A

PROJECT REPORT ON

STATISTICAL STUDY ON STOCK MARKET

SUBMITED

TO

पुण्यश्लोक अहिल्यादेवी होळकर
सोलापूर विद्यापीठ
।। विद्यया संपन्नता ।।
NAAC Accredited - 2015
'B' Grade (CGPA - 2.62)

DEPARTMENT OF STATISTICS

PUNYASHLOK AHILYADEVI HOLKAR SOLAPUR UNIVERSITY, SOLAPUR.

.

BY

BHAGWAN KIRAN PALMURI
PRANITA SATISH PHUKALE
KARUNA A. GADDAM
PRAGATI P. RESHIMKAR

UNDER GUIENDENCE
PROF. O.D. GHADAGE

Punyashlok Ahilyadevi Holkar Solapur University, Solapur

पुण्यश्लोक अहिल्यादेवी होळकर
सोलापूर विद्यापीठ
॥ विद्यया संपन्नता ॥
NAAC Accredited - 2015
'B' Grade (CGPA - 2.62)

CERTIFICATE

This is to certify that **Mr. Bhagwan Kiran Palmuri, Mrs. Karuna Amar Gaddam, Mrs.Pragati Prakash Reshimkar and Mrs. Pranita Satish Pukale** of M.Sc. Part- II (Statistics) has satisfactorily completed project on **"Statistical Study on Stock Market"** as per the syllabus Punyashlok Ahilyadevi Holkar Solapur University, Solapur during the academic year 2020-2021 under my supervision and successfully submitted the report.

Mr. O. D. Ghadge		Dr. V.B Ghute
Project guide	Examiner	Head of Department of Statistics

Place: Solapur

Date:

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all those people without whom this project could have never been completed. First and foremost, we would like to thank our respected Head of the department Dr. V. B. Ghute sir for their inexhaustible source of inspiration.

We would like to extend our gratitude to Prof. C. G. Gardi Sir, for his constant encouragement and providing very nice platform to learn.

We all are highly indebted to the teacher in charge prof. O. D. Ghadge sir for his guidance and constant supervision as well as for providing necessary information regarding our project and also for his support in completing the project..

We also thank our parents for their support.

THANK YOU.

INDEX

Introduction

The first organized stock exchange in India was started in 1875 at Bombay and it is stated to be the oldest in Asia. ... The Calcutta stock exchange was started in 1908 to provide a market for shares of plantations and jute mills. Then the madras stock exchange was started in 1920.

A market is a collection of buyers and sellers, facilitated by a stock exchange, which enables the determination of price for assets, e.g. company securities. In the share trading context, a stock market is where share prices are set, and where the organic supply and demand for shares can play out in primary and secondary markets. The sectors of the Indian economy that seem to be many investors will be buying interest in the safer sectors such as FMCG, PHARMA,BANK,REAL. Therefore, it is advisable for investors to choose sectors with good growth potential, good market share and consistent profitability.

Objective

- Effect of BANK, FMCG, PHARMA, IT, and REAL on SENSEX and NIFTY.

- To predict future value of SENSEX and NIFTY nearly 1 year

Methodology

As we know Society became queries to learn and earn saving their money in Share Markets Or in mutual funds it became important for our future to enjoy healthy lifestyle. Our approach to select this project is to study how the fluctuations occurring in share market.

Our project name is ‘Statistical study on Stock Market’ . As we know share market is collection of buyer and sellers, facilitated by a stock exchange. Share market contains different sectors like Bank FMGC, PHARMA, IT and Real. Now we are going to study How the Sensex and Nifty are affected by Bank, FMCG, PHARMA, IT and REAL also prediction for next two years. For this Statistical study we collected secondary data from BSE. India. Com and . … .. … .. …

For Analysis we used tools like Regression Model, Time series Analysis and we used software to draw the plots and to make model tables.

Software used for Analysis:

1. Ms-Excel

2. R software

3. Minitab

Statistical Analysis
On
Stock Market

Regression Analysis

Fit a multiple linear regression model

In multiple linear regression, there are p explanatory variables, and the relationship between the dependent variable and the explanatory variables is represented by the following equation:

$$\mathbf{Y}_i=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip}+\varepsilon_i$$

Where y_i=Sensex value, Nifty value.
x_i=stock values for five sectors (BANK, FMCG, IT, PHARMA, REAL).
Where, i=1,2,3,4,5
Where: β_0 is the constant term and β_1 to β_p are the coefficients relating the p explanatory variables to the variables of interest. So, multiple linear regression can be thought of an extension of simple linear regression, where there are p explanatory variables, or simple linear regression can be thought of as a special case of multiple linear regression, where p=1. The term linear is used because in multiple linear regression we assume that y is directly related to a linear combination of the explanatory variables

Regression Analysis: BSE vs BANK, FMCG, IT, PHARMA, REAL.

Hypothesis:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1: atleast one $\beta_i \neq 0$, i=1,2,3,4,5

At, $\alpha = 0.05$

ANOVA TABLE:

Source	DF	Adj SS	Adj MS	F-value	P-value
Regression	5	1.71773E+11	34354572370	128377.99	0.000
BANK	1	1575755862	1575755862	5888.37	0.000
FMCG	1	554091765	554091765	2070.56	0.000
IT	1	4453816835	4453816835	16643.26	0.000
PHARMA	1	79489483	79489483	297.04	0.000
REAL	1	264852389	264852389	989.71	0.000
Error	2446	654561458	267605		
Total	2451	1.72427E+11			

Conclusion:
In above table
p-value is less than level of significance hence we reject the null hypothesis
i.e there atleast one β_i is not equal zero
hence we conclude that BSE depends on all five sectors.

Model Summary

SS	R-Sq	R-Sq(adj)	R-sq(pred)
517.305	99.62%	99.62%	99.62%

Coefficients

Term	coef	SE coef	T-value	P-value
Constant	1981.3	96.8	20.47	0.000
BANK	0.45798	0.00597	76.74	0.000
FMCG	0.26798	0.00589	45.50	0.000
IT	0.71962	0.00558	129.01	0.000
PHARMA	0.11348	0.00658	17.23	0.000
REAL	10.962	0.348	31.46	0.000

Regression Equation

$$Y_1 = 1981.3 + 0.45798\,X_1 + 0.26798X_2 + 0.71962X_3 + 0.11348X_4 + 10.962X_5$$

Where,
Y_1= BSE
X_1= BANK
X_2= FMCG
X_3= IT
X_4 = PHARMA
X_5= REAL

Regression Analysis: NSE vs BANK, FMCG, IT, PHARMA, REAL

Analysis of Variance

Source	DF	AdjSS	AdjMS	F-value	P-value
Regression	5	15009552641	3001910528	1411.13	0.000
BANK	1	177159388	177159388	83.28	0.000
FMCG	1	38968507	38968507	18.32	0.000
IT	1	197205779	197205779	92.70	0.000
PHARMA	1	101406626	101406626	47.67	0.000
REAL	1	53932110	53932110	25.35	0.000
Error	2447	5205517544	2127306		
Lack -of -fit	2446	5205517544	2128176		
Pure Error	1	0	0		
Total	2452	20215070184			

In above table
p-value is less than level of significance hence we reject the null hypothesis i.e there atleast one β_i is not equal zero
hence we conclude that NSE depends on all five sectors.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1458.53	74.25%	74.20%	74.09%

Coefficients

Term	Coef	SE Coef	T-value	P-value
Constant	75	273	0.28	0.783
BANK	0.1536	0.0168	9.13	0.000
FMCG	0.0710	0.0166	4.28	0.000
IT	0.1514	0.0157	9.63	0.000
PHARMA	0.1282	0.0186	6.90	0.000
REAL	4.944	0.982	5.04	0.000

Regression Equation

$$\mathbf{Y_2 = 75 + 0.1536\, X_1 + 0.0710\, X_2 + 0.1514\, X_3 + 0.1282 X_4 + 4.944 X_5}$$

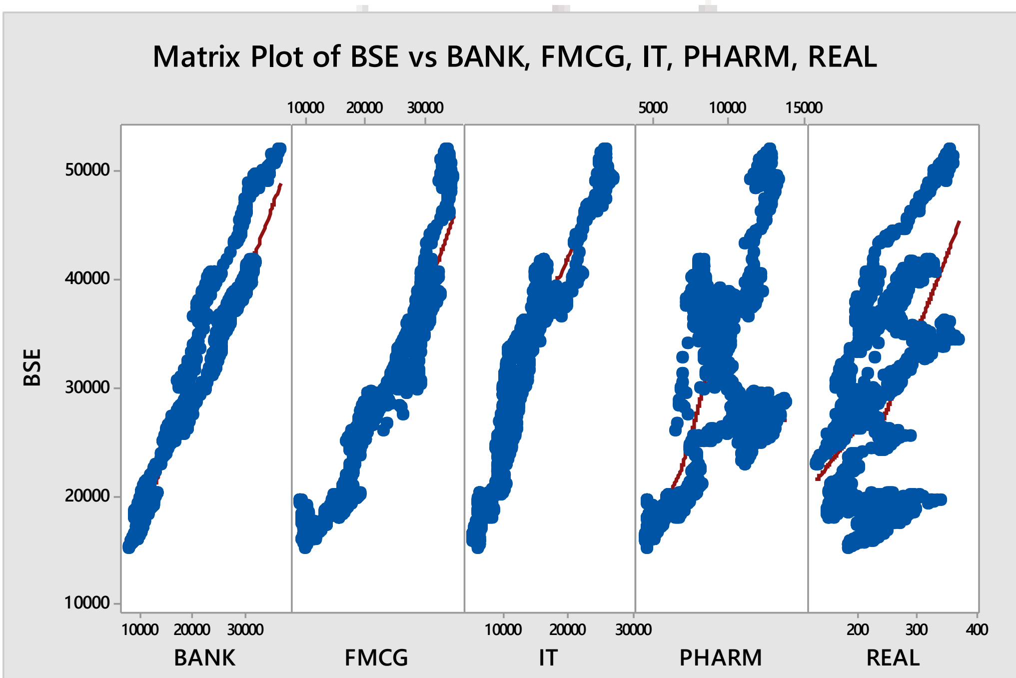
Where,
Y_2= NSE
X_1= BANK
X_2= FMCG
X_3= IT
X_4 = PHARMA
X_5= REAL

Correlation: BSE, BANK, FMCG, IT, PHARMA, REAL

	BSE	BANK	FMCG	IT	PHARMA
BSE	1				
BANK	0.971	1			
FMCG	0.965	0.944	1		
IT	0.946	0.856	0.889	1	
PHARMA	0.579	0.506	0.572	0.622	1
REAL	0.477	0.545	0.371	0.327	-0.163

In above table

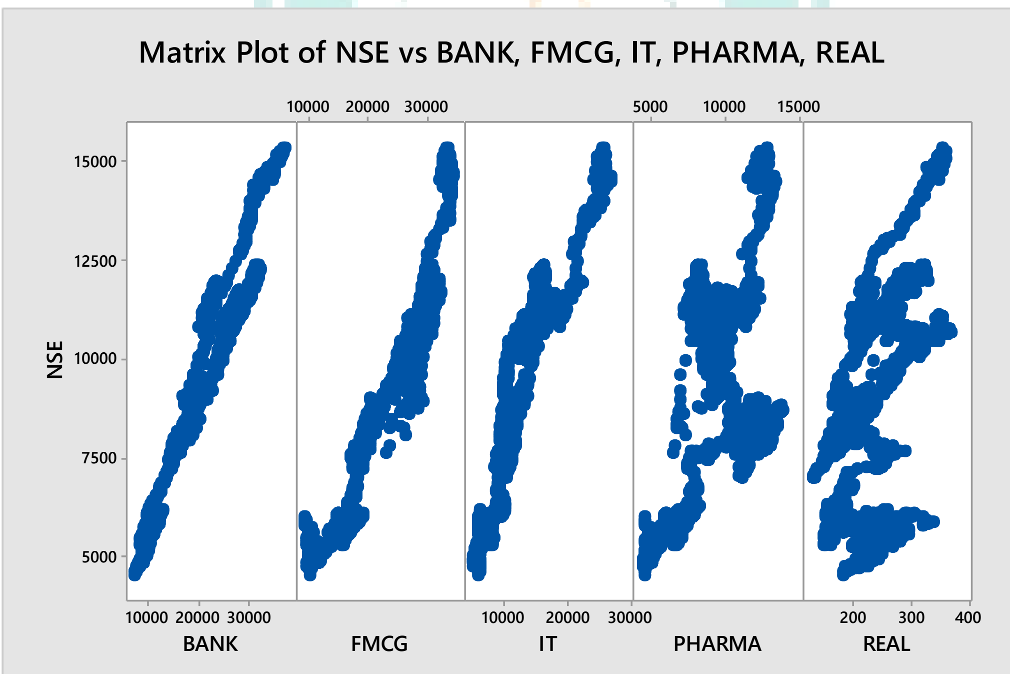
Correlation coefficients whose magnitude are between 0.5 and 0.7 indicate variables which can be considered moderately correlated. Correlation coefficients whose magnitude are between 0.3 and 0.5 indicate variables which have a low correlation and nearly 1 value indicates highly correlated i.e. BSE with BANK, FMCG, IT indicates highly correlated and PHARMA and REAL indicates negatively correlated and so on…

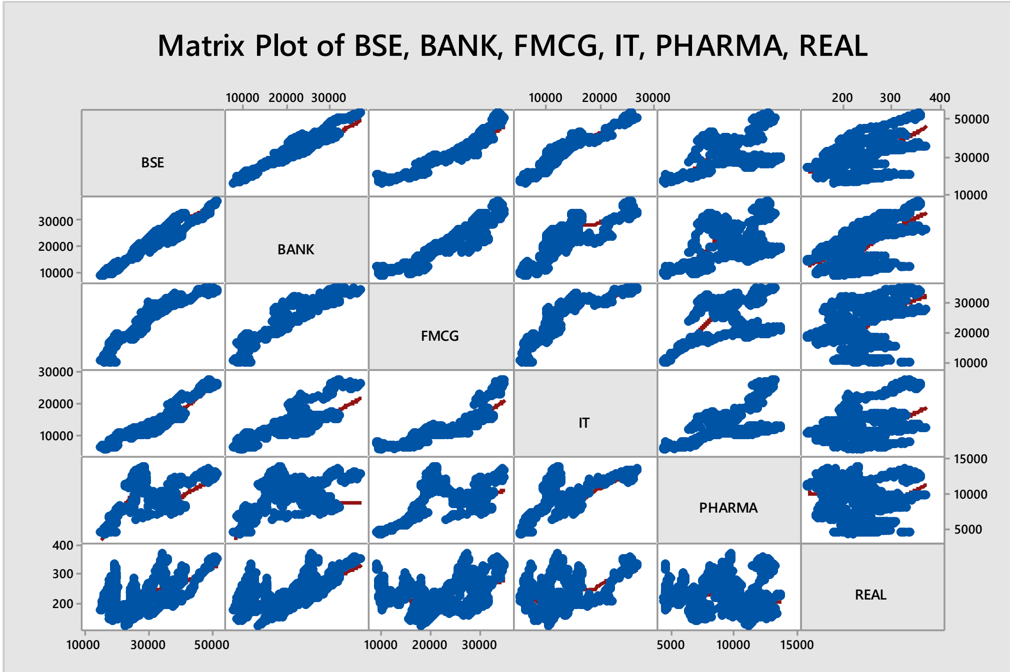


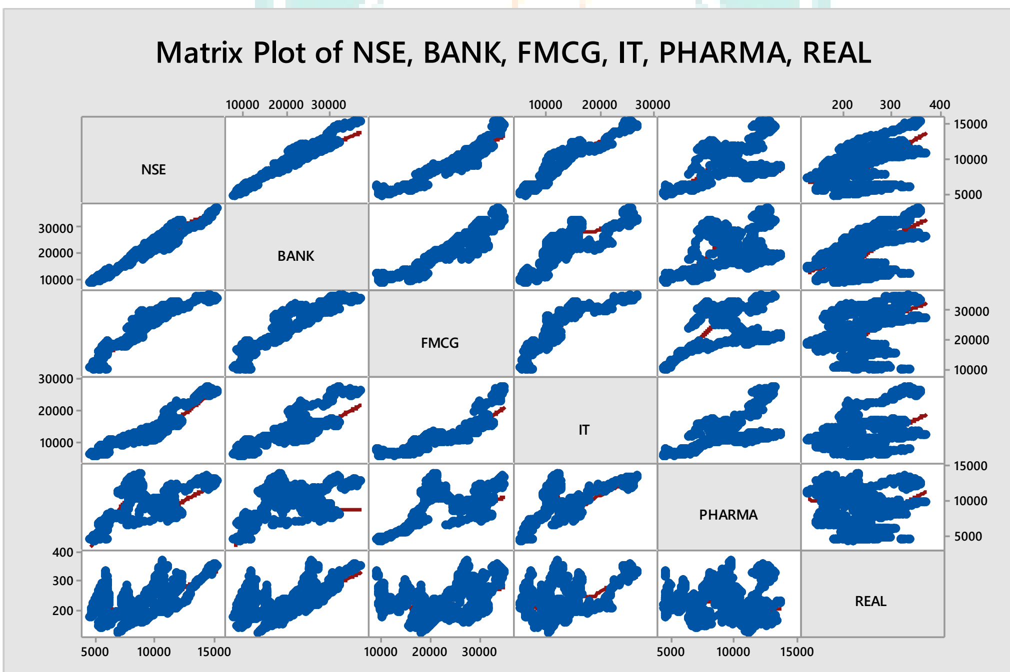
Correlation: NSE, BANK, FMCG, IT, PHARMA, REAL

	NSE	BANK	FMCG	IT	PHARMA
NSE	1				
BANK	0.978	1			
FMCG	0.830	0.944	1		
IT	0.805	0.856	0.889	1	
PHARMA	0.529	0.506	0.572	0.622	1
REAL	0.416	0.545	0.371	0.327	-0.163

NSE, BANK, FMCG are highly correlated than PHARMA and REAL.

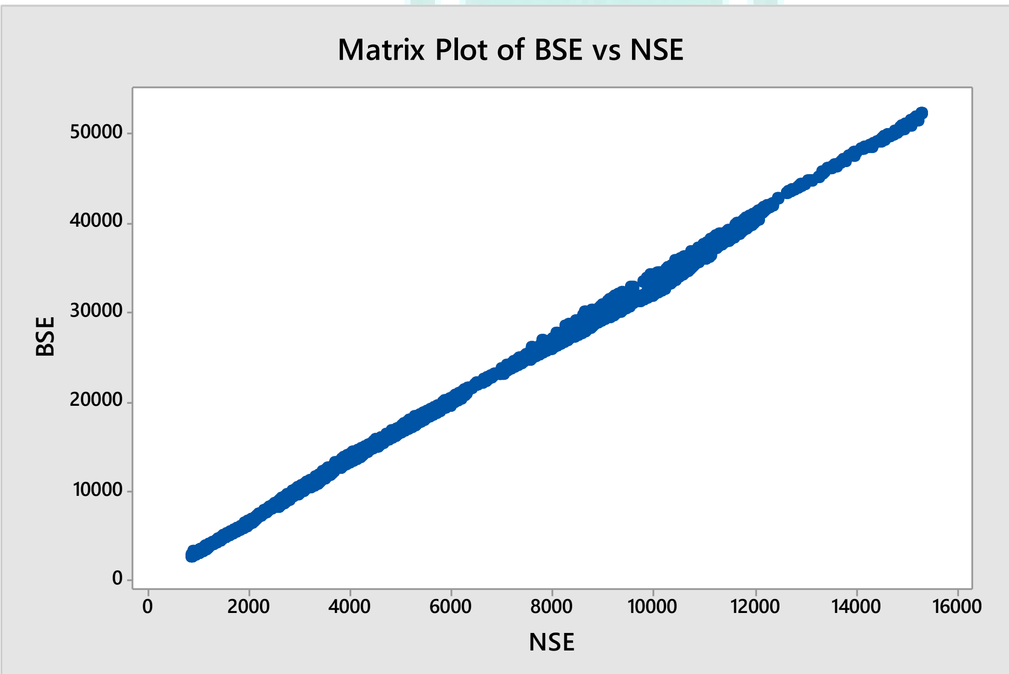



Matrix Plot of BSE, BANK, FMCG, IT, PHARMA, REAL
BSE
BANK
FMCG
IT
PHARMA
REAL


Matrix Plot of NSE, BANK, FMCG, IT, PHARMA, REAL
NSE
BANK
FMCG
IT
PHARMA
REAL

Correlation Between NSE and BSE (20 year data)

Pearson correlation of BSE and NSE = 0.998
P-Value = 0.000



Time Series Analysis

Time series analysis (TSA) requires the time series of interest to be **stationary**. What this means is that the behaviour of each timepoint (x_t1) in relation to other time points (x_t2, x_t3, …) in each series is identical throughout the progression of time (for all time points).

Time-series analysis is based on the assumption that

- random error terms are normally distributed.
- there are dependable correlations between the variable to be forecast and other independent variables.
- past patterns in the variable to be forecast will continue unchanged into the future.
- the data do not exhibit a trend

Auto-covariance and Auto-correlation

- The **auto-covariance** function of two time points (*x_s*, *x_t*) is defined as $\gamma_x(s,t) = \mathrm{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$
- Assuming that the time series mean *miu_s* = *miu_t* = 0, this provides some insight on the linear dependence of two time points, regardless of time progression. If the autocovariance function is applied to the same time point (s=t), it would simply be measuring the variance.
- Drawing parallels to the **Pearson Correlation Coefficient** equation between two series (X ,Y):

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

the **Auto-correlation** function (ACF) of two time points (*x_s*, *x_t*) is defined as:

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}.$$

With the denominator as the square root of the variance product between the two time points, the ACF provides a "normalised" correlation measure between the two time points and should range between -1 to 1.

To see the effects of ACF, let's take a look at some simulated data. Assume a time series of autoregressive AR(1) series, where *x_t* is dependent on the time point before it (t-1).

$$x_t = 0.7\, x_{t-1} + \omega_t$$

Tests

1. Augmented Dickey-Fuller (ADF) Test

Augmented Dickey Fuller test (ADF Test) is **a statistical test used to test whether a given Time series is stationary or not**

2. Ljung–Box test

The Ljung (pronounced Young) Box test (sometimes called the modified Box-Pierce, or just the Box test) is a way **to test for the absence of serial autocorrelation, up to a specified lag k.**

Arima: Autoregresive Integrating Moving Average

The fact of introducing ARIMA models comes from the assumption that we are not working with a non stationary dataset series. We say that time series datasets are stationary when their means, variance and autocovariance don't change during time. The mayority of economic time series are not stationary, but differencing them determined number times makes them stationary. Whit this previous operation we can apply ARIMA models to any stock price. In general we say that a temporal set Yt admits an integrated autoregressive representation with p, q and d moving average orders respectively. We denote this forecasting model by ARIMA(p, d, q).

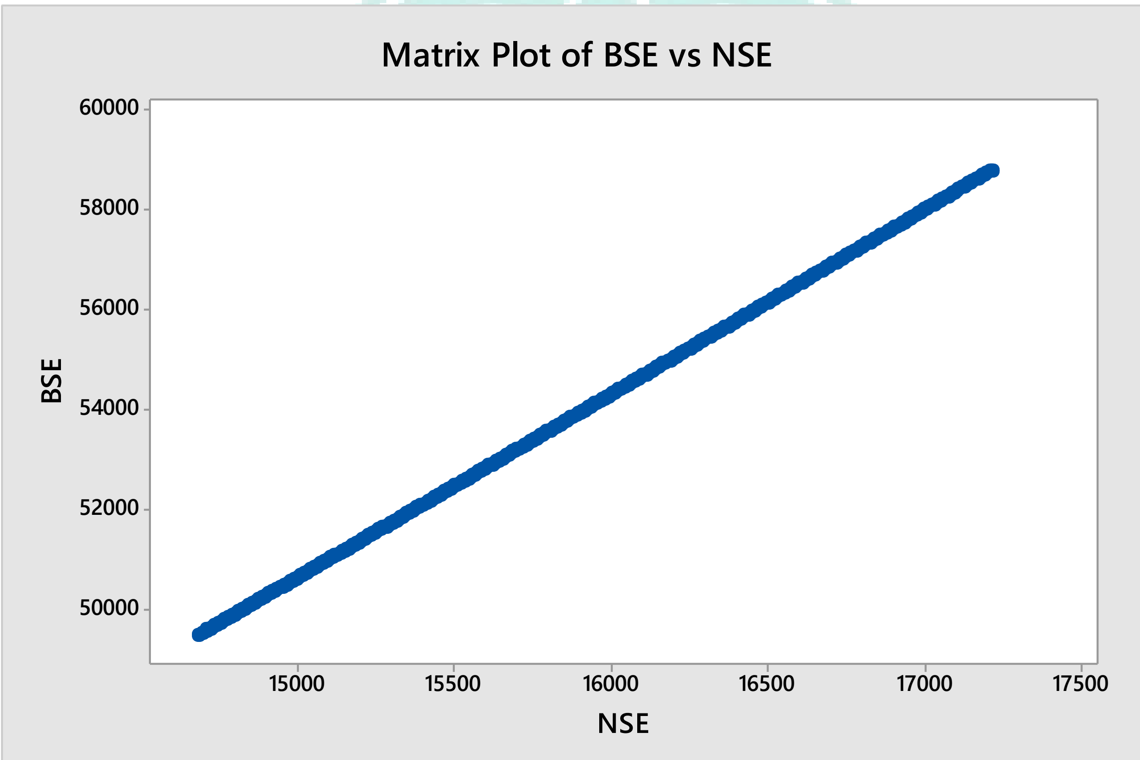
Yt=c+ϕ1ydt−1+ϕpydtp+...+θ1et−1+θqetq+etYt=c+ϕ1ydt−1+ϕpydt p+...+θ1et−1+θqetq+et

In ARIMA, p denotes the number of autoregressive terms, d denotes the number of times that the set should be differenciated for making it stationary. The last parameter q denotes the number of invertible moving average terms.

Correltion Between BSE and NSE (Predicting Values)

Pearson correlation of BSE and NSE = 1.000

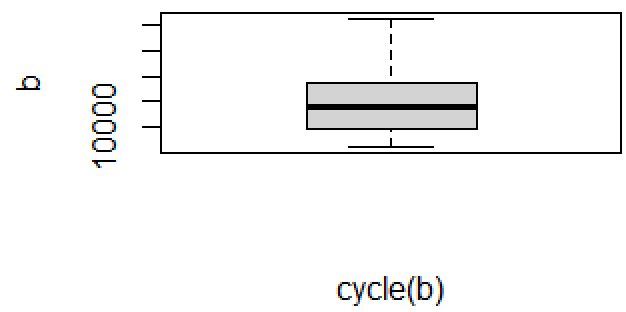
P-Value = 0.000

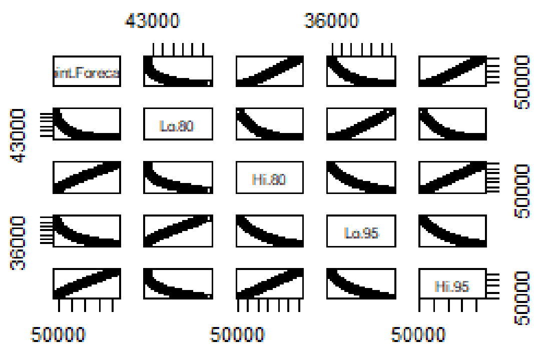


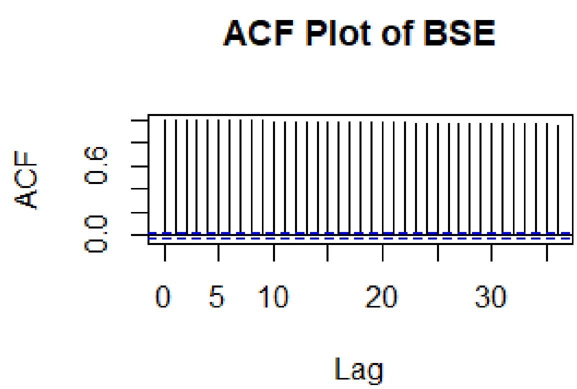
Conclusion

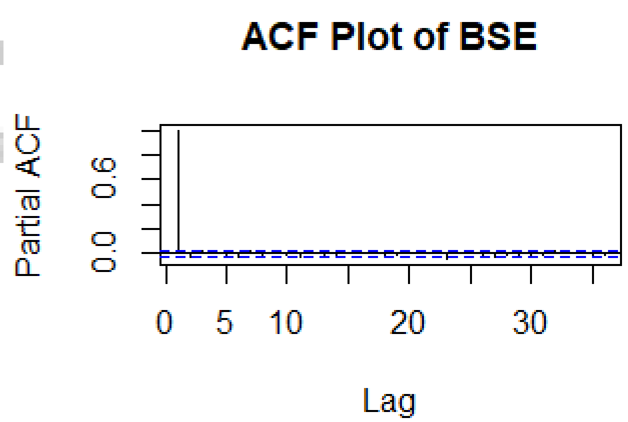
The correlation between BSE and NSE (**historical data**) is **0.998** and the correlation between BSE and NSE (**prediction values**) is **1** i.e. we calculating the prediction values are nearly correct.

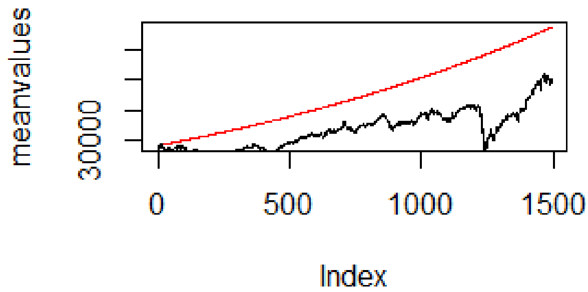
Graphs (BSE)


b
10000
cycle(b)

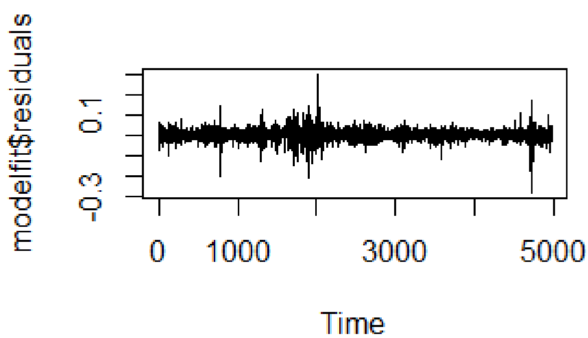

43000
36000
La.80
Hi.80
La.95
Hi.95
50000
50000
50000
50000
50000
50000
43000
36000


ACF Plot of BSE
ACF
0.0
0.6
0
5
10
20
30
Lag

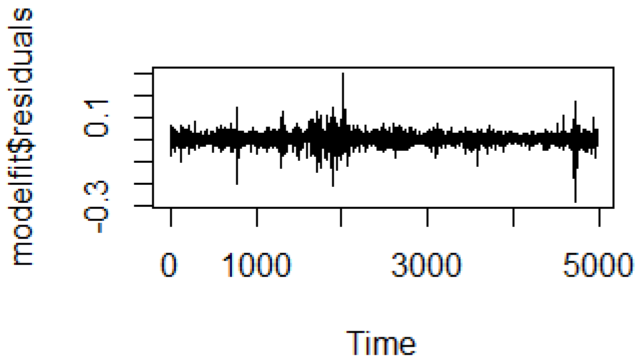

ACF Plot of BSE
Partial ACF
0.0
0.6
0
5
10
20
30
Lag


meanvalues
30000
0
500
1000
1500
Index

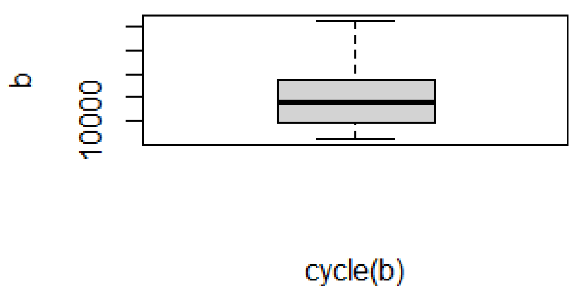

BSE
50000
40000
30000
20000
10000
2005
2010
2015
2020


modelfit$residuals
0.1
-0.3
0
1000
3000
5000
Time

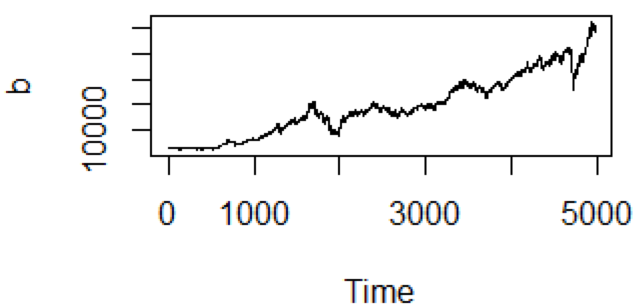
Graphs (NSE)



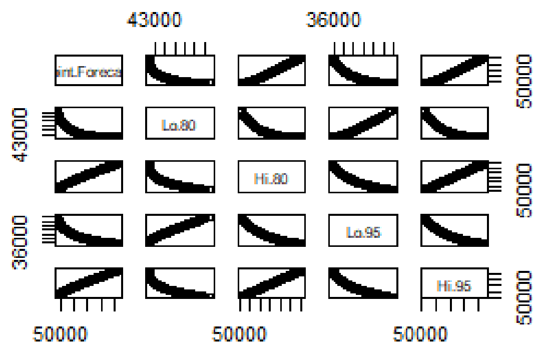
modelfit$residuals
0.1
-0.3
0
1000
3000
5000
Time

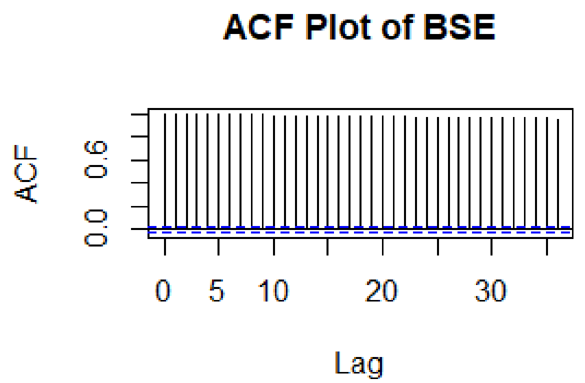


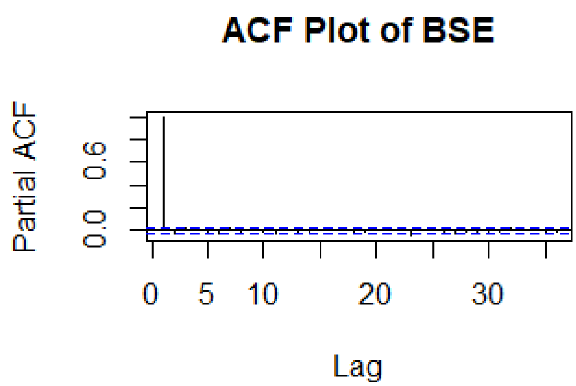
b
10000
cycle(b)

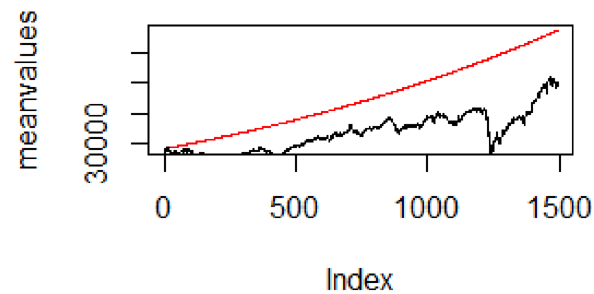


b
10000
0
1000
3000
5000
Time


43000
36000
intLForeca
Lo.80
Hi.80
Lo.95
Hi.95
50000
50000
50000
43000
36000
50000
50000
50000


ACF Plot of BSE
ACF
0.0
0.6
0
5
10
20
30
Lag


ACF Plot of BSE
Partial ACF
0.0
0.6
0
5
10
20
30
Lag


meanvalues
30000
0
500
1000
1500
Index


BSE
50000
40000
30000
20000
10000
2005
2010
2015
2020

Reference

https://www.analyticsvidhya.com/blog/

https://rpubs.com/kapage/523169

https://www.kaggle.com/rohanrao/nifty50-stock-market-data

Data collection

https://www.bseindia.com/

https://www1.nseindia.com/

https://in.finance.yahoo.com/

https://www.marketwatch.com/

R code (BSE)

```
library(readxl)

library(Quandl)

library(tidyverse)

library(ggplot2)

library(tidyquant)

library(tseries)

library(forecast)
```

```
library(lubridate)

library(openxlsx)


 BSE<-read_excel("D:/project/COMPLETE/EXCEL DATA/20 yr NSE
                              BSE.xlsx", sheet = "BSE")

a<-BSE$Close

 b<-ts(BSE$Close,start = c(2001,4),end=c(2021,3),frequency =249.05)

 b<-b[1:4974]

 plot.ts(b)

 boxplot(b~cycle(b))

p <- ggplot(BSE, aes(x=Date, y=Close)) +

  geom_line(color="blue", size=0.8) +

  theme_minimal() +

  xlab("") +

  ylab("BSE")

p##Increasing trend observed


        ####To check stationarity

adf.test(a, alternative="stationary", k=0)


#IF p-value is less than 0.05 so this implies that ts is stationary

b1<-acf(b, plot = T, main = "ACF Plot of BSE")
```

```
b2<-pacf(b, plot = T, main = "ACF Plot of BSE")

# ACF loooking good,As lag increases-fast decrease in ACF value- all in Confidence Interval


#We apply auto arima to the dataset

modelfit <- auto.arima(b, lambda = "auto");modelfit

plot(modelfit$residuals)

Box.test(modelfit$residuals,  type="Ljung-Box")

f<-forecast(modelfit, h=365)

f= data.frame(f)

q <- plot(f)

head(f$lower)

head(f$upper)

head(f$mean)


###Below work is optional######

#Dividing the data into train and test, applying the model

N = length(a)

n = 0.7*N

train =a[1:n]

test  = a[(n+1):N]

trainarimafit <- auto.arima(train, lambda = "auto")
```

```
predlen=length(test)

trainarimafit <- forecast(trainarimafit, h=predlen)

#Plotting mean predicted values vs real data

meanvalues <- as.vector(trainarimafit$mean)

precios <- as.vector(test)

plot(meanvalues, type= "l", col= "red")

lines(precios, type = "l")


####Now Exporting data####

a=write.xlsx(f,file="BSE.xlsx");a
```

+

k

9