**Exploratory Data Analysis for COVID-19**

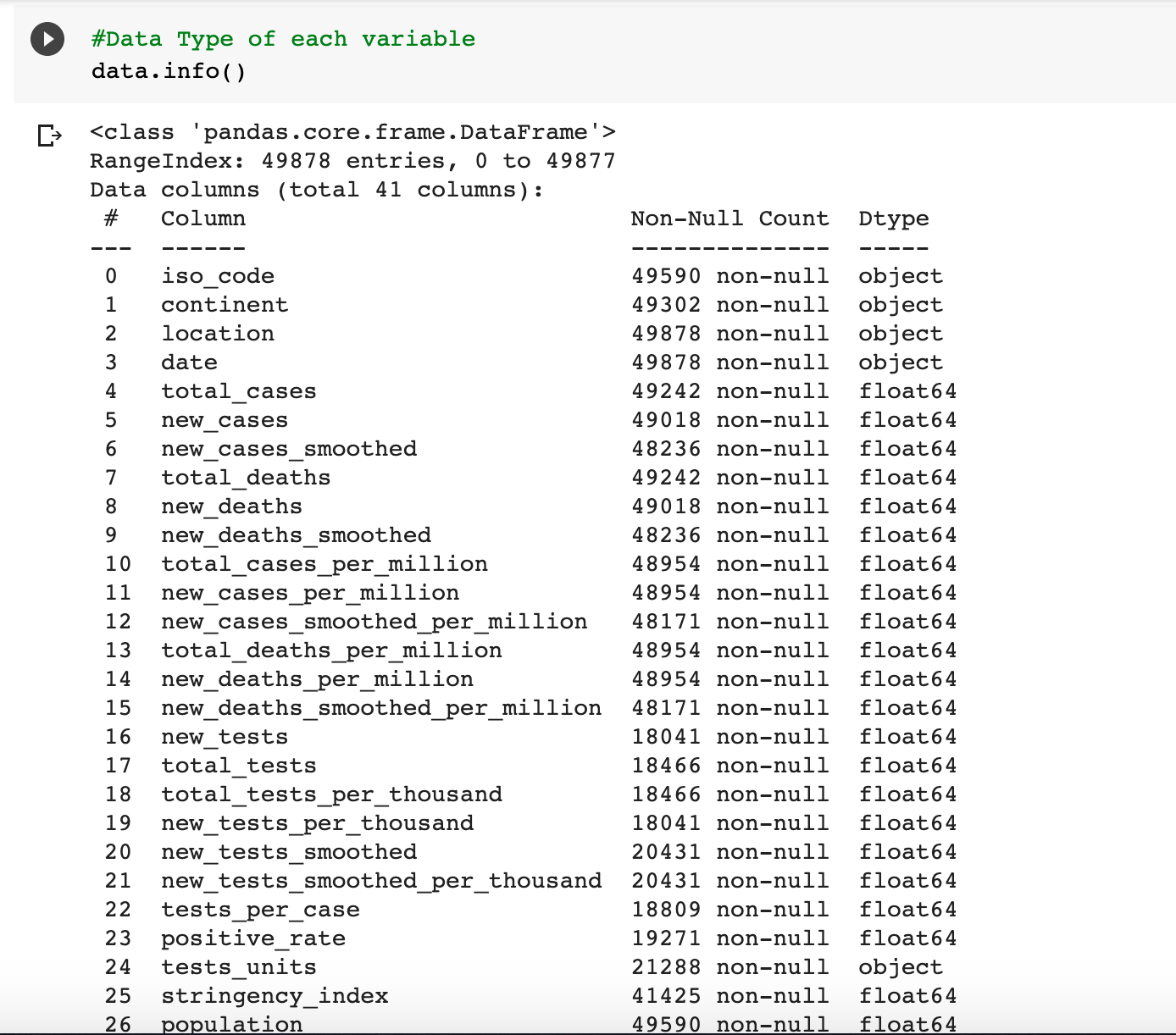
**Data Science Capstone Project   
Exploratory Data Analytics Report**

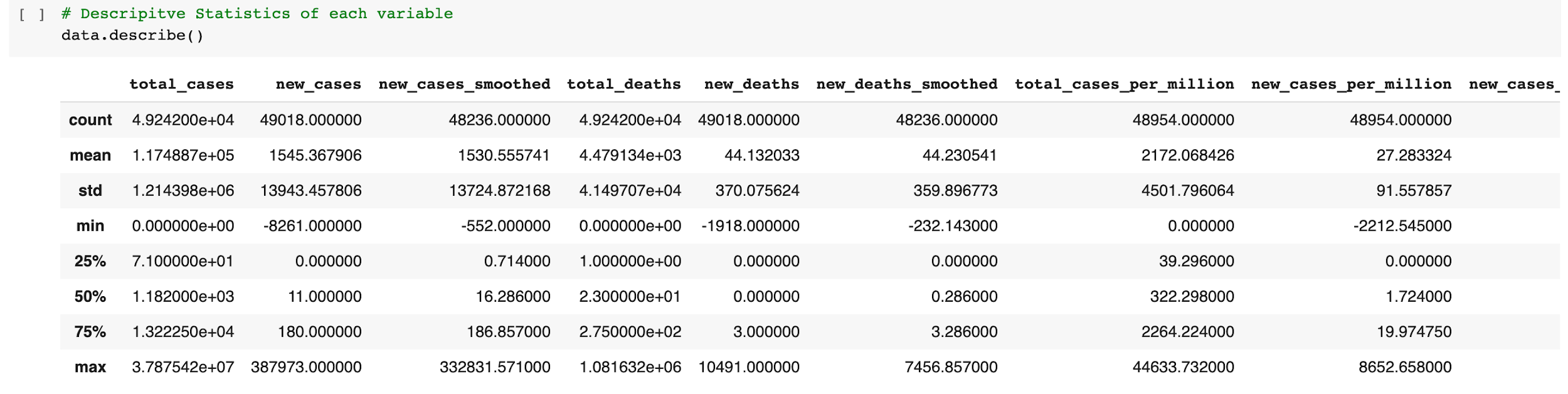
11/24/2020

**Analysis the basic metrics of variables**

* **Analysis of Basic metrics for Dataset 1(OWID):**

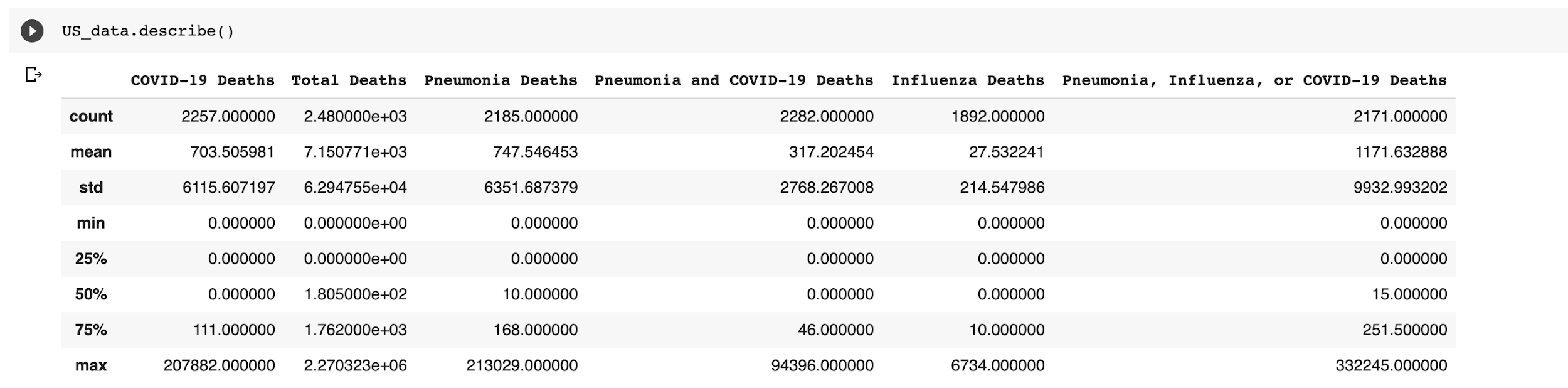
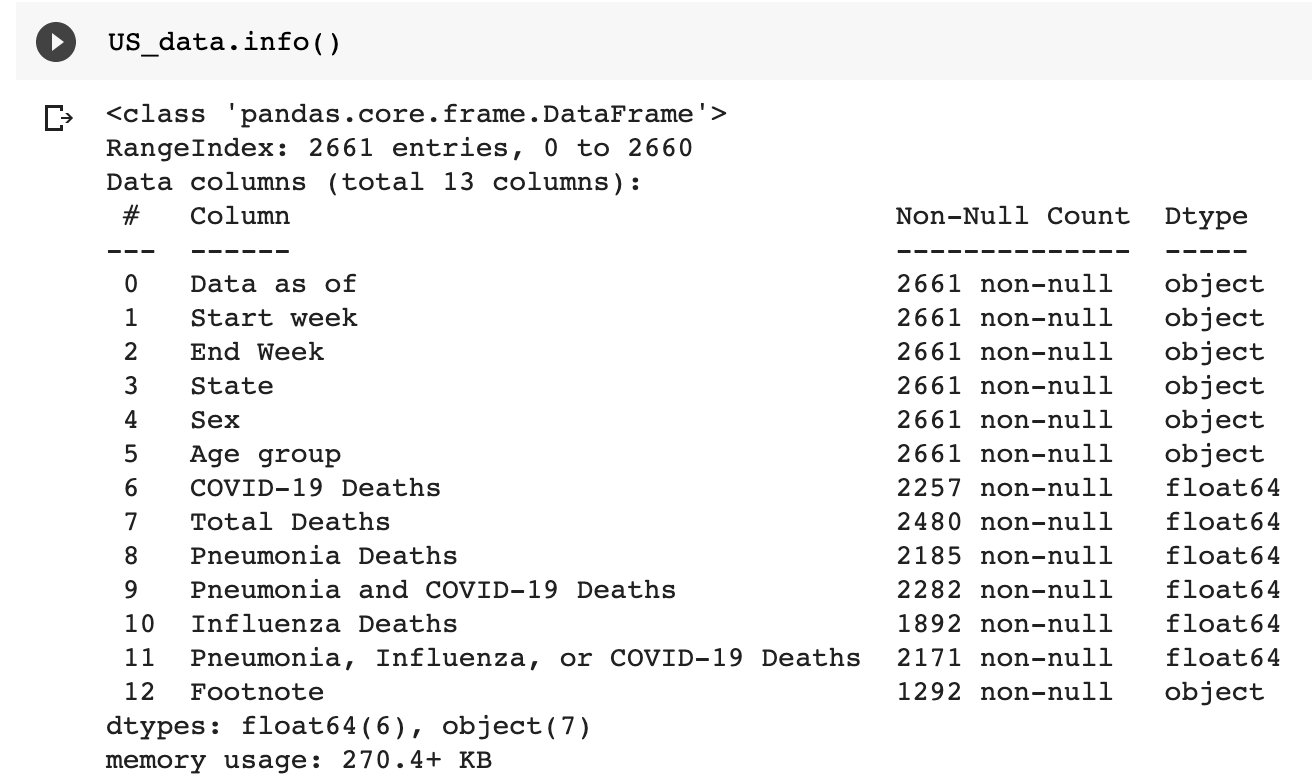
In dataset 1 there are 49878 entries with a total of 41 different columns.



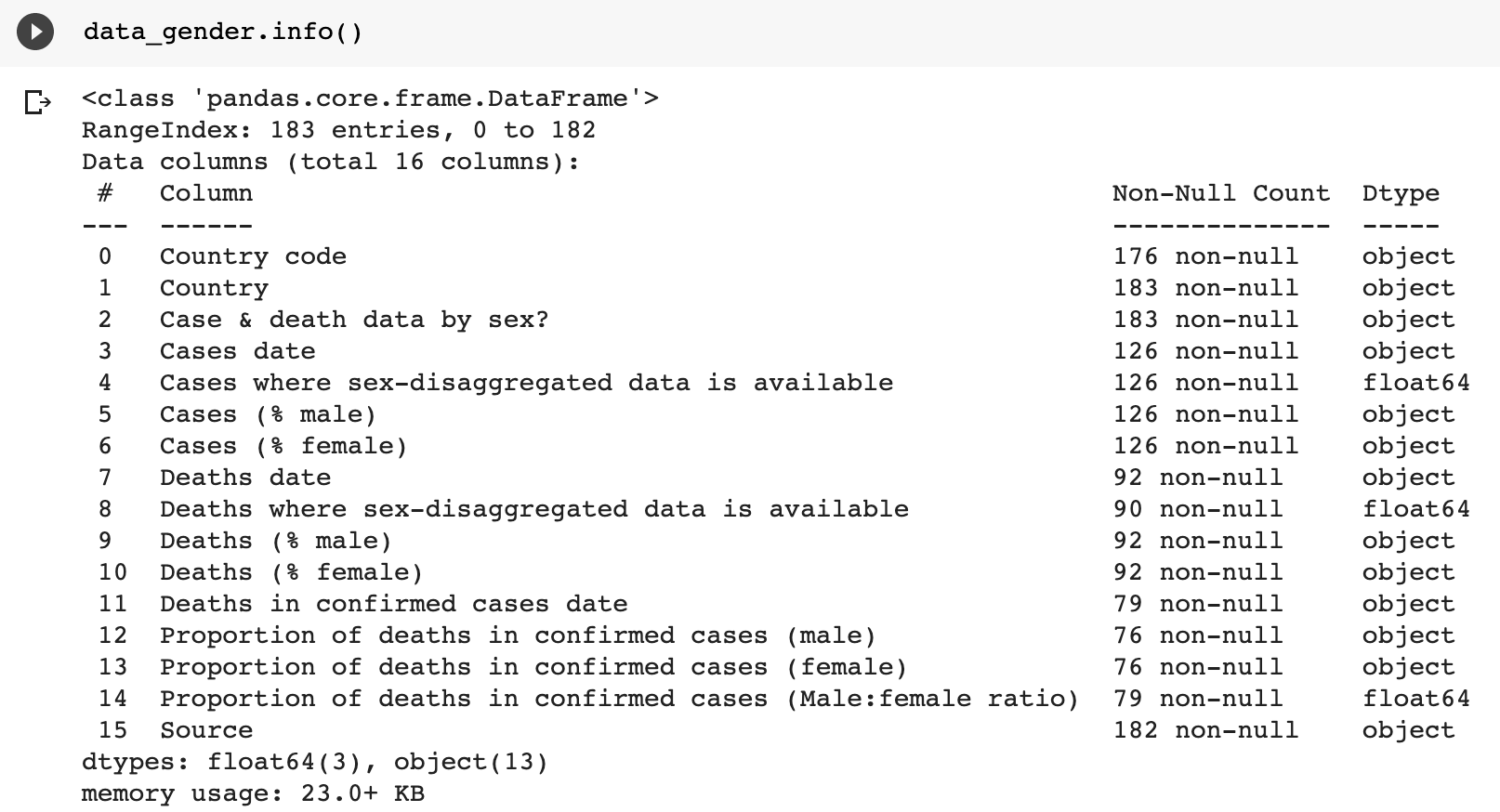


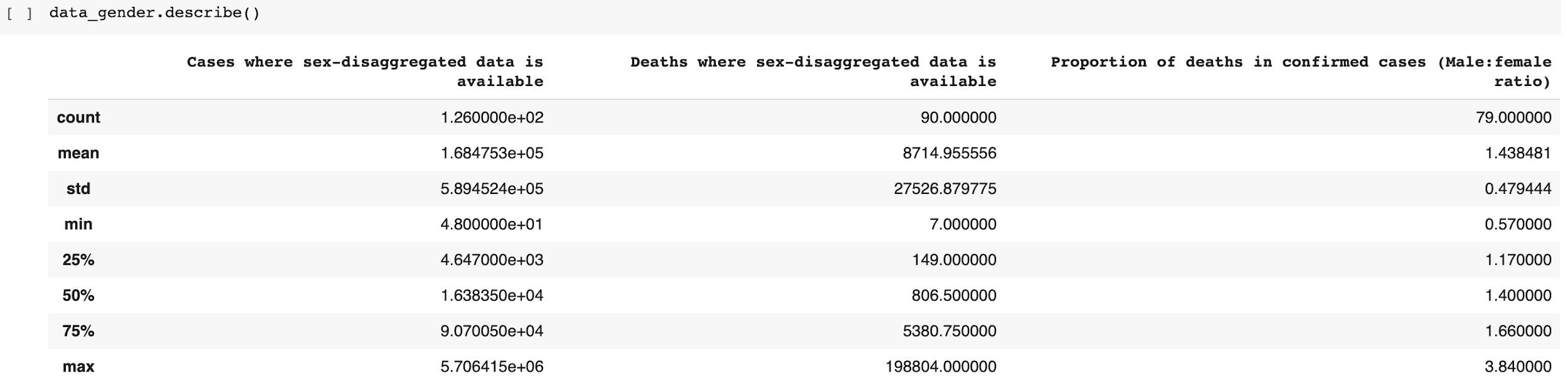
* **Analysis of Basic metrics for Dataset 2 (US age and sex data):**

In dataset 2 there are 2661 entries with a total of 13 columns.



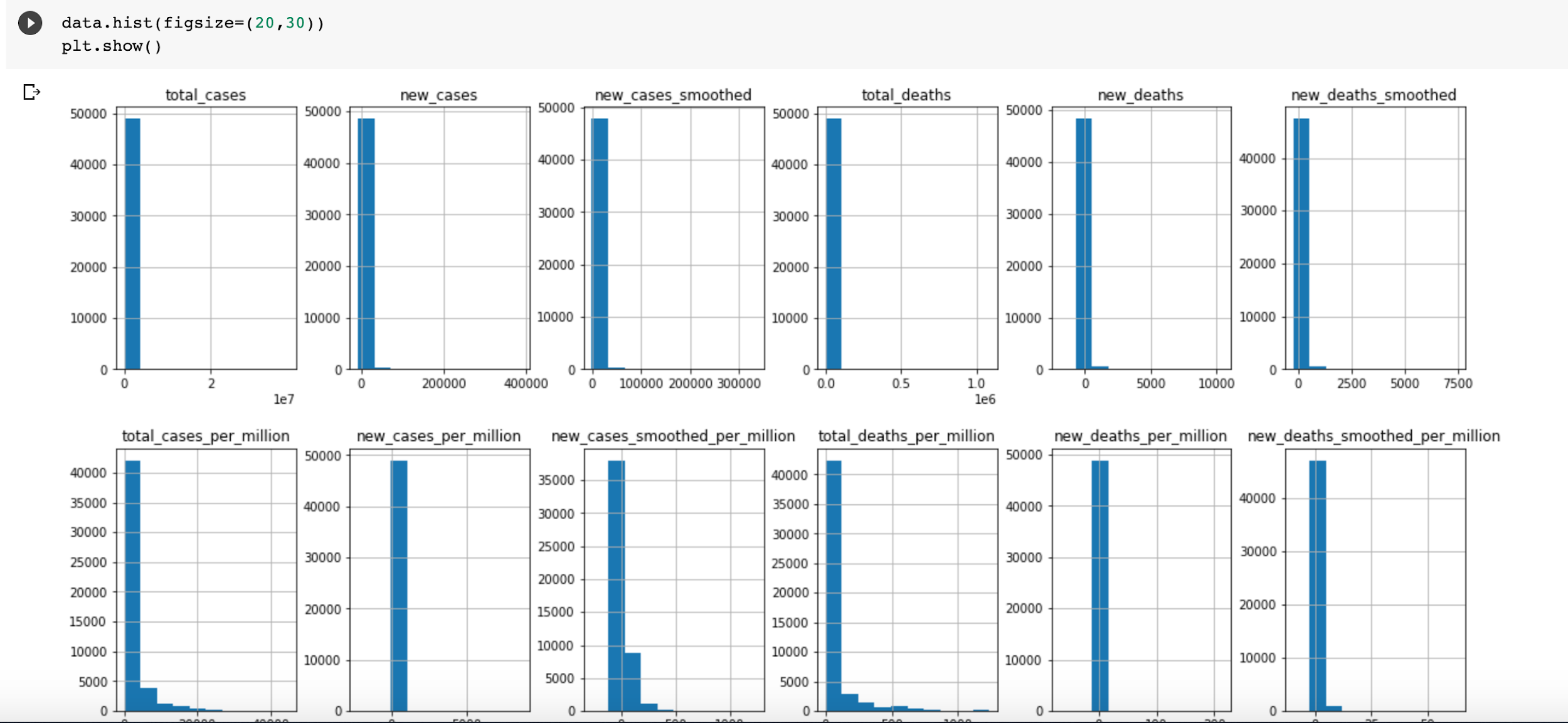
* **Analysis of Basic metrics for Dataset 3(Gender Related Data):**

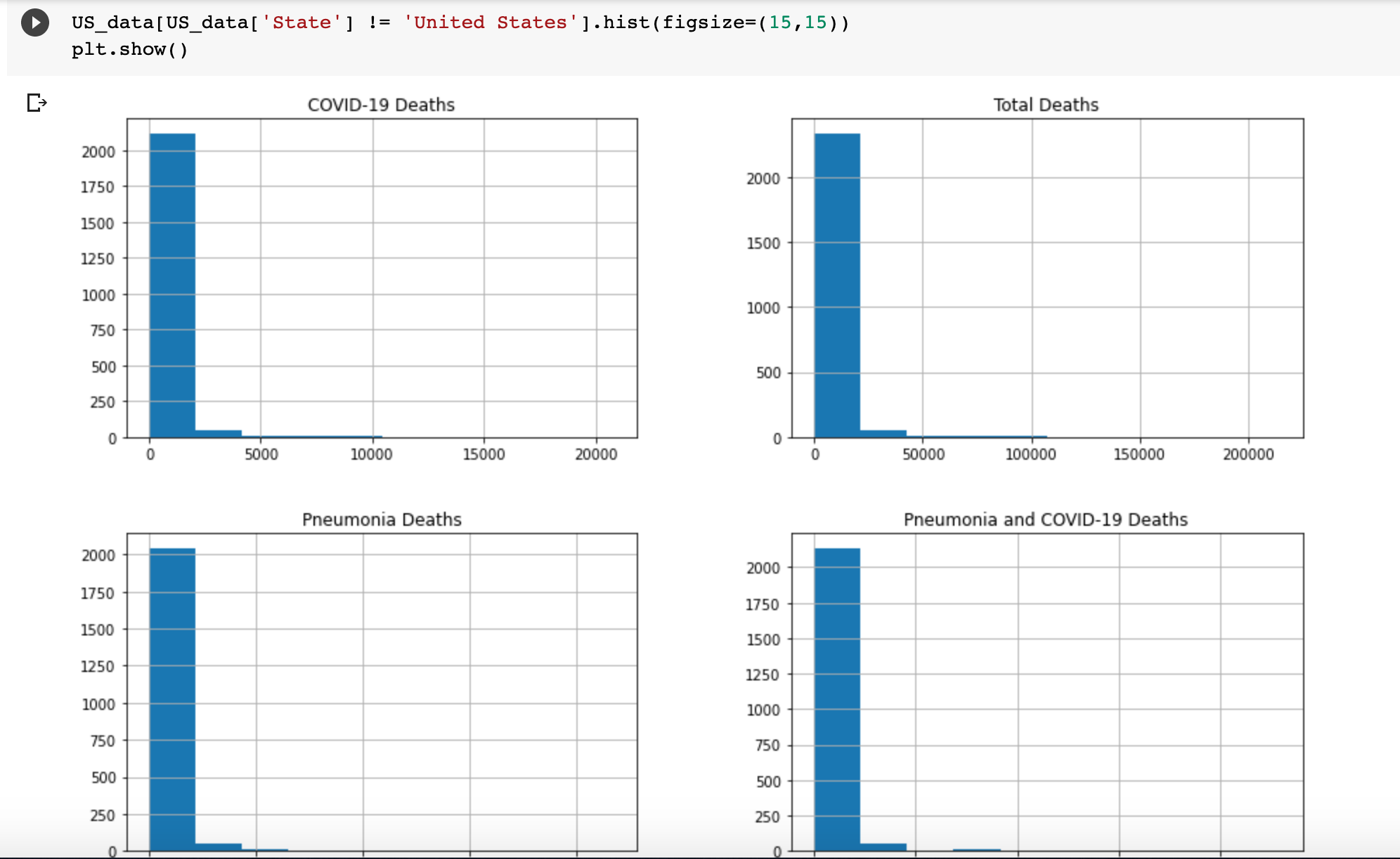
For the Dataset 3 there are 183 entries with a total of 16 columns.



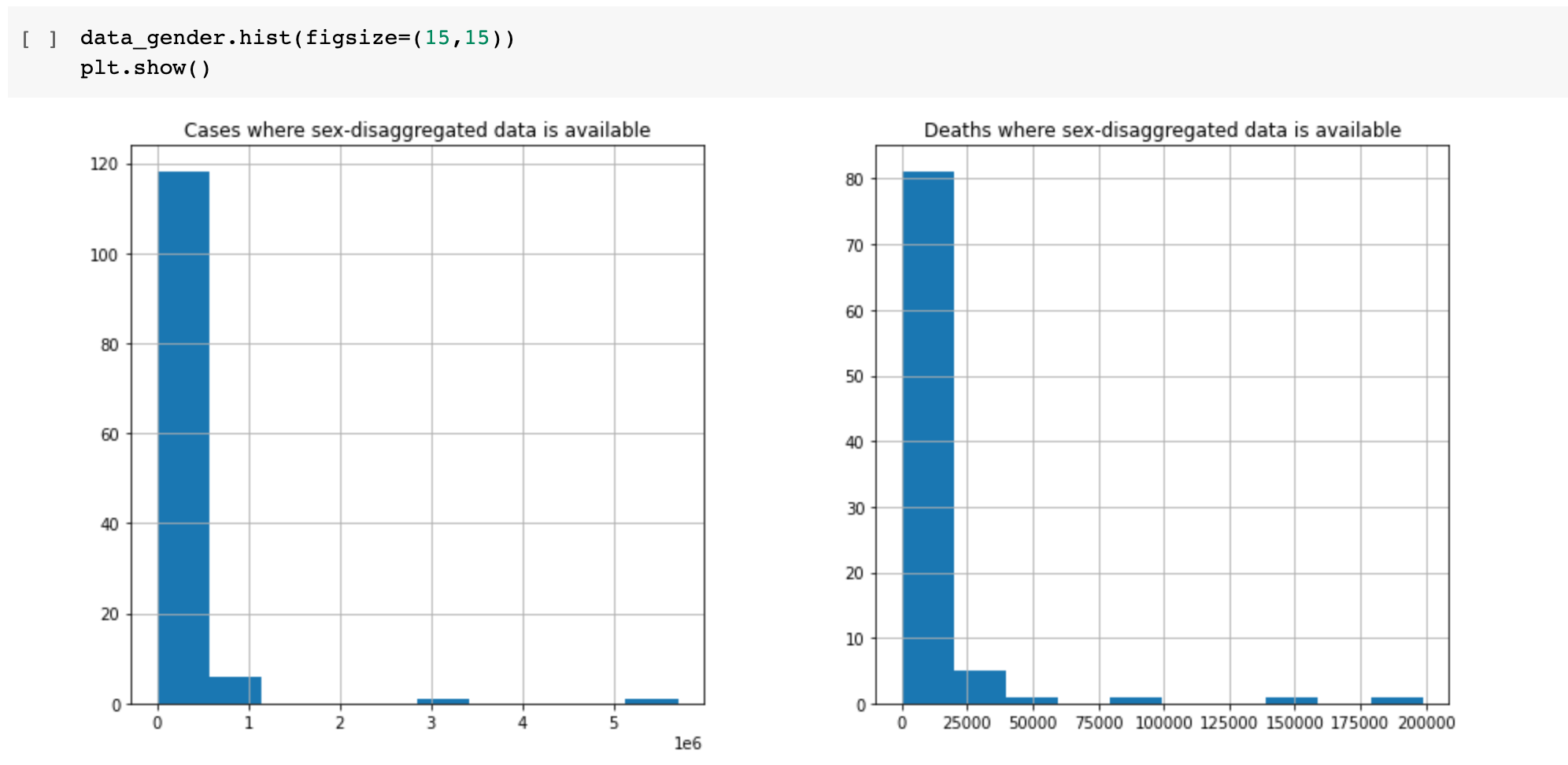
**Non-graphical and graphical univariate analysis**

**OWID:**



**Dataset 2 histogram:** 

**Dataset 3 Histogram:**



**Missing value analysis and outlier analysis**

OWID Dataset:

For the COVID-19 countries dataset, we noticed that there are 212 unique locations and there are two locations named ‘World’ and ‘International’. The website in which we downloaded the dataset from did not explain the reasoning behind International so it was decided that we would drop these rows with ‘International’ as the location from the dataset [1]. We also noticed that the columns also have the same information such as new\_deaths and new\_deaths\_smoothed for several of the variables. It was found that variables that were smoothed had a 7-day smoothing or had random variation data points removed. We decided to drop columns with smoothed data and also irrelevant columns such as test\_units and human\_development\_index. We felt that these columns were not of relevance to our project [2].

We also found that there were missing dates from the dataset in several of the countries. Some of the countries had the dates as a row and NaN as its value. But some of the countries had even omitted placing the date rows. Since there are 211 unique locations (after removal of International), we decided to create two lists of countries based on whether or not the country has all 288 dates as rows. We found that there were 67 locations that had all 288 dates and decided that these countries would still be enough to create a good analysis. We dropped all the countries that did not have the 288 dates and kept the 67 locations as our final dataset [3].

Even after removing these countries, there were still null values within the dataset [4]. We decided that since all dates were available for these 67 countries, we could use front filling to fill in the missing values for any columns related to COVID-19 information. It would not make sense to use front filling for columns such as population\_density, gdp\_per\_capita, female\_smokers, etc. So we decided to only use front filling for columns related to COVID-19 cases, deaths, total\_tests. But since front filling uses the first number in the dataset as its starting point, some of the columns have a null value, we had to remove the nulls with 0.0 into the first date: 2019-12-31 then front fill[4]. All columns that are associated with COVID-19 information no longer had null values and we believed that if we wanted to use the other information, there would be a need for another dataset[5].

Dataset 3(Gender Related Data):

For the gender related COVID-19 dataset there are 183 unique countries[6]. In the dataset there were entries where cases date was equal to 0 so we decided to delete those rows[7][8]. There are three columns which are not used for tableau analysis for this dataset so we have decided to drop those columns named “Proportion of deaths in confirmed cases (male)", "Proportion of deaths in confirmed cases (female)", and "Proportion of deaths in confirmed cases (Male:female ratio)”.

After removing the null values within the cases dates columns there were more null values in the dataset so we have decided to fill it with 0 to do the analysis in tableau.

**Data Visualizations (Tableau,Python):**

We created data visualizations focusing on exploratory data analysis using Tableau and Python. We used libraries in Python such as plotly and seaborn. There are several visualizations focusing on determining which countries and regions had the most cases and deaths overall. It was found that South America and North America are the two continents that have the largest increase over time and total deaths from COVID-19[9]. When examining countries in South America and North America, the United States has the most deaths and Brazil follows. The United States has an average of 750 deaths per day while Brazil has an average of 510 deaths per day[10]. This observation can be supported with another graph that details the total cases for the United States and Brazil as time goes on [11]. The United States and Brazil's total cases of COVID-19 both move upwards at a rapid pace. Canada and the Dominican Republic both are quite flat compared to the US and Brazil.

Another section of visualizations focuses on the mortality rate which is calculated by total deaths divided by population. From this formula, we were able to find which countries had the highest rates of mortality and which were the lowest by using the average. We observed that San Marino. Belgium, Spain, Italy, United Kingdom, Sweden, France, United States, Netherlands and Brazil were the top 10 countries with the highest mortality rates [12]. On the other hand, Cambodia, Vietnam, Taiwan, Sri Lanka, Thailand, Nigeria, Malaysia, China, Singapore and Nepal have the lowest mortality rates from COVID-19 [13]. It was also noted that within the top 10 countries with the highest mortality rates, that even though San Marino and Belgium may have high mortality rates, it does not seem like it is increasing. Brazil and the United States seem to be increasing linearly [14].

We also have a section that focuses on visualizations by case fatality rate. The case fatality rate is the total deaths divided by the total number of cases. We wanted to see which countries had the highest and lowest case fatality rate. We found the countries with the top ten highest case fatality rates are France, Italy, Belgium, United Kingdom, Netherlands, Mexico, Philippines, Spain, Sweden, and San Marino. It appears that the highest case fatality rates happened in the beginning of the pandemic before there was a solid treatment plan since the fatality rate has been going down over time in all of these countries. In comparison the ten countries with the lowest case fatality rate are Sri Lanka, Kuwait, United Arab Emirates, Oman, Iceland, Bahrain, Nepal, Qatar, and Singapore.

We also used visualizations to examine how COVID-19 affects various age groups and how it differs among male and females. It was observed that those in their 50s and older are the ones that have a higher risk of dying from Pneumonia and/or COVID-19. Those that are younger than 50 have the lowest risk in dying from these two diseases [15]. Males were found to have a higher risk of dying from COVID-19. Compared to 10 males, there may be 7 females that die from the coronavirus [16]. We plotted a bar graph that showed that females that are 85 and older were more likely to die than males that are 85 and older. Males that are 75-84 are more likely to die than females 75-84 from COVID-19. One hypothesis that may explain why is that females tend to live longer. Since there are more females that are 85 and older living than males it could explain why the deaths are higher than males 85 and older[17].

There are several visualizations focusing on determining which countries had the most cases and deaths in tableau. Primarily, we made the geo-graph in tableau which includes the geographic area of the whole world containing the total cases of COVID-19 from February to October 2020. From this graph we showed the total number of cases for each country[18]. We observed from the line graph on total cases for each country that initially there were very few cases in the whole world until March 18th where there was a sharp increase. The United States has the largest number of cases, 7.8 million, by October 12th. In comparison the Falkland Island has the least number of cases which is only 13[19]. Further, we made a dual axis graph which includes total cases and total number of deaths. From this analysis we understood that the total number of cases is extremely high which is approximately 37 million compared to deaths which is approximately 1 million[20].

We also made visualizations in python using Plotly library which enabled us to make interactive time series graphs in python. We made time series graphs which include total cases and total deaths. The largest number of cases is in the United States which has 7.8 million cases. Until June 15th there was a small number of cases in India. In July there was a sudden increase in the total number of cases making it the second largest number of total cases of COVID-19 in the world with 7.1 million[21]. When we look at the total number of deaths in the world, the United States has the largest number of deaths at 214,771. Brazil has the second largest number of deaths which is 150,488[22].

There are several visualizations focusing on determining which countries had the most cases and deaths related to gender made with tableau. For this graph we used dataset\_gender by which we made a graph between cases percentage female and male, death percentage female and male. In both, total cases and total deaths graphs, we found that the cases male percentage is higher than the female percentage. In the total cases graph male percentage is 45 and female percentage is equal to 40[23]. Whereas in the total deaths graph male percentage is 35 and female percentage is 25[24].

Further we decided to make a geo-graph for the world which shows the most cases and deaths percentage according to the gender. From these graphs we found that Vietnam, Cabo, Brazil and Canada have the highest percentage of female cases but have the highest percentage of male deaths which is different from the trend which is that all the other countries have the highest percentage of male cases and deaths over females.[25][26]

**Feature engineering and analysis**

The OWID dataset is the only one that we decided to do feature engineering and analysis on since it contained the most features and information of the three datasets. The first step that we decided to do was check the correlation matrix. We decided to look at two different ways the correlation matrix is calculated to see if there was any difference. We observed that there were slight differences between the two correlation matrices but not a significant amount. The first was through spearman correlation and the second method was through kendall correlation[27]. We found that total and new cases and total and new tests were highly positively correlated. We also found that handwashing facilities and life expectancy are highly negatively correlated with extreme poverty. This finding is not surprising since countries that have a high number of people living in extreme poverty will have a lower life expectancy. Most features we found are not correlated at all or are slightly correlated.

We decided to add two features to our dataset. The first was mortality rate which is found by dividing the total number of deaths by the population of the country. The second feature that we added was case fatality rate which is found by dividing the total number of deaths by the total number of cases.

The next thing we wanted to do was check feature importance and permutation feature importance for new and total cases and deaths. We used scikit learn’s random forest regressor to do the feature importance since the random forest will automatically give us the feature importance by how the trees are built. Scikit learn was also used in the permutation feature importance. The reason we tried both is because the feature importance from the random forest has a flaw where it favors features with many unique values and wanted to see if it was a problem with this dataset. We found the most important features to be the same from both the random forest and permutation importance.

The first step in finding the most important features was to only use the last reported date of the dataset to be able to make use of all of the features. There were still some missing values on this date so we used a random forest imputer to impute all missing values but also had to drop handwashing facilities since there were no values. When predicting total cases we found the most important feature to be new deaths. Not surprisingly population, extreme poverty, and population density were all in the top 8 most important features[28]. When predicting new cases we found that new tests were the most important feature. Interestingly human development index, extreme poverty, and population were all highly important features[29]. While predicting total deaths we found that new cases were the most important feature. We also found that positive rate, diabetes prevalence, and extreme poverty are highly important features which we would expect given what we know about COVID-19[30]. When predicting new deaths we found that new tests were the most important feature. Again given what we know about COVID-19, it is not surprising that male smokers, diabetes prevalence, and extreme poverty are all highly important features for predicting new deaths[31].

**Appendix**

[1]

Graphical user interface, text, application, email

Description automatically generated

[2]

Text

Description automatically generated

[3]

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

[4]

A close up of text on a white background

Description automatically generated

A picture containing chart

Description automatically generated

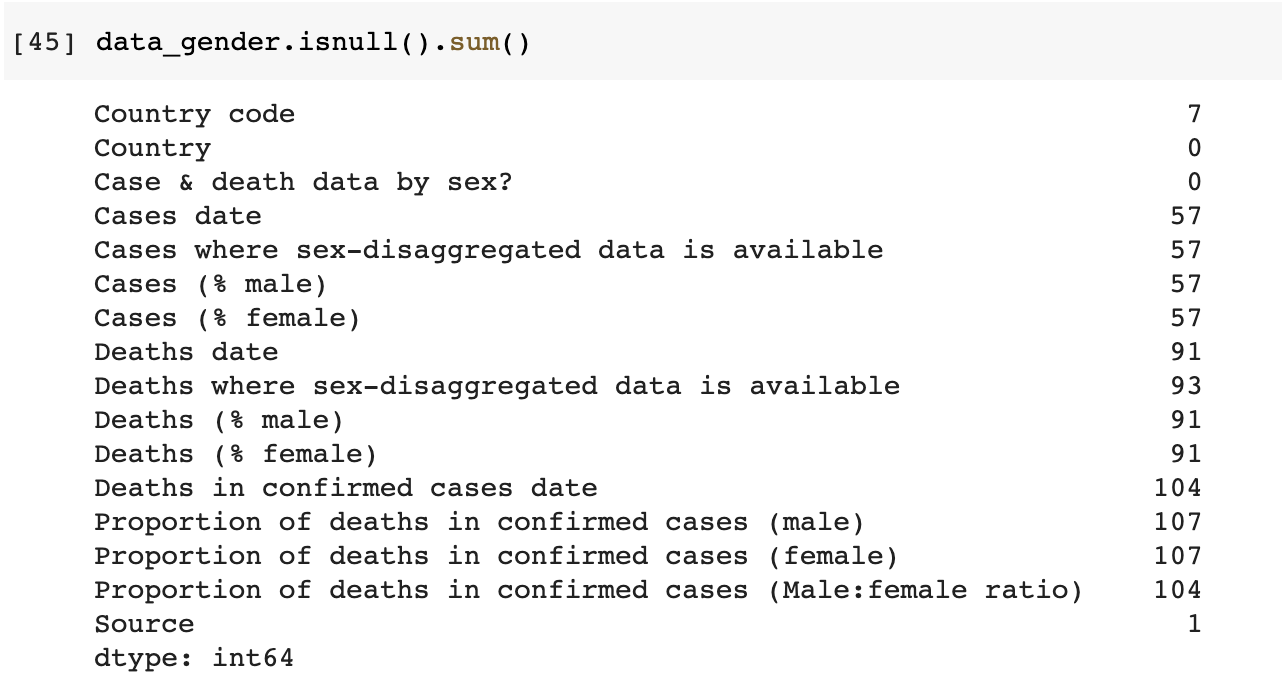
[5]

**Table

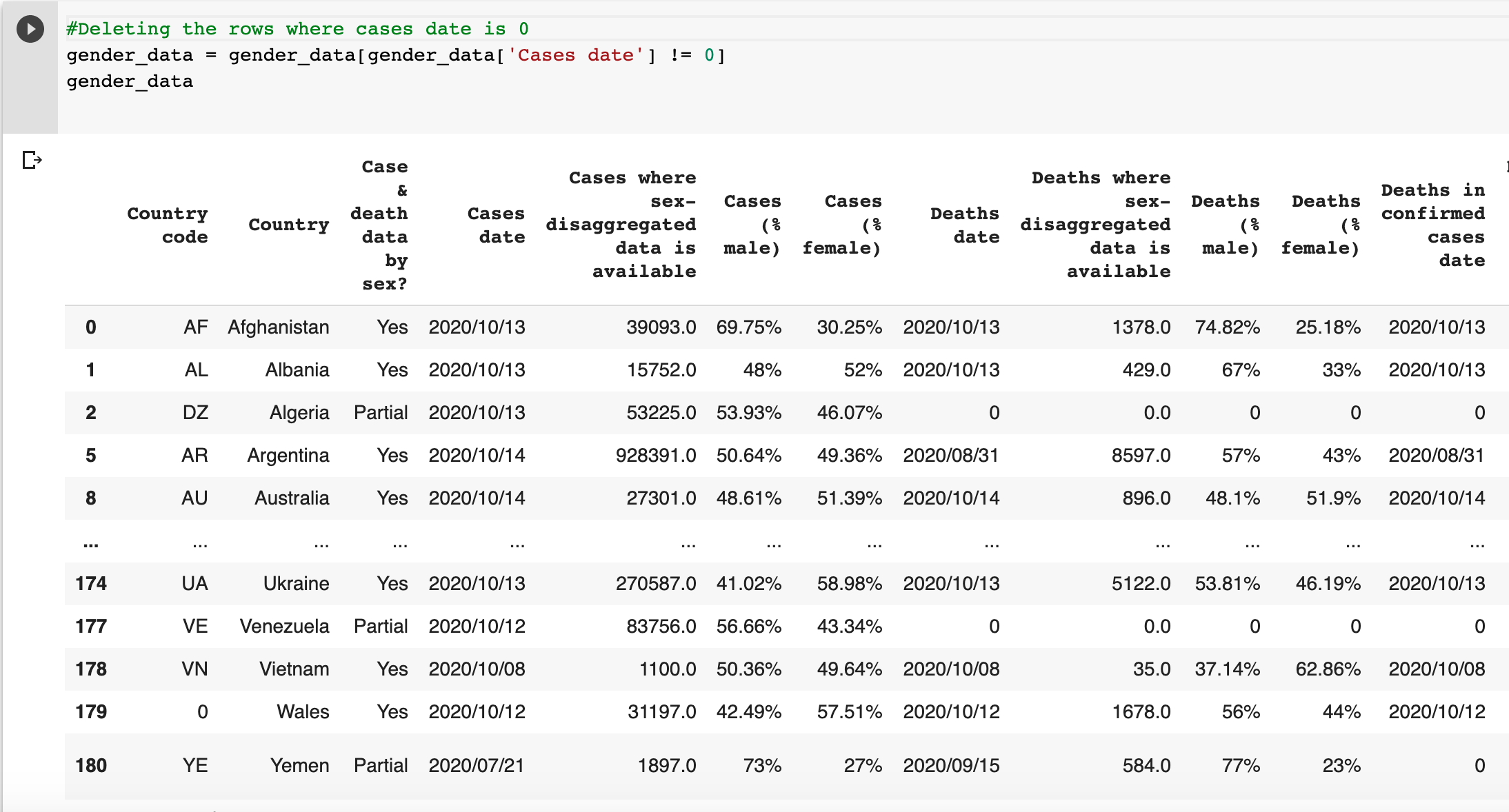
Description automatically generated**

**[6] **

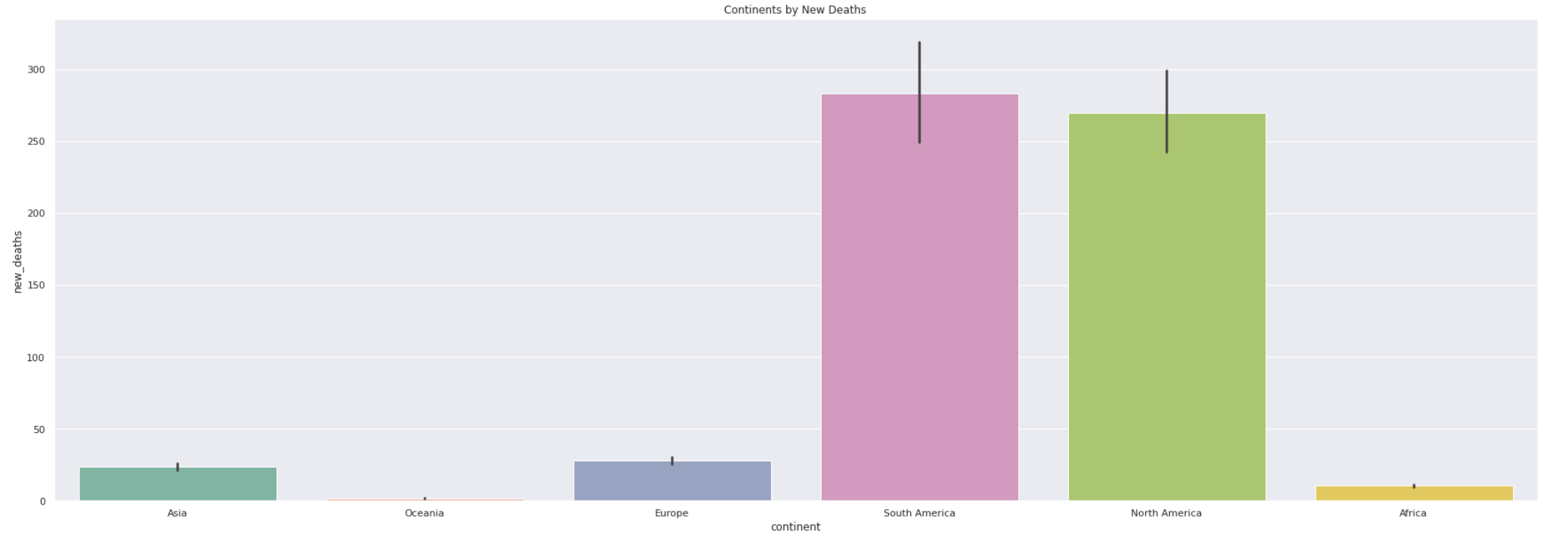
**[7]**

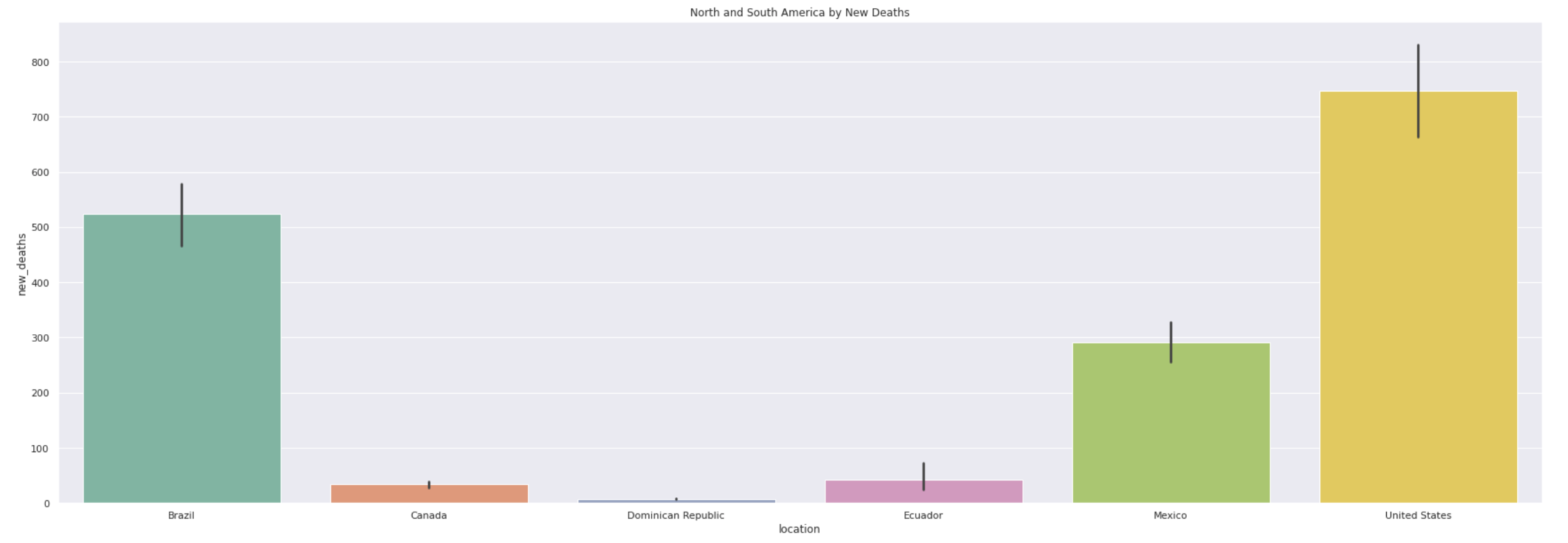
****

**[8]**

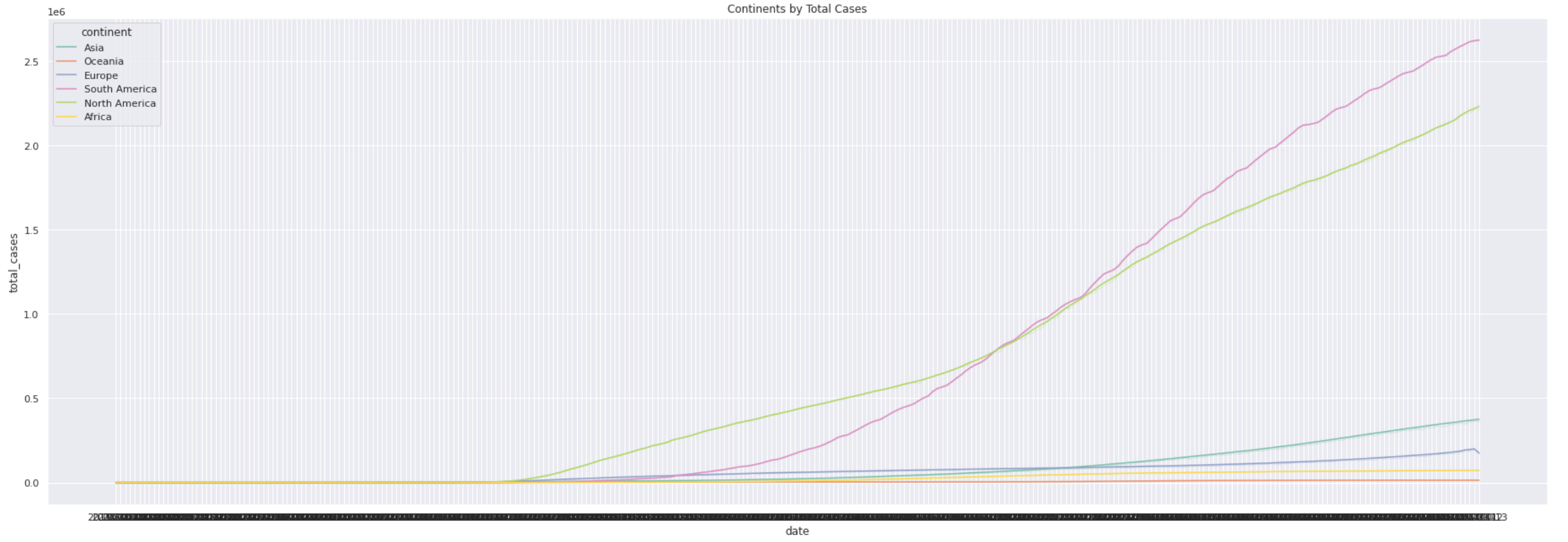
****

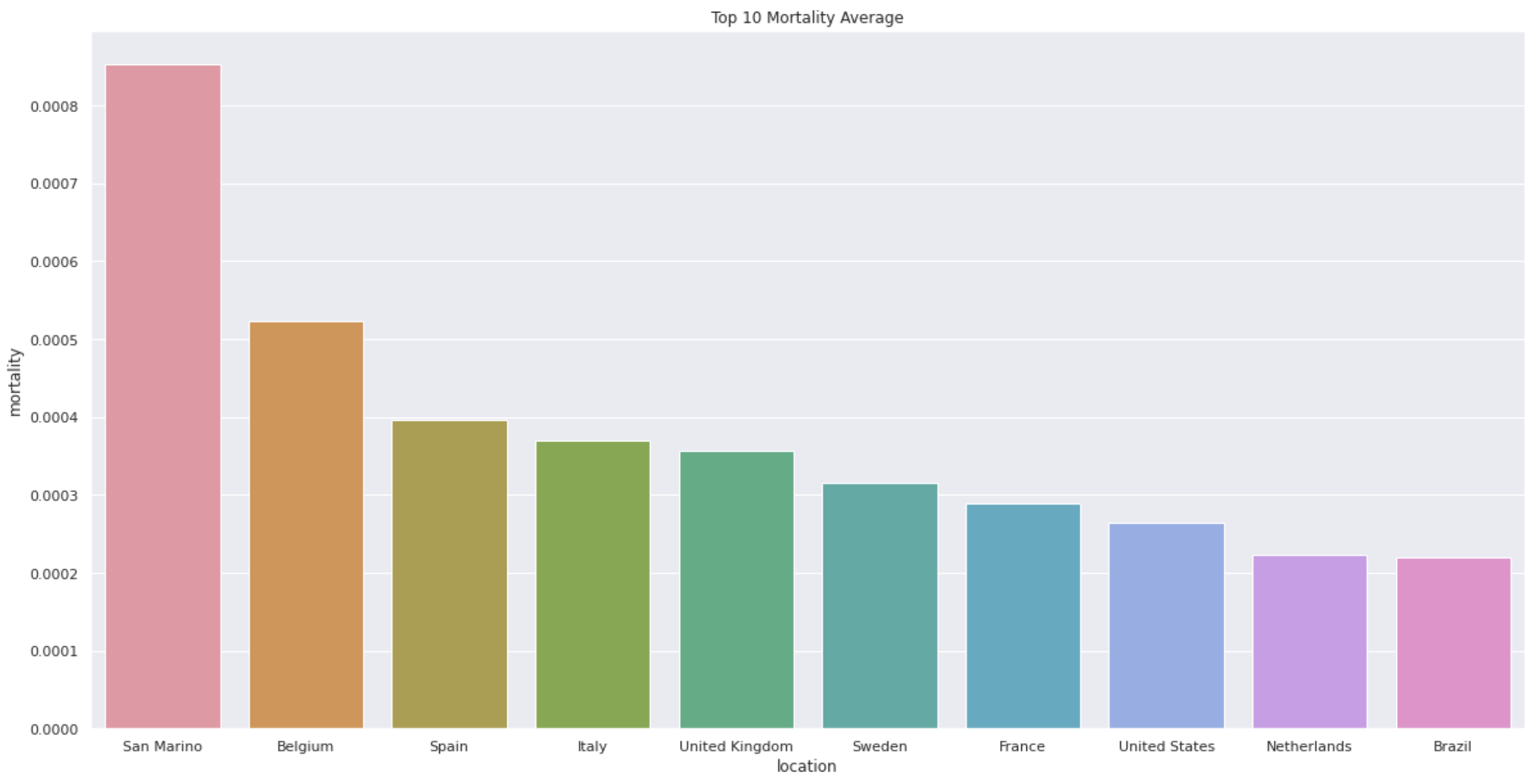
**[9]**

**[10]**

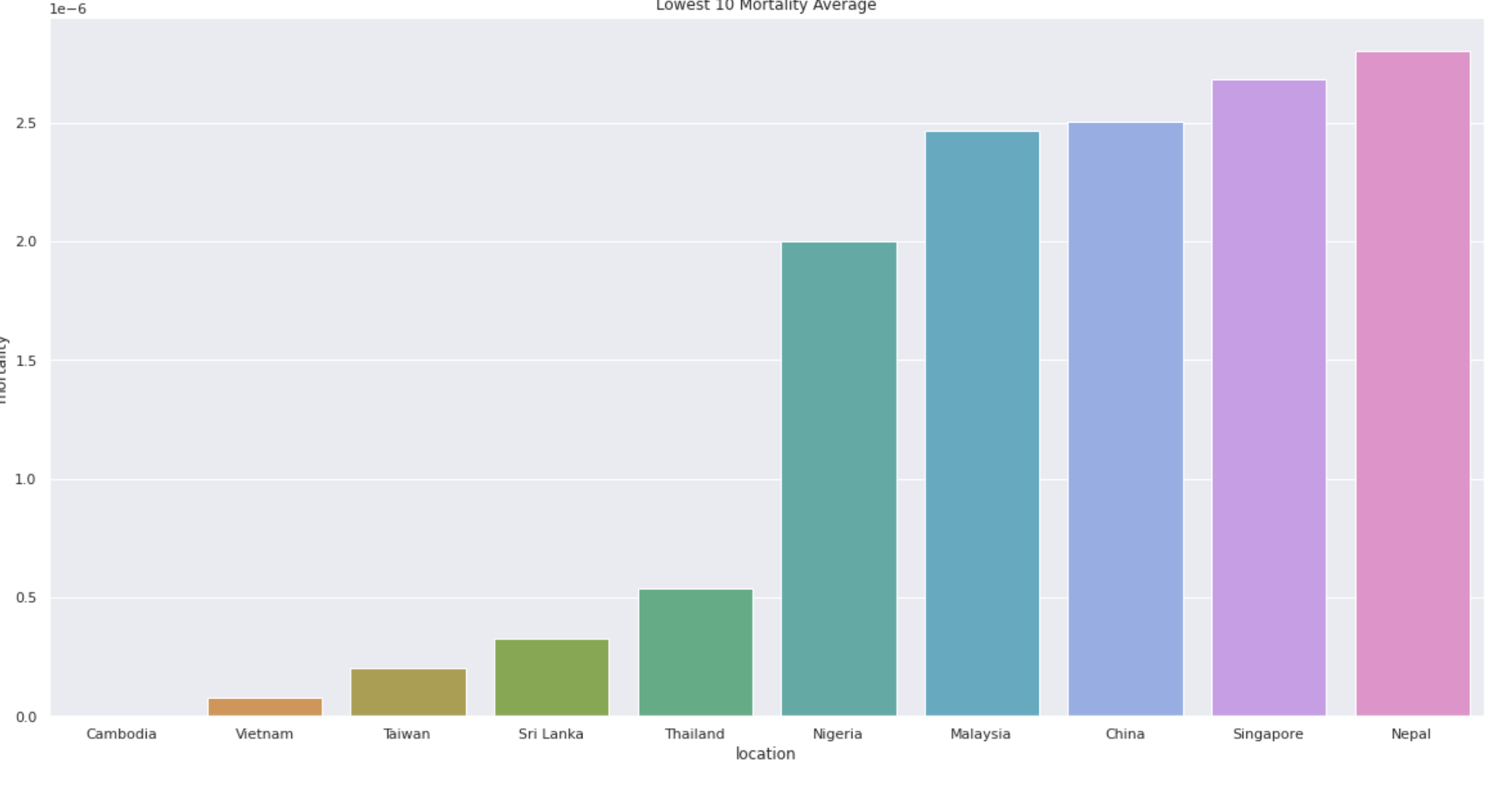


**[11]**

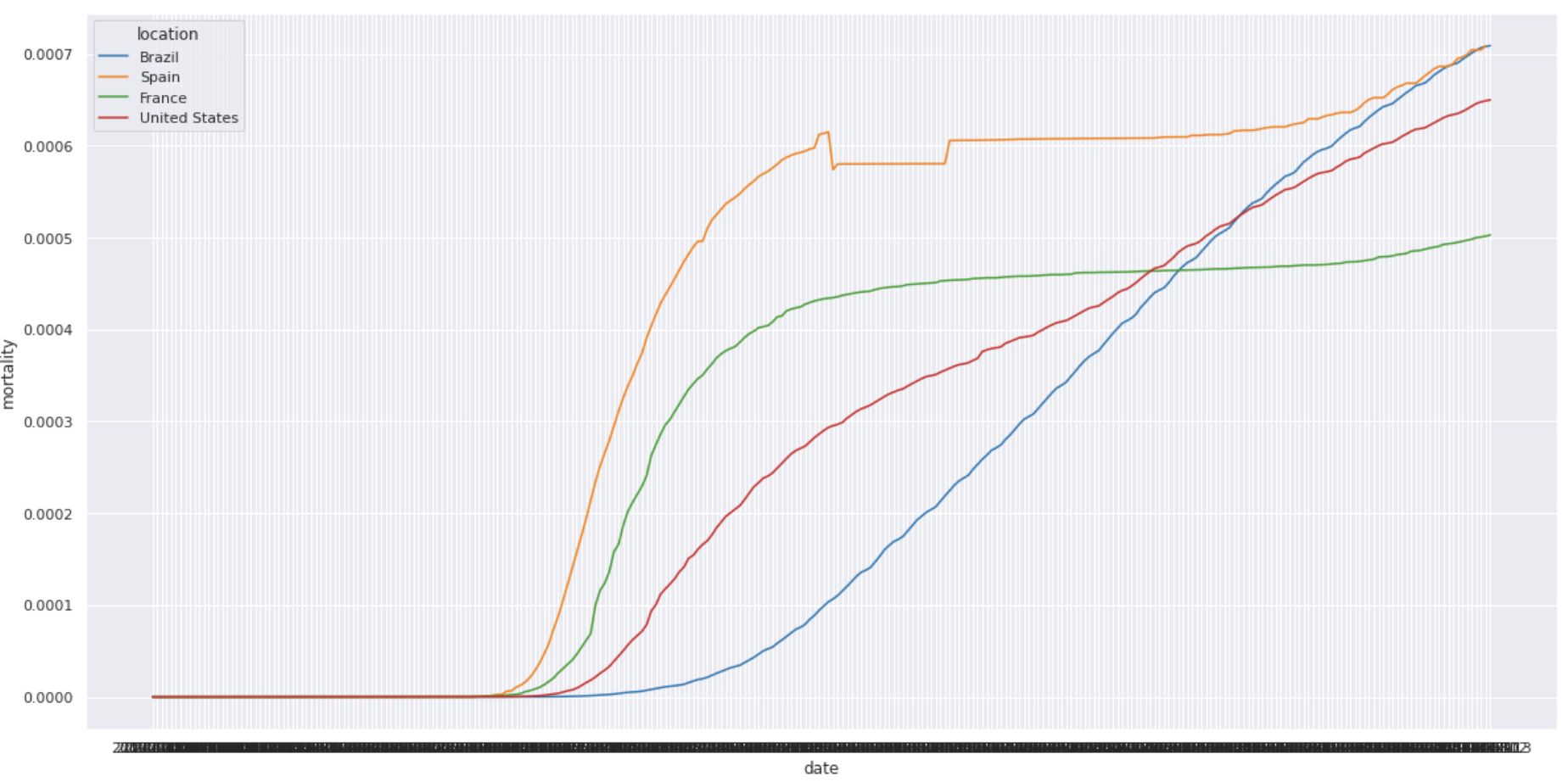
**[12]**

****

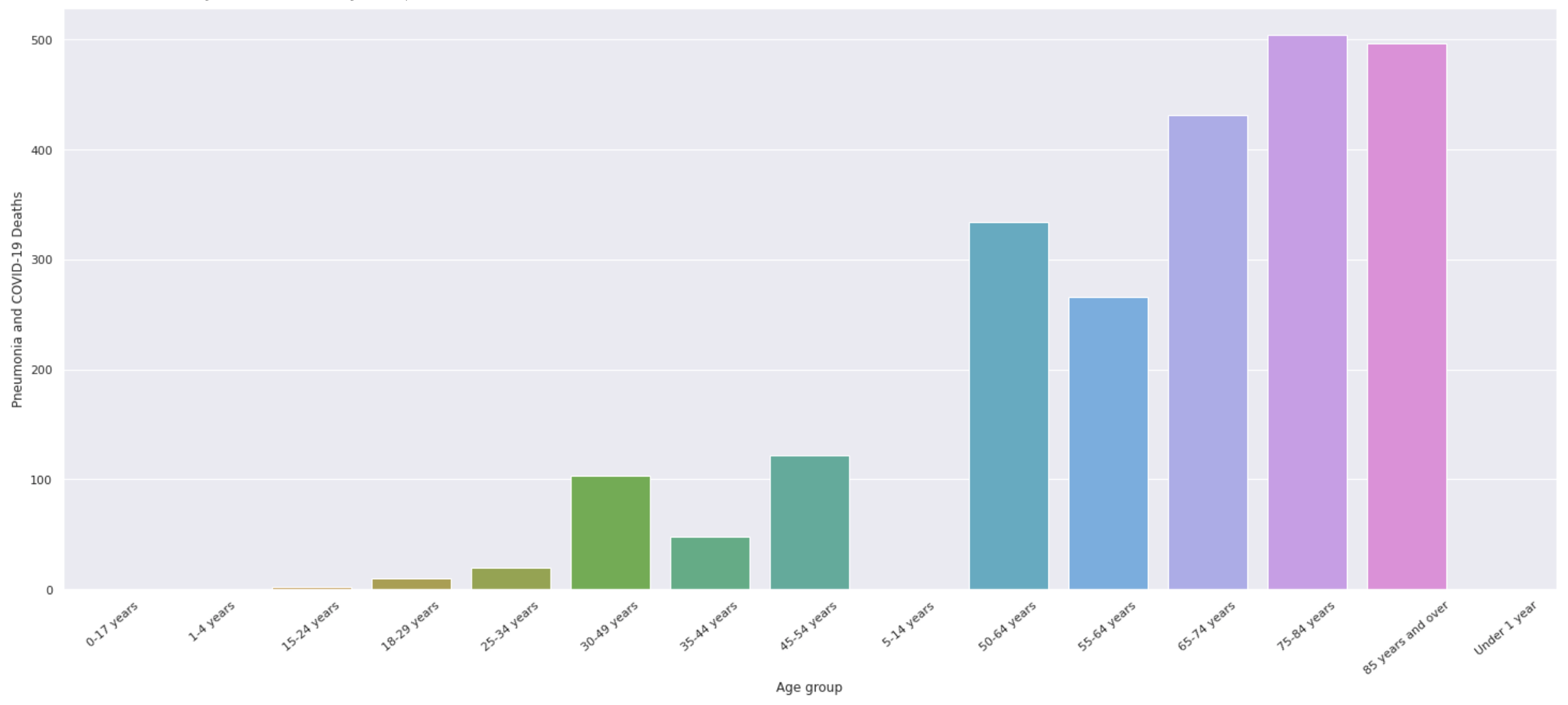
**[13]**

****

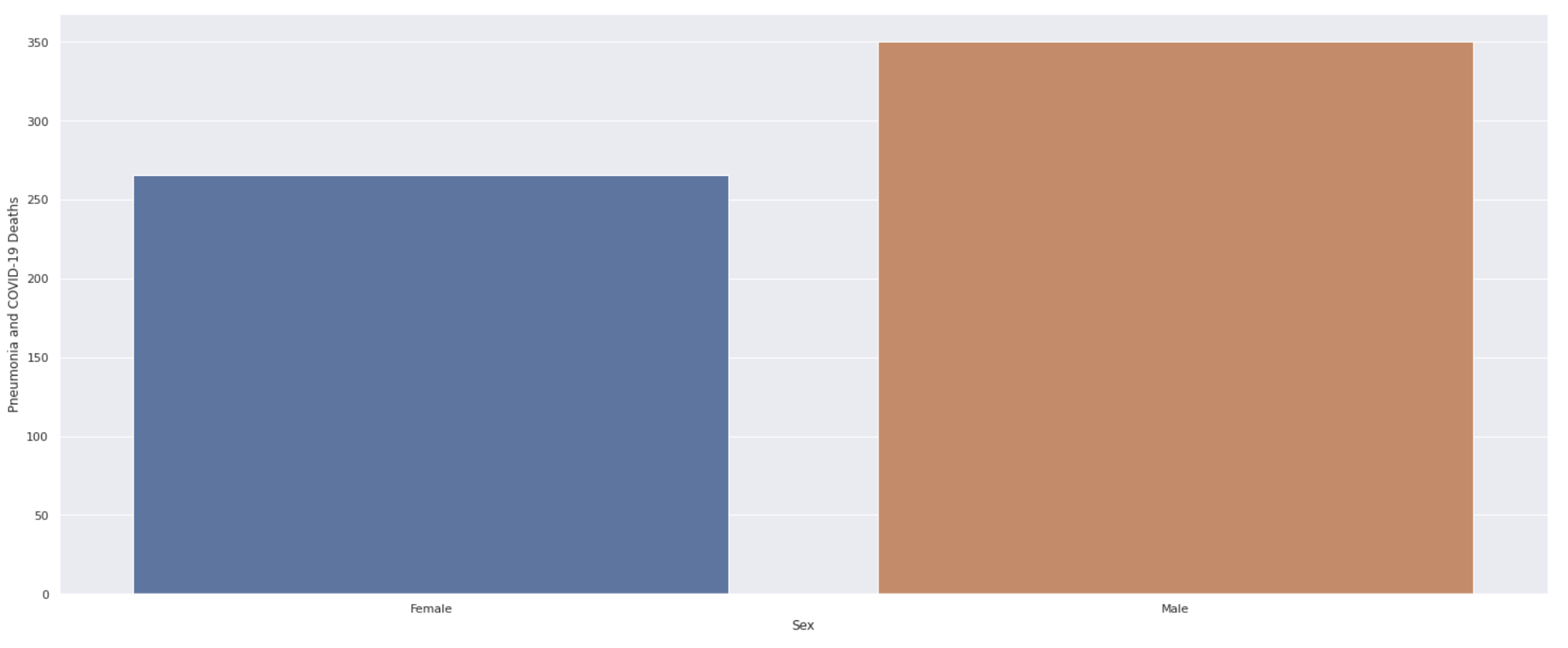
**[14]**

****

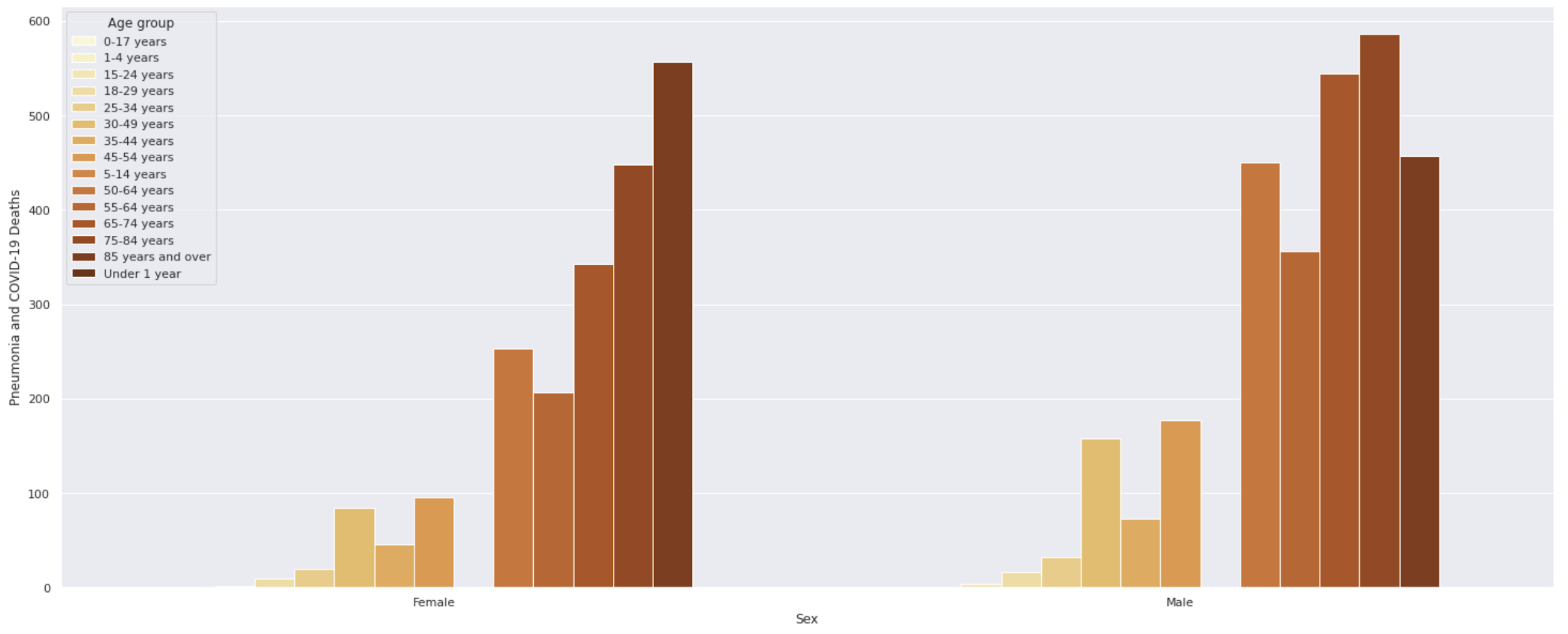
**[15]**

****

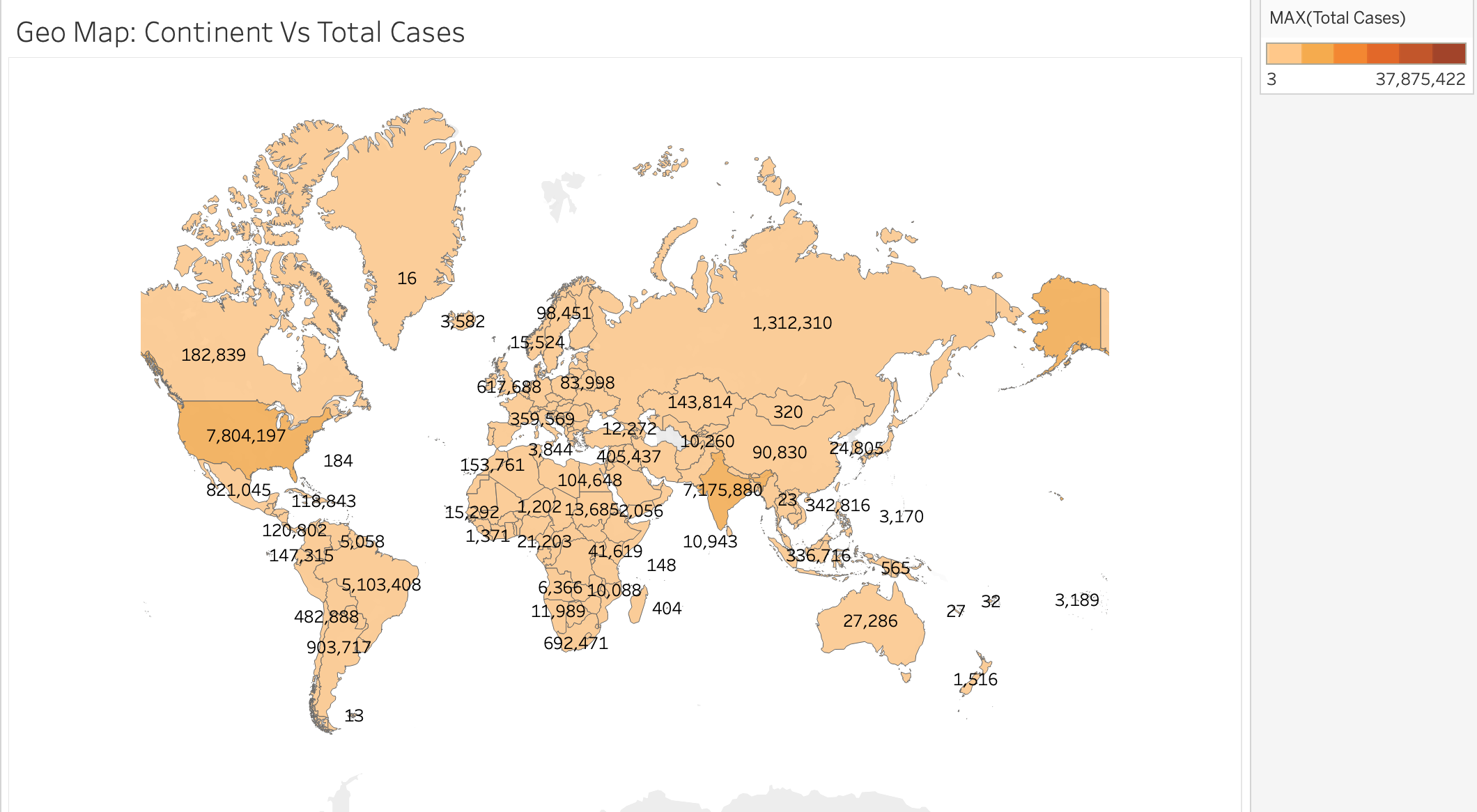
**[16]**

****

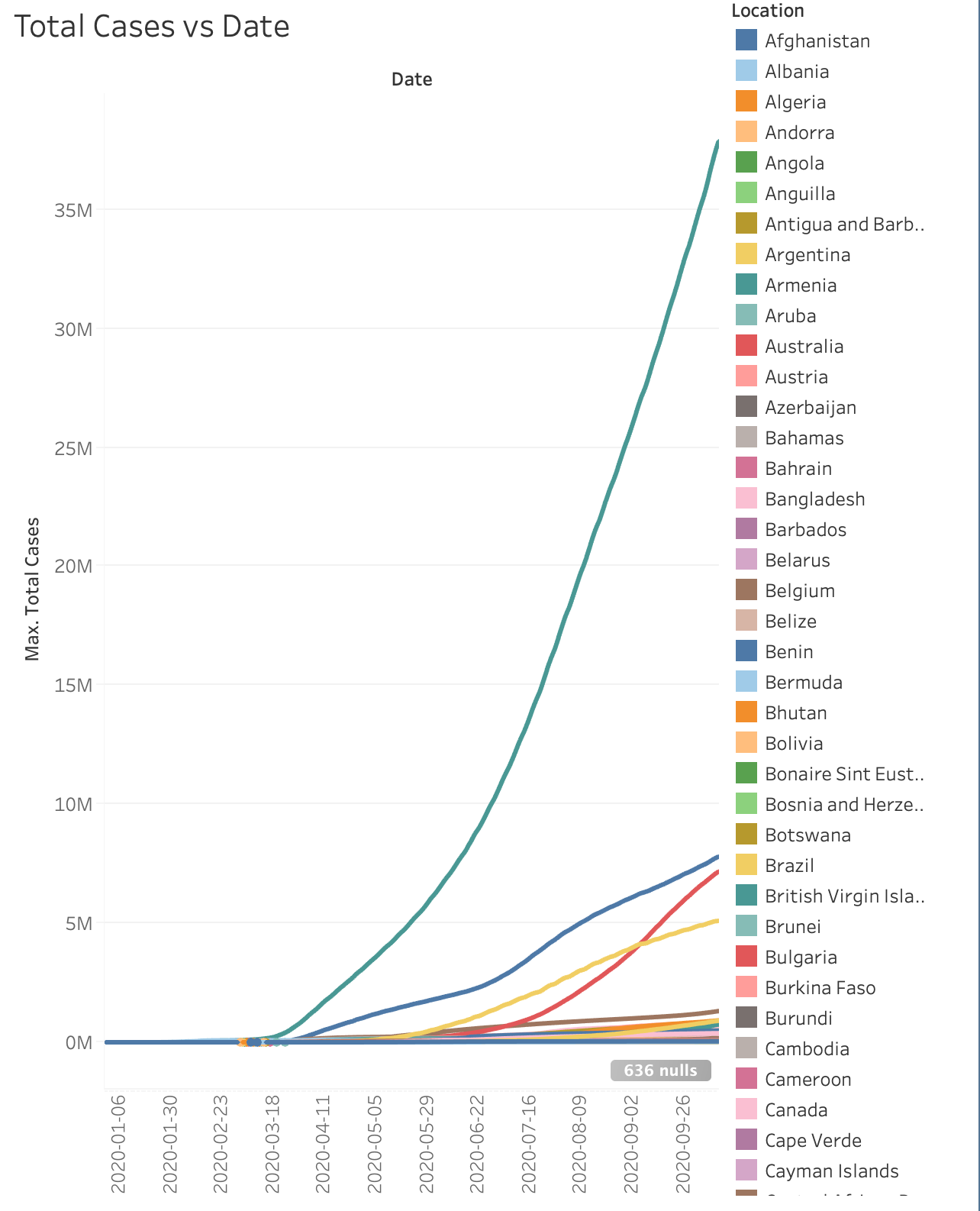
**[17]**

****

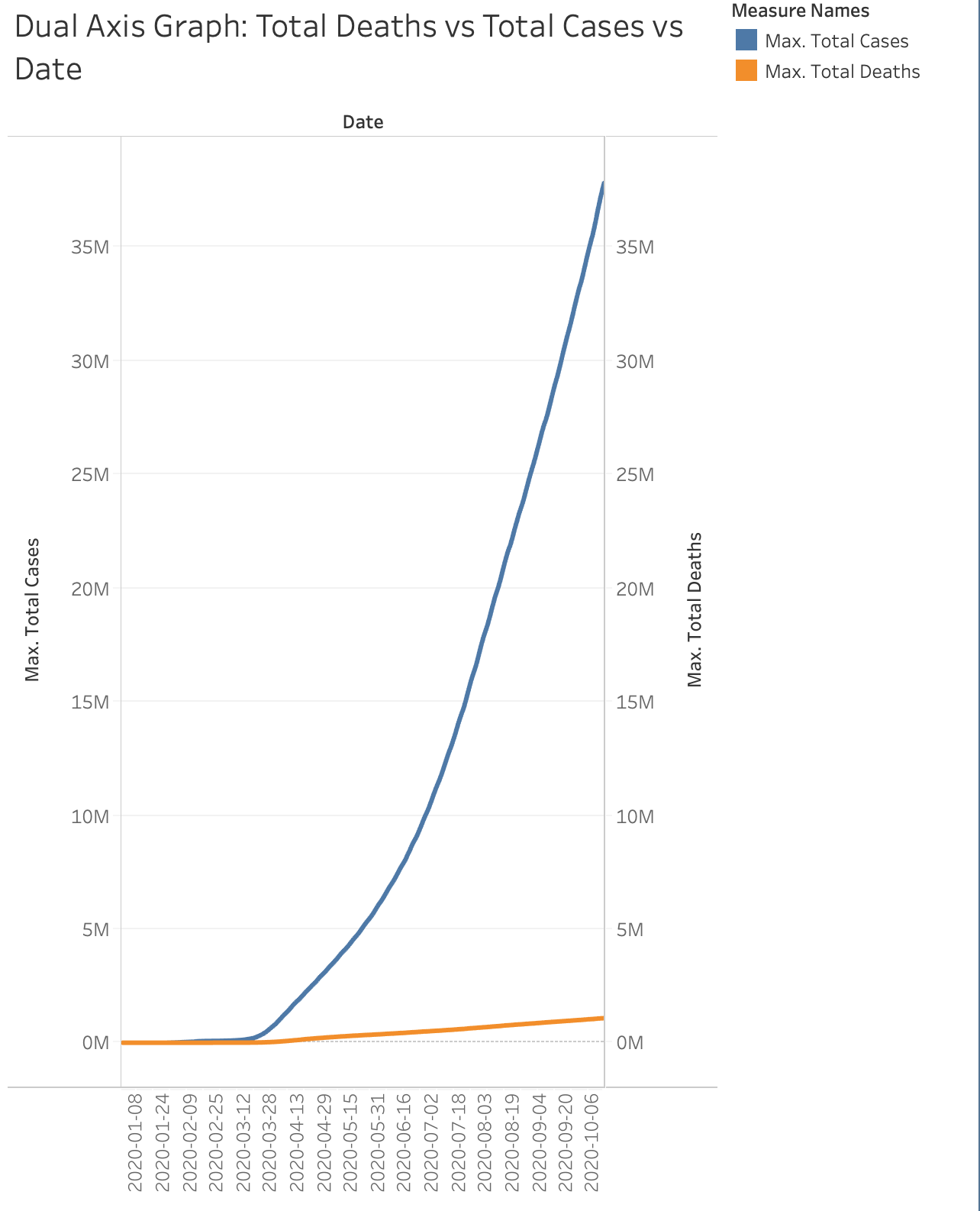
**[18]**

****

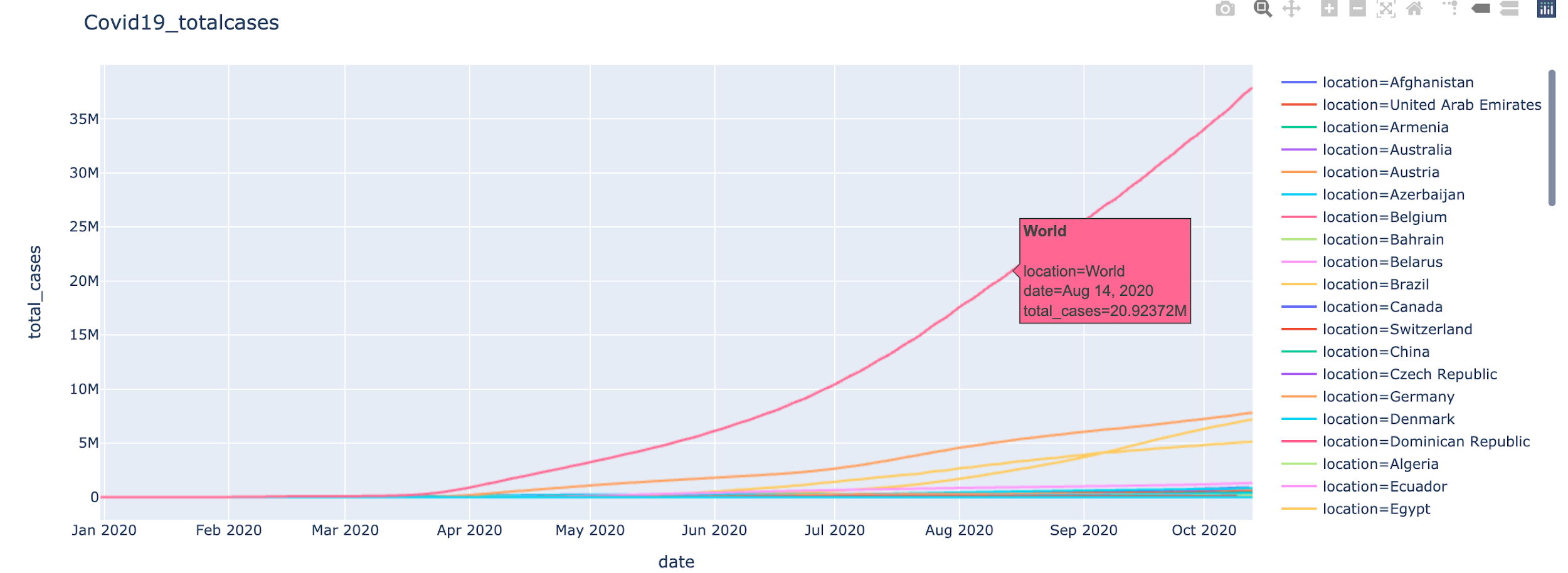
**[19]**

****

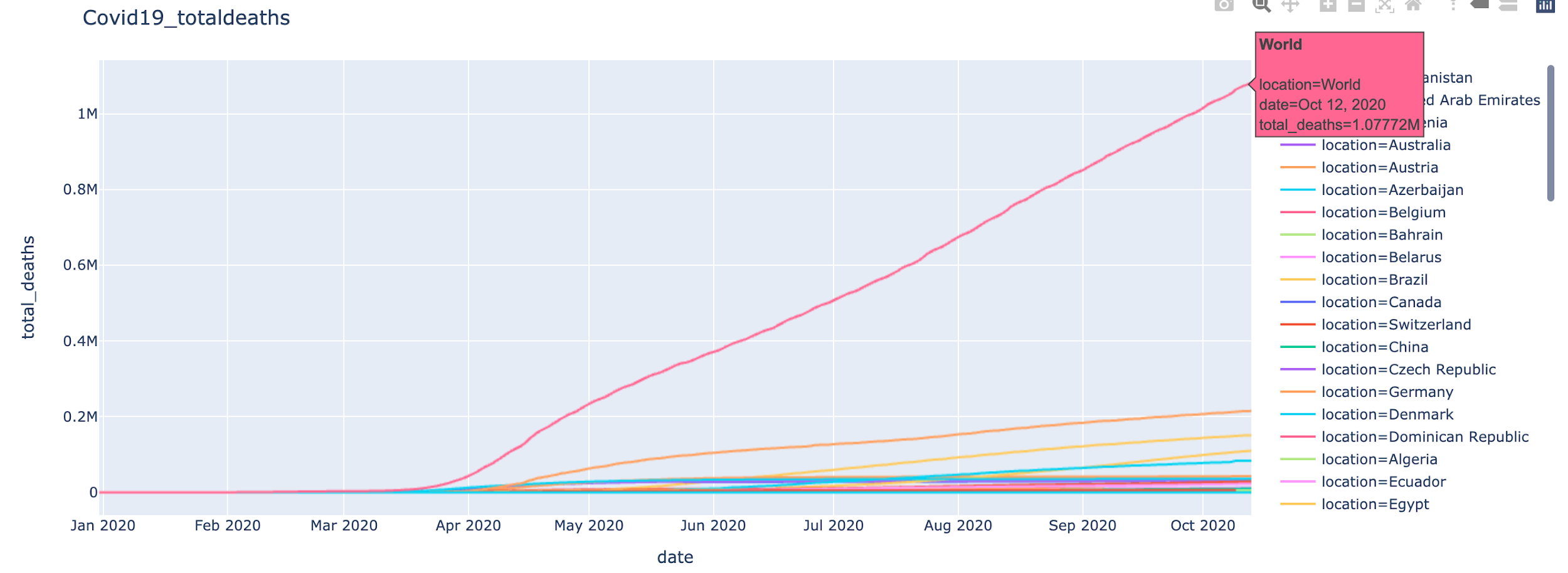
**[20]**

****

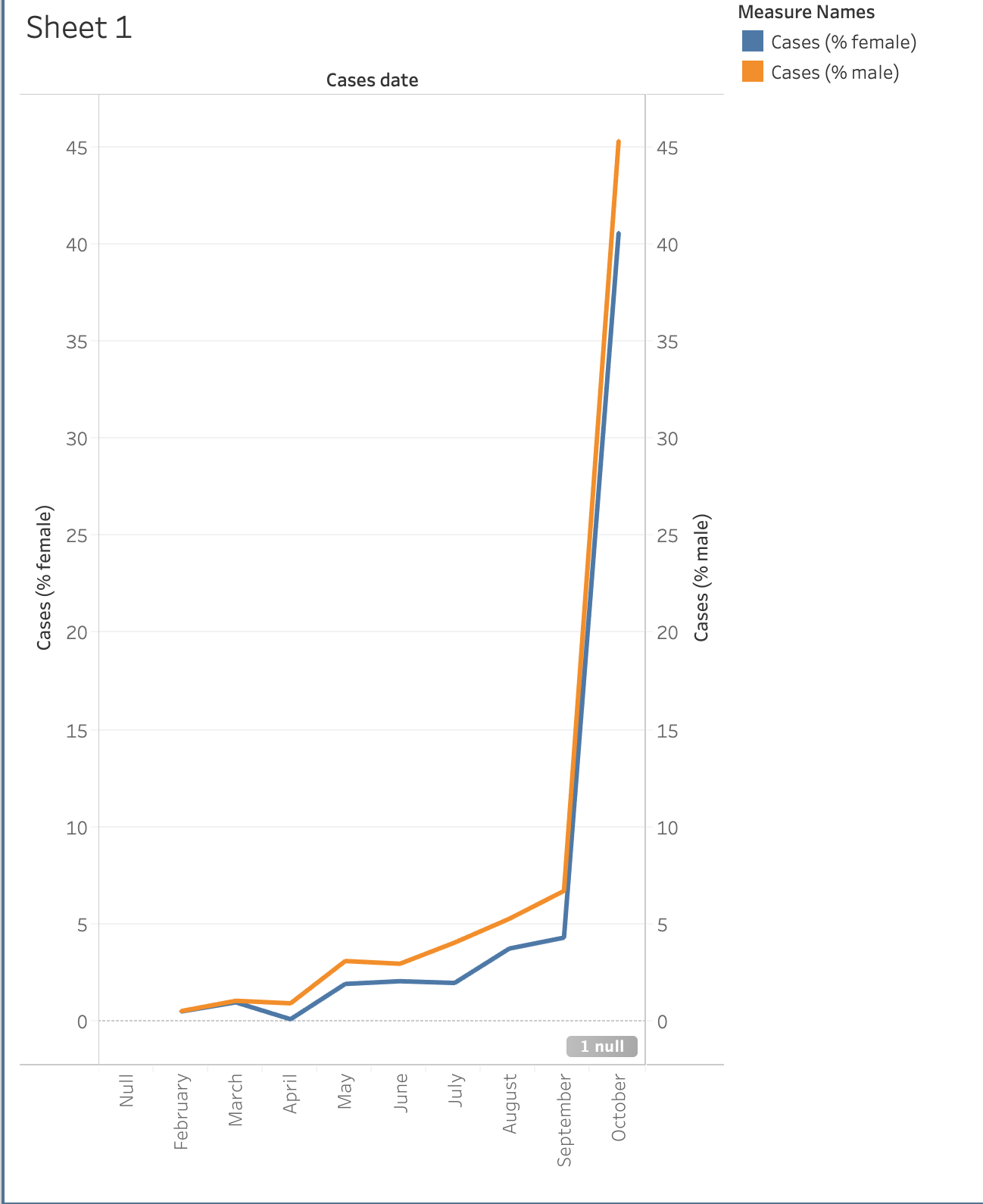
**[21]**

****

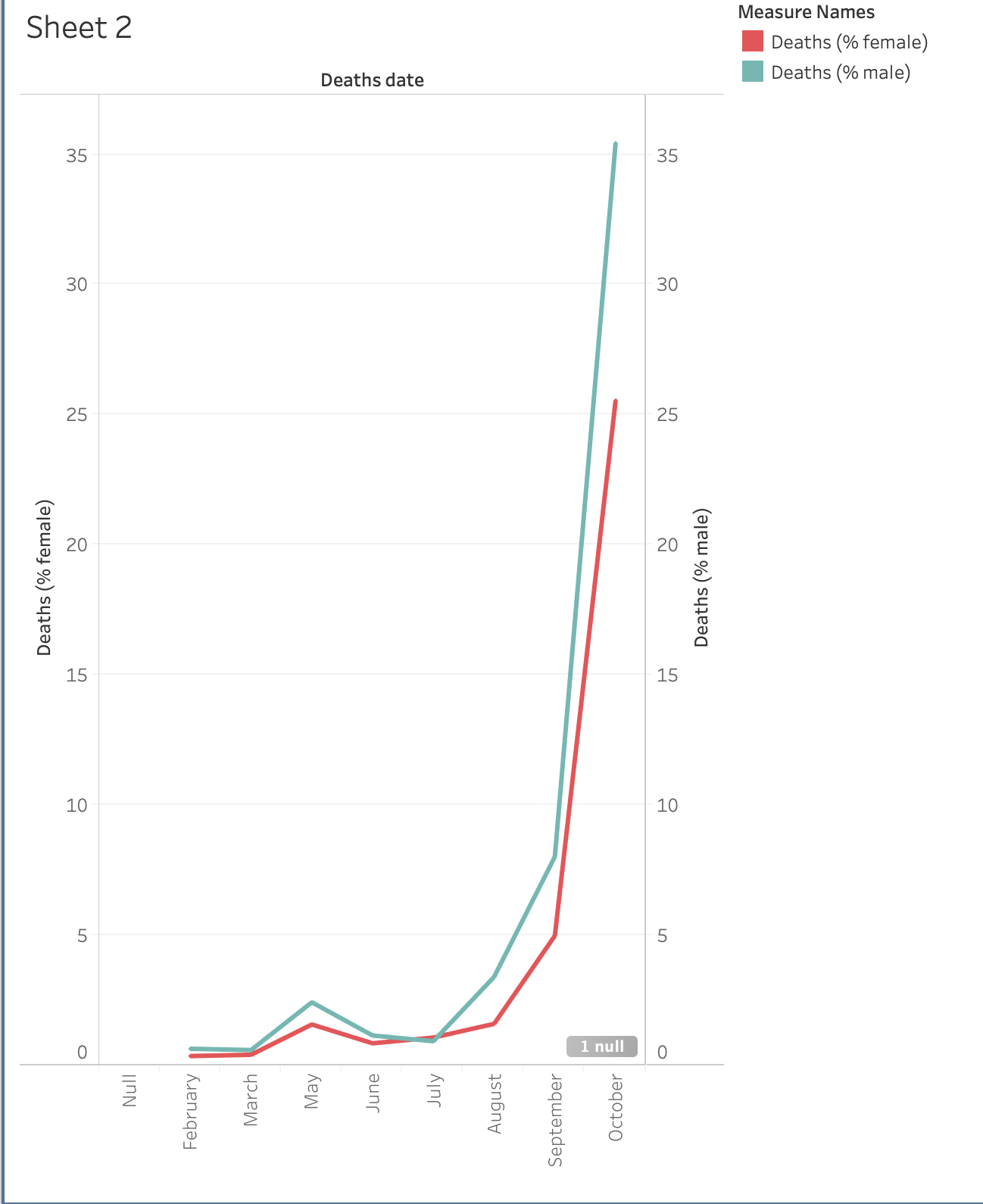
**[22]**

****

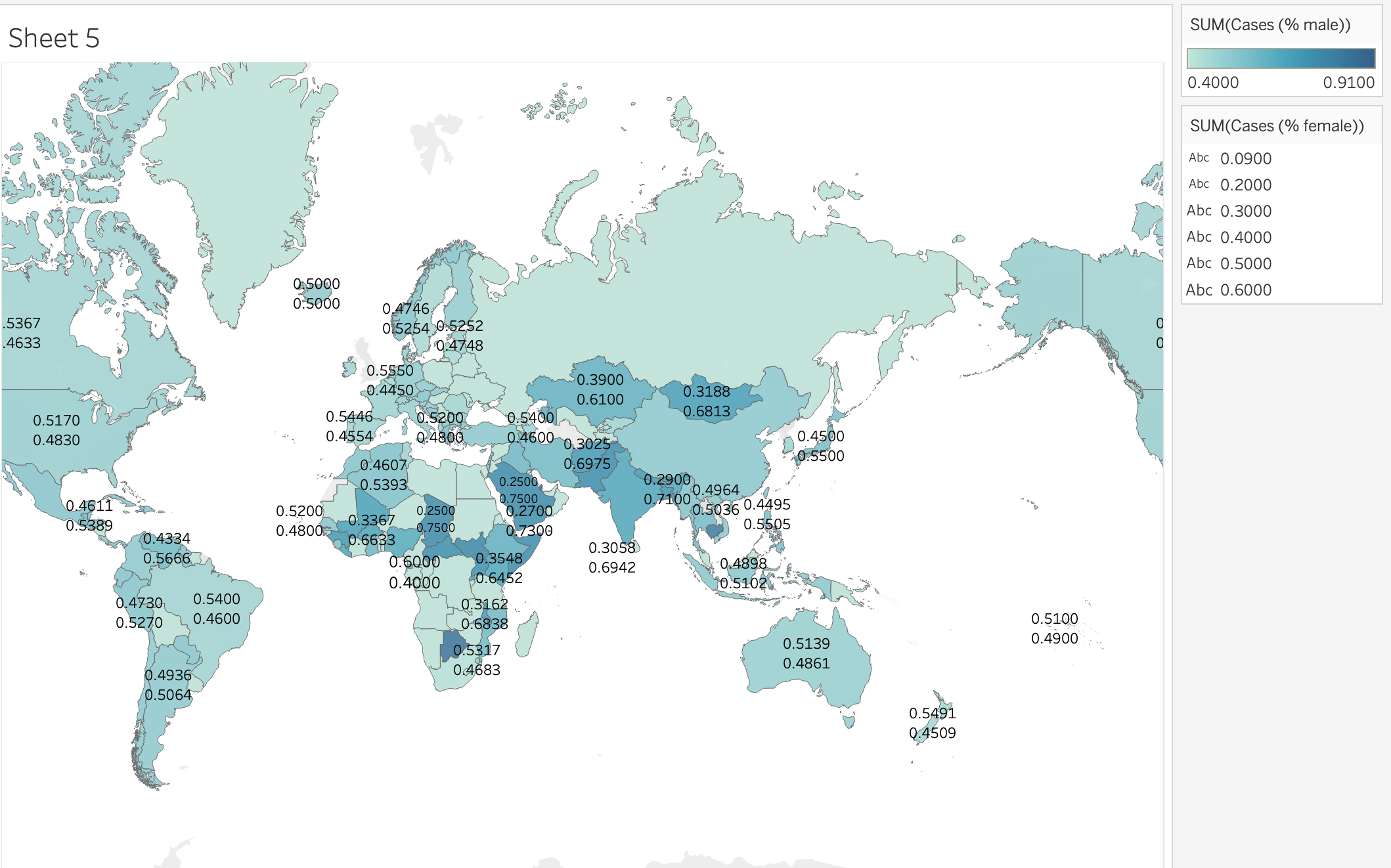
**[23]**

****

**[24]**

****

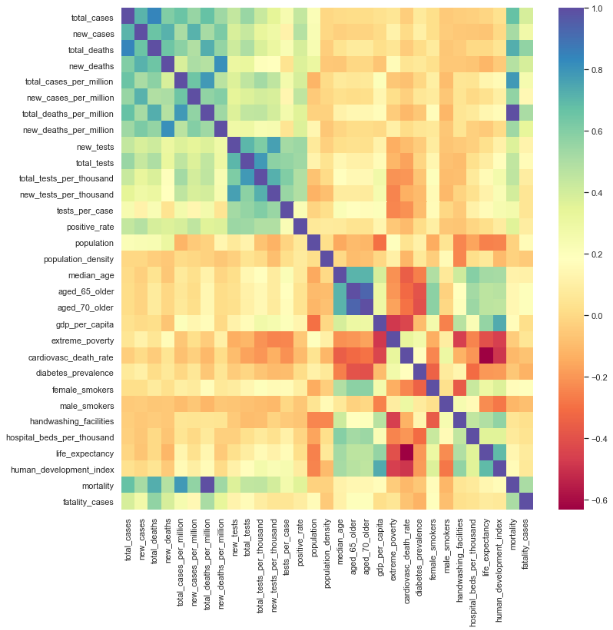
**[25]**

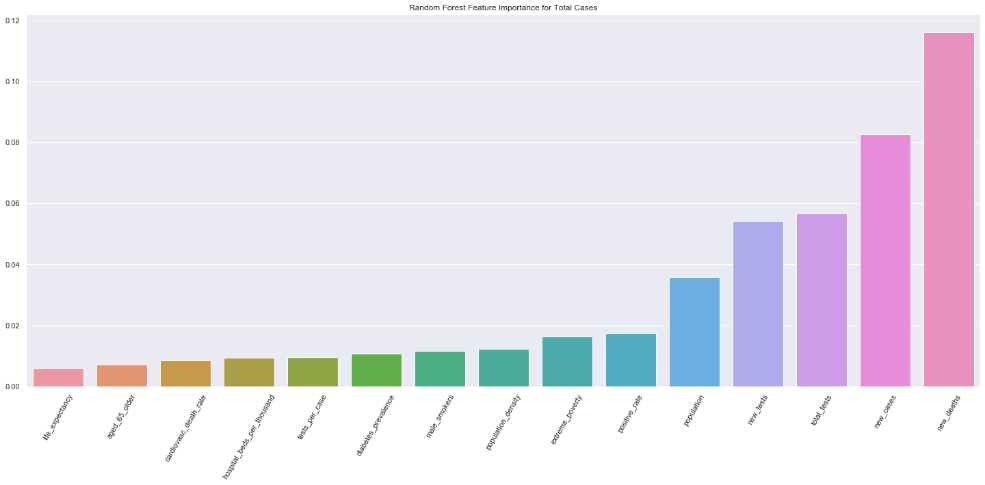
****

**[26]**

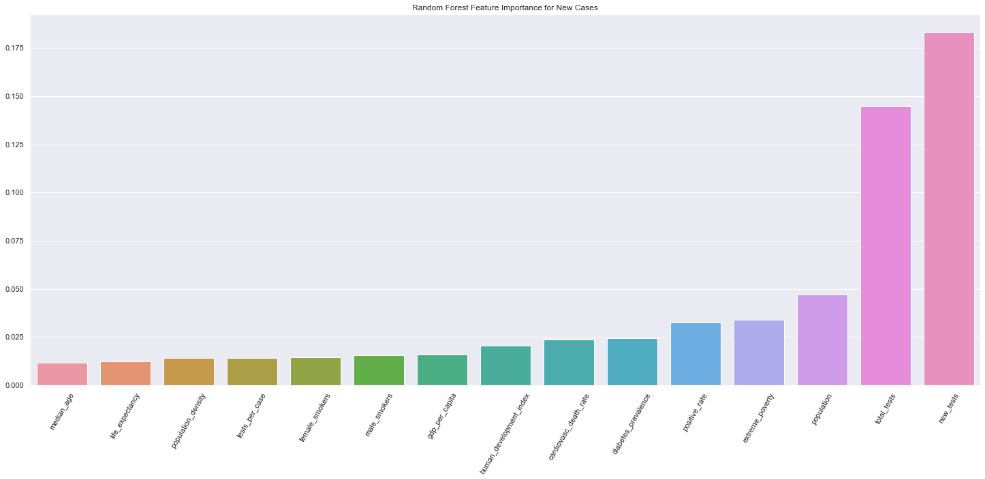
****

[27]

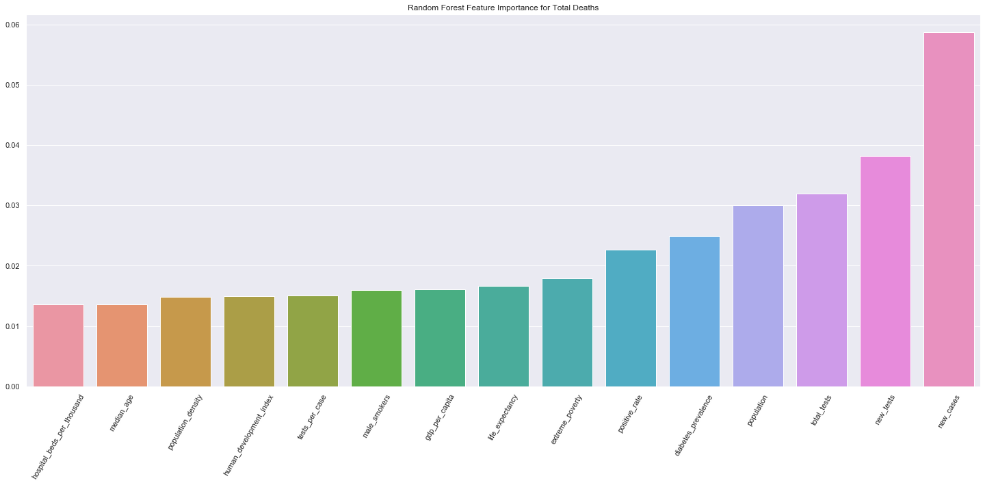


[28]

[29]



[30]



[31]

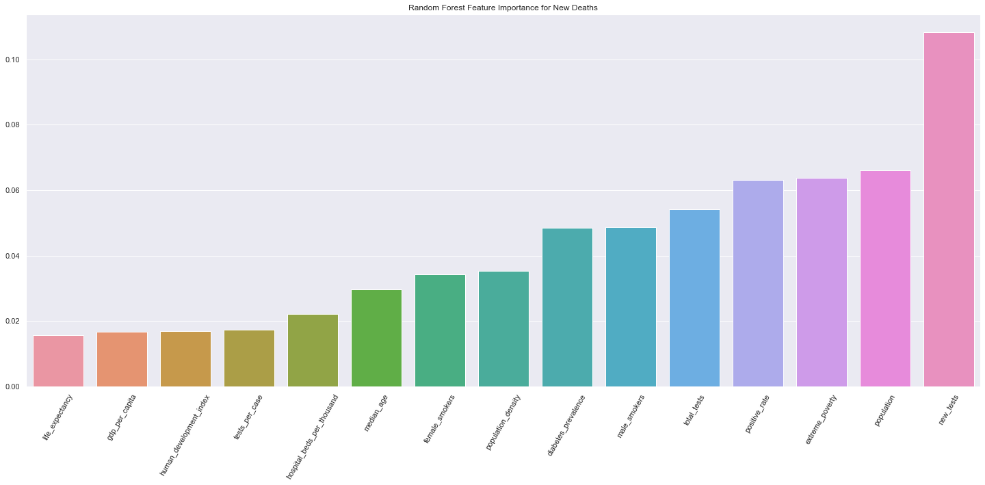


Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Analysis the basic metrics of variables** | **Avani** | **Avani** |
| **2** | **Non-graphical and graphical univariate analysis** | **Avani** | **Avani** |
| **3** | **Missing value analysis and outlier analysis** | **Toby, Avani** | **Michael** |
| **4** | **Feature engineering and analysis** | **Michael** | **Toby** |
| **5** | **Data Visualizations(Tableau,Python)** | **Toby, Michael, Avani** | **Michael** |
| **6** | **Appendix** | **Toby, Avani, Michael** | **Avani, Michael, Toby** |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.