Visualizing Covid-19 Pandemic Data

Saptarshi Mitra

BTech ECE 2nd Year (2024-28),

Institute of Engineering & Management (IEM) , Kolkata

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation,

ISI Kolkata

1. Abstract

This project presents a comprehensive data analysis and visualization of the global COVID-19 pandemic using Python. Publicly available case and death records were processed, cleaned, and filtered to focus on the period from March 2020 to August 2023. Multiple analytical perspectives were explored, including country-level, regional, and global trends. Line plots were generated to compare daily new cases across the top five most affected countries and the five least affected countries. Global daily cases were visualized with a mountain-shaped area plot to capture the intensity of outbreaks. Quarterly analyses were conducted to highlight the relationship between new cases and deaths, presented both as stacked and side-by-side bar charts. Regional disparities were explored through heatmaps of new cases and deaths across time, as well as a focused heatmap on the top ten countries by monthly cases. A pie chart highlighted the top ten countries with the highest cumulative deaths. Interactive dashboards using Plotly provided global line charts, stacked bar charts, and choropleth maps for deeper exploration. Together, these visualizations offer a multidimensional understanding of how COVID-19 spread across countries and regions over time, revealing patterns of intensity, peaks, and geographic concentration.

2. Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, is one of the most significant global health crises in modern history. Since its emergence in late 2019, it has spread across almost every country, leading to millions of confirmed cases and deaths worldwide. Understanding the spread, intensity, and impact of the pandemic is crucial not only for health professionals and policymakers but also for researchers, students, and the general public. Data-driven analysis and visualization play an essential role in interpreting the scale and dynamics of such outbreaks.

This project focuses on analyzing COVID-19 data from March 2020 to August 2023 using computational tools and visualization techniques. The dataset was obtained from publicly available WHO reports and processed using **Python**, leveraging libraries such as **Pandas** and **NumPy** for data handling, **Matplotlib** and **Seaborn** for static visualizations, and **Plotly** for interactive dashboards. These technologies allow for both high-level global overviews and country-specific insights, making the analysis flexible and engaging.

Before implementing the project, a background survey of visualization practices in epidemiology was conducted. Existing dashboards, such as the Johns Hopkins COVID-19 tracker, demonstrated the importance of clear visual representation in aiding decision-making. Inspired by these works, this project aimed to combine both traditional statistical plots and advanced interactive graphics to provide a holistic view of the pandemic.

The procedure involved cleaning and filtering raw data, aggregating values across countries, regions, and time periods, and designing multiple types of visualizations. Line charts, bar plots, pie charts, heatmaps, and choropleth maps were used to capture different perspectives of the pandemic, including daily fluctuations, quarterly summaries, regional disparities, and country-wise comparisons. Additionally, an interactive dashboard was developed with Plotly and Dash to allow users to dynamically explore cases, deaths, and cumulative trends across regions and countries.

The primary purpose of this project was to transform raw COVID-19 data into meaningful insights through visualization. By identifying trends, peaks, and regional variations, the analysis aims to highlight not only the severity of the crisis but also the importance of data-driven approaches in public health monitoring. Furthermore, the interactive elements provide an engaging learning tool, encouraging deeper exploration and understanding of the global impact of COVID-19.

On the first 2 weeks of Internship at IDEAS-TIH, ISI Kolkata we received training on-

1.Python Basics

Data Variables, Lists, Loops

Data Structures

Class,Functions, OOPS

Numpy, Pandas

2.Machine Learning Overview

3.Regression

4.Classification

5.LLM Fundamentals

6.Communication Skills

3. Project Objective

The objective of this project is to transform large-scale COVID-19 data into meaningful visual insights that highlight the dynamics and impact of the pandemic. The main goals are:

- **To analyze and visualize the temporal spread of COVID-19** by examining daily, monthly, and quarterly trends in new cases and deaths across different countries and regions.

- **To identify the most and least affected countries** through cumulative case and death counts, using pie charts, line plots, and bar charts to illustrate disparities.

- **To highlight regional and temporal intensity patterns** using heatmaps, thereby showing how outbreaks surged differently across time and geographies.

- **To develop an interactive dashboard** with Plotly and Dash that allows dynamic filtering, comparison, and exploration of cases, deaths, and cumulative trends.

- **To illustrate the importance of data-driven approaches in pandemic monitoring**, emphasizing how visualization can simplify complex datasets for better understanding and decision-making.

4. Methodology

The methodology adopted in this project followed a systematic process of **data acquisition, cleaning, preprocessing, and visualization**, supported by modern data analysis and visualization tools. The major steps are described below:

1. Data Collection

- The dataset was collected from publicly available **Google Drive repositories and WHO COVID-19 datasets**.

- The data included daily country-wise reports of **new cases, new deaths, cumulative cases, and cumulative deaths**, along with metadata such as country, WHO region, and reporting dates.

- Since the dataset was secondary in nature, no field survey was conducted.

2. Data Preprocessing

- The `pandas` library was used for loading the dataset into a structured dataframe.

- **Date parsing:** The `Date_reported` column was converted into `datetime` format for easy filtering and time-series analysis.

- **Filtering:** The dataset was trimmed to the period between **March 2020 and August 2023** to focus on the critical pandemic years.

- **Handling missing values:** Any null or missing values were identified and either removed or replaced with zeros (especially in heatmaps where consistency was required).

- **Feature engineering:** New temporal columns such as *Month-Year* and *Quarter* were created to allow monthly and quarterly trend analysis.

3. Data Analysis & Aggregation

- The dataset was **grouped and aggregated** at different levels:
  - **By country** to identify the top 5 and bottom 5 affected countries.
  - **By WHO region** to visualize regional intensity over time.
  - **By month and quarter** to generate trends and highlight peaks.
- Global aggregations were created to analyze the "mountain shape" progression of cases worldwide.

4. Data Visualization

- **Matplotlib & Seaborn:** Used to create line plots, pie charts, stacked bar charts, and heatmaps to represent case intensity and death surges.
- **Plotly & Dash:** Developed an **interactive dashboard** that allows dynamic filtering by country, region, and time. This included:
  - Line charts for daily new cases and deaths.
  - Stacked bar charts comparing cases vs deaths.
  - Choropleth maps showing cumulative global case distribution.
- Color schemes were chosen (reds for cases, blues for deaths) to make visualizations intuitive and highlight the severity of impact.

5. Tools & Technologies Used

- **Python 3.12**
- **Pandas, NumPy** – for data cleaning and aggregation.
- **Matplotlib, Seaborn** – for static data visualization.
- **Plotly, Dash** – for interactive dashboards and advanced visualizations.
- **Google Colab** – as the development environment.

6. Workflow Summary

The overall process followed in this project can be summarized in the following workflow:

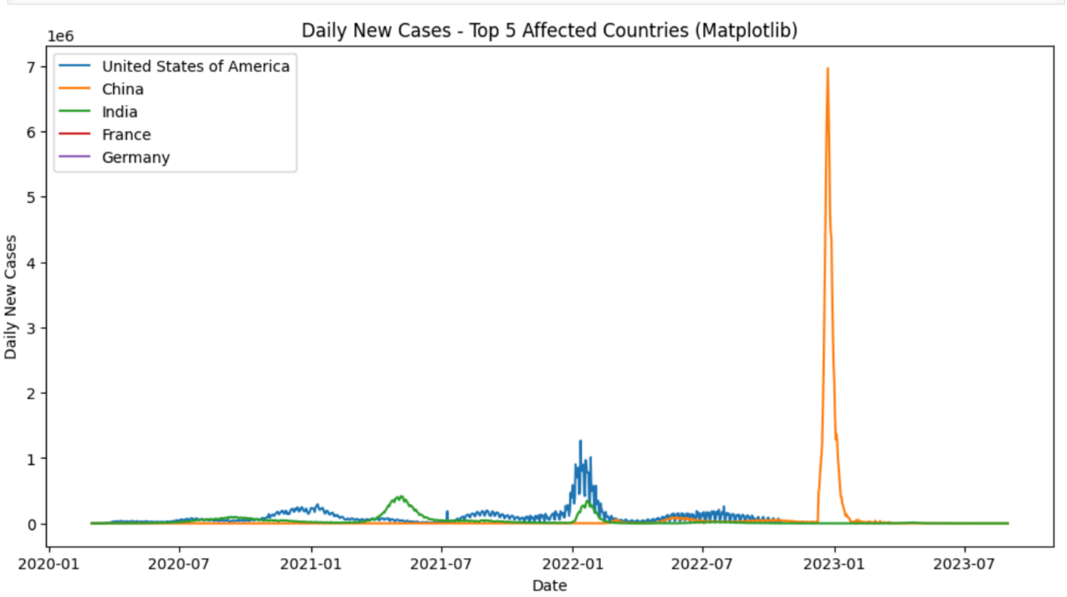
Data Collection → Data Cleaning & Preprocessing → Feature Engineering → Data Aggregation → Visualization (Static & Interactive) → Insights & Interpretation.*)*

Python Codes - https://github.com/sapcode2028/Visualizing-COVID-19-Pandemic-Data/blob/main/codebase-covid-19.py

5. Data Analysis and Results

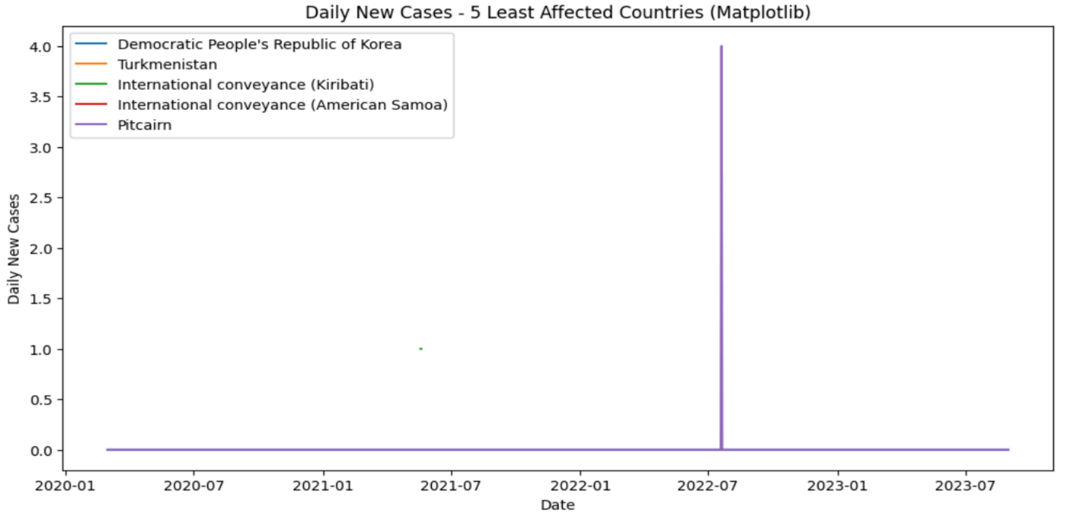
This section presents the findings obtained from the COVID-19 dataset after data preprocessing and visualization.

a) Top 5 Most Affected Countries (by cumulative cases)



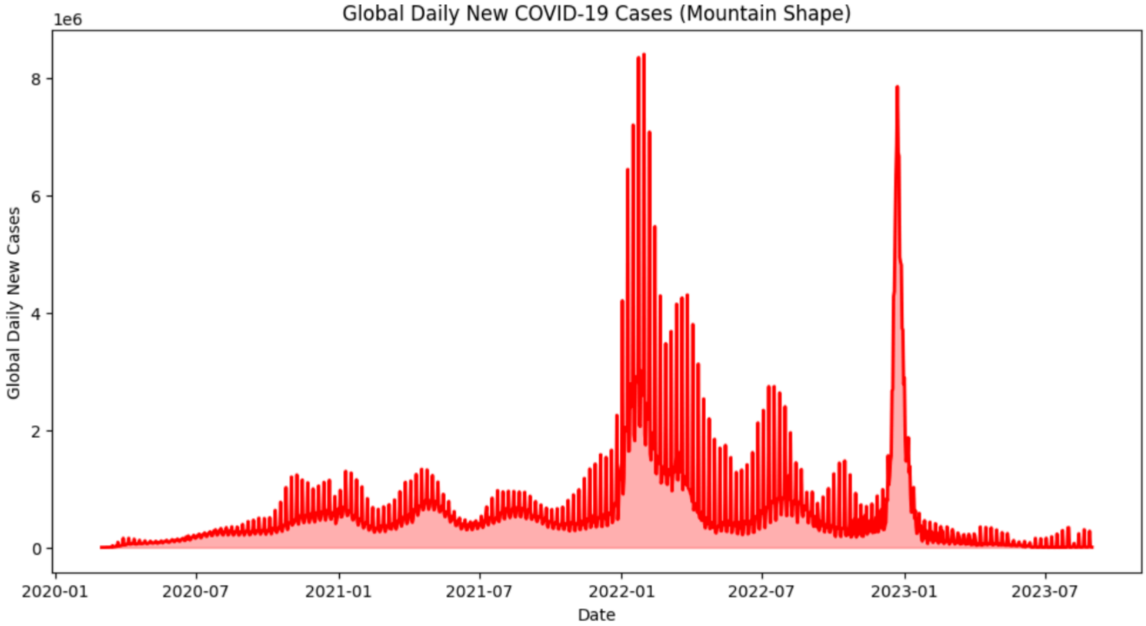
Observation: These countries consistently showed the steepest rise in daily new cases, reflecting hotspots of global COVID-19 transmission.

b)Bottom 5 Least Affected Countries

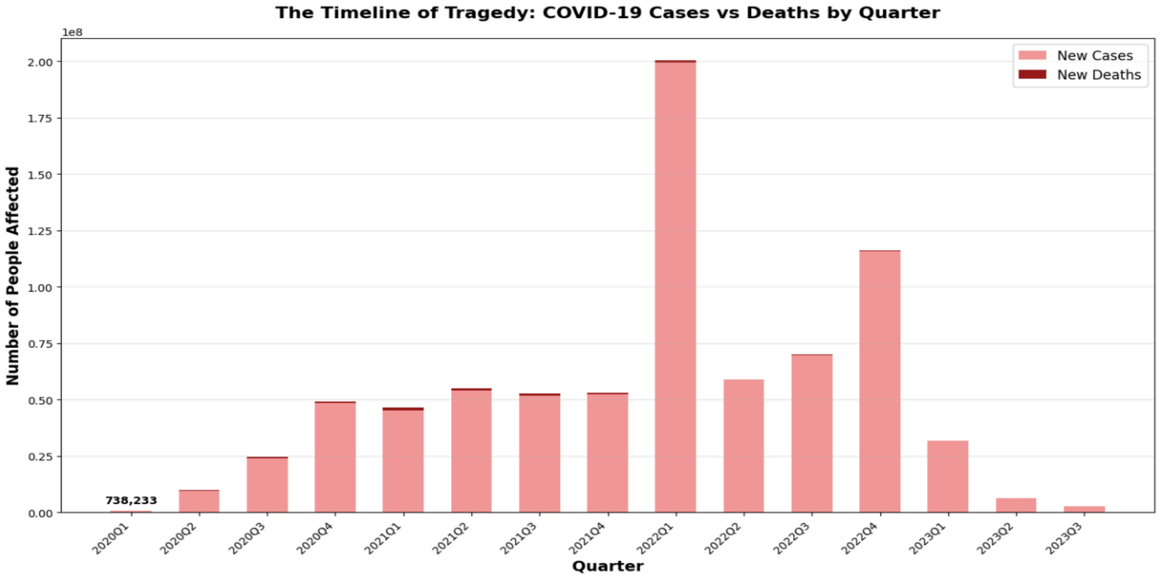


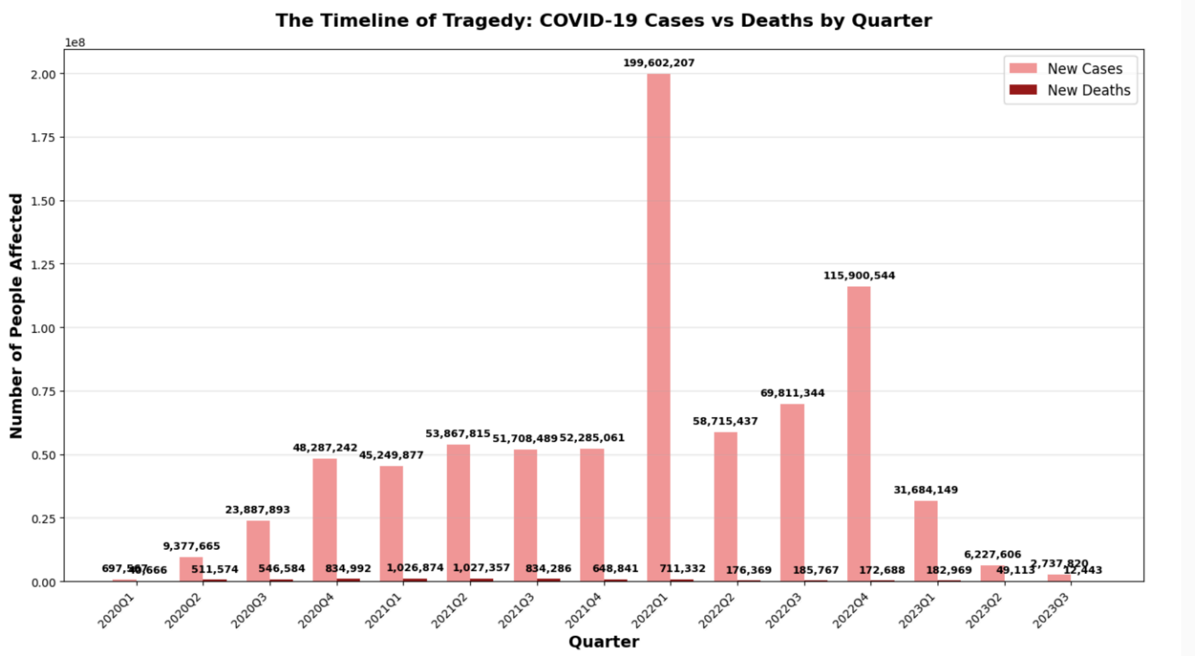
Observation: Many of these countries are either small island nations or regions with restricted international mobility, explaining their lower case counts.

c) Global Daily New Cases



d) Quarterly Cases vs Deaths

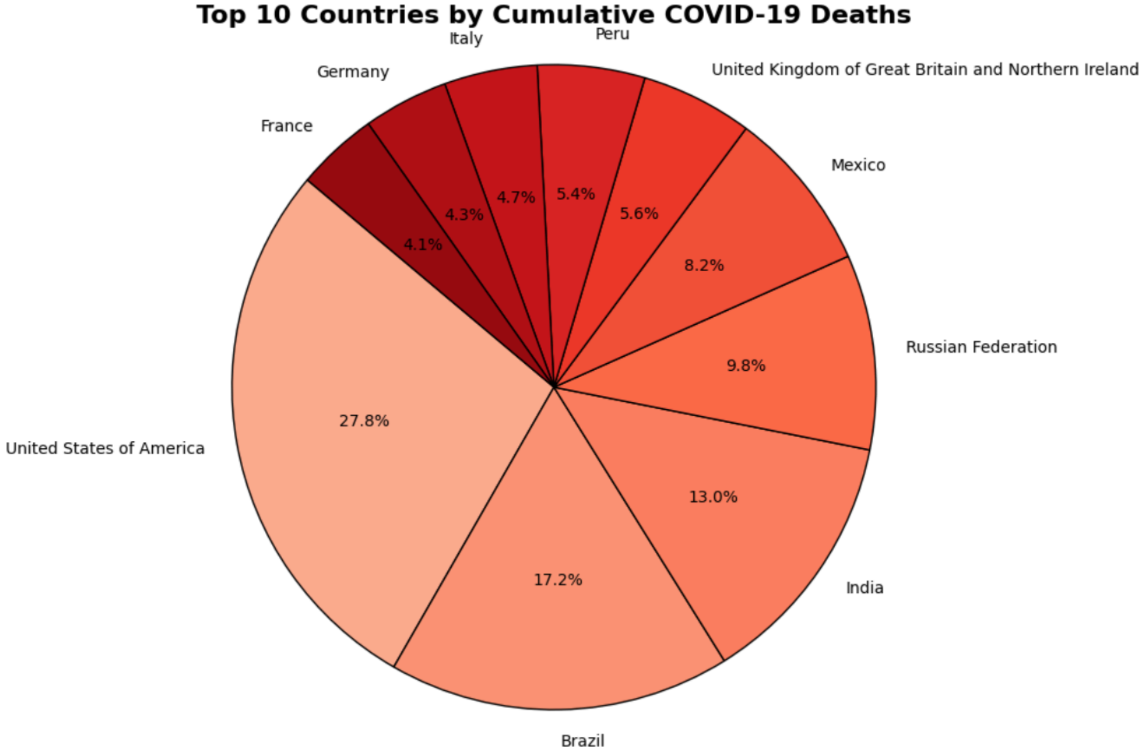




```
Quarter with highest cases: 2022Q1 (199,602,207 cases)
```

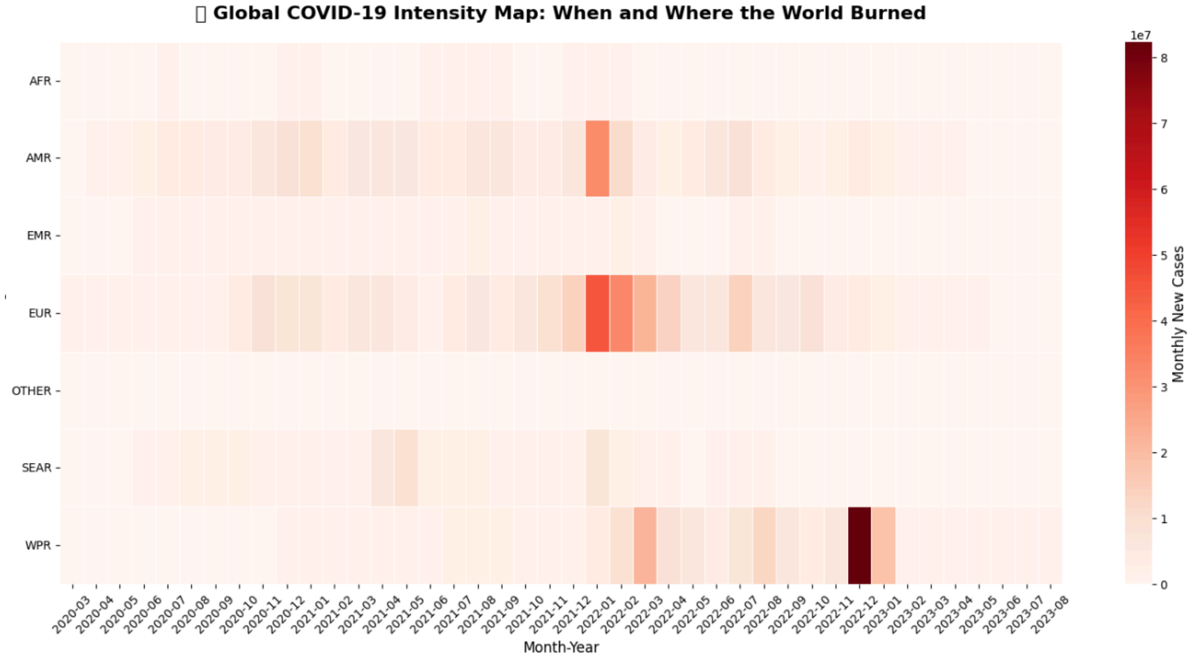
Observation: Deaths were proportionally higher during early waves, but case counts grew exponentially during later waves while deaths rose at a slower pace — reflecting vaccine rollouts and improved medical interventions.

e) Top 10 Countries by Cumulative Deaths (Pie Chart)



f) Heatmap Analyses

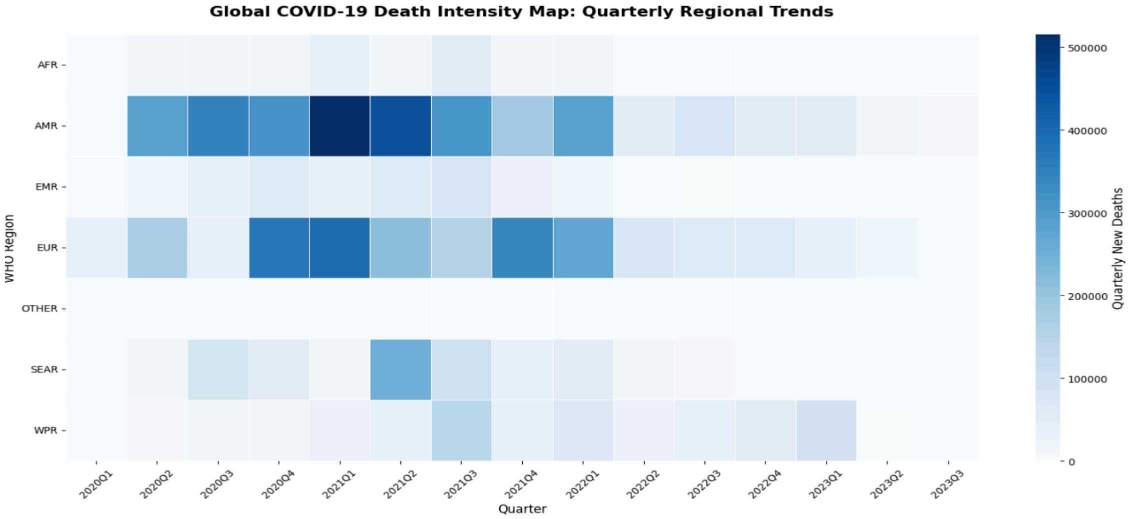
Monthly Case Intensity by WHO Region



□ Global COVID-19 Intensity Map: When and Where the World Burned

Certain regions showed **sharp monthly surges** corresponding to distinct waves. Example: Europe peaked during late 2020 and again in 2021, while the Americas saw sustained high numbers throughout 2020–2021.

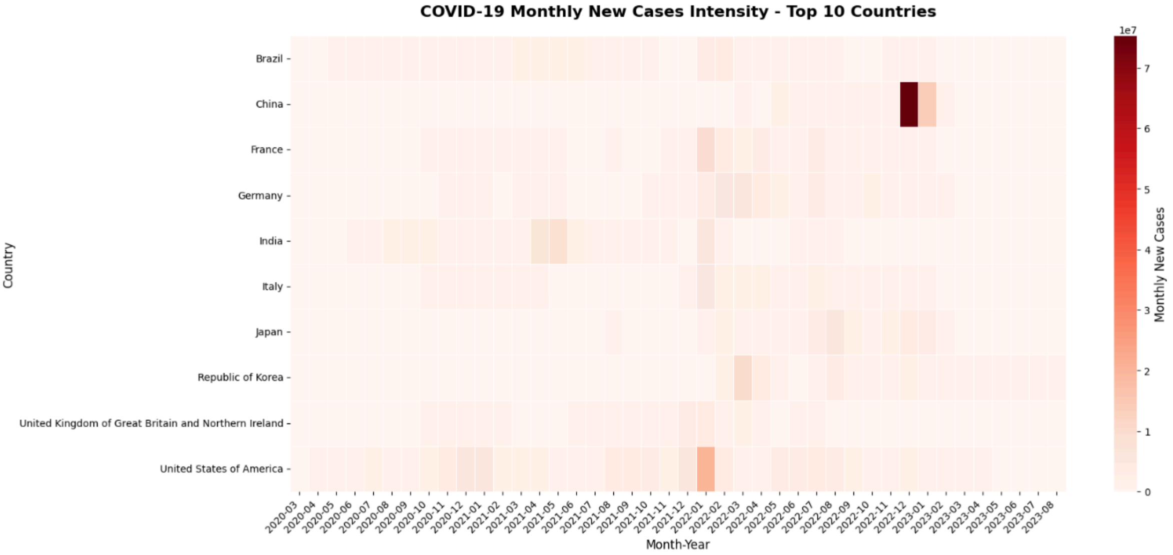
Quarterly Death Intensity by WHO Region



Global COVID-19 Death Intensity Map: Quarterly Regional Trends

Death surges aligned closely with waves in cases, but the proportion of deaths per case decreased over time.

Monthly Case Intensity by Top 10 Countries



Heatmaps clearly visualized **country-wise surges** in different time periods.
Example: India peaked in **April–May 2021** (Delta wave), while the U.S. showed multiple high-intensity phases.

2. Inferential Analysis

Since this project primarily focuses on descriptive visualization, formal **hypothesis testing** or **predictive modeling** was not performed. However, from visual analysis, the following insights can be inferred:

1. **Early vs Later Waves:** Earlier waves had a higher death-to-case ratio, suggesting weaker preparedness and lack of vaccines.

2. **Regional Variation:** Pandemic intensity varied significantly by region, influenced by healthcare infrastructure and intervention policies.

3. **Case–Death Relationship:** Although cumulative cases rose sharply after 2021, cumulative deaths rose at a slower rate, supporting the hypothesis that **vaccines reduced fatality rates**.

4. **Country-Specific Surges:** Different countries experienced waves at different times, indicating that global transmission was asynchronous and influenced by local factors.

3.Summary of Findings

- The **U.S., India, Brazil, France and Germany** were the most affected countries by case count.

- The **lowest impact** was observed in smaller nations and territories.
- The **global pandemic curve** showed three major waves, with Omicron being the largest in terms of cases but not in deaths.
- **Quarterly stacked bar charts** highlighted the disproportionate burden of cases vs deaths.
- **Pie charts** revealed concentration of deaths in a few major countries.
- **Heatmaps** effectively captured the spatio-temporal distribution of the pandemic's intensity.

6. Conclusion

The analysis of the COVID-19 dataset provided valuable insights into the global spread, intensity, and impact of the pandemic. By leveraging statistical aggregation, visual analytics, and interactive dashboards, this project successfully translated raw epidemiological data into understandable patterns and narratives.

From the findings, it is evident that the **United States, India, France, Germany, and Brazil** were the most severely affected countries in terms of cumulative cases, while other nations reported significantly lower case counts. Quarterly analysis of new cases and deaths revealed distinct peaks, particularly during the second and third waves, where both infections and mortality surged dramatically. Heatmaps further highlighted regional and temporal variations, showing when and where outbreaks intensified, while pie charts emphasized disproportionate death tolls across countries.

The interactive dashboard built with Plotly and Dash proved to be a powerful tool for dynamic exploration, enabling users to compare trends across regions and time frames. This underscores the importance of visualization in simplifying complex datasets for both researchers and policymakers.

In conclusion, the project demonstrates that **data-driven insights are crucial for monitoring pandemics**, identifying hotspots, and guiding effective interventions. Future work can extend this project by integrating vaccination data, mobility patterns, and socio-economic indicators to provide a more holistic understanding of pandemic dynamics. Additionally, predictive machine learning models could be developed to forecast future outbreaks and resource needs.

7. APPENDICES

Appendix A: References

1. World Health Organization (WHO) COVID-19 Dashboard: https://covid19.who.int/
2. Johns Hopkins University COVID-19 Data Repository: https://github.com/CSSEGISandData/COVID-19
3. Our World in Data COVID-19 Dataset: https://ourworldindata.org/covid-data
4. Official Python Libraries:
   - *pandas*: https://pandas.pydata.org/
   - *matplotlib*: https://matplotlib.org/
   - *seaborn*: https://seaborn.pydata.org/
   - *plotly*: https://plotly.com/python/

Appendix B: Survey Questionnaire

No direct survey was conducted as part of this project. The dataset used was secondary, sourced from publicly available WHO reports.

Appendix C: Github Link for Codes

The complete codebase for this project, including data preprocessing scripts, visualization notebooks, and dashboard implementation, is available at:
https://github.com/sapcode2028/Visualizing-COVID-19-Pandemic-Data

---

Appendix D: Dataset

- https://drive.google.com/file/d/1r054TYsGmBbIob_rM_KCjjbORV58ayCr/view