**PROJECT REPORT**

**ON**

***“Covid-19 Time Series Forecasting”***

Submitted in partial fulfilment of the

Requirement for the award of the degree of

**Bachelor of Computer Application**

** **

**Submitted To Submitted By**

Dr. Neha Goel Vineet Gupta

Assistant Professor 14517702018

BCA 5-C

(Batch: 2018-2021)

**Vivekananda Institute of Professional Studies**

**(Affiliated to Guru Gobind Singh Indraprastha University)**

**CERTIFICATE**

This is to certify that Vineet Gupta of BCA 5th Semester from Vivekananda Institute of Professional Studies, Delhi has presented this PROJECT REPORT entitled “Covid-19 Time Series Forecasting” in partial fulfilment of the requirements for the award of the degree of Bachelor of Computer Applications under our supervision and guidance.

Dr. Neha Goel

Assistant Professor

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to my college that gave me the golden opportunity to do this wonderful project on the topic “Covid-19 Forecasting” which also helped me in doing a lot of research and I came to know about so many new things. I am really thankful to them.

Secondly, I would also like to thank Datacamp that offered all the necessary courses which helped me to learn so many new things.

The project has offered me a great opportunity to grow and develop. It has propelled me to be able to overcome challenges and develop my career. I learnt extensively about data manipulation, visualisation, tidy data and time series. The programme has enhanced my analytic thinking and skills as well as improved my professional skills and ability to work in a multicultural environment. Working on this was not only an honour and privilege but a lifelong experience that will forever shape my professional life.

Vineet Gupta

**Table of Contents**

* Introduction
* Platform/Software
* Training & Courses
* Project Documentation
* Conclusion

**Introduction**

**Objective:**

The objective of this project is to analyse the time-series of SARS-CoV-2/Covid-19 confirmed and death cases and study the trends to forecast expected future trends in India.

**About Covid-19:**

What is corona virus Corona viruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19.

COVID-19 is the infectious disease caused by the most recently discovered corona virus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019.

**Platform/Software**

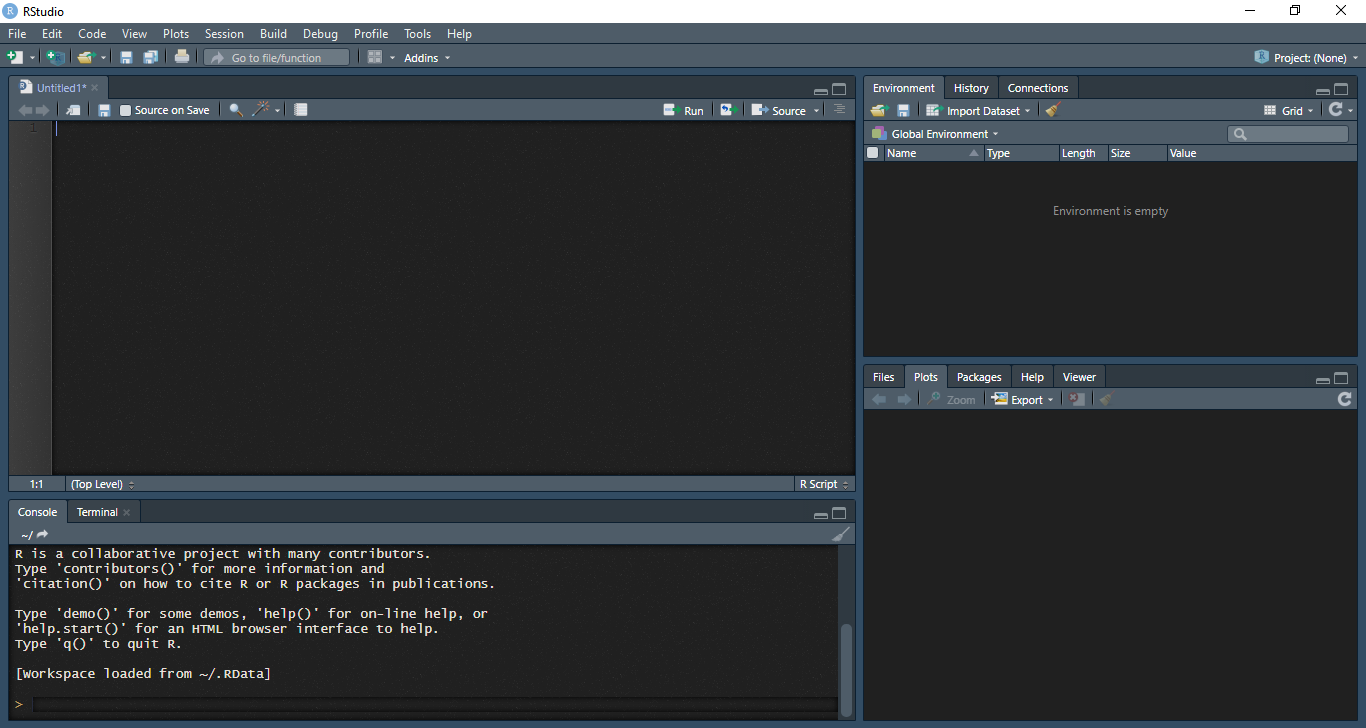
**R Programming**

R is a language and environment for statistical computing and graphics.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible.

**RStudio**

RStudio is dedicated to sustainable investment in free and open-source software for data science, to help people understand and improve the world through data.



**Trainings and Courses**

1. **Data Manipulation (using dplyr)**

Data manipulation involves modifying data to make it easier to read and to be more organized. We manipulate data for analysis and visualization. It is also used with the term ‘data exploration’ which involves organizing data using available sets of variables.  
At times, the data collection process done by machines involves a lot of errors and inaccuracies in reading. Data manipulation is also used to remove these inaccuracies and make data more accurate and precise.

1. **Data visualization (using ggplot2)**

Data visualization using R includes data plotting with R’s default graphics system, base graphics. After an introduction to base graphics, we look at a number of R plotting examples, from simple graphs such as scatterplots to plotting correlation matrices. This includes using R plot colours effectively and creating and saving complex plots in R.

R supports four different graphics systems:

base graphics, grid graphics, lattice graphics, and ggplot2.

1. **Time Series Analysis (using astsa)**

Time Series is a sequence of data in chronological order. Data is commonly recorded sequentially, over time.

Basic time series models:

1. White Noise (WN)
2. Random Walk (RW)
3. **Autoregression (AR)**
4. Moving Average (MA)

Basic assumptions:

1. Consecutive observations are equally spaced.
2. Apply a discrete-time observation index.
3. This may only hold approximately.

**Why ts() objects?**

1. Improved plotting.
2. Access to time index information.
3. Model estimation and forecasting.

**Project Documentation**

**RStudio Technical Documentation Page**

**Code:**

#covid-19 time series analysis for INDIA

#libraries

library(openxlsx) #xsls file

library(dplyr)

library(ggplot2)

library(tibble) #view()

#data sources:

#www.data.world/shad/covid-19-time-series-data

#source files: www.github.com/vineetgupta086/reimagined-panda

setwd("E:/Vineet/work/projects/project\_covid/datasets")

#processing raw data

oct\_data=read.xlsx("conf\_new\_oct.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)

nov\_data=read.xlsx("conf\_new\_nov.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)

oct\_data[,1] %>%

mutate(as.Data(oct\_data[,1]))

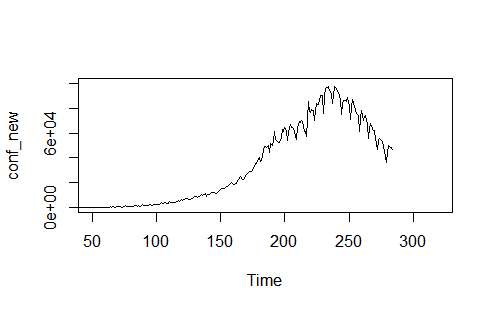
class(oct\_data[,1]) #date

#For now we do not use data\_nov as it will be later used to compare to the model

conf\_new<-as.ts(oct\_data$Confirmed)

view(conf\_new,title = "Confirmed")

plot.ts(conf\_new,xlim=c(50,320),ylim=c(0,100000))



##The data begins from 22-Jan-2020 and the x axis is the number of days after the date. The values start to show ## trends after about 60 days so the graph is cropped.

summary(conf\_new)



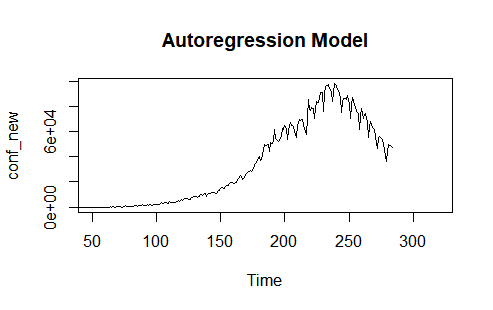
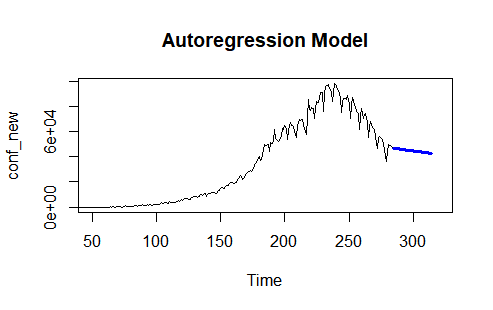
#Autoregressive model for confirmed\_new

ar\_conf<-arima(conf\_new,order=c(1,0,0)) #autoregression model

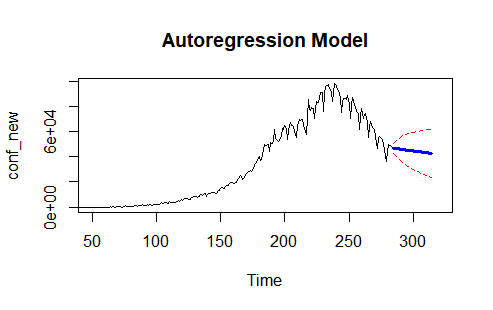
ar\_pred<-predict(ar\_conf,n.ahead=30)$pred

ar\_se<-predict(ar\_conf,n.ahead=30)$se

plot.ts(conf\_new,xlim=c(50,320),main="Autoregression Model")

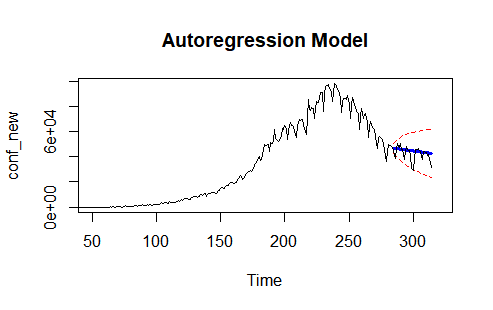
points(ar\_pred,type='l',col="blue",lwd=3)

points(ar\_pred-ar\_se,type='l',col="red",lty=2); points(ar\_pred+ar\_se,type='l',col="red",lty=2)



#comparing predicted data to the actual number of confirmed new cases during november

points(nov\_data,type='l')



##dead\_new

rm(list=ls(pattern='^ar|new$|data$'))

oct\_data2=read.xlsx("dead\_new\_oct.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)

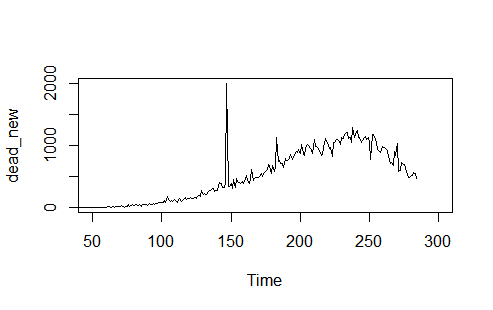
nov\_data2=read.xlsx("dead\_new\_nov.xlsx",sheet = 1,startRow = 1,colNames = T,rowNames = F,detectDates = T)

oct\_data2[,1]<-as.Date(oct\_data2[,1])

#Autoregression Model for deaths

dead\_new<-as.ts(oct\_data2$Deaths)

plot(dead\_new,ylim=c(0,2000),xlim=c(50,300))

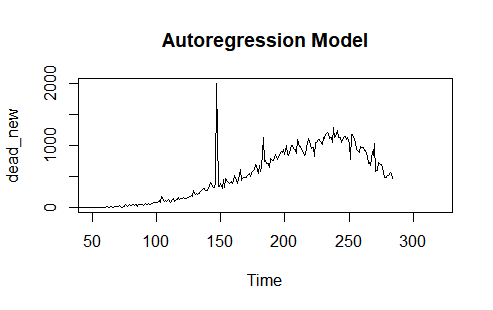


ar\_dead<-arima(dead\_new,order=c(1,0,0))

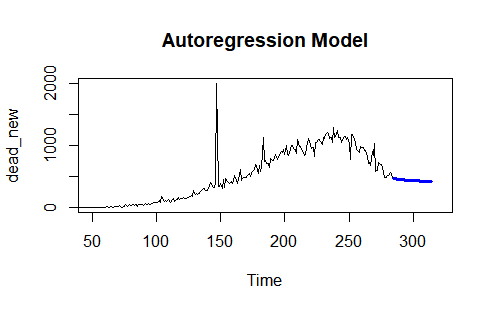
ar\_pred<-predict(ar\_dead,n.ahead=30)$pred

ar\_se<-predict(ar\_dead,n.ahead=30)$se

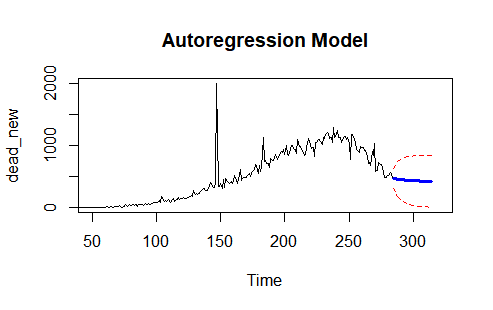
plot.ts(dead\_new,xlim=c(50,320),main="Autoregression Model")



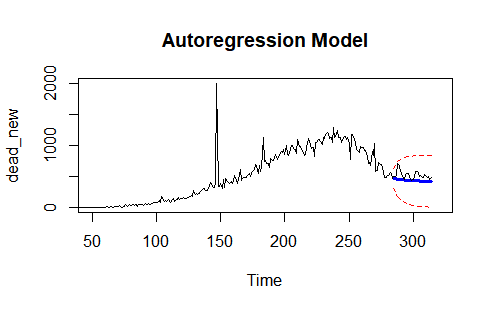
points(ar\_pred,type='l',col="blue",lwd=3)



points(ar\_pred-ar\_se,type='l',col="red",lty=2); points(ar\_pred+ar\_se,type='l',col="red",lty=2)



points(nov\_data2,type='l')



rm(list=ls(pattern='^ar|new$|data2$'))

#concluded

**CONCLUSION**

1. **Project Conclusion**

* A case study on the spread of the novel Coronavirus (Covid-19), in which we observe the trends of daily confirmed cases and deaths through time-series.
* Daily cases were modelled through Autoregression modelling for fu rther forecasting.
* The predicted numbers are displayed along with their confidence interval at 95% accuracy.
* The predicted numbers are also compared to the actual observed numbers in the month November to see how well our model fits.

1. **Limitations**

* The model suggested does not take into the account the external factors that are responsible for the trends. This suggests that the time-series forecasting can be used only for the immediate future and not for long term forecasting.
* The data had to be updated manually frequently in every cycle. This would result in resource wastage.

1. **Future scope of the project**

* A live data feeding can be added using data scrapping, and the predicted visuals can be displayed on a website.
* The timeline of the external events can be added at their respective times to observe how these external factors, if they do, change the trend.