Indian Institute of Information Technology,Sri City

Department of Computer Science engineering

A Project report on

Auto-regression analysis with time-series data for future event prediction

This a course project as part of the Python for Data Science(PDS)S course this project is submitted to under Dr.Sreeja SR

Supervised by

Dr.Sreeja SR
Department of Computer Science engineering
Indian Institute of Information Technology

Submitted by

Yalakanti Eswar S20210020332 ECE

Abstract

This project's main goal is to predict future stock values by looking at a company's past data and giving insights on its stocks using pth order auto-regression analysis. The core of the project deals with time series data, which means we're looking at values recorded at regular time intervals. First of all we start with Exploratory Data Analysis (EDA). This step helps us uncover the structure and patterns in our time series data. We plot graphs, check how different columns are related, and see if our data follows a normal pattern or not. We also look for trends, figuring out if the values are going up or down, and check for any repeating patterns.Once we've done our EDA, then we go on to data pre-processing. Here, we check for missing values, outliers, and make sure our data is normal. This step is crucial to get our data in good shape for the next phase.The main part of the project is using auto-regression analysis. This is a way of studying how past values of data are connected over time. By doing this, we aim to predict what the future stock values for a particular company over some period of time of future. As part of auto regression we use arima, auto-arima and we will find the order from the pacf plots and also we will check the accuracy using the accuracy matrix which is used to measure the best error metric to predict the data. Using the auto-arima we find the order and compare the value we got using the graph. Using the values from auto-arima we will use arima model to train and predict the future values.The main terminologies in the data set is "Open","Close","Adjclose","Volume".Here we open means the opening value of the stock during its start of the day and close represents the closing value of the stock at the end of the day and volume represents the how many stocks have been bought until now and here for the analysis we will use closing values as it is the best way to decide whether we can buy that stock the next day or not.In summary, we go through EDA, then preprocess the data, and then use auto-regression analysis to predict future stock values and finally we make the comment on which company has best stock value. .

List of contents

List of Figures

Abbreviations and List of Symbols

EDA	Exploratory Data Analysis
ROLLING STD	Rolling standard deviation
ARIMA	Auto Regressive Integrated Moving Average
ACF	Auto Correlation Function
PACF	Partial Auto Correlation Function
ADF	Augmented Dickey-Fuller

Libraries Used

ggplot2	Data visualization package
zoo	Handling and analyzing time series data
tseries	Handling and analyzing time series data
forecast	Time Series Analysis and Computational Finance
tidyverse	provides various data transformation funtions

Chapter 1

Data Preprocessing and Univariate Data analysis

Here, we have chosen one company and one column, and we are preprocessing and doing a univariate time series analysis on that information.We have developed a function that requires the data, company name, and column name as parameters in order to do univariate analysis and data preprocessing.

1.1 What happens if Data is not preprocessed?

It can lead to various challenges and issues that may adversely affect the performance of models or analyses applied to the data Quality, consistency, and usability of time series data in various applications, including forecasting, anomaly detection, and machine learning models is to be checked by preprocessing to improve reliability on the model. The nan values and outliers effect the model .

Some Data preprocessing steps are:

- Checking any missing values or nan values.
- Checking for normality.
- Checking the outliers present in the data.
- Checking whether the given data is stationary or not.
- Splitting the data set into train set and test set.

1.2 Checking any Missing Values or NAN values :

We determine whether there are any null values in the data frame, and if so, we either eliminate the samples that contain them or replace the null values with the previous value (this doesn't always work).We utilise the following code to determine the total number of null or NAN values: **sum(is.na(df))**.

In our dataset there are no null values are present.

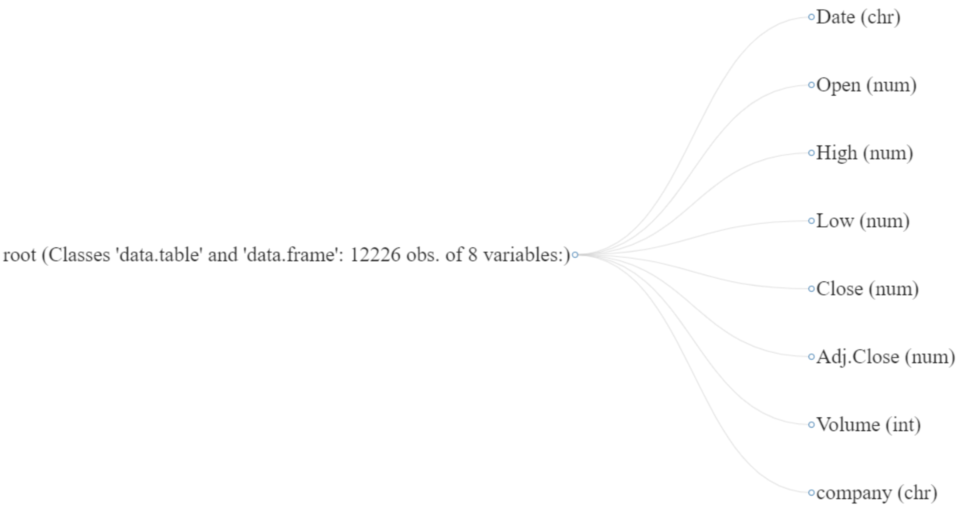


Figure 1.1: Structure of our Data Set

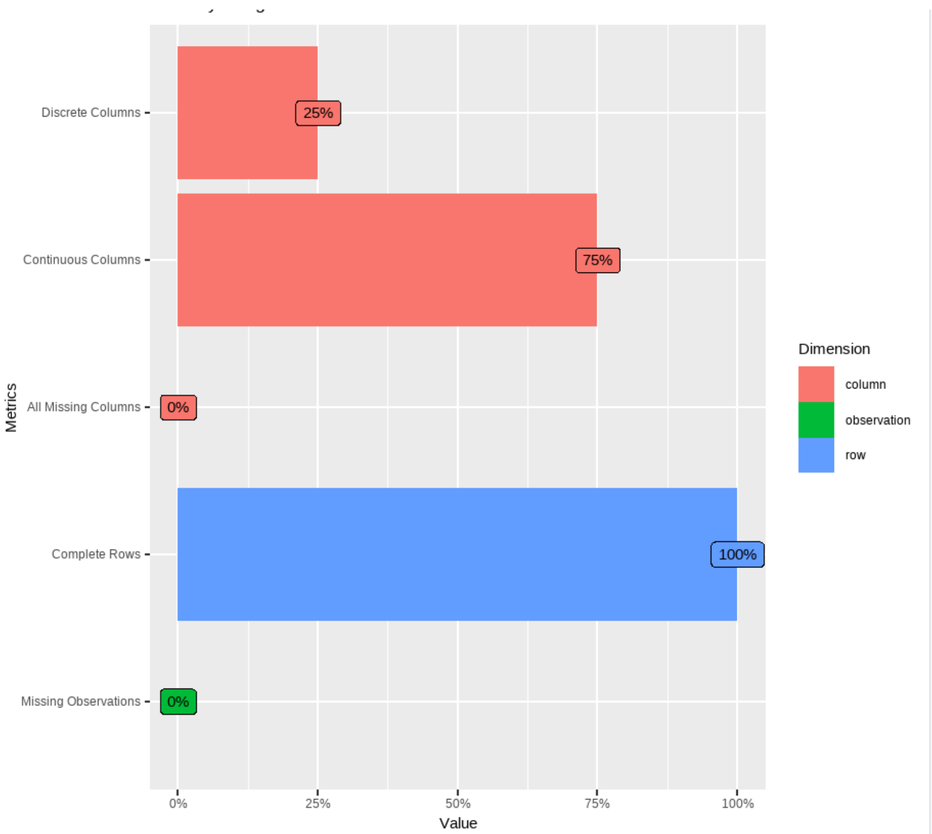


Figure 1.2: Observation on our Data Set

1.3 Checking for normality.

There are many ways to check the normality both statistically and visually

- Statistically
  - Shapiro wilk test.
- Visually.
  1 Boxplot.
  2 qqplot.
  3 Histogram.
  4 Density plot(Distribution plot).

1.3.1 Statiscally

For this we have used only the shapiro wilk-test to check for normality.

Shapiro-wilk-test

The Shapiro-Wilk test is a hypothesis test that is applied to a sample with a null hypothesis is that the sample has been generated from a normal distribution.

Null Hypothesis(H0): population is normally distributed.
Alternative Hypothesis(H1): Population is not normally distributed.

- We chose Shapiro Wilk Test as it is the most powerful test when testing for a normal distribution. It has been developed specifically for the normal distribution and it cannot be used for testing against other distributions.
- From the above test statistics, we decide normality based on p-value.
- If the P-Value of the Shapiro Wilk Test is larger than 0.05, we assume a normal distribution (null hypothesis is accepted), on the other side we won't assume as a normal distribution (null hypothesis is rejected).

The p-value of the test mentioned above is less than< 0.02. Based on this, we may conclude that the null hypothesis was rejected and does not follow a uniform normal distribution.

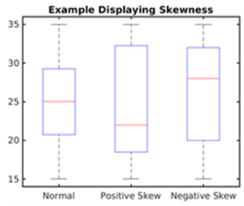
1.3.2 Visually

Box plot

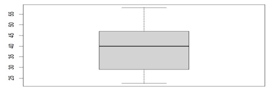
We use the box-plot to check whether the given data follows the normal distribution or not.

Q-Q plot

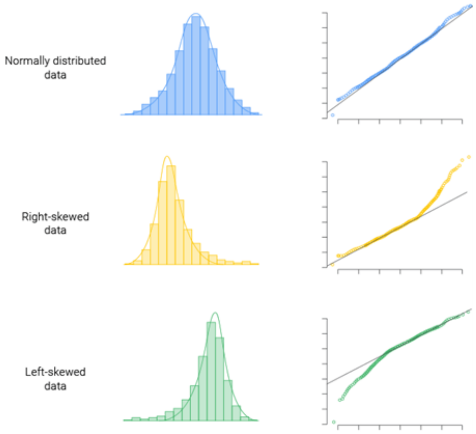
to check the normality



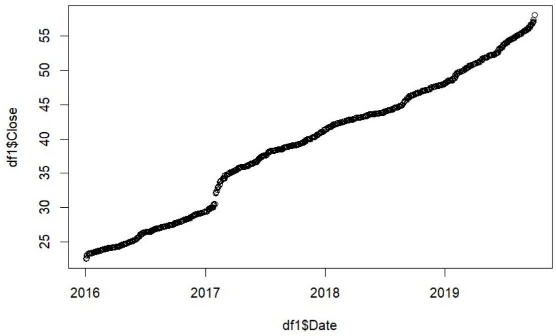
(a) fig-1: Box plot



(b) fig-2:Boxplot(Close price for apple)



(a) fig-3:Q-Q plot



(b) fig-4:Q-Q plot(Left skewed data concluded from 2.13b)

Histogram

check the normality (See Fig 1.5 Below)

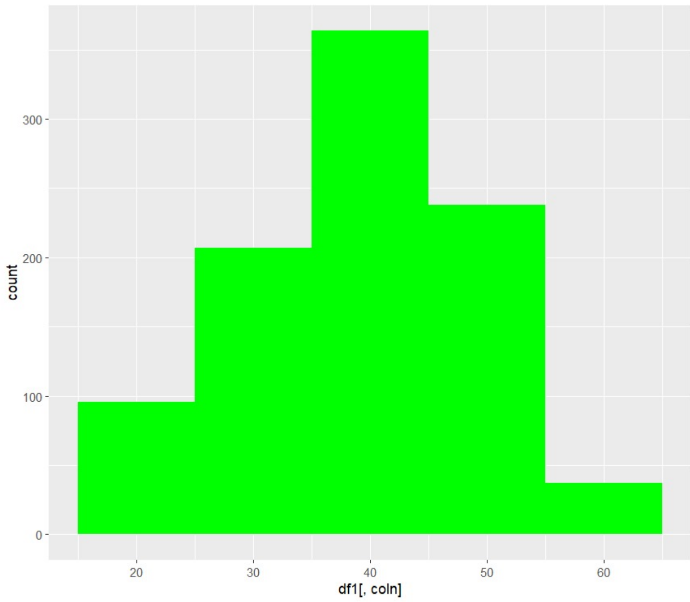


Figure 1.5: Histogram to visualize normal distribution and we can see that it is left skewed

Density plot

The density plot is used to check the normality of data.(See Fig 1.6)

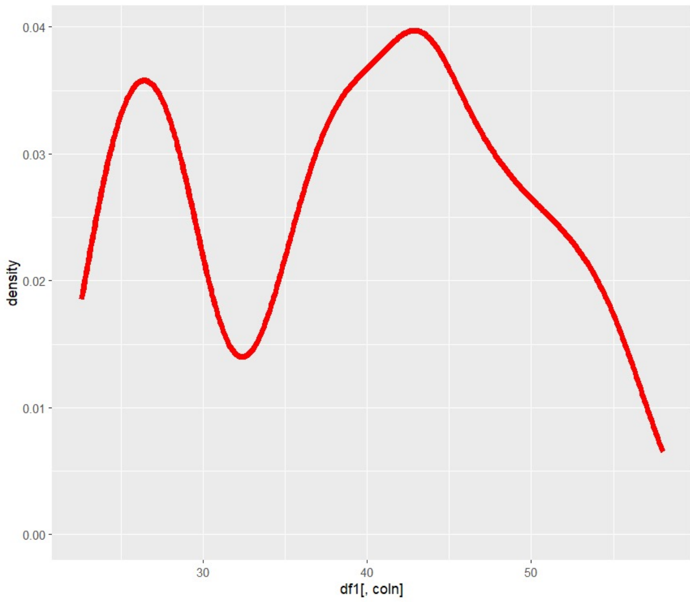


Figure 1.6: Density plot is used to check the distribution

1.4 Checking whether the data is Stationary or not

What is stationarity in data and why is it important in our project?
A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. Stationarity can be defined in precise mathematical terms, but for our purpose we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).
Why is this stationarity important in Time series data?
Because tools used in time series analysis and forecasting assume stationary. Like Arima which we used in our project. It has a huge impact on how data is perceived and predicted.
zoo series: A zoo series in R refers to a time-ordered, indexed data structure provided by the zoo package. This package is particularly designed for handling time series data, and it stands for ”Z’s ordered.observations.”

1.4.1 Rolling mean and Rolling standard deviation

Rolling mean:
For calculating the rolling mean we are converting the data into zoo series.For rolling mean calculate the mean of a rolling window of width 45 for a zoo_data object, aligning the window to the right and filling missing values with NA. Code : rollapply(zoo_data, width = 45, align = ”right”,FUN = mean,fill=NA) width: size of the window. align=’right’ : Align

means the window moves right side.

Rolling standard deviation:

For calculating the rolling standard deviation we are converting the data into zoo series.The rolling standard deviation for a time series, calculated with a window size of 45 and aligned to the right.

In time series analysis, a stationary time series is one whose statistical properties, such as mean and variance, do not change over time.

After calculating the rolling mean and rolling standard deviation we can visually say whether the data follows stationarity or not.

1.4.2 Visually

Plot the rolling mean and rolling standard deviation with the original time series. This can be done to visually assess if these values vary significantly over time and we from that we can say whether the data follows stationarity or not. A stationary time series is one whose statistical properties, such as mean and variance, do not change over time and for that graph looks like rolling mean and rolling std coincides each other.

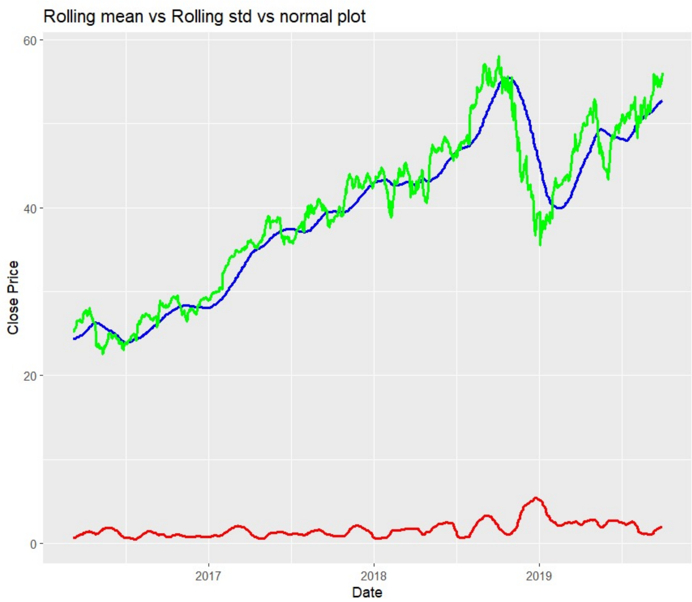


Figure 1.7: fig describes about the data is non stationary

1.4.3 statistical test

- Augmented Dickey-Fuller (ADF) Test.

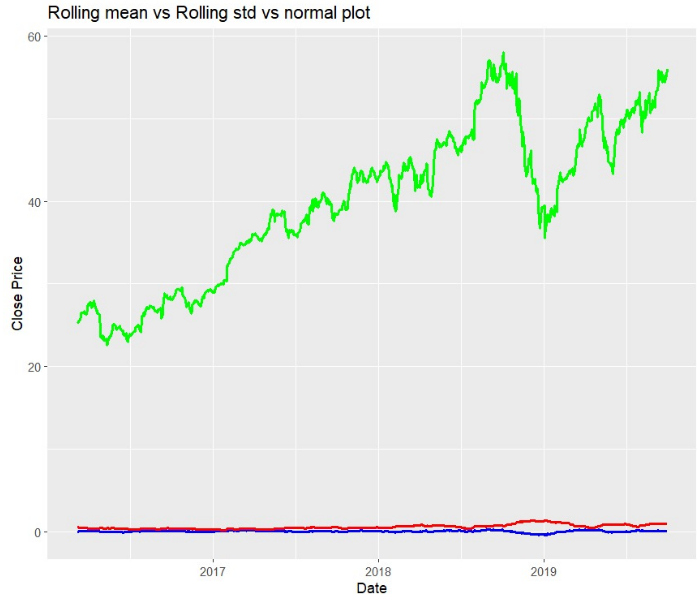


Figure 1.8: fig describes about the stationary data

Augmented Dickey-Fuller (ADF) Test:

- ADF test is also called as unit root test.
- A unit root is a feature of some stochastic processes (such as random walks) that can cause problems in statistical inference involving time series models.

ADF test is conducted with the following assumptions:

- **Null Hypothesis (HO):** Series is non-stationary, or series has a unit root.
- **Alternate Hypothesis(HA):** Series is stationary, or series has no unit root.
- The ADF test expands the Dickey-Fuller test equation to include a high-order regressive process in the model.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \dots$$

Figure 1.9: mathematical expression

Interpreting ADF Test Results:

The ADF test provides a test statistic and a p-value. The key interpretation points are:

- **Test Statistic:** If the test statistic is less than the critical values, you may reject the null hypothesis of non-stationarity.

- **p-value:** If the p-value is less than a chosen significance level (e.g., 0.05), you may reject the null hypothesis i.e. It does not have a time-dependent structure.

Applying ADF Test to Check Stationarity :

- ts() used for generating time series object from a data frame and stored in tse and then perform adf.test(tse).
- **ADF Test (adf.test):** Test we used to check for the presence of a unit root in univariate time series data, which is indicative of non-stationarity. The result is stored in the variable adf_result.

1.4.4 In our case we found out our data is non-stationary :

- **Differencing:** If the ADF test suggests non-stationarity, consider differencing the time series. Differencing involves subtracting each value from its preceding value.
- For this we use diff() function.
- Retest: Conduct the ADF test again on the differenced series to check for stationarity.
- **Further Transformations:** If differencing does not render the series stationary, consider other transformations or adjustments, such as taking the logarithm.
- For non stationary data the the rolling mean and rolling std never coincide and refer the fig 1.7

1.4.5 How we made our data stationary.

- You can proceed with auto-regression analysis and time series modeling.
- Consider differencing: If the time series is already stationary, you may not need to take further steps. However, in some cases, differencing may still be beneficial to remove trends or seasonality.
- For the stationary the rolling mean and rolling std coincide .once refer fig1.9

Model Selection: Once stationarity is achieved, proceed with selecting an appropriate auto-regression model (AR, ARIMA, etc.) based on further analysis, ACF/PACF plots, and model diagnostics.

1.5 What is Decomposition Function ?

The decompose function is used to decompose a time series into its constituent components which are trend , Seasonal , Reminder(Random/Residual).The Required library is stats.

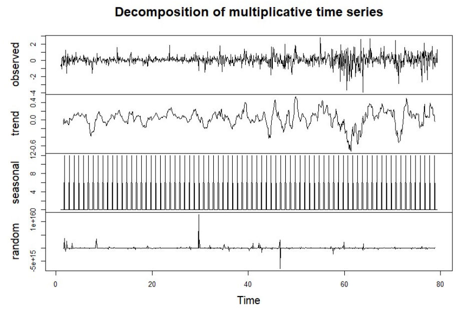
- **Trend Component:**
  Represents the long-term movement or trend in the data. It captures the overall direction in which the time series is heading.
- **Seasonal Component:**
  Captures recurring patterns or seasonality within the data. Seasonal components represent regular fluctuations that occur at consistent intervals.
- **Remainder Component (Random or Residual):**
  Accounts for the residuals or unexplained variability in the data after removing the trend and seasonal components. It represents the noise or random fluctuations in the time series.

1.5.1 Why to use decompose ?

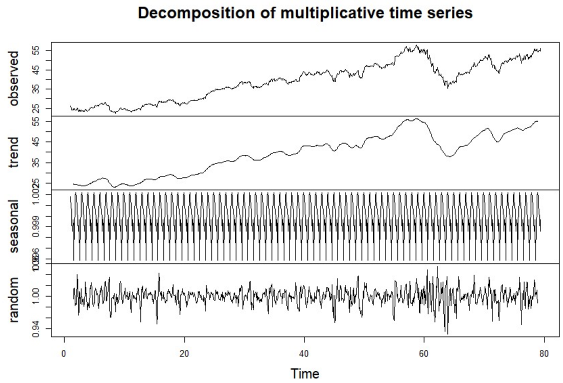
- Understanding underlying patterns.
- **Pattern Visualization:** Visualization of individual components allows for a clearer understanding of how each contributes to the overall behavior of the time series.
- **Forecasting (Future Predicting):** By isolating the trend and seasonality components, you can model and predict each separately. This leads to more accurate forecasts compared to modeling the entire time series as a single entity.
- **Modeling Stationarity:** Some time series models, like ARIMA (Auto Regressive Integrated Moving Average), require stationarity for accurate predictions. Decomposition can help transform non-stationary time series into stationary components, making them more amenable to modeling.
- **Diagnostic Checking:** Decomposition is part of the diagnostic process in time series analysis. Analysts can examine the residuals (remainder component) for patterns, autocorrelation, or heteroscedasticity, helping to identify any model misspecification.

1.6 Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF):

Partial Auto Correlation Function (PACF)
PACF stands for Partial Auto-Correlation Function, and PACF plots are graphical tools used in time series analysis to assess the direct relationship between a data point and its lags, it is a commonly used tool for identifying the order of an autoregressive model.
Auto Correlation Function(ACF):
The Autocorrelation Function (ACF) in R is a statistical tool used to measure the correlation between a time series and its own lagged values. Provides insights into the temporal dependencies(impact of previous behaviour on current behaviour) within a time series.



(a) Decomposition plot for non-stationary



(b) Decomposition plot for stationary

Required library is forecast.

1.6.1 Purpose of Autocorrelation Function (ACF):

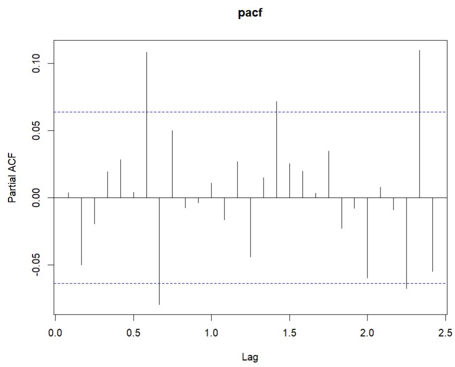
- **Detecting Seasonality:**
  Peaks in the ACF plot at regular intervals may indicate the presence of seasonality in the data.
- **Model Selection:** ACF is often used with the Partial Autocorrelation Function (PACF) to determine the appropriate lag orders for auto-regressive models (e.g., ARIMA). Peaks in the ACF and PACF plots can guide the selection of lag orders and moving average orders for the autoregressive component of a model.
- **Stationarity Analysis:**
  ACF is useful for observing the stationarity of a time series. A slowly decaying ACF may indicate non-stationarity, suggesting that stationarity measures should be taken.

1.6.2 Interpretation and Conclusions:

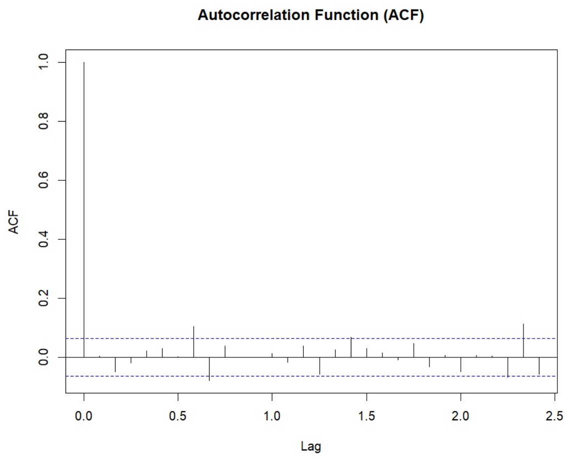
- **Significant Peaks:** Significant peaks in the ACF plot at specific lags indicate strong autocorrelation at those lags. These peaks provide insights into the repeating patterns or cycles in the data.

- **Seasonal Patterns:** If there are periodic peaks at regular intervals ,it may indicate the presence of a seasonal component in the stock prices.
- **Model Lag Orders:** Peaks in the ACF plot can be used to determine potential lag orders for auto-regressive models. For instance, if there is a significant peak at lag 1, it suggests that the current stock price is correlated with the previous day's price.

In summary , ACF is a valuable tool for understanding the temporal structure of time series data, identifying patterns, and guiding the selection of lag orders for forecasting models. Analysing the ACF plot is an essential step in the broader process of time series analysis and modelling, especially in the context of stock price predictions.



(a) Partial Auto Correlation Function (PACF)



(b) Auto Correlation Function plot for stationary

Chapter 2
Model Building

2.1 pth order auto regression analysis

We use auto regression analysis to predict the future value by studying past values and then to predict the future values based on the past values.The order of auto-regression, denoted by ”p,” specifies the number of lagged values considered in the model which will be used to predict the future values.

2.2 Finding order using ACF and PACF plots

To get the order first we will first check whether the data is stationary or not if not we convert the data into the stationary data and then plot the PACF plot from that we will find the order of the regressive model.Here we will use PACF plots to get the order of the model and ACF plot to get the correlation between the values.

2.3 what is Auto-ARIMA ?

Auto-ARIMA is a algorithm that automatically selects the optimal parameters for an ARIMA model to train and predict the future data and it gives three values as outputs p,d,q which are used to train the ARIMA model

- (p) This parameter represents the number of lagged observations included in the model which is same as the order of the model.
- (d)This parameter denotes the number of times the data needs to be differenced to achieve stationarity. Stationarity means that the statistical properties of the time series, such as mean and variance, remain constant over time.
- (q) This parameter signifies the number of lagged forecast errors included in the model. The moving average component helps capture the impact of past forecast errors on the current observation.
- By using the auto-arima we will get the above classes and these will be used to train

ARIMA model

2.4 Training ARIMA Model

- Depending on the value of the parameter-d we can check whether data is stationary or not if not we will convert it to stationary and train the model. If the original time series exhibits a trend or seasonality, it may be non-stationary.
- The differencing step involves subtracting each observation from its previous one (first-order differencing), or, in the case of seasonality, subtracting observations at regular intervals.
- The aim is to remove trends or patterns that can affect the model's performance, making the time series stationary.
- The number of differencing steps (d) is determined by Auto-ARIMA based on how many times the data needs to be differenced to achieve stationarity. After differencing, the resulting series is used for model parameter estimation and training for the ARIMA model.

2.5 Predicting future values

Now when it comes to the predicting the future values we will be doing prediction in such a way that initially we predict the very next day value and then use the overall the past data to predict the next to next day data and it goes all the way until end it means that we don't predict whole data at a time we predict it day by day.

2.5.1 Analysis and predicting values for different companies

Here we will be analysing 12 companies those are ”apple”,”TCS”, ”tesla”, ”drreddylab”,”abott”, ”IBM”, ”nvdia”, ”google”, ”accenture”, ”microsoft”, ”amazon”, ”Hp”

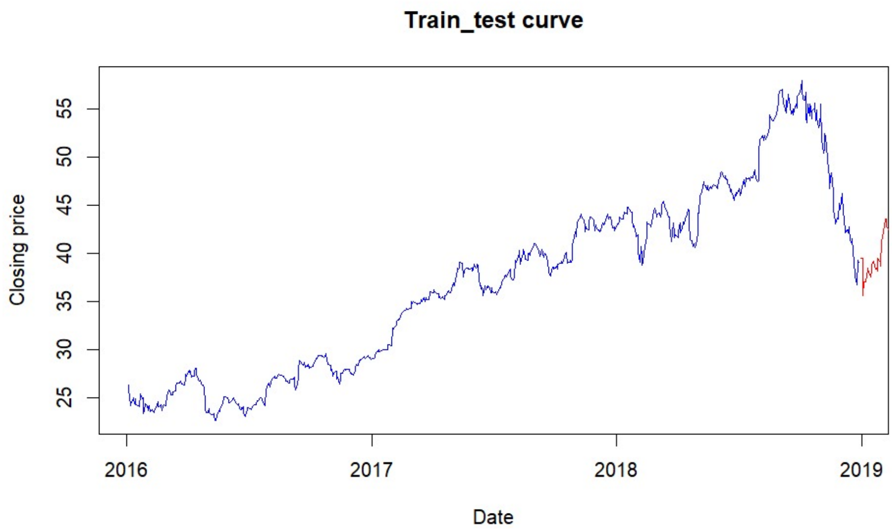
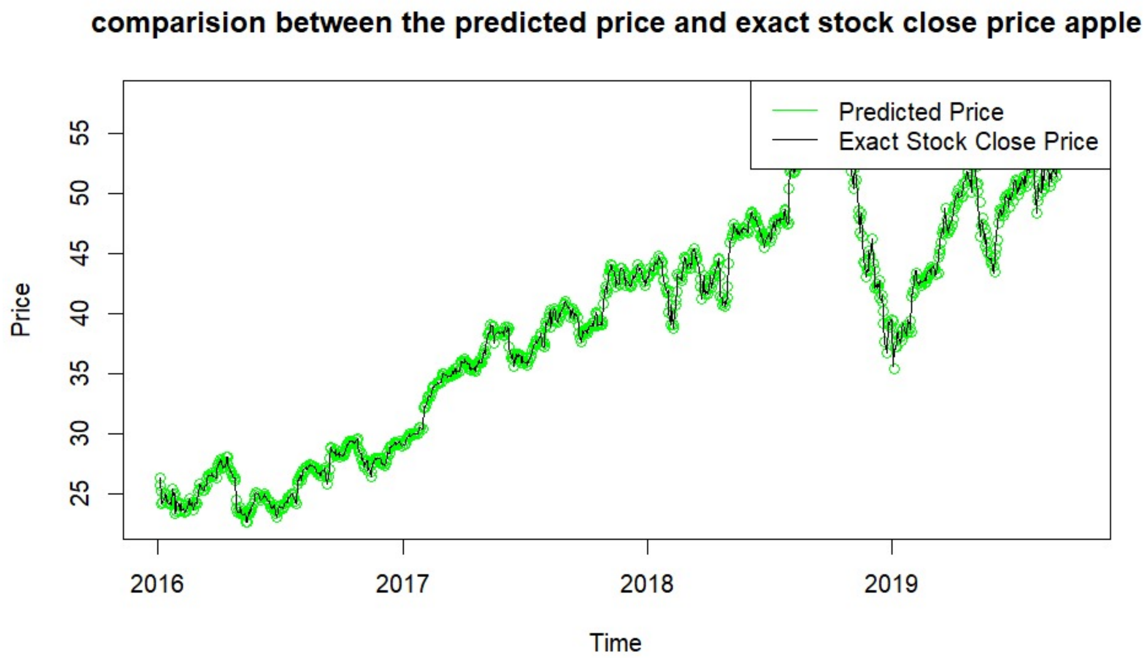
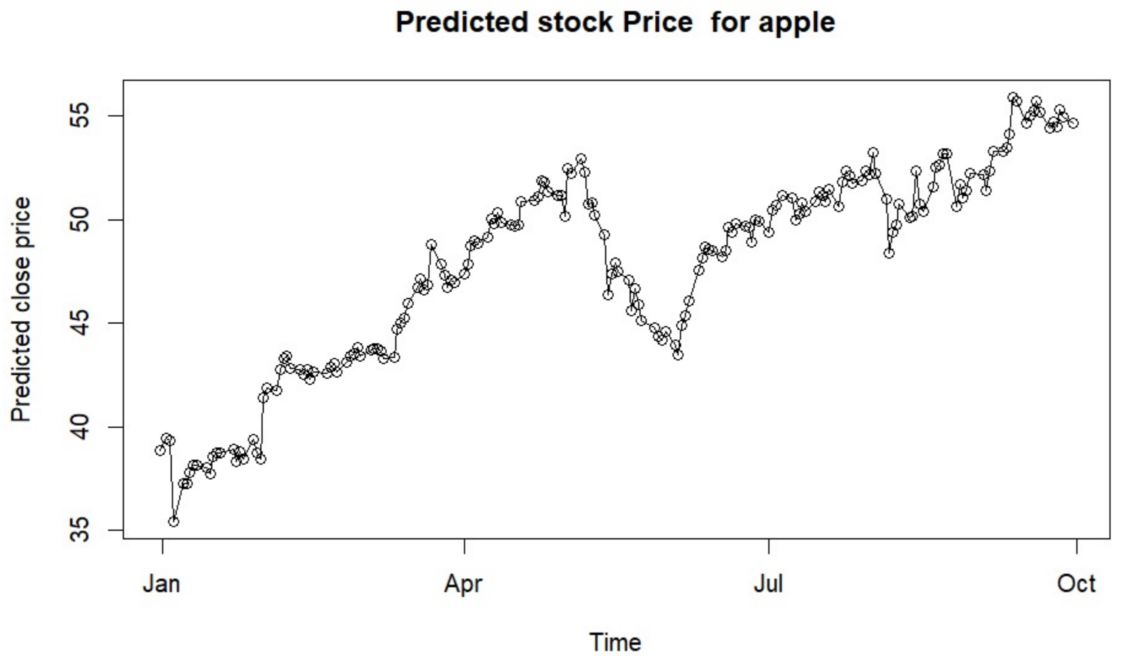


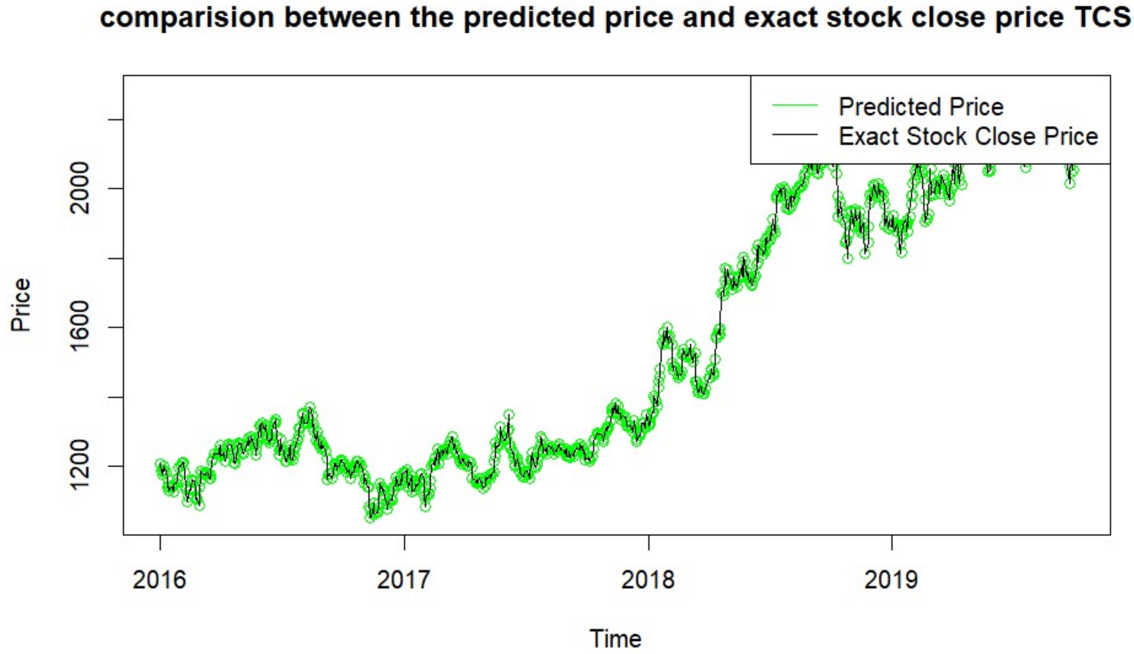
Figure 2.1: test train curve of the given data and graph of all companies data we will below this



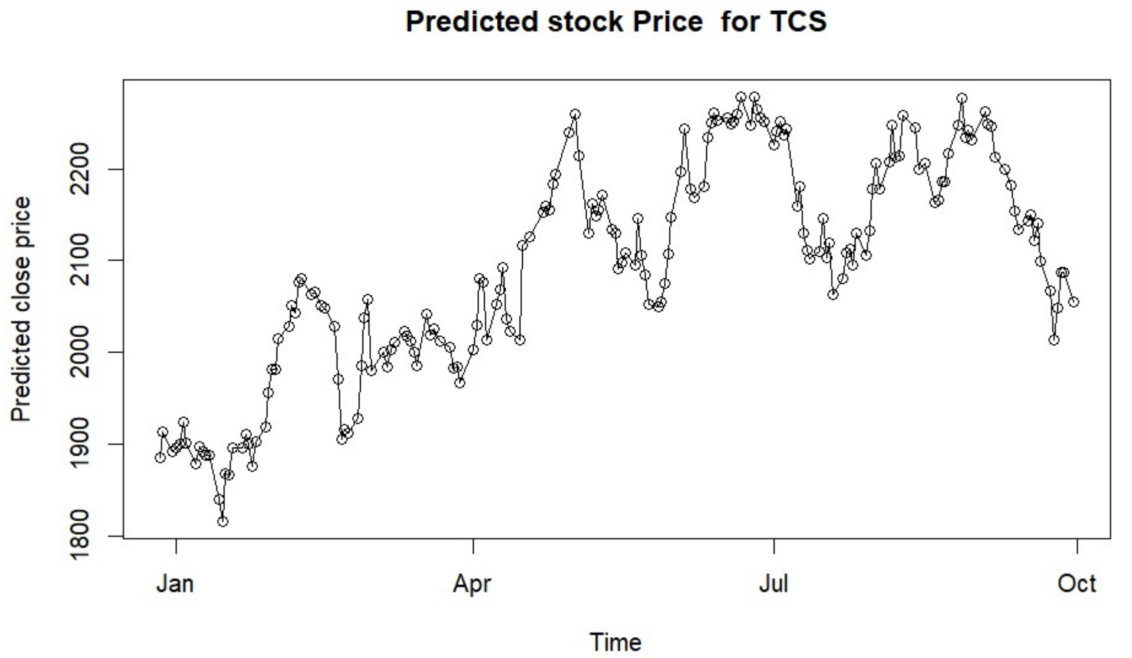
(a) comparision between predicted and exact stock close price



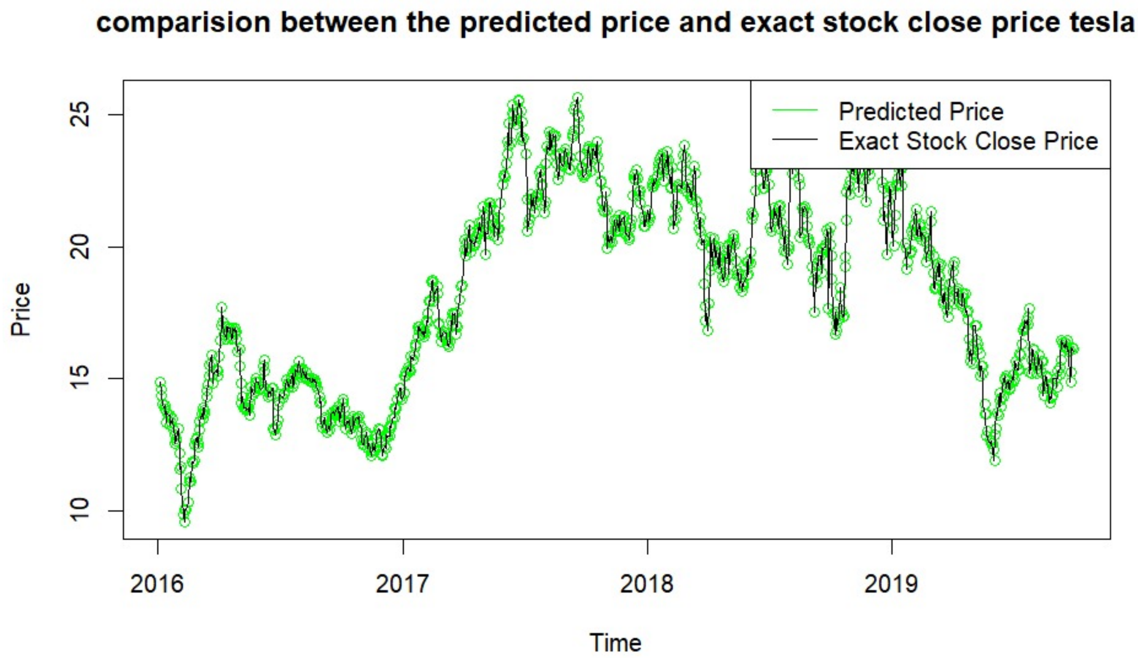
(b) Final predicted price of apple company



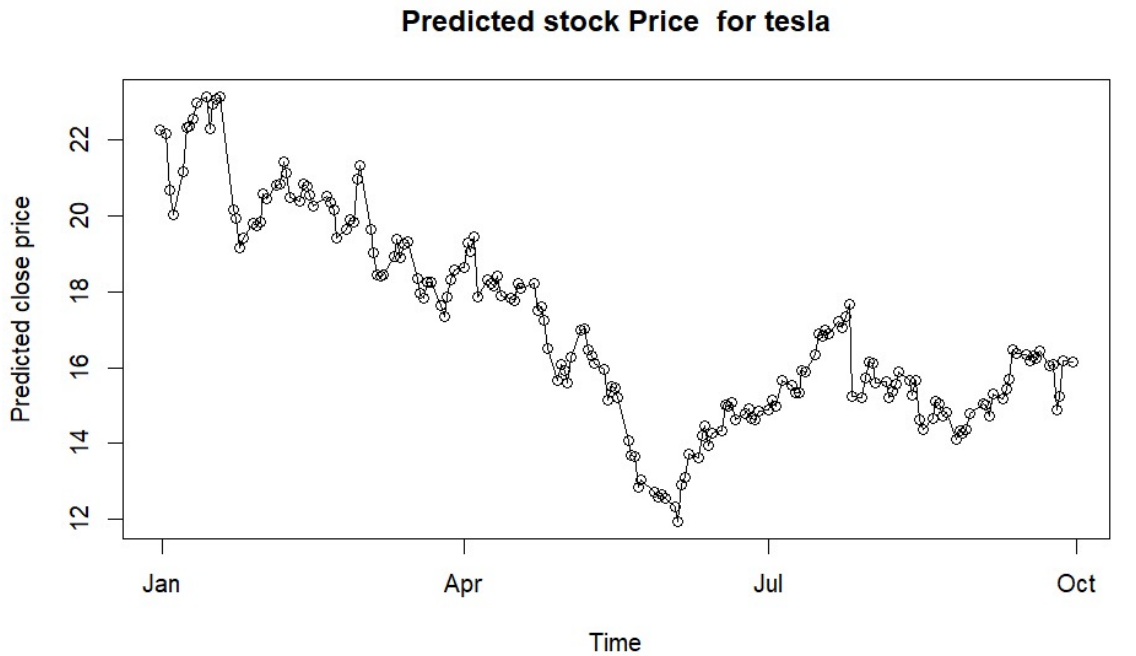
(a) comparision between predicted and exact stock close price



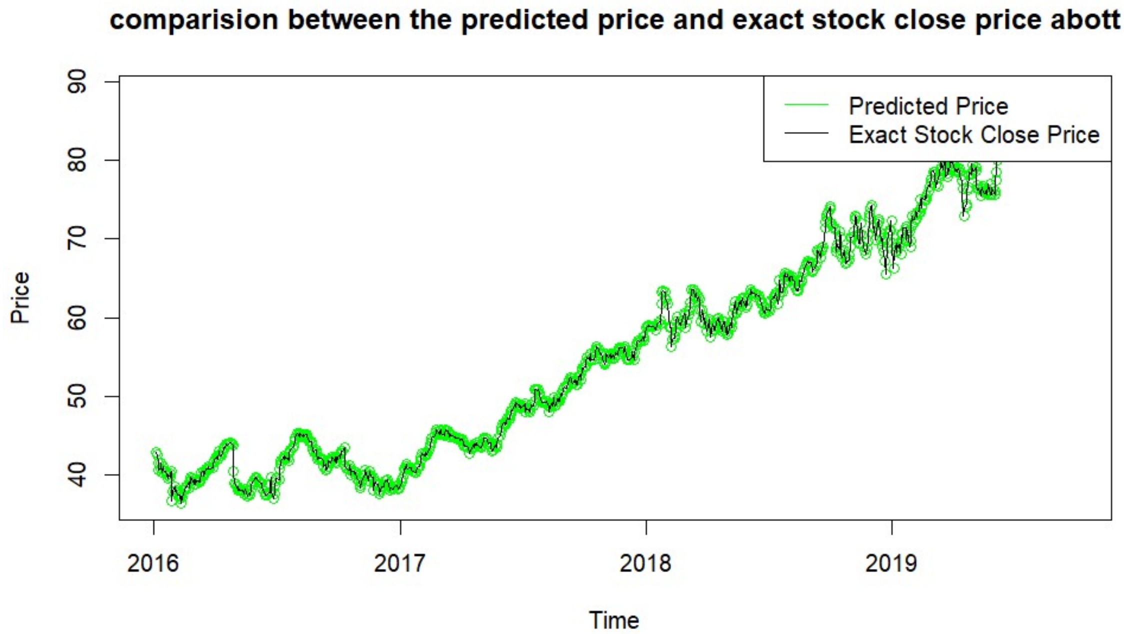
(b) Final predicted price of apple company



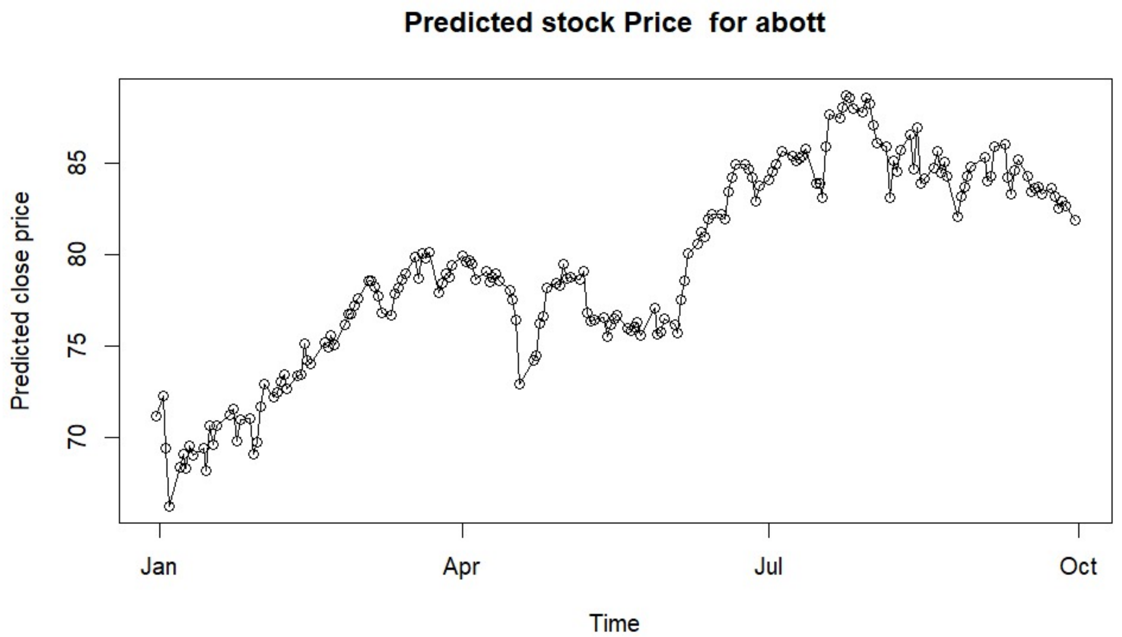
(a) comparision between predicted and exact stock close price



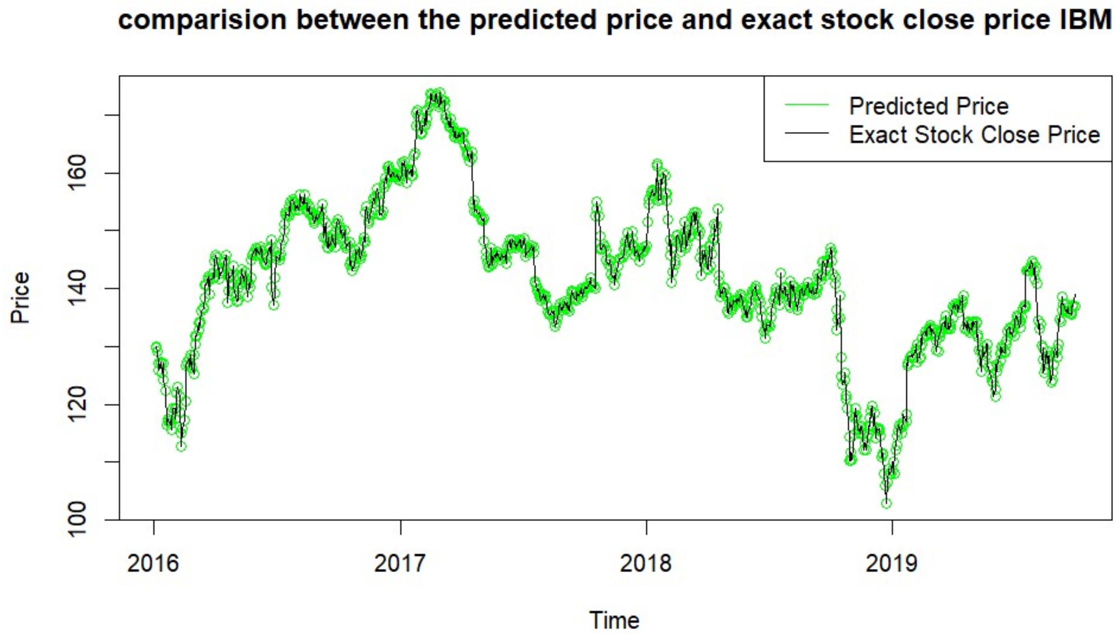
(b) Final predicted price of tesla company



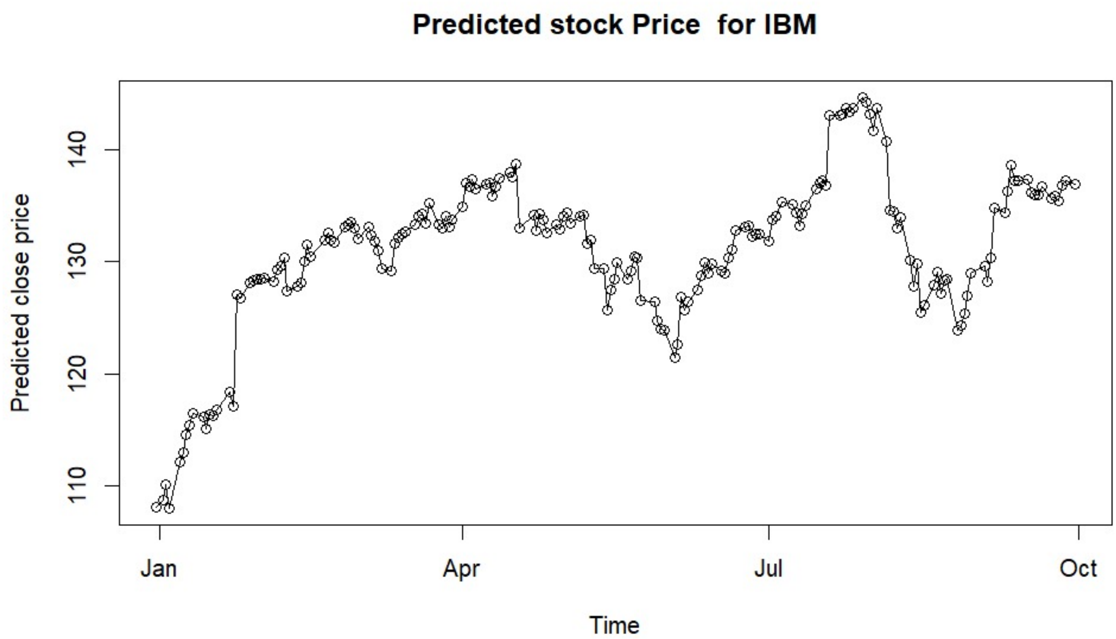
(a) comparision between predicted and exact stock close price



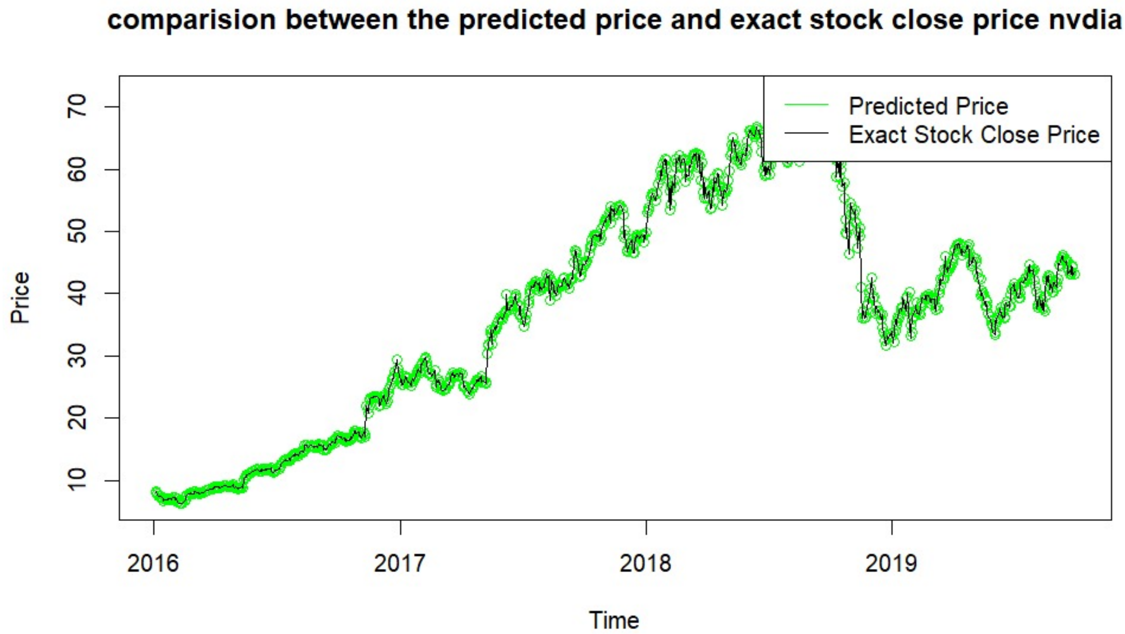
(b) Final predicted price of abott company



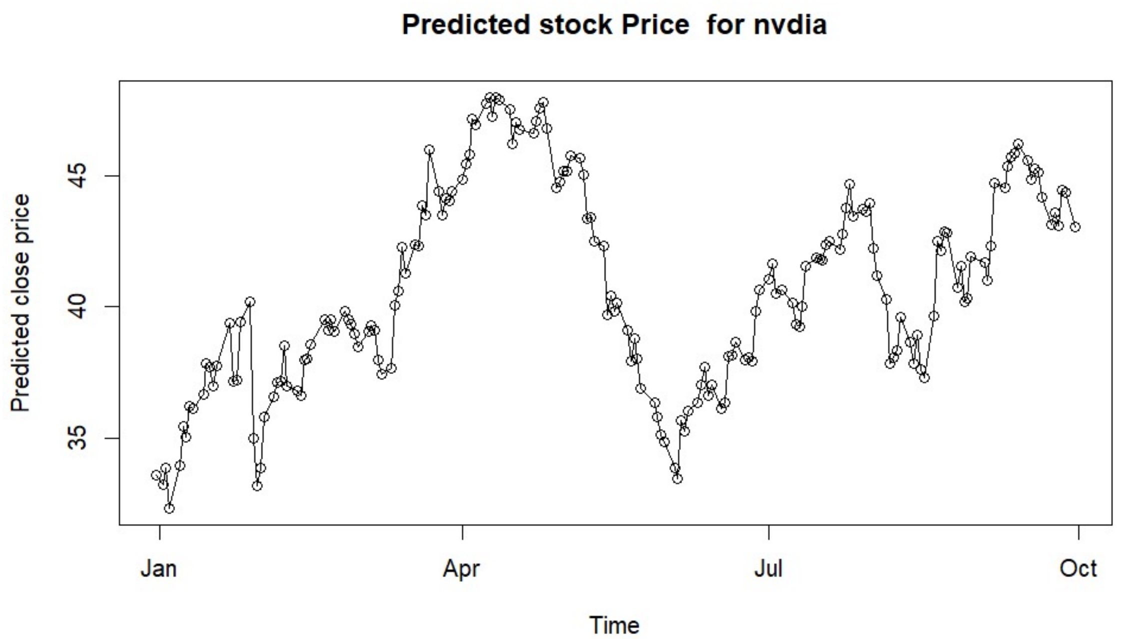
(a) comparision between predicted and exact stock close price



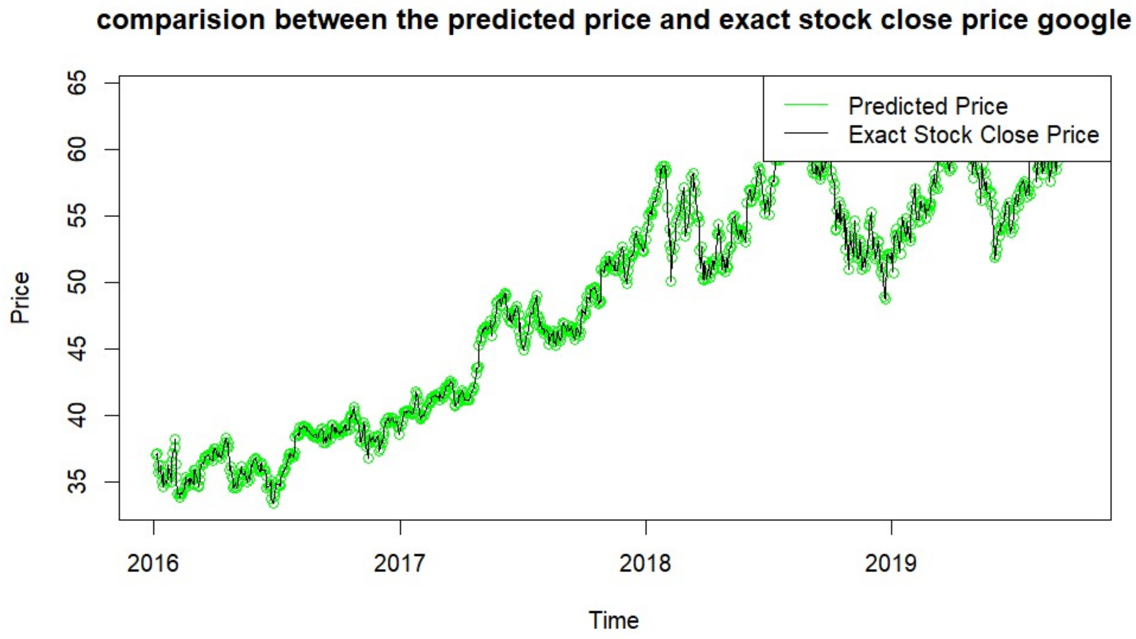
(b) Final predicted price of IBM company



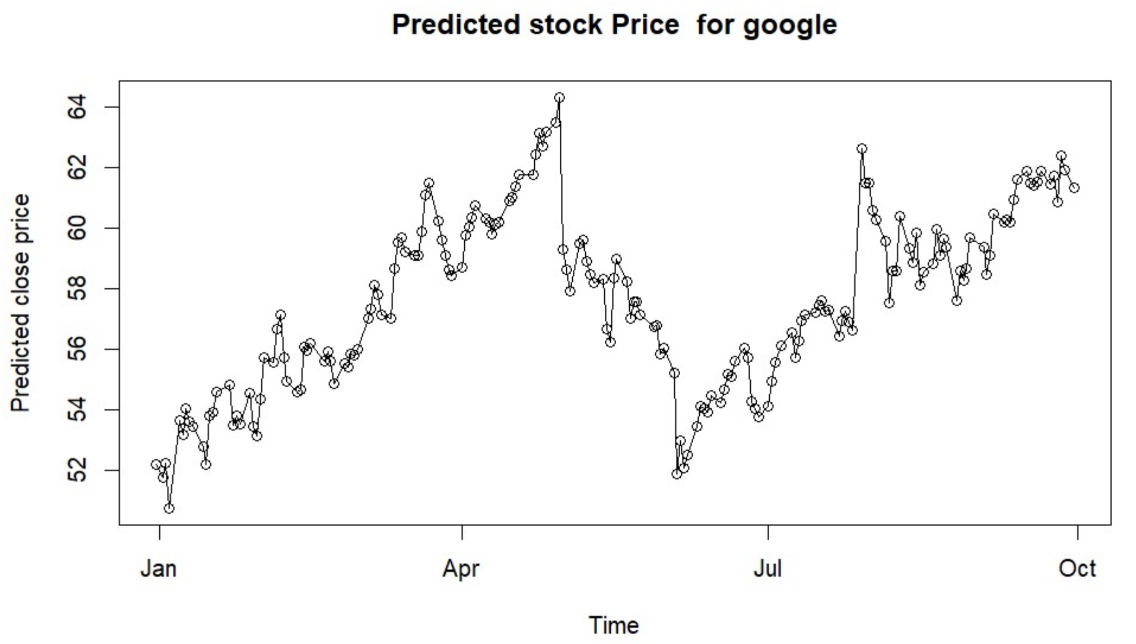
(a) comparision between predicted and exact stock close price



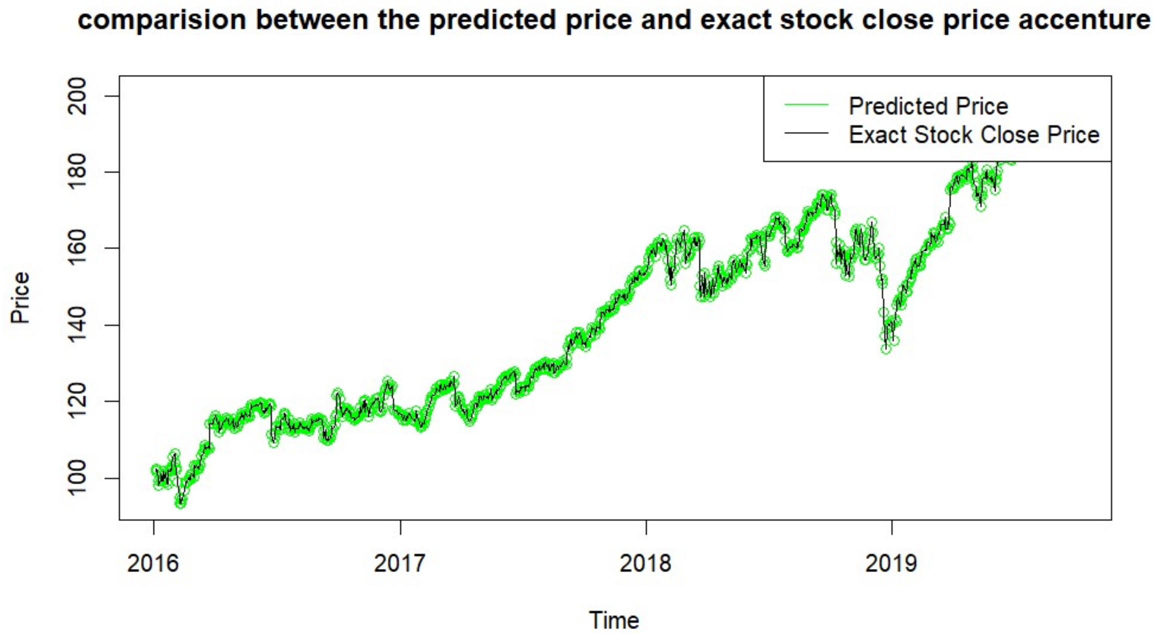
(b) Final predicted price of Nvedia company



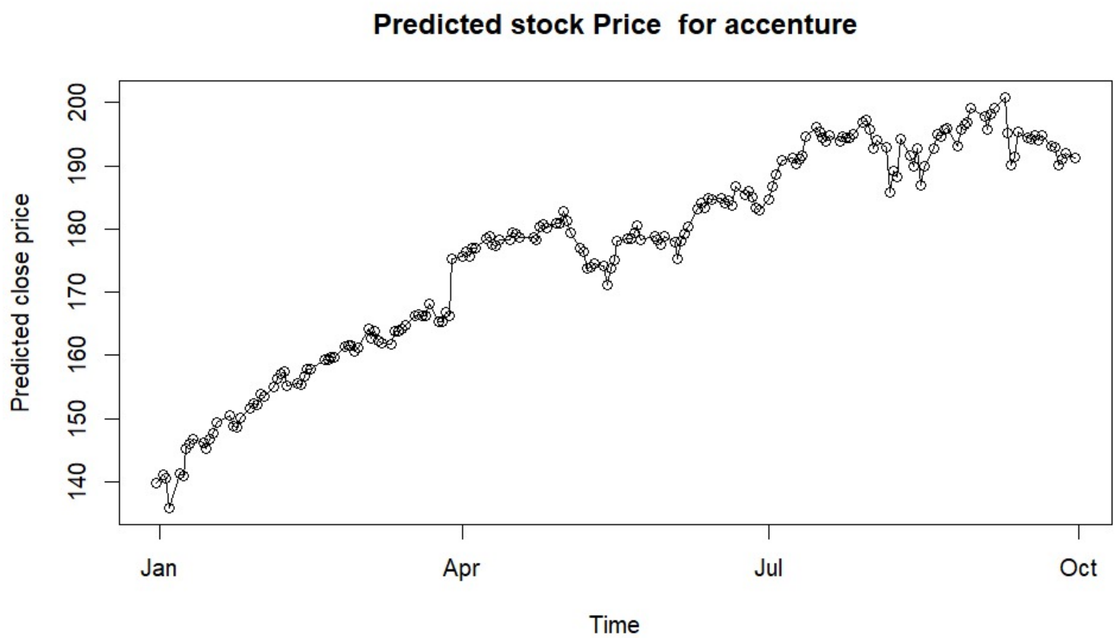
(a) comparision between predicted and exact stock close price



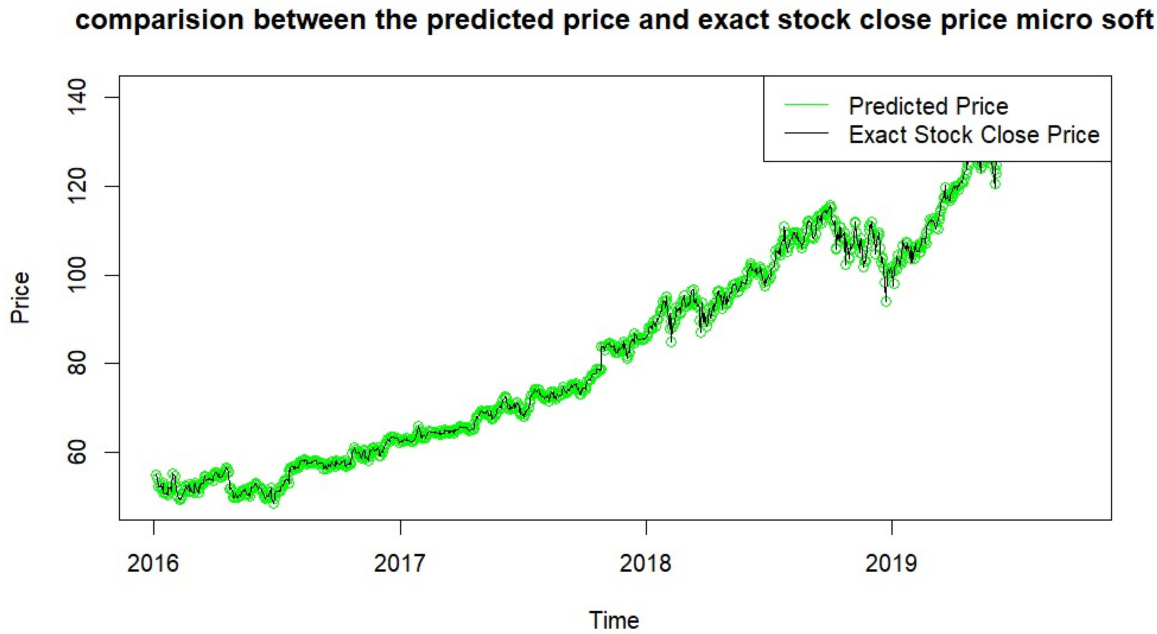
(b) Final predicted price of google company



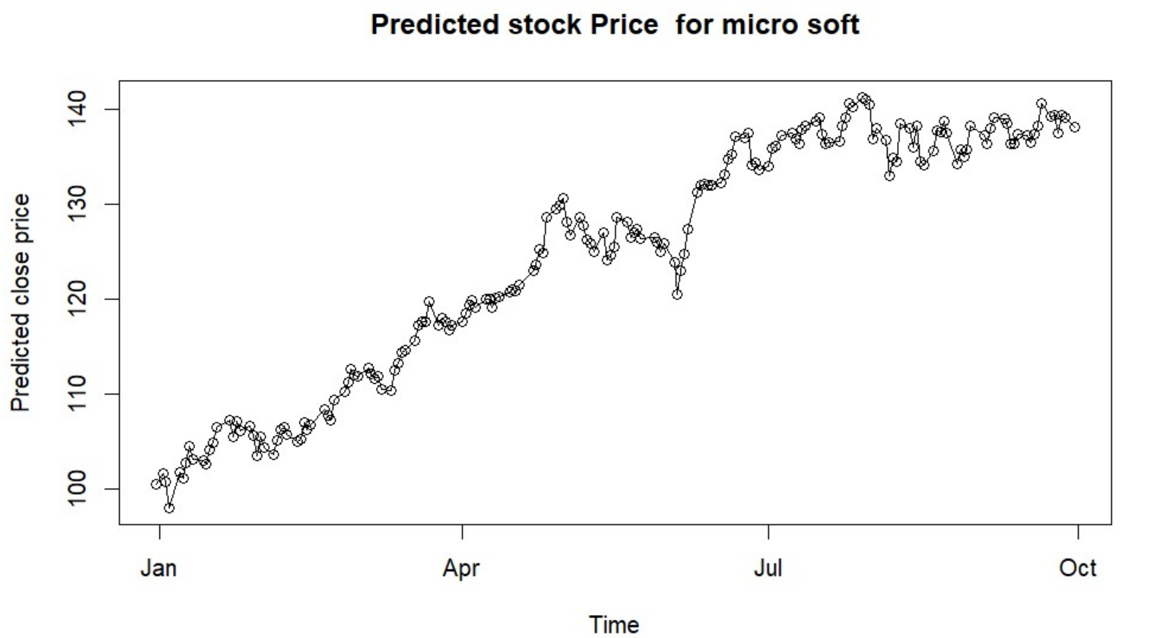
(a) comparision between predicted and exact stock close price



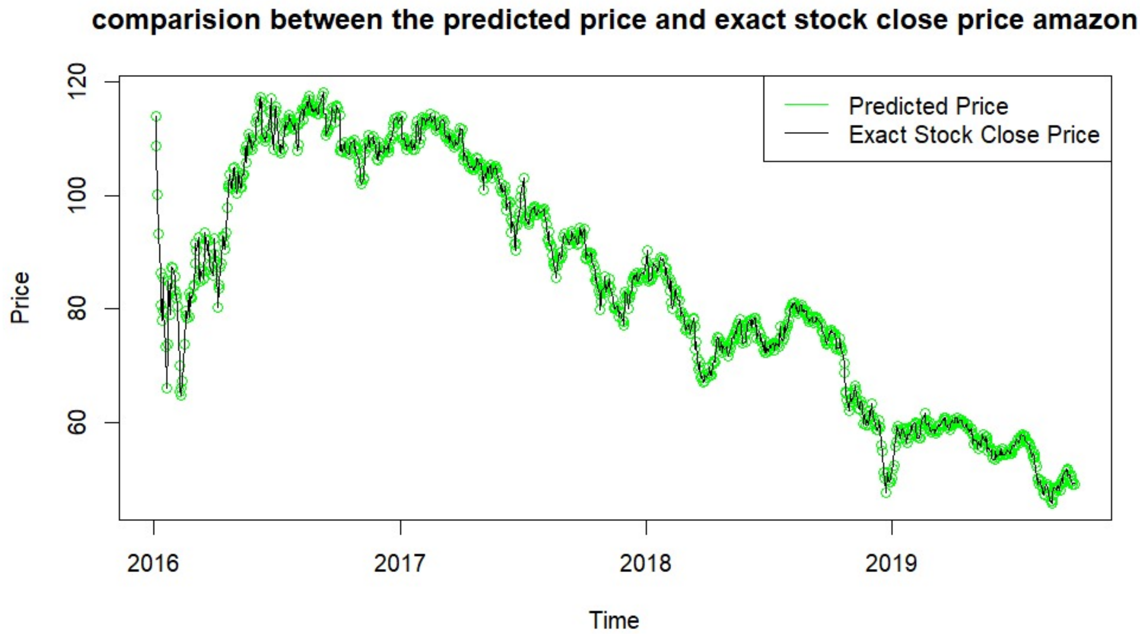
(b) Final predicted price of accenture company



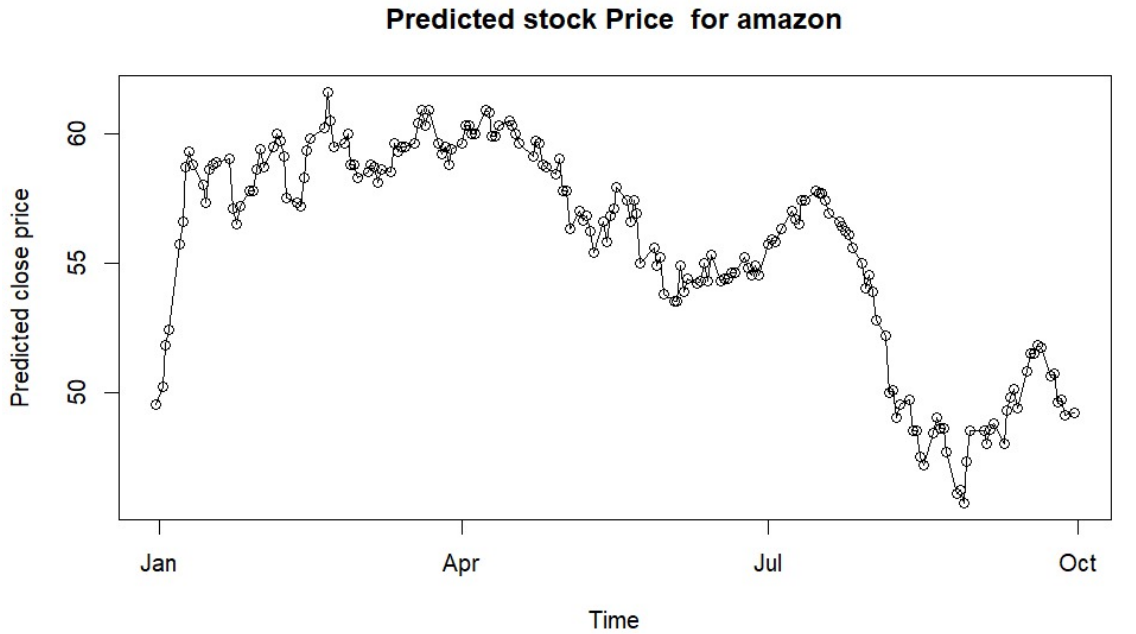
(a) comparision between predicted and exact stock close price



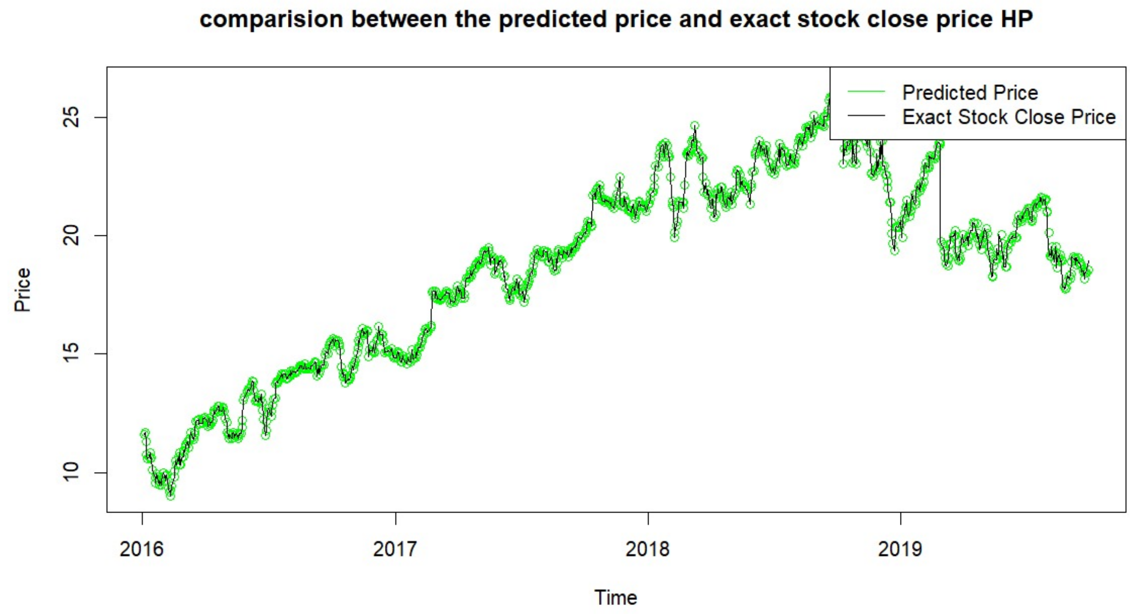
(b) Final predicted price of microsoft company



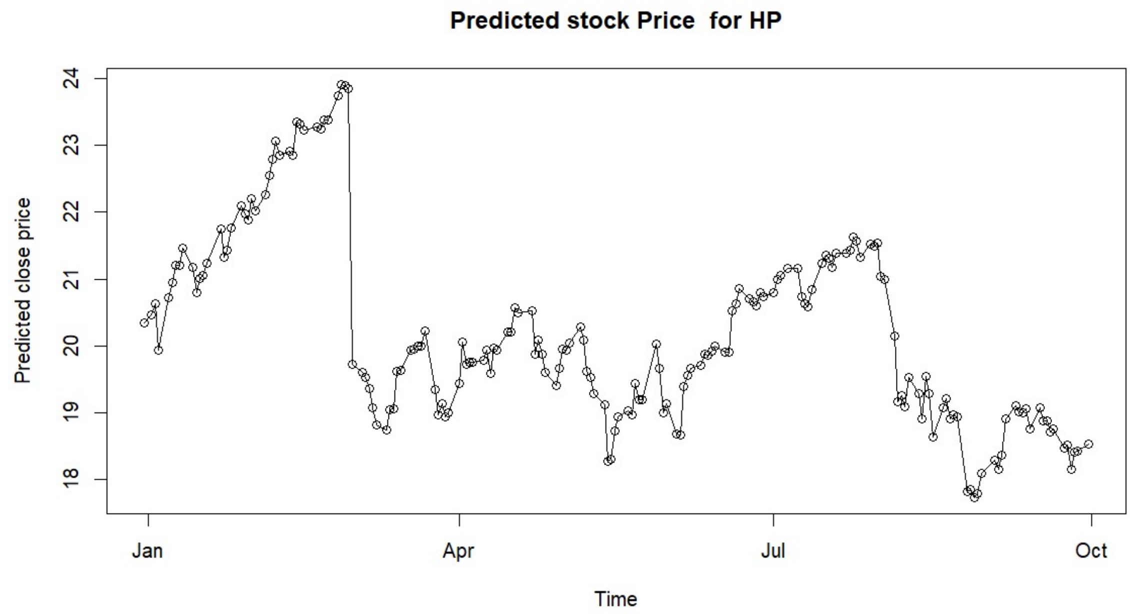
(a) comparision between predicted and exact stock close price



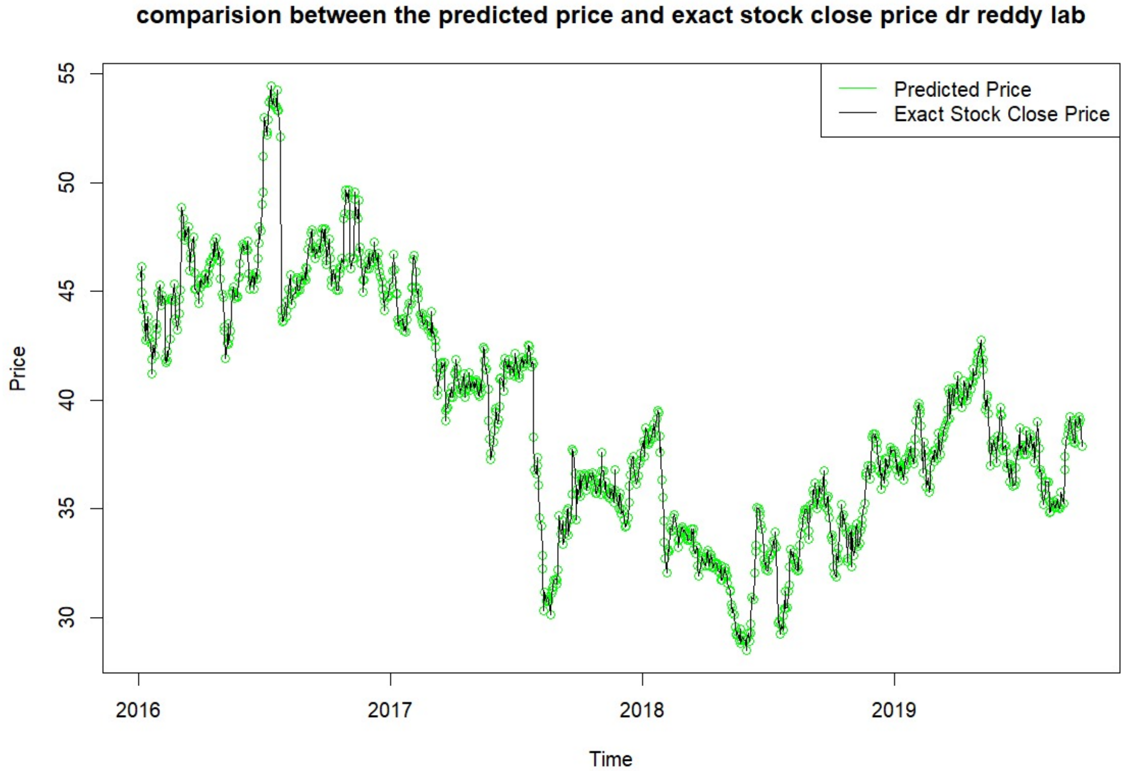
(b) Final predicted price of amazon company



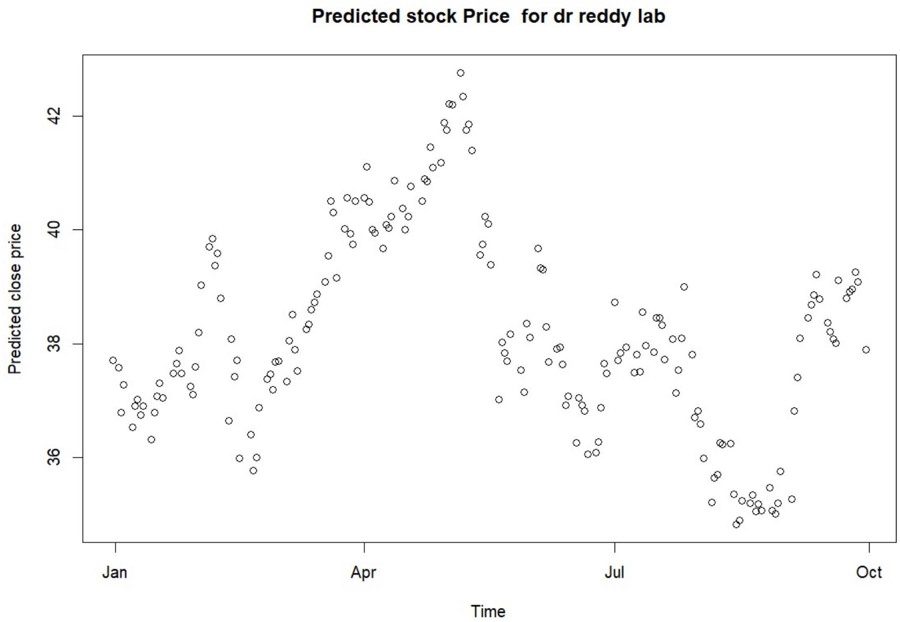
(a) comparision between predicted and exact stock close price



(b) Final predicted price of Hp company



(a) comparision between predicted and exact stock close price



(b) Final predicted price of Dr reddy's lab company

Chapter 3

Conclusion

This project focuses on the **close feature** of the dataset as the target variable(**as the stock prediction depends on close feature values**) and all the calculations and results are based from the analysis of the close column.Now using a sophisticated Auto-Regression Analysis Model to analyze time series data, specifically looking at stocks from 12 different companies. We started by carefully exploring the data through Exploratory Data Analysis (EDA) to understand the underlying patterns in the stock market.

Next, we went through a thorough Data Pre-processing phase to ensure our model's accuracy and reliability. This step involved refining the dataset, making it more robust for analysis. With a clean dataset, we used Auto-ARIMA modeling, a powerful method that considers the probability value 'p' to predict future stock values. The model was trained on historical stock records to adapt and capture the dynamic nature of the stock market.

The key part of our analysis involved examining different probability values for 'p.' We created a table highlighting the optimal 'p' values for each company based on the Auto-ARIMA model's predictions. This detailed approach helped us identify the best-performing companies.

Our results go beyond simple predictions; they offer valuable insights into market dynamics, helping stakeholders make informed decisions. The versatility of our model is evident as it predicts three different attributes for all 12 companies. This comprehensive approach ensures a deep understanding of market trends, giving users the foresight needed for successful stock trading.

As a final comment, We can observe that from the predicted values we can see from chapter-2 the predicted stock images of the MICROSOFT and ACCENTURE Companies are stocks that are increasing and we can invest in that stocks for the better results as part of stock market and we will provide thier covariance matrices and we also compare the different p values moving average values in a table and for a single company like apple we will change the p values and table out the accuracy values.During the analysis we got p value as two for both the above companies and we got those values from arima model. Here while predicting the stock we used the close column feature as

Company Name	Order value(p)	Error value(RMSE)	Moving Average
Apple	2	0.3757646	3
TCS	2	13.10417	2
Tesla	0	0.2453323	0
Dr reddy's lab	1	0.277	3
Abott	2	0.4610786	2
IBM	0	0.7818294	0
Nvedia	2	0.4956099	2
Google	4	0.43	1
Accenture	2	0.87	2
Microsoft	2	0.7382	3
Amazon	0	0.3465175	0
HP	0	0.1928274	0

Table 3.1: All twele companies with thier order,Root mean square error and moving average value

P value	RMSE(error)
2	0.3757
3	0.454712
0	0.7118294
1	0.5310786

Table 3.2: P values comparison of the Apple company and thier respective errors

an estimate.

```
Best model: ARIMA(2,1,2)

               Close Closelag1  Closelag2
Close     165069.124 -11487.43  -8993.024
Closelag1 -11487.431 165069.85 -11479.642
Closelag2  -8993.024 -11479.64 164855.434
```

Figure 3.1: Pth(Here P=2) order covariance matrix for accenture company

```
Best model: ARIMA(2,1,3)

               Close Closelag1  Closelag2
Close     165069.124 -11487.43  -8993.024
Closelag1 -11487.431 165069.85 -11479.642
Closelag2  -8993.024 -11479.64 164855.434
```

Figure 3.2: Pth (Here P=2)order covariance matrix for microsoft company

3.1 Individual contributions:

- **K Sonish** : Done Covariance matrix for pth order autoregression analysis
- **Yalakanti Eswar** : Done Exploratory data analysis and wrote function to find optimised pth-order value
- **V A V S Chakravarthy** : Done Arima model traning and testing
- **Adi Narayana Reddy M** : Done data preprossing and comprision of error of different p values.
- **Akhil sai sriram:** Done Arima model traning and testing