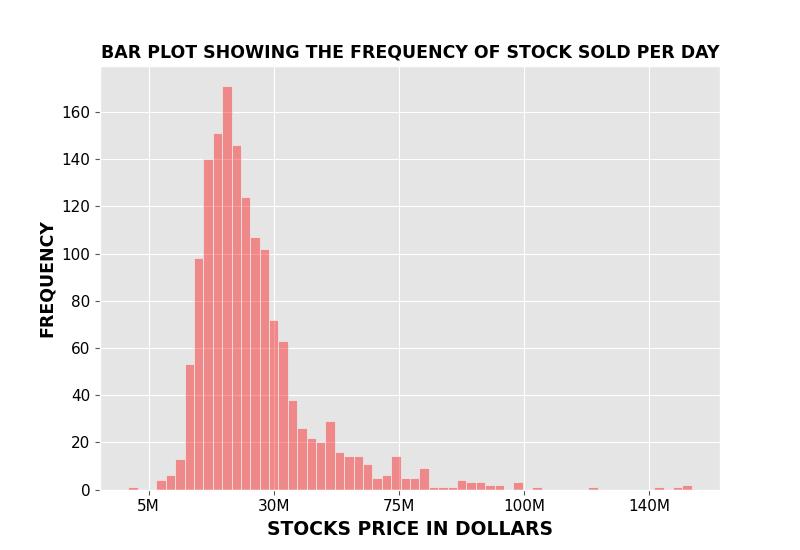
The above plot shows the general trend line behavior of stock prices over the years.

 Candlestick’s representation will be much more comprehensive;

A candlestick plot above showing the trendlines and at the same time volume, it is observed that there was a huge drop in stock price in between 2020 and 2021, though the particular date can’t be ascertained from the above graph.

The shaded area in the above graph helps us conclude that the drop in price of stocks occurred between February to March, 2020, also the volume is observed to have risen so much during that period which is quite reasonable as more people bought stocks as the price dropped.

A clearer representation above;



The above histogram highlights the distribution of the total number of stocks sold per day.

**OTHER FACTS THAT WERE GAINED DURING EXPLORATORY DATA ANALYSIS**

* NET CHANGE AND MAXIMUM INCREASE FROM OPEN PRICE

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | Open | High | Low | Close | Volume | Max Increase from Open Price | Net Change |
| 2015-04-01 16:00:00 | 40.6 | 40.76 | 40.31 | 40.72 | 36865322 | 0.16 | 0.12 |
| 2015-04-02 16:00:00 | 40.66 | 40.74 | 40.12 | 40.29 | 37487476 | 0.08 | -0.37 |
| 2015-04-06 16:00:00 | 40.34 | 41.78 | 40.18 | 41.55 | 39223692 | 1.44 | 1.21 |
| 2015-04-07 16:00:00 | 41.61 | 41.91 | 41.31 | 41.53 | 28809375 | 0.3 | -0.08 |
| 2015-04-08 16:00:00 | 41.48 | 41.69 | 41.04 | 41.42 | 24753438 | 0.21 | -0.06 |
| 2015-04-09 16:00:00 | 41.25 | 41.62 | 41.25 | 41.48 | 25723861 | 0.37 | 0.23 |
| 2015-04-10 16:00:00 | 41.63 | 41.95 | 41.41 | 41.72 | 28022002 | 0.32 | 0.09 |

The above table shows that two new columns (‘Max Increase from Open Price’, ‘Net Change’) have been added. The (‘Max Increase from Open Price’ being the difference between ‘High’ and ‘Open’ and the ‘Net Change’ being the difference between open and close.

A positive Net Change coefficient tells that for that particular day the stock price closed higher than it opened and vice versa

* The total number of times where the market closed higher than it opened was 802.

|  | **Open** | **High** | **Low** | **Close** | **Volume** |
| --- | --- | --- | --- | --- | --- |
| MIN VALUES | 40.34 | 40.74 | 39.72 | 40.29 | 101612.0 |
| MEDIAN OF VALUES | 93.99 | 95.10 | 92.92 | 93.86 | 26629615.0 |
| MAXIMUM VALUES | 245.03 | 246.13 | 242.92 | 244.99 | 135227059.0 |

**MACHINE LEARNING**

* + 1. Building the Model

The model was built for the purpose of predicting the closing price provided all other features are provided, also, not just one API was used in building the model, TensorFlow’s Keras and Scikit learn were used in building this model.

* + 1. Testing the Model

the model records a 99% percentage.