Logo

Description automatically generated

DATA VISUALIZATION REPORT

LAB 01: DATA RELATIONSHIP

GROUP 07

20127449 – Tran Quoc Bao

20127452 – Ho Dang Cao

20127476 – Do Duc Duy

20127567 – Le Nguyen Binh Nam

Instructors

Dr. Bui Tien Len

Dr. Le Ngoc Thanh

Table of Contents

[EVALUATIONS 1](#_Toc130066201)

[LEVEL OF COMPLETION OF REQUIREMENTS 1](#_Toc130066202)

[LEVEL OF COMPLETION OF MEMBERS 1](#_Toc130066203)

[DETAILED ALGORITHM 2](#_Toc130066204)

[DATA CRAWLING AND REPROCESSING 2](#_Toc130066205)

[DATA CRAWLING 2](#_Toc130066206)

[PRE-PROCESSING 4](#_Toc130066207)

[Store data and reuse 5](#_Toc130066208)

[VISUALIZATION 5](#_Toc130066209)

[A. Correlation between 2 attributes 5](#_Toc130066210)

[B. Overview of the epidemic situation in the world and continent 9](#_Toc130066211)

[C. Analysis of epidemic situation in continents 17](#_Toc130066212)

[D. Analyzing top 10 countries 27](#_Toc130066213)

[References 33](#_Toc130066214)

# EVALUATIONS

## LEVEL OF COMPLETION OF REQUIREMENTS

|  |  |  |  |
| --- | --- | --- | --- |
| ID | TASK | MEMBER | LEVEL |
| 1 | data crawling, pre-processing | 20127452 | 100% |
| 2 | scatter chart - correlation between 2 attributes | 100% |
| 3 | radar chart - attributes comparison between continents | 100% |
| 4 | % bar chart - % of attributes of each continent | 100% |
| 5 | donut chart - % cases contribution of continents | 100% |
| 6 | bar chart - reviews 4 attributes | 20127567 | 100% |
| 7 | scatter chart - correlation of 2 attributes | 100% |
| 8 | histogram and scatter chart - relationship between 3 attributes | 100% |
| 9 | scatter chart - correlation of 2 attributes | 100% |
| 10 | bar chart - test percentage in each continent | 20127476 | 100% |
| 11 | bar chart - percentage of infection case in each continent | 100% |
| 12 | bar chart - percentage of death in each continent | 100% |
| 13 | bar chart - relation between percentage of infection per person  and percentage of testing per person | 100% |
| 14 | bar chart - relation between percentage of infection per person  and percentage of death per person | 100% |
| 15 | bar chart - countries with the highest number of infections | 100% |
| 12 | Choropleth maps- visualize total cases of countries | 20127449 | 100% |
| 13 | scatter chart - distribution total case vs total deaths of countries | 100% |
| 14 | scatter chart - distribution total case vs total recovered of countries | 100% |
| 15 | bar chart - total cases, total deaths, total recovered of continents | 100% |
| 16 | bar chart - the situation of covid 19 in the world | 100% |

## LEVEL OF COMPLETION OF MEMBERS

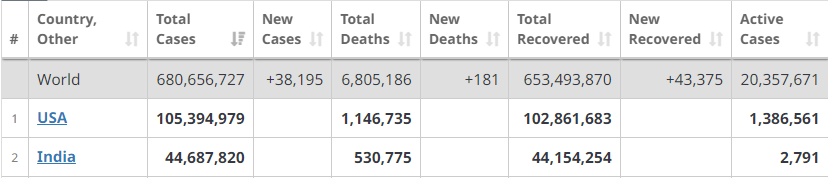
|  |  |  |
| --- | --- | --- |
| ID | FULL NAME | LEVEL |
| 20127449 | TRAN QUOC BAO | 100% |
| 20127452 | HO DANG CAO | 100% |
| 20127476 | DO DUC DUY | 100% |
| 20127567 | LE NGUYEN BINH NAM | 100% |

# DETAILED ALGORITHM

## DATA CRAWLING AND REPROCESSING

### DATA CRAWLING

In this section, we will talk about the process of getting data from the data table in the Worldometer website ([www.worldometers.info/coronavirus/](http://www.worldometers.info/coronavirus/)).



The algorithm will use 3 main libraries/modules, specifically:

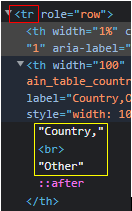
- Use selenium's webdriver function to access and get the HTML code of the website.

- Use BeautifulSoup to extract data by tag from the obtained HTML code.

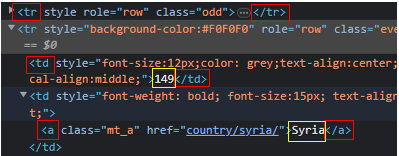
- Use re to extract the names of the data's attributes.

Observe that the data is all in the 'tr' tags. Each 'tr' tag corresponds to 1 object (1 row) in the data table. In there:

- The value of each attribute name cell is in the 'th' tag, but there is a 'br' tag in the middle.



- The value of each data cell is in the 'td' and 'a' tags.



We'll ignore hidden rows - lines with style="display: none" attribute and we take only the first 233 rows, which corresponds to the row number of the data table.



After getting the html code containing the rows in the data table, we divide it into two parts for different processing: the attribute name row and the value table.

* Attribute row:
  + Use re.findall with the syntax: >([\w\d,/ ]\*)< to get a list containing the property names as follows:



* + Due to the 'br' tag, attribute names are split into contiguous elements. Then concatenate adjacent (other than ' ') elements to get the final list:

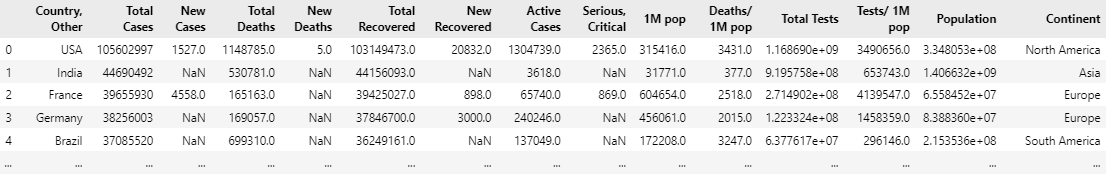


* Value table:
  + We filter out the cell data on each row by finding all the 'td' and 'a' tags.
  + Get the last data on each row using the get\_string function.
  + Illustrated problems to be handled on each row:



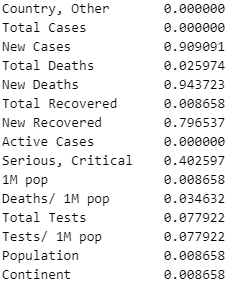
* + The get\_string function includes the following steps:
    - remove id value
    - remove redundant empty value
    - remove duplicated country name
    - get pure data by string attribute of Selenium
    - Replace ' ' (space characters) by nan values.
    - Astype numeric values from string to integer.

The crawled data is put into a data frame:

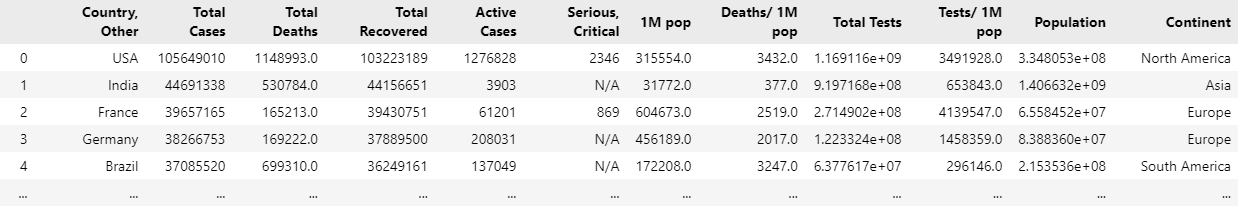
****

### PRE-PROCESSING

Observe missing ratio of attributes:



**Comment:** Because the covid situation is no longer serious, the attributes such as New Cases, New Deaths, New Recovered are not updated regularly (have high missing ratio). They wil be dropped.



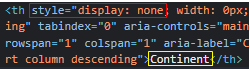
### Store data and reuse

The data frame is saved as .csv file daily.

Read the file to visualize later.

**Note that**: our data frame has 231 rows × (14 + 1) columns.

The extra column comes from the Continent attribute, which is hidden in the website.



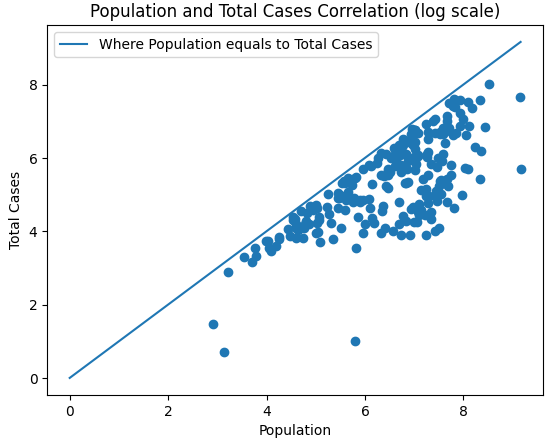
## VISUALIZATION

### Correlation between 2 attributes

We built a function to display the correlation between 2 given properties by scatter chart. The data disparity between countries is so large that the data is log scaled before visualization.

Also, visualize the mean line where the points have x coordinate equal to y coordinate to compare the magnitude of the 2 attributes.

#### Correlation between Population and Total Cases

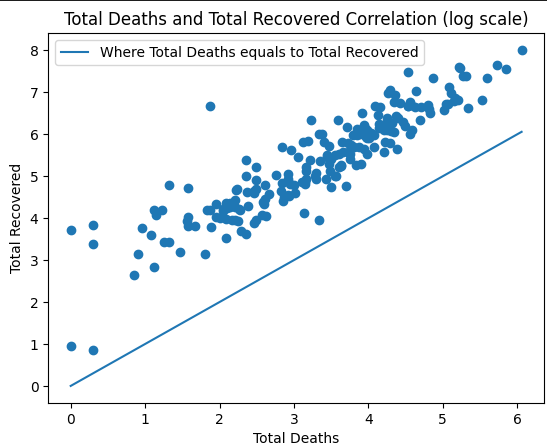


#### Comments

The higher Population, the higher Total Cases (they have positive linear correlation).

Many countries have high rates of covid infection (data points lie near the mean line).

#### Correlation between Total Deaths and Total Recovered

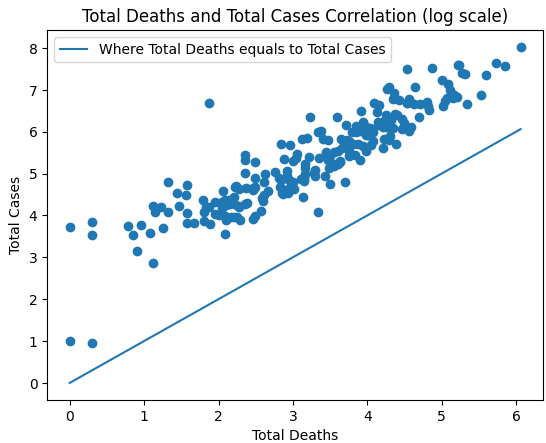


#### Comments

All countries have more recovered cases than death cases (all points are above mean line).

Recovered cases and death cases have positive linear correlation.

#### Correlation between Total Deaths and Total Cases



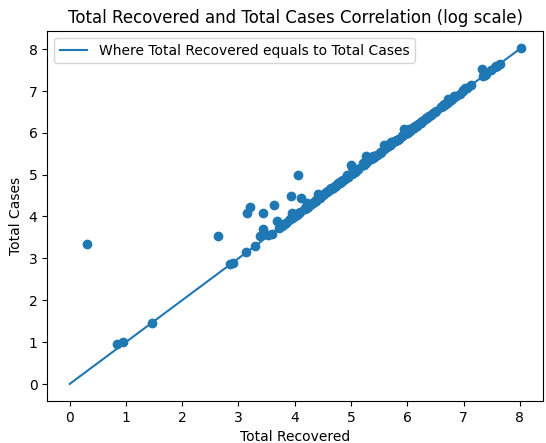
#### Comments

There doesn't seem to be a difference with the previous chart.

All points are above the mean line.

The higher Total Cases, the higher Total Deaths (positive linear correlation).

#### Correlation between Total Cases and Total Recovered



#### Comments

Most of the data points lie on the mean line that means most covid cases will recover.

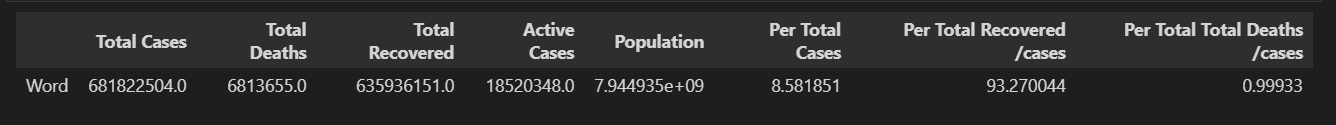
The higher Total Cases, the higher Total Recovered (highly positive linear correlation).

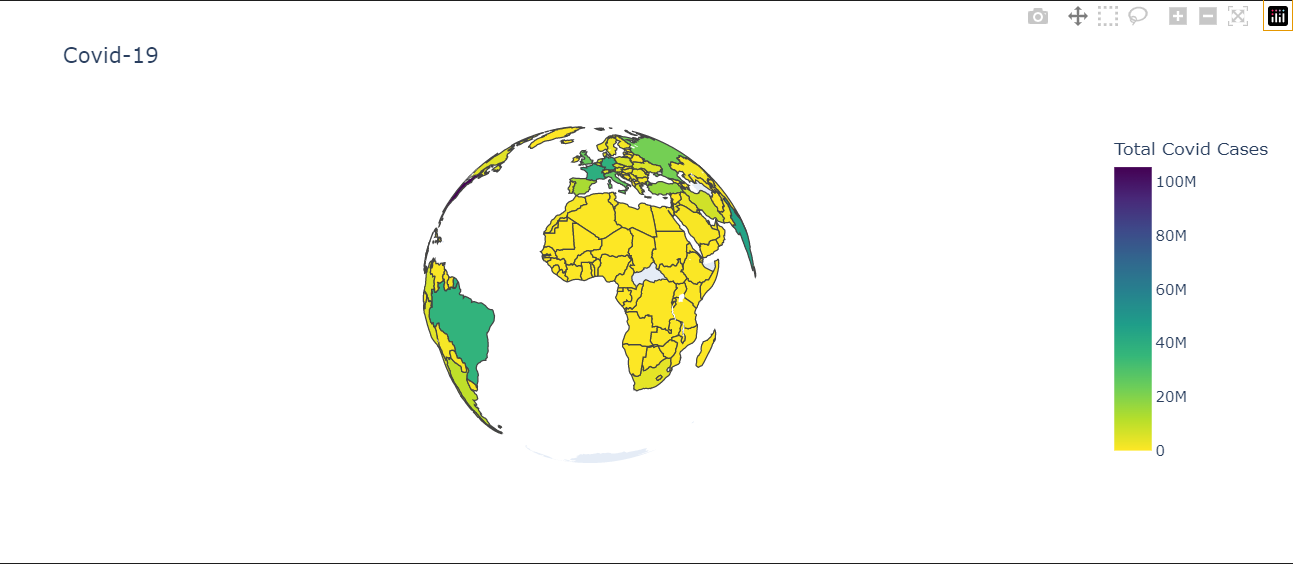
### Overview of the epidemic situation in the world and continent

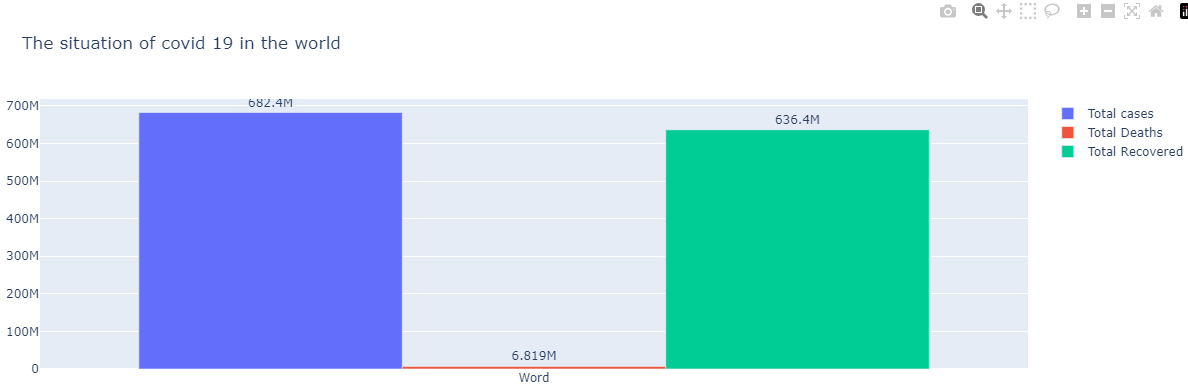
After analyzing the correlation between the attributes, let's take a look at the situation of the whole world and the continents

#### Covid 19 situation in the world

Display the world's total cases, total death and total recovered data through bar graph and Choropleth maps and analysis



**Chart**

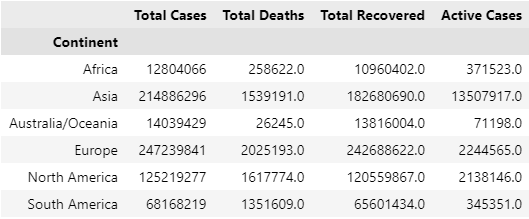


**Comment**

The largest total number of cases in a country is over 100 million.

The world has 682.4M cases of covid including 636.4M recoveries (accounted for 93.26%) and 6.819M deaths (accounted for 1%).

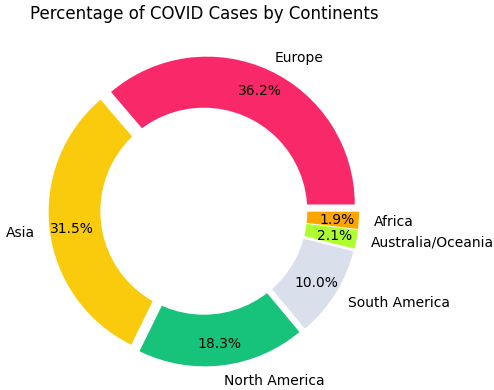
After visualizing the whole world, we recalculate attributes Total Cases, Total Deaths, Total Recovered and Active Cases to observe each continent.

******

#### Percent of cases contributed by each continent

Using a donut chart to show percentage is very suitable, arc length corresponds to percentage of cases contributed by each continent.

#### Chart



#### Comments

There is a large disparity between the largest % Total Cases (Europe - 36.2%) and the smallest % Total Cases (Africa - 1.9%). Meanwhile, Europe is the continent with the most advanced health in the world, and Africa is less developed and has a larger population.

Is the difference between the number of recorded Total Cases and the actual number of Total Cases in Africa too big?

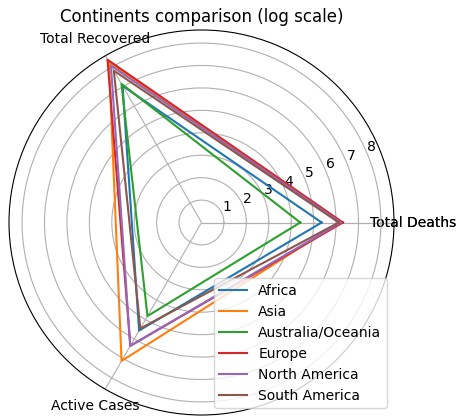
Asia is the origin of the disease and the most populous in the world, but it also ranks second in the chart with 31.5%.

#### Attributes comparison between continents

To compare continents on 3 different attributes (Total Deaths, Total Recovered and Active Cases), we use a radar chart. The area of the triangle created by the 3 attributes is correlated with the Total Cases.

The data disparity between attributes is so large that the data is log scaled before visualization.

#### Chart



#### Comments

The epidemic in Asia is still the most intense (highest Active Cases), followed by Europe and North America; Africa and South America are close together; Australia/Oceania is the lowest.

Total Deaths of Europe, Asia, North America and South America is quite close.

There is a big difference in the Total Deaths in Australia/Oceania compared to other continents.

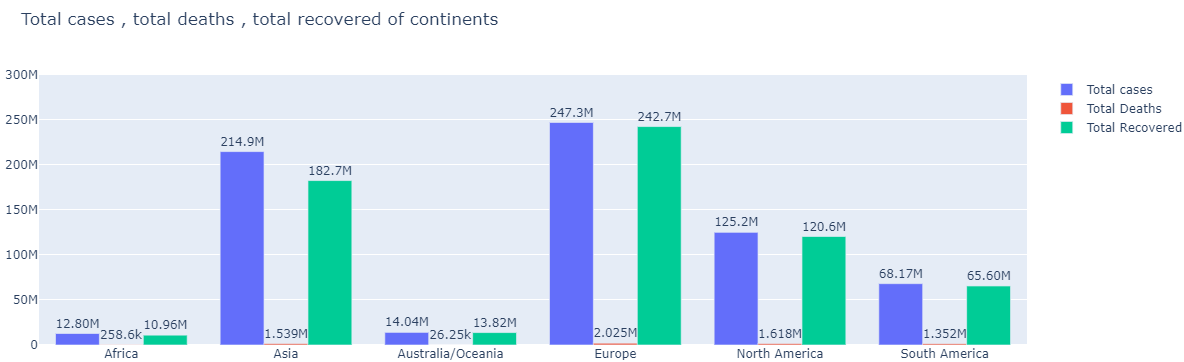
Since there are nearly equal Total Cases, the Total Deaths is lower in Australia/Oceania than in Africa, the Total Recovered in Australia/Oceania is higher than in Africa.

Total Recovered in Asia and North America is also a reversal compared to the order in the Total Deaths. The remaining continents still keep their respective order in the Total Deaths

#### Total cases, total deaths, total recovered of continents

Display data on total cases, total death and total recovered of continents, thereby giving a brief overview of the covid19 situation in continents

**Charts**



**Comment**

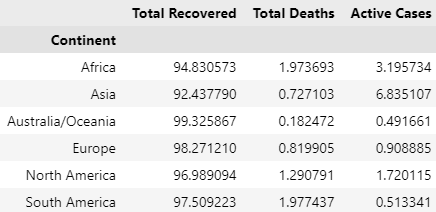
Europe is the country with the highest number of covid cases among the continents

The continent with the most total deaths is Europe

Looking at the chart, we see that most continents have total recovered close to total cases, but in Asia we see that these two columns have a clear difference. This may indicate that treating covid patients in Asia is very difficult

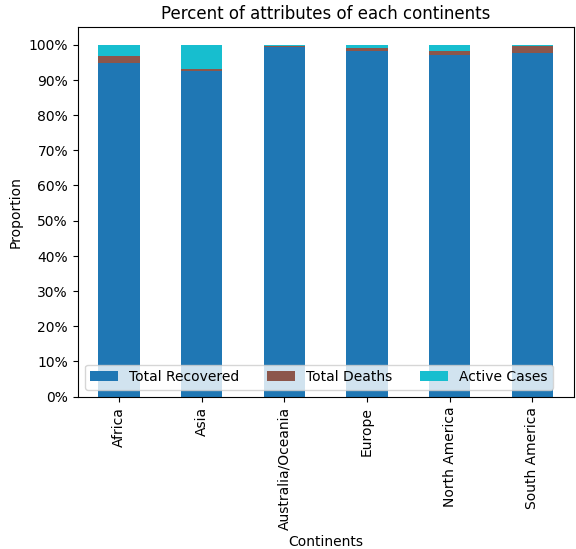
#### Percent of attributes of each continent

We recalculate attributes Total Cases, Total Deaths, Total Recovered and Active Cases to observe each continent, but at this time, we drop nan values before calculating, it makes sure that sum of percentage is 100%.



We use a percentage bar chart with equal column heights (equal to 100%) representing Totals Cases across continents.

#### Chart



#### Comment

The Recovery rate is high in all continents.

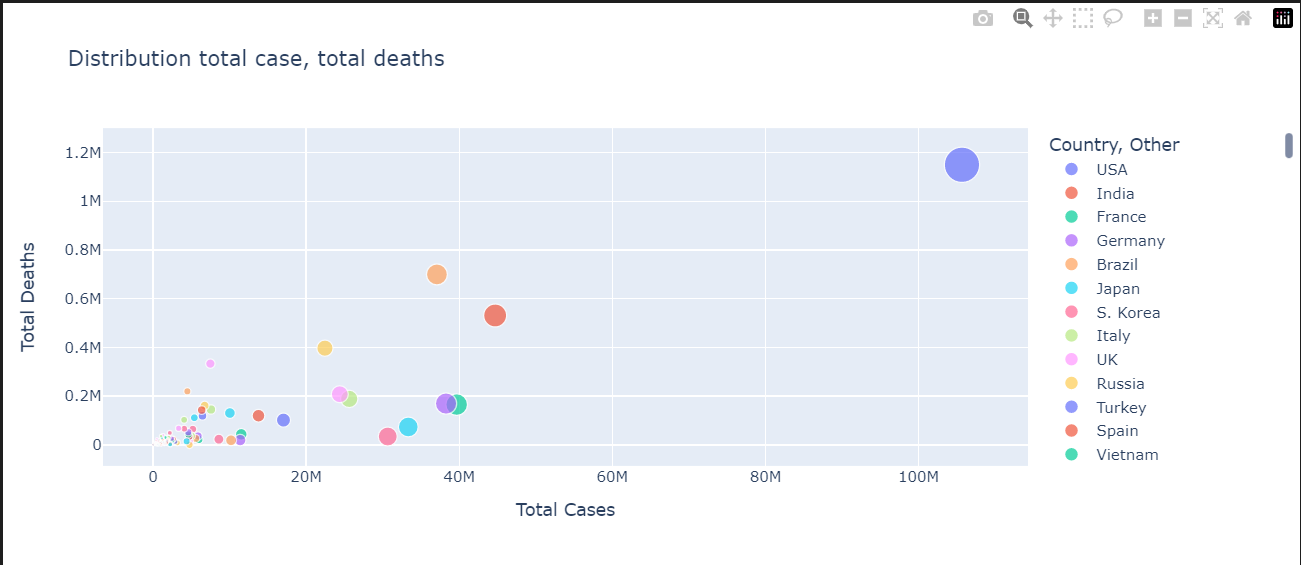
Asia and Africa still have high active cases (Asia is the highest one in countries such as Vietnam, Russia, Taiwan, China, etc).

Australia/Oceania seems to be dealing with Covid epidemic most effectively.

Europe deals with Covid epidemic well, Despites of having the highest Total Cases.

Meanwhile, Africa and South America have the highest Deaths ratio because they have less developed health systems than other continents in general.

#### Distribution total case vs total deaths of countries

****

**Comment**

The country with the most total cases and total Deaths is the US and far ahead of the rest of the countries on the chart.

USA is the first country with total deaths over 1 million people.

#### Distribution total case vs total recovered of countries

****

**Comment**

In most countries, total Case and total recovered values are roughly equal, so the distribution of the data looks like a straight line. If there are data points that are skewed, the total recovered in that country is much lower than the total cases

### Analysis of epidemic situation in continents

#### Ⅰ. Use bar charts to visualize data on: tests, infections, and deaths. To reflect in a general way the epidemic situation in the continents.

**1.1.  Rate of testing per person in each continent**

**Fields using: 'Tests/ 1M pop', 'Continent'.**

**Steps to prepare data for visualizing:**

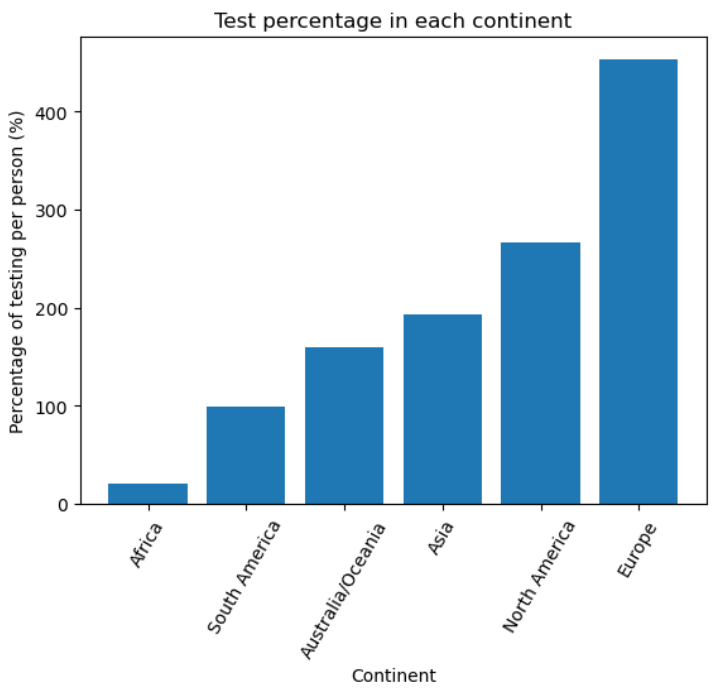
1: Extract 2 fields ‘Tests/ 1M pop’ and ‘Continent’ from dataframe.

2: Group all columns by field ‘Continent’.

3:  Using lambda function to set values of field ‘Tests/ 1M pop’ to percentage.

4. Sort values at field ‘Tests/ 1M pop’ by ascending.

**Result after visualizing:**



**Comment:**

The testing rate in Europe is the highest and in Africa it is the lowest.

There is quite a big gap between the continent with the highest testing rate and the continent with the lowest testing rate, but the continents in the range between the lowest and the highest have not too big gap with each other.

Especially here in Asia, where there is an outbreak of the disease, but the rate of testing in Asia is lower than the whole of Europe and North America.

**1.2. Infection rate per person in each continent**

**Fields using: '1M pop', 'Continent'.**

**Steps to prepare data for visualizing:**

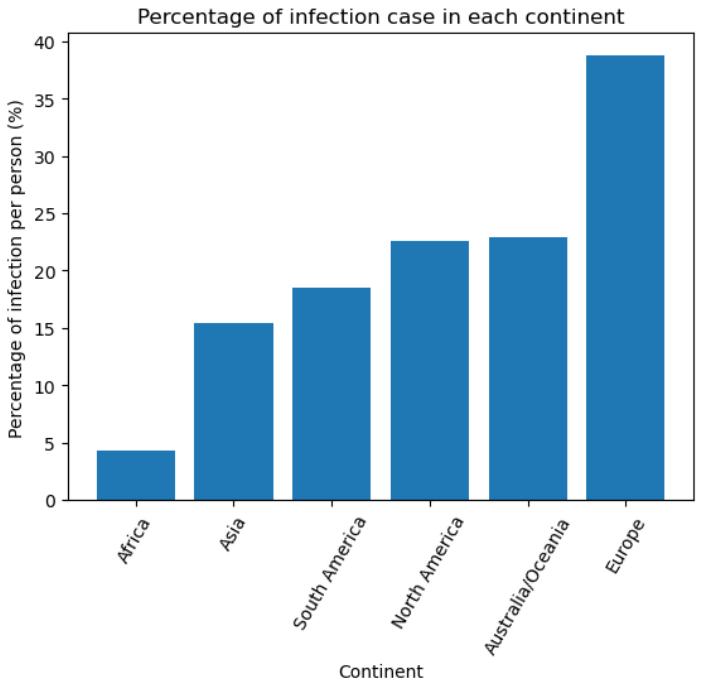
1. Extract 2 fields ‘1M pop’ and ‘Continent’ from dataframe.

2. Group all columns by field ‘Continent’.

3.  Using lambda function to set values of field ‘1M pop’ to percentage.

4. Sort values at field ‘1M pop’ by ascending.

**Result after visualizing:**



**Comment**

The infection rate in Europe leads, and is many times higher than in other continents.

The infection rate in Africa is lowest and many times lower than Europe.

The infection rate in the continents lying between the two continents with the highest and lowest have infection rate increased steadily with a relatively low increase (Asia < South America < North America < Australia/Oceania).

**1.3. The rate of death per person from the disease in each continent**

**Fields using: 'Deaths/ 1M pop', 'Continent'.**

**Steps to prepare data for visualizing:**

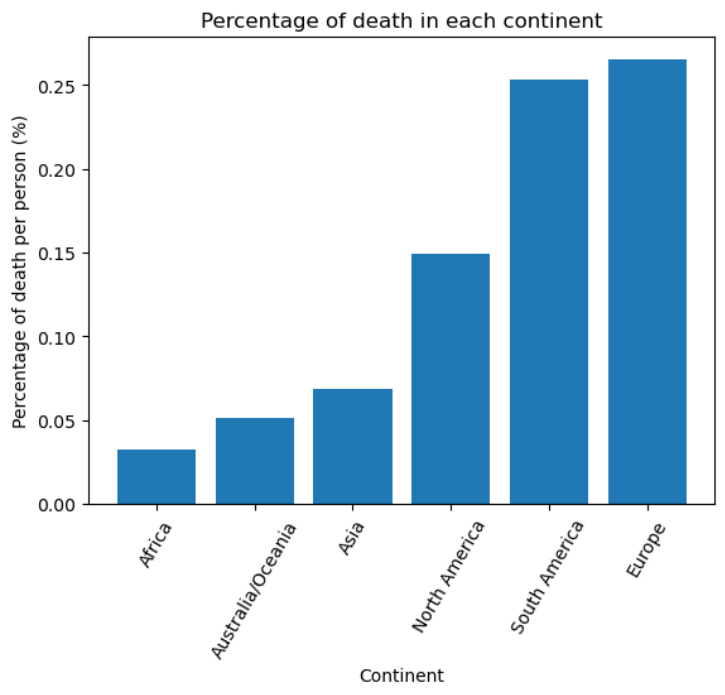
1. Extract 2 fields ‘Tests/1M pop’ and ‘Continent’ from dataframe.

2. Group all columns by field ‘Continent’.

3. Using lambda function to set values of field ‘Tests/ 1M pop’ to percentage.

4. Sort values at field ‘Tests/ 1M pop’ by ascending.

**Result after visualizing:**



**Comment**

The continents in the group with high death rate are Europe, South America and North America.

The continents in the group with low death rate rates are Africa, Australia/Oceania.

Europe is the place with the highest death rate from infection, the second highest death rate is South America, and the lowest death rate is Africa. A special feature here is when Asia is the outbreak of the disease but has a low death rate in the group of continents with low death rate (Africa, Australia/Oceania, Asia).

There is a large disparity between the group of continents with low death rate and the group of continents with high death rate.

#### Ⅱ. Visualize the relationship about rate of testing with rate of incidence and rate of incidence with rate of death.

**1.1. The relationship between the testing rate and the incidence rate per person in each continent.**

* **Fields using: 'Tests/ 1M pop', '1M pop', 'Continent'.**
* **Steps to prepare data for visualizing:**

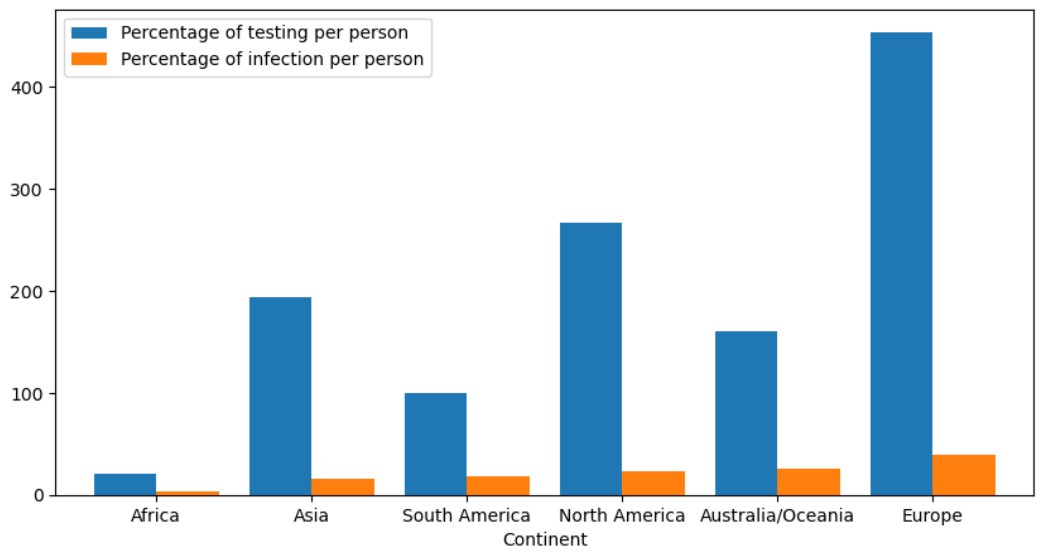
1. Extract 3 fields ‘Tests/ 1M pop’, ‘1M pop’ and ‘Continent’ from dataframe.

2. Group all columns by field ‘Continent’.

3. Use lambda function to set values of field ‘Tests/ 1M pop’ and ‘1M pop’ to percentage.

4. Sort values at field ‘1M pop’ by ascending.

* **Result after visualizing:**



**Comment:**

There is an unequal distribution between percentage of testing per person and percentage of infection per person in different continents, the percentage of testing per person in Asia is higher than South America, but percentage of infection per person in Asia is lower than South America. Therefore, these two rates have different proportions in each continent, but in general terms, as the percentage of testing per person increases, so does the percentage of infection per person.

The percentage of death per person is many times lower than the percentage of infection per person.

**1.2. The relationship between incidence and death rate per person in each continent.**

* **Fields using: '1M pop', 'Deaths/ 1M pop', 'Continent’.**
* **Steps to prepare data for visualizing:**

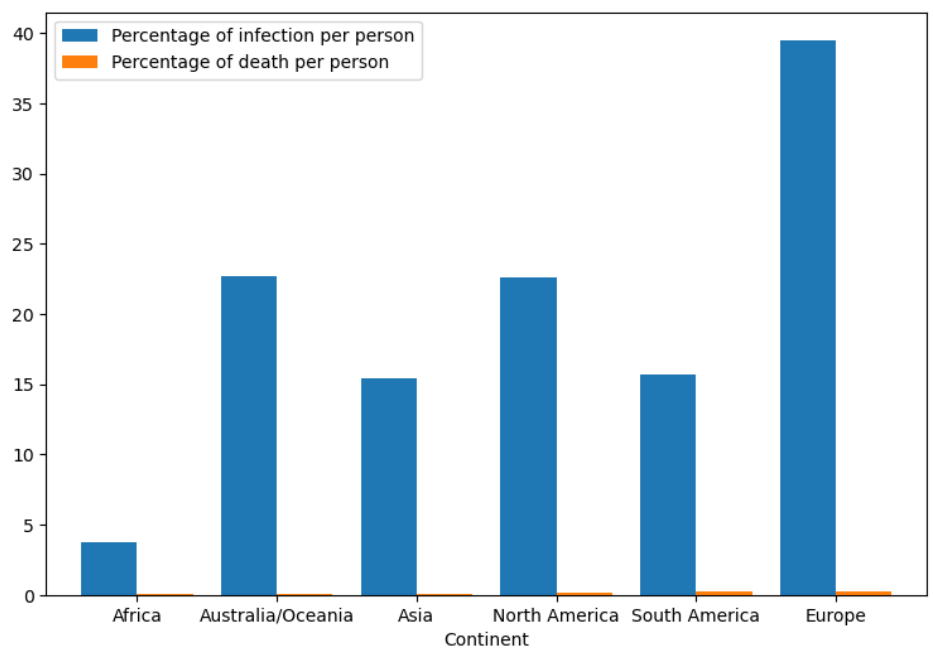
1. Extract 3 fields ‘1M pop’, ‘Deaths/ 1M pop’ and ‘Continent’ from dataframe.

2. Group all columns by field ‘Continent’.

3. Use lambda function to set values of ‘Deaths/1M pop’ and ‘1M pop’ to percentage.

4. Sort values at field ‘Deaths/ 1M pop’ by ascending.

* **Result after visualizing:**



**Comment**

There is an uneven distribution of percentage of death per person and percentage of infection in different continents, in particular in Australia/Oceania where the percentage of infection per person is higher than in Asia but the percentage of death per person in Australia/Oceania is lower than Asia. Therefore, these two rates have different proportions in each continent, but in general, the increased the percentage of infection rate per person, the percentage of death per person also increases.

The percentage of death per person is many times lower than the percentage of infection per person.

#### Ⅲ. Visualization of infection data of a number of prominent countries.

1. **Top 5 countries with the highest percentage of infected populations in the world.**

* **Fields using: 'Country, Other', 'Total Cases', 'Population', 'Continent'.**
* **Steps to prepare data for visualizing:**

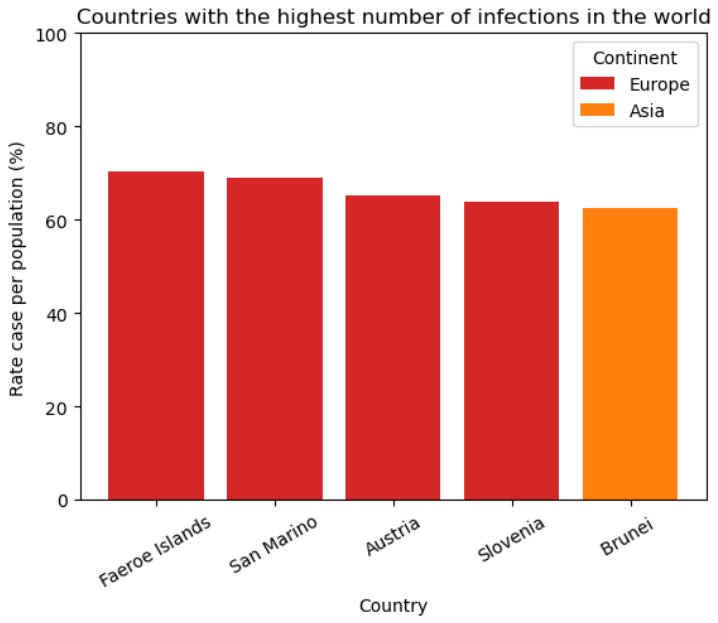
1. Extract 4 fields 'Country, Other', 'Total Cases', 'Population' and 'Continent'from dataframe.

2.  Set a new column with named ‘Rate case per population (%)' by getting values at column 'Total Cases' \* 100 divide values at column ‘Population’.

3. Sort values at field ‘Rate case per population (%)' by descending.

4. Choose the top 5 at head.

* **Result after visualizing:**



**Comment:**

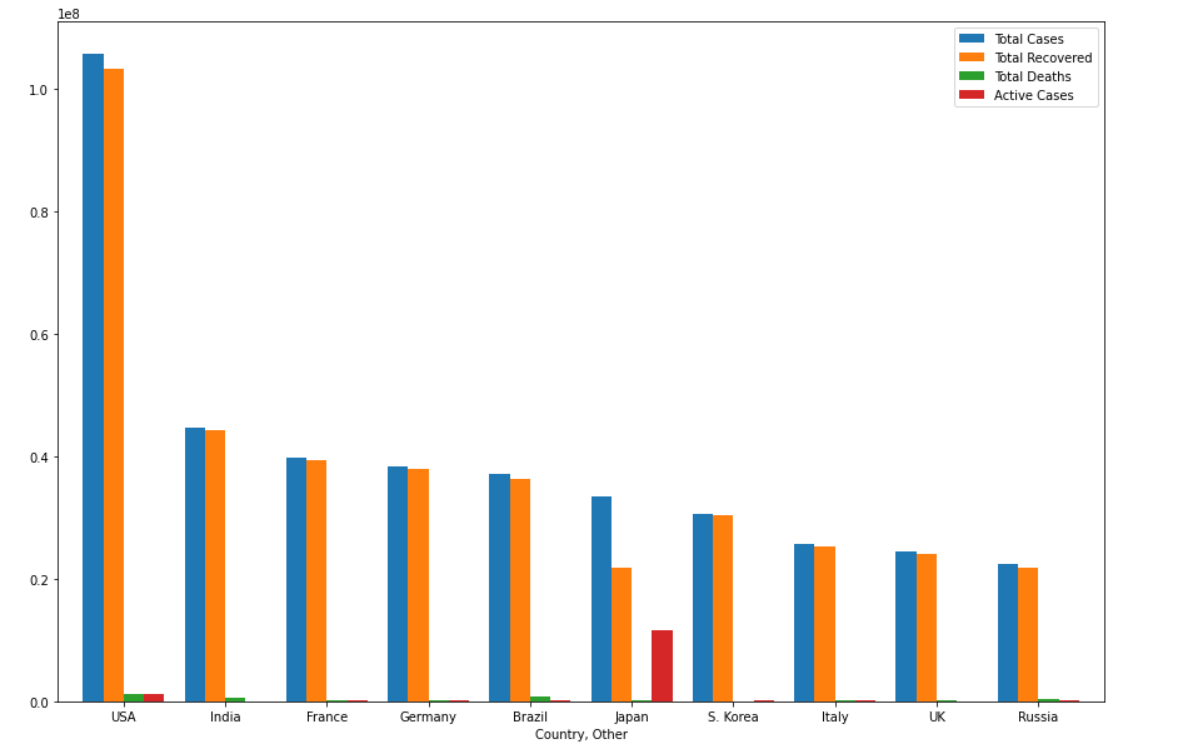
The top 5 countries with the highest percentage of infected populations with 4 countries in Europe and 1 country in Asia, the rate of population infected in the top 5 countries is quite high in the range of 62% to 70%.

The WHO recommends that when the proportion of the population infected with COVID-19 exceeds 10% of the total population of a particular area, the health system in that area may become overwhelmed and unable to provide adequate supplies to provide medical services to people with COVID-19 and other illnesses. This shows that the epidemic situation in these countries is quite serious, which does not exclude the possibility that reinfection cases exist in the data, but the number of such populations is quite high.

  This could potentially lead to an overload of the medical department.

### Analyzing top 10 countries

#### Reviews of Total Cases, Total Recovered, Total Deaths, Active Cases:



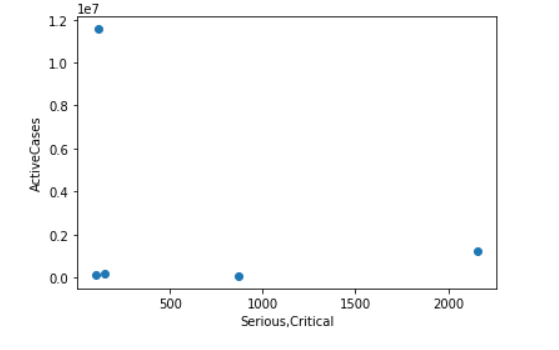
***Comment:***

The speed of disease recovery of all countries is very good ('TotalRecovered' is approximately equal to 'TotalCases'), all countries do not have any significant difference.

USA and Japan are countries that have a large percentage of cases under treatment (ActiveCases / TotalCases) in the top 10, but the most prominent is only Japan, it seems that the epidemic situation is still tense here. But in general, all countries are controlling the epidemic (the ratio of ActiveCases / TotalCases is low).

The overall mortality rate is low, indicating that the disease is not fatal

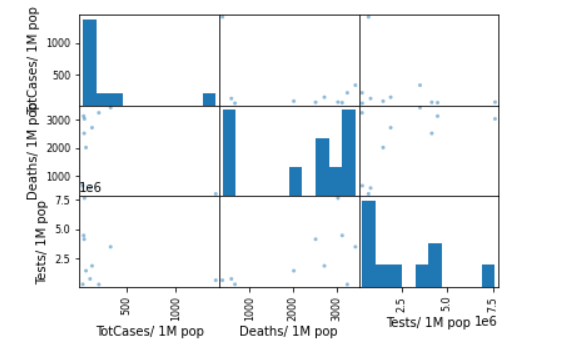
#### Correlation of Serious, Critical vs Active Cases:

****

***Comment***

We see two variables 'Serious, Critical' and 'ActiveCases' have no relationship with each other. But the chart still shows that there are still sick and severe cases, but not too many.

#### Relationship between Total Cases/1M pop, Total Deaths/1M pop, Total Tests/1M pop:

****

***Comment:***

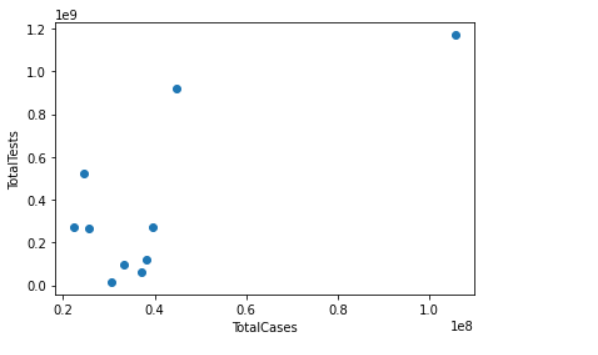
The histograms of 'TotCases/1M pop', 'Tests/1M pop', 'Deaths/1M pop' are not normally distributed.

Histograms of all 3 properties tend to increase, but not too much.

There is a positive relationship between 'TotCases/1M pop' and 'Tests/1M pop', that is, if 'Tests/1M pop' increases, 'TotCases/1M pop' also increases.

There exists a weak relationship between 'Deaths/1M pop' and the other two variables. If the total number of tests or the total number of infections increases, the mortality rate also increases, but the rate of increase is not too high.

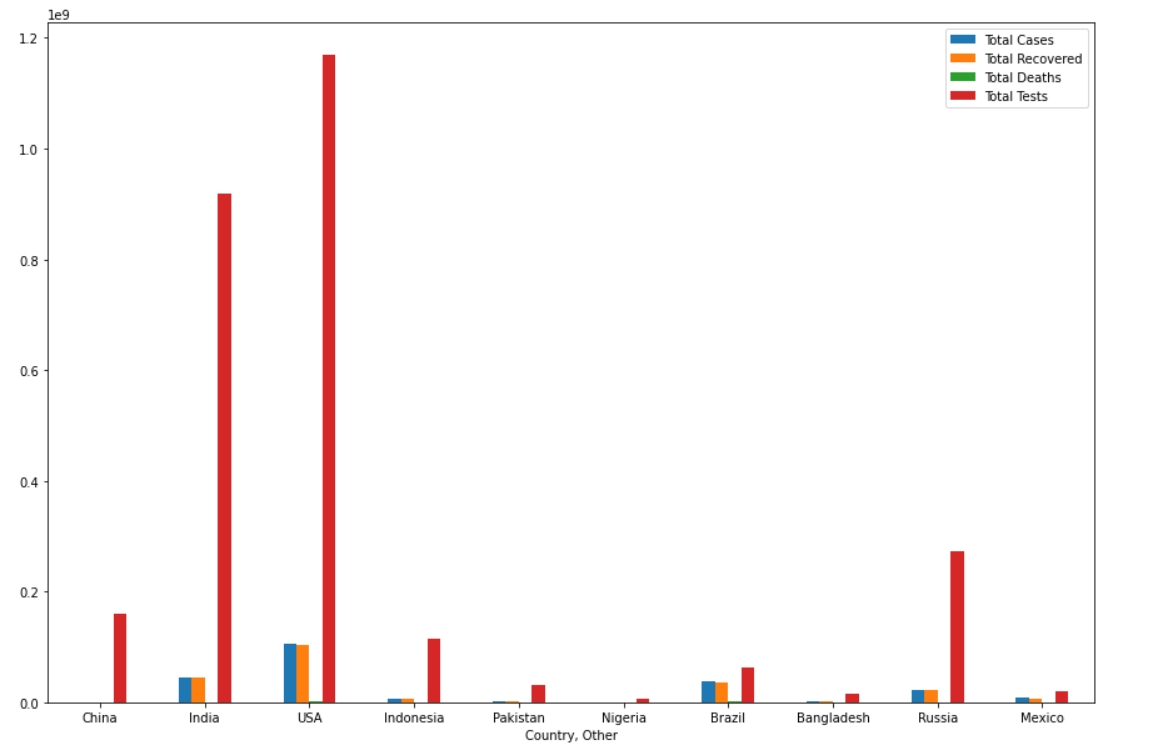
1. ***Is there a causal relationship between the two fields Total Cases and Total Tests? (The number of infections is high because there are many tests?)***

****

***Comment***

There is a positive correlation between 'Total Cases' and 'Total Tests', we see that if 'TotalCases' increases, 'TotalTests' also increases and vice versa.

1. ***Top 10 most populous countries data:***
   1. **Relationship Total Cases, Total Recovered, Total Deaths, Total Tests:**

****

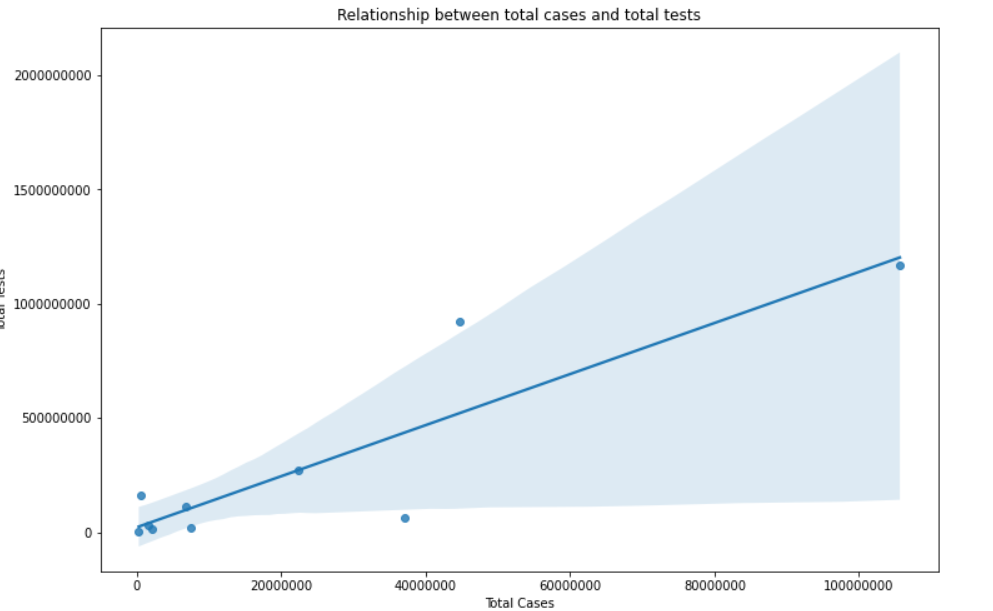
***Comment:***

Total Cases, Total Recovered, Total Deaths and Total Tests are proportional to each other. When Total Tests are higher, the remaining 3 attributes will also increase.

The speed of disease recovery in all countries is very good (TotalRecovered is approximately equal to TotalCases).

The number of deaths (TotalDeaths) is quite low compared to TotalCases, TotalRecovered.

* 1. **Relationship between Total Cases and Total Test:**

****

***Comment:***

These two attributes have a linear relationship with each other, the more people test, the higher the number of covid cases.

# References

Dario Radečić, 15 Feb, 2021. [How to Make Stunning Radar Charts with Python](https://betterdatascience.com/radar-charts-matplotlib-plotly/).

srivastavaharshit848, varshagumber28, 02 Feb, 2022. [Donut Chart using Matplotlib in Python](https://www.geeksforgeeks.org/donut-chart-using-matplotlib-in-python/).