A COMPREHENSIVE DATA EXPLORATION OF GLOBAL COVID-19 DYNAMICS AND VACCINATION IMPACT

Reza Mosavi ,400222100

Abstract

This report focuses on the Our World in Data COVID-19 dataset, a comprehensive and regularly updated resource crucial for understanding the global pandemic. The dataset covers vital metrics, including vaccinations, tests, hospital data, confirmed cases, and more, from reputable sources and is updated daily. It spans 218 countries and includes a data dictionary for clarity.Acknowledging the diligent efforts of the Our World in Data team, the dataset is made available to the Kaggle community. Its applications range from forecasting daily new cases to data analysis and visualization. This report serves as an invitation for the data science community to leverage this invaluable dataset, contributing to the collective understanding and response to the ongoing pandemic.

Introduction

In the exploration of this dataset, which pertains to the global impact of the Covid-19 disease in recent years, we delve into a dataset comprising approximately 350,000 entries and 67 variables. A significant portion of the challenges revolves around addressing missing values, a task made complex by the temporal nature of the data as it unfolds as a time series. Determining the appropriate strategy for handling missing values, whether through filling or deletion, is a critical decision. The method chosen significantly influences the conceptual integrity of the data and the information it encapsulates. Subsequently, we will conduct a comprehensive analysis of the dataset and its characteristics, ultimately making informed decisions on how to manage missing data.

EXPLORATORY DATA ANALYSIS

Missing data

Upon inspecting the dataset, a notable observation emerges: approximately 64 out of the 67 data columns contain missing values. Our initial step involves a broad assessment, identifying around 15 columns where over 80 Percent of the data is absent. In addressing this, our approach is to eliminate these columns from the dataset. Given the time series nature of our data, traditional methods like interpolation or regression prove impractical for accurate value imputation. The challenges arise from both the substantial missing data—exceeding 80 Percent in these columns—and the limited representation of the original dataset. Conventional techniques fall short in providing reliable estimates for such extensive data gaps, prompting the consideration of neural networks. In doing so, we anticipate that the remaining 20 Percent of columns with available data will present a more representative distribution of the societal context we aim to capture in our analysis.Finally, we will delete these 15 columns.

For the remaining columns, a more nuanced approach is required, taking into account the specific number of missing values in each. Despite encountering approximately 270,000 missing values across multiple columns, it's crucial to emphasize that the dataset under scrutiny follows a time series structure. Unlike scenarios where data could be simply removed or filled using general methods, the intricacies of a time series dataset necessitate careful consideration.

Given the project's primary objective of reviewing and analyzing the data, introducing imputed values poses a potential risk. Any addition might alter the fundamental understanding of the data in a singular direction. As elucidated earlier, the decision to leave missing values unaltered stems from the acknowledgment that attempting to modify them could introduce unintended biases or inaccuracies. This conservative approach aims to maintain the integrity of the dataset in its raw form, allowing for a transparent and unaltered examination of the inherent temporal patterns and trends within the time series data.

Data Analysis

In this section, our focus is on the examination of categorical features, specifically the continents and their constituent countries. The analysis unfolds in two stages: initially, a comprehensive overview of continents and countries, followed by an in-depth examination of the interplay between these columns and their mutual impact.The objective is to explore patterns, dependencies, and relationships within the dataset's categorical attributes. The examination of continents and countries provides a foundational understanding, and subsequent analysis delves into how these columns interact and influence each other. This multi-dimensional exploration aims to unveil insights into the intricate dynamics and relationships between categorical features, contributing to a holistic understanding of the dataset's structure and content.

First, we examine the number of cases in each continent.1

Certainly, your observation aligns with a common trend noted in various studies – densely populated regions often exhibit higher reported cases of infectious diseases, including COVID-19. In a preliminary analysis, it's evident that continents with higher population density tend to report more cases.

However, the relationship between reported cases and fatalities is nuanced. While one might expect a direct correlation, several factors contribute to the complexity of this association. These factors include healthcare infrastructure, access to medical resources, public health measures, and the effectiveness of governmental responses.

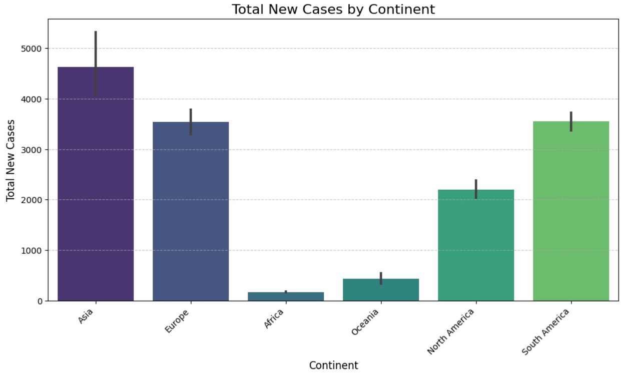

Figure 1: Number of cases per continent

To delve deeper into this relationship, a more thorough examination is needed. This involves analyzing the specific data points related to reported cases and deaths, considering additional variables that might influence the outcomes. Statistical methods, visualizations, and potentially machine learning techniques can be employed to gain a comprehensive understanding of the factors contributing to the observed patterns.By taking a data-driven approach, we can uncover insights that go beyond initial observations, providing a more nuanced perspective on the dynamics between reported cases and fatalities across densely populated continents.

Now let's look at the overall percentage of cases in the continents2

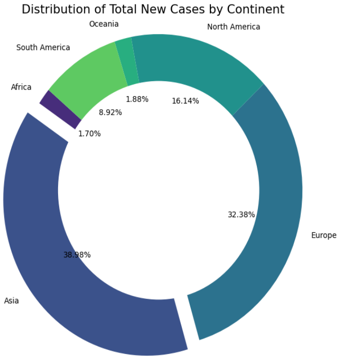

Figure 2: Percentage of cases based on each continent

It is clear that most people are from Asia, Europe and North America, which was expected. These results are based on the larger number of people in these natural areas, and the level of hygiene is also very effective.Now we examine the same graph based on the number of deaths according to the continents:3

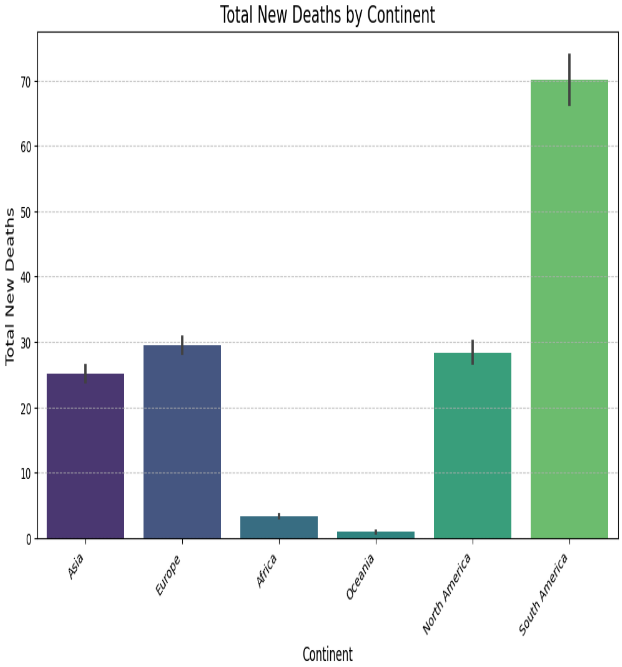

Figure 3: Number of deaths by continent

The analysis of the graphs indicates distinct patterns in the outcomes of COVID-19 cases across continents. In South America, a noteworthy observation is the relatively high proportion of cases that result in death. This suggests challenges in health conditions and the management of the disease within this continent.

Conversely, in Europe and Asia, the data reflects a more positive scenario. The number of deaths, when compared to the percentage of cases, appears to be lower, signifying better control of the disease. This observation may be indicative of more effective healthcare systems, robust public health measures, and a generally higher level of health infrastructure in these continents.These findings underscore the importance of not only considering the absolute numbers of cases and deaths but also analyzing the relative proportions. Such an approach provides a nuanced understanding of how different regions are grappling with the impact of COVID-19. It is essential for policymakers and healthcare professionals to take into account these variations to tailor effective strategies and interventions based on the specific challenges faced by each continent.Now let's examine the percentage of deaths in each continent.4

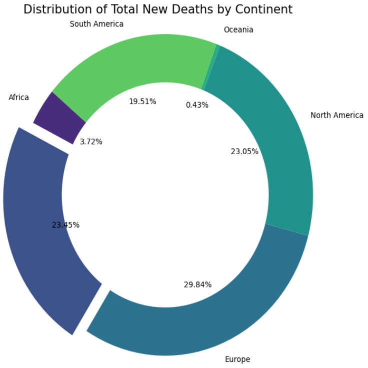

Figure 4: Percentage of deaths by continent

Your observation is insightful and highlights the importance of considering both the absolute numbers and the relative percentages when assessing the impact of COVID-19 across continents.In Europe, where the percentage of deaths may be slightly higher, the context of population density is crucial for interpretation. Higher population density can contribute to increased transmission, potentially leading to higher absolute numbers of cases and deaths, even if the proportion is relatively lower.Conversely, your point about Africa underscores that the lower number of reported cases and deaths could be influenced by factors such as lower population density and potentially less international travel, which could impede the rapid spread of the virus.

This nuanced analysis is essential for understanding how different regions are affected by and responding to the pandemic. It emphasizes the need for tailored public health strategies that consider the unique circumstances of each continent, taking into account factors such as population density, healthcare infrastructure, and cultural practices that can impact the spread and severity of the disease.Now let's examine the vaccinated people.5

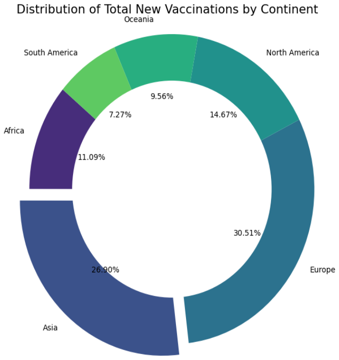

Figure 5: Percentage of vaccinated people in each continent

the analysis aligns with the global trend indicating that regions with higher vaccination rates tend to experience lower mortality rates from COVID-19. Vaccination plays a pivotal role in diminishing the severity of illness, preventing hospitalizations, and ultimately reducing the number of deaths linked to the virus.In Europe and Asia, where vaccination coverage has been notable, lower mortality rates may be attributed to successful vaccination campaigns. However, it's essential to recognize that the impact of vaccination is multifaceted, influenced by factors such as vaccine coverage, efficacy, population density, healthcare infrastructure, and public health measures.This underscores the complexity of assessing the relationship between vaccination rates and COVID-19 outcomes. The observation emphasizes the critical role of vaccination in mitigating the pandemic's impact and underscores the ongoing need for global efforts to enhance vaccine coverage.

We will examine the growth chart of disease distribution in different continents:6:

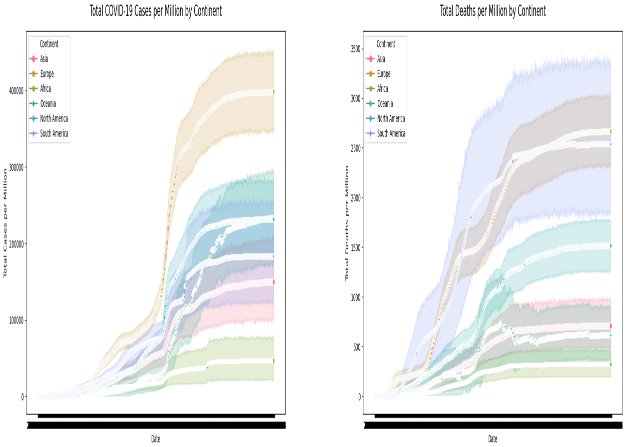

Figure 6: Growth chart of disease distribution in different continents

The analysis of the provided graph reveals a consistent prevalence of the disease across all continents over the three-year period. However, a notable deviation occurs in Europe, where there is a sharp increase in the number of cases in the middle of this time frame, likely attributed to higher population density and other contributing factors. This pattern is mirrored in continents that share similar conditions.The accompanying graph illustrates the severity of death, indicating that despite the surge in cases, Europe and comparable continents managed to control the disease, presumably through effective vaccination campaigns and robust global health measures. In contrast, South America stands out with a higher death rate, suggesting potential challenges in vaccination efforts and hygiene practices.This observation emphasizes the critical role of vaccination and global health initiatives in controlling the spread and severity of the disease. The disparities between continents underscore the need for targeted interventions and collaborative efforts to address specific challenges faced by different regions in combating the ongoing pandemic.

Now we will examine various things such as gdp and... in each continent(figure7)

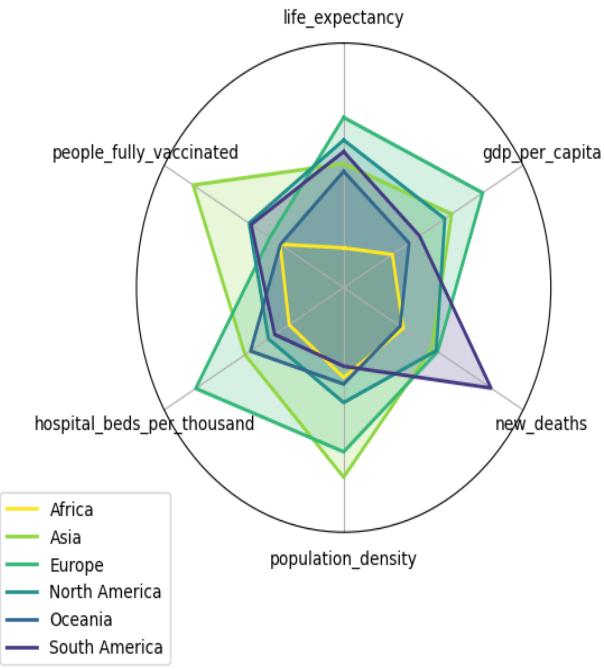

Figure 7: Examining mortality conditions based on the condition of each continent

The analysis based on the provided graph suggests that South America experiences a higher death rate, likely attributable to challenges in hospital conditions, lower GDP, and life expectancy. Despite having favorable vaccination conditions, the impact of these other factors seems to contribute to a less effective overall control of mortality in the region.In contrast, Europe appears to have successfully controlled mortality, attributed to proper health conditions and medical services. The combination of robust healthcare infrastructure, higher GDP, and presumably better life expectancy contributes to more favorable outcomes.

Similarly, Asia's ability to control mortality is linked to effective vaccination conditions. The region's success in managing the impact of the disease may be a result of widespread vaccination efforts.The impact of GDP on life expectancy and the number of hospital beds per 1000 people emerges as a crucial factor influencing mortality rates in each continent. These socio-economic indicators, along with vaccination conditions, play a pivotal role in shaping the outcomes of the ongoing pandemic. The observed patterns underscore the importance of addressing both healthcare infrastructure and socio-economic factors in formulating effective strategies for managing and mitigating the impact of global health crises.

Now let's examine the distribution of deaths in each continent(figure8)

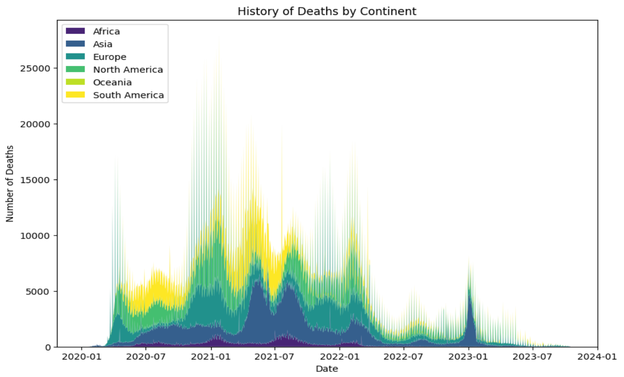

Figure 8: Distribution of deaths for each continent

The data indicates a notable spike in the death rate in 2021, particularly in the American continent. This observation aligns with the earlier analysis, emphasizing the higher death rate in South America. The American continent, as a whole, has consistently demonstrated a higher mortality rate compared to other regions.This trend underscores the persistent challenges faced by the American continent in managing and mitigating the impact of the ongoing pandemic. Factors such as healthcare infrastructure, socio-economic conditions, and vaccination efforts likely contribute to the observed patterns.Continued monitoring and targeted interventions are crucial in addressing the specific challenges faced by the American continent in the battle against COVID-19. This observation reinforces the importance of region-specific strategies and a nuanced understanding of the diverse factors influencing the trajectory of the pandemic across continents.

We will examine the distribution of vaccination rates(figure9)

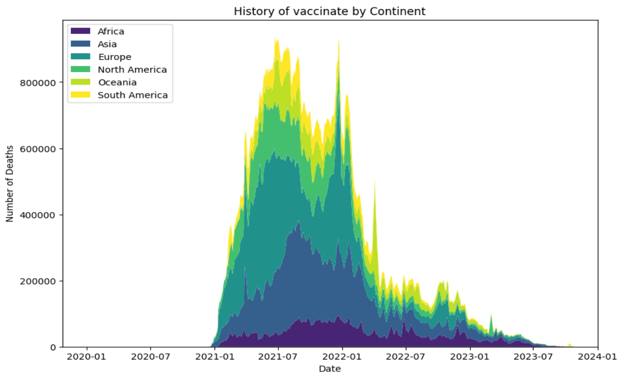

Figure 9: distribution of vaccination rates

The graph illustrates a clear correlation between the peak of the death rate in 2021 and the simultaneous peak of vaccination. This suggests a coordinated effort to increase vaccination rates during a period of heightened mortality. It is a common strategy to ramp up vaccination campaigns during surges in cases to mitigate the impact of the virus.Your observation implies that the substantial increase in vaccination during this period may lead to a significant reduction in deaths in the subsequent years, specifically in 2022 and

2023. This anticipation aligns with the expected outcomes of widespread vaccination, as immunization efforts aim to reduce the severity of illness, hospitalizations, and ultimately, fatalities.Continued vigilance and efforts to maintain high vaccination coverage are crucial for sustaining these positive trends and further controlling the impact of the ongoing pandemic. The observed patterns underscore the effectiveness of vaccination campaigns as a key tool in managing and mitigating the consequences of the COVID-19 virus. Finally, based on the graph opposite, I will discuss the total number of deaths in the world.(figure10)

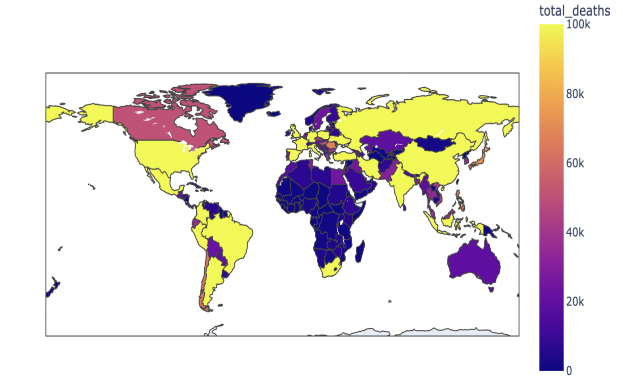

Figure 10: The total number of deaths in the world

In general, based on all the previous analysis, all the reviews can be seen in the chart below.

The chart below shows the countries with the highest number of deaths in each continent:(figure11)

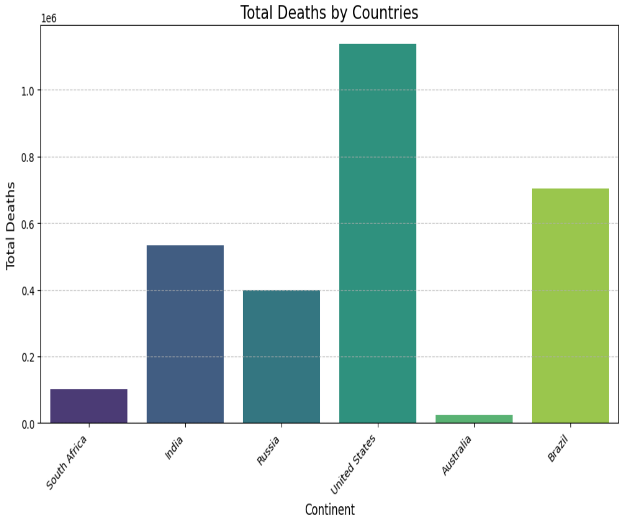

Figure 11: Highest death rate on any continent

Your observation highlights the significant disparities in COVID-19 death rates between countries, with the United States having the highest rate and India ranking second. Several factors contribute to these variations.In the case of the United States, the higher death rate could be influenced by factors such as population density, healthcare infrastructure, socio-economic conditions, and the initial impact of the virus. Regional variations, public health measures, and the effectiveness of vaccination campaigns are also critical considerations.India, as you pointed out, has a large population, and its comparatively lower level of health infrastructure might contribute to the challenges in managing the impact of the virus. Factors such as population density, access to healthcare, public health measures, and the pace of vaccination efforts all play a role in shaping the outcomes.These observations underscore the importance of considering a range of factors when interpreting COVID-19 death rates between countries. The complexities of each nation's healthcare system, population dynamics, and response strategies contribute to the observed variations in the impact of the pandemic.

Now we will examine the trend of the spread of this disease in these countries plus Iran:(figure12)

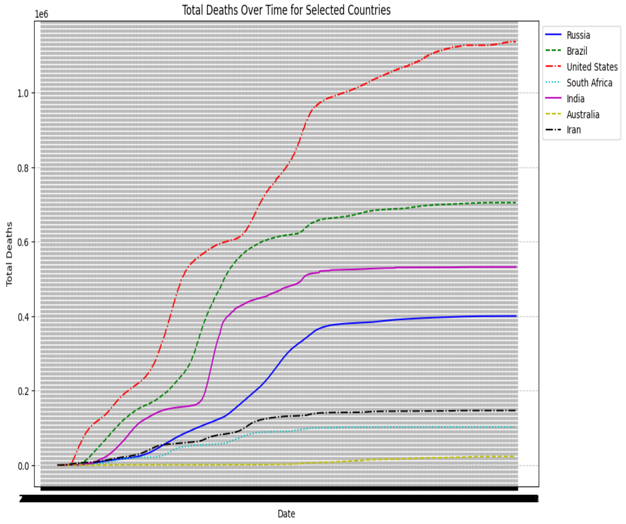

Figure 12: The spread of disease in Iran and the deadliest countries in every continent

Your observation indicates a noteworthy trend in the death rate over the three-year period. While many countries experienced a relatively stable death rate during the middle of this timeframe, the United States stands out with a continued increase in its death rate.Several factors could contribute to this trend in the United States, including the evolving nature of the pandemic, the effectiveness of public health measures, vaccination campaigns, and the emergence of new variants. Regional variations, healthcare capacity, and socio-economic conditions may also influence the trajectory of the death rate.Understanding the reasons behind the ongoing increase in the death rate in the United States requires a more detailed analysis of specific factors influencing the dynamics of the pandemic within the country. This ongoing observation underscores the importance of dynamic and adaptive public health strategies to address the evolving challenges posed by COVID-19.

Examining various factors such as gdp and health conditions based on mortality:(figure13)

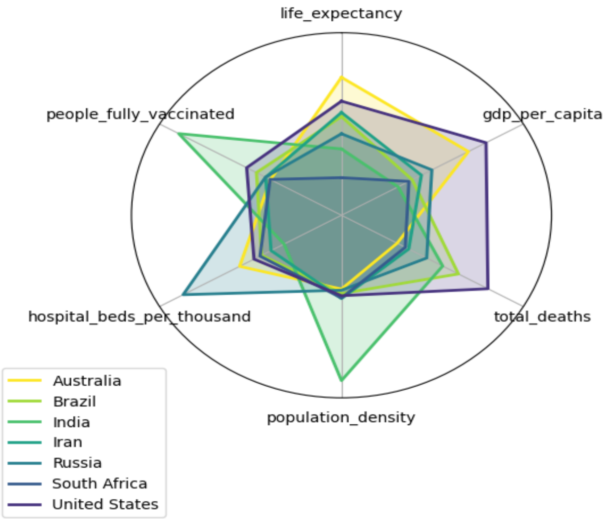

Figure 13: Various factors such as gdp and health conditions based on mortality

the analysis underscores the intricate and multifaceted nature of factors shaping COVID-19 outcomes across countries. Key observations include Russia's favorable health conditions, lower medical conditions in India and Iran, India's high population density contributing to rapid disease spread, the United States having the highest GDP, Australia boasting the highest life expectancy, and paradoxically, the United States also experiencing the highest death rate. This highlights the complexity of pandemic dynamics, emphasizing the importance of considering a range of factors, including healthcare effectiveness, public health measures, and societal responses, for a nuanced and context-specific approach to addressing the challenges posed by COVID-19 globally.Finally, we will examine the death rate in Iran:(figure14)

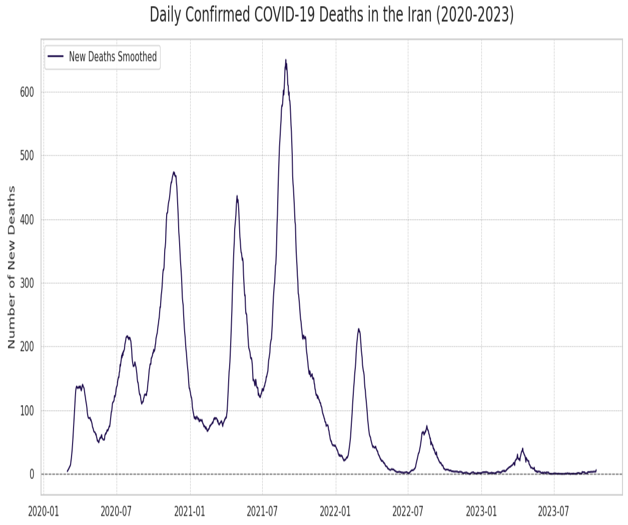

Figure 14: Death rate in Iran

Your observation points to a notable correlation between the death rate in Iran and the distribution of the world during the years 2020 to 2022. Specifically, Iran experienced its highest death rate during this period, which coincided with a significant increase in the amount of vaccination. The data suggests that after 2022, there has been a substantial decrease in the death rate.This correlation aligns with the expected outcomes of widespread vaccination efforts. As vaccination coverage increases, it is anticipated to contribute to a reduction in the severity of illness, hospitalizations, and ultimately, the death rate.Your observation underscores the potential positive impact of vaccination campaigns in mitigating the impact of COVID-19 in Iran, aligning with the global trend of vaccination being a crucial tool in controlling the spread and severity of the virus. It also emphasizes the importance of continued efforts to enhance and sustain vaccination coverage to further alleviate the impact of the ongoing pandemic.

Conclusion

observations and analyses highlight the intricate dynamics of COVID-19 outcomes across different continents and countries. Here's a brief conclusion based on the key points discussed:

1. **Global Disparities:** There are significant disparities in COVID-19 outcomes globally, influenced by factors such as healthcare infrastructure, socio-economic conditions, population density, and vaccination efforts.

2. **Vaccination Impact:** Widespread vaccination appears to be correlated with reduced death rates, emphasizing its crucial role in mitigating the severity of the virus.

3. **Regional Variances:** Different continents and countries exhibit diverse responses to the pandemic, influenced by a complex interplay of health conditions, socio-economic factors, and public health measures.

4. **Complex Factors in Death Rates:** While economic indicators like GDP are important, the death rate is also influenced by healthcare effectiveness, societal responses, and other contextual factors.

5. **Temporal Trends:** The temporal analysis reveals patterns such as the impact of vaccination on reducing death rates, indicating the importance of dynamic strategies in responding to evolving pandemic challenges.

In conclusion, a nuanced understanding of the diverse factors at play is crucial for effective strategies to combat COVID-19. Vaccination, healthcare infrastructure, and socio-economic conditions are interconnected elements that require tailored and context-specific approaches to address the unique challenges faced by each region and country.

Reference

1 : scikit-learn.org
2 : www.kaggle.com