Report on COVID19 Data

Kodai.F

2023-07-24

Analyzing the COVID-19 trends data

This is the final report for CU boulder's "Data Science as a Field" class, which analyzes COVID-19 infection trends.The data is from https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/.

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
url_in <- "https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_
file_names <- c("time_series_covid19_confirmed_global.csv?raw=true",
                "time_series_covid19_deaths_global.csv?raw=true",
                "time_series_covid19_confirmed_US.csv?raw=true",
                "time_series_covid19_deaths_US.csv?raw=true")
urls <- str_c(url_in, file_names)
```

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

Cleaning the data

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

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

Creating new vectors

```r
Japan_cases <- global_cases %>%
  filter(`Country/Region` == 'Japan')
Japan_cases
```

```
## # A tibble: 1,143 x 4
##    `Province/State` `Country/Region` date       cases
##    <chr>            <chr>            <date>     <dbl>
##  1 <NA>             Japan            2020-01-22     2
##  2 <NA>             Japan            2020-01-23     2
##  3 <NA>             Japan            2020-01-24     2
##  4 <NA>             Japan            2020-01-25     2
##  5 <NA>             Japan            2020-01-26     4
##  6 <NA>             Japan            2020-01-27     4
##  7 <NA>             Japan            2020-01-28     7
##  8 <NA>             Japan            2020-01-29     7
##  9 <NA>             Japan            2020-01-30    11
## 10 <NA>             Japan            2020-01-31    15
## # i 1,133 more rows
```

```r
Japan_deaths <- global_deaths %>%
  filter(`Country/Region` == 'Japan')
Japan_deaths
```

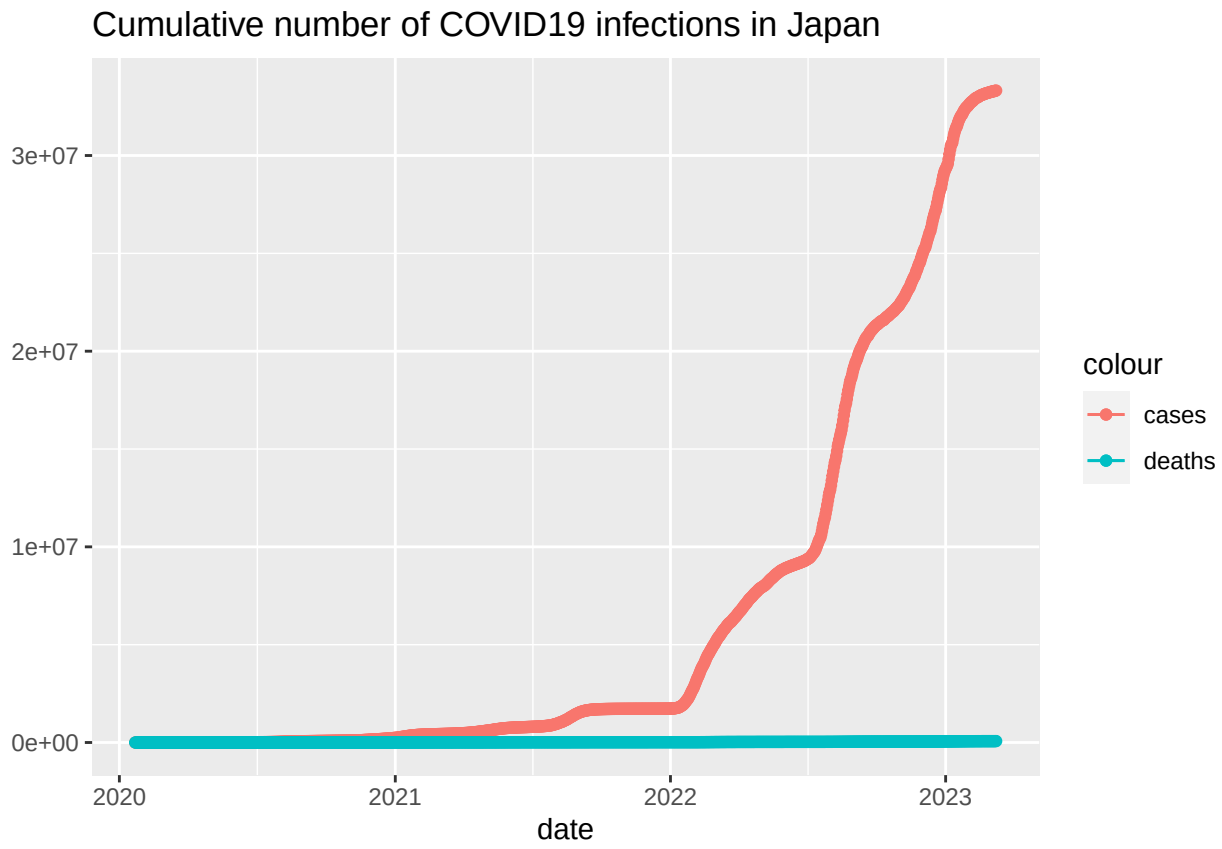
```
## # A tibble: 1,143 x 4
##    `Province/State` `Country/Region` date       deaths
##    <chr>            <chr>            <date>      <dbl>
##  1 <NA>             Japan            2020-01-22      0
##  2 <NA>             Japan            2020-01-23      0
##  3 <NA>             Japan            2020-01-24      0
##  4 <NA>             Japan            2020-01-25      0
##  5 <NA>             Japan            2020-01-26      0
##  6 <NA>             Japan            2020-01-27      0
##  7 <NA>             Japan            2020-01-28      0
##  8 <NA>             Japan            2020-01-29      0
##  9 <NA>             Japan            2020-01-30      0
## 10 <NA>             Japan            2020-01-31      0
## # i 1,133 more rows
```

```r
Japan <- Japan_cases %>%
  full_join(Japan_deaths) %>%
  rename(Country_Region = 'Country/Region',
         Province_State = 'Province/State')
```

```
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
```

Cumulative number of COVID-19 infections in Japan

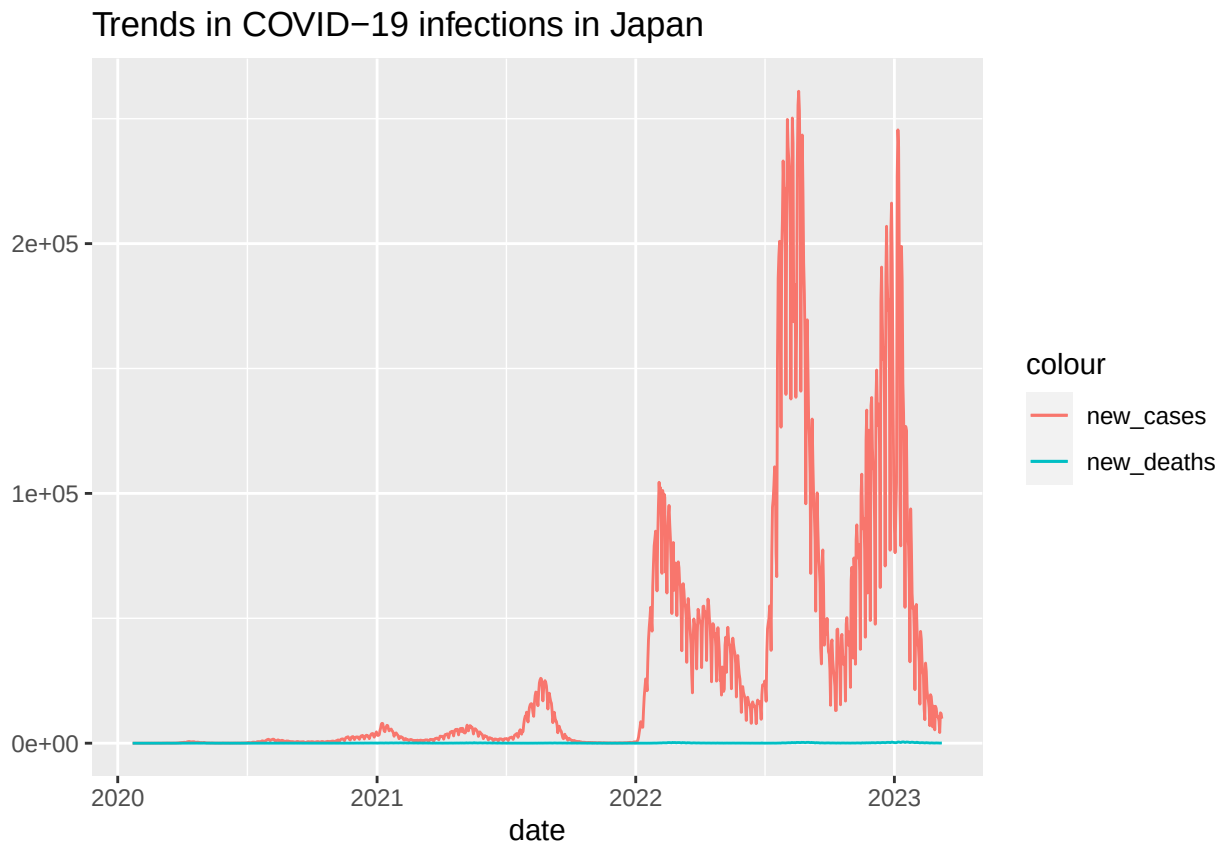
```
Japan %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  labs(title = "Cumulative number of COVID19 infections in Japan", y = NULL)
```



Trends in the number of new cases of COVID-19 infection

```
Japan <- Japan %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths)) %>%
  # Replace NA
  replace_na(list(new_cases = 2, new_deaths = 0))
```

```r
Japan %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  labs(title = "Trends in COVID-19 infections in Japan", y = NULL)
```



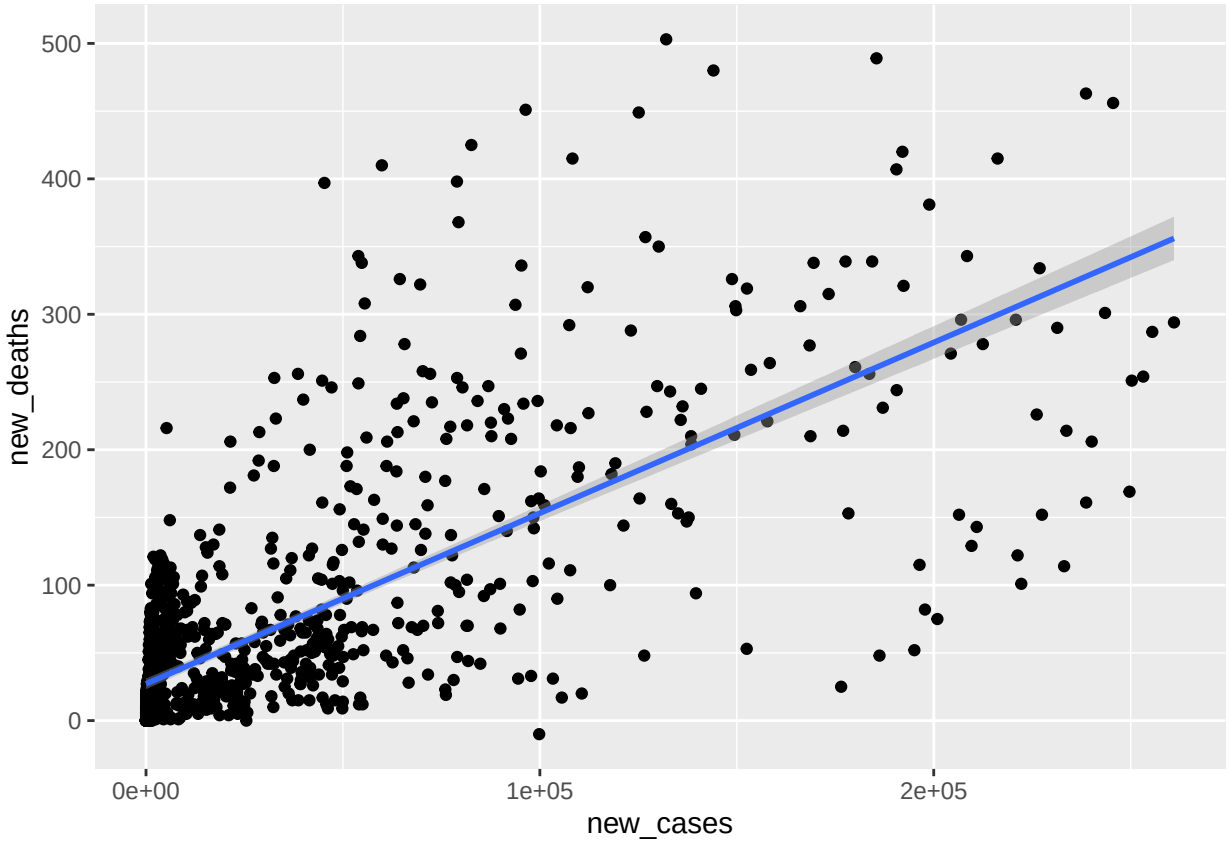
In Japan, the epidemic of COVID-19 infection experienced several waves, with a peak in the number of new daily infections in mid-2022.

Modeling(Correlation between the number of new infections and deaths)

As the number of new infections increases, the number of deaths also increases.

```r
h <- ggplot(data = Japan)+
  geom_point(mapping = aes(x = new_cases, y = new_deaths))+
  geom_smooth(mapping = aes(x = new_cases, y = new_deaths), method = 'lm')
h
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



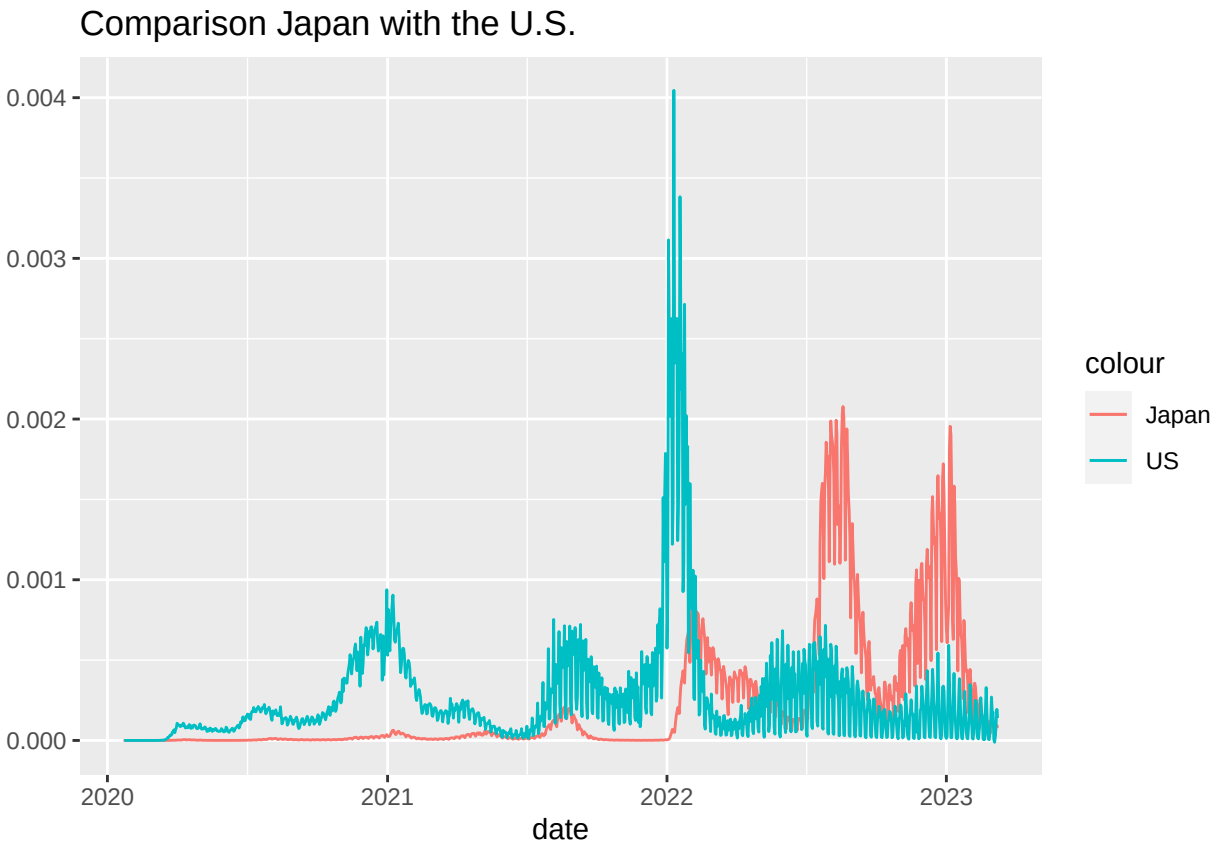
Comparison Japan with USA

To compare data from Japan and the U.S., the number of new infections in Japan and the U.S. was divided by their respective populations. In other words, I calculated the number of infections per capita.

```
Japan <- Japan %>%
  mutate(per_population = new_cases/125600000)
```

```
US <- global_cases %>%
  filter(`Country/Region` == 'US') %>%
  mutate(new_cases = cases - lag(cases),
         per_population = new_cases/334800000)
```

```
g <- ggplot(mapping = aes(x = date, y = per_population)) +
  geom_line(data = Japan, mapping = aes(x = date, y = per_population, color = "Japan")) +
  geom_line(data = US, mapping = aes(x = date, y = per_population, color = "US")) +
  labs(title = "Comparison Japan with the U.S.", y = NULL)
g
```



Conclusion

According to the COVID-19 data from Japan and the U.S., it is clear that the pandemic has had several waves of spread. Although the daily number of new cases suggests that the pandemic has passed its peak, it is possible that another wave of spread will occur in the future. There is a certain correlation between the number of new infections and the number of deaths, but it is not a strong correlation.