Covid - 19 Vaccine Analysis

Project Name	**Covid - 19 Vaccine Analysis**
Team ID	**8934**
Date	**10/10/2023**

Introduction:

In this Phase 3,Data preprocessing is a critical step in the analysis of COVID-19 Vaccine data, as it lays the foundation for extracting meaningful insights and patterns from the vast and diverse sources of information related to the pandemic.

This process involves collecting, cleaning, transforming, and structuring raw data to make it suitable for analysis.

Objective:

The goal of COVID-19 vaccine analysis in this Phase 3 is to prepare the raw data for analysis, modeling, and decision-making.

This involves several key goals:

- ***Data Transformation*** **(Convert the date into datetime object)**
- ***Data Reduction*** **(Drop the lot of Null value Rows)**
- ***Data Visualization*** **(View any visualization relationship)**
- ***Scalability and Efficiency*** **(Ready for further analysis)**

Conclusion:

Data preprocessing ensures that the raw information is reliable, consistent, and appropriately structured for analysis.

In [6]:
```
#import the required libraries
#import the required dataset
#view the dataset

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
%matplotlib inline
df=pd.read_csv('Documents/country_vaccinations.csv')
df.head()
```

Out[6]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_va
0	Afghanistan	AFG	2021-02-22	0.0	0.0	NaN	
1	Afghanistan	AFG	2021-02-23	NaN	NaN	NaN	
2	Afghanistan	AFG	2021-02-24	NaN	NaN	NaN	
3	Afghanistan	AFG	2021-02-25	NaN	NaN	NaN	
4	Afghanistan	AFG	2021-02-26	NaN	NaN	NaN	

In [7]:
```
#drop the null values in the datsets using drop()
df1=df.dropna()
print(df1)
```

```
          country iso_code        date  total_vaccinations  \
94    Afghanistan      AFG  2021-05-27            593313.0
101   Afghanistan      AFG  2021-06-03            630305.0
339   Afghanistan      AFG  2022-01-27           5081064.0
433       Albania      ALB  2021-02-18              3049.0
515       Albania      ALB  2021-05-11            622507.0
...           ...      ...         ...                 ...
86507    Zimbabwe      ZWE  2022-03-25           8691642.0
86508    Zimbabwe      ZWE  2022-03-26           8791728.0
86509    Zimbabwe      ZWE  2022-03-27           8845039.0
86510    Zimbabwe      ZWE  2022-03-28           8934360.0
86511    Zimbabwe      ZWE  2022-03-29           9039729.0

       people_vaccinated  people_fully_vaccinated  daily_vaccinations_raw  \
94              479574.0                 113739.0                  2859.0
101             481800.0                 148505.0                  4015.0
339            4517380.0                3868832.0                  6868.0
433               2438.0                    611.0                  1348.0
515             440921.0                 181586.0                  9548.0
...                  ...                      ...                     ...
86507          4814582.0                3473523.0                139213.0
86508          4886242.0                3487962.0                100086.0
86509          4918147.0                3493763.0                 53311.0
86510          4975433.0                3501493.0                 89321.0
86511          5053114.0                3510256.0                105369.0

       daily_vaccinations  total_vaccinations_per_hundred  \
94                  6487.0                            1.49
101                 5285.0                            1.58
339                 9802.0                           12.76
433                  254.0                            0.11
515                12160.0                           21.67
...                    ...                             ...
86507              69579.0                           57.59
86508              83429.0                           58.25
86509              90629.0                           58.61
86510             100614.0                           59.20
86511             103751.0                           59.90

       people_vaccinated_per_hundred  people_fully_vaccinated_per_hundred  \
94                              1.20                                 0.29
101                             1.21                                 0.37
339                            11.34                                 9.71
433                             0.08                                 0.02
515                            15.35                                 6.32
...                              ...                                  ...
86507                          31.90                                23.02
86508                          32.38                                23.11
86509                          32.59                                23.15
86510                          32.97                                23.20
86511                          33.48                                23.26

       daily_vaccinations_per_million  \
94                              163.0
101                             133.0
339                             246.0
433                              88.0
515                            4233.0
...                               ...
86507                          4610.0
```

```
86508                          5528.0
86509                          6005.0
86510                          6667.0
86511                          6874.0

                                                 vaccines  \
94     Johnson&Johnson, Oxford/AstraZeneca, Pfizer/Bi...
101    Johnson&Johnson, Oxford/AstraZeneca, Pfizer/Bi...
339    Johnson&Johnson, Oxford/AstraZeneca, Pfizer/Bi...
433    Oxford/AstraZeneca, Pfizer/BioNTech, Sinovac, ...
515    Oxford/AstraZeneca, Pfizer/BioNTech, Sinovac, ...
...                                                  ...
86507  Oxford/AstraZeneca, Sinopharm/Beijing, Sinovac...
86508  Oxford/AstraZeneca, Sinopharm/Beijing, Sinovac...
86509  Oxford/AstraZeneca, Sinopharm/Beijing, Sinovac...
86510  Oxford/AstraZeneca, Sinopharm/Beijing, Sinovac...
86511  Oxford/AstraZeneca, Sinopharm/Beijing, Sinovac...

                    source_name  \
94     World Health Organization
101    World Health Organization
339    World Health Organization
433           Ministry of Health
515           Ministry of Health
...                          ...
86507         Ministry of Health
86508         Ministry of Health
86509         Ministry of Health
86510         Ministry of Health
86511         Ministry of Health

                                       source_website
94                              https://covid19.who.int/
101                             https://covid19.who.int/
339                             https://covid19.who.int/
433    https://shendetesia.gov.al/vaksinimi-anticovid...
515    https://shendetesia.gov.al/vaksinimi-anticovid...
...                                                  ...
86507  https://www.arcgis.com/home/webmap/viewer.html...
86508  https://www.arcgis.com/home/webmap/viewer.html...
86509  https://www.arcgis.com/home/webmap/viewer.html...
86510  https://www.arcgis.com/home/webmap/viewer.html...
86511  https://www.arcgis.com/home/webmap/viewer.html...

[30847 rows x 15 columns]
```

In [8]:
```
#view the information of the dataset
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30847 entries, 94 to 86511
Data columns (total 15 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   country                              30847 non-null  object
 1   iso_code                             30847 non-null  object
 2   date                                 30847 non-null  object
 3   total_vaccinations                   30847 non-null  float64
 4   people_vaccinated                    30847 non-null  float64
 5   people_fully_vaccinated              30847 non-null  float64
 6   daily_vaccinations_raw               30847 non-null  float64
 7   daily_vaccinations                   30847 non-null  float64
 8   total_vaccinations_per_hundred       30847 non-null  float64
 9   people_vaccinated_per_hundred        30847 non-null  float64
 10  people_fully_vaccinated_per_hundred  30847 non-null  float64
 11  daily_vaccinations_per_million       30847 non-null  float64
 12  vaccines                             30847 non-null  object
 13  source_name                          30847 non-null  object
 14  source_website                       30847 non-null  object
dtypes: float64(9), object(6)
memory usage: 3.8+ MB
```

In [9]:
```
#view the statistical analysis the dataset
df1.describe()
```

Out[9]:

	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vaccinations_raw	daily_vacc
count	3.084700e+04	3.084700e+04	3.084700e+04	3.084700e+04	3.084
mean	3.980375e+07	2.177533e+07	1.579596e+07	2.021875e+05	1.975
std	1.451667e+08	8.053173e+07	5.898165e+07	7.041931e+05	6.400
min	3.000000e+00	3.000000e+00	1.000000e+00	0.000000e+00	0.000
25%	1.153332e+06	7.339795e+05	3.704450e+05	5.498000e+03	7.329
50%	6.335305e+06	3.688092e+06	2.211035e+06	2.908100e+04	3.247
75%	2.520629e+07	1.440668e+07	9.121526e+06	1.344580e+05	1.402
max	3.243599e+09	1.275541e+09	1.240777e+09	1.862727e+07	1.307

In [10]:
```
#view the columns count
df.isnull().sum()
```

Out[10]:
```
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
```

In [11]:
```
#view the columns in the dataset
df.columns
```

Out[11]:
```
Index(['country', 'iso_code', 'date', 'total_vaccinations',
       'people_vaccinated', 'people_fully_vaccinated',
       'daily_vaccinations_raw', 'daily_vaccinations',
       'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred',
       'people_fully_vaccinated_per_hundred', 'daily_vaccinations_per_million',
       'vaccines', 'source_name', 'source_website'],
      dtype='object')
```

In [12]:
```
#convert the float column into integer column

df1['people_vaccinated'] = df1['people_vaccinated'].astype(int)

df1['people_fully_vaccinated'] = df1['people_fully_vaccinated'].astype(int)

df1['daily_vaccinations_raw'] = df1['daily_vaccinations_raw'].astype(int)

df1['total_vaccinations_per_hundred'] = df1['total_vaccinations_per_hundred'].astype(i

df1['people_vaccinated_per_hundred'] = df1['people_vaccinated_per_hundred'].astype(int

df1['people_fully_vaccinated_per_hundred'] = df1['people_fully_vaccinated_per_hundred'

df1['daily_vaccinations_per_million'] = df1['daily_vaccinations_per_million'].astype(i

df1.head()
```

Out[12]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily
94	Afghanistan	AFG	2021-05-27	593313.0	479574	113739	
101	Afghanistan	AFG	2021-06-03	630305.0	481800	148505	
339	Afghanistan	AFG	2022-01-27	5081064.0	4517380	3868832	
433	Albania	ALB	2021-02-18	3049.0	2438	611	
515	Albania	ALB	2021-05-11	622507.0	440921	181586	

In [13]:
```
#again check the information of dataset
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30847 entries, 94 to 86511
Data columns (total 15 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   country                              30847 non-null  object
 1   iso_code                             30847 non-null  object
 2   date                                 30847 non-null  object
 3   total_vaccinations                   30847 non-null  float64
 4   people_vaccinated                    30847 non-null  int32
 5   people_fully_vaccinated              30847 non-null  int32
 6   daily_vaccinations_raw               30847 non-null  int32
 7   daily_vaccinations                   30847 non-null  float64
 8   total_vaccinations_per_hundred       30847 non-null  int32
 9   people_vaccinated_per_hundred        30847 non-null  int32
 10  people_fully_vaccinated_per_hundred  30847 non-null  int32
 11  daily_vaccinations_per_million       30847 non-null  int32
 12  vaccines                             30847 non-null  object
 13  source_name                          30847 non-null  object
 14  source_website                       30847 non-null  object
dtypes: float64(2), int32(7), object(6)
memory usage: 2.9+ MB
```

In [14]:
```
#drop the unwanted column in dataset

df1=df1.drop(['vaccines','source_name','source_website'],axis=1)

df1
```

Out[14]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	da
94	Afghanistan	AFG	2021-05-27	593313.0	479574	113739	
101	Afghanistan	AFG	2021-06-03	630305.0	481800	148505	
339	Afghanistan	AFG	2022-01-27	5081064.0	4517380	3868832	
433	Albania	ALB	2021-02-18	3049.0	2438	611	
515	Albania	ALB	2021-05-11	622507.0	440921	181586	
...	...	...	...	...	...	...	
86507	Zimbabwe	ZWE	2022-03-25	8691642.0	4814582	3473523	
86508	Zimbabwe	ZWE	2022-03-26	8791728.0	4886242	3487962	
86509	Zimbabwe	ZWE	2022-03-27	8845039.0	4918147	3493763	
86510	Zimbabwe	ZWE	2022-03-28	8934360.0	4975433	3501493	
86511	Zimbabwe	ZWE	2022-03-29	9039729.0	5053114	3510256	

30847 rows × 12 columns

In [39]:
```
#The date is in the 'object' format. Let us change it to Datetime format for easy hana
df1['date'] =pd.to_datetime(df['date'], format='%Y-%m-%d')
```

In [15]:
```
df1
```

Out[15]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	da
94	Afghanistan	AFG	2021-05-27	593313.0	479574	113739	
101	Afghanistan	AFG	2021-06-03	630305.0	481800	148505	
339	Afghanistan	AFG	2022-01-27	5081064.0	4517380	3868832	
433	Albania	ALB	2021-02-18	3049.0	2438	611	
515	Albania	ALB	2021-05-11	622507.0	440921	181586	
...	...	...	...	...	...	...	
86507	Zimbabwe	ZWE	2022-03-25	8691642.0	4814582	3473523	
86508	Zimbabwe	ZWE	2022-03-26	8791728.0	4886242	3487962	
86509	Zimbabwe	ZWE	2022-03-27	8845039.0	4918147	3493763	
86510	Zimbabwe	ZWE	2022-03-28	8934360.0	4975433	3501493	
86511	Zimbabwe	ZWE	2022-03-29	9039729.0	5053114	3510256	

30847 rows × 12 columns

In [46]:
```
#Group by total vaccinations given by country and sort descending to identify the top
vacc_by_country = df.groupby('country').max().sort_values('total_vaccinations', ascend
vacc_by_country = vacc_by_country.iloc[:10]
vacc_by_country
```

Out[46]:

country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vacc
Afghanistan	AFG	2022-03-22	nan	5082824.0	4420127.0	
Russia	RUS	2022-03-29	nan	79954746.0	72841232.0	
Nauru	NRU	2022-03-21	nan	9150.0	7674.0	
Nepal	NPL	2022-03-29	nan	21994736.0	19014212.0	
Netherlands	NLD	2022-03-19	nan	13455761.0	12366525.0	
New Caledonia	NCL	2022-03-28	nan	188003.0	179880.0	
New Zealand	NZL	2022-03-29	nan	4284293.0	4051832.0	
Nicaragua	NIC	2022-03-25	nan	5498389.0	4113547.0	
Niger	NER	2022-03-24	nan	2180972.0	1545630.0	
Nigeria	NGA	2022-03-27	nan	21049754.0	9565143.0	

In [47]:
```
#Now sort by total vaccinations per 100
vacc_by_country = vacc_by_country.sort_values('total_vaccinations_per_hundred', ascend
vacc_by_country
```

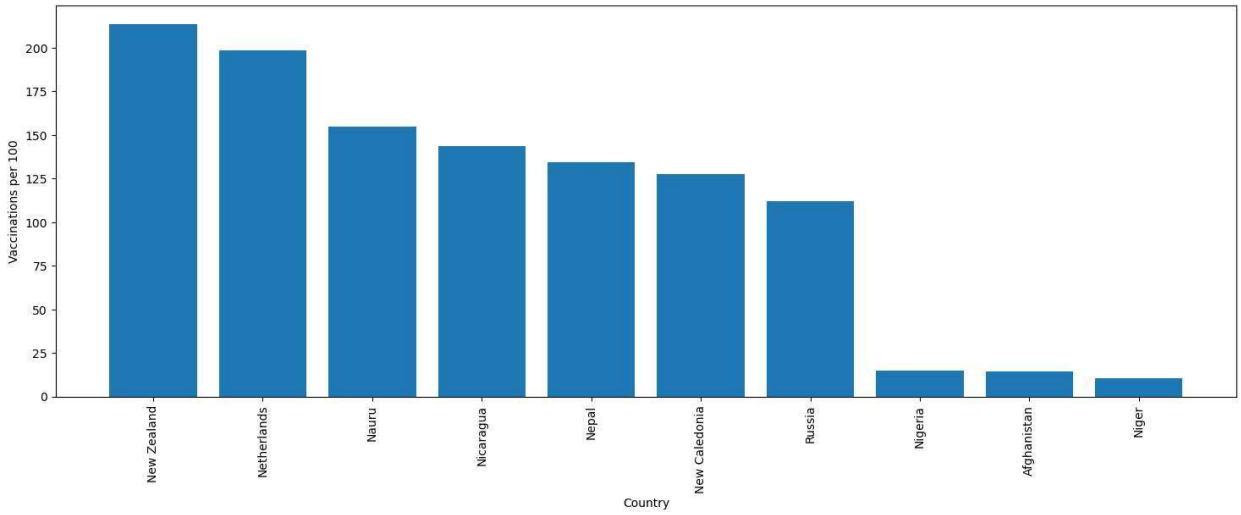
Out[47]:

country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vacc
New Zealand	NZL	2022-03-29	nan	4284293.0	4051832.0	
Netherlands	NLD	2022-03-19	nan	13455761.0	12366525.0	
Nauru	NRU	2022-03-21	nan	9150.0	7674.0	
Nicaragua	NIC	2022-03-25	nan	5498389.0	4113547.0	
Nepal	NPL	2022-03-29	nan	21994736.0	19014212.0	
New Caledonia	NCL	2022-03-28	nan	188003.0	179880.0	
Russia	RUS	2022-03-29	nan	79954746.0	72841232.0	
Nigeria	NGA	2022-03-27	nan	21049754.0	9565143.0	
Afghanistan	AFG	2022-03-22	nan	5082824.0	4420127.0	
Niger	NER	2022-03-24	nan	2180972.0	1545630.0	

In [48]:
```
#for data preprocessing step this visualization is take to further analysis

plt.figure(figsize=(18, 6))
plt.bar(vacc_by_country.index, vacc_by_country.total_vaccinations_per_hundred)

plt.xticks(rotation = 90)
plt.ylabel('Vaccinations per 100')
plt.xlabel('Country')
plt.show()
```



In [16]:
```
#this dataset is ready for further analysis
print(df1.head())
```

```
        country iso_code        date  total_vaccinations  people_vaccinated  \
94   Afghanistan      AFG  2021-05-27            593313.0              479574
101  Afghanistan      AFG  2021-06-03            630305.0              481800
339  Afghanistan      AFG  2022-01-27           5081064.0             4517380
433      Albania      ALB  2021-02-18              3049.0                2438
515      Albania      ALB  2021-05-11            622507.0              440921

     people_fully_vaccinated  daily_vaccinations_raw  daily_vaccinations  \
94                    113739                    2859              6487.0
101                   148505                    4015              5285.0
339                  3868832                    6868              9802.0
433                      611                    1348               254.0
515                   181586                    9548             12160.0

     total_vaccinations_per_hundred  people_vaccinated_per_hundred  \
94                                1                              1
101                               1                              1
339                              12                             11
433                               0                              0
515                              21                             15

     people_fully_vaccinated_per_hundred  daily_vaccinations_per_million
94                                     0                            163
101                                    0                            133
339                                    9                            246
433                                    0                             88
515                                    6                           4233
```