**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated

**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 |

1. **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 |

1. **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

Graphical user interface, application

Description automatically generated

* 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

1. **TRỰC QUAN HOÁ**
2. **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

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated  
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

Chart, line chart

Description automatically generated

A black screen with white text

Description automatically generated with low confidence

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

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

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.

1. **Graphical user interface, application, Teams

   Description automatically generatedSo 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.

Chart, histogram

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidence

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.

1. **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

Chart

Description automatically generated

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)

Calendar

Description automatically generated with low confidence

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.

1. **Mối quan hệ nhân quả**

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

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

Graphical user interface, text

Description automatically generated

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.

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

Graphical user interface

Description automatically generated

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…

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

Graphical user interface, text

Description automatically generated

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.

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

Graphical user interface, text

Description automatically generatedTừ 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.

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

Graphical user interface, text

Description automatically generated

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 .

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

Graphical user interface, text

Description automatically generated

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