Analaysis On The Stock Price data Of The Boeing Company, The Walt Disney, and The Microsoft Corporation

Sarath Chandra Gollamudi
School of Computing
National College of Ireland
Dublin, Ireland
x19193581@student.ncirl.ie

Mansi Udani
School of Computing
National College of Ireland
Dublin, Ireland
x19186088@student.ncirl.ie

Shashikumar Madihalli Byrappa
School of Computing
National College of Ireland
Dublin, Ireland
x18173837@student.ncirl.ie

***Abstract*— Tools for structuring, querying, forecasting, and visualizations are the areas full of fruitful development. In this paper, the unstructured data of three US companies (datasets) stock prices has loaded to MongoDB and after with necessary data transformation, the structured data has loaded into the PostgreSQL database. On the other hand, with the structured data from PostgreSQL a proper data analysis has performed to understand the hidden pattern present in the data and the performance of all three companies, with visualizations by using Plotly, Matplotlib, Seaborn and with necessary python libraries. However, to find the future trends of the data, the statistical analysis has performed using the ARIMA model on three datasets.**

Keywords—Stock Prices, The Boeing Company, Walt Disney, The Microsoft Corporation, EDA, ARIMA

I. INTRODUCTION

A. Objective:

The leading objective of this project is to study the variations in stock prices of three leading American companies. Predictions are based on the historical prices, with factors like open and close price, volume, dividend, split and many more. The critical analysis done during the project can help in knowing the stock trend movements of the company. Furthermore, the elaborate visualizations can help as a guide to quick decision making. The purpose of this analysis is to find the trends and forecast the future prices that will help the investors to make sound decisions.

B. Motivation of the Problem:

Stock market is one of the most enticing field as it transacts big money. With a lot of money, comes a lot of risk. Also, the unpredictable nature of the stock market makes forecasting a very tricky job. Huge amounts of data are created, due to the non-stationary nature. More the data, more is the possibility to make different kinds of visualizations. By showing visualizations in fully responsive 3D plots can give deeper insights to investors than studying the whole dataset. The responsive nature can help in precisely knowing the trend and seasonality of the stock. Understanding the data just by scrolling the mouse, initiates quick and sound decision making. These charts can provide insights to intraday traders (buying and then selling the stock on the same day) for short term profits and to investors who want make profits in the long term.

C. Research Question:

The major factors that we focused here is loading unstructured data into Mongo DB and then passing the structured data to Postgres for querying. Later, forecasting the future prices and visualizing different kinds of data. For the forecasting, we have applied an ARIMA model in the all the three datasets. Intricate visualizations to understand the dataset and to support our analysis was a priority.

II. RELATED WORK

Making investment decisions in stock market not only requires time and knowledge, but also an acute awareness of the historic data. The prices are dependent on a couple of factors like the company performance, economy of the company[1]. Besides, if same rates of prices are seen changing over long-term periods on numerous stocks, then a potential connection is derived between the corporations. Hua et al. [2], have proposed a methodology that applies a force-directed algorithm along with time-series to understand the deeper insights on potential relationships between multiple stocks with less human interventions. Hence, assisting the decision making in future with graphs.

Among the NoSQL database MongoDB is one of the classic databases. The main feature of MongoDB is the storage of data and it has large data processing functions like a relational database. MongoDB will store the data in the format of the video, documents, e-mail, pictures, and many other forms. It is not like structured data it Is more like an unstructured database and it will not store data in the form of rows and columns as data is large and it requires more space [3].

The following are the key features of MongoDB: (1) Data model is suitable for design. Since it is documented database easy to map with programming language data types and as it stores the document in array format does not need for joins and it is schema-free for easy schema evaluation. (2) MongoDB supports the following operations like queries related to embedded documents, conditional operators, regular expressions, and array. Using Map/reduce it reduces major SQL queries also we can implement complicated aggregated operations [1]. In this analysis we used MongoDB is used to store unstructured data in JSON format. Another database we used is PostgreSQL (professed as PostgreSQL), it is an open-source database and created by a group of participants [4]. In this paper, we used PostgreSQL to store the cleaned data in structured formatted.

III. METHODOLOGY

The three datasets selected for this analysis are taken from the Quandl website and below are the source website links. The datasets which are used are as follows, The Boeing Company, Walt Disney, and The Microsoft Corporation. All these are related to the stock price of the respective companies

from the year 2013 to 2017. These files are extracted from the source website in the JSON format. These JSON format files are unstructured data that consists of a lot of columns and data which are not necessary for this analysis. The datasets consist of more than 3000 records altogether. With the help of this data, the interesting hidden patterns in the data is visualized which is useful for the shareholders to understand the company performance before buying or selling the shares.

Data source links:

The Boeing Company:

https://www.quandl.com/api/v3/datasets/EOD/DIS.json?api_key=WfqN4Vbayqo_HHiZYV-w

The Microsoft Corporation:

https://www.quandl.com/api/v3/datasets/EOD/MSFT.json?api_key=WfqN4Vbayqo_HHiZYV-w

The Walt Disney:

https://www.quandl.com/api/v3/datasets/EOD/DIS.json?api_key=WfqN4Vbayqo_HHiZYV-w

A. Data Pipeline:

There are three steps involved in this complete analysis to get the knowledge from the data as shown below.

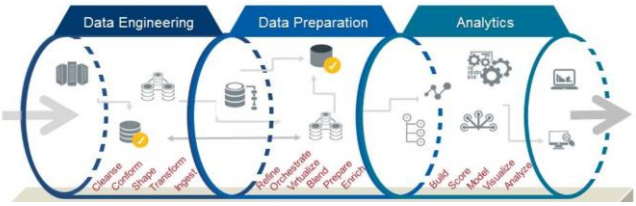
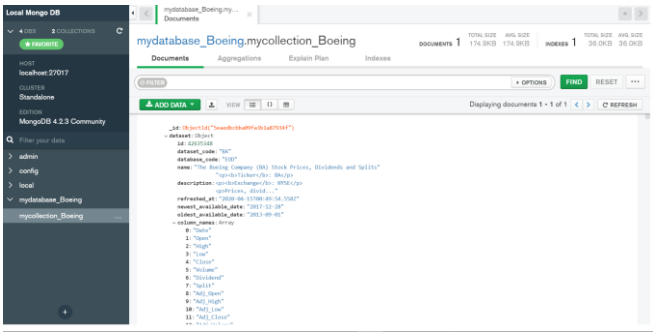


Fig-1: Process flow

From the figure above we could see that there are different steps like extraction, transformation/cleaning, visualization, and model selection. At first, the raw data of the three JSON files were loaded to the **MongoDB** by using the python libraries. Below is the screenshot of one of the dataset's databases.



Screenshot-1: MongoDB database of The Boeing Company

In the screenshot above, we could see that the dataset file was loaded to MongoDB in the form of a document. The same procedure has followed for the remaining two datasets as well.

B. Data Extraction:

Secondly, these unstructured data were extracted from MongoDB to panda's data frame using necessary python libraries to change into the structured data for our analysis as shown below.

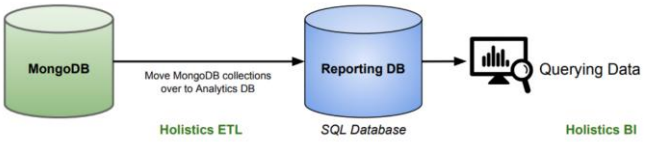


Fig-2: Data extraction from mongoDB

On the other hand, after obtaining the structured data with necessary data transformation the final data was stored in **PostgreSQL**.

C. Data Transformation/Cleaning:

Thirdly, the obtained unstructured data is converted into structured data by doing the following analysis which is Verifying the NA values, Verifying the Outliers in the data, and the correlation between all the variables.

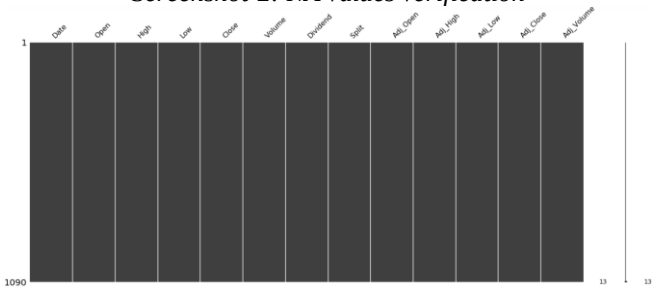
Fig-3: Data Transformation using python

Verification of NA values: This analysis has performed by verifying all the column's null values and by visualizing as shown below.

```
#Verifying the missing values
df.isnull().sum()
```

```
Date          0
Open          0
High          0
Low           0
Close         0
Volume        0
Dividend      0
Split         0
Adj_Open      0
Adj_High      0
Adj_Low       0
Adj_Close     0
Adj_Volume    0
dtype: int64
```

Screenshot-2: NA values verification

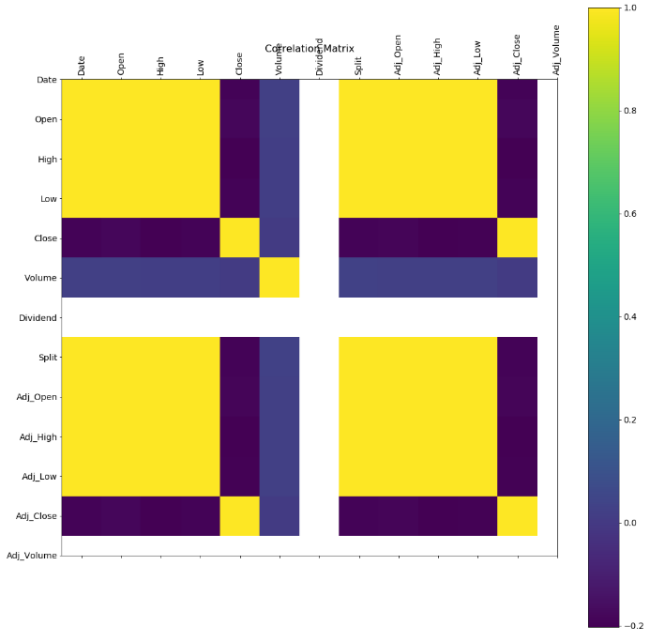


Screenshot-3: Graph of NA values

In the screenshots above, we could clearly see that there are no missing values in the data so we can skip the process of replacing the NA values with the mean or median of the respective columns. The same procedure has also followed for the remaining two datasets as well and we have observed that there are no NA values.

Correlation Verification: In this step, the correlation between all the variables has verified by visualizing, in order to remove the columns if they are highly correlated with any other

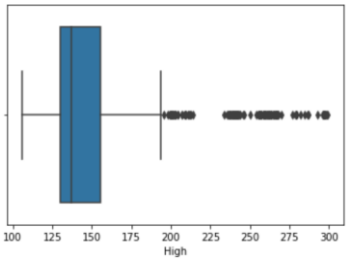
columns in the dataset. Below is the screenshot of The Boeing Company dataset verification.



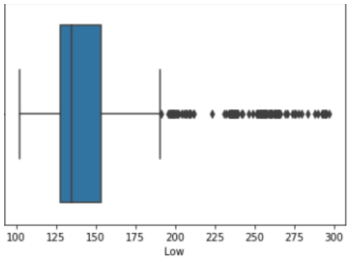
Screenshot-4: Correlation Matrix

We could see that in the above matrix there are columns that are highly correlated which are as follows (Adj_Open, Adj_High, Adj_Low, Adj_Close, and Adj_Volume). So, we have ignored these columns from the data. The same procedure has followed for the remaining two datasets as well.

Verifying the Outliers: To verify the outliers in the data the box plot has generated for some columns of the data and below is the output screenshot. This boxplot has generated using the **seaborn** library.



Screenshot-5: Outliers of column High price



Screenshot-6: Outliers of column Low price

As shown above, we have verified for all the columns in the data and we could see that there are some outliers in the data. To make this clean data, the IQR range has set for the data which are the upper fence and the lower fence as shown below.

```
df = df[~((df < (Q1 - 1.5 * IQR)) |(df > (Q3 + 1.5 * IQR))).any(axis=1)]

df.shape

(881, 8)
```

Screenshot-7: IQR range to remove the outliers

We could see that after setting the IQR range the dataset size is reduced and became the clean data which can be useful for our final analysis. These final cleaned data were stored in the PostgreSQL as shown in the *Fig-3*. The same procedure has followed for the remaining two datasets also.

PostgreSQL: The final data obtained after the necessary transformation process as shown in the above steps was stored in the PostgreSQL by using necessary python libraries (psycopg2). Below is a high-level understanding of how it looks.

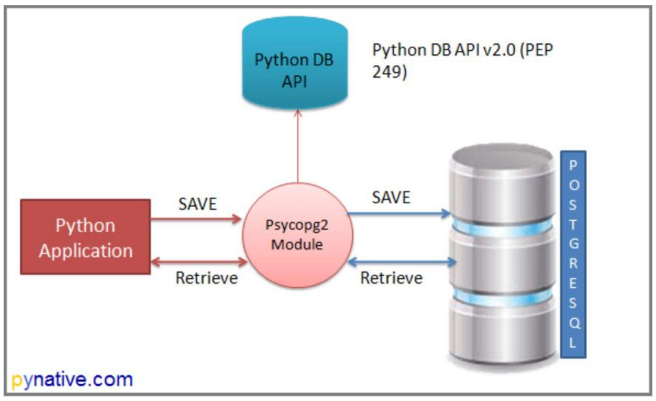
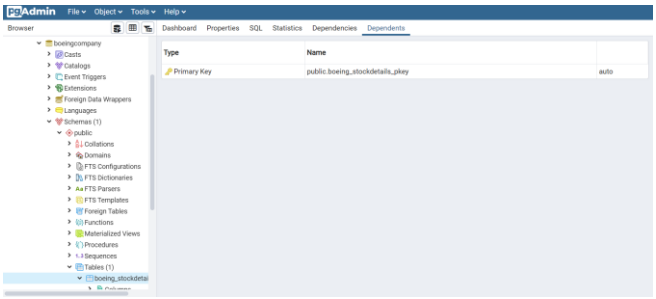


Fig-4: Storing data into PostgreSQL using python language

Based on the above figure, the cleansed data was loaded into PostgreSQL by creating a database following with table and columns. Below is the screenshot of the database.

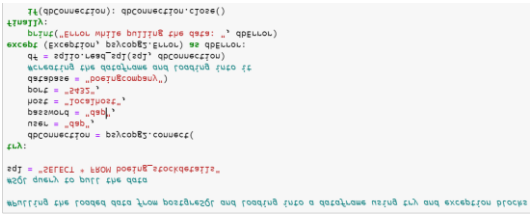


Screenshot-8: PostgreSQL Database of The Boeing Company

We could see in the above screenshot the final cleansed data was stored in the database. The same procedure has followed for the remaining two datasets as well and the respective databases were created and stored the final output.

IV. RESULTS

In this section, the actual analysis of the data is explained in the form of visualization by using the python libraries **Plotly, Matplotlib, and Seaborn.** Prior to this below are the final columns that we have pulled from the PostgreSQL for the EDA using the SQL queries with python language.



Screenshot-9: Extracting the data using queries

Feature	Description	Datatype
Date	Day of the price	datetime
Open	Opening price of the day	float
High	High price of the day	float
Low	Low price of the day	float
Close	Closing price of the day	float
Volume	Total volume of the day	integer
dividend	Dvidend on quatrely basis	integer
Split	Split of the shares	integer

Table-1: Final Features

The above are the final features and the same for all the three datasets. Using these features the final analysis has done.

Below plots are the year wise closing stock price details in the calendar format, and the all features scatter matrix plots.

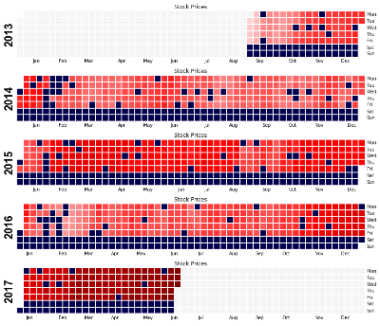


Fig-5: Year wise Closing Price of The Boeing Company data

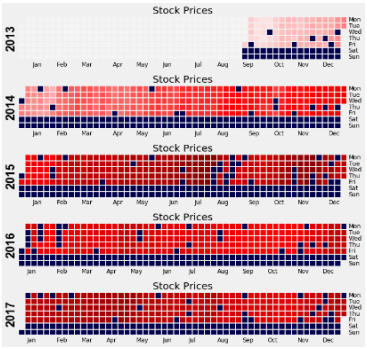


Fig-6: Year wise Closing Price of The Walt Disney data

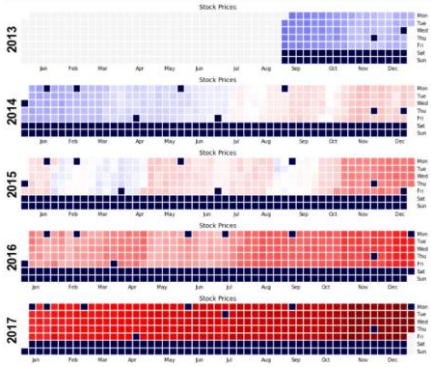


Fig-6: Year wise Closing Price of The Microsoft Corporation

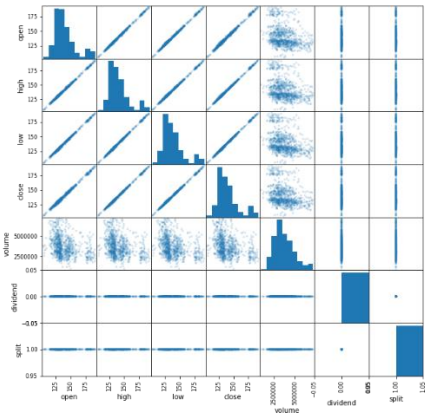


Fig-8: Scatter matrix of all features of The Boeing Company

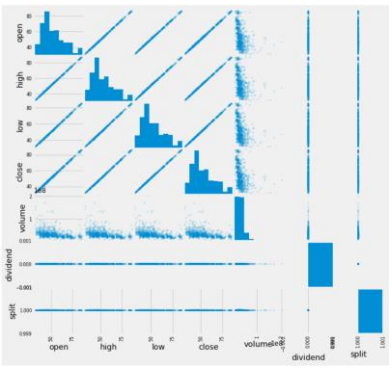


Fig-9: Scatter matrix of all features of The Microsoft Corporation

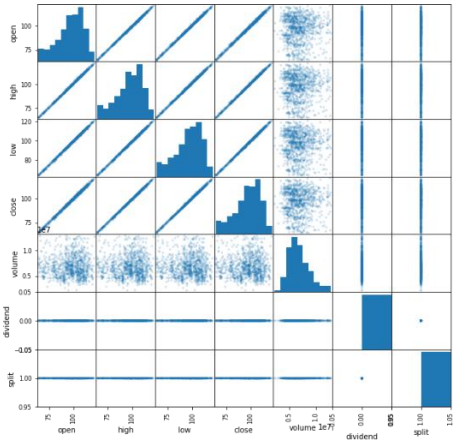


Fig-10: Scatter matrix of all features of The Walt Disney Company

Analysing the total variance between the Open and High Stock prices of the day with spread plot as shown below.

Fig-11: Spread plot of The Boeing Company between open and high columns

Fig-12: Spread plot of The Microsoft Company between open and high columns

Fig-13: Spread plot of The Walt Disney Company between open and high columns

In the above plots, we could see at the bottom of the plot is the difference between the open and high price of the respective day of all three companies (datasets).

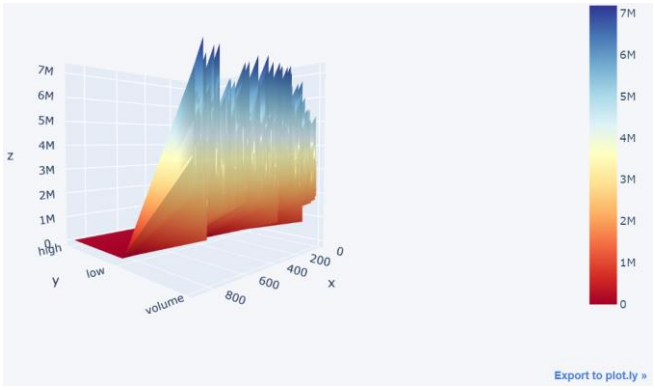


Fig-14: Relation between high, low and volume columns of The Boeing Company

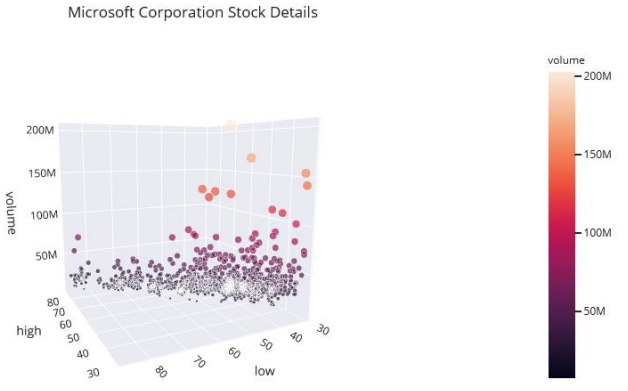


Fig-15: Relation between high, low and volume columns of The Microsoft Corporation

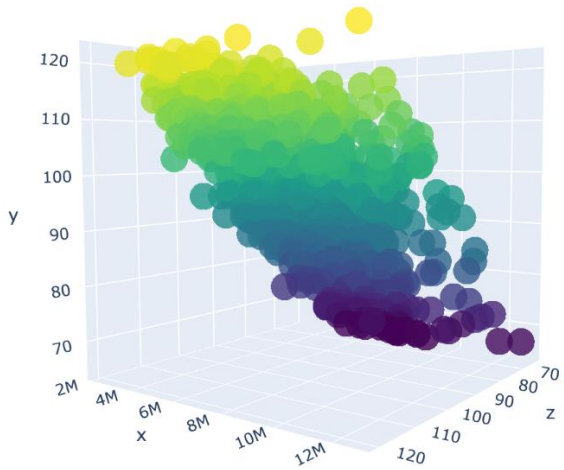


Fig-16: Relation between high, low and volume columns of The Walt Disney

We could see the relationship between the variables high, low, and volume in the above 3D interactive plots of all three datasets.

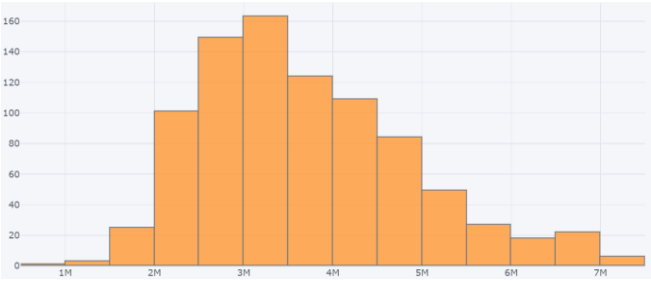


Fig-17: Volume column histogram of The Boeing Company

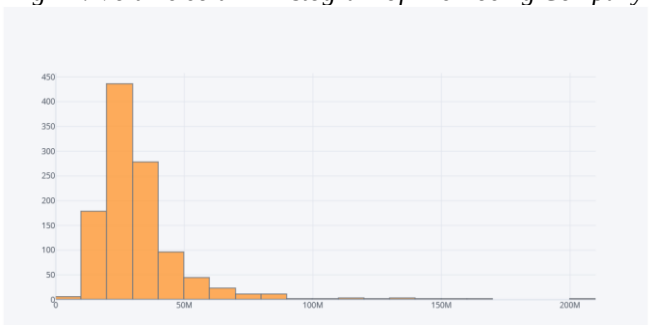


Fig-18: Volume column histogram of The Microsoft Corporation

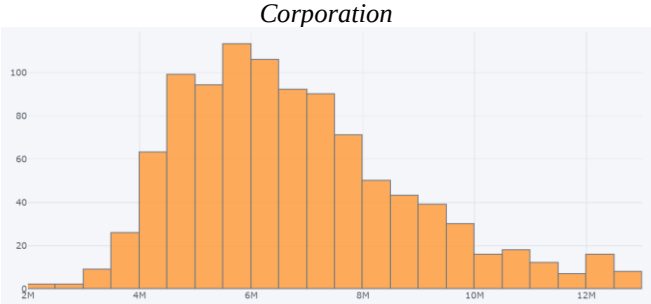


Fig-19: Volume column histogram of The Walt Disney

Based on the above volume column plots, an investor and share-holders can decide the performance of the company.

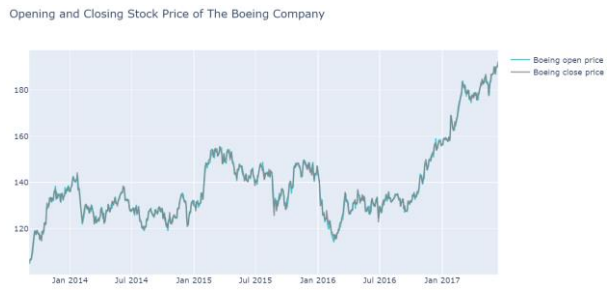


Fig-20: Opening and Closing price of The Boeing Company

Fig-21: Opening and Closing price of The Microsoft Corporation

Fig-22: Opening and Closing price of The Walt Disney

In the above figures, we could clearly see that there is not much difference between the opening and closing price of all three companies. On the other hand, we have found an interesting relation between the high price and volume columns. Below are the output plots of all three datasets.

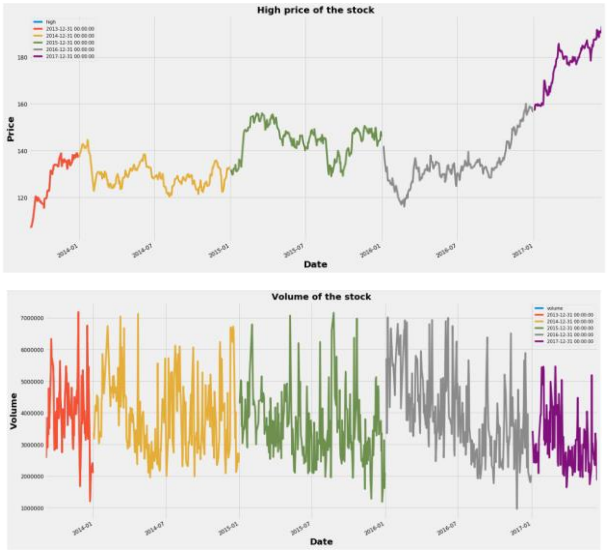


Fig-23: High and Volume column's relationship of The Boeing Company

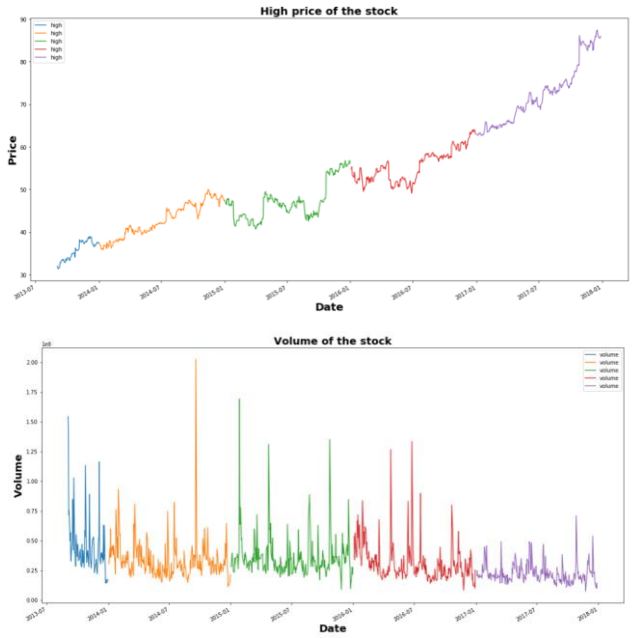


Fig-24: High and Volume column's relationship of The Microsoft Corporation

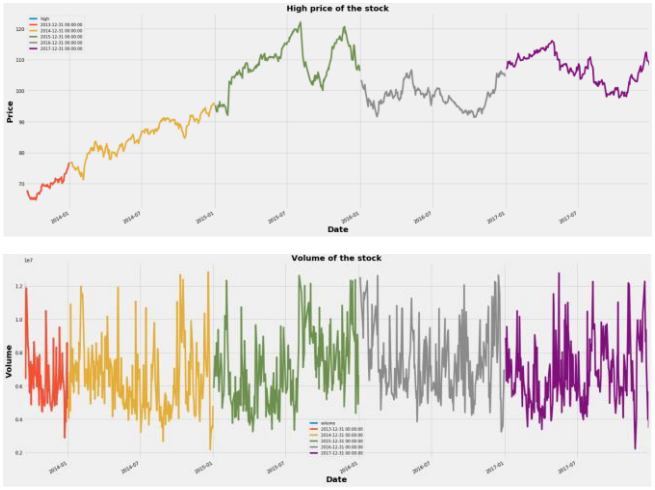


Fig-25: High and Volume column's relationship of The Walt Disney

In the above all figures of all three companies we could see that there are two observations we need to consider one is that when the price is increasing the volume may increase or decrease. This means that shareholders can easily understand when they need to sell or buy shares based on these plots.

Applying Statistical Analysis: (ARIMA Model)

The statistical analysis has performed further on the data to forecast the future values and visualizing those values so that the shareholders or people can easily understand the previous data of the company stock prices and the future predictions. As part of this, the ARIMA model was used which is the best model in prediction. This was performed by using the required libraries of the python language for all three datasets.

To this at first, the date column has changed to index and the columns which are not required for the prediction were removed except the target column. In this analysis, our target column is 'high'. After applying the necessary steps below are some of the visualizations output.

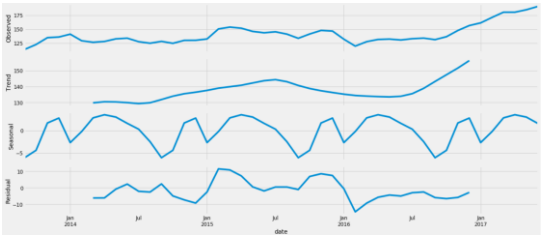


Fig-26: Trend and Seasonality observation of The Boeing Company data

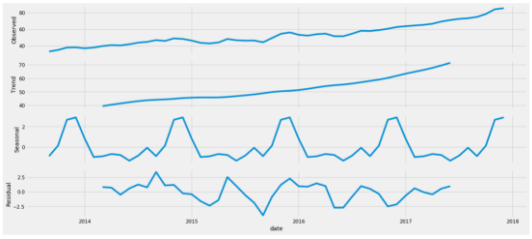


Fig-27: Trend and Seasonality observation of The Microsoft Corporation data

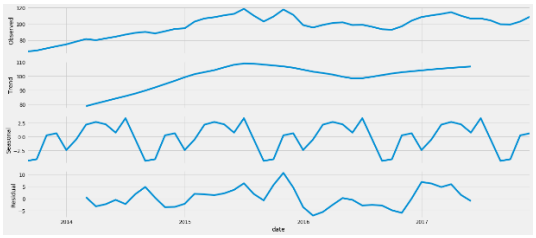


Fig-28: Trend and Seasonality observation of The Walt Disney data

After observing the above trend and seasonality the Auto ARIMA model having the lowest AUC value has applied on the data and below is the output for The Boeing Company dataset. The same method has followed for remaining three datasets as well.

```
#We can see that ARIMA has varified all the possible models and gave the ouput
#The model which is having the lowest AIC value will be the better on to fit
#From the above output we could see that ARIMA(1, 1, 1)x(1, 1, 0, 12)12 - AIC:134.96862030033222 is the best model.
#Fitting the ARIMA model on our data
bestfit = sm.tsa.statespace.SARIMAX(stockprice_forecast, order=(1, 1, 1), seasonal_order=(1, 1, 0, 12), enforce_stationarity=Fa
results = bestfit.fit()
print(results.summary().tables[1])
#Below are the standard error, coefficient and pvalue
```

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.1763      0.956     -0.184      0.854      -2.051       1.698
ma.L1          1.0000   1.23e+04   8.16e-05      1.000   -2.4e+04    2.4e+04
ar.S.L12      -0.6027      0.154     -3.918      0.000      -0.904      -0.301
sigma2        29.9435   3.67e+05   8.16e-05      1.000   -7.19e+05    7.19e+05
==============================================================================
```

Screenshot-10: ARIMA Model output of The Boeing Company

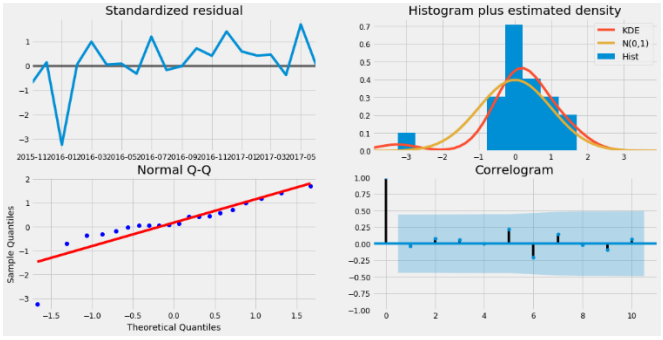


Fig-29: ARIMA Model residual plots of The Boeing Company

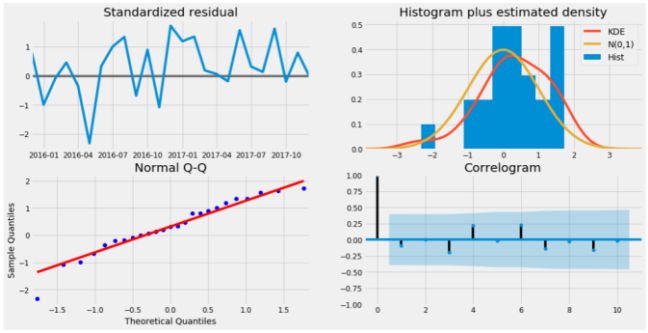


Fig-30: ARIMA Model residual plots of The Microsoft Corporation

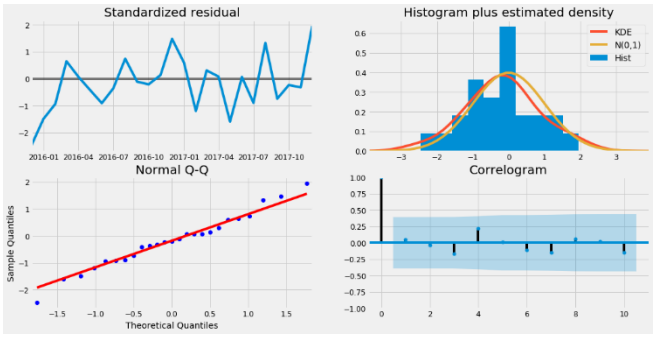


Fig-31: ARIMA Model residual plots of The Walt Disney

We could see in the above plots Normal Q-Q and the histogram that the data is normally distributed and not skewed for all three datasets.

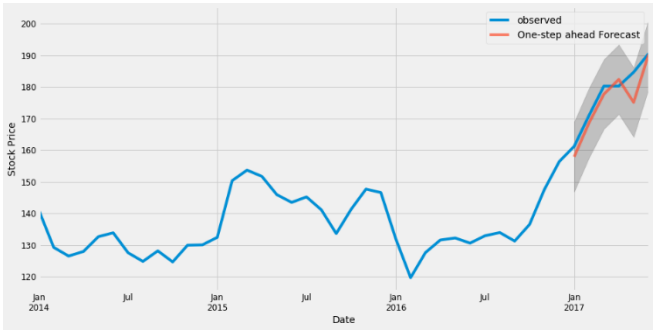


Fig-32: Observed values vs Predicted values plot of The Boeing Company data

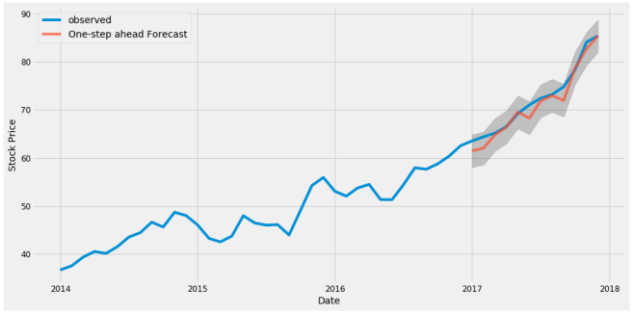


Fig-33: Observed values vs Predicted values plot of The Microsoft Corporation data

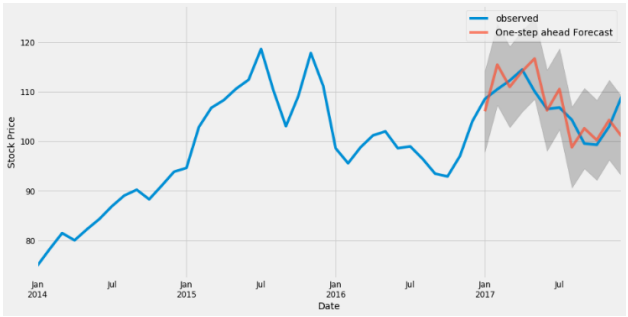


Fig-34: Observed values vs Predicted values plot of The Walt Disney data

In the figure above, we can see that the forecasting values of the model are following the previous data and there is not much difference.

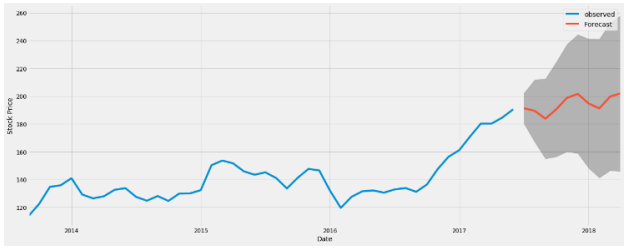


Fig-35: Forecasting output of The Boeing Company

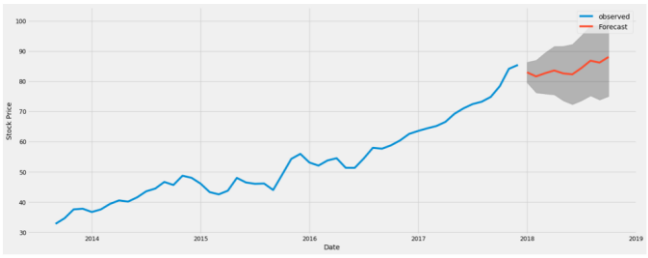


Fig-36: Forecasting output of The Microsoft Corporation

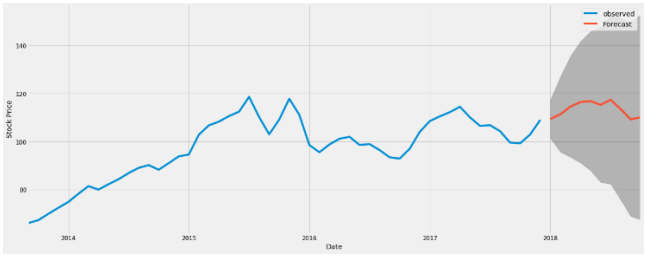


Fig-37: Forecasting output of The Walt Disney

We could clearly see in the above plots; the forecasting values graph which is in orange colour. This means that there will be a slight fall in the stock price of The Boeing company and The Microsoft Corporation but on the other side we could see there is hike in the stock price of The Walt Disney.

V. CONCLUSION AND FUTURE WORK

The ARIMA model is applied in three US companies. It proves to be the best fit for Walt Disney company as the RMSE value is less compared to other two companies. Along with this we can also see that forecasted values in the plot showing the hike in the price of Walt Disney. So, an investor is highly recommended to invest in this company for future benefits. As for the intraday traders, Walt Disney company is most advised from the (High-Volume) fluctuations graph shown above is most advised. The more is the difference in these prices, more will be the profit range. The main intension for forecasting price is to give the insight about the price to investor to buy or sell the stocks. In future work, we would like to work on the visualization by creating a webpage in support of this to give full details by creating dashboards and more interactive graphs. Moreover, Batch processing can also be applied to the stock market data for future work.

REFERENCES

[1] S. S. Pal and S. Kar, "Time series forecasting for stock market prediction through data discretization by fuzzistics and rule generation by rough set theory," Math. Comput. Simul., vol. 162, pp. 18–30, 2019, doi: 10.1016/j.matcom.2019.01.001.

[2] J. Hua, M. L. Huang, G. Wang, and M. Zreika, "Applying data visualization techniques for stock relationship analysis," Filomat, vol. 32, no. 5, pp. 1931–1936, 2018, doi: 10.2298/FIL1805931H.

[3] Gu, Y., Shen, S., Wang, J., & Kim, J.-U. (2015). Application of NoSQL database MongoDB. 2015 IEEE International Conference on Consumer Electronics - Taiwan.

[4] Sultana, S., & Dixit, S. (2017). Indexes in PostgreSQL. 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).