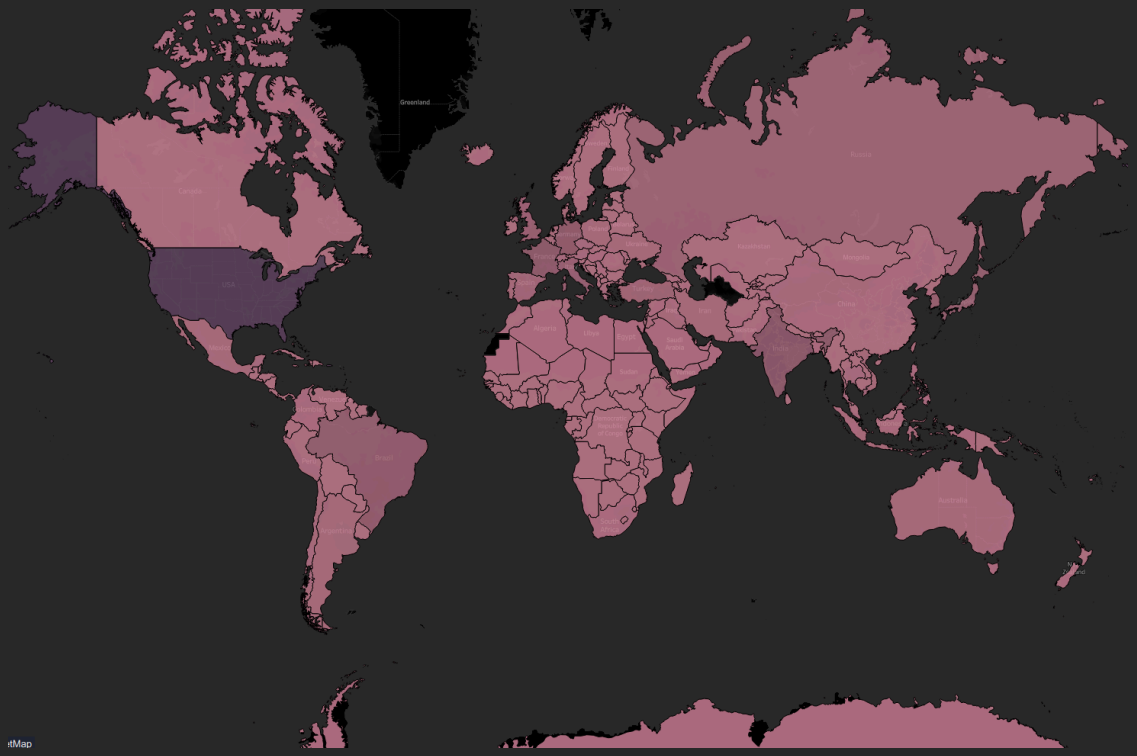
Public Health Data Trend Analysis (COVID-19)

Uncovering Pandemic Patterns for Informed Public Health Strategy



Marcelo Somma, July 27th 2025

https://github.com/MarceloSomma

Objective: Analyze global COVID-19 trends to understand the spread, impact, and recovery patterns across different countries and over time. This will involve using SQL for data preparation and aggregation, and Tableau for interactive visualization and dashboard creation.

Tools used for this project: SQL (PostgreSQL), PgAdmin, Tableau Public

Executive Summary

This report presents a comprehensive data analysis of global COVID-19 trends, leveraging publicly available Johns Hopkins University (JHU) data. The project utilized Python (Pandas) for initial data pivoting, PostgreSQL for robust data cleaning and aggregation, and Tableau Public for interactive visualization, culminating in a dynamic dashboard.

Key Findings:

- **Significant Global Impact:** The pandemic resulted in over **676 Million** confirmed cases and **6.8 Million** deaths worldwide by early 2023, with cumulative figures showing a steady, non-decreasing trend over the three-year period.
- **Distinct Waves of Infection:** Analysis of monthly new confirmed cases and deaths reveals clear waves, with notable spikes in **January 2022** (coinciding with the Omicron variant) and **January 2021** for deaths (likely linked to initial poor preparation and rapid spread).
- **Geographical Disparities:** The **United States** recorded the highest cumulative confirmed cases globally. Furthermore, countries like **Peru** experienced a particularly high mortality rate of **5.8%**, highlighting localized severe impacts.
- **Comparative Trends:** Comparative analysis of selected countries demonstrates varying infection curves and mortality patterns, suggesting diverse regional responses and outcomes.

Key Recommendations:

- **Proactive Pandemic Preparedness:** Based on observed spikes from new variants, continuous investment in and refinement of rapid response protocols for emerging health threats is crucial.
- **Targeted Resource Allocation:** Given high case counts in specific regions like the US, and elevated mortality rates in areas such as Peru, resources for testing, treatment, and healthcare infrastructure should be strategically allocated to address disproportionately affected populations.
- **Refine Crisis Management Protocols:** Insights into death spikes highlight the need for robust epidemic management protocols and efficient patient flow management in medical facilities to mitigate future health crises.

This analysis provides valuable insights for understanding past pandemic dynamics and can inform strategic planning for future public health challenges.

Phase 1: Project Setup & Data Gathering

Questions/Goals:

- What are the global trends of confirmed cases and deaths over time?
- Which countries were most affected in terms of total cases and deaths?
- How did the daily new cases and deaths evolve in key regions?
- What was the mortality rate in different countries/regions?
- Analyze recovery trends.

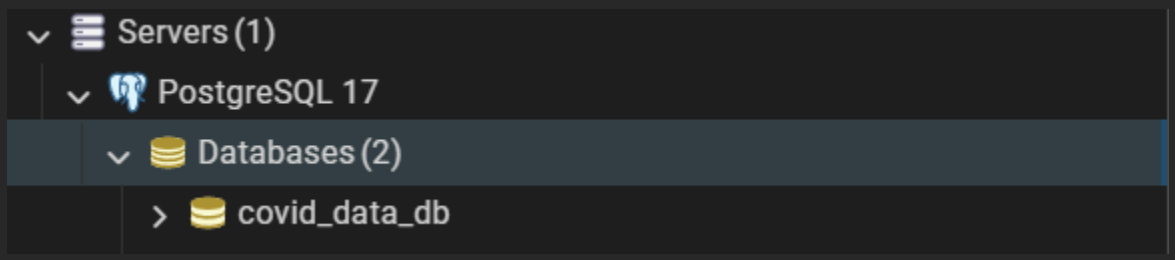
Data Gathering

The Johns Hopkins University CSSE COVID-19 data on Github

- **Confirmed Cases:** https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
- **Deaths:** https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

Setting up SQL Environment (PostgreSQL & PgAdmin):

PostgreSQL Server and Database (covid_data_db) in PgAdmin:



Phase 2: Data Preparation & Pre Processing

The Challenge with the JHU Data:

The Johns Hopkins COVID-19 CSVs (time_series_covid19_confirmed_global.csv, etc.) are in a "wide" format.

```
"Province/State","Country/Region","Lat","Long","1/22/20","1/23/20","1/24/20",...
,"Afghanistan",33.93911,67.709953,0,0,0,...
,"Albania",41.1533,20.1683,0,0,0,...
```

Solution: Pivoting to "Long" Format with Python (Pandas)

Result:

```
df_combined_long.head(10)
```

	Province/State	Country/Region	Lat	Long	Date	ConfirmedCases	Deaths
0	Alberta	Canada	53.9333	-116.5765	2020-01-22	0	0
1	Alberta	Canada	53.9333	-116.5765	2020-01-23	0	0
2	Alberta	Canada	53.9333	-116.5765	2020-01-24	0	0
3	Alberta	Canada	53.9333	-116.5765	2020-01-25	0	0
4	Alberta	Canada	53.9333	-116.5765	2020-01-26	0	0
5	Alberta	Canada	53.9333	-116.5765	2020-01-27	0	0
6	Alberta	Canada	53.9333	-116.5765	2020-01-28	0	0
7	Alberta	Canada	53.9333	-116.5765	2020-01-29	0	0
8	Alberta	Canada	53.9333	-116.5765	2020-01-30	0	0
9	Alberta	Canada	53.9333	-116.5765	2020-01-31	0	0

Full breakdown of the unpivoting process in “data_pivot.ipynb” file

Creating Table in PostgreSQL (PgAdmin4)

```
CREATE TABLE covid_data (
province_state VARCHAR(255),
country_region VARCHAR(255) NOT NULL,
latitude NUMERIC,
longitude NUMERIC,
report_date DATE NOT NULL,
confirmed_cases BIGINT NOT NULL,
deaths BIGINT NOT NULL
);
-- Adding indexes for faster query performance later
CREATE INDEX idx_country_region ON covid_data (country_region);
CREATE INDEX idx_report_date ON covid_data (report_date); CREATE
INDEX idx_country_date ON covid_data (country_region, report_date);
```

Importing data from long format .csv file

First check of the data

```
SELECT
      *
FROM
      covid_data
LIMIT
      10
```

	province_state character varying (255)	country_region character varying (255)	latitude numeric	longitude numeric	report_date date	confirmed_cases bigint	deaths bigint
1	Alberta	Canada	53.9333	-116.5765	2020-01-22	0	0
2	Alberta	Canada	53.9333	-116.5765	2020-01-23	0	0
3	Alberta	Canada	53.9333	-116.5765	2020-01-24	0	0
4	Alberta	Canada	53.9333	-116.5765	2020-01-25	0	0
5	Alberta	Canada	53.9333	-116.5765	2020-01-26	0	0
6	Alberta	Canada	53.9333	-116.5765	2020-01-27	0	0
7	Alberta	Canada	53.9333	-116.5765	2020-01-28	0	0
8	Alberta	Canada	53.9333	-116.5765	2020-01-29	0	0
9	Alberta	Canada	53.9333	-116.5765	2020-01-30	0	0
10	Alberta	Canada	53.9333	-116.5765	2020-01-31	0	0

Phase 3: Data Cleaning

Looking for missing values in key columns:

```
SELECT
      *
FROM
      covid_data
WHERE
      Country_region IS NULL OR report_date IS NULL OR
confirmed_cases IS NULL OR deaths IS NULL
```

Checking for duplicate values:

```
SELECT
      province_state,
      country_region,
      report_date,
      confirmed_cases,
      deaths,
      COUNT(*)
FROM
      covid_data
GROUP BY
      province_state,
      country_region,
      report_date,
      confirmed_cases,
      deaths,
HAVING
      COUNT(*) > 1
```

Checking for inadequate ‘country_region’ values

```
SELECT
      DISTINCT country_region
FROM
      covid_data
WHERE
      country_region LIKE '%Olympics%'
```

Deletion of rows with said values

```
DELETE
FROM
      covid_data
WHERE
      country_region LIKE '%Olympics%'
```

Checking for anomalies in minimum & maximum values

```
SELECT
      MAX(confirmed_cases),
      MIN(confirmed_cases),
      MAX(deaths),
      MIN(deaths),
FROM
      covid_data
```

Creating new table for daily new cases and deaths:

```
CREATE TABLE daily_covid_metrics AS
SELECT
    province_state,
    country_region,
    report_date,
    confirmed_cases,
    deaths,
    -- Calculate daily new confirmed cases
    confirmed_cases - LAG(confirmed_cases, 1, 0) OVER (PARTITION BY
province_state, country_region ORDER BY report_date) AS
daily_new_cases,
    -- Calculate daily new deaths
    deaths - LAG(deaths, 1, 0) OVER (PARTITION BY province_state,
country_region ORDER BY report_date) AS daily_new_deaths
FROM
    covid_data
```

This was created as a physical table to allow for direct deletion of anomalous negative daily counts, ensuring data integrity for subsequent analysis.

Deleting rows with negative daily counts

```
DELETE FROM
     daily_covid_metrics
WHERE
     daily_new_cases <= 0 OR
     daily_new_deaths <= 0
```

After this process both tables were exported into .csv format for connecting with the Tableau Public application

Phase 4: Exploratory Data Analysis & Visualization

Created a dashboard with a 2 part structure, a Global & Country Specific Trends Over Time section and a Countries Comparative Analysis

Global & Country Specific Trends Over Time charts:

- **Monthly New Confirmed Cases (Line Chart)**
  - **Observation:** A significant increase in newly confirmed cases was observed in January 2022, coinciding with the emergence of the Omicron variant.
  - **Recommendation:** Based on the observed impact, further research is suggested into alternative or more gradual lockdown implementation strategies to minimize public disruption while maintaining efficacy
- **Monthly Deaths (Line Chart)**
  - **Observation:** A significant increase in fatalities was observed in January 2021, likely attributable to inadequate preparedness and the rapid dissemination of the virus.
  - **Recommendation:** Investigate the causes of fatalities to facilitate a more effective allocation of resources to affected areas.
- **SUM of Confirmed Cases (Line Chart)**
  - **Observation:** Confirmed cases demonstrated a steady increase during the initial two years, culminating in January 2022 with the emergence of the Omicron variant, which precipitated a significant surge in new cases. Subsequent to April 2022, the rate of increase in new cases decelerated to its previous trajectory.
  - **Recommendation:** Develop a smoother transition process into a lockdown state in order for people to gather necessary supplies to stay at home for an extended period of time
- **SUM of Deaths (Line Chart)**
  - **Observation:** Fatalities demonstrated a consistent escalation until April 2022, at which point the rate of increase began to decelerate.
  - **Recommendation:** Create protocols for dealing with increased and sudden flow of patients in medical facilities

- **Confirmed Cases (Map)**
  - **Observation:** The US was the country with the most confirmed cases
  - **Recommendation:** Given the high case count in the US, continued vigilance and resource allocation for testing and treatment remain critical.
- **Mortality Rate % (Map)**
  - **Observation:** Peru had the worst Mortality Rate %
  - **Recommendation:** Given Peru is the country with the highest mortality rate, resource allocation to the health department should be considered

Country Comparative Analysis charts:

- Monthly New Cases & Deaths in Selected Countries (Line Chart)
- Total Deaths in Selected Countries (Bar Chart)
- Total Confirmed Cases in Selected Countries (Bar Chart)

Discuss differences and similarities in trends, total cases, and deaths among the countries compared.

Dashboard snapshot: