W5_NYPD analysis

Naomichi Kadota

2023/11/18

0) load the library

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- 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
```

```
library(ggplot2)
library(lubridate)
library(patchwork)
```

1) Read the CSV file

(COVID-19, by JHU_CSSE)

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- 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_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])
```

2) Data manipulation

Let's follow the manipulation demonstrated in the video of week-3, and then extract and merge the data of cases and deaths in Japan (out of my interest).

```
## simplify global data
global_cases <- global_cases %>%
  pivot_longer(cols = -c('Province.State',
                         'Country.Region',Lat,Long),
               names_to = "date",
               values_to = "case") %>%
  select(-c(Province.State, Lat, Long))
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c('Province.State',
                         'Country.Region',Lat,Long),
               names_to = "date",
               values_to = "death") %>%
  select(-c(Province.State, Lat, Long))

## extract JP data; 1143obs(20/1/12~23/3/09)
JP_cases<- global_cases %>%
  filter(Country.Region=="Japan") %>%
  select(-c(Country.Region))
JP_deaths<- global_deaths %>%
  filter(Country.Region=="Japan") %>%
  select(-c(Country.Region))
## merge
JP_data <- left_join(JP_cases, JP_deaths, by = "date")
JP_data <- JP_data %>%
  mutate(date = sub("X", "", date),  # Remove the 'X' prefix
         date = mdy(date))
```

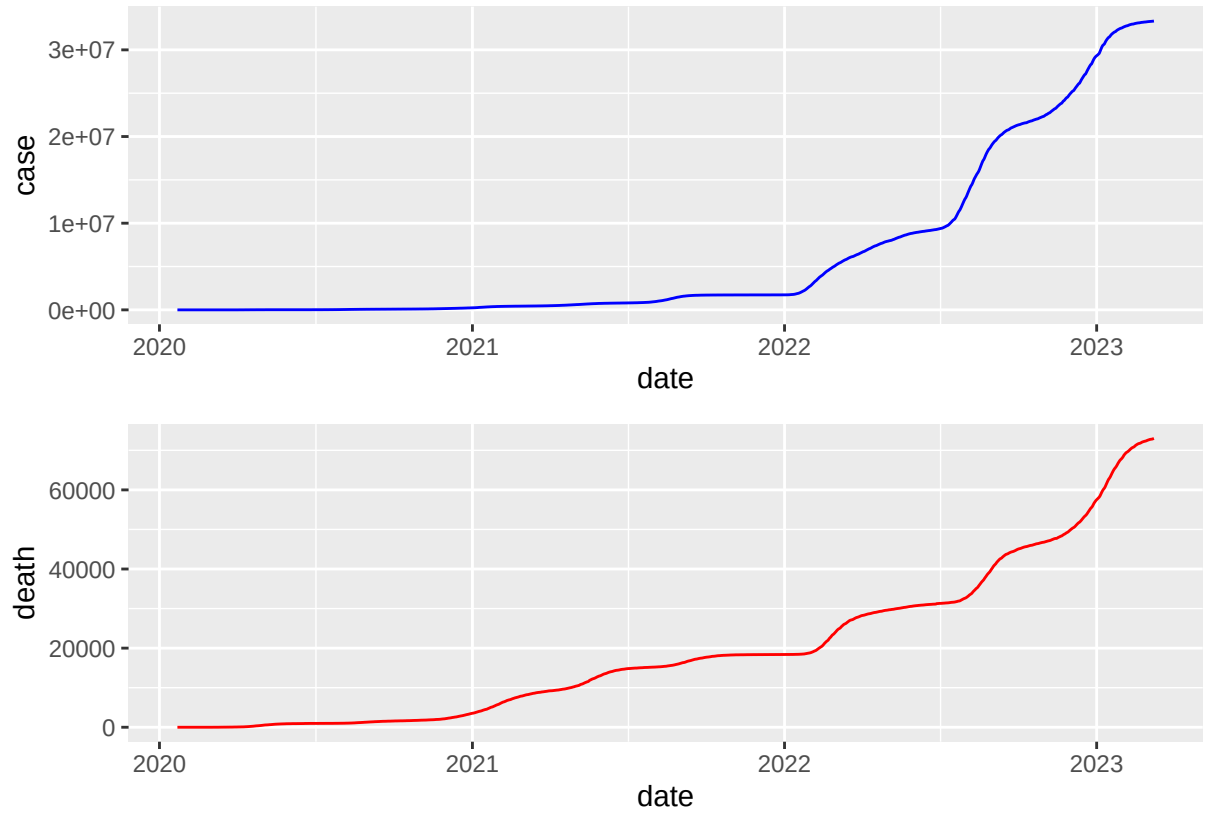
3) Overview

```
summary(JP_data)
```

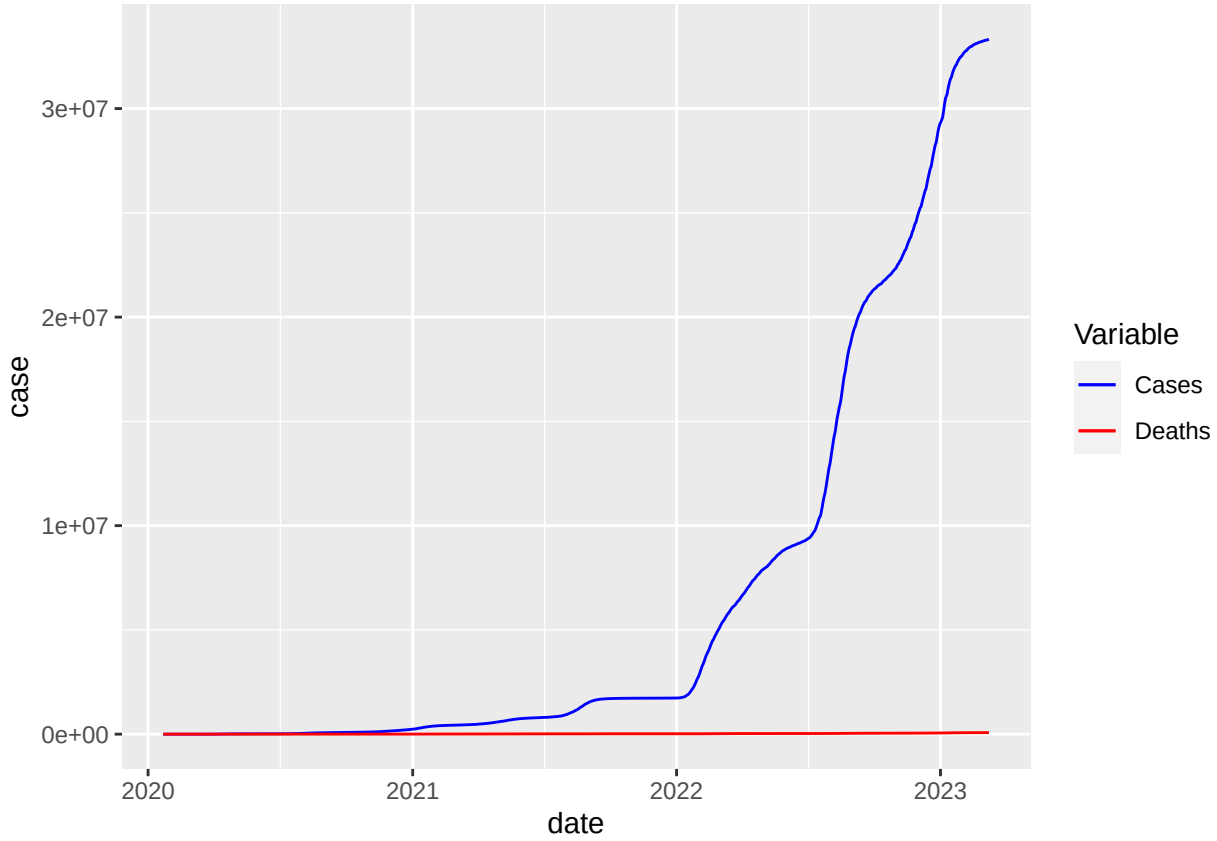
```
##       date                 case              death
##  Min.   :2020-01-22   Min.   :       2   Min.   :    0
##  1st Qu.:2020-11-02   1st Qu.:  102856   1st Qu.: 1792
##  Median :2021-08-15   Median : 1149874   Median :15412
##  Mean   :2021-08-15   Mean   : 6420133   Mean   :19573
##  3rd Qu.:2022-05-27   3rd Qu.: 8785695   3rd Qu.:30523
##  Max.   :2023-03-09   Max.   :33320438   Max.   :72997
```

```
P1.1 <- ggplot(JP_data, aes(x = date, y = case))+
  geom_line(colour="blue")
P1.2 <- ggplot(JP_data, aes(x = date, y = death))+
  geom_line(colour="red")
P1.3 <- ggplot(JP_data, aes(x = date))+
  geom_line(aes(y = case, color = "Cases")) +
```

```
  geom_line(aes(y = death, color = "Deaths")) +
  scale_color_manual(values = c("Cases" = "blue", "Deaths" = "red")) +
  labs(color = "Variable")
P1.1+P1.2 + plot_layout(ncol = 1)
```




```
P1.3
```



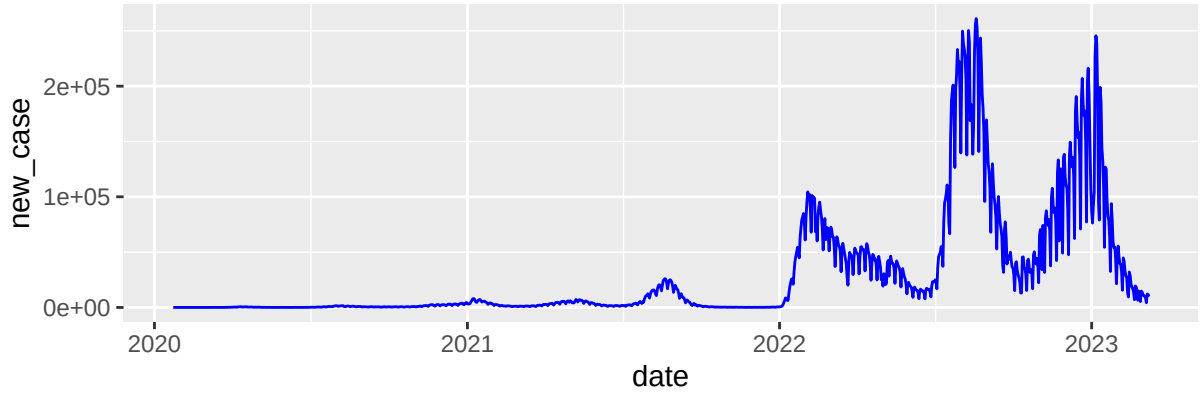
- The plots display the total number of COVID-19 cases and deaths reported. - It's noteworthy that the number of deaths is significantly lower than the total number of cases. - I hadn't realized the scale difference was so substantial; at first, I even suspected a coding error on my part. However, the plots have provided me with a clearer insight.

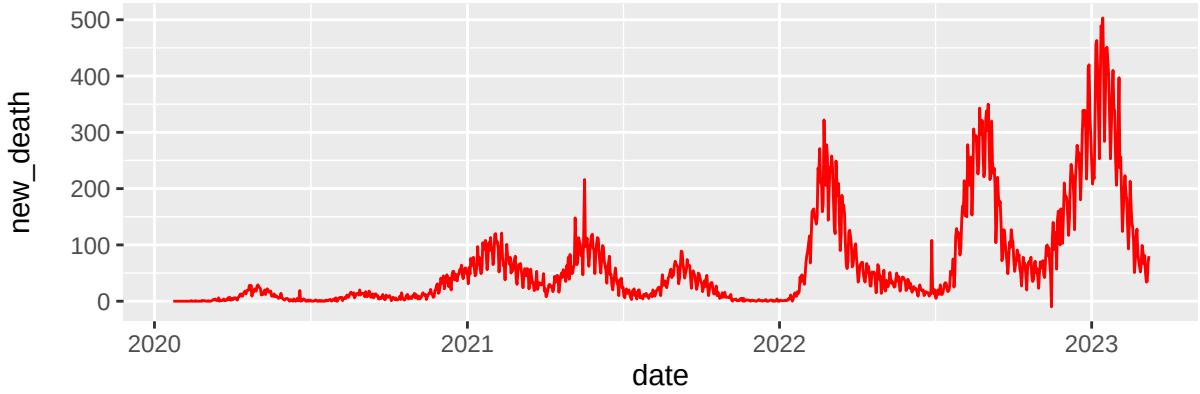
4) Analyze time series data (taking lag)

```
JP_data <- JP_data %>%
  mutate(new_case = case - lag(case),
         new_death = death - lag(death))

P2.1 <- ggplot(JP_data, aes(x = date, y = new_case))+
  geom_line(colour="blue")
P2.2 <- ggplot(JP_data, aes(x = date, y = new_death))+
  geom_line(colour="red")
P2.1+P2.2 + plot_layout(ncol = 1)
```

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





- The patterns of reported cases and deaths are roughly the same. - It might be noteworthy that the proportion of deaths prior to 2022 appears relatively high, especially when considering the fewer reported cases compared to the numbers after 2022.

5) Modeling

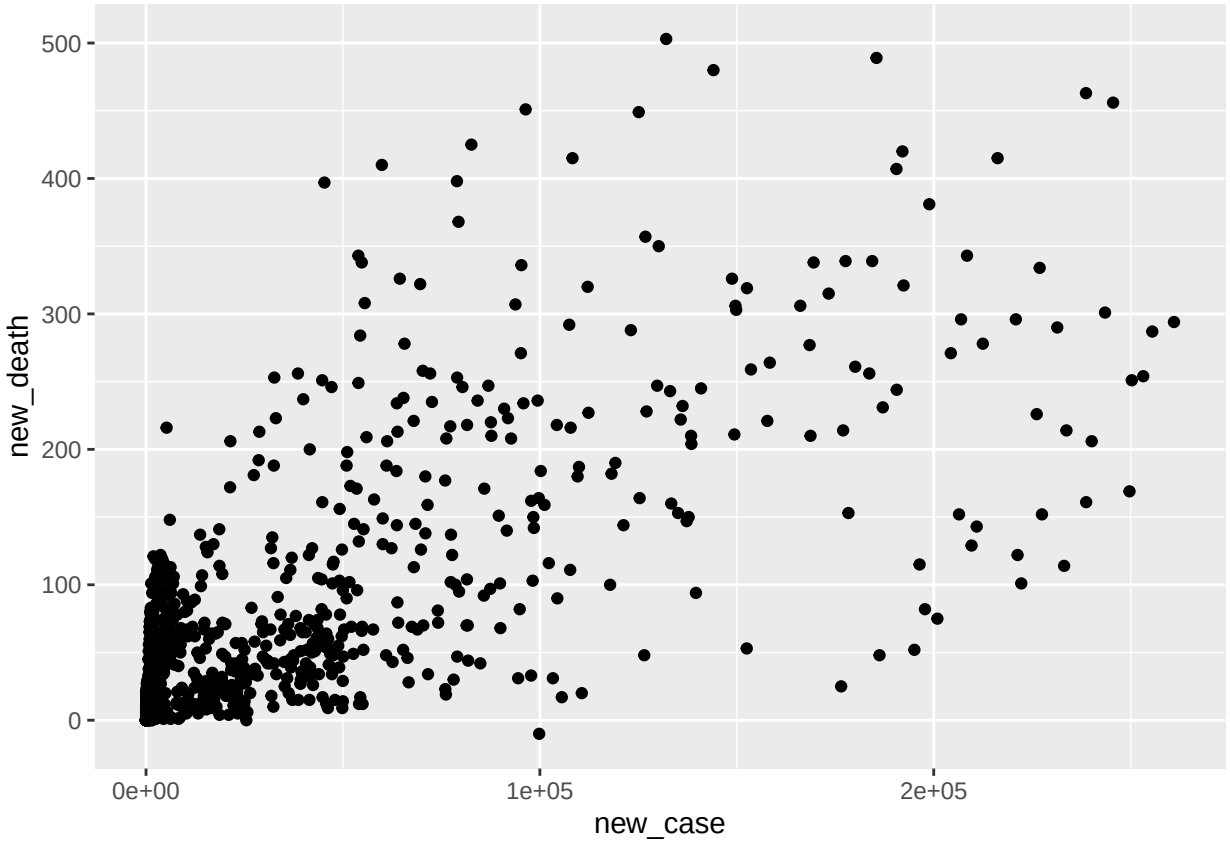
Can we predict the number of new_death from the number of new_case?

```
JP_data2 <- na.omit(JP_data)
lmod <- lm(new_death~new_case, JP_data2)
summary(lmod)
```

```
##
## Call:
## lm(formula = new_death ~ new_case, data = JP_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -224.52  -26.40  -16.26   18.20  312.77
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.716e+01  2.024e+00   13.42   <2e-16 ***
## new_case    1.260e-03  3.446e-05   36.56   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.37 on 1140 degrees of freedom
## Multiple R-squared:  0.5397, Adjusted R-squared:  0.5393
## F-statistic:  1337 on 1 and 1140 DF,  p-value: < 2.2e-16
```

```
ggplot(JP_data2, aes(x=new_case, y=new_death))+
  geom_point()
```



According to the regression analysis, the coefficient appears to have a significant impact. However, given the unevenness in the data distribution, the soundness of this interpretation may be questionable.

6) Discussion of Potential Bias

- This analysis focuses on cases in Japan, so the findings may not be generalizable to other countries.
- As a principle, the case count may include multiple reports for the same individuals, whereas the number of deaths is counted only once per individual. This difference could cause some distortion when interpreting trends.