Final Project Part 2: COVID-19 Data Analysis

2023-06-26

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

Covid-19 Historical Data Report

In this report, we will be breaking down COVID data from 2019 until 2023, analyzing COVID-19 case rates and death rates from various regions. Our objective is to answer the following questions:

1. What regions were most or least affected by COVID-19, with respect to cases and deaths?
2. How does the US compare with the rest of the world with respect to death rates as a result of COVID-19?

Get Data

```
url_in1 <-
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master"
url_in2 <-
  "/csse_covid_19_data/csse_covid_19_time_series/"

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_in1,url_in2,file_names)

uid_lookup <-
  "UID_ISO_FIPS_LookUp_Table.csv"

uid_lookup_url <- str_c(url_in1,
                        "/csse_covid_19_data/",
                        uid_lookup)
```

Read Data

```
global_cases <- read_csv(urls[2])
global_deaths <- read_csv(urls[4])
us_cases <- read_csv(urls[1])
```

```
us_deaths <- read_csv(urls[3])

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

Tidy Global Data

We will be tidying the global data. All columns are of interest, we will simply be pivoting the dates to a column in order to get accurate case counts, and appending the case and death documents to each other. We will also be summarizing total cases/deaths per country.

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

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

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)

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

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, Lat, Long) %>%
  rename(region = Country_Region)

global_totals <- global %>% group_by(region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000/Population) %>%
  select(region, date, cases,deaths,deaths_per_mill,Population) %>%
  ungroup()

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

global_by_region <- global_totals %>% group_by(region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
```

```
  mutate(deaths_per_mill = deaths * 1000000/Population) %>%
  select(region, date, cases,deaths,deaths_per_mill,Population) %>%
  ungroup()

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

global_region_totals <- global_by_region %>%
  group_by(region) %>%
  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)

summary(global_totals)
```

```
##     region              date                cases              deaths
##  Length:214113      Min.   :2020-01-22   Min.   :         1   Min.   :      0
##  Class :character   1st Qu.:2020-12-15   1st Qu.:      7504   1st Qu.:     98
##  Mode  :character   Median :2021-09-18   Median :     71705   Median :   1061
##                     Mean   :2021-09-13   Mean   :   1480108   Mean   :  20642
##                     3rd Qu.:2022-06-16   3rd Qu.:    579110   3rd Qu.:   8357
##                     Max.   :2023-03-09   Max.   :103802702   Max.   :1123836
##
##  deaths_per_mill     Population          new_cases            new_deaths
##  Min.   :   0.00   Min.   :8.090e+02   Min.   :-103802701   Min.   :-1123836
##  1st Qu.:  20.75   1st Qu.:2.083e+06   1st Qu.:         0   1st Qu.:       0
##  Median : 183.99   Median :9.006e+06   Median :        38   Median :       0
##  Mean   : 713.88   Mean   :3.413e+07   Mean   :         1   Mean   :       0
##  3rd Qu.:1059.93   3rd Qu.:2.914e+07   3rd Qu.:       660   3rd Qu.:       7
##  Max.   :6658.38   Max.   :1.418e+09   Max.   :   1354505   Max.   :   59961
##  NA's   :5861      NA's   :5861        NA's   :1            NA's   :1
```

```
summary(global_by_region)
```

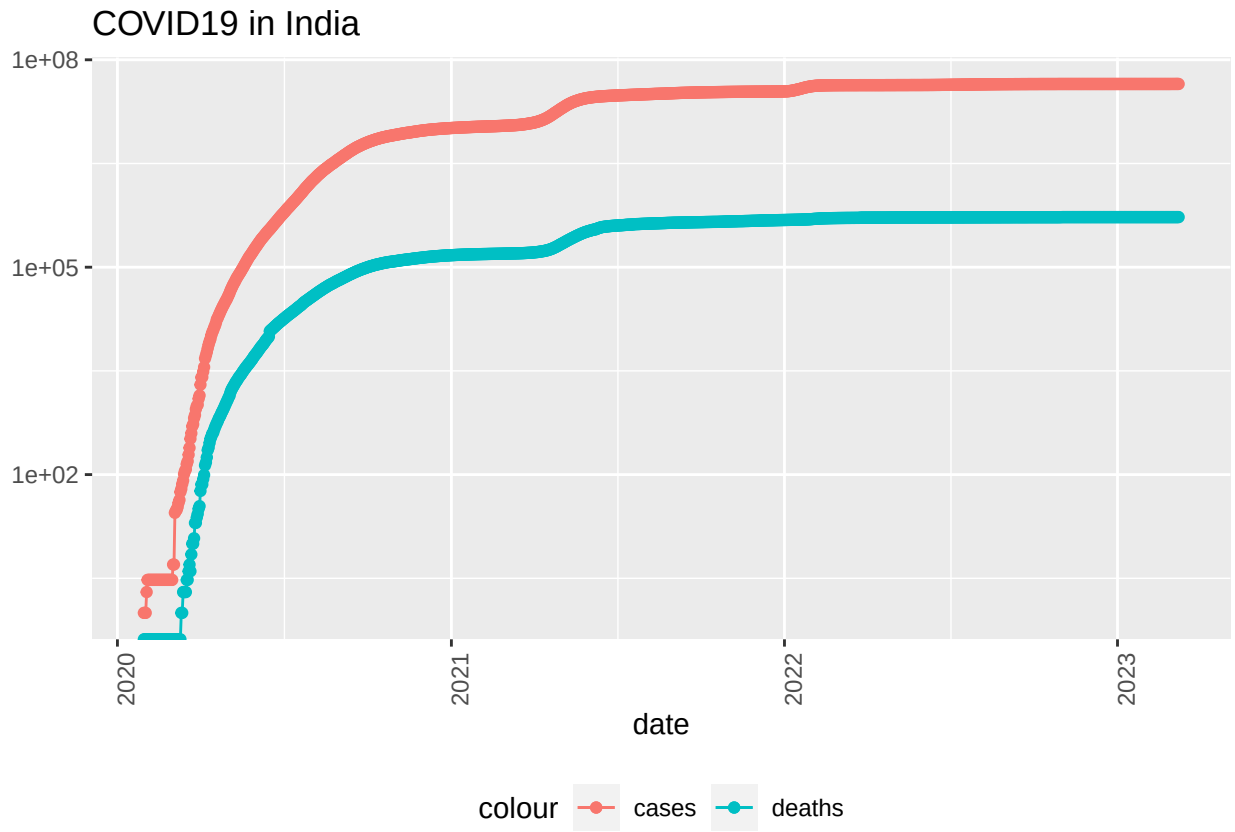
```
##     region              date                cases              deaths
##  Length:214113      Min.   :2020-01-22   Min.   :         1   Min.   :      0
##  Class :character   1st Qu.:2020-12-15   1st Qu.:      7504   1st Qu.:     98
##  Mode  :character   Median :2021-09-18   Median :     71705   Median :   1061
##                     Mean   :2021-09-13   Mean   :   1480108   Mean   :  20642
##                     3rd Qu.:2022-06-16   3rd Qu.:    579110   3rd Qu.:   8357
##                     Max.   :2023-03-09   Max.   :103802702   Max.   :1123836
##
##  deaths_per_mill     Population          new_cases            new_deaths
##  Min.   :   0.00   Min.   :8.090e+02   Min.   :-103802701   Min.   :-1123836
##  1st Qu.:  20.75   1st Qu.:2.083e+06   1st Qu.:         0   1st Qu.:       0
##  Median : 183.99   Median :9.006e+06   Median :        38   Median :       0
##  Mean   : 713.88   Mean   :3.413e+07   Mean   :         1   Mean   :       0
##  3rd Qu.:1059.93   3rd Qu.:2.914e+07   3rd Qu.:       660   3rd Qu.:       7
##  Max.   :6658.38   Max.   :1.418e+09   Max.   :   1354505   Max.   :   59961
##  NA's   :5861      NA's   :5861        NA's   :1            NA's   :1
```

Visualizing Regional Data

For a brief visual overview of our data, let's try to reduce the amount of points we look at by looking at a particular region. For example, India.

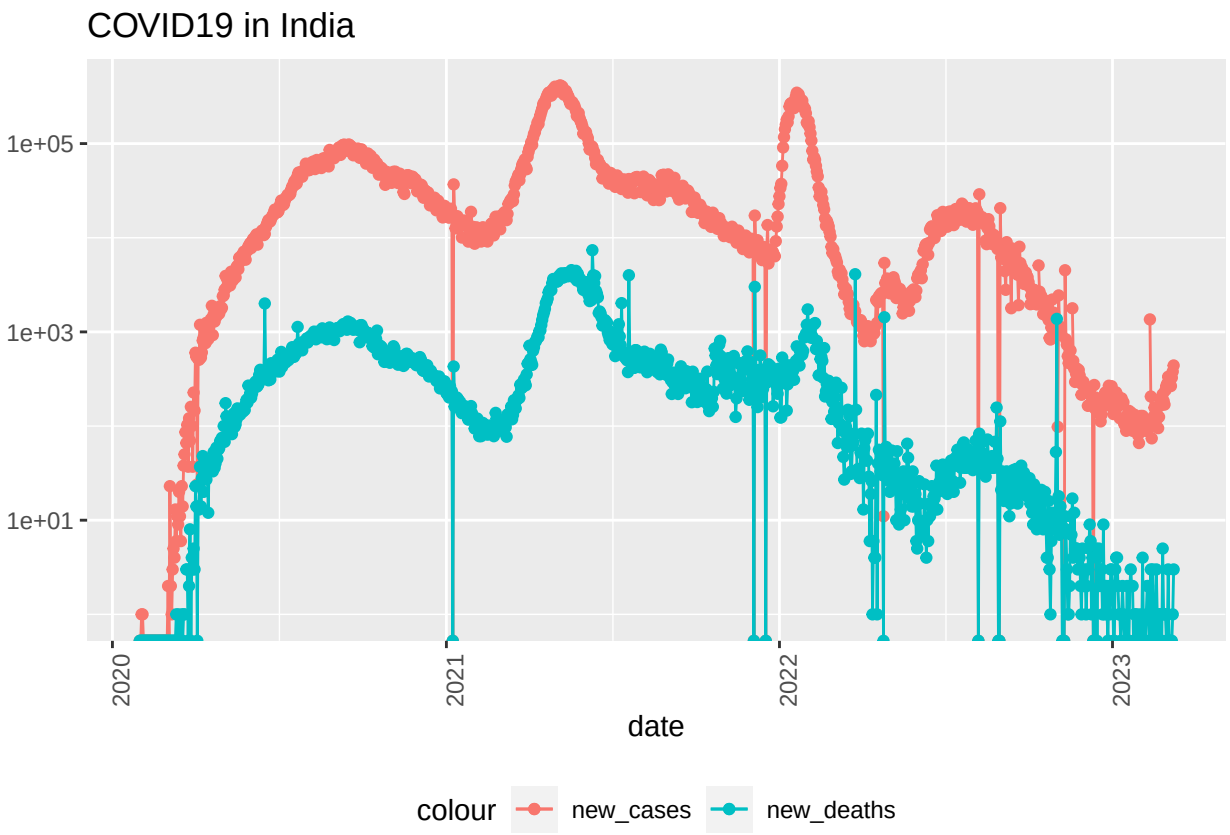
```
location <- "India"

global_by_region %>%
  filter(region == location) %>%
  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 ", location), y=NULL)
```




```
global_by_region %>%
  filter(region == location) %>%
  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("COVID19 in ", location), y=NULL)
```



Global Data (Cases)

Let's take a look at the best/worst regions with respect to global COVID-19 cases and deaths. We'll plot the 10 regions that were most/least affected by COVID-19 on a bar chart.

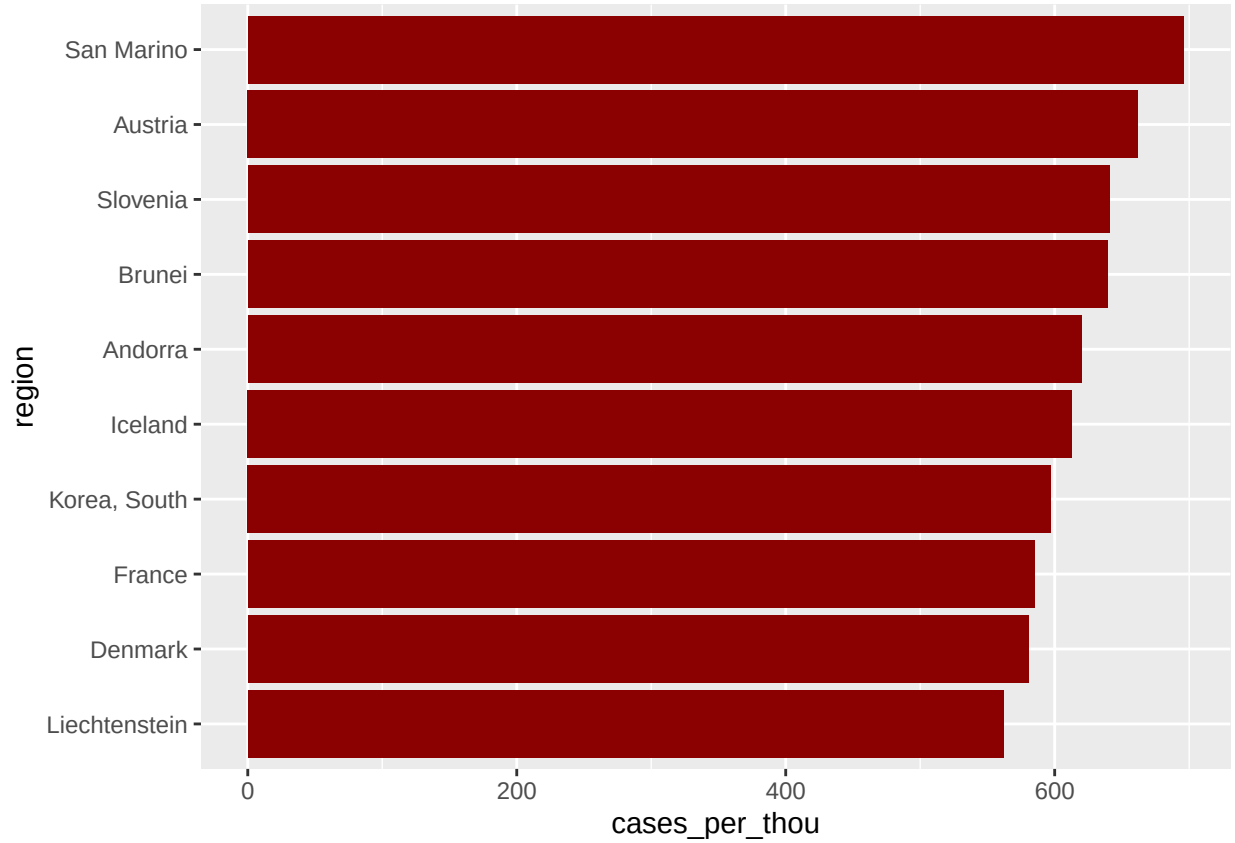
```
global_region_totals %>% slice_min(cases_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou region           deaths  cases population
##              <dbl>          <dbl> <chr>             <dbl>  <dbl>      <dbl>
##  1        0.000233      0.0000388 Korea, North          6      1   25778815
##  2        0.0130        0.393     Niger               315   9508   24206636
##  3        0.0724        0.400     Yemen              2159  11945   29825968
##  4        0.0118        0.467     Chad                194   7679   16425859
##  5        0.0142        0.718     Tanzania            846  42906   59734213
##  6        0.0158        0.973     Sierra Leone        126   7760    7976985
##  7        0.0189        1.06      Burkina Faso        396  22056   20903278
##  8        0.0163        1.07      Congo (Kinshasa)   1464  95749   89561404
##  9        0.0153        1.29      Nigeria            3155 266598  206139587
## 10        0.114         1.46      Sudan              5017  63829   43849269
```

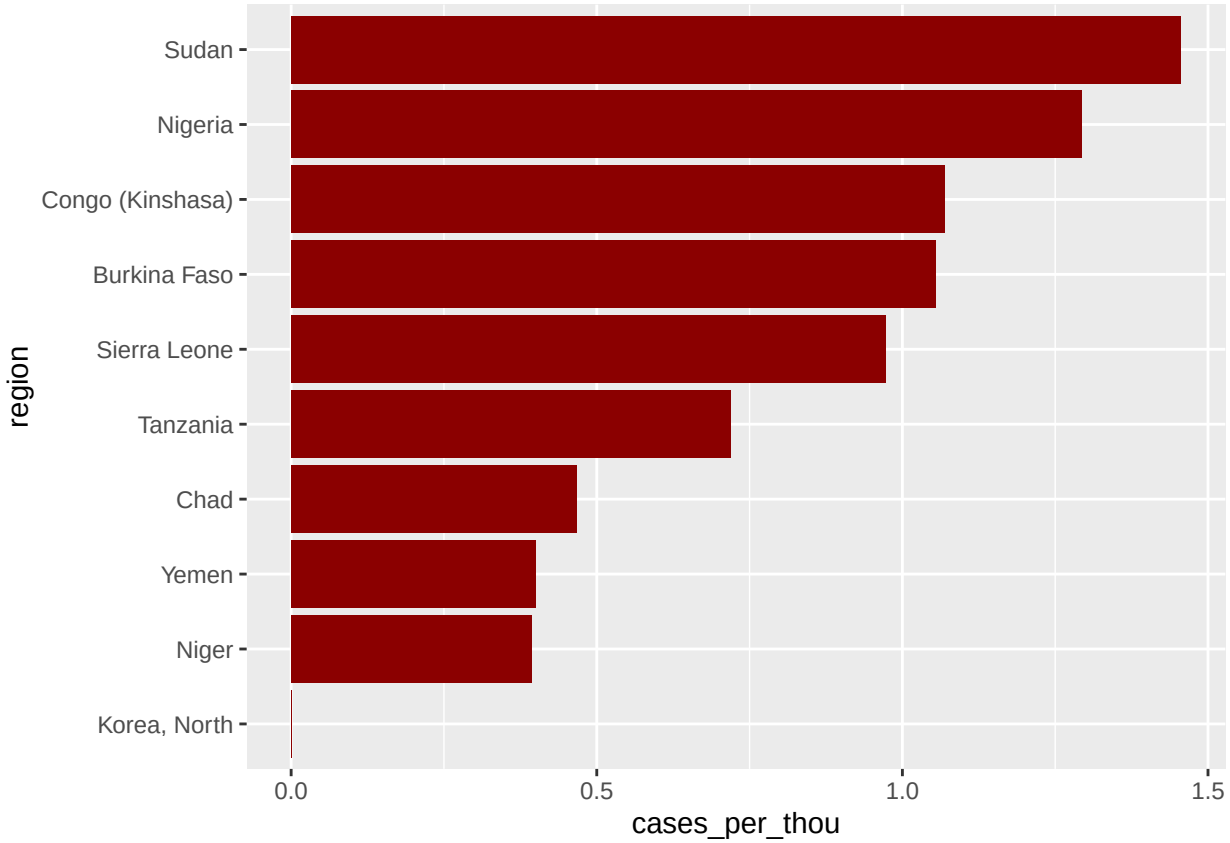
```
global_region_totals %>% slice_max(cases_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou region        deaths    cases population
##              <dbl>          <dbl> <chr>          <dbl>    <dbl>      <dbl>
##  1            3.59           696. San Marino       122    23616      33938
##  2            2.44           662. Austria        21970  5961143    9006400
##  3            3.40           641. Slovenia        7078  1331707    2078932
##  4            0.514          639. Brunei           225   279661     437483
##  5            2.14           620. Andorra          165    47890      77265
##  6            0.771          613. Iceland          263   209137     341250
##  7            0.665          597. Korea, South   34093 30615522   51269183
##  8            2.44           585. France        166176 39866718   68128061
##  9            1.40           581. Denmark         8345  3451036    5942850
## 10            2.33           562. Liechtenstein     89    21432      38137
```

```
global_region_totals %>%
  top_n(10, cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = cases_per_thou)) +
  geom_bar(stat = "identity", fill = "red4") +
  coord_flip()
```




```
global_region_totals %>%
  top_n(-10, cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = cases_per_thou)) +
  geom_bar(stat = "identity", fill = "red4") +
  coord_flip()
```


Sudan
Nigeria
Congo (Kinshasa)
Burkina Faso
Sierra Leone
Tanzania
Chad
Yemen
Niger
Korea, North
region
0.0
0.5
1.0
1.5
cases_per_thou

Global Data (Deaths)

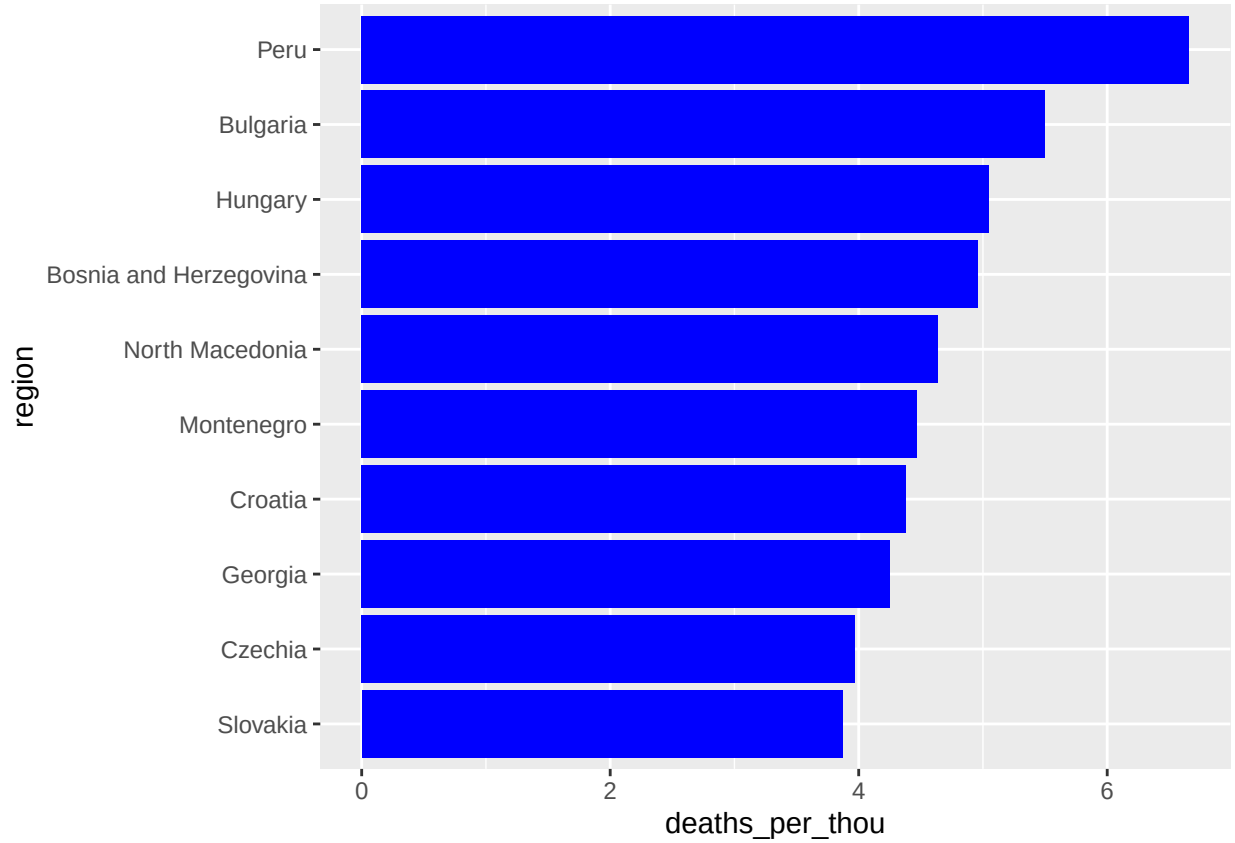
```
global_region_totals %>% slice_min(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou region       deaths cases population
##              <dbl>          <dbl> <chr>         <dbl> <dbl>      <dbl>
##  1        0            35.8       Holy See          0    29        809
##  2        0           240.        Tuvalu            0  2828      11792
##  3        0.000233      0.0000388 Korea, North      6     1   25778815
##  4        0.00320       4.51      Burundi          38 53631   11890781
##  5        0.0118        0.467     Chad            194  7679   16425859
##  6        0.0123        1.64      South Sudan     138 18368   11193729
##  7        0.0130        0.393     Niger           315  9508   24206636
##  8        0.0131        1.86      Tajikistan      125 17786    9537642
##  9        0.0134        2.31      Benin           163 27999   12123198
## 10        0.0142        0.718     Tanzania        846 42906   59734213
```

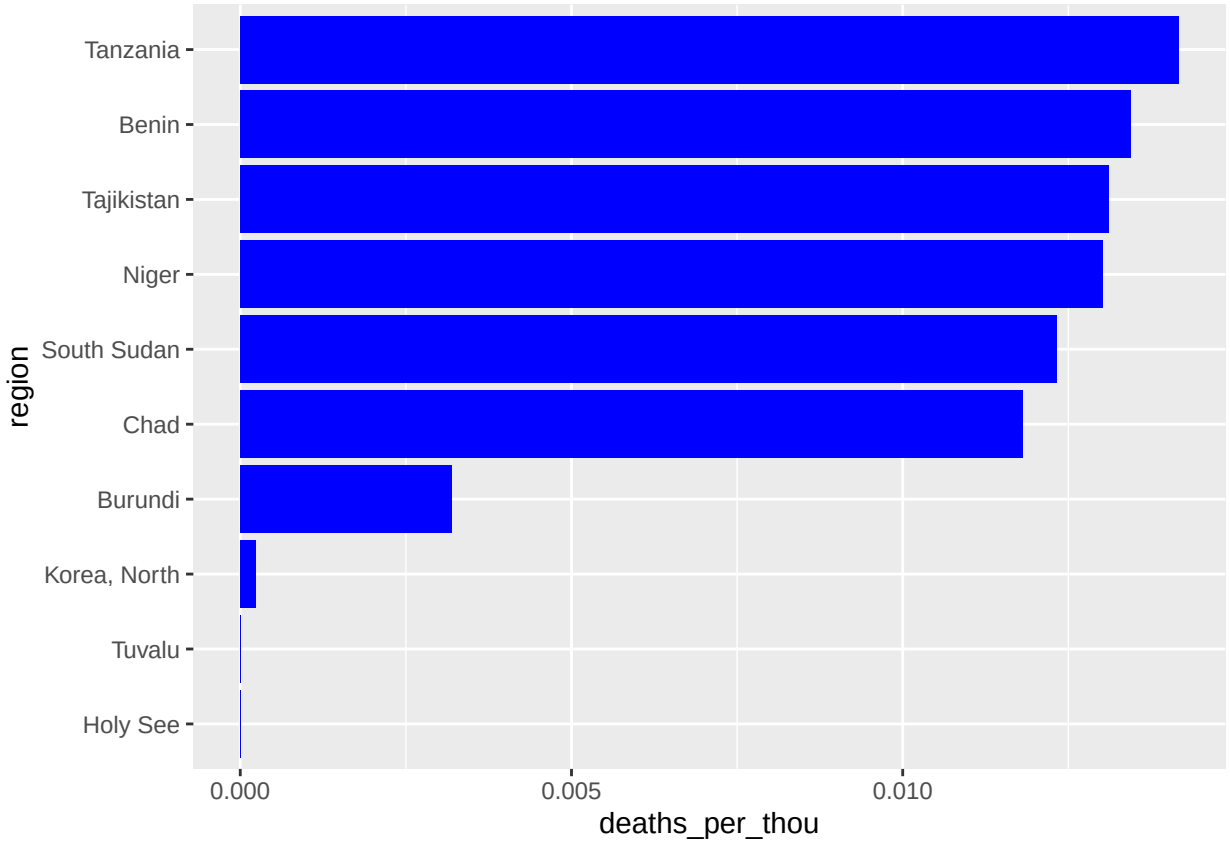
```
global_region_totals %>% slice_max(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou region                deaths  cases population
##              <dbl>          <dbl> <chr>                 <dbl>  <dbl>      <dbl>
##  1            6.66           136. Peru                 219539 4.49e6   32971846
##  2            5.50           187. Bulgaria              38228 1.30e6    6948445
##  3            5.05           227. Hungary               48762 2.20e6    9660350
##  4            4.96           122. Bosnia and Herzegovi~ 16280 4.02e5    3280815
##  5            4.64           166. North Macedonia        9662 3.47e5    2083380
##  6            4.47           460. Montenegro             2808 2.89e5     628062
##  7            4.38           309. Croatia               17987 1.27e6    4105268
##  8            4.25           458. Georgia               16971 1.83e6    3989175
##  9            3.97           431. Czechia               42491 4.62e6   10708982
## 10            3.87           491. Slovakia              21035 2.67e6    5434712
```

```
global_region_totals %>%
  top_n(10, deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = deaths_per_thou)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip()
```




```
global_region_totals %>%
  top_n(-10, deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = deaths_per_thou)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip()
```


Tanzania
Benin
Tajikistan
Niger
South Sudan
Chad
Burundi
Korea, North
Tuvalu
Holy See
region
0.000
0.005
0.010
deaths_per_thou

US Data Tidying

Similar to the Global data, we will tidy the US Data in the same fashion.

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


us_deaths <- us_deaths %>%
  pivot_longer(cols = -c(UID:Population),
                          names_to = "date",
                          values_to = "deaths") %>%
  select(Admin2:deaths) %>%
  mutate(date=mdy(date))

us <- us_cases %>% full_join(us_deaths)

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

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()

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))

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)
```

```
summary(us_by_state)
```

```
##  Province_State     Country_Region          date                cases
##  Length:63216       Length:63216       Min.   :2020-01-22   Min.   :       1
##  Class :character   Class :character   1st Qu.:2020-12-07   1st Qu.:   53858
##  Mode  :character   Mode  :character   Median :2021-09-10   Median :  320530
##                                        Mean   :2021-09-07   Mean   :  851262
##                                        3rd Qu.:2022-06-10   3rd Qu.:  999018
##                                        Max.   :2023-03-09   Max.   :12129699
##
##      deaths         deaths_per_mill   Population         new_cases
##  Min.   :     0.0   Min.   :   0    Min.   :       0   Min.   :-12129697
##  1st Qu.:   873.8   1st Qu.: 629    1st Qu.: 1068778   1st Qu.:        0
##  Median :  4551.0   Median :1779    Median : 3754939   Median :      270
##  Mean   : 11274.4   Mean   : Inf    Mean   : 5743331   Mean   :        3
##  3rd Qu.: 14388.0   3rd Qu.:2901    3rd Qu.: 6863772   3rd Qu.:     1317
##  Max.   :101159.0   Max.   : Inf    Max.   :39512223   Max.   :   207110
##                     NA's   :1110                       NA's   :1
##    new_deaths
##  Min.   :-101159.00
##  1st Qu.:      0.00
##  Median :      2.00
##  Mean   :      0.03
##  3rd Qu.:     15.00
##  Max.   :   4448.00
##  NA's   :1
```

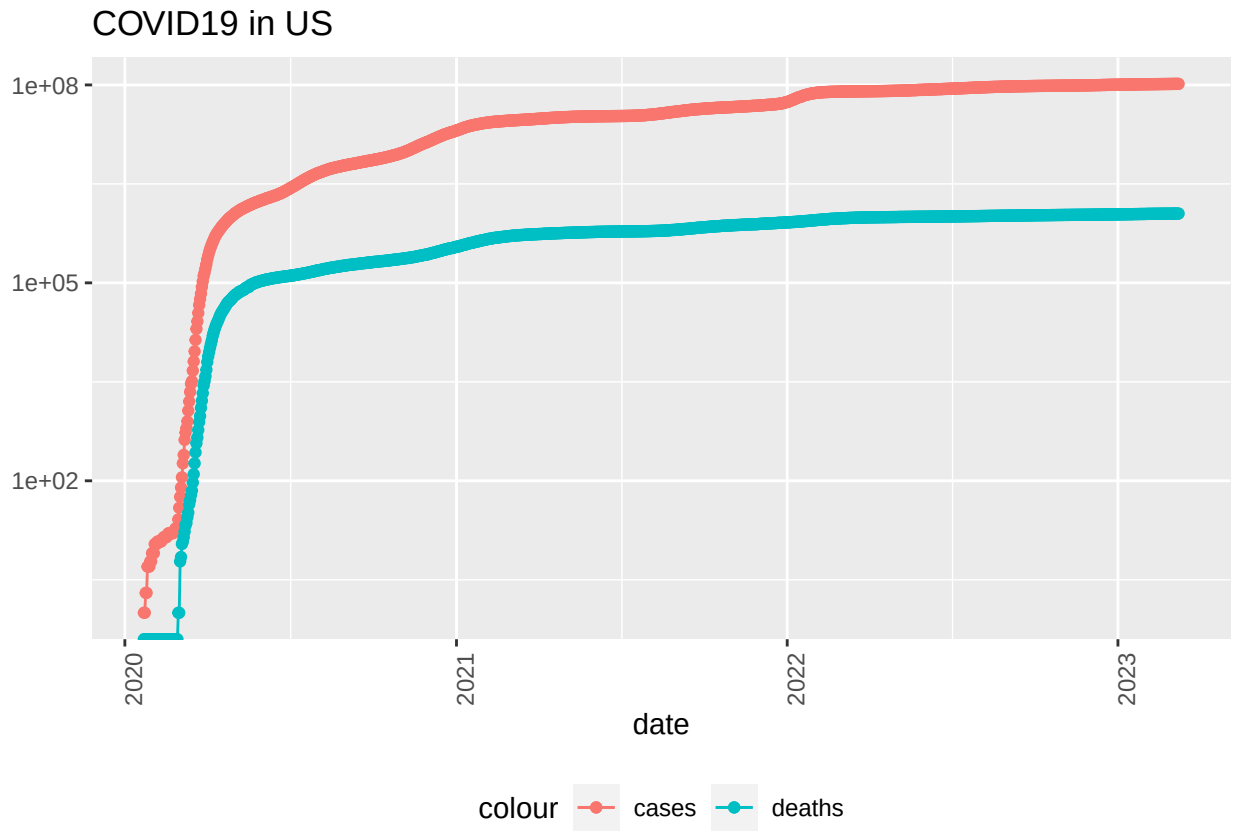
```
summary(us_totals)
```

```
##  Country_Region          date                cases               deaths
##  Length:1143        Min.   :2020-01-22   Min.   :        1   Min.   :      0
##  Class :character   1st Qu.:2020-11-02   1st Qu.:  9401880   1st Qu.: 232306
##  Mode  :character   Median :2021-08-15   Median : 36845902   Median : 617275
##                     Mean   :2021-08-15   Mean   : 47080800   Mean   : 623555
##                     3rd Qu.:2022-05-27   3rd Qu.: 84083678   3rd Qu.:1005190
##                     Max.   :2023-03-09   Max.   :103802702   Max.   :1122724
##
##  deaths_per_mill   Population          new_cases         new_deaths
##  Min.   :   0    Min.   :  2252782   Min.   :  -3862   Min.   :-1013.0
##  1st Qu.: 700    1st Qu.:331887704   1st Qu.:  25993   1st Qu.:  316.5
##  Median :1860    Median :331888491   Median :  55971   Median :  697.0
##  Mean   :1879    Mean   :317646878   Mean   :  90896   Mean   :  983.1
##  3rd Qu.:3028    3rd Qu.:331944132   3rd Qu.: 112464   3rd Qu.: 1411.0
##  Max.   :3382    Max.   :331944132   Max.   :1354508   Max.   : 5195.0
##                                      NA's   :1         NA's   :1
```

Visualizing US Data

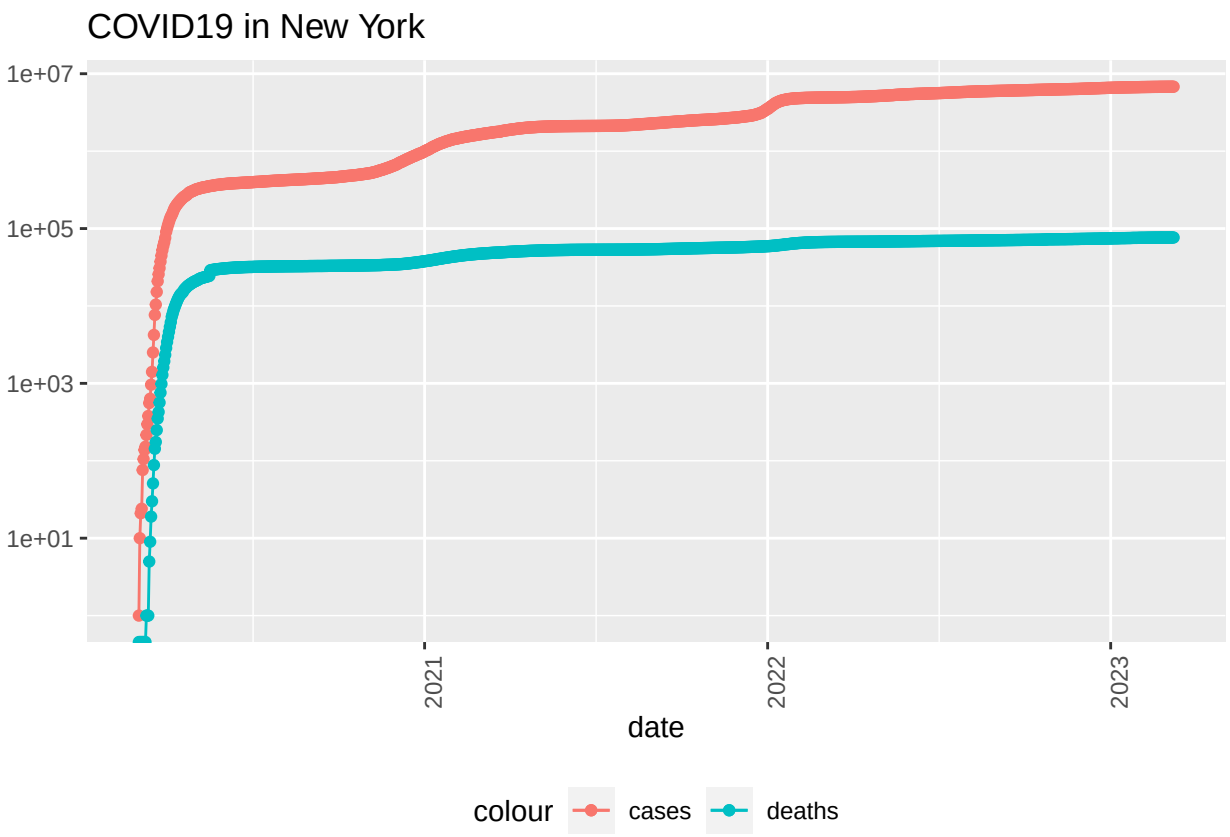
Let's take a look at total cases across the US and a selected state (New York). We can also look at a heatmap of cases and deaths

```
us_totals %>%
  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 = "COVID19 in US", y=NULL)
```




```
state <- "New York"
us_by_state %>%
  filter(Province_State == state) %>%
  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() +
```

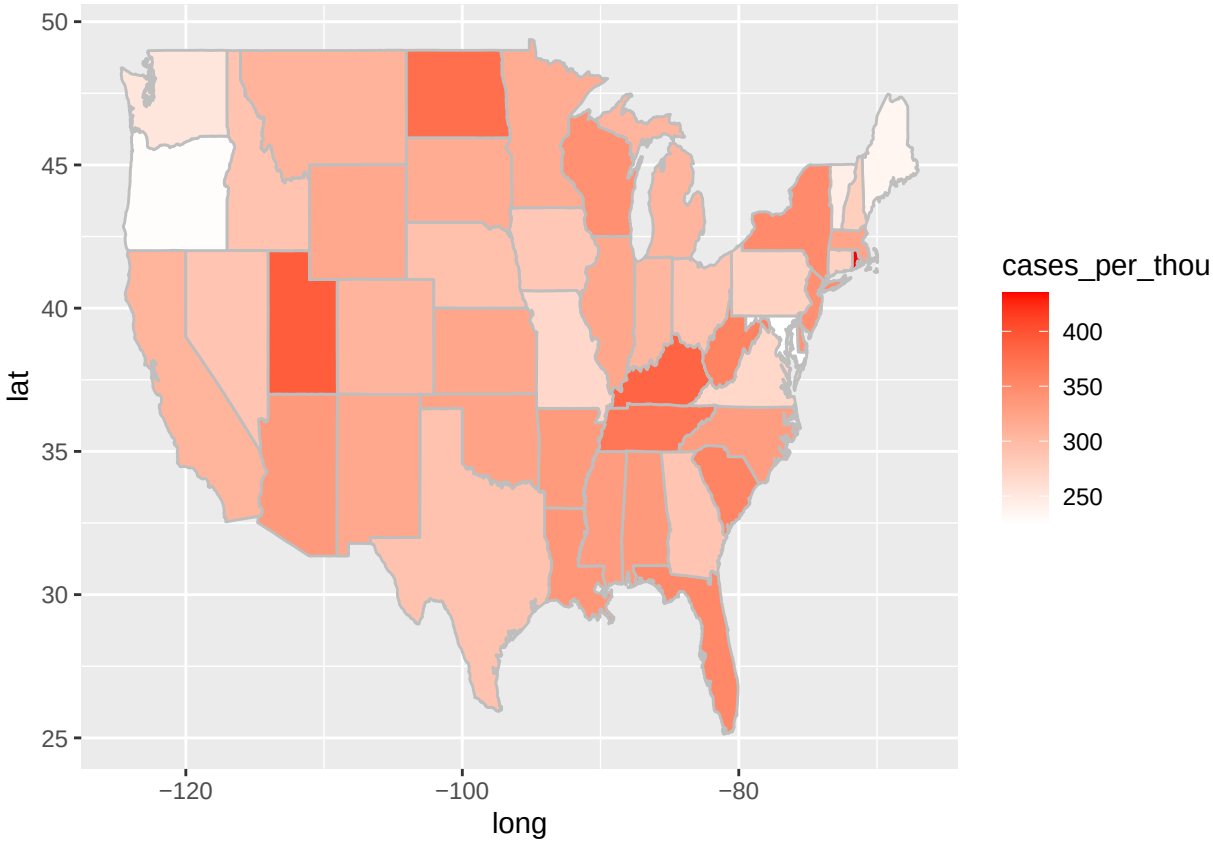
```
  theme(legend.position="bottom",
        axis.text.x = element_text(angle=90)) +
  labs(title = str_c("COVID19 in ", state), y=NULL)
```



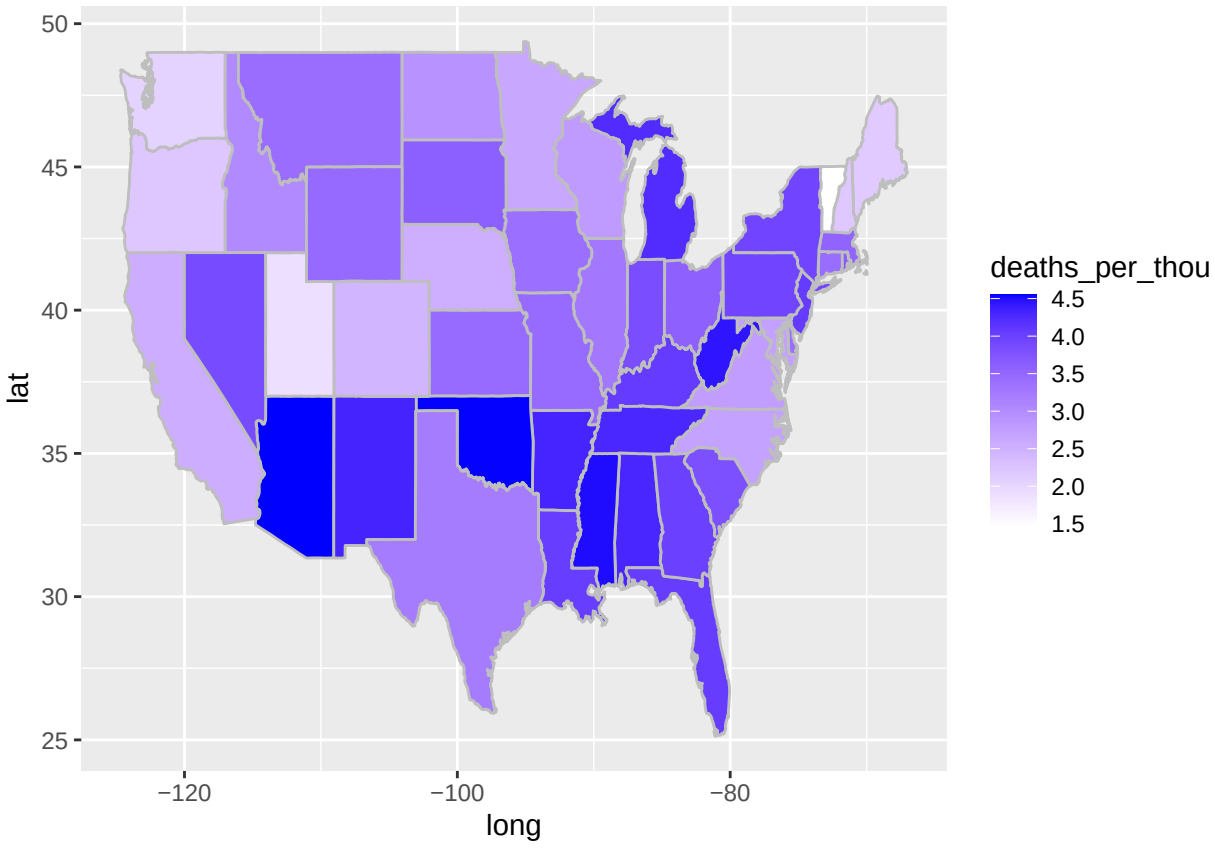

```
state_map <- map_data("state")

us_state_totals$Province_State <- tolower(us_state_totals$Province_State)
us_state_totals <- rename(us_state_totals, region = Province_State)
state_map <- left_join(state_map, us_state_totals, by = "region")

ggplot(state_map, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = cases_per_thou), color = "grey") +
  scale_fill_gradient(low = "white", high =  "red")
```



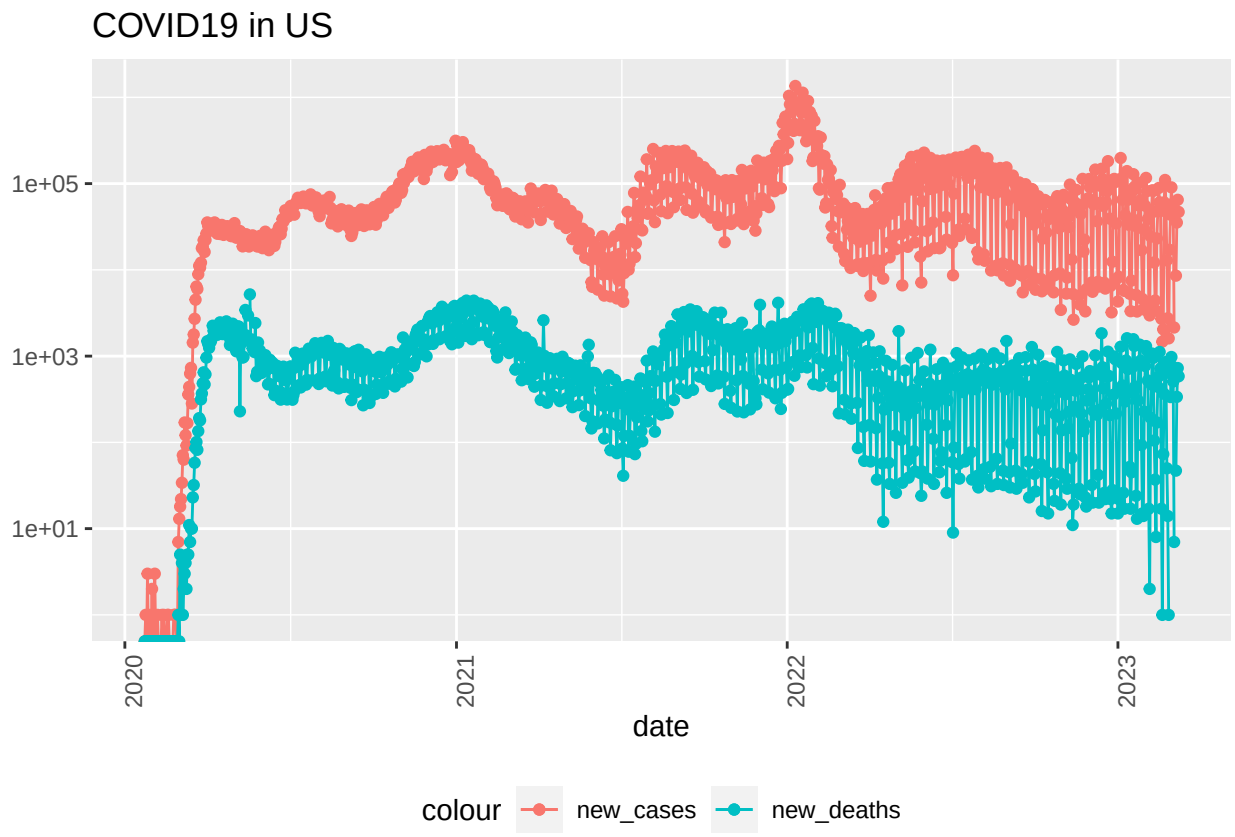

```
ggplot(state_map, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = deaths_per_thou), color = "grey") +
  scale_fill_gradient(low = "white", high =  "blue")
```


50
45
40
35
30
25
lat
−120
−100
−80
long
deaths_per_thou
4.5
4.0
3.5
3.0
2.5
2.0
1.5

US New Cases

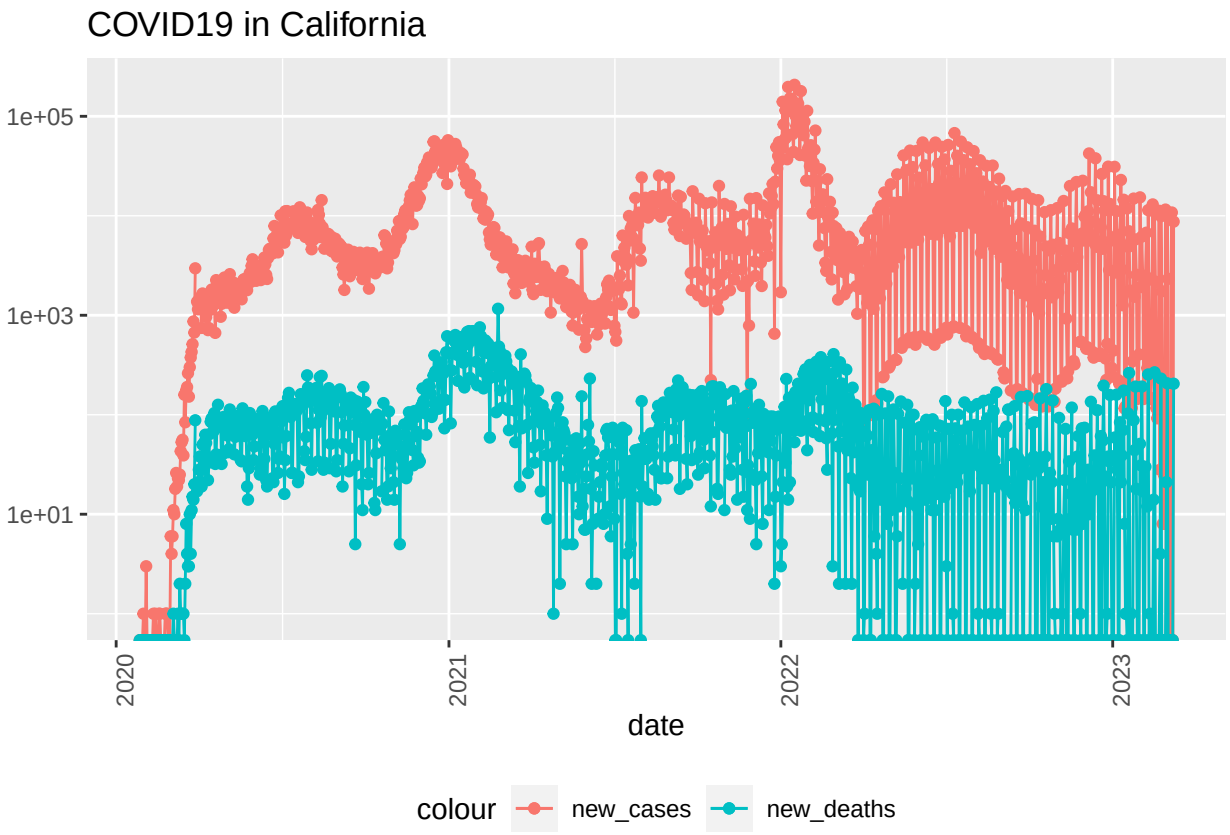
Let's see how the data looks when we plot the daily new cases.

```
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 = "COVID19 in US", y=NULL)
```




```
state <- "California"
us_by_state %>%
  filter(Province_State == state) %>%
  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("COVID19 in ", state), y=NULL)
```



US State Data: Outliers

Like we did for the global data, let's take a look at the states with the best/worst case outcomes with respect to COVID.

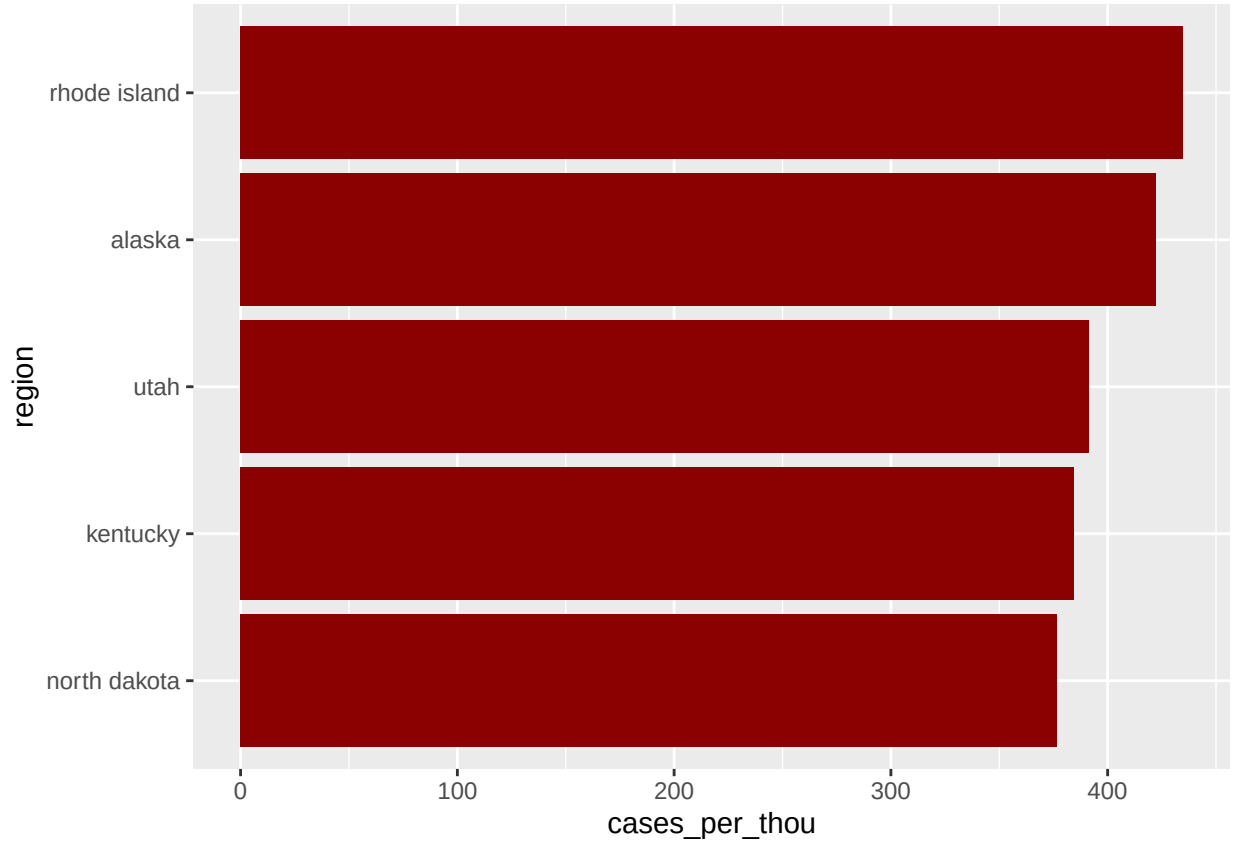
```
us_state_totals %>% slice_min(cases_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou region                deaths  cases population
##              <dbl>          <dbl> <chr>                  <dbl>  <dbl>      <dbl>
##  1           0.611           150. american samoa            34 8.32e3      55641
##  2           2.73            226. maryland               16509 1.37e6    6045680
##  3           2.22            228. oregon                  9373 9.64e5    4217737
##  4           1.21            231. virgin islands           130 2.48e4     107268
##  5           2.18            237. maine                   2928 3.18e5    1344212
##  6           1.49            245. vermont                  929 1.53e5     623989
##  7           0.744           248. northern mariana isl~     41 1.37e4      55144
##  8           2.03            252. district of columbia    1432 1.78e5     705749
##  9           2.06            253. washington             15683 1.93e6    7614893
## 10           3.45            268. missouri               22870 1.78e6    6626371
```

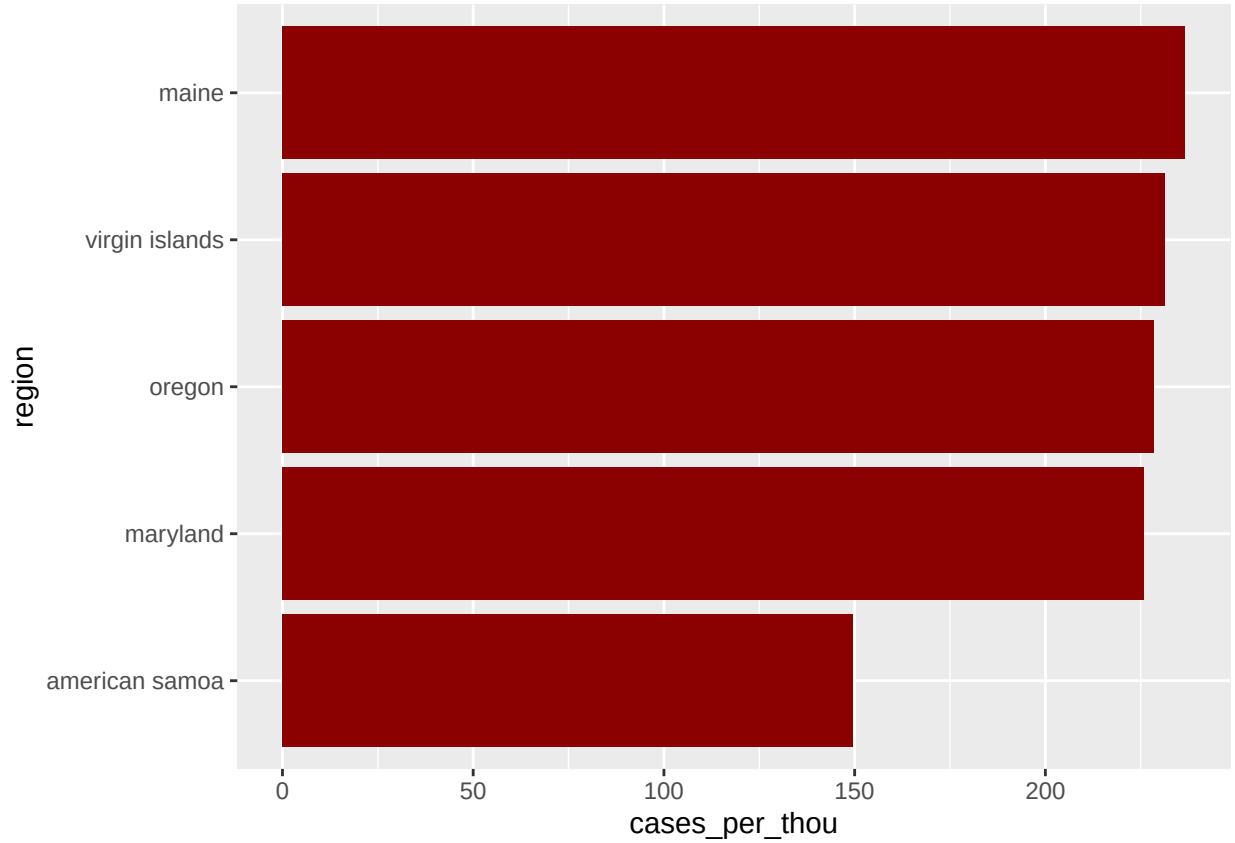
```
us_state_totals %>% slice_max(cases_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou region         deaths   cases population
##              <dbl>          <dbl> <chr>           <dbl>   <dbl>      <dbl>
##  1            3.65           435. rhode island     3870  460697    1059361
##  2            2.04           422. alaska           1486  307655     728809
##  3            1.90           391. utah             5298 1090346    2785478
##  4            4.06           385. kentucky        18130 1718471    4467673
##  5            2.93           377. north dakota     2232  286950     762062
##  6            2.56           372. guam              420   61027     164229
##  7            4.28           368. tennessee       29263 2515130    6829174
##  8            4.44           359. west virginia    7960  642760    1792147
##  9            3.81           357. south carolina  19600 1836568    5148714
## 10            4.04           353. florida         86850 7574590   21477737
```

```
us_state_totals %>%
  top_n(5, cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = cases_per_thou)) +
  geom_bar(stat = "identity", fill = "red4") +
  coord_flip()
```




```
us_state_totals %>%
  top_n(-5, cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(cases_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = cases_per_thou)) +
  geom_bar(stat = "identity", fill = "red4") +
  coord_flip()
```



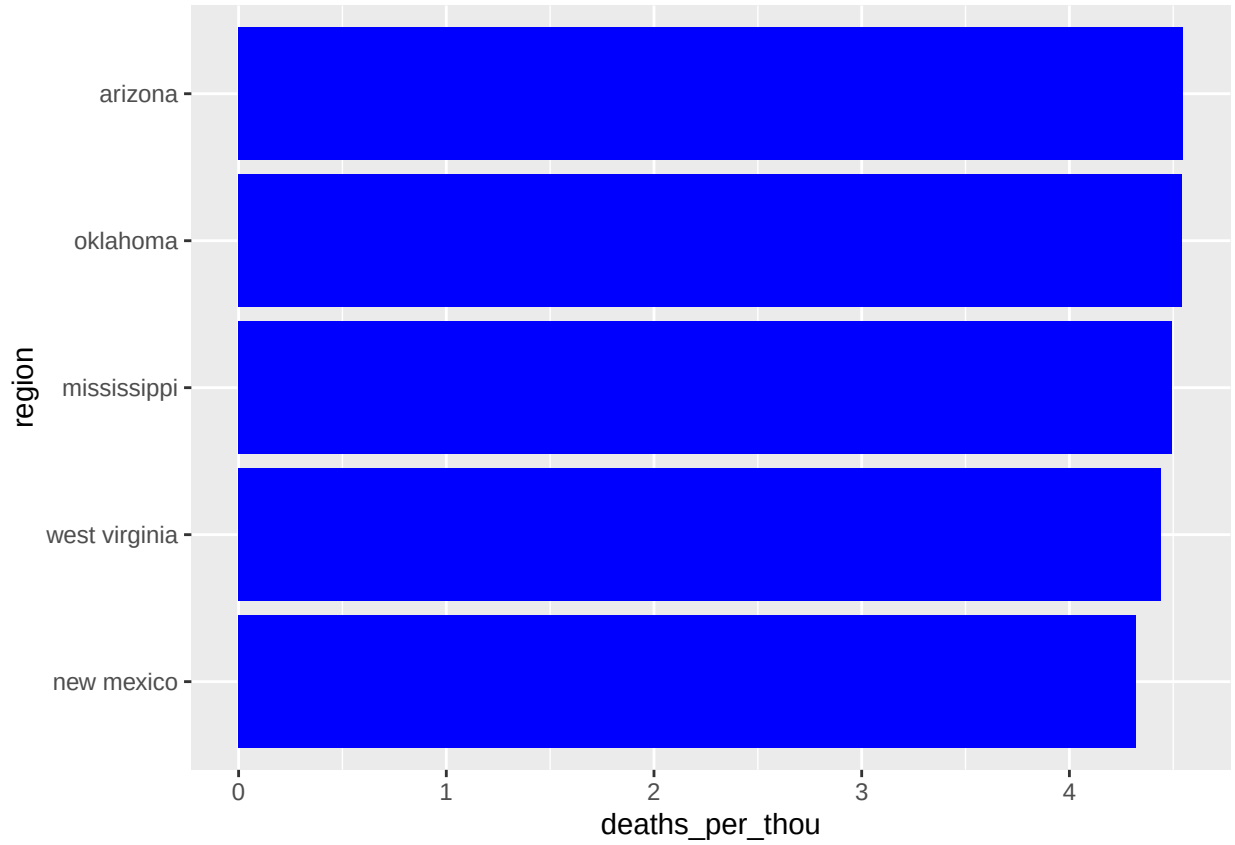

```
us_state_totals %>% slice_min(deaths_per_thou, n=5) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 5 x 6
##   deaths_per_thou cases_per_thou region                deaths  cases population
##             <dbl>          <dbl> <chr>                 <dbl>  <dbl>      <dbl>
## 1           0.611           150. american samoa           34   8320      55641
## 2           0.744           248. northern mariana isla~   41  13666      55144
## 3           1.21            231. virgin islands          130  24813     107268
## 4           1.30            269. hawaii                 1841 380608    1415872
## 5           1.49            245. vermont                 929 152618     623989
```

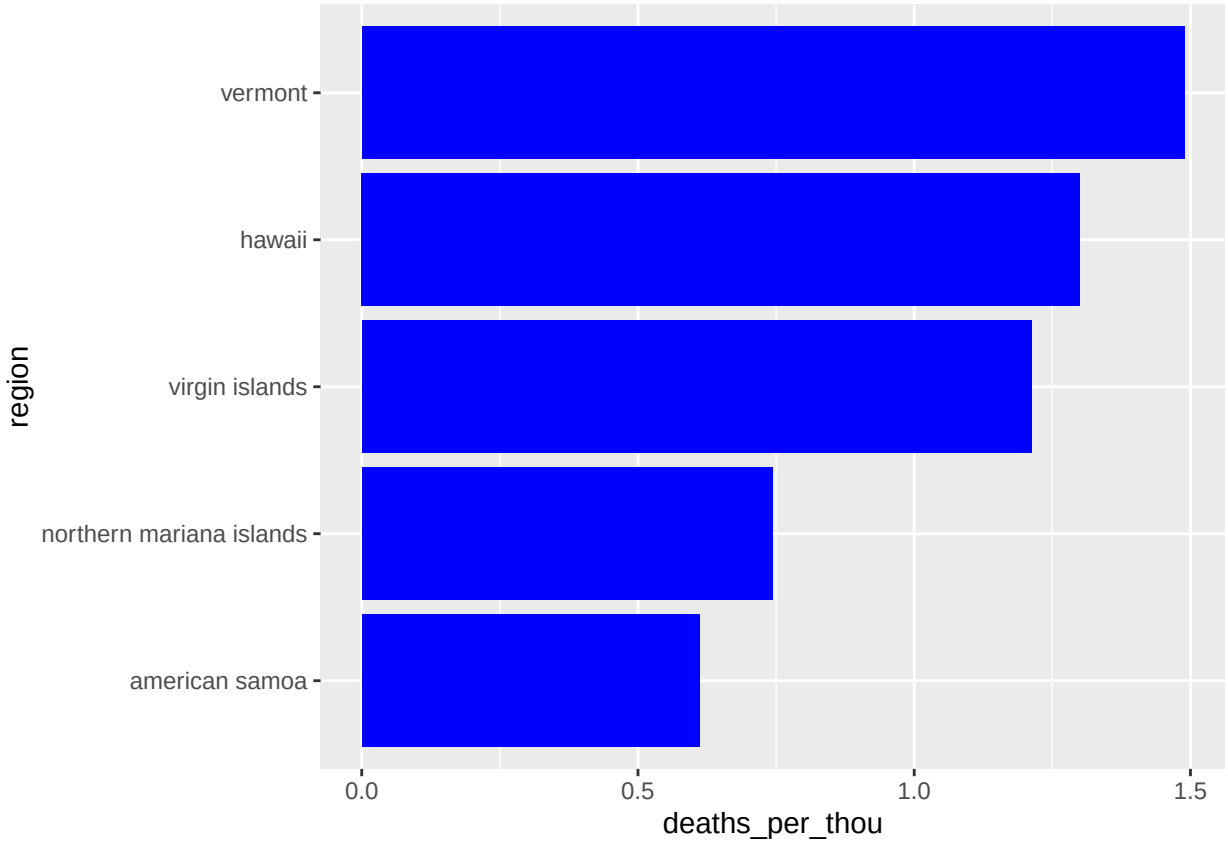
```
us_state_totals %>% slice_max(deaths_per_thou, n=5) %>%
  select(deaths_per_thou, cases_per_thou,everything())
```

```
## # A tibble: 5 x 6
##   deaths_per_thou cases_per_thou region        deaths   cases population
##             <dbl>          <dbl> <chr>          <dbl>   <dbl>      <dbl>
## 1            4.55           336. arizona        33102 2443514    7278717
## 2            4.54           326. oklahoma       17972 1290929    3956971
## 3            4.49           333. mississippi    13370  990756    2976149
## 4            4.44           359. west virginia   7960  642760    1792147
## 5            4.32           320. new mexico      9061  670929    2096829
```

```
us_state_totals %>%
  top_n(5, deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = deaths_per_thou)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip()
```




```
us_state_totals %>%
  top_n(-5, deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  arrange(deaths_per_thou) %>%
  mutate(region = factor(region, levels=region)) %>%
  ggplot(aes(x = region, y = deaths_per_thou)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip()
```


vermont
hawaii
virgin islands
northern mariana islands
american samoa
region
0.0
0.5
1.0
1.5
deaths_per_thou

Cases v. Deaths: United States

One thing we may want to verify is - do cases roughly predict deaths? Our intuition leads us to that conclusion, but lets see how that data checks out for the US data. Note that we've forced a zero intercept - it wouldn't make sense if 0 cases predicted more/less than 0 deaths.

```
mod_state <- lm(deaths_per_thou ~ 0 + cases_per_thou, data = us_state_totals)
summary(mod_state)
```

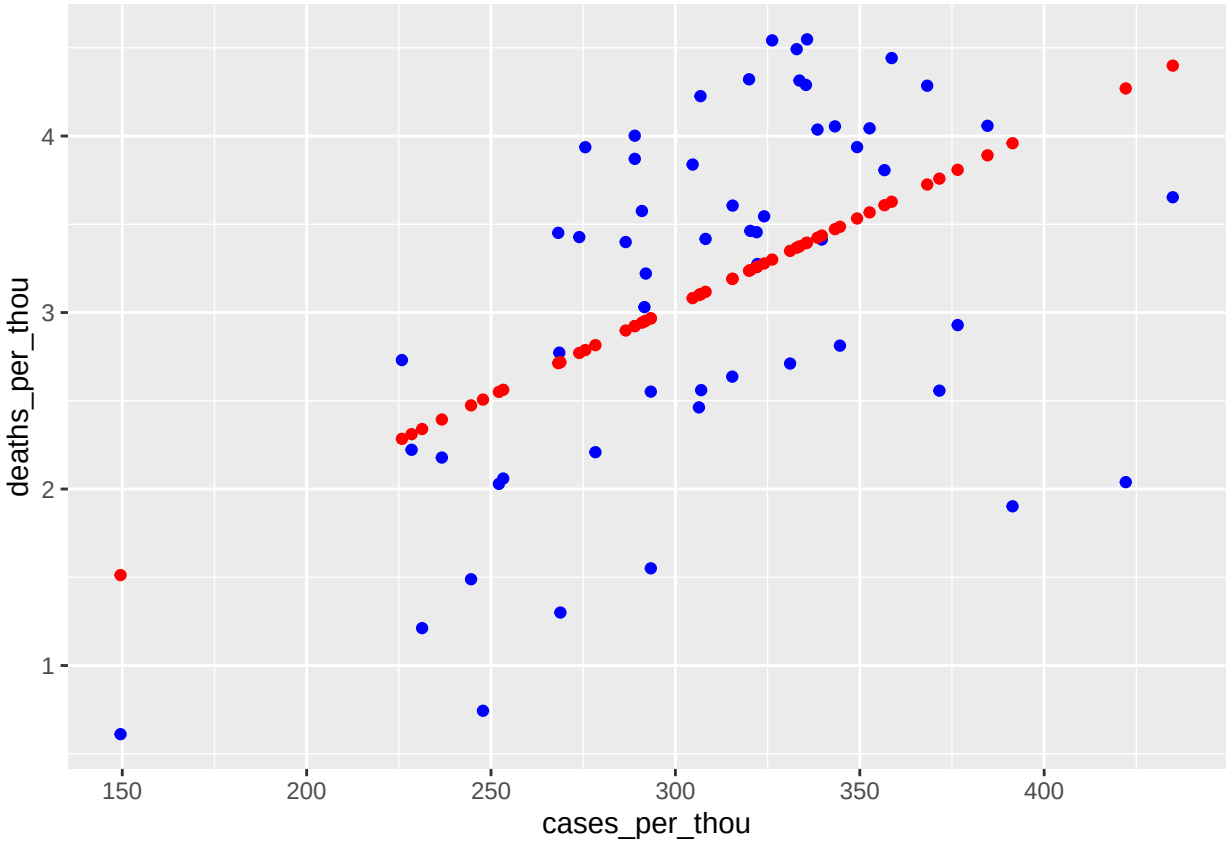
```
##
## Call:
## lm(formula = deaths_per_thou ~ 0 + cases_per_thou, data = us_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2309 -0.6141  0.1983  0.6388  1.2420
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## cases_per_thou 0.0101149  0.0003717   27.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8723 on 55 degrees of freedom
## Multiple R-squared:  0.9309, Adjusted R-squared:  0.9296
## F-statistic: 740.6 on 1 and 55 DF,  p-value: < 2.2e-16
```

```
x_grid <- seq(1,450)
us_state_totals %>% mutate(pred=predict(mod_state))
```

```
## # A tibble: 56 x 7
##    region          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   4.27
##  3 american samoa      34 8.32e3      55641           150.            0.611  1.51
##  4 arizona          33102 2.44e6    7278717           336.            4.55   3.40
##  5 arkansas         13020 1.01e6    3017804           334.            4.31   3.37
##  6 california      101159 1.21e7   39512223           307.            2.56   3.11
##  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   2.55
## # i 46 more rows
```

```
us_tot_w_pred <- us_state_totals %>% mutate(pred=predict(mod_state))

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")
```


deaths_per_thou
4
3
2
1
150
200
250
300
350
400
cases_per_thou

Cases v. Deaths: Global

We expect a similar result if we are to use the global data as well. We will perform the same linear regression:

```
mod_global <- lm(deaths_per_thou ~ 0 + cases_per_thou, data = global_region_totals)
summary(mod_global)
```

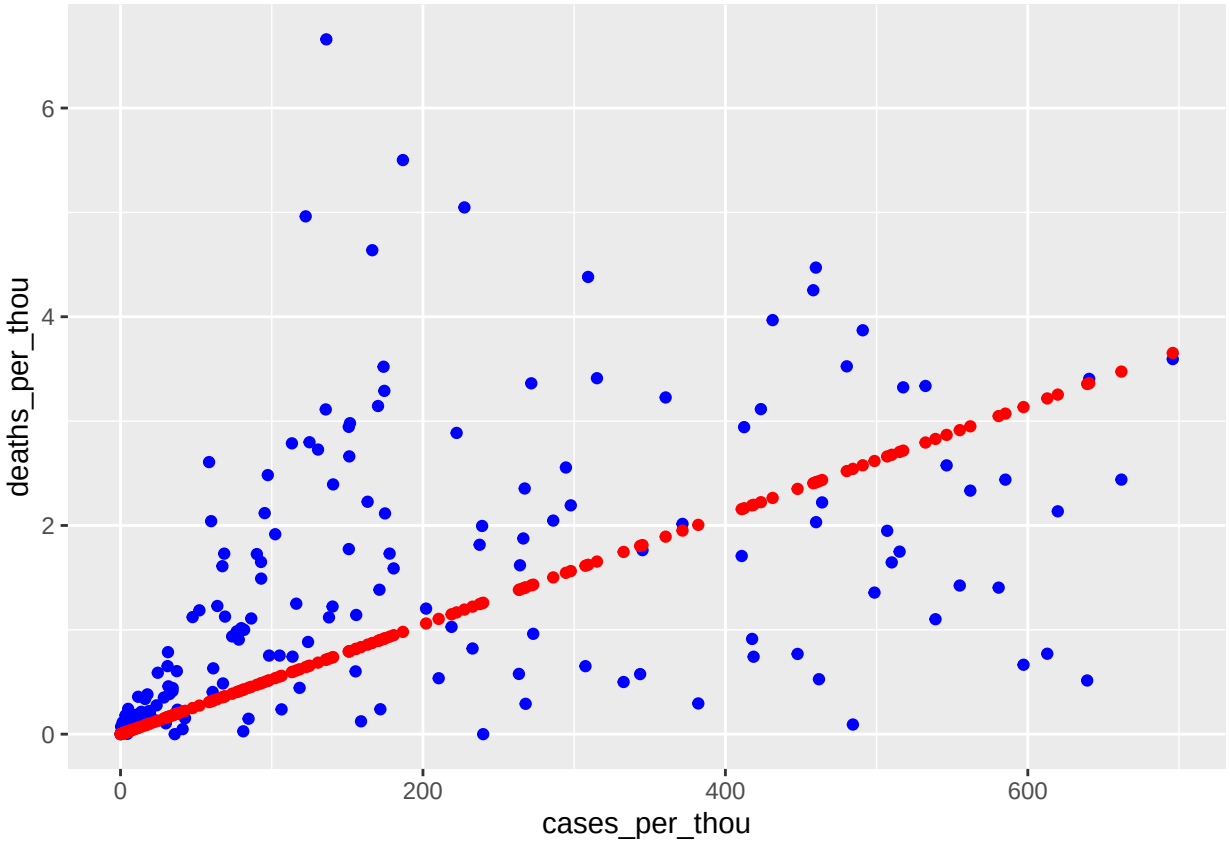
```
##
## Call:
## lm(formula = deaths_per_thou ~ 0 + cases_per_thou, data = global_region_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8414  0.0005  0.1236  0.7187  5.9439
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## cases_per_thou 0.0052495  0.0003498   15.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.238 on 193 degrees of freedom
## Multiple R-squared:  0.5385, Adjusted R-squared:  0.5361
## F-statistic: 225.2 on 1 and 193 DF,  p-value: < 2.2e-16
```

```
x_grid <- seq(1,700)
new_df <- tibble(cases_per_thou = x_grid)
global_region_totals %>% mutate(pred=predict(mod_global))
```

```
## # A tibble: 194 x 7
##    region         deaths  cases population cases_per_thou deaths_per_thou   pred
##    <chr>           <dbl>  <dbl>      <dbl>          <dbl>           <dbl>  <dbl>
##  1 Afghanistan      7896 2.09e5   38928341           5.38          0.203  0.0282
##  2 Albania          3598 3.34e5    2877800         116.            1.25   0.610
##  3 Algeria          6881 2.71e5   43851043           6.19          0.157  0.0325
##  4 Andorra           165 4.79e4      77265         620.            2.14   3.25
##  5 Angola           1933 1.05e5   32866268           3.20          0.0588 0.0168
##  6 Antigua and B~    146 9.11e3      97928          93.0           1.49   0.488
##  7 Argentina      130472 1.00e7   45195777         222.            2.89   1.17
##  8 Armenia          8727 4.47e5    2963234         151.            2.95   0.792
##  9 Australia       19574 1.14e7   25459700         448.            0.769  2.35
## 10 Austria         21970 5.96e6    9006400         662.            2.44   3.47
## # i 184 more rows
```

```
global_tot_w_pred <- global_region_totals %>% mutate(pred=predict(mod_global))

global_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")
```



Notice that the slope for the global data is much higher - we can reasonably conclude that the US was less likely to die from COVID-19. This makes some sense, given that the United States is a first world country and has access to better healthcare than many other countries. However, please take into consideration the bias below.

Bias

For our analysis on the regions most and least affected by COVID-19, we must consider that data could be voluntarily or involuntarily under reported. There may be sociopolitical disadvantages to honestly reporting COVID-19 data which could be Representative in the data. Additionally, regions that are smaller with potentially less capable health infrastructure may be less able to accurately report COVID-19 cases and deaths. One must also be aware of any bias one might have towards particular regions due to culture/race/ethnicity before coming to any conclusions about how a particular region fared with respect to COVID-19.

Conclusions

Recall the questions we asked at the beginning of the report:

1. What regions were most or least affected by COVID-19, with respect to cases and deaths?
2. How does the US compare with the rest of the world with respect to death rates as a result of COVID-19?

For question 1, we found the following:

Global Cases/Deaths:

Most COVID-19 cases per capita: San Marino, Austria, Slovenia, Brunei, Andorra, Iceland, South Korea, France, Denmark, Liechtenstein

Least COVID-19 cases per capita: Sudan, Nigeria, Congo, Burkina Faso, Sierra Leone, Tanzania, Chad, Yemen, Niger, North Korea

Most COVID-19 deaths per capita: Peru, Bulgaria, Hungary, Bosnia, North Macedonia, Montenegro, Croatia, Georgia, Czechia, Slovakia.

Least COVID-19 deaths per capita: Tanzania, Benin, Tajikistan, Niger, South Sudan, Chad, Burundi, North Korea, Tuvalu, Holy See

US Cases/Deaths:

Most COVID-19 cases per capita: Rhode Island, Alaska, Utah, Kentucky, North Dakota

Least COVID-19 cases per capita: Maine, Virgin Islands, Oregon, Maryland, American Samoa

Most COVID-19 deaths per capita: Arizona, Oklahoma, Mississippi, West Virginia, New Mexico

Least COVID-19 deaths per capita: Vermont, Hawaii, Virgin Islands, Northern Mariana Islands, American Samoa

For question 2, we found that the US fared well compared to the rest of the world in terms of death rates according to our linear model.