National Technical University of Athens

School of Electrical & Computer Engineering

Programming Tools and Technologies for Data Science 2020-21

Professors:

D. Fouskakis, N. Papaspyrou

Exploratory Data Analysis using R

Pantos Athanasios
Prospective MSc student
A.M. ΕΔΕΜΜ: 03400026
e-mail: pantos.thn@gmail.com

June 1, 2021

Contents

List of Figures

List of Tables

1 Introduction

The coronavirus pandemic is undoubtedly the greatest challenge the world has faced in over a generation.

This is an analysis report of the Novel Coronavirus (COVID-19) around the world, to demonstrate data processing and visualisation with R, dplyr, data.table, tidyverse and ggplot2.

1.1 Data Source

The data source used for this analysis is the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository built by the Center for Systems Science and Engineering, Johns Hopkins University.

1.2 R Packages

Below is a list of R packages used for this analysis. Package magrittr is for pipe and lubridate for date operations. Package tidyverse is a collection of R packages for data science, including dplyr and tidyr for data processing and ggplot2 for graphics. Package gridExtra is for arranging multiple grid-based plots on a page and kableExtra works together with kable() from knitr to build complex HTML or LaTeX tables. Tidyquant is also used for its great aesthetics possibilities.

```
library(magrittr) # pipe operations
library(lubridate) # date operations
library(tidyverse) # ggplot2, tidyr, dplyr...
library(gridExtra) # multiple grid-based plots on a page
library(ggforce) # accelerating ggplot2
library(kableExtra) # complex tables
library(data.table) # data analysis
library(tidyquant)  # for aesthetics
```

2 Loading Data

At first, the data-sets, which are two CSV files, are downloaded and saved as local files and then are loaded into R. Fread function is used which is similar to data.table but faster and more convenient.

```
#load the datasets
cases_df_raw <- fread("time_series_covid19_confirmed_global.csv",
                      sep = ",", header= TRUE)
deaths_df_raw <- fread("time_series_covid19_deaths_global.csv",
                      sep = ",", header= TRUE)

dim(cases_df_raw)
dim(deaths_df_raw)
```

[1] 271 327
[1] 271 327

Each data-set has 271 rows, corresponding to country/region/province/state. It has 150 columns. Starting from column 5, each column corresponds to a single day. Here we have a look at the first 10 rows and the first 10 columns.

Table 1: Raw Data (Confirmed, First 10 Columns only)

Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20
	Afghanistan	33.93911	67.70995	0	0	0	0	0	0
	Albania	41.15330	20.16830	0	0	0	0	0	0
	Algeria	28.03390	1.65960	0	0	0	0	0	0
	Andorra	42.50630	1.52180	0	0	0	0	0	0
	Angola	-11.20270	17.87390	0	0	0	0	0	0
	Antigua and Barbuda	17.06080	-61.79640	0	0	0	0	0	0
	Argentina	-38.41610	-63.61670	0	0	0	0	0	0
	Armenia	40.06910	45.03820	0	0	0	0	0	0
Australian Capital Territory	Australia	-35.47350	149.01240	0	0	0	0	0	0
New South Wales	Australia	-33.86880	151.20930	0	0	0	0	3	4

```
#number of columns
n.col <- ncol(cases_df_raw)
#get dates from column names
dates <- names(cases_df_raw)[5:n.col] %>% substr(1,8) %>% mdy()

range(dates)
```

[1] "2020-01-22" "2020-12-09"

It shows that the data was last downloaded on 09 Dec 2020 and all the stats and charts in this report are based on that data.

3 Data Preparation

3.1 Data Cleaning

The two data sets are converted from wide to long format and then are aggregated by country. After that, they are merged into one single data-set.Each step is well documented with comments in the code snippet below.

```
# data cleaning and transformation
cleanData <- function(data) {
  #remove columns with names Province, State, Lat and Long and
  #rename variable Country.Region to Country
  data <- data[ , -c("Province/State", "Lat", "Long")]
  data <- data %>% setnames(., old = "Country/Region",new = "Country")
  #convert from wide to long format
  data <- tidyr::gather(data, key=date, value=count, -Country)
  #convert the variable date from character to a date object
  data$date <- mdy(substr(data$date, 1, 8))
```

```
  #group by country and date
  data <- data.table(data)[, .(count = sum(count)),
  by = list(date, Country)] %>%
  as.data.frame()
  return(data)
}

#apply the function to the two dataframes
data.confirmed <- cases_df_raw %>% cleanData() %>%
  setnames(., old = "count", new = "confirmed")
data.deaths <- deaths_df_raw %>% cleanData() %>%
  setnames(., old = "count", new = "deaths")

#merge the two datasets into one
data <- merge(data.confirmed, data.deaths, all=TRUE)

#sort by country and date
data <- setorder(data Country, date)

#daily increases of deaths and recovered cases
#set NA to the increases on day1
n <- nrow(data)
day1 <- min(data$date)

#create the daily confirmed cases and daily deaths
data <- data.table(data)[, .(date, Country, confirmed, deaths,
                             confirmed.ind = ifelse(date == day1, 0,
                             confirmed - lag(confirmed, n=1)),
                             deaths.inc = ifelse(date == day1,0,
                                                 deaths -
                                                 lag(deaths, n=1)))]
#change negative number of new cases to zero
data <- data.table(tt)[, .(date, Country, confirmed, deaths,
                   confirmed.ind  = ifelse(confirmed.ind  < 0, 0,
                                           confirmed.ind ),
                   deaths.inc = ifelse(deaths.inc < 0, 0, deaths.inc))
                   ]

#counts for the whole world
data.global <- data.table(data)[, .(
                     confirmed = sum(confirmed, na.rm=T),
                     deaths = sum(deaths, na.rm=T),
                     confirmed.ind = sum(confirmed.ind, na.rm=T),
                     deaths.inc = sum(deaths.inc,  na.rm=T)
                     ), by = list(date)]

#append to main daraframe
data <- rbind(data, data.global)
```

We declare a function called cleanData. The only input argument is a data-set like the ones provided from John Hopkins University. This function is then applied to the two datasets provided by John Hopkins University. The following steps are done during the cleaning:

Below are the last 10 rows for Greece after the cleaning of the data set.

- Columns with names Province, State, Lat and Long are removed.

- Variable Country.Region is renamed to Country.
- Data is converted from wide to long format.
- Variable date is converted from character to a date object via mdy function.
- Data is then grouped by country and date.
- Variable with the cumulative confirmed cases is named confirmed and the variable with the cumulative number of deaths is named deaths.
- The two data-sets are then merged into one.
- Counts (confirmed and deaths) are calculated for the whole world.
- Two extra variables are created: confirmed.ind and deaths.inc with the daily confirmed cases and daily deaths respectively.

The last 10 rows based on Date column for Greece are displayed below after the cleaning has taken place.

Table 2: Data after Cleaning for Greece

	Country	date	confirmed	deaths	confirmed.ind	deaths.inc
314	Greece	2020-11-30	105,271	2,406	1,044	85
315	Greece	2020-12-01	107,470	2,517	2,199	111
316	Greece	2020-12-02	109,655	2,606	2,185	89
317	Greece	2020-12-03	111,537	2,706	1,882	100
318	Greece	2020-12-04	113,185	2,804	1,648	98
319	Greece	2020-12-05	114,568	2,902	1,383	98
320	Greece	2020-12-06	115,471	3,003	903	101
321	Greece	2020-12-07	116,721	3,092	1,250	89
322	Greece	2020-12-08	118,045	3,194	1,324	102
323	Greece	2020-12-09	119,720	3,289	1,675	95

3.2 Feature Engineering

```
#death rate
data %<>% mutate(death.rate = (100 * deaths / confirmed) %>% round(3))
#present confirmed cases
data %<>% mutate(present.confirmed = confirmed - deaths)
```

The idea behind feature engineering is to first discover and then bring to the surface patterns in the data. For this purpose two additional variables are calculated which could prove useful with our goal. Specifically `death.rate` is the ration between the total dead and total confirmed cases while `present.confirmed` cases is calculated by simply subtracting confirmed cases from deaths.

3.3 Variable Interpretation

The data set contains daily information about covid-19 cases for every country in the world. Specifically it contains the cumulative confirmed cases (confirmed) as well as the deaths (deaths) for each country. The absolute number of confirmed (confirmed.ind) as well as deaths (deaths.inc) have been calculated, simply by subtracting today's cumulative confirmed cases and deaths from those of the previous day.

Last but not least, death rate is simply the division of deaths by the number of confirmed cases while present confirmed is simply the subtraction of deaths from confirmed cases.

Table 3: Data Table

Variable	Variable Type	Value
date	double	
Country	character	Greece, US, Germany etc.
confirmed	integer	
deaths	integer	
confirmed.ind	integer	
deaths.inc	integer	
deaths.rate	double	
present.confirmed	integer	

Confirmed cases are the number of people who have tested positive.

One of the stealthiest characteristics of the coronavirus is that it can be spread with little to no symptoms. This means that the number of people worldwide who are actually infected is probably much higher than the official confirmed cases.If you have not been tested, you are not counted.

Every country has its own methods for reporting, which include reviewing a patient's medical records and informing family members before releasing any figures.Leading infectious disease experts have already warned that, despite strict physical distancing measures, the world could still face hundreds of thousands of deaths - if not millions. Just like confirmed cases, the actual number of deaths is almost certainly higher than what is being reported.

4 Worldwide Cases

After tidying up the data, we visualise it with various charts.

4.1 Cumulative Number of Cases

```
#convert from wide to long format, for drawing area plots
data.long <- data %>%
  select(c(Country, date, confirmed, deaths)) %>%
  tidyr::gather(key=type, value=count, -c(Country, date))
global.long <- data.long %>% filter(Country == 'World')

data.global %<>% mutate(present.confirmed = confirmed - deaths,
                        death.rate = (100 * deaths / confirmed)
                        %>% round(3))

#total cases and deaths
n_cases_worldwide <- global.long %>%
  ggplot(aes(x=date, y=count)) +
  geom_area(aes(fill=type), alpha=0.5) +
  labs(title=paste0('Numbers of Cases Worldwide')) +
  scale_fill_manual(values=c('green', 'red')) +
  theme_tq() +
  scale_color_tq()

#log scale
n_cases_worldwide_log <- global.long %>%
  ggplot(aes(x=date, y=log(count))) +
  geom_area(aes(fill=type), alpha=0.5) +
  labs(title=paste0('Numbers of Cases Worldwide (log)')) +
  scale_fill_manual(values=c('green', 'red')) +
  theme_tq() +
  scale_color_tq()
## show two plots side by side
grid.arrange(n_cases_worldwide, n_cases_worldwide_log, ncol=2)
```

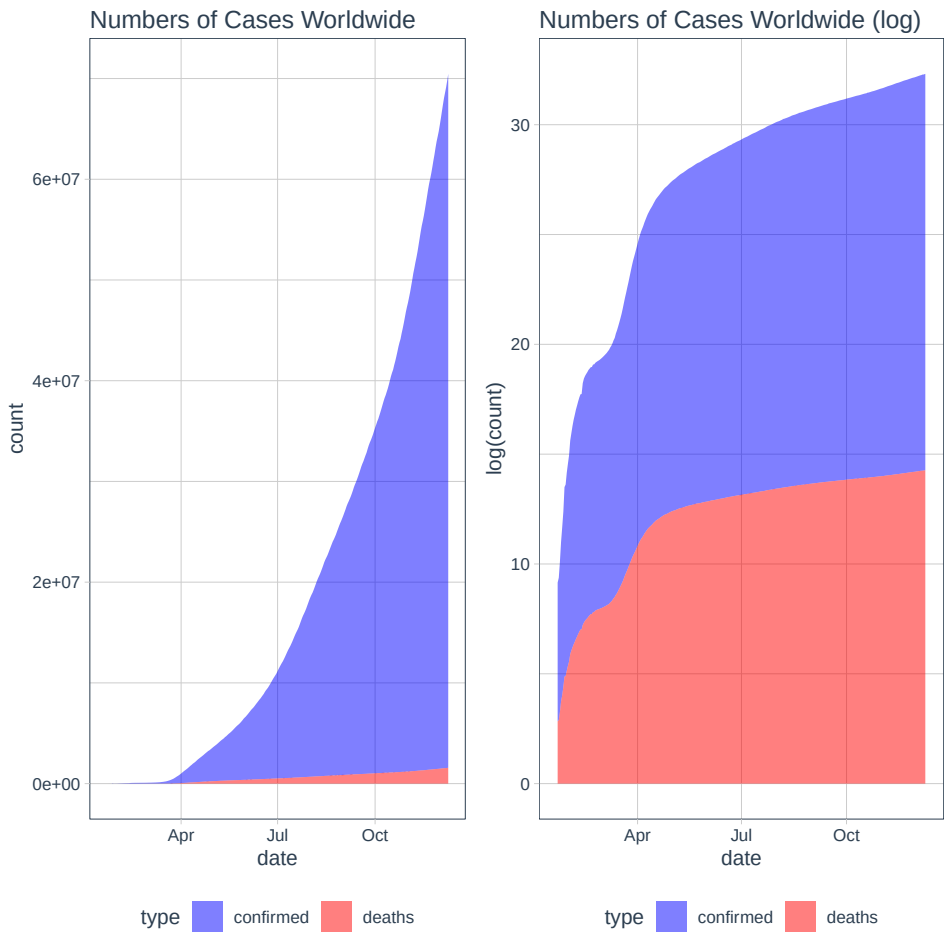

Figure 1: World Cases

As of 16:21 UTC on 22 December 2020, there are 68,894,596 confirmed cases, and 1,569,374 deaths in 191 countries/territories

As you see in Figure 1 the logarithmic scale is ideal for measuring rates of change, particularly rates of growth. On a logarithmic graph of COVID-19 infections, even though the overall numbers are still increasing, you can see the point at which the rate of growth starts to level off when that exponential growth has stopped.

In particular, it looks like that globally the rate of deaths has started to flatten and the number of cases is not growing as fast as it was during the first days of the pandemic. At that point, the logarithmic scale makes it possible to see when public health measures are starting to have the desired effect.

4.2 Daily Confirmed Cases & Deaths

```
#world confirmed cases
current_confirmed_cases <- ggplot(data.global,
                                  aes(x=date, y=present.confirmed)) +
  geom_point(shape=21) + geom_smooth(color = "blue", size =0.8) +
  geom_area(aes(fill=Country), show.legend = FALSE, alpha =0.3) +
  xlab('') + ylab('Count') +
  labs(title='Current Confirmed Cases') +
  scale_fill_manual(values =
                      c(rgb(0, 32, 150, maxColorValue = 255))) +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  theme(legend.position = "none") +
  theme_tq() +
  scale_color_tq()

#daily new cases
daily_new_confirmed_cases <- ggplot(data.global,
                                    aes(x=date, y=confirmed.ind)) +
  geom_point(size = 3, pch = 21, fill = "lightblue",
             color = "black") +
  geom_smooth(size=1, span = 0.3, level = 0.99) +
  xlab('') + ylab('Count') +
  labs(title='Daily New Confirmed Cases') +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  scale_fill_manual(values =
                      c(rgb(0, 32, 150, maxColorValue = 255))) +
  theme_tq() +
  scale_color_tq()

#world total deaths
current_deaths <- ggplot(data.global, aes(x=date, y=deaths)) +
  geom_point(shape=21) + geom_smooth(color="red",
                                     size=0.8) +
  geom_area(aes(fill=Country), show.legend = FALSE, alpha =0.3) +
  xlab('') + ylab('Count') +
  labs(title='Accumulative Deaths') +
  scale_fill_manual(values =
                      c(rgb(255, 0, 0, maxColorValue = 255)),
                    "#FF0000") +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  theme_tq() +
  scale_color_tq()

#world daily deaths
current_daily_deaths <- ggplot(data.global,
                               aes(x=date, y=deaths.inc)) +
  geom_point(size = 3, pch = 21, fill = "#FF9999",
             color = "black") +
  geom_smooth(size=1, colour = 'red', span = 0.3,
              level = 0.99) +
  xlab('') + ylab('Count') + labs(title='Daily New Deaths') +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +
  theme_tq() +
  scale_color_tq()

#display the 4 plots in a 2x2 matrix style
grid.arrange(current_confirmed_cases, daily_new_confirmed_cases,
             current_deaths, current_daily_deaths, ncol=2)
```

Figure 2: Daily New Confirmed COVID-19 Cases / Deaths

Questions Figure 2 answers:

Q. How widespread is the virus?

A. Viruses, including COVID-19, tend to spread at an exponential rate, which means they multiply really quickly. On March 6, the world had 100,000 confirmed cases, one month later there were one million, and if the virus continues to grow at this rate, this number could reach 10 million by May.

Q. Are the number of deaths increasing or decreasing?

A. To answer this, you will need to look at the overall trends. If the slope is moving upwards, then deaths are increasing, if downwards then they are decreasing. Do not draw conclusions based on looking at daily changes on their own; these can differ based on how the data was collected the day before.

5 Top 20 Countries

5.1 Based on Confirmed cases

```
## ranking by confirmed cases world excluded
data.latest.all <- data %>% filter(date == max(date) &
                                     Country != "World") %>%
  mutate(ranking = dense_rank(desc(confirmed)))

k <- 20
## top 20 countries: 21 excl. 'World'
top.countries <- data.latest.all %>%
                 filter(ranking <= k & Country != "World") %>%
                 arrange(ranking) %>% pull(Country) %>%
                 as.character()

data.latest <- data.latest.all %>%
                filter(!is.na(Country)) %>%
                mutate(Country=ifelse(ranking <= k,
                       as.character(Country), 'Others')) %>%
                mutate(Country=Country %>%
                factor(levels=c(top.countries, 'Others')))

data.latest %<>% group_by(Country) %>%
  summarise(confirmed=sum(confirmed), confirmed.ind=sum(confirmed.ind),
            present.confirmed=sum(present.confirmed),
            deaths=sum(deaths), deaths.inc=sum(deaths.inc)) %>%
  mutate(death.rate=(100 * deaths/confirmed) %>% round(1))

data.latest %<>% select(c(Country, confirmed, deaths, death.rate,
                 confirmed.ind, deaths.inc, present.confirmed))

data.latest %>% mutate(death.rate=death.rate %>%
                format(nsmall=1) %>% paste0('%')) %>%
  kable('latex', booktabs=T, row.names=T, align=c('l', rep('r', 6)),
        caption=paste0('Cases in Top 20 Countries - ', max.date,
                       '.'),
        format.args=list(big.mark=',')) %>%
  kable_styling(font_size=7, latex_options=c('striped',
                                             'hold_position',
                                             'repeat_header'))
```

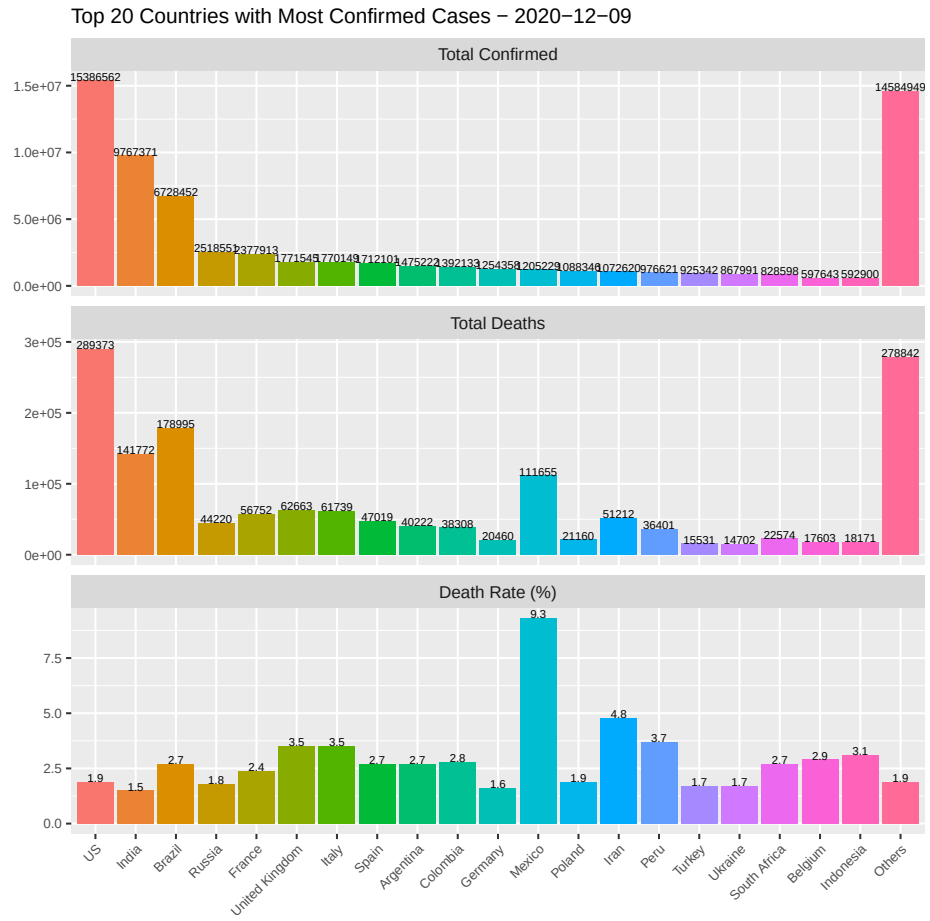

Figure 3: Bar Chart

Figure 3, depicts the countries with the most confirmed cases. In addition to that, the total number of deaths as well as the death rate is also provided. In the bar plot of total confirmed cases the first four countries are US (15,386,562 cases), India (9,767,371 cases), Brazil (6,728,452 cases) and Russia (2,518,551).

However, one would expect that those countries would also have the highest overall number of deaths.By taking a closer look at the total number of deaths one would see that Mexico is the fourth country with the highest number of deaths. It is also worth mentioning that Mexico has the highest death rate in the world.

Table 4: Number of Cases in Top 20 Countries - 2020-12-09.

	Country	confirmed	deaths	death.rate	confirmed.ind	deaths.inc	present.confirmed
1	US	15,386,562	289,373	1.9%	221,267	3,124	15,097,189
2	India	9,767,371	141,772	1.5%	31,521	412	9,625,599
3	Brazil	6,728,452	178,995	2.7%	53,453	836	6,549,457
4	Russia	2,518,551	44,220	1.8%	25,838	546	2,474,331
5	France	2,377,913	56,752	2.4%	14,717	299	2,321,161
6	United Kingdom	1,771,545	62,663	3.5%	16,634	533	1,708,882
7	Italy	1,770,149	61,739	3.5%	12,755	499	1,708,410
8	Spain	1,712,101	47,019	2.7%	9,773	373	1,665,082
9	Argentina	1,475,222	40,222	2.7%	5,303	213	1,435,000
10	Colombia	1,392,133	38,308	2.8%	7,523	150	1,353,825
11	Germany	1,254,358	20,460	1.6%	25,089	458	1,233,898
12	Mexico	1,205,229	111,655	9.3%	11,974	781	1,093,574
13	Poland	1,088,346	21,160	1.9%	12,166	568	1,067,186
14	Iran	1,072,620	51,212	4.8%	10,223	295	1,021,408
15	Peru	976,621	36,401	3.7%	2,709	127	940,220
16	Turkey	925,342	15,531	1.7%	31,712	217	909,811
17	Ukraine	867,991	14,702	1.7%	12,937	289	853,289
18	South Africa	828,598	22,574	2.7%	6,709	142	806,024
19	Belgium	597,643	17,603	2.9%	3,071	96	580,040
20	Indonesia	592,900	18,171	3.1%	6,058	171	574,729

5.2 Confirmed vs Deaths

```
confirmed_vs_deaths <- data %>%
  filter(Country %in% setdiff(top.countries,c('World'))) %>%
  mutate(Country=Country %>% factor(levels=c(top.countries))) %>%
  ggplot(aes(x=confirmed, y=deaths, group=Country)) +
  geom_line(aes(color=Country, linetype=Country)) +
  xlab('Total Confirmed') + ylab('Total Deaths') +
  scale_linetype_manual(values=linetypes) +
  theme(legend.title=element_blank(),
        legend.text=element_text(size=8),
        legend.key.size=unit(0.5, 'cm')) +
  labs(title=paste0('Top 20 Countries')) +
  theme_tq() +
  scale_color_tq()


confirmed_vs_deaths_log <- confirmed_vs_deaths +
                           scale_x_log10() +
                           scale_y_log10() +
                           labs(title=paste0('Top 20 Countries
                           (log scale)')) +
                           theme_tq() +
                           scale_color_tq()
```

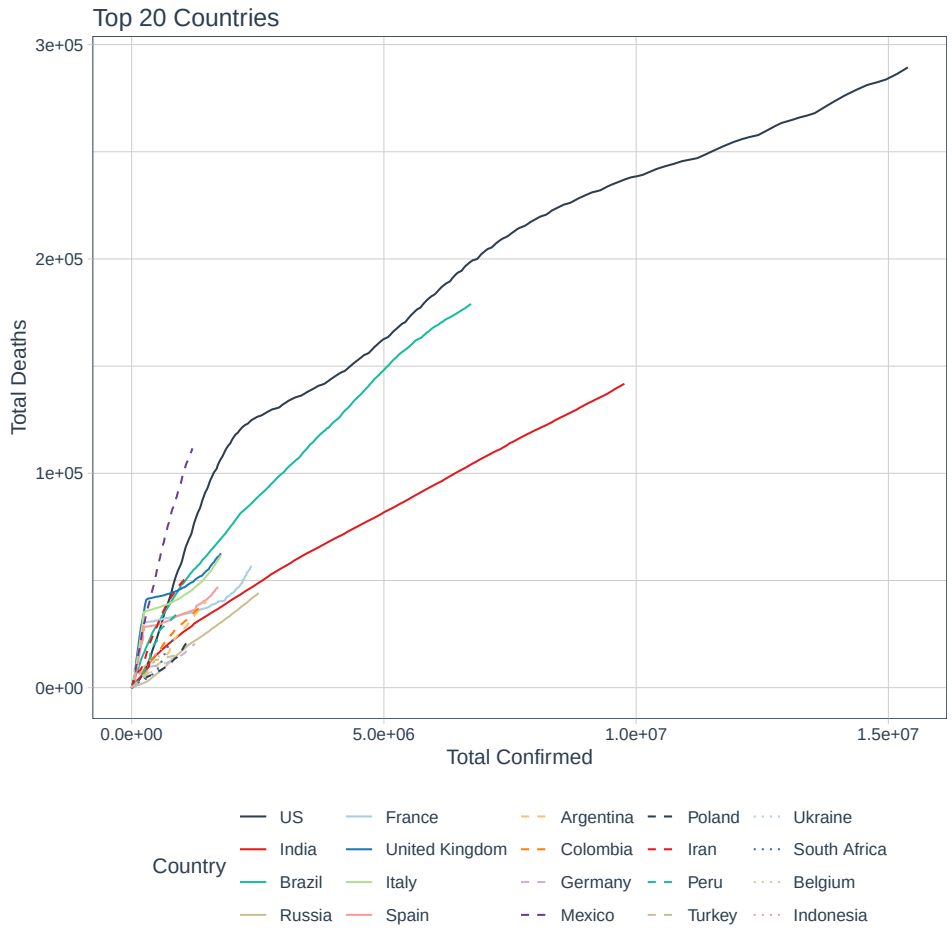

Figure 4: Total Deaths vs Total Confirmed

As we see in Figure 4, only four countries out of twenty appear to have a different trend from the rest. Mexico is the first country that seems to have the most deaths compared to confirmed cases. In other words, Mexico has the steepest line of deaths vs confirmed cases. This could indicate that Mexico is a country with a very high death rate. After Mexico, the second steepest line is that of the US (black line).

The US tends to have a similar trend with Mexico but it is the country with the highest number of confirmed cases thus leading to a big number of deaths. Finally, the green line which depicts Brazil is quite similar to that of US, but it does not have such a big number of confirmed cases and deaths as the US due to a smaller population or stricter measures.

5.3 Based on Death Rates

```
#Sort the latest data by death rate, and if tie, by confirmed
countries_with_highest_death_rate <- data %>% filter(date==max(date)
  & Country != 'World' & confirmed >= 5000) %>%
  select(Country, confirmed, confirmed.ind, present.confirmed,
         deaths, deaths.inc, death.rate) %>%
  arrange(desc(death.rate, confirmed))
  countries_with_highest_rate %>% head(20)

#Infection-Fatality-Rate Evolution
ggplot(filter(data, Country %in% top.countries & confirmed >= 5000),
              aes(x=date, y=death.rate)) +
  stat_summary(geom = "line",fun.y = mean, size= 1.2, colour='red') +
  xlab('Date') + ylab('Infection-Fatality-Rate (%)') +
  labs(title='Infection-Fatality-Rate Evolution') +
  geom_line(aes(color=Country, linetype=Country)) +
  ylim(c(0, 18)) +
  theme_tq() +
  scale_color_tq()
```

Table 5: Top 20 Countries with Highest Infection Fatality rate - 2020-12-09

	Country	confirmed	confirmed.ind	present.confirmed	deaths	deaths.inc	death.rate
1	Mexico	1,205,229	11,974	1,093,574	111,655	781	9.264%
2	Ecuador	199,228	476	185,414	13,814	20	6.934%
3	Sudan	20,084	337	18,777	1,307	6	6.508%
4	Bolivia	146,060	214	137,056	9,004	2	6.165%
5	Egypt	119,702	421	112,870	6,832	19	5.708%
6	Syria	8,675	95	8,210	465	7	5.360%
7	China	93,898	116	89,150	4,748	2	5.057%
8	Iran	1,072,620	10,223	1,021,408	51,212	295	4.774%
9	Afghanistan	47,851	135	45,932	1,919	13	4.010%
10	Peru	976,621	2,709	940,220	36,401	127	3.727%
11	United Kingdom	1,771,545	16,634	1,708,882	62,663	533	3.537%
12	Italy	1,770,149	12,755	1,708,410	61,739	499	3.488%
13	Tunisia	106,856	1,411	103,139	3,717	49	3.479%
14	Guatemala	127,127	654	122,816	4,311	25	3.391%
15	Mali	5,469	27	5,288	181	1	3.310%
16	Australia	28,000	7	27,092	908	0	3.243%
17	Bosnia and Herzegovina	97,317	1,296	94,166	3,151	70	3.238%
18	Bulgaria	171,493	3,328	166,210	5,283	127	3.081%
19	Malawi	6,051	0	5,865	186	0	3.074%
20	Indonesia	592,900	6,058	574,729	18,171	171	3.065%

Table 5 shows the top 20 countries with the highest infection-fatality rate as of 2020-12-09 while Figure 5 shows the infection-fatality rate evolution. The infection-fatality rate is the ratio between the of the number of deaths and the number of confirmed cases. The average death rate for these countries is 4.521% and the red line in Figure 5 depicts the average infection-fatality rate evolution of those 20 countries.

The country with the highest death rate is Mexico (9.264%), whereas the country with the smallest rate is Indonesia (3.065%). The death rate could be an indicator

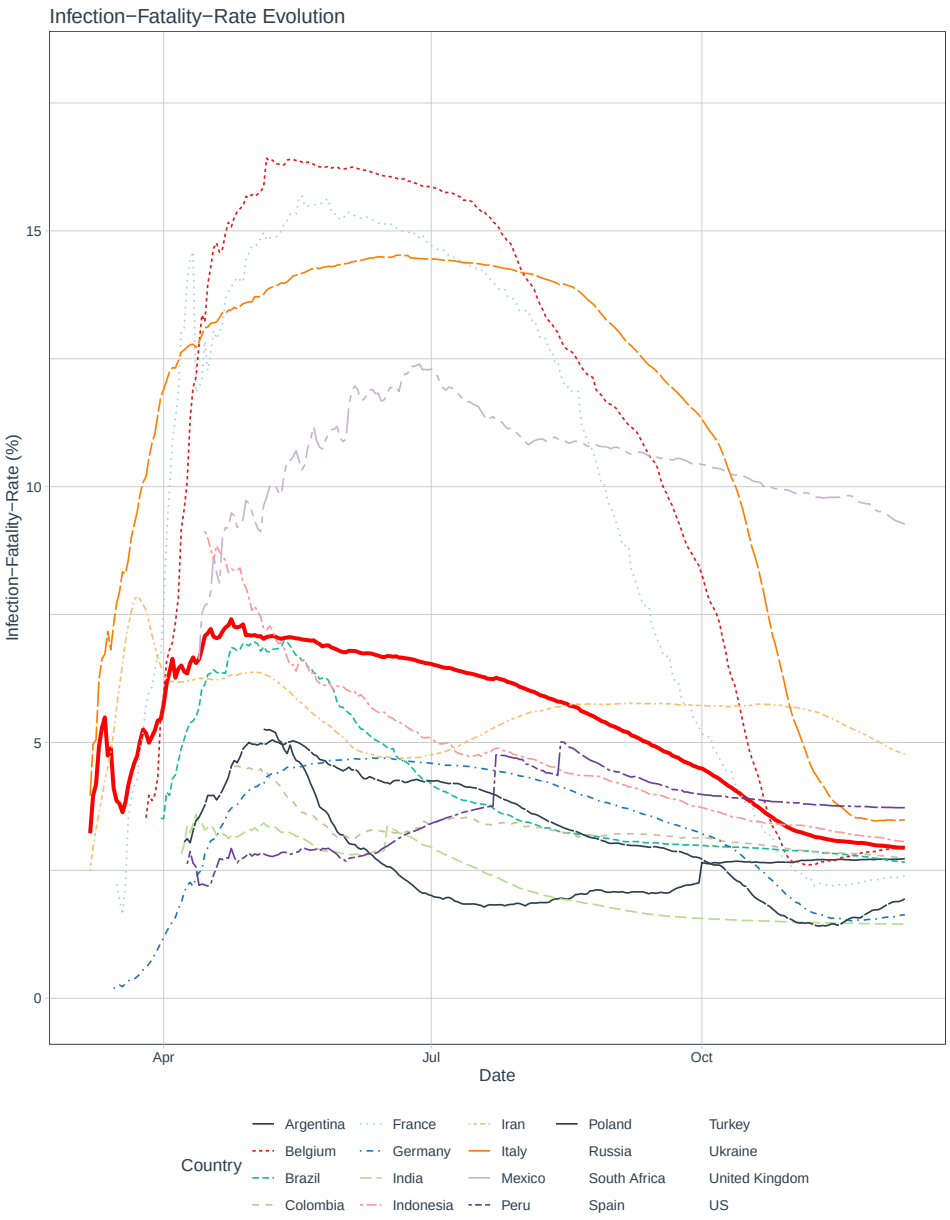

Figure 5: Infection-Fatality-Rate Evolution of top 20 countries

of the healthcare quality of each country. It is observed that most countries, with the highest death rates, are not developed ones perhaps proving that the high death rates could be attributed to lack of healthcare and perhaps poverty.