

**CAPSTONE PROJECT**

**Project Title: - COVID 19 DATA ANALYSIS**

**Course: PYTHON PROGRAMMING**

**Course Code: CAD 102**

|  |  |
| --- | --- |
| **Submitted by**  **12404584 - CHRISTIAN YARED DONDO**  **12415237 - SACHIN KUMAR SINGH** | **Submitted to**  **MR. MADHAV DHUBEY** |

**Lovely Faculty of Technology & Sciences**

**School of Computer Applications**

**Lovely Professional University**

**COVID 19 DATA ANALYSIS REPORT**

**Introduction**

The COVID-19 pandemic has had a profound impact on global health, economies, and daily life. Analyzing the available data can provide valuable insights into its spread, mortality rates, vaccination efforts, and the effectiveness of government measures. This project utilizes a dataset from Our World in Data, a trusted source for global statistics, to explore various aspects of the pandemic.

**Objective**

The project aims to analyze COVID-19 data to uncover trends, relationships, and patterns across 6 countries, India, Australia, United States, United Kingdom, Russia and Tanzania. The dataset includes critical variables such as total cases, total deaths, vaccination statistics, government response indices, and demographic information.

**Dataset Overview**

The dataset was sourced from **Our World in Data**, which provides globally standardized COVID-19 statistics.

**Source:** Our World in Data (https://ourworldindata.org/coronavirus)

The dataset came with 67 Columns and 429435 Rows then Preprocessed it to contain 14 Columns of interest to this analysis. Also, the dataset included all the countries in the world but was then filtered to contain only 6 countries(ie. India, Australia, United States, United Kingdom, Russia and Tanzania) of interest to this Data analysis.

**Columns:**

* **iso\_code:** Country code.
* **location:** Country name.
* **date:** Date of observation.
* **total\_cases:** Cumulative number of cases.
* **total\_deaths:** Cumulative number of deaths.
* **people\_fully\_vaccinated:** Number of people fully vaccinated.
* **new\_vaccinations\_smoothed:** 7-day rolling average of daily vaccinations.
* **stringency\_index:** Government response stringency.
* **population\_density:** People per square kilometer.
* **median\_age:** Median age of the population.
* **aged\_65\_older:** Percentage of population aged 65 or older.
* **aged\_70\_older:** Percentage of population aged 70 or older.
* **life\_expectancy:** Average life expectancy in years.
* **population:** Total population.

**Steps in Analysis**

1. **Data Cleaning and Preprocessing**
2. **Handling Missing Data**

Columns with Missing Values were addressed

* stringency\_index: Replaced missing values with the column's median value to preserve the dataset's integrity.
* people\_fully\_vaccinated: Applied forward fill (ffill) to carry forward the last available vaccination count.

1. **Removing Irrelevant Columns**

Certain columns were deemed unnecessary for the current analysis and were removed:

iso\_code: Not relevant for visualization or analysis.

new\_vaccinations\_smoothed: Omitted as the focus is on cumulative vaccination data.

aged\_65\_older & aged\_70\_older: Removed due to redundancy with median\_age.

1. **Filtering Rows**

To ensure meaningful analysis:

* Exclusion of rows with incomplete data in critical fields:
* Rows with missing values in total\_cases and total\_deaths were dropped.
* Focus on specific countries: Filtered specific locations (India, USA, UK, Australia, Tanzania, Russia) to focus on individual countries.

1. **Formatting Date**

The date column was converted from string format to a datetime object for best time-based analysis.

**2. Exploratory Data Analysis (EDA)**

**A. Descriptive Statistics**

Identified distributions of total cases, deaths, and vaccinations.

Key metrics for each country include:

* Mean, median, and variance of total cases and deaths.

**B. Comparative Analysis**

Total Cases vs. Total Deaths:

* Countries with high death-to-case ratios.
* Vaccination Progress:

Comparison of vaccination coverage among countries.

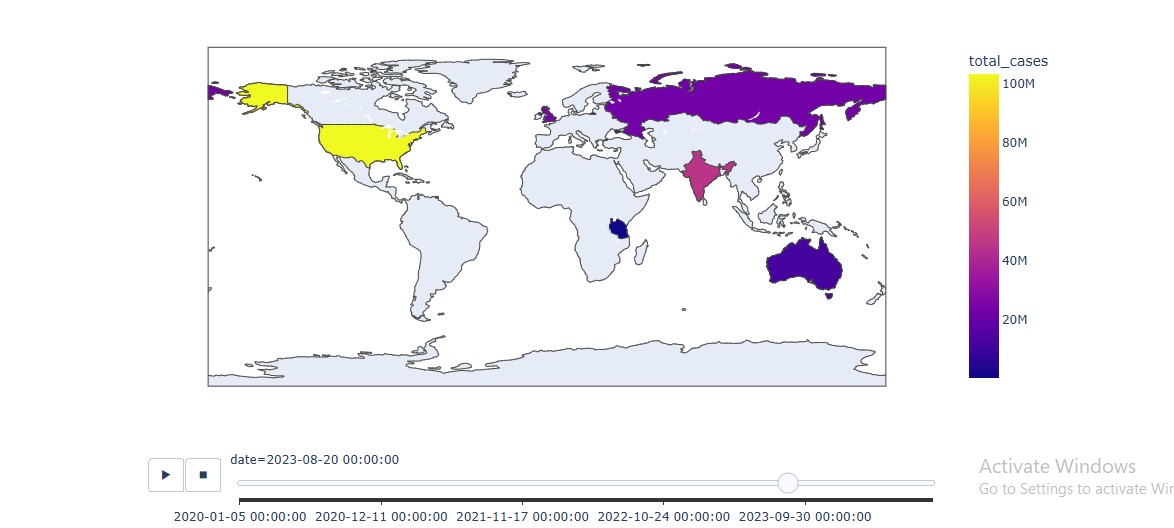
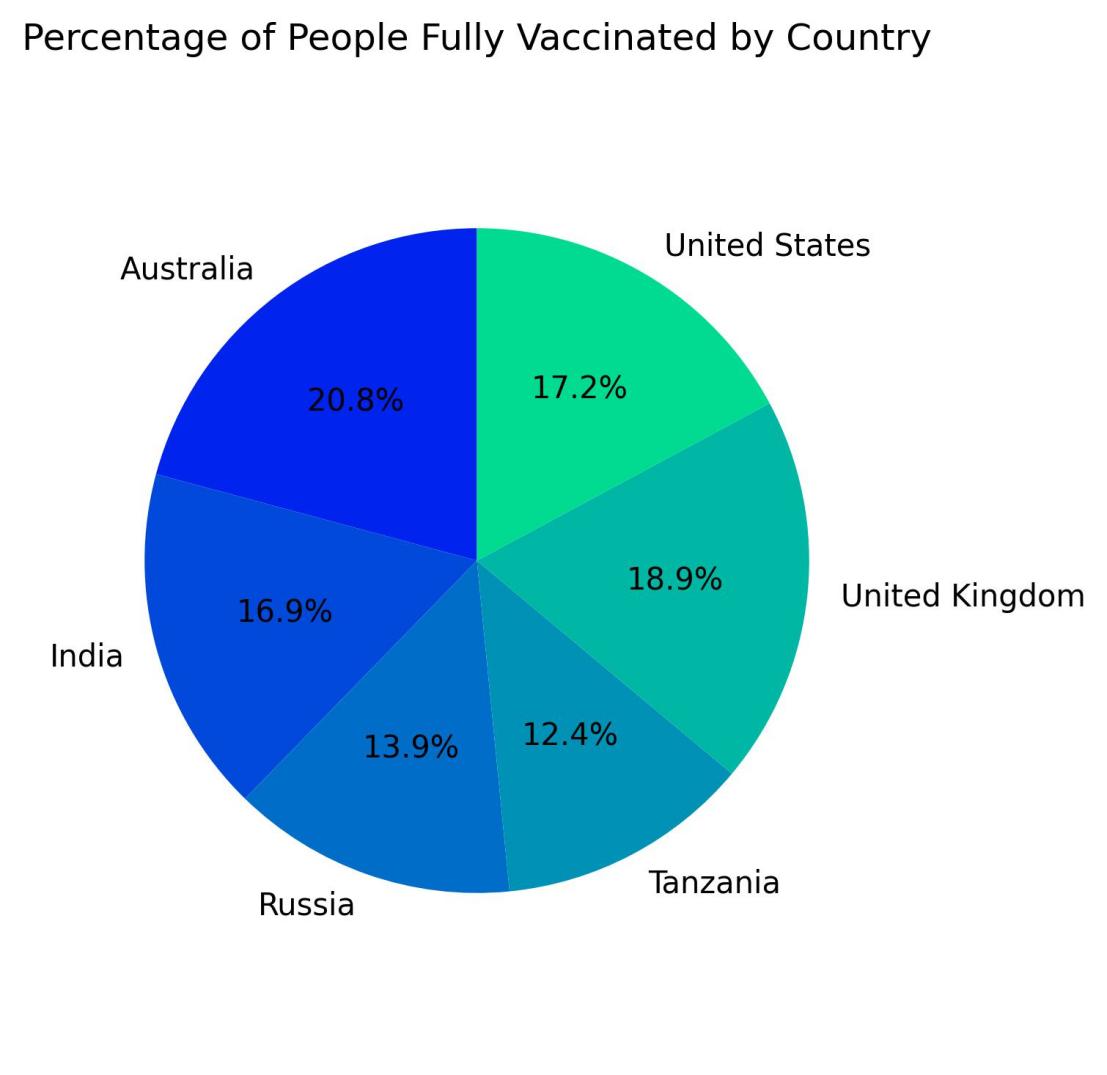
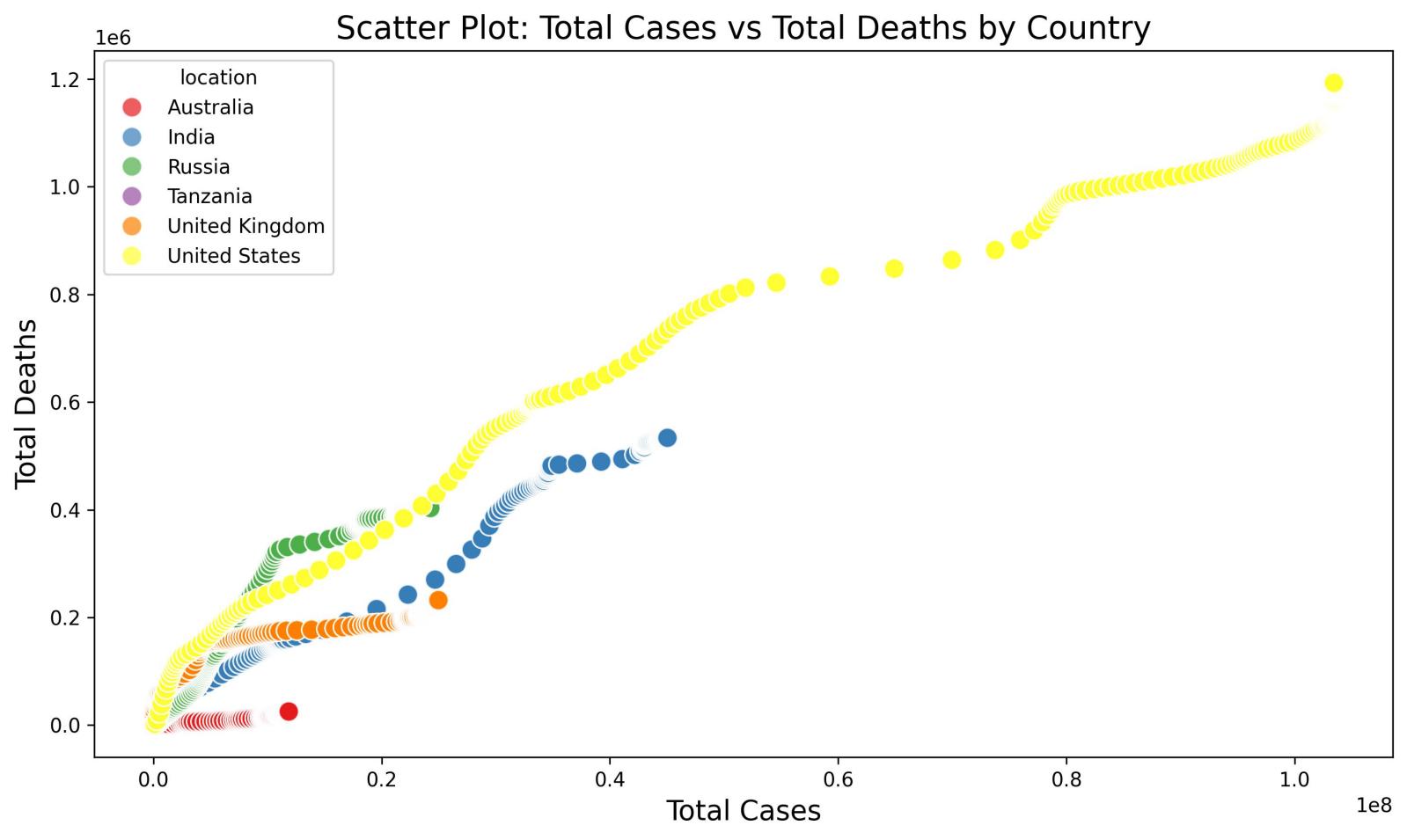
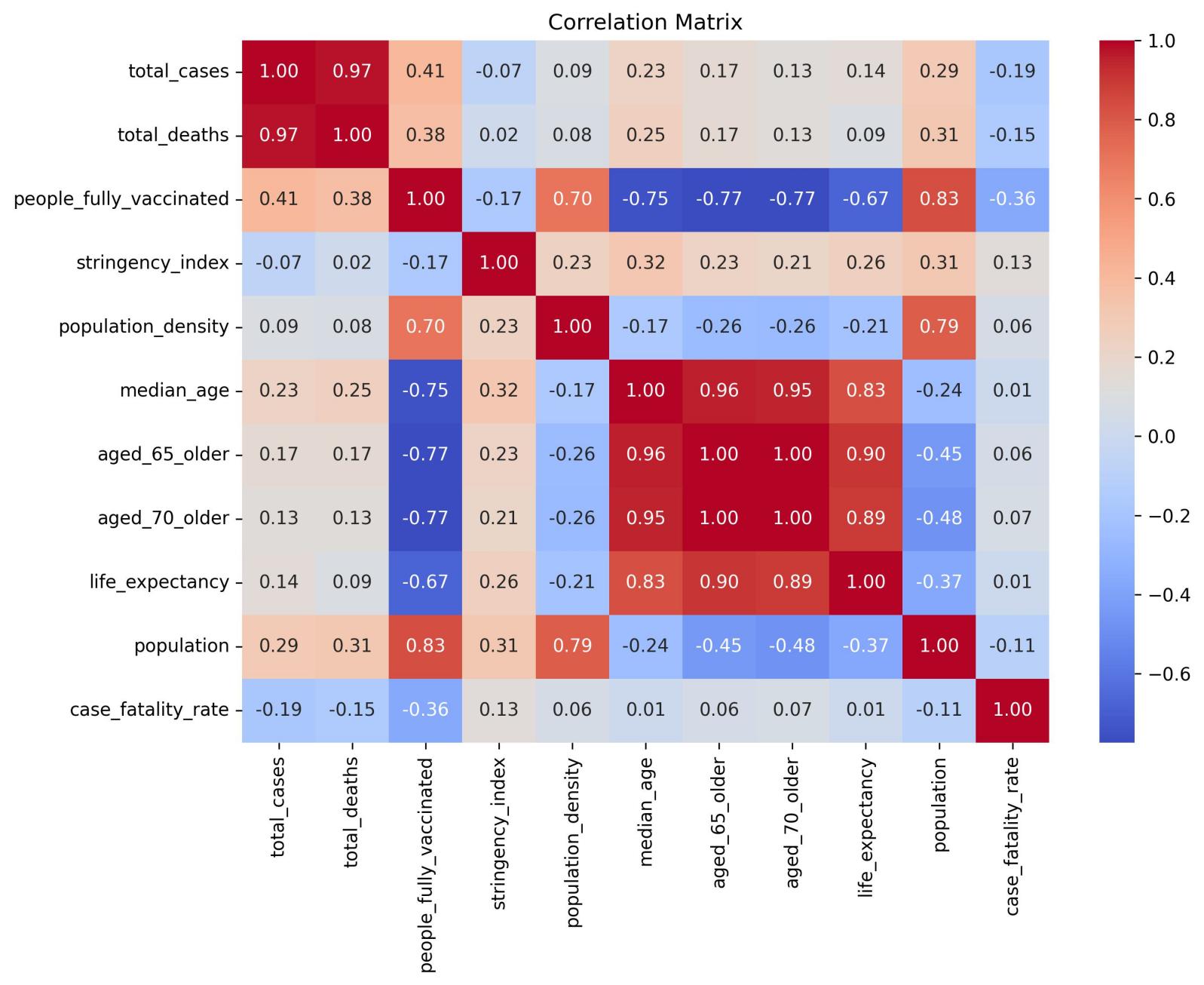
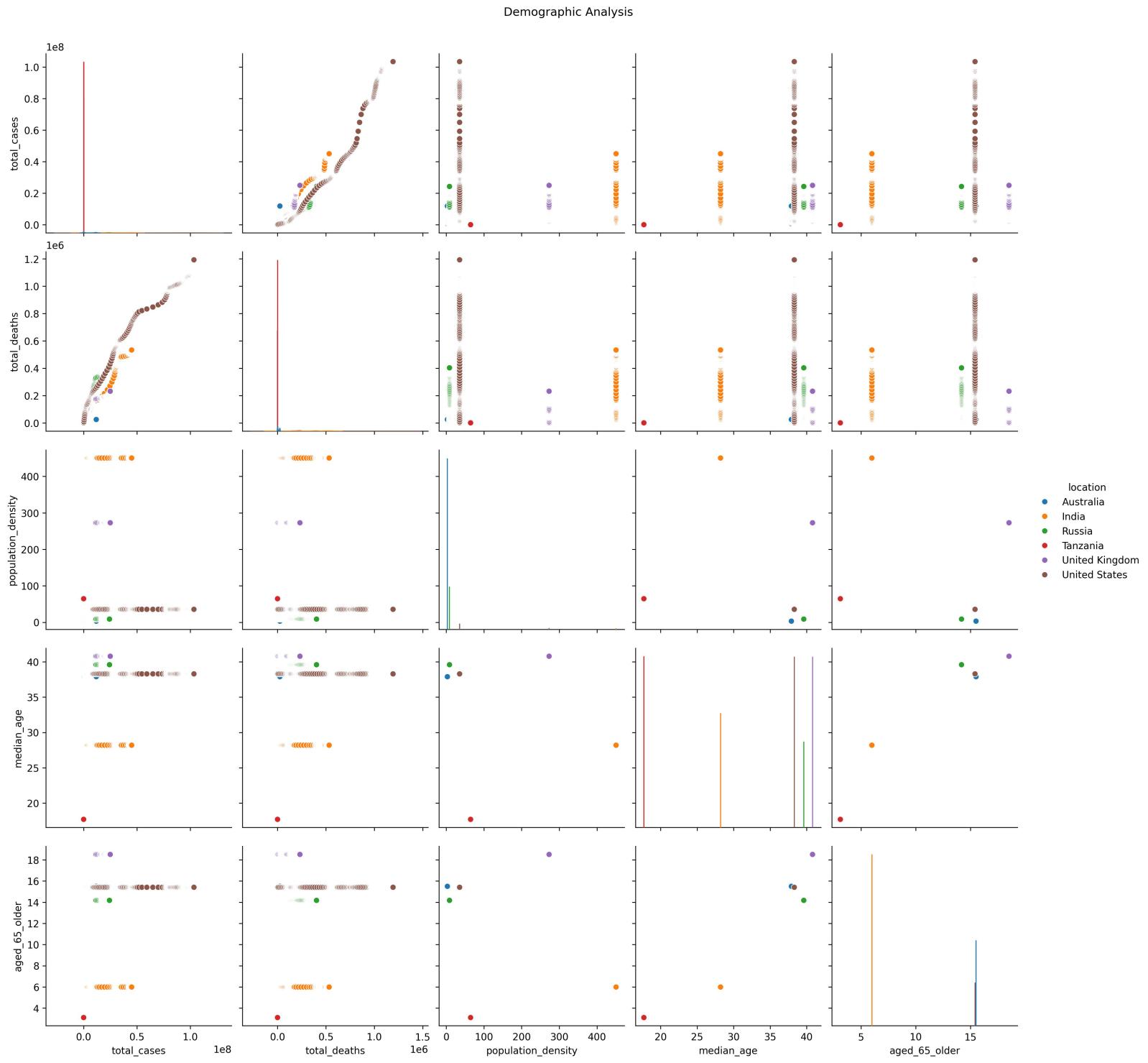
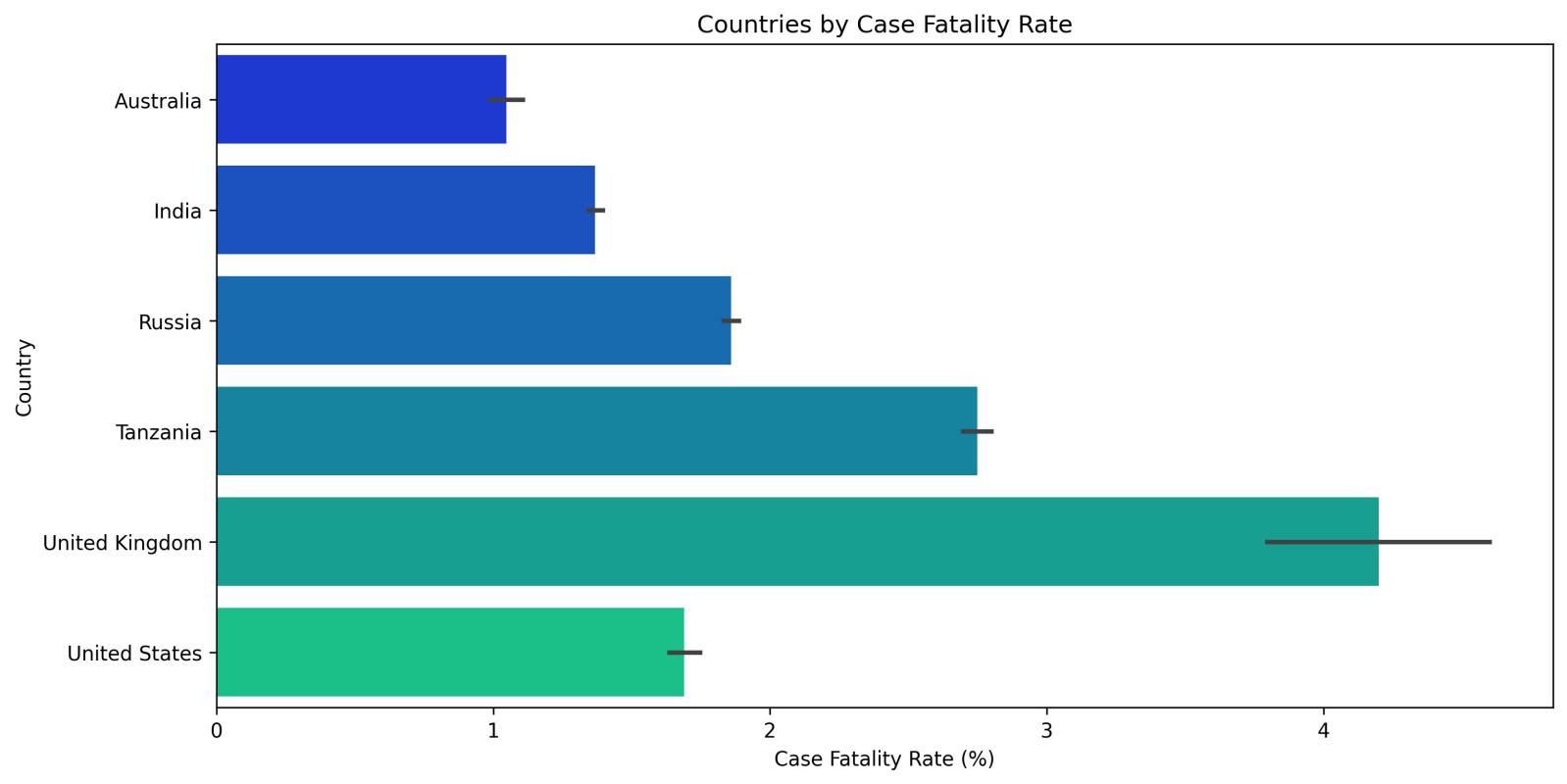
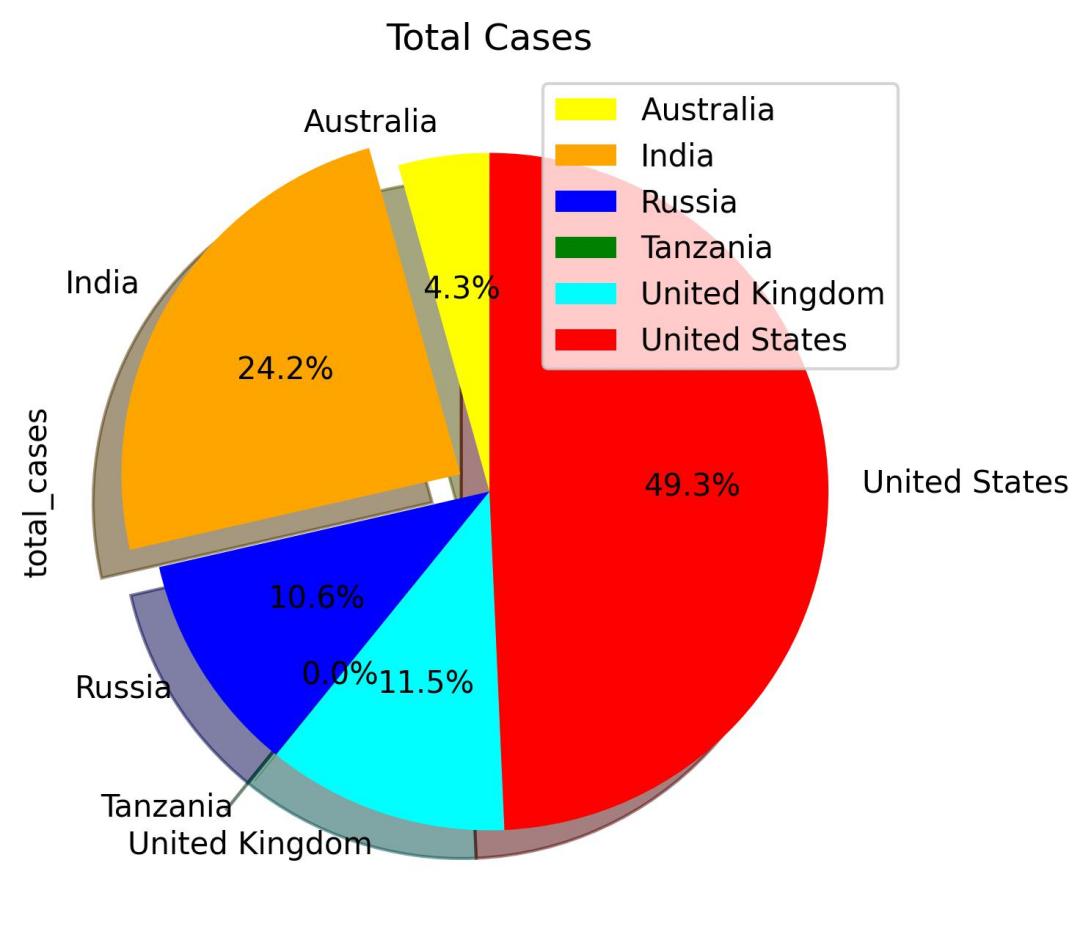
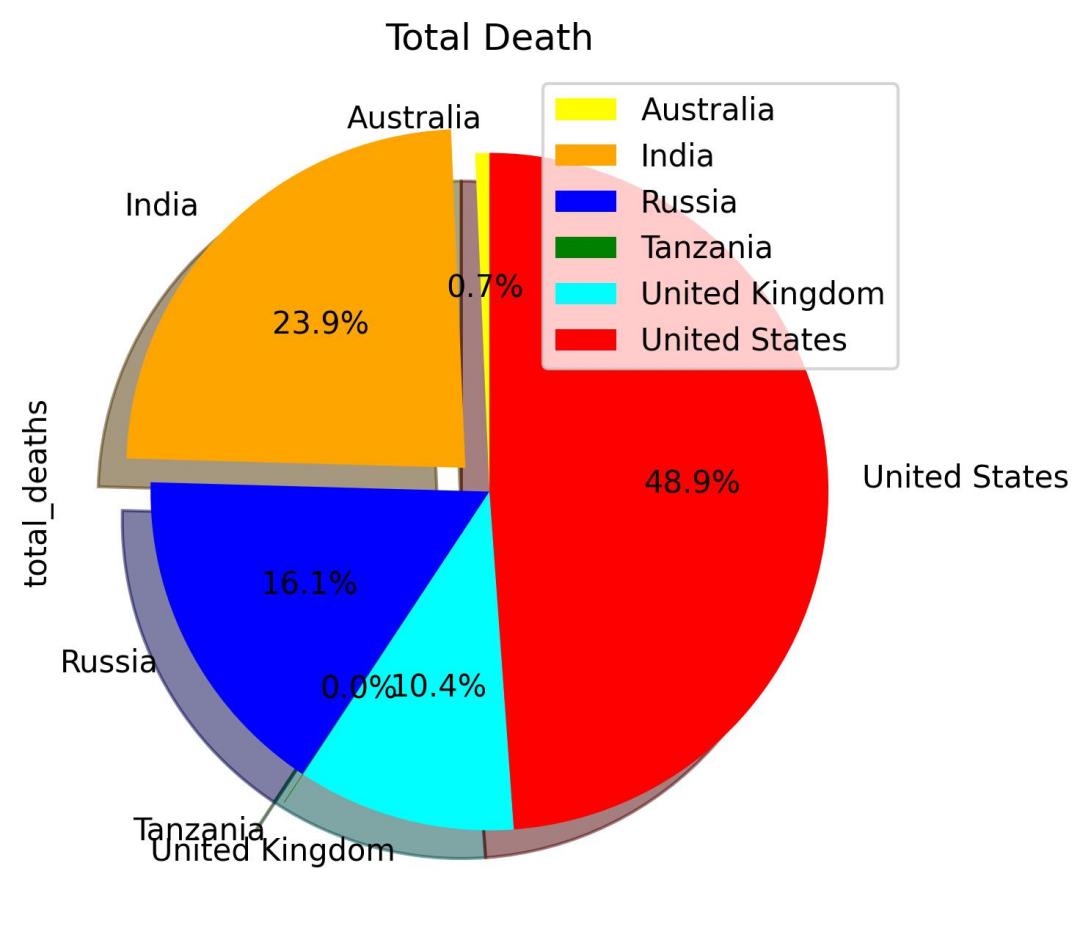
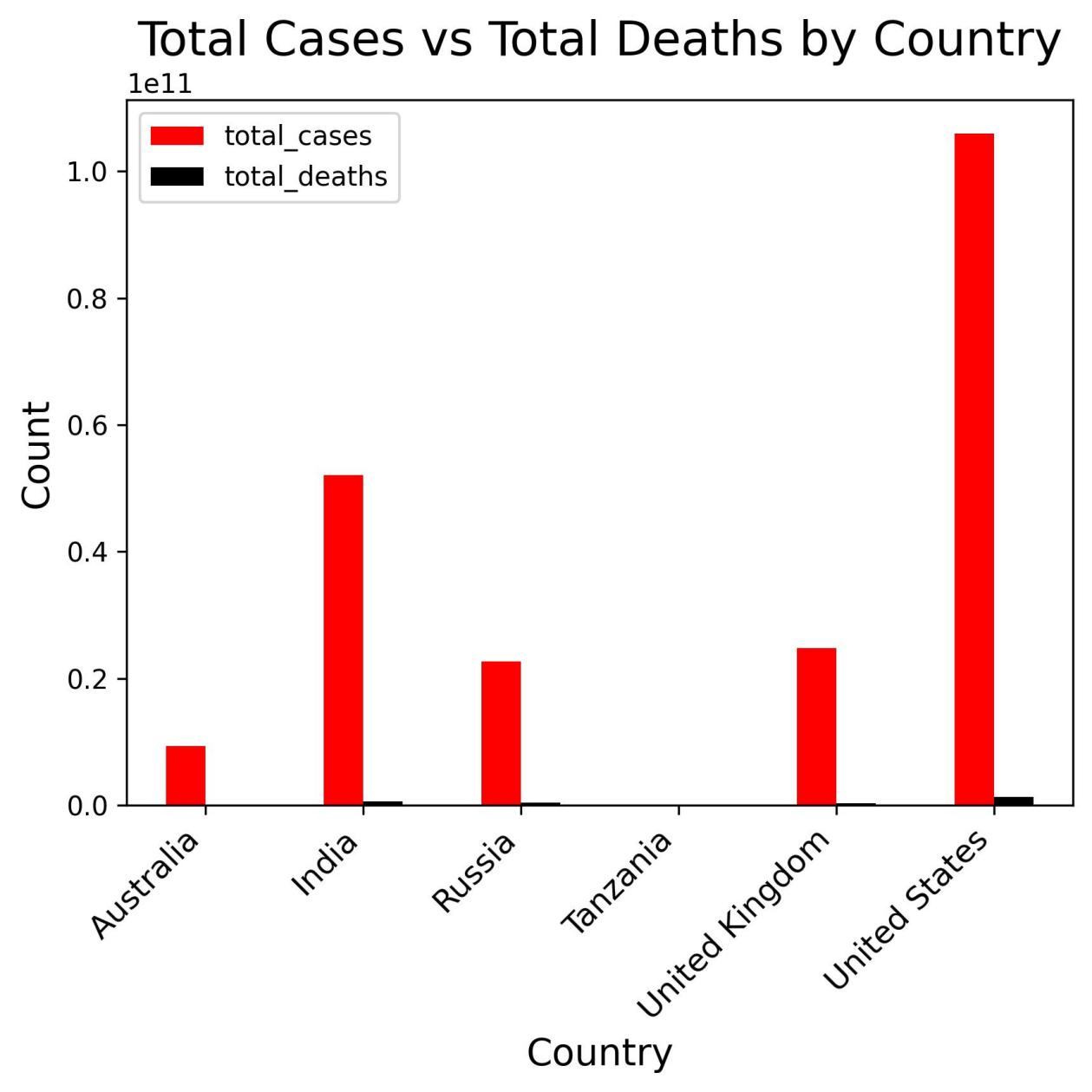
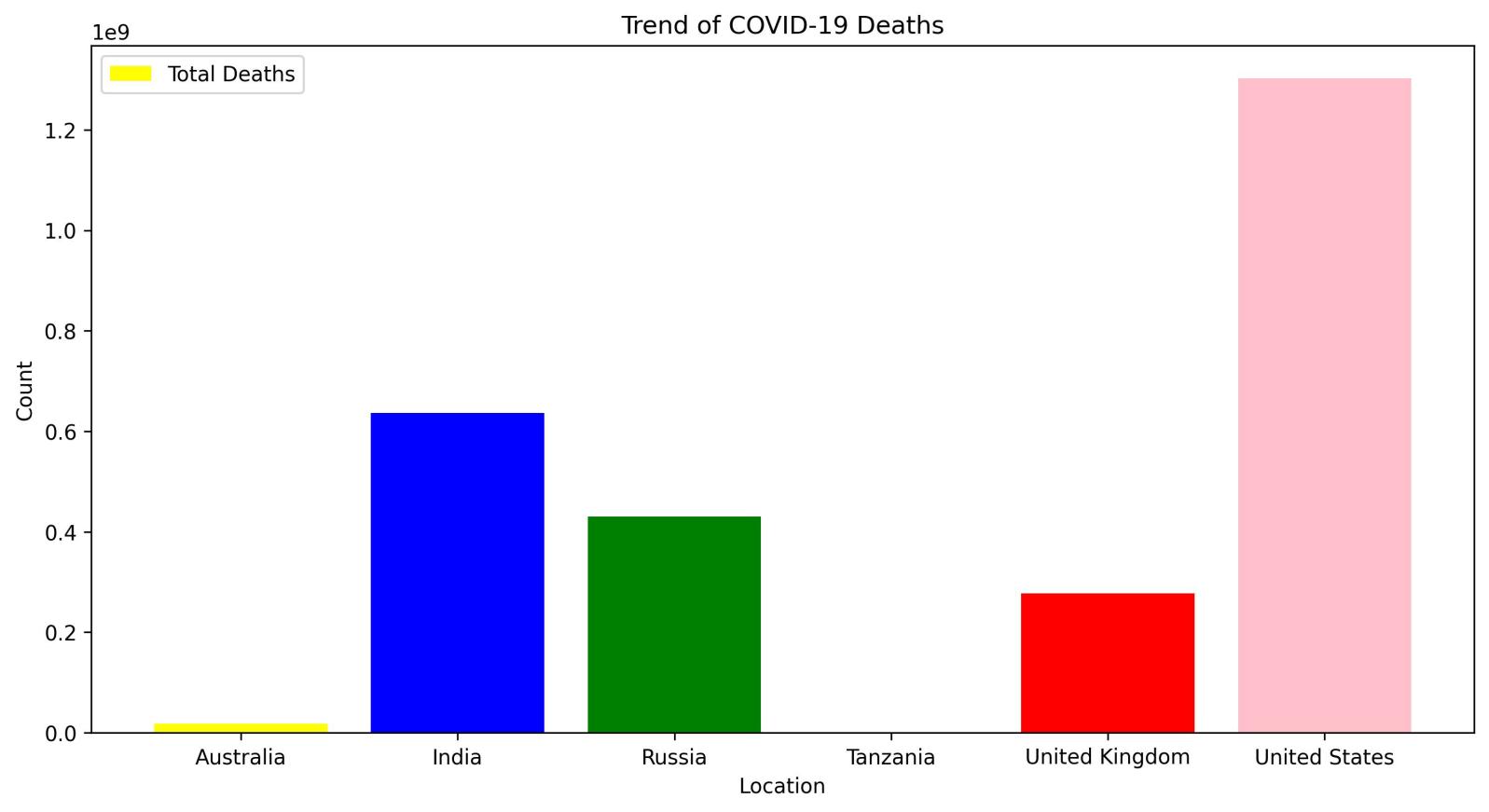
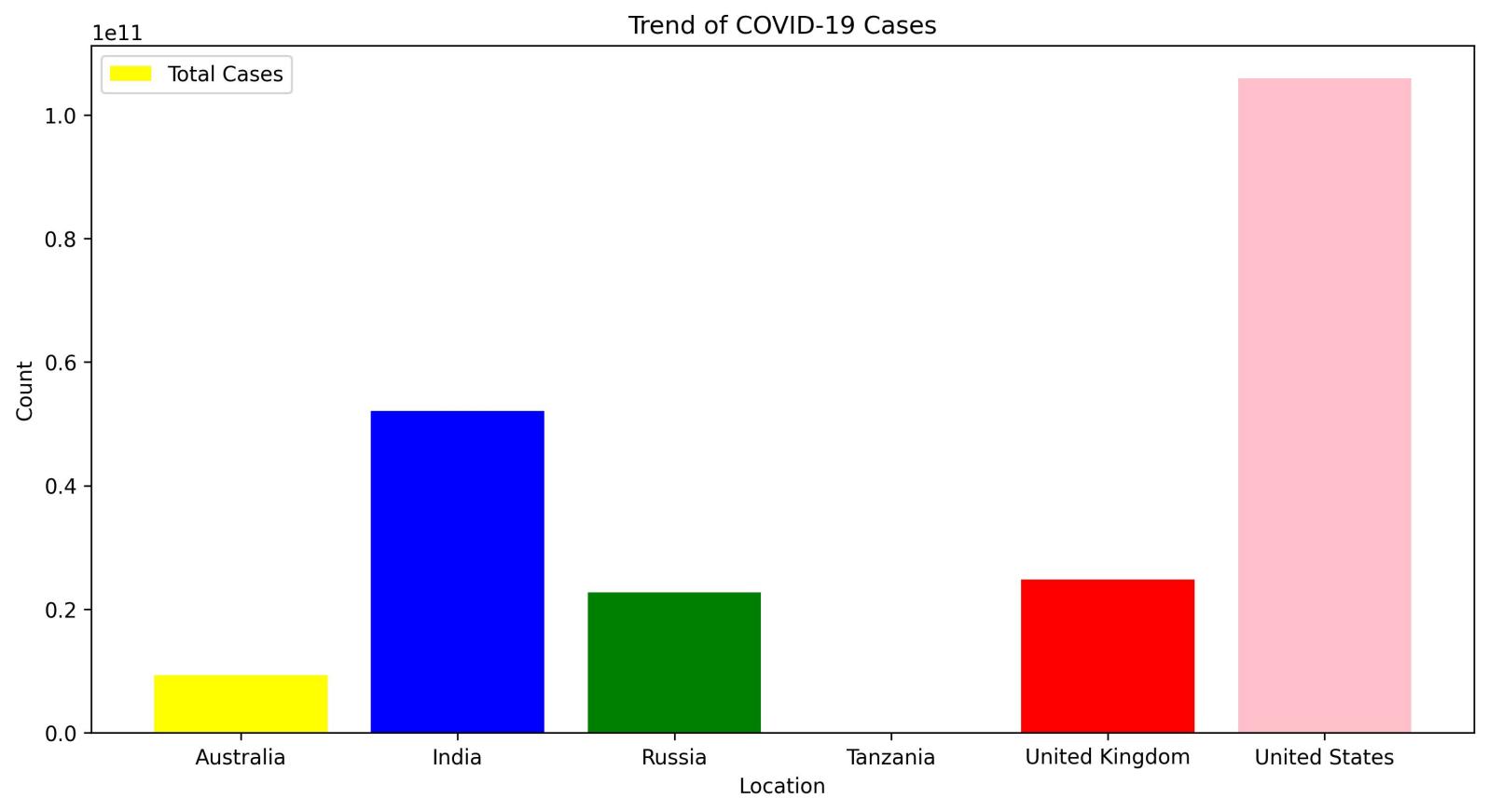
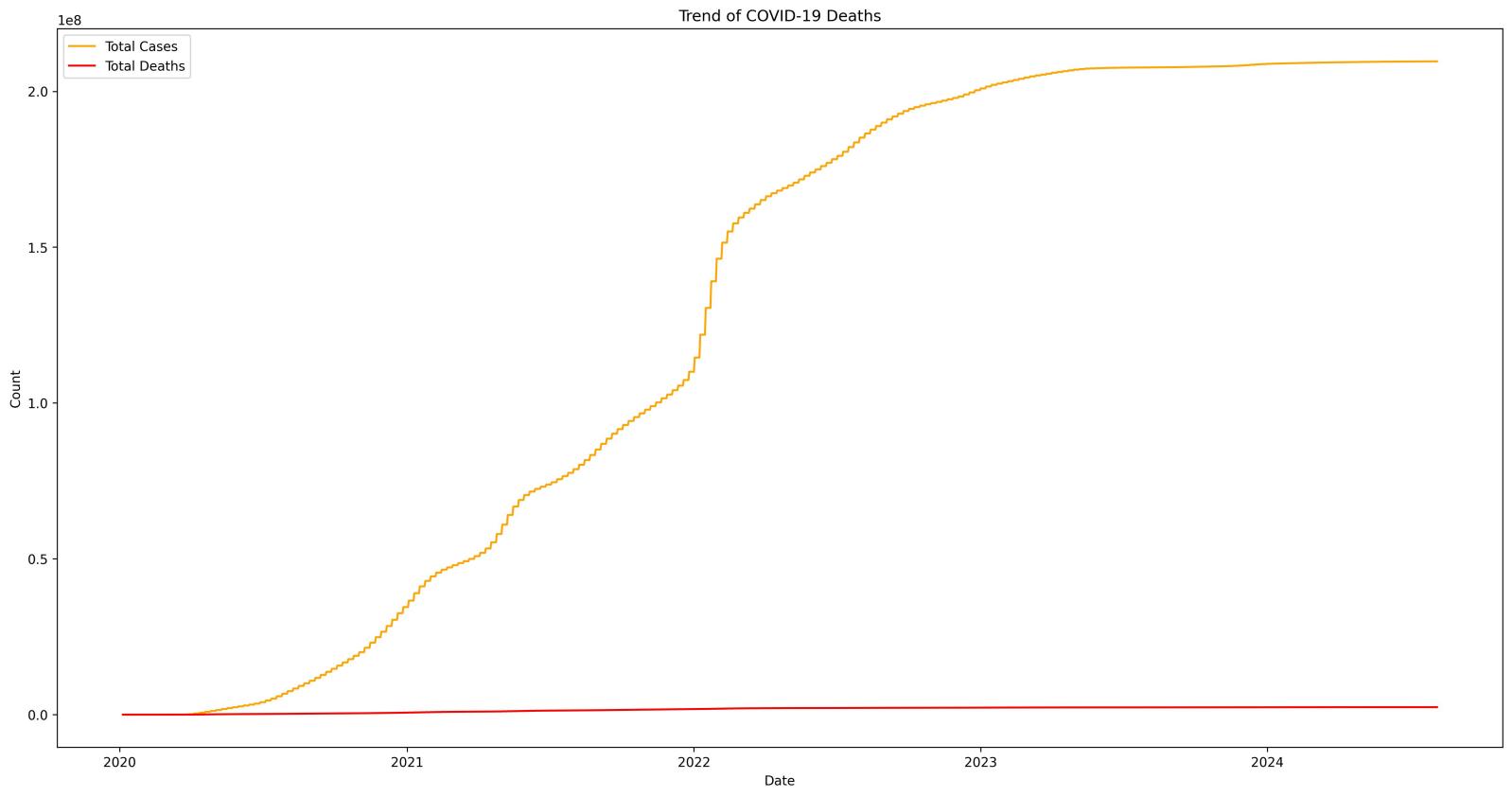
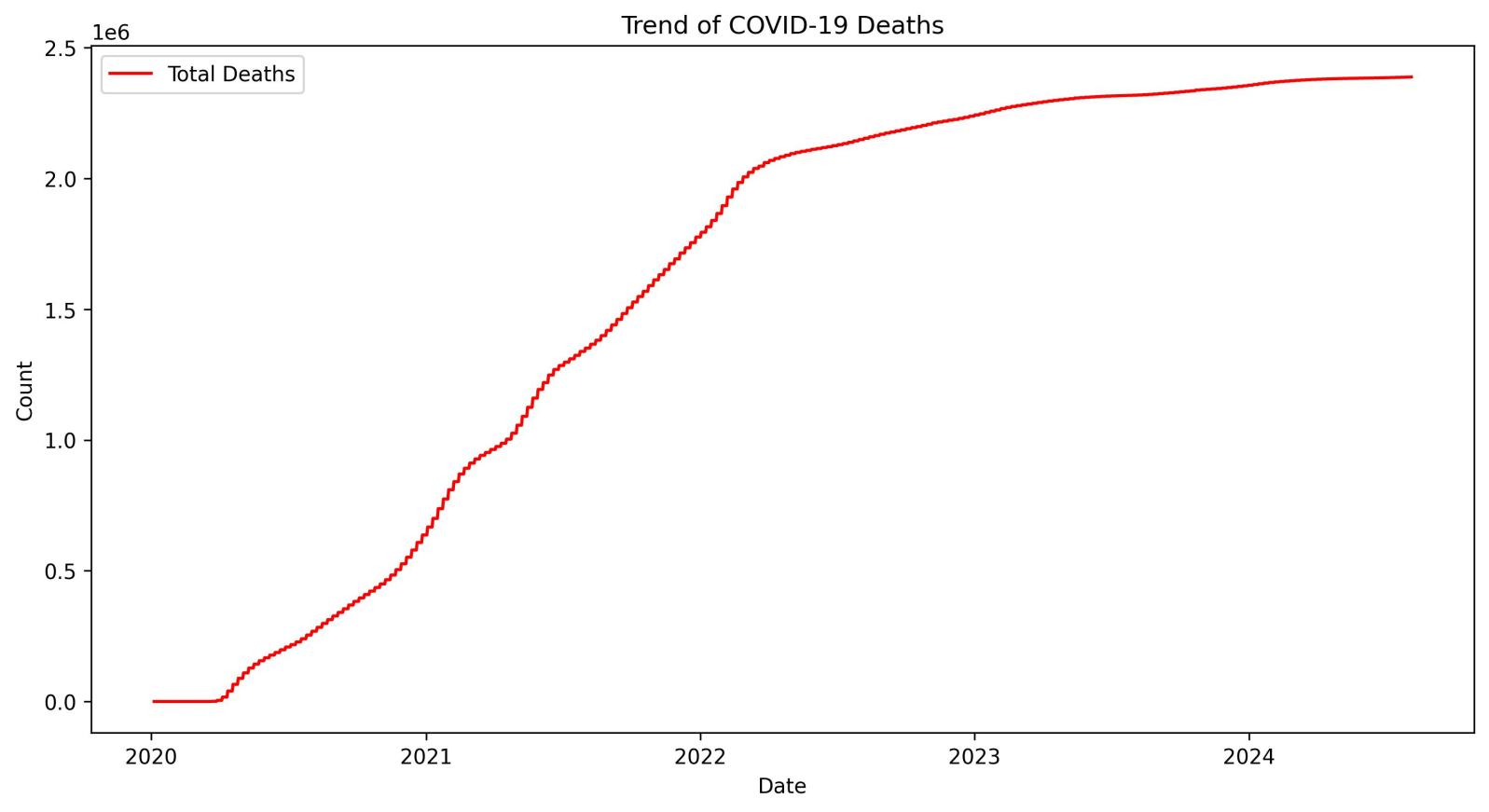
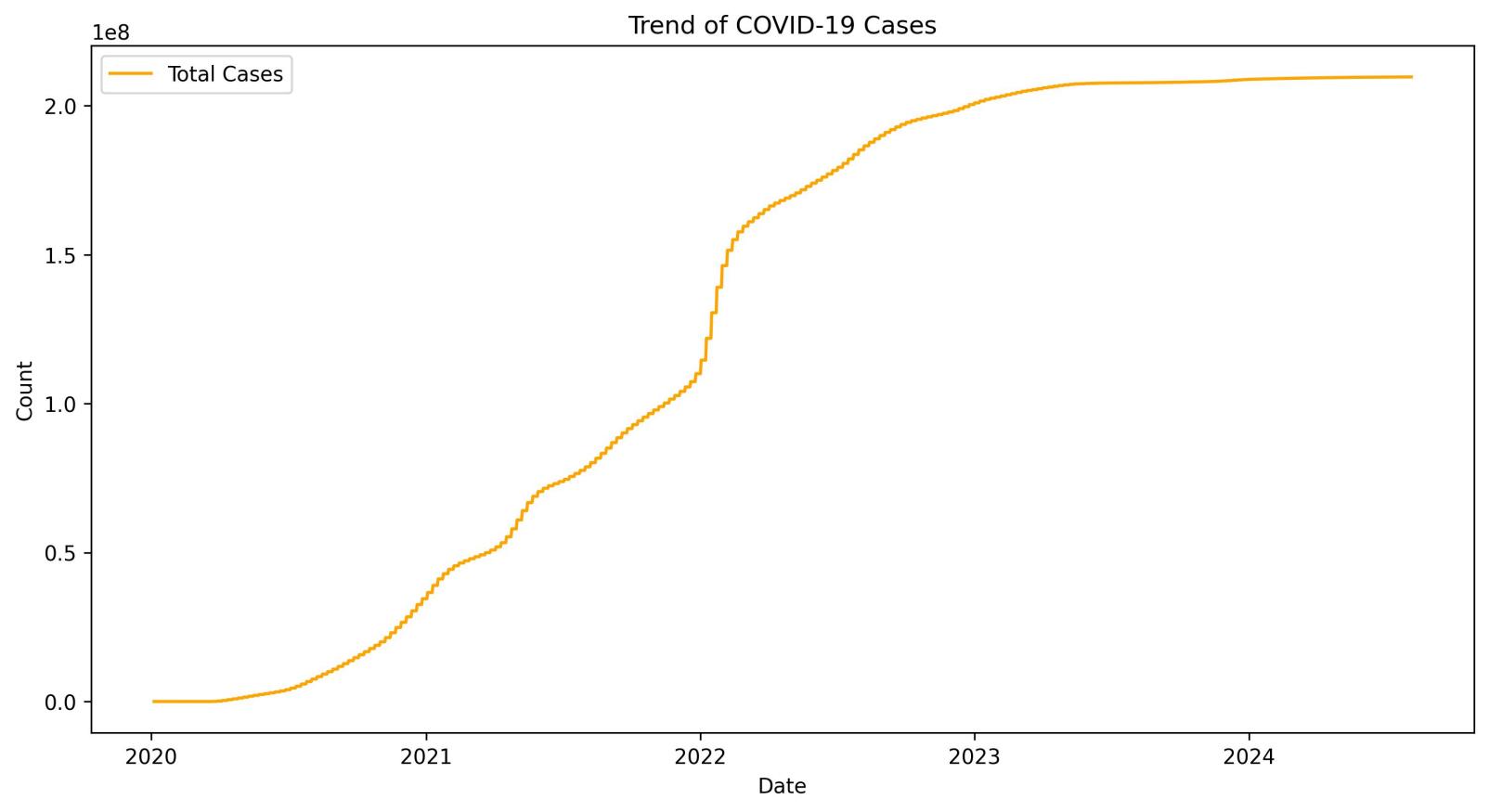
Trends in daily vaccinations over time.

**C. Correlation Analysis**

Correlation between variables:

High correlation observed between total\_cases and total\_deaths.

Weak correlation between vaccination rates and stringency\_index.

1. **Visualizations**

**Key Findings**

1. **6 Countries’ Trends comparisons**

A steep rise in total cases and deaths during specific waves (e.g., March 2020, January 2021).

Significant differences in vaccination progress among regions.

1. **Country-Specific Insights**

* Countries with high population density ( India, USA, Russia) showed higher total cases.

1. **Government Response**

Higher stringency levels correlated with slower growth in cases but not directly with vaccination rates.

1. **Vaccination Impact**

Countries with higher vaccination coverage had significantly lower death-to-case ratios during subsequent waves.

**Challenges Faced During the Analysis**

While working on this project, several challenges were encountered that required careful handling and problem-solving. Below is an outline of the main issues and how they were addressed:

1. Data Quality Issues

Problem:

The dataset contained missing values in critical columns such as total\_cases and total\_deaths. Missing data can lead to biased or incomplete analysis.

Solution:

Dropped columns with too many missing values if they were not critical to the analysis.

Filling with zero(0)

2. Inconsistent Date Formats

Problem:

The date column had inconsistent formats, causing errors when converting it to datetime objects for time-series analysis.

Solution:

Standardized the date format using Python's pd.to\_datetime() function and specified the correct format (%d-%m-%Y) to resolve parsing errors.

3. Comparing Countries with Different Population Sizes

Problem:

Countries with larger populations naturally had higher case counts, making direct comparisons unfair.

Solution:

Used per capita metrics, such as cases or deaths per 100,000 people, for fair comparisons.

4. Time Lag Between Cases and Deaths

Problem:

There is often a time lag between the increase in cases and reported deaths, complicating direct analysis.

Solution:

Used lagged correlations to understand the relationship between cases and deaths over time.

5. Complex Relationships Between Variables

Problem:

Relationships between variables (e.g., stringency index and total cases) were not always linear and required advanced techniques to explore.

Solution:

Used scatterplots, heatmaps, and statistical tests to understand these relationships better.

Explored non-linear trends where applicable.

6. Visualization Challenges

Problem:

Generating clear and interpretable visualizations with a large and complex dataset was challenging.

Solution:

* Filtered data to focus on key trends (e.g., 6 countries chosen for this analysis).
* Used advanced visualization tools like Seaborn and Matplotlib for dynamic and detailed plots.

By addressing these challenges systematically, the analysis was able to produce meaningful and reliable insights, ensuring the data's integrity and relevance to the project's objectives.

**Recommendations**

* Strengthen public health measures in densely populated regions.
* Improve vaccination strategies in low-coverage countries.
* Conduct further analysis of government policies and their effectiveness in reducing deaths.