

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**



**Logo

Description automatically generated**

**MÔN: TRỰC QUAN HÓA DỮ LIỆU**

**BÁO CÁO LAB 2:**

**SỬ DỤNG TABLEAU ĐỂ TRỰC QUAN HÓA DỮ LIỆU**

***Thứ bảy, ngày 28 tháng 5 năm 2022***

**

1. **Thông tin nhóm:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Họ và tên** | **MSSV** | **Công việc** | **Mức độ hoàn thành** |
| Hồ Hữu Ngọc | 19120602 | * Vẽ biểu đồ trực quan sử dụng tableau. * Viết báo cáo các phần trên | **100%** |
| Nguyễn Phát Minh | 19120586 | * Trực quan hóa sử dụng tableau biểu đồ. * Hỗ trợ viết báo cáo cho những phần trên. | **100%** |
| Võ Trọng Phú | 19120328 | * Tìm hiểu về Tableau * Chạy mô hình máy học * Viết báo cáo các phần trên | **100%** |

1. **Mức độ hoàn thành tổng thể mỗi yêu cầu:**

|  |  |
| --- | --- |
| **Công việc** | **Mức độ hoàn thành** |
| Tìm hiểu về Tableau | **100%** |
| Áp dụng Tableau để trực quan dữ liệu:   * Trực quan nhiều dạng biểu đồ khác nhau, dùng màu sắc để thể hiện dữ liệu * Trực quan dữ liệu biến đổi qua từng ngày. * Sử dụng các kỹ thuật được giới thiệu trong bài Manipulate View, Facet, Reduce, Embed để trình diễn trên Tableau với dữ liệu Woldometer | **100%** |
| Chạy các thuật toán máy học đơn giản để hiểu thêm về dữ liệu | **100%** |

1. **Tìm hiểu về Tableau:**
2. ***Giới thiệu về Tableau:***

* Tableau là một phần mềm được sử dụng để hỗ trợ phân tích và trực quan hóa dữ liệu, giúp người dùng tạo ra các biểu đồ trực quan dựa trên nguồn dữ liệu.
* Được sử dụng nhiều trong ngành Business Intelligence (BI).
* Tableau có nhiều ưu điểm so với các ứng dụng khác, người sử dụng có thể dễ dàng chuyển đổi những dữ liệu dạng thô sang những hình ảnh trực quan bắt mắt và dễ hình dung, ngoài ra nó có thể hoạt đông với bất kì loại dữ liệu nào (có cấu trúc, phi cấu trúc, bán cấu trúc).
* Người dùng Tableau không cần nền tảng kĩ thuật để sử dụng công cụ này.

1. ***Các tính năng hỗ trợ của Tableau và ví dụ:***
2. ***Khả năng kết nối với dữ liệu:***
   * Tableau có thể kết nối với nhiều loại dữ liệu, môi trường dữ liệu khác nhau như: File (Excel, Text file, JSON file, PDF file,…), Server (SQL Server, Oracle, MySQL,..) hoặc một Data Sources đã được lưu trước đó.
   * Ví dụ: ta thử kết nối Tableau đến một file Excel đã lưu trong máy.

A screenshot of a computer

Description automatically generated

1. ***Tính năng kéo thả:***

* Giao diện điều khiển của Tableau giúp người dùng có thể dễ dàng thực hiện thao tác kéo thả các trường, các mục điều khiển theo ý mình.
* Ví dụ: kéo thả trường Order vào vùng hoạt động.

Graphical user interface

Description automatically generated

1. ***Tính năng mô phỏng dữ liệu (Data Visualization):***

* Khi dữ liệu đã được kết nối, bằng cách kéo thả mục cần trực quan ta có thể tùy chọn trực quan hóa dữ liệu dưới nhiều tùy chọn biểu đồ khác nhau (góc bên phải màn hình)
* Ví dụ: biểu đồ bar chart trực quan Sales theo từng quốc gia.

A screenshot of a computer

Description automatically generated

1. ***Tính năng xây dựng bản đồ:***

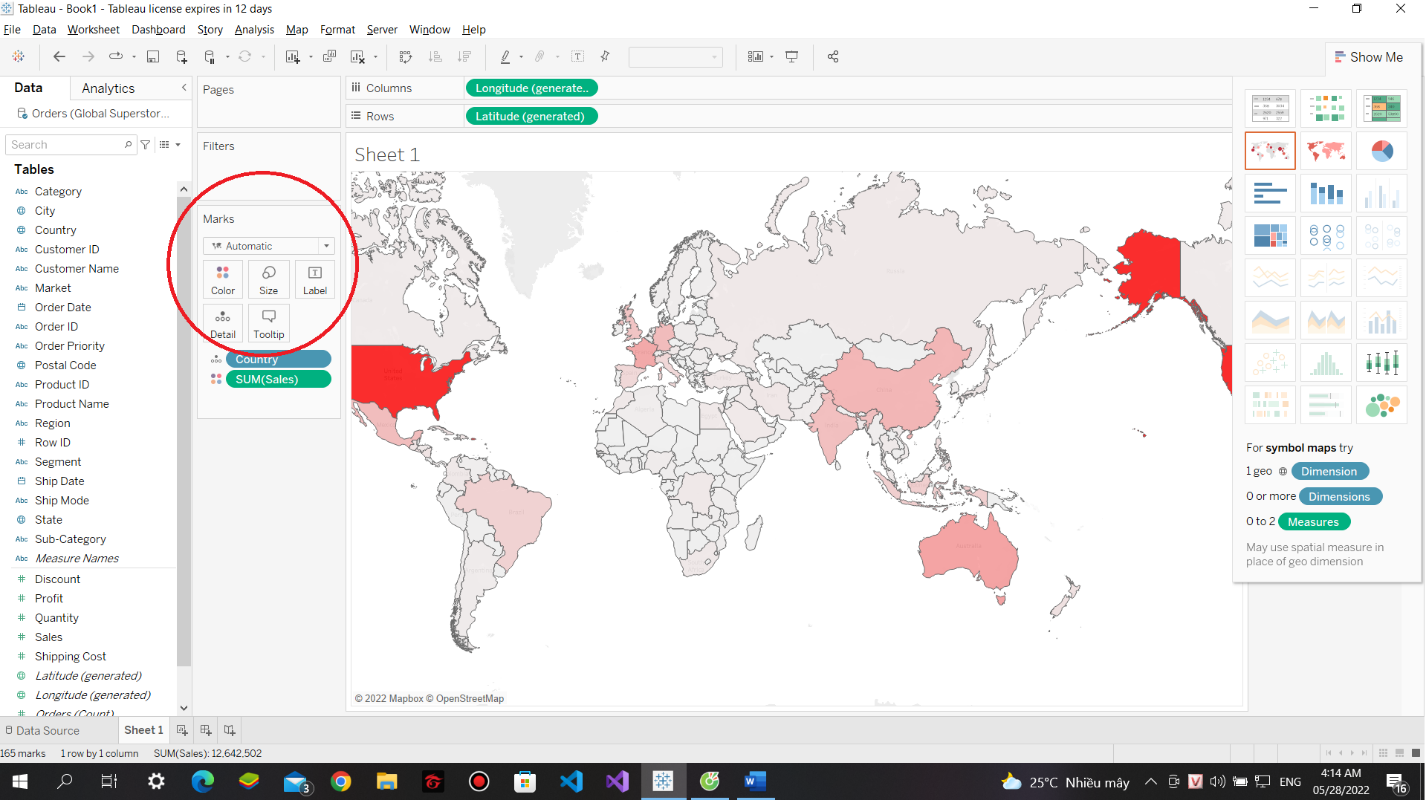
* Ta có thể xây dựng một hình thức trực quan mới, đó là vẽ bản đồ địa lí cho dữ liệu, vẫn là những lựa chọn ở góc bên phải màn hình.
* Ví dụ: trực quan Sales theo từng quốc gia bằng bản đồ.

Map

Description automatically generated

1. ***Tính năng tùy chỉnh các biểu đồ:***

* Ta có thể tùy chỉnh kích thước, màu sắc, label,...của biểu đồ ở bảng điều khiển trong hình.
* Ví dụ: thay đổi màu sắc của bản đồ trực quan ở trên sang màu đỏ.



1. ***Các tính năng khác:***

* Các tính năng hỗ trợ khác như chia sẻ, in ra file PDF, export sang powerpoint,… cũng được hỗ trợ trong Tableu ở mục tùy chọn File.
* Ví dụ: chọn File -> Share để dùng tính năng chia sẻ.

Graphical user interface, application, map

Description automatically generated

1. **Áp dụng Tableau để trực quan dữ liệu (một ngày và nhiều ngày):**
2. **Biểu đồ thể hiện tổng số ca mắc covid hiện nay trên thế giới**

Map

Description automatically generated

* Biểu đồ trên cho thấy một cái nhìn bao quát về tình trạng nhiễm Covid của các quốc gia trên thế giới và thấy rõ được một số nước có tình trạng nhiễm cao(USA,Brazil,…), và các nước vừa và thấp.

1. **Top 10 nước có số lượng ca nhiễm mới trong ngày.**

Chart

Description automatically generated

* Data được chọn là ngày 21/04/2022 sử dụng hai trường là(Country và New Case) lấy ra top 50 nước có số lượng ca nhiễm mới mỗi ngày và vẽ bằng barchar. Nhìn vào biểu đồ có thể thấy số lượng ca nhiễm vẫn tăng hàng ngày. Nhưng số ca mỗi ngày tương đối ít. Điển hình là nước Bhutan cũng không quá 1000 ca/ngày

1. **Tỉ lệ số ca chết trong ngày của top 5 nước có số người chết nhiều nhất trong ngày.**

Pie chart

Description automatically generated with medium confidence

* Dữ liệu của ngày 21/04/2022 sử dụng hai trường là (Country và New Death) cho thấy 5 nước có số lượng người chết trong ngày nhiều nhất do Covid 19. Trong ngày này thì Finland có số người chết nhiều nhất, tỷ lệ hơn hẳn các nước còn lại.

1. **Số ca nhiễm của mỗi nước được thể hiện trong ngày**

Chart, treemap chart

Description automatically generated

* Data được lấy trong ngày 28/04/2022 và thể hiện số lượng ca nhiễm của các nước có số lượng ca nhiễm nhiều nhất.

1. **Số ca nhiễm mới của các nước thể hiện qua các ngày.**

Chart, line chart

Description automatically generated

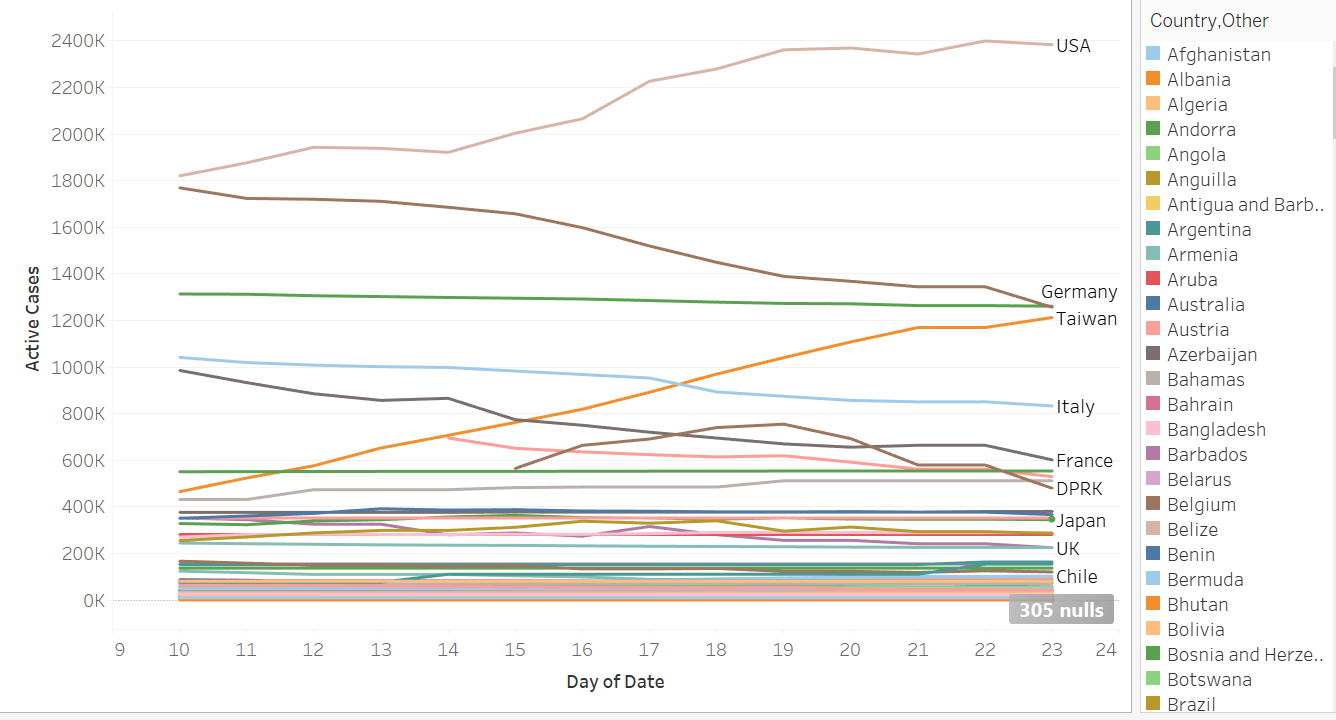
* Data được lấy từ ngày 10/05 đến ngày 23/05. Qua biểu đồ có thể thấy tình trạng các ca nhiễm của thế giới vẫn diễn ra rất bất thường chưa có sự ổn định và số ca nhiễm vẫn đang còn cao.

1. **Tương qua giữa số lượng ca chết và số lượng ca nhiễm mới và số lượng ca chết thể hiện qua các ngày.**

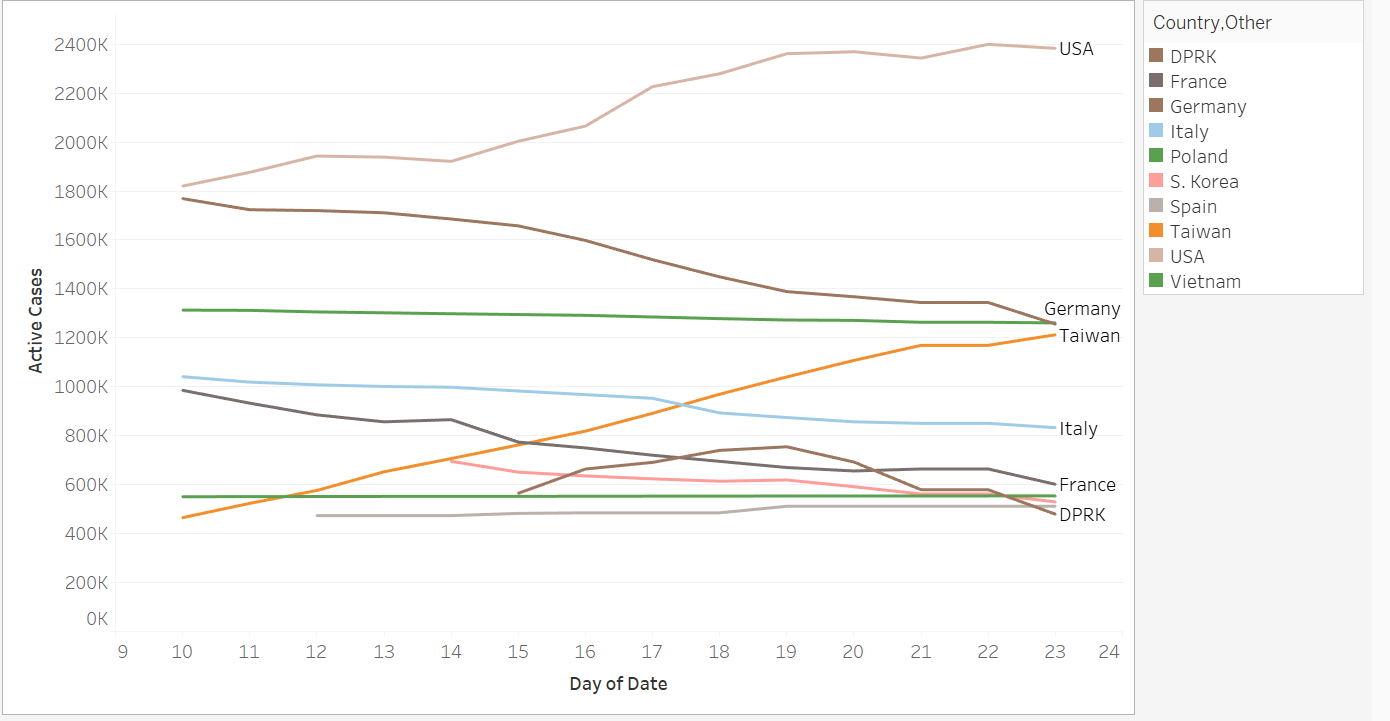
Chart, histogram

Description automatically generated

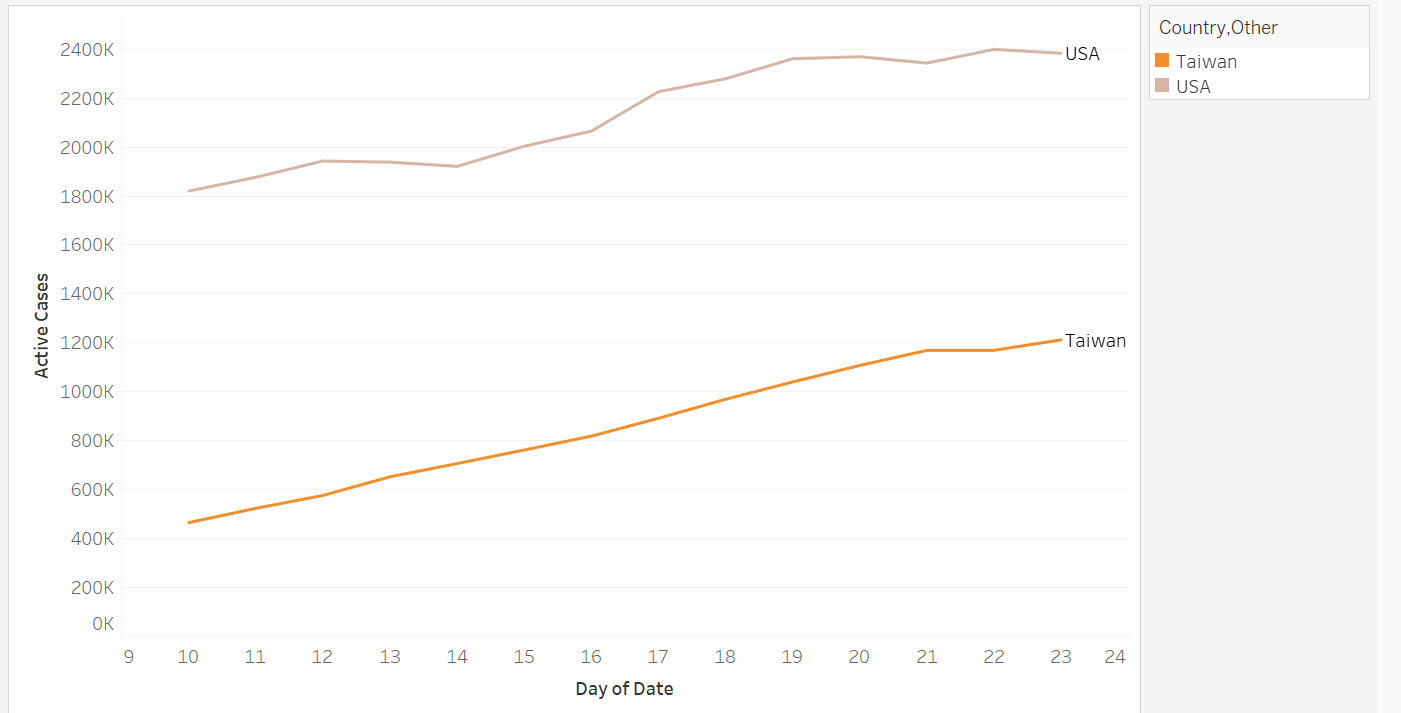
1. **Biểu đồ thể hiện tình trạng nhiễm covid giữa các nước.**



* Từ biểu đồ này ta thấy từ khoảng 400k trở xuống bị nhiễu bởi nhiều giá trị nên ta sẽ không xét đến khoảng dưới này bằng cách xóa các khoảng này đi.



* Sau khi loại bỏ ta có thể quan sát được các nước trên là những nước có tình trạng nhiễm covid cao nhất



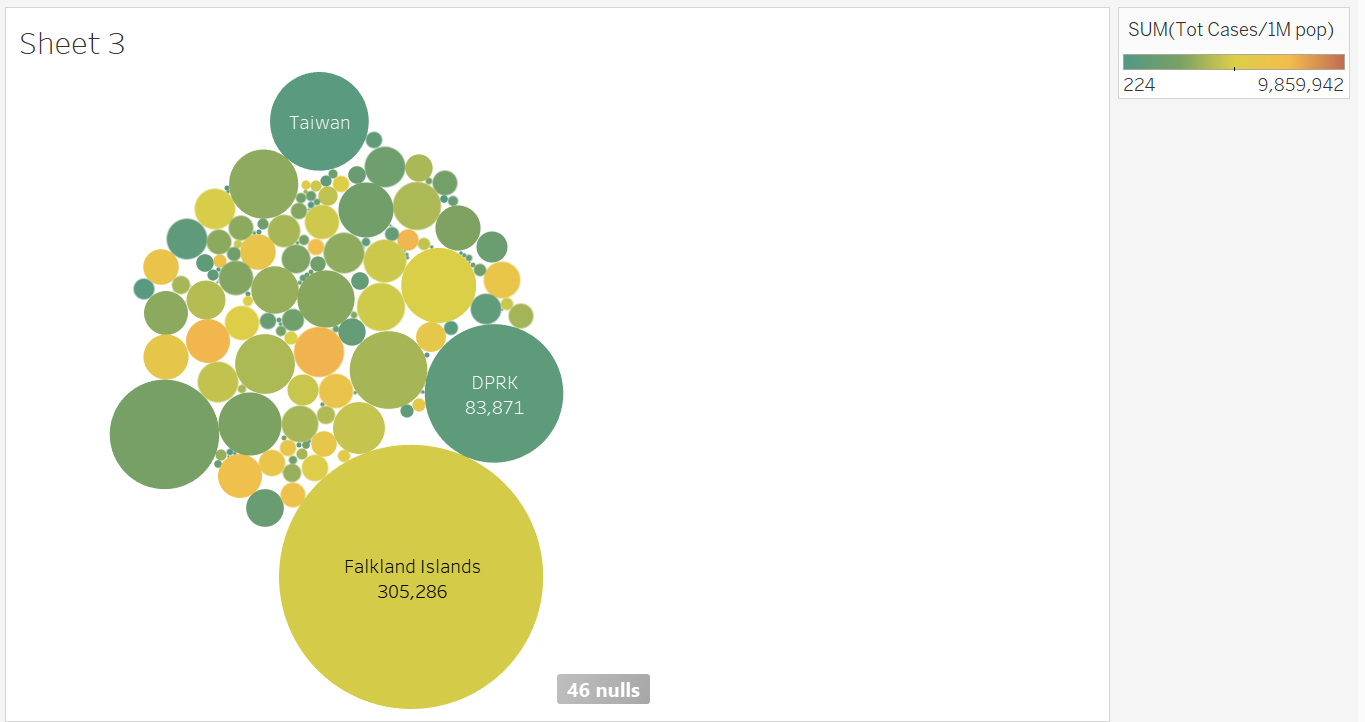
* Nhưng chỉ có Taiwan và USA là tiếp tục tăng, các nước còn lại không tăng hoặc giảm đều xuống, từ đây ta có thể đoán được tình trạng nhiễm covid trên thế giới đang giảm hoặc cập nhật không đầy đủ.

1. **Số ca nhiễm của các nước thể hiện qua từng ngày.**



* Biểu đồ thể hiện số lượng ca nhiễm hiện có của các nước trong các ngày 10-23/05/2022. Nhìn vào biểu đồ thì thấy các ca nhiễm đang giảm hoặc đứng yên qua các ngày. Mong là thế giới sẽ hoàn toàn chấm dứt đại dịch covid-19.

1. **Sự tương quan giữa tổng số lượng nhiễm/1 triệu người và số lượng ca mắc mới/ 1 triệu người**



* Màu sắc của biểu đổ thể hiện tổng số lượng nhiễm, độ to hình tròn thể hiện số lượng ca mắc mới.
* Ta thấy ngoài trừ Falkland Islands ra thì các nước có tổng số lượng ca nhiễm cao có số lượng ca mới thấp hơn so với các nước trung bình. Điều này cho thấy các nước có nhiều ca nhiễm đối phó với covid quả hiệu quả và còn một số còn nhỉnh hơn với các nước có tổng ca nhiễm thấp.

**V. Chạy một số mô hình học máy đơn giản để hiểu thêm về dữ liệu:**

1. ***Thuật toán giảm chiều Principal Component Analysis (PCA):***
2. Thuật toán PCA dùng để chuyển n chiều sang dữ liệu m chiều (m < n) mà vẫn giữ được nhiều thông tin nhất có thể.
3. Các bước thực hiện như sau:

* Load dữ liệu, lấy các dữ liệu số và bỏ đi cột ‘Country, Other’
* Chuẩn hóa dữ liệu (dùng hàm fit\_transform)
* Dùng PCA giảm chiều dữ liệu xuống 2 components, lúc này ta sẽ được kết quả giảm chiều từ 13 cột còn 2 cột:

Text

Description automatically generated

* Thêm lại cột ‘Country, Other’ vào data mới.

Graphical user interface, text, application, email

Description automatically generated

* Tiến hành trực quan dữ liệu cuối cùng bằng Scatter plot, để thấy rõ được tương quan giữa các nước, ta sẽ bỏ đi dòng World:

Chart, scatter chart

Description automatically generated

Chỉ có một chấm tròn duy nhất ở phía góc trên phải có giá trị xấp xỉ (5,8), đa số chấm tròn các nước đều tập trung ở góc trái và có Principal Components 1 âm.

1. ***Thuật toán bình phương nhỏ nhất Ordinary Least Squares (OLS):***
2. Phương pháp OLS dùng để phục vụ cho việc ước lượng tham số trong phương trình hồi quy. Mục đích của phương pháp này là tối thiểu hoá tổng bình phương của các khoảng cách theo phương thẳng đứng giữa số liệu thu thập được và đường hồi quy.
3. Trong mô hình này ta sẽ thử ước lượng mối quan hệ nhân quả giữa ‘Total Recovered’ theo ‘Total Cases’ ở lab1.
4. Các bước thực hiện như sau:

* Load dữ liệu, lấy các dữ liệu số và bỏ đi cột ‘Country, Other’, lấy 2 data dự đoán X và Y
* Dùng OLS và chuẩn hóa dữ liệu (dùng hàm fit)
* Tiến hành dự đoán ‘Total Recovered’ theo ‘Total Cases’, ta thu được kết quả:

Table

Description automatically generated with medium confidence

* Giải thích ý nghĩa một vài thông số:
* R-squared: hệ số xác định. Tỷ lệ phương sai trong biến phụ thuộc có thể dự đoán / giải thích được
* Adj. R-squared : Adjusted R-squared (R bình phương điều chỉnh) nằm trong khoảng từ 0 đến 1 và thể hiện mức độ phù hợp của mô hình, thông số này càng cao thì mô hình càng được cải thiện.
* F-statistic: thống kê F là tỷ lệ giữa sai số bình phương trung bình của mô hình với sai số bình phương trung bình của các phần dư. Nó quyết định ý nghĩa tổng thể của mô hình.
* coef: hệ số của các biến độc lập và số hạng không đổi trong phương trình, trong đó:
* const: nếu hệ số của ‘Total Cases’ bằng 0 thì của Y tức là ‘Total Recovered’ sẽ là const.
* Total Cases: tỉ lệ thay đổi của Y (Total recovered) so với Total Cases.
* t: giá trị của t-statistic, là tỷ số khác biệt giữa giá trị ước tính và giả thuyết của một tham số với sai số chuẩn
* Cuối cùng, ta thử dùng scatter plot để trực quan theo hàm

Y = 0.9446 \*data\_df\_OLS['Total Cases'] + -1.138e+05 rút ra được ở model trên:

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Kết quả thu được giống hệt scatter plot đã trực quan ở Lab 1 (hình thứ 2), vậy là model này của chúng ta đã chính xác, tức là ‘Total Recovered’ và ‘Total Cases’ có mối quan hệ nhân quả với nhau theo hàm số Y = 0.9446 \*data\_df\_OLS['Total Cases'] + -1.138e+05.