DataScienceasaFieldFinalProjectpart2

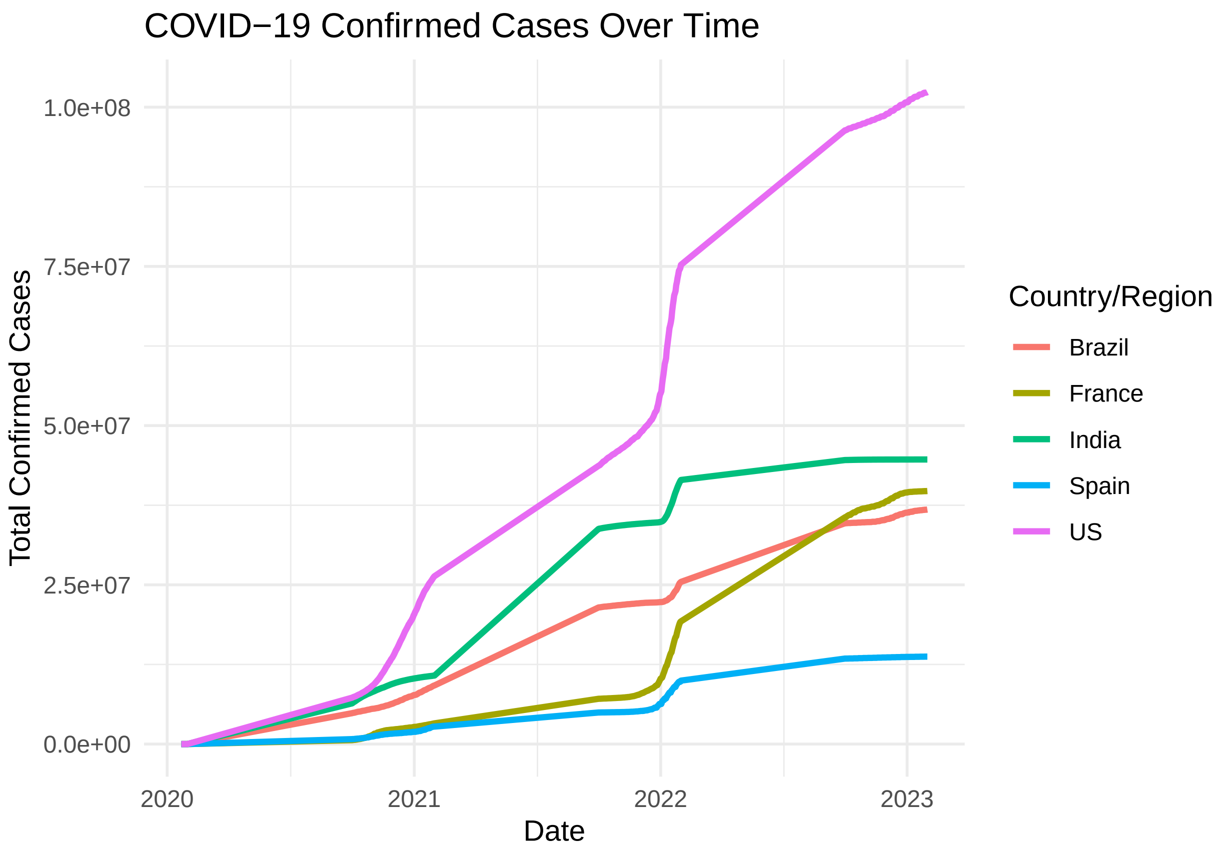
Sashaank

2025-03-24

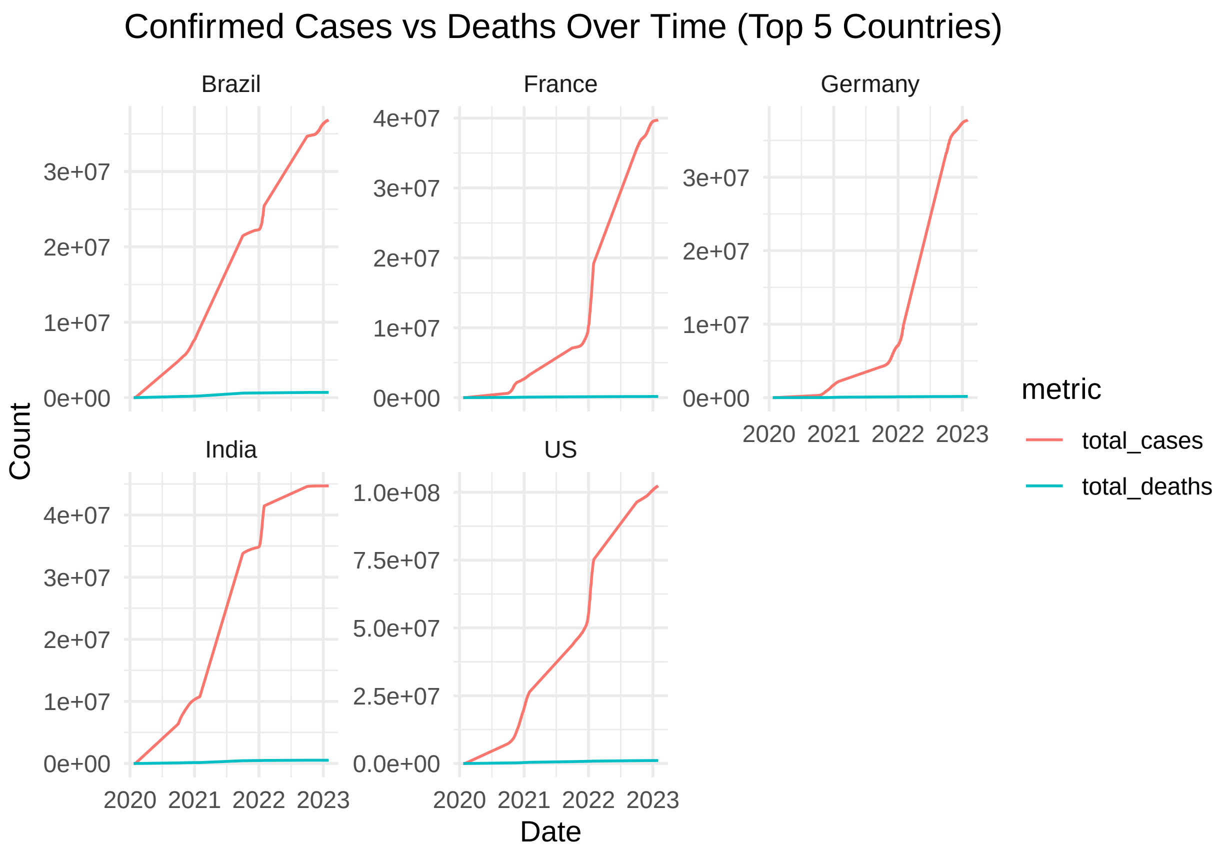
```
# Load datasets from GitHub
global_cases_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data
global_deaths_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_dat
us_cases_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/css
us_deaths_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/cs

# Read CSVs
global_cases_raw <- read_csv(global_cases_url, show_col_types = FALSE)
global_deaths_raw <- read_csv(global_deaths_url, show_col_types = FALSE)
us_cases_raw <- read_csv(us_cases_url, show_col_types = FALSE)
us_deaths_raw <- read_csv(us_deaths_url, show_col_types = FALSE)

# Clean global data
global_cases_clean <- global_cases_raw %>%
  pivot_longer(cols = starts_with("1"), names_to = "date", values_to = "total_cases") %>%
  mutate(date = mdy(date)) %>%
  group_by(`Country/Region`, date) %>%
  summarize(total_cases = sum(total_cases), .groups = "drop")

global_deaths_clean <- global_deaths_raw %>%
  pivot_longer(cols = starts_with("1"), names_to = "date", values_to = "total_deaths") %>%
  mutate(date = mdy(date)) %>%
  group_by(`Country/Region`, date) %>%
  summarize(total_deaths = sum(total_deaths), .groups = "drop")

# Merge global cases and deaths
global_summary <- left_join(global_cases_clean, global_deaths_clean,
                            by = c("Country/Region", "date"))
```

```
# Plot for selected countries
selected_nations <- c("US", "India", "Brazil", "France", "Spain")

global_cases_clean %>%
  filter(`Country/Region` %in% selected_nations) %>%
  ggplot(aes(x = date, y = total_cases, color = `Country/Region`)) +
  geom_line(size = 1.1) +
  labs(title = "COVID-19 Confirmed Cases Over Time",
       y = "Total Confirmed Cases", x = "Date") +
  theme_minimal()
```

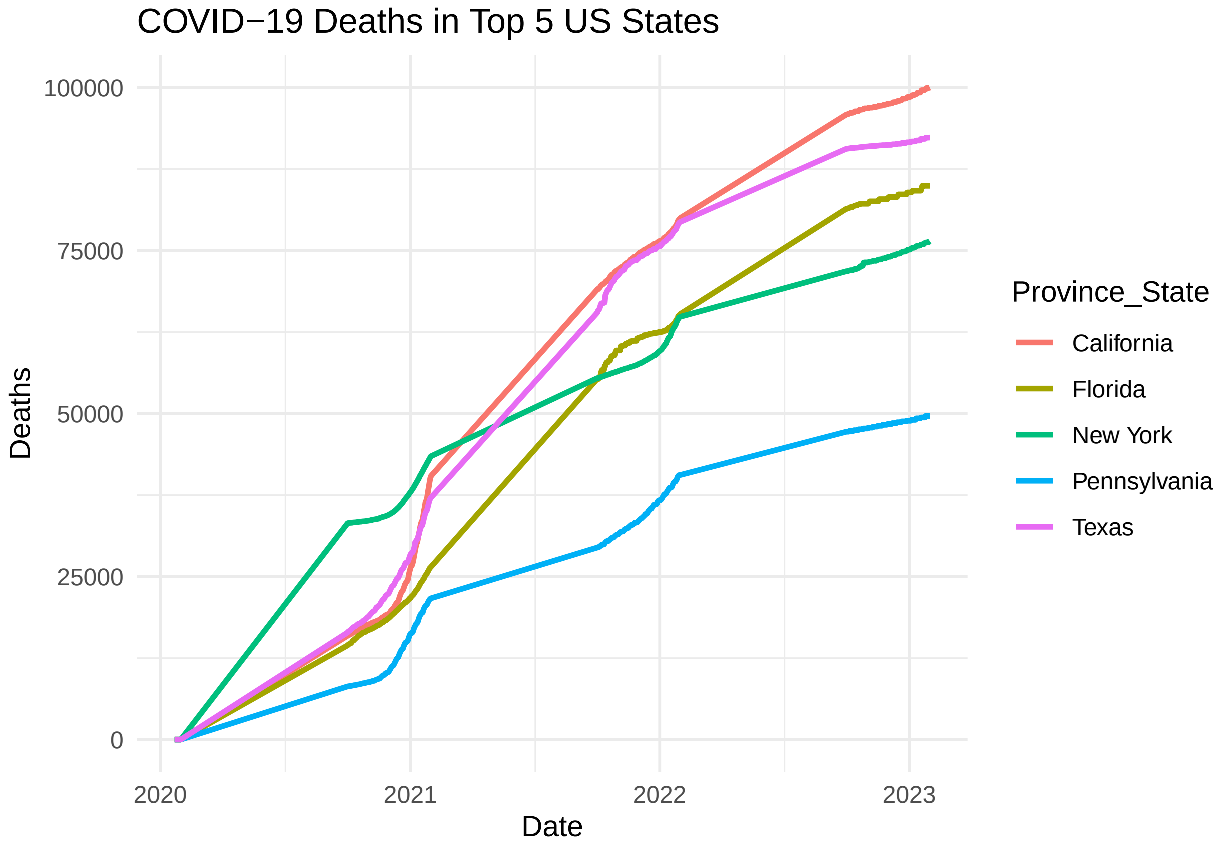
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



```r
# top 5 countries by confirmed cases
top_countries <- global_summary %>%
  filter(date == max(date)) %>%
  arrange(desc(total_cases)) %>%
  slice_head(n = 5) %>%
  pull(`Country/Region`)

# Plot confirmed vs deaths over time
global_summary %>%
  filter(`Country/Region` %in% top_countries) %>%
  pivot_longer(cols = c(total_cases, total_deaths), names_to = "metric", values_to = "value") %>%
  ggplot(aes(x = date, y = value, color = metric)) +
  geom_line() +
  facet_wrap(~ `Country/Region`, scales = "free_y") +
  labs(title = "Confirmed Cases vs Deaths Over Time (Top 5 Countries)",
       x = "Date", y = "Count") +
  theme_minimal()
```



```
# Clean US data
us_cases_clean <- us_cases_raw %>%
  pivot_longer(cols = starts_with("1"), names_to = "date", values_to = "confirmed_cases") %>%
  mutate(date = mdy(date)) %>%
  group_by(Province_State, date) %>%
  summarize(confirmed_cases = sum(confirmed_cases), .groups = "drop")

us_deaths_clean <- us_deaths_raw %>%
  pivot_longer(cols = starts_with("1"), names_to = "date", values_to = "death_count") %>%
  mutate(date = mdy(date)) %>%
  group_by(Province_State, date) %>%
  summarize(death_count = sum(death_count), .groups = "drop")

# Merge US cases and deaths
us_summary <- left_join(us_cases_clean, us_deaths_clean,
                        by = c("Province_State", "date"))

# Get top 5 states by deaths
top_states <- us_summary %>%
  filter(date == max(date)) %>%
  arrange(desc(death_count)) %>%
  slice_head(n = 5) %>%
  pull(Province_State)

# Plot deaths over time for top states
us_summary %>%
  filter(Province_State %in% top_states) %>%
  ggplot(aes(x = date, y = death_count, color = Province_State)) +
  geom_line(size = 1) +
  labs(title = "COVID-19 Deaths in Top 5 US States",
       y = "Deaths", x = "Date") +
  theme_minimal()
```



The analysis in the report relies entirely on publicly available COVID-19 data from the Johns Hopkins CSSE repository. Some biases may include underreporting and inconsistent testing, reporting delays, and other missing demographics. It lacks variables such as age, race, income, vaccination status etc which could also provide more information on the analysis.

In conclusion, we processed and analyzed global and U.S. COVID-19 time series data to examine how the pandemic evolved across different countries and states. Our key observations include The United States, India, and Brazil consistently report some of the highest totals for confirmed cases and deaths worldwide. Within the U.S., states such as California, Texas, and New York have recorded the highest cumulative death counts.Comparing confirmed cases and deaths over time highlights notable differences in case fatality rates and the overall progression of the outbreak across countries