**COVID-19 VACCINE ANALYSIS**

**Phase 5**

**TEAM MEMBERS:**

1. NISHANTH.S.D 2021115070
2. NITHYASRI INDHUMATHI.S 2021115072

3.NITHYASRI.S.R 2021115073

4.PRABHAVATHI.R 2021115074

5.YAZHINIYAN.P 2021115306

**Overall work of Naan Mudhalvan – All phases**

Phase 1

**In-Depth Analysis of Global COVID-19 Vaccine**

**Landscape: Insights for Policy and Action**

**Introduction:**

The COVID-19 pandemic has ushered humanity into an era of

unparalleled challenges, where the rapid development and deployment of

vaccines have emerged as our most potent weapon against the virus. This

ambitious project undertakes a comprehensive exploration of COVID-19

vaccine data, meticulously scrutinizing vaccine efficacy, distribution

dynamics, and the intricacies of adverse effects. Our overarching goal is to

furnish a profound understanding that empowers policymakers,

healthcare organizations, and researchers with the nuanced insights

needed to optimize vaccine deployment strategies and decisively

influence the course of the pandemic.

**Project Definition:**

This project's scope is articulated by its commitment to address

pivotal inquiries with unwavering depth:

1. **Vaccine Efficacy at Scale:** Our first undertaking is an ambitious

evaluation of the efficacy of diverse COVID-19 vaccines across the

spectrum, ranging from prevention of infection to mitigation of

severe outcomes and mortality. We traverse the data landscape to

unearth granular insights, considering a multitude of variables

including vaccine type, administration protocols, geographical

disparities, and temporal dynamics.

2. **Distribution Equality and Equity:** The achievement of global herd

immunity hinges on equitable vaccine distribution. Our analysis

extends beyond surface-level statistics, aiming to dissect the heart of

global vaccination campaigns. We pinpoint regions and

demographics facing barriers to access, and, crucially, we identify the

strategies and mechanisms that can facilitate universal vaccine

access.

3. **Adverse Effects and Safety Monitoring:** The safeguarding of public

health necessitates relentless vigilance. We delve deep into adverse

event reports, conducting a rigorous analysis to discern signals

within the data. We track trends, unearth latent patterns, and provide

actionable insights to support ongoing vaccine safety monitoring

and refinement.

**Methodology:**

This project unfolds through a meticulously orchestrated symphony

of phases:

1. **Data Collection Odyssey:** We embark on a data odyssey, harnessing

the rich and expansive COVID-19 vaccination dataset generously

made available on Kaggle. This trove of information spans the

breadth of vaccination statistics, vaccine characteristics, and adverse

event reports.

2. **Data Preprocessing Alchemy:** The journey continues with data

preprocessing, where we apply alchemical precision to ensure data

purity. Missing data are harmonized, outliers are detected and

addressed, and data are meticulously standardized for rigorous

analysis.

3**. Exploratory Data Odyssey (EDO):** The data come to life in the EDO

phase, where we conduct a panoramic survey of the dataset.

Descriptive statistics paint an intricate portrait, visualization

techniques reveal the underlying narratives, and initial correlations

are unveiled.

4. **Statistical Depth Dive:** Statistical analysis is where we plunge into

the profound. Hypothesis testing, regression analysis, and machine

learning techniques become our tools as we quantify vaccine

efficacy, expose distribution disparities, and unravel the intricacies of

adverse events.

5**. Visualization Spectacle:** Our insights are brought to life through a

spectacular array of data visualizations—immersive graphs,

interactive dashboards, and captivating charts. These visuals facilitate

not just understanding but comprehension, offering stakeholders a

vivid window into the complex web of vaccine data.

**Conclusion:**

This project aspires to be a defining testament to the collective

human response to the COVID-19 pandemic. Our mission, rooted in

exhaustive analysis and unfaltering commitment, seeks to provide a

monumental resource for those shaping the pandemic's narrative—

policymakers, healthcare luminaries, and researchers. By ploughing

through the depths of vaccine efficacy, distribution, and safety, we hope

to light the path toward evidence-based decision-making, ultimately

steering the world closer to the end of this unprecedented global crisis

**Phase 2:**

**CONTENTS:**

* Introduction
* Project Objectives
* Data Preparation
* Data Collection
* Innovation Steps
* Technology utilized

**Introduction:**

The ongoing COVID-19 pandemic has brought unprecedented

challenges to humanity. In response, the rapid development and deployment of vaccines have emerged as our most potent weapon against the virus. This ambitious project seeks to comprehensively explore COVID-19 vaccine data, focusing on vaccine efficacy, distribution dynamics, and adverse effects. Its overarching goal is to provide policymakers, healthcare organizations, and researchers with nuanced insights to optimize vaccine deployment strategies and influence the pandemic's trajectory.

**Project Objectives:**

Comprehensive analysis of COVID-19 vaccines, focusing on efficacy, distribution equity, and safety monitoring.Provide nuanced insights for policymakers and healthcare organizations, supporting evidence-based decision-making to combat the pandemic.

**Data Preparation:**

Rigorous data cleaning, harmonizing missing values, and addressing outliers ensure data purity.Standardization guarantees error-free, consistent data, laying the foundation for reliable statistical analysis and machine learning.

**Data Collection:**

Sourcing data from platforms like Kaggle and government health agencies provides a rich dataset for vaccine analysis.Inclusive data collection encompasses vaccine types, administration protocols, geographical disparities, and temporal dynamics, ensuring a diverse information set.

**INNOVATION STEPS:**

**1.Data Collection Odyssey:**

**Description:** This step involves embarking on a journey to collect comprehensive COVID-19 vaccination data. The data is sourced from platforms like Kaggle, which generously provide a vast array of information, including vaccination statistics, vaccine characteristics, and adverse event reports.

**Innovation:** The innovation lies in the meticulous collection of rich and expansive data, ensuring that a wide range of variables and dimensions are considered. This diverse dataset forms the foundation for in-depth analysis.

**2.Data Preprocessing Alchemy:**

**Description:** After collecting the data, the project continues its journey with data preprocessing. Alchemical precision is applied to purify the data. This includes tasks like harmonizing missing data, detecting and addressing outliers, and meticulously standardizing the data for rigorous analysis.

**Innovation:** The innovation here is in the attention to detail and accuracy, ensuring that the data is clean and ready for advanced analysis. The data preprocessing step is crucial for drawing meaningful insights.

**3.Exploratory Data Odyssey (EDO):**

**Description:** In the EDO phase, the data comes to life as a panoramic survey is conducted. Descriptive statistics are used to paint an intricate portrait of the dataset. Visualization techniques are employed to reveal the underlying narratives, and initial correlations among variables are unveiled.

**Innovation:** Innovation lies in the use of advanced visualization techniques to make the data more accessible and understandable. It provides the project team with a deeper understanding of the dataset.

**4.Statistical Depth Dive:**

**Description:** This is where the project plunges into profound statistical analysis. Hypothesis testing, regression analysis, and machine learning techniques become the tools of choice. The team quantifies vaccine efficacy, exposes distribution disparities, and unravels the intricacies of adverse events.

**Innovation:** The innovation here is the application of advanced statistical and machine learning techniques to extract actionable insights from the data. It enables a deeper understanding of vaccine effectiveness and safety monitoring.

**5.Visualization Spectacle:**

**Description:** In the final phase, the project's insights are brought to life through a spectacular array of data visualizations. These include immersive graphs, interactive dashboards, and captivating charts. These visuals facilitate not just understanding but true comprehension, offering stakeholders a vivid window into the complex web of vaccine data.

**Innovation:** The innovation lies in the effective communication of findings through compelling and interactive data visualizations. This step ensures that the insights are accessible and impactful for decision-makers.

**TECHNOLOGIES UTILIZED:**

**1.Python with Pandas and NumPy:**

**Tool Name:** Python, Pandas, NumPy

**Description:** Python, along with the Pandas and NumPy libraries, is used for data cleaning and manipulation. Pandas provides data structures and functions to handle structured data, while NumPy offers support for numerical operations. Python's versatility makes it an excellent choice for working with diverse datasets and data preprocessing tasks.

**2.Tableau and Power BI:**

**Tool Name:** Tableau, Power BI

**Description:** Tableau and Power BI are data visualization tools used to create interactive dashboards and visualizations. These tools enable the project team to explore data, identify trends, and present findings in a visually engaging and comprehensible format, making it easier for stakeholders to understand complex data.

**3.Natural Language Processing (NLP) Libraries - NLTK and spaCy:**

**Tool Name:** NLTK, spaCy

**Description:** Natural Language Processing (NLP) libraries like NLTK (Natural Language Toolkit) and spaCy are used for real-time safety monitoring. These libraries facilitate text analysis and machine learning techniques to process and analyze adverse event reports, enabling the detection of trends and patterns in textual data.

**4.Prophet for Time Series Forecasting:**

**Tool Name:** Prophet

**Description:** Prophet is employed for time series forecasting in the project. It is a tool developed by Facebook that is well-suited for forecasting time-dependent data, such as vaccine demand, distribution needs, and safety monitoring requirements. It uses an additive model to capture seasonal and trend patterns in data.

**5.Blockchain Technology:**

**Tool Name:** Blockchain

**Description:** Blockchain technology is utilized for secure data recording and management. It ensures the integrity and transparency of data, making it an ideal choice for recording critical information related to vaccine distribution records and safety monitoring. Blockchain's decentralized and tamper-resistant nature enhances data security**.**

**Phase 3**

**INTRODUCTION:**

In the ongoing battle against the COVID-19 pandemic, the collection and preprocessing of COVID vaccine-related data have emerged as indispensable elements in shaping effective vaccination strategies and public health policies. Diverse sources, including government health agencies, the World Health Organization, research institutions, and pharmaceutical companies, contribute to the wealth of data that informs our understanding of vaccine distribution, coverage, and efficacy. Once gathered, data preprocessing becomes imperative to ensure the data's quality and suitability for analysis. This includes addressing missing values, standardizing formats, and transforming variables. Clean and well-structured data empowers analysts and researchers to extract meaningful insights, aiding in the global effort to combat the pandemic and manage the COVID-19 vaccine deployment with greater precision and impact.

**DATA PREPROCESSING OBJECTIVE:**

The primary objectives of data preprocessing for COVID vaccine analysis include ensuring data accuracy and consistency by handling missing values and outliers, standardizing data formats and units, transforming and encoding variables for analysis, and organizing data into a structured format that facilitates meaningful insights. This process sets the foundation for effective vaccine efficacy assessments, coverage rate evaluations, and informed decision-making to support global vaccination efforts in the fight against the COVID-19 pandemic.

**CODE :**

import pandas as pd

import numpy as np

'''Reading the data set using Pandas'''

df = pd.read\_csv(r"../country\_vaccinations.csv")

'''Now extracting some basic information from the dataset'''

print("The first 5 entries in the set")

print(df.head())

print("\n\nOverall information of the file")

print(df.info())

print("\n\nNull information in the file")

print(df.isnull().sum())

'''Now removing the rows that have null values in it'''

print("\n\nRemoving rows with null values")

df.dropna(inplace=True)

df = df.reset\_index(drop=True)

print("Removed")

print("Saving the cleaned CSV in a new folder named /cleaned\_csv")

df.to\_csv(r"../cleaned\_csv/preprocessed\_country\_vaccination.csv",index=False)

'''Now creating a new CSV to store the Vaccines used in a country'''

columns\_of\_interest = ['iso\_code','vaccines']

selected\_columns\_df = df[columns\_of\_interest]

unique\_values\_df = selected\_columns\_df.drop\_duplicates()

unique\_values\_df = unique\_values\_df.reset\_index(drop=True)

print("Created a helping dataset to know about the vaccines used in each county...")

unique\_values\_df.to\_csv(r"../cleaned\_csv/vacc.csv")

**OUTPUT**

The first 5 entries in the set

country iso\_code ... source\_name source\_website

0 Afghanistan AFG ... World Health Organization https://covid19.who.int/

1 Afghanistan AFG ... World Health Organization https://covid19.who.int/

2 Afghanistan AFG ... World Health Organization https://covid19.who.int/

3 Afghanistan AFG ... World Health Organization https://covid19.who.int/

4 Afghanistan AFG ... World Health Organization https://covid19.who.int/

[5 rows x 15 columns]

* Overall information of the file

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 86512 entries, 0 to 86511

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 country 86512 non-null object

1 iso\_code 86512 non-null object

2 date 86512 non-null object

3 total\_vaccinations 43607 non-null float64

4 people\_vaccinated 41294 non-null float64

5 people\_fully\_vaccinated 38802 non-null float64

6 daily\_vaccinations\_raw 35362 non-null float64

7 daily\_vaccinations 86213 non-null float64

8 total\_vaccinations\_per\_hundred 43607 non-null float64

9 people\_vaccinated\_per\_hundred 41294 non-null float64

10 people\_fully\_vaccinated\_per\_hundred 38802 non-null float64

11 daily\_vaccinations\_per\_million 86213 non-null float64

12 vaccines 86512 non-null object

13 source\_name 86512 non-null object

14 source\_website 86512 non-null object

dtypes: float64(9), object(6)

memory usage: 9.9+ MB

None

* Null information in the file

country 0

iso\_code 0

date 0

total\_vaccinations 42905

people\_vaccinated 45218

people\_fully\_vaccinated 47710

daily\_vaccinations\_raw 51150

daily\_vaccinations 299

total\_vaccinations\_per\_hundred 42905

people\_vaccinated\_per\_hundred 45218

people\_fully\_vaccinated\_per\_hundred 47710

daily\_vaccinations\_per\_million 299

vaccines 0

source\_name 0

source\_website 0

dtype: int64

Removing rows with null values

Removed

**EXPLANATION:**

The provided Python code is designed to prepare a COVID vaccine dataset for analysis. It starts by loading the dataset from a CSV file and then conducts an initial exploration, displaying the first five entries and providing an overview of the dataset's structure. Missing values are identified and subsequently removed, ensuring that only complete and reliable data remains. The cleaned dataset is then saved to a new file for future analysis. Additionally, the code creates a separate dataset to track the unique combinations of 'iso\_code' and 'vaccines' for each country, shedding light on the vaccines used across different regions. By conducting these data preprocessing steps, the code paves the way for meaningful insights into vaccine efficacy, distribution, and coverage, contributing to more effective global efforts to combat the COVID-19 pandemic.

**CONCLUSION:**

The data preprocessing steps play a vital role in preparing the dataset for COVID vaccine analysis. By removing missing data and creating a dataset to understand vaccine usage in different countries, the code ensures that the analysis can be conducted with high-quality, complete data. These preprocessed datasets are essential for obtaining valuable insights into vaccine efficacy, distribution, and coverage, aiding global efforts to combat the COVID-19 pandemic more effectively.

**Phase 4**

***INTRODUCTION***

*In this phase, we will continue advancing our project by performing the following essential tasks:*

* *Exploratory Data Analysis (EDA)*
* *Statistical Analysis*
* *Data Visualization*

*These critical processes are carried out on the dataset we have already collected and pre-processed. During this stage, we will delve deeper into our project by conducting various analyses on the prepared dataset and using visualization techniques to present the findings for a clearer and more insightful understanding. This step marks a significant milestone in our project, helping us gain valuable insights from the data at our disposal.*

***DATA AQUISTION:***

Collecting data for COVID-19 vaccine analysis is a crucial aspect of comprehending the effectiveness and impact of vaccination campaigns, enabling us to conduct a thorough analysis. This process entails gathering a wide range of information related to vaccination efforts, including:

* Country-wise total vaccinations administered
* Number of people vaccinated
* Adverse events reported
* Distribution logistics and so on..

The primary source of data for our project is the Kaggle dataset, which provides comprehensive information related to global COVID-19 vaccination progress. You can access the dataset through the following link:

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

***DATA PREPARATION:***

Once we have collected reliable data, the next step is to clean and prepare the data for analysis, a crucial process known as Data Pre-processing. It's important to note that these pre-processing steps have already been performed and well-documented in the previous phase. Now, we will build upon this foundation and proceed with the subsequent stages of our analysis.

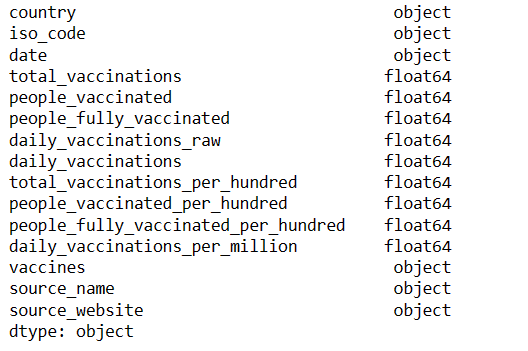
***EXPLORATORY DATA ANALYSIS (EDA)***

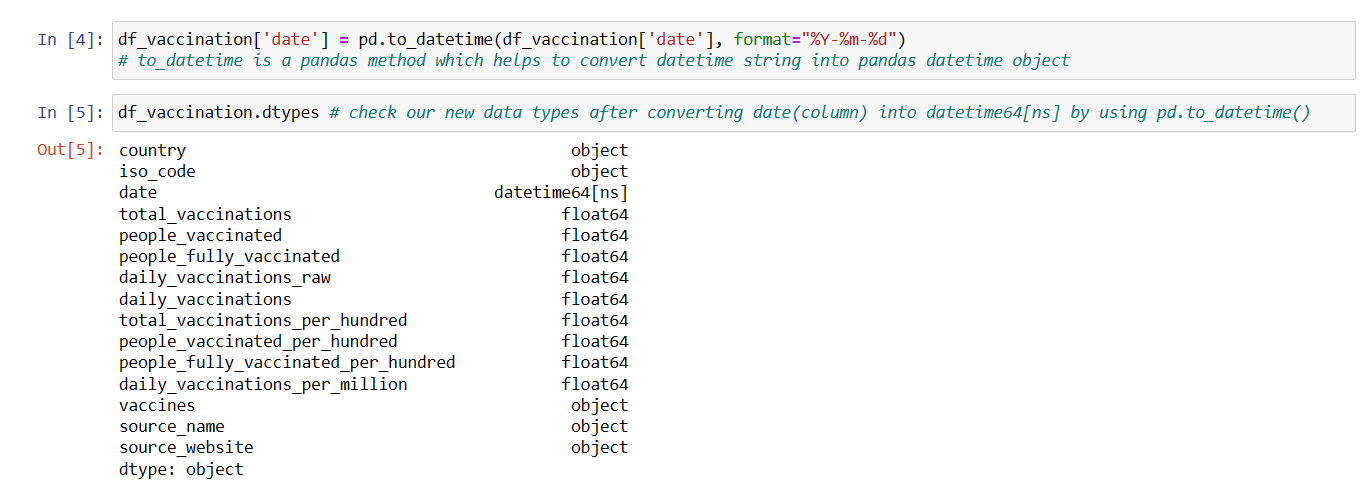
Exploratory Data Analysis (EDA) is an essential initial step in data analysis. It is the method of studying and exploring data set to recognizetheir traits, discover patterns, locate outliers, and identify relationships between variables.

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,

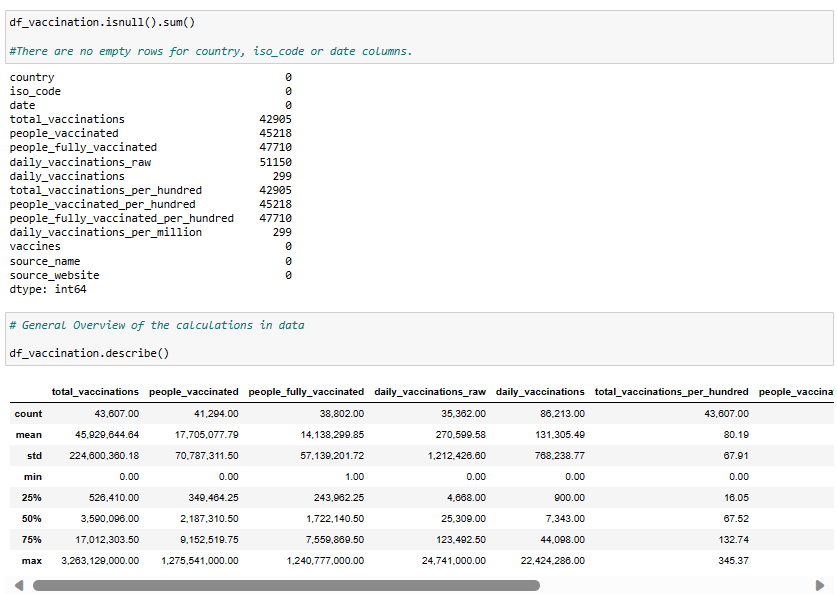
Initially, we take a look at the different types of data we have in our dataset.

The output is,



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

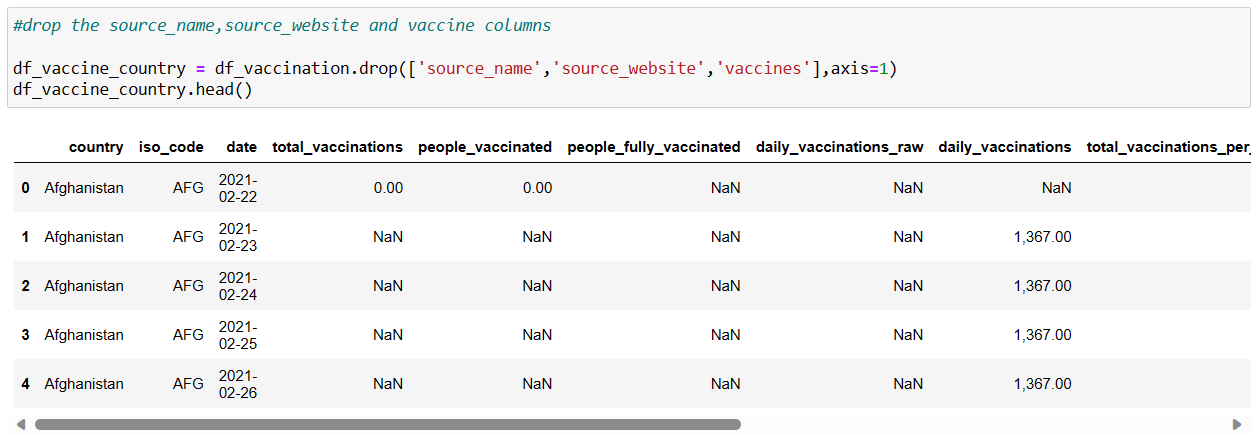
Now it can be seen that the datatype has been changed which makes it easier to work with it.

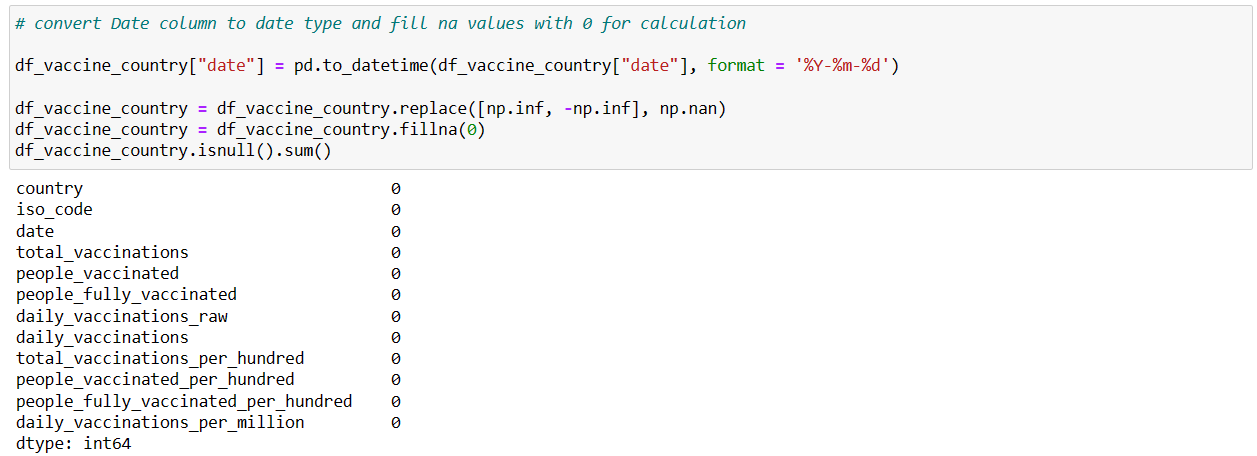
After which, various other fields are being examined to make sure we have the perfect set of data to analyze.

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.

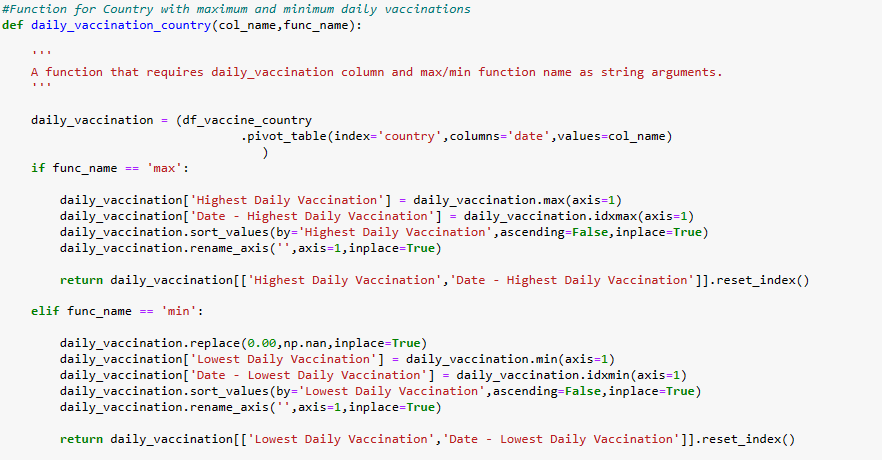


All the Nan values are then replaced by 0 to make calculations easier. From the screenshot below, it can be seen that the sum of all null values in every column is 0.

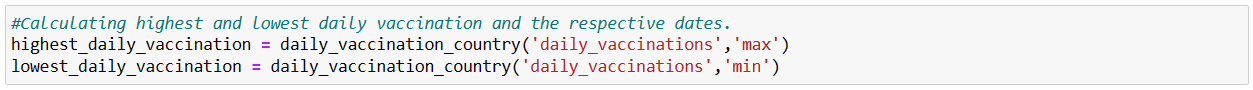
Once the dataset is prepared and ready for analysis, statistical analysis is performed on it.

***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.



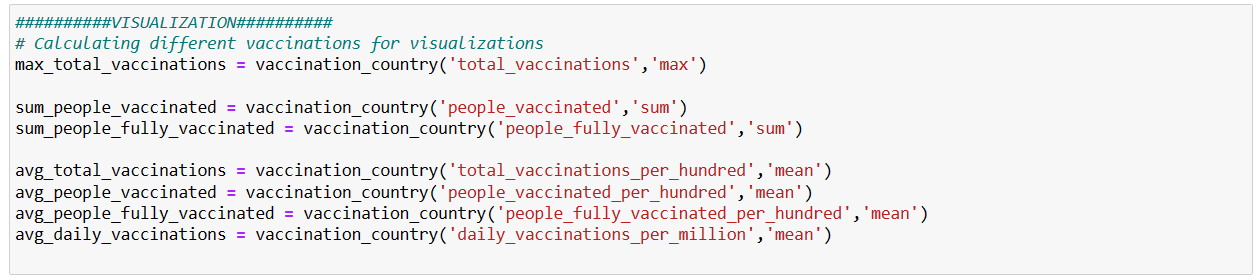
Once all necessary aspects are calculated, it now time for visualization i.e., representing the analyzed records graphically for better understanding of complex data patterns and relations.

***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.

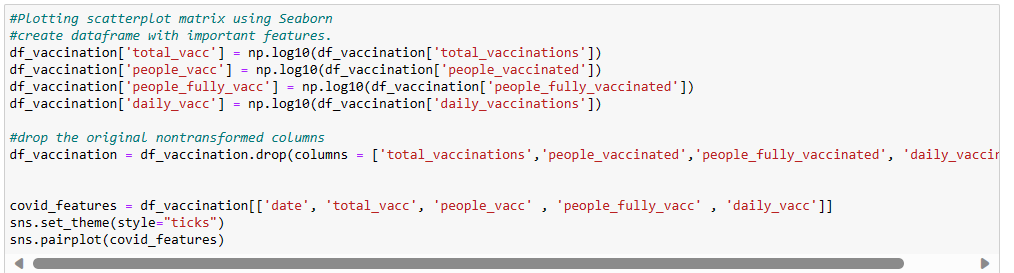
There are various techniques in data visualization. Few of them are described below,

* ***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.

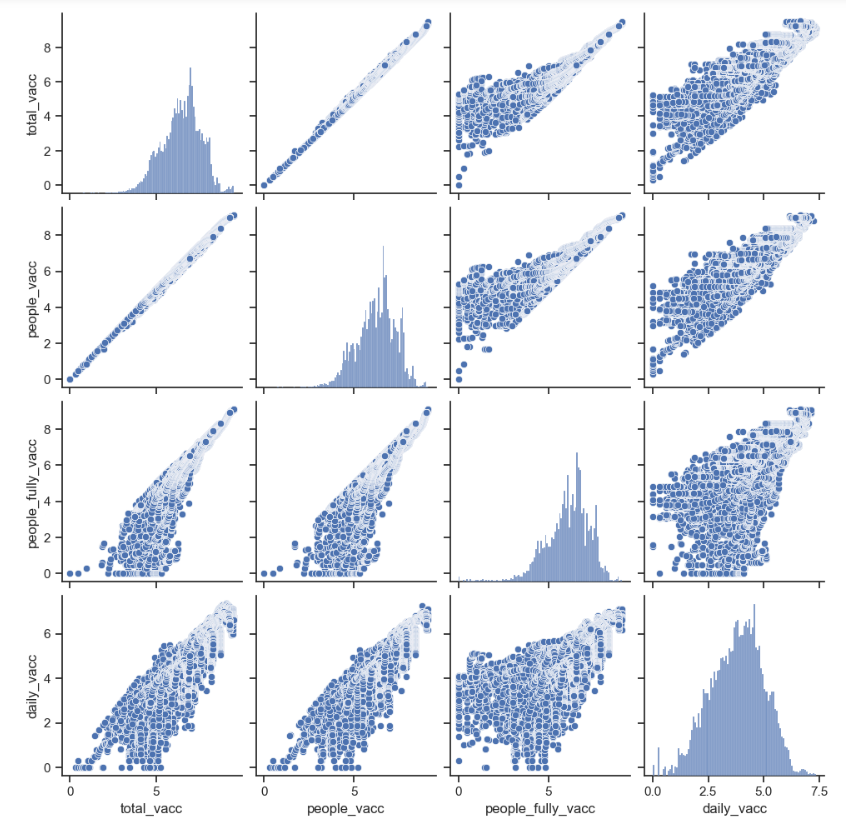
First, all required parameters are calculated using the previously created functions.



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.