Abstract

To learn more about how the virus has impacted the globe, we are delving into the COVID-19 data collected over the previous three years. Consider it as if you were looking at a large puzzle, in which every nation is a piece that, when combined, forms the overall picture of what transpired during the pandemic. To gain a better understanding of the events that transpired, we are examining the trends and significant elements found in the data. We are learning more about the COVID-19 story, from individual nations to the entire globe.

Introduction

The COVID-19 pandemic has presented the world with an unprecedented challenge over the last three years. This global health crisis has had a lasting effect on our collective consciousness, affecting individuals, communities, and nations to a great extent. Given the severity of the pandemic, a thorough analysis of its effects is imperative, not only as a retrospective exercise but also as a vital undertaking to extract lessons for an uncertain future.

The SARS-CoV-2 virus, which gave rise to COVID-19, quickly crossed national boundaries and cultural barriers. Its effects on society and on individuals have been enormous, changing the way we live our everyday lives. A thorough comprehension is necessary to address the challenges presented by the pandemic, which range from economic disruptions to fatalities.

We do an analytical journey through the large dataset that captures the COVID-19 landscape over the last three years in this quest for understanding. Our goal is to both decipher the complex narrative that the data tells and to extract lessons that can guide our future responses to unanticipated obstacles.

Tasks at Hand: EDA and Data Visualization

We use two essential tools, Data Visualisation and Exploratory Data Analysis (EDA), to help us navigate this enormous sea of information. We can find hidden trends, patterns, and anomalies in the data thanks to EDA. Comparable to removing layers from a complicated story, it reveals important details that might be missed at first glance.

On the other hand, data visualisation is our creative tool for painting the information canvas. Our goal is to make the data not only understandable but also visually appealing by turning numbers into images. We make the story hidden in the data visible to a wide audience by bringing it to life with the help of maps, graphs, and charts.

We will examine the cleaned and curated dataset in more detail in the following sections in order to draw important conclusions that advance our knowledge of the COVID-19 pandemic. Our investigation is a purposeful attempt to arm ourselves with knowledge for the potential challenges of the future, not just a nostalgic look back.

1 Data Preprocessing

1.1 Basic EDA

1. **Initial Data Inspection:** We start our analysis by taking a quick look at the dataset. We can get a basic idea of the organisation and content of the data by looking over the first fifteen rows.

2. **Dataset Dimensions:** It is essential to comprehend the size of our dataset. This dataset provides a comprehensive view of COVID-19 data over the last three years, with 355,500 entries spread across 67 columns.

3. **Summary Statistics:** Basic summary statistics are utilised to extract quantitative insights. This high-level summary gives a basic idea of the distribution of the data and includes counts, means, and standard deviations for numerical columns.

4. **Column Information:** Examining the types and non-null counts of each column helps identify potential data types and assess data completeness. This provides valuable insights into the nature of the variables at our disposal.

5. **Unique Values:** Diversity within the dataset is explored by examining the number of unique values in each column. This not only aids in identifying categorical variables but also offers insights into the granularity of certain features.

6. **Missing Values:** Finding missing data is essential for the rest of the cleaning process. Our approach to data cleaning is guided by an overview of missing values in each column, which guarantees a solid basis for additional analyses.

This thorough investigation establishes the groundwork for more focused data cleaning and later exploratory studies. We'll talk about data cleaning and outlier removal in the ensuing subsections.

1.2 Data Cleaning and Data Engneering

Several techniques were used in the data cleaning and engineering process to deal with missing values and improve the overall robustness of the dataset. Below is a thorough description of every technique:

1. **Handling NaN Values in Numeric Columns**
   Objective: To address missing values in numeric columns related to COVID-19 cases, deaths, and related metrics.

   Approach: NaN values in columns such as 'total cases' and 'total deaths' were presumed to indicate periods without recorded data, due to the different starting times for COVID-19 data recording across countries. Consequently, 0 was systematically substituted for these NaN values, signifying the lack of documented cases or fatalities within those particular time periods.

2. **Handling NaN Values in Categorical Column**
   Objective: To manage missing values in the categorical column 'tests units,' representing the units used for COVID-19 testing.

   Approach: Given that the 'tests units' column was categorical, "no test" was entered for NaN values. This simple method distinguishes situations with recorded testing information from those in which testing data was not recorded.

3. **Flagging Missing Values in Specific Columns**
   Objective: To identify missing values in demographic and environmental features and distinguish them as instances where specific information was not recorded.

   Approach: Missing values were interpreted as situations in which the corresponding feature was not recorded for a specific region for columns such as ”stringency index” and ”population density.” For increased transparency, binary indicator columns (such as ”stringency index missing”) were added to indicate whether a value was missing (1) or not (0).

4. **Identifying Countries with Missing Values**
   Objective: To pinpoint countries or territories lacking recorded information for specific demographic or environmental features.

   Approach: The countries with missing values were identified and printed after missing values were flagged in the designated columns. Understanding the geographic scope of missing data is aided by this step, which offers insights into areas that failed to report certain environmental or demographic metrics.

   All of these painstaking techniques add up to a more refined dataset, guaranteeing that missing values are handled appropriately and promoting a better understanding of the COVID-19 data in various geographic areas.

1.3 Data Splitting

A refinement was carried out in the next stage of data processing in order to improve the focus on information specific to a given country. Identification and removal of non-country entities from the main dataset was done for things like ’World,’ ’Continent’ names, and economic classifications. This step was taken in an effort to simplify the analysis and preserve a more distinct hierarchy between individual nations and higher-level entities such as continents and income groups.

The process involved the following key steps:

1. **Identification of Values to Remove:**

   - A list of values, including ’World,’ ’Asia,’ ’Europe,’ ’North America,’ ’South America,’ ’European Union,’ ’High income,’ ’Lower middle income,’ ’Low income,’ and ’Upper middle income,’ was compiled for removal.

2. **Filtering Rows:**

   - Rows in the main dataset containing the specified values in the ’location’ column were filtered out. This resulted in a new dataframe called filtered data that exclusively contained rows related to non-country entities.

3. **Creation of Separate Dataframe:**

   - A separate dataframe named ”separate dataframe” was created by copying the filtered data. This dataframe specifically focuses on the rows associated with non-country entities.

4. **Exclusion from Main Data:**

- Rows with the identified values were removed from the main dataset (data). The negation of the filter condition (data['location'].isin(values to remove)) ensured that only rows not matching these values were retained in the main dataset.

5. **Resulting Dataset Shapes:**

   - The outcome of this split revealed two distinct datasets:
     - separate dataframe with a shape of (14102, 82), encompassing information related to non-country entities.
     - data with a shape of (341398, 82), containing the refined dataset focusing exclusively on individual countries.

This separation facilitates a more granular examination of country-specific data while maintaining a comprehensive dataset for broader analyses.

2 EDA

2.1 Analyzing Missing Values: Unveiling Data Recording Patterns

During the data exploration process, important environmental and demographic features' missing values were carefully examined. Notably, fascinating trends surfaced that provided insight into various nations' recording customs. Three unique situations were found to exist:

1. **No Record from the Beginning (1407 Missing Values):** Countries with 1407 missing values for a given column indicate that the corresponding feature has not been recorded for that country since the beginning of data collection. This implies that some areas may have decided not to record this specific information, which offers important background for appreciating the dataset's limitations.

   Example:

   - Africa
   - American Samoa
   - ... (Other countries with 1407 missing values)

2. **Cessation or Delayed Start (313, 691, etc. Missing Values):** In another scenario, countries showed the number of missing values, 313 in this case. This pattern indicates that data recording either started later in the timeline or stopped for a certain amount of time. The precise number may indicate how long there was no recording or how long it took to start gathering data.

   Example:

   - Afghanistan
   - Albania
   - ... (Other countries with varying missing values)

These patterns offer important new perspectives on the temporal dimensions of data collection procedures. Comprehending the causes of missing values improves the dataset's interpretability and facilitates a more nuanced understanding of the recorded data for various nations.

2.2 Extreme Cases Analysis: Unveiling Maximums

1. **highest total reported cases:** We carried out an analysis to determine which nation had the greatest overall number of reported cases in order to obtain a deeper understanding of the scope of the COVID-19 impact. We identified the highest value using the 'total cases' column and took the relevant information out of the dataset.

   With an astonishing total of 103,436,829 cases as of May 14, 2023, the United States stands out as the nation with the highest number of reported cases. The cumulative count since the pandemic's start is shown in this figure. The extra data, which shows 93,260 new cases reported on the designated date, reflects the ongoing fight against the virus. A further indication of the seriousness of the situation is the 1,127,152 deaths in the US that have been linked to COVID-19.

   It's also important to recognize the pandemic's worldwide reach. When all cases are taken into account from all available data, the total number of reported cases globally comes to an astounding 771,820,173 cases. Together, these numbers show the enormous toll the pandemic has had on the world, highlighting the difficulties encountered and the resiliency needed to deal with a health crisis of this magnitude.

2. **highest total reported deaths:** Turning our attention to the significant impact on human lives, we carried out an analysis to determine which nation had the greatest number of COVID-19-related reported deaths overall. As with the analysis of total cases, we identified the highest value and extracted the relevant information from the dataset by using the 'total deaths' column.

   With a sad count of 1,138,309 deaths as of October 10, 2023, the United States comes out on top in terms of total reported deaths. This graph captures the total amount of loss experienced since the pandemic began. In the US, no new cases or fatalities were reported on this particular date.

   Expanding our view to a global level, the total number of deaths that have been reported globally when all recorded data is taken into account comes to a sobering 6,978,162 deaths. These numbers highlight the COVID-19 pandemic's significant human cost and stress how crucial it is to comprehend and lessen its effects.

3. **Highest death count in a day**: Examining the 'new deaths' column, we focused on the daily death of the pandemic and determined which country and date had the greatest number of newly reported deaths. In this category, Chile leads with an astounding 11,447 new deaths on March 22, 2022. This noteworthy increase in new fatalities is indicative of an important and consequential event that occurred as the pandemic in Chile spread.

   The highest number of new deaths ever reported globally, as of January 24, 2021, is 27,939 per day. This peak signifies a turning point in the global response to the pandemic and sheds light on the difficulties that nations are facing in controlling and lessening the effects of COVID-19.

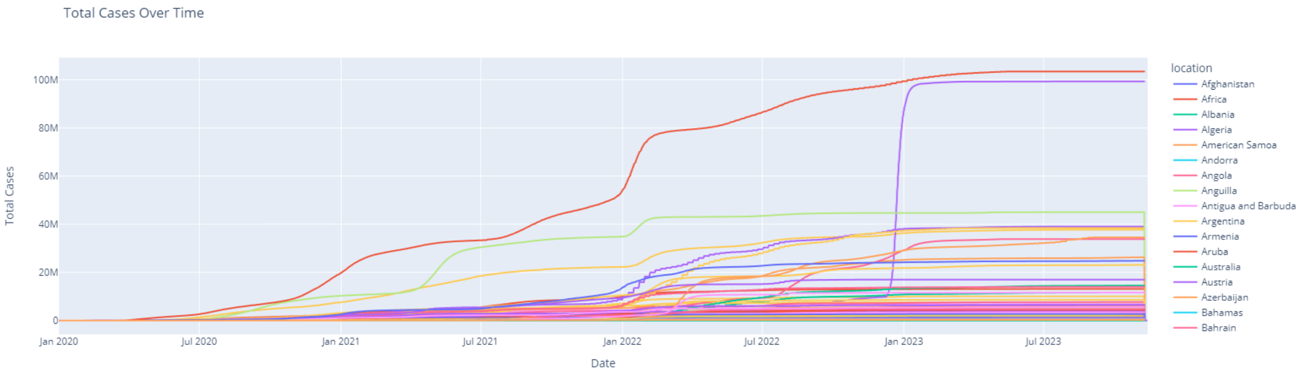
3 Visualization

3.1 Total Cases Over Time

In this line plot depicting the total COVID-19 cases over time, each country is represented by a distinct color. Here are some key observations:

- **United States (Orange Line on the top):**
  - The United States exhibits the highest total number of cases.
  - Noticeable peaks at different times suggest waves or surges in infection rates.
- **China (Purple Line):**
  - China shows a massive and sudden rise in total cases, from around 9 million on December 6, 2022, to over 98 million on January 22, 2023.
  - This drastic increase raises suspicions, possibly due to a sudden release of comprehensive medical information or a change in reporting practices.
- **India and Brazil:**
  - India and Brazil follow with substantial case numbers.
  - India's high case count may be attributed to its large population and the time it took to vaccinate a significant portion of the population.
  - Brazil's case trajectory is influenced by a slower vaccination effort and potential government response factors.

The distinct patterns among countries highlight the varying impacts of the pandemic, influenced by factors such as population size, vaccination strategies, and government interventions.

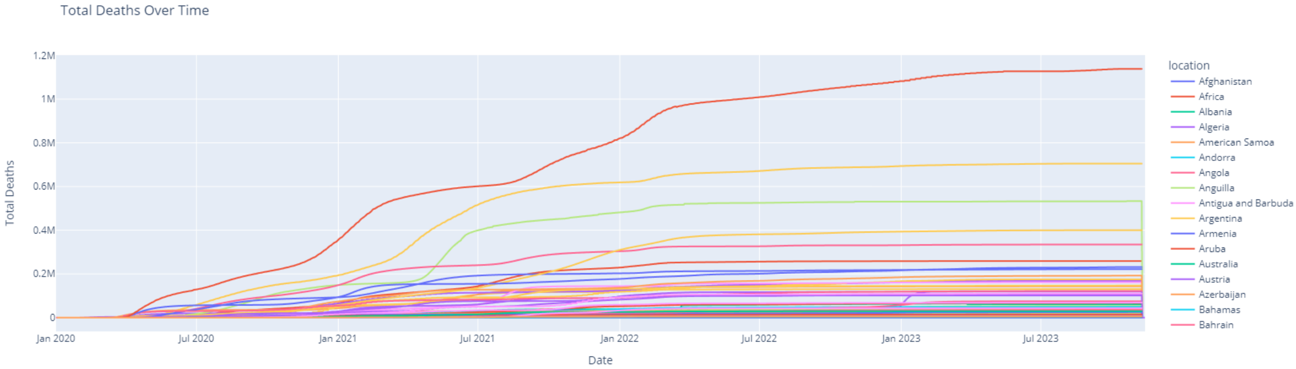


]

3.2 Total Deaths Over Time

In this line plot illustrating the total number of COVID-19 deaths over time, each country is represented by a unique color. Here are the notable observations:

- **United States (Orange Line):**
  - The United States exhibits the highest total number of deaths, surpassing one million.
  - Noticeable peaks in the line graph indicate periods of increased mortality, reflecting critical phases of the pandemic.
- **Brazil (Second Line):**
  - Brazil follows with a total death count exceeding 700,000.
  - The plot shows fewer distinct peaks compared to the United States, suggesting a different pattern of mortality events.
- **India (Third Line):**
  - India, represented by the third line, has experienced one major peak and two minor peaks in total deaths.
  - The total death count for India is around 500,000.
- **Russia (Fourth Line):**
  - Russia is represented by the fourth line, with a total death count reaching approximately 400,000.



3.3 Total Vaccinations Over Time

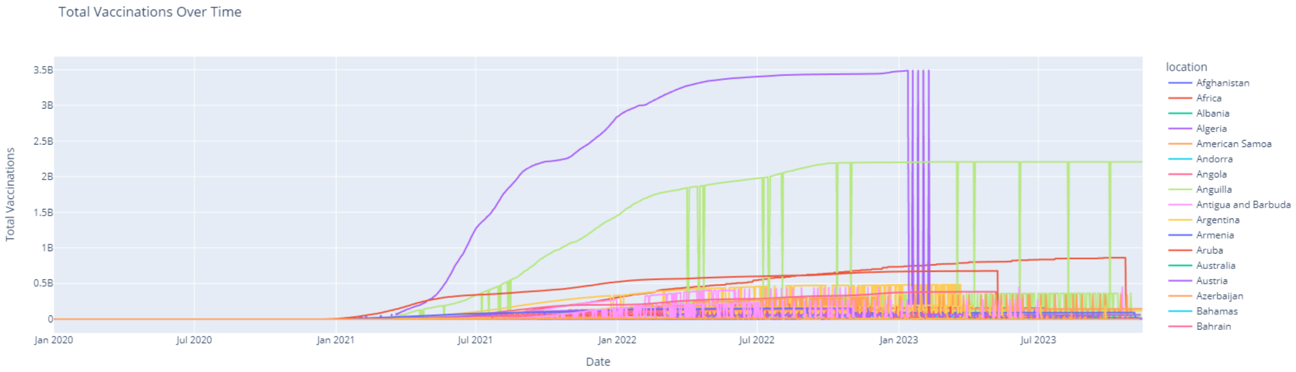
The total number of COVID-19 vaccinations given over time for various nations is represented graphically by this line plot. A particular nation is represented by each line, and noteworthy findings include:

- **China (Purple Line):**
  - China leads in total vaccinations, surpassing 3 billion doses.
  - Vaccination activity in China appears to halt around February 9, 2023, suggesting a potential completion or pause in their vaccination campaign.
- **India (Second Line):**
  - India, represented by the second line, has administered over 2.2 billion vaccinations as of its last recorded date on November 12, 2023.
  - The plot indicates ongoing vaccination efforts in India.

Africa (Third Line):

  - Africa shows a total vaccination count of slightly less than a billion.
  - Similar to China, there are instances of short-term gaps in vaccination, possibly influenced by external factors such as political or sociological considerations.

By keeping an eye on these trends, one can gain understanding of how vaccination campaigns are doing internationally. Different countries' approaches to managing the pandemic through vaccination may be influenced by logistical challenges, policy choices, or other contextual factors. These factors may also account for variations in the length of time and pauses in vaccination efforts.



]

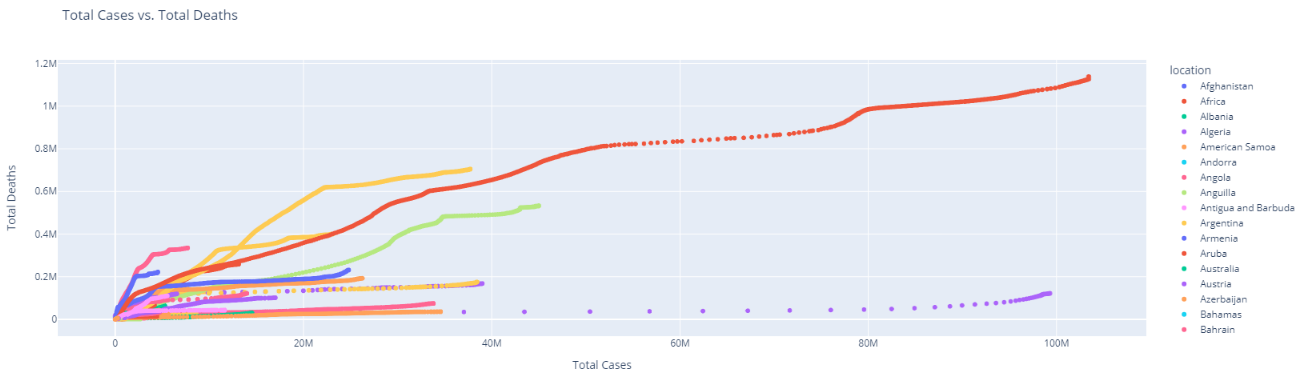
3.4 Total Cases vs. Total Deaths

This scatter plot illustrates the relationship between the total number of COVID-19 cases and the total number of deaths for various countries. Key observations include:

- **United States (Big Orange Dot):**
  - The United States stands out with a large orange dot, indicating a high total number of cases and deaths.
- **Brazil (Top Yellow Dot):**
  - Brazil is represented by the top yellow dot, showcasing a significant number of cases and deaths.

  Russia (Second Yellow Dot on the Top):
  - Russia follows Brazil with a substantial count of cases and deaths, appearing as the second yellow dot from the top.

  Mexico and Peru (Pink and Yellow Dots on the Left):
  - Mexico and Peru are notable on the left side of the plot, suggesting a comparatively high number of cases and deaths.



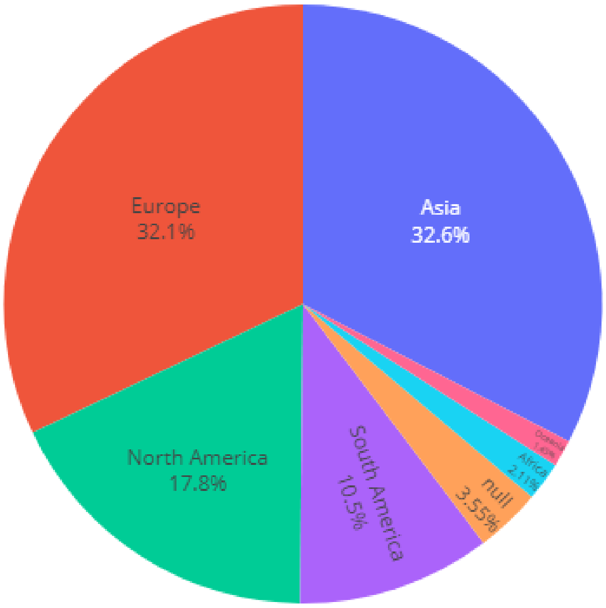
]

3.5 Continent-wise Total COVID-19 Cases

This pie chart provides an overview of the distribution of total COVID-19 cases across continents. Key observations include:

- **Europe and Asia:**
  - Europe and Asia stand out as the leading continents, each contributing approximately %32 of the total cases. These regions have experienced a substantial burden of COVID-19 infections.
- **North America and South America:**
  - Following closely, North America and South America together account for a significant portion of the total cases. The proximity of their percentages indicates a comparable impact of the pandemic on these continents.

The pie chart provides a visual depiction of the global distribution of COVID-19 cases across the continents. It draws attention to how different the effects are in different parts of the world, with Europe and Asia being most affected.

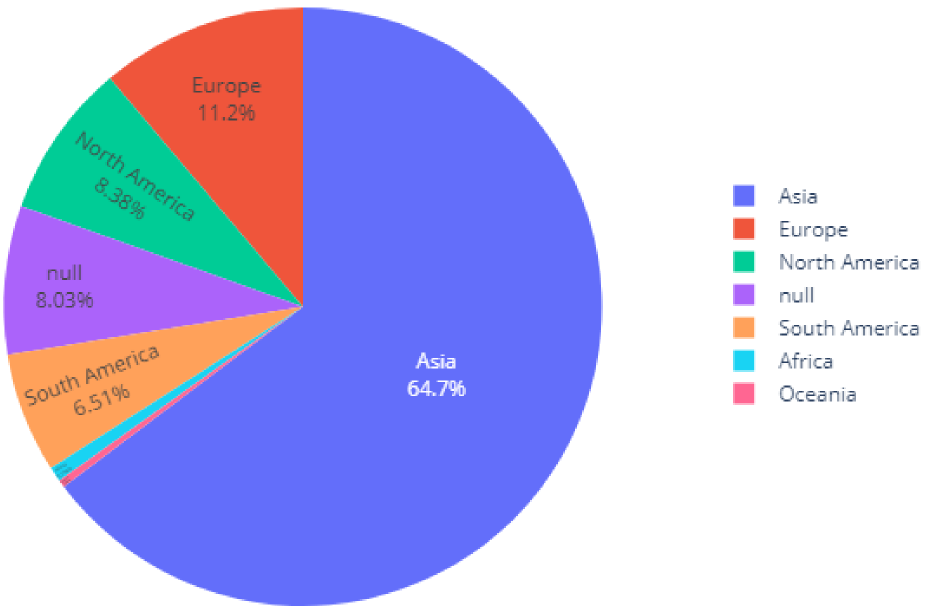


3.6 Continent-wise Total COVID-19 Vaccinations

This pie chart illustrates the distribution of total COVID-19 vaccinations across continents. Key observations include:

- **Asia**
  - Dominating the chart, Asia accounts for over %64 of total COVID-19 vaccinations. This indicates a substantial effort in the region to vaccinate its population against the virus.
- **Europe:**
  - Following at a considerable distance, Europe contributes to about %11.2 of the total vaccinations. While not as prominent as Asia, Europe's share signifies a noteworthy vaccination campaign.

The pie chart highlights the important role that Asia has played in the vaccination campaign by visualising the global distribution of COVID-19 vaccinations.



3.7 World-wise Total COVID-19 Cases over Time: Peaks and Patterns

The global scatter plot that shows the total number of COVID-19 cases over time offers important insights into the dynamics of the pandemic. Here are a few interesting findings:

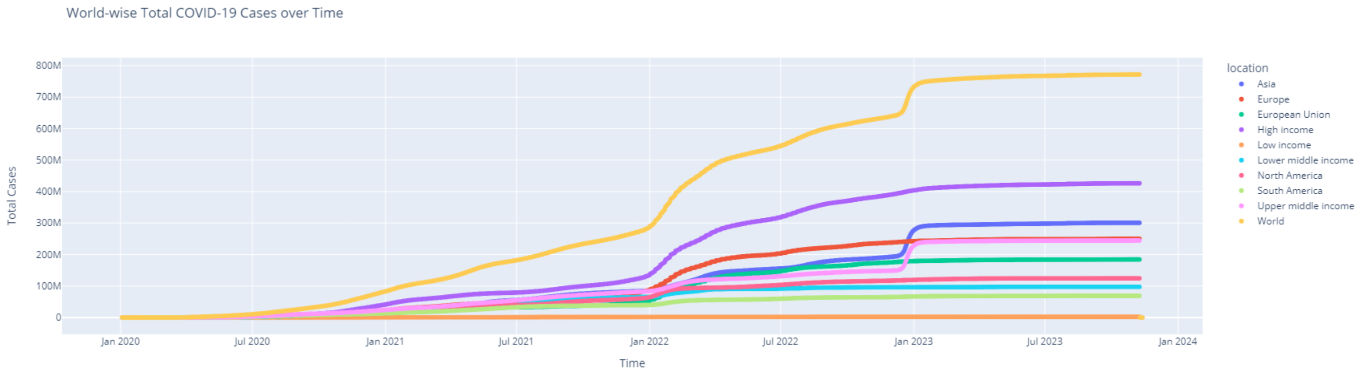
1. **Major Peaks in Jan 2022 and Jan 2023:** The plot shows two distinct peaks in the number of COVID-19 cases worldwide: one in January 2022 and the other in January 2023. These peaks point to crucial intervals of elevated reporting or transmission, highlighting the importance of these periods in the pandemic's global trajectory.

2. **Absence of Peaks in Low-Income Areas:** Interestingly, areas classified as ”Low income” have a noticeable lack of major peaks. This observation may point to a reduced impact of the virus or raise concerns about the consistency and dependability of data reporting in these areas. Additional information may be obtained by looking into low-income areas' healthcare systems and testing capacities.

3. **Synchronized Peaks for ”Upper middle income” Class and ”Asia”:** The ”upper middle income” class and the continent of ”Asia” exhibit synchronised peaks, which presents an intriguing pattern. Peak timing alignment could be a sign of coordinated regional responses or shared socioeconomic factors. Gaining insight into the relationship between pandemic dynamics and economic status can help develop more focused and efficient public health initiatives.

4. **Minor Peaks Across Continents:** Although the plot is dominated by major peaks, minor peaks can be found on different continents. These variations may be explained by regional outbreaks, differences in reporting and testing procedures, or particular sociodemographic characteristics that affect the virus's transmission.

To summarise, the examination of COVID-19 cases worldwide over time has revealed significant high points and prompted additional research on the differences in pandemic patterns between different income groups and regions. This complex knowledge is essential for developing focused interventions and improving global health plans.



3.8 World-wise Total COVID-19 Vaccinations over Time: Patterns and Disparities

The scatter plot that shows the total number of COVID-19 vaccinations worldwide provides important information about vaccination campaigns in various geographic areas and classes of society. The salient points are as follows:

1. **Commencement of Vaccination Around Jan 2021:** According to the plot, worldwide COVID-19 vaccination campaigns started in January 2021. With nations and regions starting extensive vaccination campaigns to stop the virus's spread, this is a critical turning point in the global response to the pandemic.

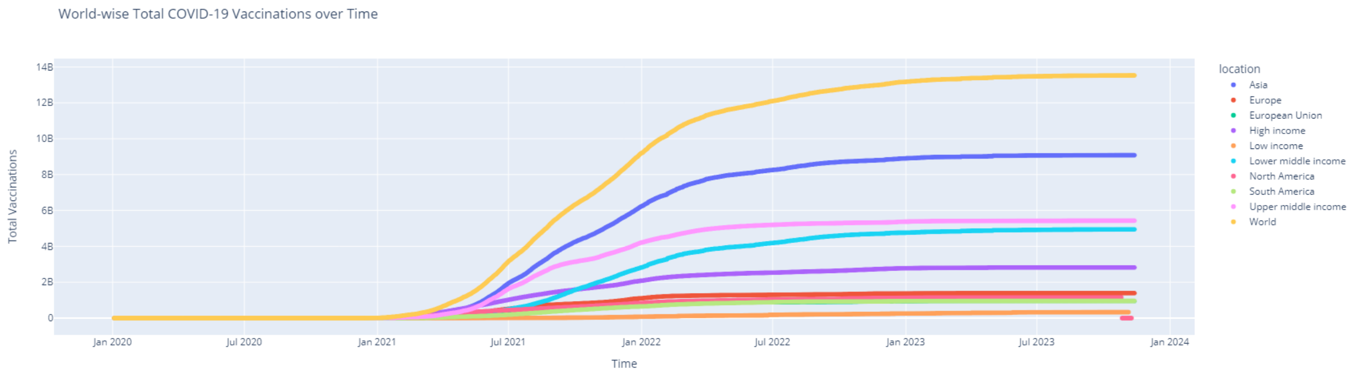
2. **Asia's Prominent Role in Vaccination:** Asia emerges as a frontrunner in the global vaccination drive, exhibiting substantial progress in total vaccinations over time. The continent's proactive approach to immunization is evident, showcasing its commitment to achieving widespread immunity.

3. **"Upper middle income" Following Asia's Lead:** Notably, in terms of total vaccinations, the "Upper middle income" class closely trails Asia. This alignment emphasises the significance of financial resources in carrying out large-scale vaccination programmes by pointing to a relationship between economic status and vaccination capabilities.

4. **"Lower middle income" as the Fourth Contributor:** The fourth largest contributor to vaccination rates worldwide is the "Lower middle income" category, which includes a sizable portion of the working class population. This finding emphasises how important this income bracket is to the vaccination landscape.

5. **Massive Disparities with Other Continents/Classes:** A striking feature of the plot is the substantial gap between the top four contributors (Asia, "Upper middle income," "Lower middle income") and the remaining continents or income classes. This gap emphasizes the concentrated efforts of a few regions and income categories in achieving mass vaccination coverage.

To sum up, the examination of COVID-19 vaccinations around the world over time reveals differences in vaccination coverage and reveals patterns of leadership. To ensure fair access to immunisation on a global scale and to optimise vaccine distribution strategies, it is imperative to comprehend these dynamics.



4 Conclusion

The thorough examination of the COVID-19 dataset over the previous three years has yielded significant insights into the complex effects of the pandemic on a worldwide level. Our investigation has uncovered patterns, trends, and notable phenomena that add to a nuanced understanding of this unprecedented health challenge, from the beginning of the crisis to the ongoing vaccination efforts.

Key Insights

Epidemiological Trends

Analysing all COVID-19 cases and fatalities over time revealed unique trends amongst nations. The country with the most recorded cases and fatalities, the United States, displayed peaks that were representative of the difficulties encountered at various stages of the pandemic. Russia, Brazil, and India—all of which had different histories—showed how different the virus's effects were depending on things like immunisation programmes, government reactions, and population density.

Vaccination Dynamics

The way that the total number of vaccinations across nations was visualised highlighted how important vaccination campaigns were in stopping the pandemic. In the global vaccination campaign, Asia came out on top, closely followed by the ”Upper middle income” class. The differences in vaccination rates brought attention to the focused efforts of particular geographic areas and socioeconomic groups, underscoring the necessity of distributing vaccines equitably.

Global and Regional Dynamics

The different ways in which the pandemic affected different regions were made clear by the examination of data by continent. The majority of COVID-19 cases worldwide were shared by Europe and Asia, with North America and South America following closely behind. The analysis of income levels revealed differences in immunisation capacities, and the path of the pandemic within each class was significantly influenced by economic standing.

Temporal Aspects and Peaks

The examination of temporal features, such as significant peaks in the overall number of cases and fatalities, produced a temporal story of the pandemic. Further investigation into the healthcare infrastructure and testing capabilities was prompted by the lack of significant peaks in low-income areas, which raised concerns about the reliability of the data. Peaks that coincided between Asia and the ”upper middle income” class suggested that socioeconomic factors common to both regions influenced pandemic dynamics.

Implications and Lessons Learned

We do more than just an in-depth review with this large dataset. It gives us insightful lessons and useful knowledge for upcoming difficulties. The necessity of consistent and open data collection procedures is highlighted by the discovery of recording trends and discrepancies in data reporting. The relationship between vaccination capacity and economic standing emphasises how crucial international cooperation is to guaranteeing universal vaccine access.