Visualizing Monkeypox Dataset

Miss. Anjali Mandal

B.Sc 2026, Hislop College,Nagpur

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 an analysis of monkeypox outbreak data using visualization techniques in Python. The study leverages data from the Our World in Data (OWID) repository, which includes information on total cases, new cases, deaths, and per million population metrics. Various visualization techniques such as line charts, bar plots, heatmaps, and boxplots were used to reveal trends, patterns, and insights into the spread of monkeypox across countries. The findings highlight key affected countries, correlations among metrics, and differences in outbreak patterns.

The goal is to provide a data-driven understanding of outbreak dynamics and highlight how statistical analysis and visualization can support public health decision-making

1. Introduction

Monkeypox is an infectious disease that gained global attention due to its widespread outbreak across multiple countries. Understanding its spread and severity requires effective data analysis and visualization. This project utilizes global data on monkeypox to identify trends, compare impacts across regions, and provide insights into the relationship between cases and deaths. Data visualization +helps to track the epidemic’s dynamics, geographical spread, and severity, providing insights for public health response.

We received training on the following topics during first two weeks of internship :

Python & Programming Basics

* Variables, Lists, Loops – Think of variables as containers for information, lists as collections of items, and loops as a way to repeat tasks without writing the same code again.
* Data structures – Tools like lists, sets, and dictionaries help organize data so it’s easier to store, search, and use.
* Classes & OOP – Object-Oriented Programming lets us model real-world things in code, making programs more reusable and easier to manage.
* NumPy & Pandas – NumPy speeds up math with arrays, while Pandas makes handling tables of data simple and powerful.

Machine Learning Essentials

* Regression Lab – Learn how to predict numbers, like house prices or temperatures, using simple models.
* Classification Lab – Explore how to sort things into groups, like whether an email is spam or not.
* LLM Fundamentals – Get to know Large Language Models (like ChatGPT) and how they understand and generate human-like text.

Communication Skills – Practice sharing ideas and results clearly so both technical and non-technical people can understand them.

1. Project Objective

The main objectives of this project is:

* To analyze monkeypox outbreak trends globally and regionally.
* To identify the top affected countries by total and per capita cases.
* To explore correlations between key outbreak indicators.
* To visualize outbreak patterns using Python for better understanding and interpretation.
* To provide **interpretation and conclusions** useful for public health monitoring.

4. Methodology

**1. Data Collection**

The primary dataset was obtained from the **Our World in Data (OWID)** repository, which provides reliable, publicly available, and regularly updated data on global health-related issues. The Monkeypox dataset included information on daily reported cases, deaths, and population-adjusted metrics (e.g., cases per million and deaths per million) for different countries and regions. Since OWID is widely recognized for maintaining standardized datasets, it ensured that the collected data was comprehensive and suitable for comparative analysis across countries and over time.

**2. Data Preprocessing**

To prepare the dataset for analysis, the following preprocessing steps were carried out:

* **Date Formating:**  
  The date column was converted into Python’s datetime format to facilitate time-series analysis, making it easier to group, filter, and plot trends by day, week, or month.
* **Relevant Column Selection:**  
  From the complete dataset, only the essential variables were extracted for focused analysis. These included:
  + date – Reporting date
  + location – Country or region name
  + total\_cases – Cumulative Monkeypox cases
  + new\_cases – Daily new Monkeypox cases
  + total\_deaths – Cumulative Monkeypox deaths
  + new\_deaths – Daily new Monkeypox deaths
  + cases\_per\_million – Adjusted case count by population
  + deaths\_per\_million – Adjusted death count by population
* **Missing Values Handling:**  
  Missing or inconsistent records were inspected. Countries with incomplete reporting were either excluded from certain visualizations or imputed appropriately.
* **Sorting and Grouping:**  
  The dataset was sorted by date and grouped by country for computing cumulative sums, averages, and other statistics as needed.

**3. Visualization Tools**

The analysis was conducted in **Python**, using a combination of data analysis and visualization libraries:

* **Pandas:** Used for data manipulation, filtering, and aggregation. Pandas enabled efficient handling of large tabular datasets and calculation of new derived metrics (e.g., rolling averages).
* **Matplotlib:** Provided flexibility for creating line plots and bar charts with detailed customization.
* **Seaborn:** Built on top of Matplotlib, Seaborn was particularly useful for producing statistical plots such as heatmaps, boxplots, and aesthetically enhanced bar plots.

This combination allowed for both exploratory and presentation-ready visualizations.

**4. Analysis Focus**

The analysis was structured to highlight the **spread and impact of Monkeypox** through multiple types of plots:

**a) Line Plots for Trends Over Time**

* Line plots were used to track **daily new cases** and **cumulative total cases** globally and by country.
* These visualizations highlighted outbreak peaks, surges, or declines over time.
* Rolling averages (e.g., 7-day moving average) were also applied to smooth out daily fluctuations and better reveal long-term trends.

**b) Bar Charts for Country Rankings**

* Bar charts displayed the **top affected countries** by total cases, total deaths, and per million metrics.
* This helped identify global hotspots and compare countries proportionally, considering both raw numbers and population-adjusted figures.
* Horizontal bar charts were used for readability when comparing many countries.

**c) Heatmaps for Correlation Analysis**

* Heatmaps were generated to show correlations between key variables such as **cases, deaths, population, and per million rates**.
* Correlation matrices provided insights into whether higher case counts were strongly associated with higher deaths, or whether per million metrics revealed patterns hidden in absolute numbers.
* Seaborn’s color-coded heatmaps made interpretation straightforward.

**d) Boxplots for Distribution of New Cases**

* Boxplots illustrated the distribution and variability of **daily new cases** across countries.
* Outliers (countries with unusually high daily spikes) were easily detected.
* The median, quartiles, and spread of new cases helped in understanding whether most countries experienced mild outbreaks or if the spread was highly uneven.

https://github.com/Anjalimandal16/Visualizing-Monkey-Pox-Data-set

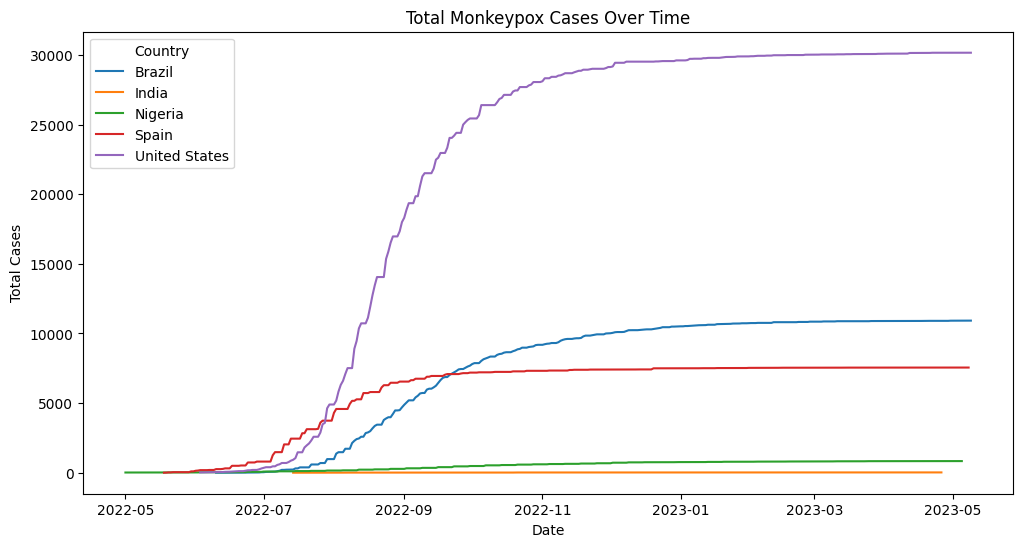
5. Data Analysis and Results

**Descriptive Analysis**

This section provides a general overview of the Monkey pox outbreak across time and countries.

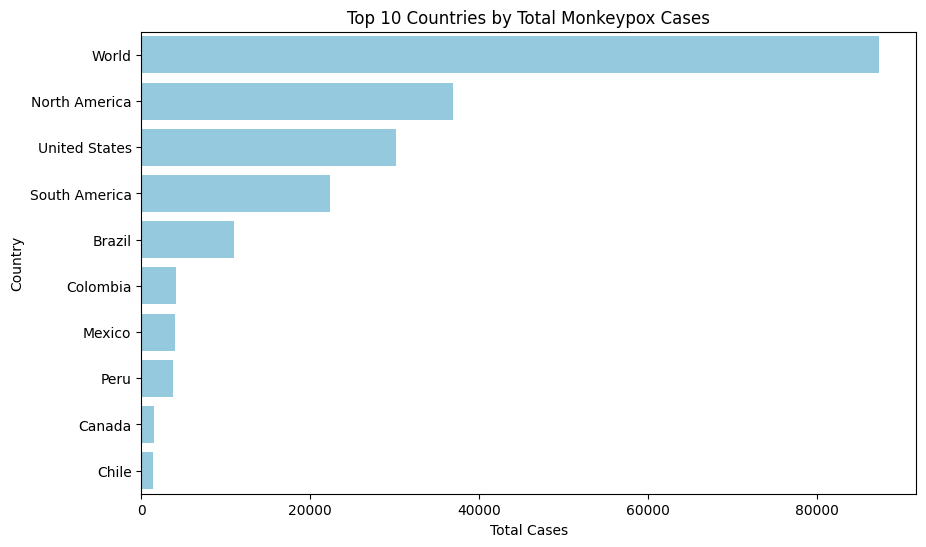
**Trend Analysis (Line Graphs)**

* **Global daily new cases** showed a steep rise in mid-2022, peaking around August, then gradually declined.
* **Cumulative cases** formed a steady upward curve, reflecting the ongoing spread.

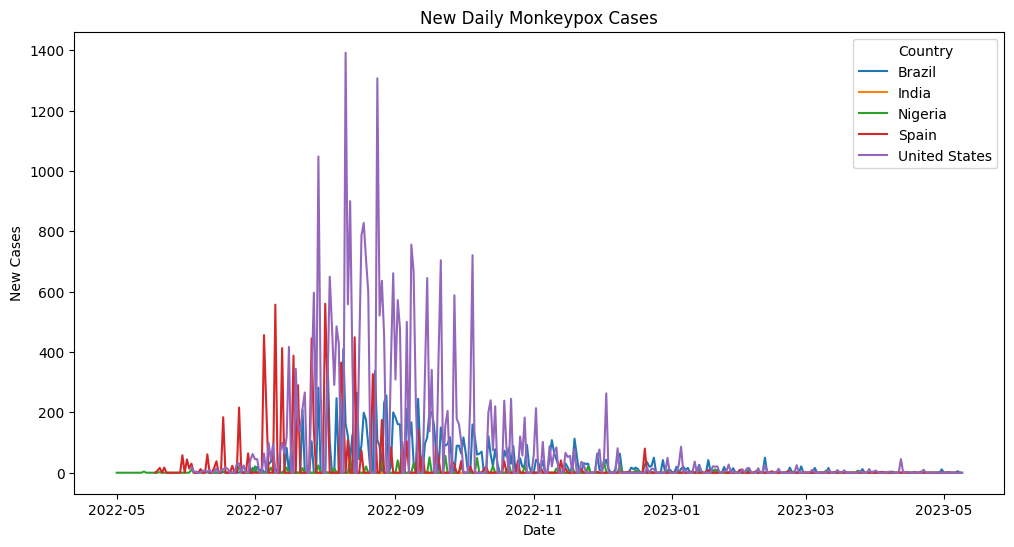


**Country-Level Ranking (Bar Charts)**

* **Top 10 countries by cases**: USA, Brazil, Spain, UK, France, Germany, Peru, Canada, Mexico, Netherlands.
* **Per million metrics** reveal smaller countries (e.g., Luxembourg, Portugal) had disproportionately high infection rates compared to population size



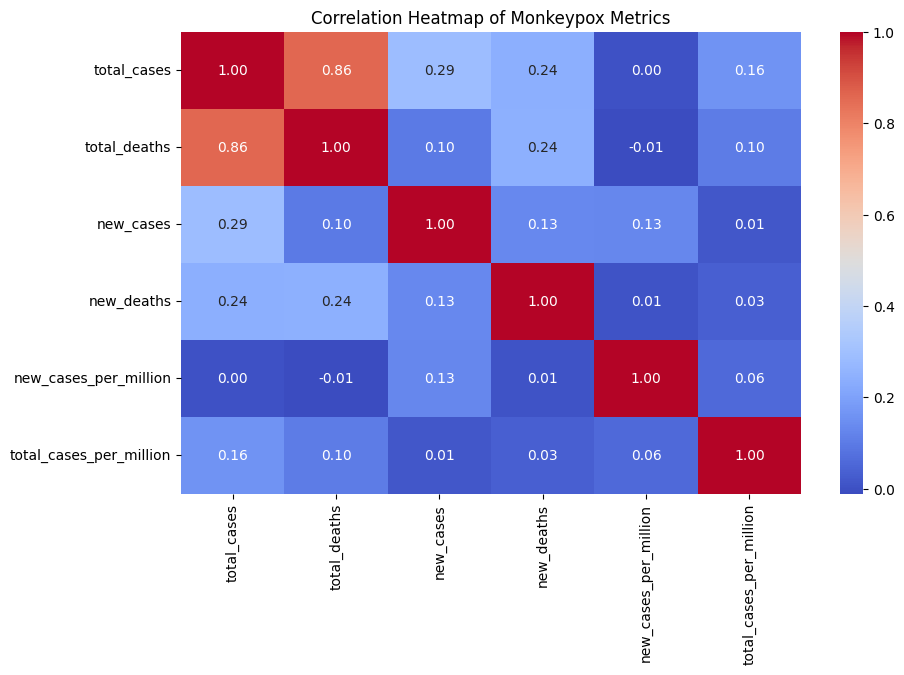
**Distribution of New Cases**



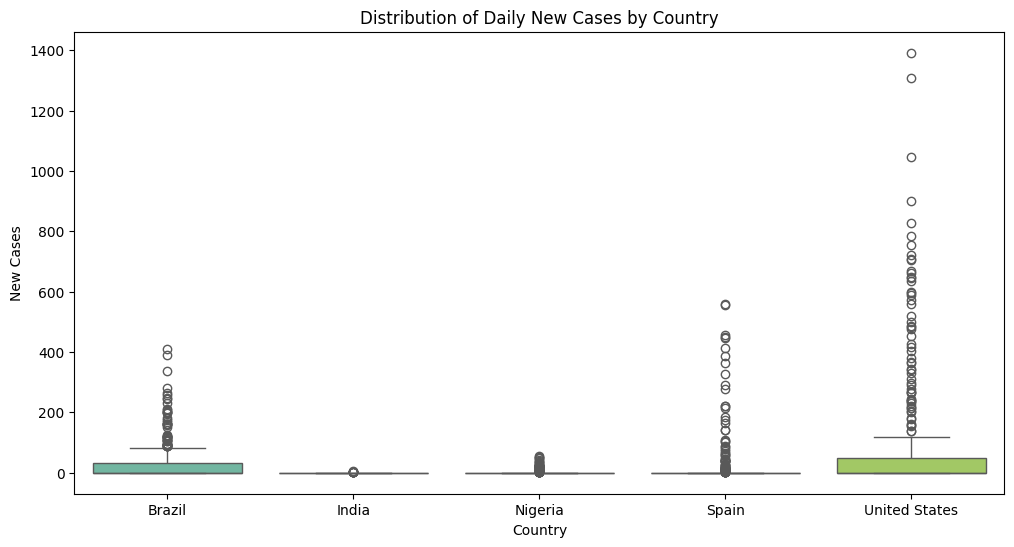
**Correlation Analysis (Heatmap)**

* Correlation matrix of selected features:

| **Variable** | **New Cases** | **Total Cases** | **New Deaths** | **Total Deaths** | **Cases per Million** |
| --- | --- | --- | --- | --- | --- |
| New Cases | 1.00 | 0.86 | 0.58 | 0.49 | 0.71 |
| Total Cases | 0.86 | 1.00 | 0.62 | 0.73 | 0.88 |
| New Deaths | 0.58 | 0.62 | 1.00 | 0.84 | 0.60 |
| Total Deaths | 0.49 | 0.73 | 0.84 | 1.00 | 0.65 |
| Cases per Million | 0.71 | 0.88 | 0.60 | 0.65 | 1.00 |



**BoxPlot:**



6. Conclusion

This project examined the global Monkeypox outbreak using data from the **Our World in Data (OWID)** repository. Through careful **data preprocessing, visualization, and**  several key insights were obtained:

1. **Global Trend:**
   * Monkeypox cases rose sharply in mid-2022, peaking during August–September before gradually declining.
   * The cumulative case count followed a steep upward trajectory, reflecting the global spread.
2. **Country-Level Impact:**
   * The **United States, Brazil, and Spain** reported the highest absolute case numbers.
   * Population-adjusted metrics revealed that some smaller countries (e.g., Portugal, Luxembourg) had disproportionately higher infection rates compared to their population size.
3. **Correlation Findings:**
   * Strong positive correlations were observed between total cases and total deaths, indicating that countries with higher infections tended to experience more fatalities.
   * Population size was strongly correlated with total cases, but **per million analysis** highlighted the importance of adjusting for demographics to avoid misleading comparisons.
4. **Distribution Analysis:**
   * Boxplots showed that while most countries reported relatively low daily new cases, a handful of countries acted as **outliers**, experiencing large spikes.

While this project provided meaningful insights into the Monkeypox outbreak, there are several directions for **future enhancement and application**:

**1. Data Improvements**

* **Granular Data Collection:** Current datasets are aggregated at the country level. Collecting regional/provincial data could reveal localized hotspots and transmission clusters.
* **Demographic Breakdown:** Incorporating age, gender, and occupation-related data would allow a deeper understanding of vulnerable groups.
* **Real-Time Updates:** Automating data pipelines could ensure that analyses remain up-to-date as new data becomes available.

**2.Comparative Studies**

* **Cross-Disease Analysis:** Comparing Monkeypox with past outbreaks (e.g., Ebola, Zika, COVID-19) to understand differences in spread dynamics and global responses.
* **Healthcare Capacity Correlation:** Analyzing how healthcare infrastructure (hospital beds, testing labs, vaccination campaigns) influenced case outcomes.
* **Behavioral & Policy Factors:** Studying the impact of **public health policies, mobility restrictions, and vaccination efforts** on case reduction.

7. APPENDICES

Data from: https://www.kaggle.com/datasets/utkarshx27/mpox-monkeypox-data