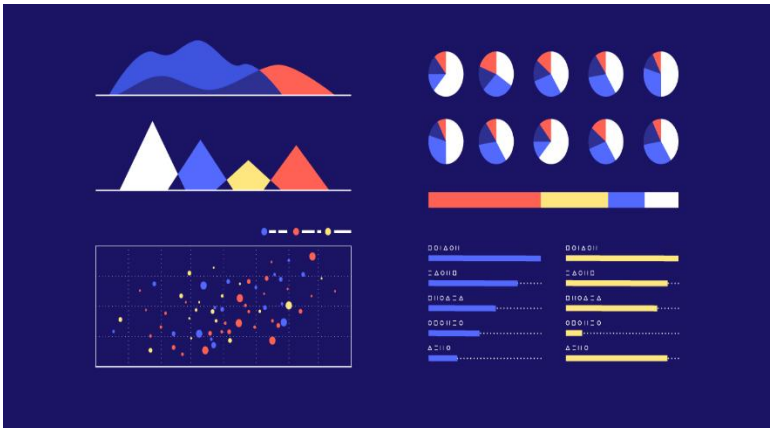
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – TP HCM

KHOA CÔNG NGHỆ THÔNG TIN

★ ★ ★

BÁO CÁO

LAB – MỐI QUAN HỆ CỦA DỮ LIỆU



THÀNH VIÊN

Lê Tiến Trí	**ID: 19127593**
Lê Hoàng Thịnh Phước	**ID: 19127518**
Hoàng Minh Đức	**ID: 19127121**

February $\mathbf{19}^{th}$*, 2022*

MỤC LỤC

I. Thu thập dữ liệu

- Sử dụng thư viện requests kết hợp với thư viện Beautifulsoup trong python để có thể lấy dữ liệu của trang www.worldometers.info

```
headers = {'User-Agent': 'Mozilla/5.0'}
page = r.get("https://www.worldometers.info/coronavirus", headers)
soup  = BeautifulSoup(page.text, 'html.parser')
```

+ Qua đó ta có thể lấy được toàn bộ dữ liệu HTML của trang web sau đó kết hợp quá trình xử lí trang HTML để lấy được dữ liệu

- Thời gian thu thập dữ liệu vào ngày: 13/03/2022
- Dữ liệu được đưa vào file *data.csv*

II. Tiền xử lí dữ liệu

❖ Quá trình tiền xử lí tập trung vào 2 vấn đề chính

- Thiếu dữ liệu ở 1 vài thuộc tính
- Kiểu dữ liệu ở 1 vài thuộc tính bị sai

1. Vấn đề thiếu dữ liệu

- Tiến hành quá trình tiến hành kiểm tra phần trăm dữ liệu bị thiếu trong mỗi thuộc tính

```
Số lượng thông tin bị thiếu trong thuộc tính  Country,Other : 0
Chiếm tỉ lệ phần trăm: 0.0 %

Số lượng thông tin bị thiếu trong thuộc tính  TotalCases : 0
Chiếm tỉ lệ phần trăm: 0.0 %

Số lượng thông tin bị thiếu trong thuộc tính  NewCases : 93
Chiếm tỉ lệ phần trăm: 40.969 %

Số lượng thông tin bị thiếu trong thuộc tính  TotalDeaths : 0
Chiếm tỉ lệ phần trăm: 0.0 %

Số lượng thông tin bị thiếu trong thuộc tính  NewDeaths : 143
Chiếm tỉ lệ phần trăm: 62.996 %

Số lượng thông tin bị thiếu trong thuộc tính  TotalRecovered : 12
Chiếm tỉ lệ phần trăm: 5.286 %

Số lượng thông tin bị thiếu trong thuộc tính  NewRecovered : 115
Chiếm tỉ lệ phần trăm: 50.661 %

Số lượng thông tin bị thiếu trong thuộc tính  ActiveCases : 11
Chiếm tỉ lệ phần trăm: 4.846 %

Số lượng thông tin bị thiếu trong thuộc tính  Serious,Critical : 6
Chiếm tỉ lệ phần trăm: 29.075 %

Số lượng thông tin bị thiếu trong thuộc tính  Tot Cases/1M pop : 2
Chiếm tỉ lệ phần trăm: 0.881 %

Số lượng thông tin bị thiếu trong thuộc tính  Deaths/1M pop : 11
Chiếm tỉ lệ phần trăm: 4.846 %
```

```
Số lượng thông tin bị thiếu trong thuộc tính  TotalTests : 16
Chiếm tỉ lệ phần trăm: 7.048 %

Số lượng thông tin bị thiếu trong thuộc tính  Tests/1M pop : 16
Chiếm tỉ lệ phần trăm: 7.048 %

Số lượng thông tin bị thiếu trong thuộc tính  Population : 0
Chiếm tỉ lệ phần trăm: 0.0 %

Số lượng thông tin bị thiếu trong thuộc tính  Continent : 2
Chiếm tỉ lệ phần trăm: 0.881 %

Số lượng thông tin bị thiếu trong thuộc tính  1 Caseevery X ppl : 2
Chiếm tỉ lệ phần trăm: 0.881 %

Số lượng thông tin bị thiếu trong thuộc tính  1 Deathevery X ppl :
Chiếm tỉ lệ phần trăm: 4.846 %

Số lượng thông tin bị thiếu trong thuộc tính  1 Testevery X ppl : 1
Chiếm tỉ lệ phần trăm: 7.048 %

Số lượng thông tin bị thiếu trong thuộc tính  New Cases/1M pop : 93
Chiếm tỉ lệ phần trăm: 40.969 %

Số lượng thông tin bị thiếu trong thuộc tính  New Deaths/1M pop : 1
Chiếm tỉ lệ phần trăm: 62.996 %

Số lượng thông tin bị thiếu trong thuộc tính  Active Cases/1M pop :
Chiếm tỉ lệ phần trăm: 2.203 %
```

- Loại bỏ các thuộc tính có phần trăm > 30%
- Loại bỏ các dòng dữ liệu của các nước có số lượng cột thiếu dữ liệu lớn hơn 3

2. Vấn đề kiểu dữ liệu bị sai

- Ta tiến hành kiểm tra các kiểu thuộc tính của các cột trong bộ dữ liệu

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Country,Other       227 non-null    object
 1   TotalCases          227 non-null    object
 2   NewCases            134 non-null    object
 3   TotalDeaths         227 non-null    object
 4   NewDeaths           84 non-null     float64
 5   TotalRecovered      215 non-null    object
 6   NewRecovered        112 non-null    object
 7   ActiveCases         216 non-null    object
 8   Serious,Critical    161 non-null    object
 9   Tot Cases/1M pop    225 non-null    object
 10  Deaths/1M pop       216 non-null    object
 11  TotalTests          211 non-null    object
 12  Tests/1M pop        211 non-null    object
 13  Population          227 non-null    object
 14  Continent           225 non-null    object
 15  New Cases/1M pop    134 non-null    object
 16  New Deaths/1M pop   84 non-null     float64
 17  Active Cases/1M pop 222 non-null    object
```

- Tiến hành thay đổi các cột về đúng định dạng dữ liệu của nó

III. Trực quan hóa dữ liệu

❖ Mục tiêu:

- Trực quan bằng các loại biểu đồ đã học (Có thể phủ được hết các loại biểu đồ)
- Thể hiện các mỗi các thuộc tính từ môi trường đơn biến đến quan hệ nhiều trường
- Khám phá ra được các mối quan hệ nhân quả trong bộ dữ liệu

1. Tiến hành so sánh đại dịch Covid-19 với các loại đại dịch trước đây trên thế giới.

   - Để có cái nhìn khách quan về đại dịch ta tiến hành lấy thông tin của đại dịch trước đây trên thế giới.

Dịch bệnh	Số Ca nhiễm	Số Ca tử vong
EBOLA	28646	11323
MERS	8096	858
SARS	2494	774

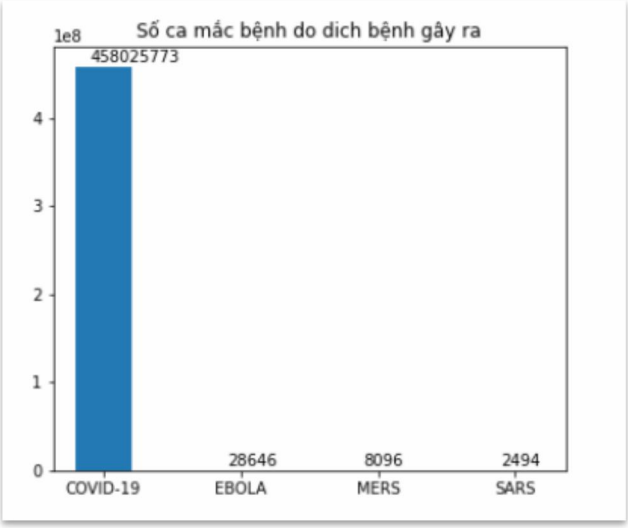
(nguồn dữ liệu từ Internet)

a. Biểu đồ Cột

- Lý do chọn biểu đồ:

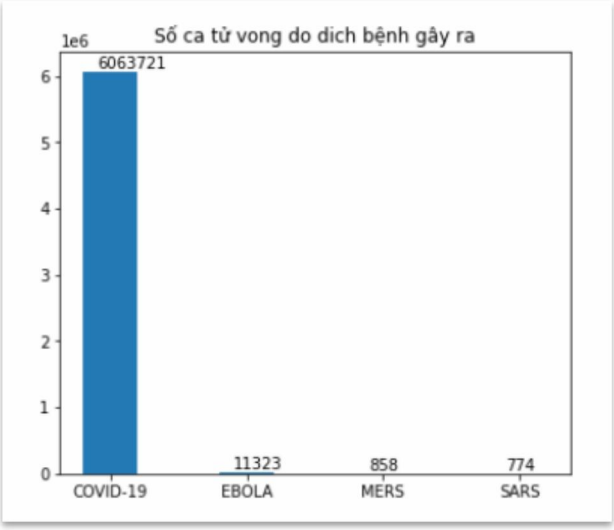
+ Biểu đồ cột dùng để so sánh thể hiện quy mô, số lượng, sản lượng hoặc khối lượng của các đối tượng. Nó rất phù hợp với mục đích trong trường hợp này, khi ta muốn sử dụng để so sánh *số lượng ca nhiễm*, *tử vong* và *tỉ lệ tử vong* của 4 loại dịch bệnh

- Biểu đồ Cột (1): Thể hiện số cac mắc bệnh do 4 đại dịch gây ra



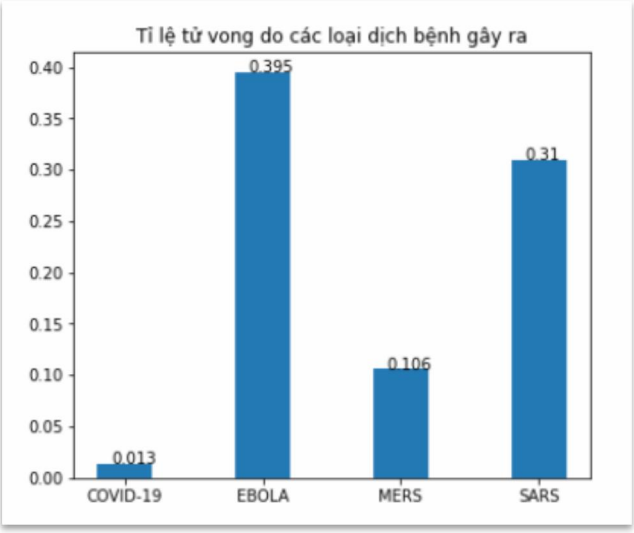
  - Ý nghĩa: Qua biểu đồ cột, ta có thể rõ số lượng mắc ca nhiễm do Covid-19 vượt trội so với 3 dịch bệnh còn lại, với 1 con số kinh khủng 458.025.773, thể hiên tốc độ lây loan nhanh chóng và kinh khủng của con viruss Corona này trên toàn thế giới.

- Biểu đồ cột (2): Thể hiện tỉ lệ do 4 đại dịch gây ra



- Ý nghĩa: Ta thấy rõ số ca tử vong do Covid19 cao ngất ngưỡng, tuy dựa vào biểu đồ ta chưa thể kết luận độ nguy hiểm của nó nhưng với lượng tử vong 6.063.721 người, ta cùng cảm nhận được mức độ hiểm của Covid19 gây ra cho người dân trên thế giới.

- Biểu đồ cột (3): Thể hiện tỉ lệ tử vong khi nhiễm bệnh



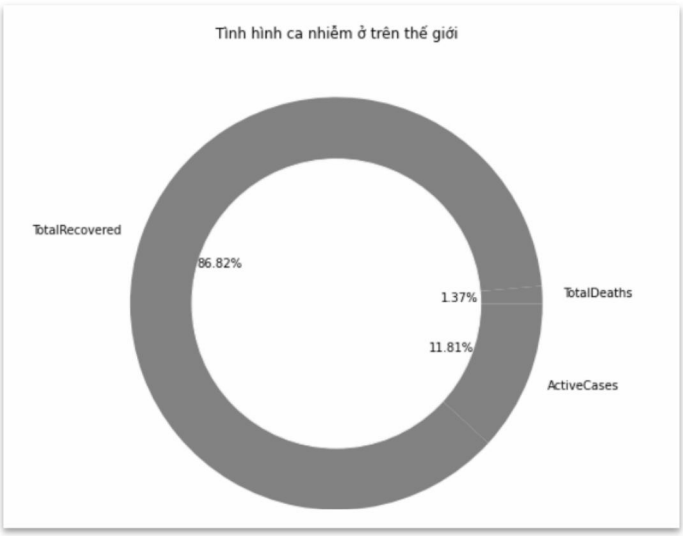
- Ý nghĩa: Ebola chiếm tỉ lệ cao nhất với 0.395% trong khi đó Covid-19 có tỉ lệ thấp nhất với chỉ 0.013%, ta có thể giải thích tại sao ở biểu đồ cột (2) số người tử vong cao của Covid19 do số lượng người nhiễm cao ngất ngưỡng. Tuy tỉ lệ thấp ta cũng có thể đưa ra suy luận do sự phát triển của y học hiện nay và các Vaccine qua đó có thể làm giảm mức độ nguy hiểm của Covid-19

b. Biểu đồ bánh Doughnut

- Lý do chọn biểu đồ

  + Là một dạng biểu đồ tròn dùng để mối quan hệ thành phần tác động đến tổng dữ liệu. Ở trong trường hợp này 3 quan hệ thành phần *totalRecovered*, *totalDeaths* và *ActiveCase* tác động đến tổng dữ liệu liệu là *totalCases.*

- Biểu đồ bánh Doughnut



- Ý nghĩa: Đa số người mắc Covid-19 đều khỏi bệnh, chỉ 1.37% là tử vong. Theo thông tin từ bộ y tế, Covid-19 ảnh hưởng mạnh đến người có bệnh nền (bệnh về hô hấp, ung thư …) và có tỉ lệ tử vong cao, có thể 1.37% liên quan người bệnh nền mắc covid-19

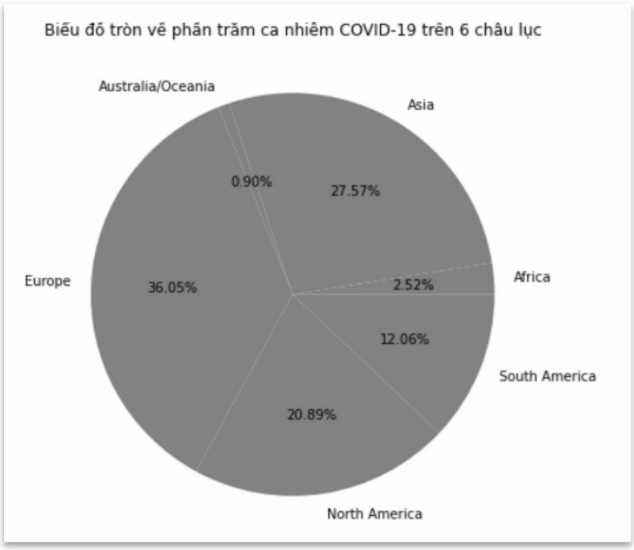
2. Xem xét tỉ lệ ca nhiễm Covid-19 ở 6 châu lục.

a. Biểu đồ tròn

- Lý do chọn đề tài

  + Biểu đồ tròn thường dùng để vẽ các biểu đồ liên quan đến cơ cấu, tỷ lệ các thành phần trong một tổng thể chung và trong trường này ta dùng để biểu thị tỷ lệ của ca nhiễm của 6 châu lục trên thế giới.

- Biểu đồ tròn



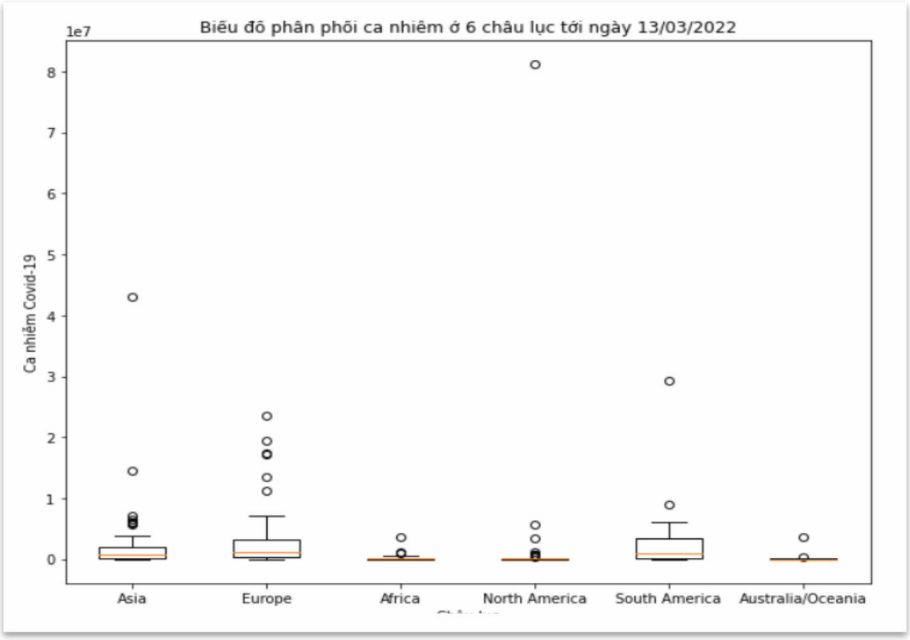
- Ý nghĩa: Tính đến ngày 13/03/2022, Châu âu chiếm tỉ lệ cao nhất 36.05% chiếm hơn 1/3 số ca nhiễm trên toàn thế giới trong khi dân số châu âu chỉ chiếm 9.43% trên thế giới. Qua đó thấy được Covid-19 đang hoành hành và tốc độc lây lan mạnh mẽ. Châu âu phải đã và đang đối diện với làn sóng Covid-19.
- Giả thuyết đặt ra: Hiện có thông tin Châu âu phải đối mặt với biến thể số ca nhiễm Omicron tăng lên, ta đặt ra 2 giả thiết về số ca nhiễm cao nhất ngưỡng ở châu âu
  - 1. Do biến thể mới phát triển mạnh ở thời tiết khí hâu Châu âu qua đó đẩy mạnh tốc độ lây lan
  - 2. Do 1 vài nước trong Châu âu phòng chống dịch không tốt và gây ra số lượng ca nhiễm tăng lên dẫn đến tổng số ca nhiễm ở Châu âu tăng theo.

  ⇨ Sử dụng *biểu đồ hộp* để thể hiện sự phân bố ca nhiễm ở các nước trong 6 châu lục

b. <u>Biểu đồ hộp</u>

- Lý do chọn biểu đồ

  + Biểu đồ hộp dùng để các đại lượng quan trọng (tứ phân vị) và trực quan các outliers, và mục đích dùng biểu đồ này thể hiện sự phân bố số lượng ca nhiễm các nước ở 1 châu lục, đồng thời xác định các nước có số ca nhiễm bất thường ở 1 châu lục để có thể trả lời giả thuyết đặt ra ở trên
- Biểu đồ hộp



- Ý nghĩa: Dựa vào box plox, ta thấy rõ ở Europe đã xuất hiện tới 5 outlier hình dung được giả thiết đầu tiên tỉ lệ ca nhiễm tăng lên đó là do 1 vài nước trong châu âu phòng chống dịch bệnh chưa tốt của ta đã đúng giải thích rõ ràng tại sao tổng số ca nhiễm của châu âu lại chiếm tới 1/3 so với thế giới. Kèm theo đó có 1 điều đặc biệt ở North America xuất hiện 1 outlier lớn qua đó do 1 mình outlier đó đã kéo số người nhiễm trung bình ở châu mĩ (North America) tăng mạnh

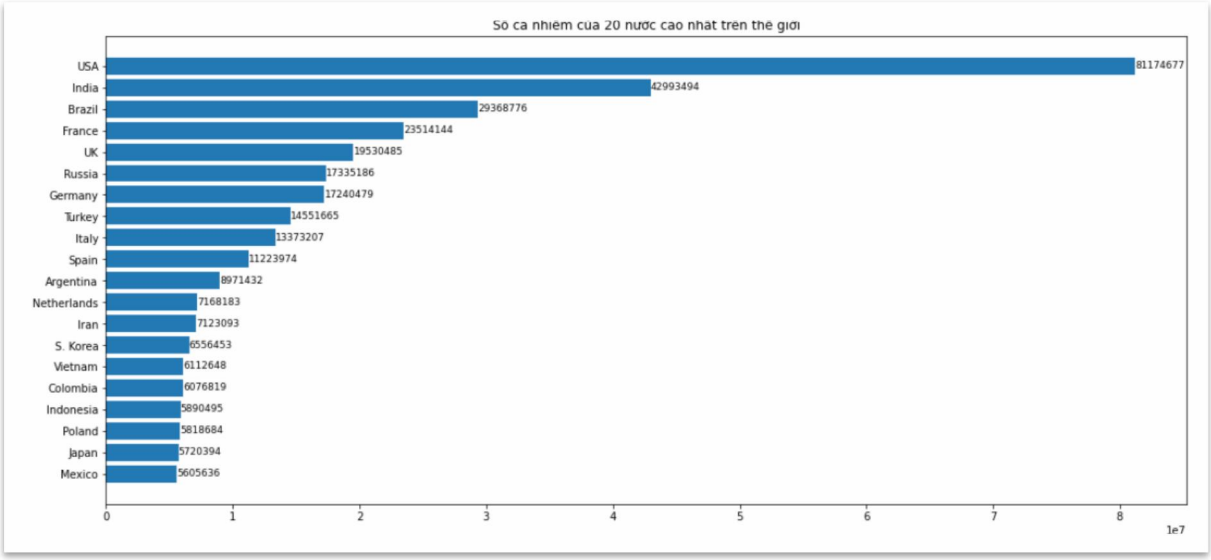
3. <u>Số ca nhiễm và tử vong của 20 nước cao nhất trên thế giới tính đến ngày 13/03/2022.</u>

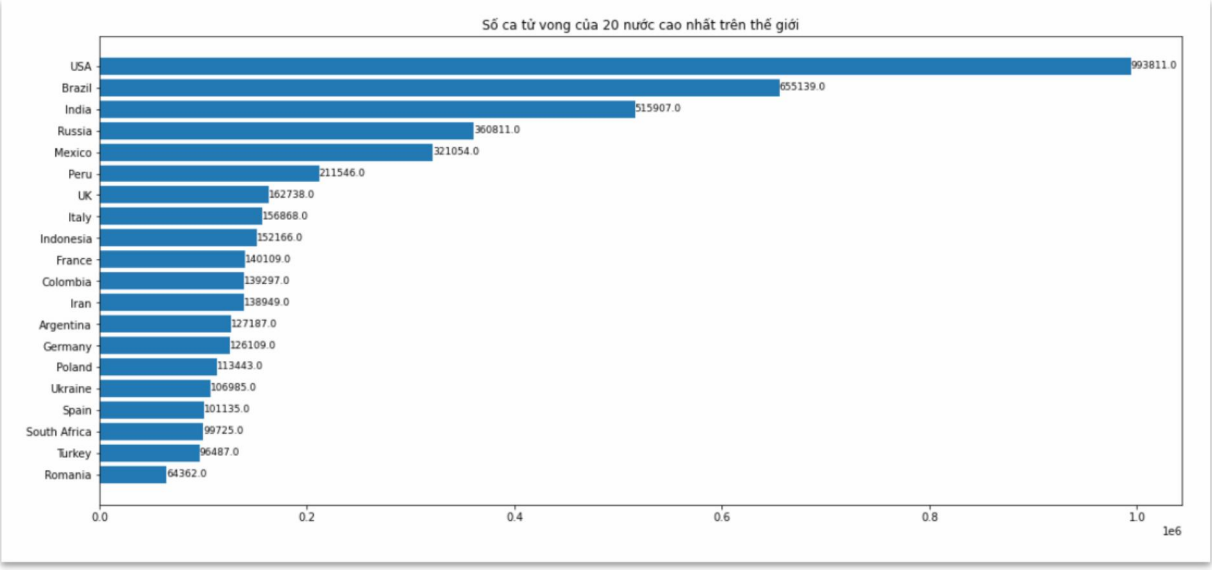
a. Biểu đồ cột ngang

- Lý do chọn biểu đồ

+ Biểu đồ ngang giúp so sánh số lượng hoặc tỉ lệ giữa các yếu tố và điều đặc biệt dùng để cho sắp xếp dữ liệu theo thứ tự, và hiện tại ta muốn biểu diễn số lượng ca nhiễm và tử vong của 20 nước cao nhất theo thứ tự có số lượng cao tới thấp.

- Biểu đồ cột ngang





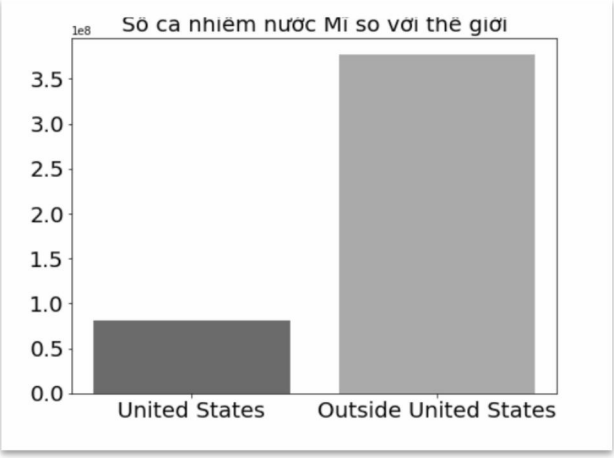
- Ý nghĩa: Mỹ chiếm ngôi vị đầu với ca nhiễm nhiều nhất trên thế giới và gần như gấp đôi so với nước đứng thứ 2 là Ấn độ. Và ta thấy hầu như các nước trong top 20 về số ca nhiễm nhiễm thì cũng nằm trong top20 về tử vong

b. Biểu đồ cột đứng đơn

- Lý do chọn biểu đồ

+ Biểu đồ cột đứng đơn dùng để so sánh thể hiện quy mô, số lượng, sản lượng hoặc khối lượng của các đối tượng. Nó rất phù hợp với mục đích trong trường hợp này, khi ta muốn sử dụng để so sánh *số lượng ca nhiễm* của của 3 nước trong bắc mĩ và tổng các nước còn lại trên thế giới

- Biểu đồ cột dứng đơn



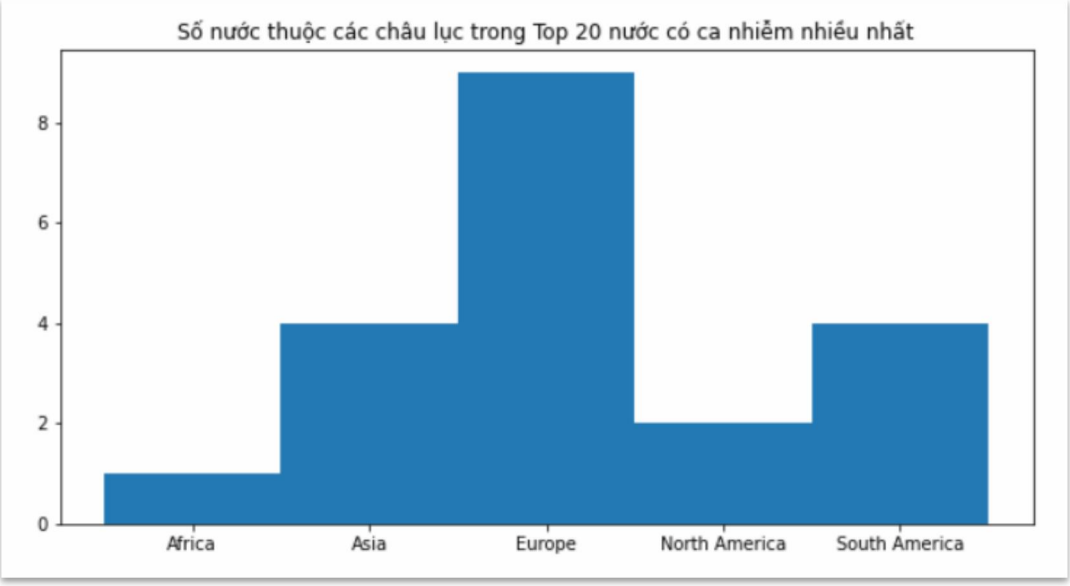
- Ý nghĩa: Ta có thể thấy số ca nhiễm của Hoa Kỳ bằng 1/5 so với thế giới vượt xa so với các nước Bắc Mĩ khác. Nguyên nhân chính chủ yếu là người dân Mĩ không tuân thủ những quy định hướng dẫn phòng chóng Covid của chính phủ. Từ đó có thể thấy nước Mĩ tăng về số ca nhiễm sẽ dẫn đến số ca nhiễm của cả Bắc Mĩ tăng đáng kể

c. Biểu đồ tần suất

- Lý do chọn biểu đồ

  + Histogram là dạng biểu đồ thể hiện tần suất dạng cột và cho thấy hình thấy phân bố dữ liệu, và mục đích của ta là thể hiện tần suất các châu lục xuất hiện trong top 20 về số ca nhiễm.

- Biểu đồ tần suất



- Ý nghĩa: Trong histogram, ta thấy châu âu có nhiều nước nhất trong top20 (9 nước) điều đó cũng bổ sung thêm thông tin làm cho giả thiết 2 đặt ra ở trên càng chắc chắn: "*Do 1 vài nước trong Châu âu phòng chống dịch không tốt và gây ra số lượng ca nhiễm tăng lên dẫn đến tổng số ca nhiễm ở Châu âu tăng theo*"

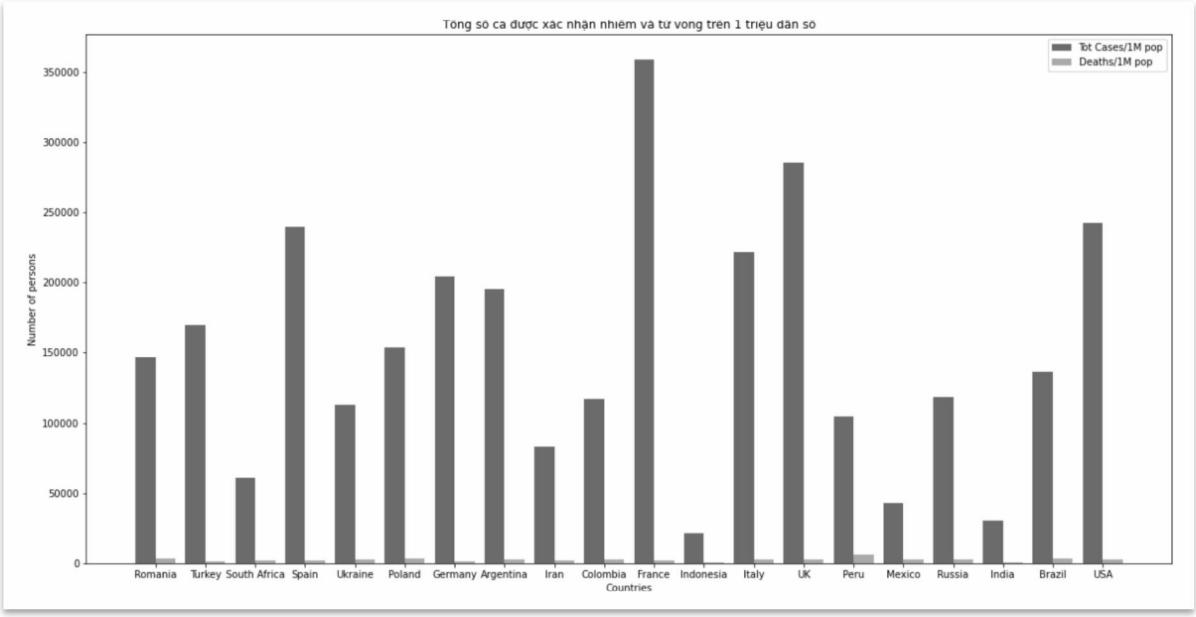
4. <u>Biểu diễn mức độ khỏi bệnh và tử vong của 20 nước có ca nhiễm cao nhất</u>

a. Biểu đồ cột ghép

- Lý do chọn biểu đồ cột ghép

  + Biểu đồ dùng để so sánh thể hiện quy mô, số lượng, sản lượng hoặc khối lượng của nhiều đối tượng trong 1 biểu đồ. Ở đây ta muốn biểu diễn 2 biến là số lượng ca bệnh và số lượng tử vong trong 1 triệu dân số trong 1 nước, đồng thời ta muốn so sánh giữa các nước với nhau.

- Biểu đồ cột



- Ý nghĩa: Tuy nước Mĩ có số ca nhiễm cao nhất và vượt trội so với các nước còn lại như có thể thấy ở trên biểu đồ trên của số người nhiễm và tử vong trên 1 triệu dân số thì nước Mĩ chỉ ở vị trí thứ 4. Nước dẫn đầu là nước Pháp với gần 350000 người nhiễm trên 1 triệu dân số. Lý giải cho vấn đề đó là dân số của nước Mĩ là 329,5 triệu còn của Pháp là 67,39 triệu qua đó khi tính toán về số ca nhiễm trên 1 triệu dân số thì Pháp lại cao hơn.

b. Bản đồ map

- Lý do chọn

  + Bản đồ map cho cái hình tổng quát hơn trên toàn thế giới qua đó hình dung rõ về vấn đề trực quan. Ở đây ta muốn trực quan ca nhiễm hiện tại trên toàn thế giới.

- Bản đồ map

- Ý nghĩa: Dựa vào map của ở trên ta thấy người bị nhiễm Covid-19 trên thế giới rất lớn và phủ khắp trên thế giới cho thấy mức độ lây loan mạnh mẽ của con viruss này

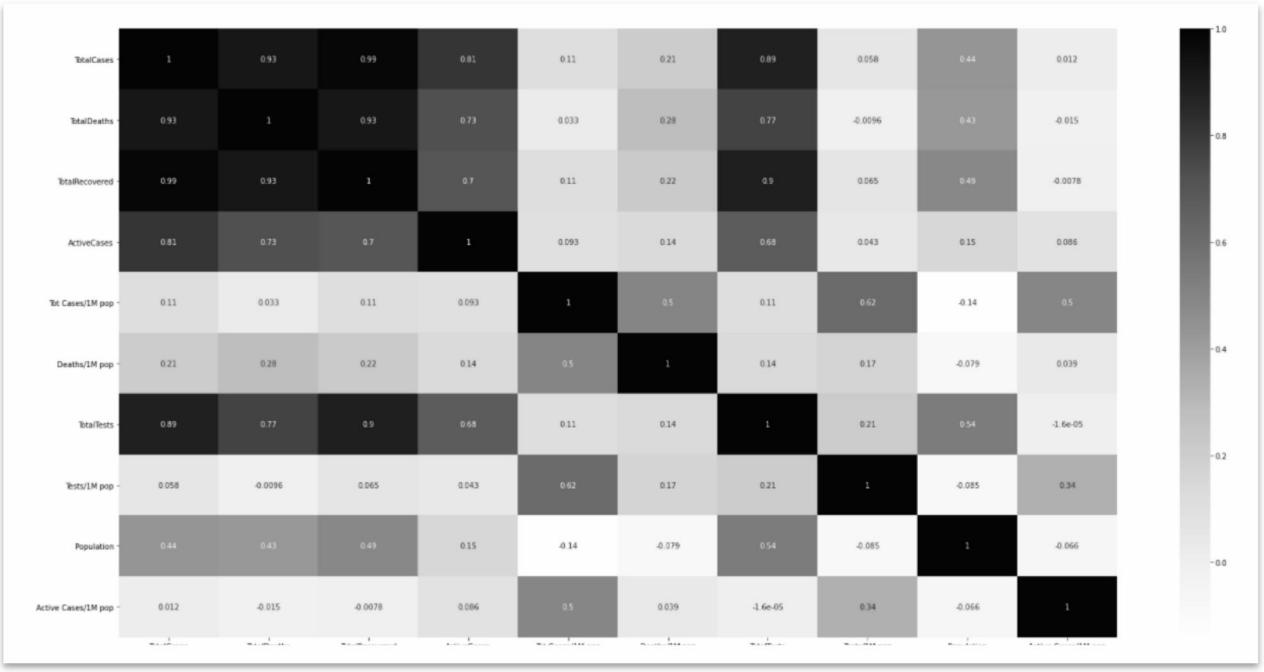
5. <u>Biểu diễn sự tương quan giữa các thuộc tính trong bộ dữ liệu.</u>

a. Bản đồ nhiệt

- Lý do chọn bản đồ nhiệt

  + HeatMap biểu diễn các thông số tương quan 2 các cặp thuôc tính qua đó ta có thể xem được 2 thuộc tính có mối quan hệ nào không. Ở đây ta muốn xem thử trong bộ dữ liệu có các cột thuộc tính qua sử dụng heatmap xem thử, có cặp nào có hệ số tương quan cao hay không.

- Bản đồ nhiệt



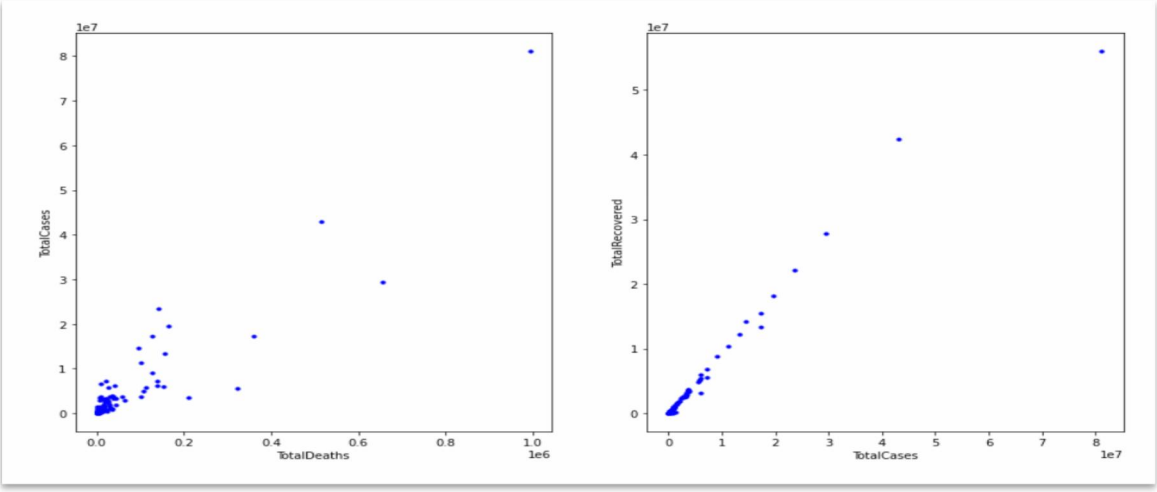
- Ý nghĩa: Dựa vào heatMap ta thấy bộ 3 TotalCases, TotalDeaths và TotalRecovered có mối quan hệ tương quan chặc chẽ với nhau với hệ số tương quan cao ngất ngưỡng.

b. Biểu đồ phân tán

- Lý do chọn biểu đồ phân tán

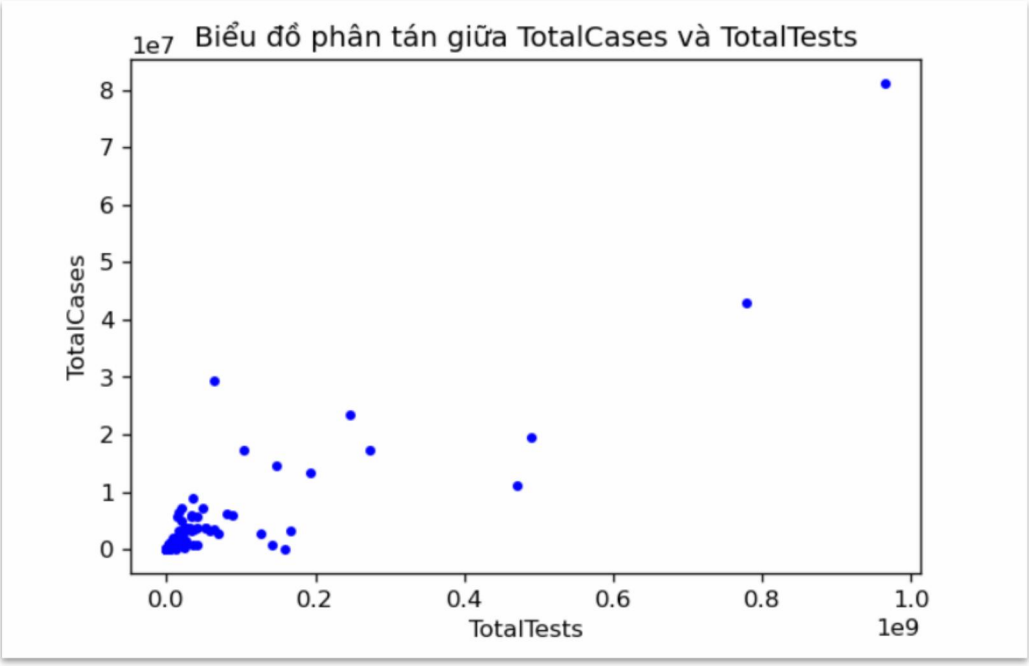
  + Biểu đồ thể hiện mối tương quan giữa nguyên nhân và kết quả hoặc giữa các yếu tố ảnh hưởng đến chất lượng. Ở đây ta dùng biểu đồ để thể mối tương quan giữa 2 cột thuộc tính

- Biểu đồ phân tán (1)



- Ý nghĩa: Ở 2 biểu đồ ta dễ dàng nhận ra khi các cặp thuộc tính gần như có mối quan hệ tuyến tính. Khi thuộc tích `TotalCases` tăng thì `TotalRecovred` và `TotalDeaths` cũng tăng theo. Cũng dễ hiểu ở tình huống này vì khi 1 người bị nhiễm bệnh chỉ xảy ra 1 trong 2 trường hợp khỏi bệnh hoặc không qua khỏi.

- Biểu đồ phân tán (2)



- Ý nghĩa: Ta có hệ số tương quan 0.89 khá cao và nhìn vào biểu đồ scatter ta thấy có mối quan hệ đồng biến. Khi ta tăng TotalTests thì qua đó TotalCases đồng thời tăng theo. Điều đó có nghĩa là khi người được

test covid-19 tăng lên qua đó phát hiện ra nhiều người bị mắc Covid-19 trong cộng đồng. Đồng thời cũng giải thích lý do tại sao ở Việt Nam khi dịch bùng phát mạnh ở thành phố hồ chí mình thì bộ y tế đã ra sức tạo ra nhiều đợt test Covid-19 ở mỗi địa phương, qua đó hi vọng có thể phát hiện sớm người mắc Covid-19 để có thể dập dịch tốt hơn.

❖ Tổng Kết:

- Độ nguy hiểm và tốc độ lây loan nhanh của Covid 19 vượt trội so với 3 loại dịch bệnh trước đây.
- Châu Âu có số ca nhiễm cao nhất chiếm hơn 1/3 thế giới và lý do Châu Âu có số ca nhiễm cao như vậy là do trong châu âu có 8 nước nằm trong top20 có số ca nhiễm cao nhất điều đó dẫn tới tổng số ca nhiễm ở cả Châu Âu tăng cao
- Nước Mĩ có số ca nhiễm và tử vong cao nhất thế giới,và vượt trột với các nước còn lại trên thế giới
- Tuy Mĩ có số ca nhiễm cao nhất thế giới nhưng Pháp lại là top1 về số ca nhiễm trên 1 triệu dân số
- Tồn tại mối quan hệ nhân quả: khi số lượng ca nhiễm tăng thì số lượng ca tử vong và hồi phục đều tăng theo
- Việt Nam giai đoạn đầu trong bùng phát dịch bệnh trên thế giới là nằm trong nhóm quốc gia phòng chống dịch tốt nhất nhưng dựa vào bảng số liệu ngày 13/03/2022 thì Việt Nam lại nằm trong Top20 nước có số ca nhiễm cao nhất trên thế giới.

IV. Phân chia công việc

Tên thành viên	Nhiệm vụ	hoàn thành(%)
Lê Tiến Trí	- Trực quan hóa đơn biến - Viết report	100%
Lê Hoàng Thịnh Phước	- Xử lí dữ liệu - Trực quan hóa đa biến	100%
Hoàng Minh Đức	- Cào dữ liệu	100%

	- Trực quan hóa đơn biến	