Covid-19 Vaccine analysis

17/10/2023

IBM

NAAN MUDHALVAN

PHASE-3

STEP-1 SETUP JUPITER NOTEBOOK, MYSQL, SPREATSHEET:

- For installing jupiter notebook, by running the following command in command prompt “pip install notebook”.
- For installing SQL, by browsing mysql.com and download the latest version of MYSQL workbench.
- For Verifying the intallation complition by running the following code.

```
mysql --version

notebook --version
```

STEP-2 PREPROCESSING THE DATA:

- Data Preprocessing: On the other hand, is the vital step where raw vaccine-related data is transformed and cleaned. This process encompasses data cleaning, transformation, and feature engineering to ensure that the data is suitable for further analysis
- Missing Data: Another common issue that we face in real-world data is the absence of data points. Most machine learning models can’t handle missing values in the data, so you need to intervene and adjust the data to be properly used inside the model.
- For Given dataset

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_
0	Afghanistan	AFG	2021-02-22	0.0	0.0	NaN	NaN	NaN
1	Afghanistan	AFG	2021-02-23	NaN	NaN	NaN	NaN	1367
2	Afghanistan	AFG	2021-02-24	NaN	NaN	NaN	NaN	1367
3	Afghanistan	AFG	2021-02-25	NaN	NaN	NaN	NaN	1367
4	Afghanistan	AFG	2021-02-26	NaN	NaN	NaN	NaN	1367
...	...	...	...	...	...	...	...	...
86507	Zimbabwe	ZWE	2022-03-25	8691642.0	4814582.0	3473523.0	139213.0	6957
86508	Zimbabwe	ZWE	2022-03-26	8791728.0	4886242.0	3487962.0	100086.0	8342
86509	Zimbabwe	ZWE	2022-03-27	8845039.0	4918147.0	3493763.0	53311.0	9062
86510	Zimbabwe	ZWE	2022-03-28	8934360.0	4975433.0	3501493.0	89321.0	1006
86511	Zimbabwe	ZWE	2022-03-29	9039729.0	5053114.0	3510256.0	105369.0	1037

- Loading the dataset: Loading the dataset using machine learning is the process of bringing the data into the machine learning environment so that it can be used to train and evaluate a model.

- Identify the dataset: The first step is to identify the dataset that you want to load. This dataset may be stored in a local file, in a database, or in a cloud storage service. Load the Dataset: Load your dataset into a Pandas DataFrame. The quality and reliability of data can significantly impact the outcomes of vaccine analysis, making it imperative to have robust data loading procedures in place.
- Program:
- vaccine_df = pd.read_csv('/content/drive/MyDrive/country_vaccinations.csv')
- vaccine_df
- Exploring data: Perform EDA to understand your data better. This includes checking for missing values, exploring the data's statistics, and visualizing it to identify patterns.
- Program: vaccine_df.isnull().sum() # Check for missing values.

Output

```
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
```

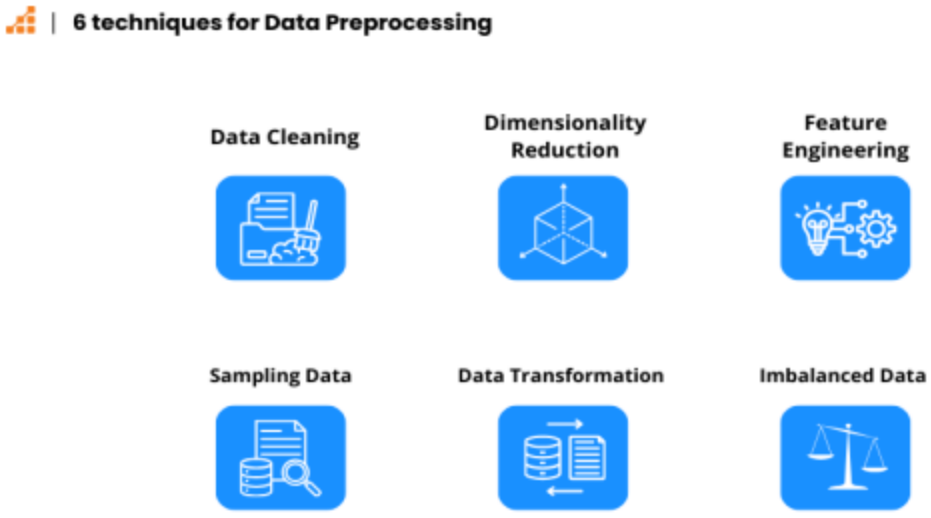
Explore statistics

vaccine_df.describe()

Output

	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_va
count	4.360700e+04	4.129400e+04	3.880200e+04	3.536200e+04	8.6213
mean	4.592964e+07	1.770508e+07	1.413830e+07	2.705996e+05	1.3130
std	2.246004e+08	7.078731e+07	5.713920e+07	1.212427e+06	7.6823
min	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.0000
25%	5.264100e+05	3.494642e+05	2.439622e+05	4.668000e+03	9.0000
50%	3.590096e+06	2.187310e+06	1.722140e+06	2.530900e+04	7.3430
75%	1.701230e+07	9.152520e+06	7.559870e+06	1.234925e+05	4.4098
max	3.263129e+09	1.275541e+09	1.240777e+09	2.474100e+07	2.2424

- Once the dataset is loaded into the machine learning environment, you may need to preprocess it before you can start training and evaluating your model. This may involve cleaning the data, transforming the data into a suitable format.

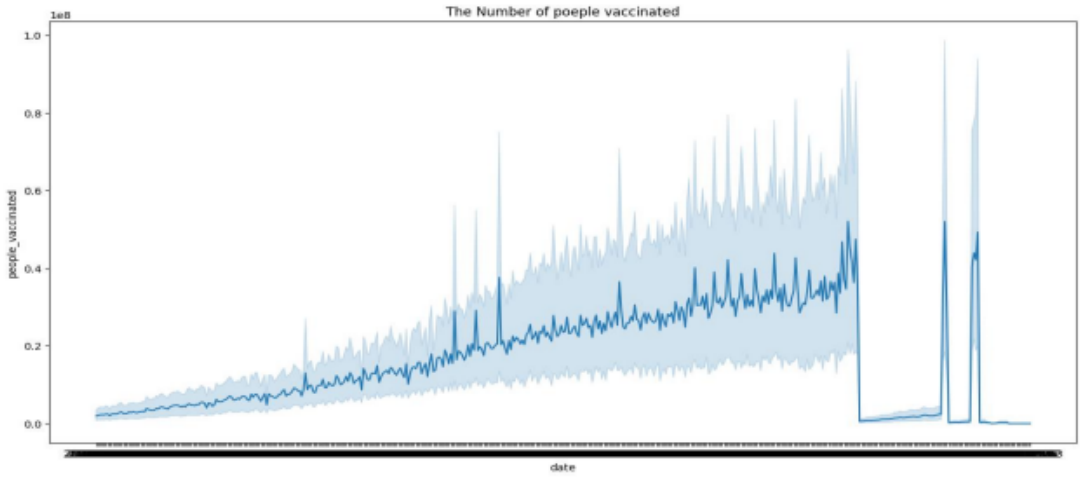


STEP-3 DATA CLEANING:

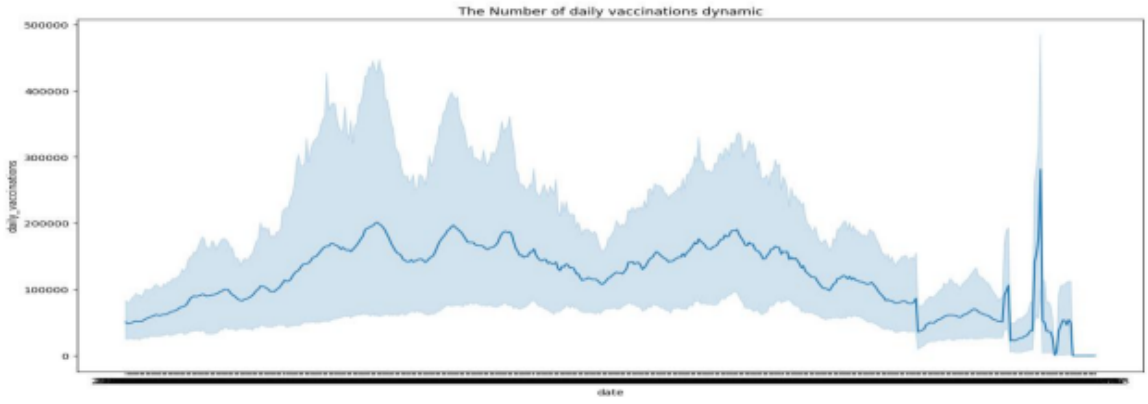
- This involves identifying and correcting errors and inconsistencies in the data. For example, this may involve removing duplicate records, correcting typos, and filling in missing values.
- Feature Scaling: Normalize or standardize numerical features to bring them to a common scale. Common methods include Min-Max scaling (scaling features to a specific range) and z-score normalization (scaling features to have a mean of 0 and a standard deviation of 1).
- Feature Engineering: Create new features or modify existing ones to capture more meaningful information from the data. This may involve mathematical transformations, interaction terms, or aggregations.
- Data transformation: It is a critical aspect of data preprocessing that involves converting and modifying the data to make it more suitable for analysis. It can help improve the performance of machine learning models, enhance the interpretability of the data, and ensure that it aligns with the assumptions of certain statistical techniques.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(16,8))
sns.lineplot(x=vaccine_df.date, y=vaccine_df.people_vaccinated)
plt.title('The Number of poeple vaccinated')
plt.show()
```

Output

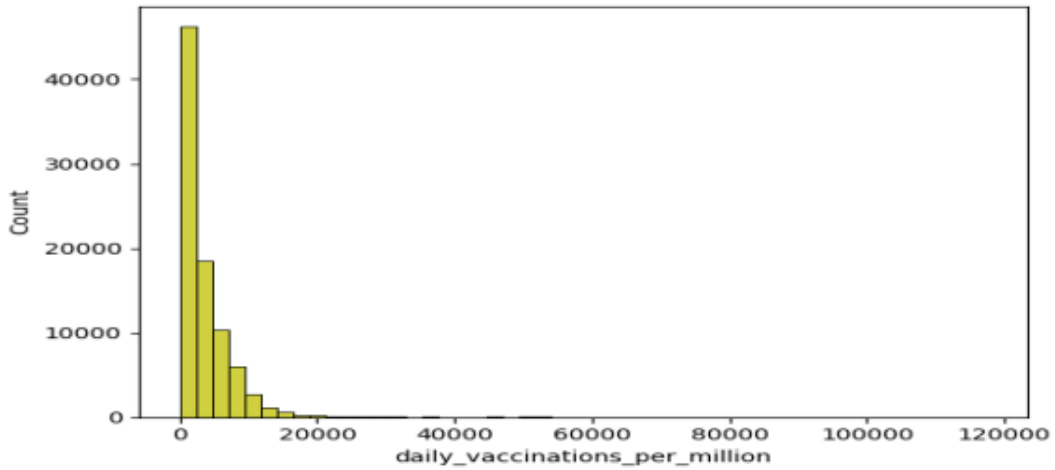


```
plt.figure(figsize=(16,8))
sns.lineplot(x=vaccine_df.date, y=vaccine_df.daily_vaccinations)
plt.title('The Number of daily vaccinations dynamic')
plt.show()
```



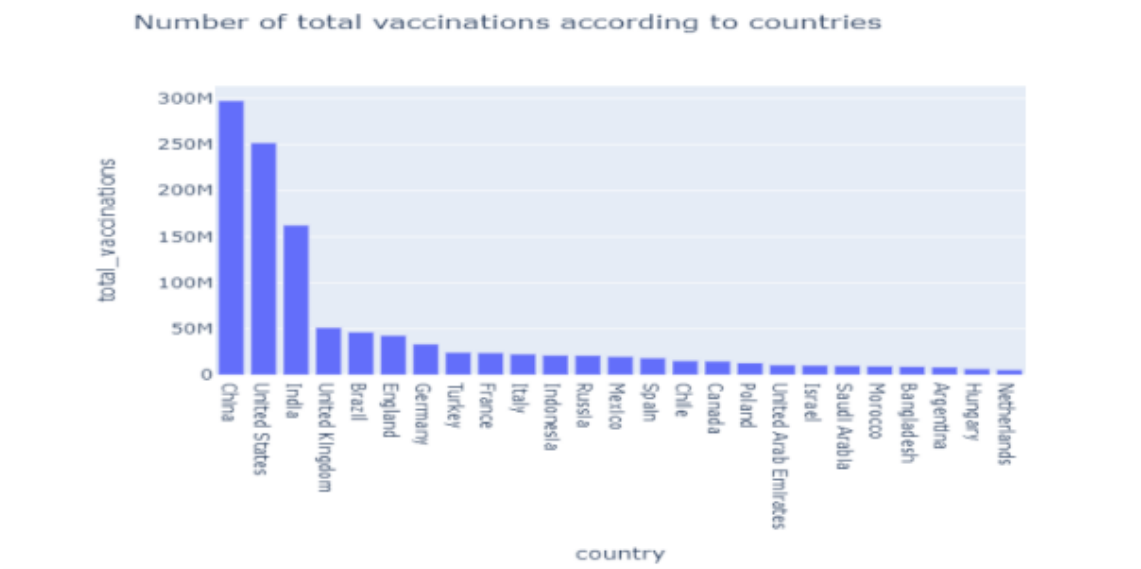

```
sns.histplot(vaccine_df, x='daily_vaccinations_per_million', bins=50, color='y')
```

Output




```
data = vaccine_df[['country','total_vaccinations']].nlargest(25,'total_vaccinations')
fig = px.bar(data, x = 'country',y = 'total_vaccinations',title="Number of total vaccinations according to countries",)
fig.show()
```

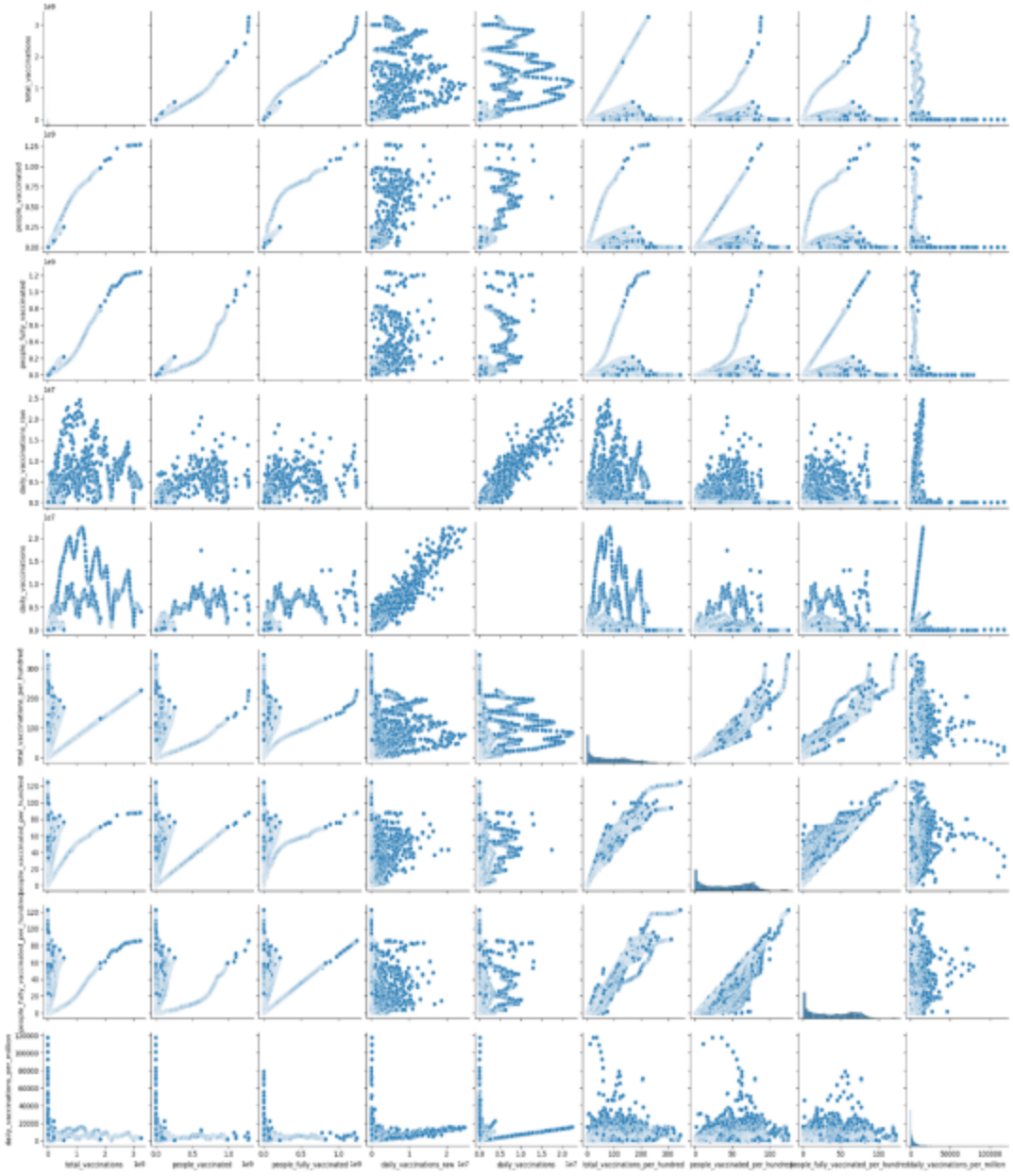
Output



plt.figure(figsize=(12,8))

sns.pairplot(vaccine_df)

OUTPUT:



vaccine_df.corr(numeric_only=True)

Output

index	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vaccinations	total_vaccinations_per_hundred
total_vaccinations	1.0	0.983438280596498	0.989681338130297	0.654711788190803	0.688501888912114	0.172297100046722
people_vaccinated	0.983438280596498	1.0	0.957599480057860	0.755540225078969	0.833433297482937	0.123938039599737
people_fully_vaccinated	0.989681338130297	0.957599480057860	1.0	0.647573972663791	0.709768165406591	0.159026253335587
daily_vaccinations_raw	0.654711788190803	0.755540225078969	0.647573972663791	1.0	0.965516572583	0.029328840593833
daily_vaccinations	0.688501888912114	0.833433297482937	0.709768165406591	0.965516572583	1.0	0.042227286198858
total_vaccinations_per_hundred	0.172297100046722	0.123938039599737	0.159026253335587	0.029328840593833	0.042227286198858	1.0
people_vaccinated_per_hundred	0.184649054590190	0.157775317674897	0.186368734714912	0.042445074564014	0.062565802459895	0.965328313791
people_fully_vaccinated_per_hundred	0.142252148700157	0.101717390787256	0.158283358797382	-0.027884824018806	-0.014054826124952	0.975454833947
daily_vaccinations_per_million	0.038299146800641	0.028720142515809	0.013220269364296	0.131078103899599	0.138822261913644	0.184908845864

CONCLUSION:

Data loading and preprocessing for vaccine analysis serve as the critical initial steps that empower researchers and healthcare professionals to make informed decisions about vaccine development, distribution, and safety. The quality, accuracy, and suitability of the data at this stage are pivotal in determining the success of subsequent analyses. Through diligent and systematic data handling, we can harness the power of data-driven insights to address public health challenges and contribute to the betterment of global health