Project Report For CS661: BIG DATA VISUAL ANALYTICS
2022-2023 Semester II
Project Title: CovidLens
Team members:
Pushkar Bhardwaj (231110605), Krishna Kumar Bais (241110038), Milan Roy (241110042), Rohan (241110057), Sevak Shekokar (241110065), Tsewang Chukey (241110092), Tsewang Namgail (241110093), Sanjay Singh Shekhawat (241230012)
Member emails:
pushkarb23@iitk.ac.in, krishnakb24@iitk.ac.in, milanr24@iitk.ac.in, rohan24@iitk.ac.in, sssanjay24@iitk.ac.in, bssevak24@iitk.ac.in, tsewangn24@iitk.ac.in, tsewang24@iitk.ac.in,
IIT Kanpur

1.Introduction :

The COVID-19 pandemic has significantly impacted global health system, education, economies, and every aspect of daily life. The large-scale availability of pandemic-related data allows for insightful visual analytics, helping researchers,governments, policymakers, and the general public understand trends and patterns. Initially they have faced an urgent need for accurate and up-to-date data to monitor the spread of the virus, healthcare capacity, evaluating intervention strategies and plan for vaccination and recovery efforts. Recognizing the critical role of data-driven decision making during a global helath-crisis, this project aim to leverage the Our World in Data COVID-19 Dataset to create a comprehensive web-based dashboard. This dashboard will enable users to explore COVID-19 trends through interactive visualizations, gaining insights into infection rates, vaccination progress, and mortality trends worldwide. The platform provides users with the ability to isolate specific regions, countries or time periods and also analyze complex patterns in a user-friendly environment.

1.1 Data Sources

Our project have used the *Our World in Data COVID-19* Dataset. The dataset includes country-wise and day-wise records of new cases, total cases, and deaths. It also contains vaccination-related data, such as new vaccinations, total vaccinations, and the vaccine manufacturers used, all categorized by country and date. Additionally, the dataset includes information on COVID-19 testing, hospitalizations, as well as Google Mobility data. For a focused analysis of the impact in India, we have obtained additional data from Kaggle.

1.3 Proposed Solution

Our platform is a standalone web-based solution built using Python and Streamlit, with Plotly used for generating the visualizations.

2. Tasks

The following sections describe our visualization tasks, design justifications, and results. We have also included snapshots of the visualizations.

2.1 Task 1 : Analysis of Disease Spread

2.1.1 Overview

The primary aim of this task is to perform a comprehensive analysis and visualization of the global spread of COVID-19, using publicly available datasets. The main file used for this task is cases_deaths.csv, which records the number of reported cases and deaths over time for countries worldwide. The core objective is to highlight the pandemic's impact across different continents and countries. This involves not only presenting the total number of cases and deaths, but also examining how rapidly these numbers increased in various regions. By analyzing both absolute counts and population-adjusted metrics, the task seeks to provide a balanced view of the pandemic's severity.

An additional aim of the task is to facilitate meaningful comparison between regions through interactive visualizations. Suitable graphical representations—such as pie charts, bar graphs, and choropleth maps—are employed to make complex data more understandable and help users draw inferences easily. By offering users the ability to filter data by region, adjust for population size, and track progression over time, the visualizations are designed to support deeper insight into how COVID-19 unfolded around the world.

2.1.2 Data Preprocessing

As discussed in the previous section, the primary file used for this task was cases_deaths.csv. However, due to its large size exceeding GitHub's 100MB upload limit, the file was preprocessed and split into two separate files: cases.csv, containing data related to cases, and deaths.csv, containing data related to deaths. Additionally, approximately 1% of the data was missing in some fields. These missing values were handled in two ways: values corresponding to new cases or deaths (including population-adjusted metrics) were filled with zeros, while values related to total cases or deaths (also including population-adjusted metrics) were forward-filled using the most recent available data.

To enable additional functionalities on the page, such as continent-wise analysis and choropleth mapping, new columns were added to each file. These include the three-letter ISO country code and the corresponding continent to which each country belongs. Although the dataset included records up to February 2025, many countries had ceased reporting data well before that point. Therefore, an additional filter was applied to exclude any entries dated after January 1st, 2024

2.1.3 Visualizations and Design Justification

The visualizations of this task are seperated into three tabs, namely Overview, Map Visualization and Timeseries Analysis.

Overview

- **Total Cases, Total Deaths (Pie Chart and Bar Graph):** Contains two section: COVID-19 Impact by Continents and COVID-19 Top Countries Impacted.
  - The former visualizes data aggregated by continent, while the latter focuses on country-level data.
  - The data presented corresponds to the most recent total recorded cases and deaths as this section emphasizes the final impact of the pandemic rather than its progression over time.
  - If a bar graph is selected, an additional option becomes available to display the data relative to population for better comparison across regions.
  - Users can choose between a pie chart or a bar graph for visualization. If a bar graph is selected, an additional option becomes available to display the data relative to population for better comparison across regions.
  - The Top Countries section allows the user to customize the number of countries to visualize—ranging from a minimum of 2 to a maximum of 8. This range adheres to the "magic number 7" rule, ensuring the visualization remains clear and easy to interpret.
  - Additionally, users can select a specific region from a dropdown menu, choosing either "World" or any individual continent. This enables focused comparisons of the top countries within a selected region.

- **Pie charts** were chosen for the visualizations because they effectively convey proportional distributions and highlight the part-to-whole relationships within the data. This format provides an intuitive understanding of how different components—such as continents or countries—were impacted relative to the whole. Unlike bar graphs or scatter plots, pie charts offer a quick visual impression of the relative magnitude of each category, making them especially suitable for summarizing final totals.

- **Bar graphs**, on the other hand, offer a clearer comparison of exact values across categories. By normalizing the data—adjusting values relative to population size—the graphs provide a fairer basis for comparison, highlighting the relative impact of COVID-19 rather than just the raw totals. This approach ensures that countries or regions with larger populations are not unfairly represented, as it's natural for areas with more people to report higher raw numbers. By adjusting for population, the visualization presents a more equitable view of the relative impact.

Map Visualization

- **Choropleth Animation:** The Map Visalization tab features an animated choropleth map to visualize the spatial and temporal spread of COVID-19.
  - Users can select the metric to visualize, with options for total cases or total deaths.
  - Essentially, this plot functions as a heatmap overlaid on a geographical map, where countries are shaded based on a color scale corresponding to the selected metric.

  - To manage the large data size and ensure smooth rendering, the visualization is updated based on monthly changes rather than daily updates.
  - The users can select the region to visualize, with options including the entire world or individual continents. This allows for a more focused analysis of a specific continent, as countries will be shaded based on the values of only the countries within that continent, enabling direct comparison between countries of the same continent rather than across the entire world.

- While temporal changes could also be animated using other visualizations such as bar graphs, they fail to capture the spatial proximity of countries. In contrast, a choropleth map overlays data onto an actual world map, making the geographic relationships between countries immediately intuitive. Furthermore, with data from over 200 countries, plotting all of them on a single bar graph would lead to visual clutter, making it difficult to draw meaningful insights.

Timeline Plots

This tab contains two sections: Time Series Analysis and Threshold Timeline.

- **Time Series Analysis (Line Graphs)** Displays line graphs for a selected metric.
  - Includes two subsections: continent-wise analysis and country-wise analysis.
  - Users can choose the metric to plot (cases or deaths), time interval (daily, weekly, biweekly or cummulative). If the cumulative interval is selected, users can enable a logarithmic scale to better visualize exponential growth patterns.
  - For both plots, users can choose which continents/countries to display, offering more control and a focused comparison.

  Line graphs were chosen for these plots as they are a widely used and effective tool for time series and trend analysis. By displaying data in chronological order, line graphs make it easier to observe rises and falls in values, as well as the rate of change. This clarity helps in identifying both long-term trends and sudden shifts in the data. Additionally, multiple lines can be plotted on the same graph, enabling easy comparison across categories such as continents or countries. For cumulative data, using a logarithmic scale helps in interpreting exponential growth patterns more effectively.

- **Threshold Timeline:** displays when different countries reached key COVID-19 threshold events:
  - 100 total cases
  - 1 case per million population
  - 5 total deaths
  - 0.1 deaths per million population

- Each event is visualized using a scatterplot, where the x-axis represents time and the y-axis lists countries. Each country is represented by a bubble, and countries positioned closer to the left (earlier on the x-axis) indicate that they reached the threshold sooner. The bubbles are also color coded by continent.

2.1.4 Customization and Interactive Features

- A radio button in the Overview tab allows users to switch between a Pie Chart and a Bar Graph.
- Multiselect boxes enable users to compare multiple categories (continents or countries) within the same plot.
- A checkbox option lets users adjust plots to show values relative to population for a fairer comparison.
- A selectbox provides an easy way to choose which metric to plot, either cases or deaths.

2.1.5 Results

- Asia has significantly higher total cases and comparative total deaths compared to Europe.1 However, Asia has much lower total cases and total deaths per million population. 2
- India reported far more COVID-19 cases and deaths than China, but China's reported deaths were disproportionately low compared to case counts, suggesting potential reporting biases.3
- In North America, the first cluster was in the United States; in South America, it was Brazil; and in Asia, it was India. In the other continents, no single country stood out as the initial epicenter.4
- China was the first country to reach 100 Total Cases, on January 19, 2020. The second and third countries—South Korea and Japan—reached the threshold about a month later, on February 21 and February 22, respectively, both in Asia. Italy was the first European country to reach it, on February 24. North America followed with the USA reaching it on March 3. The other continents also reached the threshold around the same time.5
- Interestingly, while Germany only reached the 100 Total Cases threshold on March 1, 2020, it was the second country to reach the 5 Total Deaths threshold, doing so on January 26—just five days after China. The next European country, Italy, reached it almost a month later, on February 25. Germany was also the first country to reach the 0.1 Total Deaths per Million Population threshold, achieving it on February 2, more than 20 days before the second country.

2.1.6 Key Visualizations from the Dashboard

Some figures based on which, the above conclusions were drawn.

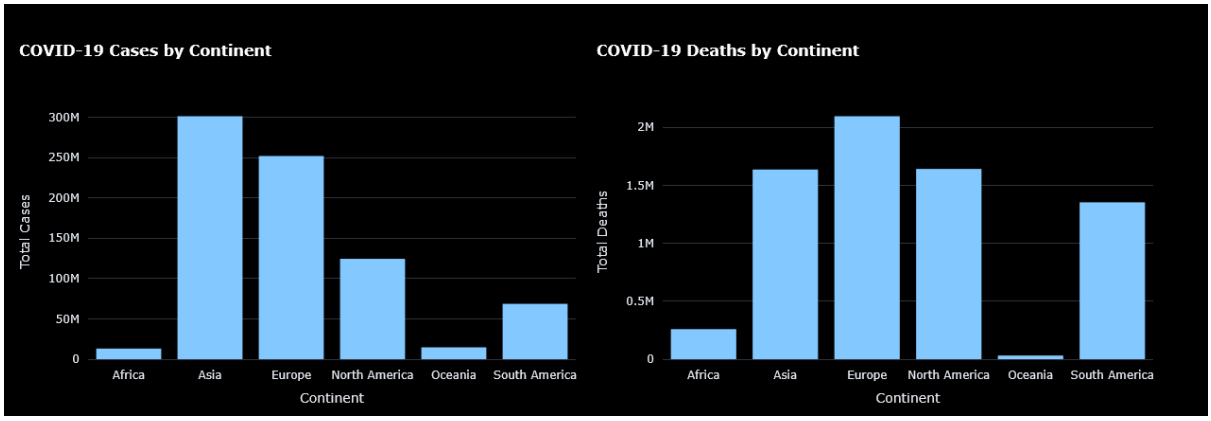


Figure 1: Absolute counts of total cases and deaths across different continents.

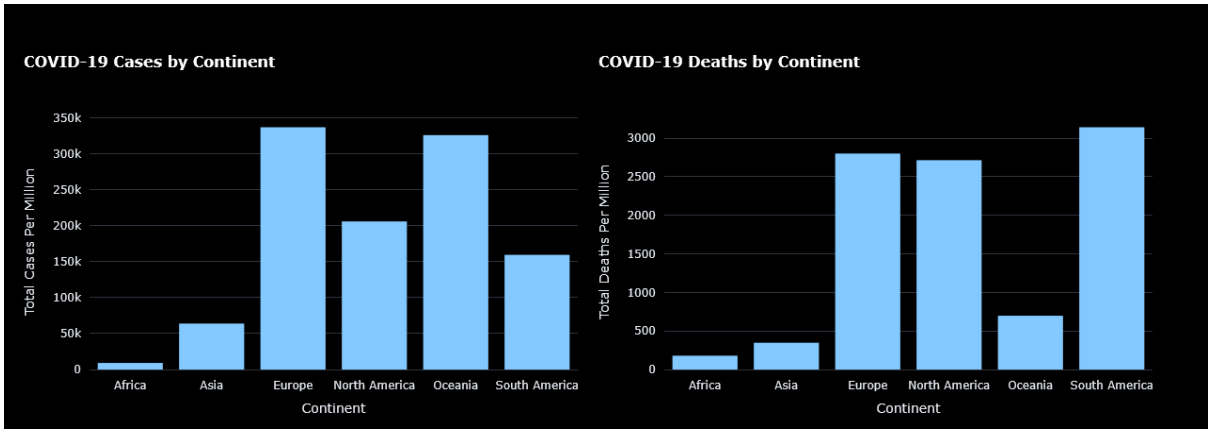


Figure 2: Counts per million (normalized) of total cases and deaths across different continents.

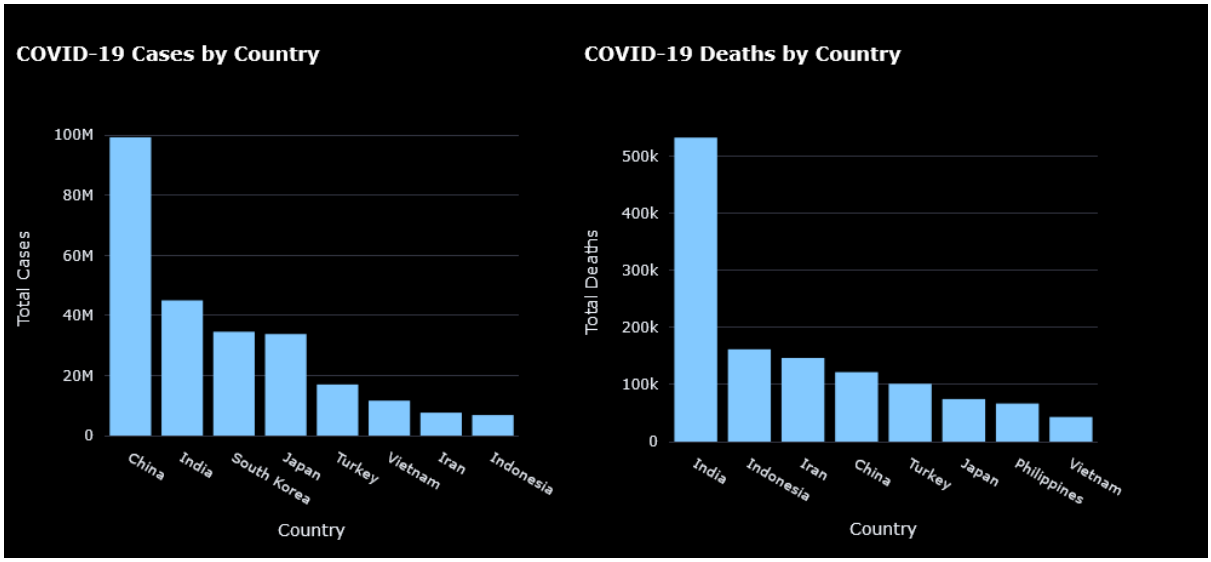


Figure 3: Top Countries in Aisa by Total Cases and Total Deaths.

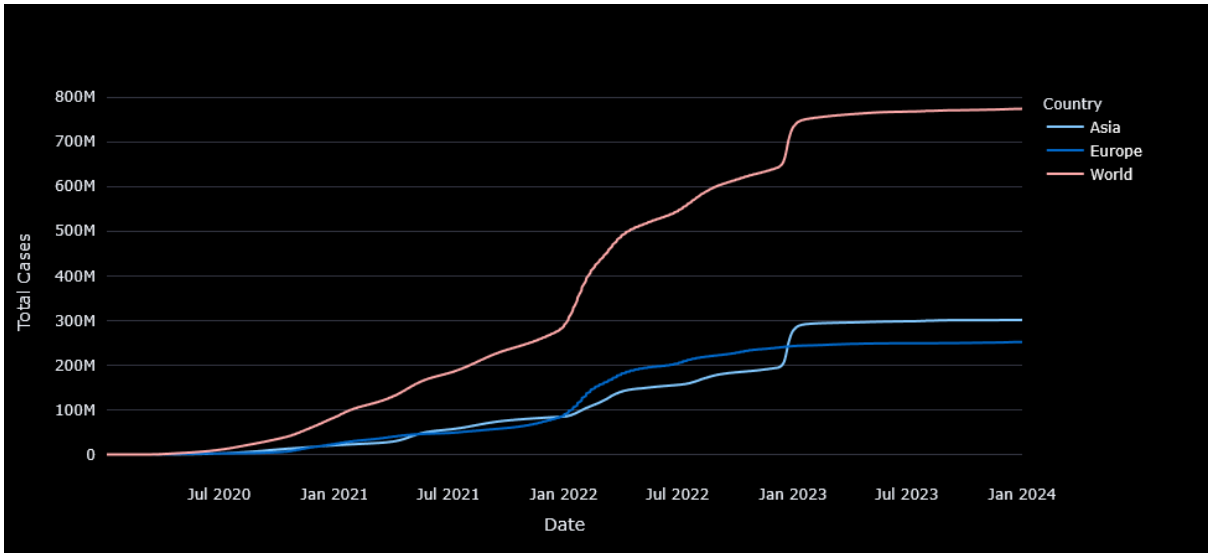


Figure 4: Timeline of Total Cases

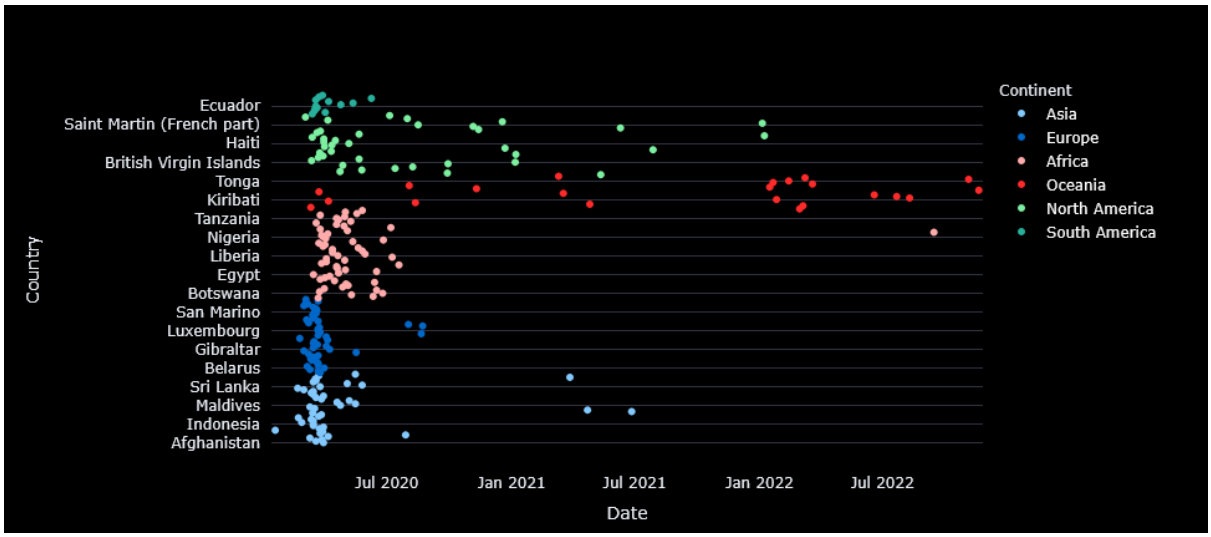


Figure 5: Timeline of countries reaching 100 cases

2.2 Task 2 : Mortality Impact Analysis

2.2.1 Overview

The compiled fatality figures for the global COVID-19 pandemic were sourced primarily from government self-reportage. Due to the massive variance in the nature of regime across nations, the definitions of COVID fatalities, people's reluctance towards getting tested, as well as national access to testing resources, this figure could be massive under-represented. To obtain a better estimate of the mortality impact of COVID-19, it makes sense to compare the quantum of total deaths during the pandemic to the expected deaths during the same period had the pandemic not occurred, based on extrapolation of historical data.
This task primarily focuses on analyzing and representing the discrepancies between the two data collection methods.

2.2.2 Data Processing & Feature Engineering

The following Processing and Engineering is carried out on the original data to reduce size by removing irrelevant or redundant data, handling missing values and feature engineering to introduce new relevant features.

- For this task, the data is sourced from two files from the original dataset: 1cases_deaths.csv' and 'excess_mortality_economist.csv'. Features not directly involved in creating the visualization have been removed, significantly reducing the dataset size. Two output files are obtained to track global and nation-wise trends separately. This introduces a small data overhead but saves on compilation-time computation.

- Additional filtering of redundant data is carried out by filtering data from 2019-01-01 to 2024-01-01 and removing continent-wise and data from various groups of nations using the 'pycountry' library.

- Linear interpolation is used to predict missing values for country-wise and global data. This maintains the variance of the data by not replacing the missing values with mean or median values.

- Scaled normalized metrics (per million) are used instead of absolute values to provide better objective comparison between countries.

- The share of non-null values in increased from 72.15% to 98.33% for the final dataset.

- Two additional features are added: Cumulative Excess Deaths to Case Deaths Ratio given by

  $$\texttt{Cumulative Excess Deaths to Case Deaths} = \frac{\texttt{Cumulative Excess Deaths}}{\texttt{Cumulative Reported Deaths}}$$

  and Estimated CFR based on Excess Deaths given by

  $$\texttt{Estimated CFR} = \frac{\texttt{Cumulative Excess Deaths} \times \texttt{Reported CFR}}{\texttt{Cumulative Reported Deaths}}$$

2.2.3 Visualization Architecture

Module	Components
Global Metrics	Real-time excess/reported death ratio (2.34:1) Cumulative excess deaths/million (2815) Reported deaths/million (1311)
Analytical Tabs	`['Reported and Excess Deaths', 'Excess to Reported Deaths', 'Reported CFR and Estimated CFR', 'Map Overlays', 'Excess Deaths Top & Bottom Nations']`

2.3.4 Visualization Techniques

- Savitzky-Golay filtering (window=64, polynomial=2) for noise reduction
- Synchronized multi-axis plots for temporal comparisons
- Adaptive choropleth color scaling using Plotly's Plasma scale

2.3.5 Results

The resultant graphs are shown below:

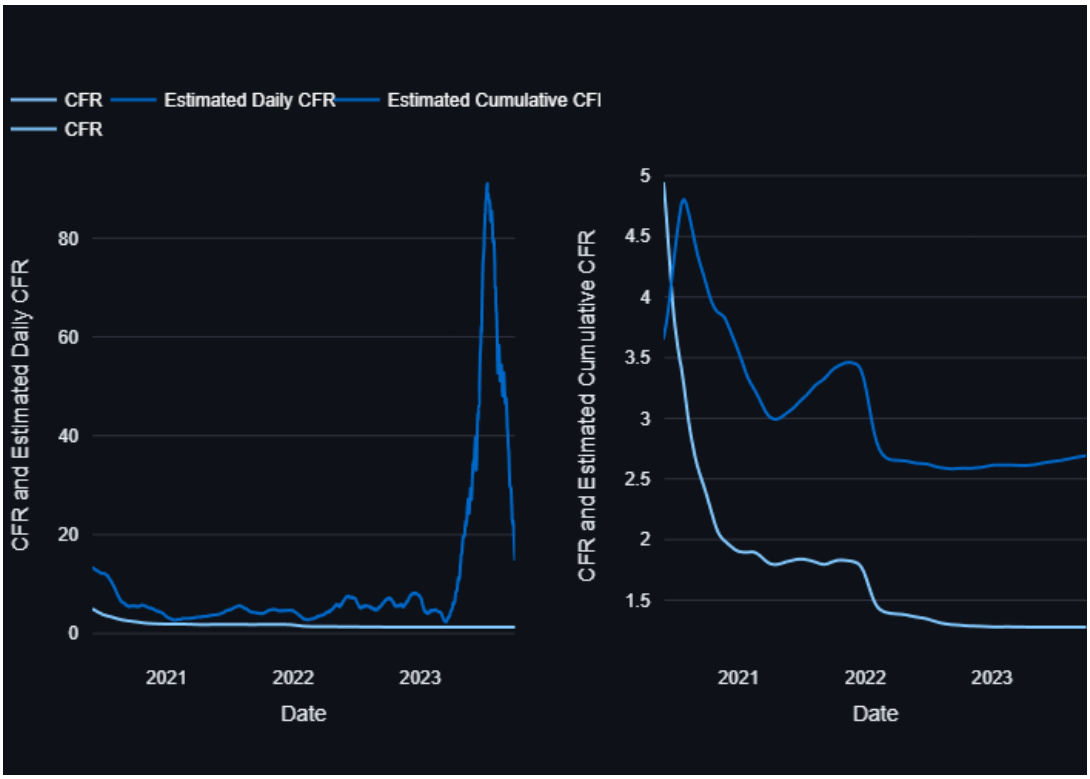


Figure 6: Reported and Excess Death Daily and Cumulative Figures over Time

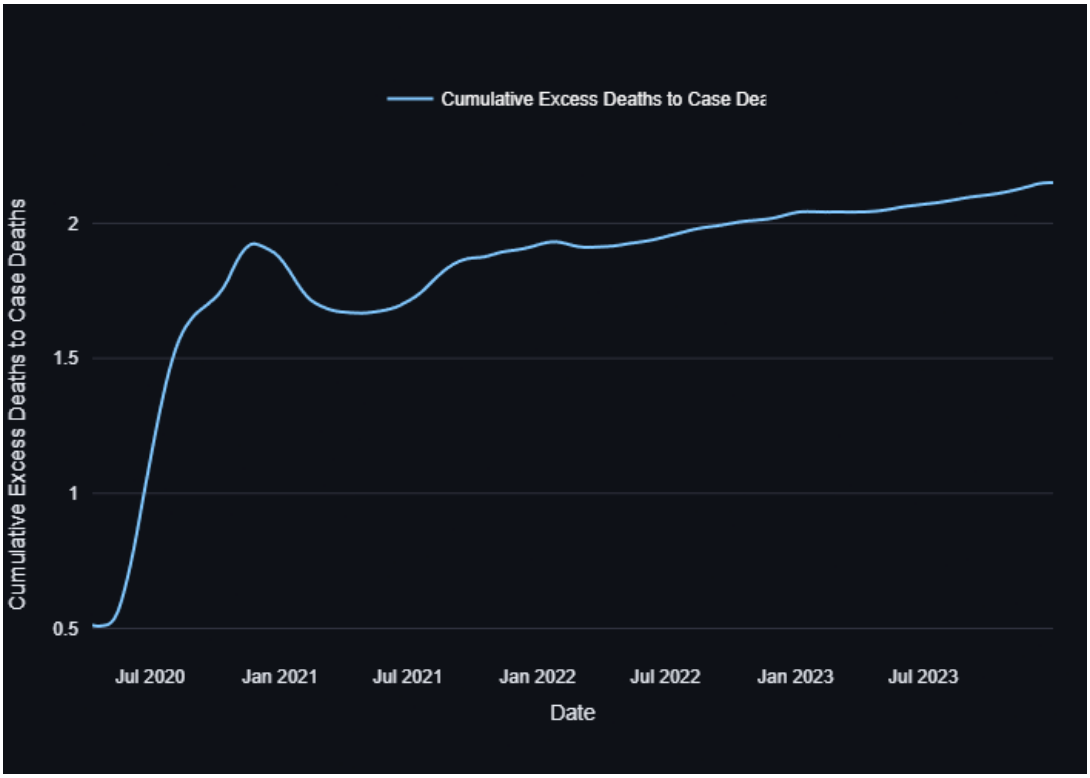


Figure 7: Cumulative Excess Deaths to Reported Deaths over Time

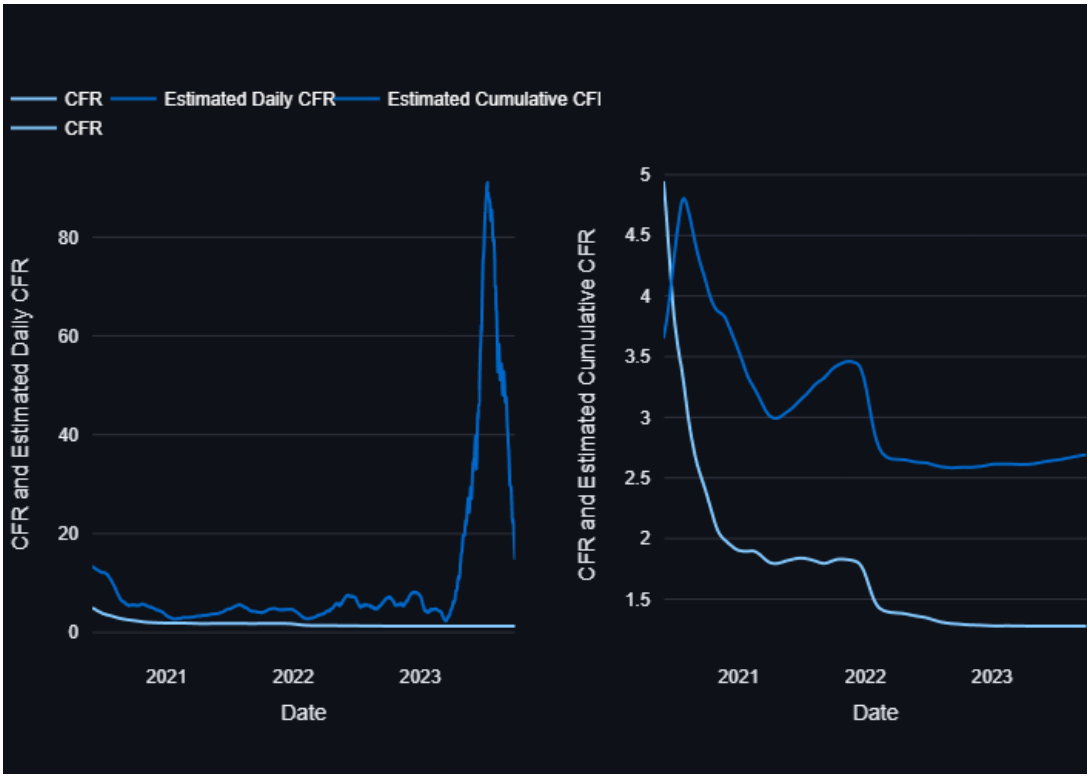


Figure 8: Reported and Estimated Case Fatality Rate and Cumulative Figures over Time

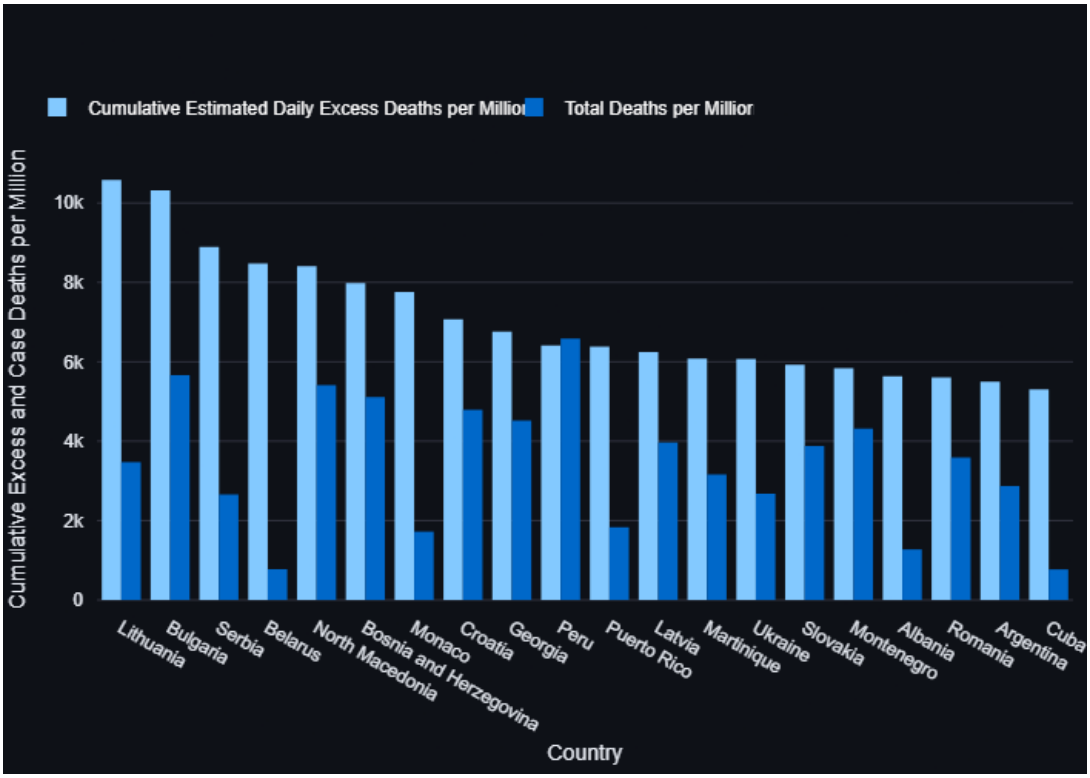


Figure 9: Top 20 Countries based on Cumulative Excess Deaths per Million

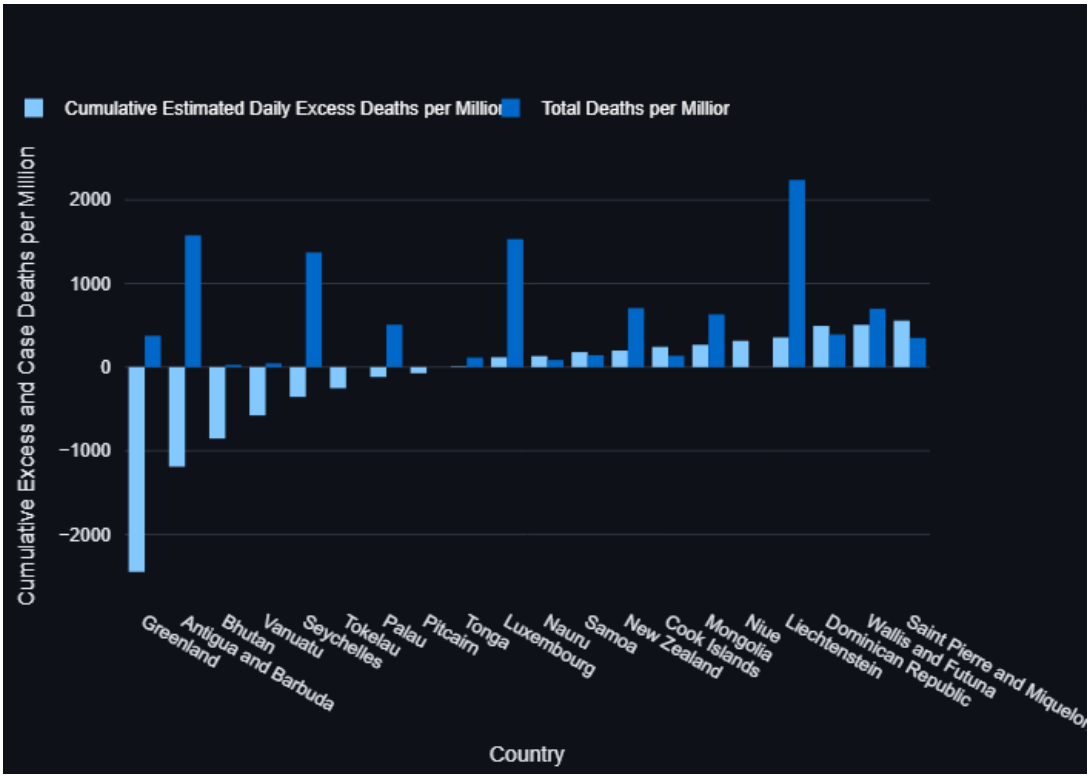


Figure 10: Bottom 20 Countries based on Cumulative Excess Deaths per Million

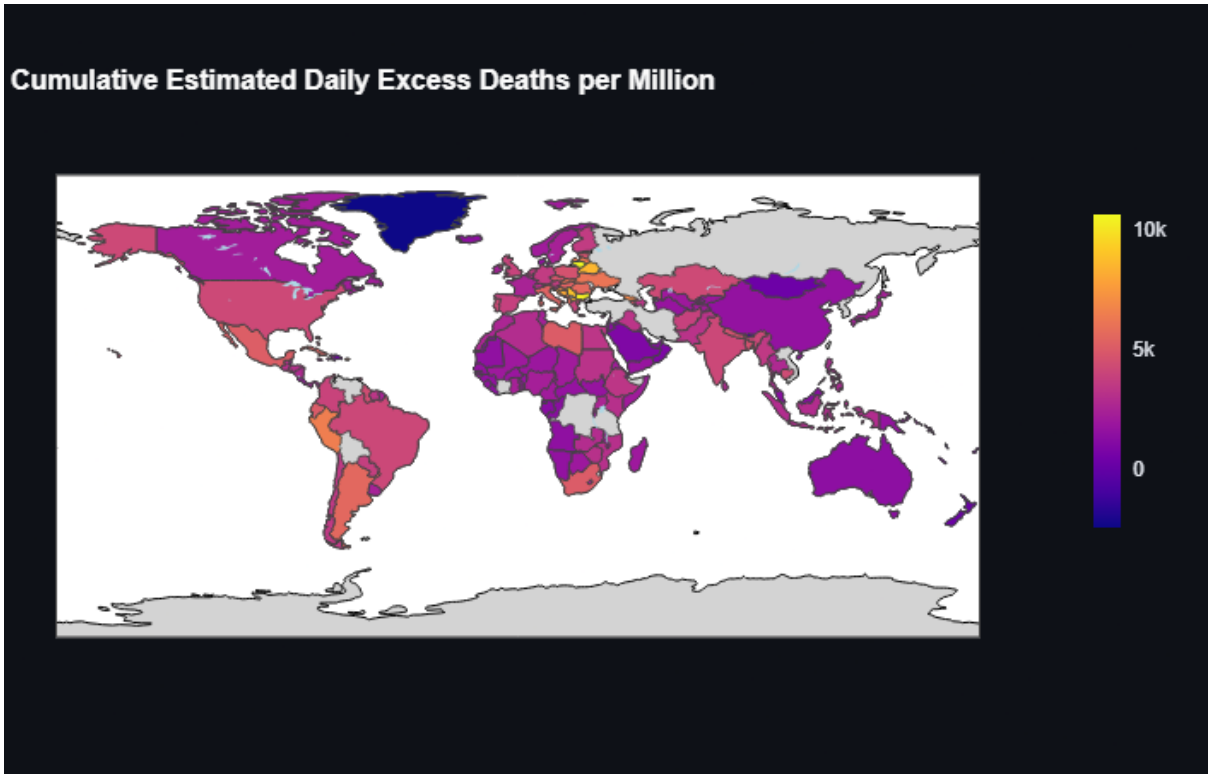


Figure 11: Country-wise Cumulative Daily Excess Deaths per Million

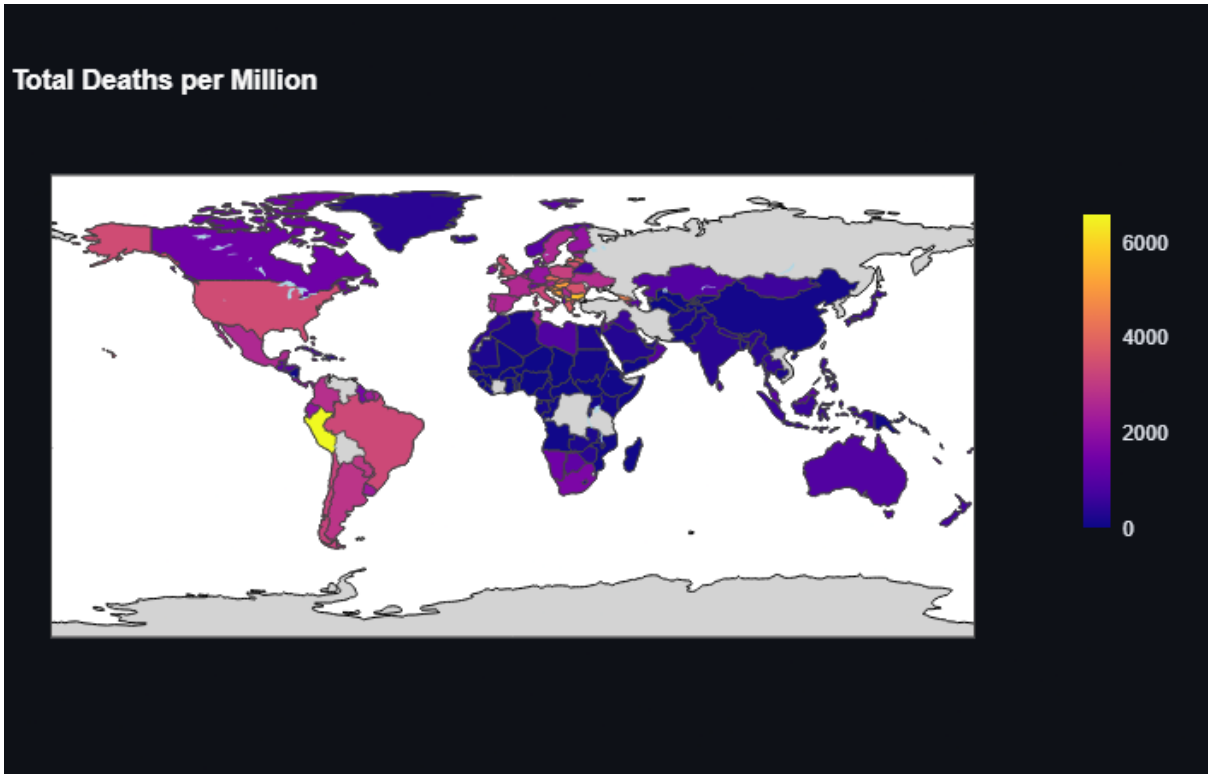


Figure 12: Country-wise Cumulative Reported Deaths per Million

The key takeaways are the following:

- Total Reported Deaths per million is about 1311.

- Total Excess Deaths per Million is about 2815.
- Averaged Across Countries without considering relative populations, Excess Deaths to Reported Deaths ratio is about 2.15.
- This figure is more stark considering that the most populous nations including India and China massively under-reported their numbers according to the excess death estimates.

2.3 Tasks 3 : Analysis of Vaccination Data

2.3.1 Overview :

The vaccination analysis task aimed to transform raw COVID-19 data into actionable insights through interactive visualizations. By focusing on global vaccination trends, country-specific progress, manufacturer dynamics, and correlations with health outcomes, this analysis provides policymakers with tools to evaluate vaccine distribution efficiency, identify disparities, and assess public health impacts. The dashboard emphasizes spatial, temporal, and comparative perspectives to answer critical questions:

- Which regions lag in vaccination coverage?
- How do vaccination rates correlate with case/death reductions?
- Which manufacturers dominate specific markets, and how does this evolve over time?

Through real-time exploration capabilities, users can navigate and interpret complex vaccination trends with ease. This tool offers a flexible, real-time view into vaccination efforts, comparing trends across countries, regions, and manufacturers, thereby supporting better insights and public health strategies.

2.3.2 Data Preprocessing

The analysis began with integrating two datasets: cases_deaths.csv - containing daily COVID-19 cases, deaths, and normalized epidemiological metrics, vaccinations_global.csv - recording vaccination metrics such as daily vaccinations, total vaccinations, and rolling averages.

Key preprocessing steps included:

- **Merging Datasets:** The cases_deaths.csv and vaccinations_global.csv files were merged using the country and date fields. This ensured that for every country and date, both epidemiological data and vaccination progress data were available side-by-side, and only required columns were selected.
- **Data Cleaning:** After merging, rows with missing values in critical vaccination or case metrics were removed to maintain data integrity and reliability for visualization.
- **Manufacturer Data Handling:** The third dataset, vaccinations_manufacturer.csv, was separately loaded for use in the manufacturer analysis tab. Upon inspection, it was determined that this file had no missing values and was fully clean and directly usable, requiring no preprocessing.
- **Temporal Filtering:** Although the dashboard allows dynamic date filtering, the data range was naturally focused on the pandemic's active vaccination phase, from December 2020 to August 2024.

Clarification: No explicit normalization (like dividing by population) was performed during preprocessing. However, normalized fields such as total_cases_per_million and weekly_deaths_per_million were already present in the raw datasets, sourced from Our World in Data.

2.3.3 Visualization Components and Design Justifications

Tab 1: Global Patterns

- **Total Cases, Total Vaccinations, Total Deaths (Choropleth Maps):** These maps visualize spatial disparities in vaccination coverage and health outcomes:
  - **Total Cases by Country** uses a Plasma color gradient where lighter yellow hues represent higher case counts, and darker hues signify lower counts.
  - **Total People Vaccinated by Country** follows the same plasma scale.
  - **Total Deaths by Country** again uses color to encode normalized mortality levels.

  Choropleths were specifically chosen for their geographic expressiveness , they allow users to immediately identify regional clusters and disparities through variations in color intensity. Bubble maps were rejected because in densely populated areas like Europe, bubble overlaps would have caused confusion, reducing readability and clarity.

- **Top Regions by Total Vaccinations (Bar Chart):** A dynamic bar chart ranks regions by total vaccination doses administered:
  - Users can toggle between viewing rankings by country, continent, or income group using a radio button.
  - **Bar charts** were chosen because they allow precise magnitude comparisons. For example, showing China's 3.51 billion doses compared to India's 2.20 billion doses clearly through bar length. Pie charts were considered but rejected because they obscure exact differences in magnitude, especially when comparing many entities.

Tab 2: Country Analysis

- **Vaccination Progress Over Time (Line Chart):** This multi-line plot tracks the progression of:People with at least one vaccine dose, Fully vaccinated people, People receiving booster shots.
  - **Line charts** were chosen over stacked area charts to better distinguish overlapping vaccination phases without occlusion problems.
  - Observations such as the early rapid rollout of first doses in the US and the slower initial booster uptake in India are made clearly visible.

- **Vaccinations vs. Health Outcomes (Scatter Plot):** A flexible scatter plot allows users to select: X-axis: Daily vaccination rates or total people vaccinated,Y-axis: New cases, deaths, or weekly metrics.
  - **LOWESS trendlines** reveal correlations — for example, correlations between vaccination rates and case/death rates. Scatter plots were preferred over heatmaps to retain per-country data granularity and allow precise trend analysis.
  - **scatter plot** was chosen over heatmaps to maintain per-country data detail and clearly show trends and correlations between vaccination rates and health outcomes

Tab 3: Comparative View

- **Vaccination Volume Comparison (Bar Charts):** This visualization presents a multi-line time series comparing daily vaccination metrics across selected countries.

  Users can dynamically select one of four different vaccination metrics to plot: Daily vaccinations smoothed, Daily people vaccinated smoothed, Daily people vaccinated smoothed per hundred, Daily vaccinations smoothed per million.

  Overlapping lines on a single timeline preserve temporal synchronicity better than separate small multiple plots.

- **Daily Vaccination Rate Trends (Line Chart):** This visualization tracks the daily vaccination rate trends across selected countries. The line chart dynamically updates to show variations in vaccination rates per country, offering insights into the speed of vaccination rollouts over time.

  Users can select a specific vaccination metric (such as daily vaccinations, people vaccinated, or vaccinations per hundred people) for a more tailored view.

  Line charts were chosen because they provide a clear, continuous view of vaccination progress over time, allowing users to easily compare trends between countries. Bar charts would not capture the temporal aspect of the data as effectively.

Tab 4: Vaccine Manufacturers

- **Vaccine Market Share (Pie Chart):** A pie chart visualizes vaccine brand dominance within a country: Pfizer/BioNTech accounted for 59.7% of vaccinations in the US, followed by Moderna (37.4%).

  Pie charts were suitable here because only a few major manufacturers dominate, making proportions easy to interpret.

- **Manufacturer Distribution Over Time (Stacked Bar Chart):** Shows how vaccine types have been administered cumulatively over time for a selected country.

  Stacked bar charts were chosen to clearly show the cumulative distribution of vaccine types over time and compare manufacturer contributions. Line charts wouldn't capture the cumulative proportions as effectively.

- **Manufacturer-Specific Analysis over Time (Line Chart):** This line chart provides a detailed view of vaccine uptake trends over time, segmented by manufacturer, within the selected country.

  Flexibility in Manufacturer Selection: Users have the option to select multiple vaccine manufacturers for comparison within the chart.

  Line chart was chosen because it clearly shows trends over time and allows for easy comparison of vaccine uptake across multiple manufacturers.

  This feature enables flexible and customizable views, allowing for direct comparisons of vaccination trends across a range of vaccine producers.

2.3.4 Customization and User Experience Features

- Interactive Date Range Selector: Users can seamlessly select specific date ranges using an interactive slider, allowing for precise analysis of vaccination data over customized time periods.
- Dynamic Metric Selection: The dashboard offers the ability to select and visualize different metrics dynamically.
- Country Comparator: The country comparison functionality allows users to select multiple countries for direct comparison of vaccination metrics.
- Hover Tooltips: To support detailed data exploration, the hover tooltips provide additional insights when users hover over data points. It dynamically displays relevant data, such as exact values and additional context. This enhances the exploration of data, ensuring clarity without overwhelming the visual.
- Data Expanders: For those interested in deeper insights, the dashboard includes expandable sections that allow users to view raw, unprocessed data alongside visualizations.

2.3.5 Results

- High-income countries achieved significantly higher vaccination coverage compared to low-income regions, with Africa remaining the least vaccinated region.
- China and the United States had similar case counts, but China experienced significantly fewer deaths, emphasizing the effectiveness of broader vaccination.
- India's delayed booster rollout contributed to a COVID surge in mid-2022, which stabilized after booster coverage improved in 2023.
- Pfizer dominated vaccine market share in the United States, while Sinovac was widely used in Asia due to easier storage requirements.
- Countries with higher booster vaccination rates showed noticeable reductions in deaths during major variant-driven waves.

2.3.6 Key Visualizations from the Dashboard

To make the analysis easier to understand, important visualizations from the developed dashboard are presented below, showing key trends and patterns in the vaccination data.

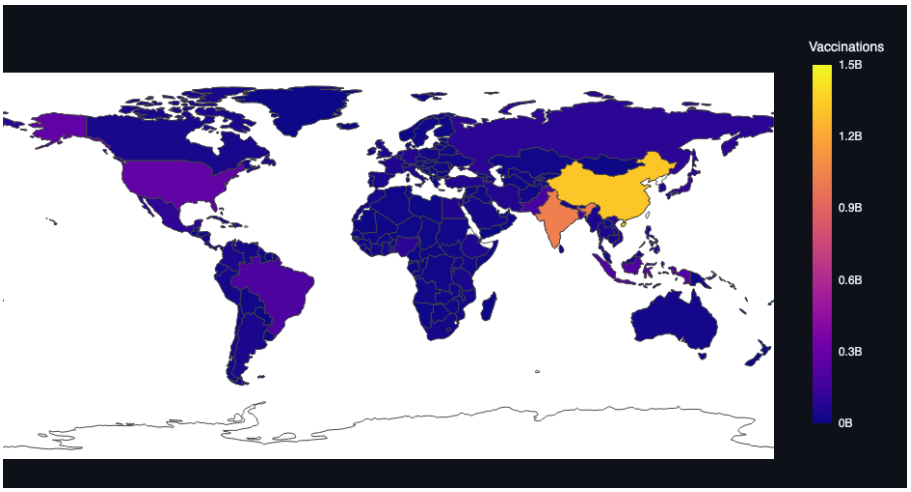


Figure 13: Global Total People Vaccinated (Choropleth Map).

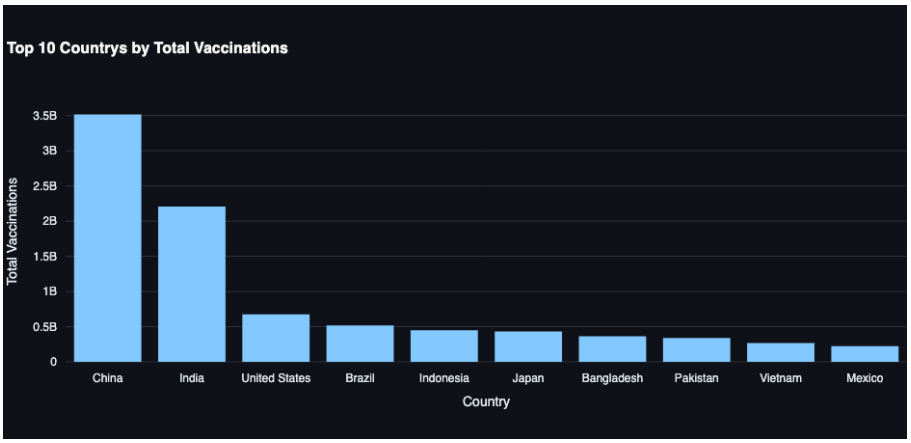


Figure 14: Top 10 countries by Total Vaccinations.

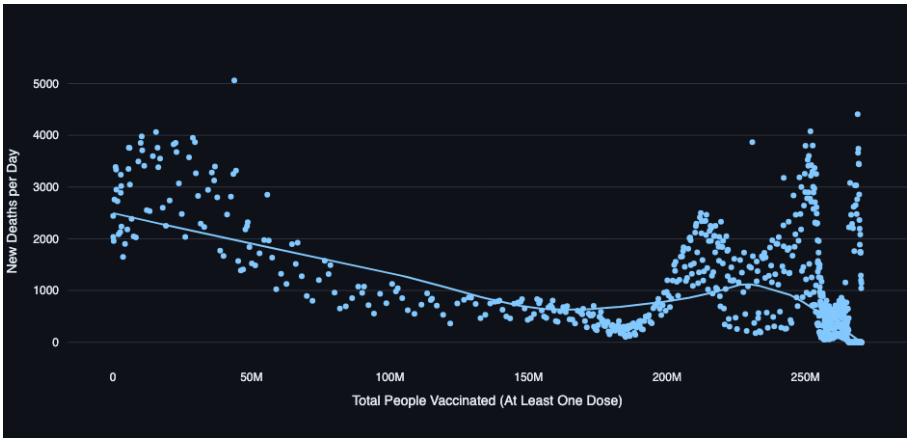


Figure 15: Vaccination Rates vs. New Deaths in the United States.

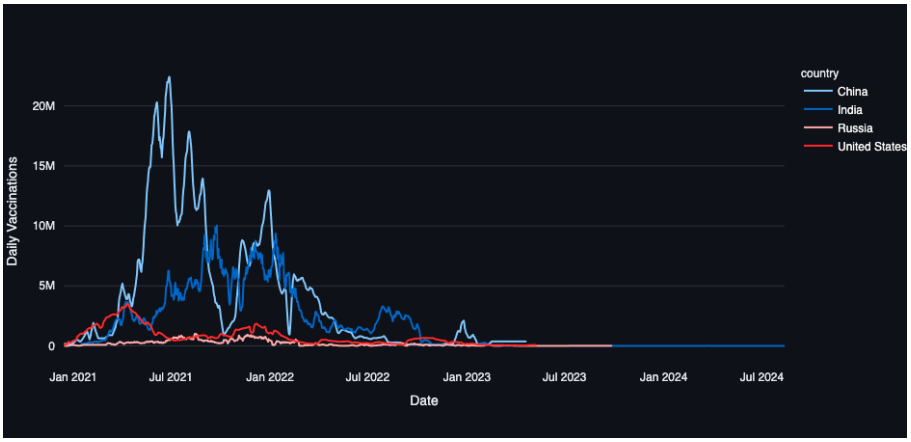


Figure 16: Daily Vaccination Rate Trends Across Selected Countries.

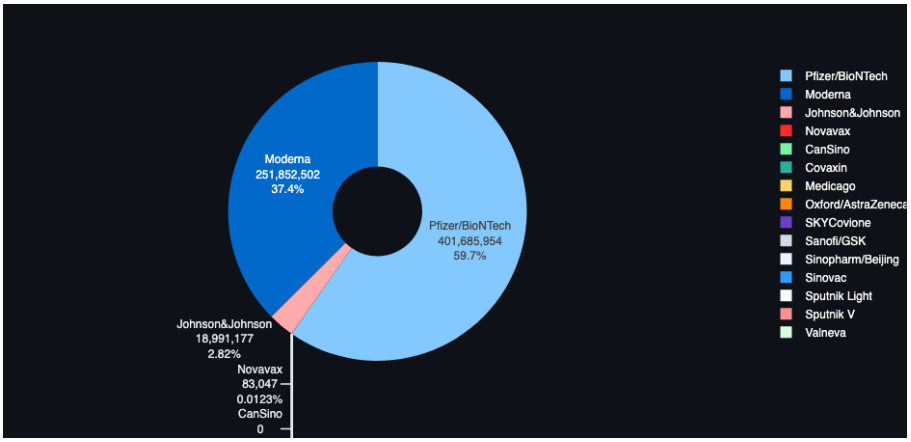


Figure 17: Vaccine Market Share in United States.

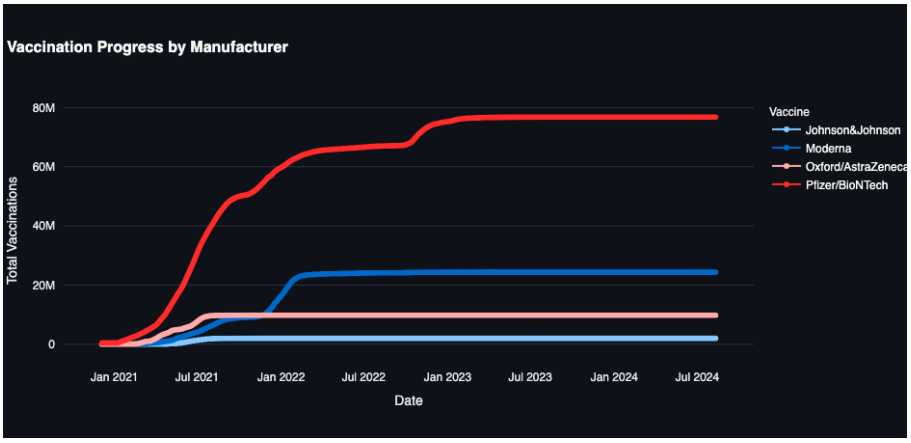


Figure 18: Vaccine Manufacturer-Specific Progress in Spain.

2.4 Task 4 : Testing Impact Analysis

2.4.1 Overview

This task focuses on analyzing the impact of COVID-19 testing globally by exploring the relationship between testing rates, new cases, and deaths. The objective was to understand how testing efforts influenced the detection and management of the pandemic across different regions.

An interactive dashboard was developed using Streamlit and Plotly, presenting multiple visual components such as choropleth maps, bar charts, area plots, and dual-axis graphs. The dashboard allows users to filter by country and date range, view daily and smoothed trends, and compare testing efforts across different countries and continents.

The analysis covers:

- Global and continental testing intensity over time.
- Cross-country comparisons of absolute and population-normalized testing.
- Trends in cumulative testing growth globally.
- Testing efficiency through positivity rates and tests-per-case metrics.

By combining multiple perspectives, the dashboard provides meaningful insights into how testing strategies evolved during the pandemic and how they contributed to outbreak control efforts.

2.4.2 Data Preprocessing

The project started with two datasets — `testing.csv` containing COVID-19 testing data and `cases_deaths.csv` containing new cases and deaths data. Both datasets were loaded using pandas, and the `date` columns were converted to datetime format for consistency.

A **left join** on `country` and `date` was performed to merge them, ensuring all available testing data was preserved. Only essential columns related to testing, cases, deaths, and their per capita versions were retained to focus the analysis.

To handle missing values:

- Rows missing critical values (`new_cases`, `new_deaths`, `new_cases_per_million`, `new_deaths_per_million`) were dropped to maintain the reliability of the analysis.
- Missing `total_tests` and `total_tests_per_thousand` were forward-filled within each country to preserve data continuity.
- Missing `new_tests` and `new_tests_per_thousand` were filled with country-wise median values to minimize the introduction of bias.

Finally, a **continent** column was added by manually mapping each country to its continent. This allowed exploration of testing trends not just globally but also at the continental level.

The cleaned dataset provided a strong foundation for analyzing the relationship between testing efforts, new cases, and deaths.

2.4.3 Visualization Components and Design Justifications

Global Overview Tab

- **Choropleth Map – Global Testing Intensity Over Time:**
  The animated choropleth map visualizes how COVID-19 testing rates per 1,000 people changed globally over time. This helps users easily identify testing intensity across different countries and periods. Choropleth maps are ideal for simultaneously representing geographic and time-series data.

- **Bar Charts – COVID-19 Testing Efforts Across Continents:**
  Three bar charts allow selection between "Total Tests," "Tests per Thousand," and "Tests per Million."

  - **Total Tests** shows absolute testing efforts by continent.
  - **Tests per Thousand/Million** show normalized rates, reflecting healthcare infrastructure strength and people's health-seeking mentality.

  Bar charts are chosen to allow easy side-by-side magnitude comparison.

- **Bar Chart – Top Countries by Total Tests:**
  This bar chart displays countries with the highest absolute number of COVID-19 tests. Although absolute values depend on population size and do not directly reflect healthcare quality, it helps users understand large-scale testing efforts globally.

- **Bar Chart – Top Countries by Tests per 1,000 People:**
  This plot highlights countries with the highest testing relative to their population, indicating better testing intensity and healthcare system responsiveness. It complements the absolute total tests chart for deeper insight.

- **Bar Chart – Bottom Countries by Tests per 1,000 People:**
  This chart identifies countries with the least testing relative to population size, highlighting healthcare gaps and under-testing risks.

- **Area Chart – Global Cumulative Tests Over Time:**
  The area plot shows how cumulative global testing scaled over the pandemic timeline. It helps users visualize the progression and intensity of testing expansion during different phases of the pandemic.

Compare Multiple Countries Tab

- **Multi-Line Chart – New Tests and New Tests per Thousand Comparison:**
  Multi-line charts allow comparison of how daily new testing efforts evolved across multiple countries. It helps users visualize healthcare responses over time both in absolute and per-population terms.

- **Line + Marker Chart – Total Tests and Total Tests per Thousand Over Time:**
  This plot tracks cumulative testing growth for selected countries, highlighting long-term testing efforts. Using line + marker styling improves traceability even when multiple countries overlap in the graph.

Country-Specific Insights Tab

- **Toggle Line Chart – Trends in Cases, Tests, and Deaths (7-Day Average vs Daily Counts):**
  The toggle between 7-day moving average and daily counts provides flexible trend exploration. Comparing cases, tests, and deaths together helps understand how testing influenced detection and outbreak severity. Yellow markers highlight peak pandemic periods based on case and death trends.

- **Dual Axis Line Chart – Testing Efficiency (Tests per Case and Positivity Rate):**
  This chart compares the 7-day moving averages of tests-per-case and positivity rates using dual Y-axes. It captures how effectively a country tested throughout the pandemic, revealing trends in infection rates and testing coverage.

- **Donut Chart – Overall Average Positivity Rate:**
  The donut chart summarizes the average positivity rate across the full selected period. It clearly shows the proportion of positive vs. negative tests, offering a quick view of testing adequacy over time.

2.4.4 Usability Enhancements

- **Sidebar Filters:**
  Dropdowns, multiselects, and date range pickers were added in the sidebar, allowing users to dynamically filter by country and time period, ensuring flexible and interactive data exploration.

- **Dynamic Metric Selection:**
  Users can select different metrics such as "Total Tests," "Tests per Thousand," or "New Tests per Thousand" for detailed comparisons, enabling multiple perspectives on testing efforts.

- **Responsive Layout:**
  Streamlit's wide layout and auto-scaling Plotly charts ensure that visualizations adjust to different screen sizes, improving user experience across devices.

- **Customized Hover Tooltips:**
  Hover labels on charts display well-formatted numbers (commas for thousands, percentage formats), improving readability and quick understanding without clicking.

- **Color Palette Selection:**
  High-contrast and intuitive color palettes (blue for healthcare metrics, green for high performance, red for poor performance) were applied to improve chart interpretability.

- **Peak Highlighting:**
  In the country-specific trends chart, peak pandemic periods were automatically detected and highlighted with yellow diamond markers to draw attention to critical phases.

2.4.5 Results

The final dashboard offers an intuitive and interactive way to explore COVID-19 testing trends globally, regionally, and nationally. Key findings from the visualization components include:

- **Testing Efforts Varied Geographically:**
  The global choropleth map and continent-level bar charts revealed significant differences in testing intensity between continents and countries. Regions like Europe and North America exhibited higher testing per thousand people compared to Africa and parts of Asia.

- **Absolute vs. Per-Population Testing:**
  Comparing top countries by total tests and tests per thousand showed that large countries conducted the most tests absolutely, but smaller countries often achieved higher per-capita testing rates, highlighting different public health strategies.

- **Cumulative Growth in Testing:**
  The area chart demonstrated the steep increase in cumulative testing worldwide, especially during major pandemic waves, emphasizing global efforts to improve testing access over time.

- **Cross-Country Testing Trends:**
  Multi-country comparison charts indicated that different countries followed different testing trajectories, with some scaling testing aggressively early on, while others expanded testing later as outbreaks worsened.

- **Testing and Case Detection Relationship:**
  Country-specific trend analysis showed that increases in testing generally corresponded with rises in detected cases, confirming the role of testing in pandemic management and outbreak monitoring.

- **Testing Efficiency Insights:**
  Analysis of positivity rates and tests per case highlighted the efficiency of testing systems. Higher tests-per-case ratios and lower positivity rates suggested more widespread and effective testing.

- **Overall Positivity Rates:**
  Donut charts summarized the positivity rates over the pandemic, providing a quick visual assessment of how adequately each country tested relative to its outbreak severity.

2.4.6 Visualization Snapshots

To better illustrate the developed dashboard, key visualizations are shown below.

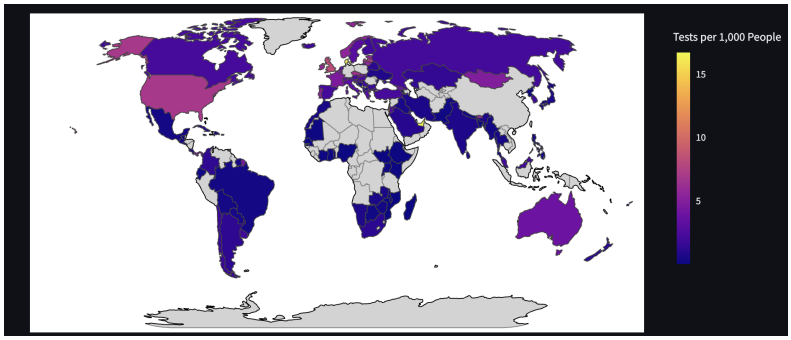


Figure 19: Choropleth Map – COVID-19 Testing Intensity Over Time, highlighting regional disparities and testing progression during major pandemic waves.

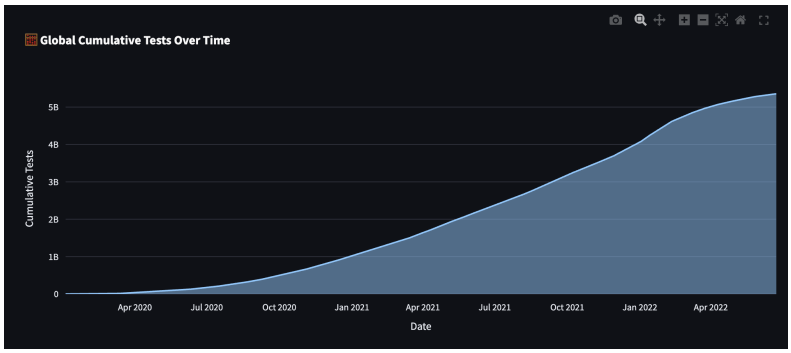


Figure 20: Area Chart – Global Cumulative Tests Over Time ,illustrating how testing scaled dramatically during the pandemic's critical phases.

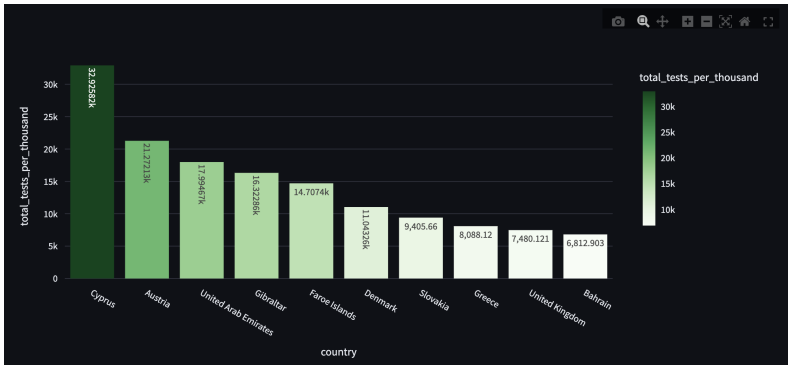


Figure 21: Bar Chart – Top Countries by Tests per 1,000 People,A bar chart ranking countries based on population-normalized testing rates.

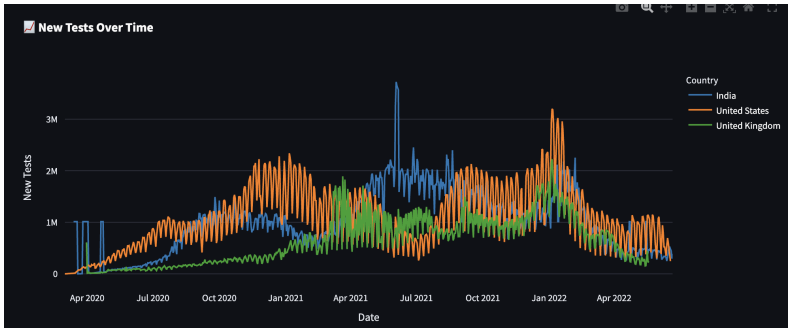


Figure 22: Multi-Line Chart – New Tests Over Time for Selected Countries, The graph highlights testing patterns across different pandemic waves, enabling cross-country comparison of testing intensity and response timing.

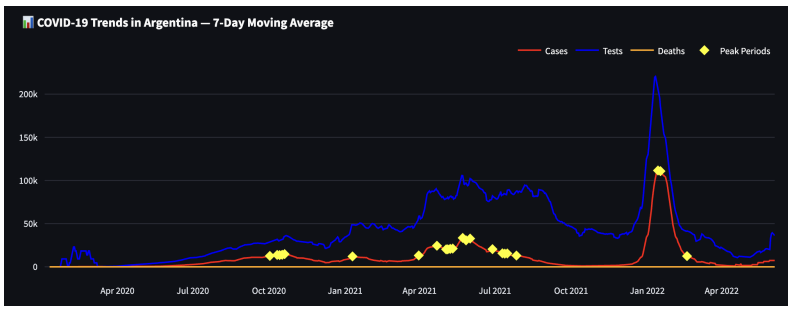


Figure 23: Toggle Line Chart – Country-Specific Trends(Cases, Tests, Deaths), with peak periods highlighted for deeper pandemic analysis.

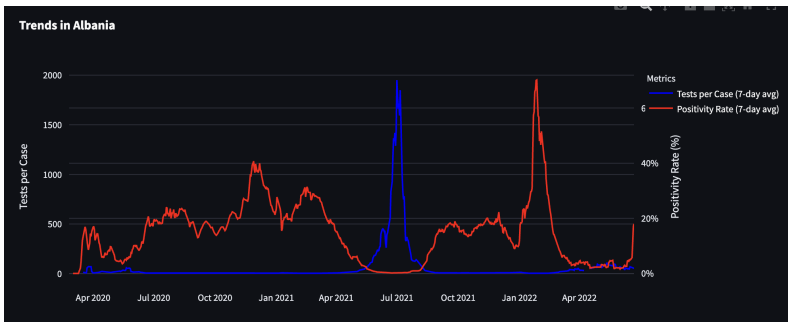


Figure 24: Dual Axis Line Chart – Testing Efficiency (Tests per Case vs Positivity Rate),A dual-axis line plot showing how testing efficiency evolved over time.

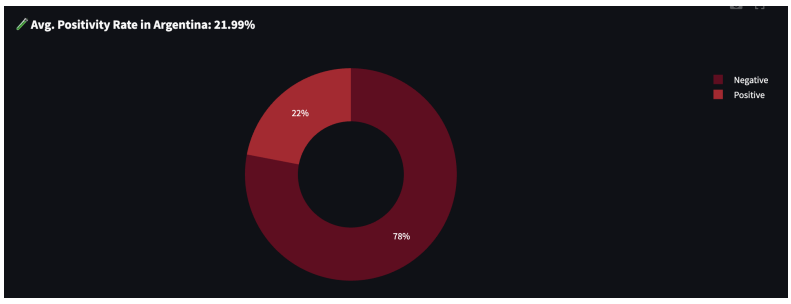


Figure 25: Donut Chart – Overall Average Positivity Rate across the selected period, offering a clear and quick visual of testing adequacy.

2.5 Task 5: Google Mobility Analysis

2.5.1 Overview

In this project, I was responsible for building the core visualization components of the COVID-19 Mobility Analysis dashboard using Streamlit and Plotly. My main goal was to transform a cleaned dataset — combining mobility trends, pandemic severity (case counts), and government response measures — into an interactive and intuitive platform. Through thoughtful design and clear storytelling, I aimed to help users explore how human movement evolved over time, across different regions, and under varying policy conditions throughout the pandemic.

2.5.2 Data Preprocessing

I began by loading the cleaned, merged dataset stored in Parquet format. To enable better temporal analysis, I derived additional features such as `month` and `year` from the date field. Missing values were carefully handled through selective filtering and aggregation, ensuring that the resulting charts remained accurate, readable, and reliable.

2.5.3 Visualization Components and Design Justifications

Global Overview Tab

- **Line Chart – Global Mobility Over Time:** A line chart was used to capture daily trends in mobility across the world. Line charts are well-suited for showing continuous changes over time, making it easy to spot pandemic-induced dips and periods of recovery.
- **Bar Chart – Yearly Comparison:** To compare mobility averages across different years, a bar chart was chosen instead of a pie chart. Bar charts present side-by-side comparisons more clearly, allowing for easier interpretation of small year-to-year differences.
- **Treemap – Mobility vs. Policy Stringency:** A treemap was selected to show two dimensions simultaneously — country size (mobility) and color intensity (policy strength). This compact visualization helped surface patterns across many countries at once.

Top Countries Tab

- **Pareto Chart – Mobility vs. COVID-19 Cases:** A dual-axis chart was used, combining bars for total case counts and a line plot for mobility percentages. This structure allowed viewers to easily compare mobility levels alongside the pandemic's severity. Mobility was plotted on the left Y-axis (as a percentage) and cases on the right (as counts) for clarity.
- **Funnel Chart – Top Mobility Rankings:** Instead of a basic bar chart, a funnel plot was used to highlight rankings and relative drop-offs between countries' average mobility. Funnels emphasize ordering naturally, making the visualization both attractive and intuitive.

Single Country Tab

- **Dual Axis Chart – Mobility vs. Cases Over Time:** For individual country analysis, a dual-axis line chart allowed users to explore the relationship between mobility fluctuations and COVID-19 case surges, with separate Y-axes for movement and infection counts to preserve scale and clarity.

Multi-Country Tab

- **Line Chart – Monthly Trend Comparison:** Aggregating data to the monthly level helped smooth out noise and made it easier to spot patterns across countries. A line chart was ideal for comparing several countries' mobility trends over time.
- **Bar Chart – Yearly Average Mobility:** A bar chart was again chosen here to rank countries by their annual average mobility. Unlike pie charts, bar charts maintain clarity even when many categories (countries) are involved and when values are close.

2.5.4 Usability Enhancements

To make the dashboard user-friendly and informative, several usability improvements were implemented:

- Customized hover tooltips for all charts, displaying percentages and large numbers in an easy-to-read format.
- Clear annotations and warnings, such as noting that Google Mobility data was discontinued after October 15, 2022.
- Interactive controls, including dropdowns, date-range sliders, and country multi-selectors, enabled users to filter and zoom into specific areas of interest.
- A consistent visual style was applied throughout, ensuring coherent legends, labels, and plot layouts for a seamless experience.

2.5.5 Results

The final visualization platform provided an effective way to explore global, national, and temporal trends in human mobility during the pandemic. Key insights uncovered include:

- A sharp global decline in mobility in early 2020, followed by uneven recovery patterns across regions.
- Some countries maintained high levels of mobility despite recording large case numbers, revealing differences in policy enforcement and public compliance.
- Countries with stricter government policies generally saw larger reductions in movement, as visualized through treemaps.
- Dual-axis charts clearly illustrated the link between COVID-19 waves and movement restrictions within individual countries.

Overall, the project successfully delivered a meaningful, visually rich, and interactive analysis that allowed users to explore the complex relationship between mobility, policies, and pandemic dynamics.

2.5.6 Visualization Snapshots

To better illustrate the developed dashboard, key visualizations are shown below.

Figure 26: Global Average Mobility Index Over Time highlighting COVID-19 lockdown impacts and recovery phases.

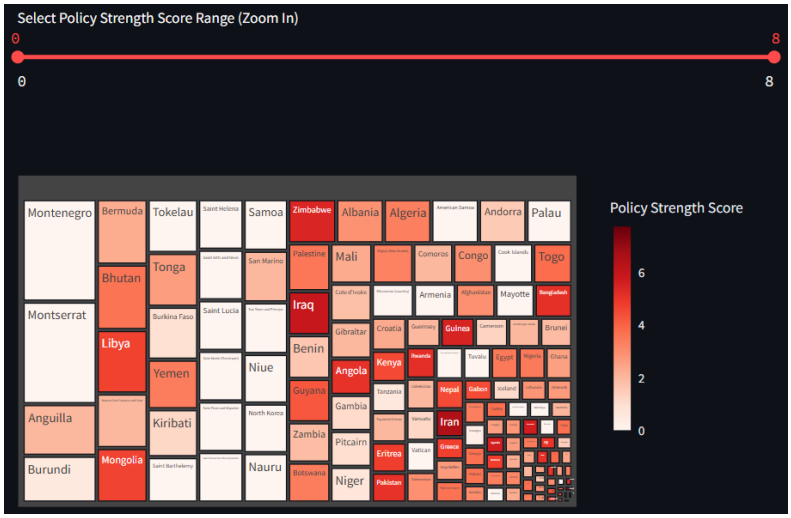


Figure 27: Treemap showing relationship between policy strictness and mobility reductions across countries.

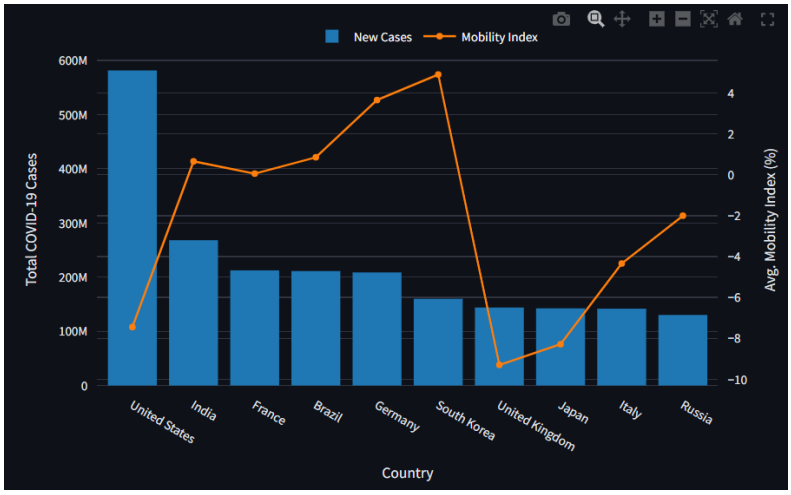


Figure 28: Pareto Chart showing the relationship between total COVID-19 cases and mobility index for the top affected countries.

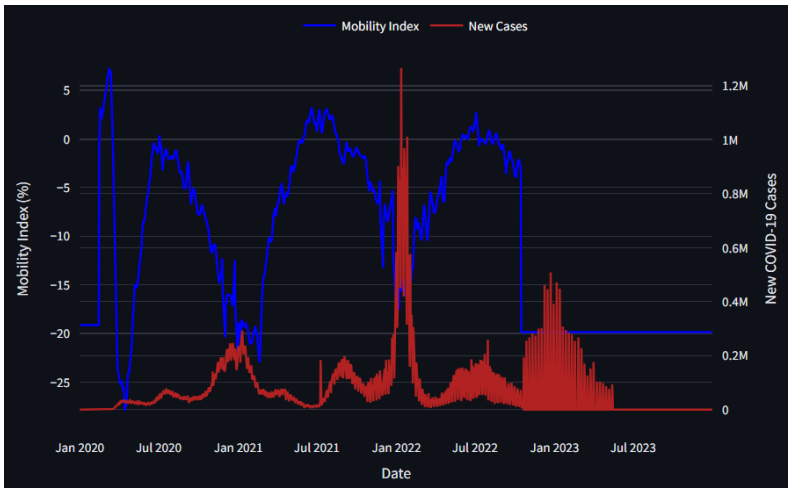


Figure 29: Dual-axis line chart comparing Mobility Index and New COVID-19 Cases for a selected country over time.

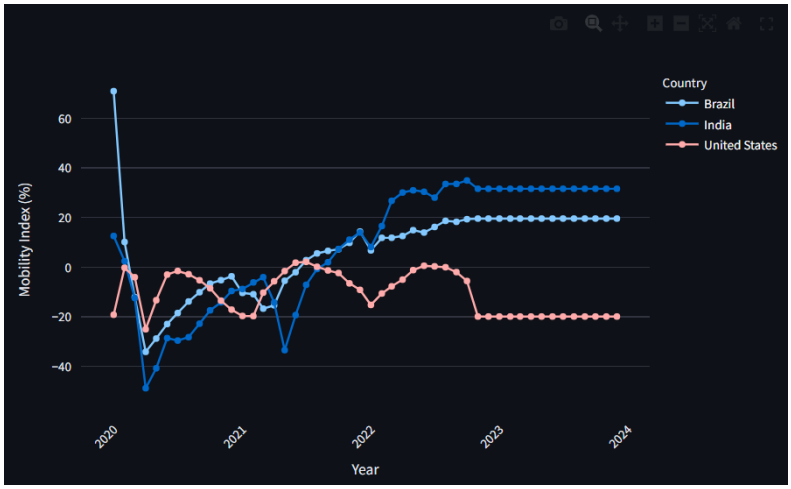


Figure 30: Line chart comparing the yearly mobility trends across multiple selected countries.

2.6 Task 6 : Analyzing COVID-19 Impacts in India

2.6.1 Overview

The primary goal of this project was to develop an interactive data visualization platform to explore and analyze the impact of COVID-19 in India, focusing on both state and district levels. The analysis spans from April 26, 2020, to February 1, 2022. The aim was to track the temporal progression of COVID-19 cases, recoveries, and deaths, across different states and districts, while offering users an intuitive interface to interact with the data through Streamlit and Plotly visualizations.

Key objectives:

- Analyze the spread of COVID-19 at the state and district levels.
- Assess age, gender, and hospitalization outcomes during the early months of the pandemic.
- Create engaging and interactive visualizations to convey meaningful insights.
- Justify the design choices for visualizations, ensuring clarity and ease of comparison.

2.6.2 Data Preprocessing

- COVID-19 data for states and districts (confirmed, recovered, deceased) were collected.
- Temporal features such as day, month, and year were extracted.
- Missing values were handled and removed.
- Data was aggregated at daily, monthly, and cumulative levels for different analysis stages.

2.6.3 Visualization Components and Design Justifications

State and District Level Analysis

- **Bar Charts:** Used to compare the total cases across states and districts, as bar charts are effective for comparing quantities across multiple categories.
- **Line Charts:** Plotted daily and monthly case counts to highlight significant waves and trends in COVID-19 spread.
- **Scatter Mapbox Plots:** Displayed district-level case intensities using point markers on a geographic map. The size and color of each bubble indicated case numbers. Scatter Mapbox was selected for its flexibility and ease of interpretation.

Demographic Analysis

- **Sunburst Charts:** Visualized the hierarchical relationships between age group, gender, and COVID-19 outcomes (hospitalized, recovered, deceased), providing a clear view of demographic impacts.

Design Choices

- Bar charts were selected to compare numerous categories in a clear manner.
- Line charts were chosen to depict trends over time effectively.
- Scatter Mapbox plots were preferred for visualizing geographic distribution based on case numbers.
- Sunburst charts were used to display hierarchical data relationships in an easy-to-understand format.
- A consistent color palette, along with well-designed tooltips and legends, ensured clarity and usability across the visualizations.

2.6.4 Results and Insights

Temporal Trends

- Two distinct waves of COVID-19 were evident: one in September 2020 and another in May 2021.

Geographic Spread

- Urban districts, especially metros, had a higher concentration of cases.
- Scatter Mapbox plots helped identify regional hotspots of the pandemic.

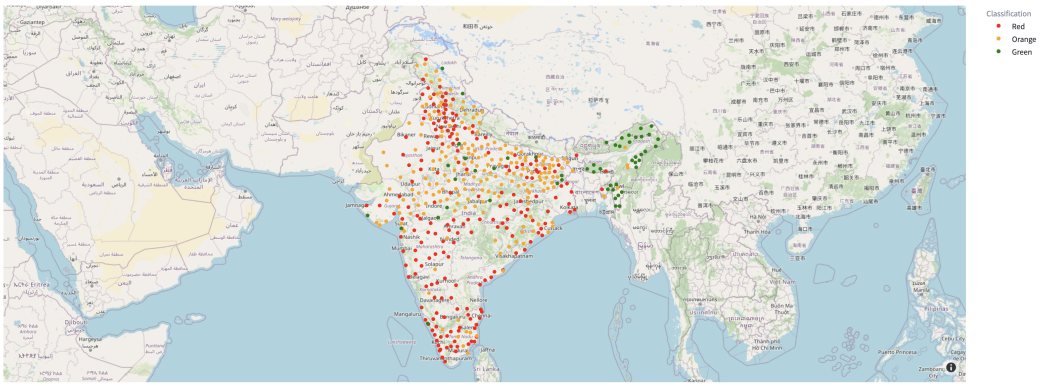
Demographic Patterns

- Older age groups exhibited higher rates of hospitalization and mortality.
- Males were marginally more affected compared to females.

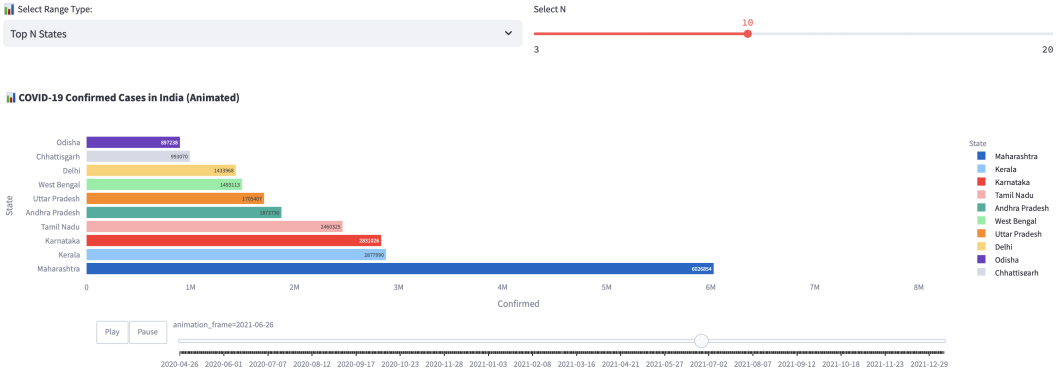
Benefits of Preprocessing

- Data cleaning ensured clearer visualizations by removing inconsistencies and errors.
- Aggregation of data at different levels facilitated smooth interaction and exploration of large datasets.

COVID-19 District Zone Classification - May 2021

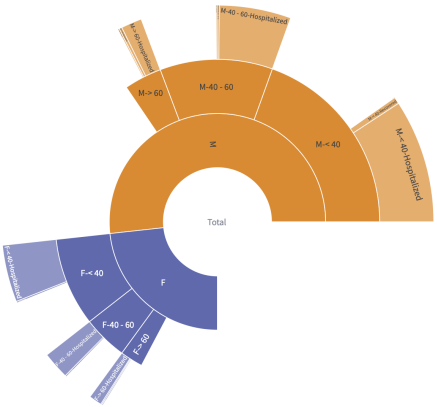


COVID-19 Confirmed Cases - Animated Race Chart



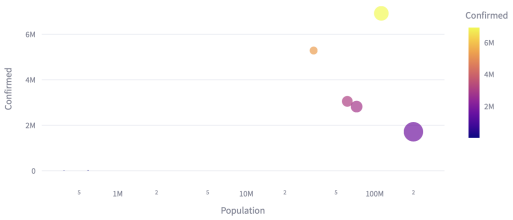
COVID-19 Age & Genderwise infection

COVID-19 Sunburst: Gender → Age Group → Status

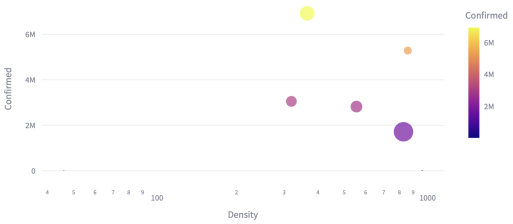


COVID-19 Impact Analysis
Confirmed Cases

Confirmed Cases vs Population



Confirmed Cases vs Population Density



2.7 Task 7: Day- Wise Visualization

2.7.1 Overview

As part of COVID LENS dashboard, the primary goal of us in this module was to develop the *Day-wise COVID-19 Statistics Visualization Page*. On any given date, our module provides an interactive and intuitive way for users to explore country - level COVID-19 metrics. The primary goal was to enable a single - day snapshots of various indicators on selected date, such as new cases, new deaths, ICU occupancy, daily vaccinations, new people vaccinated, active cases, recovered cases and recovery rate over time - across all countries.

2.7.2 Data Preprocessing

To achieve the tasks goal, we provided a solution involving custom data processing, data interpolation, estimation techniques and a responsive frontend visualization interface built using Streamlit and Plotly.

We created a three Python script, each targeting a specific data processing goal:

1. Cleaned the dataset by removing aggregated regions, filling missing values in cumulative counts, calculating daily values using 'diff()', and interpolating zeros with calculated values. Negative values were clipped, and the final dataset was streamlined for analysis.

2. Estimated recovery data based on a 14-day lag assumption. Calculated the active case by subtracting the total deaths from total cases and recoveries were estimated using past data hence addressing the common issue of missing recovery metrics in many countries.

3. Processed the original datasets, extracted all the relevant features, and then computed new daily vaccination metrics. We merged the datasets on country and date, creating a comprehensive dataset containing vaccination and ICU load indicators.

Dataset Name	Rows	Key Columns	Description
daily_analysis_data.csv	354,537	new_cases, new_deaths, total_cases, location	Cleaned, interpolated case/death dataset
active_cases_and_estimated_recovery_data.csv	361,250	active_cases, estimated_recovered, recovery_rate	Adds estimated recovery and active case statistics
daily_vaccinations_and_icu_all_countries_data.csv	68,876	daily_vaccinations, ICU occupancy, full vaccination	Merged view of immunization progress and health burden (ICU load)

Table 1: Summary of processed datasets used in the COVID-19 dashboard project.

2.7.3 Visualization Components and Design Justifications

Our visualization includes a wide range of rich, interactive features aimed at enhancing exploration, clarity, and insight generation. Here users can select any specific date using a date picker and immediately view filtered visualizations that reflect the pandemic situation for that selected day. The module offers filter modes such as showing all countries, top 10 countries by new cases or new deaths, or allowing custom country selections via multi-select dropdowns.

Daily case and Death analysis

- **Pie charts** identifying the top 10 countries by peak daily new cases and deaths, including hover tooltips showing exact numbers and dates.
- **Bubble charts** that map countries by daily new cases (x-axis) and deaths (y-axis), with bubble size indicating severity.
- An **animated bar race** showing how new cases and deaths evolve over time for the top 10 countries, complete with time sliders and play/pause controls.

Vaccination & Recovery Analysis

- **Bar charts** for daily vaccinations, new partially vaccinated, and new fully vaccinated individuals.
- **ICU occupancy visualizations** for countries reporting this metric.
- **Grouped bar charts** comparing real-time active cases with estimated recovered counts per country.
- **A monthly animated choropleth map** visualizing estimated recovery rates across the world.
- **A line chart with slider and range selectors** showing the evolution of recovery rates over time across selected countries.

2.7.4 Results

This module effectively fulfilled its objectives by enabling detailed, per-day analysis of COVID-19 statistics through a comprehensive set of interactive tools. Designed for efficiency, it supports both small and large datasets through performance enhancements such as data caching, lazy filtering, and preprocessing steps. Rigorous data cleaning ensured the reliability and accuracy of all visual outputs.

The module offers a variety of visualization formats, including:

- Pie charts to highlight peak daily cases and deaths.
- Bubble plots to compare cases and deaths by country.
- Animated bar charts showcasing the evolution of top-affected countries over time.
- Bar graphs for vaccination progress and ICU occupancy.
- Grouped bar comparisons between active cases and estimated recoveries.
- Line graphs to track recovery rates over time.
- Monthly animated choropleth maps illustrating global recovery dynamics.

With these features, users can examine daily statistics for selected countries or global trends, gain insight into the spread and intensity of the pandemic, and assess the effectiveness of national responses through vaccination efforts and healthcare capacity management.

2.7.5 Visualization Snapshots

To provide better insights, the following snapshots from the dashboard are included.

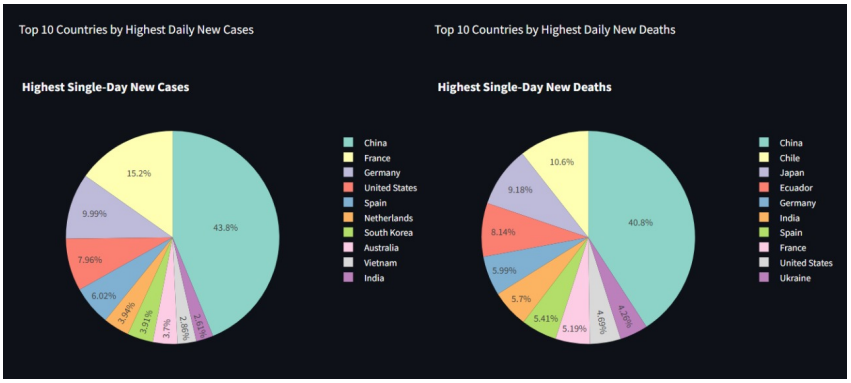


Figure 31: Pie charts illustrating the distribution of the highest recorded single-day new cases and deaths across the top 10 affected countries.

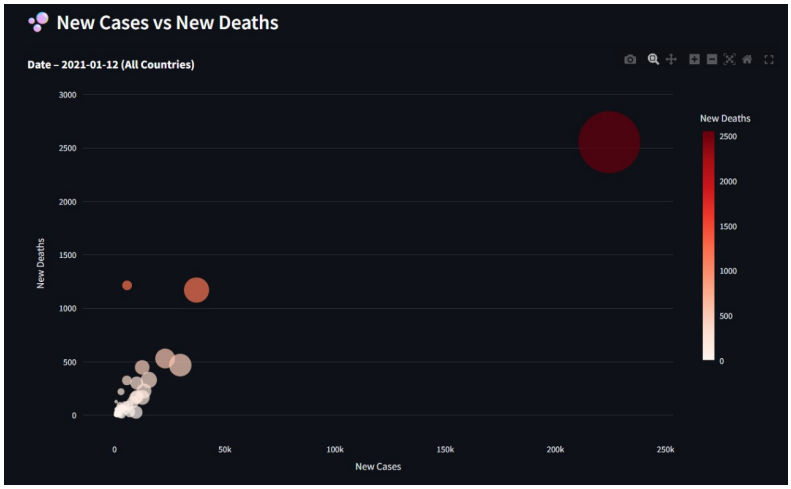


Figure 32: Scatter plot visualizing the relationship between new cases and new deaths across countries on a specific date.

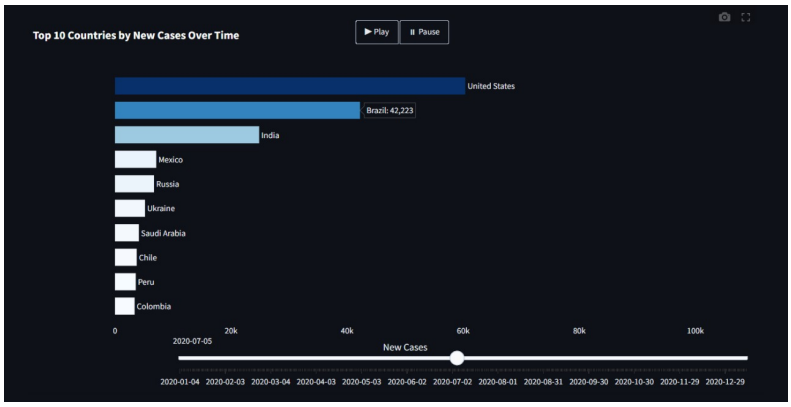


Figure 33: Animated Bar chart visualizing the top 10 countries by new cases over time, showing dynamic changes in case numbers throughout the pandemic on different dates.

Figure 34: Animated Bar chart visualizing the top 10 countries by new deaths over time, highlighting the dynamic changes in mortality trends throughout the pandemic on different dates.

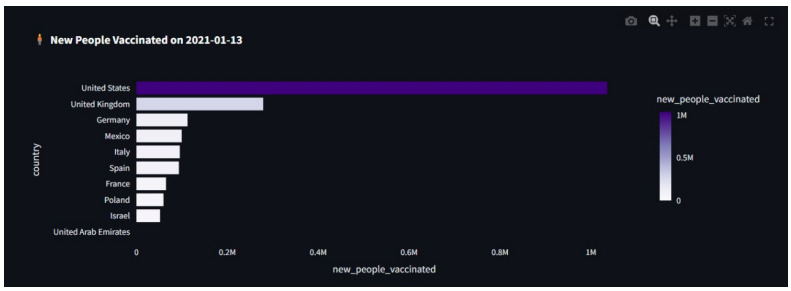


Figure 35: Bar chart displaying the number of new people vaccinated across different countries on a specific date, highlighting the global vaccination efforts.

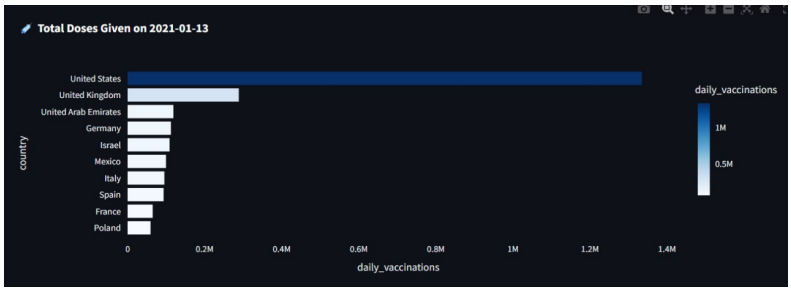


Figure 36: Bar chart showing the total number of vaccine doses administered across different countries on a specific date.

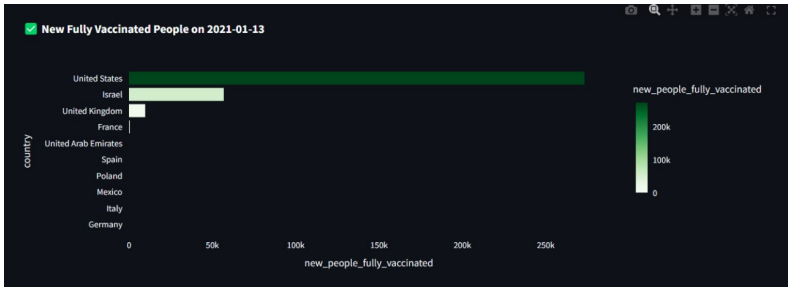


Figure 37: bar chart showing the number of newly fully vaccinated individuals across countries on a specific date.

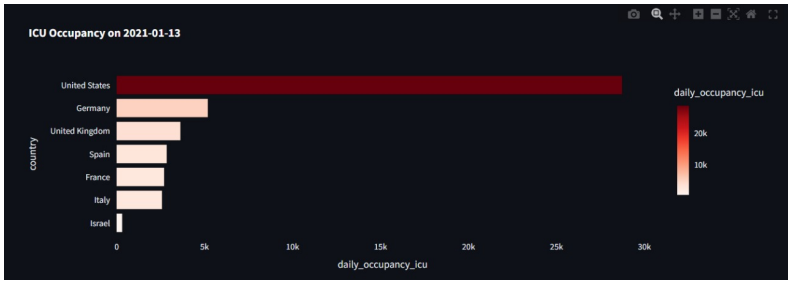


Figure 38: Bar chart illustrating the daily ICU occupancy across different countries.

Figure 39: Choropleth map showing monthly estimated recovery rates across countries over time.

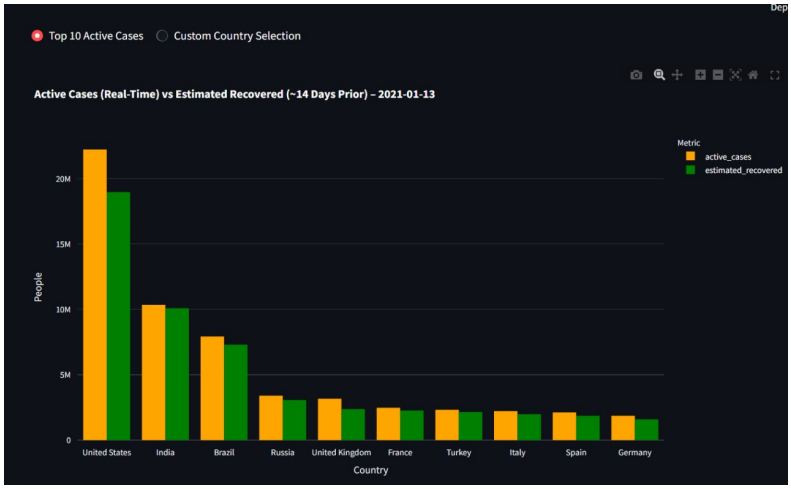


Figure 40: Bar chart comparing real-time active cases with estimated recoveries for top-affected countries.

Figure 41: Line graph visualizing the recovery rate trends over time for selected countries.

3. Conclusion:

In this project, we have developed web-based dashboard to visualize the key aspects of the COVID-19 pandemic with the data from "Our World in Data (OWID)". Our dashboard provides a dynamic platform for exploring trends in disease spread, vaccination efforts, testing, mortality and mobility patterns.

The Disease Spread and Mortality Impact analyses identified major hotspots and tracked their progression over time. The Vaccination and Testing Impact analyses offered important insights into the effectiveness of intervention strategies in controlling the pandemic's severity and spread. The Google Mobility Analysis further demonstrated how public health measures, such as lockdowns and movement restrictions, influenced the transmission dynamics.

A day-wise analysis and also detailed study of COVID-19 trends in India enhanced the understanding of localized outbreaks and short-term variations. By enabling country-specific filtering, time-based exploration, and comparative analysis, the dashboard serves as a valuable resource for researchers, policymakers, and the wider public.

Ultimately, this project provides a comprehensive tool to support informed decision-making during global health crises by making COVID-19 data more accessible, interpretable, and actionable through visual analytics.

4. Link to source code

Click here to access Source Code

[1] Python: https://www.python.org

Pandas: https://pandas.pydata.org

Numpy: https://numpy.org

Streamlit: https://streamlit.io

Dataset link: https://docs.owid.io/projects/etl/api/covid/download-data