**MALARIA DATASET**

J.JANE KARUNIYA

1832025

**PROBLEM STATEMENT:**

To Explore whether the no. of cases, deaths, CFR … of malaria increases every year. From scratch.

**DATA DESCRIPTION:**

• Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female Anopheles mosquitoes.

• It is preventable and curable.

• In 2018, there were an estimated 228 million cases of malaria worldwide.

• The estimated number of malaria deaths stood at 405 000 in 2018.

• Children aged under 5 years are the most vulnerable group affected by malaria;

• in 2018, they accounted for 67% (272 000) of all malaria deaths worldwide

• The WHO African Region carries a disproportionately high share of the global malaria burden.

• In 2018, the region was home to 93% of malaria cases and 94% of malaria deaths.

**Content of dataset:**

• reported\_numbers.csv - Reported no. of cases across the world

• estimated\_numbers.csv - Estimated no of cases across the world

• incidenceper1000popat\_risk.csv - Incidence per 1000 people at risk area

**Data Pre-processing:**

Before data can be used as input for machine learning algorithms, it must be cleaned, formatted, and maybe even restructured — this is typically known as preprocessing. Unfortunately, for this dataset, there are many invalid or missing entries(?) we must deal with, moreover, there are some qualities about certain features that must be adjusted. This preprocessing can help tremendously with the outcome and predictive power of nearly all learning algorithms.

import numpy as np

import pandas as pd

from time import time

import seaborn as sns

import matplotlib.pyplot as plt

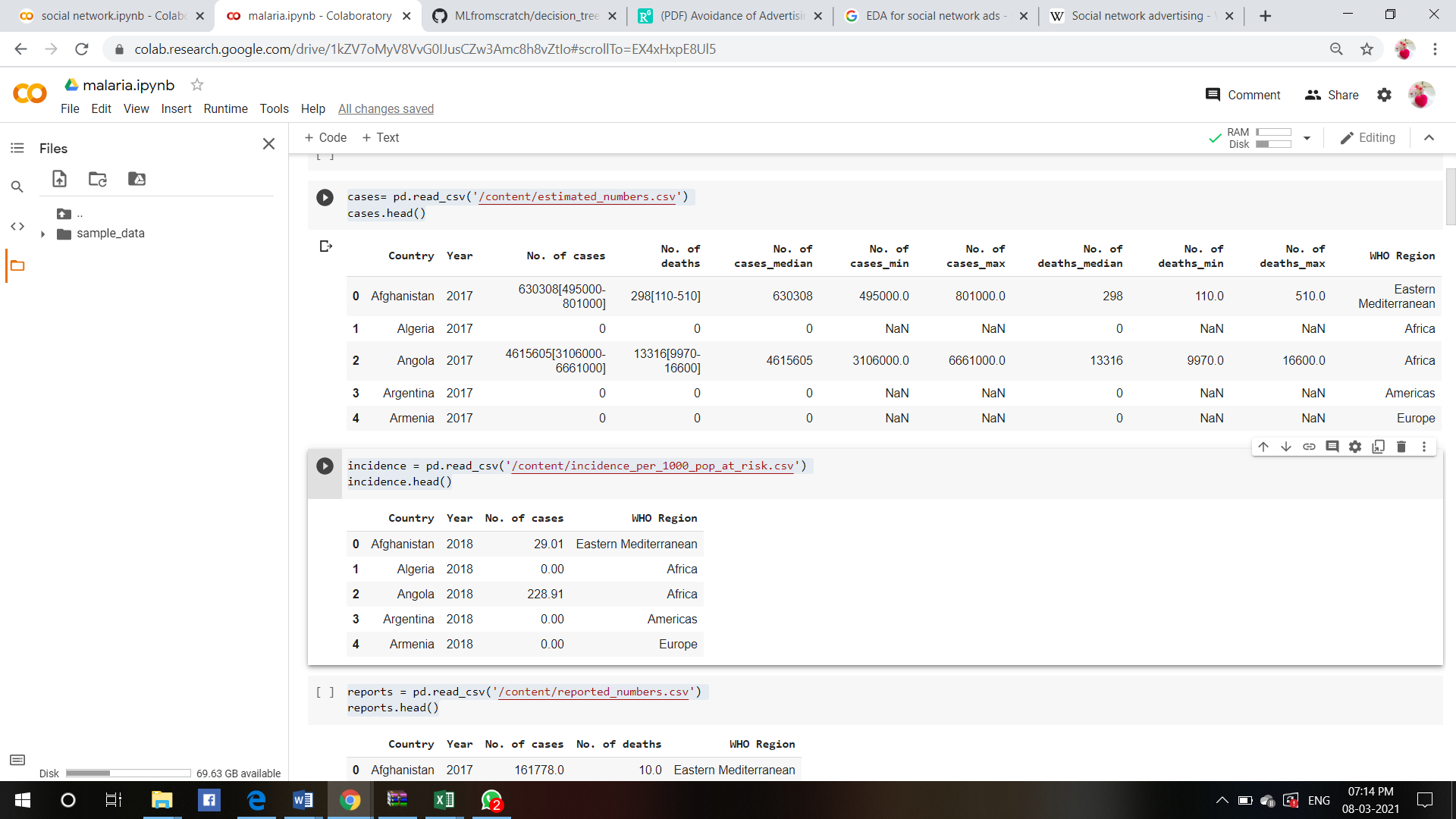
%matplotlib inline

import os

import warnings

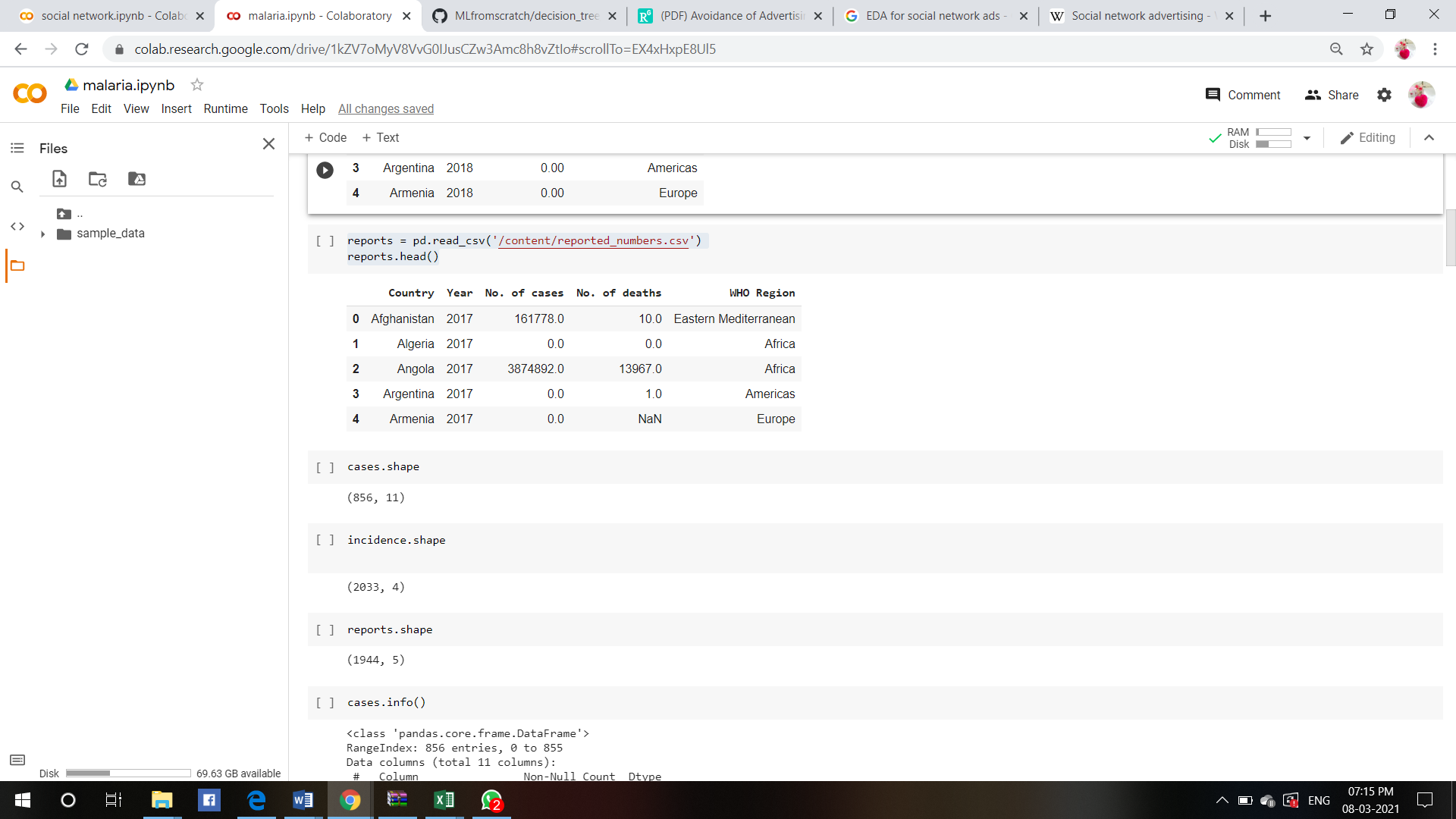
cases= pd.read\_csv('/content/estimated\_numbers.csv')

cases.head()



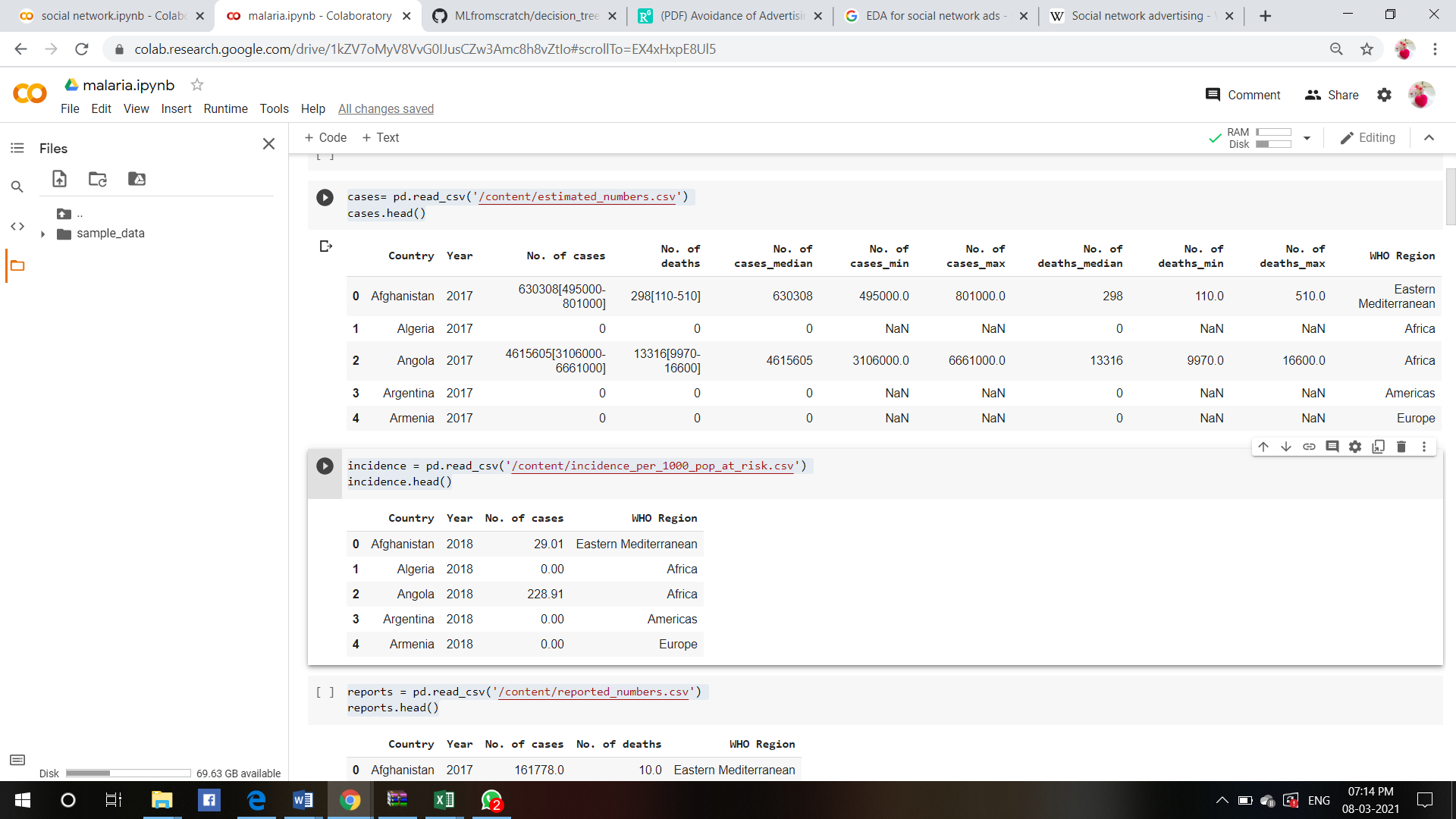
reports = pd.read\_csv('/content/reported\_numbers.csv')

reports.head()



incidence = pd.read\_csv('/content/incidence\_per\_1000\_pop\_at\_risk.csv')

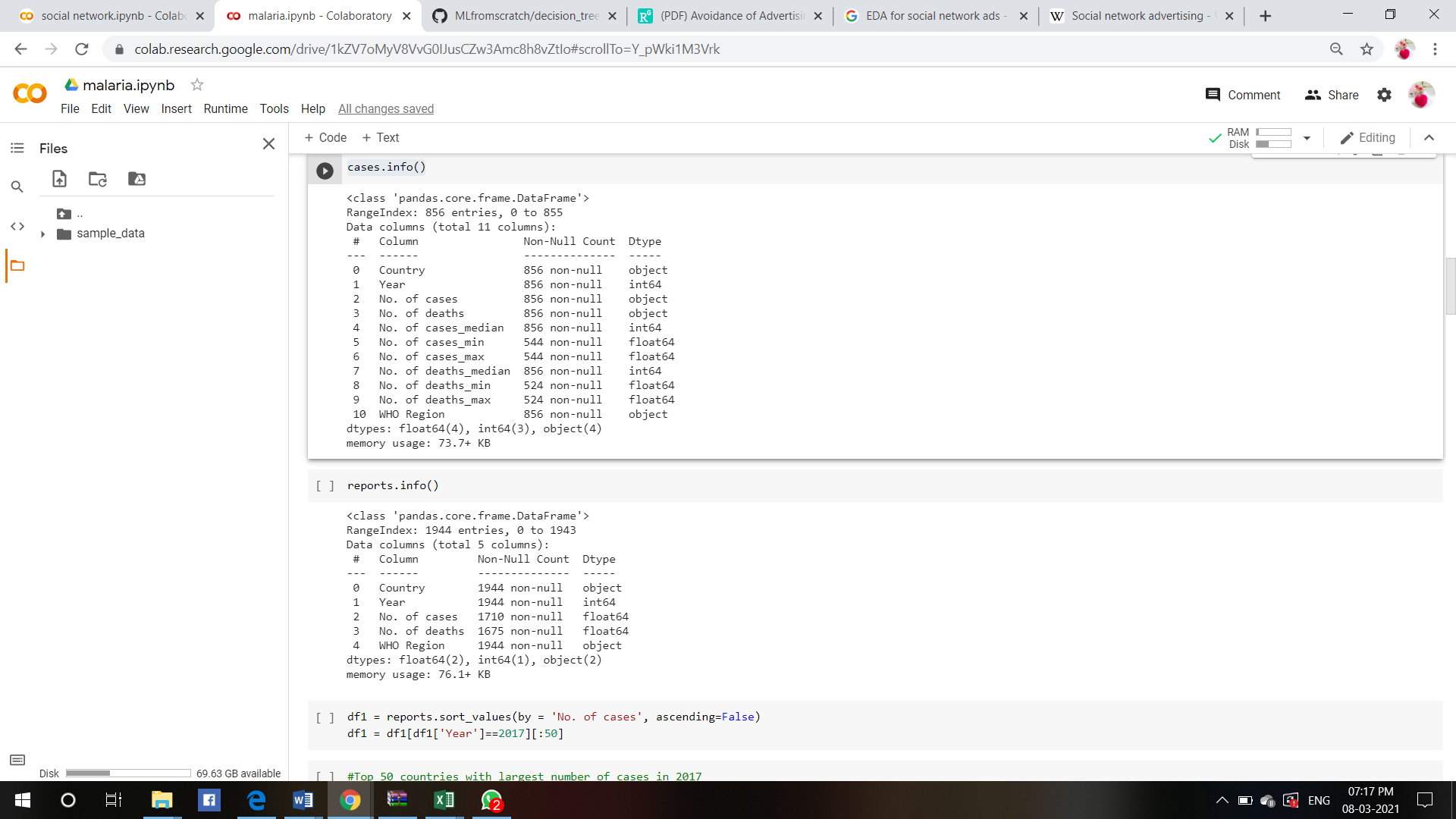
incidence.head()



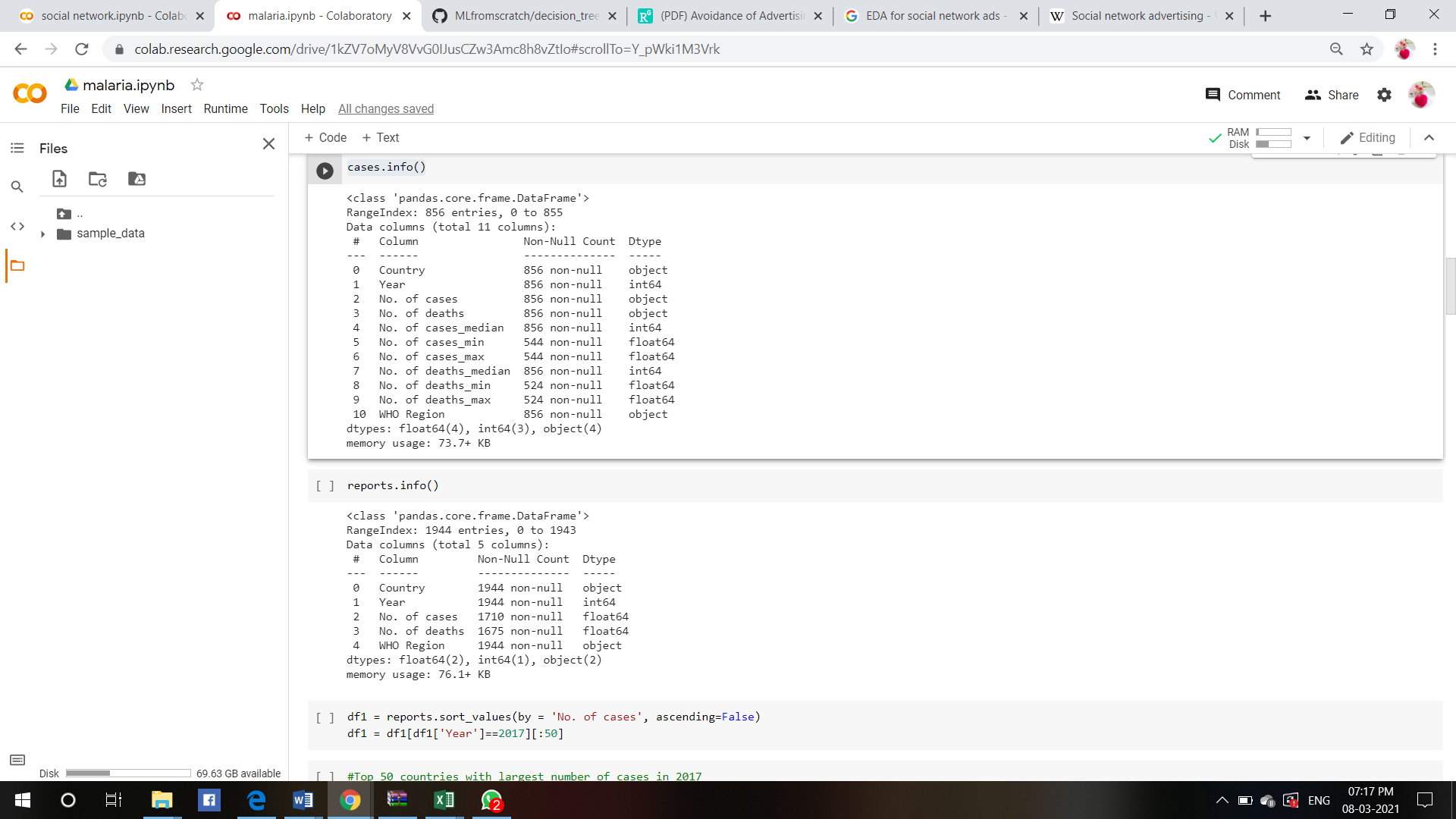
**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an approach for analysing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task. For better understanding of our chronic dataset well, we will also be performing EDAs in such a way that we can clearly see the dependencies and weightage of values clearly. After understanding the variables, it will be easy for us to implement them in the model. First let us explore the important variables that influence the accuracy of the model.

cases.info()



reports.info()

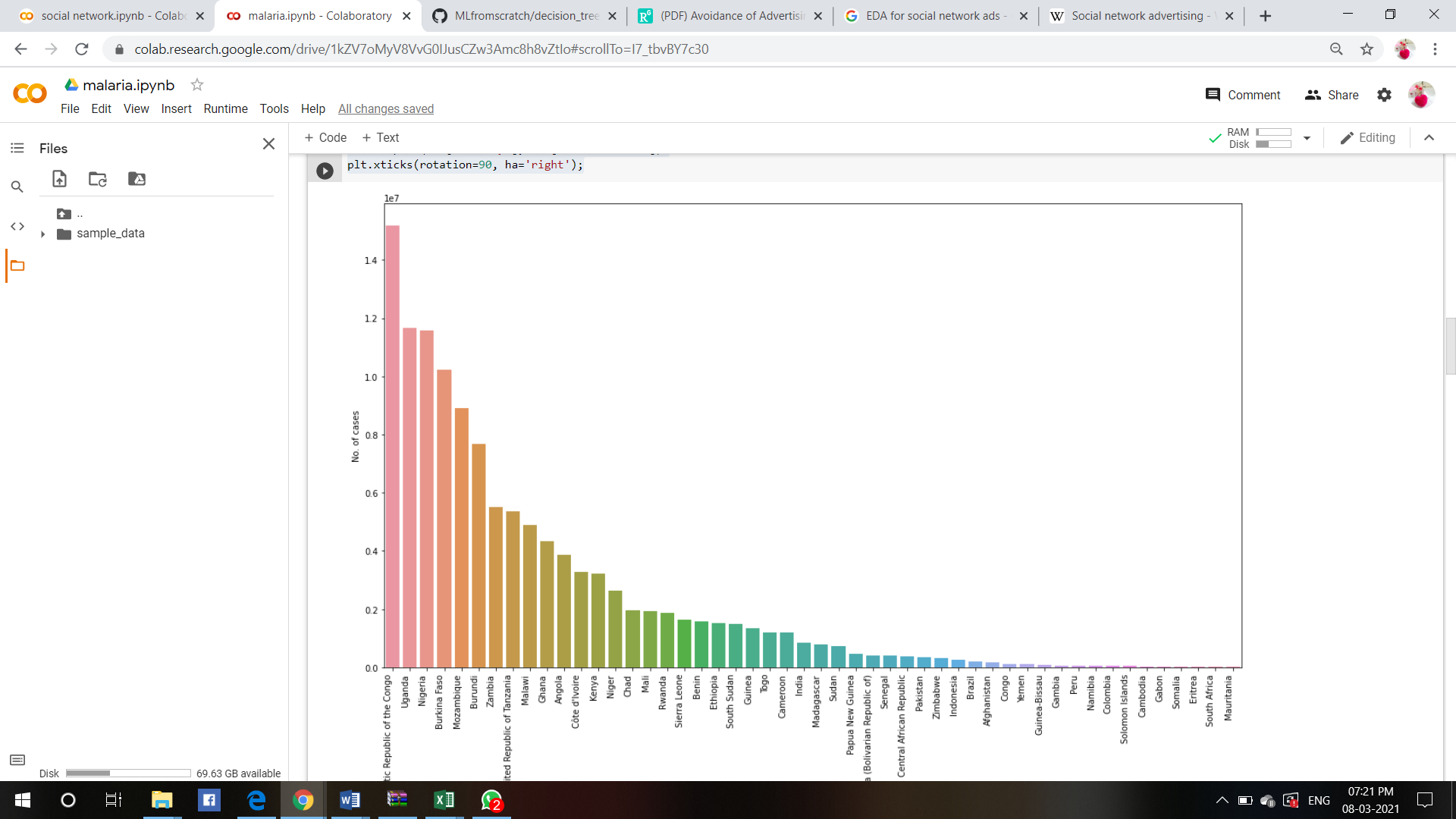


#Top 50 countries with largest number of cases in 2017

plt.figure(figsize=(18,10))

sns.barplot(df1['Country'], df1['No. of cases'])

plt.xticks(rotation=90, ha='right');

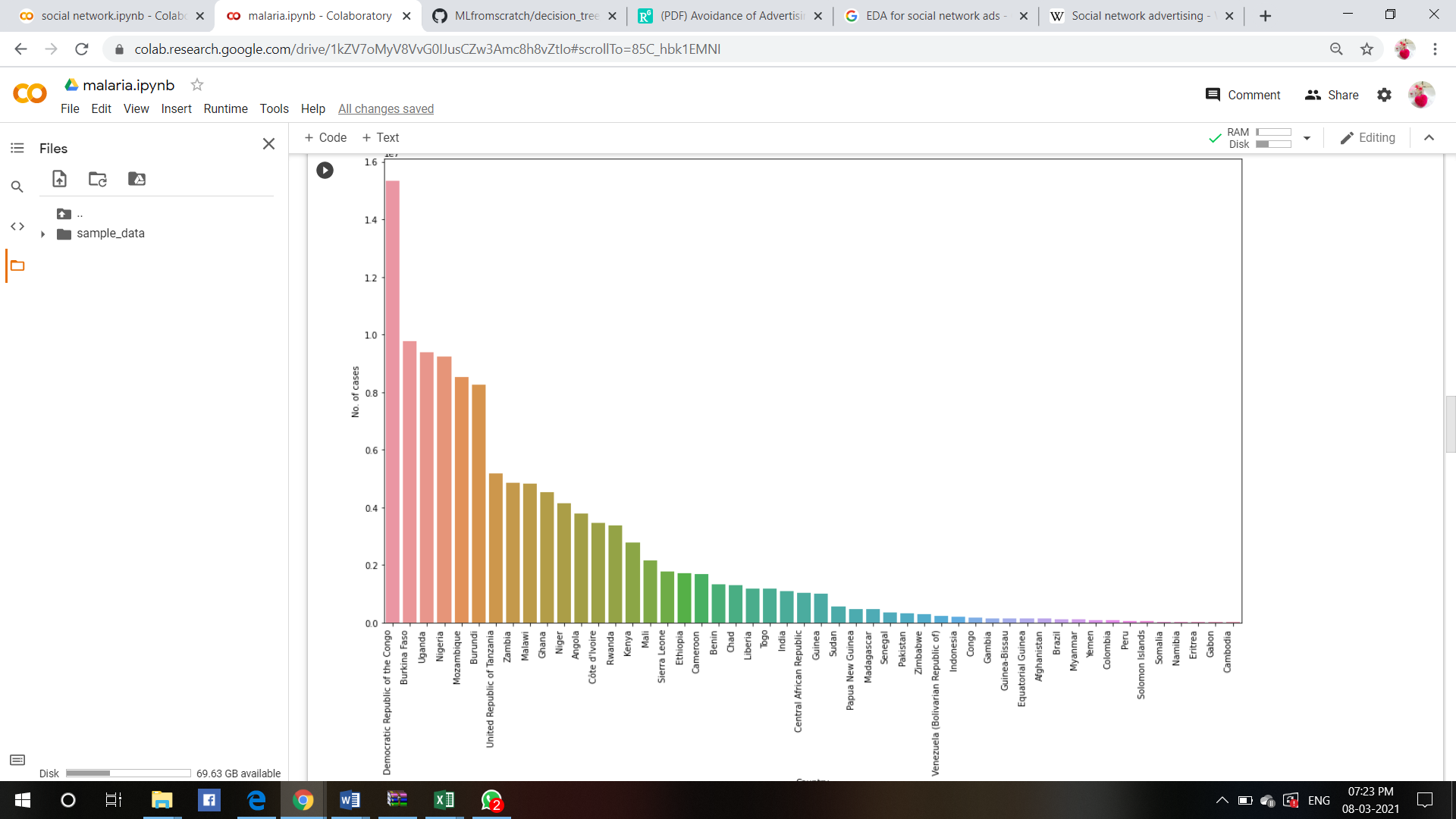


#Top 50 countries with largest number of cases in 2016

plt.figure(figsize=(18,10))

sns.barplot(df2['Country'], df2['No. of cases'])

plt.xticks(rotation=90, ha='right');

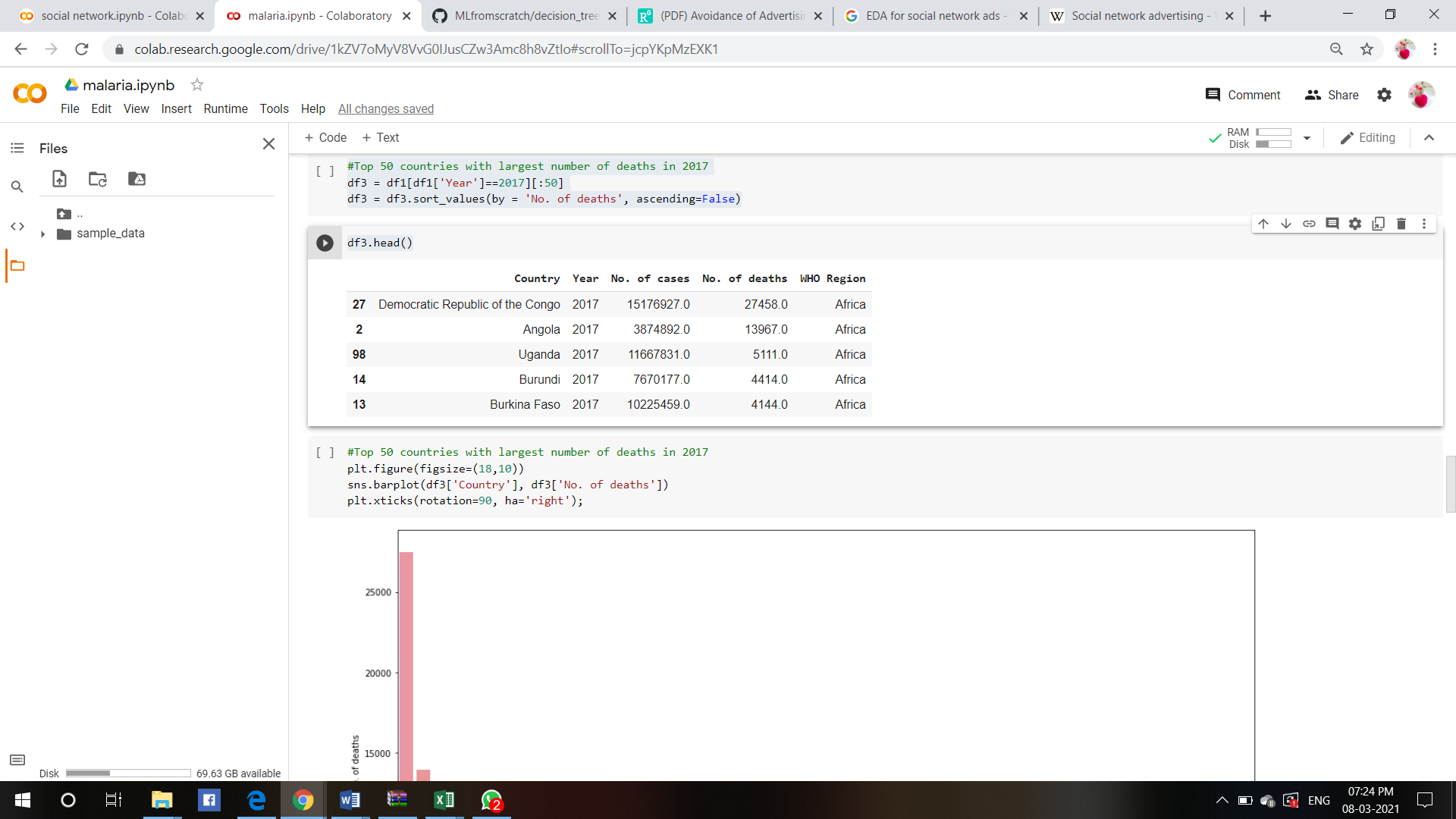


#Top 50 countries with largest number of deaths in 2017

df3 = df1[df1['Year']==2017][:50]

df3 = df3.sort\_values(by = 'No. of deaths', ascending=False)

df3.head()

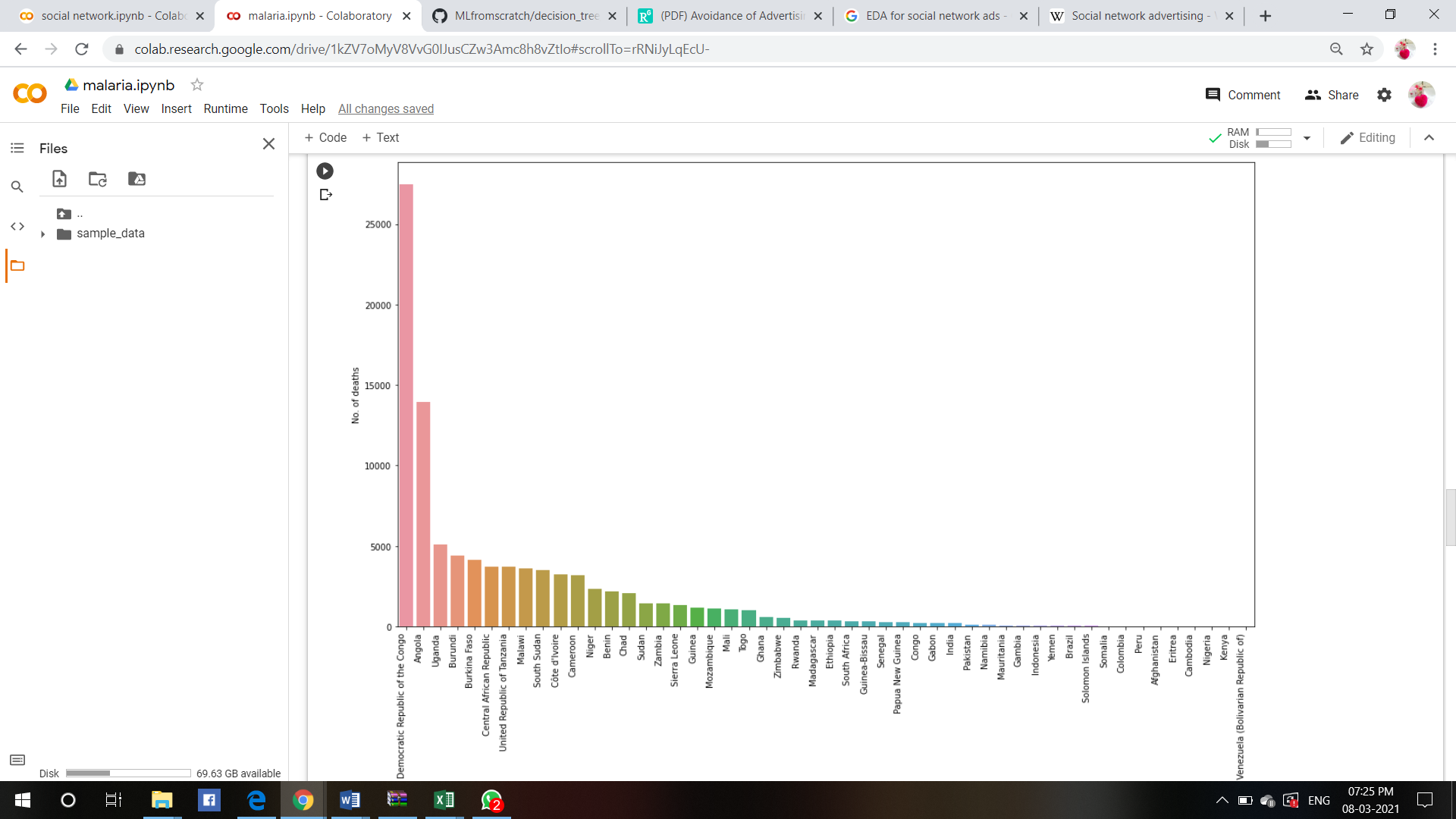


#Top 50 countries with largest number of deaths in 2017

plt.figure(figsize=(18,10))

sns.barplot(df3['Country'], df3['No. of deaths'])

plt.xticks(rotation=90, ha='right');

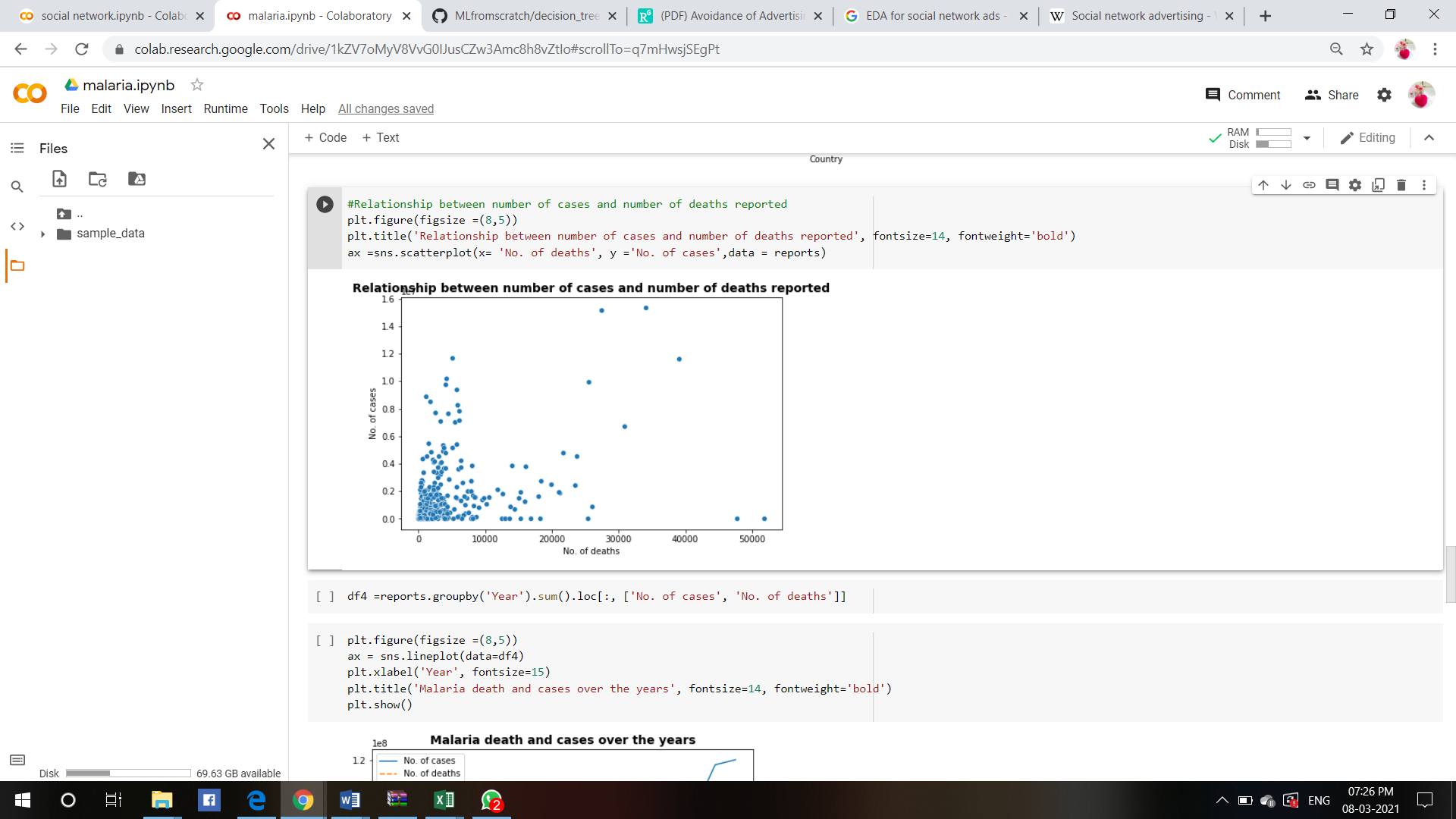


#Relationship between number of cases and number of deaths reported

plt.figure(figsize =(8,5))

plt.title('Relationship between number of cases and number of deaths reported', fontsize=14, fontweight='bold')

ax =sns.scatterplot(x= 'No. of deaths', y ='No. of cases',data = reports)



df4 =reports.groupby('Year').sum().loc[:, ['No. of cases', 'No. of deaths']]

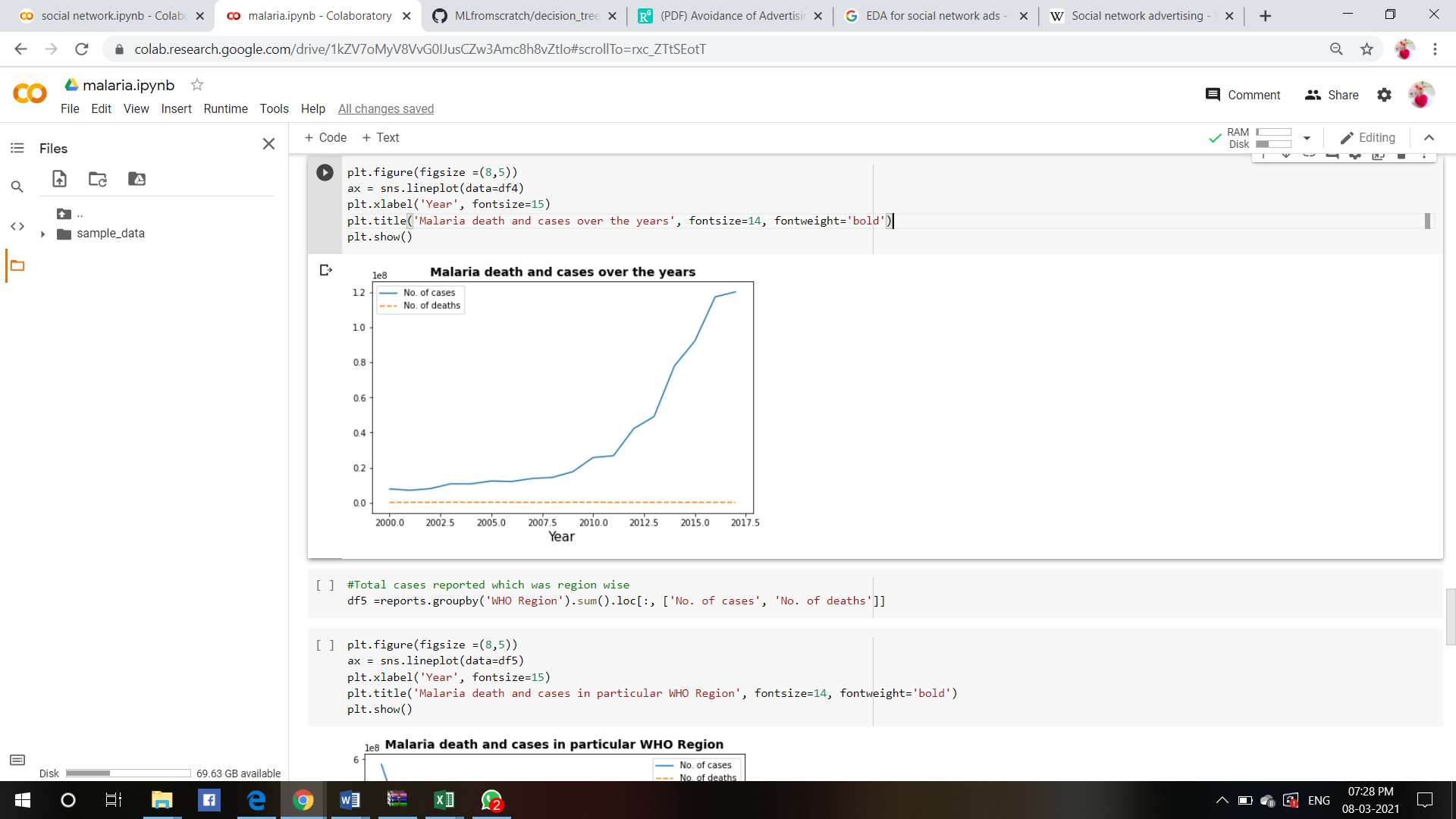
plt.figure(figsize =(8,5))

ax = sns.lineplot(data=df4)

plt.xlabel('Year', fontsize=15)

plt.title('Malaria death and cases over the years', fontsize=14, fontweight='bold')

plt.show()



#Total cases reported which was region wise

df5 =reports.groupby('WHO Region').sum().loc[:, ['No. of cases', 'No. of deaths']]

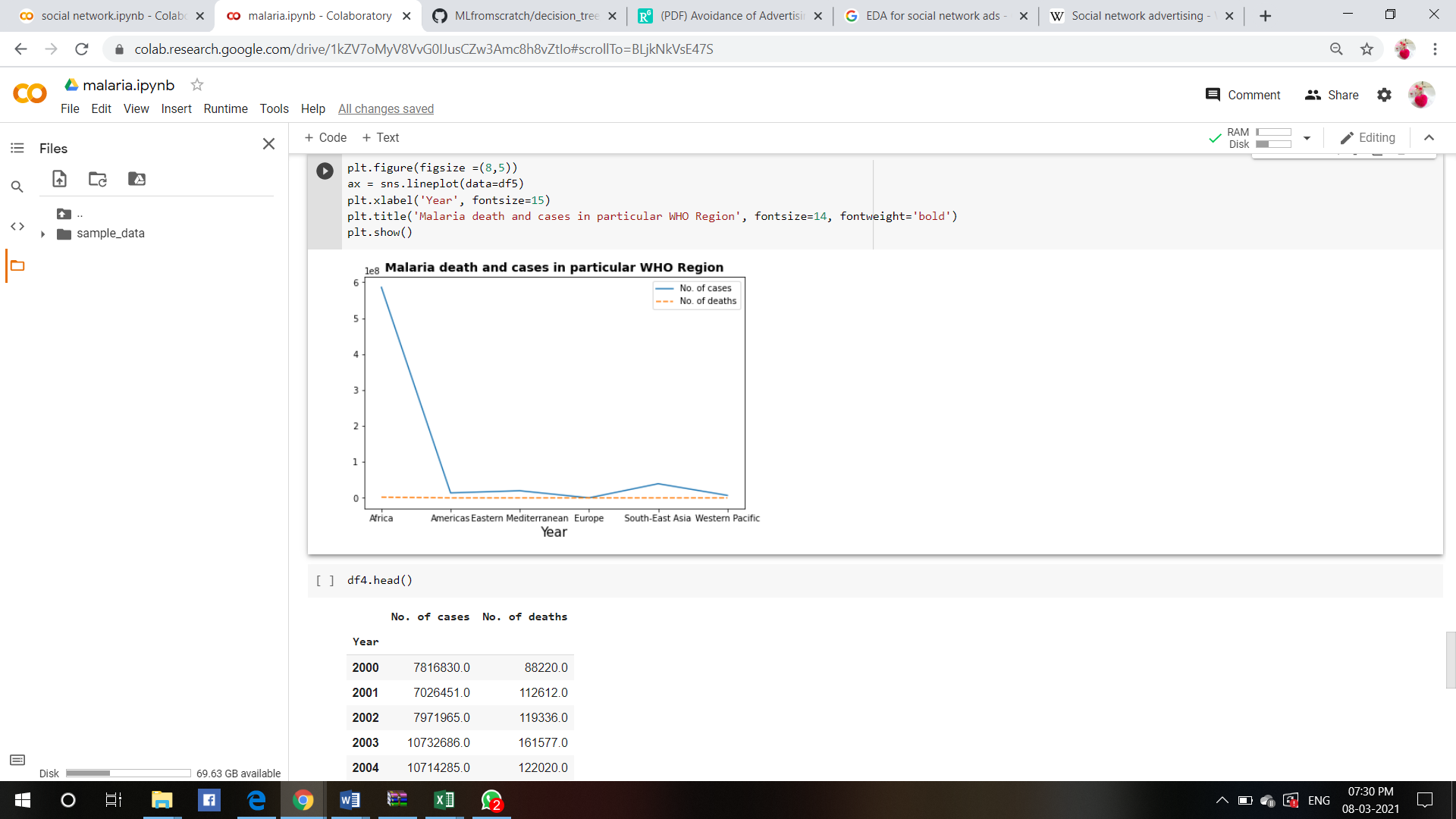
plt.figure(figsize =(8,5))

ax = sns.lineplot(data=df5)

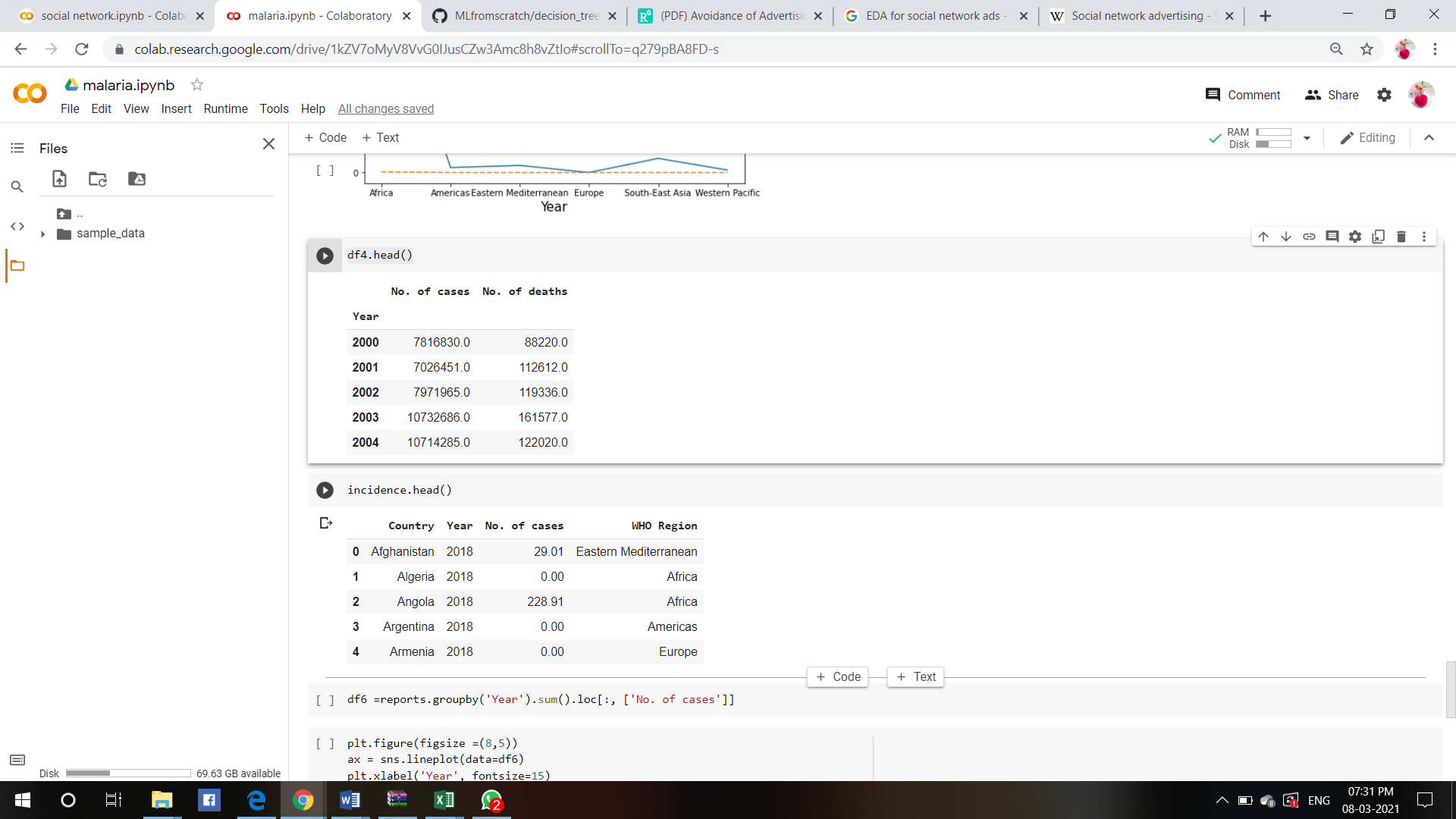
plt.xlabel('Year', fontsize=15)

plt.title('Malaria death and cases in particular WHO Region', fontsize=14, fontweight='bold')

plt.show()



df4.head()



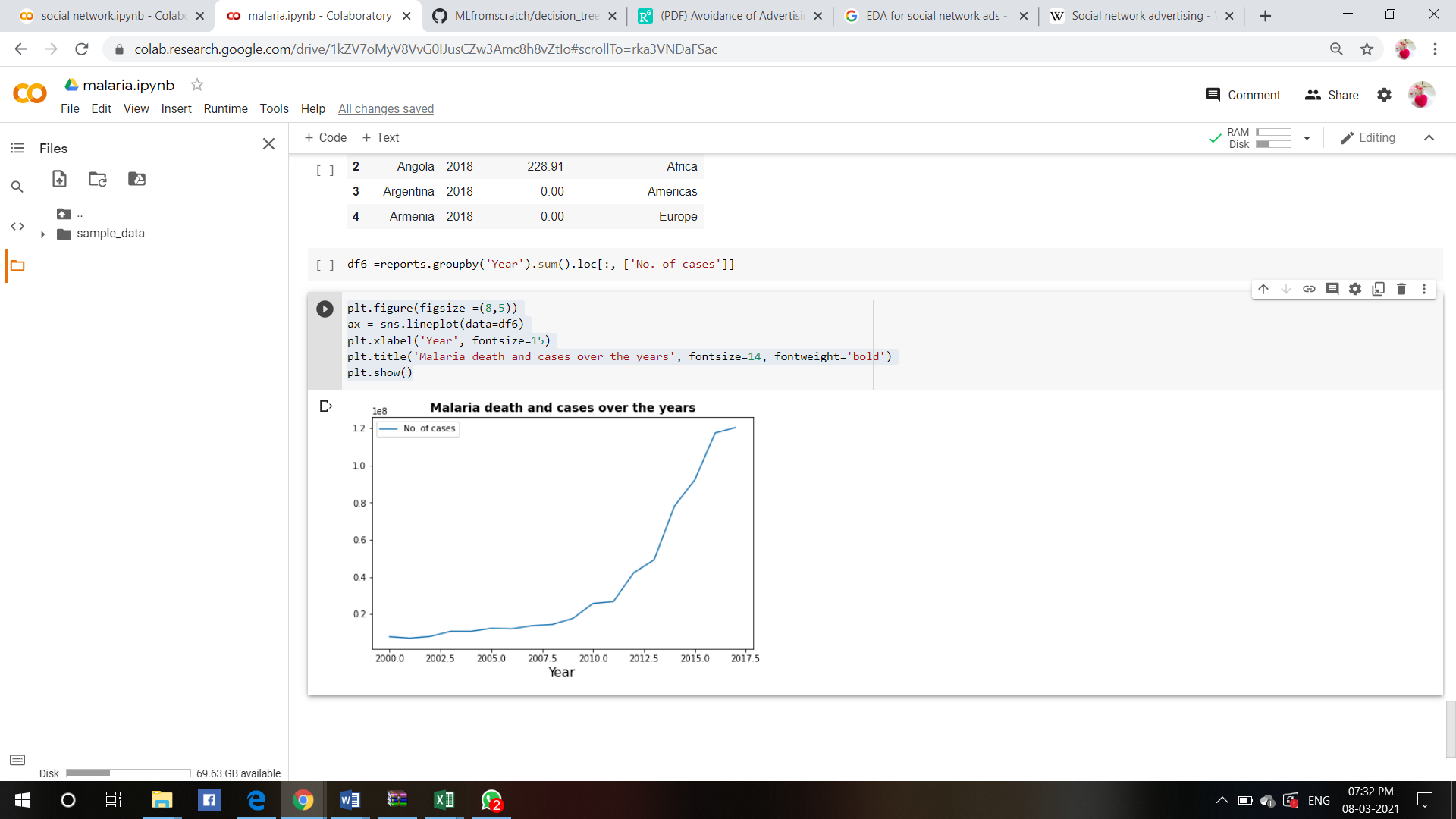
plt.figure(figsize =(8,5))

ax = sns.lineplot(data=df6)

plt.xlabel('Year', fontsize=15)

plt.title('Malaria death and cases over the years', fontsize=14, fontweight='bold')

plt.show()



**Conclusion:**

from the data interpretation I conclude that no. of cases, deaths, CFR … of malaria increases every year