ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

TRỰC QUAN HOÁ DỮ LIỆU

BÁO CÁO LAB 1

| ĐỀ TÀI |

| GIÁO VIÊN HƯỚNG DẪN |

Thầy Bùi Tiến Lên

Thầy Lê Ngọc Thành

Thành phố Hồ Chí Minh

THÀNH VIÊN NHÓM

MÃ SỐ	HỌ VÀ TÊN
19127472	Nguyễn Bá Minh
19127481	Trần Hoàng Nam
19127595	Nguyễn Minh Trí

I. MỨC ĐỘ HOÀN THÀNH

Hoàn thành thu thập dữ liệu thành công từ ngày 7/3-15/3

Trực quan hoá

Mô tả xu hướng số ca activate và số ca mới
Xem xét tương quan giữa các trường dữ liệu và trực quan hoá các trường
Sử dụng OLS model để xem xét mối quan hệ nhân quả

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.

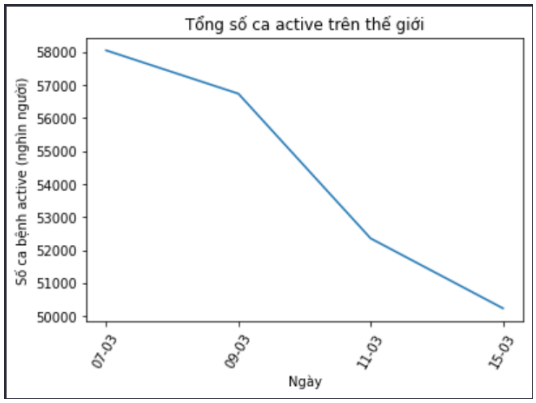
  Ví dụ hôm nay là ngày 16/03/2022 thì chúng ta sẽ thu thập dữ liệu của ngày 15/03/2022
- Kết quả dữ liệu thu thập được sẽ nằm trong thư mục Data

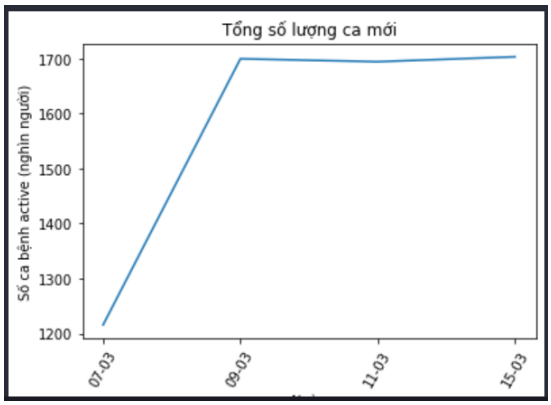
Sau khi thu thập dữ liệu bằng file PARSE_HTML, chúng ta sẽ dùng file VISUALIZE để visualize

III. TRỰC QUAN HOÁ

1. Xu hướng tăng giảm số ca trong cơn dịch

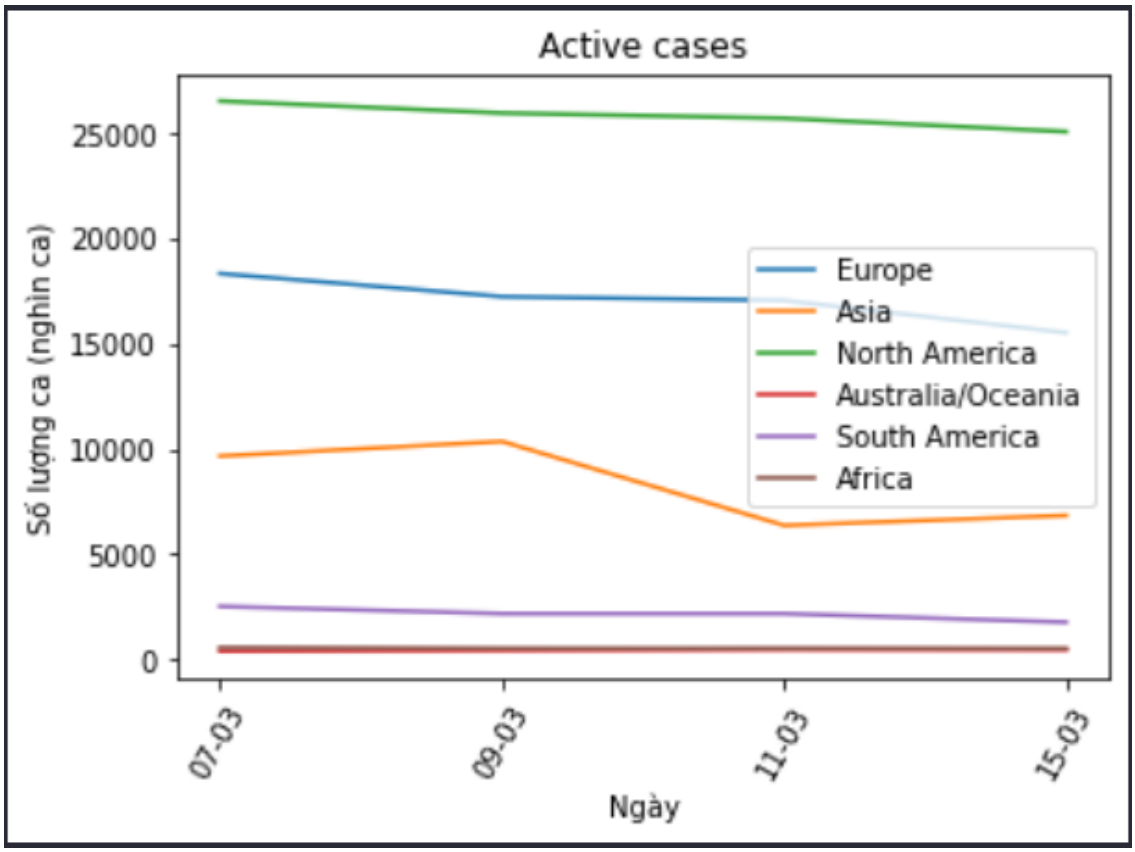
Đầu tiên, chúng ta sẽ xem xét xu hướng của số ca hiện tại đang tăng hay giảm, cũng như số lượng ca mới có xu hướng như thế nào





Nhìn chung, tổng số ca activate trên thế giới có xu hướng giảm xuống và số ca mới vẫn là một hang ngang. Dựa theo biểu đồ, có thể nói dịch bệnh đang có xu hướng biến mất, nhưng cần phải xem xét lại liệu dữ liệu trên trang web thu thập có còn cập nhật hay không, cũng như độ chính xác của dữ liệu trong tình hình hiện nay

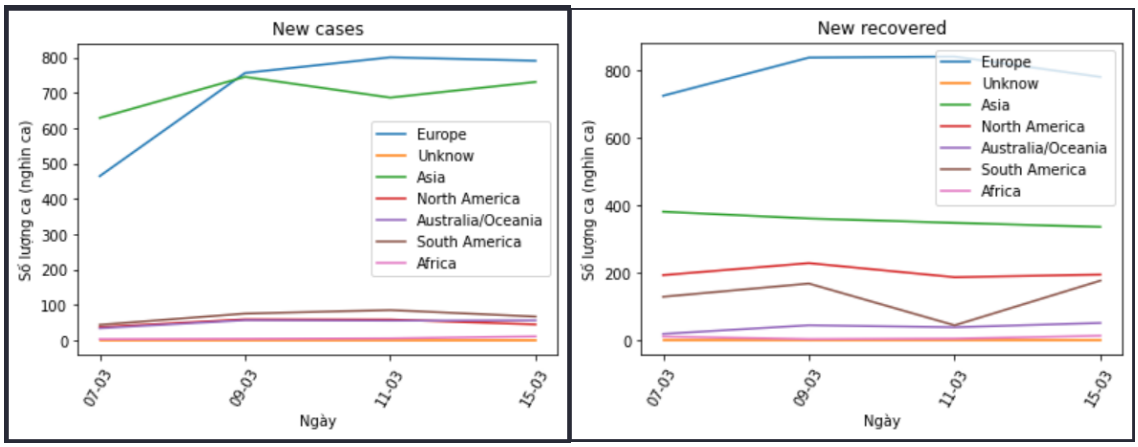
Tiếp theo, chúng ta sẽ xem xét sâu hơn vào từng châu lục



	Region	07-03	09-03	11-03	15-03
0	Europe	18,334	17,238	17,058	15,526
1	Asia	9,667	10,363	6,379	6,846
2	North America	26,520	25,951	25,701	25,067
3	Australia/Oceania	416	451	486	480
4	South America	2,532	2,191	2,190	1,782
5	Africa	580	544	540	536

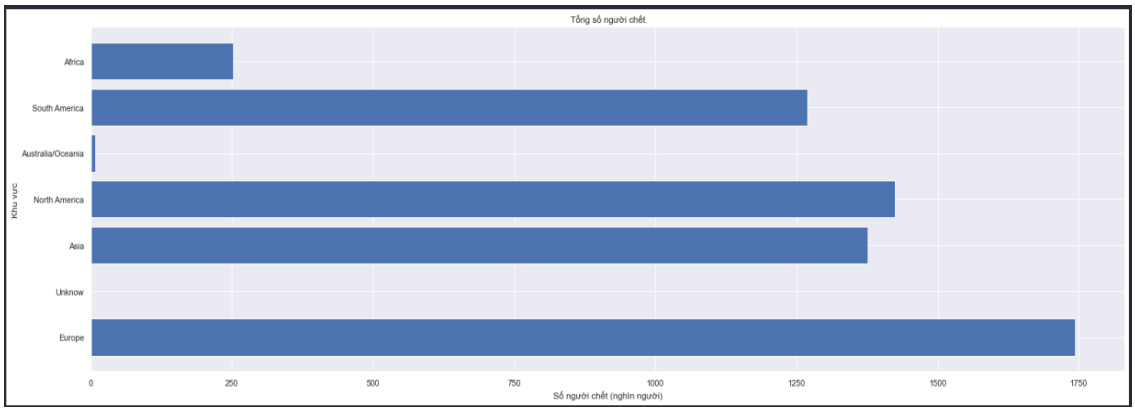
Nhìn chung, số lượng ca active hiện đang giảm dần ở các châu lục, tuy nhiên nhiên tốc độ giảm có vẻ không khả quan lắm, chúng ta có thể suy đoán rằng tốc độ giảm số ca active không chêch lệch lắm so với tốc độ lây nhiễm của bệnh dich.

Chúng ta tiếp tục trực quan để nhìn rõ hơn

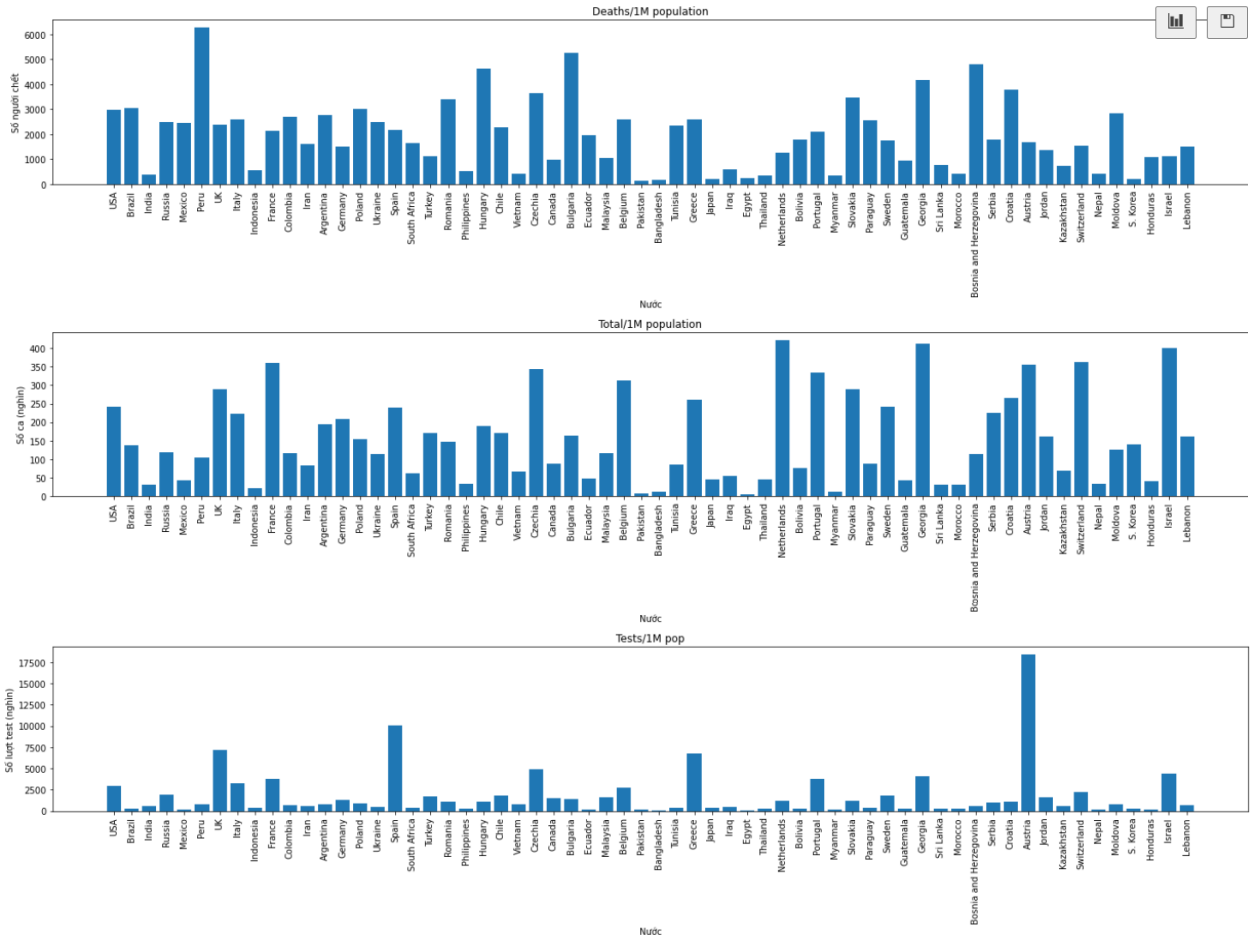


Thật vậy, số lượng ca mới và số lượng ca recovered không chênh lệch nhau mấy, cá biệc có Châu Á hiện đan có số lượng ca mới hơn hẳn số lượng ca recovered, không lạ khi ở biểu đồ Active cases giữa các châu lục thì Châu Á có xu hướng tăng rất nhẹ từ ngày 11-3 đến ngày 15-3.

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



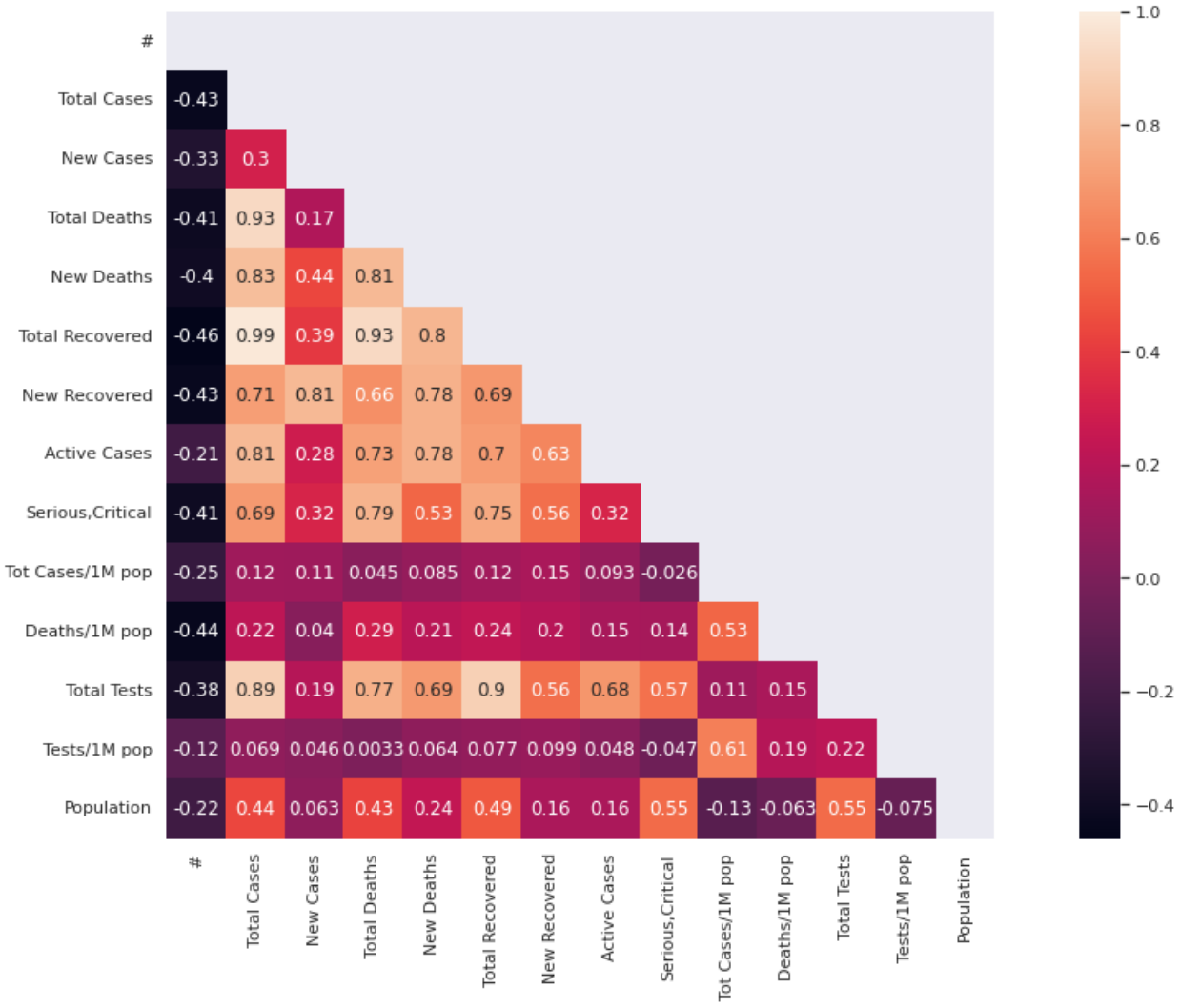
Số người chết hiện tại thì Europe vẫn là nhiều nhất và tách biệt rõ rệt so với phần còn lại, theo sau là North America, Asia, South America, cả ba cách biệt khá rõ, tuy nhiên vẫn khá sát nhau, chỉ có Africa có vẻ như có số lượng người chết vì dịch bệnh là thấp nhất, và cách biệt rất lớn so với phần còn lại của thế giới.



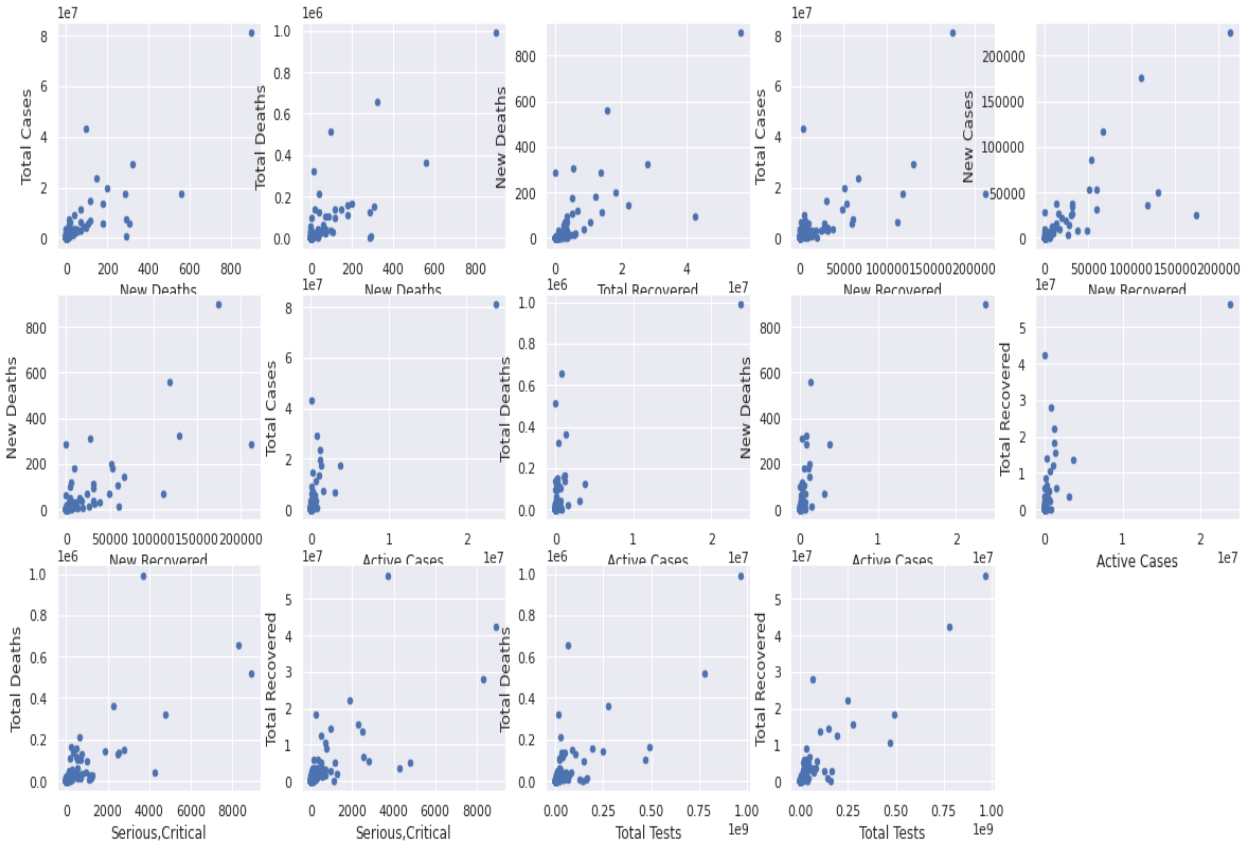
Chúng ta xem xét mối quan hệ giữa tỉ lệ test, số người chết/1M population,Số ca nhiễm/1M polulation. Quan sát kỹ thì đa số những đất nước có tỉ lệ test cao, thì dù cho tỉ lệ Activate Cases/1M Population cao thì tỉ lệ người chết ít hơn hẳn so với các nước có số lượng test ít.

3. Mối tương quan giữa các trường dữ liệu

Nhìn chung thì các trường dữ liệu không có dạng phân phối chuẩn mà có khá nhiều outliner và bị lệch về một phía, vậy nên chúng ta sẽ dùng hệ số tương quan Spearman để xem xét sự tương quan giữa các trường dữ liệu



Em sẽ dùng biểu đồ scatter để trực quan mối quan hệ giữa các biến, ở đây em sẽ chỉ chọn những cặp biến có độ tương quan từ 0.7-0.9 (có hệ số tương quan cao)



Quan sát các cặp biến trên, nhóm em có nhận xét:

+ Những cặp biến trên tương với nhau mạnh ở những giá trị nhỏ, giá trị các biến càng lớn thì mối quan hệ càng không rõ ràng.

+ Để giải thích cho vấn đề trên thì nhóm đưa ra giả thuyết: Ở những giá trị nhỏ thì ca nhiễm còn ít, trong tầm kiểm soát nên số liệu được đưa ra đầy đủ. Số ca nhiễm cao, quan điểm sống chung với bệnh, không đưa ra số liệu thống kê nữa thì làm số liệu sau này không đồng đều.

4. Mối quan hệ nhân quả

Chúng ta sẽ dùng mô hình OLS để xem xét một mối quan hệ phụ thuộc

a. Q(Serious, Critical) ~ Q(Population)?

OLS Regression Results

Dep. Variable:	Q('Serious,Critical')	R-squared:	0.298
Model:	OLS	Adj. R-squared:	0.295
Method:	Least Squares	F-statistic:	95.69
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	4.68e-19
Time:	20:56:27	Log-Likelihood:	-1850.9
No. Observations:	227	AIC:	3706.
Df Residuals:	225	BIC:	3713.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	146.5806	57.816	2.535	0.012	32.651	260.510
Q('Population')	3.966e-06	4.05e-07	9.782	0.000	3.17e-06	4.76e-06

Omnibus:	201.777	Durbin-Watson:	1.183
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14295.410
Skew:	2.864	Prob(JB):	0.00
Kurtosis:	41.453	Cond. No.	1.47e+08

Dựa vào kết quả thu được, thì có vẻ như Population chỉ trả lời được khoảng 30% cho số lượng ca Serious, Critical. Nhìn vào chỉ số trên nhóm em có một vài nhận xét:

+ Do trong cơ cấu dân số có nhiều độ tuổi khác nhau nên sức khoẻ trong từng độ tuổi cũng khác nhau.

+ Nhóm người trẻ, trung niên trở xuống có lẽ có sức khoẻ và sức trẻ tốt hơn nên ít bị chịu ảnh hưởng khi nhiễm bệnh.

+ Nhóm em nghĩ khi có có được thông tin về nhóm tuổi, nhóm người cao tuổi sẽ có nguy cơ bệnh nặng hơn và nhóm người lớn tuổi sẽ giải thích được số ca nặng (Serious, Critical) hơn.

b. Q(Serious, Critical) ~ Q(Total Cases)?

OLS Regression Results

Dep. Variable:	Q('Serious,Critical')	R-squared:	0.481
Model:	OLS	Adj. R-squared:	0.479
Method:	Least Squares	F-statistic:	208.9
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	6.19e-34
Time:	21:01:42	Log-Likelihood:	-1816.6
No. Observations:	227	AIC:	3637.
Df Residuals:	225	BIC:	3644.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	82.6718	50.183	1.647	0.101	-16.217	181.561
Q('Total Cases')	9.92e-05	6.86e-06	14.455	0.000	8.57e-05	0.000

Omnibus:	213.336	Durbin-Watson:	1.711
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9645.323
Skew:	3.347	Prob(JB):	0.00
Kurtosis:	34.224	Cond. No.	7.61e+06

Từ kết quả, chúng ta thấy Total case giải thích được ~48% Serious, Critical. Thông thường khi xét một mối quan hệ nhân quả thì tỷ lệ 48% là chưa đủ tin cậy để từ nhân suy ra quả. Nhóm em có vài nhận xét:

+ Do có nhiều tình trạng bệnh khác nhau nên việc tổng số ca nhiễm chỉ giải thích khoảng 48% số ca nặng là bình thường

+ Em nghĩ để giải thích số ca nặng thì cần biết về độ tuổi, tình trạng sức khoẻ hiện tại, có bệnh nền hay không, chế độ ăn uống, chế độ luyện tập thể thao…

c. Q(Serious, Critical) ~ Q(Total Tests)?

OLS Regression Results

Dep. Variable:	Q('Serious,Critical')	R-squared:	0.320
Model:	OLS	Adj. R-squared:	0.316
Method:	Least Squares	F-statistic:	105.6
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	1.45e-20
Time:	21:04:38	Log-Likelihood:	-1847.4
No. Observations:	227	AIC:	3699.
Df Residuals:	225	BIC:	3706.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	133.1493	57.149	2.330	0.021	20.534	245.765
Q('Total Tests')	5.755e-06	5.6e-07	10.278	0.000	4.65e-06	6.86e-06

Omnibus:	284.844	Durbin-Watson:	1.473
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18089.205
Skew:	5.357	Prob(JB):	0.00
Kurtosis:	45.399	Cond. No.	1.06e+08

Từ kết quả, chúng ta thấy Total test giải thích được ~32% Serious, Critical. Và thực đúng như thực tế chúng ta thấy được việc thực hiện test chỉ cho biết một người nhiễm hoặc không nhiễm chứ không thể đưa ra kết luận tình trạng bệnh của người đó là nhẹ, ít triệu chứng hay nặng, nguy kịch cần can thiệp y tế (thở máy, dung thuốc mạnh, …).

Việc lần hiện test quá nhiều lần làm tăng số lượng test rất nhiều, trường hợp nặng ít.

d. Q(Total Recovered) ~ Q(Total Tests)?

OLS Regression Results

Dep. Variable:	Q('Total Recovered')	R-squared:	0.805
Model:	OLS	Adj. R-squared:	0.804
Method:	Least Squares	F-statistic:	877.7
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	1.75e-77
Time:	21:25:02	Log-Likelihood:	-3479.2
No. Observations:	215	AIC:	6962.
Df Residuals:	213	BIC:	6969.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.85e+05	1.83e+05	2.102	0.037	2.4e+04	7.46e+05
Q('Total Tests')	0.0518	0.002	29.625	0.000	0.048	0.055

Omnibus:	208.556	Durbin-Watson:	1.808
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15347.906
Skew:	3.260	Prob(JB):	0.00
Kurtosis:	43.875	Cond. No.	1.09e+08

Từ kết quả, chúng ta thấy Total test giải thích được ~80% Total Recovered. Prob (F-statistic) ở mô hình này rất nhỏ < 0.05 rất nhiều nên có vẻ có ý nghĩa.

Total test sẽ chỉ ra những trường hợp dương tính với virus và những trường hợp đó sẽ có phần hồi phục trở lại. Qua mô hình chỉ ra được total test giải thích tới 80% total recovered, mức giải thích khá cao.

Điều này cho thấy tỷ lệ người hồi phục lại khá cao, nhưng vẫn có những trường hợp nặng nên vẫn cần hạn chế để nhiễm bệnh càng tốt.

e. Q(New Cases) ~ Q(Total Tests)?

OLS Regression Results

Dep. Variable:	Q('New Cases')	R-squared:	0.037
Model:	OLS	Adj. R-squared:	0.033
Method:	Least Squares	F-statistic:	8.718
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	0.00348
Time:	21:25:32	Log-Likelihood:	-2676.9
No. Observations:	227	AIC:	5358.
Df Residuals:	225	BIC:	5365.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5826.5028	2207.692	2.639	0.009	1476.106	1.02e+04
Q('Total Tests')	6.386e-05	2.16e-05	2.953	0.003	2.12e-05	0.000

Omnibus:	366.832	Durbin-Watson:	1.749
Prob(Omnibus):	0.000	Jarque-Bera (JB):	57564.540
Skew:	8.087	Prob(JB):	0.00
Kurtosis:	79.318	Cond. No.	1.06e+08

Từ kết quả, chúng ta thấy Total test chỉ giải thích được ~3% New Cases. Em khá bất ngờ trước kết quả này, theo suy đoán của em thì tỷ lệ số lượng test sẽ trả lời được cho số ca mắc mới sẽ cao nhưng thực tế qua mô hình trên thì rất thấp chỉ 3%

Về vấn đề này thì em nghĩ là do trong thời điểm lấy mẫu (tháng 3 năm 2022) thì một số nước đã không còn công bố số ca mắc mới nữa do xác định sống chung với dịch nên newcase nhỏ .

f. Q(Total Deaths) ~ Q(Total Cases)?

OLS Regression Results

Dep. Variable:	Q('Total Deaths')	R-squared:	0.858
Model:	OLS	Adj. R-squared:	0.858
Method:	Least Squares	F-statistic:	1364.
Date:	Thu, 17 Mar 2022	Prob (F-statistic):	1.80e-97
Time:	20:56:41	Log-Likelihood:	-2702.1
No. Observations:	227	AIC:	5408.
Df Residuals:	225	BIC:	5415.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1263.4753	2482.361	0.509	0.611	-3628.175	6155.125
Q('Total Cases')	0.0125	0.000	36.931	0.000	0.012	0.013

Omnibus:	230.170	Durbin-Watson:	2.423
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9677.180
Skew:	3.844	Prob(JB):	0.00
Kurtosis:	34.049	Cond. No.	7.61e+06

Từ kết quả, chúng ta thấy Total Cases giải thích được ~85% Total Deaths.

Prob (F-statistic) cũng rất thấp < 0.05 nên đây là mô hình có ý nghĩa.

Khi nhiễm virus là có khả năng tử vong do virus này, nên hãy cẩn thận, hạn chế bị nhiễm để đảm bảo sức khoẻ.

Dựa trên các kết quả thu được từ a-f, chúng ta nhận thấy rằng 2 mô hình:

- Q(Total Death) ~ Q(Total Cases): Giả thích 85.8% dữ liệu
- Q(Total Recovered) ~ Q(Total Test): giải thích 80.3% dữ liệu

  ⇨ Đây là 2 mô hình tốt nhất hiện giờ