# Data analysis of CSSE COVID-19 Time Series

Kexin Jiao

2023-11-21

```
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(wesanderson)
```

```
## Warning: package 'wesanderson' was built under R version 4.3.2
```

This is the global COVID-19 data collected and reported by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) ended on March 10,2023. Data page

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- c("time_series_covid19_confirmed_US.csv","time_series_covid19_confirmed_global.csv","time_
urls <- str_c(url_in,file_names)
confirmed_US <- read.csv(urls[1])
confirmed_global <- read.csv(urls[2])
deaths_US <- read.csv(urls[3])
deaths_global <- read.csv(urls[4])
```

## 1. Data rangling

### 1) Reshape the date frame to create a new column of cases per date.

In the original data frames, each date is a individual column. We will pivot the data frames to create a combined column of date and the column of cases per date.

```
#Drop the columns not related to further analysis
confirmed_US_long <- confirmed_US %>%
  pivot_longer(!colnames(confirmed_US)[1:11],
               names_to = "Date",
               values_to = "Cases") %>%
  select(!c(colnames(confirmed_US[1:6]),"Lat","Long_"))

confirmed_global_long <- confirmed_global %>%
  pivot_longer(!colnames(confirmed_global)[1:4],
               names_to = "Date",
```

```
              values_to = "Cases") %>%
  select(!c("Lat","Long"))

deaths_US_long <- deaths_US %>%
  pivot_longer(!colnames(deaths_US)[1:12],
               names_to = "Date",
               values_to = "Death") %>%
  select(!c(colnames(confirmed_US[1:6]),"Lat","Long_"))

deaths_global_long <- deaths_global %>%
  pivot_longer(!colnames(deaths_global)[1:4],
               names_to = "Date",
               values_to = "Death") %>%
  select(!c("Lat","Long"))
```

**2) Join the related data frames, reformat the columns, and group the columns for further analysis**

```
#Combine related data frames using join. Reformat the date columns.
global <- confirmed_global_long %>%
  full_join(deaths_global_long) %>%
  rename(Province_State = "Province.State",Country_Region = "Country.Region") %>%
  mutate(Date=substring(Date,2)) %>%
  mutate(Date = mdy(Date))
```

```
## Joining with `by = join_by(Province.State, Country.Region, Date)`
```

```
confirmed_US_long$Death = deaths_US_long$Death
confirmed_US_long$Population = deaths_US_long$Population
US <- confirmed_US_long %>% select(everything()) %>%
  mutate(Date=substring(Date,2)) %>%
  mutate(Date = mdy(Date))

#Group columns for further analysis
global_by_date<-global %>%
  group_by(Country_Region,Date) %>%
  summarize(Total_cases=sum(Cases),Total_death=sum(Death)) %>%
  select(everything()) %>%
  ungroup()
```

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

```
US_by_date<-US %>%
  group_by(Province_State,Date) %>%
  summarize(Total_cases=sum(Cases),Total_death=sum(Death),Total_population=max(Population)) %>%
  select(everything()) %>%
  ungroup()
```

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

## 2. Data visualization

### 1) Plotting of number of death in every thousand of people in some developed countries

```
country<-c("Australia","Canada","France","Germany","Japan","US","United Kingdom")
population<-c(25978935, 38929902,67935660,84079811,125124989,333287557,66971411)

# Generate data frames for each country
index=1
for (i in country){
  df <- paste0(i)
  global_by_date<- global_by_date %>%
    mutate(death_per_thousand=1000*Total_death/population[index])
  index=index+1
  assign(df,subset(global_by_date,Country_Region == i))
  print(head(df))
}
```

```
## [1] "Australia"
## [1] "Canada"
## [1] "France"
## [1] "Germany"
## [1] "Japan"
## [1] "US"
## [1] "United Kingdom"
```

```
#Plotting
color_A <- c("Canada"="purple4","Germany"="goldenrod2","Japan"="steelblue","US"="maroon","France"="spri
ggplot()+
  geom_point(data=Canada,aes(x=Date,y=death_per_thousand,color="Canada"),size=1)+
  geom_point(data=Germany,aes(x=Date,y=death_per_thousand,color="Germany"),size=1)+
  geom_point(data=Japan,aes(x=Date,y=death_per_thousand,color="Japan"),size=1)+
  geom_point(data=US,aes(x=Date,y=death_per_thousand,color="US"),size=1)+
  geom_point(data=France,aes(x=Date,y=death_per_thousand,color="France"),size=1)+
  geom_point(data=`United Kingdom`,aes(x=Date,y=death_per_thousand,color="United Kingdom"),size=1)+
  labs(title = "Death per thousand by date in developed countries",x="Date",y="Death per thousand",colo
  scale_color_manual(name="Coutries",values=color_A)
```

A

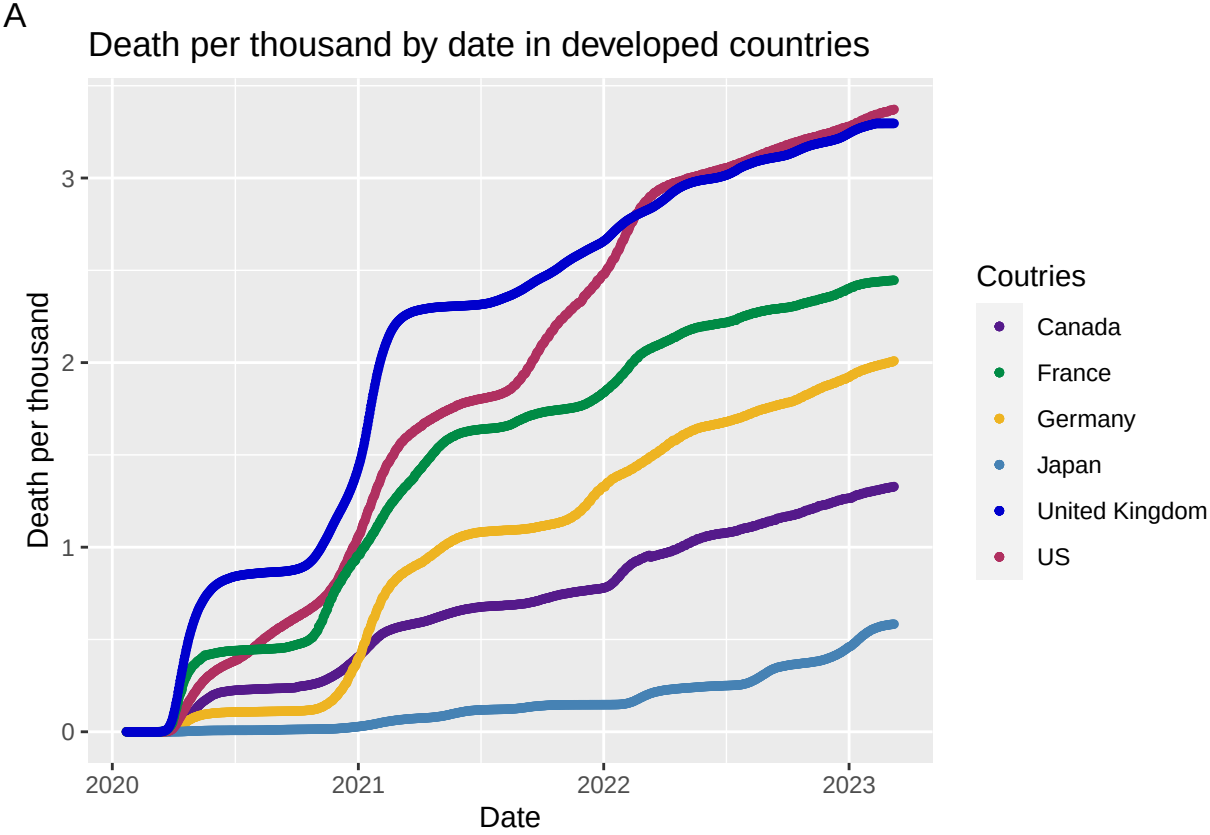



Figure A reveals that at the early stage of COVID-19 panic, the death rate rapidly increased in the United Kingdom. Both the death rates in the United Kingdom and the US were about 5-6 times higher than that for Japan by the end of the date in the data. The death rate in the UK increased rapidly at the beginning of COVID-19 while the death rate in the US became high from 2021 to 2022. The question could be asked is: why did the death rate grow rapidly in the UK and the US compare to Japan? It may be valid to investigate what happened in the UK before 2021 and what happened in the US between 2021 and 2022. We should also ask why the death rate is so high in the UK and the US but relatively low in Japan.

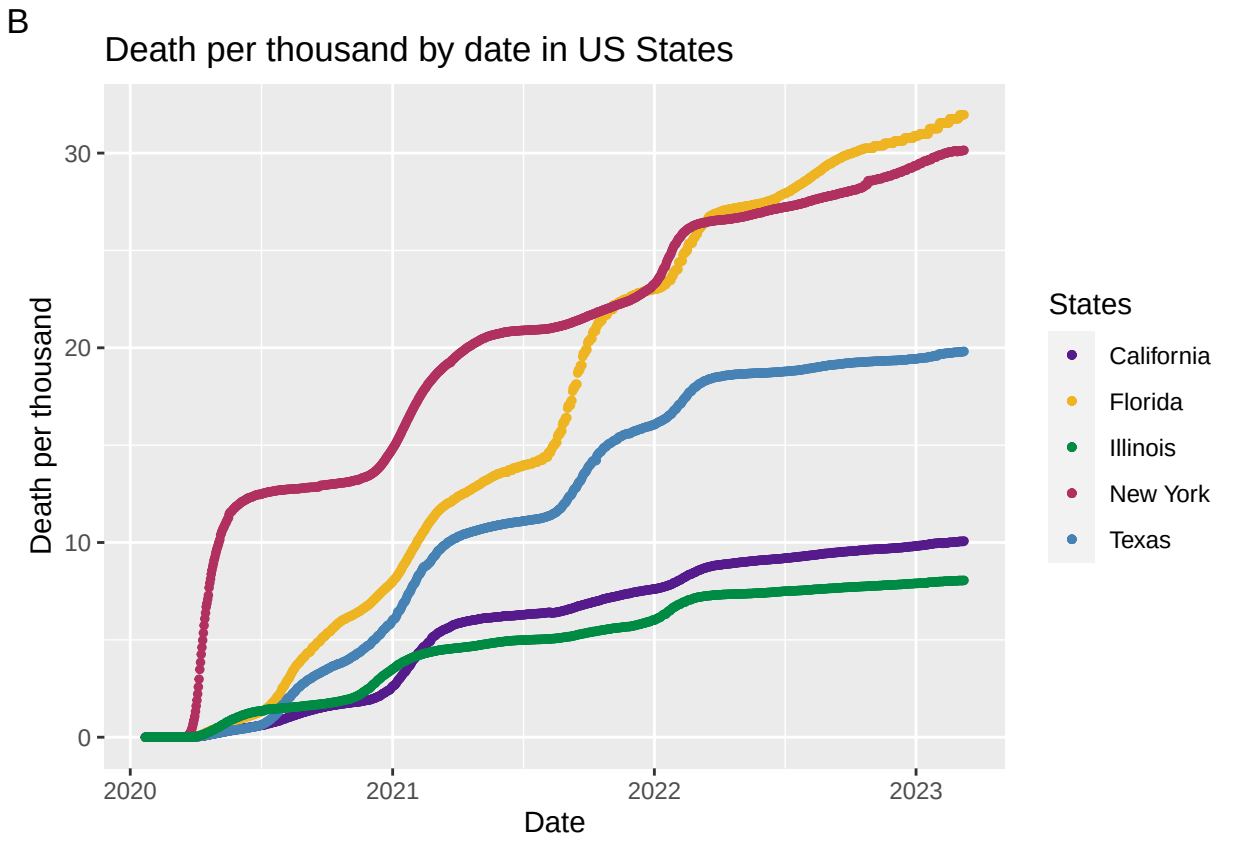
**2) Plotting of number of death in every thousand of people in some states in US**

```
state<-c("California","Florida","Texas","New York","Illinois")
# Generate data frames for each state
for (i in state){
  df <- paste0(i)
  US_by_date<-US_by_date %>%
    mutate(death_per_thousand=1000*Total_death/Total_population)
  assign(df,subset(US_by_date,Province_State == i))
  print(head(df))
}
```

```
## [1] "California"
## [1] "Florida"
## [1] "Texas"
## [1] "New York"
## [1] "Illinois"
```

```r
#plotting
color_B<-c("California"="purple4","Florida"="goldenrod2","Texas"="steelblue","New York"="maroon","Illino
ggplot()+
  geom_point(data=California,aes(x=Date,y=death_per_thousand,color="California"),size=1)+
  geom_point(data=Florida,aes(x=Date,y=death_per_thousand,color="Florida"),size=1)+
  geom_point(data=Texas,aes(x=Date,y=death_per_thousand,color="Texas"),size=1)+
  geom_point(data=`New York`,aes(x=Date,y=death_per_thousand,color="New York"),size=1)+
  geom_point(data=Illinois,aes(x=Date,y=death_per_thousand,color="Illinois"),size=1)+
  labs(title="Death per thousand by date in US States",x="Date",y="Death per thousand",color="Legend",ta
  scale_color_manual(name="States",values=color_B)
```

B




```r
selected_state_population <- US_by_date %>%
  group_by(Province_State) %>%
  subset(.,Province_State %in% c("California","Florida","Illinois","New York","Texas")) %>%
  mutate(population_for_state=max(Total_population)) %>% select(c(Province_State,population_for_state))
selected_state_population[match(unique(US_by_date$Province_State),US_by_date$Province_State),]
```

```
## # A tibble: 58 x 2
## # Groups:   Province_State [6]
##    Province_State population_for_state
##    <chr>                         <int>
##  1 California                10039107
##  2 Florida                    2716940
##  3 Illinois                   5150233
##  4 New York                   2559903
##  5 Texas                      4713325
```

```
##  6 <NA>                              NA
##  7 <NA>                              NA
##  8 <NA>                              NA
##  9 <NA>                              NA
## 10 <NA>                              NA
## # i 48 more rows
```

Figure B shows that the death rate was distinctly high in New York and was finally surpassed by that of Florida. It would be valuable to investigate the models of death rates for New York and Florida. Considering the population of each state, California has almost four times higher population than New York and Florida but a much lower death rate. The question is: why did the death rates were much higher in Florida and New York than that of California although their polulations are only about 1/4 of California? The reasons behind this need to be investigated along with more data.

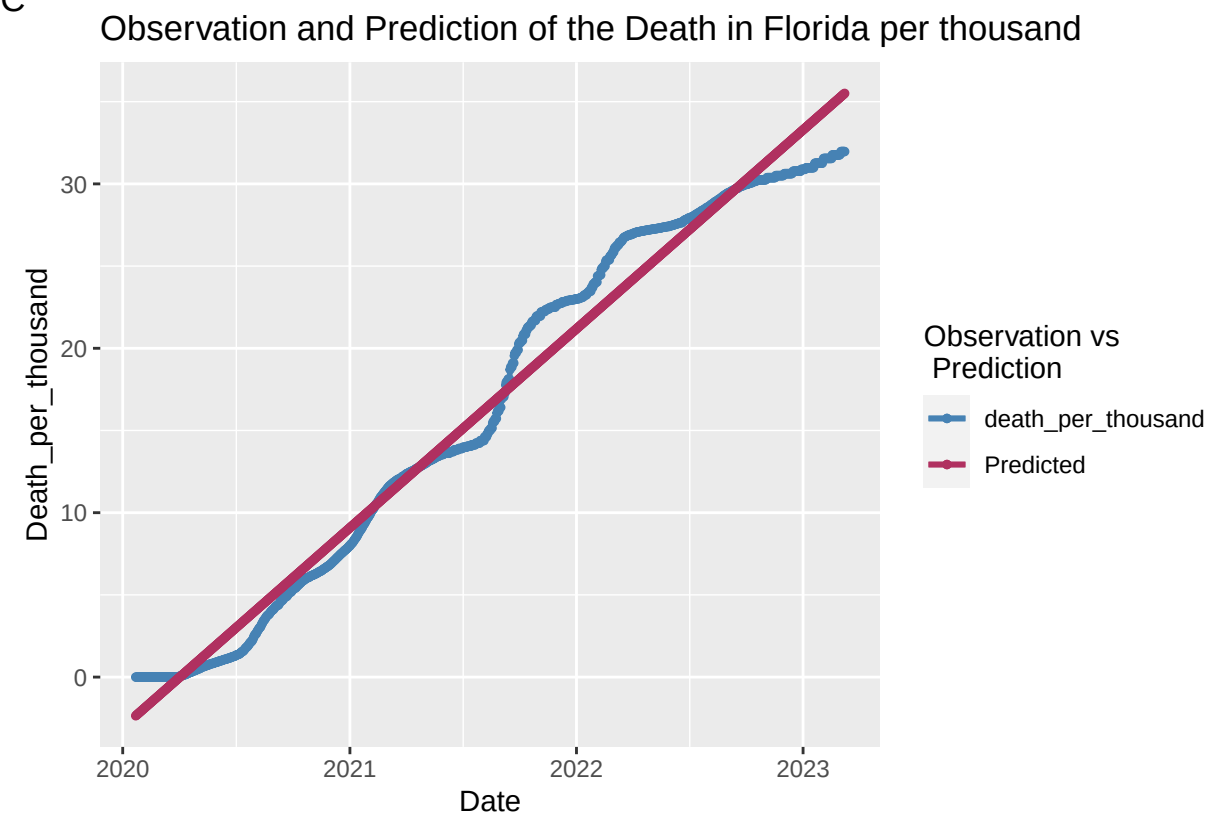
## 3. Modeling

```
mod_Florida<-lm(data=Florida,death_per_thousand~Date)
summary(mod_Florida)
```

```
##
## Call:
## lm(formula = death_per_thousand ~ Date, data = Florida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5389 -1.1079 -0.2184  1.2732  3.0231
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.085e+02  2.680e+00  -227.1   <2e-16 ***
## Date         3.315e-02  1.421e-04   233.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.585 on 1141 degrees of freedom
## Multiple R-squared:  0.9795, Adjusted R-squared:  0.9794
## F-statistic: 5.443e+04 on 1 and 1141 DF,  p-value: < 2.2e-16
```

```
Florida<-Florida %>% mutate(Predicted=predict(mod_Florida))
color_C<-c("death_per_thousand"="steelblue","Predicted"="maroon")
ggplot()+
  geom_point(data=Florida,aes(x=Date,y=death_per_thousand,color="death_per_thousand"),size=1)+
  geom_line(data=Florida,aes(x=Date,y=death_per_thousand,color="death_per_thousand"),linewidth=1)+
  geom_point(data=Florida,aes(x=Date,y=Predicted,color="Predicted"),size=1)+
  geom_line(data=Florida,aes(x=Date,y=Predicted,color="Predicted"),linewidth=1)+
  labs(title="Observation and Prediction of the Death in Florida per thousand",x="Date",y="Death_per_th
  scale_color_manual(name="Observation vs \n Prediction",values=color_C)
```

C




Here we studied how the death rate changed in Florida. Figure C shows a linear fitting of the death rate in Florida as a function of date. The slope of the fitted curve is 3.315e-02. The good fitting result indicated a steadily increased death rate in Florida. The question is: what is the mechanism/theory this fitting tells about the COVID-19 infection in Florida? One can compare this value to other states or the theoretical value predicted by epidemiology models to better understand the situation in Florida during the COVID-19 panic.

## 4. Discussion of possible bias in the data and analysis

1. The COVID-19 data is based on the reports from individual states/provinces and countries/regions. The efficiency and accuracy of data collection and statistics highly depend on the standard, statistical method, and the data collection power of health/disease control organizations in individual states/countries.
2. The mobility of population between countries and states may affect the local number of cases and deaths. For example, a region that has more economic activity, better medical institutions, or a developed tourism industry may become the preference for those people who are able to choose where to pull through during the panic. The increased population may cause a higher number of infections.