A PROJECT REPORT
ON
COVID-19 CASES

**INFO 5709 - Data Visualization**
**&**
**Communication**

**Submitted by:**

**Kaustubh Vora**

**11689537**

## Introduction:

The COVID-19 pandemic has been an unprecedented global event affecting millions of lives and altering the course of economies and societies around the world. The ability to visualize and analyze the data associated with the pandemic's progression is crucial for understanding its impact and informing public health decisions. This report details the creation of a comprehensive data visualization dashboard using Tableau Public that illustrates various aspects of the COVID-19 data from January 2020 to December 2022. Our visualizations aim to elucidate patterns, trends, and current states within the pandemic's lifecycle, facilitating a data-driven approach to comprehend the spread and effects of the virus globally.

## Description of Dataset:

The visualization project utilizes two primary datasets, **worldometer_coronavirus_summary_data.csv** and **worldometer_coronavirus_daily_data.csv**, which are extensive compilations of COVID-19-related metrics.

1. **Summary Dataset (worldometer_coronavirus_summary_data.csv)**: This dataset provides a summarized snapshot of the COVID-19 situation globally, with each row representing a different country. Key columns include:
    - **country**: Name of the country or territory.
    - **continent**: Continent where the country is located.
    - **total_confirmed**: Total number of confirmed COVID-19 cases.
    - **total_deaths**: Cumulative number of deaths attributed to COVID-19.

- **total_recovered**: Number of individuals who have recovered from the virus.
- **active_cases**: Current number of active cases.
- **serious_or_critical**: Number of serious or critical cases.
- **total_cases_per_1m_population**: Total confirmed cases per one million people.
- **total_deaths_per_1m_population**: Total deaths per one million people.
- **total_tests**: Total number of COVID-19 tests conducted.
- **total_tests_per_1m_population**: Number of tests conducted per one million people.
- **population**: Estimated population of the country.

The summary dataset is instrumental in providing a high-level view of the pandemic's reach and its current status across the globe. It allows for the assessment of overall impact, testing efforts, and the burden on healthcare systems.

2. **Daily Dataset (worldometer_coronavirus_daily_data.csv)**: This dataset captures the day-to-day development of the pandemic for each country, providing a time-series perspective. Vital columns include:
   - **date**: The date of the data record.
   - **country**: The country for which the data is reported.
   - **cumulative_total_cases**: The total number of COVID-19 cases reported up to that date.
   - **daily_new_cases**: The number of new COVID-19 cases reported on that date.
   - **active_cases**: The current number of active cases on that date.

- **cumulative_total_deaths**: The cumulative number of deaths up to that date.
- **daily_new_deaths**: The number of new deaths reported on that date.

With daily granularity, this dataset is pivotal for understanding the trajectory of the virus's spread. It provides insights into how case and death rates have evolved over time and is essential for trend analysis and pattern recognition.

The combination of these datasets enables a comprehensive analysis. While the summary dataset provides a static overview, the daily dataset adds dynamic context, revealing how the pandemic has unfolded over time. Together, they offer a multi-faceted view of the COVID-19 pandemic, from macroscopic global trends down to specific daily figures that can inform both public understanding and policy-making.

## Data Cleaning and Organization:

The foundation of any robust data visualization project is clean and well-organized data. For this analysis, we underwent a meticulous data preparation phase to ensure the integrity and usability of the datasets provided.

### Cleaning Steps:

1. **Initial Assessment**: We began by loading the datasets and performing an initial assessment to understand their structure, identify missing values, spot any obvious inaccuracies, and get a sense of the data types we were dealing with.

2. **Handling Missing Values**: Our datasets, particularly the daily data, had missing values, especially in the early stages of the pandemic when data collection was inconsistent. We addressed these missing values appropriately, which included imputing some based on available data trends and omitting others where data imputation would not have been reliable.
3. **Data Type Correction**: We ensured that each column was formatted with the correct data type (e.g., dates as datetime objects, numerical values as integers or floats, and categorical values as strings).
4. **De-duplication**: We checked for and removed any duplicate records to prevent skewed analysis, ensuring that each entry in the dataset represented a unique data point.
5. **Consistency Checks**: We standardized the naming conventions across datasets to maintain consistency. For instance, country names were regularized to align with global standards, allowing for accurate merging and comparison.
6. **Aggregation and Summarization**: For the summary dataset, we verified the aggregated values such as total cases and deaths to ensure that they matched the cumulative totals derived from the daily dataset.


summary_data.csv


Daily_data.csv

**Organizing Steps:**

1. **Sorting and Filtering**: We sorted the data where necessary to facilitate easier analysis and filtered out irrelevant records, such as entries for cruise ships or territories without recognized sovereignty.

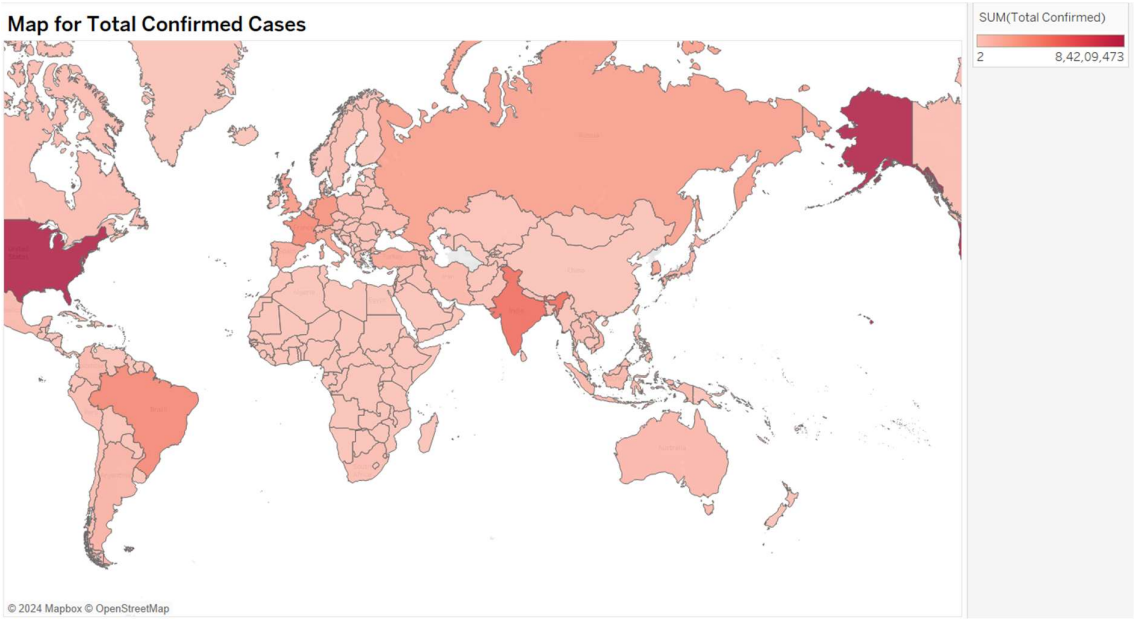
2. **Structural Adjustments**: We pivoted some columns to create a more analysis-friendly structure, particularly in the daily dataset where we wanted to analyze trends over time.
3. **Creating Calculated Fields**: We generated new calculated fields that were not originally part of the datasets but were crucial for our analysis, such as active case rates, mortality rates, and recovery rates.
4. **Segmentation**: We segmented the data into relevant groups, such as by continent or by economic status, to enable more targeted analyses within our visualizations.
5. **Joining Datasets**: Where necessary, we joined the daily and summary datasets on the country column to enrich our data and allow for a multi-dimensional analysis.

Through these cleaning and organizing steps, we addressed data quality issues that could compromise our analysis and visualizations. This rigorous preparation paved the way for the subsequent visual analytics, ensuring that our findings were grounded in the most accurate and relevant data possible.

**https://public.tableau.com/views/KaustubhVora_Project/Dashboard1?:language=en-US&publish=yes&:sid=&:display_count=n&:origin=viz_share_link**

**The link provided is a URL to a published Tableau Public visualization or dashboard titled "KaustubhVora_Project".**

## Global COVID-19 Map (Choropleth Map):




**Objective:** The primary objective of this visualization was to present a global overview of the impact of COVID-19 by illustrating the total confirmed cases or deaths across different countries. This map provides an immediate visual representation of the pandemic's distribution and intensity around the world.

## Analytical Insights:

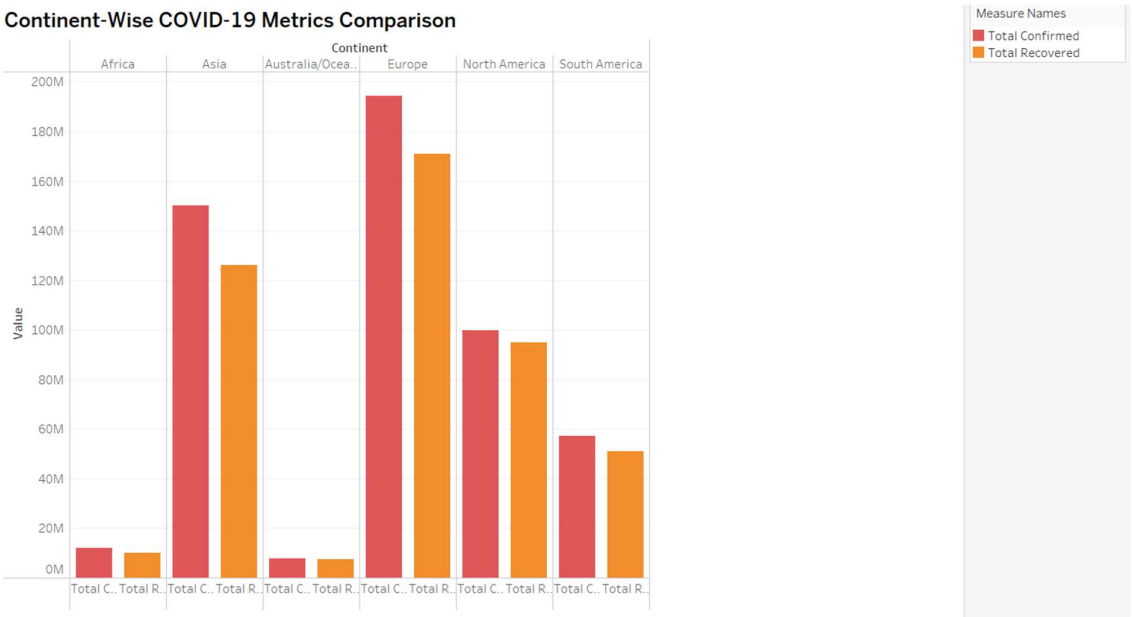
- The map showed that countries with larger populations, like the United States, India, and Brazil, had higher absolute numbers of confirmed cases and deaths, which could be partially attributed to their population size.
- Smaller countries, when considering total cases or deaths per capita, sometimes showed a disproportionately higher impact, underscoring the importance of evaluating data in the context of population density and healthcare capacity.
- The visualization also highlighted regional disparities, with some continents like Europe and the Americas showing darker shades indicative of higher case counts, while Africa

showed lighter shades, indicating fewer reported cases or deaths.

**Design and Creation:**

- **Choropleth Map**: A choropleth map was chosen as it allows for the geographical distribution of the data to be easily understood, with colors indicating the intensity or magnitude of the data (in this case, total confirmed cases or deaths).
- **Color Palette**: A gradient color palette was used, with darker shades representing higher values, enabling a quick visual assessment of which countries are most affected.
- **Interactivity**: The map was designed to be interactive. Hovering over a country reveals precise numbers for total confirmed cases or deaths, providing a detailed view on demand.

**Continent-Wise COVID-19 Metrics Comparison:**




**Objective:**

The goal of this visualization was to compare COVID-19 metrics across different continents to uncover regional differences and trends. By contrasting the total cases confirmed and recovery

rates continent-wise, we aimed to highlight the disparate impacts of the pandemic and the varied responses by different regions.

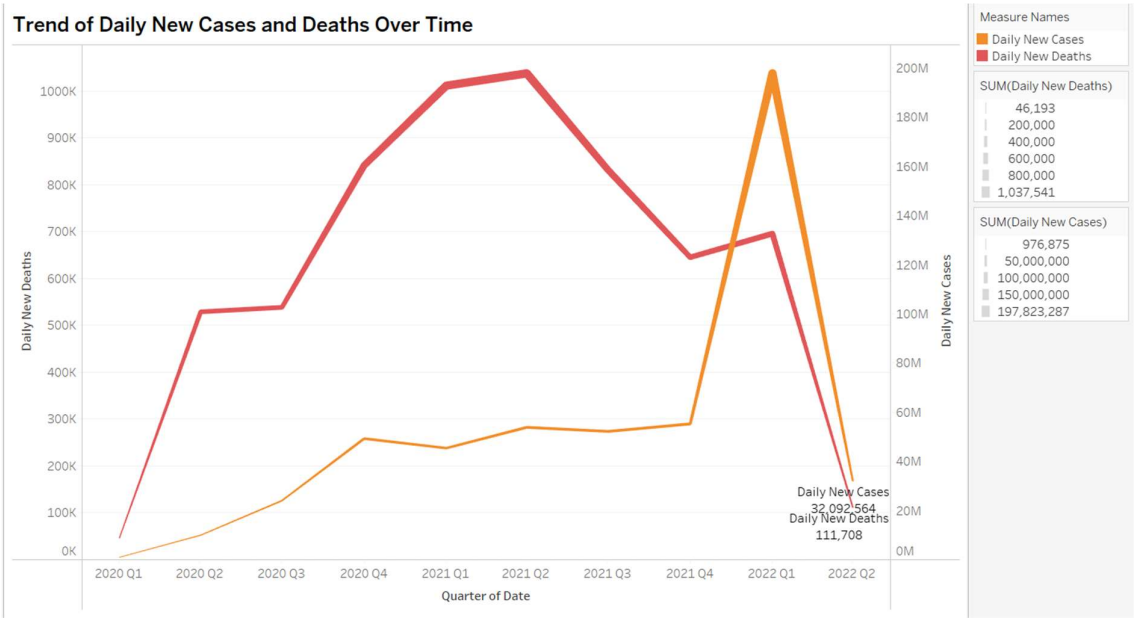
**Analytical Insights:**

- The visualization revealed significant disparities between continents. For instance, at the time of analysis, Asia showed a high number of total confirmed cases due to its large population, but when adjusted for population size, the per capita rates offered a different perspective.
- Europe and North America, despite having smaller populations, exhibited high total confirmed and death counts, indicating severe outbreaks.
- Recovery rates varied widely, with some continents showing a high percentage of recoveries, which could be reflective of the healthcare system's efficiency or differences in reporting standards.

**Design and Creation:**

Bar Charts: We utilized bar charts for their simplicity and effectiveness in comparing numerical data across categories. Each continent represents a category, and the metrics represent the numerical values.
Sorting: Continents were sorted based on the metrics, making it easier to identify which regions had the highest and lowest values.
Color Coding: A distinct color was assigned to each metric to facilitate quick differentiation. The choice of colors was made to ensure accessibility, with consideration for color-blind individuals.
Interactivity: Tooltips were added to provide additional context on hover, such as specific numbers and per capita rates.

**Trend of Daily New Cases and Deaths Over Time:**




## Objective:

The visualization's purpose was to track and compare the progression of COVID-19 through the daily new cases and deaths over time. This trend analysis is crucial for understanding the effectiveness of public health interventions, the virulence of the virus over time, and its impact on different populations.

## Analytical Insights:

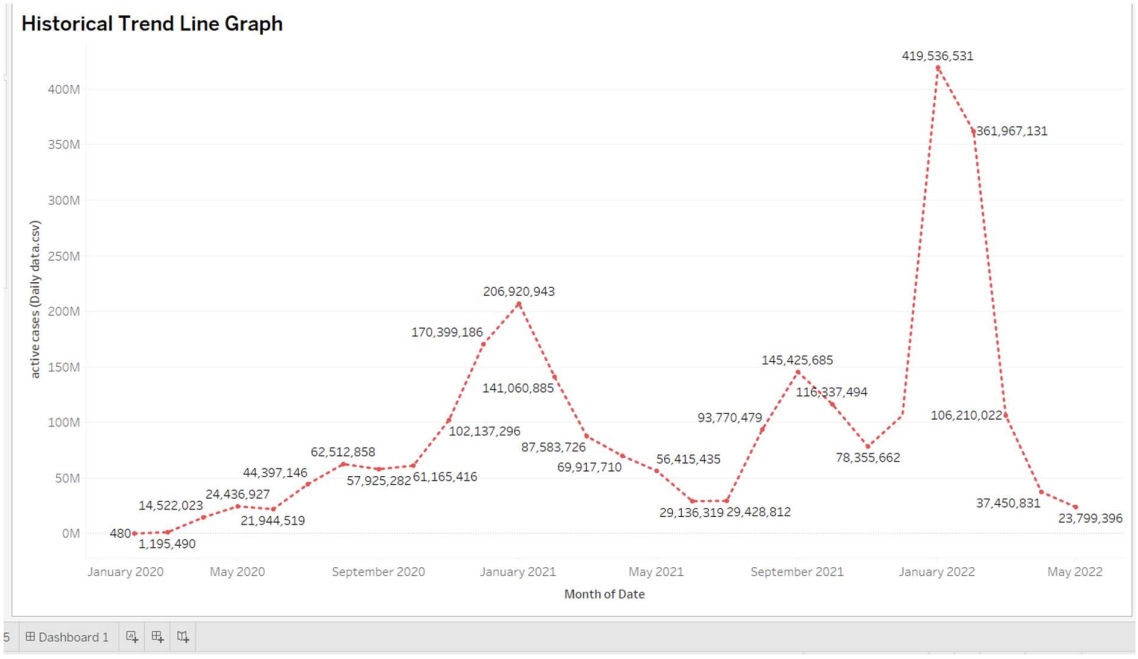
- Analysis of the trend revealed patterns such as periodic spikes that could correlate with events like holidays or the easing of restrictions.
- The graph often showed a lag between the rise in cases and subsequent increase in deaths, illustrating the progression of disease severity.
- Comparing different waves of the pandemic, we could infer the impact of interventions like lockdowns, mask mandates, and vaccinations.

## Design and Creation:

- **Line Graphs**: We employed line graphs for their effectiveness in showing changes over time. Each line represented a different metric, offering a clear visual distinction between new cases and deaths.

- **Color Differentiation**: A separate color was assigned to each metric—typically, a shade of red for deaths to symbolize urgency or danger, and a more neutral color for cases.
- **Annotations**: Key dates or events (such as the start of vaccination rollouts or new variant discoveries) were annotated to provide context to spikes or declines in the trend.
- **Axis Formatting**: The y-axis was formatted to handle large numbers by using abbreviations (e.g., 'M' for millions) to improve readability.

## Historical Trend Line Graph




## Objective:

The Historical Trend Line Graph aimed to illuminate the progression of COVID-19 over an extended period, focusing on cumulative cases, recoveries, and deaths. This visualization

sought to provide a comprehensive view of the pandemic's trajectory, showcasing how cumulative figures have evolved since the outbreak's onset.

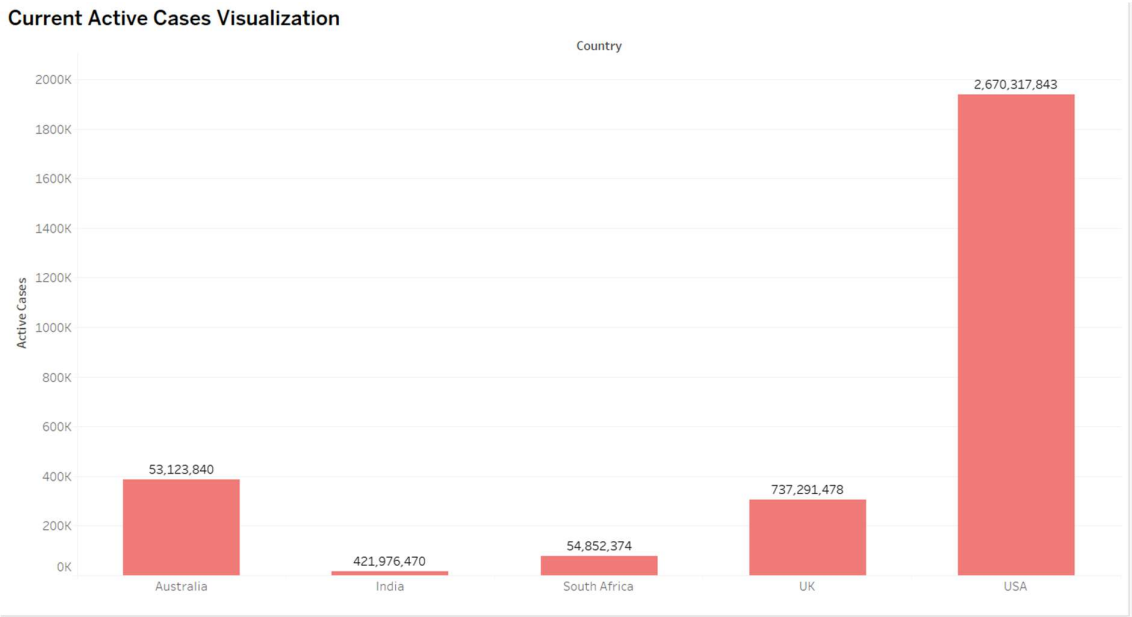
**Analytical Insights:**

- The visualization typically showed a steady incline in cumulative cases, punctuated by steeper increases during outbreak peaks.
- Cumulative deaths followed a similar, albeit less steep, trajectory, underscoring the pandemic's grave human toll.
- The recoveries line, where data permitted, offered a counter-narrative of hope, illustrating the number of individuals overcoming the virus.
- Overlaying these trends highlighted not just the challenges of controlling the virus's spread but also the impact of healthcare interventions and public compliance with health guidelines.

**Design and Creation:**

Multi-Line Graph: A multi-line graph was selected to juxtapose the trajectories of cumulative cases, deaths, and recoveries on the same timeline, facilitating direct comparison.
Color Scheme: Distinct colors were assigned to each metric for clarity—often, vibrant colors for cases, somber tones for deaths, and more hopeful hues for recoveries.
Smoothing: Given the cumulative nature of the data, lines were smoothed to emphasize trends over daily fluctuations, presenting a clearer long-term view.
Annotations: Significant milestones, such as the introduction of public health measures or vaccination milestones, were annotated to contextualize shifts in trends.

## Current Active Cases Visualization:




## Objective:

The objective of the Current Active Cases Visualization was to provide a real-time snapshot of the ongoing COVID-19 situation across different countries. By focusing on active cases, this visualization aimed to highlight regions currently under the most strain from the pandemic, offering insights into where resources and interventions might be urgently needed.

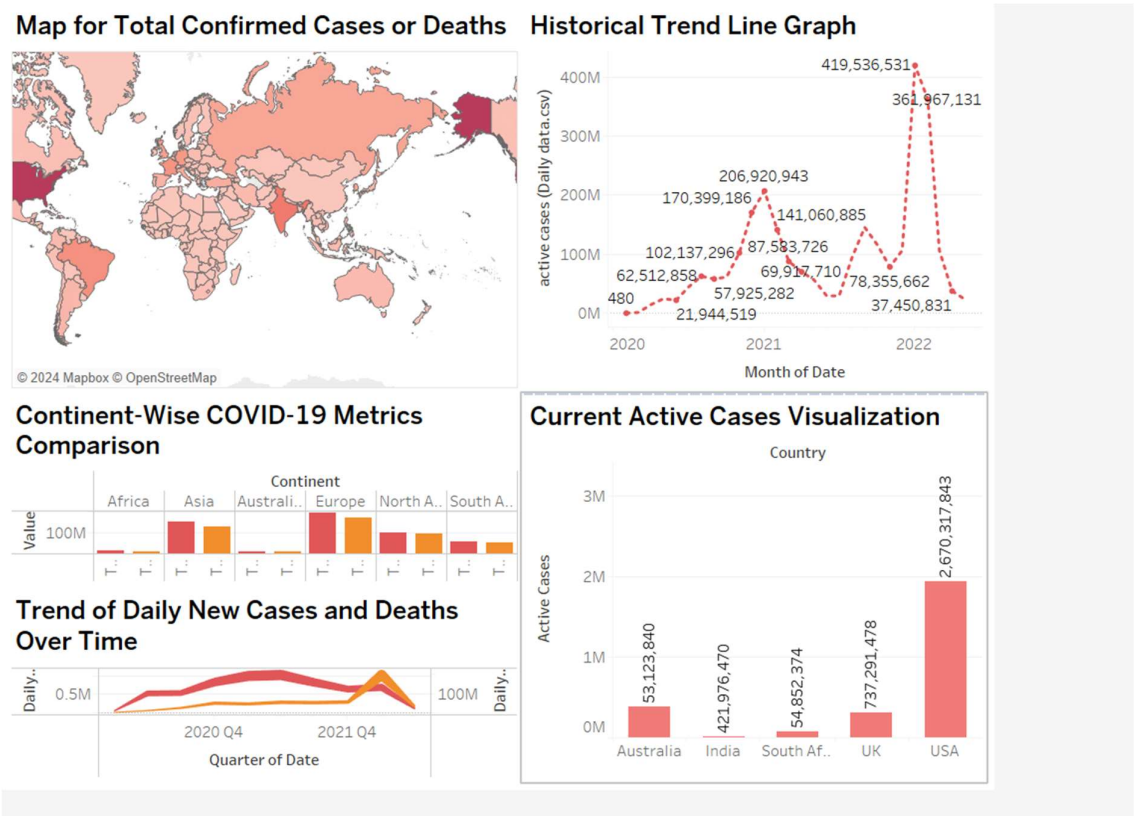
## Analytical Insights:

- This visualization highlighted the dynamic nature of the pandemic, with active cases fluctuating in response to public health measures, vaccine rollouts, and behavioral changes among the population.
- Countries with high active case counts were not always those with the highest cumulative cases, indicating varying stages of pandemic waves and the effectiveness of containment strategies.
- Regions with escalating active cases signaled potential hotspots for emerging variants or areas where health systems might be at risk of being overwhelmed.

## Design and Creation:

- **Bar Chart**: A bar chart was chosen for its straightforward representation of comparative magnitudes, making it easy to identify which countries are experiencing the highest numbers of active cases.
- **Color Coding**: Colors were carefully selected to denote the severity, with warmer colors indicating higher numbers of active cases, thereby drawing immediate attention to areas of concern.
- **Sorting**: The countries were sorted in descending order of active cases to prioritize attention on those with the most urgent situations.
- **Interactivity**: Tooltips were implemented to provide additional details on hover, such as exact numbers of active cases and comparative per capita rates, enhancing the depth of information available.

## Dashboard:



**The Cartographic Tale:**

We begin with a map, a tapestry of nations each shaded according to the burden they bear - the total confirmed cases or deaths painting a sobering picture of the virus's reach. Countries like the United States, India, and Brazil are awash in deeper hues, indicating higher numbers, while other regions hold a lighter stain, but no less affected when considering the size and resilience of their populations.

**Continental Narratives:**

The story unfolds further as we shift our gaze to a series of bars - the Continent-Wise COVID-19 Metrics Comparison. Here, the colors differentiate the statistics of total cases, deaths, and recoveries. We see the disparity, the uneven spread of this global event. Europe and North America, often in darker shades, reveal the significant impact, contrasting sharply with the lighter tones of continents like Africa.

**The Rhythm of Pandemic Waves:**

Next, our tale takes us through the undulating lines of the Trend of Daily New Cases and Deaths Over Time. The lines rise and fall, like the chest of the earth breathing under the strain of the virus - peaks corresponding to waves of infection, troughs to moments of respite, and the hope of a world collectively seeking reprieve.

**The Present State:**

Finally, we arrive at the Current Active Cases Visualization, where stark bars rise like buildings in a city skyline, each representing a country's ongoing struggle. The USA, standing tall,

reminds us of the current fight, the active battles waged in hospitals, and through public health efforts.