Project 1: CoronaVirus data analysis

Abstract

This document presents analysis on Coronavirus outbreak. It shows the trend of how virus affected from last four months

Laukik Upadhye

08133608

Dashboard

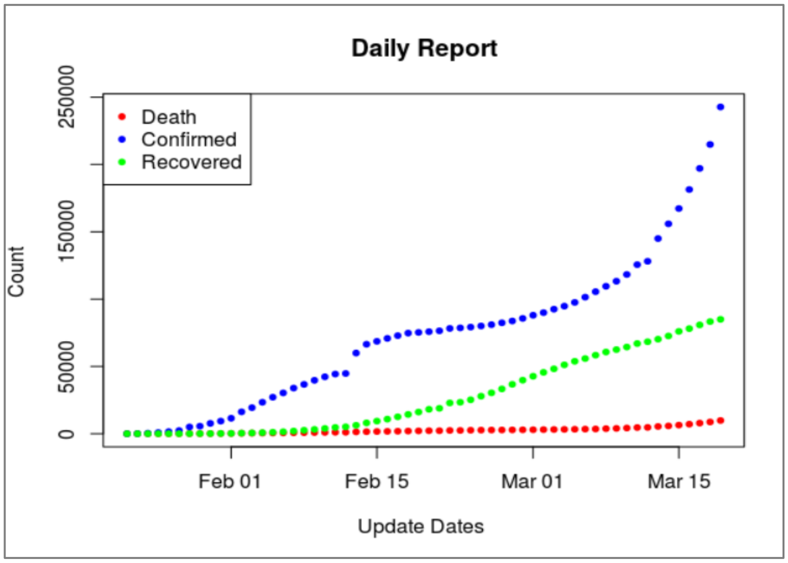

Fig (A): Dates wise death, confirm and recovered cases

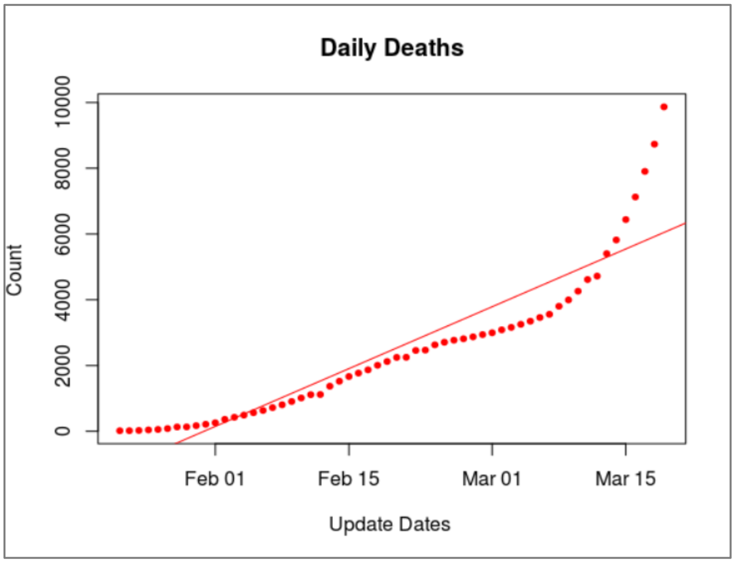

Fig (B): Regression for death count

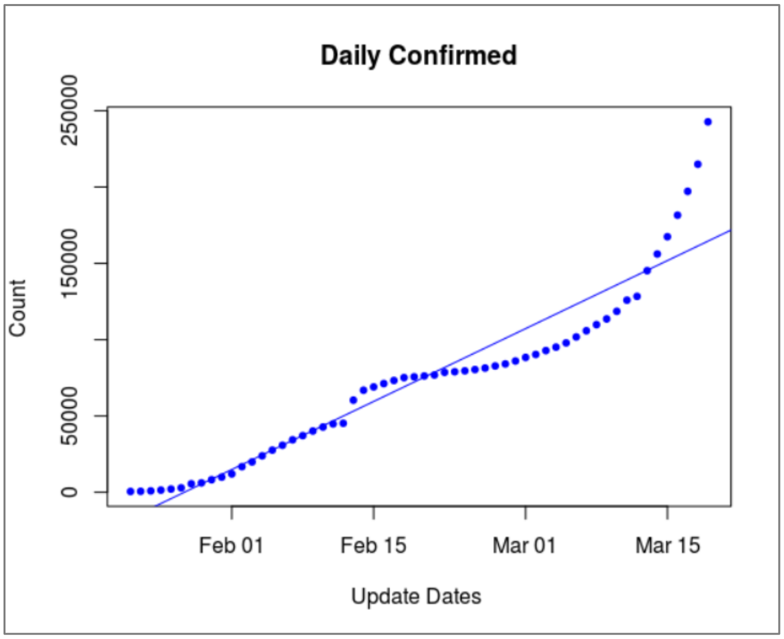

Fig (C): Regression for confirmed cases

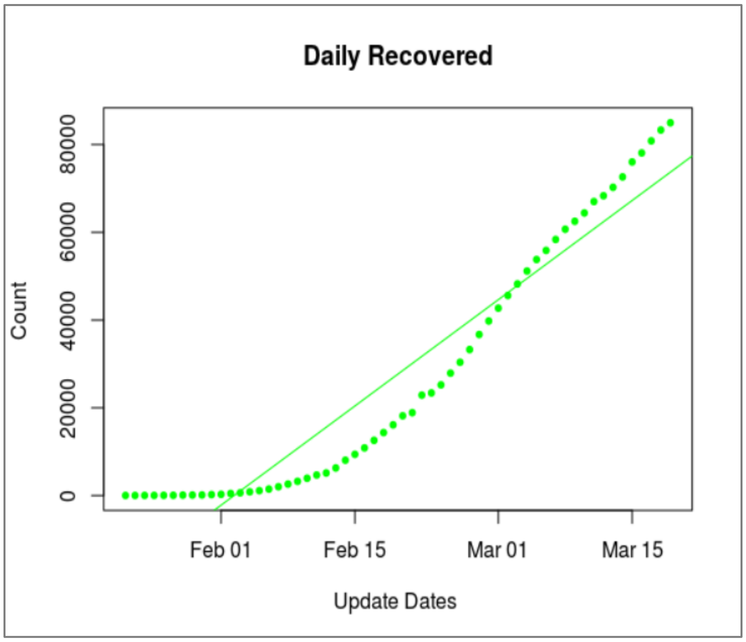

Fig (D): Regression for confirmed cases

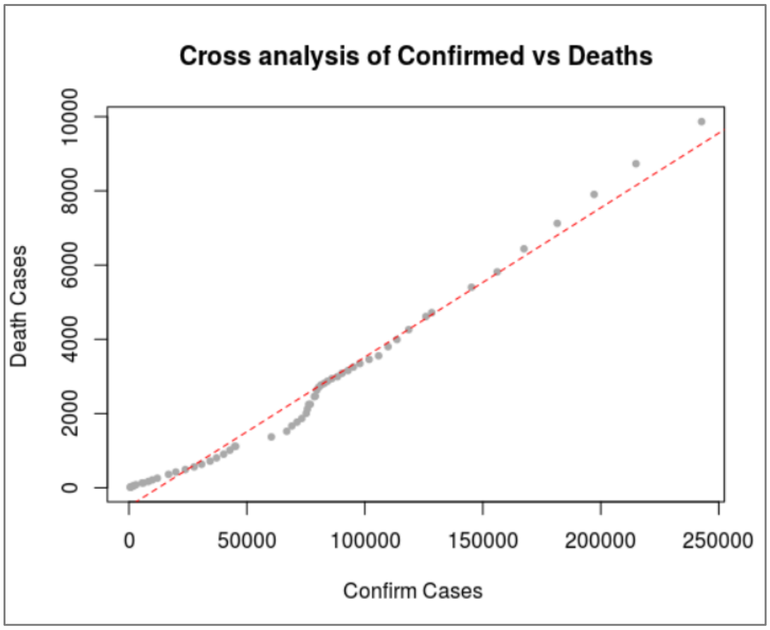

Fig (E): Estimate of expected casualties

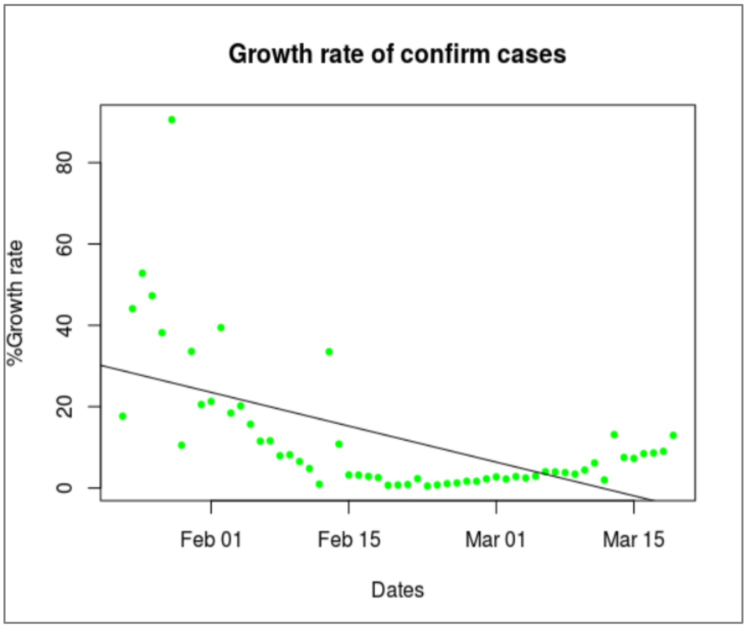

Fig (F): Growth rate of confirmed cases

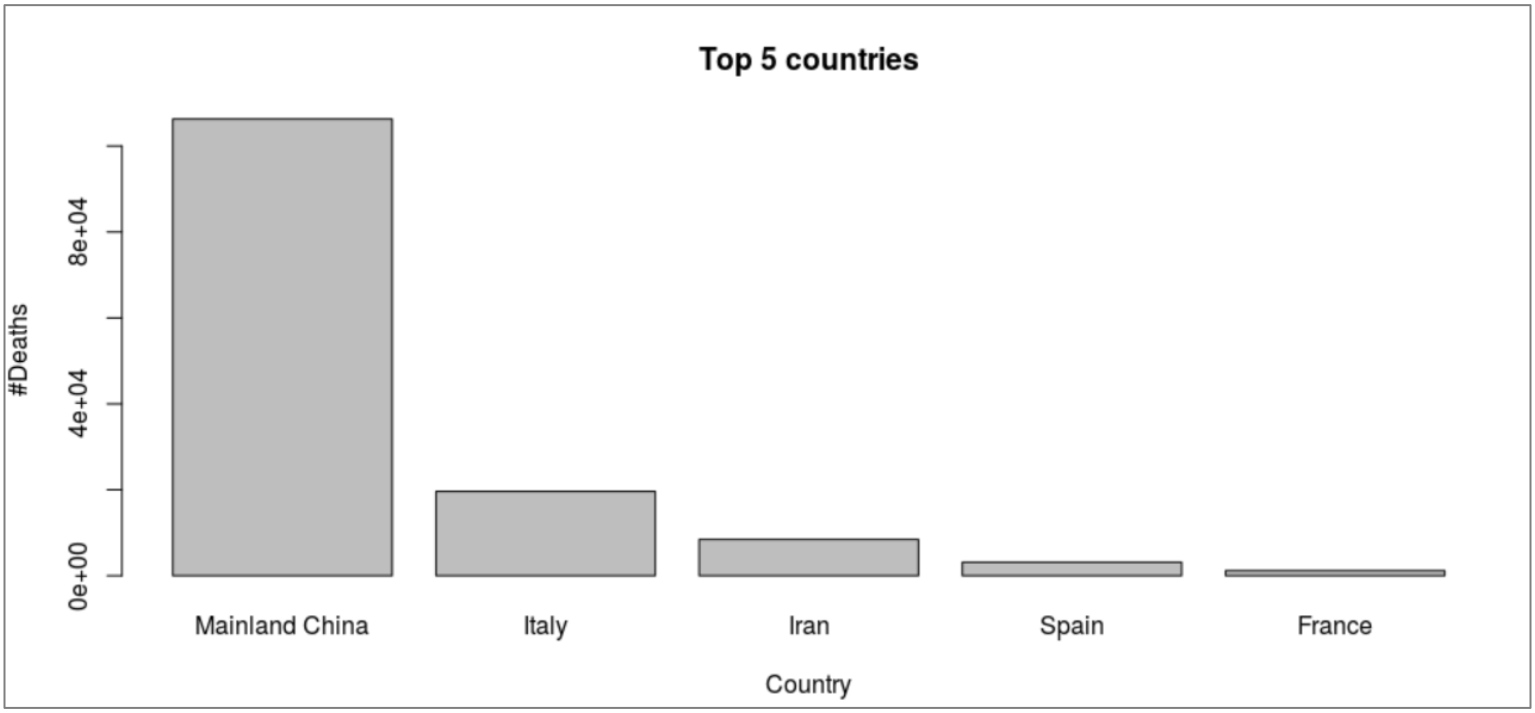

Fig (G): Top 5 countries having maximum death

Dataset:

Dataset is obtained from Kaggle Datasets repository. It has 7K observation with 8 columns. Attributes that are mostly used in this analysis are date and country.

Insights:

- With this dashboard we can identify curve of effect of corona virus started takin place from mid of January (Fig A). By that time confirmed cases, recovered cases and deaths were not so high.
- Though confirmed cases reach to 250K, 80K has recovered from the danger. However approximately 10K population has lost their lives (Fig A).

- When we look at the individual graph, it clears that the number of deaths has inflated by nearly 185% from March – 7and to make a count of 10K (Fig B)
- In case of confirmed cases count, though there were fluctuations during the period, but it is increased steadily. Where-as there was rapid growth from March – 7 and reached to 250K (Fig C).
- While the recovery rate was slow in the initial phase of virus outbreak, it went up rapidly after Feb – 15 to make total of 80K cases (Fig D).
- From cross analysis of confirmed and death couth we can estimate the more casualties. Deaths are directly proportional to confirmed numbers, and we can predict future casualties by this way (Fig E).
- We can identify the growth rate of total confirmed cases with the help of Growth rate graph (Fig F), it shows a fluctuation during the whole period, where-as the lowest rate was approximately 2%.
- In the end we can see top 5 countries suffered from death count issue listed on graph (Fig G).

Source code:

```
library(ggplot2)
library(lubridate)
library(scales)
library(grid)
library(gridExtra)
library(dplyr)

data <- read.csv(file.choose(), header = TRUE)
data$ObservationDate=as.Date(data$ObservationDate,"%m/%d/%y")


#Cumulative sum of all KPI
death_sum=aggregate(data$Deaths~data$ObservationDate,data=data,sum)
confirmed_sum=aggregate(data$Confirmed~data$ObservationDate,data=data,sum)
recovered_sum=aggregate(data$Recovered~data$ObservationDate,data=data,sum)

#death
a<-plot(death_sum$`data$ObservationDate`, (death_sum$`data$Deaths`),
        main="Daily Report",
        xlab="Update Dates ",
        ylab="Count ", pch=20, col='red',
        ylim=c(0,max(confirmed_sum$`data$Confirmed`)))
lm_death=lm((death_sum$`data$Deaths`)~death_sum$`data$ObservationDate`)
lm_death
abline(lm_death,col='red')

#confirmed
par(new=TRUE)
b<-plot(confirmed_sum$`data$ObservationDate`, (confirmed_sum$`data$Confirmed`),
        xlab="",
        ylab="",pch=20, col='blue',
        axes=FALSE)

par(new=TRUE)
lm_confi=lm((confirmed_sum$`data$Confirmed`)~confirmed_sum$`data$ObservationDate`)
abline(lm_confi,col='blue')

#Recovered
plot(recovered_sum$`data$ObservationDate`, (recovered_sum$`data$Recovered`),
     xlab="",
     ylab="", pch=20, col='green',
     ylim=c(0,max(confirmed_sum$`data$Confirmed`)),axes=FALSE)
lm_recov=lm((recovered_sum$`data$Recovered`)~recovered_sum$`data$ObservationDate`)
abline(lm_recov,col='green')
#legends
legend("topleft", legend=c("Death", "Confirmed","Recovered"),
       col=c("red", "blue","green"),pch=20, cex=1)
```

```
#Aggreate data with country level
death_con=aggregate(data$Deaths~data$Country.Region,data=data,sum)
confirmed_con=aggregate(data$Confirmed~data$Country.Region,data=data,sum)
recovered_con=aggregate(data$Recovered~data$Country.Region,data=data,sum)

#Top 5 analysis
death_con_sort=death_con[order(-death_con$`data$Deaths`),]
death_con_sort
top5cont=death_con_sort[c(1:5),]
top5cont
barplot(top5cont$`data$Deaths`,names.arg=top5cont$`data$Country.Region`,
        xlab="Country",ylab="#Deaths",main="Top 5 countries")

#grid.arrange(a,b,ncol=2,main="test")

#calculating growth rate of confirm cases to the previous day
#(curr - pre / pre) *100
data_prv=confirmed_sum[c(1:nrow(data)-1),]
data_nxt=confirmed_sum[c(2:nrow(data)),]
growthRateConfi=(((data_nxt$`data$Confirmed`-
data_prv$`data$Confirmed`)/data_prv$`data$Confirmed`)*100)
plot(data_nxt$`data$ObservationDate`,growthRateConfi,xlab = 'Dates',
     ylab = '%Growth rate',main="Growth rate of confirm cases",
     pch=20,col='green')
abline(lm(growthRateConfi~data_nxt$`data$ObservationDate`))

#calculating growth rate of recovry cases to the previous day
#(curr - pre / pre) *100
r_data_prv=recovered_sum[c(1:nrow(data)-1),]
r_data_nxt=recovered_sum[c(2:nrow(data)),]
r_growthRateConfi=(((r_data_nxt$`data$Recovered` -
r_data_prv$`data$Recovered`)/r_data_prv$`data$Recovered`)*100)
plot(r_data_nxt$`data$ObservationDate`,r_growthRateConfi,xlab = 'Dates',
     ylab = '%Growth rate',pch=20,col='Yellow')
abline(lm(growthRateConfi~data_nxt$`data$ObservationDate`))


#-----Regression Plot
par(mfcol=c(3,1))
plot(death_sum$`data$ObservationDate`, (death_sum$`data$Deaths`),
     main="Daily Deaths",
     xlab="Update Dates ",
     ylab="Count ", pch=20, col='red')
lm_death=lm((death_sum$`data$Deaths`)~death_sum$`data$ObservationDate`)
abline(lm_death,col='red')

#confirmed
plot(confirmed_sum$`data$ObservationDate`, (confirmed_sum$`data$Confirmed`),
     main="Daily Confirmed",
     xlab="Update Dates ",
     ylab="Count ",pch=20, col='blue')
lm_confi=lm((confirmed_sum$`data$Confirmed`)~confirmed_sum$`data$ObservationDate`)
abline(lm_confi,col='blue')

#Recovered
plot(recovered_sum$`data$ObservationDate`, (recovered_sum$`data$Recovered`),
     main="Daily Recovered",
     xlab="Update Dates ",
     ylab="Count ", pch=20, col='green')
lm_recov=lm((recovered_sum$`data$Recovered`)~recovered_sum$`data$ObservationDate`)

abline(lm_recov,col='green')


#relation between confirmed cases and deaths
```

```r
#linear model to expect casualties
confirmVsDeath=aggregate(.~data$ObservationDate,data=data,sum)
confirmVsDeath
plot(confirmVsDeath$Confirmed,confirmVsDeath$Deaths,
     main="Cross analysis of Confirmed vs Deaths",
     xlab="Confirm Cases ",
     ylab="Death Cases ", pch=20, col='darkgray')

expected_cas=lm(confirmVsDeath$Deaths~confirmVsDeath$Confirmed)
#expected_cas
abline(expected_cas,col='red',lty=2)
```