COVID-19 Project

Jingjing Huang

2022-06-12

COVID-19 Project

This is an analysis report based on COVID-19 data with data source of Johns Hopkins University website <"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/">.

Data Import

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(lubridate)
library(dplyr)

url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov
file_names <- c("time_series_covid19_confirmed_US.csv","time_series_covid19_confirmed_global.csv","time_
urls <- str_c(url_in, file_names)
global_cases<- read_csv(urls[2])
US_cases<- read_csv(urls[1])
US_deaths<- read_csv(urls[3])
global_deaths<- read_csv(urls[4])
```

Data Transforming and Cleaning Up

First, I would like to tidy and clean the **global** data set : global_cases and global_deaths.I put the variable in own column. Also, I delete the column: Lat and Long for creating a more tidy data set.

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

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

global <- global_cases %>%
full_join(global_deaths) %>%
rename(Country_Region = `Country/Region`, Province_State = `Province/State`) %>%
mutate(date = mdy(date))
global <- global %>% filter(cases >0)
```

Then, adding the population information into the same global data set.

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

uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U
uid <- read_csv(uid_lookup_url  ) %>%
select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))

global <- global %>%
left_join(uid, by = c("Province_State", "Country_Region")) %>%
select(-c(UID, FIPS)) %>%
select(Province_State, Country_Region, date, cases, deaths, Population, Combined_Key)
summary(global)
```

```
##  Province_State     Country_Region          date                cases
##  Length:229203      Length:229203      Min.   :2020-01-22   Min.   :       1
##  Class :character   Class :character   1st Qu.:2020-10-01   1st Qu.:     721
##  Mode  :character   Mode  :character   Median :2021-04-29   Median :   11285
##                                        Mean   :2021-04-26   Mean   :  648540
##                                        3rd Qu.:2021-11-21   3rd Qu.:  161232
##                                        Max.   :2022-06-12   Max.   :85515795
##
##      deaths          Population        Combined_Key
##  Min.   :      0   Min.   :8.090e+02   Length:229203
##  1st Qu.:      6   1st Qu.:8.696e+05   Class :character
##  Median :    134   Median :7.133e+06   Mode  :character
##  Mean   :  11433   Mean   :2.928e+07
##  3rd Qu.:   2536   3rd Qu.:2.914e+07
##  Max.   :1011275   Max.   :1.380e+09
##                    NA's   :4569
```

Same as tidying the above global data, cleaning and tidy the US data set in following below:

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

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

US <- US_cases %>%
full_join(US_deaths)
US<- US %>% filter(cases >= 0)
summary(US)
```

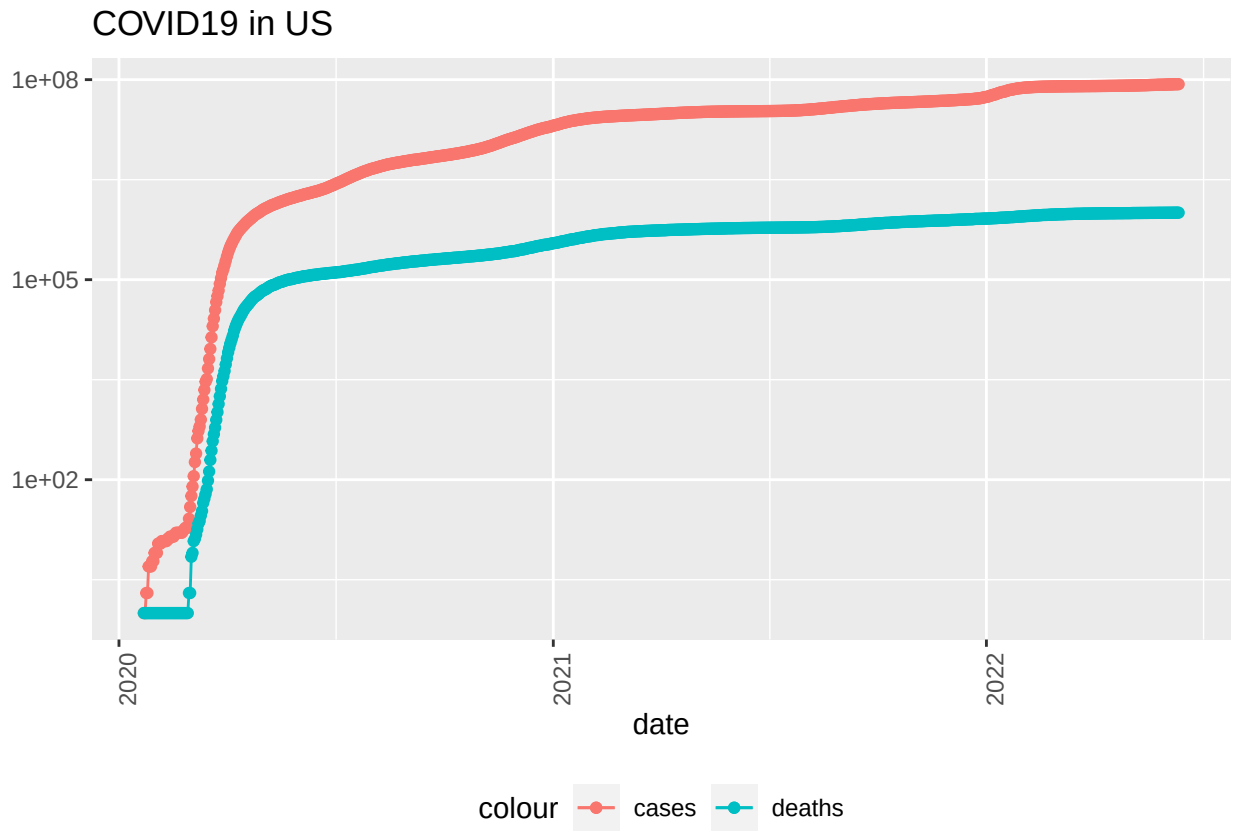
```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:2917564     Length:2917564     Length:2917564     Length:2917564
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       date                cases           Population           deaths
##  Min.   :2020-01-22   Min.   :      0   Min.   :       0   Min.   :    0.0
##  1st Qu.:2020-08-27   1st Qu.:    135   1st Qu.:    9917   1st Qu.:    1.0
##  Median :2021-04-02   Median :   1389   Median :   24909   Median :   24.0
##  Mean   :2021-04-01   Mean   :   9488   Mean   :   99604   Mean   :  145.8
##  3rd Qu.:2021-11-06   3rd Qu.:   5300   3rd Qu.:   64979   3rd Qu.:   89.0
##  Max.   :2022-06-12   Max.   :3025694   Max.   :10039107   Max.   :32201.0
```

Visualization & Analysis

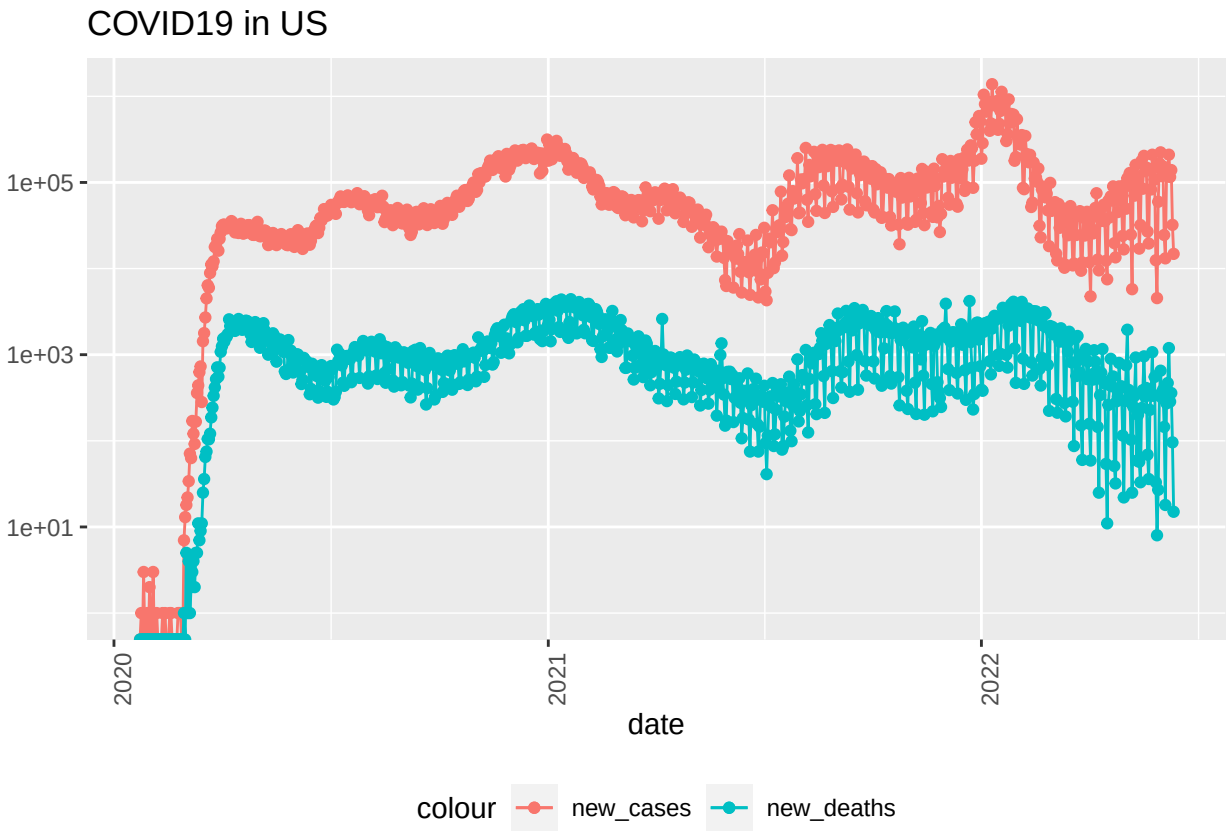
Group the US data by Province_State first and group by Country_Region next, summarize the number of cases, deaths, and population. Then calculate and add the "death_ per_mill" column in the data set.

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

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_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
```



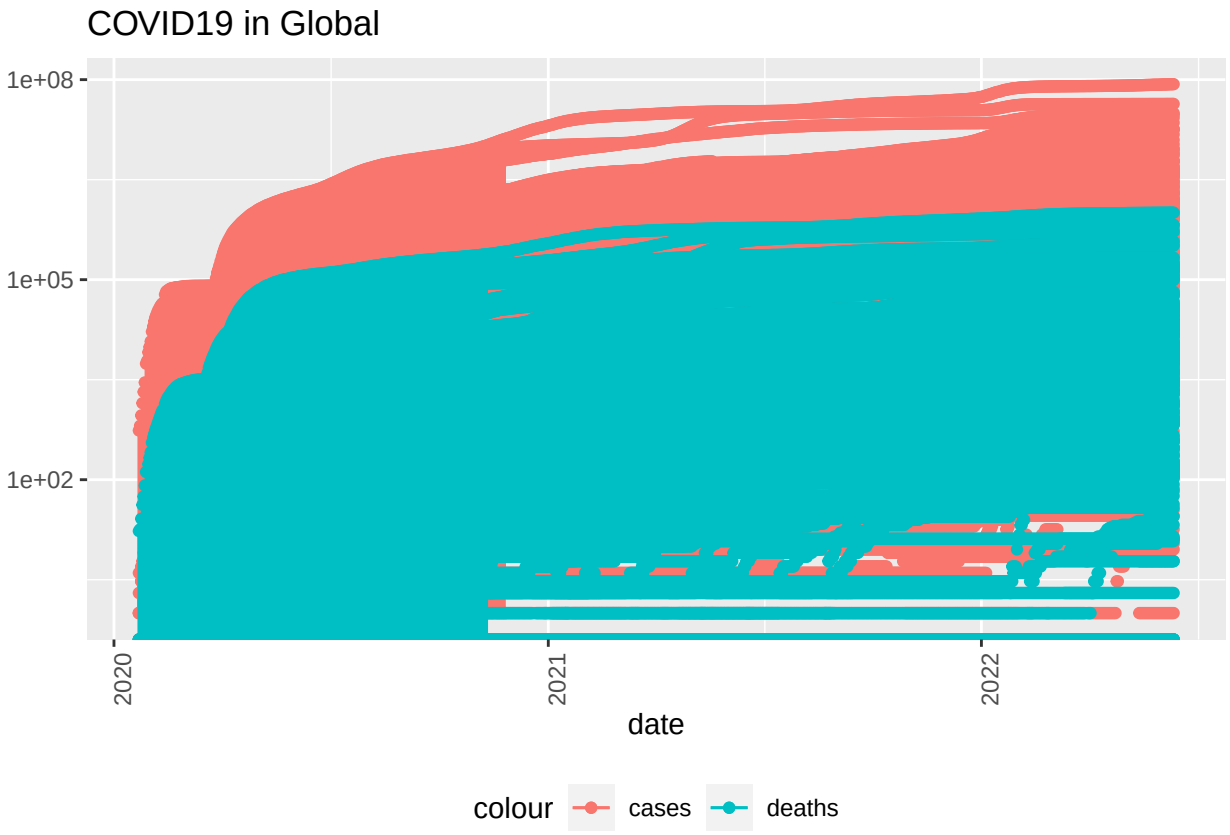
From the graph above, we can see the trend of the cases and deaths of COVID19 in US. Both of them have the almost same trend by time which increase sharply in the first three months of 2020, then increase slowing in the rest of 2020. From 2021 until present, the total cases and total deaths numbers are growing more slow then last year.

```
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_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 = "COVID19 in U
```



The above graph shows the lines of both new cases and new deaths of COVID19 in US. Both the new cases and new deaths are grow very fast until March 2020, then they have a increasing trend until end of 2020. Then, numbers dropping in the first half of 2021, then increasing until reaching their highest in the beginning of 2022.

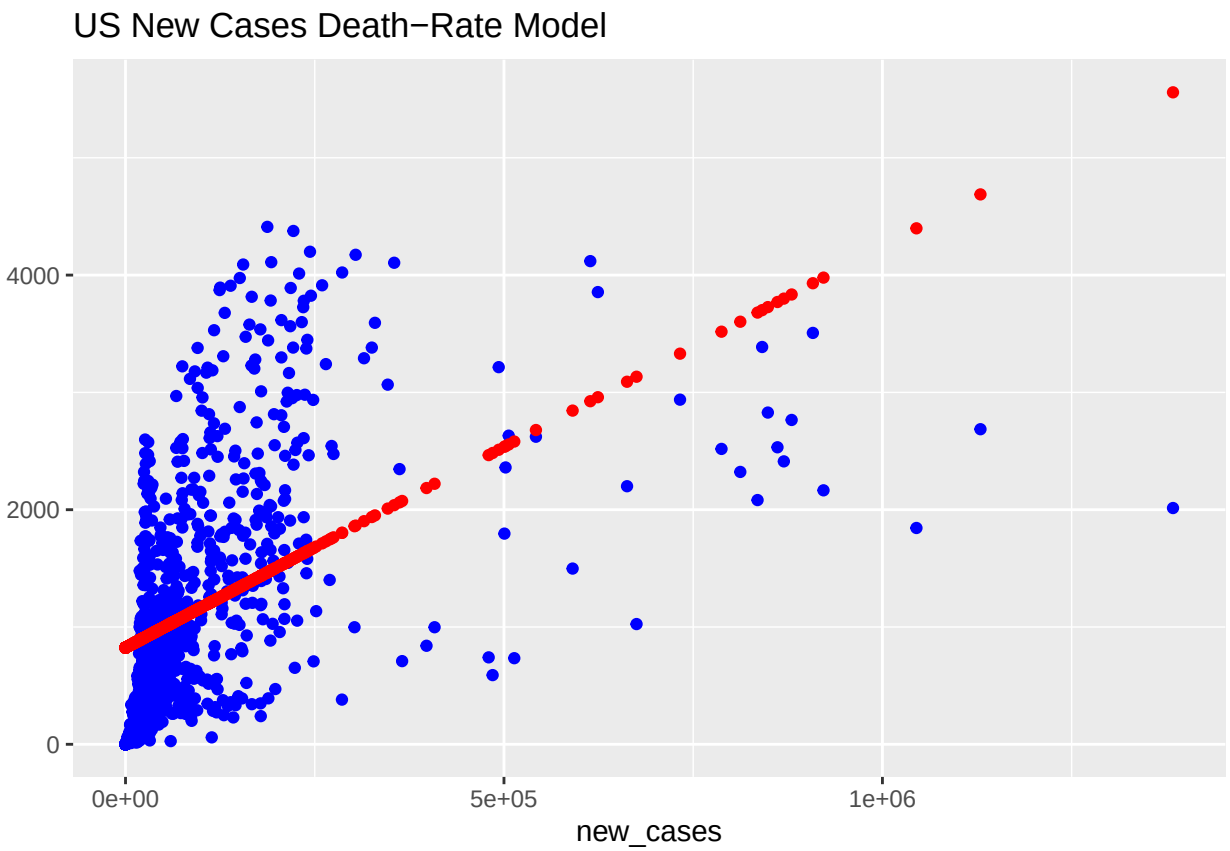
```
global_by_state <- global %>%
group_by(Province_State, Country_Region, date, Combined_Key )%>%
summarize(cases = sum(cases), deaths= sum(deaths), Population = sum(Population  ))%>%
mutate(deaths_per_mill = deaths *1000000 / Population)%>% select(Province_State, Country_Region, date,

global_totals <- global_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()

global_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 Gl
```



The global COVID19 cases seems has the same trend as the US COVID19 cases.

New Cases Death Rate Model

```
mod <- lm(new_deaths ~ new_cases, data = US_totals)
US_totals_new <- US_totals%>% filter(new_cases >= 0)
US_totals_model <-US_totals_new %>% mutate(pred = predict(mod))

US_totals_model %>%
ggplot() + geom_point(aes(x= new_cases, y= new_deaths), color = "blue" ) + geom_point(aes(x= new_cases,
```



Conclusion

From the analysis above, we can see the cases and deaths of COVID19 in US have the same trend as in global. The COVID cases and deaths increased sharply until the end of March 2020, then numbers increase slowing in the rest of 2020. Moreover, Both the new cases and new deaths of COVID19 in US reach their highest in the beginning of 2022.

Bias Identification

When cleaning the data set, both the US and global data has the data with negative number of cases and deaths. Seems the data set has some data leaking in the original data set or in the transforming process. The missing data could make me ignore much details in analysising the trand of numbers, and it could be a potential source.