Analyzing Trends in COVID-19 Death-rates in the U.S.

Ethan DeJongh

2024-06-20

```
knitr::opts_chunk$set(echo = TRUE)

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   3.5.0     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(lubridate)
```

Introduction

In this report, I will be examining how the rate of deaths from COVID-19 changed over time in the United States, both in terms of the overall death rate and deaths as compared to the number of reported cases. I am particularly interested in finding out how the trends differ between urban and rural areas.

The data I will use comes from a github repository from Johns Hopkins University. This dataset contains the daily reported numbers of cases and deaths from COVID-19 for every county in the U.S. since the start of the pandemic. It also provides county populations, which I will use to analyze differences between small and large-population counties. The first step is to import the data, in this case from two separate files containing data on U.S. cases and U.S. deaths:

```
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")

urls <- str_c(url_in,file_names)

US_cases <- read_csv(urls[1])
```

```
## Rows: 3342 Columns: 1154
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20...
##
## 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.
```

```
US_deaths <- read_csv(urls[2])
```

```
## Rows: 3342 Columns: 1155
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## 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.
```

```
#US_cases
```

The next step is tidy the data. First I pivot the data for cases and deaths so that each row represents a particular date for a particular county, whilst adjusting the date format and removing the latitude and longitude columns. Next, I join the two dataframes into a single one with both cases and deaths. Finally, I add new columns for the daily increase in cases and deaths, and filter out rows which have none.

```
US_cases <- US_cases %>%
    pivot_longer(cols = -c(UID:Combined_Key),
                         names_to = "date",
                         values_to = "cases") %>%
    select(Admin2:cases) %>%
    mutate(date = mdy(date)) %>%
    select(-c(Lat, Long_))

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

US <- US_cases %>% full_join(US_deaths) %>%
  mutate(new_cases = cases - lag(cases, default = 0), new_deaths = deaths - lag(deaths, default = 0)) %
  filter(new_cases > 0 | new_deaths > 0) %>%
  filter(Population > 0)
```

```
## Joining with `by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)`
```

```
US
```

```
## # A tibble: 1,908,069 x 10
##    Admin2 Province_State Country_Region Combined_Key date       cases Population
```

```
##    <chr>  <chr>          <chr>          <chr>        <date>     <dbl>      <dbl>
##  1 Autau~ Alabama        US             Autauga, Al~ 2020-03-24     1      55869
##  2 Autau~ Alabama        US             Autauga, Al~ 2020-03-25     5      55869
##  3 Autau~ Alabama        US             Autauga, Al~ 2020-03-26     6      55869
##  4 Autau~ Alabama        US             Autauga, Al~ 2020-03-30     8      55869
##  5 Autau~ Alabama        US             Autauga, Al~ 2020-04-01    10      55869
##  6 Autau~ Alabama        US             Autauga, Al~ 2020-04-02    12      55869
##  7 Autau~ Alabama        US             Autauga, Al~ 2020-04-07    12      55869
##  8 Autau~ Alabama        US             Autauga, Al~ 2020-04-09    17      55869
##  9 Autau~ Alabama        US             Autauga, Al~ 2020-04-10    18      55869
## 10 Autau~ Alabama        US             Autauga, Al~ 2020-04-11    19      55869
## # i 1,908,059 more rows
## # i 3 more variables: deaths <dbl>, new_cases <dbl>, new_deaths <dbl>
```

Since I am interested in broad trends over the course of the pandemic, and daily numbers can be erratic, I will group the data by month instead of day. The new dataframe contains monthly sums of new cases and deaths for each county.

```
Counties_by_month <- US %>%
  group_by(Combined_Key, month = lubridate::floor_date(date, 'month')) %>%
  summarize(cases = sum(new_cases), deaths = sum(new_deaths), population = max(Population))
```

```
## `summarise()` has grouped output by 'Combined_Key'. You can override using the
## `.groups` argument.
```

```
Counties_by_month
```

```
## # A tibble: 115,154 x 5
## # Groups:   Combined_Key [3,211]
##    Combined_Key                  month      cases deaths population
##    <chr>                         <date>     <dbl>  <dbl>      <dbl>
##  1 Abbeville, South Carolina, US 2020-03-01     5      0      24527
##  2 Abbeville, South Carolina, US 2020-04-01    28      0      24527
##  3 Abbeville, South Carolina, US 2020-05-01    12      0      24527
##  4 Abbeville, South Carolina, US 2020-06-01    74      0      24527
##  5 Abbeville, South Carolina, US 2020-07-01   173      7      24527
##  6 Abbeville, South Carolina, US 2020-08-01   142      3      24527
##  7 Abbeville, South Carolina, US 2020-09-01   175      3      24527
##  8 Abbeville, South Carolina, US 2020-10-01   210      4      24527
##  9 Abbeville, South Carolina, US 2020-11-01   167      5      24527
## 10 Abbeville, South Carolina, US 2020-12-01   308      3      24527
## # i 115,144 more rows
```

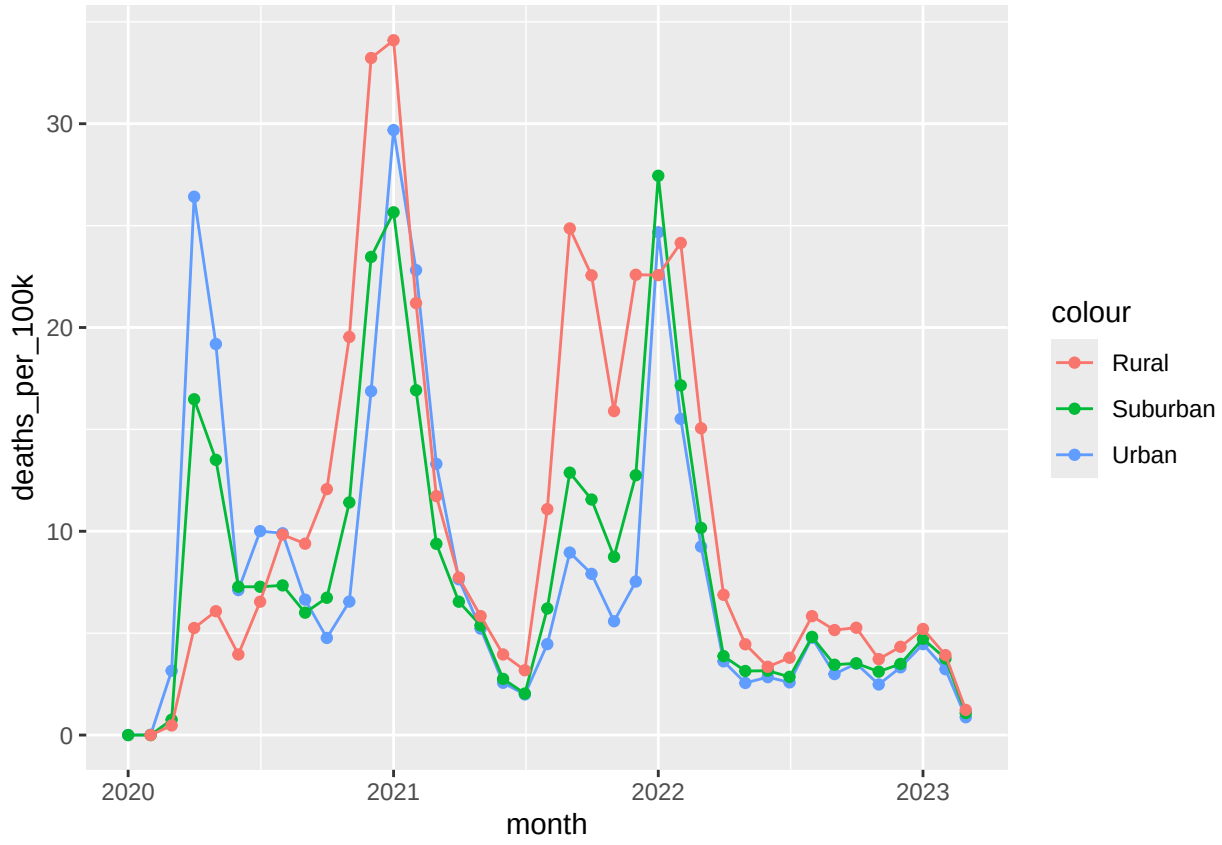
Rather than look at individual counties or states, I decided to combine the county data into 3 broad groups: urban - counties with population above 1 million, suburban - population between 100,000 and 1 million, and rural - population below 100,000. The dataframes below contain the monthly totals for these groups. I have also calculated the deaths per 100,000 population and deaths per 100 cases in separate columns.

```
Urban_totals <- Counties_by_month %>% filter(population > 1000000) %>%
  group_by(month) %>% summarize(cases = sum(cases), deaths = sum(deaths), population = sum(population))
  mutate(deaths_per_100k = 100000*deaths/population, deaths_per_100case = 100*deaths/cases)
```

```
Suburban_totals <- Counties_by_month %>% filter(population > 100000 & population <= 1000000) %>%
  group_by(month) %>% summarize(cases = sum(cases), deaths = sum(deaths), population = sum(population))
  mutate(deaths_per_100k = 100000*deaths/population, deaths_per_100case = 100*deaths/cases)

Rural_totals <- Counties_by_month %>% filter(population <= 100000) %>%
  group_by(month) %>% summarize(cases = sum(cases), deaths = sum(deaths), population = sum(population))
  mutate(deaths_per_100k = 100000*deaths/population, deaths_per_100case = 100*deaths/cases)
```

My first plot will compare the deaths per 100,000 people for urban, suburban, and rural counties over time:

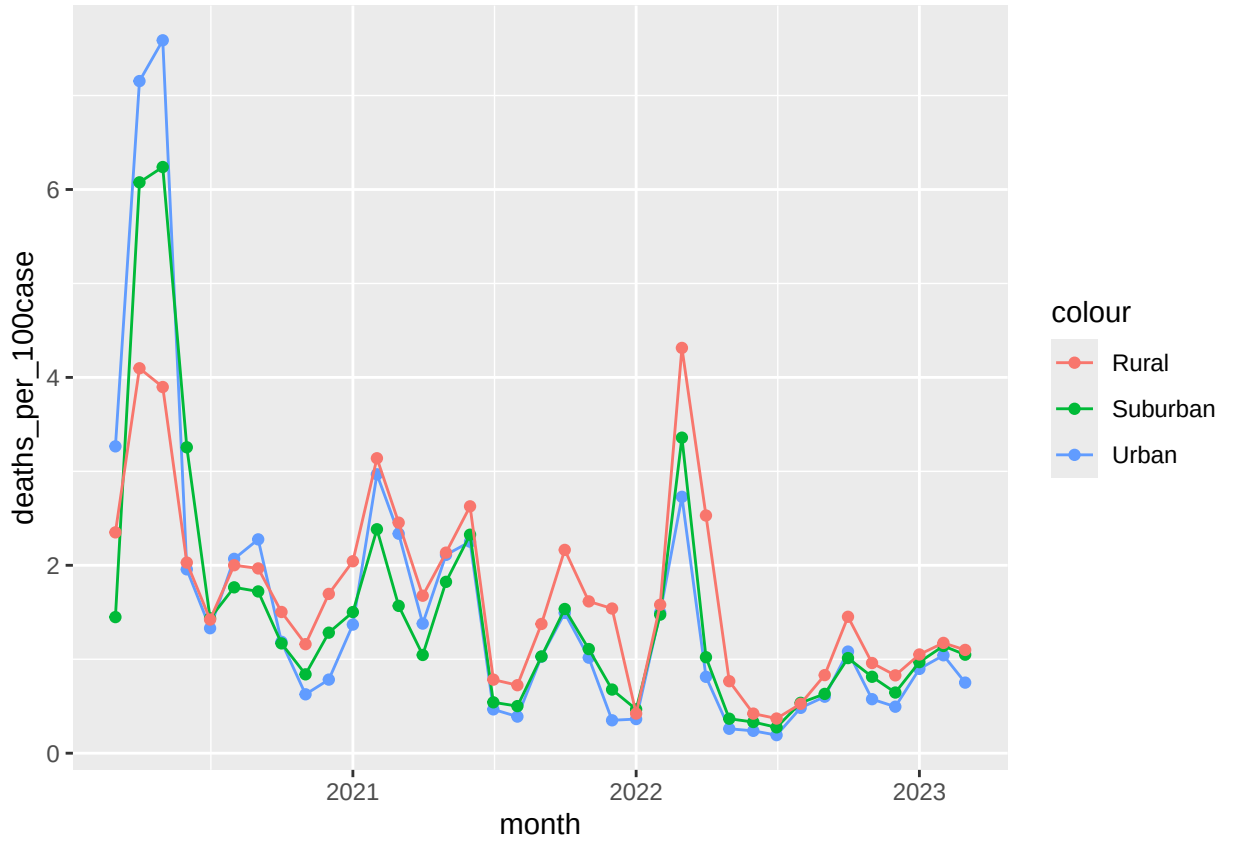
```
ggplot() +
  geom_point(data=Urban_totals, aes(month, deaths_per_100k, color='Urban')) +
  geom_line(data=Urban_totals, aes(month, deaths_per_100k, color='Urban')) +
  geom_point(data=Suburban_totals, aes(month, deaths_per_100k, color='Suburban')) +
  geom_line(data=Suburban_totals, aes(month, deaths_per_100k, color='Suburban')) +
  geom_point(data=Rural_totals, aes(month, deaths_per_100k, color='Rural')) +
  geom_line(data=Rural_totals, aes(month, deaths_per_100k, color='Rural'))
```



The results of this plot are pretty striking. We see that for the first half of 2020, death rates were far higher in urban and suburban counties compared to rural counties. By late 2020, rural death rates surged above urban and suburban. It appears that the pandemic hit large population areas hardest in the beginning, and then it gradually spread to small population areas. Death rates declined steeply across all areas in the first half of 2021, but then surged again in the second half of the year, this time affecting rural areas the most. Deaths declined again early in 2022, but then remained noticably higher for rural areas than other areas.

The next figure plots deaths per 100 cases for the three groups:

```
ggplot() +
  geom_point(data=Urban_totals %>% filter(deaths > 10), aes(month, deaths_per_100case, color='Urban')) +
  geom_line(data=Urban_totals %>% filter(deaths > 10), aes(month, deaths_per_100case, color='Urban')) +
  geom_point(data=Suburban_totals %>% filter(deaths > 10), aes(month, deaths_per_100case, color='Suburba
  geom_line(data=Suburban_totals %>% filter(deaths > 10), aes(month, deaths_per_100case, color='Suburban
  geom_point(data=Rural_totals %>% filter(deaths > 10), aes(month, deaths_per_100case, color='Rural')) +
  geom_line(data=Rural_totals %>% filter(deaths > 10), aes(month, deaths_per_100case, color='Rural'))
```



This plot suggests that the higher death rates in rural areas were not just a result of higher numbers of cases. Except for the first few months, the rate of deaths proportional to cases was consistently higher in rural areas.

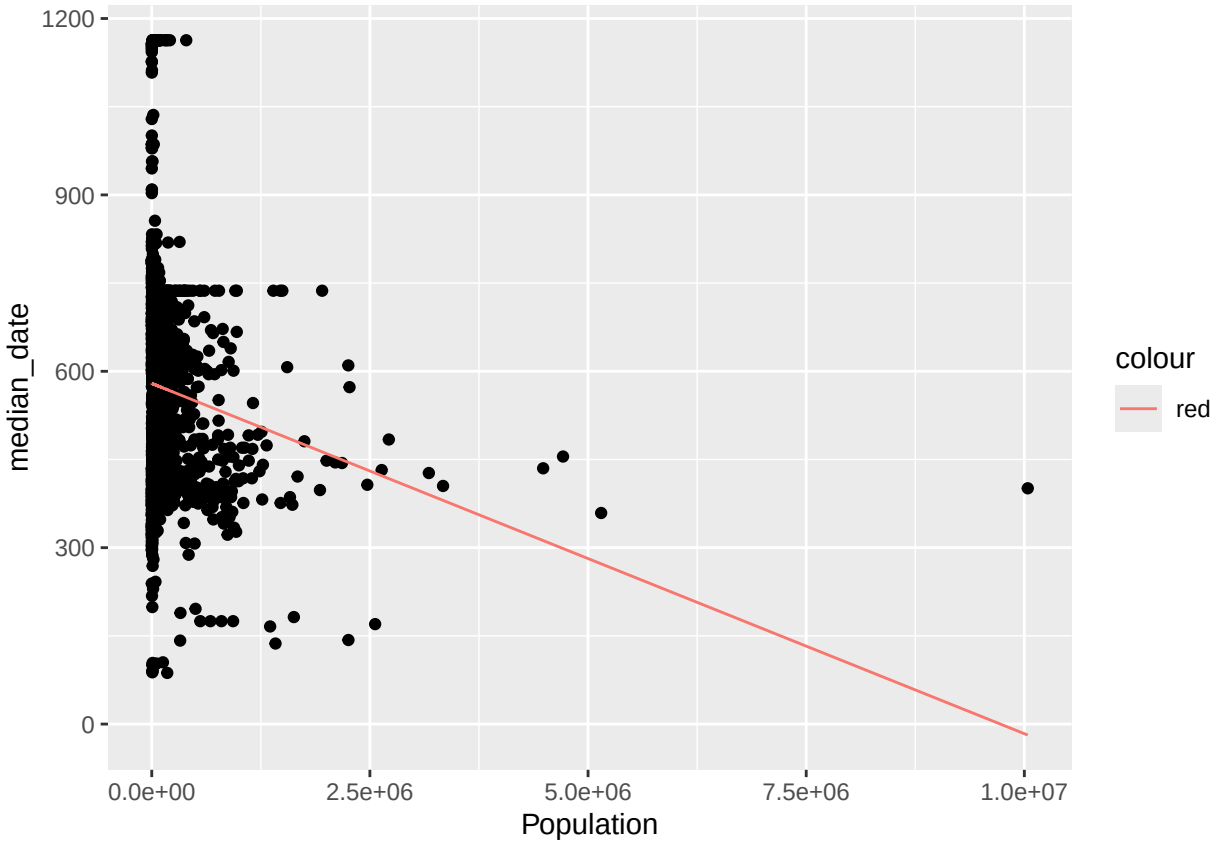
I am interested in seeing how these trends vary for individual counties. In the code below, I calculate the median death date for each county - the date at which the cumulative death total reached half of the amount at the end of the dataset - and plot it against the county population. A linear model is shown as the red line, revealing a negative correlation, but there is high variance. The y-axis is shown in days since 01-01-2020.

```
Median_death_date <- US %>% group_by(Combined_Key) %>% mutate(max_deaths = max(deaths)) %>%
  filter(deaths <= 0.5*max_deaths) %>%
  group_by(Combined_Key) %>% summarize(date = max(date), Population = max(Population)) %>%
  mutate(median_date = as.numeric(date) - 18262)

summary(Median_death_date)
```

```
##  Combined_Key            date              Population        median_date
##  Length:3211        Min.   :2020-03-28   Min.   :      86   Min.   :  87.0
```

```
##  Class :character   1st Qu.:2021-03-05   1st Qu.:   11162   1st Qu.: 429.0
##  Mode  :character   Median :2021-08-26   Median :   26277   Median : 603.0
##                     Mean   :2021-07-27   Mean   :  103526   Mean   : 573.2
##                     3rd Qu.:2021-10-22   3rd Qu.:   67647   3rd Qu.: 660.0
##                     Max.   :2023-03-09   Max.   :10039107   Max.   :1163.0
```

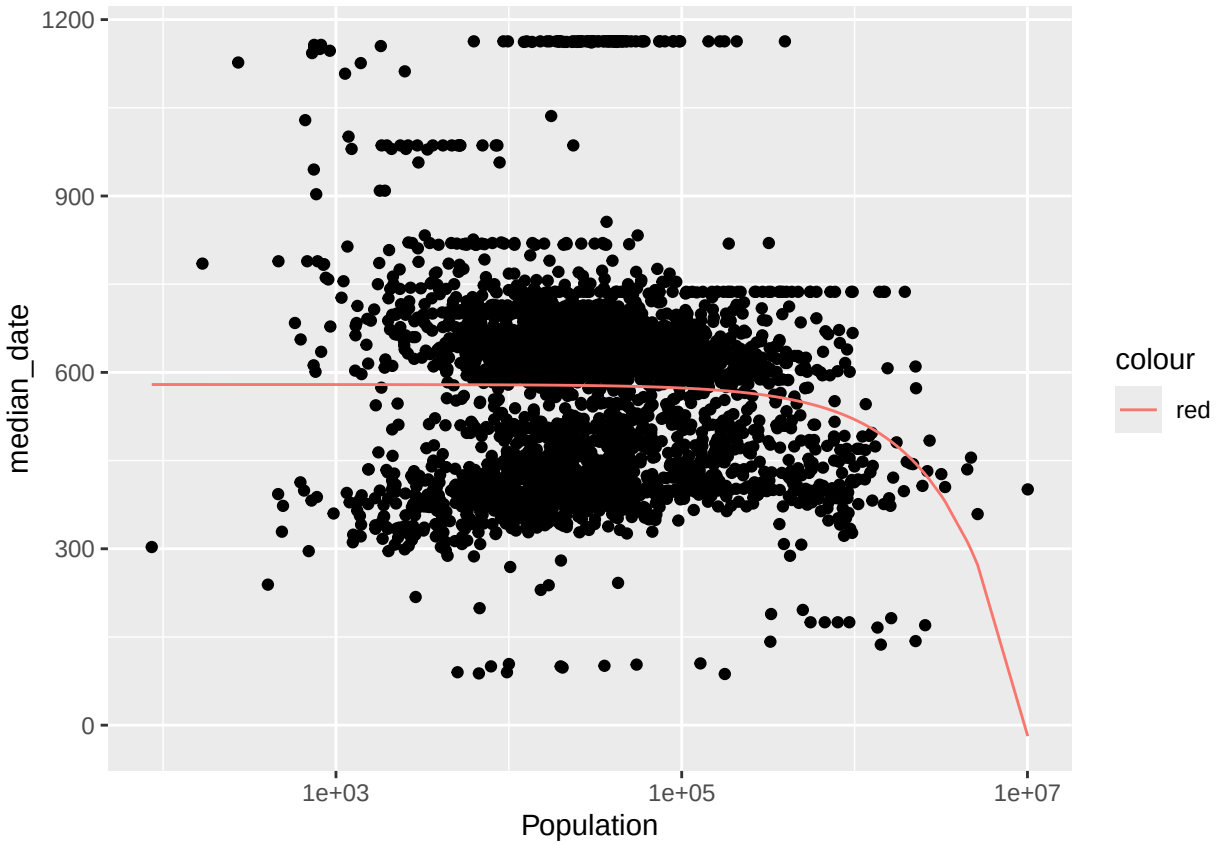
```
Median_death_date$new_date <- as.numeric(Median_death_date$date)
mod <- lm(median_date ~ Population, data = Median_death_date)

Median_death_date <- Median_death_date %>% mutate(pred = predict(mod))

ggplot() +
  geom_point(data=Median_death_date, aes(Population, median_date))  +
  geom_line(data=Median_death_date, aes(Population, pred, color='red'))
```



The plot is better visualized using a logarithmic scale for the population:

```
ggplot() +
  geom_point(data=Median_death_date, aes(Population, median_date)) + scale_x_log10() +
  geom_line(data=Median_death_date, aes(Population, pred, color='red')) + scale_x_log10()
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```



While there is a lot of variance, it is notable that all of the 10 largest population counties had their median death occur earlier than the average.

Conclusion

This report illustrates differences in how the COVID-19 pandemic affected large and small population counties across the United States. Deaths vs. population and deaths vs. cases were both much higher in urban areas at the start of the pandemic, but increased steadily in rural areas for the first year. It makes sense that the virus spread very quickly in high-population centers at the start, and took longer to spread throughout the rural population. The large variance in median death date among low population counties shows how the virus spread non-uniformly across the country, striking different areas at different times. The reason why death rates remained highest in rural areas for the later stages of the pandemic is unclear. It could be related to lower vaccination rates in rural areas or less reliable access to medical care. There is also a possible source of bias in how cases and deaths are reported. If, for example, deaths are less frequently reported in urban areas, this could partially explain the discrepancy.