**BÁO CÁO**

------------------------------------------------------------------------------------

MÔN HỌC : TRỰC QUAN HÓA DỮ LIỆU

LAB 01: MỐI QUAN HỆ CỦA DỮ LIỆU

Lớp : 19/21

Giảng viên hướng dẫn thực hành: Thầy Lê Ngọc Thành

Các sinh viên thực hiện:

Hồ Hoàng Duy – 19120207

Nguyễn Đoan Phúc - 19120330

Nguyễn Đắc Thắng – 19120364

# Phân công công việc các thành viên thực hiện

|  |  |  |
| --- | --- | --- |
| MSSV | Công việc thực hiện | Đánh giá |
| 19120207 | 1. Thực hiện phân tích dữ liệu của 10 quốc gia có dân số lớn nhất trên thế giới 2. Viết báo cáo | Hoàn thành 100%. |
| 19120330 | 1. Tiền xử lý và khám phá dữ liệu 2. Thực hiện phân tích dữ liệu của 10 quốc gia có tổng số ca nhiễm lớn nhất trên thế giới. 3. Viết khung báo cáo | Hoàn thành 100%. |
| 19120364 | 1. Thực hiện phân tích dữ liệu của 10 quốc gia được chọn ngẫu nhiên trên thế giới. 2. Viết báo cáo | Hoàn thành 100%. |

# II. Môi trường cài đặt

- Môi trường lập trình : Jupyter Notebook

- Ngôn ngữ sử dụng : Python

- Các thư viện hỗ trợ trong lab01 :

* BeautifulSoup
* requests
* numpy
* pandas
* matplotlib

# III. Giới thiệu về nguồn dữ liệu

Dữ liệu được lấy từ trang web <http://www.worldometers.info/> vào ngày 30/4/2002.

# IV. Tiền xử lý dữ liệu

* Loại bỏ bớt một số cột không định dùng trong quá trình tính toán (những cột thiếu quá nhiều dữ liệu như "NewCases","NewDeaths", "NewRecovered ")
* Chuyển đổi các cột đang xét sang dạng dữ liệu thích hợp (chuyển sang dạng float32).

# V. Khám phá dữ liệu

## Tổng ca nhiễm mỗi quốc gia chiếm bao nhiêu phần trăm so với tổng ca nhiễm của thế giới?

*+* Biểu đồ lựa chọn : Pie chart.

+ Giải thích nguyên nhân lựa chọn biểu đồ : Vì pie chart là biểu đồ hình tròn phù hợp để thể hiện mối quan hệ theo % của các biến thành phần (ở đây là ‘TotalCases’ của các quốc gia – World TotalCases) dựa trên một biến tổng thể (TotalCases của thế giới).

+ Ý tưởng thực hiện : Sau khi tính toán, đối với các quốc gia có ‘TotalCases’/ World TotalCases

< 1% thì tính tổng % của các quốc gia này và trực quan hóa trong biểu đồ qua biến ‘another\_country’ (để giảm số đối tượng phải thể hiện trong biểu đồ)

Chart, pie chart

Description automatically generated+ Kết quả thu được :

+ Nhận xét rút ra từ kết quả trực quan hóa dữ liệu: Nhìn vào biểu đồ ta thấy USA, India và Brazil là các nước có tổng số ca nhiễm (TotalCases) lớn nhất. Việt Nam ta chiếm khoảng 2.08% tổng số ca nhiễm trên thế giới

## Nhận xét về mức tử vong (TotalDeaths), mức phục hồi (TotalRecovered) và mức đang điều trị (ActiveCases) trên tổng số ca nhiễm (TotalCases).

+ Giải thích nguyên nhân lựa chọn biểu đồ : Vì pie chart là biểu đồ hình tròn phù hợp để thể hiện mối quan hệ theo % của các biến thành phần (ở đây là các biến ‘TotalDeaths’, ‘TotalRecovered’, ‘ActiveCases’) dựa trên một biến tổng thể (TotalCases).

Chart, pie chart

Description automatically generated+ Kết quả thu được :

+ Nhận xét rút ra từ kết quả trực quan hóa dữ liệu: Nhìn chung, khả năng hồi phục của bệnh là khá cao, số ca đang điều trị ở phạm vi trung bình (dịch bệnh đã được kiểm soát). Tỷ lệ tử vong của bệnh khá thấp.

# VI. Phân tích dữ liệu của 10 quốc gia có tổng số ca nhiễm (TotalCases) nhiều nhất trên thế giới

Từ dữ liệu có được, nhóm lọc ra 10 quốc gia có tổng số ca nhiễm (TotalCases) nhiều nhất trên thế giới.

A screenshot of a computer

Description automatically generated

## Nhận xét tổng quan về TotalCases, TotalRecovered, TotalDeaths, ActiveCases

## *+* Biểu đồ lựa chọn : multiple bar chart

+ Giải thích nguyên nhân lựa chọn biểu đồ : Biểu đồ multiple bar chart thích hợp để trực quan mối quan hệ giữa các biến TotalCases, TotalRecovered, TotalDeaths, ActiveCases trong cùng 1 nhóm cột (ở đây là 1 nước) từ đó đưa ra các nhận xét về tốc độ phục hồi, tổng số ca nhiễm, tổng số ca tử vong, tổng số ca đang nhiễm.... trong cùng một nước. Việc thể hiện 10 nhóm cột đại diện cho 10 nước khác nhau còn giúp ta so sánh được tình hình giữa các nước.

+ Kết quả thu được :

Chart, bar chart

Description automatically generated

+ Nhận xét rút ra từ kết quả trực quan hóa dữ liệu:

Nhìn vào biểu đồ, ta thấy :

* Tốc độ phục hồi dịch bệnh của các nước đều rất tốt ('TotalRecovered' xấp xỉ gần bằng 'TotalCases'),nhất là India và Turkey.
* Germany, France,Italy là những nước đang có tỉ lệ số ca đang điều trị ( ActiveCases/ TotalCases) lớn trong top10. Có vẻ như tình hình dịch bệnh vẫn còn đang căng thẳng ở những quốc gia này. Nhưng nhìn chung, các nước đều đang kiểm soát được dịch (tỉ lệ ActiveCases/ TotalCases thấp)
* Tỉ lệ tử vong nhìn chung thấp, chứng tỏ bệnh không gây chết nhiều.

1. **Mối tương quan Serious,Critical so với ActiveCases.**

*+* Biểu đồ lựa chọn : scatter plot

+ Giải thích nguyên nhân lựa chọn biểu đồ : scatter plot thích hợp dùng để kiểm tra mối quan hệ nào đó có tồn tại hay không giữa hai biến (mối quan hệ mạnh hay yếu, theo chiều thuận hay nghịch ....) .

🡪 Ở đây để trả lời cho các thắc mắc nếu ‘ActiveCases’ tăng thì ‘ 'Serious,Critical' có tăng theo hay giảm theo hay không (liệu có mối liên hệ nào ở đây hay không).

🡪 Và nếu hai biến này không ảnh hưởng lẫn nhau thì điều này cho thấy ý nghĩa như thế nào.

🡪 Biểu đồ scatter plot thích hợp được sử dụng ( 1 điểm phản ánh giá trị của 2 biến, dễ cho ta cái nhìn trực quan hơn).

Chart, scatter chart

Description automatically generated+ Kết quả thu được :

+ Nhận xét rút ra từ kết quả trực quan hóa dữ liệu: Ta thấy hai biến 'Serious,Critical' và 'ActiveCases' không có mối liên hệ thuận hay nghịch với nhau. Nhưng biểu đồ vẫn cho ta thấy tỷ lệ bị biến chứng nặng do bệnh gây ra không nhiều.

## Nhận xét mối quan hệ giữa TotCases/1M, Deaths/1M, Tests/1M

*+* Biểu đồ lựa chọn : scatter plot matrix.

+ Giải thích nguyên nhân lựa chọn biểu đồ : Ở câu hỏi này ta cần 1 điểm dữ liệu phản ánh được ý nghĩa cả 3 biến – đa chiều. Scatter plot matrix khắc phục được nhược điểm của scatter plot và mang lại các giá trị tương tự nếu cần đánh giá sơ bộ.

+ Kết quả thu được :

Chart

Description automatically generated

+ Nhận xét rút ra từ kết quả trực quan hóa dữ liệu:

Nhìn vào scatter plot matrix, ta thấy :

* Các histogram của 'TotCases/1M pop', 'Tests/1M pop', 'Deaths/1M pop' đều không đạt phân phối chuẩn.
* Histogram của 'TotCases/1M pop' có xu hướng lệch về bên phải, điều này cho thấy tổng số ca nhiễm đang có xu hướng giảm.
* Histogram của 'Deaths/1M pop' lệch về bên phải nhiều hơn (negative) có vẻ như số ca tử vong đang có xu hướng giảm.
* Histogram của 'Tests/1M pop' lệch về phía bên trái nhiều hơn có vẻ số lượng test vẫn đang có xu hướng tăng.
* Histogram của 'TotCases/1M pop' cho thấy tổng số ca nhiễm vẫn đang có xu hướng tăng. Biểu đồ histogram của 'Tests/1M pop'
* Giữa 2 biến 'TotCases/1M pop' và 'Tests/1M pop' có mối quan hệ thuận (yếu), nghĩa là nếu 'Tests/1M pop' tăng thì 'TotCases/1M pop' cũng tăng. Tồn tại mối liên hệ yếu giữa 'Deaths/1M pop' với hai biến còn lại.

1. **Liệu có quan hệ nhân quả giữa hai trường TotalCases và TotalTests không? (số ca nhiễm nhiều do có nhiều xét nghiệm?)**

*+* Biểu đồ lựa chọn : scatter plot.

+ Giải thích nguyên nhân lựa chọn biểu đồ : Theo cảm nhận chủ quan từ kiến thức thực tế, ta dễ nghĩ ngay đến chuyện khi tổng số ca nhiễm tăng là do tổng số xét nghiệm tăng lên, xét nghiệm nhiều giúp phát hiện các ca bệnh đang tìm ẩn trong cộng đồng. Đối với câu hỏi này có thể dùng biểu đồ cột hay scatter plot cũng được. Nhưng theo cảm nhận cá nhân, em thấy scatter plot phù hợp hơn.

+ Kết quả thu được :

Chart, scatter chart

Description automatically generated

+ Nhận xét rút ra từ kết quả trực quan hóa dữ liệu: Giữa 'TotalCases' và 'TotalTests' có mối quan hệ tương quan thuận với nhau, ta thấy nếu 'TotalCases' tăng thì 'TotalTests' cũng tăng và ngược lại.

## Thông tin thu được từ tương quan 2 biến TotalCases và Population.

*+* Biểu đồ lựa chọn : Stacked bar chart.

+ Giải thích nguyên nhân lựa chọn biểu đồ : Rõ ràng, TotalCases là “một phần” của Population. Ta đang muốn trực quan hóa tương quan tỉ lệ giữa TotalCases, Population trong một nước và giữa các nước với nhau => sử dụng Stacked bar chart.

+ Kết quả thu được :

Chart, bar chart

Description automatically generated

+ Nhận xét rút ra từ kết quả trực quan hóa dữ liệu: Nhìn vào biểu đồ ta thấy quy mô dân số không ảnh hưởng lắm đến tổng số ca nhiễm. India có tổng số dân lớn nhất nhưng tổng số ca nhiễm lại thấp hơn nhiều so với USA. Nếu xét về tỉ lệ TotalCases/Population thì có vẻ như tình hình dịch đang nghiêm trọng nhất ở France.

# VII. Phân tích dữ liệu của 10 quốc gia có dân số lớn nhất thể giới

Từ dữ liệu có được, nhóm lọc ra 10 quốc gia có dân số (Population) nhiều nhất trên thế giới.

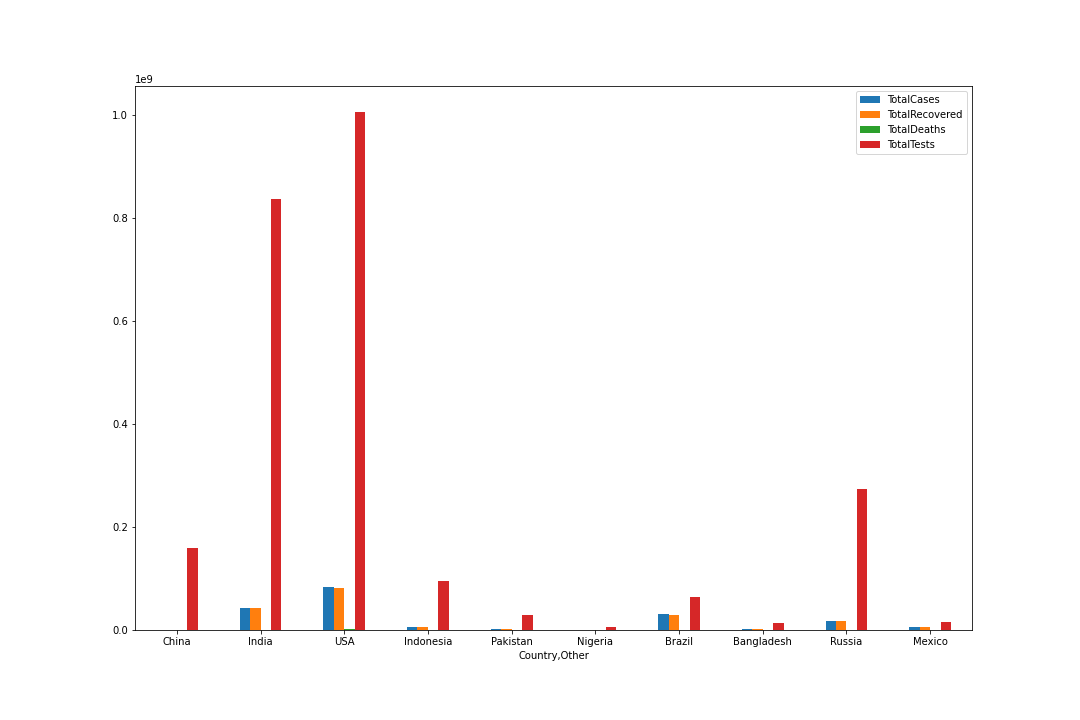
## A screenshot of a computer Description automatically generated

## Nhận xét tổng quan về TotalCases, TotalRecovered, TotalDeaths, ActiveCases

+ Biểu đồ lựa chọn: bar chart

+ Lý do chọn loại biểu đồ: bar chart thích hợp để trực quan mối quan hệ giữa các biến TotalCases, TotalRecovered, TotalDeaths, TotalTests trong cùng 1 trường dữ liệu nhất định (ở đây là các quốc gia), từ đó đưa ra các nhận xét các thông tin trong cùng một quốc gia.

+ Kết quả thu được:

+ Nhận xét rút ra từ kết quả trực quan hóa:

- Ta có thể thấy 1 sự tương quan tỷ lệ thuận với nhau giữa TotalCases, TotalRecovered, TotalDeaths và TotalTests. Tức là, số lượng thực hiện xét nghiệm càng nhiều sẽ số lượng ca nhiễm phát hiện được cũng gia tăng, dẫn đến nguy cơ có nhiều ca tử vong cũng như ca phục hồi ghi nhận.

- Tốc độ phục hồi dịch bệnh của các nước đều rất tốt (TotalRecovered xấp xỉ gần bằng TotalCases).

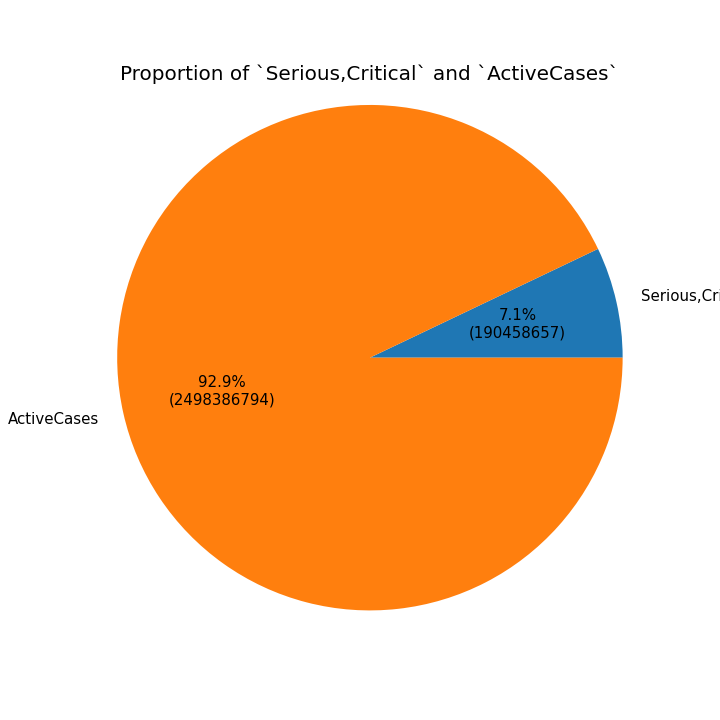
- Số ca tử vong (TotalDeaths) khá thấp nếu so sánh với TotalCases, TotalRecovered.

**2.** **Tỷ lệ Serious,Critical so với ActiveCases**

+ Biểu đồ lựa chọn: pie chart

+ Lý do chọn loại biểu đồ: do biểu đồ tròn phù hợp để thể hiện mối quan hệ theo % của các biến thành phần

+ Kết quả thu được:



+ Nhận xét rút ra từ kết quả trực quan hóa: Có thể thấy rằng tỷ lệ các ca bệnh nặng (Serious,Critical) chỉ chiếm tỷ lệ khá nhỏ (0.9% trong tổng số ca hiện tại). Có thể do biến thể mới có sức lây lan trong cộng đồng mạnh hơn nhưng độc lực đã giảm đi cộng với việc vaccine đã được phủ rộng rãi khiến tỷ lệ chuyển nặng giảm đi rất nhiều.

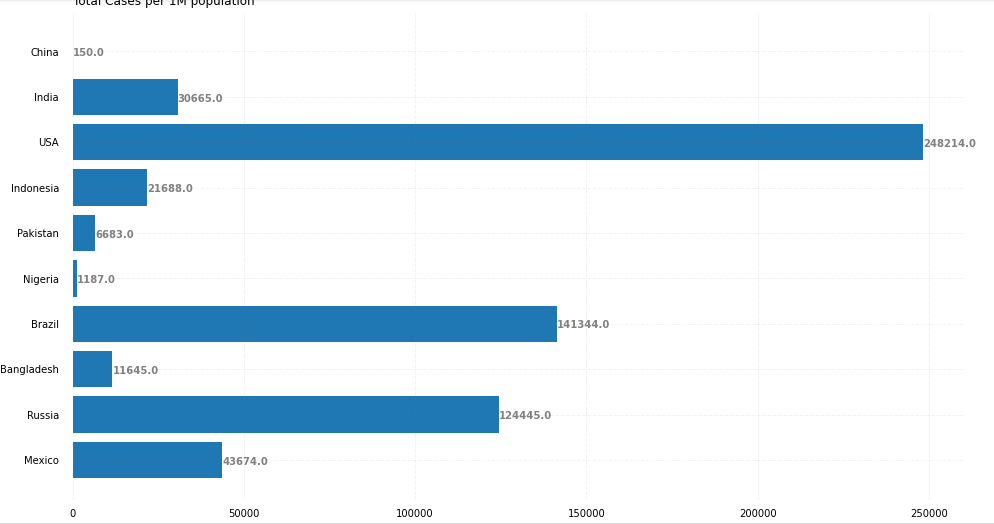
**3. Vẽ biểu đồ thể hiện TotCases/1M, Deaths/1M, Tests/1M**

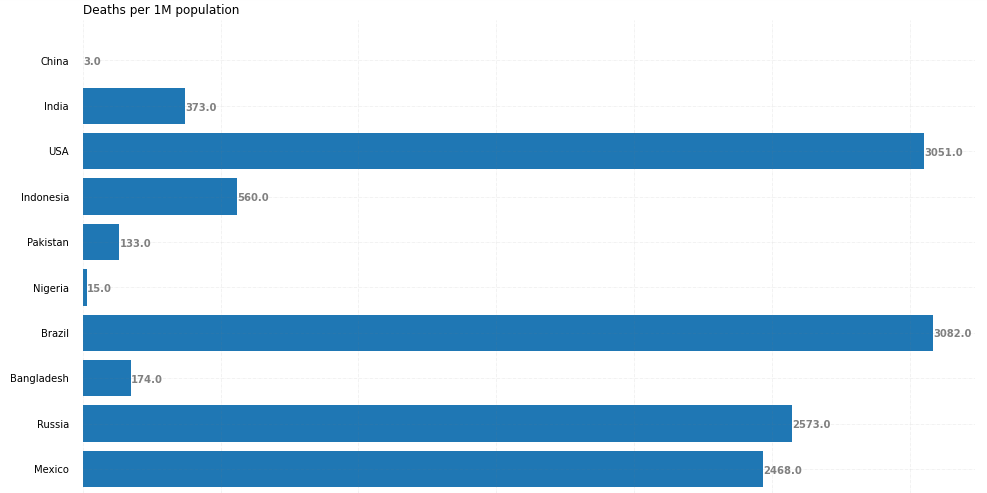
+ Biểu đồ lựa chọn: vertical bar chart + multiple bar chart (horizontal)

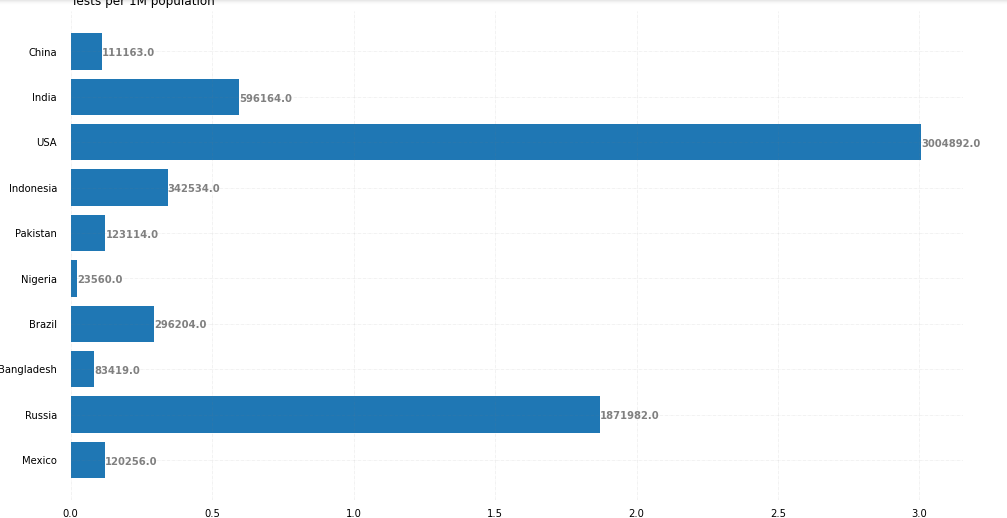
+ Lý do chọn loại biểu đồ: vertical bar chart thích hợp để trực quan mối quan hệ giữa các biến trong cùng 1 trường dữ liệu nhất định (ở đây là các quốc gia) và giúp người xem dễ quan sát do các thông tin được biểu thị theo chiều ngang, thuận tầm mắt theo dõi.

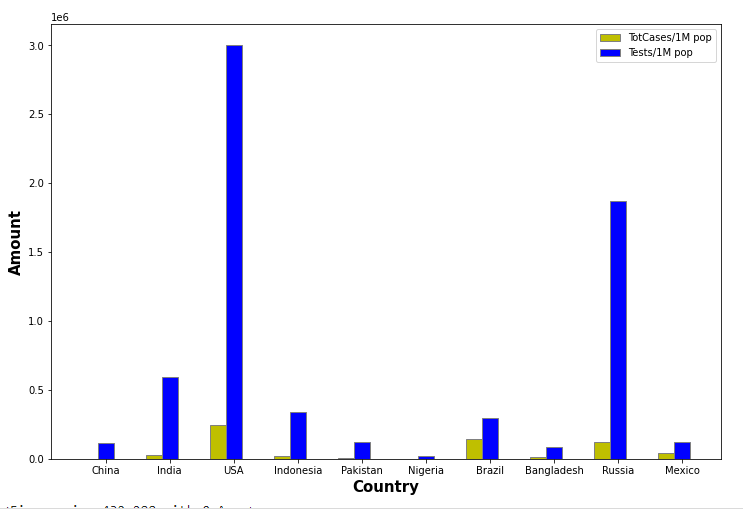
Multiple bar chart thể hiện tốt mối quan hệ giữa hai hay nhiều biến giúp người đọc rút trích các nhận xét của bản thân dựa trên kết quả trực quan.

+ Kết quả thu được:



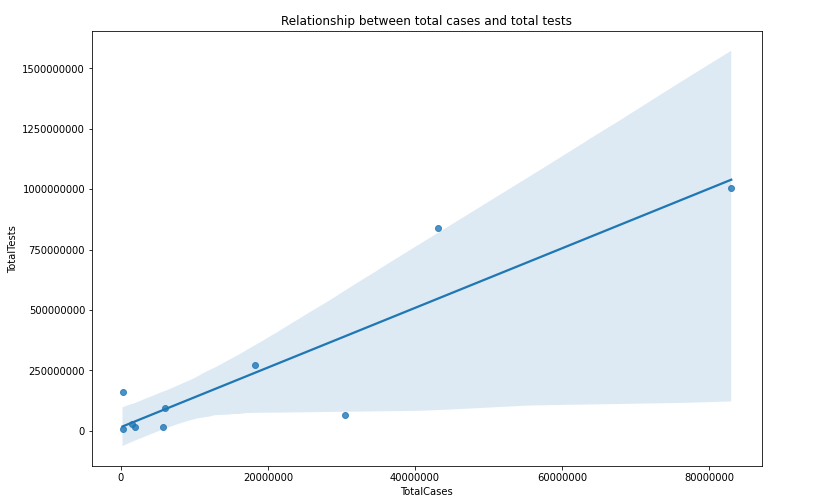






+ Nhận xét rút ra từ kết quả trực quan hóa: Từ các biểu đồ trên ta thấy Mỹ và Nga là 2 quốc gia tiến hành nhiều xét nghiệm nhất nếu tính theo tỷ lệ trên 1 triệu dân. Tuy nhiên ở 1 số quốc gia có tỷ lệ này ở mức thấp, tổng số ca nhiễm và tử vong lại chiếm tỷ trọng khá cao (như Mexico, Brazil). Cho nên 2 trường TotCases/1M pop vaf Tests/1M pop chưa thể hiện mối liên quan nào rõ ràng.

**4. Liệu có quan hệ nhân quả giữa hai trường TotalCases và TotalTests không?**

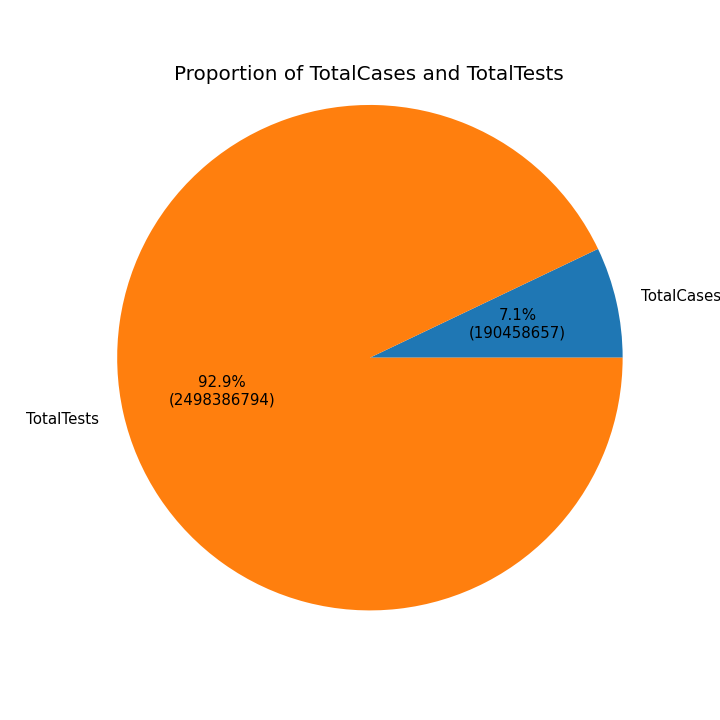


Ta thấy có một mối quan hệ tuyến tính giữa 2 cột này nên có thể đặt ra giải thuyết có mối liên hệ nhân quả giữa chúng (tức là số ca nhiễm một quốc gia ghi nhận nhiều hơn do quốc gia đó đã tiến hành nhiều xét nghiệm chẩn đoán)

**5. So sánh tỷ lệ giữa TotalCases và TotalTests**

**+** Biểu đồ sử dụng: pie chart

+ Kết quả:



+ Nhận xét rút ra từ kết quả trực quan hóa:

- Số lượng test có mối tương quan với số lượng ca nhiễm.

- Tỷ lệ tổng số ca nhiễm trên tổng số xét nghiệm là khoảng 7.1%, tức là cứ 100 xét nghiệm thì phát hiện được khoảng 7 ca bệnh. Điều này cần được đánh giá đúng hơn qua việc xác định xem xét nghiệm thực hiện là xét nghiệm nhanh hay xét nghiệm sinh học phân tử (PCR), vì độ nhạy của các xét nghiệm này là khác nhau.

**VIII. Phân tích dữ liệu của 10 quốc gia được chọn ngẫu nhiên**

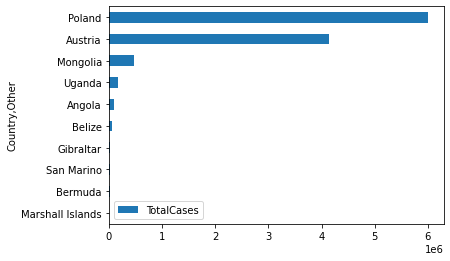
Từ dữ liệu có được, nhóm chọn ra ngẫu nhiên 10 quốc gia để phân tích.

**1. Nhận xét tổng quan về TotalCases, TotalRecovered, TotalDeaths, ActiveCases**

Biểu đồ: bar chart

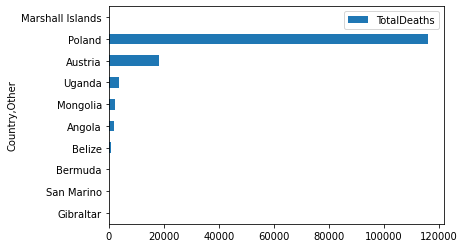
Dùng biểu đồ cột để so sánh những thuộc tính trên theo từng quốc gia.

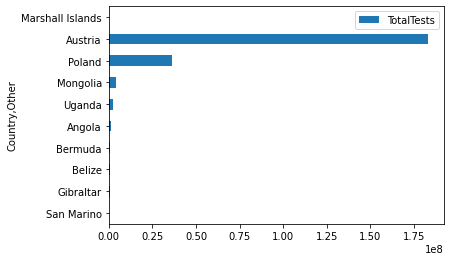
Kết quả:



Graphical user interface, text, application

Description automatically generated





Nhận xét:

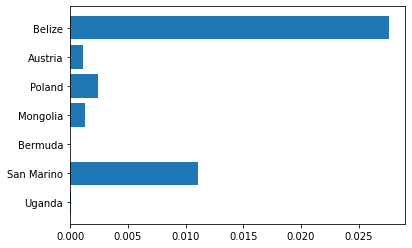
* Poland là quốc gia có nhiều ca nhiễm trong nhóm 10 quốc gia này (trên 6 triệu người vào ngày này). Theo sau là Austria với khoảng 4 triệu ca nhiễm.
* Một vài quốc gia có ca nhiễm rất ít (dưới 100.000 ca nhiễm) như Angola, Belize, Gilbratar, San Marino, Bermuda, Marshall Islands.
* Trên 5 triệu người ở Poland đã hồi phục sau khi mắc COVID-19. Con số này ở Austria là trên 4 triệu người.
* Đặc biệt: Marshall Islands không ghi nhận TotalDeaths, TotalTests vào ngày này.

#### **Mối tương quan Serious,Critical so với ActiveCases**

Biểu đồ: bar chart

Tính tỷ lệ Serious,Critical trên ActiveCases sau đó dùng biểu đồ cột để trực quan.

Kết quả:



Nhận xét:

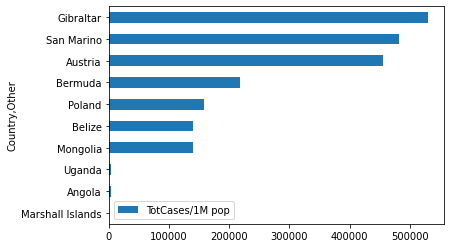
* 3 quốc gia: Austria, Poland, Mongolia có tỷ lệ Serious,Critical trên ActiveCases trung bình là 0.001
* Riêng Uganda có tỷ lệ cực thấp, chỉ 0.000033.
* 2 quốc gia có tỷ lệ cao trên biểu đồ này là Uganda và Belize lần lượt có giá trị là trên 0.01 và 0.025.
* Vào ngày này thì tỷ lệ Serious,Critical trên ActiveCases đều dưới 0.1

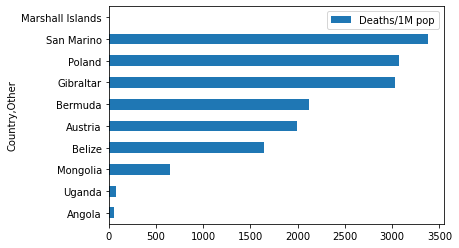
**3. Vẽ biểu đồ thể hiện TotCases/1M, Deaths/1M, Tests/1M**

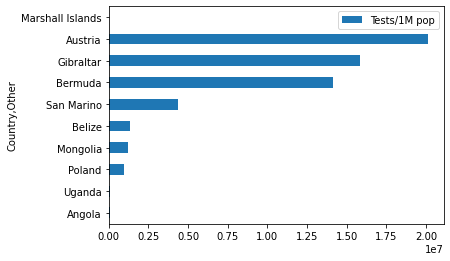
Biểu đồ: bar chart.

Dùng biểu đồ cột để so sánh những thuộc tính trên với những quốc gia khác nhau, đồng thời liệu những giá trị trên cũng có xu hướng như những thuộc tính Total khác (không có /1M).

Kết quả:







Nhận xét:

1. Số ca nhiễm trên 1 triệu dân ở 3 quốc gia Gilbraltar, San Marino và Austria là từ 400.000 đến trên 500.000 người.
2. Số ca nhiễm trên 1 triệu dân ở 4 quốc gia Bermuda, Poland, Belize, Mongolia là từ 150.000 đến trên 200.000. 3 quốc gia còn lại có số lượng rất thấp.
3. Số ca tử vong trên 1 triệu dân cao nhất ở San Marino. Tiếp sau là hai quốc gia Gilbraltar, Poland.
4. Bermuda, Austria, Belize và Mongolia có số ca tử vong trên 1 triệu dân là từ 500 đến trên 2000 người.
5. Uganda và Angola số ca tử vong trên 1 triệu dân rất thấp.
6. Marshall Island không có số liệu về số Tests/1M pop, Deaths/1M pop.
7. Austria, Gilbraltar, Bermuda có số Tests/1M pop từ 15.000.000 đến trên 20.000.000 người
8. San Marino có số Tests/1M pop gần 5.000.000. Những quốc gia còn lại có con số này rất thấp hay không có số liệu về thuộc tính này.

**4. Quan hệ nhân quả giữa hai trường TotalCases và TotalTests**

Biểu đồ: scatter plot.

Tìm mối quan hệ giữa hai biến trên.

Kết quả:

Chart, scatter chart

Description automatically generated

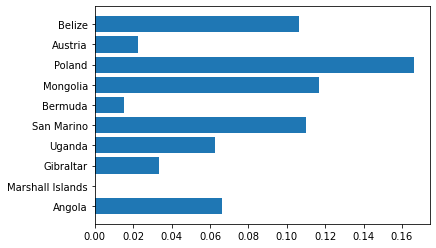
Nhận xét: do dữ liệu ít nên không thể hiện rõ được mối quan hệ giữa hai biến.

**5. Tỷ lệ TotalCases và TotalTests**

Biểu đồ: bar chart

Tính tỷ lệ TotalCases trên TotalTests để so sánh tỷ lệ giữa 10 quốc gia.

Kết quả:



Nhận xét:

* Poland có tỷ lệ TotalCases trên TotalTests cao nhất trong nhóm này với tỷ lệ là trên 0.16
* Tiếp sau là những nước Mongolia, Belize, San Marino với tỷ lệ trên 0.1
* Tỷ lệ trên 0.06 là những quốc gia như Angola, Uganda.
* Dưới 0.04 là những quốc gia như Gilbraltar, Bermudai, Austria.
* Marshall Islands không có thông tin về tỷ lệ này.