**Introduction**



Link: <https://github.com/Vamsi2155/CSCE3250-Project>

**Domain:**

COVID-19 is a global pandemic that has affected millions of lives worldwide, disrupting health systems, economies, and daily life. Understanding its spread and impact is crucial. To understand how it has impacted various countries we choose two datasets for this project. The First Dataset Contains the data of COVID-19 diseases spread across the world and the second dataset contains the population of every country, area, and density of countries. Analysis of COVID-19 data helps public health officials make informed decisions about lockdowns, social distancing measures, and healthcare resource allocation. Through this visualization and analysis, I am trying to communicate complex data understandably.

**Data Abstraction:**

**Dataset -1:**

**Dataset Type:**

Structured Data: The dataset is a structured, tabular data format containing rows and entries corresponding to daily records of COVID-19 statistics for various countries.

**Attributes:**

Geographic Identifiers: iso\_code, continent, location.

Temporal Data: date.

COVID-19 metrics: total\_cases, total\_deaths, total\_cases\_per\_million, total\_deaths\_per\_million, new cases etc

Vaccination data:total\_vaccinations,people\_vaccinated,people\_fully\_vaccinated

Health metrics: life\_expectancy, cardiovasc\_death\_rate, diabetes\_prevalence

Miscellaneous: stringency\_index, gdp\_per\_capita, median\_age, aged\_65\_older, aged\_70\_older etc

**Detailed Description of Dataset**

The COVID-19 dataset provides comprehensive day-to-day updates on the pandemic's impact globally, featuring data on cases, deaths, hospitalizations, and vaccinations. This dataset is invaluable for epidemiological analysis, public health planning, and understanding the efficacy of intervention strategies across different regions and timelines.

Key Statistics:

Entries: 166,326 records

Features: 67 total features

Coverage: Worldwide data, spanning multiple continents and countries.

**Use Cases:**

Epidemiological Research: Analyzing the spread and impact of the virus.

Public Health: Supporting decision-making for interventions and resource allocation.

Policy Analysis: Evaluating the effectiveness of different governmental policies and responses.

**Data Transformation:**

Handling Missing Values: Imputation or removal of missing entries depending on the analysis requirements.

**Dataset -2:**

**Type and Attributes:**

The dataset comprises demographic and geographic data for various countries. Here's a breakdown of the attributes and their respective types:

country (string): Name of the country.

rank (integer): A ranking value (context not specified, possibly related to population size or another demographic measure).

area (float): Total area of the country in square kilometers.

landAreaKm (float): Land area of the country in square kilometers.

cca2 (string): Two-letter country code.

cca3 (string): Three-letter country code.

netChange (float): Net change in population (context and units not specified).

growthRate (float): Annual population growth rate.

worldPercentage (float): The percentage of the world population that resides in the country.

density (float): Population density per square kilometer.

densityMi (float): Population density per square mile.

place (integer): Another rank or position identifier (context not specified).

pop1980 through pop2050 (integer): Population figures for the years 1980, 2000, 2010, 2022, 2023, 2030, and 2050 respectively.

**Detailed Description**

The Dataset has 16 attributes represent population statistics for various countries

The dataset contains a total of 238 records each representing a country's population data.

The dataset includes population data projections and historical figures, as well as geographic data for countries. It provides insights into demographic changes over time, population densities, and how much of the global population is concentrated in these countries. Each record in the dataset is related to one country, detailing not only its geographical attributes but also its demographic dynamics over several decades.

**Detail Design of Features:**

Geographic Information:

country, cca2, cca3 are identifiers.

area, landAreaKm relate directly to the physical characteristics of the country.

Demographic Information:

rank, place serve as indices or positional metrics.

netChange, growthRate, worldPercentage quantify population changes and relevance.

density, densityMi indicate population density.

pop1980, pop2000, pop2010, pop2022, pop2023, pop2030, pop2050 provide temporal population data.

**Task Classification**

The Tasks are to clean the existing data and merge the dataset with the population Data.

Removal of Date-Specific Entries: Date columns were removed, focusing the dataset on country-level aggregations.

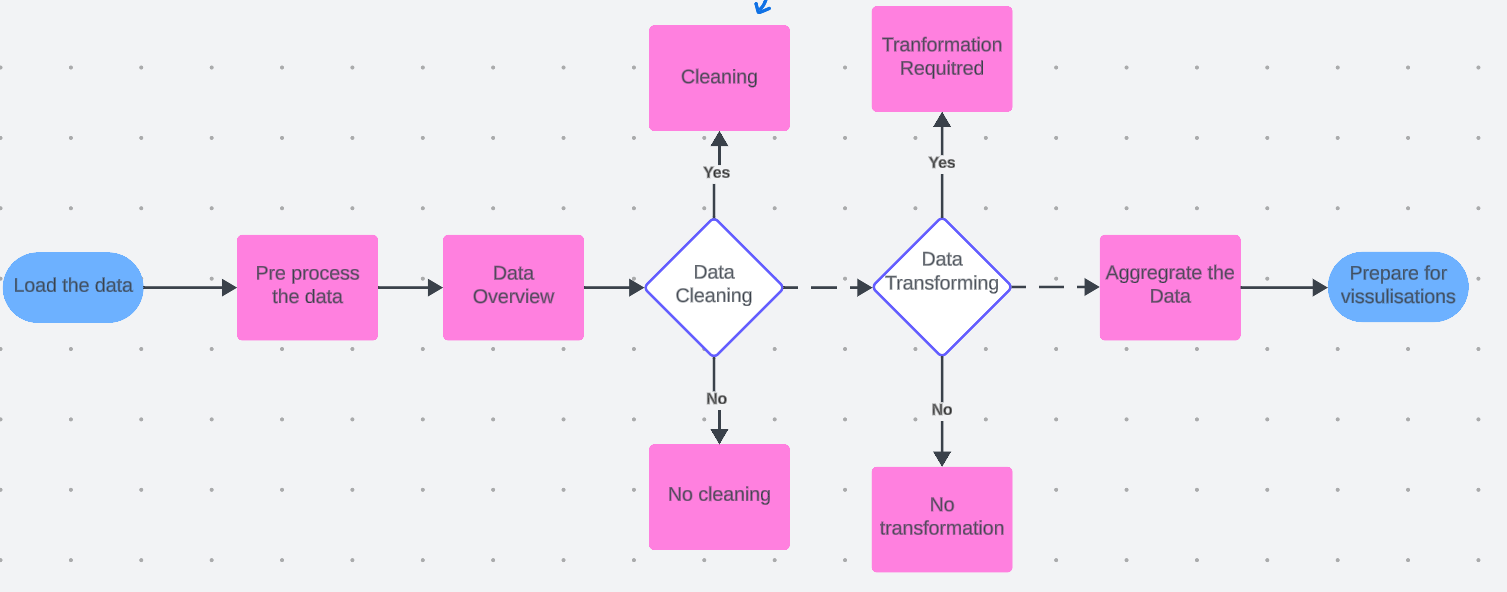
Handling of Missing Data: Rows and columns with significant amounts of missing data, such as 'weekly\_icu\_admissions' or 'weekly\_hosp\_admissions', were either removed or cleaned to ensure the quality of analyses.

Selection of Relevant Columns: Only 16 crucial columns were retained, focusing on aggregate statistics rather than daily updates and more granular temporal data.

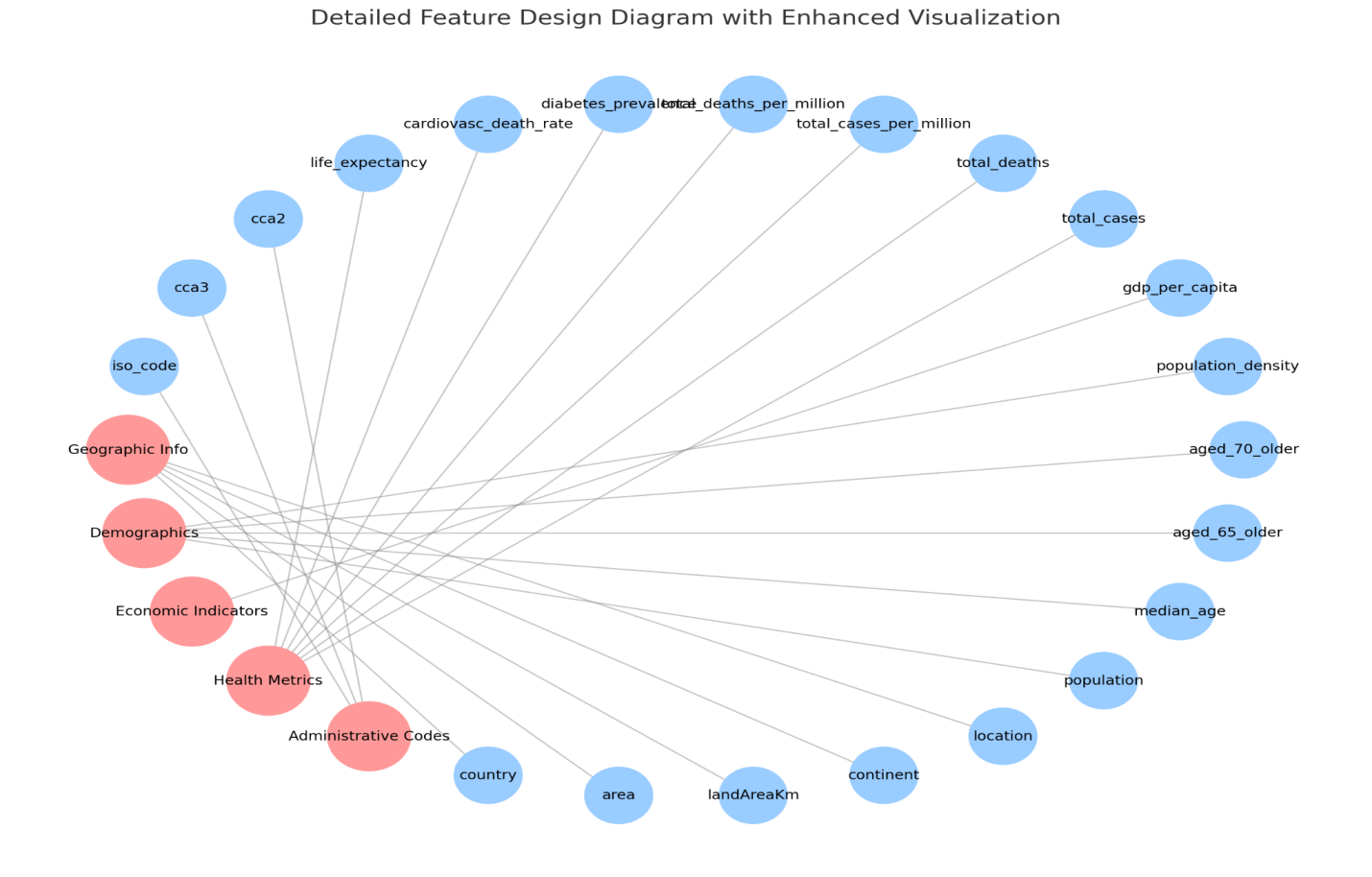
**Results of Actions:**

1. Covid Dataset is consists of Time based data. By Keeping the latest Entry of each country the cleaned dataset has only 238 records
2. Removing the attributes with high null values we de cluttered the data and results in 16 attributes
3. Once the Covid data is cleaned, I merged with the population data.
4. Merging the datasets on the "location" field yielded only 210 entries while with cca3 yielded in 220 entries
5. There are 35 attributes including country demographics (population, area, growth rate), geographic identifiers (ISO codes, continent), and health metrics (COVID-19 cases and deaths, population density, median age, GDP per capita, etc.).

Workflow Diagram:



Once the data is merged, Here is the Detail design of Features with diagram.



The diagram visually represents the structure and relationships within a dataset that combines demographic, geographic, economic, and health-related data for various countries. The design is structured to show how various data attributes (or features) are grouped into distinct categories based on their nature and use. This helps in understanding the dataset's complexity and can guide further data analysis or model building.

Description of Diagram Components

Categories:

Each category is represented by a node in a distinct color (light red in this case), emphasizing its role as a container for related features. The categories are:

Geographic Info: Includes all geographical descriptors of a country.

Demographics: Focuses on population metrics and related statistics.

Economic Indicators: Contains economic data such as GDP per capita.

Health Metrics: Encompasses health-related statistics, such as case numbers and death rates from diseases.

Administrative Codes: Deals with administrative identifiers like country codes.

Features:

Features are individual data attributes nested within categories. They are shown in a different color (light blue) to differentiate them from category nodes.

Each feature node is connected to its parent category node with a dotted line, highlighting the relationship and dependency on the category for context.

Now let’s deep dive into the visualizations:  
  
I made 8 Graphs utilizing Tableau, Power BI, and Python libraries Seaborn and Matplotlib.

**Tableau:**

Tableau is a powerful and intuitive data visualization tool widely used in business intelligence to help individuals and organizations understand and analyze their data visually. Its user-friendly drag-and-drop interface allows users to create interactive and shareable dashboards.

Typical Uses:

Interactive Dashboards: Tableau excels in building interactive visuals that allow users to manipulate variables to see real-time changes.

Geospatial Mapping: It offers robust mapping capabilities to analyze data across geographical locations.

Data Blending: It can blend data from multiple sources to create comprehensive visuals.

**Power BI:**

Power BI is a business analytics service provided by Microsoft. It provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards.

Typical Uses:

Business Dashboards: Power BI is often used for creating detailed business dashboards that include KPIs, metrics, and data points crucial for business decisions.

Real-Time Data Processing: It supports real-time processing, which is essential for dynamic business environments.

Integration with Azure: Being part of the Microsoft ecosystem, it integrates seamlessly with Azure services for extended analytics

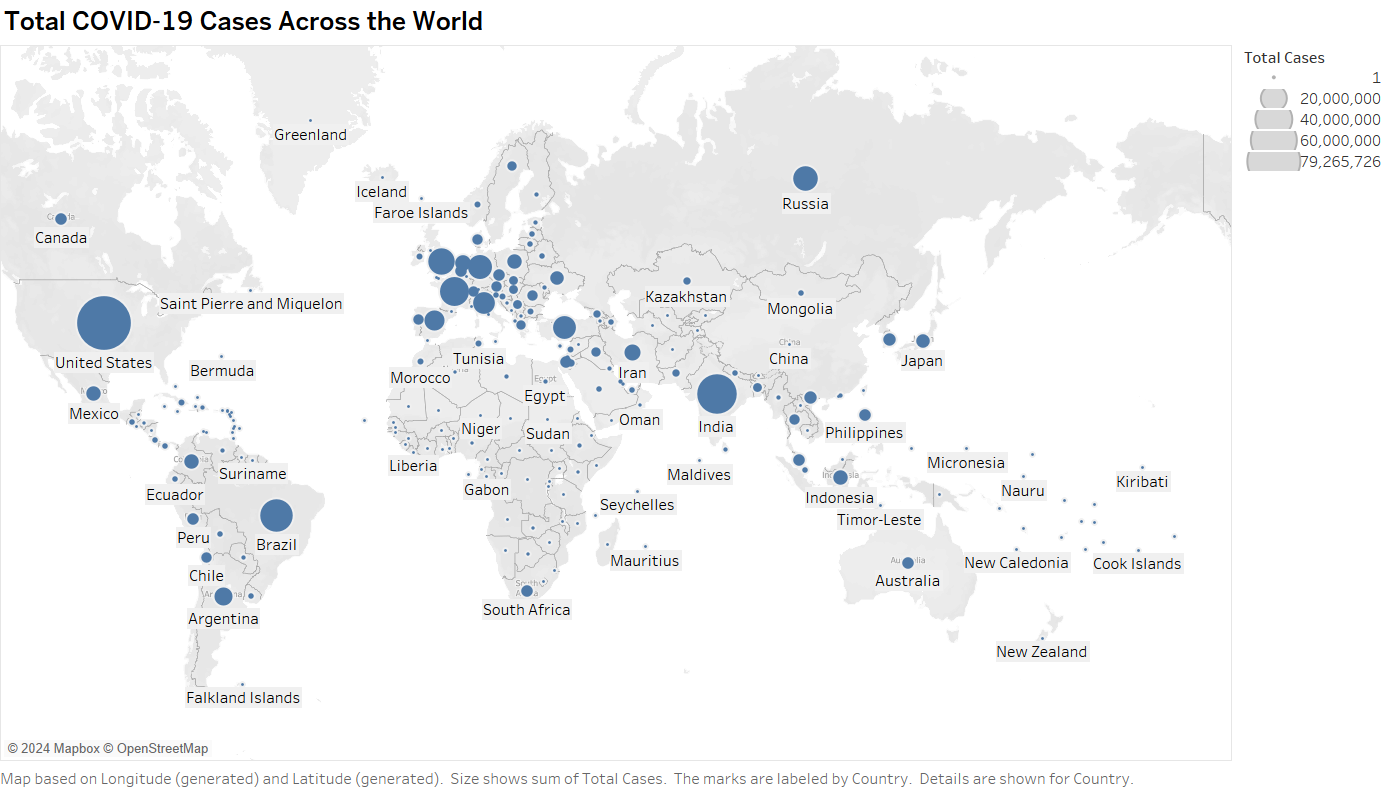
**Python:**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It is highly customizable and works well with many backends and operating systems. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Graph-1: Total COVID-19 Cases Across the World**

**Tool Used:**  Tableau

This geographic visualization maps the total number of COVID-19 cases across the world, with countries represented as circles sized proportionally to the total cases they have reported.



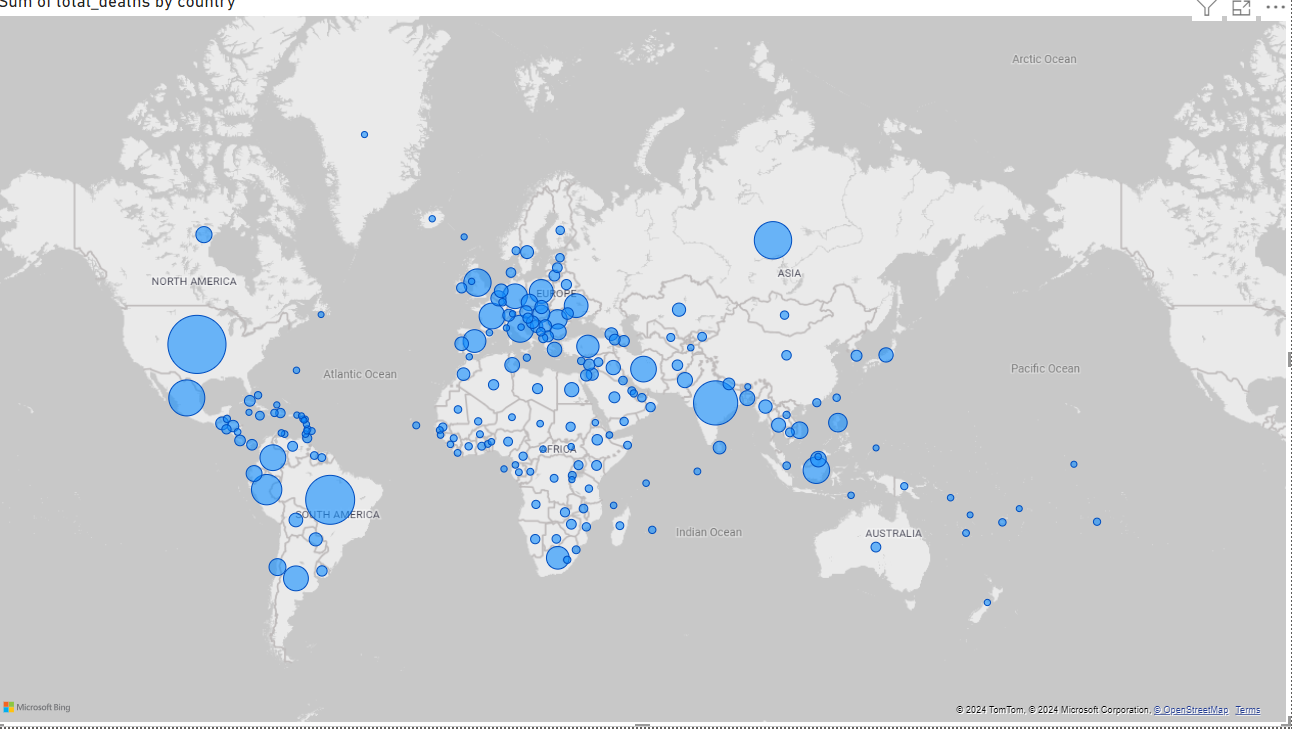
Interpretation:

* The largest circles, indicating the highest numbers of cases, are prominently visible in countries like the United States, India, and Brazil, reflecting significant outbreaks.
* Similarly in Europe, as well where countries like France Uk and Germany can be noticed Smaller circles are seen throughout Europe, Asia, and Africa, indicating a varied impact of the pandemic across different regions.
* Africa has the least impact of COVID-19 followed by Oceania

**Graph 2: Total COVID-19 deaths Across the World**

**Tool Used:** Power BI

This geographic visualization maps the total number of COVID-19 deaths across the world, with countries represented as circles sized proportionally to the total cases they have reported.



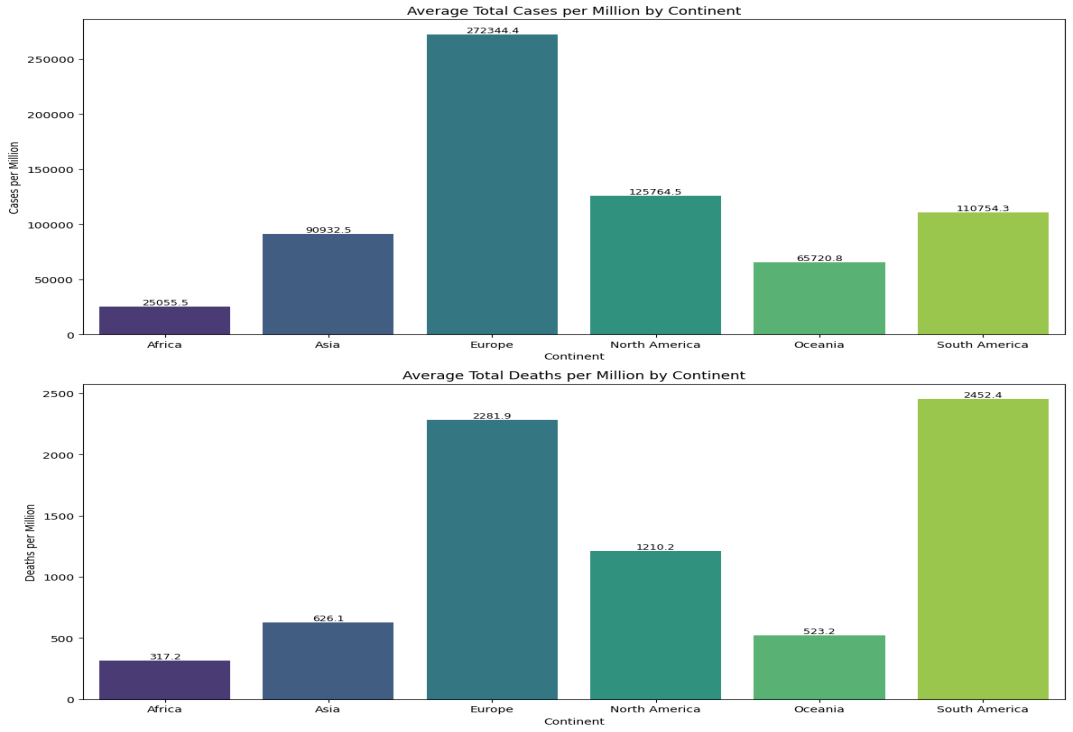
Interpretation:

Similar to the first graph, larger circles denote higher numbers of deaths, with significant visibility in the United States, Brazil, and India, countries that have also reported high case counts.

**Graph-3: Comparison of Average Total COVID-19 Cases and Deaths per Million by Continent**

Tool used: Python with Seaborn library

This pair of bar charts was created using the Seaborn library in Python, a popular data visualization library that provides a high-level interface for drawing attractive and informative statistical graphics.



Interpretation:

* Europe and North America exhibit significantly higher average cases per million compared to other continents, with Europe showing the highest.
* Africa and Asia have comparatively lower figures
* The numeric values atop each bar provide a precise measurement, emphasizing Europe's severe outbreak relative to its population size.
* In Deaths, the distribution pattern slightly differs from the cases, with Europe and South America showing the highest deaths per million. This indicates not only high transmission rates but also varying mortality impacts, which could be influenced by healthcare infrastructure and demographic factors.
* The visualization distinctly highlights the heavy impact in South America, which has a high death rate despite fewer cases per million than Europe

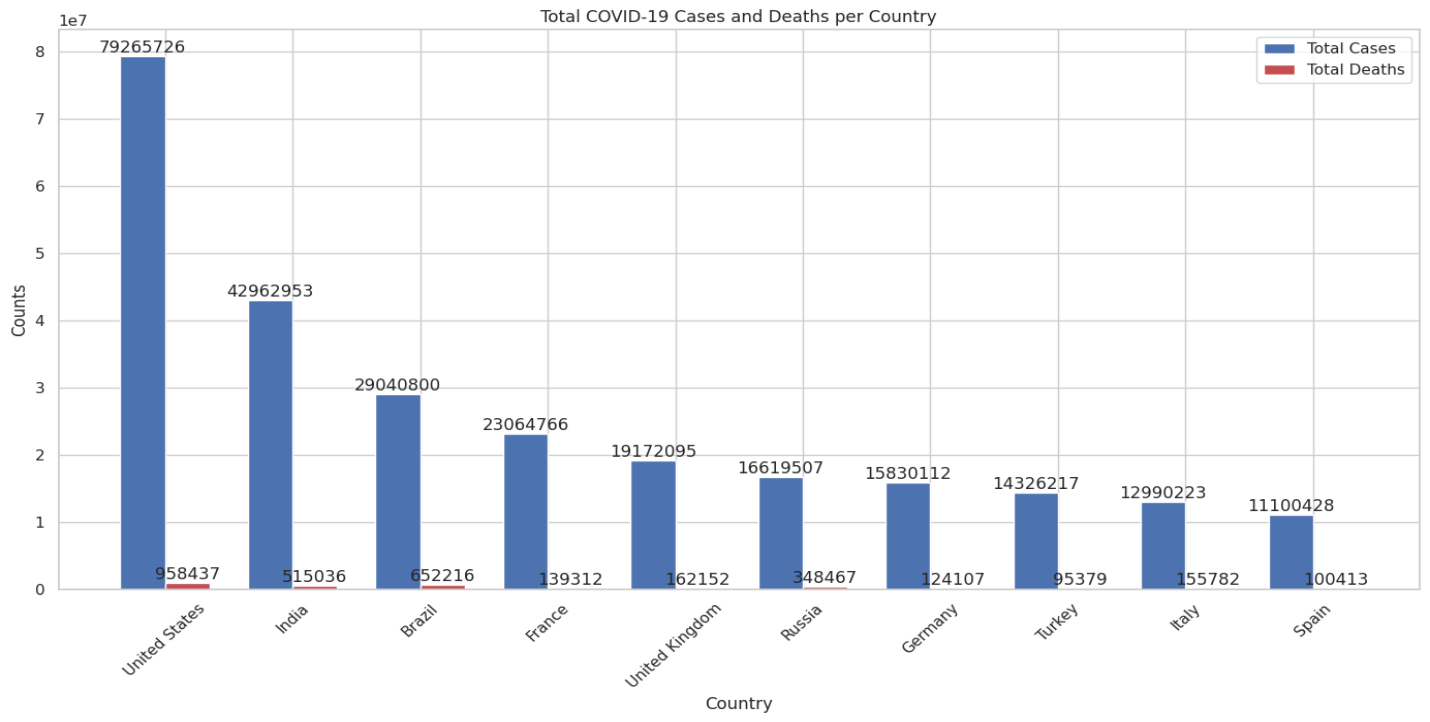
**Graph-4: Comparison of Average Total COVID-19 Cases and Deaths by Country - TOP 10**

Tool used: Python with Seaborn library

This bar chart visualizes the total number of COVID-19 cases and deaths for selected countries, using Seaborn for a clean and straightforward presentation

The chart differentiates between total cases (blue bars) and total deaths (red bars), providing a direct visual comparison of the impact in terms of both infection and fatality rates across the countries.

Each country is listed on the x-axis, with the total cases and deaths clearly labeled on the respective bars



Interpretation:

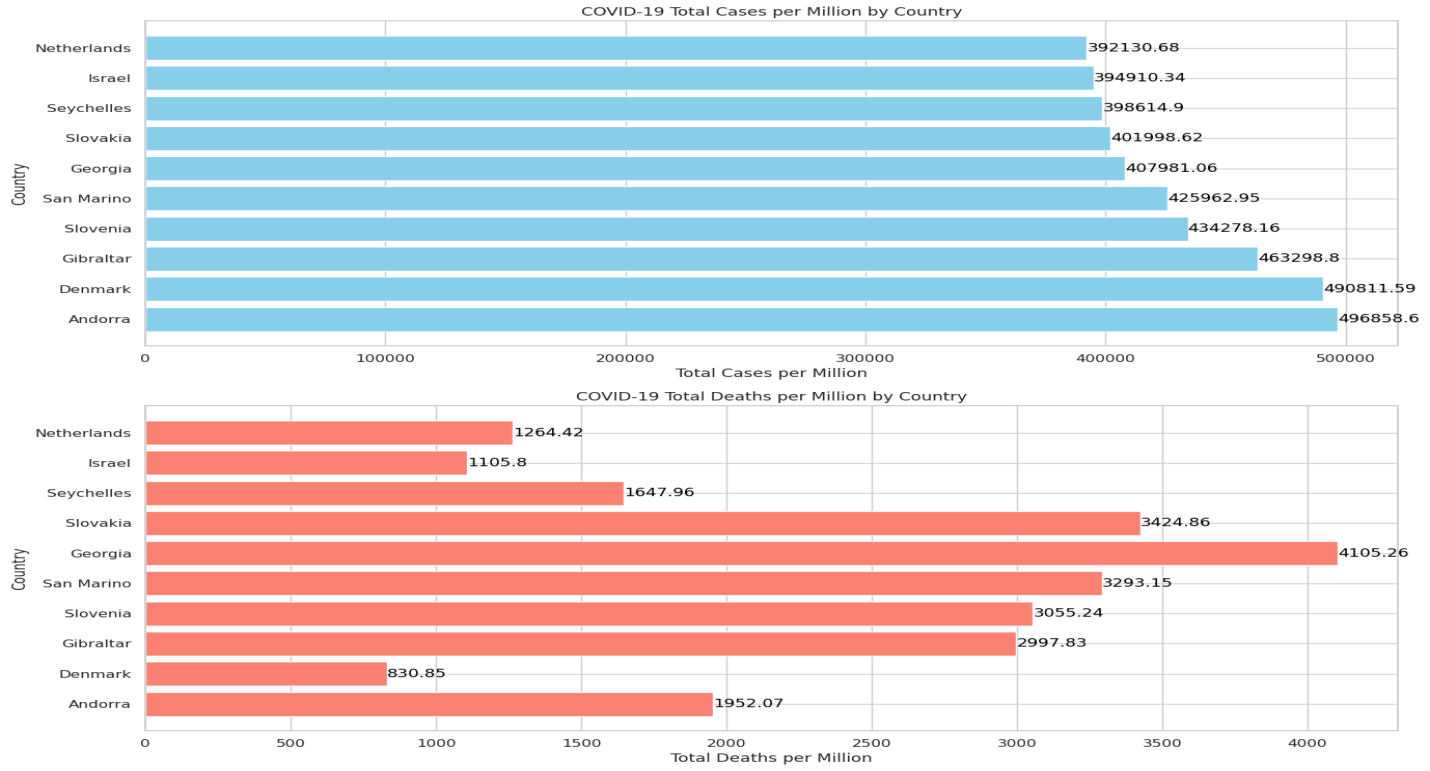
The United States, India, and Brazil show the highest numbers in both categories, reflecting their significant struggles with the pandemic. While the fatality due to the covid is very insignificant when compared to total cases

**Graph-5: Comparison of Average Total COVID-19 Cases and Deaths per million by Country - TOP 10**

Tool used: Python with Seaborn library

This set of horizontal bar charts provides a detailed breakdown of COVID-19 total cases and deaths per million by country, emphasizing the rate of infection and fatalities adjusted for population size.

Countries are sorted by cases per million from highest at the bottom to lowest at the top, facilitating a quick understanding of which countries have the highest relative infection rates and fatality rates



Interpretation:

* Some Countries reaching up to nearly 500,000 cases per million in some countries, highlighting severe outbreaks
* This chart shows that countries like Andorra, Gibraltar, and San Marino have exceedingly high numbers of cases per million. These figures suggest extremely high transmission rates within these small populations
* In Deaths, notably, San Marino, Belgium, and Slovenia appear with high death rates per million, suggesting that while the transmission was high, the death rate also reflects significant challenges in managing the disease's severity among populations.
* The disparities between the cases and deaths can be indicative of different healthcare system efficiencies

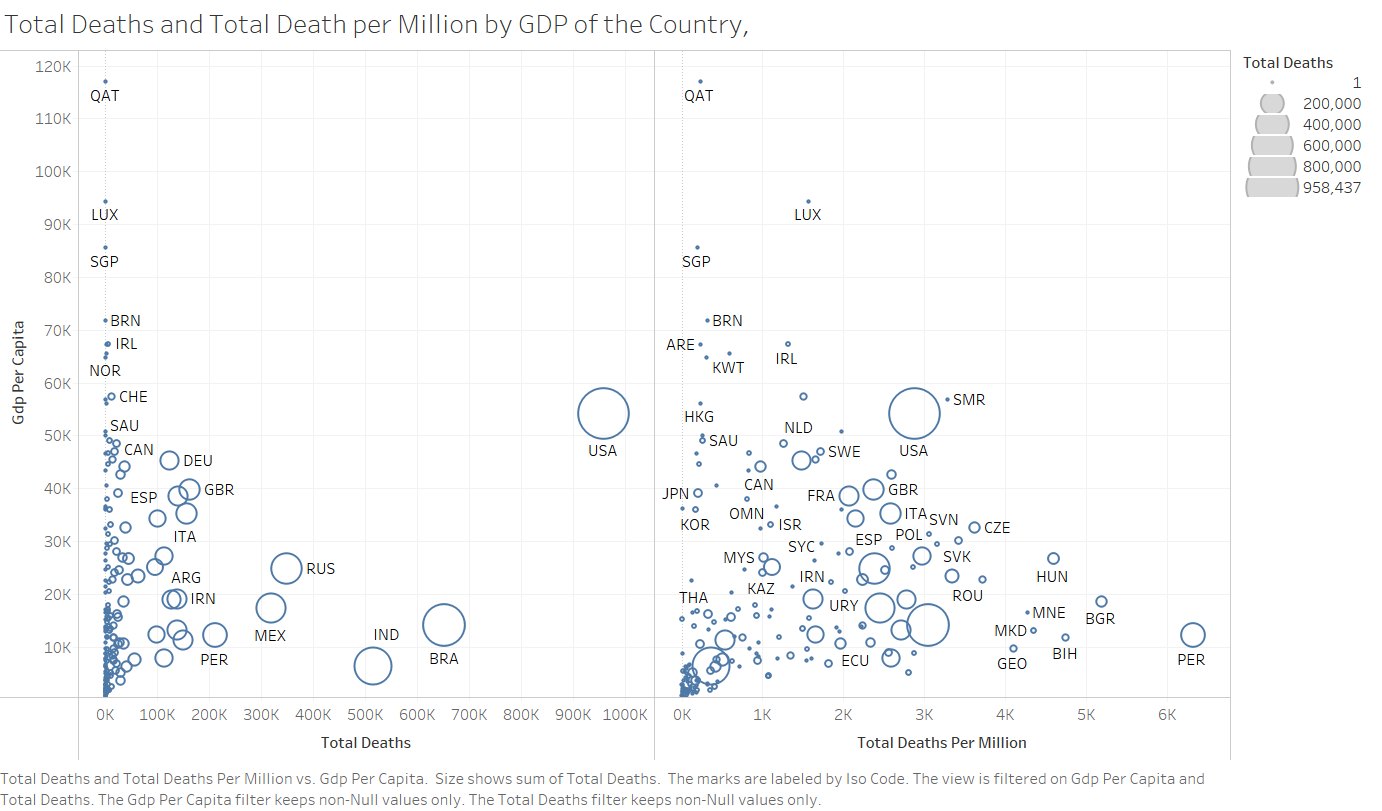
**Graph-6: Total Deaths and Total Deaths per Million by GDP of the Country**

Tool used: Tableau

This scatter plot displays the relationship between a country's GDP per capita and its COVID-19 related total deaths and deaths per million.

Each point represents a country, labeled by its ISO code, with the x-axis indicating total deaths and the y-axis showing GDP per capita.

The size of each point reflects the number of total deaths per million, providing a multi-dimensional analysis of the economic impact versus the health impact.



Interpretation:

* Wealthier nations (higher GDP per capita), such as Luxembourg and Singapore, tend to have higher total deaths but relatively lower deaths per million, suggesting better overall healthcare systems capable of managing the severity per capita.
* The USA stands out with a high number of total deaths and a significant size indicating deaths per million, despite a high GDP per capita, highlighting challenges in pandemic management.
* Countries with lower GDP per capita like India and Brazil show a large number of total deaths with a high rate per million, indicating severe impacts compounded by economic challenges.

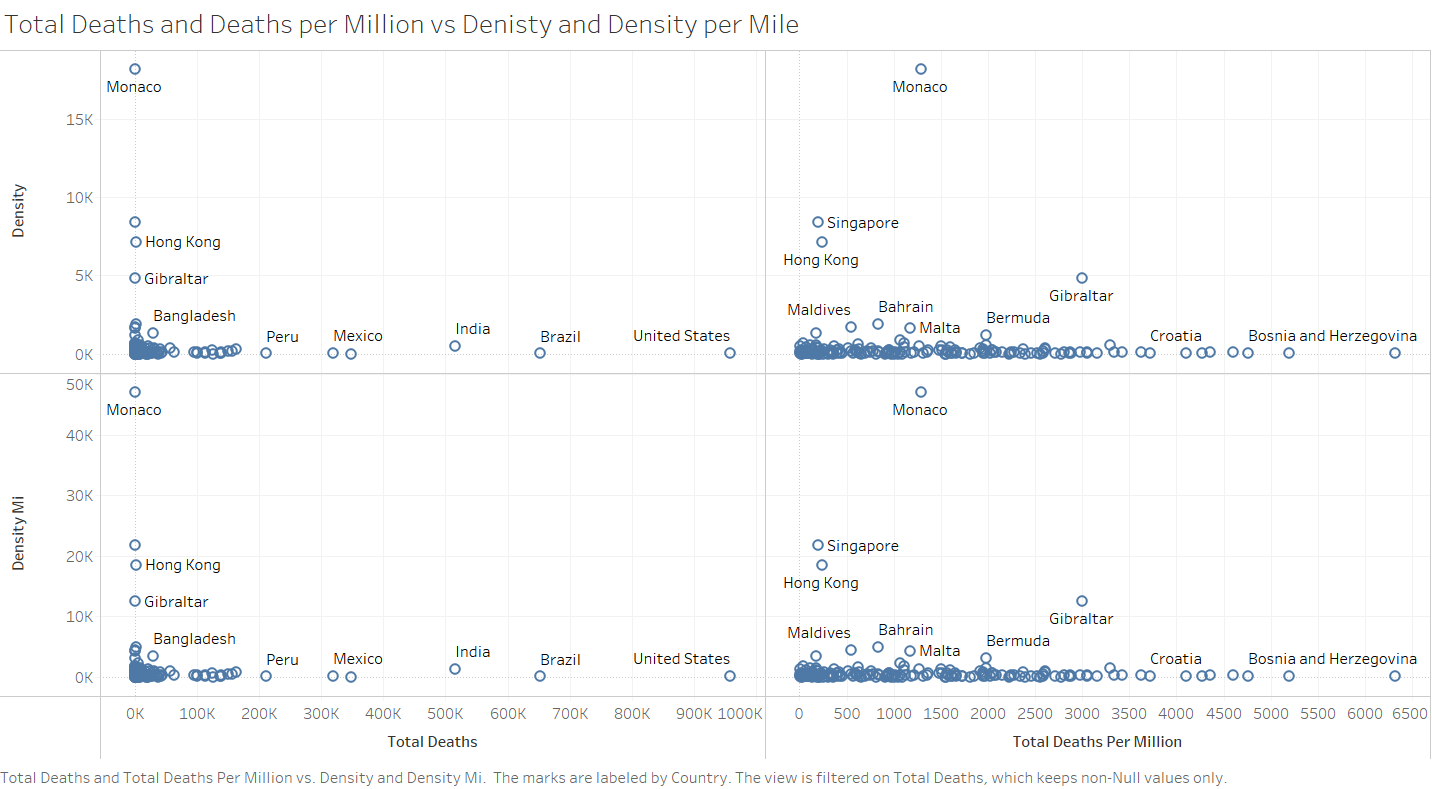
**Graph-7: Total Deaths and Deaths per Million vs. Density and Density per Mile**

Tool used: Tableau

This complex visualization compares total deaths and deaths per million against population density and density per mile.

The y-axis indicates the population density, while the x-axis shows total deaths.

The size of each point represents deaths per million, and the points are labeled by country.



Interpretation:

* Countries with high density such as Bangladesh and India show a large number of total deaths but relatively smaller deaths per million, suggesting that while the absolute numbers are high, the death rate relative to the density is managed.
* Conversely, smaller countries or regions like Monaco and Singapore, with high densities, have managed to keep both absolute and per million deaths relatively lower, likely due to effective containment and healthcare strategies.
* This graph helps visualize the challenge of managing COVID-19 in densely populated areas.

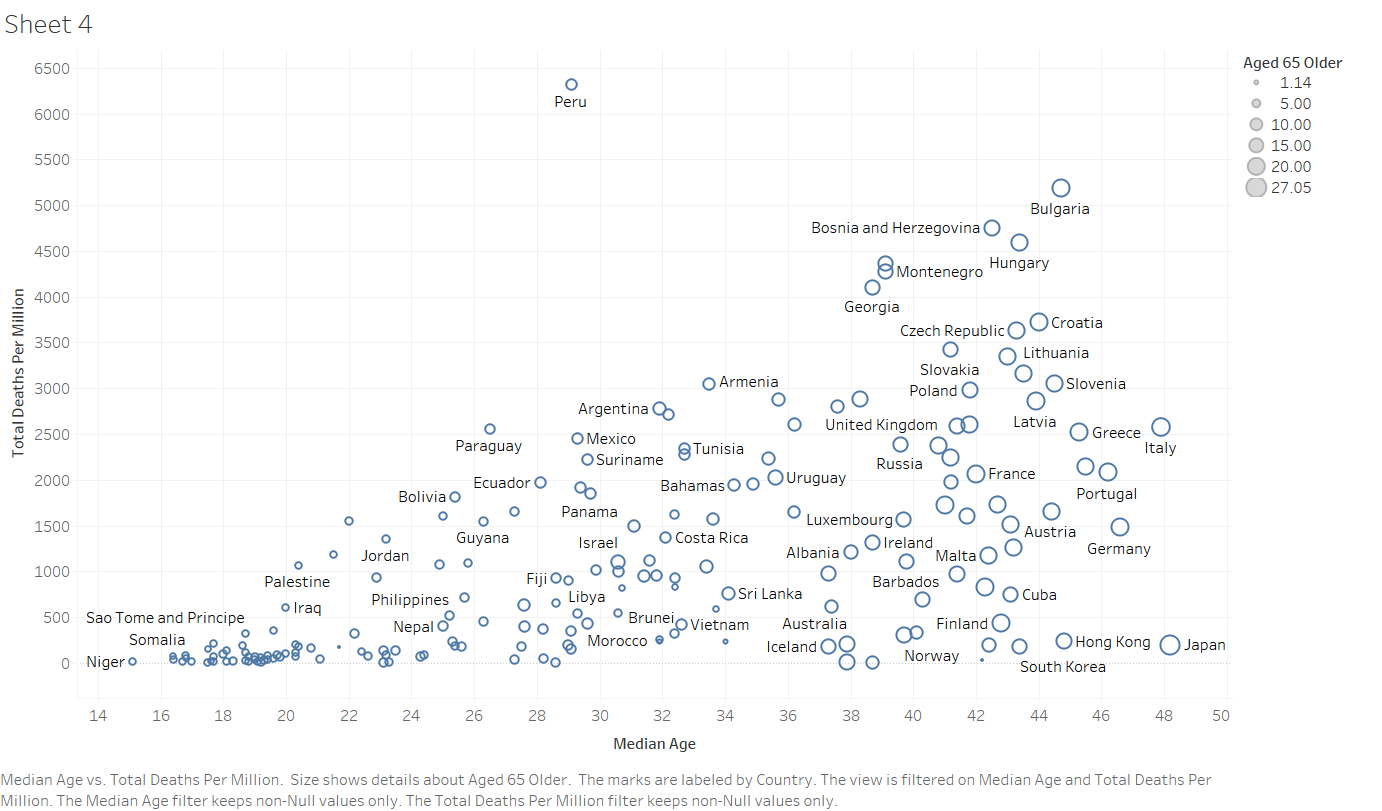
**Graph-8: Median Age vs. Total Deaths per Million**

Tool used: Tableau

This scatter plot explores the relationship between the median age of a country’s population and its COVID-19 death rates per million.

The x-axis shows the median age, and the y-axis shows total deaths per million.

The size of each point indicates the percentage of the population aged 65 or older, providing insight into how age demographics affect COVID-19 mortality rates.



Interpretation:

* Countries with older populations (higher median age and larger percentage of aged 65 or older), such as Italy, Portugal, and Germany, have higher deaths per million, underscoring the vulnerability of older individuals to COVID-19.
* Countries with younger populations like Nepal and Iraq have significantly lower death rates per million, highlighting the protective effect of a younger demographic against severe outcomes.
* The scatter plot effectively shows the correlation between age demographics and COVID-19 mortality, important for healthcare planning and resource allocation.

**Conclusion:**

1. Economic Impact: There is a notable correlation between a country's economic status (GDP per capita) and its ability to manage the health outcomes of its population, with wealthier countries generally showing better management per capita.
2. Demographic Challenges: Age demographics significantly influence COVID-19 mortality rates, with older populations experiencing higher death rates. This underscores the importance of targeted protective measures for older communities.
3. Density and Spread: High-density countries face unique challenges in controlling the spread of COVID-19, but effective management strategies can mitigate these challenges, as seen in some high-density but low mortality regions.

**References:**

1. <https://www.kaggle.com/datasets/rajkumarpandey02/2023-world-population-by-country>

2. <https://www.kaggle.com/datasets/georgesaavedra/covid19-dataset>