DSFieldFinal

Caden Zonnefeld

11/23/2022

Purpose and Nature of the Data

The question of interest is if COVID-19 equally effected the United States or if certain states/regions took a harder hit from the virus. Furthemore, how does the United States response compare to other countries around the world? I will examine trends and comparisons of the data to explore this question.

The data used to address this question is sourced from John Hopkins University. They have a public GitHub repository that contains information about number of cases and deaths from the the COVID-19 virus at a national and global level. The data is stored in four separate .csv files corresponding to global cases, US cases, global deaths, and US cases.

Data Collection

Reading the data in from the GitHub repository.

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

url_begin <- 'https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c

files <- 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_begin, files)


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

Wrangling the Data

Pivoting the data into a format that is more friendly for future analysis. End up with data that shows Province/State, Country, Date, Number of Cases, and Number of Deaths. Furthermore, a population feature is added to the global dataset.

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

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

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

US_cases <- US_cases %>%
      pivot_longer(cols = -(UID:Combined_Key),
                   names_to = 'date',
                   values_to = 'cases') %>%
      select(UID:cases) %>%
      mutate(date = mdy(date)) %>%
      select(-c(iso2, iso3, code3, FIPS, Lat, Long_))

US_deaths <- US_deaths %>%
      pivot_longer(cols = -(UID:Combined_Key),
                   names_to = 'date',
                   values_to = 'deaths') %>%
      select(UID:deaths) %>%
      mutate(date = mdy(date)) %>%
      select(-c(iso2, iso3, code3, FIPS, 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)

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

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

Initial Data Examination and Population Join

Viewing the basic summary statistics of the data and performing a gut check for the ranges of values.

```
summary(global)
```

```
##  Combined_Key       Province_State     Country_Region          date           
##  Length:299693      Length:299693      Length:299693      Min.   :2020-01-22  
##  Class :character   Class :character   Class :character   1st Qu.:2020-10-07  
##  Mode  :character   Mode  :character   Mode  :character   Median :2021-06-23  
##                                                           Mean   :2021-06-23  
##                                                           3rd Qu.:2022-03-09  
##                                                           Max.   :2022-11-23  
##      cases              deaths       
##  Min.   :       0   Min.   :      0  
##  1st Qu.:     490   1st Qu.:      3  
##  Median :   11427   Median :    125  
##  Mean   :  822645   Mean   :  12356  
##  3rd Qu.:  189639   3rd Qu.:   2626  
##  Max.   :98503462   Max.   :1078929  
```

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

summary(global)
```

```
##  Combined_Key       Province_State     Country_Region          date           
##  Length:276405      Length:276405      Length:276405      Min.   :2020-01-22  
##  Class :character   Class :character   Class :character   1st Qu.:2020-11-14  
##  Mode  :character   Mode  :character   Mode  :character   Median :2021-07-23  
##                                                           Mean   :2021-07-19  
##                                                           3rd Qu.:2022-03-27  
##                                                           Max.   :2022-11-23  
##      cases              deaths       
##  Min.   :       1   Min.   :      0  
##  1st Qu.:    1019   1st Qu.:      7  
##  Median :   16408   Median :    183  
##  Mean   :  891955   Mean   :  13398  
##  3rd Qu.:  233116   3rd Qu.:   3187  
##  Max.   :98503462   Max.   :1078929  
```

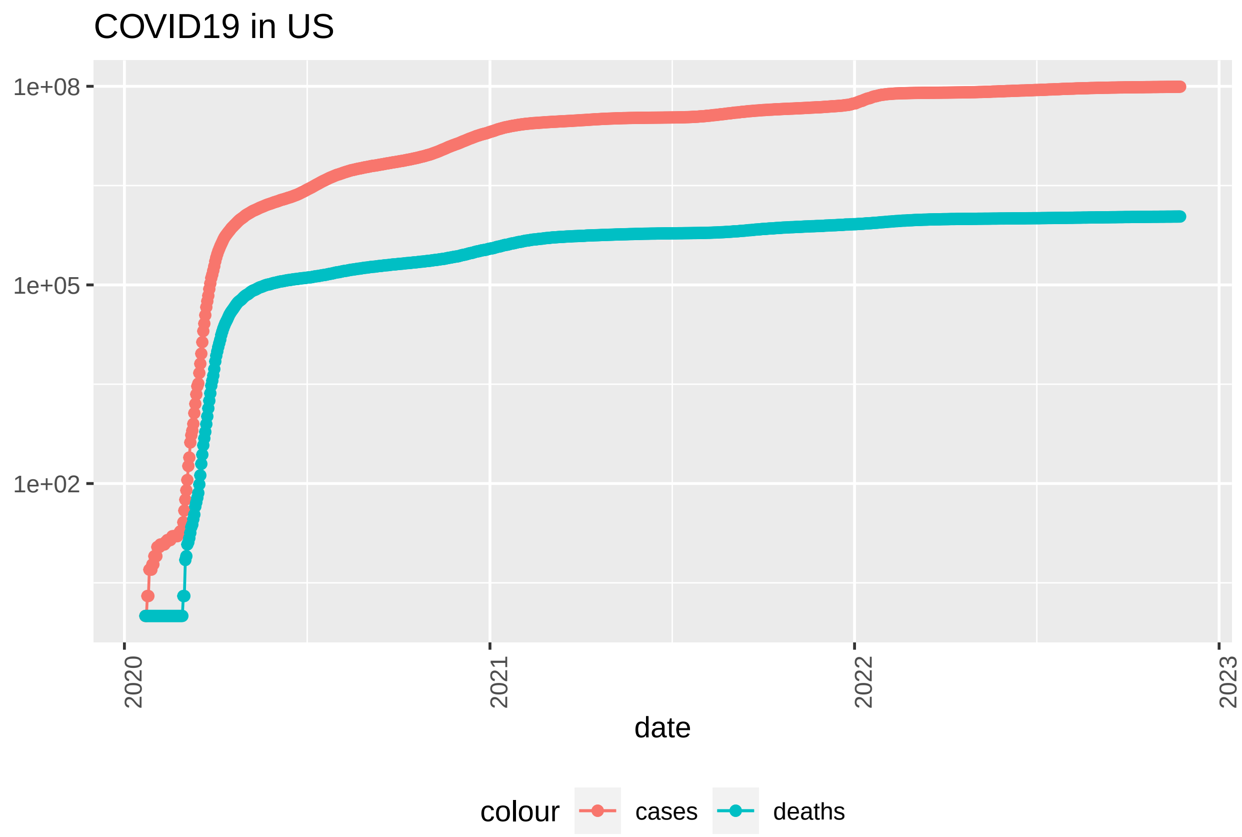
```
global %>% filter (cases > 98000000)
```

```
## # A tibble: 12 x 6
##    Combined_Key Province_State Country_Region date            cases  deaths
##    <chr>        <chr>          <chr>          <date>          <dbl>   <dbl>
##  1 US            <NA>           US             2022-11-12 98001862 1074656
##  2 US            <NA>           US             2022-11-13 98004208 1074657
##  3 US            <NA>           US             2022-11-14 98054070 1074898
##  4 US            <NA>           US             2022-11-15 98113463 1075285
##  5 US            <NA>           US             2022-11-16 98197743 1076130
##  6 US            <NA>           US             2022-11-17 98251350 1076549
##  7 US            <NA>           US             2022-11-18 98306970 1077079
##  8 US            <NA>           US             2022-11-19 98311573 1077090
##  9 US            <NA>           US             2022-11-20 98314841 1077090
## 10 US            <NA>           US             2022-11-21 98357398 1077284
## 11 US            <NA>           US             2022-11-22 98392076 1077836
## 12 US            <NA>           US             2022-11-23 98503462 1078929
```

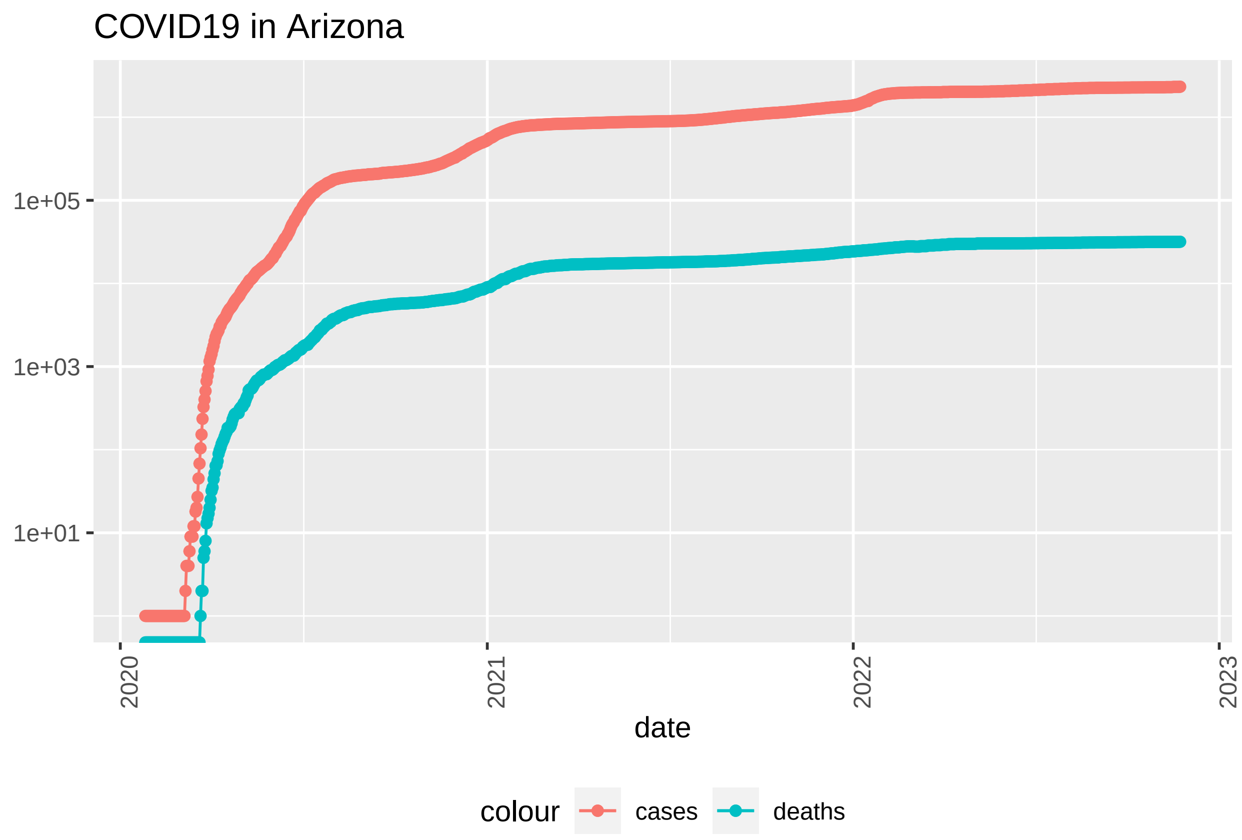
Visualizing the Cleaned Data

Showing the increase in cases and deaths for all of the United States then just Arizona and California in particular. The gradual increase over time makes sense but this covers over some of the smaller trends. I will investigate this by including some lagged features to see the microtrends that occur in the data.

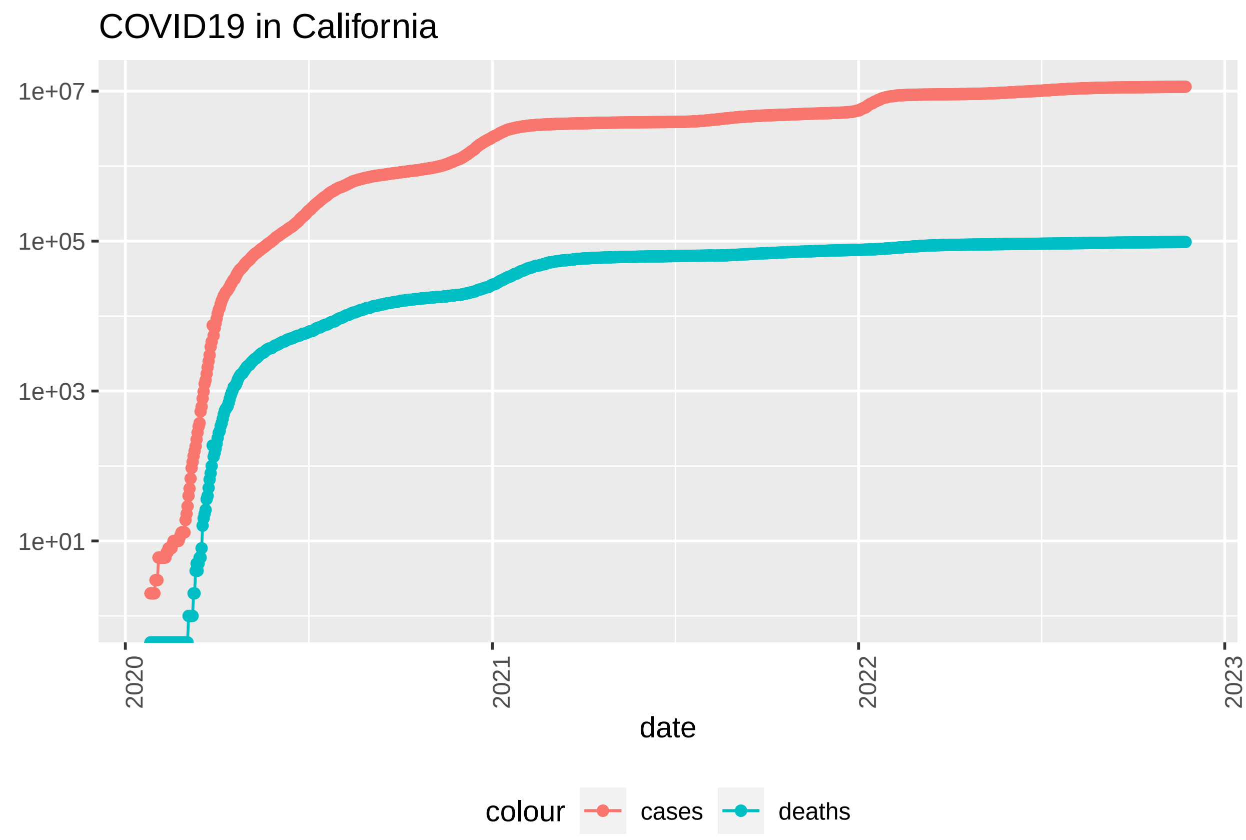
```
US_totals %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y = cases)) +
  geom_line(aes(color = 'cases')) +
  geom_point(aes(color = 'cases')) +
  geom_line(aes(y=deaths, color = 'deaths')) +
  geom_point(aes(y=deaths, color = 'deaths')) +
  scale_y_log10() +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle=90)) +
  labs(title = 'COVID19 in US', y = NULL)
```



```
US_by_state %>%
  filter(Province_State == 'Arizona') %>%
  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 = 'COVID19 in Arizona', y = NULL)
```



```
US_by_state %>%
  filter(Province_State %in% c('California')) %>%
  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 = 'COVID19 in California', y = NULL)
```



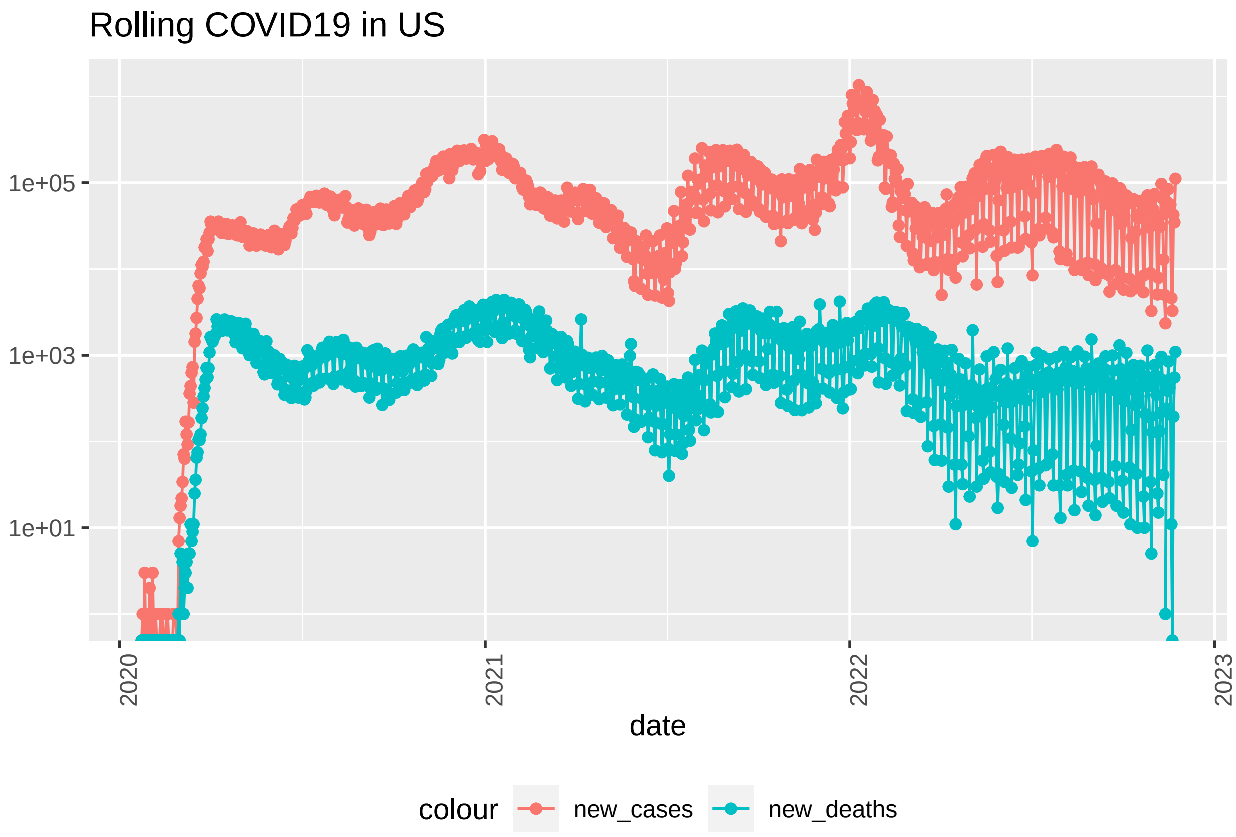
Enriching the Dataset

Adding lagged features to examine the progression of cases and deaths due to COVID19. Also visualizing the progression of new cases/deaths. This view of the data gives a more insightful perspective as it better illustrates the recent developments of the COVID19 virus. Note that the increased variability toward the present is due to inconsistency in data collection/sporadic results. For example, we can conclude that Colorado's new cases peaked around the beginning of 2022. Next I examined West Coast states and noticed a similar trend among each of the stats for infections and deaths from the COVID19 virus. Following this I considerd New York and California, two of the most populous states in the country to see if and how they responded differently. New York suffers a greater initial spike upon the onset of the virus; however, the two states follow a similar trend folliwng that. The large spike likely is a result of high population density in New York. Finally, I was surprised to see that New York (a high population state) and Wyoming (a low population state) experienced similar trends in response to COVID19. My initial hypothesis was that lower population states would experience differing effects.

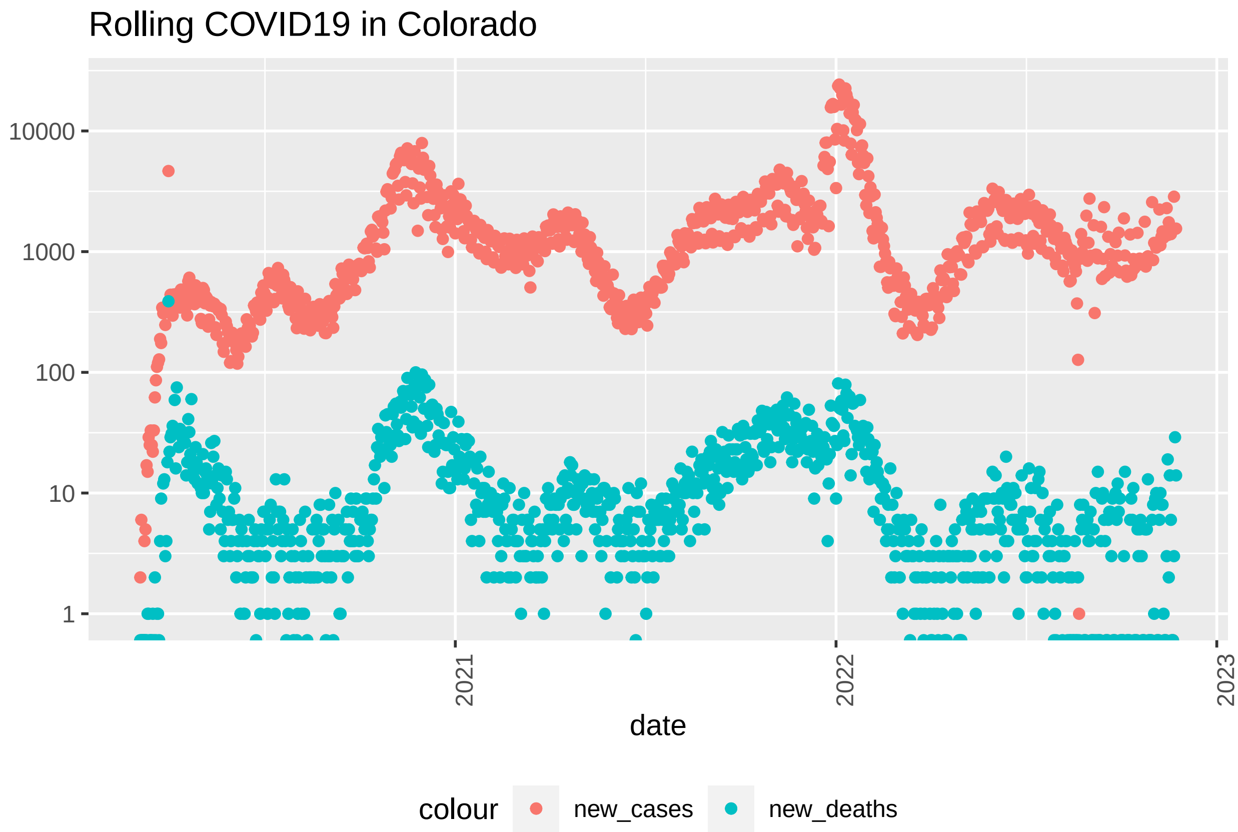
```r
US_by_state <- US_by_state %>%
        mutate(new_cases = cases - lag(cases),
               new_deaths = deaths - lag(deaths))

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

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

tail(US_totals %>% select(new_cases, new_deaths, everything()))
```

```
## # A tibble: 6 x 6
##   new_cases new_deaths Country_Region date          cases    deaths
##       <dbl>      <dbl> <chr>          <date>        <dbl>     <dbl>
## 1      4603         11 US             2022-11-19 98311573   1077090
## 2      3268          0 US             2022-11-20 98314841   1077090
## 3     42557        194 US             2022-11-21 98357398   1077284
## 4     34678        552 US             2022-11-22 98392076   1077836
## 5    111386       1093 US             2022-11-23 98503462   1078929
## 6        NA  331796208 US             NA               NA 332875137
```

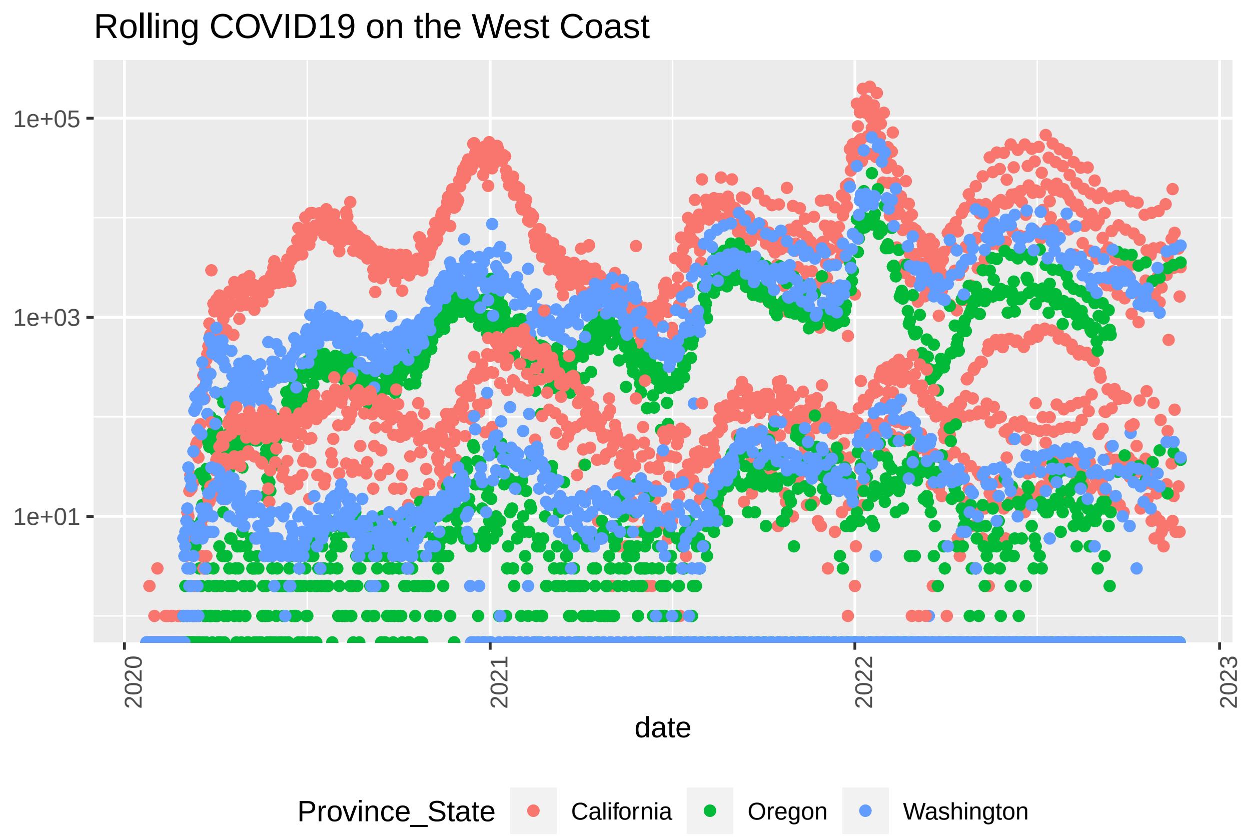
```
US_totals %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y = new_cases)) +
  geom_line(aes(color = 'new_cases')) +
  geom_point(aes(color = 'new_cases')) +
  geom_line(aes(y=new_deaths, color = 'new_deaths')) +
  geom_point(aes(y=new_deaths, color = 'new_deaths')) +
  scale_y_log10() +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle=90)) +
  labs(title = 'Rolling COVID19 in US', y = NULL)
```



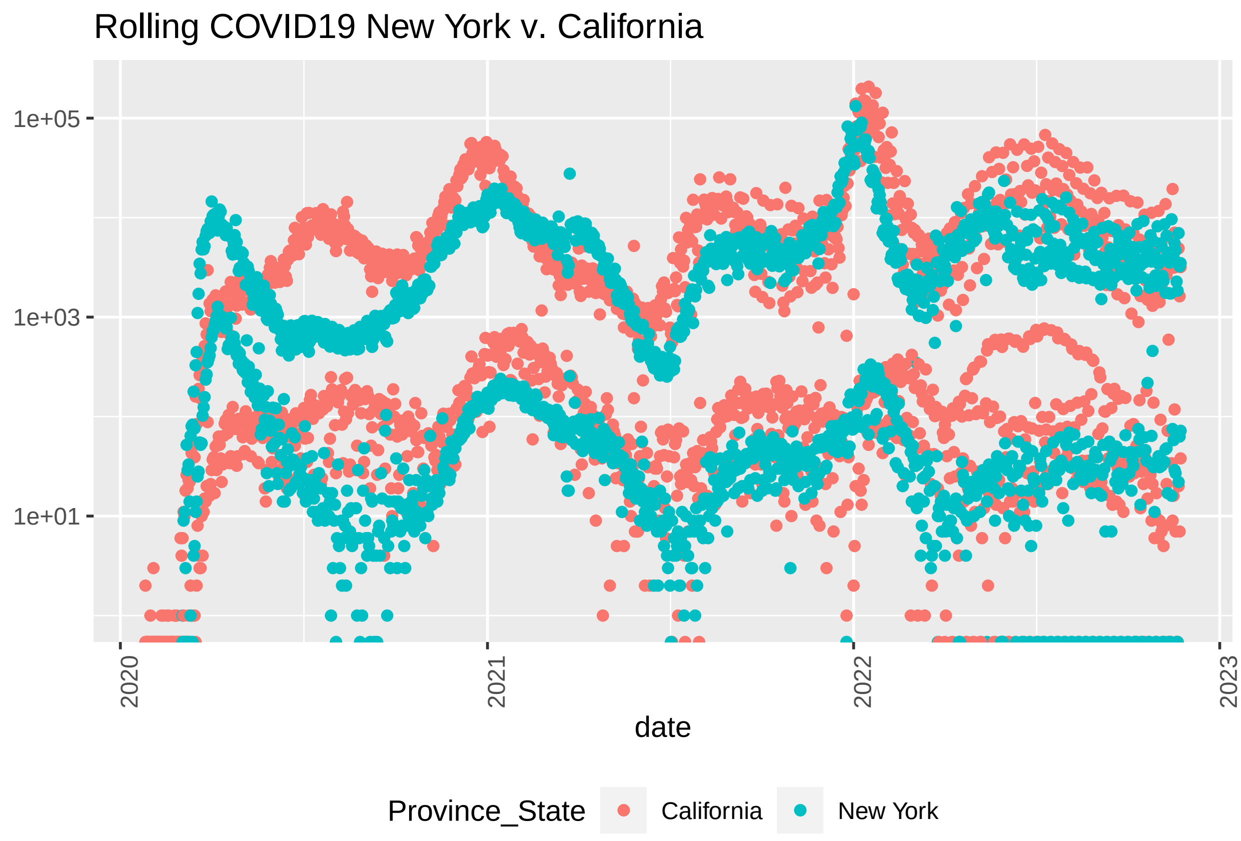
```
US_by_state %>%
  filter(Province_State == 'Colorado') %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y = new_cases)) +
  geom_point(aes(color = 'new_cases')) +
  geom_point(aes(y=new_deaths, color = 'new_deaths')) +
  scale_y_log10() +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle=90)) +
  labs(title = 'Rolling COVID19 in Colorado', y = NULL)
```



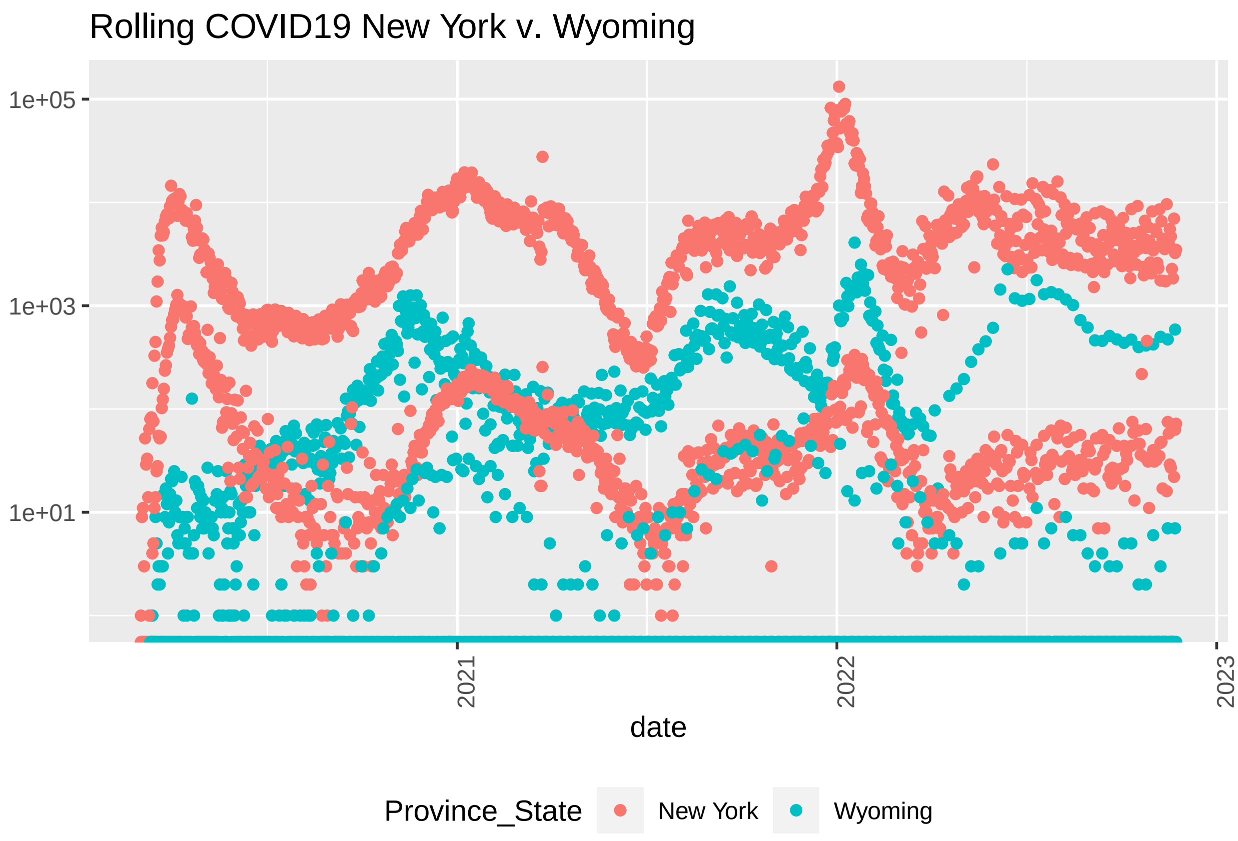
```
US_by_state %>%
  filter(Province_State %in% c('Washington', 'Oregon', 'California')) %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y = new_cases)) +
  geom_point(aes(color = Province_State)) +
  geom_point(aes(y=new_deaths, color = Province_State)) +
  scale_y_log10() +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle=90)) +
  labs(title = 'Rolling COVID19 on the West Coast', y = NULL)
```



```
US_by_state %>%
  filter(Province_State %in% c('New York', 'California')) %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y = new_cases)) +
  geom_point(aes(color = Province_State)) +
  geom_point(aes(y=new_deaths, color = Province_State)) +
  scale_y_log10() +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle=90)) +
  labs(title = 'Rolling COVID19 New York v. California', y = NULL)
```



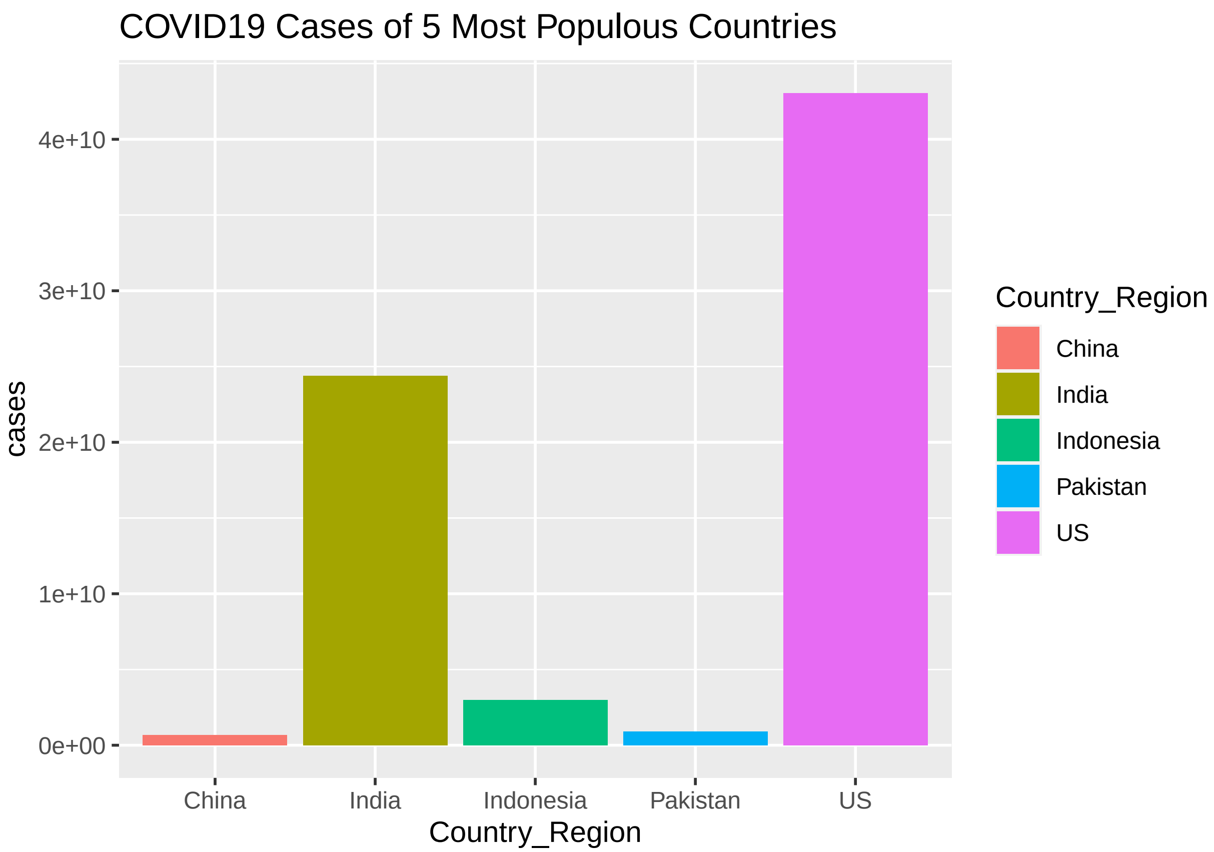
```
US_by_state %>%
  filter(Province_State %in% c('New York', 'Wyoming')) %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y = new_cases)) +
  geom_point(aes(color = Province_State)) +
  geom_point(aes(y=new_deaths, color = Province_State)) +
  scale_y_log10() +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle=90)) +
  labs(title = 'Rolling COVID19 New York v. Wyoming', y = NULL)
```



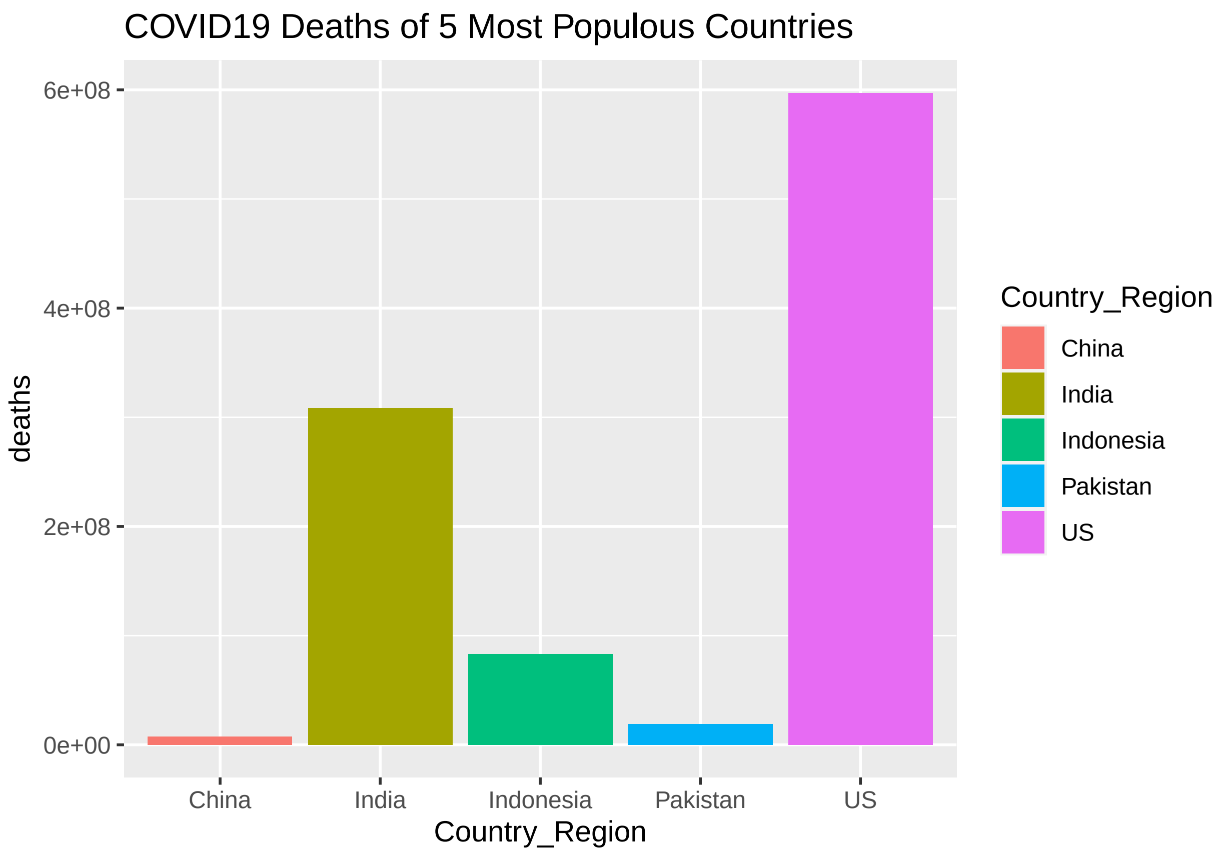
Evaluting the United States Response

This bar chart shows that the the effectiveness of the United States response to the COVID19 virus paled in comparison to that of the other most populous countries in the world. Even though China and India have vastly larger populations, they experienced far less cases. Focusing in on a comparison between the United States and India (a country with 3x the population), we can see that India has done a better job of prevention as their infection and fatality rate are decreasing while the United States rates have stayed relatively the same.

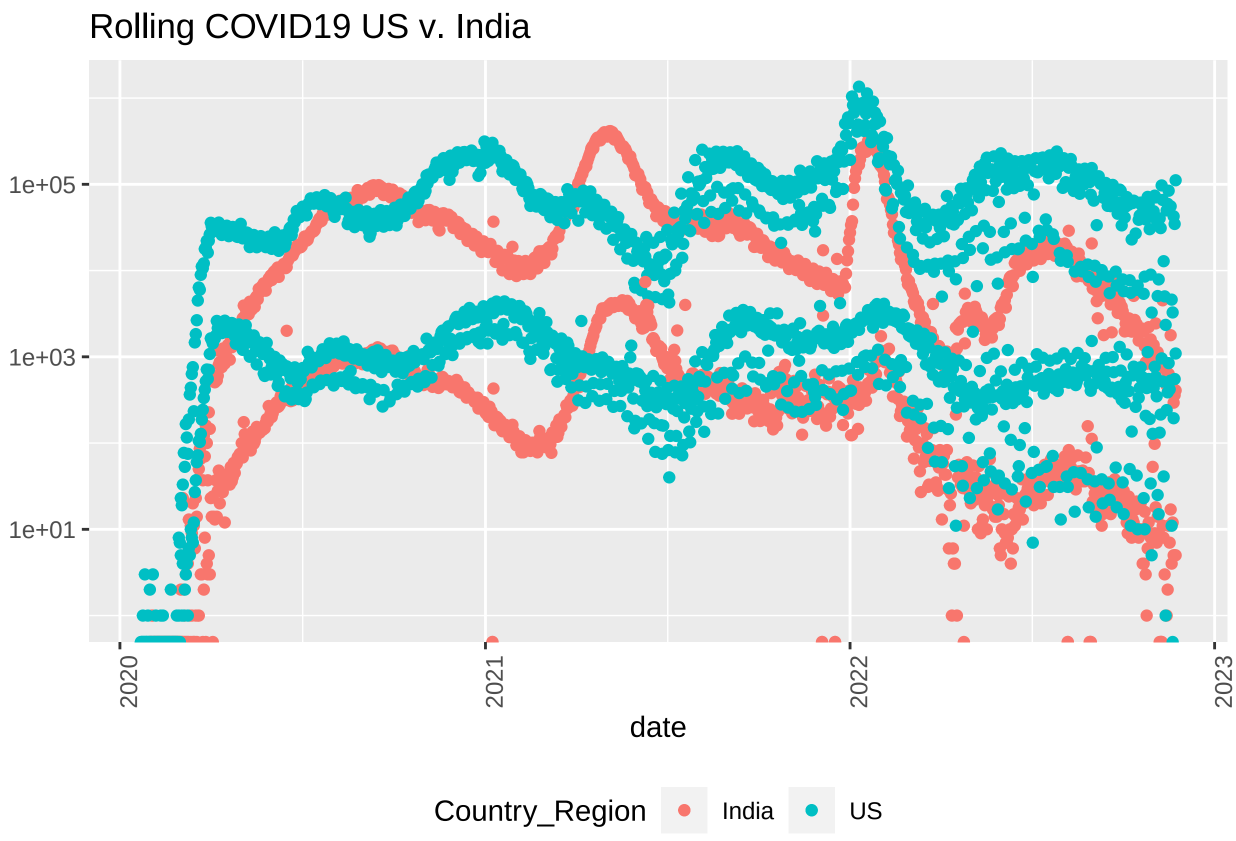
```
neighbors <- c('Pakistan', 'Indonesia', 'US', 'India', 'China')

global %>%
  filter(Country_Region %in% neighbors) %>%
  group_by(Country_Region) %>%
  summarize(cases = sum(cases)) %>%
  ggplot(aes(x=Country_Region)) +
  geom_col(aes(y=cases, fill = Country_Region)) +
  labs(title = 'COVID19 Cases of 5 Most Populous Countries')
```



```
global %>%
  filter(Country_Region %in% neighbors) %>%
  group_by(Country_Region) %>%
  summarize(deaths = sum(deaths)) %>%
  ggplot(aes(x=Country_Region)) +
  geom_col(aes(y=deaths, fill = Country_Region)) +
  labs(title = 'COVID19 Deaths of 5 Most Populous Countries')
```



```
global %>%
  filter(Country_Region %in% c('US', 'India')) %>%
  filter(cases > 0) %>%
  ggplot(aes(x=date, y = new_cases)) +
  geom_point(aes(color = Country_Region)) +
  geom_point(aes(y=new_deaths, color = Country_Region)) +
  scale_y_log10() +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle=90)) +
  labs(title = 'Rolling COVID19 US v. India', y = NULL)
```



Modeling Data

Comparing the two least populous states based on COVID19 cases/deaths. Though Wyoming experienced a statistically significant increase in cases they did not have a statistically significant difference in deaths. Furthermore, we can notice that the US had a statistically significant amount greater cases and deaths than the much larger nation of India.

```
smallest <- US_by_state %>%
        filter(Province_State %in% c('Vermont', 'Wyoming'))
linear <- lm(cases ~ Province_State, data = smallest)
summary(linear)
```

```
##
## Call:
## lm(formula = cases ~ Province_State, data = smallest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79122 -48779 -21957  68940 101304
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)               50313       1857   27.10   <2e-16 ***
## Province_StateWyoming     28809       2626   10.97   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 59790 on 2072 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.05491,    Adjusted R-squared:  0.05446
## F-statistic: 120.4 on 1 and 2072 DF,  p-value: < 2.2e-16
```

```
linear <- lm(deaths ~ Province_State, data = smallest)
summary(linear)
```

```
##
## Call:
## lm(formula = deaths ~ Province_State, data = smallest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1452   -865   -669   -233 623068
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              920.6      579.4   1.589    0.112
## Province_StateWyoming    531.0      819.5   0.648    0.517
##
## Residual standard error: 18670 on 2074 degrees of freedom
## Multiple R-squared:  0.0002024,  Adjusted R-squared:  -0.0002796
## F-statistic: 0.4199 on 1 and 2074 DF,  p-value: 0.517
```

```
neighbor_countries <- global %>%
        filter(Country_Region %in% c('US', 'India'))
linear <- lm(cases ~ Country_Region, data = neighbor_countries)
summary(linear)
```

```
##
## Call:
## lm(formula = cases ~ Country_Region, data = neighbor_countries)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -41526277 -22349791  -4082772  19399766  56977184
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23707351     834827   28.40   <2e-16 ***
## Country_RegionUS 17818927    1178344   15.12   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 26780000 on 2064 degrees of freedom
## Multiple R-squared:  0.09974,    Adjusted R-squared:  0.09931
## F-statistic: 228.7 on 1 and 2064 DF,  p-value: < 2.2e-16
```

```r
neighbor_countries <- global %>%
        filter(Country_Region %in% c('US', 'India'))
linear <- lm(deaths ~ Country_Region, data = neighbor_countries)
summary(linear)
```

```
##
## Call:
## lm(formula = deaths ~ Country_Region, data = neighbor_countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -575923 -261366   29935  224722  503006
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        299967       9200   32.61   <2e-16 ***
## Country_RegionUS   275956      12985   21.25   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 295100 on 2064 degrees of freedom
## Multiple R-squared:  0.1795, Adjusted R-squared:  0.1791
## F-statistic: 451.7 on 1 and 2064 DF,  p-value: < 2.2e-16
```

Biases and Conclusion

This dataset had several areas in which bias may have affected the results. For instance, the reporting practices of COVID19 cases differences greatly across different countries and even regions within a country. Furthermore, more desolate regions of countries like India or China may be under-counted or even flat out ignored. Another place that bias may have creeped into this data is by using raw counts instead of rates by incorporating populations. Unfortunately, my computer was unable to access the population .csv file that was used in the lectures.

Despite some anomalies, my conclusion is that the COVID19 pandemic affected the United States in a similar fashion across the board in regards to deaths, though different regions produced differing numbers of cases. On the other hand, preliminary analysis and visualization indicates that the United States did a poor job handling the pandemic compared to other populous countries in the world. This is clearly evident by vastly greater counts of cases and deaths and was confirmed with a linear model that compared to India, though the same could be done for any other peer country with similar populations.