# COVID-19 Deaths Based on Population Size

Nicole McCarthy

April 2, 2025

## Introduction

Tasked with completing some analysis on the Johns Hopkins COVID-19 datasets, I became curious about whether the size of a population affected the ratio of deaths to cases. In other words, was there a discernible trend connecting the number of people a government oversaw and the frequency with which cases became so intense that they resulted in death?I looked at the county, state, and country levels in order to see how changes in sovereignty and population scale affected my investigation.

### Data Sources

The Johns Hopkins github on the pandemic provided everything but the country population data. They began collecting data on January 22, 2020 and stopped adding data on March 9, 2023.

Country Populations were downloaded from the World Bank website.

```
Global_Population <- read_csv("WorldBank_Country_Pop_Data.csv")
Global_Population <- Global_Population[, c("Country Name", "2020")]
colnames(Global_Population)[2] <- "Population" # rename 2020
# get name column to match other data frames
colnames(Global_Population)[1] <- "Country_Region"
```

### Cleaning

The majority of this section was spent getting the data sets into plottable formats that included total cases, total deaths, population, and death:case ratio (calculated from the totals columns).
The US_Cases data frame was organized by county, but the Province_State column was preserved for later use.

```
US_Cases <- US_Cases %>%
  filter(Admin2 != "Unassigned") %>% # get rid of unassigned deaths/cases
  # get rid of deaths/cases occurring out of state
  filter(Admin2 != str_detect(Admin2, "Out of")) %>%
  # get rid of UID, iso2, iso3, code3, FIPS, Admin2, Country_Region, Lat, and Long_
  select(-c(1:6, 8:10))

head(US_Cases)
```

```
## # A tibble: 6 x 1,145
##   Province_State Combined_Key  `1/22/20` `1/23/20` `1/24/20` `1/25/20` `1/26/20`
##   <chr>          <chr>             <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Alabama        Autauga, Ala~         0         0         0         0         0
## 2 Alabama        Baldwin, Ala~         0         0         0         0         0
## 3 Alabama        Barbour, Ala~         0         0         0         0         0
## 4 Alabama        Bibb, Alabam~         0         0         0         0         0
## 5 Alabama        Blount, Alab~         0         0         0         0         0
## 6 Alabama        Bullock, Ala~         0         0         0         0         0
## # i 1,138 more variables: `1/27/20` <dbl>, `1/28/20` <dbl>, `1/29/20` <dbl>,
## #   `1/30/20` <dbl>, `1/31/20` <dbl>, `2/1/20` <dbl>, `2/2/20` <dbl>,
## #   `2/3/20` <dbl>, `2/4/20` <dbl>, `2/5/20` <dbl>, `2/6/20` <dbl>,
## #   `2/7/20` <dbl>, `2/8/20` <dbl>, `2/9/20` <dbl>, `2/10/20` <dbl>,
## #   `2/11/20` <dbl>, `2/12/20` <dbl>, `2/13/20` <dbl>, `2/14/20` <dbl>,
## #   `2/15/20` <dbl>, `2/16/20` <dbl>, `2/17/20` <dbl>, `2/18/20` <dbl>,
## #   `2/19/20` <dbl>, `2/20/20` <dbl>, `2/21/20` <dbl>, `2/22/20` <dbl>, ...
```

This data frame and US_Deaths were manipulated to produce the desired columns described above in a new data frame called County_Sums.

```
# clean up US_Deaths
US_Deaths <- US_Deaths %>%
  filter(Admin2 != "Unassigned") %>% # get rid of unassigned deaths/cases
  filter(Admin2 != str_detect(Admin2, "Out of")) %>%
  # get rid of UID, iso2, iso3, code3, FIPS, Admin2,
  # Province_State, Country_Region, Lat, and Long_
  select(-c(1:10))

# make data frame of totals of cases and deaths by county
US_Deaths_Sum <- US_Deaths %>%
  # only include Combined_Key, Population, and last recorded date (final tally)
  select(1, 2, ncol(US_Deaths))
colnames(US_Deaths_Sum)[3] <- "Total_Deaths" # rename last column

US_Cases_Sum <- US_Cases %>%
  select(Combined_Key, Province_State, ncol(US_Cases))
colnames(US_Cases_Sum)[3] <- "Total_Cases"

US_Sums <- merge(US_Deaths_Sum, US_Cases_Sum)

County_Sums <- US_Sums %>%
  mutate(Death_Case_Ratio = Total_Deaths/Total_Cases) %>%
  # get rid of data where deaths are larger than cases (incorrect)
  filter(Death_Case_Ratio <= 1) %>%
  select(1, 4, 2, 3, 5, 6) # rearrange columns

head(County_Sums)
```

```
##                    Combined_Key Province_State Population Total_Deaths
## 1 Abbeville, South Carolina, US South Carolina      24527           78
## 2         Acadia, Louisiana, US      Louisiana      62045          311
## 3        Accomack, Virginia, US       Virginia      32316          119
## 4                Ada, Idaho, US          Idaho     481587         1139
```

```
## 5               Adair, Iowa, US           Iowa       7152           52
## 6           Adair, Kentucky, US       Kentucky      19202          115
##   Total_Cases Death_Case_Ratio
## 1        7826      0.009966777
## 2       18944      0.016416807
## 3        9119      0.013049676
## 4      160373      0.007102193
## 5        1805      0.028808864
## 6        7849      0.014651548
```

Using the parent data frame for County_Sums, US_Sums, a data frame for state information was aggregated.

```
State_Sums <- aggregate(US_Sums[, c("Population", "Total_Deaths", "Total_Cases")],
                        by = list(State = US_Sums$Province_State),
                        FUN = sum)
State_Sums <- mutate(State_Sums, Death_Case_Ratio = Total_Deaths/Total_Cases)
head(State_Sums)
```

```
##        State Population Total_Deaths Total_Cases Death_Case_Ratio
## 1    Alabama    4903185        21032     1644533      0.012789041
## 2     Alaska     740995         1486      307649      0.004830180
## 3    Arizona    7278717        33102     2443514      0.013546884
## 4   Arkansas    3017804        13020      973278      0.013377473
## 5 California   39512223       101159    12125315      0.008342794
## 6   Colorado    5758736        14156     1764140      0.008024306
```

Finally, a smiliar process was completed on Global_Cases and Global_Deaths to make an aggregate data frame titled Global_Sums. This was necessary because the original version provided a column for each Province/State of the coutnries included. However, I was only interested in the country-level data from this frame.

```
# clean up data frames
Global_Cases <- select(Global_Cases, "Country/Region", "3/9/23")
colnames(Global_Cases)[2] <- "Total_Cases"

Global_Deaths <- select(Global_Deaths, "Country/Region", "3/9/23")
colnames(Global_Deaths)[2] <- "Total_Deaths"

# gather sums of all countries
Global_Cases <- aggregate(Global_Cases[, "Total_Cases"],
                          by = list(Country_Region = Global_Cases$`Country/Region`),
                          FUN = sum)

Global_Deaths <- aggregate(Global_Deaths[, "Total_Deaths"],
                          by = list(Country_Region = Global_Deaths$`Country/Region`),
                          FUN = sum)

# aggregate relevant data sets
Global_Sums <- merge(Global_Cases, Global_Deaths)
Global_Sums <- merge(Global_Sums, Global_Population)

# calculate death:case column
Global_Sums <- mutate(Global_Sums, Death_Case_Ratio = Total_Deaths/Total_Cases)
```

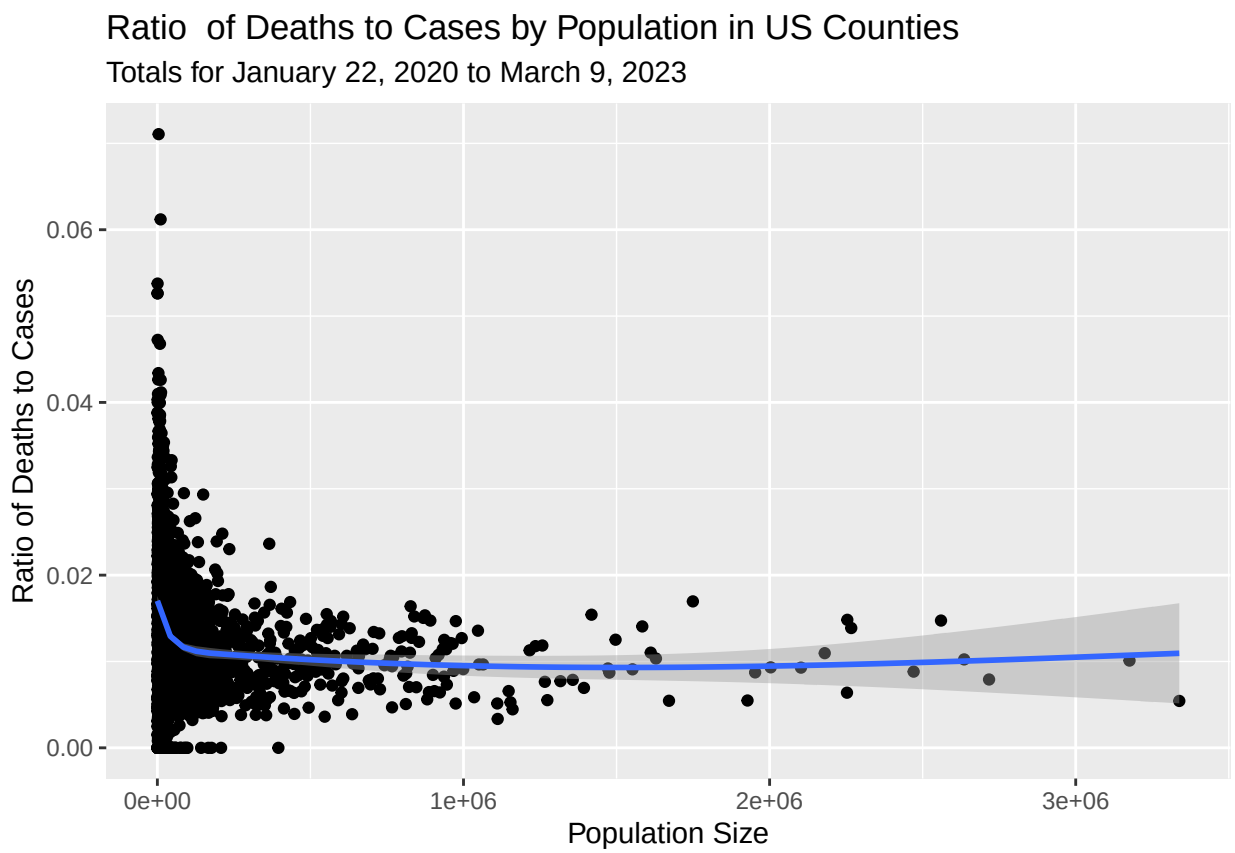
```
head(Global_Sums)
```

```
##        Country_Region Total_Cases Total_Deaths Population Death_Case_Ratio
## 1         Afghanistan      209451         7896   39068979      0.037698555
## 2             Albania      334457         3598    2837849      0.010757736
## 3             Algeria      271496         6881   44042091      0.025344756
## 4             Andorra       47890          165      77380      0.003445396
## 5              Angola      105288         1933   33451132      0.018359167
## 6 Antigua and Barbuda        9106          146      91846      0.016033385
```

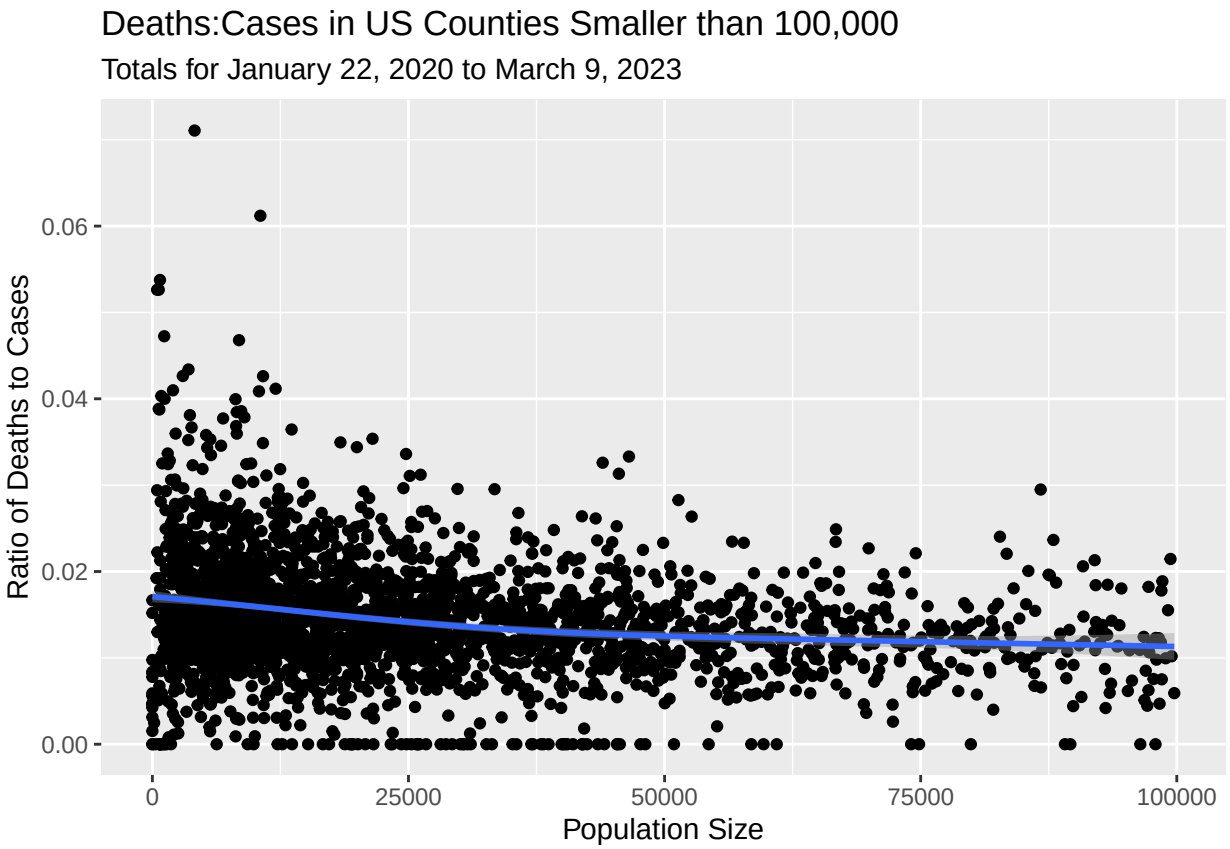
# Analysis

The plots of each level include a bit of filtering in order to eliminate outliers in population that would affect the visualizations and subsequent analysis. In each case, only a few points were eliminated.

## County Level



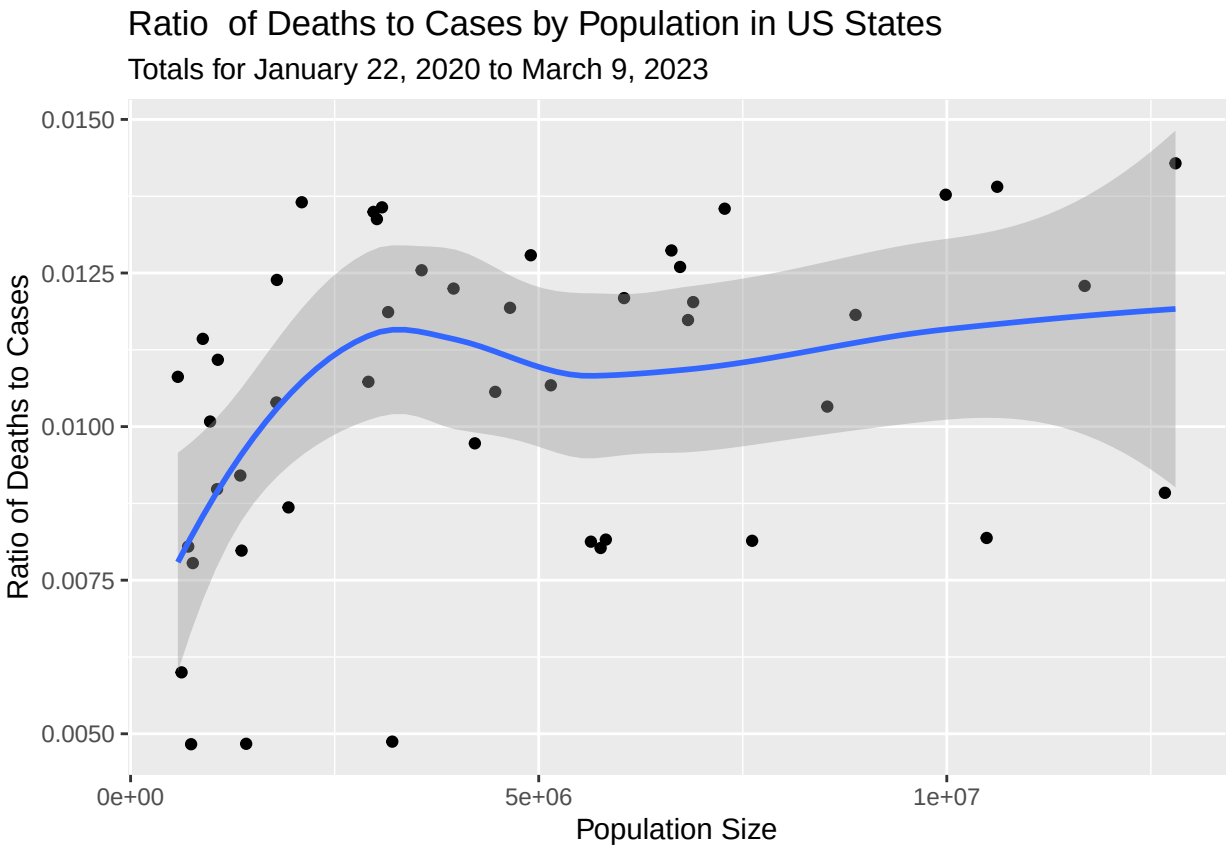

Counties larger than 4,000,000 people were not included as they were outliers that made the rest of the plot difficult to read. The counties not shown include Cook, IL, Harris, TX, Los Angeles, CA, and Maricopa, AZ. In this plot, there is a slight downward trend in among counties with populations less than 100,000. This quickly evens out. To get a better look at this these smaller counties, we'll filter our data even futher.



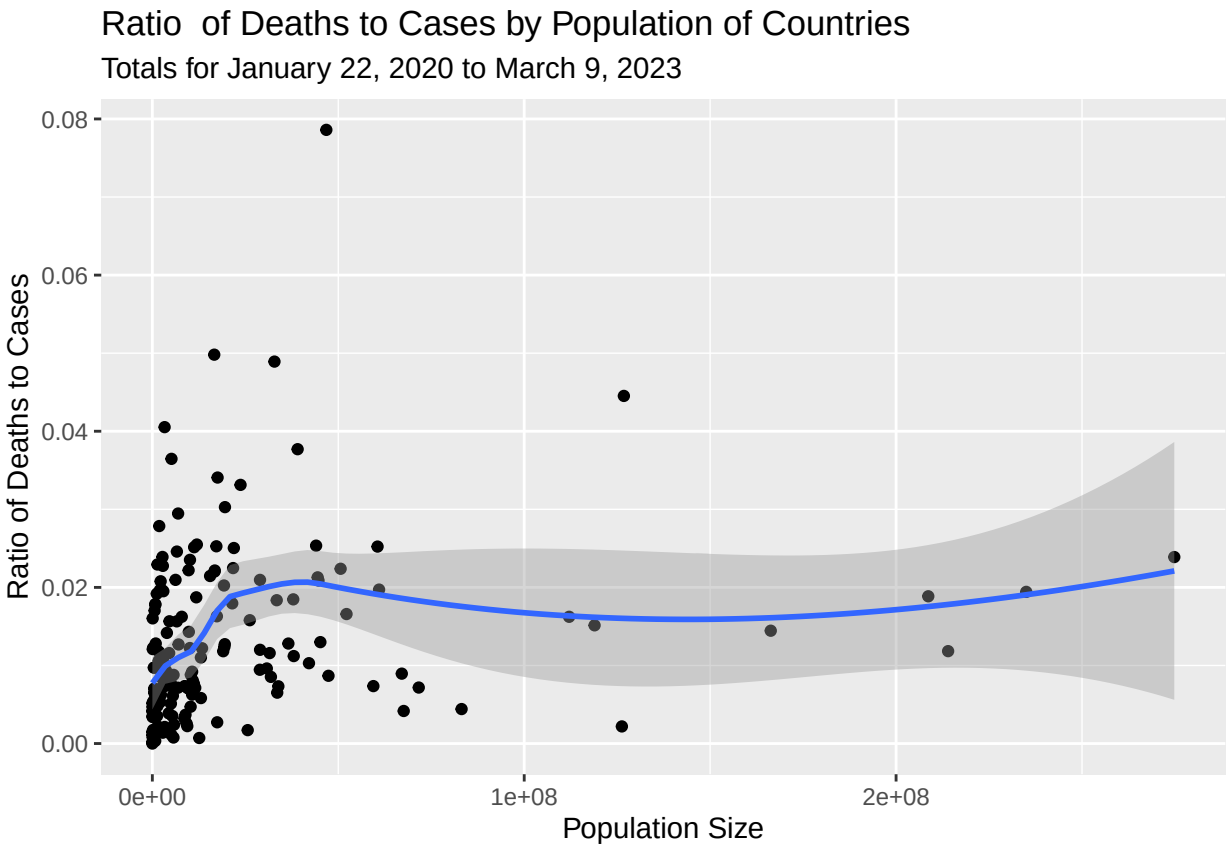

The reason for this downward trend is unclear. However, it is likely that more rural communities, ones that are more likely to have smaller populations, have higher death rates than those in urban areas. It is impossible to say from this data alone whether average population age, healthcare resources, or other factors are contributing to this.

## State Level

On the state level, we see a different trend. The larger the population, the higher the death to case ratio was. Again, we cannot say for sure from this data alone, but it seems that states with larger populations potentially had more cases than they had resources to deal with adequately.
Note that states with populations larger than 15 million are not shown. This includes California, Florida, New York, and Texas. Furthermore, Puerto Rico had a death to case ratio less than 0.003 and was not included as an outlier.


Ratio of Deaths to Cases by Population in US States
Totals for January 22, 2020 to March 9, 2023
Ratio of Deaths to Cases
0.0150
0.0125
0.0100
0.0075
0.0050
0e+00
5e+06
1e+07
Population Size

### Global Level




Finally, the global scale matches the trend seen on the state scale, but with a bit of leveling occurring once the population size reached approximately 50 million. It seems that once this threshold is met, the average death to case ratio arises no matter the resources or population.
Note that India and China are not included in this graph as they are population outliers of over 500 million.

## Conclusions

When considering the best course of action in the face of a future pandemic, we can group communities based on the following population sizes: $< 100$K and $< 50$ million. If we consider looking at populations greater than 50 million, there are no trends to help guide our decisions and it is likely we would be overwhelmed by such a larger category anyways. Resources should be divided among communities of less than 50 million such that those with larger populations get higher amounts of supplies. However, when breaking these smaller groups, those with less than 100 thousand residents will likely need more resources than were afforded during the COVID-19 pandemic. Further analysis is necessary to determine specific causes for these trends.