Exploratory Data Analysis using R

name: Dimitrios

surname: Zaridis

email: dimzaridis@gmail.com

student number: 03002986

student: PhD student

1.Abstract

In 2020 the world have faced a threat that noone expected to happen.Along with the economic crisis, the COVID-19 virus outbreak ,which started from China and has reached out the whole world, is a memorial incident of that year and that incident prevailed of any other phenomenon that took place in that year.Meanwhile, The human species has been developing new means and tools for the confrontation of the COVID-19 virus and Statistics have played a huge role since the insights they offer are of a tremendous value.A variety of WHO(World Health Organization) strategies have been created based on results extracted from statistical analysis and exploitation of the data gathered since the beginning of the outbreak.In this work, data from this source and this source have been exploited in order to extract valuable information and patterns about the COVID-19 virus.It is necessary to mention that data from this source have also been used in order to create some categorical variables and extract several other information apart from the original time-series dataset invoving spatial patterns such as density(people/km^2) and total population of a country which is crucial measurement for the calculation of relative confirmed cases and deaths.This matter of relativity will be discussed also in this work.

2.Data discussion

In this particular section some details are going to be given about the data that have been used in this work.Specifically the main datasets consist of data tables with the cumulative number of deaths and cumulative number of confirmed cases daily for 272 countries and 362 days starting from 22-01-2020 and ending at 14-01-2021.It has to be mentioned that the analysis ,apart from the main tasks, has been done for dates of the year 2020.Therefore the original datasets are a time-series datasets and they are referring to the total number of deaths for a country until a particular date, so the last date indicates the total number of deaths for this country until that date.Below is a glimpse of how the original data look like.

the knitr(kable) library 's being used in order to plot data.table's outputs nicely, not for the modification of the data tables

Table 1: Daily cumulative confirmed cases per country

Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20
	Afghanistan	33.93911	67.70995	0	0	0
	Albania	41.15330	20.16830	0	0	0
	Algeria	28.03390	1.65960	0	0	0

Table 2: Daily cumulative death cases per country

Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20
	Afghanistan	33.93911	67.70995	0	0	0
	Albania	41.15330	20.16830	0	0	0
	Algeria	28.03390	1.65960	0	0	0

The complementary dataset that it has been used in order to combine information includes median age,density(people/km^2),population etc per country, while the “countrycode” R language library helps for the matching of a country to the continent it belongs.Further explanation and discussion about the combined tables are going to be given in the next chapters.Some modifications needed to be done in order to match several countries as for instance, in the original datasets “United states” are named as US and in the complementary as “US”.Therefore some names from the complementary dataset have been changed in order to match to the originals.

2.Initial tasks

2.1 Removal of Province,State,Lat and long columns

disclaimers:
1. *In order to save some space , not all the tables are going to be shown*
2. *covid_deaths & covid_confirmed are the original datasets*

```
selected_deaths <- covid_deaths[,c("Province/State","Lat","Long"):=NULL]
selected_confirmed <- covid_confirmed[,c("Province/State","Lat","Long"):=NULL]
```

2.2 & 2.4 wide to long format & name deaths and confirmed

```
selected_deaths_long <- melt(selected_deaths,id.vars = c("Country/Region"),
                           variable.name = "Date",value.name ="deaths")
selected_confirmed_long <- melt(selected_confirmed,id.vars = c("Country/Region"),
                           variable.name = "Date",value.name ="confirmed")
```

2.3 Rename country/region to country

```
setnames(selected_deaths_long,c("Country/Region"),c("Country"))
setnames(selected_confirmed_long,c("Country/Region"),c("Country"))
```

2.5 Date conversion to the asked format dd/mm/YYYY

Argument format inside the as.Date R function indicates how the date need to be read from the interpreter (“%month/%day/%year without century”) while the second argument of format (“%day/%month/%YEAR with century”) indicates the format we would like to rewrite the date

```
selected_deaths_long[,Date:=format(as.Date(selected_deaths_long[,Date],
                                          format="%m/%d/%y"),"%d/%m/%Y")]
selected_confirmed_long[ ,Date:=format(as.Date(selected_confirmed_long[,Date],
                                             format="%m/%d/%y"),"%d/%m/%Y")]
```

2.6 Grouped by Country and Date

```
selected_deaths_long <- selected_deaths_long[,lapply(.SD,sum),
                                           by=c("Country","Date"),
                                           .SDcols=c("deaths")]
selected_confirmed_long <- selected_confirmed_long[,lapply(.SD,sum),
```

Table 3: Daily cumulative confirmed cases per country long format,renamed

Country	Date	confirmed	Country	Date	deaths
Afghanistan	22/01/2020	0	Afghanistan	22/01/2020	0
Albania	22/01/2020	0	Albania	22/01/2020	0
Algeria	22/01/2020	0	Algeria	22/01/2020	0

```
                                                  by=c("Country","Date"),
                                                  .SDcols=c("confirmed")]
```

2.7 Merging the datasets into one

The merging of the two data tables are going to be implemented by the data.table function **merge** by country and date, since those variables are included in both of the datasets.Afterwards we need to convert Date from character type to Date type.The conversion in the character type was necessary in order to bring it in the requested form as Date format is depicted as (mm-dd-yy or mm-dd-YYYY).Eventually the conversion into the Date type again is critical for the right grouping and measurements(Character type dates are not compared in the same way as Date type so the grouping will be wrong)

```
total <- merge.data.table(selected_confirmed_long,selected_deaths_long,
                          by=c("Country","Date"))
```

2.8 Calculating the global numbers for deaths and confirmed daily cases

```
total_global_d <- total[,lapply(.SD,sum),by=c("Date"),.SDcols=c("deaths")]
total_global_c <- total[,lapply(.SD,sum),by=c("Date"),.SDcols=c("confirmed")]
total_global <- merge.data.table(total_global_c,total_global_d,by=c("Date"))
```

Table 4: Global numbers for deaths and confirmed cases per day

Date	confirmed	deaths
22/01/2020	557	17
23/01/2020	655	18
24/01/2020	941	26

Table 5: Random rows of the data table to evaluate the consistency of the above code

Country	Date	confirmed	deaths	deaths.inc	confirmed.ind
Antigua and Barbuda	13/08/2020	92	3	0	0
Antigua and Barbuda	14/08/2020	93	3	0	1
Antigua and Barbuda	15/08/2020	93	3	0	0
Antigua and Barbuda	16/08/2020	93	3	0	0

Country	Date	confirmed	deaths	deaths.inc	confirmed.ind
Belgium	03/10/2020	127623	10044	7	3389
Belgium	04/10/2020	130235	10064	20	2612
Belgium	05/10/2020	132203	10078	14	1968
Belgium	06/10/2020	134291	10092	14	2088

2.9-2.10 Grouping by country and date-Creation of two new variables for the daily deaths and confirmed cases

In the chunk below firstly total data table is grouped by country and date for the task 9. Sequently,data.table function **shift** has been used in order to find the difference table.element[i]-table.element[i-1] which are *daily confirmed cases* and *deaths* for each country.The daily indicators turn to zero when the country change so the calculation of the requested indexes have been made respecting the *Country* variable.The results are shown in the table 5 above

```
total <- total[,lapply(.SD,sum),by=c("Country","Date"),
               .SDcols=c("confirmed","deaths")]
total <- total[, deaths.inc := deaths - shift(deaths, fill = first(deaths)),
             by = .(Country)]
total <- total[, confirmed.ind := confirmed - shift(confirmed,fill = first(confirmed)),
             by = .(Country)]
```

3. Addition of new variables and categories

3.1 Quarters variable

In this section of the work, our focus will target on the creation of several new variables that they will assist in the analysis and extraction of various insights.Furthermore the complementary dataset will be used in order to add variables such as *density* and *median age* into hlink.Firstly the addition of the categorical variable **quarter** will take place, which divides the year into 4 quarters by adding an extra column . Also the measurements from 2020 are going to be kept from the table,2021 measurements are dropped.

```
total <- total[year(Date)==2020]

total <- total[month(Date)>0 & month(Date)<4 & year(Date)==2020,
               quarter:="1st quarter"]

total <- total[month(Date)>3 & month(Date)<7 & year(Date)==2020,
               quarter:="2nd quarter"]
```

```
total <- total[month(Date)>6 & month(Date)<10 & year(Date)==2020,
               quarter:="3d quarter"]

total <- total[month(Date)>9 & month(Date)<13 & year(Date)==2020,
               quarter:="4th quarter"]
```

3.2 Period variable

Next, the categorical variable **period** is going to be added.The value that variable takes can be either *pre-summer* and *post-summer* and we are going to do some aggregations based on that variable such as the cumulative sum to compare how the world treated Covid-19 before and after the summer period which is also vacation perio and a plethora of individuals travel.*(summer for the most of the world is summer rather than winter due to hemispheres)*.

```
total <- total[month(Date)>0 & month(Date)<6 & year(Date)==2020,
               period:="pre-summer"]

total <- total[month(Date)>8 & month(Date)<13 & year(Date)==2020,
               period:="post-summer"]

total <- total[month(Date)>5 & month(Date)<9 & year(Date)==2020,
               period:="summer"]
```

3.3 Complementary variables

Afterwards that addition, continents will be matched to each country and also some other variables are going to be added in the table 5 and specifically *urban population percentage*, *Density*, *Median age* per country are going to be inserted into that table. Moreover some other features will be excluded such as *world share* and *fertility rate*.This complementary dataset looks like the following table.In the chunk below the matching of the continent also happens with the assistance from the *countrycode* library.The results are on table 6 which has 12 variables*(the table has been vut in hald in order to fit into the page)*

```
population<-fread("pop.csv",header = TRUE)
setnames(population,c("Country (or dependency)"),c("Country_temp"))
population[Country_temp=="United States",Country_temp := "US"]
total <- merge(population,total , by.x="Country_temp", by.y="Country")
colnames(total)[colnames(total) == 'Country_temp'] <- 'Country'
colnames(total)[colnames(total) == 'Population (2020)'] <- 'Population'
colnames(total)[colnames(total) == 'Urban Pop %'] <- 'Urban_pop'
colnames(total)[colnames(total) == 'Med. Age'] <- 'median_age'
colnames(total)[colnames(total) == 'Density (P/KmB²)'] <- 'Density'
total<-total[,-c("Yearly Change","Net Change","Land Area (KmB²)","Migrants (net)",
                 "Fert. Rate","World Share")]
total[,continent:=(countrycode(sourcevar = total[, Country],
                            origin = "country.name",destination = "continent"))]
```

Table 6: Updated table 5 with new variables added

Country	Population	Density	median_age	Urban_pop	Date	confirmed	deaths	deaths.inc	confirmed.ind	quarter	period
Afghanistan	39074280	60	18	25 %	2020-01-22	0	0	0	0	1st quarter	pre-summer
Afghanistan	39074280	60	18	25 %	2020-01-23	0	0	0	0	1st quarter	pre-summer
Afghanistan	39074280	60	18	25 %	2020-01-24	0	0	0	0	1st quarter	pre-summer

3.4 Relative deaths & confirmed variables

Meanwhile 4 new variables with relatives values are going to be added.Relative values are values of variables *confirmed*,*deaths* which are divided by the total population of each country.Comparing countries with their *confirmed*,*deaths* values is not a correct measurement .For instance Belgium cannot be compared with China in absolute numbers but in relative ones that would be more appropriate ,for 1% of China's total population is 60% of Germany's total population.That conversion is crucial in order to observe how different countries treated the situation that has been emerged.One another example is the China's and Greece's deaths.They both have around 4000 deaths but the total population of the former is 3B while the latter's is 13M.

```
total[,relative_deaths := deaths/Population]
total[,relative_confirmed := confirmed/Population]
total[,relative_deaths.inc:= deaths.inc/Population]
total[,relative_confirmed.ind := confirmed.ind/Population]
```

3.5 Categorical variable Urban population

Furthermore one more variable is created and that would be the *Urbanization* variable which divides the countries in 4 categorical values based on the percentage of people who lives in the urban centers per country. For instance 88% of Greece's population lives in Urban centers while 25% of Afghanistan lives in them.In that way we are capable of correlate how the distribution of people across each country's region has affected the Covid-19 cases. This variable has 4 categories

[urban population <25% = "Large number of Citizens live mainly at the countryside"] [24% < urban population < 50% = "Big number of Citizens live mainly at the countryside"]

[49% < urban population < 75% = "Big number of Citizens live mainly at urban centers"]

[urban population > 74% = "Big number of Citizens live mainly at urban centers"]

```
total[Urban_pop<25 ,
      Urbanization:="Large number of Citizens live mainly at the countryside"]
total[Urban_pop>24 & Urban_pop<50 ,
      Urbanization:="Big number of Citizens live mainly at the countryside"]
total[Urban_pop>49 & Urban_pop<75 ,
      Urbanization:="Big number of Citizens live mainly at  urban centers"]
total[Urban_pop>74 ,
      Urbanization:="Large number of Citizens live mainly at  urban centers"]
```

Table 7: Final table with 18 variables as columns

Country	Population	Density	median_age	Urban_pop	Date	confirmed	deaths	deaths.inc	confirmed.ind	quarter	period	continent	relative_deaths	relative_confirmed	relative_deaths.inc	relative_confirmed.ind	Urbanization
Afghanistan	39074280	60	18	25 %	2020-01-22	0	0	0	0	1st quarter	pre-summer	Asia	0	0	0	0	Big number of Citizens live mainly at the countryside
Afghanistan	39074280	60	18	25 %	2020-01-23	0	0	0	0	1st quarter	pre-summer	Asia	0	0	0	0	Big number of Citizens live mainly at the countryside
Afghanistan	39074280	60	18	25 %	2020-01-24	0	0	0	0	1st quarter	pre-summer	Asia	0	0	0	0	Big number of Citizens live mainly at the countryside

The final table that we are going to extract information is shown below and is table 7.Any plot that is going to be implemented will be based on that table.*The table figure below is seperated in order to fit into the page whole the data table*

4. Plots & Results

4.1 Deaths,Confirmed-Relative Deaths,Relative Confirmed for US,China,Pakistan,India,Indonesia(top 5 by population)

In this section of the work ,some aggregations and summaries of the table 7 are going to be made ,for usefull results could be extracted from them.It has to be mentioned that table 7 has 18 columns as described above in the section 3. The two first figures show the cases and the relative confirmed cases for the top 5 countries by population.It seems that China has treated the situation effectively compared to US and India and that is very important because China is the country where Covid-19 has started. Furthermore even if India seems to have reached US in October 2019,when it comes to relative values US has scored very poorly.Of course India has also scored poorly compared to Indonesia, Pakistan and China but all of the aforementioned countries have huge gap from the US for the relative cumulative confirmed cases category. Regarding the relative deaths thigs are far worse for US compared to the other countries.It seems that US's healthcare system have not handled well the situation. Relative numbers reveal partially some insights because number of Covid-19 incidents of a country are divided by the total population of that country and moreover those relativities reveal the impact of the outbreak in a country's society.
disclaimer:i) *We consider that the datasets we have been working on is correct,without missing values from every country.So the results are solely based on those data*

ii) *A sample of the code is shown below in order to evaluate that ggplot was used for all the plots*

```r
top5<-total[Country=="China" | Country=="India" | Country=="US" |
              Country=="Indonesia" | Country=="Pakistan"]
```

```r
top_5_plot_rel_d <- ggplot(top5,aes(Date,relative_deaths))  +
     theme_minimal() +
  geom_line(aes(color=Country),size=1.5)+ labs( x = "Date(Months)",
            y = "Relative deaths due to COVID-19",
            title ="Fig4. Relative deaths for the top 5 Countries by population")+
  theme(axis.text=element_text(size=12),
  axis.title=element_text(size=14,face="bold"))
```

Fig1. Confirmed for the top 5 Countries by population

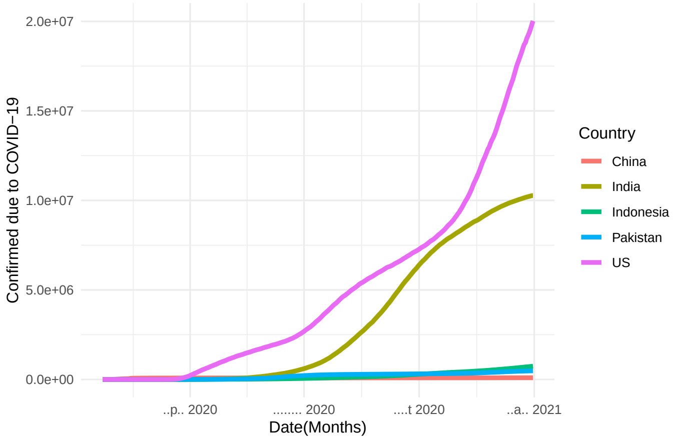


Fig2. Relative Confirmed for the top 5 Countries by population

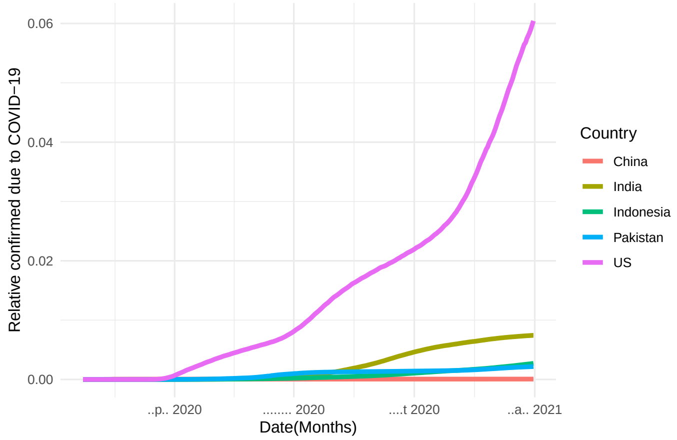


Fig3. deaths for the top 5 Countries by population

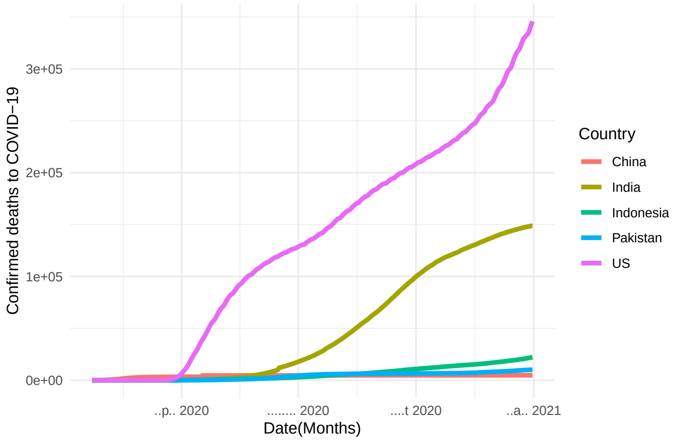
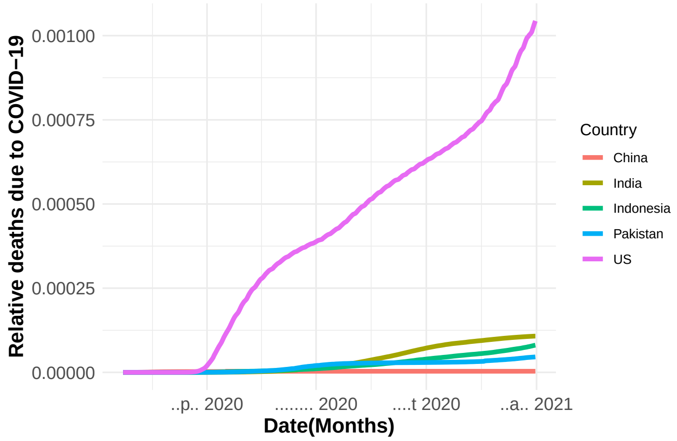


Fig4. Relative deaths for the top 5 Countries by population

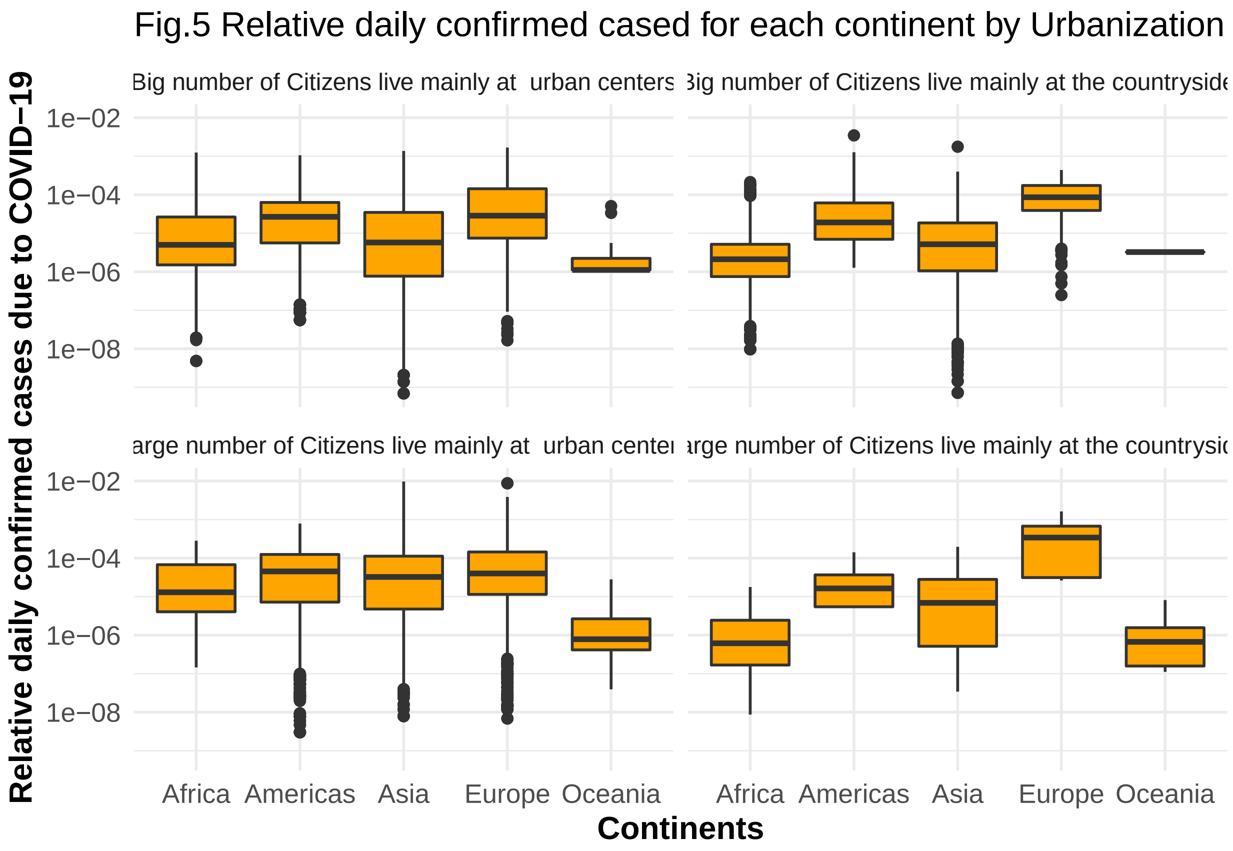


4.2 Relatived Confirmed cases by urbanization measurements

In this subsection some box plots are going to be made in order to depict how urbanization has affected the spread of the disease.The categories of urbanization variable are described in 3.5 section of this work.As someone would expect relative daily confirmed cases are highly correlated with the urbanization.The categories *Large number of citizens live mainly at urban centers*,*Big number of citizens live mainly at urban centers* has more relative confirmed cases than the other two categories.As expected Countries-Continents that have their population gathered around urban centers have higher relative confirmed cases than the other urbanization categories.Although one odd observation is the amount of relative daily confirmed cases in Europe which is high for the category*Big number of citizens live mainly at the countryside.*Therefore for Europe urbanization does not seem to be highly correlated with the relative confirmed cases.For the other continents though the above conclusion is correct.I has to be mentioned that the categories of *Urbanization* variable are 4 groups so an urbanization measurement of 49% may belong to the category of *Big number of citizens live mainly at the countryside* and 51% to the category *Big number of citizens live mainly at urban centers.*More categories should have been added for more precise results.

```
Urban <- ggplot(total,aes(continent,relative_confirmed.ind))  +
  geom_boxplot(fill="orange",size=0.5)+ labs( x = "Continents",
            y = "Relative daily confirmed cases due to COVID-19",
  title ="Fig.5 Relative daily confirmed cased for each continent by Urbanization")+
  theme_minimal()+
  theme(axis.text=element_text(size=10),
  axis.title=element_text(size=12,face="bold"))+
  facet_wrap(~ Urbanization)+
  scale_y_log10()

Urban
```

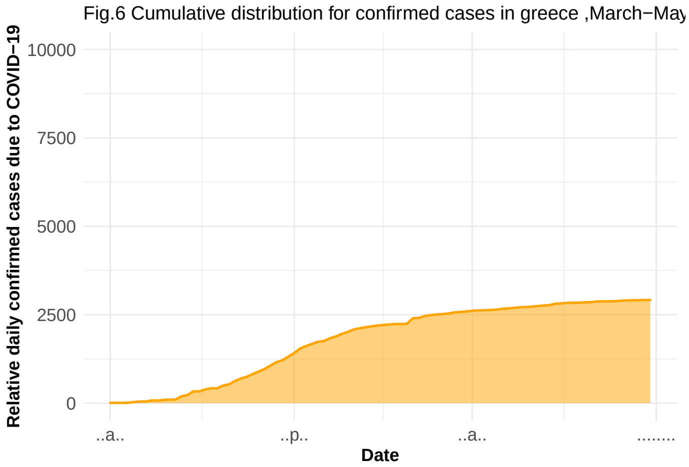


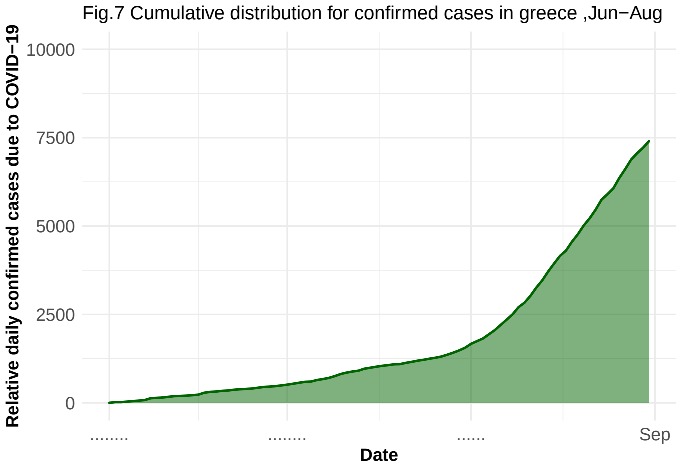
4.3 Impact of the summer in the distribution of Greece

It is time to observe how summer affected the distribution of Confirmed cases and deaths.So we are going to find the cumsum of cases for the summer months. As afore-mentioned, a variable period exists in table.7 where we categorize summer months as *summer*.Here the observation is that Greece had an excellent score from March-May when the last day of may the total confirmed cases was around 2500.We should take into consideration though the number of tests that have been made in that period.The performance of greece in summer months was also fair until mid-August when the distribution started going upwards. The scenery eventually changed dramatically in the autumn (September-November) when the daily confirmed cases were out of control. The y axes in fig.8 has been set to 60000 in order to fit the diagram in contrast with fig.6 and fig.7 where that limit has been set to 10000,therefore someone can understand that Greece's performance was very poor after the summer.The conclusion is that summer has affected Greece in a way that the country's strategy for the confrontation of the disease could not handle tourism. *A sample of the code is shown in order to save space*

In the sample code below in the tot table we create a new variable *cumulative.summer* which calculates the cumulative sum from daily confirmed cases by country and period("pre-summer","summer","post-summer")

```
tot <- total[,cumulative.summer:=cumsum(confirmed.ind),
             by=c("Country","period")]
pre_tot <-total[ ,cumulative.summer:=cumsum(confirmed.ind),
                 by=c("Country","period")]
pre_summer_greece <- ggplot(pre_tot,aes(Date,cumulative.summer))  +
  geom_area(color="orange",fill="orange",size=0.8,alpha=0.5)+ labs( x = "Date",
            y = "Relative daily confirmed cases due to COVID-19",
  title ="Fig.6 Cumulative distribution for confirmed cases in greece ,March-May")+
  theme_minimal()+
  theme(axis.text=element_text(size=12),
  axis.title=element_text(size=12,face="bold"))+
  xlim(as.Date(c('1/3/2020', '31/5/2020'), format="%d/%m/%Y"))+
  ylim(0,10000)
```





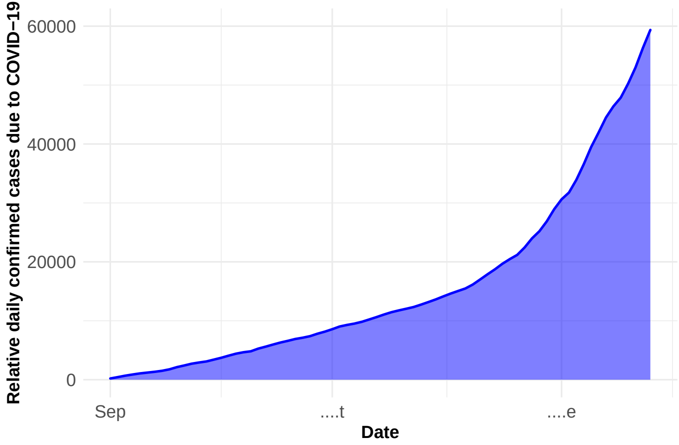
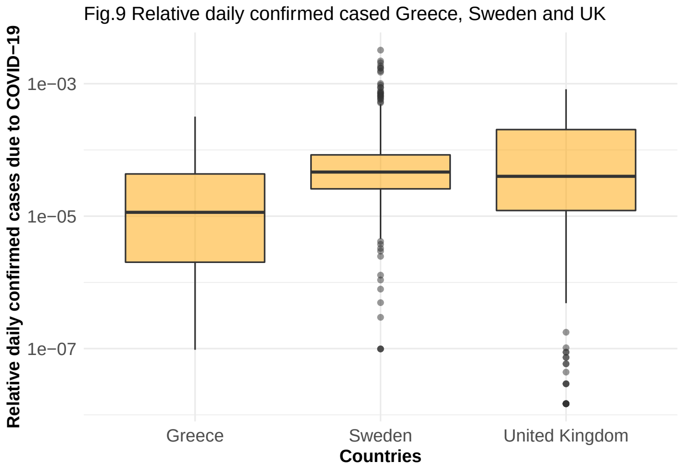


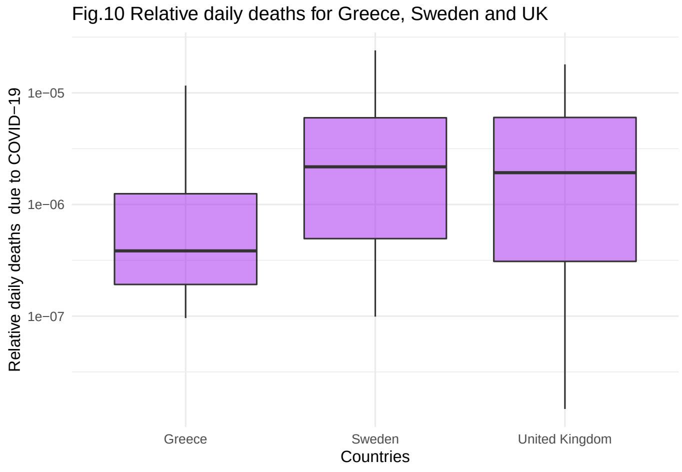
Fig.8 Cumulative distribution for confirmed cases in greece ,Sep−Nov

4.4 Sweden UK and Greece comparison

The comparison of those Countries is necessary in order to evaluate which strategy was the best fitted for the problem.Greece took strict measures such as lockdown,curfew while UK set some partial measures while Sweden was totally free without restrictions for the citizens.Again those measurements are not so accurate due to the fact that climate,citizen's behavior etc play a huge role in the values of those variables. As shown in the boxplots of fig.9 ,indeed the measures greece took were effective compared to the strategies of UK and Sweden because the relative median value for confirmed daily cases was even lower from the 25th Percentile(middle number between smallest number & the median) of the two other countries. Again as expected from the confirm cases Greece's performance was better that that of Uk's and Sweden's in terms of deaths.The logarithmic scale was used for the y axis therefore the differences are far larger than they seem , because human lives are involved so even the difference of 1 human life matters. *Sample code from the plotting below*

```
swukgr <-total[Country=="Sweden" | Country=="Greece" | Country=="United Kingdom"]
sw_uk_gr<- ggplot(swukgr,aes(Country,relative_confirmed.ind))  +
     theme_minimal() +
  geom_boxplot(fill="orange",size=0.5,alpha=0.5)+ labs( x = "Countries",
            y = "Relative daily confirmed cases due to COVID-19",
  title ="Fig.9 Relative daily confirmed cased Greece, Sweden and UK")+
  theme_minimal()+
  theme(axis.text=element_text(size=12),
  axis.title=element_text(size=12,face="bold"))+
  scale_y_log10()
```



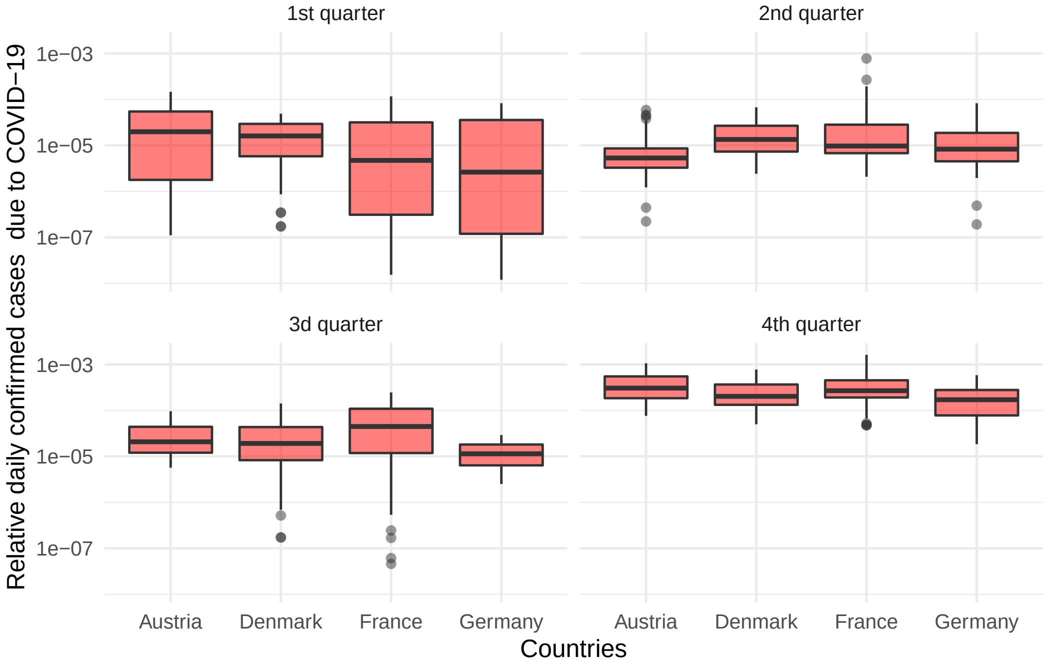


4.4 Quarters: How 4 rich European countries have been affected by the outbreak

The selection has been made from countries belonging to the central europe and those countries are France,Germany,Austria,Denmark.Remember the scale in y axis is logarithmic therefore small changes in y axis means big difference in terms of confirmed cases. Overall Germany had the best performance in every quarter except the 2nd when austria passed germany.The 4th quarter was the worst for every country compared to the other ones.Denmark's strategy was a partial lockdown and that didnt seem to have worked. *Quarters refer to 2020 quarter*

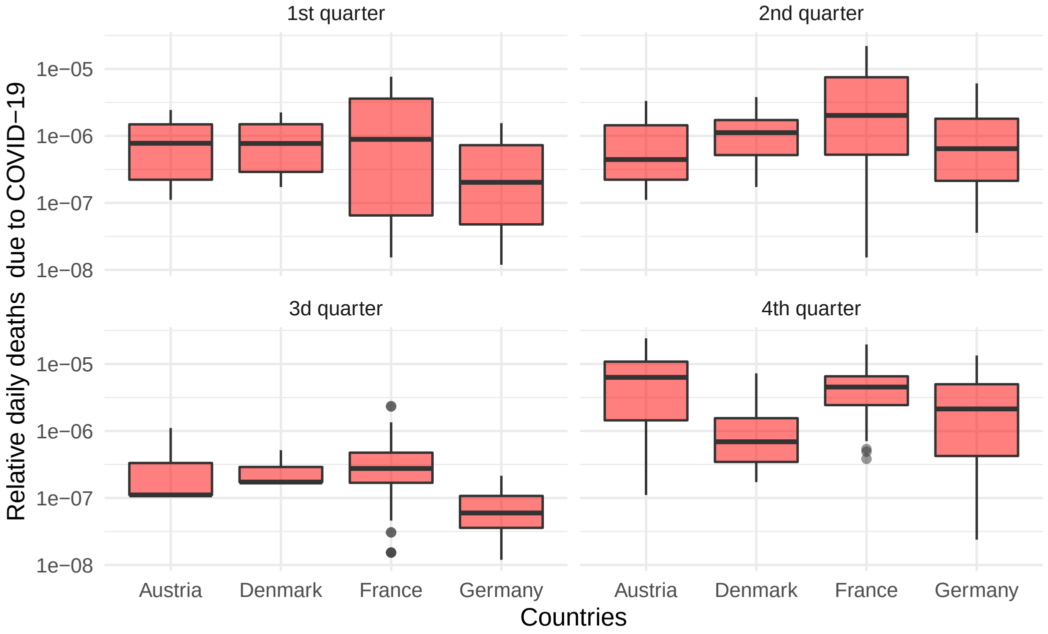
```
fgad <-total[Country=="France" | Country=="Germany" | Country=="Austria"
             | Country=="Denmark"]
fgad_d<- ggplot(fgad,aes(Country,relative_confirmed.ind))  +
  geom_boxplot(fill="red",size=0.5,alpha=0.5)+ labs( x = "Countries",
            y = "Relative daily confirmed cases  due to COVID-19",
  title ="Fig.11 Relative daily confirmed cases for  France,Germany,Austria,Denmark")+
  theme(axis.text=element_text(size=12),
  axis.title=element_text(size=12,face="bold"))+
  theme_minimal()+
  facet_wrap(~ quarter)+
  scale_y_log10()
```

Fig.11 Relative daily confirmed cases for France,Germany,Austria,Denm



In the relative daily deaths category Germany again performed very well compared to the other countries while Denmark's performance in the 4th quarter was very good compared to the others.For all the afore-mentioned countries 3d quarter was the best in terms of confirmed and deaths while the 4th quarter was the worst.

Fig.12 Relative daily deaths for France,Germany,
Austria,Denmark


1st quarter
2nd quarter
3d quarter
4th quarter
1e−05
1e−06
1e−07
1e−08
Relative daily deaths due to COVID−19
Austria
Denmark
France
Germany
Countries