COVID19

2024-03-06

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
library(knitr)

confirmed_global <- read_csv("time_series_covid19_confirmed_global.csv")
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
deaths_global <- read_csv("time_series_covid19_deaths_global.csv")
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
confirmed_global <- confirmed_global %>%
  filter(!`Country/Region` %in% c("Diamond Princess",
                                  "Holy See",
                                  "MS Zaandam",
                                  "Summer Olympics 2020",
                                  "Winter Olympics 2022"))
deaths_global <- deaths_global %>%
  filter(!`Country/Region` %in% c("Diamond Princess",
                                  "Holy See",
                                  "MS Zaandam",
                                  "Summer Olympics 2020",
                                  "Winter Olympics 2022"))
```

Inital Observations

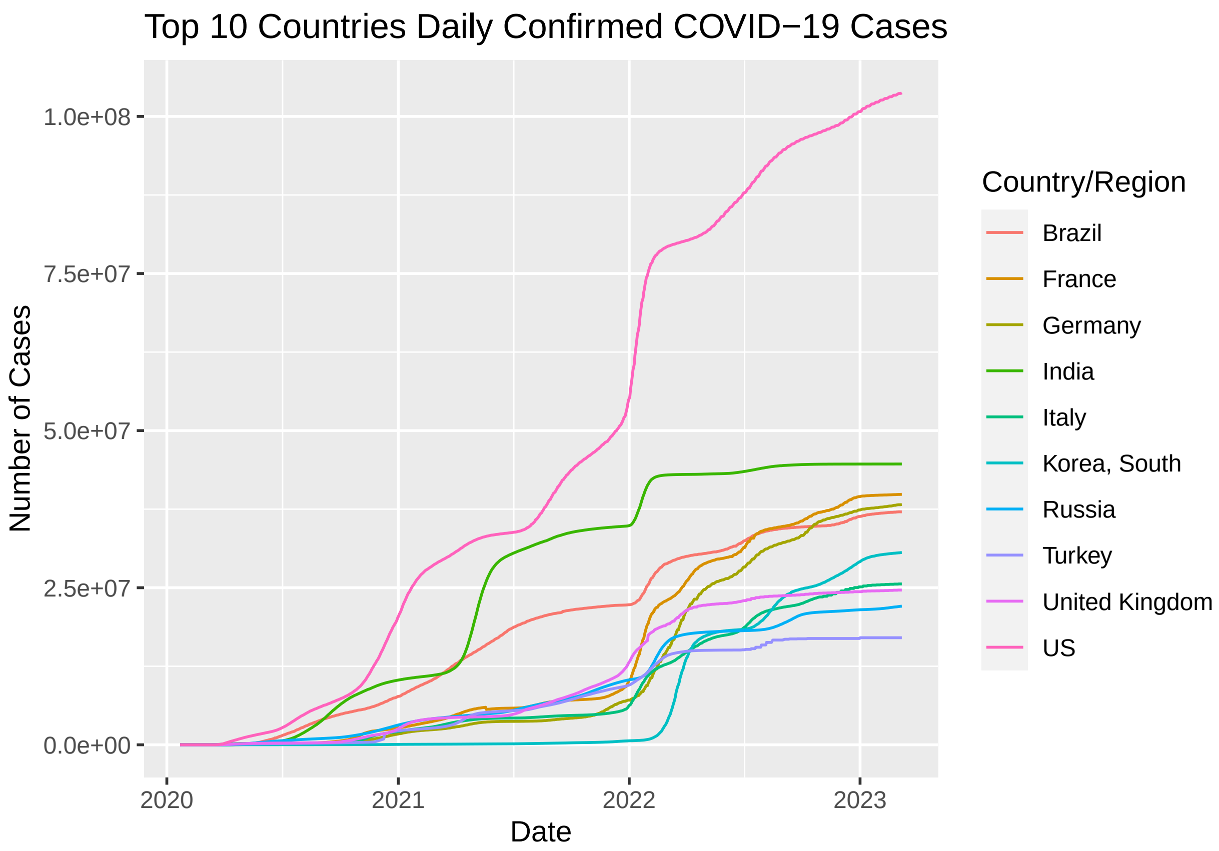
I selected the time series data on confirmed cases and deaths. Personally, I am always drawn to time series data because of the underlying theme of time ultimately ties the data together leading to interesting conclusions. My first impressions are this is an incredibly well put together data set. This being said there were some weird values in the 'Country/Region' column such as Diamond Princess, Holy See, MS Zaandam, Summer Olympics 2020, and Winter Olympics 2022. I ended up removing the data corresponding to these values as they are not countries and largely irrelevant for my analysis. Setting the ground for standard biases we would expect from a data set like this: - Data is a result of collection and various countries might have poor/inaccurate reporting - A country doesn't have the resources to allocate to data collection - A country intentionally doesn't report accurately to protect its security/privacy - Deaths from COVID might have a varied interpretation depending on the country

The first graphs I put together were time series data on cases daily and deaths daily by country. Naturally, with so many countries this graph was impossible to decipher so I selected the top 10 countries.

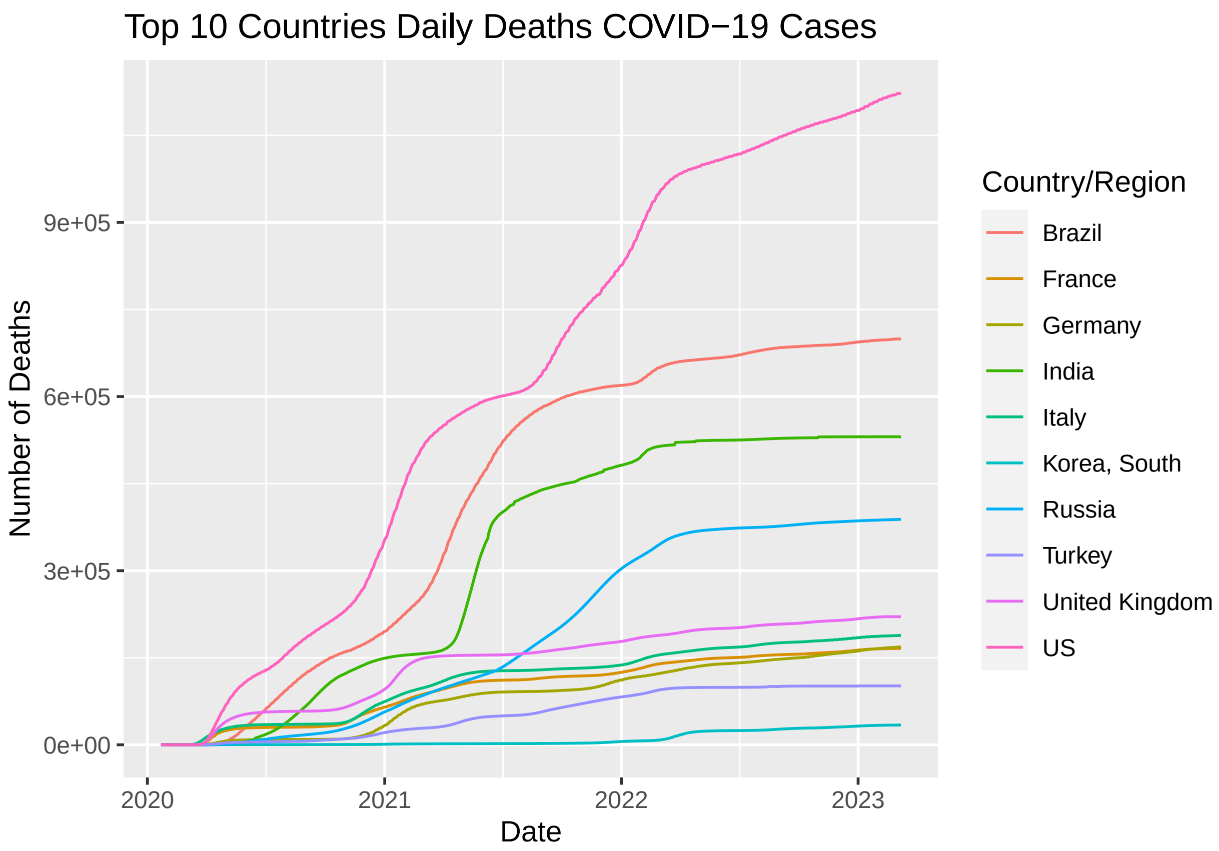
```
tots <-  confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  group_by(`Country/Region`) %>%
  summarise(total = sum(value)) %>%
  top_n(10)
```

```
## Selecting by total
```

```
confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`,
               names_to = "date",
               values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  filter(`Country/Region` %in% tots$`Country/Region`) %>%
  ggplot(aes(x = date,
             y = value,
             color = `Country/Region`,
             group = `Country/Region`)) +
  geom_line() +
  labs(x = "Date",
       y = "Number of Cases",
       title = "Top 10 Countries Daily Confirmed COVID-19 Cases")
```



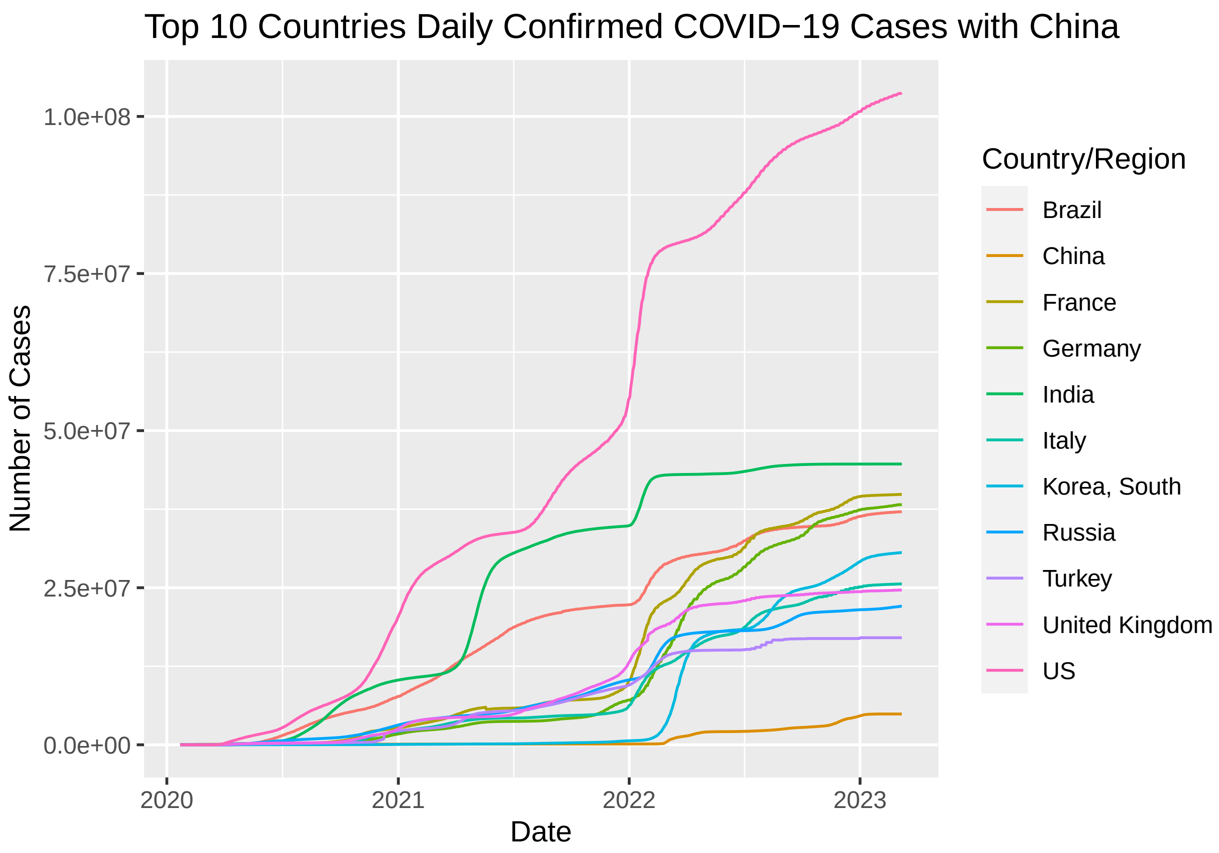
```r
deaths_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`,
               names_to = "date",
               values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  filter(`Country/Region` %in% tots$`Country/Region`) %>%
  ggplot(aes(x = date,
             y = value,
             color = `Country/Region`,
             group = `Country/Region`)) +
  geom_line() +
  labs(x = "Date",
       y = "Number of Deaths",
       title = "Top 10 Countries Daily Deaths COVID-19 Cases")
```



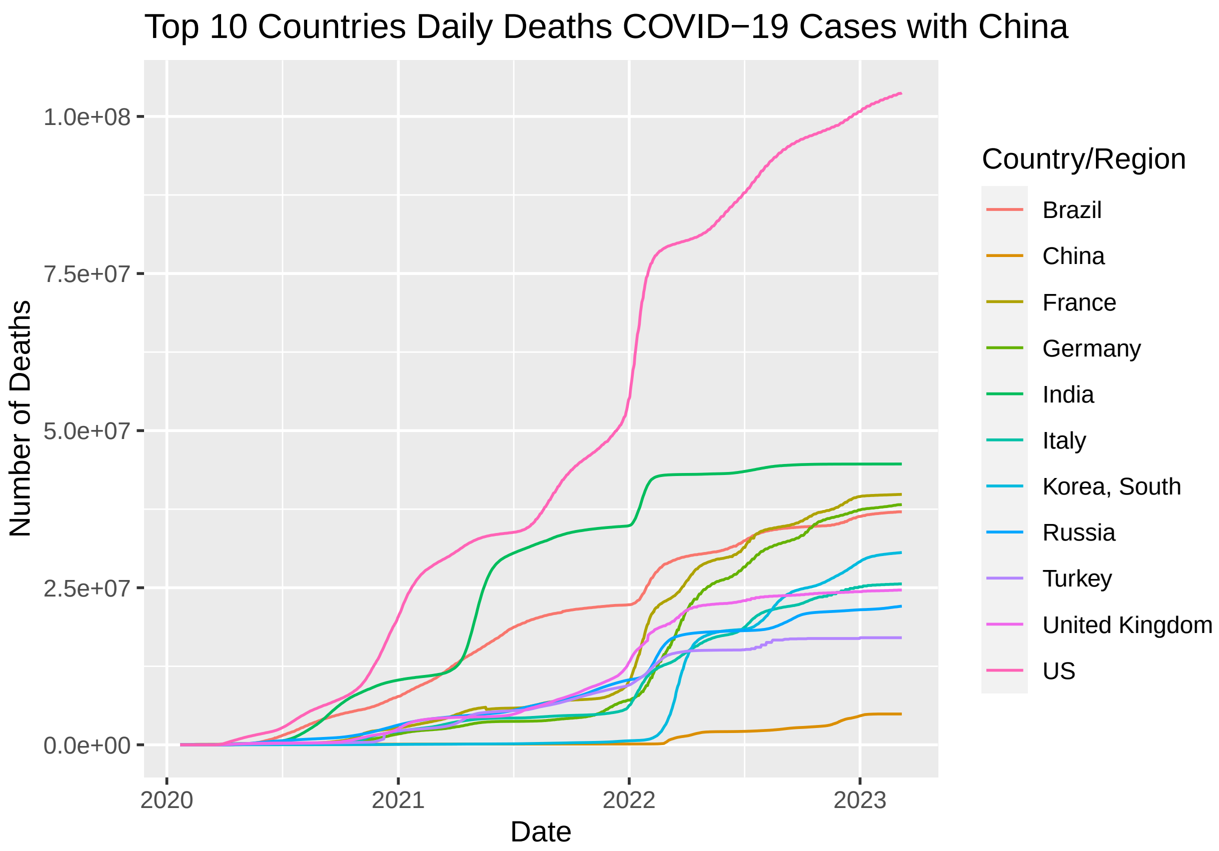
Some interesting differences I spot between these graphs is the dramatic shift in Brazil and Russia from confirmed case to deaths. Most of the other countries remain about the same between the graphs, but both Brazil and Russia shoot up on the deaths graph implying that their practices around COVID were less effective at reducing deaths.

I'll be honest I was surprised to not see China on this list given that COVID was born there. Maybe this is a bias of mine, but I decided to reproduce the prior graphs with China explicitly added.

```
confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`,
               names_to = "date",
               values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  filter(`Country/Region` %in% append(tots$`Country/Region`, c("China"))) %>%
  ggplot(aes(x = date,
             y = value,
             color = `Country/Region`,
             group = `Country/Region`)) +
  geom_line() +
  labs(x = "Date",
       y = "Number of Cases",
       title = "Top 10 Countries Daily Confirmed COVID-19 Cases with China")
```



```r
confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`,
               names_to = "date",
               values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  filter(`Country/Region` %in% append(tots$`Country/Region`, c("China"))) %>%
  ggplot(aes(x = date,
             y = value,
             color = `Country/Region`,
             group = `Country/Region`)) +
  geom_line() +
  labs(x = "Date",
       y = "Number of Deaths",
       title = "Top 10 Countries Daily Deaths COVID-19 Cases with China")
```



Now I believe these graphs essentially confirm a few of the suspected biases from my initial observations. China has hilariously few cases and deaths. I say hilariously because the numbers are so small with respect to China's enormous population and the fact that they should have been more blindsided than any other country since the virus was source from China giving other countries a chance to take precautions in advanced. I conclude from this that China has either under reported in order to hid how severe the impact of COVID effected them or they were simply unable to accurately report the number of cases and deaths.

After looking at this graphs I wanted to put some of these graphs into perspective relative to their respective populations. To do this I had to gather some more data since country populations weren't part of the original data set. I got my data from the worldbank which ultimately led to a second wave of cleaning in order to make this data cooperate with the initial data set. For the sake of simplicity, I simply aligned all of the names to the corresponding countries and joined the data sets. This involved me identifying which countries were matching under different aliases and which countries were included in one data set and not the other. In general, I used the original data set as a key and merge the new population data into it. This did lead to me dropping Antarctica, Burma, and Taiwan from the original data set since their population values were not present in the new data I was pulling in. Finally, to maintain reproduceability I have uploaded the population data set that I used post cleaning and manipulation to my Github along with this report.

```
population <- read_csv("population.csv")
```

```
## Rows: 219 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Country/Region
## dbl (1): Population
## lgl (1): Drop
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

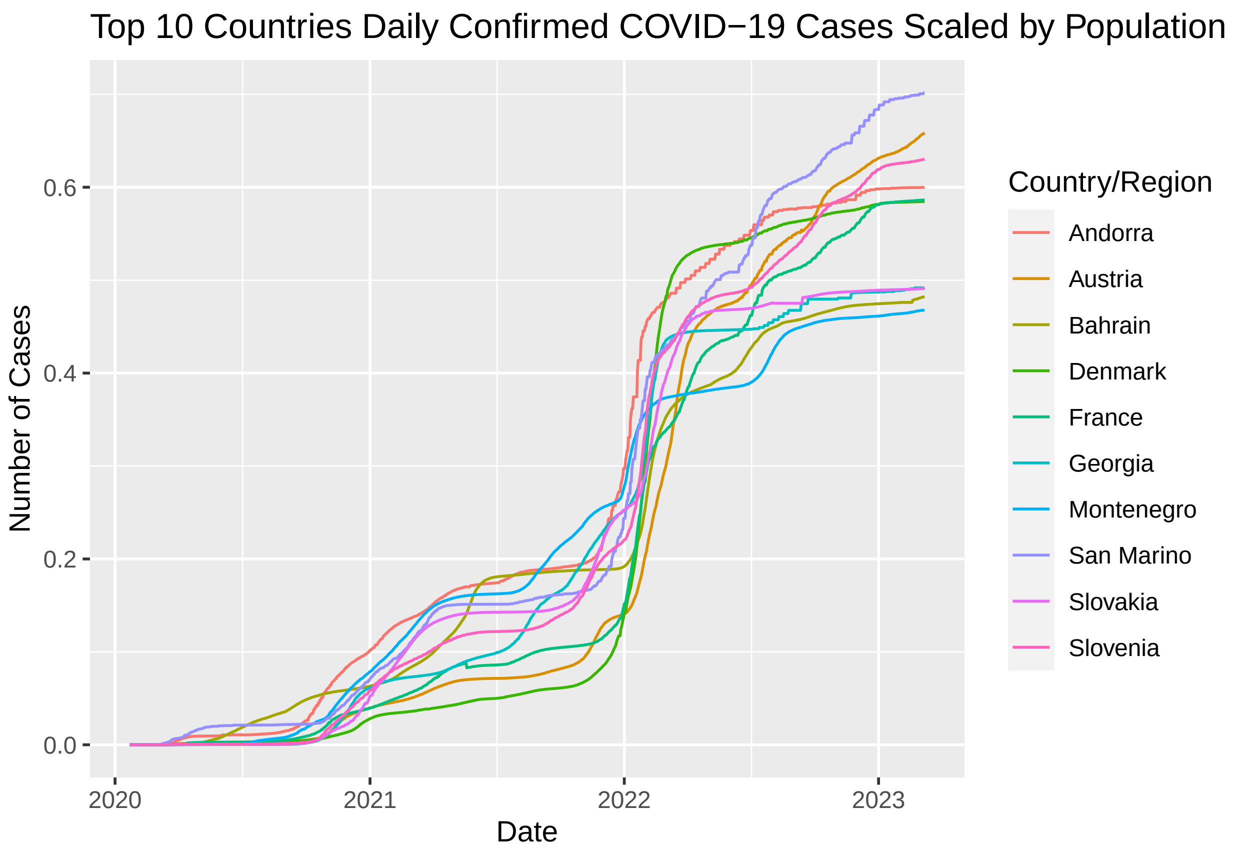
```
population <- population %>%
  filter(is.na(Drop)) %>%
  select(-Drop)
```

```
confirmed_global <- confirmed_global %>%
  filter(!`Country/Region` %in% c("Antarctica",
                                  "Burma",
                                  "Taiwan*"))
deaths_global <- deaths_global %>%
  filter(!`Country/Region` %in% c("Antarctica",
                                  "Burma",
                                  "Taiwan*"))
```

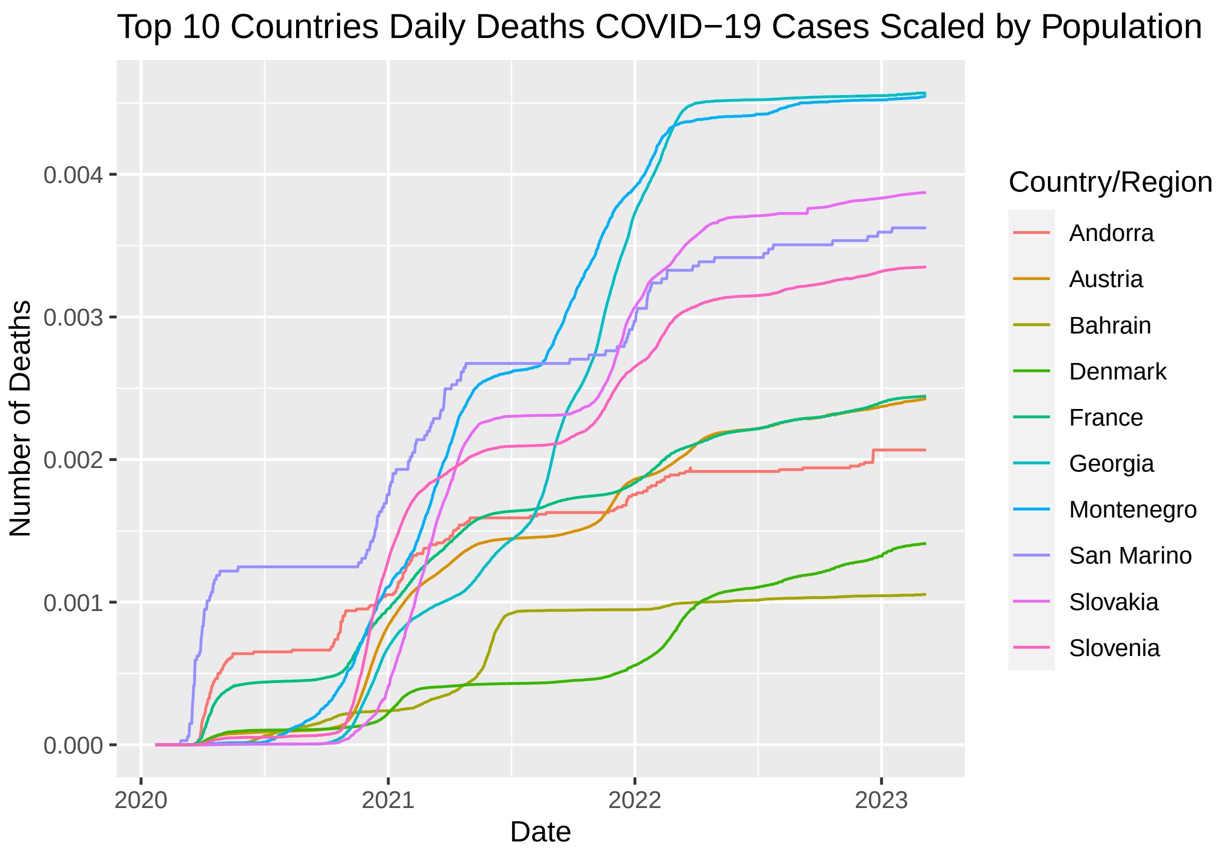
```
tots_scaled <- confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  full_join(population,
            by = join_by(`Country/Region`),
            relationship = 'many-to-many') %>%
  mutate(scaled_cases = value / Population) %>%
  group_by(`Country/Region`) %>%
  summarise(total = sum(scaled_cases)) %>%
  arrange(desc(total)) %>%
  top_n(10)
```

```
## Selecting by total
```

```
confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  full_join(population,
            by = join_by(`Country/Region`),
            relationship = 'many-to-many') %>%
  mutate(scaled_cases = value / Population) %>%
  filter(`Country/Region` %in% tots_scaled$`Country/Region`) %>%
  ggplot(aes(x = date,
             y = scaled_cases,
             color = `Country/Region`,
             group = `Country/Region`)) +
  geom_line() +
  labs(x = "Date",
       y = "Number of Cases",
       title = "Top 10 Countries Daily Confirmed COVID-19 Cases Scaled by Population")
```



```r
deaths_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  full_join(population,
            by = join_by(`Country/Region`),
            relationship = 'many-to-many') %>%
  mutate(scaled_cases = value / Population) %>%
  filter(`Country/Region` %in% tots_scaled$`Country/Region`) %>%
  ggplot(aes(x = date,
             y = scaled_cases,
             color = `Country/Region`,
             group = `Country/Region`)) +
  geom_line() +
  labs(x = "Date",
       y = "Number of Deaths",
       title = "Top 10 Countries Daily Deaths COVID-19 Cases Scaled by Population")
```



I was surprised to see nearly none of the same countries listed in the previous graphs when scaling for country population. The only exception to this was France which broke the top 10 in both graphs. Furthermore, I didn't expect to see so many European countries on this list. Once country is scaled for, this graph conveys how effectively countries were able to mitigate the spread and deaths of COVID respective to their countries population. Essentially these 10 countries were the worst performers with the most cases and deaths. This shows that the prior charts were a little misleading as countries such as the US mainly suffered such massive casualties and spreading of the virus as a result of their larger population. Also likely why we saw other large population countries such as India and Brazil.

I wanted to look at these scaled values more clearly in a table format to better display the overlap and separation.

```
tots <- confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  group_by(`Country/Region`) %>%
  summarise(total = sum(value)) %>%
  arrange(desc(total))
```

```
tots_scaled <- confirmed_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  full_join(population,
            by = join_by(`Country/Region`),
            relationship = 'many-to-many') %>%
  mutate(scaled_cases = value / Population) %>%
  group_by(`Country/Region`) %>%
  summarise(`Scaled Total` = sum(scaled_cases))

ordered_tots_scaled <- tots_scaled[order(tots_scaled$`Scaled Total`,
                                         decreasing = TRUE),] %>%
  mutate(Rank = c(1:length(`Country/Region`)))
```

```
table_10s <- ordered_tots_scaled %>%
  filter(`Country/Region` %in% append(c(ordered_tots_scaled$`Country/Region`[1:10]),
                                    c(tots$`Country/Region`[1:10])))
kable(table_10s, caption="Table of the top 10 countries by COVID cases before and after scaling for popu
```

Table 1: Table of the top 10 countries by COVID cases before and after scaling for population

Country/Region	Scaled Total	Rank
Andorra	306.92066	1
San Marino	301.89750	2
Slovenia	269.87339	3
Slovakia	248.85957	4
Montenegro	248.00182	5
Denmark	247.02576	6
Austria	243.81031	7
Bahrain	242.21437	8
Georgia	241.59012	9
France	236.36509	10
United Kingdom	180.57878	29
Italy	170.63939	31
Korea, South	163.42400	33
Germany	162.86543	34
US	161.15028	35
Turkey	103.83277	58
Brazil	98.20849	61
Russia	73.18856	79
India	20.52426	125

As we can see hear this dramatically shifts the respective telling we saw in a some of our first graphs. The countries of Russia, India, and Brazil despite the earlier spikes in COVID deaths did a solid job at mitigating their COVID cases with respect to there population.

```
tots <- deaths_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  group_by(`Country/Region`) %>%
  summarise(total = sum(value)) %>%
  arrange(desc(total))
```

```r
tots_scaled <- deaths_global %>%
  group_by(`Country/Region`) %>%
  summarize(across(colnames(confirmed_global)[5:1146], sum)) %>%
  pivot_longer(cols = -`Country/Region`, names_to = "date", values_to = "value") %>%
  mutate(date = mdy(date)) %>%
  full_join(population,
            by = join_by(`Country/Region`),
            relationship = 'many-to-many') %>%
  mutate(scaled_cases = value / Population) %>%
  group_by(`Country/Region`) %>%
  summarise(`Scaled Total` = sum(scaled_cases))

ordered_tots_scaled <- tots_scaled[order(tots_scaled$`Scaled Total`,
                                        decreasing = TRUE),] %>%
  mutate(Rank = c(1:length(`Country/Region`)))
```

```r
table_10s <- ordered_tots_scaled %>%
  filter(`Country/Region` %in% append(c(ordered_tots_scaled$`Country/Region`[1:10]),
                                       c(tots$`Country/Region`[1:10])))
kable(table_10s, caption="Table of the top 10 countries by COVID deaths before and after scaling for pop
```

Table 2: Table of the top 10 countries by COVID deaths before and after scaling for population

Country/Region	Scaled Total	Rank
Peru	5.0082929	1
Bulgaria	3.5349635	2
Bosnia and Herzegovina	3.1947465	3
Hungary	3.1260793	4
North Macedonia	2.9906871	5
Moldova	2.9664736	6
Montenegro	2.9248784	7
San Marino	2.7781937	8
Czechia	2.6438609	9
Croatia	2.6055550	10
United Kingdom	2.3982770	12
Brazil	2.2640556	14
Italy	2.1674167	18
US	2.1385538	19
Colombia	1.9379506	25
Mexico	1.8881899	26
France	1.6660585	32
Russia	1.5293941	35
India	0.2571249	110

Now looking at the same table with respect to deaths again better paints the full picture. We can see that the US actually gets substantially worse as it rises from 38th to 19th. This table does rush to the defense of France who was within the top 10 on both the scaled and unscaled totals graphs as it drops from the top 10 in cases to 32nd for deaths which is what is most important to mitigate.

One last observation after reflecting on the work I have done throughout this report, is the exceptional spike early into 2022. This seems to apply to nearly every country to some degree across all graphs cases and deaths alike. After doing some more research, I suspect this is because of the new Omicron variant of COVID 19 that was detected in November of 2021. This strain ended up being much more contagious and lethal which likely contributed to the mostly global increase we see across the graphs.