COVID19 Data Analysis With R Worldwide

Abdelrhman Mohamed Zaki

Dr. Mahmoud Hamed Abu Mossa

Faculty of Science, Department of Mathematics, Cairo University

June 11, 2020

1 INTRODUCTION

Coronaviruses are a group of related RNA viruses that cause diseases in mammals and birds. In humans, these viruses cause respiratory tract infections that can range from mild to lethal. Mild illnesses include some cases of the common cold (which is also caused by other viruses, predominantly rhinoviruses), while more lethal varieties can cause SARS, MERS, and COVID-19. Human coronaviruses were discovered in the 1960s, They were isolated using two different methods in the United Kingdom and the United States. E.C. Kendall, Malcom Byone, and David Tyrrell working at the Common Cold Unit of the British Medical Research Council in 1960 isolated from a boy a novel common cold virus B814. The virus was not able to be cultivated using standard techniques which had successfully cultivated rhinoviruses, adenoviruses and other known common cold viruses. Coronaviruses vary significantly in risk factor. Some can kill more than 30% of those infected, such as MERS-CoV, and some are relatively harmless, such as the common cold. Coronaviruses can cause colds with major symptoms, such as fever, and a sore throat from swollen adenoids.

In December 2019, a pneumonia outbreak was reported in Wuhan, China. On 31 December 2019, the outbreak was traced to a novel strain of coronavirus, which was given the interim name 2019-nCoV by the World Health Organization (WHO), later renamed SARS-CoV-2 by the International Committee on Taxonomy of Viruses. As of 4 June 2020, there have been at least 387,568 confirmed deaths and more than 6,563,099 confirmed cases in the COVID-19 pandemic.

The most common symptoms of COVID-19 are fever, dry cough, and tiredness. Other symptoms that are less common and may affect some patients include aches and pains, nasal congestion, headache, conjunctivitis, sore throat, diarrhea, loss of taste or smell or a rash on skin or discoloration of fingers or toes. These symptoms are usually mild and begin gradually. Some people become infected but only have very mild symptoms. Most people (about 80%) recover from the disease without needing hospital treatment. Around 1 out of every 5 people who gets COVID-19 becomes seriously ill and develops difficulty breathing. Older people, and those with underlying medical problems like high blood pressure, heart and lung problems, diabetes, or cancer, are at higher risk of developing serious illness. However, anyone can catch COVID-19 and become seriously ill. People of all ages who experience fever and/or cough associated withdifficulty breathing/shortness of breath, chest pain/pressure, or loss of speech or movement should seek medical attention immediately. If possible, it is recommended to call the health care provider or facility first, so the patient can be directed to the right clinic.COVID-19 is mainly spread through respiratory droplets expelled by someone who is coughing or has other symptoms such as fever or tiredness. Many people with COVID-19 experience only mild symptoms. This is particularly true in the early stages of the disease. It is possible to catch COVID-19 from someone who has just a mild cough and does not feel ill. Some reports have indicated that people with no symptoms can transmit the virus.

Data science can already provide ongoing, accurate estimates of health system demand, which is a requirement in almost all reopening plans. We need to go beyond that to a dynamic approach of data collection, analysis, and forecasting to inform policy decisions in real time and iteratively optimize public health recommendations for re-opening.

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.
So in this report, We will use the programming language R to make a simple data analysis on the COVID-19 data.

2 Loading Data

2.1 Data Source

This is an analysis report of the Novel Coronavirus (COVID-19) around the world, to demonstrate data processing and visualisation with R.
The data source used for this analysis is a collection of the COVID-19 data maintained by Our World in Data. It is updated daily and includes data on confirmed cases, deaths, and testing, as well as other variables of potential interest. We can access this data from (https://github.com/owid/covid-19-data/tree/master/public/data/).
This data has been collected, aggregated, and documented by Diana Beltekian, Daniel Gavrilov, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Edouard Mathieu, Esteban Ortiz-Ospina, Hannah Ritchie, Max Roser.

2.2 Loading Data

The easiest way to load data into memory in R is by using the R Studio menu items. R Studio has menu items for loading data in two different places. The first is in the toolbar of the upper right section of R Studio which says "Import Dataset", then you can select importing data from "From Excel File" and changing the type of dateRep variable from character to Date as ($\%Y\%m\%d$), and then importing the data.

2.3 Packages

Blow is a list of R packages used for this analysis. Package tidyverse is a collection of R packages for data science, and ggplot2 for graphics.

```
library(ggplot2)
library(tidyverse) # ggplot2, tidyr, dplyr...
```

3 Worldwide Cases

Now by using Rstudio, We will be able to visualise the data:

3.1 Number of Cases and Deaths Worldwide

At First, We need to visualise the worldwide data, we will focuse on the cases worldwide, where Figure 1 is a histogram between the time in days and the daily new number of (cases, deaths) due to COVID19 virus, which is result of this R code.

```
par(mfrow=c(1,2))
plot(covid1$dateRep,covid1$cases,
        xlab = "Time",ylab = "Number of Cases"
        ,main = paste("COVID19 Cases Worldwide"),type = "h")

plot(covid1$dateRep,covid1$deaths
        ,xlab = "Time",ylab = "Number of Deaths"
        ,main = paste("COVID19 Deaths Worldwide"),type = "h")
```

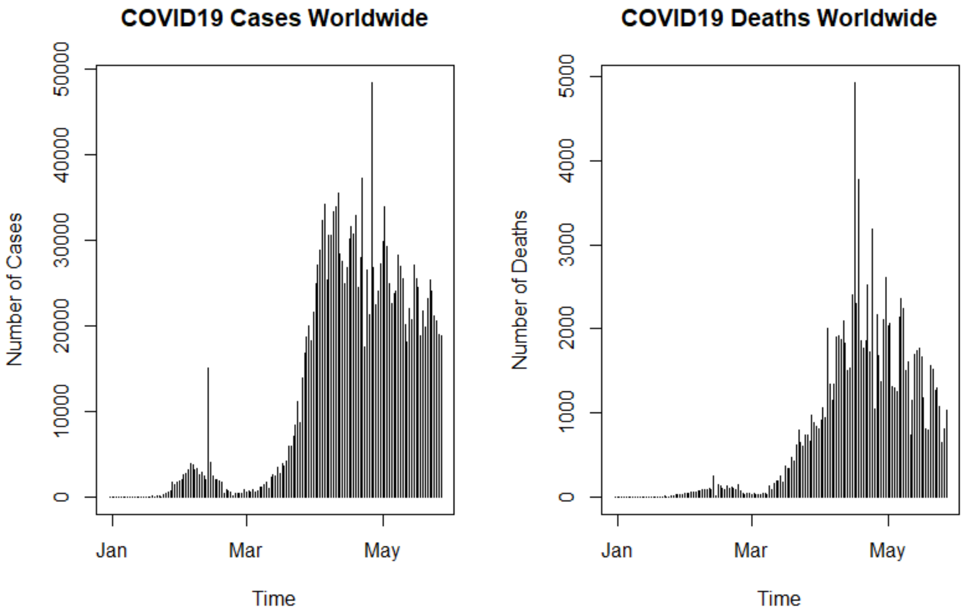

Figure 1: COVID19 Cases and Deaths Worldwide

3.2 Analysis for Every Single Country

Analysis for a single country can be done by filter the data with the corresponding country name. Here, we are going to define a function that takes the name of the country for every country in our data and then plot a line graph between the Time in days and the number of cases due to COVID19.

```
###Function which plots the number of cases in a given country.
case = function(coun){
  plot(subset(covid1$dateRep,covid1$countriesAndTerritories==coun)
       ,subset(covid1$cases,covid1$countriesAndTerritories==coun)
       ,type = ”l”,col = ”Red”
       ,main = paste(”COVID19␣in”,coun)
       ,xlab = ”Time”,ylab = ”No.of␣Cases”)

}
```

For example, if we try the ”case” function on a given country say ”Italy”, this line plot between the Time in days and the daily number of cases in ”Italy” due to COVID19 2 will be the result:

```
case(”Italy”)
```

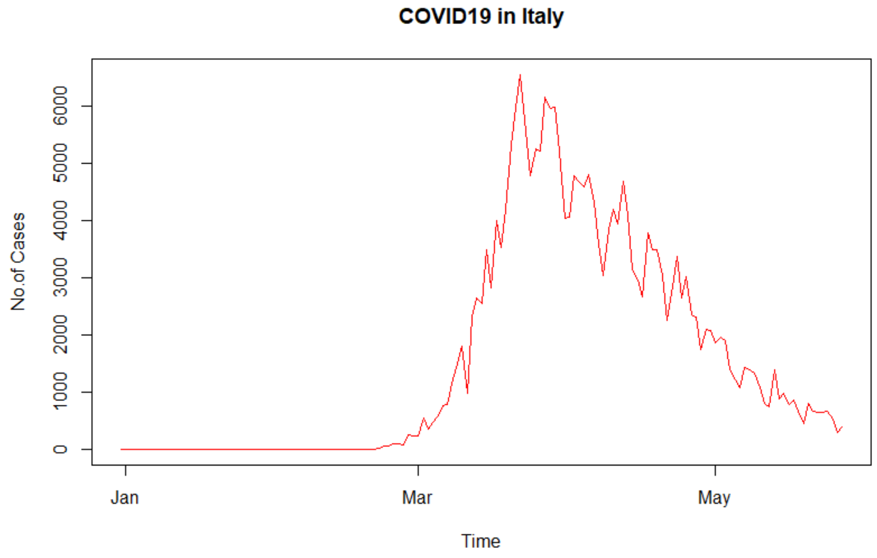

Figure 2: COVID19 Cases in Italy

We define another function that takes the name of the country for every country in our data and then plot a line graph between the Time in days and the number of deaths due to COVID19.

```
###Function which plots the number of deaths in a given country.
death = function(coun){
  plot(subset(covid1$dateRep,covid1$countriesAndTerritories==coun)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun)
       ,type = "l",col = "Blue",
       main = paste("COVID19␣in",coun)
       ,xlab = "Time",ylab = "No.of␣Deaths")

}
```

Also if we try the "death" function on a given country say "Italy", this line plot between the Time in days and the daily number of deaths in "Italy" due to COVID19 3 will be the result:

```
death("Italy")
```

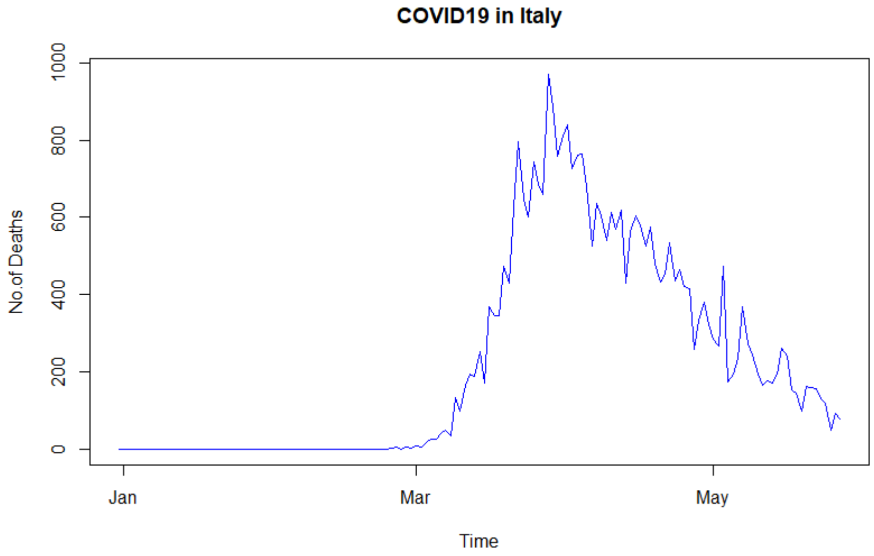

Figure 3: COVID19 Deaths in Italy

Then, we are going to define a function that plots a histogram of the number of cases and a linear graph of the number of deaths due to COVID19 in a certain country with the Time in days:

```
###Function plots the number of cases and number of deaths
 in a certain country.

casedeath = function(country){
   plot(subset(covid1$dateRep,covid1$countriesAndTerritories==country)
        ,subset(covid1$cases,covid1$countriesAndTerritories==country) ,
        main=paste(”COVID19␣cases␣and␣deaths␣in”,country),xlab = ”Time”
        ylab=”No.␣of␣casess␣and␣deaths”,
        type=”h”,
        col=”Blue4”)
   lines(subset(covid1$dateRep,covid1$countriesAndTerritories==country)
          ,subset(covid1$deaths,covid1$countriesAndTerritories==country)
   legend(”topleft”,
           c(paste(”No.of␣Cases␣in”,country)
             ,paste(”No.of␣Deaths␣in”,country)),
           fill=c(”Blue4”,”Brown”)
   )
}
```

For Example, if we try the ”casedeath” function on a given country say ”Italy”, this line plot between the Time in days and the daily number of deaths and plots also a histogram between the Time and the daily number of cases in ”Italy” due to COVID19 4 will be the result:

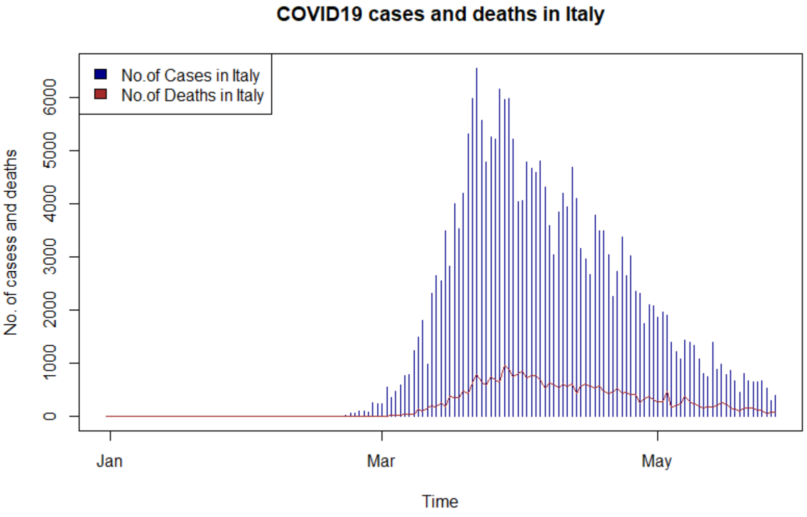

Figure 4: COVID19 Cases and Deaths in Italy

3.3 Top 10 Countries:

We have to arrange the top 10 countries by the most COVID19 cases, and then try to analyise the data of each country:

	Country	Confirmed	Deaths	Death rate
1	US	1681212	98916	5.88%
2	Brazil	391222	24512	6.26%
3	Russia	362342	3807	1.05%
4	UK	265227	37048	13.96%
5	Spain	236259	27117	11.47%
6	Italy	230555	32955	14.29%
7	France	145555	28530	15.6%
8	Germany	179364	8349	4.65%
9	Turkey	158762	4397	2.79%
10	India	151767	4337	2.85%

Now, we are going to make a barplot of the confirmed COVID19 cases of the top 10 countries, first of all, define three functions "totalcases" and "totaldeath" and "deathper", which obtain the total cases and total deaths and the death rate(%) of a given country respectively:

```
### two Functions calculate the total number of cases and
total number of deaths in a certain country.
totalcases = function(country){
  sum(subset(covid1$cases
             ,covid1$countriesAndTerritories==country),na.rm = T)
}
totaldeath = function(country){
  sum(subset(covid1$deaths
             ,covid1$countriesAndTerritories==country),na.rm = T)
}

###Function calculate the percentage of the death
according to the cases in each country.

deathper = function(country){
  sum(totaldeath(country)/totalcases(country),na.rm = T)
}
```

By using those functions, we will make barplot of the confirmed cases and deaths and the death rate of the top 10 countries 5:

```
H = c(totalcases("United_States_of_America"),totalcases("Brazil")
       ,totalcases("Russia"),totalcases("United_Kingdom")
       ,totalcases("Spain"),totalcases("Italy"),totalcases("France")
       ,totalcases("Germany"),totalcases("Turkey"),totalcases("India"))

Q = c(totaldeath("United_States_of_America"),totaldeath("Brazil")
       ,totaldeath("Russia"),totaldeath("United_Kingdom")
       ,totaldeath("Spain"),totaldeath("Italy"),totaldeath("France")
       ,totaldeath("Germany"),totaldeath("Turkey"),totaldeath("India"))

W = c(deathper("United_States_of_America"),deathper("Brazil")
       ,deathper("Russia"),deathper("United_Kingdom")
       ,deathper("Spain"),deathper("Italy"),deathper("France")
       ,deathper("Germany"),deathper("Turkey"),deathper("India"))

M = c("US","Brazil","Russia","UK","Spain"
       ,"Italy","France","Germny","Turkey","India")


par(mfrow=c(3,1))
barplot(H,names.arg = M, xlab = "Countries"
         ,ylab = "Number␣of␣Cases",col = "Green"
         ,main = paste("Top␣10␣Countries␣with␣Most␣Cases"))

barplot(Q,names.arg = M, xlab = "Countries"
         ,ylab = "Number␣of␣Deaths",col = "Red"
         ,main = paste("Top␣10␣Countries␣with␣Most␣Cases"))

barplot(W,names.arg = M, xlab = "Countries"
         ,ylab = "Death␣Rate($\%$)",col = "Blue"
         ,main = paste("Top␣10␣Countries␣with␣Most␣Cases"))
```

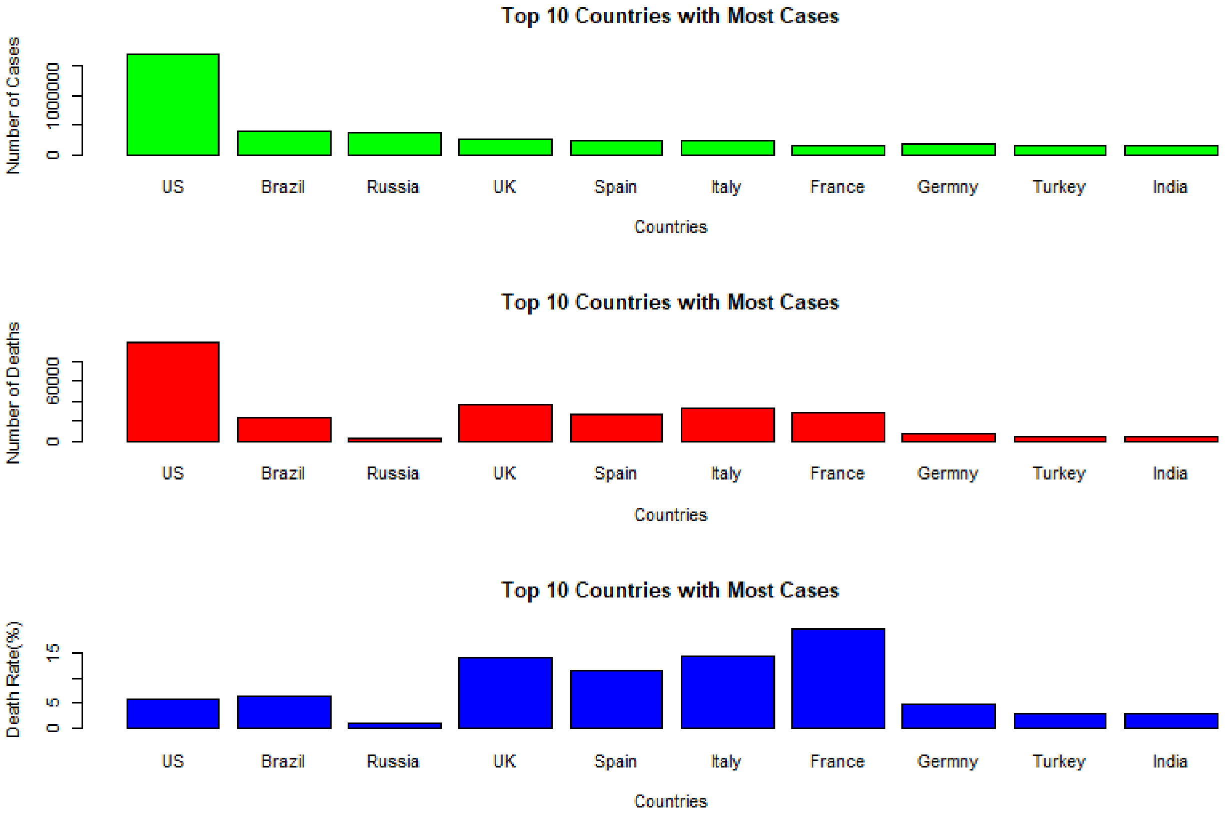

Figure 5: COVID19 Total Cases and Total Deaths and Death Rate(%) in Top 10 Countries

.

Here, we will define a function "pumkin" which plots linear graph of the daily number of cases in the top 10 countries with the time in days 6:

```
pumkin = function(coun1,coun2,coun3,coun4
                  ,coun5,coun6,coun7,coun8,coun9,coun10){
  plot(subset(covid1$dateRep,covid1$countriesAndTerritories==coun1)
       ,subset(covid1$cases,covid1$countriesAndTerritories==coun1) ,
       main=paste("COVID19␣in␣the␣top␣10␣countries"),xlab = "Time",
       ylab="No.␣of␣Cases",
       type="l",
       col="Hotpink")
```

```
lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun2)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun2)
       , col="Green4")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun3)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun3)
       , col="Red")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun4)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun4)
       , col="Blue")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun5)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun5)
       , col="Yellow")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun6)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun6)
       , col="Chocolate")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun7)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun7)
       , col="Black")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun8)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun8)
       , col="Darkorchid")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun9)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun9)
       , col="Brown")

lines(subset(covid1$dateRep, covid1$countriesAndTerritories==coun10)
       ,subset(covid1$cases, covid1$countriesAndTerritories==coun10)
       , col="Lightgreen")



 legend("topleft",
     c(paste("No. of␣Cases␣in",coun1),paste("No. of␣Cases␣in",coun2)
        ,paste("No. of␣Cases␣in",coun3),paste("No. of␣Cases␣in",coun4)
```

```
                ,paste("No. of Cases in",coun5),paste("No. of Cases in",coun6)
                ,paste("No. of Cases in",coun7),paste("No. of Cases in",coun8)
                ,paste("No. of Cases in",coun9),paste("No. of Cases in",coun10)
                , fill=c("Hotpink","Green4","Red","Blue","Yellow","Chocolate"
                    ,"Black","Darkorchid","Brown","Lightgreen")
    )
}
```

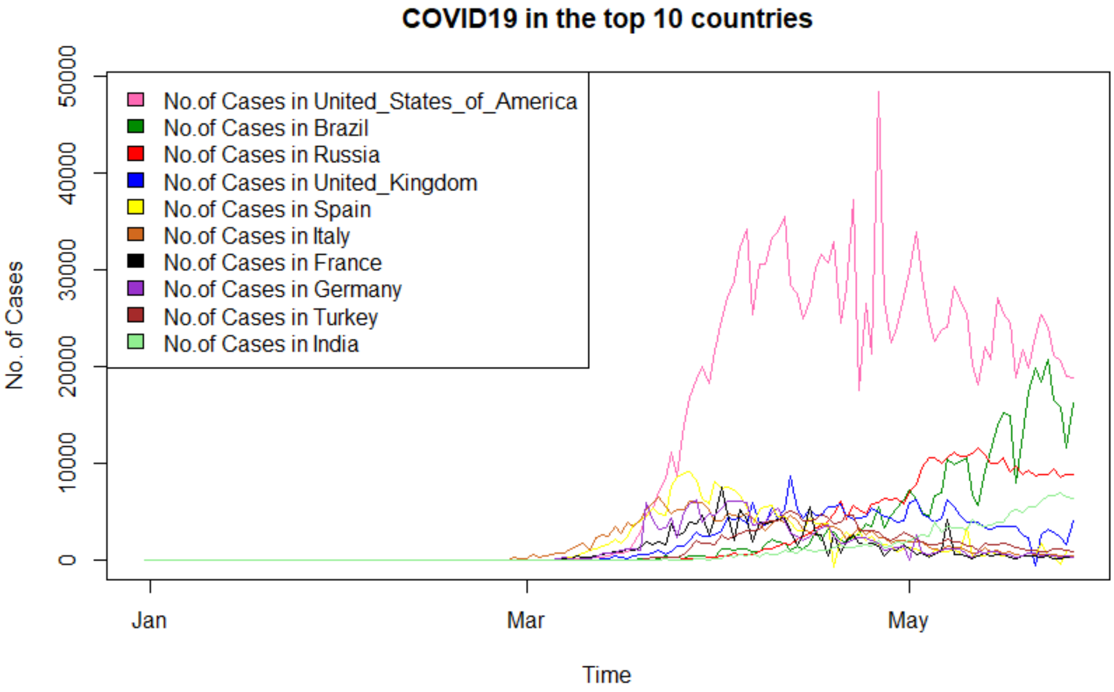

Figure 6: COVID19 Cases in Top 10 Countries

.

Also, we will define a function "topdeath" which plots linear graph of the daily number of deaths in the top 10 countries with the time in days 7:

```
topdeath = function(coun1,coun2,coun3,coun4
                    ,coun5,coun6,coun7,coun8
                    ,coun9,coun10){
```

```
plot(subset(covid1$dateRep,covid1$countriesAndTerritories==coun1)
      ,subset(covid1$deaths,covid1$countriesAndTerritories==coun1)  ,
      main=paste("COVID19␣in␣the␣top␣10␣countries"),xlab = "Time",
      ylab="No.␣of␣Deaths",
      type="l",
      col="Hotpink")
lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun2)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun2)
       ,  col="Green4")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun3)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun3)
       ,  col="Red")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun4)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun4)
       ,  col="Blue")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun5)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun5)
       ,  col="Yellow")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun6)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun6)
       ,  col="Chocolate")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun7)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun7)
       ,  col="Black")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun8)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun8)
       ,  col="Darkorchid")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun9)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun9)
       ,  col="Brown")

lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun10)
       ,subset(covid1$deaths,covid1$countriesAndTerritories==coun10)
       ,  col="Lightgreen")
```

```
  legend(”topleft”,
        c(paste(”No.of␣Deaths␣in”,coun1),paste(”No.of␣Deaths␣in”,coun2)
            ,paste(”No.of␣Deaths␣in”,coun3),paste(”No.of␣Deaths␣in”,coun4)
            ,paste(”No.of␣Deaths␣in”,coun5),paste(”No.of␣Deaths␣in”,coun6)
            ,paste(”No.of␣Deaths␣in”,coun7),paste(”No.of␣Deaths␣in”,coun8)
            ,paste(”No.of␣Deaths␣in”,coun9),paste(”No.of␣Deaths␣in”,coun10)
              fill=c(”Hotpink”,”Green4”,”Red”,”Blue”,”Yellow”,”Chocolate”
                       ,”Black”,”Darkorchid”,”Brown”,”Lightgreen”)
    )
}
```

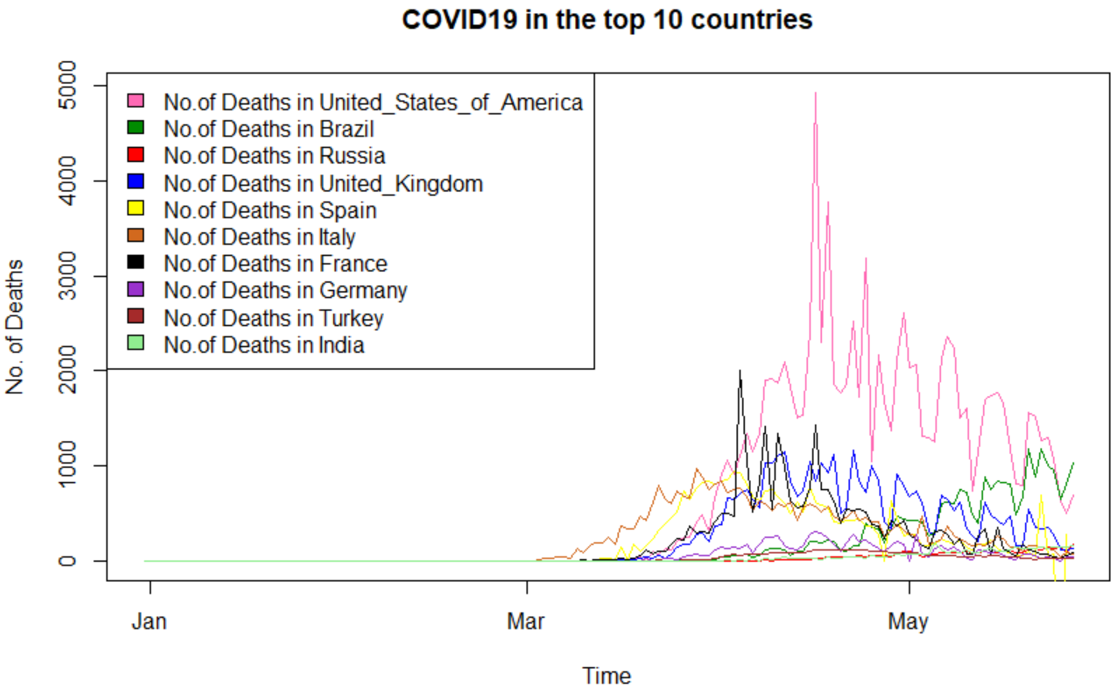

Figure 7: COVID19 Deaths in Top 10 Countries

Finally, we are going to define two functions "summ1" and "summ2" which obtain the descriptive statistics for the number of daily cases and number of daily deaths of a given country:

```
summ1 = function(coun){
  summary(subset(covid1$cases,covid1$countriesAndTerritories==coun))
}

summ2 = function(coun){
   summary(subset(covid1$deaths,covid1$countriesAndTerritories==coun))
 }
```

For Example, if we try to use the functions "summ1" and "summ2" to obtain the descriptive statistics for the number of daily cases and number of daily deaths of "United States of America", then:

```
summ1("United_States_of_America")
```

The descriptive ststistics of the daily number of cases of United States of America will be:
$Min. = 0$, $1stQu. = 0$, $Median = 511$, $Mean = 11283$, $3rdQu. = 24247$, $Max. = 48529$.

```
summ2("United_States_of_America")
```

The descriptive ststistics of the daily number of deaths of United States of America will be:
$Min. = 0$, $1stQu. = 0$, $Median = 7$, $Mean = 663.9$, $3rdQu. = 1317.0$, $Max. = 4928.0$.

4 Analysis for the COVID19 data in Egypt:

We need to analyze the COVID19 data in Egypt, to help the decision makers in our nation to make decisions about the social distancing and the quarantine procedures.

At first, we will plot a histogram of the daily number of cases and the daily number of deaths with the time in days using the functions "case" 8 and "death" 9 repectively, that we have defined earlier:

```
case("Egypt")
```

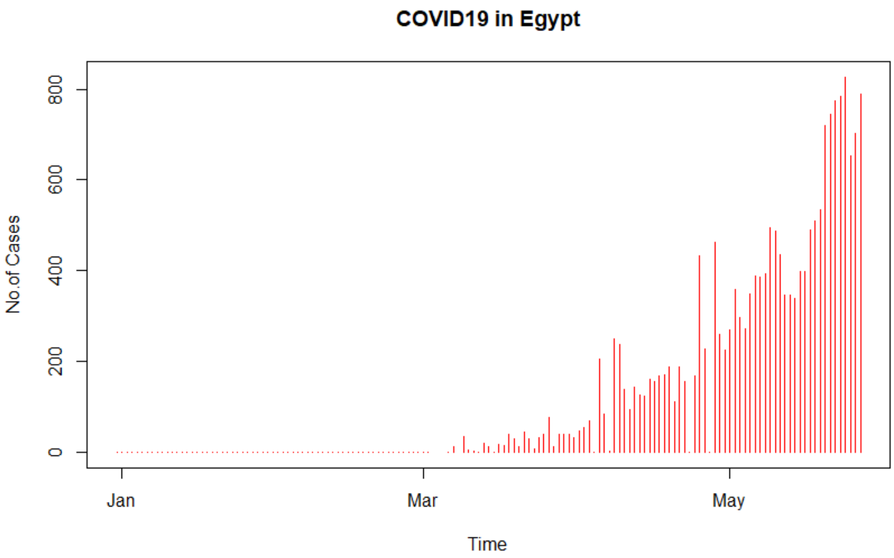

Figure 8: COVID19 daily cases in Egypt

```
death("Egypt")
```

The two graphs 8,9 showed that the number of cases and the number of deaths are still in continuous increase and that they haven't reached to the peak (highest point), so in this phase, all we need to care about is the social distancing rules and the quarantine rules for the old people and the kids in order to control the spread and to decrease the number of deaths and to decrease the pressure on the workers in the hosiptals and the medical establishments.

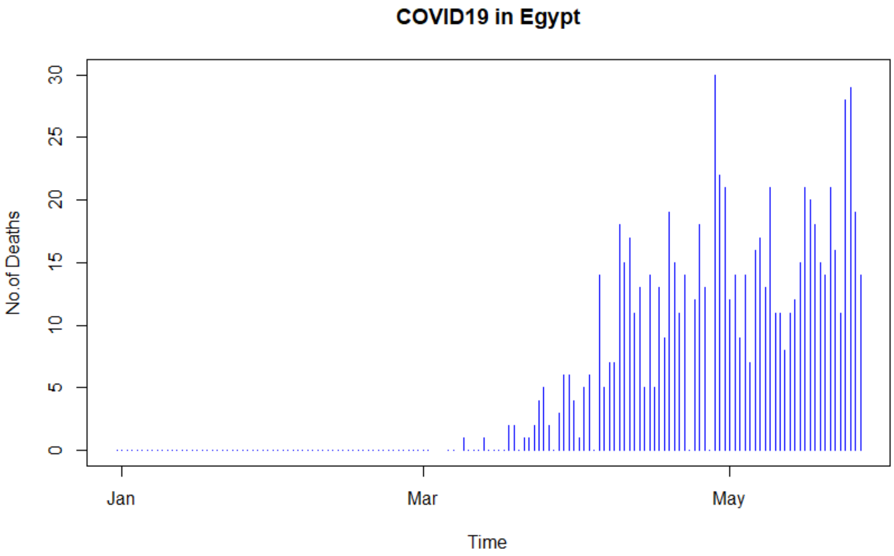

Figure 9: COVID19 daily deaths in Egypt

Now, we will use the "casedeath" function to plot the number of cases and the number of deaths with the time in days 10:

```
casedeath("Egypt")
```

Now, we will use the functions "summ1" and "summ2" to find the descriptive statistics of the daily number of cases and the daily number of deaths in Egypt:

```
summ1("Egypt")
```

The descriptive statistics of the daily number of cases in Egypt:
$Min. = 0, 1stQu. = 0, Median = 9, Mean = 129.4, 3rdQu. = 188, Max. = 827$

```
summ2("Egypt")
```

The descriptive statistics of the daily number of cases in Egypt:
$Min. = 0, 1stQu. = 0, Median = 0, Mean = 5.497, 3rdQu. = 11, Max. = 30$.

Now, we will use the function "correlation" to find the linear correlation coefficient between the number of cases and the number deaths in Egypt:

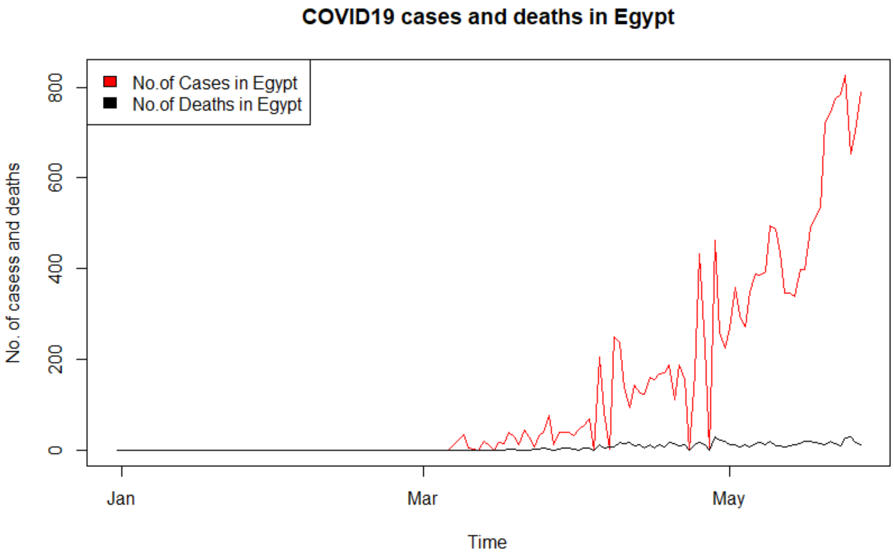

Figure 10: COVID19 cases and deaths in Egypt

```
corrleation("Egypt")
```

correlation("Egypt") = 0.8325892.

The correlation coefficient in Egypt is high due to the lack of medical service in the village and in the absence of medical service in the discrete places, also because of the lack of the social distancing procedures in the crowded public places.

we will use the function "deaper" to find the death percentage in Egypt:

```
deaper("Egypt")
```

Death rate in Egypt = 4.24%, which is little bit high for the daily numbers of cases and also refer to the lack of medical service.

At last, we need to follow the data to control the spread of the virus and to save the people lives with the social distancing and the quarantine rules specially for the older citzens, we also need to follow the data in the reopening procedures.

5 Comparison between Countries

In this section, we will define functions to plot the number of cases (deaths) of two countries or more, and to test the hypothesis of the difference between the means of the number of cases (deaths) of two countries or more, and to test the hypothesis of the ratio of variances of the number of cases (deaths) of two countries.

1) The function ”comp” that is defined to plot a linear graph of the number of daily cases of two given countries with time in days:

```
###Function plot the number of cases in two given countries.
comp = function(coun1,coun2){
  plot(subset(covid1$dateRep,covid1$countriesAndTerritories==coun1)
       ,subset(covid1$cases,covid1$countriesAndTerritories==coun1)
       ,main=paste("COVID19␣in",coun1,"and",coun2),xlab = "Time"
       ,ylab="No.␣of␣casess"
       ,type="l"
       ,col="Blue")
  lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun2)
        ,subset(covid1$cases,covid1$countriesAndTerritories==coun2)
        ,  col="Red")
  legend("topleft"
       ,c(paste("No.of␣Cases␣in",coun1),paste("No.of␣Cases␣in",coun2))
        ,fill=c("Blue","Red")
  )
}
```

For Example, if we try to input the two countries in the R code as ”United States of America” and ”Italy”, then the result will be a linear graph of the number of daily cases of two given countries with time in days 11:

```
comp("United_States_of_America","Italy")
```

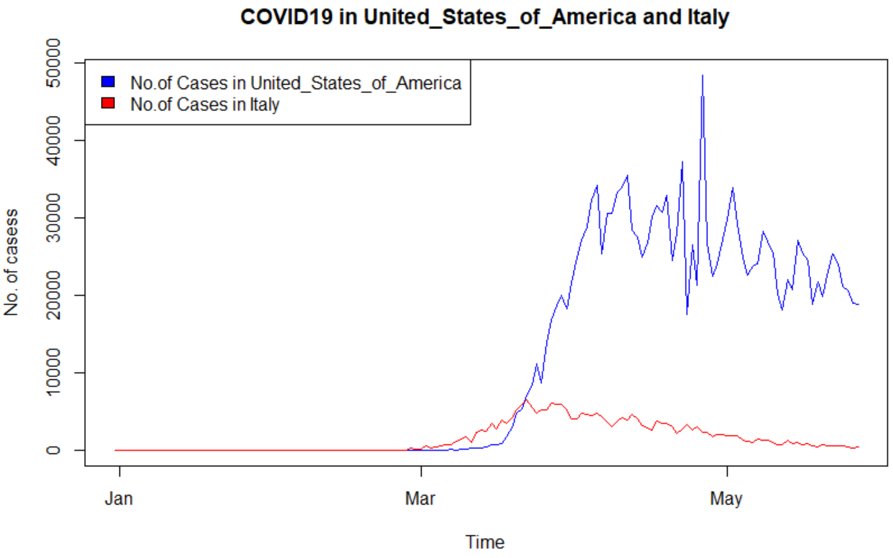

Figure 11: COVID19 Cases in US and Italy

2)The function "compdeath" that is defined to plot a linear graph of the number of daily deaths of two given countries with time in days:

```
###Function plot the number of deaths in two given countries.
compdeath = function(coun1,coun2){
  plot(subset(covid1$dateRep,covid1$countriesAndTerritories==coun1)
          ,subset(covid1$deaths,covid1$countriesAndTerritories==coun1)
          ,main=paste("COVID19␣in",coun1,"and",coun2),xlab = "Time"
          ,ylab="No.␣of␣deaths"
          ,type="l"
          ,col="Hotpink")
  lines(subset(covid1$dateRep,covid1$countriesAndTerritories==coun2)
           ,subset(covid1$deaths,covid1$countriesAndTerritories==coun2)
           , col="Green4")
  legend("topleft"
            ,c(paste("No.of␣Deaths␣in",coun1)
            ,paste("No.of␣Deaths␣in",coun2))
            ,fill=c("Hotpink","Green4")
  )
}
```

For Example, if we try to input the two countries in the R code as "United

States of America” and ”Italy”, then the result will be a linear graph of the number of daily deaths of two given countries with time in days 12:

```
compdeath(”United_States_of_America”,”Italy”)
```

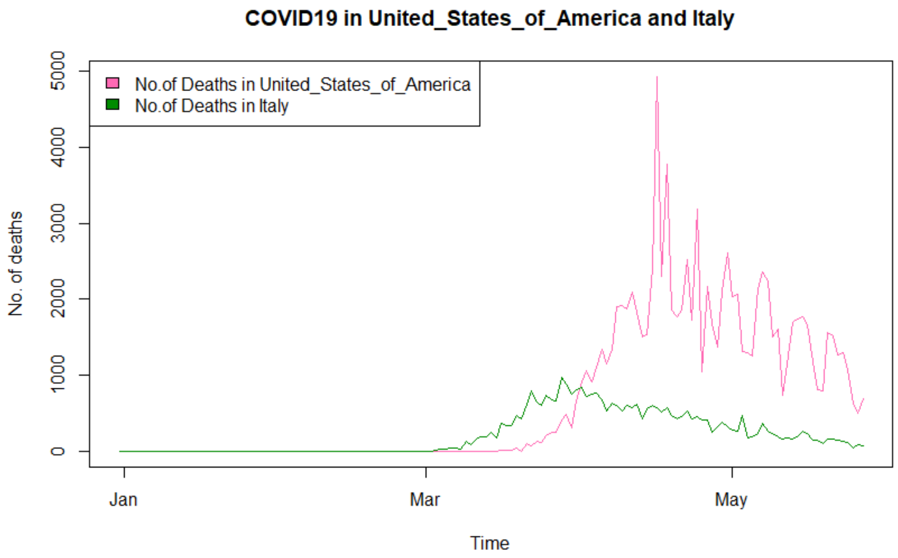

Figure 12: COVID19 Deaths in US and Italy

As previous, we can define a function that graphs the number of cases (deaths) of three countries or more with time in days.

3) The function ”meancom” defined by the following R code tests the hypothesis of difference between means of number of cases in two given countries and given alternative hypothesis whether ($\mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$) which is named (”two.sided” or ”greater” or ”less”) respectively and whether the two variances are equal or not:

```
###Function tests the hypothesis of difference between means
 of number of Deaths in two given countries and which given
 condition,and the variance of them is either equal or not.

meancom = function(coun1,coun2,cond,logiic){
   t.test(subset(covid1$cases,covid1$countriesAndTerritories==coun1)
            ,subset(covid1$cases ,covid1$countriesAndTerritories==coun2)
            ,alternative = cond,var.equal = logiic)
}
```

To test the hypothesis of the difference in the means of the daily number of cases in "Qatar" and "Egypt" using "meancom", where $H_0 : \mu_Q = \mu_E$ and the alternative hypothesis $H_A : \mu_Q \neq \mu_E$, and the variances of them are equal, so alternative = "two.sided" and var.equal = T, which is:

```
meancom("Qatar","Egypt","two.sided",T)
```

The result is:

Two Sample t-test

Data: No.of Cases in Qatar and No.of Cases in Egypt.
$t = 4.1817$, $df = 288$, $p - value = 3.843e - 05$.

Alternative hypothesis: true difference in means is not equal to 0.

95 percent confidence interval: [103.8607 , 288.5669].

Sample estimates: mean of x = 325.5655, mean of y = 129.3517.

Since, $p - value < 0.05$, so we reject H_0 and accept H_A as the two means is not equal.

4) The function "meandeathcom" defined by the following R code tests the hypothesis of the difference of means of the number of deaths in two given countries where the null hypothesis is that the means of number of deaths of the two given countries ($H_0 : \mu_1 = \mu_2$) against alternative hypotesis is variable which is ($H_A : \mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$) which is named ("two.sided" or "greater" or "less") respectively and whether the variances of the number of deaths of the two given countries is equal or not (When variance are equal,we say var.equal = T, and when they aren't, we say var.equal = F), and with level of significance $\alpha = 0.05$:

```
meandeathcom = function(coun1,coun2,cond,logiic){
  t.test(subset(covid1$deaths ,covid1$countriesAndTerritories==coun1)
            ,subset(covid1$deaths ,covid1$countriesAndTerritories==coun2)
            ,alternative = cond,var.equal = logiic)
}
```

To test the hypothesis of the difference in the means of the daily number of deaths in "Afghanistan" and "Zimbabwe" using "meancom", where $H_0 : \mu_Q = \mu_E$ and the alternative hypothesis $H_A : \mu_Q \neq \mu_E$, and the variances of them are equal, so alternative = "two.sided" and var.equal = F, which is:

```
meandeathcom("Afghanistan","Zimbabwe","two.sided",F)
```

Welch Two Sample t-test
Data: No.of Deaths in Afghanistan and No.of Deaths in Zimbabwe.

t = 4.8996, df = 140.37, p-value = 2.611e-06.

Alternative hypothesis: true difference in means is not equal to 0.

5 percent confidence interval: [1.504372 , 1.543449].

Sample estimates: mean of x = 1.58273381 mean of y= 0.05882353.

Since, $p-value < 0.05$, so we reject H_0 and accept H_A as the two means is not equal.

5) The function "compvar" defined by the following R code tests the hypothesis of the ratio between the variances of number of cases in two given countries where the null hypothesis is that the ratio between the variances of the number of cases of the two given countries equals ($H_0 : \sigma_1^2/\sigma_2^2 = 1$) against alternative hypotesis is variable which is ($H_A : \sigma_1^2/\sigma_2^2 \neq 1$ or $\sigma_1^2/\sigma_2^2 > 1$ or $\sigma_1^2/\sigma_2^2 < 1$) which is named ("two.sided" or "greater" or "less") respectively and with level of significance $\alpha = 0.05$:

```
compvar = function(country1,country2,cond){
var.test(subset(covid1$cases,covid1$countriesAndTerritories==country1)
         ,subset(covid1$cases,covid1$countriesAndTerritories==country2)
          ,ratio = 1,alternative = cond,conf.level = 0.05)

}
```

In this test, we use the statistic $F = (\frac{S_1^2}{\sigma_1^2})/(\frac{S_2^2}{\sigma_2^2})$, this test shows how far the data of number of cases of a given countries is spread out comparing with how far the data of number of cases of another country is spread out, does both number of cases of the two countries spread out equally or there is a one greater than or less than the other.

For Example, we are going to choose two countries, say ”Italy” and ”Spain” and choose the alternative hypothesis to be ”greater” that is ($H_A : \sigma_1^2/\sigma_2^2 > 1$);

```
compvar(”Italy”,”Spain”,”greater”)
```

F test to compare two variances
Data: No.of Cases in Italy and No.of Cases in Spain.

F = 0.60314, num df = 148, denom df = 147, p-value = 0.9989.

Alternative hypothesis: true ratio of variances is greater than 1.

5 percent confidence interval: [0.7914212 , ∞[.

Sample estimates: ratio of variances = 0.603144.

Since, $p-value > 0.05$, so we reject H_A as the two ratio between the two variances is less than 1, and accept H_0.

6) The function ”compvar2” defined by the following R code tests the hypothesis of the ratio between the variances of number of deaths in two given countries where the null hypothesis is that the ratio between the variances of the number of deaths of the two given countries equals ($H_0 : \sigma_1^2/\sigma_2^2 = 1$) against alternative hypotesis is variable which is ($H_A : \sigma_1^2/\sigma_2^2 \neq 1$ or $\sigma_1^2/\sigma_2^2 > 1$ or $\sigma_1^2/\sigma_2^2 < 1$) which is named (”two.sided” or ”greater” or ”less”) respectively and with level of significance $\alpha = 0.05$:

```
compvar2 = function(country1,country2,cond){
var.test(subset(covid1$deaths,covid1$countriesAndTerritories==country1
        ),subset(covid1$deaths,covid1$countriesAndTerritories==country2)
           ,ratio = 1,alternative = cond,conf.level = 0.05)
}
```

For Example, we are going to choose two countries, say ”Italy” and ”Spain” and choose the alternative hypothesis to be ”less” that is $(H_A : \sigma_1^2/\sigma_2^2 < 1)$;

```
compvar2("Italy","Spain","less")
```

F test to compare two variances:

Data: No.of Deaths in Italy and No.of Deaths in Spain.

F = 0.65878, num df = 148, denom df = 147, p-value = 0.005837.

Alternative hypothesis: true ratio of variances is less than 1.

5 percent confidence interval: [0 , 0.5019839].

Sample estimates: ratio of variances = 0.6587786.

Since, $p - value < 0.05$, so we reject H_0 and accept H_A as the two ratio between the two variances is less than 1.

5) The function ”correlation” defined by the following R code obtain the linear correlation coefficient between the number of cases and the number of deaths due to COVID19 for a given country:

```
correlation = function(country){
  cor(subset(covid1$deaths, covid1$countriesAndTerritories==country)
        ,subset(covid1$cases, covid1$countriesAndTerritories==country)
                    ,method = c("pearson", "kendall", "spearman"))
}
```

For Example, we are going to try the function for ”United States of America” and for ”Norway”:

```
correlation("United_States_of_America")
correlation("Norway")
```

correlation(”United States of America”) = 0.8741348.
Since, the numbers of cases is too high in US, so the correlation coefficient is high, it doesn’t mean that the medical facilities aren’t successful but it

means that the quarantine procedures done a little bit late, which leads to the high numbers in the cases and deaths.

correlation("Norway") = 0.3965448.
While, in Norway, the correlation coefficient is low, this means that the hospitals and the medical staff and the quarantine procedures are successful, and the deaths only occur for the older and sick people.

6 Observations:

From the graphs of the daily number of cases of each country and the time in days, it shows that the number of cases are in continuous increase at the most of the countries, some other countries manage to control the spread and already began in decrasing the number of cases due to the high of the recovery rate because of the improved medical service and the social distancing rules, these countries will begin in the reopening plan with keeping the social distancing rules.
The number of the daily deaths is correlated with the number of the cases but alot of countries manage to decrease the death rate with the social distancing and the medical service, but in Europe, which most of it's countries have high death rates due to the aging, in 2016, 19.2% of the EU population was aged 65 or over. the share of the elderly in the population differs considerably between Member States, in 2016, the highest share was recorded in Italy (22.0%) and the lowest in Ireland (13.2%) and the COVID19 virus is more dangerous for the old people and the people with the chronic diseases which leads to the countries of the lack in the medical service which have high death rate, when we showed the top 10 countries in the total number of cases, it turns that the Us which is the highest country in both number of cases and deaths doesn't have the highest death rate.

From the tests of hypothesis, we tested the difference in the means of the both number of cases and number of deaths of any two countries which shows that the means isnot the same what so ever, and we have tested the ratio between the variances of both number of cases and number of deaths of two given countries which shows how far the data of number of cases of a given countries is spread out comparing with how far the data of number of cases of another country is spread out, does both number of cases of the two countries spread out equally or there is a one greater than or less than the other, then we showed the linear correlation coefficient of any country between the number of cases and the number of deaths to show how the social distancing and the quarantine policy of every country and how some countries have a negative linear correlation coefficient between the number of cases and number of deaths like the Northern Mariana Islands.

7 Conclusion

At last, we showed a simple analysis for the COVID19 patients data, to help us understand the virus and to know the ways to stop the spread, the most important thing to beat the virus is to follow the data which tells us when to reopen. I think that the solution to stop the numbers of cases and deaths from growing is in the quarantine and social distancing both in the rich and poor countries, and to support the medical system including the hospitals and clincs across the country in a financial way to support the citzens in the village and in the discrete areas across the country and to support the medical system employees in the financial way to encourage them to be in the front line facing the dangar.

The Data is what tells us when to reopen.