COVID19 Analysis

Jordan Pelletier

2025-04-11

Question of Interest

The primary question explored in this analysis is: **How has COVID-19 spread globally over time, and what patterns can be observed in confirmed cases and death rates across different countries?**

Data Source and Description

The dataset used for this analysis is sourced from the Johns Hopkins University COVID-19 Data Repository. It contains time series data on COVID-19 confirmed cases and deaths globally, starting from the early days of the pandemic.

Get Current data

```
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_confirmed_us.csv",
"time_series_covid19_deaths_us.csv")

confirmed_global <- read.csv(paste0(url_in, file_names[1]))
confirmed_global <- read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covi

deaths_global <- read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_
```

Import Libraries and Read Files

```
library(tidyverse)
```

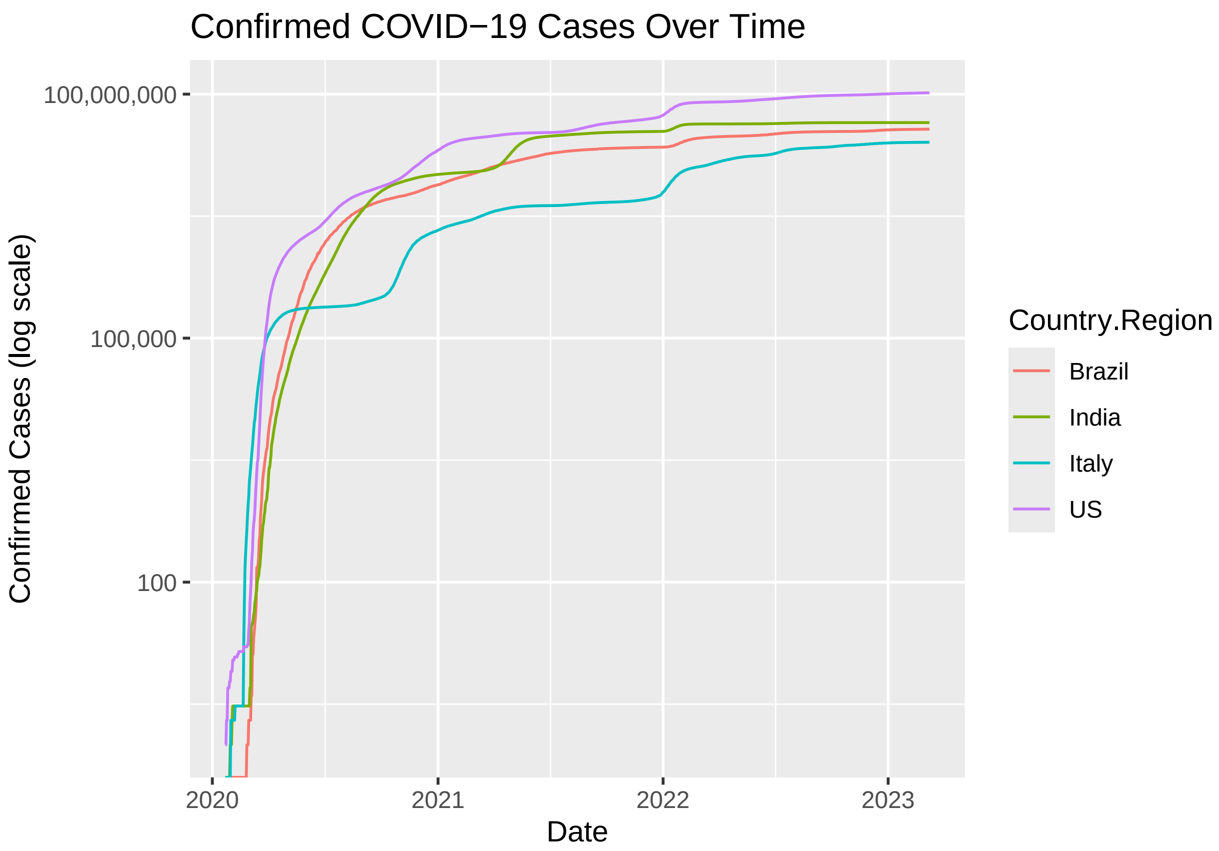
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ 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(lubridate)

# Tidy confirmed
confirmed_long <- confirmed_global %>%
  pivot_longer(cols = starts_with("X") | matches("^\\d"),
               names_to = "date", values_to = "confirmed") %>%
  mutate(date = mdy(str_remove(date, "^X"))) %>%
  group_by(`Country.Region`, date) %>%
  summarise(confirmed = sum(confirmed, na.rm = TRUE), .groups = 'drop')

# Tidy deaths
deaths_long <- deaths_global %>%
  pivot_longer(cols = starts_with("X") | matches("^\\d"),
               names_to = "date", values_to = "deaths") %>%
  mutate(date = mdy(str_remove(date, "^X"))) %>%
  group_by(`Country.Region`, date) %>%
  summarise(deaths = sum(deaths, na.rm = TRUE), .groups = 'drop')
```

Visualization 1: Confirmed Cases Over Time

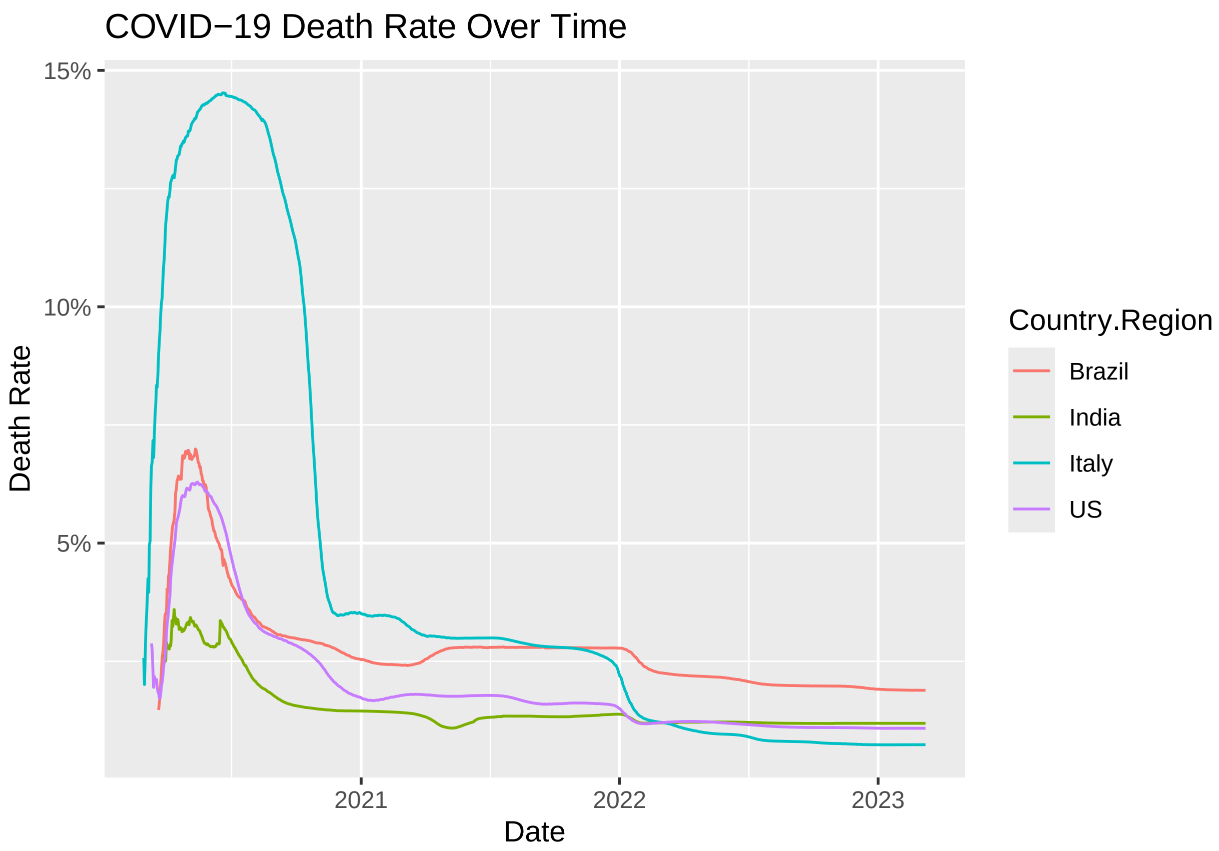
```r
countries <- c("US", "India", "Italy", "Brazil")

confirmed_long %>%
  filter(`Country.Region` %in% countries) %>%
  ggplot(aes(x = date, y = confirmed, color = `Country.Region`)) +
  geom_line() +
  scale_y_log10(labels = scales::comma) +
  labs(title = "Confirmed COVID-19 Cases Over Time",
       x = "Date", y = "Confirmed Cases (log scale)")
```

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



Visualization 2: Death Rate Over Time

```
covid_joined <- left_join(confirmed_long, deaths_long,
                          by = c("Country.Region", "date")) %>%
  filter(confirmed > 1000) %>%
  mutate(death_rate = deaths / confirmed)

covid_joined %>%
  filter(`Country.Region` %in% countries) %>%
  ggplot(aes(x = date, y = death_rate, color = `Country.Region`)) +
  geom_line() +
  labs(title = "COVID-19 Death Rate Over Time",
       y = "Death Rate", x = "Date") +
  scale_y_continuous(labels = scales::percent)
```



Model: Predicting US Cases

A linear regression model was created to predict the number of confirmed COVID-19 cases in the United States over time. The model used "days since the first confirmed case" as the independent variable and the number of confirmed cases as the dependent variable.

```
us_confirmed <- confirmed_long %>%
  filter(`Country.Region` == "US") %>%
  mutate(days_since_start = as.numeric(date - min(date)))

model <- lm(confirmed ~ days_since_start, data = us_confirmed)
summary(model)
```

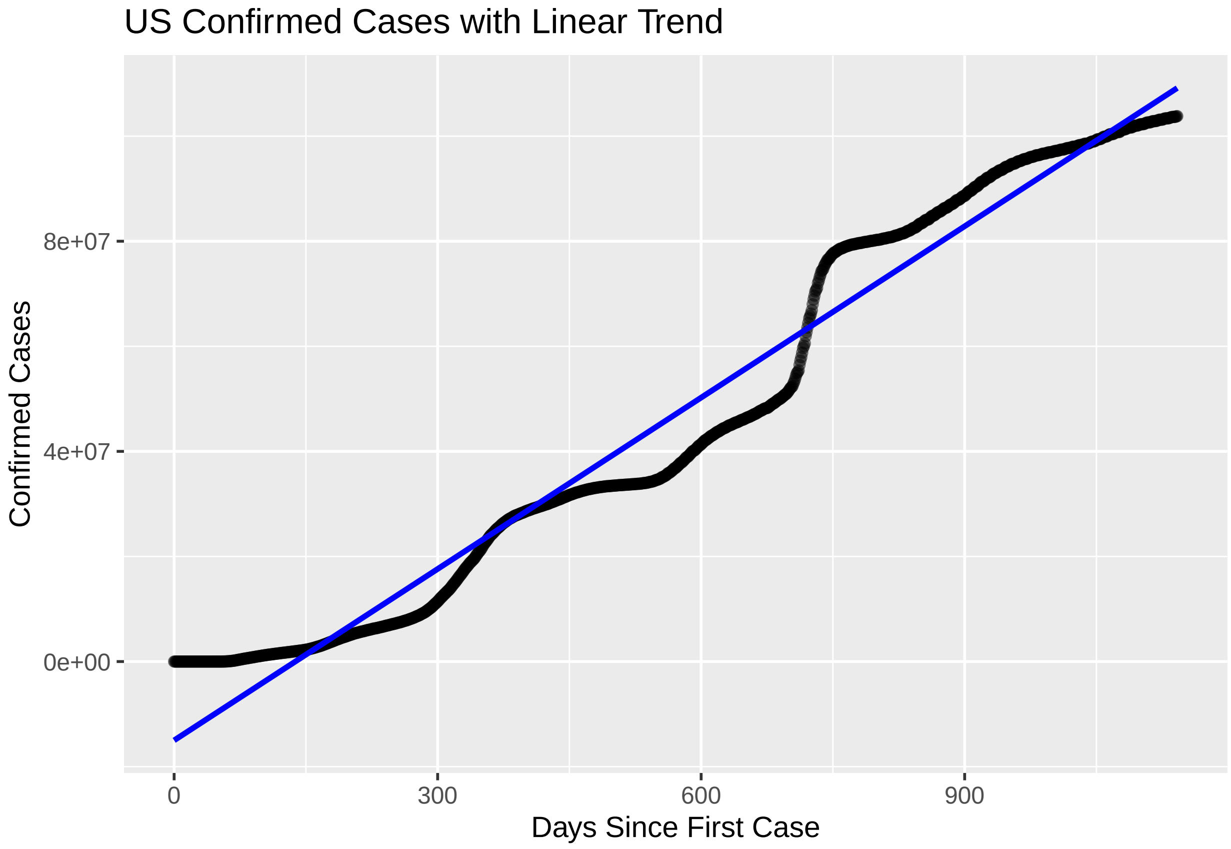
```
##
## Call:
## lm(formula = confirmed ~ days_since_start, data = us_confirmed)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10423950  -5199705   -477823   5886501  14998756
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.500e+07  3.852e+05  -38.94   <2e-16 ***
## days_since_start  1.087e+05  5.841e+02  186.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6516000 on 1141 degrees of freedom
## Multiple R-squared:  0.9681, Adjusted R-squared:  0.9681
## F-statistic: 3.464e+04 on 1 and 1141 DF,  p-value: < 2.2e-16
```

Trend Plot

```
us_confirmed %>%
  ggplot(aes(x = days_since_start, y = confirmed)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "US Confirmed Cases with Linear Trend",
       x = "Days Since First Case", y = "Confirmed Cases")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Custom Analysis

In addition to the basic trends, I also examined the **death rate** over time by calculating the ratio of deaths to confirmed cases. This provided insight into the severity of the pandemic in different countries, which can be influenced by factors such as healthcare infrastructure and government intervention.

Further, I performed a linear regression analysis to model the growth of cases in the US, which helps forecast future case trends based on historical data.

Conclusion

This analysis provides insight into the spread of COVID-19 across different countries. We observed significant growth in confirmed cases in all four countries examined, though the death rate varied significantly, likely due to healthcare differences and reporting practices. The linear model also helped forecast trends in US case growth.

Bias and Limitations

Several important limitations and sources of bias exist in the COVID-19 dataset and this analysis:

As discussed, the dataset may suffer from underreporting, inconsistent data collection practices, and reporting delays. These factors introduce potential biases that should be taken into account when interpreting the results.

- **Underreporting**: Many countries underreport due to limited testing, political pressure, or data collection issues.
- **Time lag**: Death data often lags behind confirmed case data, complicating real-time analysis.
- **Inconsistent data collection**: Different countries may have varying definitions for what constitutes a confirmed case or death.
- **Population size and density**: Raw case numbers alone do not account for differences in population sizes or density, which can affect the spread.
- **Policy and healthcare differences**: The effectiveness of public health policies, access to healthcare, and vaccine availability can strongly influence outcomes.

This analysis is meant to explore patterns, but not to draw definitive conclusions about country performance or outcomes.