**NAAN MUDHALVAN-IBM DATA ANALYTICS WITH COGNOS**

**PROJECT PHASE 5: PROJECT DOCUMENTATION & SUBMISSION**

**PROJECT TITLE:**

***COMPREHENSIVE ANALYSIS OF COVID-19 VACCINATION DATA:***

Enhancing deployment strategies for optimal public health impact

**PROVIDED DATASET:**

**<https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>**

**TEAM MEMBERS:**

|  |  |  |
| --- | --- | --- |
| **NAME** | **ROLL.NO** | **E-Mail id** |
| Shanjanaa.G | 2021115099 | [*shanju.jaya@gmail.com*](mailto:shanju.jaya@gmail.com) |
| Shantha priya.M | 2021115100 | [*mshanthapriya@gmail.com*](mailto:mshanthapriya@gmail.com) |
| Sharmitha.J | 2021115101 | [*sharmithaj23@gmail.com*](mailto:sharmithaj23@gmail.com) |
| Ramyaa.R | 2021115318 | [*ramyarajaraman31@gmail.com*](mailto:ramyarajaraman31@gmail.com) |
| Saravanakumar.P | 2021115319 | [*sksaravana012@gmail.com*](mailto:sksaravana012@gmail.com) |

***ABSTRACT***

The primary objective of this project is undertake a thorough analysis of COVID-19 vaccine data, giving priority to evaluating vaccine efficacy, tracking distribution patterns, and assessing potential adverse effects. Through this research, we aim to furnish valuable and tailored insights that can guide policymakers and healthcare organizations in making informed decisions.

These insights will play a crucial role in refining and optimizing the strategies for vaccine deployment, ensuring the most efficient and targeted approach to combat the pandemic. By focusing on these key aspects, we aspire to contribute significantly to the global efforts to control and mitigate the impact of COVID-19.

***PROBLEM STATEMENT***

The challenge at hand is to conduct an extensive analysis of COVID-19 vaccine data, focusing on vaccine efficacy, distribution, and adverse effects, in order to provide focused insights to policymakers and healthcare organizations. These insights are crucial for enhancing the precision and effectiveness of vaccine deployment strategies.

***PROJECT DESIGN***

**1.** **Data Collection**:

The data (i.e., COVID-19 vaccine data) is collected form reputable institutions including health organizations, government databases, and peer-reviewed research publications. The quality, quantity and the correctness of the data collected plays an important role in the accuracy of the analysis that is to be conducted.

**2.** **Data Preprocessing**:

Once the data is collected, rigorous cleaning of the data using appropriate preprocessing protocols is performed.

This involves managing missing values and standardizing formats, ensuring data integrity and accuracy. Data preprocessing gives out a uniformly laid out numerical representation of the data which is used for training the model.

**3. Exploratory Data Analysis (EDA):**

Then for EDA, sophisticated techniques are employed to reveal underlying trends, patterns, and potential outliers in the pre-processed data. This helps us to understand the data’s characteristics which further gives an idea of how to utilize them.

**4.Statistical Analysis:**

It is now time to implement advanced statistical tests to assess vaccine efficacy, adverse effects, and distribution trends across different populations. The statistical analysis of the data collected forms an important factor in decision-making.

**5.Visualization:**

After the analysis, various diverse visualization techniques like bar plots, line charts, heatmaps, etc. are utilized to effectively represent and hence communicate key findings and insights. The visual representation needs to be clear and appealing in order for proper understanding.

**6.Insights and Recommendations:**

Based on the analysis on the data collected, the project will yield useful insights and recommendations which will serve as a strategic guide for the policymakers and healthcare organizations to optimize vaccine deployment strategies. By tailoring deployment approaches to specific demographics and regions, the goal is to maximize the impact of vaccination efforts.

***DEVELOPMENT***

**DATA COLLECTION:**

Collecting data for COVID-19 vaccine analysis is a vital aspect of understanding the effectiveness and impact of vaccination campaigns in order to perform proper analysis.

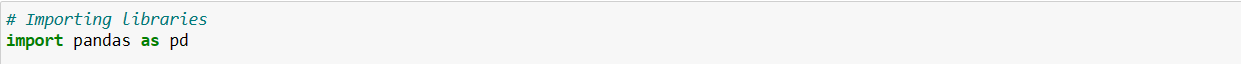
This process involves gathering a wide range of information related to vaccination efforts, including country-wise total vaccinations available, people vaccinated, adverse events, distribution logistics, and more. The source of the data for this project is,

[*https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress*](https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress)

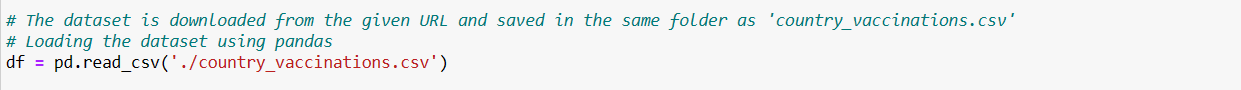
The dataset given provided information related to COVID-19 vaccinations in various countries across the globe. It had the following details for each of the countries,

* **country**: This column specifies the name of the country
* **iso\_code**: This column contains the ISO code for the country (For example, IND for India).
* **date**: This column represents the date for which the vaccination data is recorded. It is in YYYY-MM-DD format.
* **total\_vaccinations**: This column shows the total number of COVID-19 vaccine doses administered in the country on the given date.
* **people\_vaccinated**: This column indicates the total number of individuals who have received at least one vaccine dose.
* **people\_fully\_vaccinated**: This column represents the total number of individuals who have received the full recommended dose(s) of the vaccine and are considered fully vaccinated.
* **daily\_vaccinations\_raw**: This column contains data related to the daily change in total vaccinations.
* **daily\_vaccinations**: This column provides information about the daily number of vaccine doses administered on the given date.
* **total\_vaccinations\_per\_hundred**: This column calculates the total vaccinations as a percentage of the population.
* **people\_vaccinated\_per\_hundred**: This column calculates the number of people vaccinated as a percentage of the population.
* **people\_fully\_vaccinated\_per\_hundred**: This column calculates the number of fully vaccinated people as a percentage of the population.
* **daily\_vaccinations\_per\_million**: This column calculates the daily vaccinations per million people in the population.
* **vaccines**: This column specifies the types of COVID-19 vaccines used in the country. In this case, multiple vaccines are listed, including Johnson & Johnson, Oxford/AstraZeneca, Pfizer/BioNTech, and Sinopharm/Beijing.
* **source\_name**: This column mentions the source of the data, which is the World Health Organization in this dataset.
* **source\_website**: This column provides the source website or link from which the data was obtained. It links to the World Health Organization's COVID-19 data.

**DATA PRE-PROCESSING:**

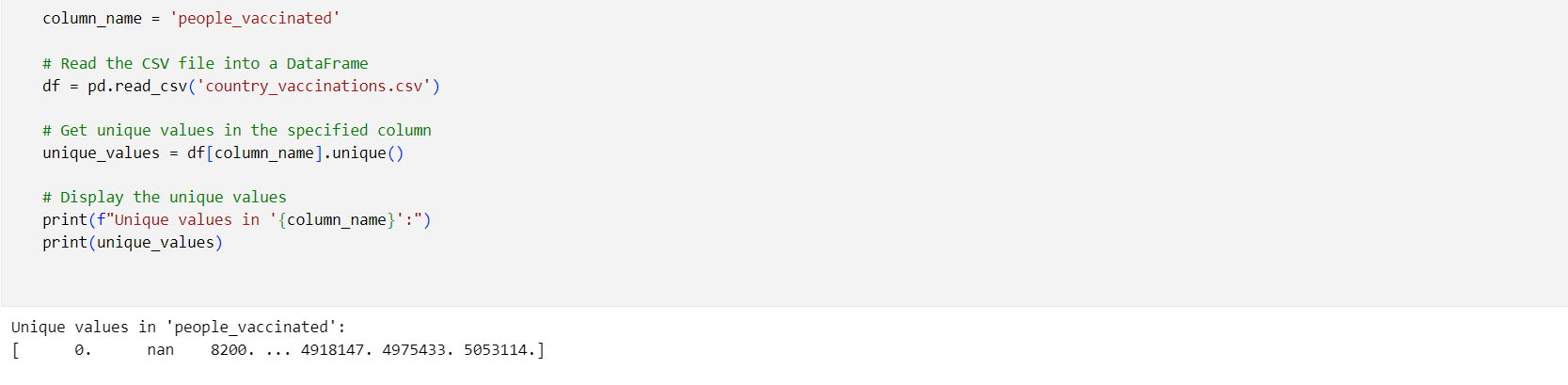
 Once reliable data has been collected, it is now time to clean and prepare the data for analysis. This process is coined ad Data Pre-processing. The same is explained below,

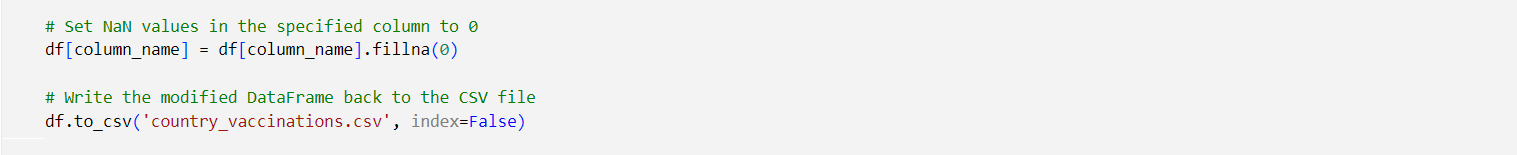
The code starts by importing necessary libraries. Here pandas library is imported. It is commonly used for data manipulation and analysis.

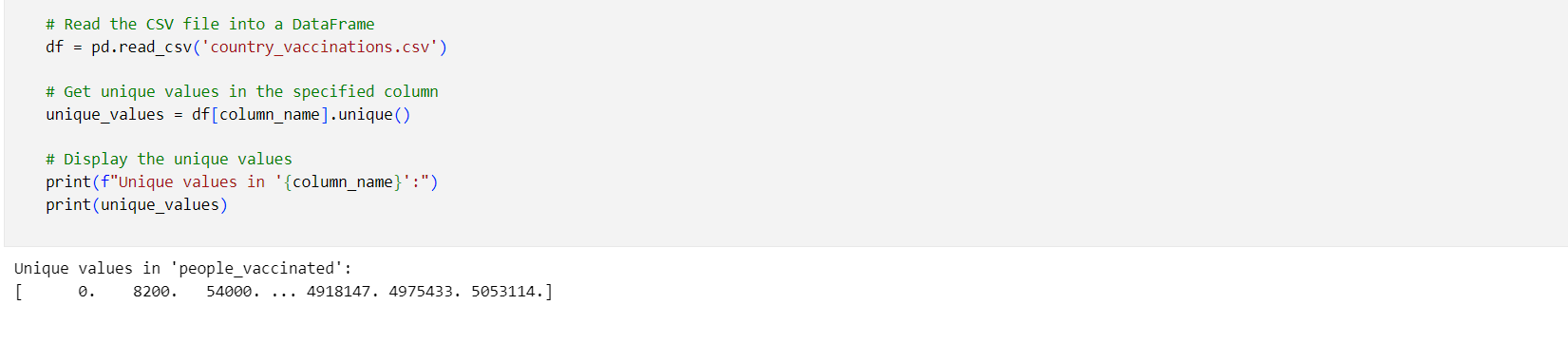


Then, the dataset is downloaded from the provided URL and saved as 'covid\_vaccine\_data.csv'.

The dataset is loaded into a pandas DataFrame ,df using the pd.read\_csv() function.

Next comes **DATA CLEANING** where the below steps are performed,

Initially, when all the unique values are displayed form the column, ”people\_vaccinated”. It can be seen that ‘nan’ value is also included in the dataset and so to make the dataset usable, all these are removed by,

After which when unique values are displayed, it can be seen that ‘nan’ value are removed.

Similarly, the same has been performed for all the columns. (Refer the code attached below.)

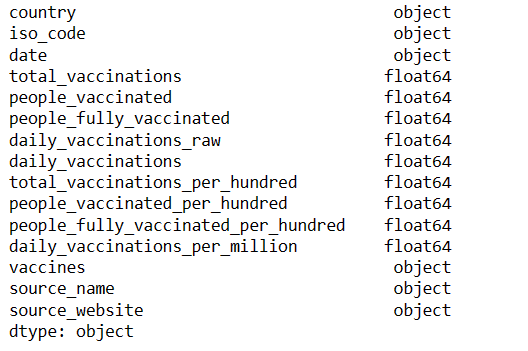
**EXPLORATORY DATA ANALYSIS (EDA):**

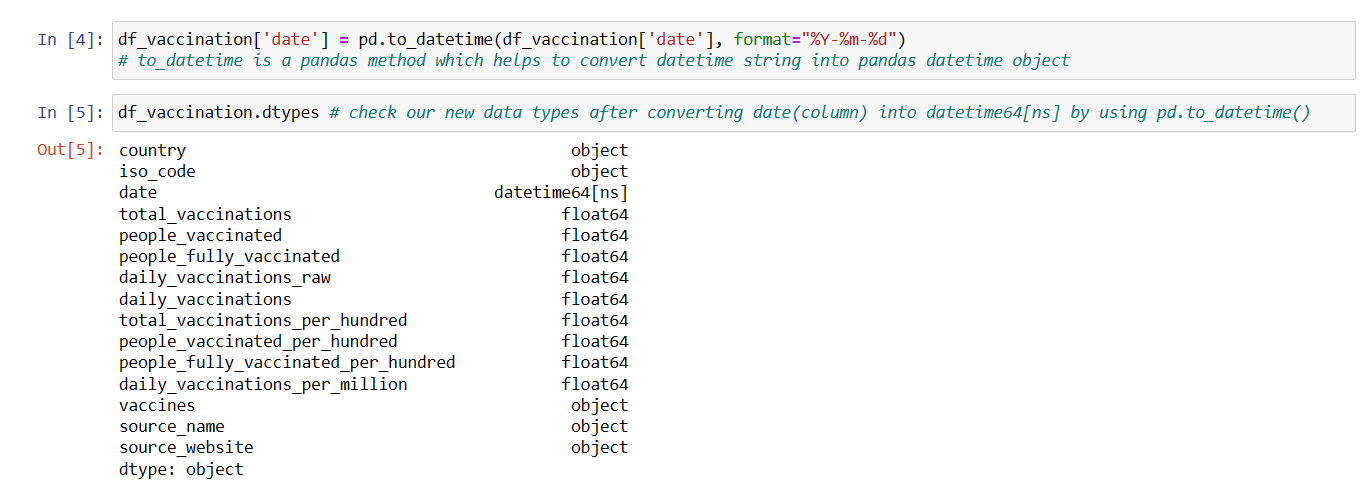
Exploratory Data Analysis (EDA) is an essential initial step in data analysis. It is the method of studying and exploring data set to recognize their traits, discover patterns, locate outliers, and identify relationships between variables.

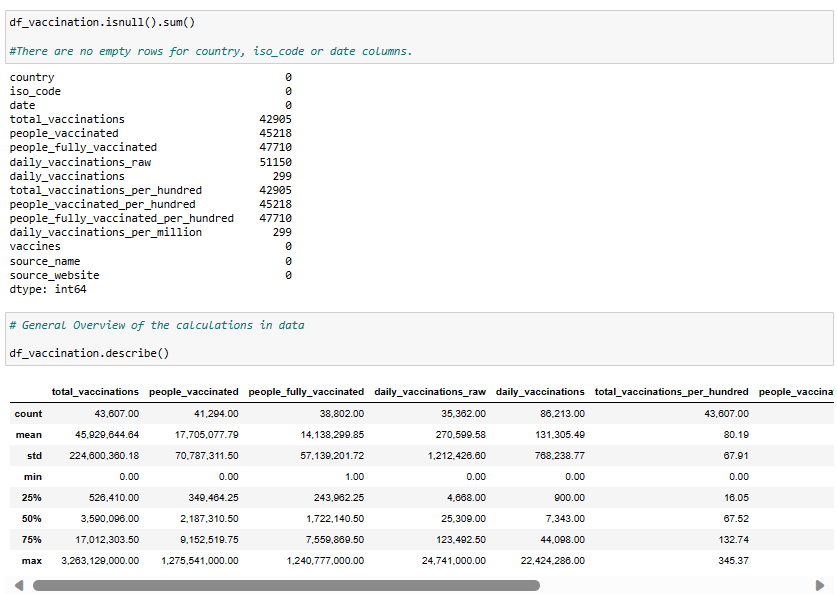
EDA is essential for getting a clear picture of the data which is useful in subsequent decision-making and can be performed using various statistical and graphical techniques. It involves multiple iterations and proves especially beneficial in prepping data for machine learning or statistical modeling. It is performed in the project as follows,

Initially, we take a look at the different types of data we have in our dataset.

The output is,



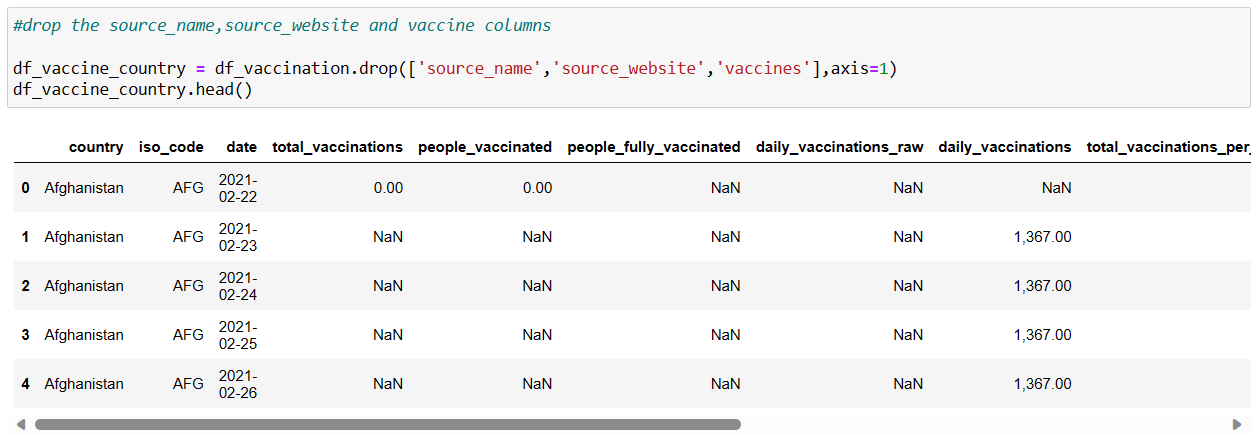
Note that the “date” field is of object datatype and so for better analysis, it is converted to datetime format by,

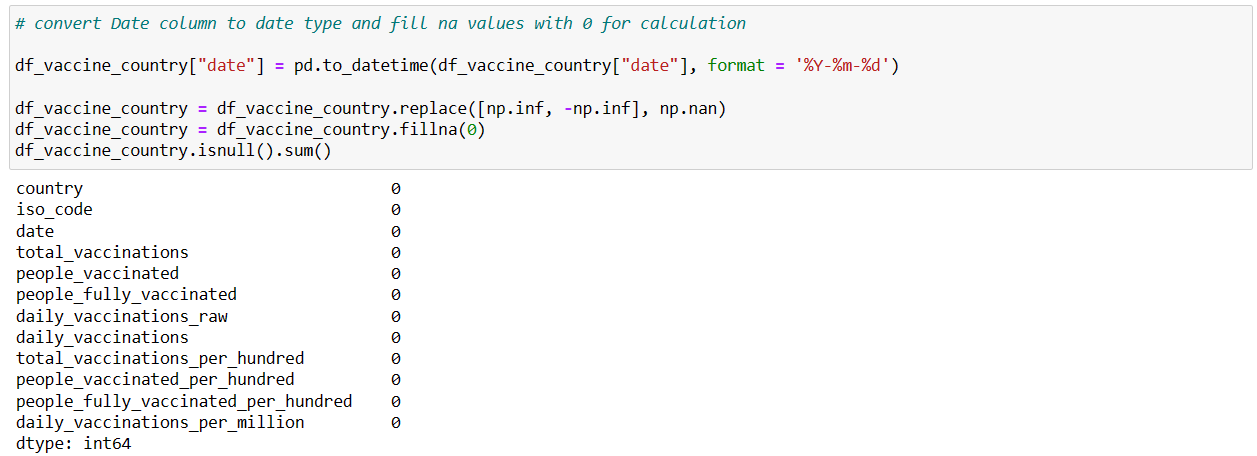
Now it can be seen that the datatype has been changed which makes it easier to work with it.

After which, various other fields are being examined to make sure we have the perfect set of data to analyze.

Followed by,

It is not always necessary that all the fields/attributes in the collected dataset is/are useful for our analysis.

 Therefore, the fields “source\_name”,”source\_website” and “vaccine\_columns” are not required and hence are dropped for more efficient analysis.

All the Nan values are then replaced by 0 to make calculations easier. From the screenshot below, it can be seen that the sum of all null values in every column is 0.

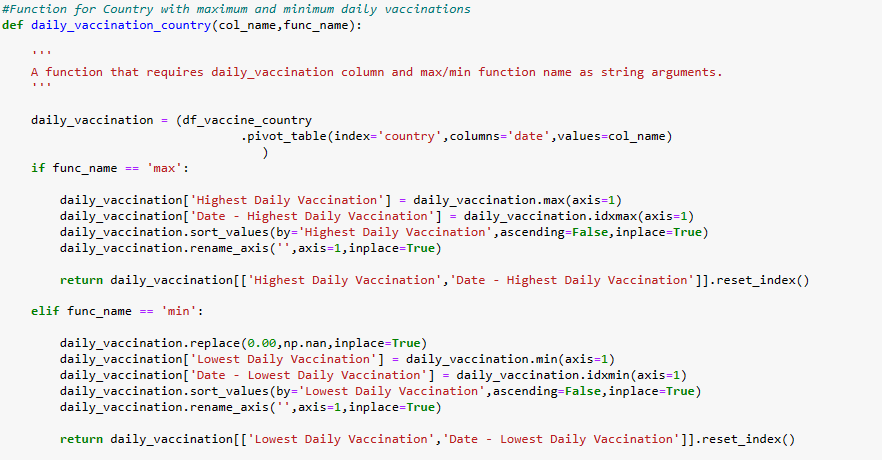
Once the dataset is prepared and ready for analysis, statistical analysis is performed on it.

**STATISTICAL ANALYSIS:**

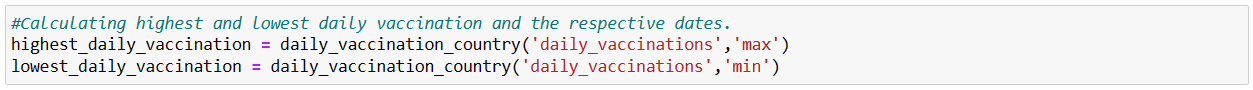
In statistical analysis, the total, average, maximum and minimum of different vaccinations status by country is calculated.



The code snippet of function for finding country with maximum and minimum daily vaccinations is,



Finally, calculating the highest and lowest daily vaccination and the respective dates.



Once all necessary aspects are calculated, it now time for visualization i.e., representing the analyzed records graphically for better understanding of complex data patterns and relations.

**VISUALIZATION**

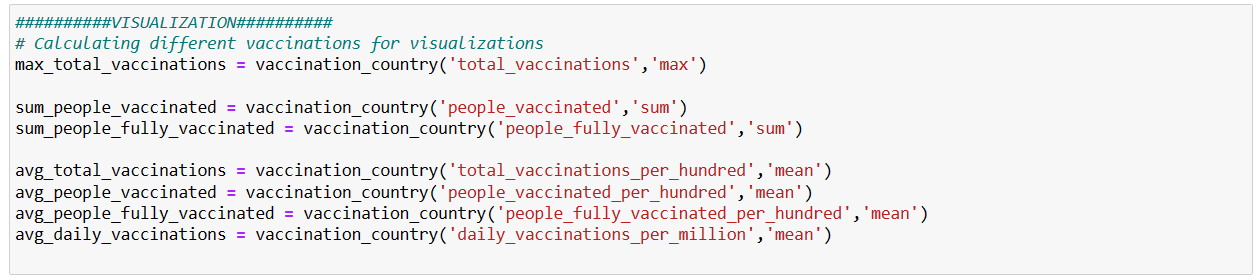
Data visualization is the use of graphical elements such as charts, graphs, and maps to represent data and information visually. The use of visualization tools provides an accessible way to see and understand trends, outliers, and patterns in data.

In machine learning (ML), visualization is a crucial tool for understanding data, model performance, and results.

It helps to gain insights, identify patterns, and communicate your findings effectively.

There are various techniques in data visualization. Few of them are described below,

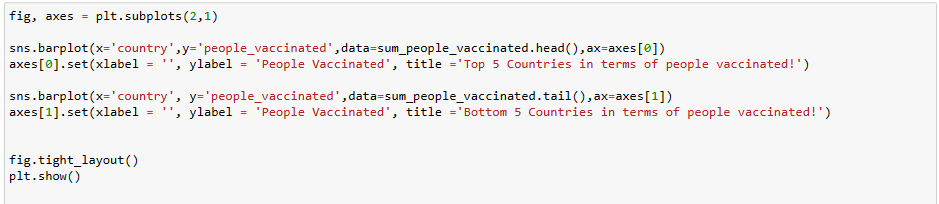
* ***Histograms***: Plot the frequency distribution of numerical variables to identify patterns and distributions.
* ***Box Plots***: Display the distribution, central tendency, and outliers in numerical data.
* ***Scatter Plots***: Visualize relationships between two numerical variables to identify correlations or patterns.
* ***Bar Charts***: Used for categorical data to show the frequency of different categories.
* ***Heatmaps***: Display the correlation between variables using color gradients.
* ***Pair Plots***: When dealing with multiple numerical variables, pair plots help visualize relationships between them.

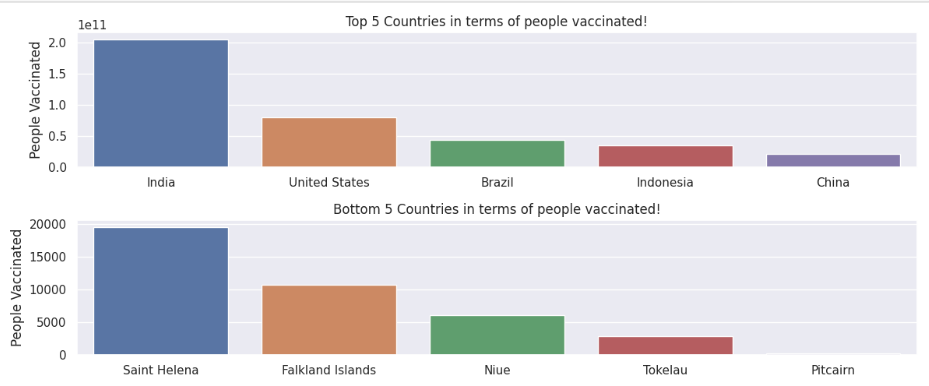


First, all required parameters are calculated using the previously created functions

Then, a bar graph is used to represent the Top 5 and Bottom 5 countries in terms of total vaccinations.

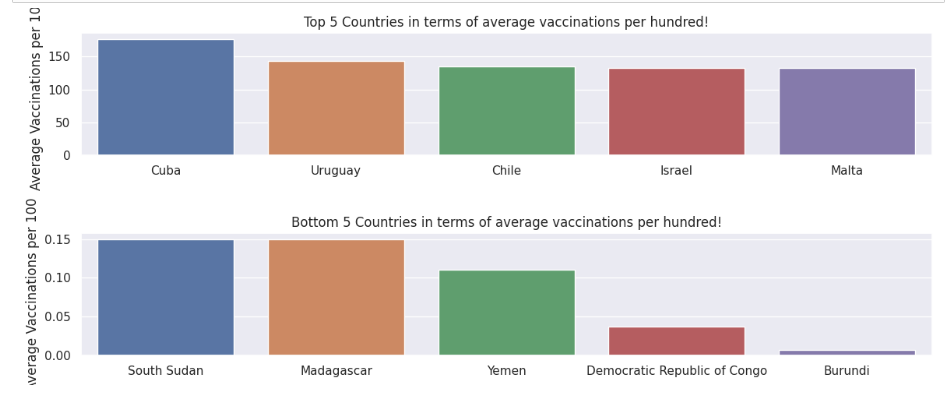
Further, the Python library Seaborn to create a vertical bar plot that displays the Top 5 and Bottom 5 countries in terms of people vaccinated.

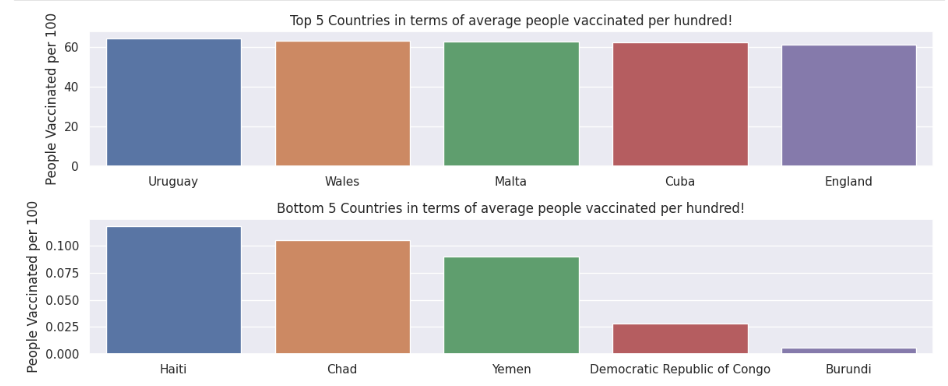


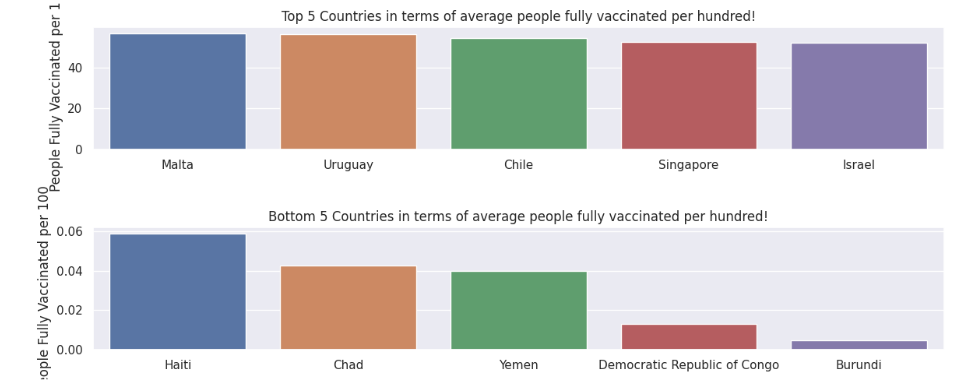


The code snippet creates a barplot visualization using the Seaborn library in Python. It is shows two barplots side by side in a single figure, comparing the top 5 and bottom 5 countries in terms of the number of people vaccinated COVID-19.

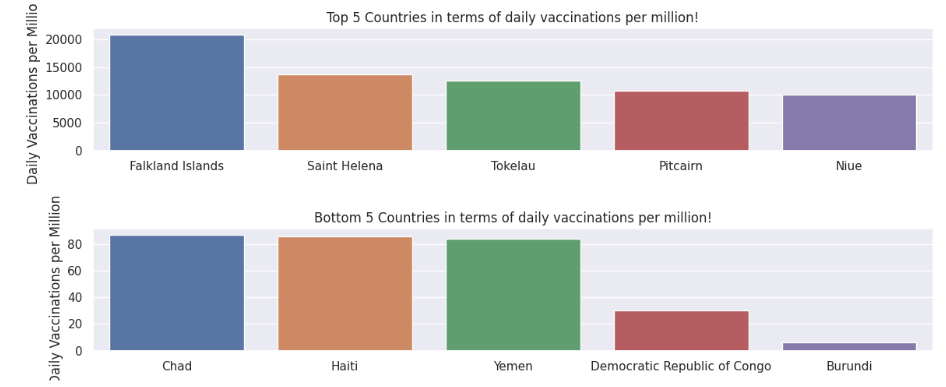
In a similar method the Top 5 Countries in terms of average vaccinations per hundred and Bottom 5 Countries in terms of average vaccinations per hundred are produced as a bar graph.

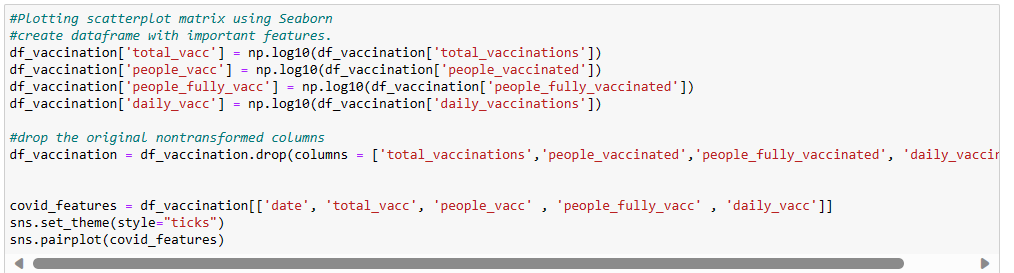


Further, the same is generated for people vaccinated per hundered for analysis,

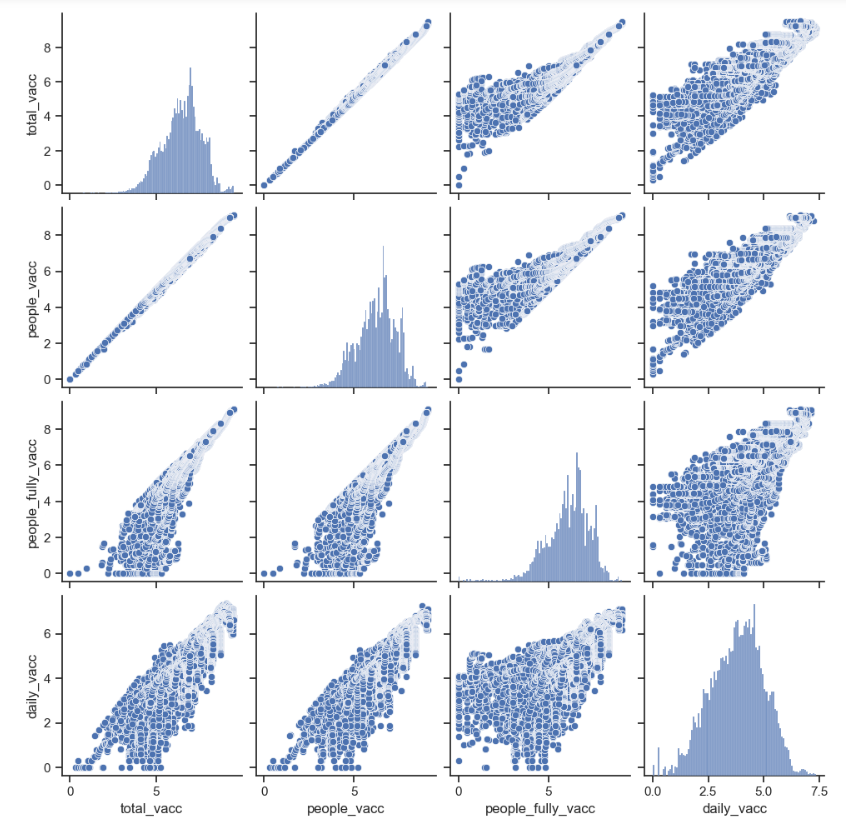
Similarly, the same process is applied to generate barplots representing the average number of people fully vaccinated per hundred for further analysis.

Once again, a similar approach is used to create barplots that display the daily vaccinations per million for the purpose of conducting further analysis.

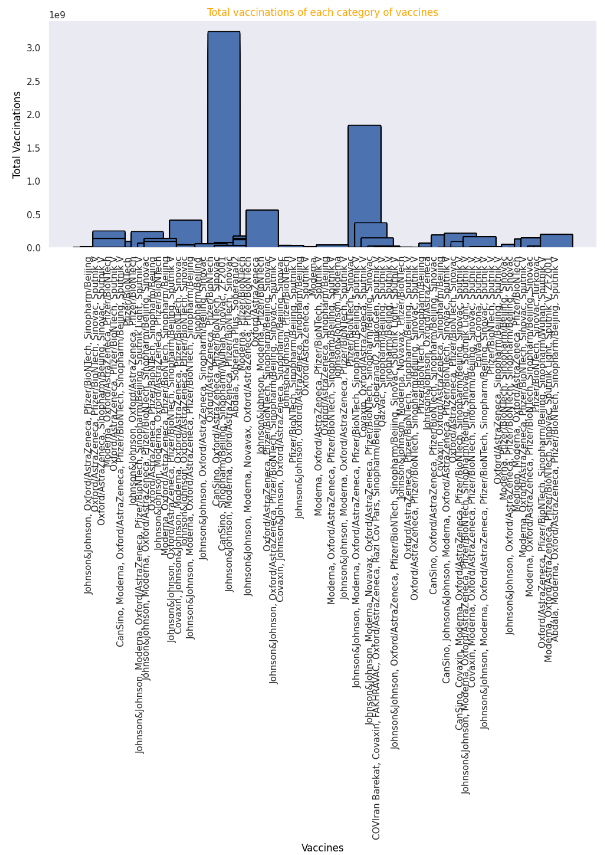


Further, a scatterplot matrix that visualizes the relationships and correlations between the transformed features in the DataFrame covid\_features is generated by,

Here scatter plot is used for which the output is,



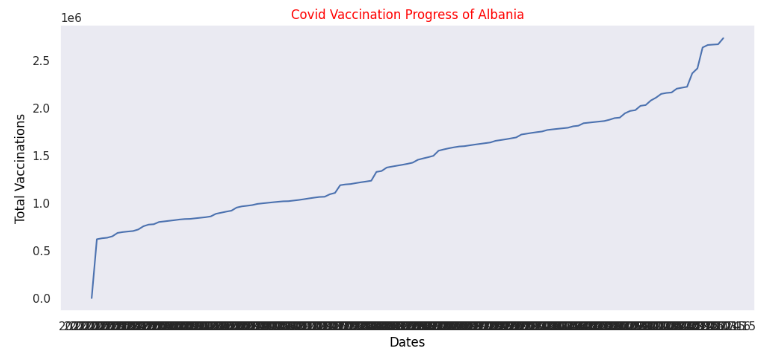
Further, additional visualization techniques are used on the preprocessed dataset. They are,

The above code snippet produces a bar chart that displays the total vaccinations for each category of vaccines, with vaccine names on the x-axis and the corresponding total vaccinations on the y-axis. It provides a visual representation of the distribution of total vaccinations across different vaccine categories. The output is,

Further in order to represent the data of one particular country (here Albania) the following is done,

The line plot generated illustrates the progress of COVID-19 vaccinations in Albania over time, with dates on the x-axis and the total number of vaccinations on the y-axis.

The plot provides a visual representation of how vaccination numbers have changed over a period.



***KEY FINDINGS AND INSIGHTS***

* The project identifies the countries with the highest and lowest daily vaccination rates and provides the respective dates when these extremes were recorded.
* It provides functionality to rank countries based on different vaccination statistics, such as total vaccinations, average vaccinations, maximum daily vaccinations, and minimum daily vaccinations.
* The analysis reveals trends in vaccination rates, helping to understand how quickly the population is getting vaccinated.
* It provides information on the distribution of different vaccine types, which can be critical for monitoring the effectiveness of specific vaccines in the region.
* By comparing the top and bottom 5 countries in terms of total vaccinations, you can gain insights into the variation in vaccination efforts among different nations.
* The analysis performed shifts the focus to the number of people vaccinated in various countries, providing insights into the extent of population coverage in terms of vaccination.
* By comparing the top and bottom countries in terms of people vaccinated, it is possible to gain insights into how different countries are faring in terms of vaccinating their populations.
* The analysis helps in understanding which countries are leading or lagging in terms of people vaccinated can have policy implications for allocating resources, improving distribution strategies, and addressing vaccination disparities.
* The bar plots visualizing the top 5 countries with the highest number of people fully vaccinated and the bottom 5 countries with the lowest number of people fully vaccinated highlights the progress made in achieving complete vaccination coverage in different countries.
* The next set of bar plots displaying the top 5 countries with the highest average vaccinations per hundred people and the bottom 5 countries with the lowest average vaccinations per hundred people allows for the assessment of the vaccination rate relative to the population size in different countries.
* The final set of bar plots illustrates the top 5 countries with the highest average number of people vaccinated per hundred people and the bottom 5 countries with the lowest average number of people vaccinated per hundred people which provides insights into the proportion of the population that has received at least one vaccine dose in different countries.
* Further,after the data transformation, a pairplot is created to explore the relationships between the transformed features, including 'total\_vacc,' 'people\_vacc,' 'people\_fully\_vacc,' and 'daily\_vacc.' The pairplot provides a visual representation of correlations or patterns in the data.
* The pairplot can help identify potential correlations or dependencies between different vaccination-related features. For example, it can show whether there is a positive relationship between the number of total vaccinations and the number of people fully vaccinated. It can also reveal any outliers or unusual patterns in the data.
* The pairplot analysis can be a starting point for more in-depth exploration and statistical analysis of the relationships between the transformed features. It may also lead to additional insights about the effectiveness of vaccination campaigns and their impact on different metrics.
* In the line plot,by plotting the total vaccinations against dates, the plot provides a time series view of the vaccination progress. This allows viewers to observe trends, fluctuations, and milestones in the vaccination campaign.
* These visualizations can help policymakers and health authorities identify countries that have successfully vaccinated a large proportion of their population, as well as those that need to ramp up their vaccination efforts. It can aid in understanding the effectiveness of vaccination campaigns and the impact of different strategies on population coverage.

***RECOMMENDATIONS***

* The project visualizations showing the total vaccinations, people vaccinated, people fully vaccinated, and daily vaccinations in various countries.

Policymakers and health authorities can use these insights to assess the effectiveness of vaccination campaigns and allocate resources accordingly.

* It includes a bar plot that shows the total vaccinations for different categories of vaccines, health organizations can use this information to optimize vaccine distribution strategies and assess the performance of different vaccine types.
* It also provides a time series plot of total vaccinations in Albania, health authorities can monitor and adjust vaccination strategies based on time series data, ensuring timely and efficient vaccination campaigns.
* The pair plots generated to explore relationships between multiple vaccination-related features can help researchers and analysts to identify key variables that may influence vaccination outcomes and further investigate these relationships.
* The logarithmic transformations that are applied to certain columns in the dataset, so analysts should consider the purpose and implications of the transformation and interpret the results accordingly.
* Further analysis should be conducted to answer specific research questions or to support decision-making related to vaccination campaigns.
* Further we can conduct a regional analysis to identify disparities in vaccination rates and access to vaccines among different areas within a country. This can help tailor vaccination strategies to address specific regional needs.
* And, can investigate the effectiveness of different vaccine categories and their impact on the reduction of COVID-19 cases, hospitalizations, and mortality. Such studies can guide vaccine distribution and public health policies.
* Also, time series forecasting techniques can be applied to predict future vaccination rates and the potential timing of reaching herd immunity. This information is critical for resource allocation and planning.
* It is recommended to develop and implement long-term vaccination strategies that address the potential need for booster shots, maintaining immunity over time, and adapting to evolving variants of the virus.

In summary, the analysis performed offers valuable tools for in-depth analysis of COVID-19 vaccination data. Policymakers, health authorities, and researchers can leverage this analysis to optimize vaccination strategies, address regional disparities, assess vaccine efficacy, and forecast future vaccination rates.. These recommendations not only aid the current pandemic but also future public health challenges, promoting informed, evidence-based decision-making to safeguard public health and well-being.

***CONCLUSION***

In conclusion, the analysis performed offers a valuable set of tools for exploring and understanding COVID-19 vaccination data. These tools empower policymakers, health authorities, and researchers to gain critical insights into vaccination progress, optimize strategies, and make data-driven decisions in the ongoing battle against the COVID-19 pandemic.

The analyses covers a broad spectrum of factors, including the evaluation of total vaccinations and the number of people vaccinated, as well as a detailed examination of vaccination categories and disparities among different regions. Moreover, through the visualization of time series data and the exploration of relationships among multiple variables, the provided code facilitates a holistic comprehension of the intricate dynamics in the dataset.

It is worth emphasizing the importance of conducting further research and considering the broader context to draw comprehensive conclusions. This approach goes beyond managing the pandemic and extends to addressing future public health challenges. The recommendations outlined in this analysis function as a guiding path for making decisions grounded in empirical evidence. These decisions are of paramount importance in preserving public health and the welfare of populations.

In a world where data-driven insights are becoming increasingly inevitable, this adds on to the array of instruments that can aid in addressing not just the crisis but also forthcoming public health challenges, ultimately saving lives and promoting the welfare of communities worldwide.