COVID-19 Data Analysis Report

**1. Introduction**

**Project Overview**

The COVID-19 pandemic had a significant impact worldwide, affecting millions of people and leading to severe economic and healthcare challenges. This analysis aims to uncover key insights from COVID-19 data, focusing on total cases, deaths, death rates, and monthly trends across different countries. By leveraging Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn, this project provides a data-driven understanding of how the virus spread over time.

**Objectives**

The primary objectives of this analysis are:

* To analyze the total cases and deaths across different countries.
* To calculate death rates based on available data.
* To transform daily data into monthly aggregated values for better trend analysis.
* To visualize trends using bar charts, line graphs, and other relevant plots.
* To identify anomalies or inconsistencies in data recording across countries.

**2. Data Cleaning and Processing**

**2.1 Data Sources**

The datasets used in this analysis include:

* **time\_series\_covid19\_confirmed\_global.csv** (for confirmed cases)
* **time\_series\_covid19\_deaths\_global.csv** (for death cases)

Both datasets were sourced from Johns Hopkins University's GitHub repository.

**2.2 Data Cleaning and Preprocessing Steps**

**Step 1: Checking Data Consistency**

* Checked the shapes of both datasets to ensure they had matching structures.
* Verified that the column names and country names were consistent across both datasets.

**Step 2: Handling Missing and Unnecessary Columns**

* Dropped the columns **'Province/State', 'Lat', and 'Long'** since they were not needed for country-level analysis.
* Checked for missing values in key columns to ensure data integrity.

**Step 3: Converting Date Columns**

* The dataset originally had date columns in string format (e.g., '1/22/20').
* Converted these columns to **datetime format** for accurate time-based analysis.

**Step 4: Merging Cases and Deaths Data**

* Aggregated total cases and deaths per country.
* Merged the datasets using the **'Country/Region'** column, ensuring each country had corresponding case and death counts.

**Step 5: Transforming Data to Monthly Aggregation**

* The original dataset contained daily case counts, which made analysis complex.
* Instead of summing up cases, we extracted the **last available day's value per month** to maintain cumulative case tracking.
* Created separate monthly datasets for cases and deaths.

**Step 6: Calculating Death Rate**

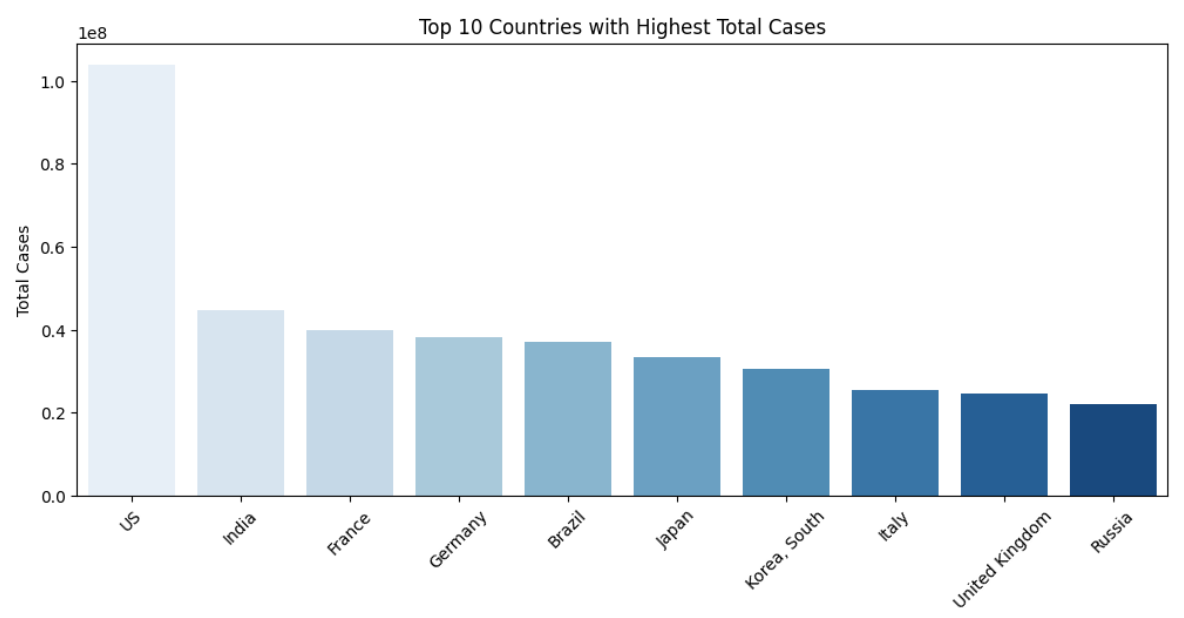
* Since the data was cumulative, the total number of cases and deaths could be determined from the last recorded month's data.
* The **Death Rate** was calculated as: (Total Deaths/Total Cases)\*100
* This calculation was performed using the final month’s values for accuracy.

**Step 7: Finalizing the Dataset**

* Sorted the dataset with months in chronological order.
* Ensured country names were properly formatted.
* Added a 'Total' row summarizing the global statistics.
* Saved the cleaned dataset as **cleaned\_covid19\_data.csv** for further analysis and visualization.

**3. Data Analysis and Insights**

**3.1 Top 10 Countries with Highest Total Cases**

****

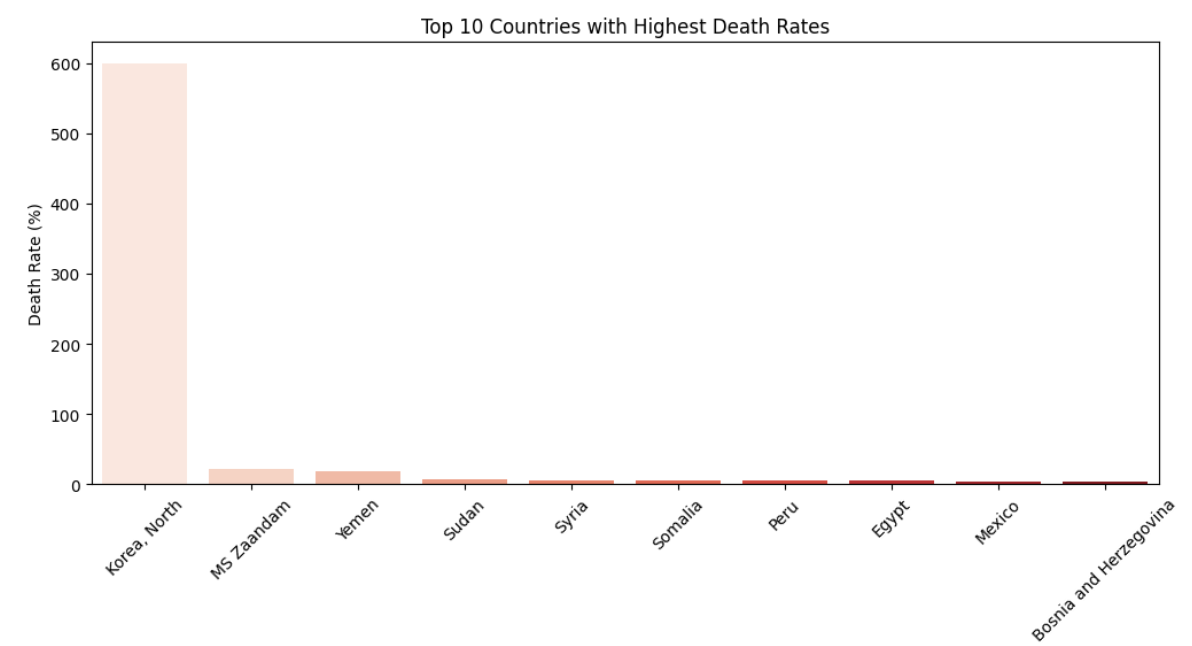
**Insight:**

* The **United States (US)** had the highest total cases, followed by **India, France, Germany, and Brazil**.
* The case counts significantly dropped after the first few waves, particularly after widespread vaccination campaigns.
* Countries like **Japan, South Korea, and Italy** also showed high case numbers but a relatively moderate death rate.

**Possible Explanation:**

* Countries with high population density and large international travel hubs tended to have higher total cases.
* Efficient vaccination drives and lockdown measures may have impacted the spread of the virus.

**3.2 Top 10 Countries with Highest Death Rate**

****

**Insight:**

* Surprisingly, **North Korea** reported a death rate of around **600%**, which appears highly unrealistic. This anomaly could indicate either a data entry error or misrepresented records.
* Other countries with high death rates included **MS Zaandam (a cruise ship)**, **Yemen, Sudan, and Syria**, where healthcare infrastructure may have been severely affected by conflict or underdevelopment.

**Recommendation:**

* This unusually high death rate in some regions may require further investigation.
* Countries like **Yemen and Sudan** may benefit from international health assistance to reduce the mortality rate.

**3.3 Countries with Lowest Death Rate**

**A graph of a number of people

AI-generated content may be incorrect.**

**Insight:**

* Several countries reported a **0% death rate**, such as **Antarctica, Holy See, and some small island nations**. This could imply:
  1. **No COVID-19 cases reported.**
  2. **No deaths recorded due to minimal outbreaks.**
  3. **Data not being reported accurately.**
* Countries like **Singapore and Brunei** had the lowest death rates at around **0.08%**, which demonstrates the effectiveness of their healthcare systems and preventive measures.

**Recommendation:**

* Verify the authenticity of countries with a reported **0% death rate**.
* Study the healthcare models of countries with low death rates to identify effective strategies for future pandemics.

**3.4 Monthly Trends of Cases and Deaths**

**A graph of the covid-19 virus

AI-generated content may be incorrect.**

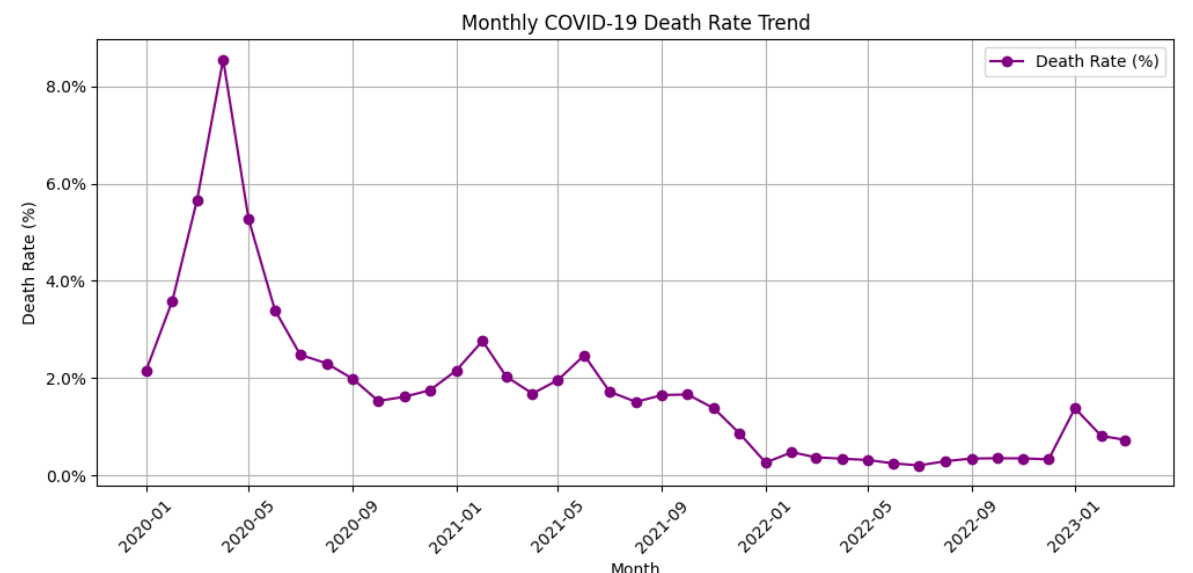
**Insight:**

* The line chart comparing monthly cases and deaths revealed that:
  + **Cases peaked in multiple waves**, particularly around **March 2021 and January 2022**.
  + **Deaths also peaked** but with a slight lag behind the rise in cases, indicating delays in reporting deaths or slower impact.
  + Vaccination rollouts seem to have played a significant role in reducing death rates after the initial peaks.

**Recommendation:**

* Continuously track monthly data for potential future waves.
* Correlate death peaks with healthcare infrastructure challenges.

**3.5 Death Rate Trend**

****

**Insight:**

* The death rate showed a sharp increase during the early phase of the pandemic (March 2020 to May 2020).
* After May 2020, the death rate started declining steadily, likely due to:
  + Improved healthcare infrastructure.
  + Mass vaccination campaigns.
  + Enhanced public awareness and preventive measures.

**Recommendation:**

* Track future death rate trends to measure the long-term impact of vaccination and policy measures.

**4. Key Findings and Recommendations**

**Key Findings:**

* **The United States** consistently reported the highest number of cases.
* **North Korea's death rate of 600%** is likely a data anomaly or error.
* Some countries reported **0% death rate**, which may require further investigation.
* Countries with robust healthcare infrastructure like **Singapore and Brunei** had the lowest death rates.
* Major case spikes were observed around **March 2021 and January 2022**, correlating with virus variants and lockdown policies.

**Recommendations:**

* Future research should focus on the role of vaccination and public policy in reducing mortality rates.
* Investigate the data anomalies in countries like **North Korea** and **Antarctica**.
* Provide international health assistance to countries with high death rates such as **Yemen, Sudan, and Syria**.
* Monitor long-term post-pandemic effects on healthcare and economic systems.

**3.5 Correlation Heatmap**

**A screenshot of a graph

AI-generated content may be incorrect.**

**Insight:**

* + A **strong positive correlation (0.88)** was observed between **Total Cases** and **Total Deaths**, indicating that as COVID-19 cases increased in a country, the number of deaths also increased.
  + The **Death Rate** showed a **near-zero correlation** with **Total Cases (-0.03)**, suggesting that **having high cases did not necessarily result in high death rates**. This may indicate the effectiveness of **healthcare systems, testing, and vaccination** in some countries.
  + A **very weak negative correlation (-0.02)** was found between **Total Deaths** and **Death Rate**, implying that some countries reported **high death counts** but maintained **lower death rates**, possibly due to **massive testing, fast treatment, and healthcare capacity**.
  + This heatmap also highlights that countries with the highest **Total Cases** did not always correspond to countries with the highest **Death Rate**, emphasizing the **importance of healthcare infrastructure** in reducing fatalities.

**Recommendation:**

* Conduct further research to analyze the impact of **healthcare infrastructure, policy measures, and vaccination campaigns** in reducing death rates despite high case counts.
* Examine the healthcare strategies in countries with a **high death rate** to identify opportunities for international health support.
* Include vaccination data in future analyses to understand its direct impact on reducing COVID-19 deaths.
* Perform a case study on countries like **Yemen, Sudan, and Syria**, which reported high death rates, to identify potential gaps in healthcare response and emergency services.

**3.5 Heatmap**

**A graph with red squares

AI-generated content may be incorrect.**

**Insight:**

* The United States had an exceptionally high peak of **over 20 million new cases** in **January 2022**, representing the **largest single-month case spike** among the top 10 countries.
* India experienced a **major spike in May 2021**, coinciding with the **Delta variant outbreak**, resulting in over **10 million cases** in a single month.
* Several countries like **Japan, Korea, and Italy** experienced periodic case spikes every few months, but none reached the intensity of the US or India.
* **South Korea** showed a sudden sharp increase in cases in **May 2022**, highlighting a potential outbreak or post-lockdown surge.
* Some countries like **Russia, Italy, and Brazil** had relatively consistent case levels without extreme spikes, suggesting better control measures or underreporting.

**Recommendation:**

* Further investigation is needed for the **United States' peak in January 2022** to identify the key factors (e.g., new variant, holiday season gatherings, etc.).
* India's sudden spike in May 2021 correlates with the **Delta variant spread**, suggesting that countries should prepare **healthcare infrastructure** when new variants emerge.
* Countries like **Japan, Korea, and Italy** should be analyzed for their repeated but moderate spikes, which may indicate **wave patterns** rather than one-time outbreaks.
* It would be helpful to conduct a trend analysis on countries like **Brazil and Russia**, which showed relatively stable but sustained high numbers of new cases.
* Understanding the factors behind **South Korea’s sudden spike in May 2022** could provide insights into post-pandemic surges and policy adjustments.

**5. Conclusion**

This analysis provided an in-depth look into the spread of COVID-19 across different countries, focusing on total cases, deaths, and monthly trends. By cleaning and transforming the dataset, we achieved a clearer view of how the pandemic evolved over time. Future research can expand on these findings by incorporating vaccination rates and policy measures to gain more actionable insights.