ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO EXERCISE 2

MÔN HỌC: NHẬP MÔN KHOA HỌC DỮ LIỆU

Chuyên ngành: Khoa học dữ liệu

Thành phố Hồ Chí Minh – 2021

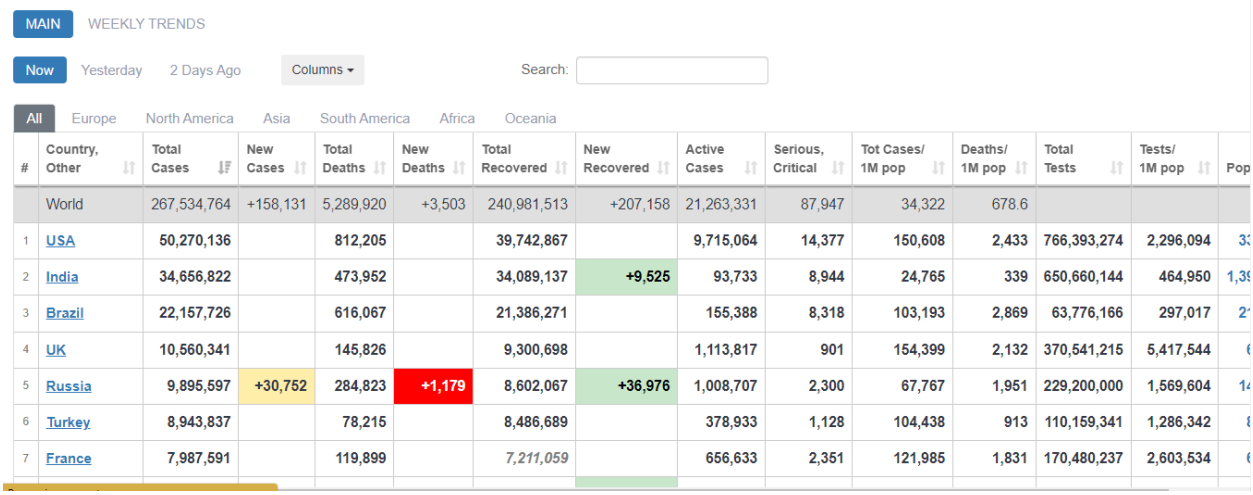
THÔNG TIN CÁNHÂN

MSSV	Họ tên	Email
19127083	Nguyễn Hữu Tuấn	19127083@student.hcmus.edu.vn

MỤC LỤC

1. Thu thập dữ liệu

- Bài tập này sẽ thu thập dữ liệu từ trang Worldmeter, trang cho biết thông tin về dịch COVID toàn thế giới.
- Ở phần này, ta sẽ thu thập toàn bộ bảng dữ liệu có định dạng sau:



MAIN | WEEKLY TRENDS

Now | Yesterday | 2 Days Ago | Columns | Search:

All | Europe | North America | Asia | South America | Africa | Oceania

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	New Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Pop
	World	267,534,764	+158,131	5,289,920	+3,503	240,981,513	+207,158	21,263,331	87,947	34,322	678.6			
1	USA	50,270,136		812,205		39,742,867		9,715,064	14,377	150,608	2,433	766,393,274	2,296,094	3
2	India	34,656,822		473,952		34,089,137	+9,525	93,733	8,944	24,765	339	650,660,144	464,950	1,3
3	Brazil	22,157,726		616,067		21,386,271		155,388	8,318	103,193	2,869	63,776,166	297,017	2
4	UK	10,560,341		145,826		9,300,698		1,113,817	901	154,399	2,132	370,541,215	5,417,544	
5	Russia	9,895,597	+30,752	284,823	+1,179	8,602,067	+36,976	1,008,707	2,300	67,767	1,951	229,200,000	1,569,604	1
6	Turkey	8,943,837		78,215		8,486,689		378,933	1,128	104,438	913	110,159,341	1,286,342	
7	France	7,987,591		119,899		*7,211,059*		656,633	2,351	121,985	1,831	170,480,237	2,603,534	

- Ta sẽ làm việc này sử dụng phương pháp parse HTML, và lần lượt truy xuất dữ liệu của 3 ngày gần nhất. Để dữ liệu được chính xác, thu thập vào cuối mỗi ngày sẽ cho ta thông tin chính xác nhất (vì thu thập ban ngày sẽ có những nước chưa cập nhật)
- Để thu thập dữ liệu, truy cập file *19127083_crawl.ipynb* và chọn "Run all" để quá trình thu thập dữ liệu bắt đầu.
- Sau khi thu thập dữ liệu, toàn bộ data thô chưa được xử lí sẽ lưu vào thư mục "data" và sẽ đến bước tiền xử lý dữ liệu.

2. Tiền xử lý dữ liệu

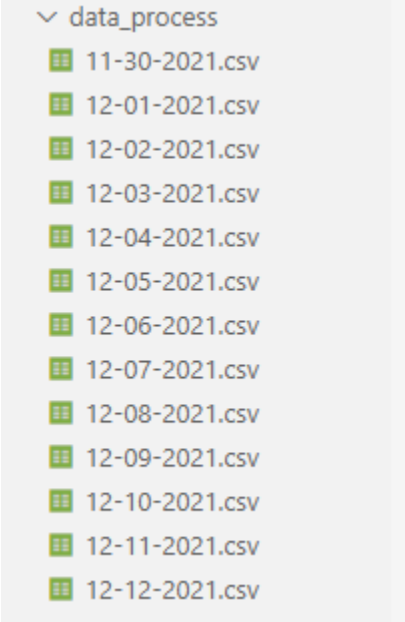
- Ở bước này, ta sẽ xử lí vấn đề thiếu dữ liệu đối với một số cột dữ liệu là số.
- Các cột dữ liệu kiểu chuỗi đều không thiếu dữ liệu, tuy nhiên ở các cột dữ liệu là số thì thiếu khá nhiều.
- Đối với các cột có kiểu là int nhưng khi đọc về lại ở kiểu float, ta sẽ chuyển lại về kiểu int. Và ở các cột này, đối với các cột bị thiếu dạng "NaN", ta sẽ đổi về dạng số 0 (tương đương với số ca ở mỗi cột bằng 0). Ta sẽ xử lí các cột sau:

```
change = ['TotalDeaths', 'NewCases',  'NewDeaths', 'TotalRecovered', 'NewRecovered', 'ActiveCases', 'Serious,Critical',\
      'TotalTests', 'Population', 'TotCases/1M pop', 'Deaths/1M pop', 'Tests/1M pop', '1 Caseevery X ppl',\
      '1 Deathevery X ppl', '1 Testevery X ppl', 'Active Cases/1M pop']
change_float = ['New Cases/1M pop', 'New Deaths/1M pop']
```

- Ta sẽ thay "NaN" bằng 0 và đổi lại các cột ở sai kiểu dữ liệu.

```
for col in change:
    df[col] = df[col].replace(np.nan, 0)
    df[col] = df[col].astype(int)
for col in change_float:
    df[col] = df[col].replace(np.nan,0.0)
```

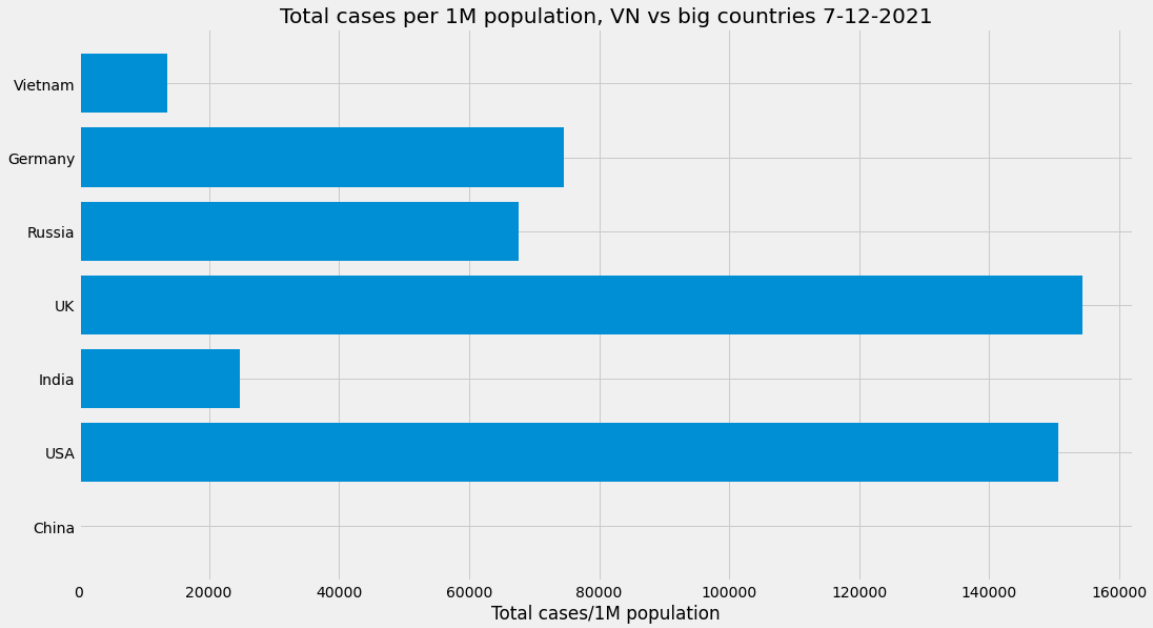
- Sau khi đổi, ta sẽ lưu theo định dạng "MM-DD-YYYY" để windows tự động sắp xếp các file theo đúng thứ tự ngày tháng. Kết quả thu được sau khi chạy file *19127083_dataprocessing.ipynb* như sau:



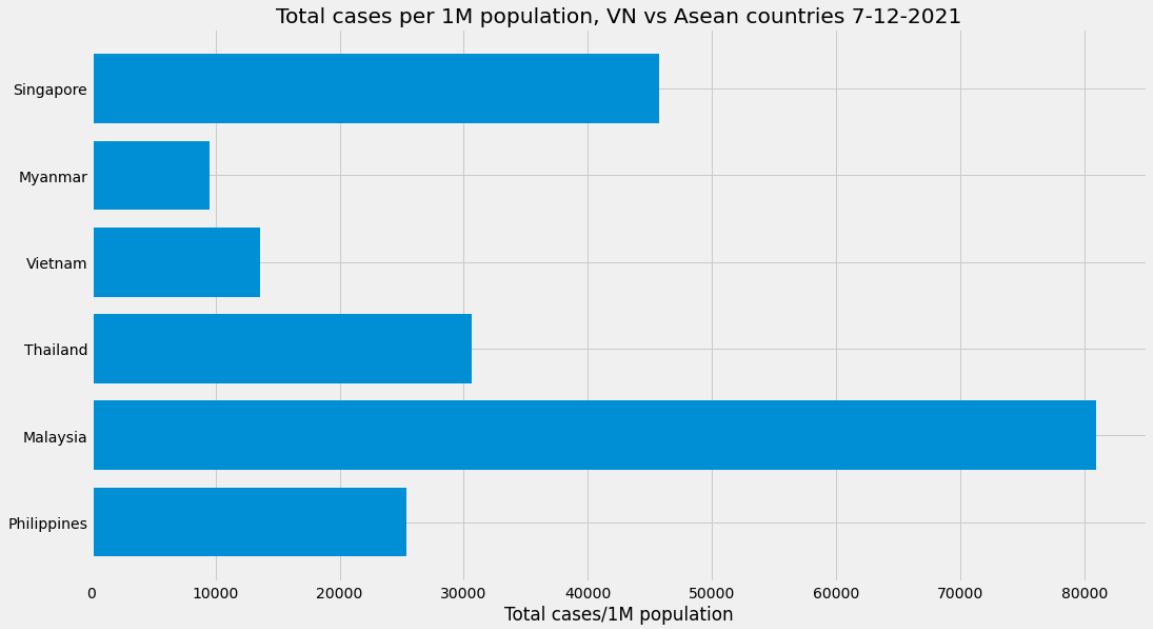
- Sau đó, sẽ đến bước phân tích hay trực quan hoá dữ liệu.

3. Trực quan hoá dữ liệu

- Trước tiên, để thấy được số ca nhiễm COVID khi so giữa Việt Nam và các nước khác, ta sẽ sử dụng bar chart để so sánh (vì các ca nhiễm giữa các nước là rời rạc và ta cần so sánh mức độ lây nhiễm giữa các nước). Ngoài ra, ta sẽ sử dụng thuộc tính "TotCases/1M pop" để so sánh, bởi thuộc tính này sẽ thể hiện đúng nhất sự tương quan giữa các nước (Tổng số ca sẽ không phản ánh chính xác bởi có nước đông dân, có nước ít dân). Kết quả thu được khi so sánh Việt Nam với các nước lớn khác vào ngày 7-12-2021 như sau:



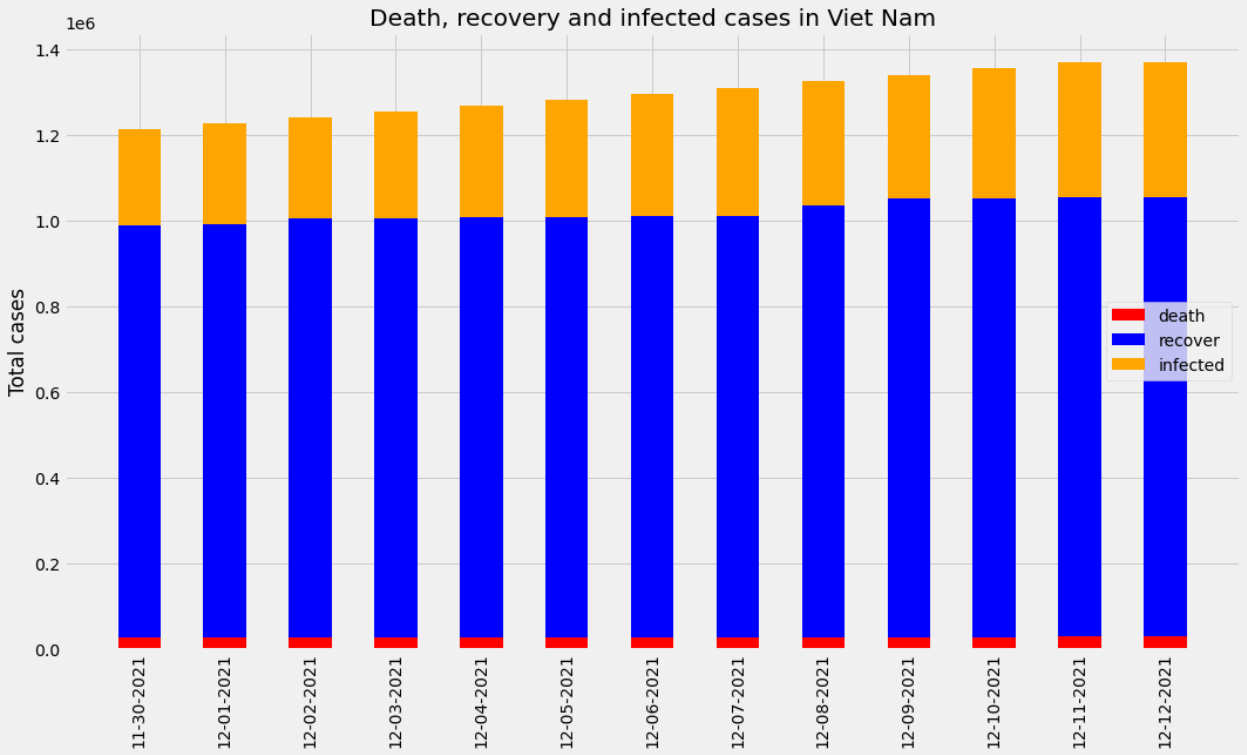
- Khi nhìn vào dữ liệu trên, có thể thấy số ca nhiễm khi so với các nước lớn của Việt Nam là khá nhỏ, điều này phản ánh đúng thực trạng của đất nước, khi mọi người hạn chế ra đường khi không cần thiết. Tại các nước lớn, khi người dân vẫn ra đường rất nhiều bất chấp dịch, tỉ lệ số ca nhiễm sẽ tăng cao.
- Khi so với các nước trong khu vực Đông Nam Á, ta được đồ thị sau:



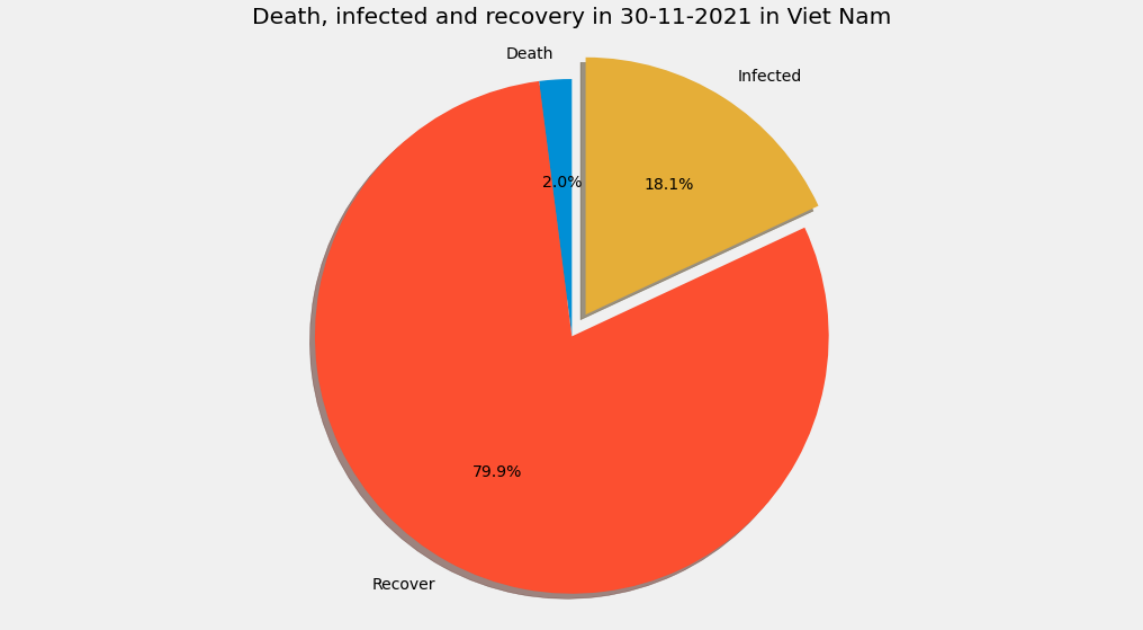
- Khi nhìn vào đồ thị, số ca của nước ta cũng khá nhỏ, cho thấy sự phòng chống dịch khá tốt cho dù trang thiết bị còn hạn chế.

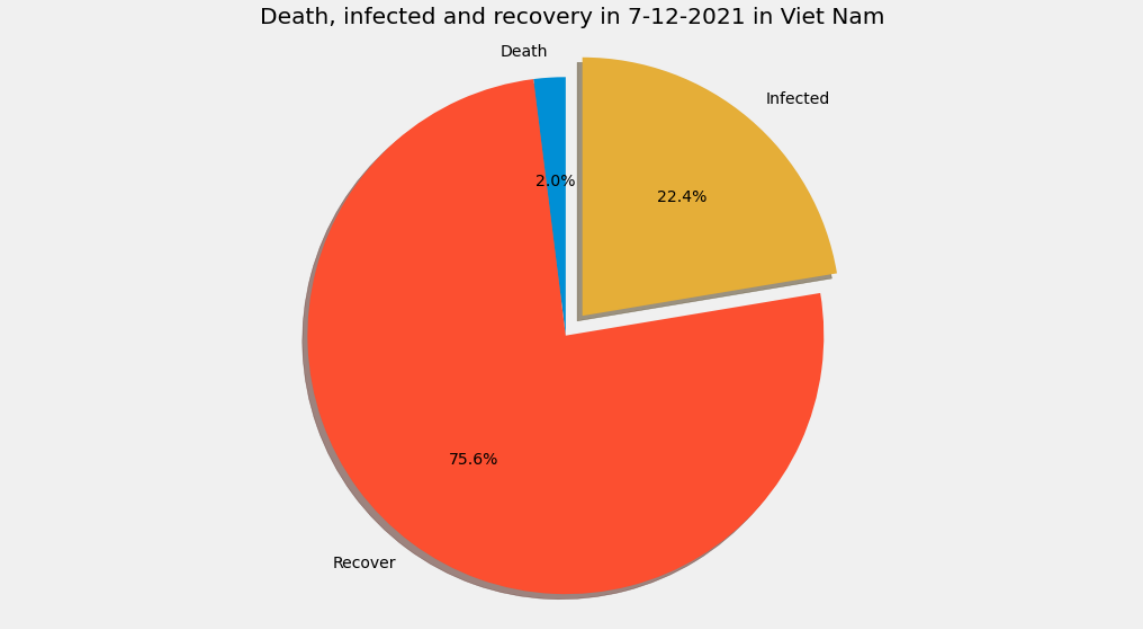
- Để trực quan hơn về tình hình Việt Nam, ta sẽ xem xét về tỉ lệ số ca tử, đã hồi phục và đang dương tính ở nước ta. Mặc dù bảng dữ liệu không cho ta tình trạng số ca đang

nhiễm, nhưng ta có thể tính được dựa trên tổng số ca nhiễm, số ca hồi phục và số ca chết ($số\ ca\ đang\ nhiễm\ =\ tổng\ số\ ca\ -\ số\ ca\ hồi\ phục\ -\ số\ ca\ chết$):



- Khi nhìn vào đồ thị trên, ta có thể thấy rằng, trong tổng số ca mắc, số ca tử vong chiếm tỉ lệ khá thấp, chưa tới 5%, và số ca hồi phục đang tăng cao, đó là một dấu hiệu đáng mừng. Tuy nhiên, qua từng ngày phần màu vàng lại càng tăng, và chiều dài của mỗi cột cũng đều tăng. Điều này có nghĩa là mấy ngày vừa qua, tình hình dịch ở nước ta đang khá căng thẳng. Ta sử dụng "stacked bar chart" để thấy được sự tương quan giữa 3 nhóm là số ca chết, số ca đang nhiễm và số ca hồi phục. Tuy nhiên, để thấy rõ hơn giữa tỉ lệ giữa các nhóm, ta sẽ sử dụng pie chart để làm rõ vấn đề này. Dưới đây là đồ thị biểu diễn phần trăm các ca theo nhóm ở Việt Nam vào 2 ngày cụ thể:

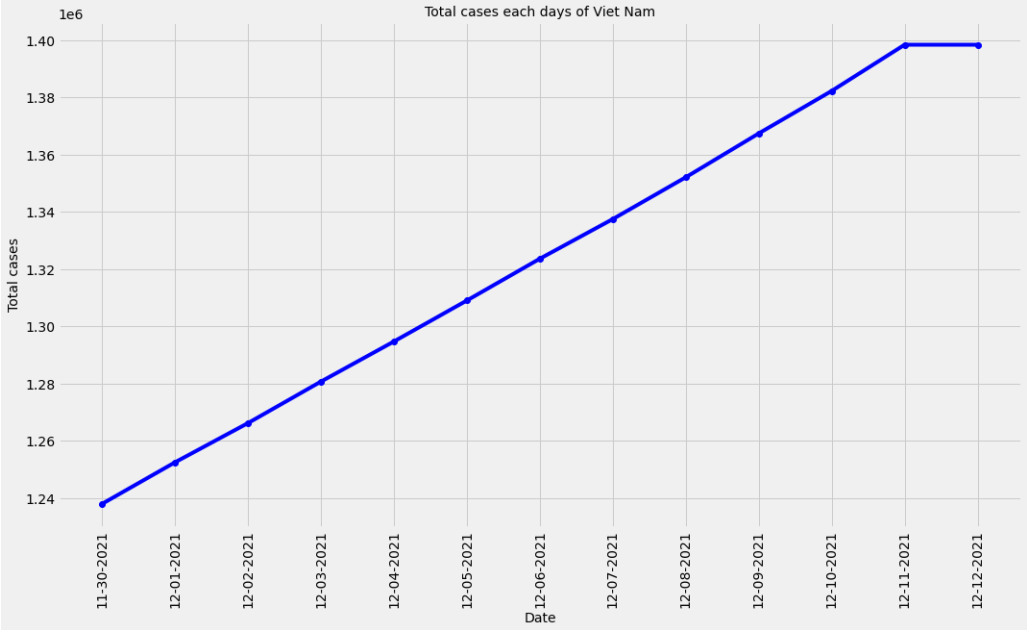




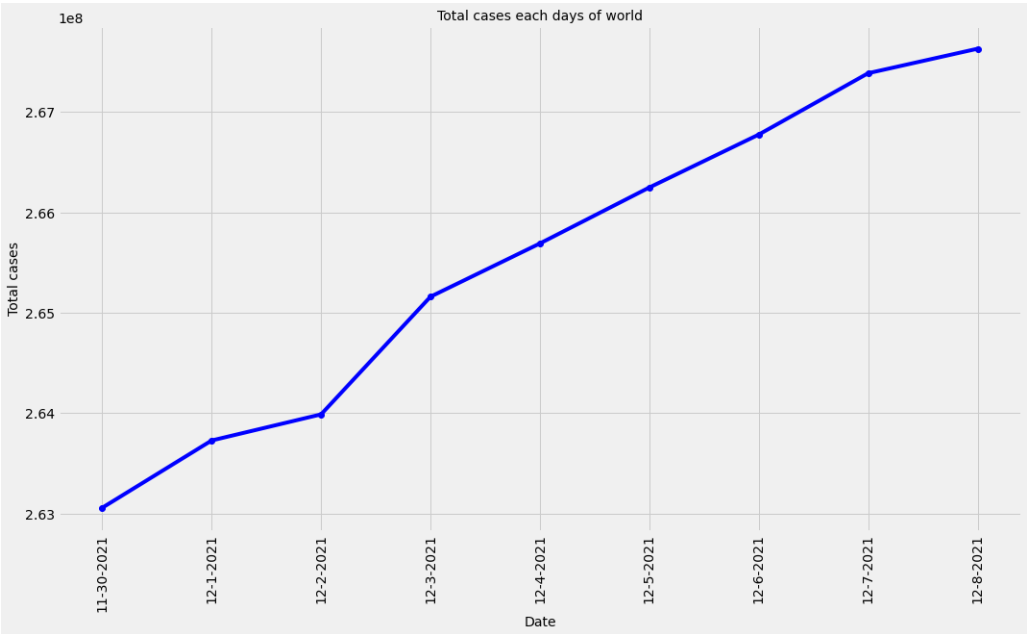
- Khi nhìn vào 2 đồ thị trên, có thể thấy rõ ràng rằng, khi số ca tăng qua từng ngày nhưng tỉ lệ phần trăm ca chết vẫn không đổi, chứng tỏ số ca chết tăng qua 1 tuần. Phần màu vàng là số ca nhiễm tăng mạnh qua 7 ngày phản ánh đúng tình hình dịch cả nước hiện nay: mở cửa dần dần các tỉnh thành phố để phát triển kinh tế, dẫn đến số ca nhiễm tăng nhanh. Phần trăm số ca hồi phục giảm cũng là do số ca đang tăng trong mấy ngày gần đây.
- Và dựa vào cả 2 loại đồ thị (đồ thị thể hiện tỉ lệ chết, hồi phục và nhiễm ở các ngày và 2 đồ thị thể hiện rõ tỉ lệ ở 2 ngày cụ thể), ta có thể thấy rằng, khi số ca

nhiễm tăng cao, số ca tử cũng sẽ tăng theo. 2 thuộc tính này có vẻ tỉ lệ thuận với nhau và điều này có vẻ đúng, xuyên suốt từ 30-11 đến 12-12

- Để thấy rõ hơn về sự gia tăng số ca nhiễm ở Việt Nam cũng như thế giới, ta sẽ xét 2 đồ thị sau:
    - Đối với Việt Nam:

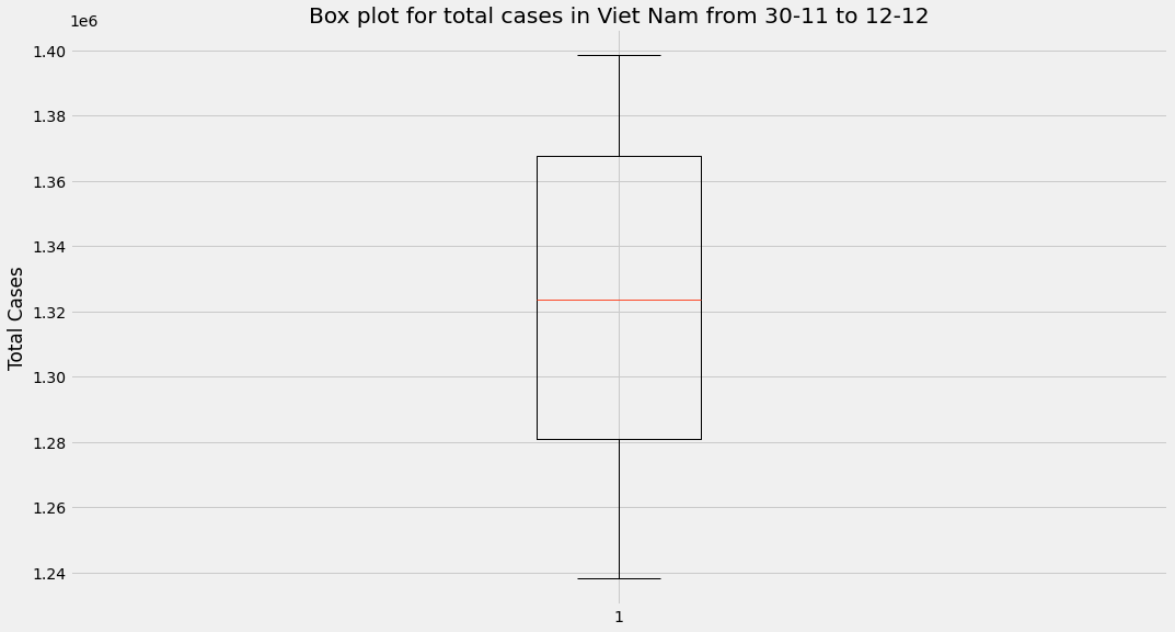


    - Đồ thị tăng thẳng đứng từ ngày 30-11 đến 12-12 (số liệu ngày 12 có thể được lấy ban ngày nên chưa cập nhật), trùng với lúc chủng COVID mới Omicron xuất hiện. Để có cái nhìn tổng quát hơn, ta sẽ xét đến tốc độ tăng của cả thế giới:



- Ta sử dụng line chart để thể hiện rõ sự thay đổi số ca theo thời gian, và theo 2 đồ thị, đồ thị tăng và chưa biết khi nào sẽ đạt được đỉnh. Bởi vậy, cả thế giới vẫn nên đề phòng với dịch chứ không nên quá chủ quan.

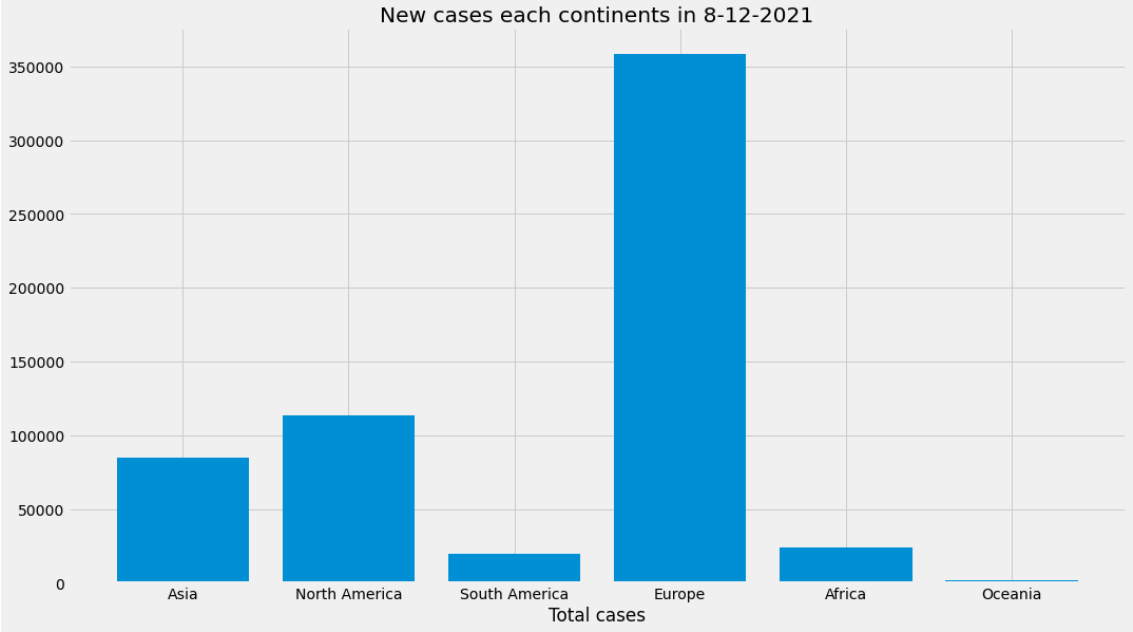
- Để cụ thể hơn với số ca ở Việt Nam, ta sẽ dùng box plot để có thể khám phá thêm các điểm thú vị của dữ liệu:




```
Median:  1323683.0
Upper quartile:  1367433.0
Lower quartile:  1280780.0
```

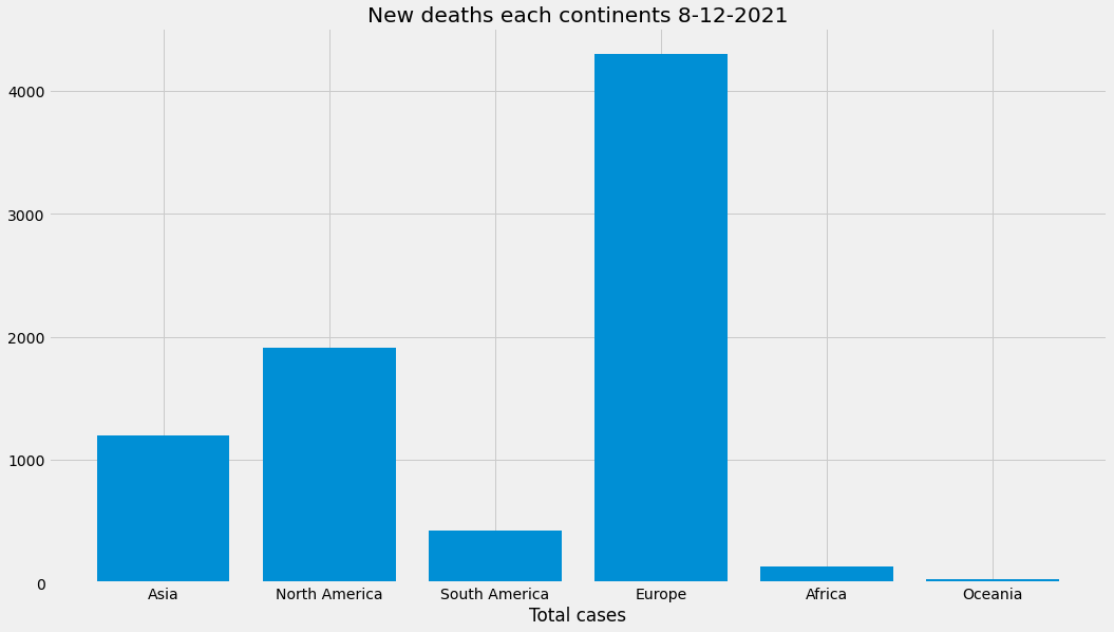
  - Từ đồ thị box ở trên, ta thấy rằng từ khoảng thời gian 30-11 đến 12-12, tổng số ca tăng từ khoảng 1.24 đến 1.4 tỷ ca.
  - Số ca tăng khá đều, bên trong hộp không lệch về phía nào của đường mean nên đây là tập dữ liệu đối xứng
  - Trong các ngày ở giữa, ta thấy số lượng của tứ phân vị thứ nhất và thứ ba hơi lệch lên phía trên, ta thấy số ca tăng trong những ngày này cũng khá nhanh.
  - Lý do sử dụng biểu đồ dạng hộp là để thấy được một số thông tin quan trọng như min, max, Q1, Q3, mean.

- Cuối cùng, ta sẽ xét mối tương quan về số ca tử vong và số ca mắc mới ở mỗi châu lục với nhau vào ngày 8-12-2021:
  - Với số ca mắc mới:



- Từ đồ thị trên, ta thấy được rằng Châu Âu đang là châu lục phức tạp nhất, bởi đây là Châu lục đã mở cửa lại biên giới giữa các nước, và tổ chức nhiều các hoạt động tập trung đông người.
- Cộng với biến chủng mới lan nhanh, đây là một trong những lý do khiến số ca ở châu Âu bùng lên trông thấy.
- Ở các châu lục khác, số liệu cho thấy số ca vẫn ở mức "cao" ngoại trừ châu đại dương (vì châu lục này có khá ít quốc gia)
- Bắc Mỹ có số ca mắc cao thứ 2 bởi ở đây có nước Mỹ lớn mạnh, với tình hình những ngày gần đây khá phức tạp

o Với số ca chết mới:



- Số liệu vẫn là nhiều nhất nghiêng về Châu Âu, Bắc Mỹ và Châu Á.
- Như phân tích đối với số ca mắc mới, số ca tử vong mới cũng như thế.
- Từ 2 đồ thị trên, ta càng thấy được mối quan hệ giữa số ca mắc và số ca tử vong, khi số ca mắc càng nhiều thì số ca tử vong càng nhiều.

THAM KHẢO

Boxplot - Geeksforgeeks. (n.d.). Retrieved from https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/

Drive - Introduction to DS, visualization. (n.d.). Retrieved from https://drive.google.com/drive/folders/1BKItKiX7qsEzLzUfGPhPjssAvySOjS5S

Pie chart - Geeksforgeeks. (n.d.). Retrieved from https://www.geeksforgeeks.org/plot-a-pie-chart-in-python-using-matplotlib/

Stacked bar chart - Geeksforgeeks. (n.d.). Retrieved from https://www.geeksforgeeks.org/create-a-stacked-bar-plot-in-matplotlib/