

Covid_19_deaths_JH_data

JG

2023-08-22

Setup

```
### install.packages("tidyverse"), install.packages("readr"), install.packages("knitr"), install.packages("rmarkdown")

library(tidyverse)
```

```
## -- 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    3.4.3      v tibble     3.2.1
## v lubridate  1.9.2      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(lubridate)
library(stringr)
library(readr)
library(selectr)
```

Import Data

#note:website error prohibited importing directly from site, save to local directory

```
global_cases_url<-"C:\\Temp\\time_series_covid19_confirmed_global.csv"
global_deaths_url<-"C:\\Temp\\time_series_covid19_deaths_global.csv"
US_cases_url<-"C:\\Temp\\time_series_covid19_confirmed_US.csv"
US_deaths_url<-"C:\\Temp\\time_series_covid19_deaths_US.csv"
global_cases<-read_csv(global_cases_url)
```

```
## Rows: 289 Columns: 1147
## -- Column specification -----
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (1145): 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.
```

```
global_deaths<-read_csv(global_deaths_url)
```

```
## Rows: 289 Columns: 1147
## -- Column specification -----
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): 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(US_cases_url)
```

```
## Rows: 3342 Columns: 1154
## -- Column specification -----
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): 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(US_deaths_url)
```

```
## Rows: 3342 Columns: 1155
## -- Column specification -----
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): 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.
```

Tidy data

```
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))

```

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

```

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)

```

Joining with 'by = join_by(Admin2, Province_State, Country_Region,
Combined_Key, date)'

```

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

```

Import population data

```

#add population data
uid_lookup_url<-"C:\\Temp\\UID_ISO_FIPS_LookUp_Table.csv"
uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))

```

```
## Rows: 4321 Columns: 12
## -- Column specification -----
## Delimiter: ","
## chr (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (6): UID, code3, FIPS, 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.
```

```
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)
```

Visualize

```
##Visualize
US_by_state<-US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(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()
```

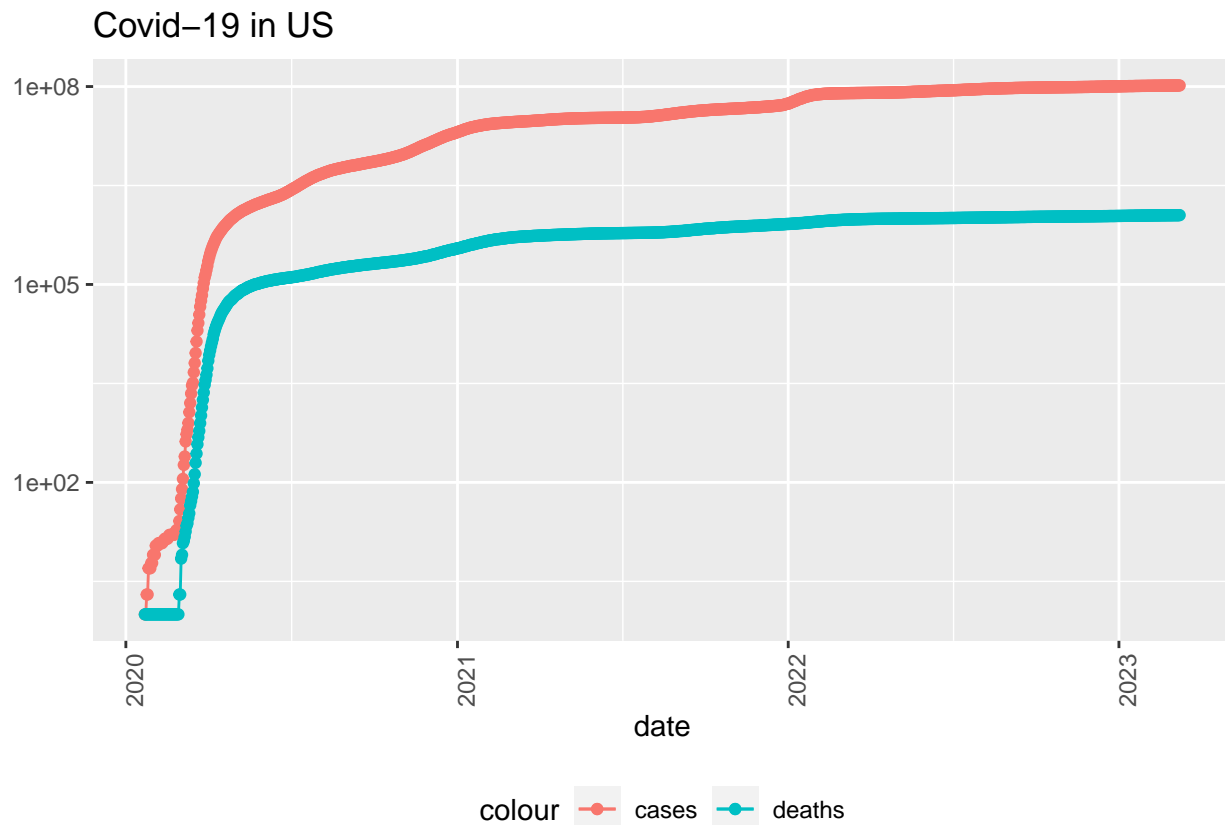
'summarise()' has grouped output by 'Province_State', 'Country_Region'. You can
override using the '.groups' argument.

```
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()
```

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

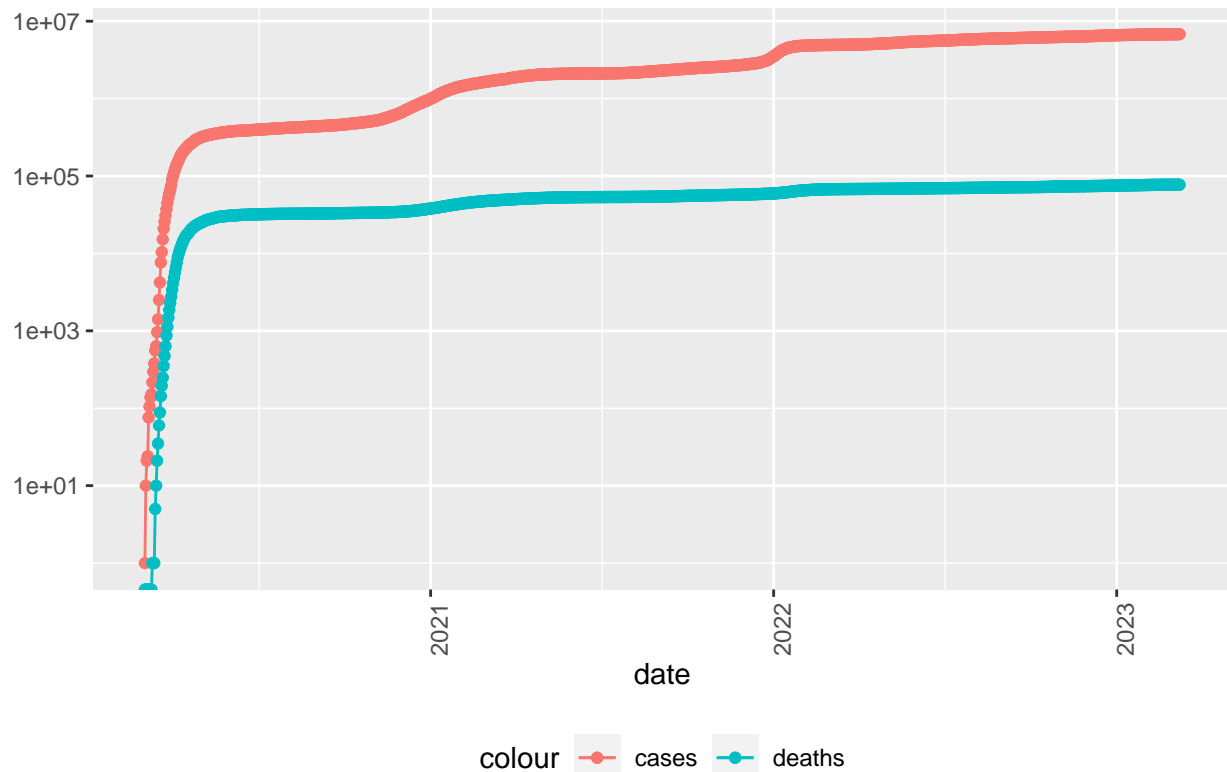
```
US_totals %>%
  filter(cases>0) %>%
  ggplot(aes(x=date,y=cases))+
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y=deaths,color="deaths")) +
```

```
geom_point(aes(y=deaths, color = "deaths")) +
scale_y_log10() +
theme(legend.position="bottom",
      axis.text.x=element_text(angle=90)) +
labs(title="Covid-19 in US", y=NULL)
```



```
state<-"New York"
US_by_state %>%
  filter(Province_State==state) %>%
  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 = str_c("COVID19 in", state), y=NULL)
```

COVID19 in New York

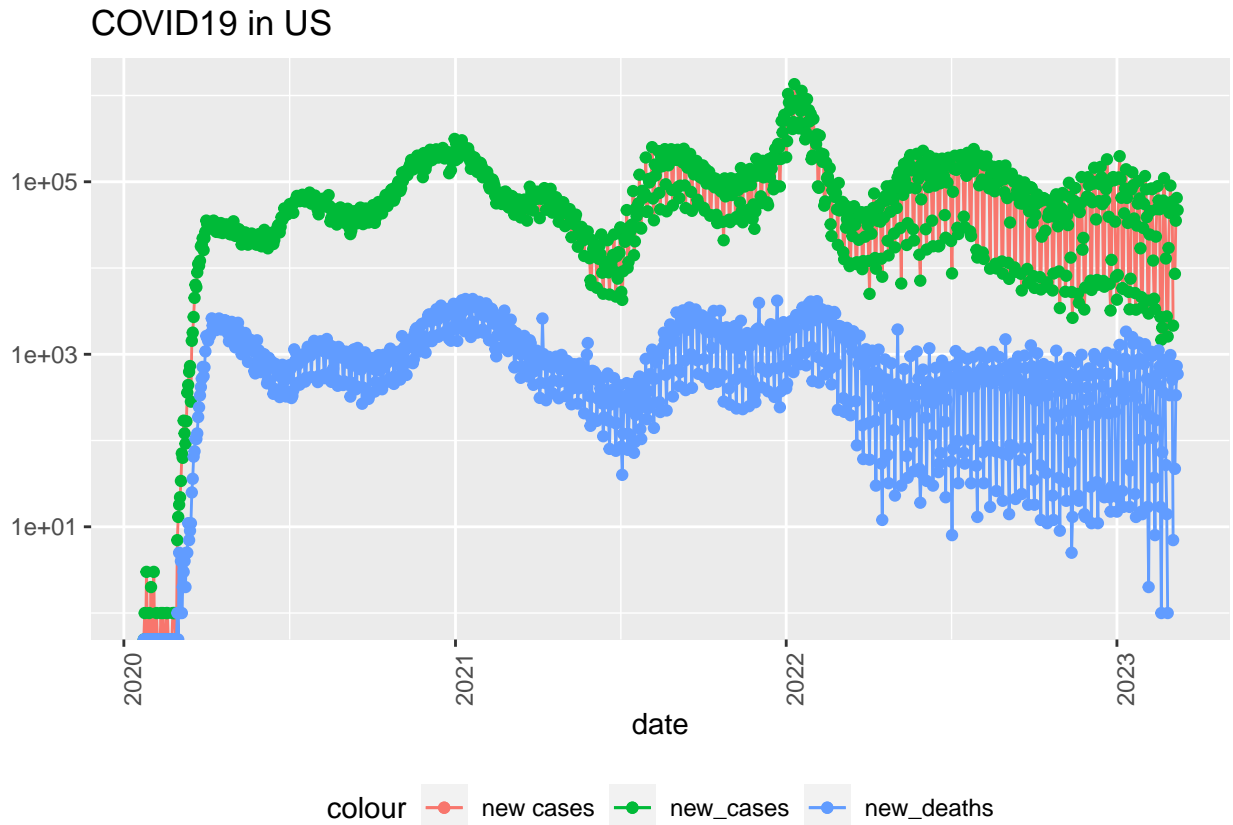


```
US_by_state <-US_by_state %>%
  mutate(new_cases=cases-lag(cases),
         new_deaths=deaths-lag(deaths))
US_totals<-US_totals %>%
  mutate(new_cases =cases-lag(cases),
         new_deaths=deaths - lag (deaths))
tail(US_totals %>%
     select(new_cases, new_deaths,everything()))
```

```
## # A tibble: 6 x 8
##   new_cases new_deaths Country_Region date      cases deaths deaths_per_mill
##   <dbl>      <dbl> <chr>      <date>      <dbl> <dbl>      <dbl>
## 1      2147         7 US        2023-03-04  1.04e8  1.12e6      3371.
## 2     -3862        -38 US        2023-03-05  1.04e8  1.12e6      3371.
## 3      8564         47 US        2023-03-06  1.04e8  1.12e6      3371.
## 4     35371        335 US        2023-03-07  1.04e8  1.12e6      3372.
## 5     64861        730 US        2023-03-08  1.04e8  1.12e6      3374.
## 6     46931        590 US        2023-03-09  1.04e8  1.12e6      3376.
## # i 1 more variable: Population <dbl>
```

```
US_totals %>%
  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 US", y=NULL)
```



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

US_states_totals %>%
  slice_min(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##   deaths_per_thou cases_per_thou Province_State deaths cases population
##   <dbl> <dbl> <chr> <dbl> <dbl> <dbl>
## 1 0.611 150. American Samoa 34 8.32e3 55641
## 2 0.744 248. Northern Mariana Isl~ 41 1.37e4 55144
## 3 1.21 231. Virgin Islands 130 2.48e4 107268
## 4 1.30 269. Hawaii 1841 3.81e5 1415872
```

##	5	1.49	245. Vermont	929	1.53e5	623989
##	6	1.55	293. Puerto Rico	5823	1.10e6	3754939
##	7	1.65	340. Utah	5298	1.09e6	3205958
##	8	2.01	415. Alaska	1486	3.08e5	740995
##	9	2.03	252. District of Columbia	1432	1.78e5	705749
##	10	2.06	253. Washington	15683	1.93e6	7614893

Model

```
mod<-lm(deaths_per_thou~cases_per_thou, data=US_states_totals)
summary(mod)
US_states_totals %>% slice_min(cases_per_thou)
US_states_totals %>% slice_max(cases_per_thou)
x_grid<-seq(25,150)
new_df<-tibble(cases_per_thou =x_grid)
US_tot_w_pred<-US_states_totals %>% mutate(pred=predict(mod))
US_tot_w_pred

US_tot_w_pred %>%
  ggplot()+
  geom_point(aes(x=cases_per_thou,y=deaths_per_thou), color="blue") +
  geom_point(aes(x=cases_per_thou, y=pred), color="red")
```

