# Final Project

Dhruvkumar Shah ID :-300318529

22/04/2021

## Question 1

### Introduction

In the following report I will be discussing evolution of Covid 19 infection and death rates for USA and compare the curve with China, Italy,France, Germany, South Korea, UK.

### Analysis

We will start the analysis by looking at the evolution of Covid cases overtime.

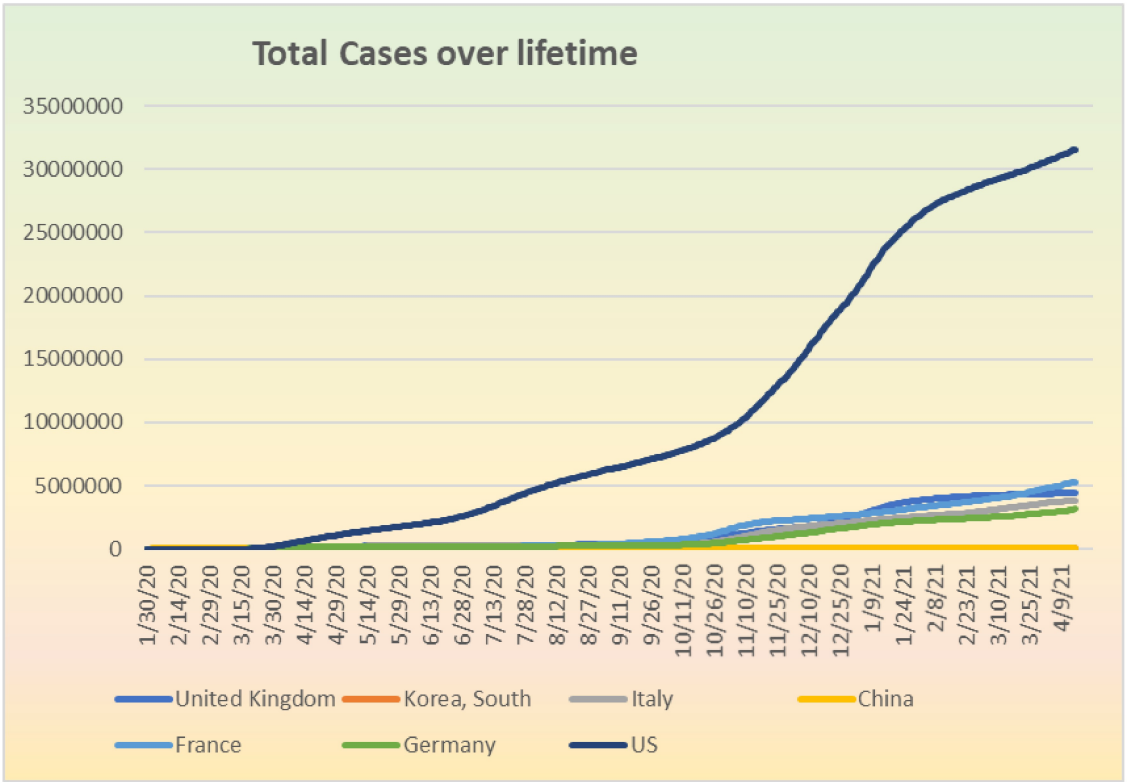



Figure 1:

There appears to be steep increase in the number of cases in USA around 30th of March 2020.

Looking at the death rate overtime.

A lot of fluctuations can be observed in death rate from 10th of March to 17th of June of different countries.

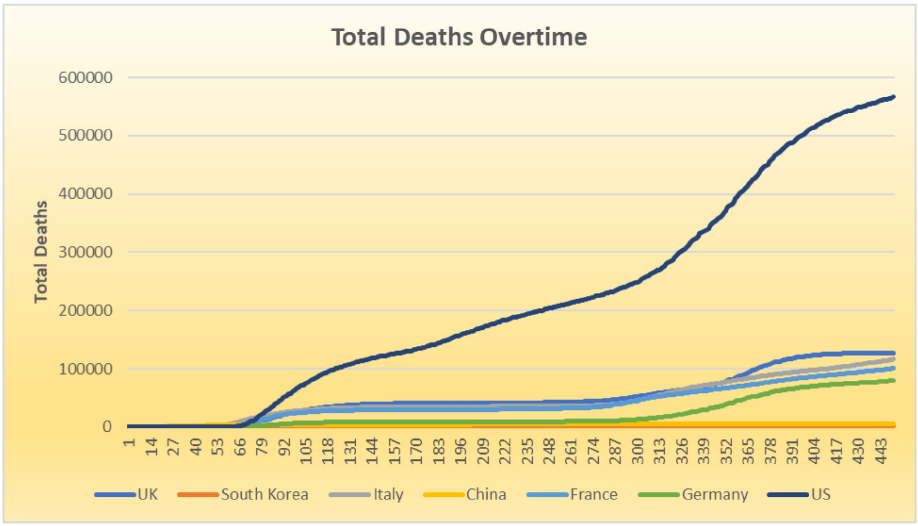



Figure 2:

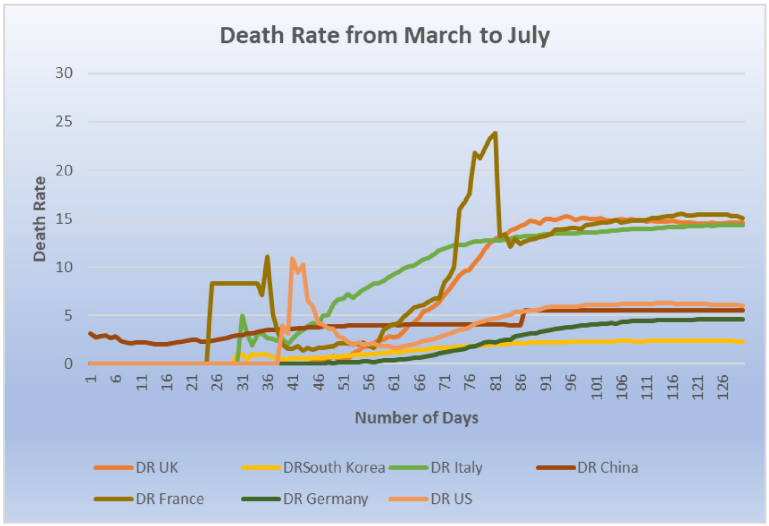



Figure 3:

Sudden surge in Death rate is appearing in France, US and Italy.

France reached the highest Death Rate of 22-23 around the end of May 2020, which fell down to about 13-15.

France, UK, Italy have death rate of 13-15 while China, USA, Germany and South Korea have Death rate less than 5.

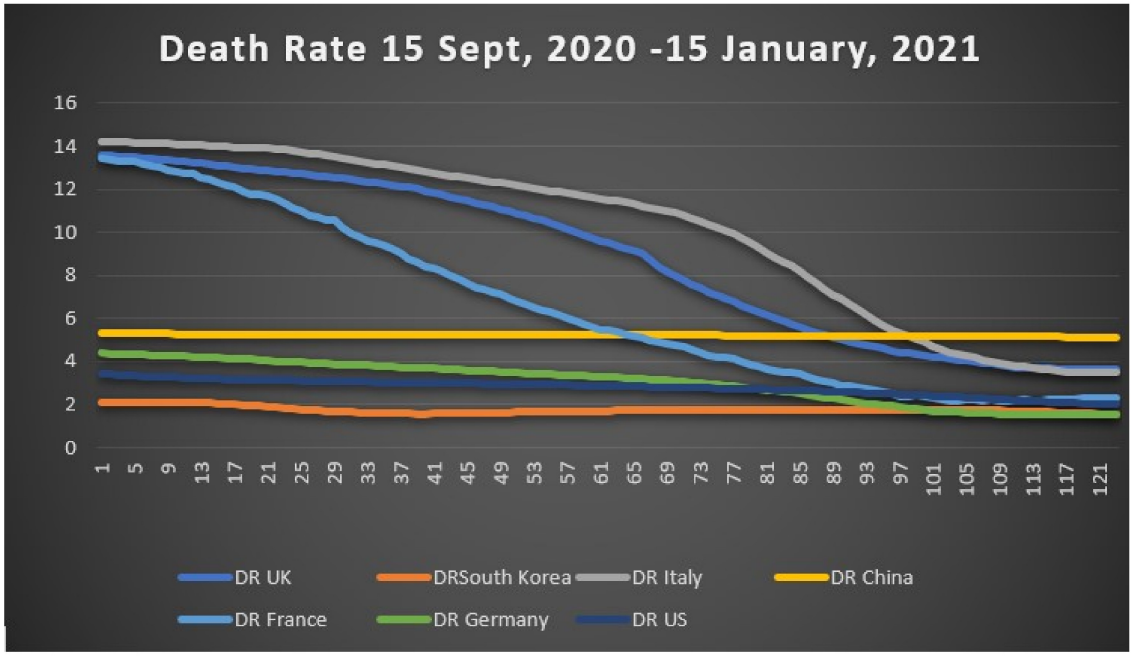



Figure 4:

Decrease in Death Rates from 15th September to 15th January for Italy, UK, France to less than 5. Meanwhile death rate remained almost constand for other countries.

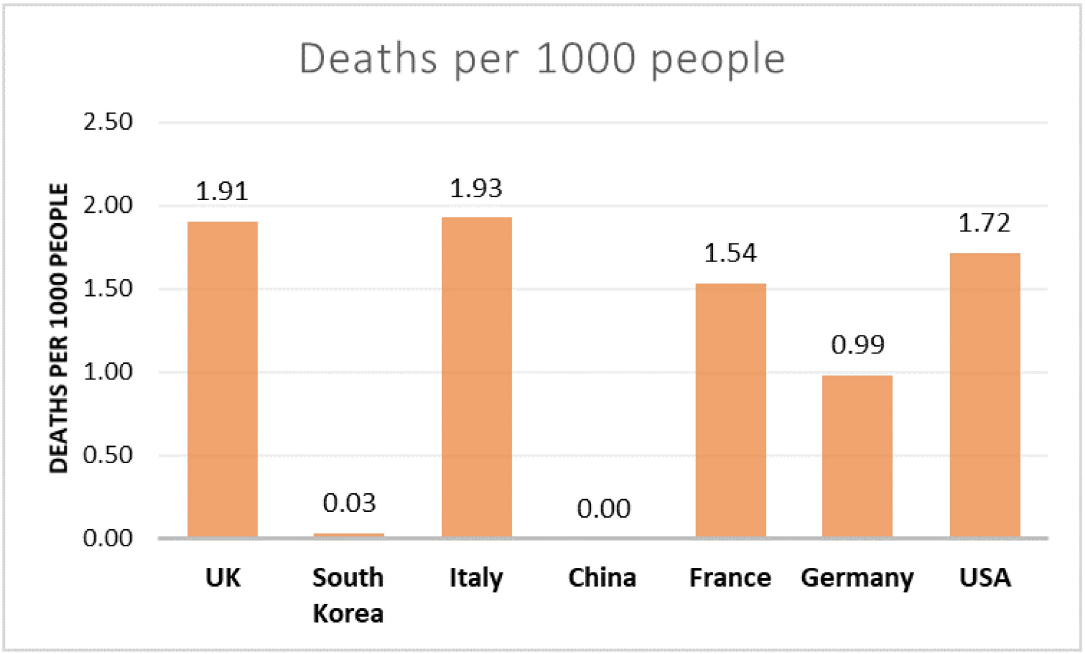



Figure 5: Deaths per 1000 as of 16 April 2021

In the following bar chart the death rates of all countries can be observed.

## Conclusions

While comparing USA to other countries the death rates in USA did not reach as high as countries like Italy, UK and France

The huge number of deaths in USA are a result of increased high Covid cases.

Looks like except China and South Korea, which have very low death rate and Germany which has death rate of 0.99, USA is in similar situation as other European countries.

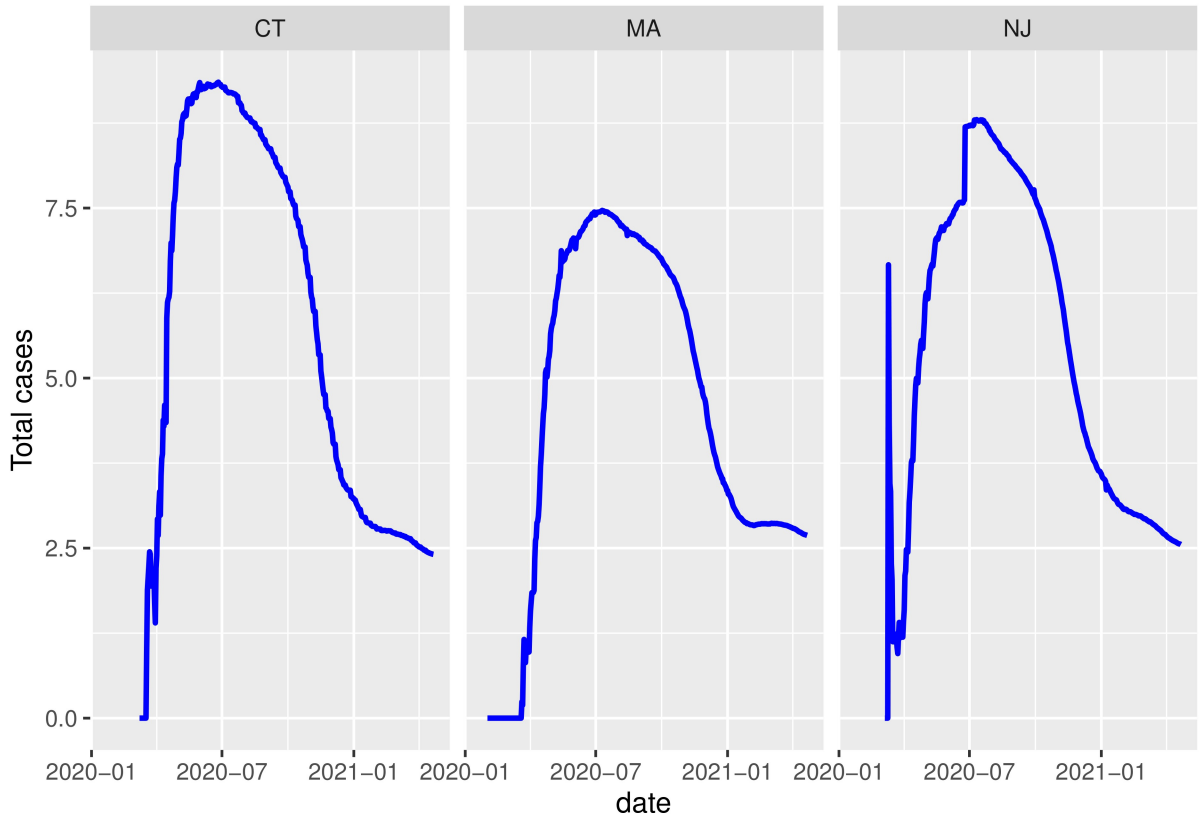
# Question 2

```
## [1] 0
```

States with maximum Death rate

```
## # A tibble: 3 x 2
##   state deathrate
##   <chr>     <dbl>
## 1 MA         2.69
## 2 NJ         2.55
## 3 CT         2.41
```

Massachusetts, New Jersey and Connecticut are the states with maximum death rate




Plotting the curve overtime, it can be observed that each of these states reached their peaks around July where CT had a death rate of almost 10 while MA had about 7.5.

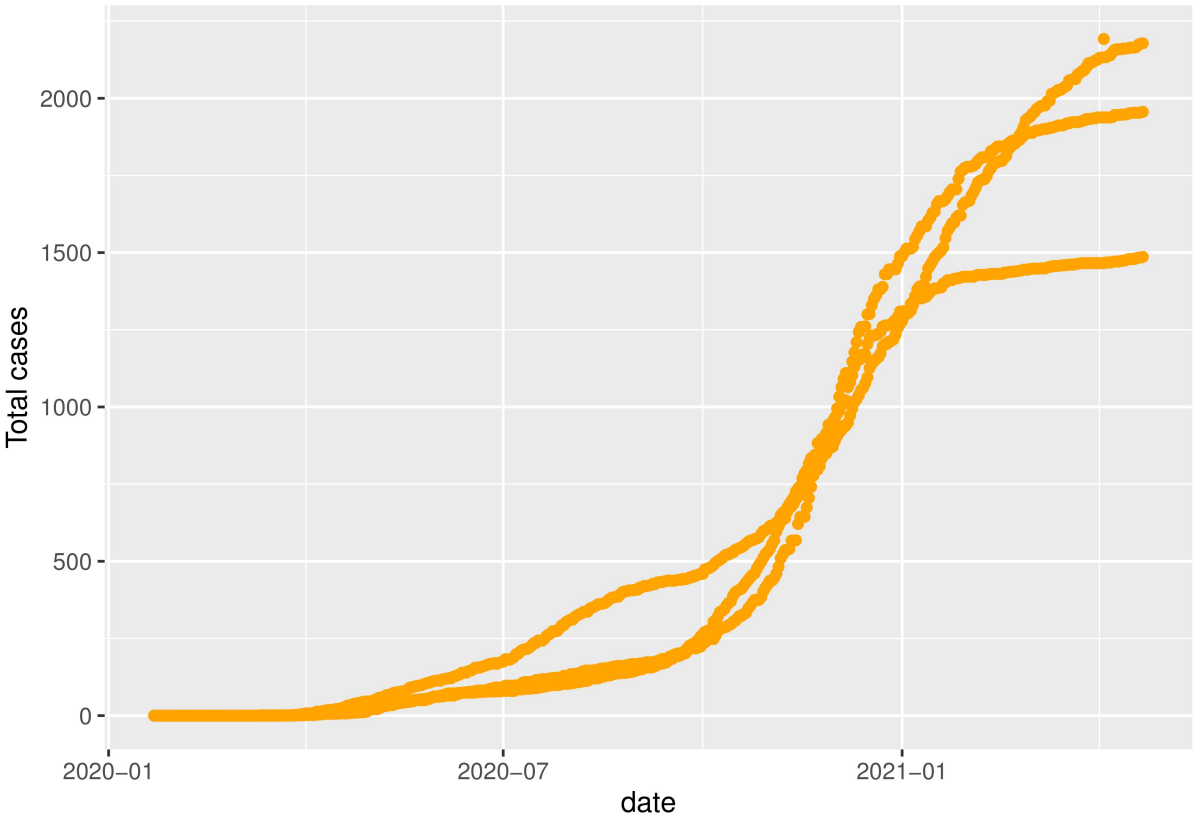
Also unique spike can be observed during the start of the pandemic in NJ.

States with maximum Infection Rate rate

```
## # A tibble: 3 x 2
##   state infectionrate
##   <chr>         <dbl>
## 1 ND             15.8
## 2 SD             14.9
```

```
## 3 UT              14.3
```

North Dakota State has maximum infection rate followed by South Dakota(14.92) and Utah



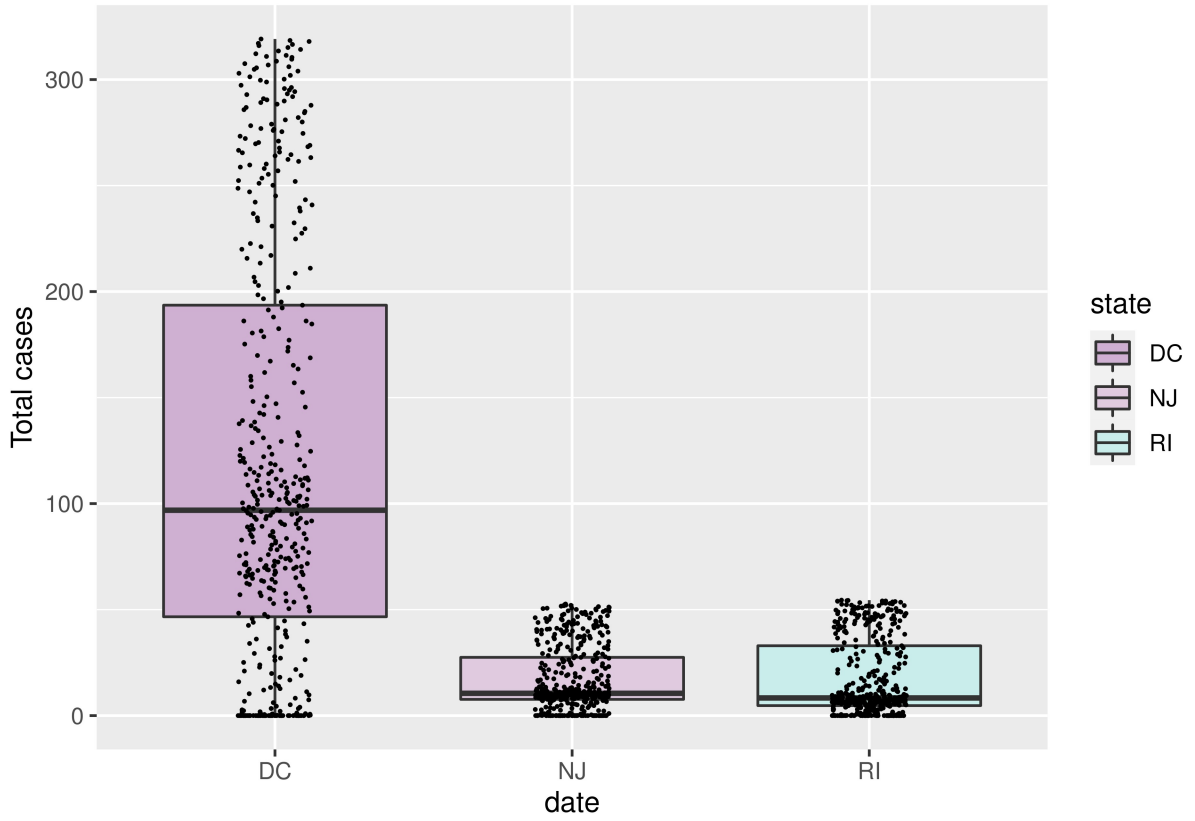

Though the total cases are not as high in states like ND and SD , the infection rate is amoung the highest which might suggest lack of awareness and self

Maximum cases per square KM

```
## # A tibble: 3 x 2
##   state cases_per_sq_km
##   <chr>           <dbl>
## 1 DC              319.
## 2 RI               54.5
## 3 NJ               52.6
```

DC has maximum cases per sq km followed by New Jersey and Rhode ISland.

These states also have maximum population density which makes high cases per sq km more understandable.




DC has extremely high cases per square km compared to the second highest state as well, which looks like needs further investigation.

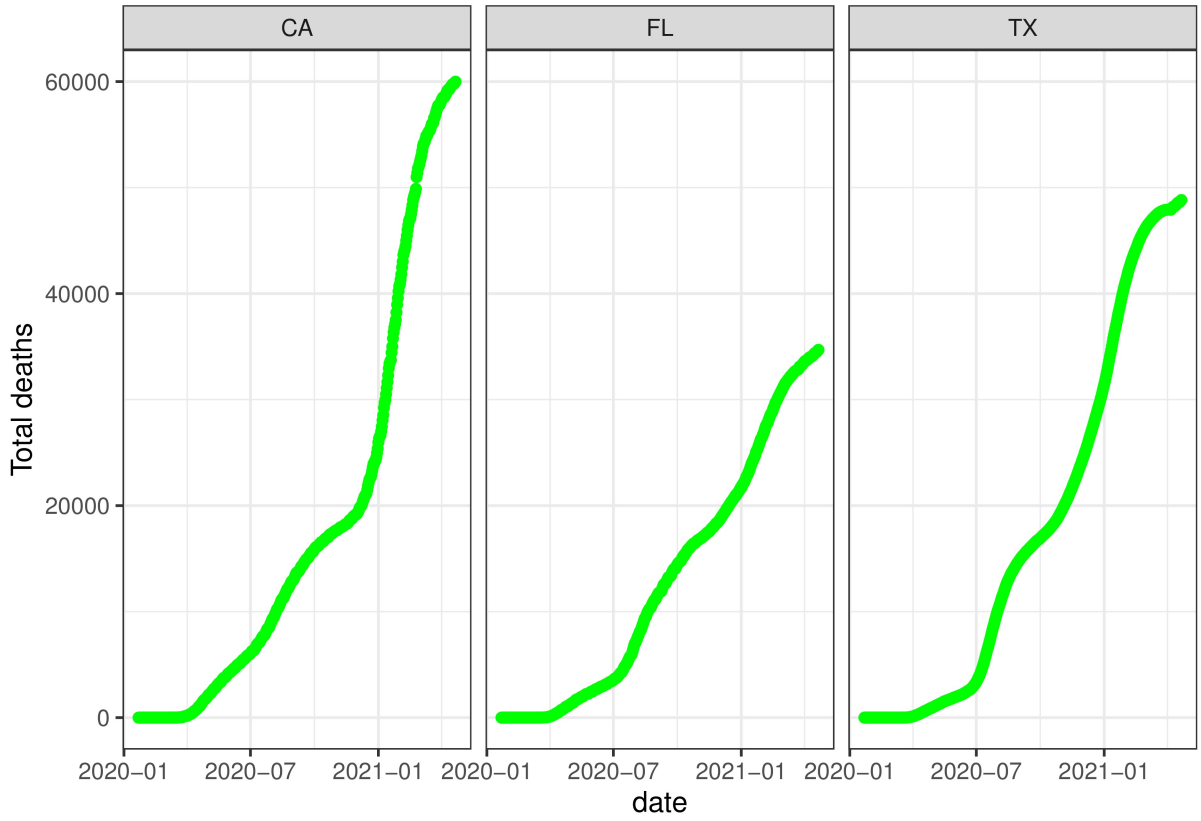
New Jersey also has second highest Death rate which looks like an area of concern.

Maximum total cases

```
## # A tibble: 3 x 2
##   state tot_cases
##   <chr>     <dbl>
## 1 CA      3624838
## 2 TX      2857017
## 3 FL      2149932
```

Maximum total deaths

```
## # A tibble: 3 x 2
##   state tot_death
##   <chr>     <dbl>
## 1 CA        59992
## 2 TX        48828
## 3 FL        34696
```




Looking at total cases and total deaths California has highest deaths and total cases followed by Texas and Florida. Not surprisingly these States also have the highest population.

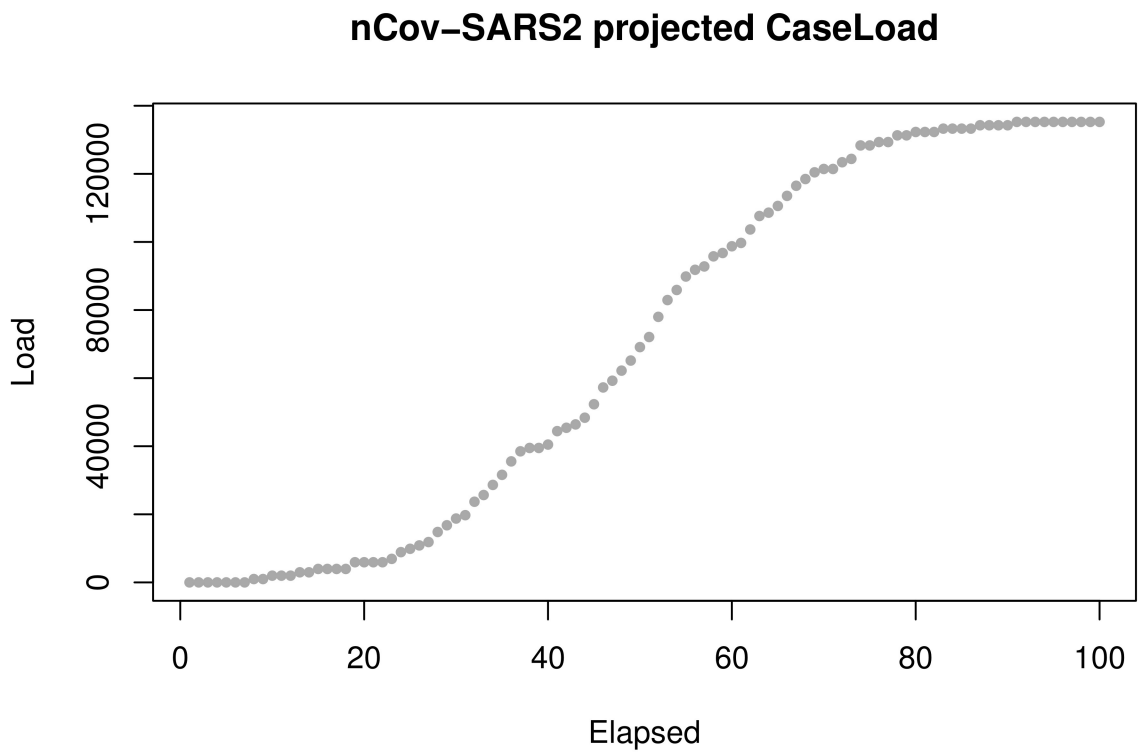
## Conclusion

The states with high population density tend to have high infection rate and high density of cases.

## Question 3

### Reviewing the client's code

```
a<-5.32
b<-49.3
c<-51.6
d<- 987.32
num_days = 100
days <- 1:num_days
lambda_sim <- exp(-a*((days-b)**2/c**2))
W <- 987.32*rpois(num_days,pi*lambda_sim)
plot(cumsum(W), xlab = "Elapsed", ylab = "Load",
main = 'nCov-SARS2 projected CaseLoad', pch = 16, cex = 0.75, col = "darkgrey")
```




The following code tries to predict Covid 19 case load for first 100 days from the the first case.

The simulation uses Poisson distribution to predict the case load.

Lambda is generated using exponential equation.

The input varaible is number of day while the output is the cummulative case load for those days.

*Effect of different variables on the curve*

Value of $a$ impacts the steepness of the curve. The higher value of a we use, less steeper will be the rise in case load.

As the value of $b$ increases the later will the caseload start to show exponential increase.

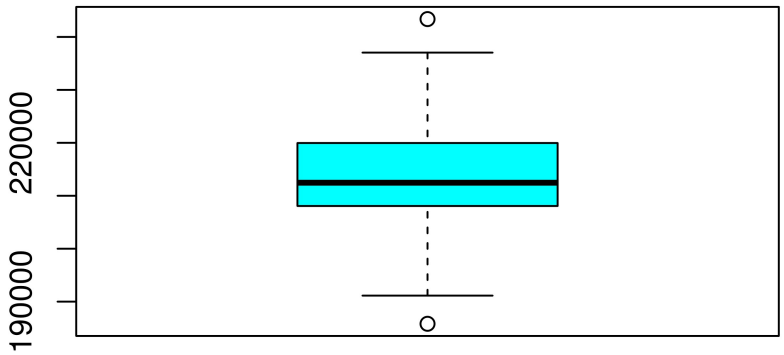
Value of $c$ impacts the exponential nature of the curve. As value of $c$ increases the the curve becomes less exponential and more linear.

Value of $d$ directly affects the value of case load and increasing it will impact the total case load.

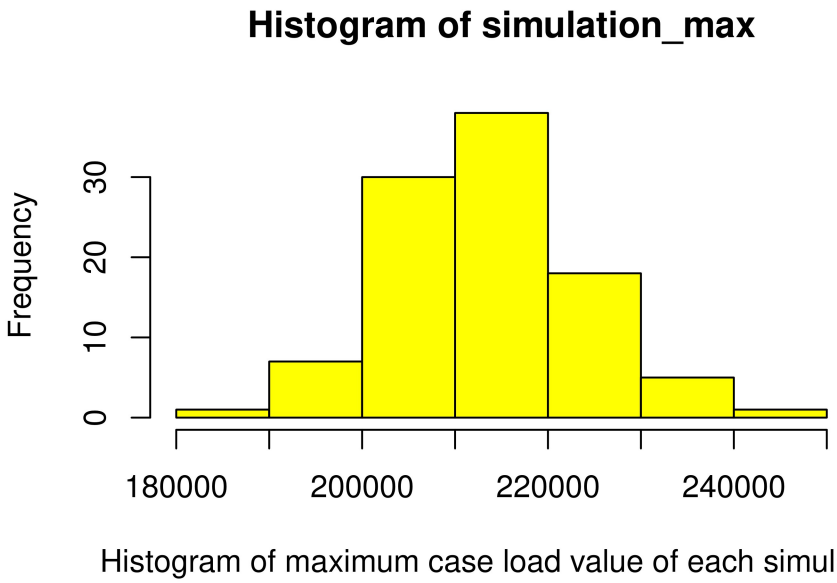
I tried using normal distribution in place of poisson distribution and got the following curve.

Normal distribution is not useful since it spits out some negative values of W which but value of covid cases cannot be negative.

Running this simulation 100 times and analyzing the maximum values each time, can give an idea about the maximum possible values of case load




Boxplot of maximum case load value of each simulation




Histogram of maximum case load value of each simulation

Statistical summary of the simulation_max.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  185819  208140  212463  213452  219947  243374
```

The following simulation suggests that we can expect a peak value of around 235000 cases.

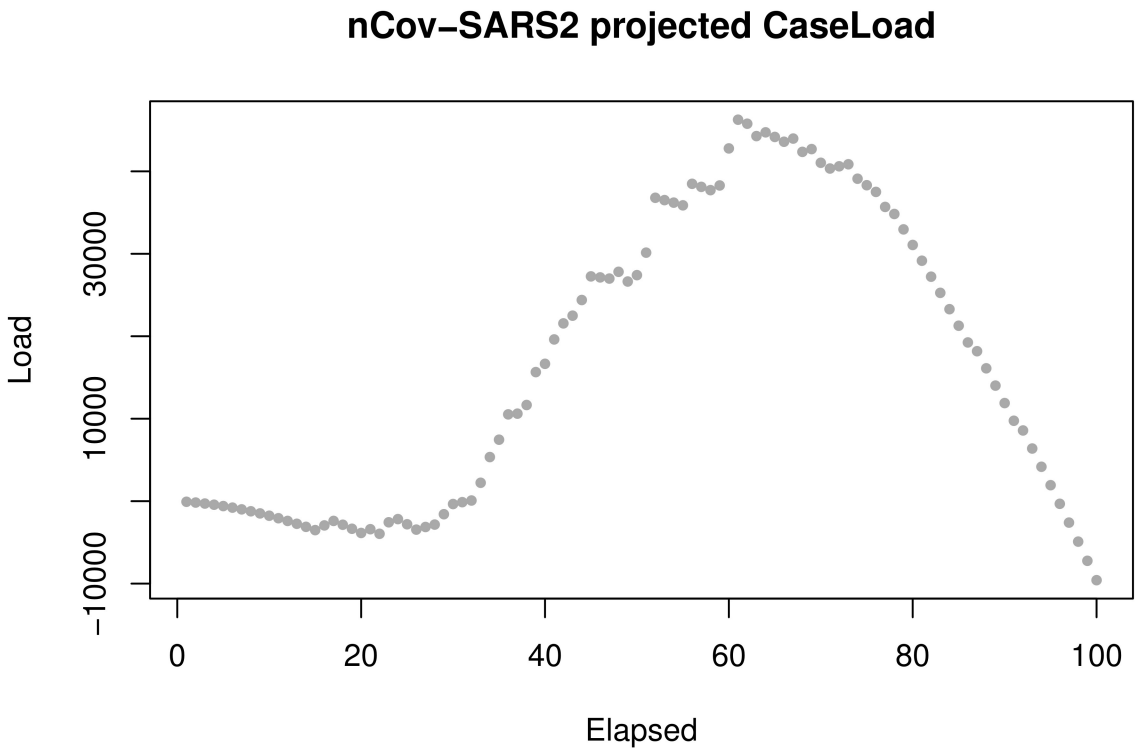
### Observed drawbacks and suggestions.

This simulation does not seem to take into account the factor of recoveries into account therefore this model might overpredict the maximum actual cases at a given time.

Below is an attempt to simulate the case load with entering a decaying factor in the simulation.

Decay looks linear and is dependent on the the number of days as variable.

```
num_days = 100
days <- 1:num_days
lambda_sim <- exp(-5.32*((days-49.3)**2/51.6**2))

W <- 987.32*rpois(num_days,pi*lambda_sim) -(23*days + 50)
plot(cumsum(W), xlab = "Elapsed", ylab = "Load",
main = 'nCov-SARS2 projected CaseLoad', pch = 16, cex = 0.75, col = "darkgrey")
```




Above is an attempt to simulate the case load with entering a decaying factor in the simulation. Decay looks linear and is dependent on the the number of days as variable.

## Code Appendix

```
# Question 2

#Loading required libraries

library(readr)
library(plotly)
library(ggplot2)
library(dplyr)
library(readxl)

#basic data cleaning
df <- read_csv("covid_data.csv")
df$date <-as.Date(df$date,format= "%d-%m-%Y")


sum(is.na(df))
```

```
## [1] 0
```

```
dftest<-df %>% group_by(date) %>% summarise(new_case  = sum(new_case),
                                            tot_case = sum(tot_cases),
                                            new_death = sum(new_death),
                                            tot_death = sum(tot_death))

df_population <- read_excel("PerfCarSales.xlsx")
a<-0
#getting population values from perfcar sales
df$population <- NA
for (a in 1:nrow(df)){
  df$population[a] <- df_population[df_population$Abb==df$state[a],4]
}

#Deriving useful columns using available data

df$area <- as.numeric(df$population)/as.numeric(df$Population_per_sqyare_km)

df$cases_per_sq_km <-as.numeric(df$tot_cases)/as.numeric(df$area)



df$deathrate<-as.numeric(df$tot_death)*100/as.numeric(df$tot_cases)
df$recoveryrate<- 100-df$deathrate

# Subsetting data of the latest day from the data frame
df_latest <- df[df$date == "2021-04-22",]

#states with maximum death rate
max_deathrate <-c("HI","VT","ME")
min_deathrate <- c("ND","SD","UT")
max_pop_density <- c("AK","WY","MT")
df_latest$infectionrate <-as.numeric(df_latest$tot_cases)/as.numeric(df_latest$population)*100
```

```
head(df_latest[order(-df_latest$deathrate),c(3,4,6,9,10,11,12,13,14)],3)
```

```
## # A tibble: 3 x 9
##   state tot_cases tot_death Population_per_~ population   area cases_per_sq_km
##   <chr>     <dbl>     <dbl>            <dbl> <list>      <dbl>           <dbl>
## 1 MA       646372     17376             332. <dbl [1]>  19746.            32.7
## 2 NJ       990580     25301             467. <dbl [1]>  18819.            52.6
## 3 CT       333732      8039             287. <dbl [1]>  12468.            26.8
## # ... with 2 more variables: deathrate <dbl>, recoveryrate <dbl>
```

```
head(df_latest[order(-df_latest$deathrate),c(3,4,6,9,10,11,12,13,14)],3)
```

```
## # A tibble: 3 x 9
##   state tot_cases tot_death Population_per_~ population   area cases_per_sq_km
##   <chr>     <dbl>     <dbl>            <dbl> <list>      <dbl>           <dbl>
## 1 MA       646372     17376             332. <dbl [1]>  19746.            32.7
## 2 NJ       990580     25301             467. <dbl [1]>  18819.            52.6
## 3 CT       333732      8039             287. <dbl [1]>  12468.            26.8
## # ... with 2 more variables: deathrate <dbl>, recoveryrate <dbl>
```

```
head(df_latest[order(-df_latest$infectionrate),],3)
```

```
## # A tibble: 3 x 15
##   Column1 date       state tot_cases new_case tot_death new_death Region
##     <dbl> <date>     <chr>     <dbl>    <dbl>     <dbl>     <dbl> <chr>
## 1   27376 2021-04-22 ND       106537      152      1486         2 North~
## 2   27398 2021-04-22 SD       121651      134      1956         2 North~
## 3   27401 2021-04-22 UT       394334      472      2178         1 West
## # ... with 7 more variables: Population_per_sqyare_km <dbl>, population <list>,
## #   area <dbl>, cases_per_sq_km <dbl>, deathrate <dbl>, recoveryrate <dbl>,
## #   infectionrate <dbl>
```

```
head(df_latest[order(-df_latest$cases_per_sq_km),],3)
```

```
## # A tibble: 3 x 15
##   Column1 date       state tot_cases new_case tot_death new_death Region
##     <dbl> <date>     <chr>     <dbl>    <dbl>     <dbl>     <dbl> <chr>
## 1   27410 2021-04-22 DC        47040       99      1098         1 South
## 2   27418 2021-04-22 RI       146028      381      2660         2 North~
## 3   27383 2021-04-22 NJ       990580     3230     25301        30 North~
## # ... with 7 more variables: Population_per_sqyare_km <dbl>, population <list>,
## #   area <dbl>, cases_per_sq_km <dbl>, deathrate <dbl>, recoveryrate <dbl>,
## #   infectionrate <dbl>
```

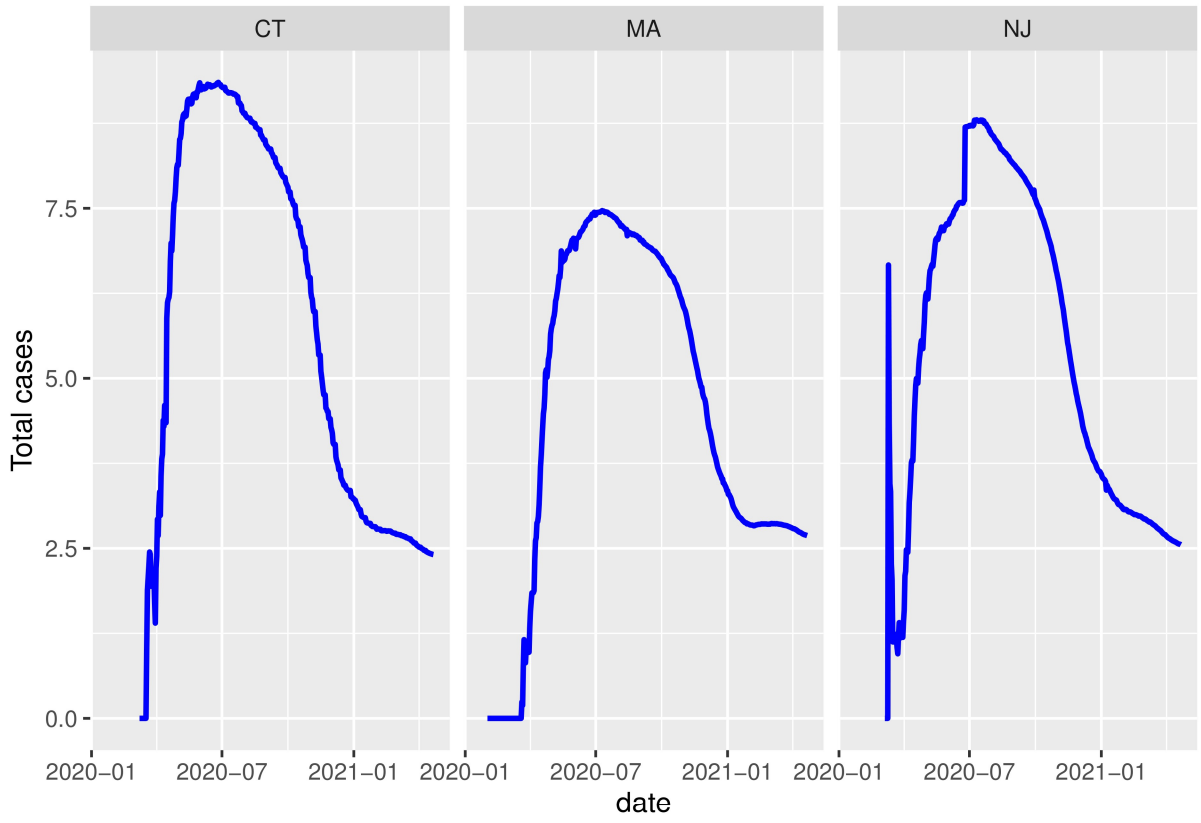
```
head(df_latest[order(-df_latest$tot_cases),],3)
```

```
## # A tibble: 3 x 15
##   Column1 date       state tot_cases new_case tot_death new_death Region
##     <dbl> <date>     <chr>     <dbl>    <dbl>     <dbl>     <dbl> <chr>
## 1   27368 2021-04-22 CA      3624838     2411     59992       102 West
## 2   27408 2021-04-22 TX      2857017     2990     48828        69 South
## 3   27391 2021-04-22 FL      2149932     6574     34696        80 South
## # ... with 7 more variables: Population_per_sqyare_km <dbl>, population <list>,
## #   area <dbl>, cases_per_sq_km <dbl>, deathrate <dbl>, recoveryrate <dbl>,
## #   infectionrate <dbl>
```

```r
head(df_latest[order(-df_latest$tot_death),],3)
```

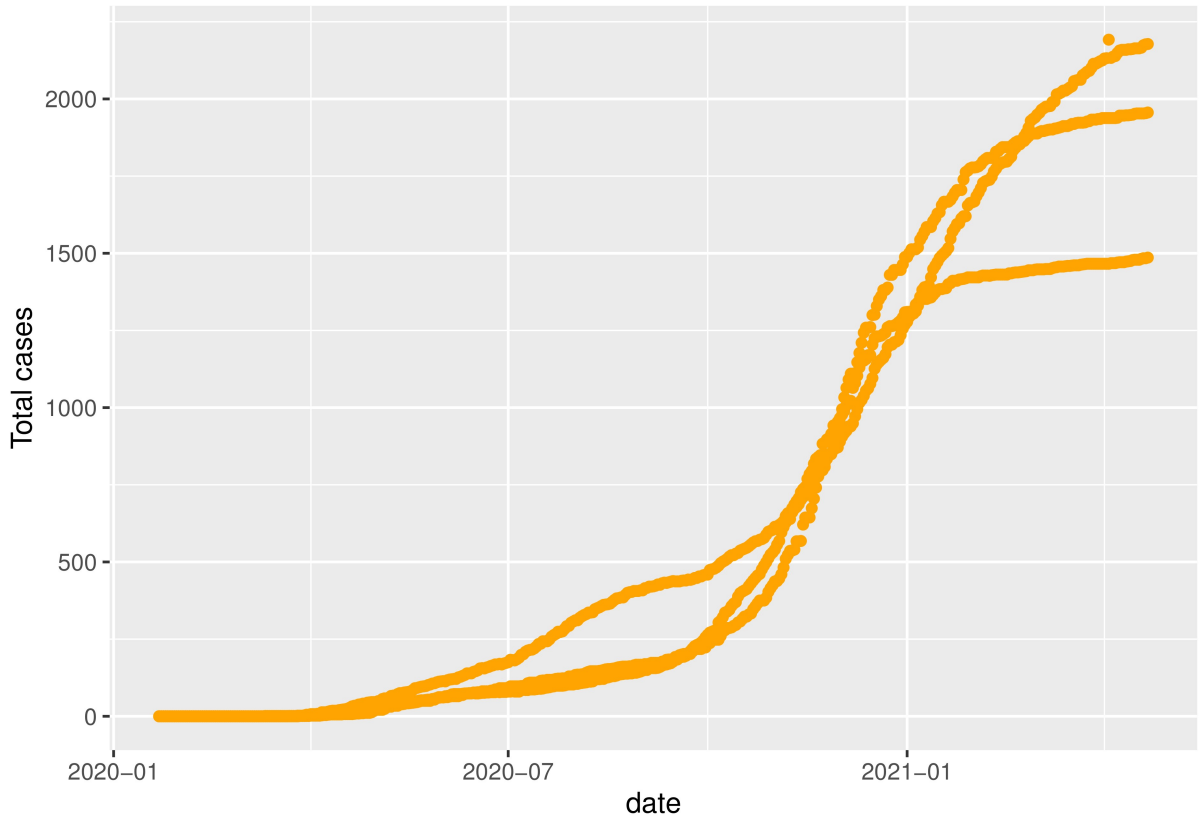
```
## # A tibble: 3 x 15
##   Column1 date       state tot_cases new_case tot_death new_death Region
##     <dbl> <date>     <chr>     <dbl>    <dbl>     <dbl>     <dbl> <chr>
## 1   27368 2021-04-22 CA      3624838     2411     59992       102 West
## 2   27408 2021-04-22 TX      2857017     2990     48828        69 South
## 3   27391 2021-04-22 FL      2149932     6574     34696        80 South
## # ... with 7 more variables: Population_per_sqyare_km <dbl>, population <list>,
## #   area <dbl>, cases_per_sq_km <dbl>, deathrate <dbl>, recoveryrate <dbl>,
## #   infectionrate <dbl>
```

```r
#All the visualization for question 2

#Graph 1

x<-df[df$state =="MA" |df$state =="NJ"|df$state =="CT", ]$date

y<-df[df$state =="MA" |df$state =="NJ"|df$state =="CT" ,]$deathrate

ggplot(df[df$state =="MA" |df$state =="NJ"|df$state =="CT" ,],aes(x=x,y=y),color = state) + geom_line(c
  xlab("date")+ylab("Total cases")
```

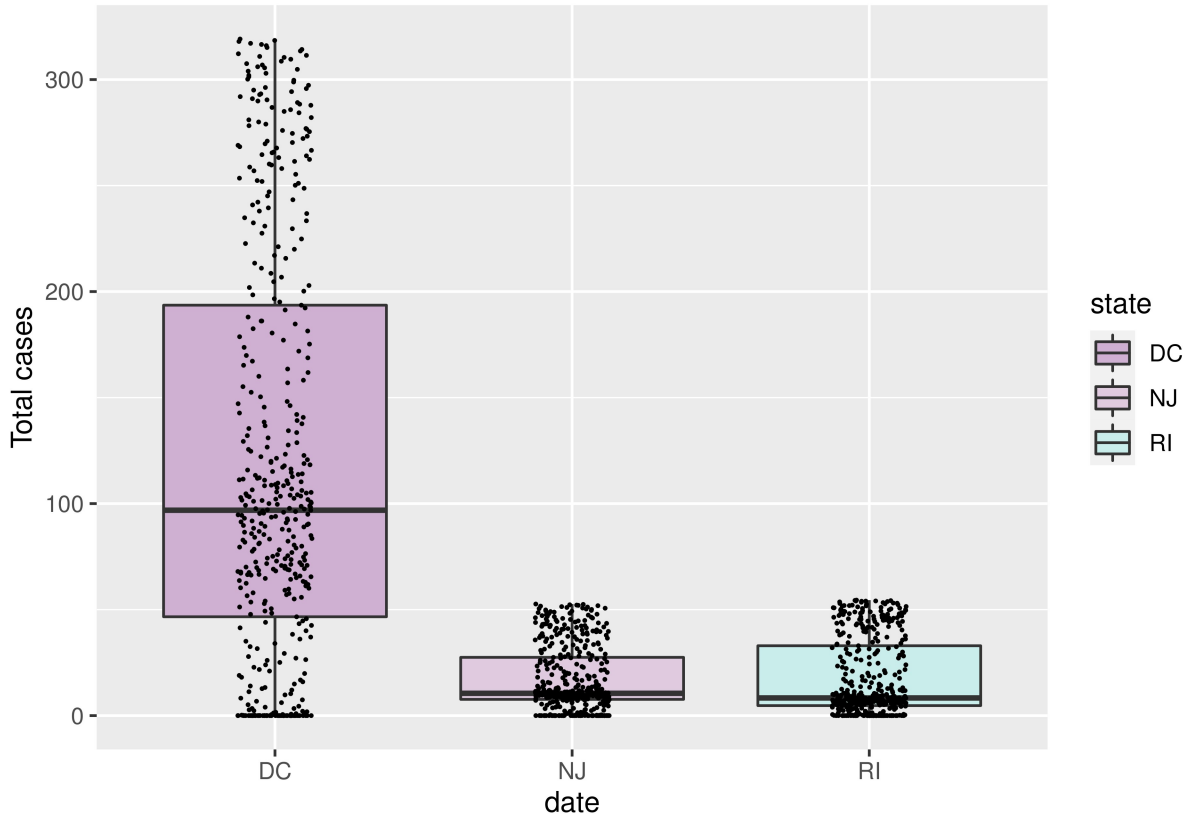



```r
#Graph 2
x<-df[df$state =="ND" |df$state =="SD"|df$state =="UT", ]$date

y<-df[df$state =="ND" |df$state =="SD"|df$state =="UT" ,]$tot_death
```

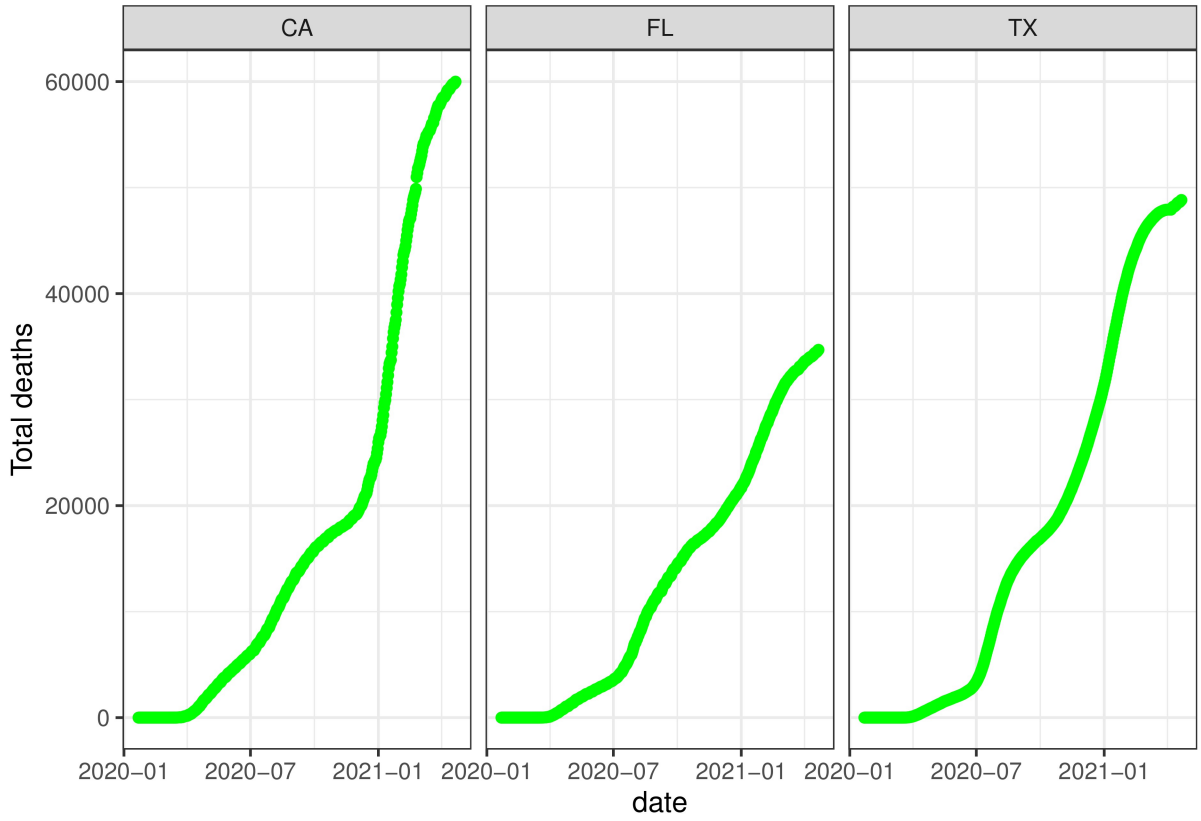
```r
ggplot(df[df$state =="ND" |df$state =="SD"|df$state =="UT" ,],aes(x=x,y=y)) + geom_point(color = c("ora
  xlab("date")+ylab("Total cases")
```



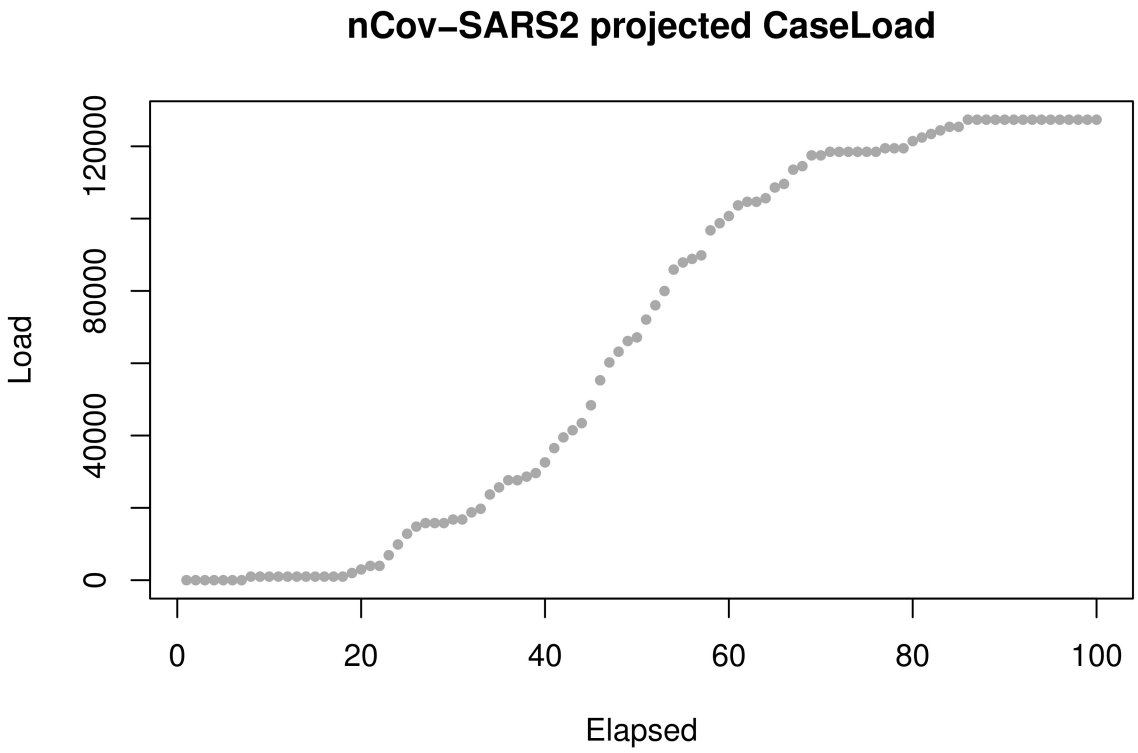

```r
#Graph 3
x<-df[df$state =="DC" |df$state =="RI"|df$state =="NJ", ]$state

y<-df[df$state =="DC" |df$state =="RI"|df$state =="NJ" ,]$cases_per_sq_km
color = c("#cfb1d2","#dfcbe1","#caeeea","#bde6d1","#fefec8")

ggplot(df[df$state =="DC" |df$state =="RI"|df$state =="NJ" ,],aes(x=x,y=y),color = state) + geom_boxplo
  xlab("date")+ylab("Total cases")+
  scale_fill_manual(values = color)+
  geom_jitter(size = 0.2, width = 0.125)
```



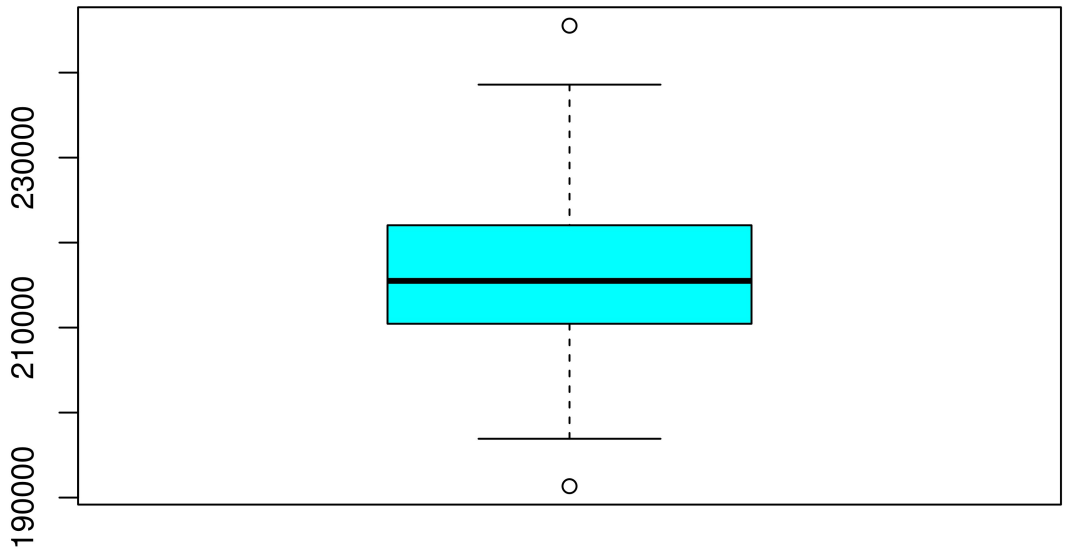

```
#Graph 4

x<-df[df$state =="CA" |df$state =="TX"|df$state =="FL", ]$date

y<-df[df$state =="CA" |df$state =="TX"|df$state =="FL" ,]$tot_death

ggplot(df[df$state =="CA" |df$state =="TX"|df$state =="FL" ,],aes(x=x,y=y)) +
  geom_point(color = c("green")) +
  facet_wrap(~state)+
  xlab("date")+
  ylab("Total deaths")+
  theme_bw()
```




```r
#Question 3

a<-5.32
b<-49.3
c<-51.6
d<- 987.32
num_days = 100
days <- 1:num_days
lambda_sim <- exp(-a*((days-b)**2/c**2))
W <- 987.32*rpois(num_days,pi*lambda_sim)
plot(cumsum(W), xlab = "Elapsed", ylab = "Load",
main = 'nCov-SARS2 projected CaseLoad', pch = 16, cex = 0.75, col = "darkgrey")
```

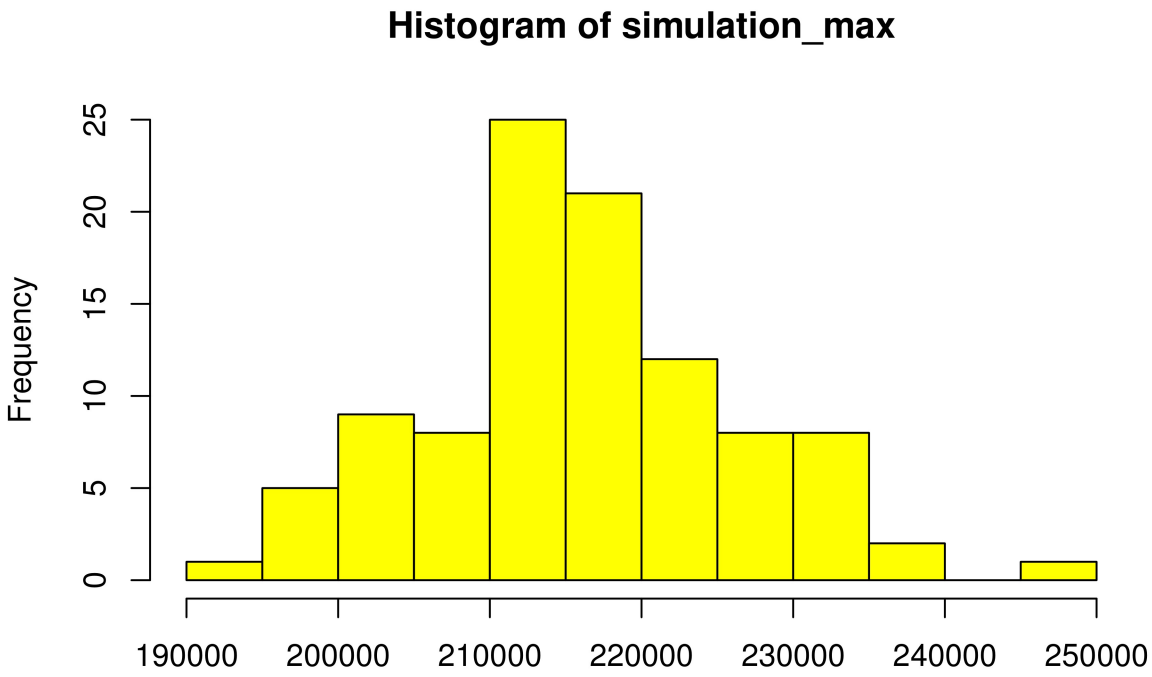



```
simulations_mean <- vector()
simulation_max <- vector()

for (a in 1:100){
num_days = 100
days <- 1:num_days
lambda_sim <- exp(-5.32*((days-49.3)**2/100**2))
W <- 987.32*rnorm(num_days,pi*lambda_sim)

simulations_mean[a] <- mean(cumsum(W))

simulation_max[a]<- max(cumsum(W))
}

par(mfrow=c(1,1))

boxplot(simulation_max,col="cyan",xlab = "Boxplot of maximum case load value of each simulation")
```




Boxplot of maximum case load value of each simulation

```
hist(simulation_max,col = "yellow",xlab ="Histogram of maximum case load value of each simulation")
```

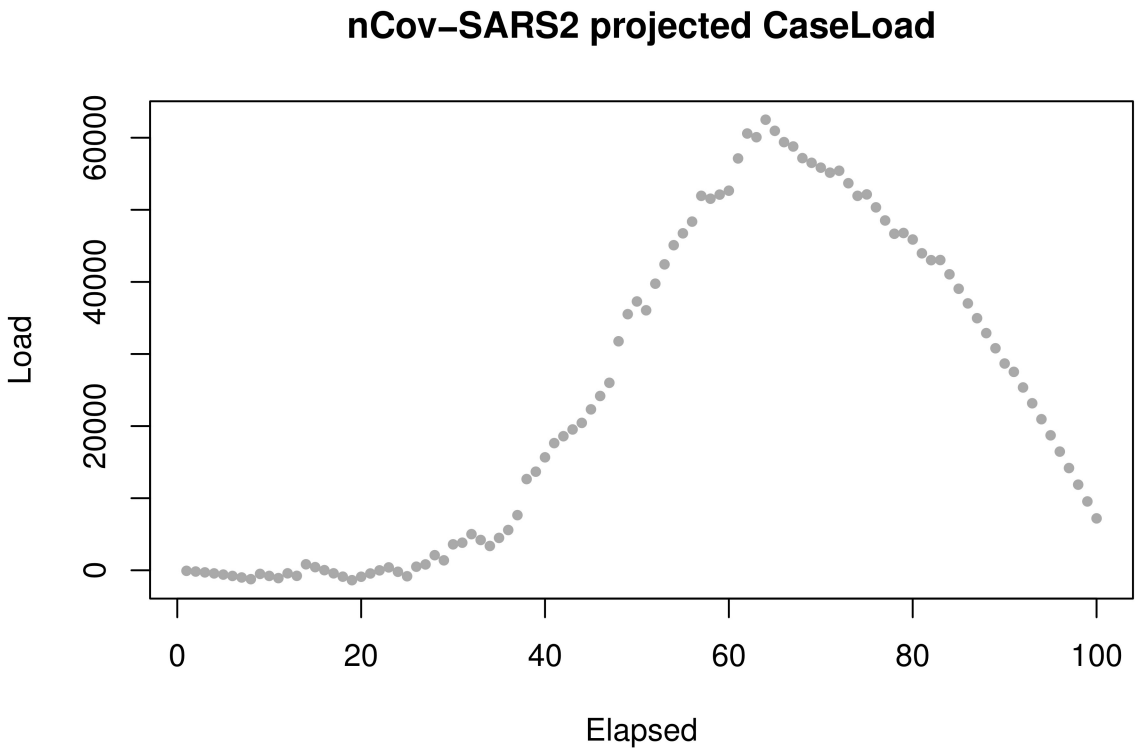



Histogram of maximum case load value of each simulation

```
summary(simulation_max)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  191343  210494  215497  216073  222028  245522
```

```
num_days = 100
days <- 1:num_days
lambda_sim <- exp(-5.32*((days-49.3)**2/51.6**2))

W <- 987.32*rpois(num_days,pi*lambda_sim) -(23*days + 50)
plot(cumsum(W), xlab = "Elapsed", ylab = "Load",
main = 'nCov-SARS2 projected CaseLoad', pch = 16, cex = 0.75, col = "darkgrey")
```


nCov−SARS2 projected CaseLoad
Load
0
20000
40000
60000
0
20
40
60
80
100
Elapsed