2019 - 2020 Covid-19 outbreak

Linlin Sun

3/10/2020

Background

The 2019-2020 Covid-19 outbreak is an ongoing global outbreak of Covid-19 disease 2019 that has been declared a Public Health Emergency of International Concern. It is caused by the SARS-CoV-2 Covid-19, first identified in Wuhan, Hubei, China. Over 100 countries and territories have been affected at the beginning of March 2020 with major outbreaks in central China, South Korea, Italy, Iran, France, and Germany.

Background of the author

As a newbie in the data science world, I would like to use the Covid-19 data to practice my data wrangling skills. As a Chinese-American, I have spent more than half of my life here in US than in China. I have paid attention to the news in China while Wuhan and other places in China were experiencing the covid-19 spreading. Now, I am experiencing covid-19 spread in US. Beside trying my best to do the social distancing, I would like to make my small portion of contribution to the covid-19 analysis.

I am hoping we will get through this soon and wish the best for everyone!

Data files

I have been referring to Johns Hopkins CSSE Covid-19

https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf for affected cases. From their website, they provided a github link containing the data they are using to populate the website.

I got the world population information from https://www.worldometers.info/world-population/population-by-country/

Including the library.

```
library(tidyverse)
library(gridExtra)
library(lubridate)
library(matrixStats)
library(kableExtra)
options(digits=2)
```

Loading the data.

```
path <- getwd()
covid_Confirmed_ts <- "data/time_series_covid_19_confirmed_global.csv"
covid_Deaths_ts <- "data/time_series_covid_19_deaths_global.csv"

c_ts <- read.csv(paste(path, covid_Confirmed_ts, sep = "/"), header = TRUE)
c_ts_col_count <- length(colnames(c_ts))
d_ts <- read.csv(paste(path, covid_Deaths_ts, sep = "/"), header = TRUE)
d_ts_col_count <- length(colnames(d_ts))
```

Set some constant variables.

```
select_top <- 15

day0_count <- 100

colors <- c("aquamarine2", "brown", "blue", "orange", "chartreuse",
            "darkgoldenrod1", "cyan", "darkgreen", "yellow", "darkseagreen",
            "grey39", "deepskyblue4", "darkorchid", "pink2", "red1")
```

First, I get the confirmed and deaths data ready and combine them.

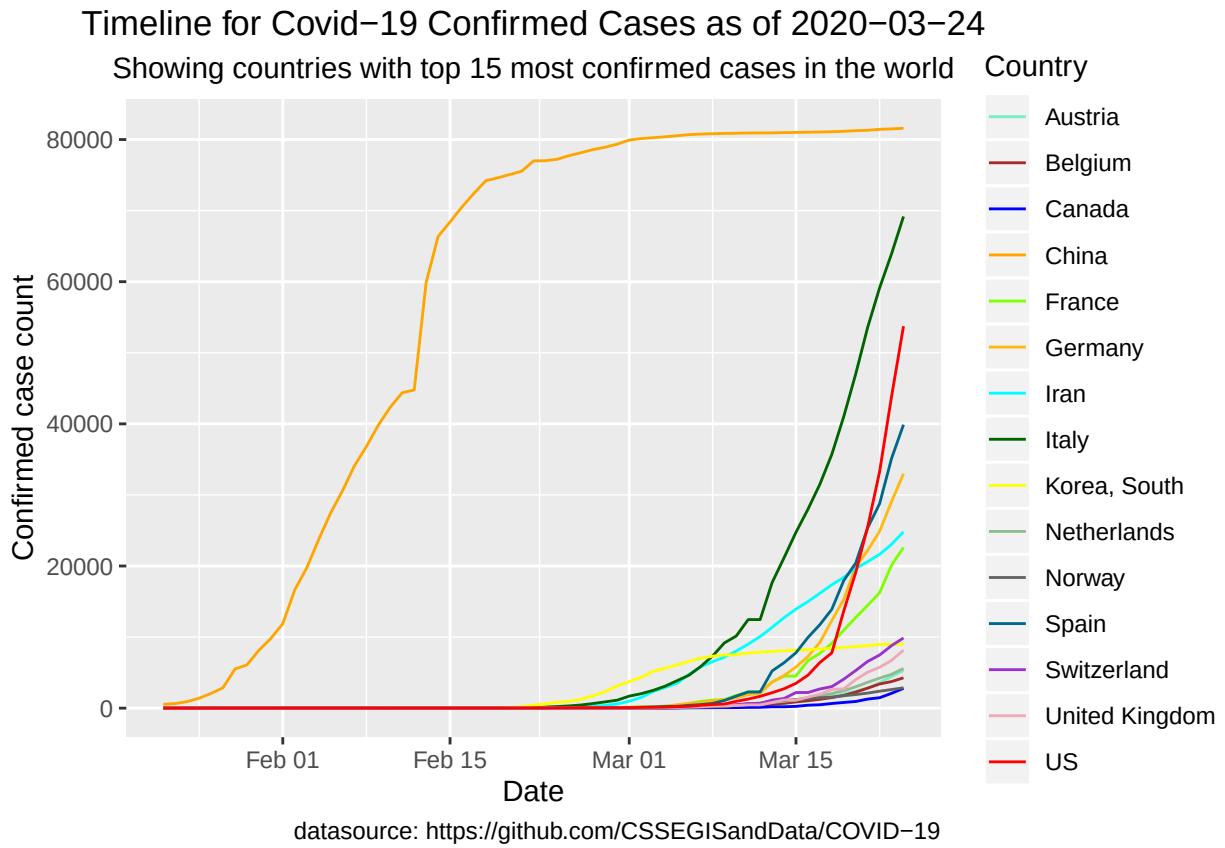
```
confirmed <- c_ts %>% select(-Province.State, -Lat, -Long) %>%
    gather(Date_temp, value, -Country.Region) %>%
    group_by(Country.Region, Date_temp) %>%
    summarize(Confirmed = sum(value)) %>%
    ungroup() %>%
    mutate(Date_temp = str_replace(Date_temp, "X", "")) %>%
    mutate(Date = as.POSIXct(strptime(Date_temp, "%m.%d.%y")))

deaths <- d_ts %>% select(-Province.State, -Lat, -Long) %>%
    gather(Date_temp, value, -Country.Region) %>%
    group_by(Country.Region, Date_temp) %>%
    summarize(Deaths = sum(value)) %>%
    ungroup() %>%
    mutate(Date_temp = str_replace(Date_temp, "X", "")) %>%
    mutate(Date = as.POSIXct(strptime(Date_temp, "%m.%d.%y")))

confirmed_deaths <- cbind(confirmed[, c(1, 4)], confirmed[, 3], deaths[3])

str(confirmed_deaths)
```

```
## 'data.frame':    10710 obs. of  4 variables:
##  $ Country.Region: Factor w/ 170 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date          : POSIXct, format: "2020-01-22" "2020-01-23" ...
##  $ Confirmed     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths        : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
today <- max(confirmed$Date)
```

I would like to plot the time series data for a few countries with the most confirmed cases.

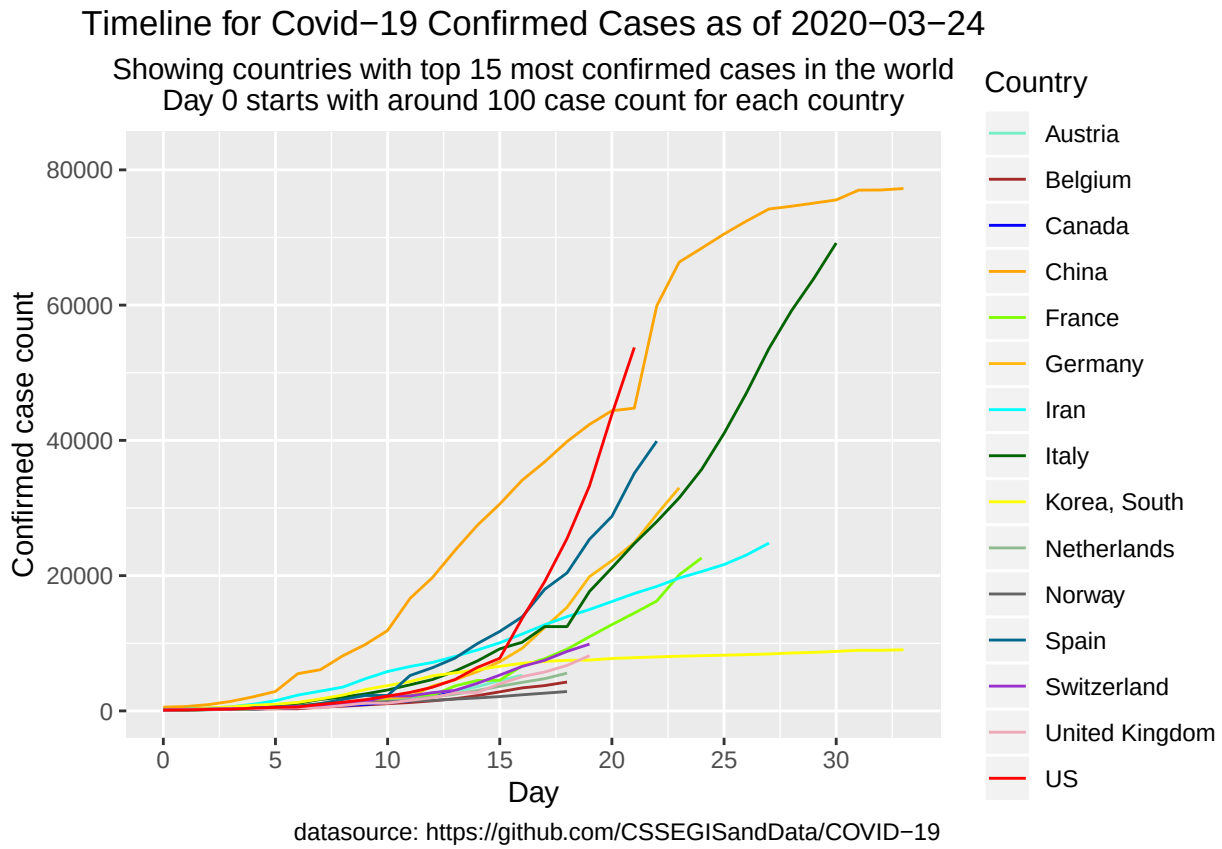
```r
top_confirmed_countries <- confirmed %>% filter(Date == today) %>%
    arrange(desc(Confirmed)) %>%
    top_n(select_top, Confirmed) %>%
    mutate(Country.Region = as.character(Country.Region)) %>%
    .$Country.Region

# confirmed %>% filter(Country.Region == "United Kingdom")

confirmed %>% filter(Country.Region %in% top_confirmed_countries) %>%
    ggplot(aes(Date, Confirmed, color = Country.Region)) +
    geom_line() +
    theme(plot.title = element_text(hjust = 0.5),
          plot.subtitle = element_text(hjust = 0.5)) +
    scale_color_manual(values = colors) +
    labs(title = paste("Timeline for Covid-19 Confirmed Cases as of", today, sep = " "),
         subtitle = paste("Showing countries with top", select_top,
                          "most confirmed cases in the world", sep = " "),
         x = "Date",
         y = "Confirmed case count",
         caption = "datasource: https://github.com/CSSEGISandData/COVID-19",
         color = "Country")
```



The above plot show show the covid-19 progressed in the past few months for each country displayed. I am going to find out the day that the country has around 100 confirmed cases, then plot each country at the same starting point. This way, it is easier to do visual comparison among the countries.

```r
# day0 plot


confirmed_day0 <- confirmed %>% filter(Confirmed >= day0_count) %>% group_by(Country.Region) %>%
    arrange(Confirmed) %>% mutate(Day = row_number() - 1) %>% ungroup()

x_lim_max <- confirmed_day0 %>% filter(Country.Region %in% top_confirmed_countries) %>%
  filter(Country.Region != top_confirmed_countries[1]) %>%
  arrange(desc(Day)) %>%
  select(Day) %>%
  summarize(max_day = max(Day)) %>%
  pull(max_day)


p_day0_base <- confirmed_day0 %>% filter(Country.Region %in% top_confirmed_countries) %>%
    ggplot(aes(Day, Confirmed, color = Country.Region)) +
    geom_line() +
    theme(plot.title = element_text(hjust = 0.5),
          plot.subtitle = element_text(hjust = 0.5)) +
    scale_x_continuous(breaks = seq(0, x_lim_max, 5), lim = c(0, x_lim_max)) +
    scale_color_manual(values = colors)
p_day0 <- p_day0_base +
      labs(title = paste("Timeline for Covid-19 Confirmed Cases as of", today, sep = " "),
           subtitle = paste("Showing countries with top", select_top,
    "most confirmed cases in the world\nDay 0 starts with around 100 case count for each country",
                     sep = " "),
           x = "Day",
           y = "Confirmed case count",
           caption = "datasource: https://github.com/CSSEGISandData/COVID-19",
           color = "Country")
p_day0
```

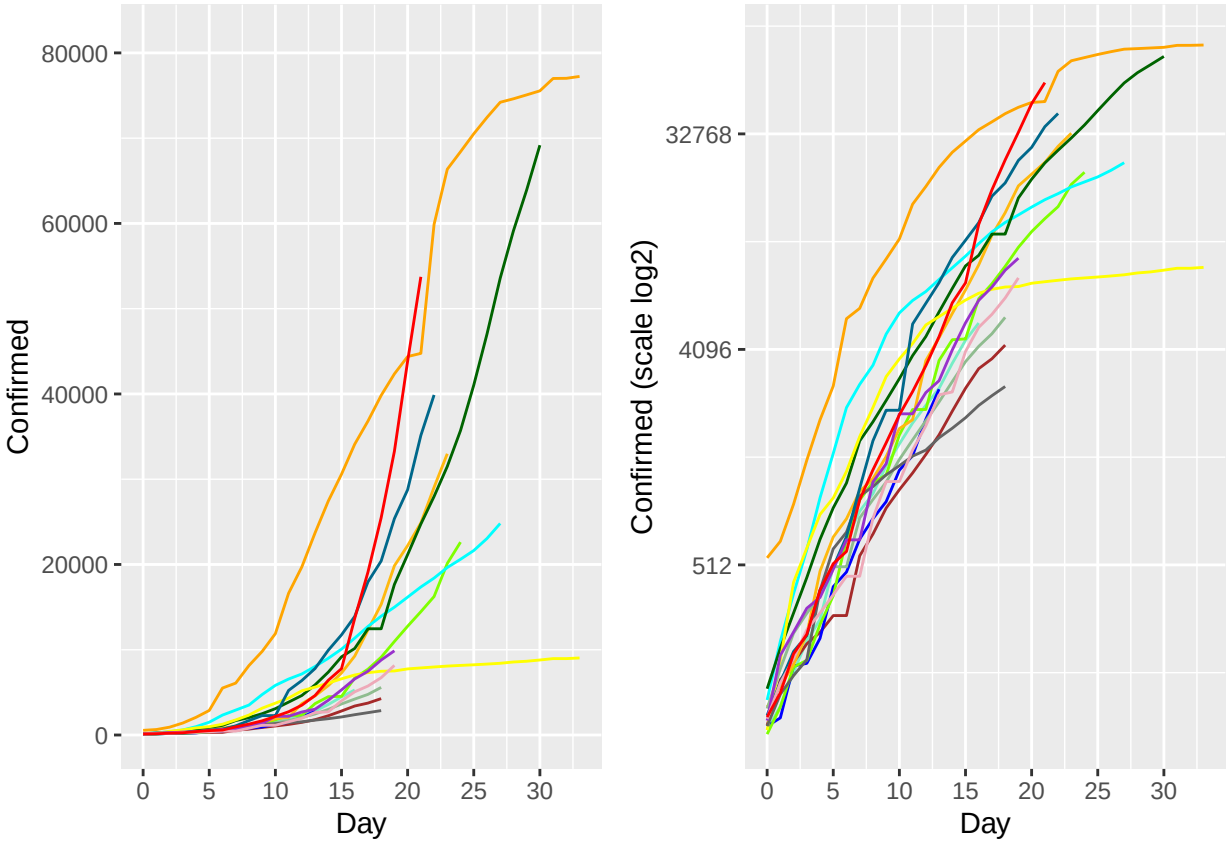
```
## Warning: Removed 29 rows containing missing values (geom_path).
```



Show the same plot side by side with the confirmed case being transformed on log2 scale

```
p_day0_log2 <- p_day0_base +
  scale_y_continuous(trans = "log2") +
  theme(legend.position = "none") +
  labs(y = "Confirmed (scale log2)")
grid.arrange(p_day0_base + theme(legend.position = "none"), p_day0_log2, ncol = 2)
```

```
## Warning: Removed 29 rows containing missing values (geom_path).

## Warning: Removed 29 rows containing missing values (geom_path).
```



Below is what I see from the above plots. 1. China had the fastest growth at the beginning, most likely due to the dense population and no prepareness being the first being hit.

2. European countries were the next that got most impacted. The countries are Spain, Germany and Italy.
3. US started after European countries got impacted. The confirmed case growth rate of US exceed the other three European countries.
4. From the log2 scaled plot, all countries confirmed growth rate seems to be similar until the situation got controlled.

I would like to combine country information such as population, density and median_age into my analysis.

I am getting the world population data.

```
source("WorldPopulation.R")
wp <- getWorldPopulation()
```

I am joining the world population data with the covid-19 data by country name.

```
wppd <- wp %>%
    mutate(Density = str_replace_all(Density, ",", "")) %>%
    mutate(Density = as.numeric(Density)) %>%
    mutate(Population = str_replace_all(Population, ",", "")) %>%
    mutate(Population = as.numeric(Population)) %>%
    select(Country.Region, Density, Population, Median_Age)
```

```
confirmed_deaths <- confirmed_deaths %>% mutate(Country.Region = as.character(Country.Region))

all <- left_join(confirmed_deaths %>% filter(Date == today), wppd, by = "Country.Region")

# sapply(all, function(col) {sum(is.na(col))})
unique(all[is.na(all$Density),]$Country.Region)
```

```
## [1] "Congo (Brazzaville)" "Congo (Kinshasa)"    "Diamond Princess"
```

```
# Three countries do not get a matching density

# There are two Congo entries in covid-19 data, one congo entry from the world population data
all$Density[str_detect(all$Country.Region, "Congo")] <-
    wppd$Density[wppd$Country.Region == "Congo"]

all$Population[str_detect(all$Country.Region, "Congo")] <-
    wppd$Population[wppd$Country.Region == "Congo"]/2

# Cruise ship is not a real country.
all$Density[str_detect(all$Country.Region, "Cruise Ship")] <- 10
all$Population[str_detect(all$Country.Region, "Cruise Ship")] <-2000
```

I am calculating below columns.

C/Population and D/Population shows number of cases per 10,000 people.

"D/C by Density" is calculated by Deaths/Confirmed divided by country density. (Density here is number of people per square km). So this rate removes the density factor. If a country has higher density, that makes the virus to be transmitted more easily.

```
all <- all %>%
    mutate("D/C %" = (Deaths/Confirmed)*100) %>%
    mutate("C/Population" = (Confirmed/Population)*10000,
           "D/Population" = (Deaths/Population)*10000) %>%
    mutate("D/C % by Density" = `D/C %`/Density) %>%
    arrange(desc(Confirmed))

knitr::kable(slice(all, 1:select_top),
  caption = paste("World Covid-19 Summary", today, "C/Population, D/Population, R/Population are per 10
  format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

Below are my personal opinion from this table. I could be wrong. Please let me know what you think. I will really appreciate it.

1. If not considering the population and density factors, China has the highest confirmed case. Italy has the highest Deaths/Confirmed rate (D/C %).
2. Adding consideration of the countries' population, Italy has the highest Confirmed and Death cases per 10,000 people.
3. Adding the consideration of population density, Iran has relative high Deaths/Confirmed by density. Italy is not on the top of this list.

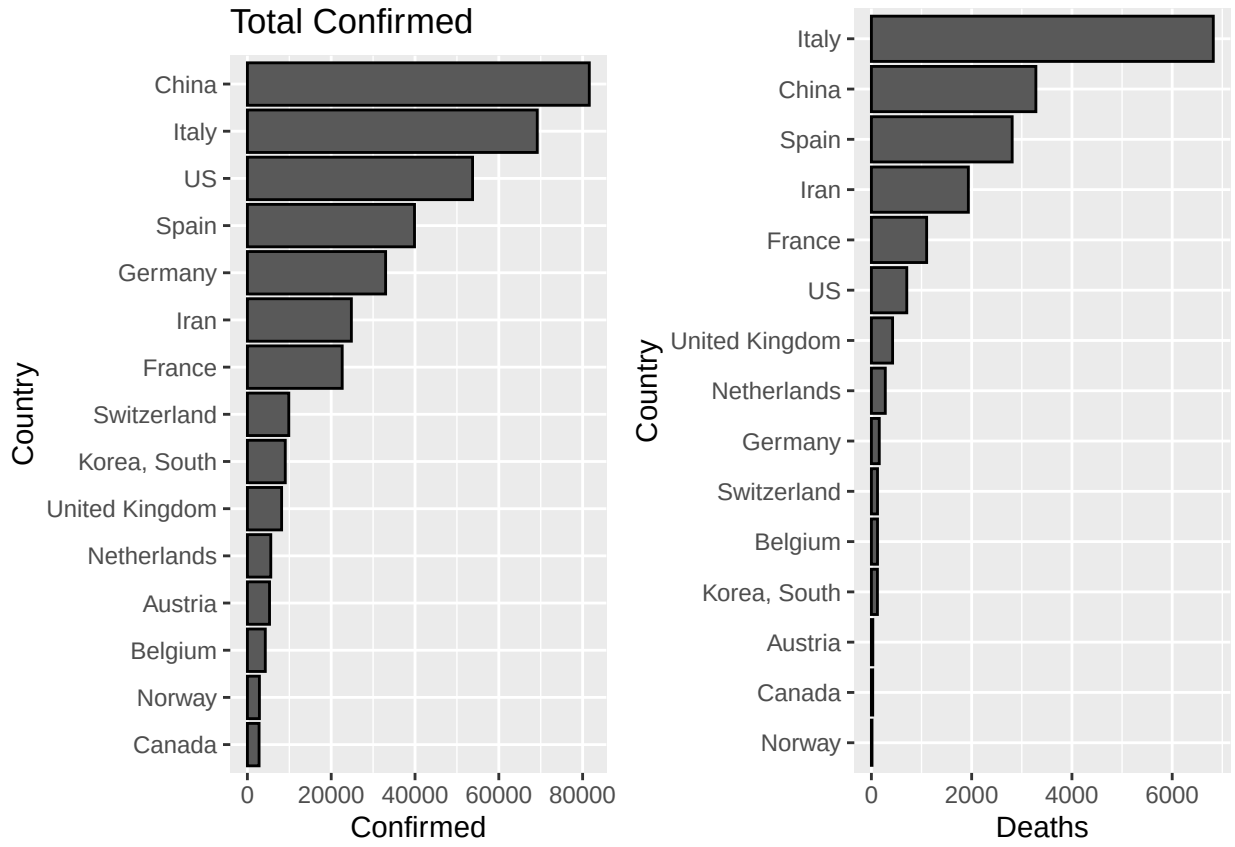
Table 1: World Covid-19 Summary 2020-03-24 C/Population, D/Population, R/Population are per 10,000 people

Country.Region	Date	Confirmed	Deaths	Density	Population	Median_Age	D/C %	C/Population	D/Population	D/C % by Density
China	2020-03-24	81591	3281	153	1.4e+09	38	4.02	0.57	0.02	0.03
Italy	2020-03-24	69176	6820	206	6.0e+07	47	9.86	11.44	1.13	0.05
US	2020-03-24	53740	706	36	3.3e+08	38	1.31	1.62	0.02	0.04
Spain	2020-03-24	39885	2808	94	4.7e+07	45	7.04	8.53	0.60	0.07
Germany	2020-03-24	32986	157	240	8.4e+07	46	0.48	3.94	0.02	0.00
Iran	2020-03-24	24811	1934	52	8.4e+07	32	7.79	2.95	0.23	0.15
France	2020-03-24	22622	1102	119	6.5e+07	42	4.87	3.47	0.17	0.04
Switzerland	2020-03-24	9877	122	219	8.7e+06	43	1.24	11.41	0.14	0.01
Korea, South	2020-03-24	9037	120	527	5.1e+07	44	1.33	1.76	0.02	0.00
United Kingdom	2020-03-24	8164	423	281	6.8e+07	40	5.18	1.20	0.06	0.02
Netherlands	2020-03-24	5580	277	508	1.7e+07	43	4.96	3.26	0.16	0.01
Austria	2020-03-24	5283	28	109	9.0e+06	43	0.53	5.87	0.03	0.00
Belgium	2020-03-24	4269	122	383	1.2e+07	42	2.86	3.68	0.11	0.01
Norway	2020-03-24	2863	12	15	5.4e+06	40	0.42	5.28	0.02	0.03
Canada	2020-03-24	2790	26	4	3.8e+07	41	0.93	0.74	0.01	0.23

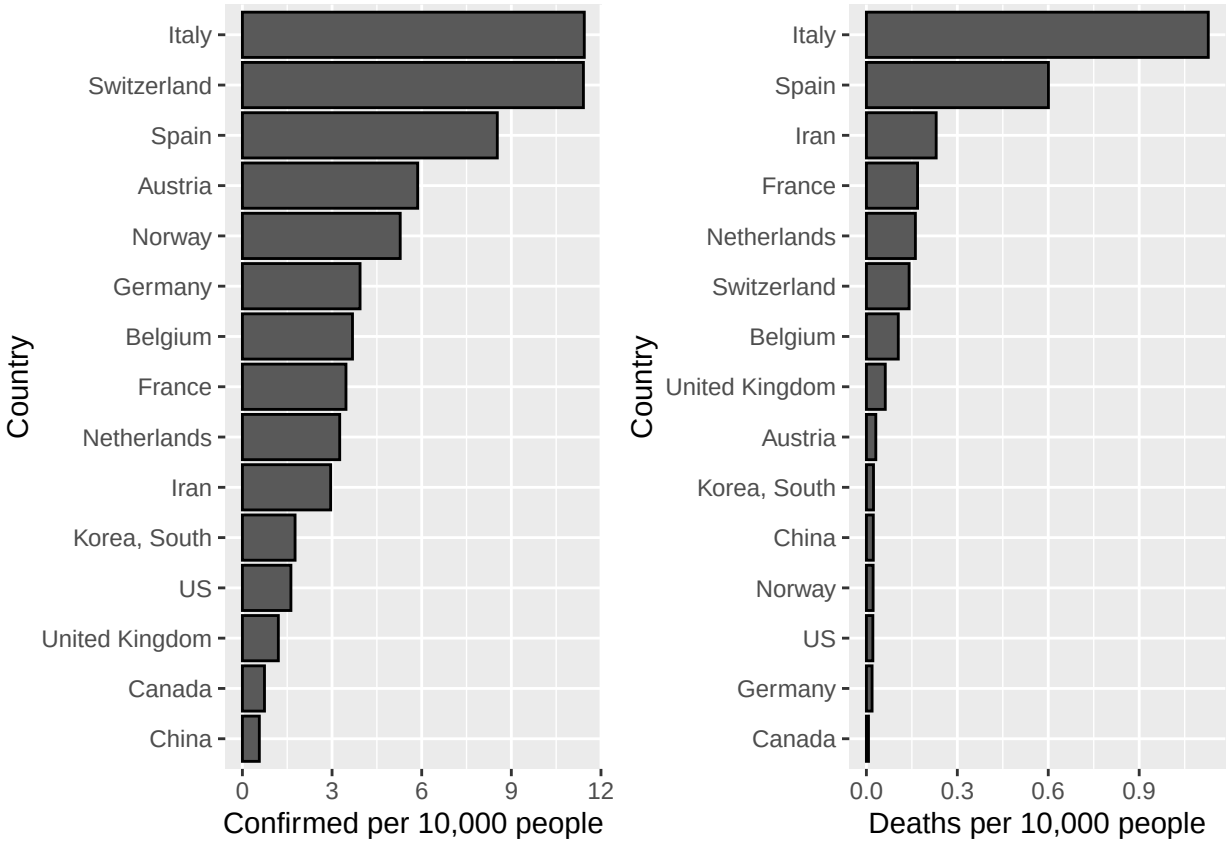
4. For most of the countries which have higher confirmed cases, most of them have D/C % by Density within 0.05. This means to me that the virus transmission seems to be similar across all countries.

5. There can be other factors that contribute to D/C % by Density. If a country has more people older than a certain age, it will be more affected since covid-19 has much worse impact on older people. Italy is a great exmaple. Better medical facilities will have positive effect on this.

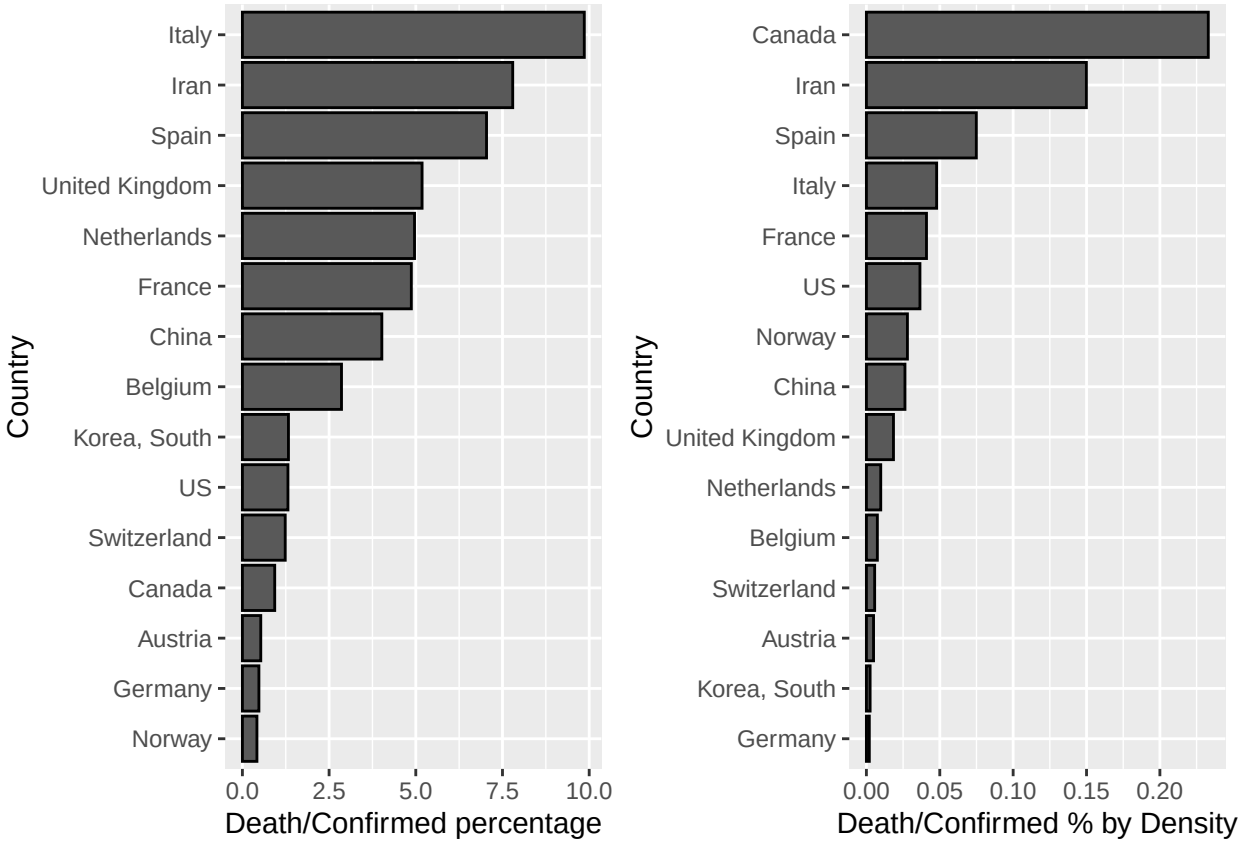
```
all <- all %>%
  mutate(Country.Region = as.factor(Country.Region)) %>%
  slice(1:select_top)

p_confirmed <- all %>%
  mutate(Country.Region = reorder(Country.Region, Confirmed, FUN = mean)) %>%
  ggplot(aes(Country.Region, Confirmed)) +
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Total Confirmed",
       y = "Confirmed",
       x = "Country") +
  coord_flip()

p_deaths <- all %>%
  mutate(Country.Region = reorder(Country.Region, Deaths, FUN = mean)) %>%
  ggplot(aes(Country.Region, Deaths)) +
  geom_bar(stat = "identity", color = "black") +
  labs(y = "Deaths",
       x = "Country") +
  coord_flip()

grid.arrange(p_confirmed, p_deaths, ncol = 2)
```



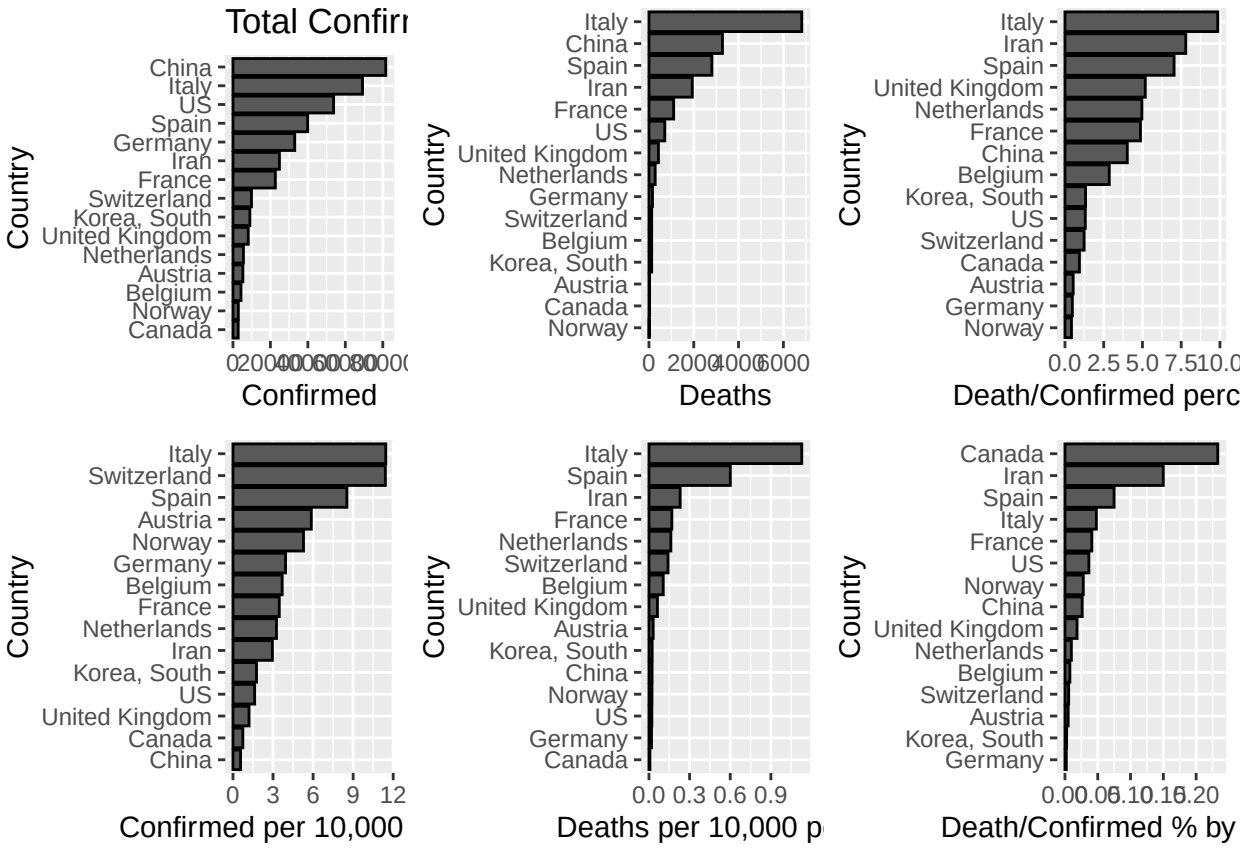

```r
p_c_population <- all %>%
  mutate(Country.Region = reorder(Country.Region, `C/Population`, FUN = mean)) %>%
  ggplot(aes(Country.Region, `C/Population`)) +
  geom_bar(stat = "identity", color = "black") +
  labs(y = "Confirmed per 10,000 people",
       x = "Country") +
  coord_flip()

p_d_population <- all %>%
  mutate(Country.Region = reorder(Country.Region, `D/Population`, FUN = mean)) %>%
  ggplot(aes(Country.Region, `D/Population`)) +
  geom_bar(stat = "identity", color = "black") +
  labs(y = "Deaths per 10,000 people",
       x = "Country") +
  coord_flip()

grid.arrange(p_c_population, p_d_population, ncol = 2)
```




```r
p_deaths_confirmed <- all %>%
  mutate(Country.Region = reorder(Country.Region, `D/C %`, FUN = mean)) %>%
  ggplot(aes(Country.Region, `D/C %`)) +
  geom_bar(stat = "identity", color = "black") +
  labs(y = "Death/Confirmed percentage",
       x = "Country") +
  coord_flip()

p_deaths_confirmed_by_density <- all %>%
  mutate(Country.Region = reorder(Country.Region, `D/C % by Density`, FUN = mean)) %>%
  ggplot(aes(Country.Region, `D/C % by Density`)) +
  geom_bar(stat = "identity", color = "black") +
  labs(y = "Death/Confirmed % by Density",
       x = "Country") +
  coord_flip()

grid.arrange(p_deaths_confirmed, p_deaths_confirmed_by_density, ncol = 2)
```




```
grid.arrange(p_confirmed, p_deaths, p_deaths_confirmed,
             p_c_population, p_d_population, p_deaths_confirmed_by_density,
             ncol = 3, nrow = 2)
```


Total Confir
China
Italy
US
Spain
Germany
Iran
France
Switzerland
Korea, South
United Kingdom
Netherlands
Austria
Belgium
Norway
Canada
Country
Confirmed
Italy
China
Spain
Iran
France
US
United Kingdom
Netherlands
Germany
Switzerland
Belgium
Korea, South
Austria
Canada
Norway
0 2000 4000 6000
Deaths
Italy
Iran
Spain
United Kingdom
Netherlands
France
China
Belgium
Korea, South
US
Switzerland
Canada
Austria
Germany
Norway
0.0 2.5 5.0 7.5 10.0
Death/Confirmed perc
Italy
Switzerland
Spain
Austria
Norway
Germany
Belgium
France
Netherlands
Iran
Korea, South
US
United Kingdom
Canada
China
0 3 6 9 12
Confirmed per 10,000
Italy
Spain
Iran
France
Netherlands
Switzerland
Belgium
United Kingdom
Austria
Korea, South
China
Norway
US
Germany
Canada
0.0 0.3 0.6 0.9
Deaths per 10,000 p
Canada
Iran
Spain
Italy
France
US
Norway
China
United Kingdom
Netherlands
Belgium
Switzerland
Austria
Korea, South
Germany
0.00 0.05 0.10 0.15 0.20
Death/Confirmed % by