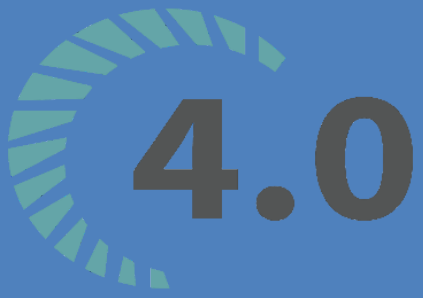
KHOA CÔNG NGHỆ THÔNG TIN

ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM

TRỰC QUAN HÓA DỮ LIỆU



Nhóm thực hiện: Nhóm 11

19127476 - Trần Thị Huế Minh

19127486 - Nguyễn Lê Nguyên

19127125 - Lê Duy Dũng

MỤC LỤC

I. Phân chia công việc

BẢNG THÀNH VIÊN

Họ và tên	MSSV	Công việc				Đánh giá công việc
		Thu thập dữ liệu	Tiền xử lý dữ liệu	Trực quan hóa dữ liệu	Báo cáo	
Trần Thị Huế Minh	19127476		Xử lý missing data	Xem xét tình hình, số liệu, mối tương quan dữ liệu	Viết Báo Cáo	100%
Lê Duy Dũng	19127125	Thu thập dữ liệu			Viết Báo Cáo	80%
Nguyễn Lê Nguyên	19127486		Chuyển đổi kiểu dữ liệu	Phân phối dữ liệu	Viết Báo Cáo	100%

II. Thu thập dữ liệu

1. Thiết lập kết nối đến trang web và phương thức thu thập dữ liệu

Sử dụng thư viện Selenium kết hợp với thư viện BeautifulSoup để lấy dữ liệu từ trang https://www.worldometers.info/coronavirus bằng cách parse HTML. Trong đó:

- Sử dụng các gói của thư viện selenium để tạo kết nối đến website 'https://www.worldometers.info/coronavirus' , tạo một phiên bản mới của Chrome WebDriver từ mô-đun webdriver, cho phép tập lệnh Python điều khiển trình duyệt web Chrome thông qua các tương tác tự động.

```
browser = webdriver.Chrome(service=Service(ChromeDriverManager().install()))
browser.maximize_window()
browser.get('https://www.worldometers.info/coronavirus')
```

- Sử dụng thư viện BeautifulSoup để phân tích cú pháp mã nguồn HTML của trang web.

```
soup = BeautifulSoup(browser.page_source, 'html.parser')
```

2. Tiến hành lấy dữ liệu

Sau khi đã tạo được kết nối và cài đặt phương thức thu thập dữ liệu từ trang web, ta tiến hành thu thập các dữ liệu cần thiết:

- Chỉ lấy các cột dữ liệu được hiển thị trên bảng dữ liệu của trang web, bên cạnh đó thêm một trường dữ liệu "Continent" cung cấp thông tin quốc gia đó thuộc Châu lục nào:

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	New Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	682,403,916	+39,347	6,818,879	+131	655,331,438	+53,808	20,253,599	40,194	87,546	874.8			
1	USA	105,820,390		1,151,253		103,485,147		1,183,990	2,261	316,065	3,439	1,169,976,712	3,494,499	334,805,269
2	India	44,694,349		530,799		44,158,161		5,389	N/A	31,774	377	920,111,661	654,124	1,406,631,776
3	France	39,690,610		165,314		*39,447,483*		77,813	869	605,183	2,521	271,490,188	4,139,547	65,584,518
4	Germany	38,297,037		169,661		37,934,100	+2,800	193,276	N/A	456,550	2,023	122,332,384	1,458,359	83,883,596
5	Brazil	37,145,514		699,634		36,249,161		196,719	N/A	172,486	3,249	63,776,166	296,146	215,353,593
6	Japan	33,368,365	+7,066	73,477	+38	21,708,854	+1,922	11,586,034	86	265,704	585	96,765,558	770,519	125,584,838
7	S. Korea	30,690,223	+8,995	34,159	+4	*30,479,566*	+10,810	176,498	129	597,901	665	15,804,065	307,892	51,329,899
8	Italy	25,651,205		188,750		25,320,467		141,988	104	425,656	3,132	269,127,054	4,465,893	60,262,770
9	UK	24,423,396		208,458		*24,149,508*	+2,385	65,430	N/A	356,557	3,043	522,526,476	7,628,357	68,497,907
10	Russia	22,493,866	+13,009	396,801	+32	21,841,765	+12,155	255,300	N/A	154,273	2,721	273,400,000	1,875,095	145,805,947
11	Turkey	17,042,722		101,492		N/A	N/A	N/A		199,186	1,186	162,743,369	1,902,052	85,561,976
12	Spain	13,783,163		119,872		13,632,061		31,230	231	295,022	2,566	471,036,328	10,082,298	46,719,142
13	Vietnam	11,527,116	+6	43,186		10,614,847	+6	869,083	5	116,490	436	85,826,548	867,342	98,953,541
14	Australia	11,385,534		19,459		11,332,576		33,499	27	436,750	746	78,835,048	3,024,116	26,068,792

- Có thể lấy dữ liệu theo một trong 3 ngày sau:
  - Dữ liệu được lấy hiện tại của ngày 8/3/2023 được lưu vào file csv với tên 'covid_data_now.csv'
  - Dữ liệu của ngày trước đó là 7/3/2023, được lưu vào file csv với tên 'covid_data_yesterday.csv'
  - Dữ liệu của 2 ngày trước đó là 6/3/2023, được lưu vào file csv với tên 'covid_data_yesterday2.csv'

```python
def get_data(index_time):
    if index_time == 0:
        table = soup.find("table", {"id" : "main_table_countries_today"})
    else:
        if index_time == 1:
            table = soup.find("table", {"id" : "main_table_countries_yesterday"})
        else:
            table = soup.find("table", {"id" : "main_table_countries_yesterday2"})

    headers = get_header(table)
    df = pd.DataFrame(columns = headers)

    #Loại bỏ các dòng không muốn dùng đến
    list_row = table.find_all('tr', class_="total_row_world") + table.find_all('tr', class_="total_row")

    for j in table.find_all('tr')[1:]:
        if j in list_row:
            continue
        else:
            row_data = j.find_all('td')
            row = [i.text for i in row_data]

            new_row = pd.Series(row, index=df.columns[:len(row)])
            df = pd.concat([df, new_row.to_frame().T], ignore_index=True, axis=0)

    #Xuất dữ liệu ra file csv
    if index_time == 0:
        df.to_csv('covid_data_now.csv', index=False)
    else:
        if index_time == 1:
            df.to_csv('covid_data_yesterday.csv', index=False)
        else:
            df.to_csv('covid_data_yesterday2.csv', index=False)
```

Sau khi thu thập dữ liệu vào file csv ta tiếp tục đọc dữ liệu từ file và có thể tiền xử lý dữ liệu trước khi tiến hành trực quan hóa

Dữ liệu nhóm sử dụng để trực quan được thu thập vào ngày 8/3/2023. Vì vậy file dữ liệu được sử dụng trong các phần sau sẽ là “‘covid_data_now.csv’”

III. Tiền xử lý dữ liệu

1. Quan sát dữ liệu tổng quát

a. Nhận xét về kiểu dữ liệu

Theo quan sát các trường dữ liệu có mô tả thuộc tính như sau:

- #: STT
- Country,Other:Tên quốc gia
- TotalCases: Tổng số lượng ca nhiễm
- NewCases: Số ca nhiễm mới trong ngày
- TotalDeaths: Tổng số ca tử vong do Covid-19
- NewDeaths: Số ca tử vong mới được phát hiện trong ngày
- TotalRecovered: Tổng số ca hồi phục khi bị nhiễm Covid-19

- NewRecovered: Số ca hồi phục trong ngày
- ActiveCases: Số ca đang bị nhiễm và chưa khỏi bệnh
- Serious,Critical: Số lượng ca nhiễm Covid-19 đang trong nguy kịch
- Tot Cases/1M pop: Tổng số ca được xác nhận nhiễm trên 1 triệu dân số
- Deaths/1M pop: Tổng số ca tử vong trên 1 triệu dân số
- TotalTests: Tổng số lượng được xác nghiệm Covid-19
- Tests/1M pop: Tổng số lượng được xác nghiệm Covid-19 trên 1 triệu dân số
- Population: Tổng dân số của quốc gia
- Continent: Tên châu lục

Trong đó:

- Các cột "Country,Other", và continent có kiểu dữ liệu là chuỗi.
- Các cột còn lại đều là kiểu số nguyên.

b. Nhận xét về thành phần quốc gia và châu lục

- Có tổng cộng 231 giá trị khác nhau trong cột Country,Other.
- Có tổng cộng 6 giá trị khác nhau trong cột Continent

2. Tiền xử lý dữ liệu

a. Xóa cột không cần thiết

- Sau khi lấy dữ liệu từ website, ta sẽ xóa các cột dữ liệu không cần thiết. Trong đó cột "#", không sử dụng vì nó chỉ đánh dấu số thứ tự của các dòng và dữ liệu trong python được lưu dưới dạng dataframe đã cung cấp sẵn index vì vậy ta có thể loại bỏ để tập trung đến các trường dữ liệu quan trọng khác
- Thực hiện câu lệnh drop với axis = 1

```
df = df.drop('#', axis=1)
df
```

b. Chuyển kiểu dữ liệu

- Dữ liệu được đọc từ file csv có thể không đúng với dữ liệu mong muốn, vì vậy ta cần kiểm tra và chuyển đổi để đảm bảo cho việc phân tích dữ liệu

```
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Country,Other     231 non-null    object
 1   TotalCases        231 non-null    object
 2   NewCases          46 non-null     object
 3   TotalDeaths       231 non-null    object
 4   NewDeaths         23 non-null     float64
 5   TotalRecovered    210 non-null    object
 6   NewRecovered      50 non-null     object
 7   ActiveCases       212 non-null    object
 8   Serious,Critical  127 non-null    object
 9   Tot Cases/1M pop  229 non-null    object
 10  Deaths/1M pop     223 non-null    object
 11  TotalTests        213 non-null    object
 12  Tests/1M pop      213 non-null    object
 13  Population        231 non-null    object
 14  Continent         229 non-null    object
```

- Kiểm tra kiểu dữ liệu ta thấy chỉ có cột NewDeaths là có kiểu dữ liệu float64 còn các cột còn lại là kiểu *object*.
- Để phân tích và thống kê dữ liệu ta cần chuyển kiểu dữ liệu của các cột như đã nhận xét.
    - Cột 'Country,Other' và 'Continent' có kiểu dữ liệu là *string*.
    - Các cột còn lại có kiểu dữ liệu là *integer*.

c. Kiểm tra và xử lý dữ liệu missing value

- Để kiểm tra missing value, ta sẽ check số lượng NaN trong mỗi thuộc tính và tính tỉ lệ phần trăm của NaN

	Mumbers	Percent missing
Country,Other	0	0.000
TotalCases	0	0.000
NewCases	185	80.087
TotalDeaths	6	2.597
NewDeaths	208	90.043
TotalRecovered	21	9.091
NewRecovered	181	78.355
ActiveCases	19	8.225
Serious,Critical	104	45.022
Tot Cases/1M pop	2	0.866
Deaths/1M pop	8	3.463
TotalTests	18	7.792
Tests/1M pop	18	7.792
Population	2	0.866
Continent	2	0.866

- Nhận xét về kết quả:

  - Các cột có tỉ lệ missing value cao hơn 70% là:
    - NewCases : 80,087%
    - NewDeaths : 90,043%
    - NewRecovered : 78.355%

  - Các cột có tỉ lệ missing value trung bình là:
    - SeriousCritical : 45.022%

  - Các cột có tỉ lệ missing value thấp dưới 10% là:
    - TotalRecovered : 9.091%
    - ActiveCases : 8.225%
    - Deaths/1M pop : 3.463%
    - TotalTests : 7.792%
    - Tests/1M pop : 7.792%
    - Continent : 0.866%
    - Tot Cases/1M pop : 0.866%

- Các không có missing value: Country,Other, TotalCases, TotalDeaths, Population

d. Xử lý missing value

- **Đối với các cột dữ liệu có tỉ lệ missing value cao hơn 70%:**
  - Dữ liệu missing ở các cột này khá cao, đồng thời những dữ liệu này cho biết số ca nhiễm, số ca tử vong, số ca hồi phục mới trong ngày và những số liệu này không đóng góp nhiều vào quá trình phân tích và thống kê --> Chọn loại bỏ các cột dữ liệu này

- **Đối với các cột dữ liệu có tỉ lệ missing value trung bình**:
  - Vì số lượng missing value trong cột Serious,Critical cũng ở mức cao, nhưng cột dữ liệu này có ý nghĩa thống kê quan trọng vì vậy nhóm chọn phương pháp điền giá trị trống trong cột này là 0 để mục đích khi phân tích có thể so sánh được mức độ Serious,Critical của các quốc gia

- **Đối với các cột dữ liệu có tỉ lệ missing value thấp:**
  - Nhận thấy đối với 2 cặp cột dữ liệu [Continent,Tot Cases/1M pop] và [TotalTests ; Tests/1M pop] có tỉ lệ missing value bằng nhau lần lượt là 0.866% và 7.792%
  - Đặt giả thuyết rằng các giá trị missing tại các cặp cột này là của cùng một quốc gia. Ta kiểm tra như sau:
    - [Continent,Tot Cases/1M pop]

```
country_Continent = []

for i in df.index:

    try:
        if df.isnull().iloc[i, 11]:
            country_Continent.append(i)
            print('Continent _', i, ': ', df.loc[i, "Country,Other"])
    except:
        None

    try:
        if df.isnull().iloc[i, 6]:
            print('Tot Cases/1M pop _', i, ': ', df.loc[i, "Country,Other"])
    except:
        None
```

```
Continent _ 226 :  Diamond Princess
Tot Cases/1M pop _ 226 :  Diamond Princess
Continent _ 229 :  MS Zaandam
Tot Cases/1M pop _ 229 :  MS Zaandam
```

- ***Nhận xét:*** Đúng với giả thuyết các ô dữ liệu bị thiếu thuộc cùng một Quốc gia, tại bộ [Continent,Tot Cases/1M pop] cột dữ liệu Continent xác định Châu lục mà quốc gia đó thuộc về bị null, điều này có thể gây ảnh hưởng đến việc phân tích và thống kê nếu điền giá trị null bằng phương pháp thay thế. Vì vậy nhóm lựa chọn phương án loại bỏ các sample này và xem chúng như outlier của bộ dữ liệu.

    - [TotalTests ; Tests/1M pop]

```
TotalTests _ 29 :  DPRK :  25990679
Tests1Mpop _ 29 :  DPRK :  25990679
TotalTests _ 138 :  French Polynesia :  284164
Tests1Mpop _ 138 :  French Polynesia :  284164
TotalTests _ 150 :  Seychelles :  99426
Tests1Mpop _ 150 :  Seychelles :  99426
TotalTests _ 156 :  Tanzania :  63298550
Tests1Mpop _ 156 :  Tanzania :  63298550
TotalTests _ 172 :  Solomon Islands :  721159
Tests1Mpop _ 172 :  Solomon Islands :  721159
TotalTests _ 181 :  Nicaragua :  6779100
Tests1Mpop _ 181 :  Nicaragua :  6779100
TotalTests _ 183 :  Tajikistan :  9957464
Tests1Mpop _ 183 :  Tajikistan :  9957464
TotalTests _ 190 :  Marshall Islands :  60057
Tests1Mpop _ 190 :  Marshall Islands :  60057
TotalTests _ 203 :  Comoros :  907419
Tests1Mpop _ 203 :  Comoros :  907419
TotalTests _ 216 :  Kiribati :  123419
Tests1Mpop _ 216 :  Kiribati :  123419
TotalTests _ 221 :  Tuvalu :  12066
Tests1Mpop _ 221 :  Tuvalu :  12066
TotalTests _ 222 :  Saint Helena :  6115
Tests1Mpop _ 222 :  Saint Helena :  6115
TotalTests _ 225 :  Niue :  1622
Tests1Mpop _ 225 :  Niue :  1622
TotalTests _ 226 :  Vatican City :  799
Tests1Mpop _ 226 :  Vatican City :  799
TotalTests _ 227 :  Western Sahara :  626161
Tests1Mpop _ 227 :  Western Sahara :  626161
TotalTests _ 228 :  Tokelau :  1378
Tests1Mpop _ 228 :  Tokelau :  1378
```

- ***Nhận xét:*** Các sample bị thiếu thuộc các quốc gia có số lượng dân số thấp, và nhận thấy rằng nếu loại bỏ các sample này sẽ không làm ảnh hưởng quá nhiều đến việc thống kê sau này, bên cạnh đó còn hạn chế được các giá trị missing đáng kể.

- Kiểm tra lại phần trăm missing value ta thấy các cột còn lại có missing value khá thấp, nhóm lựa chọn phương pháp giữ nguyên các dữ liệu đó để tránh làm sai lệch dữ liệu quá nhiều.

	Mumbers	Percent missing
Country,Other	0	0.000
TotalCases	0	0.000
TotalDeaths	1	0.469
TotalRecovered	16	7.512
ActiveCases	16	7.512
Serious,Critical	0	0.000
Tot Cases/1M pop	0	0.000
Deaths/1M pop	1	0.469
TotalTests	0	0.000
Tests/1M pop	0	0.000
Population	0	0.000
Continent	0	0.000

Sau khi hoàn tất quá trình tiền xử lý dữ liệu , ta có thể thấy dữ liệu đã có các thay đổi so với dữ liệu ban đầu:

- Số lượng dòng: 213
- Số lượng cột: 12
- Số lượng quốc gia: 213
- Số lượng châu lục: 6

Dữ liệu này sẽ được lưu sang file mới với tên "new_covid_data_now.csv", và phần trực quan dữ liệu sẽ sử dụng dữ liệu từ file này

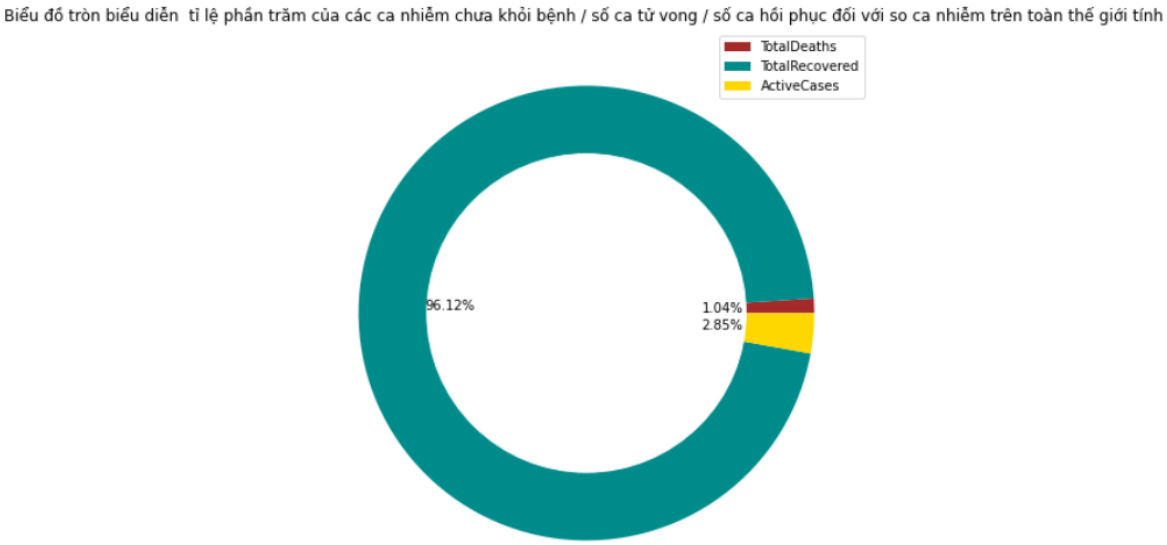
IV. Trực quan hóa dữ liệu

1. Xem xét tình hình số ca nhiễm đang điều trị COVID chưa khỏi bệnh / số ca tử vong / số ca hồi phục trên toàn thế giới tính cho đến ngày 8/3/2023

Tính tổng số ca nhiễm, tổng số ca phục hồi, tổng số ca tử vong và số ca đang điều trị từ tập dữ liệu thu thập được, ta có:

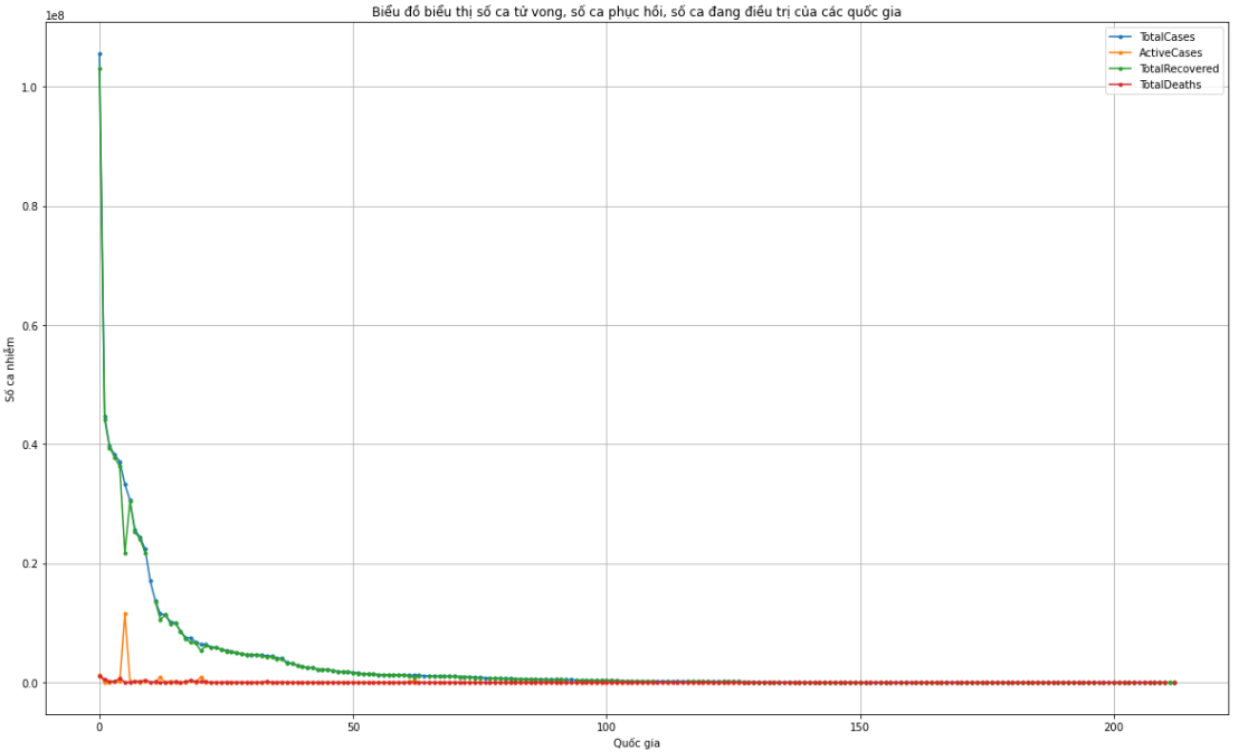
```
Số ca nhiễm covid-19:  676160585
Số ca tử vong do covid-19:  6807411
Số ca hồi phục :  630277141
Số ca đang điều trị:  18668000
```

- Nhận xét dữ liệu ta thấy: *Số ca nhiễm = Số ca tử vong + Số ca hồi phục + Số ca đang điều trị*, vì vậy nhóm quyết định sử dụng **pie chart** để biểu diễn tỉ lệ phần trăm của các ca nhiễm đang điều trị COVID chưa khỏi bệnh / số ca tử vong / số ca hồi phục trên toàn thế giới



- **Nhận xét biểu đồ:** Vừa nhìn vào biểu đồ ta đã có thể nhận xét rằng tình hình dịch Covid19 hiện nay đã hoàn toàn được kiểm soát khi tỉ lệ hồi phục chiếm tới 96.12% trong khi tỉ lệ đang điều trị là 2.85% và tỉ lệ chết chỉ có 1.04%.

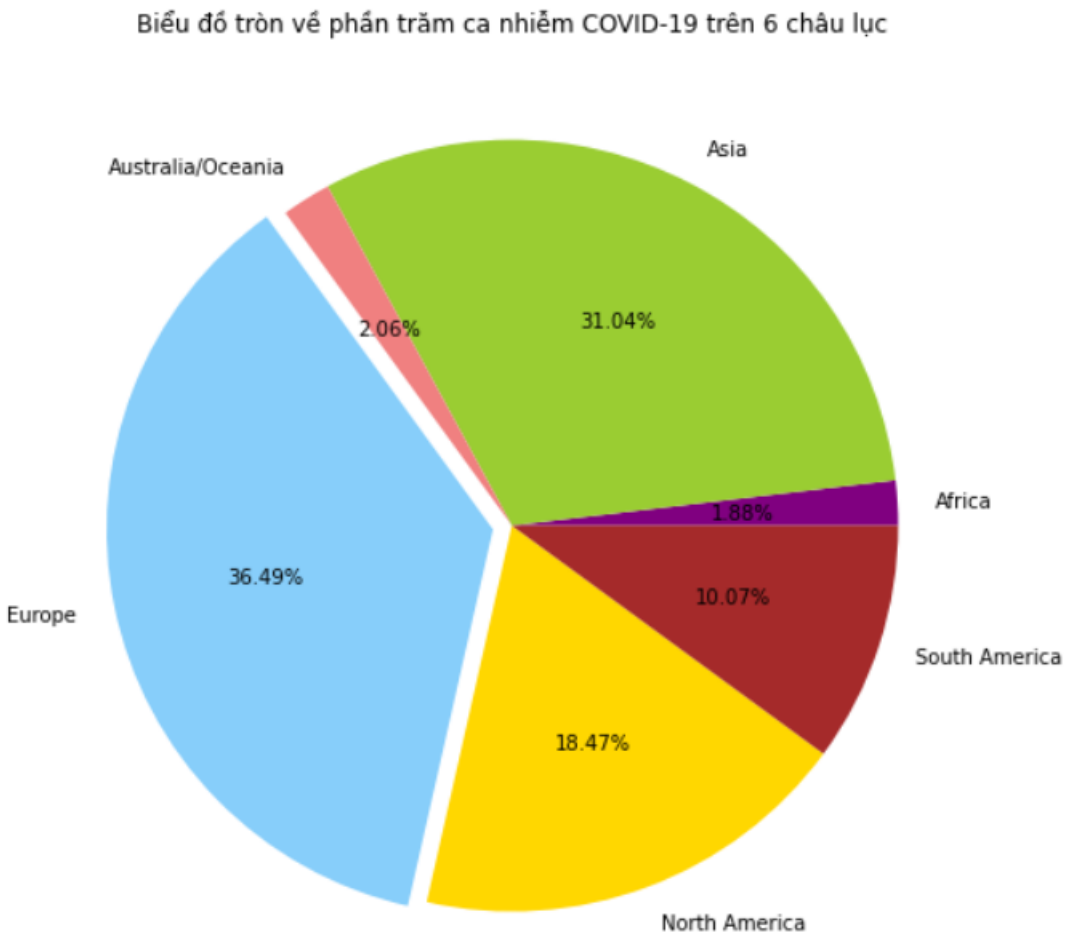
Để quan sát rõ hơn về số lượng ca nhiễm cũng như số lượng ca tử vong, số ca phục hồi, số ca đang điều trị của các quốc gia trên thế giới có trong tập dữ liệu, ta có thể sử dụng **Line chart** để biểu diễn, khi đó ta có thể xem mỗi một điểm trên đường biểu diễn là một quốc gia và các đường biểu diễn có màu khác nhau đại diện cho số liệu về tổng số ca nhiễm, tổng số ca phục hồi, tổng số ca tử vong và số ca đang điều trị



- **Nhận xét về biểu đồ:** Nhìn chung, các quốc gia có số ca nhiễm càng cao thì số ca phục hồi cũng cao. Dựa vào biểu đồ trên ta có thể thấy:
    - Số ca nhiễm cao nhất của các quốc gia là trên 100 triệu ca.
    - Số ca phục hồi cao nhất của các quốc gia cũng trên 100 triệu ca.
    - Số ca đang được điều trị cao nhất là trên 10 triệu ca
    - Số ca tử vong cao nhất cũng trên 1 triệu ca

2. Xem xét có tỉ lệ ca nhiễm ở các châu lục trên thế giới

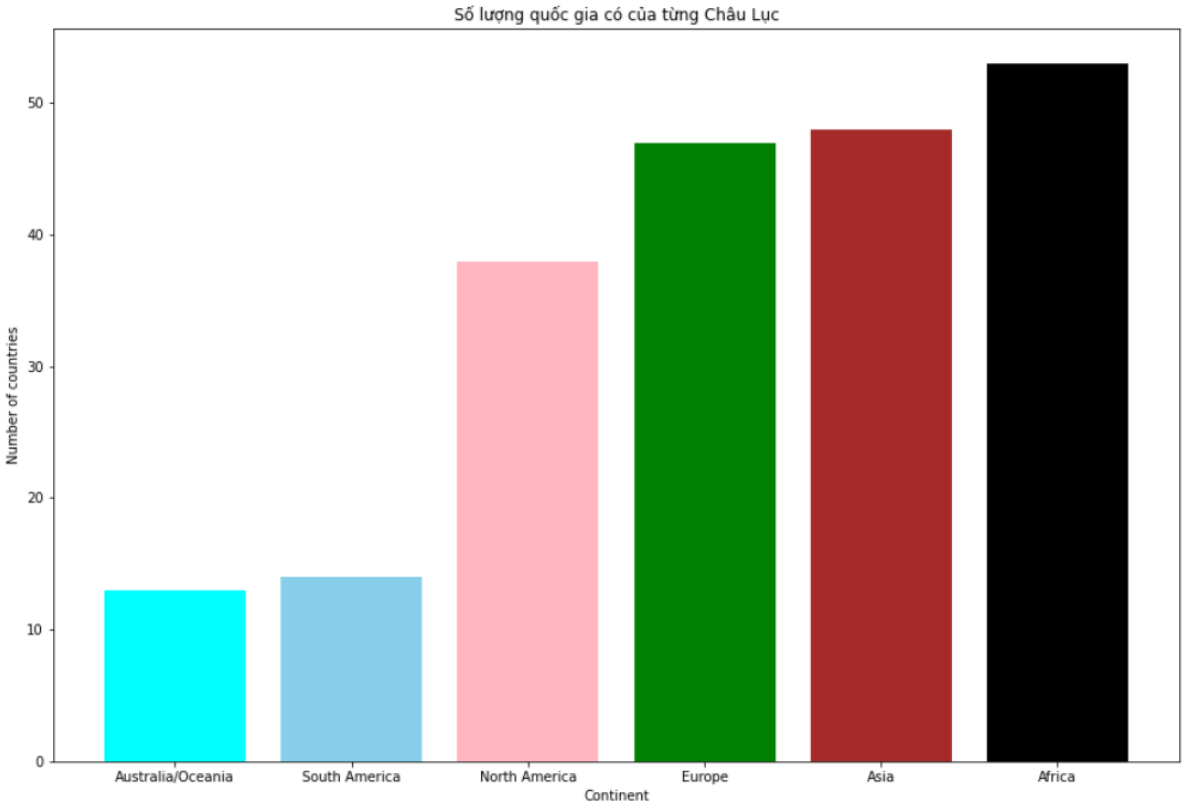
Đối với việc xem xét tỉ lệ ca nhiễm ở các châu lục, nhóm quyết định sử dụng **pie chart** để biểu diễn tỉ lệ phần trăm của tổng số ca nhiễm của các Châu lục trên toàn thế giới, mỗi Châu lục được biểu diễn là một phần của hình tròn.



Nhận xét biểu đồ:

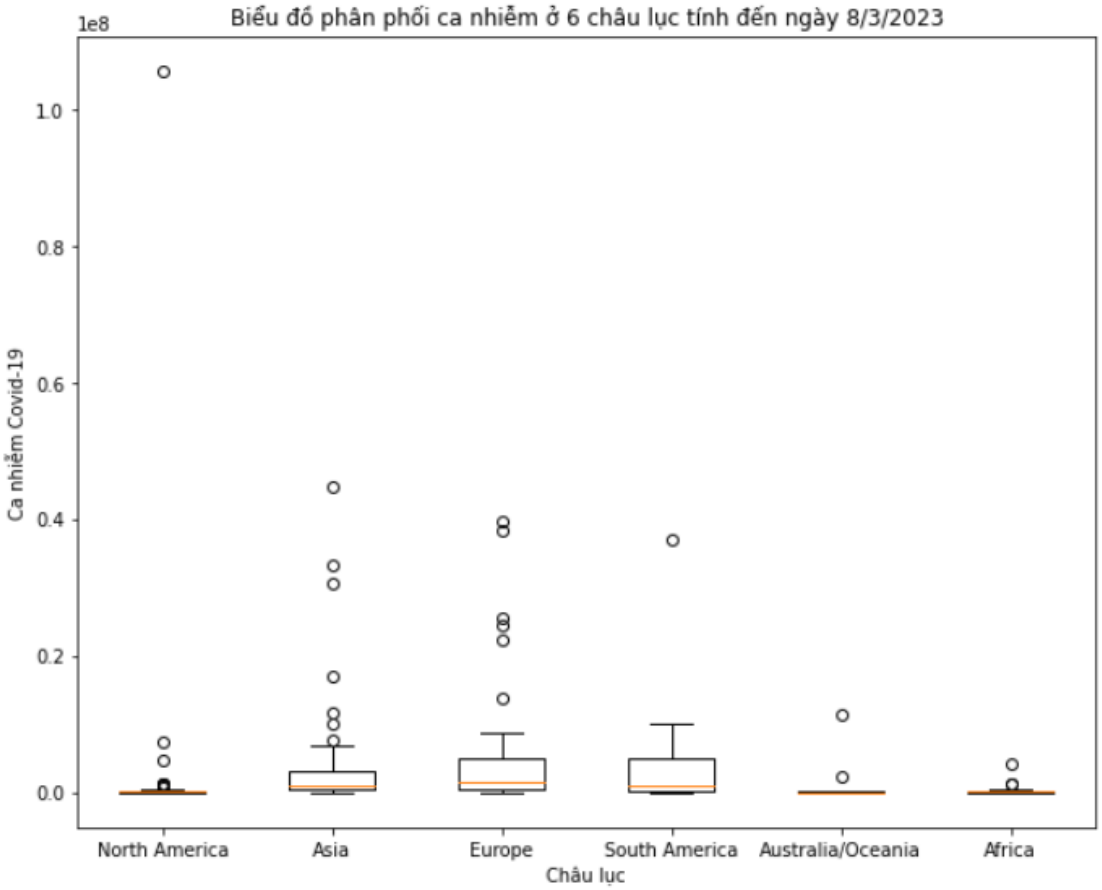
- Nhìn vào biểu đồ này ta có thể thấy ngay rằng tính tới ngày 8/3/2023, Châu âu (Europe) chiếm tỉ lệ cao nhất 36.49% (phần được tách rời để phân biệt với các phần còn lại), chiếm hơn 1/3 số ca nhiễm trên toàn thế giới, đến ngày nay Châu Âu được xem là tâm điểm của đại dịch.

Từ nhận xét trên ta có thể đưa ra giả thuyết rằng tỉ lệ số ca nhiễm của một Châu lục sẽ phụ thuộc vào số lượng các quốc gia thuộc Châu lục đó được ghi nhận trong bảng dữ liệu. Ta sử dụng biểu đồ **bar chart** để quan sát số quốc gia của các Châu lục được ghi nhận trong dữ liệu này. Lí do sử dụng biểu đồ bar chart vì các biến châu lục là kiểu biến phân loại, mỗi cột đại diện cho một châu lục.



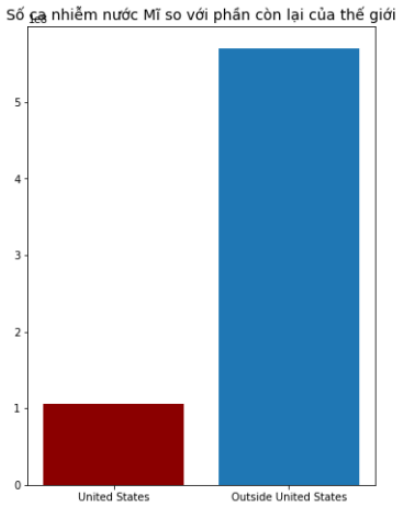
Nhận xét:

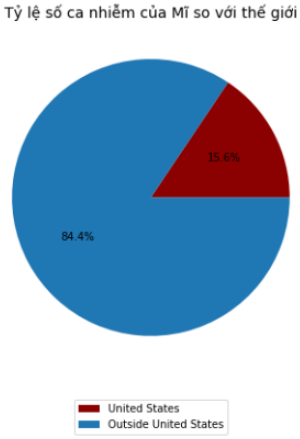
- Dựa vào biểu đồ Bar chart như trên, ta có thể thấy số lượng quốc gia thuộc Châu Phi là nhiều nhất, kế tiếp là Châu Á. Trong khi số lượng ca nhiễm ở Châu Âu mới là cao nhất.
- Điều này đã cho thấy giả thuyết về : Số ca nhiễm tại các Châu Lục phụ thuộc vào số lượng quốc gia được ghi nhận là sai
- Để tìm hiểu về lý do số ca nhiễm của Châu Âu chiếm tỉ lệ lớn nhất trong các Châu lục trên thế giới, ta tiến hành trực quan sự phân phối của số ca nhiễm được ghi nhận ở các Châu Lục bằng sơ đồ **box plot** sau:

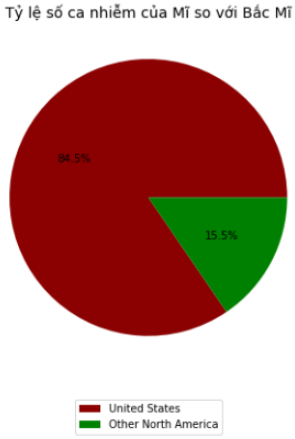


Nhận xét:

- Dựa vào box plox, ta thấy tại Europe đã xuất hiện tới 6 outlier với mật độ khá cao, có đến 5/6 outlier lớn hơn 20000000 điều này có thể giải thích răng Châu Âu có số lượng ca nhiễm chiếm 1/3 số ca nhiễm trên toàn thế giới là do có một số quốc gia thuộc Châu Âu có thể chưa có biện pháp phòng chống dịch COVID-19 tốt dẫn đến số lượng ca nhiễm tại quốc gia đó cao vượt trội, vì vậy tổng số ca nhiễm của Châu Âu đã cao hơn so với các Châu Lục khác.
- Ngoài ra, ta cũng có thể thấy trên biểu đồ Boxplot trên, tại North America có 1 outlier lớn hơn 100000000 , có thể đây là quốc gia có số lượng ca nhiễm cao nhất thế giới, và outlier này cũng đã kéo theo số người nhiễm COVID-19 trung bình tại Bắc Mỹ tăng mạnh
- Vì vậy, ta có thể xem xét về số lượng ca nhiễm của Mỹ so với thế giới như sau:







Nhận xét:

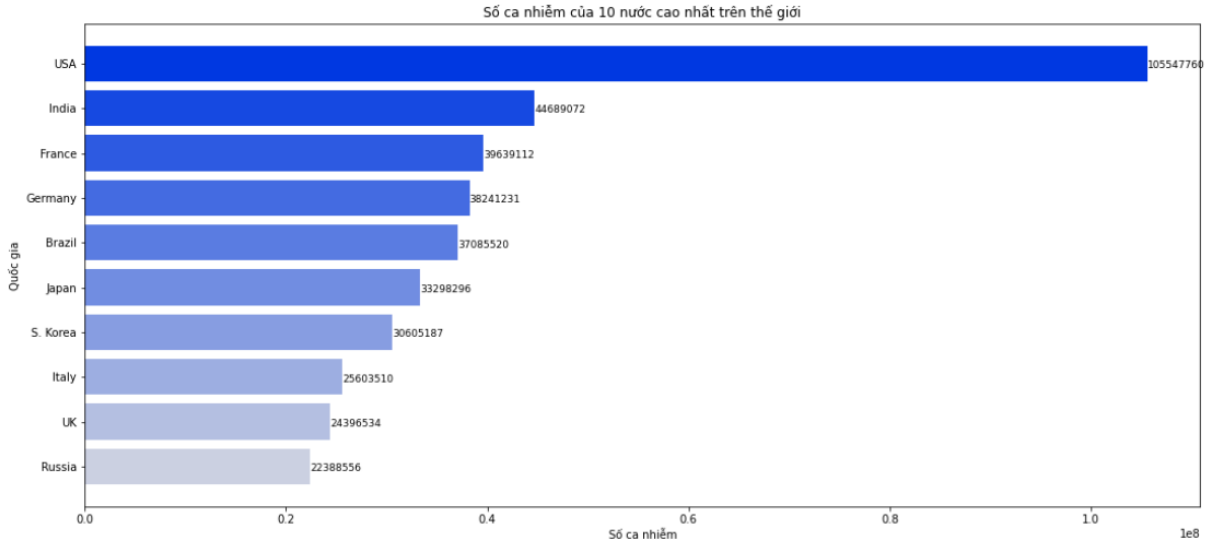
- Ta có thể thấy số ca nhiễm của Hoa Kỳ chiếm 15,6% so với thế giới, chiếm 84,5% so với Bắc Mỹ, vượt xa so với các nước khác. Nguyên nhân có thể bắt nguồn từ thông tin vào thời điểm dịch bệnh chưa phát triển mạnh ở Hoa Kỳ, người dân của nước Mỹ không quan tâm việc tiêm ngừa dịch bệnh, và chính phủ nước này chưa có sự chuẩn bị chu đáo trước khi tiếp nhận những còn đại bùng dịch do COVID-19
- Vì vậy ta có thể thấy nước Mĩ tăng về số ca nhiễm sẽ dẫn đến số ca nhiễm của cả Bắc Mĩ tăng đáng kể

3. Xem xét số ca nhiễm của 10 nước cao nhất vào ngày 8/03/2023

Để biểu diễn số lượng ca nhiễm của 10 quốc gia có số lượng ca nhiễm cao nhất thế giới, nhóm sử dụng biểu đồ **"Bar chart ngang"** bởi vì Quốc gia là biến phân loại, mỗi quốc gia sẽ được biểu diễn bằng một cột, tại đây có thể thấy tên của một quốc gia có số lượng dài ngắn khác nhau, nên việc sử dụng biểu đồ ngang để có thể dễ dàng đọc được tên quốc gia.

Thêm vào đó nhóm sử dụng phương pháp ưu tiên độ dài, sắp xếp số lượng quốc gia giảm dần để có thể thấy các cột dữ liệu ngắn dần theo thứ tự từ trên xuống dưới.

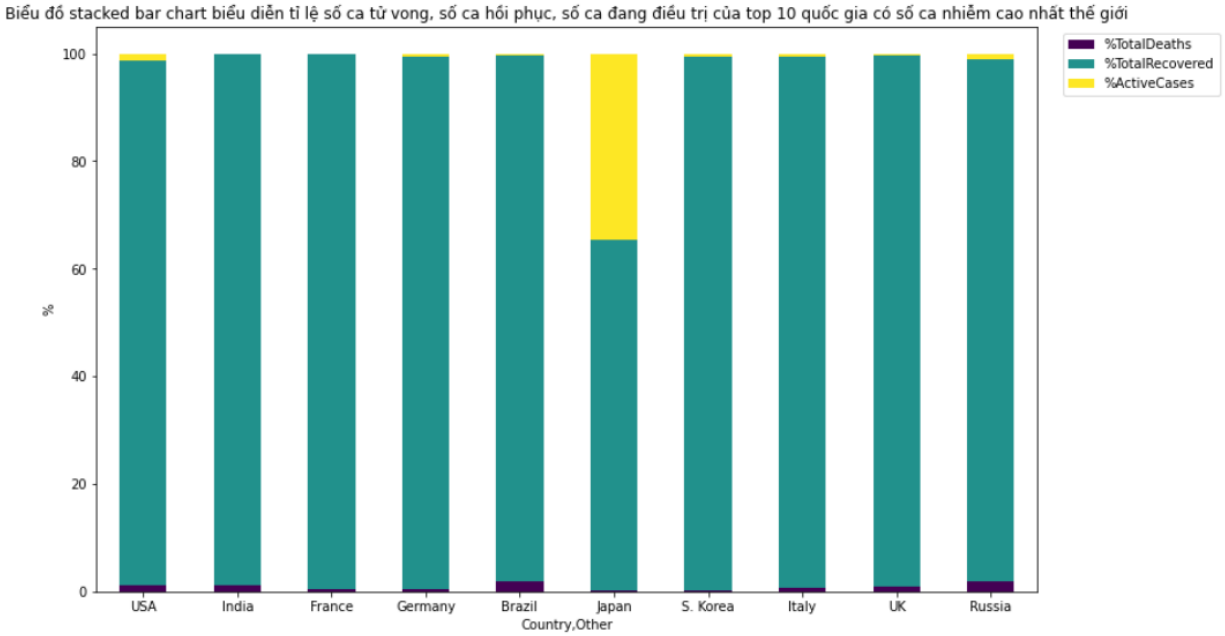
Một điểm bổ sung là màu sắc của cột dữ liệu cũng giảm dần theo mức độ tương phản để làm nổi bật cột dữ liệu có số ca nhiễm cao



Nhận xét:

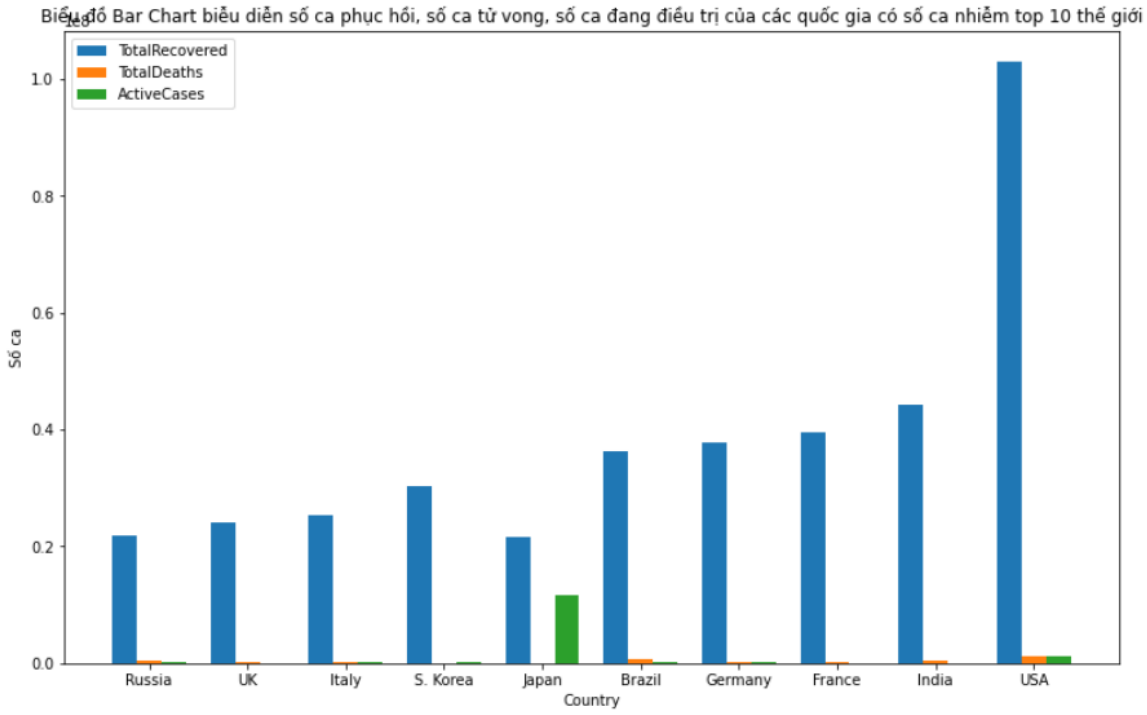
- Như biểu đồ bar chart như trên đã biểu diễn, ta có thể thấy USA là quốc gia có số ca nhiễm cao nhất thế giới với số ca nhiễm lên đến 105547760 ca
- Tiếp đến là India và France, dựa vào biểu đồ cũng thấy được số lượng ca nhiễm vượt trội của USA, với cột biểu diễn dài gấp đôi so với quốc gia đứng thứ 2 thế giới về số ca nhiễm

Tiếp theo, để tìm hiểu nhiều hơn về thông tin dịch bệnh tại các quốc gia này, ta kiểm tra số ca tử vong, số ca hồi phục, số ca đang điều trị có trong số ca nhiễm của các quốc gia đó. Nhóm sử dụng biểu đồ **stacked bar chart**, vì chúng biểu diễn được phần trăm của từng thành phần như số ca tử vong, số ca hồi phục, số ca đang điều trị của nhiều đối tượng trong cùng một biểu đồ



Nhận xét:

- Nhìn vào biểu đồ trên, ta có thể thấy trong số 10 quốc gia có số ca nhiễm COVID-19 cao nhất thế giới,
    - Brazil là quốc gia có số ca tử vong vì COVID chiếm tỉ lệ so với tổng số ca nhiễm của quốc gia đó cao hơn so với các nước khác trong top 10 quốc gia. Từ đó ta có thể suy đoán nguyên nhân do y tế của quốc gia chưa đủ mạnh để hạn chế số lượng ca tử vong từ COVID-19
    - Japan là quốc gia có tỉ lệ số ca đang điều trị so với tổng số ca nhiễm trên 30%. So với các quốc gia trong top 10 các quốc gia có số ca nhiễm cao nhất thế giới, quốc gia này vẫn đang trong quá trình chống chọi với dịch bệnh căng thẳng.
- Để quan sát rõ hơn về số ca số ca tử vong, số ca hồi phục, số ca đang điều trị tại top 10 quốc gia có số ca nhiễm cao nhất thế giới ta có thể xem dưới dạng biểu đồ **Bar chart với nhiều trường dữ liệu** như sau



Ở biểu đồ này, mỗi quốc gia được bởi tập hợp 3 cột dữ liệu liên tiếp có thứ tự, mỗi cột biểu diễn một trong 3 thông số của ca tử vong, ca hồi phục, ca đang điều trị

Nhận xét:

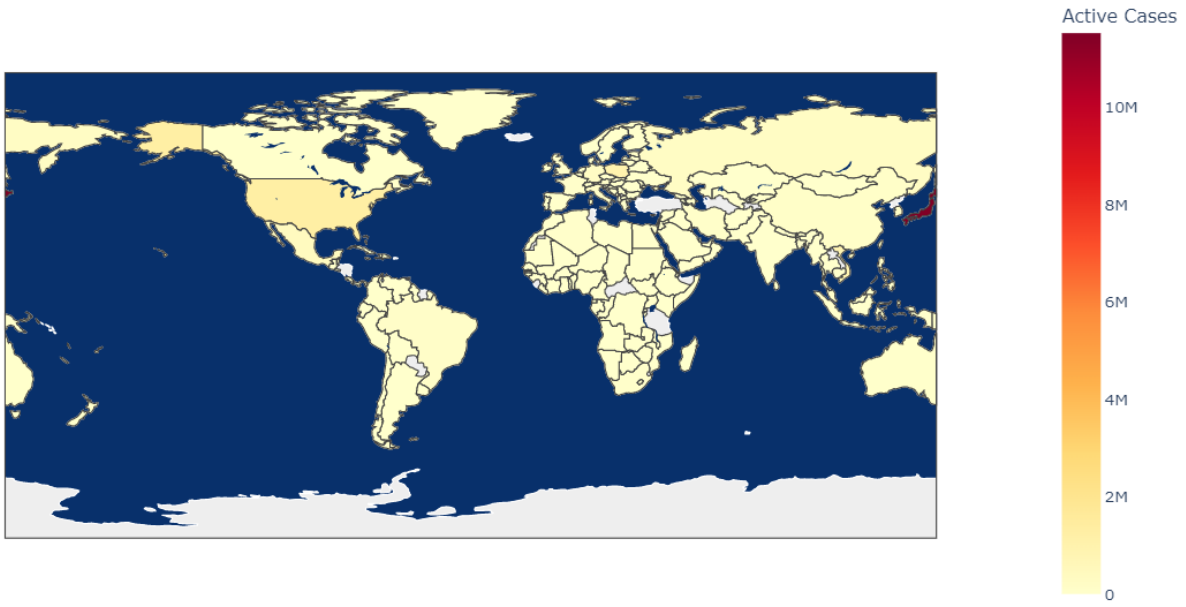
- Biểu đồ Bar chart biểu diễn rõ về số ca của từng quốc gia, từ đó ta dễ dàng ta thấy Mỹ là quốc gia số ca nhiễm cao nhất thế giới và có số ca tử vong đứng đầu trong top 10 quốc gia có số ca nhiễm cao nhất. Và số ca nhiễm còn đang điều trị tại Nhật Bản đạt ngưỡng cao nhất so với top 10 quốc gia đang quan sát. Có khoảng 10000000 số ca nhiễm đang điều trị ở Nhật Bản

4. Xem xét các quốc gia đang còn tồn tại dịch bệnh

Tiếp theo ta sẽ phân tích về "Số ca nhiễm đang điều trị" - ActiveCase, số liệu này biểu diễn số ca nhiễm COVID-19 chưa được khỏi bệnh, và trong số đó còn có những ca nguy kịch cần chăm sóc được thống kê thành cột Serious,Critical

ActiveCase thể hiện rằng quốc gia đó đang có số ca nhiễm đang điều trị là bao nhiêu, nghĩa là quốc gia này đang chống chọi với bao nhiêu ca nhiễm Dương tính. Ta sử dụng các **"Biểu đồ map"** để quan sát thống kê số ca đang điều trị của tất cả các quốc gia trên đất nước.

Biểu đồ map biểu diễn số lượng ca mắc COVID-19 chưa khỏi ở các quốc gia trên thế giới



Nhận xét:

- Dựa vào biểu đồ Map trên, ta có thể thấy tuy dịch bệnh đã gần như không còn trong giai đoạn bùng nổ nghiêm trọng như khoảng thời gian 3 năm qua, nhưng vẫn còn rất nhiều quốc gia trên toàn thế giới trong đó có cả Việt Nam vẫn còn dịch bệnh. Trong đó, quốc gia đang còn đối mặt với sự đe dọa của COVID-19 lớn nhất hiện nay là Nhật Bản với số ca nhiễm chưa khỏi bệnh là trên 11 triệu ca

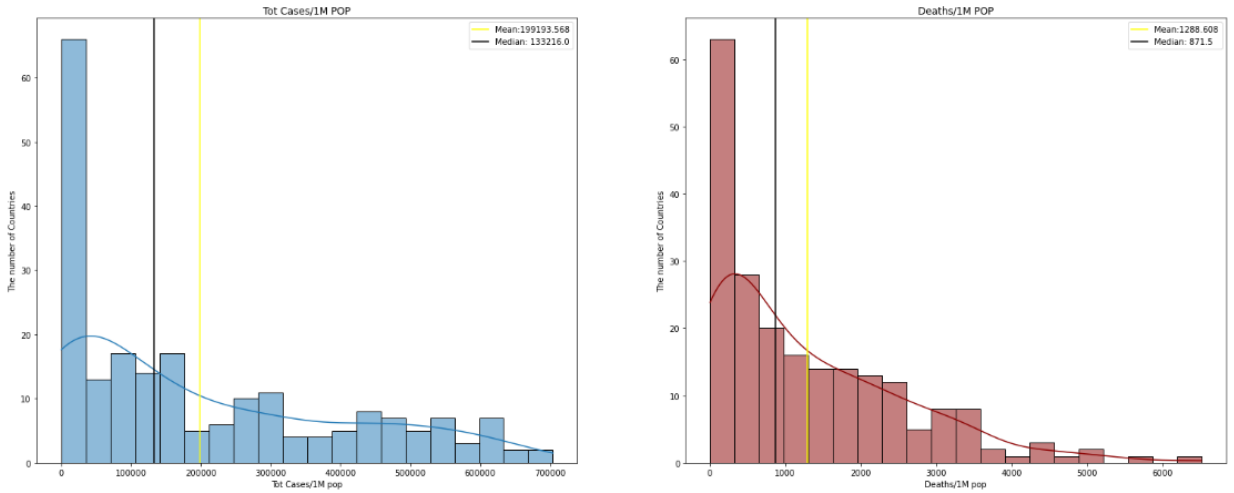
5. Quan sát phân phối dữ liệu của số ca nhiễm và số ca tử vong trên 1 triệu dân

Ngoài các dữ liệu đã được phân tích trên, một trong những số liệu cho thấy một quốc gia đã chịu áp lực từ dịch bệnh COVID-19 như thế nào đó là Số ca nhiễm trên 1 triệu dân, số ca tử vong trên 1 triệu dân.

Nhóm sử dụng biểu đồ **Histogram** để biểu diễn sự phân phối của số ca nhiễm và số ca tử vong trên 1 triệu dân. Số cột biểu diễn được chọn ở đây là 20 bins, vì thử nghiệm cho thấy với số khoảng chia này biểu đồ histogram cho thấy được dữ liệu được phân phối chuẩn

Đường phân phối Gaussian trên biểu đồ sẽ biểu diễn dạng phân phối của biểu đồ histogram đó, đỉnh là điểm dữ liệu được xuất hiện nhiều nhất trong tập dữ liệu đang quan sát. Bên cạnh đó 2 đường biểu diễn mean và median của dữ liệu giúp nhóm so sánh được số quốc gia vượt trên trung bình hoặc dưới trung bình

Biểu đồ Histogram biểu diễn phân phối dữ liệu của Tot Cases/1M pop và Deaths/1M pop cho tới ngày 8/3/2023



Nhận xét:

❖ **Nhận xét về biểu đồ Histogram của Tot cases/1M pop**

Với số khoảng chia bằng 20, biểu đồ histogram cho thấy phân phối của dữ liệu Tot Cases/1M pop trên các quốc gia. Điều này cho thấy phân phối dữ liệu gần với phân phối chuẩn, có đỉnh hình chuông tại khoảng giá trị 0-5000 và có đuôi phân phối dài về bên phải, tương ứng với các quốc gia có số ca nhiễm một triệu dân cao hơn.

Ngoài ra, biểu đồ còn cho thấy có một số quốc gia có số ca nhiễm một triệu dân rất cao, vượt xa số liệu của phần lớn các quốc gia khác, với số lượng quốc gia này giảm dần khi giá trị số ca nhiễm tăng lên.

Từ đó ta có thể rút ra nhận định như sau:

- Phần lớn các quốc gia có số ca nhiễm trên một triệu dân cao
- Chưa đến 100 quốc gia có Tot Cases/1M pop dưới 100000 nghĩa là số ca nhiễm của các quốc gia này chiếm 10% dân số.
- Và ngược lại, có khoảng 12 quốc gia có Tot Cases/1M pop trên 600000, chiếm hơn 60% dân số của quốc gia.
- Quốc gia có số Tot Cases/1M pop cao nhất có thể lên đến 700000, nghĩa là cứ một triệu dân sẽ có đến 700000 người bị nhiễm bệnh. Điều này cho thấy quốc gia đó đã trải qua nhiều đợt bùng phát dịch nghiêm trọng, dẫn đến mật độ nhiễm bệnh của quốc gia đó ở mức báo động.

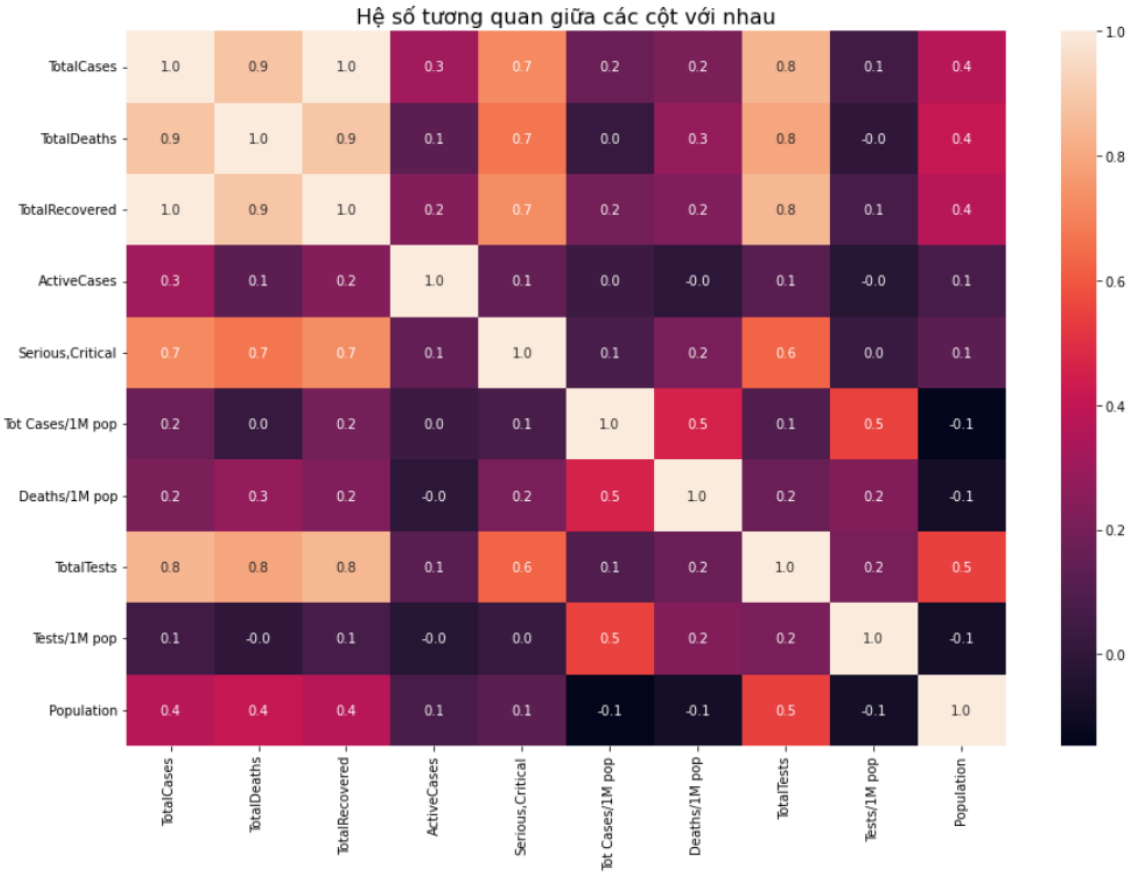
❖ **Nhận xét về biểu đồ Histogram của Deaths/1M pop**

Tương tự biểu đồ Histogram của Tot cases/1M pop, với khoảng chia bằng 20 ta thấy được sự phân phối dữ liệu Deaths/1M pop nghiêng về bên trái biểu đồ và có phần đuôi dài về bên phải

Nhìn vào biểu đồ ta có thể thấy trung bình các quốc gia có số ca tử vong trên 1 triệu dân là khoảng 1200 ca, có phần lớn các quốc gia có khoảng 300 Deaths/1M

Tuy nhiên, vẫn có một vài quốc gia có mật độ ca tử vong trên 1 triệu dân cao trên 6000 Deaths/1M pop, có nghĩa là cứ 1 triệu dân sẽ có 6000 người tử vong vì COVID, những quốc gia này đã bị mất mát rất lớn về người.

6. Xét mối quan hệ tương quan giữa các trường dữ liệu



Nhận xét:

- Dựa vào biểu đồ heatmap trên, ta có thể thấy được các mối tương quan giữa các trường dữ liệu trong bảng dữ liệu ban đầu. Trong đó, nếu 2 trường dữ liệu có hệ số tương quan khác 0 là có sự tương quan với nhau. Các hệ số tương qua gần bằng 1 nghĩa là chúng có sự tương quan càng mạnh.

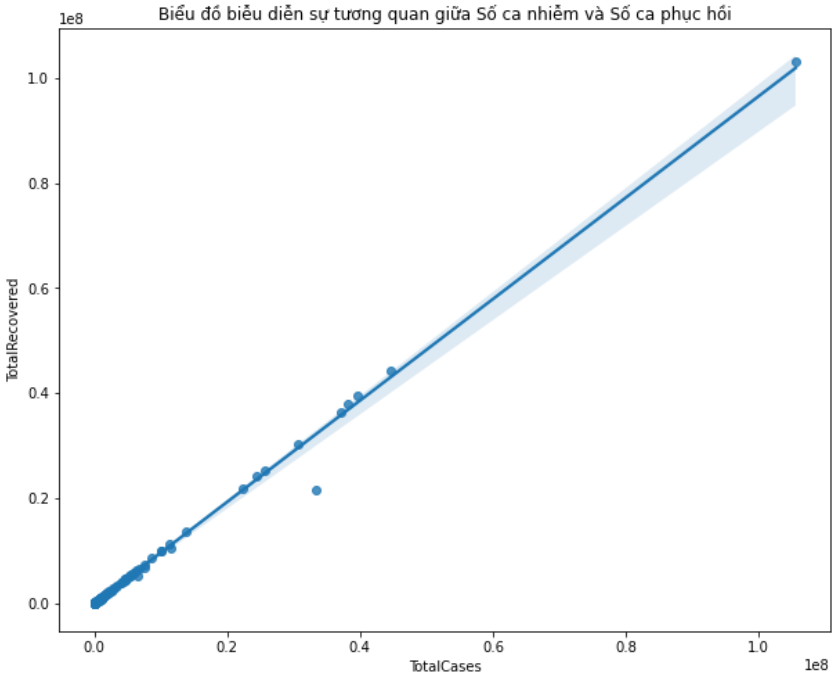
- Từ đó, ta có thể nhận xét như sau:
    - Các trường dữ liệu có mối tương quan cao là:
        + TotalCase và TotalRecovered: 1.0
        + TotalCase và TotalDeath: 0.9
        + TotalDeath và TotalRecovered: 0.9

+ TotalTest với [TotalCase,TotalDeath,TotalRecovered] : 0.8
+ Population và TotalTest: 0.5
+ Ngoài ra dù không phải là các cặp trường dữ liệu có mối tương quan cao nhất nhưng đáng chú ý khi của Population với các trường TotalCase, TotalDeath, TotalRecoverd là bằng nhau và bằng 0.4

- Và dựa vào những nhận xét về số liệu trên, ta có thể đưa ra những giả thuyết phù hợp với độ tương quan mạnh mẽ của các cặp trường dữ liệu

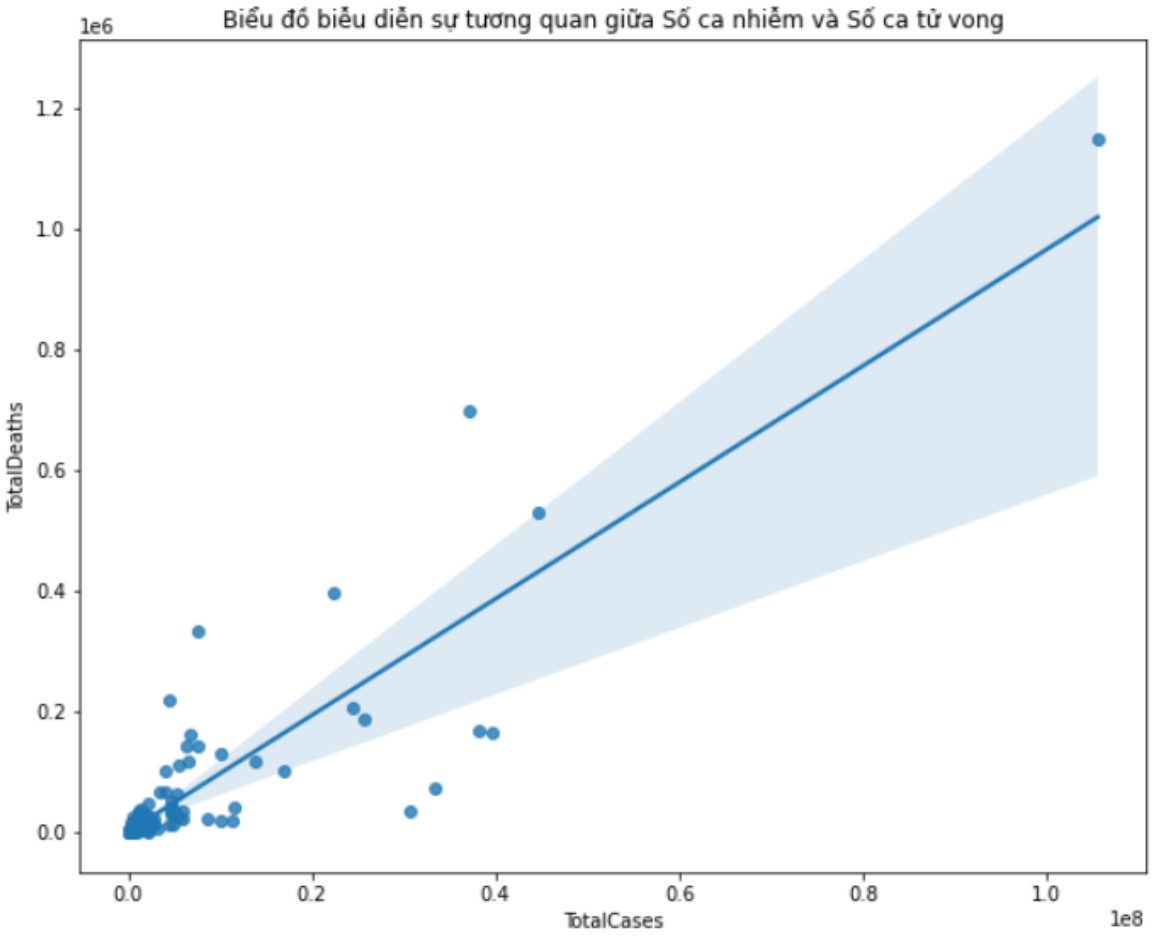
a. Số ca nhiễm tăng sẽ dẫn đến số ca phục hồi tăng

- Dựa vào mô hình **scatterplot**:



- **Nhận xét**: Ta có thể thấy khi số ca nhiễm càng cao, thì số ca phục hồi cũng càng cao. Vì vậy giả thuyết đưa ra là đúng. Điều này có thể hiểu rằng tại các quốc gia có số ca nhiễm càng lớn thì hệ thống y tế, chế độ dinh dưỡng và chăm sóc sức khỏe của các quốc gia đó được đẩy mạnh và phát triển, và dẫn đến việc khỏi bệnh sẽ diễn ra nhanh hơn so với các quốc gia bình thường, vì thế dẫn đến số lượng ca phục hồi cũng tăng theo.

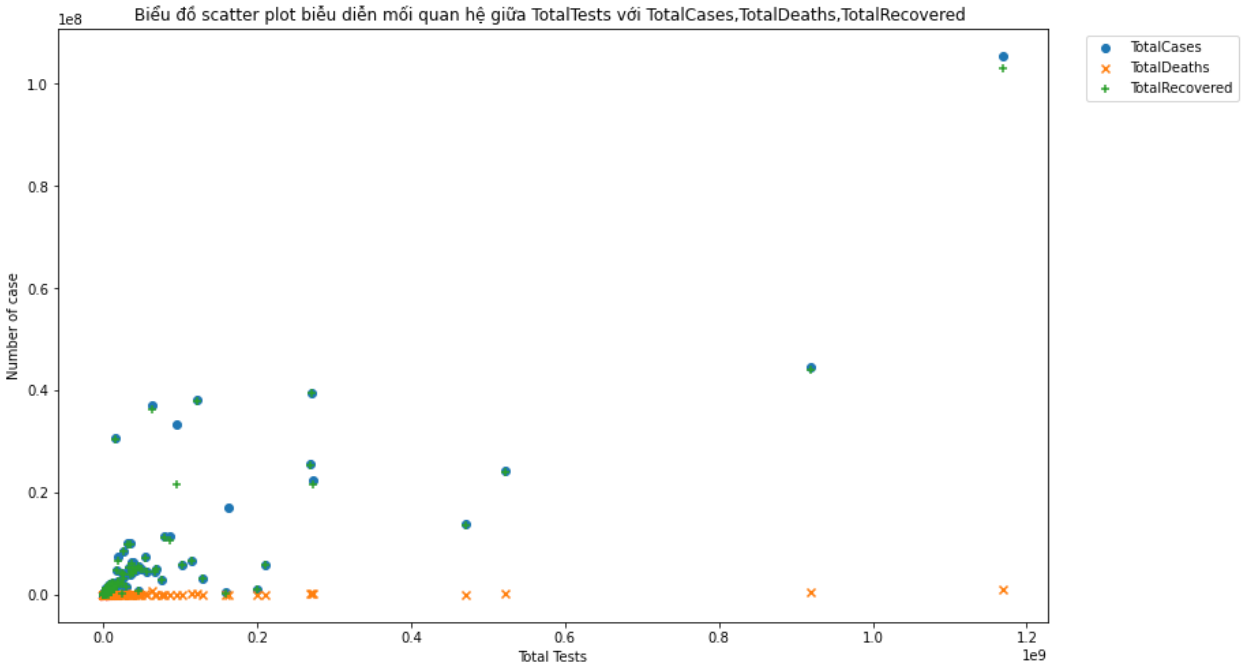
b. Số ca nhiễm tăng dẫn đến số ca tử vong cũng tăng



Nhận xét:

- Dựa vào biểu đồ **Scatter plot** trên, ta có thể thấy rõ: "Khi số ca nhiễm càng cao, thì số ca tử vong cũng càng cao."
- Ở đây, ta có thể hiểu : Khi số ca nhiễm tăng, cơ hội để virus lây lan và gây ra nhiều biến chứng và các tình trạng bệnh lý nặng cũng tăng lên, dẫn đến số ca tử vong tăng. Ngoài ra, khi số ca nhiễm tăng, hệ thống y tế của một quốc gia có thể bị quá tải, dẫn đến khả năng đáp ứng của hệ thống y tế giảm xuống, điều này có thể ảnh hưởng đến khả năng chữa trị và chăm sóc bệnh nhân, dẫn đến số ca tử vong tăng.

c. Số ca test càng nhiều thì các số liệu ca nhiễm, ca tử vong, ca phục hồi càng tăng

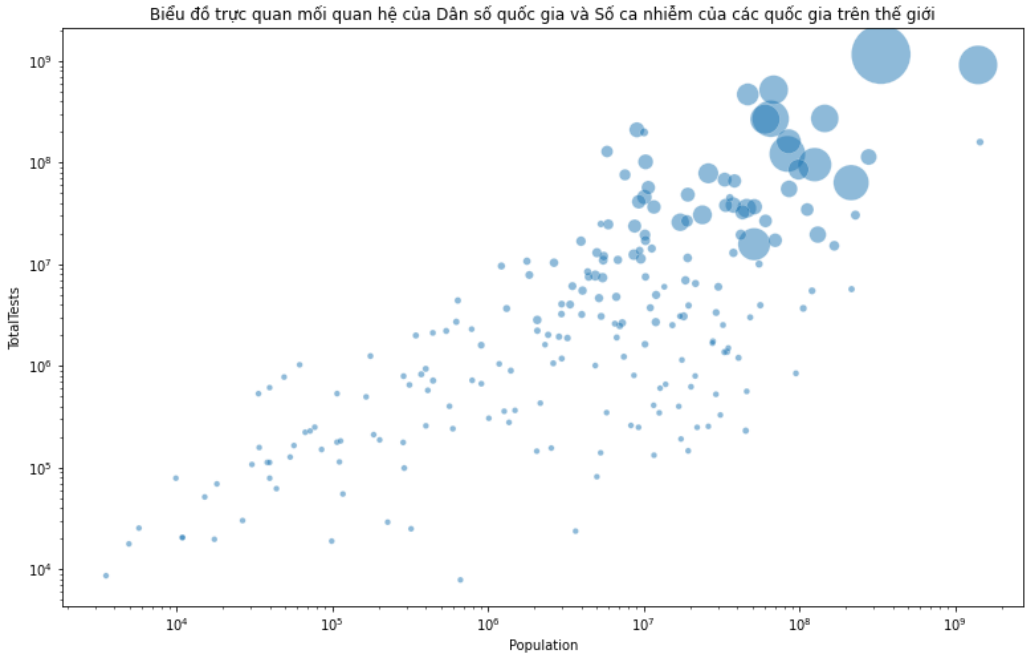


Nhận Xét:

- Nhìn chung, số ca nhiễm, số ca tử vong và số ca phục hồi đề tăng khi số lần xét nghiệm tăng
- Điều này cũng dễ hiểu khi các quốc gia tăng cường xét nghiệm thì số ca được ghi nhận cũng tăng theo và dựa vào mối tương quan dương của Số ca nhiễm với số ca tử vong, Số ca nhiễm với số ca hồi phục , thì khi số lần xét nghiệm tăng cũng kéo theo 2 số liệu về ca tử vong và ca phục hồi tăng theo

d. Quan sát về số ca tử vong do Covid tại tất cả các quốc gia

- Xét về các ca tử vong do dịch bệnh của các quốc, dựa vào biểu đồ **Bubble Plot**:



- **Nhận xét:**
  - Ta có thể thấy khi dân số càng cao thì số lần test càng cao, điều này dễ hiểu vì với các quốc gia có dân số càng đông thì việc triển khai thử nghiệm virus COVID-19 sẽ được quan tâm nhiều hơn để đảm bảo khả năng kiểm soát dịch bệnh, và số dân nhiều nên số lần test cũng sẽ cao.
  - Biểu đồ trên còn cho thấy kích thước các điểm cũng tăng dần theo dân số, kích thước này được biểu diễn theo số ca nhiễm được ghi nhận. Và tất nhiên, khi dân số càng đông thì khả năng lây nhiễm càng cao, và số ca nhiễm cũng sẽ tăng theo.

V. Nguồn tham khảo

[1]https://baotintuc.vn/the-gioi/lieu-bien-the-xbb15-co-gay-ra-lan-song-covid19-moi-o-chau-au-20230107100001469.htm

[2] https://www.worldometers.info/coronavirus

[3]https://soyte.hanoi.gov.vn/diem-bao/-/asset_publisher/4IVkx5Jltnbg/content/thong-tin-y-te-tren-cac-bao-ngay-8-3-2023

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[5]https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python