{AUTUMN INTERNSHIP PROJECT REPORT}

Comparative Data Analysis and Visualization of COVID-19 and Ebola Outbreaks(Notebook - 6)

Dip Kumar Majumder,

BS in IT(Data Science)/2023-27 and MAKAUT(Inhouse Campus)

Period of Internship: 25th August 2025 - 19th September 2025 (Do not change the dates)

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

1. Abstract

This project analyzes and visualizes data from two significant global health crises: the COVID-19 pandemic and the 2014-2016 Western Africa Ebola outbreak. Utilizing Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Plotly, the project explores key trends, geographical distribution, and temporal patterns of these epidemics. The analysis includes identifying the most and least affected countries, visualizing daily and cumulative case and death counts, and creating heatmaps to show regional and country-specific intensity over time. Interactive visualizations using Plotly provide a dynamic way to explore the data. The project demonstrates the power of data visualization in transforming raw numbers into understandable insights, aiding in comprehending the impact and spread of infectious diseases.

2. Introduction

The COVID-19 pandemic and the 2014-2016 Ebola outbreak were two of the most impactful global health events in recent history. Understanding the dynamics of such epidemics is crucial for public health response and preparedness. This project delves into datasets from both outbreaks to illustrate how data analysis and visualization can provide valuable insights into their spread, severity, and impact across different regions and over time.

The project utilizes Python as the primary programming language, leveraging its rich ecosystem of data science and visualization libraries. The procedure involves loading, cleaning, subsetting, and transforming the data to prepare it for analysis. Various visualization techniques are then applied to highlight key trends and patterns.

During the first two weeks of the internship, training was received on the following topics:

- Python Basics(Data, Variable, Lists, Loop, Data Structures, Class, Functions, OOPs)
- Python Libraries(Numpy, Pandas)
- Machine Learning Overview
- Regression Lab
- Classification Lab
- LLM Fundamentals
- Communication Skills

3. Project Objective

The objectives of this project are to:

- Load and preprocess epidemic datasets (COVID-19 and Ebola).
- Subset the datasets to focus on relevant information and time periods.
- Perform exploratory data analysis to understand key statistics and trends.
- Visualize the data to illustrate the spread and impact of the diseases.
- Identify and compare the most and least affected regions/countries.

- Create interactive visualizations for enhanced data exploration.

4. Methodology

The methodology employed in this project involved the following steps:

1. **Data Loading:** The COVID-19 dataset was loaded directly from a publicly shared Google Drive link into a pandas DataFrame. Similarly, the Ebola dataset was loaded from another Google Drive link.
2. **Data Subsetting:** Relevant columns such as 'Date_reported', 'Country', 'WHO_region', 'New_cases', 'New_deaths', 'Cumulative_cases', and 'Cumulative_deaths' were selected for the COVID-19 dataset. For the Ebola dataset, 'Country', 'Date', 'Cumulative no. of confirmed, probable and suspected cases', and 'Cumulative no. of confirmed, probable and suspected deaths' were selected and renamed to 'Cumulative_cases' and 'Cumulative_deaths'.
3. **Date Trimming (COVID-19):** The COVID-19 dataset was filtered to include data within a specific date range (2020-03-01 to 2023-08-31) to focus on the main period of the pandemic.
4. **Data Cleaning and Preparation (Ebola):** The 'Date' column in the Ebola dataset was converted to datetime objects. The data was sorted by 'Country' and 'Date'. Daily new cases and deaths were calculated by taking the difference of cumulative values within each country group. Negative values resulting from reporting inconsistencies were replaced with 0.
5. **Exploratory Data Analysis (EDA) and Visualization (Matplotlib/Seaborn):**
    - Line plots were generated using Matplotlib to show daily new cases for the top 5 most and least affected countries by cumulative cases for COVID-19.
    - A line plot was created to visualize daily global new COVID-19 cases, showing the overall trend.
    - Quarterly new COVID-19 cases and deaths were aggregated and visualized using both stacked and double bar charts with Matplotlib to compare the impact over time.
    - A pie chart was constructed using Matplotlib to show the proportion of cumulative COVID-19 deaths among the top 10 most affected countries.
    - Heatmaps were created using Seaborn to visualize the intensity of monthly new COVID-19 cases by WHO region and monthly new cases by the top 10 affected countries.
    - For the Ebola dataset, a line plot showed daily new cases for the top 5 most affected countries by cumulative cases.
    - A double bar chart visualized cumulative Ebola cases and deaths for the top 5 most affected countries.
    - A heatmap showed monthly new Ebola case intensity by country.
    - A pie chart displayed the proportion of cumulative Ebola deaths among the top 4 most affected countries.
6. **Interactive Visualization (Plotly):**

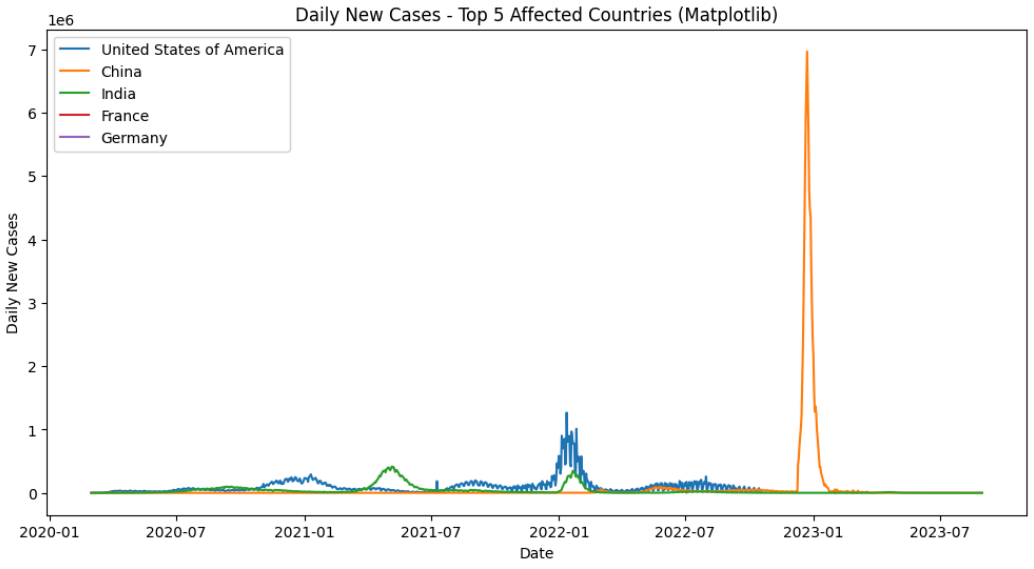
- Line charts were generated using Plotly to show global daily new COVID-19 cases and deaths over time, allowing for interactive exploration.
- A stacked bar chart using Plotly showed new COVID-19 cases vs. new deaths grouped by WHO region over time.
- A Choropleth map using Plotly visualized the global distribution of total cumulative COVID-19 cases by country.
- For the Ebola dataset, line charts using Plotly showed daily new cases and deaths over time globally.
- A Choropleth map using Plotly visualized the global distribution of total cumulative Ebola cases by country.

The code for this project is hosted on GitHub at: [Insert GitHub Link Here]

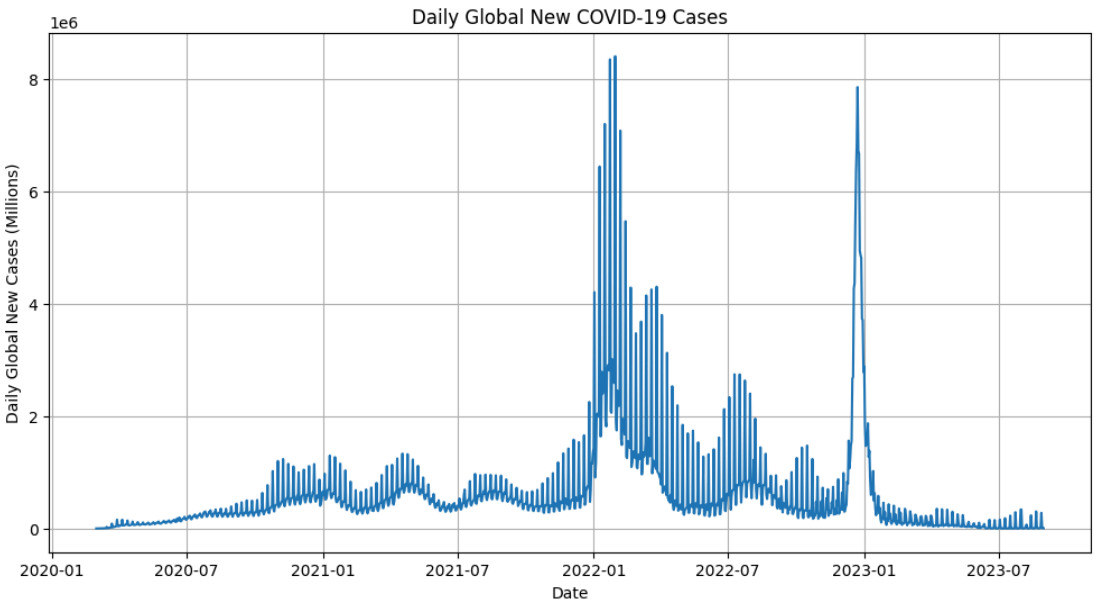
5. Data Analysis and Results

COVID-19 Analysis:

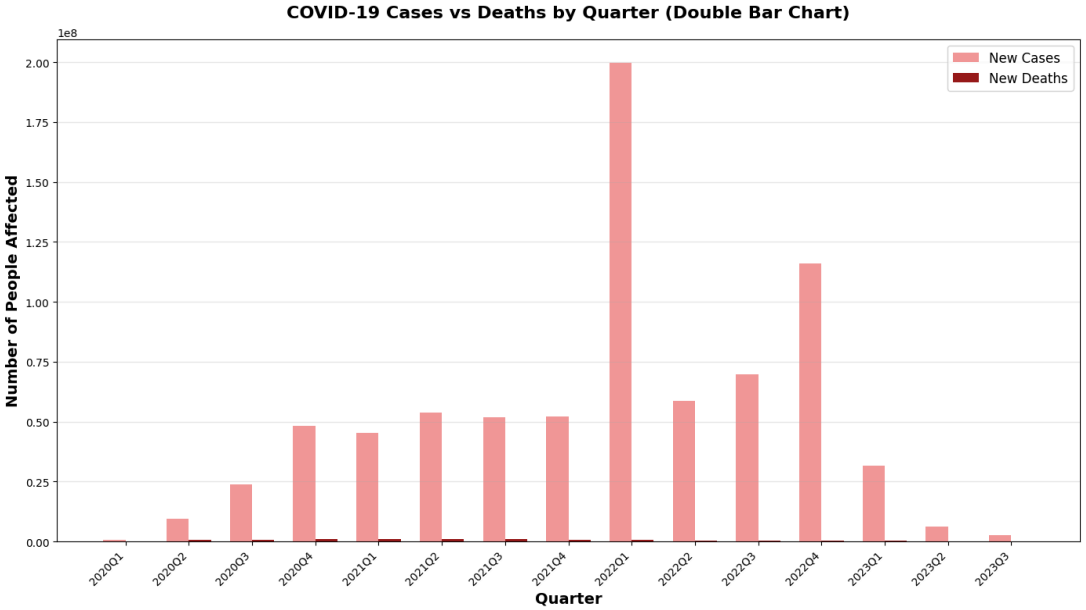
- **Top/Least Affected Countries (Line Plots):** The line plots for the top 5 most affected countries (United States of America, China, India, France, Germany) clearly show significant waves of daily new cases at different times. The plot for the least affected countries (Democratic People's Republic of Korea, Turkmenistan, International conveyance (Kiribati), International conveyance (American Samoa), Pitcairn) shows very low or no reported cases, highlighting the vast disparity in the pandemic's impact.



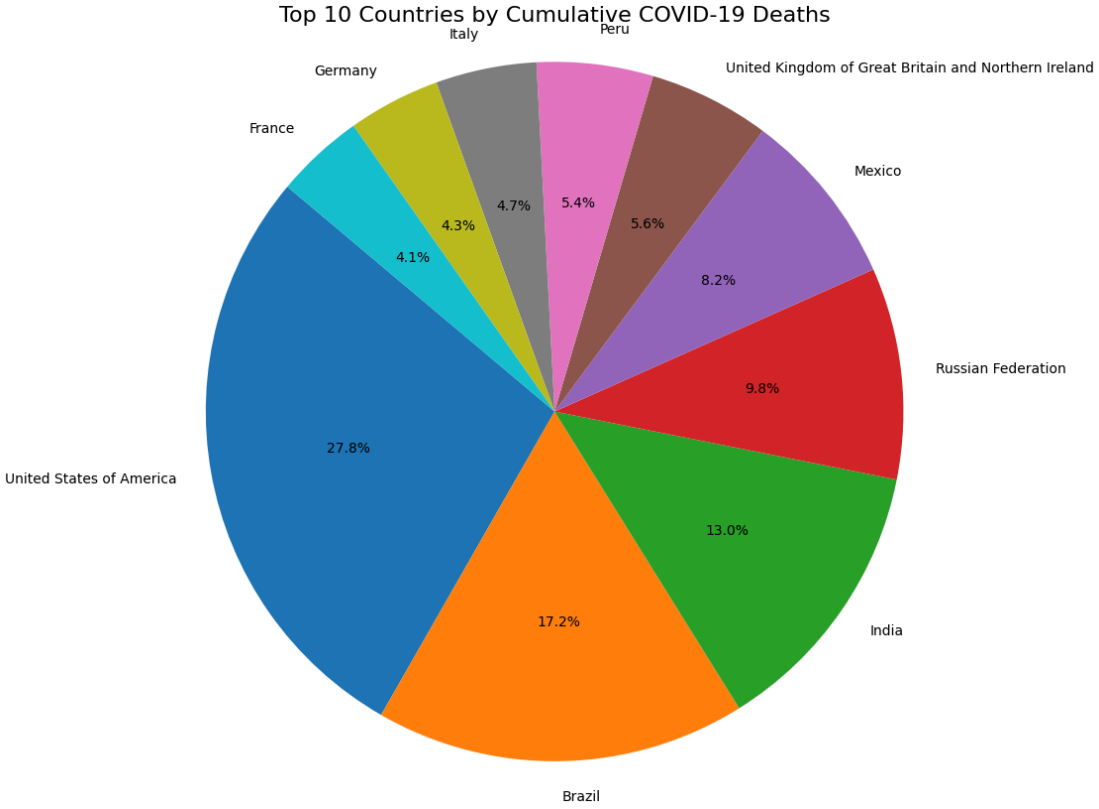
- **Global Daily Cases (Line Plot):** The global daily new cases plot shows distinct peaks, illustrating the different waves of the pandemic worldwide.



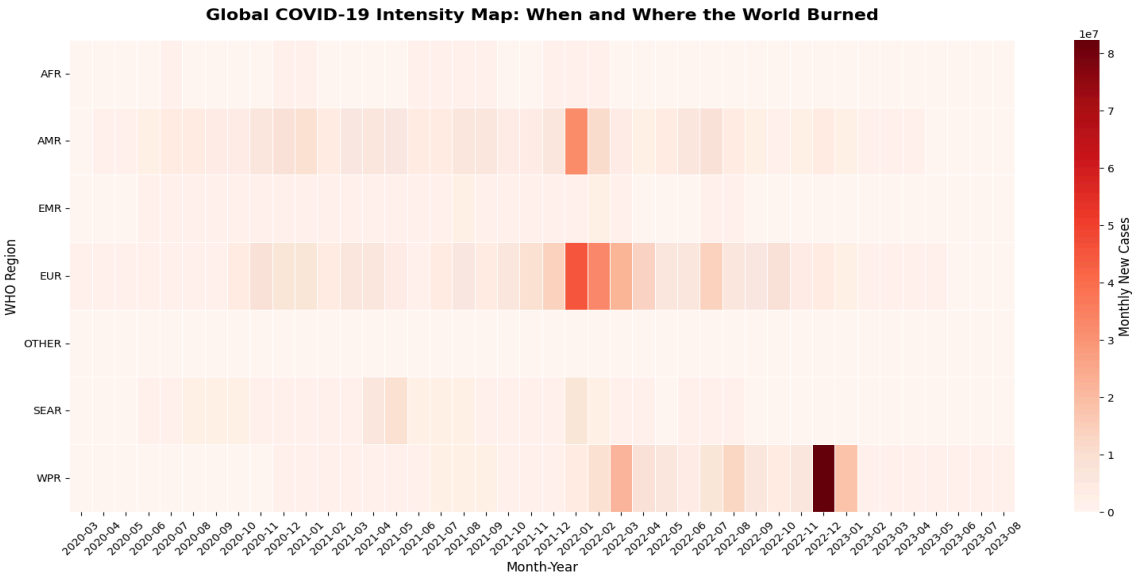
- **Cases vs Deaths by Quarter (Bar Charts):** Both the stacked and double bar charts effectively demonstrate the quarterly trends in new cases and deaths. The stacked chart shows the total impact, while the double bar chart allows for direct comparison of cases and deaths within each quarter. The peak in cases is clearly visible in early 2022.



- **Top 10 Countries by Cumulative Deaths (Pie Chart):** The pie chart highlights that a large percentage of cumulative COVID-19 deaths were concentrated in a few countries, with the United States of America, Brazil, and India accounting for a significant portion.

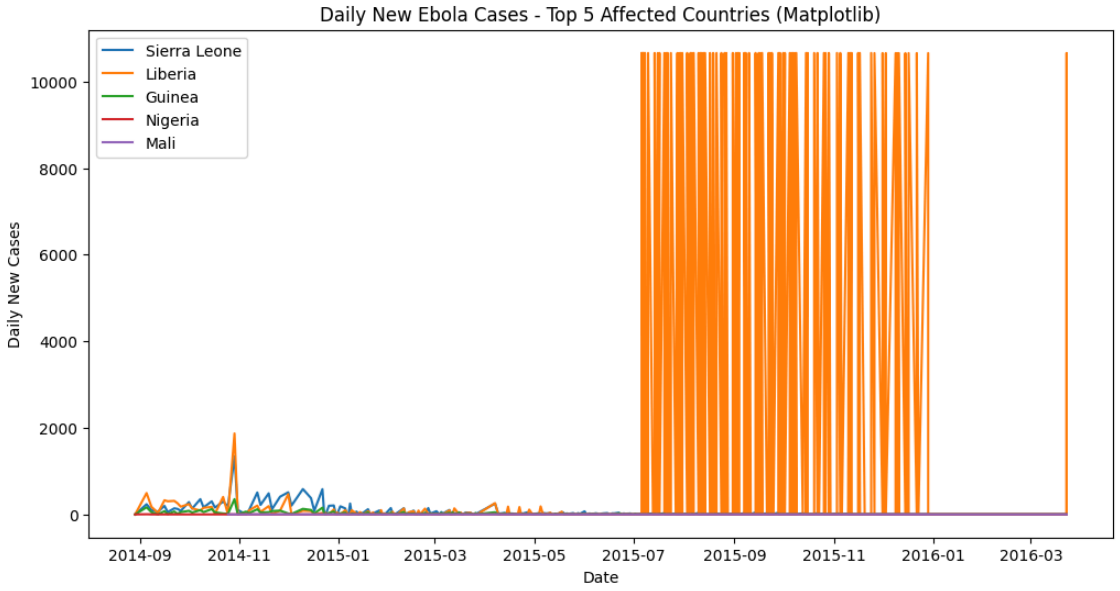


- **Heatmaps (Regional and Country):** The regional heatmap shows when and where case surges were most intense across different WHO regions, with AMR (Americas) and EUR (Europe) showing high intensity at various periods. The country-level heatmap for the top 10 affected countries provides a more granular view, pinpointing specific countries and months with high case loads.

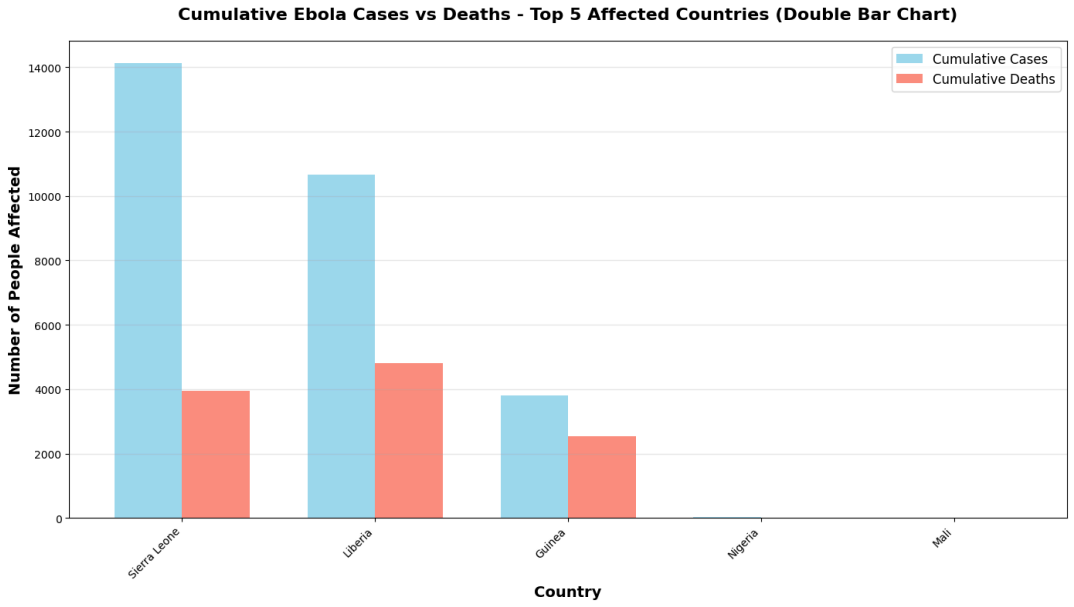


Ebola Analysis:

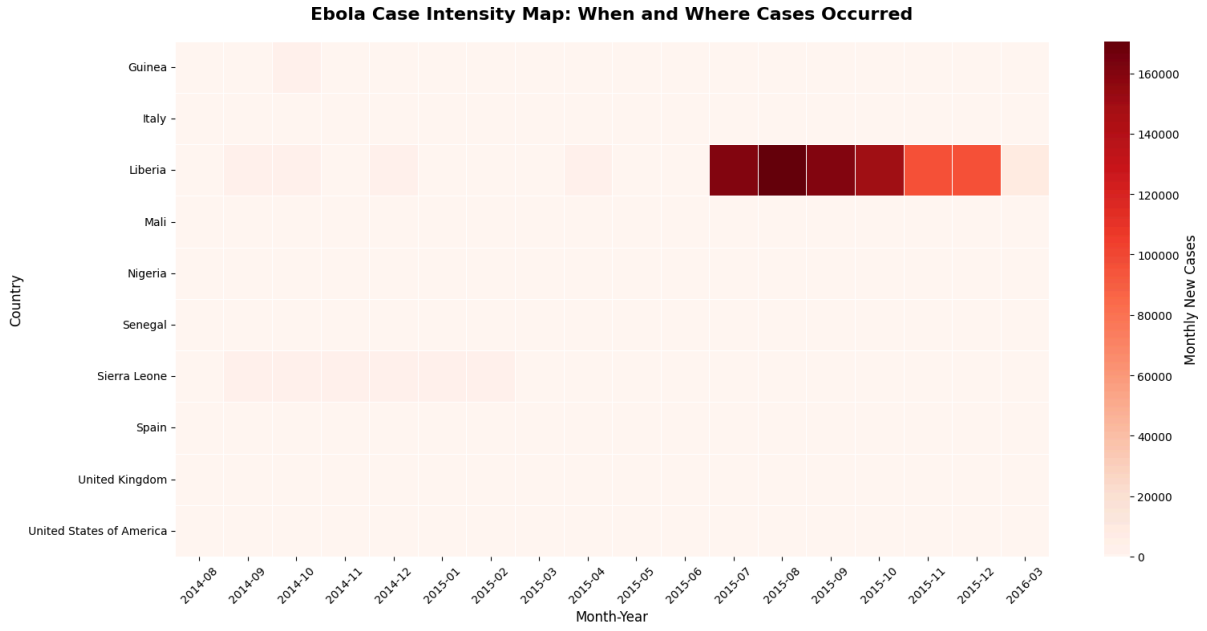
- **Top 5 Affected Countries (Line Plot):** The line plot for daily new Ebola cases in the top 5 affected countries (Sierra Leone, Liberia, Guinea, Nigeria, Mali) shows that the primary outbreak was concentrated in Sierra Leone, Liberia, and Guinea, with distinct peaks in case reporting. Nigeria and Mali had much lower case counts.



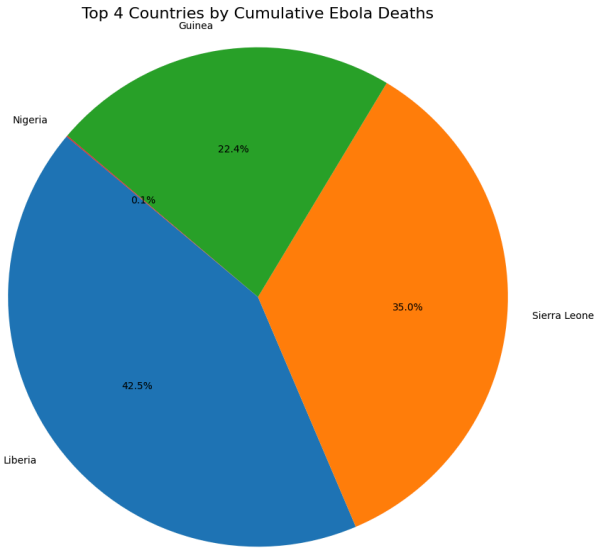
- **Cumulative Cases vs Deaths (Double Bar Chart):** The double bar chart clearly shows that Sierra Leone and Liberia had the highest cumulative cases and deaths during the Ebola outbreak, followed by Guinea. This reinforces the geographical concentration of the epidemic.



- **Monthly Case Intensity (Heatmap):** The Ebola heatmap illustrates the temporal and geographical spread of new cases. It clearly shows that the outbreak was most intense in Liberia, Sierra Leone, and Guinea during specific periods in 2014 and 2015. Other countries had minimal or no reported cases.



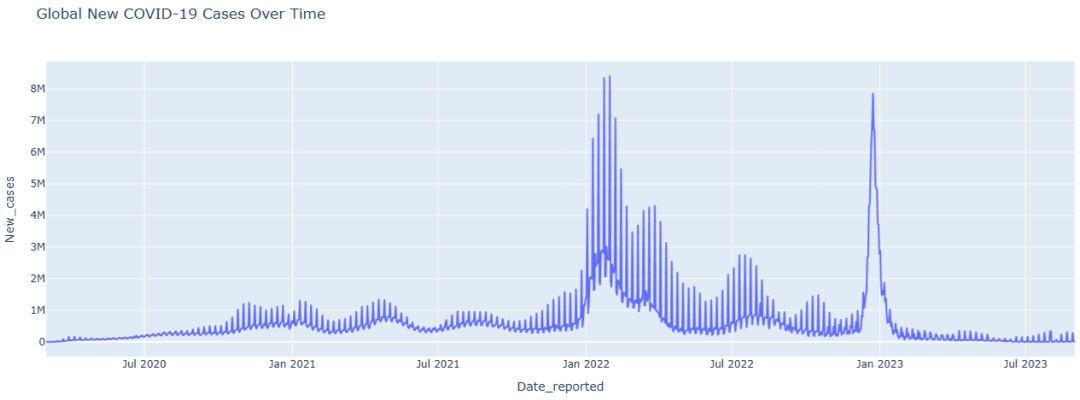
- **Top 4 Countries by Cumulative Deaths (Pie Chart):** The pie chart demonstrates that the vast majority of cumulative Ebola deaths occurred in Liberia, Sierra Leone, and Guinea, underscoring the severe impact on these West African nations.



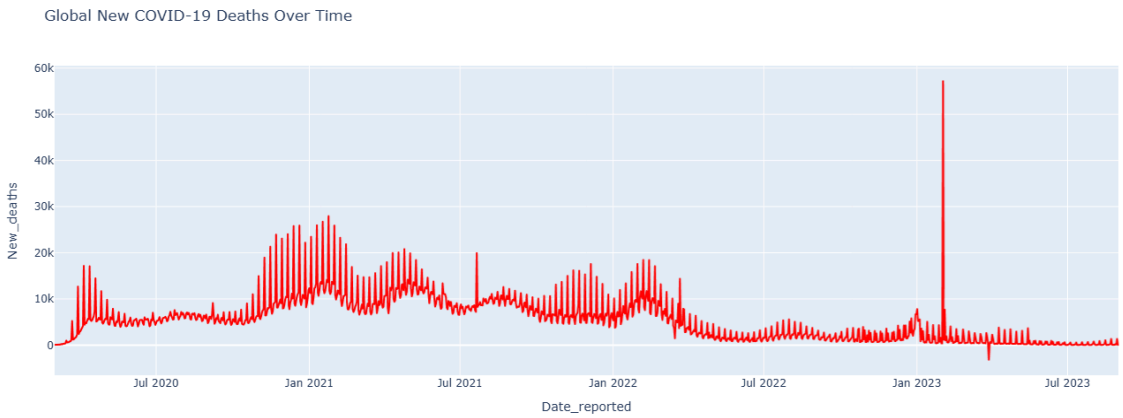
Interactive Visualizations (Plotly):

Plotly was utilized to create interactive visualizations, allowing for dynamic exploration of the data. These charts provide enhanced capabilities for identifying trends and patterns compared to static plots.

- **Line Chart – Global Daily New COVID-19 Cases Over Time:** This interactive line plot displays the total number of new COVID-19 cases reported globally each day. Users can zoom in and pan across the timeline to observe the peaks and valleys of the pandemic waves, gaining insights into the temporal progression of the outbreak worldwide. Hovering over data points reveals the exact date and number of cases.

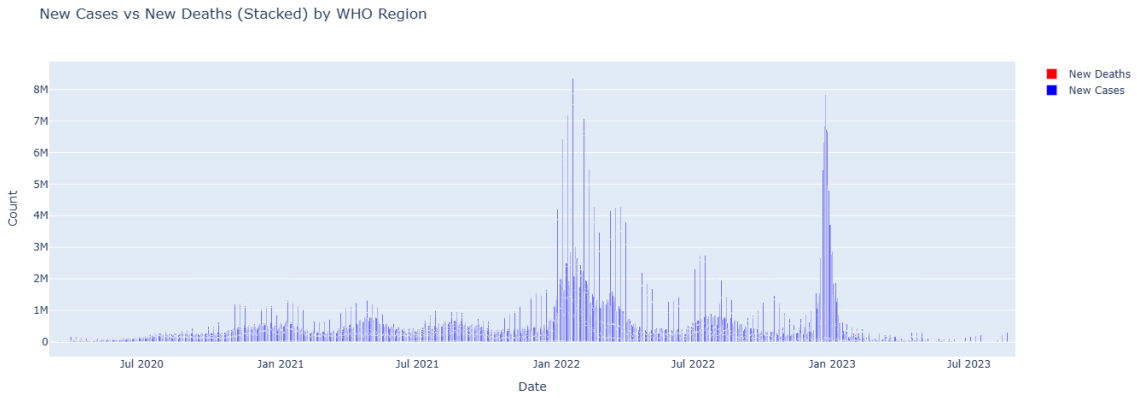


- **Line Chart – Global Daily New COVID-19 Deaths Over Time:** Similar to the cases chart, this line plot visualizes the total number of new COVID-19 deaths reported globally per day. Plotted in red for clear distinction, it allows for easy identification of periods with high mortality rates. The interactive features enable close examination of death trends over specific timeframes.

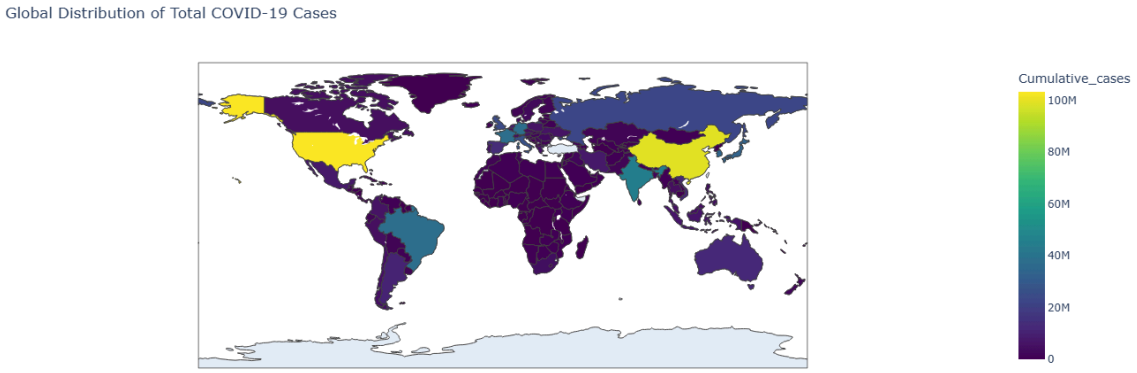


- **Stacked Bar Chart – New COVID-19 Cases vs New Deaths by WHO Region:** This interactive stacked bar chart shows the daily new cases and new deaths, grouped and stacked by WHO region over time. It provides a dynamic view of how the pandemic's impact varied across different regions and how the proportion of cases to

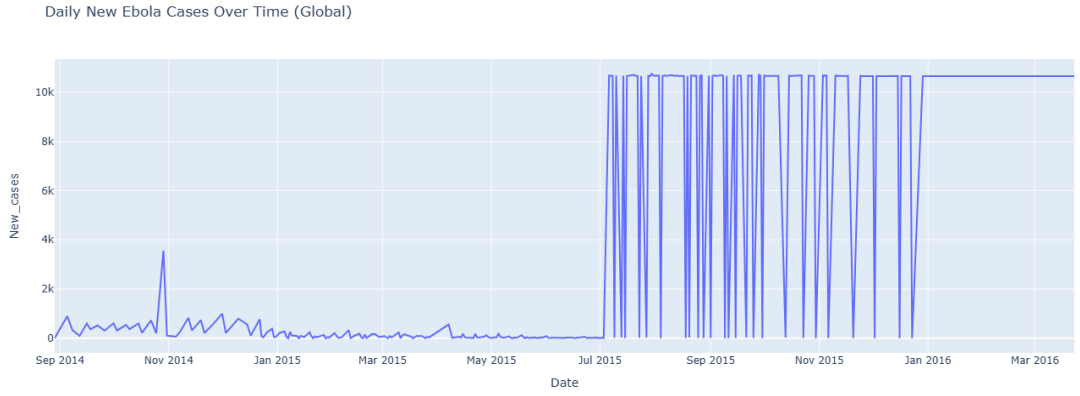
deaths changed over time within each region. Hovering provides details on case and death counts for each region on a given date.



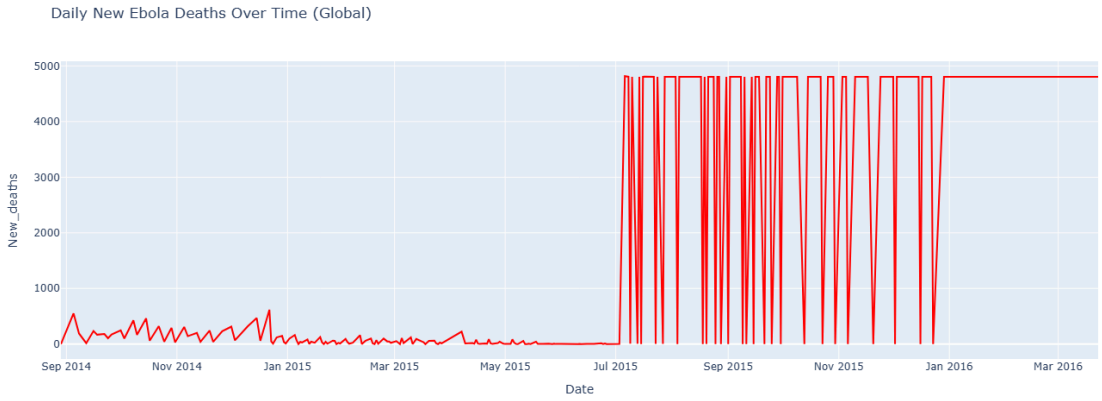
- **Choropleth Map – Global Distribution of Total COVID-19 Cases:** This interactive world map visualizes the cumulative COVID-19 cases by country. The color intensity of each country represents its total case count, allowing for immediate visual comparison of the pandemic's severity across different nations. Users can hover over a country to see its name and the exact number of cumulative cases.



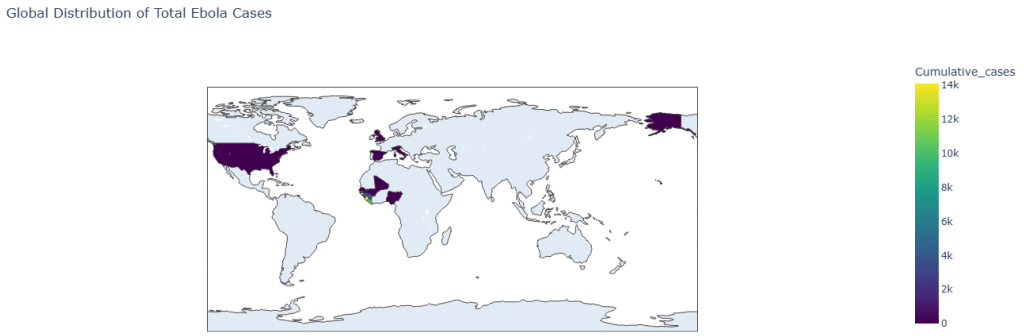
- **Line Chart – Daily New Ebola Cases Over Time (Global):** This interactive line chart displays the daily reported new cases for the Ebola outbreak globally. It allows users to observe the overall trend of the Ebola epidemic, identifying key periods of increased transmission. The interactive features help in examining the timeline of the outbreak's progression.



- **Line Chart – Daily New Ebola Deaths Over Time (Global):** This line chart shows the daily reported new deaths for the Ebola outbreak globally. Similar to the COVID-19 deaths chart, it highlights the mortality trend of the Ebola epidemic over time. The red color signifies deaths, and interactivity allows for detailed inspection of the death counts on specific dates.



- **Choropleth Map – Global Distribution of Total Ebola Cases:** This interactive world map visualizes the cumulative Ebola cases by country. The color intensity indicates the total case count in each country, clearly showing the geographical concentration of the outbreak, primarily in West Africa. Hovering over a country displays its name and the cumulative case count.



6. Conclusion

This project successfully demonstrates the application of data analysis and visualization techniques to understand the dynamics of the COVID-19 pandemic and the 2014-2016 Ebola outbreak. By visualizing key metrics like daily new cases, deaths, and cumulative totals across different regions and countries, we were able to identify trends, hotspots, and the overall impact of these epidemics. The use of both static and interactive visualizations provided a comprehensive view of the data, allowing for both detailed analysis and broad overviews. The project highlights how data-driven approaches are essential for monitoring, understanding, and responding to public health crises.

Recommendations

- Incorporate more advanced time-series analysis techniques (e.g., forecasting models) to predict future trends.
- Explore the relationship between various factors (e.g., population density, healthcare infrastructure intervention measures) and the spread and severity of the outbreaks.
- Develop a more comprehensive interactive dashboard with additional filtering and drill-down capabilities.

7. APPENDICES

1. References

a. COVID-19 Data Source: [https://drive.google.com/uc?export=download&id=1Sj3Il94NXun9owedSWNGrxszjpAXTDEQ]
b. Ebola Data Source: [https://www.kaggle.com/datasets/imdevskp/ebola-outbreak-20142016-complete-dataset]
c. Pandas Documentation: [https://pandas.pydata.org/]
d. NumPy Documentation: [https://numpy.org/]
e. Matplotlib Documentation: [https://matplotlib.org/]
f. Seaborn Documentation: [https://seaborn.pydata.org/]
g. Plotly Documentation: [https://plotly.com/python/]

2. Colab Notebook Link:

Visualizing_Time_Series_Dataset_COVID_19_Data_V2 _Dip.ipynb