

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – TP HỒ CHÍ MINH  
KHOA CÔNG NGHỆ THÔNG TIN

**BÁO CÁO**

LAB 2- TRỰC QUAN HÓA DỮ LIỆU VỚI TABLEAU



**Giáo viên hướng dẫn**    Lê Ngọc Thành

**Sinh viên thực hiện**    Lê Tiến Trí

Hoàng Minh Đức

Lê Hoàng Thịnh Phước

*Ngày 16, Tháng 4, Năm 2022*

# MỤC LỤC

<b>I. Thông tin thành viên và phân công việc</b>	<b>3</b>
1. Thông tin thành viên.....	3
2. Bảng phân công công việc.....	3
<b>II. Tìm hiểu về Tableau</b>	<b>3</b>
<b>III. Sử dụng Tableau để trực quan dữ liệu về Covid-19</b>	<b>9</b>
1. Dữ liệu chuẩn bị cho trực quan trên Tableau.....	9
2. Trực quan dữ liệu .....	9
<b>IV. Thuật toán Học máy</b> .....	<b>18</b>
1. Bộ dữ liệu cho quá trình training .....	18
2. Thuật toán Support vector machine .....	18
<b>V. Tài liệu tham khảo</b> .....	<b>20</b>

## I. Thông tin thành viên và phân công việc

### 1. Thông tin thành viên

Họ và tên	MSSV	Email
Lê Tiến Trí	19127593	lttri19@clc.fitus.edu.vn
Lê Hoàng Thịnh Phước	19127518	lhtphuoc19@clc.fitus.edu.vn
Hoàng Minh Đức	19127121	19127121@student.hcmus.edu.vn

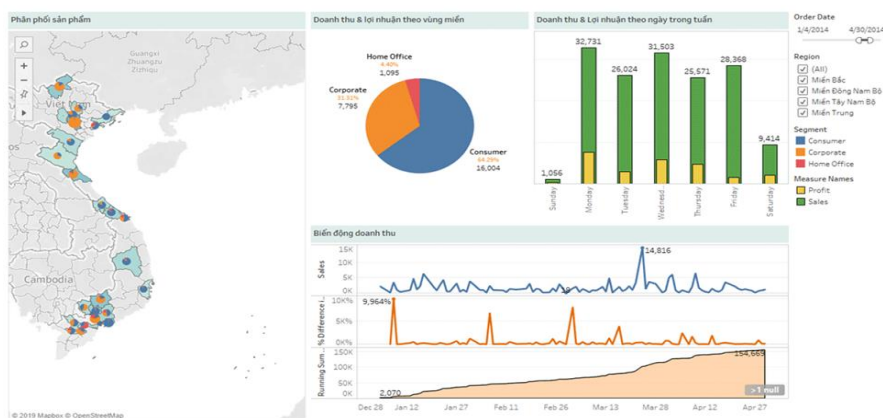
### 2. Bảng phân công công việc

Thành viên	Công việc	(%) hoàn thành
Lê Tiến Trí	<ul style="list-style-type: none"><li>- Trực quan hóa dữ liệu bằng tableau</li><li>- Tiến hành code và chạy mô hình học máy (SVM)</li></ul>	100%
Lê Hoàng Thịnh Phước	<ul style="list-style-type: none"><li>- Tìm hiểu công cụ Tableau</li><li>- Trực quan hóa dữ liệu</li></ul>	100%
Hoàng Minh Đức	<ul style="list-style-type: none"><li>- Tìm hiểu công cụ Tableau</li><li>- Trực quan hóa dữ liệu</li></ul>	100%

## II. Tìm hiểu về Tableau

### 1. Tableau là gì?

- Tableau là phần mềm hỗ trợ phân tích và trực quan hóa dữ liệu (Data Visualization), được dùng nhiều trong ngành BI (Business Intelligence). Cũng giống như Excel, Tableau giúp tổng hợp các dữ liệu nhưng ở một cấp độ cao hơn khi chuyển những liệu này từ các dãy số thành những hình ảnh, biểu đồ trực quan.



Phân tích dữ liệu dưới định dạng hình ảnh cho hiệu quả cao hơn

## 2. Tầm quan trọng của Tableau

- Nhu cầu phân tích dữ liệu ngày nay đã trở nên vô cùng cấp thiết, các dữ liệu không chỉ đơn thuần là tập hợp những con số. Thay vào đó, việc trực quan những dữ liệu giúp người xem so sánh, tổng kết, đánh giá và đưa ra những quyết định chính xác.
- Công việc này càng trở nên quan trọng hơn đối với các doanh nghiệp quy mô. Khi những báo cáo ngày một nhiều mất quá nhiều thời gian để đánh giá thông qua những báo cáo truyền thống. Đó cũng là lúc những công cụ phân tích và trực quan dữ liệu Tableau được sử dụng

AutoSave ☐ Off

File

Home

Insert

Draw

Page Layout

Formulas

Data

Review

View

Developer

Help

PDFelement

PivotTable

Recommended PivotTables

Table

Pictures

Online Pictures

Shapes

Icons

3D Models

SmartArt

Screenshot

Get Add-ins

My Add-ins

Recommended Charts

Charts

Auto

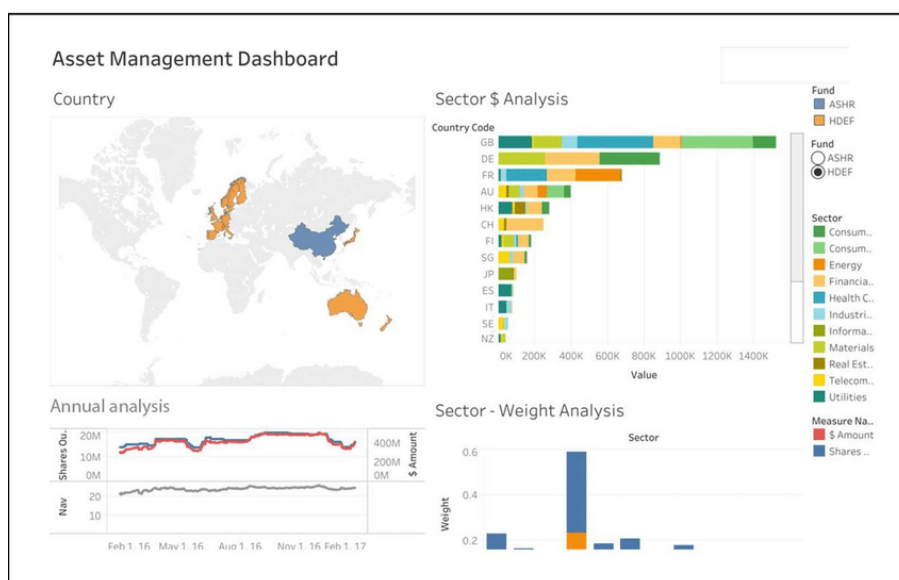
A1

</

Báo cáo dưới dạng số liệu bằng Excel truyền thống

## 3. Tính năng của Tableau

### a. Tableau Dashboard



- Tableau Dashboard cung cấp một cái nhìn đầy đủ về dữ liệu với đa dạng các định dạng và cách sắp xếp. Nó cũng hỗ trợ các bộ lọc và khả năng sao chép các thành phần bảng biểu cho các mục đích sử dụng khác.

## **b. Cộng tác và chia sẻ**

- Tableau cho phép người dùng cộng tác và chia sẻ luồng công việc với nhau trong thời gian thực một cách an toàn, giúp tăng hiệu quả công việc khi làm việc theo nhóm.

## **c. Dữ liệu trực tiếp và In-memory**

- Tableau có khả năng kết nối và sử dụng các nguồn dữ liệu thời gian thực, hoặc lưu trữ thông tin từ thiết bị ngoại vi vào bộ nhớ máy tính để xử lý.

## **d. Kết nối đến nguồn cấp dữ liệu**

- Tableau hỗ trợ nhiều nguồn cấp dữ liệu khác nhau như tập tin, cơ sở dữ liệu quan hệ/phi quan hệ, dữ liệu trên đám mây,...



## **e. Trực quan hóa nâng cao dưới dạng biểu đồ**

- Đây là một trong các tính năng chủ chốt của công cụ, hỗ trợ đa dạng các loại biểu đồ như biểu đồ cột, biểu đồ tròn, gantt chart, cây,... việc chuyển đổi giữa các dạng biểu đồ cũng đơn giản chỉ bằng một cú nhấp chuột.

## **f. Trình diễn thông tin trên bản đồ**

- Tableau được cài sẵn nhiều dạng thông tin như tên các địa danh, mã bưu chính,... hỗ trợ rất tốt cho việc thể hiện thông tin chi tiết và chính xác trên bản đồ. Các dạng bản đồ hỗ trợ cũng đa dạng như bản đồ nhiệt, bản đồ mật độ điểm, bản đồ luồng...

## **g. Xem trên thiết bị di động**

- Các thiết bị di động ngày càng có chỗ đứng quan trọng trong cuộc sống hàng ngày và là những thiết bị được sử dụng thường xuyên nhất. Do đó Tableau hỗ trợ cả phiên bản ứng dụng di động tương thích cho từng hệ điều hành giúp người sử dụng có được trải nghiệm linh hoạt, tự do hơn.

## **h. Truy vấn dữ liệu**

- Người dùng có thể truy vấn dữ liệu từ Tableau chỉ bằng ngôn ngữ tự nhiên, công cụ sẽ trả về thông tin cả dạng thô và dạng trực quan hóa.

### i. Phân tích, dự đoán

- Không chỉ thể hiện các dữ liệu sẵn có mà Tableau cũng giúp đưa ra các dự đoán xu hướng dữ liệu dựa trên thuật toán, tạo tiền đề cho việc đưa ra quyết định của con người.

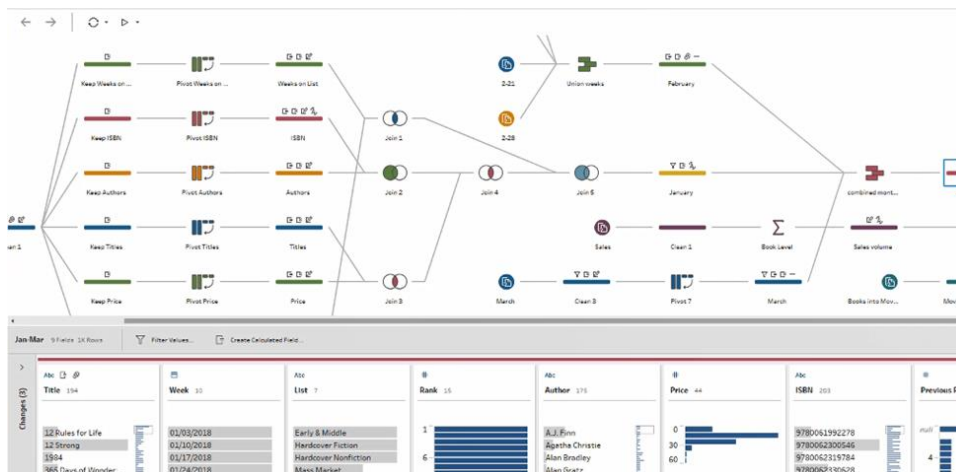
## 4. Tableau có những phiên bản nào



*Các sản phẩm của Tableau*

Danh sách các dịch vụ sau là những phần rất quan trọng mà bạn cần nắm. Đây là những công cụ được sử dụng trong suốt quá trình xây dựng báo cáo và trực quan dữ liệu.

### a. Tableau Prep



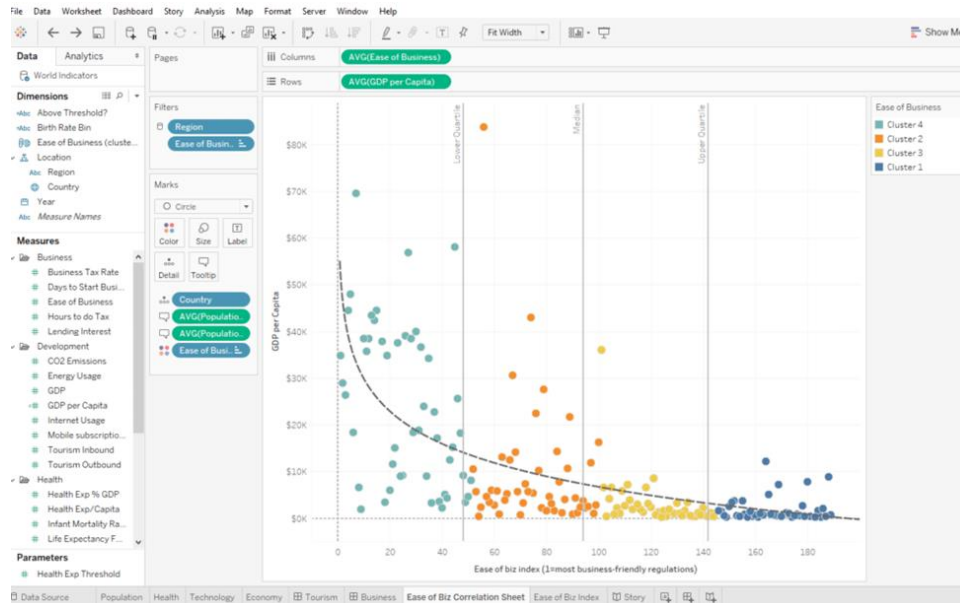
*Tableau Prep là nơi để chuẩn bị dữ liệu*

Đúng với tên gọi của mình, Tableau Prep có thể được hiểu là công cụ được dùng để chuẩn bị dữ liệu. Tableau Prep mang đến sự thay đổi quan trọng trong việc tổ chức dữ liệu, so với phương pháp truyền thống có nhiều cải tiến.

Cụ thể, ứng dụng giúp người dùng doanh nghiệp và nhà phân tích định hình dữ liệu nhanh chóng. Cho phép thực hiện các truy vấn, kết hợp và làm sạch dữ liệu cực kỳ đơn giản và tiện lợi.

Sử dụng Tableau Prep giúp dữ liệu có tổ chức, rõ ràng, dễ quản lý hơn. Hiện có hai công cụ là Tableau Prep Builder để xây dựng luồng dữ liệu và Tableau Prep Conductor để quản lý các luồng.

## b. Tableau Desktop



*Tableau Desktop là nơi thực hiện các phân tích và trực quan dữ liệu*

Sau khi đã hoàn tất bước chuẩn bị, công cụ tiếp theo sẽ giúp bạn phân tích các dữ liệu, Tableau Desktop. Cung cấp giao diện trực quan cùng các tính năng đa dạng để mã hóa và phân tích dữ liệu.

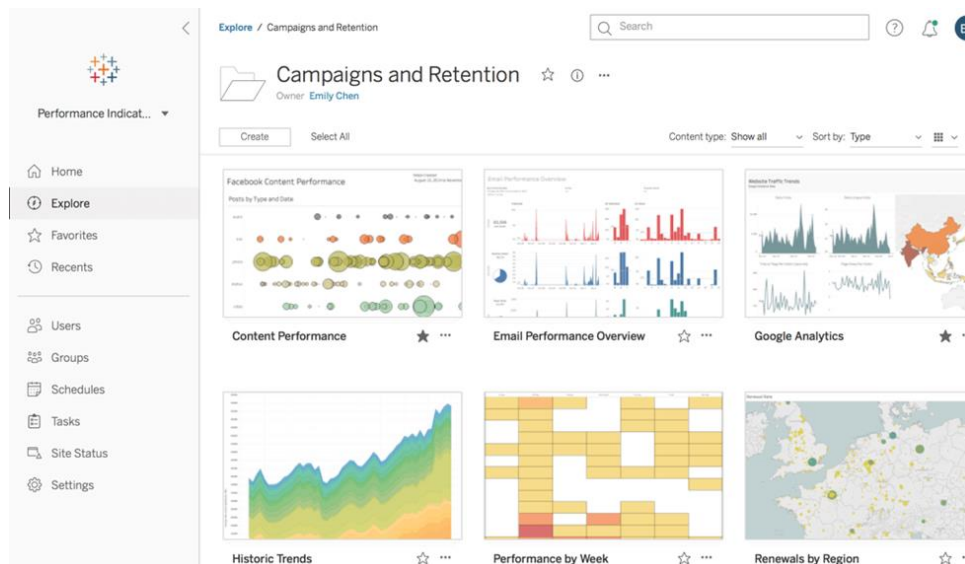
Các thao tác phần lớn là kéo thả và không yêu cầu quá nhiều am hiểu về mặt kỹ thuật hay lập trình. Tableau Desktop có khả năng kết nối rộng rãi đến nhiều định dạng file khác nhau, để đáp ứng tốt nhất nhu cầu phân tích trong nhiều ngành nghề, lĩnh vực.

### Tableau Desktop có thể chia thành 2 loại:

**Tableau Desktop Personal:** Tính năng phát triển tương tự Tableau Desktop nhưng quyền truy cập bị hạn chế. Báo cáo không thể xuất bản trực tuyến, nên được phân phối ngoại tuyến hoặc trong Tableau public.

**Tableau Desktop Professional:** Báo cáo có thể xuất bản trực tuyến hoặc trong máy chủ Tableau. Sở hữu quyền truy cập toàn bộ tất cả các loại dữ liệu, dành cho những người muốn chia sẻ báo cáo trên máy chủ Tableau.

## c. Tableau Online



*Tableau Online là dịch vụ hoàn toàn miễn phí*

Không cần đến máy chủ, không giới hạn lưu trữ, cho phép liên kết đến hơn 40 nguồn dữ liệu khác nhau. Tuy nhiên, để có thể xuất bản, vẫn cần đến Tableau Desktop, có thể hình dung nó giống một server miễn phí.

Một điều cần lưu ý là Tableau Online chia sẻ các xuất bản của bạn đến tất cả mọi người, không nên đặt các dữ liệu quan trọng trên đây. Dù vậy, Tableau Online vẫn cho phép bạn mời các đối tác, khách hàng xem báo cáo trực tuyến qua trình duyệt và ứng dụng di động. Phần lớn người dùng Tableau Online sử dụng cho mục đích học tập.

#### d. Tableau Server

Là nơi chia sẻ các phân tích của doanh nghiệp được bảo mật cẩn thận và cấp quyền truy cập. Giúp mọi người cùng chia sẻ và quản lý dữ liệu trên đám mây, một sản phẩm dành cho các doanh nghiệp và tất nhiên nó có phí.

Cũng giống như Tableau Online, Tableau Server cần đến Tableau Desktop để xuất bản. Tuy nhiên, khi đầu tư chi phí cho Server nhà xuất bản có thể quản lý, bảo mật dữ liệu, cấp quyền truy cập...

Ngoài việc truy cập trực tiếp vào Server để đọc báo cáo, Tableau còn cho phép chia sẻ đến người dùng khác các bảng điều khiển dưới dạng tĩnh. Người nhận được bảng điều khiển này có thể sử dụng Tableau Reader để đọc báo cáo.

## 5. Ưu điểm và nhược điểm Tableau

### a. Ưu điểm

- Dễ dàng thao tác và xây dựng các Dashboard và cá bản phân tích bắt mắt
- Có thể sử dụng cho mọi phòng ban và mọi nhân viên trong bất cứ ngành nghề nào
- Tốc độ xử lý dữ liệu cực kỳ nhanh với công nghệ In-memory
- Khả năng mở rộng cơ sở dữ liệu và mức độ phức tạp cho doanh nghiệp đang phát triển nhanh
- Có khả năng quản lý toàn bộ các công tác, chia sẻ và mức độ bảo mật cao
- Khả năng kết nối và làm việc với nhiều loại dữ liệu cùng lúc
- Đáp ứng được các công nghệ mạnh mẽ như Big Data, AI và khả năng tích hợp cao



- Dữ liệu có thể được chia sẻ với nhau và đưa đến tay người cần để họ tự xử lý
- Tạo ra một môi trường làm việc dựa trên dữ liệu và phân tích dữ liệu
- Luôn có dữ liệu và phân tích mọi lúc mọi nơi








## b. Nhược điểm

- Hạn chế các hỗ trợ truy vấn SQL nâng cao
- Tốn nhiều thời gian khi làm quen với sử dụng
- Chi phí giá thành cao

## III. Sử dụng Tableau để trực quan dữ liệu về Covid-19

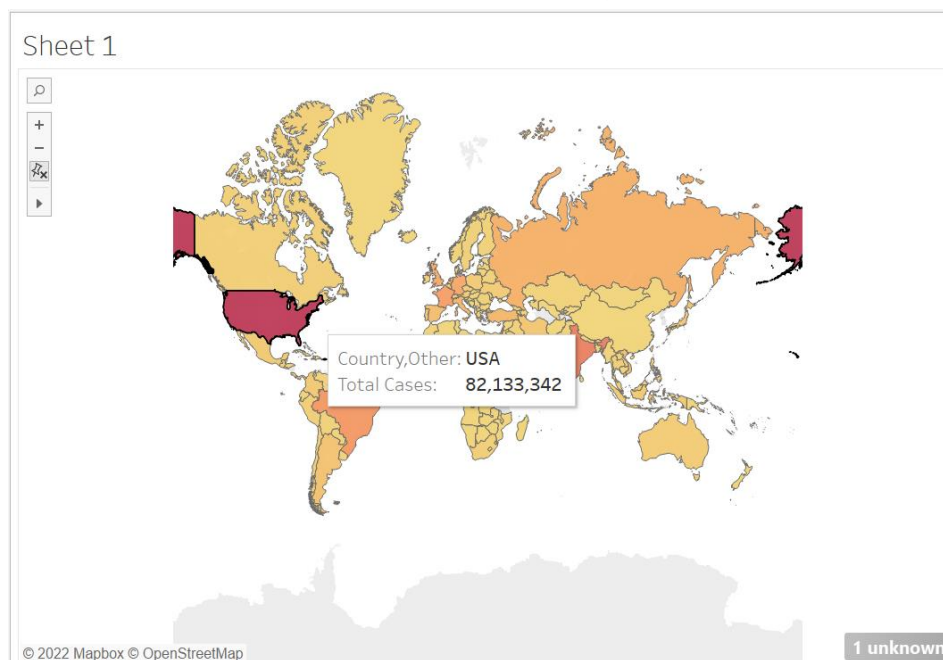
### 1. Dữ liệu chuẩn bị cho trực quan trên Tableau

- Nguồn dữ liệu từ trang: <https://www.worldometers.info/>
- Thời gian lấy dữ liệu từ: 07/04/2022 – 13/04/2022

Name	Date modified	Type	Size
 Covid-07-04.csv	9/4/2022 4:00 pm	XLS Worksheet	31 KB
 Covid-08-04.csv	9/4/2022 4:01 pm	XLS Worksheet	31 KB
 Covid-09-04.csv	10/4/2022 9:24 pm	XLS Worksheet	31 KB
 Covid-10-04.csv	11/4/2022 4:47 pm	XLS Worksheet	30 KB
 Covid-11-04.csv	12/4/2022 2:09 pm	XLS Worksheet	31 KB
 Covid-12-04.csv	13/4/2022 6:47 pm	XLS Worksheet	31 KB
 Covid-13-04.csv	14/4/2022 1:21 am	XLS Worksheet	31 KB

### 2. Trực quan dữ liệu

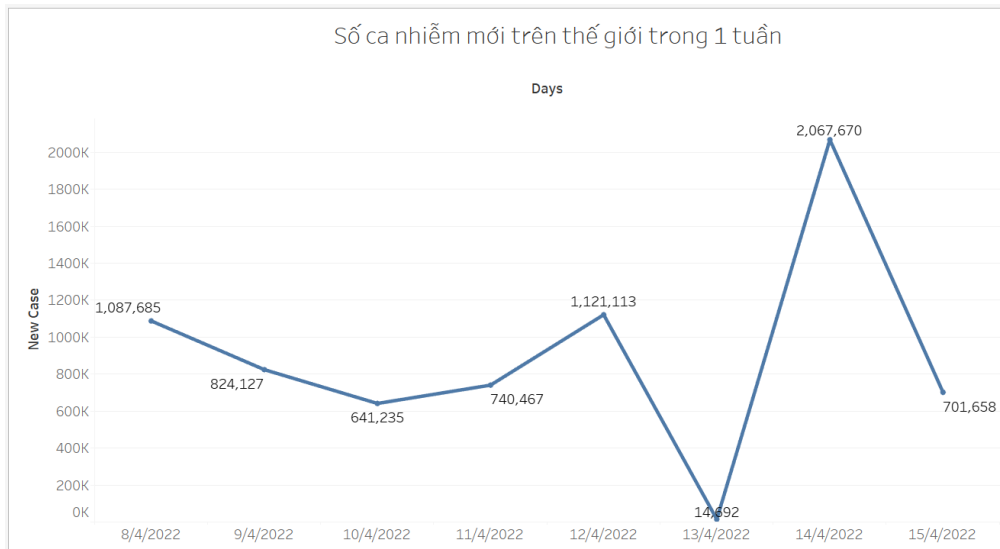
#### a. Tình hình ca nhiễm Covid-19 trên thế giới



#### ▪ Ý nghĩa của biểu đồ:

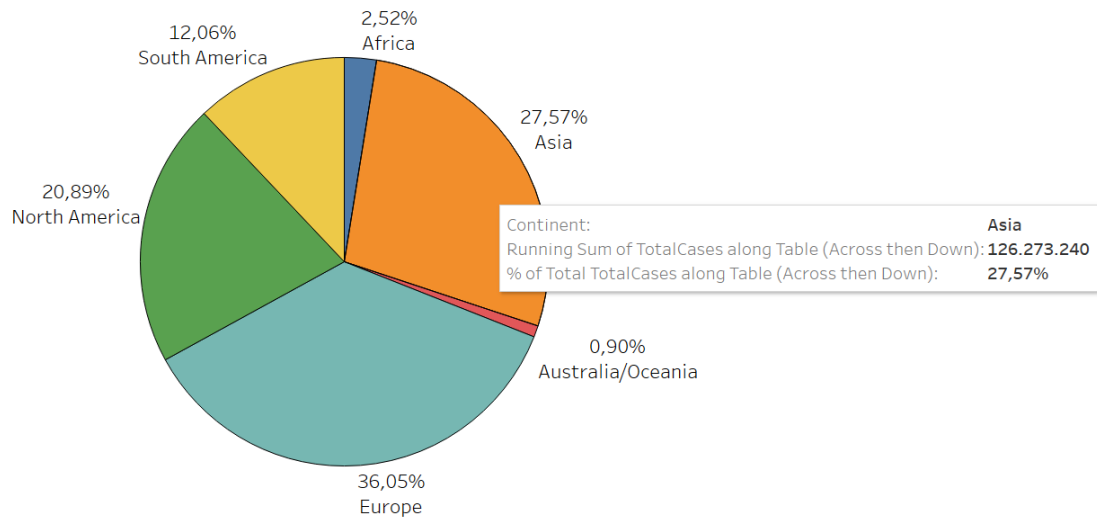
- Dựa map ở trên ta thấy số lượng người nhiễm Covid-19 trên thế giới rất lớn và phủ khắp trên thế giới cho thấy mức độ lây lan mạnh mẽ và nhanh chóng của con viruss này.

- Màu sắc được áp dụng trong biểu đồ sử dụng gam màu nóng với ý đồ thể hiện mức độ nguy hiểm của Covid-19 và ý nghĩa khi áp dụng: khi ta nhìn vào biểu đồ có áp dụng màu ta có thể so sánh số ca nhiễm ở các nước khác nhau thông qua màu sắc mà không cần phải rê chuột vào để xem thông tin số ca nhiễm. Khi ta nhìn vào biểu đồ Map trên ta thấy rõ nước Mỹ là số ca nhiễm nhiều nhất vì có màu sắc là đậm nhất.
- Kỹ thuật được sử dụng Geometric Zooming vì lượng thông tin cả các nước trên thế giới rất nhiều để có thể có cái nhìn rõ hơn về từng khu vực ta có thể zoom lên lên để nhìn rõ hơn.

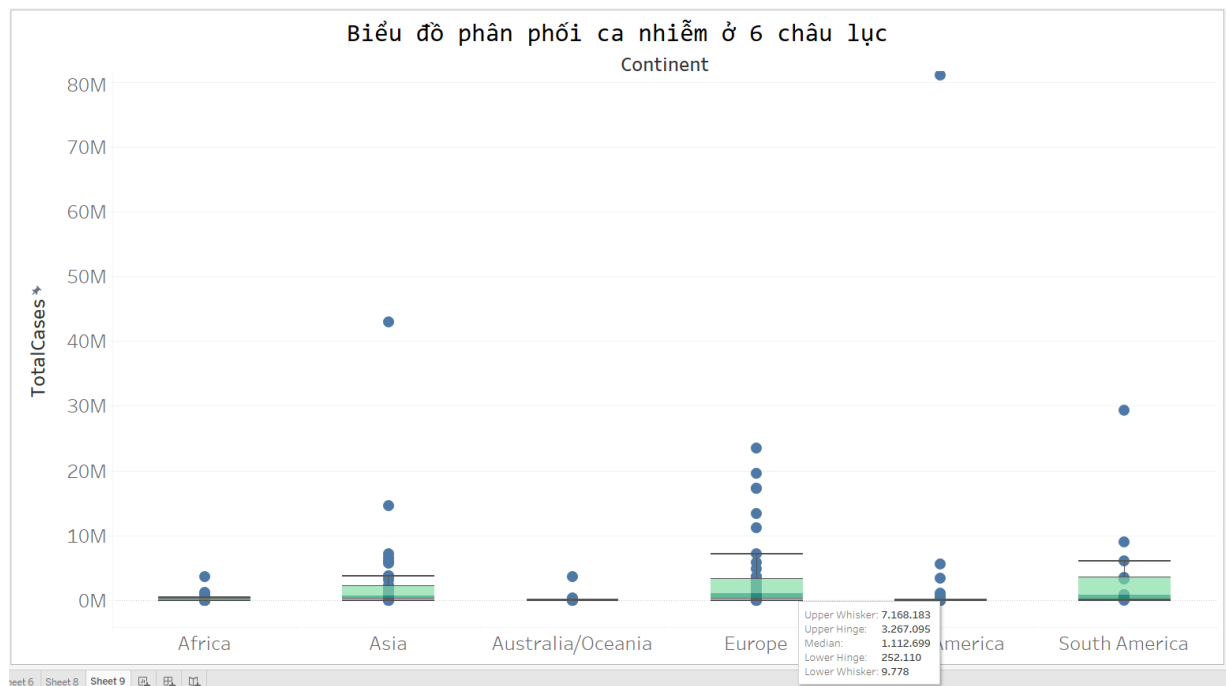


- Ý nghĩa của biểu đồ:
  - Nhìn vào biểu đồ đường ta thấy trong 1 tuần tuy số ca nhiễm mới không có xu hướng tăng hoặc giảm liên tục trong 1 tuần nhưng ta có quan ngại sâu sắc khi số ca nhiễm mới ở ngày thấp nhất cũng rất cao 14,692 ca. Tốc độ lây lan của Covid-19 vẫn chưa có dấu hiệu giảm.
- Màu sắc được sử dụng là mặc định màu xanh dương vì ta chỉ biểu diễn 1 đường duy nhất, do đó màu sắc trong biểu đồ này không đóng góp gì cho quá trình trực quan hóa.

b. Xem xét tỉ lệ ca nhiễm Covid-19 ở 6 châu lục

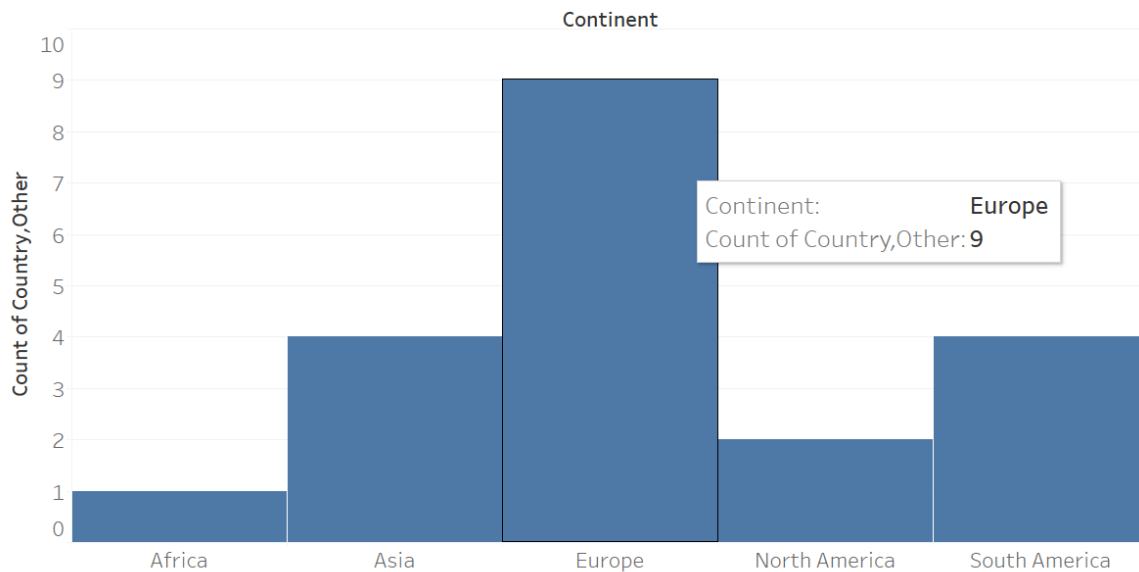


- Ý nghĩa của biểu đồ:
  - Tính tới ngày 08/04/2022, Châu Âu (Europe) chiếm tỉ lệ cao nhất 36.05%, chiếm hơn 1/3 số ca nhiễm trên toàn thế giới, hiện nay châu Âu được xem là tâm điểm của đại dịch. Theo thông tin của nhiều nguồn cho biết Châu Âu phải đối mặt với số ca nhiễm biến thể Omicron tăng lên.
  - Giả thuyết đặt ra: Hiện có thông tin Châu Âu phải đối mặt với biến thể số ca nhiễm Omicron tăng lên, ta đặt ra 2 giả thiết về số ca nhiễm cao nhất ngưỡng ở châu Âu
    1. Do biến thể mới phát triển mạnh ở thời tiết khí hậu Châu Âu qua đó đẩy mạnh tốc độ lây lan.
    2. Do 1 vài nước trong Châu Âu phòng chống dịch không tốt và gây ra số lượng ca nhiễm tăng lên dẫn đến tổng số ca nhiễm ở Châu Âu tăng theo.
 Sử dụng *biểu đồ hộp* để thể hiện sự phân bố ca nhiễm ở các nước trong 6 châu lục
- Màu sắc được áp dụng trong biểu đồ gồm nhiều màu với độ tương phản cao để dễ dàng nhận biết cũng như phân biệt sự phân bố giữa các dữ liệu
- Ta có thể chọn một mảnh của pie chart để thể hiện được % cũng như con số của thể của từng châu lục.



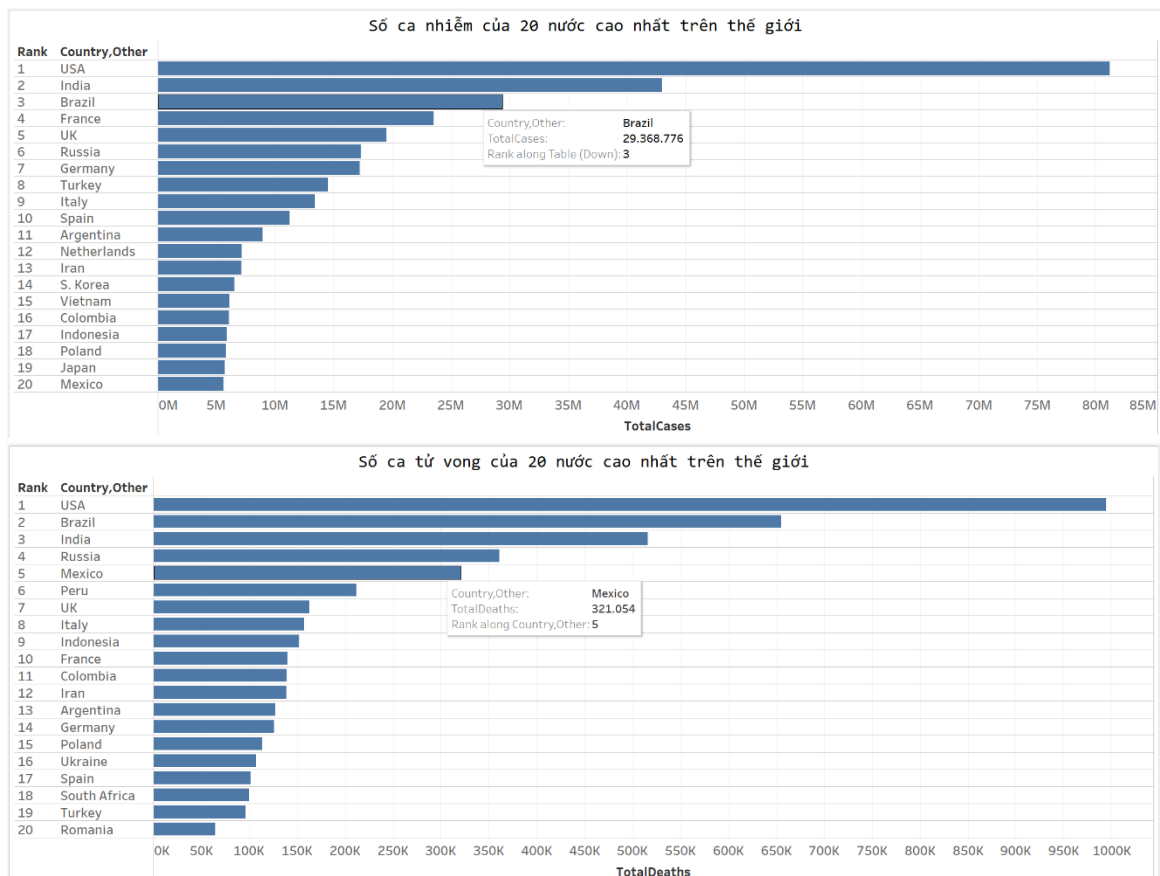
- Ý nghĩa của biểu đồ:
  - Dựa vào box plot, ta thấy rõ ở Europe đã xuất hiện tới năm outlier hình dung được giả thiết đầu tiên tỉ lệ ca nhiễm tăng lên đó là do 1 vài nước trong châu âu phòng chống dịch bệnh chưa tốt của ta đã đúng giải thích rõ ràng tại sao tổng số ca nhiễm của châu âu lại chiếm tới 1/3 so với thế giới. Kèm theo đó có 1 điều đặc biệt ở North America xuất hiện 1 outlier lớn qua đó do 1 mình outlier đó đã kéo số người nhiễm trung bình ở châu mỹ (North America) tăng mạnh.
- Màu sắc được áp dụng trong biểu đồ: các chấm màu xanh biển, đường kẻ giới hạn quartile là màu đen còn vùng quartile được tô gam màu xanh lá thể hiện vùng interquartile range (IQR) của số ca nhiễm ở từng châu lục
- Ta có thể chọn các điểm cũng như vùng quartile để xem rõ các thông số của quartile và trung vị về số ca nhiễm các nước trong một châu lục

### Số nước thuộc các châu lục trong Top 20 nước có ca nhiễm nhiều nhất

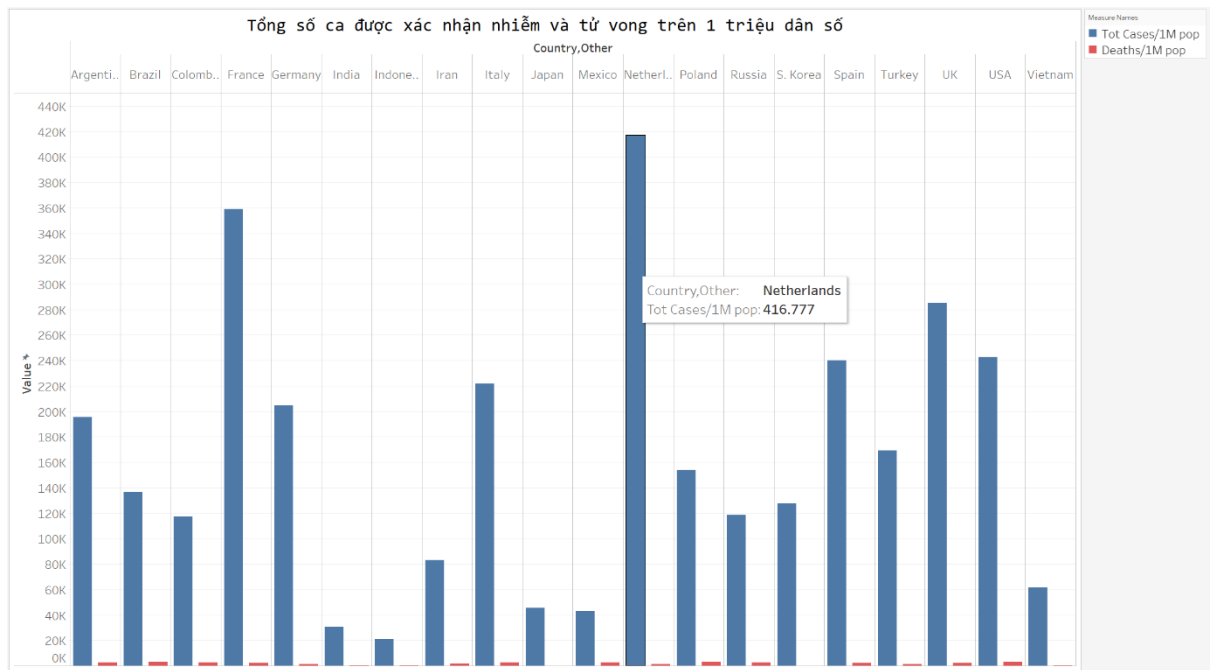


- Ý nghĩa của biểu đồ:
  - Mỹ vẫn chiếm ngôi vị đầu với số ca nhiễm nhiều nhất trên thế giới và gần như là gấp đôi so với nước đứng thứ 2 là Ấn độ. Ở trong histogram ta thấy châu âu (Europe) có nhiều nước trong top 20 nhất (9 nước) => Qua đó ta có thể giải thích lý do tại sao châu âu lại chiếm tỉ lệ nhiễm 1/3 so với thế giới. Đồng thời giả thiết đầu tiên của ta đặt ra là đúng: tỉ lệ ca nhiễm tăng lên đó là do một vài nước trong châu âu phòng chống dịch bệnh chưa tốt
- Màu sắc được áp dụng trong biểu đồ: là một màu xanh do việc tô màu cho khác nhau cho từng châu lục không mang thêm lại thông tin gì khác ngoài ra nó còn gây ra việc khó so sánh độ cao giữa các cột
- Ta có thể chọn một cột để xem được số liệu cụ thể hơn như ví dụ trong hình cột được chọn là cột tổng ca nhiễm trên một triệu dân số của nước Netherlands

#### c. Biểu diễn mức độ tử vong ở 20 nước

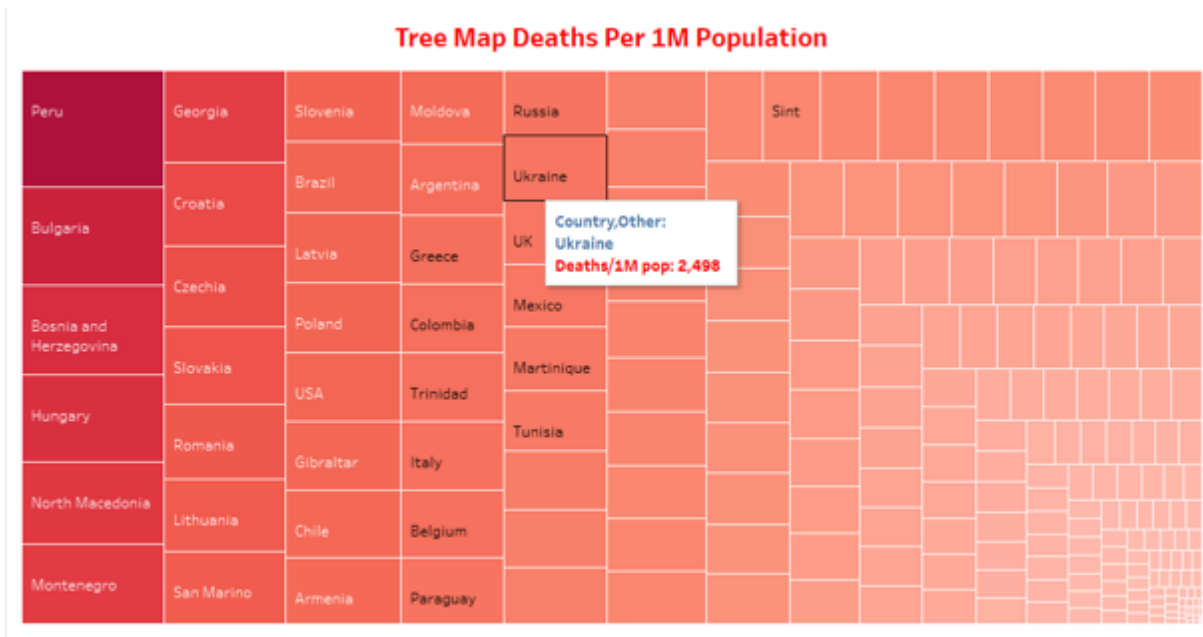


- Ý nghĩa của biểu đồ:
  - Mỹ chiếm ngôi vị đầu với ca nhiễm nhiều nhất trên thế giới và gần như gấp đôi so với nước đứng thứ 2 là Ấn độ. Và ta thấy hầu như các nước trong top 20 về số ca nhiễm nhiễm thì cũng nằm trong top 20 về tử vong.
- Màu sắc được áp dụng trong biểu đồ là một màu do việc tô màu cho 20 nước khác nhau không mang thêm lại thông tin gì khác ngoài ra nó còn gây khó nhìn
- Ta có thể chọn một cột để xem được số liệu cụ thể cũng như rank của nước đó ví dụ trong hình cột được chọn là nước Brazil với Mexico

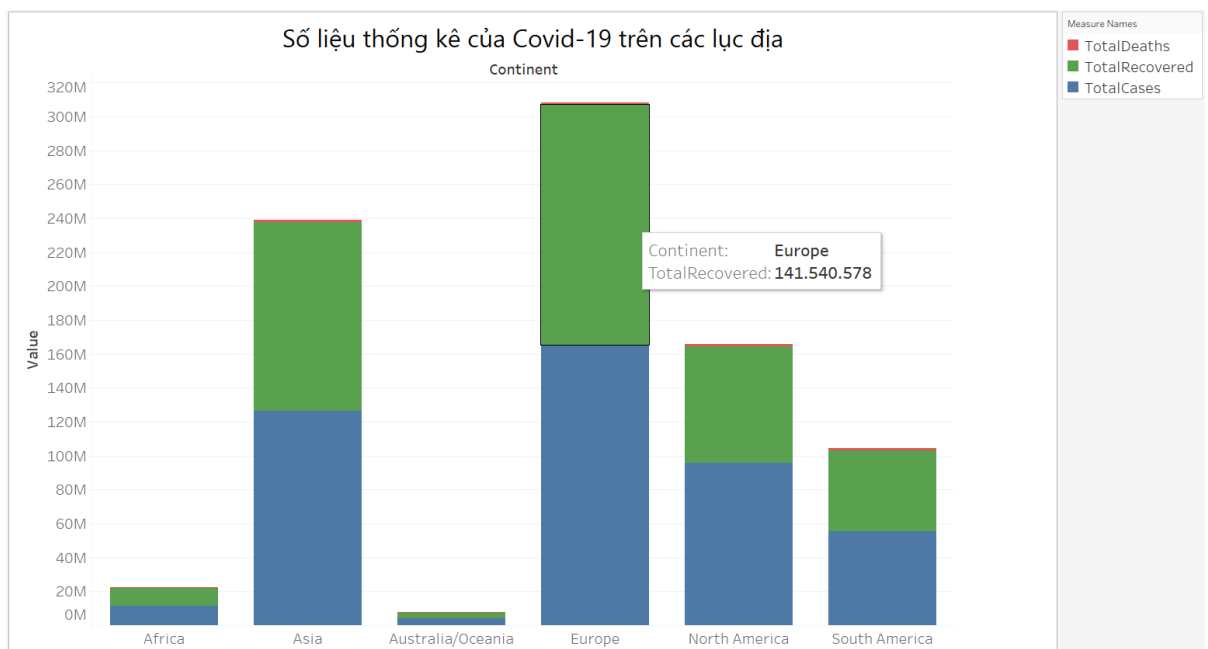


- Ý nghĩa của biểu đồ:
  - Dựa vào biểu đồ “Tổng số ca được xác nhận nhiễm và tử vong trên một triệu dân số” ta có thể thấy tuy nước Mỹ có số ca nhiễm cao nhất và vượt trội so với các nước còn lại như có thể thấy ở trên biểu đồ trên của số người nhiễm và tử vong trên 1 triệu dân số thì nước Mỹ chỉ ở vị trí thứ 4. Nước dẫn đầu là nước Pháp với gần 350000 người nhiễm trên 1 triệu dân số. Lý giải cho vấn đề đó là dân số của nước Mỹ là 329,5 triệu còn của pháp là 67,39 triệu.
- Màu sắc được áp dụng trong biểu đồ: thông tin tổng ca nhiễm trên một triệu dân số có màu xanh còn số ca tử vong trên một triệu dân số còn số ca tử vong thể hiện màu đỏ
- Ta có thể chọn một cột để xem được số liệu cụ thể hơn như ví dụ trong hình cột được chọn là cột tổng ca nhiễm trên một triệu dân số của nước Netherlands

#### d. TreeMap



- Ý nghĩa của biểu đồ: Tree Map cho chúng ta thấy được số ca nhiễm trên 1 triệu dân số. Dữ liệu được sắp xếp từ trái sang phải theo thứ tự giảm dần. Ta còn biết được những nước có dân số ít sẽ có tỷ lệ trên số ca nhiễm cao.
- Màu sắc được áp dụng trong biểu đồ: là màu đỏ thể hiện gram màu nóng mang ý nghĩa nguy hiểm đối với virus Covid-19.
- Ta có thể chọn một cột để xem được số liệu cụ thể hơn như ví dụ trong hình cột được chọn là cột tổng ca nhiễm trên một triệu dân số của nước Ukraine



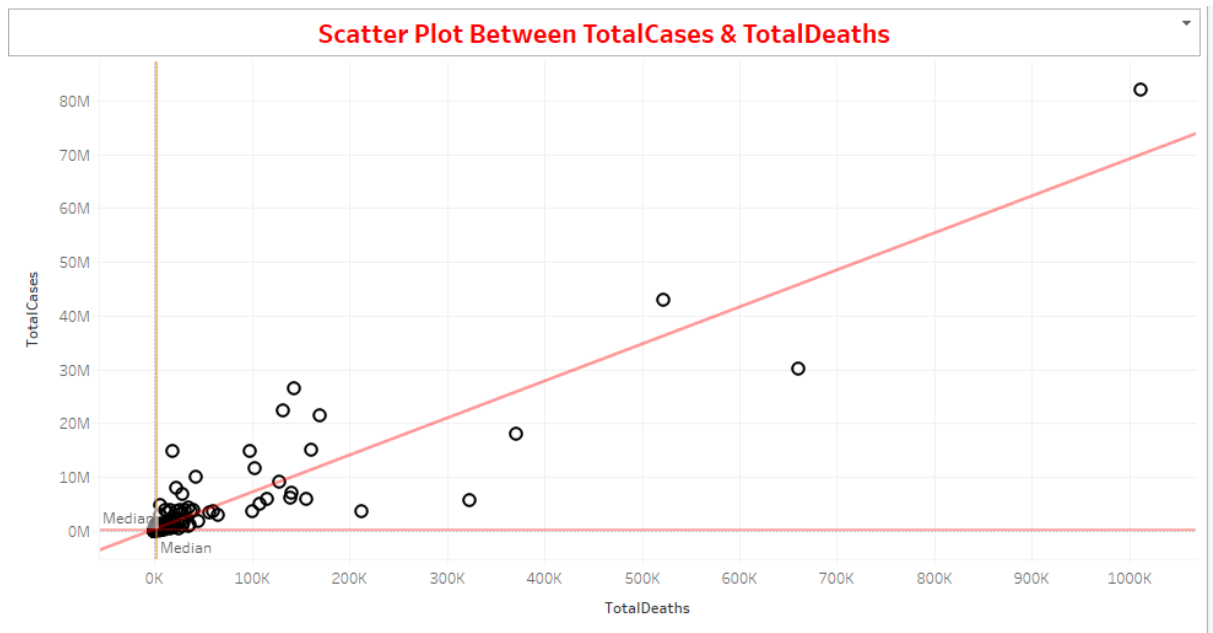
- Ý nghĩa của biểu đồ:
  - Dựa vào biểu đồ cho thấy được số lượng các ca nhiễm, ca tử vong, ca hồi phục của từng lục địa qua đó ta rút ra được tỉ lệ ca tử vong chiếm



một tỉ lệ cực kì thấp trên mọi lục địa. Ta thấy được châu Âu có số ca nhiễm cao nhất mặc dù châu Á có dân số đông nhất thế giới

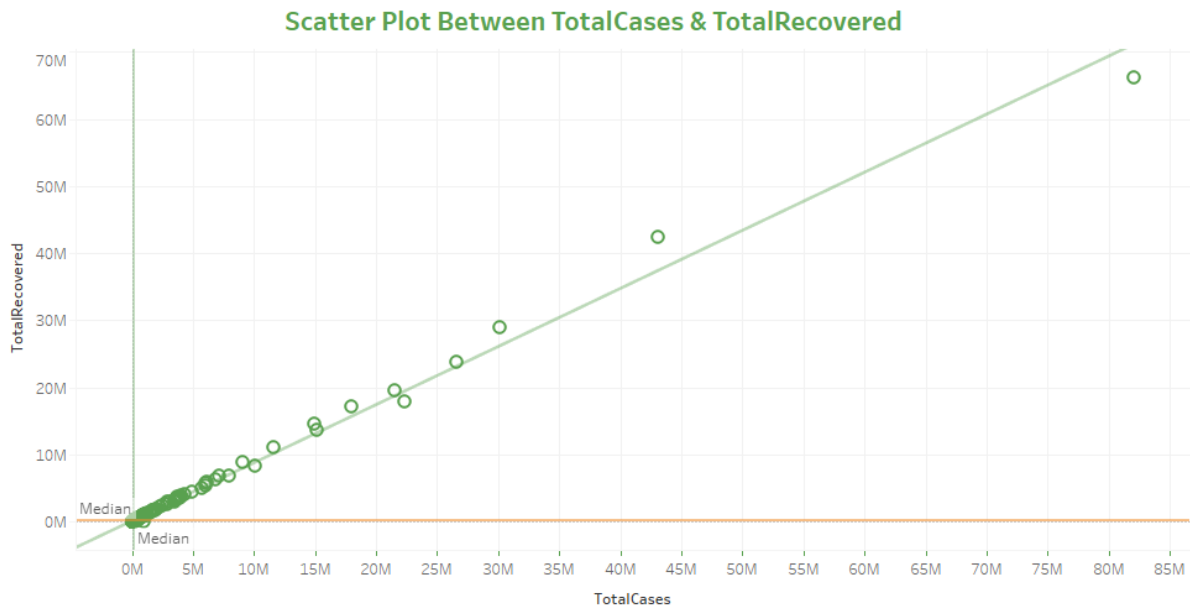
- Màu sắc được áp dụng trong biểu đồ: từng thông tin được thể hiện một màu khác nhau với màu đỏ là ca tử vong, xanh lá là ca hồi phục, xanh biển là tổng số ca để dễ dàng phân biệt giữa ba loại.
- Kỹ thuật được sử dụng ta gộp ba biểu đồ ca tử vong, ca hồi phục, tổng số ca lại thành một biểu đồ để trực quan, ngoài ra còn có thể chỉ vào từng cột để thể hiện con số chính xác

e. Biểu thị sự tương quan giữa các thuộc tính



Ý nghĩa của biểu đồ: Ta thấy sự đồng biến của 2 trường dữ liệu TotalCases và TotalDeaths. Khi TotalCase tăng sẽ dẫn đến TotalDeaths tăng và ngược lại.

- Màu sắc được áp dụng trong biểu đồ sử dụng màu đỏ và cam là những gam màu nóng mang ý nghĩa sự nguy hiểm của virus gây nên số lượng lớn tử vong cho con người.
- Ta có thể click chọn vào từng điểm để xem số liệu cụ thể.



- Ý nghĩa của biểu đồ: Ta thấy 2 biến dữ liệu TotalCases và TotalDeaths đồng biến với nhau. Thật vậy số ca nhiễm càng nhiều thì tỷ lệ khỏi bệnh cũng sẽ tăng theo.
- Màu sắc được áp dụng trong biểu đồ sử dụng màu xanh lục và cam. Màu cam thể hiện sự nguy hiểm của virus Covid-19 và màu xanh lục thể hiện sự sống hồi phục của những người bị nhiễm virus.

## IV. Thuật toán Học máy

### 1. Bộ dữ liệu cho quá trình training

- Nhóm đã thu thập thêm dữ liệu để có thể chạy mô hình học máy với tổng là 10 ngày
- Dữ liệu được dành toàn bộ cho quá trình training và cũng được dùng cho quá trình testing

### 2. Thuật toán Support vector machine

#### a. Mục đích thuật toán

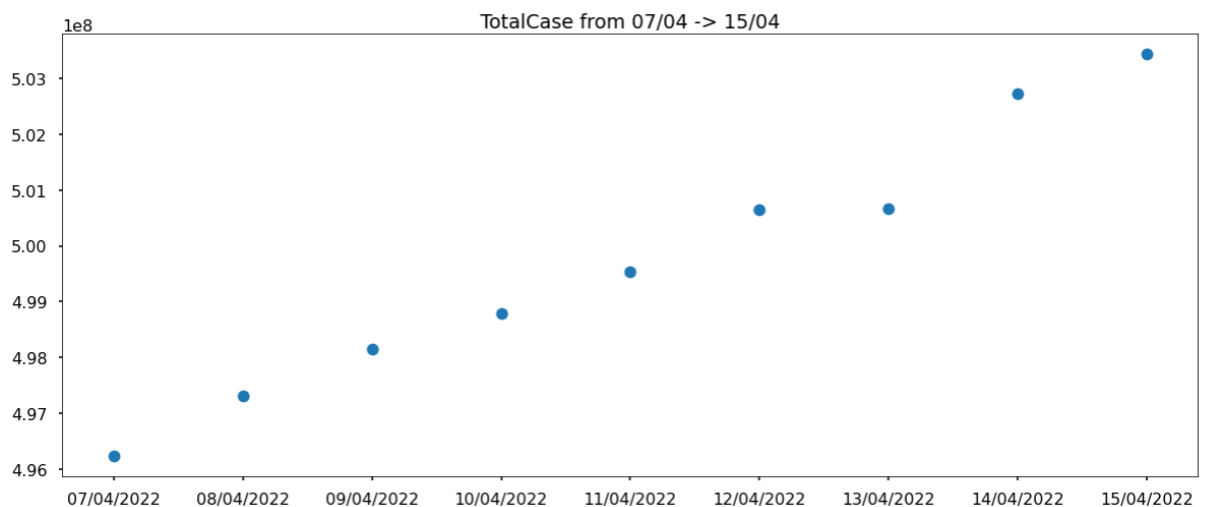
- Nhóm sử dụng SVM cho phân tích hồi quy qua đó dựa vào bộ dữ liệu đã thu thập để *dự đoán tổng số ca nhiễm Covid-19 trên thế giới*, với bộ dữ liệu là tổng số ca nhiễm trên thế giới từ ngày 07/04/2022 -> 15/04/2022

#### b. Giải thích mã nguồn

- Thư viện sử dụng cho SVM: Sklearn

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.svm import SVR
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import mean_squared_error, mean_absolute_error
import datetime
%matplotlib inline
```

- Từ bộ dữ liệu, ta tiến hành trích xuất và tính tổng số ca nhiễm trên toàn thế giới ở từng ngày



(dữ liệu sau trích xuất, xử lý và được trực quan hóa)

- Từ dữ liệu đã được trích xuất, ta sẽ sử dụng SVM cho hồi quy trong Sklearn và train

```
kernel = ['poly', 'sigmoid', 'rbf']
c = [0.01, 0.1, 1, 10]
gamma = [0.01, 0.1, 1]
epsilon = [0.01, 0.1, 1]
shrinking = [True, False] 1
svm_grid = {'kernel': kernel, 'C': c, 'gamma': gamma, 'epsilon': epsilon, 'shrinking': shrinking}

svm = SVR()
svm_search = RandomizedSearchCV(svm, svm_grid, scoring='neg_mean_squared_error', cv=3, return_train_score=True, n_jobs=-1, n_iter=2)
svm_search.fit(np.array(X).reshape(-1, 1), Y_totalcases)
```

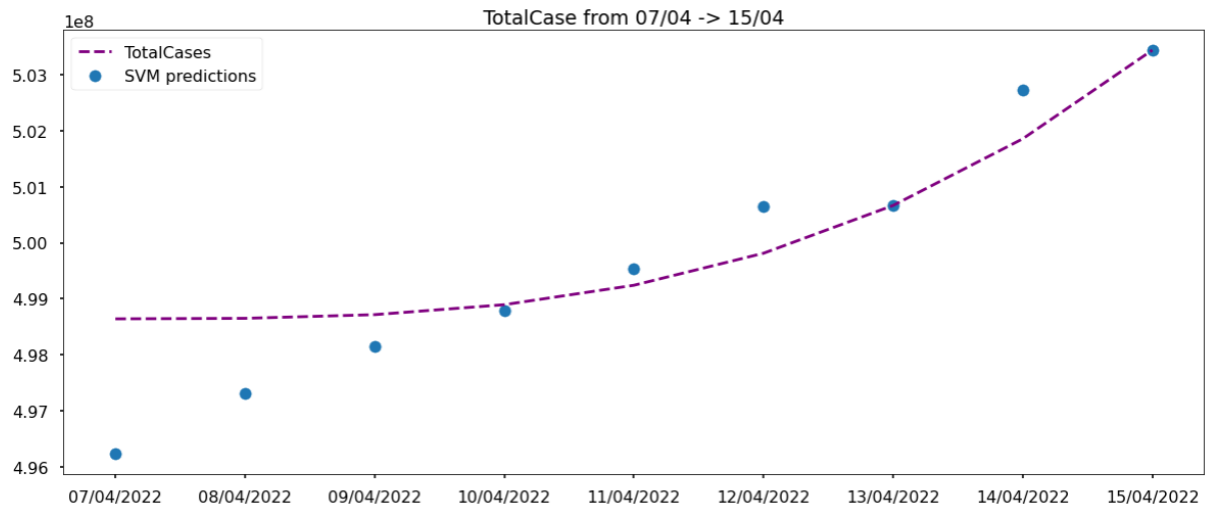
- Trong mô hình SVM thì có các tham số, dựa vào bộ dữ liệu và yêu cầu bài toán cần chọn các cặp tham số cho phù hợp, nhóm đã chọn giải pháp Random Search (2) để chọn ngẫu nhiên các cặp tham số (1) để chọn ra cặp tham số cho ra độ chính xác cao nhất cho mô hình SVM

```
svm_confirmed = svm_search.best_estimator_
```

```
svm_confirmed
```

```
SVR(C=10, epsilon=1, gamma=1, kernel='poly', shrinking=False)
```

- Kết quả thu được sau khi quá trình training



## V. Tài liệu tham khảo

[1] <https://migoda.v/cong-cu-tableau-1648597730>

[2] <https://nhanhua.com/0074in-tuc/tableau-la-gi.html>

[3] <https://bsdinsight.com/phan-mem-tableau-va-nhung-dieu-can-biet/>