ML Lab 1: Exploratory Data Analysis (EDA)

Objective:

Import the dataset and 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, explore dimensionality, type the mean or average value, and using seaborn library to plot different graphs. Consider one of the datasets given below.

Code:

```
from google.colab import files
data = files.upload()
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import io
import pandas as pd
from matplotlib.pyplot import figure as fig
df = pd.read_csv(io.StringIO(data['who_data.csv'].decode('utf-8')))
Date_reported = df.Date_reported
Country = df.Country
New_cases = df.New_cases
Cumulative_cases = df.Cumulative_cases
New_deaths = df.New_deaths
Cumulative_deaths = df.Cumulative_deaths

Date_reported = np.array(Date_reported)
Country = np.array(Country)
New_cases = np.array(New_cases)
Cumulative_cases = np.array(Cumulative_cases)
New_deaths = df.New_deaths
Cumulative_deaths = df.Cumulative_deaths
ndf = df.copy()
df_india = ndf[(ndf["Country"] == "India")]
df_usa = ndf[(ndf["Country"] == "United States of America")]
df_italy = ndf[(ndf["Country"] == "Italy")]
df_china = ndf[(ndf["Country"] == "China")]
```

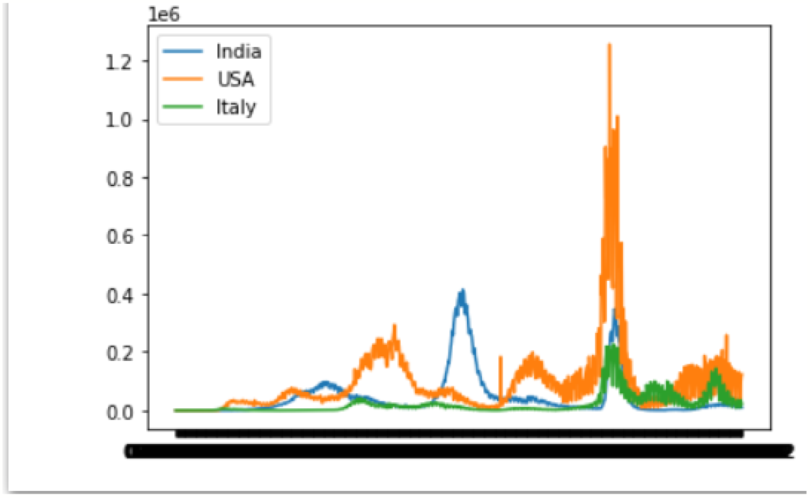
```
IDate_reported = df_india.Date_reported
ICountry = df_india.Country
INew_cases = df_india.New_cases
ICumulative_cases = df_india.Cumulative_cases
INew_deaths = df_india.New_deaths
ICumulative_deaths = df_india.Cumulative_deaths

UDate_reported = df_usa.Date_reported
UCountry = df_usa.Country
UNew_cases = df_usa.New_cases
UCumulative_cases = df_usa.Cumulative_cases
UNew_deaths = df_usa.New_deaths
UCumulative_deaths = df_usa.Cumulative_deaths

YDate_reported = df_italy.Date_reported
YCountry = df_italy.Country
YNew_cases = df_italy.New_cases
YCumulative_cases = df_italy.Cumulative_cases
YNew_deaths = df_italy.New_deaths
YCumulative_deaths = df_italy.Cumulative_deaths

plt.plot(Date_reported,INew_cases)
plt.plot(Date_reported,UNew_cases)
plt.plot(Date_reported,YNew_cases)
plt.legend(["India", "USA", "Italy"], loc ="upper left")
```

Plot output:



Comments on code flow:

- Import the csv file which consists of data regarding the new cases, cumulative cases and cumulative deaths due to COVID19 complications in each country on each day since January 2020
- Using pandas library, read the csv file using UTF-8 decoding
- Convert the columns of date_modified, country, new_cases, cumulative_cases, new_deaths and cumulative_deaths into numPy arrays
- Create objects only considering specific countries
- Plot date_modified vs new_cases using matplotlib library

Inference:

- By the plot we can observe different waves of COVID19 pandemic in different countries