COVID Analysis - Data Preprocessing

Import libraries

In [1]:

```
import pandas as pd
import numpy as np
```

Loading dataset

In [2]:

```
df = pd.read_csv('country_vaccinations.csv')
df.head()
```

Out[2]:

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	c
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	

Checking null values

In [3]:

```
df.isnull().sum()
```

Out[3]:

```
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 Null values and rechecking

In [4]:

```
df = df.dropna()
df.isnull().sum()
```

Out[4]:

```
country                                0
iso_code                               0
date                                   0
total_vaccinations                     0
people_vaccinated                      0
people_fully_vaccinated                0
daily_vaccinations_raw                 0
daily_vaccinations                     0
total_vaccinations_per_hundred         0
people_vaccinated_per_hundred          0
people_fully_vaccinated_per_hundred    0
daily_vaccinations_per_million         0
vaccines                               0
source_name                            0
source_website                         0
dtype: int64
```

Adding Year Column in the dataset from Date Column.

In [5]:

```
def year(date):
    return int(date.split('-')[0])

df['year'] = df.date.apply(year)
df['year'].unique()
```

Out[5]:

```
array([2021, 2022, 2020], dtype=int64)
```

Adding Month Column in the dataset from Date Column (1-12)

In [6]:

```
def month(date):
    return int(date.split('-')[1])

df['month'] = df.date.apply(month)
df['month'].unique()
```

Out[6]:

```
array([ 5,  6,  1,  2,  7,  8,  9, 10, 11, 12,  3,  4], dtype=int64)
```

Add Day Column in the dataset from Date Column

In [7]:

```
def day(date):
    return int(date.split('-')[-1])

df['day'] = df.date.apply(day)
df['day'].unique()
```

Out[7]:

```
array([27,  3, 18, 11, 12, 13, 14, 19, 20, 21, 24, 31,  1,  7,  8,  9, 10,
       15, 16, 17, 22, 23, 29, 30,  2,  5,  6,  4, 25, 26, 28],
      dtype=int64)
```

In [8]:

```
df[['country','iso_code','date','year','month','day']].head()
```

Out[8]:

	country	iso_code	date	year	month	day
94	Afghanistan	AFG	2021-05-27	2021	5	27
101	Afghanistan	AFG	2021-06-03	2021	6	3
339	Afghanistan	AFG	2022-01-27	2022	1	27
433	Albania	ALB	2021-02-18	2021	2	18
515	Albania	ALB	2021-05-11	2021	5	11

In [9]:

```
df.to_csv('Files/country_vaccination_preprocessed.csv',index = False )
```