Exploratory Data Analysis using R

Evaggelos Tsogkas
MSc Student
Email: evaggelostsogkas@mail.ntua.gr
Student number: 03400120

Data Science and Machine Learning
National Technical University of Athens
December 2020

1 Introduction

With the emergence of the novel coronavirus disease (COVID-19) and its rapid spread around the world, it has become an important task to analyze and monitor the characteristics and the growth of the new disease, in order to understand it and take protective measures against it. It is worth noting, however, that the data concerning the spread of COVID-19 are subject to biases, e.g., the number of available tests and who is tested. In this work, I focus on data visualizations of the confirmed cases and deaths caused by COVID-19 worldwide and I explore the situation and mortality rate between different countries. For this purpose two data-sets are used that track the cumulative confirmed cases and deaths respectively, accumulated by John Hopkins University.

2 Data Preprocessing

In order to facilitate the analysis, the data-sets are first restructured and preprocessed. The preprocessing steps include the conversion from wide to long format, the removal of unused variables and renaming of the remaining ones and the conversion of the variable date from character to date object. The data-sets are then sorted and merged and two extra variables are created with the daily confirmed cases and deaths calculated using the cumulative counts. The code for the preprocessing steps is shown in Listing 1. Before creating the extra variables I calculated the total confirmed cases and deaths world wide by using the cumulative counts and selecting the last date. The data are always up-to-date, but I use the version last updated on December 24, 2020. The total confirmed cases on that date are **79,368,142** and the deaths are **1,742,271**. These counts give us a first insight of how much COVID-19 has spread.

```
1  library(data.table)
2  library(ggplot2)
3  library(lubridate)
4  library(dplyr)
5  library(DT)
6
7  # read data
8  confirmed_raw <- fread("time_series_covid19_confirmed_global.
       csv")
9  deaths_raw <- fread("time_series_covid19_deaths_global.csv")
10
11 # remove columns
```

```r
dt_confirmed <- confirmed_raw[, -c("Province/State", "Lat", "Long")]
dt_deaths <- deaths_raw[, -c("Province/State", "Lat", "Long")]

# convert data from wide to long format
dt_confirmed <- melt(dt_confirmed)
dt_deaths <- melt(dt_deaths)

# rename variables
setnames(dt_confirmed, c("country", "date", "confirmed"))
setnames(dt_deaths, c("country", "date", "deaths"))

# convert date from character to a date object
dt_confirmed$date <- mdy(dt_confirmed$date)
dt_deaths$date <- mdy(dt_deaths$date)

# group counts by country and date
dt_confirmed <- dt_confirmed[, .(confirmed=sum(confirmed)), by = .(country, date)]
dt_deaths <- dt_deaths[, .(deaths=sum(deaths)), by = .(country, date)]

# merge the two data-sets
covid19_data <- merge(dt_confirmed, dt_deaths, by=c("country", "date"))

# calculate counts for the whole world
last_date <- tail(covid19_data$date, 1)
total_confirmed <- covid19_data[date==last_date, sum(confirmed)]
total_deaths <- covid19_data[date==last_date, sum(deaths)]

# create variables confirmed.ind and deaths.ind with daily cases
covid19_data <- cbind(covid19_data, covid19_data[, .(confirmed.ind = confirmed - c(0, lag(confirmed)[-1]), deaths.ind = deaths - c(0, lag(deaths)[-1])), by=country][, c("confirmed.ind", "deaths.ind")])
```

Listing 1: Data preprocesing.

3 Worldwide growth of COVID-19

In this section, I visualize the universal growth of COVID-19 over time. Figure 1 shows the cumulative confirmed cases in linear scale. Unfortunately, it seems like the growth follows an exponential curve, which must not be mistaken for merely "fast" as the situation might quickly get out of hand over time.

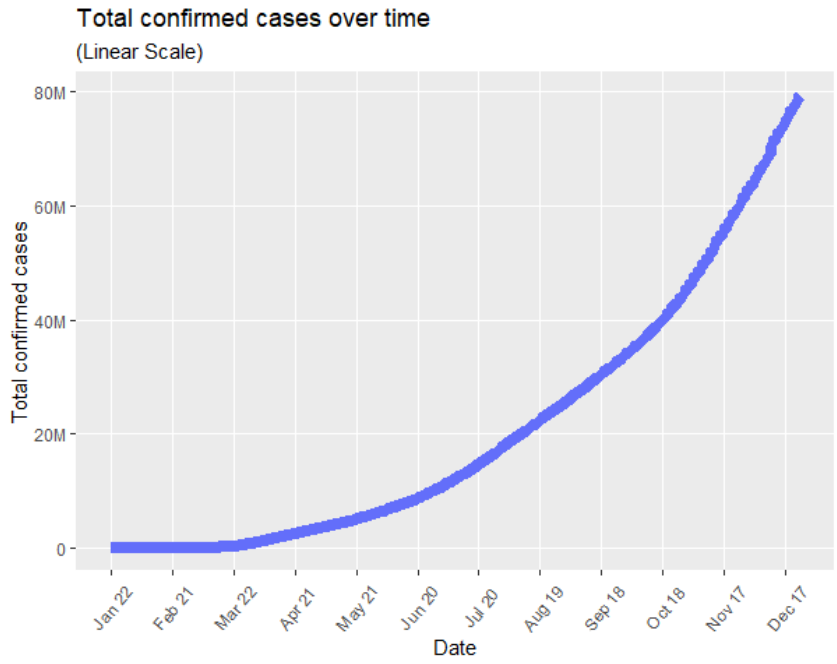

Figure 1: Cumulative confirmed cases of COVID-19 worldwide, plotted in linear scale.

```
# line plot of cumulative confirmed cases in linear scale
world_confirmed <- covid19_data[, .(confirmed=sum(confirmed)),
    by=date]
# axis labels and breaks
x_brks <- world_confirmed$date[seq(1, length(world_confirmed$
    date), 30)]
x_lbls <- paste(month.abb[month(x_brks)], lubridate::day(x_brks
    ))
y_brks <- c(seq(0, 8e+7, 2e+7))
y_lbls <- c(0, paste(y_brks[-1] / 1e+6, 'M', sep=''))
# plot
ggplot(world_confirmed, aes(x = date, y=confirmed)) +
  geom_line(size=3, color="#636efa") +
  labs(title="Total confirmed cases over time", subtitle="(
    Linear Scale)", x="Date", y="Total confirmed cases") +
```

```
  scale_x_date(labels = x_lbls, breaks = x_brks) +
  scale_y_continuous(labels = y_lbls, breaks = y_brks) +
  theme(axis.text.x = element_text(angle = 50, vjust=0.5),
   panel.grid.minor = element_blank())
```

Listing 2: Code to produce Figure 1.

A more accurate way to recognise exponential growth is to plot the cumulative counts in logarithmic scale as shown in Figure 2. Since the logarithm is the inverse of the exponential function, if the growth is exponential then the logarithm of the cumulative counts will follow an approximately straight line. It can be seen that during March the growth of COVID-19 was clearly exponential, although later the slope of the line is smaller as a result of the measures taken by the governments around the world.

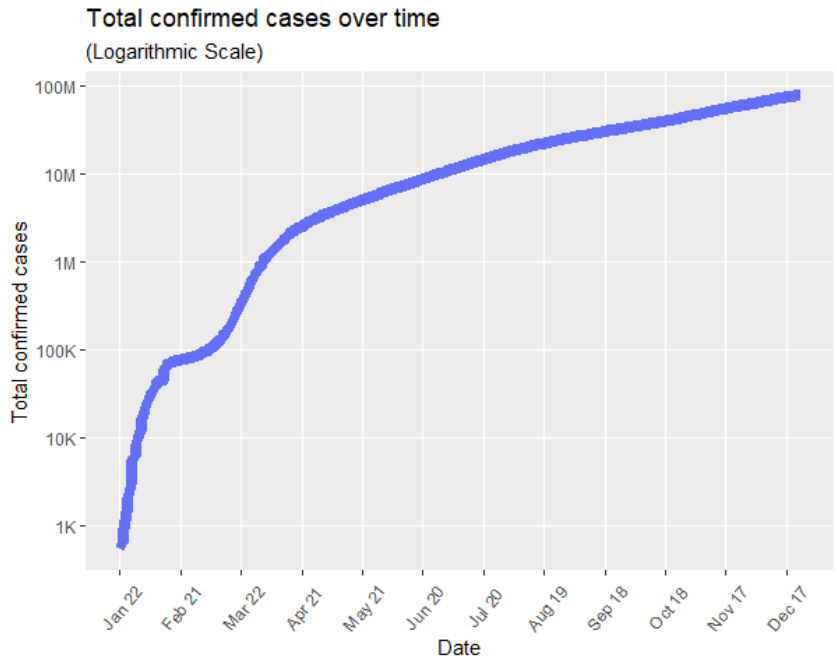

Figure 2: Cumulative confirmed cases of COVID-19 worldwide, plotted in logarithmic scale.

```
# line plot of cumulative confirmed cases in logarithmic scale
y_brks <- c(1e+3, 1e+4, 1e+5, 1e+6,  1e+7, 1e+8)
y_lbls <- c("1K", "10K", "100K", "1M", "10M", "100M")

ggplot(world_confirmed, aes(x = date, y=confirmed)) +
```

```
  geom_line(size=3, color="#636efa") +
  labs(title="Total confirmed cases over time", subtitle="(
   Logarithmic Scale)", x="Date", y="Total confirmed cases") +
  scale_x_date(labels = x_lbls, breaks = x_brks) +
  scale_y_log10(labels = y_lbls, breaks = y_brks) +
  theme(axis.text.x = element_text(angle = 50, vjust=0.5),
   panel.grid.minor = element_blank())
```

Listing 3: Code to produce Figure 2.

In Figure 3 we take a closer look by visualizing the daily confirmed cases. During January, COVID-19 was only present in China and after the lockdown there were not many new cases in February. But when the disease was spread in other countries the cases continued to increase through March. However, many countries enforced lockdown in March and that is why later the daily cases do not increase exponentially. The same pattern seems to repeat over time, as the daily count increases again during July and October due to the relaxation of measures.

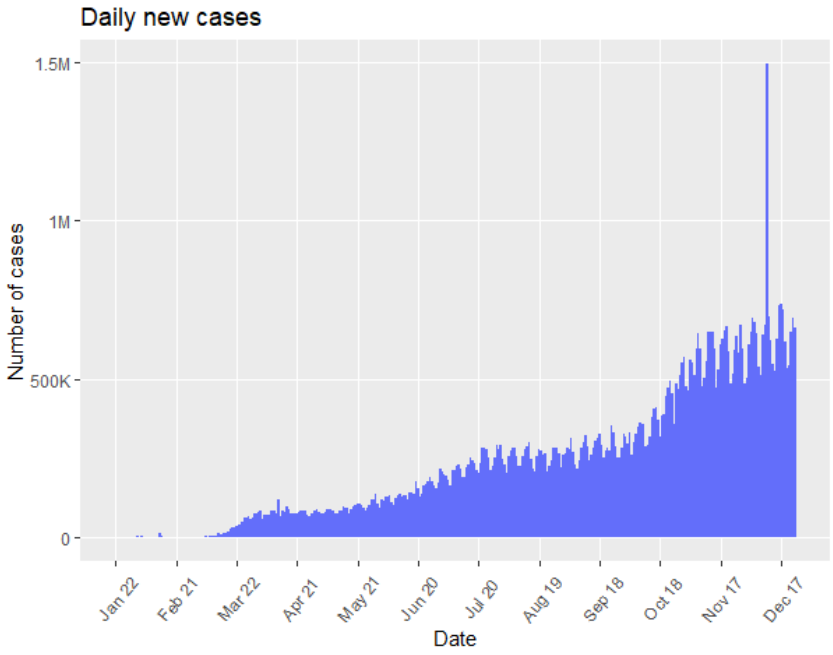

Figure 3: Daily confirmed cases of COVID-19 worldwide.

```r
# bar plot of daily cases
world_daily_confirmed <- covid19_data[, .(confirmed.ind=sum(
    confirmed.ind)), by=date]

y_brks <- c(0, 5e+5, 1e+6, 1.5e+6)
y_lbls <- c('0', '500K', '1M', '1.5M')
ggplot(world_daily_confirmed, aes(x = date, y=confirmed.ind)) +
  geom_bar(stat="identity", width=1, fill="#636efa") +
  labs(title="Daily new cases", x="Date", y="Number of cases")
   +
  scale_x_date(labels = x_lbls, breaks = x_brks) +
  scale_y_continuous(labels = y_lbls, breaks = y_brks) +
  theme(axis.text.x = element_text(angle = 50, vjust=0.5),
   panel.grid.minor = element_blank())
```

Listing 4: Code to produce Figure 3.

Figure 4 shows the cumulative deaths over-time. Thankfully, it seems that the increase of deaths is not as fast as the spread of COVID-19. Additionally, in Figure 5 we can see that the daily counts of deaths follow the same patterns of increasing and decreasing as the confirmed cases, between the same time periods.

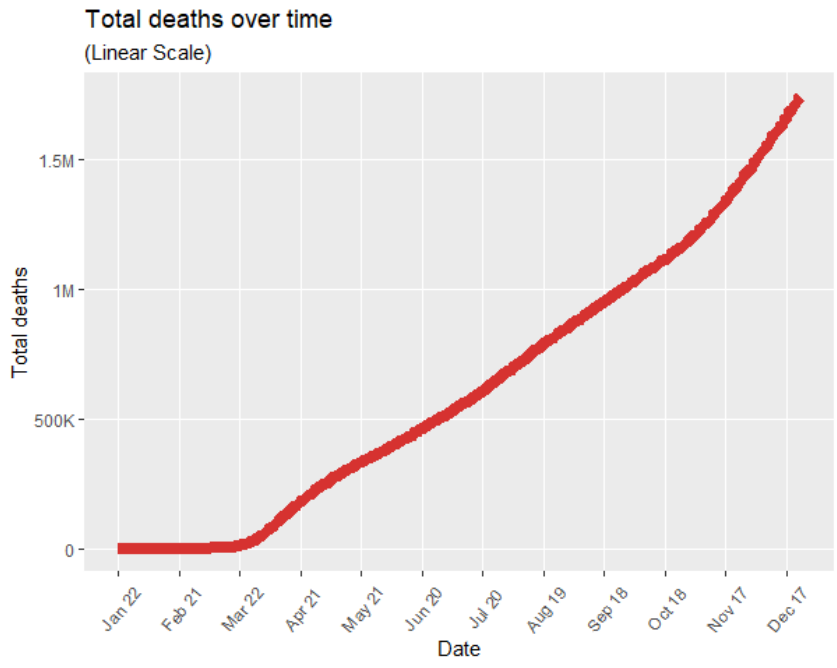

Figure 4: Cumulative deaths due to COVID-19 worldwide.

```
# line plot of cumulative deaths in linear scale
world_deaths <- covid19_data[, .(deaths=sum(deaths)), by=date]
y_brks <- c(0, 5e+5, 1e+6, 1.5e+6)
y_lbls <- c('0', '500K', '1M', '1.5M')

ggplot(world_deaths, aes(x = date, y=deaths)) +
  geom_line(size=3, color="#D63230") +
  labs(title="Total deaths over time", subtitle="(Linear Scale)
   ", x="Date", y="Total deaths") +
  scale_x_date(labels = x_lbls, breaks = x_brks) +
  scale_y_continuous(labels = y_lbls, breaks = y_brks) +
  theme(axis.text.x = element_text(angle = 50, vjust=0.5),
   panel.grid.minor = element_blank())
```

Listing 5: Code to produce Figure 4.

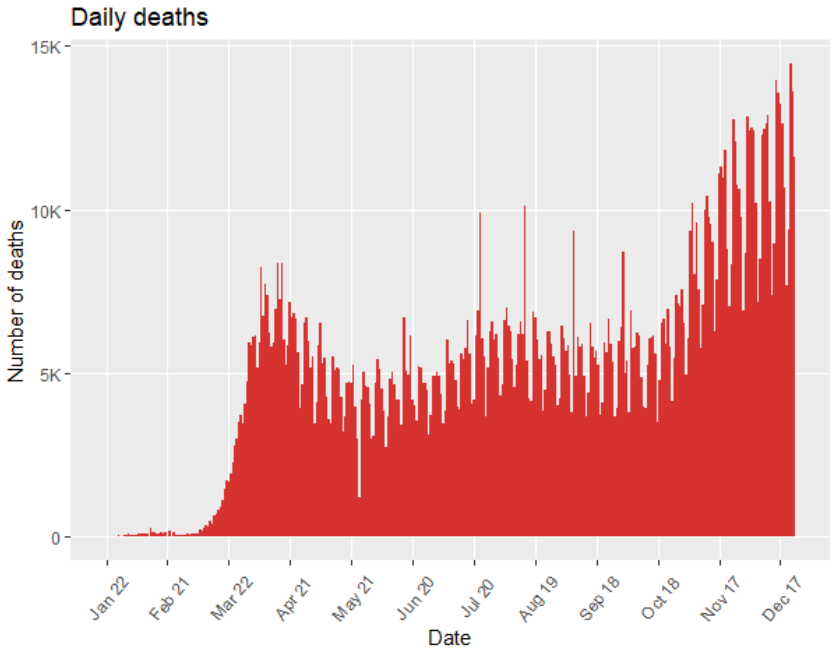

Figure 5: Daily deaths due to COVID-19 worldwide.

```
# bar plot of daily deaths
world_daily_deaths <- covid19_data[, .(deaths.ind=sum(deaths.
   ind)), by=date]

y_brks <- c(0, 5e+3, 10e+3, 15e+3)
y_lbls <- c('0', '5K', '10K', '15K')
```

```
ggplot(world_daily_deaths, aes(x = date, y=deaths.ind)) +
  geom_bar(stat="identity", width=1, fill="#D63230") +
  labs(title="Daily deaths", x="Date", y="Number of deaths") +
  scale_x_date(labels = x_lbls, breaks = x_brks) +
  scale_y_continuous(labels = y_lbls, breaks = y_brks) +
  theme(axis.text.x = element_text(angle = 50, vjust=0.5),
   panel.grid.minor = element_blank())
```

Listing 6: Code to produce Figure 5.

4 COVID-19 situation in different countries

In this section, I explore which countries are most affected by COVID-19 and which handled the situation better. Figure 6 shows the countries with the most confirmed cases. US, India and Brazil have by far the most COVID-19 cases, but it seems that many European countries are affected a lot too.

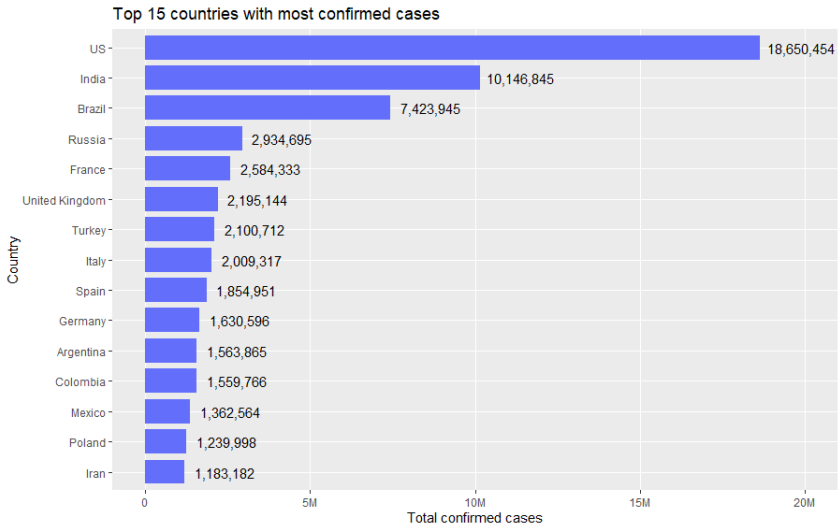

Figure 6: Countries with the most confirmed cases of COVID-19.

```
# bar plot of top 15 countries with most confirmed cases
most_confirmed <- covid19_data[, .(confirmed=sum(confirmed.ind)
   ), by=country][order(-confirmed)][1:15]

ggplot(most_confirmed, aes(x=confirmed, y=country)) +
  geom_bar(stat="identity", width=0.8, fill="#636efa") +
  labs(title="Top 15 countries with most confirmed cases", x="
   Total confirmed cases", y="Country") +
```

```
  scale_x_continuous(labels = c('0', '5M', '10M', '15M', '20M')
   , breaks = c(0, 5e+6, 1e+7, 1.5e+7, 2e+7), limits = c(0, 2e
   +7)) +
  scale_y_discrete(limits = rev(most_confirmed$country)) +
  geom_text(aes(label = format(confirmed, big.mark=","), hjust
   = - 0.1), color = "black") +
  theme(panel.grid.minor = element_blank())
```

Listing 7: Code to produce Figure 6.

Figure 7 shows the countries with the most deaths due to COVID-19. The top three are again the US, India and Brazil, but its interesting to note that Mexico and Italy have a lot of deaths compared to confirmed cases. Italy, in fact, was the first European country to be affected so badly by COVID-19.

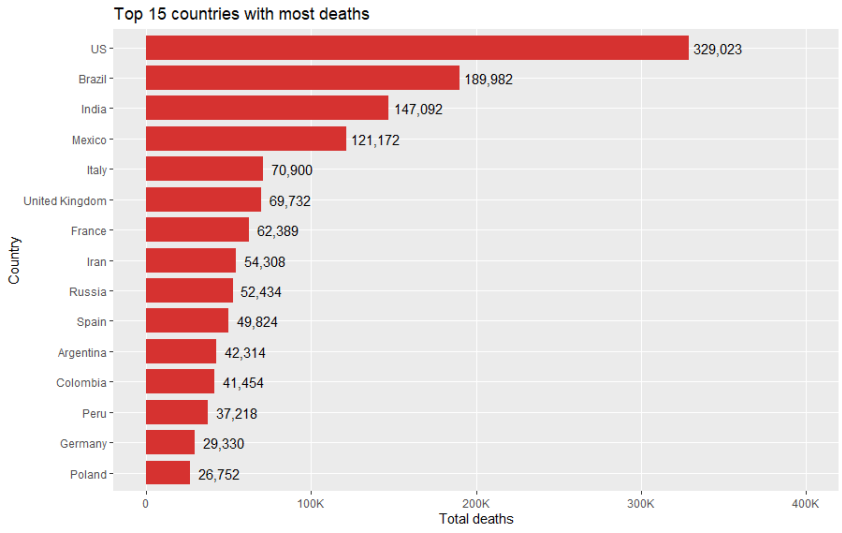

Figure 7: Countries with the most deaths due to COVID-19.

```
# bar plot of top 15 countries with most deaths
most_deaths <- covid19_data[, .(deaths=sum(deaths.ind)), by=
   country][order(-deaths)][1:15]
x_brks <- c(seq(0, 4e+5, 1e+5))
x_lbls <- c(0, paste(x_brks[-1] / 1e+3, 'K', sep=''))

ggplot(most_deaths, aes(x=deaths, y=country)) +
  geom_bar(stat="identity", width=0.8, fill="#D63230") +
  labs(title="Top 15 countries with most deaths", x="Total
   deaths", y="Country") +
  scale_x_continuous(labels = x_lbls, breaks =x_brks, limits =
   c(0, 4e+5)) +
```

```
  scale_y_discrete(limits = rev(most_deaths$country)) +
  geom_text(aes(label = format(deaths, big.mark=","), hjust = -
    0.1), color = "black") +
  theme(panel.grid.minor = element_blank())
```

Listing 8: Code to produce Figure 7.

A more accurate way to compare how serious are the effects of COVID-19 in different countries is by calculating the mortality rate, i.e. the number of deaths per 100 confirmed cases. Figure 8 shows the countries with the highest mortality rate. It is no surprise that Mexico is so high in the list as per our previous observations. The first in the list, however, is Yemen[1], whose actual number of cases in unknown, but as a country still in war with a collapsed health system, it is certainly badly affected.

country	confirmed	deaths	mortality_rate
Yemen	2092	606	28.97
Mexico	1362564	121172	8.89
Ecuador	208010	13977	6.72
Sudan	23316	1468	6.3
Syria	10701	652	6.09
Bolivia	153121	9076	5.93
Egypt	128993	7260	5.63
Chad	1958	102	5.21
China	95383	4769	5
Liberia	1779	83	4.67

Figure 8: Countries with the highest mortality rate due to COVID-19. Countries with less than 1000 confirmed cases are excluded.

```
# mortality rate
mortality_rate <- covid19_data[date==last_date, .(country,
    confirmed, deaths, mortality_rate=round(deaths/confirmed*
    100, 2))]

# top 10 countries with highest mortality rate
# (only countries with more than 1000 confirmed cases
    considered)
```

[1] www.bbc.com/news/world-middle-east-53106164

```
highest_mortality <- head(mortality_rate[confirmed>1000][order
    (-mortality_rate)], 10)

# format style
brks <- quantile(highest_mortality$mortality_rate, probs = seq
    (.05, .95, .05))
clrs <- round(seq(255, 40, length.out = length(brks) + 1), 0)
    %>% {paste0("rgb(255,", ., ",", ., ")")}

datatable(highest_mortality, rownames = FALSE) %>% formatStyle(
    'mortality_rate', backgroundColor = styleInterval(brks,
    clrs))
```

Listing 9: Code to produce Figure 8.

Figure 9 shows the countries with the lowest mortality rate. It is interesting that both Singapore and Qatar, although they have many more confirmed cases than Mongolia they also have a very low mortality rate. The fact that they are two of the wealthiest countries in the world is probably an important factor in how well they can handle the disease, but the decisive factor is that they had a great and early response to the pandemic, swiftly taking measures against it.

country	confirmed	deaths	mortality_rate
Mongolia	1075	0	0
Singapore	58495	29	0.05
Qatar	142605	243	0.17
Botswana	14025	40	0.29
United Arab Emirates	198435	647	0.33
Maldives	13558	48	0.35
Bahrain	91070	350	0.38
Malaysia	100318	446	0.44
Sri Lanka	39231	185	0.47
Iceland	5683	28	0.49

Figure 9: Countries with the lowest mortality rate due to COVID-19. Countries with less than 1000 confirmed cases are excluded.

```r
# top 10 countries with lowest mortality rate
# (only countries with more than 1000 confirmed cases
    considered)
lowest_mortality <- head(mortality_rate[confirmed>1000][order(
    mortality_rate)], 10)

# format style
brks <- quantile(lowest_mortality$mortality_rate, probs = seq
    (.05, .95, .05))
clrs <- round(seq(40, 255, length.out = length(brks) + 1), 0)
    %>% {paste0("rgb(", ., ",255,", ., ")")}

datatable(lowest_mortality, rownames = FALSE) %>% formatStyle('
    mortality_rate', backgroundColor = styleInterval(brks, clrs
    ))
```

Listing 10: Code to produce Figure 9.

5 Conclusions

Although many measures are taken and they have some positive impact in the containment of COVID-19, the disease is still spreading and the cases and deaths are increasing over time. It is important, however, that we continue to enforce protective measures, as when the cases are increasing steadily and not exponentially, the health systems can handle more people in critical condition and the mortality rate is reduced.