Name : Sara Sheth UID : 2019120058 Course : Data Analytics Lab

Objectives: Perform EDA such as number of data samples, number of features, number of classes, number of data samples per class, removing missing values, conversion to numbers, using seaborn library to plot different graphs.

Importing Liberaries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from seaborn import load_dataset
```

Theory: EDA(Exploratory Data Analysis) Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

Loading data into Dataframe

```
df = pd.read_csv('vaccination-data.csv')
```

```
import pandas as pd
import io

print(df)
```

```
                COUNTRY ISO3 WHO_REGION DATA_SOURCE DATE_UPDATED  \
0           Afghanistan  AFG       EMRO   REPORTING   2022-02-07
1               Albania  ALB       EURO   REPORTING   2022-01-30
2               Algeria  DZA       AFRO   REPORTING   2022-01-09
3        American Samoa  ASM       WPRO   REPORTING   2022-01-28
4               Andorra  AND       EURO   REPORTING   2022-01-23
..                  ...  ...        ...         ...          ...
223            Viet Nam  VNM       WPRO   REPORTING   2022-01-27
```

```
224  Wallis and Futuna  WLF       WPRO   REPORTING   2022-01-11
225              Yemen  YEM       EMRO   REPORTING   2022-02-07
226             Zambia  ZMB       AFRO   REPORTING   2021-11-18
227           Zimbabwe  ZWE       AFRO   REPORTING   2022-01-29

     TOTAL_VACCINATIONS  PERSONS_VACCINATED_1PLUS_DOSE  \
0               5216998                      4634282.0   
1               2613974                      1261272.0   
2              12974545                      7247787.0   
3                 82992                        41820.0   
4                140193                        57709.0   
..                  ...                            ...   
223           180366266                     79023934.0   
224               12287                         6151.0   
225              758480                       600559.0   
226             1041441                       832532.0   
227             7608063                      4263080.0   

     TOTAL_VACCINATIONS_PER100  PERSONS_VACCINATED_1PLUS_DOSE_PER100  \
0                       13.402                                11.905   
1                       90.800                                44.318   
2                       29.588                                16.528   
3                      150.356                                75.765   
4                      181.400                                75.756   
..                         ...                                   ...   
223                    185.298                                81.185   
224                    109.257                                54.695   
225                      2.543                                 2.014   
226                      5.665                                 4.529   
227                     51.188                                28.683   

     PERSONS_FULLY_VACCINATED  PERSONS_FULLY_VACCINATED_PER100  \
0                   3959887.0                           10.172   
1                   1127431.0                           40.377   
2                   5796432.0                           13.218   
3                     36804.0                           66.678   
4                     53046.0                           69.711   
..                        ...                              ...   
223                74011623.0                           76.035   
224                    6136.0                           54.562   
225                  358824.0                            1.203   
226                  651965.0                            3.546   
227                 3291261.0                           22.144   

                                         VACCINES_USED FIRST_VACCINE_DATE  \
0    Beijing CNBG - BBIBP-CorV,Janssen - Ad26.COV 2...         2021-02-22   
1    AstraZeneca - Vaxzevria,Gamaleya - Gam-Covid-V...         2021-01-13   
2    Beijing CNBG - BBIBP-CorV,Gamaleya - Gam-Covid...         2021-01-30   
3    Janssen - Ad26.COV 2-S,Moderna - Spikevax,Pfiz...         2020-12-21   
4    AstraZeneca - Vaxzevria,Moderna - Spikevax,Pfi...         2021-01-20   
```

displaying first 5 rows of data

```
df.head()
```

	COUNTRY	ISO3	WHO_REGION	DATA_SOURCE	DATE_UPDATED	TOTAL_VACCINATIONS	PERSONS_
0	Afghanistan	AFG	EMRO	REPORTING	2022-02-07	5216998	
1	Albania	ALB	EURO	REPORTING	2022-01-30	2613974	
2	Algeria	DZA	AFRO	REPORTING	2022-01-09	12974545	
3	American Samoa	ASM	WPRO	REPORTING	2022-01-28	82992	
4	Andorra	AND	EURO	REPORTING	2022-01-23	140193	

df.tail() displays the last 5 rows of Data

```
df.tail()
```

	COUNTRY	ISO3	WHO_REGION	DATA_SOURCE	DATE_UPDATED	TOTAL_VACCINATIONS	PERSONS
223	Viet Nam	VNM	WPRO	REPORTING	2022-01-27	180366266	
224	Wallis and [illegible]	WLF	WPRO	REPORTING	2022-01-11	12287	
225	Zambia	ZMB	AFRO	REPORTING	[illegible]	[illegible]	

Number of rows and columns present in data

```
df.shape
```

```
(228, 14)
```

Checking if there is any duplicate rows in data and removing them

```
duplicate_rows_df = df[df.duplicated()]
print(" Number of duplicate rows: ",duplicate_rows_df.shape)
```

```
 Number of duplicate rows:  (0, 14)
```

Checking the number of rows that each column contains

```
df.count()
```

```
COUNTRY                                 228
ISO3                                    228
WHO_REGION                              228
DATA_SOURCE                             228
DATE_UPDATED                            228
TOTAL_VACCINATIONS                      228
PERSONS_VACCINATED_1PLUS_DOSE           225
TOTAL_VACCINATIONS_PER100               228
PERSONS_VACCINATED_1PLUS_DOSE_PER100    225
PERSONS_FULLY_VACCINATED                225
PERSONS_FULLY_VACCINATED_PER100         225
VACCINES_USED                           225
FIRST_VACCINE_DATE                      208
NUMBER_VACCINES_TYPES_USED              225
dtype: int64
```

deleting the duplicate rows and displaying the first 5 rows of data

```
df = df.drop_duplicates()
df.head(5)
```

	COUNTRY	ISO3	WHO_REGION	DATA_SOURCE	DATE_UPDATED	TOTAL_VACCINATIONS	PERSONS_
0	Afghanistan	AFG	EMRO	REPORTING	2022-02-07	5216998	
1	Albania	ALB	EURO	REPORTING	2022-01-30	2613974	
2	Algeria	DZA	AFRO	REPORTING	2022-01-09	12974545	
3	American Samoa	ASM	WPRO	REPORTING	2022-01-28	82992	
4	Andorra	AND	EURO	REPORTING	2022-01-23	140193	

```
df.count()
```

```
COUNTRY                                 228
ISO3                                    228
WHO_REGION                              228
DATA_SOURCE                             228
DATE_UPDATED                            228
TOTAL_VACCINATIONS                      228
PERSONS_VACCINATED_1PLUS_DOSE           225
TOTAL_VACCINATIONS_PER100               228
PERSONS_VACCINATED_1PLUS_DOSE_PER100    225
PERSONS_FULLY_VACCINATED                225
PERSONS_FULLY_VACCINATED_PER100         225
VACCINES_USED                           225
FIRST_VACCINE_DATE                      208
NUMBER_VACCINES_TYPES_USED              225
dtype: int64
```

```
df['TOTAL_VACCINATIONS'].value_counts()
```

```
5216998     1
33411666    1
9369918     1
8881477     1
1798575     1
           ..
111323      1
19303999    1
79577       1
84708       1
```

```
7608063     1
Name: TOTAL_VACCINATIONS, Length: 228, dtype: int64
```

Checking for null values

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

```
COUNTRY                                  0
ISO3                                     0
WHO_REGION                               0
DATA_SOURCE                              0
DATE_UPDATED                             0
TOTAL_VACCINATIONS                       0
PERSONS_VACCINATED_1PLUS_DOSE            3
TOTAL_VACCINATIONS_PER100                0
PERSONS_VACCINATED_1PLUS_DOSE_PER100     3
PERSONS_FULLY_VACCINATED                 3
PERSONS_FULLY_VACCINATED_PER100          3
VACCINES_USED                            3
FIRST_VACCINE_DATE                      20
NUMBER_VACCINES_TYPES_USED               3
dtype: int64
```

Removing Null/Missing Values

```
df = df.dropna()
df.count()
```

```
COUNTRY                                 206
ISO3                                    206
WHO_REGION                              206
DATA_SOURCE                             206
DATE_UPDATED                            206
TOTAL_VACCINATIONS                      206
PERSONS_VACCINATED_1PLUS_DOSE           206
TOTAL_VACCINATIONS_PER100               206
PERSONS_VACCINATED_1PLUS_DOSE_PER100    206
PERSONS_FULLY_VACCINATED                206
PERSONS_FULLY_VACCINATED_PER100         206
VACCINES_USED                           206
FIRST_VACCINE_DATE                      206
NUMBER_VACCINES_TYPES_USED              206
dtype: int64
```

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

```
COUNTRY                                 0
ISO3                                    0
WHO_REGION                              0
DATA_SOURCE                             0
DATE_UPDATED                            0
TOTAL_VACCINATIONS                      0
```

```
PERSONS_VACCINATED_1PLUS_DOSE          0
TOTAL_VACCINATIONS_PER100              0
PERSONS_VACCINATED_1PLUS_DOSE_PER100   0
PERSONS_FULLY_VACCINATED               0
PERSONS_FULLY_VACCINATED_PER100        0
VACCINES_USED                          0
FIRST_VACCINE_DATE                     0
NUMBER_VACCINES_TYPES_USED             0
dtype: int64
```

To check the actual dimension of the box,we can use the .describe() method

```
df.describe()
```

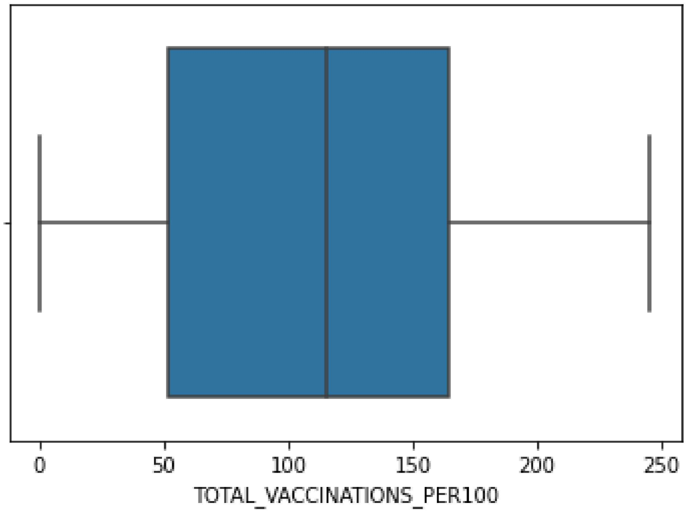
	TOTAL_VACCINATIONS	PERSONS_VACCINATED_1PLUS_DOSE	TOTAL_VACCINATIONS_PER100	PE
count	2.060000e+02	2.060000e+02	206.000000	
mean	4.720353e+07	2.260903e+07	110.768981	
std	2.447806e+08	1.131227e+08	66.928409	
min	7.400000e+01	3.700000e+01	0.081000	
25%	4.513935e+05	2.772375e+05	51.969500	
50%	2.777090e+06	1.719060e+06	115.291500	
75%	1.569372e+07	7.674768e+06	164.329250	
max	3.009902e+09	1.275814e+09	245.275000	

```
sns.boxplot(x=df['TOTAL_VACCINATIONS'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff67d9acd50>
```

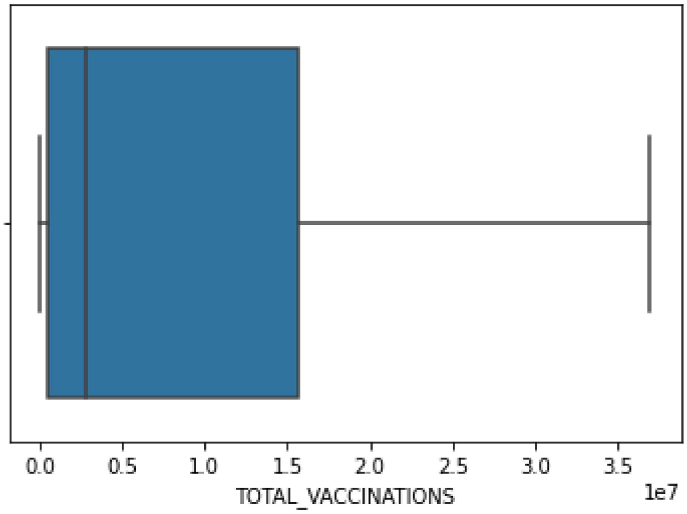
```
sns.boxplot(x=df['TOTAL_VACCINATIONS_PER100'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff67d874710>
```



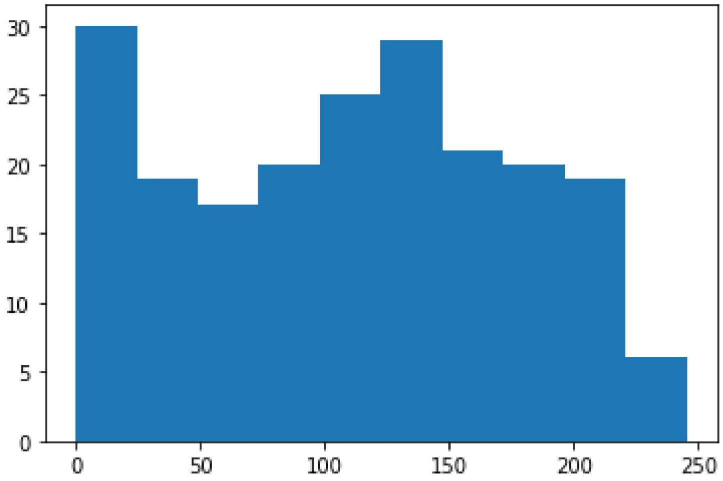
```
sns.boxplot(x=df['TOTAL_VACCINATIONS'], showfliers = False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff67d3ad090>
```

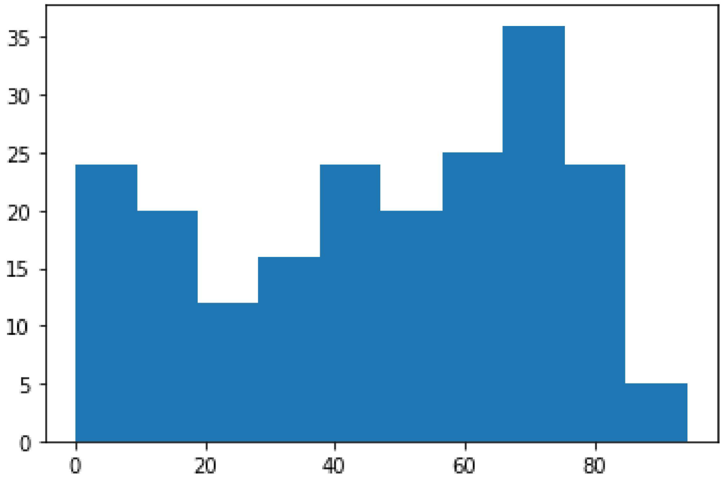


```
plt.hist(df['TOTAL_VACCINATIONS'])
plt.show()
```

```
plt.hist(df['TOTAL_VACCINATIONS_PER100'])
plt.show()
```

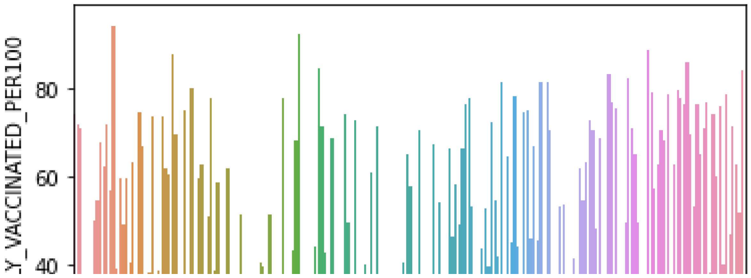


```
plt.hist(df['PERSONS_FULLY_VACCINATED_PER100'])
plt.show()
```



```
sns.barplot(df['TOTAL_VACCINATIONS'],df['PERSONS_FULLY_VACCINATED_PER100'])
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass tl
  FutureWarning
```

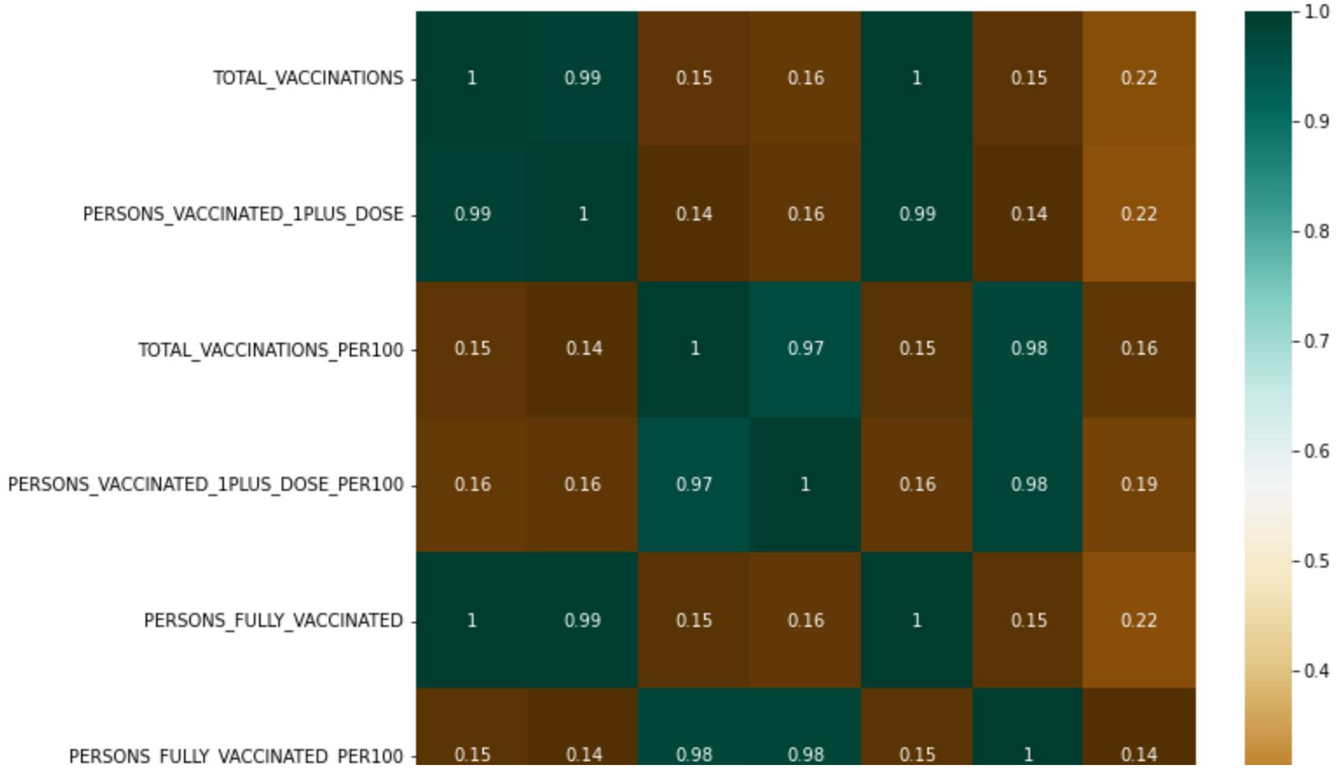


```
print(df.isnull().sum())
```

```
COUNTRY                                 0
ISO3                                    0
WHO_REGION                              0
DATA_SOURCE                             0
DATE_UPDATED                            0
TOTAL_VACCINATIONS                      0
PERSONS_VACCINATED_1PLUS_DOSE           0
TOTAL_VACCINATIONS_PER100               0
PERSONS_VACCINATED_1PLUS_DOSE_PER100    0
PERSONS_FULLY_VACCINATED                0
PERSONS_FULLY_VACCINATED_PER100         0
VACCINES_USED                           0
FIRST_VACCINE_DATE                      0
NUMBER_VACCINES_TYPES_USED              0
dtype: int64
```

```
plt.figure(figsize=(10,10))
c=df.corr()
sns.heatmap(c,cmap="BrBG",annot = True)
c
```

	TOTAL_VACCINATIONS	PERSONS_VACCINATED_1PL
TOTAL_VACCINATIONS	1.000000	C
PERSONS_VACCINATED_1PLUS_DOSE	0.991250	1
TOTAL_VACCINATIONS_PER100	0.150696	C
PERSONS_VACCINATED_1PLUS_DOSE_PER100	0.161909	C
PERSONS_FULLY_VACCINATED	0.999605	C
PERSONS_FULLY_VACCINATED_PER100	0.149195	C
NUMBER_VACCINES_TYPES_USED	0.216776	C



conclusion: In this experiment I learnt how to perform Exploratory Data Analysis with the help of different python libraries such as pandas, seaborn, matplotlib,etc. by the performed EDA we can conclude that persons fully vaccinated per 100 people is very low.

TOTAL_VACCINATIONS
PERSONS_VACCINATED_1PLUS_DOSE
TOTAL_VACCINATIONS_PER100
PERSONS_VACCINATED_1PLUS_DOSE_PER100
PERSONS_FULLY_VACCINATED
PERSONS_FULLY_VACCINATED_PER100
NUMBER_VACCINES_TYPES_USED