Are COVID19 new daily deaths per thousand affected by a state's political affiliation?

MG

April 13, 2024

Summary

This report analyses the COVID19 Dataset from Center for Systems Science and Engineering (CSSE) at Johns Hopkins University https://github.com/CSSEGISandData/COVID-19.

The dataset contains global COVID19 confirmed cases and deaths, and the data was collected between the beginning of the pandemic in March of 2020 until March 10, 2023.

This report will focus on USA data only, and will explore if a state's political affiliation, as determined by presidential vote outcome in 2020 election (i.e. "red" - republican vs. "blue" - democratic state), has any effect on COVID19 daily deaths per thousand.

Importing and Cleaning the data

```
library(tidyverse)
library(lubridate)
library(readxl)
library(ggplot2)
```

First, we will load the US data and perform some basic cleanup. We will use the CSV files for US confirmed cases and 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])
US_deaths <- read_csv(urls[2])
```

The cases and deaths CSV files contain several columns which we will not be using, for example: "UID", "Lat", "Long", and others, and also use a column for each date, so we will remove the unused columns, and convert the date columns into a single column called "date" and convert their values into a Date object.

```
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_))

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_))
```

Next, we will merge US cases and deaths data into a single dataframe.

```
US <- US_cases %>%
  full_join(US_deaths)
```

Now, let's look at the summary of the combined cases and deaths dataset.

```
summary(US)
```

```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:3819906     Length:3819906     Length:3819906     Length:3819906
##  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.   :  -3073   Min.   :       0   Min.   :  -82.0
##  1st Qu.:2020-11-02   1st Qu.:    330   1st Qu.:    9917   1st Qu.:    4.0
##  Median :2021-08-15   Median :   2272   Median :   24892   Median :   37.0
##  Mean   :2021-08-15   Mean   :  14088   Mean   :   99604   Mean   :  186.9
##  3rd Qu.:2022-05-28   3rd Qu.:   8159   3rd Qu.:   64979   3rd Qu.:  122.0
##  Max.   :2023-03-09   Max.   :3710586   Max.   :10039107   Max.   :35545.0
```

From the summary we can see that Population column has rows with 0 as the value, and there are also rows where cases or deaths are less than 0, so let's remove those rows.

Additionally, since we are using 2020 election to determine political affiliation of the sates, let's use the data starting January 1, 2021.

```
US <- US %>%
  filter(Population > 0, cases >= 0, deaths >= 0, date >= '2021-01-01')

summary(US)
```

```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:2575146     Length:2575146     Length:2575146     Length:2575146
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       date                cases           Population          deaths
##  Min.   :2021-01-01   Min.   :      0   Min.   :      86   Min.   :    0.0
##  1st Qu.:2021-07-19   1st Qu.:   1722   1st Qu.:   11137   1st Qu.:   26.0
##  Median :2022-02-03   Median :   4626   Median :   26205   Median :   71.0
##  Mean   :2022-02-03   Mean   :  20025   Mean   :  103153   Mean   :  253.6
##  3rd Qu.:2022-08-22   3rd Qu.:  12823   3rd Qu.:   67493   3rd Qu.:  179.0
```

```
##  Max.   :2023-03-09   Max.   :3710586   Max.   :10039107   Max.   :35545.0
```

Since we will be analyzing data for each of the US states, we will next create a dataframe with cases and deaths statistics for each state.

```
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population) %>%
  ungroup()
```

Data Exploration and Feature Engineering

Adding deaths per thousand data

The numbers for deaths in the dataset are cumulative, so let's see which states are top 10 by the end of the data collection period.

```
US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), population = max(Population)) %>%
  slice_max(deaths, n = 10)
```

```
## # A tibble: 10 x 3
##    Province_State deaths population
##    <chr>           <dbl>      <dbl>
##  1 California     101159   39512223
##  2 Texas           93355   28995881
##  3 Florida         86454   21477737
##  4 New York        76592   19453561
##  5 Pennsylvania    50398   12801989
##  6 Michigan        41964    9986857
##  7 Ohio            41794   11689100
##  8 Georgia         40833   10617423
##  9 Illinois        36431   12671821
## 10 New Jersey      36015    8882190
```

Sate population varies significantly, and deaths will tend to be higher in states with larger population, so we will add a new variable "deaths_per_thou" to have a better way to compare individual state's numbers.

```
US_by_state <- US_by_state %>%
  mutate(deaths_per_thou = deaths * 1000 / Population)
```

We can now see what are the top 10 states with highest total deaths per thousand people.

```
US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths_per_thou = max(deaths_per_thou), population = max(Population)) %>%
  slice_max(deaths_per_thou, n = 10)
```

```
## # A tibble: 10 x 3
##    Province_State deaths_per_thou population
##    <chr>                    <dbl>      <dbl>
##  1 Arizona                   4.55    7278717
##  2 Mississippi               4.49    2976149
##  3 West Virginia             4.44    1792147
##  4 New Mexico                4.32    2096829
```

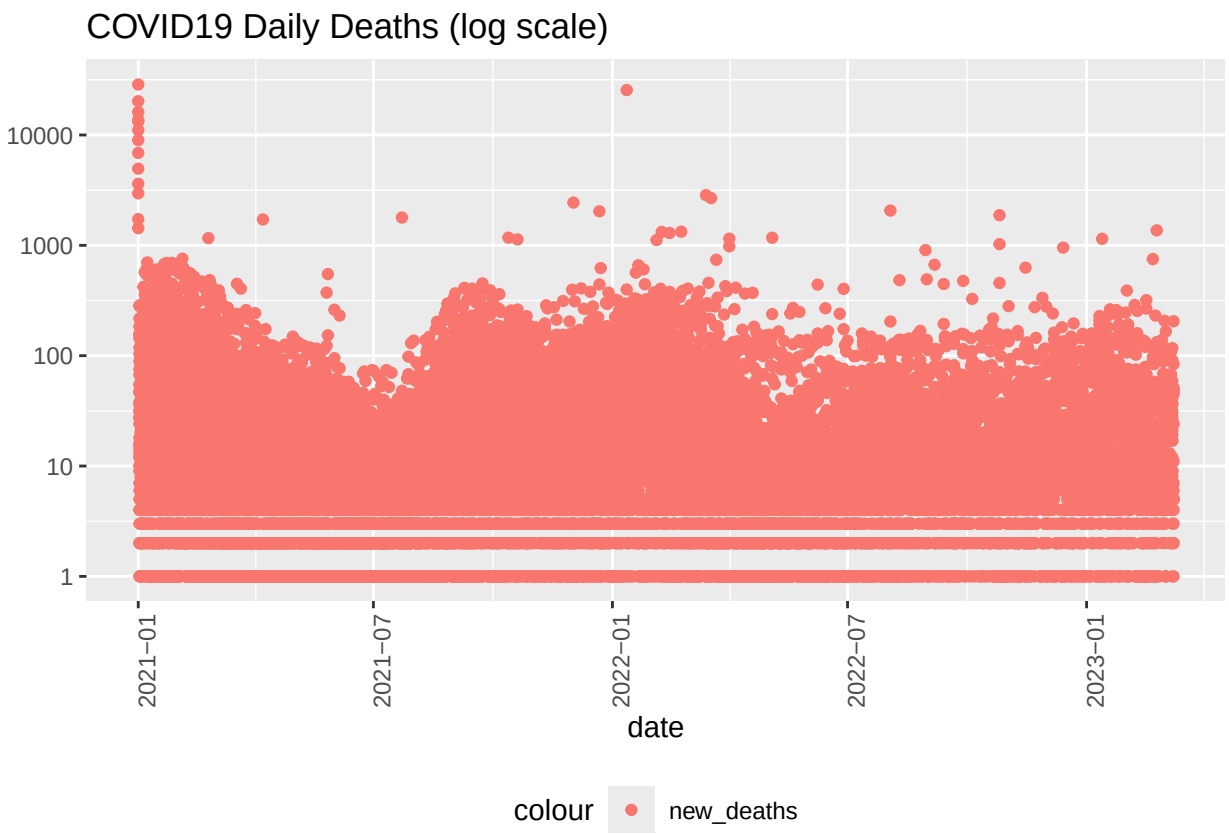
```
##  5 Arkansas                 4.31    3017804
##  6 Alabama                  4.29    4903185
##  7 Tennessee                4.21    6829174
##  8 Michigan                 4.20    9986857
##  9 Kentucky                 4.06    4467673
## 10 New Jersey               4.05    8882190
```

Adding daily new deaths data

We are specifically interested in daily deaths in each state, so we will add two new columns "new_deaths".

```
US_by_state <- US_by_state %>%
  mutate(new_deaths = deaths - lag(deaths))
```

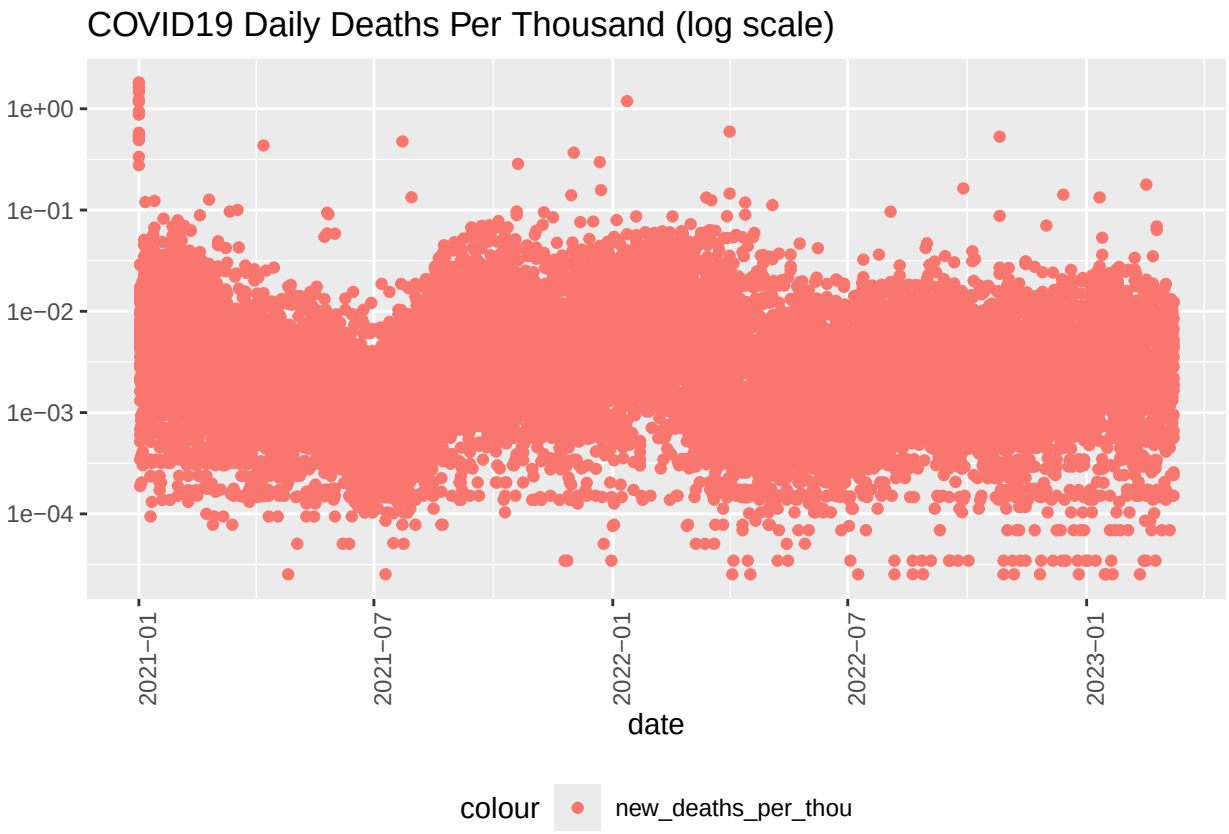
```
US_by_state %>%
  filter(new_deaths > 0) %>%
  ggplot(aes(x = date, y = new_deaths)) +
  #geom_line(aes(color = "new_deaths")) +
  geom_point(aes(color = "new_deaths")) +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title = str_c("COVID19 Daily Deaths (log scale)"), y = NULL)
```



Now lets add another column, "new_deaths_per_thou" to account for differences in states' population.

```
US_by_state <- US_by_state %>%
  mutate(new_deaths_per_thou = new_deaths * 1000 / Population)
```

```
US_by_state %>%
  filter(new_deaths_per_thou > 0) %>%
  ggplot(aes(x = date, y = new_deaths_per_thou)) +
  #geom_line(aes(color = "new_deaths_per_thou")) +
  geom_point(aes(color = "new_deaths_per_thou")) +
  scale_y_log10() +
  theme(legend.position = "bottom", axis.text.x = element_text(angle=90)) +
  labs(title = str_c("COVID19 Daily Deaths Per Thousand (log scale)"), y = NULL)
```



Adding state party affiliation

Next, we will add a new field to mark a state as "red" or "blue". We will official data from the Federal Election Commission. The full dataset is an Excel spreadsheet, available here: https://www.fec.gov/documents/4228/federalelections2020.xlsx.

The spreadsheet includes multiple sheets, we will be using sheet 9, "2020 Pres General Results". We will also remove spaces from the column names.

```
download.file('https://www.fec.gov/documents/4228/federalelections2020.xlsx', destfile = './federalelec

election_results <- read_excel('./federalelections2020-downloaded.xlsx', sheet = 9)

# remove spaces from column names
names(election_results) <- make.names(names(election_results), unique = TRUE)

colnames(election_results)
```

```
##  [1] "X1"                          "FEC.ID"
```

```
##  [3] "STATE"                          "STATE.ABBREVIATION"
##  [5] "GENERAL.ELECTION.DATE"          "FIRST.NAME"
##  [7] "LAST.NAME"                      "LAST.NAME...FIRST"
##  [9] "TOTAL.VOTES"                    "PARTY"
## [11] "GENERAL.RESULTS"                "GENERAL.."
## [13] "TOTAL.VOTES.."                  "COMBINED.GE.PARTY.TOTALS..NY."
## [15] "COMBINED....NY."                "WINNER.INDICATOR"
## [17] "ELECTORAL.VOTES"                "FOOTNOTES"
```

We will be using the "WINNER.INDICATOR" column with value of "W" or "W*" (for Maine) to get the winner for each state, and the "PARTY" column to determine the party affiliation for that winner. "STATE" column will be used to map the winner and their party to a specific state.

```
election_results_clean <- election_results %>%
  select(STATE, PARTY, WINNER.INDICATOR) %>%
  filter(WINNER.INDICATOR %in% c('W', 'W*'))
```

"PARTY" column has several different values, not just "D" or "R":

```
unique(election_results_clean$PARTY)
```

```
## [1] "R"                 "D"                 "DFL"
## [4] "Combined Parties:" "WF"
```

"DFL" is the Minnesota Democratic Party (https://dfl.org/about/), so we can treat value "DFL" as "D". New York state's "WF" is the Working Families party (https://workingfamilies.org/state/new-york/) and is also a Democratic party, so we can omit that row, since New York already has a winner entry with value "D". Additionally, "Combined Parties:" for New York appears to be a special indicator to mark both Democratic and Working Families parties as the winner, so we can also omit this row. Finally, we will drop "WINNER.INDICATOR" column, and will rename values of "D" to "blue", and "R" to "red" to make the results more readable for the final analysis.

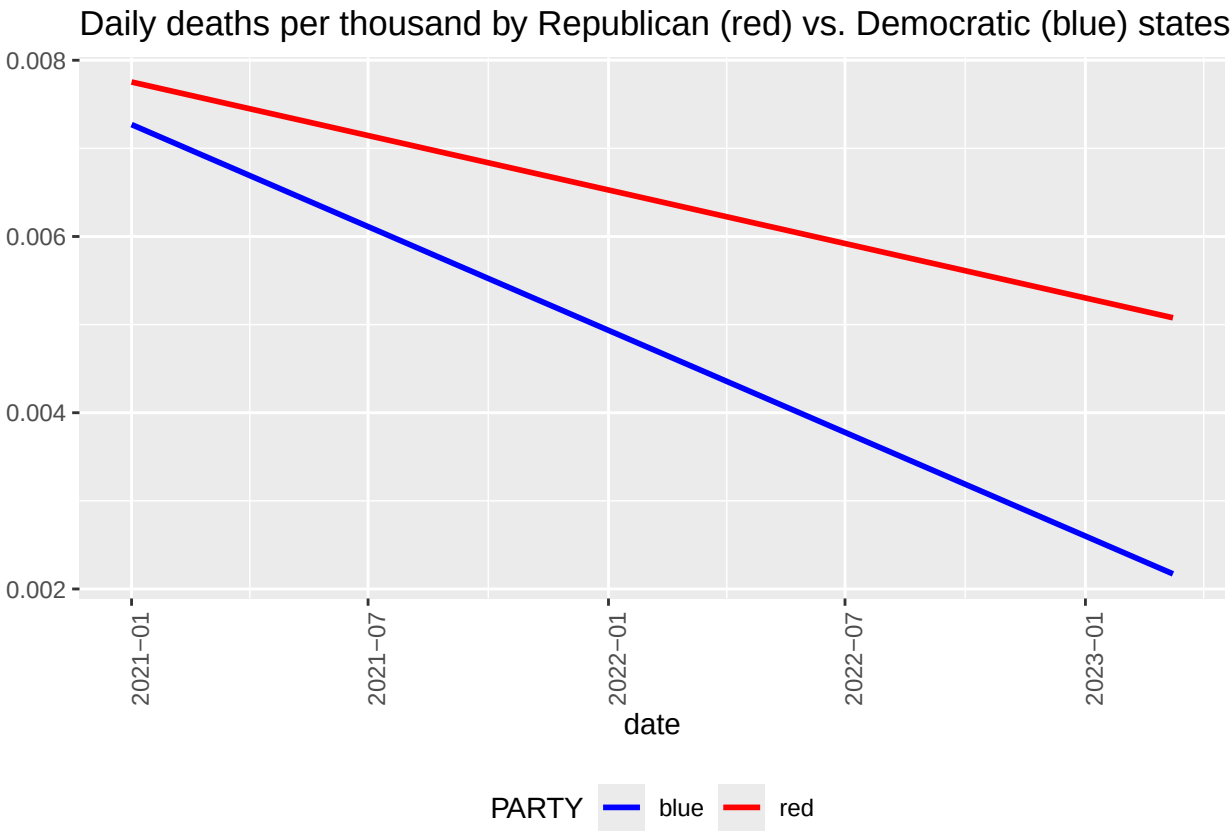
```
election_results_clean <- election_results_clean %>%
  filter(PARTY %in% c('D', 'R', 'DFL')) %>%
  mutate(PARTY = recode(PARTY, DFL = 'blue', R = 'red', D = 'blue')) %>%
  select(-WINNER.INDICATOR)
```

Next, we will merge the political affiliation data with our COVID19 by state dataset.

```
US_by_state <- US_by_state %>%
  full_join(election_results_clean, by = join_by(Province_State == STATE))
```

Now we can visualize the changes in daily deaths per thousand broken out by Democratic vs. Republican states. We will use a regression a.k.a. "trend" line against the state daily deaths per thousand data split by "red" vs. "blue" states.

```
US_by_state %>%
  filter(new_deaths_per_thou > 0, PARTY %in% c('red', 'blue')) %>%
  ggplot(aes(x = date, y = new_deaths_per_thou, color = PARTY)) +
  geom_smooth(method = "lm", se = FALSE, aes(group = PARTY)) +
  scale_color_manual(values = c("red" = "red", "blue" = "blue")) +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Daily deaths per thousand by Republican (red) vs. Democratic (blue) states", y = NULL)
```



Building a model to predict daily death rate per thousand based on states' political affiliation

The plot above strongly suggests that there's an observable difference in daily death rates per thousand between Republican and Democratic states. We will now build a linear model to get an idea of statistical significance of the "red" vs. "blue" classification, and how it affects the death rate.

```
mod <- lm(new_deaths_per_thou ~ PARTY, data = US_by_state)
summary(mod)
```

```
##
## Call:
## lm(formula = new_deaths_per_thou ~ PARTY, data = US_by_state)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.584   0.003   0.008   0.009   1.821
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.001858   0.002660  -0.698    0.485
## PARTYred    -0.005830   0.003800  -1.534    0.125
##
## Residual standard error: 0.3832 on 40695 degrees of freedom
##   (3991 observations deleted due to missingness)
## Multiple R-squared:  5.785e-05,  Adjusted R-squared:  3.328e-05
## F-statistic: 2.354 on 1 and 40695 DF,  p-value: 0.1249
```

The resulting model has very low R-squared number, and high p-values for the coefficients, and so we cannot use "red" vs. "blue" state classification alone to predict the death rate.

Conclusion and sources of bias

The plot of regression lines of "red" vs. "blue" state daily death rate per thousand does suggests that there is some sort of a correlation between the political affiliation of the state and the rates, but a simple linear model did not find any statistically significant correlation. This is most likely due to many other factors not present in this dataset which somehow relate to the political affiliation of the states.

A future study should consider adding new features to the data, for example:

- Policies and timelines surrounding mitigation efforts (e.g. mask mandates, social distancing) and their enforcement
- Vaccination rates
- Medical care availability
- Population density
- Various health metrics of a given state's population (e.g. obesity rates, smoking rates, etc.).

Potential sources of bias

The dataset used in this analysis may have various issues which introduce some sort of a bias, for example:

- Difficulty in attributing deaths specifically to COVID19:
  - https://www.aamc.org/news/how-are-covid-19-deaths-counted-it-s-complicated
- Issues with data reporting and collection from the various government agencies:
  - https://www.cidrap.umn.edu/covid-19/study-confusing-government-covid-reporting-requirements-led-disparities-hospital-data
- Classification of "red" vs. "blue" states based on 2020 election data can be problematic, for example, Joe Biden (D) won in Arizona making it a "blue" state, but the difference in popular votes was very small: 1,672,143 for Biden (D) vs. 1,661,686 for Trump (R), which is a difference of only 10,457 votes. Additionally, some states as "swing" or "purple" states, so their affiliation can change between elections, and their voter affiliation (Democrat vs. Republican) numbers may be very close.
  - https://www.fec.gov/documents/4228/federalelections2020.xlsx