

LAB

THỐNG KÊ VÀ TRỰC QUAN HÓA DỮ LIỆU

Họ tên: Phạm Bảo Hân

MSSV: 19127135

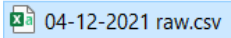
I – Đánh giá mức độ hoàn thành:

1. Crawl data: 100%
2. Thống kê và trực trực quan hóa dữ liệu: 100%
3. Báo cáo: 100%

II – Chi tiết thuật toán:

★ Những biểu đồ dùng để hiển thị kết quả bên dưới được lấy từ dữ liệu của ngày 01-12-2021.

1. Crawl data: *Crawl_data.ipynb*

- Dùng Selenium để mở trang <https://www.worldometers.info/coronavirus/>
- Dùng BeautifulSoup để phân tích HTML và tìm data.
- Crawl data thuộc tab 2 *Days Ago* bao gồm:
 - 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
 - Test/1M pop
 - Population
 - Continent
- Ghi lại data lấy được vào file csv.
- Đặt tên file csv là “*thời gian (ngày-tháng-năm) lấy data trừ đi 2 ngày raw.csv*”.
- Kết quả:
 - File csv chứa data () và in ra màn hình tên của file csv đó (*04-12-2021 raw.csv*).
 - Hình minh họa 1 phần của file csv:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	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	Test/1M pop	Population	Continent	
2	World	265,159,969	704,991	5,260,517	7,934	238,992,967	452,429	20,906,485	86,587	34,018	674.9				All	
3	Asia	82,386,300	84,213	1,220,663	1,455	79,655,539	72,306	1,510,098	29,832						Asia	
4	Europe	74,956,650	420,494	1,428,277	4,245	65,745,631	275,590	7,782,742	22,289						Europe	
5	North America	59,603,335	156,854	1,199,939	1,711	47,981,417	83,473	10,421,979	19,747						North America	
6	USA	49,878,049	147,434	808,116	1,353	39,463,245	73,599	9,606,688	13,714	149,438	2,421	759,601,375	2,275,818	333,770,690	North America	
7	South America	39,053,513	21,028	1,183,281	393	37,128,758	13,163	741,474	12,732						South America	
8	India	34,624,360	8,603	472,954	417	34,045,666	8,612	105,740	8,944	24,743	338	644,668,082	460,692	1,399,346,585	Asia	
9	Brazil	22,129,409	10,627	615,454	229	21,357,412	5,907	156,543	8,318	103,065	2,866	63,776,166	297,028	214,713,983	South America	
10	UK	10,377,785	49,884	145,424	143	9,156,066	29,938	1,076,295	895	151,734	2,126	366,108,215	5,352,883	68,394,590	Europe	
11	Russia	9,736,037	32,930	278,857	1,217	8,436,631	36,514	1,020,549	2,300	66,674	1,910	226,800,000	1,553,172	146,023,734	Europe	

2. Thống kê và trực trực quan hóa dữ liệu: *Ex_02.ipynb*

a. Set up:

- Ta có 1 class Region để chứa dữ liệu:

```

class Region:
    # Data of World, Asia, Africa, Europe, Oceania, North America, South America
    world = asia = africa = europe = oceania = north_america = south_america = ''
    countries_asia = []          # Data of all countries in Asia
    countries_africa = []        # Data of all countries in Africa
    countries_europe = []        # Data of all countries in Europe
    countries_oceania = []       # Data of all countries in Oceania
    countries_north_america = [] # Data of all countries in North America
    countries_south_america = [] # Data of all countries in South America
    countries_other = []         # Data of other countries
    all_countries = []          # Data of all countries in the World

    def set_all_countries(self):
        self.all_countries = (self.countries_asia + self.countries_africa + self.
                               countries_europe + self.countries_north_america + self.
                               countries_south_america + self.countries_oceania + self.countries_other)

```

- Tạo instance *region* của class *Region*.

b. Lấy dữ liệu từ file csv cho trước và lưu vào *region*:

```
def open_file(csv_file: str, rg: Region, header: list)
```

- Dòng đầu tiên là header sẽ được lưu vào array *header*.
- Xem giá trị của các trường dữ liệu trong một dòng data, nếu giá trị bằng rỗng hay “N/A” thì ta thay giá trị 0 vào ô đó, nếu là giá trị số thì ta cần xóa các dấu “,” và chuyển đổi từ string sang float.
- Gán các dòng dữ liệu vào biến thích hợp trong *region*:
 - Những dòng chứa dữ liệu tổng hợp sẽ được gán vào các biến tương ứng (*world, asia, africa, europe, oceania, north_america, south_america*).
 - Những dòng dữ liệu của quốc gia sẽ được phân vào list dựa vào châu lục quốc gia đó (*countries_asia, countries_africa, countries_europe, countries_oceania, countries_north_america, countries_south_america*), nếu quốc gia đó không thuộc châu lục nào thì ta đưa vào list *countries_other*.
- Riêng dòng dữ liệu vô nghĩa (không có tên và tên châu lục) sẽ được bỏ qua.
- Gọi method *set_all_countries* của *region* để biến *all_region* được gán giá trị là data của tất cả các quốc gia trên Thế giới.

c. Kiểm tra số liệu:

```
def check_total(rg: Region)
```

- Ta lấy tổng số liệu của tất cả châu lục và các nước không nằm trong châu lục nào so với số liệu của thế giới. Chỉ lấy và so sánh các trường mà châu lục có tổng hợp dữ liệu.

```
## Calculate the sum of data
for i in range(1, 9):
    sum[i] = rg.asia[i]+rg.africa[i]+rg.europe[i]+rg.oceania[i]+rg.north_america[i]+rg.south_america[i]
    for location in rg.countries_other:
        sum[i]+=location[i]
```

- Nếu các trường dữ liệu bằng nhau thì ta in ra tên trường dữ liệu và chữ “Same” bên cạnh, ngược lại là chữ “Different”.

```
## Check data
print('Continents - World')
for i in range(1, 9):
    if sum[i]==rg.world[i]:
        print(f"\t{header[i]}: Same")
    else:
        print(f"\t{header[i]}: Different")
```

Kết quả:

```
Continents - World
Total cases: Same
New cases: Same
Total deaths: Same
New deaths: Same
Total recovered: Same
New recovered: Same
Active cases: Same
Serious, Critical: Same
```

- Ta lấy tổng dữ liệu của các quốc gia và vùng lãnh thổ thuộc một châu lục nhất định và so với dữ liệu tổng hợp của châu lục đó.

```
## Calculate the sum of data
for i in range(1, 9):
    for country in rg.countries_asia:
        sum_asia[i] += country[i]
    for country in rg.countries_africa:
        sum_africa[i] += country[i]
    for country in rg.countries_europe:
        sum_europe[i] += country[i]
    for country in rg.countries_oceania:
        sum_oceania[i] += country[i]
    for country in rg.countries_north_america:
        sum_north_america[i] += country[i]
    for country in rg.countries_south_america:
        sum_south_america[i] += country[i]
```

- Nếu các trường dữ liệu ta so sánh khác nhau thì ta in tên trường dữ liệu, tên châu lục đó bên cạnh là chữ “Different”.

```

## Check data
print('Countries - Continent')
for i in range(1, 9):
    flag = True
    if not sum_asia[i]==rg.asia[i]:
        print(header[i])
        print('\tAsia: Different')
        flag = False
    if not sum_africa[i]==rg.africa[i]:
        if flag:
            print(header[i])
            flag = False
        print('\tAfrica: Different')
    if not sum_europe[i]==rg.europe[i]:
        if flag:
            print(header[i])
            flag = False
        print('\tEurope: Different')
    if not sum_oceania[i]==rg.oceania[i]:
        if flag:
            print(header[i])
            flag = False
        print('\tOceania/Australia: Different')
    if not sum_north_america[i]==rg.north_america[i]:
        if flag:
            print(header[i])
            flag = False
        print('\tNorth America: Different')
    if not sum_south_america[i]==rg.south_america[i]:
        if flag:
            print(header[i])
            flag = False
        print('\tSouth America: Different')

```

Kết quả:

```

Countries - Continent
Total recovered
    Africa: Different
    Oceania/Australia: Different
    North America: Different
    South America: Different
Active cases
    Africa: Different
    Oceania/Australia: Different
    North America: Different
    South America: Different

```

d. Thể hiện dữ liệu bằng Histogram:

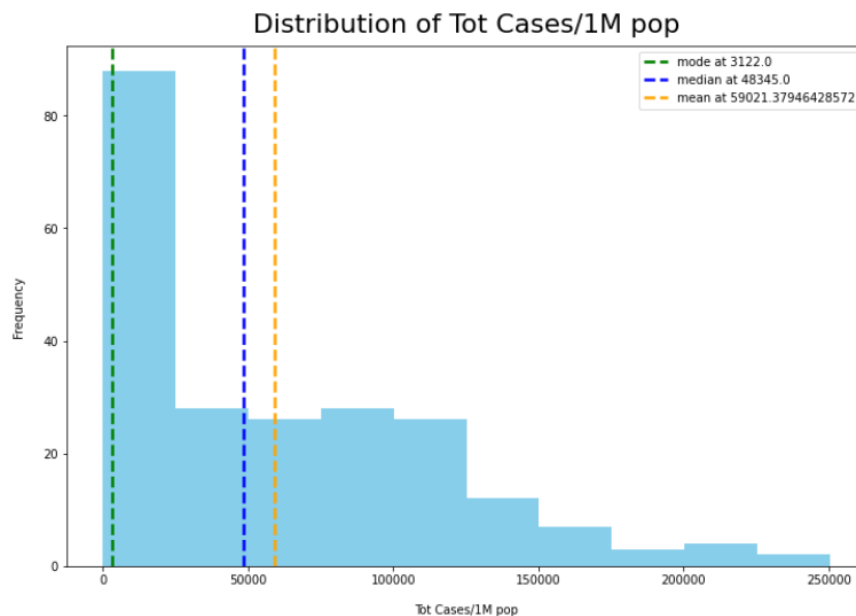
- Histogram có thể thể hiện tốt sự phân bố, giá trị mode, mean, median của một trường dữ liệu.
- Input:
 - index: index của trường dữ liệu đang xét.
- Đầu tiên, ta tạo một array chứa tất cả giá trị trong trường dữ liệu ta muốn tính. Ta lấy các giá trị này từ array *all_countries* của *region*.
- Tính giá trị mean, mode, median của của array vừa tạo ở trên.
- Sau đó ta hiển thị sự phân bố, giá trị mean, mode, median bằng histogram

```
def histogram(index: int):
    # Get the array with specified index
    arr = np.array(region.all_countries)[:,:index].astype(float)

    plt.figure(figsize=(12,8))
    plt.hist(arr, color="skyblue")
    plt.xlabel(header[index], labelpad=15)
    plt.ylabel("Frequency", labelpad=15)
    plt.title(f"Distribution of {header[index]}", y=1.012, fontsize=22)
    measurements = [mode(arr), np.median(arr), np.mean(arr)]
    names = ["mode", "median", "mean"]
    colors = ['green', 'blue', 'orange']
    for measurement, name, color in zip(measurements, names, colors):
        plt.axvline(x=measurement, linestyle='--', linewidth=2.5, label='{0} at {1}'.format(name, measurement), c=color)
    plt.legend()
```

Kết quả:

Hình minh họa cho trường dữ liệu Tot Cases/1M pop.



e. Thể hiện dữ liệu bằng Scatter plot:

- Scatter plot có thể thể hiện tốt mối quan hệ giữa 2 hoặc 3 trường dữ liệu.
- Do sự chênh lệch dữ liệu và giá trị dữ liệu lớn nên ta scale trục x, y theo giá trị 'log'.
- Input:
 - x_idx: index của trường dữ liệu mà ta muốn gán cho Ox.
 - y_idx: index của trường dữ liệu mà ta muốn gán cho Oy.
 - color: list các màu sắc để phân biệt các châu lục, nếu chỉ thể hiện 2 trường dữ liệu thì ta để các màu sắc giống nhau.

- name: tên biểu đồ
- Nếu thể hiện 2 trường dữ liệu, ta cho 2 trường đó là giá trị x, y của biểu đồ.
- Nếu thể hiện 3 trường dữ liệu thì trường thứ 3 là Continent. Lúc này ta phân biệt bằng màu sắc của các chấm.

```
def scatter_plot(x_idx: int, y_idx: int, color: list):
    asia = np.array(region.countries_asia[:, [x_idx, y_idx]].astype(float)
    africa = np.array(region.countries_africa[:, [x_idx, y_idx]].astype(float)
    europe = np.array(region.countries_europe[:, [x_idx, y_idx]].astype(float)
    north_america = np.array(region.countries_north_america[:, [x_idx, y_idx]].astype(float)
    south_america = np.array(region.countries_south_america[:, [x_idx, y_idx]].astype(float)
    oceania = np.array(region.countries_oceania[:, [x_idx, y_idx]].astype(float)
    other = np.array(region.countries_other[:, [x_idx, y_idx]].astype(float)

    plt.figure(figsize=(17,12))
    if not color[0] == color[1]: # If you need to classified each continent
        plt.title(f'Relationship between: {header[y_idx]} - {header[x_idx]} - Continent', y=1.012, fontsize=22)
    else:
        plt.title(f'Relationship between: {header[y_idx]} - {header[x_idx]}', y=1.012, fontsize=22)
    plt.xlabel(header[x_idx])
    plt.ylabel(header[y_idx])

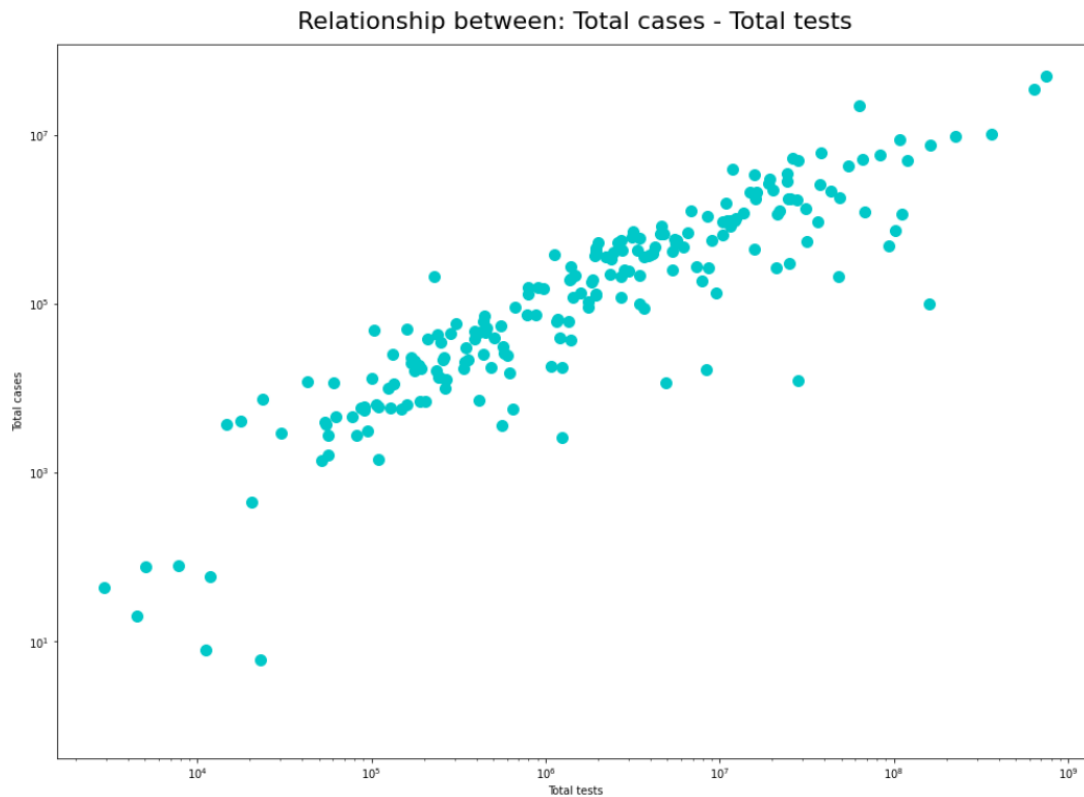
    plt.xscale(value='log')
    plt.yscale(value='log')

    plt.scatter(asia[:,0], asia[:,1], color=color[0], label="Asia", s=100)
    plt.scatter(africa[:,0], africa[:,1], color=color[1], label="Africa", s=100)
    plt.scatter(europe[:,0], europe[:,1], color=color[2], label="Europe", s=100)
    plt.scatter(north_america[:,0], north_america[:,1], color=color[3], label="North America", s=100)
    plt.scatter(south_america[:,0], south_america[:,1], color=color[4], label="South America", s=100)
    plt.scatter(oceania[:,0], oceania[:,1], color=color[5], label="Oceania/Australia", s=100)
    plt.scatter(other[:,0], other[:,1], color=color[6], label="Other", s=100)

    if not color[0] == color[1]: # If you need to classified each continent
        plt.legend()
```

Kết quả:

```
scatter_plot(11,1,["#00c8c8" for i in range(7)],"Total cases - Total test")
```



f. Thể hiện dữ liệu bằng Horizontal bar chart:

```
def ratio_hor_bar_chart(id1: int, id2: int, name: str)
```

- Trong bài này, trường dữ liệu numeric không có sẵn mà được tạo ra từ phép chia của 2 trường dữ liệu có sẵn.
- Thể hiện mối quan hệ giữa 2 trường dữ liệu trong đó 1 trường là Continent.
- Horizontal bar chart được sử dụng trong trường hợp này vì:
 - Phù hợp để biểu thị thứ tự độ lớn của giá trị. Từ đó ta có thể so sánh giá trị của trường dữ liệu giữa các Continent.
 - Tên một số châu lục khá dài nên ta hiển thị theo chiều ngang sẽ dễ nhìn hơn.
 - Đa dạng hóa loại biểu đồ trong bài làm.
- Input:
 - id1: index của trường dữ liệu là số bị chia.
 - id2: index của trường dữ liệu là số chia.
 - name: tên biểu đồ.
- Đầu tiên ta thực hiện phép chia giữa 2 trường dữ liệu theo từng châu lục, sau đó ta sắp xếp thứ tự giảm dần và hiển thị lên biểu đồ.


```
def ratio_hor_bar_chart(id1: int, id2: int, name: str):
    dic = {}
    dic['Asia'] = region.asia[id1]/region.asia[id2]
    dic['Africa'] = region.africa[id1]/region.africa[1]
    dic['Europe'] = region.europe[id1]/region.europe[1]
    dic['North America'] = region.north_america[id1]/region.north_america[1]
    dic['South America'] = region.south_america[id1]/region.south_america[1]
    dic['Oceania/Australia'] = region.oceania[id1]/region.oceania[1]
    sorted_dic = sorted(dic.items(), key=lambda x: x[1])

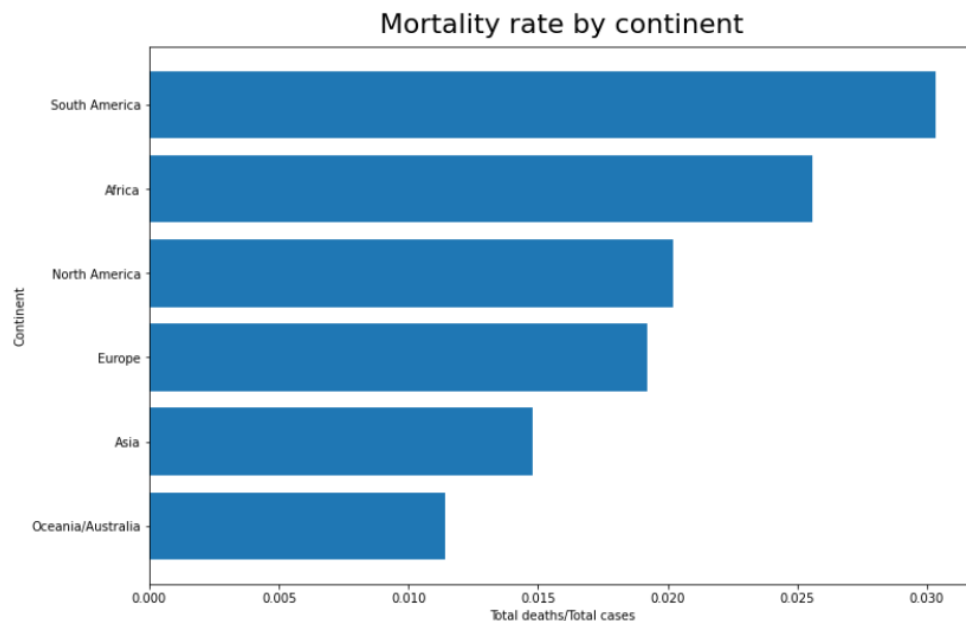
    plt.figure(figsize=(12,8))
    plt.barh(np.array(sorted_dic)[: ,0], np.array(sorted_dic)[: ,1].astype(float), align='center')

    plt.xlabel(f'{header[id1]}/{header[id2]}')
    plt.ylabel('Continent')
    plt.title(name, y=1.012, fontsize=22)

    plt.show()
```

Kết quả: Hình minh họa Morality rate by Continent/

```
ratio_hor_bar_chart(3,1,'Mortality rate by continent')
```



g. Thể hiện dữ liệu bằng Grouped vertical bar chart:

```
def grouped_ver_bar_chart(list_index: list, name: str)
```

- Grouped vertical bar chart có thể thể hiện nhiều trường dữ liệu có liên quan cùng một lúc. Mỗi trường dữ liệu là một nhóm các cột, mỗi cột biểu thị giá trị trường dữ liệu đó trong một Continent. Ta có thể dễ dàng so sánh giá trị giữa các cột trong một nhóm và mối tương quan giữa các nhóm với nhau.

- Do sự chênh lệch giá trị của các nhóm có thể rất lớn nên ở những nhóm giá trị nhỏ, ta có thể không thấy được cột trong nhóm đó.
- Biểu đồ chủ yếu để so sánh các giá trị.
- Input:
 - o lst_id: index của các trường dữ liệu numeric.
 - o name: tên biểu đồ.

```
def grouped_ver_bar_chart(lst_id: list, name: str):
    index = np.arange(len(lst_id))
    bar_width = 0.1275

    fig, ax = plt.subplots()
    fig.set_size_inches(18, 12)

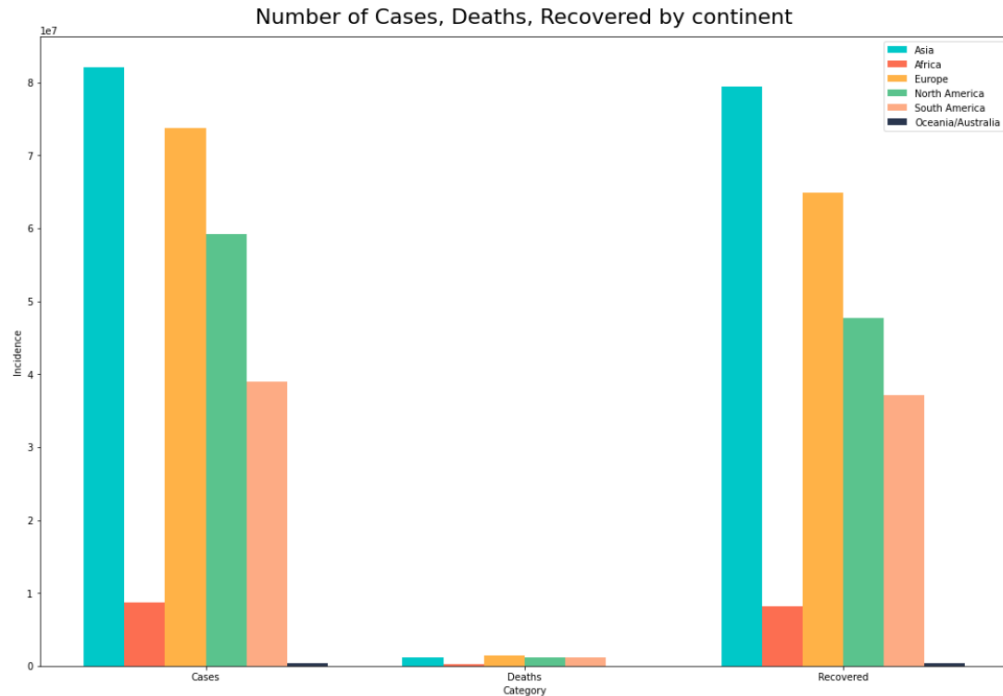
    ax.bar(index, [region.asia[i] for i in lst_id], bar_width, label="Asia", color="#00c8c8")
    ax.bar(index+bar_width, [region.africa[i] for i in lst_id], bar_width, label="Africa", color="#fc6e51")
    ax.bar(index+bar_width*2, [region.europe[i] for i in lst_id], bar_width, label="Europe", color="#ffb247")
    ax.bar(index+bar_width*3, [region.north_america[i] for i in lst_id], bar_width, label="North America", color="#5AC38D")
    ax.bar(index+bar_width*4, [region.south_america[i] for i in lst_id], bar_width, label="South America", color="#fdab84")
    ax.bar(index+bar_width*5, [region.oceania[i] for i in lst_id], bar_width, label="Oceania/Australia", color="#29364e")

    ax.set_xlabel('Category')
    ax.set_ylabel('Incidence')
    ax.set_title(name, y=1.012, fontsize=22)
    ax.set_xticks(index + bar_width*5/2)
    ax.set_xticklabels([header[i] for i in lst_id])
    ax.legend()

    plt.show()
```

Kết quả: Hình minh họa Total case, Deaths, Recovered theo Continent

```
grouped_ver_bar_chart([1,3,5], 'Number of Cases, Deaths, Recovered by continent')
```



h. Thể hiện dữ liệu bằng Stacked bar chart:

```
def stacked_ver_bar_chart(lst_id: list, name: str)
```

- Stacked bar chart dùng để thể hiện cùng lúc nhiều trường dữ liệu có quan hệ với nhau thành các cột.
- Ta dùng kiểu biểu đồ này để xem xét tổng quan mối tương quan giữa một trường dữ liệu với các trường còn lại hoặc trong một mối quan hệ tổng quát. Ngoài ra ta cũng có thể so sánh các cột với nhau.
- Do sự chênh lệch giá trị giữa các thành phần hoặc có thể quá lớn dẫn đến biểu đồ không thể hiện được thành phần hoặc cột có giá trị nhỏ so với phần còn lại.
- Ở đây ta kết hợp 3 trường dữ liệu numeric thành một cột và mỗi cột thể hiện cho các trường dữ liệu đó trong một Continent.
- Input:
 - lst_id: index của các trường dữ liệu numeric.

○ name: tên biểu đồ.

```
def stacked_ver_bar_chart(lst_id: list, name: str):
    index = np.arange(6)
    bar_width = 0.3

    fig, ax = plt.subplots()
    fig.set_size_inches(18, 12)

    ax.bar(index, [region.asia[lst_id[0]], region.africa[lst_id[0]], region.europe[lst_id[0]], region.north_america[lst_id[0]], region.south_america[lst_id[0]], region.oceania[lst_id[0]]], bar_width, label=header[lst_id[0]], color="#29364e")

    ax.bar(index, [region.asia[lst_id[1]], region.africa[lst_id[1]], region.europe[lst_id[1]], region.north_america[lst_id[1]], region.south_america[lst_id[1]], region.oceania[lst_id[1]]], bar_width, bottom=[region.asia[lst_id[0]], region.africa[lst_id[0]], region.europe[lst_id[0]], region.north_america[lst_id[0]], region.south_america[lst_id[0]], region.oceania[lst_id[0]]], label=header[lst_id[1]], color="#ffb247")

    ax.bar(index, [region.asia[lst_id[2]], region.africa[lst_id[2]], region.europe[lst_id[2]], region.north_america[lst_id[2]], region.south_america[lst_id[2]], region.oceania[lst_id[2]]], bar_width, bottom=[region.asia[lst_id[0]]+region.asia[lst_id[1]], region.africa[lst_id[0]]+region.africa[lst_id[1]], region.europe[lst_id[0]]+region.europe[lst_id[1]], region.north_america[lst_id[0]]+region.north_america[lst_id[1]], region.south_america[lst_id[0]]+region.south_america[lst_id[1]], region.oceania[lst_id[0]]+region.oceania[lst_id[1]]], label=header[lst_id[2]], color="#5AC38D")

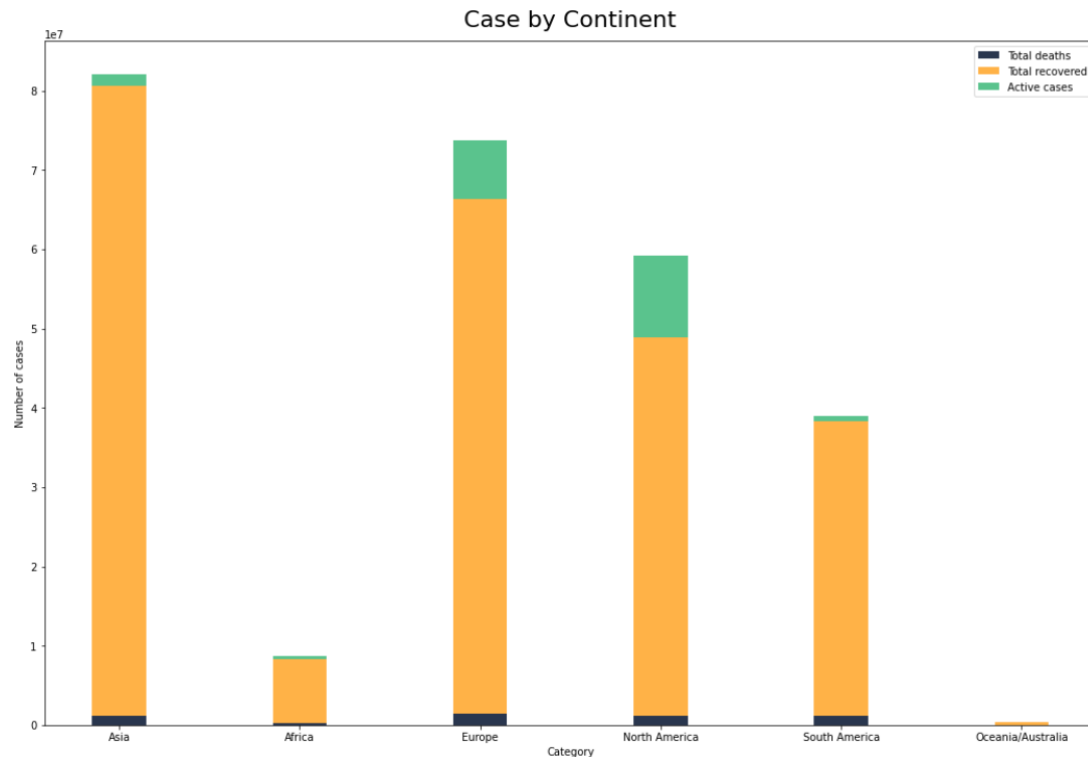
    ax.set_xlabel('Category')
    ax.set_ylabel('Number of cases')
    ax.set_title(name, y=1.012, fontsize=22)
    ax.set_xticks(index)
    ax.set_xticklabels(["Asia", "Africa", "Europe", "North America", "South America", "Oceania/Australia"])
    ax.legend()

    plt.show()
```

Kết quả:

Hình minh họa Total case, Deaths, Recovered theo Continent.

```
stacked_bar_chart([3,5,7], "Case by Continent")
```



III – Nhận xét:

★ Những biểu đồ bên dưới là dữ liệu của ngày 01-12-2021, nếu lấy từ ngày khác thì sẽ được chú thích.

1. Crawl Data:

- Trong file csv có số liệu của 224 quốc gia và vùng lãnh thổ, đồng thời có số liệu tổng hợp của 6 khu vực và của toàn thế giới.
- Có một dòng dữ liệu vô nghĩa (không có tên và tên khu vực):

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	721		15		706		0	0						

2. Sự chênh lệch dữ liệu:

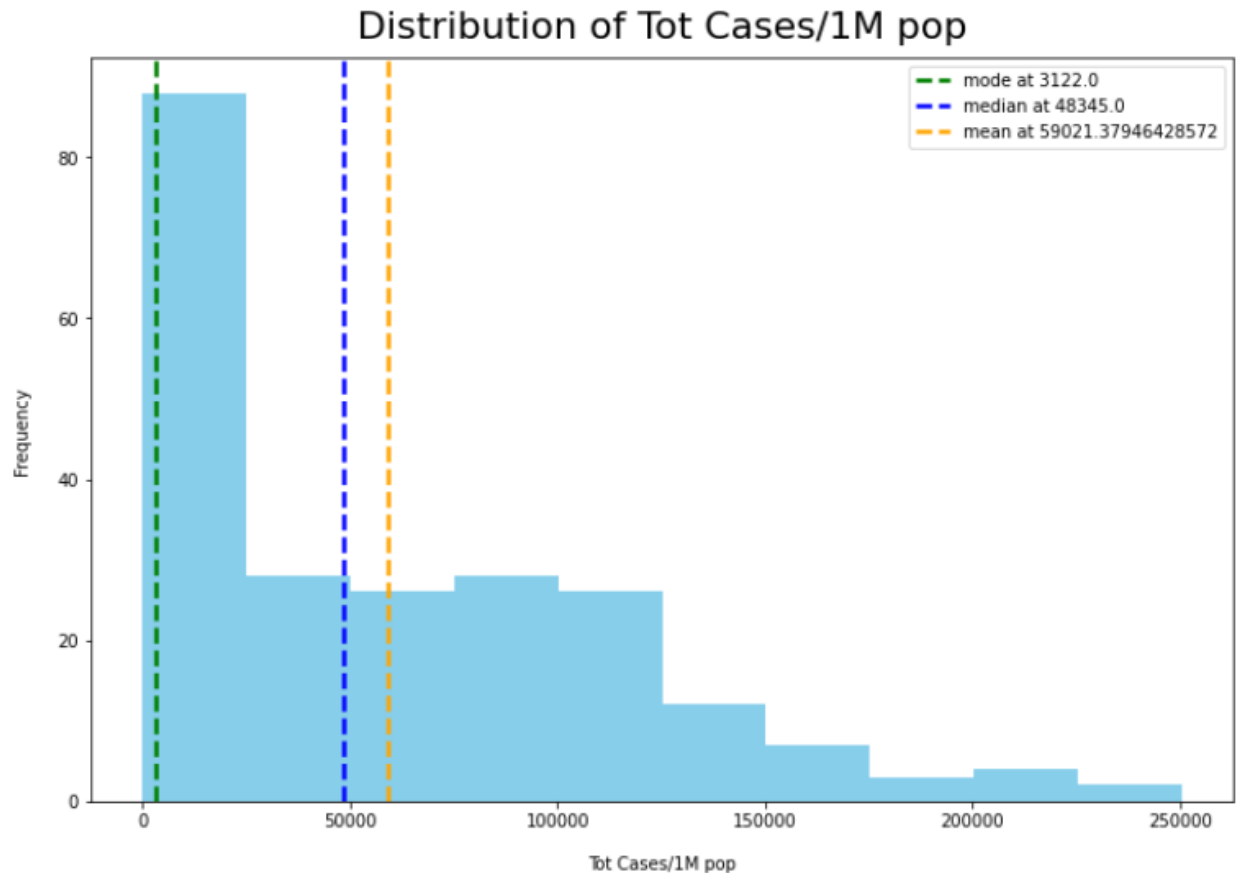
- Dữ liệu tổng hợp của các châu lục khớp với dữ liệu tổng hợp của thế giới.
- Khi so dữ liệu của các nước thuộc một châu lục với dữ liệu tổng hợp của châu lục đó, có diễn ra sự chênh lệch dữ liệu ở các trường có xuất hiện giá trị “N/A”: Total recovered, New recovered, Active cases.

Continents - World	Continents - World
Total cases: Same	Total cases: Same
New cases: Same	New cases: Same
Total deaths: Same	Total deaths: Same
New deaths: Same	New deaths: Same
Total recovered: Same	Total recovered: Same
New recovered: Same	New recovered: Same
Active cases: Same	Active cases: Same
Serious, Critical: Same	Serious, Critical: Same
Countries - Continent	Countries - Continent
Total recovered	Total recovered
Africa: Different	Africa: Different
Oceania/Australia: Different	Oceania/Australia: Different
North America: Different	North America: Different
South America: Different	South America: Different
New recovered	New recovered
South America: Different	South America: Different
Active cases	Active cases
Africa: Different	Africa: Different
Oceania/Australia: Different	Oceania/Australia: Different
North America: Different	North America: Different
South America: Different	South America: Different

Dữ liệu của ngày 01-12-2021 và 02-12-2021

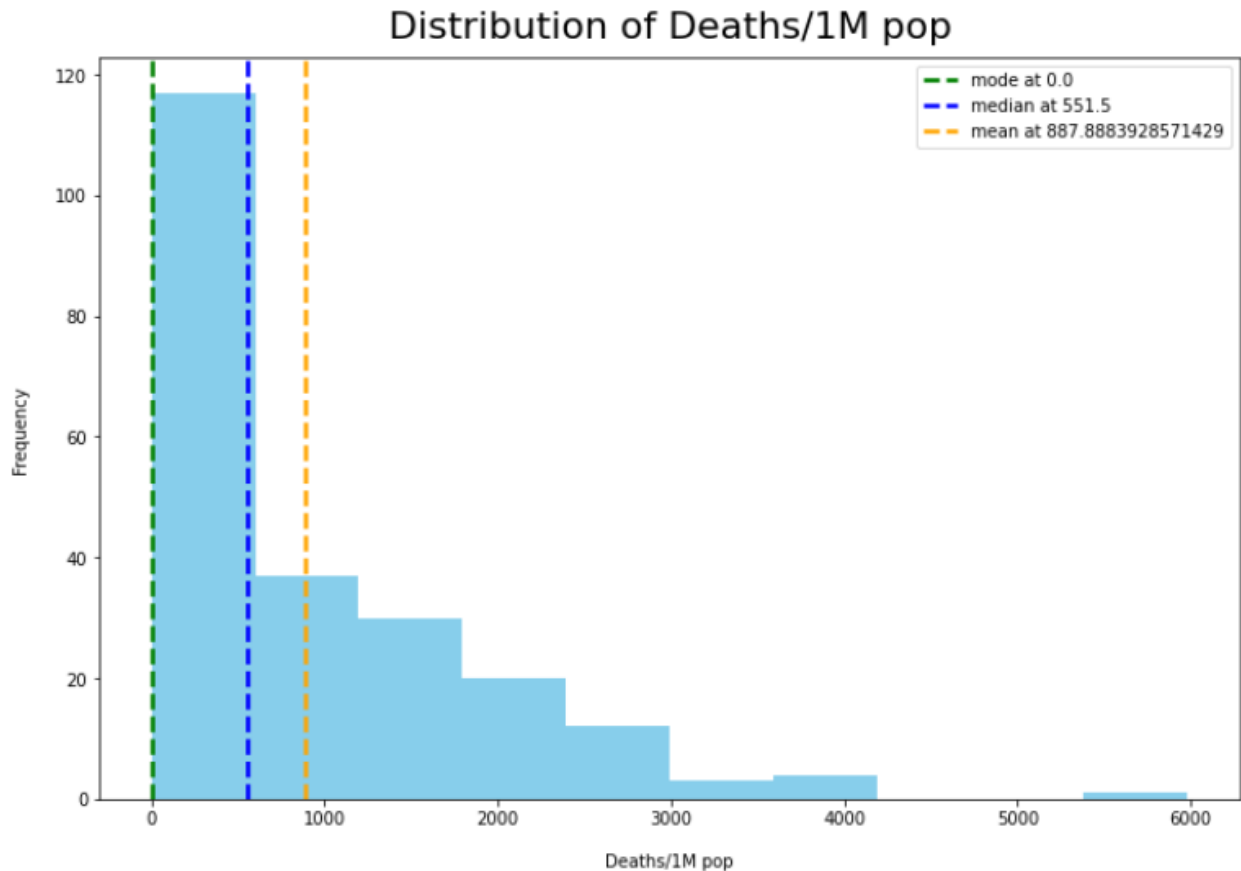
3. Distribution of Tot Cases/1M pop:

- Biểu đồ dùng để xem xét sự phân bố giá trị của trường dữ liệu Tot Cases/1M pop.
- Giá trị:
 - o Mode: 3122
 - o Median: 48345
 - o Mean: ~59021.38
- Biểu đồ này lệch dương, các giá trị chủ yếu tập hợp từ 0-50000.
- Một nửa số quốc gia có số ca nhiễm/ 1000000 dân số khá thấp (0-48345) tuy nhiên nửa kia lại số ca nhiễm có thể đạt tới gấp ~5 lần nửa trước đó (48345 - ~250000).



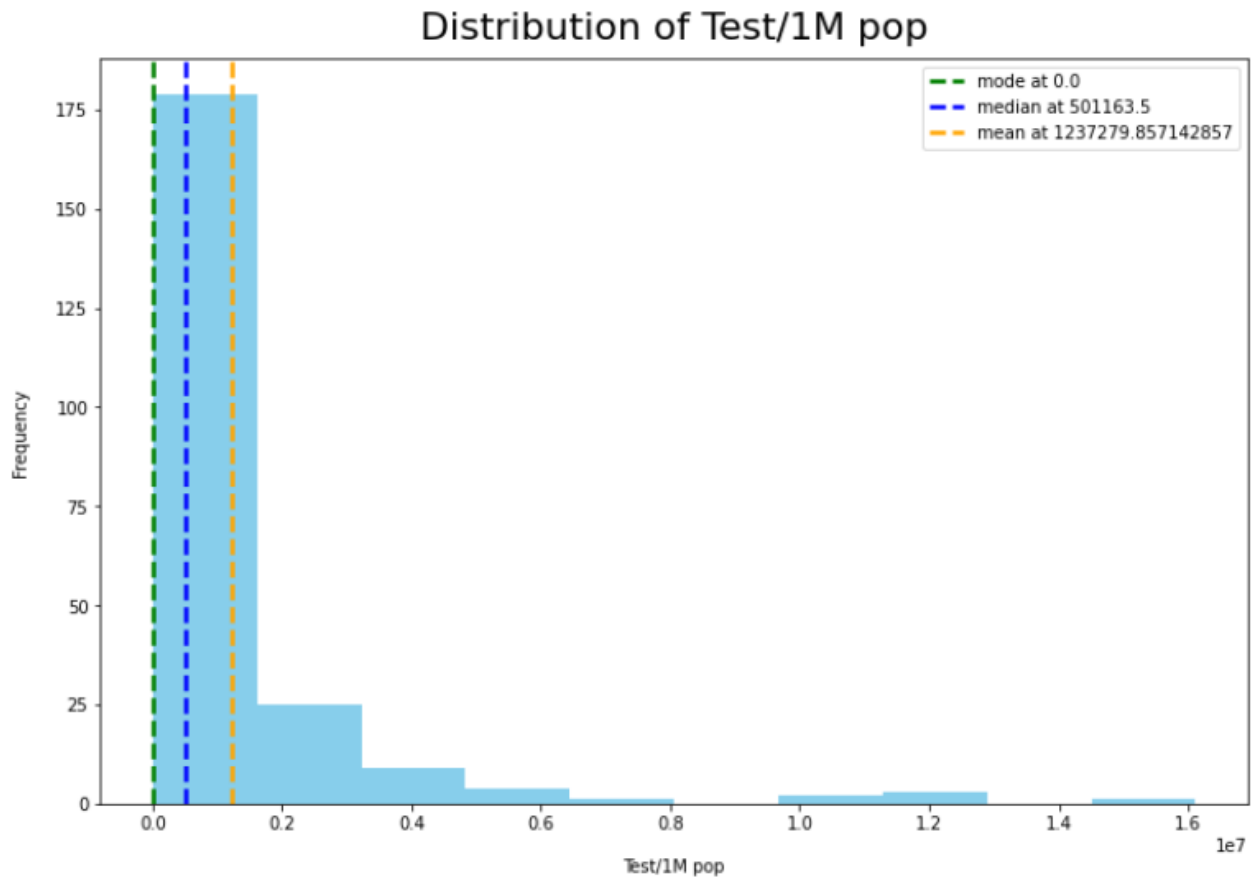
4. Distribution of Deaths/1M pop:

- Biểu đồ dùng để xem xét sự phân bố giá trị của trường dữ liệu Deaths/1M pop.
- Giá trị:
 - Mode: 0
 - Median: 551.5
 - Mean: ~887.88
- Biểu đồ này lệch dương, các giá trị chủ yếu tập hợp từ 0-10000.
- Một nửa số quốc gia có số ca tử vong/ 1000000 dân số từ 0-551,5. Tuy nhiên, nửa còn lại có số ca tử vong cao gấp ~7 lần (551,6 - ~4100) nửa trước đó và đặc biệt có một số quốc gia cao gấp ~10 lần (~5500 - ~6000).



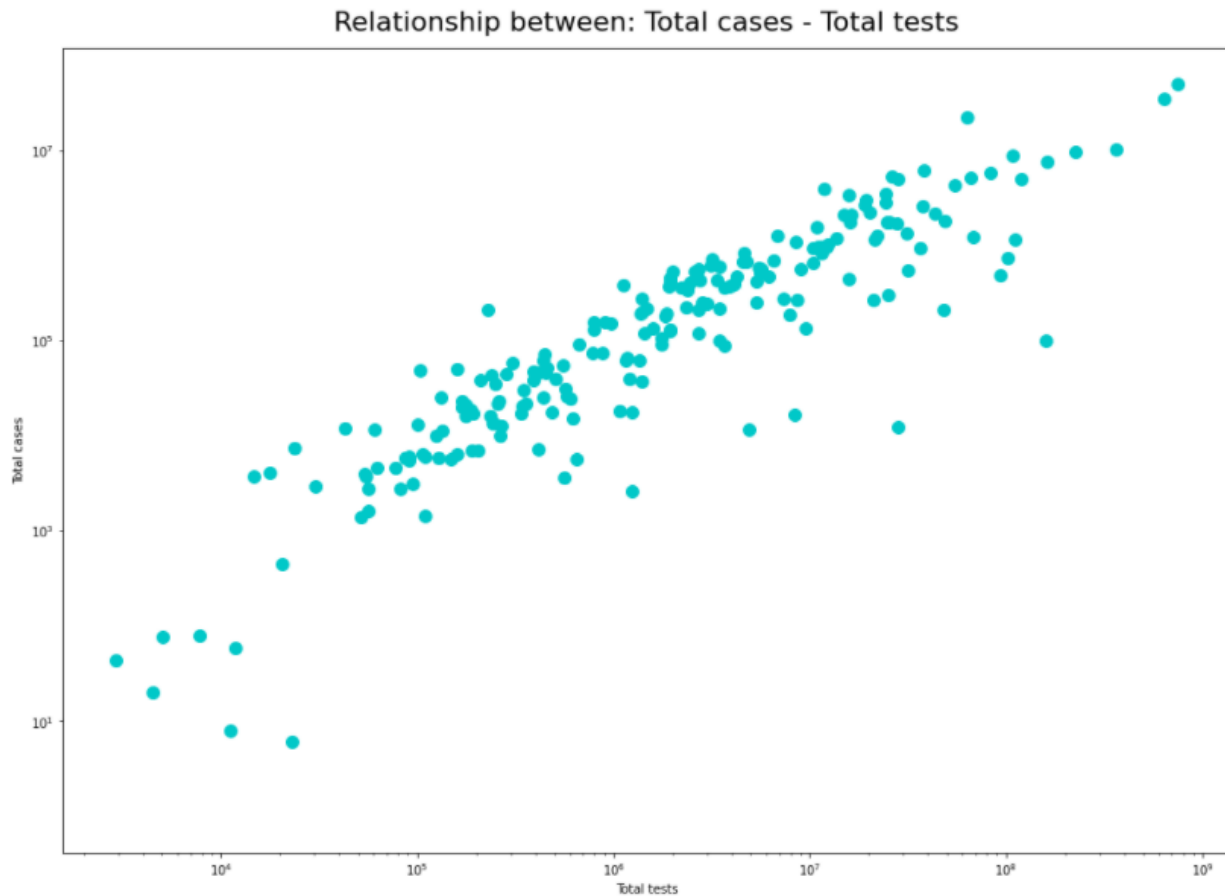
5. Distribution of Test/1M pop:

- Biểu đồ dùng để xem xét sự phân bố giá trị của trường dữ liệu Tests/1M pop.
- Giá trị:
 - Mode: 0
 - Median: 501163.5
 - Mean: ~887.88
- Biểu đồ này lệch dương, các giá trị chủ yếu tập hợp từ 0 - $\sim 0,2 \times 10^7$.
- Một nửa số quốc gia có số xét nghiệm/ 1000000 dân số từ 0-501163,5. Nửa còn lại có số xét nghiệm từ 501163,5 đến 1237279,857.
- Phần lớn quốc gia có tỉ lệ xét nghiệm khá nhỏ, nguyên nhân có thể là do năng lực xét nghiệm, điều kiện y tế còn hạn chế. Trong khi đó, những quốc gia điều kiện y tế dồi dào thì có tỉ lệ xét nghiệm chênh lệch hoàn toàn với phần còn lại ($\sim 1,0 \times 10^7 - \sim 1,3 \times 10^7$ và $\sim 1,4 \times 10^7 - \sim 1.6 \times 10^7$)



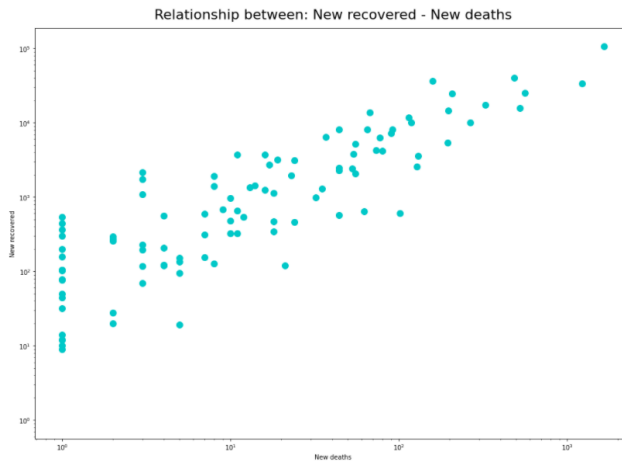
6. Relationship between: Total Cases – Total Test:

- Biểu đồ dùng để thể hiện mối quan hệ giữa 2 trường dữ liệu Total Cases và Total Test.
- Từ biểu đồ, ta nhận thấy mối quan hệ giữa Total Cases và Total Test là positive correlation. Đây là mối tương quan khá lớn và rõ ràng.
- Trong thực tế, những quốc gia có số ca nhiễm tăng thì sẽ tăng cường xét nghiệm hơn.

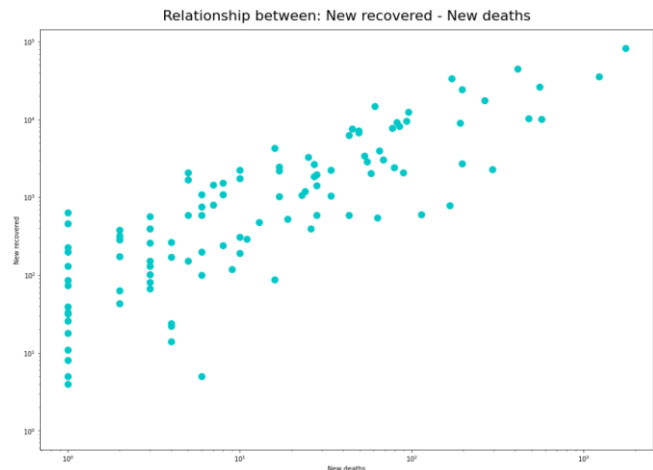


7. Relationship between: New recovered – New deaths:

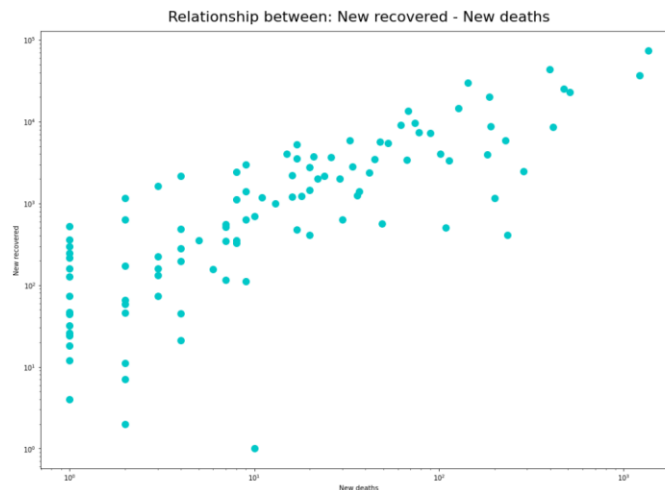
- Biểu đồ dùng để thể hiện mối quan hệ giữa 2 trường dữ liệu New recovered và New deaths.
- Từ biểu đồ, ta nhận thấy mối quan hệ giữa New recovered và New deaths là positive correlation. Đây là mối tương quan không quá lớn và rõ ràng.
- Trong thực tế, số ca tử vong và số ca hồi phục đều tỉ lệ với số ca nhiễm nên chúng cũng đồng thời tỉ lệ với nhau. Tuy nhiên, nhiều quốc gia có năng lực điều trị tốt hơn thì đạt được số ca hồi phục cao hơn các quốc gia có cùng số ca tử vong.



Dữ liệu ngày 01-12-2021



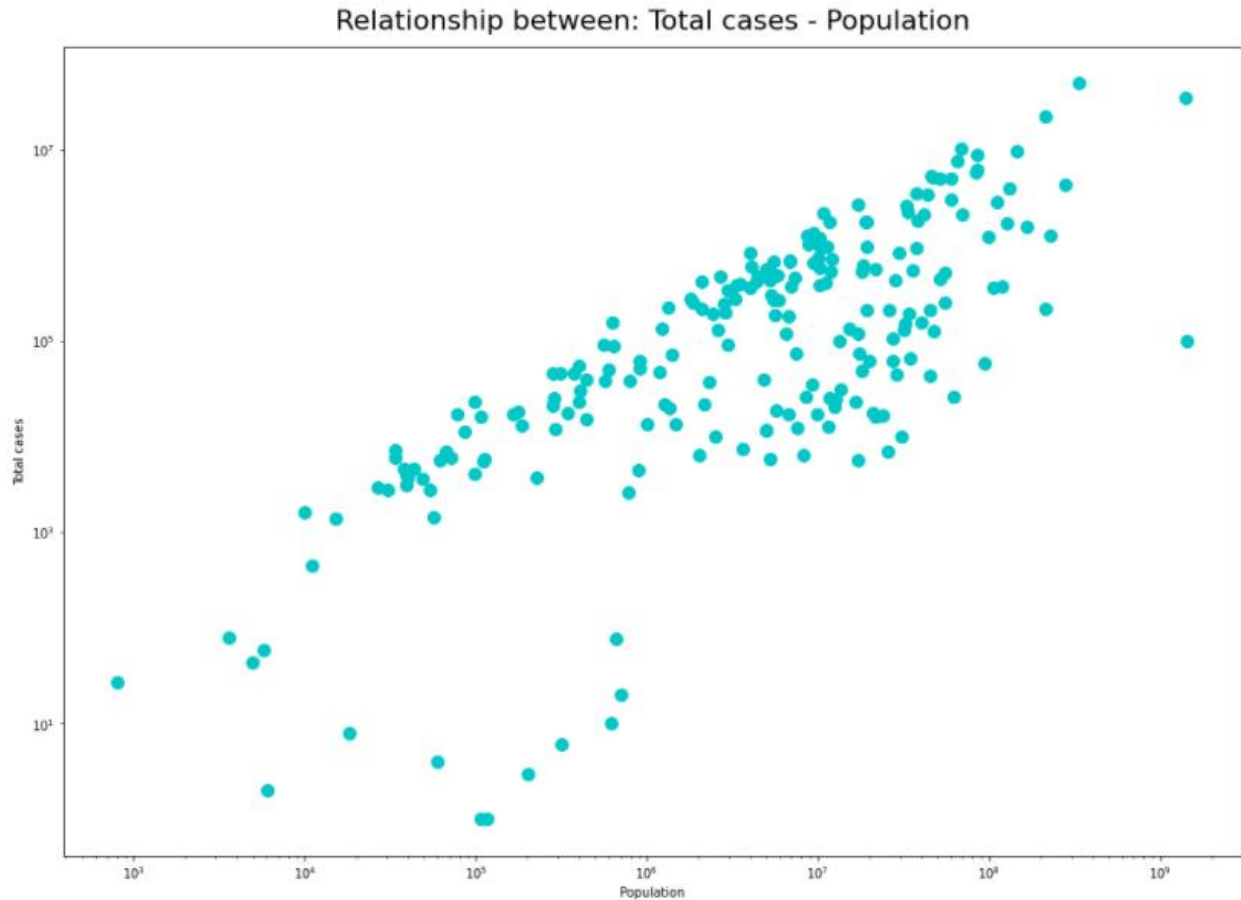
Dữ liệu ngày 02-12-2021



Dữ liệu ngày 03-12-2021

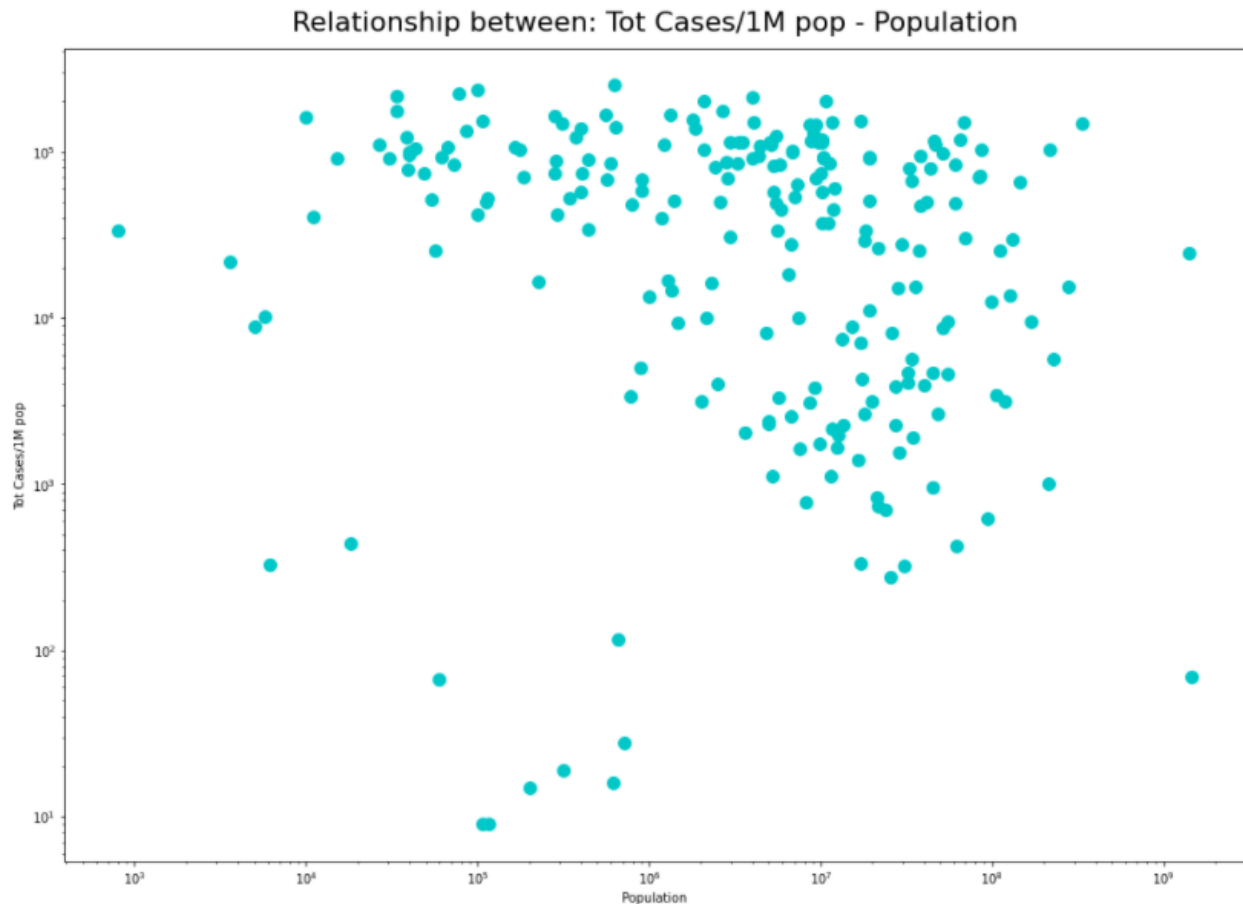
8. Relationship between: Total cases – Population:

- Biểu đồ dùng để thể hiện mối quan hệ giữa 2 trường dữ liệu Total Cases và Population.
- Từ biểu đồ, ta nhận thấy mối quan hệ giữa Total case và Population là positive correlation. Đây là mối tương quan không quá lớn và rõ ràng.
- Trong thực tế, những quốc gia có số dân đông thường có số ca nhiễm nhiều hơn.
- Lý do có thể do dân số đông thì diện tích đất trung bình trên đầu người thấp, dịch bệnh dễ dàng lây lan. Đồng thời, các quốc gia có dân số đông thường là các quốc gia thuộc thế giới thứ ba, dân trí không cao, khả năng chống dịch kém, điều kiện y tế không đủ.
- Bên cạnh đó vẫn xuất hiện nhiều quốc gia có số ca nhiễm ít hơn rất nhiều so với các quốc gia cùng dân số.



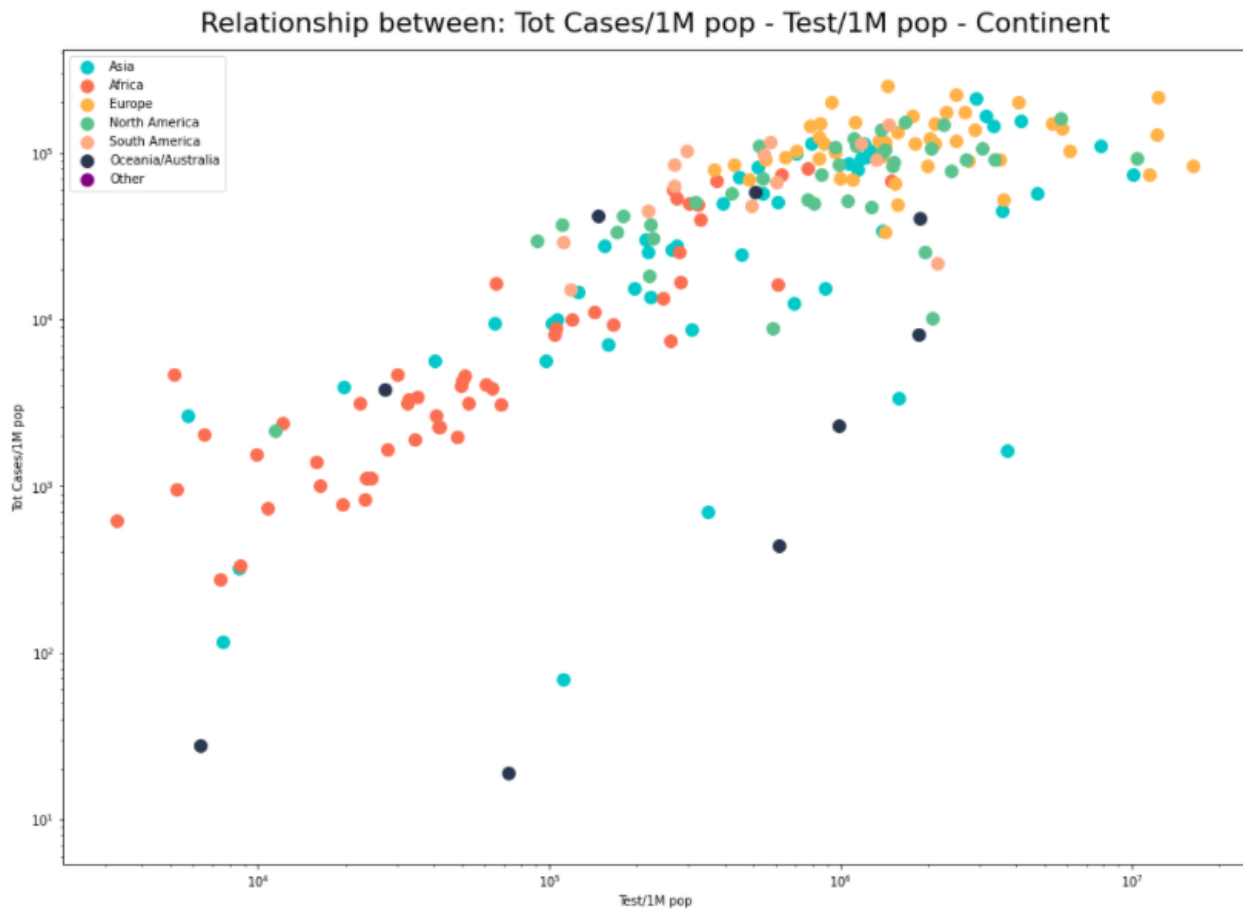
9. Relationship between: Tot case/1M pop – Population:

- Biểu đồ dùng để thể hiện mối quan hệ giữa 2 trường dữ liệu Tot Cases/1M pop và Population.
- Từ biểu đồ, ta nhận thấy không có mối quan hệ giữa Tot case/1M pop và Population.
- Tuy ở biểu đồ trên ta thấy được mối quan hệ giữa Tot Cases và Population là positive correlation nhưng thực tế, những quốc gia có đông dân số hơn không đồng nghĩa với việc có tỉ lệ nhiễm cao hơn.



10. Relationship between: Total cases/1M pop – Test/1M pop – Continent:

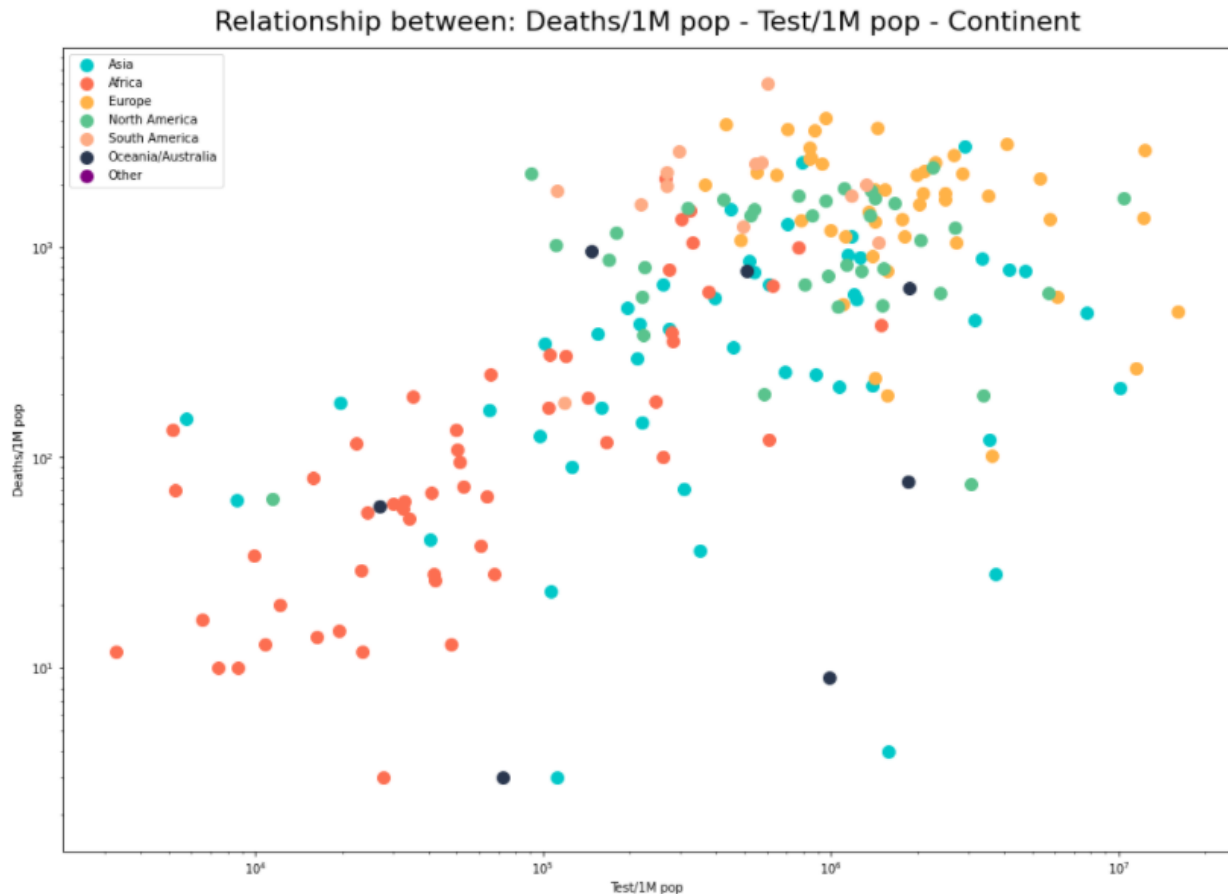
- Biểu đồ dùng để thể hiện mối quan hệ giữa 3 trường dữ liệu Total cases/1M pop, Test/1M pop và Continent.
- Từ biểu đồ, ta nhận thấy mối quan hệ giữa Total cases/1M pop và Test/1M pop là positive correlation. Đây là mối tương quan trung bình.
- Trong thực tế, ở những nước có tỉ lệ nhiễm tăng thì chính phủ cũng tăng cường xét nghiệm dẫn đến tỉ lệ xét nghiệm cũng tăng theo.
- Tuy nhiên, ở một số nước do điều kiện y tế không phát triển nên tỉ lệ xét nghiệm lại thấp hơn so với các quốc gia có cùng tỉ lệ nhiễm.
- Nhìn chung, những quốc gia thuộc Châu Phi có tỉ lệ nhiễm và xét nghiệm ít hơn các châu lục còn lại.
- Các quốc gia Châu Âu đều có tỉ lệ nhiễm và tỉ lệ xét nghiệm cao.
- Các quốc gia thuộc Châu Á có tỉ lệ nhiễm và tỉ lệ xét nghiệm đa dạng từ thấp đến cao.



11. Relationship between: Deaths/1M pop – Test/1M pop – Continent:

- Biểu đồ dùng để thể hiện mối quan hệ giữa 3 trường dữ liệu Deaths/1M pop, Test/1M pop và Continent.
- Từ biểu đồ, ta thấy được mối quan hệ giữa Deaths/1M pop và Test/1M pop là positive correlation. Đây là mối tương quan rất yếu và không rõ ràng.
- Ở những nước có tỉ lệ tử vong cao thường do tỉ lệ mắc bệnh cao, từ đó dẫn đến việc tăng cường xét nghiệm hơn (tỉ lệ mắc bệnh và tỉ lệ xét nghiệm được phân tích ở biểu đồ trên).
- Tuy nhiên trên thực tế, tùy vào nhiều yếu tố khác nhau như phương án phòng chống dịch và điều kiện y tế của từng quốc gia mà tỉ lệ xét nghiệm khác nhau dù tỉ lệ tử vong giống nhau.
- Tại phần lớn các quốc gia Châu Phi, cả tỉ lệ tử vong và tỉ lệ xét nghiệm đều thấp.
- Các quốc gia Bắc Mỹ có tỉ lệ tử vong và tỉ lệ xét nghiệm đều trên mức trung bình.

- Các quốc gia Nam Mỹ có tỉ lệ tử vong cao nhưng tỉ lệ xét nghiệm chỉ ở mức trung bình.
- Nhìn chung các quốc gia Châu Âu có tỉ lệ tử vong khá cao và tỉ lệ xét nghiệm cũng cao hơn so với các châu lục khác.



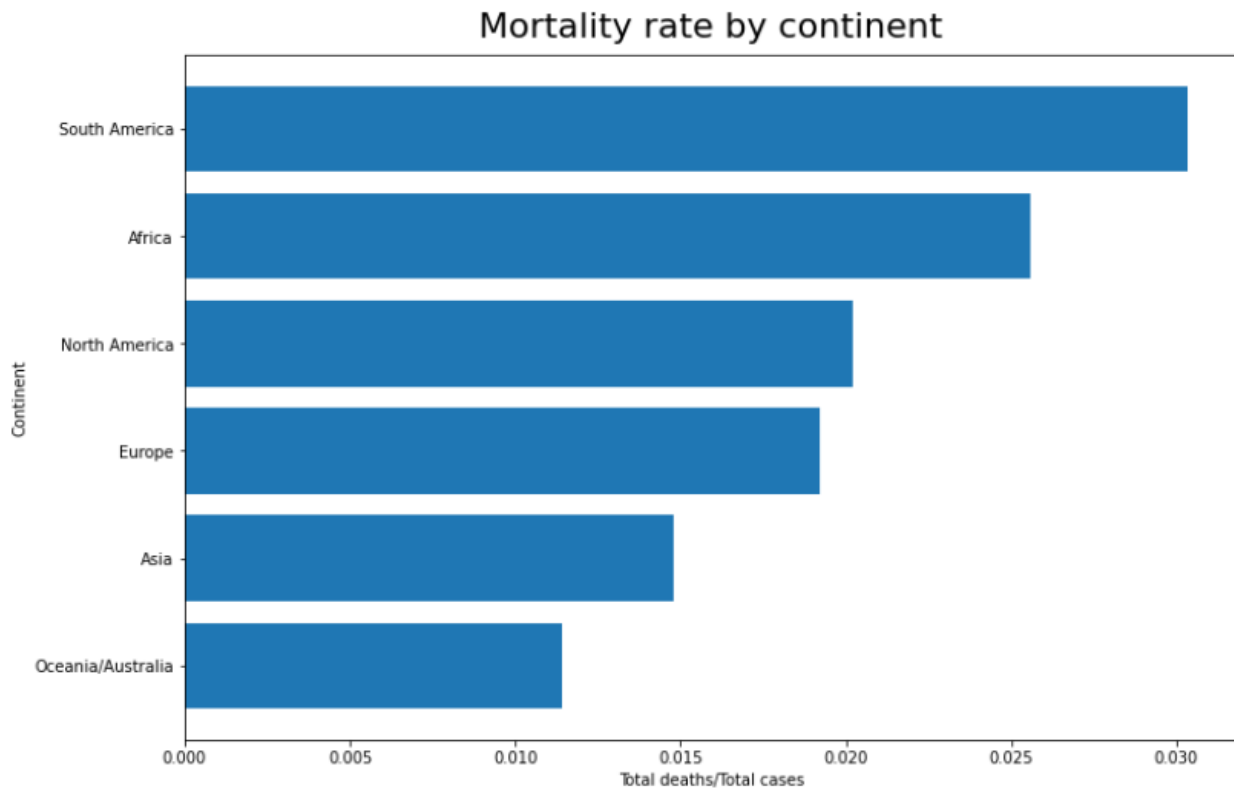
12. Mortality rate by Continent:

- Biểu đồ thể hiện tỉ lệ tử vong sau khi mắc COVID của các châu lục.

$$\text{Mortality rate} = \frac{\text{Deaths}}{\text{Total cases}}$$

- Dựa vào biểu đồ, ta thấy được các quốc gia Nam Mỹ có tỉ lệ tử vong cao nhất, sau đó đến Châu Phi, Bắc Mỹ, Châu Âu, Châu Á và thấp nhất là Châu Úc.
- Các phân tích trên đều cho Châu Phi là một quốc gia có tỉ lệ nhiễm thấp hơn các châu lục còn lại, nhưng biểu đồ này cho thấy tỉ lệ tử vong sau khi mắc COVID ở Châu Phi rất cao. Điều này thể hiện các quốc gia Châu phi

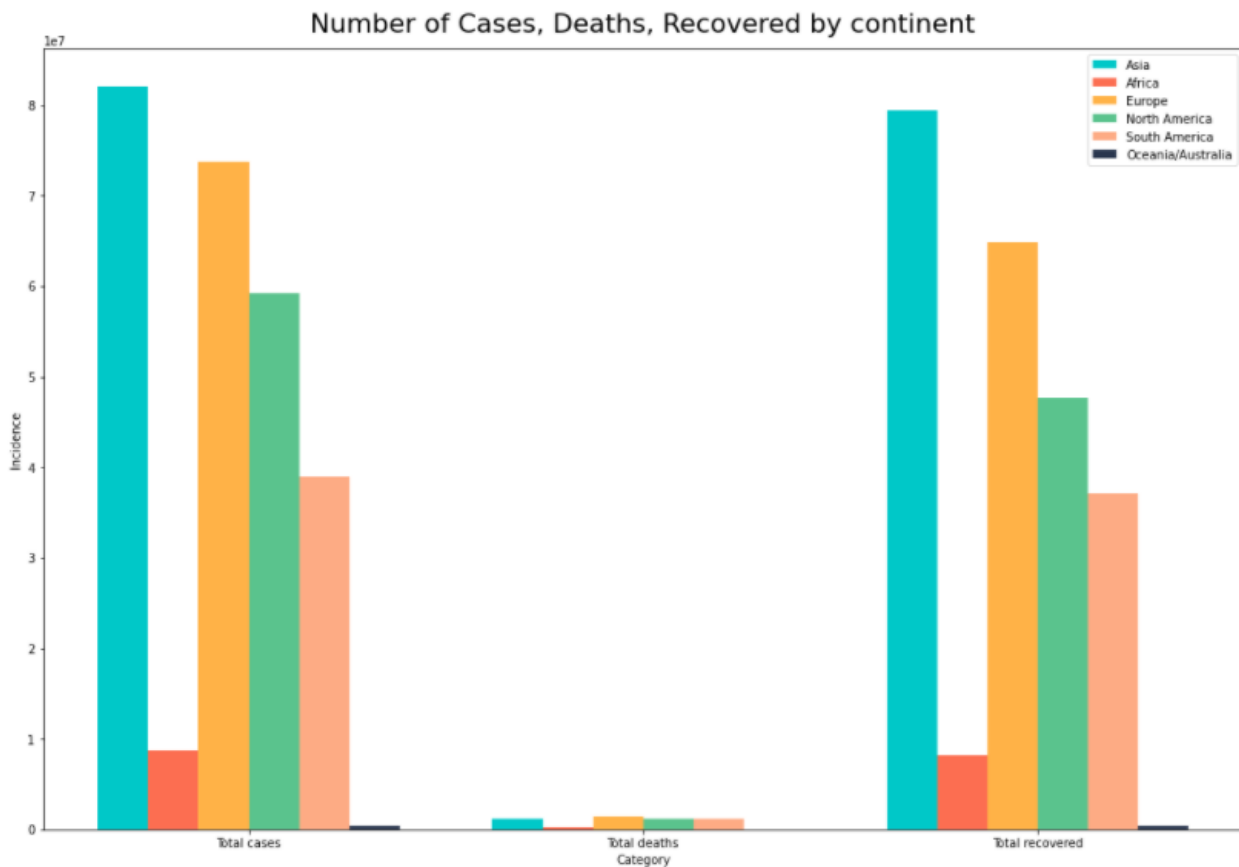
- có tỉ lệ mắc thấp không phải do điều kiện y tế tốt mà là do các yếu tố khác như tự nhiên, dân số,...
- Các quốc gia Nam Mỹ có tỷ lệ nhiễm bệnh/ 1000000 dân số như đã phân tích ở trên, nhìn chung cao hơn các quốc gia thuộc châu lục khác và từ biểu đồ này, ta cũng thấy tỉ lệ tử vong sau khi mắc COVID cũng lớn nhất trong các châu lục. Điều này chứng tỏ dịch bệnh COVID ảnh hưởng lớn nhất đến các quốc gia ở Nam Mỹ. Nguyên nhân có thể là do chính phủ yếu kém, vốn đã nổi tiếng với tham nhũng, băng đảng, ma túy,..., cùng nền kinh tế, điều kiện y tế kém phát triển, dân trí thấp.
 - Các quốc gia thuộc Châu Úc gồm 2 quốc gia phát triển nổi bật là Úc và New Zealand có điều kiện kinh tế, y tế phát triển, chính sách thực hiện phong tỏa kéo dài ở các thành phố lớn của Úc đã hạn chế được dịch bệnh. Các quốc gia còn lại có dân cư đặc trưng thường là các bộ tộc, sống ở các đảo, quần đảo hoặc quốc gia có diện tích nhỏ, ít tiếp xúc với bên ngoài nên nguy cơ mắc bệnh thấp.



13. Number of Cases, Deaths, Recovered by continent:

- Biểu đồ thể hiện số ca nhiễm, tử vong, phục hồi theo Châu lục.
- Ta thấy có sự giống nhau về thứ tự ở ca nhiễm và ca phục hồi.

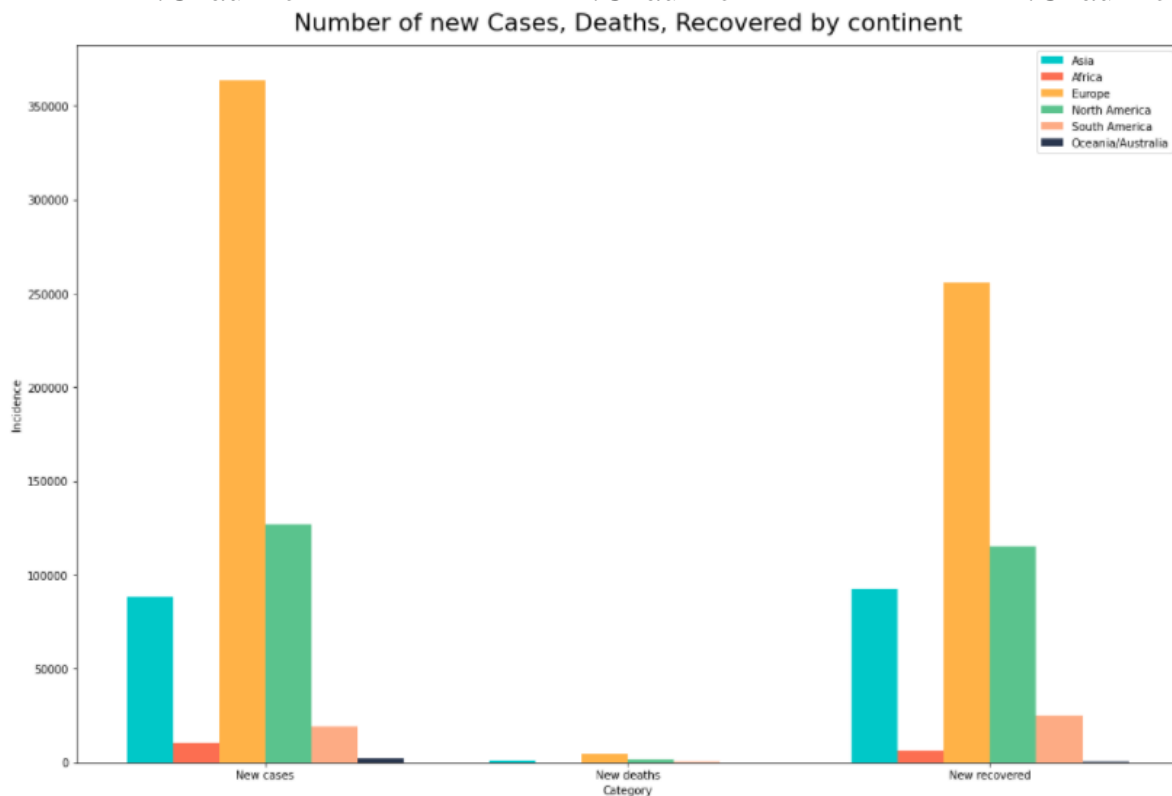
- Riêng ở ca tử vong, Châu Âu đứng đầu về số lượng thay vì Châu Á.
- Nguyên nhân Châu Á đứng đầu về ca nhiễm bệnh có thể nguyên nhân do đây là châu lục có số dân đông nhất. (Mỗi quan hệ số ca nhiễm, dân số được phân tích ở trên).
- Số ca nhiễm:
 1. Châu Á
 2. Châu Âu
 3. Bắc Mỹ
 4. Nam Mỹ
 5. Châu Phi
 6. Châu Úc
- Số ca phục hồi:
 1. Châu Á
 2. Châu Âu
 3. Bắc Mỹ
 4. Nam Mỹ
 5. Châu Phi
 6. Châu Úc
- Số ca tử vong:
 1. Châu Âu
 2. Châu Á
 3. Bắc Mỹ
 4. Nam Mỹ
 5. Châu Phi
 6. Châu Úc



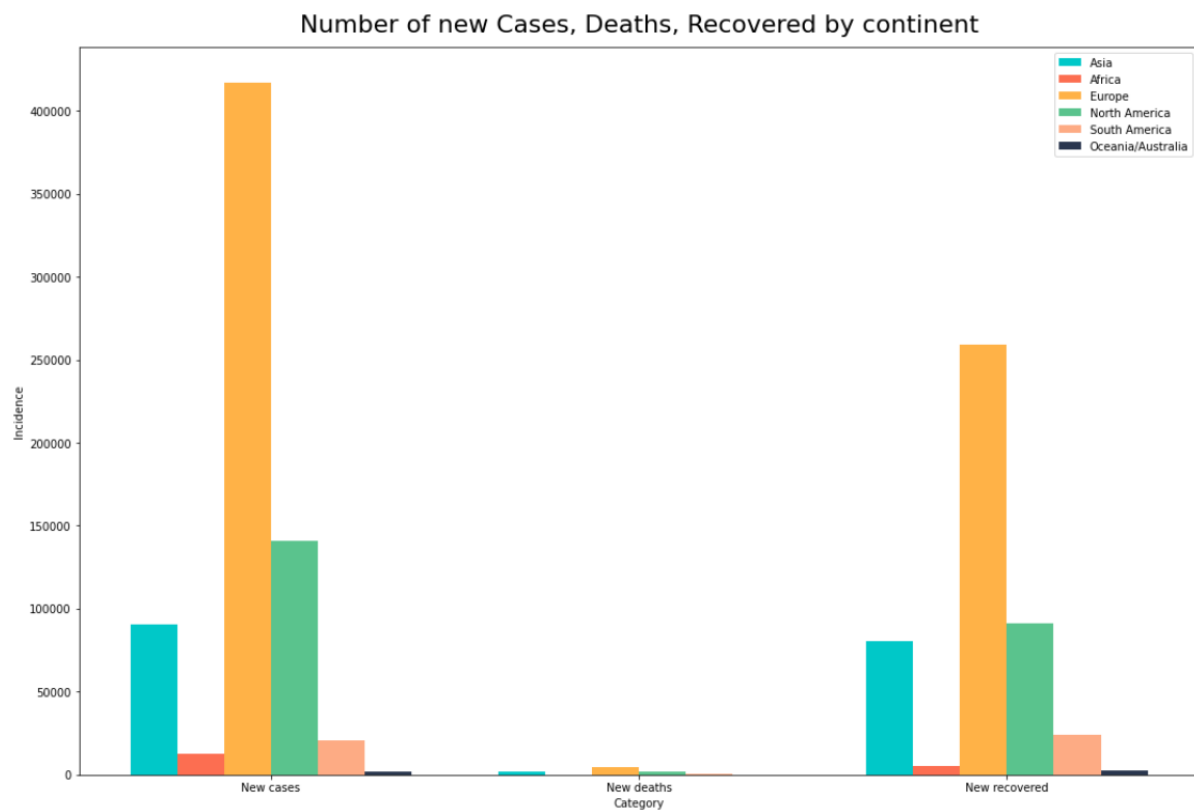
14. Number of New Cases, Deaths, Recovered by continent:

- Biểu đồ thể hiện số ca nhiễm, tử vong, phục hồi mới trong ngày 01-12-2021 theo từng Châu lục.
- Ta thấy có sự giống nhau về thứ tự ở số lượng ca nhiễm, ca phục hồi và ca tử vong.

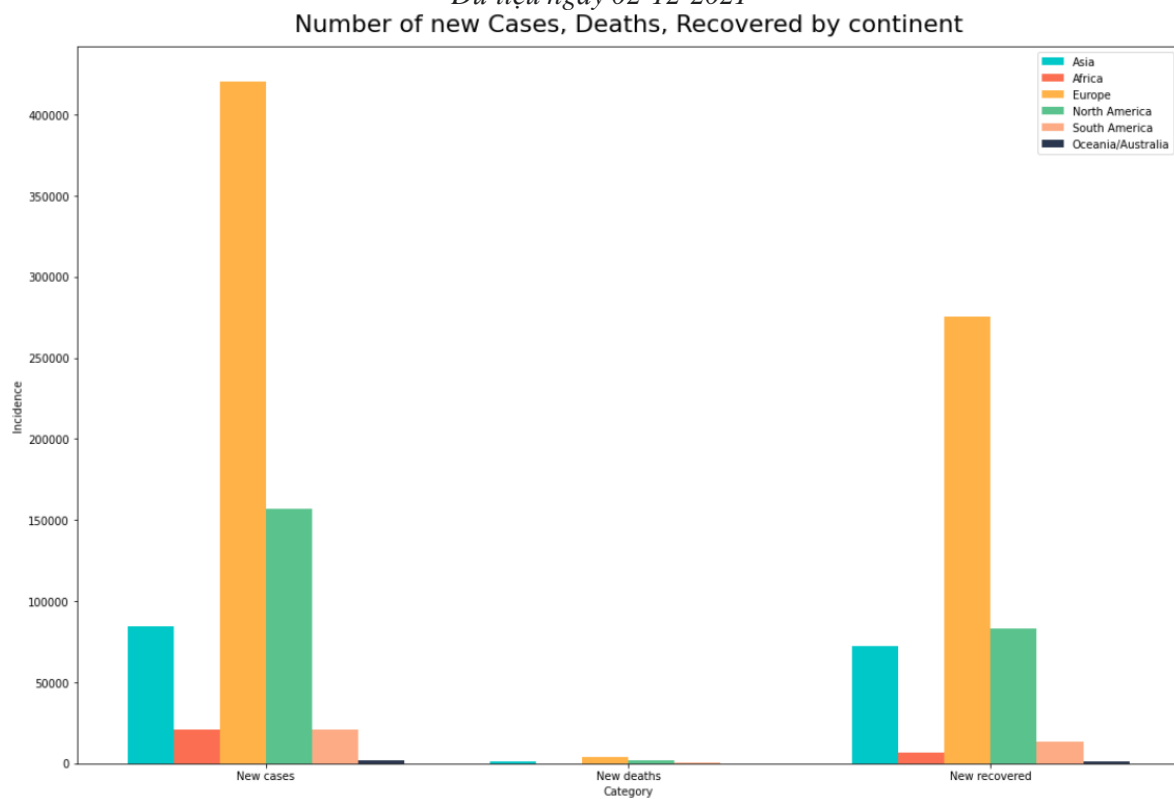
- Các quốc gia Châu Âu tăng mạnh về số ca nhiễm, ca phục hồi và ca tử vong. Nguyên nhân do các quốc gia này đang dần nới lỏng các lệnh hạn chế. Hơn nữa, có thể đang vào mùa Đông và lễ Giáng Sinh nhiều hoạt động vui chơi diễn ra bên trong nhà, dẫn đến dễ lây lan dịch bệnh hơn.
- Số ca nhiễm:
 - 7. Châu Á
 - 8. Châu Âu
 - 9. Bắc Mỹ
 - 10. Nam Mỹ
 - 11. Châu Phi
 - 12. Châu Úc
- Số ca phục hồi:
 - 7. Châu Á
 - 8. Châu Âu
 - 9. Bắc Mỹ
 - 10. Nam Mỹ
 - 11. Châu Phi
 - 12. Châu Úc
- Số ca tử vong:
 - 7. Châu Âu
 - 8. Châu Á
 - 9. Bắc Mỹ
 - 10. Nam Mỹ
 - 11. Châu Phi
 - 12. Châu Úc



Dữ liệu ngày 01-12-2021



Dữ liệu ngày 02-12-2021



Dữ liệu ngày 03-12-2021

15. Case by Continent:

- Biểu đồ thể hiện mối tương quan giữa Total deaths, Recovered, Active cases ở các châu lục.

$$Total\ cases = Active\ cases + Total\ Recovered + Total\ Deaths$$

- Tuy dẫn đầu về số ca nhiễm nhưng nhìn sơ ta có thể thấy Châu Á có số ca nhiễm hiện tại rất thấp.
- Trong khi đó Châu Âu và đặc biệt là Bắc Mỹ có số ca nhiễm hiện tại cao đáng kể, mặc dù nơi đây tập trung nhiều quốc gia thuộc thế giới thứ nhất.

