Đại học Quốc gia TP. HCM
Trường Đại học Khoa học Tự nhiên

BÁO CÁO LAB 01
MỐI QUAN HỆ TRONG DỮ LIỆU

Trực quan hoá dữ liệu (CSC10108)

Nhóm 17

TP Hồ Chí Minh, ngày 29/04/2021

Contents

1 Thông tin nhóm

STT	MSSV	Họ tên	Email	SĐT
1	18120078	Ngô Phù Hữu Đại Sơn	18120078@student.hcmus.edu.vn	0919070940
2	18120201	Nguyễn Bảo Long	18120201@student.hcmus.edu.vn	0981850699
3	18120227	Phạm Văn Minh Phương	18120227@student.hcmus.edu.vn	0981850699
4	18120253	Mai Ngọc Tú	18120253@student.hcmus.edu.vn	0981850699
5	1712424	Hàn Văn Gia Hiên	1712424@student.hcmus.edu.vn	0911572108

Table 1: Bảng danh sách thành viên nhóm

2 Phân tích hoàn thiện yêu cầu

2.1 Tổng quan mức độ hoàn thành mỗi yêu cầu

STT	Yêu cầu	Công việc	Hoàn thành (%)
1	Thu thập dữ liệu	- Cài đặt chương trình thu thập dữ liệu - Tiền xử lý dữ liệu	100/100
2	Trực quan mối quan hệ	- Chọn trường dữ liệu cần trực quan - Trực quan mối quan hệ, giải thích ý nghĩa, nhận xét biểu đồ	90/100

Table 2: Bảng phân tích đóng góp cá nhân

2.2 Mức độ hoàn thành của thành viên nhóm

STT	Họ tên	Công việc tham gia	Hoàn thành (%)
1	Ngô Phù Hữu Đại Sơn	- Thu thập dữ liệu - Hồi quy tuyến tính cho cá quan hệ - Biểu diễn quan hệ 2 biến & 4 biến	100/100
2	Nguyễn Bảo Long	- Biểu diễn quan hệ 3 biến & 4 biến - Giải thích lý do sử dụng biểu đồ đường - Nhận xét dữ liệu 3 biến & 4 biến	100/100
5	Phạm Văn Minh Phương	- Giải thích biểu đồ stacked bar chart - Nhận xét dữ liệu stacked bar chart	100/100
5	Mai Ngọc Tú	- Giải thích biểu đồ scatter - Nhận xét quan hệ dữ liệu 2 biến	100/100
5	Hàn Văn Gia Hiên	- Giải thích biểu đồ Worldmap - Nhận xét dữ liệu biểu đồ Worldmap	100/100

Table 3: Bảng phân tích đóng góp cá nhân

3 Thu thập dữ liệu

3.1 Thu thập số liệu thống kê từng ngày

- Nguồn dữ liệu: `https://www.worldometers.info/coronavirus/`
- Viết chương trình thu thập dữ liệu
  - Ngôn ngữ: Python
  - Framework: Scrapy - framework của python cho phép ta có thể cạo dữ liệu từ web site bằng các lớp đã được định nghĩa sẵn. Scrapy có 5 thành phần:
    1. Spiders: Spiders là lớp đươc định nghĩa sẵn giúp ta có thể lấy dữ liệu liệu dự trên cấu trúc của website. Pipelines: Xử lý các thao tác tiền xử lý dữ liệu như: Làm sạch dữ liệu, xóa các phần tử trùng lặp. Middlewares: Xử lý các request gửi đến website và các response mà ta nhận lại được. Engine: Chịu trách nhiệm điều phối hoạt động của các thành phần khác. Sheduler: Chịu trách nhiệm duy trì thứ tự thực hiện của các công việc
  - Các bước thực hiện
    1. Cài đặt Scrapy
       ```
       pip install Scrapy
       ```
    2. Tạo Scrapy project
       ```
       scrapy startproject corona
       ```
    3. Tạo spider trong Scrapy
       ```
       cd corona
       scrapy genspider covid www.worldometers.info/coronavirus
       ```
    4. Lấy dữ liệu từ xpath của trang HTML sau đó lưu vào Dataframe (Pandas). Cuối cùng, xuất dataframe ra file *.csv. Chi tiết cài đặt xem tại folder **crawler**.
    5. Lặp lại quá trình lấy dữ liệu theo ngày bằng đoạn script sau:
       ```
       cd corona
       now=$(date +%d-%m)
       scrapy crawl covid -o ../../data/"$now".csv
       cd ..
       ```

3.2 Tiền xử lý

- Dữ liệu dạng số sau khi lấy về có chứa dấu "," (ví dụ: 12,000) khiến cho máy tính "hiểu nhầm" là kiểu chuỗi.
  → Xoá dấu "," và các khoảng trắng dư thừa trong dữ liệu
  → Chuyển dữ liệu về dạng số → Điền giá trị 0 vào các ô dữ liệu trống
- Sau khi thực hiện xong các bước trên, xuất dữ liệu "sạch" ra một file *.csv mới.
- Chi tiết xem tại file **preprocess.py**

3.3 Khám phá dữ liệu

- Total Cases: Tổng số các ca nhiễm (gồm các ca nhiễm hiện có, đã tử vong, đã hồi phục).
- Total Deaths: Tổng số các ca nhiễm đã tử vong.
- Total Recovered: Tổng số các ca nhiễm đã hồi phục.
- Active Cases: Các ca nhiễm đang được điều trị.
- Total Tests: Tổng số lần xét nghiệm được thực hiện.
- Population: Dân số của một quốc gia.

4 Trực quan hoá mối quan hệ giữa các trường dữ liệu

4.1 Chọn trường dữ liệu

- Với mỗi biến X, Y của tập dữ liệu, ta tính hệ số tương quan

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}} = \frac{Cov(X,Y)}{\sigma(X) \times \sigma(Y)} \tag{1}$$

  Với $Cov(X,Y) = E[XY] - E[X]E[Y]$ là hiệp phương sai của 2 biến X, Y.

- Từ đó, tính được ma trận hệ số tương quan giữa các cặp biến X, Y trong tập dữ liệu. Biểu diễn ma trận này bằng biểu đồ heatmap để xác định các trường dữ liệu trực quan.

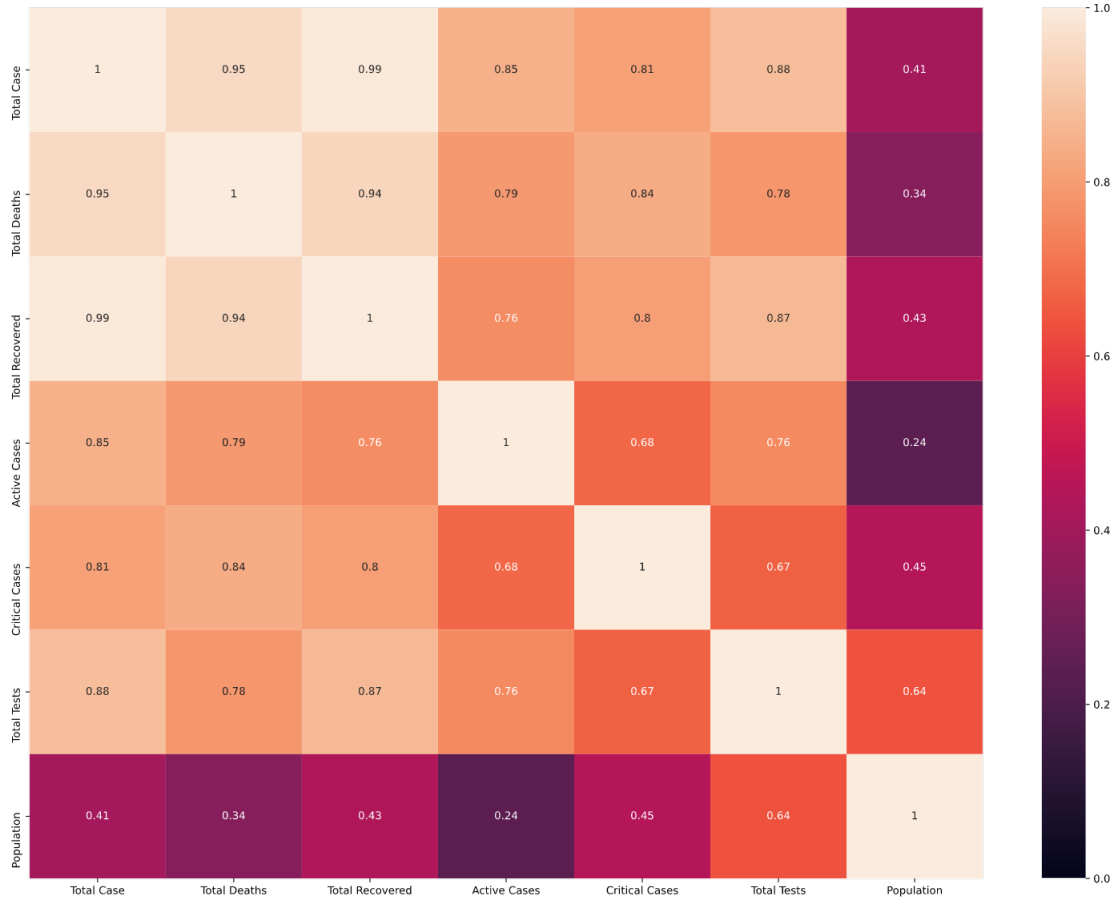


Figure 1: Biểu đồ thể hiện hệ số tương quan giữa các trường dữ liệu

- Từ biểu đồ trên, chọn ra các cặp thuộc tính có hệ số tương quan lớn hơn 0.7, ta được bảng các cặp thuộc tính sau

STT	Var1	Var2	Hệ số tương quan
1	Total Deaths	Total Cases	0.95
2	Total Recovered	Total Cases	0.99
3	Total Recovered	Total Deaths	0.94
4	Active Cases	Total Cases	0.85
5	Active Cases	Total Deaths	0.79
6	Active Cases	Total Recovered	0.76
7	Critical Cases	Total Cases	0.81
8	Critical Cases	Total Deaths	0.84
9	Critical Cases	Total Recovered	0.8
10	Total Test	Total Cases	0.88
11	Total Test	Total Deaths	0.78
12	Total Test	Total Recovered	0.87
13	Total Test	Active Cases	0.76

Table 4: Bảng các thuộc tính có hệ số tương quan trên 0.7

- Từ bảng trên, ta thấy có 6 thuộc tính có mức độ tương quan với nhau cao là **Total Case, Total Deaths, Total Recovered, Active Cases, Critical Cases, Total Tests**. Do đó, ta tập trung vào tìm hiểu mối quan hệ giữa các thuộc tính này.

4.2 Biểu diễn quan hệ giữa 2 trường dữ liệu

- Với mỗi 2 trường dữ liệu X, Y trong danh sách các trường dữ liệu xuất hiện trong Table 4, trực quan các điểm dữ liệu thuộc 2 trường đó bằng biểu đồ scatter.
- Tiếp đó, tìm một quy luật xấp xỉ phù hợp nhất với dữ liệu quan sát bằng một đường thẳng $\bar{y} = w_0 + xw_1$ với w_0, w_1 được tính theo công thức sau

$$\begin{cases} w_1 = \frac{\overline{XY} - \overline{X}.\overline{Y}}{\overline{X^2} - \overline{X}^2} \\ w_0 = \overline{Y} - w_1\overline{X} \end{cases} \tag{2}$$

- Lý do sử dụng biểu đồ scatter kết hợp biểu diễn quy luật bằng một đường tuyến tính: Sử dụng biểu đồ scatter để trực quan giúp người đọc dễ dàng nhận ra một sự xấp xỉ tuyến tính (nếu có) tồn tại giữa 2 biến dữ liệu. Cộng thêm việc biểu diễn quy luật bằng một đường tuyến tính, người đọc có thể thấy rõ được "chất lượng" của đường tuyến tính này trong việc giải thích mối quan hệ giữa 2 trường dữ liệu.
- Kết quả thu được $C_6^2 = 15$ biểu đồ như sau

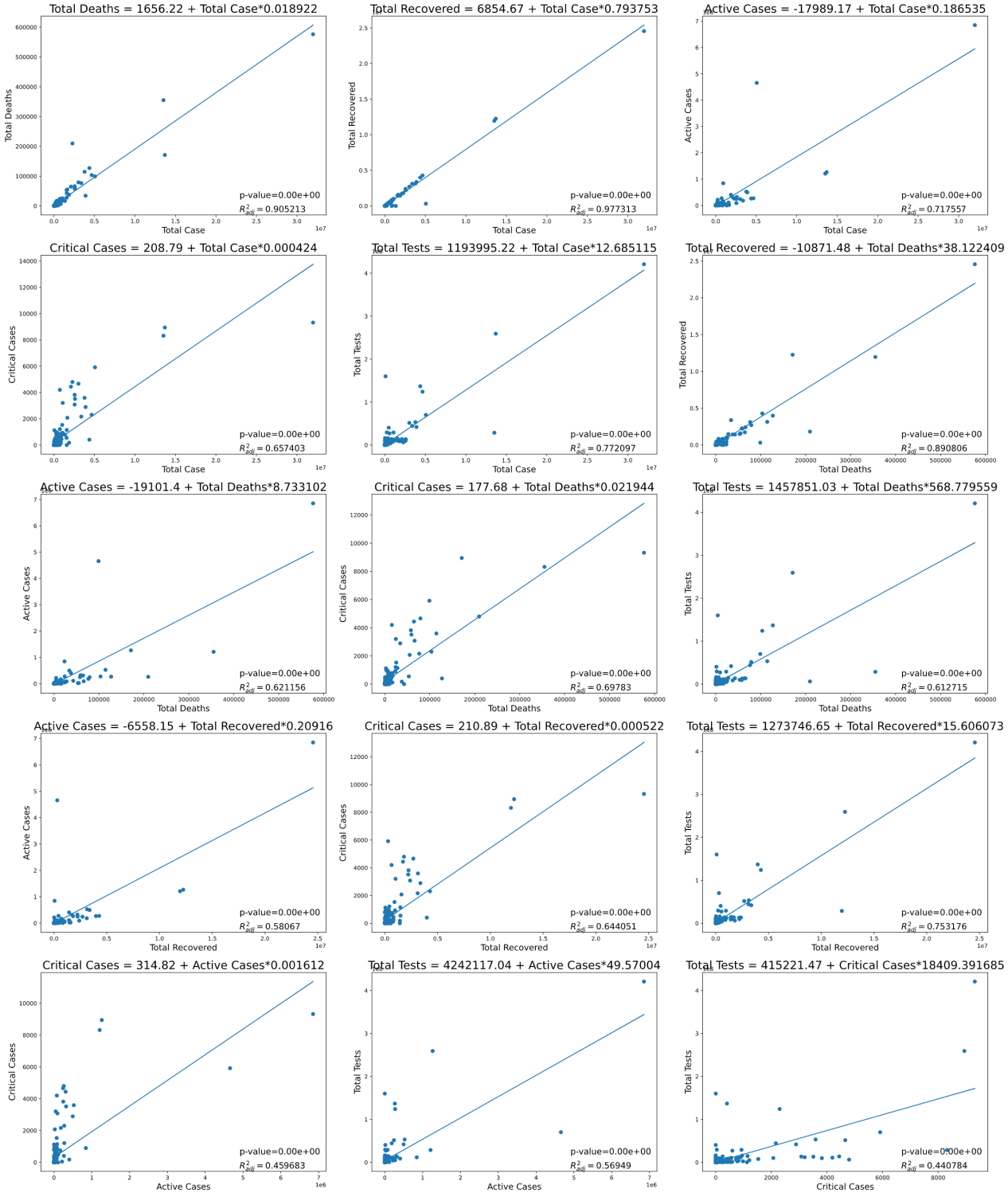


Figure 2: Biểu đồ thể hiện mỗi quan hệ giữa 2 biến dữ liệu

- Nhận xét dữ liệu
  - Với 5 cặp biến (Độc lập - Phụ thuộc): **(Total Case, Total Deaths), (Total Case, Total Recovered), (Total Case, Total Tests), (Total Deaths, Total Recovered), (Total Recovered, Total Tests)**, ta có các kết luận sau
    - Chỉ số p-value của các cặp biến này rất nhỏ (trong khoảng $[10^{-180}, 10^{-68}]$). Do đó, các biến độc lập nêu trên có ý nghĩa về mặt thống kê.

  - * Mô hình phù hợp tốt với dữ liệu quan sát về mặt thống kê (p-value trong khoảng $[10^{-180}, 10^{-68}]$).
  - * Hệ số xác định hiệu chỉnh R^2_{adj} của mỗi bộ dữ liệu đều nằm trong khoảng $[75\%, 98\%]$. Do đó, các biến độc lập giải thích được từ 75% đến 98% sự thay đổi của các biến phụ thuộc.
  - * Cụ thể phương trình hồi của các cặp biến được ghi trên biểu đồ.
- Với các cặp biến còn lại, chỉ số $R^2_{adj} < 75\%$. Do đó, việc giải thích sự thay đổi của biến phụ thuộc dựa trên biến độc lập là không đủ tin cậy.

4.3 Biểu diễn quan hệ giữa 3 trường dữ liệu

- Với mỗi 3 trường dữ liệu X, Y, Z trong danh sách các trường dữ liệu xuất hiện trong Table 4, sử dụng kỹ thuật hồi quy tuyến tính để xác định một mặt phẳng xấp xỉ các điểm dữ liệu:

$$w = (X^T X)^{-1} X^T y, \text{ với X là ma trận các biến độc lập, y là biến phụ thuộc} \quad (3)$$

- Sau đó tính hệ số xác định hiệu chỉnh R^2_{adj}, ta xác định được các bộ 3 biến có $R^2_{adj} > 70\%$ như sau

STT	**Var1 (độc lập)**	**Var2 (độc lập)**	**Var3 (phụ thuộc)**	R^2_{adj}
1	Total Case	Total Deaths	Total Recovered	0.97
2	Total Case	Total Deaths	Active Cases	0.71
3	Total Case	Total Deaths	Total Tests	0.80
4	Total Case	Total Recovered	Active Cases	0.96
5	Total Case	Total Recovered	Total Tests	0.77
6	Total Case	Active Cases	Total Tests	0.77
7	Total Case	Critical Cases	Total Tests	0.77
8	Total Deaths	Total Recovered	Total Tests	0.76
8	Total Recovered	Active Cases	Total Tests	0.77
10	Total Recovered	Critical Cases	Total Tests	0.75

Table 5: Bảng các bộ 3 thuộc tính có mối tương quan trên 0.7

- Sử dụng biểu đồ đường để biểu diễn mối quan hệ giữa 3 biến dữ liệu như sau: Mỗi thuộc tính có một đường biểu diễn riêng. Giá trị này ứng với giá trị trên trục tung. Trục hoành là tên các quốc gia tương ứng.
- Lý do sử dụng biểu đồ đường
  - Trước hết, chúng ta cần xác định mối quan hệ giữa 3 trường dữ liệu (bao gồm 2 trường độc lập và 1 trường phụ thuộc). Bằng kỹ thuật hồi quy tuyến tính (áp dụng phương trình (3)), giả sử ta tìm ra được hàm số sau:

    $$TotalRecovered = f(TotalCases, TotalDeaths) = 4556 + 0.8(TotalCases) + 1.4(TotalDeaths)$$

  - Lấy đạo hàm riêng của Total Recovered theo các biến độc lập, ta được
    - * $\frac{\partial \text{ Total Recovered}}{\partial \text{ Total Cases}} = 0.8 > 0$: Total Cases và Total Recovered biến thiên đồng biến. Nghĩa là khi Total Cases tăng thì Total Recovered cũng tăng theo (xét trong trường hợp Total Deaths là hằng số hoặc có biến động nhưng biến động này rất nhỏ so với các biến còn lại).

* $\frac{\partial \text{ Total Recovered}}{\partial \text{ Total Deaths}} = 1.4 > 0$: Total Deaths và Total Recovered biến thiên đồng biến. Nghĩa là khi Total Deaths tăng thì Total Recovered cũng tăng theo (xét trong trường hợp Total Cases là hằng số hoặc có biến động nhưng biến động này rất nhỏ so với các biến còn lại).
* Ngoài ra, $(\frac{\partial \text{ Total Recovered}}{\partial \text{ Total Deaths}} = 1.4) > (\frac{\partial \text{ Total Recovered}}{\partial \text{ Total Cases}} = 0.8)$: Total Recovered phụ thuộc vào Total Deaths nhiều hơn Total Cases.

– Đối chiếu vào biểu đồ: Quan sát các tổng thể biểu đồ, ta thấy

* Total Deaths biến động rất nhẹ so với 2 biến còn lại. Do đó, sự thay đổi của biến này không ảnh hưởng nhiều đến Total Recovered.
* Total Cases biến động rất mạnh, và khi nó giảm/tăng thì Total Recovered (biến phụ thuộc) cũng giảm/tăng theo.
* Vậy ta đã biểu diễn được quan hệ biến thiên đồng biến giữa các biến Total Cases và Total Deaths với biến Total Recovered

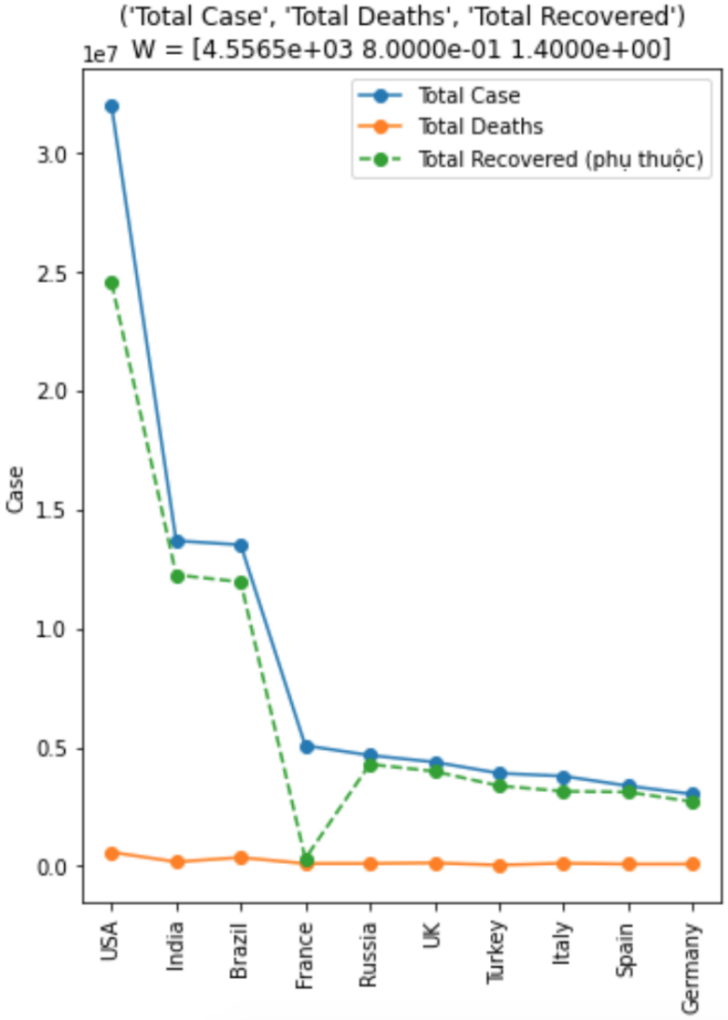
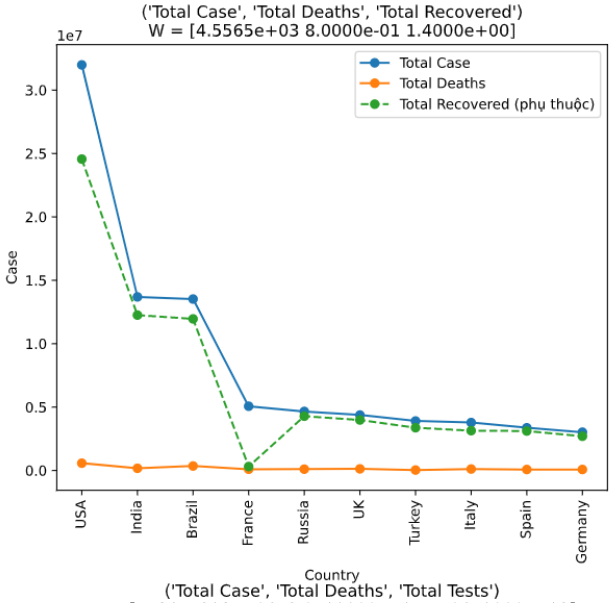
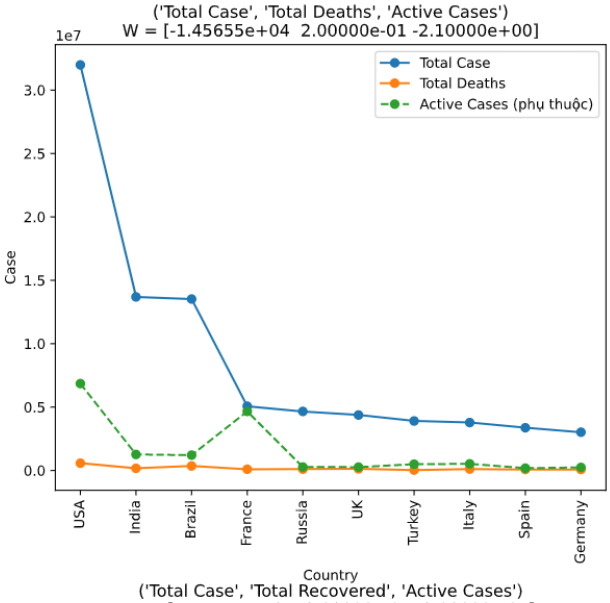


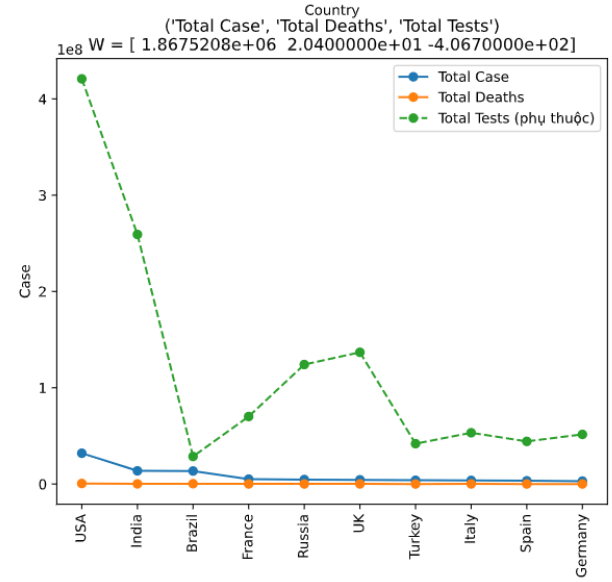
Figure 3: Biểu đồ biểu diễn sự mối quan hệ của Total Cases, Total Deaths và Total Recovered

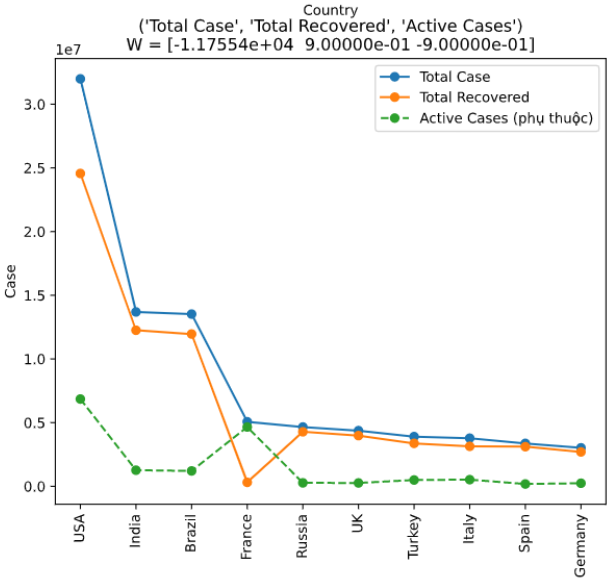
– Tóm lại, việc sử dụng biểu đồ đường để thể hiện mối quan hệ tuyến tính giữa các biến dữ liệu được nhóm cho là hợp lý vì đã biểu diễn được sự biến thiên phụ thuộc giữa các biến phụ thuộc và biến độc lập.

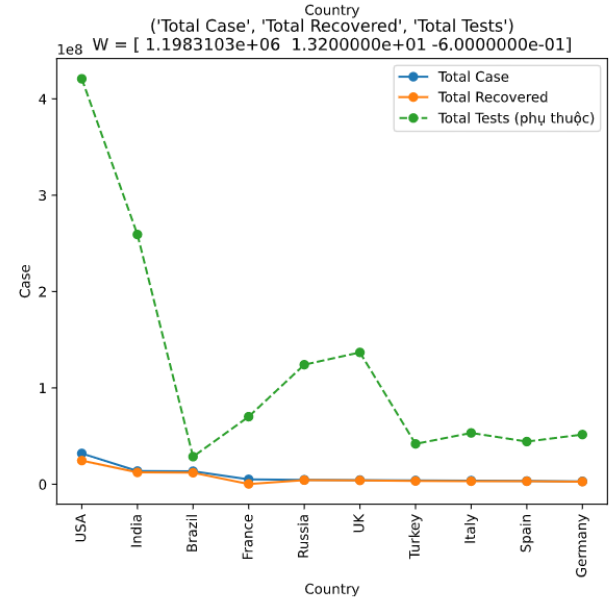
- Trực quan các bộ 3 biến

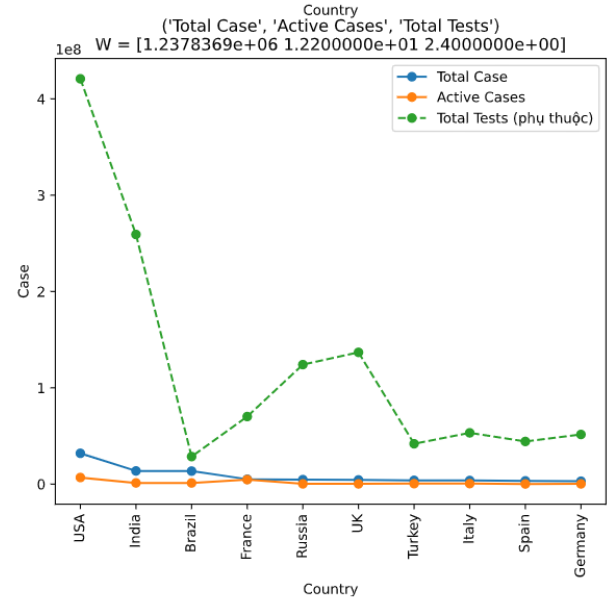

('Total Case', 'Total Deaths', 'Total Recovered')
W = [4.5565e+03 8.0000e-01 1.4000e+00]
1e7
Total Case
Total Deaths
Total Recovered (phụ thuộc)
Case
Country
USA
India
Brazil
France
Russia
UK
Turkey
Italy
Spain
Germany


('Total Case', 'Total Deaths', 'Active Cases')
W = [-1.45655e+04 2.00000e-01 -2.10000e+00]
1e7
Total Case
Total Deaths
Active Cases (phụ thuộc)
Case
Country
USA
India
Brazil
France
Russia
UK
Turkey
Italy
Spain
Germany


('Total Case', 'Total Deaths', 'Total Tests')
W = [1.8675208e+06 2.0400000e+01 -4.0670000e+02]
1e8
Total Case
Total Deaths
Total Tests (phụ thuộc)
Case
Country
USA
India
Brazil
France
Russia
UK
Turkey
Italy
Spain
Germany


('Total Case', 'Total Recovered', 'Active Cases')
W = [-1.17554e+04 9.00000e-01 -9.00000e-01]
1e7
Total Case
Total Recovered
Active Cases (phụ thuộc)
Case
Country
USA
India
Brazil
France
Russia
UK
Turkey
Italy
Spain
Germany


('Total Case', 'Total Recovered', 'Total Tests')
W = [1.1983103e+06 1.3200000e+01 -6.0000000e-01]
1e8
Total Case
Total Recovered
Total Tests (phụ thuộc)
Case
Country
USA
India
Brazil
France
Russia
UK
Turkey
Italy
Spain
Germany


('Total Case', 'Active Cases', 'Total Tests')
W = [1.2378369e+06 1.2200000e+01 2.4000000e+00]
1e8
Total Case
Active Cases
Total Tests (phụ thuộc)
Case
Country
USA
India
Brazil
France
Russia
UK
Turkey
Italy
Spain
Germany

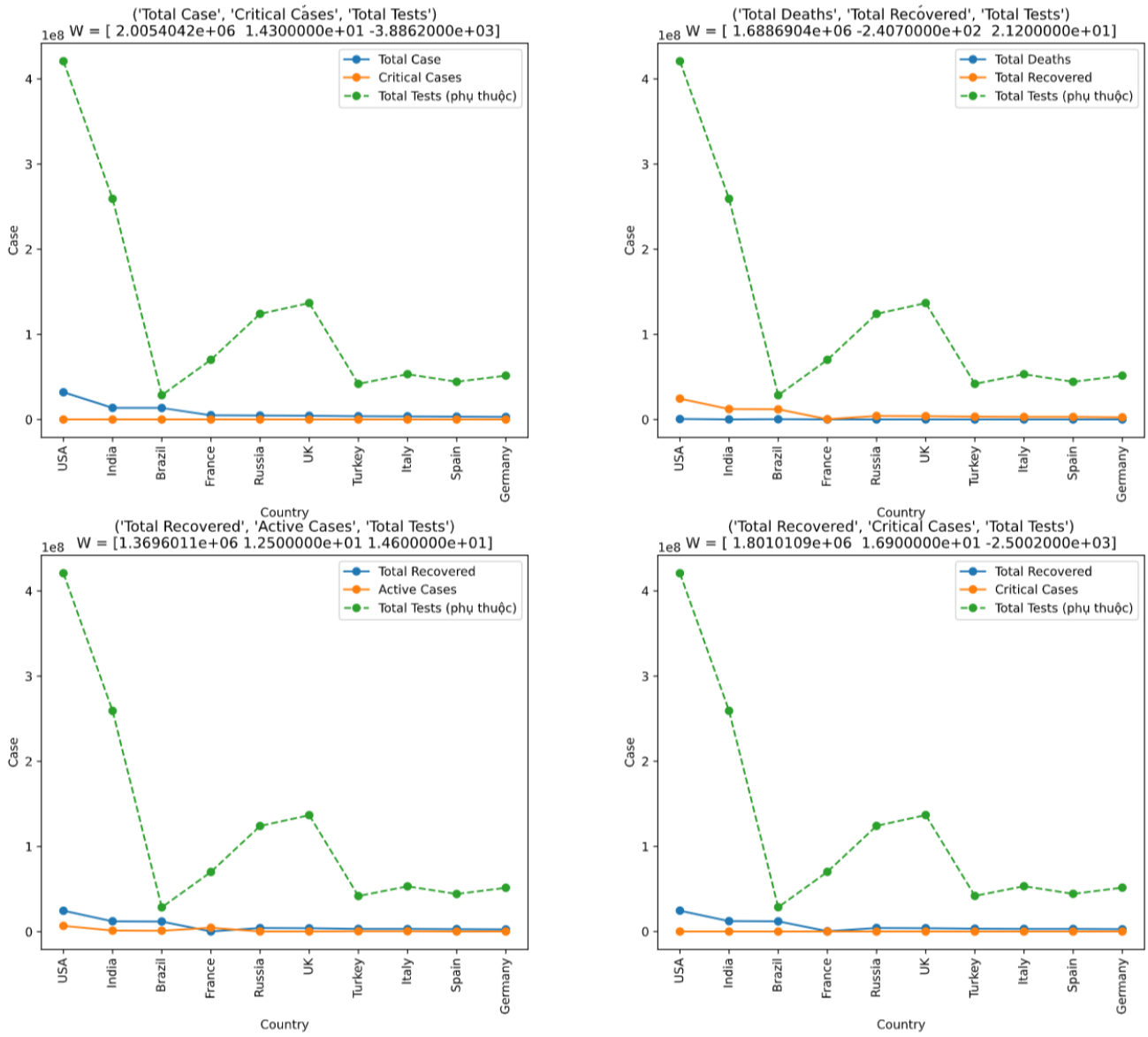


Figure 4: Biểu đồ biểu diễn mỗi quan hệ giữa 3 biến dữ liệu

- Nhận xét dữ liệu: Từ Table 5, ta có thể kết luận rằng Var1 và Var2 có thể giải thích được từ 75% đến 97% sự thay đổi của Var3.

4.4 Biểu diễn quan hệ giữa 4 trường dữ liệu

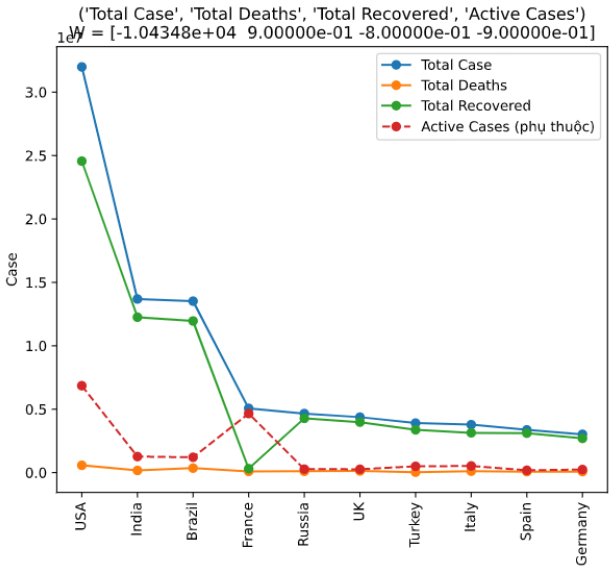
- Tương tự như biểu diễn quan hệ giữa 3 biến dữ liệu, việc sử dụng biểu đồ đường cho biểu diễn quan hệ 4 biến cũng có thể trực quan được mỗi quan hệ này một cách rõ ràng với cơ sở toán học như đã trình bày ở phần *4.3. Biểu diễn quan hệ giữa 3 trường dữ liệu*

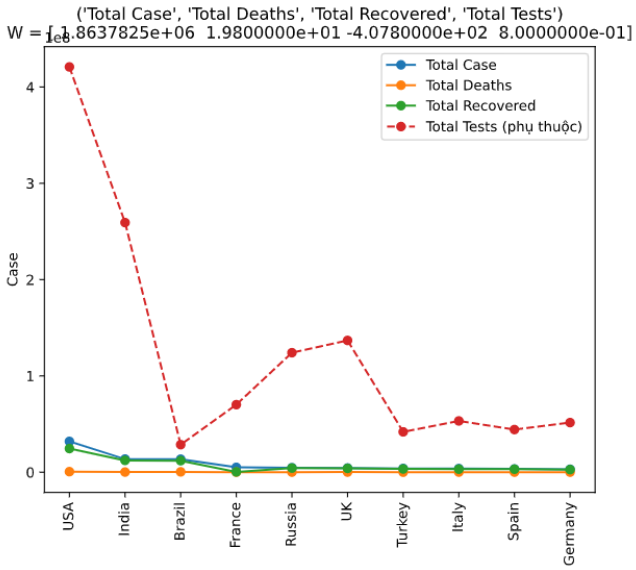
- Các bộ 4 biến được xét tại đây thoả mãn điều kiện $R^2_{adj} > 0.7$. Ta có bảng thống kê sau

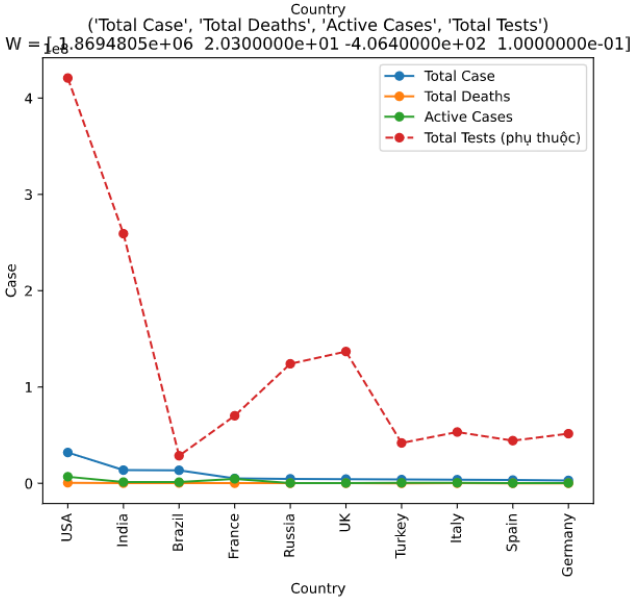
STT	Var1 (độc lập)	Var2 (độc lập)	Var3 (độc lập)	Var4 (phụ thuộc)	R^2_{adj}
1	Total Case	Total Deaths	Total Recovered	Active Cases	0.96
2	Total Case	Total Deaths	Total Recovered	Total Tests	0.80
3	Total Case	Total Deaths	Active Cases	Total Tests	0.80
4	Total Case	Total Deaths	Critical Cases	Total Tests	0.80
5	Total Case	Total Recovered	Active Cases	Total Tests	0.77
6	Total Case	Total Recovered	Critical Cases	Total Tests	0.77
7	Total Case	Active Cases	Critical Cases	Total Tests	0.77
8	Total Deaths	Total Recovered	Active Cases	Total Tests	0.79
8	Total Deaths	Total Recovered	Critical Cases	Total Tests	0.76
10	Total Recovered	Active Cases	Critical Cases	Total Tests	0.77

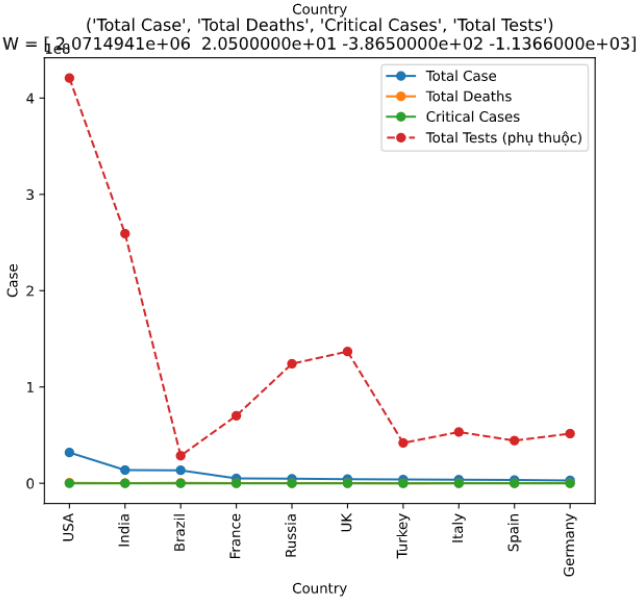
Table 6: Bảng các bộ 4 thuộc tính có mối tương quan trên 0.7

- Trực quan các bộ 4 biến









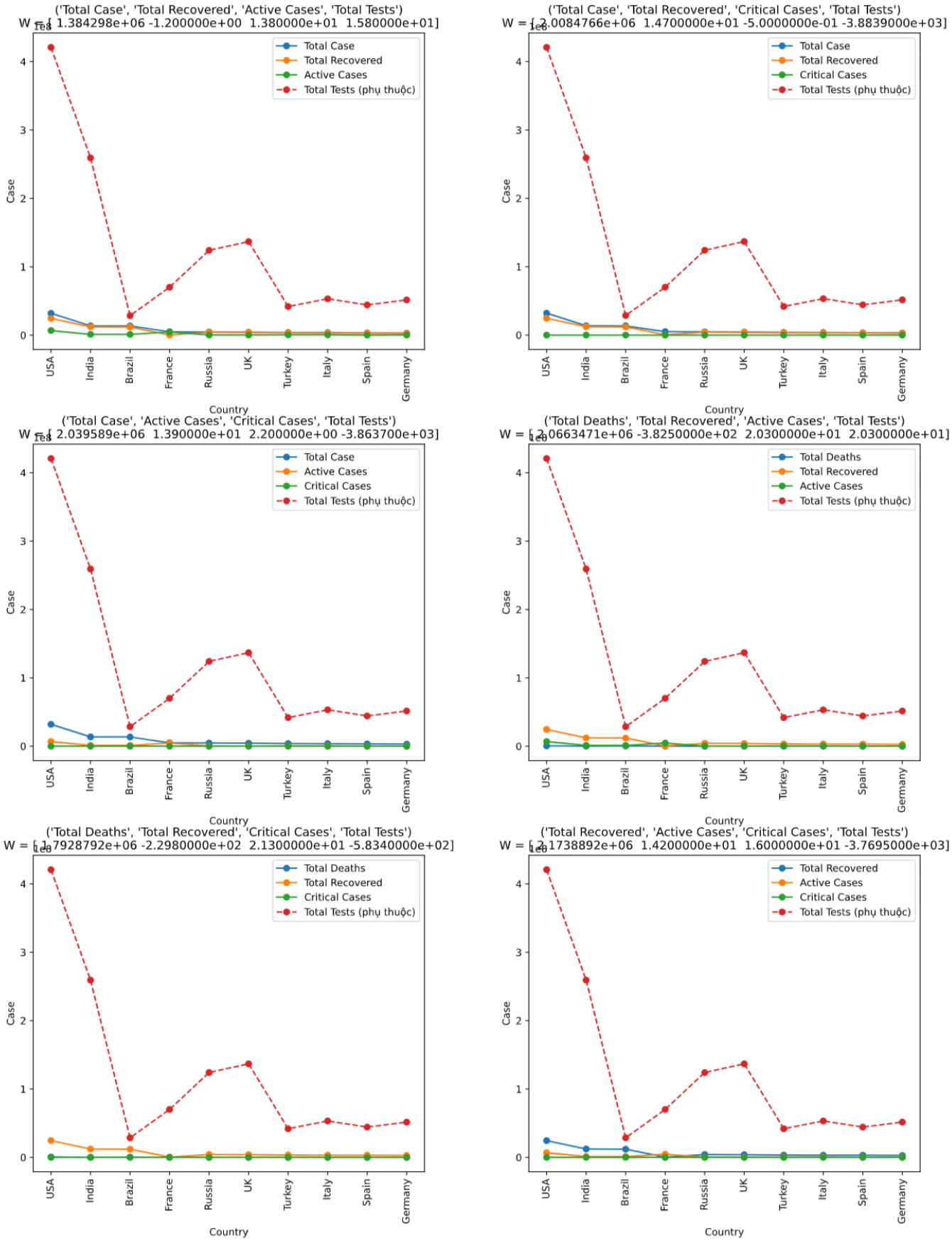


Figure 5: Biểu đồ biểu diễn mối quan hệ giữa 4 biến dữ liệu

- Nhận xét dữ liệu: Từ Table 5, ta có thể kết luận rằng Var1, Var2 và Var3 có thể giải thích được từ 76% đến 96% sự thay đổi của Var4.

5 Một vài biểu diễn khác

5.1 Stacked bar chart

- Stacked bar chart dùng để biểu diễn một quan hệ 4 biến có sẵn (do `https://www.worldometers.info/` cung cấp)

$$Totalcases = TotalDeaths + TotalRecovered + ActiveCases \tag{4}$$

- Trực quan

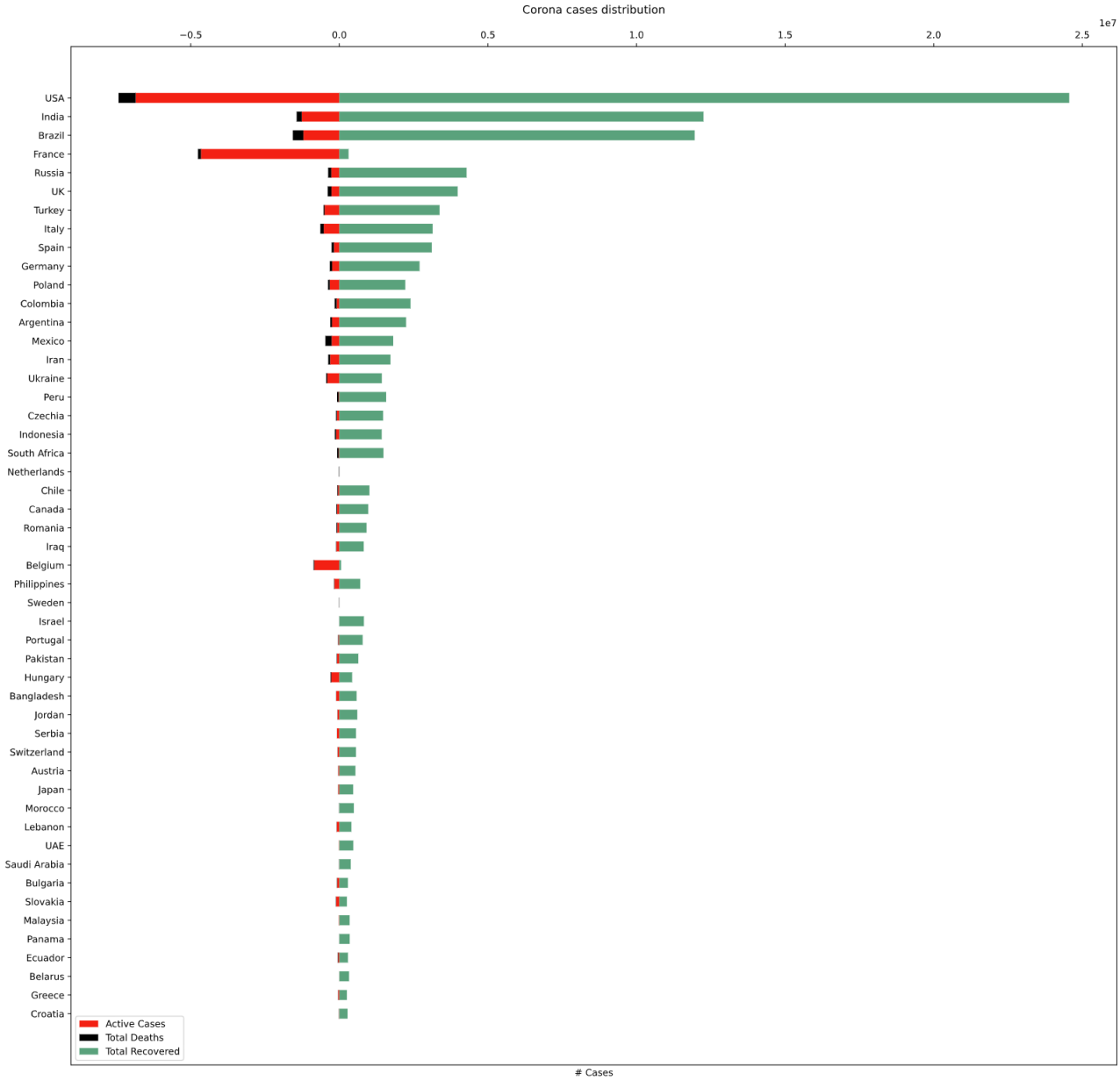


Figure 6: Biểu đồ thể hiện quan hệ theo phương trình (4)

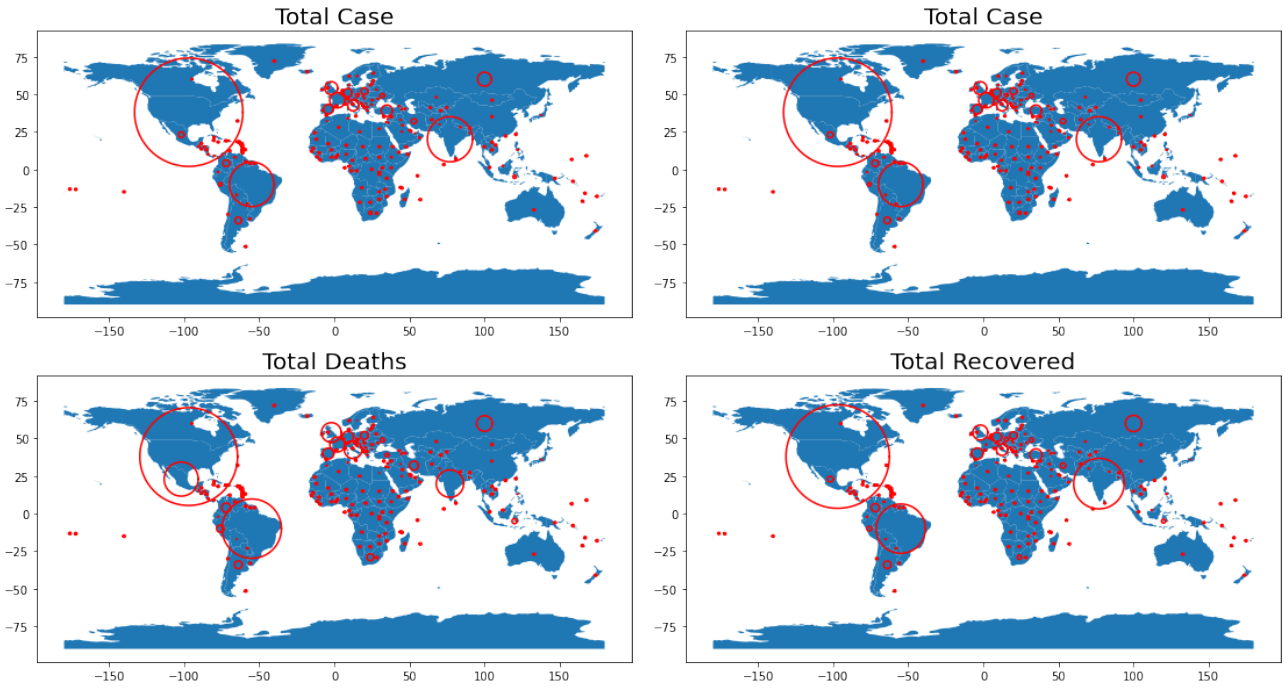
- Giải thích biểu đồ: Với mỗi nước được trực quan trên biểu đồ, chiều dài của cả cột chính bằng Total Cases của nước đó. Trong mỗi cột lại phân chia ra nhiều cột nhỏ như đã chú thích trong biểu đồ. Phần Total Deaths và Active Cases được vẽ hướng về phía âm - biểu thị số ca đã chết hoặc đang điều trị. Phần Total Recovered được tô màu xanh và hướng

về phía dương - biểu thị số ca đã phục hồi. Ở đây, các nước được sắp xếp theo thứ tự giảm dần Total Cases.

- Lý do sử dụng stacked bar chart: Biểu đồ cột chồng rất có ích trong việc so sánh, xếp hạng dữ liệu tổng cũng như cho ta biết tỉ lệ của các thành phần tạo nên tổng. Trong trường hợp này, nó trực quan được rất rõ tỷ lệ hồi phục và tỷ lệ chết hoặc đang điều trị, thích hợp cho việc so sánh tổng số ca mắc giữa các quốc gia với nhau.

5.2 World map

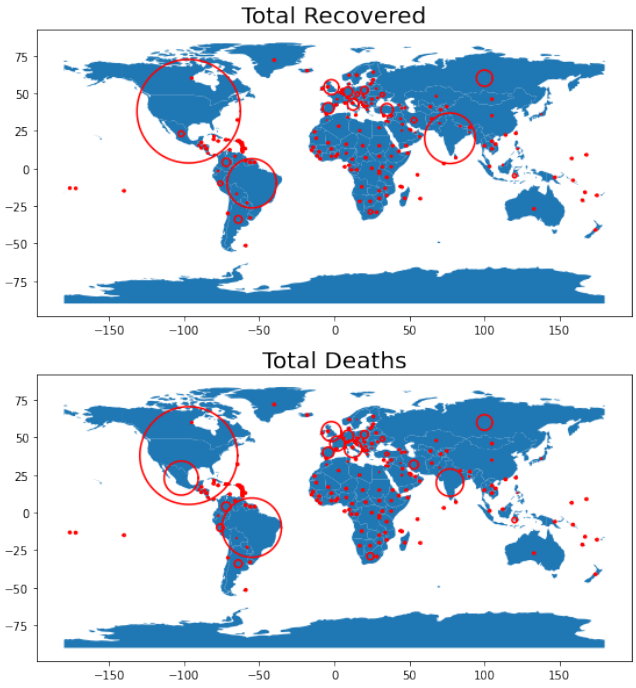
- Nếu việc trực quan bằng biểu đồ cột chồng là chưa đủ rõ ràng trong việc so sánh số liệu giữa các quốc gia cũng như đối chứng quan hệ giữa các biến dữ liệu thì biểu đồ world map như sắp trình bày dưới đây sẽ phần nào mang lại cảm giác trực quan rõ ràng hơn.
- Việc sử dụng world map cho ta thấy được cái nhìn tổng quan của mỗi biến dữ liệu giữa các quốc gia và thấy rõ được sự khác biệt giữa các quốc gia. Đồng thời cho ta dễ dàng thấy được những nét tương đồng giữa hai biến dữ liệu giữa các quốc gia, từ đó có thể nhận định được có sự tương quan xảy ra giữa hai biến dữ liệu đang xét không.
- Trực quan
  - Tương quan giữa Total Cases với Total Deaths và Total Recovered



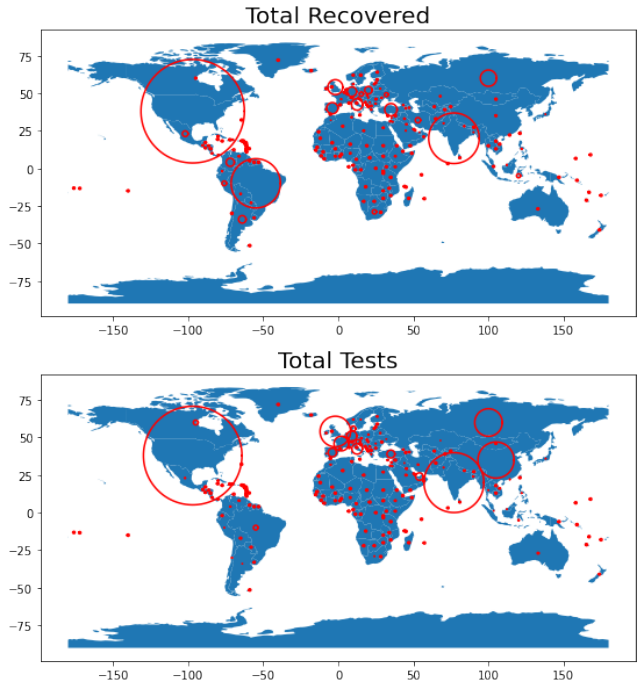
(a) Total cases - Total Deaths (b) Total cases - Total Recovered

Figure 7: Tương quan giữa Total Cases với Total Deaths và Total Recovered

  - Tương quan giữa Total Recovered với Total Deaths và Total Test



(a) Total Recovered - Total Deaths



(b) Total Recovered - Total Test

Figure 8: Tương quan giữa Total Recovered với Total Deaths và Total Test

– Tương quan giữa Total Cases với Total Test

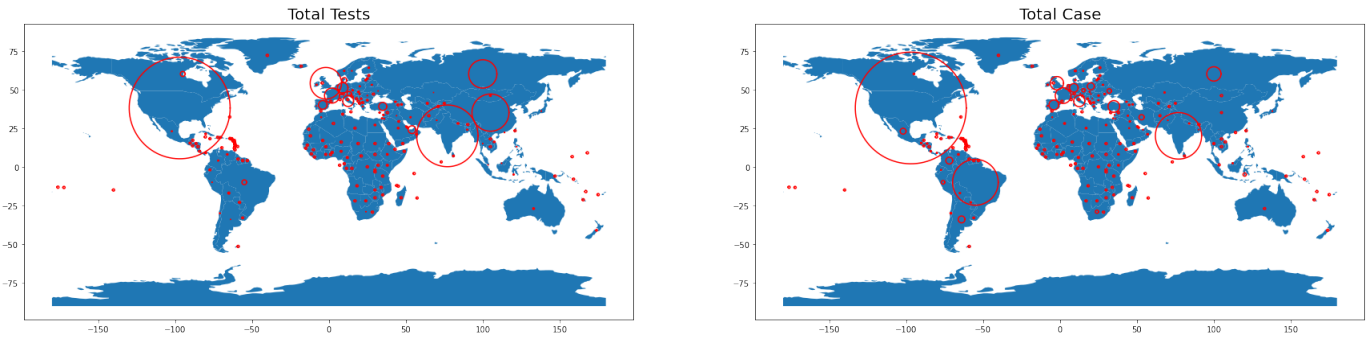


Figure 9: Tương quan giữa Total Cases với Total Test

6 Tham khảo

1. Các cơ sở toán học về thống kê và xấp xỉ tuyến tính được tham khảo trong cuốn "Thống kê máy tính tóm tắt" - TS. Bùi Tiến Lên

2. Sử dụng framework Crawler: `https://gaire-crisna.medium.com/corona-data-scraping-with-s`