**Visualizing Time Series Dataset Of COVID-19**

Souhardya Mukherjee

Btech Biotechnology

Amity University 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 analyzes the global impact of the COVID-19 pandemic through comprehensive data visualization. Using datasets of confirmed cases and deaths, We applied Python libraries such as Pandas, Matplotlib, Seaborn, and Plotly to uncover key trends and patterns. Static visualizations included line plots, pie charts, and heat maps to highlight daily case surges, country-level death distributions, and regional temporal dynamics. An interactive dashboard built with Plotly provided flexibility to explore new cases, deaths, and cumulative trends across different countries and WHO regions.These visualizations not only enhance understanding of the scale and progression of COVID-19 but also demonstrate the value of data-driven approaches in monitoring global health crises.

1. **Introduction**

The COVID-19 pandemic has been one of the most significant global health crises in recent history, affecting millions of lives and disrupting societies worldwide. Understanding its progression, impact, and patterns requires not only reliable data but also effective visualization methods. This project was undertaken to analyze global COVID-19 datasets and transform raw data into meaningful insights through visualizations.

The project involved the use of Python as the primary programming language due to its extensive ecosystem of data analysis and visualization libraries. Key technologies included Pandas for data processing, Matplotlib and Seaborn for static visualizations, and Plotly for interactive dashboards. Choropleth maps and heat maps were also used to capture the spatial and temporal intensity of cases and deaths.

Before starting the main tasks, background material surveys were conducted to study existing COVID-19 dashboards developed by organizations such as the WHO, Johns Hopkins University, and Our World in Data. These references helped in framing the scope of the project and guided decisions on the types of plots and dashboards that would be most informative.

The procedure followed included:

1. Cleaning and preparing the dataset.

2. Aggregating data by different time intervals (daily, monthly, quarterly).

3. Creating static visualizations such as line plots, bar charts, and pie charts.

4. Developing comparative and distribution-based charts like stacked bars and heat maps.

5. Building an interactive COVID-19 dashboard with Plotly to allow flexible exploration of trends.

The purpose of this project is to demonstrate how data visualization can make large-scale health data easier to interpret, compare, and communicate. By using multiple visualization approaches, this work highlights the scale, intensity, and geographic spread of the pandemic while providing a platform for further exploration.

Training Topics During First Two Weeks of Internship

During the initial phase of the internship, we received training on a range of foundational and technical topics that equipped us to carry out this project effectively. These included:

1. Basic Python operations – covering data types, functions, and control structures.

2. Object-Oriented Programming (OOP) in Python – classes, objects, inheritance, and modular coding.

3. Machine Learning fundamentals – introduction to supervised and unsupervised learning.

4. Communication skills – effective technical writing, presentation skills, and teamwork strategies.

This training provided both the technical and professional foundation needed to execute the project successfully.

1. **Project Objective**

The main objectives of this project are:

* To analyze and visualize the global impact of COVID-19 by using reliable datasets of new cases and deaths, highlighting temporal and regional variations.
* To illustrate key patterns and trends such as daily surges, quarterly intensity, regional differences, and the distribution of deaths across the top affected countries.
* To develop interactive tools (using Plotly dashboards) that allow users to dynamically filter, compare, and explore the data from multiple perspectives.
* To evaluate the effectiveness of visualization in simplifying large-scale health data for better interpretation, communication, and decision-making.

***No hypothesis testing or sample surveys were performed in this project, as the analysis was entirely based on secondary datasets***.

1. **Methodology**

The methodology of this project was designed to systematically analyze COVID-19 data and visualize patterns at global, regional, and country levels. The work involved multiple phases, from data collection and cleaning to visualization and dashboard creation. The following subsections describe each step in detail:

1. Data Collection

The dataset was sourced from publicly available repositories (e.g., WHO COVID-19 dataset).

The data included attributes such as Date\_reported, Country/Region, WHO\_region, New\_cases, Cumulative\_cases, New\_deaths, and Cumulative\_deaths.Since no primary survey was conducted, no questionnaire or sampling methodology was applied. The project relied entirely on secondary data.

2. Data Preprocessing and Cleaning

* Inspected the dataset for missing values and inconsistencies.
* Handled null or invalid values (e.g., replacing with 0 where appropriate).
* Converted date fields (Date\_reported) into proper datetime format for time-series analysis.
* Added additional derived columns such as month and quarter for aggregation.
* Grouped data by countries, regions, and time intervals (daily, monthly, quarterly).

3. Tools and Technologies Used

* Python: The primary programming language used for analysis.
* Libraries:
* Pandas → Data manipulation and aggregation.
* Matplotlib & Seaborn → Static plots (line charts, bar charts, pie charts, heatmaps).
* Plotly → Interactive visualizations and dashboard creation.
* Google Colab: Coding environment for development and testing.

4. Data Analysis and Visualization Steps

* Line plots → To show daily new cases globally and for least/most affected countries.
* Bar charts → To compare quarterly cases and deaths across regions.
* Double bar charts → Side-by-side comparison of cases vs deaths.
* Pie charts → To illustrate top 10 countries by cumulative deaths.
* Heat maps → For quarterly deaths by region and monthly cases by top 10 countries
* Interactive dashboard with Plotly → To allow users to filter, compare, and explore cases and deaths dynamically.

1. **Data Analysis and Results**

**COVID-19 data analysis, results and insights:**

**COVID-19 - Descriptive Analysis**

The dataset covers COVID-19 cases from 2020-03-01 to 2023-08-31, across 240 countries.

A total of 770,122,340 new cases and 6,968,439 deaths are recorded.

**COVID-19 - Analysis and Insights**

The country most affected by COVID-19 cases is United States of America with 103,436,760 cases.

The country with the highest deaths is United States of America with 1,140,868 deaths.

The highest daily surge in cases was 8,401,963 on 2022-01-30.

The peak quarterly cases occurred in 2022Q1 with 199,602,568 cases.

**COVID-19 - Data Analysis and Results**

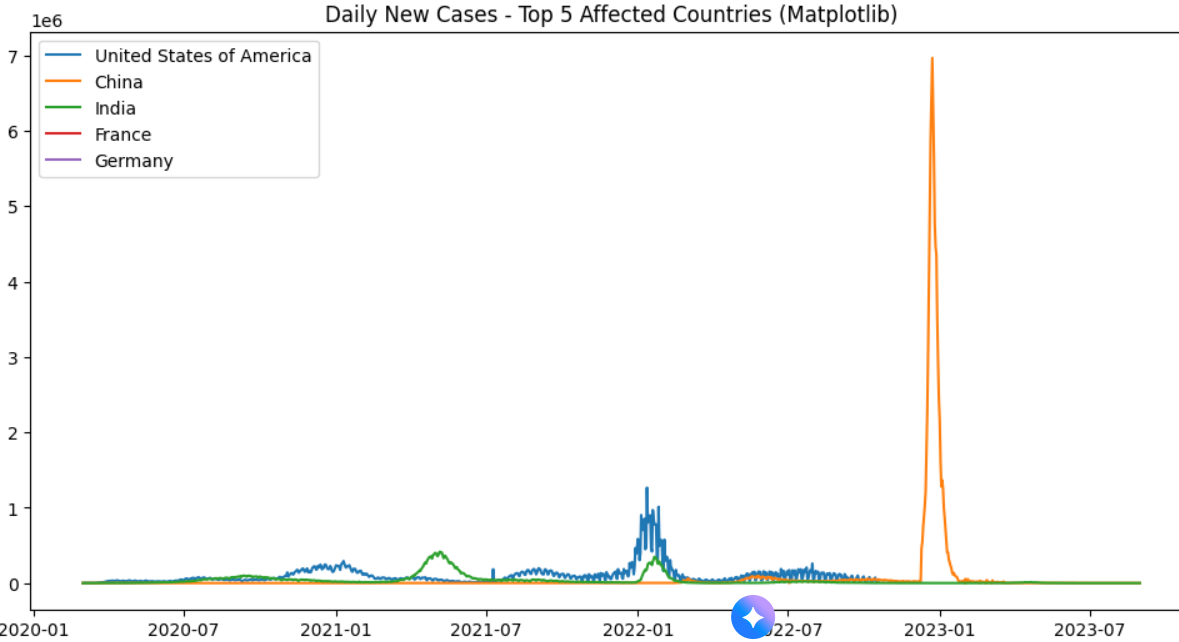
The temporal analysis shows waves of outbreak with clear peaks and declines.

Stacked bar charts indicate that deaths follow similar trends as cases but at lower magnitude.

Heatmaps highlight the most affected countries and months, showing concentrated surges in specific regions.

The interactive dashboard provides a comprehensive tool to explore cases, deaths, and distributions dynamically.

**GRAPHS AND CHARTS GENERATED:**

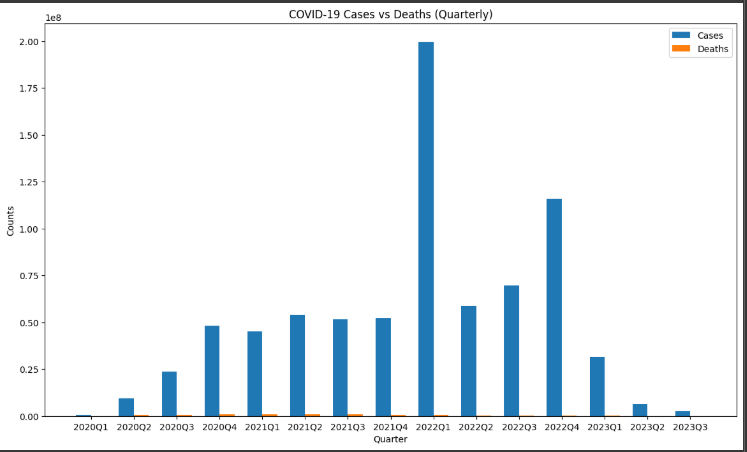
**A graph of a number of cases

AI-generated content may be incorrect.A graph of covid-19

AI-generated content may be incorrect.*A graph of a number of cases

AI-generated content may be incorrect.A map of the world with different colored areas

AI-generated content may be incorrect.***

****

**A graph of a virus

AI-generated content may be incorrect.**

**A colorful pie chart with text

AI-generated content may be incorrect.**

**A screenshot of a graph

AI-generated content may be incorrect.**

**A graph with numbers and a red line

AI-generated content may be incorrect.**

**A screenshot of a graph

AI-generated content may be incorrect.**

**A graph of covid-19 cases

AI-generated content may be incorrect.**

**A graph of a number of people

AI-generated content may be incorrect.**

**Ebola data analysis, results and insights:**

**Ebola - Descriptive Analysis**

The dataset covers Ebola cases from 2014-08-29 to 2016-03-23, across 10 countries.

A total of 836,618 new cases and 375,702 deaths are recorded.

**Ebola - Analysis and Insights**

The country most affected by Ebola cases is Liberia with 819,411 cases.

The country with the highest deaths is Liberia with 369,614 deaths.

The highest daily surge in cases was 10,749 on 2015-07-30.

The peak quarterly cases occurred in 2015Q3 with 491,253 cases.

**Ebola - Data Analysis and Results**

The temporal analysis shows waves of outbreak with clear peaks and declines.

Stacked bar charts indicate that deaths follow similar trends as cases but at lower magnitude.

**GRAPHS AND CHARTS GENERATED:**

**A graph of a number of people

AI-generated content may be incorrect.**

**A graph of a graph with numbers and a bar

AI-generated content may be incorrect.**

**A screen shot of a graph

AI-generated content may be incorrect.**

**A screenshot of a graph

AI-generated content may be incorrect.**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

Heatmaps highlight the most affected countries and months, showing concentrated surges in specific regions.

The interactive dashboard provides a comprehensive tool to explore cases, deaths, and distributions dynamically.

1. **Conclusion**

The project successfully demonstrated how systematic data analysis and visualization can uncover meaningful insights into the COVID-19 pandemic. By applying Python tools and visualization libraries, we were able to transform raw datasets into clear, interpretable patterns.

Our findings showed that:

The top 10 most affected countries by cumulative deaths were consistently those with high population density and international connectivity, confirming global observations. The quarterly heat map of deaths clearly highlighted peak mortality periods during major pandemic waves, with certain WHO regions experiencing more severe surges.The monthly heat map of new cases for top countries revealed contrasting patterns of outbreak intensity, reflecting differences in containment measures and vaccination progress. The global daily cases line plot resembled a mountain-like trajectory, emphasizing how the pandemic unfolded in successive waves.The interactive dashboard proved effective in offering dynamic exploration, enabling policymakers and researchers to compare trends across regions and timelines with greater flexibility.

From these outcomes, it is concluded that visualization and dashboard-based approaches make pandemic data more understandable, actionable, and engaging compared to static reports.

1. **APPENDICES**

References

1. World Health Organization (WHO). Coronavirus (COVID-19) Dashboard. Retrieved from: <https://covid19.who.int>

2. Johns Hopkins University. COVID-19 Global Data Repository.

3. Python Libraries: Pandas, Numpy, Matplotlib, Seaborn, Plotly documentation.

4. Kaggle Ebola Dataset

GitHub Repository

All Python scripts and Jupyter Notebooks developed for this project have been uploaded to GitHub:

🔗 [GitHub Repository Link – https://github.com/Souhardya-Mukherjee/IDEAS-TIH-ISI-Kol-covid-19-dataset-project.git]

Contents include:

Data preprocessing scripts

Visualization codes (line plots, bar charts, heat maps, pie charts)

Plotly interactive dashboard code

PDF copy of the final report