BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Khoa CNTT CLC

DATA VISUALIZATION

Lab 1: Data Relationship

|Giáo viên hướng dẫn|

Bùi Tiến Lên

Lữ Thế Vỹ - 19127009

Nguyễn Ngọc Uyên Trang - 19127074

Mạc Văn Hưng - 19127416

THÀNH PHỐ HỒ CHÍ MINH - THÁNG 03 NĂM 2022

MỤC LỤC

Mục 1: Mức độ hoàn thành

Sinh viên	Yêu cầu	Mức độ hoàn thành
Lữ Thế Vỹ	Thu thập số liệu thống kê từ api Disease, sử dụng code/thuật toán để thống kê, trực quan hoá	100%
Mạc Văn Hưng	Thu thập số liệu thống kê từ api Novel Covid 19, sử dụng code/thuật toán để thống kê, trực quan hoá	100%
Nguyễn Ngọc Uyên Trang	Thu thập số liệu thống kê từng ngày trên worldometers, sử dụng code/thuật toán để thống kê, trực quan hoá	100%

Mục 2: Mô tả dữ liệu

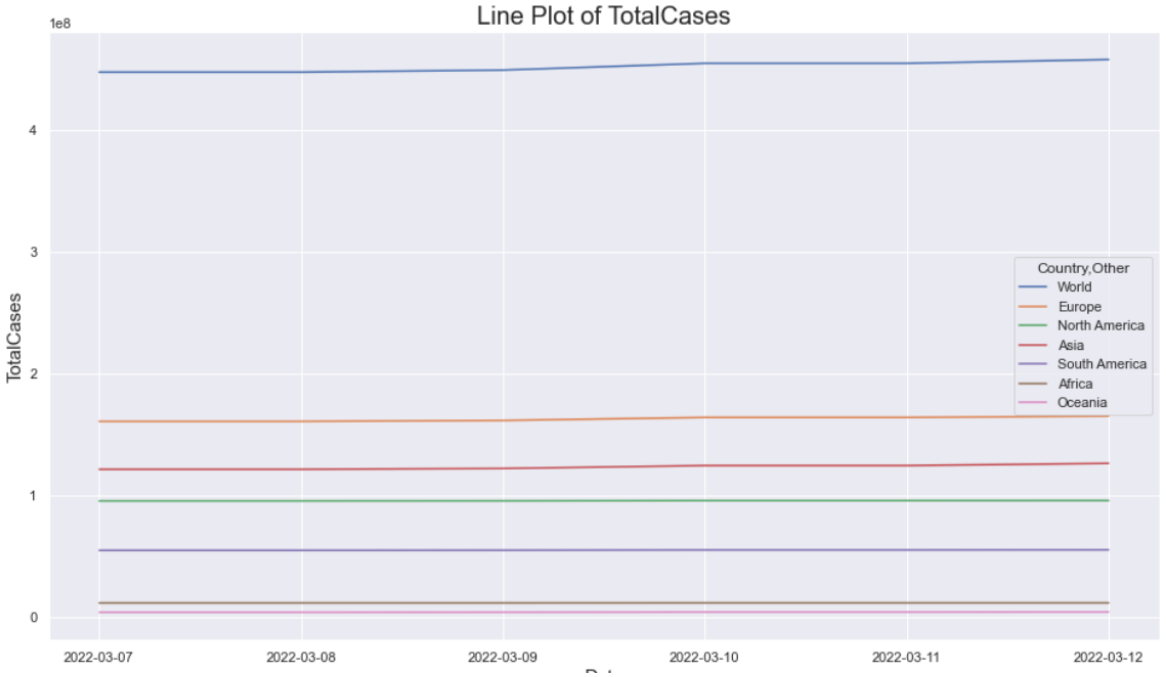
1. **Nguồn gốc dữ liệu**
    - Dữ liệu HTML được lấy từ trang web của Tổ chức Worldometer (www.worldometers.info), được thu thập từ ngày 07/03/2022 đến ngày 12/03/2022
    - Dữ liệu API được lấy từ các trang web api: Novel Covid 19 (ở trên postman) và Disease. Cả 2 trang web này đều dựa vào trang web của Tổ chức Worldometer và các trang web khác mà được tạo ra. Trường dữ liệu API đa số giống với Worldometer nhưng có thêm nhiều thông số khác.
2. **Ý nghĩa của các trường dữ liệu Worldometer**
    - Country, Other: Quốc gia hoặc các châu lục
    - Total Cases: Tổng số ca nhiễm Covid 19 được phát hiện tính tới ngày thu thập
    - New Cases: Số ca nhiễm mới được ghi nhận trong ngày
    - Total Deaths: Tổng số ca tử vong tích lũy trong số các trường hợp được phát hiện tính tới ngày thu thập
    - New Deaths: Số ca tử vong được ghi nhận trong ngày
    - Total Recovered: Tổng số ca hồi phục tích lũy trong số các trường hợp được phát hiện tính tới ngày thu thập
    - New Recoverd: Số ca hồi phục được ghi nhận trong ngày
    - Active Cases: Số ca đang được điều trị. Active Cases = Total Cases - Total Deaths - Total Recoverd
    - Serious, Critical: Số ca nhiễm nghiêm trọng hoặc trở nặng tích lũy trong số các trường hợp được phát hiện tính tới ngày thu thập
    - TotCases/1M pop: Tỉ lệ người nhiễm Covid 19 so với dân số quốc gia. TotCases/1M pop = (Total Cases / Popular) * 1000
    - Deaths/1M pop: Tỉ lệ người tử vong do Covid 19 so với dân số quốc gia. Deaths/1M pop = (Total Deaths / Popular) * 1000
    - Total Tests: Tổng số ca nghi nhiễm Covid 19 tích lũy trong số các trường hợp được phát hiện tính tới ngày thu thập
    - Tests/1M pop: Tỉ lệ các ca nghi nhiễm Covid 19 so với dân số quốc gia. Tests/1M pop = (Total Tests / Popular) * 1000
    - Population: Dân số của một quốc gia
3. **Ý nghĩa của các trường dữ liệu API**
    - updated: mã cập nhật mới nhất của quốc gia đó
    - country: tên quốc gia
    - countryInfo: thông tin của quốc gia
    - cases: tổng số ca nhiễm Covid 19 được phát hiện tính tới ngày thu thập

- todayCases: số ca nhiễm mới được ghi nhận trong ngày
- deaths: tổng số ca tử vong tính tới ngày thu thập
- todayDeaths: số ca tử vong được ghi nhận trong ngày
- recovered: tổng số ca hồi phục được phát hiện tính tới ngày thu thập
- todayRecovered: số ca hồi phục được ghi nhận trong ngày
- active: số ca nhiễm đang được điều trị
- critical: số ca nhiễm nặng hoặc nghiêm trọng
- casesPerOneMillion: tỉ lệ người nhiễm Covid 19 so với dân số quốc gia
- deathsPerOneMillion: tỉ lệ người tử vong so với dân số quốc gia
- tests: tổng số ca nghi nhiễm Covid 19 được phát hiện tính tới ngày thu thập
- testsPerOneMillion: tỉ lệ người nghi nhiễm Covid 19 so với dân số quốc gia
- population: dân số của quốc gia
- continent: châu lục mà quốc gia này trực thuộc
- oneCasePerPeople: tổng số lượng người trong đó có 1 ca nhiễm Covid 19
- oneDeathPerPeople: tổng số lượng người trong đó có 1 ca tử vong
- oneTestPerPeople: tổng số lượng người trong đó có 1 ca nghi nhiễm
- activePerOneMillion: tỉ lệ người đang được điều trị so với dân số quốc gia
- recoveredPerOneMillion: tỉ lệ người hồi phục so với dân số quốc gia
- criticalPerOneMillion: tỉ lệ người nhiễm Covid 19 nặng hoặc nghiêm trọng so với dân số quốc gia

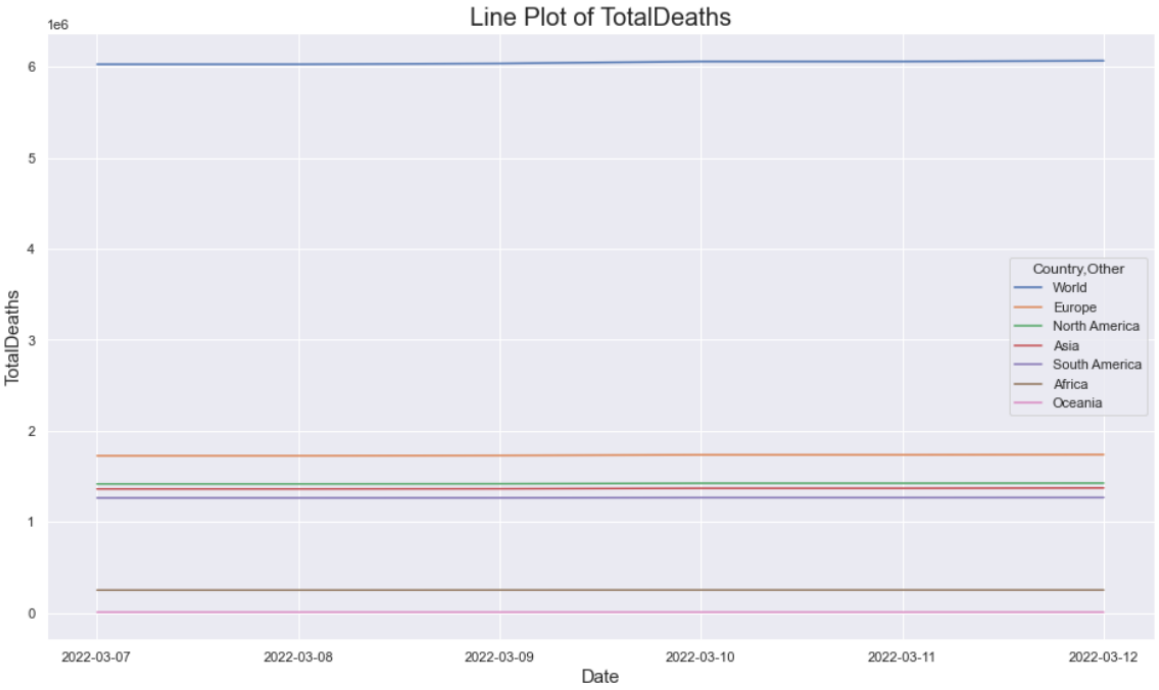
Mục 3: Trực quan hoá dữ liệu

1. Biểu đồ đường

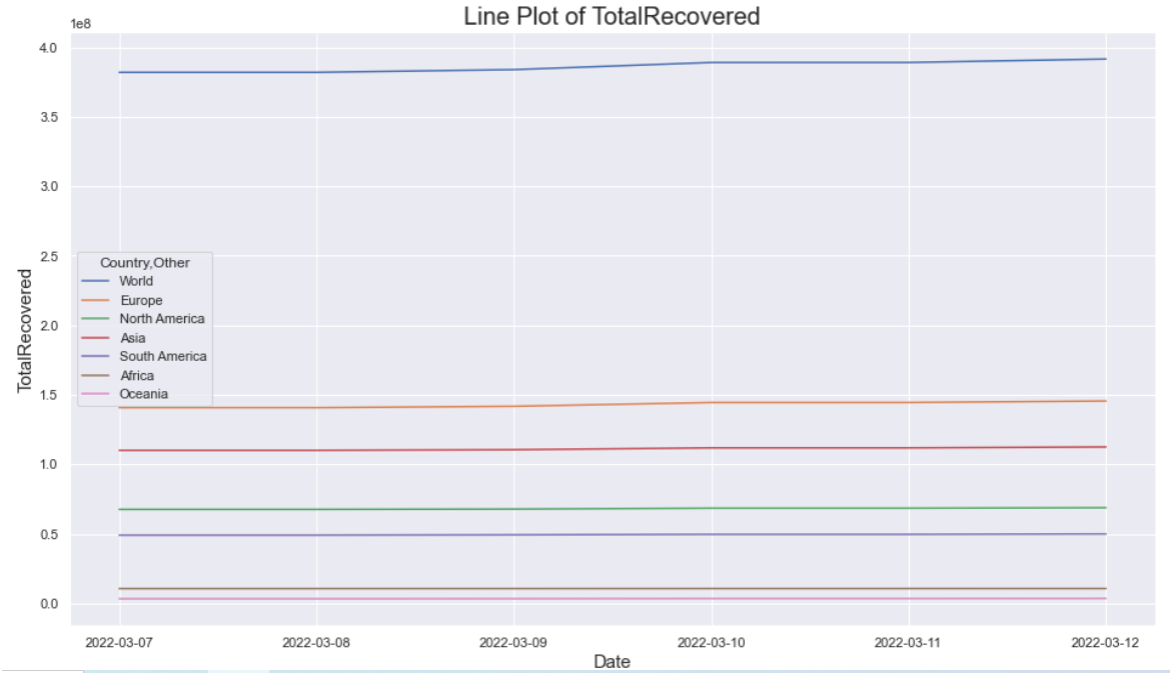
- Đầu tiên, ta sử dụng biểu đồ đường (line chart) để so sánh sự gia tăng của dịch bệnh Covid 19 ở các châu lục nói riêng và cả Thế giới nói chung.
- Để có các nhìn tổng quan về dịch bệnh, ta sẽ chọn 3 trường Total Cases, Total Deaths và Total Recoverd để trực quan hóa.
- Biểu đồ đường sẽ cho ta thấy cái nhìn trực quan về tổng số ca nhiễm bệnh, tổng số ca tử vong và tổng số ca hồi phục trong thời gian 6 ngày, từ ngày 07/03/2022 đến ngày 12/03/2022.



H1. Biểu đồ so sánh sự gia tăng số ca nhiễm bệnh



H2. Biểu đồ so sánh sự gia tăng số ca tử vong

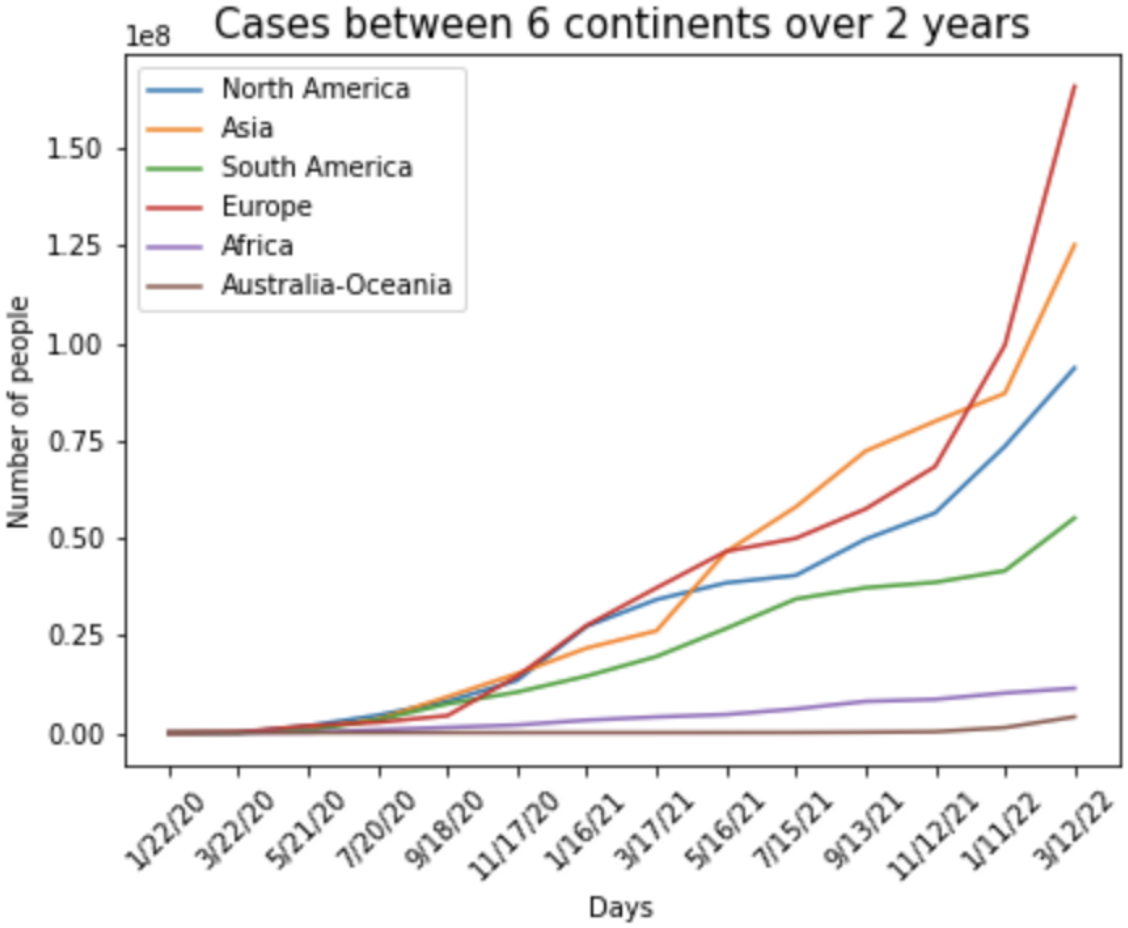


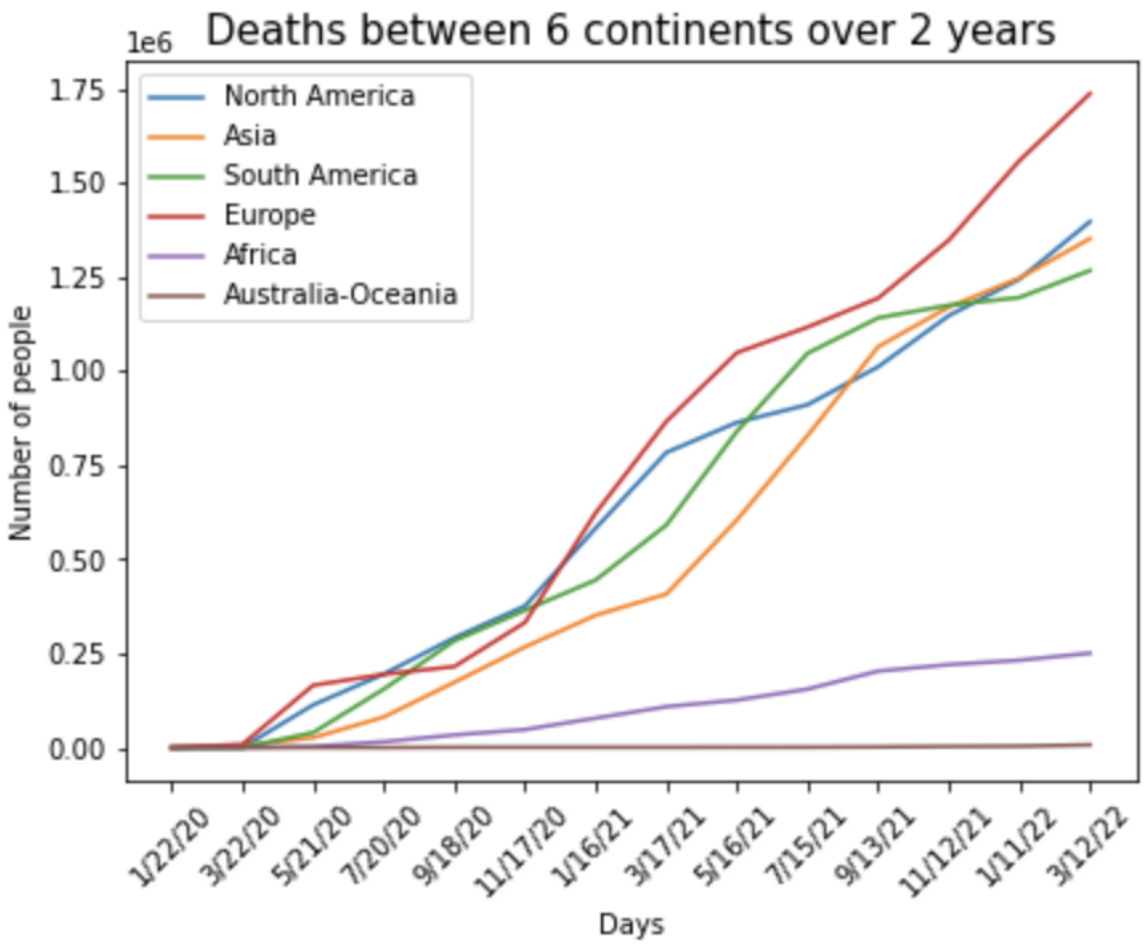
H3. Biểu đồ so sánh sự gia tăng số ca hồi phục

- Trên đây là ba biểu đồ đường so sánh sự gia tăng của 1 trường dữ liệu nào đó giữa Thế giới và các lục địa, và giữa các lục địa với nhau. Trục thẳng đứng là số liệu của trường dữ liệu TotalCases/TotalDeaths/TotalRecovered. Trục nằm ngang là các ngày thu thập dữ liệu.
- Dựa vào 3 biểu đồ, ta có thể thấy số ca nhiễm bệnh và số ca hồi phục của cả thế giới nói chung và của các châu lục nói riêng có sự gia tăng rõ rệt, còn số ca tử vong dường như không thay đổi. Ta có thể lý giải điều này

là do hiện nay xuất hiện nhiều biến chủng mới, có khả năng lây lan nhanh hơn, tuy nhiên chính phủ các nước đã cho người dân tiêm vaccine ít nhất là 2 mũi nên tỉ lệ hồi phục cao hơn. Ta sẽ sử dụng tập dữ liệu của ngày thu thập gần nhất, tức ngày 12/03/2022, để trực quan hóa sự diễn biến dịch bệnh Covid 19 của cả Thế giới cũng như các châu lục.

- Tiếp theo, ta sử dụng biểu đồ đường (line chart) để so sánh sự gia tăng rõ hơn của dịch bệnh Covid 19 giữa các châu lục.
- Để có các nhìn tổng quan về dịch bệnh, ta sẽ chọn 2 trường Cases, Deaths để trực quan hóa.
- Biểu đồ đường sẽ cho ta thấy cái nhìn trực quan về tổng số ca nhiễm bệnh, tổng số ca tử vong trong thời gian khoảng 2 năm, từ ngày 22/01/2020 đến ngày 12/01/2022.

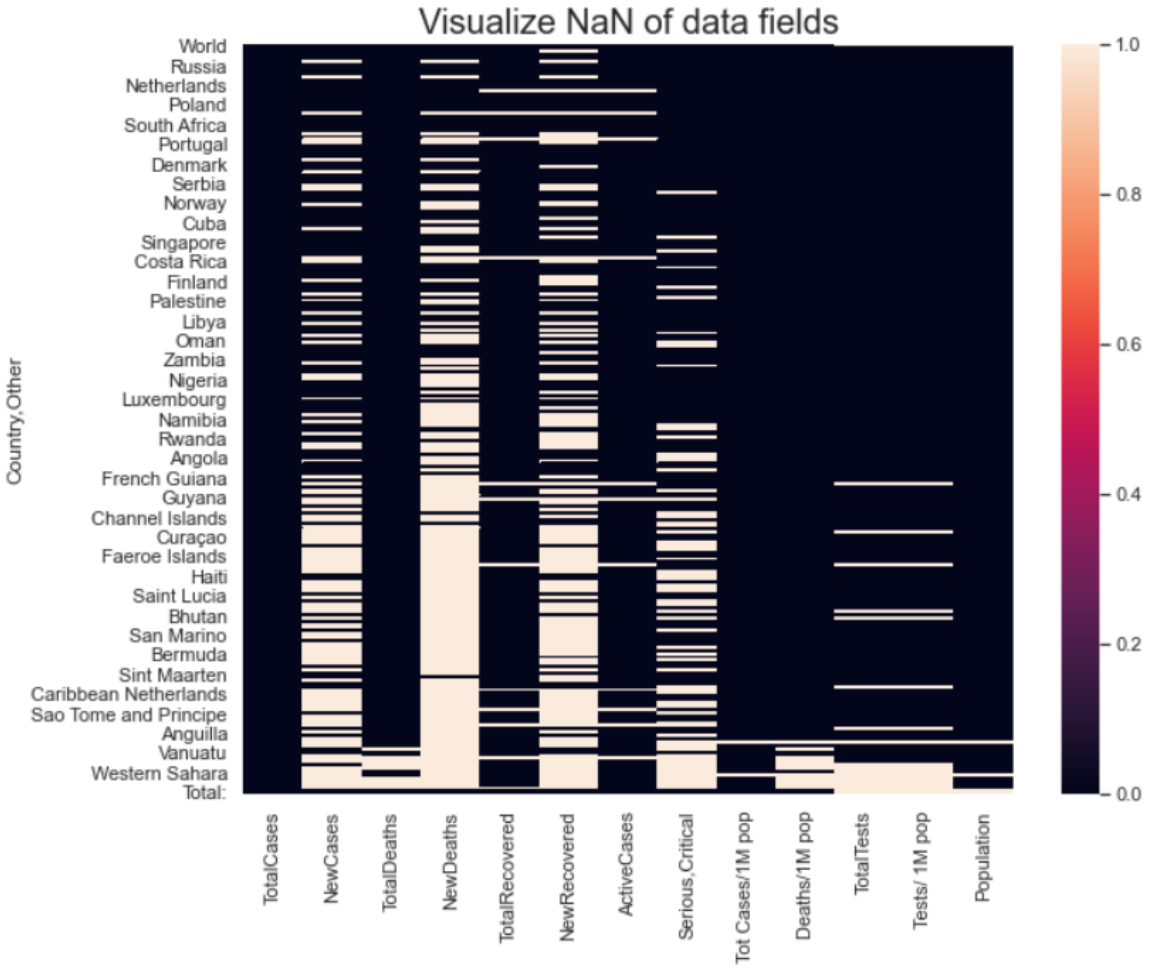




- Cả 2 trường Cases và Deaths đều cho thấy sự gia tăng rõ rệt của các châu lục.
- Trong đó, Europe (Châu Âu) gia tăng nhiều nhất, tiếp theo là Asia (Châu Á), North America (Bắc Mỹ), South America (Nam Mỹ), Africa (Châu Phi) và cuối cùng là Australia-Oceania (Châu Úc-Châu Đại Dương).

2. Bản đồ nhiệt

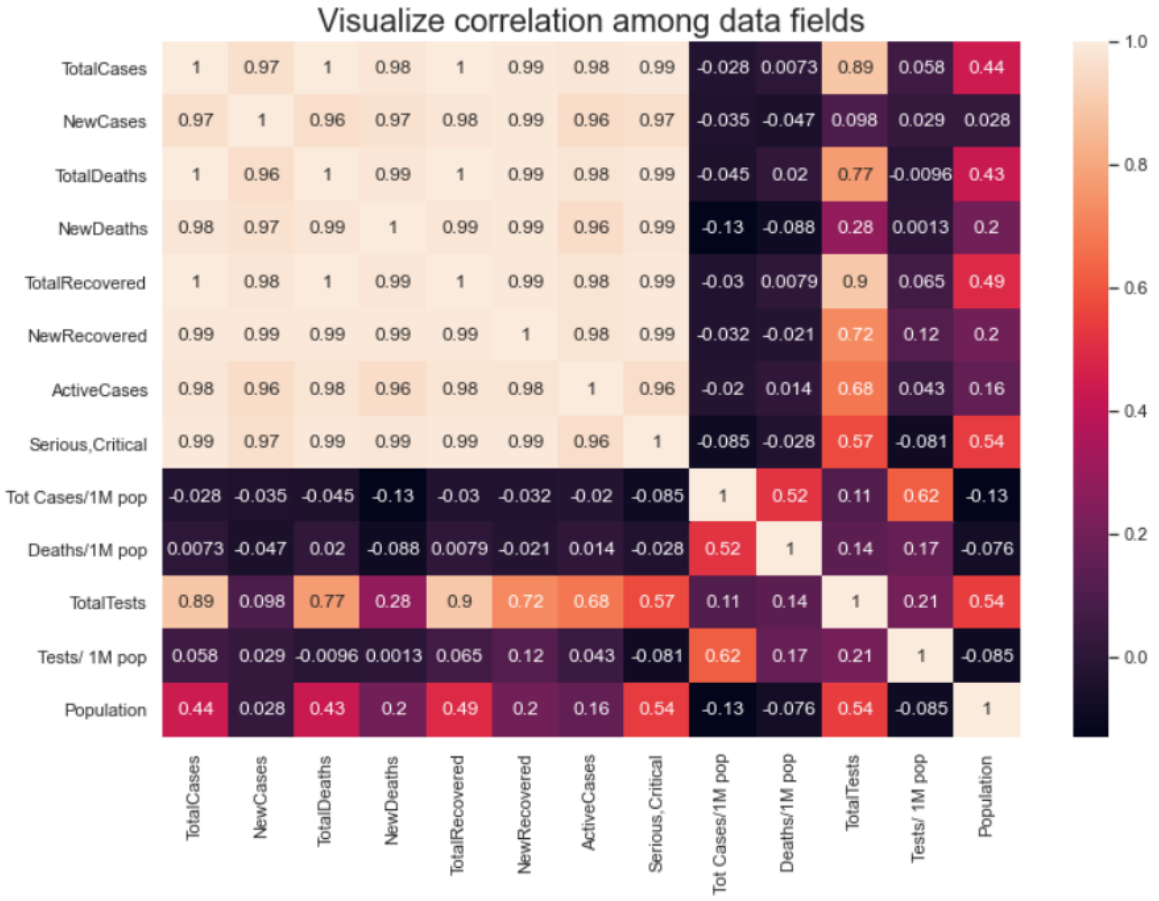
a. Trực quan hóa giá trị NaN của các trường dữ liệu



H4. Biểu đồ trực quan hóa giá trị NaN trong tập dữ liệu

- Trên đây là bản đồ nhiệt trực quan hóa các giá trị NaN của các trường dữ liệu trong tập dữ liệu tất cả các quốc gia trên thế giới vào ngày gần nhất (12/03/2022). Trục thẳng đứng là tên của các quốc gia. Trục nằm ngang là các trường dữ liệu.
- Nhìn vào bản đồ, ta có thể thấy trường NewDeaths có số lượng giá trị NaN nhiều nhất. Ta có thể lý giải điều này là do các nước đã thực hiện tiêm vaccine cho người dân, vì vậy số ca tử vong của các quốc gia đã giảm đáng kể, cùng với đó, số ca nhiễm mới (NewCases) cũng giảm rõ rệt.
- Các giá trị NaN xuất hiện nhiều nhất ở các quốc gia châu Phi. Điều này cho thấy dịch Covid 19 không diễn biến mạnh ở các quốc gia châu Phi.

b. Trực quan hóa sự tương quan giữa các trường dữ liệu

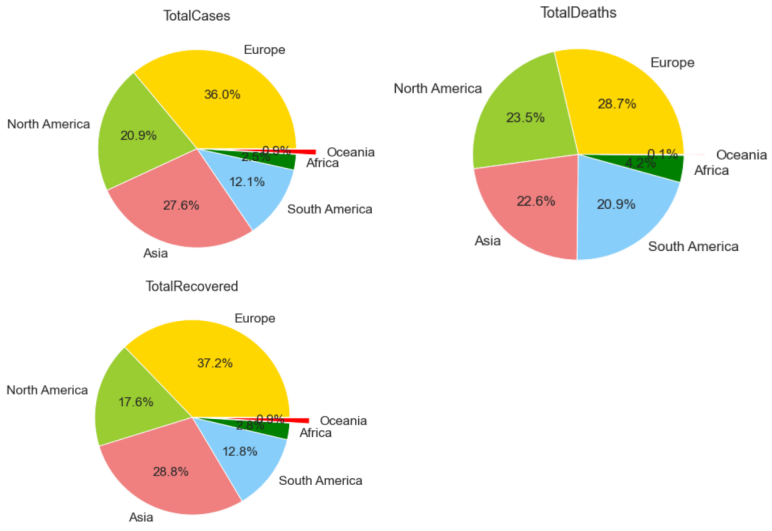


H5. Biểu đồ thể hiện sự tương quan giữa các trường dữ liệu

- Trên đây là bản đồ nhiệt thể hiện sự tương quan giữa các trường dữ liệu của tập dữ liệu tất cả các quốc gia trên thế giới vào ngày thu thập gần nhất (12/03/2022).

Từ biểu đồ ta có thể thấy trường TotalCases có sự tương quan với các trường NewCases, TotalDeaths, NewDeaths, TotalRecovered, NewRecovered, ActiveCases và Serious, Critical. Điều này hoàn toàn hợp lý.

3. Biểu đồ tròn

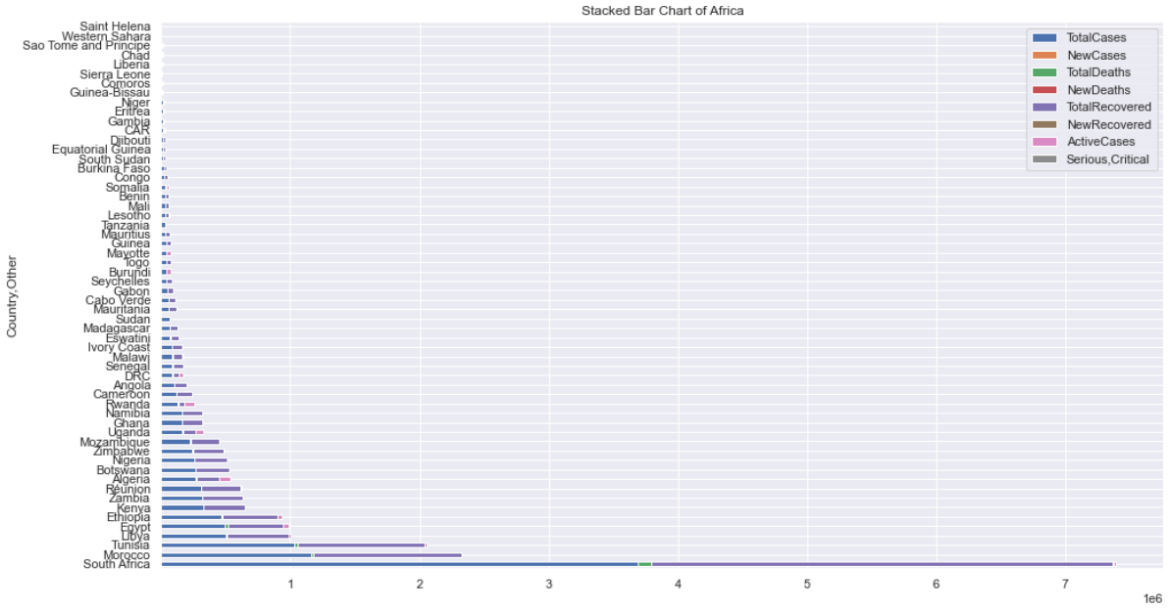


H6. Biểu đồ so sánh số liệu giữa các châu lục về 1 số trường nhất định

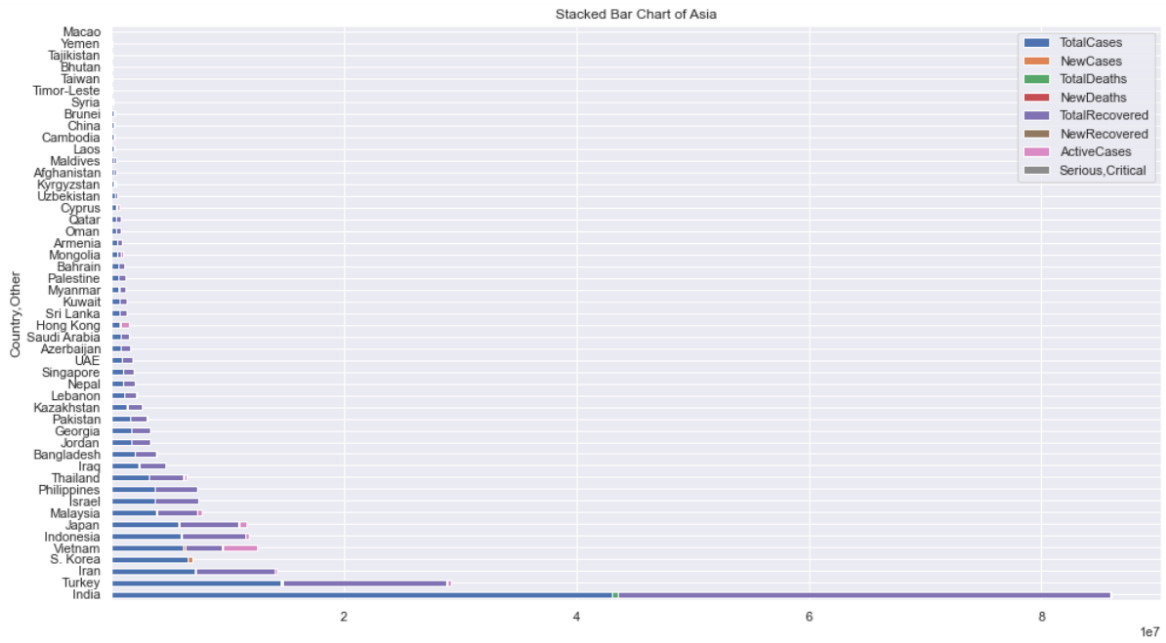
- Trên đây là ba biểu đồ so sánh số ca nhiễm bệnh (TotalCases), số ca tử vong (TotalDeaths) và số ca hồi phục (TotalRecovered) giữa các châu lục.
- Từ 3 biểu đồ trên ta có thể thấy, số ca nhiễm Covid 19 (TotalCases), số ca tử vong (TotalDeaths) và số ca hồi phục (TotalRecovered) ở các quốc gia châu Âu là nhiều hơn cả. Các quốc gia thuộc châu Đại dương có số ca nhiễm bệnh (TotalCases), số ca tử vong (TotalDeaths) và số ca hồi phục (TotalRecovered) ít nhất.

4. Biểu đồ cột chồng

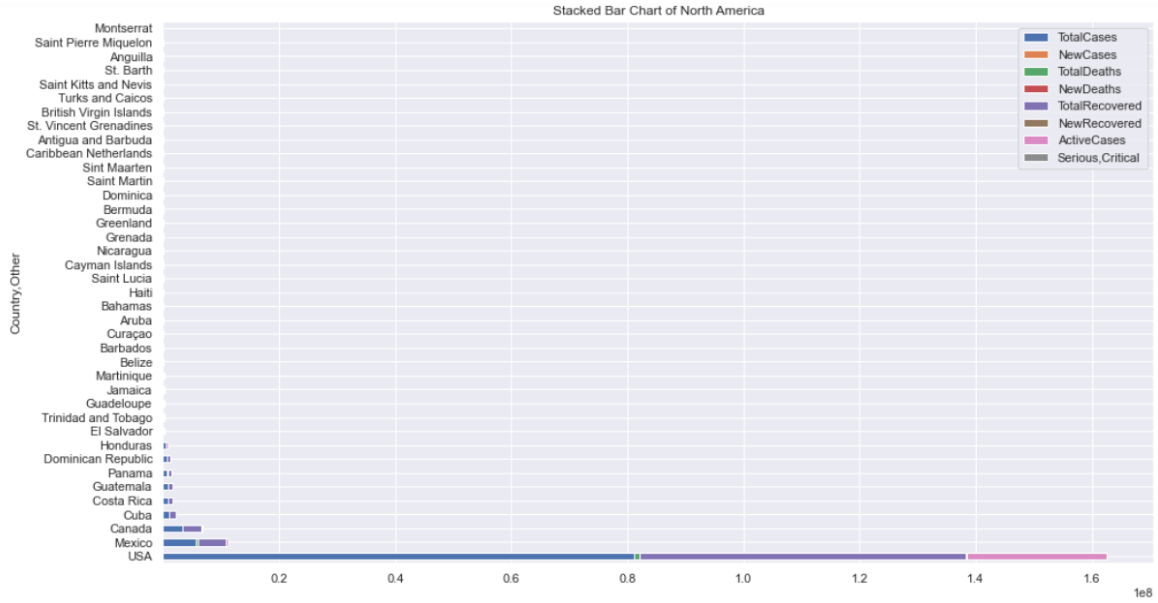
- Dưới đây là các biểu đồ cột chồng so sánh số liệu giữa các quốc gia trong cùng 1 châu lục về 1 số trường được chỉ định. Trục thẳng đứng là tên các quốc gia. Trục nằm ngang là số liệu thu thập được.



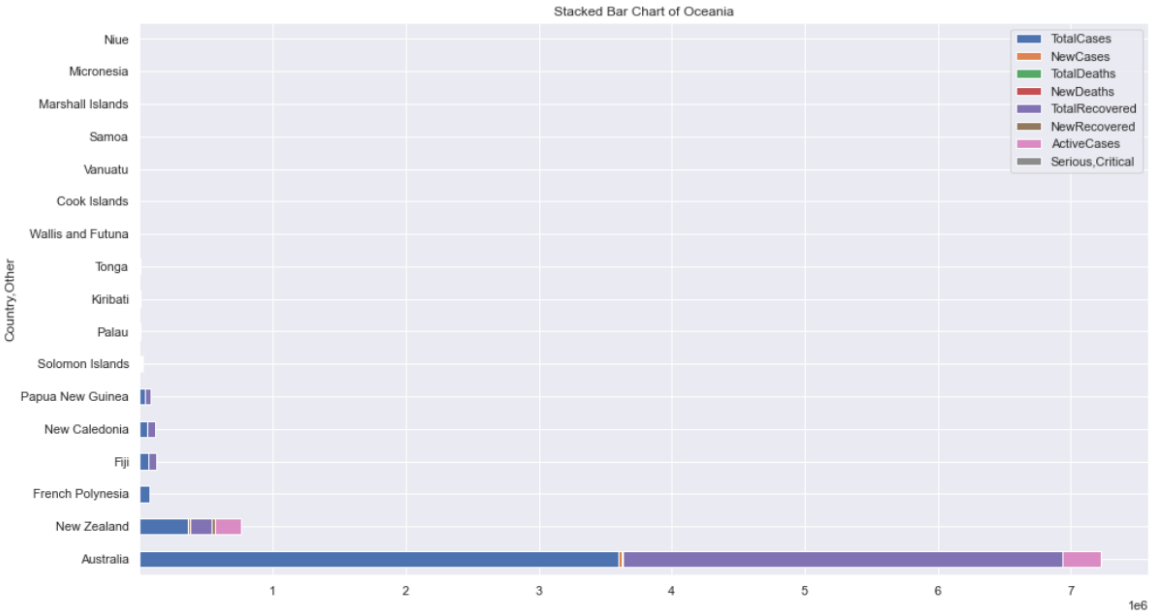
H7. Biểu đồ so sánh số liệu giữa các quốc gia châu Phi về 1 số trường nhất định



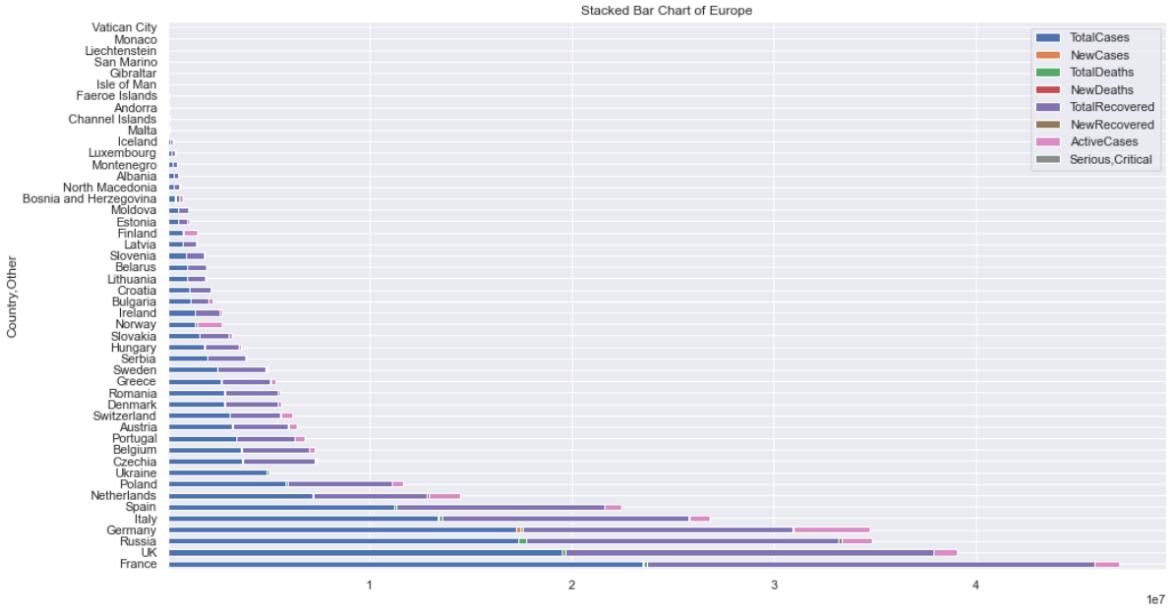
H8. Biểu đồ so sánh số liệu giữa các quốc gia châu Á về 1 số trường nhất định



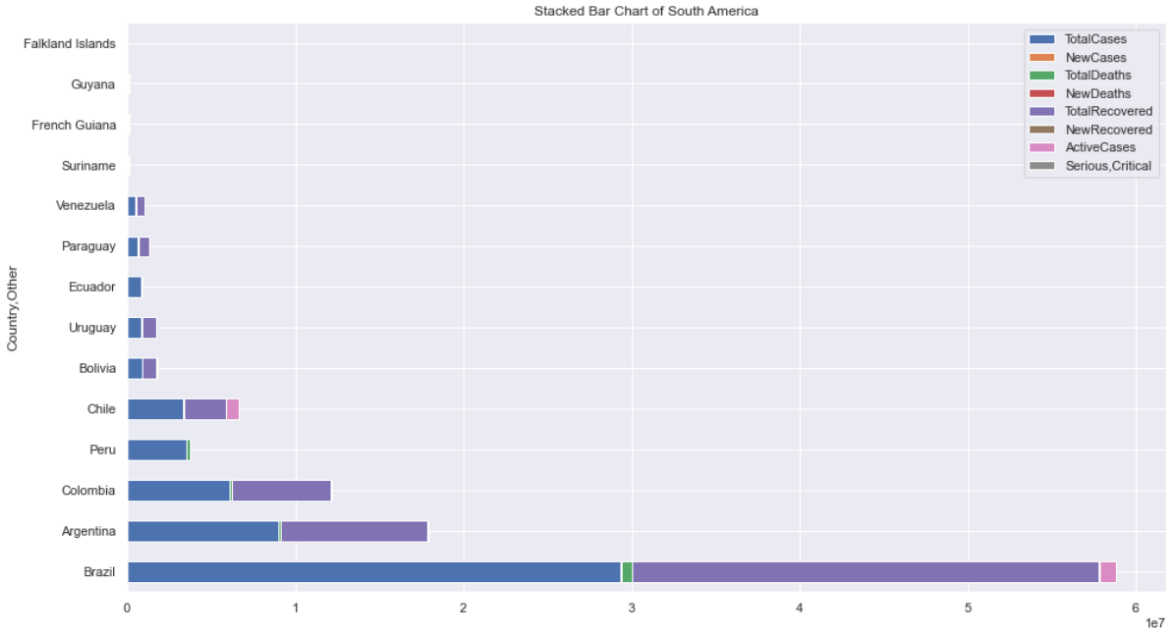
H9. Biểu đồ so sánh số liệu giữa các quốc gia Bắc Mỹ về 1 số trường nhất định



H10. Biểu đồ so sánh số liệu giữa các quốc gia châu Úc về 1 số trường nhất định



H11. Biểu đồ so sánh số liệu giữa các quốc gia châu Âu về 1 số trường nhất định



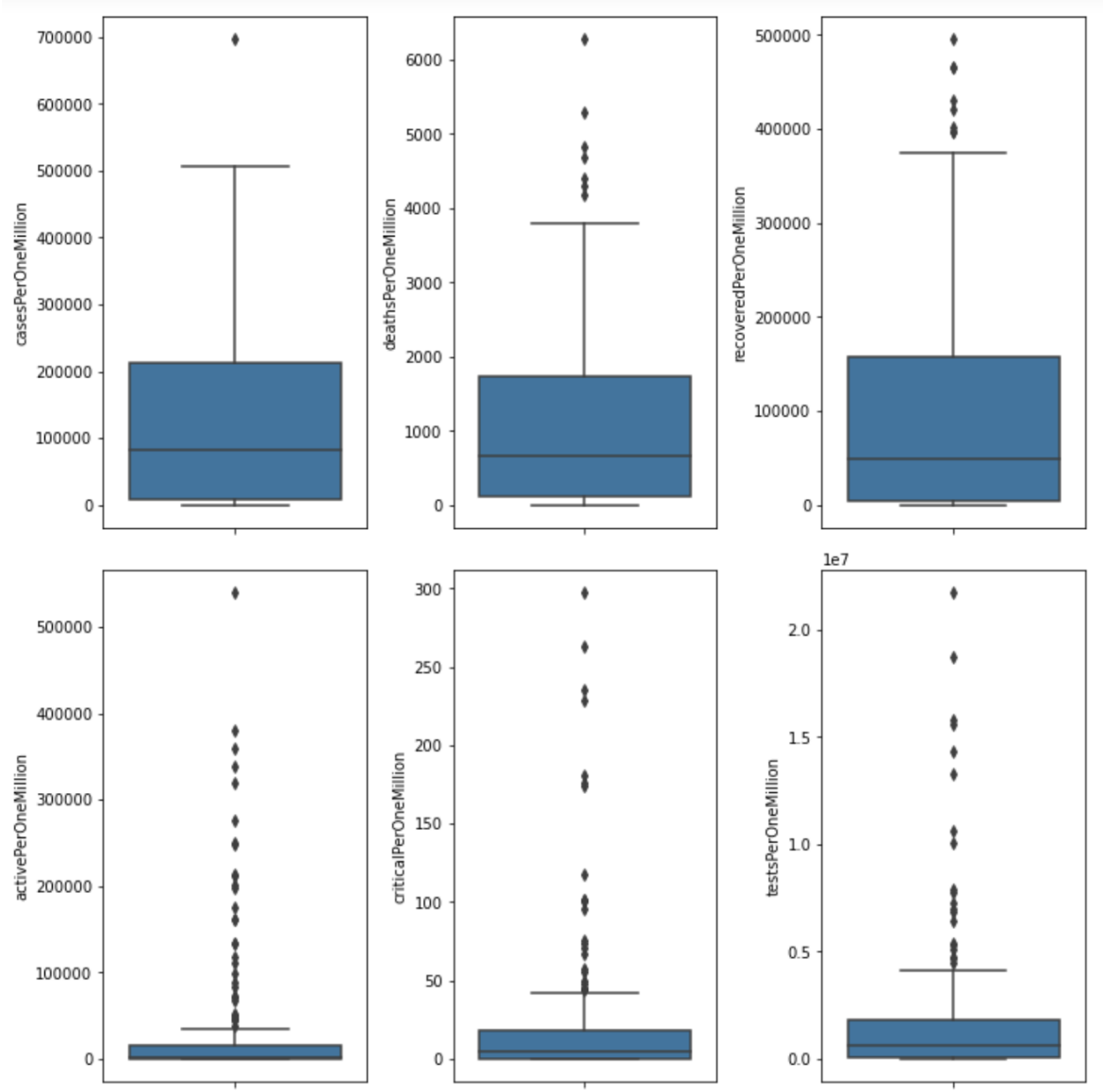
H12. Biểu đồ so sánh số liệu giữa các quốc gia Nam Mỹ về 1 số trường nhất định

- Nhìn chung, dịch bệnh Covid 19 có sự phân bố không đồng đều giữa các quốc gia trong cùng 1 châu lục. Cụ thể, các quốc gia có nền kinh tế phát triển hơn cả trong 1 châu lục sẽ có tình hình dịch bệnh Covid 19 diễn biến phức tạp hơn.

5. Biểu đồ hộp

- Ta sử dụng biểu đồ hộp để biểu diễn sự phân bố so sánh các quốc gia trên thế giới, đồng thời biểu diễn được các outliers của từng trường dữ liệu.

- Ta sẽ chọn các trường casesPerOneMillion, deathsPerOneMillion, recoveredPerOneMillion, activePerOneMillion, criticalPerOneMillion và testsPerOneMillion để trực quan hoá.

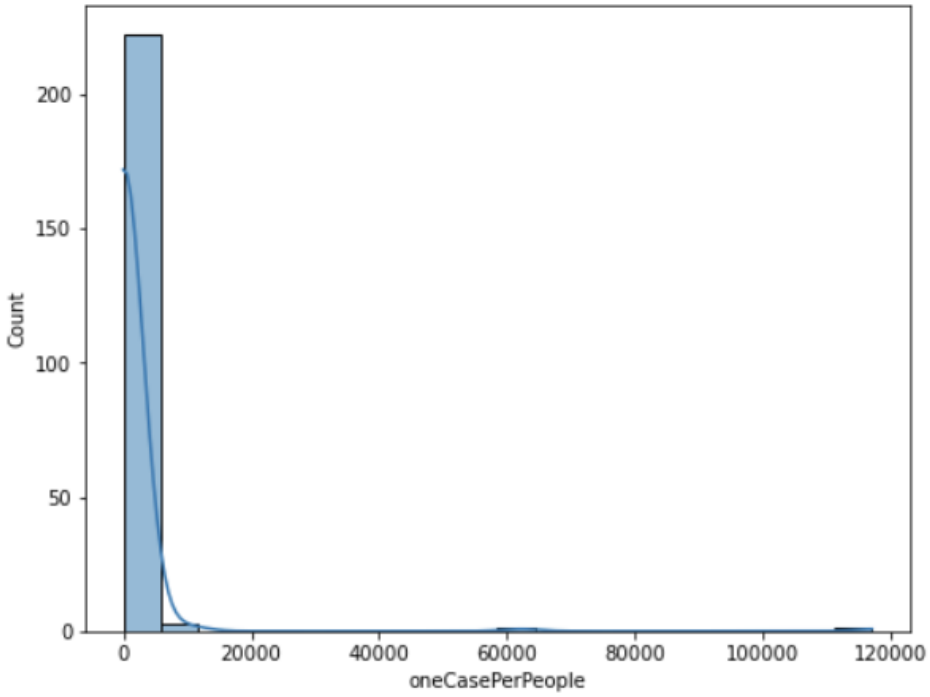


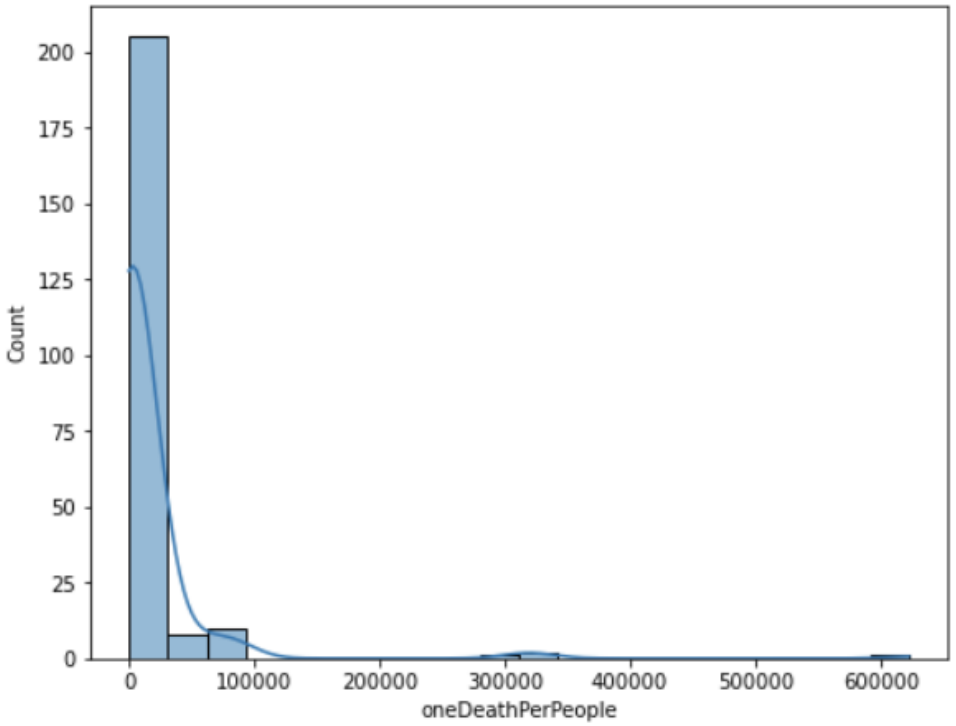
- Cả 6 trường dữ liệu kể trên đều có outliers => Mức độ Covid 19 của quốc gia đó rất cao so với mức trung bình.
- casesPerOneMillion chủ yếu tập trung ở mức khoảng 100000 người trong 1 triệu người, có vài outliers ở mức 700000.
- deathsPerOneMillion chủ yếu tập trung ở mức khoảng 7000 người trong 1 triệu người, có vài outliers nằm rải rác từ 4200 đến 6300 người.
- recoveredPerOneMillion chủ yếu tập trung ở mức khoảng 700000 người trong 1 triệu người, có vài outliers ở mức rất cao 10 triệu đến 25 triệu.

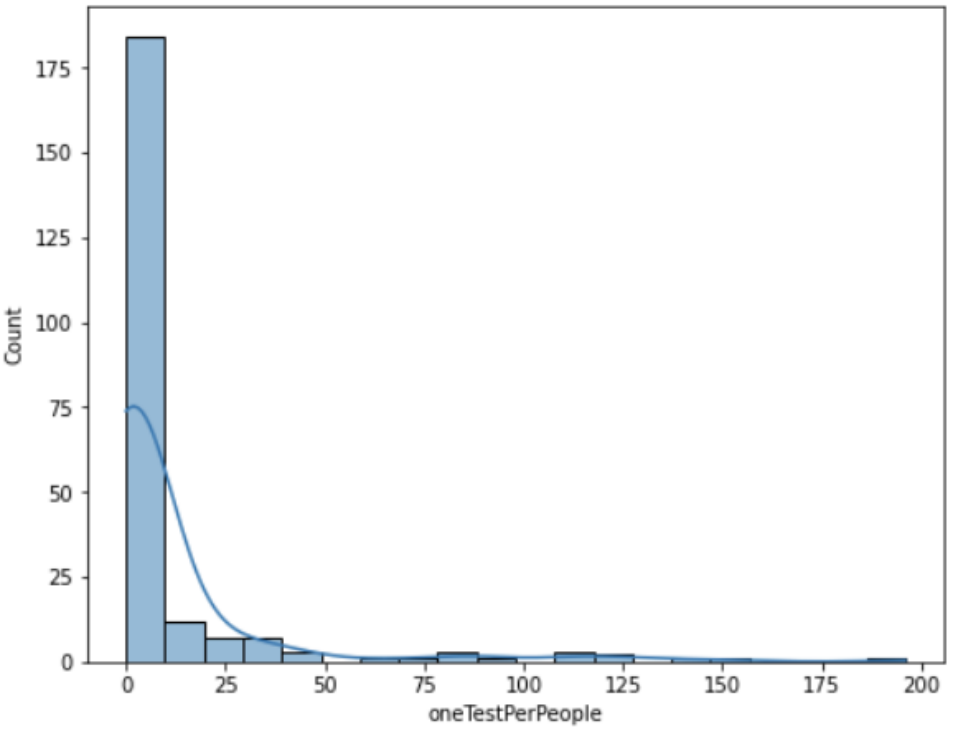
- activePerOneMillion chủ yếu tập trung ở mức khoảng 1000 người trong 1 triệu người, có rất nhiều outliers nằm rải rác từ 50000 trở lên => Mức độ đang nhiễm Covid 19 rất cao.
- criticalPerOneMillion chủ yếu tập trung ở mức khoảng 5 người trong 1 triệu người, có vài outliers nằm rải rác từ 50 đến 125.
- testsPerOneMillion chủ yếu tập trung ở mức khoảng 5 người trong 1 triệu người, có vài outliers nằm rải rác từ 50 đến 125.

6. Biểu đồ tần suất

- Ta sử dụng biểu đồ tần suất để biểu diễn sự phân bố so sánh các quốc gia trên thế giới.
- Ta sẽ chọn các trường oneCasePerPeople, oneDeathPerPeople và oneTestPerPeople để trực quan hoá.





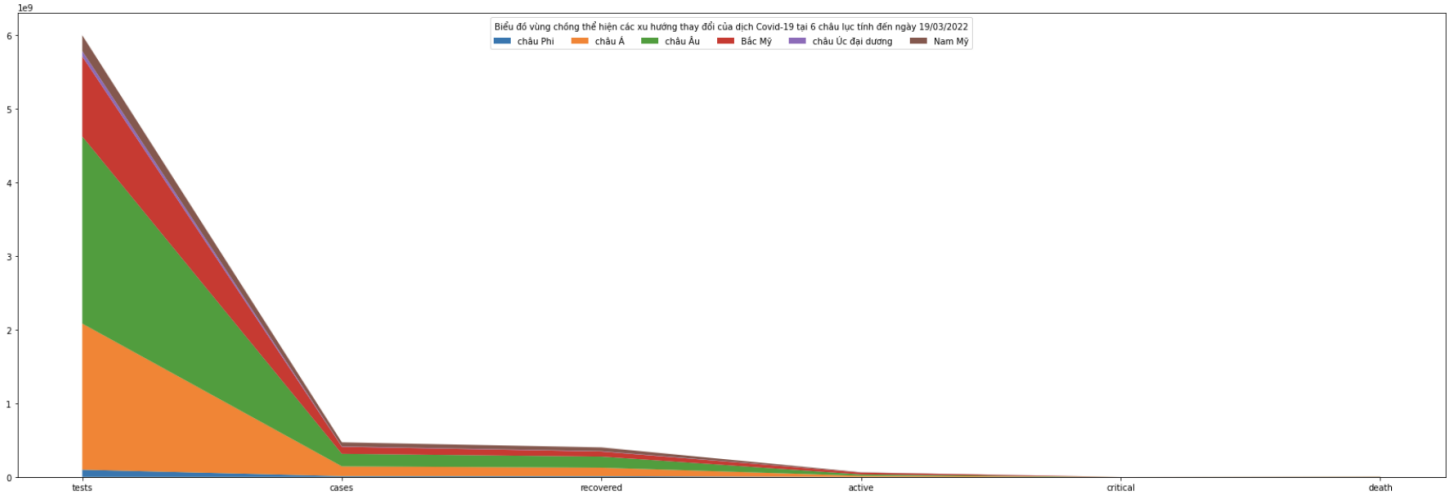


- Đa số các quốc gia có số lượng người (số người mà trong đó có ít nhất 1 người bị) thấp. Có rất ít quốc gia có chỉ số cao, số lượng không nhiều.
- Số lượng người càng thấp tương ứng với tỷ lệ bị càng cao, do thông được tính rằng trong số lượng người này thì có 1 người bị, vì vậy nếu càng thấp tức tỷ lệ mắc càng cao.

- Tỷ lệ Case (ca nhiễm) cao, tỷ lệ Death cũng cao => Mức độ mắc bệnh của đa số quốc gia này rất nguy hiểm.

7. Biểu đồ vùng xếp chồng

- Ta sử dụng biểu đồ vùng xếp chồng để so sánh và biểu diễn các xu hướng thay đổi của dịch Covid-19 tại 6 châu lục tính đến ngày 19/03/2022.
- Ta sẽ chọn các trường tests, cases, deaths, recovered, active và critical để trực quan hoá.

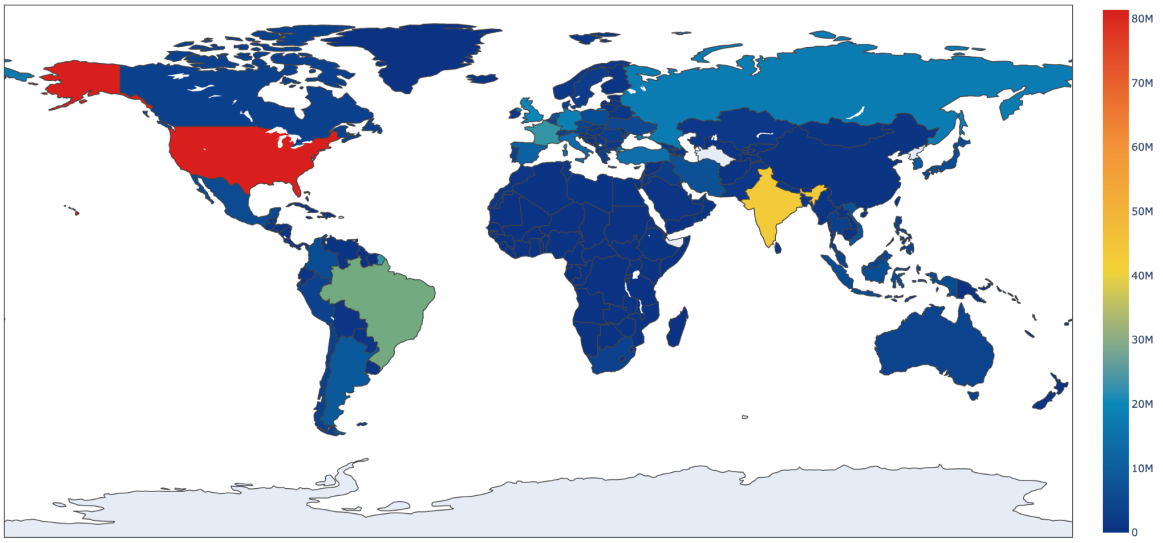


- Ta có thể thấy số lượng tests cao => Các châu lục đang tập trung đi test Covid 19. Cases và Recovered gần như bằng nhau => Tỷ lệ hồi phục sau khi bị Covid 19 tăng cao. Active thấp => số lượng bị mắc bệnh hiện nay thấp. Critical và Death gần bằng nhau => Do tỷ lệ hồi phục sau khi bị Covid 19 cao nên tỷ lệ tử vong thấp.
- Châu Âu có số lượng cao nhất, sau đó là Châu Á, Bắc Mỹ, Nam Mỹ, Châu Phi và cuối cùng là Châu Úc-Đại Dương.

8. Choropleth map

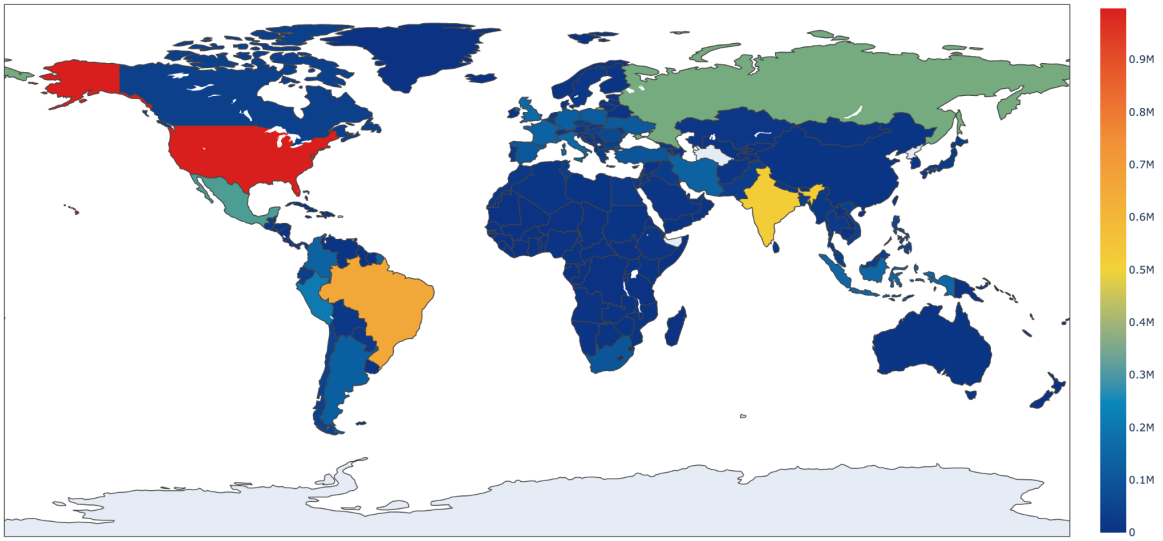
- Ta sử dụng Choropleth map để biểu diễn sự phân bố của dịch Covid-19 ở tất cả các quốc gia trên thế giới tính đến ngày 19/03/2022.
- Ta sẽ chọn các trường cases, deaths và tests để trực quan hoá.

Bản đồ thể hiện tổng số ca nhiễm của các quốc gia thuộc 6 châu lục tính đến ngày 19/03/2022 (đơn vị: ca)



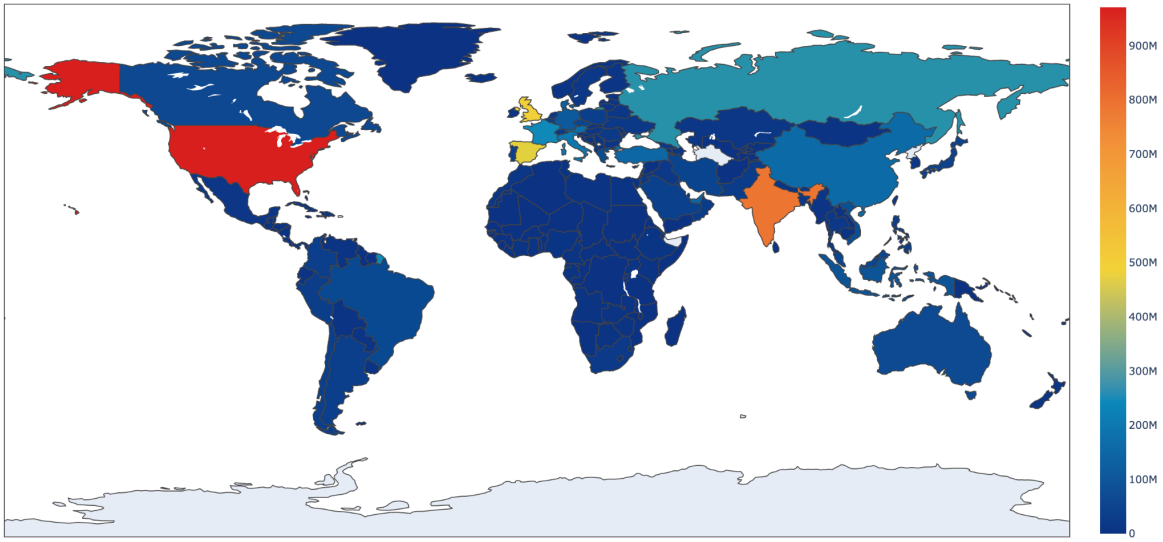
- Có thể thấy Mỹ là quốc gia có số lượng ca nhiễm cao nhất (khoảng 80 triệu trở lên), tiếp theo là Ấn Độ (40 - 50 triệu), Brazil (30 triệu), các quốc gia châu Âu (10 - 20 triệu).

Bản đồ thể hiện tổng số ca tử vong của các quốc gia thuộc 6 châu lục tính đến ngày 19/03/2022 (đơn vị: ca)



- Có thể thấy Mỹ là quốc gia có số lượng ca tử vong cao nhất (khoảng 900000 trở lên), tiếp theo là Brazil (700000), Ấn Độ (500000), Nga (400000) và rải rác ở các quốc gia châu Âu (200000).

Bản đồ thể hiện tổng số lượt xét nghiệm của các quốc gia thuộc 6 châu lục tính đến ngày 19/03/2022 (đơn vị: ca)



- Có thể thấy Mỹ là quốc gia có số lượng xét nghiệm cao nhất (khoảng 900 triệu trở lên), tiếp theo là Ấn Độ (800 triệu), các quốc gia châu Âu (300 triệu trở lên), Trung Quốc (200 triệu).