Final-Covid19

Ryan Talbot

2023-11-20

Description of Data

This COVID-19 dataset is from the Johns Hopkins Github site and contains daily time series summary tables, including confirmed, deaths, and recovered. The COVID-19 data repository is operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Since January 21, 2020, this dataset has collected data from sources such as the World Health Organization (WHO), Los Angeles Times, and QQ News, etc. On March 10, 2023, the Johns Hopkins Coronavirus Resource Center ceased its collecting and reporting of global COVID-19 data.

(Please refer to https://github.com/CSSEGISandData/COVID-19 for additional information about this dataset.)

Import Packages

```r
# Add libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## 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
```

```r
library(readr)
library(dplyr)
library(lubridate)
library(ggplot2)
```

Import the Data

Copy the link address of the csv file from github.

```
## [1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [3] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [4] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
```

```
# `read_csv()` usecd to read in the data to variables
global_cases = read_csv(urls[1])
global_deaths = read_csv(urls[2])
```

```
# first rows of csv files get better understanding for tidy
head(global_cases)
```

```
## # A tibble: 6 x 1,147
##   ‘Province/State‘ ‘Country/Region‘   Lat  Long ‘1/22/20‘ ‘1/23/20‘ ‘1/24/20‘
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: ‘1/25/20‘ <dbl>, ‘1/26/20‘ <dbl>, ‘1/27/20‘ <dbl>,
## #   ‘1/28/20‘ <dbl>, ‘1/29/20‘ <dbl>, ‘1/30/20‘ <dbl>, ‘1/31/20‘ <dbl>,
## #   ‘2/1/20‘ <dbl>, ‘2/2/20‘ <dbl>, ‘2/3/20‘ <dbl>, ‘2/4/20‘ <dbl>,
## #   ‘2/5/20‘ <dbl>, ‘2/6/20‘ <dbl>, ‘2/7/20‘ <dbl>, ‘2/8/20‘ <dbl>,
## #   ‘2/9/20‘ <dbl>, ‘2/10/20‘ <dbl>, ‘2/11/20‘ <dbl>, ‘2/12/20‘ <dbl>,
## #   ‘2/13/20‘ <dbl>, ‘2/14/20‘ <dbl>, ‘2/15/20‘ <dbl>, ‘2/16/20‘ <dbl>,
## #   ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>, ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ...
```

```
head(global_deaths)
```

```
## # A tibble: 6 x 1,147
##   ‘Province/State‘ ‘Country/Region‘   Lat  Long ‘1/22/20‘ ‘1/23/20‘ ‘1/24/20‘
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: ‘1/25/20‘ <dbl>, ‘1/26/20‘ <dbl>, ‘1/27/20‘ <dbl>,
## #   ‘1/28/20‘ <dbl>, ‘1/29/20‘ <dbl>, ‘1/30/20‘ <dbl>, ‘1/31/20‘ <dbl>,
## #   ‘2/1/20‘ <dbl>, ‘2/2/20‘ <dbl>, ‘2/3/20‘ <dbl>, ‘2/4/20‘ <dbl>,
## #   ‘2/5/20‘ <dbl>, ‘2/6/20‘ <dbl>, ‘2/7/20‘ <dbl>, ‘2/8/20‘ <dbl>,
## #   ‘2/9/20‘ <dbl>, ‘2/10/20‘ <dbl>, ‘2/11/20‘ <dbl>, ‘2/12/20‘ <dbl>,
## #   ‘2/13/20‘ <dbl>, ‘2/14/20‘ <dbl>, ‘2/15/20‘ <dbl>, ‘2/16/20‘ <dbl>,
## #   ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>, ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ...
```

Tidy Data

1. **Tidy the Data** - Put each variable (**date**, **cases**, and **deaths**) in their own column.

- Remove columns: **Lat** and **Long**.
- Rename columns: **Province/State** and **Country/Region**.
- Convert column **date** to date object.
-

```
# Pivot wide-format data for dates and sum totals for each state
covid_global_deaths = global_deaths %>%
  pivot_longer(cols = 13:ncol(global_deaths), names_to = "date") %>%
  group_by(`Country/Region`, `Province/State`, date) %>%
  summarise("cumulative_deaths" = sum(value, na.rm = TRUE), .groups = 'drop')

covid_global_cases = global_cases %>%
  pivot_longer(cols = 13:ncol(global_cases), names_to = "date") %>%
  group_by(`Country/Region`, `Province/State`, date) %>%
  summarise("cumulative_cases" = sum(value, na.rm = TRUE), .groups = 'drop')


# Convert dates to datetime object
covid_global_deaths$date = lubridate::mdy(covid_global_deaths$date)
covid_global_cases$date = lubridate::mdy(covid_global_cases$date)

# Rename columns from Province_State -> State & Admin2 -> County
covid_global_deaths = covid_global_deaths %>%
  rename_at('Province/State', ~'State') %>%
  rename_at('Country/Region', ~'Country')

covid_global_cases = covid_global_cases %>%
  rename_at('Province/State', ~'State') %>%
  rename_at('Country/Region', ~'Country')


# check global deaths and cases data
head(covid_global_deaths)
```

```
## # A tibble: 6 x 4
##   Country     State date       cumulative_deaths
##   <chr>       <chr> <date>                 <dbl>
## 1 Afghanistan <NA>  2021-01-01              2201
## 2 Afghanistan <NA>  2022-01-01              7356
## 3 Afghanistan <NA>  2023-01-01              7849
## 4 Afghanistan <NA>  2021-01-10              2277
## 5 Afghanistan <NA>  2022-01-10              7373
## 6 Afghanistan <NA>  2023-01-10              7854
```

```
head(covid_global_cases)
```

```
## # A tibble: 6 x 4
##   Country     State date       cumulative_cases
##   <chr>       <chr> <date>                <dbl>
## 1 Afghanistan <NA>  2021-01-01            52513
## 2 Afghanistan <NA>  2022-01-01           158107
## 3 Afghanistan <NA>  2023-01-01           207616
## 4 Afghanistan <NA>  2021-01-10            53489
## 5 Afghanistan <NA>  2022-01-10           158394
## 6 Afghanistan <NA>  2023-01-10           207866
```

```r
# merge global data sets and filter to get data just for Switzerland and for germany
world = merge(x=covid_global_deaths, y=covid_global_cases, all.x=TRUE)
ch <- world[world$Country == "Switzerland", ]
de <- world[world$Country == "Germany", ]


de_tidy <- de %>% select(-State)
ch_tidy <- ch %>% select(-State)

# View first several lines of each data set

head(ch_tidy)
```

```
##            Country       date cumulative_deaths cumulative_cases
## 280346 Switzerland 2020-01-30                 0                0
## 280347 Switzerland 2020-01-31                 0                0
## 280348 Switzerland 2020-02-01                 0                0
## 280349 Switzerland 2020-02-02                 0                0
## 280350 Switzerland 2020-02-03                 0                0
## 280351 Switzerland 2020-02-04                 0                0
```

```r
head(de_tidy)
```

```
##        Country       date cumulative_deaths cumulative_cases
## 153226 Germany 2020-01-30                 0                4
## 153227 Germany 2020-01-31                 0                5
## 153228 Germany 2020-02-01                 0                8
## 153229 Germany 2020-02-02                 0               10
## 153230 Germany 2020-02-03                 0               12
## 153231 Germany 2020-02-04                 0               12
```

Step 3: Add Visualizations and Analysis

Question 1: What are the trends for cases and deaths of COVID-19 comparing, Germany and Switzerland, viewing per capita, deaths per case

more than half of the population of switzerland (8.703 million) had covid. while less than half of germanys population (83.2 million) had covid. .203% of the population died from covid in germany, while Switzerland had .163% of population. This is interesting, you'd think that the more cases per capita would result in more deaths per capita.

```r
# Adding a per capita column to both data sets and viewing their summaries
de_per_cap <- de_tidy %>%
  mutate(
    de_deaths_per_capita = cumulative_deaths / 83.2e6,
    de_cases_per_capita = cumulative_cases / 83.2e6
  )


ch_per_cap <- ch_tidy %>%
  mutate(
    ch_deaths_per_capita = cumulative_deaths / 8703000,
    ch_cases_per_capita = cumulative_cases / 8703000
  )

summary(de_per_cap)
```

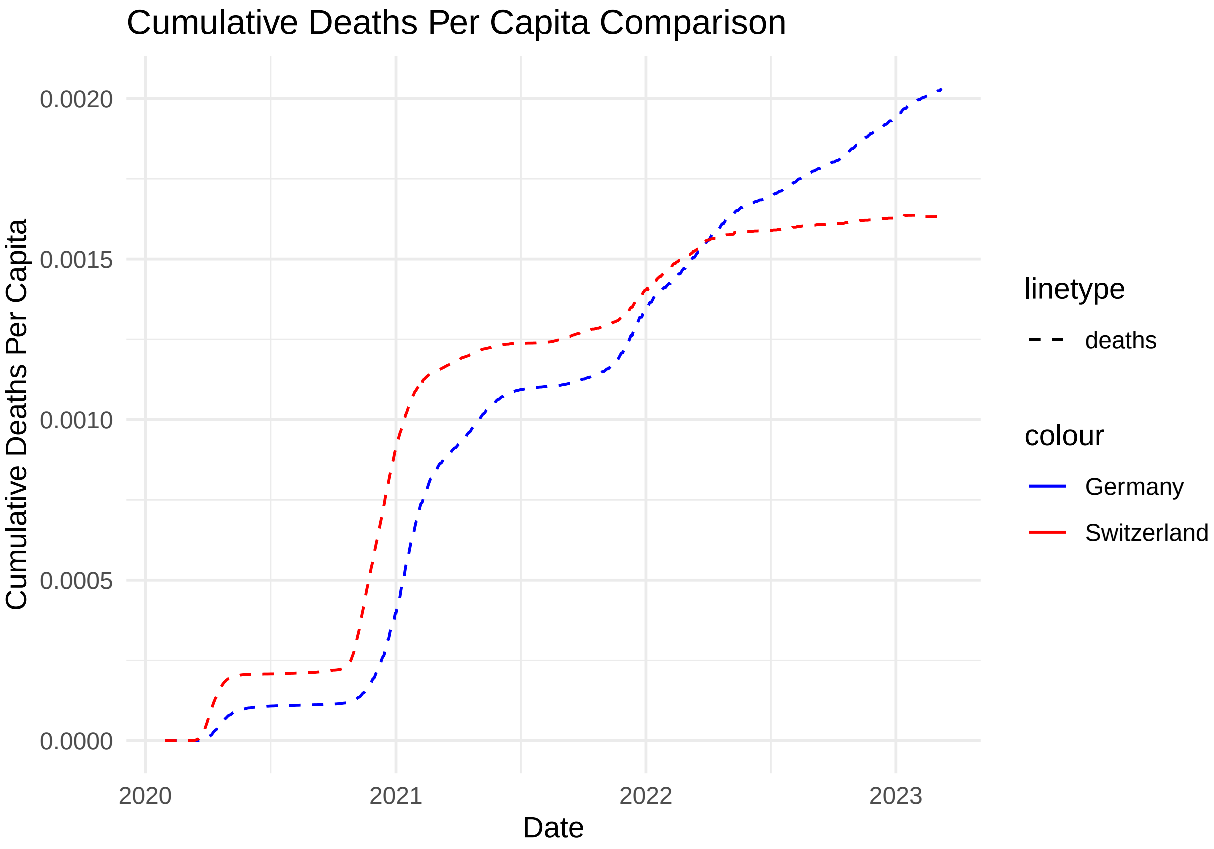
```
##    Country               date            cumulative_deaths cumulative_cases
##  Length:1135        Min.   :2020-01-30   Min.   :     0    Min.   :       4
##  Class :character   1st Qu.:2020-11-08   1st Qu.: 11320    1st Qu.:  665186
##  Mode  :character   Median :2021-08-19   Median : 91943    Median : 3843775
##                     Mean   :2021-08-19   Mean   : 84633    Mean   :12058188
##                     3rd Qu.:2022-05-29   3rd Qu.:138864    3rd Qu.:26244107
##                     Max.   :2023-03-09   Max.   :168935    Max.   :38249060
##  de_deaths_per_capita de_cases_per_capita
##  Min.   :0.0000000    Min.   :0.000000
##  1st Qu.:0.0001361    1st Qu.:0.007995
##  Median :0.0011051    Median :0.046199
##  Mean   :0.0010172    Mean   :0.144930
##  3rd Qu.:0.0016690    3rd Qu.:0.315434
##  Max.   :0.0020305    Max.   :0.459724
```

```r
summary(ch_per_cap)
```

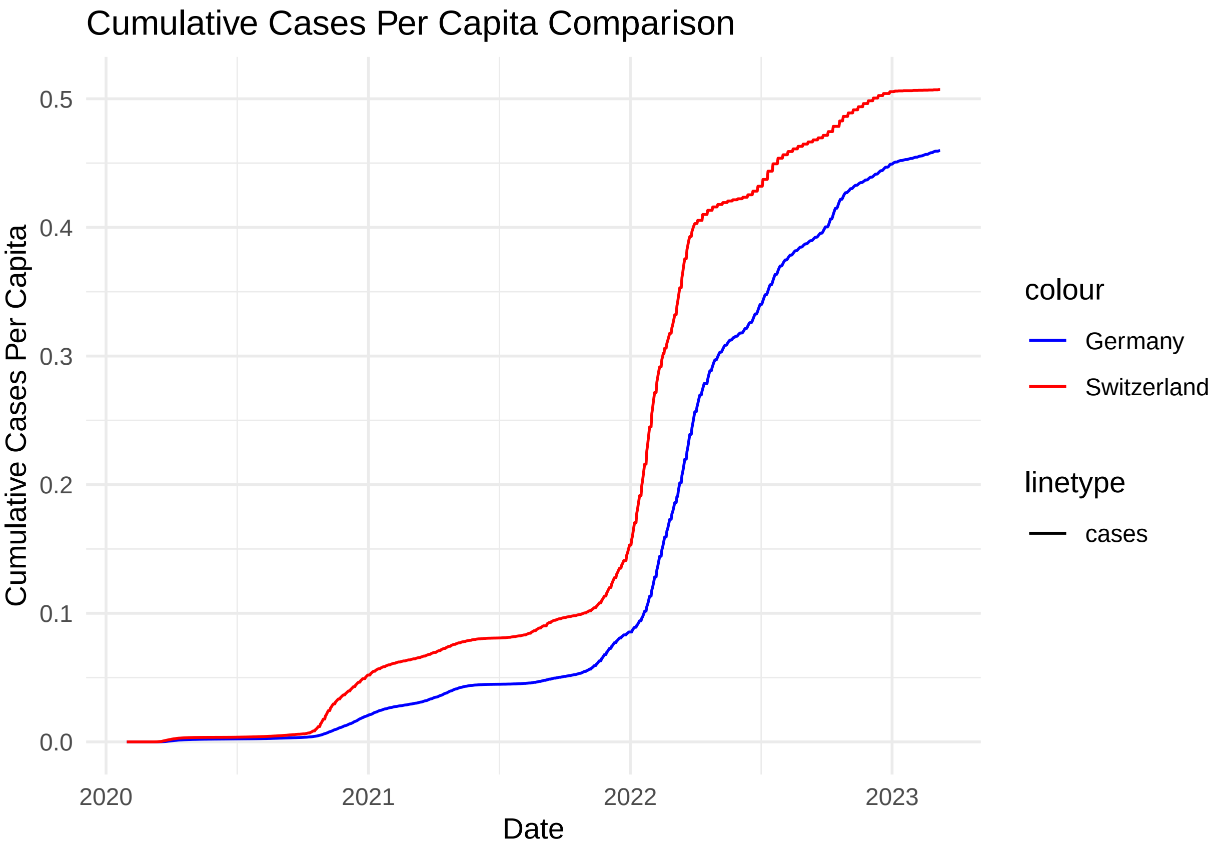
```
##    Country               date            cumulative_deaths cumulative_cases
##  Length:1135        Min.   :2020-01-30   Min.   :    0     Min.   :      0
##  Class :character   1st Qu.:2020-11-08   1st Qu.: 3047     1st Qu.: 220568
##  Mode  :character   Median :2021-08-19   Median :10828     Median : 750186
##                     Mean   :2021-08-19   Mean   : 9283     Mean   :1685445
##                     3rd Qu.:2022-05-29   3rd Qu.:13796     3rd Qu.:3668054
##                     Max.   :2023-03-09   Max.   :14244     Max.   :4413911
##  ch_deaths_per_capita ch_cases_per_capita
##  Min.   :0.0000000    Min.   :0.00000
##  1st Qu.:0.0003501    1st Qu.:0.02534
##  Median :0.0012442    Median :0.08620
##  Mean   :0.0010667    Mean   :0.19366
##  3rd Qu.:0.0015853    3rd Qu.:0.42147
##  Max.   :0.0016367    Max.   :0.50717
```

```r
# Merge Germany and Switzerland per capita data
combined_data <- merge(x = de_per_cap, y = ch_per_cap, by = "date", all = TRUE)

# Create line plots for deaths per capita
ggplot(combined_data, aes(x = date, y = de_deaths_per_capita, color = "Germany", linetype = "deaths"))
  geom_line() +
  geom_line(aes(y = ch_deaths_per_capita, color = "Switzerland", linetype = "deaths")) +
  labs(title = "Cumulative Deaths Per Capita Comparison",
       x = "Date",
       y = "Cumulative Deaths Per Capita") +
  scale_color_manual(values = c("Germany" = "blue", "Switzerland" = "red")) +
  scale_linetype_manual(values = c("dashed")) +
  theme_minimal()
```

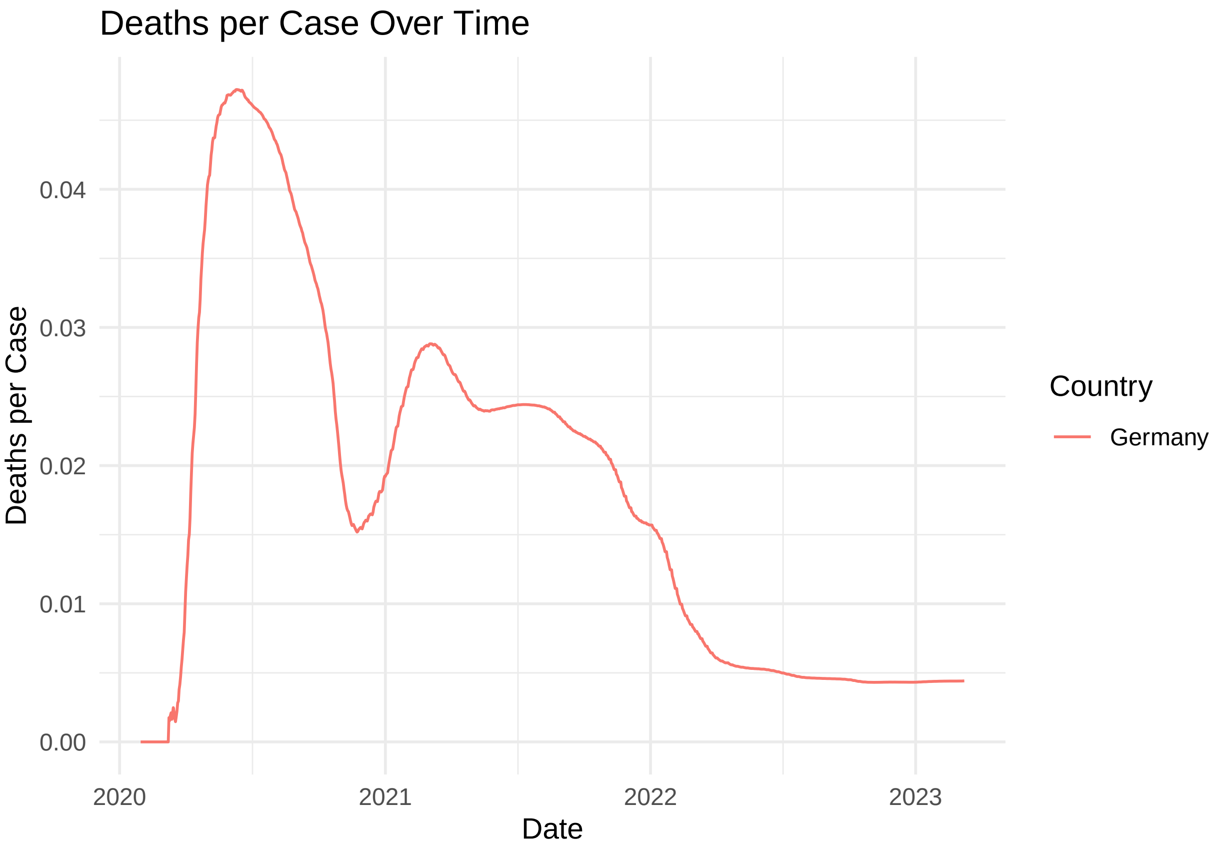


```
# Create line plots for cases per capita
ggplot(combined_data, aes(x = date, y = de_cases_per_capita, color = "Germany", linetype = "cases")) +
  geom_line() +
  geom_line(aes(y = ch_cases_per_capita, color = "Switzerland", linetype = "cases")) +
  labs(title = "Cumulative Cases Per Capita Comparison",
       x = "Date",
       y = "Cumulative Cases Per Capita") +
  scale_color_manual(values = c("Germany" = "blue", "Switzerland" = "red")) +
  scale_linetype_manual(values = c("solid")) +
  theme_minimal()
```



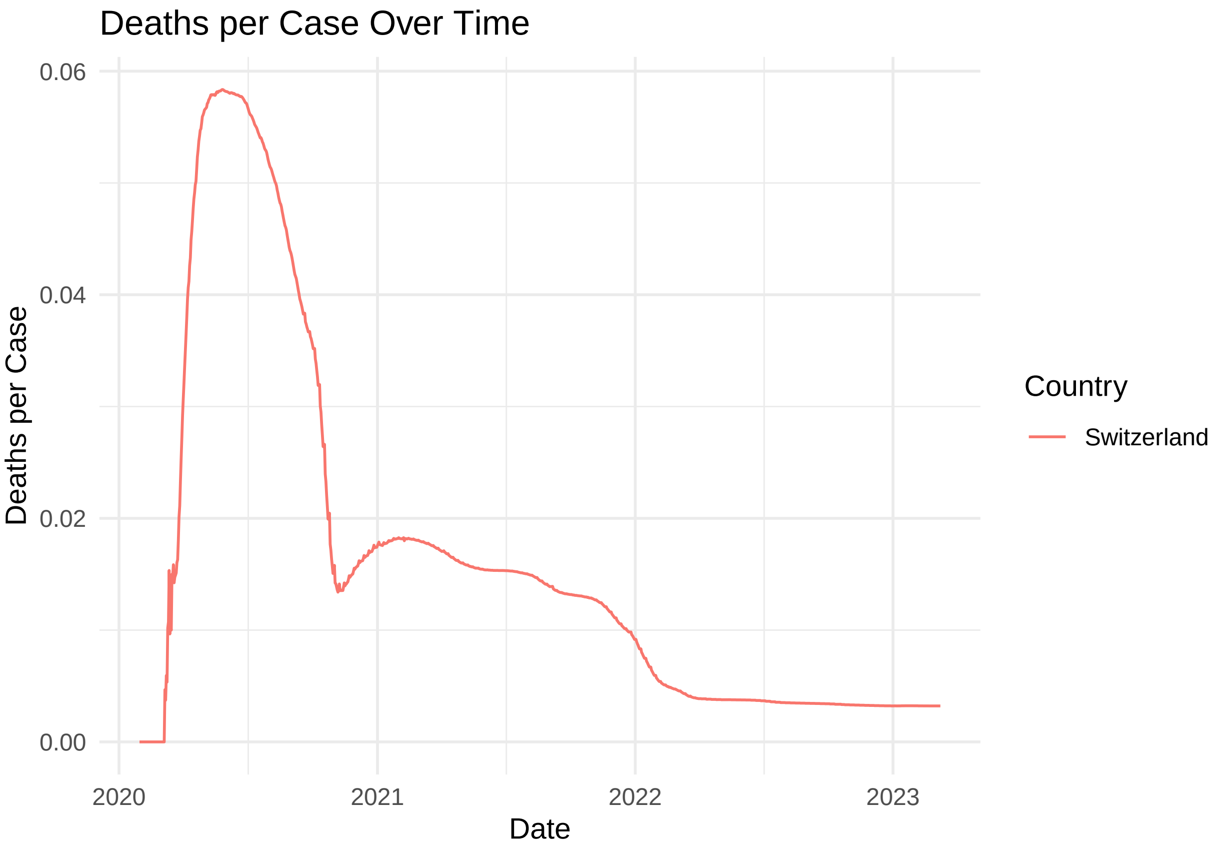
```r
# Calculate deaths per case with a check for division by zero
de_tidy$deaths_per_case <- ifelse(de_tidy$cumulative_cases > 0,
                                  de_tidy$cumulative_deaths / de_tidy$cumulative_cases,
                                  0)

# Create a time series plot
ggplot(de_tidy, aes(x = date, y = deaths_per_case, color = Country)) +
  geom_line() +
  labs(title = "Deaths per Case Over Time",
       x = "Date",
       y = "Deaths per Case",
       color = "Country") +
  theme_minimal()
```



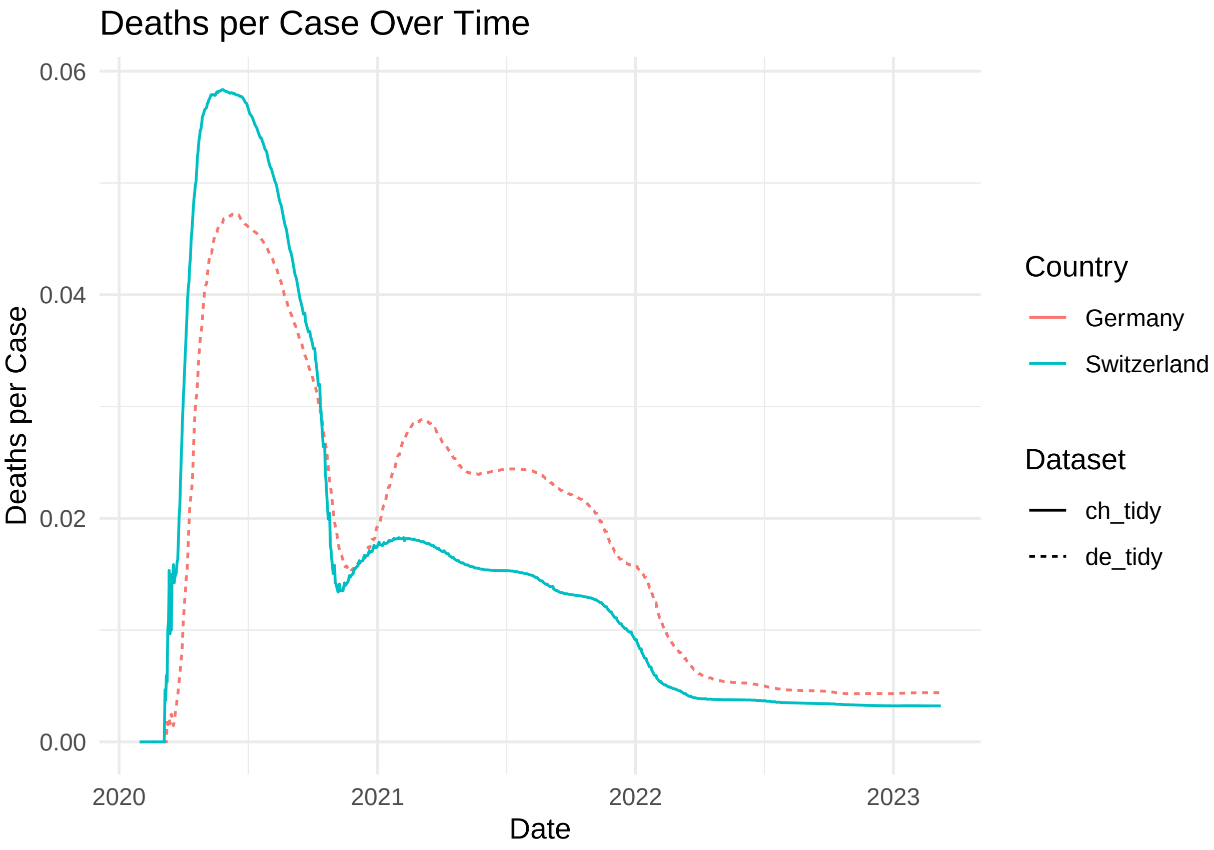
```
ch_tidy$deaths_per_case <- ifelse(ch_tidy$cumulative_cases > 0,
                                   ch_tidy$cumulative_deaths / ch_tidy$cumulative_cases,
                                   0)

# Create a time series plot
ggplot(ch_tidy, aes(x = date, y = deaths_per_case, color = Country)) +
  geom_line() +
  labs(title = "Deaths per Case Over Time",
       x = "Date",
       y = "Deaths per Case",
       color = "Country") +
  theme_minimal()
```



```r
combined_df <- rbind(transform(de_tidy, dataset = "de_tidy"),
                     transform(ch_tidy, dataset = "ch_tidy"))

# Create a time series plot with multiple datasets
ggplot(combined_df, aes(x = date, y = deaths_per_case, color = Country, linetype = dataset)) +
  geom_line() +
  labs(title = "Deaths per Case Over Time",
       x = "Date",
       y = "Deaths per Case",
       color = "Country",
       linetype = "Dataset") +
  theme_minimal()
```


Deaths per Case Over Time
0.06
0.04
0.02
0.00
Deaths per Case
2020
2021
2022
2023
Date
Country
Germany
Switzerland
Dataset
ch_tidy
de_tidy