# VISUALIZATION AND ANALYSIS OF EPIDEMIC OUTBREAKS

*A Comparative Study of Ebola (2014–2016) and COVID-19 (2020–2023)*

**NIMA JIJO**

M.Sc. Statistics/2023 – 25,

Nirmala College, Muvattupuzha

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 focuses on the visual analytics of epidemic outbreaks, with case studies on the Ebola outbreak (2014–2016) and the COVID-19 pandemic (2020–2023). Both datasets were obtained from the World Health Organization (WHO) and processed to ensure consistency and accuracy. Preprocessing involved renaming variables, handling missing values, formatting date fields, and organizing data into structured time series. Visualization techniques included line plots, stacked and double bar charts, pie charts, heatmaps, and interactive dashboards.

For Ebola, analysis revealed that Liberia, Sierra Leone, and Guinea were the most affected countries, with high fatality rates and concentrated spread in West Africa. In contrast, COVID-19 exhibited global waves of infections, with the United States, China, and India emerging as epicentres at different times. Interactive dashboards allowed dynamic exploration of cumulative cases and deaths, improving interpretability. The results emphasize the critical role of modern data analytics in supporting public health surveillance, policy-making, and epidemic preparedness.

# 2. INTRODUCTION

The 21st century has witnessed two major infectious disease emergencies that tested global health security systems: the 2014-2016 West Africa Ebola outbreak and the 2020-2023 COVID-19 pandemic. These events, while vastly different in scale and characteristics, provide valuable insights into disease transmission dynamics, public health response effectiveness, and the critical role of data analysis in emergency management.

The emergence of global epidemics such as Ebola (2014–2016) and COVID-19 (2020–2023) has demonstrated the vital role of data analytics in understanding, tracking, and mitigating health crises. Reliable and timely information about cases and deaths is essential for decision-making in healthcare, government, and international organizations. However, the raw data collected during such outbreaks is often large, complex, and difficult to interpret directly. Effective data preprocessing, visualization, and time-series analysis are therefore crucial to extract actionable insights from epidemiological data.

This project leverages Python-based data analysis and visualization techniques to study and compare two major outbreaks: Ebola and COVID-19. The Ebola dataset (sourced from WHO) focuses on the cumulative number of confirmed, probable, and suspected cases and deaths in affected African countries between 2014 and 2016. In contrast, the COVID-19 dataset provides a global perspective, capturing daily new and cumulative cases and deaths from March 2020 to August 2023.

The work was implemented using Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Plotly, with Google Colab as the development platform. The methodology included data collection, preprocessing (renaming variables, handling missing values, formatting dates), exploratory data analysis (EDA), and visualization through line plots, bar charts, pie charts, heatmaps, and interactive dashboards. Comparative analysis between the two outbreaks was also performed, highlighting both differences and common patterns in epidemic progression.

Background research involved studying previous works on epidemic modelling, visualization techniques, and time-series forecasting. While Ebola was a geographically concentrated but highly fatal outbreak, COVID-19 was a global pandemic with multiple waves but lower relative fatality. The study emphasizes how visualization tools can simplify epidemic monitoring, provide regional comparisons, and support resource allocation.

This report documents the end-to-end workflow, including dataset handling, descriptive and inferential analysis, visualization, and results interpretation. It demonstrates how computational techniques contribute to public health monitoring and builds a foundation for integrating predictive models in the future.

Training Received in Internship (first 2 weeks):

Programming & Data Handling:

1. Data, Variables, Lists, Loops (Data Structures) – Basics of storing and manipulating data, understanding variable types, lists for collection, and loops for repetitive tasks.
2. Classes, Functions, Object-Oriented Programming (OOPs) – Creating reusable code, designing classes, and understanding inheritance, encapsulation, and polymorphism.
3. NumPy – Efficient numerical computations with arrays and mathematical operations.
4. Pandas – Data manipulation using DataFrames, handling missing data, filtering, grouping, and summarizing datasets.

Machine Learning & AI:

5. Machine Learning Overview – Introduction to supervised vs. unsupervised learning, model training, evaluation metrics, and ML workflows.
6. Regression Lab – Implementation of regression algorithms to predict numerical outcomes and evaluation using RMSE and $R^2$.
7. Classification Lab – Hands-on experience with classification algorithms, performance metrics such as accuracy, precision, recall, and F1-score.
8. LLM (Large Language Model) Fundamentals – Understanding the basics of LLMs, their architecture, and applications.

Soft Skills:

9. Communication Skills – Training on effective communication, presentation skills, and professional collaboration in team environments.

# 3. PROJECT OBJECTIVE

**3.1** Objectives of Project 1: Ebola Outbreak (2014–2016)

i) Preprocess and clean the Ebola dataset

- Import dataset and inspect structure.
- Rename lengthy columns for readability.
- Convert Date into proper datetime format.
- Handle missing values and sort data by Country and Date.
- Subset into key variables: Country, Date, Cumulative Cases, Cumulative Deaths.

ii) Transform data into analysable structures

- Aggregate cumulative cases and deaths at country and global levels.
- Create new features such as Quarter and Month-Year.
- Generate subsets for top 5 most affected and least 5 affected countries.

iii) Perform exploratory data analysis (EDA)

- Line plots of cumulative cases and deaths for top/least affected countries.
- Global mountain-shaped epidemic curve.
- Quarterly stacked and double bar charts for cases vs deaths.
- Pie chart of top 10 countries by deaths.
- Heatmaps of monthly cumulative cases by country.

iv) Interpret and communicate insights

- Highlight epicentres of the.
- Compare mortality vs case burden regionally.
- Document key findings and public health implications.

**3.2** Objectives of Project 2: COVID-19 Pandemic (2020–2023)

i) Preprocess and clean the COVID-19 dataset

- Import WHO dataset covering global daily cases and deaths.
- Filter records between March 2020 and August 2023.

- Select relevant columns (Date, Country, Region, Cases, Deaths).
- Handle missing values and standardize time series.

ii) Transform data into structured formats

- Aggregate daily cases/deaths into quarterly summaries.
- Create new features such as WHO Region and Month-Year.
- Generate subsets for top 5 and least 5 affected countries.

iii) Perform exploratory data analysis (EDA)

- Line plots showing cumulative cases and deaths across countries.
- Global mountain-shaped curve to capture pandemic waves.
- Quarterly stacked vs double bar charts for new cases and deaths.
- Pie charts of top 10 countries by deaths.
- Heatmaps for case intensity by WHO region and countries.
- Interactive dashboards (line plots, choropleth maps, stacked bars).

iv) Interpret and communicate insights

- Identify pandemic peaks and waves.
- Compare mortality trends across regions.
- Show contrasts between early and late pandemic phases.
- Document findings with implications for global health monitoring.

# 4. METHODOLOGY

The methodology describes the step-by-step process followed to analyse and visualize epidemic datasets for Ebola (2014–2016) and COVID-19 (2020–2023). The workflow included data collection, cleaning and preprocessing, exploratory data analysis (EDA), transformation for visualization, interactive dashboards, and interpretation.

### Data Collection

Two datasets were used for this project:

- Ebola Dataset (2014–2016): Obtained from WHO records, containing variables such as Country, Date, Cumulative Cases, and Cumulative Deaths.
- COVID-19 Dataset (2020–2023): Sourced from WHO's global time-series records, with variables including Date, Country, WHO Region, Daily Cases, Daily Deaths, Cumulative Cases, and Cumulative Deaths.

Both datasets were imported into Google Colab using the pandas library. These datasets served as the foundation for visualization and trend analysis.

### Data Cleaning and Preprocessing

The datasets underwent thorough preprocessing to ensure uniformity and consistency.

- Long column names were renamed into shorter, descriptive labels (e.g., "Cumulative no. of confirmed, probable and suspected cases" → "Cumulative_cases").
- Dates were converted into datetime format.
- Missing values were handled appropriately (imputation or removal).
- Datasets were sorted by Country and Date to maintain chronological sequence.
- Unnecessary or redundant columns were dropped, retaining only relevant variables.
- New features such as Quarter and Month-Year were created to allow time-based aggregation.
- Data subsets were prepared for top and least affected countries for comparative analysis.

### Exploratory Data Analysis (EDA)

EDA was performed to uncover trends, patterns, and anomalies in the datasets. Key steps included:

- Descriptive Statistics: Summary of cumulative cases and deaths by country.
- Line Graphs: Temporal progression of cases for top 5 and least 5 affected countries.
- Global Trend Plots: Mountain-shaped epidemic curves showing cumulative global cases.
- Quarterly Analysis: Stacked and double bar charts comparing quarterly cases vs deaths.
- Pie Charts: Top 10 countries by cumulative deaths.
- Heatmaps: Monthly or regional progression of cases to identify hotspots.

EDA revealed key patterns such as the epidemic peaks, country-level disparities, and global waves of infection.

### Data Transformation for Visualization

The datasets were transformed into suitable structures for visualization:

- Grouped data by Country and Date for time-series plots.
- Aggregated global cases and deaths for overall epidemic curves.
- Created pivot tables for heatmaps (Country × Month-Year).
- Added categorical grouping by WHO Region for COVID-19.
- Extracted maximum cumulative values for pie chart comparisons.

This transformation step enabled clear representation of epidemic progression.

### Interactive Dashboard Development

Using Plotly Express and Plotly Graph Objects, interactive dashboards were created for deeper exploration:

- Global Line Plots: Interactive cumulative cases and deaths over time.
- Stacked Bar Charts: Dynamic view of cases vs deaths by country.
- Choropleth Maps: Geospatial visualization of outbreak intensity across countries.

These dashboards allowed users to zoom, filter, and hover for specific data, making the analysis more engaging and insightful.

## Results Interpretation

The results were analysed for both outbreaks:

- Ebola analysis showed West Africa (Guinea, Liberia, Sierra Leone) as the epicentre, with high case-to-death ratios.
- COVID-19 analysis identified global pandemic waves, with United States, India, and Brazil among the most impacted.
- Comparative insights highlighted Ebola's localized but highly fatal spread versus COVID-19's global reach with multiple waves but lower fatality rates.

Reproducibility and Reusability

- All code and analysis were conducted in Google Colab notebooks.
- Datasets were cleaned and saved in processed form.
- Interactive dashboards were preserved for reuse in new datasets.
- The workflow ensures reproducibility, efficiency, and scalability for future epidemic studies.

Project Code Repository:

# 5. DATA ANALYSIS AND RESULTS

## Descriptive Analysis

Dataset Overview:

- Ebola Dataset (2014–2016):
  - Variables: Country, Date, Cumulative Cases, Cumulative Deaths
  - Coverage: Primarily West African countries
- COVID-19 Dataset (2020–2023):
  - Variables: Date, Country, WHO Region, Daily Cases, Daily Deaths, Cumulative Cases, Cumulative Deaths
  - Coverage: Global (200+ countries)
  - Time Span: March 2020 – August 2023

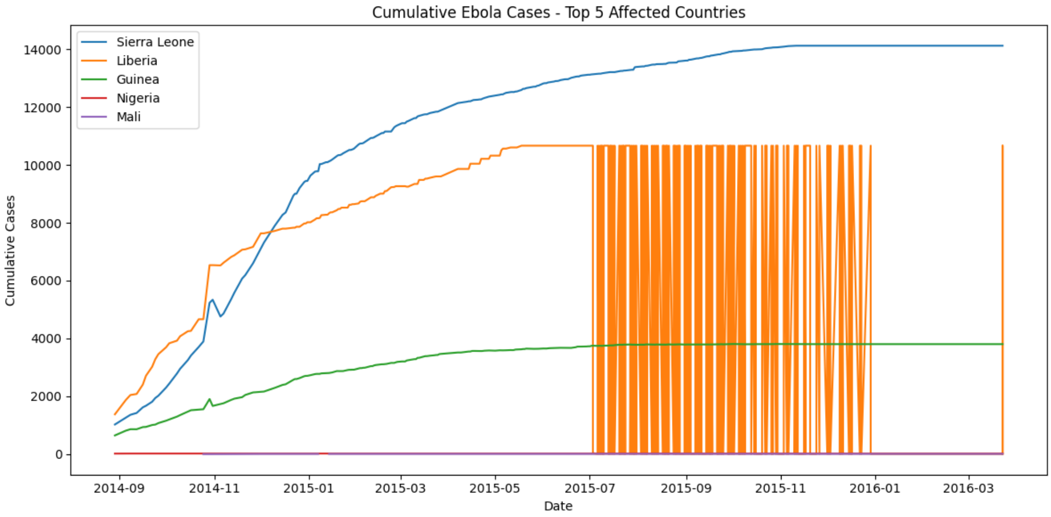
Observations:

- Ebola data is more localized but fatal, concentrated in Liberia, Sierra Leone, and Guinea.
- COVID-19 data is global, with large case volumes, recorded in multiple waves.

## Line Graphs

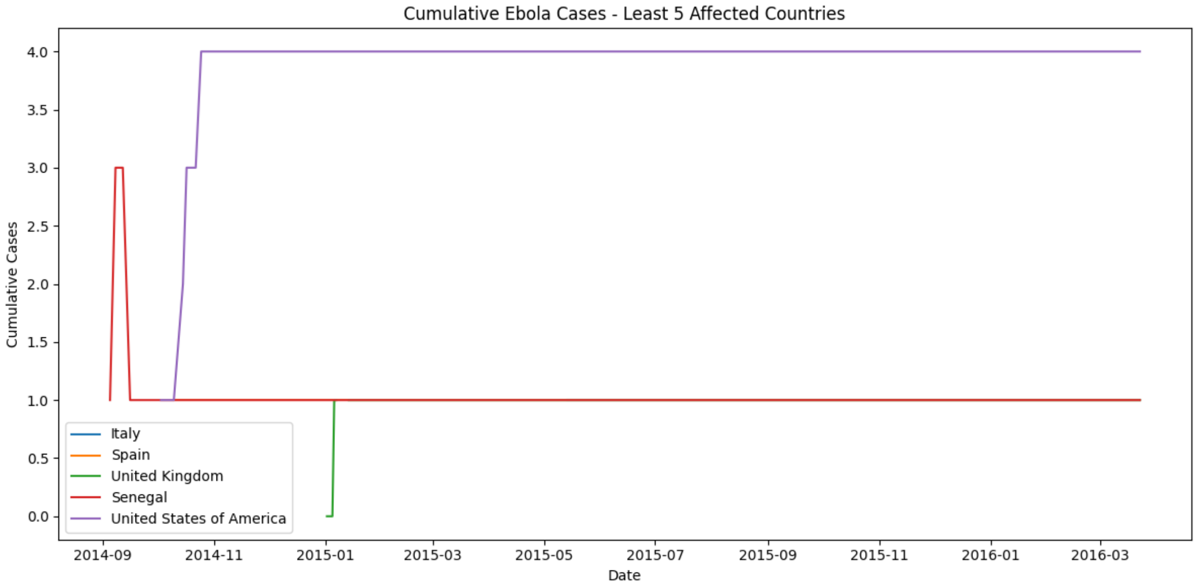
A line graph shows data points connected over time, highlighting temporal progression. It is especially useful for identifying peaks, trends, and growth rates.

## EBOLA



## Explanation of graph:

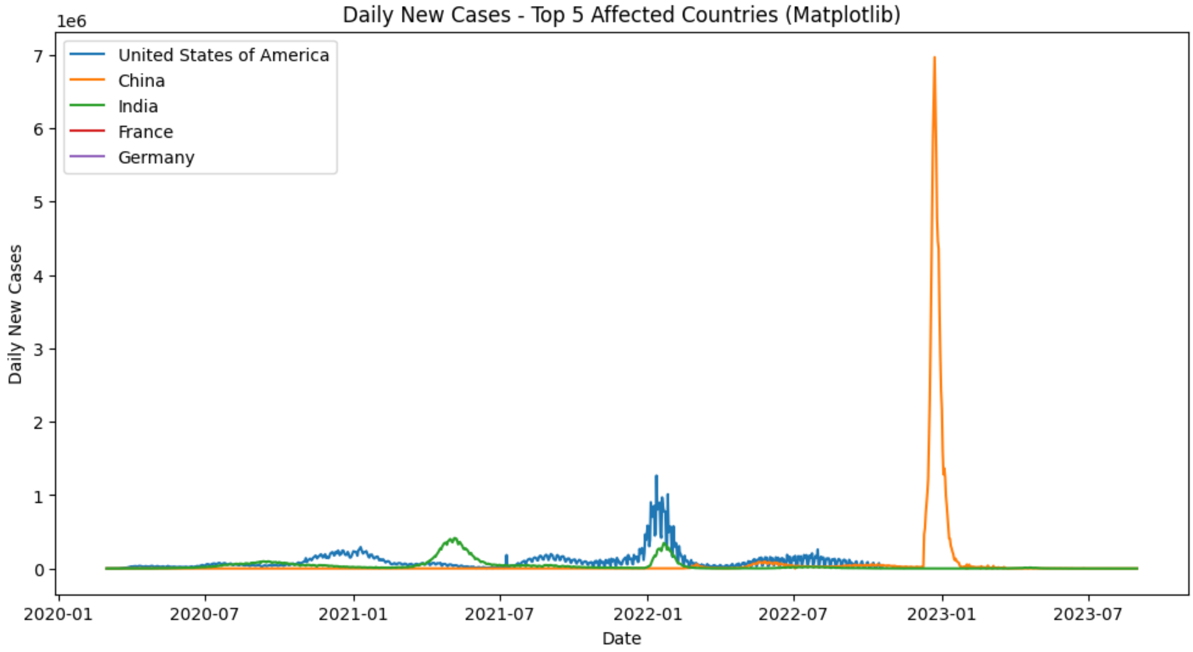
The line graph above illustrates the progression of cumulative Ebola cases in the five most affected countries during the 2014–2016 outbreak. Sierra Leone and Liberia recorded the highest case counts, with Sierra Leone surpassing 14,000 cases and Liberia stabilizing around 10,600 cases. Guinea followed as the third most affected nation, with cases peaking at nearly 4,000. In contrast, Nigeria and Mali reported minimal spread, and their curves remain nearly flat throughout the period, reflecting effective containment efforts.



## Explanation of graph:

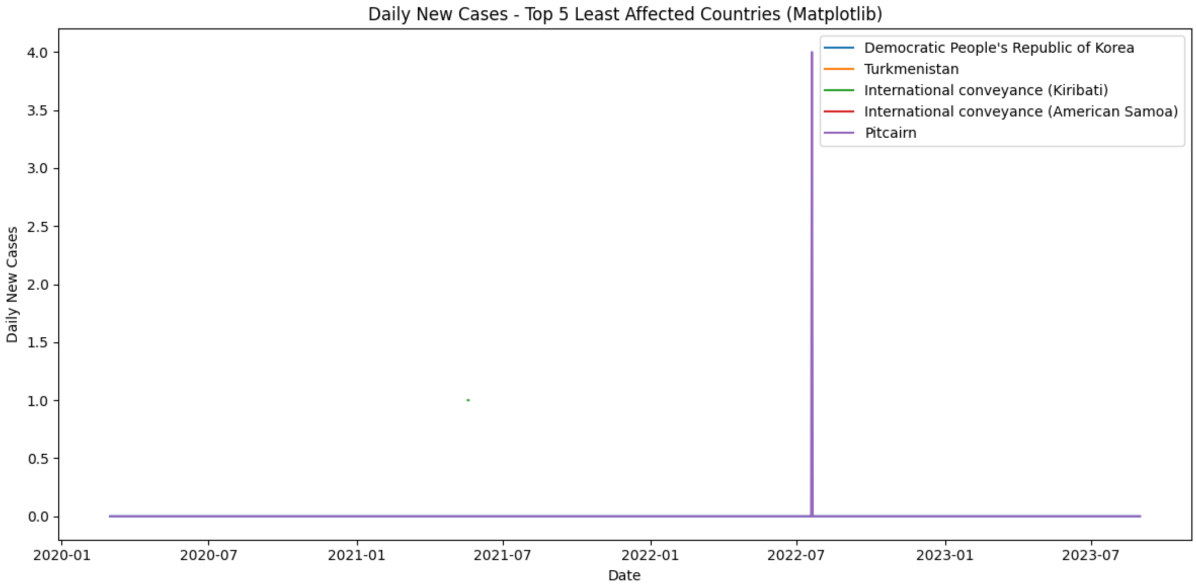
The line graph above presents the cumulative Ebola cases in the five least affected countries during the 2014–2016 outbreak. Italy, Spain, the United Kingdom, Senegal, and the United States reported only isolated cases, with totals never exceeding four. The flat curves indicate that secondary transmission was successfully prevented, reflecting strong containment measures and effective healthcare responses in these nations. Unlike the steep trajectories observed in West Africa, these countries experienced minimal spread, underscoring the localized nature of the outbreak and the importance of rapid intervention in halting global transmission.

## COVID-19



## Explanation of graph:

This line graph tracks the progression of cumulative COVID-19 cases over time for the five most affected countries: **United States, China, India, France, and Germany**. The curves clearly show multiple waves of infection throughout the pandemic. India and China also recorded steep increases. The visualization highlights both the scale and intensity of COVID-19 spread across major global epicentres.



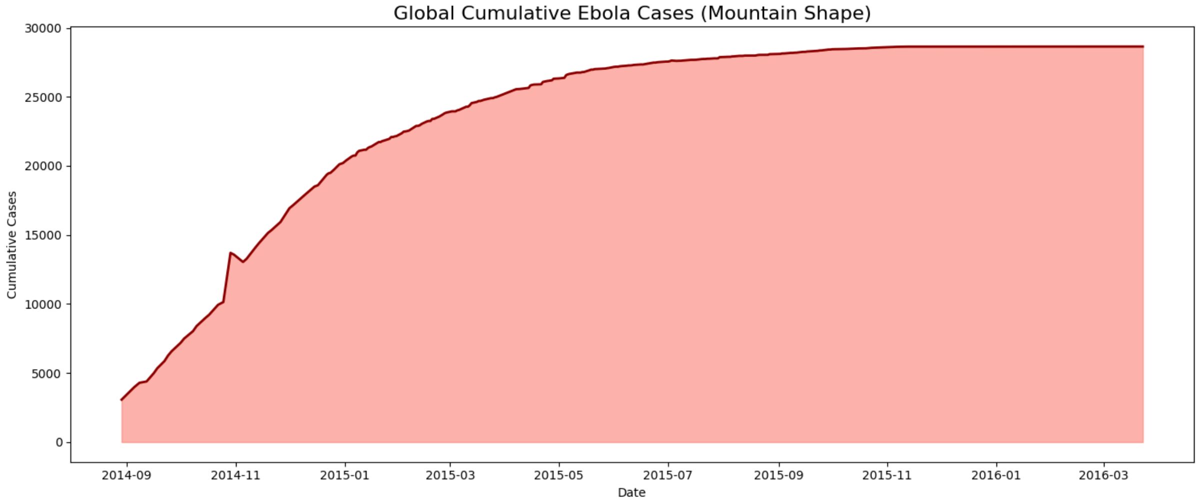
**Explanation of graph:**

This line graph displays cumulative cases in the five least affected countries during the same period. Unlike the steep trajectories of the top countries, these curves remain almost flat, indicating minimal caseloads. Many of these countries reported only a few hundred or thousand cases in total.

## Mountain-Shaped Global Curves

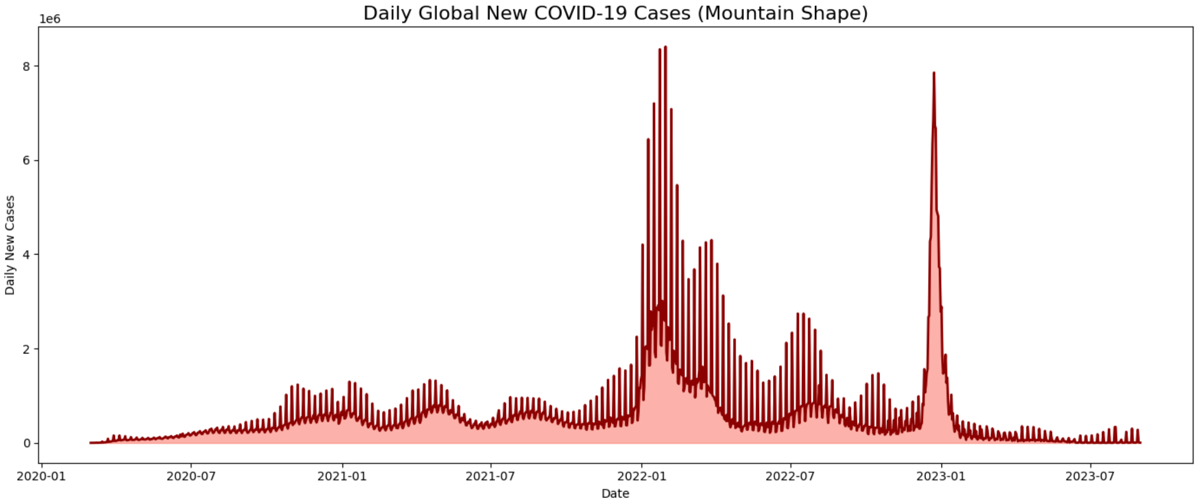
These plots aggregate global cumulative cases over time, resembling epidemic waves.

## EBOLA



The mountain-shaped graph of cumulative Ebola cases illustrates a sharp rise in infections during mid-2014, followed by a gradual slowing in 2015 and stabilization by 2016. The steep slope during the peak months reflects the uncontrolled spread in Guinea, Liberia, and Sierra Leone.

## COVID-19



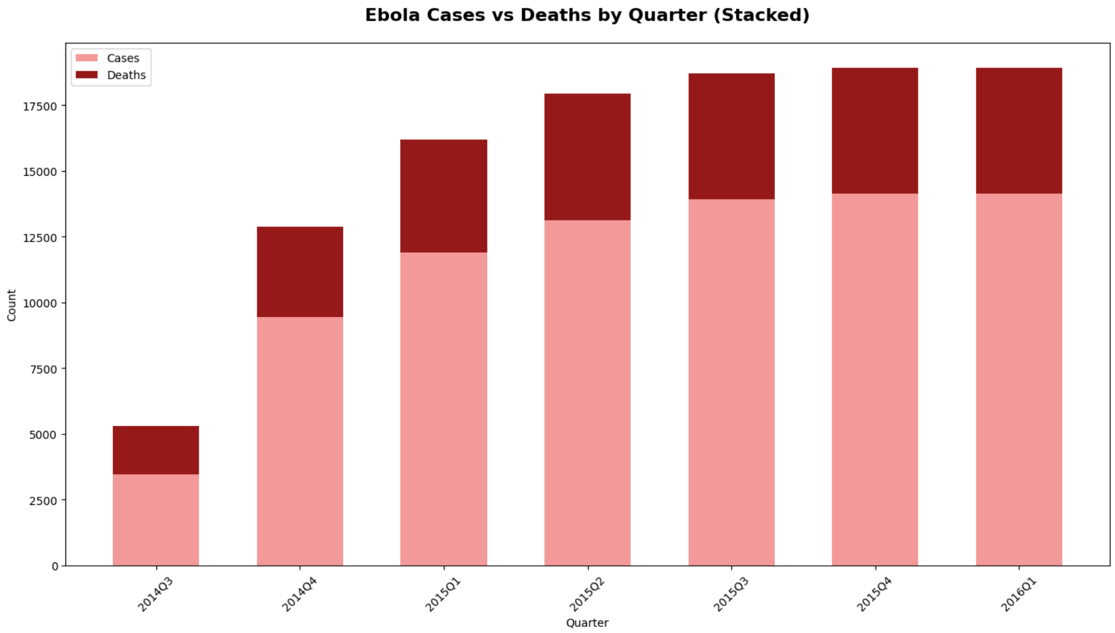
In contrast, the mountain-shaped curve for COVID-19 shows multiple steep waves across the years 2020–2023. The first rise in early 2020 reflects the initial outbreak, followed by a much larger surge in 2021 corresponding to the Delta variant. The highest peak occurs in late 2021 and early 2022 during the Omicron wave, when cases increased explosively worldwide.

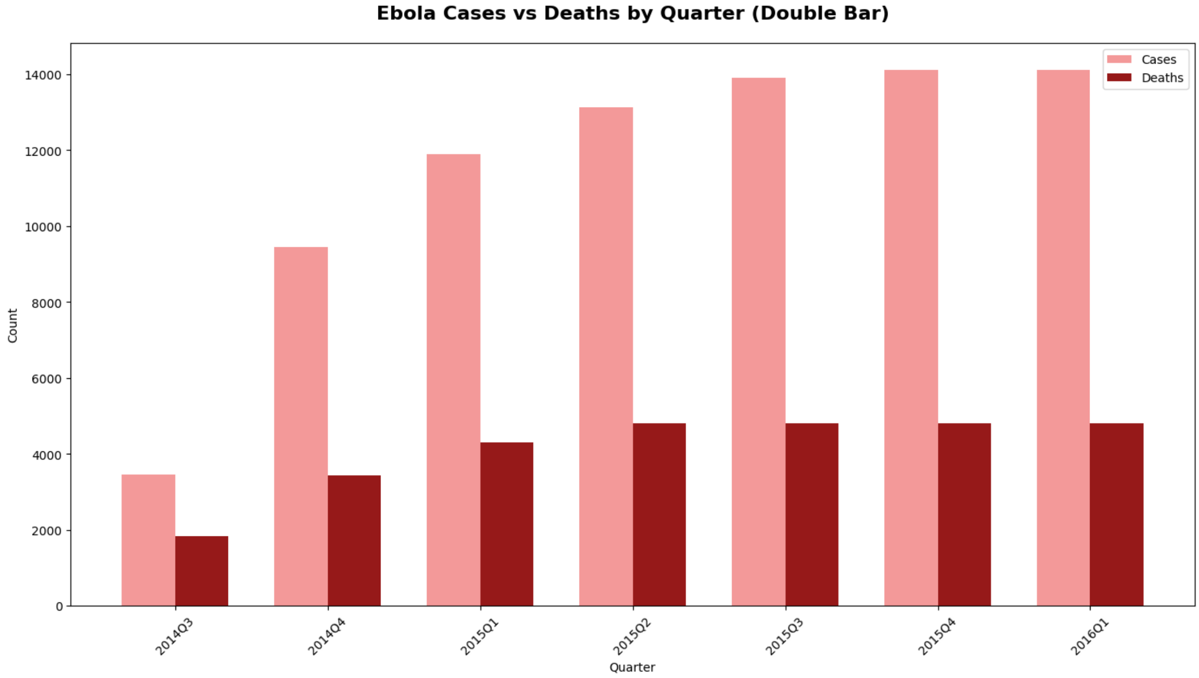
### Quarterly Bar Charts

Bar charts summarize quarterly maximums of cases and deaths.

- **Stacked Bar Chart:** Shows cumulative relationship of cases vs deaths.
- **Double Bar Chart:** Directly compares case counts against death counts per quarter.

# EBOLA

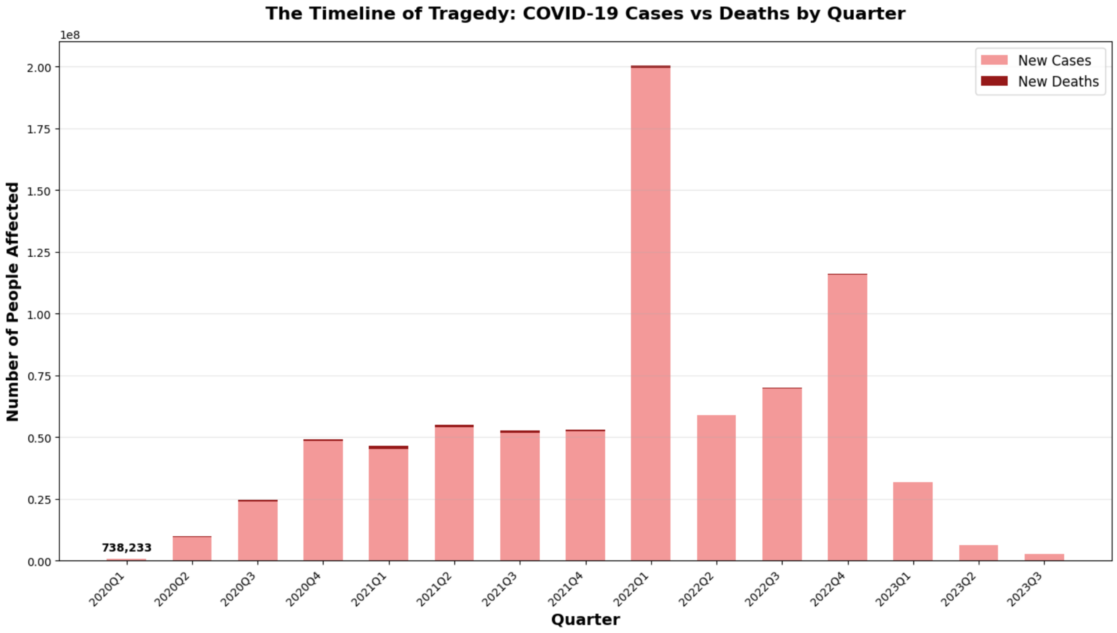


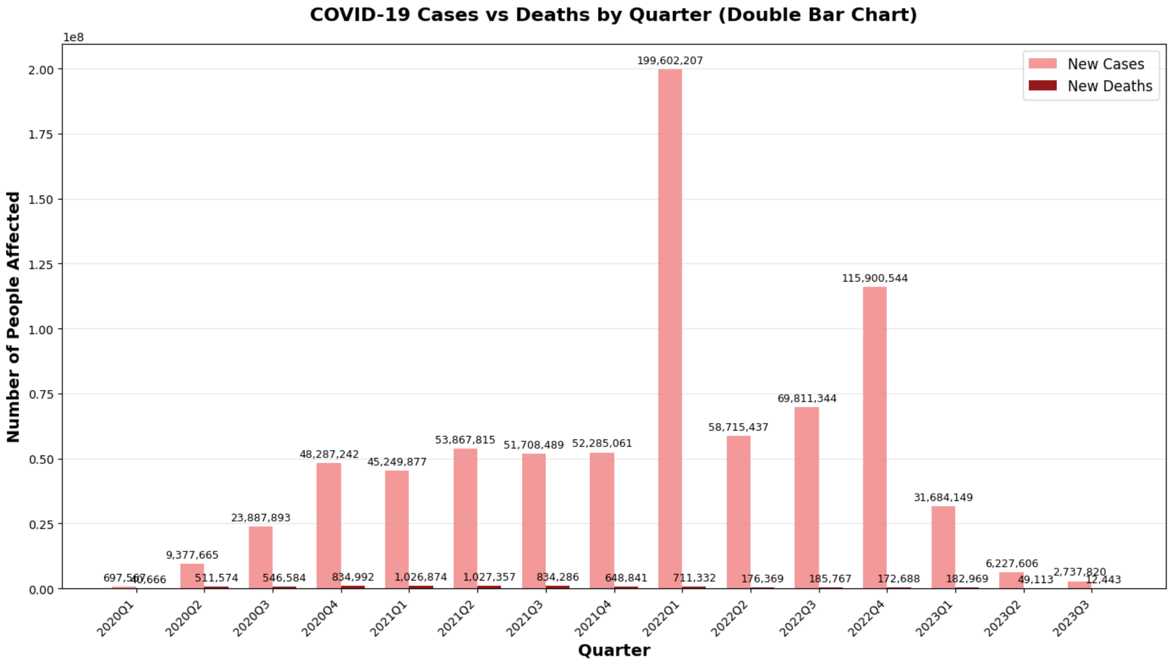


The stacked bar chart provides a combined view of cumulative Ebola cases and deaths on a quarterly basis. Each bar represents the total cases, with deaths stacked on top, making it easy to compare the relative burden of mortality within the overall outbreak. The visualization shows that as cases increased sharply in 2014, deaths also rose proportionally, highlighting the high fatality rate of Ebola. By late 2015 and 2016, both cases and deaths stabilized, and the bars level off, indicating that the epidemic had been brought under control. This chart emphasizes the close relationship between cases and deaths, underscoring Ebola’s severity.

The double bar chart separates cases and deaths into side-by-side bars for each quarter, allowing a clearer comparison between the two outcomes. The height difference between the case bars and death bars illustrates the magnitude of mortality relative to infections. During the peak quarters of 2014, the gap between cases and deaths was narrower, reflecting Ebola's high case fatality ratio. In later quarters, the bars show smaller increases, indicating that both infections and deaths had slowed significantly. This chart highlights the proportional impact of deaths across different time periods, offering a direct view of how the epidemic evolved.

## COVID-19





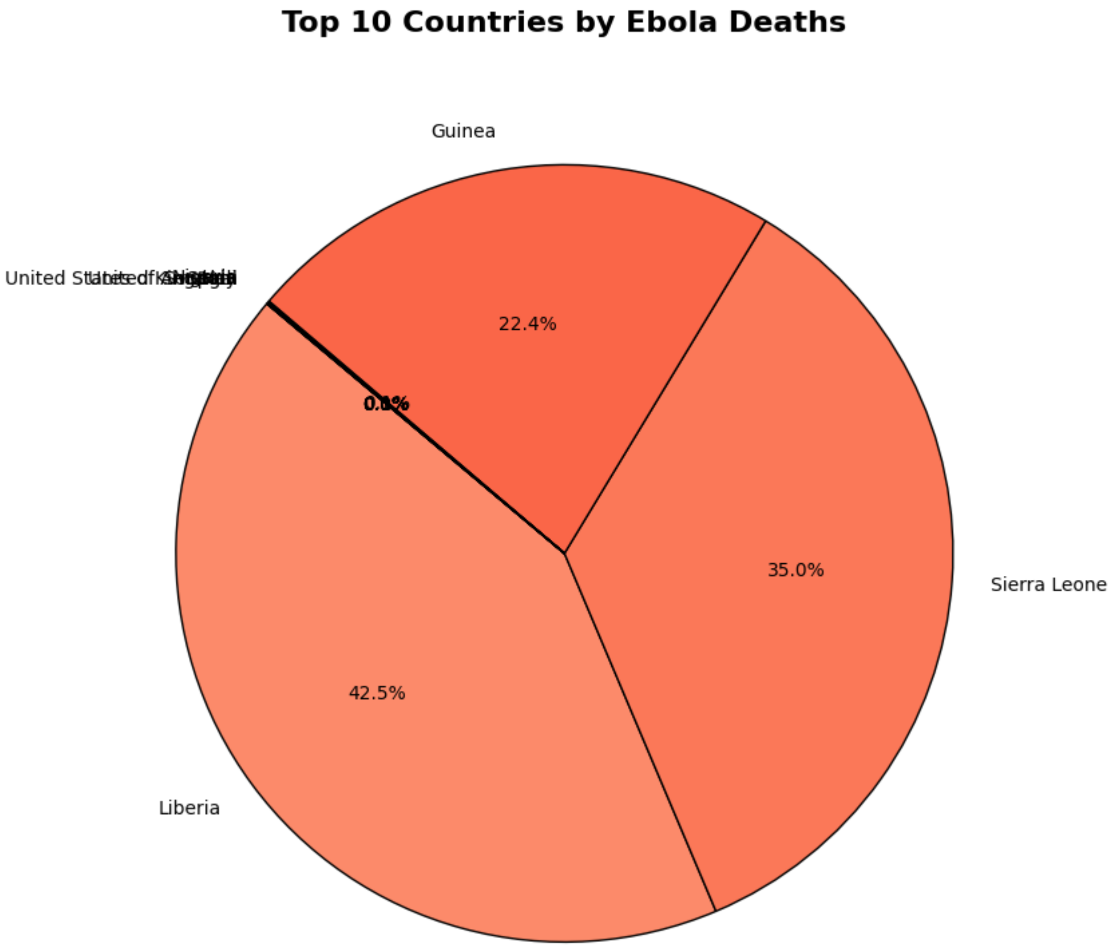
The stacked bar chart illustrates the quarterly progression of global COVID-19 cases and deaths. Each bar combines the number of cases and deaths, showing their cumulative contribution over time. The visualization highlights the massive spikes in 2020 and 2021, corresponding to the early waves and the Delta variant surge.

The double bar chart places quarterly cases and deaths side by side, making the contrast between infections and fatalities more explicit. The chart shows that while cases soared dramatically in successive waves, deaths did not rise at the same rate, leading to a widening gap between the two bars in later quarters.

### Pie Charts

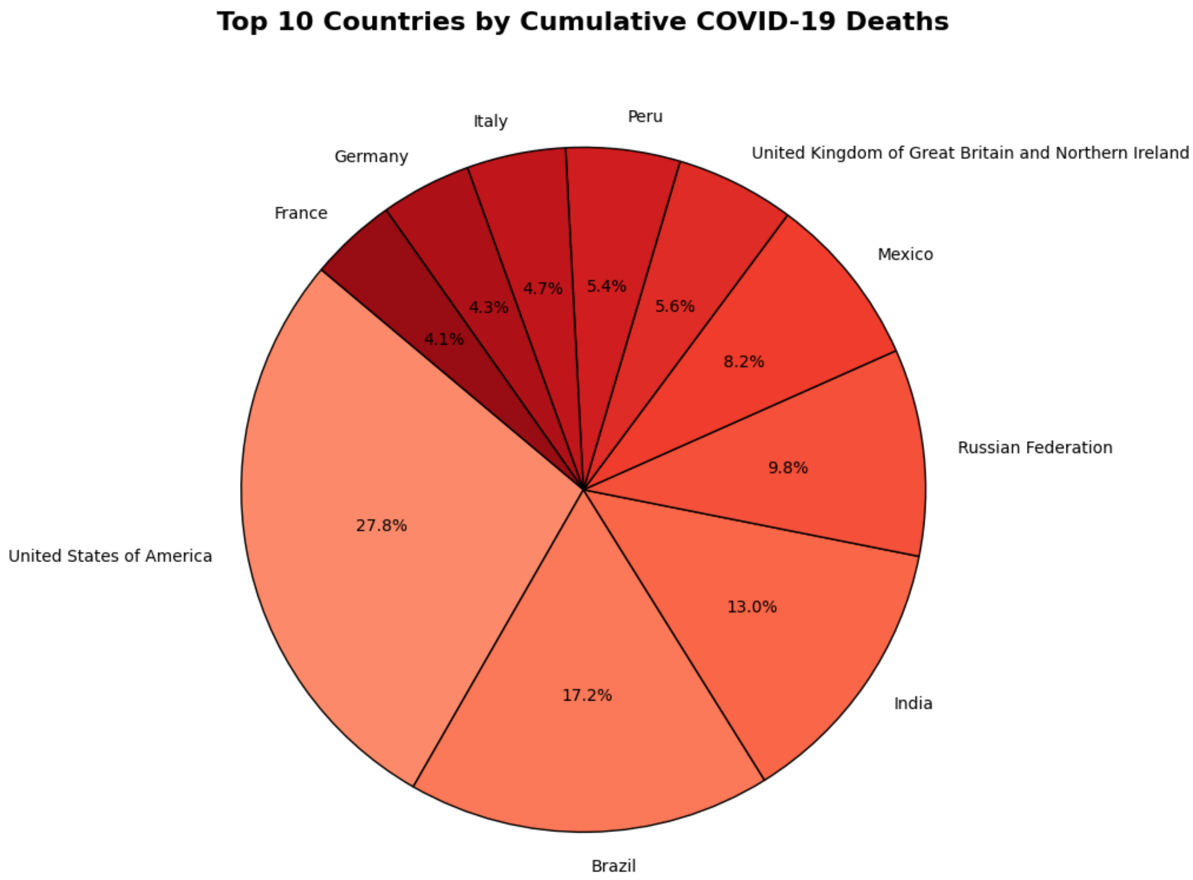
Pie charts summarize proportional contribution of countries.

### EBOLA



The pie chart illustrates the distribution of Ebola deaths across the ten most affected countries during the 2014–2016 outbreak. Liberia, Sierra Leone, and Guinea dominate the chart, together accounting for more than 85% of all reported deaths. The remaining countries contribute only small slices, reflecting the highly localized nature of the epidemic.

## COVID-19

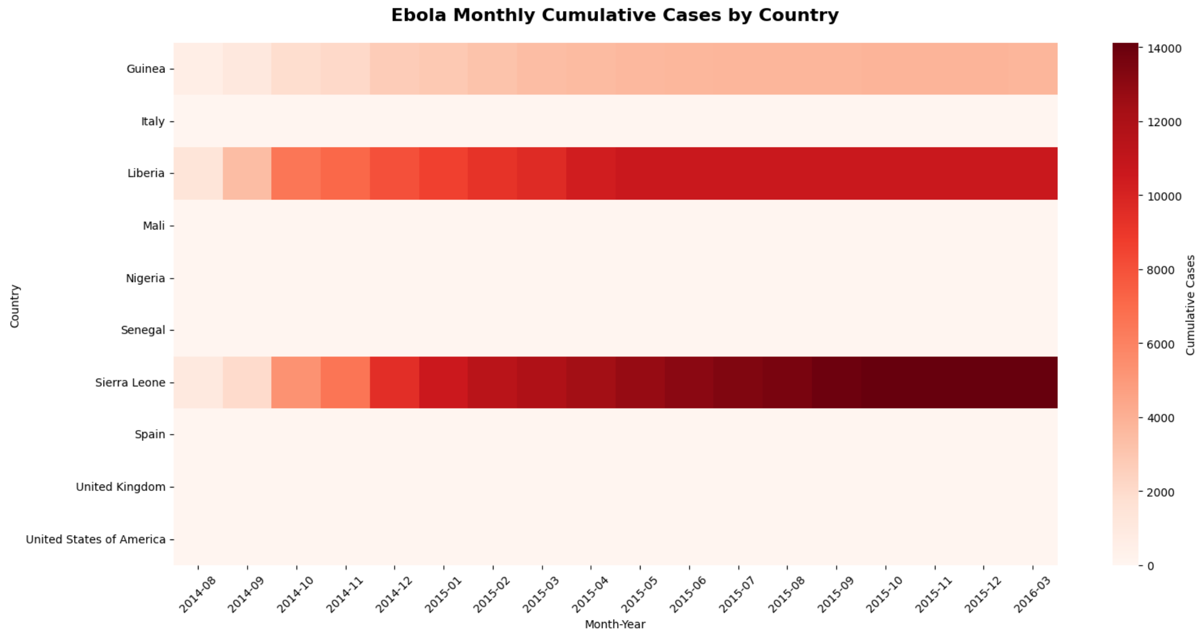


The COVID-19 pie chart highlights the countries with the highest cumulative death tolls between 2020 and 2023. The United States occupies the largest slice, followed by significant shares for India, Brazil, Russia, and Mexico. Other countries in the top ten contribute smaller but still considerable portions.

## Heatmaps

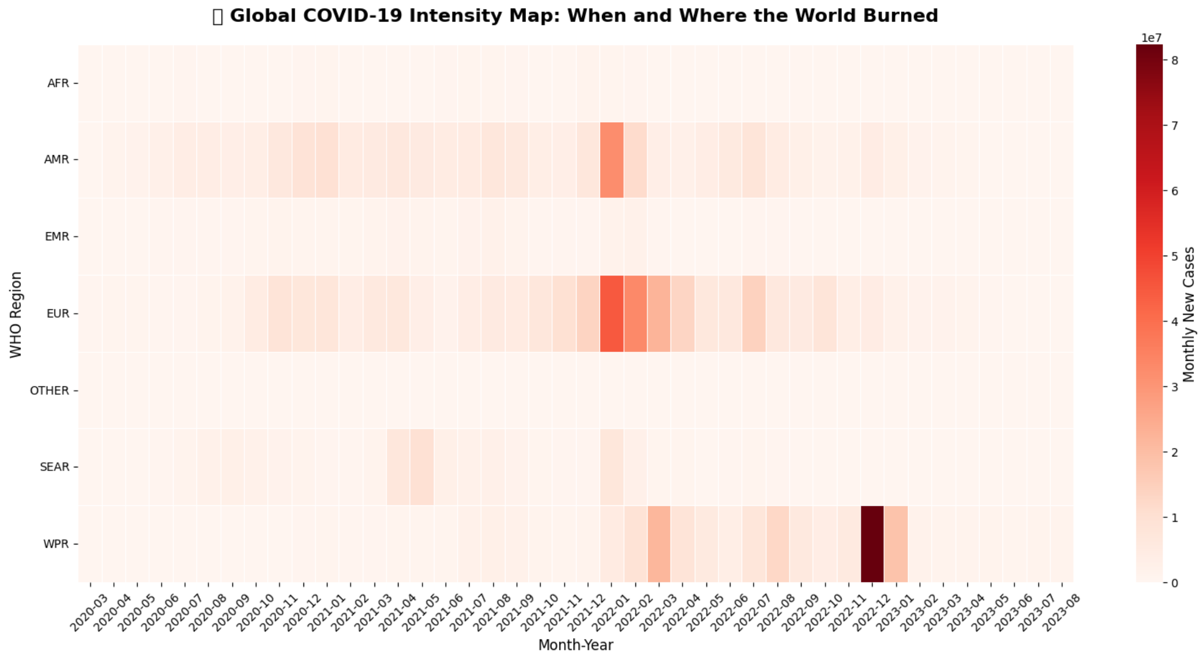
Heatmaps show spread intensity across time and geography.

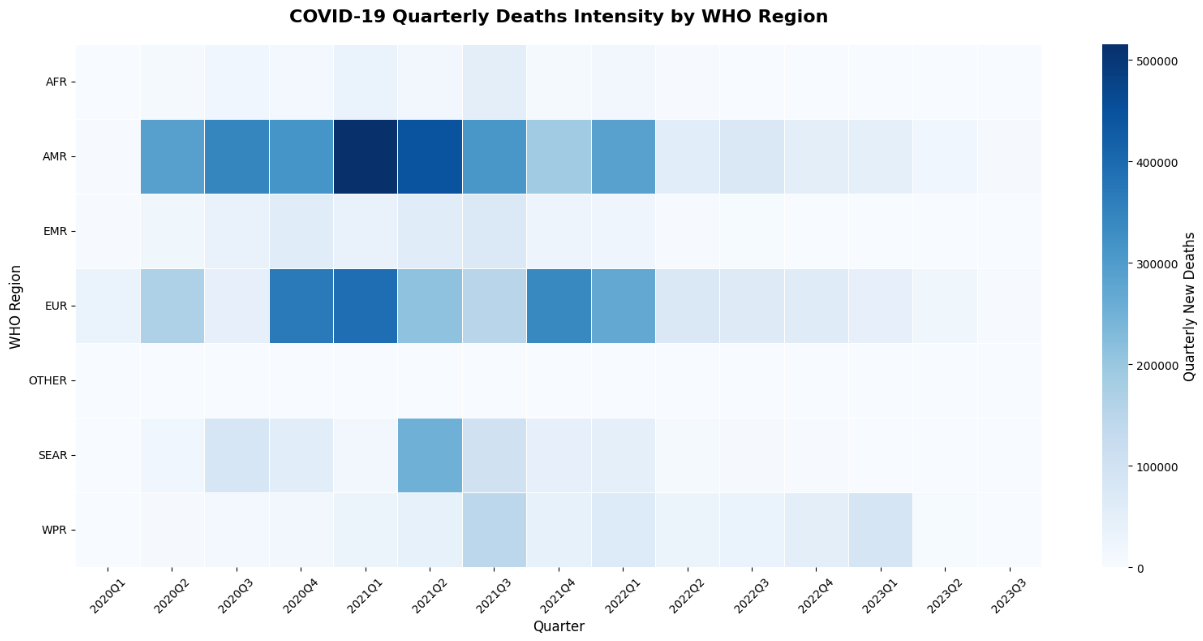
## EBOLA



The Ebola heatmap highlights the intensity of cases and deaths across affected countries. The darkest shades are concentrated in Guinea, Sierra Leone, and Liberia, confirming these nations as the epicentre of the 2014–2016 outbreak. Neighbouring regions show lighter shades, reflecting lower spillover, while most countries outside West Africa have negligible or no recorded cases.

## COVID-19

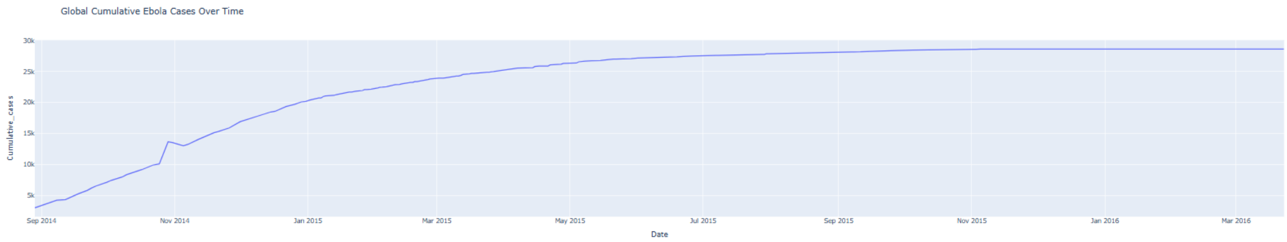


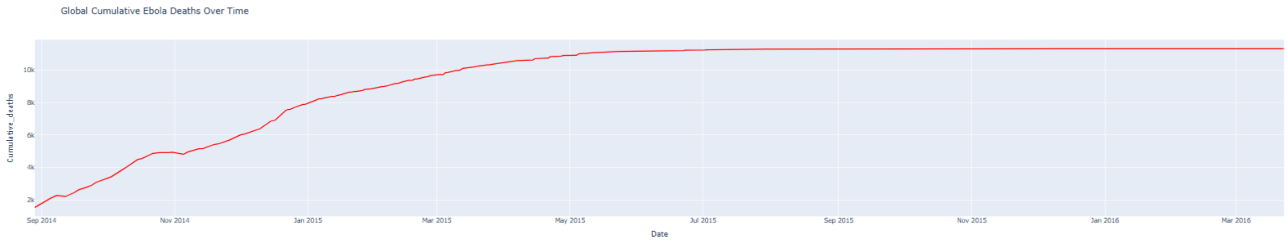


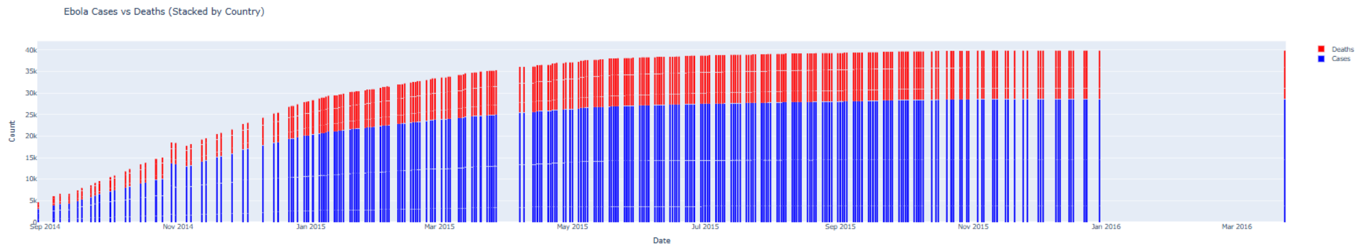
In contrast, the COVID-19 heatmap displays a far broader and denser distribution of cases and deaths worldwide. Intense colours are visible across North and South America, Europe, South Asia, and parts of Africa, illustrating the pandemic's global penetration.

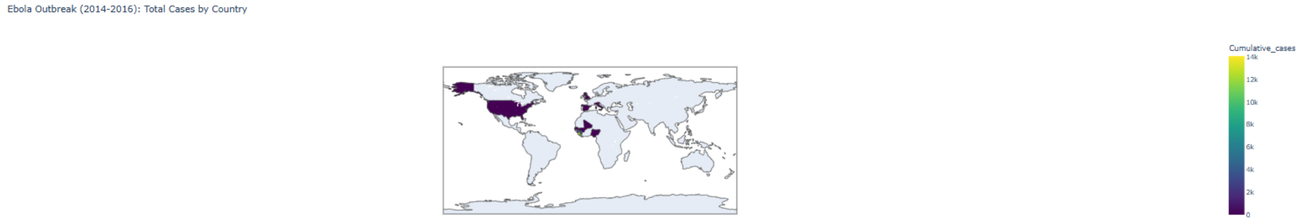
## Interactive plots

## EBOLA

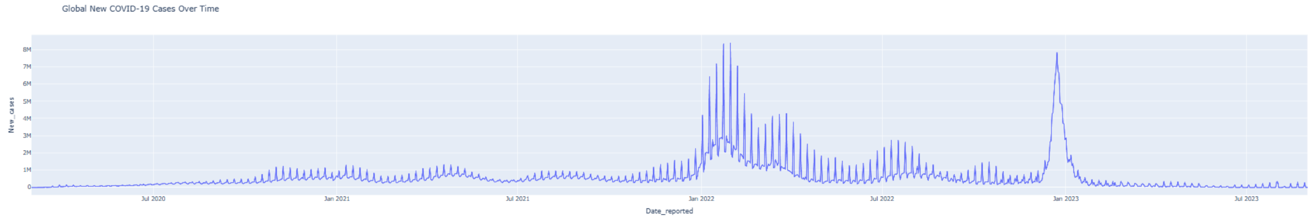


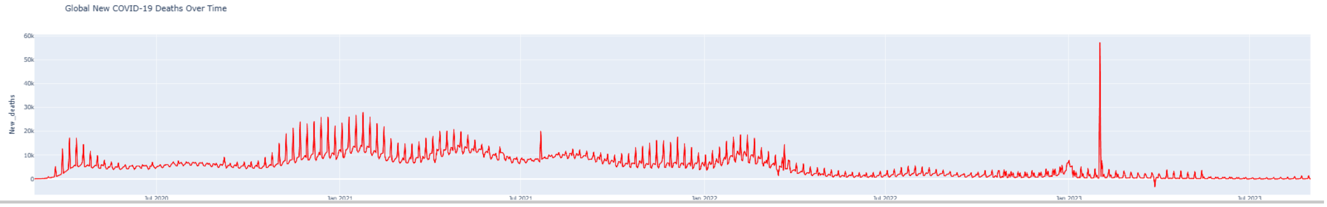


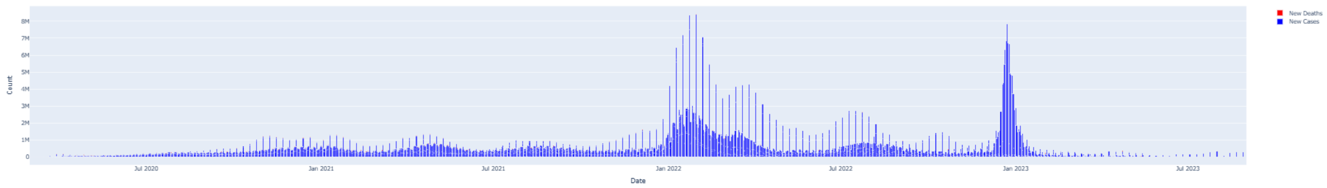

Ebola Cases vs Deaths (Stacked by Country)

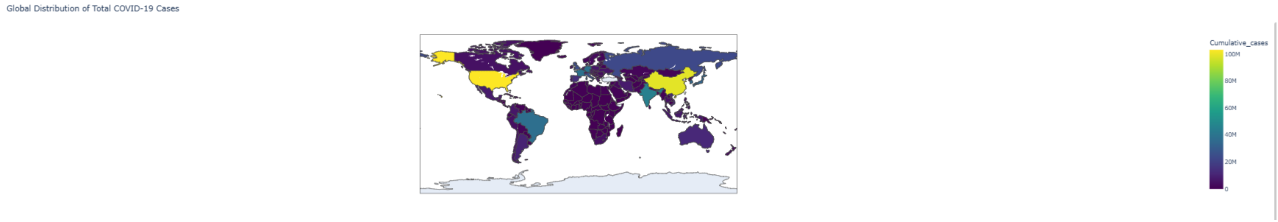

Ebola Outbreak (2014-2016): Total Cases by Country

# COVID-19


Global New COVID-19 Cases Over Time


Global New COVID-19 Deaths Over Time




Global Distribution of Total COVID-19 Cases

# 6. CONCLUSION

The comparative analysis of the Ebola outbreak (2014–2016) and the COVID-19 pandemic (2020–2023) has provided valuable insights into the contrasting dynamics of an epidemic versus a global pandemic. Through the use of exploratory data analysis, visualization techniques such as line charts, bar charts, pie charts, stacked and double bar charts, heatmaps, and advanced time-series plots, the study revealed how the two diseases spread differently in terms of geography, intensity, and impact.

The Ebola outbreak, as seen in the mountain-shaped curve, showed a sharp but regionally confined rise and fall in cases, concentrated mainly in Sierra Leone, Liberia, and Guinea. Visualizations such as the pie chart and heatmap confirmed that over 98% of deaths were restricted to these three nations, while other countries reported minimal impact. This indicated Ebola's high fatality but limited spread, heavily burdening local health systems but not escalating into a global crisis.

In contrast, COVID-19 displayed a prolonged and globally distributed curve, with multiple waves of rising and falling cases across years. The line and stacked bar charts highlighted how heavily populated nations like the United States, India, and Brazil consistently reported the largest caseloads and fatalities. The pie chart showed a more balanced distribution across many countries, emphasizing the truly global nature of the pandemic. Heatmaps further reinforced this, showing dense hotspots across almost all continents, underscoring COVID-19's unprecedented worldwide reach.

From this analysis, it can be concluded that while Ebola was intense but localized, COVID-19 was persistent and universal, affecting every aspect of global health, economy, and society. The visual comparisons between the two diseases not only demonstrate the role of public health interventions and preparedness but also highlight the importance of timely detection, international cooperation, and data-driven decision-making in containing future outbreaks.

Looking ahead, such comparative studies serve as critical references for strengthening global surveillance systems, enhancing healthcare infrastructure, and leveraging data visualization to communicate risks effectively. Future work could expand on this analysis by incorporating socio-economic indicators, vaccination coverage, and real-time mobility data to better understand the multifaceted impact of pandemics and to prepare more resilient responses for the future.

# 7. APPENDICES

## Appendix A: References

- WHO Ebola Situation Reports (2014–2016)
- Johns Hopkins University COVID-19 Data Repository (2020–2023)
- Kaggle Dataset: Ebola Outbreak Data (2014–2016)
- Kaggle Dataset: COVID-19 Global Time Series Data
- Scikit-learn Documentation: https://scikit-learn.org
- Pandas Documentation: https://pandas.pydata.org
- Matplotlib Documentation: https://matplotlib.org
- Seaborn Documentation: https://seaborn.pydata.org
- Plotly Documentation: https://plotly.com/python
- World Health Organization (WHO). (2016). Ebola Virus Disease – Key Facts.
- World Health Organization (WHO). (2021). COVID-19 Weekly Epidemiological Update.

## Appendix B: GitHub Repository

- Project Code Repository: Link

## Appendix C: Additional Documents

- Dataset links:
  - Ebola Dataset – Data