Analyzing and Forecasting using Tata Consultancy Services Stock Market Data

Submitted by:

Sampriti Dutta (211366)
Sankhadeep Mitra (211369)
Soumya Paul (211391)

Under the guidance of:

Prof. Amit Mitra
Professor, Department of Mathematics and Statistics, IIT Kanpur

Abstract

The stock market is a place where buying and selling of business stocks may be done easily. There is a stock index value for each stock exchange. The average value obtained by averaging many equities is known as an index. As a result, it is easier to represent the whole stock market and forecast the market's long-term trajectory. The equity market has a significant bearing on both the general populace and the nation's economy. Therefore, a successful stock trend prediction can reduce investment risk and increase profit. In this project, we forecast and visualize the results using the Time Series Forecasting approach. We'll base our predictions on a technical study of historical data and the ARIMA model. Since the Autoregressive Integrated Moving Average (ARIMA) model is well-known for being reliable, effective, and capable of making accurate short-term share market predictions, it has been widely employed in the fields of finance and economics.

Acknowledgement

We are delighted to give a project report on "***Analyzing and Forecasting using Tata Consultancy Services Stock Market Data***". Every accomplishment has received ongoing encouragement and counsel from excellent and noble minds to aid us in directing our efforts in the appropriate direction to bring the project to completion. We would like to thank our professor, **Prof. Amit Mitra**, for his consistent assistance and support during the project's completion. It would have been practically impossible to work on this project as a team and comprehend the practical component of the course *MTH517A: TIME SERIES ANALYSIS* without his invaluable instruction and enthusiasm. We are also grateful to all faculty members and seniors who provided assistance at various phases of the project. Finally, my heartfelt gratitude goes to our friends, who were always there for me when we needed them, no matter what.

Sampriti Dutta

Sankhadeep Mirta

Soumya Paul

Contents

1 Introduction 5

2 Exploratory Data Analysis 6

2.1 Dataset Description . . . 6

2.1.1 Datatypes of the variables . . . 7

2.2 Dataset Visualization . . . 7

2.3 Structural Break Analysis . . . 8

2.3.1 Chow's Test . . . 9

2.4 Train-Test Splitting . . . 9

3 Presence of time series components 10

3.1 Test for Randomness . . . 10

3.2 Test for Trend . . . 11

3.2.1 De-trending the data . . . 12

4 Model Selection 13

4.1 Test for Stationarity . . . 13

4.2 ACF and PACF plots . . . 14

4.3 Model Selection for Original Data . . . 15

5 Residual Analysis 15

5.1 Ljung-Box test . . . 16

5.2 Lagrange Multiplier Test: . . . 18

6 Forecasting 19

7 Conclusion 20

References 21

1 Introduction

A collection of marketplaces where stocks, bonds, and other types of securities are issued and traded over the counter as well as through different physical and electronic exchanges are together referred to as the stock market. One of the most crucial elements of a market economy is the stock market, which gives businesses access to cash by enabling investors to purchase shares of firm ownership. The stock market industry is continually evolving via a process of improvement. Investors must carefully prepare their strategies in light of the daily fluctuations it delivers in order to earn.

When predicting stock market data, it is assumed that current publicly available data has some predictive correlations to future stock returns. Because of the difficulties in navigating the extremely complex world of the stock market, stock trend forecasting is one of the most challenging tasks to complete in the financial sector. Investors in the stock market are constantly looking for a method that would limit their risk while ensuring simple profits by predicting stock movements. This encourages forecasting model developers to create new forecasting techniques.

Stock prices may be thought of as a discrete time series model that is based on a set of well-defined numerical data items gathered at successive points at regular intervals of time rather than being viewed as values that are created at random. Since it is crucial to find a model to evaluate stock price movements with sufficient information for decision-making, it is advised that utilising ARIMA to convert the time series is a better algorithmic method than just predicting since it produces more accurate findings. Prior to analysis, the Autoregressive Integrated Moving Average (ARIMA) Model transforms non-stationary data into stationary data. It is one of the most often used models for predicting data from linear time series.

R is a language and environment for statistical computing and graphics. For statistical programming and data analysis, data analysts typically use the R dialect. The R development core team now oversees the maintenance of the R language, which was developed by Ross Ihaka and Robert Gentleman at the University of Auckland in New Zealand. The ARIMA model may be used with R-Studio, a free and capable integrated programming environment for the R language.

In this project, we have primarily concentrated on the degree of accuracy in predicting stock prices for various sectors, which will help new investors understand the market and make an informed decision about investing in the stock market.

2 Exploratory Data Analysis

2.1 Dataset Description

We have fetched our dataset from https://finance.yahoo.com/quote/TCS.NS/history/. The head and tail part of our data set look like the following.

```
        Date    Open     High      Low    Close Adj.Close  Volume
1 2018-01-01 1341.15 1347.400 1317.500 1322.800  1211.359 1351760
2 2018-01-02 1330.00 1334.800 1310.100 1315.600  1204.766 1920290
3 2018-01-03 1316.00 1334.500 1315.600 1319.325  1208.177 1257120
4 2018-01-04 1325.00 1331.000 1320.000 1328.550  1216.625  913082
5 2018-01-05 1325.00 1349.750 1325.000 1344.600  1231.323 1153706
6 2018-01-08 1350.00 1363.425 1340.925 1357.200  1242.861 1242220
```

⋮

```
          Date    Open    High     Low   Close Adj.Close  Volume
982 2021-12-23 3648.00 3670.50 3630.00 3662.70  3622.250 1792861
983 2021-12-24 3685.00 3705.00 3644.80 3670.90  3630.359 2209923
984 2021-12-27 3671.00 3700.00 3653.10 3696.10  3655.281 1534135
985 2021-12-28 3710.00 3725.00 3693.85 3706.55  3665.616 1456218
986 2021-12-29 3692.25 3719.95 3685.00 3694.70  3653.896 1456923
987 2021-12-30 3681.35 3740.00 3680.00 3733.75  3692.515 1966475
```

So, we have 904 observations on 7 variables with 0 missing values. The description of the variables are placed below.

Date: Date of trading

Open: The opening price of a stock is the price at which the share opens at the beginning of trading hours of the stock market

High: Maximum price of stock during the day

Low: Minimum price of stock during the day

Close: The closing price of a stock is the price at which the share closes at the end of trading hours of the stock market

Adj.Close: Adjusted close price adjusted for both dividends and splits.

Volume: The number of shares that changed hands during a given day.

2.1.1 Datatypes of the variables

```
'data.frame': 987 obs. of  7 variables:
 $ Date     : Date, format: "2018-01-01" "2018-01-02" ...
 $ Open     : num  1341 1330 1316 1325 1325 ...
 $ High     : num  1347 1335 1334 1331 1350 ...
 $ Low      : num  1318 1310 1316 1320 1325 ...
 $ Close    : num  1323 1316 1319 1329 1345 ...
 $ Adj.Close: num  1211 1205 1208 1217 1231 ...
 $ Volume   : int  1351760 1920290 1257120 913082 1153706 1242220 2149396 3365850 252074
```

So, our dataset is a data-frame with “Open”, “High”, “Low”, “Close”, “Adj.Close” as numerical variables, “Date” as date-time and “Volume” as integer variable.

2.2 Dataset Visualization

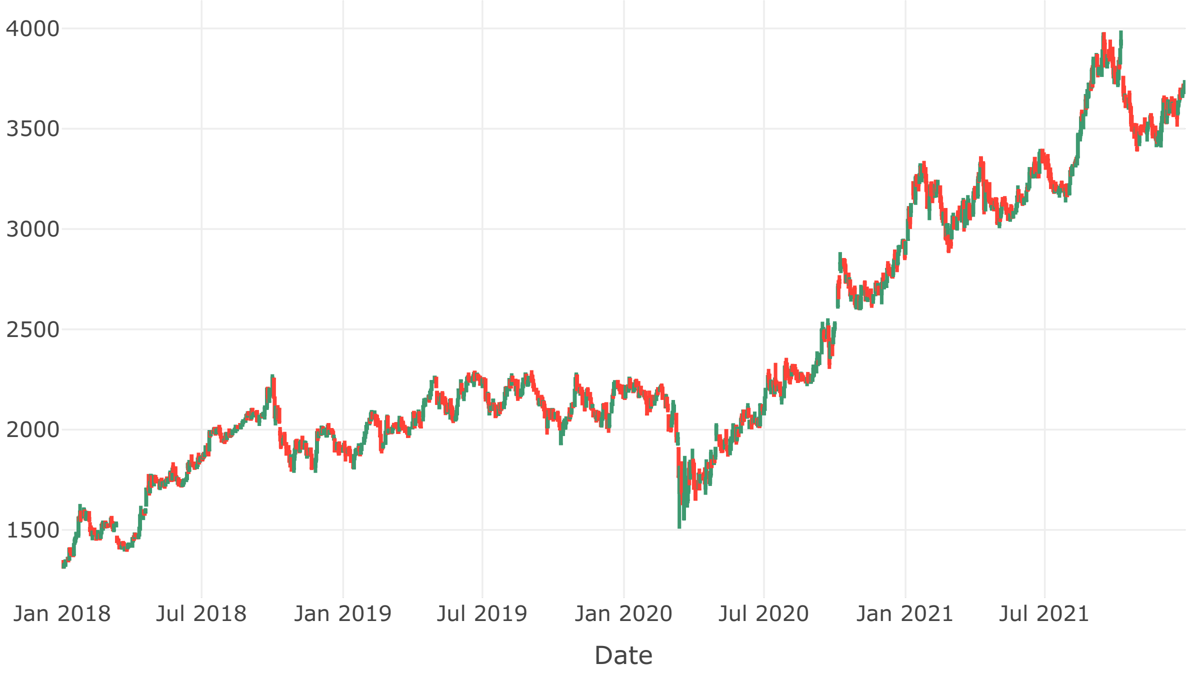

Figure 1: Candlestick Chart for dataset

A daily candlestick displays the market's open, high, low, and closing prices for the day, much like a bar chart does. The "true body" of the candlestick is the candlestick's widest portion.

The price range between the opening and closing prices of that trading day is represented by this actual body. It indicates that the close was lower than the open when the true body is filled in or became *red*. The close was higher than the open if the genuine body is filled in or become *green*.

Here we will perform our analysis only with the Adjusted Close price and its line diagram is represented below.

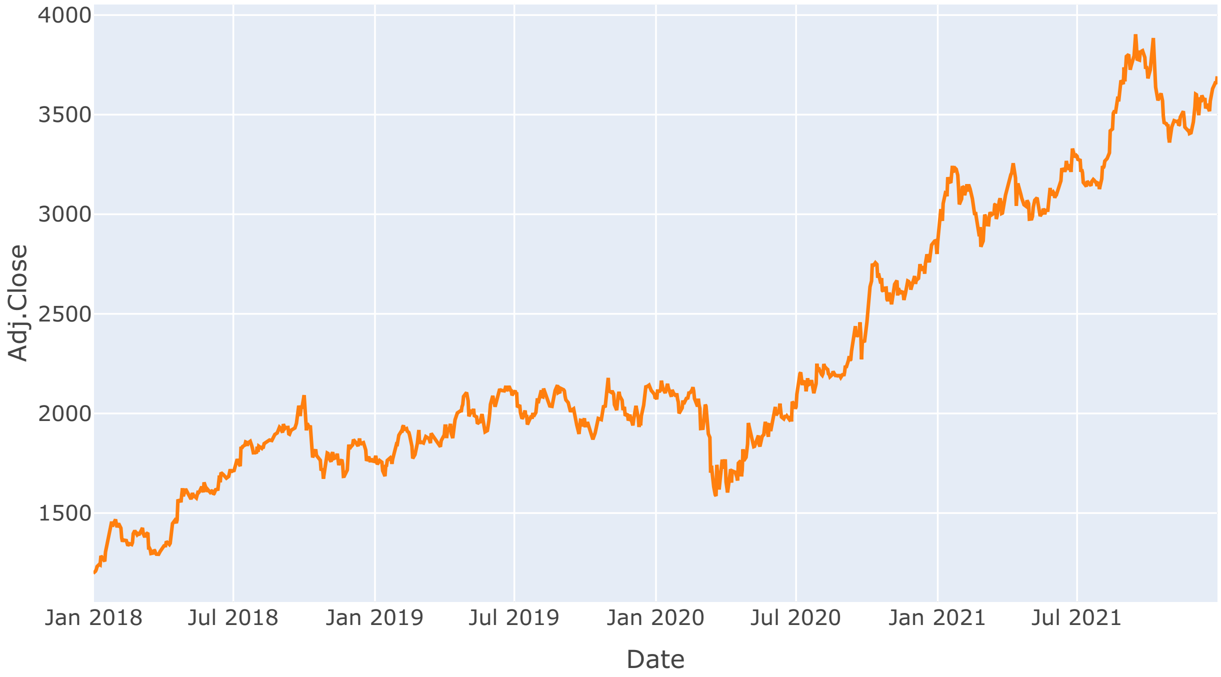

Figure 2: Adjusted close price over time

2.3 Structural Break Analysis

We can see from (2)at around 1st April, 2020 there is a structural change in our series. To test the existence of structural break mathematically we will use *Chow's test.*

2.3.1 Chow's Test

Suppose that we can model our whole data as

$$Y_t = \alpha + \beta t + \epsilon_t$$

If we split our dataset into two groups, then we have two different models

$$Y_t = \alpha_1 + \beta_1 t + \epsilon_t \text{ and } Y_t = \alpha_2 + \beta_2 t + \epsilon_t$$

where ϵ_t's are assumed to be independent and identically distributed from a normal distribution with unknown variance.

Here our test is

$$H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2 \text{ against } H_1 : \text{not } H_0$$

If RSS_C is the residual sum of square for the whole model, RSS_1 and RSS_2 are the residual sum of squares for *Model 1* and *Model 2* respectively and there are 553 and 434 observations which will be used for *Model 1* and *Model 2* respectively. The Chow's test statistic is given by

$$F = \frac{(RSS_C - (RSS_1 + RSS_2))\,/2}{(RSS_1 + RSS_2)\,/\,(987 - 4)} \overset{H_0}{\sim} F_{2,983}$$

We reject null hypothesis with α level of significance if observed value of F is greater than ,$F_{\frac{\alpha}{2};2,983}$ and less than $F_{1-\frac{\alpha}{2};2,983}$, where $F_{\alpha;\,2,\,983}$ is the upper αth point of F distribution.

```
The approx. breakpoint is 553

Chow test


data:  raw.data$Adj.Close ~ raw.data$Date
F = 1078.7, p-value < 2.2e-16
```

For our dataset p-value=$2.2 \times 10^{-16} < 0.05$. So, our null hypothesis got rejected and hence we can conclude the existence of structural break.

2.4 Train-Test Splitting

We have splitted each part of the dataset into two parts for assessing the quality of our model and prediction purpose.. For the training purpose we have considered data from "2020-04-02" to "2021-08-01" and rest of it will go for testing purpose.

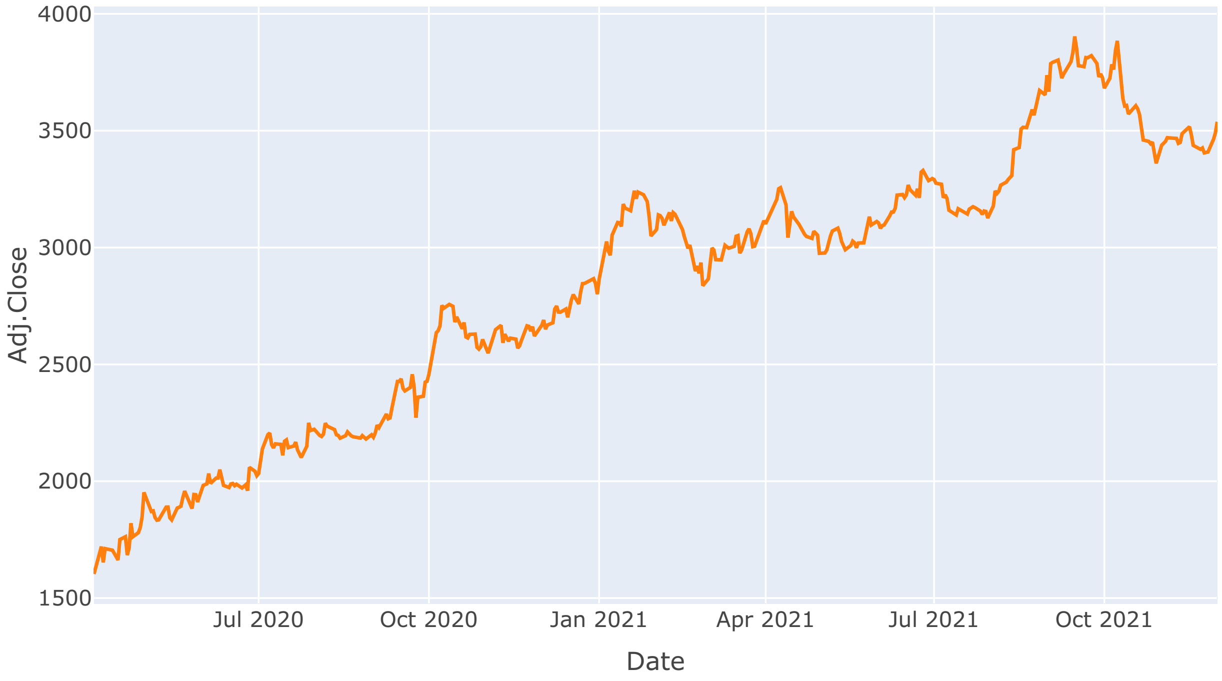

Figure 3: Plot of training sets

3 Presence of time series components

3.1 Test for Randomness

To check whether there is a deterministic components in our time series $\{X_t\}$, we have to perform test for randomness. We will use *"Turning Point Test"* which is a non-parametric test procedure to check randomness of the time series.

Here we are going to test

$$H_0 \text{ :The series is purely random } vs \ H_1 : \text{not } H_0.$$

Define,

$$U_i = \begin{cases} 1 & \text{If } X_i \text{ is a turning point} \\ 0 & \text{otherwise} \end{cases}$$

Suppose P is the total number of turning points, which does mean $P = \sum_{i=2}^{n-1} U_i$. An asymptotic test for H_0 is based on the test statistic ,

$$Z = \frac{P - \frac{2}{3}(n-2)}{\sqrt{\frac{16n-29}{90}}} \underset{H_0}{\overset{a}{\sim}} N(0,1)$$

We reject null hypothesis with α level of significance if observed value of $|Z|$ is greater than $\tau_{\frac{\alpha}{2}}$, where $\tau_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$th point of standard normal distribution.

```
Turning point test of independence


data:  x
T = -9.5244, p-value < 2.2e-16
```

For our dataset, p-value=$2.2 \times 10^{-16} < 0.05$. So, our null hypothesis got rejected and hence we can conclude that the series is not random.

3.2 Test for Trend

We can see from figure (2) there is an increasing trend in the series. To check the presence of trend in our time series $\{X_t\}$, we have to perform test for exixtence of trend. We will use *"Relative ordering test"* which is a non-parametric test procedure to check presence of trend of the time series.

Here we are going to test

$$H_0 : \text{There is no trend } vs \ H_1 : \text{not } H_0.$$

Define,

$$q_{ij} = \begin{cases} 1 & \text{If } X_i > X_j \text{ for } i < j \\ 0 & \text{otherwise} \end{cases}$$

Suppose Q is the total number of decreasing points, which does mean $Q = \sum_{\substack{i,j \\ i<j}} q_{ij}$. Q is related to Kendall's τ test statistic ,

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

Under the null hypothesis,$E(\tau) = 0$ and $Var(\tau) = \frac{2n(2n+5)}{9n(n-1)}$. An asymptotic test for H_0 is based on the test statistic ,

$$Z = \frac{\tau - E\left(\tau\right)}{\sqrt{Var\left(\tau\right)}} \overset{a}{\underset{H_0}{\sim}} N\left(0,1\right)$$

We reject null hypothesis with α level of significance if observed value of $|Z|$ is greater than $\tau_{\frac{\alpha}{2}}$, where $\tau_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$th point of standard normal distribution.

```
#############Relative Ordering Test for Presence of Trend#############

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 25.5183
p_value: 0
No. of Discordants: 6778
Expected No. of Discordants: 42539
```

For our dataset, p-value=0 $<$ 0.05. So, our null hypothesis got rejected and hence we can conclude that trend component is present in our time series.

3.2.1 De-trending the data

For de-trending the series we will use *"Differencing Method"*. We first applied differencing method of lag 1 in our series. In most of the cases, the stock prices are assumed to be dependent on its immediate past value.

If we apply order 1 differencing our time series reduces to

$$Z_t = \nabla X_t = X_t - X_{t-1}$$

```
#############Relative Ordering Test for Presence of Trend#############

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: -0.6009
p_value: 0.274
No. of Discordants: 43172
Expected No. of Discordants: 42333
```

For our dataset, p-value=0.274 > 0.05. So, we cannot reject null hypothesis and hence we can conclude that trend component is no longer present in our time series. To get a visualization, we have the figure below.

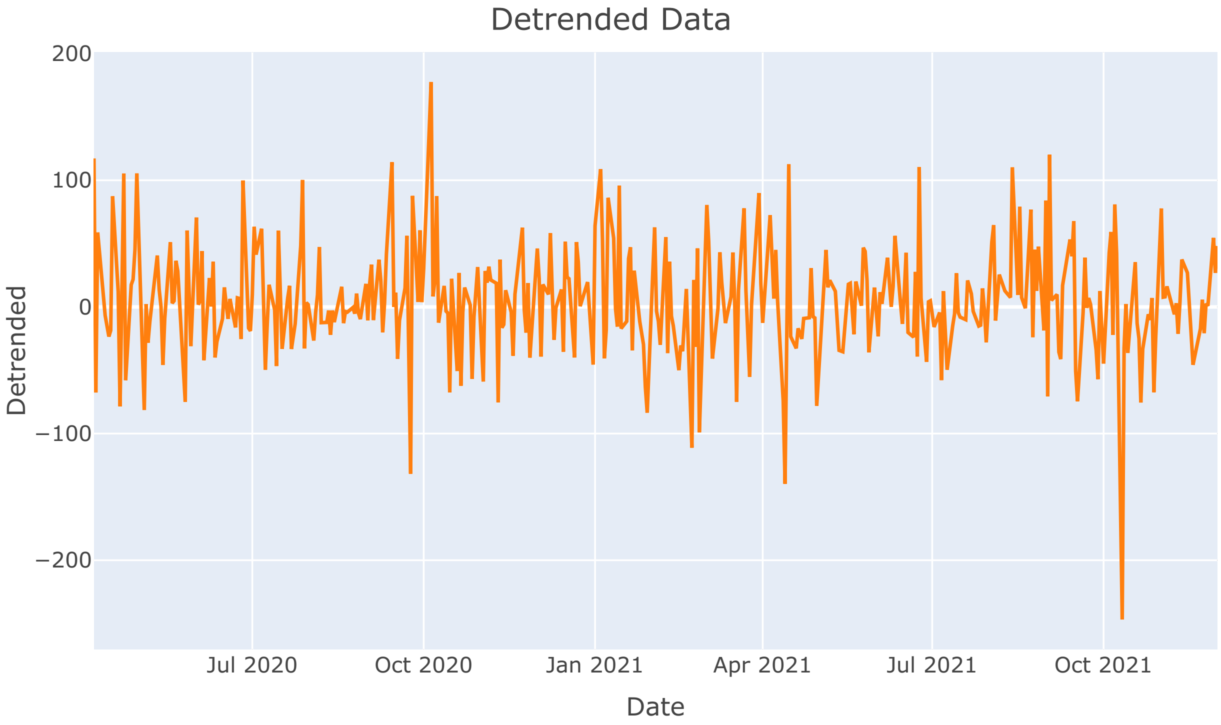

Figure 4: Plot of De-trended data

4 Model Selection

4.1 Test for Stationarity

Here we will use **Augmented Dickey-Fuller Test (ADF Test)**. ADF test is a unit root test that tests the null hypothesis $H_0 : \alpha = 0$ against $H_1 : \alpha < 0$ in the following model equation.

$$\nabla Z_t = c + \beta t + \alpha Z_{t-1} + \phi_1 \nabla Z_{t-1} + ... + \phi_p \nabla Z_{t-p} + \epsilon_t$$

which is same as testing the null hypothesis $H_0 : \alpha^* = 1$ against $H_1 : \alpha^* < 1$ in the following model equation

$$\begin{aligned} Z_t &= c+\beta t+(\alpha+1)\,Z_{t-1}+\phi_1\nabla Z_{t-1}+....+\phi_p\nabla Z_{t-p}+\epsilon_t \\ &= c+\beta t+\alpha^* Z_{t-1}+\phi_1\nabla Z_{t-1}+....+\phi_p\nabla Z_{t-p}+\epsilon_t \end{aligned}$$

The test's underlying premise is that the lagged level of the series Z_{t-1} will not be useful in forecasting the change in Z_t if the series is characterized by a unit root process.

```
Augmented Dickey-Fuller Test

data:  detrended
Dickey-Fuller = -6.3487, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Here our p-value comes out to be $0.01 < 0.05$. So, we reject H_0 and conclude that the series is stationary.

4.2 ACF and PACF plots

To visually represent the degree of correlation between an observation of a time series and observations made at earlier time steps, autocorrelation and partial autocorrelation plots are utilized. We may observe time series from various angles using plots of the **autocorrelation function (ACF) and partial autocorrelation function (PACF)**.

Only the direct correlation between an observation and its lag adjusted by its intervening observations is described by PACF. While ACF defines the autocorrelation between one observation and another observation at a preceding time step that contains direct and indirect dependency information, this would imply that there would be no correlation for lag values exceeding k.

To establish the arrangement of AR and MA, we utilize the ACF and PACF plots. We calculate the lag for both graphs, which is the time period after which the ACF and PACF values are not substantially different from 0. From the ACF plot, we get the order of $MA\,(q)$ i.e. the value of q and from PACF plot we get the order of $AR\,(p)$ i.e. the value of p.

From the graph below, we can see that in both the cases after lag 2 the values of ACF and PACF can be considered indifferent from 0. So, we consider all the 36 models with p and q both less than or equal to 5.

From these 36 models we will choose that model for which we get the minimum value of *AIC (Akaike Information Criterion).*

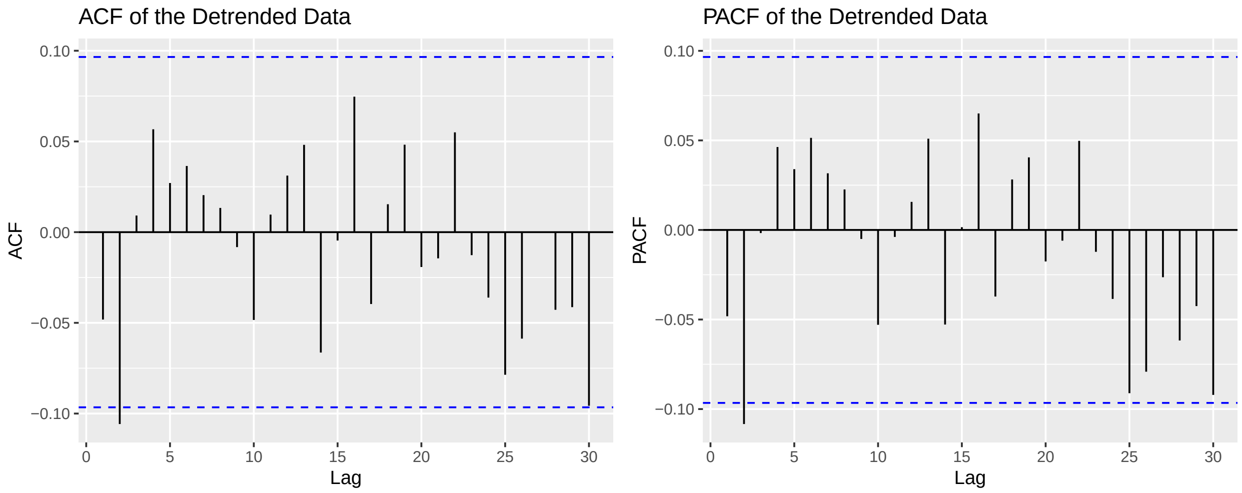

Figure 5: ACF and PACF plot of de-trended data

```
The best order of data is ( 2 1 0 )
The minimum AIC is 4267.75613505433
```

The result we have obtained shows that the minimum value of AIC attains for $ARIMA\,(2, 1, 0)$ model. So, we choose the value of p and q as 2 and 0 respectively.

4.3 Model Selection for Original Data

$\{X_t\}$ is said to follow an Auto Regressive Integrated Moving Average (ARIMA) of order (p, d, q), if,

$$Y_t = \nabla^d X_t = (1 - B)^d \, X_t \sim ARMA\,(p, q)$$

where d is the smallest integer for which $\nabla^d X_t$ is stationary. For our data, $d = 1$; so we can use $ARIMA(2, 1, 0)$ for fitting and forecasting procedure.

5 Residual Analysis

For the first part of our data, we have fitted, the $ARIMA(2, 1, 0)$ model. The model coefficients are

```
Coefficients for the model are
        ar1         ar2
-0.04018710 -0.09672109
```

The model is given by,

$$\left(1 + 0.04018710B + 0.09672109B^2\right) \nabla X_t = \epsilon_t$$

Information that is not explained by the fitted model is referred to as residuals. To determine if errors follow a *White Noise Process,* residual analysis is carried out. The autocorrelation function having values within the confidence interval of the corresponding estimates will indicate whether the residuals are uncorrelated.

5.1 Ljung-Box test

We perform ***Ljung-Box* Test** to check whether the errors are independently distributed.

To test

$$H_0 \text{ :The data is independently distributed against } H_1 : \text{not } H_0$$

The test statistic is

$$Q = n\left(n+2\right)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

where n is the sample size, $\hat{\rho}_k$ is the sample autocorrelation at lag k, and h is the number of lags being tested. Under H_0 the statistic Q asymptotically follows a $\chi^2_{(h)}$. For significance level α, the critical region for rejection of the hypothesis of randomness is:

$$Q > \chi^2_{1-\alpha,h}$$

```
Box-Ljung test

data:  res
X-squared = 10.18, df = 20, p-value = 0.9648
```

Here we have tested for 20 lags and p-value of the test came out to be 0.9648> 0.05. Thus we accept H_0 at 5% level of significance and infer that the residuals are independent.

Let's plot ACF and PACF of ϵ_t

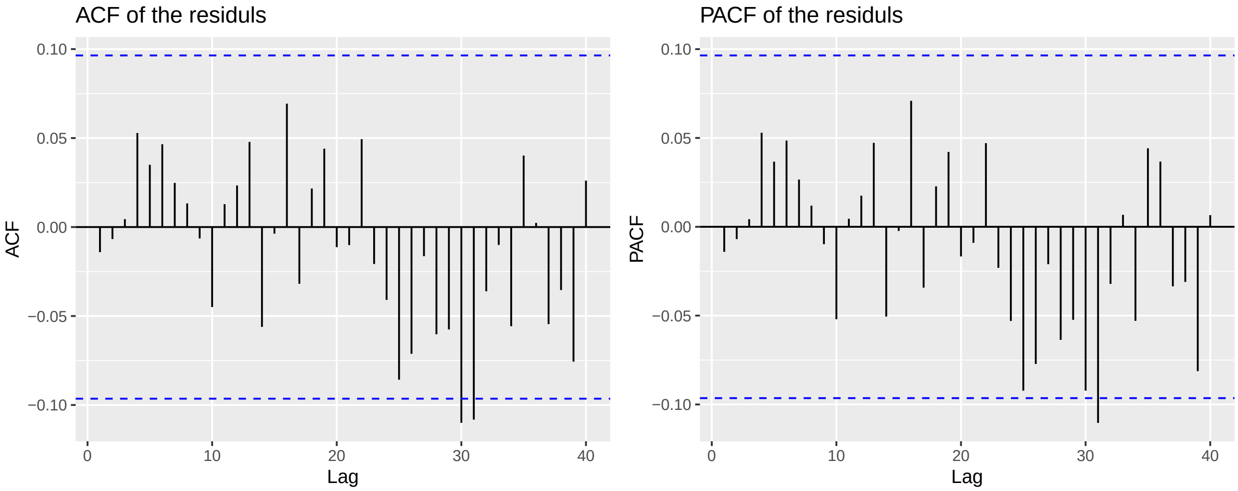

Figure 6: ACF and PACF of Residuals

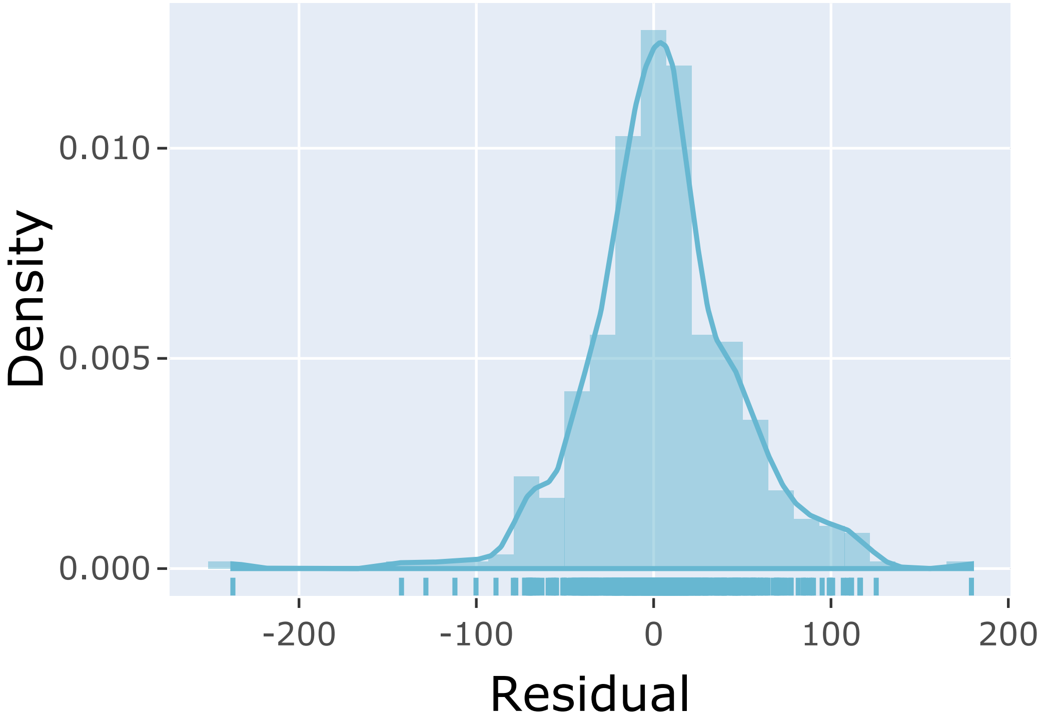

Figure 7: Density estimation of Residuals

From the density plot of the residuals, we observe that the residuals are concentrated at 0, which suggests that the mean of the residuals is close to zero, i.e. we have,

$$E\left(X_t\right) = 0,\ \forall t$$

Now we wish to check whether the residuals are homoscedastic or not, i.e.,

$$V\left(X_t\right) = constant, \forall t$$

Checking for the ARCH effect in the model is necessary for this reason. A statistical model for time series data called the Autoregressive Conditional Heteroscedasticity (ARCH) model depicts the conditional variance of the current error component as a function of the lagged values of the residuals in the past.

A test of serial independence is finding the ARCH effect in a time series. We have assumed that an effective model eliminates the linear serial dependence found in the original series. Any additional serial dependence must therefore result from a nonlinear mechanism that the model has not yet identified. Conditional heteroscedasticity is the nonlinear mechanism in this case that we are interested in.

5.2 Lagrange Multiplier Test:

In Lagrange Multiplier Testing Process,

$$H_0 : ARCH \text{ effect is not present ag. } H_1 : ARCH \text{ effect is present}$$

This procedure simply involves obtaining the squares of the residual from fitted model and regress them on a constant and p lagged values, where is the $ARCH$ lags.Let us consider the equation:

$$\epsilon_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + V_t$$

,where V_t is random error. The hypothesis is that, in the absence of ARCH components, we have $\alpha_i = 0, \forall i = 1, 2, ..., p$,against the alternative that, in the presence of ARCH components, at least one of the estimated α_i,must be significant.

```
ARCH LM-test; Null hypothesis: no ARCH effects

data:  res
Chi-squared = 19.9, df = 26, p-value = 0.7963
```

Here p-value is $0.7963 > 0.05$, so we fail to reject the null hypothesis and conclude that errors are homoscedastic and hence we will go with our old ARIMA model to forecast future values.

6 Forecasting

We are going to forecast with the ARIMA model

$$\left(1 + 0.04018710B + 0.09672109B^2\right) \nabla X_t = \epsilon_t$$

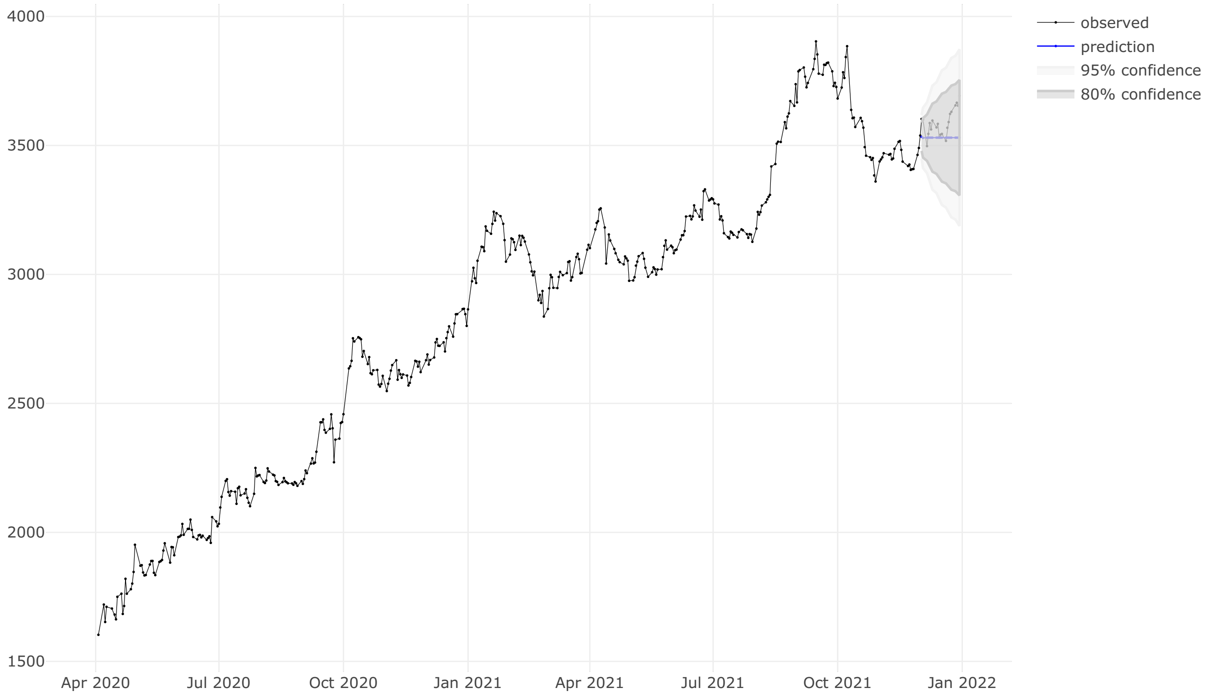

Figure 8: Forecasting of test data

We have also obtained

```
Mean Absolute Percentage Error is 1.140068
```

which is quite low.

7 Conclusion

After all the statistical analysis we found that $ARIMA\,(2,1,0)$ i.e. $AR\,(3)$ fits the best on the raw data. Also, the residuals follow a *White Noise* process. Forecasting for the future 16 days with the final $AR\,(3)$ model resulted predicted values close to the observed ones with *Mean Absolute Percentage Error* of 1.140068.

References

[1] Englex2019;s ARCH Test - MATLAB amp; Simulink — mathworks.com. https://www.mathworks.com/help/econ/engles-arch-test.html. [Accessed 10-Nov-2022].

[2] Peter J. Brockwell and Richard A. Davis. *Introduction*, pages 1–37. Springer International Publishing, Cham, 2016.

[3] Hardikkumar Dhaduk. Stock market forecasting using Time Series analysis With ARIMA model — analyticsvidhya.com. https://www.analyticsvidhya.com/blog/2021/07/stock-market-forecasting-using-time-series-analysis-with-arima-model/. [Accessed 10-Nov-2022].

[4] Wayne A Fuller. *Introduction to statistical time series*. Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, TN, 2 edition, December 1995.

[5] James Douglas Hamilton. *Time series analysis*. Princeton university press, 2020.