TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG TP.HCM

BÁO CÁO LAB 1:
MỐI QUAN HỆ CỦA DỮ LIỆU

Môn học: Trực quan hóa dữ liệu

Thông tin nhóm **3**

Mức độ hoàn thành **3**

Mức độ hoàn thành tổng thể các yêu cầu 3

Mức độ hoàn thành của các thành viên 4

Chi tiết **4**

Thu thập dữ liệu 4

Tiền xử lý dữ liệu sau khi thu thập 6

Trực quan hóa và phân tích 6

Trực quan trên một trường đơn 6

Các trường liên quan đến tổng 6

Các trường liên quan đến tỷ lệ 8

Pie chart 10

Trực quan và phân tích mối quan hệ giữa nhiều trường dữ liệu 11

Mối quan hệ tương quan 11

Xét sự phân bố các điểm dữ liệu khi bắt cặp các cột thuộc tính với nhau 13

Xét một số mối quan hệ nhân quả với đường hồi quy tuyến tính 14

Giữa tỉ lệ test trên tổng dân số và tỉ lệ tử vong trên tổng số ca mắc. 14

Giữa tỉ lệ mắc bệnh và tỉ lệ tử vong 14

So sánh tỉ lệ số ca hiện tại giữa các châu lục 15

So sánh tỉ lệ tử vong giữa các châu lục với nhau và giữa các nước trong châu lục với nhau 17

Trực quan và phân tích dựa trên vị trí địa lý 18

Với trường liên quan đến tổng 18

Với trường liên quan đến tỉ lệ 20

1. Thông tin nhóm

Danh sách thành viên:

- 19120068 - Dương Nam Hải
- 19120267 - Hoàng Dược Lam
- 19120298 - Mai Duy Nam

2. Mức độ hoàn thành

2.1. Mức độ hoàn thành tổng thể các yêu cầu

STT	Yêu cầu	Mức độ hoàn thành
1	Thu thập dữ liệu	100%
2	Trực quan hóa các mối quan hệ	100%

2.2. Mức độ hoàn thành của các thành viên

STT	Công việc	Người phụ trách	Mức độ hoàn thành
1	Viết code thu thập dữ liệu, quản lý thùng chứa github	Dương Nam Hải	100%
2	Xem xét một số quan hệ nhân quả, xem xét tình trạng dịch bệnh hiện thời và tỉ lệ tử vong giữa các châu lục.		100%
3	Xem xét, trực quan hóa nhiều thuộc tính	Hoàng Dược Lam	100%
4	Tìm kiếm, trực quan hóa dựa trên thuộc tính địa lý		100%
5	Tổng hợp các file notebook thành một file tổng		100%
6	Thực hiện tiền xử lý dữ liệu	Mai Duy Nam	100%
7	Xem xét, trực quan hóa trên các thuộc tính đơn		100%
8	Xem xét, trực quan hóa cơ cấu dịch bệnh của một số quốc gia		100%
9	Thực hiện báo cáo	Cả nhóm	100%

3. Chi tiết

3.1. Thu thập dữ liệu

Dữ liệu được thu thập từ html của trang https://www.worldometers.info/coronavirus/.

Các bước thực hiện:

- Sử dụng thư viện requests và bs4 (BeautifulSoup) để lấy dữ liệu html của trang web

```
URL = "https://www.worldometers.info/coronavirus/"
r = requests.get(URL)

soup = BeautifulSoup(r.text, features="lxml") # If
```

- Bảng dữ liệu nằm trong thẻ <tbody>. Sau khi lấy được thẻ này gồm các thẻ <t> khác nhau, ta sẽ split các thẻ <tr> ra và được 236 item, với 8 item đầu là của dòng total.

```
table_body = soup.tbody #Find the tab
data = table_body.findAll('tr')[8:]
```

- Trong từng thẻ <tr> vừa tìm được, có các thẻ <td> khác nhau chứa từng thuộc tính và item thứ nhất là index. Ta nạp tất cả vào trong một mảng 2 chiều.

```
dataframe = []
for i, row in enumerate(data):
    try:
        information = row.findAll('td')[1:]
        item = []
        for j in information:
            item.append(j.text)

        dataframe.append(item)
    except:
        continue

dataframe = np.array(dataframe)
```

- Biến dữ liệu thành kiểu DataFrame, thêm thuộc tính “Date” để phục vụ cho các lab sau và lưu với mode = “append” vào file “covid.csv”

```
df = pd.DataFrame(data=dataframe, columns=columns)
date = datetime.today().strftime('%Y-%m-%d')
temp = pd.read_csv('covid.csv')

if date not in temp['Date'].unique():

    df['Date'] = date
    print(df)
    df.to_csv('covid.csv', mode='a', header=False)
```

3.2. Tiền xử lý dữ liệu sau khi thu thập

Các bước thực hiện:

- Chuyển đổi kiểu dữ liệu cho các cột có kiểu dữ liệu chưa phù hợp (Các cột numeric và datetime mang kiểu dữ liệu str)
- Loại những dòng thiếu từ 40% số thuộc tính trở lên
- Loại bỏ các cột 'New Recovered', 'New Cases', 'New Cases/1M pop', 'New Deaths', 'New Deaths/1M pop' vì tỉ lệ thiếu quá cao.
- Loại bỏ các cột '1 Death every X ppl', '1 Case every X ppl', '1 Test every X ppl' vì hoàn toàn có thể được tính từ những cột sẵn có.
- Loại bỏ khoảng trắng ở đầu và cuối mỗi chuỗi trong cột 'Country' (Xuất hiện khi thực hiện crawl html)
- Lưu lại thành file mới (covid_preprocessed.csv)

3.3. Trực quan hóa và phân tích

Tiền xử lý trước khi phân tích:

- Thêm các cột thuộc tính "Death rate", "Recovery rate", "Active rate", "Serious rate" với ý nghĩa là các tỉ lệ tử vong, phục hồi, đang mắc bệnh, ca nghiêm trọng trên tổng số trường hợp mắc, và "Cases per test" là tỉ lệ số ca tìm thấy trên tổng số ca test.
- Thêm các cột "Country Code" và "Continent Code" bằng các dữ liệu json api từ các nguồn trên internet. Các cột này sẽ phục vụ cho việc visualize trên world map.

3.3.1. Trực quan trên một trường đơn

Đối với các mối quan hệ trên các trường đơn, nhóm chia ra làm hai nhóm: các trường liên quan đến tổng các ca (ví dụ Total Cases, Total Deaths, Total Recovered, v.v.) và các trường liên quan đến tỷ lệ (ví dụ Recovery Rate, Death Rate, Active Rate, v.v.). Nhóm sử dụng biểu đồ histogram cho *các trường liên quan đến tổng* và box plot cho *các trường liên quan đến tỷ lệ*.

3.3.1.1. Các trường liên quan đến tổng

Có năm trường liên quan đến tổng: Total Cases, Total Tests, Total Recovered, Total Deaths và Active Cases. Nhóm gom hai trường Total Cases và Total Tests vẽ chung trên một biểu đồ do hai trường này có range tương tự nhau, ba trường còn lại vẽ trên một biểu đồ khác.
Về trực quan hai trường Total Cases và Total Tests:

- Để vẽ được biểu đồ này, nhóm tận dụng lớp FacetGrid của seaborn. Sử dụng phương thức melt() của pandas để đưa dữ liệu từ dạng wide-form về dạng long-form.

	Total Cases	Total Tests
0	83170407	1.006270e+09
1	43085166	8.386282e+08
2	30460997	6.377617e+07
3	28690748	2.698169e+08
4	24791190	1.223324e+08
...	...	...
212	2788	5.138200e+04
213	2701	2.344600e+04
215	454	2.050800e+04
216	273	1.013500e+04
227	217836	1.600000e+08

melt()

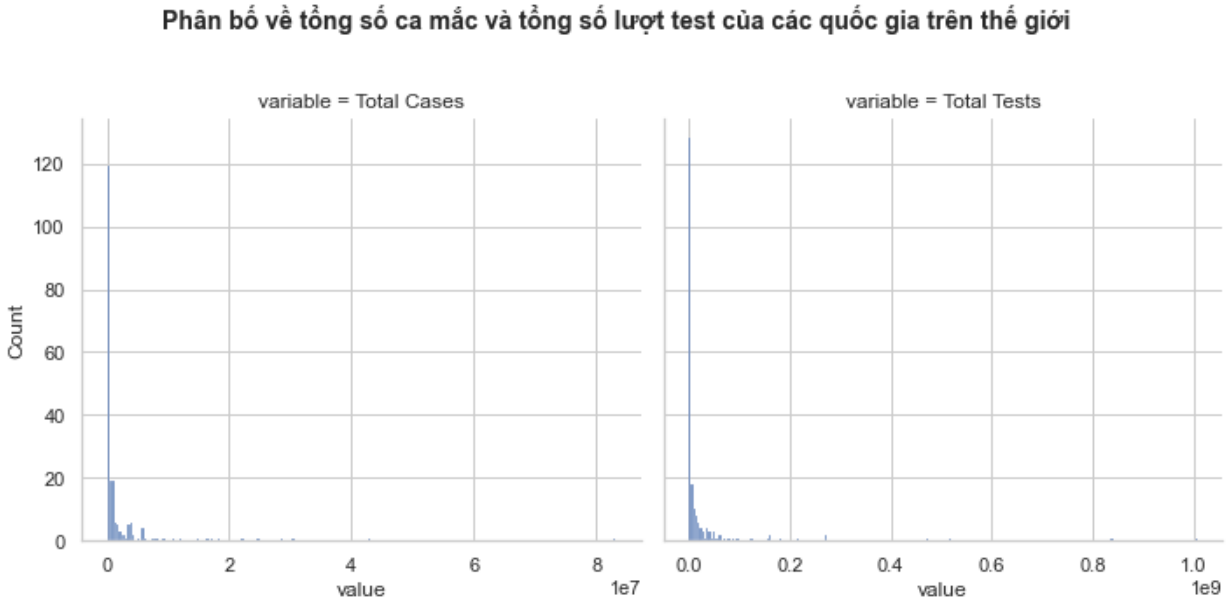
	variable	value
0	Total Cases	83170407.0
1	Total Cases	43085166.0
2	Total Cases	30460997.0
3	Total Cases	28690748.0
4	Total Cases	24791190.0
...	...	...
417	Total Tests	51382.0
418	Total Tests	23446.0
419	Total Tests	20508.0
420	Total Tests	10135.0
421	Total Tests	160000000.0

- Filter trên cột variable bằng tham số col, sau đó map mỗi nhóm bằng hàm histplot để vẽ histogram

```
f = sns.FacetGrid(
    # melt để đưa dữ dữ liệu từ dạng wide-form về dạng long-form
    data=df[['Total Cases', 'Total Tests']].melt(),
    col='variable',
    sharex=False,
    height=5,
)

f.map_dataframe(sns.histplot, x='value')
f.fig.subplots_adjust(top=0.8)
f.fig.suptitle(
    t='Phân bố về tổng số ca mắc và tổng số lượt test của các quốc gia trên thế giới',
    fontsize=14,
    fontweight='bold'
)
plt.show()
```

- Kết quả trực quan thu được:



Về trực quan hóa ba trường Total Recovered, Total Deaths và Active Cases, ta cũng thực hiện tương tự như trên. Kết quả trực quan thu được:

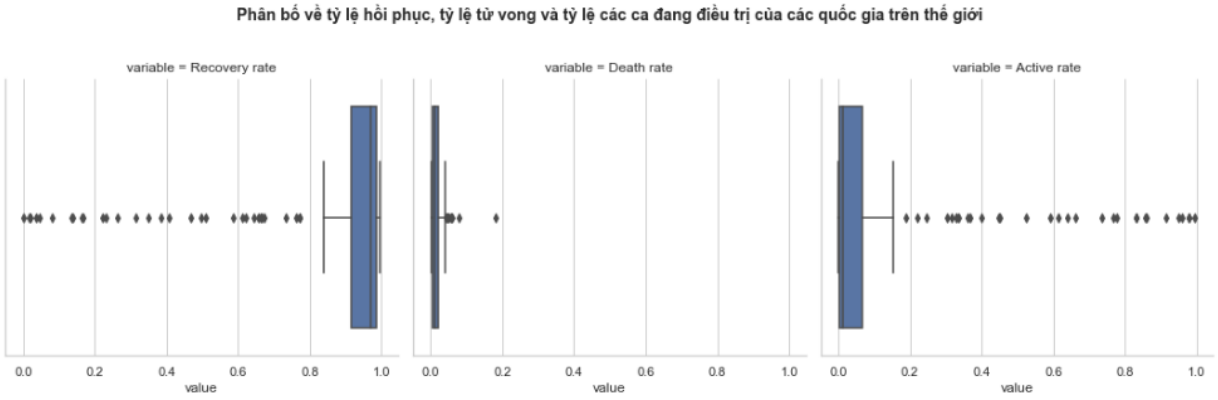
Nhận xét chung: cả năm thuộc tính đều có phân phối bị lệch dương nhiều, các giá trị của thuộc tính tập trung chủ yếu ở các giá trị nhỏ.

3.3.1.2. Các trường liên quan đến tỷ lệ

Trong các trường liên quan đến tỷ lệ:

- Gom các trường Recovery rate, Death rate và Active rate lại để trực quan hóa chung với nhau do các trường này có cùng khoảng giá trị (từ 0 đến 1). Gom ba trường Total Cases/1M pop, Tests/1M pop và Deaths/1M pop để trực quan hóa chung vì các trường này có cùng đơn vị.
- Recovery, Death và Active là ba thành phần đóng góp vào tổng số ca mắc của các quốc gia. Do đó, Recovery rate, Death rate và Active rate là tỷ lệ phần trăm của ba thành phần này.

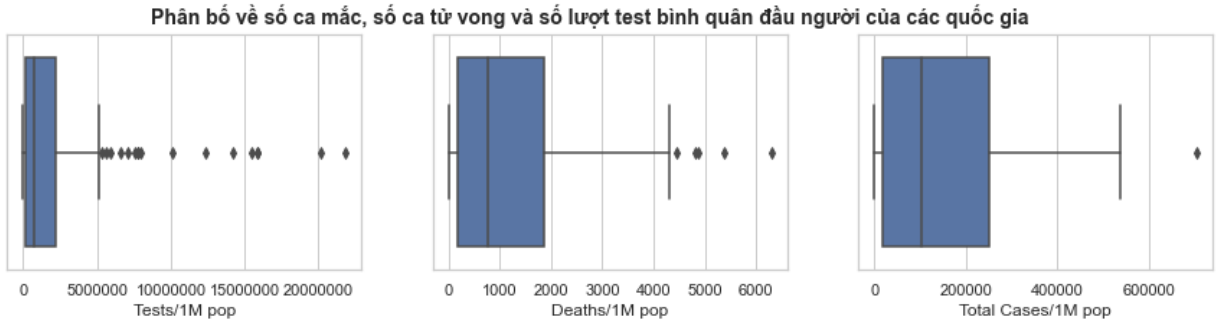
Về các trường Recovery rate, Death rate, Active rate:



- Tương tự với các biểu đồ histogram ở trên, ta cũng sử dụng lớp FacetGrid của seaborn, kết hợp với hàm melt để trực quan
- Nhận xét chung:
    - Thuộc tính Recovered rate bị lệch âm, các giá trị quy tụ chủ yếu ở quanh mức 90-100% chứng tỏ phần lớn các ca nhiễm ở các quốc gia đều đã được điều trị khỏi. Tuy nhiên phân bố của thuộc tính này cũng có nhiều ngoại lai. Ta sẽ khảo sát những ngoại lai này bằng biểu đồ pie chart ở phần dưới.
    - Thuộc tính Death rate và Active rate bị lệch âm. Điều này cũng dễ hiểu do Recovered rate bị lệch dương và tổng ba thuộc tính này ở từng quốc gia là 100%. Do phần lớn các ca mắc đều đã được điều trị khỏi do đó số ca mắc hiện còn điều trị ở mức thấp.

Về các trường còn lại:

- Thay vì dùng FacetGrid, ta tạo 3 subplot và vẽ box plot của từng trường lên 3 subplot này
- Kết quả trực quan:



- Nhận xét chung:
    - Phân bố của tỷ lệ ca mắc và ca tử trong trên 1 triệu dân bị lệch dương và tập trung chủ yếu ở mức gần 0.

- Số lượt test trên 1 triệu dân cũng có phân phối tương tự, nhưng có nhiều ngoại lai hơn và đặc biệt có một vài nước có số lượt test là hơn 20 triệu/1 triệu dân, tức một người dân trung bình được test trên 20 lần kể từ khi bắt đầu thống kê đại dịch.

3.3.1.3. Pie chart

Như đã chỉ ra ở trên bằng box plot, hầu hết các quốc gia đều có Recovery rate cao, Death rate và Active rate thấp. Tuy nhiên, trên thực tế vẫn có nhiều ngoại lai với các quốc gia có Recovery rate thấp hoặc rất thấp. Để so sánh cơ cấu dịch bệnh của các quốc gia này, nhóm chọn ra 12 quốc gia để trực quan hóa, trong đó:

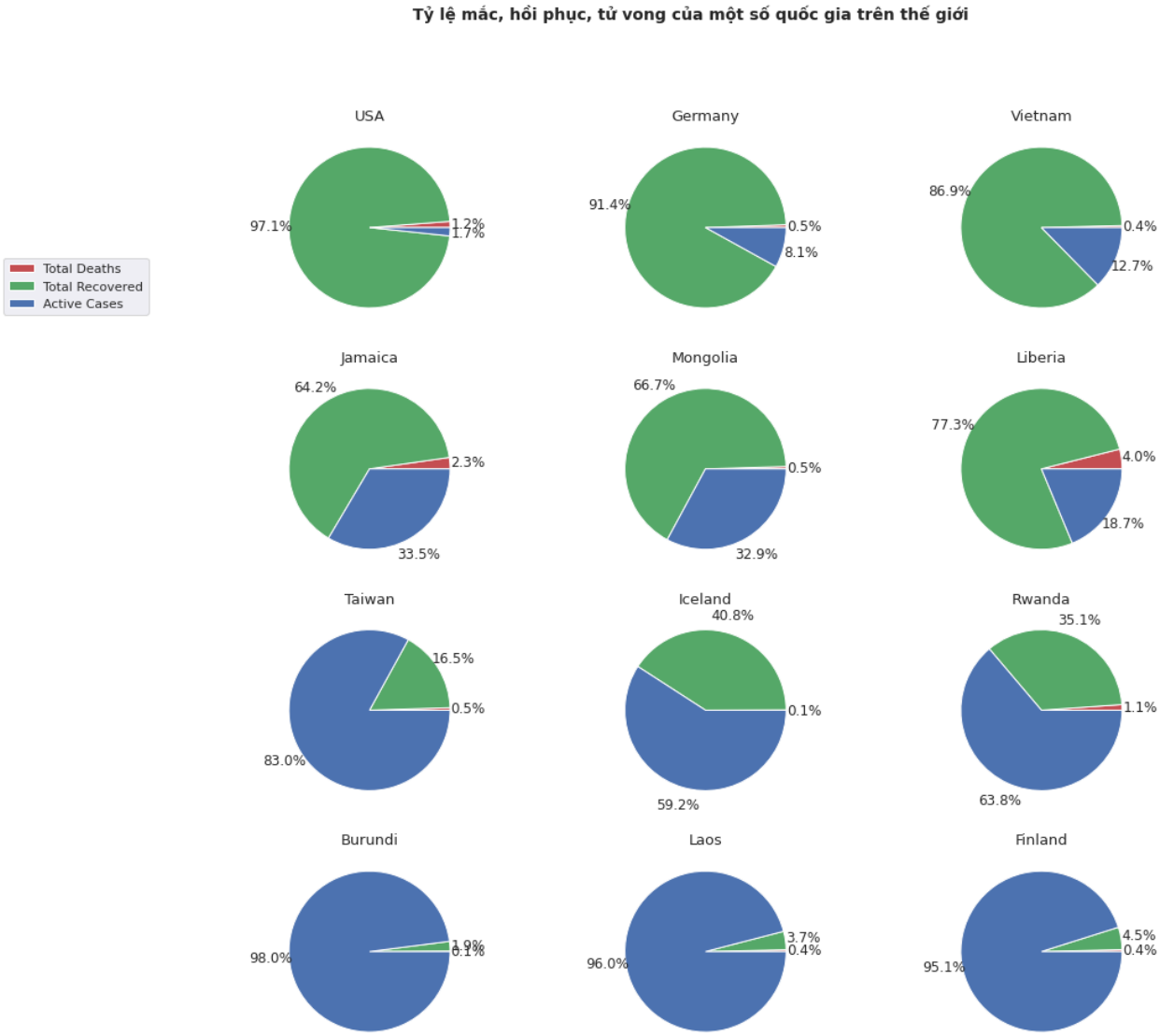
- 3 nước đại diện cho nhóm “bình thường" là Mỹ, Đức và Việt Nam
- 9 nước còn lại chọn ra từ danh sách các nước có Recovery rate nhỏ hơn 0.8%

```
In [11]: low_recovery_rate = new_df[new_df['Recovery rate'] < 0.8].sort_values(by='Recovery rate')
         low_recovery_rate
```

Out[11]:

	Country	Total Cases	Total Deaths	Total Recovered	Active Cases	Serious, Critical	Total Cases/1M pop	Deaths/1M pop	Total Tests	Tests/1M pop	Population	Continent	Active Cases/1M pop	D
117	Martinique	153253	926.0	104.0	152223.0	8.0	408966.0	2471.0	828928.0	2212050.0	374733.0	North America	406217.0	20 05
120	Guadeloupe	140130	854.0	2250.0	137026.0	19.0	350107.0	2134.0	938039.0	2343639.0	400249.0	North America	342352.0	20 05
152	Burundi	39998	38.0	773.0	39187.0	NaN	3190.0	3.0	345742.0	27575.0	12538337.0	Africa	3125.0	20 05
107	Laos	207867	745.0	7660.0	199462.0	NaN	27832.0	100.0	1232128.0	164972.0	7468707.0	Asia	26706.0	20 05

Để trực quan hóa, ta tạo một subplot có 12 ô với 4 dòng và 3 cột. Lặp trên mỗi quốc gia tương ứng với từng ô, ta trích ra giá trị của thuộc tính Recovery rate, Death rate và Active rate của quốc gia đó và trực quan bằng pie chart. Kết quả trực quan thu được:



3.3.2. Trực quan và phân tích mối quan hệ giữa nhiều trường dữ liệu

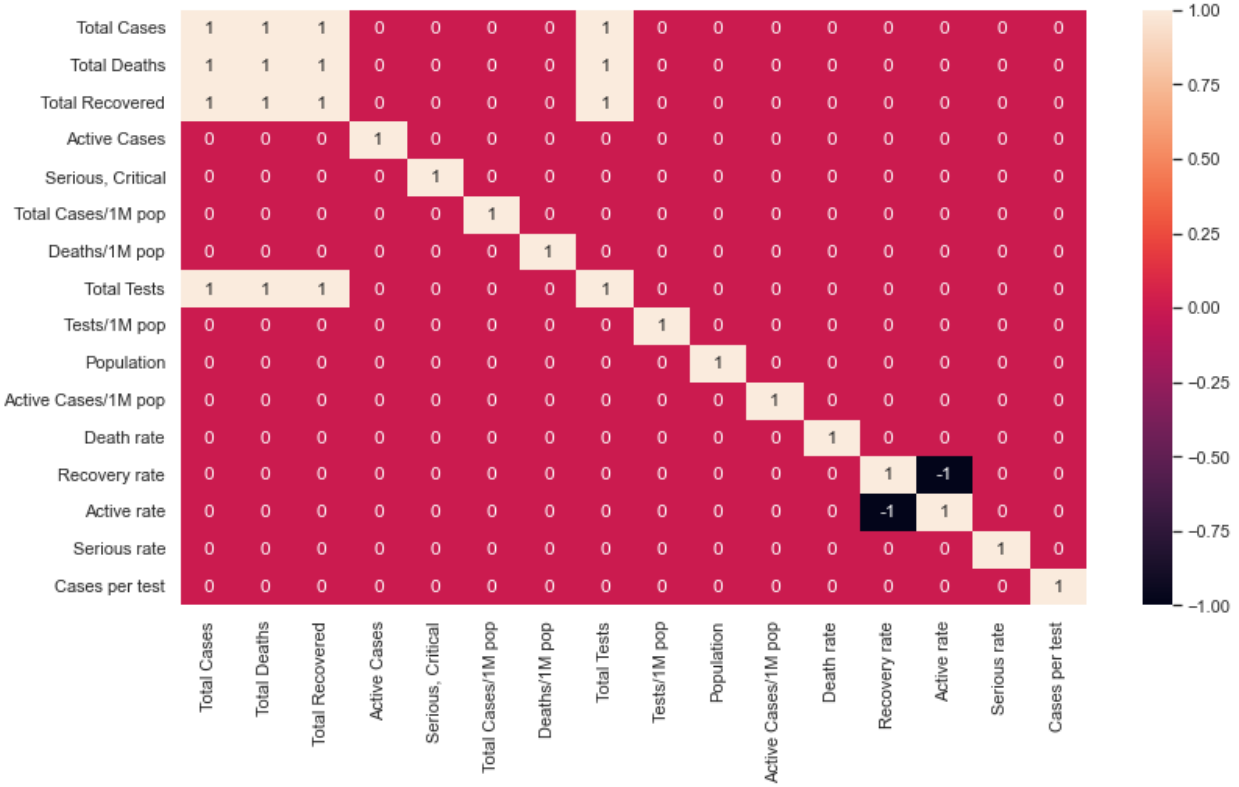
3.3.2.1. Mối quan hệ tương quan

Ta xây dựng correlation matrix và vẽ heatmap để trực quan mối quan hệ giữa các cặp thuộc tính numeric.

Do cần hạn chế sử dụng màu sắc để trực quan, ta sẽ chỉ xét những cặp có độ tương quan lớn 0.7 là tương quan thuận cao, bé hơn -0.7 là tương quan nghịch cao, còn lại là không tương quan.

```
def set_value_for_corr(x):
    if x > 0.7:
        return 1
    elif x < -0.7:
        return -1
    else:
        return 0
corr = df.corr()
corr = corr.applymap(set_value_for_corr)
plt.figure(figsize=(13, 7))
sns.heatmap(corr, annot=True)
```

Kết quả trực quan:

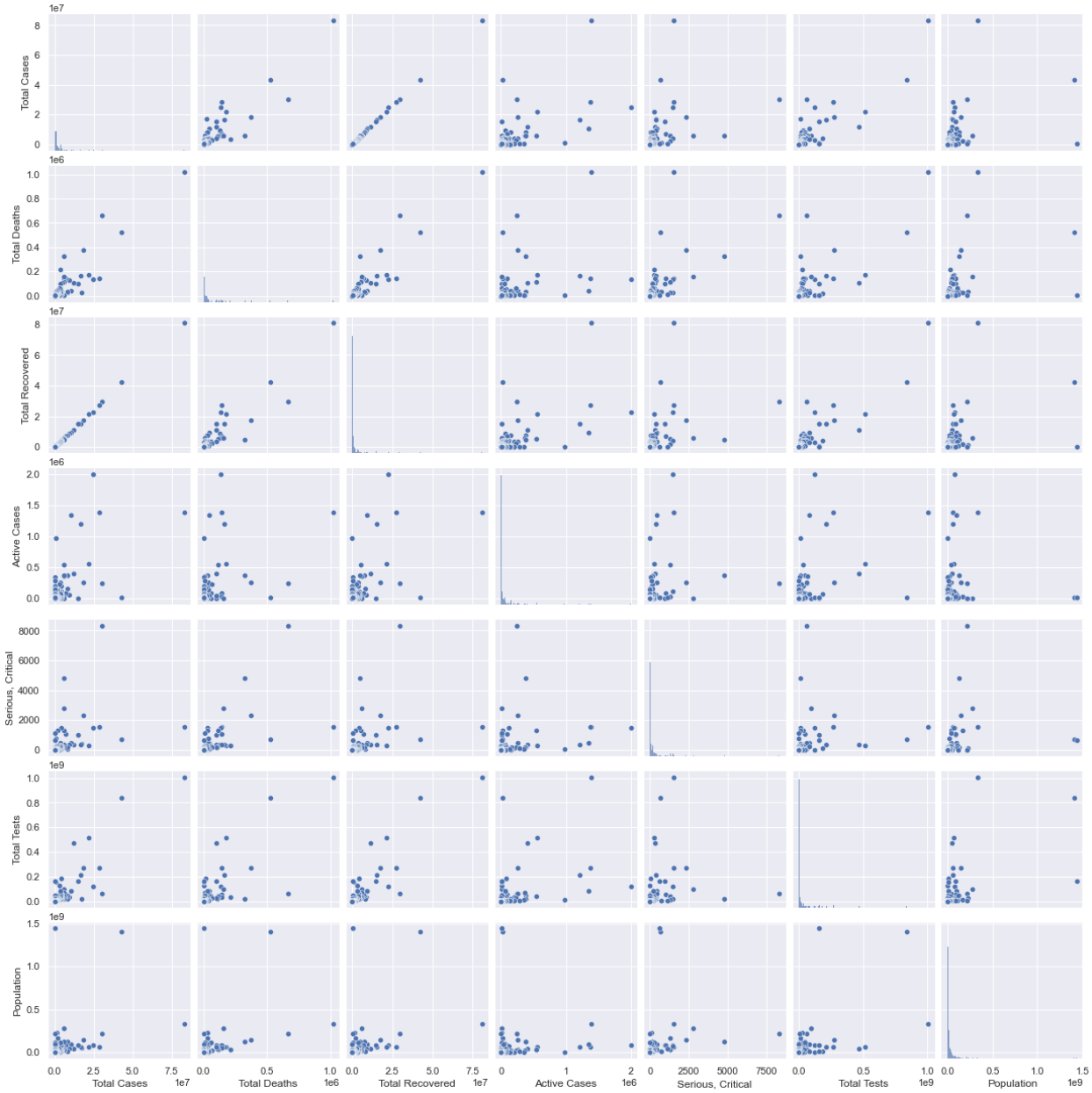


Ta có mối quan hệ tính toán sau: Total Cases = Active Cases + Total Deaths + Total Recovered. Vì vậy nếu Death rate ổn định, tỉ lệ Active rate và Recovery rate sẽ có mức tương quan nghịch cao.

- *<u>Lý do không sử dụng covariance matrix</u>: Do giá trị của các cột dữ liệu chênh lệch quá lớn, có cột lên đến hàng chục triệu, có cột giá trị chỉ khoảng từ 0 đến 1. Covariance matrix sẽ cho ra giá trị trực quan không tốt.*

3.3.2.2. Xét sự phân bố các điểm dữ liệu khi bắt cặp các cột thuộc tính với nhau

Ta sẽ thực hiện pairplot cho các cột dữ liệu mà không thể nội suy từ các cột khác. Các cột sẽ không thực hiện pair plot là những cột có thể tính được như “Death rate”, “Active rate”, “Deaths/1m pop”,...



Có thể thấy các thuộc tính của chúng ta bị lệch trái khá nhiều nên dữ liệu bị gom cụm ngay tại trục tọa độ. Vì vậy khi xem xét cụ thể một cặp thuộc tính, ta phải áp dụng một số biện pháp scaling để có thể trực quan chúng tốt hơn.

3.3.2.3. Xét một số mối quan hệ nhân quả với đường hồi quy tuyến tính

3.3.2.3.1. Giữa tỉ lệ test trên tổng dân số và tỉ lệ tử vong trên tổng số ca mắc.

- Cột "Test/1M pop" là tỉ lệ test được thực hiện trên 1 triệu dân, tính bằng việc lấy cột "Total Tests" chia cho "Population".
- Cột "Death rate" là tỉ lệ tử vong trên tổng số ca mắc, được tính bằng việc lấy cột "Total Deaths" chia cho "Total Cases".

Như vậy khi trực quan mối quan hệ giữa 2 cột "Test/1M pop" và "Death rate", ta sẽ xem được mối quan hệ của cả 4 trường dữ liệu thành phần.

Qua việc scatterplot các cặp dữ liệu đã thực hiện trước đó ta cũng biết các trường dữ liệu bị lệch trái, vì vậy phương pháp tốt nhất để giúp cho dễ trực quan hơn là scale với hàm log.

```
In [11]: scatter_with_regression_plot(np.log(df['Tests/1M pop']), np.log(df['Death rate']))
```

- **Nhận xét:** Đường hồi quy dựa trên các điểm dữ liệu có chiều hướng đi lùi, như vậy khi tỉ lệ test càng cao thì tỉ lệ tử vong càng giảm. Điều này thể hiện rằng khi một quốc gia tích cực thực hiện test để phát hiện ngăn chặn dịch bệnh, quốc gia đó sẽ giảm thiểu được số ca tử vong vì bệnh Covid.

3.3.2.3.2. Giữa tỉ lệ mắc bệnh và tỉ lệ tử vong

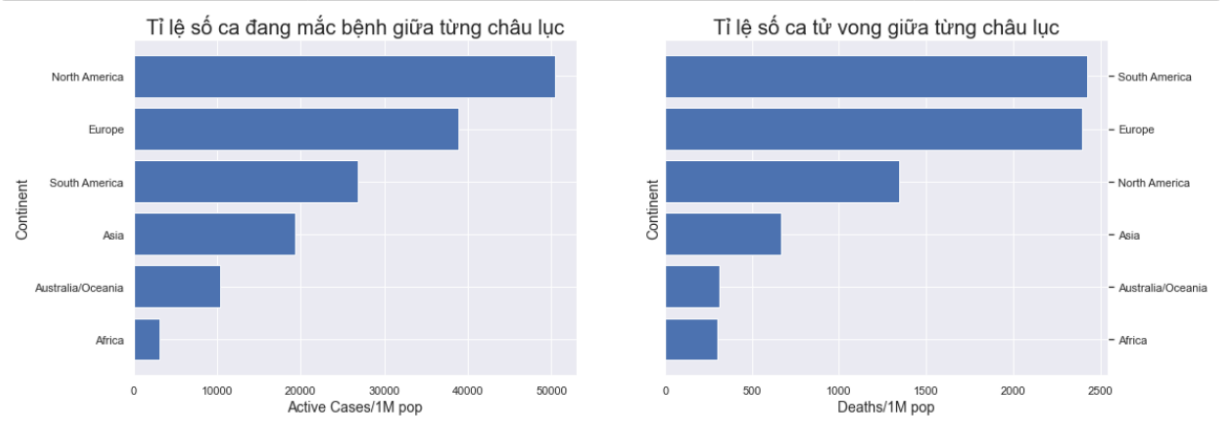
Ta tiến hành trực quan trên 2 cột "Total Cases/1M pop" và "Deaths/1M pop". Ta không chọn 2 cột Total vì giữa các quốc gia có sự chênh lệch dân số với nhau, việc phân tích trên các cột tỉ lệ sẽ cho kết quả chính xác hơn.

```
In [12]: scatter_with_regression_plot(np.log(df['Total Cases/1M pop']), np.log(df['Deaths/1M pop']))
```

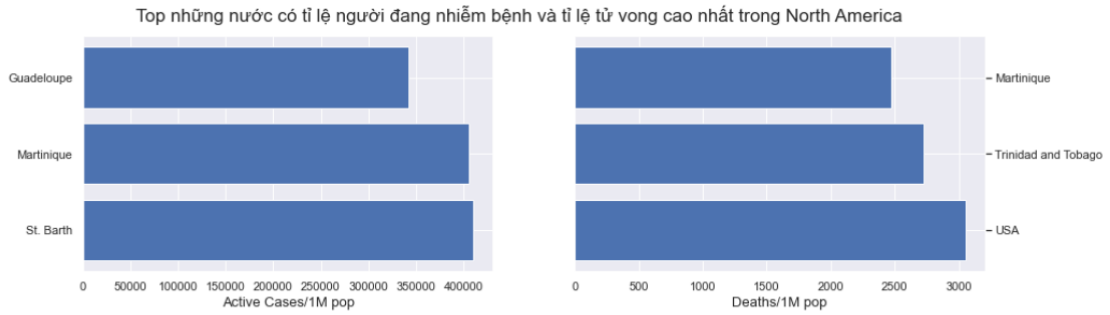
- **Nhận xét:** Đường hồi quy có xu hướng đi lên, chứng tỏ khi tỉ lệ mắc bệnh trong một quốc gia càng tăng thì tỉ lệ tử vong vì Covid cũng tăng theo. Chính phủ của các quốc gia nên thực hiện phòng chống dịch bệnh để bảo toàn sức khỏe cho người dân của họ.

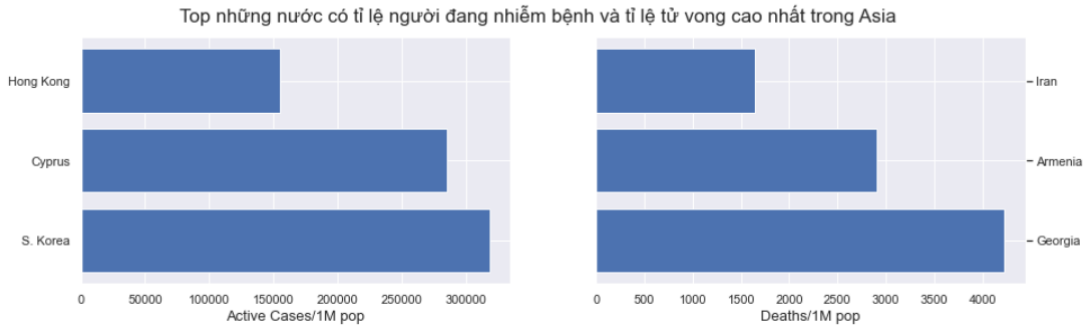
3.3.2.4. So sánh tỉ lệ số ca hiện tại giữa các châu lục

Ta tiến hành nhóm các dòng dữ liệu theo châu lục và tính mean để có các tỉ lệ trung bình theo từng châu lục. Sau đó ta xem xét các châu lục có tỉ lệ số ca đang mắc bệnh và tỉ lệ tử vong giảm dần

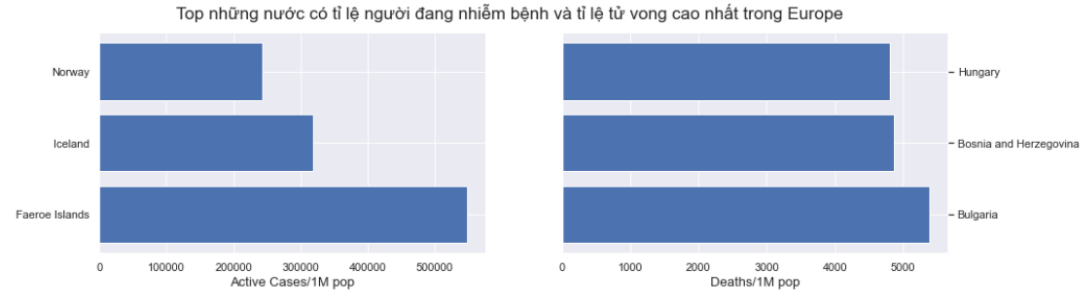


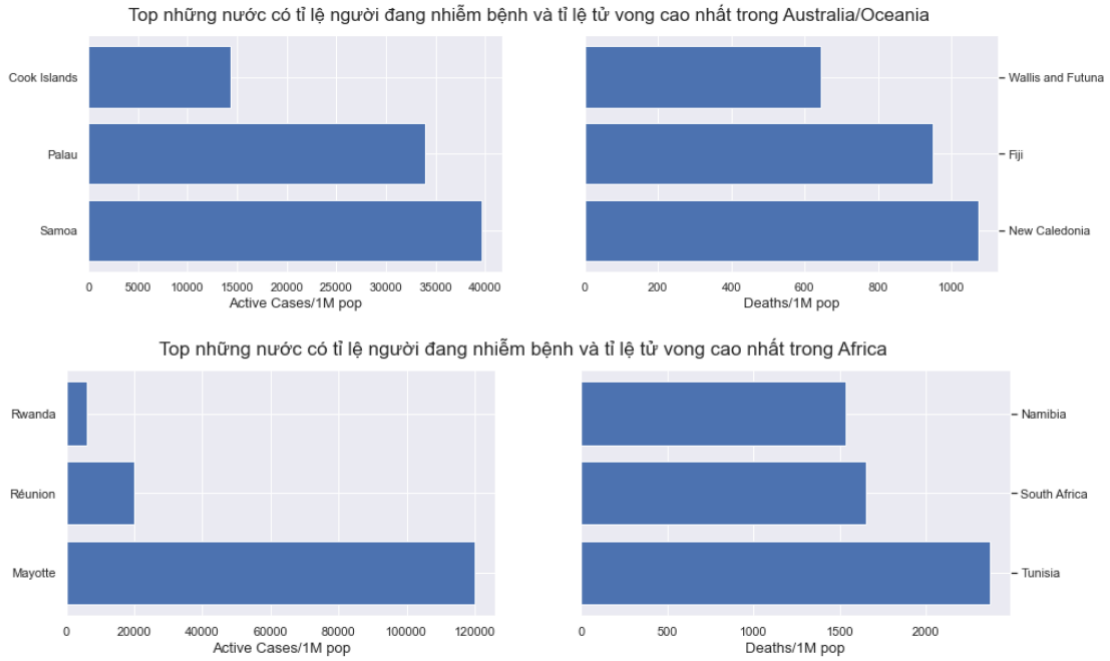
Đi sâu hơn vào từng châu lục, ta xem top 3 nước thuộc châu lục đó có tỉ lệ active case và tỉ lệ tử vong cao nhất


Top những nước có tỉ lệ người đang nhiễm bệnh và tỉ lệ tử vong cao nhất trong North America
Guadeloupe
Martinique
St. Barth
0 50000 100000 150000 200000 250000 300000 350000 400000
Active Cases/1M pop
Martinique
Trinidad and Tobago
USA
0 500 1000 1500 2000 2500 3000
Deaths/1M pop


Top những nước có tỉ lệ người đang nhiễm bệnh và tỉ lệ tử vong cao nhất trong Asia
Hong Kong
Cyprus
S. Korea
0 50000 100000 150000 200000 250000 300000
Active Cases/1M pop
Iran
Armenia
Georgia
0 500 1000 1500 2000 2500 3000 3500 4000
Deaths/1M pop

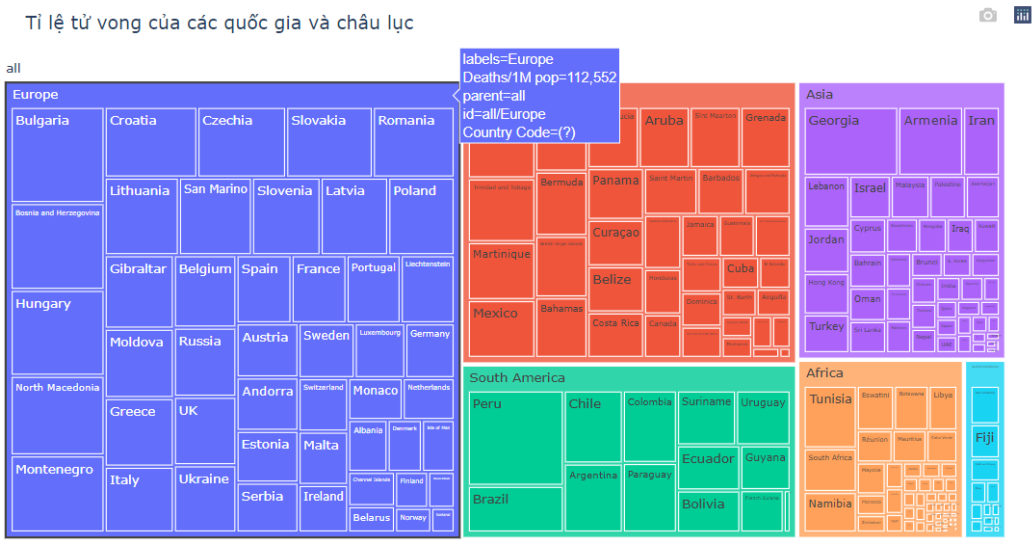

Top những nước có tỉ lệ người đang nhiễm bệnh và tỉ lệ tử vong cao nhất trong South America
Suriname
Peru
French Guiana
0 50000 100000 150000 200000
Active Cases/1M pop
Chile
Brazil
Peru
0 1000 2000 3000 4000 5000 6000
Deaths/1M pop


Top những nước có tỉ lệ người đang nhiễm bệnh và tỉ lệ tử vong cao nhất trong Europe
Norway
Iceland
Faeroe Islands
0 100000 200000 300000 400000 500000
Active Cases/1M pop
Hungary
Bosnia and Herzegovina
Bulgaria
0 1000 2000 3000 4000 5000
Deaths/1M pop



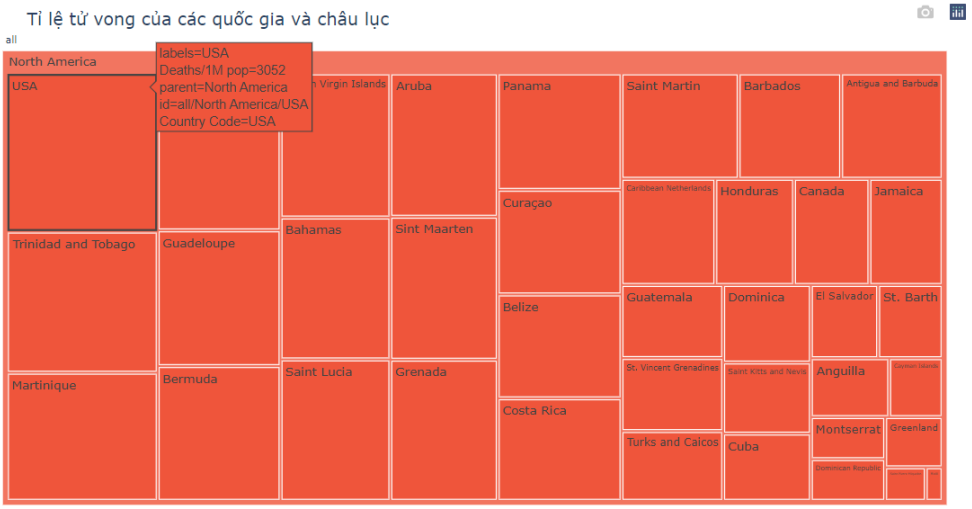
3.3.2.5. So sánh tỉ lệ tử vong giữa các châu lục với nhau và giữa các nước trong châu lục với nhau

Sau khi đã có Bar Chart thể hiện top những nước có tỉ lệ người đang nhiễm bệnh và tỉ lệ tử vong cao nhất trong từng khu vực, nhóm đã nảy ra ý tưởng gom nhóm các quốc gia trong châu lục lại với nhau thành một nhóm, và gom các nhóm trên thành một nhóm lớn hơn. Từ ý tưởng trên, nhóm quyết định sử dụng Treemap cho trường ***Deaths/1M pop***



Biểu đồ thể hiện tổng quan cho dữ liệu của trường ***Deaths/1M pop*** giữa các châu lục với nhau. Biểu đồ cũng có thể hiện kèm số liệu tổng ***Deaths/1M pop*** của các quốc gia trong châu lục.

Khi ấn vào mỗi châu lục, biểu đồ sẽ thể hiện so sánh tổng quan của trường ***Deaths/1M pop*** giữa các nước trong châu lục với nhau. Khi tương tác với các quốc gia, ta cũng có thể thấy số liệu của mỗi quốc gia



3.3.3. Trực quan và phân tích dựa trên vị trí địa lý

Do dữ liệu được nhóm sử dụng khá phù hợp để trực quan dựa trên vị trí địa lý, do có bao gồm các quốc gia, và châu lục, từ đó có thể thêm vào mã quốc gia và mã châu lục, giúp cho ta có thể trực quan dữ liệu của các quốc gia trên bản đồ địa lý thế giới.

Nhóm sử dụng Bubble Map để trực quan một số trường dữ liệu đơn, có kết hợp trực quan vị trí địa lý của các dòng (các quốc gia) trên bản đồ thế giới, có thể thấy rõ được tình hình tổng thể của các quốc gia trên thế giới với nhau, cũng như điều chỉnh góc nhìn hợp lý cho bản thân để có cái nhìn tổng thể về nhóm các quốc gia trong cùng 1 khu vực, 1 châu lục.

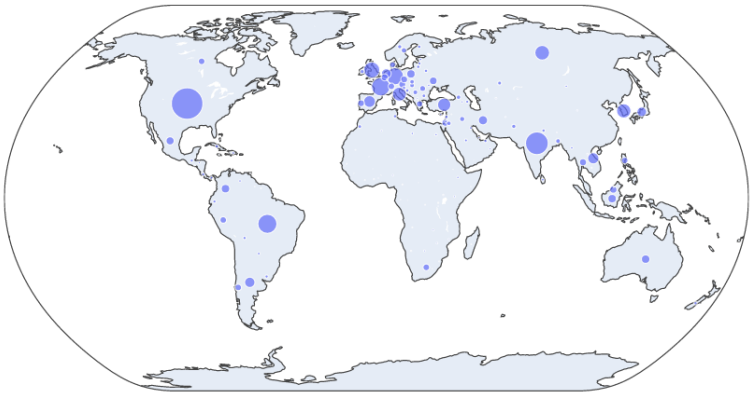
Do đã phân tích từ trước, trường dữ liệu đơn được chia thành 2 nhóm: các trường liên quan đến tổng và các trường liên quan đến tỷ lệ. Với mỗi nhóm, ta sẽ trực quan đại diện 1 trường trong nhóm đó, cụ thể như sau:

- Với trường liên quan đến tổng, nhóm sẽ trực quan trường ***Total Cases***.
- Với trường liên quan đến tỷ lệ, nhóm sẽ trực quan trường ***Death Rate***.

3.3.3.1. Với trường liên quan đến tổng

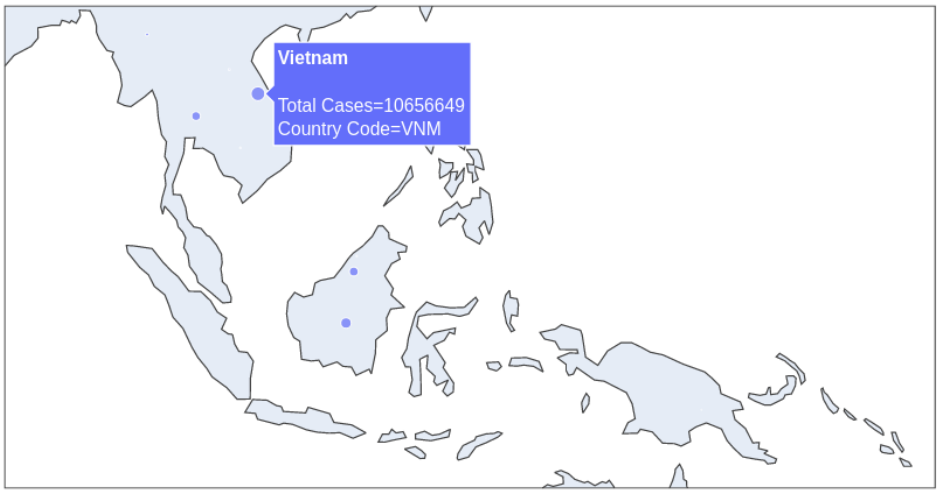
Sử dụng hàm *scatter_geo* của thư viện plotly.express, với location được xác định bằng Country Code, và tên của mỗi bubble, tức hover_name là tên mỗi quốc gia

Tổng số ca mắc Covid của mỗi quốc gia trên toàn thế giới



Ta có được một bản đồ thế giới với các bong bóng có kích cỡ khác nhau, thể hiện sự khác nhau về số lượng tổng số ca mắc covid của mỗi quốc gia trên toàn thế giới. Ta cũng có thể zoom in và zoom out, sao cho có góc nhìn hợp lý mà bản thân mong muốn trên bản đồ. Ví dụ dưới đây ta có thể zoom in để so sánh số liệu giữa các nước Đông Nam Á với nhau:

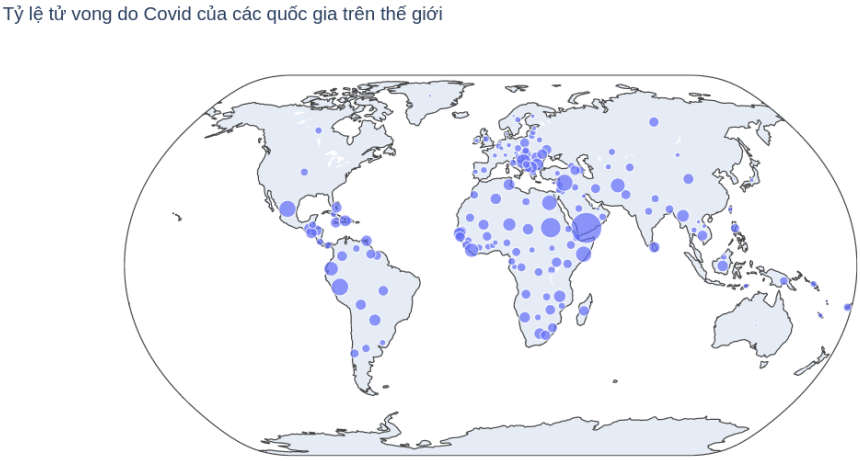
Khi di chuyển con trỏ chuột tới các bubble trên bản đồ, ta cũng có thể có được thêm một số thông tin về quốc gia mà bubble đó thể hiện: tên quốc gia, tổng số ca mắc covid cụ thể, mã quốc gia.



Nhận xét: Ta thấy được rằng, Châu Âu có các nhiều quốc gia có số lượng mắc Covid khá lớn. Một số quốc gia đông dân như Mỹ, Ấn Độ, Nga, Brazil cũng có số lượng người mắc Covid cao. Tuy nhiên, Trung Quốc với số dân đông nhất thế giới lại có lượng người mắc Covid rất nhỏ (phải zoom in mới có thể thấy được)

3.3.3.2. Với trường liên quan đến tỉ lệ

Tương tự như trên, sử dụng hàm *scatter_geo* của thư viện plotly.express, với location được xác định bằng Country Code, và tên của mỗi bubble, tức hover_name là tên mỗi quốc gia



Ta cũng có thể zoom in vào những khu vực mong muốn trên bản đồ, và tương tác với các bubble thể hiện các thông tin: tên quốc gia, tỷ lệ tử vong cụ thể, mã quốc gia.

Nhận xét: Nhìn chung, ta có thể thấy được các quốc gia trong khu vực Bắc Phi và Nam Trung Mỹ có tỉ lệ người tử vong do mắc bệnh khá cao, có thể là do một số lý do như điều kiện y tế vẫn chưa được cao, kinh tế vẫn còn khó khăn, ...