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

**PROJECT PHASE 4: DEVELOPMENT PART 2**

**PROJECT TITLE:**

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

Enhancing deployment strategies for optimal public health impact

**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) |

***INTRODUCTION***

In this phase, we will continue building the project by performing the following:

* Exploratory data analysis (EDA)
* Statistical analysis
* Visualization

The above processes are performed on the already loaded and pre-processed dataset. During this stage, we will carry on with the project by conducting various analysis on the prepared dataset and finally visualization techniques are used to represent the same for better and clearer understandings.

This step marks the next milestone of our project in understanding the data we have.

***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)

***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 above processes have already been performed and documented in the previous phase. Now,

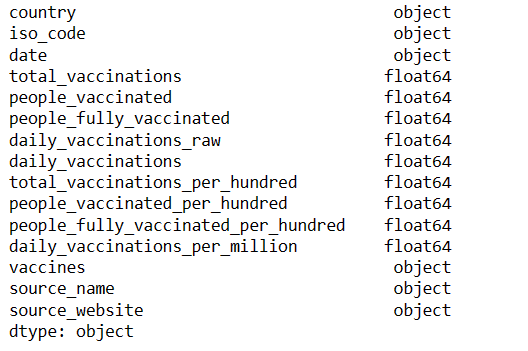
***EXPLORATORY DATA ANALYSIS (EDA)***

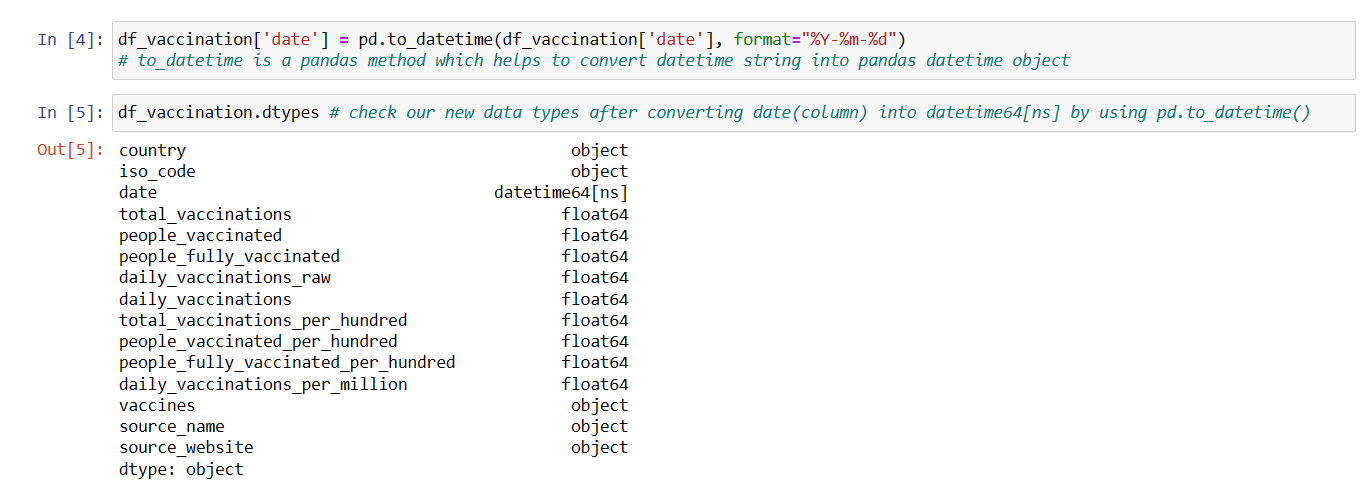
Exploratory Data Analysis (EDA) is an essential initial step in data analysis. It is the method of studying and exploring data set to recognize their traits, discover patterns, locate outliers, and identify relationships between variables.

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,

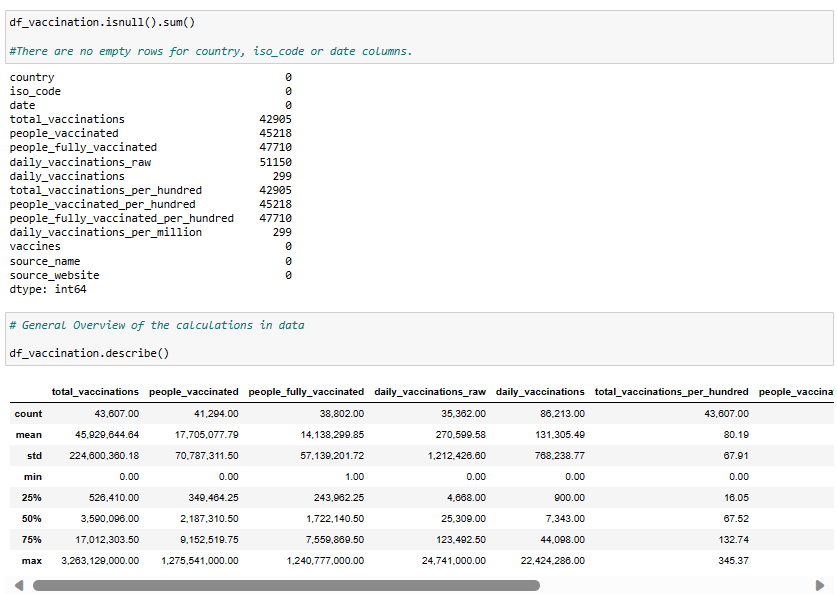
Initially, we take a look at the different types of data we have in our dataset.

The output is,



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

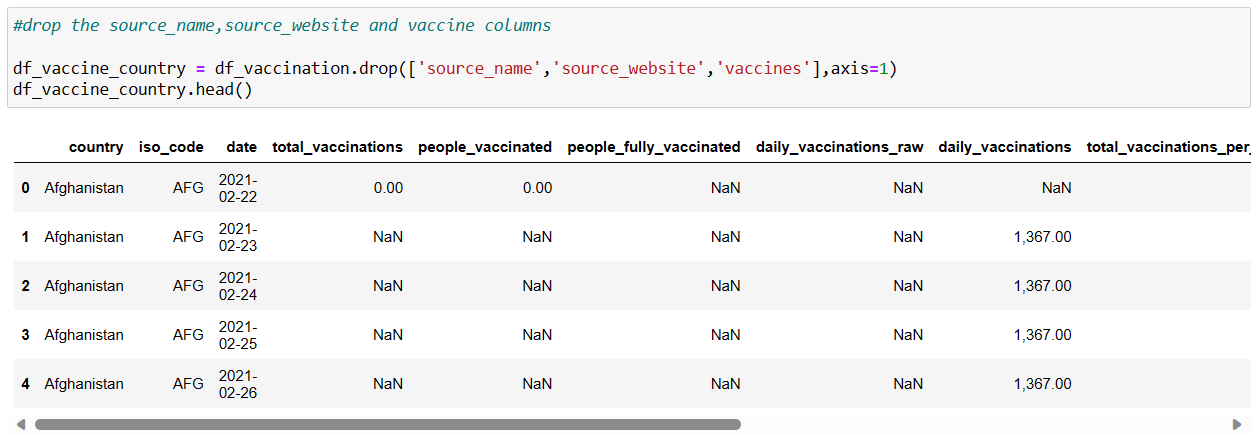
Now it can be seen that the datatype has been changed which makes it easier to work with it.

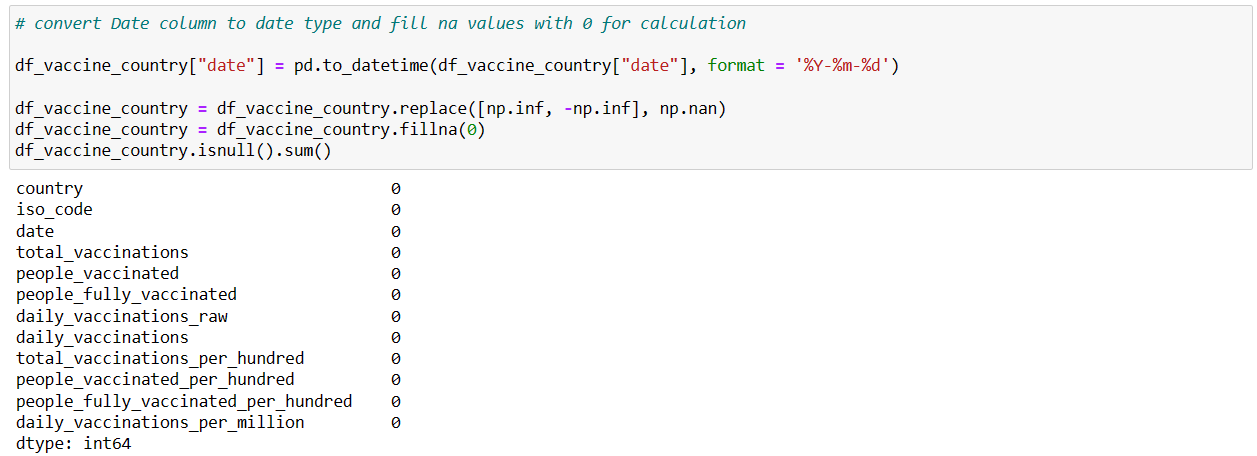
After which, various other fields are being examined to make sure we have the perfect set of data to analyze.

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.

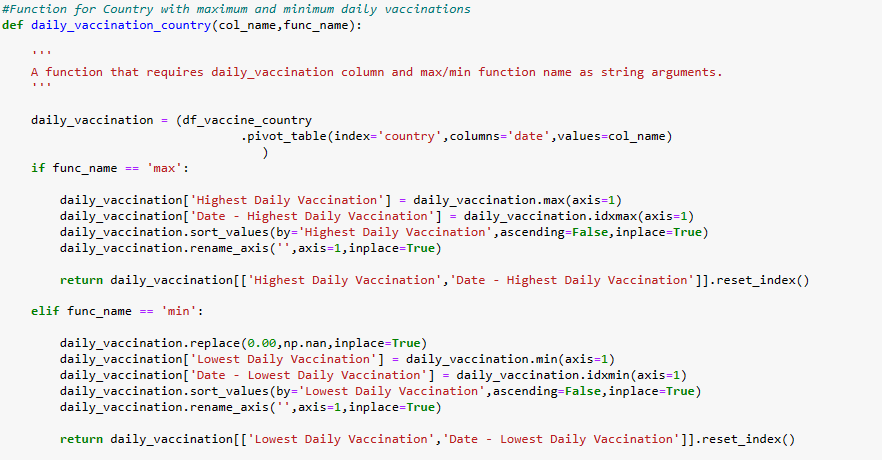


All the Nan values are then replaced by 0 to make calculations easier. From the screenshot below, it can be seen that the sum of all null values in every column is 0.

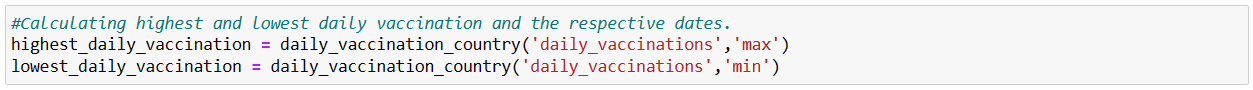
Once the dataset is prepared and ready for analysis, statistical analysis is performed on it.

***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.



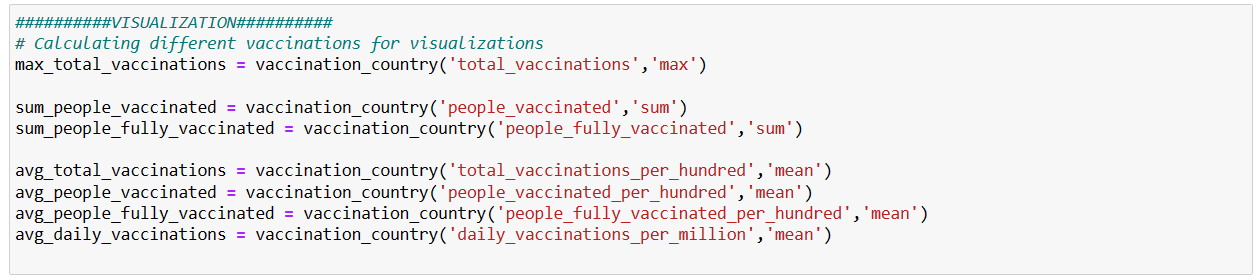
Once all necessary aspects are calculated, it now time for visualization i.e., representing the analyzed records graphically for better understanding of complex data patterns and relations.

***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.

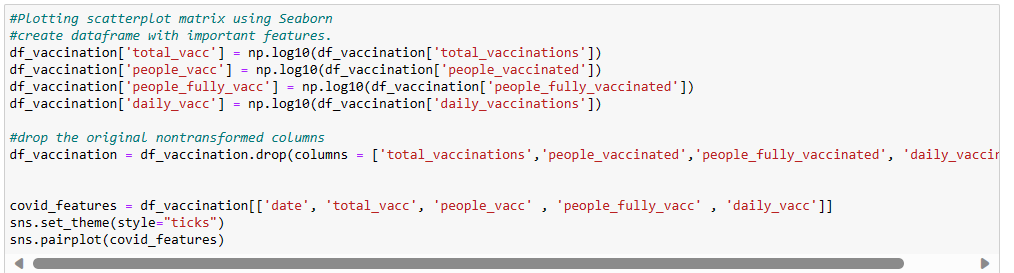
There are various techniques in data visualization. Few of them are described below,

* ***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.

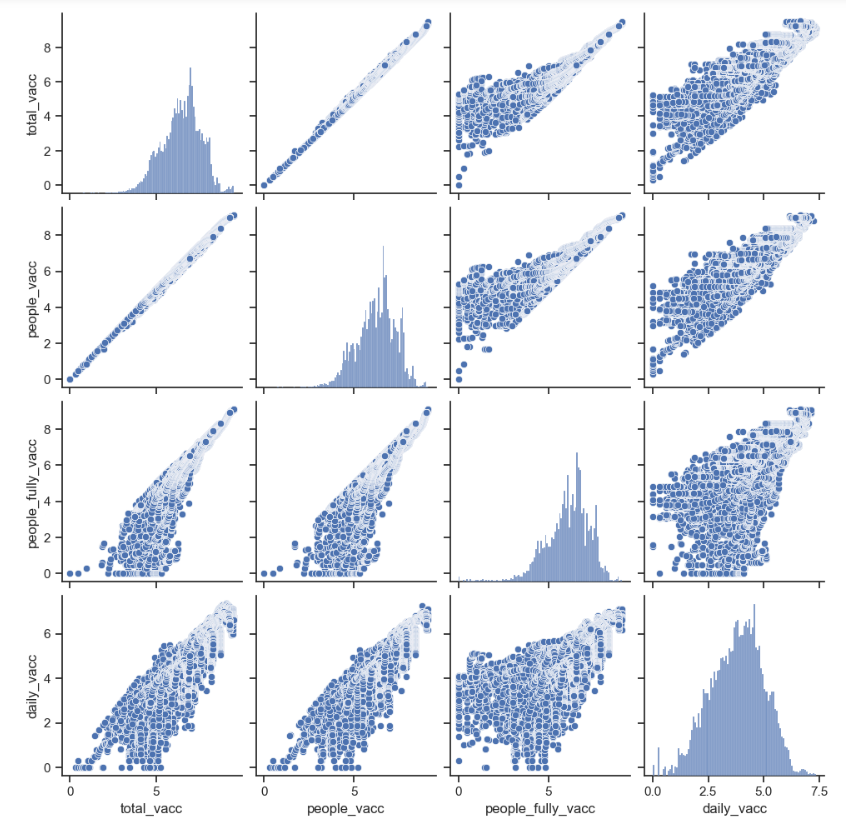
First, all required parameters are calculated using the previously created functions.



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



Here scatter plot is used for which the output is,



***CONCLUSION:***

At the end of this phase, the collected and prepared data has been gone through Exploratory data analysis (EDA) and Statistical analysis. And finally, visualization tools have been used to graphically represent the analyzed data which helps in deeper insights on the same.