COVID19 - Final

N. Vastine

2024-04-27

Introduction

COVID19 has left a lasting mark long beyond the Pandemic. The severity of the pandemic for different regions depended significantly on state/country policy for lockdown requirements and access to vaccines.

This report seeks to answer the following questions by analyzing the case and death results of COVID of various states and countries prior to widespread vaccination:

- Which 5 US states and 5 countries had the highest COVID deaths per thousand?
- Which 5 US states and 5 countries had the lowest COVID deaths per thousand?
- How does deaths per thousand correlate to cases per thousand for the 5 highest and 5 lowest states? How does that compare to the 5 highest states and countries?

Identifying states with less devastating COVID results allows policymakers to adopt proven strategies given another global pandemic. While this report does not investigate the governance policies directly, it provides a framework to evaluate these results en masse and can be extended to isolate the effects of specific policies.

The Data

The data is sourced from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University", linked below:

https://github.com/CSSEGISandData/COVID-19

The report will isolate reporting to December 31, 2020 totals to represent cases and deaths prior to mass vaccination.

We will also use a World Population table sourced from the World Bank Dataset to perform case and death per thousand analysis for each country.

https://github.com/datasets/population/tree/main

The following code imports cases and deaths for the US and globally and the population data.

```
library(tidyverse)
library(lubridate)

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

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])

global_population <- read_csv("https://raw.githubusercontent.com/datasets/population/main/data/populati
```

US State Data

The below code pivots the data for reporting and further manipulation to aggregate the data to each state. The code also calculate the rates of cases and deaths per thousand.

```
US_cases_tidy <- US_cases %>%
  pivot_longer(cols = -c(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select (Admin2:cases) %>% # Select a range of headers
  mutate (date = mdy(date)) %>%
  filter(date == "2020-12-31") %>%
  select(-c(Lat,Long_))

US_deaths_tidy <- US_deaths %>%
  pivot_longer(cols = -c(UID:Population),
               names_to = "date",
               values_to = "deaths") %>%
  select (Admin2:deaths) %>%
  mutate (date = mdy(date), year = year(date)) %>%
  filter(date == "2020-12-31") %>%
  select(-c(Lat,Long_))

US_tidy <- US_cases_tidy %>%
  full_join(US_deaths_tidy)

US_by_state <- US_tidy %>%
  group_by(Province_State) %>%
  summarize(cases=sum(cases),deaths=sum(deaths),
            Population=sum(Population)) %>%
  bind_rows(summarise(., across(where(is.numeric), sum),
                          across(where(is.character), ~'US Total'))) %>%
  select(Province_State, cases, deaths, Population)

US_by_state <- US_by_state %>%
  mutate(deaths_per_thou = deaths * 1000 / Population,
         cases_per_thou = cases * 1000 / Population,
         deaths_per_100_cases = deaths * 100/ cases) %>%
  filter(Population > 500000)
```

Global Data

The following code performs similar analysis but with global data, and takes the extra step to join the population table into the tidy global data.

```
global_cases_tidy <- global_cases %>%
  pivot_longer(cols = -c('Province/State',
                         'Country/Region',Lat,Long),
               names_to = "date",
               values_to = "cases") %>%
  mutate (date = mdy(date)) %>%
  filter(date == "2020-12-31") %>%
  select(-c(Lat,Long))


global_deaths_tidy <- global_deaths %>%
  pivot_longer(cols = -c('Province/State',
                         'Country/Region',Lat,Long),
               names_to = "date",
               values_to = "deaths") %>%
  mutate (date = mdy(date)) %>%
  filter(date == "2020-12-31") %>%
  select(-c(Lat,Long))

global_tidy <- global_cases_tidy %>%
  full_join(global_deaths_tidy) %>%
  rename(Country_Region = `Country/Region`,
         Province_State = `Province/State`)

global_population_tidy <- global_population %>%
  filter (Year == 2020) %>%
  select(`Country Name`,Value) %>%
  rename(Country_Region = `Country Name`,
         Population = Value)

global_by_country <- global_tidy %>%
  group_by(Country_Region) %>%
  summarize(cases=sum(cases),deaths=sum(deaths)) %>%
  select(Country_Region, cases, deaths)

global_by_country <- full_join(global_by_country,
                               global_population_tidy) %>%
  filter(Population > 1000,
         cases > 0,
         deaths > 0)

global_by_country <- global_by_country %>%
  mutate(deaths_per_thou = deaths * 1000 / Population,
         cases_per_thou = cases * 1000 / Population,
         deaths_per_100_cases = deaths * 100/ cases)
```

Top 5 / Bottom 5 by Deaths per Thousand

We will further isolate the top 5 and bottom 5 based on deaths per thousand for both US states and globally.

```
US_top_5 <- US_by_state %>%
  slice_min(deaths_per_thou, n = 5) %>%
  select(Province_State,deaths_per_thou,cases_per_thou,
         everything()) %>%
  arrange(deaths_per_thou)

US_bottom_5 <- US_by_state %>%
  slice_max(deaths_per_thou, n = 5) %>%
  select(Province_State,deaths_per_thou,cases_per_thou,
         everything()) %>%
  arrange(desc(deaths_per_thou))

US_overall <- US_by_state %>%
  filter(Province_State == "US Total")

global_top_5 <- global_by_country %>%
  slice_min(deaths_per_thou, n = 5) %>%
  select(Country_Region,deaths_per_thou,cases_per_thou,
         everything()) %>%
  arrange(deaths_per_thou)

global_bottom_5 <- global_by_country %>%
  slice_max(deaths_per_thou, n = 5) %>%
  select(Country_Region,deaths_per_thou,cases_per_thou,
         everything()) %>%
  arrange(desc(deaths_per_thou))

global_US_overall <- US_overall %>%
  rename(Country_Region = Province_State)
```

Analysis

US State Analysis

The below table isolates the top 5 and bottom 5 for the US. Naturally, the overall US response bisects the top and bottom 5 as the average performance of the entire nation. Hawaii had the lowest deaths per thousand, likely because of its inherent isolation as an island.

Surprisingly population size does not sway the deaths per case in either direction. Of the 10 least populated states, Vermont, Maine, and Alaska were in the top 5 fewest deaths per thousand, while North Dakota, South Dakota, and Rhode Island were in the bottom 5 with the most deaths per thousand.

```
US_top_bottom_5 <- rbind(US_top_5,US_bottom_5)
US_top_bottom_5 <- rbind(US_top_bottom_5,US_overall)%>%
  arrange(deaths_per_thou)

head(US_top_bottom_5,11)
```

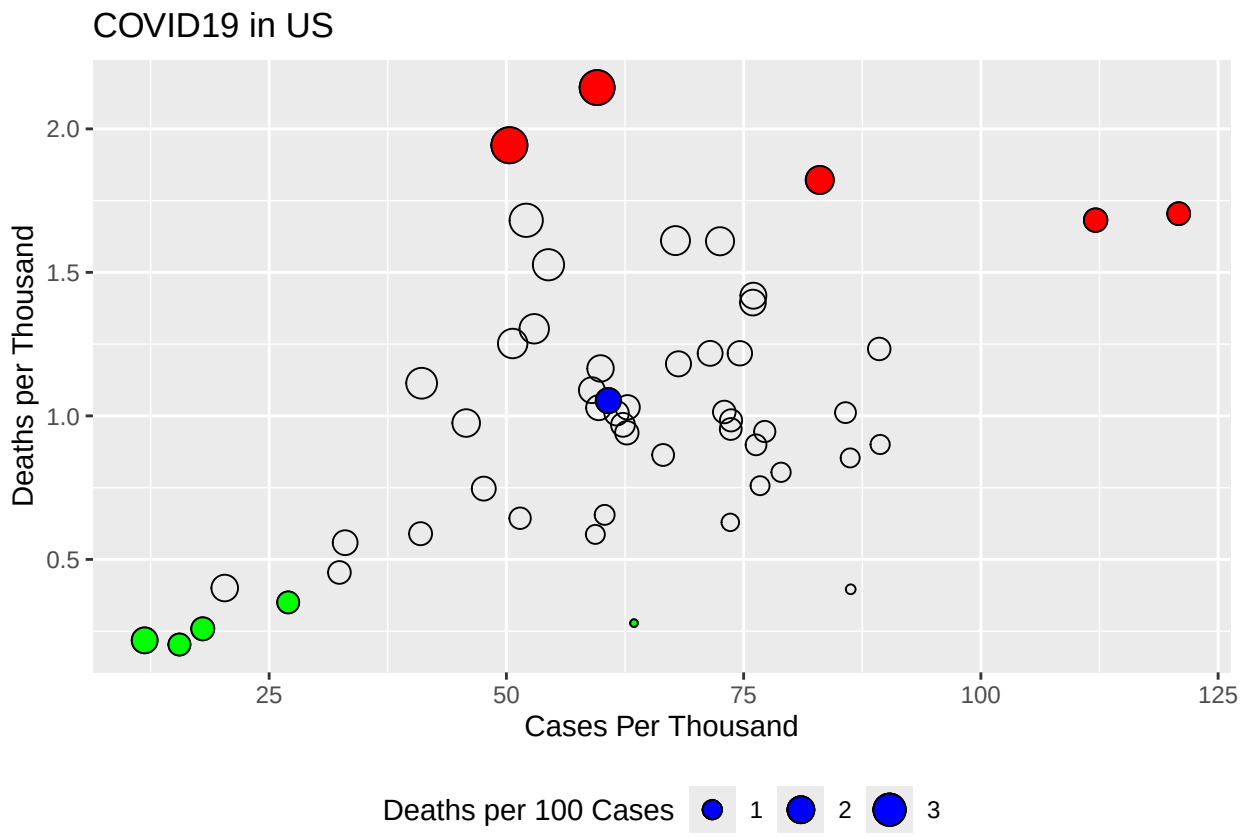
```
## # A tibble: 11 x 7
```

```
##    Province_State deaths_per_thou cases_per_thou    cases deaths Population
##    <chr>                    <dbl>          <dbl>    <dbl>  <dbl>      <dbl>
##  1 Hawaii                   0.203           15.5    22007    288    1415872
##  2 Vermont                  0.218           11.9     7412    136     623989
##  3 Maine                    0.258           18.0    24201    347    1344212
##  4 Alaska                   0.278           63.4    47014    206     740995
##  5 Oregon                   0.350           27.0   113909   1477    4217737
##  6 US Total                 1.05            60.7 20219881 350604  332875137
##  7 South Dakota             1.68           112.     99164   1488     884659
##  8 North Dakota             1.70           121.     92091   1299     762062
##  9 Rhode Island             1.82            83.0    87949   1930    1059361
## 10 New York                 1.94            50.3   978783  37799   19453561
## 11 New Jersey               2.14            59.6   529027  19042    8882190
## # i 1 more variable: deaths_per_100_cases <dbl>
```

Below visualizes cases and deaths per thousand for each state through Dec. 31 2020. The top and bottom 5 states by deaths per thousand are colored green and red respectively, and the overall US response is represented in blue. We can also see the effects of deaths per 100 cases by the size of each point, indicating if there might be limited healthcare or a more susceptible population (age, demographics, wealth, etc).

The graph shows how New York and New Jersey (the red points to the farthest left) may lost more lives from a more susceptible population, confirmed by their higher deaths per 100 cases. Considering both New York and New Jersey are much denser social centers, there may have been higher transmission to more vulnerable populations.

```
ggplot() +
  geom_point(data = US_by_state,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),shape = 21) +
  geom_point(data = US_top_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "green", shape = 21) +
  geom_point(data = US_bottom_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "red", shape = 21) +
  geom_point(data = US_overall,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "blue", shape = 21 ) +
  theme(legend.position = "bottom") +
  labs(title = "COVID19 in US",
       x = "Cases Per Thousand",
       y = "Deaths per Thousand",
       size = "Deaths per 100 Cases")
```



Global Analysis

We can perform a similar analysis on countries, with the top 5 and bottom 5 listed compared to the total US response. Note like New York and New Jersey, Peru had a dramatically high rate of death per 100 cases of 9.16. Peru has had the most deaths per thousand despite having relatively low cases per thousand, which is made apparent in the next graphic.

```
global_top_bottom_5 <- rbind(global_top_5,global_bottom_5)
global_top_bottom_5 <- rbind(global_top_bottom_5,global_US_overall)%>%
  arrange(deaths_per_thou)

head(global_top_bottom_5, n = 11)
```

```
## # A tibble: 11 x 7
##    Country_Region  deaths_per_thou cases_per_thou    cases deaths Population
##    <chr>                     <dbl>          <dbl>    <dbl>  <dbl>      <dbl>
##  1 Burundi                0.000159         0.0648      818      2   12617036
##  2 Mongolia               0.000301         0.359      1195      1    3327204
##  3 Tanzania               0.000344         0.00835     509     21   60972798
##  4 Thailand               0.000851         0.0961     6884     61   71641484
##  5 Eritrea                0.000912         0.401      1320      3    3291271
##  6 US Total               1.05            60.7     20219881 350604  332875137
##  7 North Macedonia        1.35            44.9        83329   2503    1856124
##  8 United Kingdom         1.42            37.2      2496187  95107   67081234
##  9 Belgium                1.69            56.0       646496  19528   11538604
```

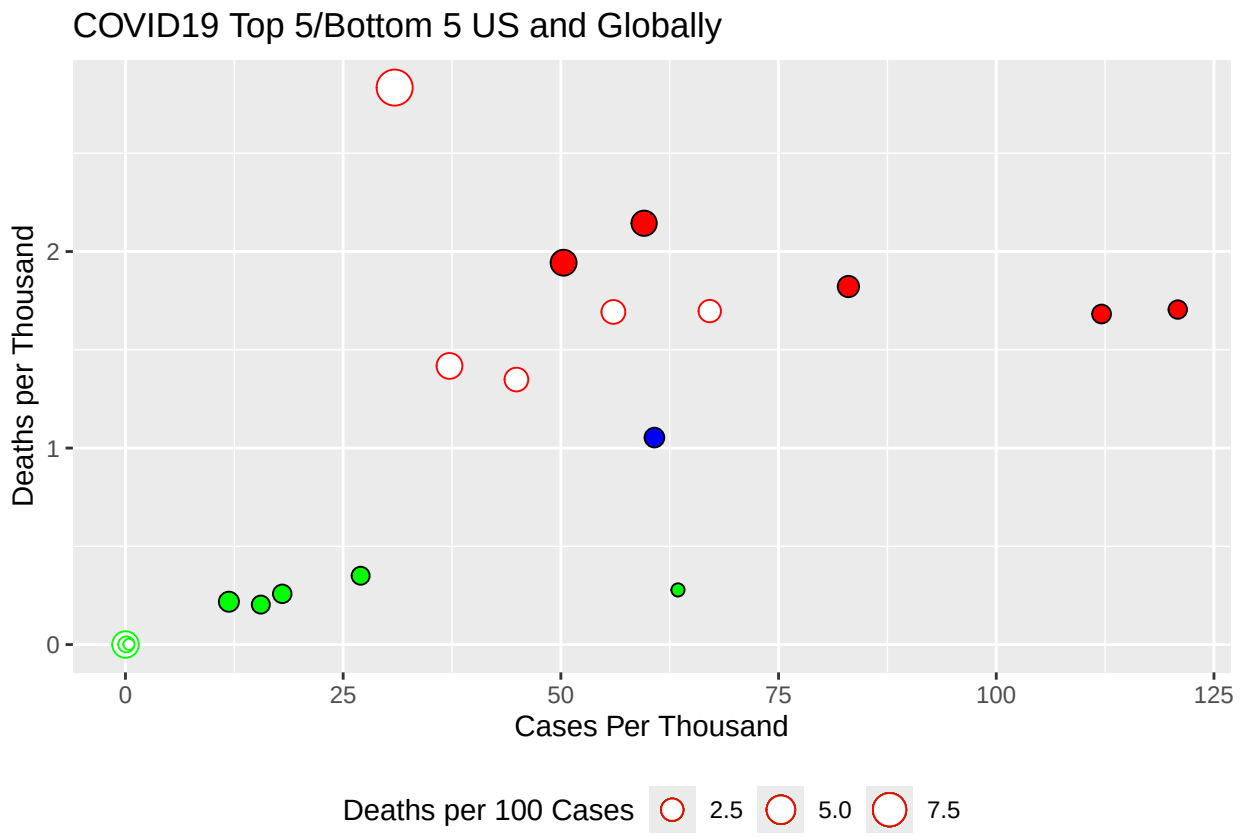
```
## 10 San Marino             1.70           67.1         2333     59      34770
## 11 Peru                   2.83           30.9      1015137  93070   32838579
## # i 1 more variable: deaths_per_100_cases <dbl>
```

US/Global Comparison

Below plots the top 5 and bottom 5 of both the US States and countries for comparison, with the US in blue.

Note the US states are filled with the color, while the countries are outlined.

```
ggplot() +
  geom_point(data = US_top_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "green", shape = 21) +
  geom_point(data = US_bottom_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "red", shape = 21) +
  geom_point(data = US_overall,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "blue", shape = 21) +
  geom_point(data = global_top_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             color = "green", shape = 21,fill = "white") +
  geom_point(data = global_bottom_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             color = "red", shape = 21, fill = "white") +
  theme(legend.position = "bottom") +
  labs(title = "COVID19 Top 5/Bottom 5 US and Globally",
       x = "Cases Per Thousand",
       y = "Deaths per Thousand",
       size = "Deaths per 100 Cases")
```



Modeling

We can develop linear models for the top 5 and bottom 5 of each category, which is calculated below. Note the data point 0 cases per thousand and 0 deaths per thousand are added to each dataset as logically there should be 0 deaths if there are 0 cases.

The below plot adds the models as lines to our plot, with solid lines for the US states and dashed lines indicating countries models.

Based on the plots alone, we can see the aforementioned points for New York and New Jersey performed worse than the model based on the bottom 5 countries. We also see limited value from the top 5 countries model since the data points are very clustered.

The US as expected falls somewhere between the extremes, though closer to the bottom 5 fit than the top 5 fit in both cases.

```
US_top_5[nrow(US_top_5) + 1,] = list("Zero",0,0,0,0,0,0)
US_bottom_5[nrow(US_bottom_5) + 1,] = list("Zero",0,0,0,0,0,0)
global_top_5[nrow(global_top_5) + 1,] = list("Zero",0,0,0,0,0,0)
global_bottom_5[nrow(global_bottom_5) + 1,] = list("Zero",0,0,0,0,0,0)

US_top_5_model <- lm(deaths_per_thou ~ cases_per_thou,
            data = US_top_5)
US_bottom_5_model <- lm(deaths_per_thou ~ cases_per_thou,
            data = US_bottom_5)
global_top_5_model <- lm(deaths_per_thou ~ cases_per_thou,
```

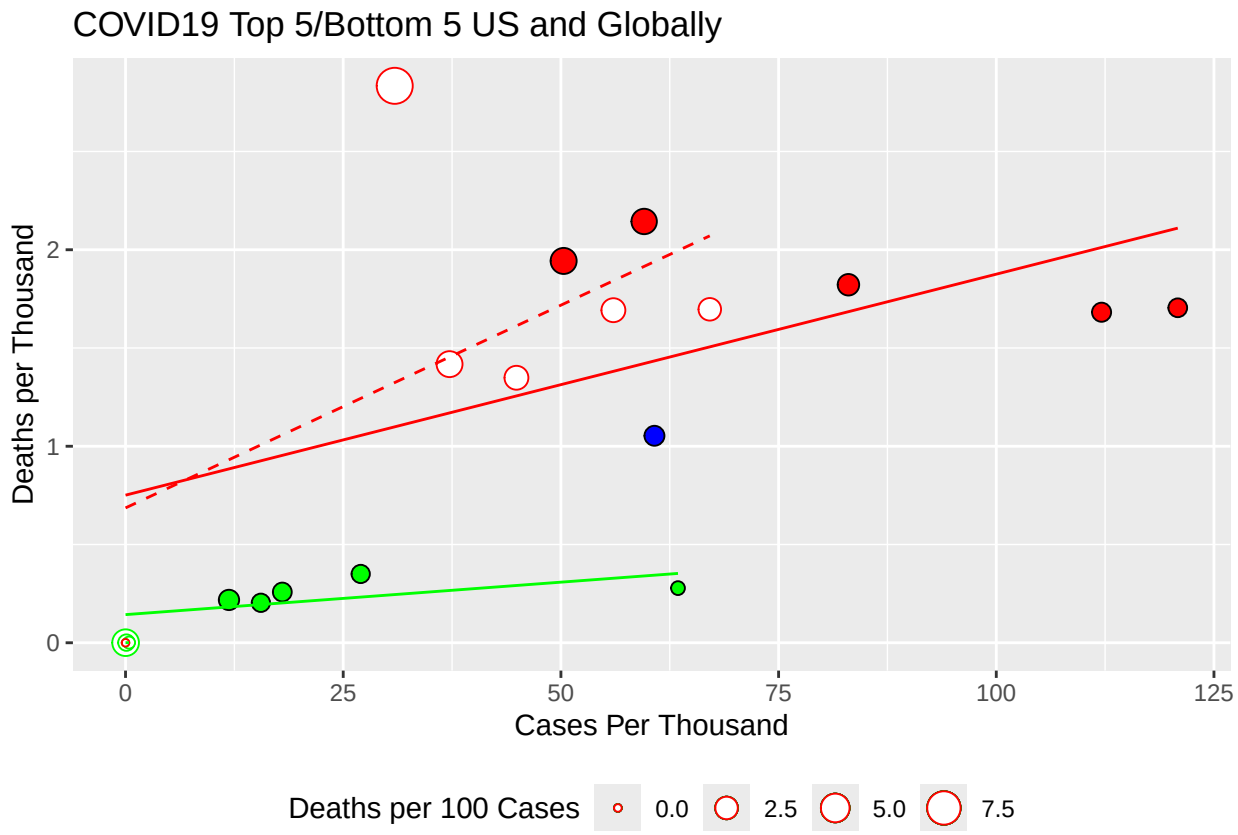
```
            data = global_top_5)
global_bottom_5_model <- lm(deaths_per_thou ~ cases_per_thou,
            data = global_bottom_5)

# US_top_5[nrow(US_bottom_5),] = NULL
# US_bottom_5[nrow(US_bottom_5),] = NULL
# US_top_5[nrow(US_bottom_5),] = NULL
# US_bottom_5[nrow(US_bottom_5),] = NULL


US_top_5_pred <- US_top_5 %>%
  mutate(pred = predict(US_top_5_model))
US_bottom_5_pred <- US_bottom_5 %>%
  mutate(pred = predict(US_bottom_5_model))
global_top_5_pred <- global_top_5 %>%
  mutate(pred = predict(global_top_5_model))
global_bottom_5_pred <- global_bottom_5 %>%
  mutate(pred = predict(global_bottom_5_model))

ggplot() +
  geom_point(data = US_top_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "green", shape = 21) +
  geom_point(data = US_bottom_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "red", shape = 21) +
  geom_point(data = US_overall,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "blue", shape = 21) +
  geom_point(data = global_top_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             color = "green", shape = 21,fill = "white") +
  geom_point(data = global_bottom_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             color = "red", shape = 21, fill = "white") +
  geom_line(data = US_top_5_pred,
            aes(x = cases_per_thou,
                y = pred),
            color = "green") +
  geom_line(data = US_bottom_5_pred,
            aes(x = cases_per_thou,
                y = pred),
            color = "red") +
```

```
  geom_line(data = global_top_5_pred,
            aes(x = cases_per_thou,
                y = pred),
            color = "green",linetype=2) +
  geom_line(data = global_bottom_5_pred,
            aes(x = cases_per_thou,
                y = pred,),
            color = "red",linetype=2) +
  theme(legend.position = "bottom") +
  labs(title = "COVID19 Top 5/Bottom 5 US and Globally",
       x = "Cases Per Thousand",
       y = "Deaths per Thousand",
       size = "Deaths per 100 Cases")
```



Uncertainty & Bias

There was a common storyline during the pandemic that most people did not self report, or cases could not be accurately identified without proper testing. This uncertainty implies the tabulated cases are underreported, which may vary dramaticall across different countries as well. For example, during the Pandemic there were suspicions China was intentionally underrepresenting its figures. Thus, there is a mix of reporting uncertainty or intentional bias build into these reports.

Another storyline from COVID was deaths were inflated because any death was reported as a COVID death even if COVID was not the direct cause. This would inflate the corresponding deaths for a given number of cases in our model.

The uncertainty in both of these variables combined provide a conservative estimate of the corresponding deaths per cases, which in my opinion is safer than underestimating the effects of the Pandemic.

Conclusions

This analysis found the top 5 and bottom 5 states based on the number of deaths. At a higher level, Hawaii had the fewest deaths per thousand (0.2) while New Jersey had the most (2.1). These fell within the global deaths per thousand extremes of the countries of Burundi (0.0002) and Peru (2.9)

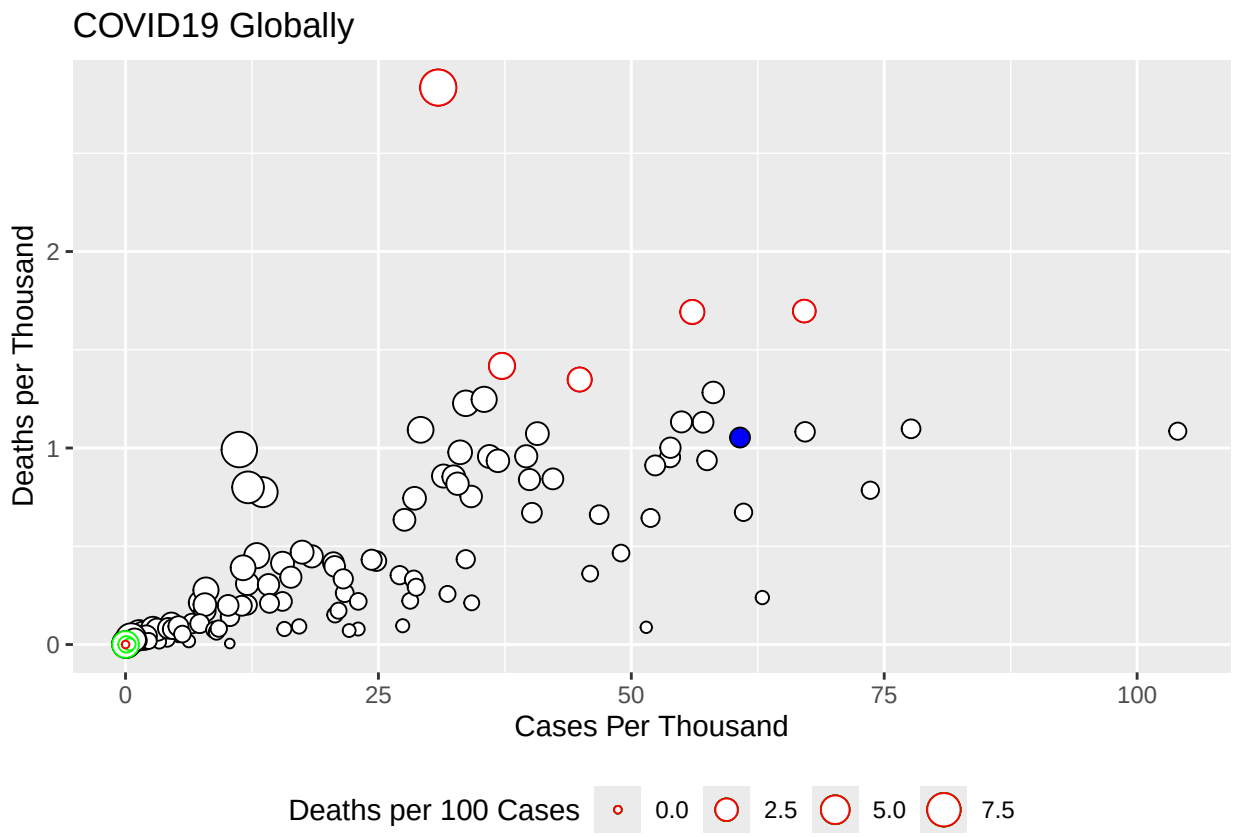
While fitting the data to linear models, it was found the bottom 5 states performed better than the bottom 5 countries, though New York and New Jersey performed worse individually than the bottom 5 countries' model. This appears to be due to the high rate of death per hundred cases, potentially because of higher population density exposing susceptible populations to the virus.

Further analysis could explore the effects of population density and government policies from the identified states to inform best practices for future pandemic responses.

Appendix

Below is a comparison across all countries with the top 5 outlined in green, the bottom five outlined in red, and the US colored blue, indicating the US performing about on average for deaths per cases though with a high number of cases per thousand people.

```
ggplot() +
  geom_point(data = global_by_country,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),shape = 21, fill = "white") +
    geom_point(data = US_overall,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             fill = "blue", shape = 21 ) +
    geom_point(data = global_top_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             color = "green", shape = 21,fill = "white") +
    geom_point(data = global_bottom_5,
             aes(x = cases_per_thou,
                 y = deaths_per_thou,
                 size = deaths_per_100_cases),
             color = "red", shape = 21, fill = "white") +
  theme(legend.position = "bottom") +
  labs(title = "COVID19 Globally",
       x = "Cases Per Thousand",
       y = "Deaths per Thousand",
       size = "Deaths per 100 Cases")
```


COVID19 Globally
Deaths per Thousand
2
1
0
0
25
50
75
100
Cases Per Thousand
Deaths per 100 Cases
0.0
2.5
5.0
7.5