COVI-19 ANALYSIS AND GLOBAL FORCAST

A PROJECT REPORT

Submitted by

MAULISHREE AWASTHI – 19BCE1864
REKHA – 19BCE1871

Course Title:

FOUNDATION OF DATA ANALYTICS

SLOT: F2

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

INDEX

ABSTRACT:

The aim of the project is to provide data analysis of covid-19 (a pandemic started in December 2019). Through plotting of data, various cases have been studied like most affected countries due to this pandemic. Study of data from various countries is combined to show the growth of cases and recovery graph. In this project, the predictions on various cases have been done and finally, the accuracy of the algorithm has been determined. The project mainly focusses on data analysis of country India. And the project also focusses on the analysis of vaccines administered by India and various visualizations techniques.

OBJECTIVE AND SCOPE:

The pandemic has already taken grip over peoples' life. Since the start of the pandemic, some countries are facing problem of ever-increasing cases. Through the data analysis of cases one can analyses how countries all over the world are doing in terms of controlling the pandemic. In this project we will be diving deep into 'What does data say about Covid-19 situation in World?'. And with available data we come up with some observations and conclusions.

This analysis mainly focuses on:

- Which countries got affected by COVID-19 the most?
- Which countries observed the most hike in COVID-cases in recent months?
- Which countries recovered from COVID-19 till date?
- Which country suffered the most deaths due to COVID-19?

Analyzing data leads to adapt the prevention model of the countries that are doing great in terms of lowering the graph. Predictions are made with the dataset available to the individual/country, thus helping them to decide how far they are able to control the pandemic or up to how much extent they should guide preventive measures.

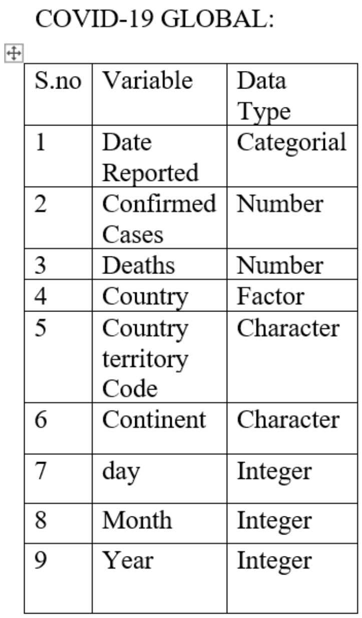
DATA SOURES:

The following sources of data are used:

- COVID19 Global Forecasting from Kaggle
- https://opendata.ecdc.europa.eu/covid19/casedistribution/csv
- COVID_19_INDIA from Kaggle
- covid_vaccine_statewise from Kaggle

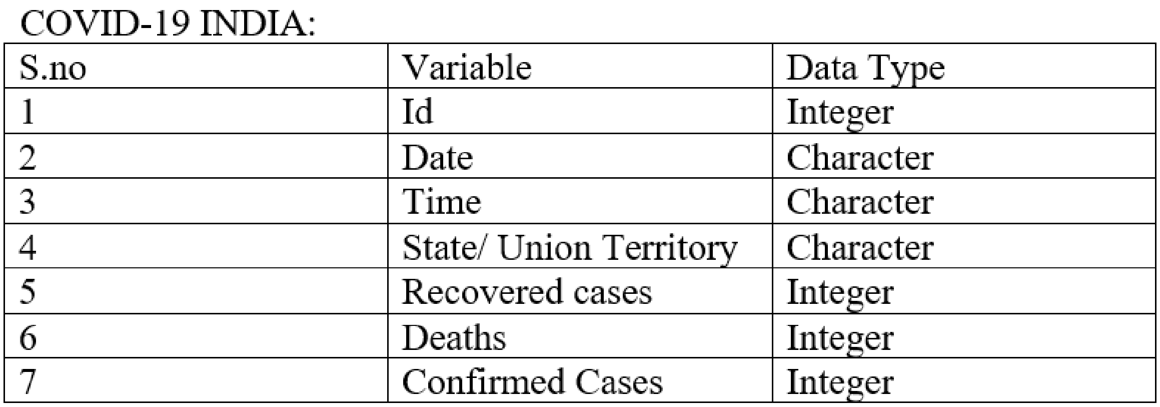
We have extracted data for a year across 12 variables. The variables are date-reported, day, month, year, cases, deaths, countries And Territories, geo-Id, country-territory-Code etc. The Global forecasting dataset consists of variables like data reported, number of covid cases, number of deaths, countries and territories. The COVID_19_INDIA dataset consists of variables like date, time, State and union territory, number of cured people, number of deaths and number of confirmed cases.

Blow is the list of variables and their types:



COVID-19 GLOBAL:

S.no	Variable	Data Type
1	Date Reported	Categorial
2	Confirmed Cases	Number
3	Deaths	Number
4	Country	Factor
5	Country territory Code	Character
6	Continent	Character
7	day	Integer
8	Month	Integer
9	Year	Integer



COVID-19 INDIA:

S.no	Variable	Data Type
1	Id	Integer
2	Date	Character
3	Time	Character
4	State/ Union Territory	Character
5	Recovered cases	Integer
6	Deaths	Integer
7	Confirmed Cases	Integer

TOOLS AND TECHNIQUES:

We have used the following data Analytics technique / methodology for analyzing the Data:

- Summary of Statistics for each variable
- Structure of the dataset
- Using Graphs and density Plots to visually represent them

Tools used: RStudio, Kaggle and Excel.

Techniques: Time Series Plot, Density plot, Bar Chart, Line Chart, Correlation, Heat map. We have used R Programming environment and Microsoft Excel for our analysis and Tableau for data.

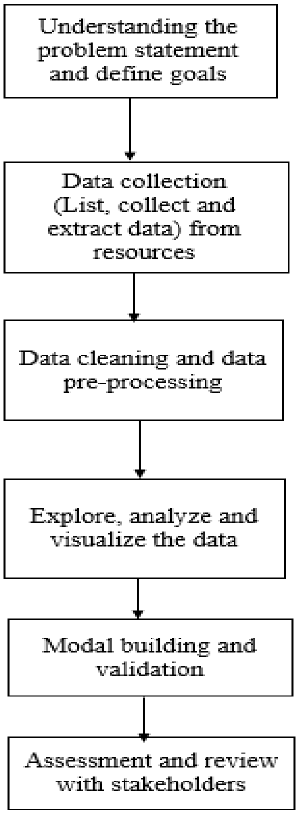
Analytics Approach:

The Analytical Approach will involve the following activities:

- Data extraction from Primary Data source as well as secondary data sources
- Data quality check
- Data cleaning and data preparation
- Study each of the variables by exploring the data
- Division of data into train and test
- Model Development

- Final Model
- Model Validation

The below figure shows the flow of the project:

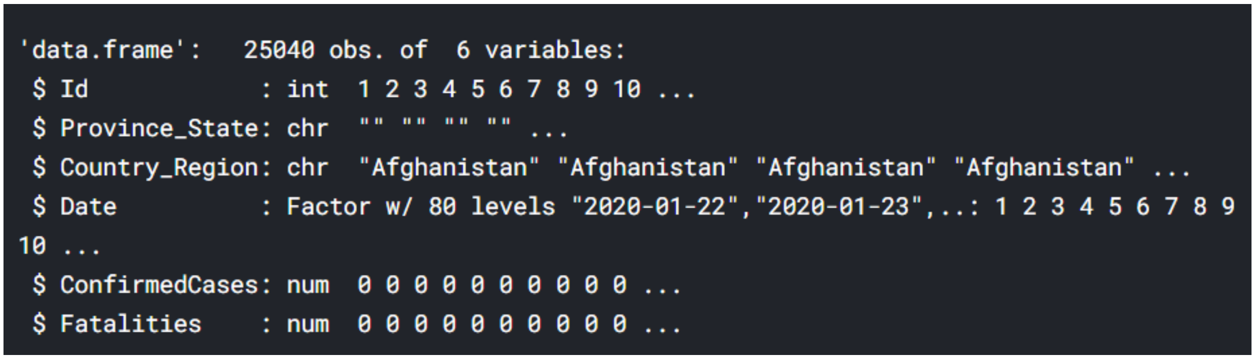

Understanding the problem statement and define goals
Data collection (List, collect and extract data) from resources
Data cleaning and data pre-processing
Explore, analyze and visualize the data
Modal building and validation
Assessment and review with stakeholders

Data Description and Preparation:

For the COVID-19 GLOBAL dataset:

First the conversion of factor data types to character data types.

```
covid_data_train[["Province_State"]] <- as.character(covid_data_train[["Province_State"]] )

covid_data_train[["Country_Region"]] <- as.character(covid_data_train[["Country_Region"]] )

str(covid_data_train)
```

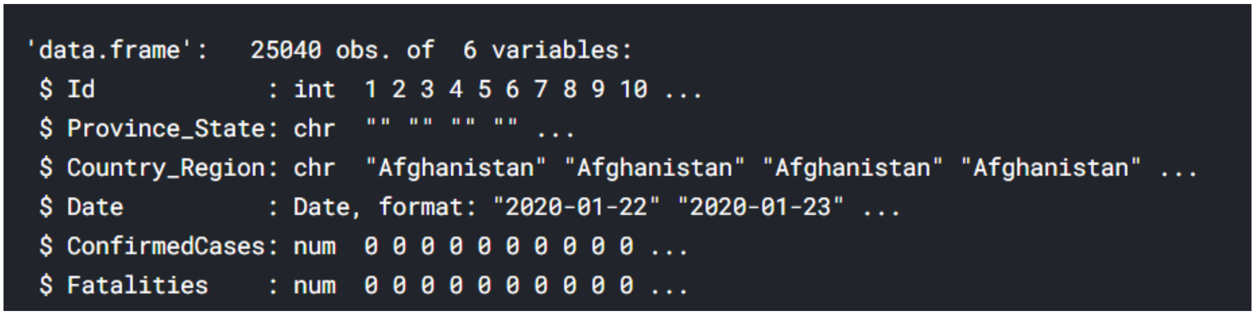


```
'data.frame':   25040 obs. of  6 variables:
 $ Id            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Province_State: chr  "" "" "" "" ...
 $ Country_Region: chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ Date          : Factor w/ 80 levels "2020-01-22","2020-01-23",..: 1 2 3 4 5 6 7 8 9
10 ...
 $ ConfirmedCases: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Fatalities    : num  0 0 0 0 0 0 0 0 0 0 ...
```

Convert the date from categorial to Date format:

covid_data_train[["Date"]] <- as.Date(covid_data_train[["Date"]], format = "%Y-%m-%d")

str(covid_data_train)



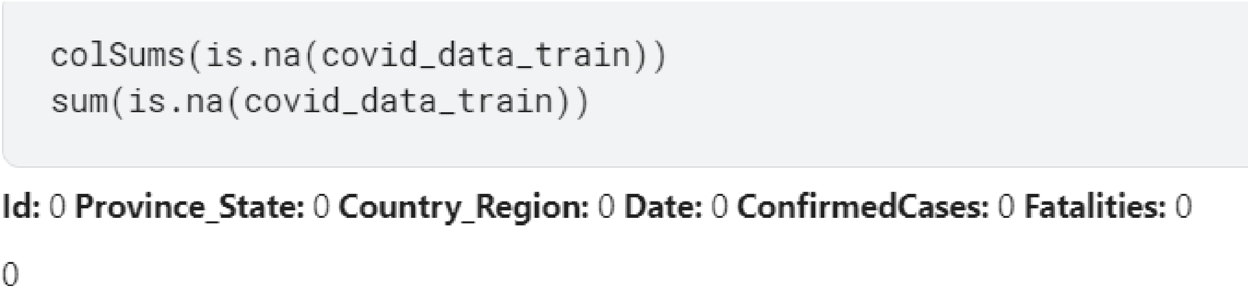
```
'data.frame':   25040 obs. of  6 variables:
 $ Id            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Province_State: chr  "" "" "" "" ...
 $ Country_Region: chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ Date          : Date, format: "2020-01-22" "2020-01-23" ...
 $ ConfirmedCases: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Fatalities    : num  0 0 0 0 0 0 0 0 0 0 ...
```

Variables with missing values:

colSums(is.na(covid_data_train))

sum(is.na(covid_data_train))

There are zero missing values.



```
colSums(is.na(covid_data_train))
sum(is.na(covid_data_train))
```

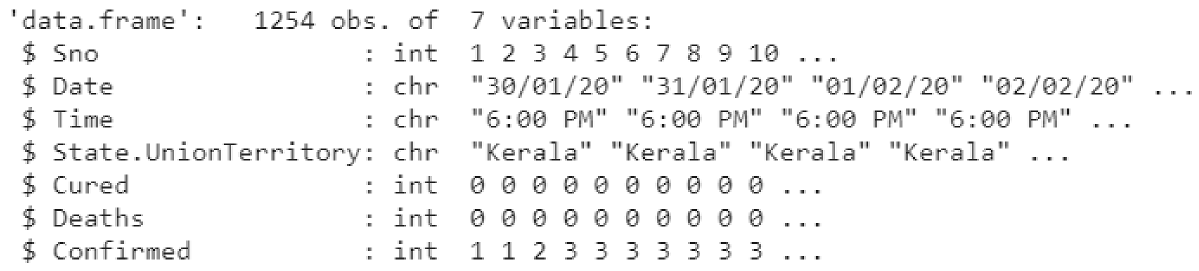
Id: 0 **Province_State:** 0 **Country_Region:** 0 **Date:** 0 **ConfirmedCases:** 0 **Fatalities:** 0

0

For the COVID-19-INDIA dataset:

First the un-wanted columns are removed.

covid <- subset(covid,select=c(ConfirmedIndianNational,ConfirmedForeignNational))



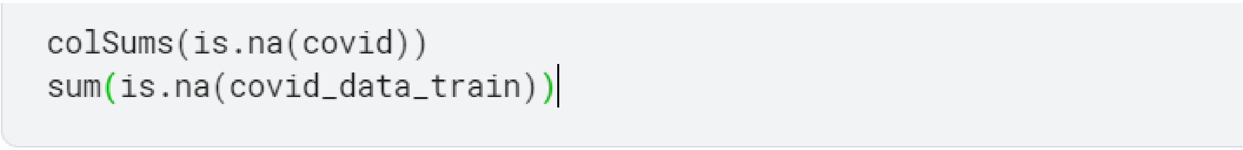
```
'data.frame':	1254 obs. of  7 variables:
 $ Sno                    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Date                   : chr  "30/01/20" "31/01/20" "01/02/20" "02/02/20" ...
 $ Time                   : chr  "6:00 PM" "6:00 PM" "6:00 PM" "6:00 PM" ...
 $ State.UnionTerritory: chr  "Kerala" "Kerala" "Kerala" "Kerala" ...
 $ Cured                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Deaths                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Confirmed              : int  1 1 2 3 3 3 3 3 3 3 ...
```

Searching for missing values:

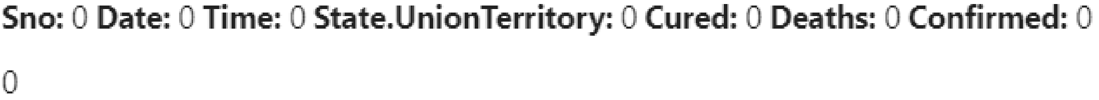
colSums(is.na(covid))

sum(is.na(covid_data_train))

- No missing values in the dataset



```
colSums(is.na(covid))
sum(is.na(covid_data_train))
```

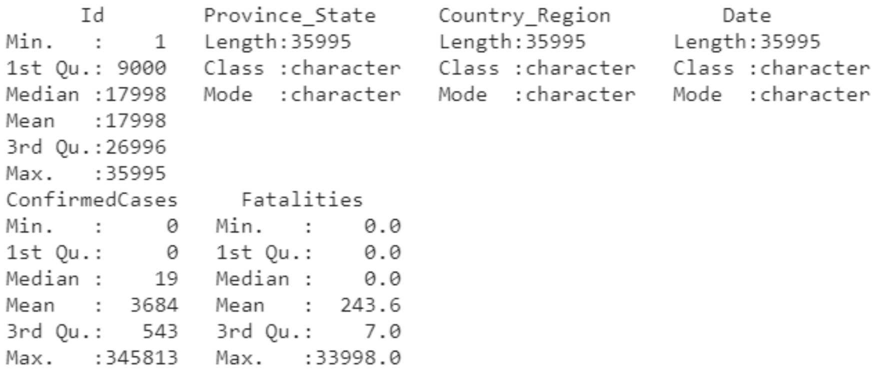


Sno: 0 **Date:** 0 **Time:** 0 **State.UnionTerritory:** 0 **Cured:** 0 **Deaths:** 0 **Confirmed:** 0

0

The statistical analysis of COVID-19 Global dataset

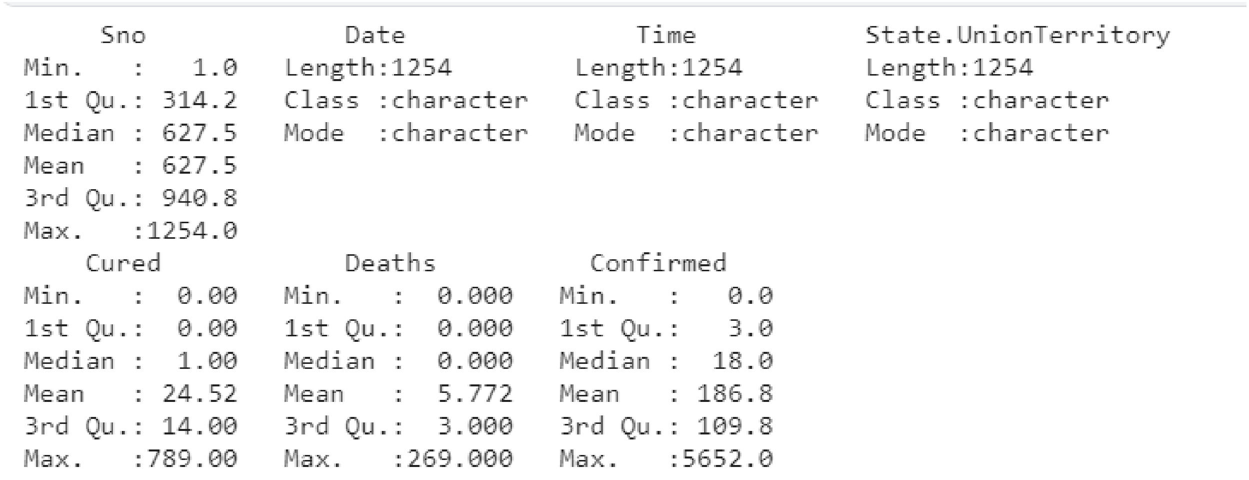
Summary(covid_data_train)



```
       Id        Province_State     Country_Region          Date          
 Min.   :    1   Length:35995       Length:35995       Length:35995      
 1st Qu.: 9000   Class :character   Class :character   Class :character  
 Median :17998   Mode  :character   Mode  :character   Mode  :character  
 Mean   :17998                                                            
 3rd Qu.:26996                                                            
 Max.   :35995                                                            
 ConfirmedCases     Fatalities     
 Min.   :     0   Min.   :    0.0  
 1st Qu.:     0   1st Qu.:    0.0  
 Median :    19   Median :    0.0  
 Mean   :  3684   Mean   :  243.6  
 3rd Qu.:   543   3rd Qu.:    7.0  
 Max.   :345813   Max.   :33998.0  
```

The statistical analysis of COVID-19 India dataset

Summary(covid_data_train)



```
      Sno               Date               Time            State.UnionTerritory
 Min.   :   1.0   Length:1254        Length:1254        Length:1254
 1st Qu.: 314.2   Class :character   Class :character   Class :character
 Median : 627.5   Mode  :character   Mode  :character   Mode  :character
 Mean   : 627.5
 3rd Qu.: 940.8
 Max.   :1254.0
     Cured             Deaths          Confirmed
 Min.   :  0.00   Min.   :  0.000   Min.   :   0.0
 1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:   3.0
 Median :  1.00   Median :  0.000   Median :  18.0
 Mean   : 24.52   Mean   :  5.772   Mean   : 186.8
 3rd Qu.: 14.00   3rd Qu.:  3.000   3rd Qu.: 109.8
 Max.   :789.00   Max.   :269.000   Max.   :5652.0
```

Exploratory Data Analysis for Covid-19 Global:

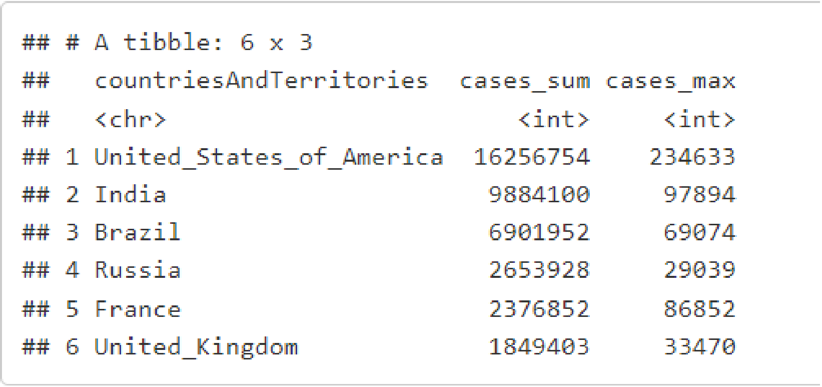
Total number of cases and max single day by country:

head(data %>%

group_by(countriesAndTerritories) %>%

summarise(cases_sum = sum(cases), cases_max = max(cases)) %>%

arrange(desc(cases_sum)))



```
## # A tibble: 6 x 3
##   countriesAndTerritories  cases_sum cases_max
##   <chr>                        <int>     <int>
## 1 United_States_of_America  16256754    234633
## 2 India                      9884100     97894
## 3 Brazil                     6901952     69074
## 4 Russia                     2653928     29039
## 5 France                     2376852     86852
## 6 United_Kingdom             1849403     33470
```

- From this we can observe that USA has the greatest number of confirmed cases.

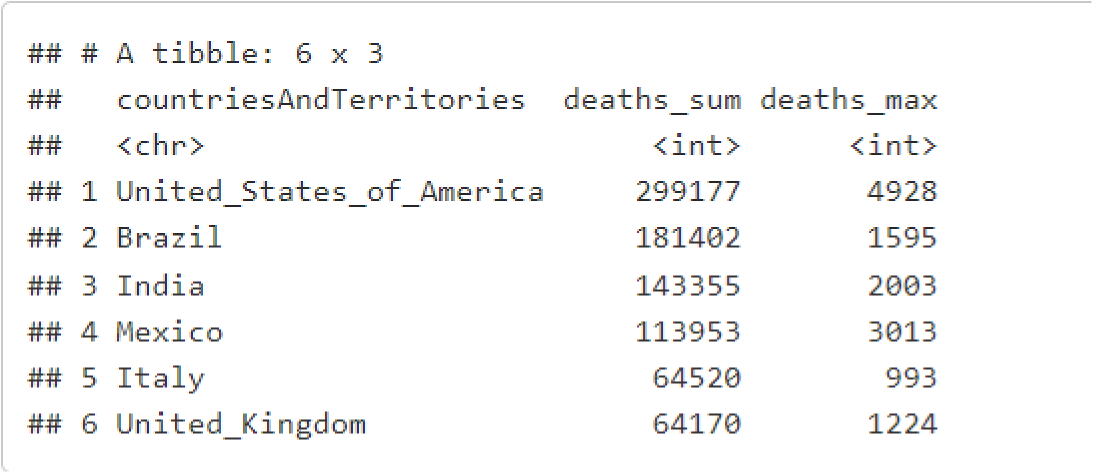
Total number of deaths and max single day by country:

head(data %>%

group_by(countriesAndTerritories) %>%

summarise(deaths_sum = sum(deaths), deaths_max = max(deaths)) %>%

arrange(desc(deaths_sum)))

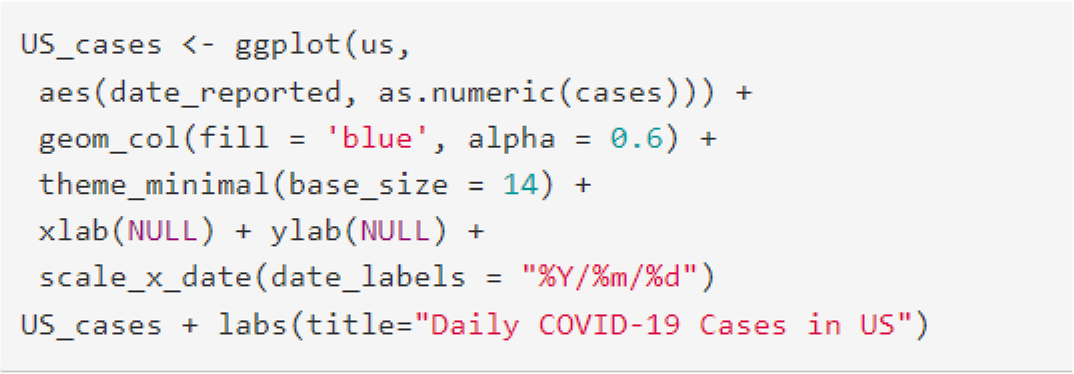


```
## # A tibble: 6 x 3
##   countriesAndTerritories  deaths_sum deaths_max
##   <chr>                         <int>      <int>
## 1 United_States_of_America     299177       4928
## 2 Brazil                       181402       1595
## 3 India                        143355       2003
## 4 Mexico                       113953       3013
## 5 Italy                         64520        993
## 6 United_Kingdom                64170       1224
```

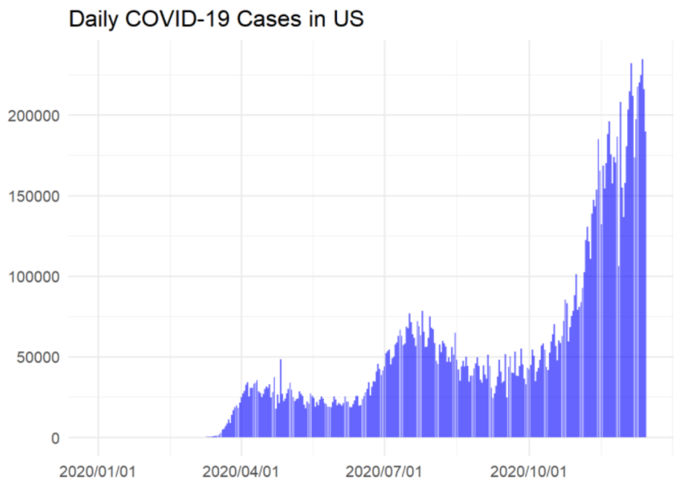
- Most number of deaths is observed in USA followed by Brazil, India, Mexico, Italy and United Kingdom.

Visualization of COVID-19 cases in different countries:

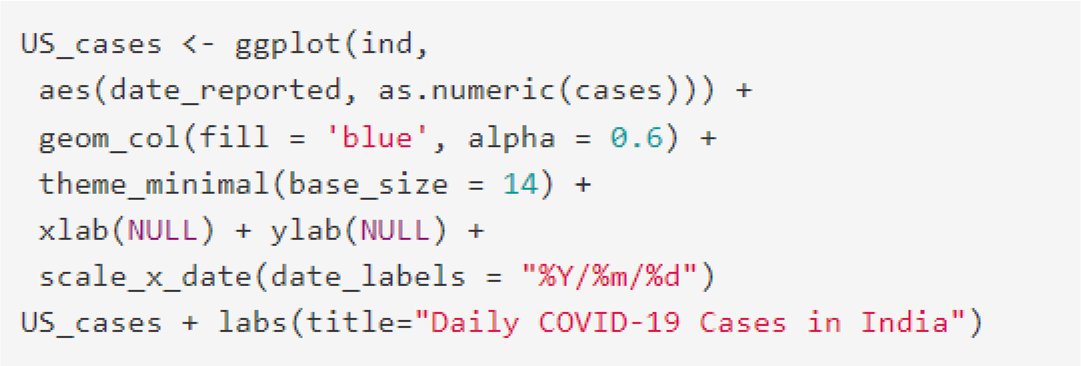
Daily confirmed cases in USA:



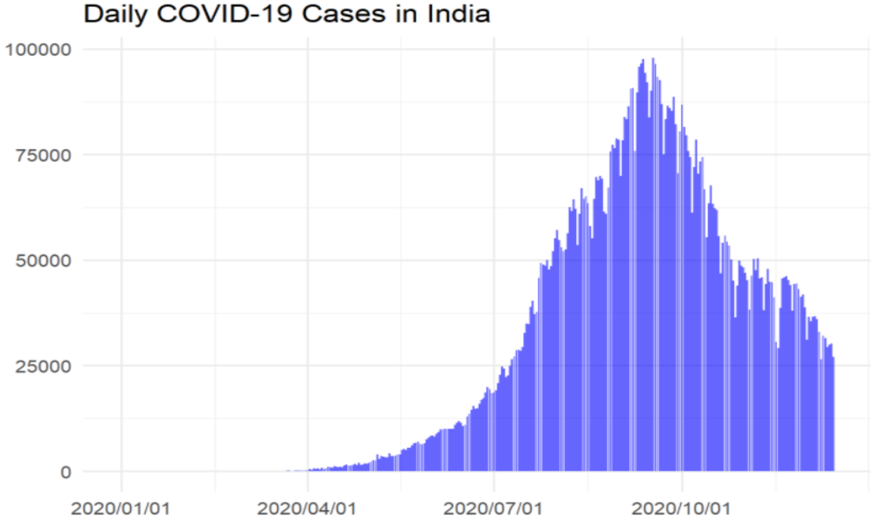
```
US_cases <- ggplot(us,
 aes(date_reported, as.numeric(cases))) +
 geom_col(fill = 'blue', alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")
US_cases + labs(title="Daily COVID-19 Cases in US")
```



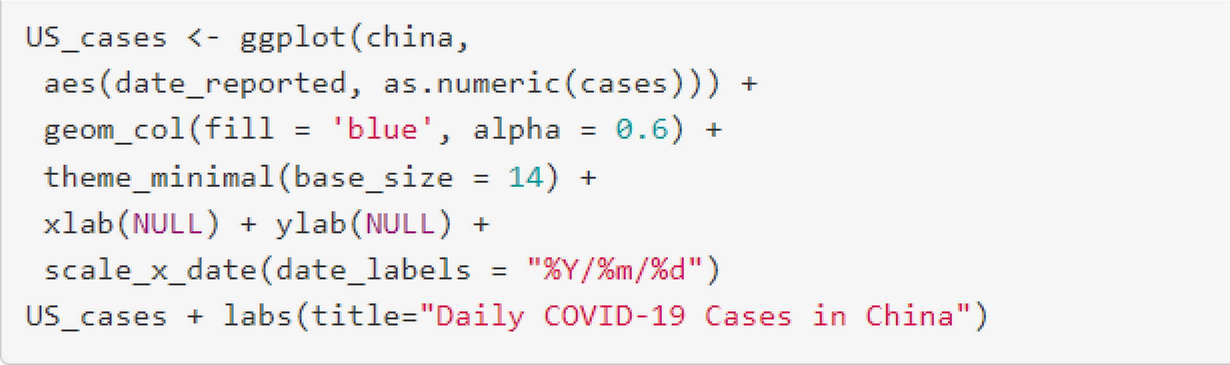
Daily confirmed cases in INDIA:



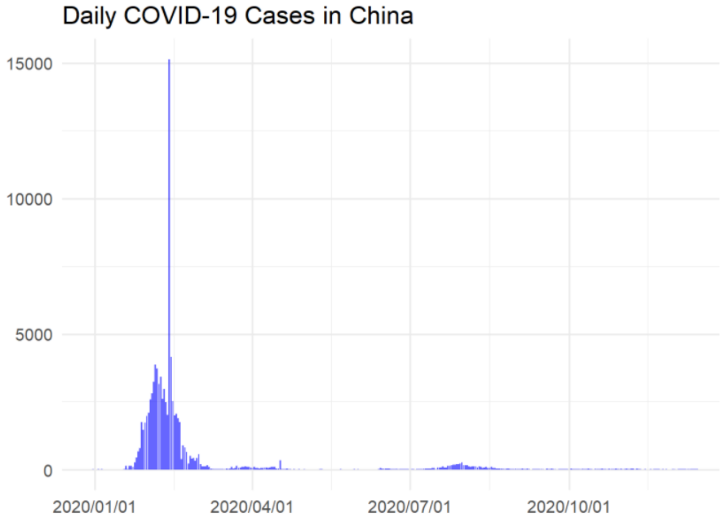
```
US_cases <- ggplot(ind,
 aes(date_reported, as.numeric(cases))) +
 geom_col(fill = 'blue', alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")
US_cases + labs(title="Daily COVID-19 Cases in India")
```



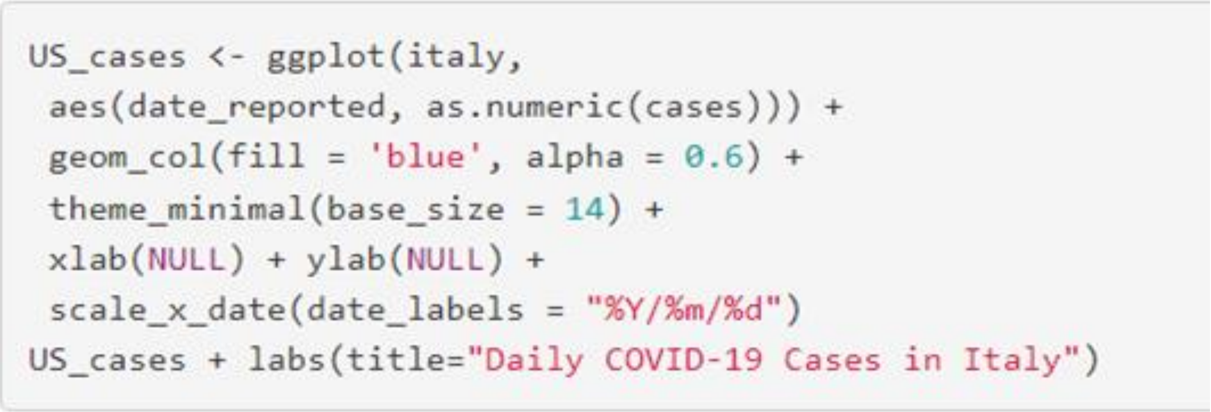
Daily confirmed cases in CHINA:



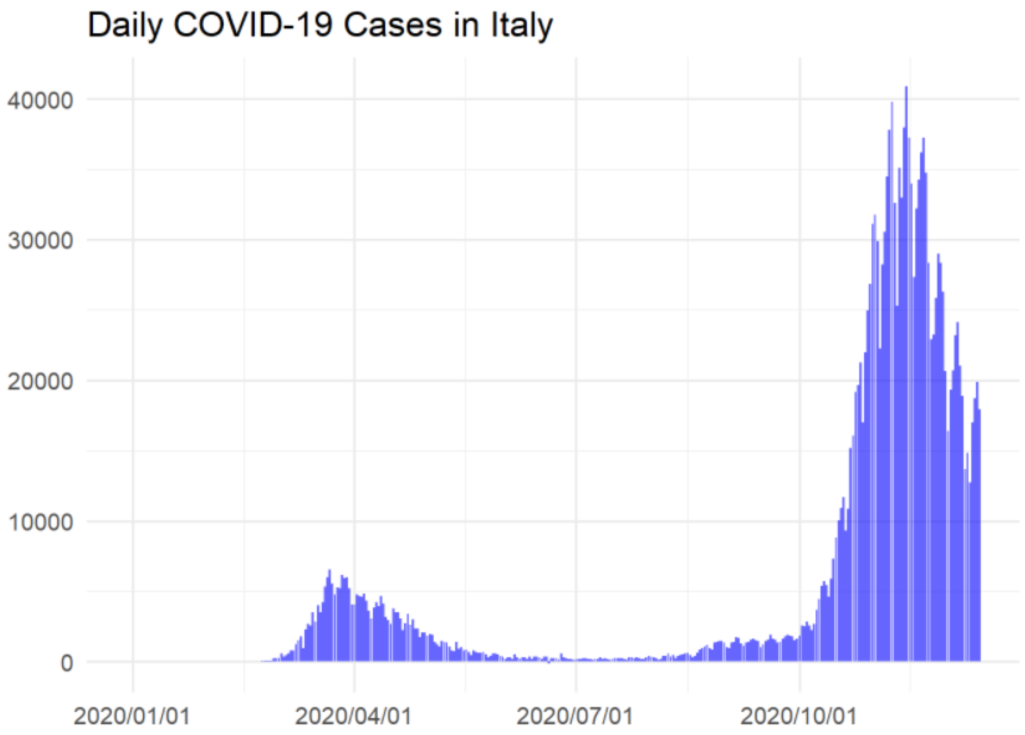
```
US_cases <- ggplot(china,
 aes(date_reported, as.numeric(cases))) +
 geom_col(fill = 'blue', alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")
US_cases + labs(title="Daily COVID-19 Cases in China")
```



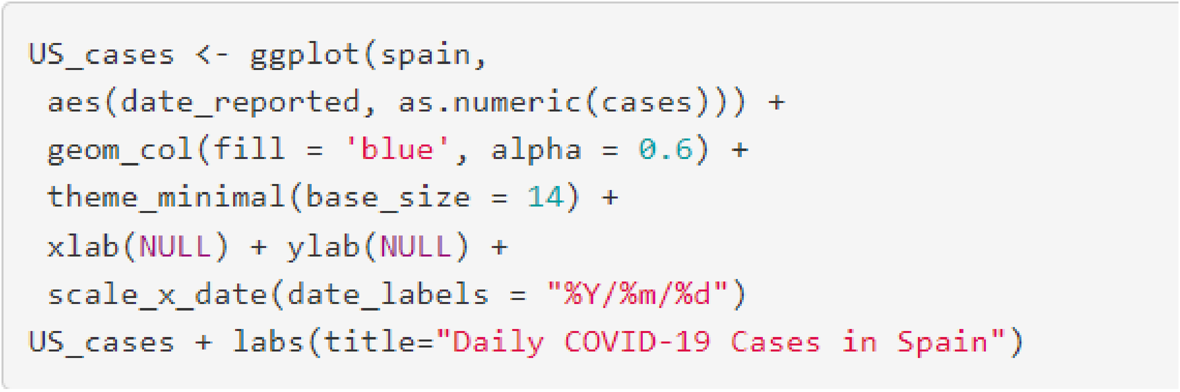
Daily confirmed cases in ITALY:



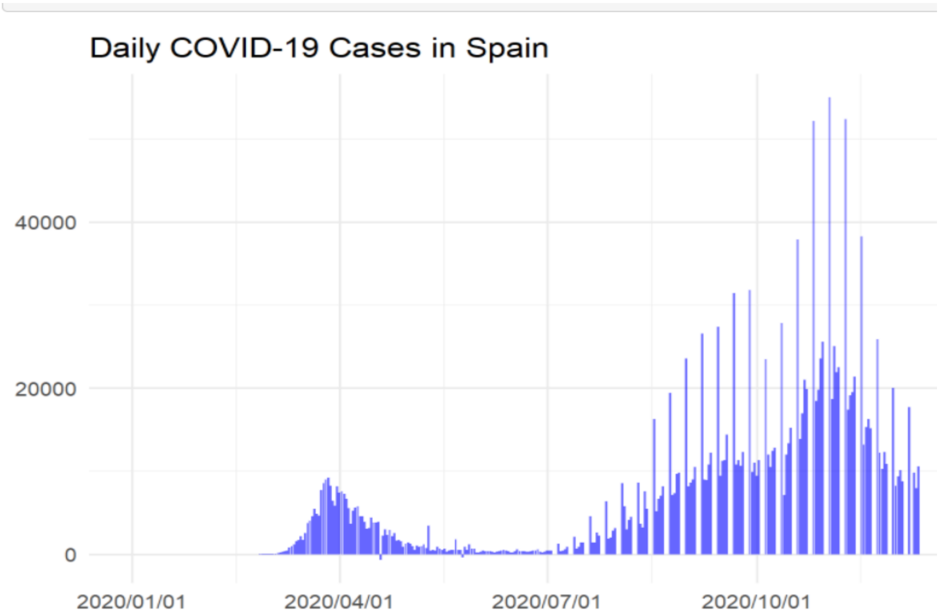
```
US_cases <- ggplot(italy,
 aes(date_reported, as.numeric(cases))) +
 geom_col(fill = 'blue', alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")
US_cases + labs(title="Daily COVID-19 Cases in Italy")
```



Daily confirmed cases in SPAIN:

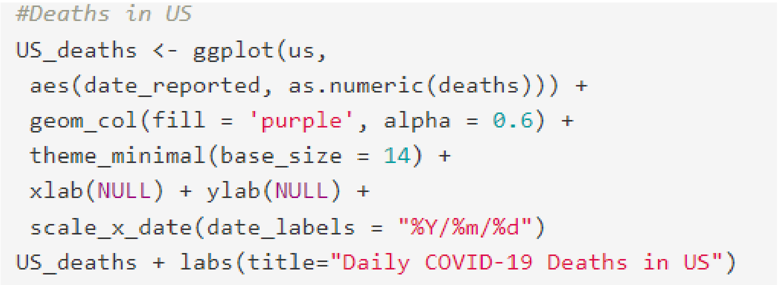


```
US_cases <- ggplot(spain,
 aes(date_reported, as.numeric(cases))) +
 geom_col(fill = 'blue', alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")
US_cases + labs(title="Daily COVID-19 Cases in Spain")
```

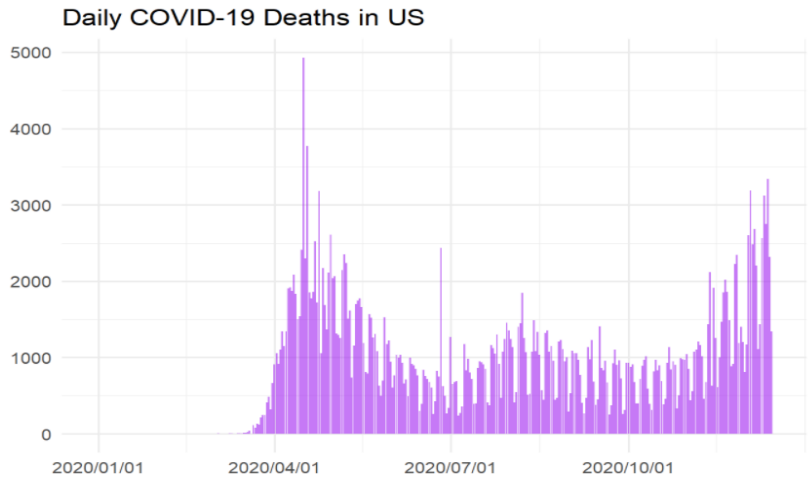


Visualization of COVID-19 Deaths in different countries:

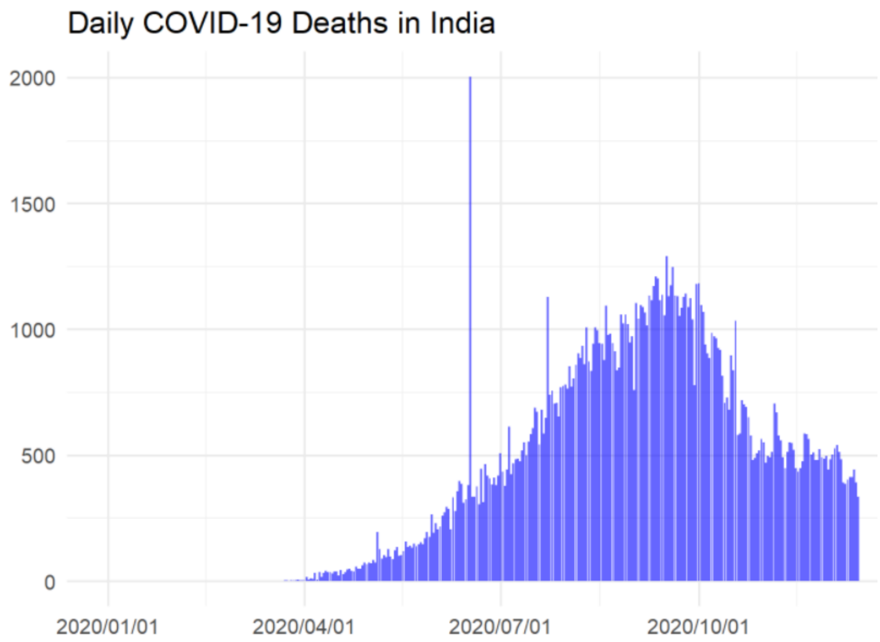
Daily COVI-19 Deaths in USA:



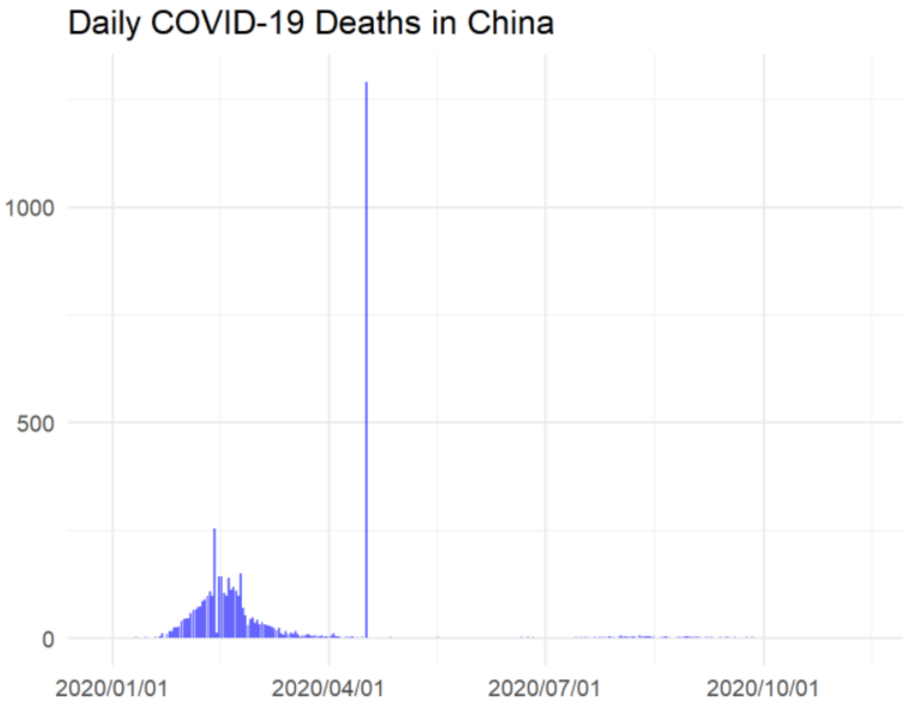
```
#Deaths in US
US_deaths <- ggplot(us,
 aes(date_reported, as.numeric(deaths))) +
 geom_col(fill = 'purple', alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")
US_deaths + labs(title="Daily COVID-19 Deaths in US")
```



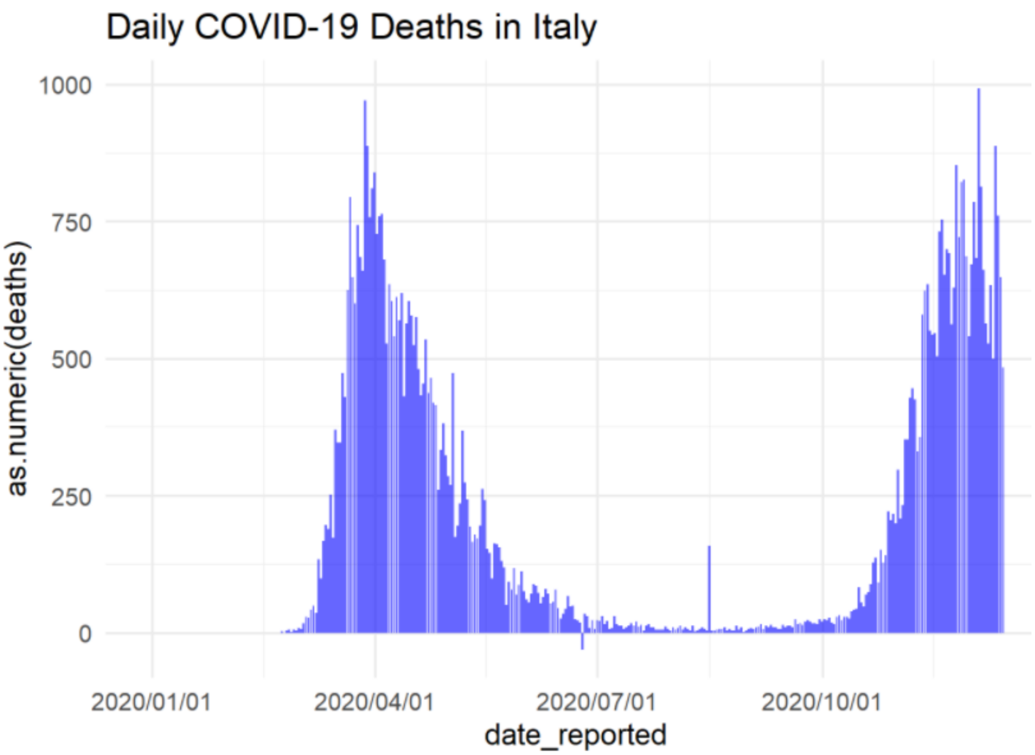
Daily COVI-19 Deaths in India:



Daily COVI-19 Deaths in China:



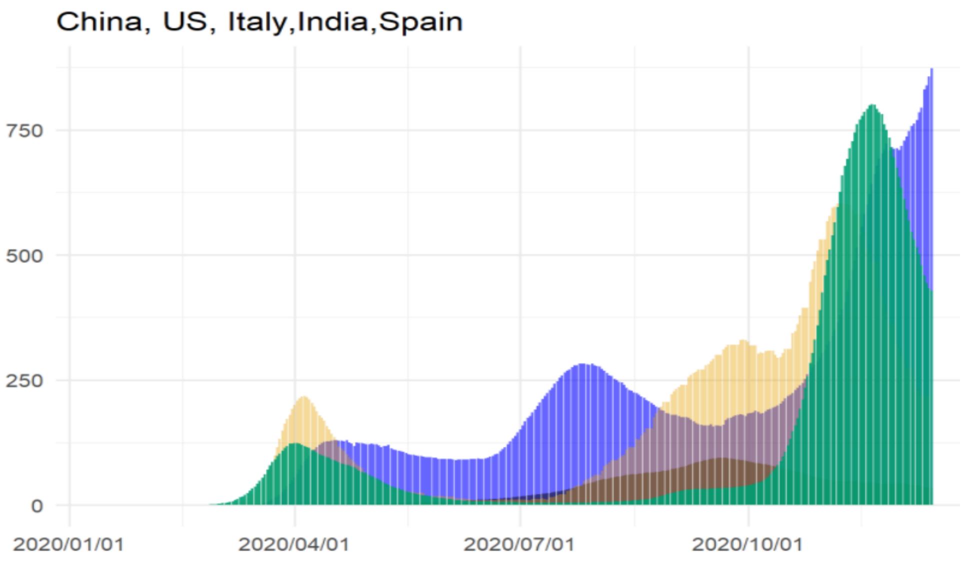
Daily COVI-19 Deaths in Italy:



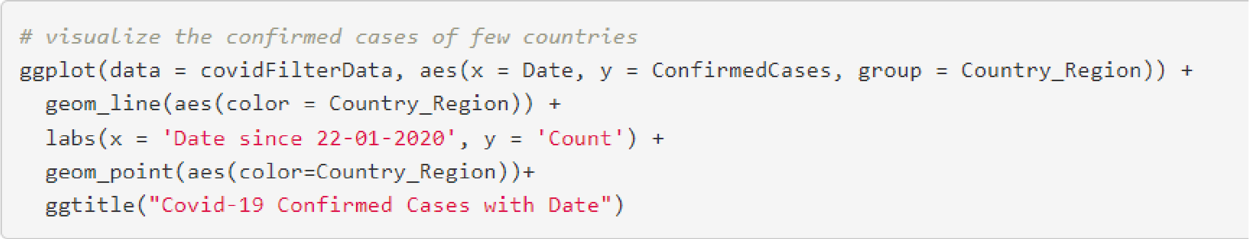
Density Plot for all countries:

CHINA: RED | USA: BLUE | INDIA: BLACK | ITALY: GREEN |

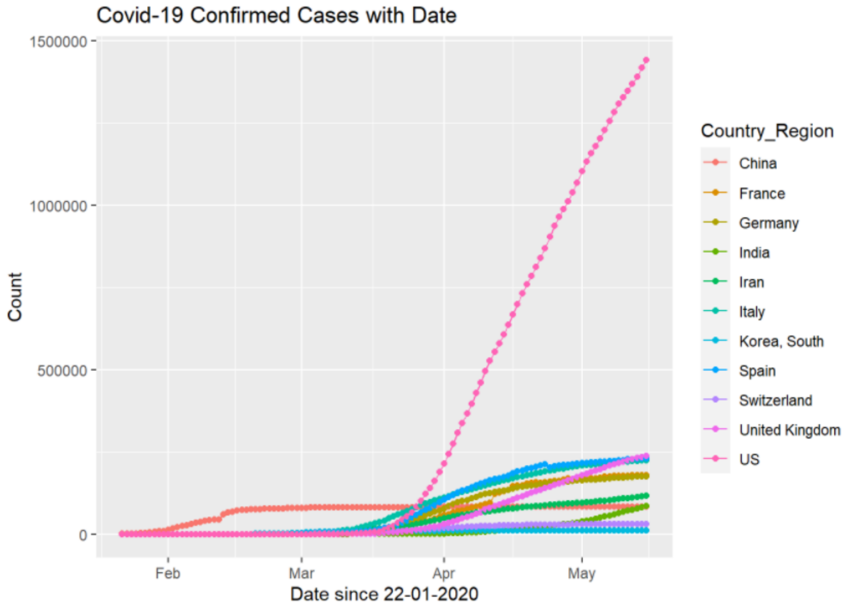
SPAIN: ORANGE



Visualization of confirmed cases for few countries:

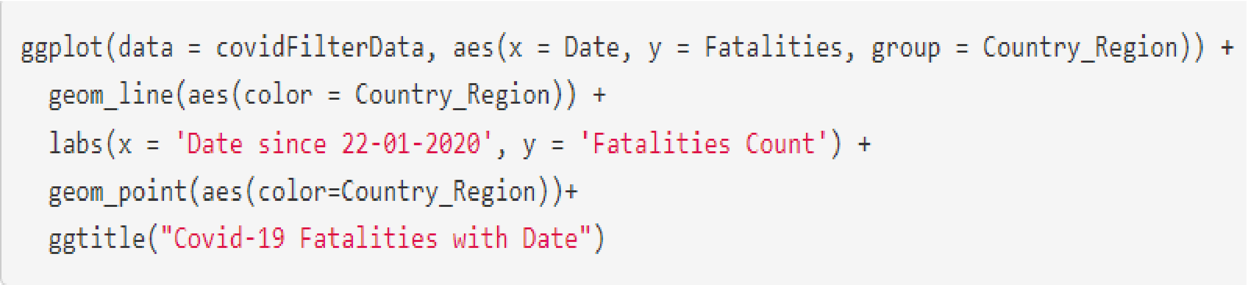


```
# visualize the confirmed cases of few countries
ggplot(data = covidFilterData, aes(x = Date, y = ConfirmedCases, group = Country_Region)) +
  geom_line(aes(color = Country_Region)) +
  labs(x = 'Date since 22-01-2020', y = 'Count') +
  geom_point(aes(color=Country_Region))+
  ggtitle("Covid-19 Confirmed Cases with Date")
```

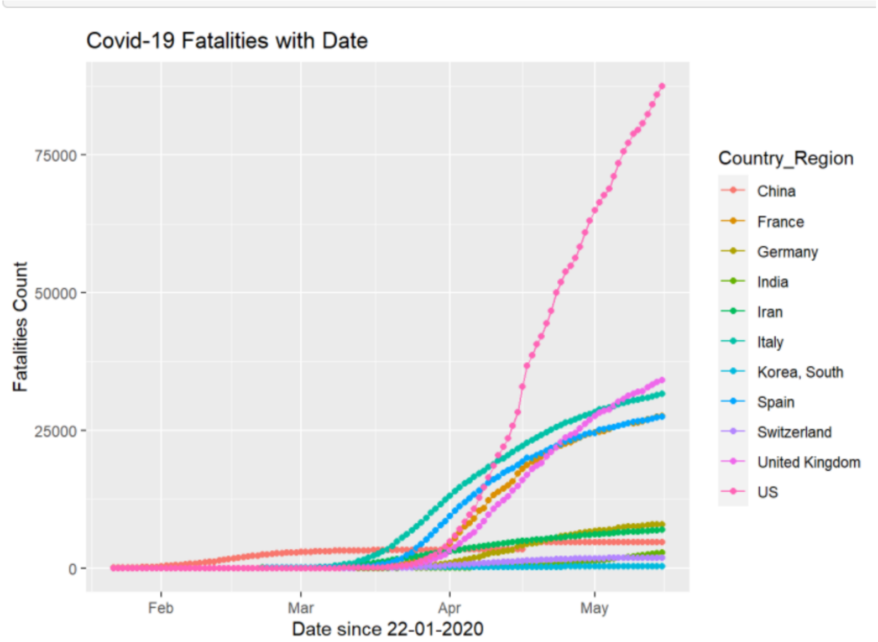


- From the graph we can observe that the US has the greatest number of Covid-19 cases.

Visualization of deaths for few countries:

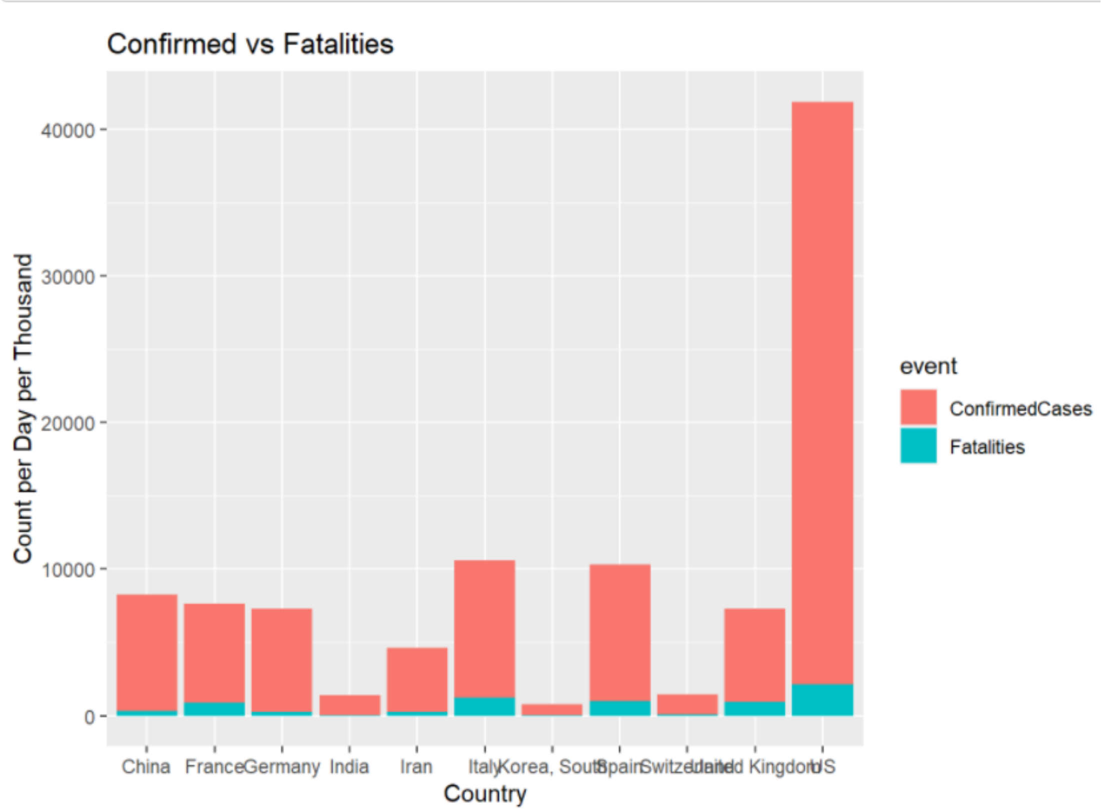


```
ggplot(data = covidFilterData, aes(x = Date, y = Fatalities, group = Country_Region)) +
  geom_line(aes(color = Country_Region)) +
  labs(x = 'Date since 22-01-2020', y = 'Fatalities Count') +
  geom_point(aes(color=Country_Region))+
  ggtitle("Covid-19 Fatalities with Date")
```



- From the graph we can observe that the US has the greatest number of Covid-19 deaths.

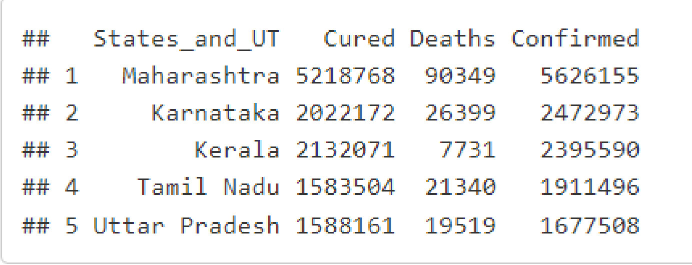
Visualization of confirmed cases Vs Fatalities:



Exploratory Data Analysis for Covid-19 in INDIA:

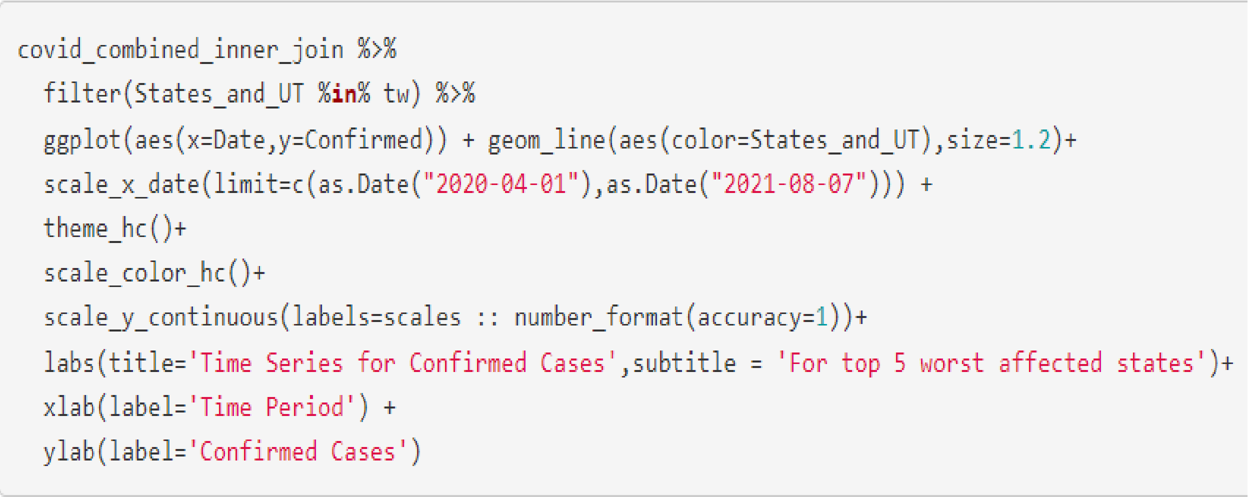
First finding the top most affected states in India:

```
top_worst <- covid_combined_inner_join %>%
  filter(Date == "2021-05-26" ) %>%
  select(States_and_UT,Cured,Deaths,Confirmed) %>%
  arrange(desc(Confirmed)) %>%
  top_n(5)
```

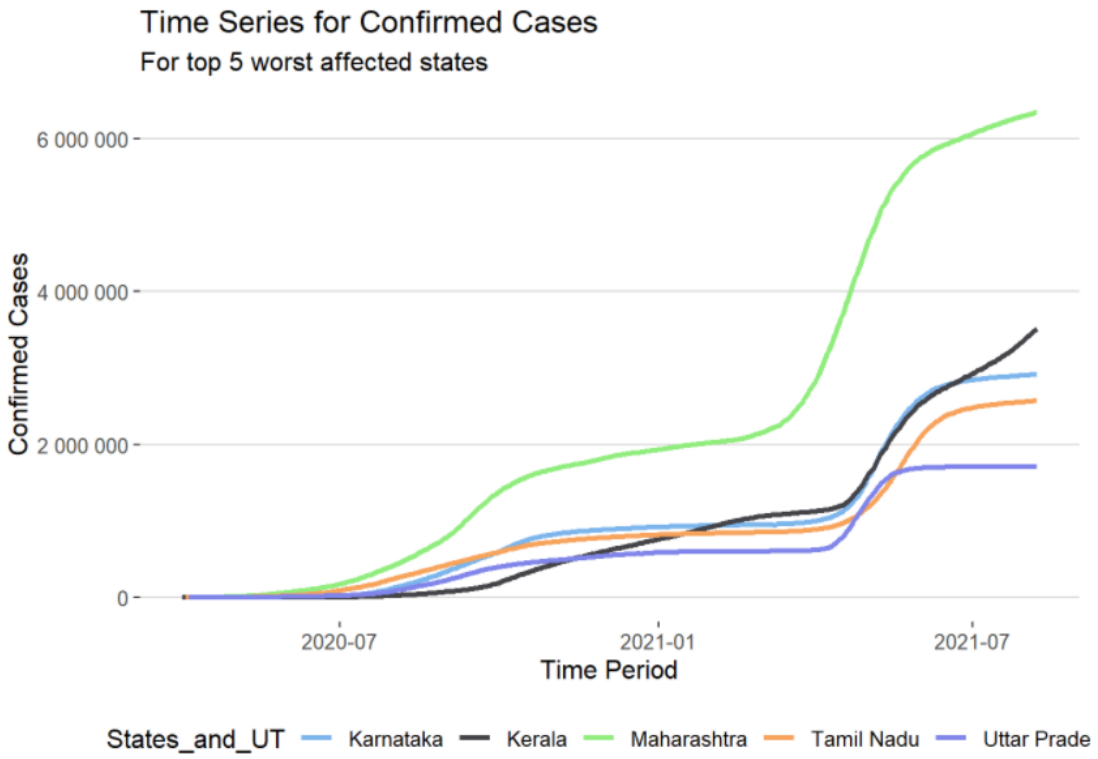


```
##   States_and_UT   Cured Deaths Confirmed
## 1   Maharashtra 5218768  90349   5626155
## 2     Karnataka 2022172  26399   2472973
## 3        Kerala 2132071   7731   2395590
## 4    Tamil Nadu 1583504  21340   1911496
## 5 Uttar Pradesh 1588161  19519   1677508
```

Time Series Graph for Confirmed Cases for Top most affected states in India:

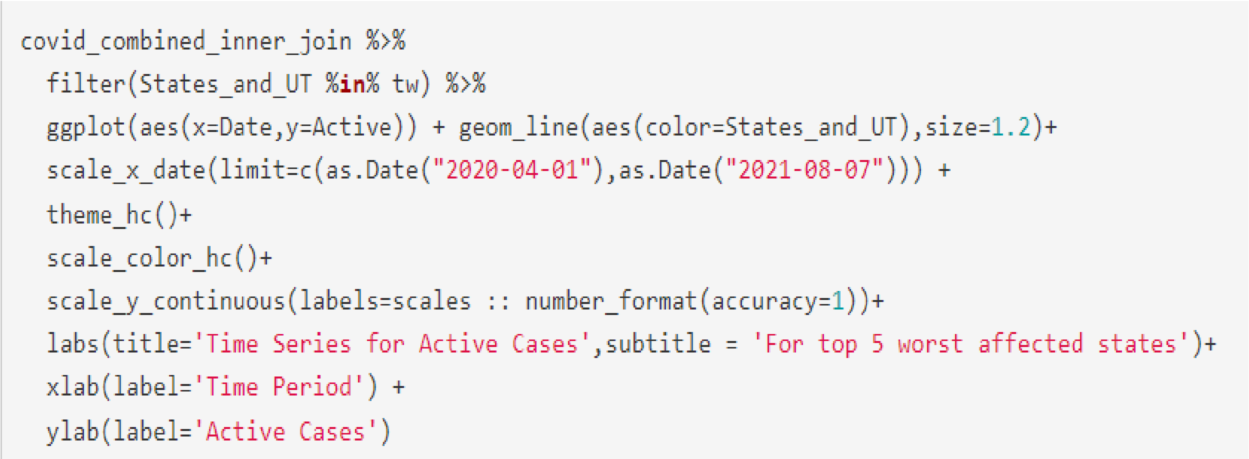


```
covid_combined_inner_join %>%
  filter(States_and_UT %in% tw) %>%
  ggplot(aes(x=Date,y=Confirmed)) + geom_line(aes(color=States_and_UT),size=1.2)+
  scale_x_date(limit=c(as.Date("2020-04-01"),as.Date("2021-08-07"))) +
  theme_hc()+
  scale_color_hc()+
  scale_y_continuous(labels=scales :: number_format(accuracy=1))+
  labs(title='Time Series for Confirmed Cases',subtitle = 'For top 5 worst affected states')+
  xlab(label='Time Period') +
  ylab(label='Confirmed Cases')
```

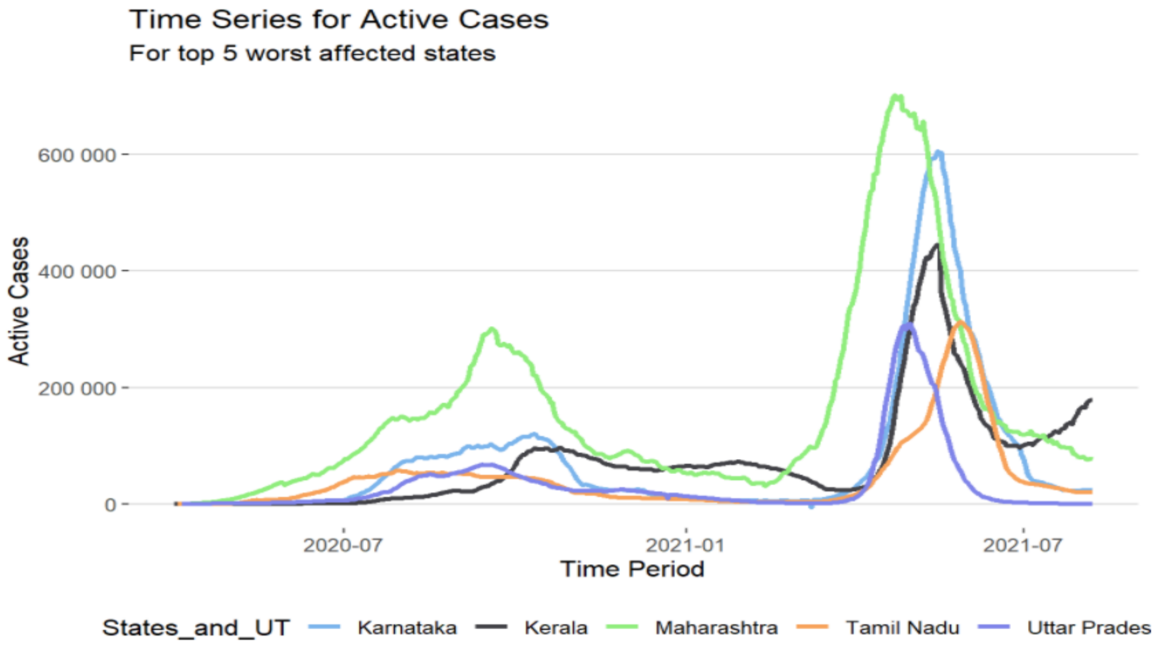


From the above graph we observe that the Maharashtra state has the greatest number of confirmed cases. The number of cases recorded are greater than 600000. The second highest is Kerala followed by Karnataka, Tamil Nadu and Uttar Pradesh.

Time Series Graph for Active Cases for Top most affected states in India:

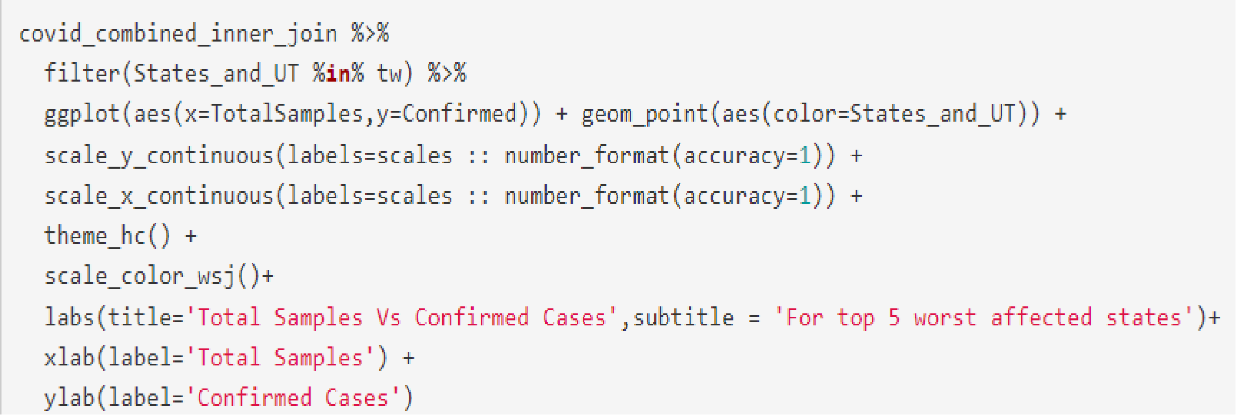


```
covid_combined_inner_join %>%
  filter(States_and_UT %in% tw) %>%
  ggplot(aes(x=Date,y=Active)) + geom_line(aes(color=States_and_UT),size=1.2)+
  scale_x_date(limit=c(as.Date("2020-04-01"),as.Date("2021-08-07"))) +
  theme_hc()+
  scale_color_hc()+
  scale_y_continuous(labels=scales :: number_format(accuracy=1))+
  labs(title='Time Series for Active Cases',subtitle = 'For top 5 worst affected states')+
  xlab(label='Time Period') +
  ylab(label='Active Cases')
```

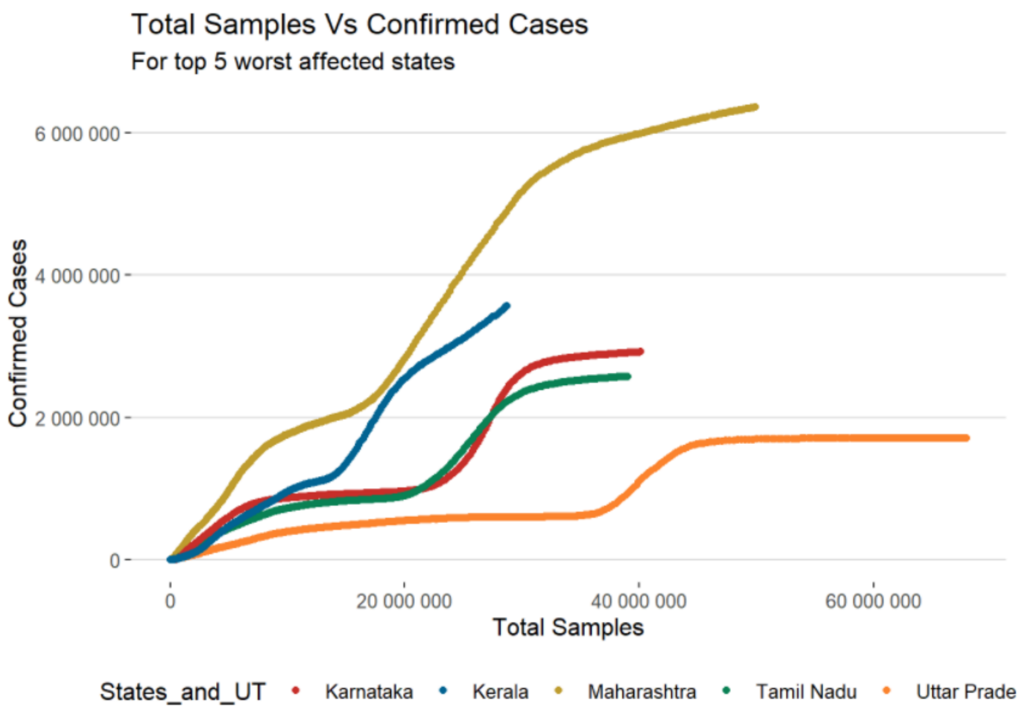


The time series graph is from year 2020 to August 2021. From the graph we can see there are two peeks. One peek in the year of 2020 which denotes the first wave which hit India. In the first wave we can observe that the number of Active cases recorded was around 300000. The second peek denotes the second wave which hit India in April 2021 and lasted till July 2021. Maharashtra has recorded the greatest number of Active cases. The number cases recorded were greater than 600000.

Total sample Vs Confirmed Cases:

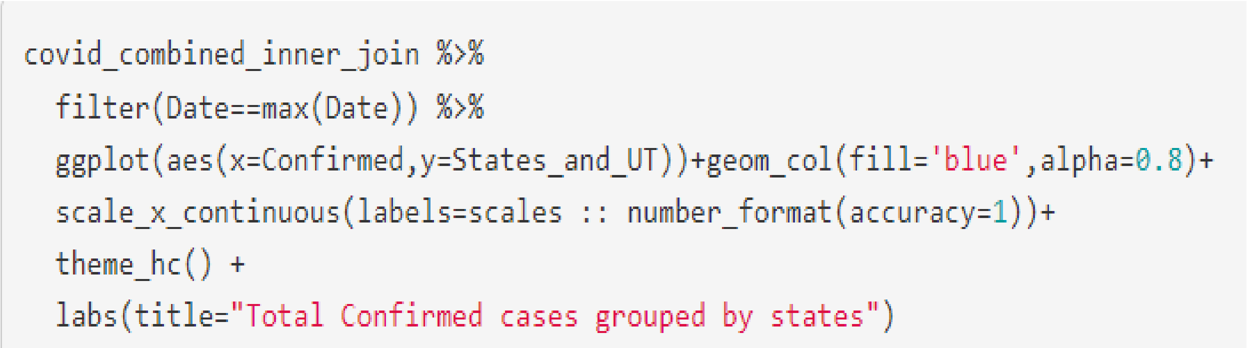


```
covid_combined_inner_join %>%
  filter(States_and_UT %in% tw) %>%
  ggplot(aes(x=TotalSamples,y=Confirmed)) + geom_point(aes(color=States_and_UT)) +
  scale_y_continuous(labels=scales :: number_format(accuracy=1)) +
  scale_x_continuous(labels=scales :: number_format(accuracy=1)) +
  theme_hc() +
  scale_color_wsj()+
  labs(title='Total Samples Vs Confirmed Cases',subtitle = 'For top 5 worst affected states')+
  xlab(label='Total Samples') +
  ylab(label='Confirmed Cases')
```

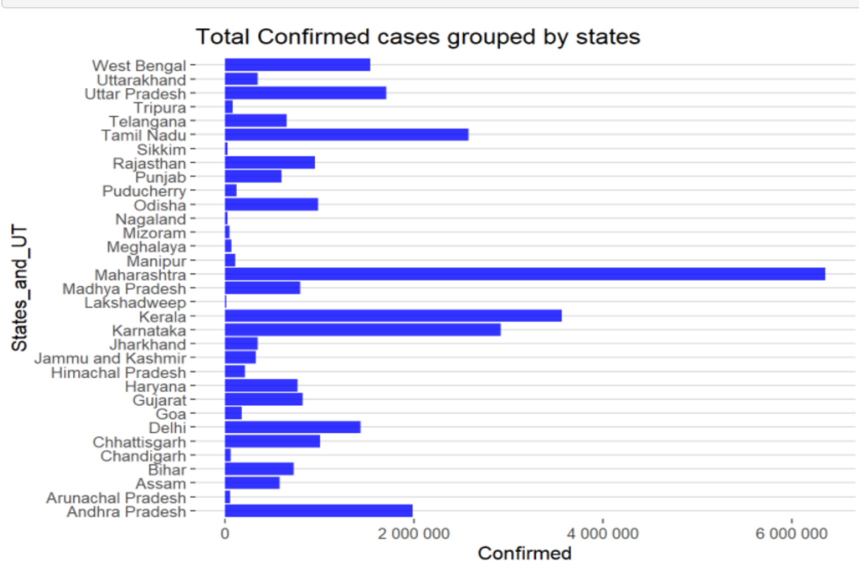


From the graph we can infer that the state of Utter Pradesh has done good job by taking maximum number of sample and at the same time maintain a smaller number of Covid cases. Maharashtra has maximum number of cases and less testing samples.

Total Confirmed cases grouped by states:

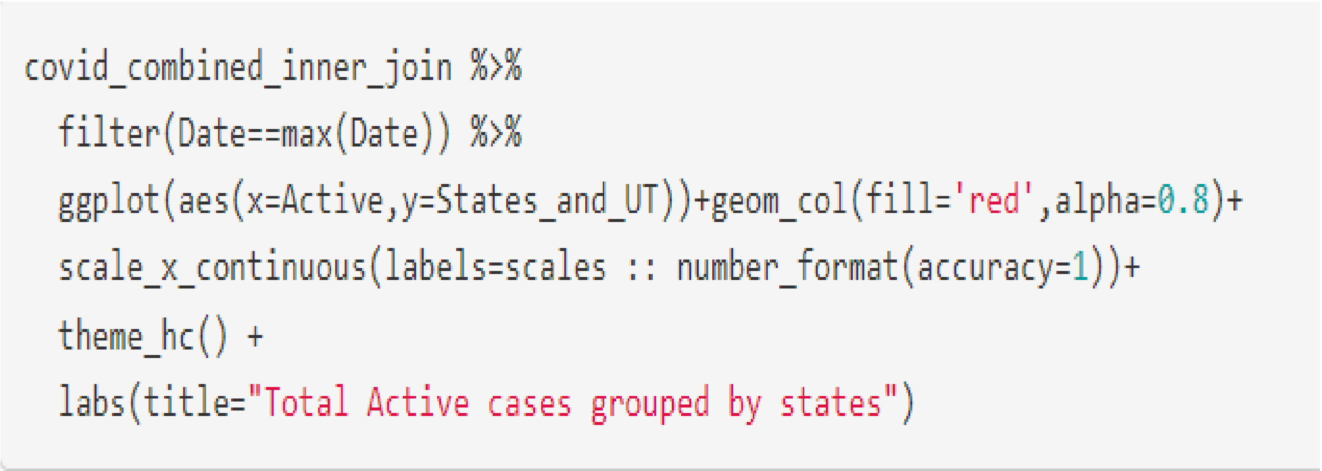


```
covid_combined_inner_join %>%
  filter(Date==max(Date)) %>%
  ggplot(aes(x=Confirmed,y=States_and_UT))+geom_col(fill='blue',alpha=0.8)+
  scale_x_continuous(labels=scales :: number_format(accuracy=1))+
  theme_hc() +
  labs(title="Total Confirmed cases grouped by states")
```

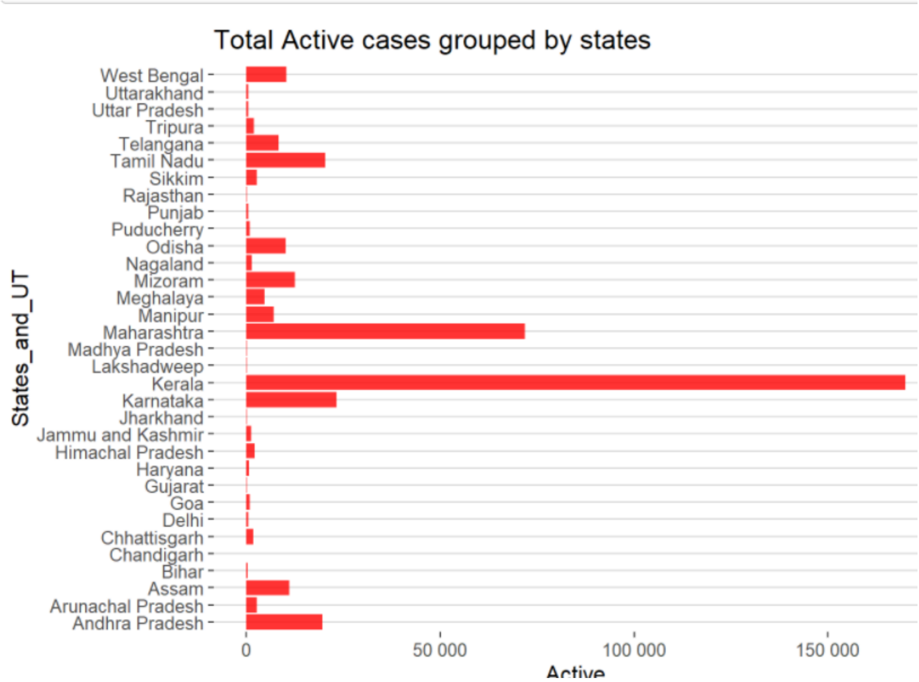


The confirmed cases in Maharashtra are maximum. The other states with maximum cases are Kerala, Karnataka, Tamil Nadu, Andhra Pradesh, Uttar Pradesh and West Bengal.

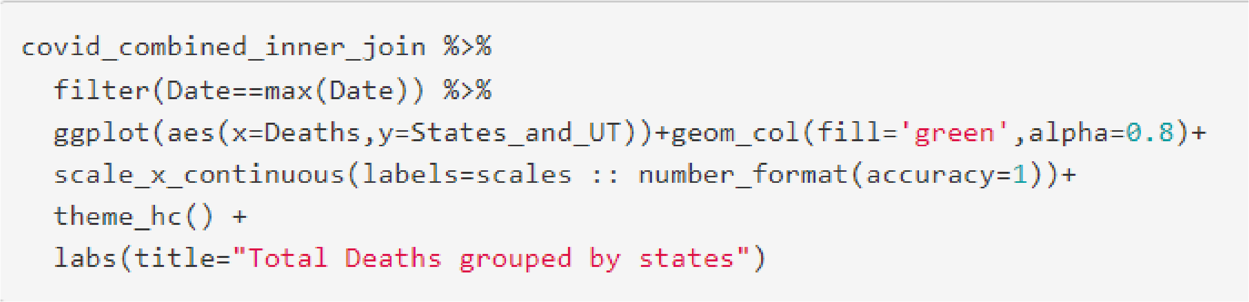
Total Active cases grouped by states:



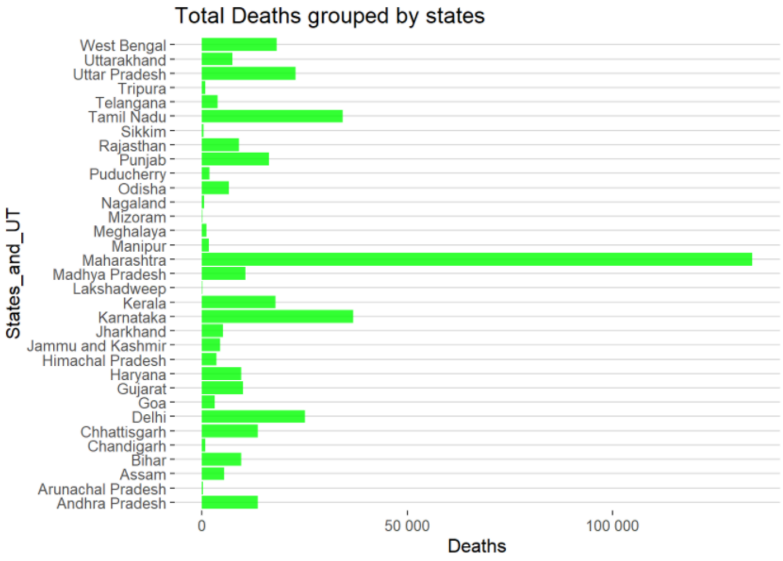
```
covid_combined_inner_join %>%
  filter(Date==max(Date)) %>%
  ggplot(aes(x=Active,y=States_and_UT))+geom_col(fill='red',alpha=0.8)+
  scale_x_continuous(labels=scales :: number_format(accuracy=1))+
  theme_hc() +
  labs(title="Total Active cases grouped by states")
```



The Active cases in Kerala is maximum followed by Maharashtra, Andhra Pradesh, Tamil Nadu and Karnataka.

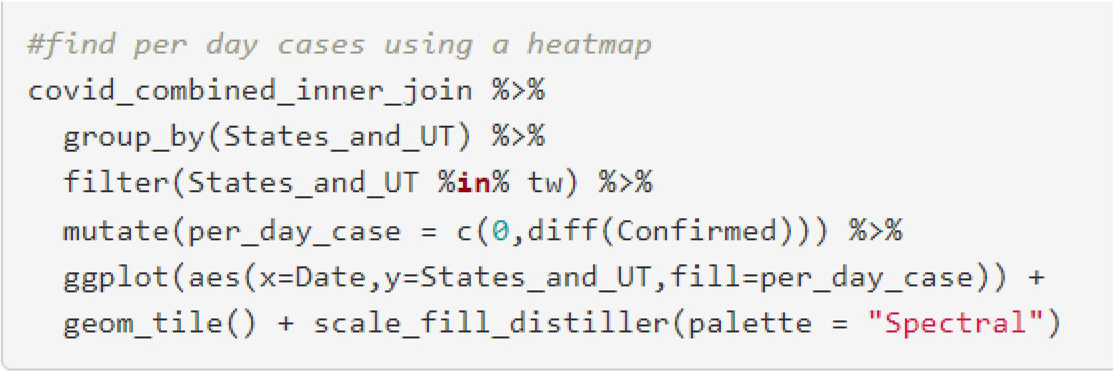


```
covid_combined_inner_join %>%
  filter(Date==max(Date)) %>%
  ggplot(aes(x=Deaths,y=States_and_UT))+geom_col(fill='green',alpha=0.8)+
  scale_x_continuous(labels=scales :: number_format(accuracy=1))+
  theme_hc() +
  labs(title="Total Deaths grouped by states")
```

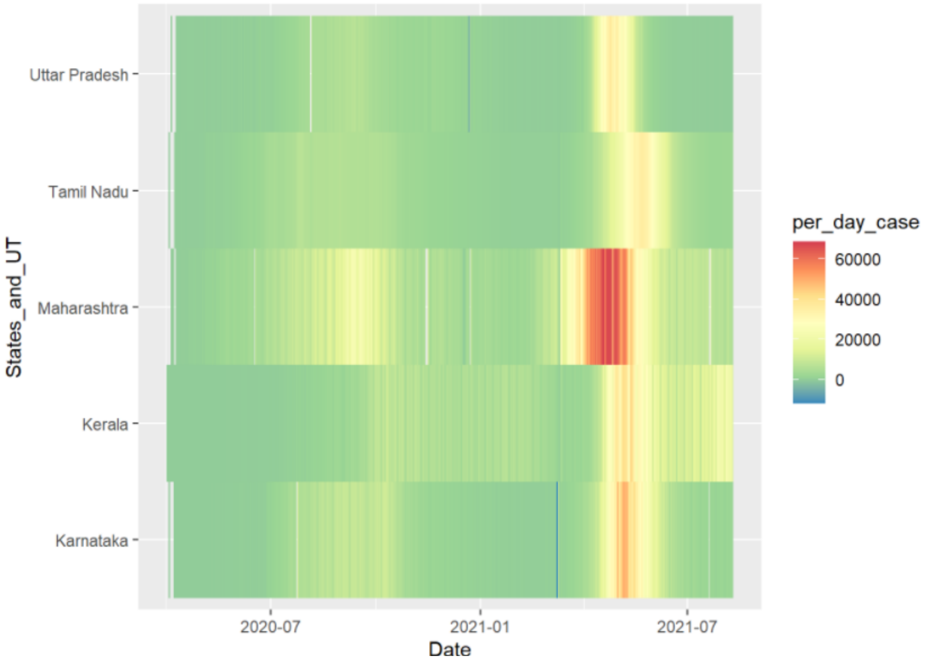


Most number of deaths were recorded in Maharashtra state followed by Karnataka, Kerala, Tamil Nadu, Delhi and Uttar Pradesh.

Heat-Map for Cases Per Day:

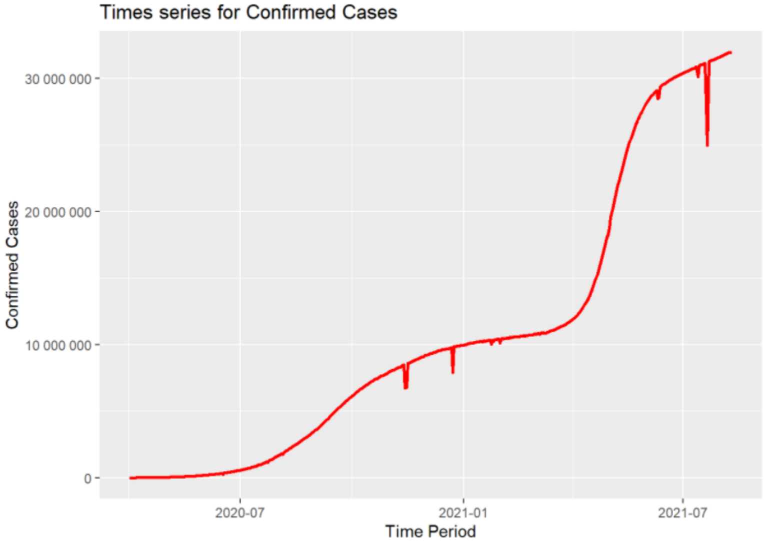


```
#find per day cases using a heatmap
covid_combined_inner_join %>%
  group_by(States_and_UT) %>%
  filter(States_and_UT %in% tw) %>%
  mutate(per_day_case = c(0,diff(Confirmed))) %>%
  ggplot(aes(x=Date,y=States_and_UT,fill=per_day_case)) +
  geom_tile() + scale_fill_distiller(palette = "Spectral")
```

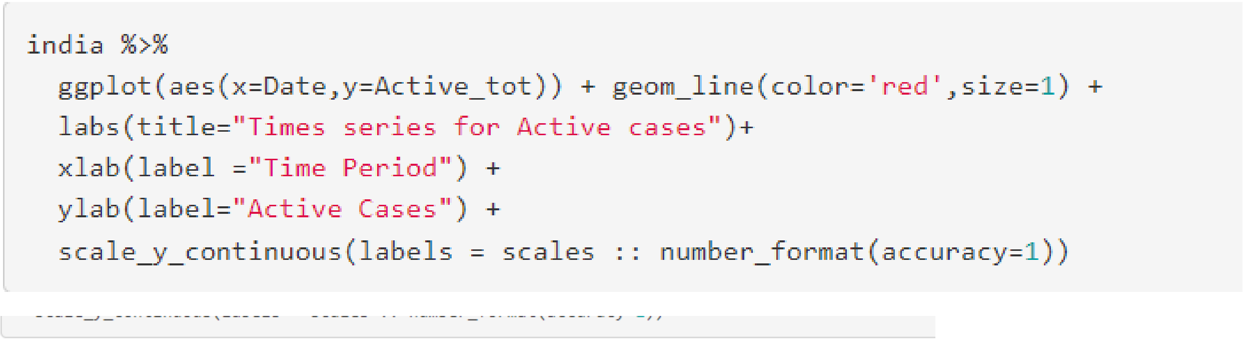


The cases per day is high in the Maharashtra state which is denoted by red color. We can also observe that the number of cases in 2020 is less compared to the cases in the year of 2021.

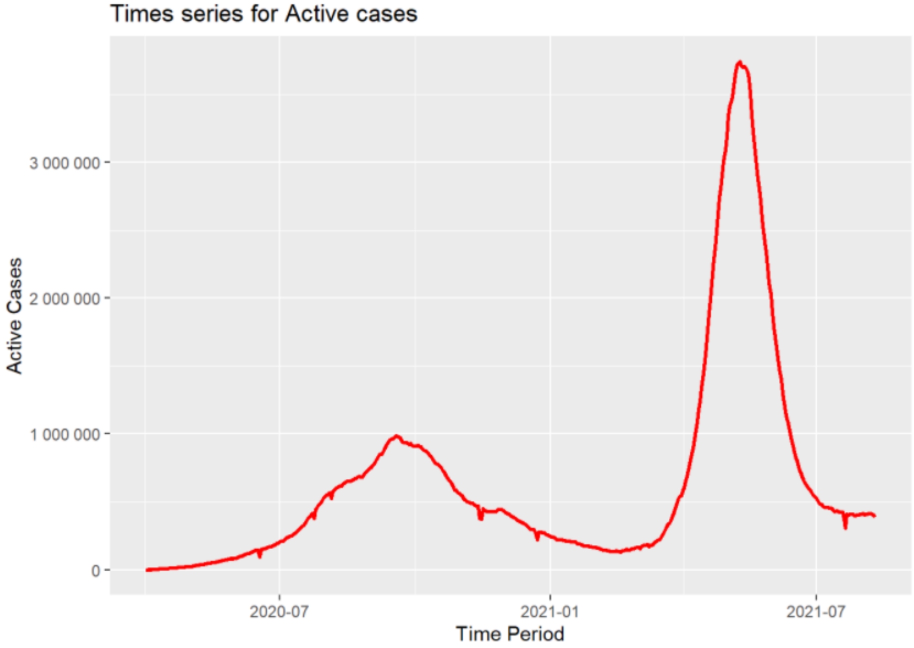
Times series for Confirmed Cases for Whole India:



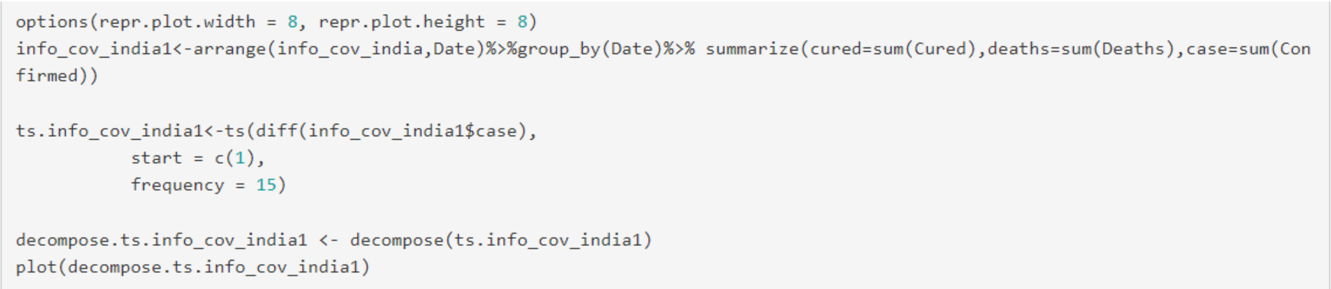
Times series for Active Cases for Whole India:



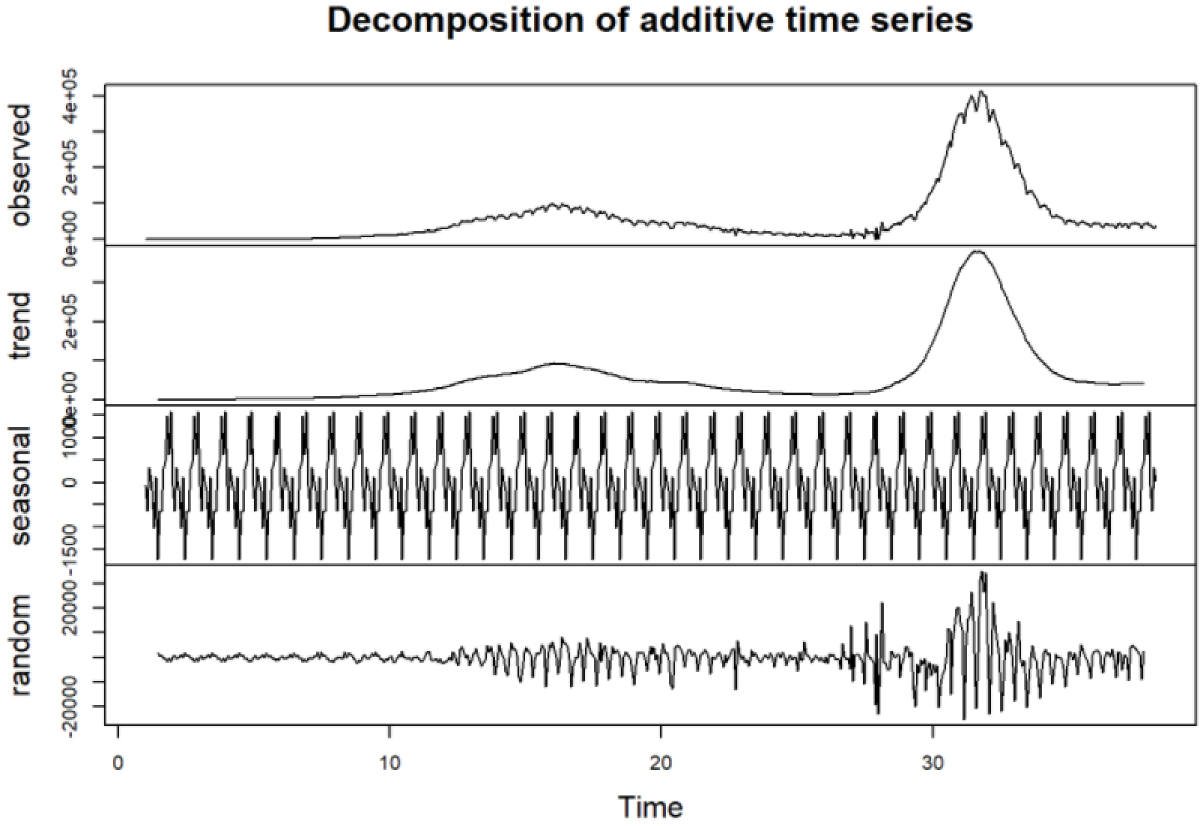
```
india %>%
  ggplot(aes(x=Date,y=Active_tot)) + geom_line(color='red',size=1) +
  labs(title="Times series for Active cases")+
  xlab(label ="Time Period") +
  ylab(label="Active Cases") +
  scale_y_continuous(labels = scales :: number_format(accuracy=1))
```



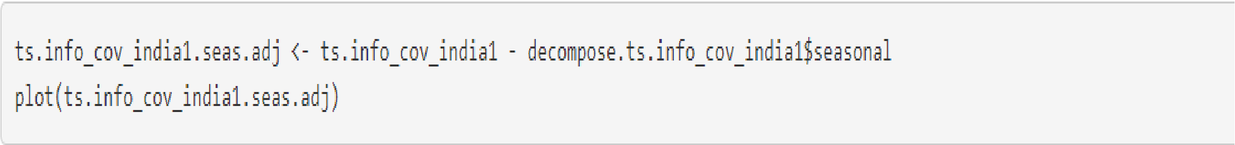
PREDICTION MODEL:



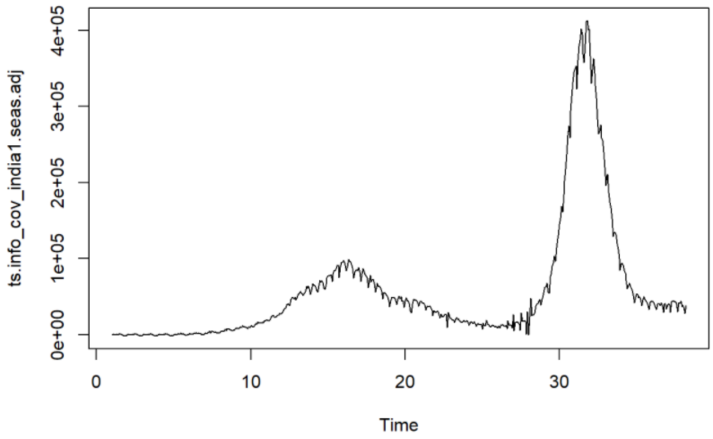
```
options(repr.plot.width = 8, repr.plot.height = 8)
info_cov_india1<-arrange(info_cov_india,Date)%>%group_by(Date)%>% summarize(cured=sum(Cured),deaths=sum(Deaths),case=sum(Con
firmed))

ts.info_cov_india1<-ts(diff(info_cov_india1$case),
           start = c(1),
           frequency = 15)

decompose.ts.info_cov_india1 <- decompose(ts.info_cov_india1)
plot(decompose.ts.info_cov_india1)
```



Time series model is being created and modified for better prediction.

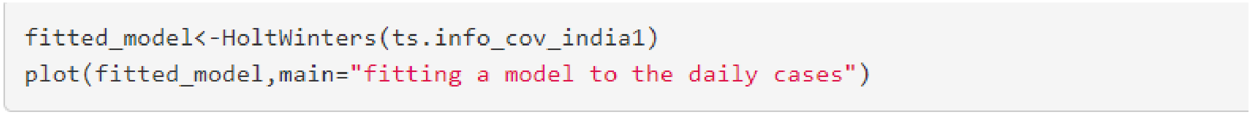


```
ts.info_cov_india1.seas.adj <- ts.info_cov_india1 - decompose.ts.info_cov_india1$seasonal
plot(ts.info_cov_india1.seas.adj)
```

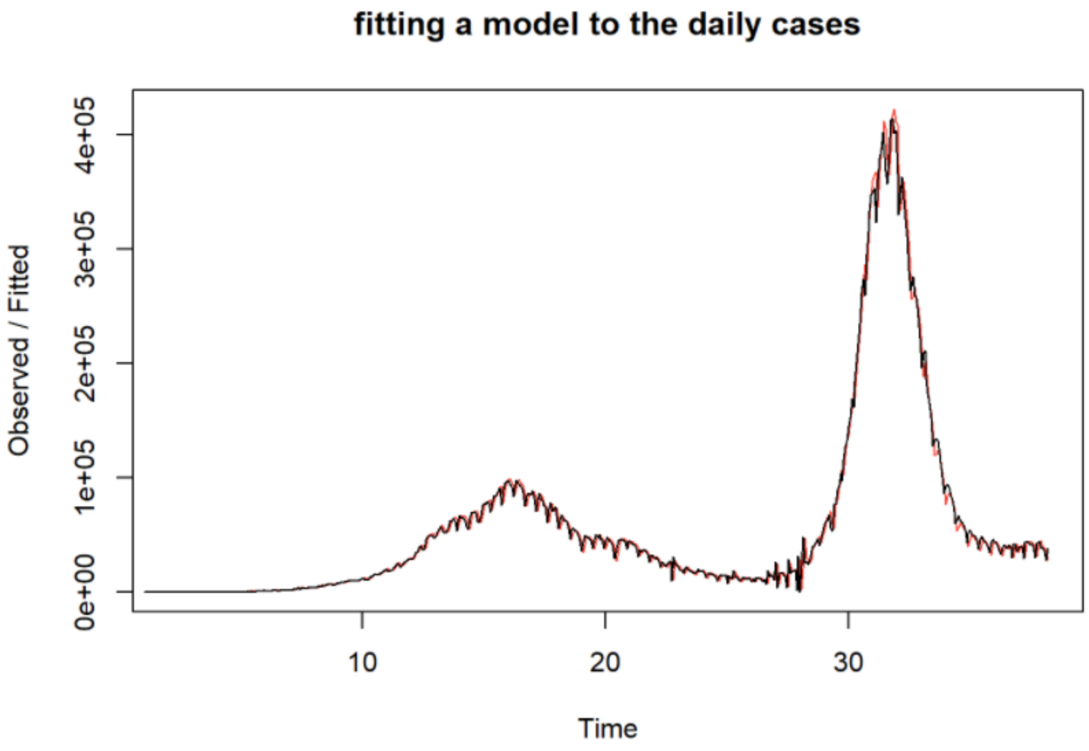


The Above plot for the Time Series datasets for number of cases in India.

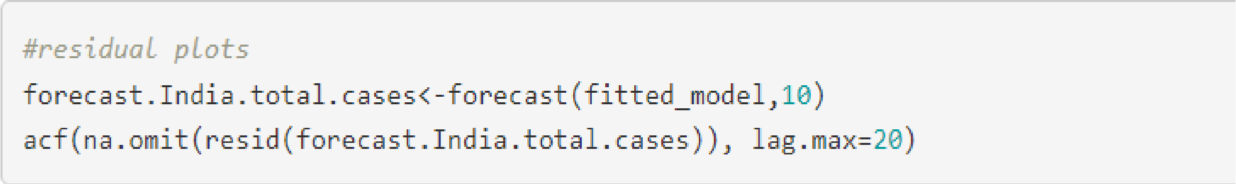
Fitting the Holt-Winters model for the time series dataset and visualizing the fitted model.



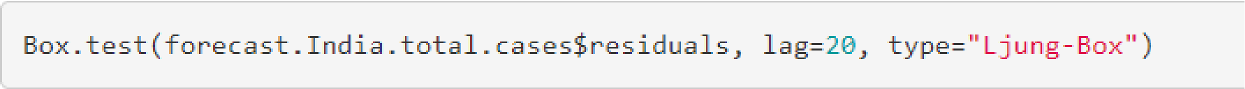
```
fitted_model<-HoltWinters(ts.info_cov_india1)
plot(fitted_model,main="fitting a model to the daily cases")
```



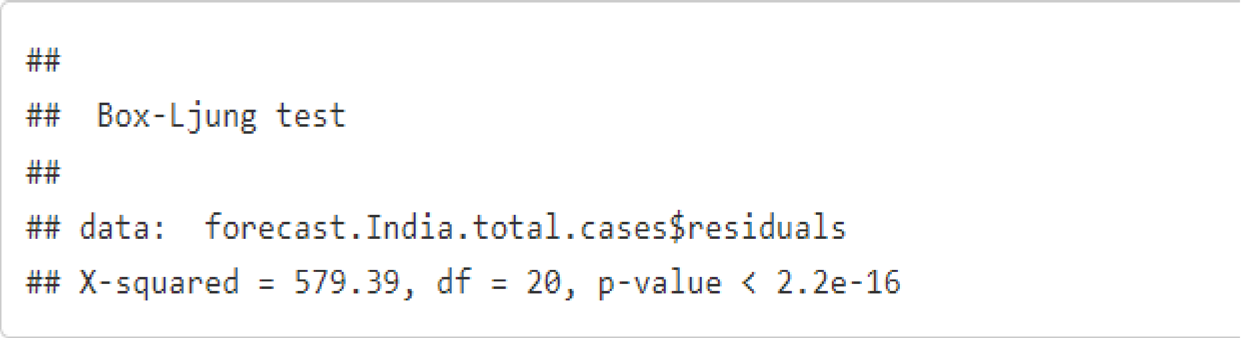
Forecasting using the fitted model:



```
#residual plots
forecast.India.total.cases<-forecast(fitted_model,10)
acf(na.omit(resid(forecast.India.total.cases)), lag.max=20)
```



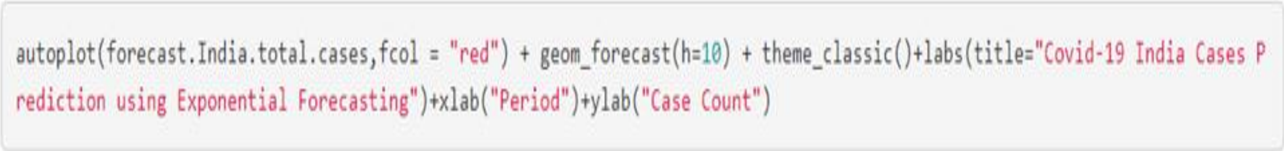
```
Box.test(forecast.India.total.cases$residuals, lag=20, type="Ljung-Box")
```



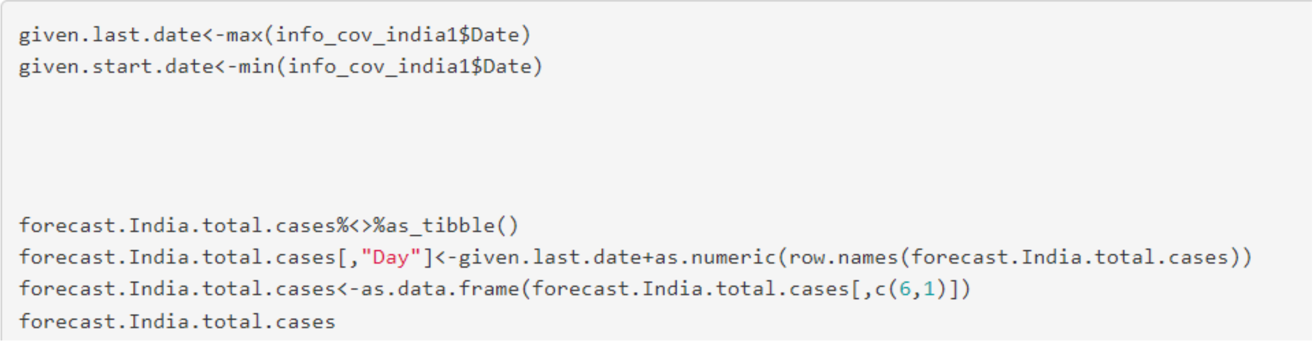
```
##
##  Box-Ljung test
##
## data:  forecast.India.total.cases$residuals
## X-squared = 579.39, df = 20, p-value < 2.2e-16
```

The small p-value tells that the dataset is non-stationary. The non-stationary dataset is the one in which a variable value changes for different time.

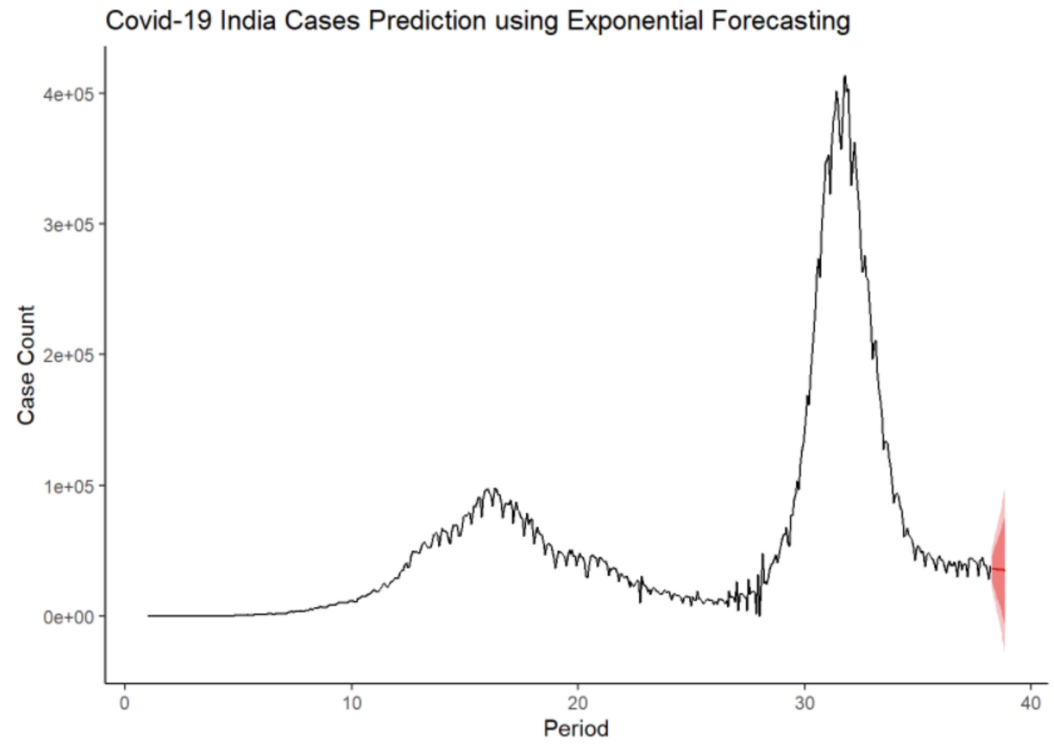
The plotting is done using the auto-plot to plot the forecasted results.

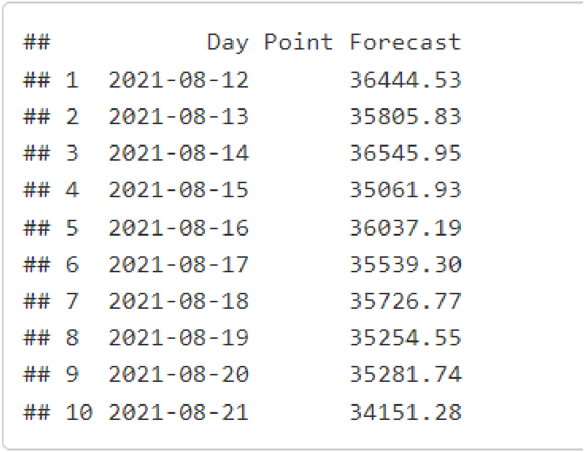


```
autoplot(forecast.India.total.cases,fcol = "red") + geom_forecast(h=10) + theme_classic()+labs(title="Covid-19 India Cases Prediction using Exponential Forecasting")+xlab("Period")+ylab("Case Count")
```



```
given.last.date<-max(info_cov_india1$Date)
given.start.date<-min(info_cov_india1$Date)




forecast.India.total.cases%<>%as_tibble()
forecast.India.total.cases[,"Day"]<-given.last.date+as.numeric(row.names(forecast.India.total.cases))
forecast.India.total.cases<-as.data.frame(forecast.India.total.cases[,c(6,1)])
forecast.India.total.cases
```



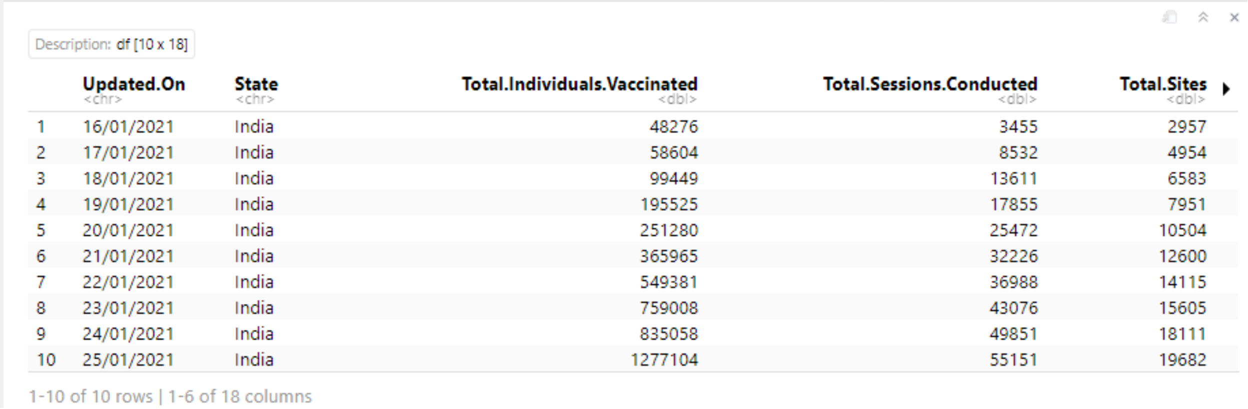


```
##          Day Point Forecast
## 1  2021-08-12       36444.53
## 2  2021-08-13       35805.83
## 3  2021-08-14       36545.95
## 4  2021-08-15       35061.93
## 5  2021-08-16       36037.19
## 6  2021-08-17       35539.30
## 7  2021-08-18       35726.77
## 8  2021-08-19       35254.55
## 9  2021-08-20       35281.74
## 10 2021-08-21       34151.28
```

The forecasting is done for 10 days in the month of August 2021. The accuracy of the model is greater than 90%.

Exploratory Data Analysis for Vaccines in India:

For the analyzing the vaccines administered in the India we have used a dataset named “covid_vaccines_statewise” from Kaggle. The dataset contains 18 variables that are Updated Date, State, Total Individuals Vaccinated, Total Sessions Conducted, First Dose Administered, Second Dose Administered etc... First the summary and structure of the dataset is explored and then the data preprocessing is done for the vaccine’s dataset.



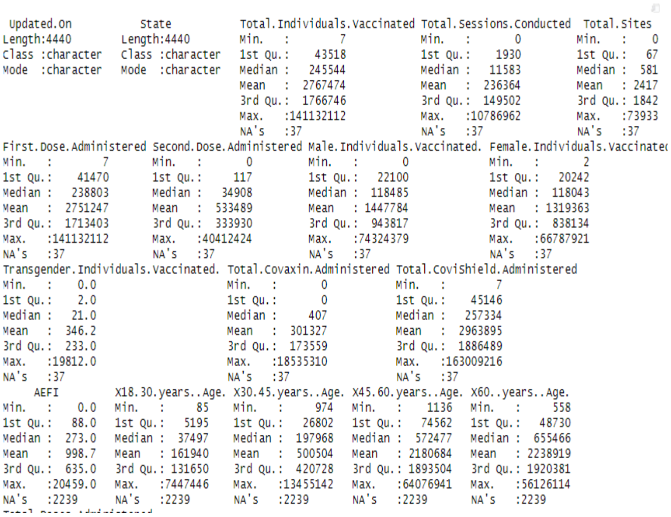
Description: df [10 x 18]

	Updated.On <chr>	State <chr>	Total.Individuals.Vaccinated <dbl>	Total.Sessions.Conducted <dbl>	Total.Sites <dbl>
1	16/01/2021	India	48276	3455	2957
2	17/01/2021	India	58604	8532	4954
3	18/01/2021	India	99449	13611	6583
4	19/01/2021	India	195525	17855	7951
5	20/01/2021	India	251280	25472	10504
6	21/01/2021	India	365965	32226	12600
7	22/01/2021	India	549381	36988	14115
8	23/01/2021	India	759008	43076	15605
9	24/01/2021	India	835058	49851	18111
10	25/01/2021	India	1277104	55151	19682

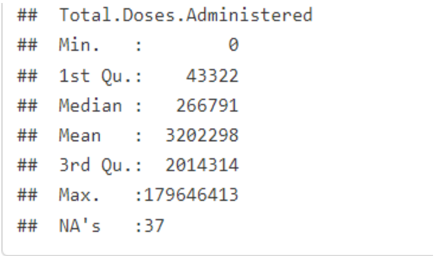
1-10 of 10 rows | 1-6 of 18 columns

Figure above shows the dataset

summary(vaccine)

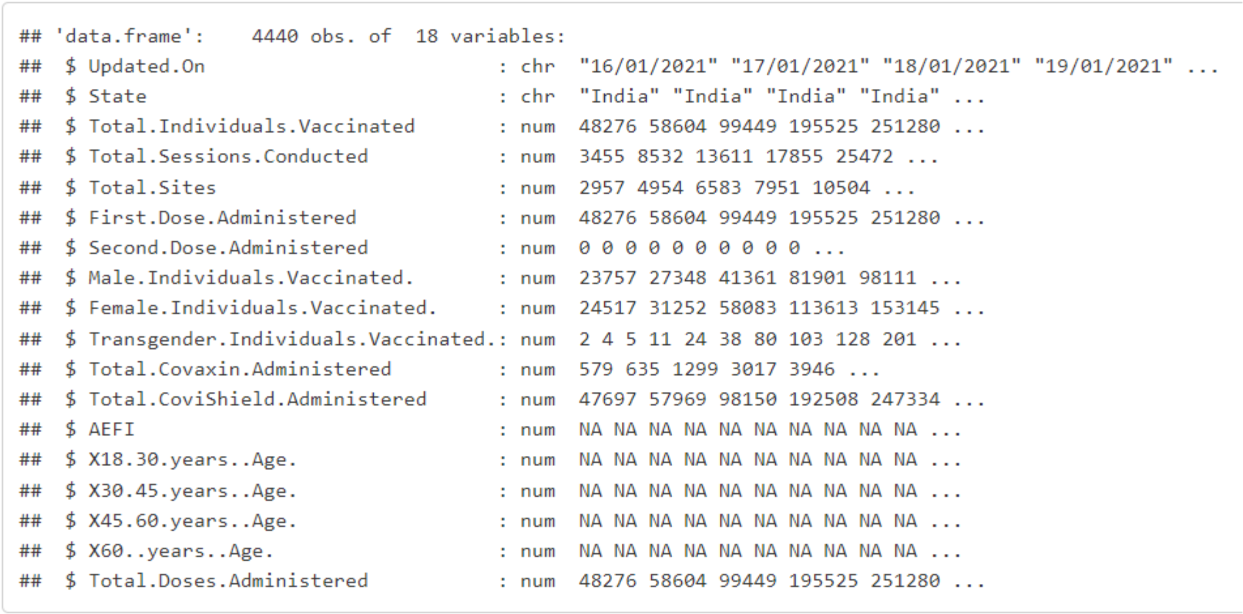


```
 Updated.On           State           Total.Individuals.Vaccinated Total.Sessions.Conducted  Total.Sites
Length:4440        Length:4440        Min.   :        7            Min.   :        0         Min.   :    0
Class :character   Class :character   1st Qu.:    43518            1st Qu.:     1930         1st Qu.:   67
Mode  :character   Mode  :character   Median :   245544            Median :    11583         Median :  581
                                      Mean   :  2767474            Mean   :   236364         Mean   : 2417
                                      3rd Qu.:  1766746            3rd Qu.:   149502         3rd Qu.: 1842
                                      Max.   :141132112            Max.   :10786962          Max.   :73933
                                      NA's   :37                   NA's   :37                NA's   :37
First.Dose.Administered Second.Dose.Administered Male.Individuals.Vaccinated. Female.Individuals.Vaccinate
Min.   :        7       Min.   :       0         Min.   :       0             Min.   :       2
1st Qu.:    41470       1st Qu.:     117         1st Qu.:   22100             1st Qu.:   20242
Median :   238803       Median :   34908         Median :  118485             Median :  118043
Mean   :  2751247       Mean   :  533489         Mean   : 1447784             Mean   : 1319363
3rd Qu.:  1713403       3rd Qu.:  333930         3rd Qu.:  943817             3rd Qu.:  838134
Max.   :141132112       Max.   :40412424         Max.   :74324379             Max.   :66787921
NA's   :37              NA's   :37               NA's   :37                   NA's   :37
Transgender.Individuals.Vaccinated. Total.Covaxin.Administered Total.CoviShield.Administered
Min.   :    0.0                     Min.   :       0           Min.   :        7
1st Qu.:    2.0                     1st Qu.:       0           1st Qu.:    45146
Median :   21.0                     Median :     407           Median :   257334
Mean   :  346.2                     Mean   :  301327           Mean   :  2963895
3rd Qu.:  233.0                     3rd Qu.:  173559           3rd Qu.:  1886489
Max.   :19812.0                     Max.   :18535310           Max.   :163009216
NA's   :37                          NA's   :37                 NA's   :37
     AEFI          X18.30.years..Age. X30.45.years..Age. X45.60.years..Age. X60..years..Age.
Min.   :    0.0   Min.   :     85    Min.   :     974   Min.   :    1136   Min.   :     558
1st Qu.:   88.0   1st Qu.:   5195    1st Qu.:   26802   1st Qu.:   74562   1st Qu.:   48730
Median :  273.0   Median :  37497    Median :  197968   Median :  572477   Median :  655466
Mean   :  998.7   Mean   : 161940    Mean   :  500504   Mean   : 2180684   Mean   : 2238919
3rd Qu.:  635.0   3rd Qu.: 131650    3rd Qu.:  420728   3rd Qu.: 1893504   3rd Qu.: 1920381
Max.   :20459.0   Max.   :7447446    Max.   :13455142   Max.   :64076941   Max.   :56126114
NA's   :2239      NA's   :2239       NA's   :2239       NA's   :2239       NA's   :2239
Total.Doses.Administered
```



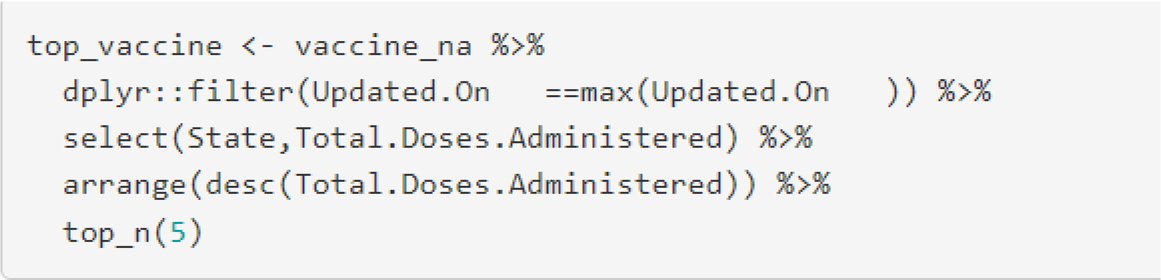
```
##  Total.Doses.Administered
##  Min.   :        0
##  1st Qu.:    43322
##  Median :   266791
##  Mean   :  3202298
##  3rd Qu.:  2014314
##  Max.   :179646413
##  NA's   :37
```

str(vaccine)

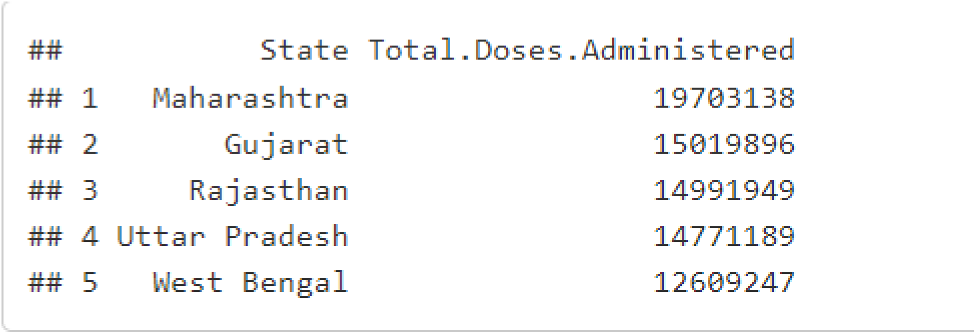


```
## 'data.frame':    4440 obs. of  18 variables:
##  $ Updated.On                         : chr  "16/01/2021" "17/01/2021" "18/01/2021" "19/01/2021" ...
##  $ State                              : chr  "India" "India" "India" "India" ...
##  $ Total.Individuals.Vaccinated       : num  48276 58604 99449 195525 251280 ...
##  $ Total.Sessions.Conducted           : num  3455 8532 13611 17855 25472 ...
##  $ Total.Sites                        : num  2957 4954 6583 7951 10504 ...
##  $ First.Dose.Administered            : num  48276 58604 99449 195525 251280 ...
##  $ Second.Dose.Administered           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Male.Individuals.Vaccinated.       : num  23757 27348 41361 81901 98111 ...
##  $ Female.Individuals.Vaccinated.     : num  24517 31252 58083 113613 153145 ...
##  $ Transgender.Individuals.Vaccinated.: num  2 4 5 11 24 38 80 103 128 201 ...
##  $ Total.Covaxin.Administered         : num  579 635 1299 3017 3946 ...
##  $ Total.CoviShield.Administered      : num  47697 57969 98150 192508 247334 ...
##  $ AEFI                               : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ X18.30.years..Age.                 : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ X30.45.years..Age.                 : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ X45.60.years..Age.                 : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ X60..years..Age.                   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Total.Doses.Administered           : num  48276 58604 99449 195525 251280 ...
```

Top states which are vaccinated:



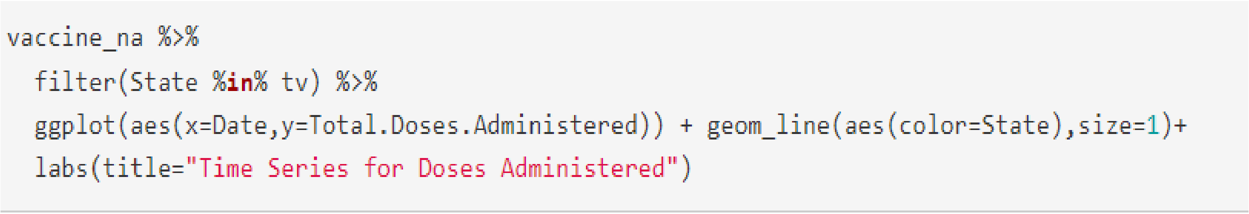
```
top_vaccine <- vaccine_na %>%
  dplyr::filter(Updated.On    ==max(Updated.On    )) %>%
  select(State,Total.Doses.Administered) %>%
  arrange(desc(Total.Doses.Administered)) %>%
  top_n(5)
```



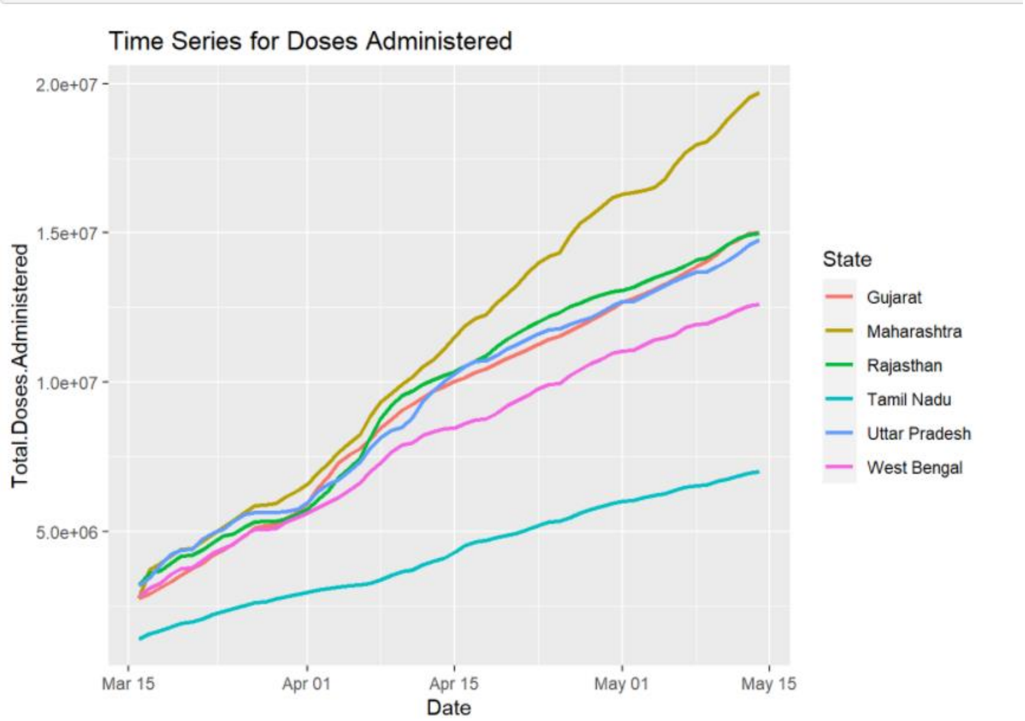
```
##             State Total.Doses.Administered
## 1     Maharashtra                 19703138
## 2         Gujarat                 15019896
## 3       Rajasthan                 14991949
## 4   Uttar Pradesh                 14771189
## 5     West Bengal                 12609247
```

From the above we can observe that Maharashtra state has highest number of doses administered followed by Gujarat, Rajasthan, Uttar Pradesh and West Bengal.

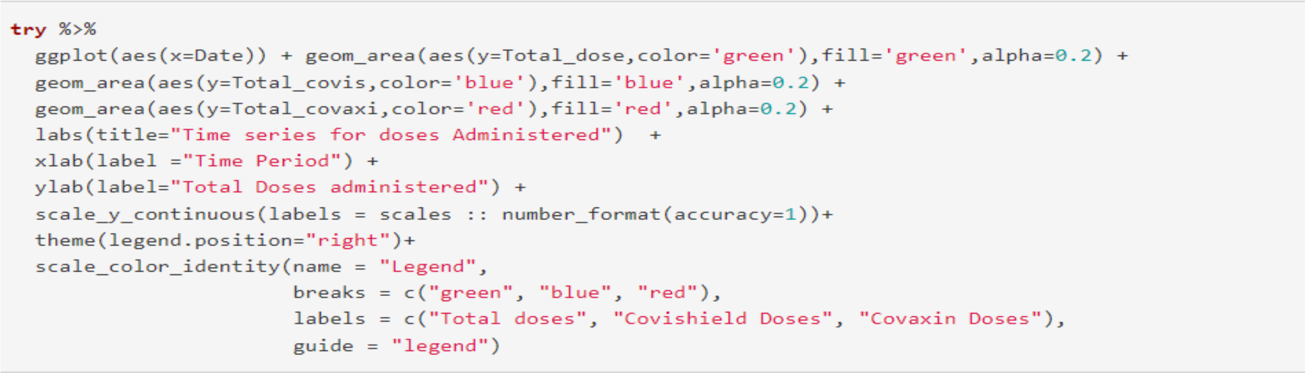
Time Series for Doses Administered:



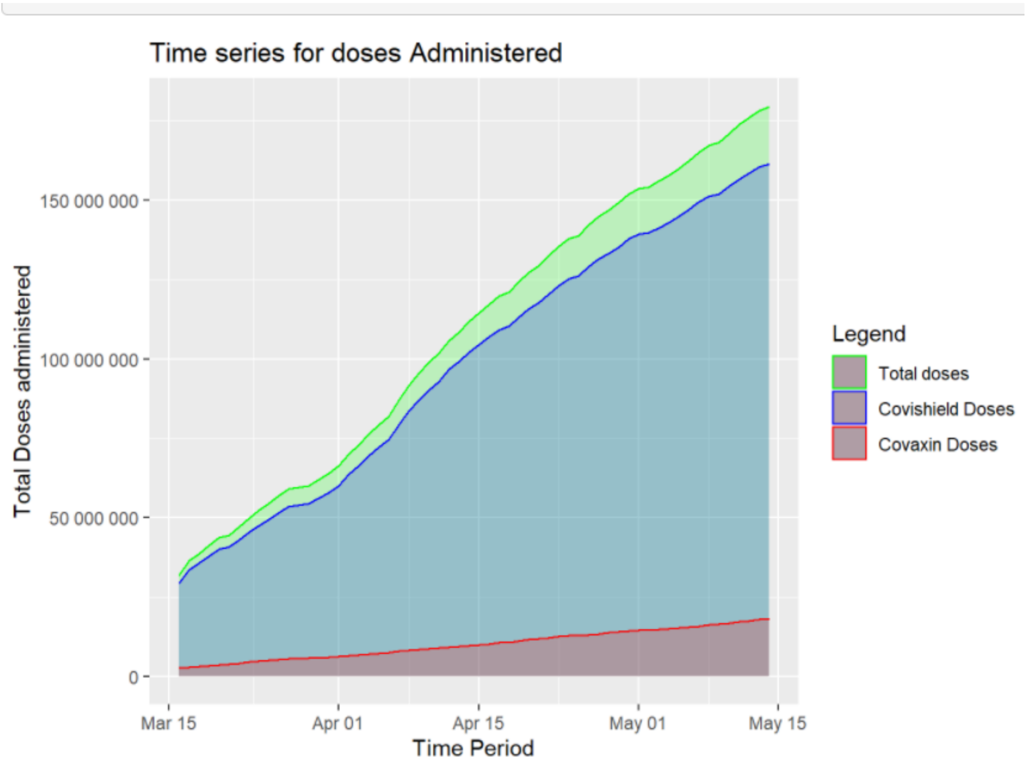
```
vaccine_na %>%
  filter(State %in% tv) %>%
  ggplot(aes(x=Date,y=Total.Doses.Administered)) + geom_line(aes(color=State),size=1)+
  labs(title="Time Series for Doses Administered")
```



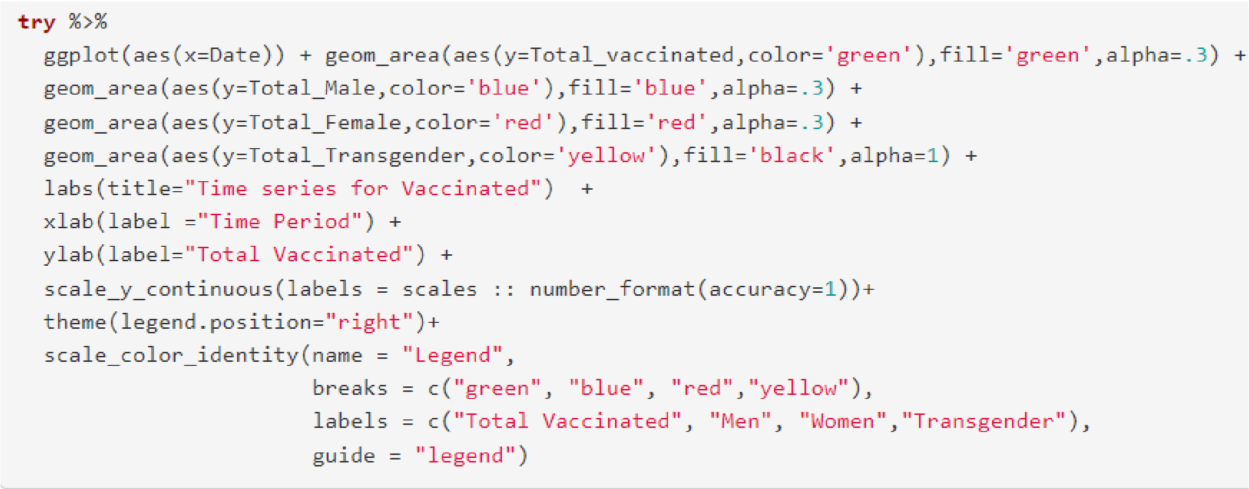
From the above graph we can infer that Tamil Nadu state has a smaller number of Doses Administered compared to other states.



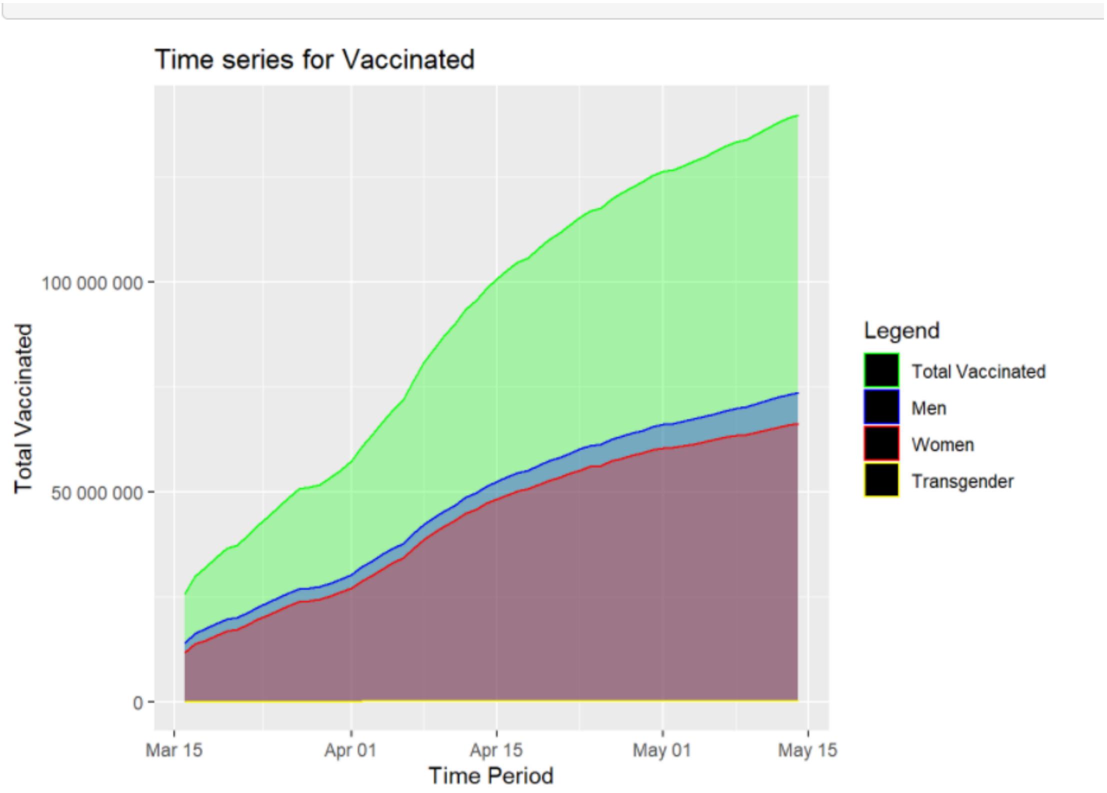
```
try %>%
  ggplot(aes(x=Date)) + geom_area(aes(y=Total_dose,color='green'),fill='green',alpha=0.2) +
  geom_area(aes(y=Total_covis,color='blue'),fill='blue',alpha=0.2) +
  geom_area(aes(y=Total_covaxi,color='red'),fill='red',alpha=0.2) +
  labs(title="Time series for doses Administered")  +
  xlab(label ="Time Period") +
  ylab(label="Total Doses administered") +
  scale_y_continuous(labels = scales :: number_format(accuracy=1))+
  theme(legend.position="right")+
  scale_color_identity(name = "Legend",
                     breaks = c("green", "blue", "red"),
                     labels = c("Total doses", "Covishield Doses", "Covaxin Doses"),
                     guide = "legend")
```



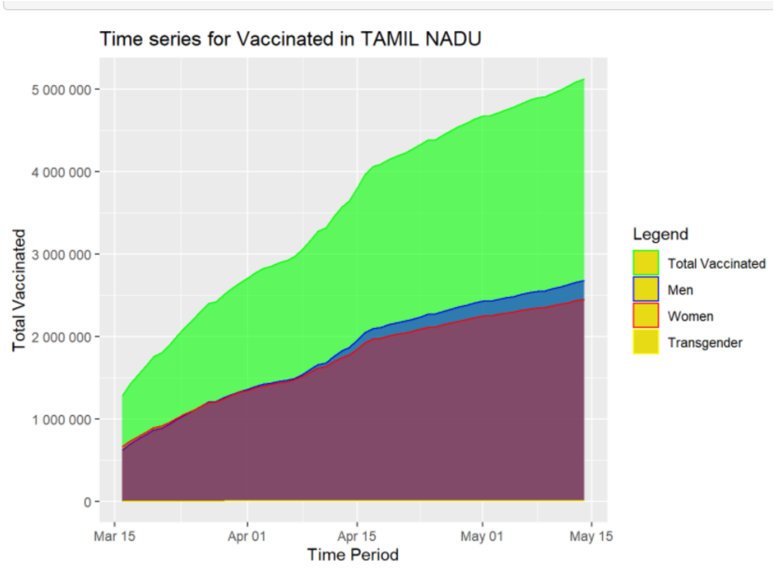
From March 15, 2021 to May 15,2021 the total number of doses administered are greater than 150000000. The Covid-shield doses administered are greater than the covaxin doses in India.



```
try %>%
  ggplot(aes(x=Date)) + geom_area(aes(y=Total_vaccinated,color='green'),fill='green',alpha=.3) +
  geom_area(aes(y=Total_Male,color='blue'),fill='blue',alpha=.3) +
  geom_area(aes(y=Total_Female,color='red'),fill='red',alpha=.3) +
  geom_area(aes(y=Total_Transgender,color='yellow'),fill='black',alpha=1) +
  labs(title="Time series for Vaccinated")  +
  xlab(label ="Time Period") +
  ylab(label="Total Vaccinated") +
  scale_y_continuous(labels = scales :: number_format(accuracy=1))+
  theme(legend.position="right")+
  scale_color_identity(name = "Legend",
                       breaks = c("green", "blue", "red","yellow"),
                       labels = c("Total Vaccinated", "Men", "Women","Transgender"),
                       guide = "legend")
```

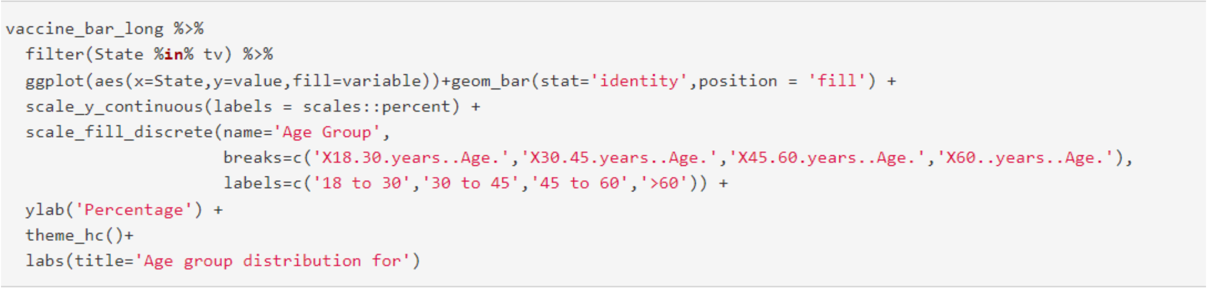


In this time period we can observe that the number of men vaccinated are quite greater than the number of women vaccinated in India. The number of transgenders vaccinated are very less.

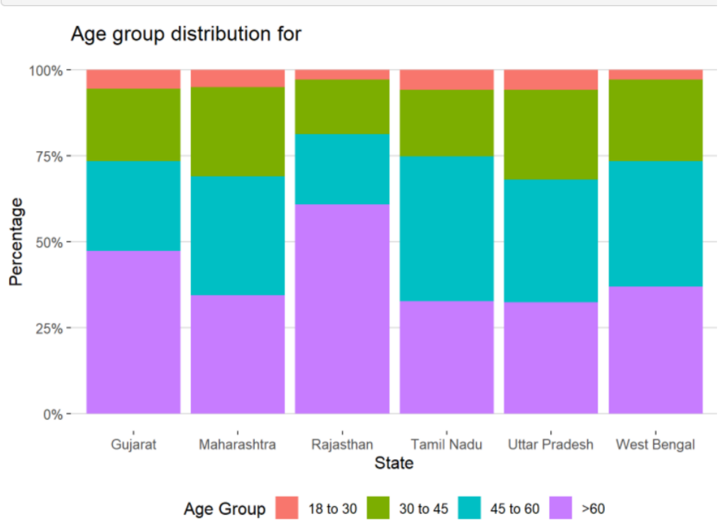


For the Time Series for the vaccinated people in TAMIL NADU:

From the above graph we can observe that the total does administer are greater than 5000000. The number of men and women vaccinated are almost equal.



```
vaccine_bar_long %>%
  filter(State %in% tv) %>%
  ggplot(aes(x=State,y=value,fill=variable))+geom_bar(stat='identity',position = 'fill') +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_discrete(name='Age Group',
                      breaks=c('X18.30.years..Age.','X30.45.years..Age.','X45.60.years..Age.','X60..years..Age.'),
                      labels=c('18 to 30','30 to 45','45 to 60','>60')) +
  ylab('Percentage') +
  theme_hc()+
  labs(title='Age group distribution for')
```



CONCLUSION:

Through this project, the analysis on COVID-19 data has been performed successfully. The analysis on this pandemic spread has been done and compared between different countries. The analysis of confirmed cases, active cases, recovered cases and deaths are done separately to give a clear look on how the virus is spreading, which countries are getting affected mostly and how different countries are recovering. A separate analysis on cases of INDIA has been done and predictions of different cases both around the world and INDIA has been done. The analysis on the vaccines administered in INDIA is also performed successfully.