ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

TRẦN HOÀNG NAM

19127481

BÁO CÁO ĐỒ ÁN

VISUALIZATION

Chuyên ngành: Khoa Học Dữ Liệu

Thành phố Hồ Chí Minh – 2021

I. MỨC ĐỘ HOÀN THÀNH

Thu thập dữ liệu (50/50): Đã hoàn toàn thu thập dữ liệu thành công trong khoảng ngày 28/11 – 9/12

Visualization

Mô tả xu hướng	Sử dụng line chart(5 biểu đồ)
So sánh số lượng người chết ở các khu vực	Sử dụng bar chart (1 biểu đồ)
Xem xét tỉ lệ Deaths/Total case các khu vực	Sử dụng pie chart (7 biểu đồ)
Xem xét mối quan hệ nhân quả của tỉ lệ test và tỉ lệ người chết	Sử dụng bar chart(3 biểu đồ)

II. THU THẬP DỮ LIỆU

Chúng ta sẽ thu thập dữ liệu từ trang web https://www.worldometers.info/coronavirus/, dữ liệu chúng ta thu thập sẽ là bảng Reported Cases and Deaths by Country or Territory bằng file PARSE_HTML notebook

#	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	Population
	World	269,985,626	+14,781	5,317,622	+302	242,718,281	+10,709	21,949,723	88,823	34,637	682.2			
1	USA	50,762,671		817,789		39,986,483		9,958,399	14,935	152,074	2,450	773,300,851	2,316,642	333,802,450
2	India	34,684,396		475,128		34,114,331		94,937	8,944	24,782	339	654,627,300	467,735	1,399,568,826
3	Brazil	22,188,179		616,859		21,414,318		157,002	8,318	103,326	2,873	63,776,166	296,994	214,738,715
4	UK	10,771,444		146,387		9,453,429		1,171,628	900	157,476	2,140	374,468,898	5,474,658	68,400,424
5	Russia	9,986,967		288,351		8,709,964		988,652	2,300	68,392	1,975	231,000,000	1,581,924	146,024,756

- Dữ liệu chúng ta thu thập được sẽ được lưu vào file theo định dạng ‘ngày-tháng-năm.csv’
- Chúng ta sẽ tiến hành thu thập dữ liệu ngày hôm trước của ngày hiện tại, nguyên nhân là số liệu hiện tại trong ngày có thể bị thay đổi theo gian, do đó, để đảm bảo dữ liệu không thay đổi theo thời gian, chúng ta sẽ tiến hành thu thập dữ liệu của ngày trước đó.

  Ví dụ hôm nay là ngày 10/12/2021 thì chúng ta sẽ thu thập dữ liệu của ngày 9/12/2021
- Kết quả dữ liệu thu thập được sẽ nằm trong thư mục Data

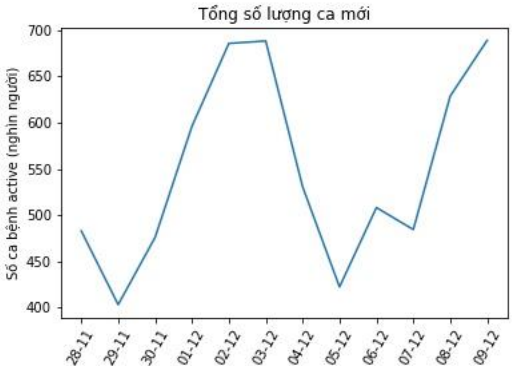
Sau khi thu thập dữ liệu bằng file PARSE_HTML, chúng ta sẽ dùng file PHANTICH để visualize

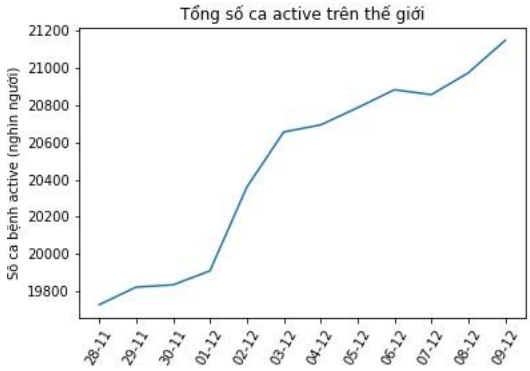
III. VISUALIZE

1. Xu hướng của số lượng ca bệnh

Đầu tiên, chúng ta sẽ bắt đầu xem xu hướng của số ca active hiện nay, tổng số lượng ca mới trên thế giới

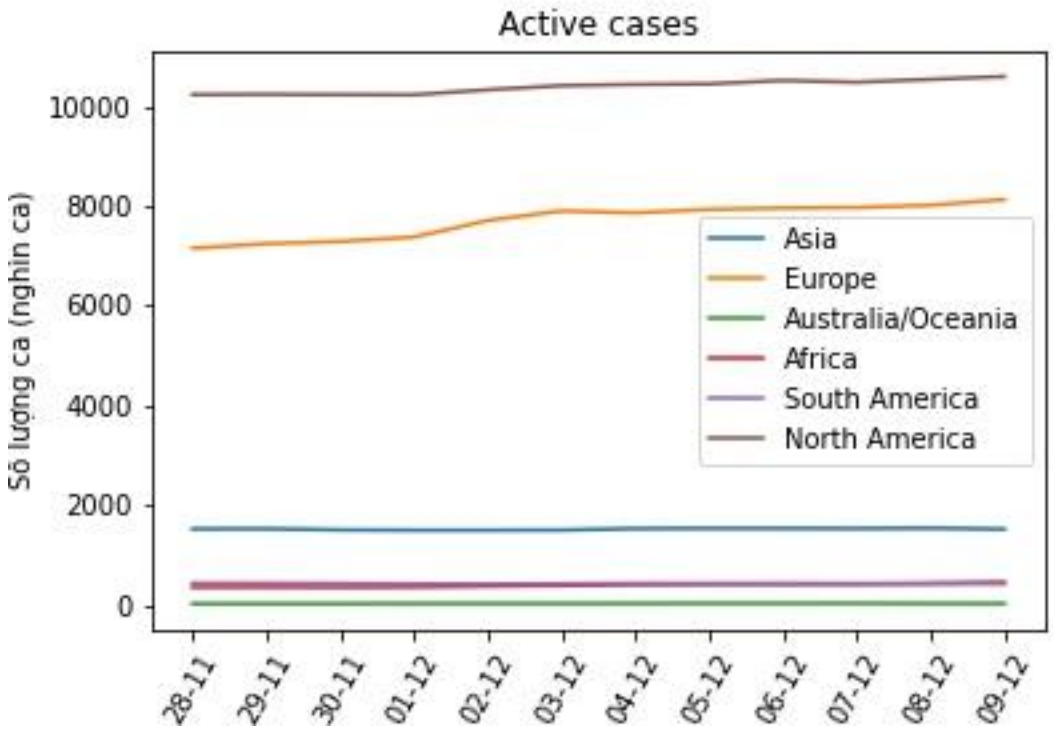
Do chúng ta cần thể hiện xu hướng của số lượng ca bệnh nên chúng ta sẽ dùng biểu đồ đường (line chart) để thể hiện





Như chúng ta thấy ở trên, tổng số ca active hiện tại trên thế giới có xu hướng tăng và không có dấu hiệu nào sẽ giảm, tuy nhiên, tổng số lượng ca mới mỗi ngày thì lại có thời điểm tăng giảm không rõ ràng.

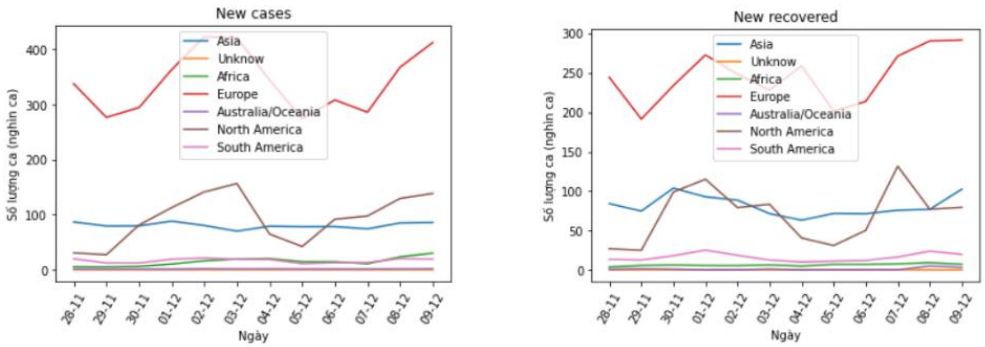
Để tìm hiểu kỹ hơn, chúng ta sẽ xem xét riêng từng khu vực bao gồm EU,Asia,Africa,North America, South America,Australia/Oceania, ở đây do một số nước không cập nhật khu vực trên web, nên chúng ta sẽ để những nước đó ở một khu vực riêng là unknow, chúng ta tạm thời không xét đến các nước này do thiếu hụt thông tin về khu vực



Bảng số liệu (nghìn ca)

Region	28-11	29-11	30-11	01-12	02-12	03-12	04-12	05-12	06-12	07-12	08-12	09-12
Asia	1,521	1,525	1,502	1,496	1,496	1,499	1,527	1,531	1,531	1,529	1,536	1,518
Africa	353	352	351	356	373	386	402	410	410	413	430	453
Europe	7,157	7,246	7,293	7,379	7,714	7,905	7,866	7,936	7,962	7,973	8,018	8,135
Australia/Oceania	22	22	22	23	24	24	25	26	26	28	24	23
North America	10,241	10,245	10,240	10,236	10,337	10,421	10,444	10,455	10,523	10,489	10,546	10,604
South America	433	432	427	419	417	423	429	429	431	426	421	418

Dựa theo biểu đồ và bảng kết quả trên thì số lượng ở các khu vực ít ca như south America và Australia đang có xu hướng giảm nhẹ, tuy nhiên các khu vực có nhiều ca như EU,NA lại có xu hướng tăng cực kỳ nhanh, khá bất ngờ là Asia lại giảm mạnh số ca active



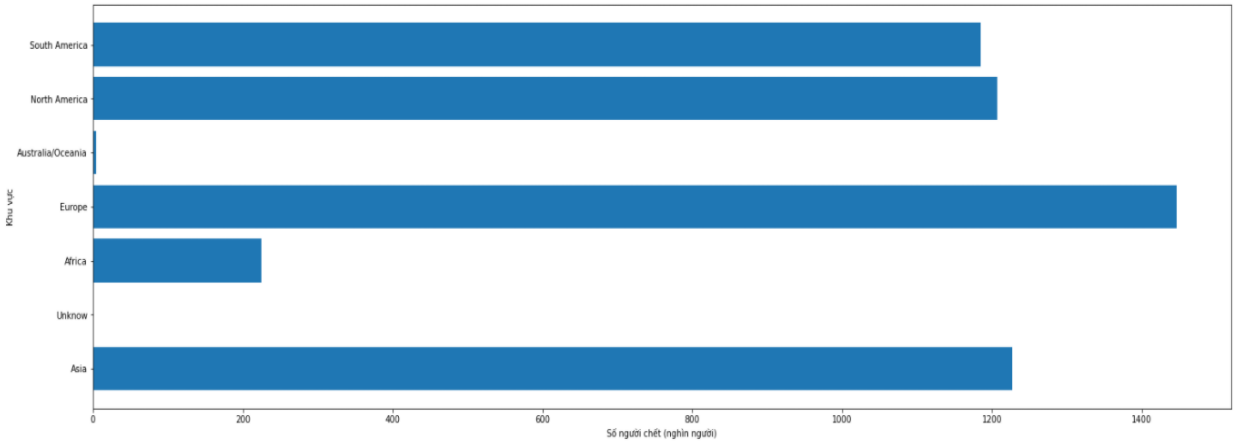
- Chúng ta sẽ xem xét đến số ca mới mỗi ngày và các ca hồi phục mỗi ngày ở các khu vực

Như chúng ta thấy trên hình, số ca mới nhiễm mỗi ngày có thể lên tới khoảng 400 nghìn ca, trong khi đó, số ca hồi phục mỗi ngày tối đa chỉ khoảng 300 nghìn

⇨ có thể nói covid 19 có tốc độ lây lan rất khủng khiếp, không lạ khi tổng số ca active luôn có xu hướng tăng chứ không hề có xu hướng giảm, dù cho số lượng ca mới mỗi ngày có thể tăng giảm nhưng luôn vượt trội hơn số ca hồi phục mỗi ngày

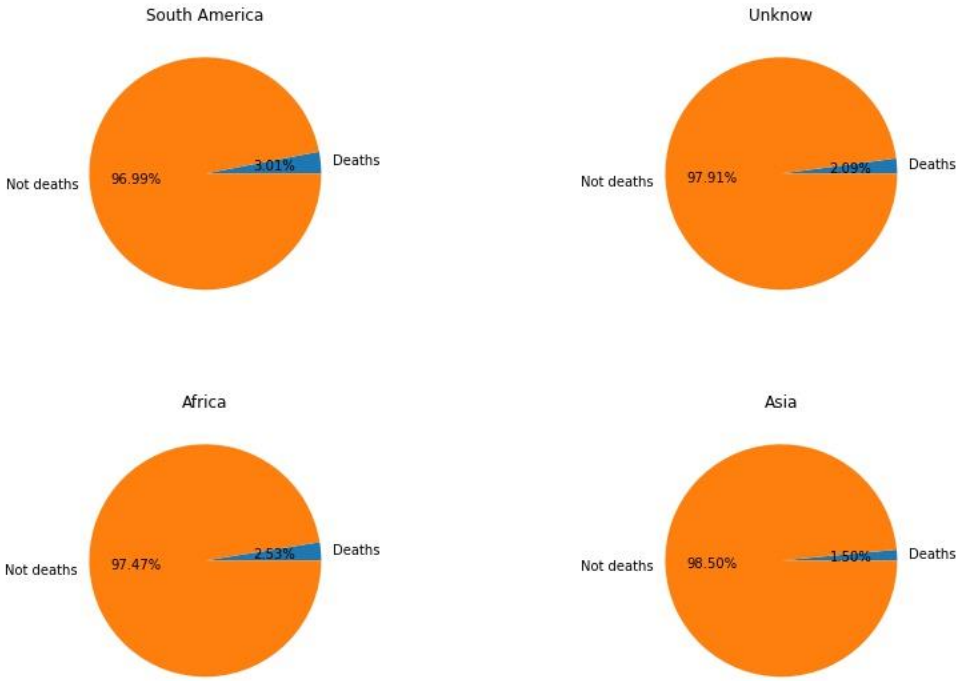
2. So sánh số lượng người chết giữa các khu vực

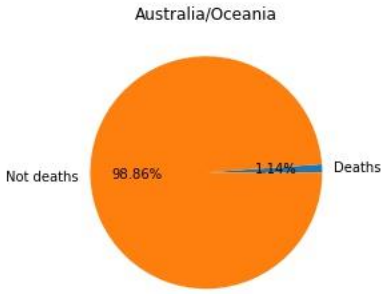
Vì đây là so sánh giữa các category nên chúng ta sẽ dùng biểu đồ cột, bar chart để dễ so sánh

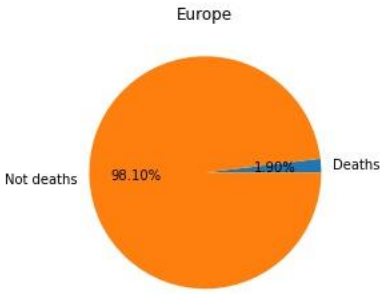


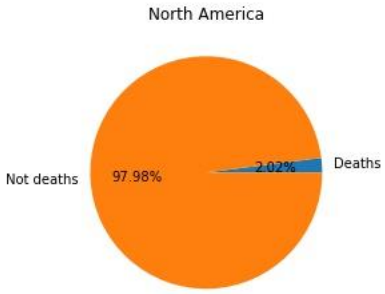
Tính tới thời điểm hiện tại số lượng người chết ở khu vực EU là cao nhất và khá cách biệt so với phần còn lại. Tuy nhiên số ca chết ở South America, North America và Asia là khá gần nhau

Chúng ta sẽ dùng biểu đồ tròn (pie chart) để xem xét tỉ lệ số ca chết/tổng số vì biểu đồ tròn thể hiện tỉ lệ









Tổng kết

Continent	
Australia/Oceania	1.1276936441
Asia	1.4819742956
Europe	1.8860503102
North America	2.0056568232
Unknow	2.0804438280
Africa	2.5281675983
South America	3.0277418001

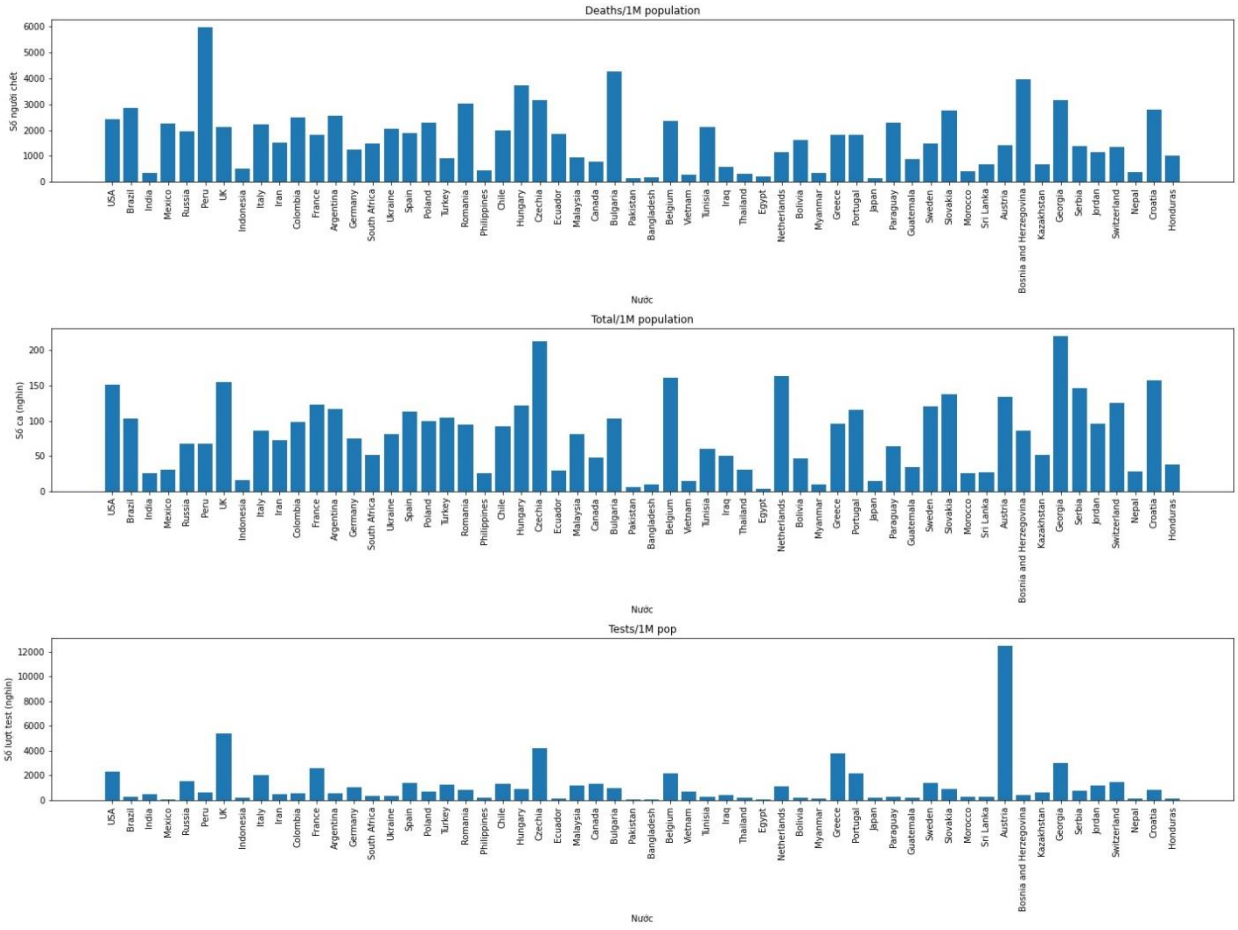
⇨ Rõ ràng EU dù đang có số lượng ca chết do nhiễm bệnh cao nhất thế giới, thế nhưng tỉ lệ chết lại đứng thứ 5, chỉ cao hơn Australia và Asia

⇨ Ngược lại với điều đó, Africa tuy có số lượng nhiễm bệnh thấp, nhưng tỉ lệ chết lại rất cao

⇨ Liệu có thể kết luận rằng thể trạng người châu Phi rất khoẻ, do đó mà người miễn dịch thì sẽ không mắc bệnh, nhưng người mắc bệnh thì là do hệ miễn dịch không đủ khoẻ không?
South America có số lượng người chết cao thứ 4, chênh lệnh không đáng kể với NA và Asia, nhưng lại có tỉ lệ chết cao nhất trong số các khu vực

3. Mối quan hệ nhân quả

Chúng ta sẽ nhìn sơ qua về cột tests/1m pop, cột tot/1m pop, death/1m pop để kiểm tra xem, liệu có mối liên hệ nào giữ tỉ lệ test và tỉ lệ nhiễm bệnh, tỉ lệ chết do bệnh hay không, do số lượng nước khá nhiều(trên 200 nước) do đó chúng ta chỉ lấy những nước có số lượng người chết >10000 để xem xét



Như chúng ta thấy trên hình, dường như nước nào có tỉ lệ tests/1M thì dù cho số ca nhiễm bệnh Total/1M cao thì số ca chết lại thấp, trong khi ngược

lại, số lượt test/1M càng thấp thì tỉ lệ Deaths/1M lại càng cao dù cho số ca nhiễm bệnh thấp.