ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN 1

MỐI QUAN HỆ CỦA DỮ LIỆU

TRỰC QUAN HÓA DỮ LIỆU

Thành viên nhóm:

Nguyễn Phạm Quang Dũng– 19120485

Nguyễn Thị Tiểu Mi - 19120577

Nguyễn Thị Kim Ngân – 19120598

Giảng viên hướng dẫn:

Lê Ngọc Thành

Thành phố Hồ Chí Minh – 2022

MỤC LỤC

1. Phân công

Họ tên	MSSV	Công việc	Hoàn thành
Nguyễn Phạm Quang Dũng	19120485	Crawl dữ liệu, tiền xử lý dữ liệu, trực quan hóa dữ liệu	100%
Nguyễn Thị Tiểu Mi	19120577	Trực quan hóa dữ liệu, báo cáo	100%
Nguyễn Thị Kim Ngân	19120598	Trực quan hóa dữ liệu, báo cáo	100%

2. Tiền xử lý dữ liệu

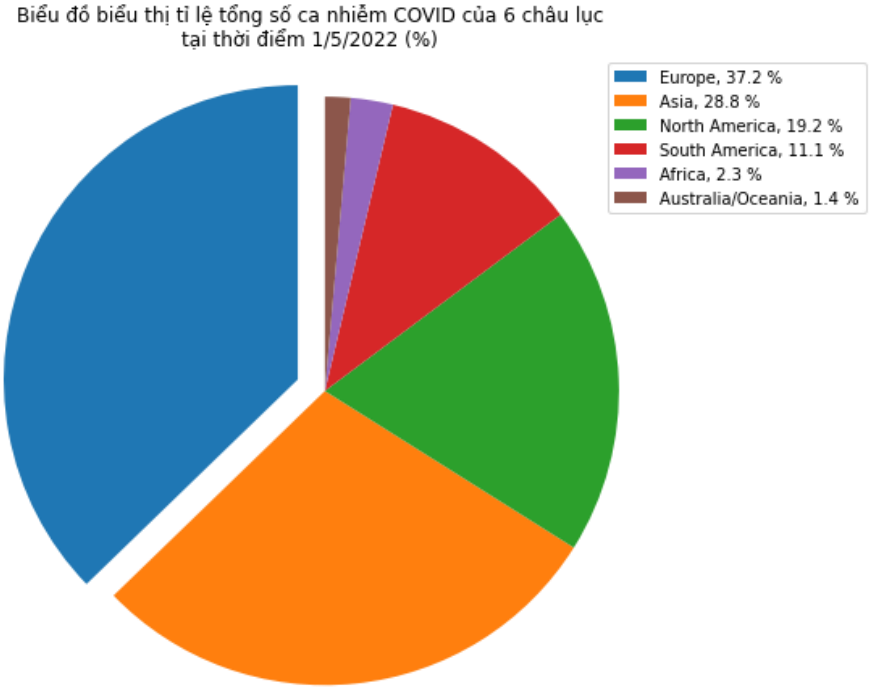
Ta sẽ loại bỏ kí tự "+" và "," ở các ô dữ liệu, đồng thời cũng thay các ô trông(Na/N) thành 0.

Thay các dòng ở cột Continent bị trống thành "The remaining areas".

Tuy nhiên các trường dữ liệu đều đang ở dạng object nên ta sẽ đưa các trường không phải Country,Other và Continent về đúng kiểu dữ liệu thực sự của nó (int, float).

3. Trực quan hóa dữ liệu

3.1. Tỉ lệ ca nhiễm COVID của 6 châu lục tính đến hết ngày 1/5/2022:



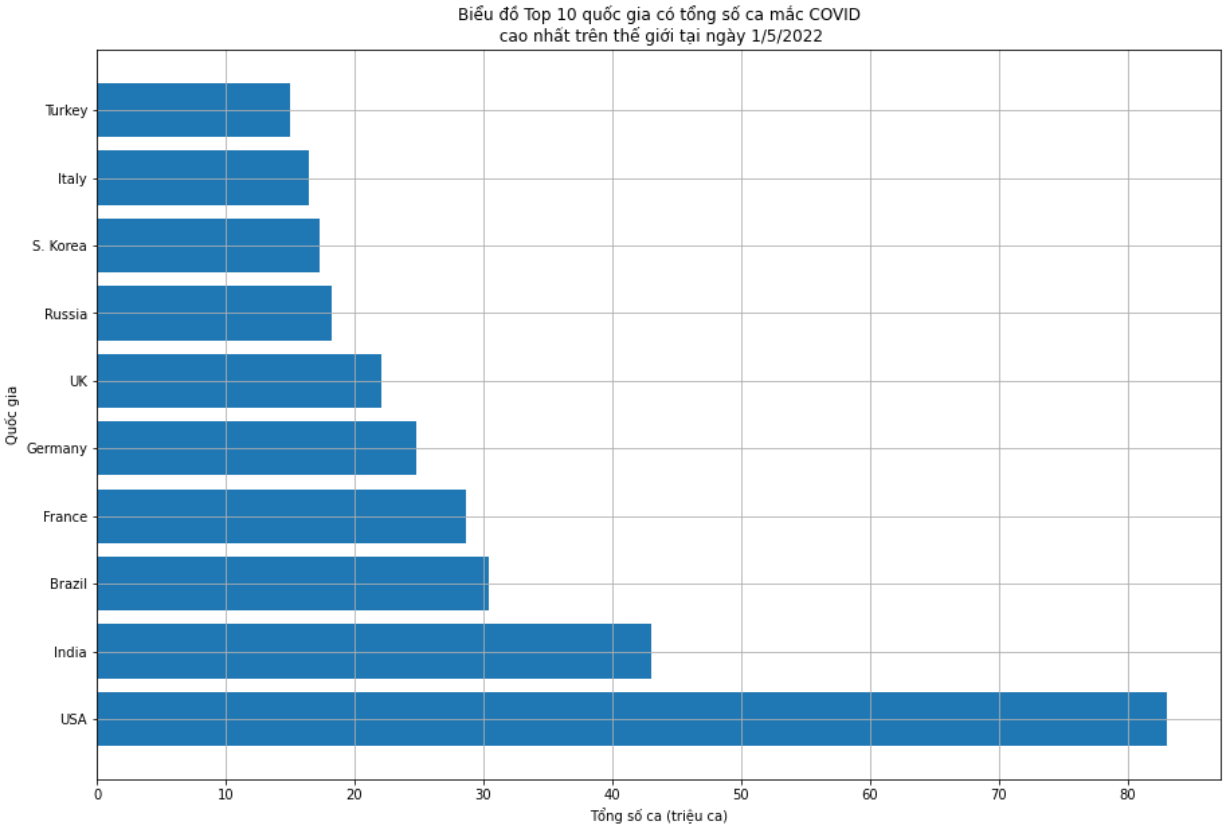
Kiểu biểu đồ: Pie Chart

Giải thích: Việc sử dụng Pie Chart sẽ cho ta thấy rõ hơn về tỉ lệ % về tổng số ca nhiễm giữa các châu lục.

Nhận xét:

- Europe là châu lục đứng đầu với hơn 37% cách khá xa các châu lục còn lại.
- Kế đến là Asia – châu lục đông dân nhất thế giới vẫn đang trong giai đoạn chống dịch.
- Thứ tự ca nhiễm trên pie chart cũng phản ánh rõ thực tế tình trạng dịch bệnh ở thế giới, kết quả cho ra là đúng với kỳ vọng.

3.2. Top 10 quốc gia có số ca nhiễm COVID-19 cao nhất thế giới (tính đến hết ngày 1/5/2022)



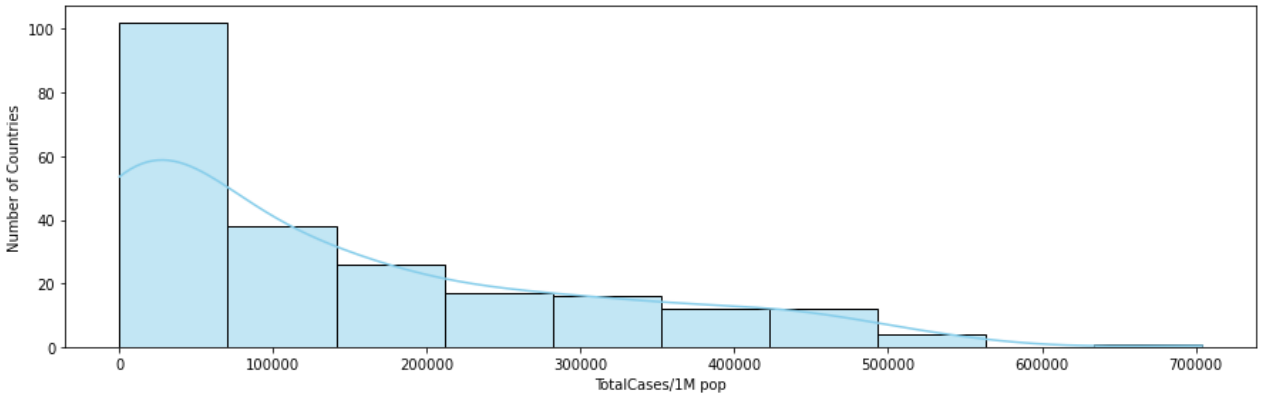
Kiểu biểu đồ: Barh Chart

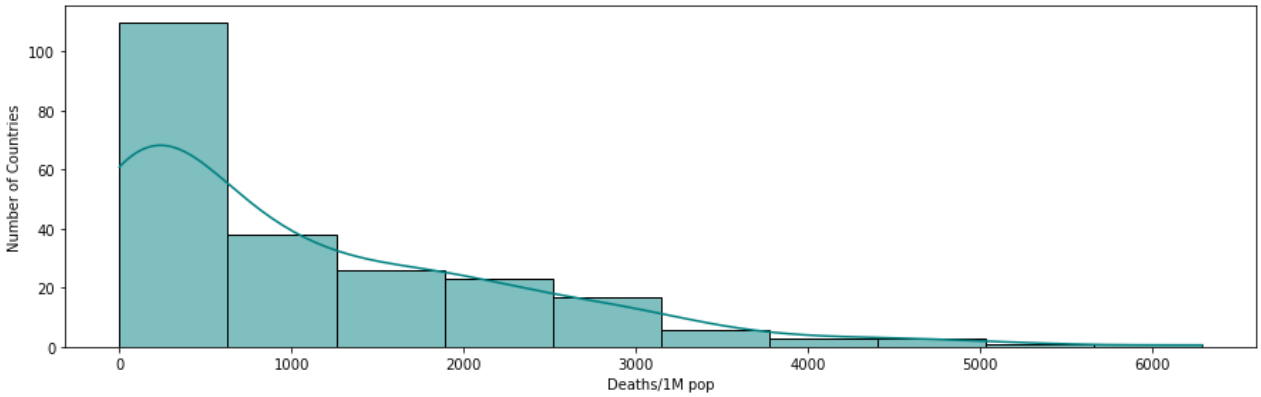
Giải thích: Bar chart phù hợp với việc thống kê số liệu tổng số ca nhiễm của từng nước, tuy nhiên để dễ dàng quan sát ta sẽ sử dụng Barh Chart.

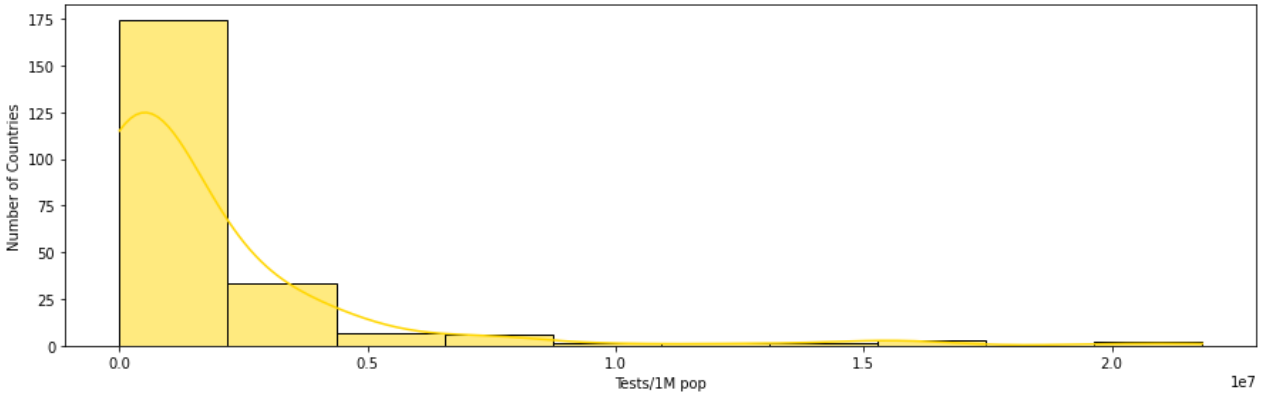
Nhận xét:

- Các quốc gia dẫn đầu về số ca nhiễm đa phần là ở Châu Âu với Châu Mỹ.
- Mỹ với tổng số ca nhiễm ca ngất ngưỡng (gấp đôi Ấn Độ) với số ca lên đến hơn 80 triệu ca.

3.3. Tình hình dịch bệnh của các nước trên thế giới (TotalCases/1M pop, Deaths/1M pop, Tests/1M pop)







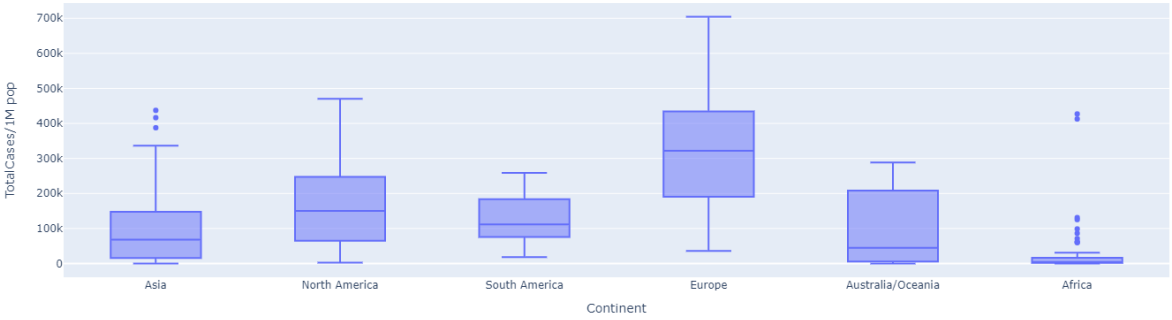
Kiểu biểu đồ: Histogram

Giải thích: Để thể hiện được số quốc gia với phân bố của các giá trị ta sẽ dùng Histogram.

Nhận xét:

- Hầu hết các nước trên thế giới đều có chỉ số ca nhiễm, ca tử vong/1 triệu dân khá thấp. Chỉ một vài nước có ca nhiễm và tử vong cao.
- Số ca test/1 triệu dân cũng nằm ở mức khoảng dưới 2,5 triệu lần test/1 triệu dân ở nhiều quốc gia. Trung bình một người dân sẽ test khoảng 2 lần.

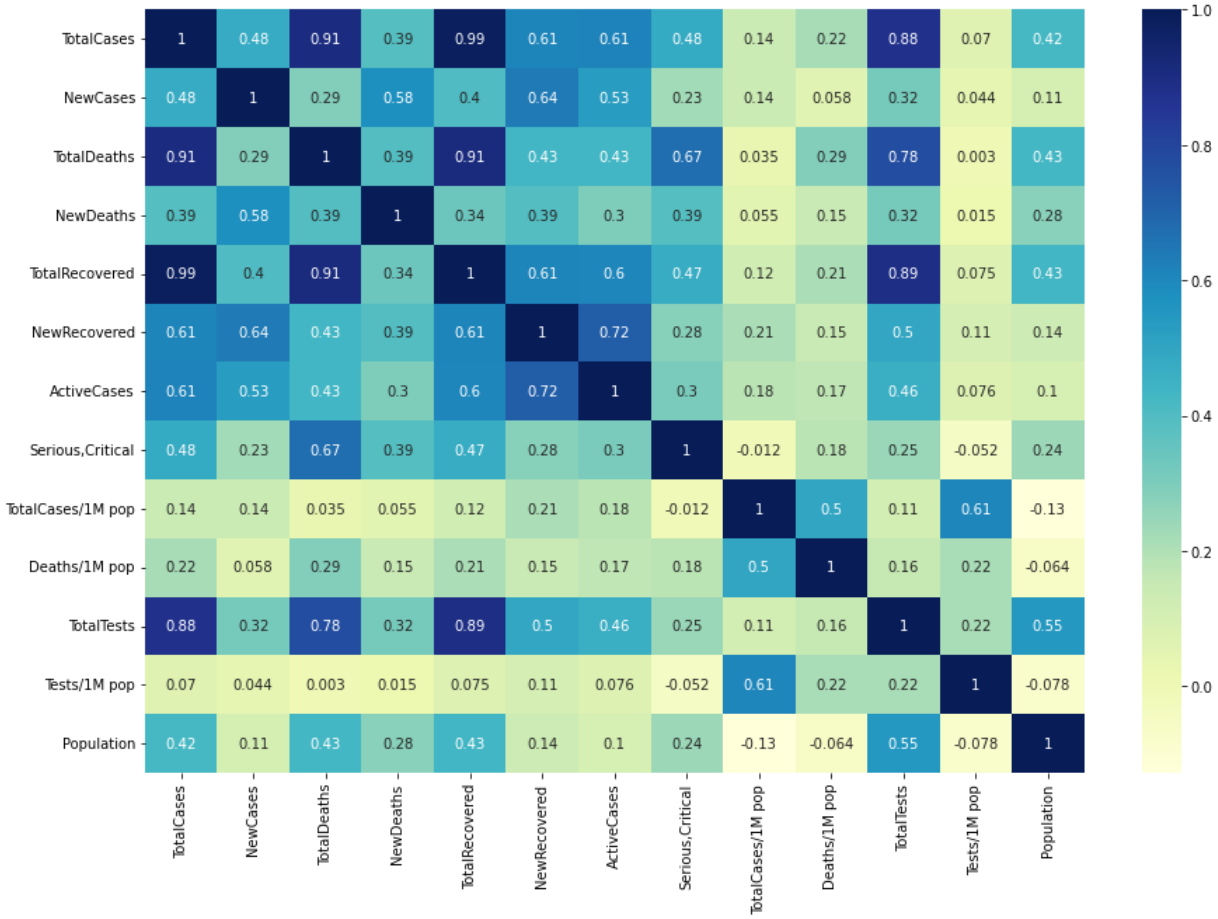
3.4. Với tình hình diễn biến dịch với nhiều chủng virus mới thì tỉ lệ số ca nhiễm ở các châu lục đang diễn biến như thế nào?



Kiểu biểu đồ: Box plot
Giải thích: Chọn Box plot để thể hiện sự phân bố của từng châu lục một cách cụ thể nhất
Nhận xét:

- Dựa theo box plot, ta có thể thấy sự chênh lệch về tỉ lệ số ca nhiễm giữa các châu lục. Châu Phi là châu lục có tỉ lệ ca nhiễm thấp nhất, còn Châu Âu là nơi có tỉ lệ ca nhiễm cao nhất. Tuy nhiên thì vẫn có một vài quốc gia ở Châu Phi có tỉ lệ ca nhiễm cao ngất ngưỡng (~400000 ca nhiễm/ 1 triệu dân).
- Châu Âu là châu lục có tỉ lệ ca nhiễm rất cao, thậm chí còn có nước lên đến 700000 ca nhiễm/1 triệu dân. Mặc dù là châu lục với rất nhiều nước phát triển nhưng có lẽ ý thức và suy nghĩ phòng chống dịch của người dân ở Châu Âu là không cao. Họ thường xem nhẹ việc nhiễm Covid-19. Nhờ vào việc có nền y tế phát triển và việc tiêm phòng cũng nhanh chóng.

3.5. Mối tương quan giữa các trường dữ liệu



Kiểu biểu đồ: Heat map

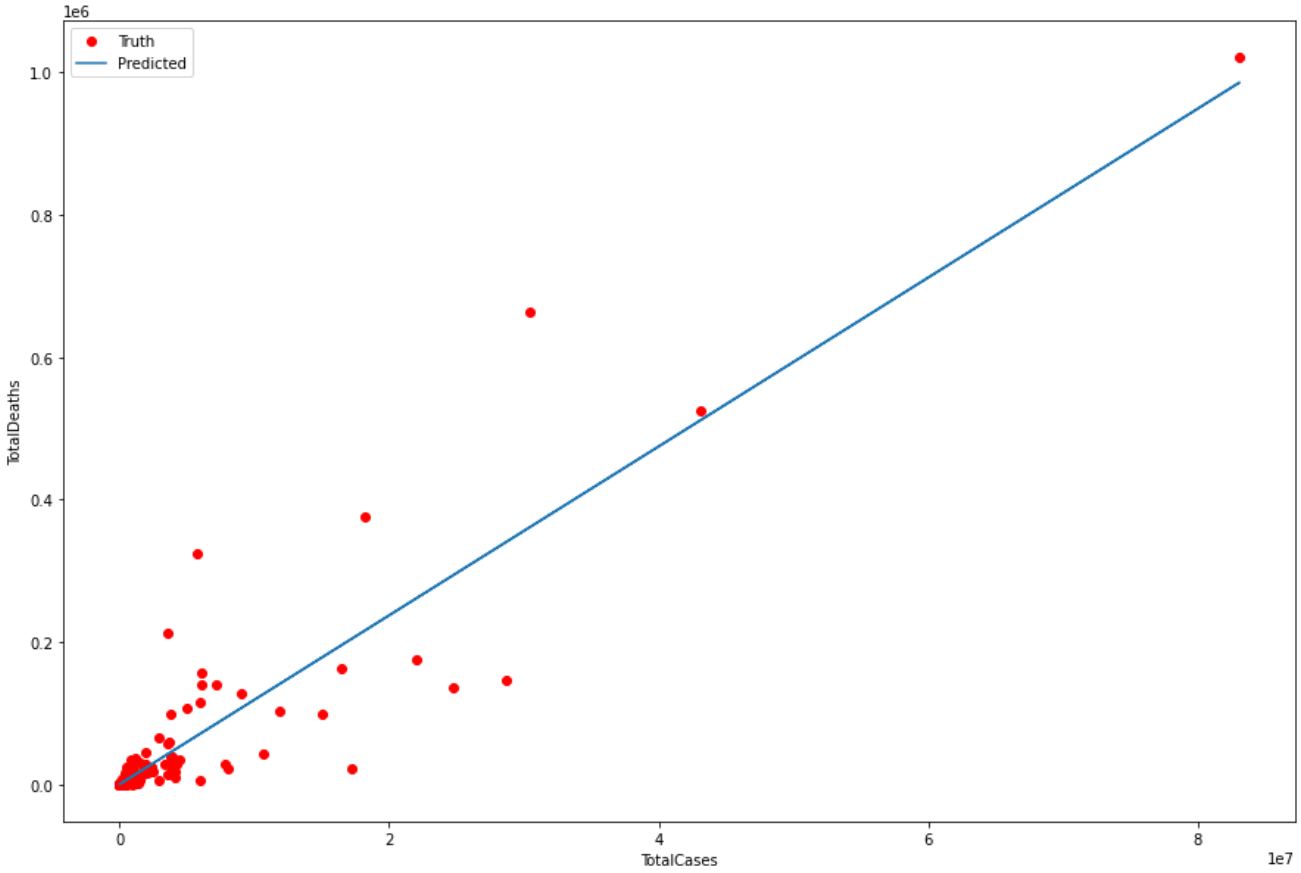
Giải thích: Việc sử dụng Correlation matrix sẽ cho ta thấy được mức độ tương quan giữa các trường dữ liệu.

Nhận xét:

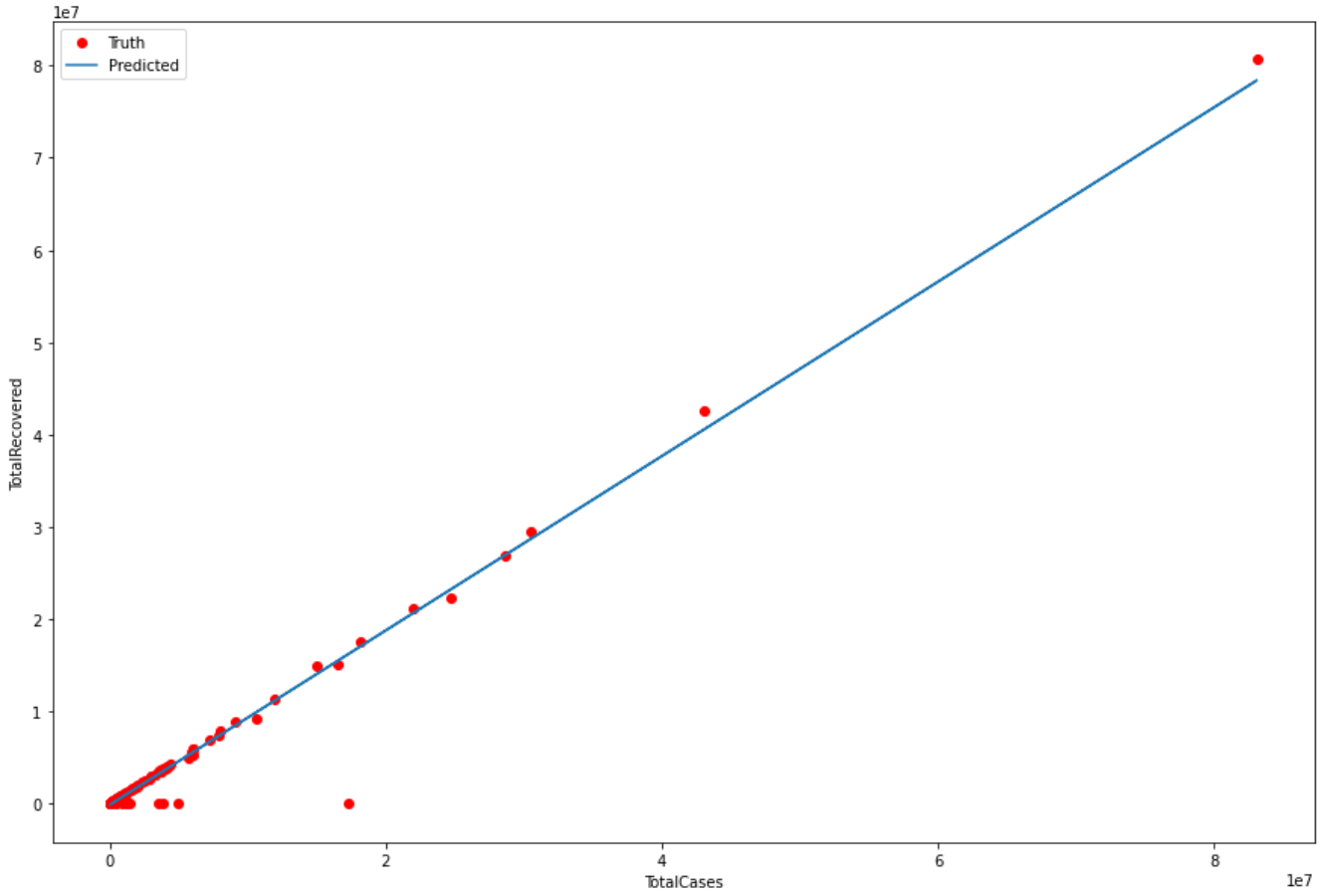
- Chúng ta có thể thấy được giữa một số trường dữ liệu có mối quan hệ tuyến tính với nhau: TotalCases, TotalDeaths, TotalRecovered, NewRecovered, ActiveCases, Serious, Critical, TotalTests
- Ta có thể thấy đây là một căn bệnh khá nguy hiểm, việc nhiễm bệnh nặng có thể dẫn đến việc tử vong xảy ra khá nhiều. Tuy nhiên, với những ca gần đây thì tỉ lệ nhiễm bệnh nặng, nghiêm trọng đã được giảm đi rất nhiều. Cho thấy chất lượng đi lên của y tế và hiệu quả của vacxin.

3.6. Kiểm thử hệ số tương quan giữa các trường dữ liệu TotalCases, TotalDeaths, TotalRecovered, TotalTests bằng hồi quy tuyến tính.

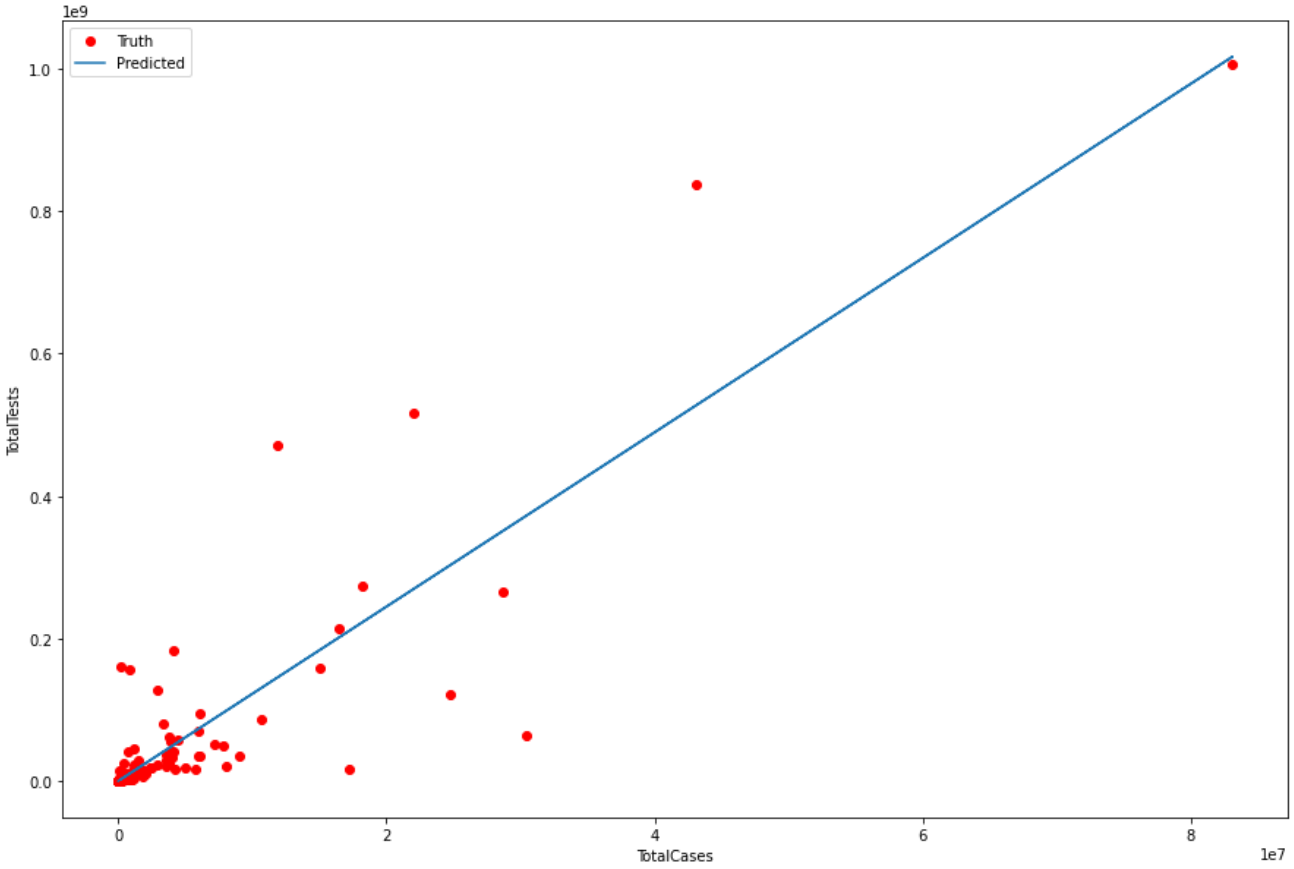
TotalCase, TotalDeaths: Mức độ biểu diễn cho 2 trường dữ liệu này là : 0.8230249736194025



TotalCases, TotalRecovered: Mực độ biểu diễn cho 2 trường dữ liệu này là : 0.9716457514194253.

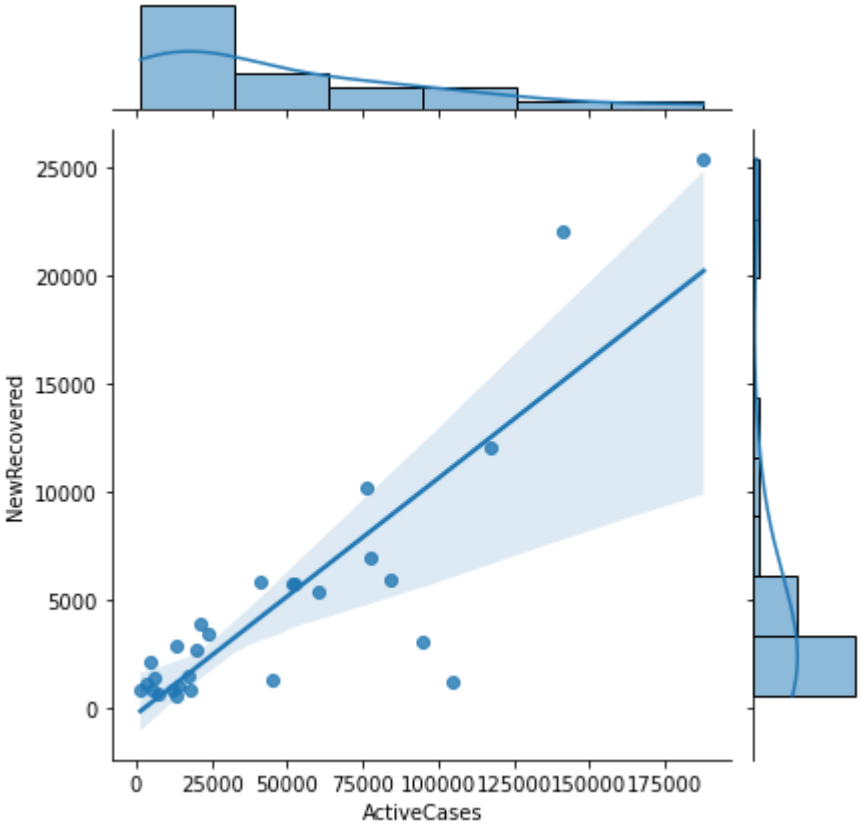


TotalCases, TotalTests: Mức độ biểu diễn cho 2 trường dữ liệu này là: 0.774869808325452



Nhận xét: Hệ số biểu diễn của trường dữ liệu TotalCases và 3 trường TotalDeaths, TotalTests, TotalRecovered là rất cao. Từ đó ta có thể dữ đoán được số ca tử vong/hồi phục/test dựa trên số liệu của số ca nhiễm.

3.7. Ở các nước vẫn còn đang dịch bệnh, sự tương quan giữa số ca hồi phục và số ca nhiễm là như thế nào?



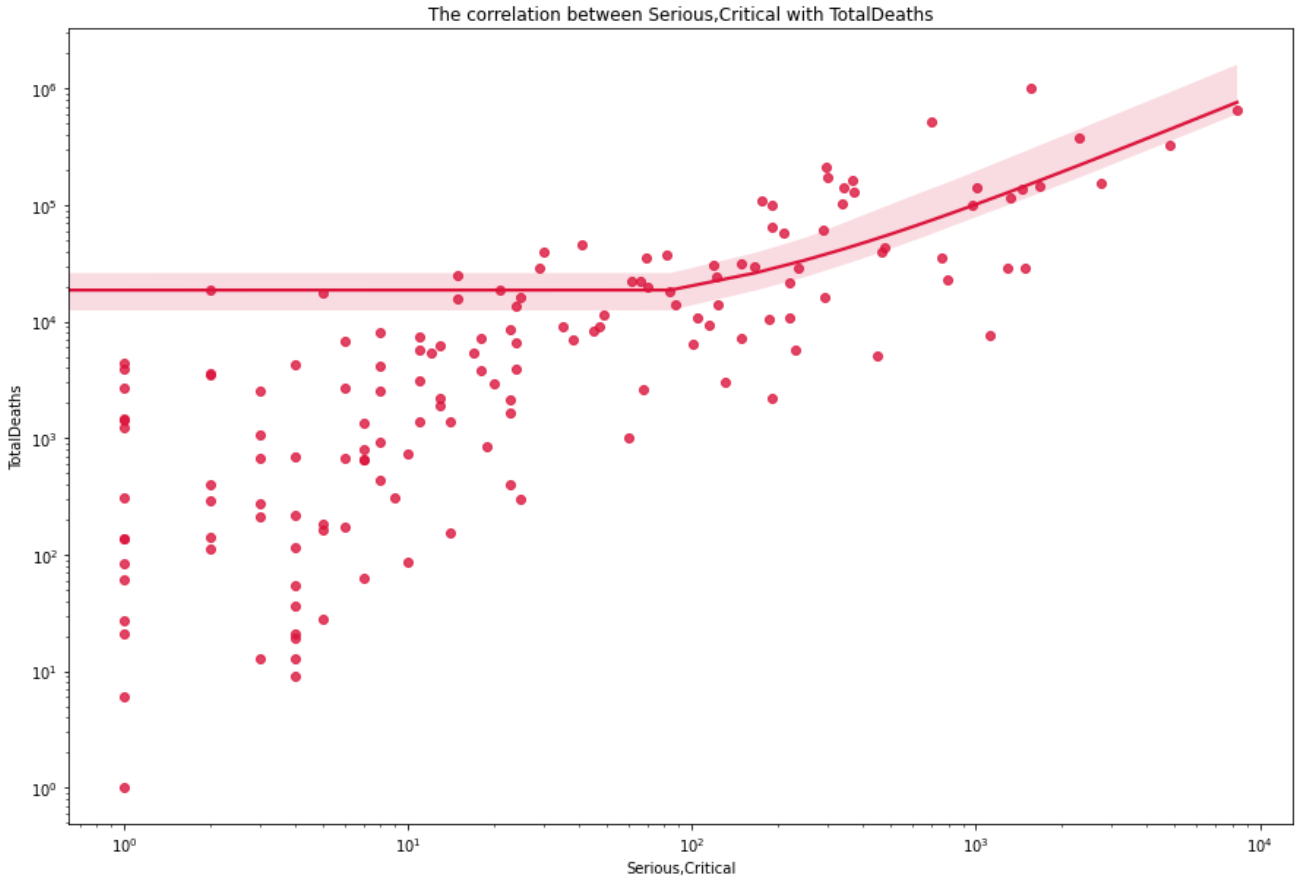
Kiểu biểu đồ: Joint plot

Giải thích: Dạng biểu đồ Joint Plot nhằm cho ta thấy được sự phụ thuộc của 1 biến NewRecovered vào 1 biến độc lập khác ActiveCases.

Nhận xét:

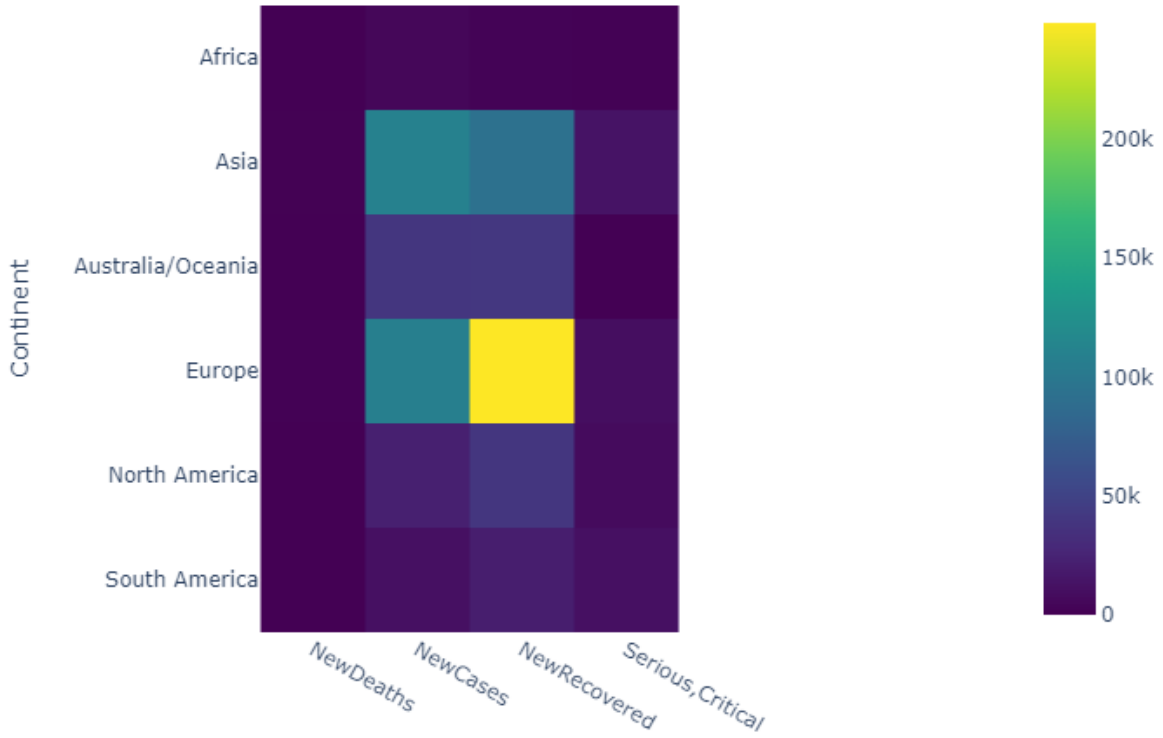
- Quan sát biểu đồ, ta thấy được 2 biến này có sự tương quan với nhau, mức độ tập trung dữ liệu cũng tương đối.
- Số điểm nằm dưới đường thẳng hồi quy tương đối ít hơn những điểm nằm trên và lên cao dần, điều này cho thấy rằng tình hình hồi phục sau khi nhiễm của các quốc gia đang có chiều hướng tốt lên. Có thể 1 phần khẳng định các quốc gia đã có kinh nghiệm và đủ y tế để gia tăng mức độ hồi phục, ngoài ra thì việc tiêm vaccine cũng có thể giúp 1 phần trong việc này.

3.8. Tương quan giữa số ca nhiễm nặng và số ca tử vong.



Kiểu biểu đồ: Scatter plot
Giải thích: Biểu đồ Scatter plot nhằm thể hiện mối quan hệ giữa 2 biến số Serious,Critical và TotalDeaths, từ đó đánh giá được sự tương quan giữa chúng, các điểm trong biểu đồ cho ta cái nhìn tổng quát hơn về tập dữ liệu chung.
Nhận xét: Biểu đồ trên biểu diễn các điểm là các quốc gia trên thế giới dựa vào 2 thuộc tính Serious,Critical và TotalDeaths, ta thấy được có rất nhiều outlier. Vì vậy mức độ tương quan giữa 2 thuộc tính không cao, hay nói cách khác không phải số ca nhiễm nặng tăng thì số ca tử vong sẽ tăng theo.

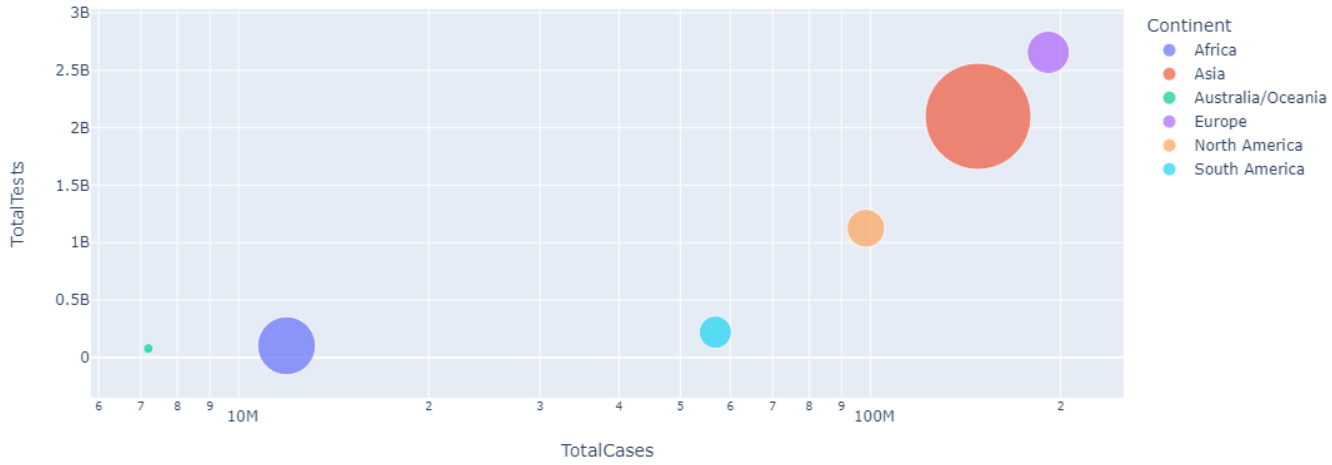
3.9. Tình hình chung của các châu lục tính đến hết ngày 1/5/2022.



Kiểu biểu đồ: Heatmap
Giải thích: Heatmap giúp trực quan dữ liệu 1 trách dễ dàng bằng thể hiện màu sắc bằng tone màu nóng và lạnh, người xem có thể nhanh chóng đánh giá và so sánh số lượng ca nhiễm, tình hình dịch bệnh ở các châu lục.
Nhận xét:

- Số lượng ca nhiễm mới xảy ra nhiều ở khu vực châu Á và châu Âu. Tuy nhiên số ca hồi phục tại châu Âu lại cao gấp 3 lần so với châu Á, vì hầu hết là các nước phát triển tại châu Âu sẽ có nền y tế tiên tiến hơn, còn châu Á tập trung nhiều các nước đang phát triển, do đó không đảm bảo về chất lượng y tế.
- Các chỉ số dữ liệu còn lại không có sự chênh lệch quá cao giữa các châu lục, vì sắc độ màu tương đối gần nhau.

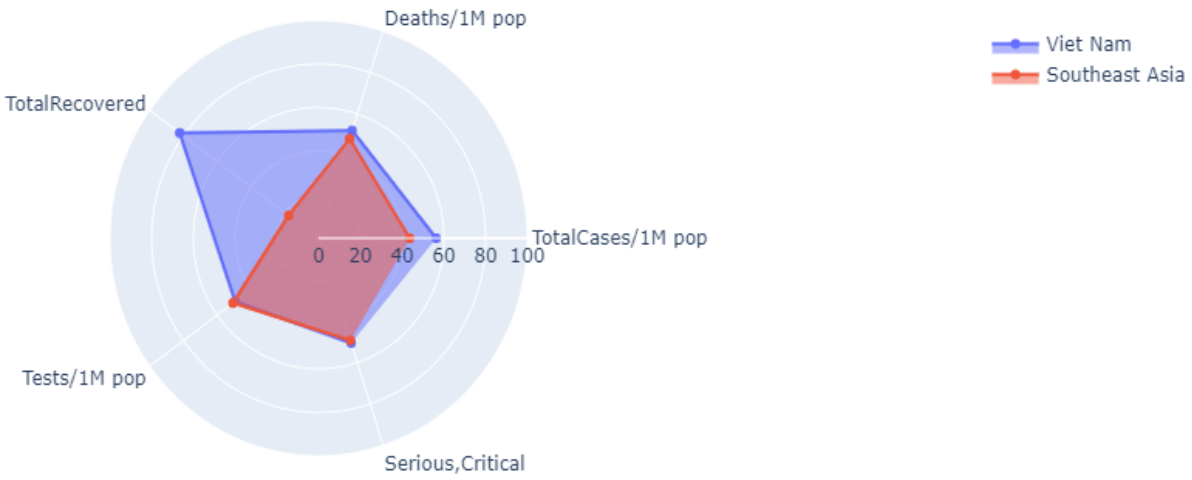
3.10. Liệu dân số của các châu lục có ảnh hưởng đến số lượng ca nhiễm và số lượng ca test của câu lục đó hay không?



Kiểu biểu đồ: Bubble plot.
Giải thích: Biểu đồ Bubble hữu ích trong việc trực quan hóa đồng thời 3 trường dữ liệu, thuận tiện cho việc so sánh giữa chúng.
Nhận xét: Ta dùng 3 trường dữ liệu Population, TotalTests, TotalCases so sánh giữa các châu lục.

- Châu Á là châu lục có dân số đông nhất, tương đương với việc tổng số ca nhiễm và số lần test sẽ cao.
- Châu Âu là châu lục có dân số đứng thứ 3 nhưng lại là châu lục có tổng số ca nhiễm và số lần test nhiều nhất mặc dù dân số chỉ bằng 1/6 châu Á.
- Châu Phi tuy là châu lục có dân số đứng thứ 2 nhưng lại có tổng số ca nhiễm và số lần test ít nhất.
- Vì vậy ta có thể nhận xét tương đối rằng yếu tố dân số không phải là yếu tố quyết định, ảnh hưởng nhiều đến tổng số ca nhiễm và số lần test của châu lục.

3.11. Tình hình dịch bệnh của Việt Nam so với khu vực Đông Nam Á tính đến hết ngày 1/5/2022.



Kiểu biểu đồ: Multi Radar Chart.
Giải thích: Biểu đồ này giúp ta thực hiện việc so sánh nhiều thông số của nhiều đối tượng khác nhau cùng 1 lúc.
Nhận xét:

- Các chỉ số về Tests/1M pop, TotalCases/1M pop, Deaths/1M pop, Serious,Critical ở Việt Nam nằm ở mức tương đương so với các nước còn lại trong khu vực Đông Nam Á, không trội hơn quá nhiều.
- Tuy nhiên biểu đồ cho thấy số ca hồi phục của Việt Nam lại cao gấp 4 lần so với các nước khu vực. Điều này cho thấy được sự nổ lực của nền y tế Việt Nam trong việc điều trị covid, hiệu quả hơn rất nhiều so với mặt bằng chung các nước khác của khu vực Đông Nam Á.

4. Tài liệu tham khảo:

- Slide lí thuyết của thầy Bùi Tiến Lên
- [Matplotlib Document] (https://matplotlib.org/stable/tutorials/index)
- [Seaborn Document] (https://seaborn.pydata.org/)
- [Plotly Express Document] (https://plotly.com/python/plotly-express/)