ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO

Ex02: Thống kê và trực quan hóa dữ liệu

Họ và tên: Võ Nhất Huy

MSSV: 19127642

Nhập môn khoa học dữ liệu

1. **Tổng quan bài làm**

   - Ngôn ngữ: Python
   - Môi trường code: min_ds-env + thư viện plotly 5.2.1 + selenium
   - Các file bài làm: crawl_corona.ipynb và processing.ipynb
   - Files chứa data: Corona_day.tsv và Corona_week.tsv

2. **Thu thập dữ liệu**

   - Dữ liệu được thu thập từ 2 trang https://www.worldometers.info/coronavirus/#main_table và https://www.worldometers.info/coronavirus/weekly-trends/#weekly_table .
   - Sử dụng selenium để load web page, sau đó xuất ra source html rồi sử dụng Beautiful Soup để lấy dữ liệu.

```
driver.get("https://www.worldometers.info/coronavirus/")

html = driver.page_source
```

```
html
soup = BeautifulSoup(html, 'html.parser')
df = crawl_raw_data(soup)
```

   - Dữ liệu trong html được driver trả ra bao gồm dữ liệu của cả 3 ngày, tương ứng với 3 bảng “Now", “Yesterday" và “2 Days Ago". Tuy nhiên để đảm bảo chất lượng dữ liệu, ta không lấy bảng “Now" mà chỉ lấy “Yesterday" và “2 Days Ago".
   - Lí do là bởi dữ liệu tại bảng “Now" không đầy đủ và vẫn được cập nhật trong ngày. Nếu như lấy thì sau đó cũng sẽ phải bổ sung lại, như thế là không cần thiết.
   - Do vậy, chỉ cần lấy dữ liệu của 2 ngày trước, sau đó lặp lại chu kì 2 ngày lấy dữ liệu 1 lần để thu thập dữ liệu trong nhiều ngày.
   - Đối với dữ liệu ở “Weekly trends", thực hiện lấy dữ liệu vào mỗi thứ 2 hàng tuần.

- Khi thu thập dữ liệu tuần, cần markdown lại cell chứa code thu thập dữ liệu ngày, và ngược lại (vì dữ liệu thu thập được cũng khá nhiều, nên cá nhân em cũng không muốn phải sửa code để đề phòng bất trắc)

3. Tiền xử lí dữ liệu

- Sau khi thu thập dữ liệu xong, ta thực hiện bước tiền xử lí.
- Dữ liệu ban đầu được thu thập đều nằm ở dạng string, đồng thời cũng có một số dòng dữ liệu bị thiếu. Do đó để đảm bảo cho công việc tính toán được thuận lợi, ta xử lí các ô trống và chuyển các cột chứa số liệu về dạng numeric. Các quốc gia với tên châu lục bị bỏ trống sẽ được thay bằng “Other"

```
# read file
def read_file():
    df_day = pd.read_csv("Corona_day.tsv", sep = '\t', skipinitialspace = True, thousands = ',')
    df_week = pd.read_csv("Corona_week.tsv", sep = '\t' , skipinitialspace = True, thousands = ',')
    return df_day, df_week

def auto_divide_dataframe(df_, n_diff, by_column):
    if n_diff == 1:
        return df_
    list_country = df_[by_column].unique()
    list_df = list()
    grouped = df_.groupby(df_[by_column])
    for i in range(len(list_country)):
        list_df.append(grouped.get_group(list_country[i]))
    return list_df
```

✓ 0.2s

```
df_day, df_week = read_file()
df_day["Continent"].replace({np.nan: 'Other'}, inplace=True)
df_day.fillna(0, inplace = True)
```

✓ 0.6s

- 2 data frame df_day và df_week sau đó được tách ra thành các list chứa dictionary để phụ vụ cho việc trực quan hóa dữ liệu. Vì dữ liệu được thu thập trong nhiều ngày, đồng thời có quá nhiều quốc gia nên chúng cần được tách ra để xử lí tùy theo mục đích của người thực hiện thống kê và trực quan hóa.
- Nhìn chung, với mỗi data frame, sẽ có 3 nhóm dữ liệu có thể trích ra từ đó: dữ liệu của cả thế giới trong 1 chu kì thời gian, của châu lục và của các quốc gia.

Chi tiết code nằm bên trong từng mục markdown

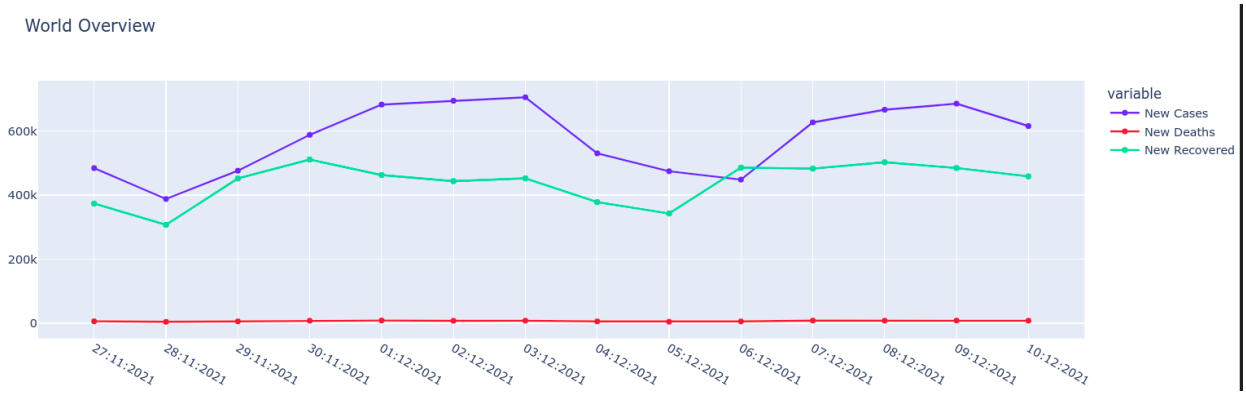
4. Thống kê và trực quan hóa

❖ *Tổng quan về tình hình thế giới*

- Trước hết, ta cần phải có cái nhìn tổng quát về tính hình dịch bệnh trên thế giới trong những ngày gần đây. Các trường dữ liệu được sử dụng ở đây sẽ là “New Cases", “New Deaths" và “New Recovered" được biểu diễn qua từng ngày (“Date")

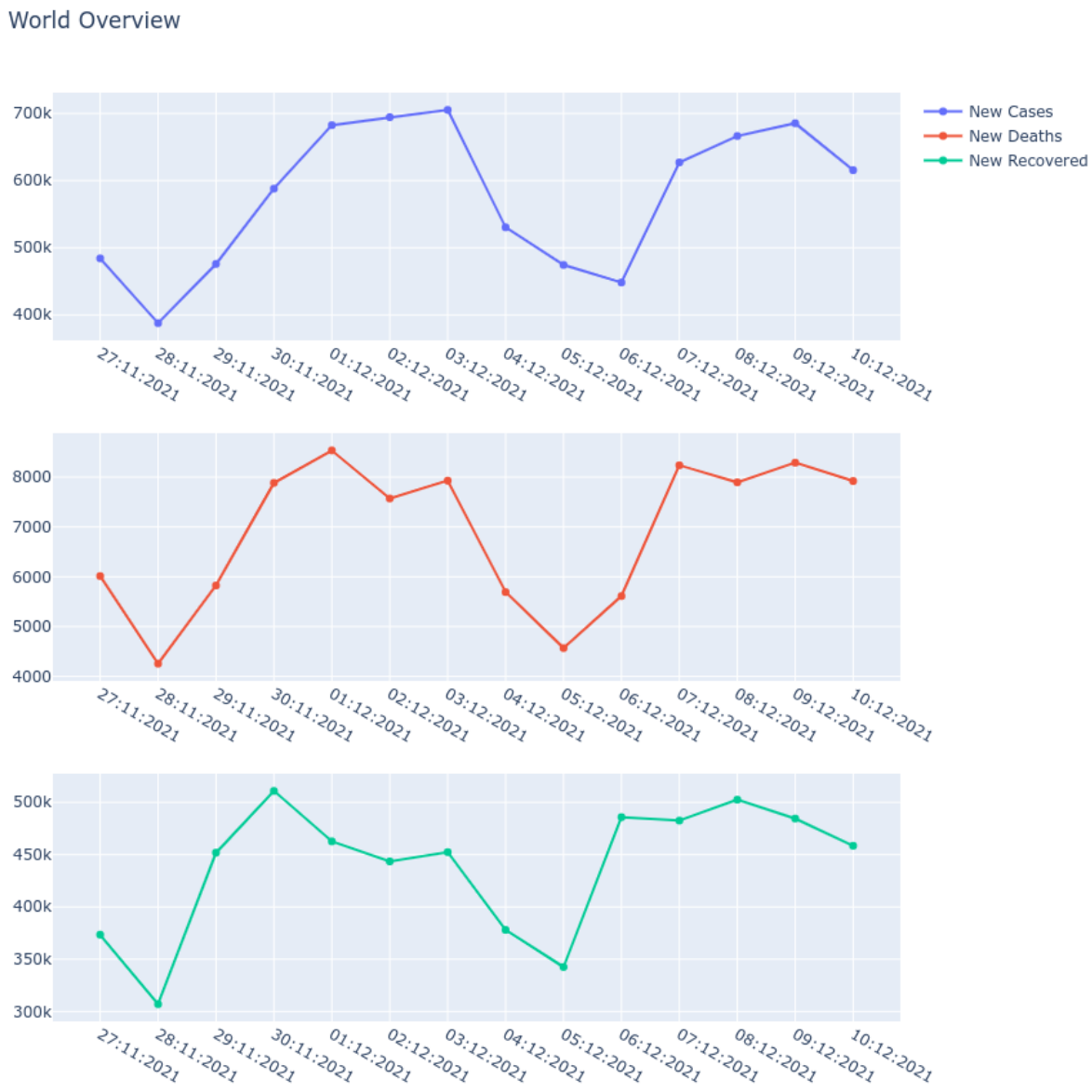
	Date	New Cases	New Deaths	New Recovered
0	27:11:2021	484205.0	6014.0	373473.0
1	28:11:2021	387788.0	4257.0	307047.0
2	29:11:2021	475969.0	5825.0	451875.0
3	30:11:2021	587982.0	7883.0	510886.0
4	01:12:2021	682525.0	8535.0	462696.0
5	02:12:2021	694394.0	7571.0	443518.0
6	03:12:2021	705204.0	7928.0	452414.0
7	04:12:2021	530432.0	5692.0	378011.0
8	05:12:2021	474423.0	4569.0	342523.0
9	06:12:2021	448248.0	5612.0	485818.0
10	07:12:2021	627031.0	8237.0	482526.0
11	08:12:2021	666364.0	7895.0	502513.0
12	09:12:2021	685385.0	8289.0	484503.0
13	10:12:2021	615462.0	7921.0	458478.0

- Khi đã chọn được các trường dữ liệu, giờ ta chỉ việc trực quan hóa nó lên. Biểu đồ được sử dụng ở đây là line chart.



- Tuy nhiên khi trực quan hóa biểu đồ lên, ta dễ dàng thấy được rằng chúng không có nhiều sự liên quan. Đồng thời cũng không thể hiện rõ được trường “New Deaths", trong khi giữa các ngày ta đều thấy được sự biến động về số liệu trong trường này

- Ví thế, ta cần tách 3 trường này ra thành 3 biểu đồ độc lập.



- Nhận xét:
  - Số các ca mắc trên thế giới vẫn còn nhiều, chứng tỏ tình hình dịch bệnh vẫn còn đang rất nguy hiểm
  - Tuy nhiên tỉ lệ ca tử vong là rất ít, điều này có thể thấy được phần nào thông qua biểu đồ "lỗi" ở trên.
  - Có sự biến động tương đối đồng đều ở cả 3 biểu đồ. Khả năng sự suy giảm đột ngột về số ca trong cùng 1 ngày là do có tác động của yếu tố khác, khiến dữ liệu không đầy

đủ?!(Chẳng hạn như ngày 28/11 và 5/12 đều là chủ nhật, nên việc cập nhật số liệu có vấn đề…)

❖ *Tổng quan về tình hình các châu lục*

- Sau khi đã có được cái nhìn tổng quát về tình hình dịch trên thế giới, ta tìm hiểu xem số liệu của các châu lục như thế nào.

	Continent	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	New Recovered
0	North Ameica	834441750	1494673.0	16797510.0	18033.0	6.715734e+08	1048838.0
1	Asia	1153840887	1162477.0	17076963.0	18395.0	1.115540e+09	1152745.0
2	South America	546805148	248206.0	16567358.0	4962.0	4.957339e+08	231136.0
3	Europe	1050597407	4893967.0	20008024.0	52910.0	9.220848e+08	3588960.0
4	Africa	123271500	242710.0	3138079.0	1791.0	1.140664e+08	94543.0
5	Australi/Ocen	5242858	23379.0	59539.0	137.0	4.215709e+06	20059.0

- Cũng với 3 trường dữ liệu trên, và thêm vào đó là các trường dữ liệu "Total…" tương ứng. Chúng ta dựa vào biểu đồ có thể dễ dàng thấy được sự khác biệt giữa tình hình dịch bệnh trong những ngày gần đây, so với cả quá trình từ khi virus xuất hiện.

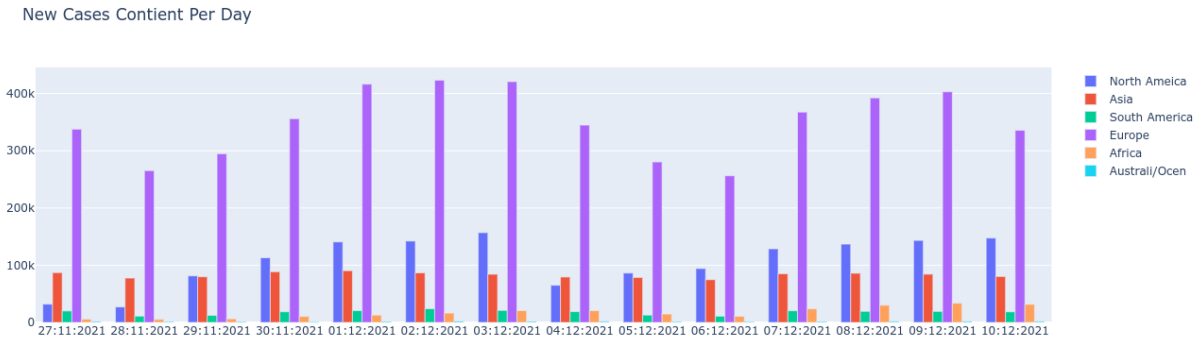
- Nhận xét:

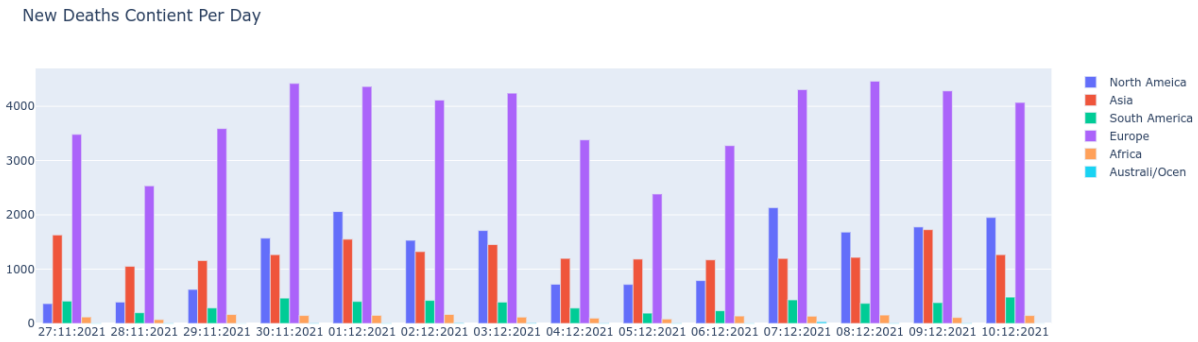
  - Các quốc gia châu Âu đang dần chịu ảnh hưởng nặng nề hơn hẳn so với các nước từ châu lục khác.

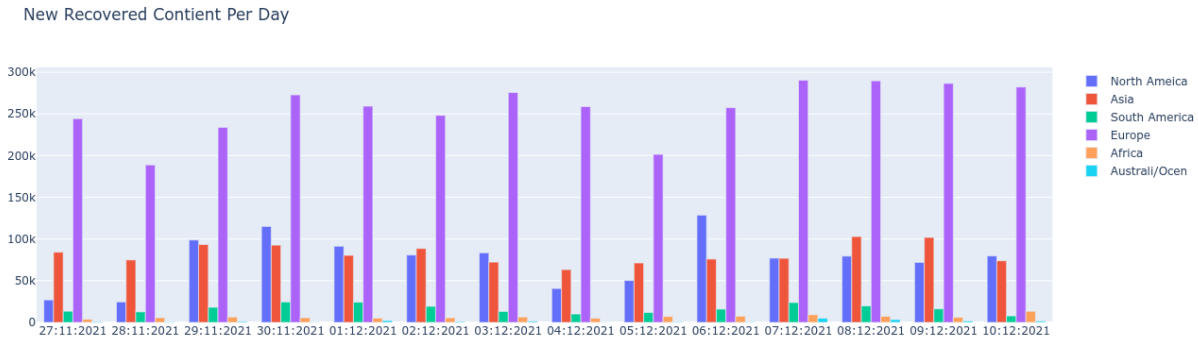
  - Châu Á vì chịu ảnh hưởng trực tiếp từ Trung Quốc và Ấn Độ, nên ở các biểu đồ "Total…" ta có thể thấy số ca mắc và tử vong chiếm tỉ lệ khá lớn.

  - Tuy nhiên, do tỉ lệ người cao tuổi không cao bằng so với các nước châu Âu, cũng như nhờ các biện pháp cứng rắn từ những ngày đầu có dịch nên tổng thể số ca khỏi bệnh ở châu Á vẫn cao hơn, và ngược lại với số ca tử vong.

❖ *Số các ca mắc, nhiễm và tử vong mỗi ngày*



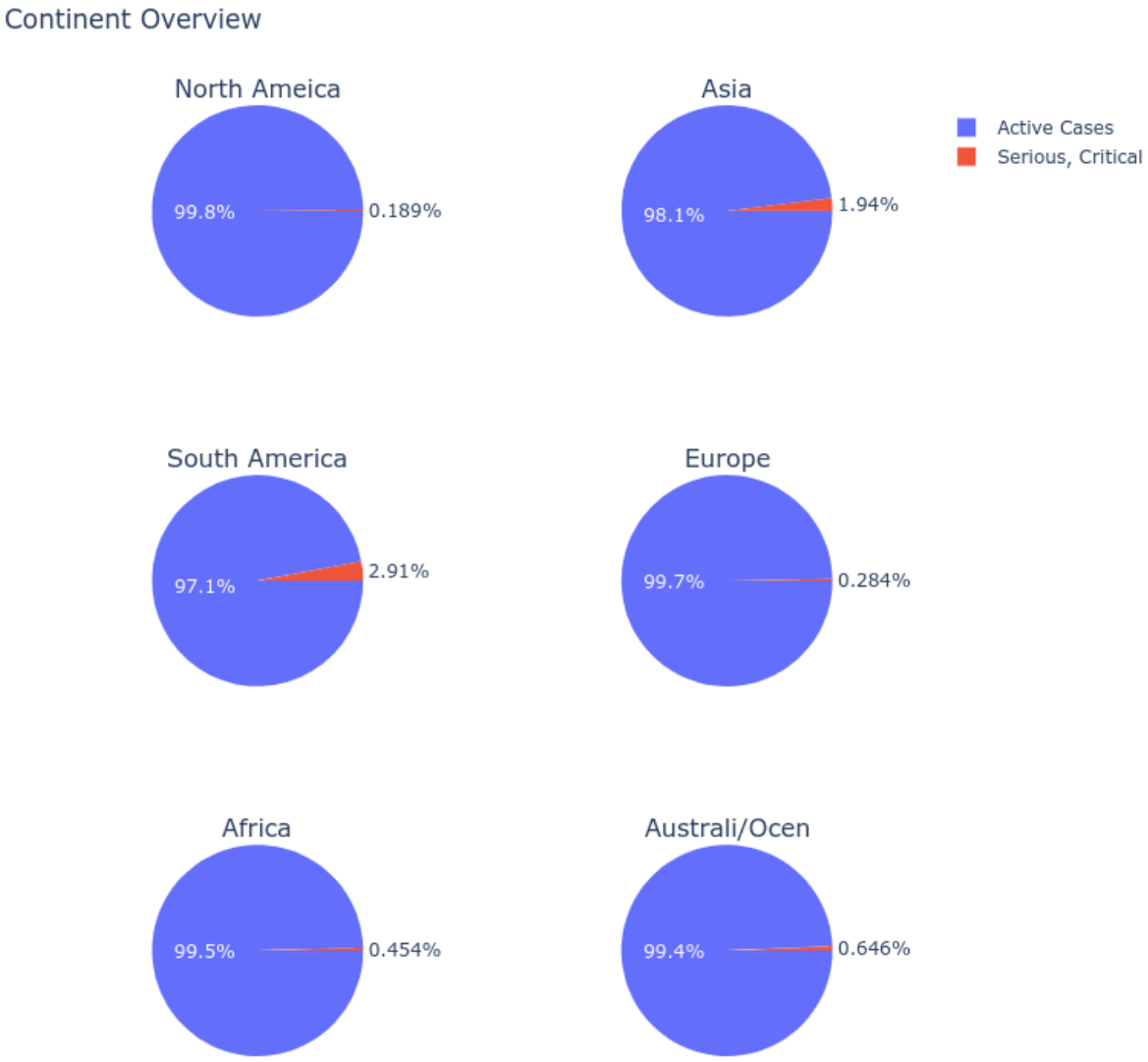




- Nhận xét:
  - Nhìn chung, số liệu từ những ngày gần đây khá tương đồng với những gì ta thấy được trong biểu đồ chung về tình hình các châu lục phía trên
  - Sự biến động về số ca ở châu Âu cũng tương đồng với tình hình thế giới. Có lẽ vì số ca ở cả 3 biểu đồ thể hiện đều chiếm tỉ lệ rất lớn, do đó cũng ảnh hưởng tới biểu đồ về tình hình thế giới.

❖ *Tỉ lệ số ca nguy kịch so với các ca bệnh hiện tại*

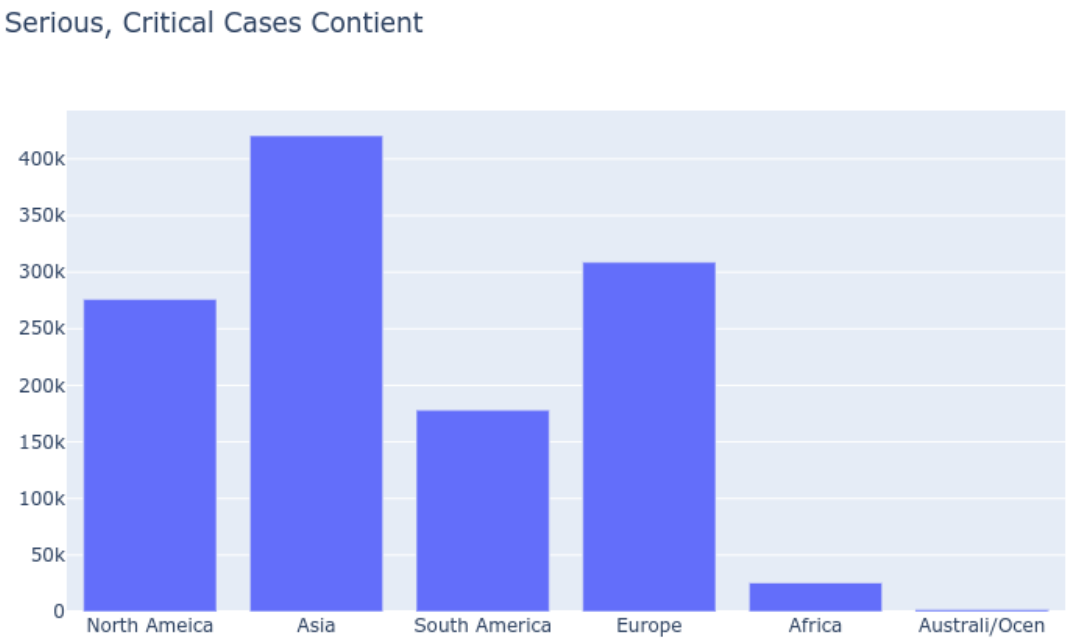
	Continent	Serious, Critical	Active Cases	%
0	North Ameica	276141.0	146009841.0	0.19%
1	Asia	420570.0	21223596.0	1.98%
2	South America	178224.0	5940878.0	3.0%
3	Europe	308793.0	108504561.0	0.28%
4	Africa	25639.0	5621380.0	0.46%
5	Australi/Ocen	2152.0	330821.0	0.65%



- Đến đây chúng ta lại nhìn thấy một số khía cạnh khác về tình hình dịch bệnh ở từng châu lục. Cụ thể:
  - Khác với suy đoán, châu Âu không phải là nơi có tỉ lệ các ca nguy kịch nhiều nhất, nếu như dựa vào tỉ lệ tử vong trong những ngày gần đây.
  - Thay vào đó, Nam Mỹ và châu Á lại có các ca bệnh nguy kịch chiếm phần lớn hơn so với các châu lục còn lại.
  - Lý giải cho điều này có thể có những nguyên nhân như: tuy tỉ lệ ca nguy kịch cao hơn, nhưng số ca mắc ít hơn khiến cho số lượng các ca tử vong không nhiều bằng;

hoặc do điều kiện y tế không được tốt, dẫn đến tỉ lệ các ca đang nguy kịch nhiều hơn; hoặc số liệu về các ca nguy kịch vì chưa tử vong nên không được tính vào ca chết…

- Để chứng minh cho một số suy luận trên, ta có thể nhìn vào biểu đồ sau:

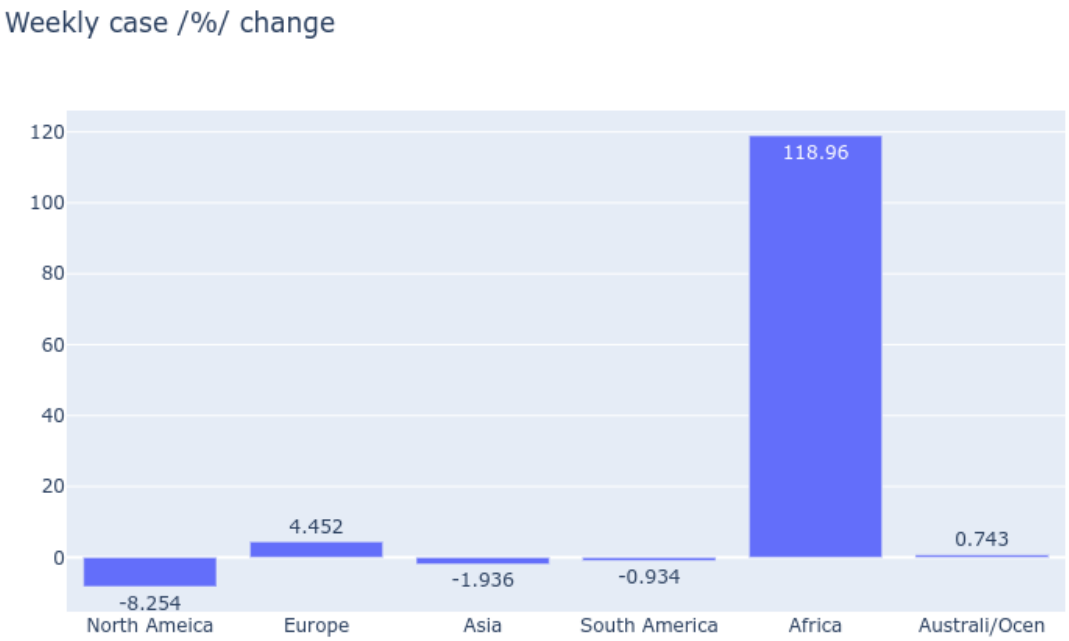


❖ *So sánh tỉ lệ ca mắc và tử vong với tuần trước đó*

- Ở trên chúng ta đã nói về tình hình của châu lục trong 2 tuần thu thập dữ liệu, vậy so với trong tuần gần nhất, tỉ lệ các ca mắc có sự thay đổi như nào, so với tuần trước đó?

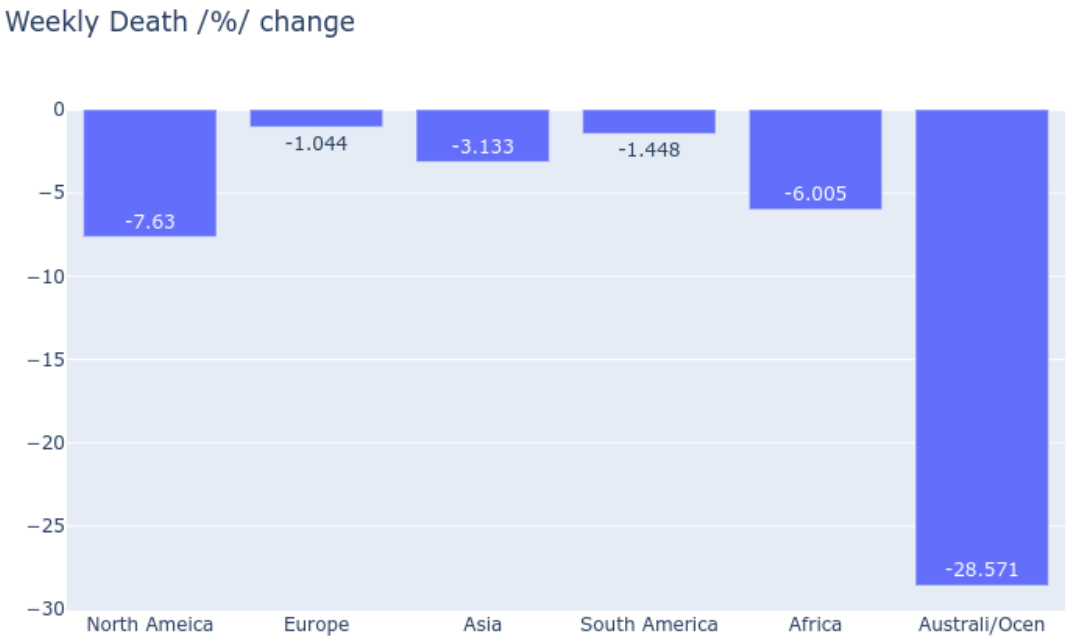
	Continent	Weekly case /%/ change
0	North Ameica	-8.254
1	Europe	4.452
2	Asia	-1.936
3	South America	-0.934
4	Africa	118.960
5	Australi/Ocen	0.743

- Bất ngờ thay, tỉ lệ ca mắc ở châu Phi lại cao bất thường. Có thể nói là số ca mắc ở châu Phi đã tăng hơn gấp đôi so với tuần trước đó.
- Điều này cũng dễ hiểu khi mà đây là châu lục hiện đang ít được viện trợ về y tế nhất, cũng như hệ thống y tế sẵn có tại đây đã không tốt, khiến dịch bệnh lây lan mạnh mẽ



- Bên cạnh đó, hầu hết các châu lục khác đều có tỉ lệ ca mắc giảm, hoặc tăng rất nhẹ.
- Riêng đối với châu Âu, sự tăng nhẹ này lại đáng quan ngại hơn, bởi lẽ bản thân các nước ở đây đã có số ca nhiễm cao sẵn, điều này càng chứng tỏ tình hình dịch bệnh phức tạp nơi đây
- Ngoài số ca mắc ra, chúng ta cũng cần nhìn vào sự thay đổi số ca tử vong

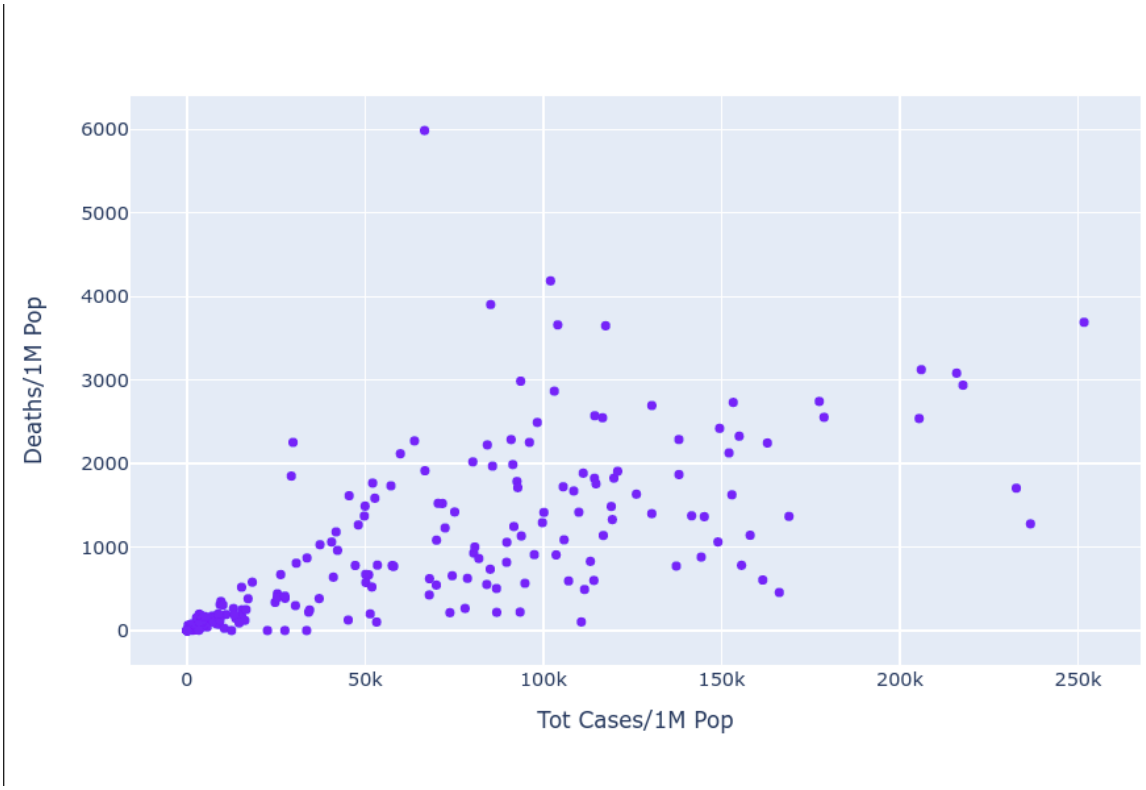
	Continent	Weekly Death /%/ change
0	North Ameica	-7.630
1	Europe	-1.044
2	Asia	-3.133
3	South America	-1.448
4	Africa	-6.005
5	Australi/Ocen	-28.571



- Nhìn chung, số ca tử vong ở tất cả châu lục đều giảm, đây là một tín hiệu lạc quan cho tình hình dịch bệnh hiện nay.

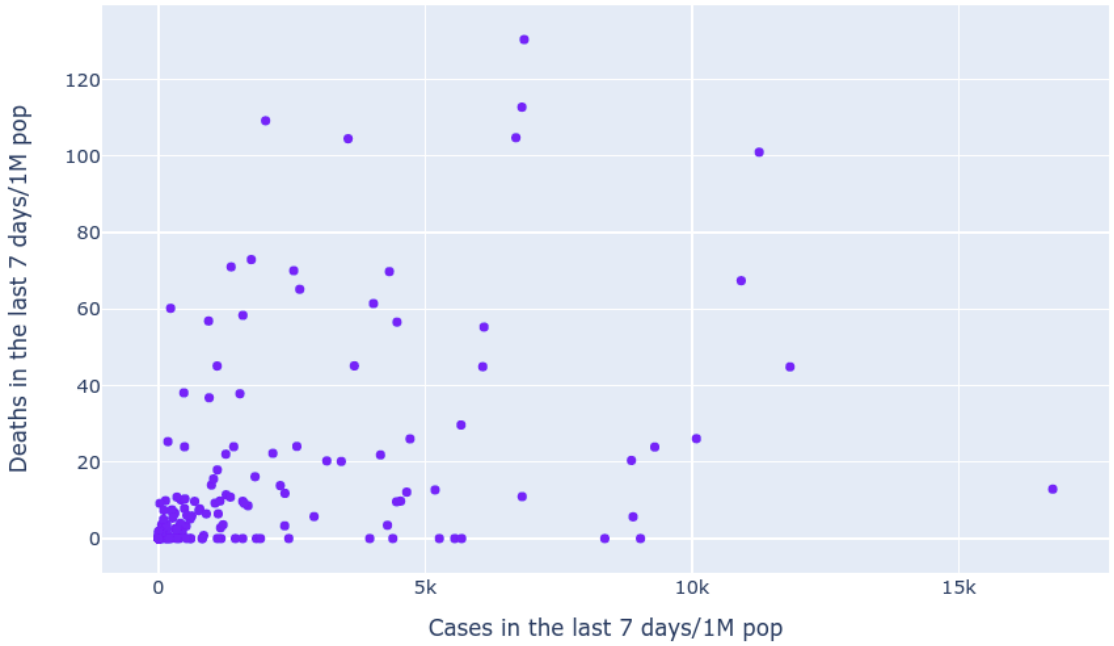
❖ *Mối liên hệ giữa tỉ lệ ca mắc và tỉ lệ ca tử vong*

- Trong quá trình tìm kiếm mối quan hệ nhân quả giữa các trường dữ liệu, dễ nhận ra nhất có lẽ là giữa 2 trường “Tot Cases/1M Pop” và “Tot Deaths/1M Pop"



- Điều này khá là dễ dàng suy luận, vì đơn giản với cùng 1 mật độ dân cư, số ca mắc nhiều hơn sẽ khiến nhiều người tử vong hơn.

- Nhưng đó là với tình hình dịch bệnh tổng thể từ trước tới nay. Nếu so với những ngày gần đây thì như thế nào?



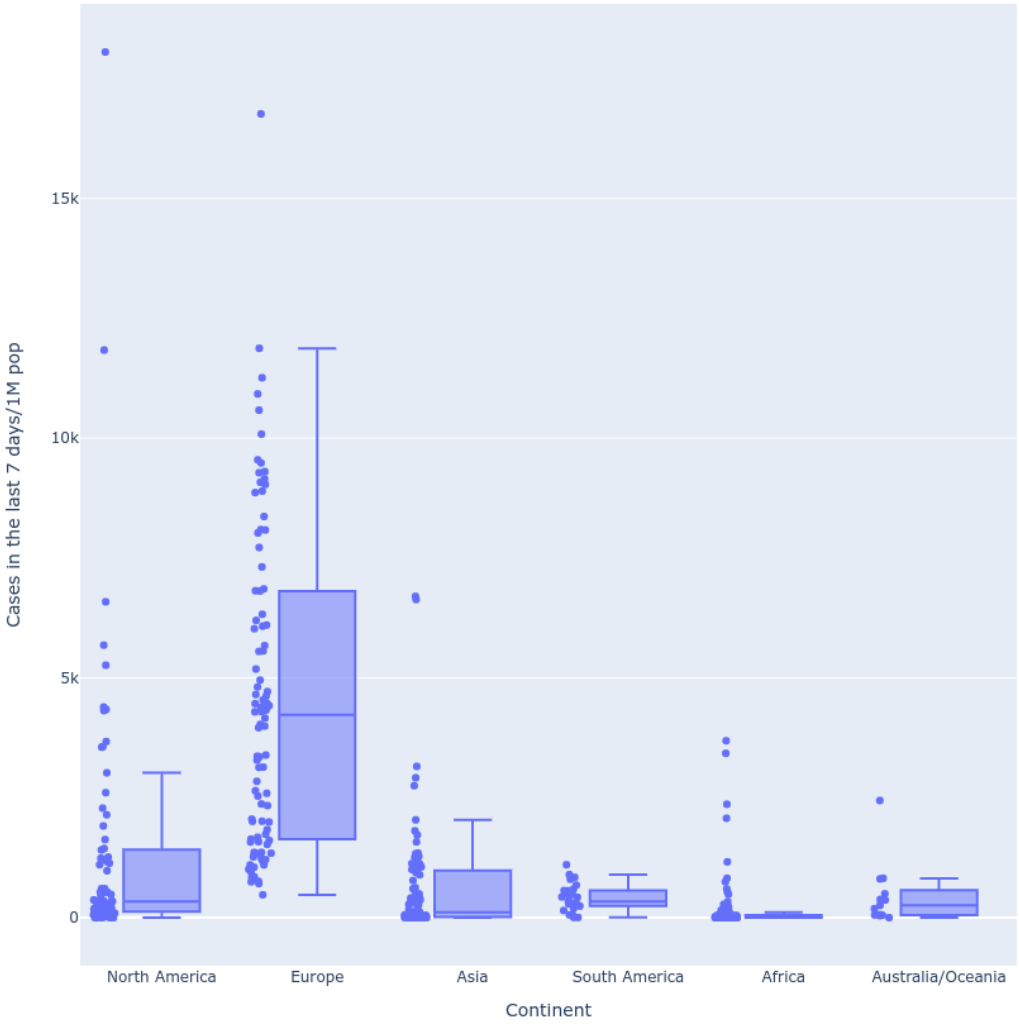
- Có thể thấy được, đối với dữ liệu từ tuần vừa rồi, dường như số ca tử vong và số ca mắc vẫn còn có mối liên hệ, nhưng đã không còn rõ ràng như trước

- Lí do có thể là bởi biến chủng Delta với độc tính mạnh hơn hẳn so với những "phiên bản đời đầu". Tỉ lệ một người mắc bệnh tử vong giờ đây cao hơn, khi mà các ca bệnh diễn biến nhanh và nặng hơn.

❖ *Phân bố số lượng ca mắc giữa các quốc gia ở mỗi châu lục trong 1 tuần vừa qua*

- Chúng ta đã đánh giá về tình hình của các châu lục tương đối nhiều, nhưng liệu có phải tất cả các quốc gia trong châu lục đều chịu chung tình hình?

	Cases in the last 7 days/1M pop			
	median	std	q1	q3
Continent				
Africa	10.5990	575.557166	3.43950	65.44475
Asia	113.8110	1161.603657	10.77200	857.20850
Australi/Ocen	380.3780	844.397688	121.15850	659.44000
Europe	4291.9280	3510.539980	1644.35700	6457.20900
North Ameica	347.5665	2309.022198	128.33850	1434.13675
South America	367.4475	309.640077	263.57125	569.77200



- Nhận xét:
  - Ở biểu đồ này ta lại thêm một lần nữa thấy được sự bùng phát dịch bệnh ở châu Âu, khi mà nhìn chung số các ca bệnh mới ở mỗi quốc gia là khá cao.
  - Nhưng giữa các châu lục thì số lượng ca nhiễm mới có xu hướng nghiêng về phía dưới số ca mắc trung bình trong mỗi châu lục. Điều này được thấy rõ nhất ở Bắc Mỹ và châu Á.
  - Số lượng các ca mắc mới là không đồng đều, thậm chí là chênh lệch rất lớn ở một số nước. Ta có thể thấy ở hầu hết các châu lục đều có một vài, thậm chí là khá nhiều giá trị outliers.

5. Tổng kết, đánh giá

- Nhìn chung bài làm đánh giá được nhiều khía cạnh của dữ liệu thu thập được
- Cách xử lí dữ liệu ở mức ổn, cho phép người phân tích có thể tái sử dụng dữ liệu đã được chuẩn bị để thống kê và trực quan hóa với nhiều trường khác nhau.
- Tuy nhiên bên cạnh đó vẫn còn một số vấn đề như:
  - tổ chức dữ liệu sau khi tiền xử lí chưa tốt, các kiểu cấu trúc dữ liệu bị chồng chéo
  - Nhận định, đánh giá về các biểu đồ chưa đủ sâu để làm rõ vấn đề
  - Code để thu thập dữ liệu chưa hoàn thiện tốt
- Đánh giá mức độ hoàn thành tổng thể: 99/100
  - Thu thập dữ liệu: 98/100
  - Thống kê và trực quan hóa: 100/100

6. Nguồn tham khảo

https://plotly.com/python/box-plots/

https://plotly.com/python/line-and-scatter/

https://plotly.com/python/line-charts/

https://plotly.com/python/bar-charts/

https://plotly.com/python/pie-charts/

https://stackoverflow.com/questions/47637774/pandas-groupby-quantile-values