# John Hopkins COVID-19 Project

student

7/22/2021

**John Hopkins COVID-19:**

This is a breakdown of COVID-19 data from John Hopkins github.com site. (source: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series)

**Loading R Packages:**

```
## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.0.4

## Warning: package 'tidyr' was built under R version 4.0.4

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'purrr' was built under R version 4.0.3

## Warning: package 'dplyr' was built under R version 4.0.4

## Warning: package 'stringr' was built under R version 4.0.3

## Warning: package 'forcats' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Warning: package 'lubridate' was built under R version 4.0.5

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

**Reading-in: Data:**

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov

file_names <- c("time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_global.csv",
                "time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")

urls <- str_c(url_in, file_names)

global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
US_cases <- read_csv(urls[3])
US_deaths <- read_csv(urls[4])
```

**Summary of Data:**

```
head(US_cases)
```

```
## # A tibble: 6 x 562
##        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 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # ... with 553 more variables: Long_ <dbl>, Combined_Key <chr>, 1/22/20 <dbl>,
## #   1/23/20 <dbl>, 1/24/20 <dbl>, 1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>,
## #   1/28/20 <dbl>, 1/29/20 <dbl>, 1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>,
## #   2/2/20 <dbl>, 2/3/20 <dbl>, 2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>,
## #   2/7/20 <dbl>, 2/8/20 <dbl>, 2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>,
## #   2/12/20 <dbl>, 2/13/20 <dbl>, 2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>,
## #   2/22/20 <dbl>, 2/23/20 <dbl>, 2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>,
## #   2/27/20 <dbl>, 2/28/20 <dbl>, 2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>,
## #   3/3/20 <dbl>, 3/4/20 <dbl>, 3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>,
## #   3/8/20 <dbl>, 3/9/20 <dbl>, 3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>,
## #   3/13/20 <dbl>, 3/14/20 <dbl>, 3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>,
## #   3/18/20 <dbl>, 3/19/20 <dbl>, 3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>,
## #   3/23/20 <dbl>, 3/24/20 <dbl>, 3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>,
## #   3/28/20 <dbl>, 3/29/20 <dbl>, 3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>,
## #   4/2/20 <dbl>, 4/3/20 <dbl>, 4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>,
## #   4/7/20 <dbl>, 4/8/20 <dbl>, 4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>,
## #   4/12/20 <dbl>, 4/13/20 <dbl>, 4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>,
## #   4/17/20 <dbl>, 4/18/20 <dbl>, 4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>,
## #   4/22/20 <dbl>, 4/23/20 <dbl>, 4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>,
## #   4/27/20 <dbl>, 4/28/20 <dbl>, ...
```

```r
head(US_deaths)
```

```
## # A tibble: 6 x 563
##        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 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # ... with 554 more variables: Long_ <dbl>, Combined_Key <chr>,
## #   Population <dbl>, 1/22/20 <dbl>, 1/23/20 <dbl>, 1/24/20 <dbl>,
## #   1/25/20 <dbl>, 1/26/20 <dbl>, 1/27/20 <dbl>, 1/28/20 <dbl>, 1/29/20 <dbl>,
## #   1/30/20 <dbl>, 1/31/20 <dbl>, 2/1/20 <dbl>, 2/2/20 <dbl>, 2/3/20 <dbl>,
## #   2/4/20 <dbl>, 2/5/20 <dbl>, 2/6/20 <dbl>, 2/7/20 <dbl>, 2/8/20 <dbl>,
## #   2/9/20 <dbl>, 2/10/20 <dbl>, 2/11/20 <dbl>, 2/12/20 <dbl>, 2/13/20 <dbl>,
## #   2/14/20 <dbl>, 2/15/20 <dbl>, 2/16/20 <dbl>, 2/17/20 <dbl>, 2/18/20 <dbl>,
## #   2/19/20 <dbl>, 2/20/20 <dbl>, 2/21/20 <dbl>, 2/22/20 <dbl>, 2/23/20 <dbl>,
## #   2/24/20 <dbl>, 2/25/20 <dbl>, 2/26/20 <dbl>, 2/27/20 <dbl>, 2/28/20 <dbl>,
## #   2/29/20 <dbl>, 3/1/20 <dbl>, 3/2/20 <dbl>, 3/3/20 <dbl>, 3/4/20 <dbl>,
## #   3/5/20 <dbl>, 3/6/20 <dbl>, 3/7/20 <dbl>, 3/8/20 <dbl>, 3/9/20 <dbl>,
## #   3/10/20 <dbl>, 3/11/20 <dbl>, 3/12/20 <dbl>, 3/13/20 <dbl>, 3/14/20 <dbl>,
## #   3/15/20 <dbl>, 3/16/20 <dbl>, 3/17/20 <dbl>, 3/18/20 <dbl>, 3/19/20 <dbl>,
## #   3/20/20 <dbl>, 3/21/20 <dbl>, 3/22/20 <dbl>, 3/23/20 <dbl>, 3/24/20 <dbl>,
## #   3/25/20 <dbl>, 3/26/20 <dbl>, 3/27/20 <dbl>, 3/28/20 <dbl>, 3/29/20 <dbl>,
## #   3/30/20 <dbl>, 3/31/20 <dbl>, 4/1/20 <dbl>, 4/2/20 <dbl>, 4/3/20 <dbl>,
## #   4/4/20 <dbl>, 4/5/20 <dbl>, 4/6/20 <dbl>, 4/7/20 <dbl>, 4/8/20 <dbl>,
## #   4/9/20 <dbl>, 4/10/20 <dbl>, 4/11/20 <dbl>, 4/12/20 <dbl>, 4/13/20 <dbl>,
## #   4/14/20 <dbl>, 4/15/20 <dbl>, 4/16/20 <dbl>, 4/17/20 <dbl>, 4/18/20 <dbl>,
## #   4/19/20 <dbl>, 4/20/20 <dbl>, 4/21/20 <dbl>, 4/22/20 <dbl>, 4/23/20 <dbl>,
## #   4/24/20 <dbl>, 4/25/20 <dbl>, 4/26/20 <dbl>, 4/27/20 <dbl>, ...
```

```r
# Top Fields (short list):
# * Province_State: US State
# * Country_Region: Country Part
# * Last_Update: Date of last update
# * Lat: Global Coordinates
# * Long: Global Coordinates
# * Confirmed: Number of Cases
# * Deaths: Number of deaths
# * Recovered: Number of recovered
# * Active: Number of active cases
# * Incident_Rate: Incidents of cases
# * Total_Test_Results: Number of tests (have been)
# * People_Hospitalized: Number of people need to be put in hospital
```

**# TIDY DATA (a wee bit):**

```r
global_cases <- global_cases %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long), names_to = "date", values_to =
```

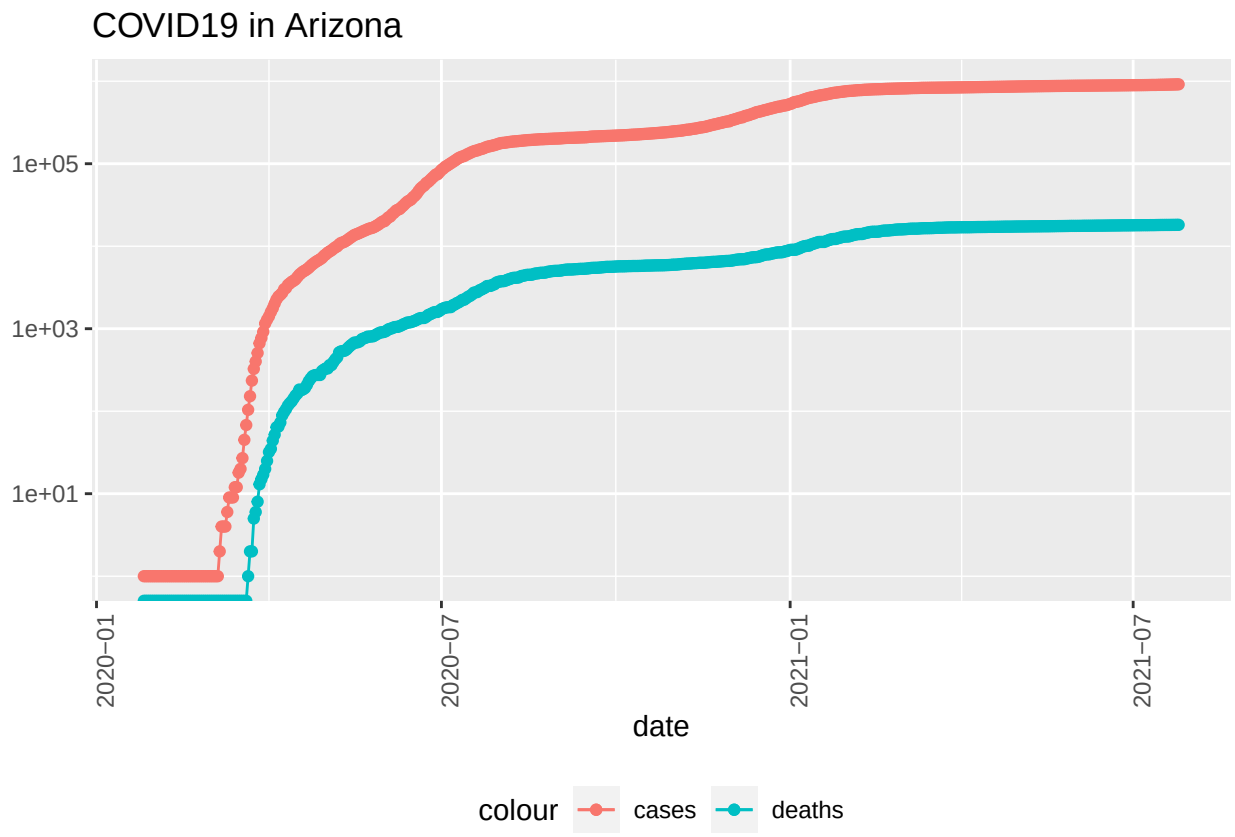
```
  select (-c(Lat, Long))

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

global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = 'Country/Region', Province_State = 'Province/State') %>%
  mutate(date = mdy(date))

global <- global %>% filter(cases > 0)

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)

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

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

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

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)
```

**Visualization Prep:**

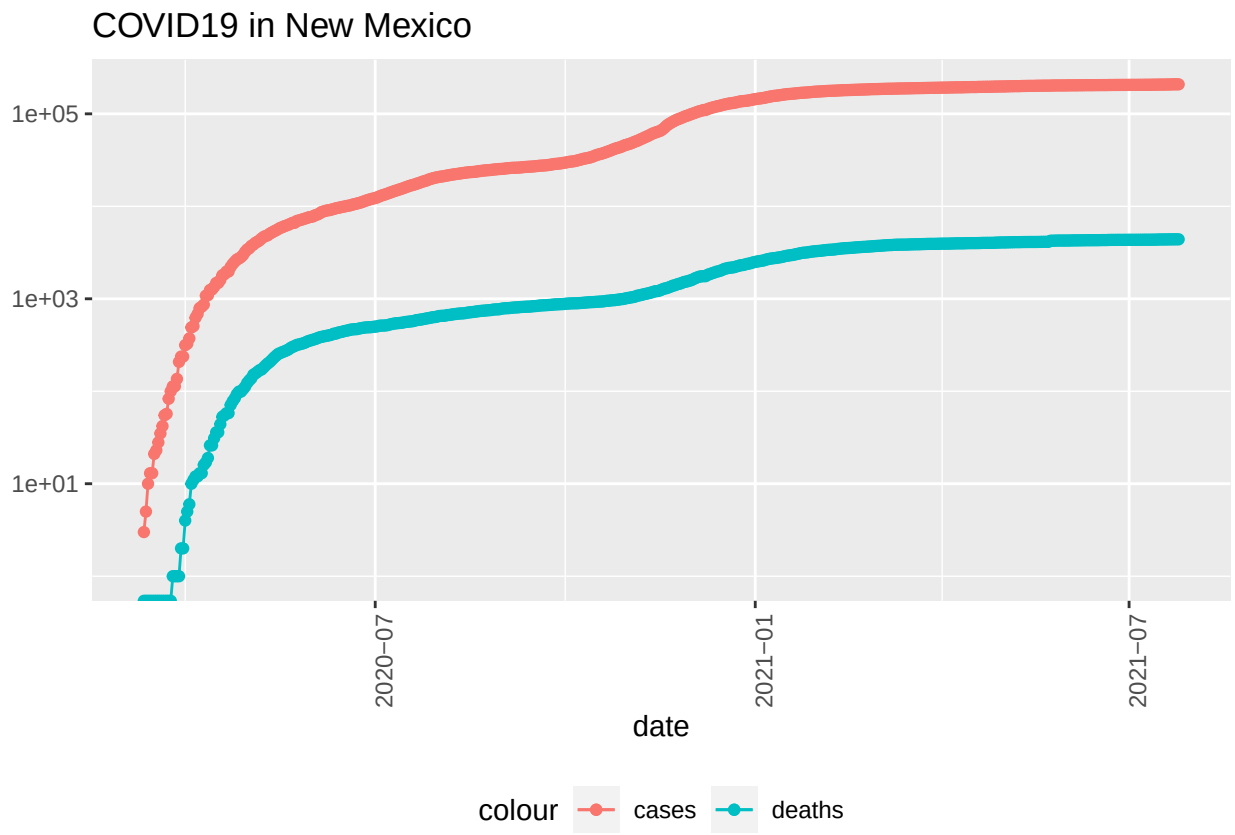
```
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths *1000000 / Population)  %>%
  select(Province_State, Country_Region, date,
         cases,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

US_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()
```

**Three Visualizations:**

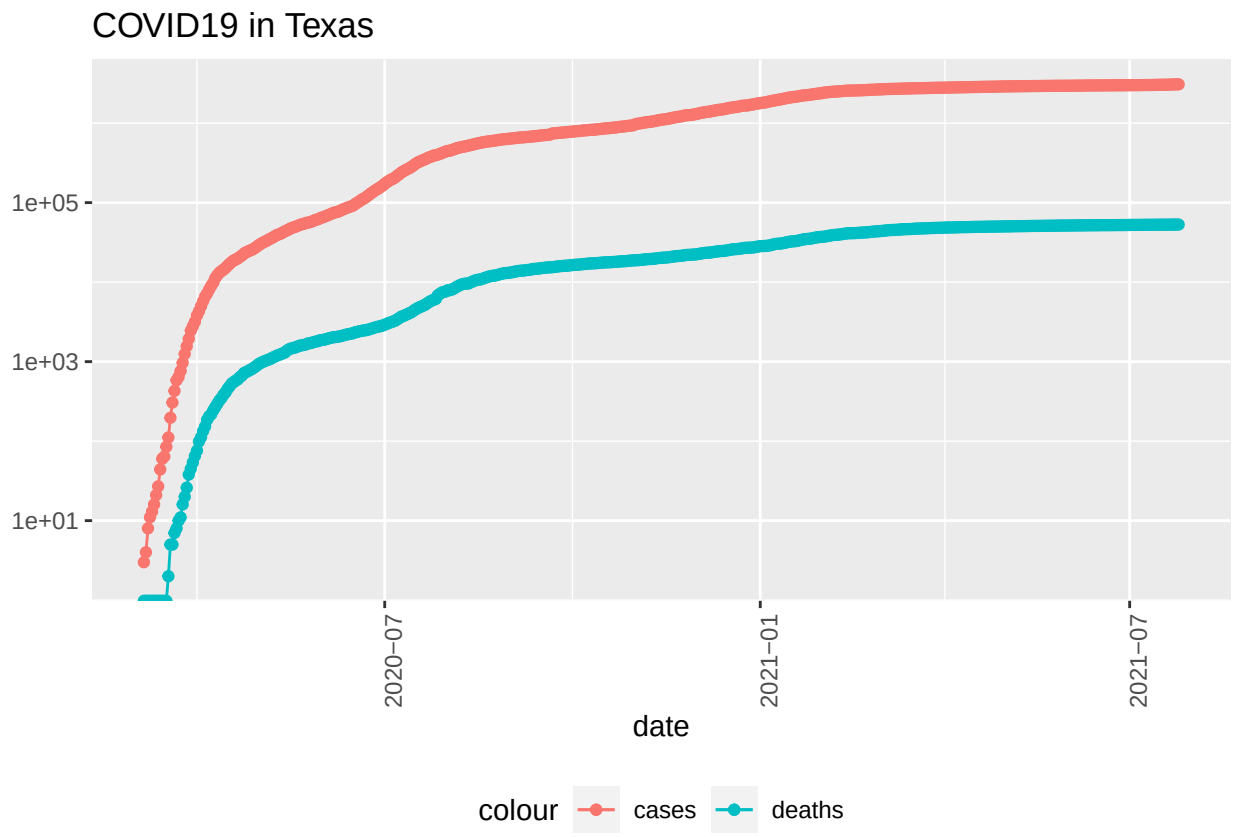
```
state <- "Arizona"
US_by_state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID19 in " , state), y = NULL)
```



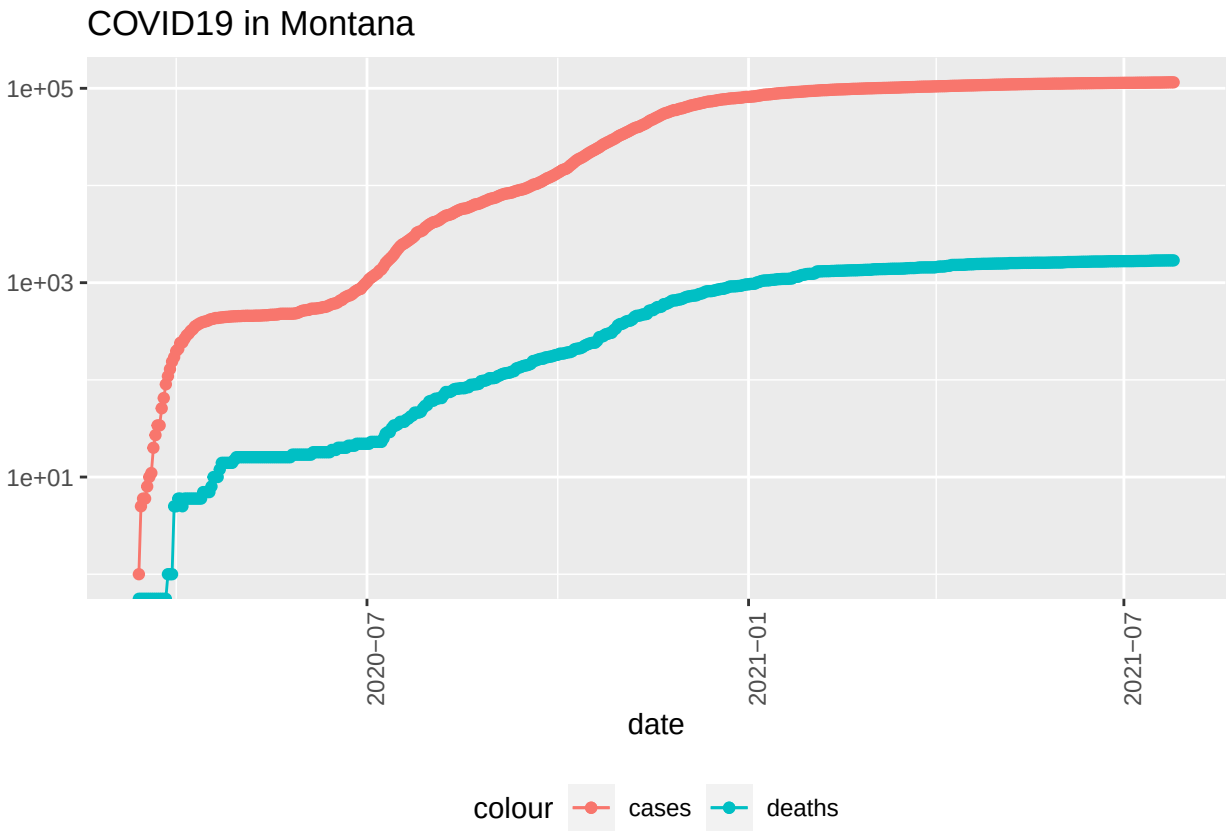

```
state <- "New Mexico"
US_by_state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID19 in " , state), y = NULL)
```



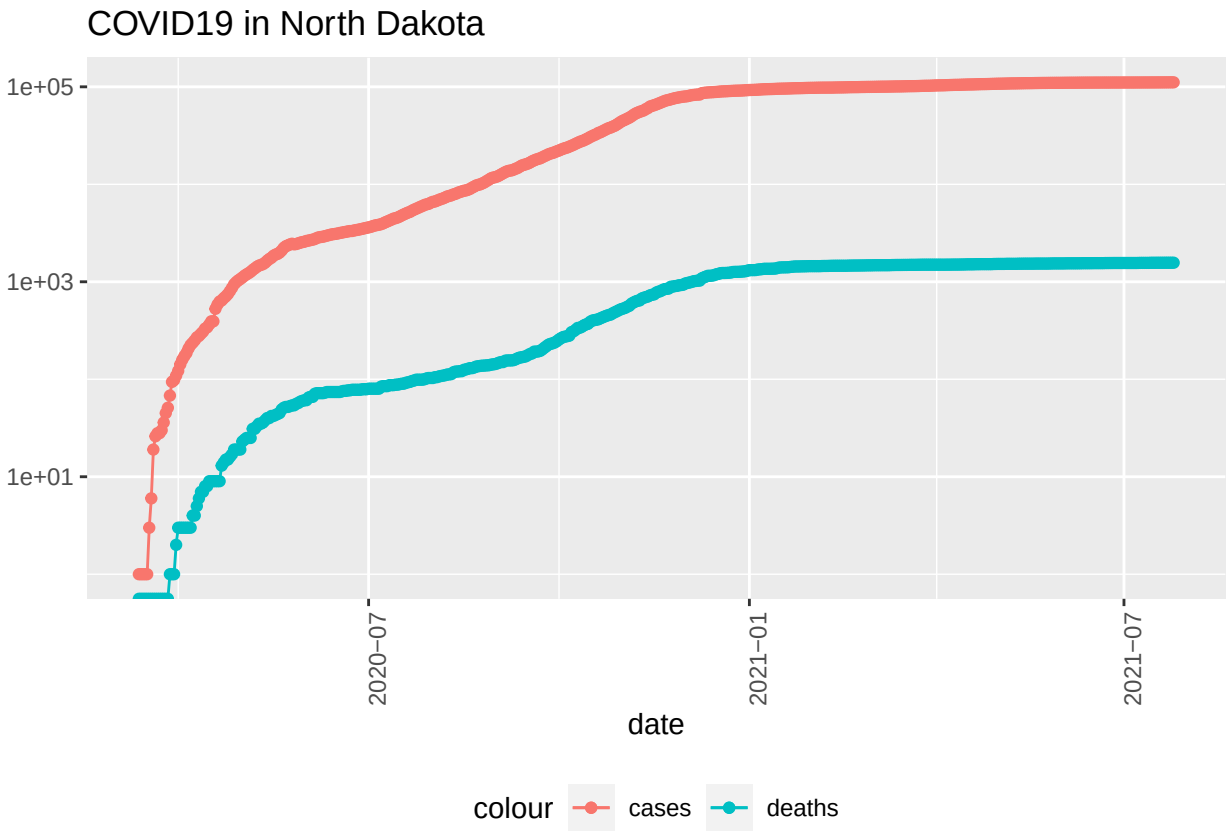

```
state <- "Texas"
US_by_state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID19 in " , state), y = NULL)
```



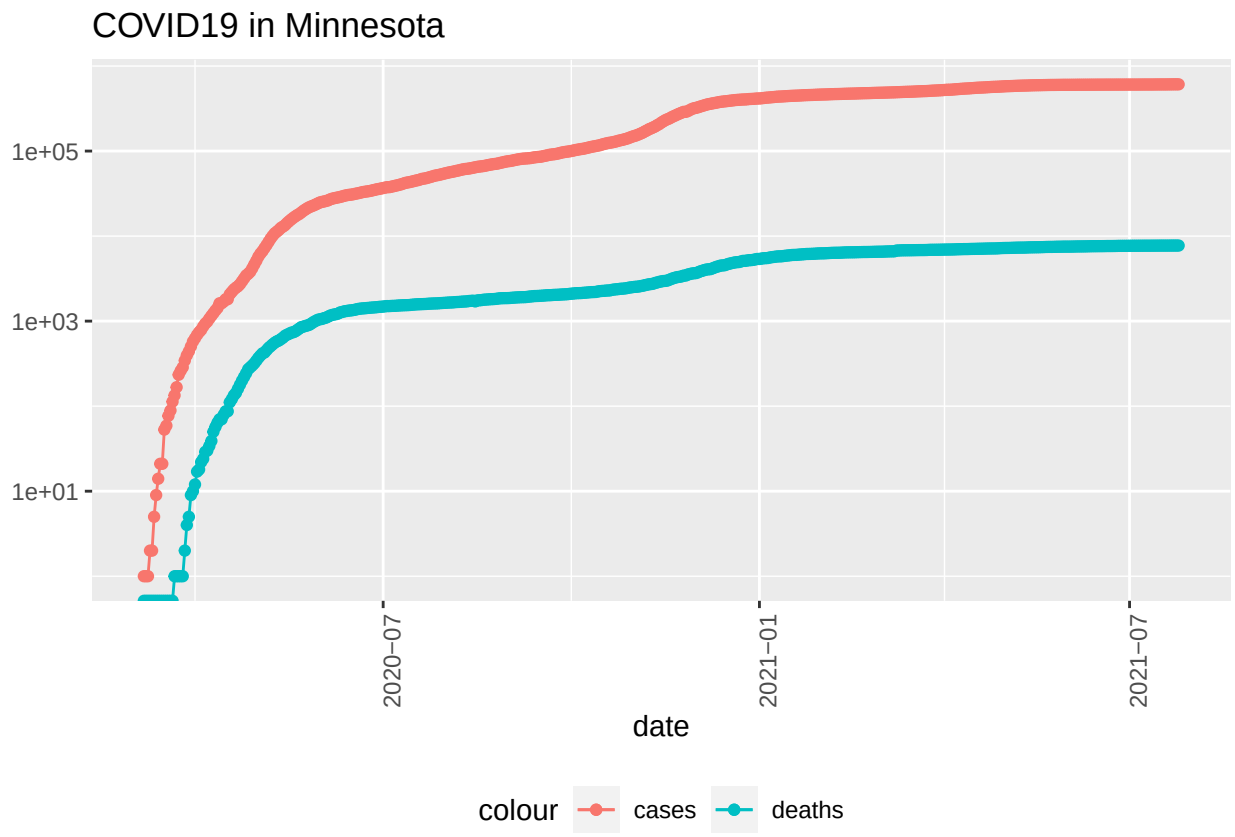

```
state <- "Montana"
US_by_state %>%
  filter(Province_State == state) %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID19 in " , state), y = NULL)
```




```
#-------------------------------------------------------------------------------
state <- "North Dakota"
US_by_state %>%
  filter(Province_State == state) %>%
  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 = str_c("COVID19 in " , state), y = NULL)
```




```
#------------------------------------------------------------------------------
state <- "Minnesota"
US_by_state %>%
  filter(Province_State == state) %>%
  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 = str_c("COVID19 in " , state), y = NULL)
```




### MODEL:

```
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)


#-------------------------------------------------------------------------------
# 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          0.0363           3.32 Northern Mariana Isl~     2    183      55144
##  2          0.326           40.8  Virgin Islands           35   4375     107268
##  3          0.374           28.7  Hawaii                  529  40659    1415872
##  4          0.415           39.5  Vermont                 259  24676     623989
##  5          0.518           99.6  Alaska                  384  73818     740995
```

```
##  6          0.667           52.0  Maine                  897  69905    1344212
##  7          0.672           50.9  Oregon                2836 214869    4217737
##  8          0.684           38.4  Puerto Rico           2568 144140    3754939
##  9          0.756          133.   Utah                  2425 426418    3205958
## 10          0.798           61.2  Washington            6078 466235    7614893
```

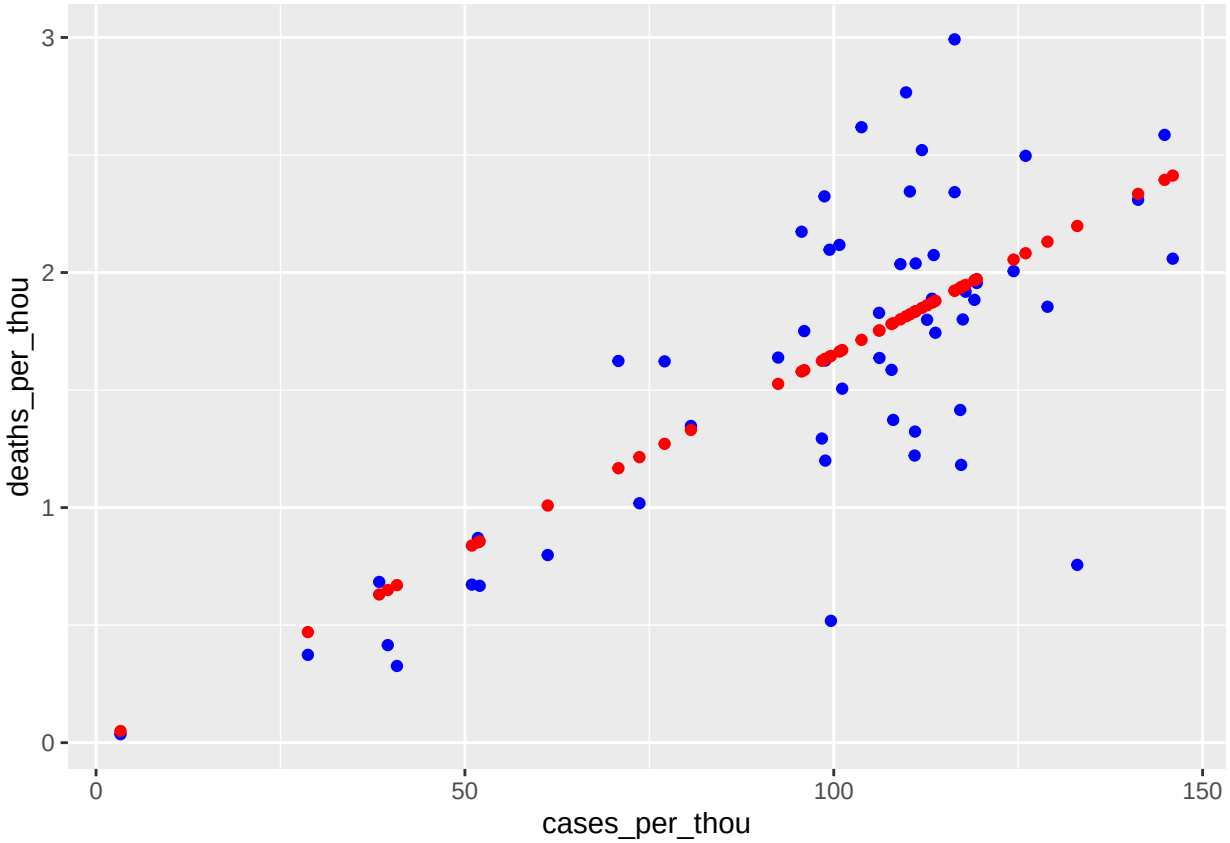
```
#-------------------------------------------------------------------------------
# Worst states:
#-------------------------------------------------------------------------------
US_state_totals %>%
  slice_max(deaths_per_thou, n = 10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State deaths   cases population
##              <dbl>          <dbl> <chr>           <dbl>   <dbl>      <dbl>
##  1            2.99          116.  New Jersey      26575 1033712    8882190
##  2            2.77          110.  New York        53813 2136032   19453561
##  3            2.62          104.  Massachusetts   18046  715180    6892503
##  4            2.59          145.  Rhode Island     2739  153447    1059361
##  5            2.52          112.  Mississippi      7502  333180    2976149
##  6            2.50          126.  Arizona         18171  917168    7278717
##  7            2.34          110.  Louisiana       10900  512843    4648794
##  8            2.34          116.  Alabama         11483  570667    4903185
##  9            2.32           98.7 Connecticut      8286  352037    3565287
## 10            2.31          141.  South Dakota     2043  124960     884659
```

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

x_grid <- seq(1, 151)
new_df <- tibble(cases_per_thou = x_grid)
US_state_totals %>% mutate(pred = predict(mod))
```

```
## # A tibble: 55 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          11483 5.71e5    4903185          116.            2.34   1.92
##  2 Alaska             384 7.38e4     740995           99.6           0.518  1.64
##  3 Arizona          18171 9.17e5    7278717          126.            2.50   2.08
##  4 Arkansas          6054 3.75e5    3017804          124.            2.01   2.06
##  5 California       64235 3.91e6   39512223           98.9           1.63   1.63
##  6 Colorado          6910 5.69e5    5758736           98.9           1.20   1.63
##  7 Connecticut       8286 3.52e5    3565287           98.7           2.32   1.63
##  8 Delaware          1698 1.11e5     973764          114.            1.74   1.88
##  9 District of Co~   1146 5.00e4     705749           70.8           1.62   1.17
## 10 Florida          38670 2.52e6   21477737          117.            1.80   1.94
## # ... with 45 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")
```




**Plots: Snap Short of Three States Review:**

I was interested in the three upper northern states and three lower states, partly because of the climate. As you can see from the three upper states, there seems to be a difference in the onset of the COVID19 between Minnesota compared with North Dakota and Montana, which both of these states were a little bit more gradual compared to Minnesota. Minnesota had very similar onset attributes to Texas and Arizona.

**Model:**

Earlier-on in the model, you can see that the predictions are quite accurate, however as time goes on the spread becomes more defined, also outliers are seen readily.

**Conclusion:**

Conclusion of the John Hopkins (JH) COVID19 data: As expected, the data has many facets that can be shown in various ways. Though, an impressive amount of data, one needs to question if all variables are accounted for some real understanding of what makes up the details of COVID19. Tracking and understanding the data is shown here but I found myself questioning what this was really telling me. Obviously, you can see that the start of COVID19 was very fast, and rather maintained a consistent pattern, which holds true throughout the period examined. In regards to properties that would be interesting in the future to study would be the correlation between: mask wearing, school interactions (of all grades and classification), public gatherings and such said items that would provide depth.

**Bias:**

This entire project is very much subject to all sorts of biasness. The data collection methods need to be examined to see if data was indeed collected correctly, personal bias from the collectors need to be accounted for as well. The nature of COVID19 leads itself also to people not reporting said issues, affects, interactions, as well as wrong doings.