**PROJECT REPORT**

Title:

**Exploratory Data Preparation and Visualization of Global COVID-19 Statistics**

Submitted by:

**Troy Tristan Jacob**  
BSCS-3A

Submitted to:

**Mr. Michael H. Indico, MM, MEng**

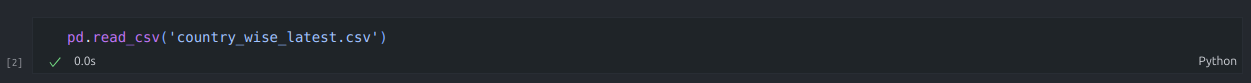
Instructor

## **Project Overview**

This project performs an in-depth **data analysis** on global COVID-19 data, focusing on total cases, deaths, recoveries, and population impacts by country and region. The analysis includes data cleaning, transformation, outlier handling, and visualization to uncover insights and trends.

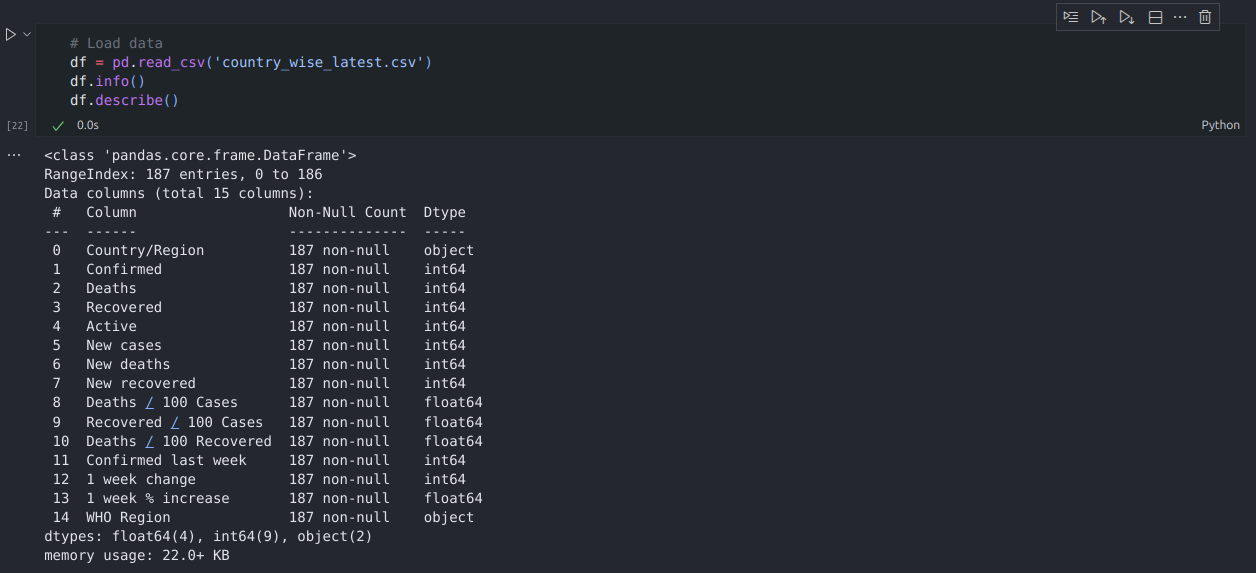
**I. Data Acquisition**

The CSV file, named country\_wise\_lastest.csv, was download from google classroom and loaded it in python using pd.read\_csv()



**II. Data Inspection**

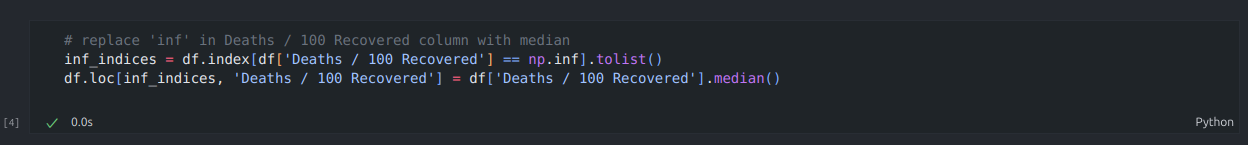
To identify inconsistencies, data structure issues, and missing or incomplete values in the dataset, Pandas' info() and describe() functions were used to inspect the structure, data types, and summary statistics.



**III. Data Cleaning**

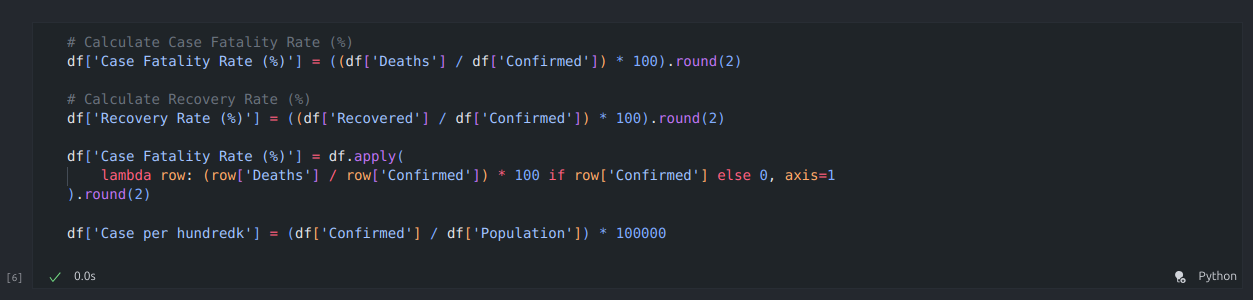
The dataset had only two inconsistent country names: "The United States" was labeled as "US", and "Taiwan" contained an extra character at the end. These inconsistencies were corrected manually.

Additionally, five missing values were identified during data inspection in the column “Deaths / 100 Recovered”, which were filled using Pandas’ median() function.



**IV. Data Transformation**

Three new columns were added to the dataset: Case Fatality Rate, Recovery Rate, and Cases per hundredk People. However, upon examining the dataset, it was found to be missing a crucial column — "Population", which was necessary for calculating population-based metrics.

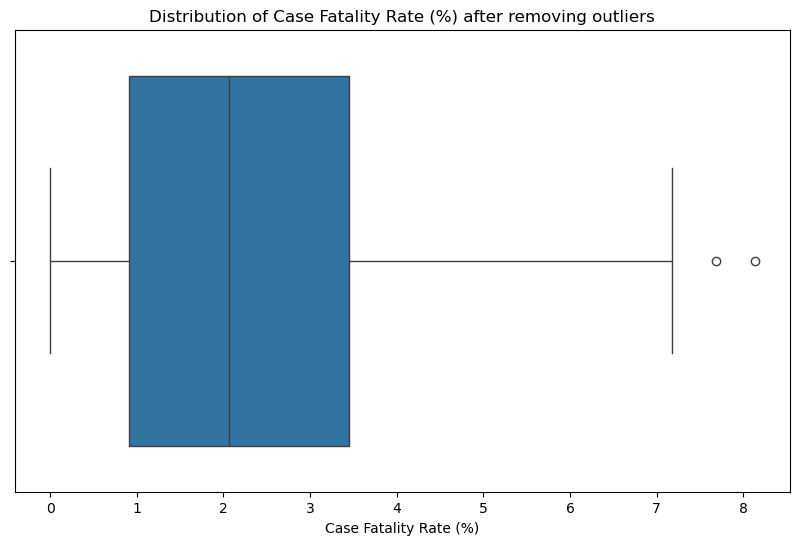
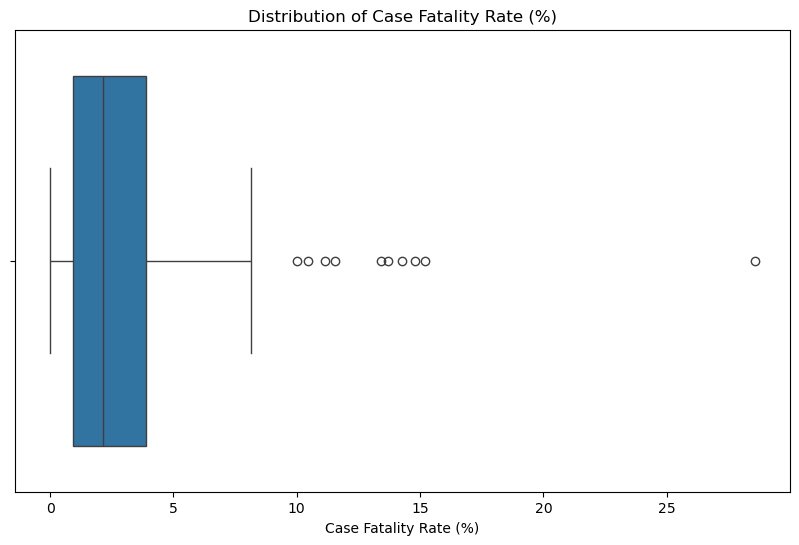


To address this, the Population column was added using the pycountry and pypopulation libraries to retrieve the population of each country. For countries that were not identified by the program, their population values were filled manually using the median() function.



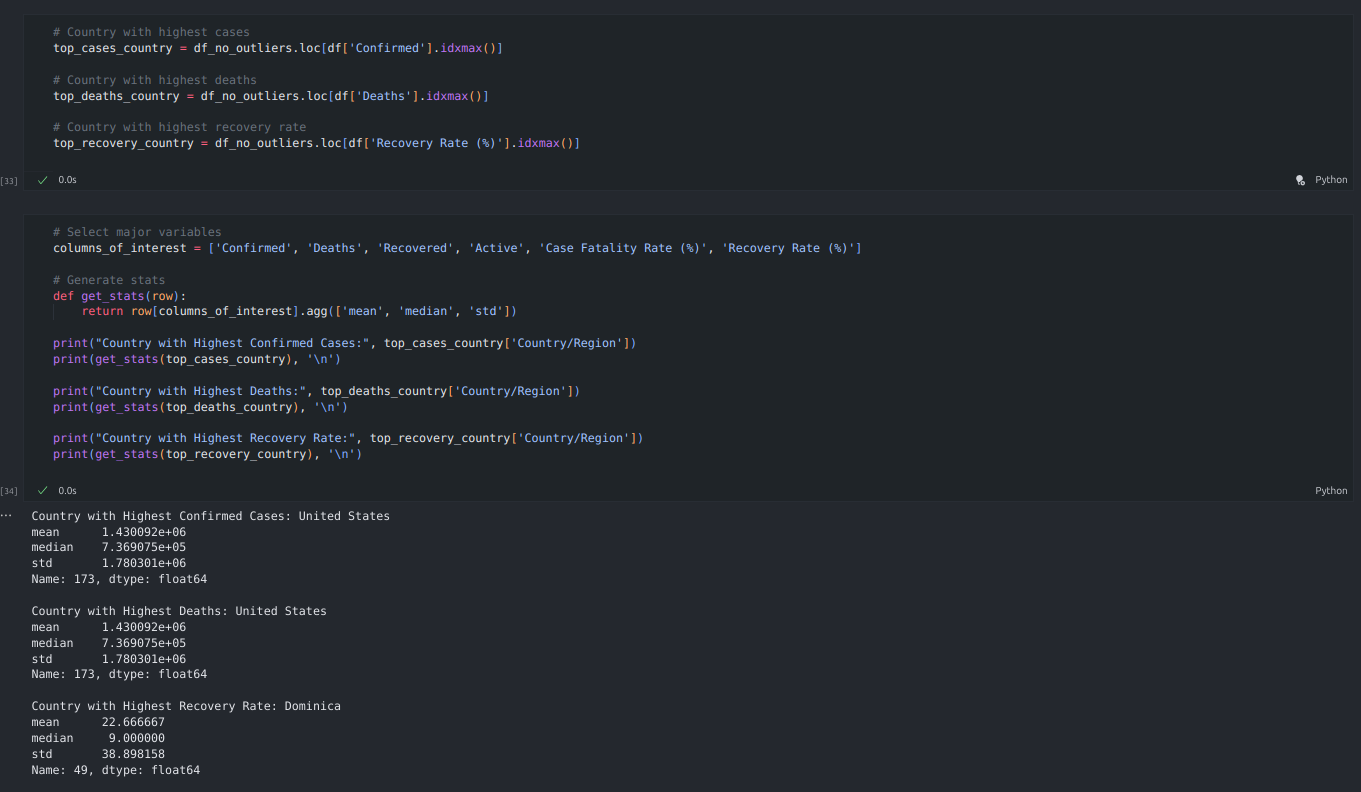
**V. Outlier Detection and Removal**

Outliers in numerical columns were identified through boxplots, and handled using the Interquartile Range (IQR) method, which filters out values that fall significantly outside the typical data range.



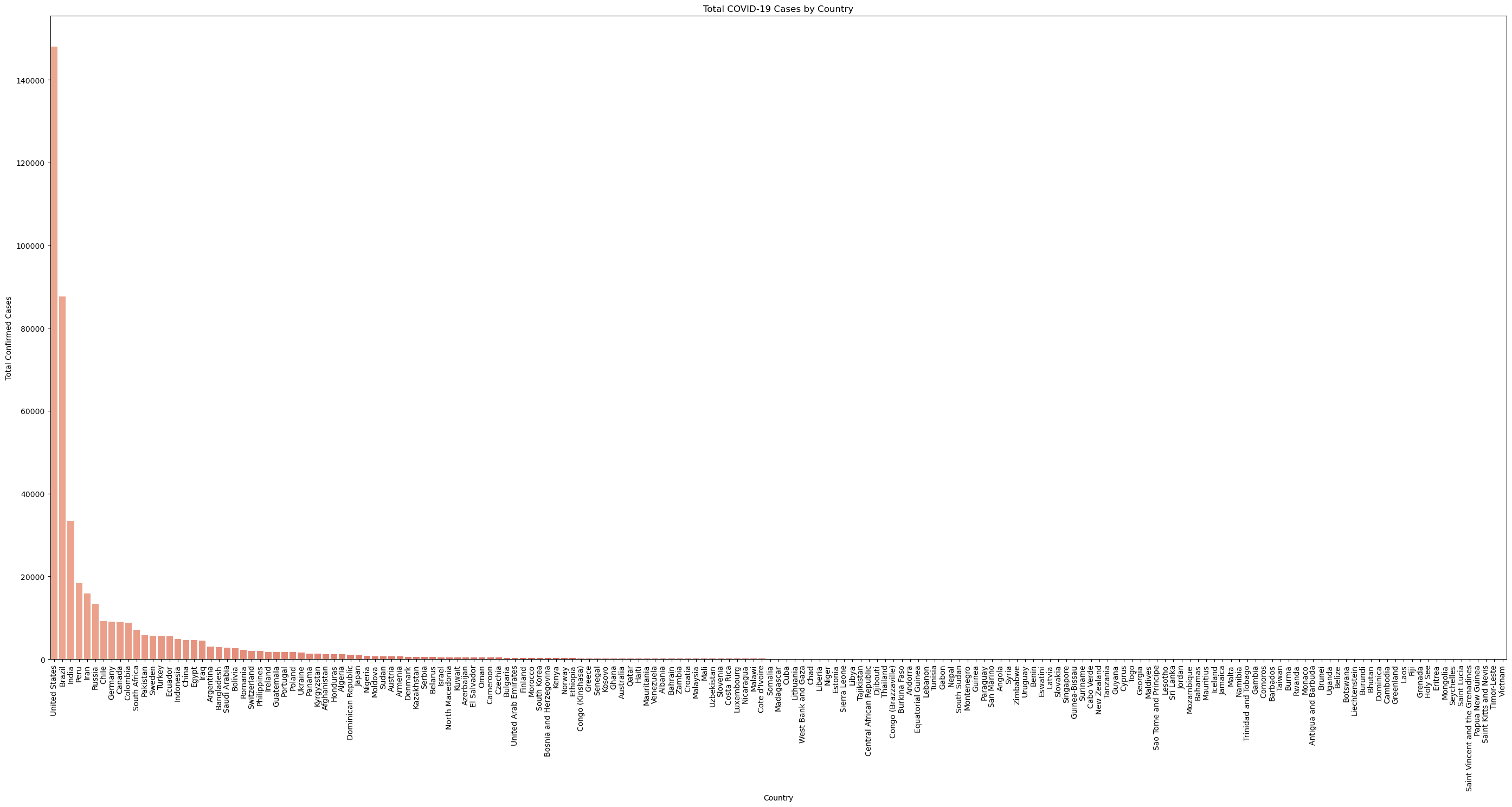
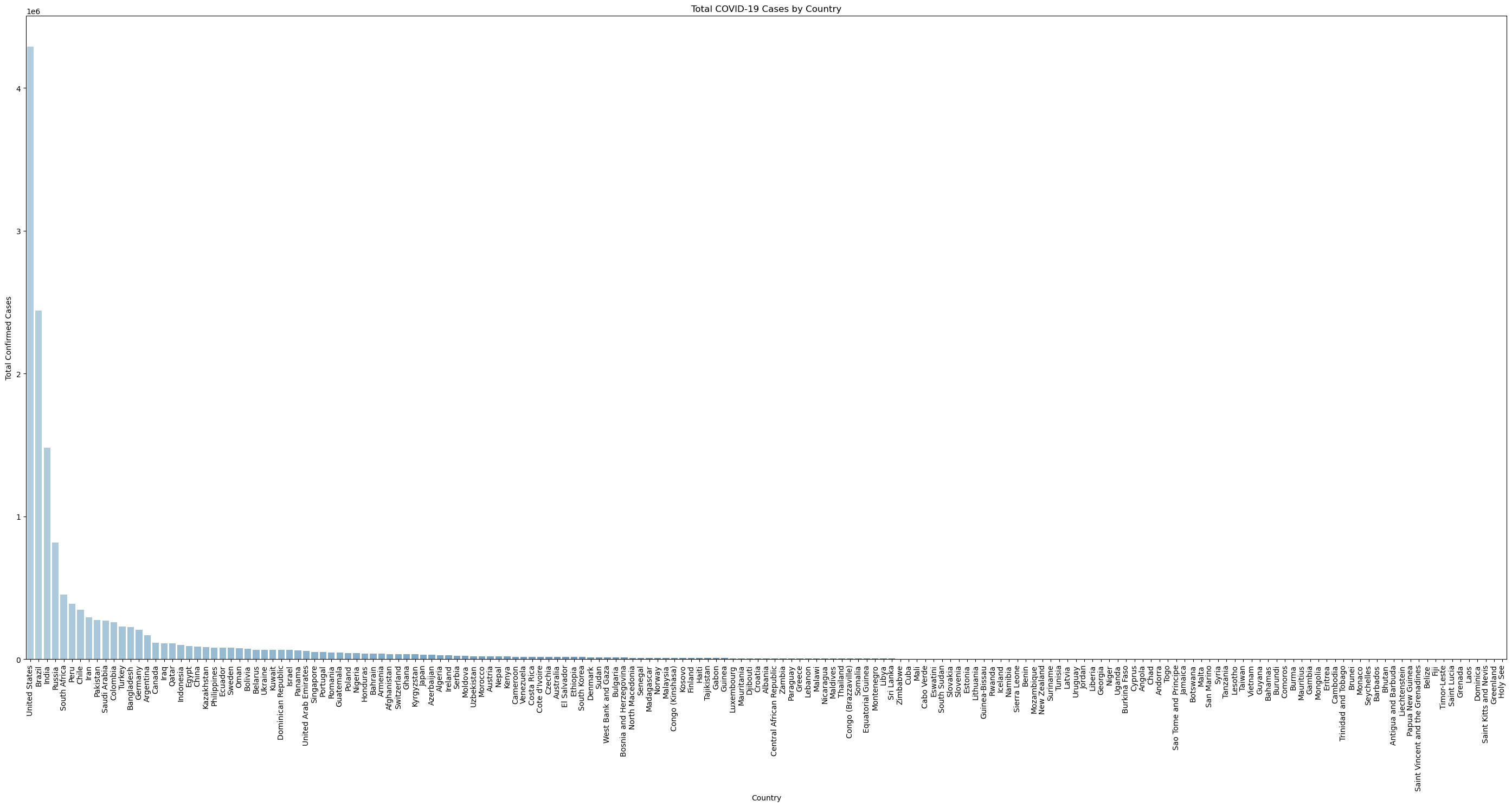
**VI. Descriptive Statistics**

To generate summary statistics for the countries with the highest cases, deaths, and recovery rates, Pandas’ agg() function was used. The analysis revealed that the United States had the highest number of confirmed cases and deaths. Furthermore, Dominica had the highest recovery rate among all countries.

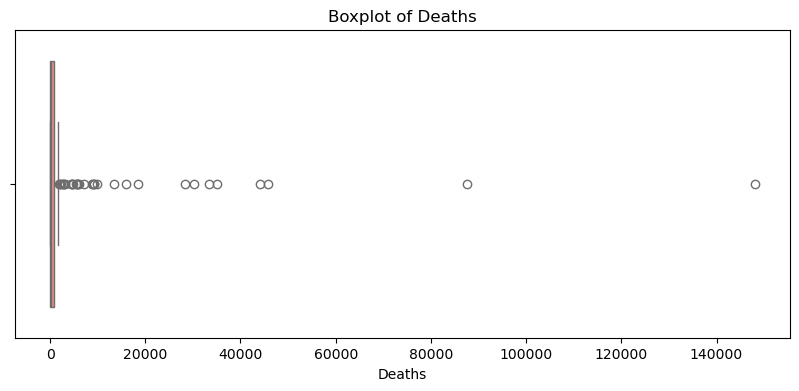
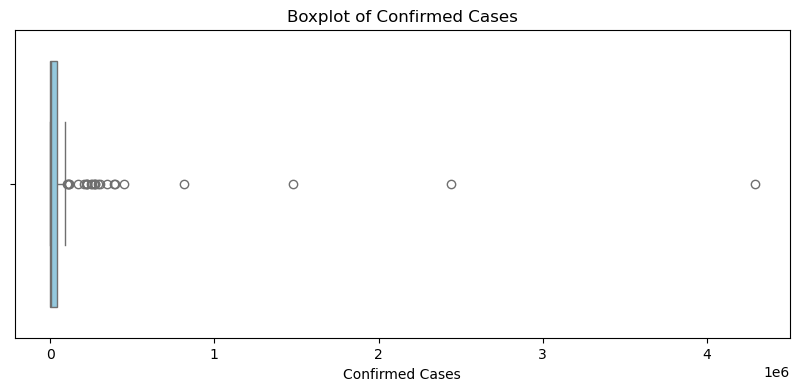


**VII. Data Visualization**

**Univariate Analysis**



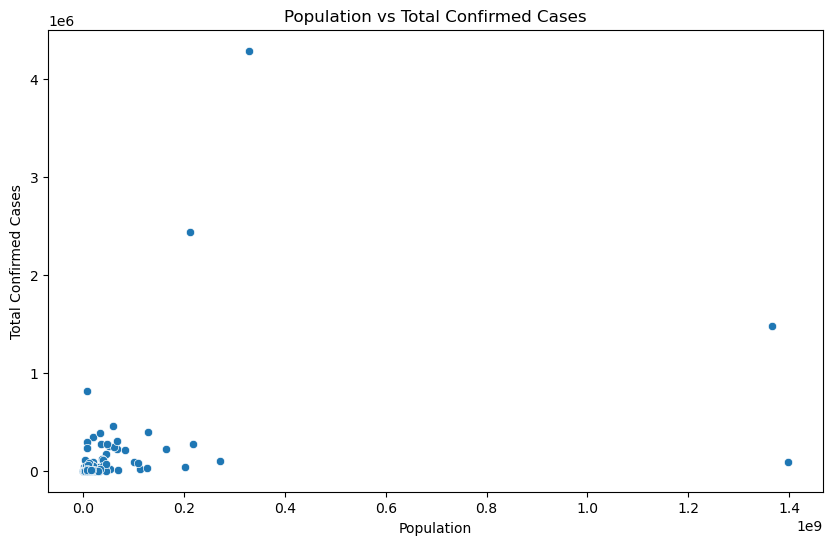
*Histogram: Distribution of total cases and deaths.*



*Boxplot: Detect outliers in deaths and cases.*

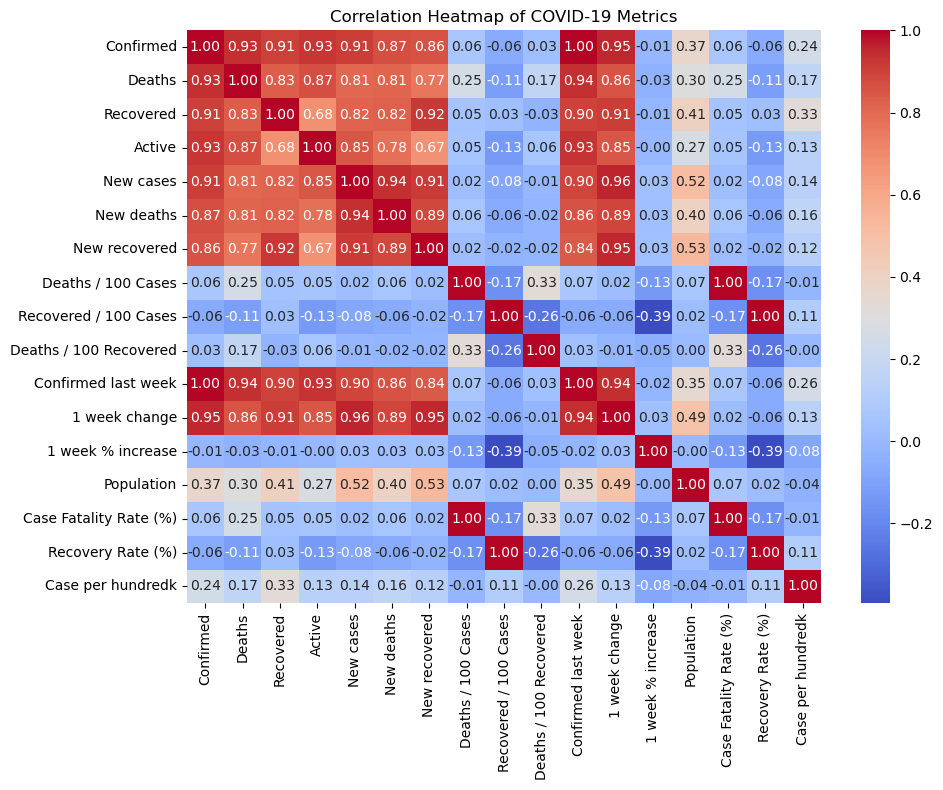
**Bivariate Analysis**

The scatter plot below reveals that a country with large population does not necessarily mean had high number of confirm cases.



*Scatter Plot: Relationship between population and total cases.*

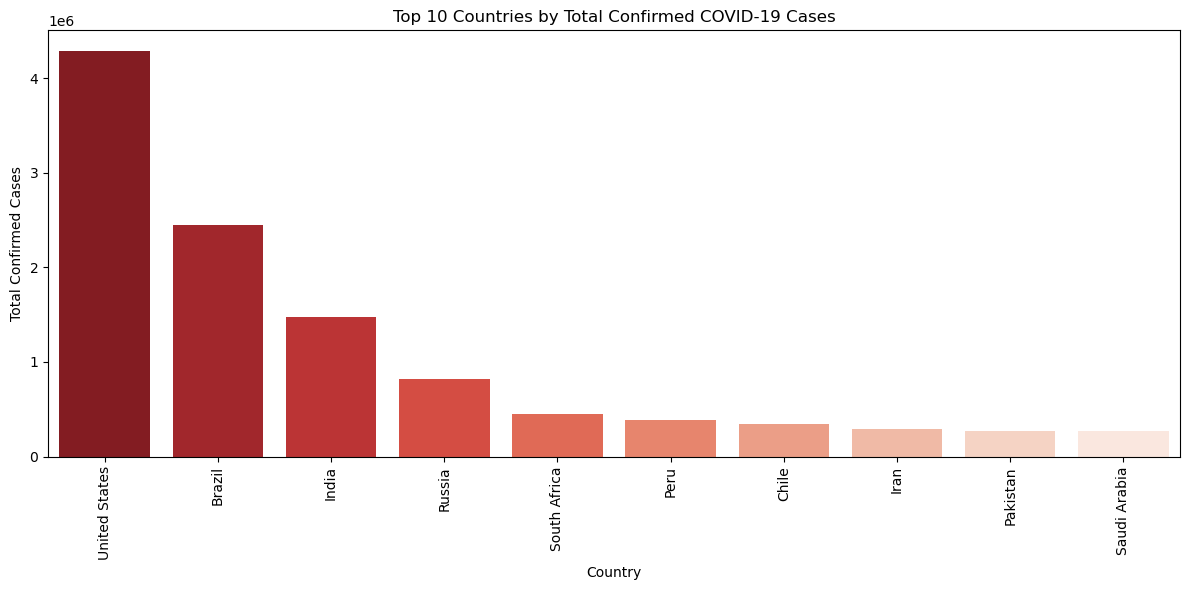
The heatmap shows that majority of at the top-left side of the matrix had a positive correlation meaning they have strong positive relationship.



*Heatmap: Correlation between numeric variables.*

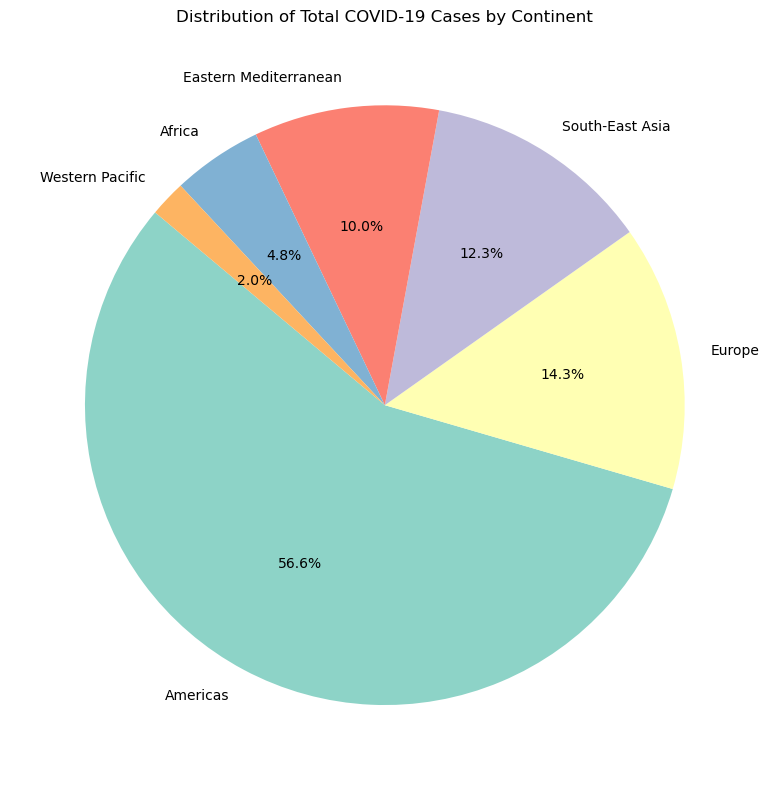
**Comparative Visualizations**

The analysis uncovers that United State had the most Covid-19 cases recorded.

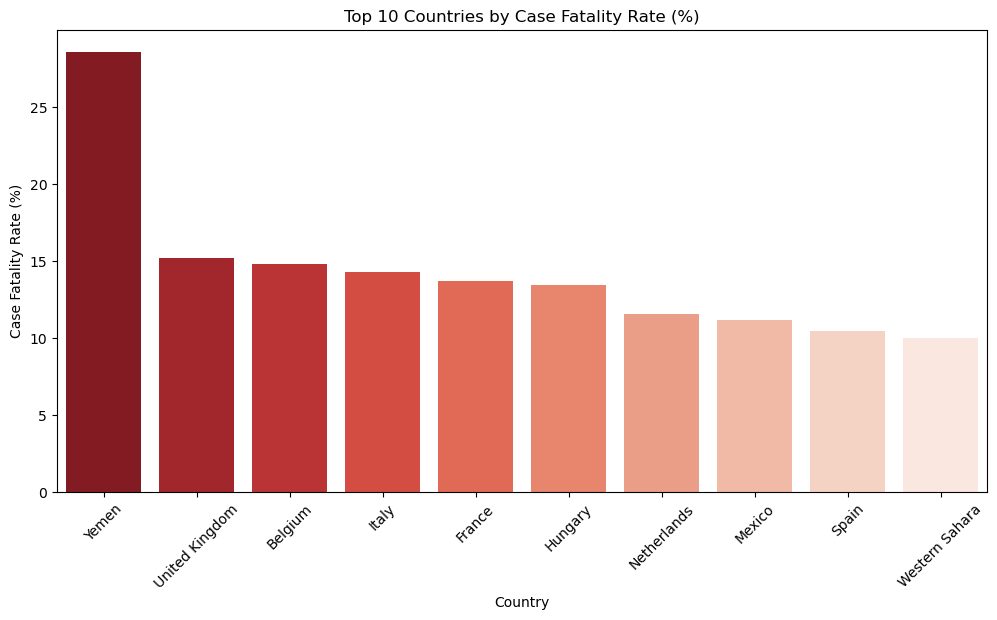


*Bar Chart: Top 10 countries by total cases.*

The region that had the most Covid-19 confirmed cases in the world were North and South America. Meanwhile, Western Pacific had the lowest confirmed cases.



*Pie Chart: Distribution of total cases by continent or region.*



*Bar Chart: Top 10 countries by CFR.*

**VIII. Insights and Interpretation**

The analysis revealed that among all countries, Yemen had an extremely high Case Fatality Rate (CFR). This indicates the severity of COVID-19 in the country. Additionally, the heatmap showed a strong positive correlation between confirmed cases and deaths, reinforcing that the virus was highly dangerous and often fatal in areas with widespread infection. Furthermore, the region of Americas recorded the highest confirmed cases of COVID-19.

**IX. Conclusion**

Data preparation — including cleaning, handling missing values, and adding population data — greatly improved the dataset’s accuracy and allowed for more meaningful analysis. New columns like Case Fatality Rate and Cases per 100k people enabled fair comparisons across countries.

The analysis revealed that Yemen had the highest fatality rate, the United States had the most cases and deaths, and the Americas region was the most affected overall. A strong correlation was found between confirmed cases and deaths.

However, limitations such as missing time-series, vaccination, and testing data reduced the depth of the analysis. Future studies should include these elements, along with daily trends and geospatial data, for a more comprehensive understanding.