Post Graduate Teaching Department of Statistics
RTMNU
(2022 – 2023)

A project on

Infosys Stock Price Prediction using ARIMA Model

Guide
Ms. Manjiri Gajarlawar
Mrs. Vishakha Jaiswal

Submitted by
Ms. Bhagyashri Lakhadive
Ms. Riya Dehariya
Mr. Swapnil Wankhede

Outline

1. **Problem Statement**
2. **Stock Market, Data Info.**
3. **Exploratory Data Analysis (Interpretation of closing price)**
4. **Time series analysis**
   - ✓ Components of time series analysis (Interpretation)
   - ✓ Mathematical model of time series
   - ✓ Stationarity (ADF test to check stationarity)
   - ✓ Autocorrelation Analysis (Interpretation)
   - ✓ ARIMA Model (Diagnosis) (Interpretation)
5. **Limitation of Study**
6. **Challenges**
7. **Conclusion**
8. **References**

Problem Statement

To predict the **"Stock Closing price of January-2023 of Infosys Company"** on the basis of past trends and patterns in the data with the help of past observations of one year using the **ARIMA Model**.

Stock Market

- ✓ The stock market is considered to be one of the most highly complex financial systems which consist of various components or stocks, the price of which fluctuates greatly with respect to time.
- ✓ All the stock market investors aim to maximize the returns over their investments and minimize the risk associated.
- ✓ Stock markets being highly sensitive and susceptible to quick changes, the main aim of stock-trend prediction is to develop new innovative approaches to forsee the stocks that result in high profits.
- ✓ Here we are trying to analyze the time series data of the Indian Stock market and build a statistical model that could efficiently predict the future stocks.

Stock Exchange

- ✓ A stock exchange is a regulated market for trading.
- ✓ If a company wishes to sell its shares, it should be registered in the stock exchange.
- ✓ There are two major stock exchanges in India:
    1. Bombay Stock Exchange (BSE).
    2. National Stock Exchange (NSE).
- ✓ Once registered, the company can list its shares and sell them at a price to the investor.
- ✓ The stock market in India operates during a specific time window. The regular market trading hours are from 09:15 AM and close at 03:30 PM.

Information about Company and Dataset

- ✓ Infosys Limited is an Indian Multinational Company. It deals with Information technology, Consulting and Outsourcing.
- ✓ The company is Headquartered in Bangalore, Karnataka in India.
- ✓ The main services include Application Development & Maintenance, Business Process Management, Consulting Services, Incubating Emerging offerings.

The data from **1st January 2022 to 31st January 2023 of Bombay stock exchange (BSE**) has been collected. The dataset consist of 269 observations.

Date	The date of particular day
Open	The opening price of stock of the particular day
Close	The closing price of stock of the particular day
High	The Highest price at which a stock traded during the period.
Low	The Lowest price at which a stock traded during the period

Exploratory Data Analysis (EDA)

EDA is an important first step in any data analysis. EDA is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected. Excel and Python are used for the exploratory data analysis and time series analysis.

Histogram, Scatter plot, Box whisker plot are used.

Moving Average, Weighted Moving Average, Exponentially Moving Average are used **to check the smoothness of time series data.**

- ✓ **No missing observation**
- ✓ **Ordered data**
- ✓ **Can not be independent.**
- ✓ **In regular interval**

Graphical representation of Decreasing Trend in Closing Price

- The stock prices are declining due to Global Recession.
- A major downward trend from March 2022, due to consecutive top-level resignations.
- Shares fell in the last week of August, after it was reported that Infosys has reduced the variable pay out for all its employees.
- The price gained after the Company announced to consider a proposal for a share buyback on Thursday, 13 October.
- Factors including Covid-19 pandemic, Russia's invasion of Ukraine, and rising inflation.

Time series

- ✓ A time series can be defined as " **A set of observation of a variable recorded at successive intervals or point of time** ".
- ✓ Mathematically, time series is defined by functional relationship

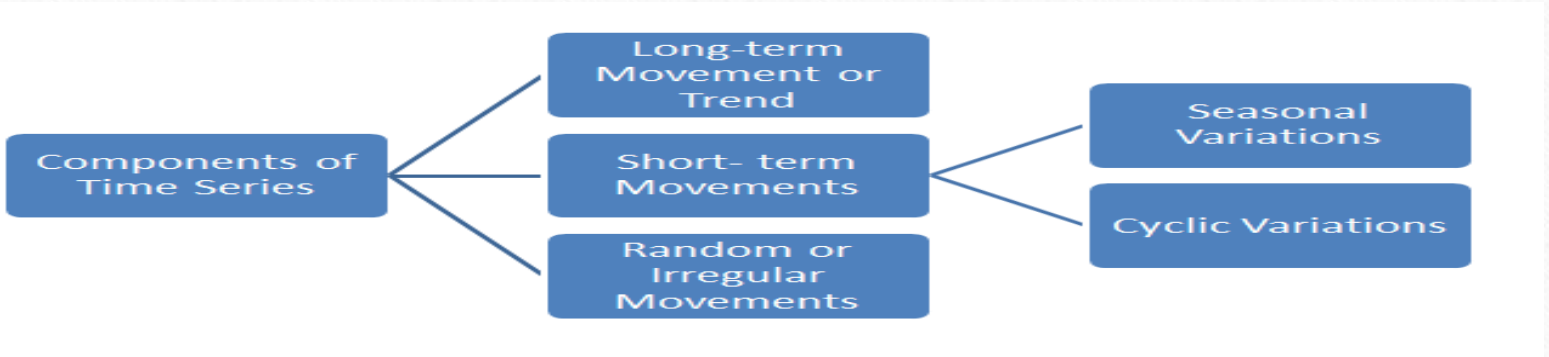
$$y_t = f(t) \quad ; \quad t = t_1, t_2, t_3 \ \ldots\ldots\ldots\ldots$$

Where, y_t - value of variable under consideration at time t

- ✓ The purpose of Time series analysis in stock analysis is to forecast the future stock prices on the basis of past trends and patterns in the data.
- ✓ Time series analysis is well suited for stock analysis because the stock prices are often influenced by the no of factors that change over time, factors including economic indicators, market trends, company specific events, etc.

Components of Time series

Time series data can exhibit a variety of patterns, and it is often helpful to split a time series into several components.



- ✓ **Trend** is the general tendency of the data to increase or decrease over period of time.
- ✓ **Seasonal** variations are the regular pattern of up and down fluctuations which operates in a regular and periodic manner.
- ✓ **Cyclic** pattern exists when data exhibit rises and falls that are not of fixed period. The duration of these fluctuations is usually of at least 2 years.
- ✓ **Random** movements are purely random, unpredictable and are due to factors which are beyond the control of human hands.

Mathematical Models for Time Series

Two models are commonly used for decomposition of a time series into its components.

1) Additive model :

Additive models assume that the components of the time series can be added together to produce the final result.

An additive model for a time series might be represented as:

$$y_t = \text{Trend(t)} + \text{Seasonality(t)} + \text{Cyclicity(t)} + \text{Residual(t)}$$

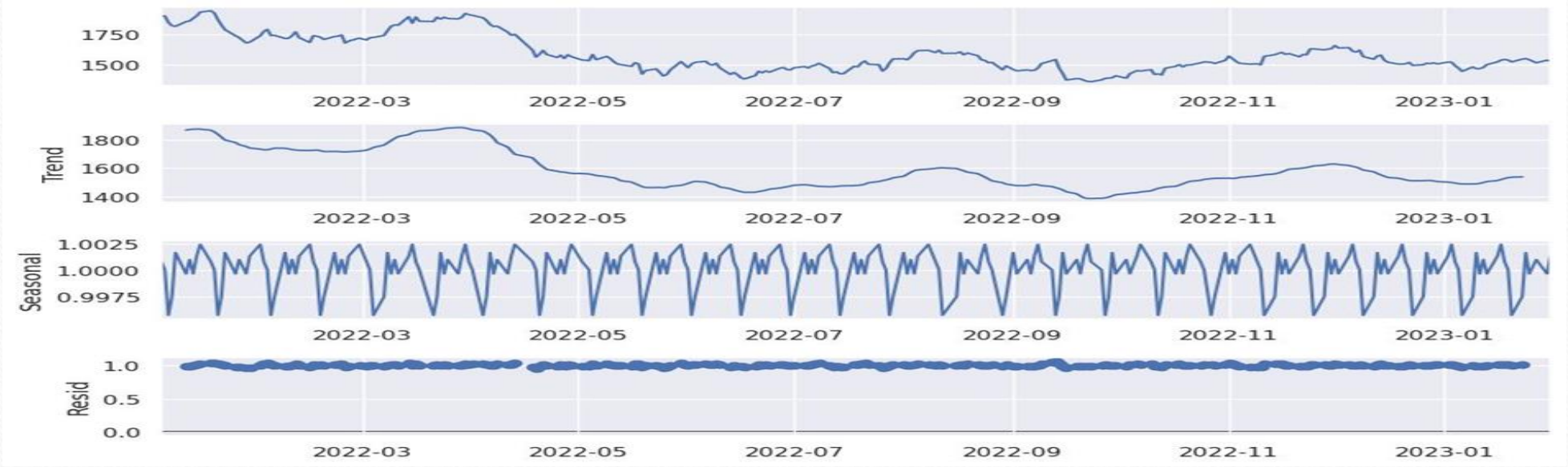
2) Multiplicative Model :

Multiplicative models assume that the components of the time series multiply to produce the final result.

A multiplicative model for a time series might be represented as:

$$y_t = \text{Trend(t)} \times \text{Seasonality(t)} \times \text{Cyclicity(t)} \times \text{Residual(t)}$$

Interpretation of Components of Time Series analysis



- ✓ The above graph shows a multiplicative decomposition of the data.
- ✓ The trend shows the overall movement in the series. The data have an overall decreasing trend.
- ✓ The stock market tends to cycle between periods of high and low values, but there is no set amount of time between those fluctuations.
- ✓ The residual component shown in the bottom panel is what is left over when the cyclic and trend components are removed.

Stationarity in time series

- ✓ A stationary time series is one whose statistical properties such as mean, variance and covariance are all constant over time.
- ✓ It is easy to make predictions on a stationary series since we can assume that the future statistical properties will not be different from those currently observed.
- ✓ A time series whose statistical properties change over time is called a non-stationary time series.

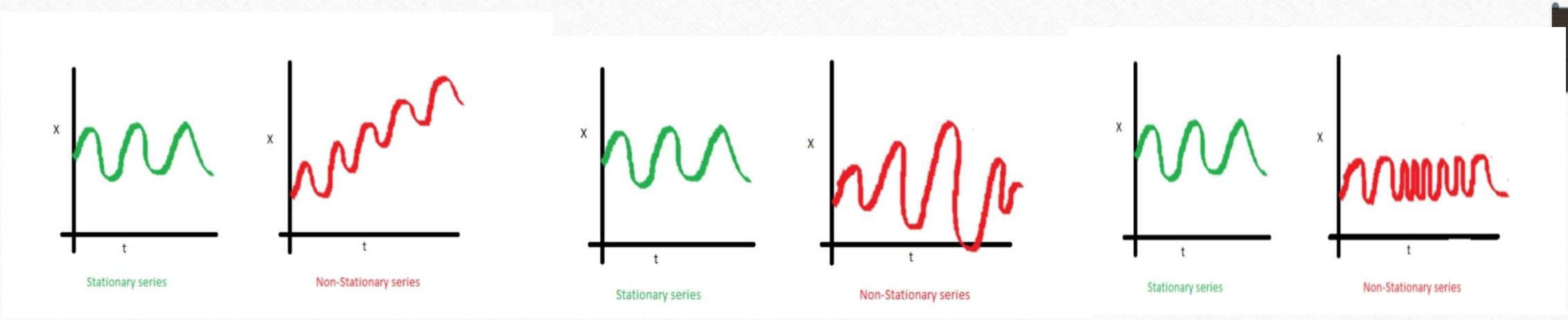
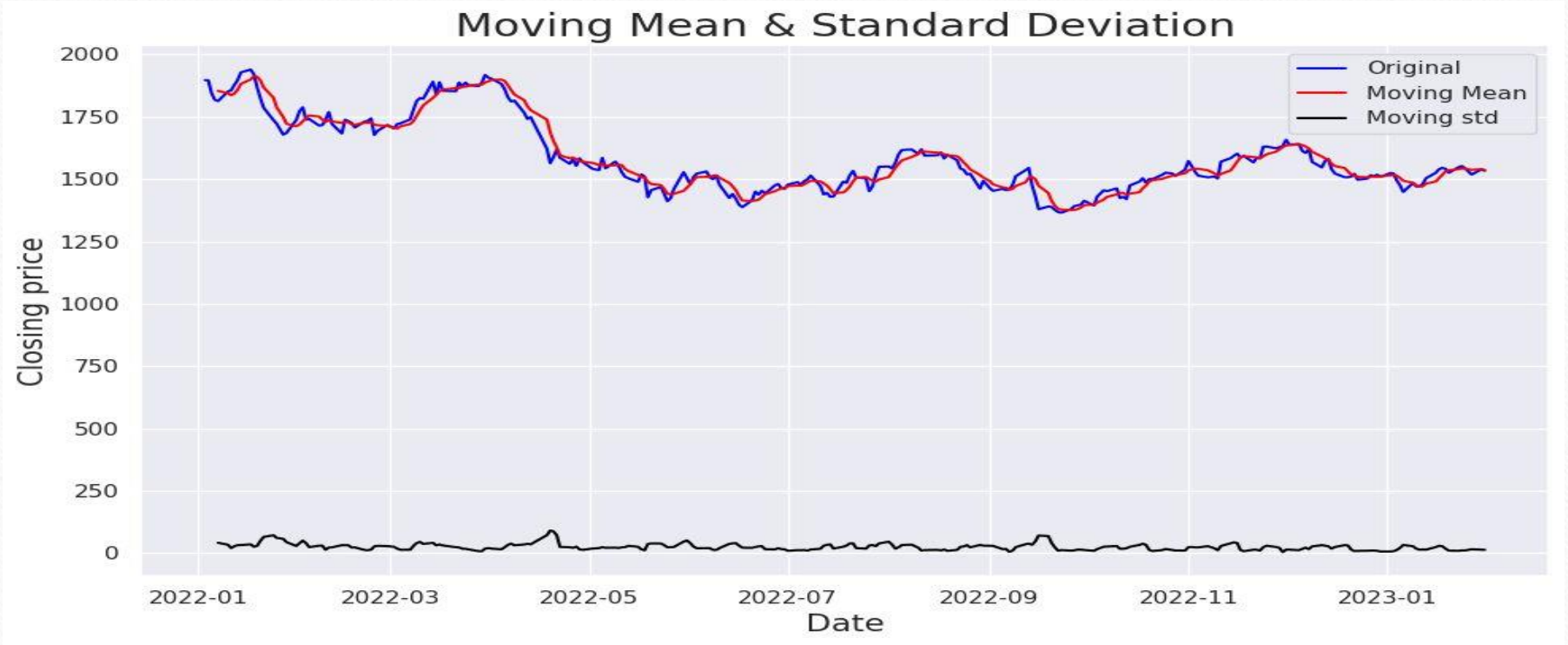


fig (A) fig(B) fig(C) fig(D) fig(E) fig(F)

Analyzing the Stationarity of a Time Series:

- **Moving Statistics**

To analyse stationarity we have taken moving mean and moving standard deviation. we used window of 10 for moving mean and standard deviation



Augmented Dickey-Fuller test(ADF) or unit root test

- ✓ ADF test is based on Dickey-fuller test, which is used to test for the presence of a unit root in time series.
- ✓ A unit root is a feature of non-stationary time series data that causes the data to have a trend or drift over time.
- ✓ To check the series is stationary or not by ADF test or Unit root test we set up the hypothesis,

  H_0 : The data is non-stationary or There is a presence of unit root.

  v/s

  H_1 : The data is stationary or There is absence of unit root.
- ✓ If the null hypothesis can be rejected, we can conclude that the time series is stationary.

There are two ways to rejects the null hypothesis:

On the one hand, the null hypothesis can be rejected if the p-value is below a set significance level. The defaults significance level is 5%

p-value > (default: 0.05)

(Fail to reject (H0), the data is non-stationary.)

p-value <= (default: 0.05)

(Reject the (H0), the data is stationary.)

p-value: The probability of obtaining the ADF statistic if the null hypothesis is true. The lower the p-value, the stronger the evidence against the null hypothesis.

On the other hand, the null hypothesis can be rejects if the test statistic is less than the critical value.

ADF statistic > critical value

(Fail to reject (H0), the data is non-stationary.)

ADF statistic < critical value:

(Reject the (H0), the data is stationary.)

Critical Values(at 5%): The critical value for the ADF test at, 5% significance level are used to determine whether the null hypothesis should be rejected or not.

The ADF test statistic is calculated as:

$$\mathrm{ADF}_t = \frac{(\Delta y_t - \alpha - \beta_t)}{\sigma}$$

. It measures how many standard deviations the time series is away from being stationary. The more negative this value is, the more likely it is that the time series is stationary.

Results of Dickey-Fuller Test:

Test Statistic	-2.289897
P-value	0.175247
Critical value (5%)	-2.87236

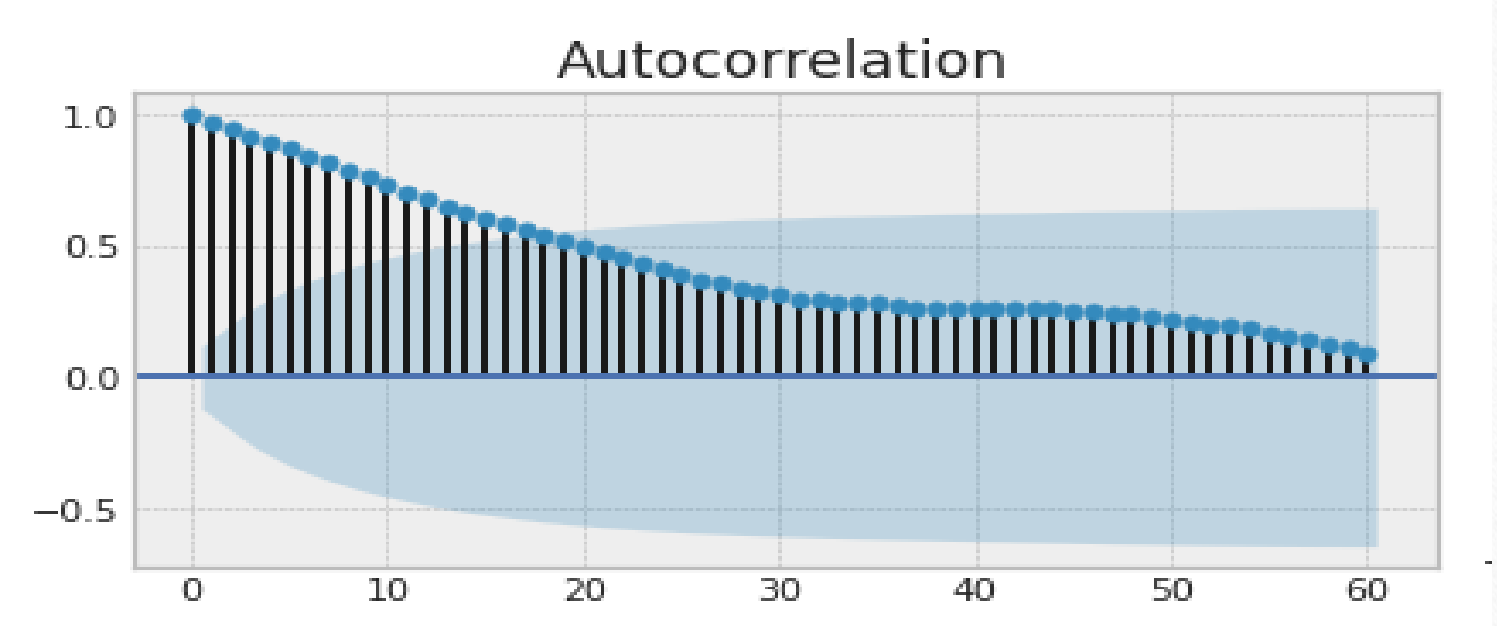
Looking at the test statistic value, we can see that it is greater than a 5% critical value score of -2.87386 and the p values is also greater than 0.05 (significance level) Therefore, we cannot reject the null-hypothesis and can conclude that this data is non-stationary.

Autocorrelation analysis

Autocorrelation analysis is an important step in the Exploratory Data Analysis of time series forecasting. The autocorrelation analysis helps detect patterns and check for randomness. It's especially important when you intend to use an autoregressive–moving-average (ARMA) model for forecasting because it helps to determine its parameters. The analysis involves looking at the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

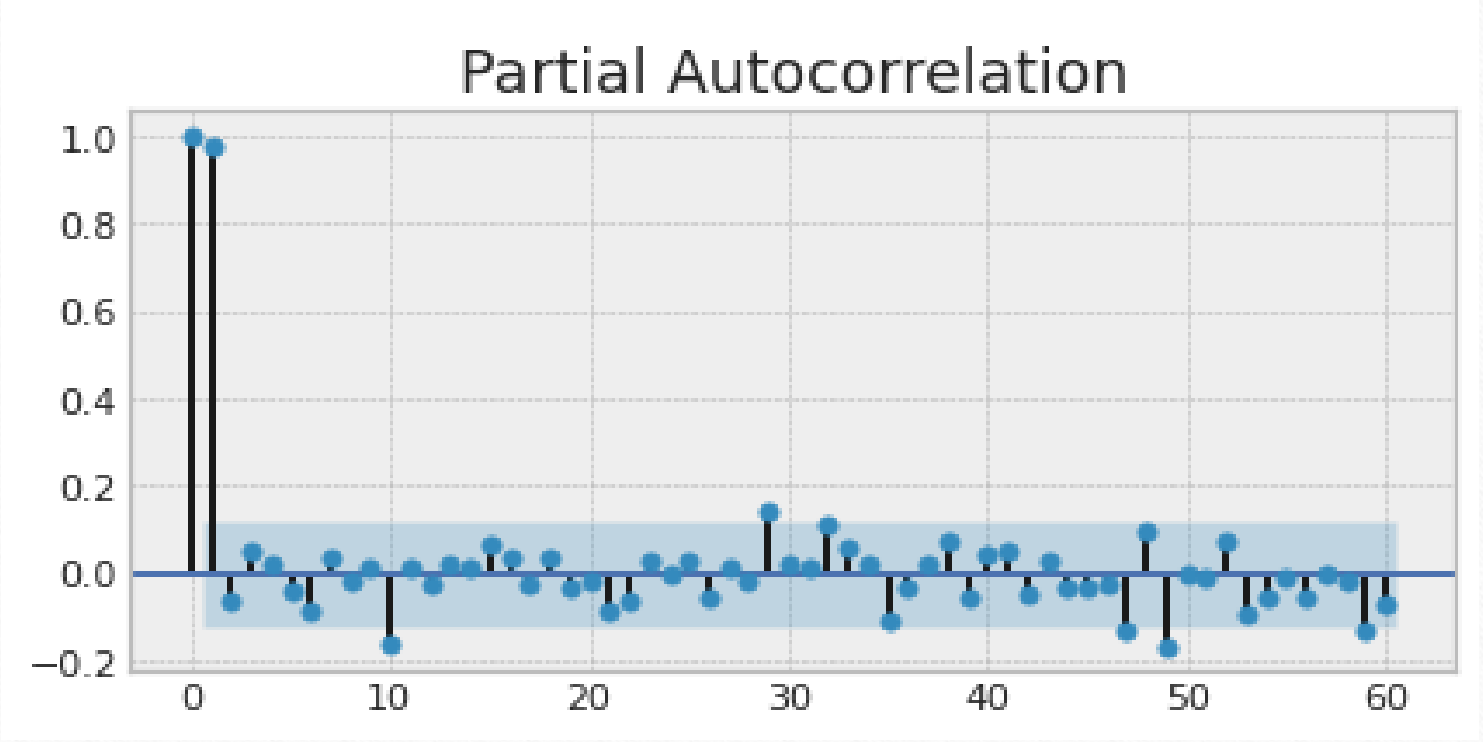
Autocorrelation Function (ACF)

- ✓ Autocorrelation is the correlation between a time series with a lagged version of itself. The ACF starts at a lag of 0, which is the correlation of the time series with itself and therefore results in a correlation of 1. From the ACF, you can assess the randomness and stationarity of a time series.
- ✓ The autocorrelation function declines to near zero rapidly for a stationary time series. the ACF drops slowly for a non-stationary time series.

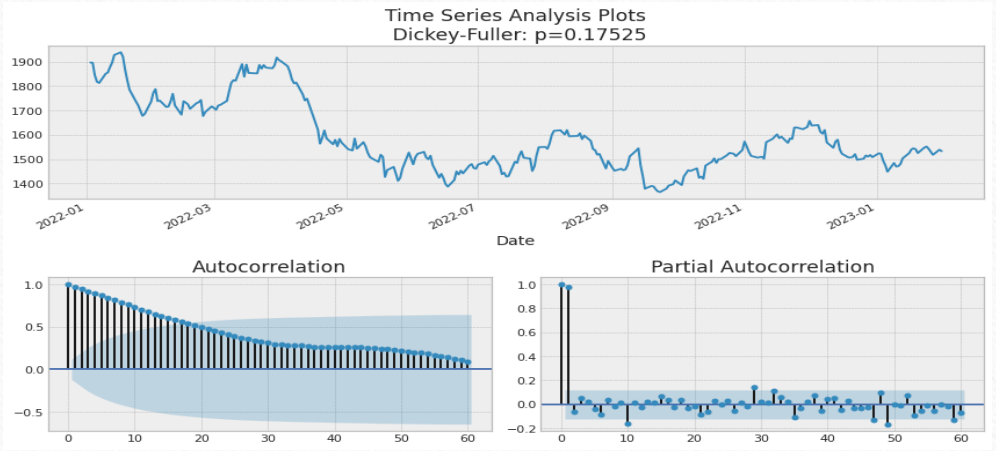


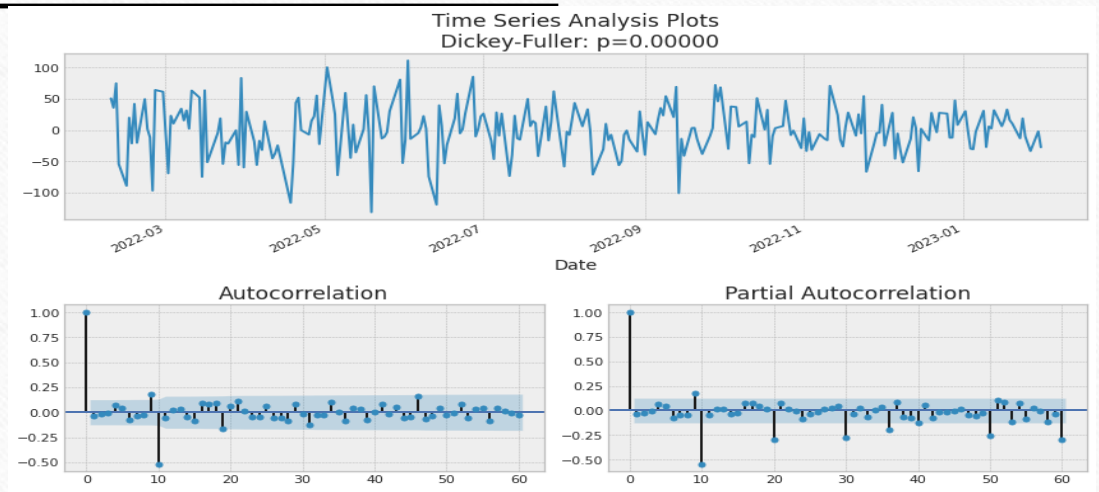
Partial autocorrelation function(PACF)

- ✓ The partial autocorrelation function is similar to the ACF except that it displays only the correlation between two observations that the shorter lags between those observations do not explain. The autocorrelation function helps assess the properties of a time series.



Interpretation of ACF and PACF





Before differencing

After differencing

The Partial Autocorrelation Function (PACF) plot can be used to draw a correlation between the time series and its lag. From the above Partial Autocorrelation plot, we can see that the 9^{th} lag is significantly out of the limit so we can select the order of the p (AR) as 9.To find out the value of q we can use the Autocorrelation Function (ACF) plot, Here, in the above auto correlation plot, we can see that 9^{th} lag is out of the significance limit so we can say that the optimal value of q (MA) is 9.

ARIMA: Auto Regression Integrated Moving Average

Auto Regression (AR) Model

Differencing operation to convert non stationary series to stationary. $y'_t = y_t - y_{t-1}$

Moving Average (MA) Model

1. Auto Regression model

- ✓ The current series values depend on its previous values with some lag.
- ✓ Regression of the time series onto itself
- ✓ **AR model of order p:** $y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$
- ✓ The maximum lag in the model is referred to as p (PACF plot)

2. Moving Average (MA)

- ✓ Time series is regressed with residuals of the past observations.
- ✓ The order of the error lag is denoted as **q, (ACF) plot.**
- ✓ **MA model of order q:** $y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$

ARIMA

- ✓ Statistical model used for forecasting time series data.
- ✓ Regression type equation in which the independent variables are lags of the dependent variable and lags of the forecast errors.

$$y'_t = c + \varphi_1 y'_{t-1} + \cdots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Interpretation of ARIMA

1. Mean Absolute Percentage Error

Metrics use to calculate the accuracy of a method for constructing fitted time series values.

$$MAPE = \frac{100}{N} \times \sum_{i=1}^{N} \left| \frac{Y_i - \widehat{Y_i}}{Y_i} \right|$$

N – number of non-missing data points.

2. Akaike's Information Criteria

To determine useful order of an ARIMA model.

$$AIC = -2log(L) + 2(p + q + k + 1)$$

where L is the likelihood of the data,

$$k = 1 \ if \ c \neq 0 \ and \ k = 0 \ if \ c = 0.$$

3. Corrected AICc

$$AIC_c = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{\mathrm{N}-p-q-k-2}$$

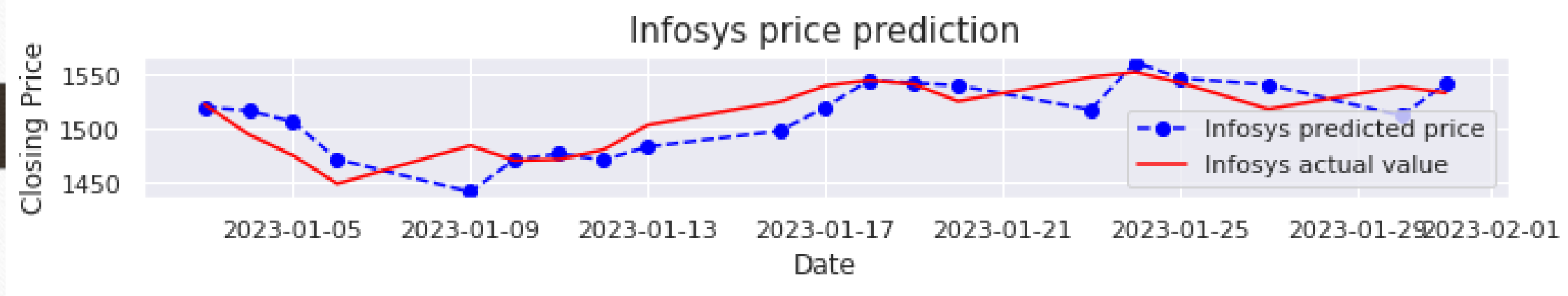
4. Bayesian Information Criterion

✓ Criterion for model selection among a finite set of models

$$BIC = AIC + [\log(\mathrm{N})-2]\,(p+q+k+1)$$

Good models are obtained by minimising the AIC, AICc or BIC

- ✓ No. of Observation = 249
- ✓ ARIMA(p,d,q)=(9,1,9)
- ✓ AIC = 2357.022; BIC = 2360.022 less value among all possible combinations



- ✓ **Graph**: Around 3.94 % MAPE (mean absolute percentage Error) implies the model is about 96.06% accurate in predicting the test set observations.

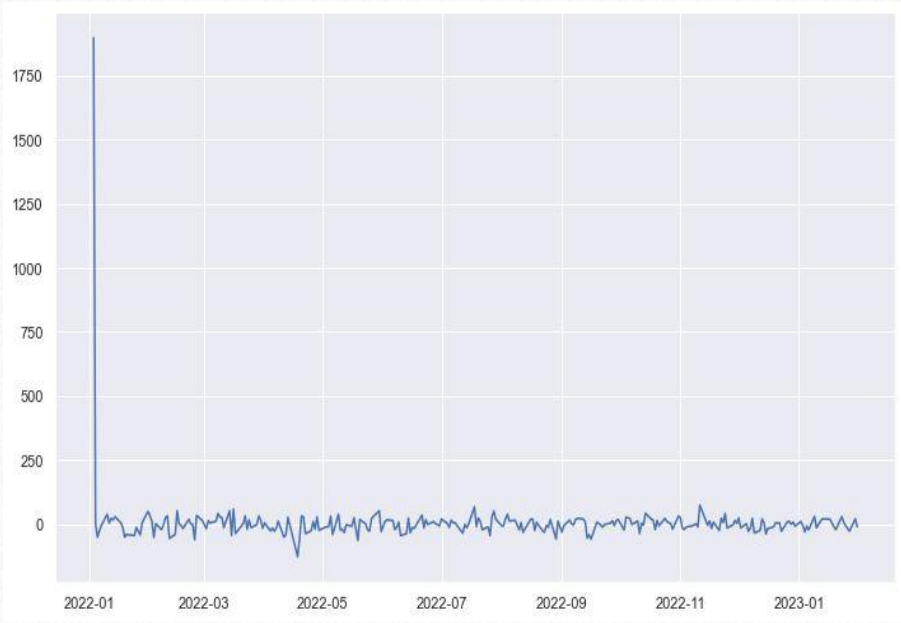
Diagnosis of ARIMA Model

The aim of diagnostics is to ensure that the ARIMA model is a good fit for the data and that the model's assumptions should be satisfied.

1. Residual Analysis:

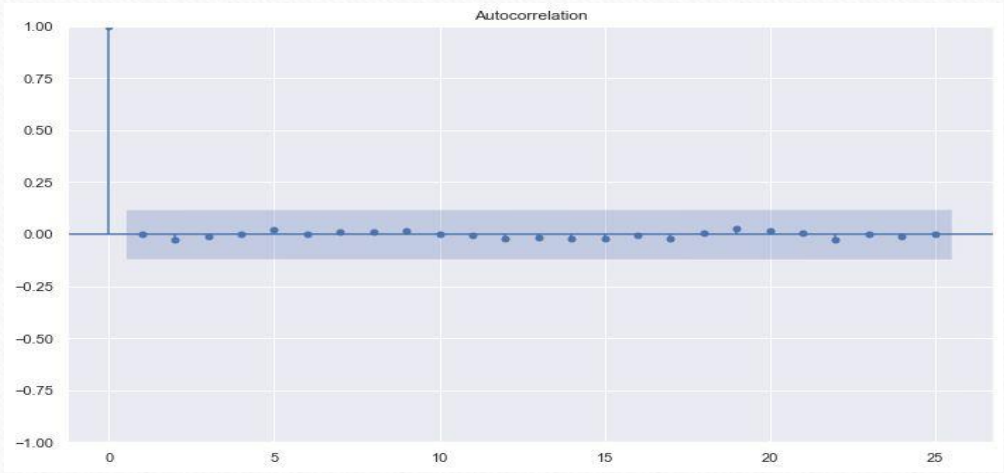
Residual are equal to the difference between the observations and the corresponding fitted values.

$$e_t = y_t - \hat{y}_t$$

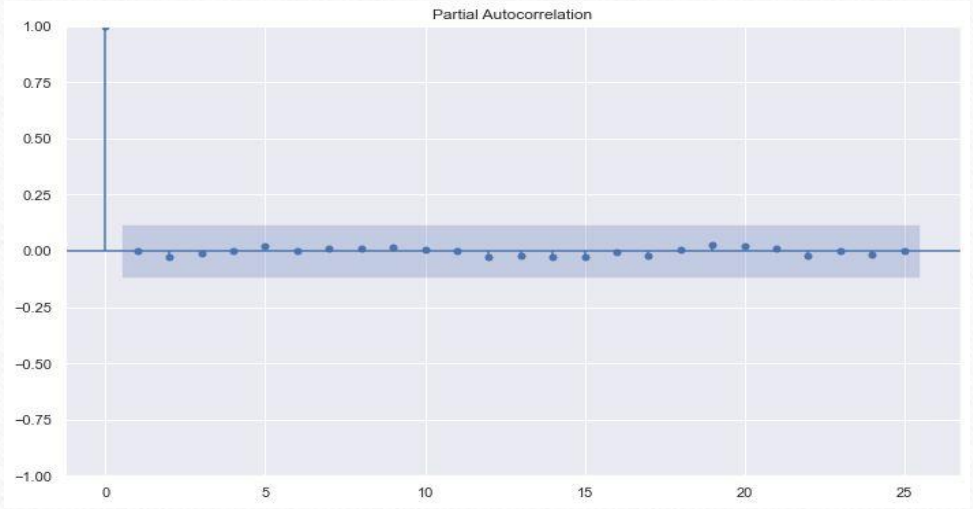


Graph: Residuals are stationary with mean zero and constant variance. This means that ARIMA model is good fit for the data, and residuals which are random error does not show any kind of trend or pattern in the data.

2. **ACF and PCF plots**



ACF plots:



PACF plots:

- ✓ All residuals are lying within the blue area of ACF and PACF plots.
- ✓ No significant correlation in the residuals series.
- ✓ All the residuals are statistically in control and the residuals are stationary

Limitation of Study

1. Only one year of Infosys stock's dataset is considered for the study purpose due to time circumstance. The study can be strengthen by increasing size of given dataset.

2. We have studied stock data, It is observed that observation of weekends are not available as stock market is closed in weekends.

3. Other affecting factors on stock prices are not considered deeply while interpreting data visualization.

Challenges

- ✓ Interpreting the decreasing trend of Infosys stock was the challenges
- ✓ Use of proper test to remove the stationarity, Implementing the ARIMA model and Study of Time series analysis and Python language while performing simultaneously were the challenge. To overcome these challenges we divided the work in teammates, understanding the concepts from one another.

Conclusion

- ✓ Closing stock price of Infosys company is showing decreasing trend.
- ✓ Augmented Dickey fuller test is used to check given series is stationary or not. The given series is non-stationary.
- ✓ Around 3.94% percentage MAPE (mean absolute percentage Error) implies the model is about 96.06% accurate in predicting the test set observations. The model with the lowest AIC and BIC value is generally considered to be the best-fitting model.
- ✓ In diagnosis of ARIMA model, residuals are stationary with mean zero, constant variance, no significant correlation.

References:

1. S.C. Gupta & V.K. Kapoor, (2007), Fundamentals of Applied Statistics; Sultan Chand & Sons Educational Publishers, Volume 4, chapter 2: Analysis of Time Series, 2.1 – 2.80.
2. Medhi.j. (July 1984), stochastic processes, published by Mohinder Singh Sejwal for wiley Eastern limited, (223-240)
3. Brockwell, P.J. and Davis, R. A. (2003). Introduction to Time Series Analysis, Springer
4. Chatfield, C. (2001). Time Series Forecasting, Chapmann & Hall, London
5. www.Investopedia.com (stock related)
6. www. python.org (programming language)
7. www.moneycontrol.com
8. www.infosys.com

Thank you!