COVID-19 Data

Holly White

2025-04-12

Raw Data

The data sets I will be using comes from this Johns Hopkins Github page. They combines data from the Centers for Disease Control (CDC) and the World Health Organization (WHO) for for data on the number of cases, number of deaths, and number of recoveries bot in the United States and globally. As of March 10, 2023, Johns Hopkins ceased updating the data sets.

```
url_in <-
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_ti

file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv",
                "time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_global.csv")

urls <- str_c(url_in, file_names)

US_cases <- read.csv(urls[1])
US_deaths <- read.csv(urls[2])
global_cases <- read.csv(urls[3])
global_deaths <- read.csv(urls[4])
```

Motivating Questions

After briefly scanning through the available data, a few questions arose.

1. Which US states were impacted the most, both in terms of cases and deaths? How did these top states compare?
2. Which countries were impacted the most by COVID-19, both in terms of number of cases and number of deaths? How did these countries compare?
3. Both in the United States and globally, where was there the least impact?

Tidying the Data

Starting with the global data sets (global_deaths, global_cases), I can see that there are columns for Province.State, Country.Region, Latitude (Lat), Longitude (Long), and a column for each date beginning at January 22, 2020. I know that I will not need the latitude and longitude for my analysis, so I will remove that. I also want to create a date column so that each date value has its own row, change the dates into the correct format, and join the two data sets into one with all the columns together. In looking at the cases column, there are a lot of rows with zeroes, so I chose to filter these out so that the data only contains information starting with the first case for each country.

```r
global_cases <- global_cases %>%
  pivot_longer(cols = -c("Province.State", "Country.Region", "Lat", "Long") ,
               names_to = "date", values_to = "cases") %>%
  select(-c(Lat, Long))

global_deaths <- global_deaths %>%
  pivot_longer(cols = -c("Province.State", "Country.Region", "Lat", "Long") ,
               names_to = "date", values_to = "deaths") %>%
  select(-c(Lat, Long))

global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = "Country.Region",
         Province_State = "Province.State") %>%
  mutate(date = gsub("^X", "", date),
         date = gsub("\\.", "/", date),
         date = as.Date(date, format = "%m/%d/%y")) %>%
  filter(cases > 0 )
```

I wanted to do the same thing with the US data, which contains a few extra columns that are not necessary to the analysis I want to perform. I also noticed that there is data on two US cruise ships, which I do not think is necessary for my analysis, so I removed those two.

```r
US_cases <- US_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
   mutate(date = gsub("^X", "", date),
         date = gsub("\\.", "/", date),
         date = as.Date(date, format = "%m/%d/%y")) %>%
  select(-c(Lat, Long_))

US_deaths <- US_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to = "date",
               values_to = "deaths") %>%
  select(-c(Lat, Long_)) %>%
   mutate(date = gsub("^X", "", date),
         date = gsub("\\.", "/", date),
         date = as.Date(date, format = "%m/%d/%y")) %>%
  select(Admin2:deaths)

US <- US_cases %>%
  full_join(US_deaths) %>%
  filter(!Province_State %in% c("Grand Princess", "Diamond Princess")) %>%
  filter(cases > 0 )
```

After combining the US data, I needed to create a "Combined_Key" variable in my global data set. Also, I noticed that it contained population data, while the global data set does not, so I added it using another csv file from the same Johns Hopkins github.

```
global <- global %>%
  unite("Combined_Key",
        c(Province_State, Country_Region),
        sep = ", ",
        na.rm = TRUE,
        remove = FALSE)
```

```
uid_lookup_url <-
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_ISO_FIPS_Loo
uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))
global <- global %>%
  mutate(Province_State = ifelse(Province_State == "" | is.na(Province_State),
                                 NA, Province_State)) %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>%
  select(-c(UID, FIPS)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population, Combined_Key)
```

Finally, I wanted to add columns for the number of cases per million and the number of deaths per million for both US states and globally.

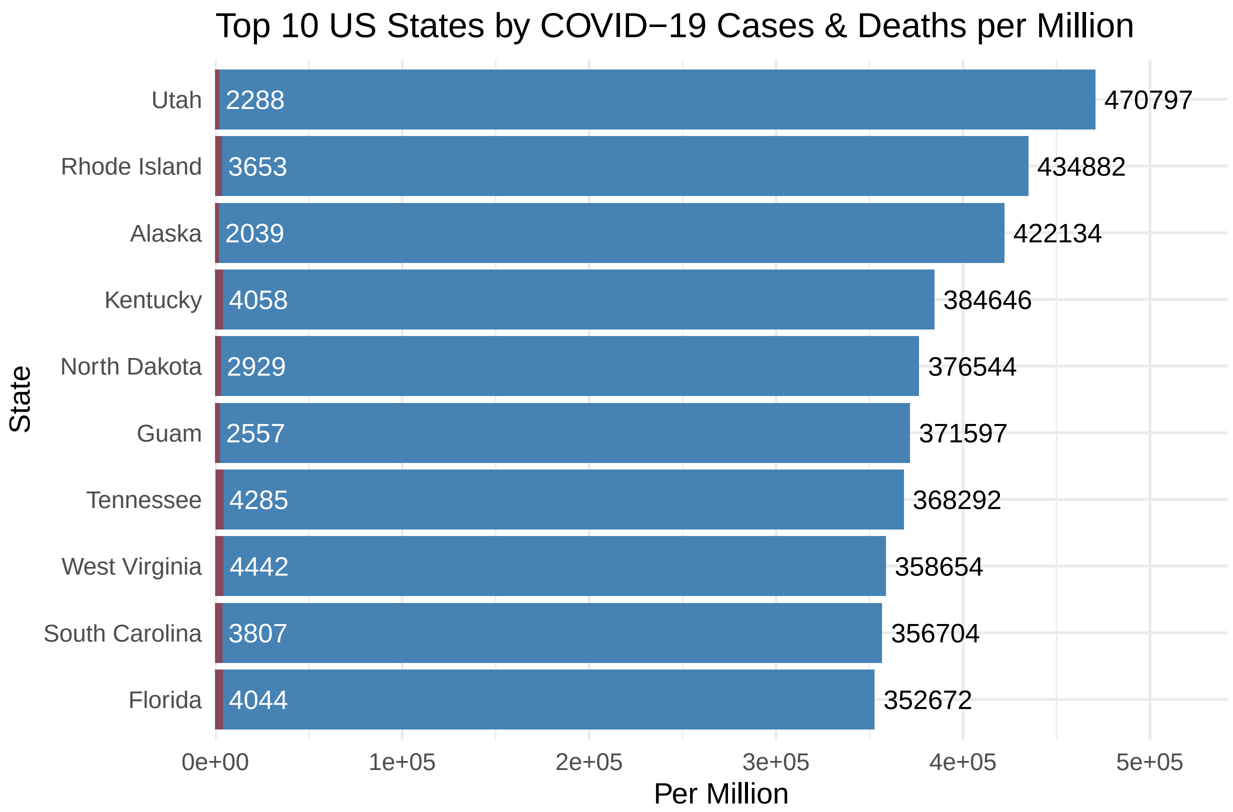
```
global_totals <- global %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
   mutate(
    deaths = ifelse(is.na(deaths), 0, deaths),
    cases_per_mill = ifelse(Population > 0, cases * 1e6 / Population, NA),
    deaths_per_mill = ifelse(Population > 0, deaths * 1e6 / Population, NA)
  ) %>%
  select(Country_Region, date,
         cases, deaths, cases_per_mill, deaths_per_mill, Population) %>%
  ungroup()
```

```
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases=sum(cases), deaths=sum(deaths),
            Population= sum(Population)) %>%
  mutate(
    deaths = ifelse(is.na(deaths), 0, deaths),
    cases_per_mill = ifelse(Population > 0, cases * 1e6 / Population, NA),
    deaths_per_mill = ifelse(Population > 0, deaths * 1e6 / Population, NA)
  ) %>%
  select(Province_State, Country_Region, date,
         cases, deaths, deaths_per_mill, cases_per_mill, Population) %>%
  ungroup()

US_totals <- US_by_state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
  mutate(cases_per_mill = cases * 1000000 / Population) %>%
  select(Country_Region, date,
         cases, deaths, cases_per_mill, deaths_per_mill, Population) %>%
  ungroup()
```

Visualizations

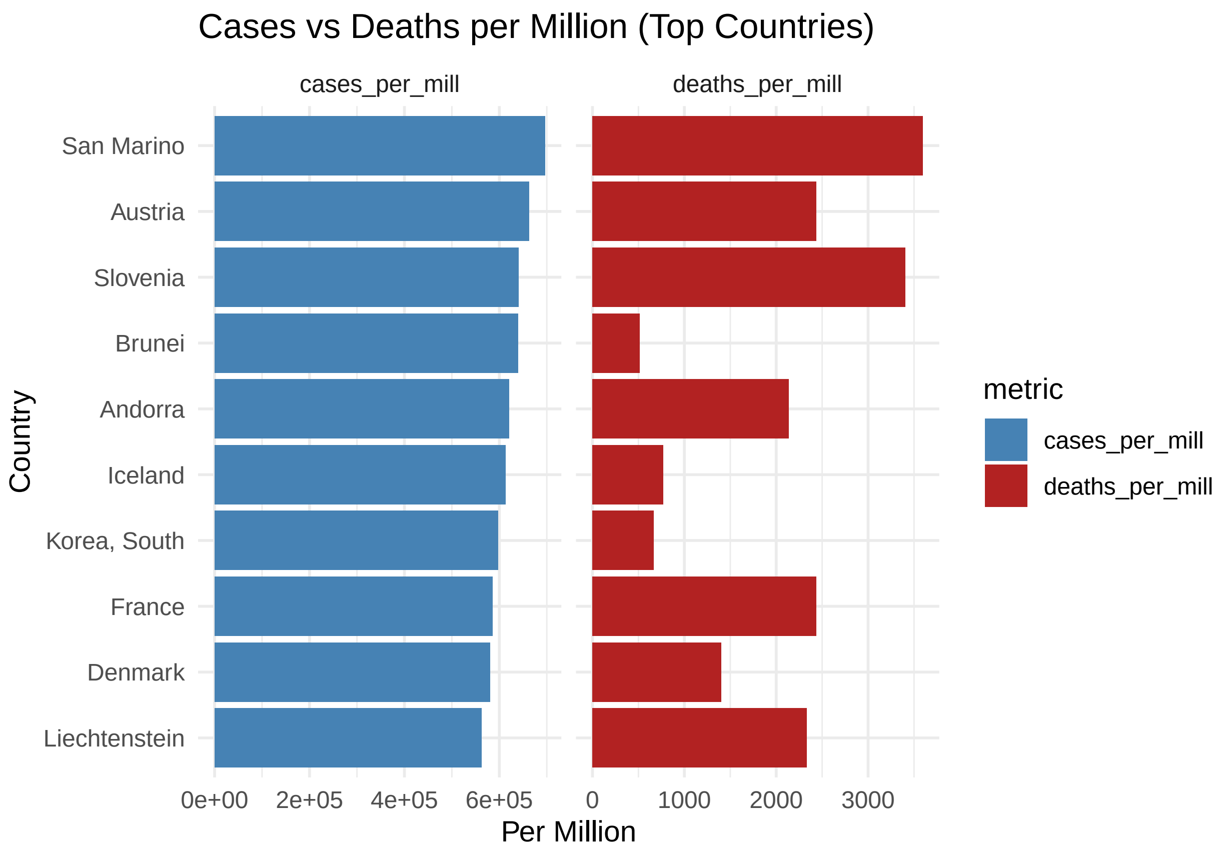
As my first motivating question was *Which US states were impacted the most, both in terms of cases and deaths? How did these top states compare?* I will need to look at each state's data based on the cases and deaths.

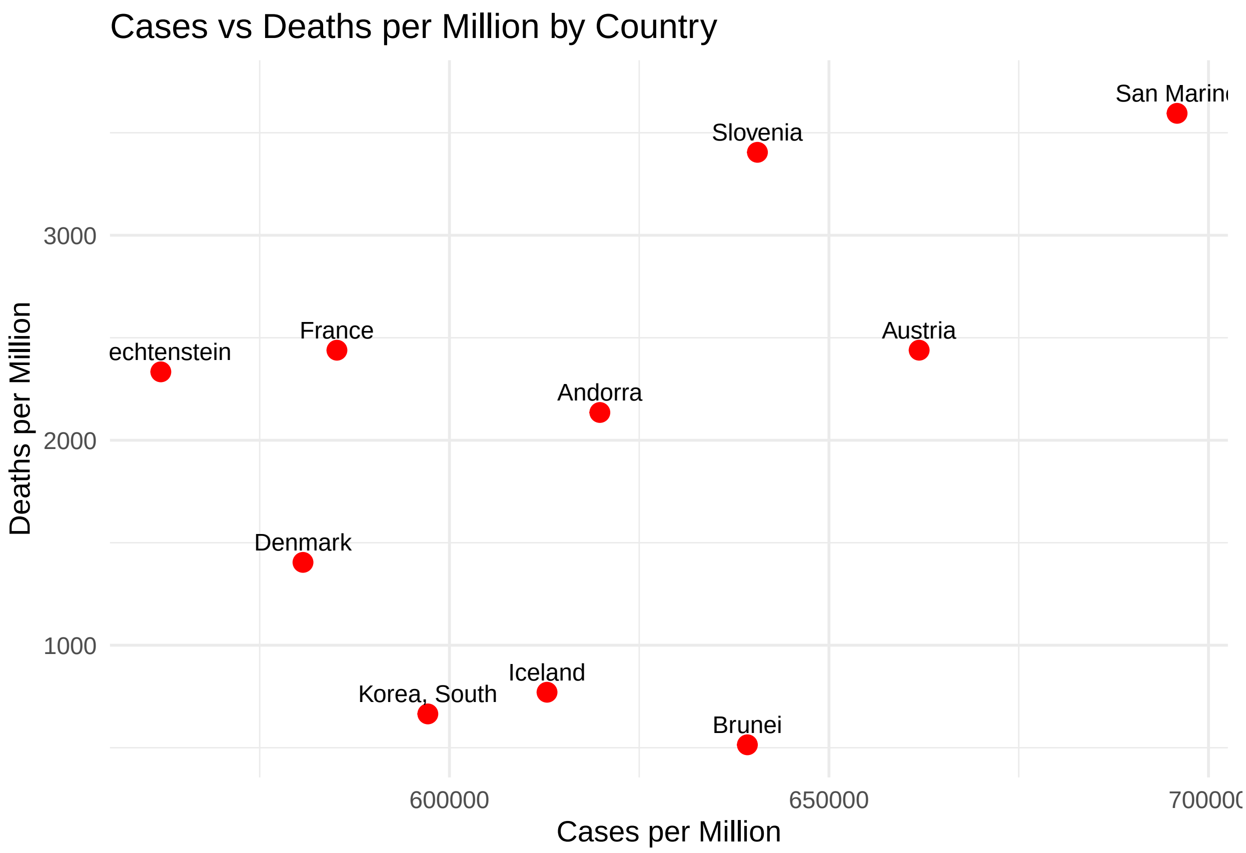


Top 10 US States by COVID−19 Cases & Deaths per Million

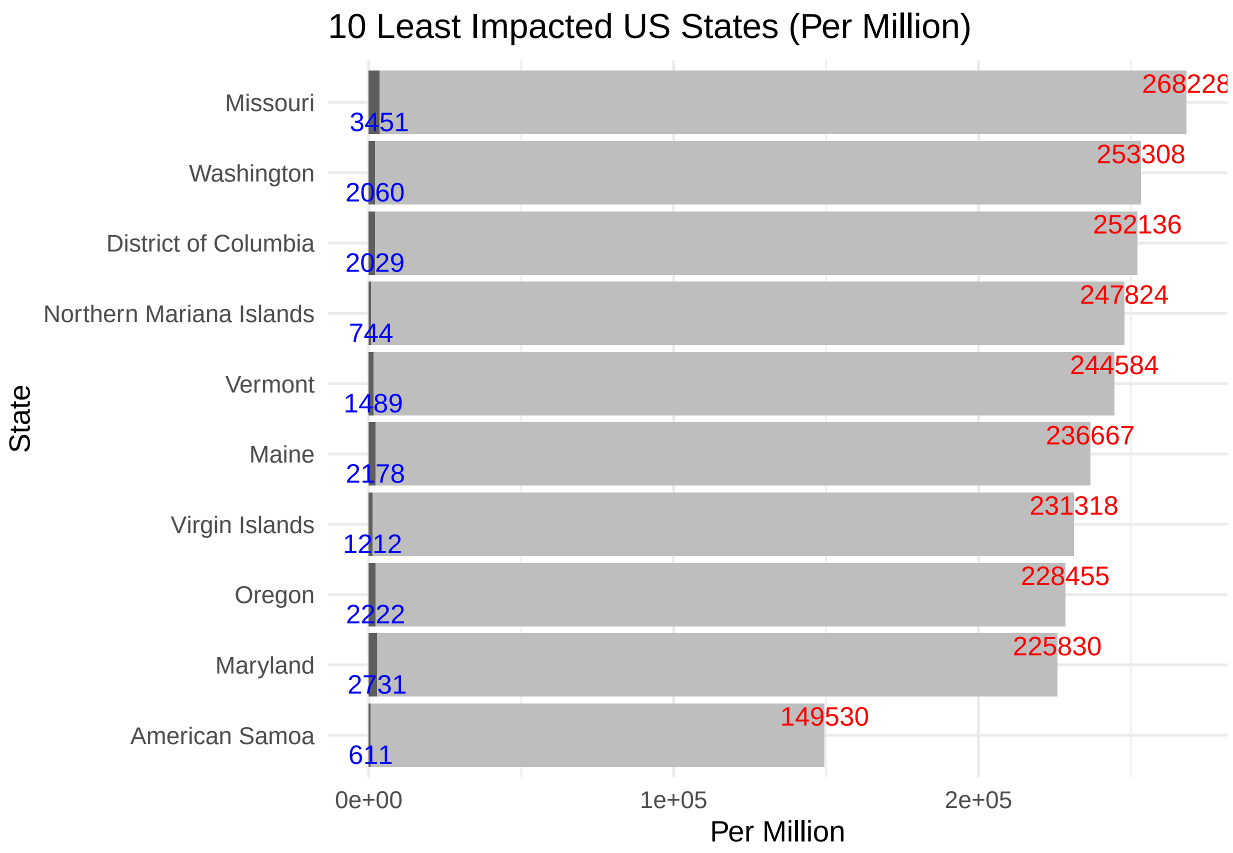
Blue = Cases per Million, Red = Deaths per Million

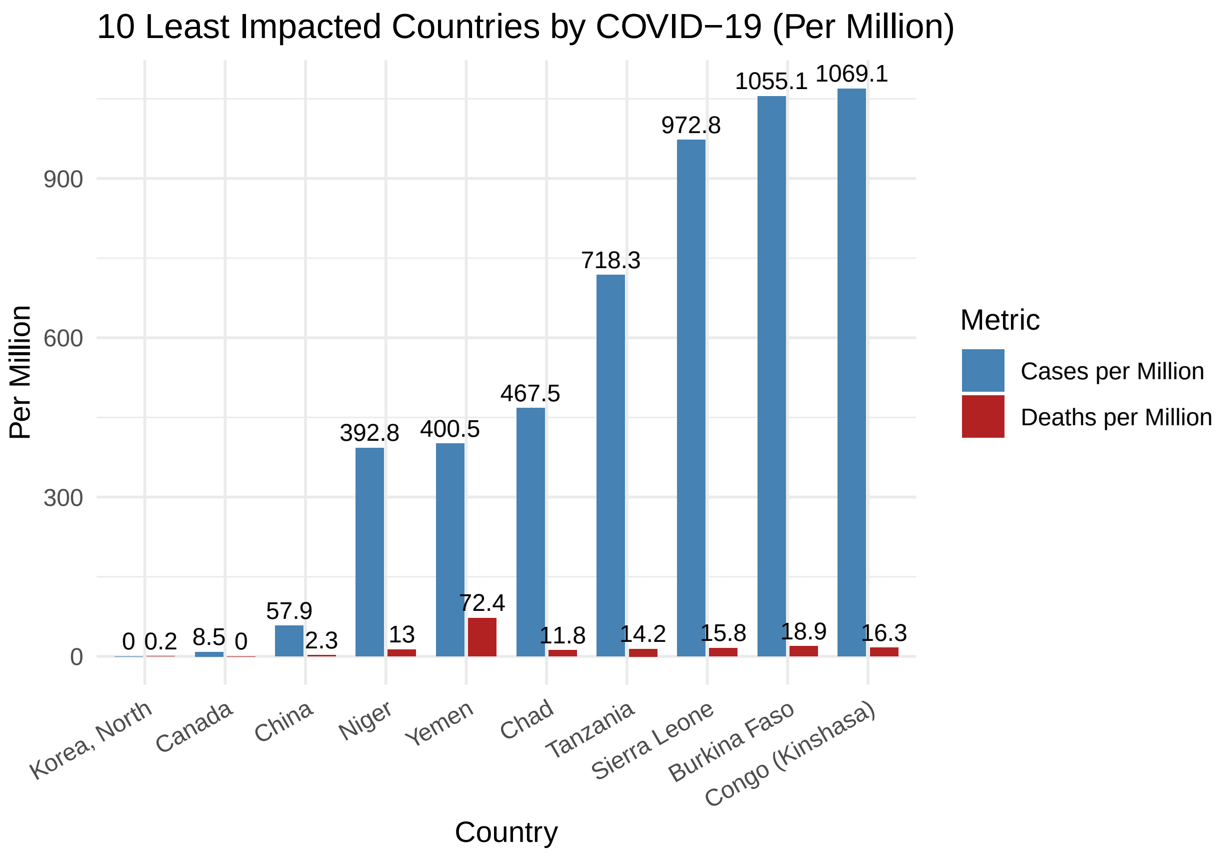
My second motivating question was *Which countries were impacted the most by COVID-19, both in terms of number of cases and number of deaths? How did these countries compare?*


Cases vs Deaths per Million (Top Countries)
cases_per_mill
deaths_per_mill
Country
San Marino
Austria
Slovenia
Brunei
Andorra
Iceland
Korea, South
France
Denmark
Liechtenstein
0e+00
2e+05
4e+05
6e+05
0
1000
2000
3000
Per Million
metric
cases_per_mill
deaths_per_mill



My final motivating question was *Both in the United States and globally, where was there the least impact?*


10 Least Impacted US States (Per Million)
State
Per Million
Missouri
Washington
District of Columbia
Northern Mariana Islands
Vermont
Maine
Virgin Islands
Oregon
Maryland
American Samoa
268228
3451
253308
2060
252136
2029
247824
744
244584
1489
236667
2178
231318
1212
228455
2222
225830
2731
149530
611
0e+00
1e+05
2e+05



Analysis

I was also curious as to whether the population and the number of cases in a country could predict the number of deaths.

```
model <- lm(deaths_per_mill ~ cases_per_mill + Population, data = global_totals)
summary(model)
```

```
##
## Call:
## lm(formula = deaths_per_mill ~ cases_per_mill + Population, data = global_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3184.3  -337.6  -282.9   174.6  5601.5
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.376e+02  2.188e+00 154.309  < 2e-16 ***
## cases_per_mill 5.258e-03  1.499e-05 350.841  < 2e-16 ***
## Population     1.127e-07  1.602e-08   7.033 2.03e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 826.8 on 208249 degrees of freedom
##   (5861 observations deleted due to missingness)
## Multiple R-squared:  0.372,  Adjusted R-squared:  0.372
## F-statistic: 6.168e+04 on 2 and 208249 DF,  p-value: < 2.2e-16
```

This linear regression model shows that the relationship between population, cases per million, and deaths per million is statistically significant, though there are other important factors not accounted for in this relationship.

Conclusion

Based on my analysis, I was able to find answers to the questions I had at the start. I found that, in the United States, Utah, Rhode Island, and Alaska were the most impacted, while Missouri, Washington, and District of Columbia were the least impacted. These findings surprised me, partially due to my personal experience during the COVID-19 pandemic, as I was living in NYC and facing first hand the large numbers of cases and deaths. However, it makes sense, as there is such a large population in New York, not only in the city but the state as a whole, so the overall percentage is relatively low.

Globally, I found that San Marino, Austria, and Slovenia were the most highly impacted, and North Korea, Canada, and China were the least impacted countries. This also shocked me, possibly due to biased thinking, as China was the first country with reported cases of COVID-19. However, this also makes sense due to the large population density in China. Moreover, differing lock down requirements could have played a role in limiting the spread of COVID-19.

Finally, my linear regression model indicates that population and number of cases per million residents have a statistically significant relationship with the number of deaths per million, though there are other factors not accounted for that could explain the number of deaths. These could be severity of lock down protocols, overall access to proper medical care, among other things. An analysis on the most impacted and least impacted countries, including data on these other factors, could be beneficial in explaining why these were the effects of COVID-19, as well as inform governments globally as to the most efective plan of attack for future pandemics.