TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO LAB 01

DATA RELATIONSHIP

Học phần: Trực quan hóa dữ liệu

Lớp: 19_21

Giáo viên hướng dẫn: Lê Ngọc Thành

Sinh viên thực hiện:

- Lê Kiệt – 19120554
- Nguyễn Minh Lương – 19120571
- Lưu Trường Dương – 19120489

Hồ Chí Minh, ngày 3 tháng 5 năm 2021

MỤC LỤC

I. MỨC ĐỘ HOÀN THÀNH 3
1. MỨC ĐỘ HOÀN THÀNH CỦA MỖI YÊU CẦU 3
2. MỨC ĐỘ HOÀN THÀNH CỦA MỖI THÀNH VIÊN 3
II. DATASET 3
III. TRỰC QUAN HÓA CÁC QUAN HỆ 4
1. Trường đơn 4
a) Total Cases 4
b) Total Deaths 4
c) New Cases, New Deaths, New Recovered 6
d) Active Cases 8
e) New Deaths 8
2. Nhiều trường 9
a) Total Deaths ~ Total Recovered 9
b) Total Deaths ~ Total Recovered 11
c) Total Deaths ~ Total Recovered ~ Active Cases 11
d) Total Cases ~ Total Deaths ~ Total Recovered ~ Total Tests 12
e) Total Cases ~ (Active Cases, Total Deaths, Total Recovered) 13
f) New Deaths & các cột còn lại 17
g) New Cases & Density 18
3. Quan hệ nhân quả 19
a) Khảo sát mối quan hệ nhân quả [Total Recovered, Serious] → Total Tests 20
b) Khảo sát mối quan hệ nhân quả giữa 2 cột (Total Cases, Total Recovered) 20
c) Khảo sát mối quan hệ nhân quả giữa New Deaths & New Cases 22
III. THAM KHẢO 25

I. MỨC ĐỘ HOÀN THÀNH

1. MỨC ĐỘ HOÀN THÀNH CỦA MỖI YÊU CẦU

STT	Tên yêu cầu	Mức độ hoàn thành
1	Thu thập dữ liệu	100%
2	Tiền xử lý dữ liệu, dữ liệu 1 vài ngày sau khi tiền xử lý được lưu trong tệp Worldometer-data-preprocessed	100%
3	Thể hiện quan hệ trường đơn	100%
4	Thể hiện quan hệ giữa nhiều trường dữ liệu	100%
5	Thể hiện mối quan hệ nhân quả	100%
6	Sử dụng nhiều hơn 1 loại biểu đồ cho vài quan hệ	100%

2. MỨC ĐỘ HOÀN THÀNH CỦA MỖI THÀNH VIÊN

Tên	Nhiệm vụ	Mức độ hoàn thành
Lê Kiệt	- Xây dựng pipeline cào dữ liệu từ trang https://www.worldometers.info/coronavirus/ - Xây dựng pipeline tiền xử lý dữ liệu - Tìm và trực quan các quan hệ trường đơn (2 quan hệ) /nhiều trường (2 quan hệ) /nhân quả (1 quan hệ)	100%
Nguyễn Minh Lương	- Bổ sung vào pipeline tiền xử lý dữ liệu - Tìm và trực quan các quan hệ trường đơn (2 quan hệ) /nhiều trường (1 quan hệ) /nhân quả (1 quan hệ)	100%
Lưu Trường Dương	- Bổ sung và chỉnh sửa vào pipeline tiền xử lý dữ liệu - Tìm và trực quan các quan hệ trường đơn (1 quan hệ) /nhiều trường (3 quan hệ) /nhân quả (1 quan hệ)	100%

II. DATASET

- Dataset được cào từ trang https://www.worldometers.info/coronavirus/, từ ngày 18/04/2022-04/05/2022, source code dùng để cào dữ liệu nằm trong file **Worldometer-crawler.ipynb**; dữ liệu thô cào được được lưu dưới dạng file csv nằm trong thư mục Worldometer-data
- Một vài file dữ liệu thô được tiền xử lý và lưu trong thư mục Worldometer-data-preprocessed, source code để tiền xử lý dữ liệu nằm trong file **Preprocess.ipynb**. Ngoài ra thư mục Worldometer-data-preprocessed còn chứa dữ liệu cho 1 tuần "data_1_week.csv" dùng cho trực quan hóa các quan hệ liên quan tới thời gian lúc sau
- Thư mục utils chứa các file
  - **continents.csv**: chứa thông tin châu lục của từng nước, file này được cào từ chính trang https://www.worldometers.info/coronavirus/, source code dùng để cào thông tin châu lục nằm trong **Worldometer-crawler.ipynb**

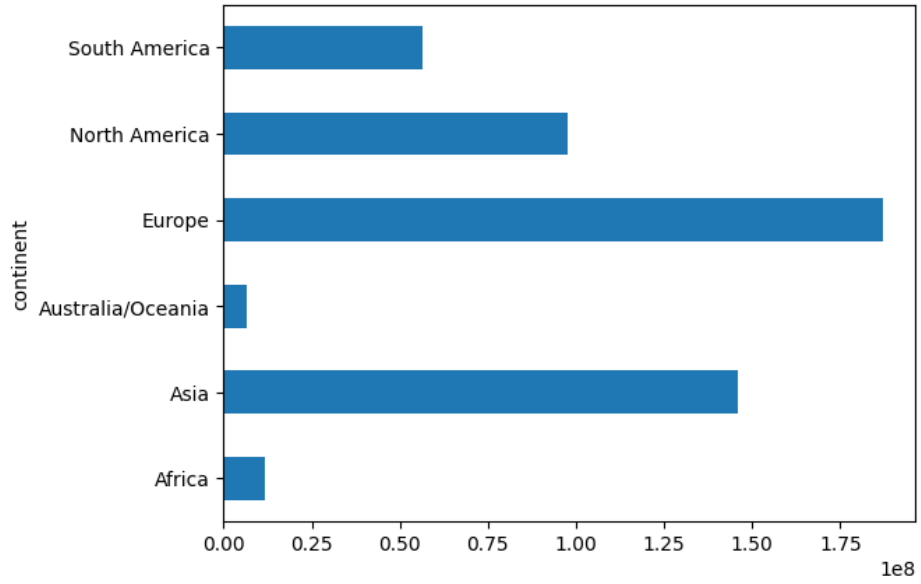
- **population density.csv**: Dữ liệu mật độ dân số theo từng nước, lấy từ nguồn https://worldpopulationreview.com/country-rankings/countries-by-density

III. TRỰC QUAN HÓA CÁC QUAN HỆ

1. Trường đơn

a) Total Cases

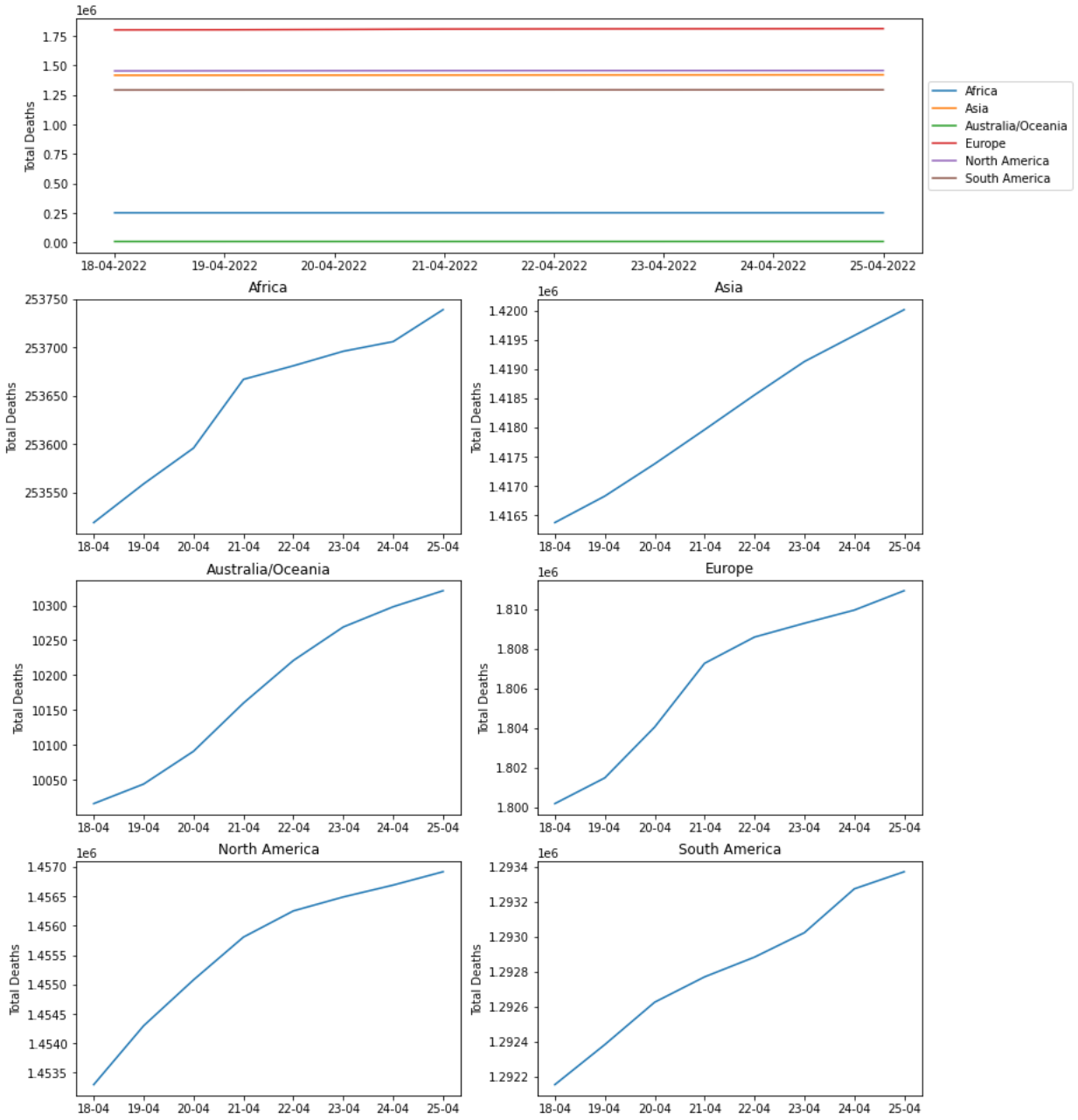
- Ta sẽ gom nhóm theo 6 châu lục, mỗi châu lục sẽ tính tổng Total Cases. Vì vậy ta chọn biểu đồ bar chart để so sánh Total Cases giữa các châu



➔ Nhận xét: tổng số ca nhiễm ở châu Âu và châu Á là nhiều nhất so với các châu lục còn lại. Điều này khá dễ hiểu vì châu Á có Trung Quốc là nơi đầu tiên bùng phát dịch bệnh; còn châu Âu thì đa phần người dân không đeo khẩu trang

~~a)~~b) Total Deaths

- Đầu tiên, ta gom nhóm theo 2 thuộc tính continent & day, sau đó lấy tổng trên thuộc tính Total Deaths để cho biết với mỗi châu lục vào mỗi ngày từ 18-25/4/2022 thì tổng số người chết là bao nhiêu
- Vì tính chất dữ liệu theo thời gian nên ta dùng biểu đồ line chart để thể hiện sự thay đổi tổng số người chết - Total Deaths trong 1 tuần từ 18-25/4/2022



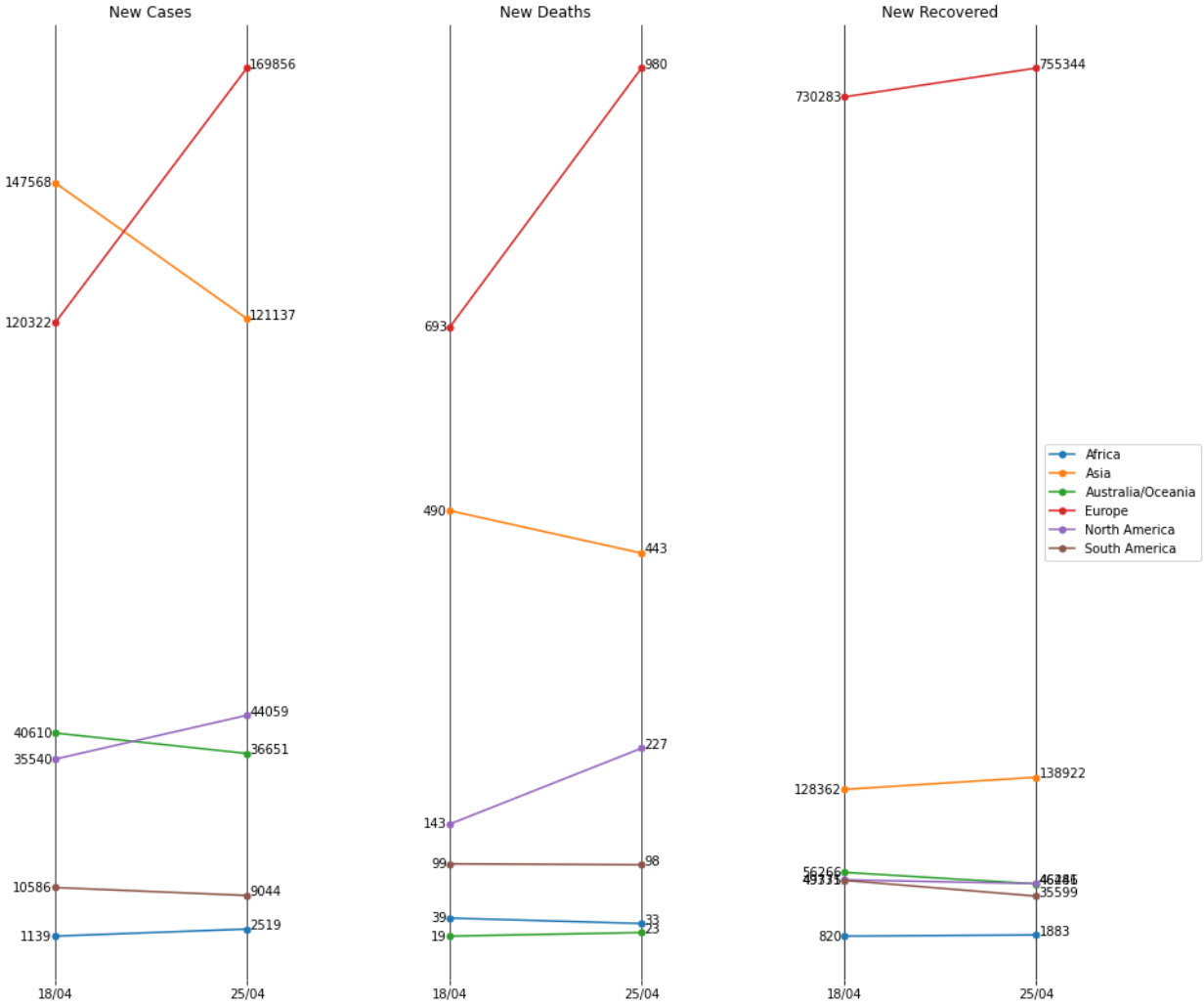
- Biểu đồ đầu tiên là biểu đồ tổng quát về sự biến thiên Total Deaths cho cả 6 châu lục
- 6 biểu đồ còn lại là chi tiết biến thiên Total Deaths của từng châu lục
- Nhận xét:

  ➔ Với biểu đồ đầu tiên: Nhận thấy rằng 1 tuần vừa qua, số người chết hầu như không đổi ở 6 châu lục, ta sẽ xét cụ thể từng châu lục trong các biểu đồ dưới. Ngoài ra, dễ thấy tổng số ca chết tập trung cao ở 4 châu (giảm dần): Europe, North America, Asia và South America và tập trung thấp ở 2 châu còn lại: Africa & Australia/Oceania

➔ Qua các biểu đồ còn lại, thấy rằng thuộc tính Total Deaths có xu hướng tăng qua từng ngày bất kể châu lục nào, ví dụ Africa số ca chết tăng 200 người so với đầu tuần, Asia tăng khoảng 10000 ca chết so với đầu tuần, ...

~~b)~~c) New Cases, New Deaths, New Recovered

- Ta sẽ xem xét riêng lẻ 3 cột này trong 1 tuần, từ 18/4/2022 - 25/4/2022
- **Góc nhìn từ slope graph**: Ta mong muốn có 1 cái nhìn trực tiếp về tốc độ phát sinh các ca nhiễm mới (New Deaths)/ ca chết mới (New Deaths)/ ca hồi phục mới (New Recovered) theo thời gian. Đây là 1 thể loại con của line graph

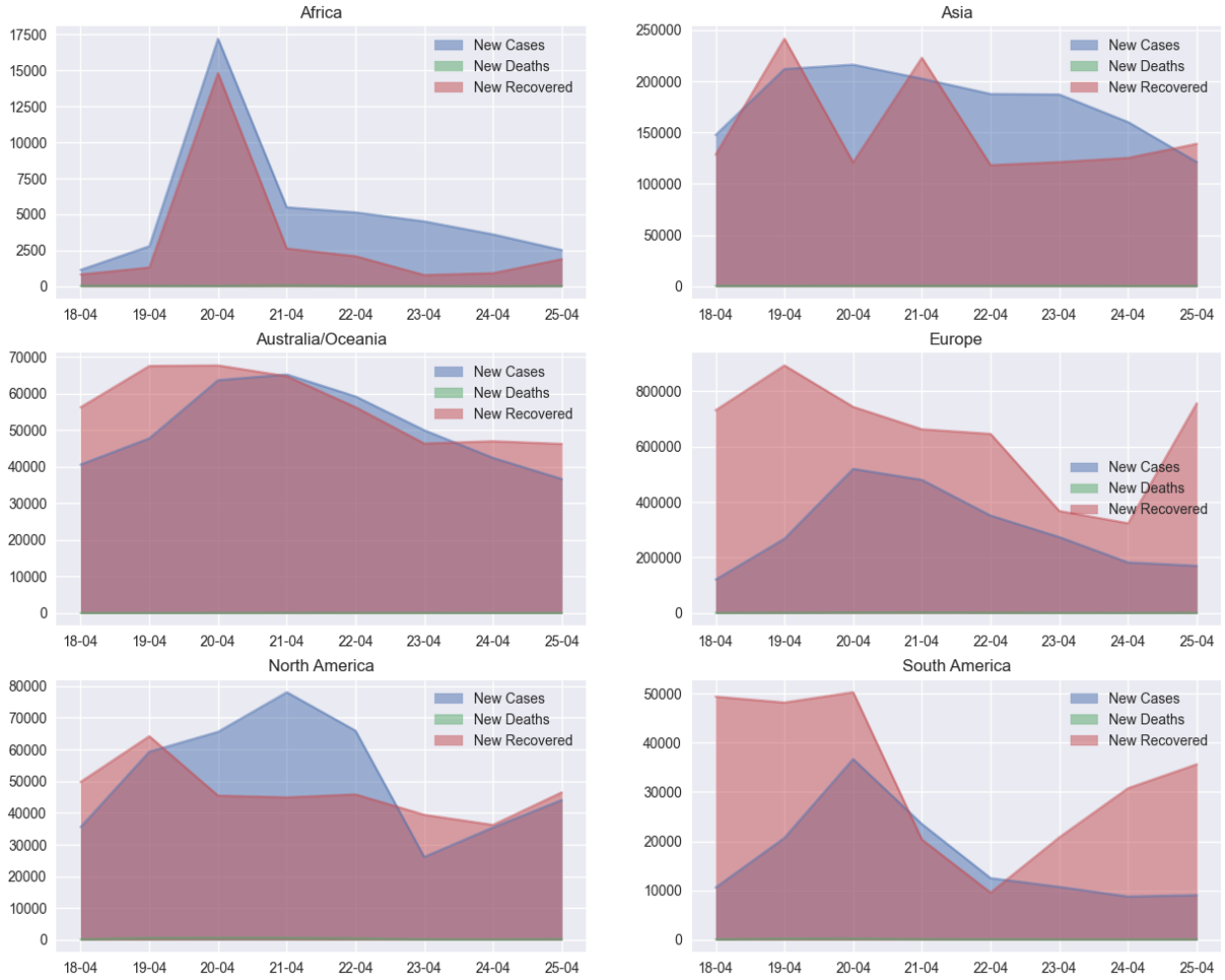


- Nhận xét:

➔ New Cases: thuộc tính này có tốc độ thay đổi diễn ra mạnh hơn 2 thuộc tính còn lại. Europe trong 1 tuần tăng khoảng 49000 ca nhiễm mới, tốc độ lây lan nhanh nhất so với các châu còn lại, kế tiếp là North America. Asia và Australia/Ocenia là 2 châu lục giảm mạnh trong tuần, tuy vậy nhưng Asia vẫn xếp thứ 2 về số lượng lượng ca nhiễm mới. Còn Africa và South America lần lượt tăng và giảm 1 lượng không nhiều

➔ New Deaths: tương tự thì Europe và Asia lần lượt là 2 châu lục tăng và giảm nhanh so với các châu còn lại, nhưng vẫn xếp hạng có New Deaths cao. North America tăng hơn 100 ca chết trong 1 tuần. Các châu lục còn lại chết dưới 10 người trong tuần qua
➔ New Recovered: thuộc tính này nhìn chung tốc độ tăng rất chậm, đôi khi là giảm. Dễ thấy tuy Europe có số ca nhiễm & ca chết tăng mạnh, nhưng tốc độ hồi phục cũng rất nhanh

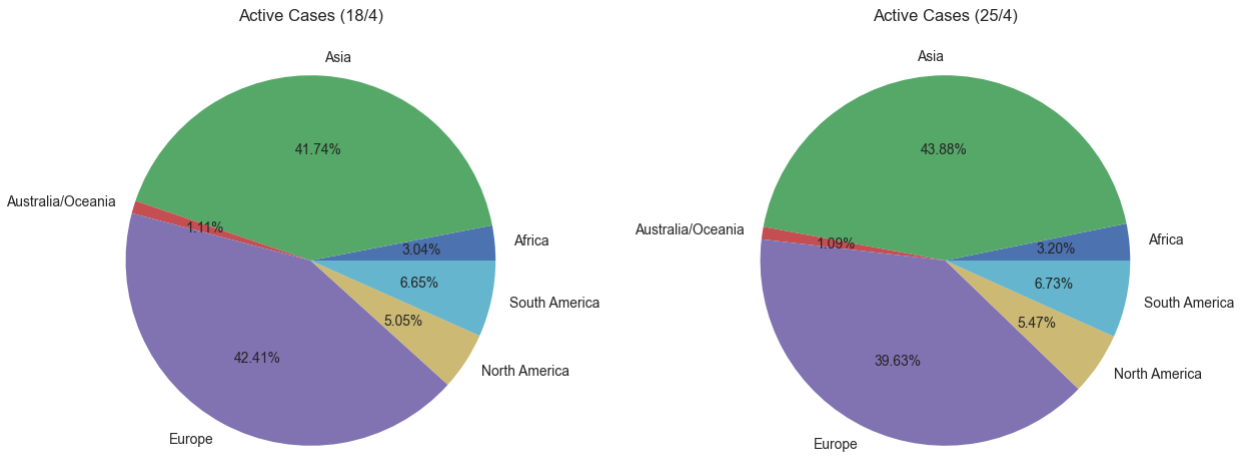
- **Góc nhìn từ area plot**: slope graph cho ta biết về tốc độ thay đổi của 3 cột New Cases, New Deaths, New Recovered. Ngoài ra ta còn muốn biết sự thay đổi này đáng kể hay không đáng kể so với tổng thể chung, chính vì vậy mà area chart được dùng



- Nhận xét:
  ➔ Tổng số người chết - New Deaths trên cả thế giới (6 châu lục) thay đổi rất nhỏ so với lượng New Cases & New Recovered
  ➔ Chiếm phần nhiều nhất hầu như là New Recovered, tuy tốc độ hồi phục theo như slope graph là chậm, nhưng số lượng người hồi phục thì rất nhiều, và nhiều hơn New Cases
  ➔ 1 điểm nhỏ nổi bật ở Asia: ngày 19 và 21/4/2022 là 2 ngày châu Á đạt 2 đỉnh

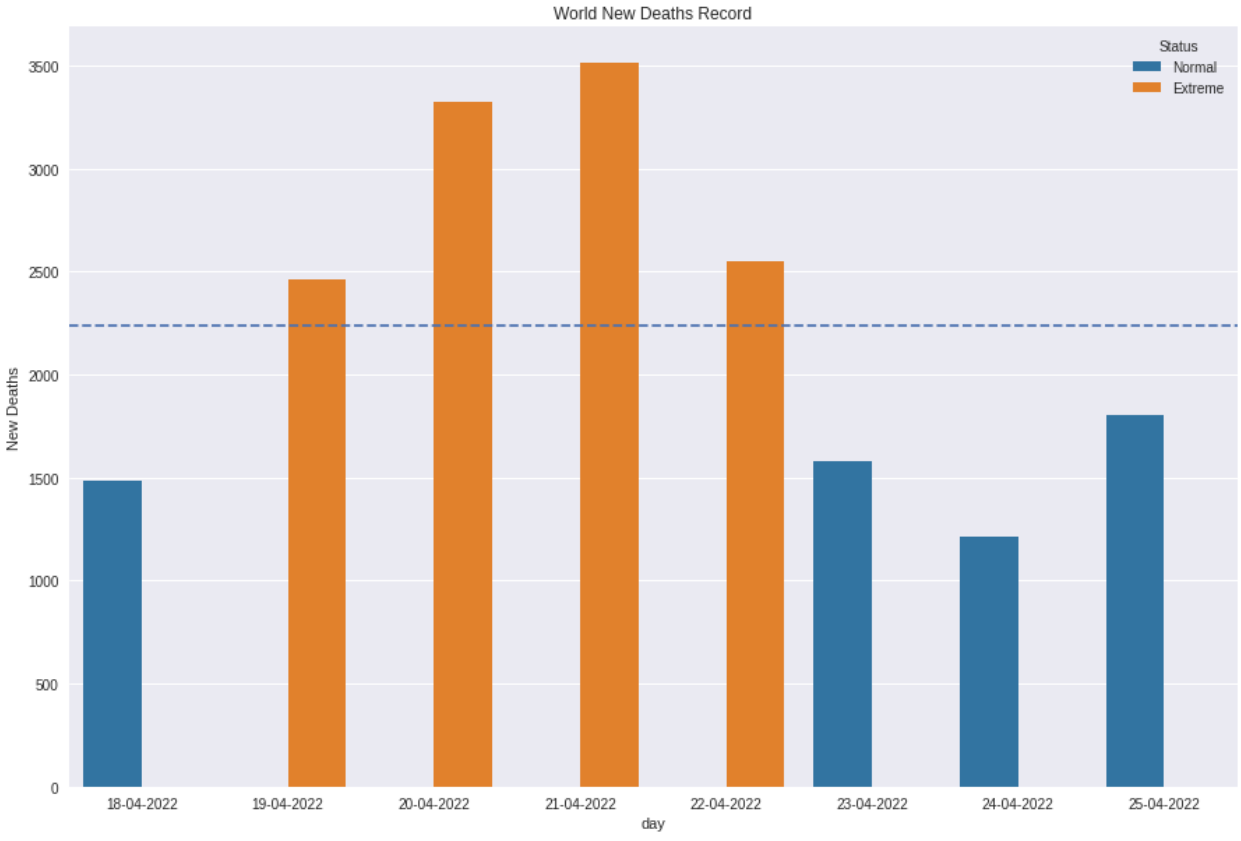
c)d) Active Cases

- Ta so sánh Active Cases của mỗi châu lục thay đổi thế nào sau 1 tuần, vì vậy nên pie chart được dùng

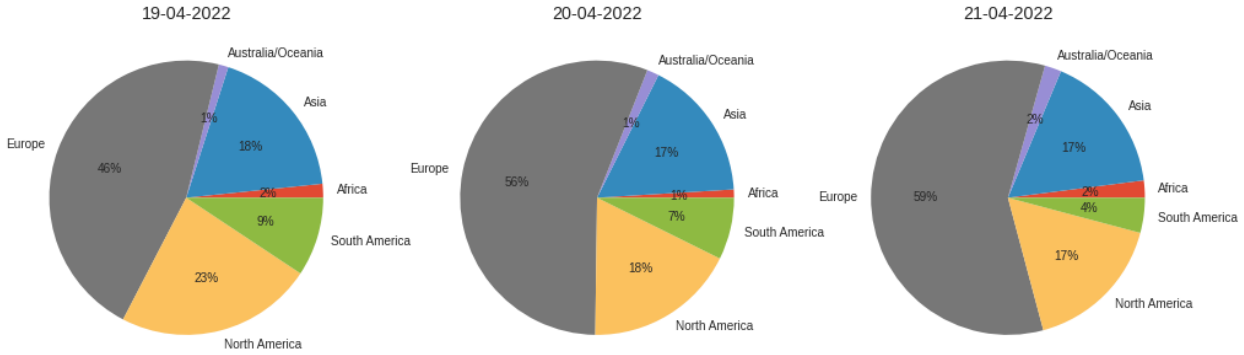


- Nhận xét: Active Cases trong 1 tuần của các châu lục tăng, trừ Europe & Australia/Oceania giảm

d)e) New Deaths



- Chọn biểu đồ cột vì đây là biểu đồ đơn giản để so sánh dữ liệu trong 1 khoảng thời gian ngắn. Có thể lấy số liệu của 1 tuần để có thể thể hiện rõ ràng hơn sự biến đổi dữ liệu trong 1 khoảng thời gian.
- Nhận xét: Trong 3 ngày 19/04/2022 → 21/04/2022, số lượng ca tử vong mới tăng nhiều hơn hẳn các ngày còn lại. Ta sẽ tiến hành xem kỹ hơn số liệu của 3 ngày này.

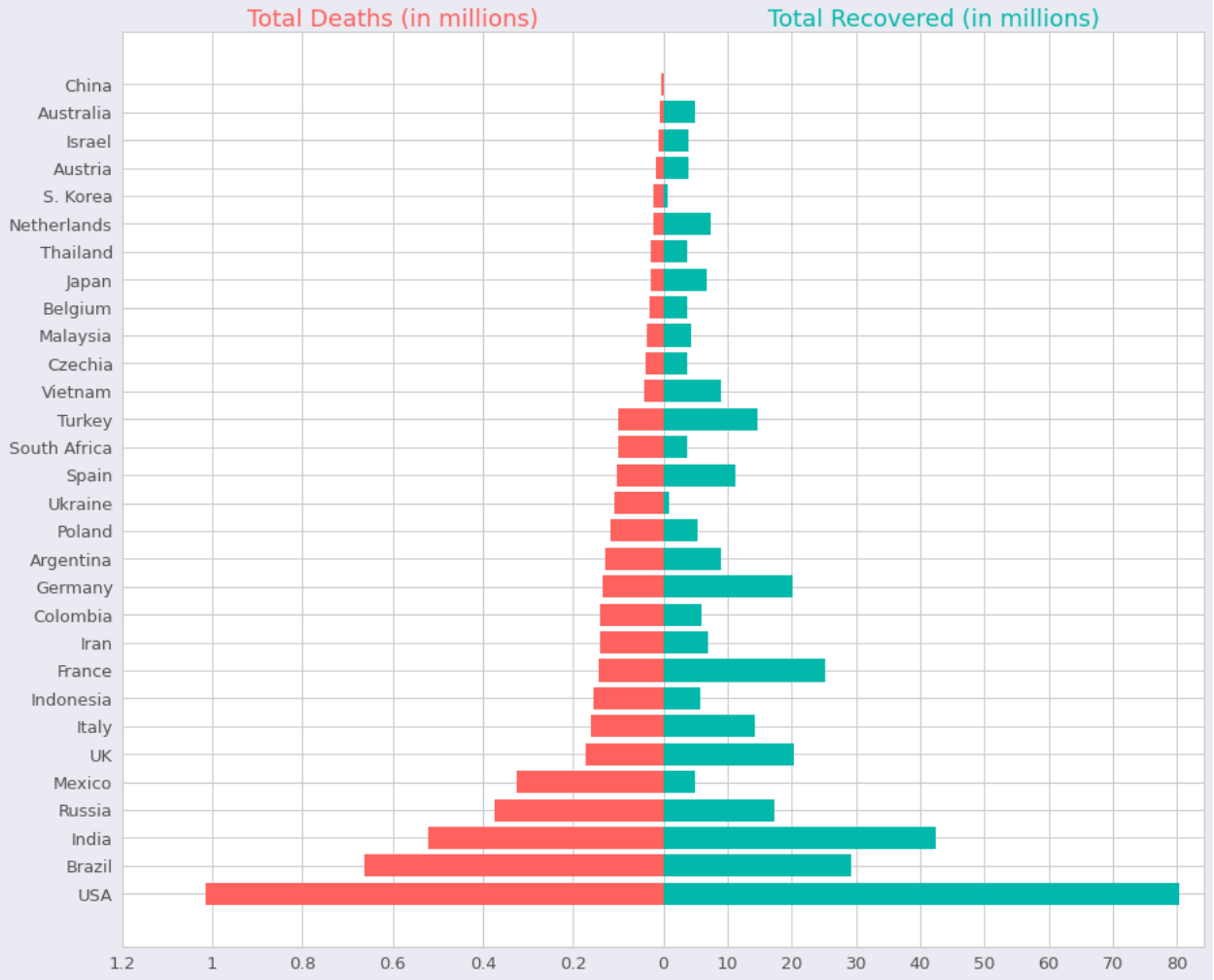


- Sử dụng biểu đồ pie chart vì đây là dạng biểu đồ dùng để thể hiện phần trăm của từng thành phần cấu tạo nên 1 quần thể. Từ đó chúng ta có thể dễ dàng so sánh sự đóng góp vào quần thể của từng thành phần.
- Nhận xét: Trong cả 3 ngày 19, 20, 21/04/2022 thì **Europe** đều có số lượng ca tử vong mới cao nhất, tiếp đó là **North America** và **Asia** là 3 châu lục có số lượng ca tử vong mới cao nhất trên toàn thế giới.

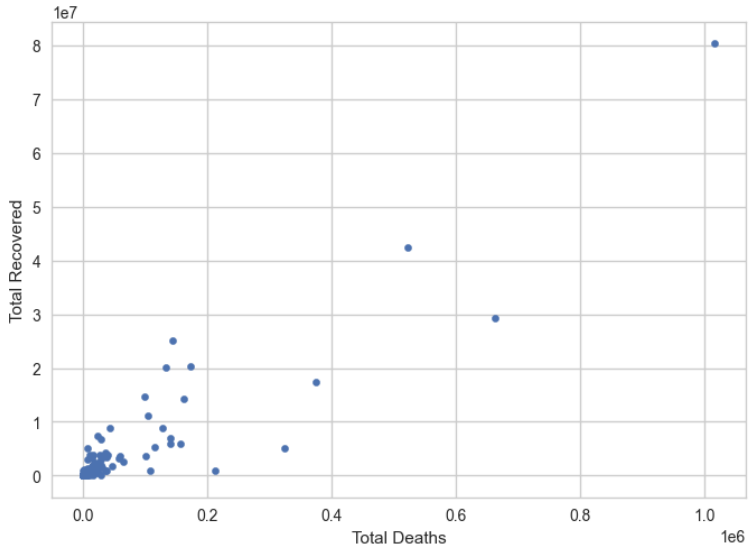
2. Nhiều trường

a) Total Deaths ~ Total Recovered

- **Góc nhìn từ bidirectional bar chart:** Vì Total Deaths và Total Recovered là 2 thái cực trái nhau nên nhóm quyết định dùng bidirectional bar chart để thể hiện mối quan hệ giữa 2 trường này, với 30 nước đầu tiên được xếp theo thứ tự tăng dần của Total Deaths.

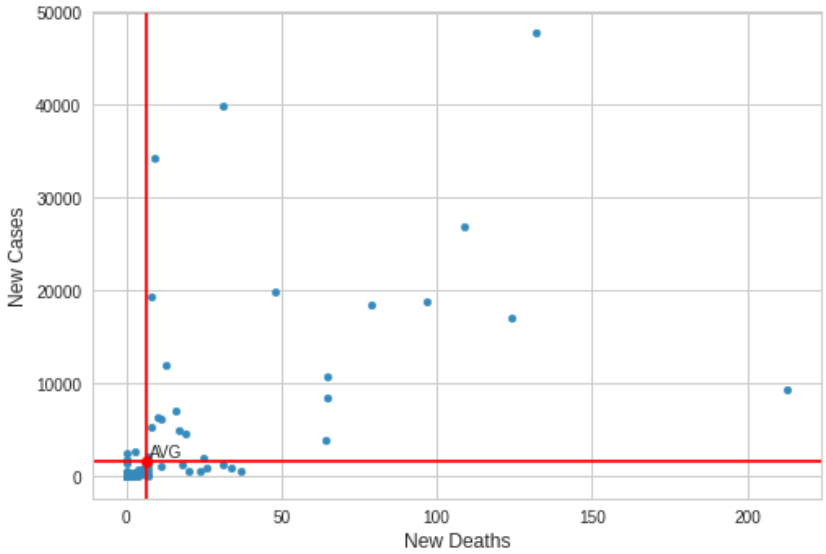


- Nhận xét: Các nước được xếp theo thứ tự tăng dần Total Deaths. Nhìn chung Total Deaths tăng thì Total Recovered nhiều khả năng cũng tăng theo, tuy nhiên mối quan hệ này không quá mạnh (ta có thể kiểm chứng bằng biểu đồ scatter phía dưới). Ngoài ra thì tỉ lệ Recovered luôn cao gấp hơn 10 lần so với tỉ lệ Deaths (Ví dự ở nước USA thì Total Recovered cao gấp 80 lần so với Total Deaths)
- **Góc nhìn từ scatter plot**: Mục đích là để kiểm chứng lại xem mối tương quan giữa Total Deaths & Total Recovered có phải tương quan dương và nếu có thì tương quan này mạnh hay yếu



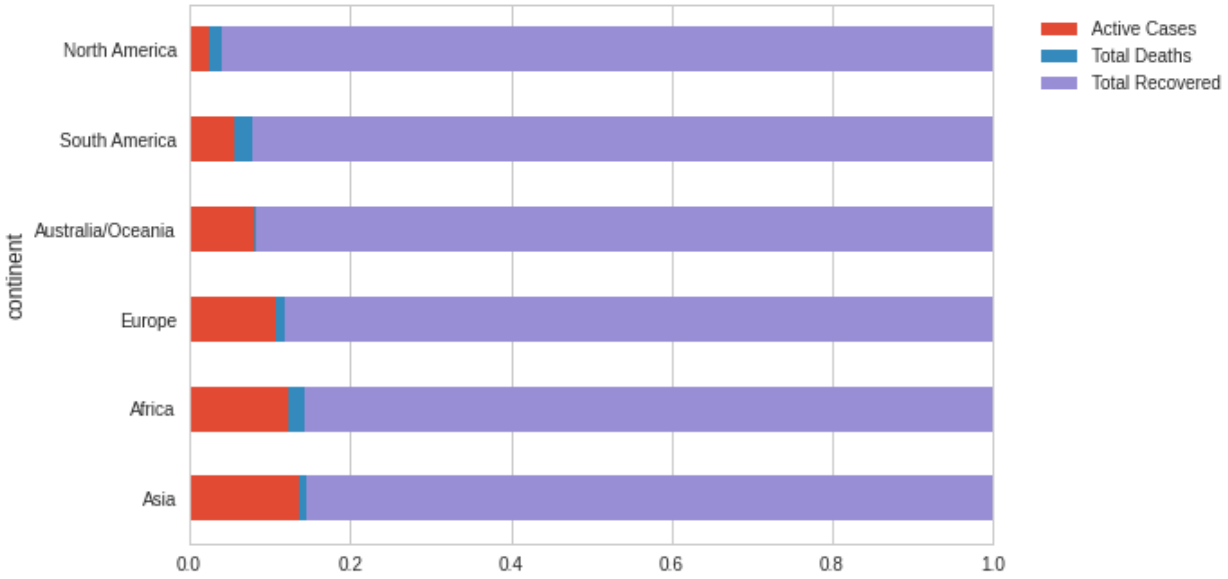
- Nhận xét: đây là mối tương quan dương nhưng càng yếu dần khi Total Deaths > 0.2

b) Total Deaths ~ Total Recovered
- Xét quan hệ giữa New Cases và New Deaths bằng scatter plot



- Nhận xét: đây là mối tương quan dương nhưng không mạnh. Thấy rằng lượng New Deaths lớn hơn trung bình khi New Cases đạt ít nhất khoảng 1000 ca; tuy nhiên, từ đồ thị ta thấy không phải cứ New Cases > 1000 ca thì New Deaths trên trung bình. Tương tự, New Cases trên trung bình khi **đa số** New Deaths > 6.56.

c) Total Deaths ~ Total Recovered ~ Active Cases
- Theo đường link này, Total Cases cấu tạo bởi 3 thành phần: Active Cases, Total Deaths & Total Recovered, ta sẽ dùng 100% stacked bar chart để xét tỉ trọng của 3 thành phần.

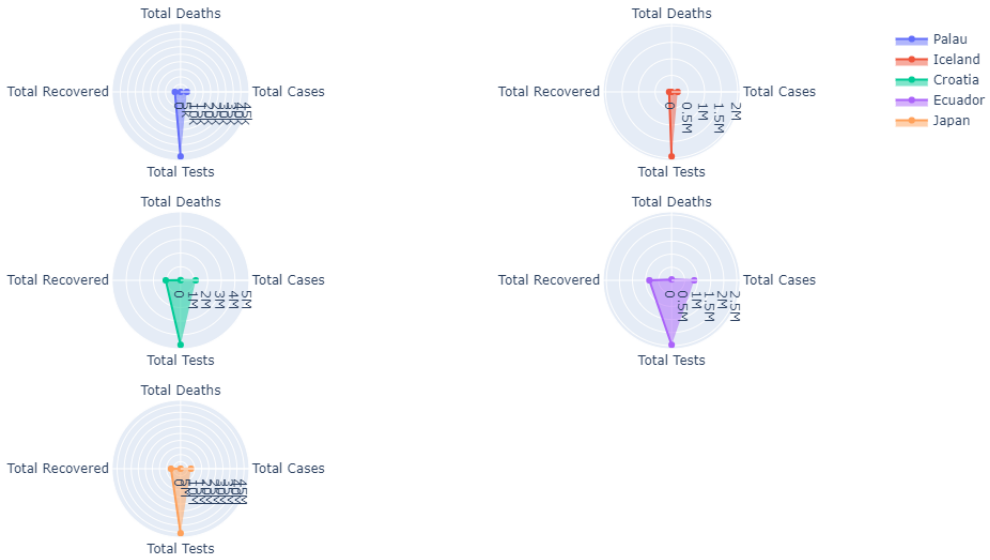


- Nhận xét: Total Recovered chiếm tỉ trọng nhiều nhất trong 3 thành phần. Các châu lục được sắp xếp theo chiều giảm dần của Total Recovered, nhờ đó thấy rằng Total Recovered càng giảm thì Active Cases càng tăng.

d) Total Cases ~ Total Deaths ~ Total Recovered ~ Total Tests

- Dùng biểu đồ **circular coordinates** có 4 góc (4 thuộc tính trên); với 5 polygon là 5 nước, lần lượt có Population thuộc [0, 20%], [20, 40%],...,[80, 100%]. Mục đích để kiểm tra giả thiết: "Population càng lớn thì thuộc tính x càng tăng".

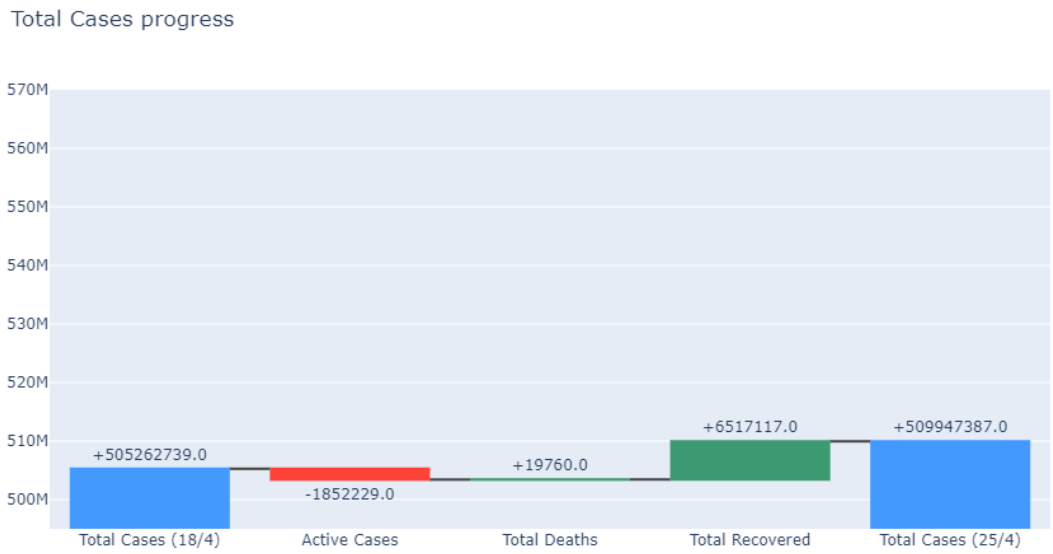
- Đầu tiên, cần chọn ra nhóm các nước thuộc vào các khoảng phân vị [0, 20%], [20, 40%],..., [80, 100%] lần lượt tương ứng với các nhãn 5,4,3,2,1. Từ mỗi nhóm, ta lấy 1 nước đại diện, nước đại diện này có Population là trung vị trong nhóm đó. Cuối cùng ta thu dươc 5 nước đại diện cho 5 khoảng phân vị.



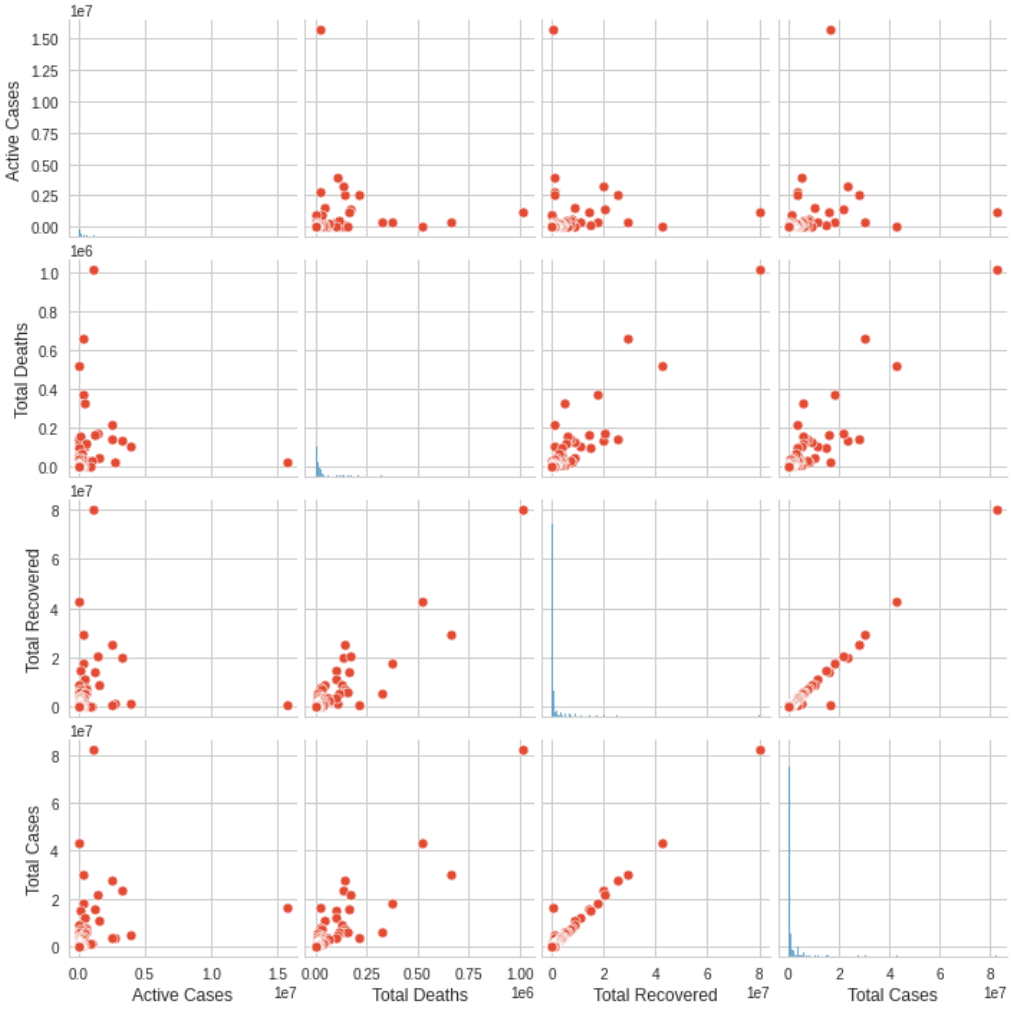
- Nhận xét:
  - Total Population ~ Total Tests: Dân số càng tăng, Total Tests cũng gần như tăng theo, trừ Ecuador dân đông nhưng Total Tests lại ít hơn Croatia
  - Total Population ~ Total Cases: Dân số càng tăng, số ca nhiễm cũng tăng theo, điều này tương đối dễ hiểu
  - Total Recovered lên xuống thất thường khi Total Population tăng, chứng tỏ còn nhiều yếu tố khác ảnh hưởng tới trường dữ liệu này

e) Total Cases ~ (Active Cases, Total Deaths, Total Recovered)

- Ta sẽ xét tổng số ca nhiễm Total Cases trên cả thế giới (World) từ trong 1 tuần, từ 18-25/4/2022, và xem cụ thể các trường Active Cases, Total Deaths, Total Recovered ảnh hưởng như thế nào tới tổng số ca nhiễm 1 tuần như công thức đã được đề cập ở link này: '"**Active Cases = (total cases) - (total deaths) - (recovered)**"'.
- Đầu tiên, ta lấy dữ liệu ngày 18 và 25 cùng các cột liên quan. Sau đó, tính sự chênh lệch (hiệu) của các cột Active Cases, Total Deaths, Total Recovered giữa ngày 18 và 25 để biết trong 1 tuần thì các trường dữ liệu này tăng/giảm ra sao. Vì sự thay đổi của từng thành phần (Active Cases, Total Deaths, Total Recovered) sẽ dẫn tới sự thay đổi của phần tổng thể (Total Cases), vì thế biểu đồ waterfall được sử dụng để nhấn mạnh vào sự thay đổi thành phần.

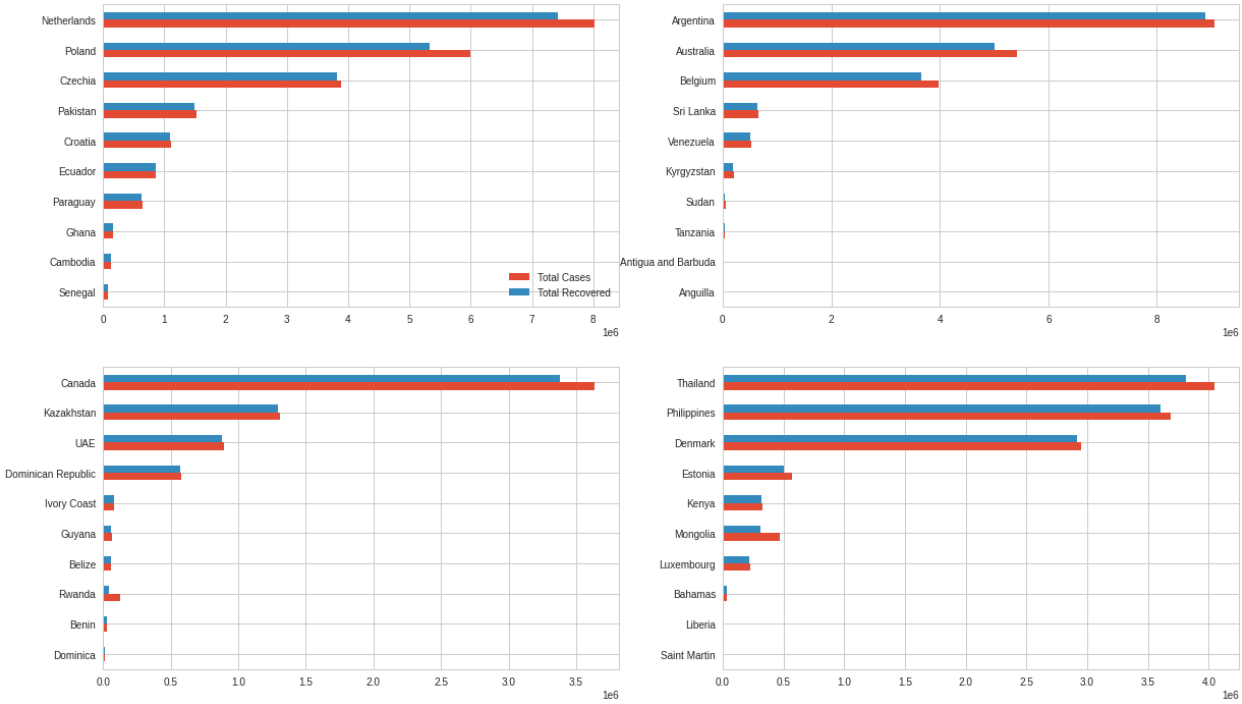


- Nhận xét: Total Cases sau 1 tuần tăng, cụ thể sự tăng này là do Active Cases giảm, và 2 trường Total Deaths và Total Recovered tăng (với Total Recovered tăng mạnh).
- Sẵn tiện ta coi kỹ từng quan hệ giữa các cặp thuộc tính ['Active Cases', 'Total Deaths', 'Total Recovered', 'Total Cases'].



- Nhận xét: Total Recovered & Total Cases có quan hệ gần như tuyến tính. Cụ thể là nếu Total Cases tăng/giảm thì Total Recovered cũng tăng/giảm, ta sẽ thử chọn ngẫu nhiên mỗi 10 nước và xây dựng biểu đồ **group bar chart** để xem thử có đúng như vậy không.

❖ *Total Cases ~ Total Recovered*

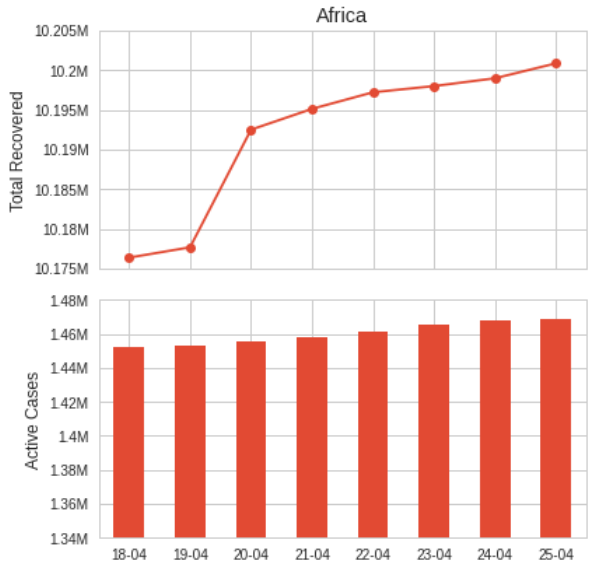


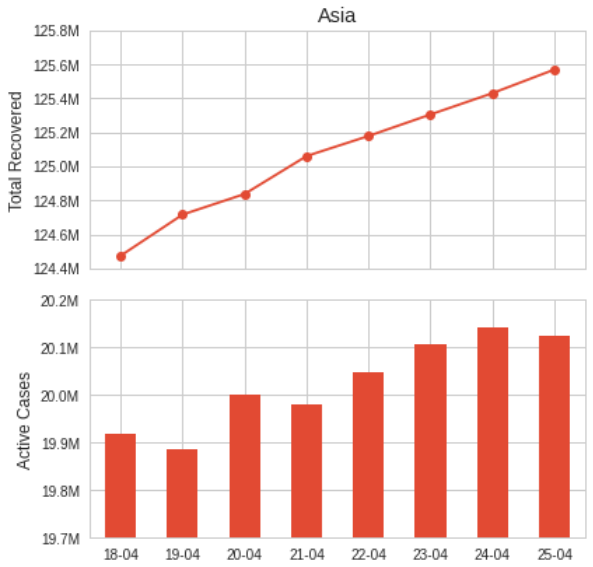
⇨ Mối quan hệ tuyến tính được thể hiện tương đối đúng ở hầu hết các nước, trừ 1 số nước đặc biệt thì có Total Cases cao hơn 1 số nước nhưng Total Recovered lại thấp hơn (VD: Ukraine, Tunisia, ...).

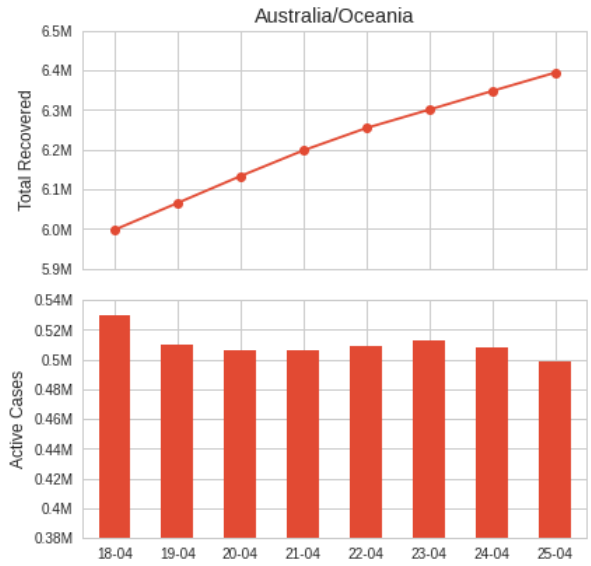
❖ *Active Cases & Total Recovered*

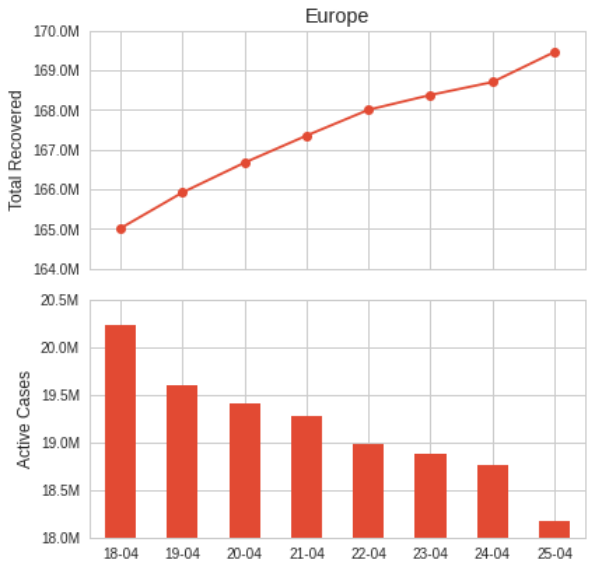
- Với mỗi châu lục, ta muốn biết rằng liệu sự thay đổi Active Cases theo thời gian 1 tuần từ 18-25/4/2022 có liên quan tới sự thay đổi của trường dữ liệu Total Recovered hay không.
- Ta kết hợp 2 loại biểu đồ: bar chart cho Active Cases & line chart cho Total Recovered. Lý do không dùng grouped bar chart với mỗi 2 cột cho 1 ngày là vì Active Cases & Total Recovered có giá trị cách nhau khá xa, vì dụ Active Cases có giá trị 1.46M trong khi Total Recovered lại có giá trị 10.175M, biểu đồ sẽ bị nén và khó nhìn ra xu hướng.

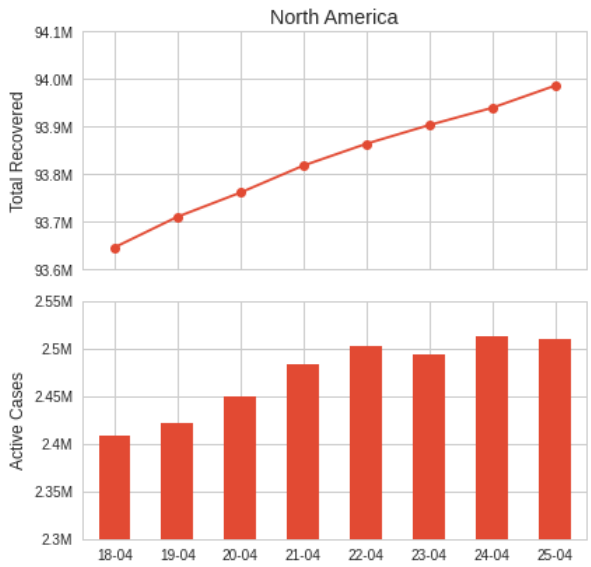
KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện Thoại: (08) 38.354.266 - Fax:(08) 38.350.096

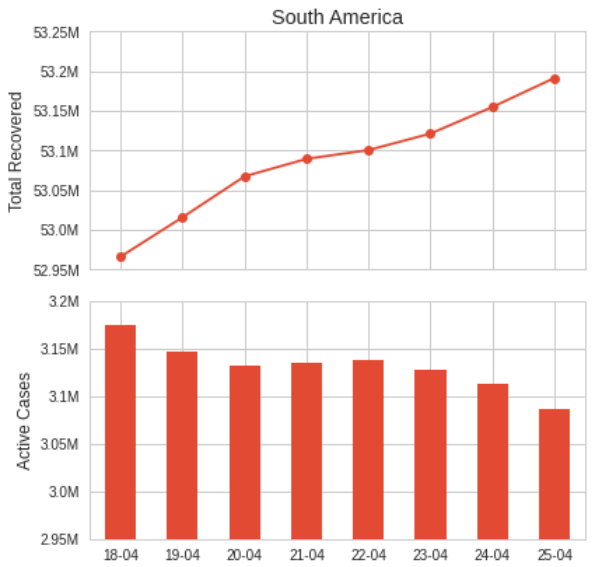




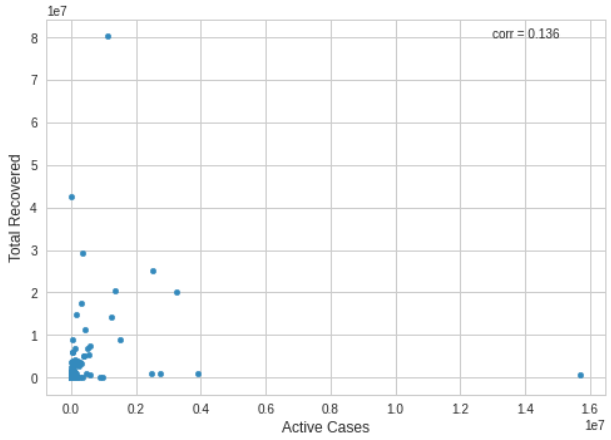








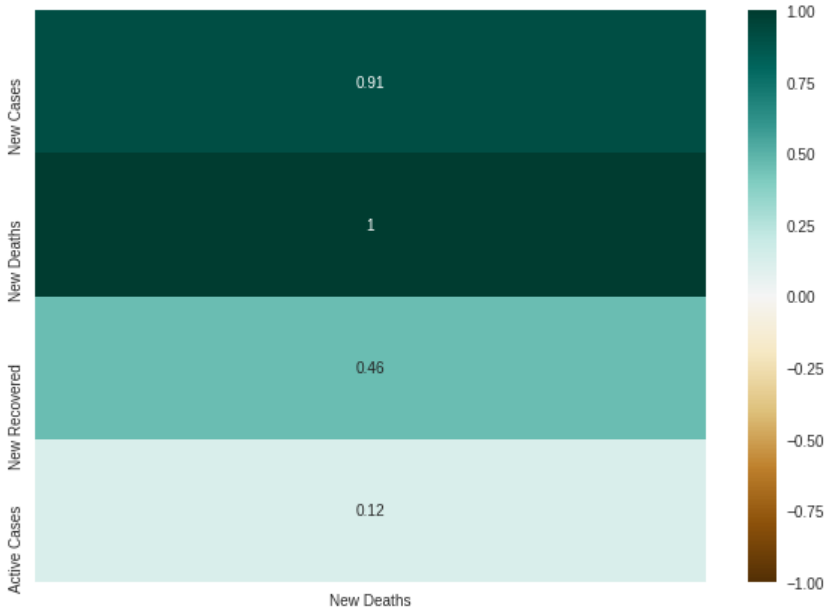
- Nhận xét: Total Recovered tăng qua từng ngày bất kể Active Cases có tăng hay giảm. Điều này có thể chứng tỏ không có mối quan hệ nào giữa 2 trường này. Ta thử kiểm chứng bằng biểu đồ scatter.



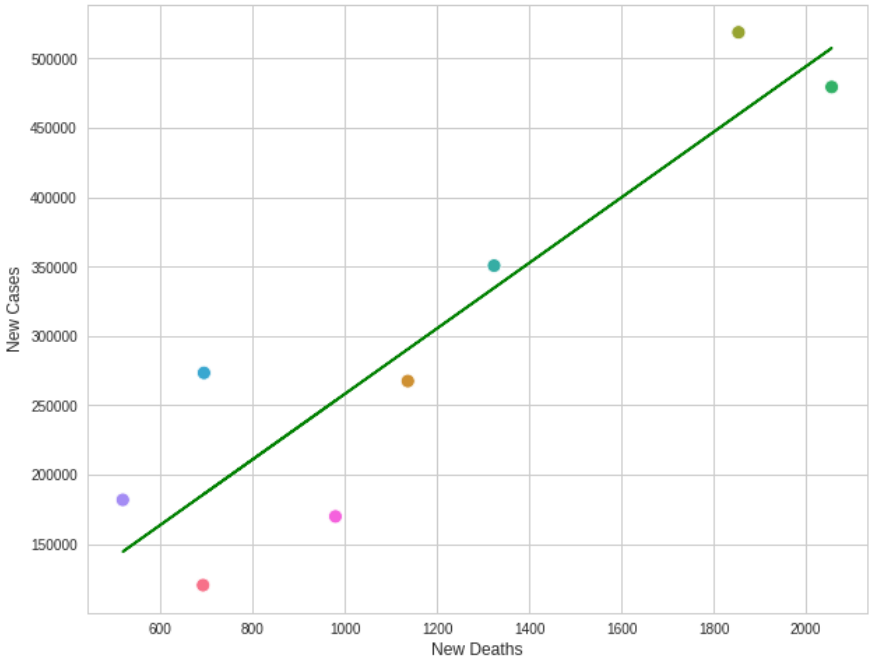
- Nhận xét: 2 trường này có tương quan dương nhưng mối quan hệ này lại rất yếu vì độ tương quan Pearson = 0.136

f) New Deaths & các cột còn lại

- Tìm mối liên hệ giữa New Deaths và các thuộc tính khác có trong data_df. Chúng ta chỉ xét trên châu lục **Europe.**



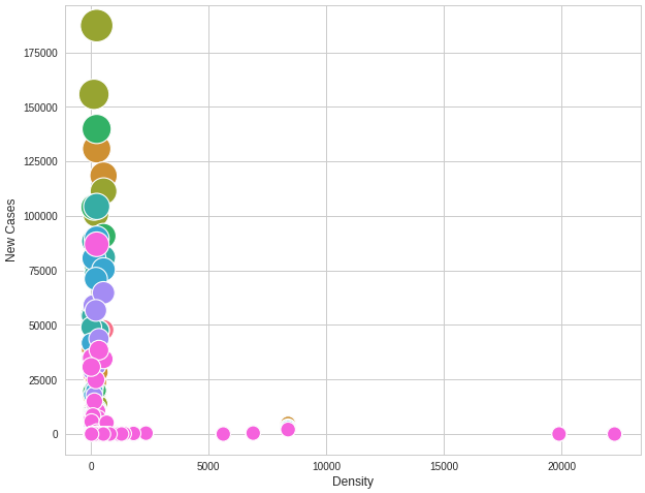
- Sử dụng biểu đồ heatmap vì biểu đồ này có thể giúp dễ dàng phân biệt sự quan trọng của dữ liệu bằng tone màu khác nhau.
- Nhận xét: Ta thấy rằng New Cases và New Deaths có độ tương quan khá là cao. Xem xét dưới góc nhìn scatter plot để kiểm chứng điều này:



- Sử dụng biểu đồ scatter vì nó có thể biểu diễn được quan hệ của 2 trường giá trị.
- Bên cạnh đó sử dụng thể 1 đường Regression để có thể biểu diên rõ hơn xu hướng của quan hệ của 2 trường giá trị.
- Nhận xét: ta có thể thì số ca nhiễm mới càng tăng thì số ca tử vong mới cũng càng tăng, cho thấy giữa số ca nhiễm mới và số ca chết mới có quan hệ tuyến tính với nhau và quan hệ này được thể hiện rõ ràng qua việc đường Regression có thể fit gần khớp với các điểm dữ liệu.

g) New Cases & Density

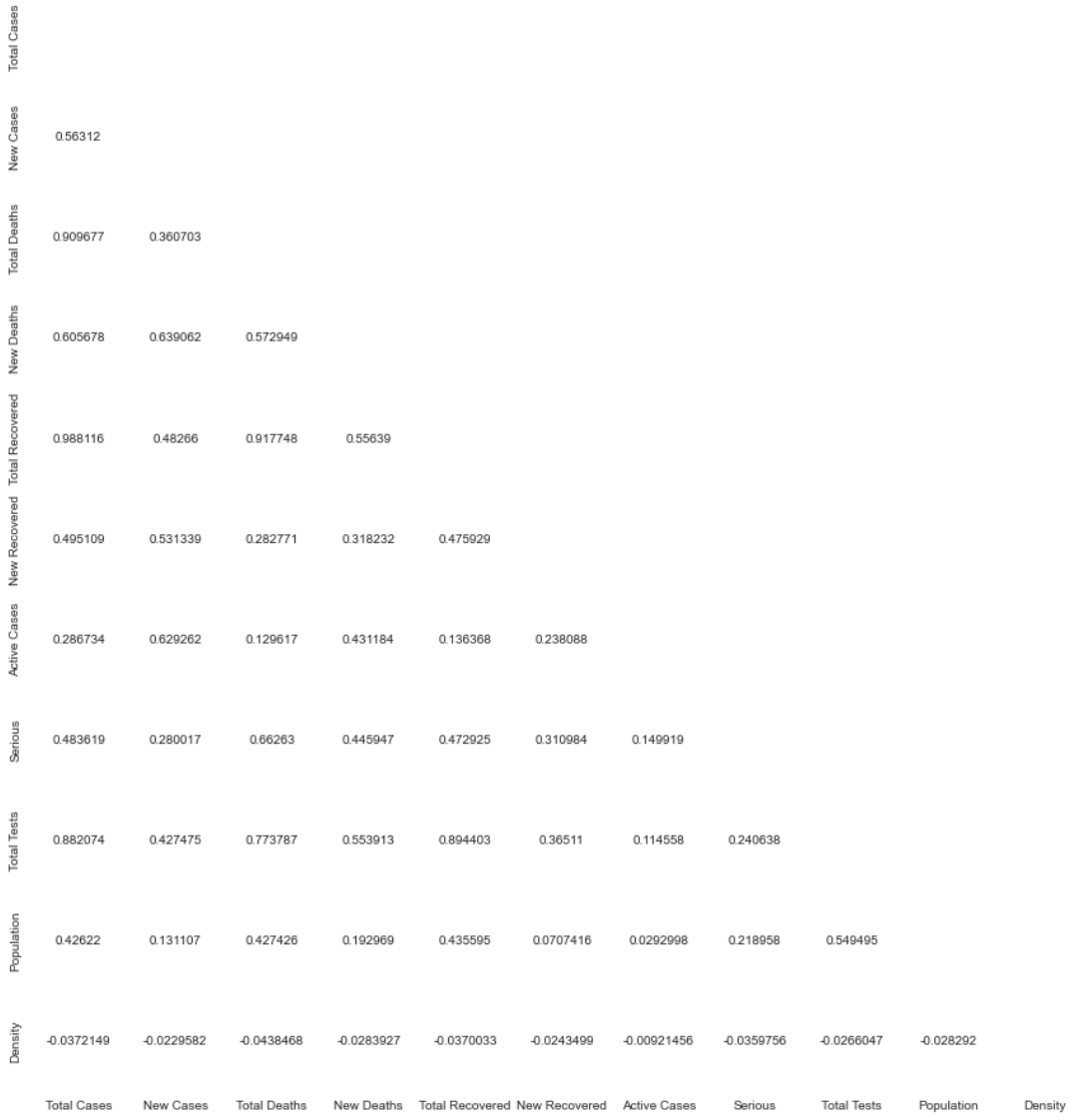
- Tìm mối liên hệ giữa Total Cases và Density. Hypothesis là nếu Mật độ dân số càng cao thì có số ca nhiễm càng nhiều và ngược lại.



- Sử dụng Bubble Chart vì biểu đồ này vừa có thể biểu diễn quên hệ giữa 2 trường dữ liệu và vừa có thể biểu hiện độ lớn của quan hệ đố qua kích cỡ của điểm dữ liệu.
- Nhận xét: Nhìn vào biểu đồ thì chúng ta có thể dễ dàng thấy số ca nhiễm không có mối liên hệ nào với mật độ dân số.

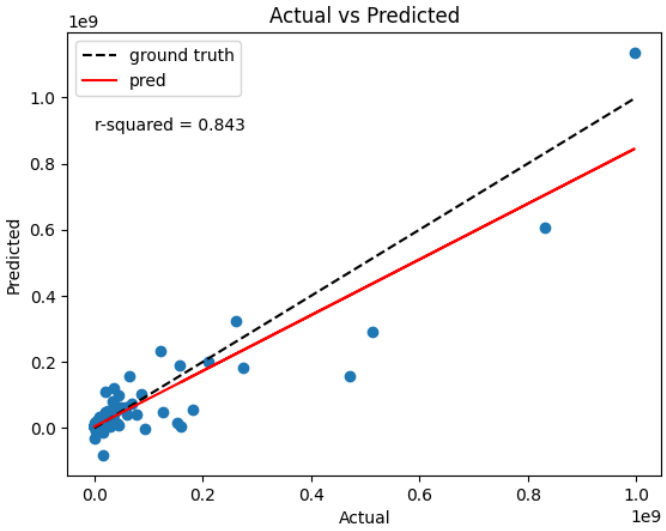
3. Quan hệ nhân quả

- Xét tính tương quan của các cột bằng biểu đồ heatmap, do yêu cầu không dùng màu để hiểu thông tin nên heatmap bên dưới sẽ không có màu mà chỉ chứa giá trị Pearson correlation



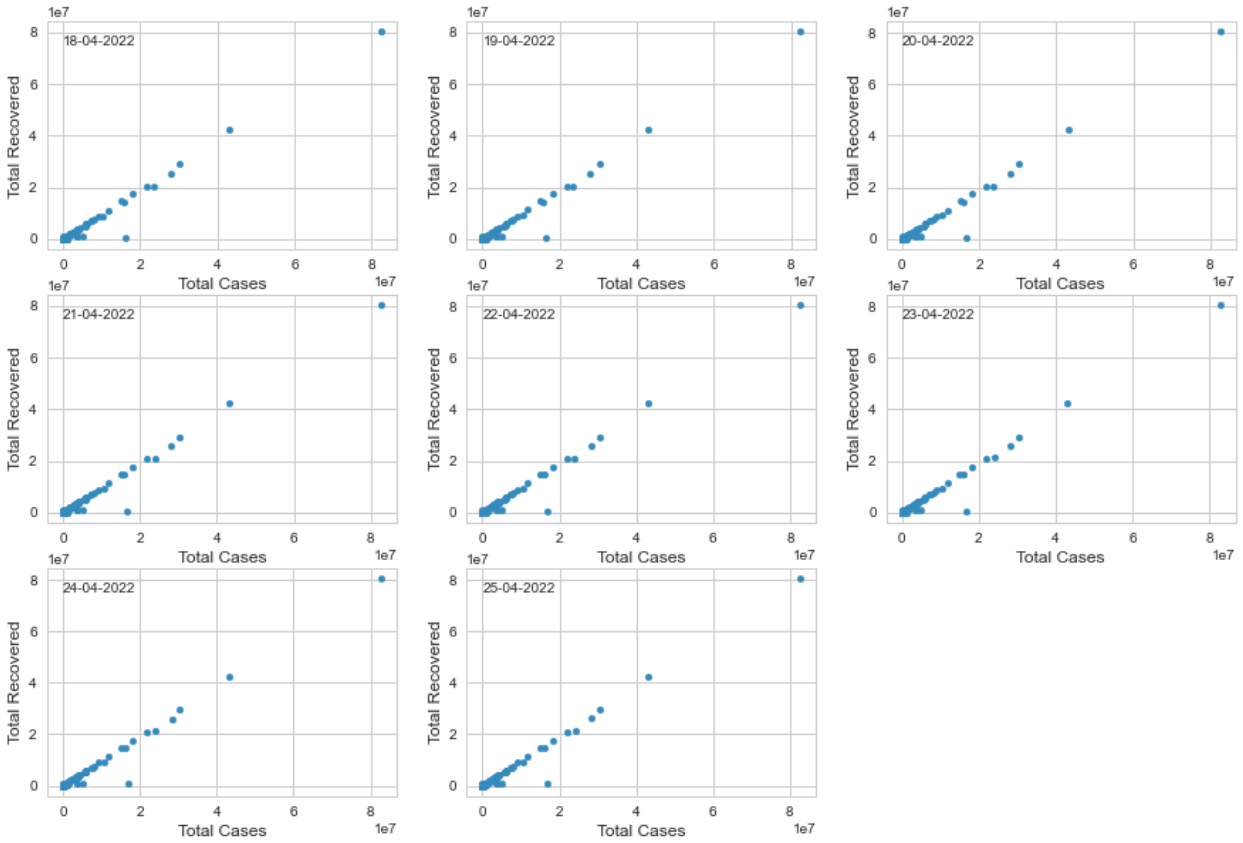
- Nhận xét: có thể thấy gần như các trường dữ liệu đều tỉ lệ thuận với nhau. Trường dữ liệu Total Cases ảnh hưởng rõ ràng nhất đối với các trường dữ liệu khác. Các cặp trường dữ liệu có ảnh hưởng lẫn nhau nổi bật nhất là: (Total Cases, Total Tests), (Total Cases, Total Recovered), (Total Cases, Total Deaths), (Total Deaths, Total Recovered), (Total Recovered, Total Tests)

a) Khảo sát mối quan hệ nhân quả [Total Recovered, Serious] → Total Tests
   - Ta sẽ xây dựng mô hình hồi quy tuyến tính với các biến độc lập là các trường Total Recovered, Serious và biến phụ thuộc là Total Tests
   - Để đơn giản thì thay vì xây dựng mặt phẳng 3D và vẽ mặt phẳng hồi quy tuyến tính thì ta sẽ so sánh giá trị thực tế và giá trị dự đoán trên hệ trục 2D như hình dưới, cũng như vẽ đường nét đứt thể hiện đường đúng (ground truth) và đường nét liền thể hiện kết quả dự đoán mô hình

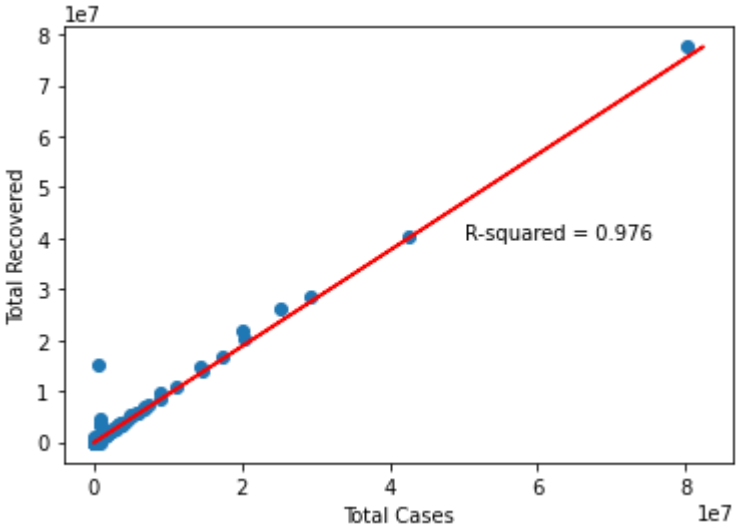


   - Nhận xét: Đường hồi quy có thể xấp xỉ được đường ground truth (nét đứt) khi Actual Total Test < 0.4, khi Total Test >= 0.4 thì đường hồi quy có xu hướng bị ảnh hưởng bởi các điểm ngoại lai. Ngoài ra, giá trị r-squared=0.843, điều này chứng tỏ 2 biến Total Recovered, Serious giải thích được 84.3% dữ liệu. Vậy Total Recovered, Serious là 2 nguyên nhân chính trong nhiều nguyên nhân gây ra ảnh hưởng lên trường Total Tests

b) Khảo sát mối quan hệ nhân quả giữa 2 cột (Total Cases, Total Recovered)
   - Theo nguồn này: "**There is a causal relationship between two variables if a change in the level of one variable causes a change in the other variable**" → Ta sẽ kiểm tra mối quan hệ này trong nhiều ngày liền để coi thử mối tương quan tuyến tính có còn giữ được sự tuyến tính của nó trong nhiều ngày không



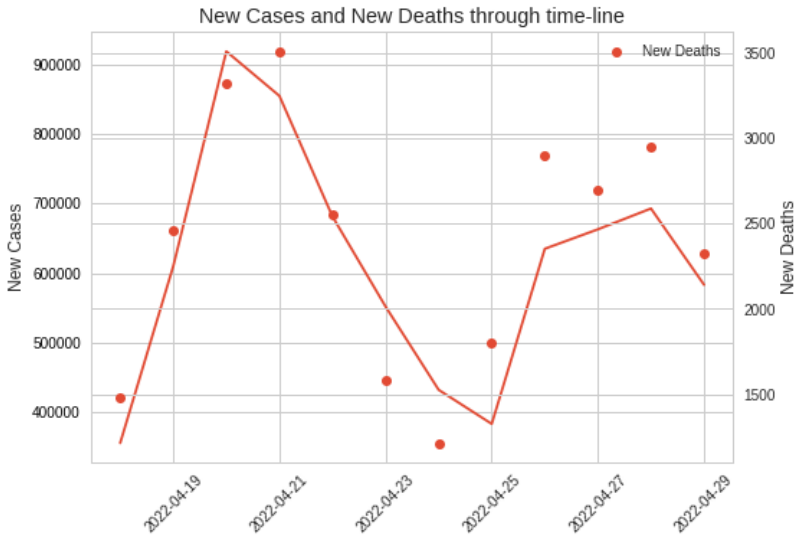
- Nhận xét: qua nhiều ngày, mối tương quan tuyến tính vẫn được giữ nguyên. Tiếp theo, ta muốn biết khi Total Cases tăng 1 đơn vị thì Total Recovered tăng/giảm thế nào → Sử dụng mô hình hồi quy tuyến tính với biến phụ thuộc là Total Recovered và biến độc lập là Total Cases
- Ta sẽ vẽ đường hồi quy lên dữ liệu và tính giá trị r-squared



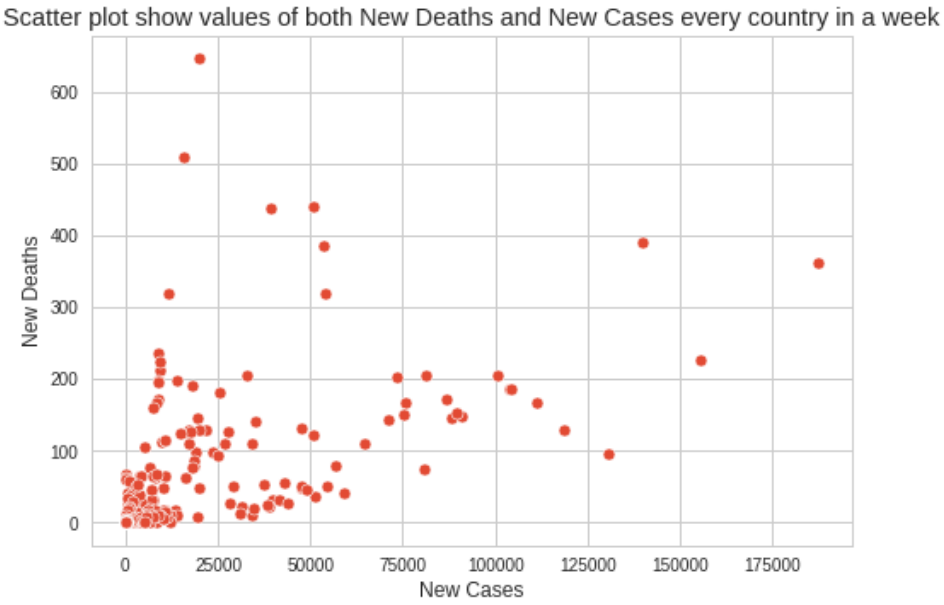
⇨ Mô hình giải thích được 97.6% dữ liệu, chứng tỏ biến Total Cases giải thích được phần lớn xu hướng của biến Total Recovered. Ta có thể tự tin nói rằng Total Cases là nguyên nhân chính gây ra Total Recovered

c) Khảo sát mối quan hệ nhân quả giữa New Deaths & New Cases

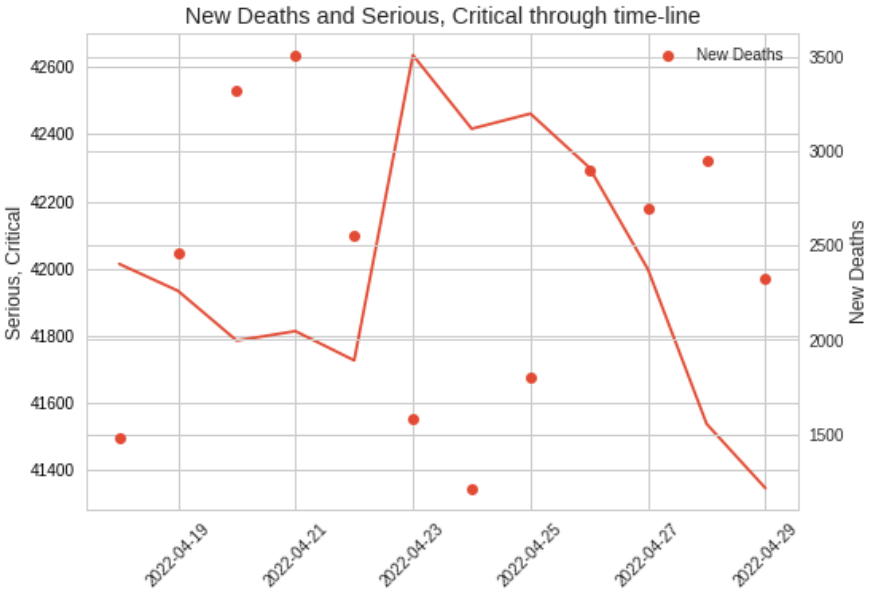
- Đầu tiên ta sẽ xem thử sự thay đổi của New Deaths và New Cases của toàn thế giới qua thời gian. Để trực quan tốt nhất trường hợp này ta sẽ dùng biểu đồ đường đẻ có thể trực quan tốt nhất sự thay đổi của giá trị theo thời gian, với 2 trục tung với 2 bước giá trị khác nhau để có thể dễ so sánh, trong khi đó trục hoành là thời gian của các dữ liệu.



- Nhận xét: Có thể thấy dường như số ca nhiễm mới tăng thì số ca tử vong cũng tăng, còn những ngày số ca nhiễm mới giảm thì số ca tử vong cũng giảm theo, đây là nhìn chung số liệu của thế giới, vậy còn với các nước riêng thì sao, liệu có mối quan hệ nào giữa 2 trường giá trị này rõ ràng không?
- Tiếp theo ta sẽ lấy dữ liệu trong 1 tuần của tất cả các nước trên thế giới, từ đó xem thử có mối quan hệ nào giữa 2 trường dữ liệu. Chọn biểu đồ scatter vì đây là biểu đồ tốt và đơn giản nhất để có thể thể hiện được mối quan hệ giữa 2 trường giá trị. Có thể lấy số liệu của 1 ngày để vẽ nhưng để có thể có cái nhìn rộng hơn, thì ta sẽ lấy số liệu của cả 1 tuần để có thể thể hiện rõ ràng hơn mối quan hệ của 2 trường dữ liệu.



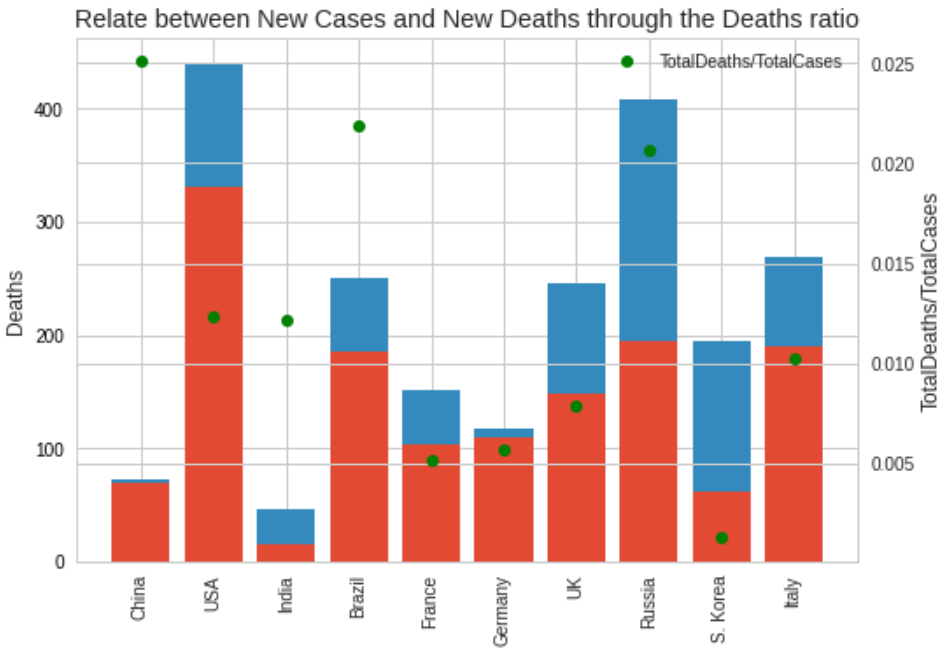
- Nhận xét: Khi nhìn vào tương quan giữa ca nhiễm mới và ca tử vong mới của các nước trong vòng 1 tuần thì có vẻ như là số ca nhiễm mới càng cao thì số ca tử vong mới cũng có xu hướng tăng, nhưng không phải là chính xác hoàn toàn vì có thể thấy có nhiều điểm mà tại đó New Cases thấp nhưng New Deaths vẫn cao, vậy còn yếu tố nào ảnh hưởng đến số ca tử vong mới?
- Ta sẽ xét tiếp Serious, Critical với New Deaths, có thể chúng có mối liên quan nào đó với nhau, chẳng hạn người bệnh ở tình trạng khẩn cấp thì sẽ có tỉ lệ tử vong cao hơn chẳng hạn?
- Tương tự khi so sánh New Cases và New Deaths của thế giới qua thời gian, ta sẽ sử dụng biểu đồ đường với 2 trục tung.



- Nhận xét: Nhìn có vẻ như là New Deaths diễn biến chậm hơn Serious, Critical vài ngày. Có thể đúng bởi vì một người trong giai đoạn nguy hiểm chưa thể tử vong ngay được. Nhưng số liệu chưa đủ để có thể cho cái nhìn chính xác hơn
- Nếu số liệu quá ít thì ta sẽ sử dụng giá trị được tổng hợp từ trước đến nay như Total Deaths và Total Cases để tính tỉ lệ tử vong Total Deaths / Total Cases
- Tiếp đó ta sẽ lấy tỉ lệ trung bình tính được này để tính dự đoán số ca tử vong từ số ca mắc mới:

Predict New Deaths = New Cases * Total Deaths / Total Cases

- Với tỉ lệ tử vong chung, ta hy vọng có thể xác định được trình độ y tế của các nước, từ đó có thể đánh giá mối quan hệ tương quan này theo nhiều khía cạnh hơn.
- Ta dùng biểu đồ cột chồng kết hợp với biểu đồ đường để có thể trực quan được quan hệ này. Sử dụng cột chồng để có thể dễ so sánh số liệu chênh lệch giữa 2 giá trị Predict New Deaths và New Deaths, và thêm biểu đồ dường để có thể thể hiện được tỉ lệ Total Deaths / Total Cases của các nước.



- Nhận xét: Với 10 nước ví dụ, ta có thể thấy được dù sử dụng tỉ lệ Total Deaths / Total Cases trung bình từ trước đến nay nhưng vẫn có sự khác biệt rất lớn giữa số ca tử vong và số ca tử vong dự đoán. Do mỗi giai đoạn dịch bệnh diễn biến khác nhau, chủng loại covid khác nhau, số Vaccine được tiêm khác nhau, thậm chí là số liệu thống kê có thể không còn chính xác nữa. Nên không thể thông qua tỉ lệ tử vong chung mà có thể đoán từ số ca mắc mới thành số ca tử vong được
- Kết luận: New Cases không phải là nguyên nhân chính dẫn đến New Deaths vì khi New Cases thay đổi thì New Deaths có thể không bị thay đổi theo. Mà New Deaths có thê chịu ảnh hưởng từ nhiều yếu tố khác như Serius Cases, chủng loại virus, số vaccine đã tiêm, số người mắc bệnh nền,… nên 2 trường dữ liệu này không có quan hệ nhân quả.

III. THAM KHẢO

- Cào dữ liệu dạng bảng với beautifulsoup: https://www.pluralsight.com/guides/extracting-data-html-beautifulsoup
- Tạo waterfall chart: https://plotly.com/python/waterfall-charts/
- Tạo slope graph: https://towardsdatascience.com/slope-charts-with-pythons-matplotlib-2c3456c137b8
- Tạo bidirectional chart: https://sharkcoder.com/data-visualization/mpl-bidirectional