Final Project 2: Reproducible Report on COVID-19 Data

Student

2025-09-08

Introduction

This analysis is to answer the question on how COVID-19 cases are trending in the US, and whether the number of deaths from COVID-19 improved.

The sources of data are the COVID-19 time series datasets from John Hopkins (link: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data). This data contains a list of global confirmed cases and deaths at a country level.

Importing Data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- 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(readr)
library(dplyr)
library(ggplot2)
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")
urls <- str_c(url_in, file_names)
global_cases <- read_csv(urls[1])
```

```
## 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.
```

```
death_cases <- read_csv(urls[2])
```

```
## 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.
```

Transforming Data

```
# rename columns
global_cases <- global_cases %>% rename(Country_Region = 'Country/Region',
                                        Province_State = 'Province/State')
death_cases <- death_cases %>% rename(Country_Region = 'Country/Region',
                                      Province_State = 'Province/State')

# focus data on USA only
US_cases <- global_cases %>% filter(Country_Region == 'US')
US_death_cases <- death_cases %>% filter(Country_Region == 'US')

# transpose dates into columns and remove unnecessary Lat, Long columns
# remove State column as it is blank
global_cases <- global_cases %>%
  pivot_longer(cols=-c('Country_Region', 'Province_State', Lat, Long),
               names_to = "dates", values_to = "cases") %>%
  select(-c(Lat, Long, 'Province_State'))

death_cases <- death_cases %>%
  pivot_longer(cols=-c('Country_Region', 'Province_State', Lat, Long),
               names_to = "dates", values_to = "cases") %>%
  select(-c(Lat, Long, 'Province_State'))

US_cases <- US_cases %>%
  pivot_longer(cols=-c('Country_Region', 'Province_State', Lat, Long),
               names_to = "dates", values_to = "cases") %>%
  select(-c(Lat, Long, 'Province_State'))

US_death_cases <- US_death_cases %>%
  pivot_longer(cols=-c('Country_Region', 'Province_State', Lat, Long),
               names_to = "dates", values_to = "death_cases") %>%
  select(-c(Lat, Long, 'Province_State'))

# change factor columns to factor type and dates to date type
global_cases <- global_cases %>%
  mutate(Country_Region = factor(Country_Region),
         dates = mdy(dates))

death_cases <- death_cases %>%
```

```
  mutate(Country_Region = factor(Country_Region),
         dates = mdy(dates))

US_cases <- US_cases %>%
  mutate(Country_Region = factor(Country_Region),
         dates = mdy(dates))

US_death_cases <- US_death_cases %>%
  mutate(Country_Region = factor(Country_Region),
         dates = mdy(dates))

# for global data, get as ratio of death cases to global cases
global_death_cases <- (global_cases %>% group_by(dates) %>%
                         summarize(g_cases=sum(cases))) %>%
  full_join (death_cases %>% group_by(dates) %>%
               summarize(g_d_cases=sum(cases)))
```

```
## Joining with `by = join_by(dates)`
```

```
global_death_cases <- global_death_cases %>% mutate("percent_death_to_total_cases"=g_d_cases / g_cases)

# show summary of transformed data
summary(global_death_cases)
```

```
##      dates                g_cases             g_d_cases
##  Min.   :2020-01-22   Min.   :      557   Min.   :     17
##  1st Qu.:2020-11-02   1st Qu.: 47426060   1st Qu.:1282415
##  Median :2021-08-15   Median :207815449   Median :4388700
##  Mean   :2021-08-15   Mean   :277261852   Mean   :3866857
##  3rd Qu.:2022-05-27   3rd Qu.:528830664   3rd Qu.:6312702
##  Max.   :2023-03-09   Max.   :676570149   Max.   :6881802
##  percent_death_to_total_cases
##  Min.   :0.01009
##  1st Qu.:0.01194
##  Median :0.02101
##  Mean   :0.02398
##  3rd Qu.:0.02488
##  Max.   :0.07730
```

```
summary(US_cases)
```

```
##  Country_Region     dates                cases
##  US:1143        Min.   :2020-01-22   Min.   :        1
##                 1st Qu.:2020-11-02   1st Qu.:  9401879
##                 Median :2021-08-15   Median : 36845900
##                 Mean   :2021-08-15   Mean   : 47080651
##                 3rd Qu.:2022-05-27   3rd Qu.: 84083598
##                 Max.   :2023-03-09   Max.   :103802702
```

```
summary(US_death_cases)
```

```
##  Country_Region     dates             death_cases
##  US:1143        Min.   :2020-01-22   Min.   :      0
##                 1st Qu.:2020-11-02   1st Qu.: 232595
##                 Median :2021-08-15   Median : 618029
##                 Mean   :2021-08-15   Mean   : 624564
##                 3rd Qu.:2022-05-27   3rd Qu.:1006626
##                 Max.   :2023-03-09   Max.   :1123836
```

Visualizing Data

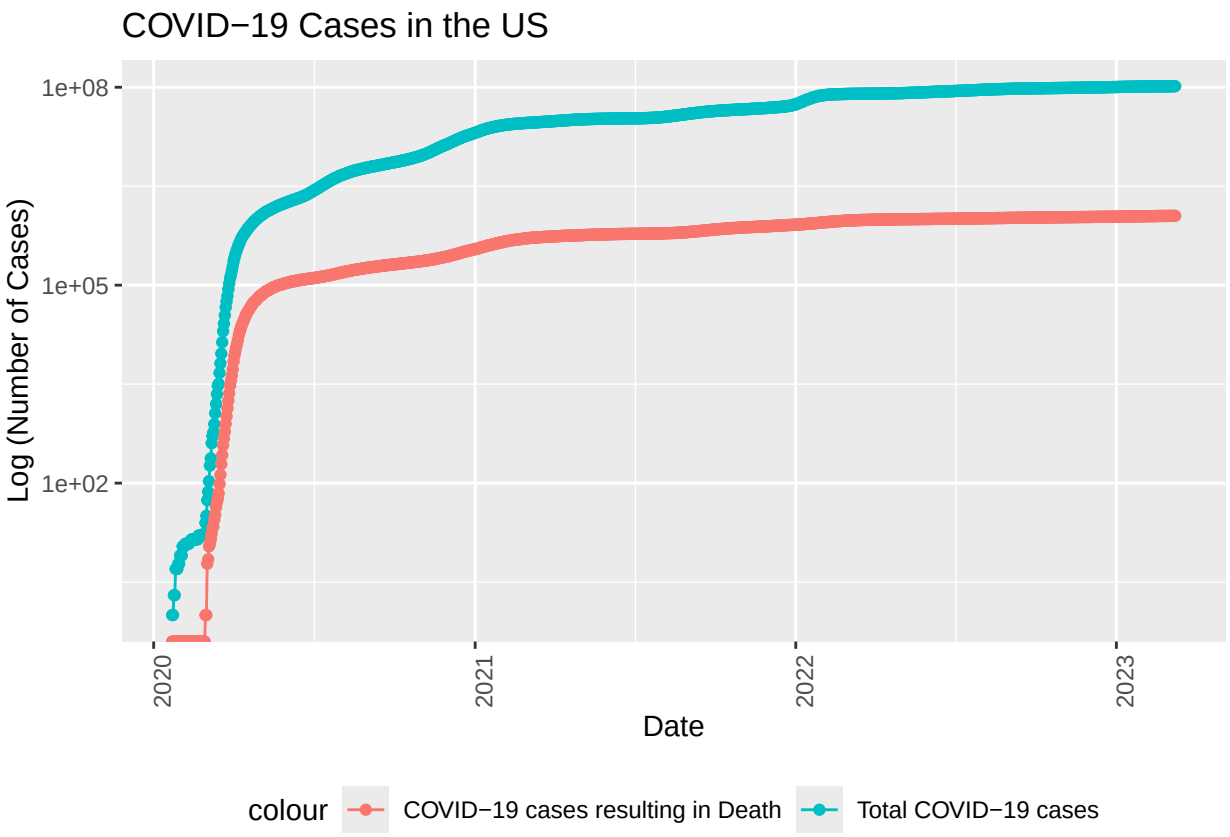
To answer the question on how COVID-19 cases are trending in the US, and whether there are improvements in the deaths from COVID-19, the number of cases and deaths were plotted in a time series.

```
US_cases_death_cases <- US_cases %>% full_join(US_death_cases)
```

```
## Joining with `by = join_by(Country_Region, dates)`
```

```
US_cases_death_cases %>% ggplot(aes(x = dates, y = cases)) +
  geom_line(aes(color="Total COVID-19 cases")) +
  geom_point(aes(color="Total COVID-19 cases")) +
  geom_line(aes(y = death_cases, color="COVID-19 cases resulting in Death")) +
  geom_point(aes(y = death_cases, color="COVID-19 cases resulting in Death")) +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle=90)) +
  scale_y_log10() +
  labs(title = "COVID-19 Cases in the US", y="Log (Number of Cases)", x="Date")
```

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
## log-10 transformation introduced infinite values.
```



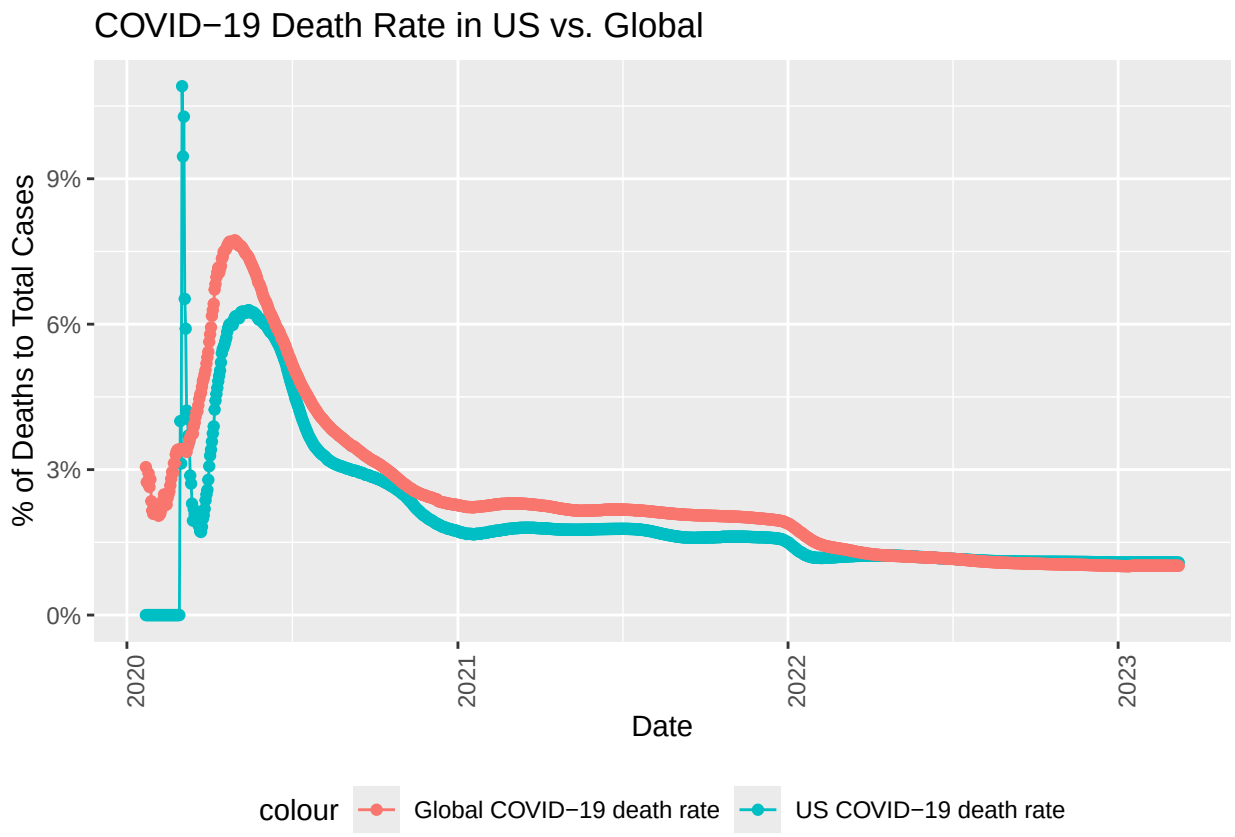
The US COVID-19 death rate was compared against the global death rate.

```
US_cases_death_cases <- US_cases_death_cases %>%
  mutate("US_percent_death_to_total_cases"= death_cases / cases)

global_vs_US_death_rate <- US_cases_death_cases %>% full_join(global_death_cases)
```

```
## Joining with `by = join_by(dates)`
```

```
global_vs_US_death_rate %>% ggplot(aes(x = dates, y = US_percent_death_to_total_cases)) +
  geom_line(aes(color="US COVID-19 death rate")) +
  geom_point(aes(color="US COVID-19 death rate")) +
  geom_line(aes(y = percent_death_to_total_cases, color="Global COVID-19 death rate")) +
  geom_point(aes(y = percent_death_to_total_cases, color="Global COVID-19 death rate")) +
  theme(legend.position="bottom",
        axis.text.x = element_text(angle=90)) +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "COVID-19 Death Rate in US vs. Global",
       y="% of Deaths to Total Cases", x="Date")
```



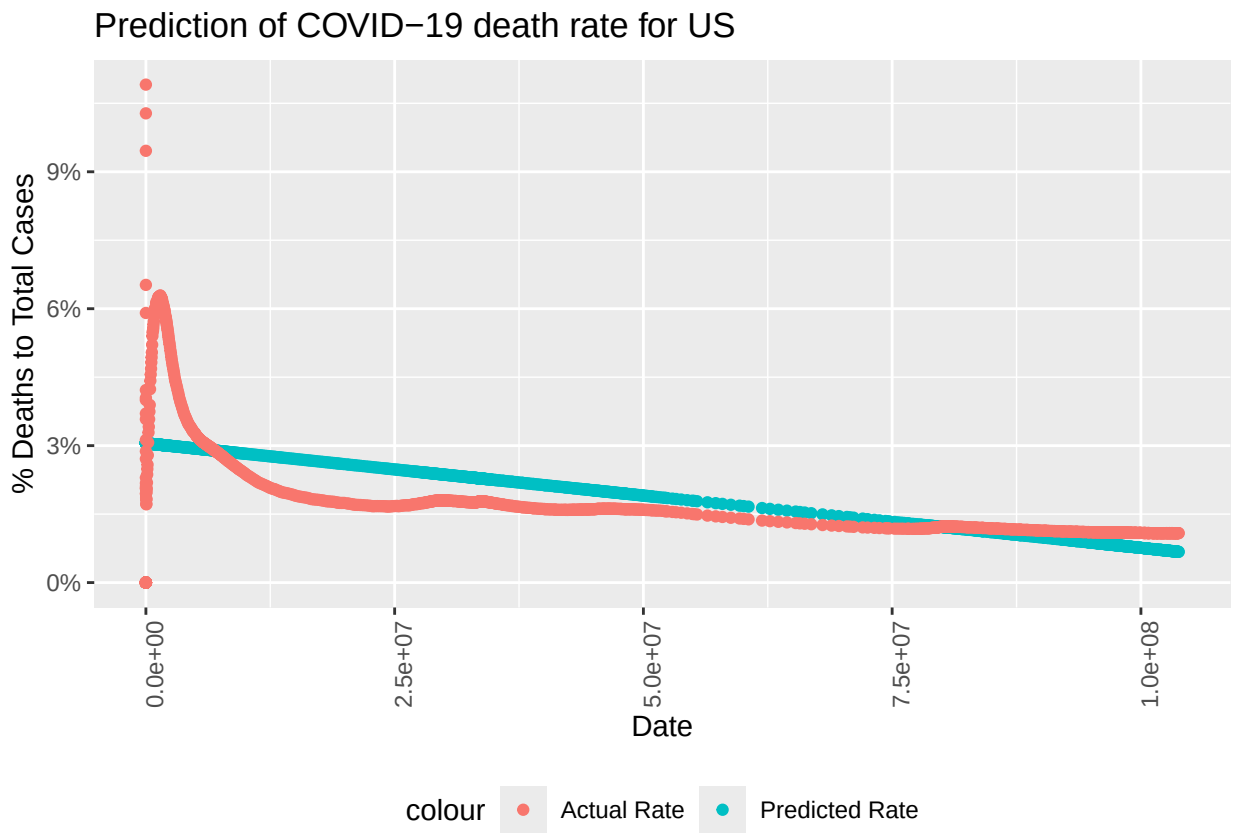
Modeling

To predict the COVID-19 death rate for US and global cases, the below linear model was created.

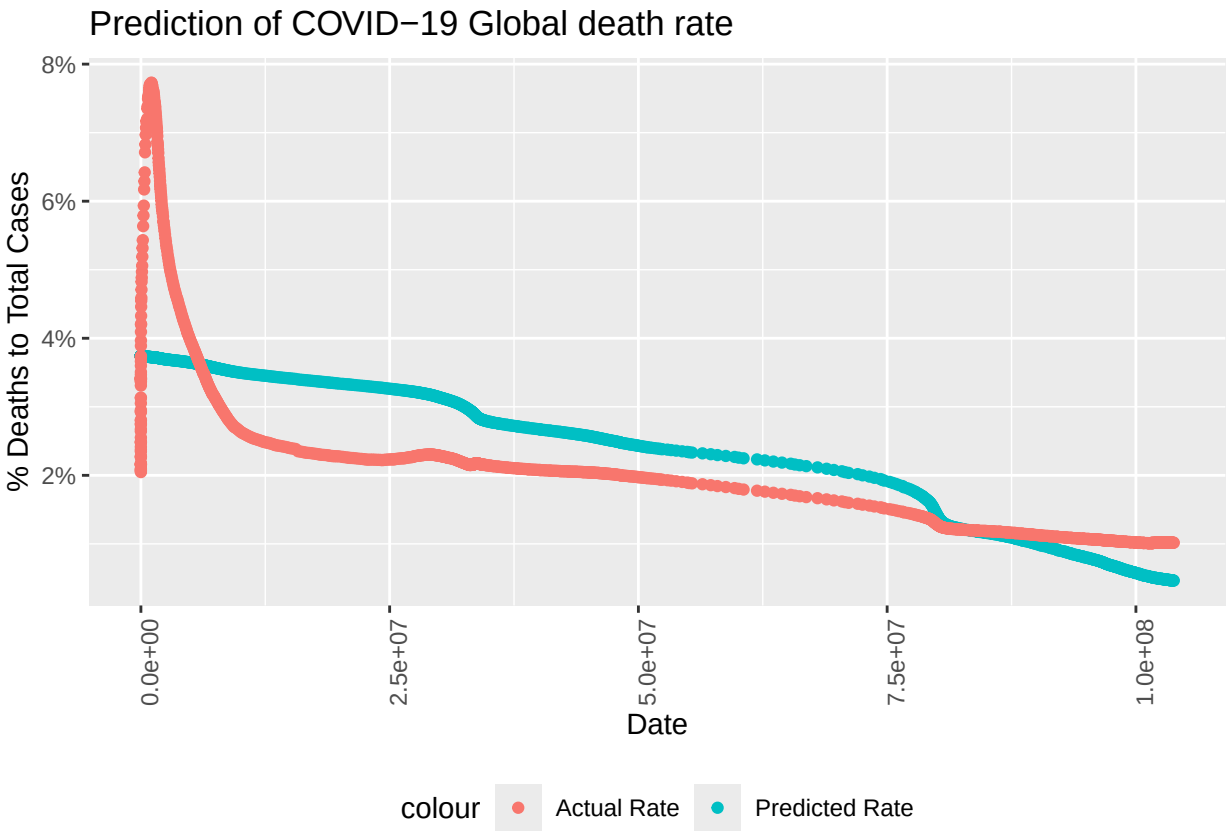
```
mod_US <-lm(US_percent_death_to_total_cases~cases,
            data=global_vs_US_death_rate)
death_rate_pred_us <- global_vs_US_death_rate %>%
  mutate(pred=predict(mod_US))

mod_global <-lm(percent_death_to_total_cases~g_cases,
            data=global_vs_US_death_rate)
death_rate_pred_glob <- global_vs_US_death_rate %>%
  mutate(pred=predict(mod_global))

death_rate_pred_us %>%ggplot(aes(x = cases, y = pred)) +
  geom_point(aes(color = "Predicted Rate"))+
  geom_point(aes(y = US_percent_death_to_total_cases
                 ,color="Actual Rate"))+
  theme(legend.position="bottom",
      axis.text.x=element_text(angle=90))+
  scale_y_continuous(labels = scales::percent) +
  labs(title="Prediction of COVID-19 death rate for US", x="Date",
       y="% Deaths to Total Cases")
```




```
death_rate_pred_glob %>%ggplot(aes(x = cases, y = pred)) +
  geom_point(aes(color = "Predicted Rate"))+
  geom_point(aes(y = percent_death_to_total_cases
                 ,color="Actual Rate"))+
  theme(legend.position="bottom",
      axis.text.x=element_text(angle=90))+
  scale_y_continuous(labels = scales::percent) +
  labs(title="Prediction of COVID-19 Global death rate", x="Date",
       y="% Deaths to Total Cases")
```



Conclusion and Bias Identification

The analysis shows that the amount of COVID-19 cases and deaths from COVID-19 in the US have leveled off since 2021. This likely indicates that the virus is under control. From the start of 2020 to the first quarter, the US COVID-19 death rate was increasing. However, after the first quarter of 2020, the US COVID-19 death rate started decreasing significantly. Since then, the US death rate has decreased to a stable less than 1.5% death rate after 2022. This indicates that the death rate improved by mid 2020 and has leveled off as well. Comparing the US death rate against the global death rate shows that the US death rate was lower between mid 2020 to 2022. This likely indicates that the US COVID-19 cases were less deadly than the global average.

In the US death rate data, there were three outliers identified in the beginning of 2020. Further analysis is required to understand the factors behind the unusual spikes on those dates. A source of bias could be the data collection method, as there may be more testing conducted in some countries compared to others. A lack of testing would result in under-reporting of COVID-19 cases and COVID-19 deaths. To mitigate bias in the results, a log scale was used to visualize the large amount of US COVID-19 cases and deaths. The log scale allowed for a clearer visual on the trends of cases and deaths. Additionally, as the number of death cases are dependent on the number of total cases identified, I chose to use the percentage of deaths to cases when comparing US vs. global trends. This allowed for a more accurate comparison that was not skewed by the differences in total number of cases.