Covid-19 Tracking and Analysis during Lockdown using R Studio

A PROJECT REPORT

ABSTRACT

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, deep learning and big data. RStudio is an integrated development environment (IDE) for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser.

The RStudio IDE is partly written in the C++ programming language and uses the Qt framework for its graphical user interface. The bigger percentage of the code is written in Java. JavaScript is also amongst the languages used.

The scope of this project is to investigate and visualize the dataset of COVID-19 cases using Data science techniques. When identifying infected, the values are more plotted.
This also can predict the outcome of a future pandemic curve. This curve of case rise and fall can be taken as a study to provide more data for future use.

Thorough this project we want to showcase the importance and power of R and why it is so popular and recommended by data science professionals. This project was made to give a better visualization of the covid-19 pandemic, how it spread across India…? and what was the effect of the lockdown in slowing the spread…?

This gives us an opportunity to explore this huge amount of data. Analyse it and get meaningful results from it. This brings an opportunity to our door steps. Through deep studying of the data and information generated by the governments across the world we can get a better and visual understanding of this pandemic.

ACKNOWLEDGEMENTS

We express our sincere thanks to the Head of the Department, Department of Computer Science and Engineering, **Dr. B. Amutha**, for all the help and infrastructure provided to us to complete this project successfully and her valuable guidance. We also owe our profound gratitude to our project guide **Dr. R. S. Ponmagal**, who took keen interest in our project work and guided us all along, till the completion of our project work by providing all the necessary information for completing this task. We are thankful and fortunate enough to get constant encouragement, support and guidance from all the Teaching staff of the Department of Computer Science and Engineering which helped us in successfully completing our major project work. Also, we would like to extend our sincere regards to all the non-teaching staff of the department of Computer Science and Engineering for their timely assistance.

Arul Saxena, Karan Jhajj

CONTENTS

INTRODUCTION

1.1 What is Data Science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, deep learning and big data.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science. Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge

1.2 What is R studio?

RStudio is an integrated development environment (IDE) for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser.

The RStudio IDE is partly written in the C++ programming language and uses the Qt framework for its graphical user interface. The bigger percentage of the code is written in Java. JavaScript is also amongst the languages used.

Work on the RStudio IDE started around December 2010, and the first public beta version (v0.92) was officially announced in February 2011. Version 1.0 was released on 1 November 2016. Version 1.1 was released on 9 October 2017.

In April 2018, RStudio PBC (at the time RStudio, Inc.) announced that it will provide operational and infrastructure support to Ursa Labs in support of the Labs focus on building a new data science runtime powered by Apache Arrow.

In April 2019, RStudio PBC (at the time RStudio, Inc.) released a new product, the RStudio Job Launcher. The Job Launcher is an adjunct to RStudio Server. The launcher provides the ability to start processes within various batch processing systems (e.g. Slurm) and container orchestration platforms (e.g. Kubernetes). This function is only available in RStudio Server Pro (fee-based application).

1.3 COVID-19

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It was first identified in December 2019 in Wuhan, China, and has since spread globally, resulting in an ongoing pandemic. As of 19 May 2020, more than 4.8 million cases have been reported across 188 countries and territories, resulting in more than 318,000 deaths. More than 1.78 million people have recovered.

Common symptoms include fever, cough, fatigue, shortness of breath, and loss of smell and taste. While the majority of cases result in mild symptoms, some progress to acute respiratory distress syndrome (ARDS) likely precipitated by cytokine storm, multi-organ failure, septic shock, and blood clots. The time from exposure to onset of symptoms is typically around five days but may range from two to fourteen days.

Proposed Work

2.1 An Efficient Approach of Tracking Covid-19

Tech companies, governments, and international agencies have all announced measures to help contain the spread of the COVID-19, otherwise known as the Coronavirus.

Some of these measures impose severe restrictions on people's freedoms, including to their privacy and other human rights. Unprecedented levels of surveillance, data exploitation, and misinformation are being tested across the world.

Many of those measures are based on extraordinary powers, only to be used temporarily in emergencies. Others use exemptions in data protection laws to share data.

Some may be effective and based on advice from epidemiologists, others will not be. But all of them must be temporary, necessary, and proportionate.

It is essential to keep track of them. When the pandemic is over, such extraordinary measures must be put to an end and held to account.

2.2 Representing Data

It involves producing images that communicate relationships among the represented data to viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization. This mapping establishes how data values will be represented visually, determining how and to what extent a property of a graphic mark, such as size or colour, will change to reflect changes in the value of a datum.

2.3 R-studio Modelling

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, data mining surveys, and studies of scholarly literature databases show substantial increases in popularity; as of February 2020, R ranks 13th in the TIOBE index, a measure of popularity of programming languages.
The most specialized integrated development environment (IDE) for R is RStudio.[53] A similar development interface is R Tools for Visual Studio. Some generic IDEs like Eclipse, [54] also offer features to work with R.

Graphical user interfaces with more of a point-and-click approach include Rattle GUI, R Commander, and RKWard.
Some of the more common editors with varying levels of support for R include Emacs (Emacs Speaks Statistics), Vim (Nvim-R plugin), Neovim (Nvim-R plugin), Kate, LyX, Notepad++, Visual Studio Code, WinEdt, and Tinn-R.
R functionality is accessible from several scripting languages such as Python, Perl, Ruby, F#, and Julia. Interfaces to other, high-level programming languages, like Java and .NET C# are available as well.

PROPOSED MODEL

3.1 Scope

The scope of this project is to investigate and visualize the dataset of COVID-19 cases using Data science techniques. When identifying infected, the values are more plotted. This also can predict the outcome of a future pandemic curve. This curve of case rise and fall can be taken as a study to provide more data for future use.

3.2 Description of the work

The aim is of this project is to track covid-19 pandemic in our country. As we know the government of India have taken a great number of steps to handle this unprecedented crisis. But this kind of crisis can’t be handled just by lockdown measures and social distancing. What is necessary is proper tracking of cases and real time update of the same. Testing is of no use if proper data base is not maintained for keep track of the progress.

This gives us an opportunity to explore this huge amount of data. Analyse it and get meaningful results from it. This brings an opportunity to our door steps. Through deep studying of the data and information generated by the governments across the world we can get a better and visual understanding of this pandemic.

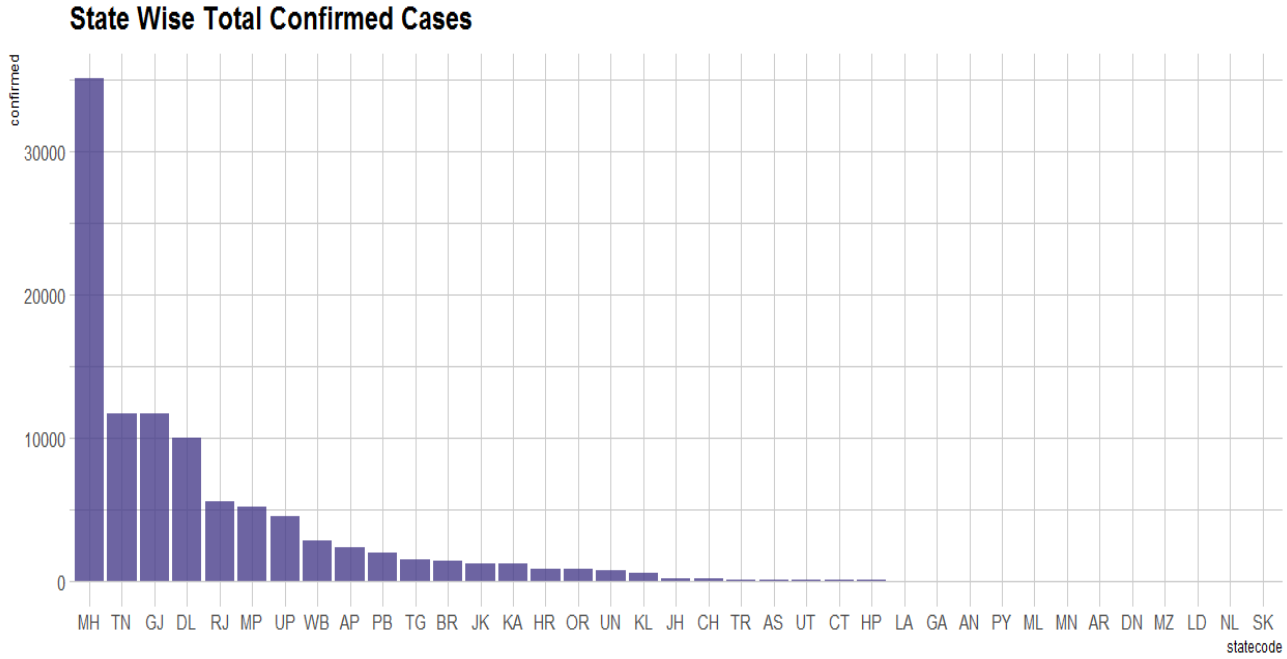
3.3 Project Goals

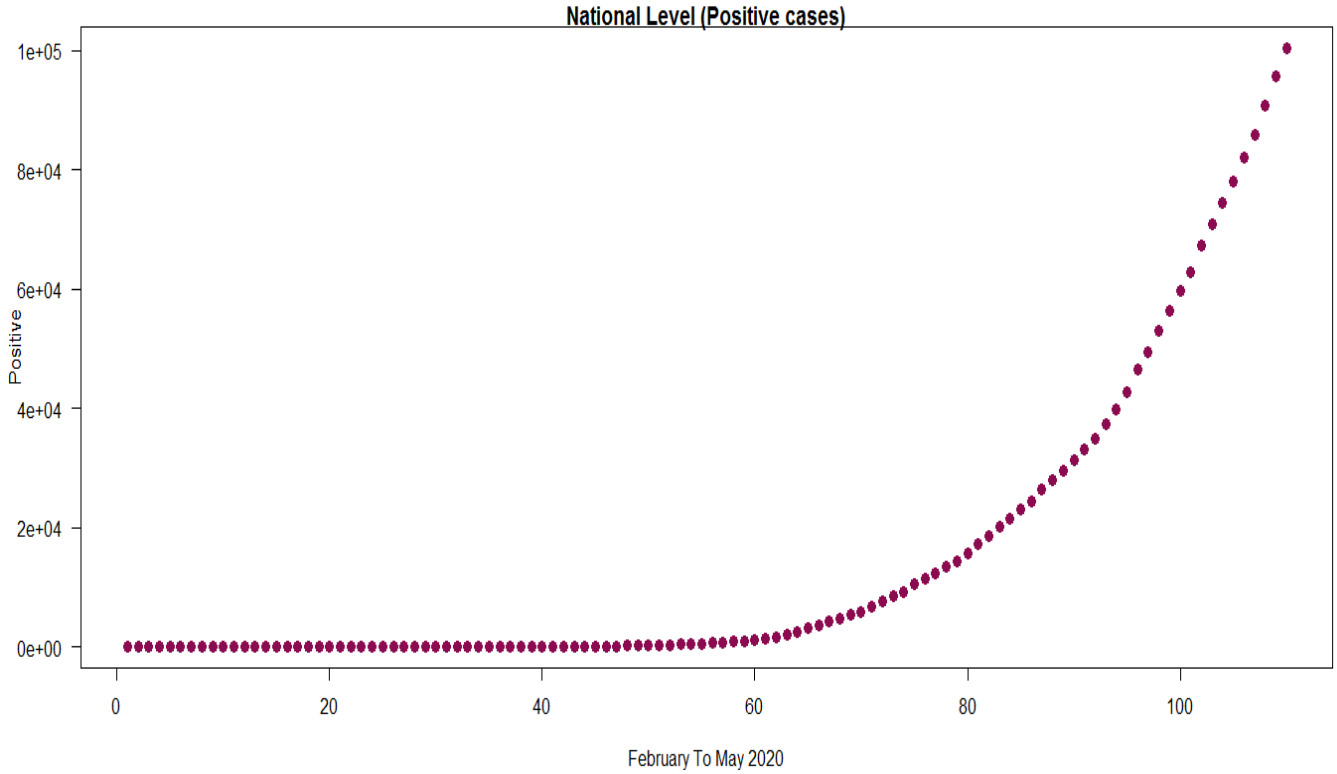
Thorough this project we want to showcase the importance and power of R and why it is so popular and recommended by data science professionals. This project was made to give a better visualization of the covid-19 pandemic, how it spread across India…? and what was the effect of the lockdown in slowing the spread…?

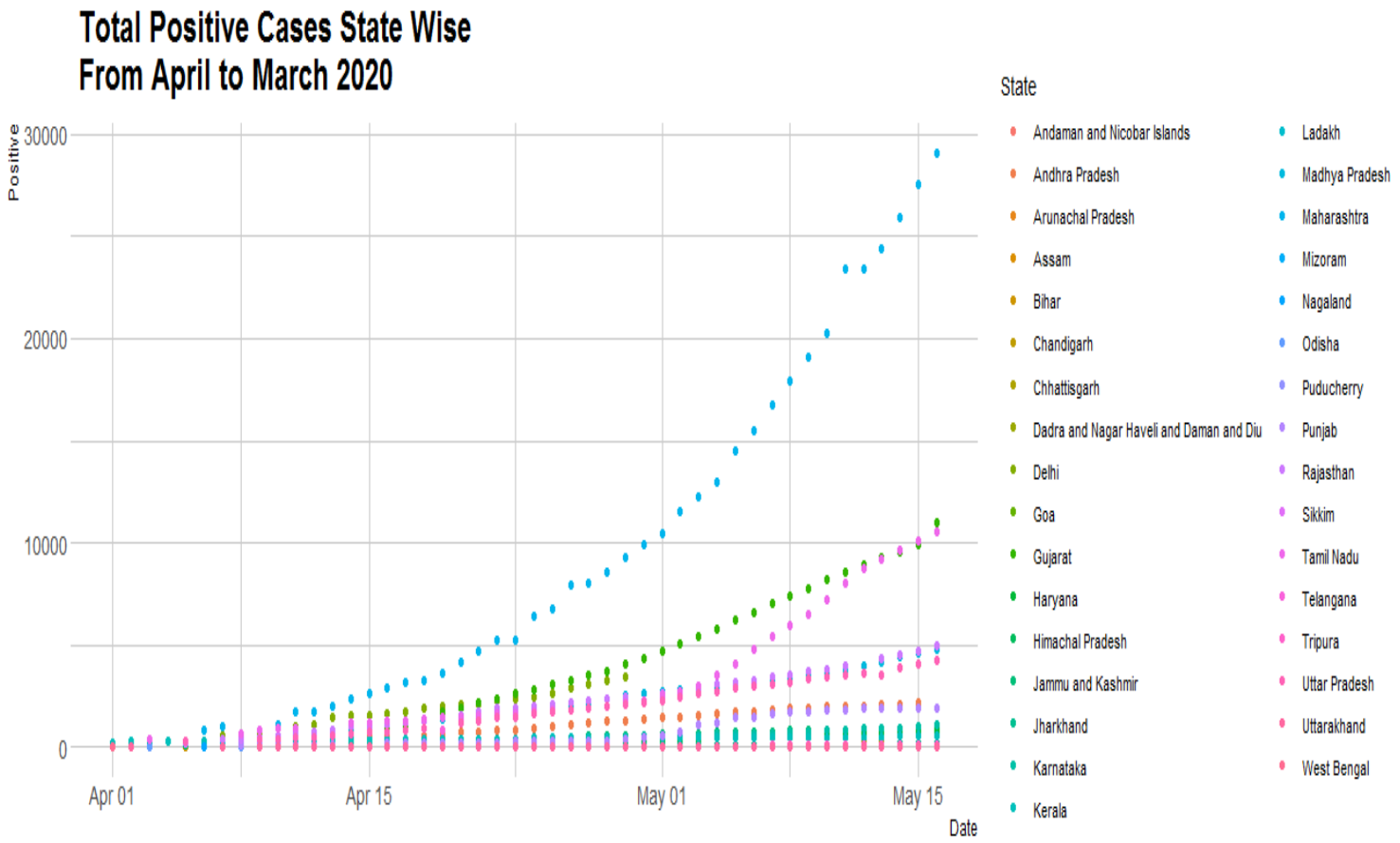
All this was done by using datasets provided by the government and some datasets from Kaggle which were easy to use. The following are the goals achieved by the project:

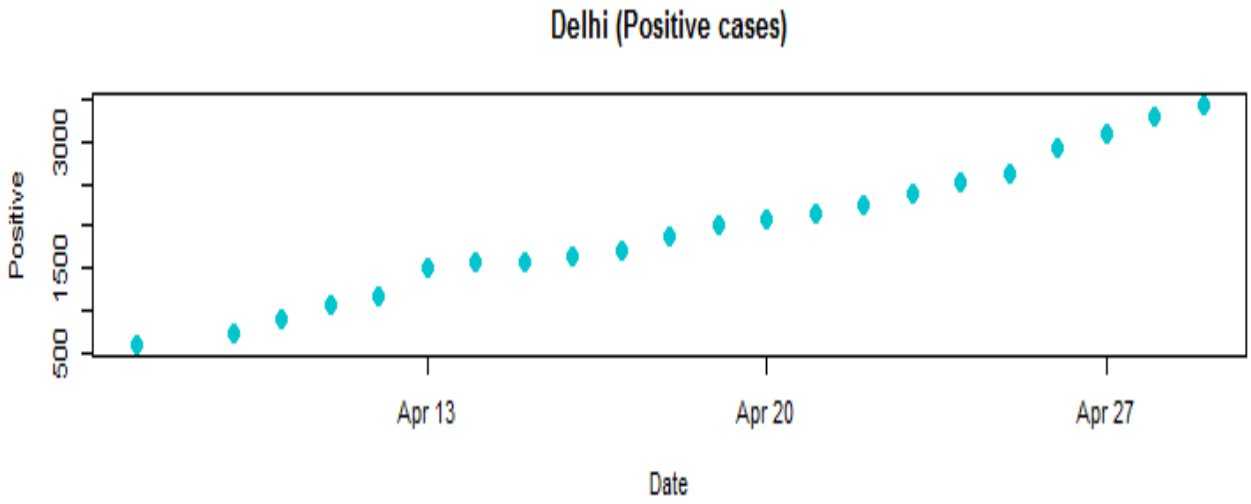
- We have used state wise data to represent the curves and the rate of spread in the main cities.
- We have data from ICMR to track the labs and testing centres across the states.
- We have also used patient wise data to track the age group that is most affected by this virus.
- We have also represented the ICMR testing and results together to get a better understanding of the spreading rate.
- We have used the ploty Library to plot some interactive graphs.
- Which can be used on websites as widgets.
- Using the ICMR data we have also visualized the number of government and private centres and how they are distributed across the states.

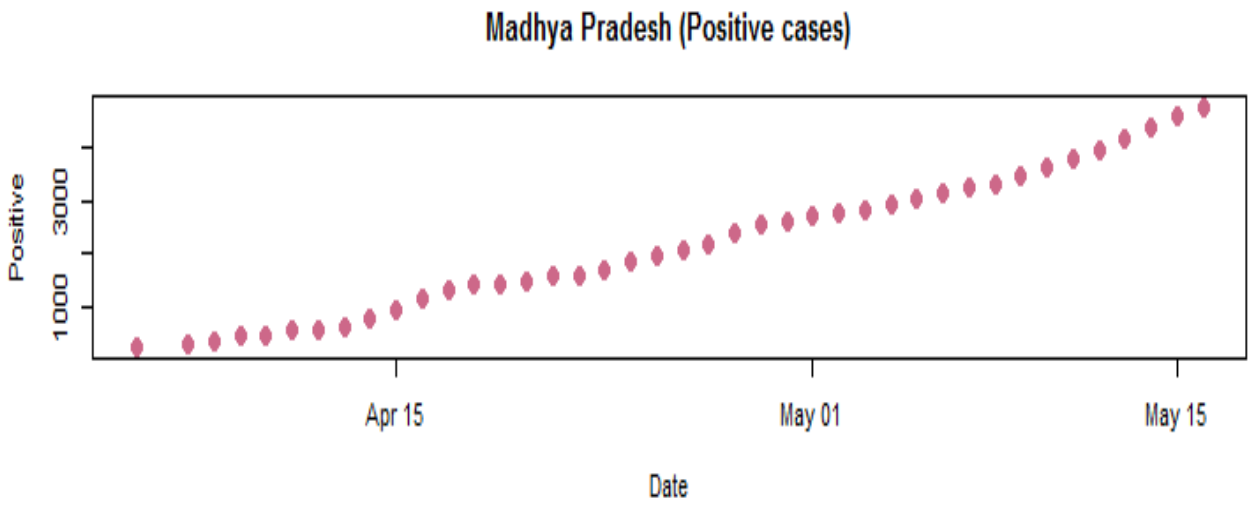
Snapshots

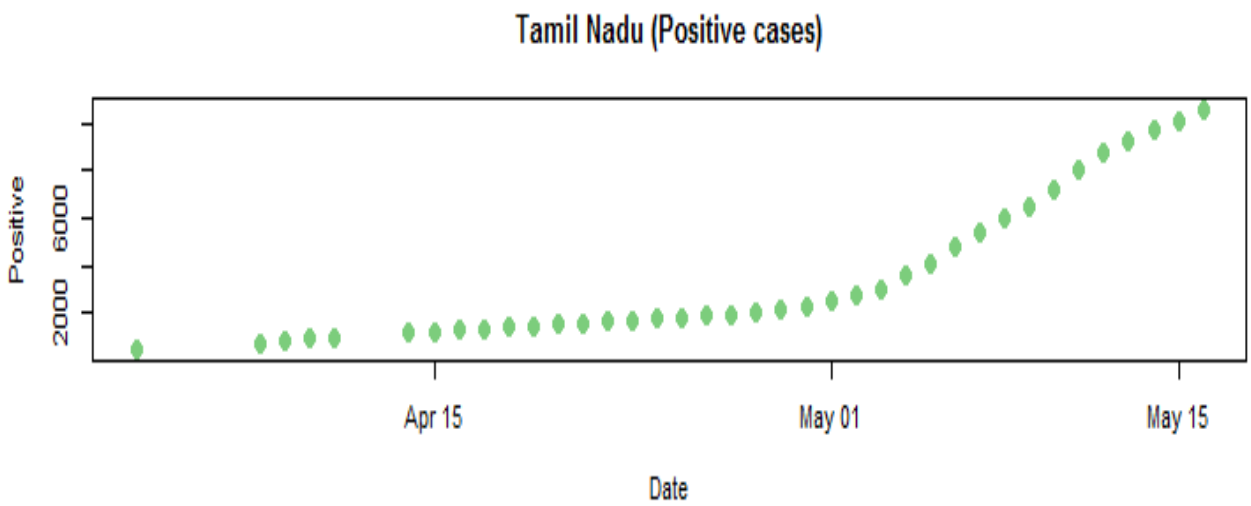

State Wise Total Confirmed Cases
confirmed
30000
20000
10000
0
MH TN GJ DL RJ MP UP WB AP PB TG BR JK KA HR OR UN KL JH CH TR AS UT CT HP LA GA AN PY ML MN AR DN MZ LD NL SK
statecode

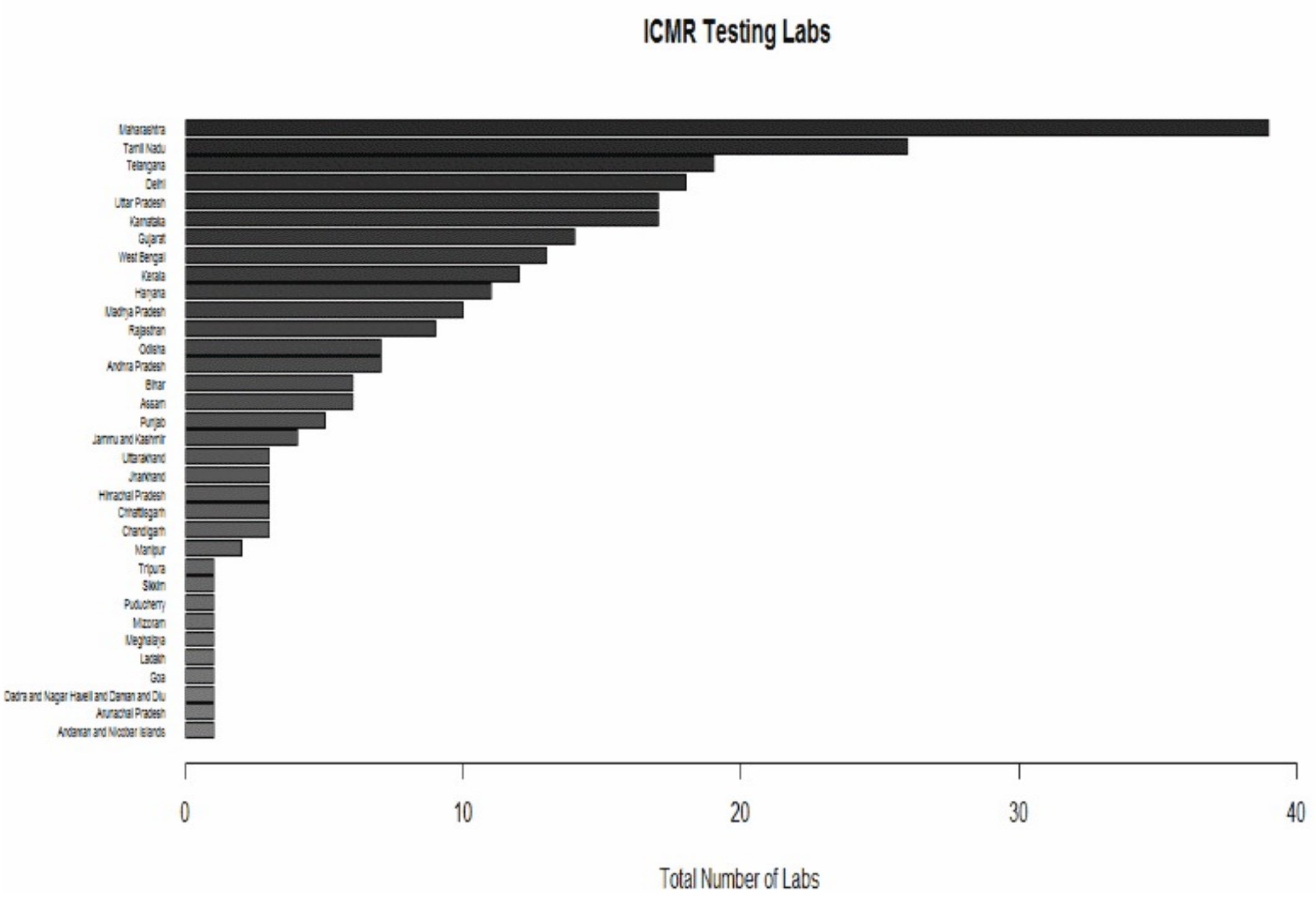

National Level (Positive cases)
1e+05
8e+04
6e+04
4e+04
2e+04
0e+00
Positive
0
20
40
60
80
100
February To May 2020

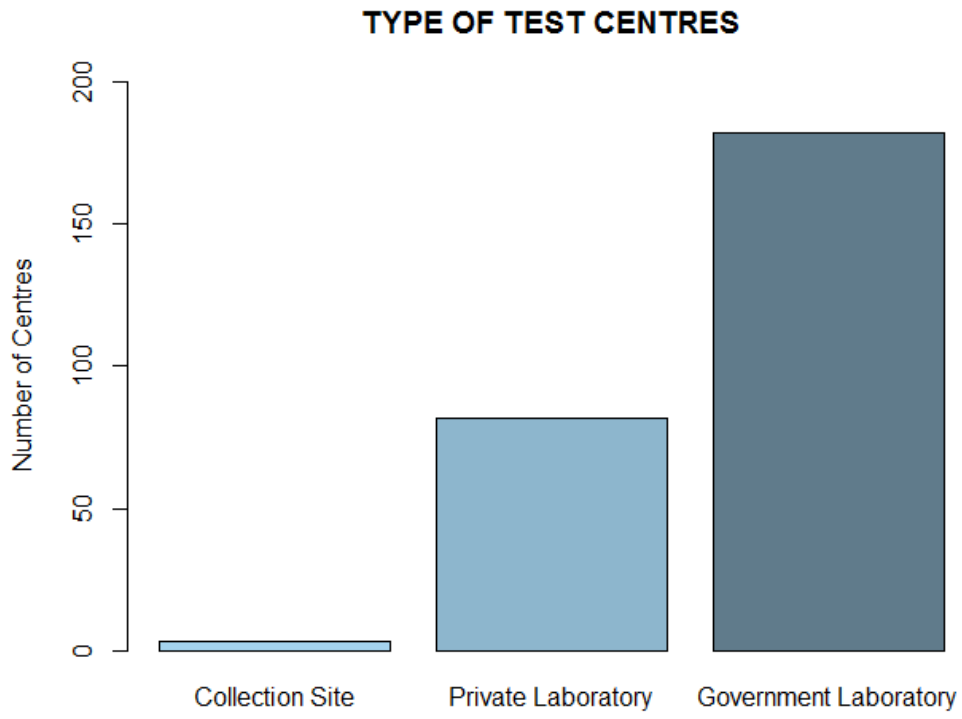

Total Positive Cases State Wise
From April to March 2020
Positive
30000
20000
10000
0
Apr 01
Apr 15
May 01
May 15
Date
State
Andaman and Nicobar Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra and Nagar Haveli and Daman and Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Madhya Pradesh
Maharashtra
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal

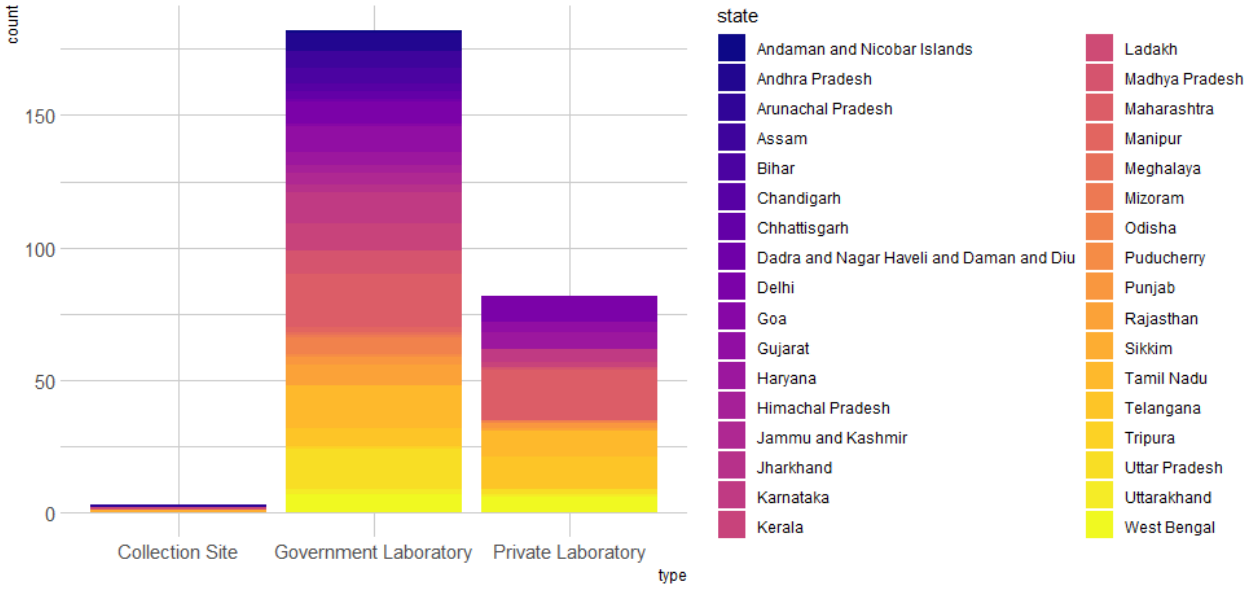

Delhi (Positive cases)
Positive
500
1500
3000
Apr 13
Apr 20
Apr 27
Date

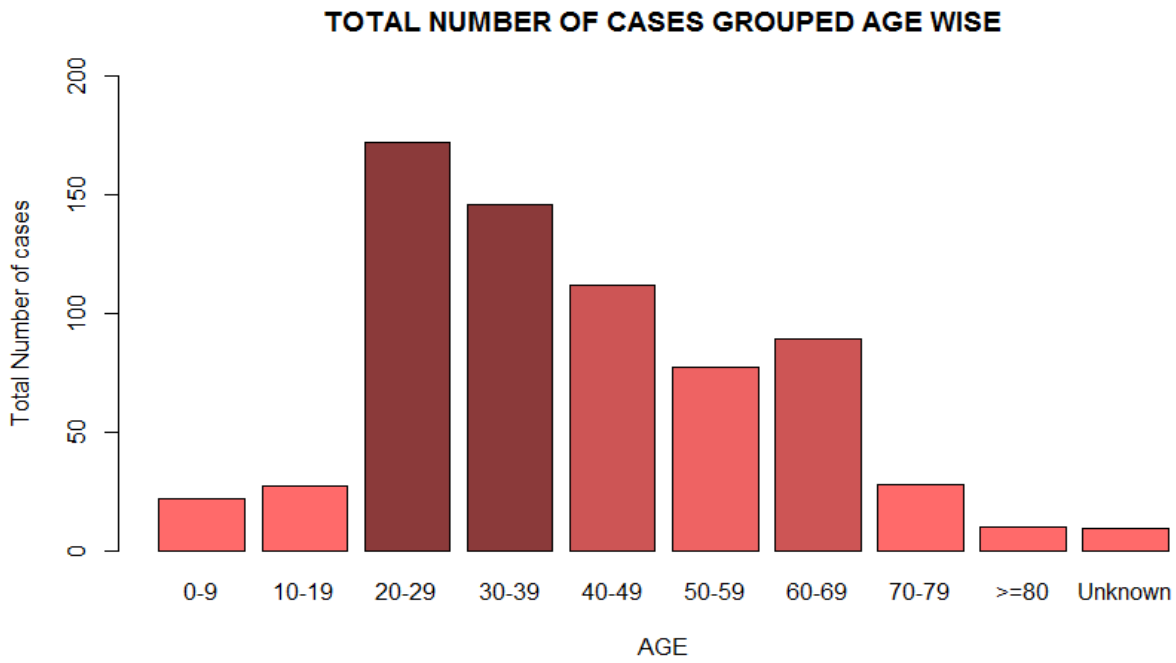

Madhya Pradesh (Positive cases)
Positive
1000
3000
Apr 15
May 01
May 15
Date


Tamil Nadu (Positive cases)
Positive
2000
6000
Apr 15
May 01
May 15
Date

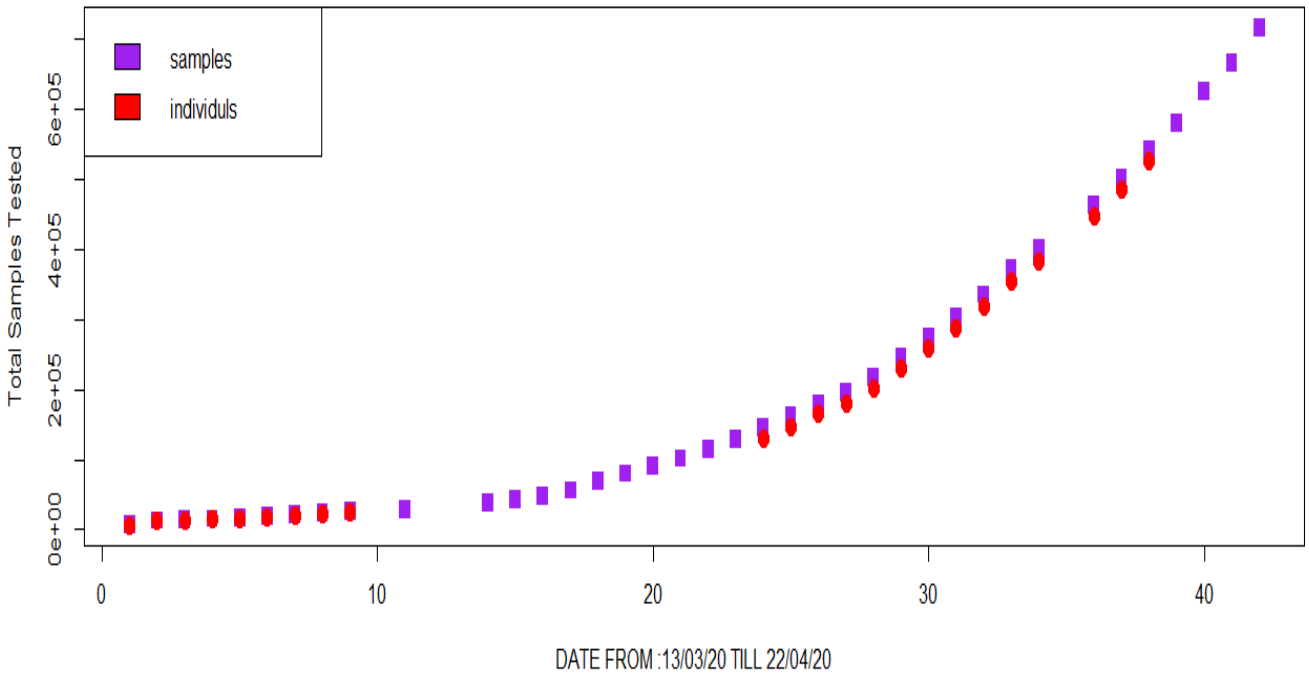

ICMR Testing Labs
Maharashtra
Tamil Nadu
Telangana
Delhi
Uttar Pradesh
Karnataka
Gujarat
West Bengal
Kerala
Haryana
Madhya Pradesh
Rajasthan
Odisha
Andhra Pradesh
Bihar
Assam
Punjab
Jammu and Kashmir
Uttarakhand
Jharkhand
Himachal Pradesh
Chhattisgarh
Chandigarh
Manipur
Tripura
Sikkim
Puducherry
Mizoram
Meghalaya
Ladakh
Goa
Dadra and Nagar Haveli and Daman and Diu
Arunachal Pradesh
Andaman and Nicobar Islands
0
10
20
30
40
Total Number of Labs


TYPE OF TEST CENTRES
Number of Centres
0
50
100
150
200
Collection Site
Private Laboratory
Government Laboratory

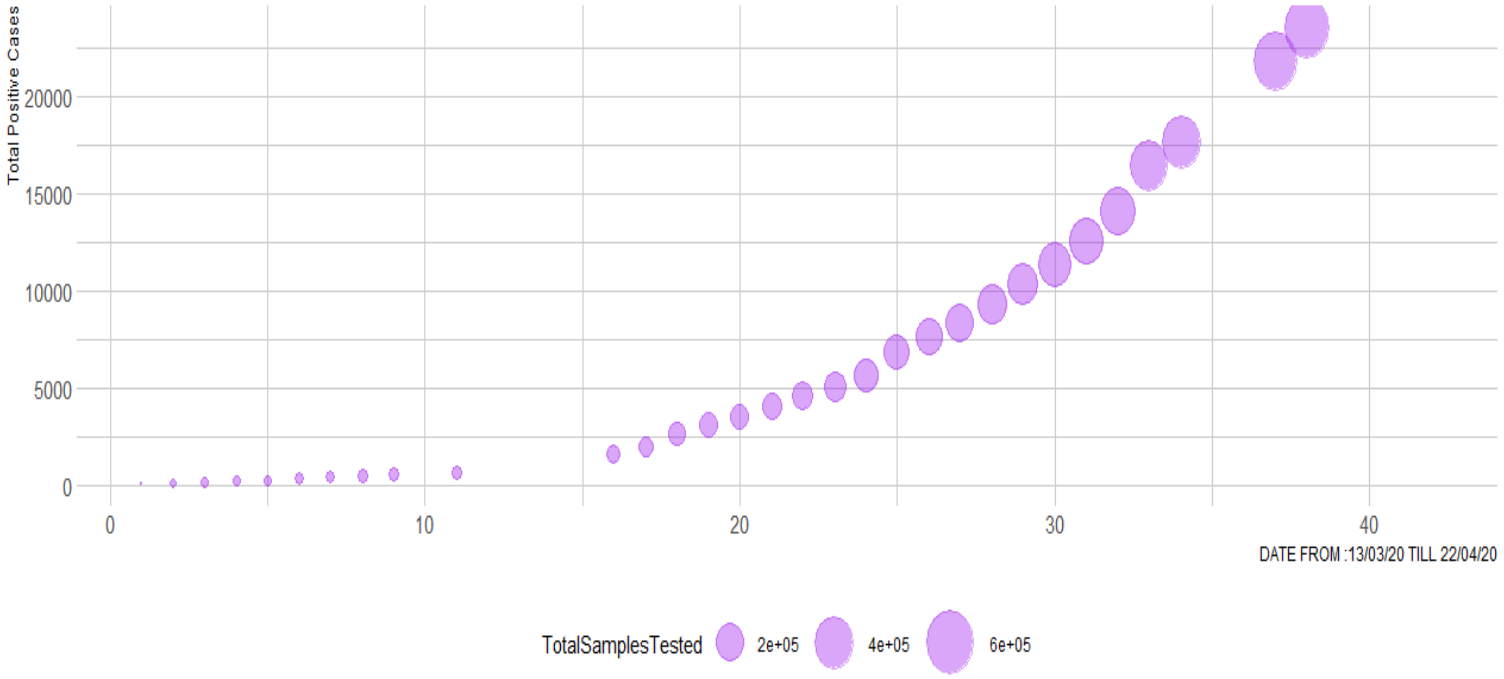

count
0
50
100
150
Collection Site
Government Laboratory
Private Laboratory
type
state
Andaman and Nicobar Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra and Nagar Haveli and Daman and Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal



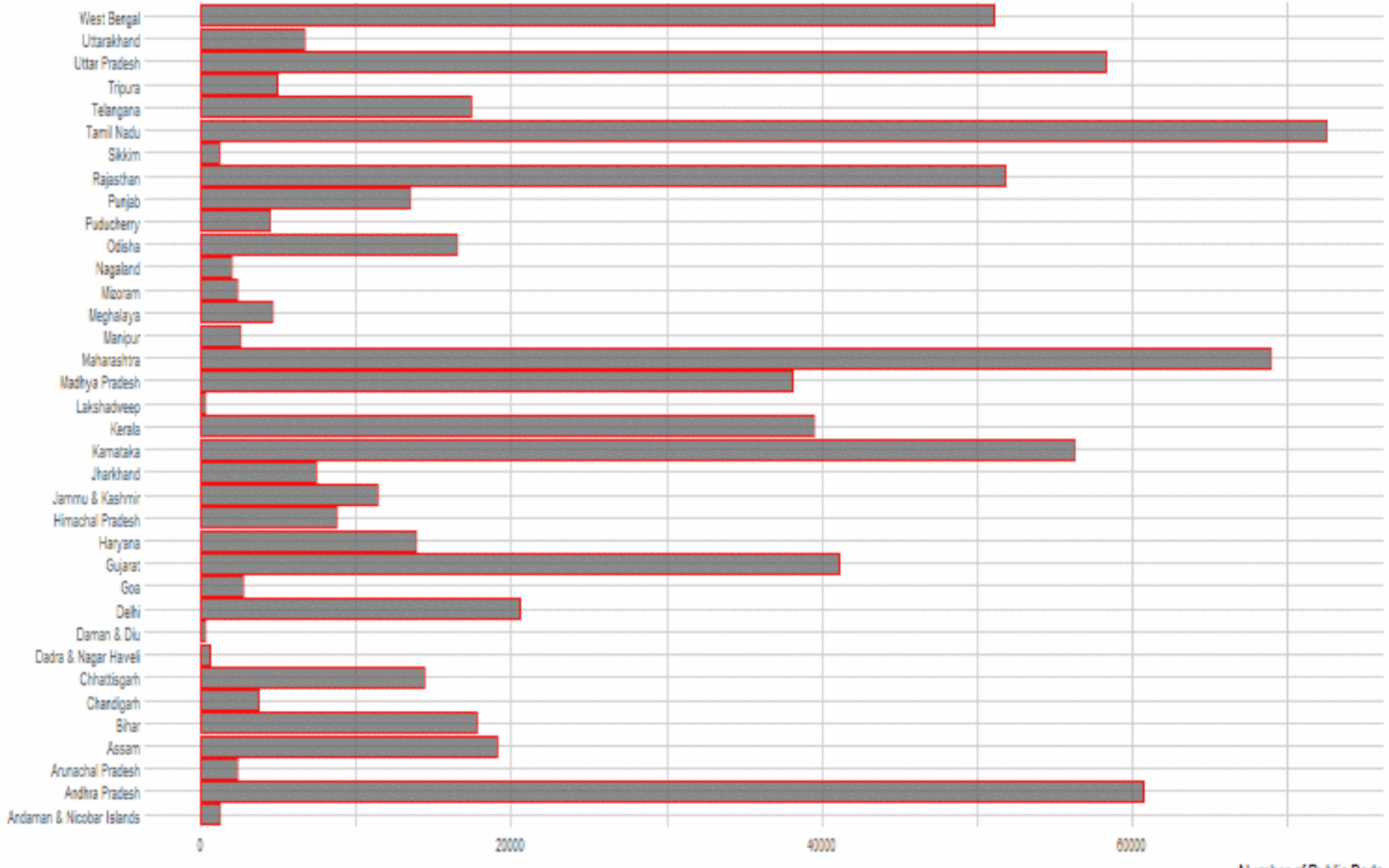
ICMR Samples and Individuals Tested



No. of Total Positive Cases From March to April 2020



Public Beds State Wise



CONCLUSION

Through this project we wanted to showcase the power of R studio and the importance of good data representation.

We could conclude that the growth of the virus was slowed down by the lockdown making it a somewhat success. But also, the testing in the initial phase was less and not up to the mark. That also created the illusion of the growth rate to be linear rather than exponential. Only after India bumped up the testing, we got to see the real picture.
We could also determine the hotspot states in India and the testing centres too. Also, the age group that was most effected was found out. The individual state curves representing the growth rate of the spread was also observed successfully.

In conclusion we can say that we have successfully slowed the spread but the cases are still rising and self-imposed measures are a must.

References

- Web References:
  - https://www.r-graph-gallery.com/index.html
  - https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=StatewiseTestingDetails.csv
  - https://www.kaggle.com/sudalairajkumar/covid19-in-india
  - https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset
  - https://www.youtube.com/watch?v=hr2X7rmkprM
  - https://www.youtube.com/watch?v=N5gYo43oLE8
  - https://www.datamentor.io/r-programming/
  - https://www.youtube.com/watch?v=_V8eKsto3Ug&list=PLW8UG5jpmjwH0JBGIhalKIXvY_PNbmmcQ&index=2&t=0s

Appendix (source code)

```
#Importing Libraries----------
library(ggplot2)
library(forcats)
library(dplyr)
library(hrbrthemes)
library(viridis)
library(plotly)

# table one ------------------------------------------------------------

df<-read.csv("project(data science)\\covid_19_india.csv", header = TRUE)
head(df)

tail(df)
plot(df)
summary(df)
plot(df$ï..Sno,df$Deaths)



# Need a table with frequencies for each category
states_india <- table(df$State.UnionTerritory)  # Create table
barplot(states_india)             # Bar chart
plot(states_india)               # Default X-Y plot (lines)



barplot(states_india,
      main = "COVID-19 State wise Distribution in India",
      xlab = "Cases",
      las=1,
      col = "darkred",
      horiz = TRUE)
summary(df$State.UnionTerritory)
summary(df$ConfirmedIndianNational)




plot(df$Date,df$Deaths)
```

```
hist(df$Cured[df$Date==19/04/20])

dev.off()
cat("\014")




#Table Two----------------------------------------------------------------------

df2<-read.csv("AgeGroupDetails.csv", header = TRUE)

#Head-----

head(df2)

#Tail-----

tail(df2)

#Summary---

summary(df2)


#Total number of cases grouped age wise ----------------------------------------------------------

barplot(df2$TotalCases,
      names.arg = c("0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69","70-79",">=80","Unknown"),

col=c("indianred1","indianred1","indianred4","indianred4","indianred3","indianred2","indianred3","indianred1"
,"indianred1","indianred1"),
      ylim = c(0,200),
      xlab="AGE",
      ylab = "Total Number of cases",
      main = "TOTAL NUMBER OF CASES GROUPED AGE WISE")
```

```
# Table
Three=============================================================================-

df3<-read.csv("HospitalBedsIndia.csv", header = TRUE)
# Head--------------------------------------------------------------------------------------------
head(df3)
#ploting number of public beds according to states
barplot(df3$NumPublicBeds_HMIS,
        ylim = c(500,300000))

#Removing the last row containing all india stats
df3<- df3[-c(37),]
tail(df3)



df3_1%>%
mutate(name = fct_reorder(State.UT,desc(NumPublicBeds_HMIS)))%>%
ggplot(aes(x=State.UT,y=NumPublicBeds_HMIS))+
              geom_bar(stat = "identity",alpha=0.7,colour="red")+
              coord_flip()+
              xlab(" ")+
              theme_ipsum(axis_text_size = 7)+
              ylab("Number of Public Beds")+
              ggtitle("Public Beds State Wise")




#=============================================================================
==

df5<-read.csv("ICMRTestingDetails.csv", header = TRUE)

head(df5)
tail(df5)
plot(df5$ï..SNo,df5$TotalPositiveCases,
     xlab="DATE FROM :13/03/20 TILL 22/04/20",
     ylab="Total positive cases",

     col="purple",
     pch=19,
     cex=1.5,
     main="Total Positive cases")
```

```
p<-df5 %>%

      mutate(text = paste("Date/Time: ", DateTime, "\nTotalSamplesTested: ", TotalSamplesTested, sep="")) 
%>%
      ggplot(aes(x=ï..SNo,y=TotalPositiveCases,size=TotalSamplesTested,text=text))+
      geom_point(alpha=0.4,color="purple")+
      scale_size(range=c(.5,15))+
      scale_fill_viridis(discrete=TRUE, guide=FALSE) +
      theme_ipsum() +
      theme(legend.position="bottom") +
      ylab("Total Positive Cases") +
      xlab("DATE FROM :13/03/20 TILL 22/04/20")+
      ggtitle("No. of Total Positive Cases \n From March to April 2020")
p
pp<-ggplotly(p,tooltip="text")
pp




plot(df5$TotalSamplesTested,
    xlab="DATE FROM :13/03/20 TILL 22/04/20",
    ylab="Total Samples Tested",
    col="purple",
    pch=15,
    cex=1.5,
    type = "p",)
    points(df5$TotalIndividualsTested,col="red",pch=19,cex=1.5)
    legend("topleft",
        c("samples","individuls"),
        fill = c("purple","red"))




#----------Table 6----------------------------------------
```

```
df6<-read.csv("ICMRTestingLabs.csv", header = TRUE)
head(df6)

?barplot
states_testcenters<-table(df6$state)

par(mar=c(4,9,4,4))
barplot(states_testcenters[order(states_testcenters,decreasing = FALSE)],
        horiz = TRUE,
        las=1,
        cex.names = 0.5,
        cex.axis = 1.0,
        main = "ICMR Testing Labs ",
        xlim = c(0,40),
        xlab = "Total Number of Labs",
        col=c("gray49","gray48","gray47","gray46","gray45","gray44","gray43","gray42","gray41",
              "gray40","gray39","gray38","gray37","gray36","gray35","gray34","gray33","gray32",
              "gray31","gray30","gray29","gray28","gray27","gray26","gray25","gray24","gray23",
              "gray22","gray21","gray20","gray19","gray18","gray17","gray16")



        )


# Center Type-------------------------------------
centre_type<-table(df6$type)
barplot(centre_type[order(centre_type,decreasing = FALSE)],
        main = "TYPE OF TEST CENTRES",
        col = c("lightskyblue2","lightskyblue3","lightskyblue4"),
        ylim = c(0,200),
        ylab = " Number of Centres ",
        )

#ggplot version of the above graph-------------------------
ggplot(df6,aes(x=type,fill=state))+
        geom_bar()+
        theme_ipsum()+
        scale_fill_viridis_d(option="plasma")



dev.off()
cat("\014")
```

```
#======================Table 7=======================================
df7<-read.csv("StatewiseTestingDetails.csv", header = TRUE)

#---------------------removing null values------------


df7<-df7[!(df7$State=="Meghalaya" | df7$Positive==552),]
#------head-----------2020-02-16       Meghalaya          552        299        7
head(df7)
#--------Tail---------

tail(df7)

#----------changing date column to type date----------
str(df7$Date)
df7$Date<-as.Date(df7$Date)

par(mfrow = c(3, 1))

plot(df7$Date[df7$State=="Delhi"],df7$Positive[df7$State == "Delhi"],
    pch = 19,
    cex = 1.5,
    ylab = "Positive",
    xlab = "Date",
    main = "Delhi (Positive cases)",
    col="turquoise3")

plot(df7$Date[df7$State=="Madhya Pradesh"],df7$Positive[df7$State == "Madhya Pradesh"],
    pch = 19,
    cex = 1.5,
    ylab = "Positive",
    xlab = "Date",
    main = "Madhya Pradesh (Positive cases)",
    col="palevioletred3")


plot(df7$Date[df7$State=="Tamil Nadu"],df7$Positive[df7$State == "Tamil Nadu"],
    pch = 19,
    cex = 1.5,
    ylab = "Positive",
    xlab = "Date",
    main = "Tamil Nadu (Positive cases)",
    col="palegreen3")
```

```
# Restore graphic parameter
par(mfrow=c(1, 1))



#---------------All States----------------------------------


ggplot(data=df7,aes(y=Positive,x=Date,col=State))+geom_point()+
       theme_ipsum_pub()+
       scale_fill_viridis_d(option="plasma")+
       ggtitle(" Total Positive Cases State Wise\n From April to March 2020")




#-----------------------------------------------------

df8<-read.csv("new_data\\nation_level_daily.csv", header = TRUE)

#-------head-------

head(df8)
#--------------Tail
tail(df8)

#--------------

str(df8)
library(lubridate)
ymd(df8$date)
as.Date(df8$date)
parse_date_time2(df8$date,orders="mdy")

par(mar=c(4,4,1,1))
plot(df8$totalconfirmed,
    ylim = c(500,100000),
    las=1,
    pch = 16,
    cex = 1.3,
    ylab = "Positive",
    xlab = "February To May 2020",
```

```
    main = "National Level (Positive cases)",
    col="deeppink4",

    )

#=====================================================

df9<-read.csv("new_data\\state_level_latest.csv", header = TRUE)

head(df9)
df9<- df9[-c(1),]

df9%>%
      mutate(statecode=fct_reorder(statecode,desc(confirmed)))%>%
      ggplot(aes(x=statecode,y=confirmed))+
      geom_bar(stat = "identity",alpha=0.8,fill="darkslateblue")+
      theme_ipsum()+
      ggtitle("State Wise Total Confirmed Cases")
```