Trực quan hóa dữ liệu

LAB

Mối quan hệ của dữ liệu

1. Tổng quan đồ án:
   - Ngôn ngữ: Python
   - Công cụ: Jupyter Notebook
   - File bài làm: crawl_data.ipynb và visualize_data.ipynb
   - File chứa dữ liệu: Corona_by_day.csv và Corona_by_week.csv

2. Thu nhập dữ liệu:
   - Dữ liệu được lấy từ trang web: https://www.worldometers.info/coronavirus/
   - Sử dụng selenium để load web, save cache cũng như crawl dữ liệu thô từ các thẻ HTML trong trang web.

```
requests_cache.install_cache('demo_cache',expire_after=None,allowable_methods=['GET'])
driver = webdriver.Chrome(ChromeDriverManager().install())
```

```
link = "https://www.worldometers.info/coronavirus/"

driver.get(link)
```

   - Dữ liệu trả về được chia thành 3 phần là Now, Yesterday và 2 Days Ago nên ta sẽ lưu toàn bộ với từng phần có các khoảng thời gian khác nhau.
   - Tuy nhiên, bởi vì Now có nghĩa là thời điểm bây giờ, các trường dữ liệu vẫn chưa cập nhật một cách chính xác vì vẫn chưa cuối ngày nên ta chỉ lấy hai phần là Yesterday và 2 Days Ago.

```
today = date.today()
yesterday = (today - timedelta(1)).strftime("%d-%m-%Y")
two_days_ago = (today - timedelta(2)).strftime("%d-%m-%Y")

for row in rows_yesterday:

    row.insert(2,yesterday)
for row in rows_two_days_ago:

    row.insert(2,two_days_ago)

total_covid = rows_two_days_ago + rows_yesterday
```

- Có một lỗi nhỏ ở đây khi chạy một vài lần đầu mà nhóm chưa khám phá ra được lí do khi trong một số dòng có thể có một hoặc vài dữ liệu rỗng khiến số lượng trường dữ liệu không khớp. Để thể hiện rõ thì nhóm có đặt một dòng for nhỏ ở đây để check. Nếu thật sự có lỗi thì chạy lại chương trình vài lần sẽ hết.
- Dữ liệu được lưu vào:
  + Corona_by_day.csv
  + Corona_by_week.csv

```
for i in total_covid:
    if len(i) > 17:
        if i[-1] == '' or i[-1] == ' ':
            i.pop(-1)
        else:
            print(i)
```

3. Tiền xử lý dữ liệu:

- Tiền xử lý dữ liệu chủ yếu thực hiện ở các bước đổi dạng(type) của các trường dữ liệu thành số(int, float).
- Ta sẽ không thực hiện chuyển hóa các dữ liệu thiếu, rỗng hay data cleaning vì các dữ liệu lấy trực tiếp, nếu thực hiện các kĩ thuật này có thể khiến sai thông tin quá nhiều. Chung quy cũng vì các thông tin về dịch bệnh nên được minh bạch và thực tế.
- Tiếp theo, ta phân chia thành các tập dữ liệu phục vụ cho các mục đích trực quan hóa khác nhau. Ví dụ:

  + Tập dữ liệu sau khi chia theo 6 châu lục khác nhau vào ngày HÔM QUA(Yesterday)

Continent	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	TotalTests	Population	Date
Africa	11568729	5573	250060	70	10690988	7228	587319	1701	96534723	1394200091	2022-03-05
Asia	120150238	660777	1358458	1960	109432424	347517	9359356	31211	1949215809	4666025143	2022-03-05
Australia/Oceania	3799100	42607	8022	52	3323045	24940	400250	158	71712046	43396187	2022-03-05
Europe	159615176	539851	1721782	1888	138621259	719040	19272135	13579	2505120376	748384748	2022-03-05
NorthAmerica	95295244	34089	1414249	734	67201767	182292	26669425	12328	1070194927	596825025	2022-03-05
SouthAmerica	54694359	88234	1262482	947	46697084	126908	2620218	13037	210736462	436700401	2022-03-05

+ Tập dữ liệu theo thời gian đặc biệt cho line chart.

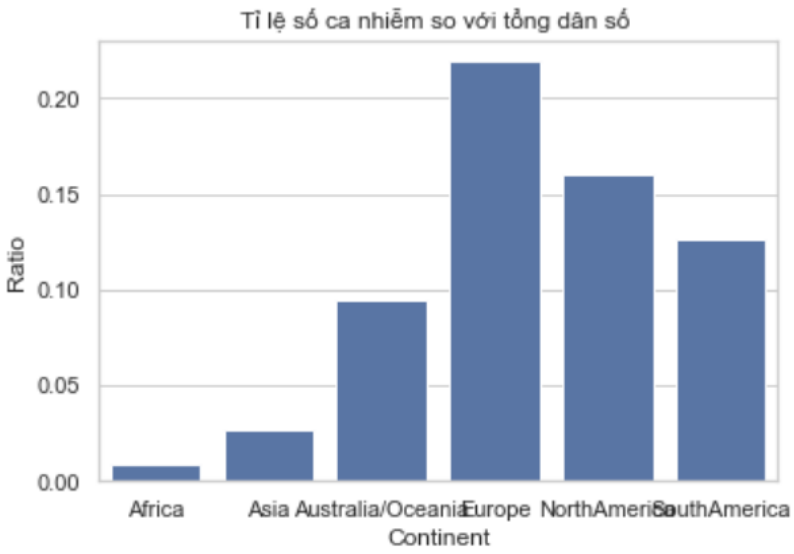
Date	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	TotalTests	Population
2022-02-28	437173429	1128262	5976416	6045	366621406	2154754	60360176	74759	5864353102	7884712500
2022-03-01	438510827	1337398	5983229	6813	368695782	2074376	59615435	74737	5869547415	7884712500
2022-03-02	440290515	1590173	5992432	7756	370783860	1939726	59289172	75953	5881445441	7884917270
2022-03-03	440377632	1599844	5993186	7872	370787000	1942866	59371827	75953	5881556015	7885122041
2022-03-04	443707497	1646436	6009345	8022	374469187	1751635	58996442	71862	5897737771	7885326827
2022-03-05	445123567	1371131	6015068	5651	375967273	1407925	58908703	72014	5903514343	7885531595

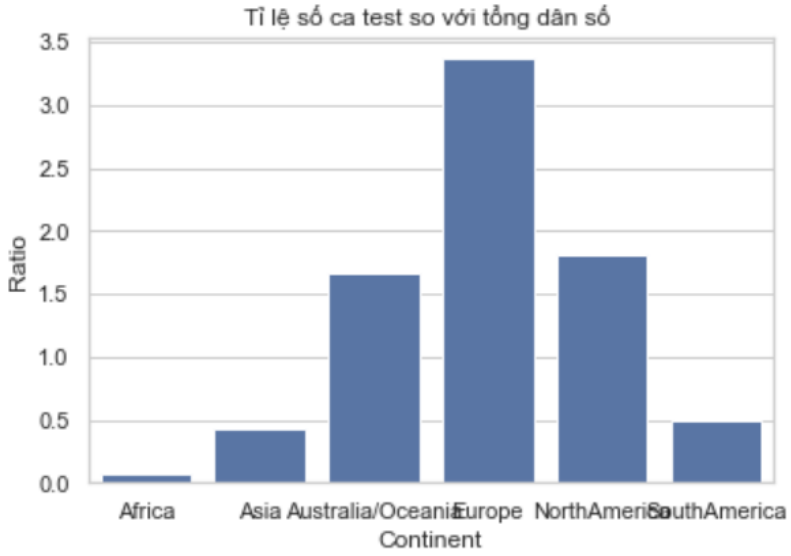
4. Trực quan hóa và nhận xét:

1. Biểu đồ cột về một vài thông số tổng quát của 6 châu lục như tổng ca nhiễm, tổng ca test, tổng ca chết, hồi phục… (Trương Chí Toàn)

Lý do chọn:

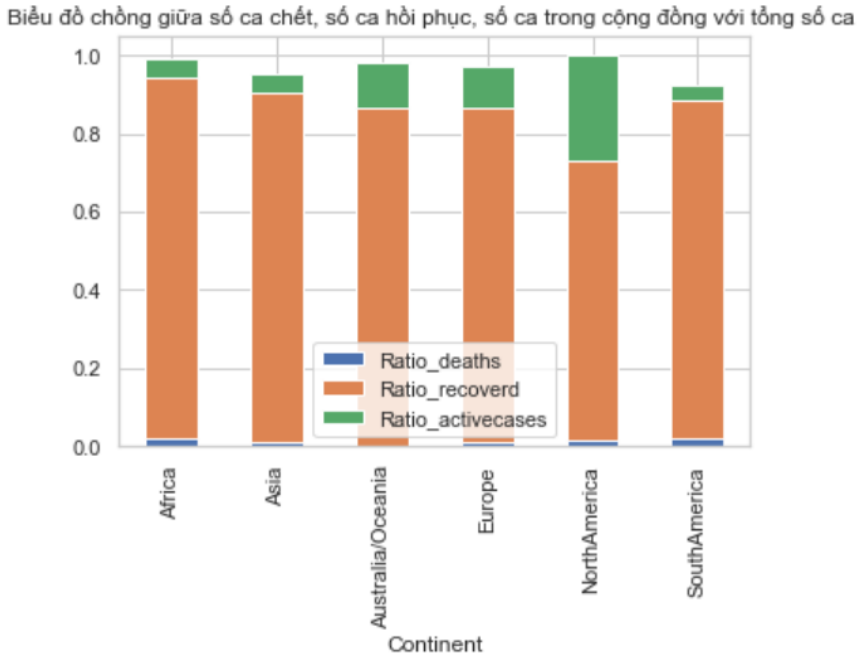
- Biểu đồ cột dùng để so sánh tương quan về độ lớn giữa các đại lượng. Cụ thể ở đây là tổng số ca nhiễm, tổng ca chết, hồi phục giữa các châu lục.
- Ngoài ra ở đây ta có sử dụng thêm biểu đồ dạng cột chồng để thể hiện mối quan hệ tổng số ca chết, số ca hồi phục và số ca trong cộng đồng với tổng số ca nhiễm để có cái nhìn trực quan về tình trạng dịch bệnh ở mỗi châu lục.





Nhận xét:

- Ta thấy châu Âu là châu có số ca nhiễm và số ca test lớn nhất thế giới. Chứng tỏ châu Âu đã có khoảng thời gian để dịch bệnh lây lan, hoành hành, sau đó đã có những biện pháp cố gắng khắc phục rõ ràng.
- Các châu lục khác (kể cả châu Âu), đều có tỉ lệ số ca nhiễm và số ca test gần như là ngang nhau, trong khi đó Nam Mỹ, theo như biểu đồ thì có tỉ lệ ca nhiễm khá lớn trong khi tỉ lệ số ca test lại rất thấp. Chứng tỏ Nam Mỹ đầu tư cho phòng chống covid khá tệ, có thể 1 phần là do tỉ lệ người nghèo ở Nam Mỹ rất là lớn.



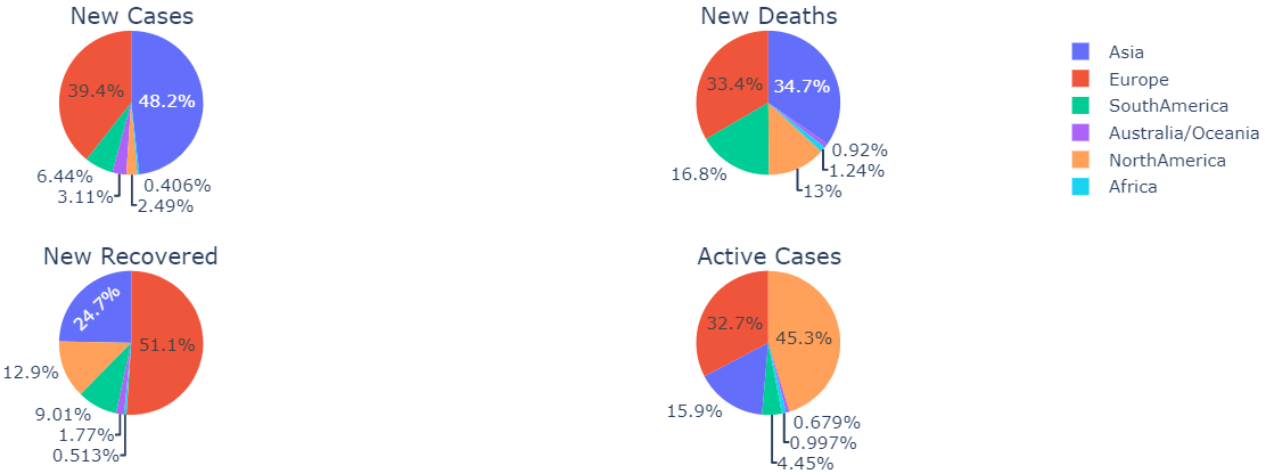
Nhận xét:

- Tuy châu Âu có tỉ lệ ca nhiễm cao nhất nhưng Bắc Mỹ lại là nơi có tỉ lệ số ca ngoài cộng đồng cao nhất, hơn gấp đôi so với châu có tỉ lệ ca nhiễm ngoài cộng đồng đứng thứ nhì. Chứng tỏ Bắc Mỹ đã áp dụng miễn dịch cộng đồng lên trên hầu hết các thành phố của họ.

2. **Biểu đồ tròn trực quan tỷ lệ về các thông số ca nhiễm, ca chết, ca hồi phục MỚI và thêm cả số ca nhiễm trong cộng đồng trên toàn thế giới.**

   Lý do chọn:
   - Biểu đồ tròn thường được trực quan để biểu diễn cơ cấu, tỷ lệ giữa các thành phần trong một tổng thể nhất định.
   - Trong trường hợp này, tổng thể chính là toàn bộ thế giới và các thành phần là 6 châu lục. Như vậy, với từng thông số ta có các biểu đồ khác nhau thể hiện các tình hình khác nhau ở 6 châu lục.

- Ngoài ra, các thông số được nêu đều là MỚI theo từng ngày, điều này cũng thể hiện tình hình thế giới ngày hôm qua biến động như thế nào.



Nhận xét:

- Đối với ca nhiễm mới, Châu Á và Âu vẫn chiếm rất cao chứng tỏ ở hai châu lục này vẫn đang bùng dịch hơn so với phần còn lại.
- Đối với ca chết, vẫn là Châu Á và Âu, một phần cũng có thể đoán ra là do ca nhiễm mới như đã nêu quá cao kéo theo tỷ lệ tử vong. Ngoài ra, Nam Mĩ cũng khá cao(16,8%) bằng một nửa so với hai châu lục đứng đầu.
- Đối với ca hồi phục, ta có một lời khen cho Châu Âu khi tỷ lệ người hết bệnh tới hơn 50%. Châu Á ở hai chỉ số đã nêu cao bằng hoặc hơn Châu Âu nhưng ca hồi phục chỉ bằng một nửa.
- Trong khi ở ba biểu đồ đã nêu, Châu Á và Âu chiếm tỷ lệ cao thì ở biểu đồ cuối cùng là ca nhiễm trong cộng đồng, Bắc Mĩ có số lượng người nhiễm bệnh trong cộng đồng đứng đầu với 45,3% và tiếp theo sau là Châu Âu.

3. **Biểu đồ đường trực quan theo thời gian 3 thông số ca nhiễm/chết/hồi phục MỚI.**

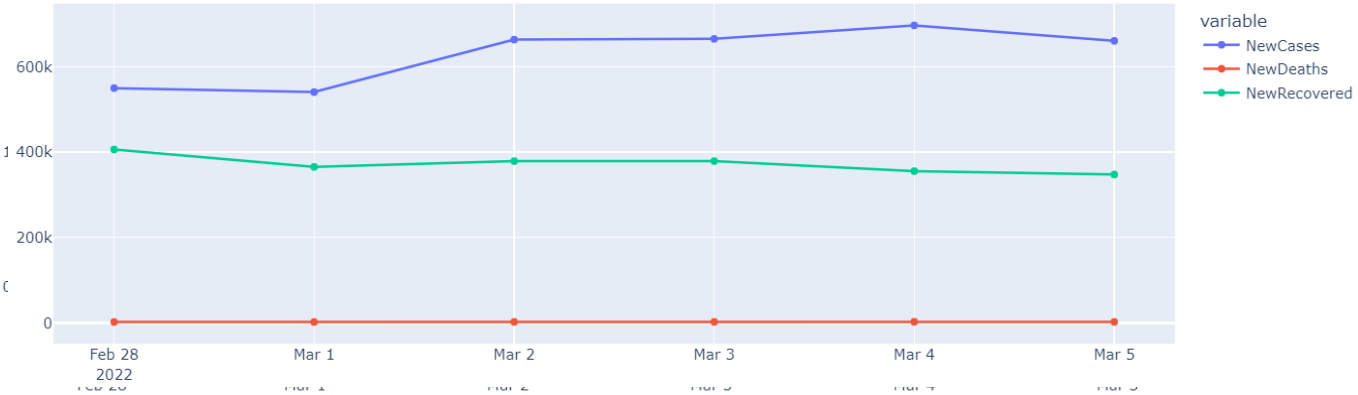
Lý do chọn:

- Biểu đồ đường được sử dụng để trực quan dữ liệu trong tiến trình phát triển theo thời gian của một hoặc nhiều dữ liệu khác nhau.
- Tuy nhiên, nếu biểu đồ tròn đã nêu chỉ trực quan trong khoảng thời gian ngày hôm qua, biểu đồ đường có thể trực quan toàn bộ từ khi nhóm thu nhập dữ liệu đến thời điểm nhóm nộp bài.
- Bởi vì tập dữ liệu của nhóm có lưu theo thời gian nên có thể chọn bất kì một trường dữ liệu nào để áp dụng thoải mái. Ở đây, nhóm quyết định chọn các thông số MỚI để trực quan rõ ràng hơn theo thời gian mà biểu đồ tròn không làm được ở câu số 2.

Ta sẽ trực quan toàn bộ thế giới trước rồi tới từng châu lục.

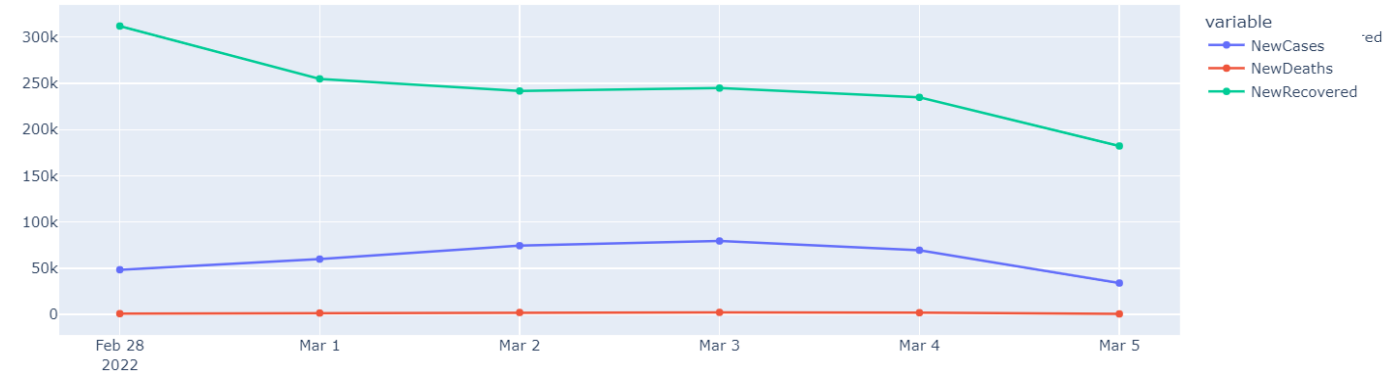
Date	NewCases	NewDeaths	NewRecovered
2022-02-28	1128262	6045	2154754
2022-03-01	1337398	6813	2074376
2022-03-02	1590173	7756	1939726
2022-03-03	1599844	7872	1942866
2022-03-04	1646436	8022	1751635

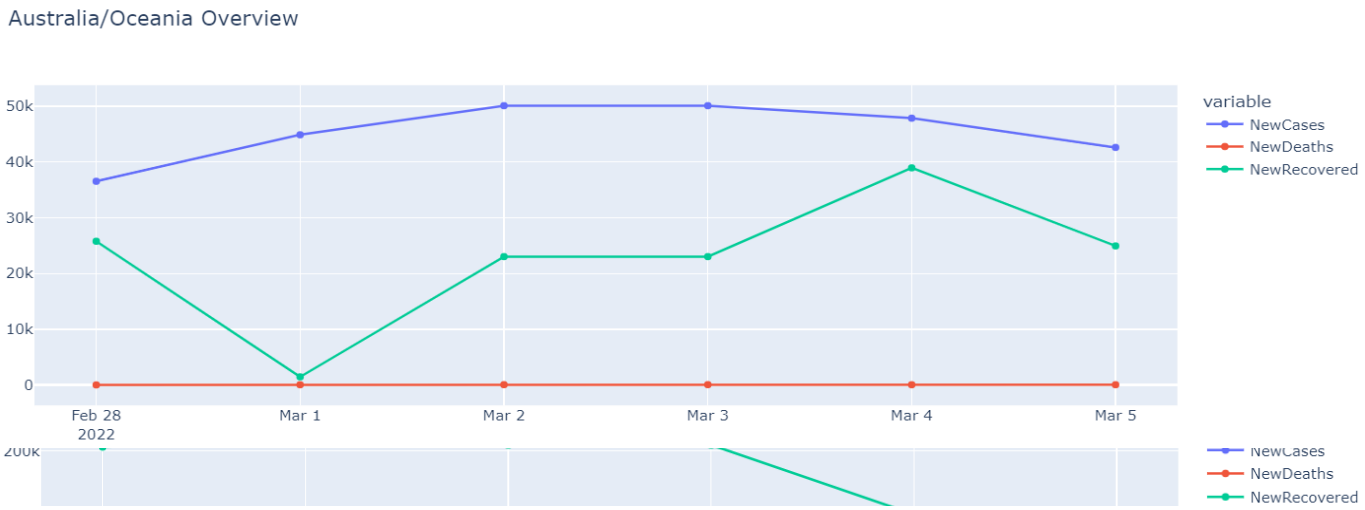
Asia Overview

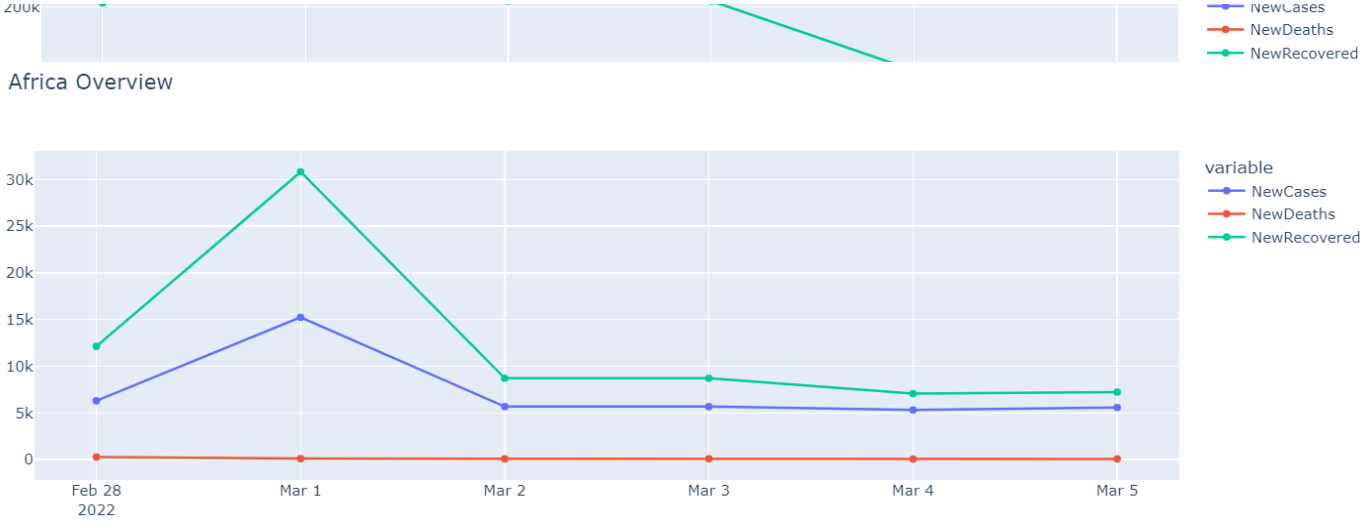


Europe Overview

NorthAmerica Overview







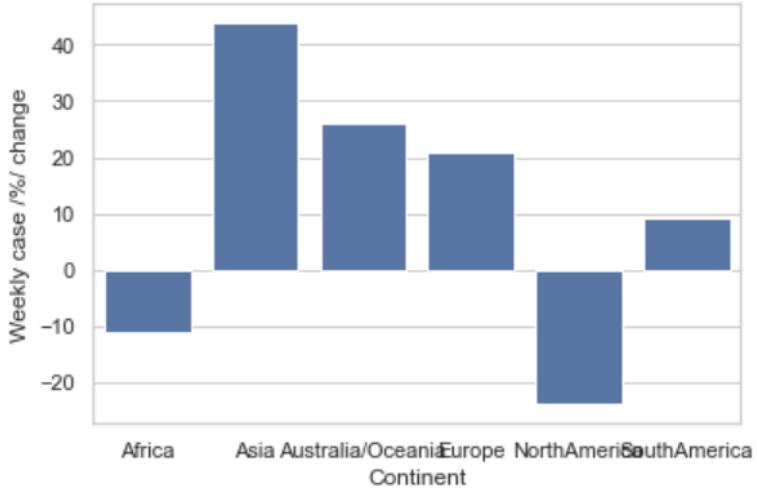
Nhận xét:

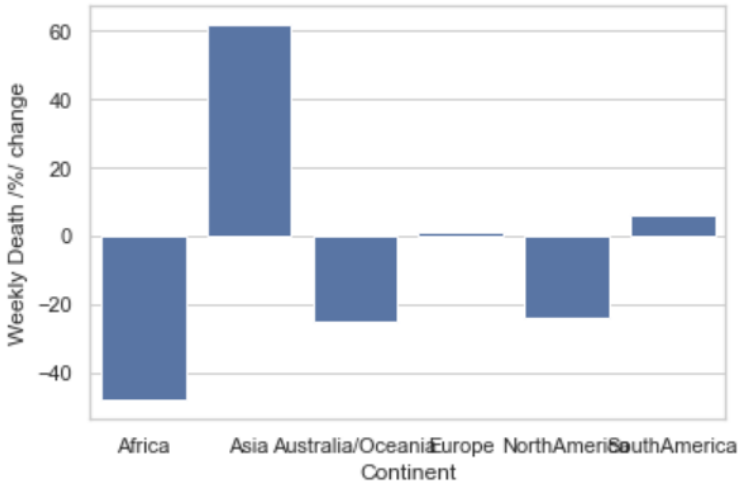
- Ta có thể thấy mỗi châu lục đều có mô hình biểu đồ khác nhau. Châu Âu và Thế giới có mô hình khá giống nhau khi hai đường số ca nhiễm và hồi phục đều hợp lại từ giữa tới cuối, các chỉ số đều ổn định.
- Châu Á như đã trực quan ở biểu đồ tròn khi có 48% tỷ lệ ca nhiễm mới thì ở đây số ca nhiễm mới mỗi ngày đều vượt trội hơn so với hai chỉ số còn lại. Bên cạnh đó thì cũng có số ca hồi phục mới cũng cao nên có phần yên tâm.
- Trong khi đó, Bắc Mĩ có chuyển hướng vô cùng tích cực khi số ca hồi phục gấp mấy lần hai ca còn lại. Nam Mĩ cũng phần nào giống Bắc Mĩ tuy ca nhiễm vào ngày 4/3 có tăng.

- Châu Đại Dương có xu hướng giống Châu Á khi số ca nhiễm vượt trội tuy ca hồi phục cũng có tăng cao.
- Châu Phi có tăng cao ở 1/3 và giảm mạnh, ổn định ở các ngày còn lại.

4. **Biểu đồ cột hai trường Weekly Case/%/Change và Weekly Death/%/Change của 6 châu lục. (Trương Chí Toàn)**

   Lý do chọn:
   - Hai trường dữ liệu khá đặc biệt khi có số âm tức tuần này, các nước có tình hướng chuyển biến tích cực và ngược lại.
   - Với sự trợ giúp của biểu đồ cột, ta có thể trực quan sự thay đổi số lượng ca nhiễm và số ca chết so với tuần trước của 6 châu lục theo chiều hướng tích cực hay tiêu cực một cách dễ dàng.

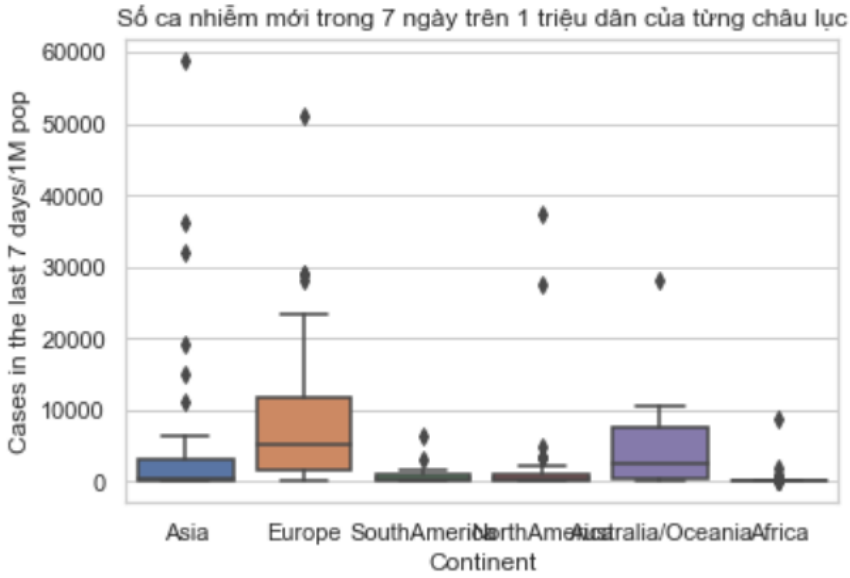




Nhận xét:

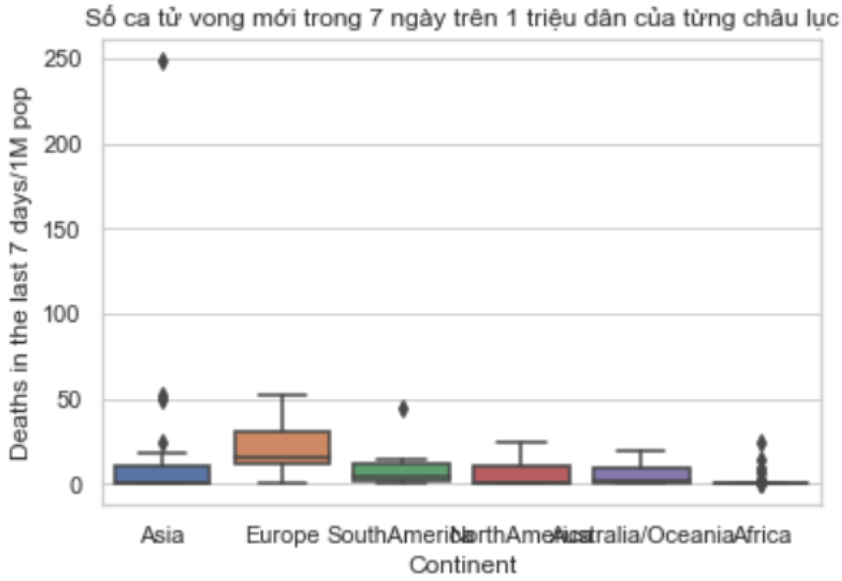
- Như phân tích ở câu 1, Bắc Mỹ đang thực hiện miễn dịch cộng đồng, và như thông số ở trên thì Bắc Mỹ đang có số ca mắc mới chuyển biến tích cực nhất, tiếp sau đó là châu Phi. Có khi nào việc thực hiện miễn dịch cộng đồng lại hợp lý. 4 châu còn lại đều có % số ca mắc mới dương, và cao nhất là Châu Á.
- Ở bảng số ca chết mỗi tuần, châu Phi có chuyển biến tích cực nhất, theo sau là châu Úc và Bắc Mỹ. Châu Á lại có tỉ lệ ca chết cao vượt trội so với cả châu Âu và Nam Mỹ. Như vậy dịch covid lại đang diễn biến rất là phức tạp ở châu Á. Trong khi châu Phi và Bắc Mỹ là 2 châu đang có diễn biến tích cực nhất.

5. **Biểu đồ boxplot về hai trường dữ liệu chính trên tập dữ liệu tuần về số ca nhiễm và số ca chết trên 1 triệu dân trong vòng 7 ngày toàn thế giới. (Trương Chí Toàn)**
   Lý do chọn:
   - Sử dụng boxplot để thể hiện sự phân phối của dữ liệu, độ dàn trải dữ liệu giữa các châu lục, dữ liệu có đối xứng hay không, phân bố rộng hay hẹp và thể hiện ra các điểm ngoại lệ
   - Ở đây ta thể hiện số ca nhiễm mới của các thành phố ở mỗi châu lục trên boxplot



   Nhận xét:

- Ta thấy có 3 boxplot nhìn rõ ràng, và cả 3 đều có Q3 lớn hơn Q1, chứng tỏ số thành phố có ca nhiễm lớn nhiều hơn số có ít ca, nghĩa là tốc độ lây lan ở các thành phố lớn nhiều hơn thành phố nhỏ.
- Châu Á tuy có dạng hộp nhỏ nhất so với 2 hộp trên nhưng lại có số lượng outlier nhiều nhất.



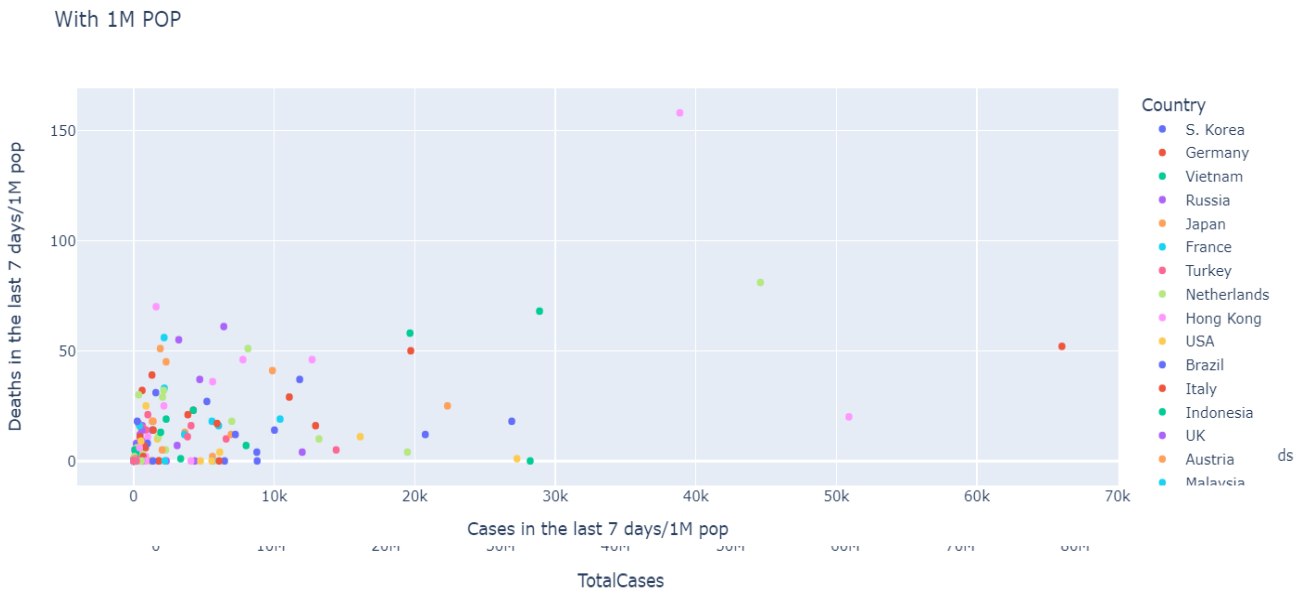
Nhận xét:

- Ta thấy châu Á vẫn có lượng outlier cao nhất về số ca tử vong, rõ ràng có 1 outlier khiến cho cả biểu đồ bị kéo cao lên.
- Từ 2 biểu đồ trên, đây có thể liên quan tới việc tại sao câu 5 tại châu Á có tỉ lệ ca mắc mới và tử vong cao như vậy.

6. **Biểu đồ phân tán trực quan hai cặp trường dữ liệu:**
    - **Tổng ca nhiễm với ca chết ngày hôm qua toàn thế giới.**
    - **Số ca nhiễm và số ca chết trên 1 triệu dân trong vòng 7 ngày toàn thế giới**

Lý do chọn:

- Biểu đồ phân tán được sử dụng để thể hiện mối quan hệ giữa hai trường đã nêu trên tập dữ liệu tuần vì giữa ca nhiễm và ca chết chắc chắn có mối liên hệ nào đó như nhân quả(cause-effect) khi số ca nhiễm tăng thì ca chết cũng tăng trên một phương diện nào đó.
- Ta lựa chọn hai tập dữ liệu xét trên tổng và trên triệu dân vì nếu xét trên tổng, ta có cái nhìn tổng quát trước. Trong khi triệu dân có cái nhìn khách quan hơn khi cùng một mật độ dân cư.



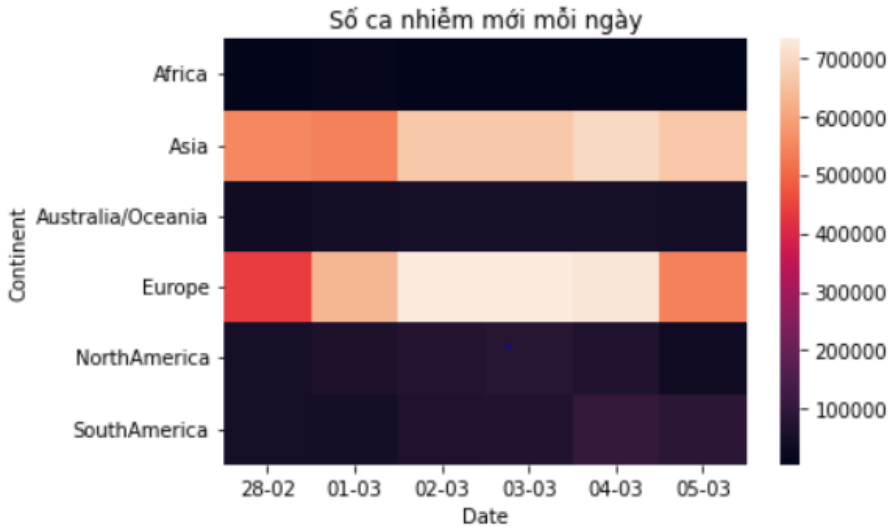
Nhận xét:

- Ở biểu đồ phân tán thế giới, hầu hết các nước đều tập trung ở phần dưới, ngoại trừ các nước top đầu về ca nhiễm đồng thời ca chết. Vậy ta có thể ngầm nhận định rằng các nước có số lượng ca nhiễm cao thì tỷ lệ ca chết cũng cao nhưng giới hạn lại ở các nước top đầu.
- Với cùng mật độ dân cư là 1 triệu dân, biểu đồ dương như phân tán hơn ở các nước nhóm dưới và biểu thị rõ hơn về quan hệ nhân quả như đã nêu. Có một vài nước ngoài cùng khi có số lượng ca nhiễm từ 50-70k nhưng chết chỉ khoảng 50 trở xuống.

7. **Biểu đồ nhiệt (heatmap) trực quan độ nóng của 6 châu lục theo thời gian về số lượng ca nhiễm mới.**

Lý do chọn:

- Biểu đồ nhiệt trực quan sự tương tác giữa số ca nhiễm mới tác động đến các châu lục với màu nóng là ít tương tác, màu nhạt là nhiều tương tác.
- Biểu đồ tròn trong câu 2 có thể trực quan tỷ lệ giữa các châu lục nhưng khi trực quan biểu đồ dạng này, ta có thể thấy rõ hơn sự tương quan, tương tác giữa các châu lục hơn.



Nhận xét:

- Có hai châu lục đen tức tương tác quá ít là Châu Phi và Châu Đại Dương. Nếu liên hệ với các biểu đồ khác thì khá dễ hiểu khi các châu lục này có các chỉ số đều rất ít so với phần còn lại.
- Châu Mỹ gồm Bắc Mỹ và Nam Mỹ thì sáng hơn tức tương tác cao hơn một chút so với hai châu lục "đen" đã nhắc vào những ngày cuối được hiển thị.
- Không có gì lạ khi hai châu lục tương tác nhiều nhất là Châu Âu và Châu Á vì ta đã có khá nhiều kết luận khi hai châu lục này đứng đầu về các chỉ số khác nhau trong biểu đồ tròn ở câu 2.