ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

MÔN HỌC: TRỰC QUAN HÓA DỮ LIỆU

Báo cáo Lab 1: Mối quan hệ giữa các trường trong dữ liệu

Nhóm sinh viên
Nguyễn Huy Hải – 18120023
Phạm Công Minh – 18120058
Nguyễn Thanh Tùng – 18120104
Nguyễn Văn Hậu – 18120359
Trần Đại Tài – 18120543

Contents

1 Thông tin nhóm

1.1 Mức độ hoàn thành tổng thể của mỗi yêu cầu

Yêu cầu	Mức độ hoàn thành
Thu thập và tiền xử lý dữ liệu	100%
Chọn lựa, giải thích, trực quan các trường và mối quan hệ giữa chúng	100%
Rút ra ý nghĩa hợp lý sau mỗi biểu đồ trực quan	100%
Xem xét trên nhiều quan hệ, nhiều góc nhìn khác nhau	100%
Báo cáo, trình bày bố cục và định dạng hợp lý, rõ ràng	100%

1.2 Công việc của mỗi thành viên

- **Nguyễn Huy Hải - 18120023**
  - Chọn câu hỏi để vẽ biểu đồ scatter, giải thích lý do vẽ biểu đồ và nhận xét.
  - Sử dụng thuật toán máy học đơn giản để dự đoán.
- **Phạm Công Minh - 18120058**
  - Chọn câu hỏi để vẽ biểu đồ histogram, giải thích lý do vẽ biểu đồ và nhận xét.
  - Chọn câu hỏi để vẽ biểu đồ cột đơn giải thích lý do vẽ biểu đồ và nhận xét.
- **Nguyễn Thanh Tùng - 18120104**
  - Thu thập dữ liệu, tiền xử lý dữ liệu.
  - Chọn câu hỏi để vẽ biểu đồ đường, giải thích lý do vẽ biểu đồ và nhận xét.
- **Nguyễn Văn Hậu - 18120359**
  - Tìm hiểu mối tương quan giữa các trường trong dữ liệu.

  - Chọn câu hỏi để vẽ biểu đồ tròn, giải thích lý do vẽ biểu đồ và nhận xét.

- **Trần Đại Tài - 18120543**
  - Chọn câu hỏi để vẽ biểu đồ đôi, giải thích lý do vẽ biểu đồ và nhận xét.
  - Tổng hợp lại và viết báo cáo.

2 Chi tiết từng yêu cầu

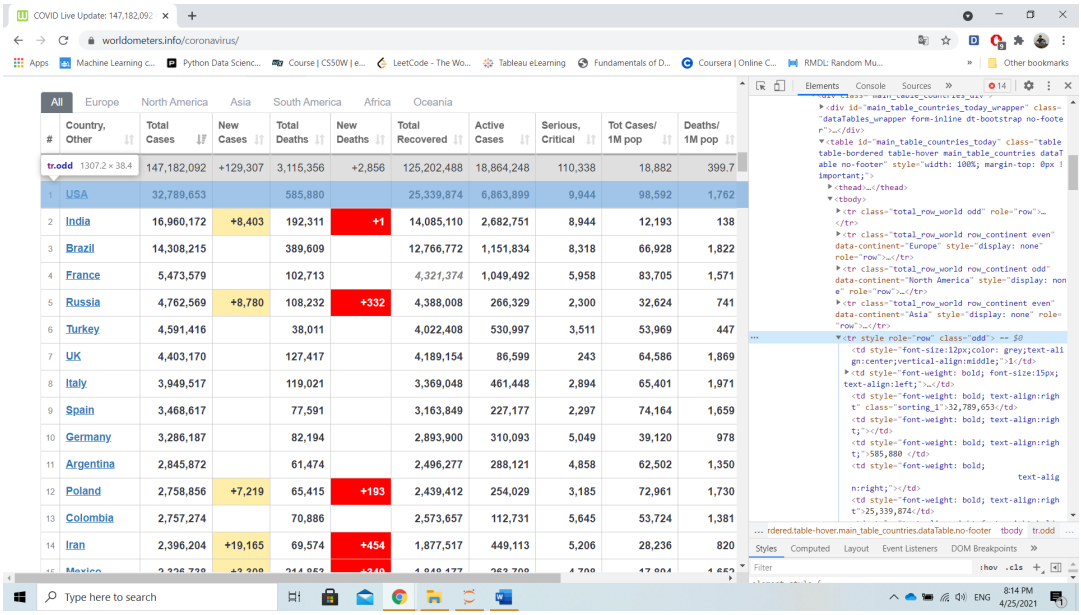
2.1 Thu thập và tiền xử lý dữ liệu

- Thu thập dữ liệu.
  - Thư viện sử dụng: request, selenium, BeauttifulSoup.
  - Ngày thu thập dữ liệu: 16/4/2000 - 21/4/2000.
  - Tiến hành thu thập dữ liệu.

Bước 1: Sử dụng trình duyệt web Google chrome để mở địa chỉ url.

```
url = f"https://www.worldometers.info/coronavirus/"
browser = webdriver.Chrome(executable_path="Driver/chromedriver.exe")
browser.get(url)
```

Bước 2: Ta nhận thấy toàn bộ dữ liệu được để trong thẻ <tbody>, mỗi dòng tương ứng với một thẻ <trow> với role =”row".



Ta tiến hành lấy html_text của trang web và tìm tất cả những thẻ <tr> có role = “row”. Vì với role = “row” sẽ có dữ liệu của cả thế giới và các châu lục nhưng mục đích của ta là thu thập theo từng nước nên ta sẽ loại bỏ dữ liệu của cả thế giới và các châu lục khác đi.

```
html_text = browser.page_source
tree = BeautifulSoup(html_text, 'html.parser')
tbody = tree.find("tbody")
trows = tbody.find_all("tr",{"role": "row"})
total_row_world = tbody.find_all("tr",{"class" : "total_row_world"})
list_data_country = [trow.text.split("\n")[1:-2] for trow in trows if trow not in total_row_world]
```

Bước 3: Vì em chỉ thu thập được dữ liệu của ngày hiện tại và dữ liệu sẽ được reset lúc 0h GMT +0 nên em sẽ thu thập dữ liệu vào lúc 7h sáng GMT +7 và thêm vào dữ liệu 1 cột "Time" để biết thời gian thu thập dữ liệu.

Bước 4: Lưu dữ liệu vào file "raw_data.xlsx" với sheet = "country". Dữ liệu sau khi thu thập được như hình:

#	Country, Other	Total Cases	New Cases	otal Death	New Death	Total Recovered	Active Case	Serious Critical	Tot Cases/1M pop	Death/1M pop	Total test	Test/1M pop	Population	Time
1	USA	32,297,655	+81,600	580,018	+902	24,840,359	6,877,278	9,816	97,123	1,744	426,562,606	1,282,731	332,542,637	2021-04-16 00:00:00
2	India	14,521,683	+233,943	175,673	+1,338	12,666,889	1,679,121	8,944	10,442	126	263,476,625	189,449	1,390,753,234	2021-04-16 00:00:00
3	Brazil	13,834,342	+76,249	369,024	+3,070	12,298,863	1,166,455	8,318	64,720	1,726	28,600,000	133,796	213,757,680	2021-04-16 00:00:00
4	France	5,224,321	+36,442	100,404	+313	4,046,518	1,077,399	5,914	79,897	1,536	71,424,719	1,092,317	65,388,253	2021-04-16 00:00:00
5	Russia	4,684,148	+8,995	104,795	+397	4,310,557	268,796	2,300	32,087	718	125,100,000	856,942	145,984,203	2021-04-16 00:00:00

- Tiền xử lý dữ liệu.
  - Nhận xét về dữ liệu.
    - Thay thấy những ô trống bằng '0'.
    - Xóa ',' giữa các số.
    - Kiểm tra loại dữ liệu thì các cột có dữ liệu số đều là chuỗi nên ta cần phải chuyển về số.
  - Tiến hành tiền xử lý dữ liệu
    - Thay thấy những ô trống bằng '0'.

```
df = df.replace(["","  "], "0")
```

    - Xóa đi "," và chuyển các cột dữ liệu số nhưng đang ở dạng chuỗi về lại số.

```
# Chuẩn hóa các cột dữ liệu dạng số
num_cols=['Total Cases', 'New Cases', 'Total Deaths',
          'New Death', 'Total Recovered', 'Active Case',
          'Serious Critical', 'Tot Cases/1M pop', 'Death/1M pop',
          'Total test', 'Test/1M pop', 'Population']

for col in num_cols:
    df[col]=df[col].str.replace(',', '')
    df[col]=pd.to_numeric(df[col])
```

Sau khi xử lý xong, ta lưu dữ liệu lại vào file "data.xlsx".

#	Country, Other	Total Cases	New Cases	Total Deaths	New Death	Total Recovered	Active Case	Serious Critical	Tot Cases/1M pop	Death/1M pop	Total test	Test/1M pop	Population	Time
1	USA	32297655	81600	580018	902	24840359	6877278	9816	97123	1744	426562606	1282731	332542637	2021-04-16
2	India	14521683	233943	175673	1338	12666889	1679121	8944	10442	126	263476625	189449	1390753234	2021-04-16
3	Brazil	13834342	76249	369024	3070	12298863	1166455	8318	64720	1726	28600000	133796	213757680	2021-04-16
4	France	5224321	36442	100404	313	4046518	1077399	5914	79897	1536	71424719	1092317	65388253	2021-04-16
5	Russia	4684148	8995	104795	397	4310557	268796	2300	32087	718	125100000	856942	145984203	2021-04-16
6	UK	4383732	2756	127225	34	4139553	116954	329	64307	1866	140944028	2067568	68169005	2021-04-16
7	Turkey	4150039	63082	35320	289	3591550	523169	3205	48791	415	43148200	507281	85057795	2021-04-16
8	Italy	3842073	15937	116366	429	3218975	506732	3366	63620	1927	54532594	902990	60391111	2021-04-16
9	Spain	3407283	10598	76981	99	3129234	201068	2180	72853	1646	44285495	946895	46769173	2021-04-16
10	Germany	3116950	21934	80387	246	2752000	284563	4740	37108	957	52737238	627849	83996696	2021-04-16

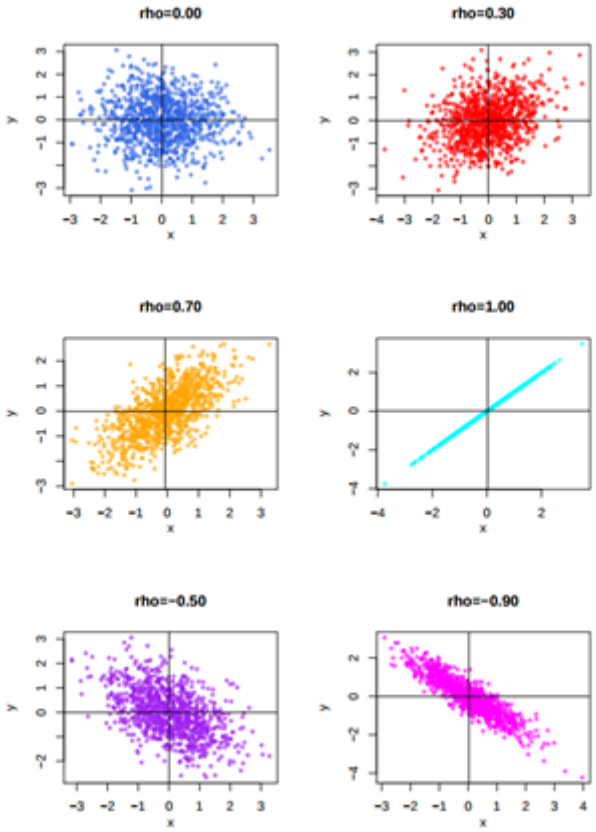
2.2 Mối quan hệ giữa các trường dữ liệu.

- Để tìm mối quan hệ giữa các trường trong dữ liệu, nhóm em sử dụng Correlation.
  Correlation là Covariance được chuẩn hóa của hai biến X, Y.

  – Theo định nghĩa: Correlation coefficient của hai biến X và Y được tính theo công thức:

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sigma(X)+\sigma(Y)}$$

    Trong đó:
    * $Cov(X,Y) = E((X-\mu(X))(Y-\mu(Y)))$
    * $\mu(X), \mu(Y)$ là kỳ vọng của hai biến X, Y.



  – Quan sát hình ta có thể thấy:
    * Nếu $Cov(X,Y) > 0$ hai biến tương quan thuận (càng gần 1 thì mối tương quan càng mạnh).
    * Nếu $Cov(X,Y) < 0$ hai biến tương quan thuận (càng gần 1 thì mối tương quan càng mạnh)

- Khi sử dụng Correlation lên các trường trong dữ liệu, ta có bảng sau:

	Total Cases	New Cases	Total Deaths	New Death	Total Recovered	Active Case	Serious Critical	Tot Cases/1M pop	Death/1M pop	Total test	Test/1M pop	Population
Total Cases	1.000000	0.630262	0.947141	0.667614	0.997698	0.951754	0.817420	0.207022	0.271891	0.882230	0.043141	0.423837
New Cases	0.630262	1.000000	0.495320	0.708727	0.653851	0.501049	0.677265	0.055733	0.081219	0.619027	-0.019567	0.665509
Total Deaths	0.947141	0.495320	1.000000	0.706097	0.950043	0.868171	0.835851	0.209139	0.345832	0.779281	0.028717	0.340833
New Death	0.667614	0.708727	0.706097	1.000000	0.702657	0.467288	0.757277	0.125331	0.243671	0.448873	-0.040077	0.442863
Total Recovered	0.997698	0.653851	0.950043	0.702657	1.000000	0.928877	0.832649	0.211381	0.280926	0.876575	0.043539	0.439430
Active Case	0.951754	0.501049	0.868171	0.467288	0.928877	1.000000	0.696441	0.174455	0.207010	0.861569	0.040181	0.337297
Serious Critical	0.817420	0.677265	0.835851	0.757277	0.832649	0.696441	1.000000	0.203724	0.343710	0.663301	-0.012505	0.435931
Tot Cases/1M pop	0.207022	0.055733	0.209139	0.125331	0.211381	0.174455	0.203724	1.000000	0.819820	0.144954	0.476245	-0.083659
Death/1M pop	0.271891	0.081219	0.345832	0.243671	0.280926	0.207010	0.343710	0.819820	1.000000	0.181774	0.296420	-0.038400
Total test	0.882230	0.619027	0.779281	0.448873	0.876575	0.861569	0.663301	0.144954	0.181774	1.000000	0.109905	0.635215
Test/1M pop	0.043141	-0.019567	0.028717	-0.040077	0.043539	0.040181	-0.012505	0.476245	0.296420	0.109905	1.000000	-0.078441
Population	0.423837	0.665509	0.340833	0.442863	0.439430	0.337297	0.435931	-0.083659	-0.038400	0.635215	-0.078441	1.000000

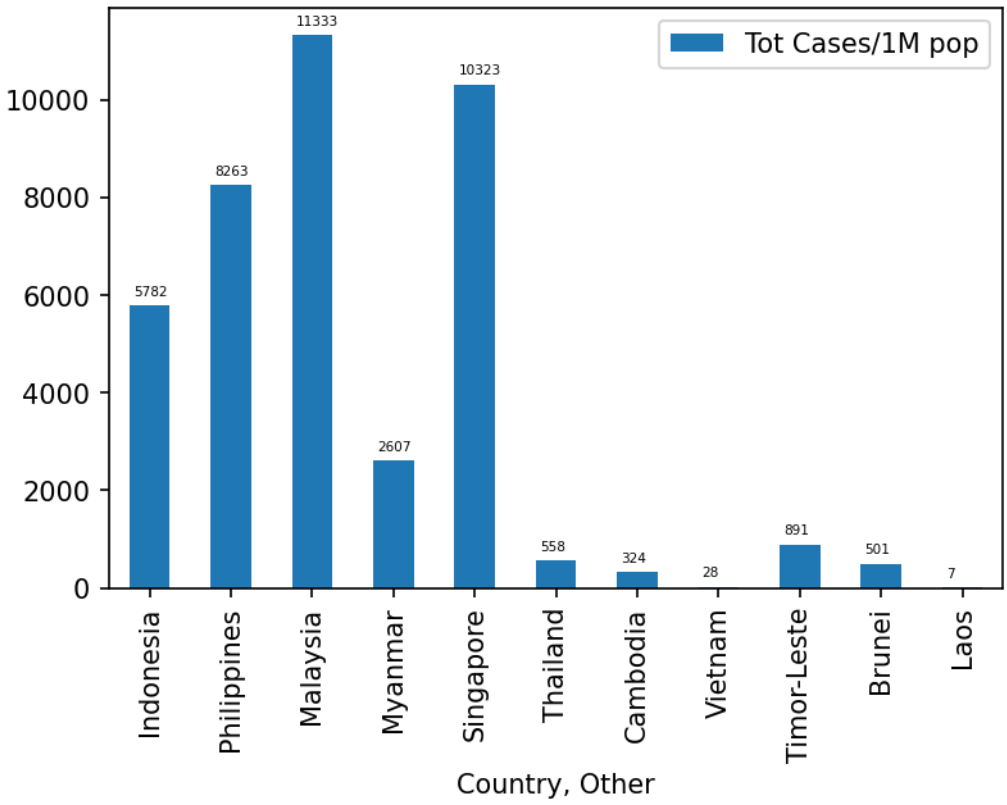
Dễ dàng nhìn thấy một số trường có mối tương quan thuận mạnh trên 0.8 như như:

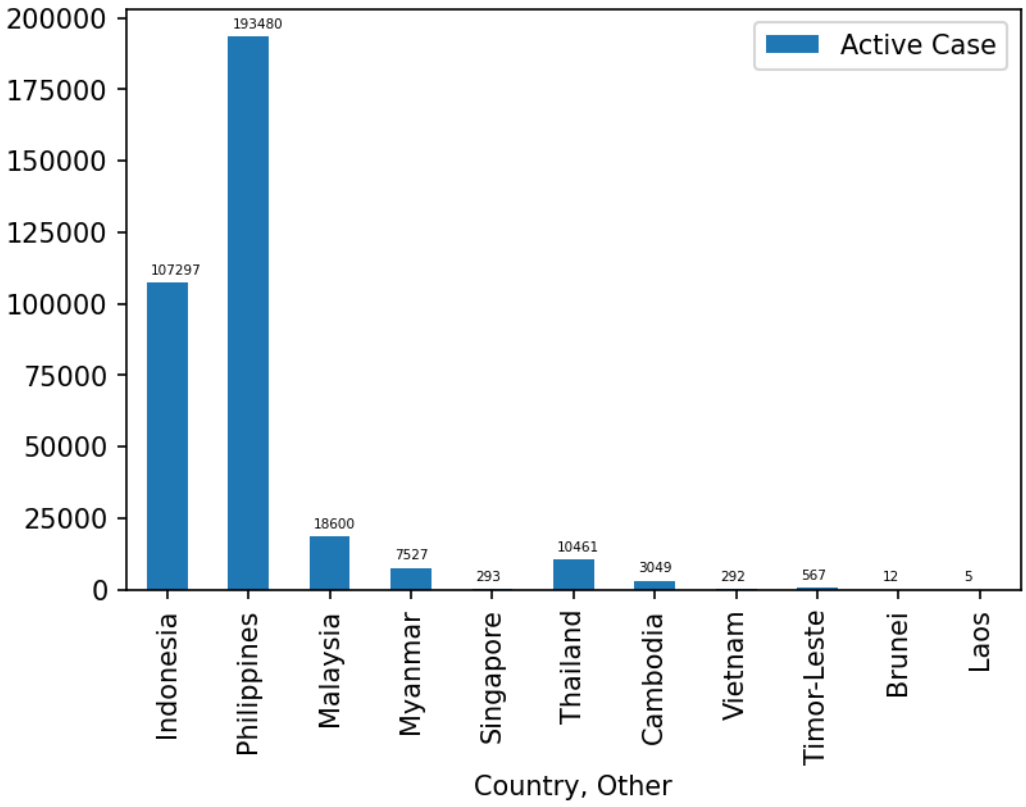
- Total Cases với Total Deaths, Total Recovered, Active Case, Serious Critical, Total test.
- Total Death với Total Recovered, Serious Critical, Active case.
- Total Recovered với Active Case, Serious Critical, Total test.
- Total test với Active Case.

2.3 Các câu hỏi dựa trên dữ liệu.

2.3.1 Xem xét và so sánh tình hình dịch bệnh từ trước đến giờ giữa các nước Đông Nam Á với nhau?

- Chọn trường dữ liệu: Tot Cases/1M và Active Case.
  - Vì các nước có dân số chệnh lệch nhau, về mặt thống kê ta không nên đem Total Cases để đánh giá tình hình dịch ở 1 nước và so sánh với nước khá. Các số liệu hợp lý hơn nên là Total Cases/số người hoặc là Total Cases/diện tích => Ở đây ta có dữ liệu Tot Cases/1M pop nên sẽ chọn Tot Cases/1M pop để trực quan.
  - Để thể hiện tình hình dịch bệnh thì cần xem xét số ca nhiễm hiện giờ và phải chi tiết hơn như số ca ngoài cộng đồng, số ca F1,.. Ở đây chỉ có Active Cases là thể hiện được phần nào tình hình dịch bệnh hiện nay nên sẽ chọn trường này.
- Chọn biểu đồ: Chỉ thể hiện 1 trường dữ liệu nên để cho trực quan và dễ dàng so sánh ta chọn biểu đồ cột.
- Biểu đồ:





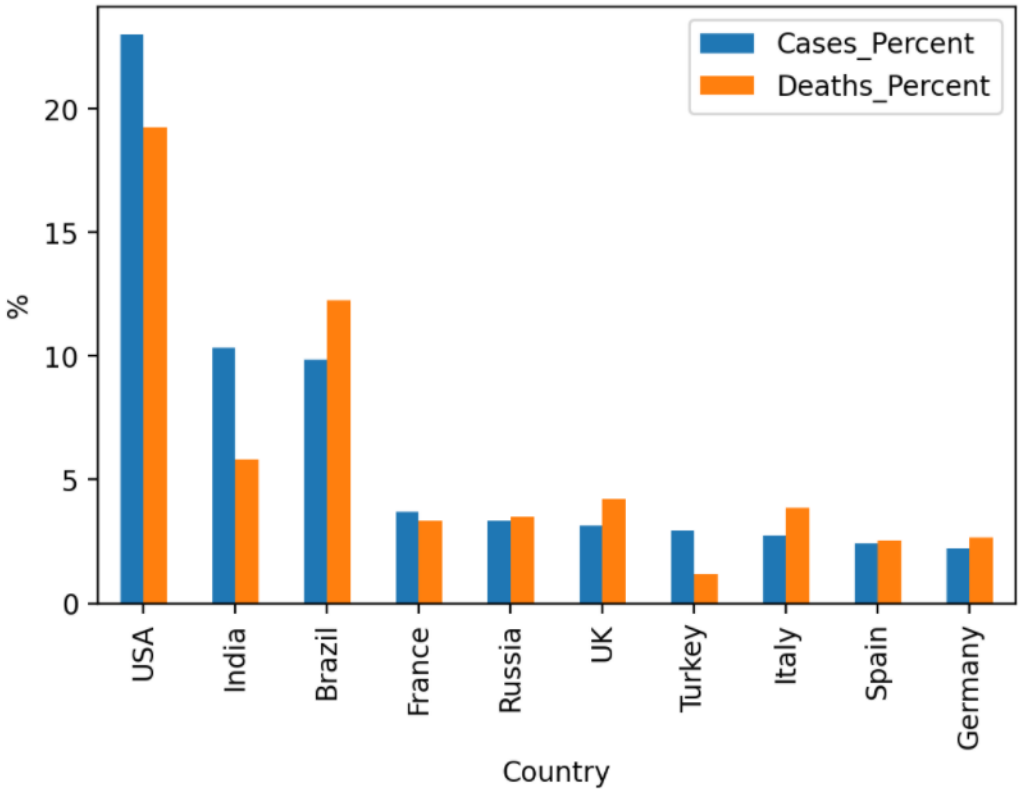
- Nhận xét:
  - Hai biểu đồ cột ở trên lần lượt thể hiện Trung bình số ca mắc trên một triệu người (Tot Cases/1M pop) và số ca bệnh hiện đang có (Active Case) ở mỗi nước thuộc khu vực Đông Nam Á.
  - Tuy nhiên nó lại chỉ thể hiện đơn lẻ từng thuộc tính mà k giúp chúng ta thấy được sự tương quan của các thuộc tính với nhau.

2.3.2 Xem xét số ca tử vong ở các quốc gia có số ca nhiễm cao?

- Chọn trường dữ liệu: Total Cases và Death Cases.
  Trong các trường dữ liệu chỉ có Death Cases là thể hiện được các ca tử vong của các quốc gia. Tuy nhiên, chỉ sử dụng riêng trường dữ liệu này thì không có ý nghĩa so sánh. Các giải quyết ở đây là xem xét Death Cases đồng thời với Total Cases. Ta có thể dựa vào sự chênh lệch giữa Death Cases và Total Cases để so sánh giữa các quốc gia

- Chọn biểu đồ: Biểu đồ cột đôi. Tuy nhiên độ chênh lệch giữa Death Cases và Total Cases quá lớn, nếu để nguyên thì khi trực quan lên sẽ không thấy được cột Death Cases. Cách giải quyết: chuyển các giá trị thành % so với thế giới. Như vậy ta vừa có thể thấy được lượng ca tử vong ở các quốc gia và so sánh giữa các quốc gia với nhau.

- Biểu đồ:



- Nhận xét:
  - Ta có thể thấy 2 thuộc tính này về cơ bản là có tương quan với nhau, cùng cao hoặc cùng thấp ở mỗi nước. Nhưng có sự chênh lệch về tỉ lệ (Với 1 vài nước thì Cases Percent cao hơn Deaths Percent và ngược lại), điều này có thể lý giải là do trình độ phát triển của mỗi nước và khả năng phòng chống dịch bệnh khác nhau.
  - Biểu đồ vẫn chưa hoàn toàn thể hiện mối tương quan giữa các thuộc tính.

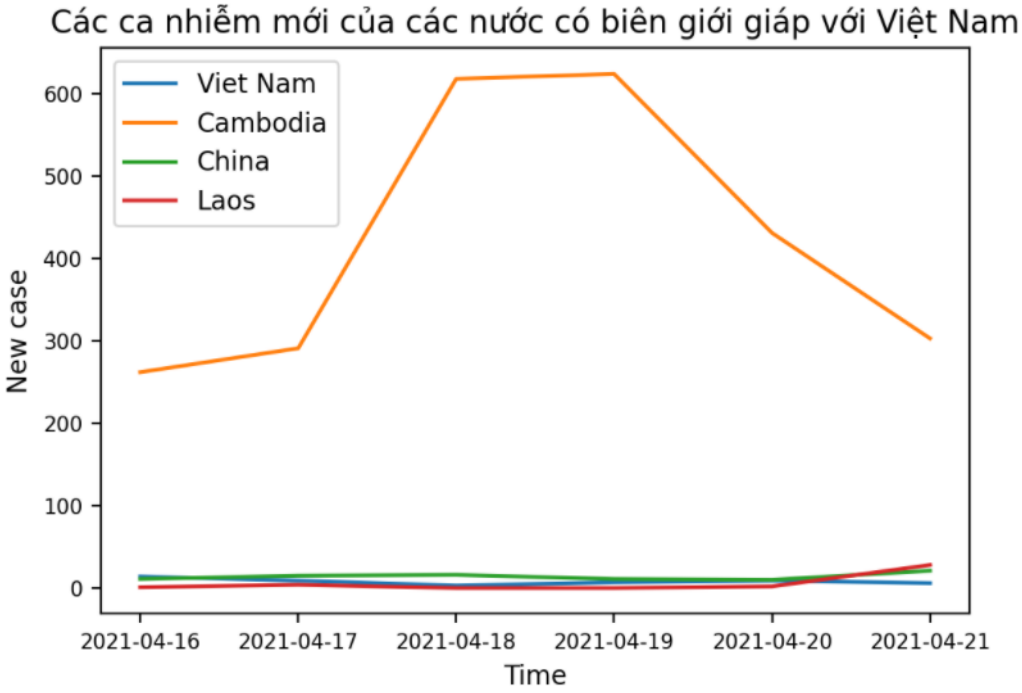
2.3.3 Tình hình diễn biến dịch bệnh ở Việt Nam và các nước có đường biên giới giáp với Việt Nam trong thời gian qua?

- Chọn trường dữ liệu: New Cases.
  Trong các trường dữ liệu chỉ có New Cases là thể hiện được số ca nhiễm mới của mỗi quốc gia trong từng ngày.

- Chọn biểu đồ: Biểu đồ đường. Vì biểu đồ này vừa có các điểm thể hiện rõ số ca mắc cụ thể ở mỗi ngày (thông tin rõ ràng), vừa có các

đường nối để thể hiện xu hướng tăng giảm các ca mắc, giúp ta có cái nhìn trực quan hơn về tình hình dịch bệnh ở các nước .

- Biểu đồ:



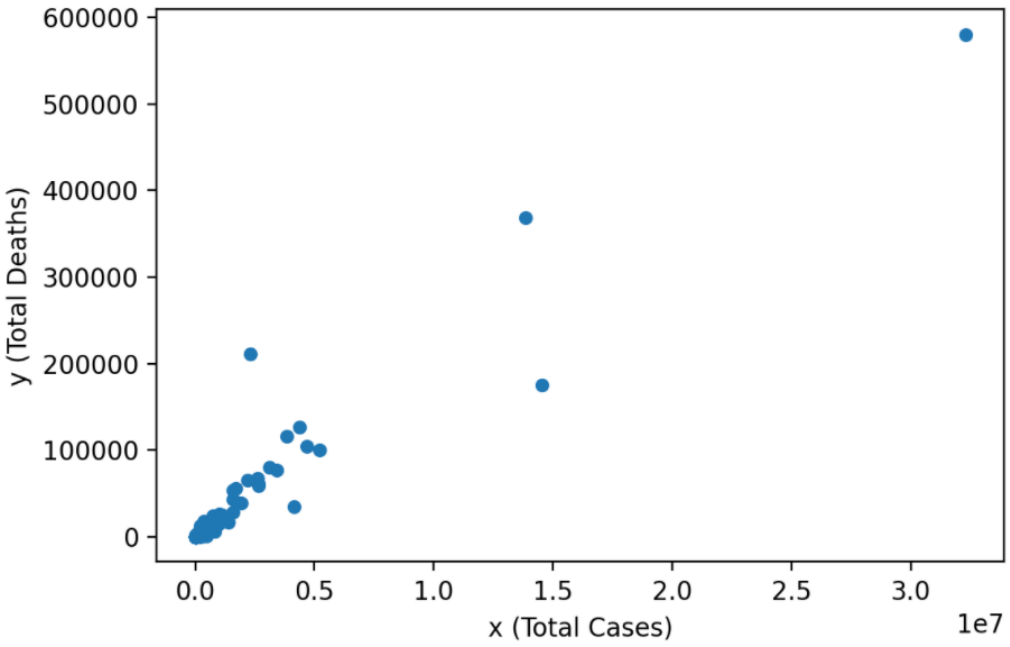
- Nhận xét:
  - Số ca nhiễm mới ở các nước Việt Nam, Lào, Trung Quốc không nhiều. Nhưng ở Campuchia thì lại rất cao, có ngày hơn 600 ca nhiễm.
  - Cần phải kiểm tra kĩ những khu vực biên giới để đề phòng tình trạng vượt biên trái phép, khai báo y tế không trung thực khi nhập cảnh vào nước ta hơn nữa, nhất là khu vực giáp với Campuchia. Từ đó có thể giảm thiểu khả năng bùng dịch trở lại.

2.3.4 Có mối quan hệ nào giữa tổng số ca nhiễm và tổng số người chết hay không?

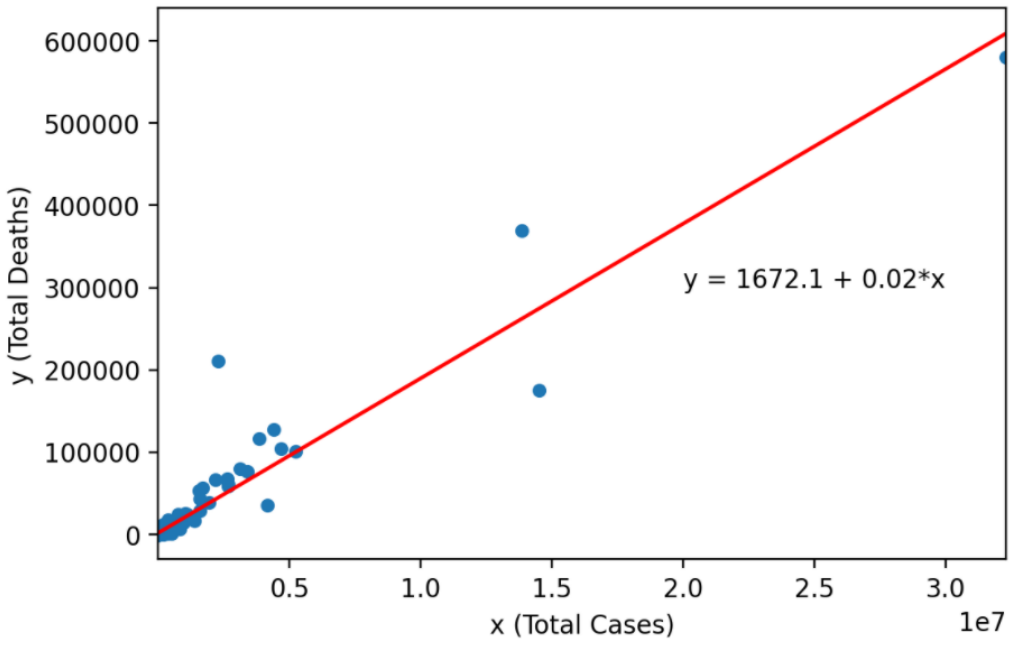
- Chọn trường dữ liệu: Total Cases và Total Deaths.
  Trong các trường dữ liệu có Total Cases thể hiện đượ tổng số ca nhiễm, còn Total Deaths thể hiện tổng số người chết của mỗi quốc gia trong từng ngày.

- Chọn biểu đồ: Biểu đồ scatter. Vì biểu đồ này ta có thể dễ dàng quan sát được mối quan hệ giữa hai trường dữ liệu .

- Biểu đồ:



- Nhận xét:
  - Ta có thể thấy các điểm dữ liệu gần như tạo thành đường thẳng, do đó hai trường dữ liệu này là liên quan mật thiết. Điều này đúng khi ta tính được Correlation giữa hai trường là 0.947141 (có mối tương quan thuận cao).
  - Tuy nhiên vẫn có những điểm Outlier (Ngoại lệ), đây có thể là do các nước đặc biệt (Nước kém phát triển, hoặc không xử lý dịch bệnh đúng cách dẫn đến số ca tử vong cao bất thường), hoặc cũng có thể là do các ca tử vong được cập nhật sớm hoặc những nguyên nhân khác. Nhưng số lượng các điểm này rất ít nên không ảnh hưởng đến tình tổng quan của dữ liệu.
- Sử dụng Linear Regression để thể mối tương quan giữa Total Cases và Total Deaths.



Ta tìm được đường thẳng $y = 1672.1 + 0.02 * x$
Mô hình tìm được có điểm dự đoán chính xác là: 0.9007.

2.3.5 Tỉ lệ giữa số ca đang nhiễm, số người chết, số người được chữa trị giữa Việt Nam và Italy ngày 16/4/2021?

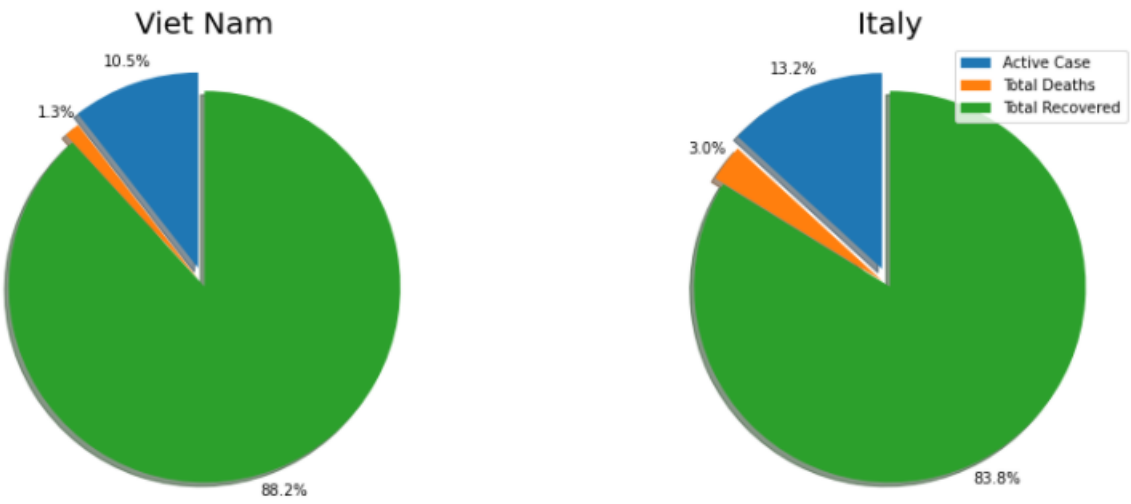
- Chọn trường dữ liệu: Active Cases, Total Deaths, Total Recovered, Time lấy ngày 16/4/2021.
  Trong các trường dữ liệu có Active Cases thể hiện số ca đang nhiễm, Total Deaths thể hiện số người chết, Total Recovered thể hiện số người được chữa trị của mỗi quốc gia trong từng ngày.

- Chọn biểu đồ: Biểu đồ tròn. Vì

$$ActiveCases + TotalDeaths + TotalRecovered = TotalCases$$

  nên ta có thể tính được trong số trường hợp nhiễm đó thì số ca đang nhiễm, số người chết, số người được chữa trị chiếm tỉ lệ như thế nào?. Ngoài ra, ta cũng có thể so sánh được tỉ lệ giữa Việt Nam và Italy.

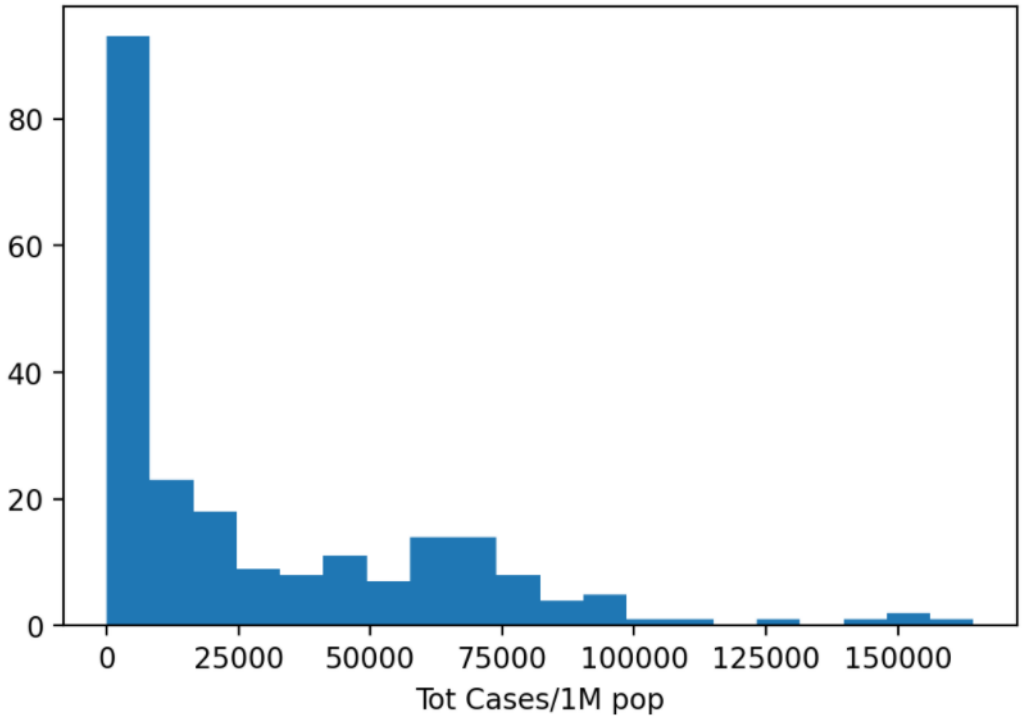
- Biểu đồ:

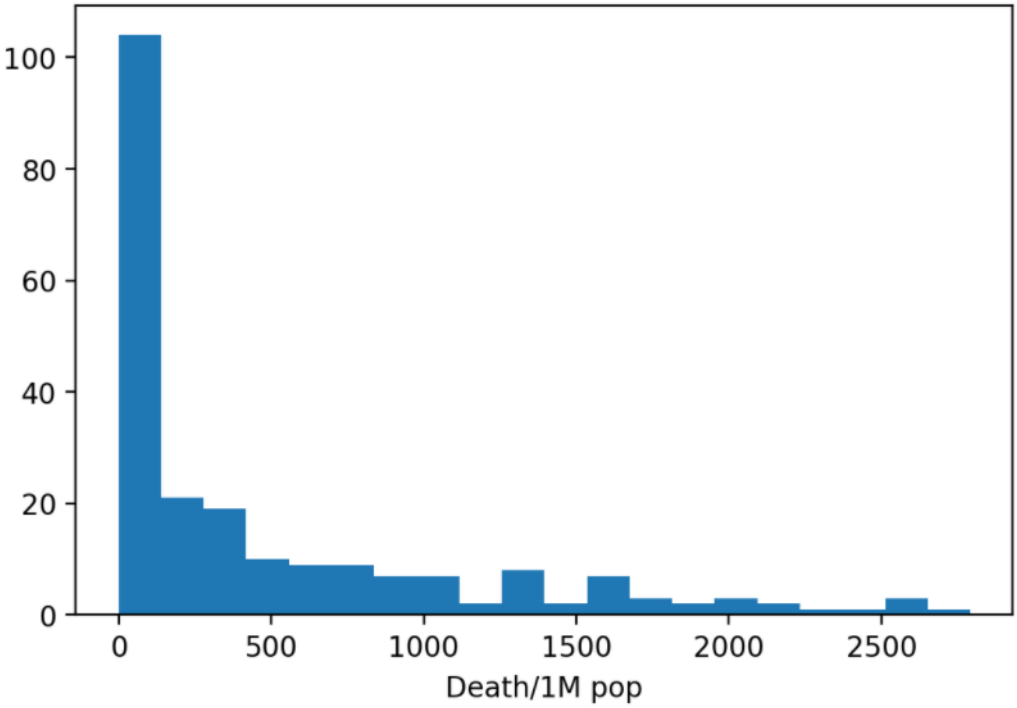


- Nhận xét:
  - Về cơ bản thì tỉ lệ phần trăm giữa ba dữ liệu Active Case, Total Deaths và Total Recovered của cả hai nước là gần như tương đồng nhau.
  - Về cơ bản thì tỉ lệ phần trăm giữa ba dữ liệu Active Case, Total Deaths và Total Recovered của cả hai nước là gần như tương đồng nhau.

2.3.6 Sự phân bố giá trị của Tot Cases/1M pop và Death/1M pop ngày 16/4/2021.

- Chọn trường dữ liệu: Tot Cases/1M pop và Death/1M pop của ngày 16/4/2021.
- Chọn biểu đồ: Biểu đồ histogram để thể hiện sự phân bố.
- Biểu đồ:





- Nhận xét riêng:
  - Nhận xét về Tot Cases/1M pop:
    * Các giá trị tập trung chủ yếu dưới 25000. Đặc biệt là dưới 6000.
    * Phân bố với giá trị trên 100000 rất thưa. Tuy nhiên ở khoảng giá trị 150000 thì sự phân bố lại có sự gia tăng,

đây là khoảng giá trị rất cao => Có thể đây là giá trị Tot Cases/1M pop ở các nước đã vỡ trận trong trận dịch này.

- Nhận xét về Death/1M pop:
  * Các giá trị tập trung chủ yếu dưới 500. Đặc biệt là dưới 150.
  * Phân bố giá trị ở khoảng cuối (khoảng 2500) cũng có sự nhô lên bất thường => Cũng như nhận xét ở Tot Cases/1M pop đây cũng có thể đây là giá trị Death/1M pop ở các nước đã vỡ trận trong trận dịch này.

- Nhận xét chung:
  Có thể thấy 2 biểu đồ gần như tương tự nhau: cá giá trị tập chung phần lớn ở bin đầu tiên và ở các bin sau có xu hướng giảm dần về không có sự thay đổi đột ngột
  => Có thể thấy là chia rõ ra thành 2 nhóm nước (kiểm soát tốt dịch và dịch đã bùng phát). Range của giá trị Tot Cases/1M pop và Death/1M pop của 2 nhóm nước có sự chênh lệch lớn. Tot Cases/1M pop có thể chia thành 2 nhóm giá trị nhóm 0 - 25000 và nhóm từ 25000 - hơn 150000. Khoảng giá trị của 2 nhóm này là lớn.
  => Từ đó có thể rút ra được tính chất của dịch: nếu để dịch bùng phát thì số ca nhiễm sẽ tăng rất đột biến và rất khó kiểm soát

3 Link code thu thập, tiền xử lý dữ liệu, vẽ biểu đồ.

Link github: https://github.com/nttung2603/Lab1-Relationship

4 Tài liệu tham khảo.

- Thư viện sử dụng trong đồ án:
  - pandas.
  - datetime.
  - matplotlib.
  - sklearn.
  - selenium.
  - BeautifulSoup
  - request.