High COVID19 Death Rates in Indiana, Michigan, and Ohio

Jacob Moore

2022-10-18

Introduction

The objective of this study is to present and analyze COVID19 data for the US states of Indiana, Michigan, and Ohio, in order to determine if, despite their proximity, there may be some non-geographical variable which caused a notable difference in either case numbers or deaths.

This study will first present case and death numbers for the US as a whole in order to contextualize the state-level data. Then, the individual data for each state is presented and compared.

All written code and console outputs are preserved within this report in order to facilitate their examination.

```
library(tidyverse)
library(lubridate)
```

```
### Get current Data in the four files
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_deaths_US.csv")

urls <- str_c(url_in, file_names)
```

```
US_cases <- read_csv(urls[1], show_col_types = FALSE)
US_deaths <- read_csv((urls[2]), show_col_types = FALSE)

uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U

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

```
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_))

US <- US_cases %>%
  full_join(US_deaths)

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()

US_by_state <- US_by_state %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))

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) %>%
  ungroup()

US_totals <- US_totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))
```

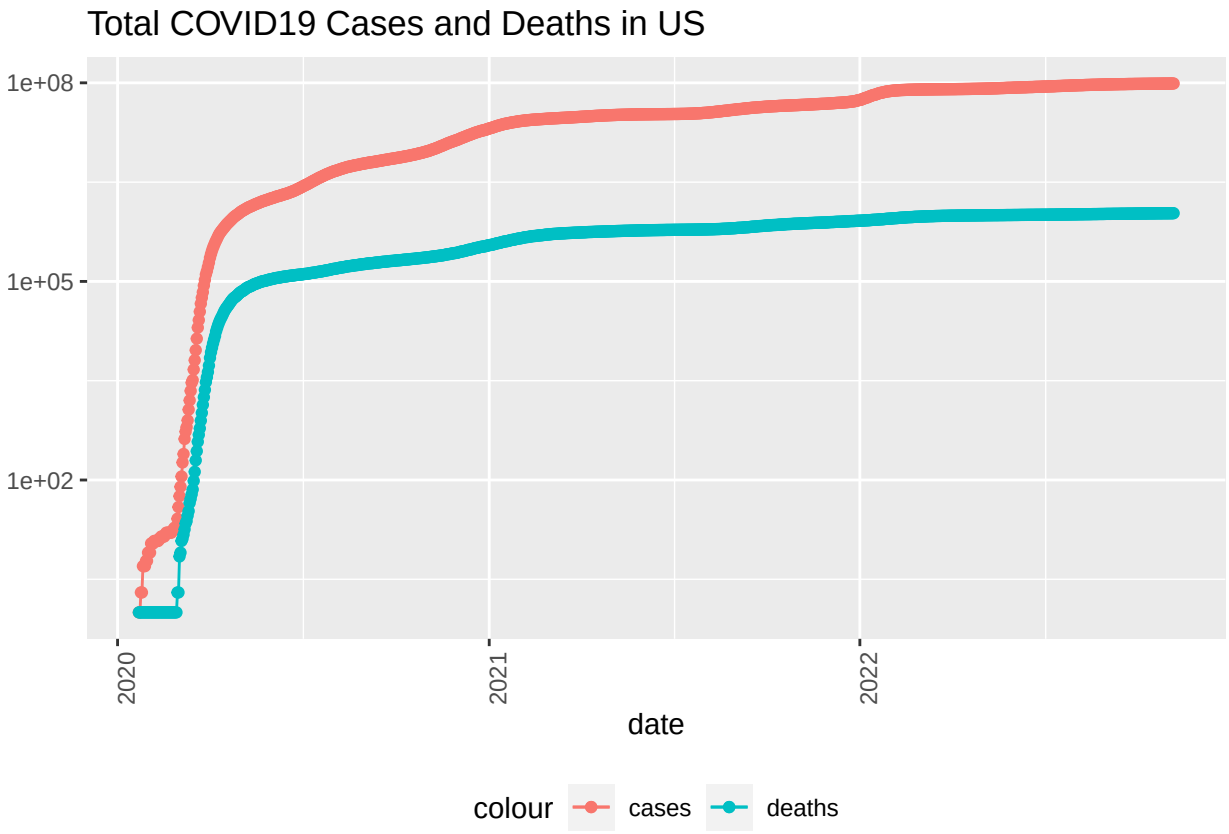
US Data

Below is the visualization of total COVID19 cases and deaths. Note that the visualization is presented on a logarithmic scale to preserve legibility. We can see that while growth continues to the present in both cases and deaths, growth overall has not reached the frighteningly rapid pace of 2020 in either 2021 or 2022.

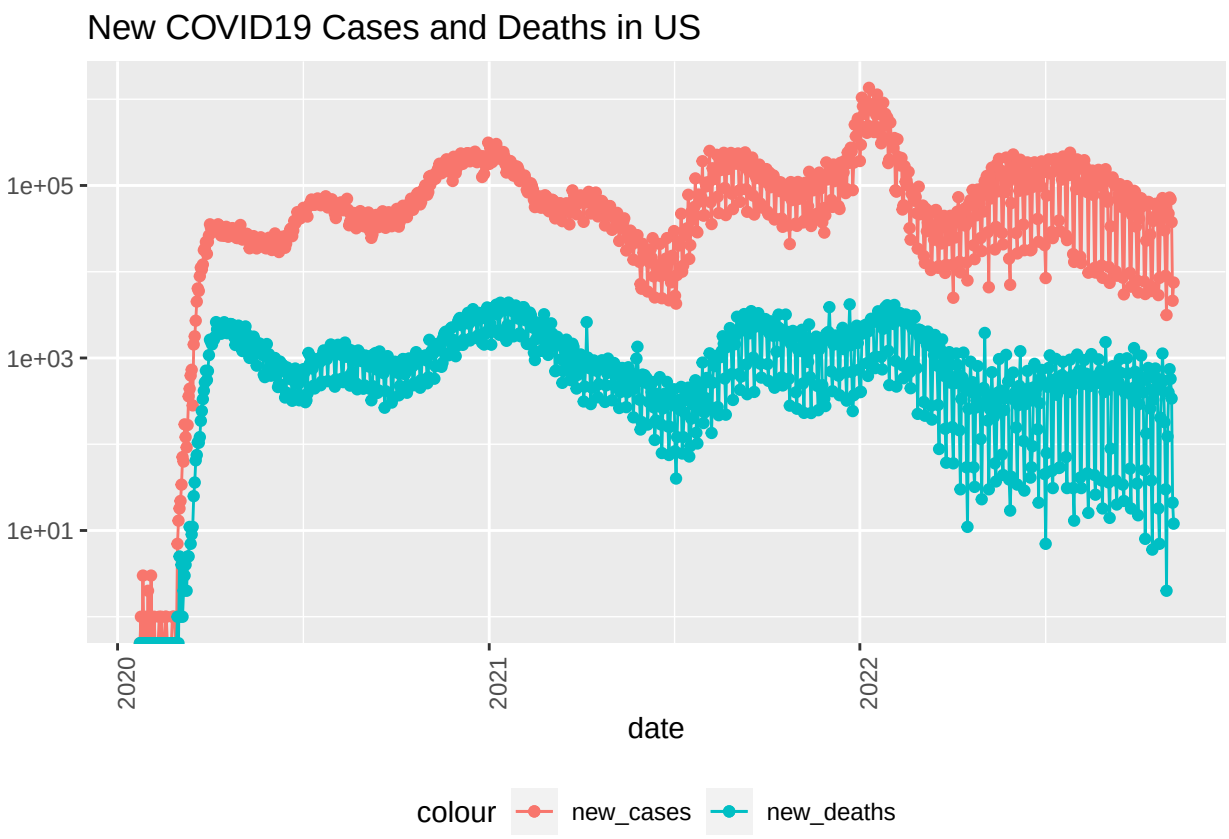
Further below is the visualization for new cases and deaths in the US as a whole. We can see, as expected from the visualization of totals, that new cases level off in mid-2020, though not without some spikes, especially in the winter months.

```
# graphs for US
US_totals %>%
  filter(cases > 0 ) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
```

```
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Total COVID19 Cases and Deaths in US", y=NULL)
```




```
US_totals %>%
  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 = "New COVID19 Cases and Deaths in US", y=NULL)
```



Best and Worst Death Rates in US States and Territories

Below is a table showing the states and territories within the US with the lowest death rate as measured by deaths per 1000 people in the population. Below that is a similar table showing those states with the highest death rate. Note the presence of Michigan in the second table as the state with the tenth highest death rate. Note also that death rates and case rates appear to correlate.

```
#transform 'by state' data
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)

# 10 best states
US_state_totals %>%
  slice_min(deaths_per_thou, n = 10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State        deaths  cases population
##              <dbl>          <dbl> <chr>                 <dbl>  <dbl>      <dbl>
##  1           0.611           148. American Samoa           34 8.26e3      55641
##  2           0.744           240. Northern Mariana Isl~    41 1.32e4      55144
```

```
##  3           1.16            218. Virgin Islands           124 2.34e4     107268
##  4           1.21            233. Vermont                  754 1.45e5     623989
##  5           1.21            256. Hawaii                  1711 3.63e5    1415872
##  6           1.41            263. Puerto Rico             5279 9.88e5    3754939
##  7           1.58            327. Utah                    5065 1.05e6    3205958
##  8           1.91            406. Alaska                  1413 3.01e5     740995
##  9           1.92            242. Washington             14597 1.84e6    7614893
## 10           1.98            241. District of Columbia    1397 1.70e5     705749
```

```
# 10 worst states
US_state_totals %>%
  slice_max(deaths_per_thou, n = 10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State deaths   cases population
##              <dbl>          <dbl> <chr>           <dbl>   <dbl>      <dbl>
##  1            4.37           314. Mississippi     12992  934401    2976149
##  2            4.34           315. Arizona         31573 2293015    7278717
##  3            4.33           306. Oklahoma        17138 1211210    3956971
##  4            4.20           340. West Virginia    7534  609356    1792147
##  5            4.19           313. Alabama         20558 1534287    4903185
##  6            4.15           319. Arkansas        12521  961765    3017804
##  7            4.13           301. New Mexico       8664  630704    2096829
##  8            4.12           346. Tennessee       28122 2364399    6829174
##  9            3.95           290. Michigan        39406 2897827    9986857
## 10            3.93           316. New Jersey      34938 2809400    8882190
```

Modeling Death Rates and Case Rates

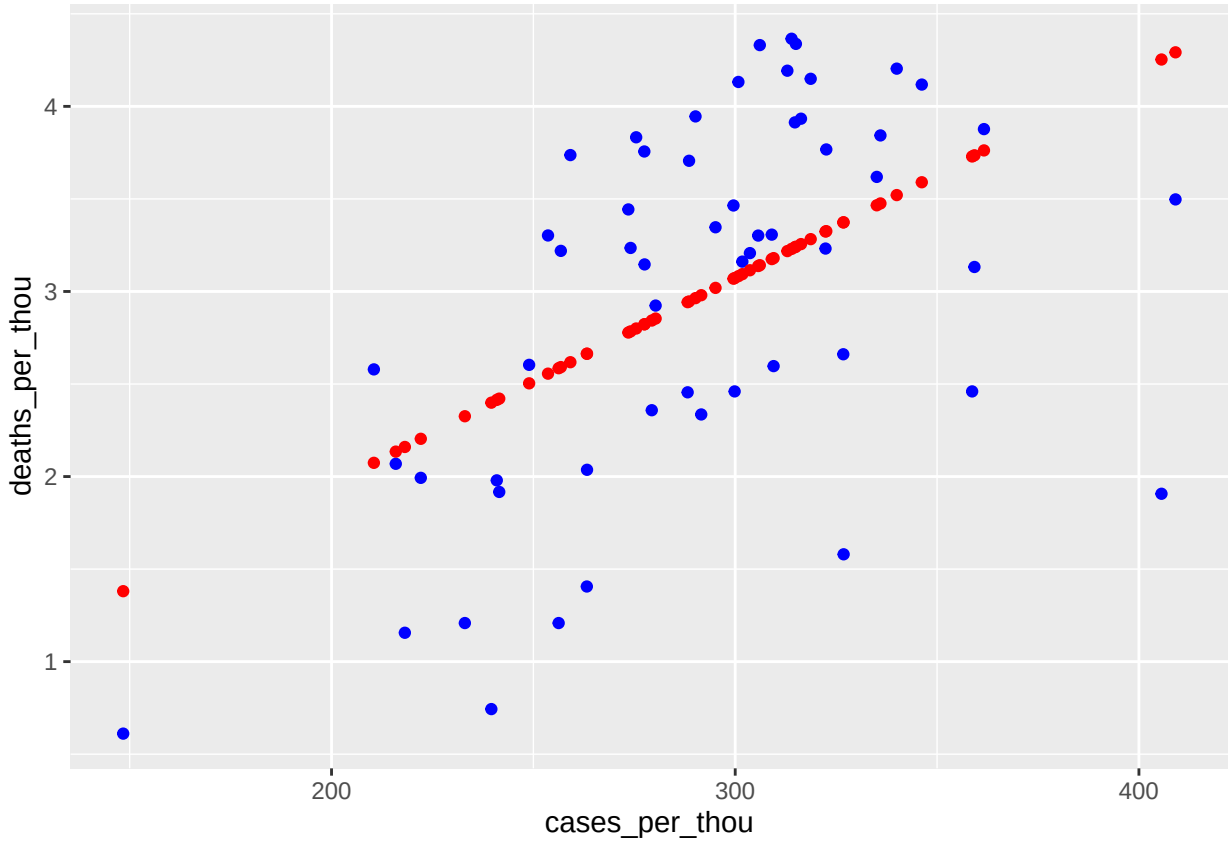
The model below formalizes the apparent relationship between death rate and case rate. We can see a clear correlation but the relationship is, of course, not perfectly linear.

```
mod <- lm(deaths_per_thou ~ cases_per_thou, data = US_state_totals)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3464 -0.6057  0.1236  0.6753  1.1892
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.276939   0.716325  -0.387    0.701
## cases_per_thou  0.011169   0.002424   4.608 2.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8516 on 54 degrees of freedom
```

```
## Multiple R-squared:  0.2823, Adjusted R-squared:  0.269
## F-statistic: 21.24 on 1 and 54 DF,  p-value: 2.52e-05
```

```
US_state_totals_w_pred <- US_state_totals %>% mutate(pred = predict(mod))
US_state_totals_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")
```

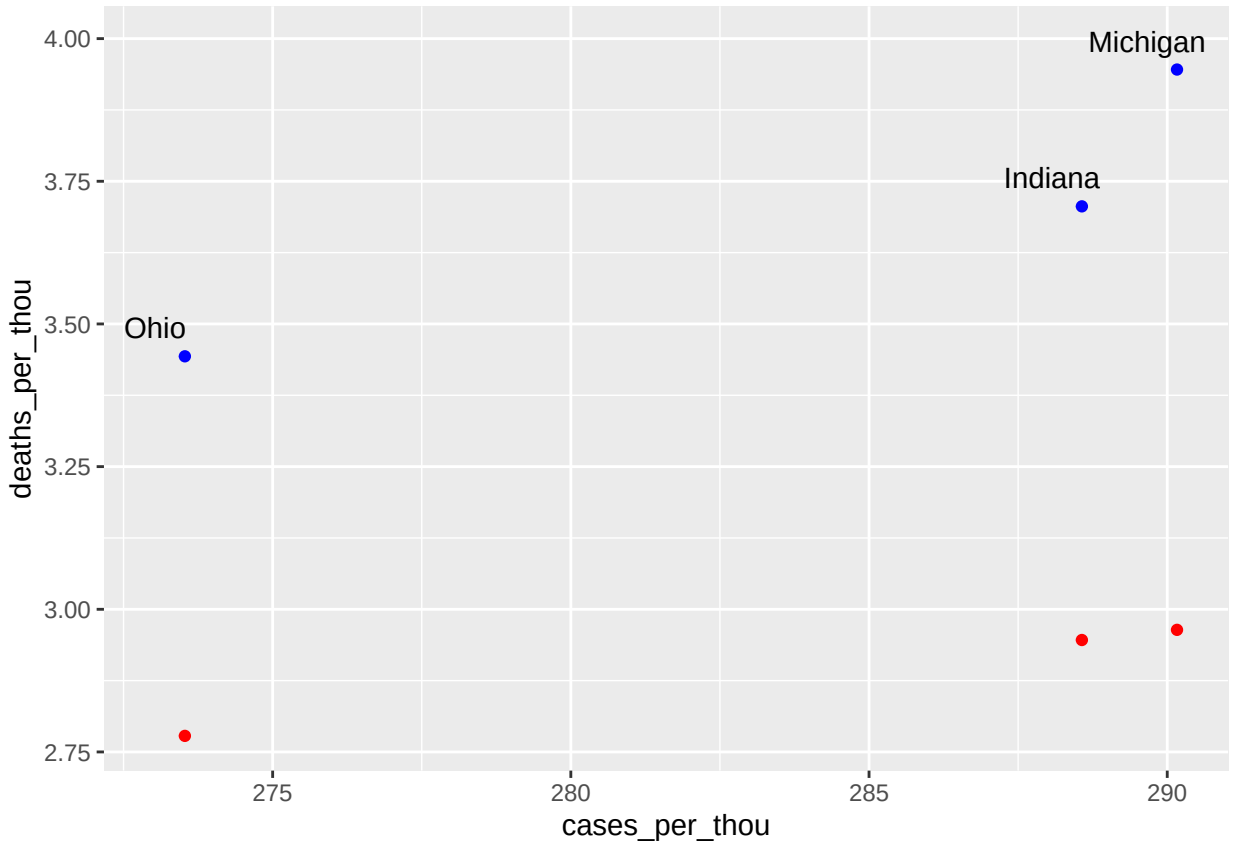


Regional Clusters

When we isolate only those states in which we are interested, Indiana, Michigan, and Ohio, we can see that all three have more deaths than predicted by our model. We also see that Ohio has fewer cases, but performs no better with respect to deaths.

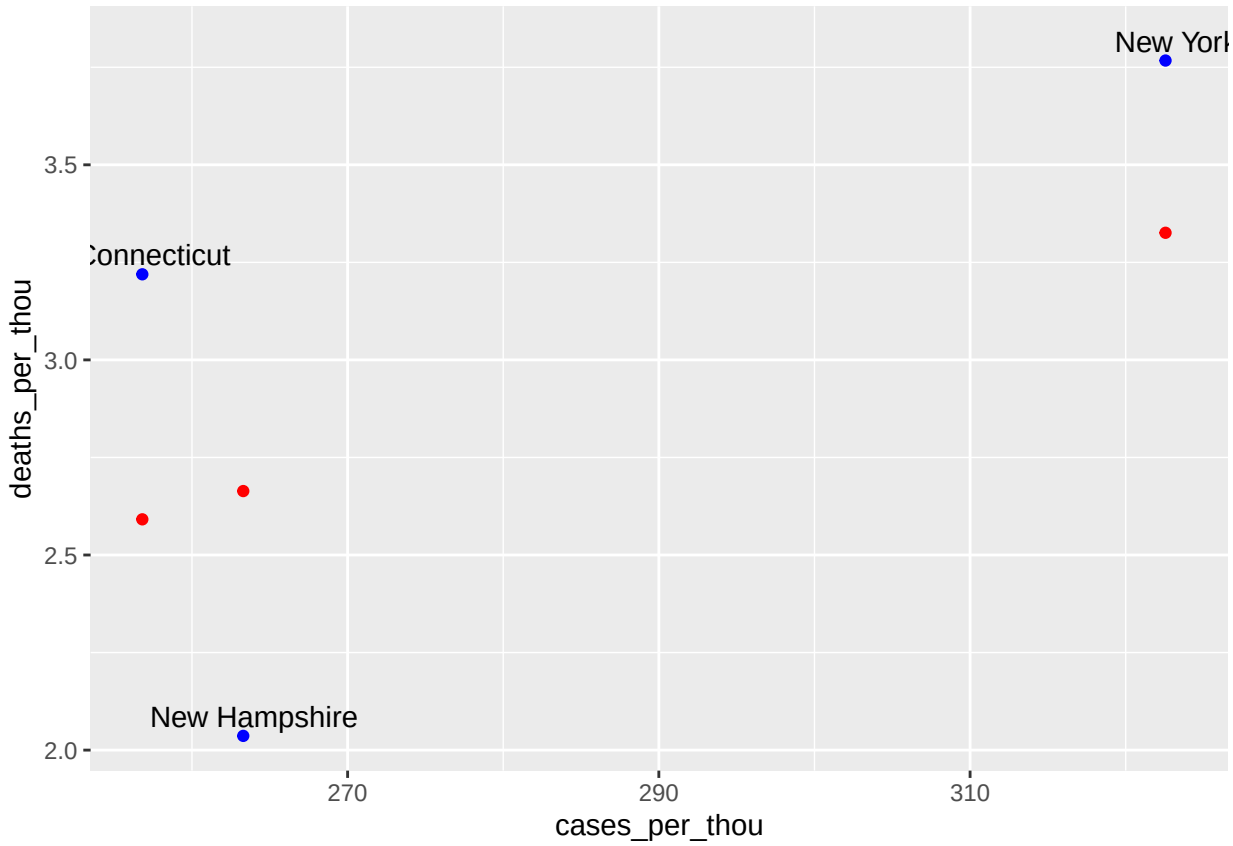
These results may point to some shared regional factor or factors that increase the risk of death for those who contracted the disease.

```
# add model that only includes my states!
my_state_totals_w_pred <- US_state_totals_w_pred %>% filter(Province_State == "Michigan" | Province_Sta

my_state_totals_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") +
  geom_text(aes(x = cases_per_thou, y = deaths_per_thou, label = Province_State), nudge_x = -.5, nudge_
```



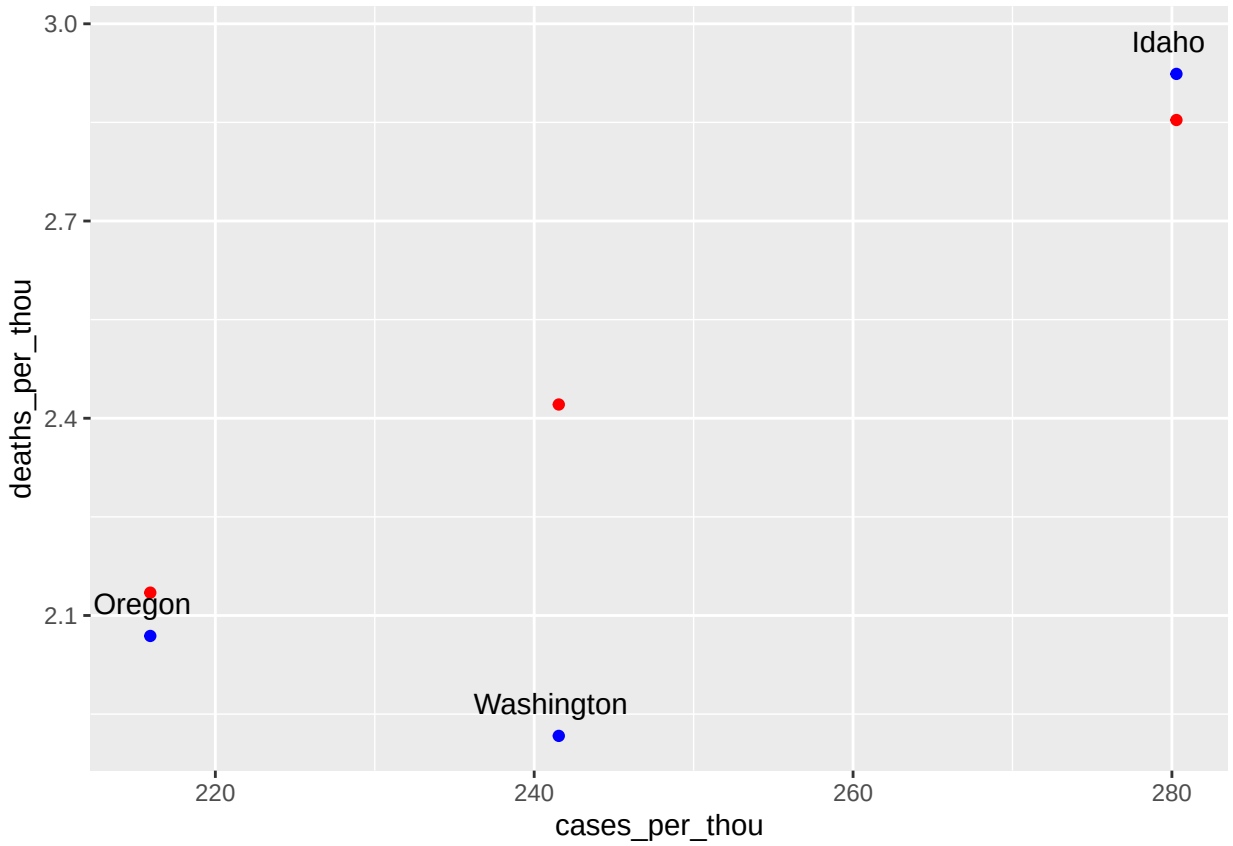
Other regional clusters of states do not have this same tendency. For example, in New England, if we look at New York, Connecticut, and Rhode Island:

```
# add model that only includes my states!
new_england_state_totals_w_pred <- US_state_totals_w_pred %>% filter(Province_State == "New York" | Pro

new_england_state_totals_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") +
  geom_text(aes(x = cases_per_thou, y = deaths_per_thou, label = Province_State), nudge_x = .7, nudge_y
```



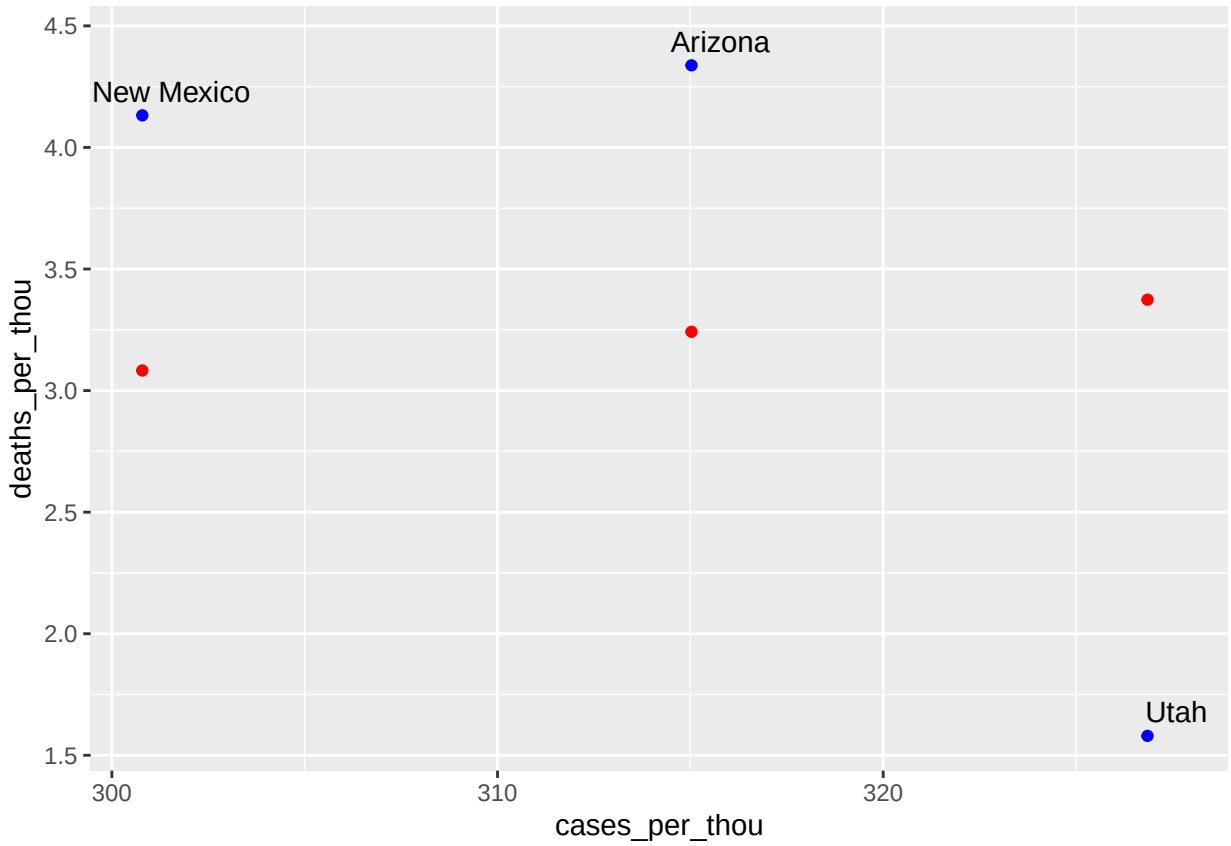
If we look to the Pacific Northwest in Washington, Oregon, and Idaho:

```
# add model that only includes my states!
northwest_state_totals_w_pred <- US_state_totals_w_pred %>% filter(Province_State == "Washington" | Prov

northwest_state_totals_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") +
  geom_text(aes(x = cases_per_thou, y = deaths_per_thou, label = Province_State), nudge_x = -.5, nudge_
```



Finally, looking at the Southwestern states of New Mexico, Utah, and Arizona:

```
# add model that only includes my states!
southwest_state_totals_w_pred <- US_state_totals_w_pred %>% filter(Province_State == "New Mexico" | Prov

southwest_state_totals_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") +
  geom_text(aes(x = cases_per_thou, y = deaths_per_thou, label = Province_State), nudge_x = .75, nudge_
```

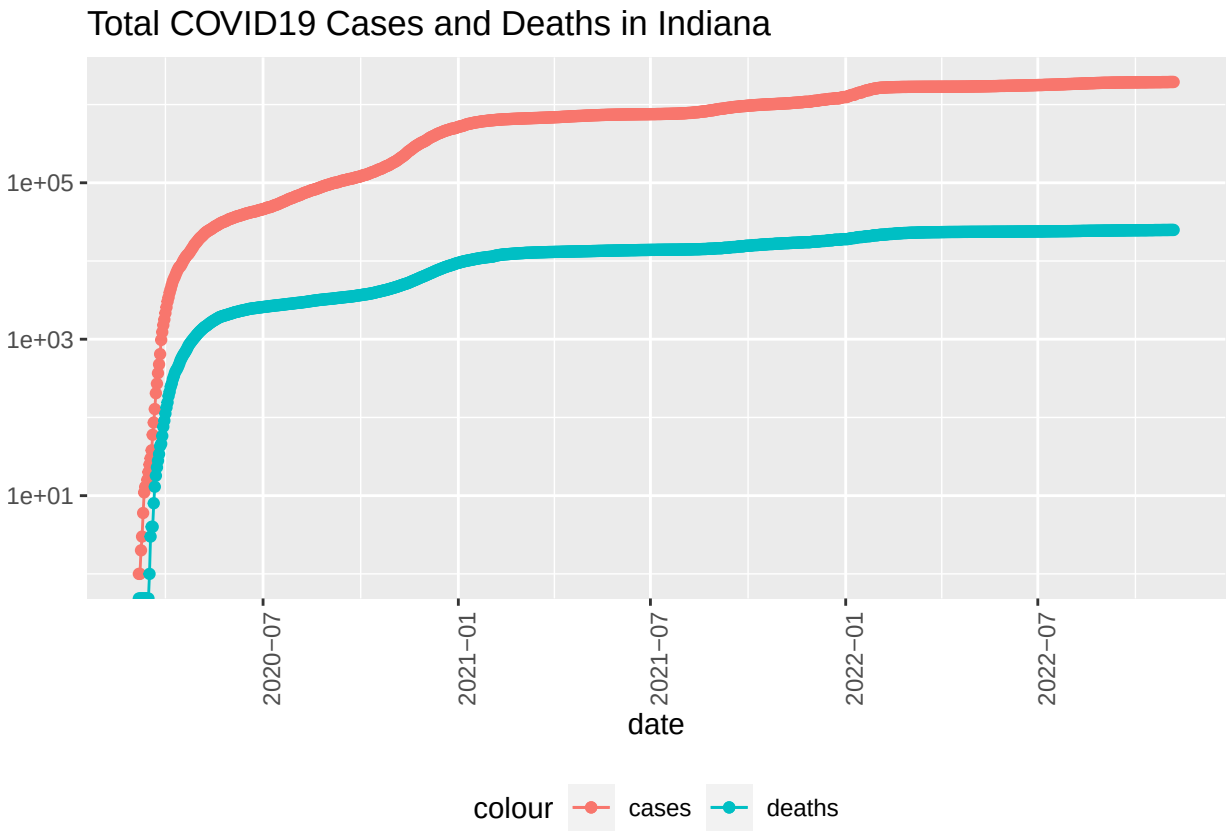


Clearly, none of these other regions show the same sort of tendency toward a clustering of death rates. The Midwestern tri-state area of Ohio, Indiana, and Michigan is unique in the way that they all exceed predicted death rates.

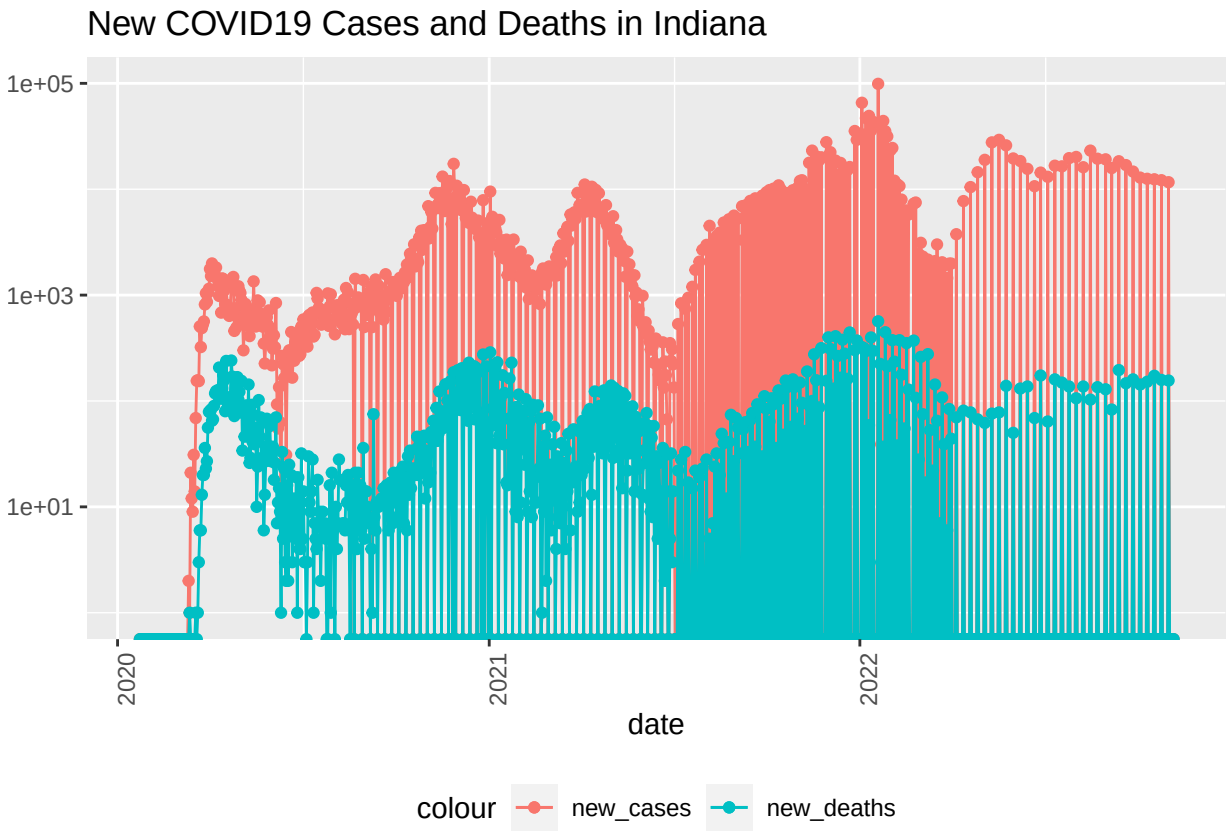
State Level Data for Indiana

Looking at the states individually we see that, despite the high death rates, each follows the relative national trends in deaths and cases relatively closely. First, we have total and new deaths and cases for Indiana.

```
US_by_state %>%
  filter(Province_State == "Indiana") %>%
  filter(cases > 0 ) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Total COVID19 Cases and Deaths in Indiana", y=NULL)
```



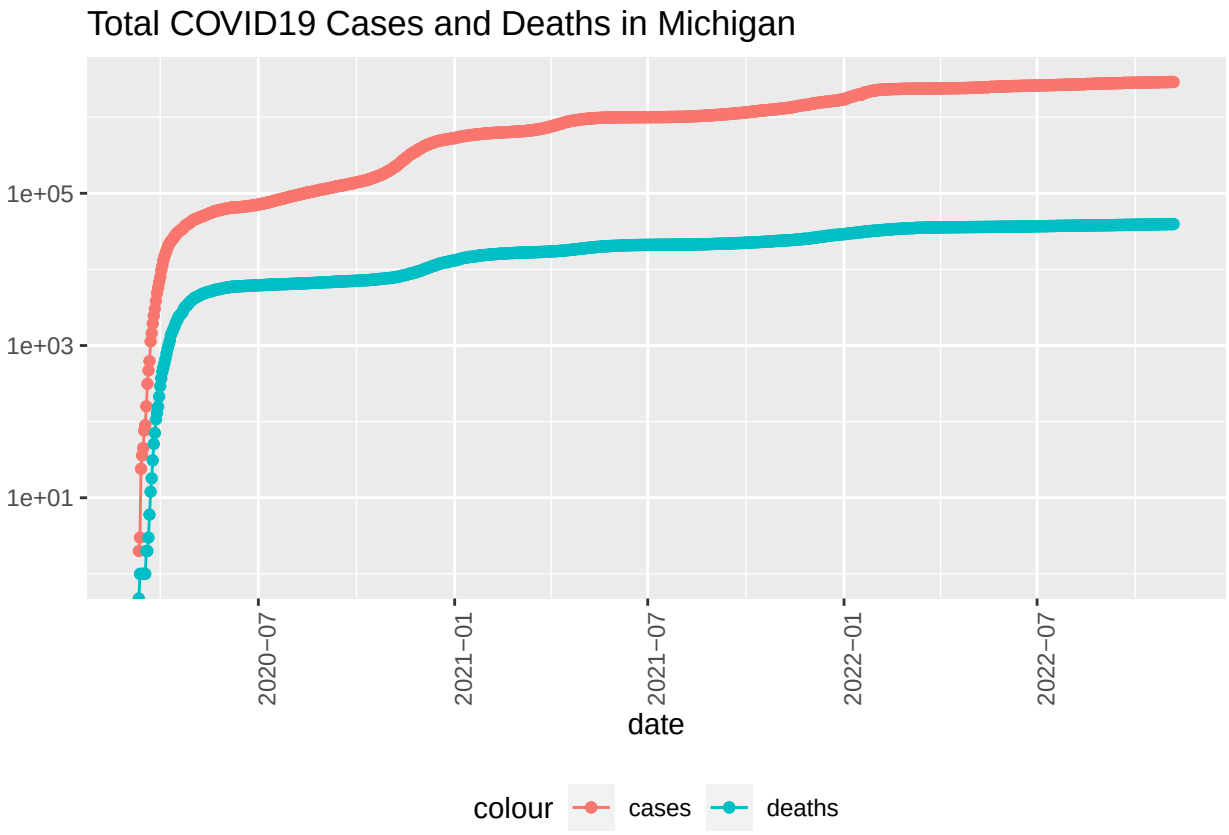

```
US_by_state %>%
  filter(Province_State == "Michigan") %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New COVID19 Cases and Deaths in Indiana", y=NULL)
```



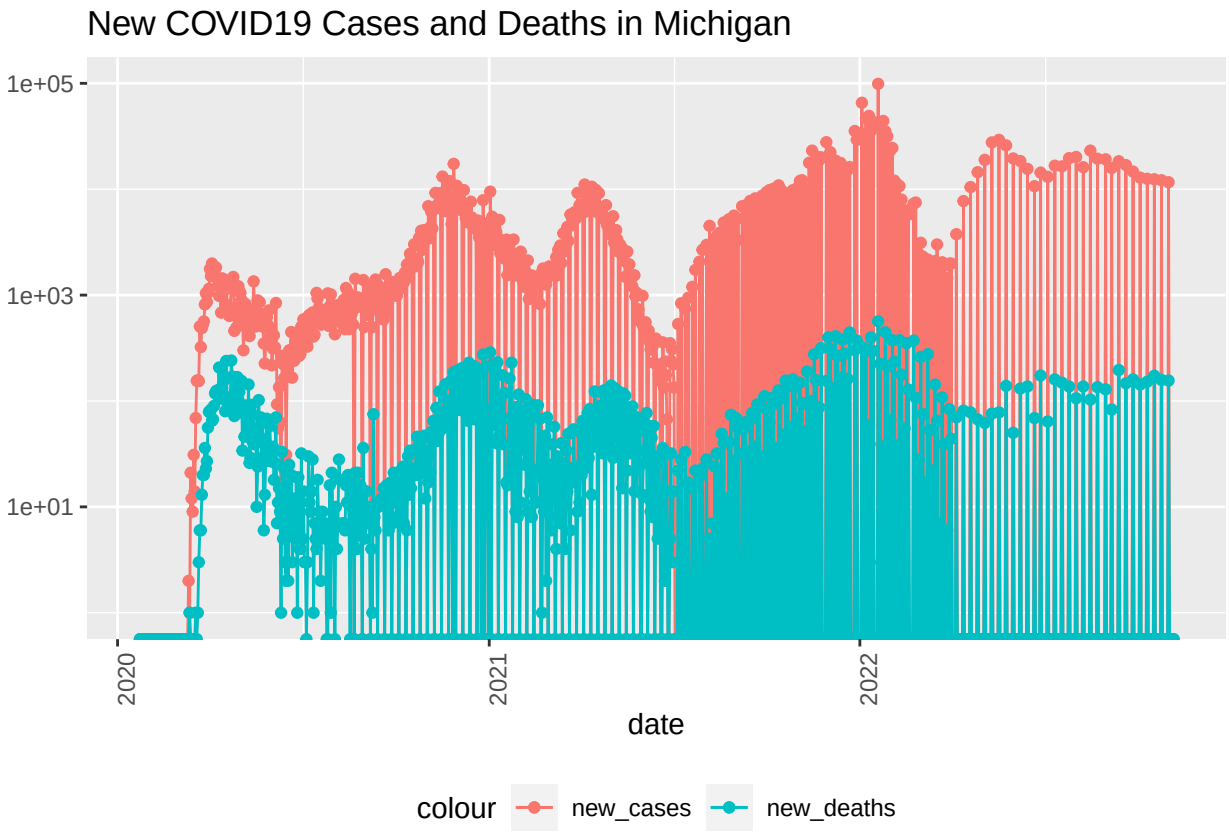
State Level Data for Michigan

The same visualizations for Michigan show much the same pattern.

```
# graphs for Michigan
US_by_state %>%
  filter(Province_State == "Michigan") %>%
  filter(cases > 0 ) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Total COVID19 Cases and Deaths in Michigan", y=NULL)
```



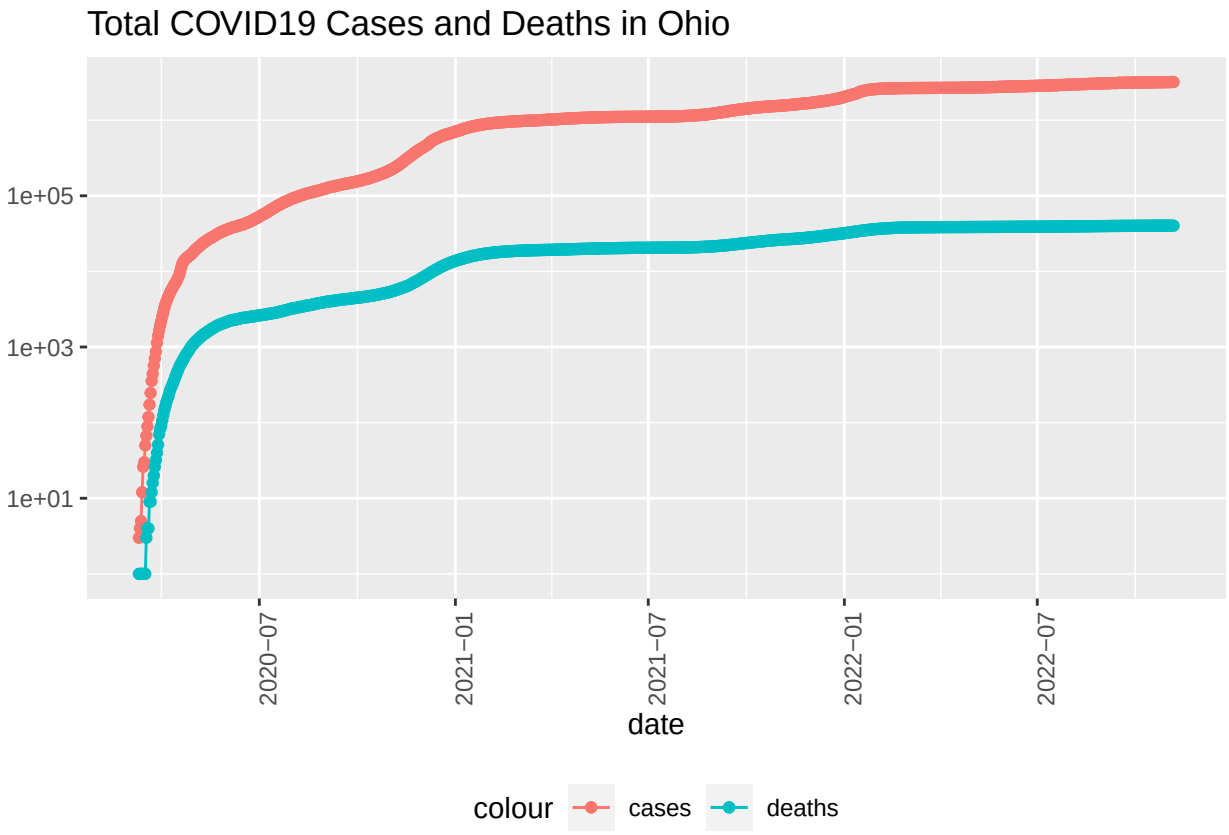

```
US_by_state %>%
  filter(Province_State == "Michigan") %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New COVID19 Cases and Deaths in Michigan", y=NULL)
```



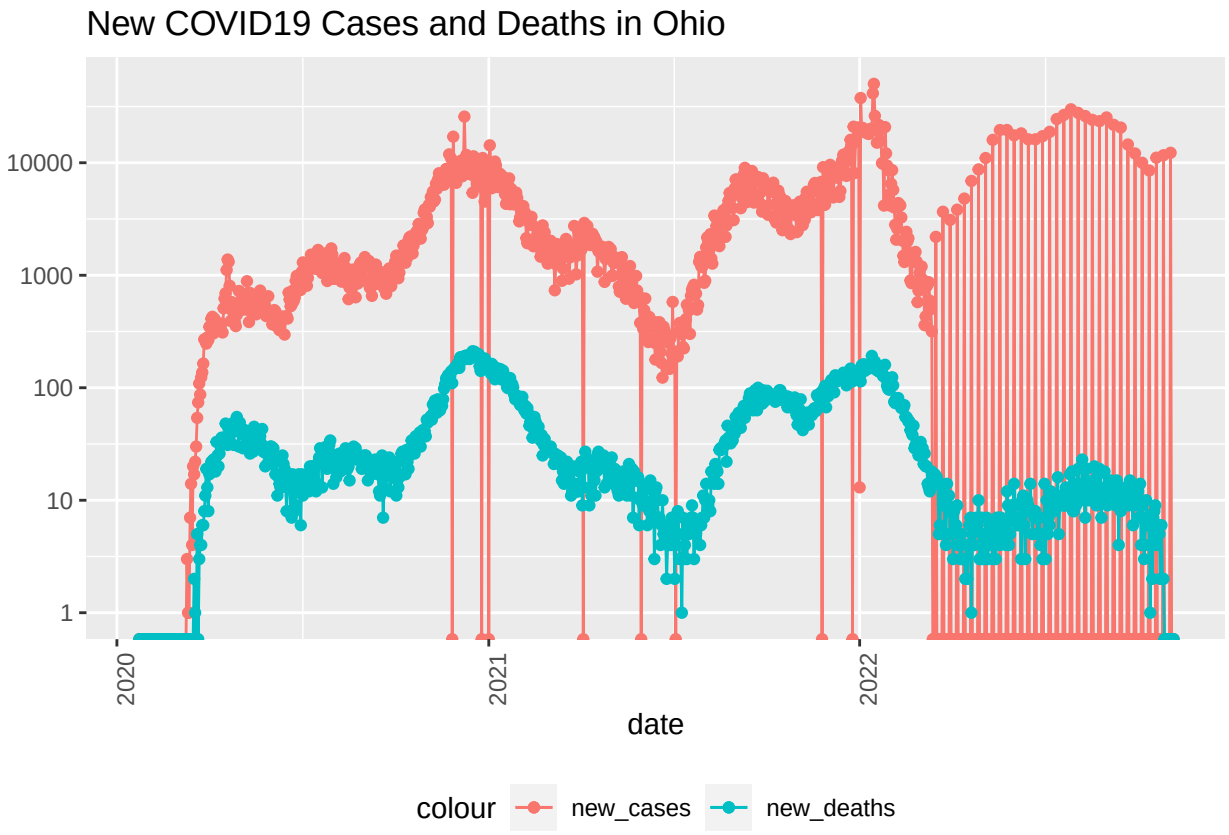
State Level Data for Ohio

Finally, Ohio follows our pattern as well.

```
US_by_state %>%
  filter(Province_State == "Ohio") %>%
  filter(cases > 0 ) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Total COVID19 Cases and Deaths in Ohio", y=NULL)
```




```
US_by_state %>%
  filter(Province_State == "Ohio") %>%
  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 = "New COVID19 Cases and Deaths in Ohio", y=NULL)
```



Biases

The collected data has several possible sources of bias. We can see that the states vary dramatically in frequency of dates with 0 new cases. This may be the result of a lack of regular reporting. This irregularity may be a confounding factor in other respects as well.

The collected data also only reflects reported cases and deaths. The federal government no longer mandates data reporting in the same way as early in the pandemic. This may lead us to believe that more recent data is less reliable.

Conclusion

We can see that while Indiana, Michigan, and Ohio follow, in general, the same general trends in case numbers and deaths as the rest of the United States, the death rate is higher than we would expect based on our model. This may point to issues in regional healthcare infrastructure or regional cultural make up that leads to distinctive lifestyle factors.

Suggestions for Further Research

Additional research ought to be conducted to analyse the relationship in this region between COVID deaths and other factors to determine which, if any, are playing a decisive role in these outcomes.

This is imperative to prevent possible deaths as we continue to deal with the ongoing COVID situation and for any possible future pandemics or epidemics that may effect the region.