KLEF
KONERU LAKSHMAIAH EDUCATION FOUNDATION
(Deemed to be university estd, u/s, 3 of the UGC Act, 1956)
(NAAC Accredited "A" Grade University)

PROJECT BASED LAB REPORT

On

FORECASTING THE CORONA OUTBREAK MARCH- 2020

Submitted in partial fulfilment of the

Requirements for the award of the Degree

of Bachelor of Technology

In

Computer science and Engineering

Under the esteemed guidance of

V. Bhavani
(Assistant Professor)

By

M.V.S. Pavan Kalyan

(170030860)

(DST-FIST Sponsored Department)

K L EDUCATION FOUNDATION

Green Fields, Vaddeswaram, Guntur District-522 502

2019-2020

K L EDUCATION FOUNDATION

DEPARTMENT OF COMPUTER SCIENCE ANDENGINEERING

(DST-FIST Sponsored Department)

CERTIFICATE

This is to certify that this project based lab report entitled **"FORECASTING THE CORONA OUTBREAK MARCH – 2020"**is a bonafide work done by M.V.S. Pavan Kalyan (170030860) in the course **17CS3065R Big Data Analytics** in partial fulfilment of the requirements for the award of Degree in Bachelor of Technology in **COMPUTER SCIENCE AND ENGINEERING** during the Even Semester of Academic year 2020-2021.

Faculty in Charge	**Head of the Department**
V. Bhavani	**Dr. V. Hari Kiran**

K L EDUCATION FOUNDATION

DEPT OF COMPUTER SCINCE AND ENGINEERING

(DST-FIST Sponsored Department)

DECLARATION

We hereby declare that this project based lab report entitled **"FORECASTING THE CORONA OUTBREAK MARCH - 2020"** has been prepared by us in the course **17CS3065 Big Data Analytics** in partial fulfilment of the requirement for the award of degree bachelor of technology in **COMPUTER SCIENCE AND ENGINEERING** during the Even Semester of the academic year 2019-2020.We also declare that this project-based lab report is of our own effort and it has not been submitted to any other university for the award of any degree.

Date: 05-05-2020

Place: Vijayawada

ACKNOWLEDGEMENT

Our sincere thanks to **V. Bhavani** in the Lab for her outstanding support throughout the project for the successful completion of the work.

We express our gratitude to **Dr P.Vidyullatha** Course Co-ordinator for **17CS3065R Big Data Analytics** course in the Computer Science and Engineering Department for providing us with adequate planning and support and means by which we can complete this project-based Lab.

We express our gratitude to **Dr. V. HARIKIRAN,** Head of the Department for computer science and Engineering for providing us with adequate facilities, ways and means by which we can complete this project-based Lab.

We would like to place on record the deep sense of gratitude to the honourable Vice Chancellor, K L University for providing the necessary facilities to carry the project-based Lab.

Last but not the least, we thank all Teaching and Non-Teaching Staff of our department and especially our classmates and our friends for their support in the completion of our project-based Lab.

Name of the student

M.V.S. Pavan Kalyan
170030860

TABLE OF CONTENTS

ABSTRACT

What will be the global impact of the novel coronavirus (COVID-19)? Accurate forecasting the spread of confirmed cases as well as analysis of the number of deaths and recoveries.

Forecasting, however, requires ample historical data. At the same time, no prediction is certain as the future rarely repeats itself in the same way as the past. Moreover, forecasts are influenced by the reliability of the data, vested interests, and what variables are being predicted. Also, psychological factor play a significant role in how people perceive and react to the danger from the disease and the fear that it may affect them personally. It introduces an objective approach to predicting the continuation of the COVID-19 using a simple, but powerful method to do so.

Assuming that the data used is reliable and that the future will continue to follow the past pattern of the disease, our forecasts suggest a continuing increase in the confirmed COVID-19 cases with sizable associated uncertainty. The risks are far from symmetric as underestimating its spread like a pandemic and not doing enough to contain it is much more severe than overspending and being over careful when it will not be needed. It describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides objective forecasts for the confirmed cases of COVID-19.

1. INTRODUCTION

1.1 INTRODUCTION:

The accuracy of traditional forecasting largely depends on the availability of data to base its predictions and estimates of uncertainty. In outbreaks of epidemics there is no data at all in the beginning and then limited as time passes, making predictions widely uncertain. Besides, there are concerns that the data may not be reliable, as was the case of bird flu and SARS when the number of affected people and deaths were misreported to hide the extent of the epidemic.

Related to forecasting accuracy and uncertainty, there is a more severe problem that has to do the perception of epidemics and pandemics. Politicians are concerned with regards to the measures to be taken while the general population fears about the impact on the epidemic on their health/lives. Of course, the big problem is the asymmetry of risks and the irrational fear of a pandemic with its possible catastrophic consequences, as happened with the 1918 Spanish flu that killed an estimated 50 million worldwide.

In contrast, the SARS killed a total of 774 in 2003 and the bird flu around 100 in 1997. COVID-19 has resulted in an estimated 5.8 thousand deaths until (15/03/2020) At the same time, there is much less concern over the seasonal flu that kills about 646,000 people worldwide each year Medical predictions are often not accurate while their uncertainty is seriously underestimated.

Predicting the future of epidemics and pandemics is much more difficult as the number of cases to be studied can be measured in one hand. At one end of the scale is the case of SARS where the fear of becoming a pandemic was overblown, resulting in overspending and the application of restrictive measures to be contained that it turned out to be unnecessary.

At the other end is the Spanish flu that turned out to be a serious pandemic with catastrophic consequences, 8 arguably in a different era when communication and the ability to raise public awareness were limited. Despite the inaccuracies associated with medical predictions, still forecasting is valuable in allowing us to better understand the current situation and plan for the future.

1.2 PROBLEM DEFINITION:

- Forecasting the outbreak of corona in march-2020.

- To analyze corona outbreak.

- We focus on the cumulative daily figures aggregated globally of the three main variables of interest: confirmed cases, deaths and recoveries. These were retrieved from the Kaggle. While all three data patterns show an exponential increase, the trends of both the confirmed cases and the deaths were reduced in the mid of February; a second exponential increase is observed in late March and April as a result of the increased number of cases in INDIA. At the same time, the number of recovered cases is steadily increasing.

1.3 SCOPE:

- To analyze the effect of the outbreak of corona in the month march 2020 worldwide.

1.4 PURPOSE:

- The uncertainty surrounding an unknown, novel coronavirus can spark a global alarm, leading a Harvard Professor stating that 40-70% of the global population might be infected in the coming year Regardless of what one’s beliefs are, we believe that forecasts and their associated uncertainty can and should be an integral part of the decision-making process, especially in high-risk cases. Apart from the significant public health concerns, the dangers imposed on global supply chains and the economy as a whole are also considerable.

1.5 PROBLEM AND EXISTING TECHNOLOGY:

- This Problem states about forecasting the corona outbreak. Here We used ARIMA model to forecast the dataset for achieving the result.

1.6 PROPOSED SYSTEM:

- Analyzing the corona infected patients date. With implementation in R.
- Data Collection-past data is collected from Kaggle.
- Three datasets are considered ‘TotalConfirmedCases’, ‘Cured’,’Deaths’ .

2. REQUIREMENTS AND ANALYSIS

2.1 PLATFORM REQUIREMENTS

Hardware/ Software	Hardware / Software element	Specification /version
	Processor	Intel Core i5
	RAM	8GB
	Hard Disk	10GB
	OS	
	R language	

2.2 MODULE DESCRIPTION

Dataset Description:

The dimensions of the dataset used is 1286 * 4. It consist of four columns (Date, Total Confirmed Cases, Cured, Deaths).The total dataset consist of no.of Confirmed Cases, no.of Deaths, no.of cured cases on each date at the month of the march-2020.

1. Total Confirmed Cases ------- Count of Total Victims of Corona.
2. Cured -------- Count of Cured Cases from Corona.
3. Deaths -------- Count of Death Cases by Corona.

3. DESIGN & IMPLEMENTATION

3.1 ALGORITHM

ARIMA:

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modelled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q

where,

p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. More on that once we finish ARIMA.

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

ARIMA model in words:

Predicted Yt = Constant + Linear combination Lags of Y (up to p lags) + Linear Combination of Lagged forecast errors (up to q lags).

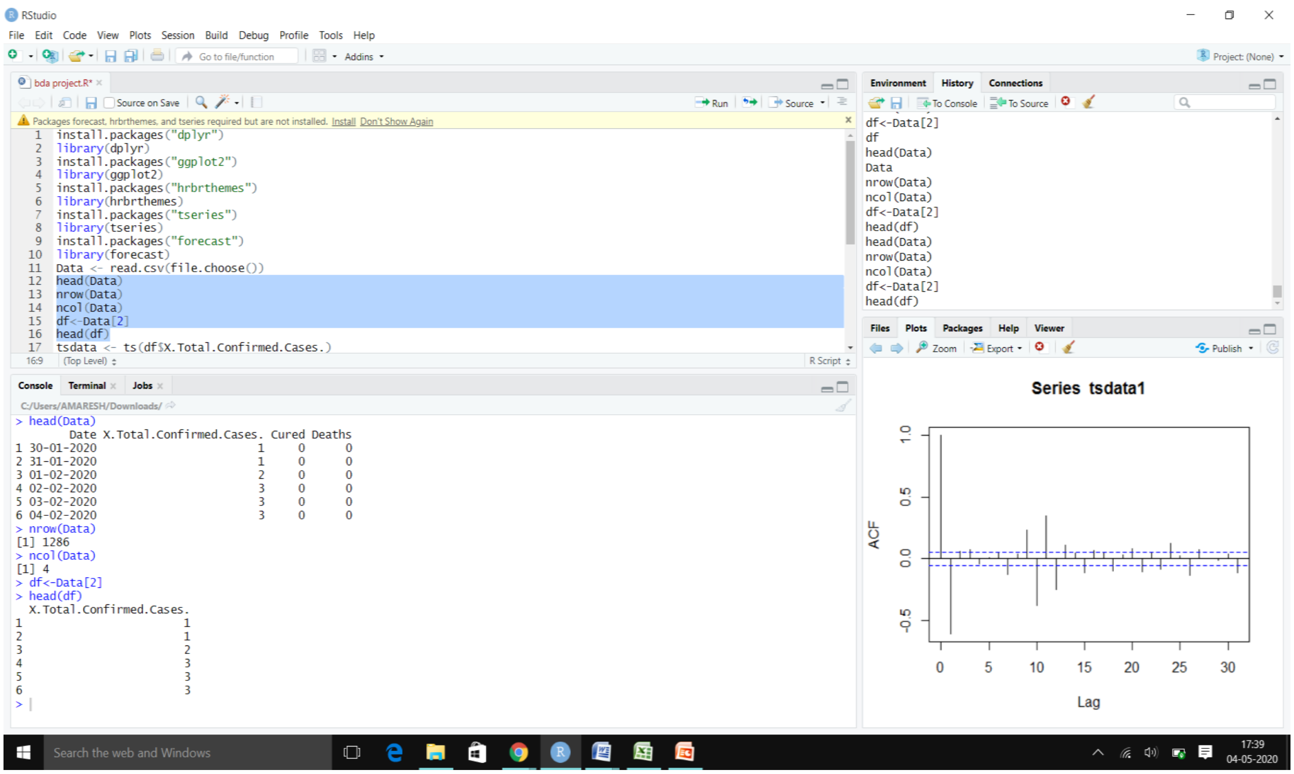
3.2 PSEUDOCODE

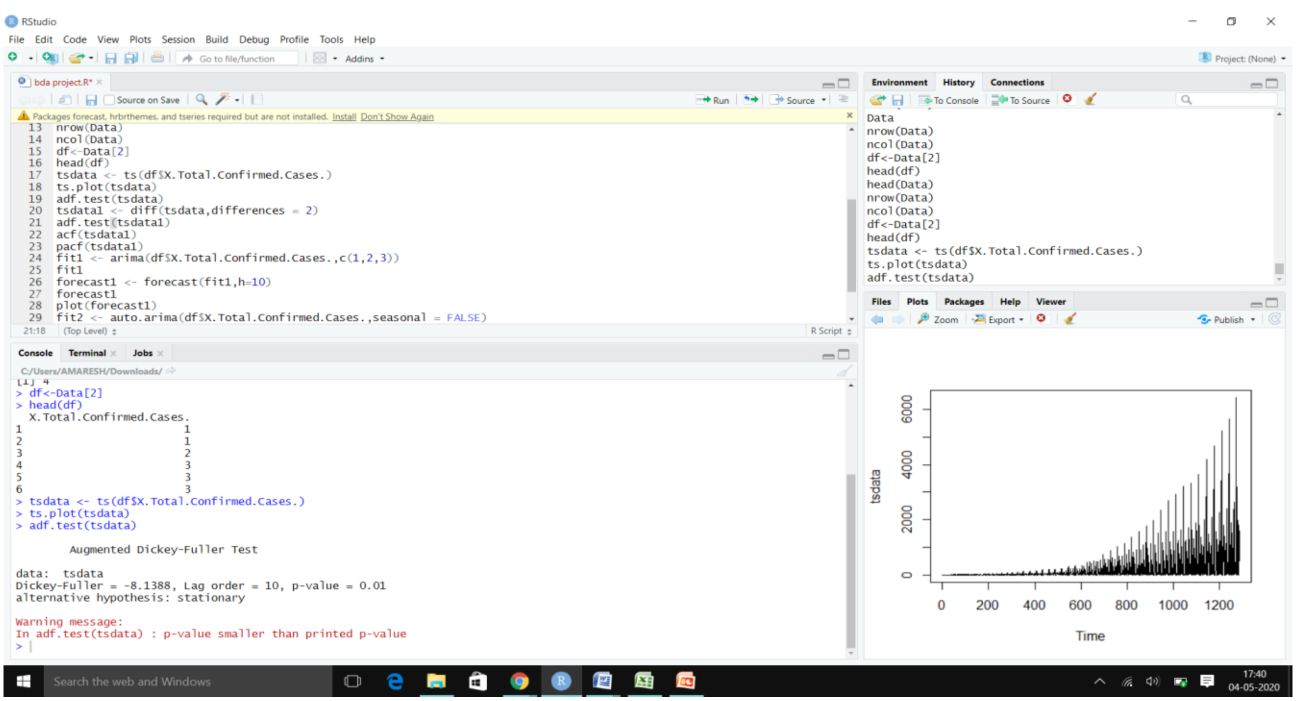
```
install.packages("dplyr")

library(dplyr)

install.packages("ggplot2")

library(ggplot2)

install.packages("hrbrthemes")

library(hrbrthemes)

install.packages("tseries")

library(tseries)

install.packages("forecast")

library(forecast)

Data <- read.csv(file.choose())

head(Data)
```

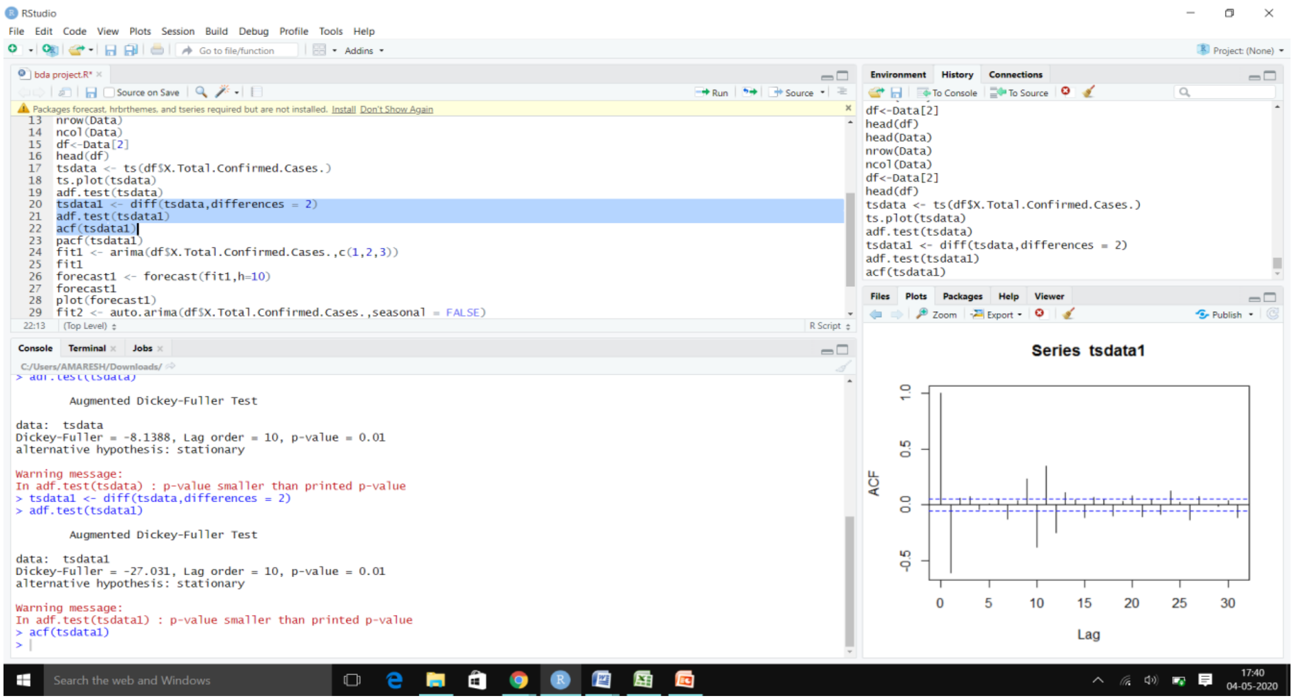
```
Data

nrow(Data)

ncol(Data)

df<-Data[2]

df

tsdata <- ts(df$X.Total.Confirmed.Cases.)

ts.plot(tsdata)

adf.test(tsdata)

tsdata1 <- diff(tsdata,differences = 2)

adf.test(tsdata1)

acf(tsdata1)

pacf(tsdata1)

fit1 <- arima(df$X.Total.Confirmed.Cases.,c(1,2,3))

fit1

forecast1 <- forecast(fit1,h=10)

forecast1
```

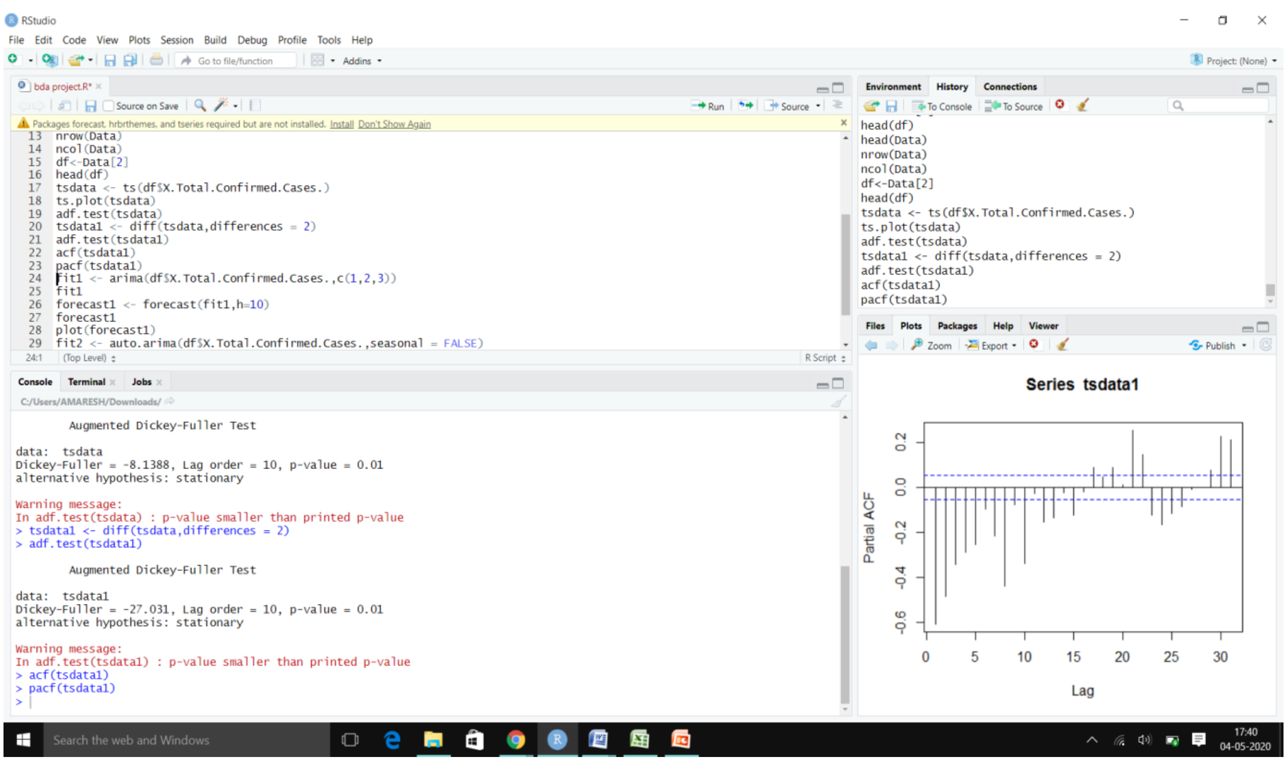
```
plot(forecast1)

fit2 <- auto.arima(df$X.Total.Confirmed.Cases.,seasonal = FALSE)

fit2

forecast2 <- forecast(fit2, h = 9)

forecast2

plot(forecast2)
```

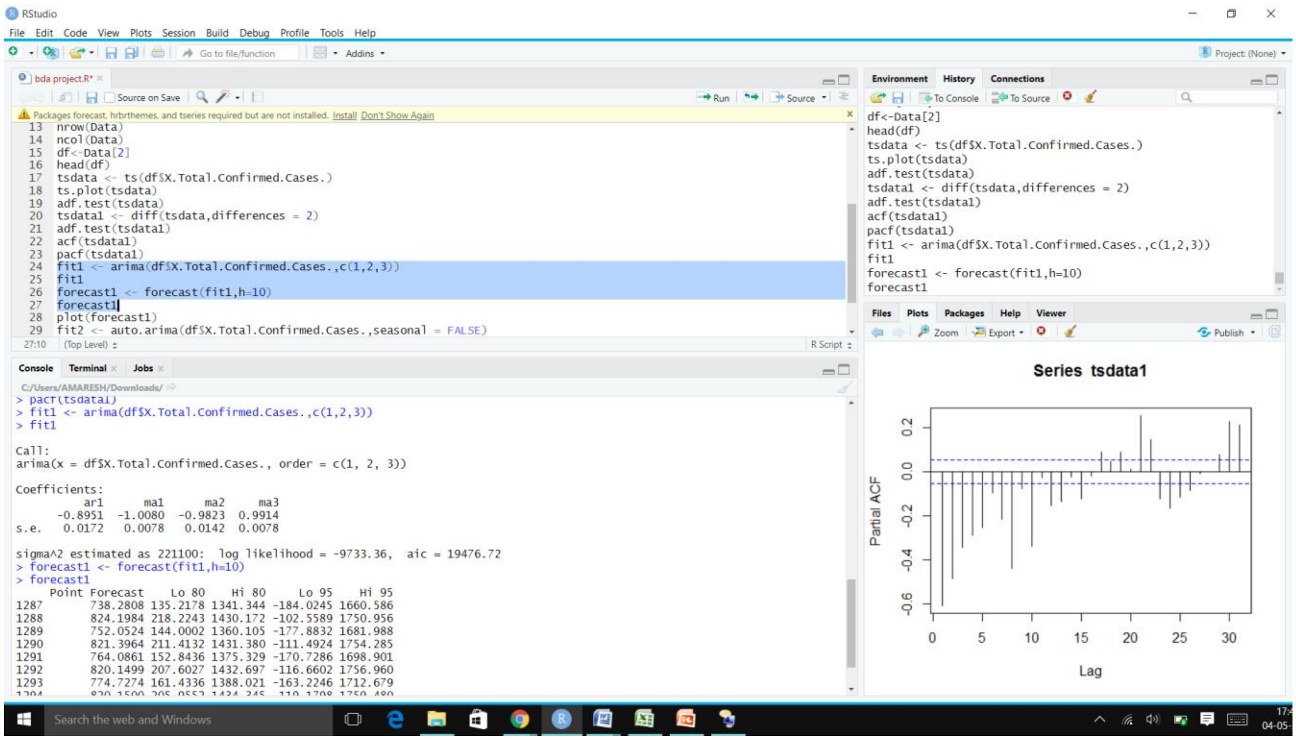
4. SCREENSHOTS

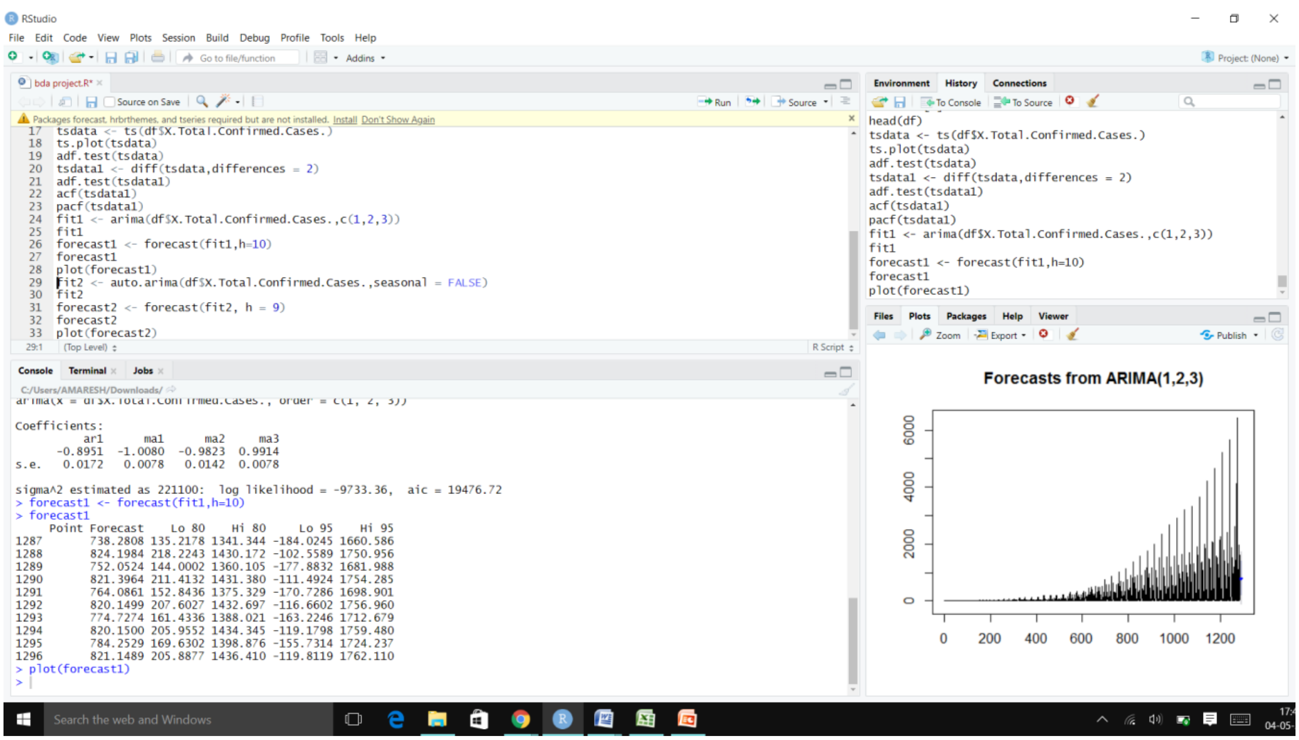


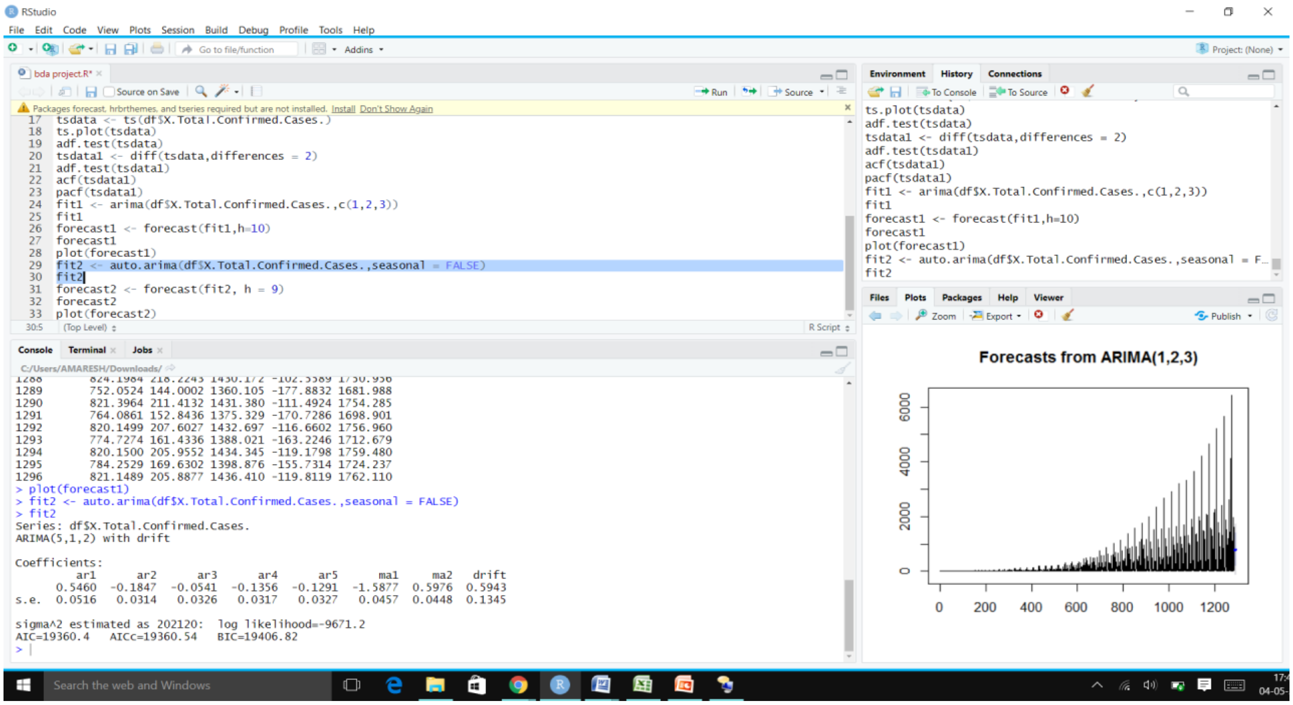



Series tsdata1
ACF
Lag

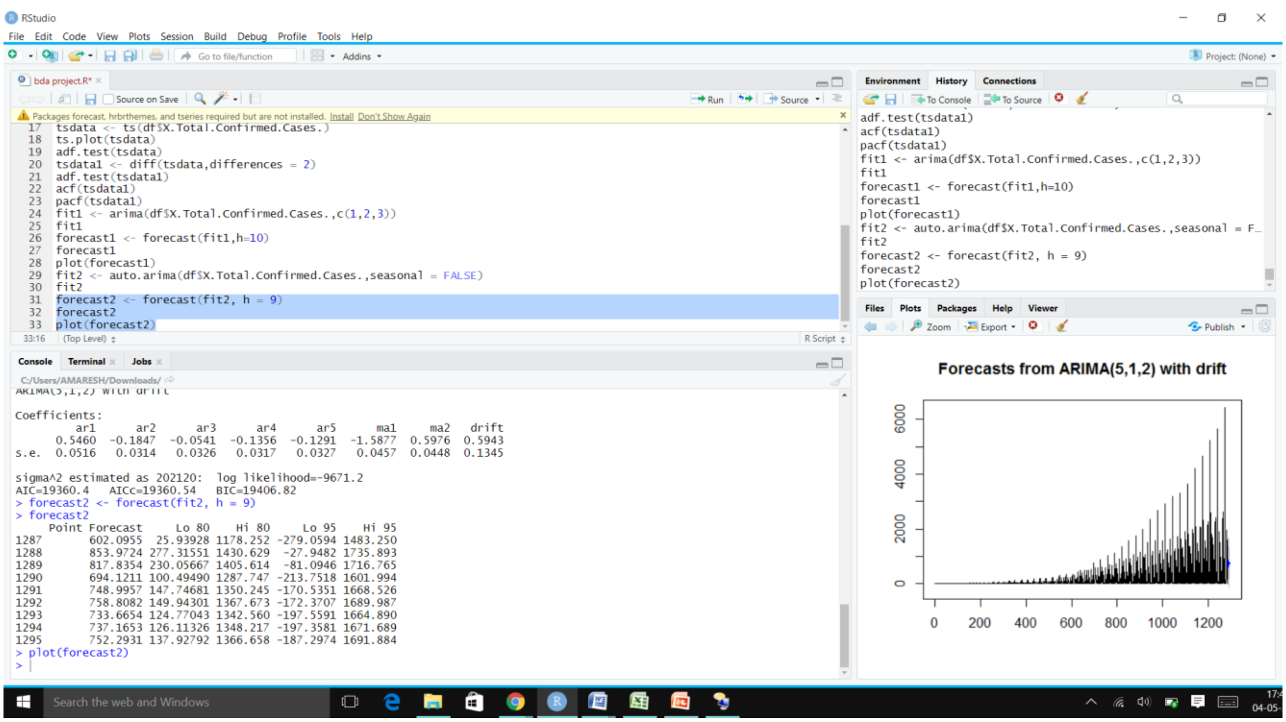

Series tsdata1
Partial ACF
Lag


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
bda project.R*
Packages forecast, hrbrthemes, and tseries required but are not installed. Install Don't Show Again
nrow(Data)
ncol(Data)
df<-Data[2]
head(df)
tsdata <- ts(df$X.Total.Confirmed.Cases.)
ts.plot(tsdata)
adf.test(tsdata)
tsdata1 <- diff(tsdata,differences = 2)
adf.test(tsdata1)
acf(tsdata1)
pacf(tsdata1)
fit1 <- arima(df$X.Total.Confirmed.Cases.,c(1,2,3))
fit1
forecast1 <- forecast(fit1,h=10)
forecast1
plot(forecast1)
fit2 <- auto.arima(df$X.Total.Confirmed.Cases.,seasonal = FALSE)
Console Terminal Jobs
C:/Users/AMARESH/Downloads/
> fit1 <- arima(df$X.Total.Confirmed.Cases.,c(1,2,3))
> fit1
Call:
arima(x = df$X.Total.Confirmed.Cases., order = c(1, 2, 3))
Coefficients:
ar1 ma1 ma2 ma3
-0.8951 -1.0080 -0.9823 0.9914
s.e. 0.0172 0.0078 0.0142 0.0078
sigma^2 estimated as 221100: log likelihood = -9733.36, aic = 19476.72
> forecast1 <- forecast(fit1,h=10)
> forecast1
Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
1287 738.2808 135.2178 1341.344 -184.0245 1660.586
1288 824.1984 218.2243 1430.172 -102.5589 1750.956
1289 752.0524 144.0002 1360.105 -177.8832 1681.988
1290 821.3964 211.4132 1431.380 -111.4924 1754.285
1291 764.0861 152.8436 1375.329 -170.7286 1698.901
1292 820.1499 207.6027 1432.697 -116.6602 1756.960
1293 774.7274 161.4336 1388.021 -163.2246 1712.679
Environment History Connections
Files Plots Packages Help Viewer
Series tsdata1
Partial ACF
Lag
Search the web and Windows


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
bda project.R*
tsdata <- ts(df$X.Total.Confirmed.Cases.)
ts.plot(tsdata)
adf.test(tsdata)
tsdata1 <- diff(tsdata,differences = 2)
adf.test(tsdata1)
acf(tsdata1)
pacf(tsdata1)
fit1 <- arima(df$X.Total.Confirmed.Cases.,c(1,2,3))
fit1
forecast1 <- forecast(fit1,h=10)
forecast1
plot(forecast1)
fit2 <- auto.arima(df$X.Total.Confirmed.Cases.,seasonal = FALSE)
fit2
forecast2 <- forecast(fit2, h = 9)
forecast2
plot(forecast2)
Console Terminal Jobs
Coefficients:
ar1 ma1 ma2 ma3
-0.8951 -1.0080 -0.9823 0.9914
s.e. 0.0172 0.0078 0.0142 0.0078
sigma^2 estimated as 221100: log likelihood = -9733.36, aic = 19476.72
> forecast1 <- forecast(fit1,h=10)
> forecast1
Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
1287 738.2808 135.2178 1341.344 -184.0245 1660.586
1288 824.1984 218.2243 1430.172 -102.5589 1750.956
1289 752.0524 144.0002 1360.105 -177.8832 1681.988
1290 821.3964 211.4132 1431.380 -111.4924 1754.285
1291 764.0861 152.8436 1375.329 -170.7286 1698.901
1292 820.1499 207.6027 1432.697 -116.6602 1756.960
1293 774.7274 161.4336 1388.021 -163.2246 1712.679
1294 820.1500 205.9552 1434.345 -119.1798 1759.480
1295 784.2529 169.6302 1398.876 -155.7314 1724.237
1296 821.1489 205.8877 1436.410 -119.8119 1762.110
> plot(forecast1)
Forecasts from ARIMA(1,2,3)
Search the web and Windows



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
bda project.R*
Packages forecast, hrbrthemes, and tseries required but are not installed. Install Don't Show Again
tsdata <- ts(df$X.Total.Confirmed.Cases.)
ts.plot(tsdata)
adf.test(tsdata)
tsdata1 <- diff(tsdata,differences = 2)
adf.test(tsdata1)
acf(tsdata1)
pacf(tsdata1)
fit1 <- arima(df$X.Total.Confirmed.Cases.,c(1,2,3))
fit1
forecast1 <- forecast(fit1,h=10)
forecast1
plot(forecast1)
fit2 <- auto.arima(df$X.Total.Confirmed.Cases.,seasonal = FALSE)
fit2
forecast2 <- forecast(fit2, h = 9)
forecast2
plot(forecast2)
C:/Users/AMARESH/Downloads/
1289 752.0524 144.0002 1360.105 -177.8832 1681.988
1290 821.3964 211.4132 1431.380 -111.4924 1754.285
1291 764.0861 152.8436 1375.329 -170.7286 1698.901
1292 820.1499 207.6027 1432.697 -116.6602 1756.960
1293 774.7274 161.4336 1388.021 -163.2246 1712.679
1294 820.1500 205.9552 1434.345 -119.1798 1759.480
1295 784.2529 169.6302 1398.876 -155.7314 1724.237
1296 821.1489 205.8877 1436.410 -119.8119 1762.110
> plot(forecast1)
> fit2 <- auto.arima(df$X.Total.Confirmed.Cases.,seasonal = FALSE)
> fit2
Series: df$X.Total.Confirmed.Cases.
ARIMA(5,1,2) with drift
Coefficients:
ar1 ar2 ar3 ar4 ar5 ma1 ma2 drift
0.5460 -0.1847 -0.0541 -0.1356 -0.1291 -1.5877 0.5976 0.5943
s.e. 0.0516 0.0314 0.0326 0.0317 0.0327 0.0457 0.0448 0.1345
sigma^2 estimated as 202120: log likelihood=-9671.2
AIC=19360.4 AICc=19360.54 BIC=19406.82
Environment History Connections
Files Plots Packages Help Viewer
Forecasts from ARIMA(1,2,3)
0 2000 4000 6000
0 200 400 600 800 1000 1200
Search the web and Windows



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
bda project.R*
tsdata <- ts(df$X.Total.Confirmed.Cases.)
ts.plot(tsdata)
adf.test(tsdata)
tsdata1 <- diff(tsdata,differences = 2)
adf.test(tsdata1)
acf(tsdata1)
pacf(tsdata1)
fit1 <- arima(df$X.Total.Confirmed.Cases.,c(1,2,3))
fit1
forecast1 <- forecast(fit1,h=10)
forecast1
plot(forecast1)
fit2 <- auto.arima(df$X.Total.Confirmed.Cases.,seasonal = FALSE)
fit2
forecast2 <- forecast(fit2, h = 9)
forecast2
plot(forecast2)
C:/Users/AMARESH/Downloads/
Coefficients:
ar1 ar2 ar3 ar4 ar5 ma1 ma2 drift
0.5460 -0.1847 -0.0541 -0.1356 -0.1291 -1.5877 0.5976 0.5943
s.e. 0.0516 0.0314 0.0326 0.0317 0.0327 0.0457 0.0448 0.1345
sigma^2 estimated as 202120: log likelihood=-9671.2
AIC=19360.4 AICc=19360.54 BIC=19406.82
> forecast2 <- forecast(fit2, h = 9)
> forecast2
Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
1287 602.0955 25.93928 1178.252 -279.0594 1483.250
1288 853.9724 277.31551 1430.629 -27.9482 1735.893
1289 817.8354 230.05667 1405.614 -81.0946 1716.765
1290 694.1211 100.49490 1287.747 -213.7518 1601.994
1291 748.9957 147.74681 1350.245 -170.5351 1668.526
1292 758.8082 149.94301 1367.673 -172.3707 1689.987
1293 733.6654 124.77043 1342.560 -197.5591 1664.890
1294 737.1653 126.11326 1348.217 -197.3581 1671.689
1295 752.2931 137.92792 1366.658 -187.2974 1691.884
> plot(forecast2)
Forecasts from ARIMA(5,1,2) with drift
0 2000 4000 6000
0 200 400 600 800 1000 1200
Search the web and Windows

5. CONCLUSION

In this project, we calculated moving average as the dataset gets updated with the new cases enrolled.

The slowing down of the trend during this period suggested that COVID-19 would not cause any serious problems, particularly in INDIA. Unfortunately, that was not the case.

The last two sets of forecasts that cover the period 02/03/2020 to 23/04/2020 show a significant increase in the trend of cases globally coupled with an increase in the associated uncertainty.

6. REFERENCES

- https://www.machinehack.com/course/covid-19-forecasting-the-corona-outbreak/
- https://www.livescience.com/topics/live/coronavirus-live-updates