Week 5 - JHU CSSE COVID-19 Data Report

Anubhav Sharma

22/04/2022

COVID19 Report

This report has been made on the basis of the public COVID-19 Data Repository maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

JHU CSSE COVID-19 Data is the is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).

This huge data set has aggregated various sources of information about COVID19 like WHO, ECDC, US CDC, COVID Tracking Project, WorldoMeters, and many others. JHU CSSE COVID 19 is a data set used by various organizations and governments to monitor and analyse data globally and is licensed for fair use under Creative Commons Attribution 4.0 International. You can find more about this data at this link: https://github.com/CSSEGISandData/COVID-19.

This report focuses on four data sets in the JHU CSSE COVID-19: Global Cases, Global Deaths, US Cases, and US Deaths updated till 21/04/2022. We will tidy the data to suit the need of the further analysis as needed during the report.

Report has 5 Parts:

1. Part 1: Tidying the data
2. Part 2: Visualizations and Analysis
3. Part 3: Model
4. Part 4: Bias Sources
5. Conclusion

Part 1: Tidying the data

Loading up the libraries required for this report.

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
```

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.1.3
```

Importing the URLs for the required data sets from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

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

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_raw,file_names)

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

Reading the data into datasets for further processing.

```
global_cases <- read_csv(urls[2])
```

```
## Rows: 284 Columns: 826
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (824): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
US_cases <- read_csv(urls[1])
```

```
## Rows: 3342 Columns: 833
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (827): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20,...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
## Rows: 3342 Columns: 834

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (828): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24/...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
global_deaths <- read_csv(urls[4])
```

```
## Rows: 284 Columns: 826

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (824): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
## Rows: 4316 Columns: 12

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

In the next two steps we will focus on the Global Data only. We will tidy up the two data sets and merge them into one dataset "global" while adding the population variable from UID Data set for further analysis in Part 2.

Geo-Location variables has been removed along with others considered not necessary for the ansalysis.

```
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)) %>%
  filter(cases >  0)
```

```
## Joining, by = c("Province/State", "Country/Region", "date")
```

Adding the population variable from UID Data set so that US and Global could have similar variables.

```
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)
```

Similar to global data in the following steps will be carried out to combine the US data while removing the variables containing geo-location data and others not considered necessary for the analysis in Part 2.

```
US_cases <- US_cases %>%
  pivot_longer(cols = -c(UID:Combined_Key),
                         names_to = 'date',
                         values_to = 'cases')
US_cases <- US_cases %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat,Long_))

US_deaths <- US_deaths %>%
```

```
  pivot_longer(cols = -c(UID:Population),
               names_to = 'date',
               values_to = 'deaths') %>%
  select(Admin2:deaths) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat,Long_))

US_total <- US_cases %>%
  full_join(US_deaths)
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key", "date")
```

US Data set scrictly contains the data from the US as a whole while US_State have state-wise data which we will further use in Part 2 and Part 3.

```
US_state <- US_total %>%
  group_by(Province_State, Country_Region, date) %>%
  summarise(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()
```

```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can override using the `
```

```
US <- US_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()
```

```
## `summarise()` has grouped output by 'Country_Region'. You can override using the `.groups` argument.
```

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

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

Part 2: Visualizations and Analysis

We will divide this part into parts with each part having spaces for analysis and visualizations to answer the questions of interest for Global and US data separately.

Analysis Part 1: Global Data

Major questions of interest for the Global Data:

1. What are the Top 10 days when most new cases were recorded by a country during the pandemic?
2. What are the Top 10 days when most new deaths were recorded by a country during the pandemic?
3. Which countries recorded highest new deaths during the pandemic?
4. Which countries have most cases and which countries has most deaths during the pandemic? Are they same or are they totally different countries? If they are different what could be the source of difference which caused expected linear model of more cases leading to more deaths differ?
5. Further analyzing the point 4 and draw various graphs to understand the finding of point 4.

Finding Top 10 additions to the cases of countires around the world on a particular day.

```
summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:214791      Length:214791      Min.   :2020-01-22   Min.   :       1
##  Class :character   Class :character   1st Qu.:2020-09-18   1st Qu.:     659
##  Mode  :character   Mode  :character   Median :2021-04-03   Median :    9381
##                                        Mean   :2021-03-31   Mean   :  567897
##                                        3rd Qu.:2021-10-13   3rd Qu.:  140588
##                                        Max.   :2022-04-22   Max.   :80952268
##
##      deaths          Population        Combined_Key
##  Min.   :     0   Min.   :8.090e+02   Length:214791
##  1st Qu.:     4   1st Qu.:8.696e+05   Class :character
##  Median :   121   Median :7.203e+06   Mode  :character
##  Mean   : 10704   Mean   :2.936e+07
##  3rd Qu.:  2292   3rd Qu.:2.983e+07
##  Max.   :991169   Max.   :1.380e+09
##                   NA's   :4161
```

```
global %>% mutate(new_cases = cases - lag(cases)) %>%
  top_n(10, wt = new_cases) %>%
  select(Country_Region, date, new_cases) %>%
  arrange(desc(new_cases))
```

```
## # A tibble: 10 x 3
##    Country_Region date       new_cases
##    <chr>          <date>         <dbl>
##  1 US             2022-01-10   1383823
##  2 US             2022-01-18   1129422
##  3 US             2022-01-03   1044895
##  4 US             2022-01-24    921994
##  5 US             2022-01-19    908016
##  6 US             2022-01-14    879999
##  7 US             2022-01-07    869692
##  8 US             2022-01-13    861302
##  9 US             2022-01-12    848569
## 10 United Kingdom 2022-01-31    847371
```

Nine of the top ten additions to the cases belong to US and all of them happeed in the month of January 2022. This corrospond to the reports from other sources as new cases in US and UK did rise a lot during that time due to the recent COVID wave arising from new variant of COVID.

Now we will do a similar analysis to find the top 10 recordings of the addition to the number of deaths from the countries around the world

```
global %>% mutate(new_deaths = deaths - lag(deaths)) %>%
  top_n(10, wt = new_deaths) %>%
  select(Country_Region, date, new_deaths) %>%
  arrange(desc(new_deaths))
```

```
## # A tibble: 10 x 3
##    Country_Region date       new_deaths
##    <chr>          <date>          <dbl>
##  1 Chile          2022-03-21      11447
##  2 Ecuador        2021-07-20       8786
##  3 India          2021-06-10       7374
##  4 India          2021-05-18       4529
##  5 India          2021-05-23       4454
##  6 US             2021-01-20       4431
##  7 US             2021-01-12       4368
##  8 India          2021-05-17       4329
##  9 Mexico         2021-06-01       4272
## 10 India          2021-05-20       4209
```

Following our previous bias that number of deaths and number of cases fall in a linear model we believed that top 10 additions to the deaths for countries would be similar to the top 10 additions to the cases, but the table above is showing totally different stories.

First of all, US is taking 2 slots in the top 10 this time, and that too happened last year with none of the dates being even close to the dates when US recorded most new cases.

After this, we can see that India takes dominant place here with 5 slots and rest being filled by South American nations like Mexico, Chile, and Ecuador.

This shows that even if the developed nations like the US recorded most cases they are better equipped to prevent deaths of their citizens as compared to the developing nations.

Next, we will find the 10 countries with most cases and deaths until 21-04-2022.

```
global %>%
  select(-c("Province_State", "Combined_Key", "date")) %>%
  group_by(Country_Region) %>%
  top_n(1, wt = cases) %>%
  arrange(desc(cases)) %>%
  filter(!duplicated(Country_Region))
```

```
## # A tibble: 198 x 4
## # Groups:   Country_Region [198]
##    Country_Region    cases deaths Population
##    <chr>             <dbl>  <dbl>      <dbl>
##  1 US             80952268 991169  329466283
##  2 India          43054952 522149 1380004385
##  3 Brazil         30338697 662802  212559409
```

```
##  4 France         27416764 141763   65249843
##  5 Germany        24141333 134155   83155031
##  6 United Kingdom 21933206 173352   67886004
##  7 Russia         17855661 367036  145934460
##  8 Korea, South   16830469  22024   51269183
##  9 Italy          16008181 162466   60461828
## 10 Turkey         15013616  98660   84339067
## # ... with 188 more rows
```

```
global %>%
  select(-c("Province_State", "Combined_Key", "date")) %>%
  group_by(Country_Region) %>%
  arrange(desc(deaths)) %>%
  top_n(1, wt =  deaths) %>%
  filter(!duplicated(Country_Region))
```

```
## # A tibble: 198 x 4
## # Groups:   Country_Region [198]
##    Country_Region    cases deaths Population
##    <chr>             <dbl>  <dbl>      <dbl>
##  1 US             80952268 991169  329466283
##  2 Brazil         30338697 662802  212559409
##  3 India          43054952 522149 1380004385
##  4 Russia         17855661 367036  145934460
##  5 Mexico          5731635 324033  127792286
##  6 Peru            3559343 212724   32971846
##  7 United Kingdom 21933206 173352   67886004
##  8 Italy          16008181 162466   60461828
##  9 Indonesia       6043246 156040  273523621
## 10 France         27416764 141763   65249843
## # ... with 188 more rows
```

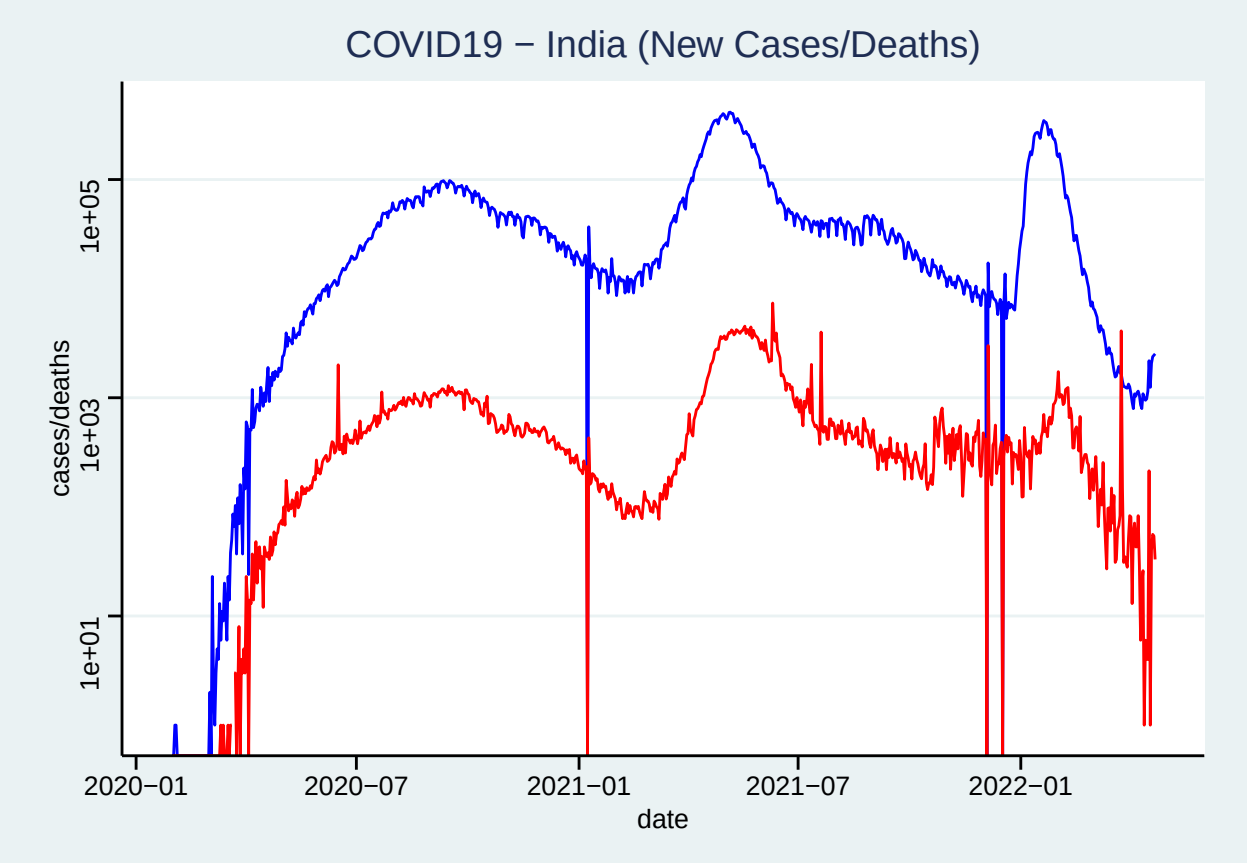
Just like our previous analysis we could see that while the US and Europe populates the top 10 ranking for most number of cases during the pandemic developing nations of Asia and South America populates the top 10 ranking of most number of deaths recorded by a nation.

This is a clear indication that people living in different countries had different treatments available to them with quality of treatments depending upon the economic prosperity of a nation they were living in.

Now we will move to further analyse the difference in cases and deaths between developed nations and developing nations. For this we will consider US and India, both of which fall under top 10 countries with most cases and death.

```
global %>%
  filter(Country_Region == "India") %>%
  select(-c(Province_State,Population, Combined_Key)) %>%
  mutate(cases = cases - lag(cases),
         deaths = deaths - lag(deaths)) %>%
  ggplot()+
  geom_line(aes(x = date, y = cases), color = "blue")+
  geom_line(aes(x = date, y = deaths), color = "red")+
  scale_y_log10()+
  theme_stata()+
  ggtitle("COVID19 - India (New Cases/Deaths)")+
  ylab("cases/deaths")
```

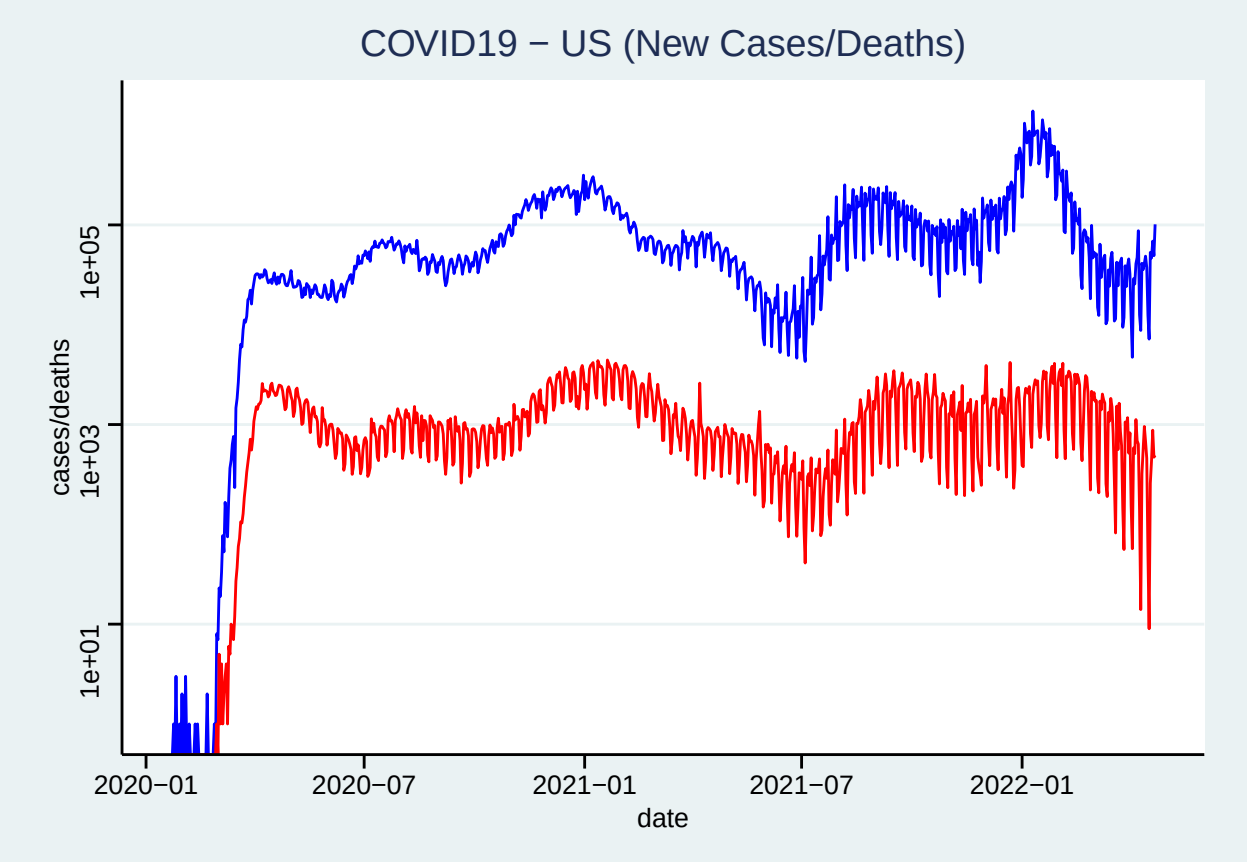
```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 1 row(s) containing missing values (geom_path).
```



From the graph above we can see that peaks in cases and deaths came together during the new waves. In third wave there is a spike in the recording of the deaths after the peak possibly because of the delay in collecting and entering the data. But overall, both cases and deaths follow each other in a visible linear relationship in developing country like India.

```
global %>%
  filter(Country_Region == "US") %>%
  select(-c(Province_State,Population, Combined_Key)) %>%
  mutate(cases = cases - lag(cases),
         deaths = deaths - lag(deaths)) %>%
  ggplot()+
  geom_line(aes(x = date, y = cases), color = "blue")+
  geom_line(aes(x = date, y = deaths), color = "red")+
  scale_y_log10()+
  theme_stata()+
  ggtitle("COVID19 - US (New Cases/Deaths)")+
  ylab("cases/deaths")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 1 row(s) containing missing values (geom_path).
```



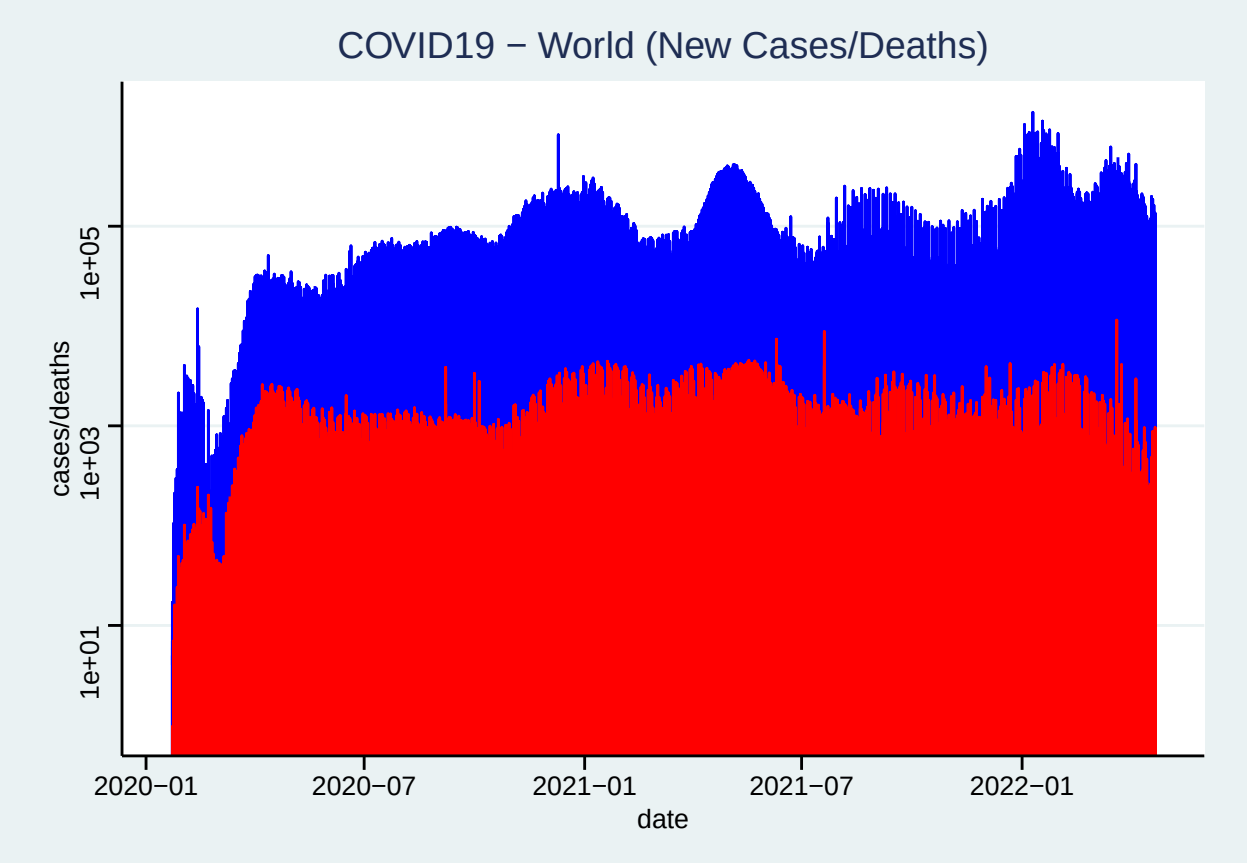

```
global %>%
  select(-c(Province_State,Population, Combined_Key)) %>%
  mutate(cases = cases - lag(cases),
         deaths = deaths - lag(deaths)) %>%
  ggplot()+
  geom_line(aes(x = date, y = cases), color = "blue")+
  geom_line(aes(x = date, y = deaths), color = "red")+
  scale_y_log10()+
  theme_stata()+
  ggtitle("COVID19 - World (New Cases/Deaths)")+
  ylab("cases/deaths")
```

```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 30 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 13 row(s) containing missing values (geom_path).
```



While it is tempting to consider graphs of US and India as similar: deaths increase when cases increase, we can see that the US graph have suppressed peaks after the first wave. This means that while in India number of deaths generally increase with number of cases with both reaching peaks together in US deaths is relative flat.

This difference can be attributed to the fact that even if whole world face the similar intensity during the first wave the developed countries were able to reduce a great number of deaths with their superior technology and health care, if we compare them to the developing nations which also improved their abilities to deal with the pandemic in later waves but still fell short of reducing number of deaths like developed nations.

And if we look at graph for the whole world we can see that even though most of the times cases and deaths reach the peaks together the number of deaths actually fell during the last peak of the cases. This is optimistic because it shows that even though the number of cases may reach a new peak world is better equipped to reduce the number of deaths from COVID19 due to the collective efforts of the humanity as a whole.

Analysis Part 2: US Data

In this section we will analyse the US Data. It contains many workings carried out during the course along with the additional analysis.

Questions of interest for the US Data:

1. What are the top 5 states with most cases and deaths during the pandemic?
2. Which states fall in top 5 places for the highest number of new deaths and new cases during the pandemic?
3. Which states have most deaths per millions and cases per millions in the US? Is there a visible linear relationship here and if not, what are the observations?
4. Graph showing new cases and deaths of US from the Global data set and US data set is similar? And whether state of Washington have different graph as compared to the US as a whole?

```
US_state %>%
  group_by(Province_State) %>%
  summarize(cases = sum(cases), deaths = sum(deaths)) %>%
  top_n(5, wt = cases) %>%
  arrange(desc(cases))
```

```
## # A tibble: 5 x 3
##   Province_State      cases   deaths
##   <chr>               <dbl>    <dbl>
## 1 California     2633141593 34786991
## 2 Texas          2075202395 32325250
## 3 Florida        1745916491 25681046
## 4 New York       1439451662 34904098
## 5 Illinois        935096891 15292594
```

```
US_state %>%
  group_by(Province_State) %>%
  summarize(cases = sum(cases), deaths = sum(deaths)) %>%
  top_n(5, wt = deaths) %>%
  arrange(desc(deaths))
```

```
## # A tibble: 5 x 3
##   Province_State      cases   deaths
##   <chr>               <dbl>    <dbl>
## 1 New York       1439451662 34904098
## 2 California     2633141593 34786991
## 3 Texas          2075202395 32325250
## 4 Florida        1745916491 25681046
## 5 New Jersey      667082506 16951397
```

From this basic analysis we can see that top 5 states in both categories are different with New York recording most deaths when the it had fourth highest cases during the period. This shows that quality of COVID treatment varies across the States even in developed countries like the US.

```
US_state %>%
  group_by(Province_State) %>%
  summarize(new_cases = max(new_cases), new_deaths = max(new_deaths)) %>%
  top_n(5, wt = new_cases) %>%
  arrange(desc(new_cases))
```

```
## # A tibble: 5 x 3
```

```
##   Province_State new_cases new_deaths
##   <chr>              <dbl>      <dbl>
## 1 California        207094       1174
## 2 Florida           150251       1554
## 3 New York          132093       1271
## 4 North Carolina    119101        181
## 5 Michigan           98927        566
```

```
US_state %>%
  group_by(Province_State) %>%
  summarize(new_cases = max(new_cases), new_deaths = max(new_deaths)) %>%
  top_n(5, wt = new_deaths) %>%
  arrange(desc(new_deaths))
```

```
## # A tibble: 5 x 3
##   Province_State new_cases new_deaths
##   <chr>              <dbl>      <dbl>
## 1 Missouri           26220       2441
## 2 Tennessee          22098       2116
## 3 Oklahoma           27754       1716
## 4 Florida           150251       1554
## 5 New York          132093       1271
```

This additional analysis was done to find out about the highest new cases/deaths recorded on a single day across the states. Unlike the previous analysis where mostly same states occupied the top 5 slots here we can see other states like Missouri and Michigan coming in the top 5. Especially Missouri and Tennessee which recorded over 2000 deaths on a single day, around 10% of the maximum new cases recorded during the pandemic.

```
US_state %>%
  group_by(Province_State) %>%
  summarize(deaths_per_mill = max(deaths_per_mill),
            cases_per_mill = max(cases/Population*1000000)) %>%
  top_n(10, deaths_per_mill) %>%
  arrange(desc(deaths_per_mill))
```

```
## # A tibble: 10 x 3
##    Province_State deaths_per_mill cases_per_mill
##    <chr>                    <dbl>          <dbl>
##  1 Mississippi              4179.        267742.
##  2 Arizona                  4101.        277408.
##  3 Oklahoma                 3997.        262524.
##  4 Alabama                  3985.        264958.
##  5 Tennessee                3826.        296779.
##  6 West Virginia            3816.        278881.
##  7 Arkansas                 3767.        276646.
##  8 New Jersey               3757.        252269.
##  9 Louisiana                3705.        252029.
## 10 Michigan                 3598.        241464.
```

```
US_state %>%
  group_by(Province_State) %>%
```

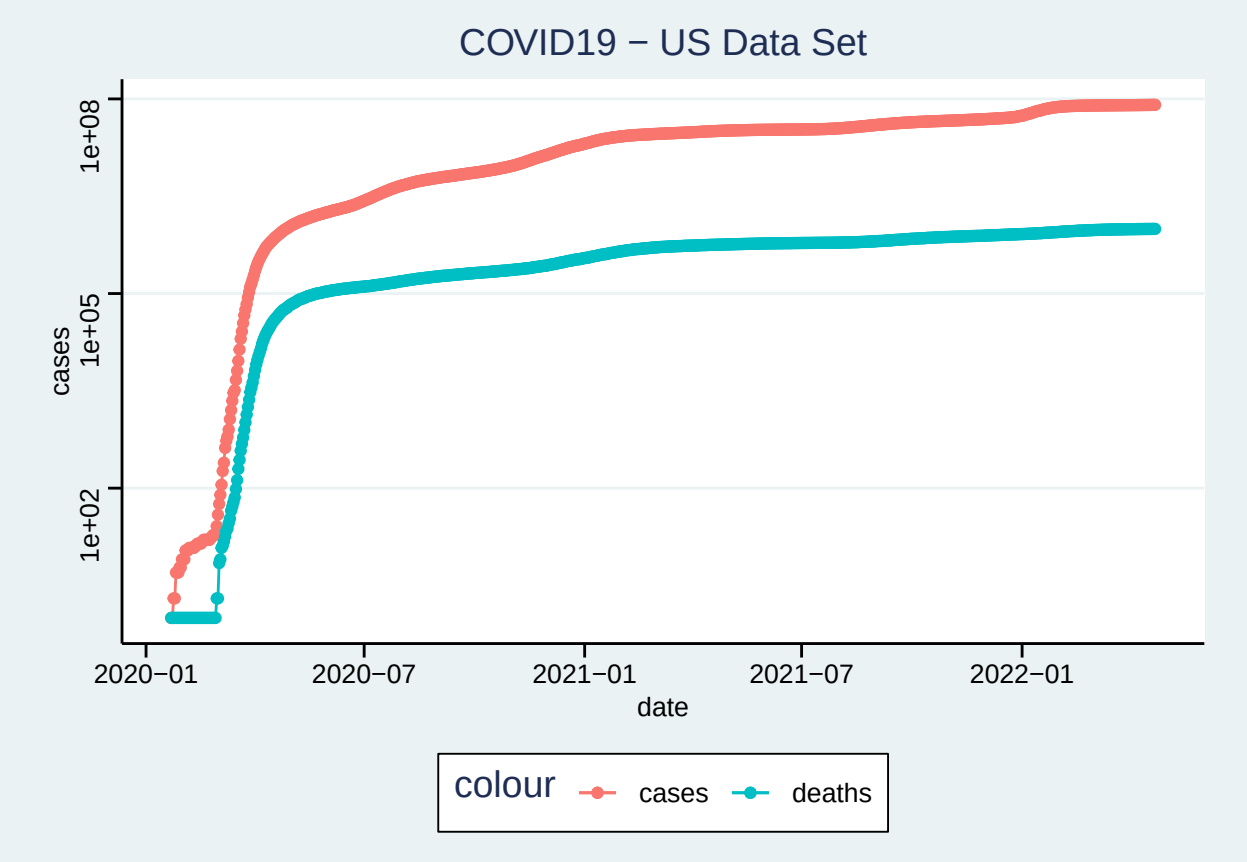
```
  summarize(deaths_per_mill = max(deaths_per_mill),
            cases_per_mill = max(cases/Population*1000000)) %>%
  top_n(10, cases_per_mill) %>%
  arrange(desc(cases_per_mill))
```

```
## # A tibble: 10 x 3
##    Province_State deaths_per_mill cases_per_mill
##    <chr>                    <dbl>          <dbl>
##  1 Rhode Island             3334.        347611.
##  2 Alaska                   1684.        339417.
##  3 North Dakota             2966.        315871.
##  4 Tennessee                3826.        296779.
##  5 Kentucky                 3438.        296472.
##  6 Guam                     2156.        291903.
##  7 Utah                     1478.        290259.
##  8 South Carolina           3444.        285915.
##  9 West Virginia            3816.        278881.
## 10 Arizona                  4101.        277408.
```

This additional analysis to find more about the states like Mississippi which had highest number of new deaths recorded on a single day. From the above table we could see that while many international media organizations highlighted the New York as most effected city of US it neither have most cases or deaths as per the data.

Lesser known states appears to have faced the most brunt of deaths and cases per millions due to COVID in US. There can be many reasons why States like Utah, which have more cases per million than Mississippi, but have only 1478.497 deaths per million, which pales in front of the 4178.554 deaths per million in Mississippi. Whatever the reasons may be, we can be sure that even in developed nations like US quality of the treatment for COVID varied a lot.

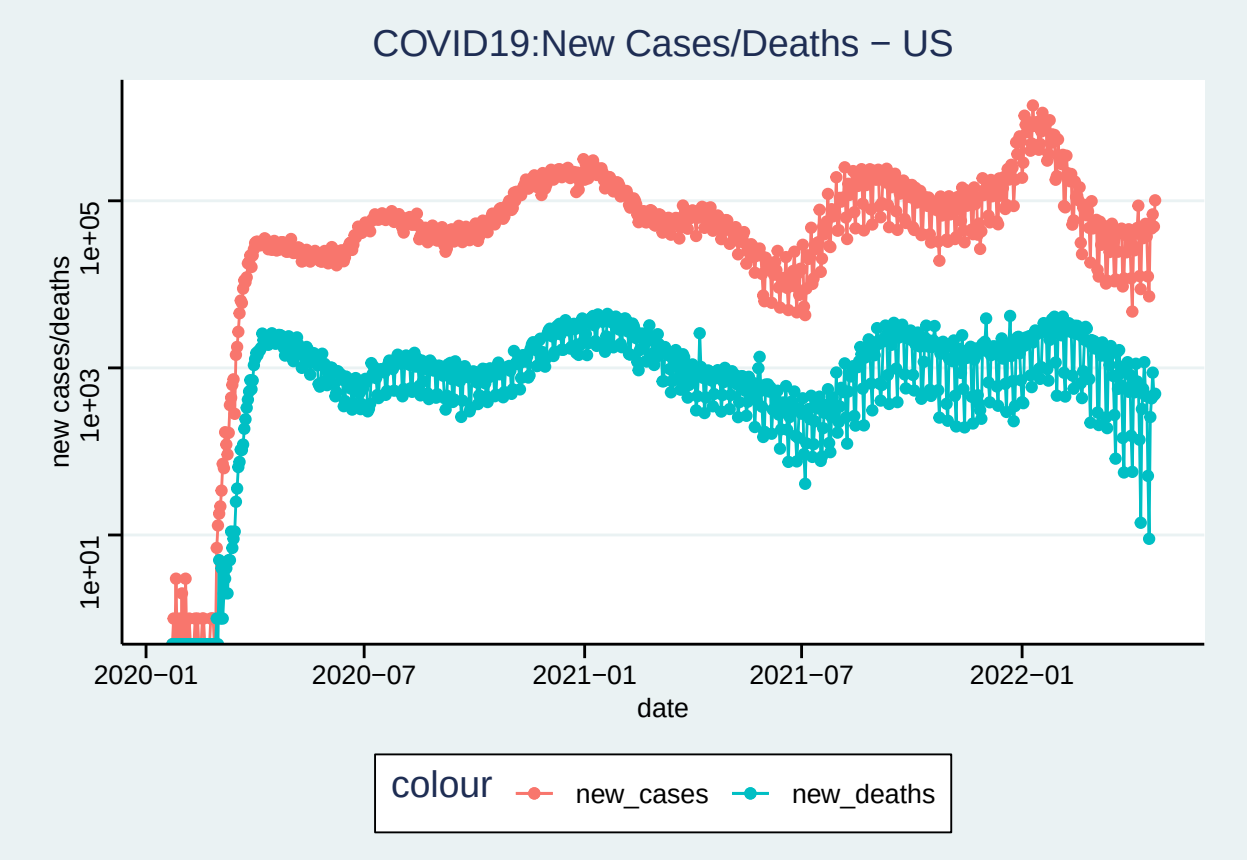
```
US %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  ggtitle('COVID19 - US Data Set')+
  theme_stata()
```




```
US %>%
  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() +
  ggtitle('COVID19:New Cases/Deaths - US')+
  ylab("new cases/deaths")+
  theme_stata()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1 row(s) containing missing values (geom_path).

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

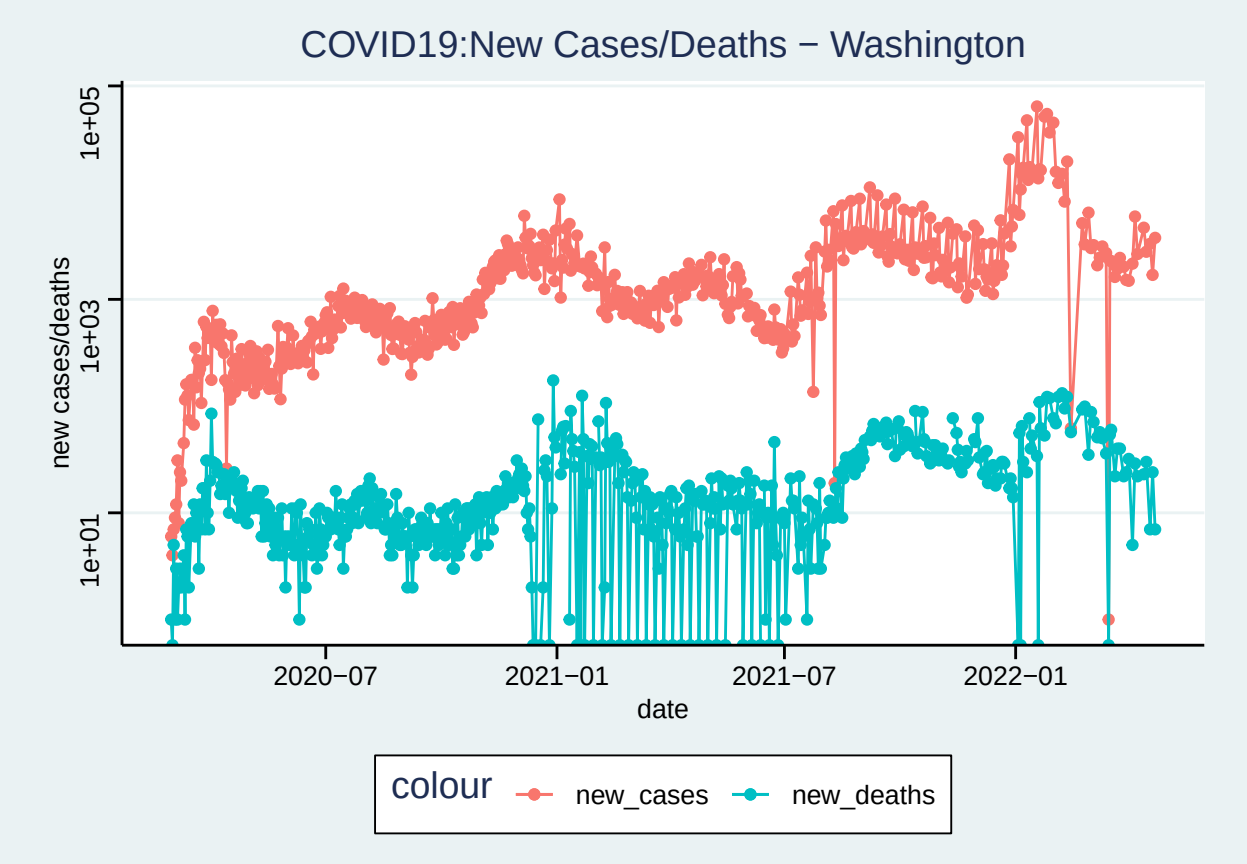


This analysis was to check if both data set have same data for the US and as we can confirm it after seeing the similar graphs for US new deaths and cases from both of the data sets, this was mainly done to check the integrity of the source data.

```
US_state %>%
  filter(new_cases > 0, Province_State == "Washington") %>%
  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() +
  ggtitle("COVID19:New Cases/Deaths - Washington")+
  ylab("new cases/deaths")+
  theme_stata()
```

```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```

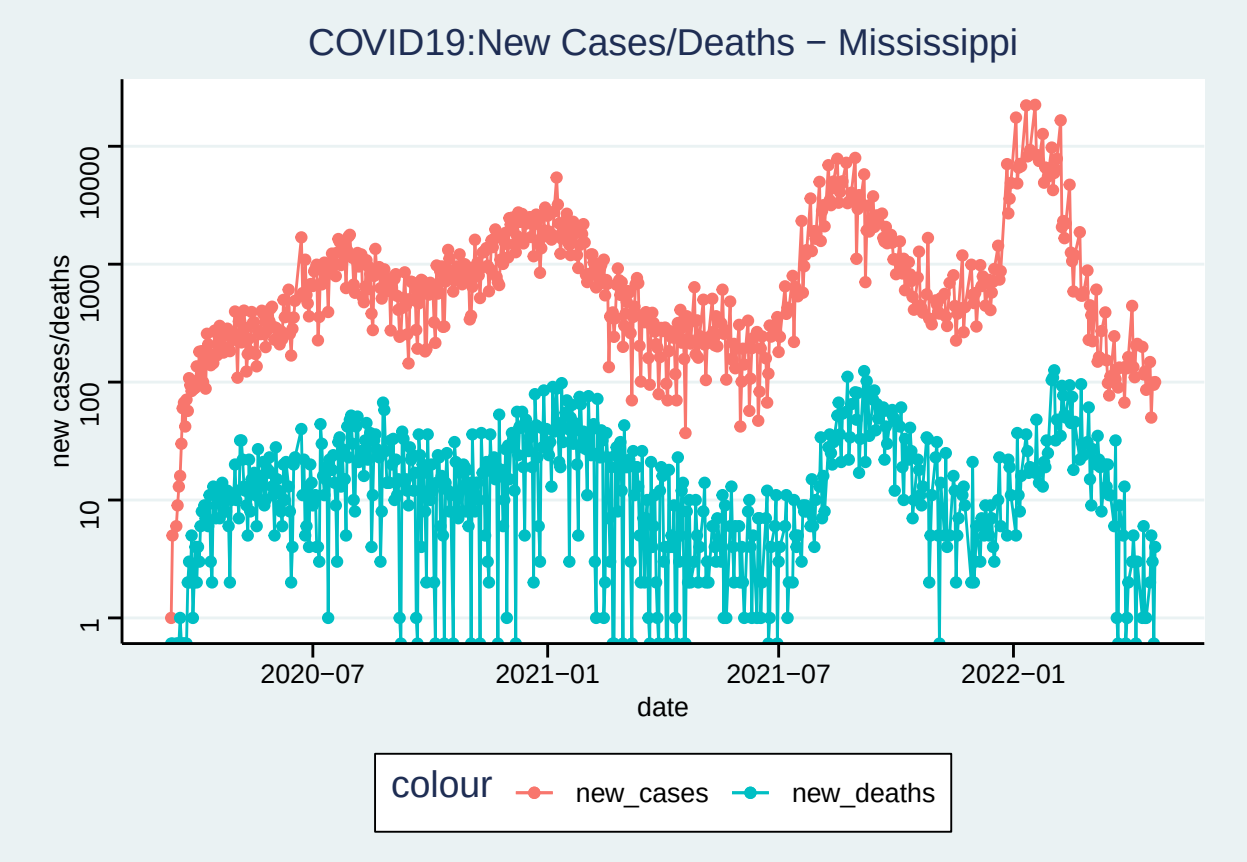


State of Washington have similar pattern to the the US total graph with new deaths line appearing as not increasing corresponding to the new cases.

Now, let's see how the state of Mississippi, we found as one of the most affected state in US.

```
US_state %>%
  filter(new_cases > 0, Province_State == "Mississippi") %>%
  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() +
  ggtitle("COVID19:New Cases/Deaths - Mississippi")+
  ylab("new cases/deaths")+
  theme_stata()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis
```



Compared to Washington and US total graph, graph of Mississippi is showing more peaks in cases and deaths with more evident linear relationship between new cases and deaths as compared to the previous two graphs. This also shows that while US in general did good job in curbing the deaths after initial wave states like Mississippi were not able to do similar level of job in this regard and essentially produced a graph similar to the graphs produced by the developing countries like India.

Part 3: Model

In this part we will create a model to predict cases per thousand on the basis of deaths per thousands. This model is similar to the one shown in the course but as we can see from the previous analysis US has done a considerable good job to reduce the linear relationship between cases and deaths due to COVID during the period.

To check whether a linear relationship really changed during the period we will compare two models: US_model_firstwave with only data until 30-06-2021 and US_model with complete data.

```
US_model_firstwave <- US_state %>%
  group_by(Province_State) %>%
  filter(date < "2021-07-01")%>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population), cases_per_thou = cases/population*1000,
            deaths_per_thou = deaths/population*1000) %>%
  filter(cases>0, population >0)
```

```
model_firstwave <- lm(cases_per_thou ~ deaths_per_thou, data = US_model_firstwave)
summary(model_firstwave)
```

```
##
## Call:
## lm(formula = cases_per_thou ~ deaths_per_thou, data = US_model_firstwave)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.194 -15.927  -0.536  11.227  61.689
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       43.312      7.030   6.161 1.00e-07 ***
## deaths_per_thou   33.113      4.074   8.128 7.01e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.13 on 53 degrees of freedom
## Multiple R-squared:  0.5548, Adjusted R-squared:  0.5464
## F-statistic: 66.06 on 1 and 53 DF,  p-value: 7.005e-11
```

```
US_model <- US_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population), cases_per_thou = cases/population*1000,
            deaths_per_thou = deaths/population*1000) %>%
  filter(cases>0, population >0)
```

```
model <- lm(cases_per_thou ~ deaths_per_thou, data = US_model)
summary(model)
```

```
##
## Call:
## lm(formula = cases_per_thou ~ deaths_per_thou, data = US_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.382 -21.360  -0.763  17.775 127.831
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      162.428     15.447  10.515 1.12e-14 ***
## deaths_per_thou   29.187      5.328   5.478 1.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.07 on 54 degrees of freedom
## Multiple R-squared:  0.3572, Adjusted R-squared:  0.3453
## F-statistic: 30.01 on 1 and 54 DF,  p-value: 1.155e-06
```

We can see from the data that while in the first wave model based on the deaths and cases had a respectable adjusted R Square of 0.5464 and could be considered quite a good fit the model with updated data has adjusted R Square of only 0.3453.

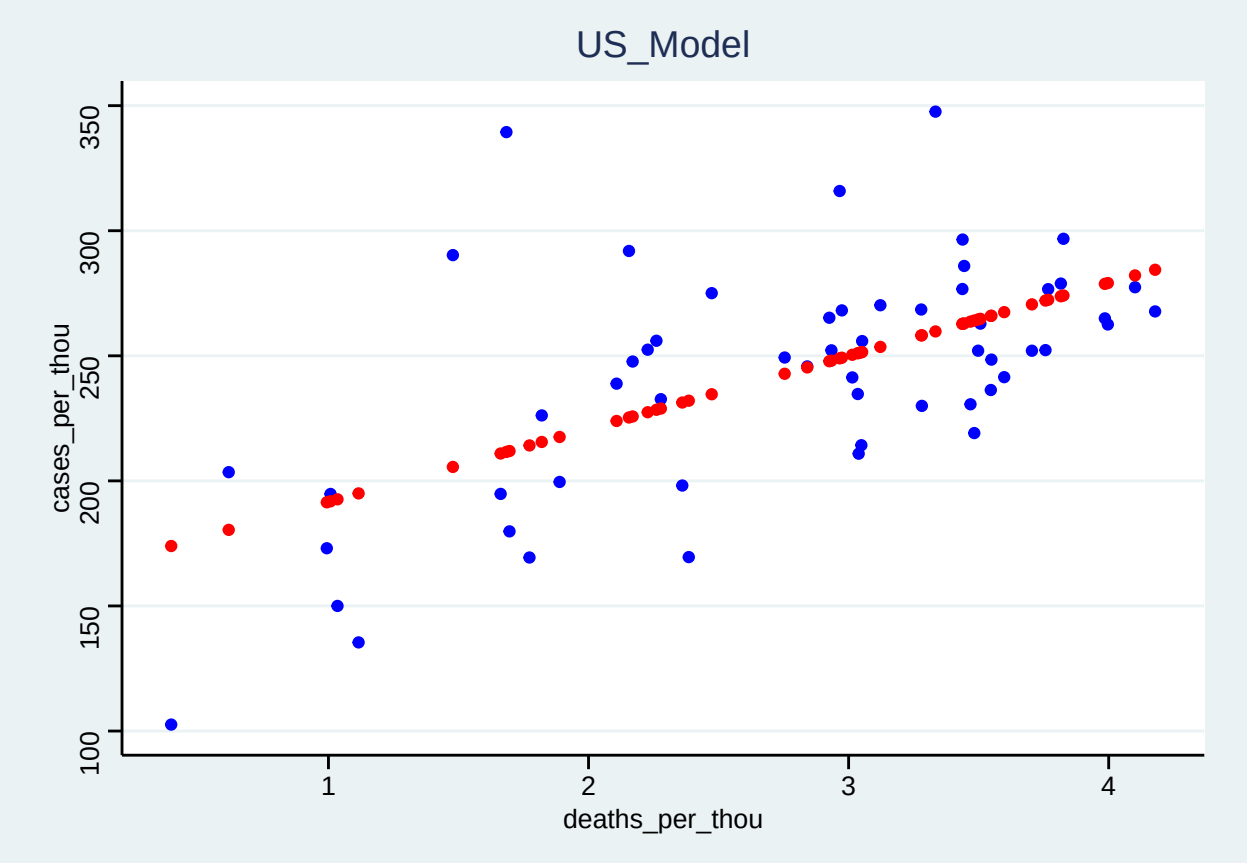
This shows clearly that while in the first wave deaths increased with the cases in the other US was able to eliminate this linear relationship to a great degree. It could be due to the influx of vaccines and better

understanding of the COVID in general beside many more variables we can not account for with the data we have.

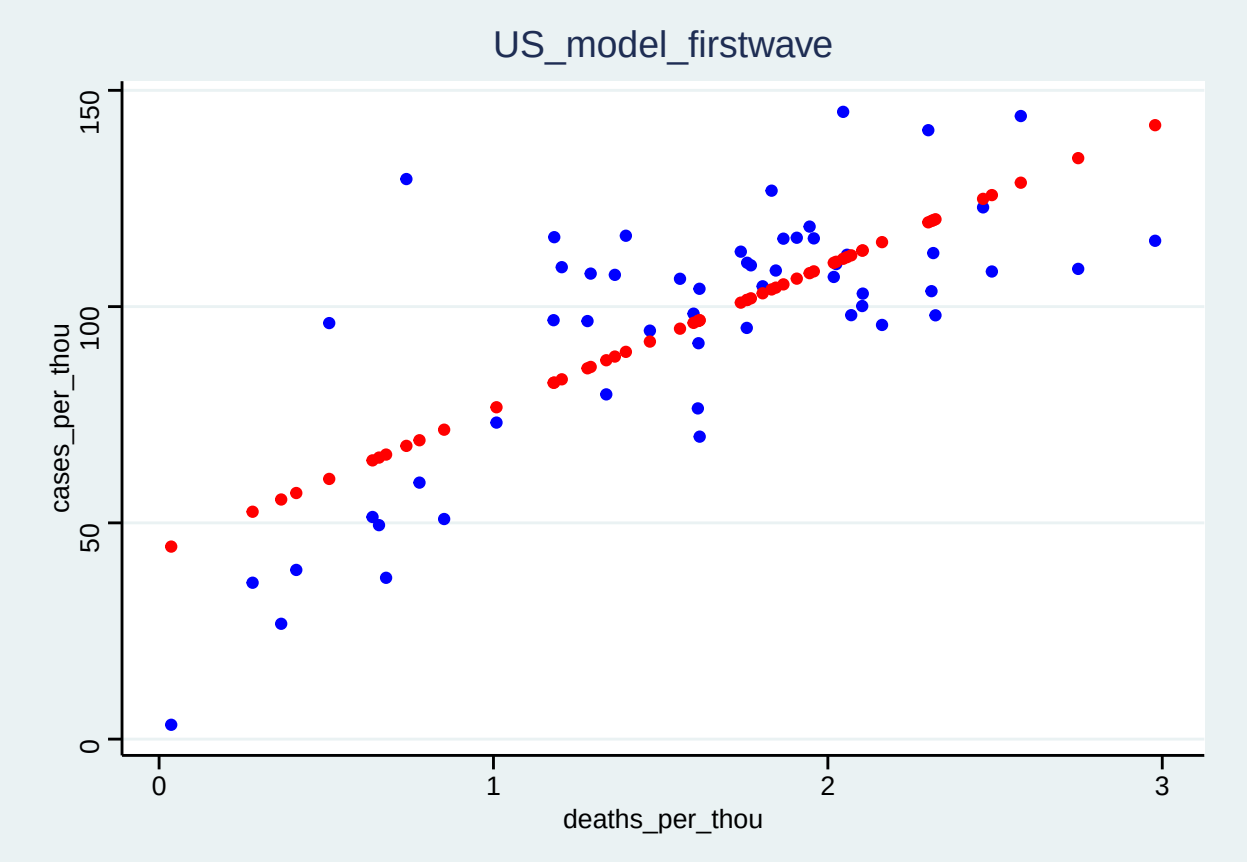
Now, we will create a prediction on the basis of the two models and compare the graphs resulting from them.

```
US_state_pred <- US_model %>% mutate(pred_cases =predict(model))

US_firstwave_pred <- US_model_firstwave  %>% mutate(pred_cases =predict(model_firstwave))
```

```
US_state_pred %>%
  ggplot()+
  geom_point(aes(x = deaths_per_thou, y = cases_per_thou), color = "blue")+
  geom_point(aes(x = deaths_per_thou, y = pred_cases), color = "red")+
  ggtitle("US_Model")+
  theme_stata()
```




```
US_firstwave_pred %>%
  ggplot()+
  geom_point(aes(x = deaths_per_thou, y = cases_per_thou), color = "blue")+
  geom_point(aes(x = deaths_per_thou, y = pred_cases), color = "red")+
  ggtitle("US_model_firstwave")+
  theme_stata()
```



From the graphs we could see that the graph based on first wave showed more strong relationship between deaths and cases with less number of outliers as compared to the graph based on the model of complete US data.

US_model shows a very weak linear relationship due to which we can not predict number of cases on the basis of number of deaths with quite a confidence. We should look for other models which could fit the data in a much better way.

Note: The above finding is not conclusive and may change if data sets from other sources are also included in the analysis.

Special Note:

During the report various analysis and visualizations beyond the class have been carried out, especially in the Part 2 and Part 3.

Examples of such additional analysis are complete separate part for analyzing and visualizing Global Data, finding more about the states with most cases and deaths in US, and comparing models used for predicting number of cases on the basis of deaths based on the complete data until 21-04-2022 and the the data only for the first wave until 30-06-2021.

Part 4: Bias Sources

Bias identification is an important part of any Data Science Project. Below are the Bias Sources identified throughout the whole reporting process:

1. A preconceived bias that there exist a linear relationship between deaths and cases due to COVID19 across the world. This led me to further analyse this aspect across different levels.

2. A preconceived bias that New York was the most affected city/province in US due to COVID19. This bias originated from the various media reports I obtained in my country which depicted severity of situation in New York specifically.
3. This report is an extended version of the workings I carried out during a course. While most of the analysis and visualizations in this report are different from the ones in the course this report is based on personal workings I carried out along with the instructor.
4. Various variables pertaining to identification keys used by other sources and geo-location data has been removed because they were deemed not important for the analysis. Special notes have been included where such filters are used.

Part 5: Conclusion

JHU CSSE COVID-19 Data is a huge data set which is nothing short of being a standardized global level data set for the COVID19. During our analysis we were able to obtain a lot of interesting findings:

1. While developed countries took most of the spots in Top 10 countries with most number of COVID19 cases the developing countries took most of the spots in Top 10 countries with most number of deaths due to COVID19.
2. Developing Country like India shows more evidences of the existence of a linear relationship between cases and deaths as compared to the Developed Country like US. Implying that across the different waves of COVID19 developed countries were better equipped in reducing the number of deaths due to COVID19.
3. That being said, the world as a whole is better equipped in reducing the deaths even when cases are increasing with incremental waves. Even though deveoloped countries have an edge developing countries are progressing to reduce the linear relationship between cases and deaths.
4. Severity of COVID19 varies across the US with lesser known states bearing the most number of deaths per millions. This shows that just like difference in quality of treatment available to people on a global scale in US too there exist a difference in quality of treatment available to people across the different States.
5. There existed a good linear relationship between cases and deaths per thousand in US during the first wave but as time passed on due to the Vaccination Programs and other factors US has eliminated this linear relationship to a commendable level. And same should be true for other developed nations.

All the above conclusions are based on the process followed in the report and are subjected to the the bias sources identified in Part 4. Hence, we must be cautious in accepting these conclusions as they could change if one could use data from other sources and introduces the variables we didn't consider in this report.