MTH522 Mathematical Statistics Project 1

Anushree Chopde 01865258

3/19/2020

Corona Virus Data Analysis:

Coronavirus also called as COVID-19 is the loudest topic around the world.

Data Source : novel-corona-virus-2019-dataset

Datasets used :

1) covid_19_data.csv - 7014 obervations and 9 variables
2) COVID19_line_list_data.csv - 1085 observations and 27 variables

The motive of this project is to show the trends in COVID-19 spread. Following are the points on which Data Analysis and Visualizations are done in this project :

- Plot of Observation Date vs Number of Confirmed cases of COVID-19.
- Plot of Observation Date vs Number of Deaths due to COVID-19.
- Plot of Observation Date vs Number of Recovery cases of COVID-19.
- Plot of Number of Confirmed cases vs Number of Deaths due to COVID-19.
- Plot of Number of Confirmed cases vs Number of Recovery cases of COVID-19.
- Visualization of Observation Date vs Number of Confirmed cases in China and the rest of the world.
- Visualization of Observation Date vs Number of Deaths in China and the rest of the world.
- Visualization of Observation Date vs Number of Recovery cases in China and the rest of the world.
- Mortality Rate China vs the rest of the world.
- Healed among Infected China vs the rest of the world.
- Analysis of Deaths due to COVID-19 based on age and gender.

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(readr)
library(DT)
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

library(spatstat)

## Loading required package: spatstat.data

## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## Loading required package: rpart

##
## spatstat 1.63-3       (nickname: 'Wet paint')
## For an introduction to spatstat, type 'beginner'

##
## Attaching package: 'spatstat'

## The following objects are masked from 'package:car':
##
##     bc, ellipse

library(tidyverse)

## Registered S3 method overwritten by 'cli':
##   method     from
##   print.boxx spatstat

## ── Attaching packages
──────────────────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ tibble  2.1.3     ✓ stringr 1.4.0
## ✓ tidyr   1.0.2     ✓ forcats 0.5.0
## ✓ purrr   0.3.3

## ── Conflicts ─────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## x nlme::collapse() masks dplyr::collapse()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x car::recode()    masks dplyr::recode()
## x purrr::some()    masks car::some()

library(scales)

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:spatstat':
##
##     rescale

## The following object is masked from 'package:readr':
##
##     col_factor

library(RColorBrewer)
library(ggthemes)
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

library(ggrepel)
library(lubridate)

##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date

library(arsenal)
```

```
##
## Attaching package: 'arsenal'

## The following object is masked from 'package:lubridate':
##
##     is.Date

## The following object is masked from 'package:scales':
##
##     ordinal

library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa

library(party)

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

##
## Attaching package: 'modeltools'

## The following object is masked from 'package:spatstat':
##
##     parameters

## The following object is masked from 'package:car':
##
##     Predict

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

##
## Attaching package: 'strucchange'

## The following object is masked from 'package:stringr':
##
##     boundary
```

Analysis of COVID-19

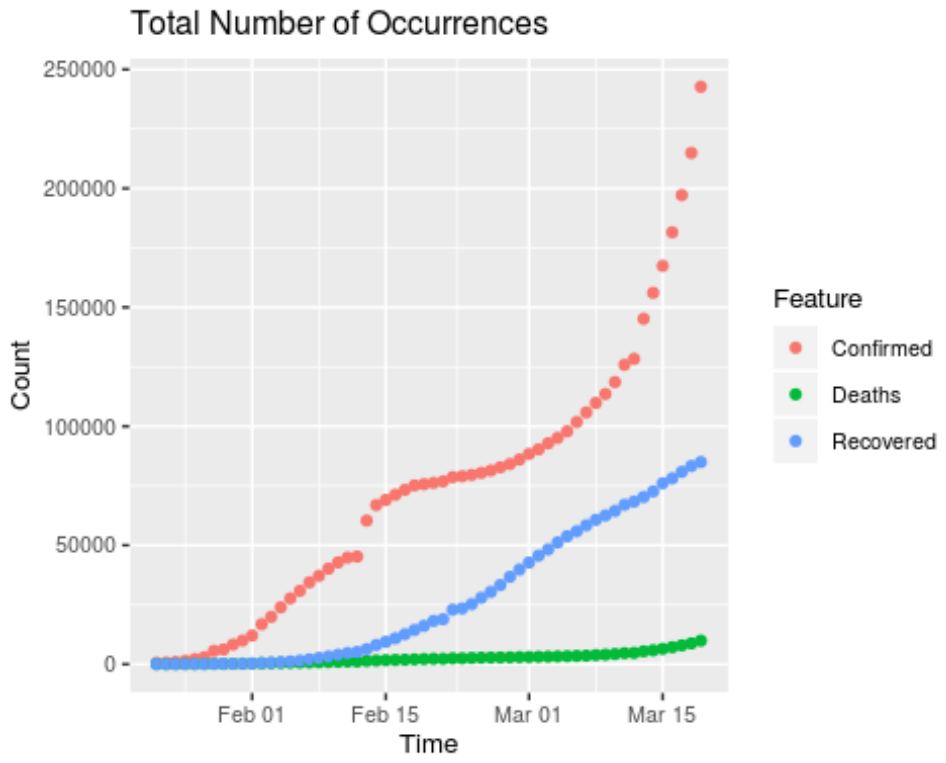
Reading data covid_19_data.csv in R to read inputs and to perform the analysis.

```
#data1 <- covid_19_data
data1 <- read.csv("covid_19_data.csv")
dim(data1)

## [1] 7014    8

PRINT_MODE = 'datatable'
print_df <- function(df_in) {
  if (PRINT_MODE == 'datatable') {
    return(datatable(df_in))
  }
  return(df_in)
}
colnames(data1) <- make.names(colnames(data1))
data1 %>% print_df()

year <- function(date_in) {
    if (nchar(str_split(date_in, "/")[[1]][[3]]) == 2) {
        return(str_c(date_in, '20'))
    }
    date_in
}

data1$ObservationDate <- sapply(data1$ObservationDate, year)

data1 <- data1 %>%
    mutate(ObservationDate = as.Date(ObservationDate, '%m/%d/%Y'))

data1_time_aggregate <- data1 %>%
  group_by(ObservationDate) %>%
  summarise_at(vars(Confirmed, Deaths, Recovered), sum) %>%
  pivot_longer(cols = c('Confirmed', 'Deaths', 'Recovered'),
               names_to = 'Feature', values_to = 'Value')
```

```
data1_time_aggregate %>% print_df()
```

The below plot shows the Total Number of Occurrences of COVID-19 from February 1, 2020. The plot shows the Number of Confirmed cases, Number of Deaths and Number of Recovery Cases.

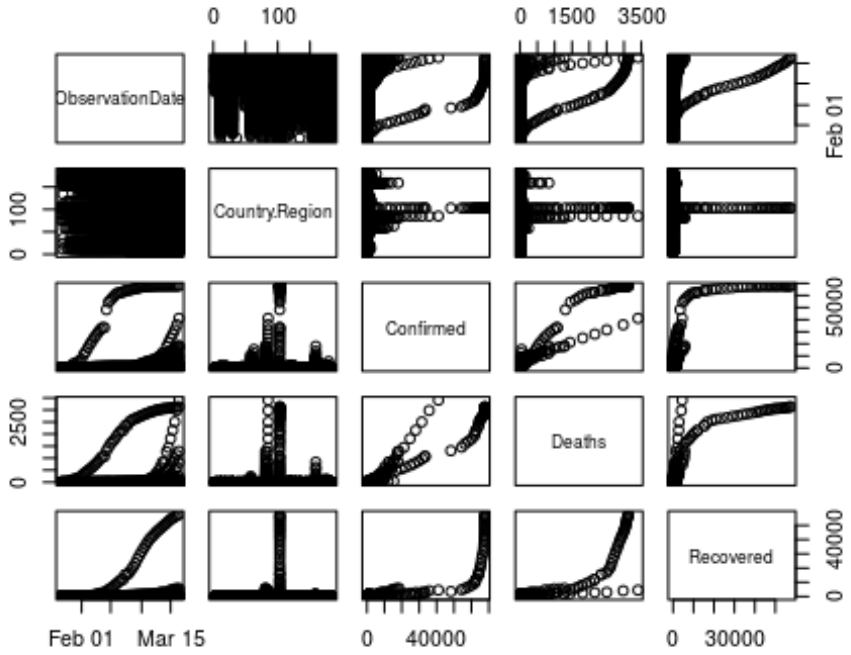
```
ggplot(data1_time_aggregate,
       aes(x = ObservationDate, y = Value,
           group = Feature, fill = Feature, colour = Feature), pch = 9) +
  geom_point() +
  theme(text = element_text(size = 10)) + xlab('Time') + ylab('Count') +
  ggtitle('Total Number of Occurrences')
```



From the above plot, we can observe that there is a huge increase in the number of Confirmed cases.

Further for Prediction Analysis using Linear Regression, we need to observe the data and find which variables have strong correlation. This can be done by using pairs()in R.

```
pairs(data1[, c(2,4,6,7,8)])
```



From the above graph we got from pairs(), we observe that there is : * a correlation between - Observation Date with Confirmed, Deaths and Recovered cases * a strong correlation between - Confirmed and Deaths and Confirmed and Recovered cases

So my further prediction analysis using linear regression are based on the above parameters.

Linear regression :

- Scatter plot is used.
- Equation : y = b0 + b1x where b0 is y-intercept

The R function predict() and data.frame() is used to specify the particular dependent variable value like a particular Observation Date / Number of Confirmed cases for which we are predicting the value of our dependent variable.

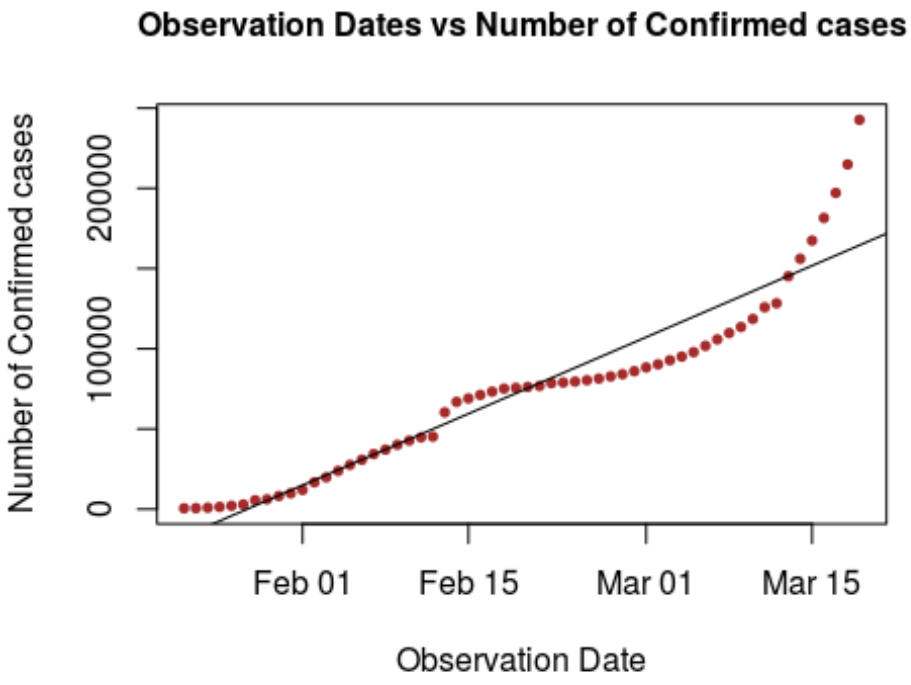
Observation Date with Confirmed, Deaths and Recovered cases

Prediction of Number of Confirmed Cases:

Following is the analysis Number of Confirmed cases from February 1, 2020.

```
data1$ObservationDate <- as.Date(data1$ObservationDate, format = "%Y/%m/%d")
#data1$Date <- as.numeric(as.Date(data1$ObservationDate, format =
"%m/%d/%y"))

Conf_confirmed <- data1 %>%
```

```
  group_by(ObservationDate) %>%
  summarise(totalConfirmed = sum(Confirmed))

plot(Conf_confirmed, xlab = "Observation Date", ylab = "Number of Confirmed
cases", main = "Observation Dates vs Number of Confirmed cases", pch = 20,
col = "Brown", font = 1.5, cex.main = 1.0, cex.axis = 1.0, cex.lab=1.0,
col.main = "black", cex = 0.9)
abline(lm(totalConfirmed ~ ObservationDate, data = Conf_confirmed))
```




```
results_confirmed <- lm(totalConfirmed ~ ObservationDate, data =
Conf_confirmed)
results_confirmed

##
## Call:
## lm(formula = totalConfirmed ~ ObservationDate, data = Conf_confirmed)
##
## Coefficients:
##     (Intercept)  ObservationDate
##       -58204276             3183

summary(results_confirmed)

##
## Call:
## lm(formula = totalConfirmed ~ ObservationDate, data = Conf_confirmed)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -22123 -13217   -511   6954  78148
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.820e+07  2.608e+06  -22.32   <2e-16 ***
## ObservationDate  3.183e+03  1.424e+02   22.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18160 on 56 degrees of freedom
## Multiple R-squared:  0.8992, Adjusted R-squared:  0.8974
## F-statistic: 499.5 on 1 and 56 DF,  p-value: < 2.2e-16
```

```
new_date <- data.frame(ObservationDate = "2020/05/01")
new_date$ObservationDate <- as.Date(new_date$ObservationDate, format =
"%Y/%m/%d")

predict(results_confirmed, new_date)
```

```
##      1
## 301417
```

The plot of Observation Date vs Number of Confirmed cases shows the linear regression on the two parameters.

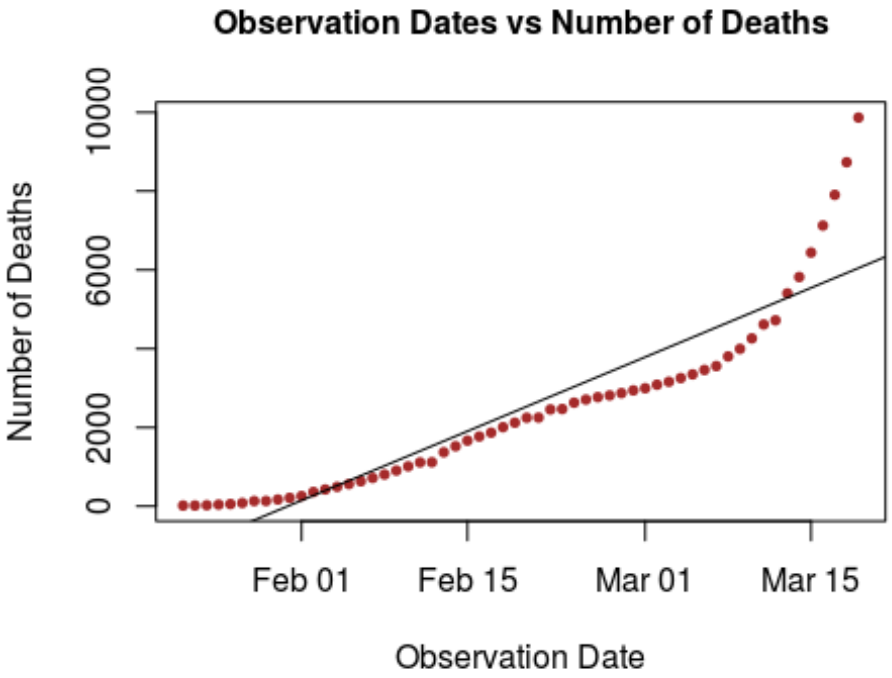
From summary we can see that the variables are significant.

The above results show that by the month of May there will be 301417 COVID-19 Confirmed cases all over the world.

Prediction of Number of Deaths :

The following is the prediction analysis of Number of Deaths from February 1, 2020.

```
data1$ObservationDate <- as.Date(data1$ObservationDate, format = "%Y/%m/%d")
#data1$Date <- as.numeric(as.Date(data1$ObservationDate, format =
"%m/%d/%y"))

Conf_deaths <- data1 %>%
  group_by(ObservationDate) %>%
  summarise(totalDeaths = sum(Deaths))

plot(Conf_deaths, xlab = "Observation Date", ylab = "Number of Deaths", main
= "Observation Dates vs Number of Deaths", pch = 20, col = "Brown", font =
1.5, cex.main = 1.0, cex.axis = 1.0, cex.lab=1.0, col.main = "black", cex =
0.9)
abline(lm(totalDeaths ~ ObservationDate, data = Conf_deaths))
```




```
results_deaths <- lm(totalDeaths ~ ObservationDate, data = Conf_deaths)
results_deaths

##
## Call:
## lm(formula = totalDeaths ~ ObservationDate, data = Conf_deaths)
##
## Coefficients:
##     (Intercept)  ObservationDate
##      -2295792.3            125.5

summary(results_deaths)

##
## Call:
## lm(formula = totalDeaths ~ ObservationDate, data = Conf_deaths)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -984.2 -452.7 -264.0  251.7 3818.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.296e+06  1.306e+05  -17.58   <2e-16 ***
## ObservationDate  1.255e+02  7.132e+00   17.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 909.3 on 56 degrees of freedom
## Multiple R-squared:  0.8469, Adjusted R-squared:  0.8441
## F-statistic: 309.7 on 1 and 56 DF,  p-value: < 2.2e-16

new_date <- data.frame(ObservationDate = "2020/05/01")
new_date$ObservationDate <- as.Date(new_date$ObservationDate, format =
"%Y/%m/%d")

predict(results_deaths, new_date)

##        1
## 11445.26
```

The plot of Observation Date vs Number of Deaths shows the linear regression on the two parameters.

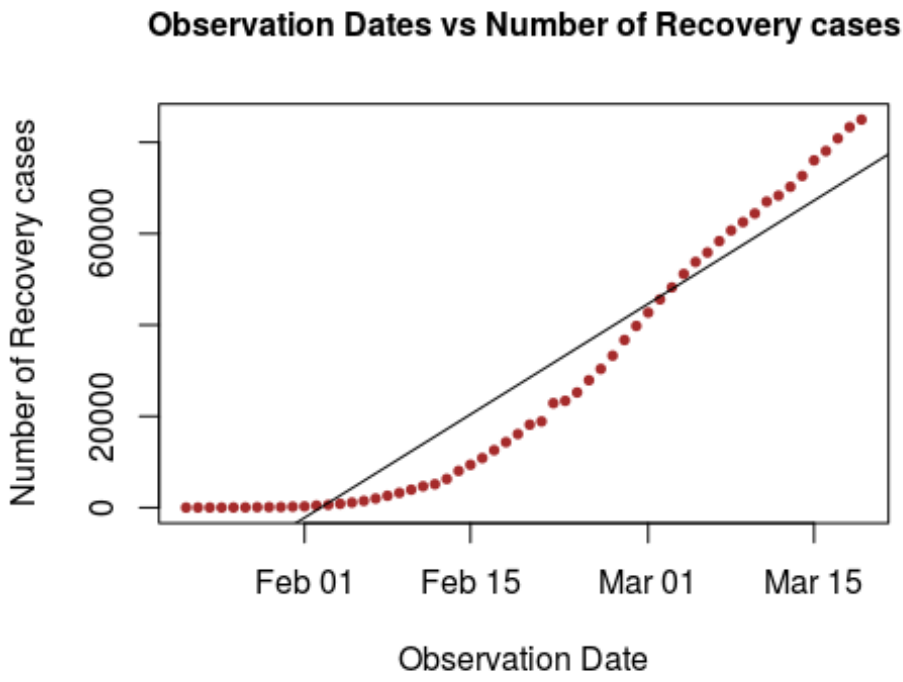
From summary we can see that the variables are significant.

It is predicted that by the month of May there will be 11445.26 deaths all over the world due to COVID-19.

Prediction of Number of Recovery cases :

The following is the prediction analysis of Number of Deaths from February 1, 2020.

```
data1$ObservationDate <- as.Date(data1$ObservationDate, format = "%Y/%m/%d")
#data1$Date <- as.numeric(as.Date(data1$ObservationDate, format =
"%m/%d/%y"))

Conf_recovered <- data1 %>%
  group_by(ObservationDate) %>%
  summarise(totalRecovered = sum(Recovered))

plot(Conf_recovered, xlab = "Observation Date", ylab = "Number of Recovery
cases", main = "Observation Dates vs Number of Recovery cases", pch = 20, col
= "Brown", font = 1.5, cex.main = 1.0, cex.axis = 1.0, cex.lab=1.0, col.main
= "black", cex = 0.9)
abline(lm(totalRecovered ~ ObservationDate, data = Conf_recovered))
```




```
results_recovered <- lm(totalRecovered ~ ObservationDate, data =
Conf_recovered)
results_recovered

##
## Call:
## lm(formula = totalRecovered ~ ObservationDate, data = Conf_recovered)
##
## Coefficients:
##     (Intercept)  ObservationDate
##       -29539130             1615

summary(results_recovered)

##
## Call:
## lm(formula = totalRecovered ~ ObservationDate, data = Conf_recovered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11229.0  -8607.4    -41.7   6219.5  18348.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.954e+07  1.250e+06  -23.64   <2e-16 ***
## ObservationDate  1.615e+03  6.824e+01   23.66   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8700 on 56 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9074
## F-statistic: 559.9 on 1 and 56 DF,  p-value: < 2.2e-16
```

```
new_date <- data.frame(ObservationDate = "2020/05/01")
new_date$ObservationDate <- as.Date(new_date$ObservationDate, format =
"%Y/%m/%d")

predict(results_recovered, new_date)
```

```
##        1
## 143145.1
```

The plot of Observation Date vs Number of Recovery cases shows the linear regression on the two parameters.

From summary we can see that the variables are significant.

It is predicted that by the month of May there will be 143145.1 Recovery cases from COVID-19 infection all over the world.

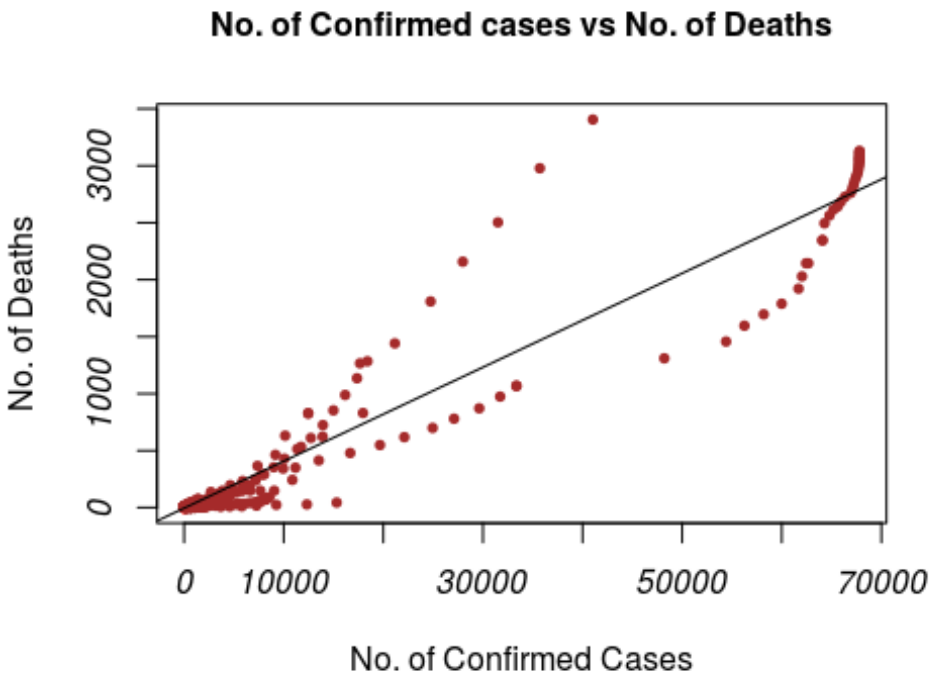
Next is the prediction of Number of Deaths from Number of Confirmed cases and prediction of Number of Recovery cases from Number of Confirmed cases.

Prediction of Deaths :

The below analysis shows the linear regression between Number of Confirmed cases and Number of Deaths.

Here we are going to predict number of Deaths due to COVID-19 at a particlar number of Confirmed cases. For this I have used the R function predict and data.frame to specify number of Confirmed cases.

```
plot(data1$Confirmed, data1$Deaths, xlab = "No. of Confirmed Cases", ylab =
"No. of Deaths", main = "No. of Confirmed cases vs No. of Deaths", pch = 20,
col = "Brown", font = 3, cex.main = 1.0, cex.axis = 1.0, cex.lab=1.0,
col.main = "black", cex = 0.9)
abline(lm(Deaths ~ Confirmed, data = data1))
```




```
results_confirmedvsdeaths <- lm(Deaths ~ Confirmed, data = data1)
results_confirmedvsdeaths

##
## Call:
## lm(formula = Deaths ~ Confirmed, data = data1)
##
## Coefficients:
## (Intercept)    Confirmed
##    -4.74352      0.04123

summary(results_confirmedvsdeaths)

##
## Call:
## lm(formula = Deaths ~ Confirmed, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -781.18    1.67    4.34    4.70 1718.05
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.7435230  0.6160209    -7.7 1.54e-14 ***
## Confirmed    0.0412257  0.0001248   330.5  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 51.2 on 7012 degrees of freedom
## Multiple R-squared:  0.9397, Adjusted R-squared:  0.9397
## F-statistic: 1.092e+05 on 1 and 7012 DF,  p-value: < 2.2e-16

predict(results_confirmedvsdeaths, data.frame(Confirmed = 500000))

##        1
## 20608.12
```

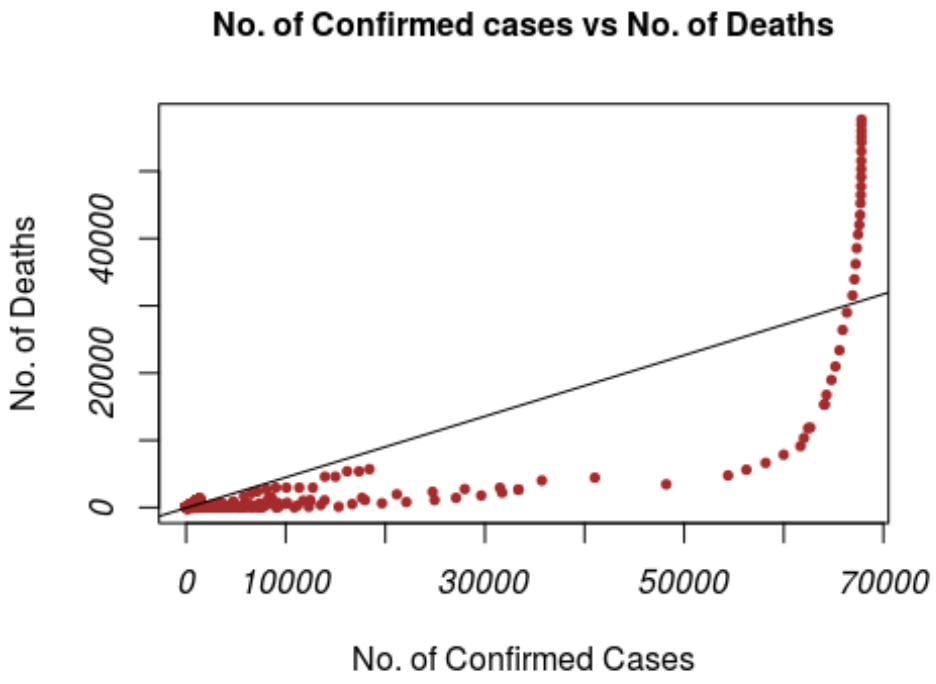
From the above linear regression plot of Number of Confirmed cases vs Number of Deaths, it is predicted that when there will be 500000 cases are confirmed, 20608.12 will die due to COVID-19 infection.

Prediction of Recovery cases:

The below analysis shows the linear regression between Number of Confirmed cases and Number of Deaths.

Here we are going to predict number of Recovery cases from COVID-19 at a particlar number of Confirmed cases. For this I have used the R function predict and data.frame to specify number of Confirmed cases.

```
plot(data1$Confirmed, data1$Recovered, xlab = "No. of Confirmed Cases", ylab
= "No. of Deaths", main = "No. of Confirmed cases vs No. of Deaths", pch =
20, col = "Brown", font = 3, cex.main = 1.0, cex.axis = 1.0, cex.lab=1.0,
col.main = "black", cex = 0.9)
abline(lm(Recovered ~ Confirmed, data = data1))
```




```
results_confirmedvsrecovered <- lm(Recovered ~ Confirmed, data = data1)
results_confirmedvsrecovered

##
## Call:
## lm(formula = Recovered ~ Confirmed, data = data1)
##
## Coefficients:
## (Intercept)    Confirmed
##    -48.3411       0.4541

summary(results_confirmedvsrecovered)

##
## Call:
## lm(formula = Recovered ~ Confirmed, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19885.3     34.7     47.0     47.9  26940.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -48.34112   16.09682  -3.003  0.00268 **
## Confirmed     0.45414    0.00326 139.312  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1338 on 7012 degrees of freedom
## Multiple R-squared:  0.7346, Adjusted R-squared:  0.7346
## F-statistic: 1.941e+04 on 1 and 7012 DF,  p-value: < 2.2e-16

predict(results_confirmedvsrecovered, data.frame(Confirmed = 500000))

##        1
## 227019.2
```

From the above linear regression plot of Number of Confirmed cases vs Number of Recovery cases, it is predicted that when there will be 500000 cases are confirmed, 227019.2 will be recovered from COVID-19 infection.

Visualizations:

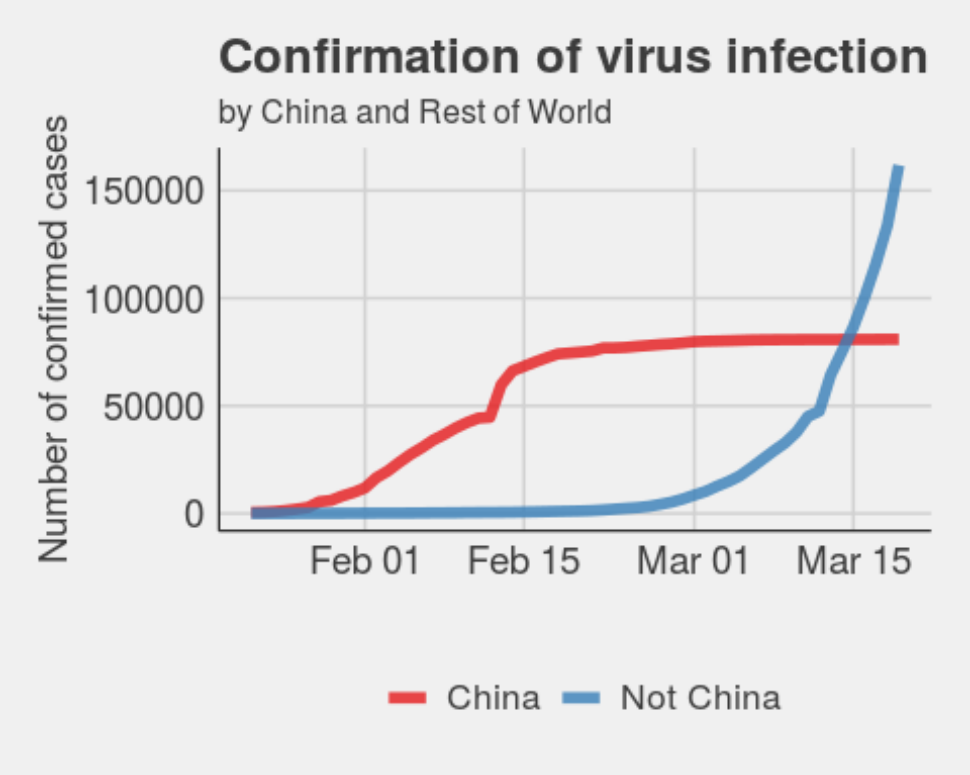
Further more, by using the visualizations of various factors, my analysis explains the Number of Confirmation cases, Number of Deaths and Number of Recovery cases from February 1, 2020. The analysis shows the cases in China and in rest of the world.

China vs Other Countries :

Visualization of Observation Date vs Number of Confirmed cases in China and the rest of the world:

Below is the visualization of Observation Date vs Number of Confirmed cases.

```
data1$isChina <- ifelse(data1$Country %in% c("Mainland China",
"China"),"China","Not China")
data1$ObservationDate <- as.Date(data1$ObservationDate, format = "%m/%d/%y")

Conf1 <- data1 %>%
  group_by(isChina, ObservationDate) %>%
  summarise(Confirmed = sum(Confirmed))

ggplot(Conf1, aes(ObservationDate, Confirmed, colour = isChina))+
  geom_line(size = 2, alpha = 0.8)+
 # geom_point(size = 2.7)+
  labs(x = "", y = "Number of confirmed cases", title =  "Confirmation of
virus infection", subtitle = "by China and Rest of World")+
  scale_colour_brewer(palette = "Set1")+
  theme_fivethirtyeight()+
  theme(legend.position="bottom", legend.direction="horizontal", legend.title
= element_blank(), axis.text = element_text(size = 14),
        legend.text = element_text(size = 13), axis.title = element_text(size
= 14), axis.line = element_line(size = 0.4, colour = "grey10"))
```

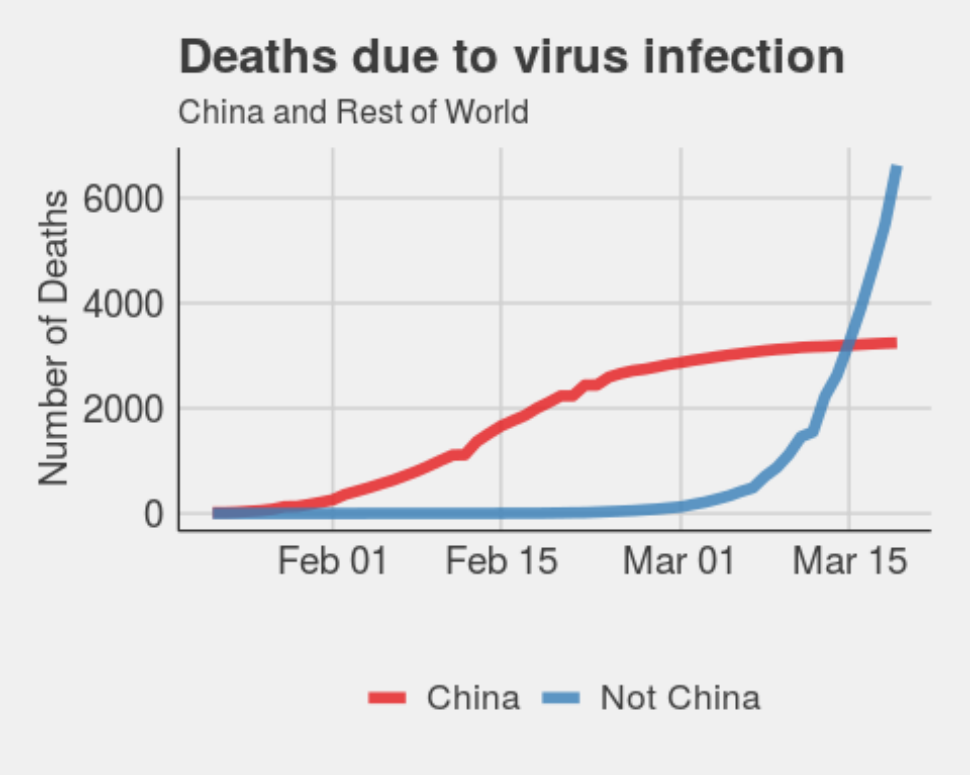


The plot shows that in China the number of Confirmed cases increased from February 1, 2020 and has stabilized to the count of 75000 from February 18, 2020. Whereas, it is increasing drastically and continuously in the rest of the world.

Visualization of Observation Date vs Number of Deaths in China and the rest of the world:

Below is the visualization of Visualization of Observation Date vs Number of Deaths.

```
data1$isChina <- ifelse(data1$Country %in% c("Mainland China",
"China"),"China","Not China")
data1$ObservationDate <- as.Date(data1$ObservationDate, format = "%m/%d/%y")

Conf2 <- data1 %>%
  group_by(isChina, ObservationDate) %>%
  summarise(Deaths = sum(Deaths))

ggplot(Conf2, aes(ObservationDate, Deaths, colour = isChina))+
  geom_line(size = 2, alpha = 0.8)+
  labs(x = "", y = "Number of Deaths", title =  "Deaths due to virus
infection", subtitle = "China and Rest of World")+
  scale_colour_brewer(palette = "Set1")+
  theme_fivethirtyeight()+
  theme(legend.position="bottom", legend.direction="horizontal", legend.title
= element_blank(), axis.text = element_text(size = 14),
        legend.text = element_text(size = 13), axis.title = element_text(size
= 14), axis.line = element_line(size = 0.4, colour = "grey10"))
```



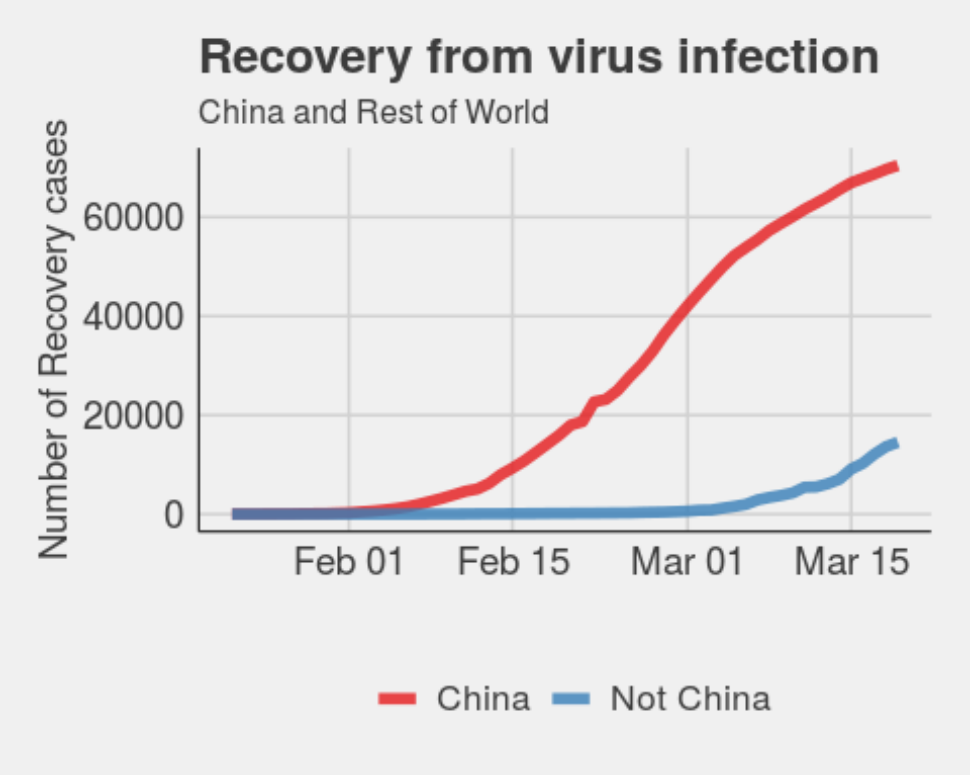
From the above graph, we can see that, on one hand, the number of Deaths in China is beginning to stabilize by March 15, 2020, on the other hand, number of Deaths have is seen to have hiked up from March 15, 2020 for rest of the world.

Visualization of Observation Date vs Number of Recovery cases in China and the rest of the world:

Below is the visualization of Visualization of Observation Date vs Number of Recovery cases.

```
data1$isChina <- ifelse(data1$Country %in% c("Mainland China",
"China"),"China","Not China")
data1$ObservationDate <- as.Date(data1$ObservationDate, format = "%m/%d/%y")

Conf3 <- data1 %>%
  group_by(isChina, ObservationDate) %>%
  summarise(Recovered = sum(Recovered))

ggplot(Conf3, aes(ObservationDate, Recovered, colour = isChina))+
  geom_line(size = 2, alpha = 0.8)+
  labs(x = "", y = "Number of Recovery cases", title =  "Recovery from virus
infection", subtitle = "China and Rest of World")+
  scale_colour_brewer(palette = "Set1")+
  theme_fivethirtyeight()+
  theme(legend.position="bottom", legend.direction="horizontal", legend.title
= element_blank(), axis.text = element_text(size = 14),
```

```
        legend.text = element_text(size = 13), axis.title = element_text(size
= 14), axis.line = element_line(size = 0.4, colour = "grey10"))
```



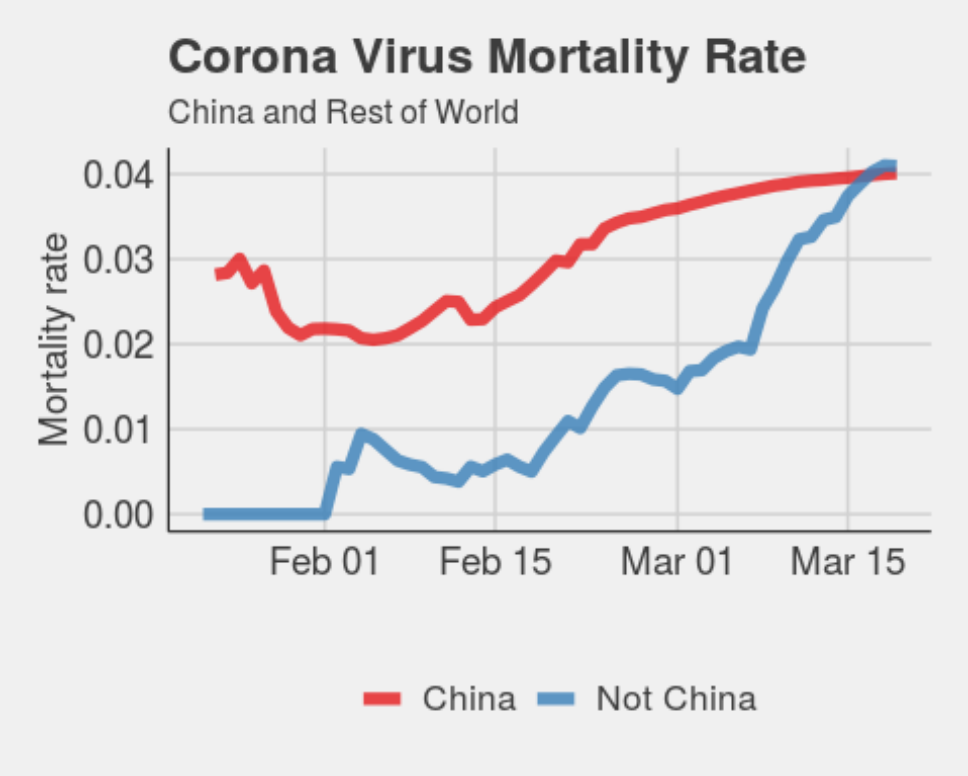
From the above visualization, China's number of Recovered cases have increased, whereas it is quite stagnant for the rest of the world.

Mortality Rate China vs the rest of the world :

The next visualization is of mortality rate.

```
Conf1 <- data1 %>%
  group_by(isChina, ObservationDate) %>%
  summarise(Confirmed = sum(Confirmed))
Conf2 <- data1 %>%
  group_by(isChina, ObservationDate) %>%
  summarise(Deaths = sum(Deaths))

Mortality_rate <- cbind(Conf2, Conf1)
Mortality_rate <- Mortality_rate[,c(1,2,3,6)]
names(Mortality_rate)[3:4] <- c("Deaths", "Total")
Mortality_rate$Deaths_to_all <- Mortality_rate$Deaths/Mortality_rate$Total

ggplot(Mortality_rate[-c(1,47),], aes(ObservationDate,Deaths_to_all, colour =
isChina))+
  geom_line(size = 2.2, alpha = 0.8)+
  labs(x = "", y = "Mortality rate", title =  "Corona Virus Mortality Rate",
subtitle = "China and Rest of World", colour = "")+
```

```
  scale_colour_brewer(palette = "Set1")+
  theme_fivethirtyeight()+
  theme(legend.position="bottom", legend.direction="horizontal", axis.text =
element_text(size = 14), axis.title = element_text(size = 14),
        legend.text = element_text(size = 13), axis.line = element_line(size
= 0.4, colour = "grey10"))
```



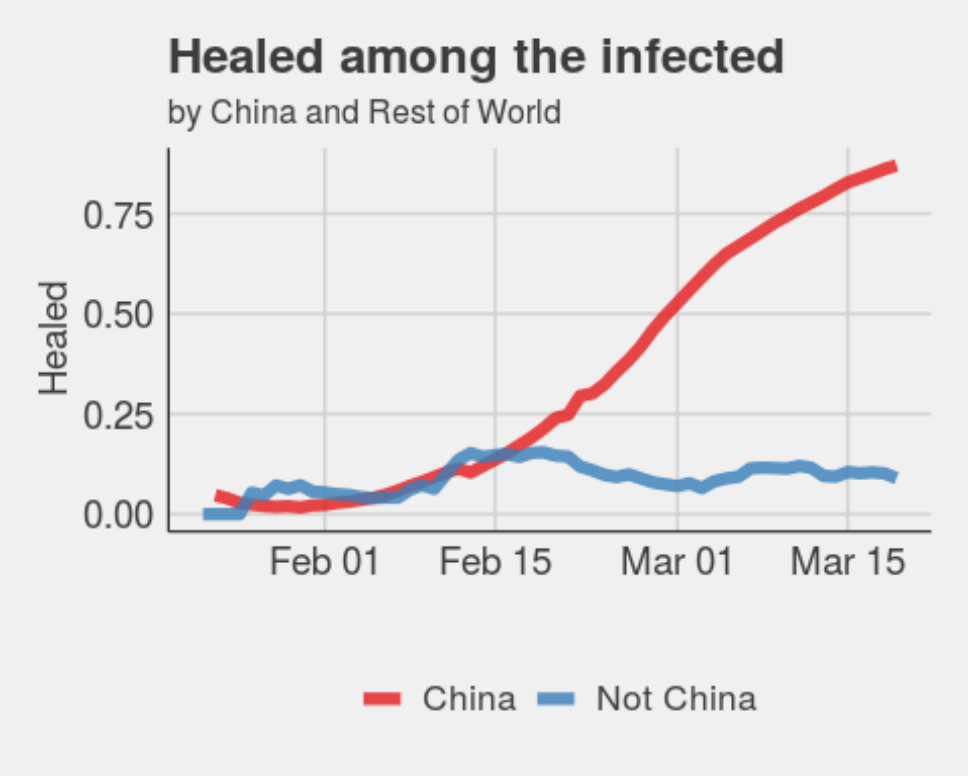
With mortality rate that is the the ratio of people who died of the virus to all infected, we can observe and vizualise the Mortality rate. Throughout the history of the virus to date, this ratio is around 2 - 3 percent in China is near constans and what should be treated as a hope of curing the disease in its entirety. From January 31, we can count this ratio for the rest of the world as there was a fatalities outside of China found. Initially,the mortality rate increased, but it began to decrease over time and never exceeded 1%. It is worth observing that the victims are usually an elderly person or with other diseases, so this value cannot be treated as the probability of death of a newly infected person.

Healed among Infected China vs the rest of the world :

It would be interesting to visualize number of cases of those who were infected got healed.

```
Conf1 <- data1 %>%
  group_by(isChina, ObservationDate) %>%
  summarise(Confirmed = sum(Confirmed))
Conf3 <- data1 %>%
  group_by(isChina, ObservationDate) %>%
  summarise(Recovered = sum(Recovered))
```

```
Healed_rate <- cbind(Conf3, Conf1)
Healed_rate <- Healed_rate[,c(1,2,3,6)]
names(Healed_rate)[3:4] <- c("Rec", "Total")
Healed_rate$Recovered_to_all <- Healed_rate$Rec/Healed_rate$Total
ggplot(Healed_rate[-c(1,47),], aes(ObservationDate,Recovered_to_all, colour =
isChina))+
  geom_line(size = 2.2, alpha = 0.8)+
  #geom_point(size = 3)+
  labs(x = "", y = "Healed", title =  "Healed among the infected", subtitle =
"by China and Rest of World", colour = "")+
  scale_colour_brewer(palette = "Set1")+
  theme_fivethirtyeight()+
  theme(legend.position="bottom", legend.direction="horizontal", axis.text =
element_text(size = 14), axis.title = element_text(size = 14), legend.text =
element_text(size = 13),
        axis.line = element_line(size = 0.4, colour = "grey10"))
```

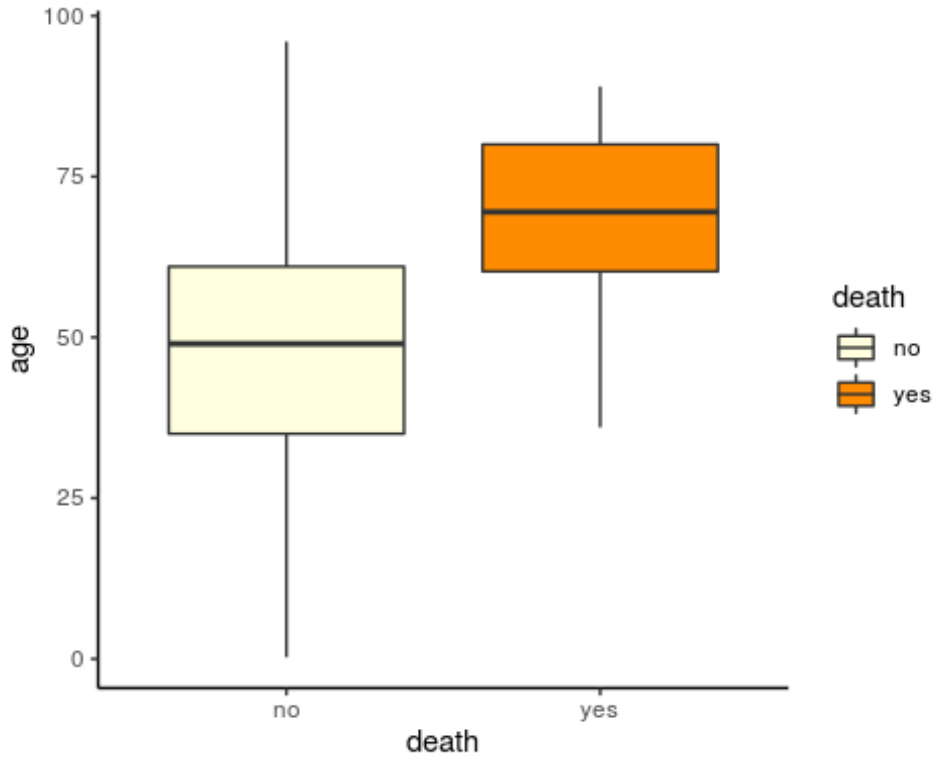


From the above plot we can see that the percentage of people completely cured of the virus in China and the rest of the world. Initially, the value of this coefficient in China was high, but it decreased over time to reach its bottom on 27 January. Since then, we have seen a moderate increase in what should be considered positively (an increase from 2 to more than 10% is optimistic). Outside of China, the coefficient value is seen fuctuating more irregularly, however, it is indicating an upward trend.

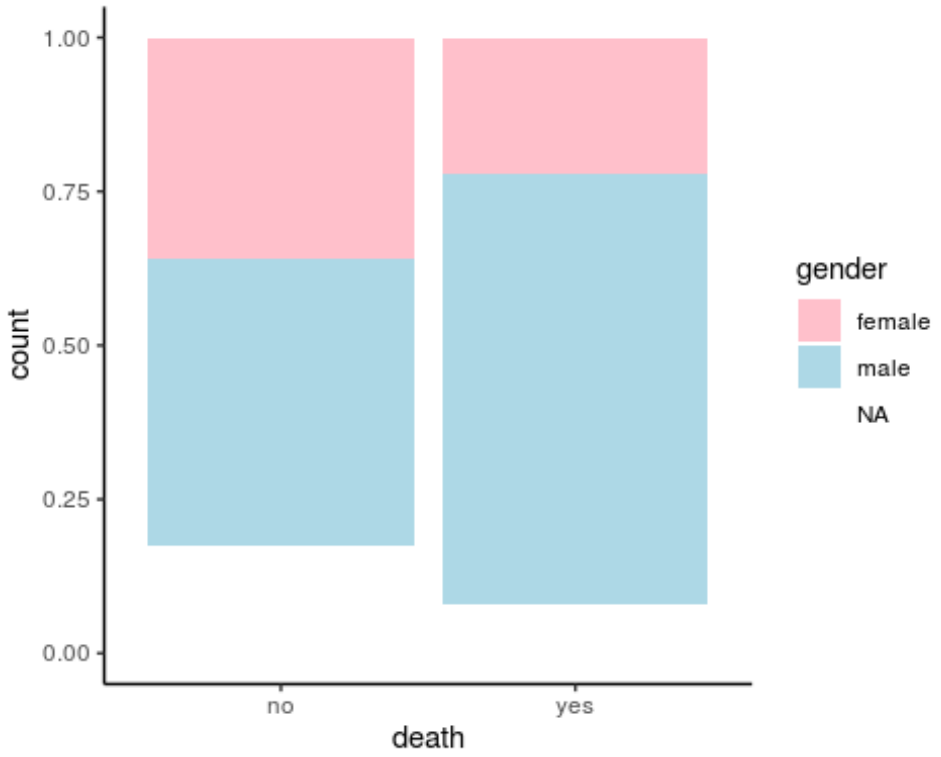
Analysis of Deaths due to Corona Virus

The next module shows the vizualization and analysis of number of Deaths due to Corona virus using the COVID19_line_list_data.csv dataset. This will be analysis on the basis of age and gender. The analysis will tell the number of and which age and gender people died due to COVID-19.

```
data2 <- read.csv("COVID19_line_list_data.csv")
dim(data2)

## [1] 1085   27

## Symptoms

data2$respiratory <- str_detect(data2$symptom,
'breath|pneumonia|breathlessness|dyspnea|respiratory')
data2$abdominal <- str_detect(data2$symptom, 'abdominal|diarrhea|vomiting')

data2 <- data2 %>%
  select(reporting.date, country, gender, age, death, respiratory, abdominal,
symptom)  %>%
  mutate(
    death = ifelse(death == '0', 0, 1),
    country = factor(country),
    gender = factor(gender),
    death = factor(death, label = c('no','yes')),
    reporting.date = as.Date(reporting.date, format = c('$d/$m/$Y')),
    respiratory = factor(respiratory, label = c('no','yes')),
    abdominal = factor(abdominal, label = c('no','yes'))
  )

ggplot(data2, aes(death, age, fill = death))+
  geom_boxplot()+
  theme_classic()+
  scale_fill_manual(values = c('light yellow','dark orange'))

## Warning: Removed 242 rows containing non-finite values (stat_boxplot).
```




```
ggplot(data2, aes(death, fill = gender))+
  geom_bar(position ='fill')+
  theme_classic()+
  scale_fill_manual(values = c('pink','light blue','black'))
```



The first plot shows on x axis the person died or not and on y axis his/her age. From the graph we observe that the people in the age group of around 30 to 60 did not die, whereas people in the age group of around 65 to 80 died due to COVID-19. The mean Age of Patients who died was 70 years.

The second plot shows the deaths based on gender that is on x axis whether the person died or not and on y axis the count. The pink color on the box plot indicates females and the light blue color indicates males. From the plot we observe that there are 20 to 60 percent chance that a male with COVID-19 will not die and 10 to 75 percent chance that he will die due to COVID-19.

```
Analysis1 <- tableby(death ~ age + gender, total = FALSE, data = data2)
summary(Analysis1, text = TRUE)

##
##
## |             |   no (N=1022)    |   yes (N=63)     | p value|
## |:------------|:----------------:|:----------------:|-------:|
## |age          |                  |                  | < 0.001|
## |-  N-Miss    |       237        |         5        |        |
## |-  Mean (SD) | 48.072 (17.763)  | 68.586 (13.582)  |        |
## |-  Range     | 0.250 - 96.000   | 36.000 - 89.000  |        |
## |gender       |                  |                  |   0.004|
## |-  N-Miss    |       178        |         5        |        |
## |-  female    |   368 (43.6%)    |   14 (24.1%)     |        |
## |-  male      |   476 (56.4%)    |   44 (75.9%)     |        |
```

From the above table we can say that the COVID-19 virus is more fatal to older males. We observed more deaths in male patients due to COVID-19.