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

**PROJECT PHASE 3: DEVELOPMENT PART 1**

**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 begin building the project by loading and preprocessing the dataset. During this stage, we will initiate the project by acquiring and preparing the dataset. This step marks the commencement of our analysis of COVID-19 vaccine data as we gather and preprocess the necessary information.

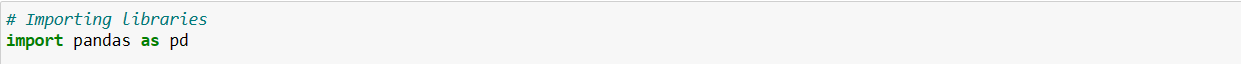
***DATA COLLECTION***

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.

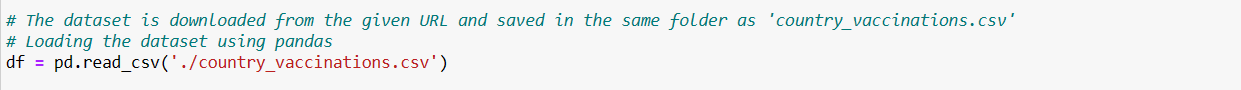
This process involves gathering a wide range of information related to vaccination efforts, including country-wise total vaccinations available, people vaccinated, adverse events, distribution logistics, and more. The source of the data for this project is,

[*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 same is explained below,

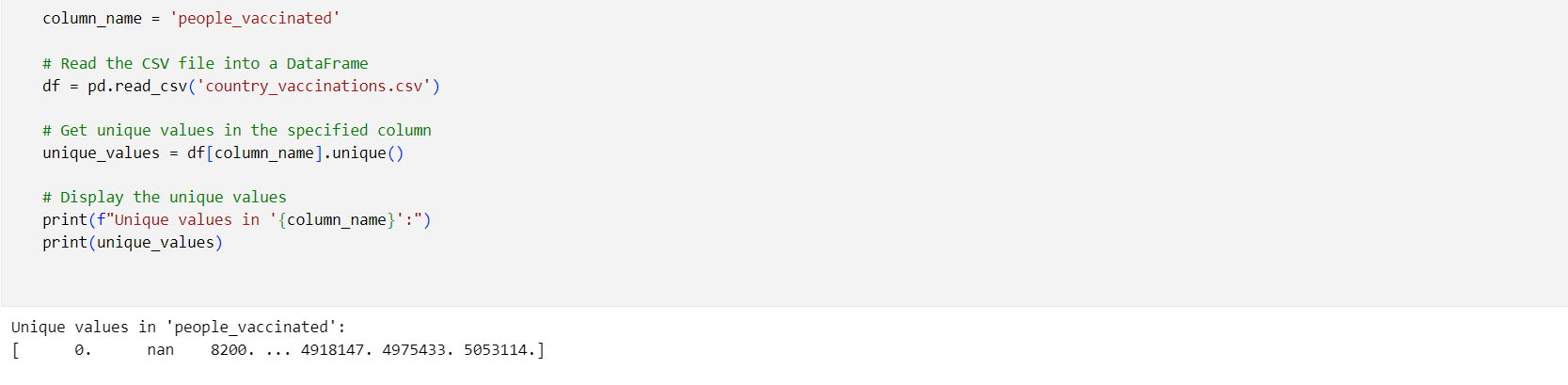
The code starts by importing necessary libraries. Here pandas library is imported. It is commonly used for data manipulation and analysis.



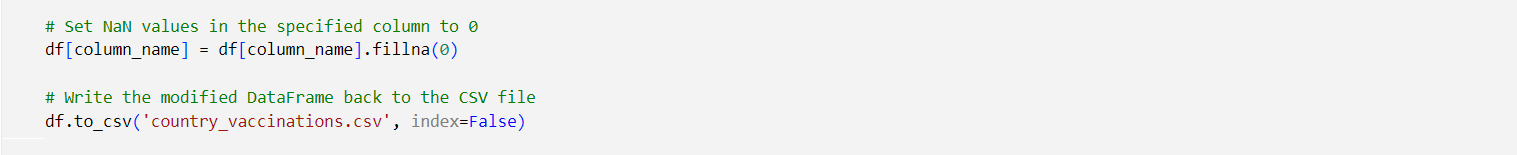
Then, the dataset is downloaded from the provided URL and saved as 'covid\_vaccine\_data.csv'.

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

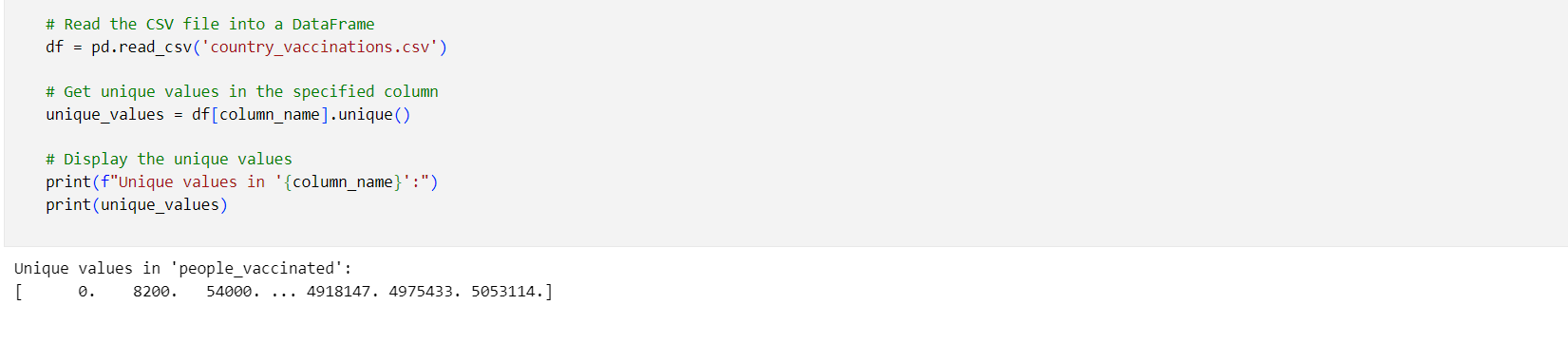
Next comes **DATA CLEANING** where the below steps are performed,



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.)

***CONCLUSION:***

At the end of this phase, the data has been collected, cleaned and made consistent. Now, it is ready for analysis that is to be performed in the upcoming phases.

***CODE:***

import pandas as pd

column\_name = 'total\_vaccinations'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'total\_vaccinations':

[0.000000e+00 nan 8.200000e+03 ... 8.845039e+06 8.934360e+06

9.039729e+06]

import pandas as pd

column\_name = 'total\_vaccinations'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'total\_vaccinations'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'total\_vaccinations':

[0.000000e+00 8.200000e+03 5.400000e+04 ... 8.845039e+06 8.934360e+06

9.039729e+06]

import pandas as pd

column\_name = 'people\_vaccinated'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_vaccinated':

[ 0. nan 8200. ... 4918147. 4975433. 5053114.]

import pandas as pd

column\_name = 'people\_vaccinated'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

import pandas as pd

column\_name = 'people\_vaccinated'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_vaccinated':

[ 0. 8200. 54000. ... 4918147. 4975433. 5053114.]

import pandas as pd

column\_name = 'people\_fully\_vaccinated'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_fully\_vaccinated':

[ nan 55624. 77560. ... 3493763. 3501493. 3510256.]

import pandas as pd

column\_name = 'people\_fully\_vaccinated'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'people\_fully\_vaccinated'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_fully\_vaccinated':

[ 0. 55624. 77560. ... 3493763. 3501493. 3510256.]

import pandas as pd

column\_name = 'daily\_vaccinations\_raw'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'daily\_vaccinations\_raw':

[ nan 2859. 4015. ... 100086. 89321. 105369.]

import pandas as pd

column\_name = 'daily\_vaccinations\_raw'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'daily\_vaccinations\_raw'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'daily\_vaccinations\_raw':

[ 0. 2859. 4015. ... 100086. 89321. 105369.]

import pandas as pd

column\_name = 'daily\_vaccinations'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'daily\_vaccinations':

[ nan 1367. 1580. ... 90629. 100614. 103751.]

import pandas as pd

column\_name = 'daily\_vaccinations'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'daily\_vaccinations'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'daily\_vaccinations':

[ 0. 1367. 1580. ... 90629. 100614. 103751.]

import pandas as pd

column\_name = 'total\_vaccinations\_per\_hundred'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'total\_vaccinations\_per\_hundred':

[0.000e+00 nan 2.000e-02 ... 5.453e+01 5.861e+01 5.990e+01]

import pandas as pd

column\_name = 'total\_vaccinations\_per\_hundred'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'total\_vaccinations\_per\_hundred'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'total\_vaccinations\_per\_hundred':

[0.000e+00 2.000e-02 1.400e-01 ... 5.453e+01 5.861e+01 5.990e+01]

import pandas as pd

column\_name = 'people\_vaccinated\_per\_hundred'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_vaccinated\_per\_hundred':

[0.000e+00 nan 2.000e-02 ... 2.162e+01 2.249e+01 2.783e+01]

import pandas as pd

column\_name = 'people\_vaccinated\_per\_hundred'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'people\_vaccinated\_per\_hundred'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_vaccinated\_per\_hundred':

[0.000e+00 2.000e-02 1.400e-01 ... 2.162e+01 2.249e+01 2.783e+01]

import pandas as pd

column\_name = 'people\_fully\_vaccinated\_per\_hundred'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_fully\_vaccinated\_per\_hundred':

[ nan 0.14 0.19 ... 17.69 19.62 23.26]

import pandas as pd

column\_name = 'people\_fully\_vaccinated\_per\_hundred'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'people\_fully\_vaccinated\_per\_hundred'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'people\_fully\_vaccinated\_per\_hundred':

[ 0. 0.14 0.19 ... 17.69 19.62 23.26]

import pandas as pd

column\_name = 'daily\_vaccinations\_per\_million'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'daily\_vaccinations\_per\_million':

[ nan 34. 40. ... 22445. 18298. 14062.]

import pandas as pd

column\_name = 'daily\_vaccinations\_per\_million'

df = pd.read\_csv('country\_vaccinations.csv')

# Set NaN values in the specified column to 0

df[column\_name] = df[column\_name].fillna(0)

# Write the modified DataFrame back to the CSV file

df.to\_csv('country\_vaccinations.csv', index=False)

import pandas as pd

column\_name = 'daily\_vaccinations\_per\_million'

# Read the CSV file into a DataFrame

df = pd.read\_csv('country\_vaccinations.csv')

# Get unique values in the specified column

unique\_values = df[column\_name].unique()

# Display the unique values

print(f"Unique values in '{column\_name}':")

print(unique\_values)

Unique values in 'daily\_vaccinations\_per\_million':

[ 0. 34. 40. ... 22445. 18298. 14062.]