JHU COVID-19 Data

K. Katzmann

2024-08-05

Introduction

This analysis aims to identify trends in new COVID-19 cases and deaths over time, and to better understand how the ratio of deaths to confirmed cases changes over time. Is there evidence, based on the ratio of confirmed cases to deaths over time, that treatments have improved throughout the pandemic?

To answer this question, we will explore COVID-19 datasets from the Johns Hopkins University COVID-19 data repository. These datasets contain information on the number of confirmed COVID-19 cases and deaths at both the US and global levels from January 2020 to March 2023.

Importing Data

We'll begin the analysis by installing necessary packages and importing data from the Johns Hopkins University COVID-19 data repository.

```
# load necessary libraries
library(tidyverse)
library(lubridate)

# set theme for ggplot
theme_set(theme_minimal())
```

We'll import the datasets for US confirmed cases, US deaths, global confirmed cases and global deaths.

```
covid_start <- "2020-01-01"

base_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c

file_names <- c(
  "time_series_covid19_confirmed_US.csv",
  "time_series_covid19_deaths_US.csv",
  "time_series_covid19_confirmed_global.csv",
  "time_series_covid19_deaths_global.csv"
)

confirmed_us_raw <- read_csv(paste0(base_url, file_names[1]))
deaths_us_raw <- read_csv(paste0(base_url, file_names[2]))
confirmed_global_raw <- read_csv(paste0(base_url, file_names[3]))
deaths_global_raw <- read_csv(paste0(base_url, file_names[4]))
```

Next, let's inspect the data to understand its structure and contents.

```
confirmed_us_raw
```

```
## # A tibble: 3,342 x 1,154
##         UID iso2  iso3  code3  FIPS Admin2   Province_State Country_Region   Lat
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>    <chr>          <chr>          <dbl>
##  1 84001001 US    USA     840  1001 Autauga  Alabama        US              32.5
##  2 84001003 US    USA     840  1003 Baldwin  Alabama        US              30.7
##  3 84001005 US    USA     840  1005 Barbour  Alabama        US              31.9
##  4 84001007 US    USA     840  1007 Bibb     Alabama        US              33.0
##  5 84001009 US    USA     840  1009 Blount   Alabama        US              34.0
##  6 84001011 US    USA     840  1011 Bullock  Alabama        US              32.1
##  7 84001013 US    USA     840  1013 Butler   Alabama        US              31.8
##  8 84001015 US    USA     840  1015 Calhoun  Alabama        US              33.8
##  9 84001017 US    USA     840  1017 Chambers Alabama        US              32.9
## 10 84001019 US    USA     840  1019 Cherokee Alabama        US              34.2
## # i 3,332 more rows
## # i 1,145 more variables: Long_ <dbl>, Combined_Key <chr>, `1/22/20` <dbl>,
## #   `1/23/20` <dbl>, `1/24/20` <dbl>, `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>, ...
```

```
deaths_us_raw
```

```
## # A tibble: 3,342 x 1,155
##         UID iso2  iso3  code3  FIPS Admin2   Province_State Country_Region   Lat
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>    <chr>          <chr>          <dbl>
##  1 84001001 US    USA     840  1001 Autauga  Alabama        US              32.5
##  2 84001003 US    USA     840  1003 Baldwin  Alabama        US              30.7
##  3 84001005 US    USA     840  1005 Barbour  Alabama        US              31.9
##  4 84001007 US    USA     840  1007 Bibb     Alabama        US              33.0
##  5 84001009 US    USA     840  1009 Blount   Alabama        US              34.0
##  6 84001011 US    USA     840  1011 Bullock  Alabama        US              32.1
##  7 84001013 US    USA     840  1013 Butler   Alabama        US              31.8
##  8 84001015 US    USA     840  1015 Calhoun  Alabama        US              33.8
##  9 84001017 US    USA     840  1017 Chambers Alabama        US              32.9
## 10 84001019 US    USA     840  1019 Cherokee Alabama        US              34.2
## # i 3,332 more rows
## # i 1,146 more variables: Long_ <dbl>, Combined_Key <chr>, Population <dbl>,
## #   `1/22/20` <dbl>, `1/23/20` <dbl>, `1/24/20` <dbl>, `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>, ...
```

```
confirmed_global_raw
```

```
## # A tibble: 289 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
##  7 <NA>               Antigua and Bar~  17.1 -61.8          0         0         0
##  8 <NA>               Argentina        -38.4 -63.6          0         0         0
##  9 <NA>               Armenia           40.1  45.0          0         0         0
## 10 Australian Capit~ Australia        -35.5 149.           0         0         0
## # i 279 more rows
## # 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>, ...
```

```
deaths_global_raw
```

```
## # A tibble: 289 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
##  7 <NA>               Antigua and Bar~  17.1 -61.8          0         0         0
##  8 <NA>               Argentina        -38.4 -63.6          0         0         0
##  9 <NA>               Armenia           40.1  45.0          0         0         0
## 10 Australian Capit~ Australia        -35.5 149.           0         0         0
## # i 279 more rows
## # 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>, ...
```

Tidying and Transforming Data

After inspecting the data, we see that there is one row per location and columns for each date with the number of confirmed cases and deaths on that date. We also see that the number is cumulative, so we will need to calculate the new cases and deaths per day. From there, we will calculate the ratio of deaths to confirmed cases to see if there are any trends over time.

We'll start with the US confirmed cases and deaths datasets. We'll begin tidying by removing unassigned locations and columns that are not needed for our analysis. Then we'll pivot the date columns to rows and calculate new cases and deaths per day. Finally, we'll join the two datasets together and calculate the ratio of deaths to confirmed cases.

```
# tidy US data
confirmed_us <- confirmed_us_raw %>%
  # remove unneeded columns
  select(-c(iso2, iso3, code3, FIPS, Lat, Long_, Combined_Key)) %>%
  # filter out bad or unneeded data
  filter(Admin2 != 'Unassigned',
         Country_Region == 'US') %>%
  # pivot date columns to rows
  pivot_longer(cols = -c(UID, Admin2, Province_State, Country_Region),
               names_to = "Date", values_to = "Confirmed") %>%
  # convert data types
  mutate(Date = mdy(Date)) %>%
  # calculate new cases per day
  arrange(UID, Date) %>%
  mutate(Lag_Confirmed = lag(Confirmed),
         # we need to lag the UID as well so that we don't compare cases between different places!
         Lag_UID = lag(UID),
         New_Confirmed = case_when(
             UID == Lag_UID ~ (Confirmed - Lag_Confirmed),
             TRUE ~ NA
         ))

deaths_us <- deaths_us_raw %>%
  # remove unneeded columns
  select(-c(iso2, iso3, code3, FIPS, Lat, Long_, Combined_Key)) %>%
  # filter out bad or unneeded data
  filter(Admin2 != 'Unassigned',
         Country_Region == 'US') %>%
  # pivot date columns to rows
  pivot_longer(cols = -c(UID, Admin2, Province_State, Country_Region, Population),
              names_to = "Date", values_to = "Deaths") %>%
  # convert data types
  mutate(Date = mdy(Date)) %>%
  # calculate new deaths per day
  arrange(UID, Date) %>%
  mutate(Lag_Deaths = lag(Deaths),
         # we need to lag the UID as well so that we don't compare cases between different places!
         Lag_UID = lag(UID),
         New_Deaths = case_when(
             UID == Lag_UID ~ (Deaths - Lag_Deaths),
             TRUE ~ NA
         ))

all_us <- confirmed_us %>%
  # join deaths data. We only need a few columns
  left_join(deaths_us %>%
              select(UID, Date, Deaths, New_Deaths),
            by = c("UID", "Date")) %>%
  # add a couple of columns for analysis
  mutate(Month_Year = format(Date, '%Y-%m'),
         Months_Since_Start = interval(covid_start, Date) %/% months(1)) %>%
  # aggregating to reduce variance in ratios
  group_by(Country_Region, Province_State, Month_Year, Months_Since_Start) %>%
```

```r
  summarise(New_Deaths = sum(New_Deaths),
            New_Confirmed = sum(New_Confirmed)) %>%
  # calculate ratio of deaths to confirmed cases
  mutate(Deaths_to_Cases_Ratio = ifelse(New_Confirmed == 0, NA, (New_Deaths * 1000 / New_Confirmed))) %>
  # cases less than 100 can affect ratio too much
  filter(New_Confirmed >= 100)
```

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

```r
all_us
```

```
## # A tibble: 1,915 x 7
## # Groups:   Country_Region, Province_State, Month_Year [1,915]
##    Country_Region Province_State Month_Year Months_Since_Start New_Deaths
##    <chr>          <chr>          <chr>                   <dbl>      <dbl>
##  1 US             Alabama        2020-03                     2         15
##  2 US             Alabama        2020-04                     3        257
##  3 US             Alabama        2020-05                     4        358
##  4 US             Alabama        2020-06                     5        296
##  5 US             Alabama        2020-07                     6        605
##  6 US             Alabama        2020-08                     7        552
##  7 US             Alabama        2020-09                     8        457
##  8 US             Alabama        2020-10                     9        427
##  9 US             Alabama        2020-11                    10        611
## 10 US             Alabama        2020-12                    11       1249
## # i 1,905 more rows
## # i 2 more variables: New_Confirmed <dbl>, Deaths_to_Cases_Ratio <dbl>
```

Now we'll repeat the same process for the global datasets. We'll remove rows and columns that we won't need for our analysis, then we'll pivot the date columns to rows and calculate new cases and deaths per day. Finally, we'll join the two datasets together and calculate the ratio of deaths to confirmed cases.

```r
# tidy global data
confirmed_global <- confirmed_global_raw %>%
  # rename columns
  rename(Province_State = `Province/State`, Country_Region = `Country/Region`) %>%
  # filter out bad or unneeded data
  filter(!Country_Region %in% c("Korea, North", "Antarctica")) %>%
  # pivot date columns to rows
  pivot_longer(cols = -c(Province_State, Country_Region, Lat, Long),
               names_to = "Date", values_to = "Confirmed") %>%
  # convert data types and create an ID column to use later on
  mutate(Date = mdy(Date),
         Global_ID = paste(Country_Region, Province_State, sep = "|")) %>%
  # calculate new cases per day
  arrange(Global_ID, Date) %>%
  mutate(Lag_Confirmed = lag(Confirmed),
         # we need to lag the Global_ID as well so that we don't compare cases between different places
         Lag_Global_ID = lag(Global_ID),
         New_Confirmed = case_when(
           Global_ID == Lag_Global_ID ~ (Confirmed - Lag_Confirmed),
```

```r
          TRUE ~ NA
        ))

deaths_global <- deaths_global_raw %>%
  # rename columns
  rename(Province_State = `Province/State`, Country_Region = `Country/Region`) %>%
  # filter out bad or unneeded data
  filter(!Country_Region %in% c("Korea, North", "Antarctica")) %>%
  # pivot date columns to rows
  pivot_longer(cols = -c(Province_State, Country_Region, Lat, Long),
               names_to = "Date", values_to = "Deaths") %>%
  # convert data types and create an ID column to use later on
  mutate(Date = mdy(Date),
         Global_ID = paste(Country_Region, Province_State, sep = "|")) %>%
  # calculate new deaths per day
  arrange(Global_ID, Date) %>%
  mutate(Lag_Deaths = lag(Deaths),
         # we need to lag the UID as well so that we don't compare cases between different places!
         Lag_Global_ID = lag(Global_ID),
         New_Deaths = case_when(
           Global_ID == Lag_Global_ID ~ (Deaths - Lag_Deaths),
           TRUE ~ NA
         ))

all_global <- confirmed_global %>%
  # join deaths data. We only need a few columns
  left_join(deaths_global %>%
              select(Global_ID, Date, Deaths, New_Deaths),
            by = c("Global_ID", "Date")) %>%
  # add a couple of columns for analysis
  mutate(Month_Year = format(Date, '%Y-%m'),
         Months_Since_Start = interval(covid_start, Date) %/% months(1)) %>%
  # aggregating to reduce variance in ratios
  group_by(Country_Region, Month_Year, Months_Since_Start) %>%
  summarise(New_Deaths = sum(New_Deaths),
            New_Confirmed = sum(New_Confirmed)) %>%
  # calculate ratio of deaths to confirmed cases
  mutate(Deaths_to_Cases_Ratio = ifelse(New_Confirmed == 0, NA, (New_Deaths * 1000/ New_Confirmed))) %>%
  # cases less than 100 can affect ratio too much
  filter(New_Confirmed >= 100)
```

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

```r
all_global
```

```
## # A tibble: 5,814 x 6
## # Groups:   Country_Region, Month_Year [5,814]
##    Country_Region Month_Year Months_Since_Start New_Deaths New_Confirmed
##    <chr>          <chr>                   <dbl>      <dbl>         <dbl>
##  1 Afghanistan    2020-03                     2          4           161
##  2 Afghanistan    2020-04                     3         56          1661
##  3 Afghanistan    2020-05                     4        194         13353
```

```
## 4 Afghanistan    2020-06                          5        485         16265
## 5 Afghanistan    2020-07                          6        536          5183
## 6 Afghanistan    2020-08                          7        131          1620
## 7 Afghanistan    2020-09                          8         56          1106
## 8 Afghanistan    2020-10                          9         71          1980
## 9 Afghanistan    2020-11                         10        230          4881
## 10 Afghanistan   2020-12                         11        426          6115
## # i 5,804 more rows
## # i 1 more variable: Deaths_to_Cases_Ratio <dbl>
```

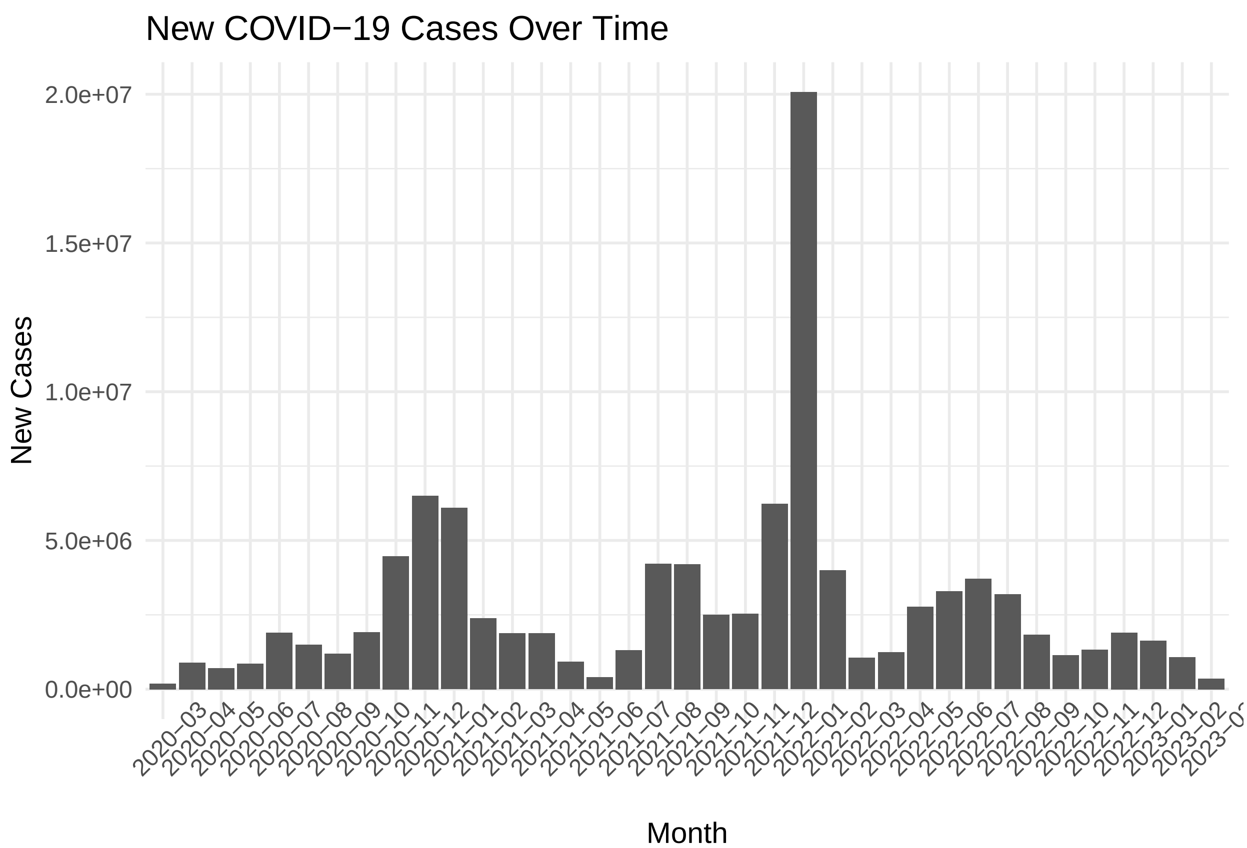
Exploratory Data Analysis

US Analysis

We'll start our exploratory data analysis with the US data to see whether there is any evidence of improvements in the ratio of deaths to confirmed cases over time.

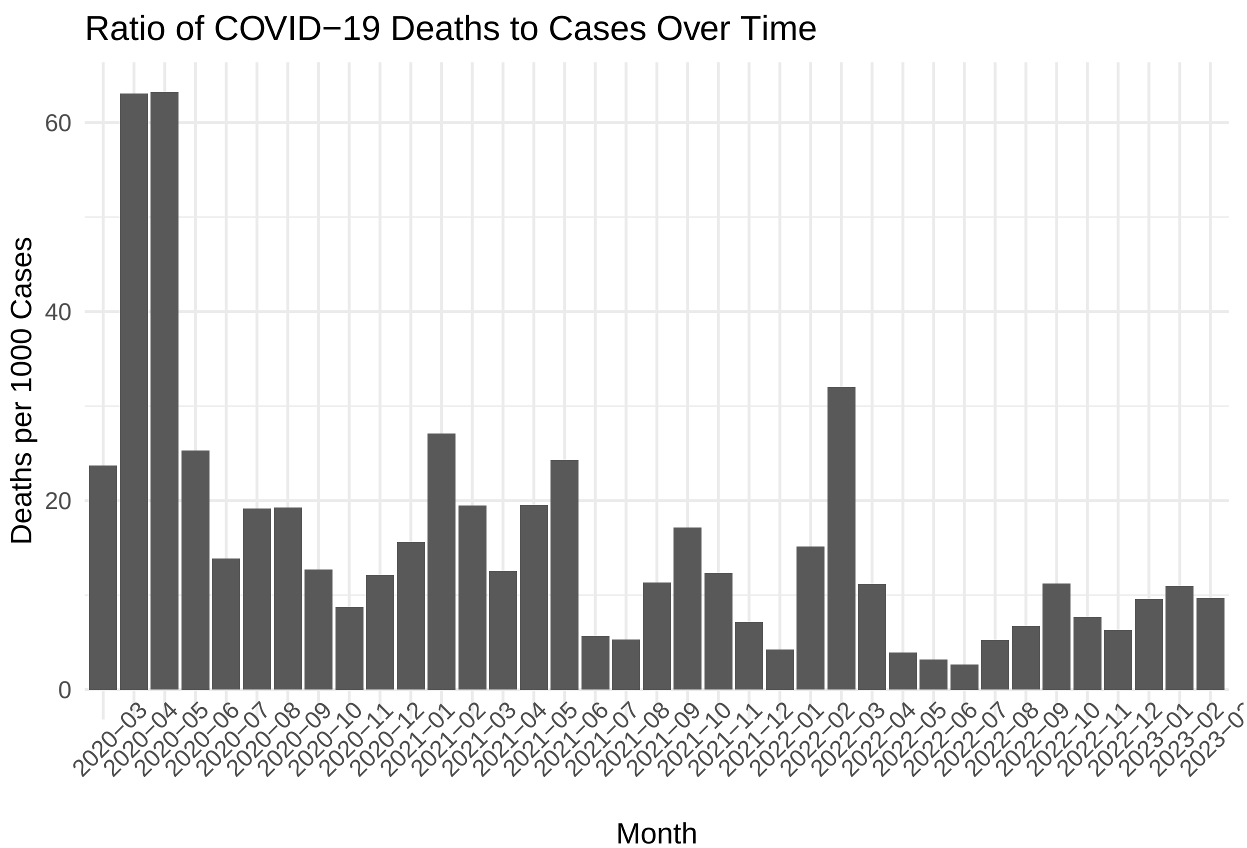
First, let's plot the total number of new cases over time.

```
all_us %>%
  group_by(Month_Year) %>%
  summarise(Mean_New_Confirmed = mean(New_Confirmed),
            Sum_New_Confirmed = sum(New_Confirmed)) %>%
  ggplot(aes(x = Month_Year, y = Sum_New_Confirmed)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "New COVID-19 Cases Over Time",
       x = "Month",
       y = "New Cases")
```



Next, let's plot the ratio of deaths to confirmed cases over time.

```
all_us %>%
  group_by(Month_Year) %>%
  summarise(Ratio = sum(New_Deaths) * 1000 / sum(New_Confirmed)) %>%
  ggplot(aes(x = Month_Year, y = Ratio)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Ratio of COVID-19 Deaths to Cases Over Time",
       x = "Month",
       y = "Deaths per 1000 Cases")
```



Finally, let's fit a linear regression model to the US data to see if there is any evidence of improvements in the ratio of deaths to confirmed cases over time. We'll try to predict the ratio of deaths to cases based on the number of months since the start of the pandemic and the number of new confirmed cases.

```
model <- lm(Deaths_to_Cases_Ratio ~ Months_Since_Start + New_Confirmed, data = all_us)
summary(model)
```

```
##
## Call:
## lm(formula = Deaths_to_Cases_Ratio ~ Months_Since_Start + New_Confirmed,
##     data = all_us)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.023  -9.137  -4.264   3.236 289.542
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.692e+01  9.289e-01  28.982  < 2e-16 ***
## Months_Since_Start -5.012e-01  4.034e-02 -12.423  < 2e-16 ***
## New_Confirmed      -1.730e-05  3.496e-06  -4.948 8.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.82 on 1912 degrees of freedom
## Multiple R-squared:  0.08718,    Adjusted R-squared:  0.08622
```

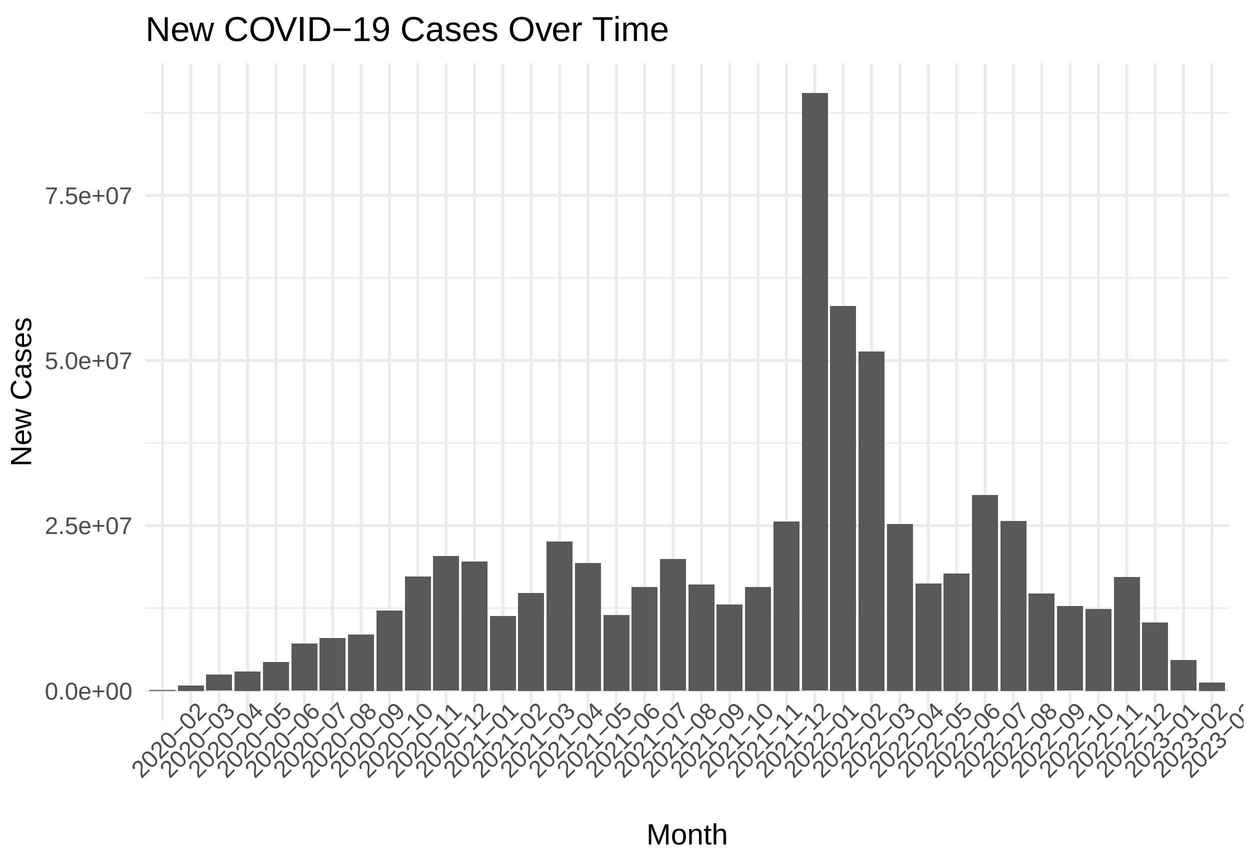
```
## F-statistic:  91.3 on 2 and 1912 DF,  p-value: < 2.2e-16
```

Global Analysis

Next, we'll repeat the same analysis for the global data to see if there are any trends in the ratio of deaths to confirmed cases over time.

Again, we'll start by plotting the total number of new cases over time.

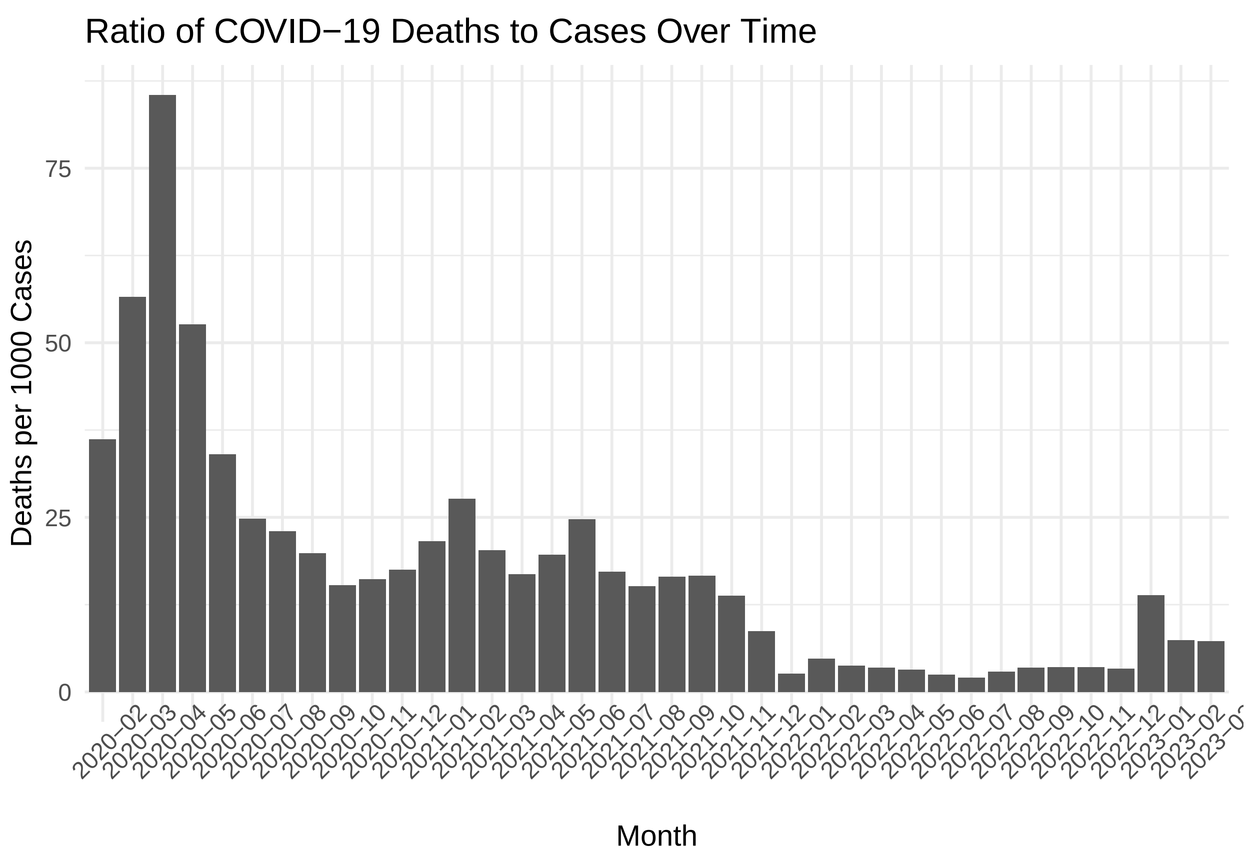
```
all_global %>%
  group_by(Month_Year) %>%
  summarise(Mean_New_Confirmed = mean(New_Confirmed),
            Sum_New_Confirmed = sum(New_Confirmed)) %>%
  ggplot(aes(x = Month_Year, y = Sum_New_Confirmed)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "New COVID-19 Cases Over Time",
       x = "Month",
       y = "New Cases")
```



Next, let's plot the ratio of deaths to confirmed cases over time.

```
all_global %>%
  group_by(Month_Year) %>%
  summarise(Ratio = sum(New_Deaths) * 1000 / sum(New_Confirmed)) %>%
  ggplot(aes(x = Month_Year, y = Ratio)) +
```

```r
  geom_col() +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Ratio of COVID-19 Deaths to Cases Over Time",
       x = "Month",
       y = "Deaths per 1000 Cases")
```



Finally, let's fit a linear regression model to the global data to see if there is any evidence of improvements in the ratio of deaths to confirmed cases over time. As we did with the US data, we'll try to predict the ratio of deaths to cases based on the number of months since the start of the pandemic and the number of new confirmed cases.

```r
model <- lm(Deaths_to_Cases_Ratio ~ Months_Since_Start + New_Confirmed, data = all_global)
summary(model)
```

```
##
## Call:
## lm(formula = Deaths_to_Cases_Ratio ~ Months_Since_Start + New_Confirmed,
##     data = all_global)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -67.68  -13.87   -8.28    2.10 2172.38
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          3.265e+01  1.473e+00  22.171   <2e-16 ***
## Months_Since_Start -6.416e-01  6.615e-02  -9.699   <2e-16 ***
## New_Confirmed      -3.057e-06  1.225e-06  -2.495   0.0126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.76 on 5811 degrees of freedom
## Multiple R-squared:  0.01745,    Adjusted R-squared:  0.01711
## F-statistic: 51.59 on 2 and 5811 DF,  p-value: < 2.2e-16
```

Bias and Limitations

Before discussing the results, it's important to acknowledge the potential biases and limitations of this analysis.

One limitation is that the data is prone to reporitng bias and it may not be complete or accurate. For instance, early on in the pandemic, testing capacity was limited, which may have led to underreporting of cases, which would increase the ratio of deaths to confirmed cases at that time.

Another limitation is that the data may not be consistent across states and countries; Reporting practices may vary, which could affect the accuracy of the data and bias our results.

Finally, the analysis assumes that the ratio of deaths to confirmed cases is a good indicator of treatment effectiveness, but there are many other factors that could influence this ratio, include the presence of new variants and the overall health of the population.

Conclusion

After analyzing the US and global COVID-19 data, we uncovered a few interesting insights.

In the US and in the global data, the number of new COVID-19 cases fluctuated over time. These fluctuations could be caused by outbreaks in different regions and of different variants. The total number of new cases could also be affected by the availability of testing.

We also found that the ratio of deaths to confirmed cases decreased over time in both the US and global data. We found a negative relationship between the ratio of deaths to confirmed cases and the number of months since the start of the pandemic in the US data (estimate = -0.50, $p < 0.001$) as well as in the global data (estimate = -0.64, $p < 0.001$). This trend suggests that treatments may have improved over time, leading to a lower mortality rate among confirmed cases.

However, as mentioned in the previous section, it is important to note that the ratio of deaths to confirmed cases is not the only indicator of treatment effectiveness. There are likely many other factors that affect the ratio of deaths to confirmed cases.

Overall, this analysis provides some evidence that treatments for COVID-19 may have improved over time. Future analyses could explore other factors that may influence the ratio of deaths to confirmed cases, such as the availability of testing and the presence of new variants.

Session Info

```r
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
```

```
## Platform: aarch64-apple-darwin23.4.0
## Running under: macOS Sonoma 14.6
##
## Matrix products: default
## BLAS:   /opt/homebrew/Cellar/openblas/0.3.27/lib/libopenblasp-r0.3.27.dylib
## LAPACK: /opt/homebrew/Cellar/r/4.4.0_1/lib/R/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.5      highr_0.10        crayon_1.5.2
##  [5] compiler_4.4.0    tidyselect_1.2.1  parallel_4.4.0    scales_1.3.0
##  [9] yaml_2.3.8        fastmap_1.1.1     R6_2.5.1          labeling_0.4.3
## [13] generics_0.1.3    curl_5.2.1        knitr_1.46        munsell_0.5.1
## [17] pillar_1.9.0      tzdb_0.4.0        rlang_1.1.3       utf8_1.2.4
## [21] stringi_1.8.4     xfun_0.43         bit64_4.0.5       timechange_0.3.0
## [25] cli_3.6.2         withr_3.0.0       magrittr_2.0.3    digest_0.6.35
## [29] grid_4.4.0        vroom_1.6.5       rstudioapi_0.16.0 hms_1.1.3
## [33] lifecycle_1.0.4   vctrs_0.6.5       evaluate_0.23     glue_1.7.0
## [37] farver_2.1.2      fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.26
## [41] tools_4.4.0       pkgconfig_2.0.3   htmltools_0.5.8.1
```