CovidProject

A. Coles

2024-04-29

```
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   3.5.1     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)
```

Importing Covid 19 Data

This project uses the Johns Hopkins Covid-19 dataset, which begins on Jan. 22, 2020 and ends March 9, 2023. The data was archived on Mar. 10, 2023 on github.com. The data is provided by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE): https://systems.jhu.edu/ and is described in:

Dong, Du, and Gardner. "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet,* 20.5 (May 2020). DOI: https://doi.org/10.1016/S1473-3099(20)30120-1

Dong, et al. "The Johns Hopkins University Center for Systems Science and Engineering COVID-19 Dashboard: data collection process, challenges faced, and lessons learned," *The Lancet,* 22.12 (Dec. 2022). DOI: https://doi.org/10.1016/S1473-3099(22)00434-0

```
##Import csv files from github URL,
##and concatenate URL with file names into strings

url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov
file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_US.csv",
                "time_series_covid19_deaths_global.csv")
urls <- str_c(url_in, file_names)

##Read in data, making sure name matches data
global_cases <- read_csv(urls[2])
```

```
## 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.
```

```
global_deaths <- read_csv(urls[4])
```

```
## 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.
```

```
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[3])
```

```
## 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.
```

Tidying Global Covid 19 Data

The data was tidied to

```
##Tidy global_cases file
global_cases <- global_cases %>%
    pivot_longer(cols = -c(`Province/State`,
                           `Country/Region`, Lat, Long),
                 names_to = "date",
```

```
                 values_to = "cases") %>%
    select(-c(Lat,Long))

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

##Combine global_cases and global_deaths, rename files to be R friendly
global <- global_cases %>%
    full_join(global_deaths) %>%
    rename(Country_Region = `Country/Region`,
           Province_State = `Province/State`) %>%
    mutate(date = mdy(date))
```

```
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
```

```
##Filter out cases = 0
## To data check with filters: global %>% filter(cases > 103800000))
global <- global %>% filter(cases > 0)

##Create Combined_Key with province, country
global <- global %>%
    unite("Combined_Key",
          c(Province_State, Country_Region),
          sep = ", ",
          na.rm = TRUE,
          remove = FALSE)

##Find file with global population data
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U

uid <- read_csv(uid_lookup_url) %>%
   select(-c(Lat, Long_, Combined_Key,
             code3, iso2, iso3, Admin2))
```

```
## Rows: 4321 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population
##
## 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.
```

```
##Join population information to global
global <- global %>%
    left_join(uid, by = c("Province_State",
                          "Country_Region")) %>%
```

```
    select(-c(UID, FIPS)) %>%
    select(Province_State, Country_Region, date, cases,
           deaths, Population, Combined_Key)
```

Summary of Global Data

```
##Show global data for project file
summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:306827      Length:306827      Min.   :2020-01-22   Min.   :        1
##  Class :character   Class :character   1st Qu.:2020-12-12   1st Qu.:     1316
##  Mode  :character   Mode  :character   Median :2021-09-16   Median :    20365
##                                        Mean   :2021-09-11   Mean   :  1032863
##                                        3rd Qu.:2022-06-15   3rd Qu.:   271281
##                                        Max.   :2023-03-09   Max.   :103802702
##
##      deaths          Population        Combined_Key
##  Min.   :      0   Min.   :6.700e+01   Length:306827
##  1st Qu.:      7   1st Qu.:7.866e+05   Class :character
##  Median :    214   Median :6.948e+06   Mode  :character
##  Mean   :  14405   Mean   :2.890e+07
##  3rd Qu.:   3665   3rd Qu.:2.914e+07
##  Max.   :1123836   Max.   :1.380e+09
##                    NA's   :6729
```

Tidying Covid 19 Data for the US

```
##Tidy US cases file
US_cases%>%
    pivot_longer(cols = -(UID:Combined_Key),
                        names_to = "date",
                        values_to = "cases")
```

```
## # A tibble: 3,819,906 x 13
##         UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
##  1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  2 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  3 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  4 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  5 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  6 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  7 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  8 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
##  9 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 10 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## # i 3,819,896 more rows
## # i 4 more variables: Long_ <dbl>, Combined_Key <chr>, date <chr>, cases <dbl>
```

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

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

##Combine two US files
US <- US_cases %>%
     full_join(US_deaths)
```

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

```
##Filter out US cases less than 0
US <- US %>% filter(cases > 0)
```

Summary of US Data

```
##Show US data for project file
summary(US)
```

```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:3474292     Length:3474292     Length:3474292     Length:3474292
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       date                cases           Population           deaths
##  Min.   :2020-01-22   Min.   :      1   Min.   :       0   Min.   :    0.0
##  1st Qu.:2020-12-27   1st Qu.:    687   1st Qu.:   10953   1st Qu.:   10.0
##  Median :2021-09-20   Median :   2849   Median :   26248   Median :   47.0
##  Mean   :2021-09-19   Mean   :  15489   Mean   :  104502   Mean   :  205.1
##  3rd Qu.:2022-06-15   3rd Qu.:   9345   3rd Qu.:   68098   3rd Qu.:  137.0
##  Max.   :2023-03-09   Max.   :3710586   Max.   :10039107   Max.   :35545.0
```

Analysis and Visualization

Analysis of US Data

```
##Summarize data by state
US_by_state <- US %>%
    group_by(Province_State, Country_Region, date) %>%
    summarize(cases = sum(cases), deaths = sum(deaths),
              Population = sum(Population)) %>%
    mutate(deaths_per_mill = deaths*1000000 / Population) %>%
    select(Province_State, Country_Region, date,
             cases, deaths, deaths_per_mill, Population) %>%
    ungroup()
```

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

```
##Summarize data for US
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) %>%
  select(Country_Region, date, cases, deaths,
         deaths_per_mill, Population) %>%
  ungroup()
```

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

```
##Calculate new cases and new deaths
US_by_state <- US_by_state %>%
     mutate(new_cases = cases - lag(cases),
            new_deaths = deaths - lag(deaths))
US_totals <- US_totals %>%
     mutate(new_cases = cases - lag(cases),
            new_deaths = deaths - lag(deaths))

##Calculate cases and deaths per thousand
US_state_totals <- US_by_state %>%
         group_by(Province_State) %>%
         summarize(deaths = max(deaths), cases = max(cases),
                   population = max(Population),
                   cases_per_thou = 1000*cases / population,
                   deaths_per_thou = 1000*deaths / population) %>%
     filter(cases > 0, population > 0)

##Look at states with least deaths per thousand
US_state_totals %>%
      slice_min(deaths_per_thou, n = 10)
```

```
## # A tibble: 10 x 6
##    Province_State        deaths  cases population cases_per_thou deaths_per_thou
##    <chr>                  <dbl>  <dbl>      <dbl>          <dbl>           <dbl>
##  1 American Samoa            34 8.32e3      55641           150.           0.611
##  2 Northern Mariana Isl~     41 1.37e4      55144           248.           0.744
##  3 Virgin Islands           130 2.48e4     107268           231.           1.21
##  4 Hawaii                  1841 3.81e5    1415872           269.           1.30
##  5 Vermont                  929 1.53e5     623989           245.           1.49
##  6 Puerto Rico             5823 1.10e6    3754939           293.           1.55
##  7 Utah                    5298 1.09e6    2785478           391.           1.90
##  8 District of Columbia    1432 1.78e5     705749           252.           2.03
##  9 Alaska                  1486 3.08e5     728809           422.           2.04
## 10 Washington             15683 1.93e6    7614893           253.           2.06
```

```
##Look at states with most deaths per thousand
US_state_totals %>%
      slice_max(deaths_per_thou, n = 10)
```

```
## # A tibble: 10 x 6
##    Province_State deaths   cases population cases_per_thou deaths_per_thou
##    <chr>           <dbl>   <dbl>      <dbl>          <dbl>           <dbl>
##  1 Arizona         33102 2443514    7278717           336.            4.55
##  2 Oklahoma        17972 1290929    3956971           326.            4.54
##  3 Mississippi     13370  990756    2976149           333.            4.49
##  4 West Virginia    7960  642760    1792147           359.            4.44
##  5 New Mexico       9061  670929    2096829           320.            4.32
##  6 Arkansas        13020 1006883    3017804           334.            4.31
##  7 Alabama         21032 1644533    4903185           335.            4.29
##  8 Tennessee       29263 2515130    6829174           368.            4.28
##  9 Michigan        42205 3064125    9986857           307.            4.23
## 10 Kentucky        18130 1718471    4467673           385.            4.06
```

Plots of Cases v. Deaths

US new cases and new deaths

```
US_totals %>%
     filter(new_cases > 0) %>%
     ggplot(aes(x = date, y = new_cases)) +
     geom_line(aes(color = "new_cases")) +
     geom_point(aes(color = "new_cases")) +
     geom_line(aes(y = new_deaths, color = "new_deaths")) +
     geom_point(aes(y = new_deaths, color = "new_deaths")) +
     scale_y_log10() +
     theme(legend.position="bottom",
           axis.text.x = element_text(angle = 90)) +
     labs(title = "Covid-19 in US", y = NULL)
```

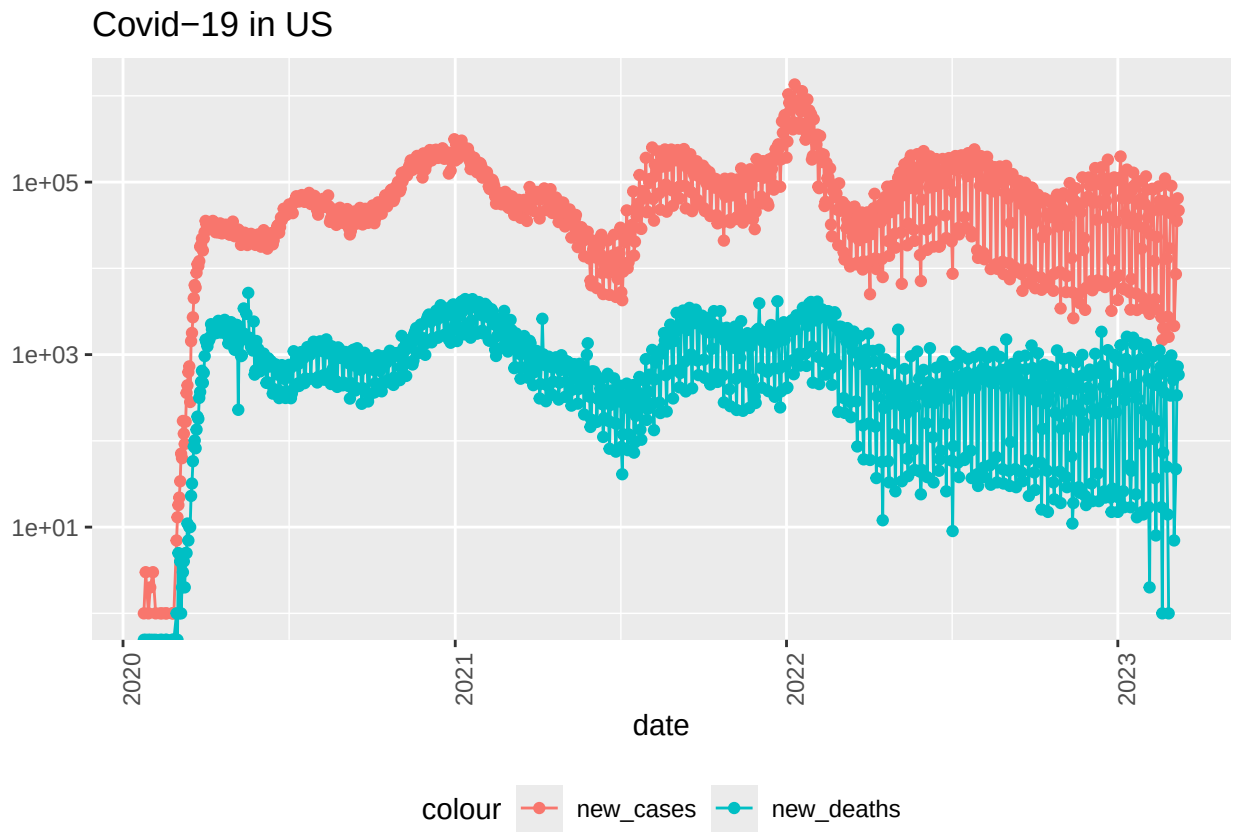
```
## Warning in transformation$transform(x): NaNs produced
```

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

```
## Warning in transformation$transform(x): NaNs produced

## Warning in scale_y_log10(): log-10 transformation introduced infinite values.

## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Illinois new cases and new deaths

```
state <- "Illinois"
 US_by_state %>%
     filter(Province_State == state) %>%
     filter(new_cases > 0) %>%
     ggplot(aes(x = date, y = new_cases)) +
     geom_line(aes(color = "new_cases")) +
     geom_point(aes(color = "new_cases")) +
     geom_line(aes(y = new_deaths, color = "new_deaths")) +
     geom_point(aes(y = new_deaths, color = "new_deaths")) +
     scale_y_log10() +
     theme(legend.position="bottom",
           axis.text.x = element_text(angle = 90)) +
     labs(title = str_c("Covid-19 in ", state), y = NULL)
```

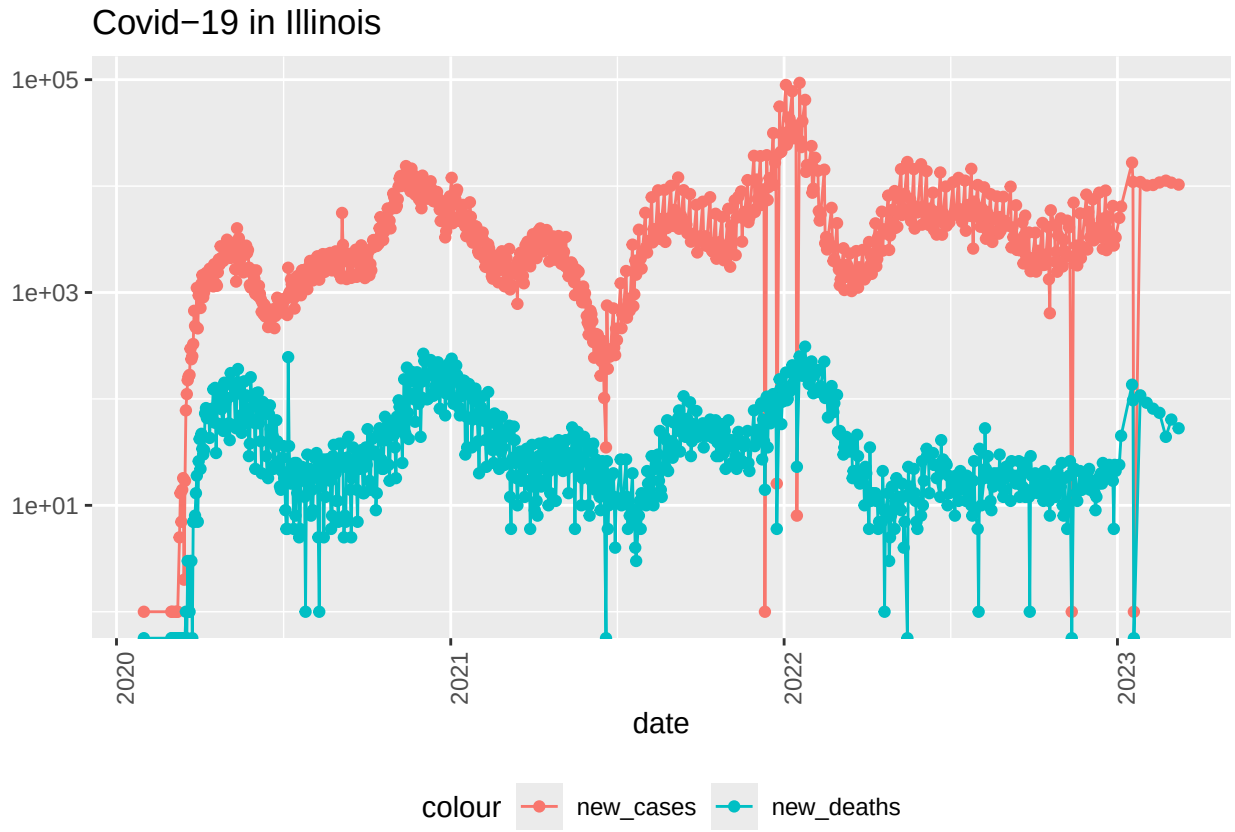
```
## Warning in transformation$transform(x): NaNs produced
```

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.

## Warning in transformation$transform(x): NaNs produced

## Warning in scale_y_log10(): log-10 transformation introduced infinite values.

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



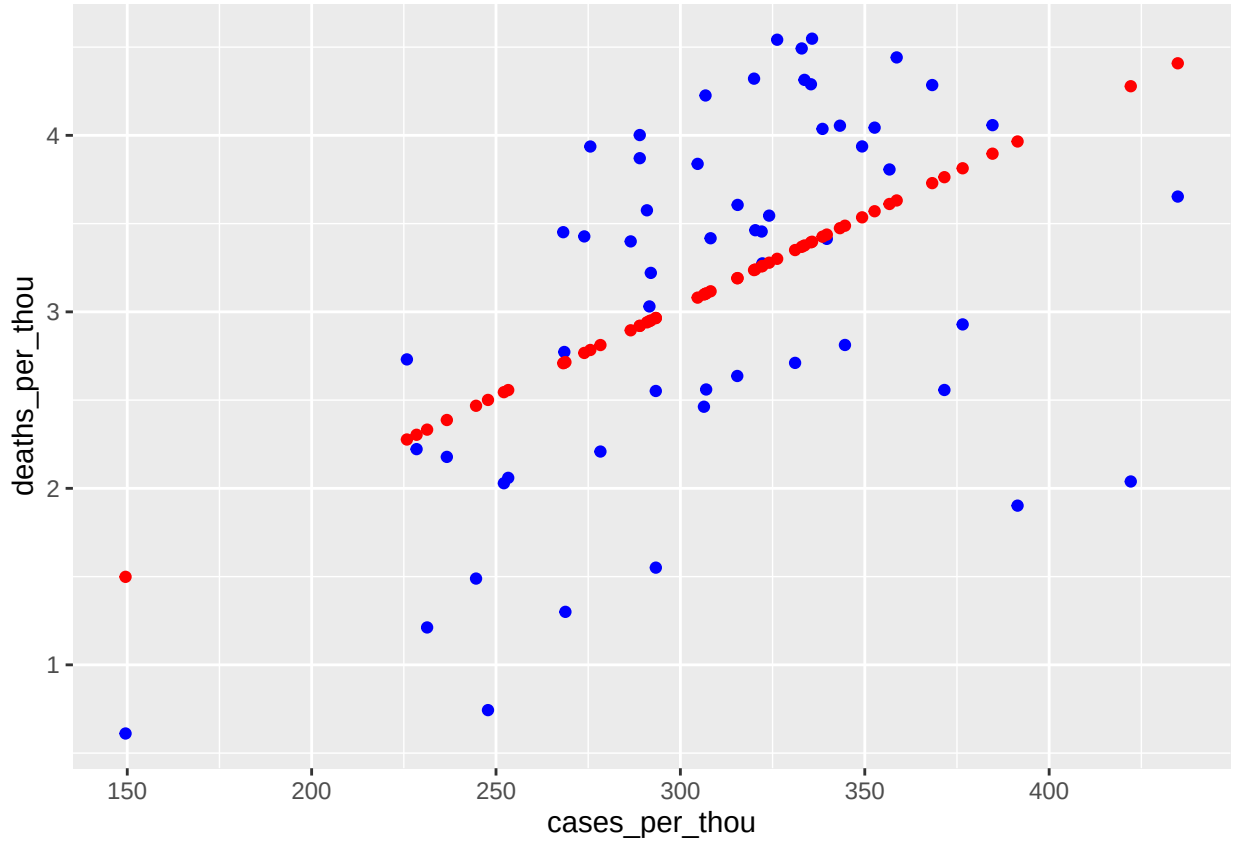
Modelling Data

```
mod <- lm(deaths_per_thou ~ cases_per_thou, data = US_state_totals)
US_state_totals %>% mutate(pred = predict(mod))
```

```
## # A tibble: 56 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          21032 1.64e6    4903185           335.            4.29  3.39
##  2 Alaska            1486 3.08e5     728809           422.            2.04  4.28
##  3 American Samoa      34 8.32e3      55641           150.            0.611 1.50
##  4 Arizona          33102 2.44e6    7278717           336.            4.55  3.40
##  5 Arkansas         13020 1.01e6    3017804           334.            4.31  3.38
```

```
##  6 California      101159 1.21e7   39512223          307.          2.56   3.10
##  7 Colorado         14181 1.76e6    5758736          306.          2.46   3.10
##  8 Connecticut      12220 9.77e5    3565287          274.          3.43   2.77
##  9 Delaware          3324 3.31e5     973764          340.          3.41   3.44
## 10 District of Co~   1432 1.78e5     705749          252.          2.03   2.54
## # i 46 more rows
```

```
US_tot_w_pred <- US_state_totals %>% mutate(pred = predict(mod))
US_tot_w_pred %>% ggplot() +
     geom_point(aes(x = cases_per_thou, y = deaths_per_thou),
                color = "blue") +
     geom_point(aes(x = cases_per_thou, y = pred),
                color = "red")
```



Conclusions:

Overall, the data show that the most Covid deaths per thousand occurred in Arizona, Oklahoma, Mississippi, West Virginia, New Mexico, Arkansas, Alabama, Tennessee, Michigan, and Kentucky. The state legislatures in these states as of 2023 were predominately Republican, with the exception of Michigan and New Mexico (NCSL Map, May 25, 2023). Michigan, however, had a divided government in 2022 (https://www.multistate.us/issues/2022-state-trifectas). Other research also suggested that political messaging around Covid resulted in higher death rates for Republican constituencies in most, but not all cases (https://ncnewsline.com/2022/10/04/study-more-republicans-than-democrats-likely-died-of-covid/) Politics is a definite source of bias, both in analysis and also, perhaps, in reporting numbers of deaths. While Republican states dominated the top 10 list of most Covid deaths, this doesn't mean no Republican governor protected their state well, e.g. in Ohio and North Carolina.