COVID-19 Historical Data

V. Wingo

2025-04-18

COVID-19 Data for the US

While watching the lectures and considering what I wanted to do with this data, I recalled the more human aspects of the pandemic. Namely, the the theories about how it was handled and why, such as generational or political differences among those responsible for mitigating the public health emergency. For the purposes of this analysis, I will be focusing on how the political affiliation of the state governorship affected the population during the pandemic. While there are many other socioeconomic and political aspects that had effects on how the virus spread that I would like to investigate, I would like to start with this simpler factor

Obtaining and Cleaning Data

I am most interested in the US data so I will read in only the pertinent files. At this point, I am unsure what features I want to compare, so I am starting with the states, dates, cases, deaths, and populations from the data. Seeing as we have variable population densities, it's most likely a good idea to include the deaths per million as well.

```
url_in <-
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_data/csse_
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]) %>%
  select(-c(UID, iso2, iso3, code3, Country_Region, Lat, Long_, Combined_Key))
us_deaths <- read_csv(urls[2]) %>%
  select(-c(UID, iso2, iso3, code3, Country_Region, Lat, Long_, Combined_Key))
```

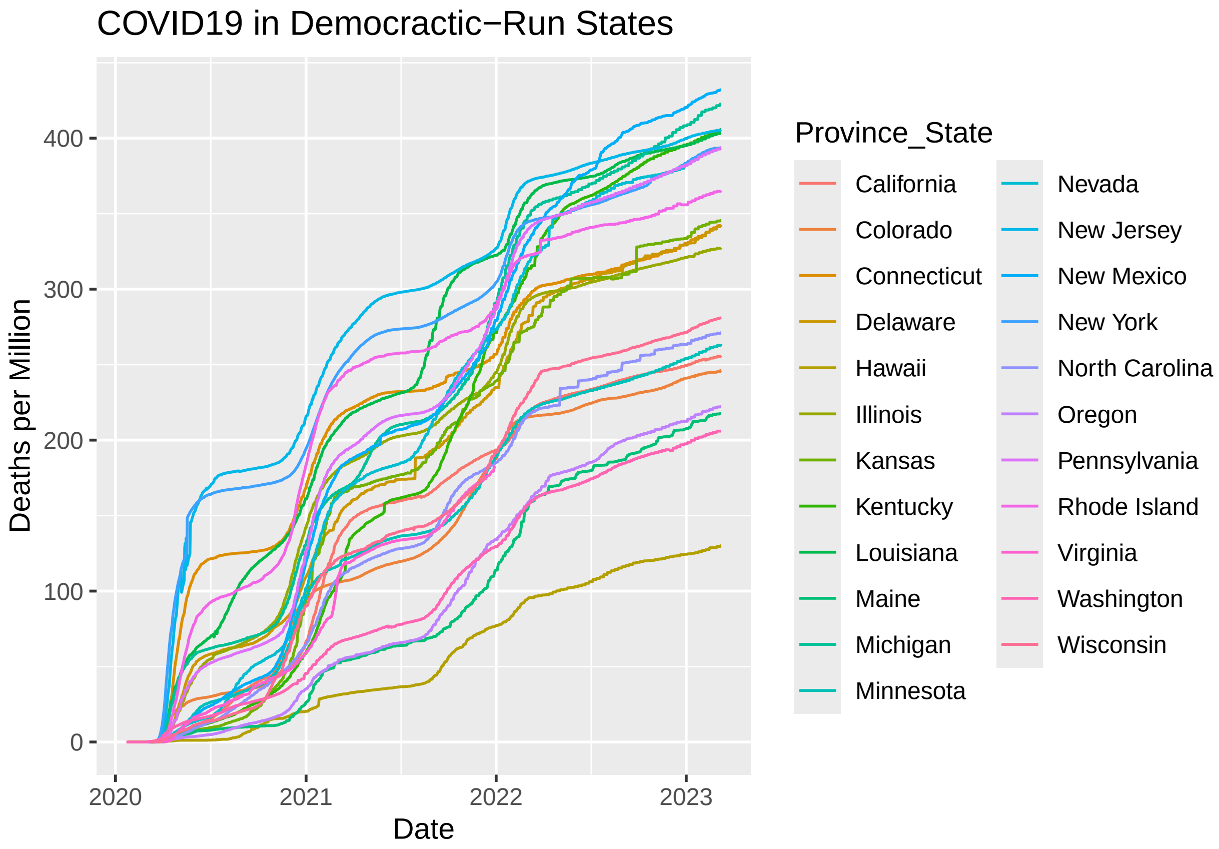
```
us_c <- us_cases %>%
  pivot_longer(cols = -c('FIPS', 'Admin2', 'Province_State'),
               names_to = "date",
               values_to = "cases")

us_d <- us_deaths %>%
  pivot_longer(cols = -c('FIPS', 'Admin2', 'Province_State', 'Population'),
               names_to = "date",
               values_to = "deaths")

US <- us_c %>%
  full_join(us_d) %>%
  filter(cases > 0) %>%
  select(-c(FIPS, Admin2)) %>%
  mutate(date = mdy(date))

US_by_state <- US %>%
  group_by(Province_State, date) %>%
  summarize(cases = sum(cases), deaths=sum(deaths),
           Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 100000 / Population) %>%
  select(Province_State, date, cases, deaths, deaths_per_mill)
```

In the following code chunk, I create vectors delineating between states whose governors during the time period were affiliated with the Democratic or Republican party. Some states changed governorship and party affiliation. I decided to remove some of these states from the list due to the short amount of time they had these affiliations. For those left in, I divided the data based upon when the affiliation changed.

```
R <- c("Alabama", "Alaska", "Arkansas", 'Florida', 'Georgia', 'Idaho', 'Iowa', 'Ohio',
       'Oklahoma', 'South Carolina', 'Tennessee', 'Texas', 'Vermont', 'Wyoming',
       'Indiana', 'Mississippi', 'Missouri', 'Nebraska', 'New Hampshire',
       'North Dakota', 'South Dakota', 'Utah', 'West Virginia')
D <- c("California", "Colorado", "Connecticut", 'Delaware', 'Illinois', 'Kansas',
       'Kentucky', 'Maine', 'Michigan', 'Minnesota', 'New Jersey', 'New Mexico',
       'Wisconsin', 'Hawaii', 'New York', 'North Carolina', 'Oregon',
       'Pennsylvania', 'Rhode Island', 'Washington', 'Louisiana')

R_States <- US_by_state[which(US_by_state$Province_State %in% R), ]
R_2 <- US_by_state[
  (which(US_by_state$Province_State %in% c("Arizona", "Maryland", "Massachusetts")
         & year(US_by_state$date) < 2023)), ]
R_3 <- US_by_state[(which(US_by_state$Province_State == 'Montana'
                          & year(US_by_state$date) >= 2021)), ]
#R_4 <- US_by_state[(which(US_by_state$Province_State == 'Nevada'
# & year(US_by_state$date) >= 2023)), ]
R_5 <- US_by_state[(which(US_by_state$Province_State == 'Virginia'
                          & year(US_by_state$date) >= 2022)), ]
R_States <- R_States %>% rbind(R_2, R_3, R_5)

D_States <- US_by_state[which(US_by_state$Province_State %in% D), ]
#D_2 <- US_by_state[
# (which(US_by_state$Province_State %in% c("Arizona", "Maryland", "Massachusetts")
# & year(US_by_state$date) >= 2023)), ]
#D_3 <- US_by_state[(which(US_by_state$Province_State == 'Montana'
# & year(US_by_state$date) < 2021)), ]
D_4 <- US_by_state[(which(US_by_state$Province_State == 'Nevada'
                          & year(US_by_state$date) < 2023)), ]
D_5 <- US_by_state[(which(US_by_state$Province_State == 'Virginia'
                          & year(US_by_state$date) < 2022)), ]
D_States <- D_States %>% rbind(D_4, D_5)
```

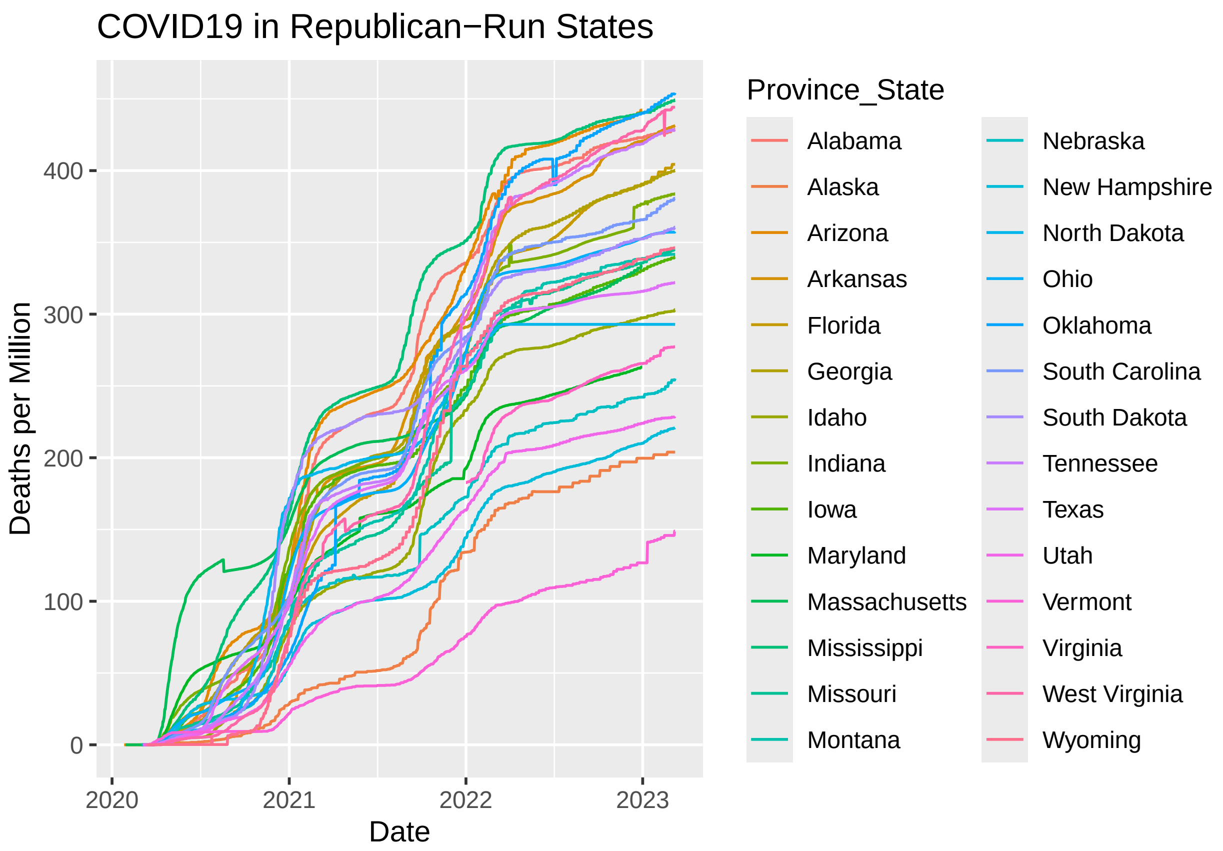
In the following plot, we graph the deaths per million over time in the states headed by a Democractic governor. As you can undoubtedly see in this and the next plot, where we investigate the Republican-run states, the number of lines and their indistinct colors can make it difficult to understand what the data is showing us and how they compare to one another. There are clearer "jumps" where almost all states see a similar sudden increase in deaths and then a period of stabilization.

```
D_States %>%
  ggplot(aes(x = date, color = Province_State)) +
  geom_line(aes(y = deaths_per_mill)) +
  theme(legend.position = "right") +
  labs(title = str_c("COVID19 in Democractic-Run States"),
       x = "Date",
       y = "Deaths per Million")
```



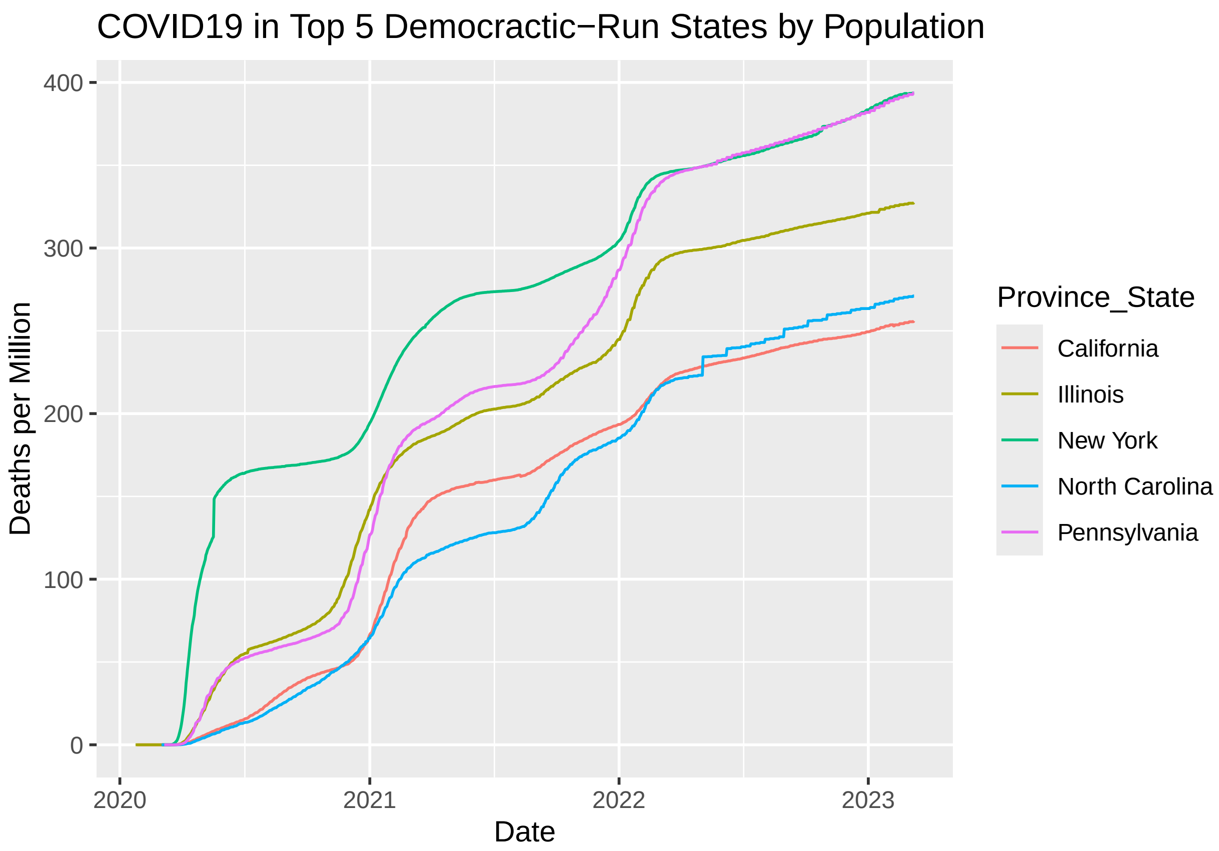
```
R_States %>%
  ggplot(aes(x = date, color = Province_State)) +
  geom_line(aes(y = deaths_per_mill)) +
  theme(legend.position = "right") +
  labs(title = str_c("COVID19 in Republican-Run States"),
       x = "Date",
       y = "Deaths per Million")
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```

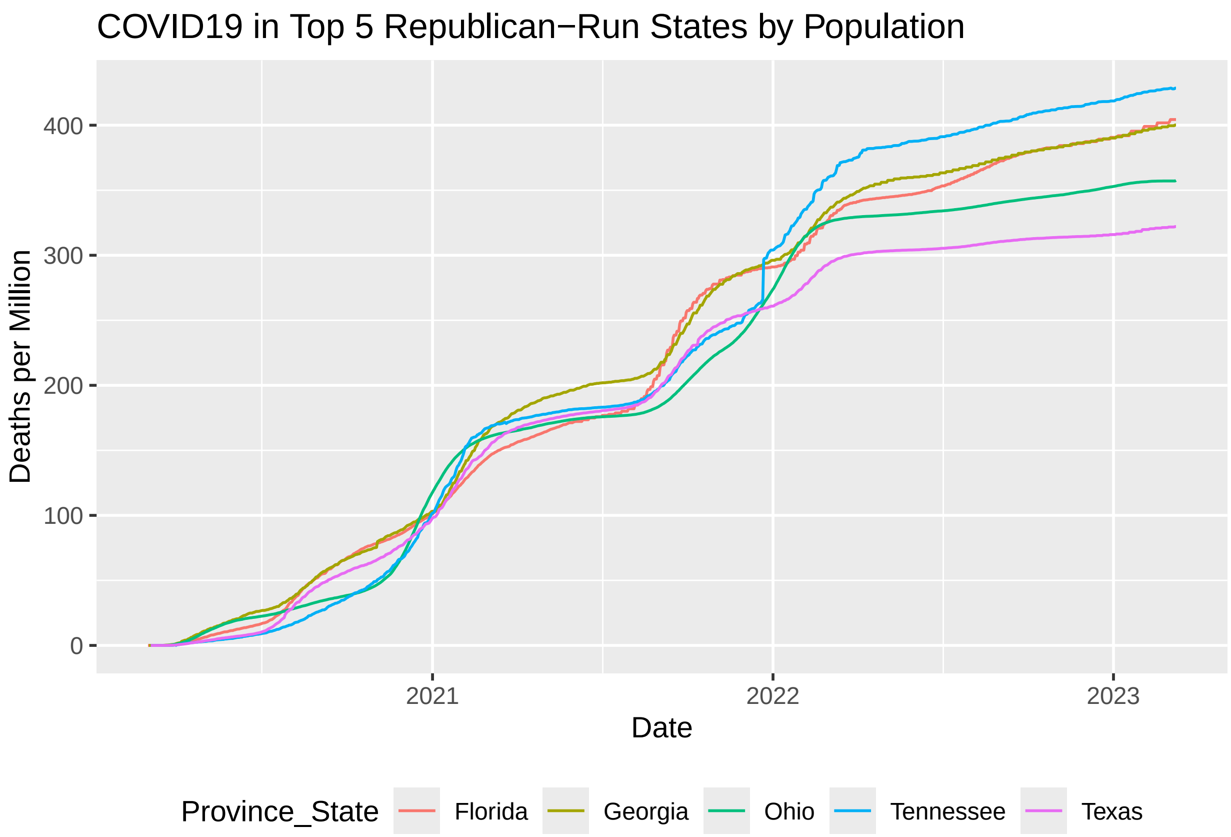


In the following plots, I aimed to narrow the data in order to have a clearer picture of how the data looks and make it easier to compare our two categories. In order to do this, I graphed only the top 5 most populous states from each category, excluding states that changed party affiliation for the sake of simplicity.

```
D_States[
  which(D_States$Province_State %in%
          c('California', 'New York', 'Pennsylvania', 'Illinois',
            'North Carolina')), ] %>%
  ggplot(aes(x = date, color = Province_State)) +
  geom_line(aes(y = deaths_per_mill)) +
  theme(legend.position = "right") +
  labs(title = str_c("COVID19 in Top 5 Democractic-Run States by Population"),
       x = "Date",
       y = "Deaths per Million")
```

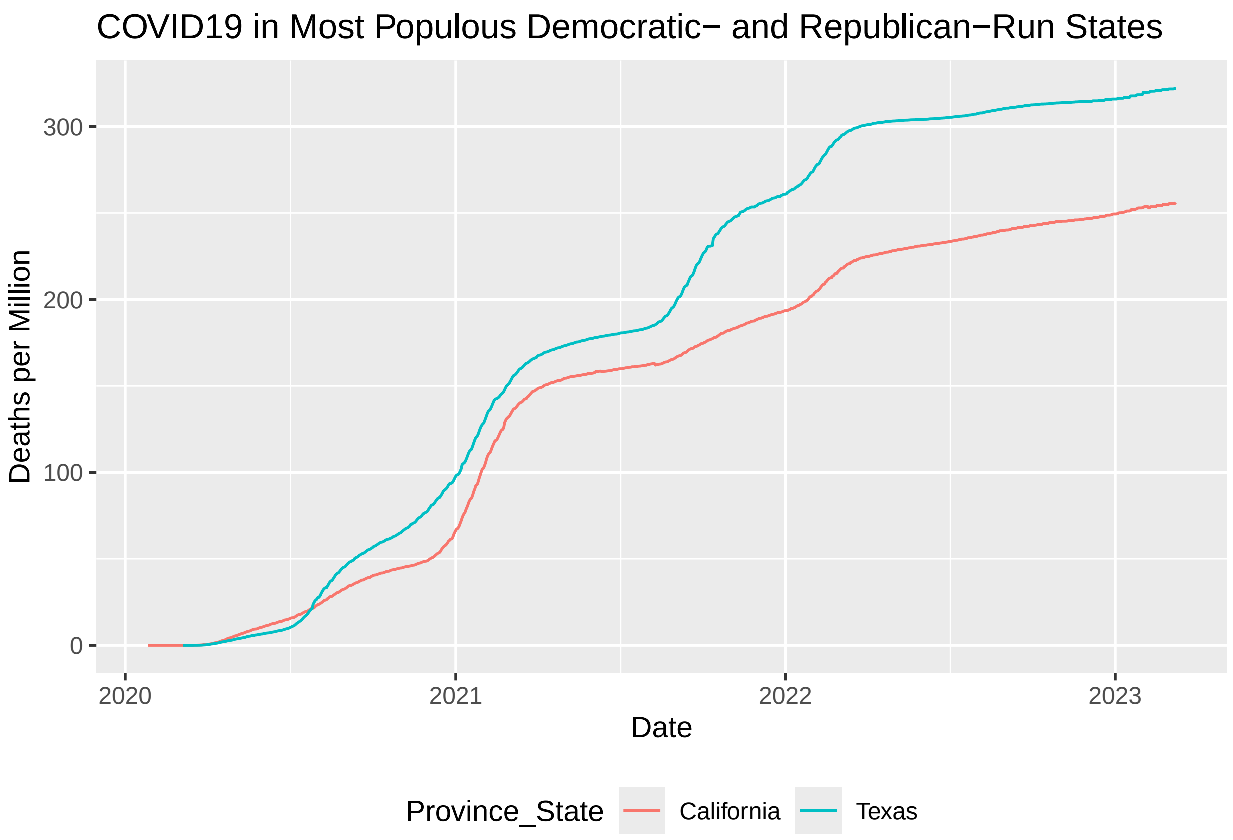


```
R_States[which(R_States$Province_State
               %in% c('Texas', 'Florida', 'Ohio', 'Georgia', 'Tennessee')), ] %>%
  ggplot(aes(x = date, color = Province_State)) +
  geom_line(aes(y = deaths_per_mill)) +
  theme(legend.position = "bottom") +
  labs(title = str_c("COVID19 in Top 5 Republican-Run States by Population"),
       x = "Date",
       y = "Deaths per Million")
```



Now we can see that Republican-run states have a consistent shape and, while they appear lower in some places compared to the Democratic-run states, they end higher. Meanwhile, their Democratic counterparts have similar shapes but are more spread, with New York having a steep initial rise. It is still not clear how these two categories really compare, so in our next graph, we will plot the most populous states from both categories, California and Texas.

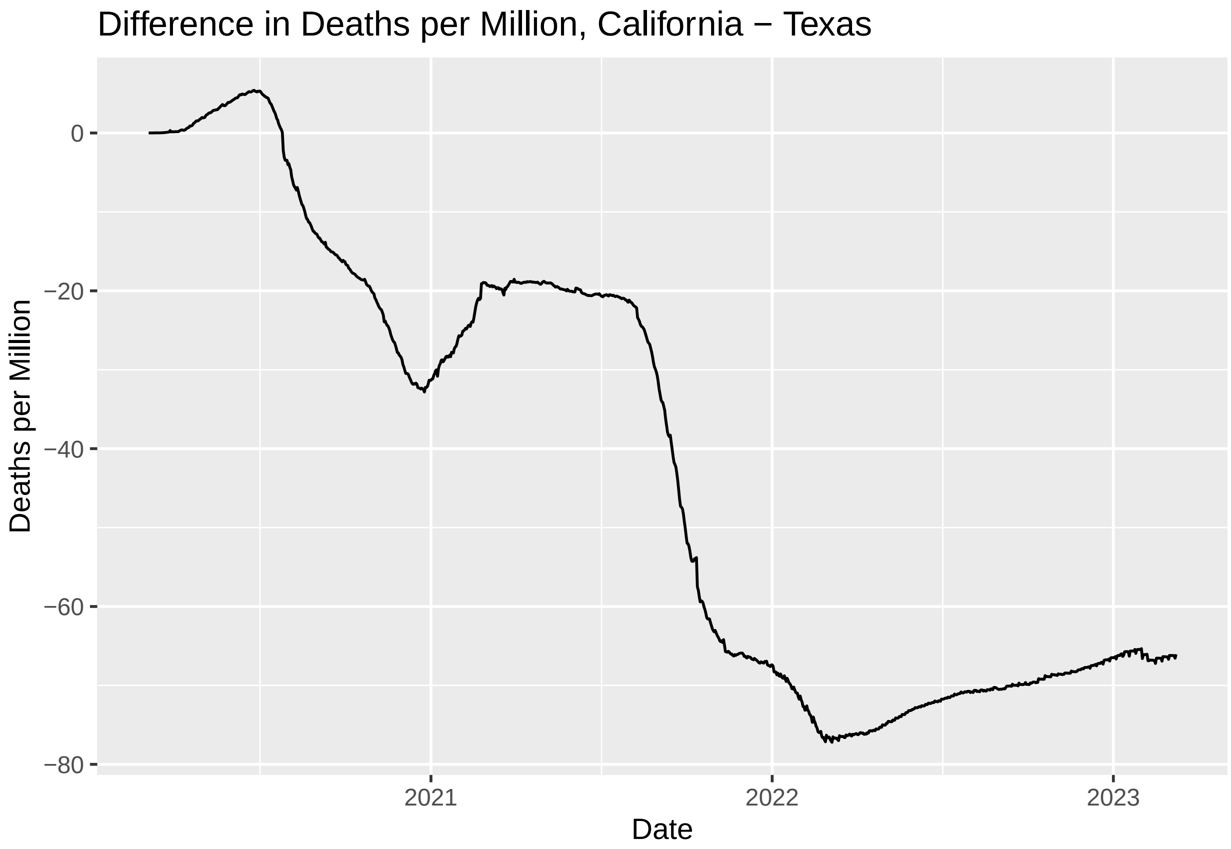
```r
US_by_state[which(US_by_state$Province_State
                  %in% c('California', 'Texas')), ] %>%
  ggplot(aes(x = date, color = Province_State)) +
  geom_line(aes(y = deaths_per_mill)) +
  theme(legend.position = "bottom") +
  labs(title = str_c("COVID19 in Most Populous Democratic- and Republican-Run States"),
       x = "Date", y = "Deaths per Million")
```



Here we can see more interesting patterns in the data. While California's deaths per million were initially higher, there is a point in mid-2020 when it is overtaken by Texas numbers and remains so. There are also fewer steep jumps with marginal slopes in between as compared to the Texas data. The lines remain close until mid-2021 where we see a sharp increase in deaths, reflected by a much smaller increase in California.

As a less skilled statistician, it was difficult for me to think of a way to best model these differences. So in the following plot and subsequent model, I consider the differences between the data from these two states.

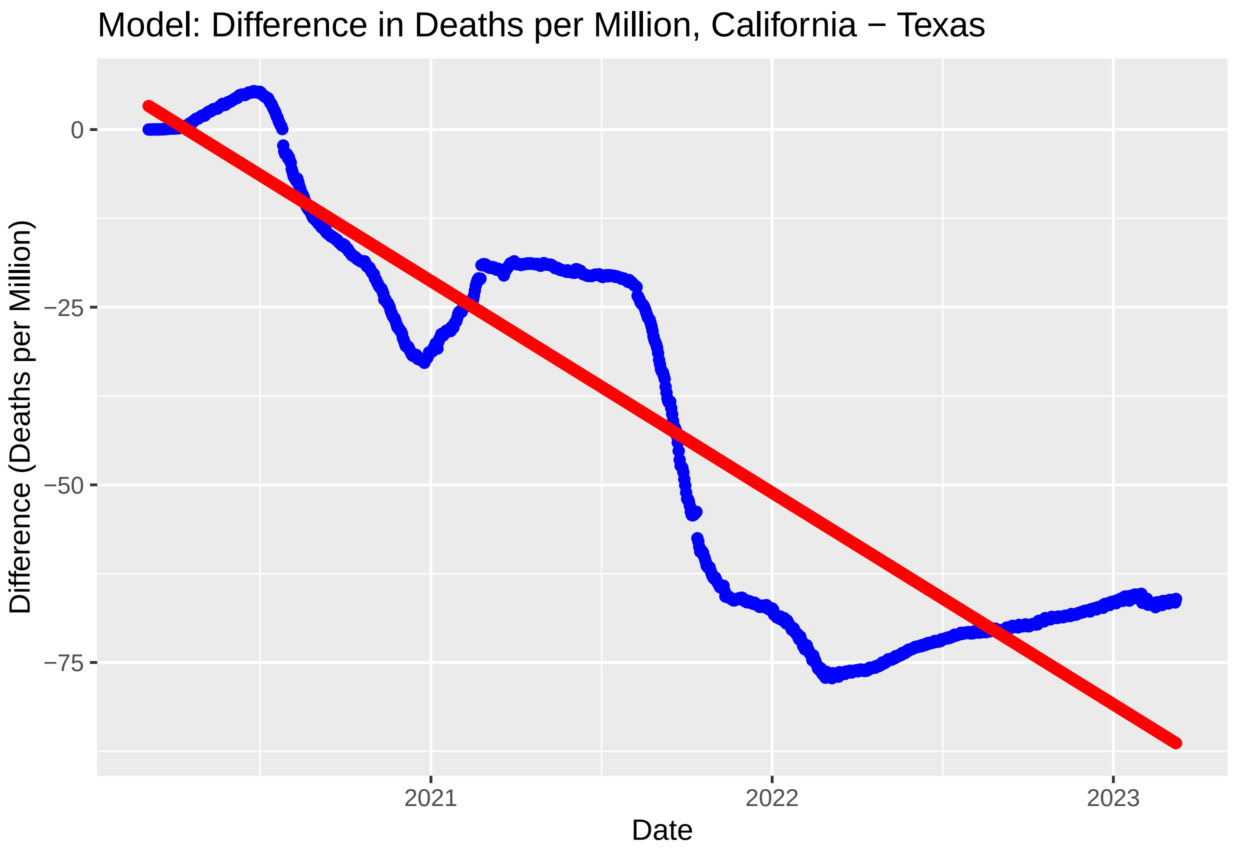
```
cali <- US_by_state[which(US_by_state$Province_State == 'California'), ]
texas <- US_by_state[which(US_by_state$Province_State == 'Texas'), ]
dates <- unique(texas$date)
cali <- cali[which(cali$date %in% dates),]
dates <- unique(cali$date)
texas <- texas[which(texas$date %in% dates),]
diff <- c()
for (i in 1:length(dates)) {
  date <- dates[i]
  diff[i] <- cali$deaths_per_mill[which(cali$date == date)] -
    texas$deaths_per_mill[which(texas$date == date)]
}
diff_df <- data.frame(
  date = dates,
  diff = diff
)

diff_df %>%
  ggplot(aes(x=date, y = diff)) +
  geom_line() +
  labs(title = str_c("Difference in Deaths per Million, California - Texas"),
       x = "Date",
       y = "Deaths per Million")
```



```
mod <- lm(diff ~ date, data = diff_df)
summary(mod)
```

```
##
## Call:
## lm(formula = diff ~ date, data = diff_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3885  -9.4735   0.0236  10.3024  20.2958
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.499e+03  2.084e+01   71.93   <2e-16 ***
## date        -8.160e-02  1.104e-03  -73.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.62 on 1098 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8326
## F-statistic:  5466 on 1 and 1098 DF,  p-value: < 2.2e-16
```

```
diff_df$pred = predict(mod, data = diff_df)
diff_df %>%
  ggplot() +
  geom_point(aes(x = date, y = diff), color = "blue") +
  geom_point(aes(x = date, y = pred), color="red") +
  labs(title = str_c("Model: Difference in Deaths per Million, California - Texas"),
       x = "Date",
       y = "Difference (Deaths per Million)")
```



Conclusions

With my limited expertise, it is difficult to confirm one way or the other if political affiliation of the state governorship had a significant affect on COVID-related deaths. From what I have seen here, it appears worthy of investigating and I would like to revisit it in the future when I have more nuanced and appropriate tools at my disposal. There are other related factors at play, such as available resources for medical care, which can often be influenced by state leadership.

Bias

Of course, during the height of the COVID-19 pandemic, there was a lot of controversy about this data itself. There is bias in both testing and reporting. There was a time before we knew about the virus and there were tests. We cannot guarantee that potential cases were tested due to the belief of those infected or hospital staff. There were times when supplies were low and a test couldn't be administered. Then there were issues with the numbers; reports came out that states weren't supplying accurate numbers or numbers at all and the CDC was under threat from the administration. And this wasn't just an issue in the US, but other countries as well. There is no way of knowing if we have an accurate picture of how the pandemic affected the world.

There is also bias in what I chose to investigate. I chose the US because I live there and I chose to investigate how the politics of state governors affected how the pandemic spread because I believed that it determined whether or not it was handled effectively. This could extend to investigating different countries based on their politics at the time or their dominant ethnicity.