



VNUHCM - University of Science  
Information Technology Department  
Subject: Data Mining and Applications.

**Report:**  
**Lab 02**  
**Data Visualization with Tableau**

**Information:**

Đoàn Ánh Dương – 20127474

Bùi Tấn Phương – 20127597

Trần Quốc Trung – 20127625

Trịnh Thế Sơn – 20127617

Trần Thị Thanh Phú – 20127279

**Instructors:**

Bùi Tiến Lên

Lê Ngọc Thành

Ho Chi Minh City, April 18<sup>th</sup> 2023

## Contents

1.	Mức độ hoàn thành .....	3
2.	Một số lưu ý .....	3
3.	Tìm hiểu công cụ Tableau: .....	3
3.1.	Trực quan hóa dữ liệu .....	3
3.2.	Kết nối đến nhiều nguồn dữ liệu .....	7
3.3.	Tích hợp các công cụ phân tích khác .....	9
3.4.	Hỗ trợ cộng tác .....	11
3.5.	Bảo mật dữ liệu .....	13
4.	Thực hành: Vận dụng Tableau để trực quan hóa dữ liệu Worldometer .....	14
4.1.	Khả năng y tế của các Châu lục trong dịch bệnh.....	14
4.2.	Tình hình dịch bệnh hiện tại của các châu lục trên thế giới .....	45
4.3.	Số ca mắc cao thì sẽ dẫn đến số ca tử vong cao? .....	49
4.4.	Tốc độ khỏi bệnh của Covid-19 là bao nhiêu? Những quốc gia nào đang có tỷ lệ khỏi bệnh cao nhất? .....	51
4.5.	So sánh tình hình dịch bệnh và khả năng kiểm soát bệnh ở giữa các nước phát triển và các nước đang phát triển .....	54
4.6.	Tốc độ lây lan của virus Covid-19 có khác biệt giữa các khu vực không? Những khu vực nào đang ghi nhận tốc độ lây lan cao nhất? .....	64
4.7.	Có mối tương quan nào giữa mật độ dân số của một quốc gia và số ca nhiễm COVID-19 của quốc gia đó không? .....	71
5.	Áp dụng một số kỹ thuật máy học để hiểu rõ hơn về dữ liệu.....	77
6.	References .....	83

# 1. Mức độ hoàn thành

MSSV	Họ và tên	Công việc	Hoàn Thành
20127597	Bùi Tân Phương	3.3, 4.1, 4.4	100 %
20127625	Trần Quốc Trung	3.2, 4.5, 4.3	100 %
20127474	Đoàn Ánh Dương	3.1, 5, 4.2	100 %
20127617	Trịnh Thê Sơn	3.4, Doc, 4.2, 4.5	100 %
20127279	Trần Thị Thanh Phú	3.5, 4.6, 4.7	100 %

## 2. Một số lưu ý

Để có thể chạy Tableau, nhóm em bổ sung và cài đặt một số công cụ tích hợp của Tableau:

- Cài đặt Tabpy và kết nối với Tableau thông qua port 9004.
- Thư viện python: sklearn, numpy, pandas,...

Trước khi mở file Tableau, cần phải chạy tabpy trước và chạy code trong file “model.py”.

Cấu trúc tập tin gồm có:

- Data: Chứa toàn bộ file data dạng \*.csv từ Lab01.
- Source Code:
  - model.py: file code để chạy với tabpy.
  - preprocessing.ipynb: xử lý data, chuẩn bị cho mô hình học máy. Khi chạy cần để ngoài folder Source Code.
- Visualization.twbx: file tổng hợp chart của nhóm.
- Report.pdf: file báo cáo.

## 3. Tìm hiểu công cụ Tableau:

Tableau là một công cụ trực quan hóa dữ liệu và trí tuệ doanh nghiệp (business intelligence) mạnh mẽ cho phép người dùng dễ dàng kết nối, phân tích và chia sẻ dữ liệu bằng cách trực quan và tương tác. Nó cho phép người dùng tạo và chia sẻ các bảng điều khiển, báo cáo và biểu đồ tương tác thông qua một giao diện kéo và thả mà không cần có kiến thức về công nghệ quá rộng.

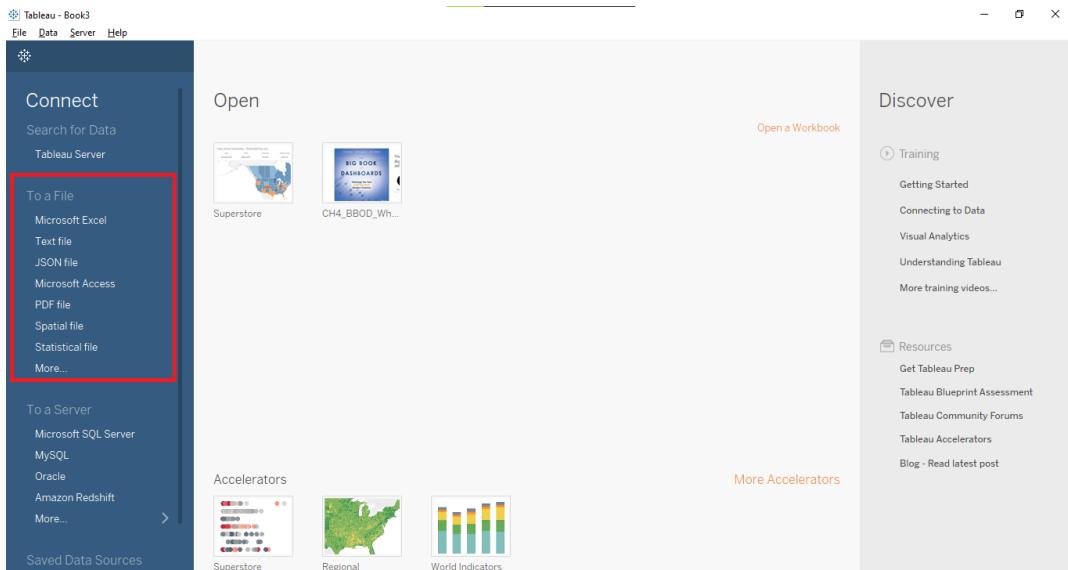
Các chức năng chính của Tableau:

### 3.1. Trực quan hóa dữ liệu

Tính năng mạnh mẽ nhất đáng kể đến trong Tableau là trực quan hóa dữ liệu, cũng là tính năng quan trọng nhất của Tableau. Không chỉ cho phép doanh nghiệp hay cá nhân sử dụng nhiều thư viện đồ thị, hình ảnh phong phú mà còn cho phép tương tác trực tiếp với dữ liệu thông qua Dashboard. Ngoài ra, xét về mặt tốc độ xử lý, Tableau cũng là một trong những công cụ ưu thế trong việc xử lý tạo đồ thị, biểu đồ, ...

Để có thể trực quan hóa dữ liệu với Tableau trước hết ta cần phải thực hiện theo các bước sau:

B1: Load dataset vào Tableau, Tableau hỗ trợ nhiều loại file khác nhau như \*.xlsx, \*.csv,...



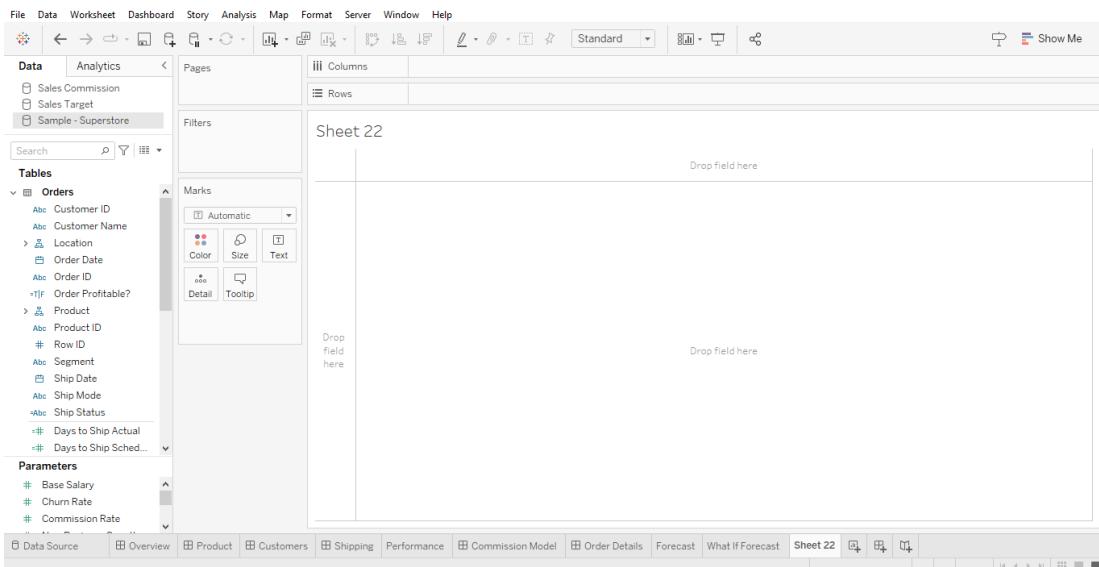
B2: Sau khi mở được dataset rồi sẽ có giao diện như ở hình 2, có 3 chế độ:

- Một Worksheet chứa một view (visualization) đơn.
- Một Dashboard là tập hợp của các views từ nhiều worksheet khác nhau. Ngoài ra Dashboard còn hỗ trợ các tính năng khác như text,...
- Một Story chứa một chuỗi các worksheets hoặc dashboards làm việc cùng nhau để truyền tải thông tin.

Category	Order Date	Segment	Sales Target
Office Supplies	1/3/2020	Consumer	18
Office Supplies	1/4/2020	Home Office	300
Office Supplies	1/5/2020	Consumer	21
Furniture	1/6/2020	Home Office	2316
Office Supplies	1/6/2020	Consumer	17

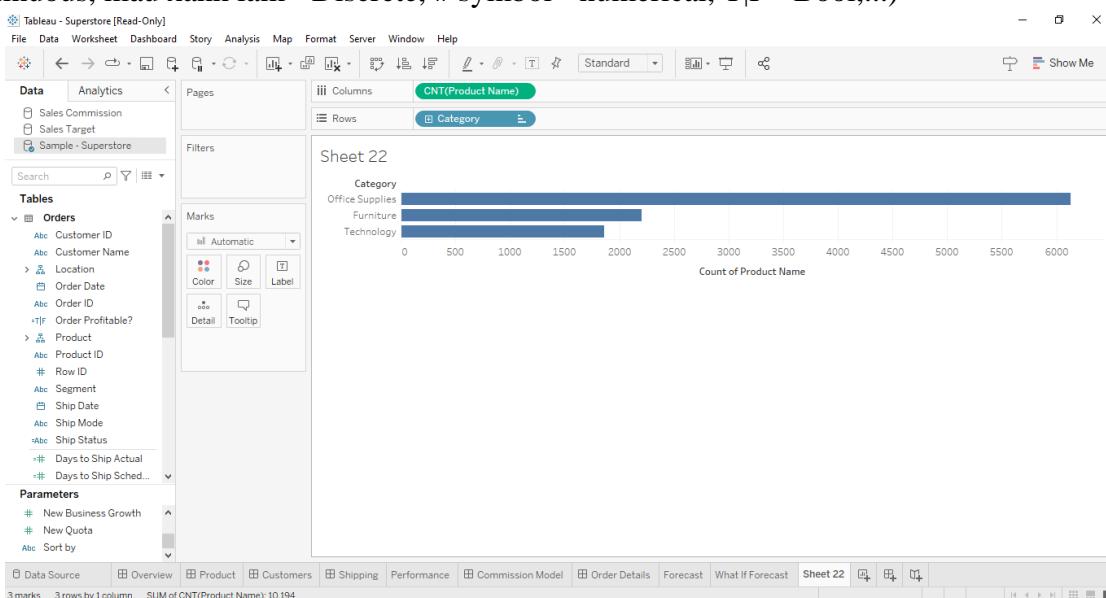
Để có thể trực quan hóa với Tableau, ta có thể chọn sheet có sẵn như trong hình.

B3: Sau khi chọn hoặc tạo một sheet bất kỳ, Tableau sẽ hiển thị giao diện như hình 3. Ta tiến hành lọc (chọn) những trường dữ liệu phù hợp cho việc trực quan hóa, tùy vào nhu cầu và mục đích khác nhau mà ta sẽ chọn những trường khác nhau trong Tag Table. Sau đó kéo thả vào Columns và Rows.

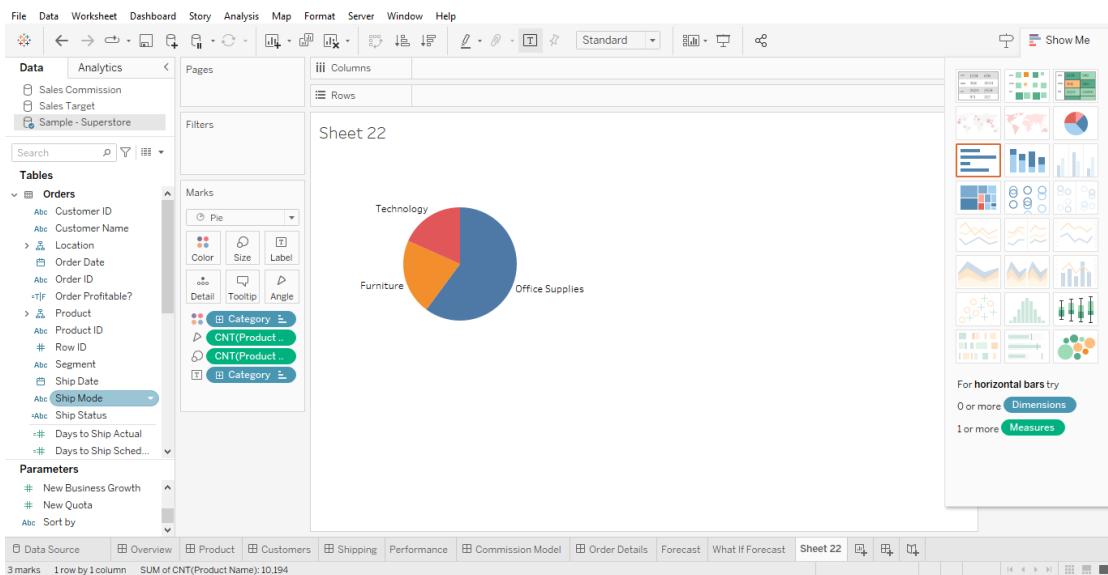


B4: Lấy ví dụ với Dataset có sẵn là UpperStore. Xét trường hợp nhóm em muốn xem số lượng Product được bán ra. Nhóm em cần phải đưa Product Field vào Rows và COUNT(Product) vào Columns.

Một số lưu ý khi trực quan, ta cần để ý tới data type của dữ liệu đó (Ví dụ: màu xanh lục - Continuous, màu xanh lam - Discrete, # symbol - numerical, T|F - Bool,...)



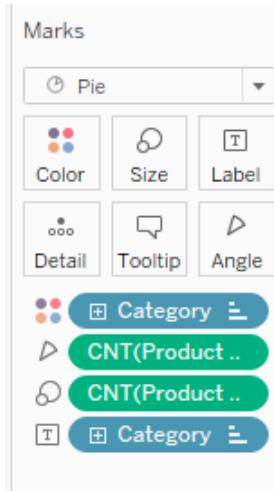
B5: Ở trên ta đã có một horizontal chart, chúng ta có thể tùy chọn nhiều loại chart khác nhau để visualize tại tag Show me. Ví dụ chọn Pie Chart, Ta có được biểu đồ như hình sau.



Ngoài ra, trong Tag “Show me”, Tableau còn hỗ trợ nhiều loại biểu đồ khác nhau tùy theo nhu cầu của người dùng hoặc doanh nghiệp. Tuy nhiên mỗi biểu đồ cần có yêu cầu khác nhau (số lượng Dimension, Số lượng measure, ...) thì mới có thể sử dụng được. Ví dụ như Pie Chart cần 1-2 Measures và 1-n Dimensions.



Người dùng hoặc doanh nghiệp còn có thể tùy chỉnh biểu đồ theo nhu ý muốn của mình (màu sắc, kích thước,...) trong tag Mark sau khi tạo ra biểu đồ. Ví dụ, Xét ở trường hợp trên, ta cần thêm label cho các pieces, bằng cách kéo thả Category vào Label của Marks Tag, hoặc chỉnh sửa màu sắc thì sẽ kéo vào Color,...



Trong tag Filter, Filters có thể giúp giảm thiểu kích thước của tập dữ liệu để sử dụng hiệu quả, loại bỏ các phần tử dimension không liên quan, dọn dẹp dữ liệu cơ bản, đặt phạm vi ngày và các measure theo yêu cầu, đơn giản hóa và tổ chức dữ liệu,... Một số Filters phổ biến như: Context, Extract, Dimension,...

### 3.2. Kết nối đến nhiều nguồn dữ liệu

Tableau hỗ trợ kết nối đến các nguồn dữ liệu web, bao gồm API, web và các tệp JSON. Tableau không kết nối trực tiếp với dữ liệu của web, mà thông qua một middleware là **Web Data Connector**. Web Data Connector là một trang web được viết bằng html và javascript dùng để lưu dữ liệu của một trang web. Sau đây là cách tableau lấy dữ liệu từ web

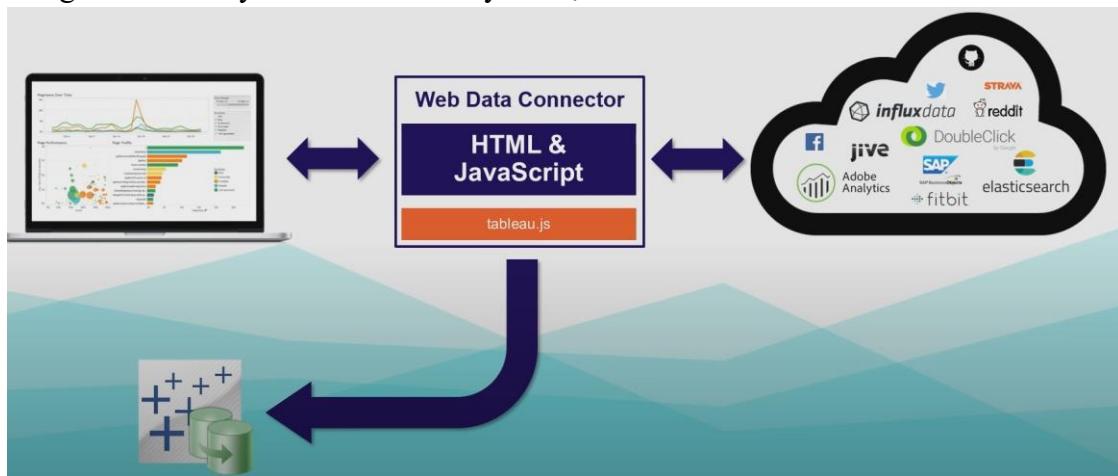


Tableau truy cập vào Web Data Connector -> **Web Data Connector** gửi request đến trang web cần lấy dữ liệu -> Trang web trả về dữ liệu và được lưu vào dữ liệu của tableau

Ví dụ minh họa:

B1: Tạo Web Data Connector để truy vấn đến trang web mong muốn. Ở đây ta sẽ dùng những Web Data Connector đã được viết sẵn bởi Tableau Community ở trang web <https://tableau.github.io/webdataconnector/community/>

The screenshot shows the Tableau GitHub repository page. In the center, there's a section titled "Community Portal" which says: "The following connectors have been written by the Tableau Community and made available to use. If you write a connector, please contribute!" Below this, there are several cards for different connectors:

- AdStage**: Written by David Haslem, Source Code Available, Description: A basic adapter to pull last month of campaign data from AdStage API.
- Alpha Vantage**: Written by Satish C., Source Code Available, Description: A connector for stock data.
- Alpha Vantage Official**: Written by Patrick Collins, Source Code Available, Description: A connector for stock, forex, cryptocurrency, technical indicators, and sector performance.
- Almibro API**: Written by Joshua Frisby, Source Code Available, Description: A Tableau Web Data Connector for AlmibroAPI (<http://www.almibroapi.com/>).
- Appbot Data Connector**: Written by Appbot, Description: A Tableau Web Data Connector for Appbot (<https://appbot.com>). Examine correlation between marketing activities and app reviews/ratings; create your own multi-series or aggregate Ratings data charts from Appbot data, and much more. Learn more: <https://support.appbot.com/help/doc/tableau>
- Aprimo Platform Connector**: Written by Solutions Plus, Description: A Tableau WDC for the Aprimo Platform.
- Army Corps. of Engineers - U.S.**: Description: A Tableau WDC for the Army Corps of Engineers.

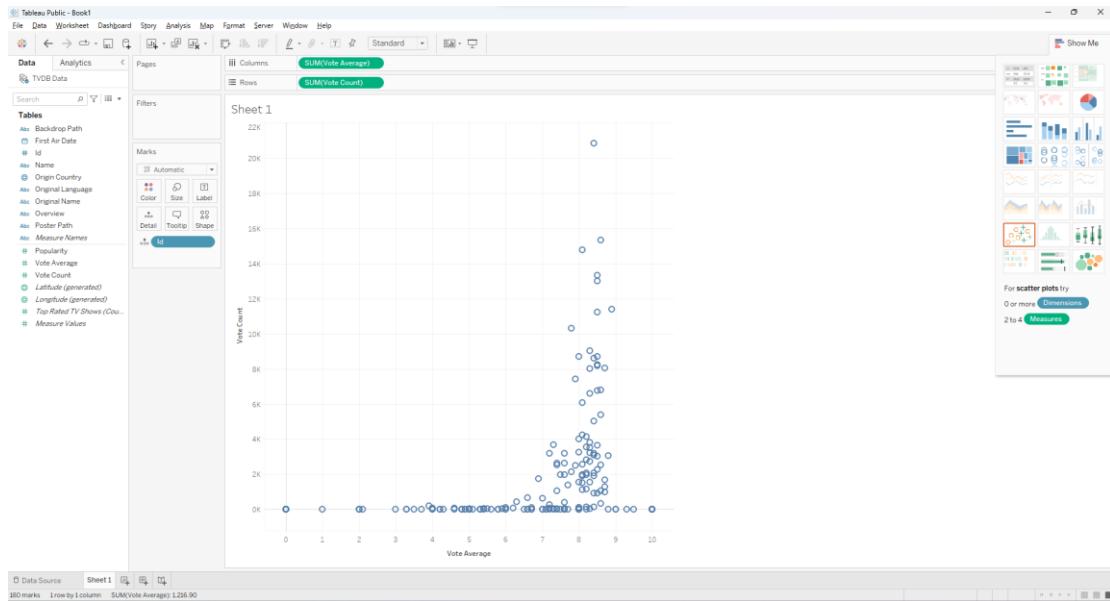
B2: Chọn một Web Data Connector và dán link vào ô input. Nhấn Enter

The screenshot shows the Tableau Public interface. On the left, there's a sidebar with "Connect" selected. Under "To a File", there are options like Microsoft Excel, Text file, JSON file, Microsoft Access, PDF file, Spatial file, Statistical file. Under "To a Server", there are options like OData, Web Data Connector (deprecated), and More... A modal window titled "Web Data Connector" is open, showing a preview of a bubble chart and the URL <https://brendan.github.io/tvWDC.html>.

Nhấn Get Data (hoặc tên khác tùy vào người lập trình) để thực hiện request data đến trang web. Kết quả

The screenshot shows the Tableau interface with the "TVDB Data" connection selected. On the left, there's a sidebar with "Connections" and "Add". Under "Table", there's a table named "Top Rated TV Shows". A preview of the table is shown with columns: Name, Poster Path, Popularity, Id, Backdrop Path, Vote Average, Overview, First Air Date, Origin Country, Original Language, and Vote Count. At the bottom, there are buttons for "Update Now" and "Update Automatically". A "Go to Worksheet" button is also visible.

B3: Thực hiện trực quan trên dữ liệu thu được



### 3.3. Tích hợp các công cụ phân tích khác

- a. Các công cụ tích hợp: Tableau cung cấp khả năng tích hợp mạnh mẽ và rộng rãi với các công cụ phân tích khác để đáp ứng nhu cầu phân tích dữ liệu phức tạp hơn:
- Tích hợp với các ngôn ngữ lập trình như Python, R, SQL
  - Tích hợp với các công cụ Big data như Hadoop, Spark để thực hiện phân tích dữ liệu lớn và đa dạng
  - Tích hợp với các công cụ ETL(Extract, Transform, Load - là các công cụ chuẩn bị và chuyển đổi dữ liệu từ các nguồn khác nhau để phục vụ cho các mục đích phân tích dữ liệu.)
  - Ngoài ra, Tableau còn có thể tích hợp với công cụ quản lý dữ liệu như SAP và Oracle để truy cập và phân tích dữ liệu trong các hệ thống quản lý dữ liệu doanh nghiệp.
- b. Ví dụ minh họa: ở đây, chúng ta sẽ minh họa trên khả năng tích hợp với ngôn ngữ lập trình Python để xử lý và phân tích dữ liệu trên Tableau

B1: Cài đặt Tabpy và chạy Tabpy

```
Anaconda Prompt - tabpy
Requirement already satisfied: pycparser in c:\users\dell\anaconda3\lib\site-packages (from cffi>=1.12->cryptography<40, >=38.0.0->pyopenssl->tabpy) (2.21)
Installing collected packages: genson, docopt, simplejson, exceptiongroup, coverage, configparser, hypothesis, coveralls, textblob, pytest-cov, tabpy
Successfully installed configparser-5.3.0 coverage-6.5.0 coveralls-3.3.1 docopt-0.6.2 exceptiongroup-1.1.1 genson-1.2.2 hypothesis-6.70.2 pytest-cov-4.0.0 simplejson-3.18.4 tabpy-2.6.0 textblob-0.17.1

(base) C:\Users\DELL>tabpy
2023-04-04,16:20:54 [INFO] (app.py:app:248): Parsing config file C:\Users\DELL\anaconda3\lib\site-packages\tabpy\tabpy_server\app..\common\default.conf
2023-04-04,16:20:54 [INFO] (app.py:app:443): Loading state from state file C:\Users\DELL\anaconda3\Lib\site-packages\tabpy\tabpy_server\state.ini
2023-04-04,16:20:54 [INFO] (app.py:app:338): Password file is not specified: Authentication is not enabled
2023-04-04,16:20:54 [INFO] (app.py:app:353): Call context logging is disabled
2023-04-04,16:20:54 [INFO] (app.py:app:129): Initializing TabPy...
2023-04-04,16:20:54 [INFO] (callbacks.py:callbacks:43): Initializing TabPy Server...
2023-04-04,16:20:54 [INFO] (app.py:app:133): Done initializing TabPy.
2023-04-04,16:20:54 [INFO] (app.py:app:83): Setting max request size to 104857600 bytes
2023-04-04,16:20:54 [INFO] (callbacks.py:callbacks:64): Initializing models...
2023-04-04,16:20:54 [INFO] (app.py:app:110): Web service listening on port 9004
2023-04-04,16:21:19 [INFO] (web.py:web:2239): 200 GET / (::1) 72.90ms
2023-04-04,16:21:19 [INFO] (base_handler.py:base_handler:115): Authentication is not a required feature for API "v1"
2023-04-04,16:21:19 [INFO] (base_handler.py:base_handler:115): Authorization header not found
2023-04-04,16:21:19 [INFO] (web.py:web:2239): 200 GET /info (::1) 14.69ms
2023-04-04,16:21:19 [INFO] (base_handler.py:base_handler:115): Authentication is not a required feature for API "v1"
2023-04-04,16:21:19 [INFO] (base_handler.py:base_handler:115): Authorization header not found
2023-04-04,16:21:19 [INFO] (web.py:web:2239): 200 GET /endpoints (::1) 12.57ms
2023-04-04,16:21:19 [INFO] (web.py:web:2239): 200 GET /tableau.png (::1) 11.11ms
2023-04-04,16:21:20 [WARNING] (web.py:web:2239): 404 GET /favicon.ico (::1) 1.15ms
```

## TabPy Server Info:

```
{  
    "description": "",  
    "creation_time": "0",  
    "state_path": "C:\\Users\\DELL\\anaconda3\\Lib\\site-packages\\tabpy\\tabpy_server",  
    "server_version": "2.6.0",  
    "name": "TabPy Server",  
    "versions": {  
        "v1": {  
            "features": {  
                "evaluate_enabled": true,  
                "gzip_enabled": true  
            }  
        }  
    }  
}
```



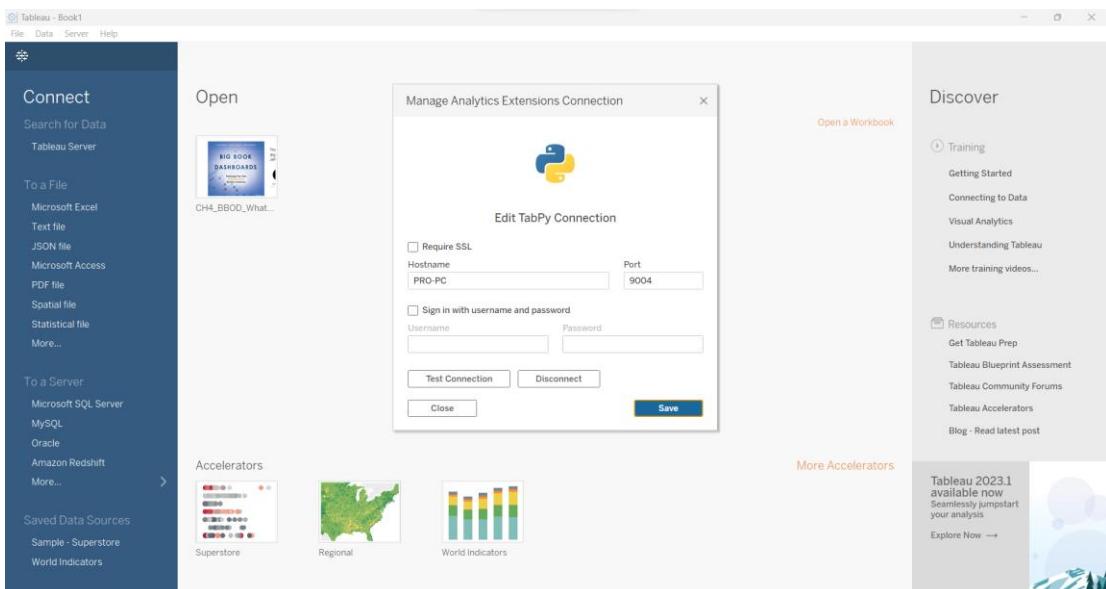
## Deployed Models:

```
{}
```

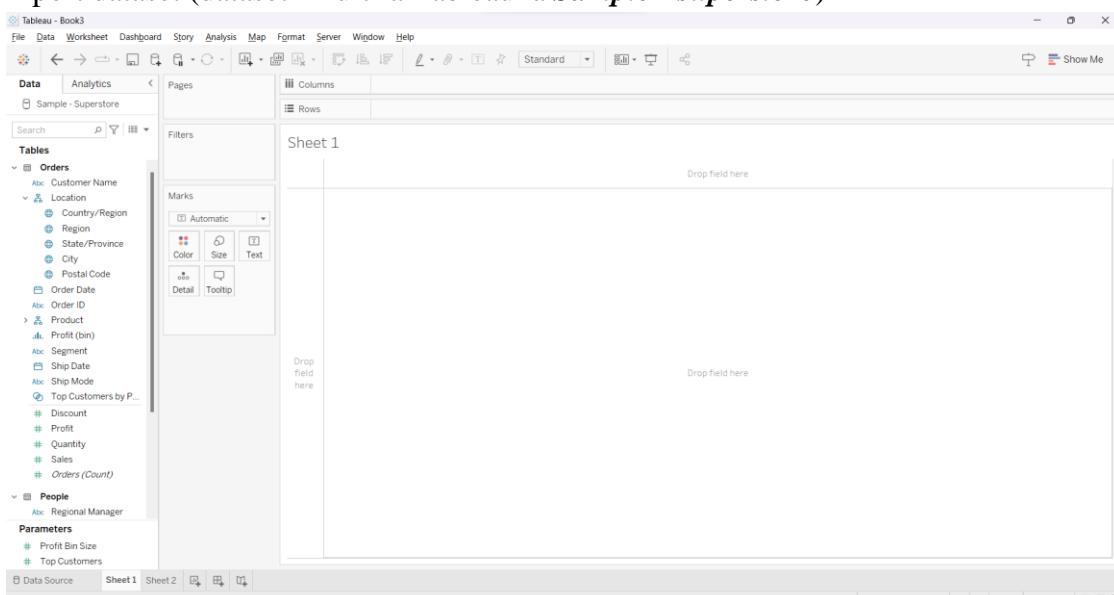
## Useful links:

- [TabPy Documentation](#)
- [TabPy Source Code](#)
- [TabPy PyPi](#)

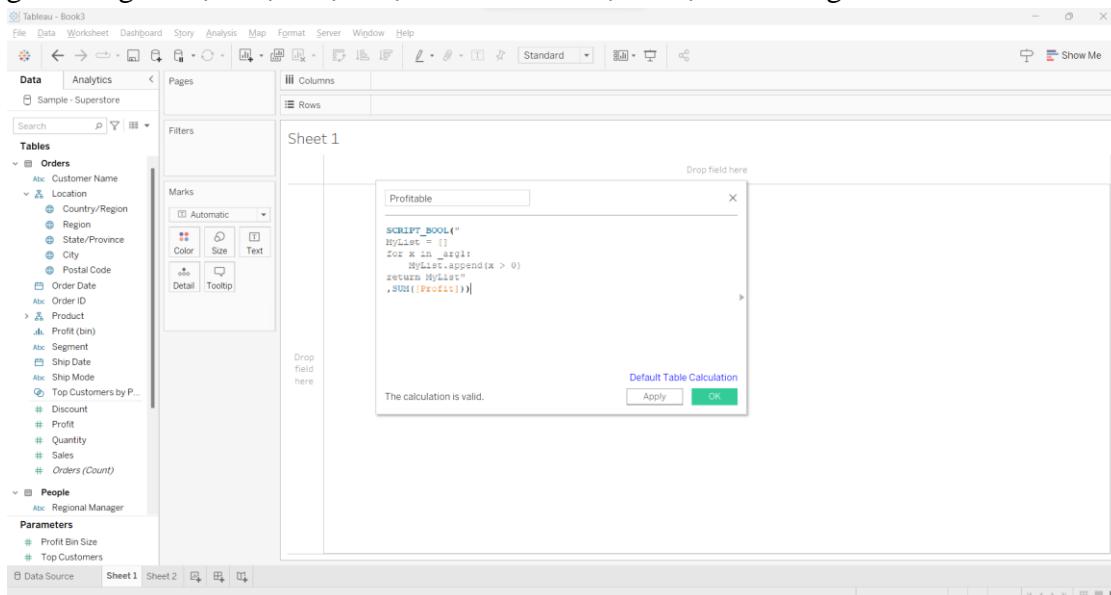
## B2: Liên kết với Tableau



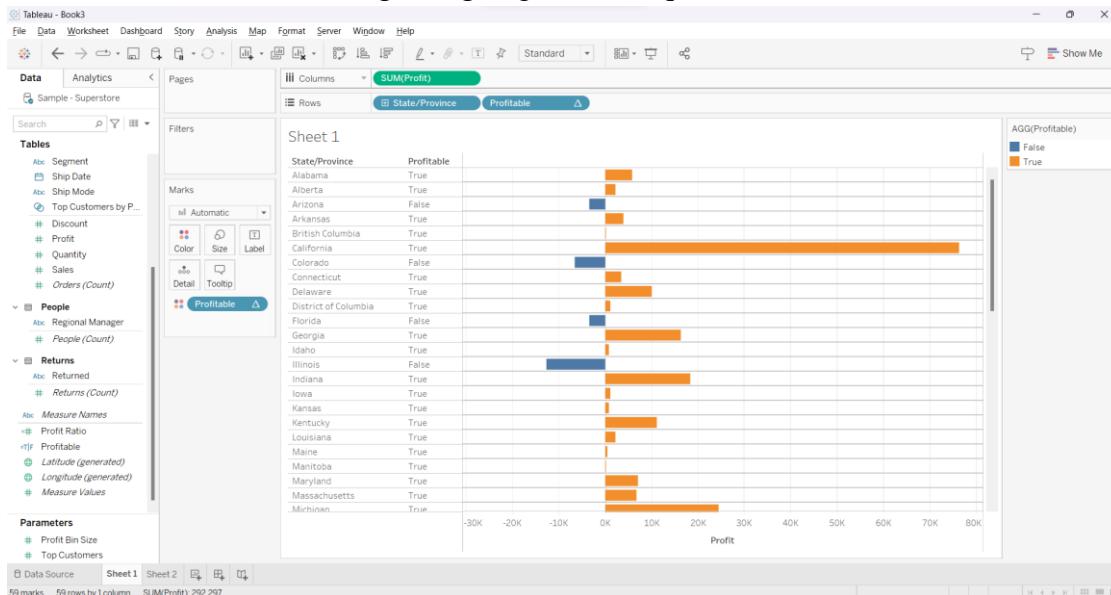
## B3: import dataset (dataset mẫu của Tableau là *Sample - superstore*)



#### B4: giờ chúng ta thực hiện một đoạn code để tính lợi nhuận mỗi bang



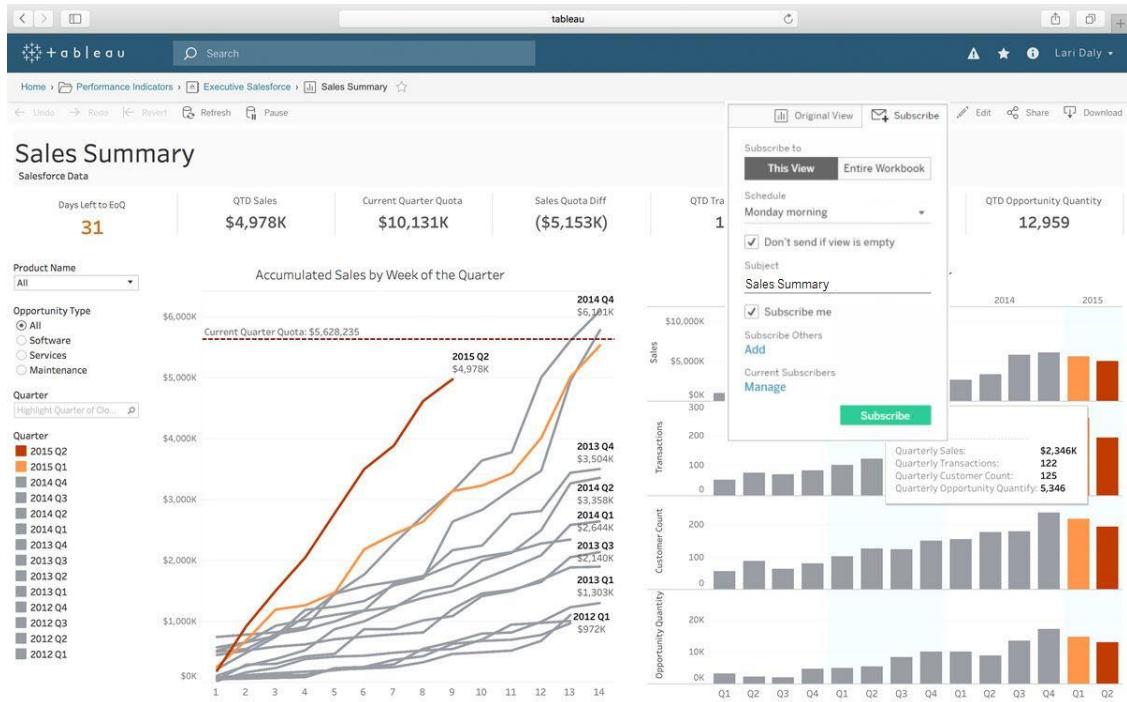
#### B5: Đưa các miền vào cột và hàng tương ứng ta có kết quả



### 3.4. Hỗ trợ cộng tác

Tableau được xây dựng để cộng tác. Các thành viên trong nhóm có thể chia sẻ dữ liệu, thực hiện các truy vấn tiếp theo và chuyển tiếp các hình ảnh trực quan dễ hiểu cho những người khác có thể thu được giá trị từ dữ liệu.

Bằng cách chia sẻ workbook của bạn lên Tableau Server hoặc Tableau Cloud với quyền chỉnh sửa web, các nhóm có thể làm việc cùng nhau để mang lại cho họ khả năng làm việc hiệu quả và cộng tác.



- Nếu bạn hoặc công ty của bạn không sử dụng Tableau Server hoặc nếu bạn muốn tìm hiểu về tùy chọn chia sẻ thay thế, miễn phí, hãy chuyển sang sử dụng Tableau Public.
- Nếu bạn hoặc công ty của bạn sử dụng Tableau Server và bạn đã quen với những quyền được gán cho mình, hãy chuyển sang Sử dụng Tableau Server.

Trước khi chia sẻ workbook của bạn, hãy lưu nó vào một vị trí mà tất cả các thành viên trong nhóm có thể truy cập được.

B1: Để chia sẻ workbook của bạn với người khác, bạn cần publish nó lên Tableau Server hoặc Tableau Online. Nhấp vào menu "Server" trong Tableau Desktop và chọn “Publish workbook”.

B2: Chọn project: bạn sẽ cần chọn một project để chia sẻ.

B3: Sau khi chọn 1 project, bạn có thể đặt quyền cho workbook. Quyền xác định ai có thể xem, chỉnh sửa và tương tác workbook. Bạn có thể đặt quyền cho từng người dùng hoặc nhóm người dùng và bạn có thể cấp các cấp độ truy cập khác nhau, chẳng hạn như quyền truy cập chỉ xem hoặc quyền chỉnh sửa.

B4: Publish: Sau khi bạn đã đặt quyền, hãy nhấp vào nút "Publish" để upload workbook của bạn lên Tableau Server hoặc Tableau Online, nó sẽ sẵn dùng cho những người dùng khác truy nhập.

B5: Chia sẻ liên kết số làm việc: Để chia sẻ số làm việc với người khác, bạn cần chia sẻ liên kết tới số làm việc trên Tableau Server hoặc Tableau Online. Bạn có thể tìm thấy liên kết bằng cách nhấp vào nút "Share" bên workbook trên Server. Bạn có thể chia sẻ liên kết qua email hoặc qua các kênh liên lạc khác.

Sau khi số làm việc được chia sẻ, nhiều người dùng có thể cộng tác đồng thời trên số làm việc đó. Họ có thể xem và tương tác với các phần trực quan hóa, chỉnh sửa số làm việc cũng như để lại nhận xét và phản hồi cho những người dùng khác.

### 3.5. Bảo mật dữ liệu

**Để bảo mật dữ liệu trong Tableau, có thể áp dụng các phương pháp sau:**

Sử dụng mật khẩu: Tableau cho phép tạo mật khẩu cho các bảng dữ liệu, tập tin hoặc bảng trên máy chủ. Bằng cách này, chỉ người dùng được cấp quyền mới có thể truy cập vào dữ liệu.

B1: Truy cập vào bảng dữ liệu trong Tableau và chọn "Bảo mật" từ menu.

B2: Tạo mật khẩu mới và xác nhận lại mật khẩu đó.

B3: Lưu mật khẩu và thoát khỏi bảng dữ liệu.

Khi người dùng muốn truy cập vào bảng dữ liệu đó, họ phải nhập mật khẩu để mở bảng dữ liệu. Nếu mật khẩu nhập vào chính xác, người dùng có thể truy cập vào bảng dữ liệu và xem dữ liệu. Bằng cách này, bạn có thể bảo mật bảng dữ liệu trên Tableau bằng cách sử dụng mật khẩu. Chỉ những người dùng biết mật khẩu mới có thể truy cập vào bảng dữ liệu.

Sử dụng kết nối bảo mật: Sử dụng kết nối bảo mật có thể giúp giảm thiểu rủi ro an ninh dữ liệu. Bạn có thể sử dụng kết nối bảo mật SSL/TLS cho các kết nối đến máy chủ Tableau.

Cấu hình quyền truy cập: Cấu hình quyền truy cập theo cách phù hợp để chỉ cho phép những người dùng được cấp quyền có thể truy cập vào dữ liệu.

Sử dụng tính năng mã hóa: Tableau cung cấp tính năng mã hóa dữ liệu để bảo vệ các dữ liệu nhạy cảm của bạn.

B1: Trong Tableau, chọn bảng dữ liệu mà bạn muốn mã hóa và chọn "Mã hóa" từ menu.

B2: Chọn phương thức mã hóa mà bạn muốn sử dụng. Nếu bạn sử dụng phương thức mã hóa theo khóa, bạn cần tạo một khóa mã hóa và lưu nó lại trên máy tính của bạn.

B3: Sau khi hoàn thành cấu hình, Tableau sẽ tiến hành mã hóa dữ liệu trên máy chủ hoặc trong tập tin Tableau và lưu trữ dữ liệu dưới dạng được mã hóa.

B4: Khi bạn muốn truy cập dữ liệu, Tableau sẽ tự động giải mã dữ liệu và hiển thị dữ liệu đúng theo yêu cầu của bạn.

Bằng cách này, bạn có thể bảo mật dữ liệu trên máy chủ và trong các tập tin Tableau bằng cách sử dụng tính năng mã hóa. Chỉ những người được cấp quyền truy cập và biết cách giải mã dữ liệu mới có thể truy cập và xem dữ liệu.

Kiểm soát truy cập: Kiểm soát truy cập bằng cách cho phép chỉ những người dùng có quyền truy cập vào bảng dữ liệu được cấp phép có thể truy cập vào dữ liệu.

B1: Truy cập vào trang quản lý truy cập của Tableau Server và tạo các nhóm người dùng, và gán các quyền truy cập cho mỗi nhóm người dùng tùy theo vai trò của họ.

B2: Cấu hình các quyền truy cập cho các người dùng cụ thể bằng cách gán họ vào các nhóm người dùng tương ứng. Bạn cũng có thể cấu hình các quyền truy cập cho người dùng đơn lẻ, bằng cách cấu hình các quyền truy cập của họ trực tiếp thay vì gán họ vào một nhóm người dùng.

B3: Lưu lại các cấu hình của bạn và kiểm tra lại để đảm bảo rằng người dùng có quyền truy cập đúng theo yêu cầu.

Bằng cách này, bạn có thể quản lý quyền truy cập của người dùng đến các tài nguyên trên máy chủ Tableau và bảo mật dữ liệu một cách hiệu quả hơn.

Sử dụng trình quản lý quyền: Sử dụng trình quản lý quyền để quản lý các quyền truy cập của người dùng đến dữ liệu. Tableau Server cung cấp tính năng quản lý quyền để quản lý quyền truy cập của người dùng đến các bảng dữ liệu trên máy chủ.

B1: Đăng nhập vào Tableau Server và truy cập vào trang quản lý trình quản lý quyền.

B2: Chọn danh sách bảng dữ liệu hoặc tập tin mà bạn muốn cấu hình quyền truy cập cho người dùng.

B3: Tạo một nhóm người dùng và cấu hình các quyền truy cập cho nhóm đó. Bạn có thể cho phép nhóm truy cập chỉ đọc hoặc truy cập đầy đủ vào bảng dữ liệu.

B4: Chọn người dùng cần cấu hình quyền truy cập và gán người dùng đó vào nhóm vừa tạo.

B5: Lưu các cấu hình của bạn và kiểm tra lại xem người dùng đã được cấp quyền truy cập đúng theo yêu cầu chưa.

Bằng cách này, bạn có thể quản lý quyền truy cập của người dùng đến các bảng dữ liệu trên Tableau Server một cách hiệu quả và bảo mật hơn

→ Tuy nhiên, để đảm bảo an toàn tuyệt đối cho dữ liệu, nên sử dụng một loạt các biện pháp bảo mật khác nhau để bảo vệ dữ liệu của bạn.

## 4. Thực hành: Vận dụng Tableau để trực quan hóa dữ liệu Worldometer

### 4.1. Khả năng y tế của các Châu lục trong dịch bệnh

#### 4.1.1. Tiếp cận vấn đề:

Ở câu hỏi này, ta sẽ sử dụng dữ liệu được thu thập vào ngày 17/03/2023 để trả lời câu hỏi, câu trả lời sẽ được chia ra 2 ý nhỏ để trả lời là:

- Khả năng chống dịch
- Khả năng chữa bệnh

Về phương pháp làm thì sẽ là quy trình cơ bản:

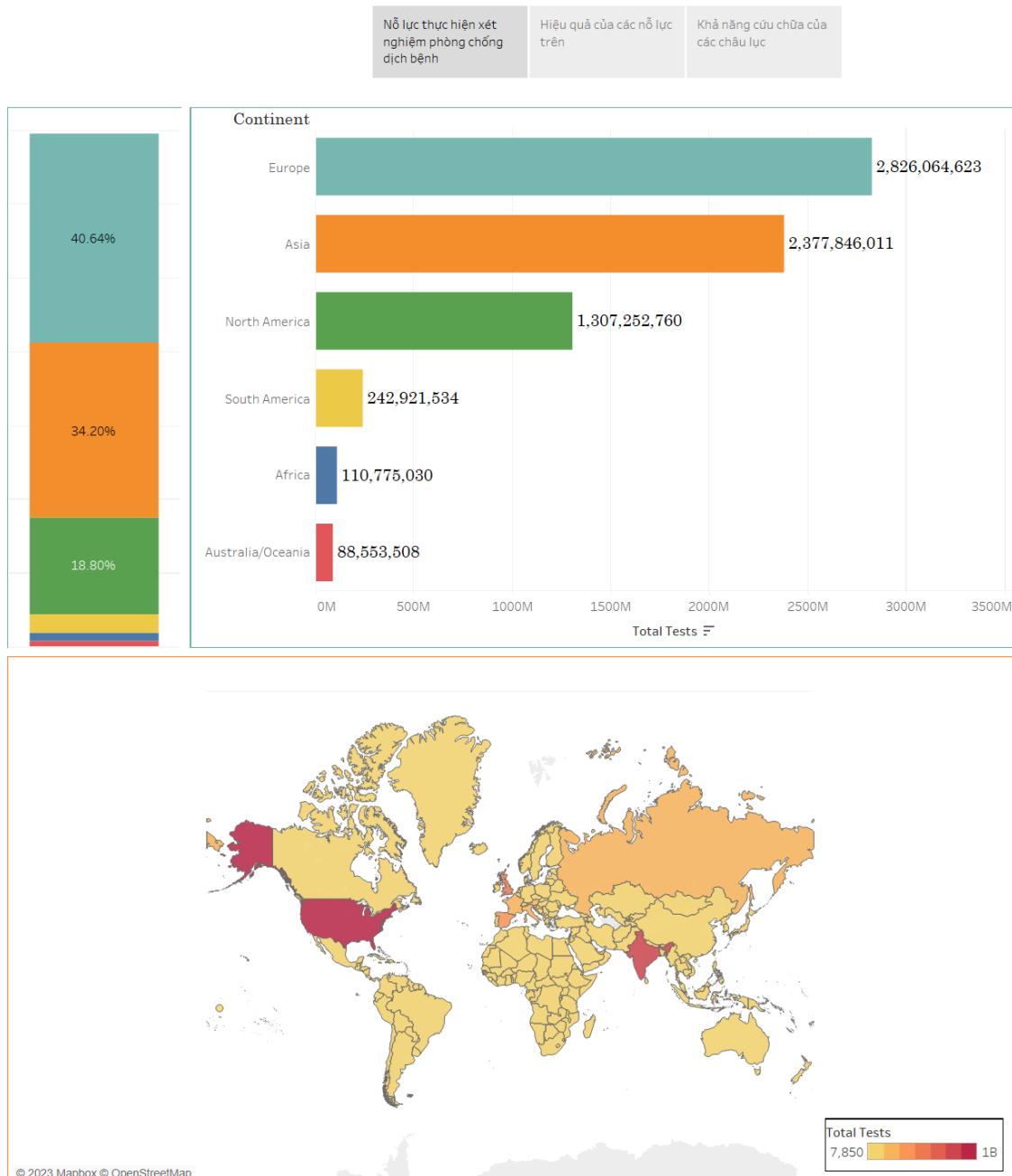
- Thực hiện trực quan hóa từng sheet (sheet là một đơn vị dùng để trực quan hóa 1 biểu đồ nhất định)
- Đưa các biểu đồ liên quan với nhau vào chung một dashboard (là một trang tổng hợp các sheet và các thành phần khác như Text, Extension, v.v)

- Và cuối cùng, nếu dashboard đó chỉ là 1 phần nhỏ cho câu trả lời thì ta gộp các dashboard và story của tableau

#### 4.1.2. Trực quan trên Tableau:

##### a. Nỗ lực thực hiện xét nghiệm phòng chống dịch bệnh

Khả năng y tế của các châu lục trên thế giới



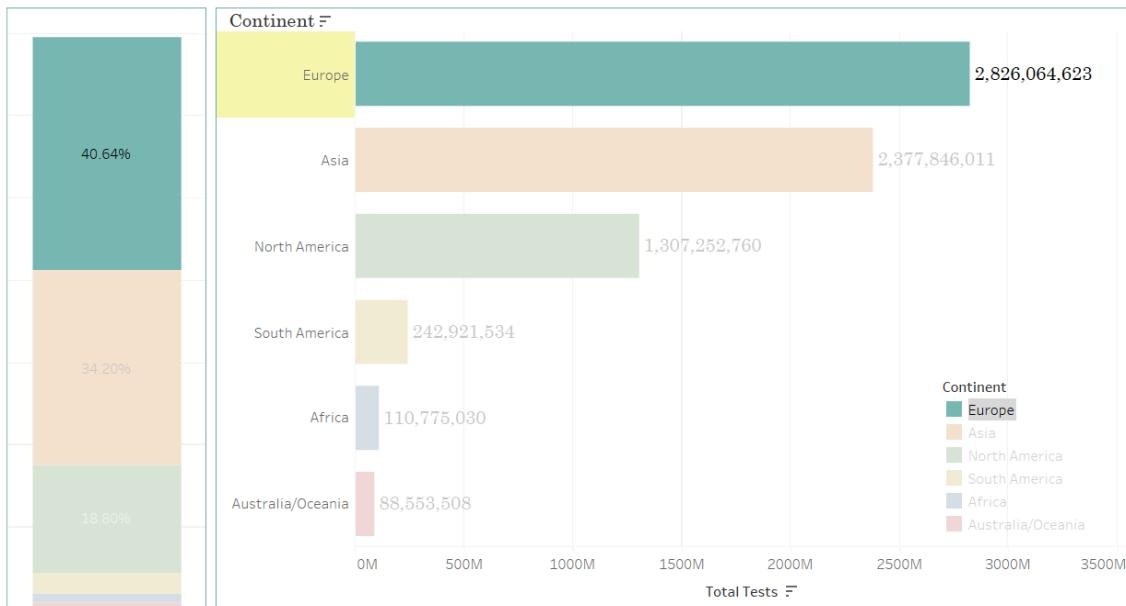
Từ các biểu đồ và bản đồ trên, ta có thể nhận xét:

- Thứ tự thực hiện xét nghiệm từ nhiều nhất đến ít nhất là châu Âu, châu Á, Bắc Mỹ, Nam Mỹ, châu Phi và Úc / Đại Tây Dương.
- Châu Âu chiếm phần lớn số lần thực hiện xét nghiệm phát hiện covid-19 với 40,64% số lần của thế giới, tiếp theo là châu Á chiếm phần trăm nhỏ hơn một ít với 34,20% của thế giới. Ngoài ra các châu lục khác cũng không chiếm quá nhiều.
- Ngoài ra, nỗ lực thực hiện các chiến dịch xét nghiệm phát hiện dịch bệnh giữa các quốc gia không quá lớn dựa trên gam màu sắc của các quốc gia không khác biệt quá nhiều, trừ một số điểm ngoại lệ (ví dụ như Mỹ, Ấn Độ hoặc Nga).

##### ● Nhận xét về dữ liệu:

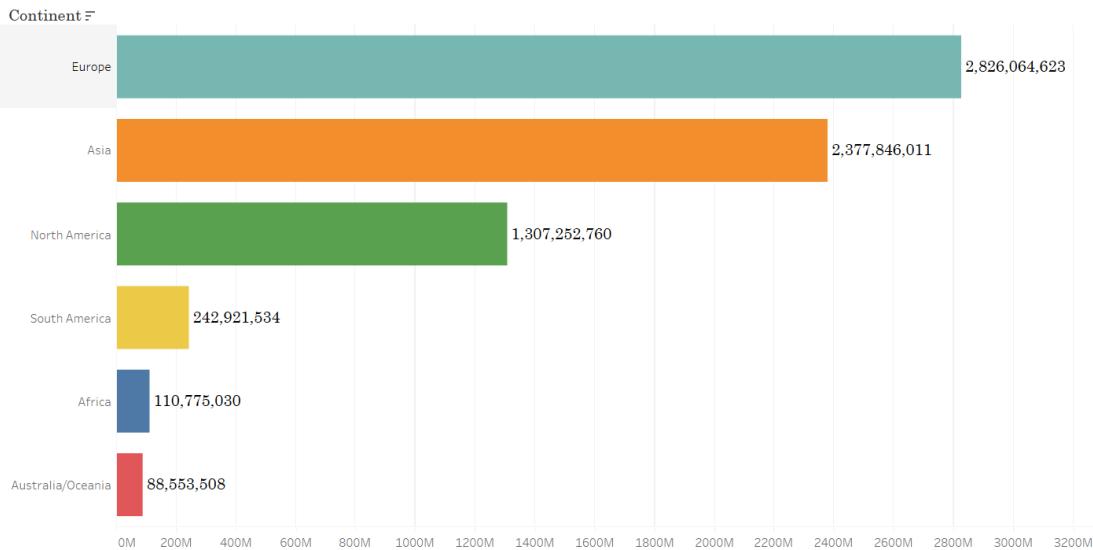
- Thứ tự thực hiện xét nghiệm từ nhiều nhất đến ít nhất là châu Âu, châu Á, Bắc Mỹ, Nam Mỹ, châu Phi và Úc / Đại Tây Dương.
- Châu Âu chiếm phần lớn số lần thực hiện xét nghiệm phát hiện covid-19 với 40,64% số lần của thế giới, tiếp theo là châu Á chiếm phần trăm nhỏ hơn một ít với 34,20% của thế giới. Ngoài ra các châu lục khác cũng không chiếm quá nhiều.

- Ngoài ra, nỗ lực thực hiện các chiến dịch xét nghiệm phát hiện dịch bệnh giữa các quốc gia không quá lớn dựa trên gam màu sắc của các quốc gia không khác biệt quá nhiều, trừ một số điểm ngoại lệ (ví dụ như Mỹ, Ấn Độ hoặc Nga).
- **Nhận xét về tính trực quan:**
  - Biểu đồ Horizontal bar chart ở hàng đầu bên phải dùng để cung cấp các số liệu cho biết tổng số lần thực hiện xét nghiệm của từng châu lục.
  - Ngoài ra, stacked bar-chart và map dùng để cung cấp thêm nhiều thông tin về số % lượt test mà mỗi châu lục chiếm tỉ lệ trên toàn thế giới và nỗ lực thực hiện xét nghiệm trên toàn thế giới (through qua màu sắc trên từng quốc gia)
  - Màu sắc ở 2 biểu đồ trên hàng đầu tiên hỗ trợ liên kết giữa 2 biểu đồ và cũng giúp phân biệt các châu lục khác nhau.
  - Ngoài ra, ở trang story này, ta còn có chức năng hỗ trợ highlight châu lục mà chúng ta muốn xem kỹ



## Các bước thực hiện trực quan:

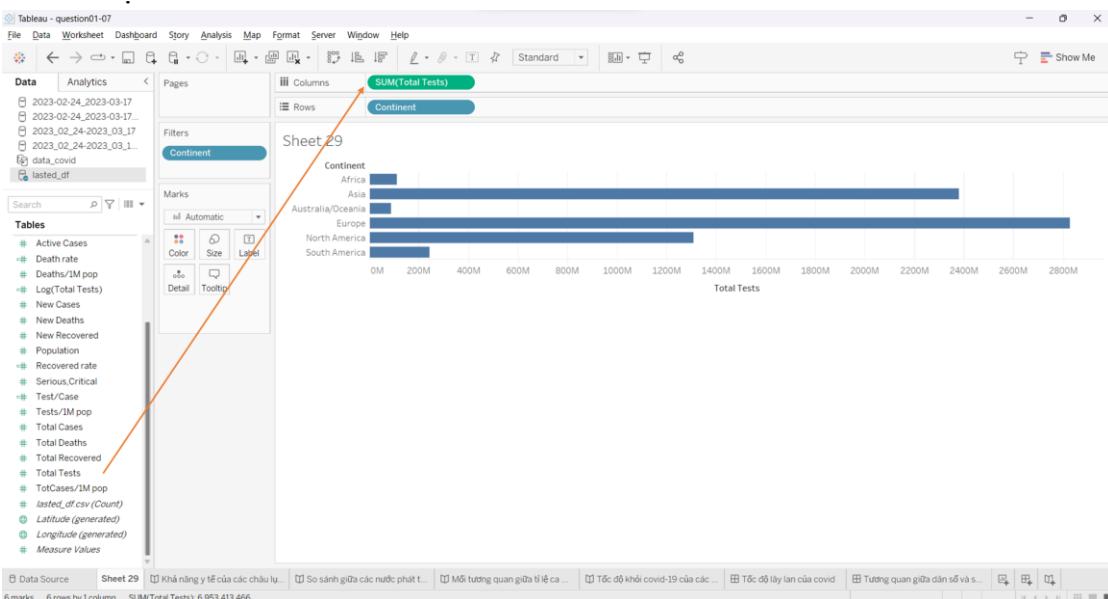
Biểu đồ 1: bar chart



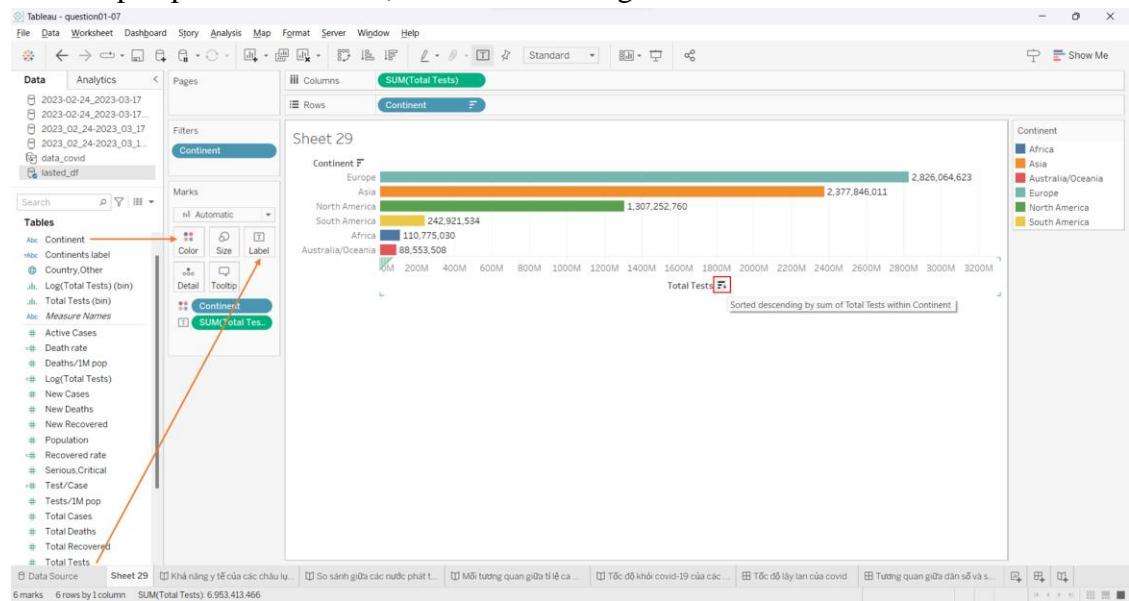
B1: ta sẽ đưa trường dữ liệu **Continent** vào *Rows* trong Tableau và thực hiện exclude các châu lục không hợp lệ (ở đây là All):

The screenshot shows the Tableau interface with the 'Continent' dimension selected in the Rows shelf. A context menu is open over the 'Continent' field, with the 'Exclude' option highlighted by a red arrow. Other options visible in the menu include 'All', 'Keep only', 'Hide', 'Format...', 'Rotate Label', 'Show Header', 'Edit Alias...', and 'Split'.

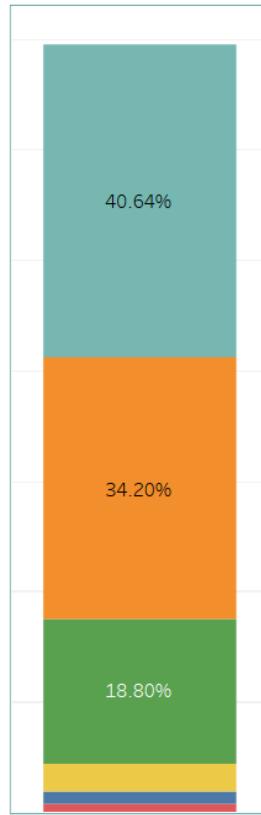
B2: ta đưa thuộc tính **Total Tests** vào Row:



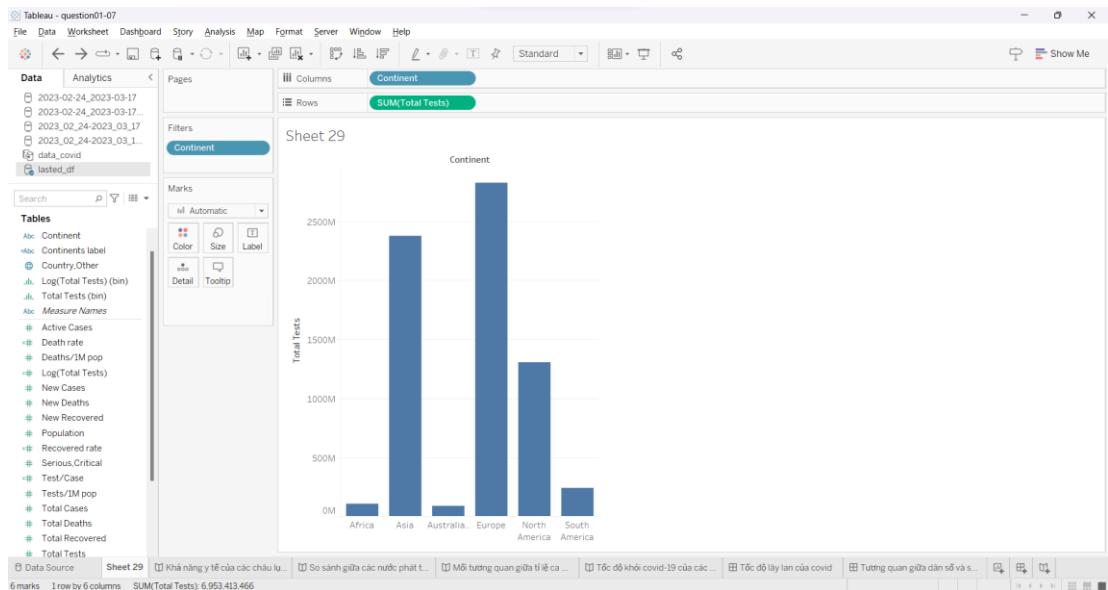
B3: ta tinh chỉnh thêm bằng cách đưa giá trị màu vào các thanh bar, thêm các giá trị số vào mỗi đầu bar và sắp xếp các thanh bar lại cho dễ nhìn bằng các thao tác mô tả như bên dưới:



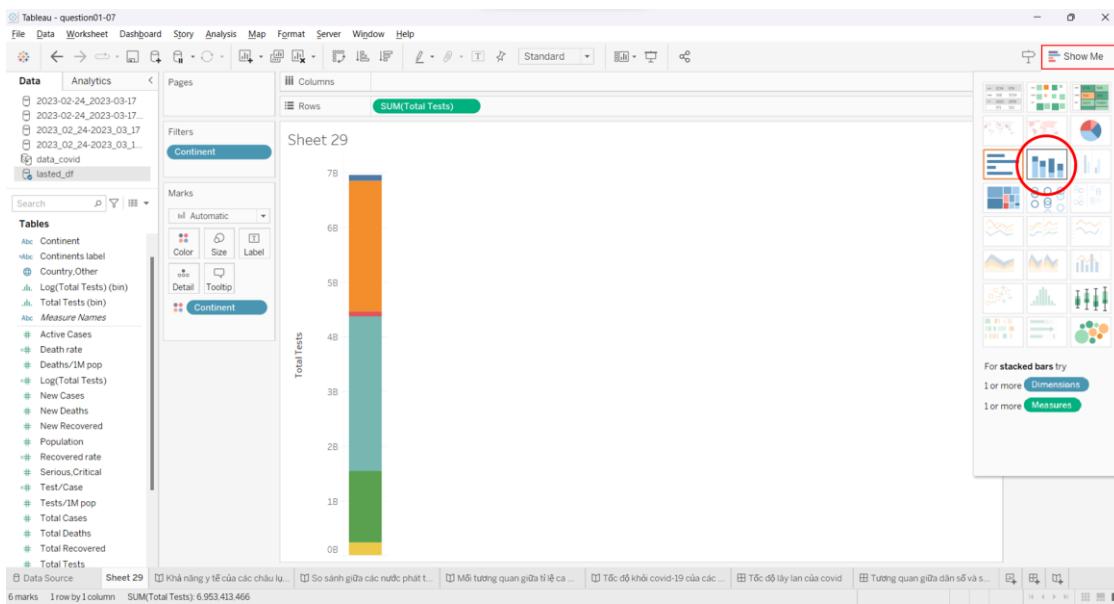
Biểu đồ 2: Stacked bar chart



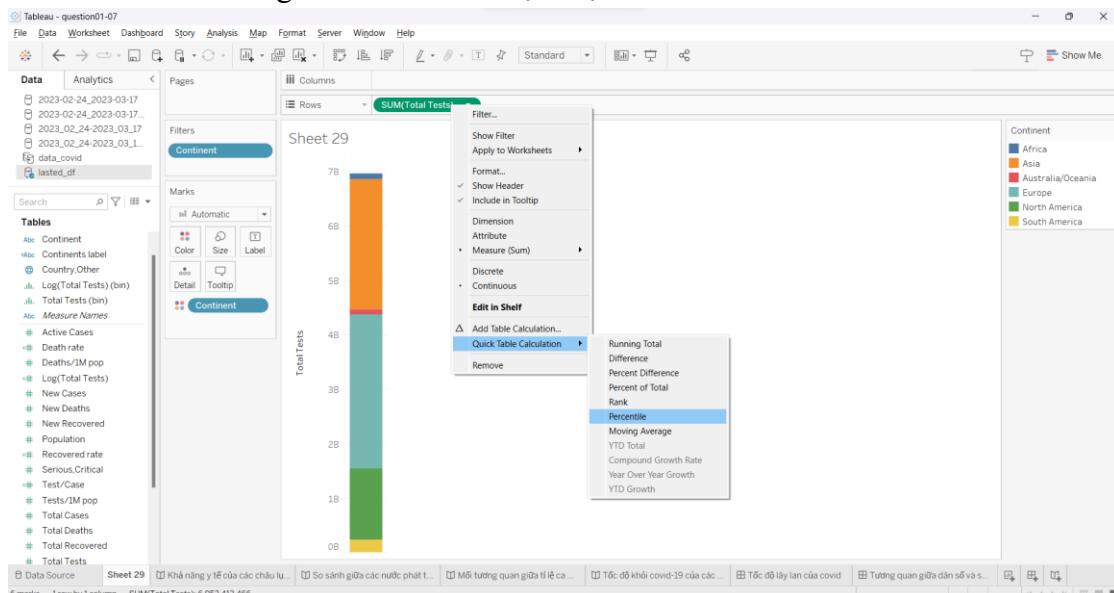
B1, B2: Ta cũng sẽ thực hiện tương tự như Bước 1, 2 ở trên nhưng sẽ là đảo ngược nhau, tức là trường **Continent** → Column, **Total Tests** → Row, và cũng thực hiện loại bỏ các châu lục không hợp lệ:



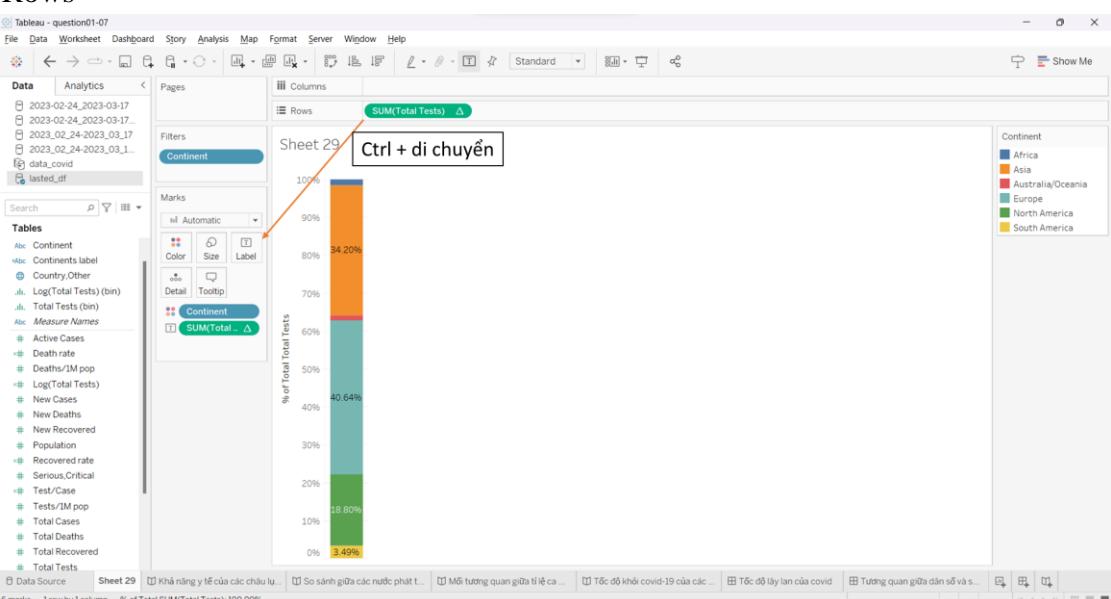
B3: ta click vào nút Show Me trên cùng bên phải màn hình và chọn vào biểu đồ stacked bar chart



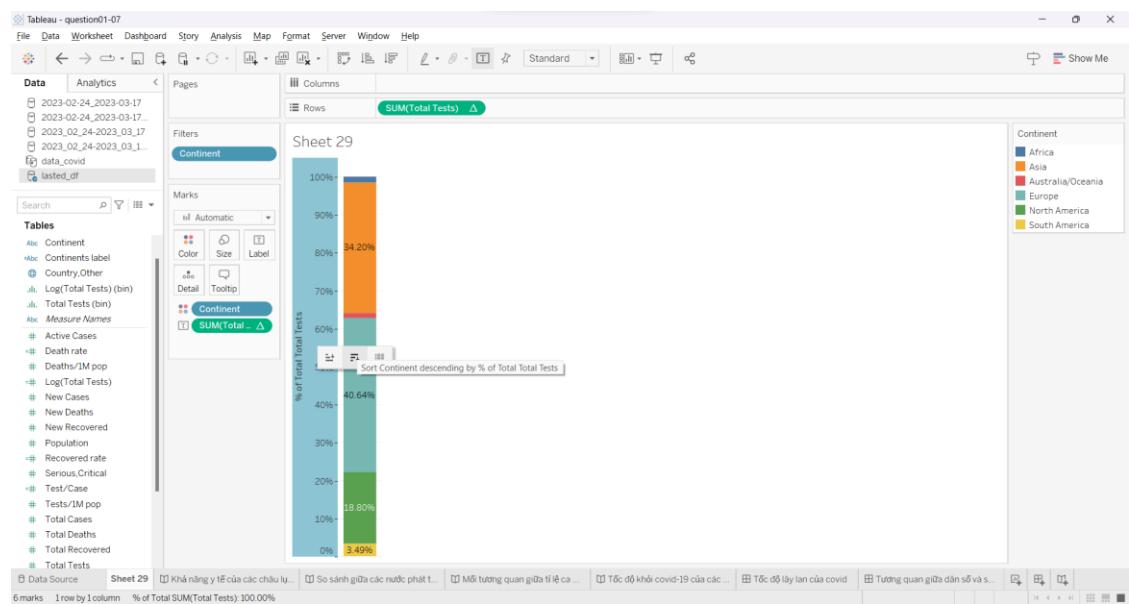
B4: ta chuyển giá trị trong bar chart thành giá trị % trên toàn bộ bằng cách nhấp phải chuột vào thuộc tính **Total tests** đang ở trên Row và chọn mục như hình bên dưới:



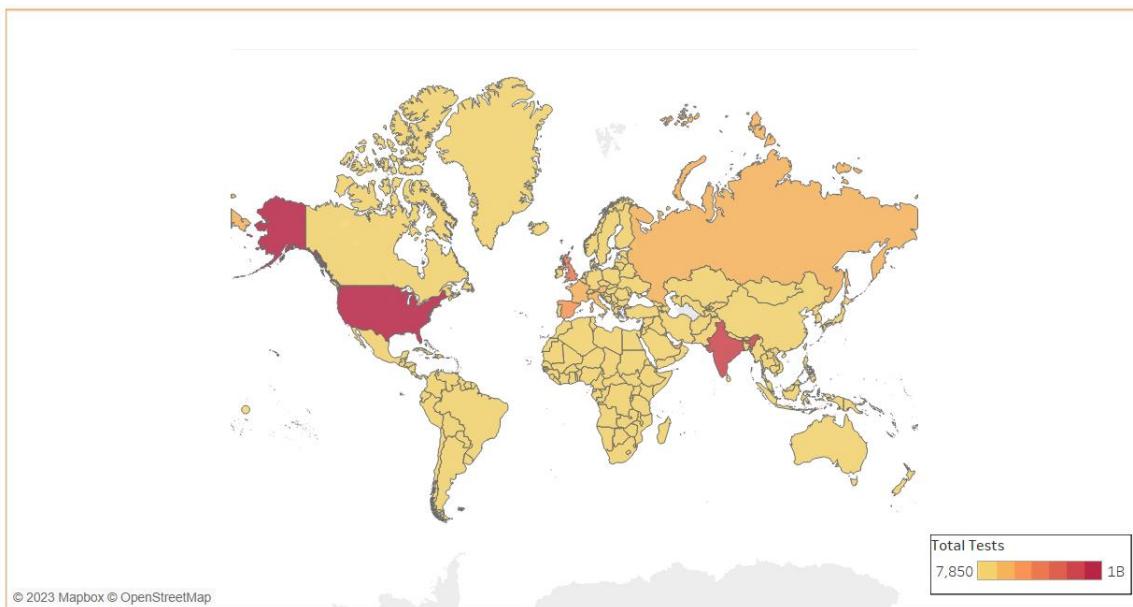
B5: ta thêm số % vào mỗi thanh bằng cách giữ Ctrl, di chuyển trường **Total tests** đang ở trên Rows



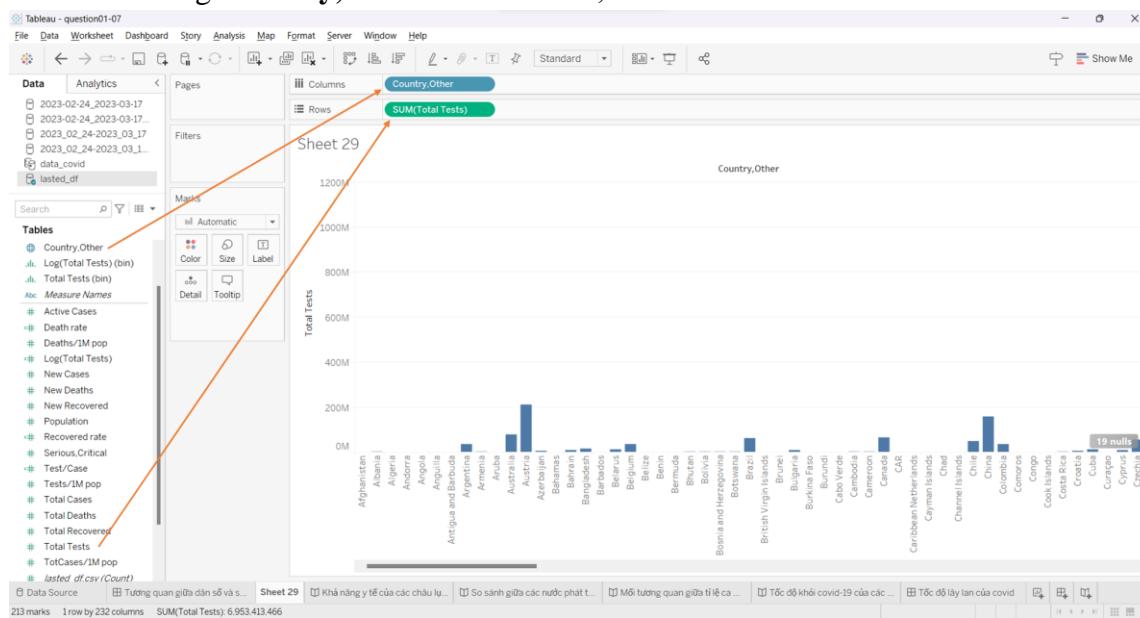
B6: cuối cùng, ta chỉ cần sắp xếp các thứ tự thanh bar bằng cách click chuột vào header trục y và nhấn nút như hình dưới là hoàn thành:



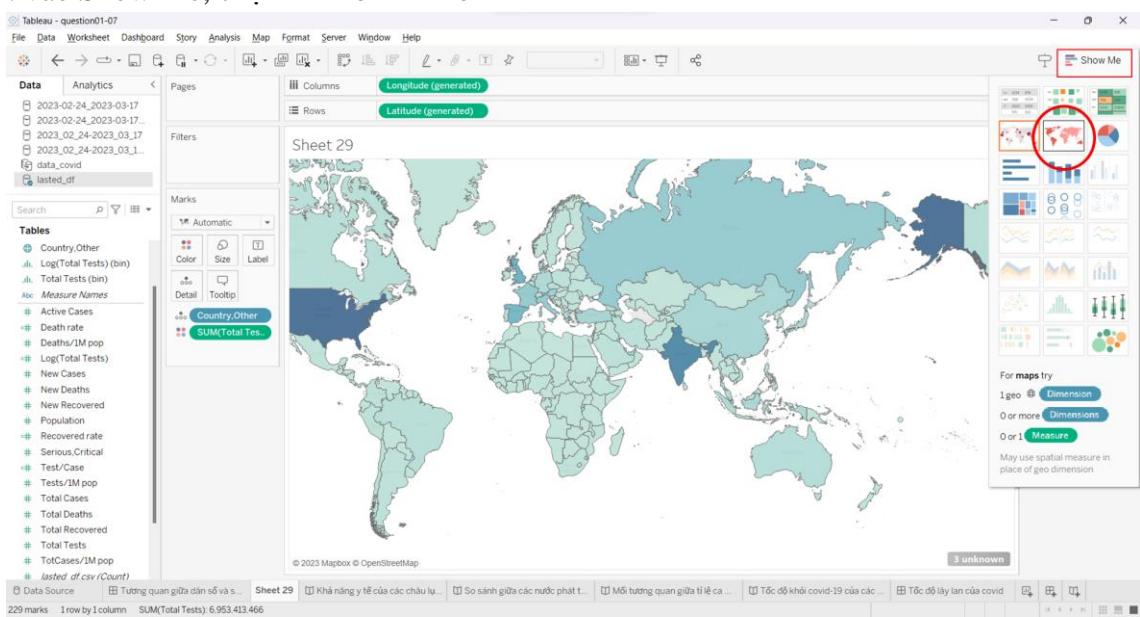
Biểu đồ 3: biểu đồ bản đồ (map)



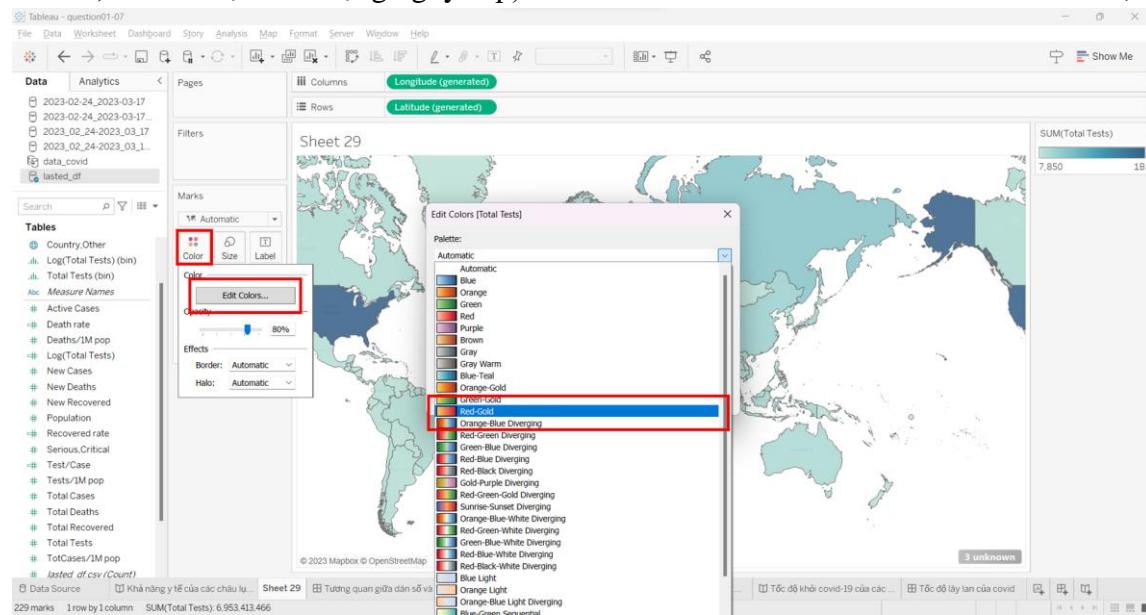
B1: kéo thả trường Country, Other → Columns; Total Tests → Rows



B2: vào Show Me, chọn hình biểu đồ bản đồ

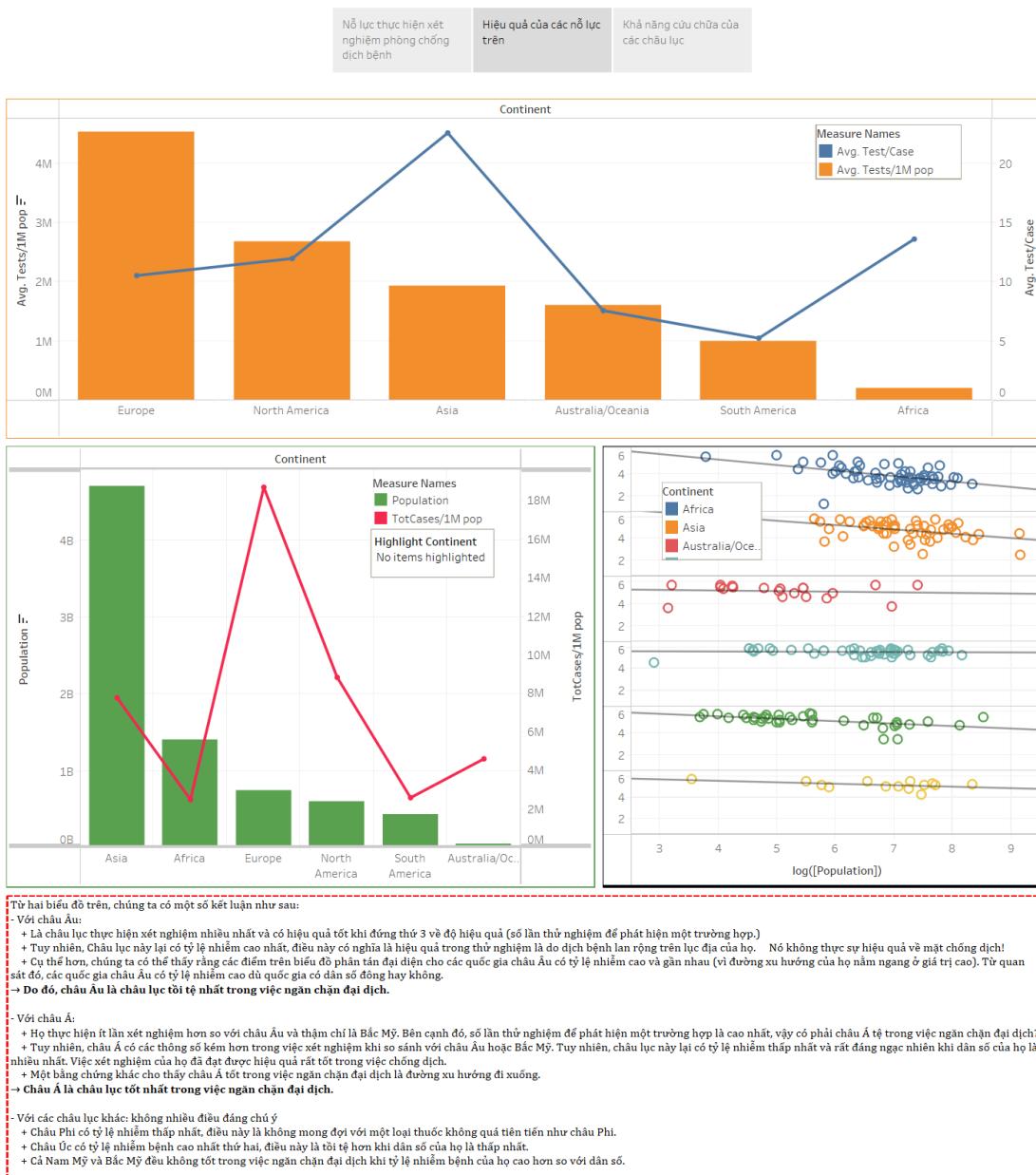


B3: ta thực hiện chọn lại màu sắc cho phù hợp với ngữ cảnh là dịch bệnh (nên sử dụng các tông màu **cam**, **đỏ** thể hiện tình trạng nguy cấp). Ta vào Color → Edit Color → Palette → Red, Gold



## b. Hiệu quả của việc xét nghiệm phát hiện bệnh:

Khả năng y tế của các châu lục trên thế giới



### • Nhận xét dữ liệu:

Với châu Âu:

- Là châu lục thực hiện xét nghiệm nhiều nhất và có hiệu quả tốt khi đứng thứ 3 về độ hiệu quả (số lần thử nghiệm để phát hiện một trường hợp.)
- Tuy nhiên, Châu lục này lại có tỷ lệ nhiễm cao nhất, điều này có nghĩa là hiệu quả trong thử nghiệm là do dịch bệnh lan rộng trên lục địa của họ. Nó không thực sự hiệu quả về mặt chống dịch!
- Cụ thể hơn, chúng ta có thể thấy rằng các điểm trên biểu đồ phản ánh đại diện cho các quốc gia châu Âu có tỷ lệ nhiễm cao và gần nhau (vì đường xu hướng của họ nằm ngang ở giá trị cao). Từ quan sát đó, các quốc gia châu Âu có tỷ lệ nhiễm cao dù quốc gia có dân số đông hay không.

→ Do đó, châu Âu là châu lục tồi tệ nhất trong việc ngăn chặn đại dịch.

Với châu Á:

- Họ thực hiện ít lần xét nghiệm hơn so với châu Âu và thậm chí là Bắc Mỹ. Bên cạnh đó, số lần thử nghiệm để phát hiện một trường hợp là cao nhất, vậy có phải châu Á tệ trong việc ngăn chặn đại dịch?
- Tuy nhiên, châu Á có các thông số kém hơn trong việc xét nghiệm khi so sánh với châu Âu hoặc Bắc Mỹ. Tuy nhiên, châu lục này lại có tỷ lệ nhiễm thấp nhất và rất đáng ngạc nhiên khi dân số của họ là nhiều nhất. Việc xét nghiệm của họ đã đạt được hiệu quả rất tốt trong việc chống dịch.
- Một bằng chứng khác cho thấy châu Á tốt trong việc ngăn chặn đại dịch là đường xu hướng đi xuống.

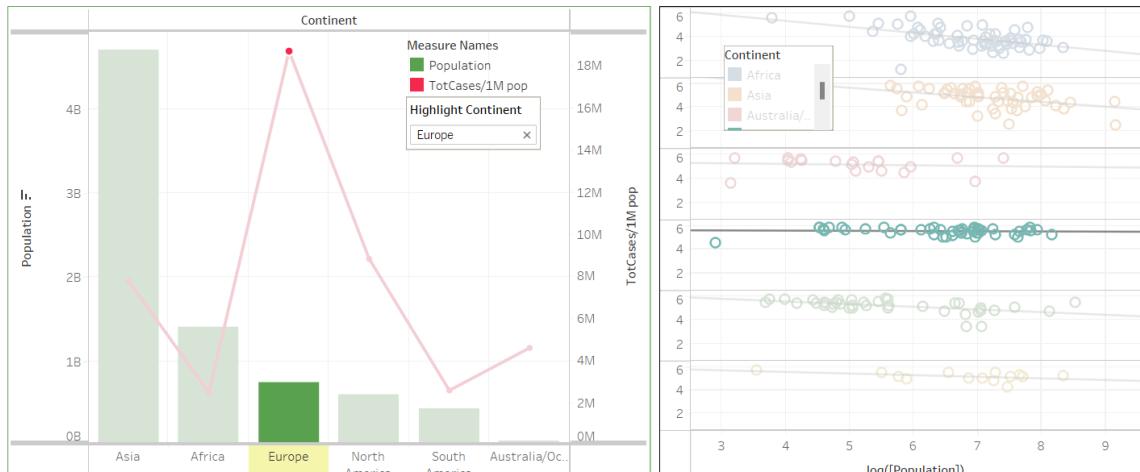
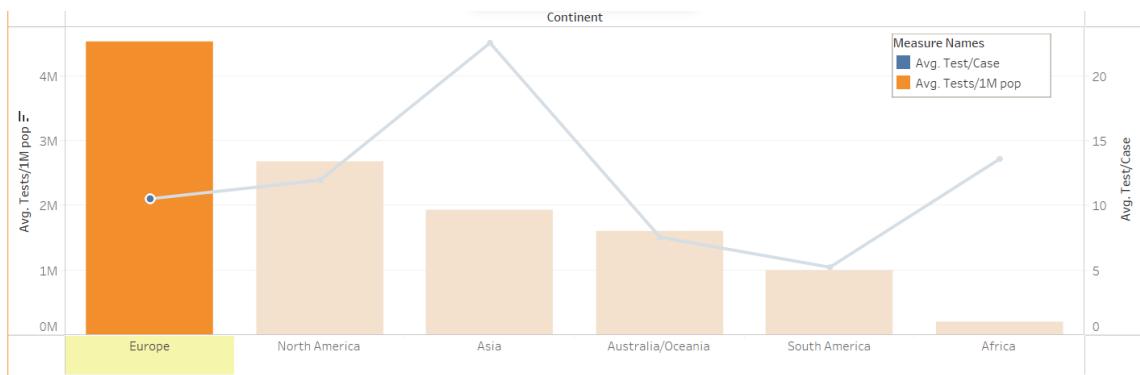
→ **Châu Á là châu lục tốt nhất trong việc ngăn chặn đại dịch.**

Với các châu lục khác: không nhiều điều đáng chú ý

- Châu Phi có tỷ lệ nhiễm thấp nhất, điều này là không mong đợi với một loại thuốc không quá tiên tiến như châu Phi.
- Châu Úc có tỷ lệ nhiễm bệnh cao nhất thứ hai, điều này là tồi tệ hơn khi dân số của họ là thấp nhất.
- Cá Nam Mỹ và Bắc Mỹ đều không tốt trong việc ngăn chặn đại dịch khi tỷ lệ nhiễm bệnh của họ cao hơn so với dân số.

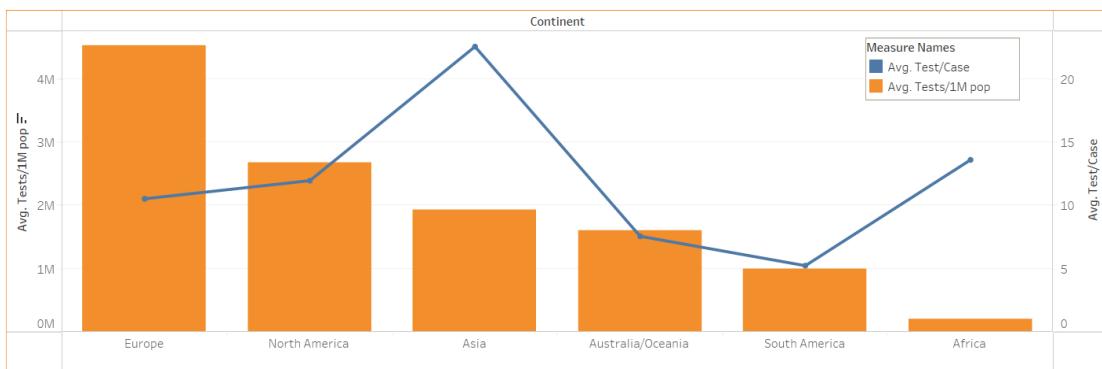
#### • Nhận xét trực quan:

- Ở bài này ta cũng sẽ sử dụng 3 biểu đồ để trả lời cho câu hỏi của chúng ta, 2 biểu đồ bar + line chart và một scatterplot.
- Ở biểu đồ bar + line chart đầu tiên ta so sánh **trung bình một triệu người của một châu lục được test bao nhiêu lần** (bar chart) và sự hiệu quả của nó bằng giá trị **số lần test để phát hiện ra một ca nhiễm bệnh** (line chart). Ta kết hợp 2 biểu đồ này với lý do là muốn so sánh song song cả 2 yếu tố trên.
- Tuy nhiên, với biểu đồ đầu tiên ta chưa có đủ góc nhìn về sự hiệu quả đó (ví dụ: nếu tỉ lệ test/case cao có thể có 2 nguyên nhân là **khả năng xét nghiệm phát hiện bệnh hiệu quả cao** hoặc có thể là **dịch bệnh đã tràn lan dẫn tới việc phát hiện ra bệnh nhân nhiễm bệnh cũng dễ dàng**). Vì vậy với biểu đồ thứ 2 ta so sánh song song 2 thông tin, tổng dân số của từng châu lục (bar chart) và tỉ lệ nhiễm bệnh của châu lục đó (trường dữ liệu **Total Cases/1M pop**, line chart)
- Cuối cùng là scatter plot, biểu đồ này ứng dụng kĩ thuật **Partition into Views** trong bài **Facet**. Biểu đồ này giúp ta quan sát rõ hơn mối liên hệ giữa **Population vs Total Cases/1M pop** của các quốc gia trong châu lục đó liệu có điều gì đặc biệt không?
- Ngoài ra, còn cung cấp khả năng **highlighter** để ta dễ dàng quan sát và thuận tiện trong việc phân tích chi tiết từng châu lục bằng cách chọn vào khung highlighter và chọn châu lục bạn muốn highlighter (hình ở dưới là minh họa highlight cho châu Âu)

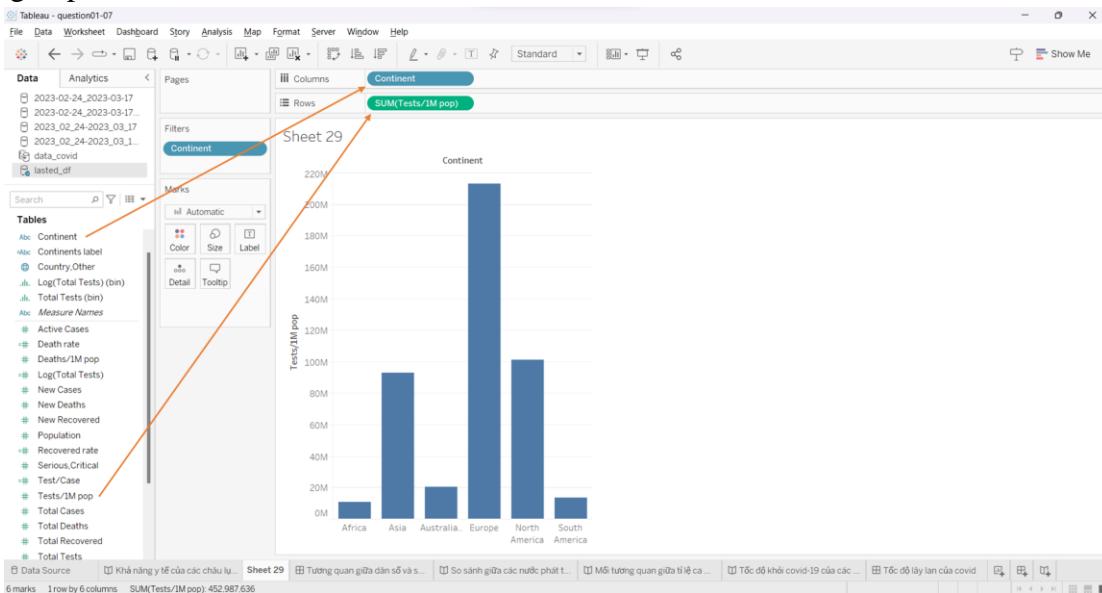


## Các bước thực hiện trực quan:

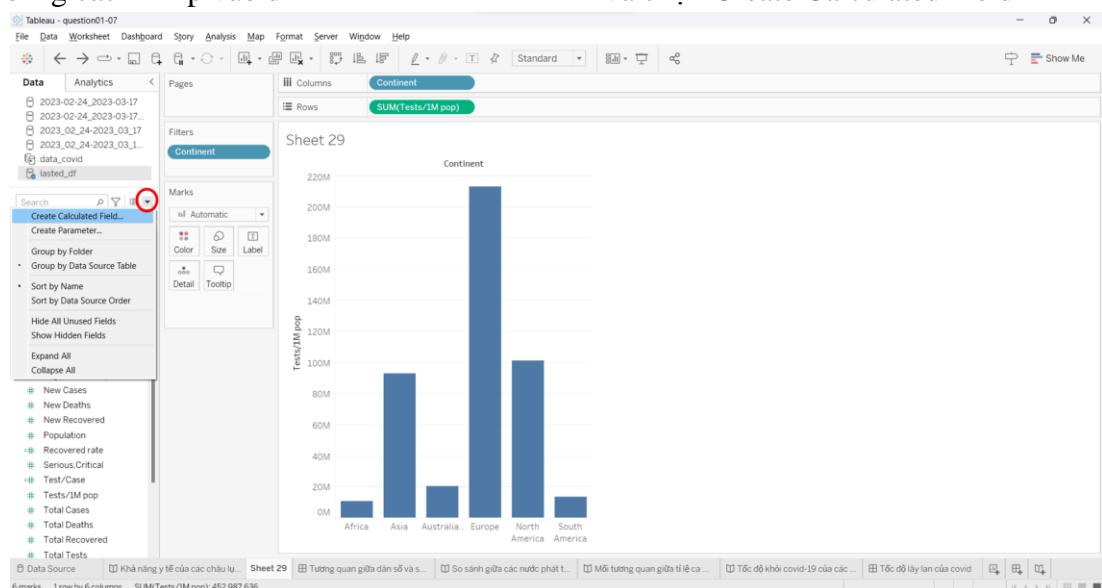
Biểu đồ 1: bar + line chart trực quan số lượt test trên 1 triệu người và số lượt test để phát hiện 1 ca nhiễm bệnh:



B1: Ta kéo thả **Continent** → Columns, **Tests/1M pop** → Rows, nhớ là ta vẫn sẽ lọc các châu lục không hợp lệ:



B2: vì bộ dữ liệu ban đầu không có trường dữ liệu **Test/Case**, nên ta sẽ phải tạo trường dữ liệu mới bằng cách nhập vào dấu mũi tên như bên dưới và chọn Create Calculated Field



B3: ta nhập code để tạo ra trường dữ liệu mới từ khung Calculated Field

Tên trường dữ liệu mới

[Test/Case]

[Total Tests]/[Total Cases]

Phép tính từ trường dữ liệu cũ tạo ra trường dữ liệu mới

The calculation is valid.

3 Dependencies

OK

All

Search

ABS

ACOS

AND

AREA

ASCII

ASIN

ATAN

ATAN2

ATTR

Avg

BUFFER

CASE

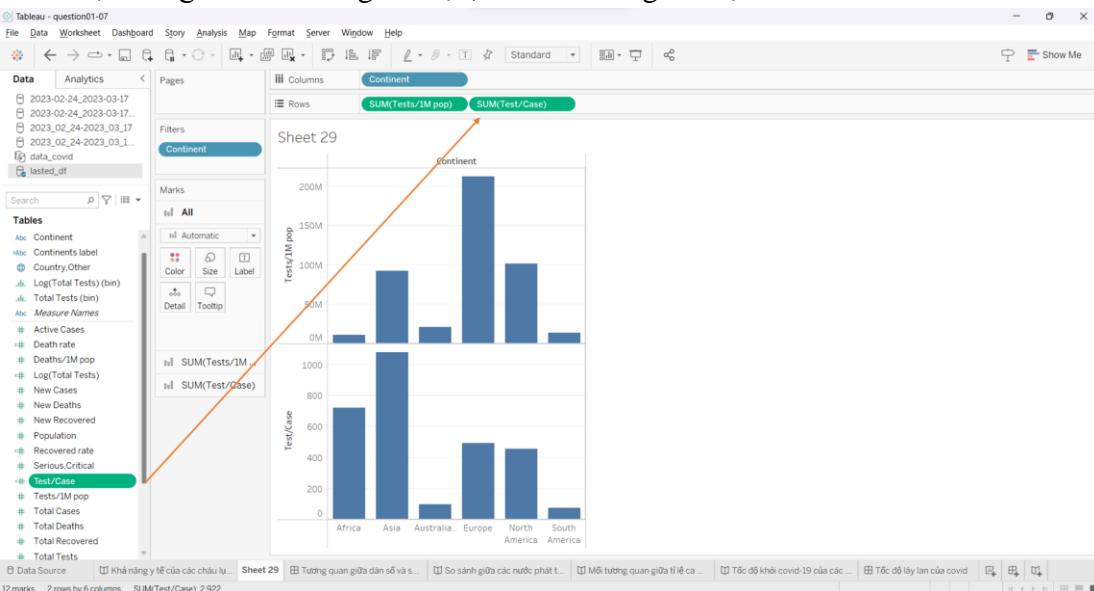
ABS (number)

Returns the absolute value of the given number.

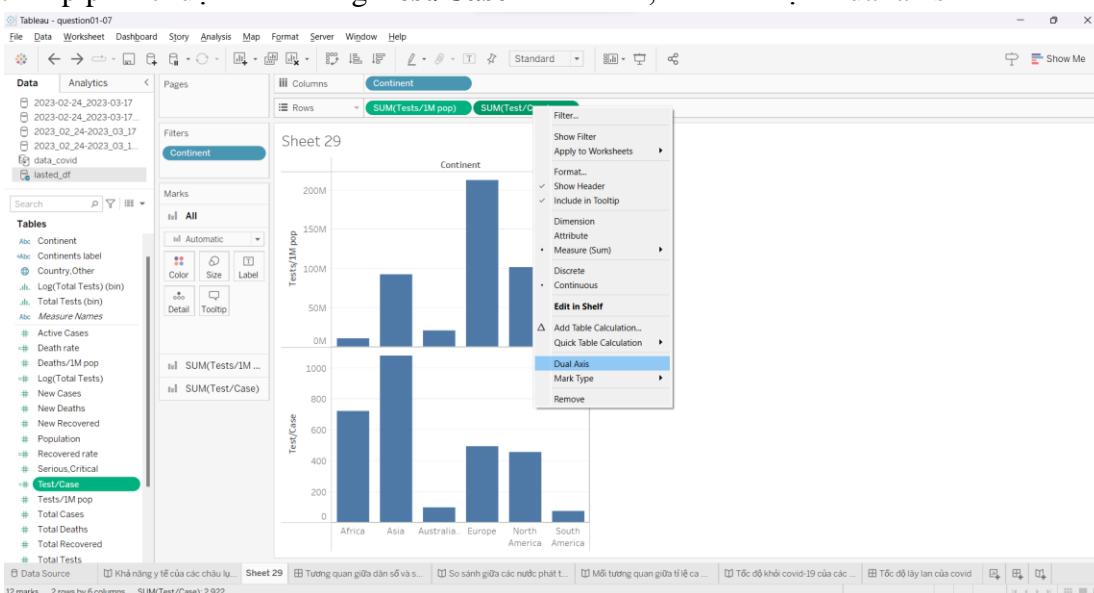
Example: ABS(-7) = 7

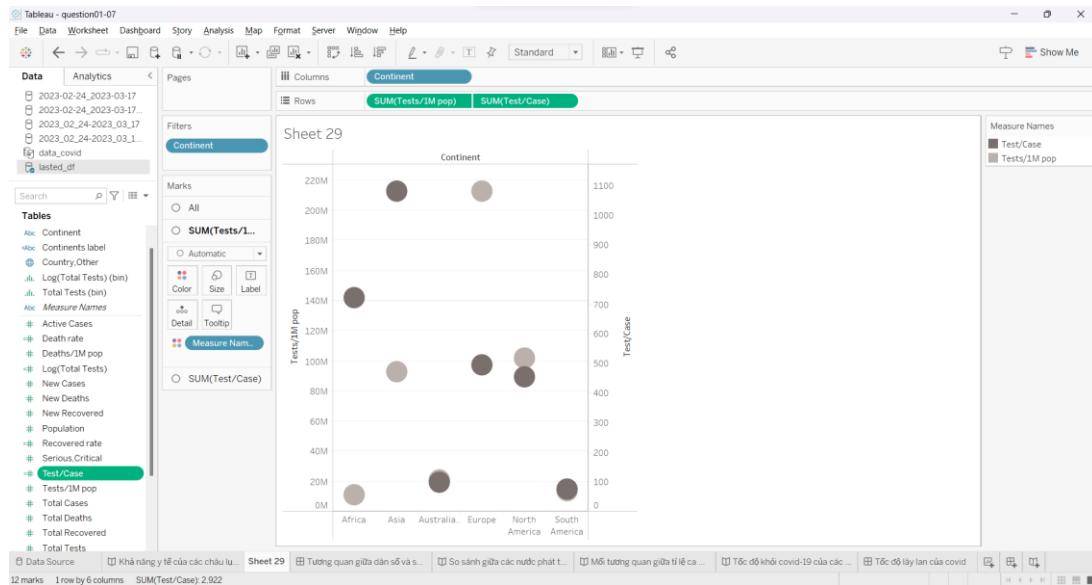
Một số hàm mà Tableau hỗ trợ

#### B4: sau khi tạo xong ta vào trường dữ liệu, ta kéo trường dữ liệu Test/Case → Rows

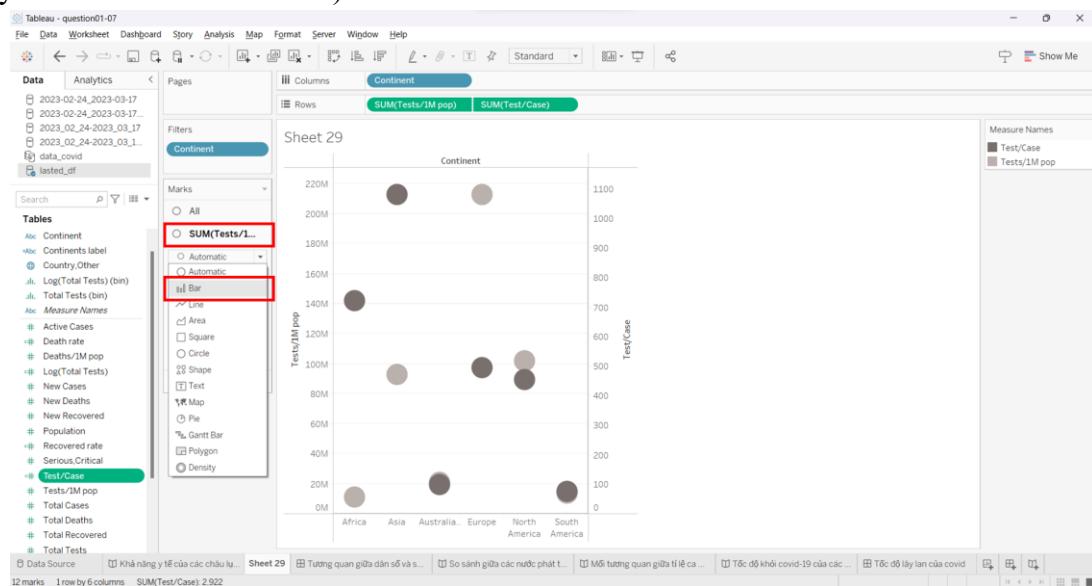


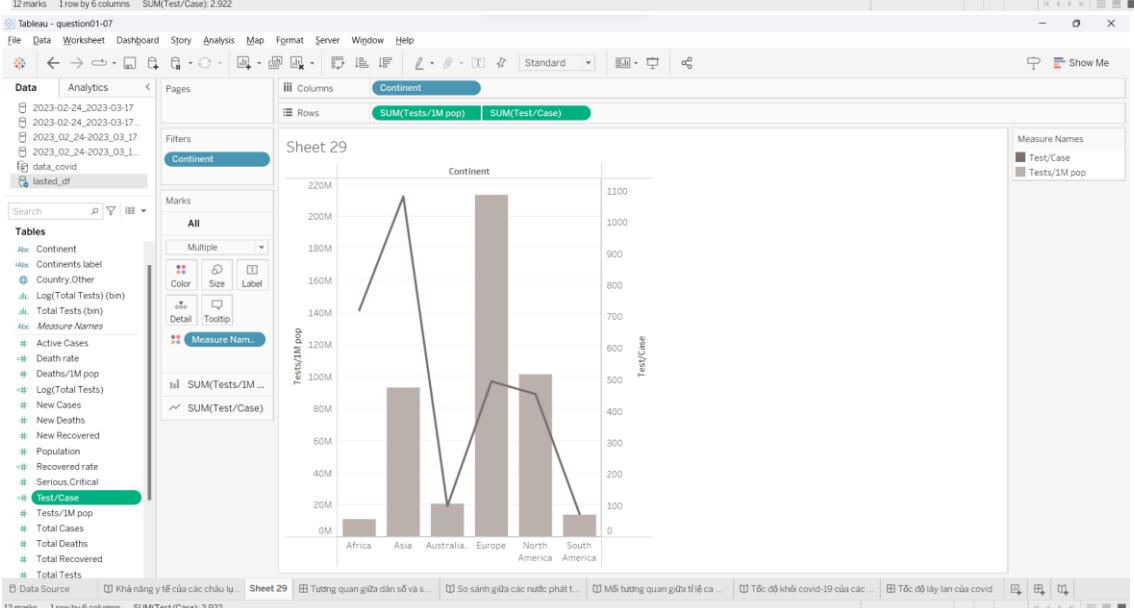
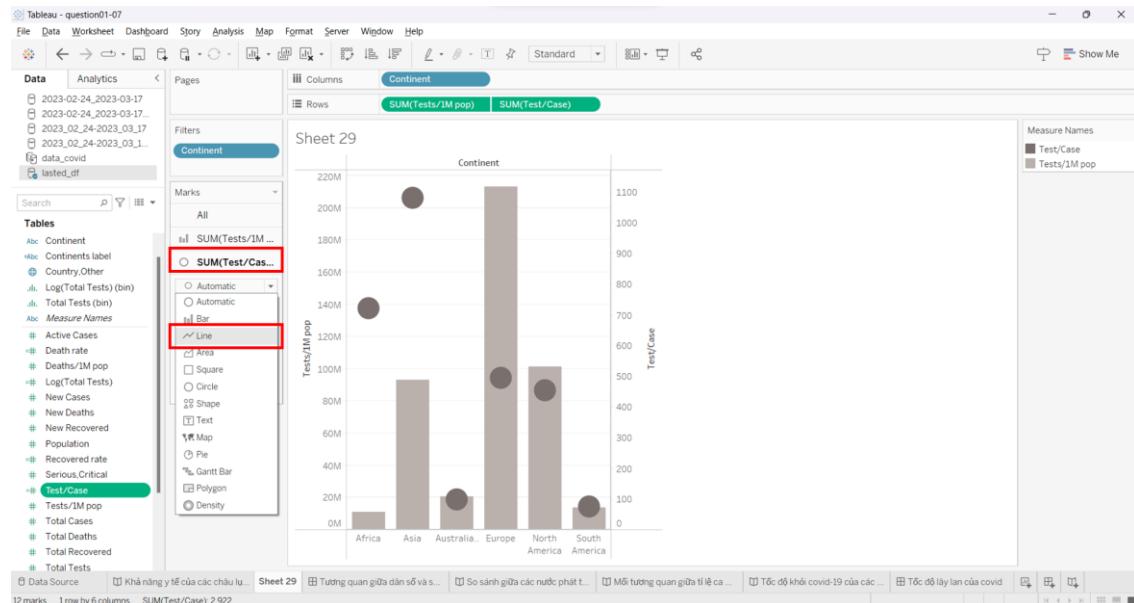
#### B5: ta nhấp phải chuột vào trường Test/Case trên Rows, sau đó chọn Dual axis





B6: ta vào từng vùng như hình ảnh ở bên dưới và chọn lại dạng biểu đồ mà chúng ta mong muốn (ở đây là bar chart và line chart)





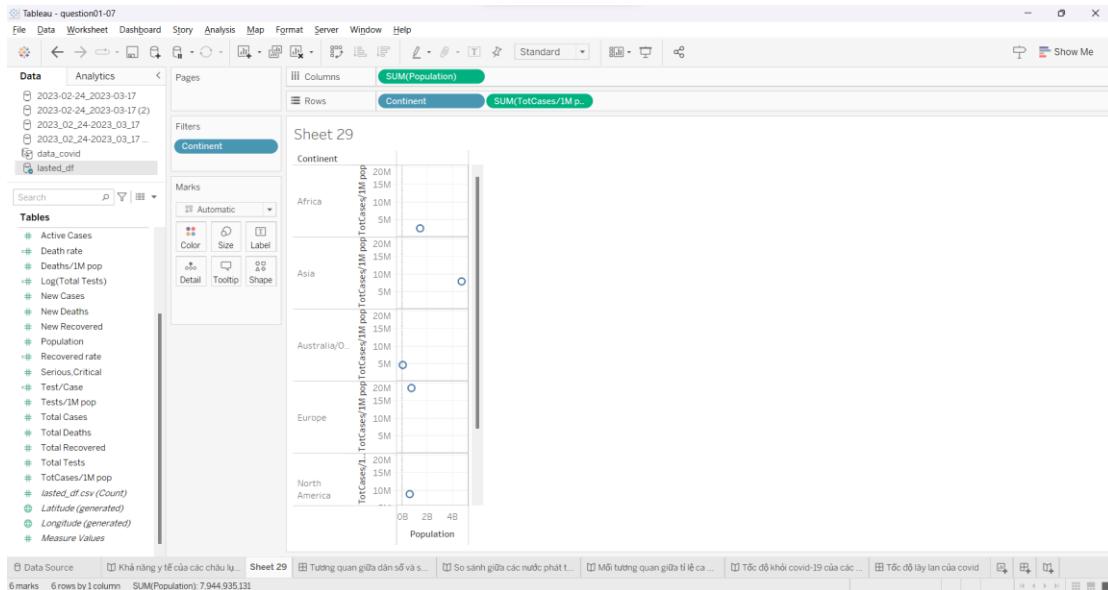
B6: Ta tinh chỉnh lại màu sắc và sắp xếp lại các thanh bar là hoàn thành (bước này đã trình bày ở trên nên không trình bày lại nữa)

Biểu đồ 2: biểu đồ line + bar chart trực quan **Population** và **Total Cases/ 1M pop**

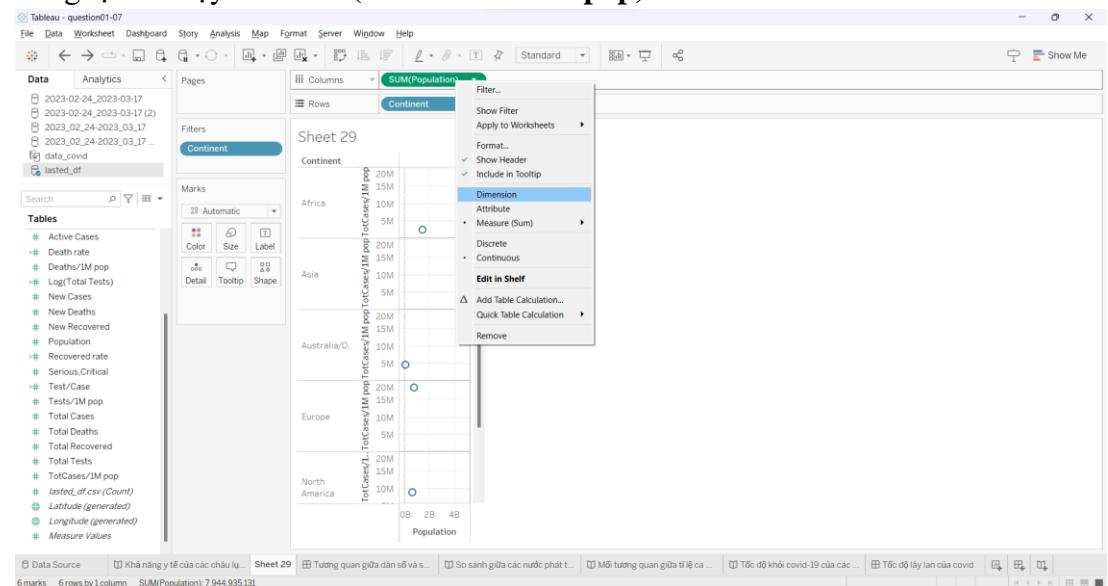
Ta cũng sẽ thực hiện các bước trực quan như vừa trình bày ở trên nhưng đơn giản hơn vì ta sẽ không cần tạo ra thêm trường dữ liệu mới. Vì vậy chúng ta sẽ không trình bày lại nữa.

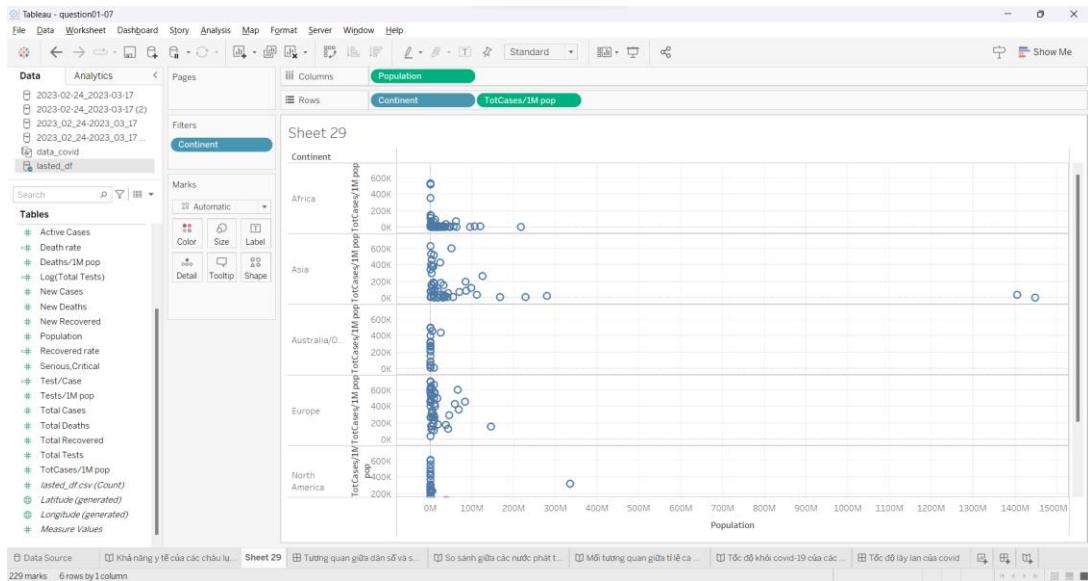
Biểu đồ 3: các scatterplot thể hiện mối quan hệ giữa **Population vs Total Cases/1M pop** của từng châu lục

B1: kéo thả **Total Cases/1M pop, Continent** → Rows, lọc châu lục hợp lệ; **Population** → Columns



B2: nhấp chuột phải vào trường dữ liệu **Sum(Population)** ở hàng Columns, chọn Dimension, ta làm tương tự như vậy cho **Sum(Total Cases/1M pop)**





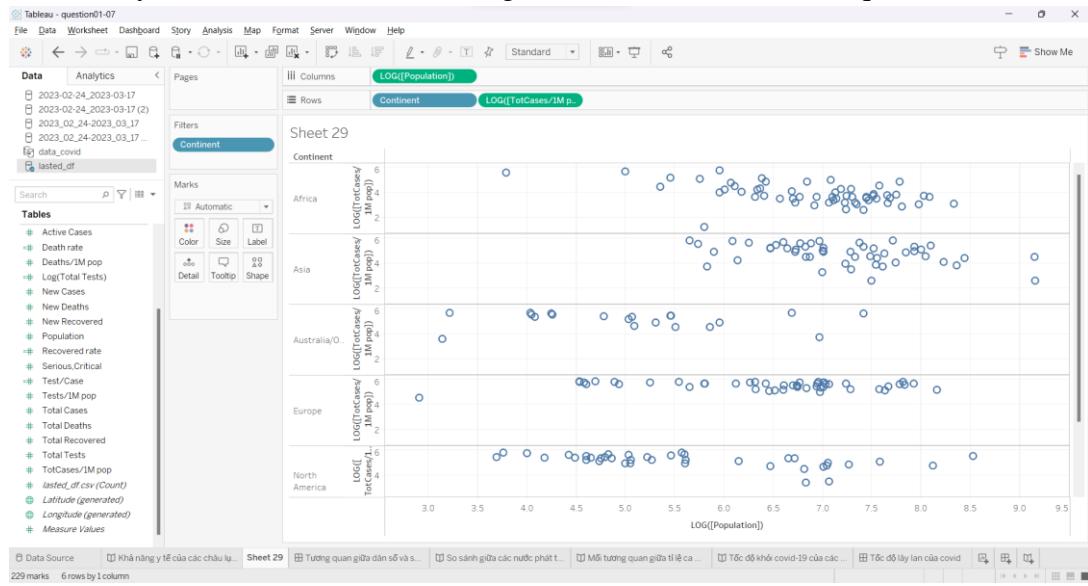
B3: bây giờ ta đã có scatterplot, nhưng nếu nhìn kĩ thì ta nhận ra các scatterplot này tụ lại với nhau và không nhìn thấy xu hướng của chúng, để giải quyết vấn đề này ta sẽ thực hiện rã cụm chúng bằng cách lấy **LOG(Population)**, **LOG(Total Cases/1M pop)**

Đầu tiên, ta nhấp đúp chuột và thuộc tính **Population** ở trên Columns, chỉnh lại thành

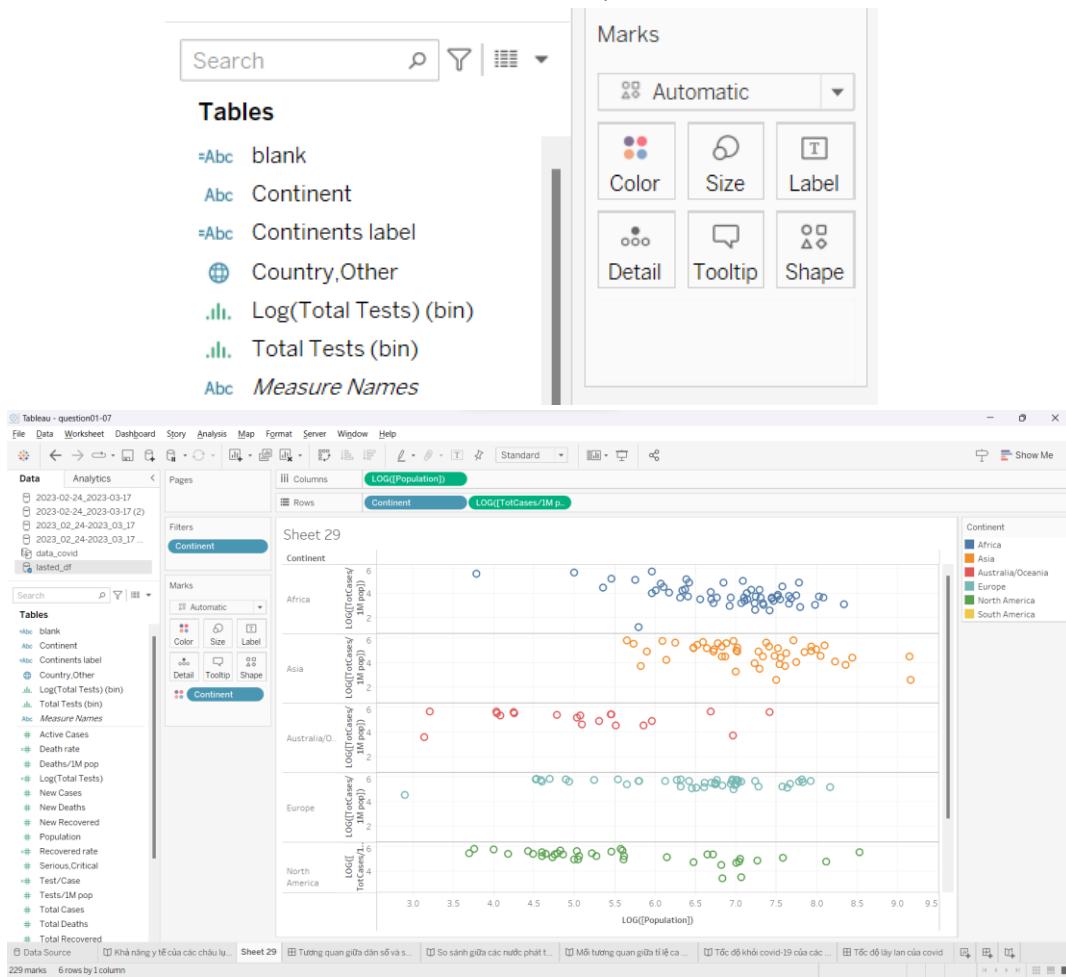
**LOG([Population])**, sau đó nhấn Enter. Ta cũng làm tương tự như vậy với **Total Cases/1M pop**.

Sau khi xong, các trường dữ liệu có thể bị tự động lấy **Sum**

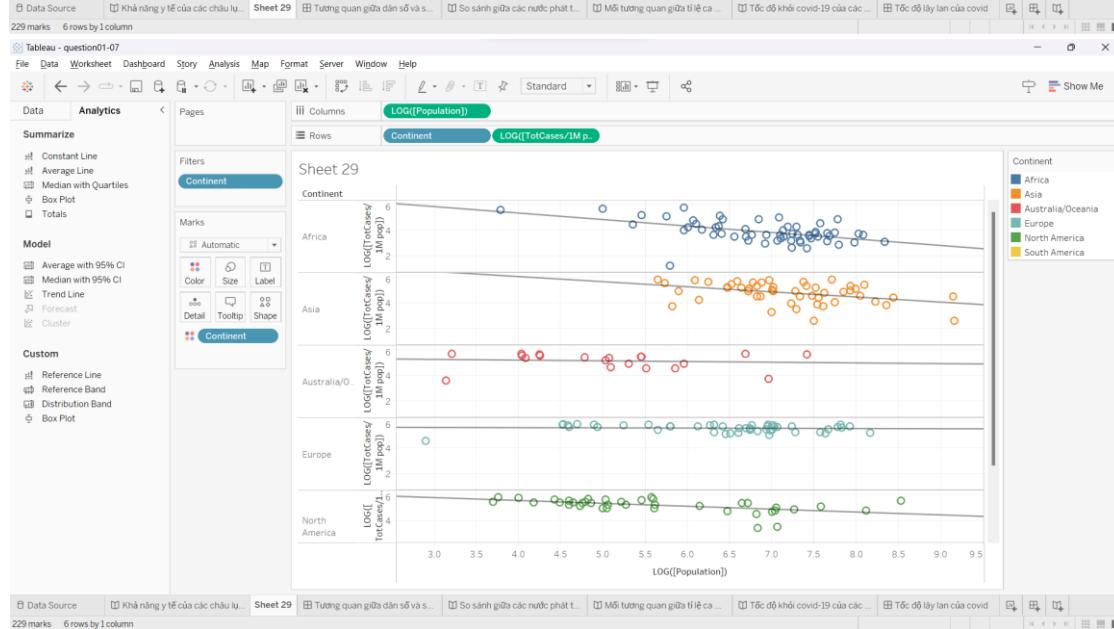
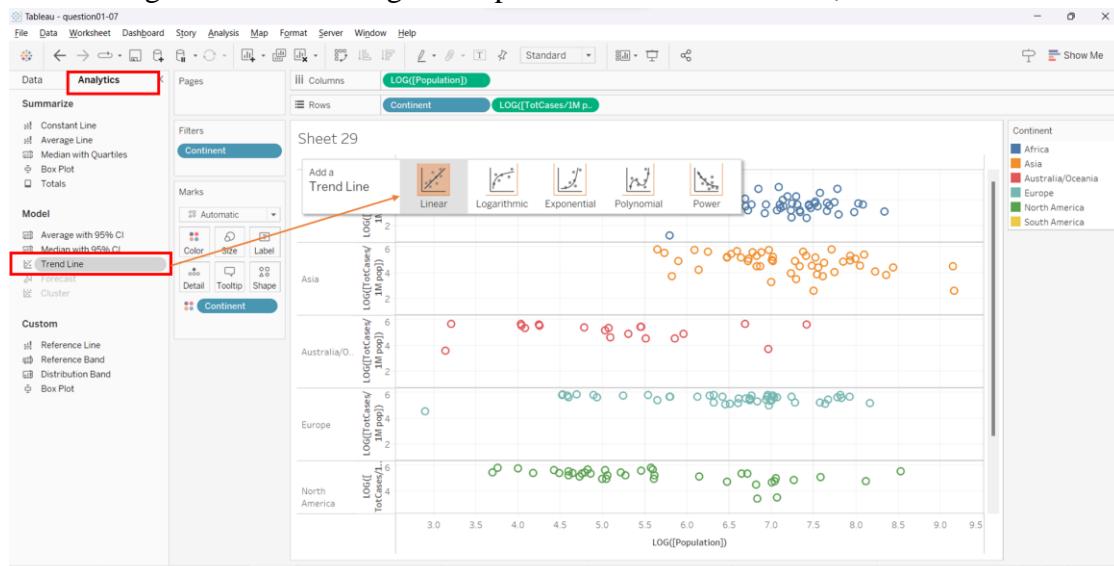
Ta chỉ cần chuyển về Dimension như hướng dẫn ở trên thì ta sẽ có kết quả



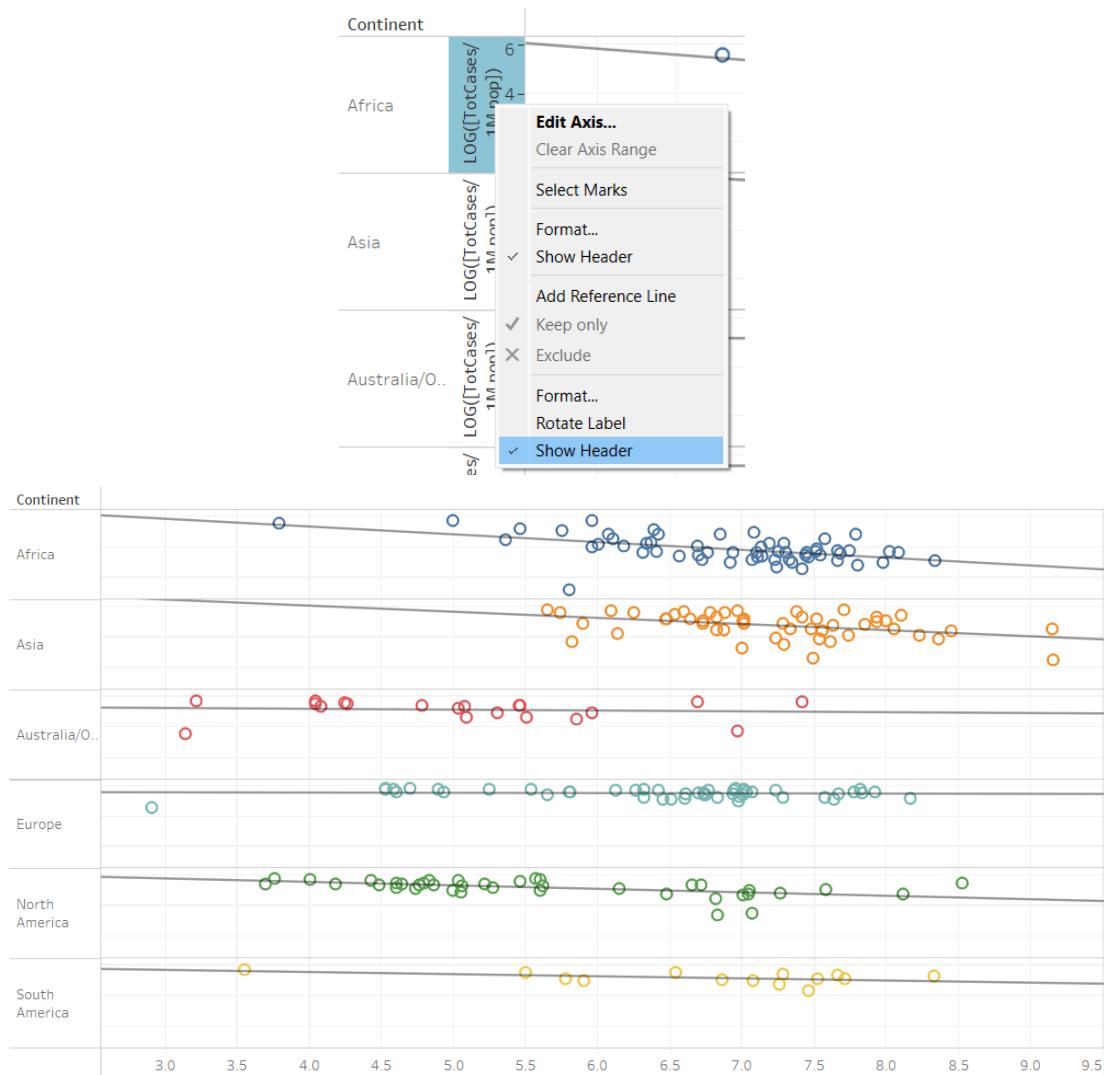
B6: ta kéo Continent vào color để tô màu cho dữ liệu:



B7: ta vẽ đường trendline cho từng scatterplot theo các bước minh họa ở bên dưới



B8: ta xóa bỏ các header lặp liên tục bằng cách click phải chuột vào header, bỏ tích Show Header là hoàn thành



c. Khả năng chữa bệnh của các Châu lục

## Khả năng y tế của các châu lục trên thế giới



Kết luận:

- Đối với tất cả các châu lục, phần lớn các trường hợp đều đã hồi phục được chứng minh bằng màu đỏ chiếm phần lớn thanh bar trong biểu đồ stacked bar chart. Đây là điều rất tích cực trong tình hình đại dịch đang lây lan nhanh.
- Chúng ta chỉ thấy một phần nhỏ màu xanh lá cây của thanh bar châu Âu, châu Á và Bắc Mỹ, có nghĩa là đại dịch có thể vẫn còn đang diễn ra trên các lục địa này.
- Cả tỷ lệ phục hồi và tỷ lệ tử vong đều rất thấp, chứng tỏ khả năng chữa bệnh của toàn thế giới là đáng kinh ngạc.
- VỚI CHÂU ÂU:
  - + Mặc dù không tốt trong việc ngăn chặn đại dịch, các nước châu Âu lại tốt trong việc chữa trị, được chứng minh từ biểu đồ trên khi có số trường hợp hồi phục cao nhất cùng với tỷ lệ hồi phục trên 95%.
  - + Tuy nhiên, châu Âu cũng có số lượng ca tử vong cao nhất.
- VỚI CHÂU Á:
  - + Sau đó là châu Á, đứng thứ hai cả về tổng số ca hồi phục và tỷ lệ hồi phục.
  - + Mặc dù có số lượng ca tử vong lớn nhưng đây là điều tốt nhất mà họ có thể làm với lục địa có dân số đông nhất.
- VỚI CÁC CHÂU LỤC KHÁC:
  - + Ngạc nhiên là châu Phi đứng thứ ba về tỷ lệ hồi phục, cao hơn cả Bắc/Nam Mỹ có y tế và khoa học tiên tiến hơn.
  - + VÀ CHÂU ĐẠI DƯƠNG là châu lục tồi nhất trong việc chữa trị với tỷ lệ phục hồi xấp xỉ 70%.
  - + Mặc dù có dân số thấp hơn châu Á nhưng cả Bắc/Nam Mỹ đều có quá nhiều ca tử vong so với châu Á. Điều này cho thấy khả năng chữa lành của châu Á tốt hơn Bắc/Nam Mỹ.

### • Nhận xét về dữ liệu:

- Đối với tất cả các châu lục, phần lớn các trường hợp đều đã hồi phục được chứng minh bằng màu đỏ chiếm phần lớn thanh bar trong biểu đồ stacked bar chart. Đây là điều rất tích cực trong tình hình đại dịch đang lây lan nhanh.
- Chúng ta chỉ thấy một phần nhỏ màu xanh lá cây của thanh bar châu Âu, châu Á và Bắc Mỹ, có nghĩa là đại dịch có thể vẫn còn đang diễn ra trên các lục địa này.
- Cả tỷ lệ phục hồi và tỷ lệ tử vong đều rất thấp, chứng tỏ khả năng chữa bệnh của toàn thế giới là đáng kinh ngạc.

Với châu Âu:

- Mặc dù không tốt trong việc ngăn chặn đại dịch, các nước châu Âu lại tốt trong việc chữa trị, được chứng minh từ biểu đồ trên khi có số trường hợp hồi phục cao nhất cùng với tỷ lệ hồi phục trên 95%.
- Tuy nhiên, châu Âu cũng có số lượng ca tử vong cao nhất.

Với châu Á:

- Sau đó là châu Á, đứng thứ hai cả về tổng số ca hồi phục và tỷ lệ hồi phục.
- Mặc dù có số lượng ca tử vong lớn nhưng đây là điều tốt nhất mà họ có thể làm với lục địa có dân số đông nhất.

Với các châu lục khác:

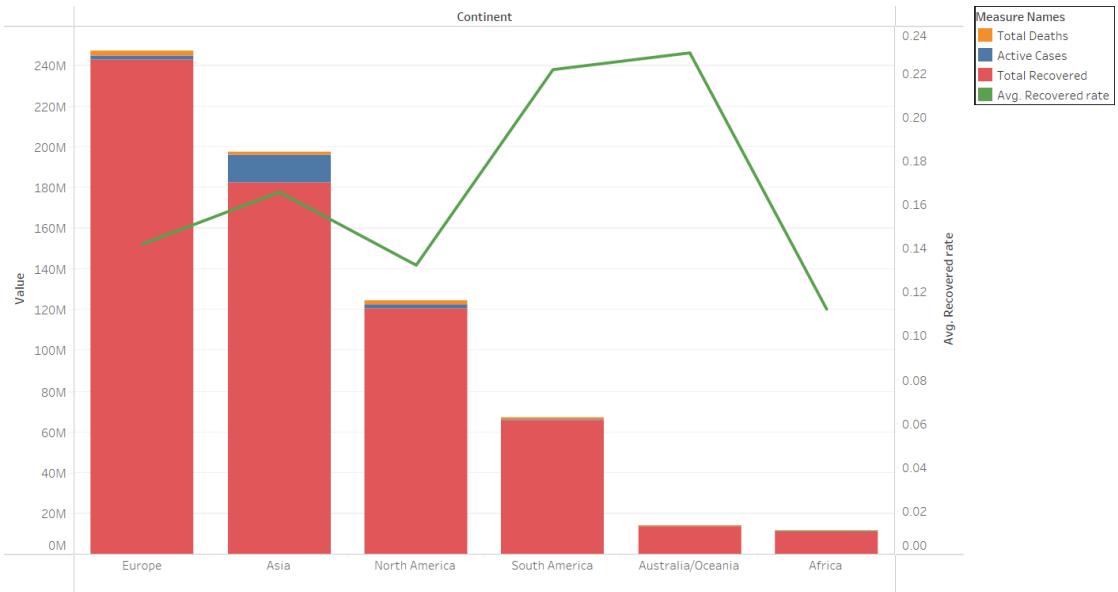
- Ngạc nhiên là châu Phi đứng thứ ba về tỷ lệ hồi phục, cao hơn cả Bắc/Nam Mỹ có y tế và khoa học tiên tiến hơn.
- Và châu Đại Dương là châu lục tồi nhất trong việc chữa trị với tỷ lệ phục hồi xấp xỉ 70%.
- Mặc dù có dân số thấp hơn châu Á nhưng cả Bắc/Nam Mỹ đều có quá nhiều ca tử vong so với châu Á. Điều này cho thấy khả năng chữa lành của châu Á tốt hơn Bắc/Nam Mỹ.

- **Nhận xét về tính trực quan:**

- Ở trang story này, ta chỉ sử dụng 2 sheet: 1 line + stacked bar chart bên trái dùng để nêu lên tổng quan khả năng chữa bệnh của từng châu lục; các bar, line chart bên trái dùng để so sánh chi tiết hơn về các thông số (mà sheet bên trái không trực quan được)
- Ở đây, màu sắc của các thanh bar chart đóng vai trò phân biệt giữa các thông tin dữ liệu và cho biết % mà màu đó chiếm trên tổng thể. Ngoài ra màu sắc còn đóng vai trò liên kết sheet bên trái và các biểu đồ bên phải (bar chart bên trái mô tả cho phần cùng màu trên thanh bar của sheet bên phải)
- Và ta cũng sử dụng kĩ thuật **Partition into Views**

## Các bước thực hiện trực quan:

Biểu đồ 1: stacked bar chart + line chart



B1: đầu tiên, ta thấy rằng chưa có thuộc tính **Recovered rate**, vì vậy ta sẽ thực tạo ra trường dữ liệu đó

B2: kéo thả **Continent** → Columns, lọc các châu lục không hợp lệ; **Measure Values** → Rows

B3: ta xóa các thuộc tính ở khung như hình bên dưới và chỉ để lại các trường **Sum(Active Cases)**, **Sum(Total Death)**, **Sum(Total Recovered)**

Measure Values

- SUM(Total Deaths)**
- SUM(Active Cases)**
- SUM(Total Recovered)**

B4: Ta kéo trường **Measure Name** → Color và thu được stacked bar chart như mong muốn:

Marks

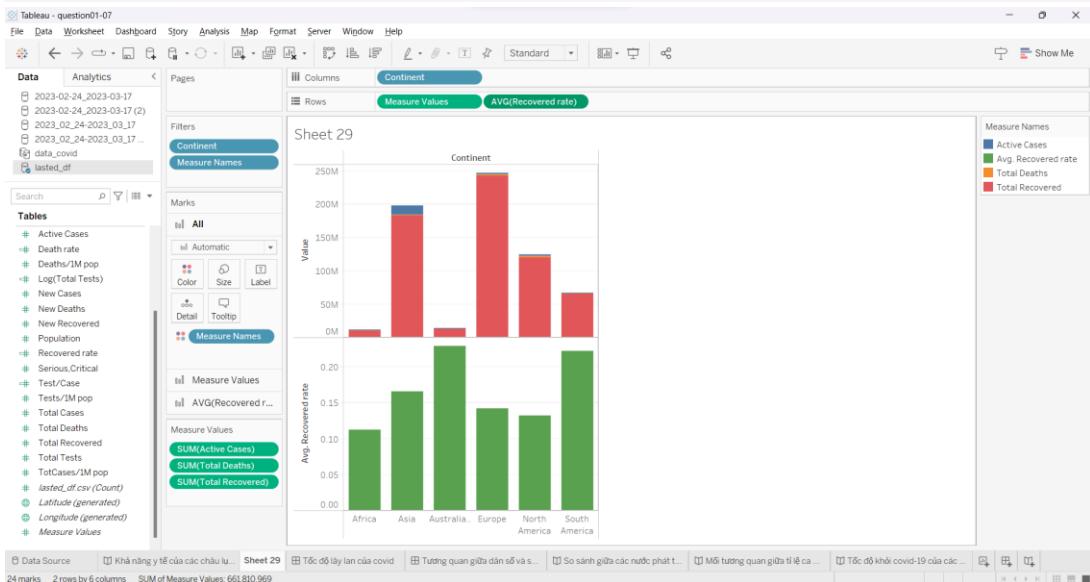
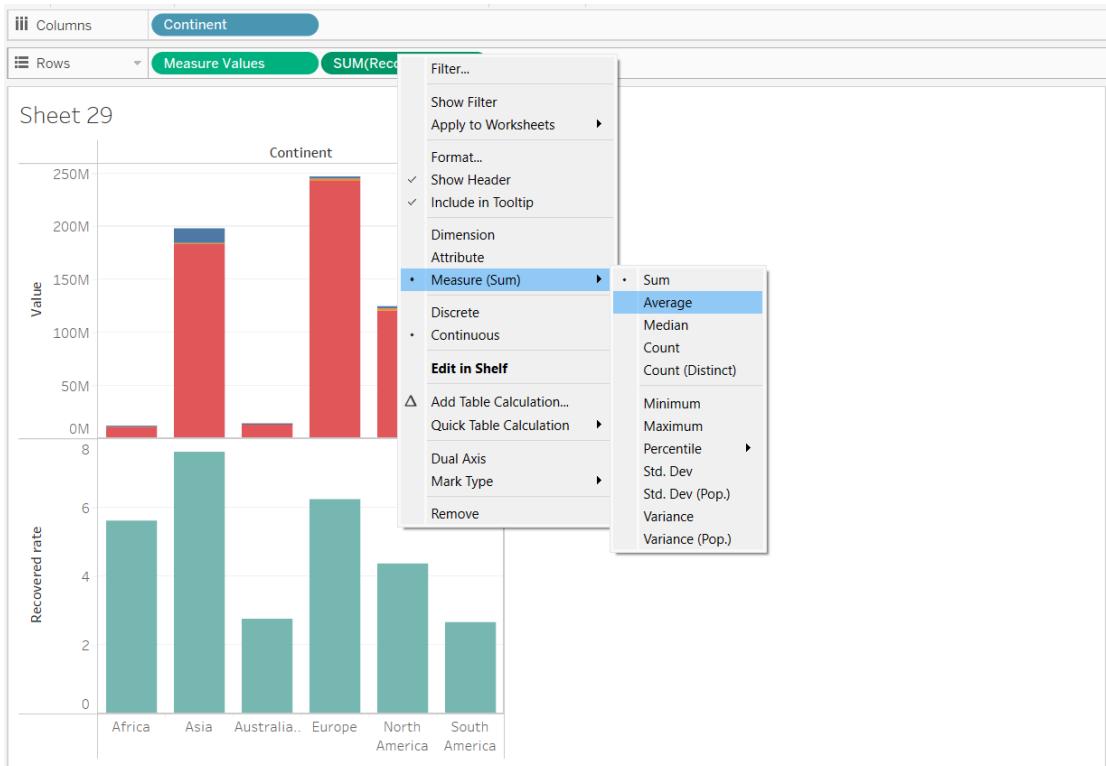
Automatic

Color	Size	Label
Detail	Tooltip	

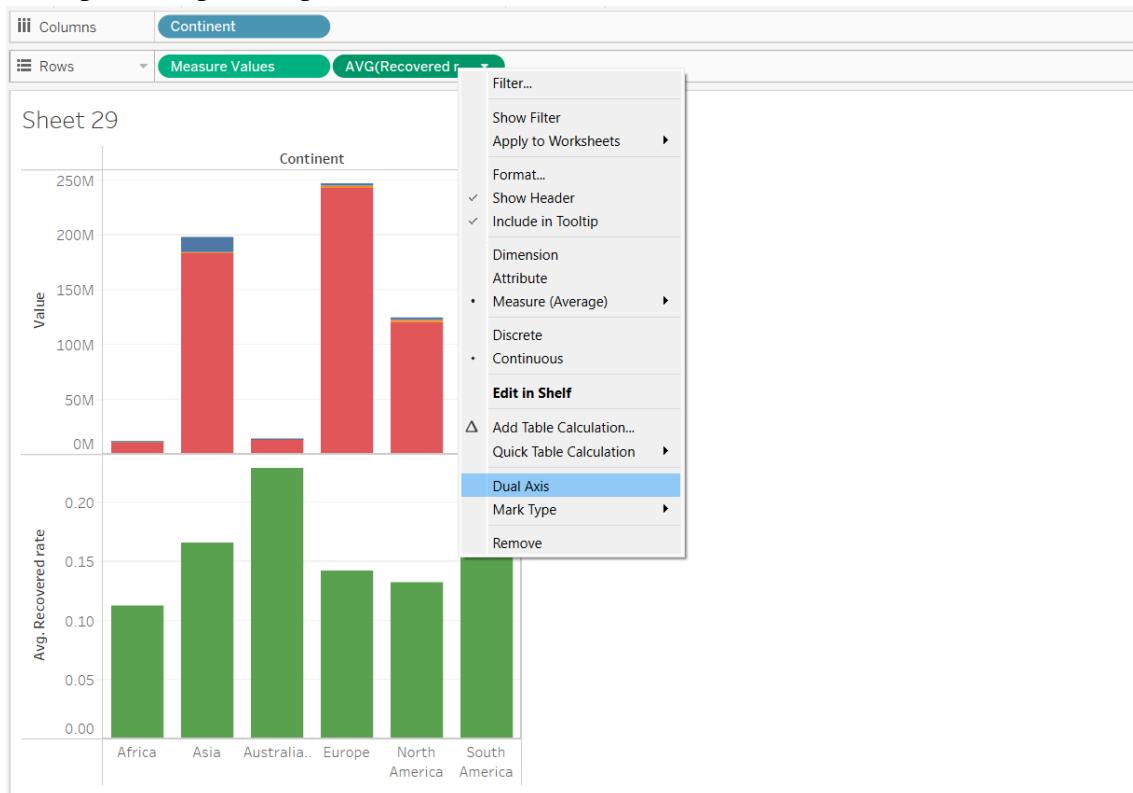
Measure Names



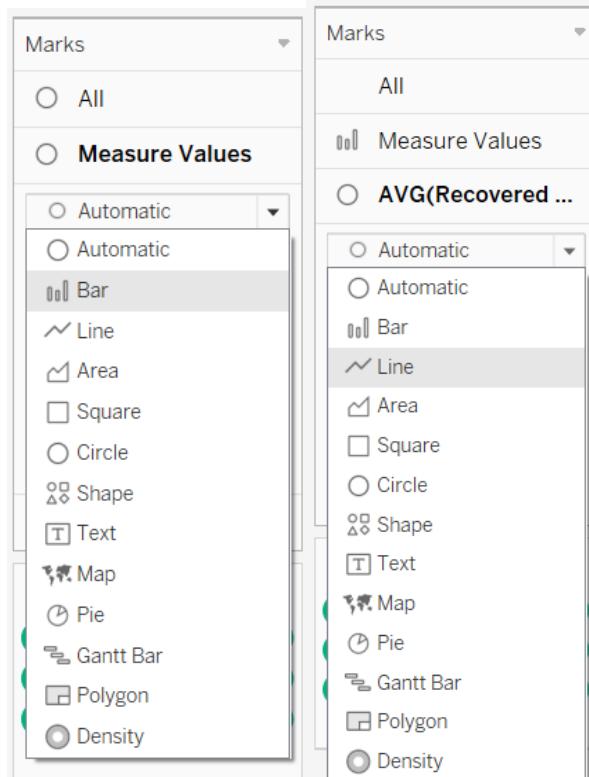
B5: ta tiếp tục vẽ thêm line chart bằng cách kéo trường **Recovered rate** → Rows, nhập phải chuột vào **Sum(Recovered rate)**, Measure → Average



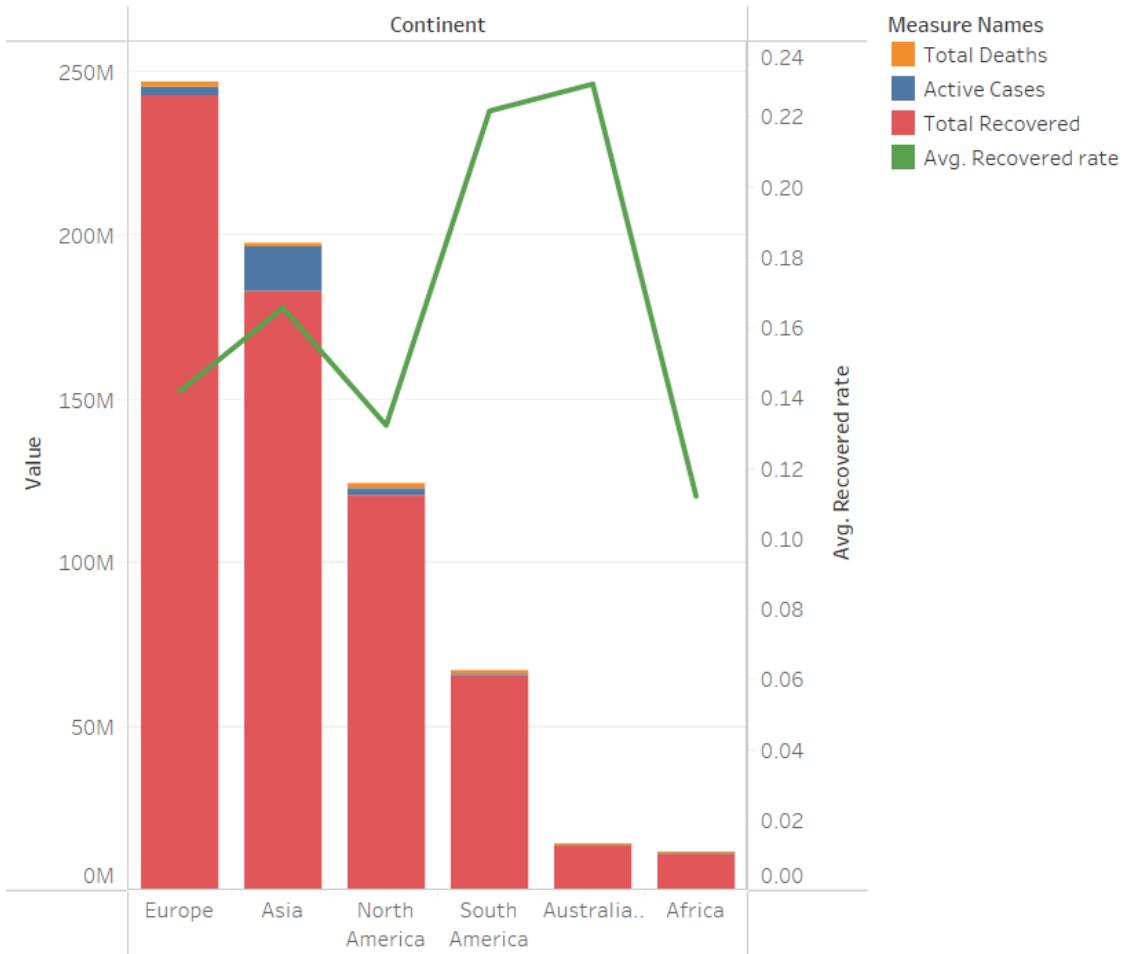
B6: ta tiếp tục nhập chuột phải vào AVG(Recovered rate), chọn Dual axis.



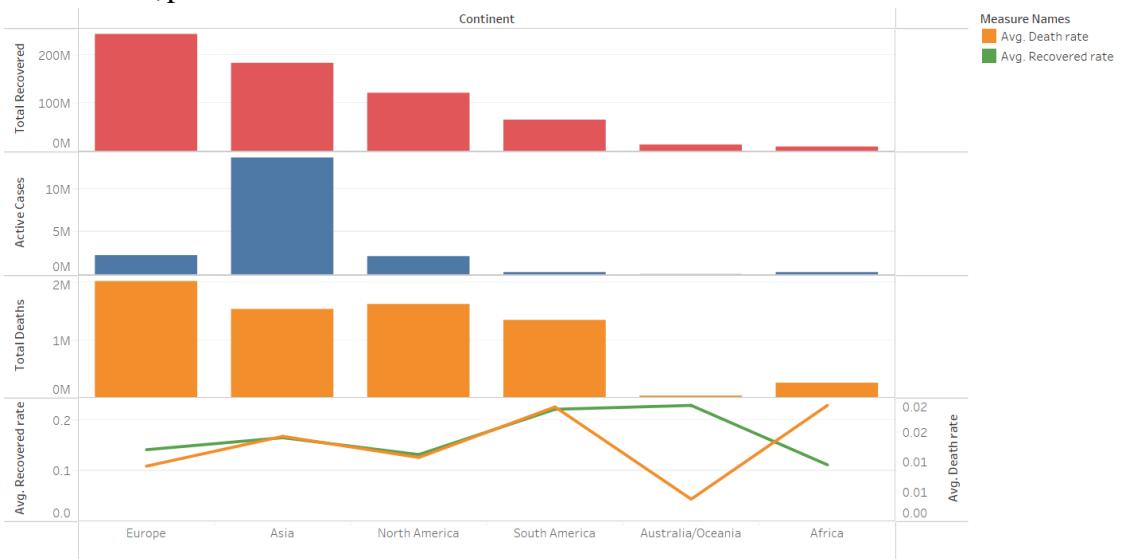
Nếu đồ thị tự động chuyển sang các chấm tròn, ta chỉ cần vào từng phần trong mark và chọn lại kiểu biểu đồ: Bar cho Measure Values và Line cho AVG(Recovered rate)



B7: ta thực hiện tinh chỉnh cuối cùng như sắp xếp các thanh bar, màu sắc,... và ta được kết quả cuối cùng



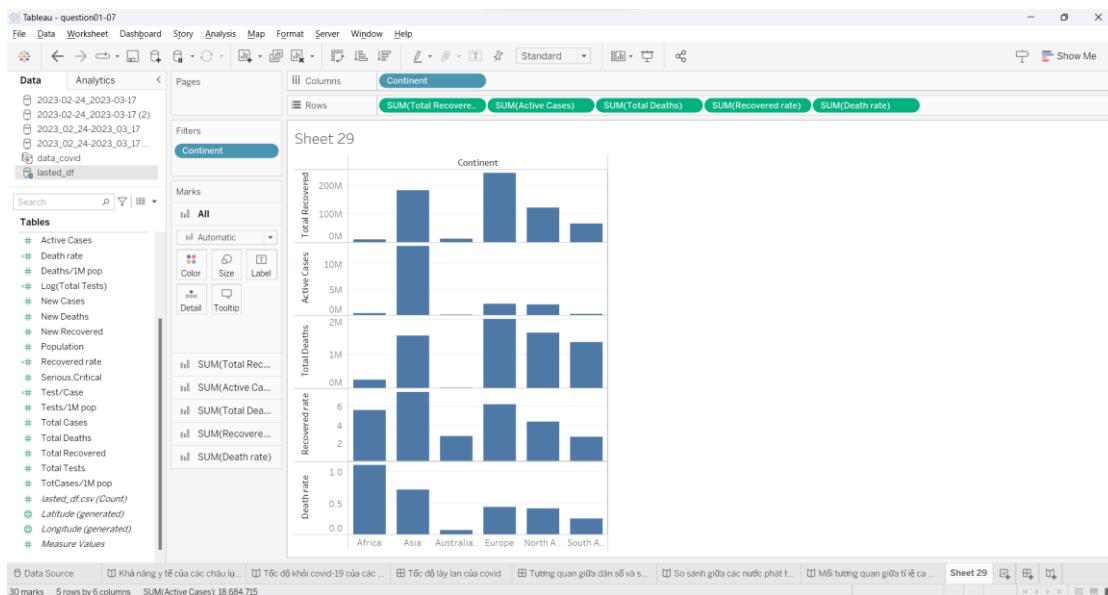
Biểu đồ 2: Tô hợp các bar và line chart



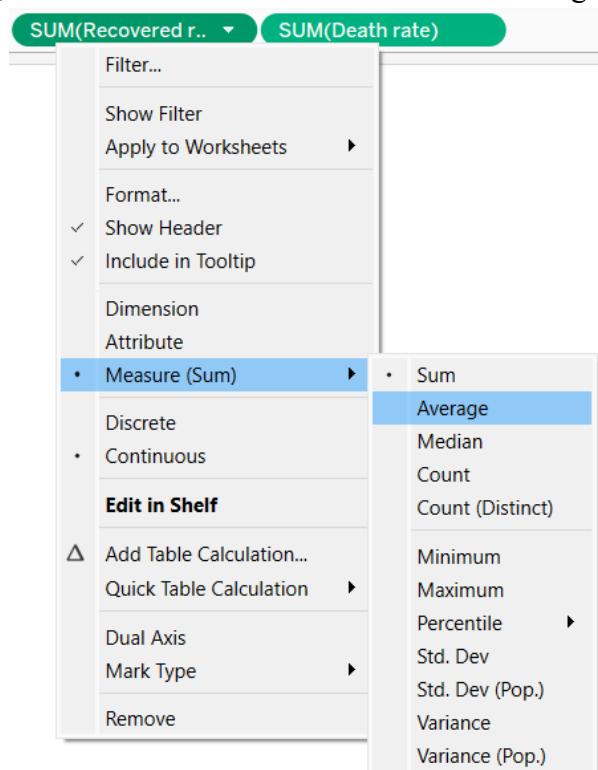
B1: ta sẽ tạo thêm trường mới là **Death rate** tương tự như **Recovered rate** (nên sẽ không trình bày lại)

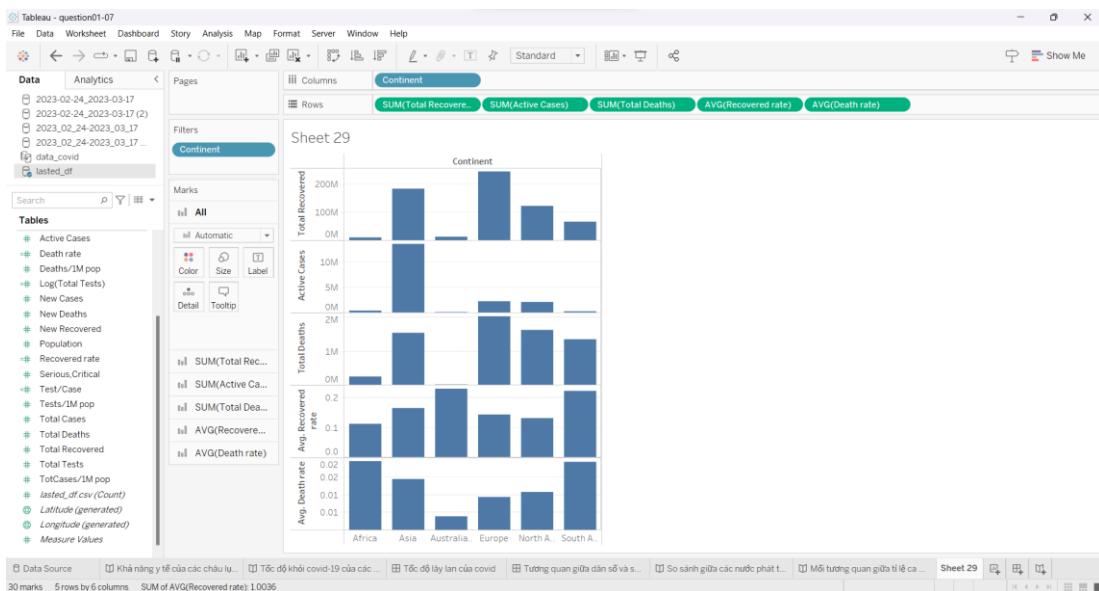
B2: ta kéo các trường vào Rows và Columns như hình bên dưới (lưu ý là vẫn lọc châu lục hợp lệ)

iii Columns	Continent
Rows	SUM(Total Recovered..) SUM(Active Cases) SUM(Total Deaths) SUM(Recovered rate) SUM(Death rate)

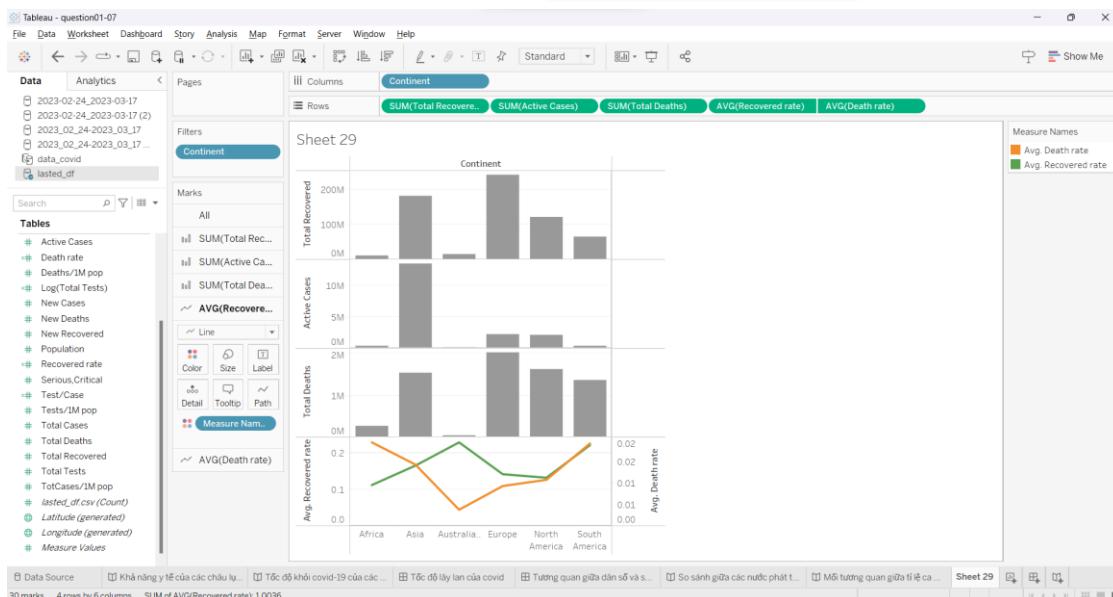
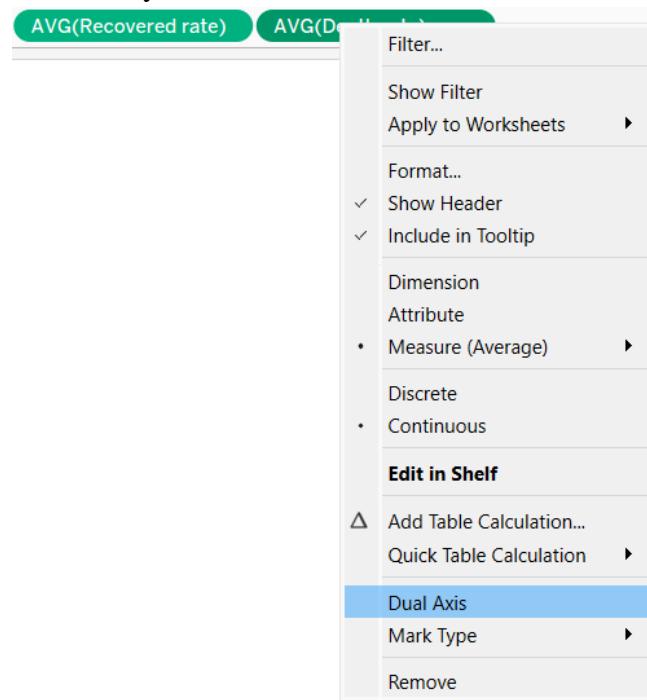


B3: nhấp phải chuột vào Recovered rate → Measure → Average, làm tương tự với Death rate

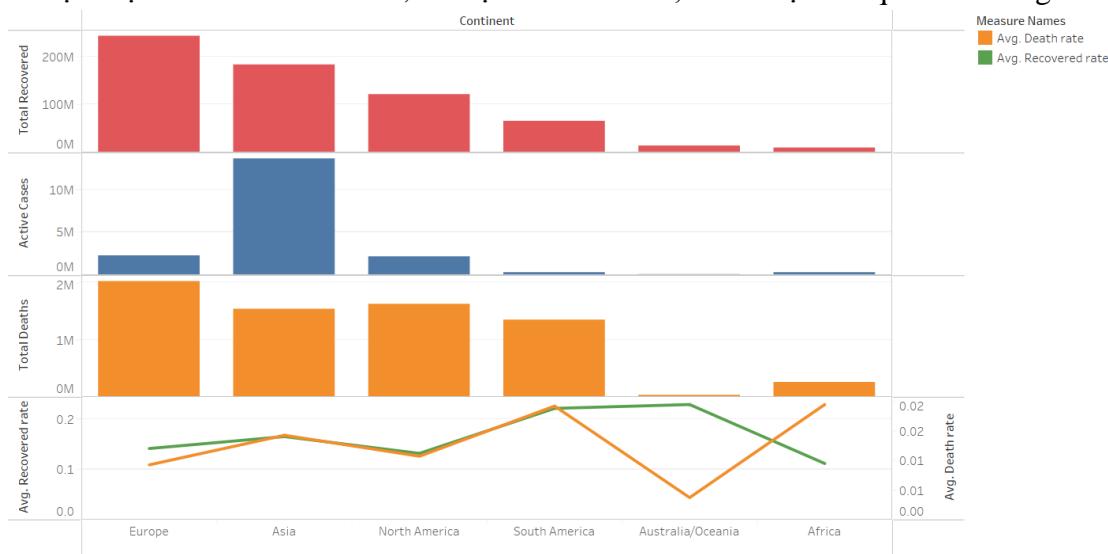




B4: nhấp phải chuột vào AVG(Death rate) → Dual axis. Nếu các đồ thị chuyển về dạng Circle chart thì ta cũng thực hiện chuyển đổi về bar và line chart như đã trình bày ở trên.



B5: ta thực hiện tinh chỉnh màu sắc, thứ tự các thanh bar,... và được kết quả cuối cùng



## 4.2. Tình hình dịch bệnh hiện tại của các châu lục trên thế giới

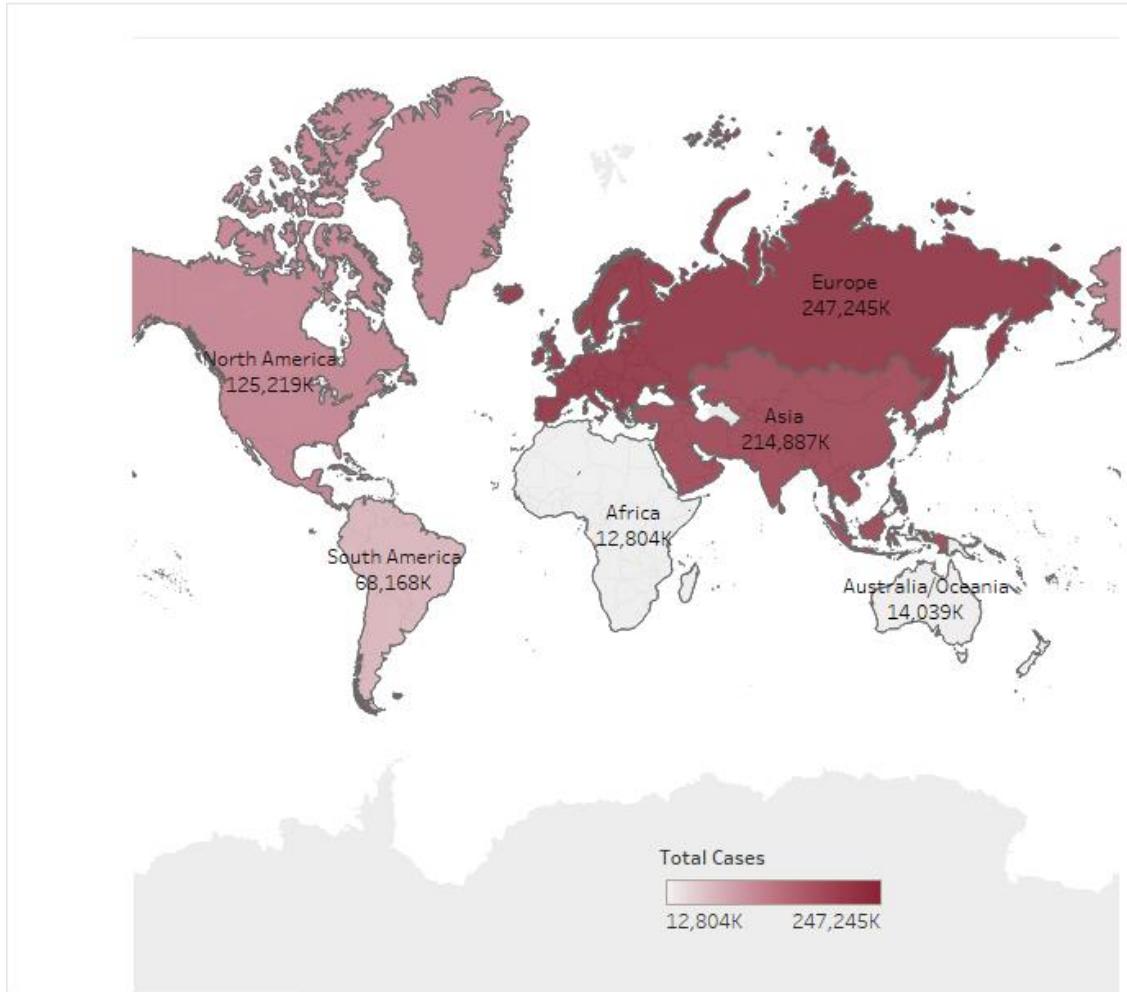
### 4.2.1. Tiếp cận vấn đề

Khảo sát tổng số ca nhiễm và tổng số ca nhiễm chưa khỏi tính đến thời điểm gần nhất của bộ dữ liệu (17/03/2023). Từ đó có được cái nhìn tổng quan về dịch bệnh

### 4.2.2. Trực quan bằng biểu đồ trên tableau

#### a. Tổng số ca nhiễm các châu lục tính đến ngày 17/03/2023

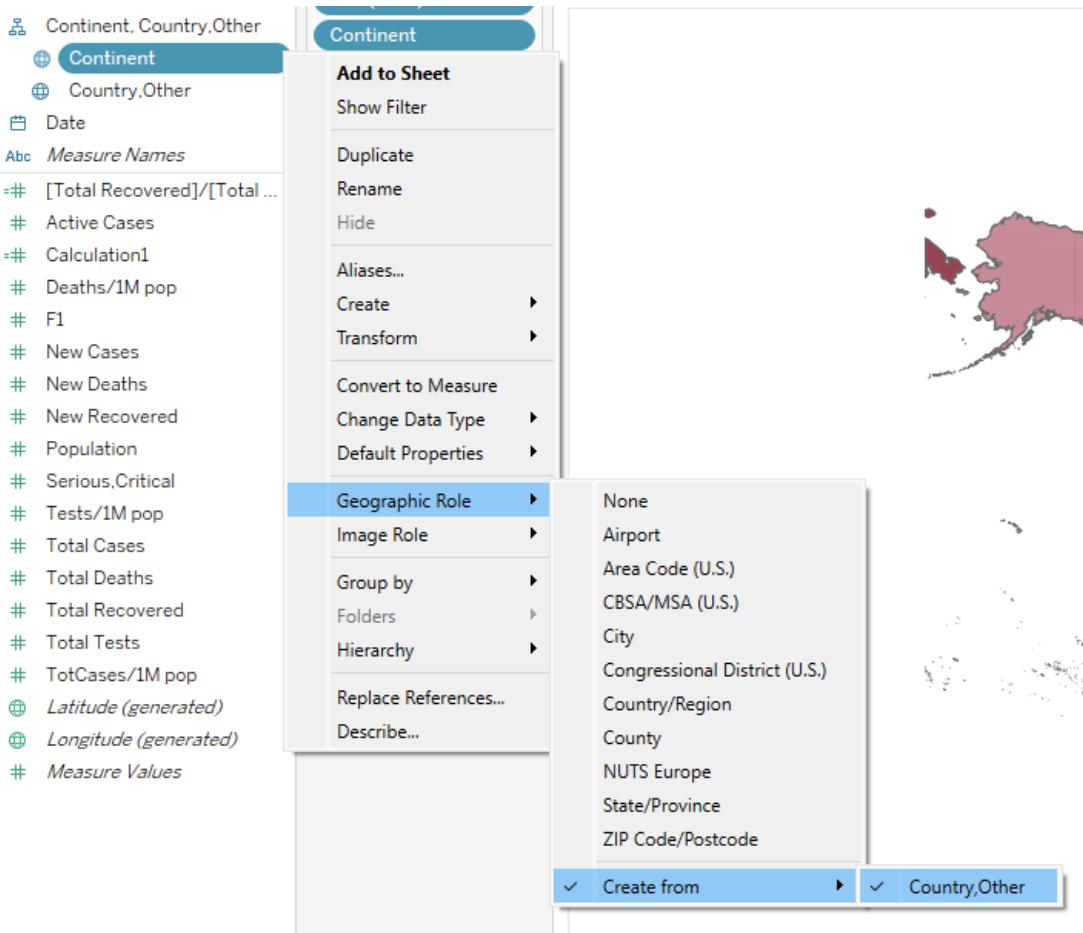
Total cases by continents up to 17/03/2023



- **Nhận xét về dữ liệu:**
  - Châu Âu và Châu Á là hai châu lục có số ca nhiễm nhiều và chênh lệch nhất với hơn 200 triệu ca nhiễm
  - Châu Đại Dương và châu Phi ghi nhận số ca ít nhất.
- **Nhận xét tính trực quan:**
  - Sử dụng biểu đồ dạng map không những cho thêm nhiều thông tin hơn (vị trí các châu lục) mà còn mang tính thẩm mỹ hơn.
  - Màu sắc được sử dụng sao cho nổi bật với màu trắng (phần nước biển) giúp nổi bật được châu lục. Chọn màu đỏ đô để thể hiện được sự nghiêm trọng của dịch bệnh

#### Các bước thực hiện trực quan biểu đồ:

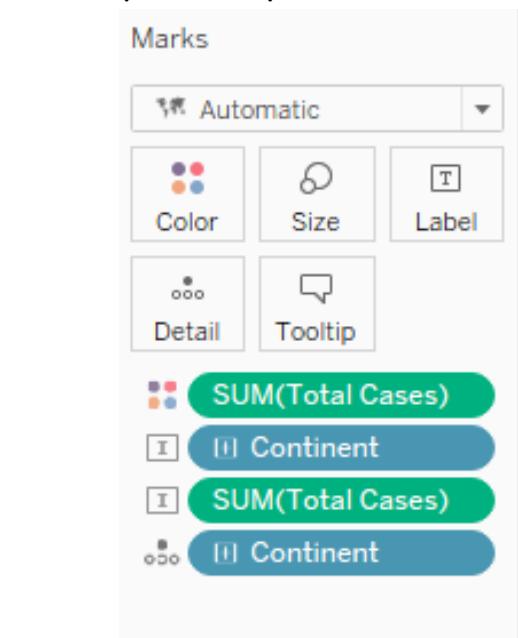
B1: Tạo vai trò địa lý giữa thuộc tính châu lục và quốc gia



B2: Sau khi thực hiện bước 1, tableau sẽ hiển thị vĩ độ và kinh độ ở các thuộc tính. Sử kinh độ ở vùng column và vĩ độ ở vùng row

Columns	Longitude (generated)
Rows	Latitude (generated)

B3: Cấu hình marks. Sử dụng thuộc tính *TotalCases* để hiển thị màu và số. Thuộc tính *Continent* cũng được sử dụng để hiển thị tên châu lục



B4: Sử dụng Filter để lọc lấy ngày 17/03/2023 và bỏ đi châu lục “All”

**Filter [Date]**

Relative dates      Range of dates      Starting date      Ending date      Special

Range of dates

3/17/2023      3/17/2023

2/24/2023      3/17/2023

Show: Only Relevant Values       Include Null Values

Reset      OK      Cancel      Apply

**Filter [Continent]**

General      Wildcard      Condition      Top

Select from list  Custom value list  Use all

Enter search text

<input type="checkbox"/> Africa
<input checked="" type="checkbox"/> All
<input type="checkbox"/> Asia
<input type="checkbox"/> Australia/Oceania
<input type="checkbox"/> Europe
<input type="checkbox"/> North America
<input type="checkbox"/> South America

All      None       Exclude

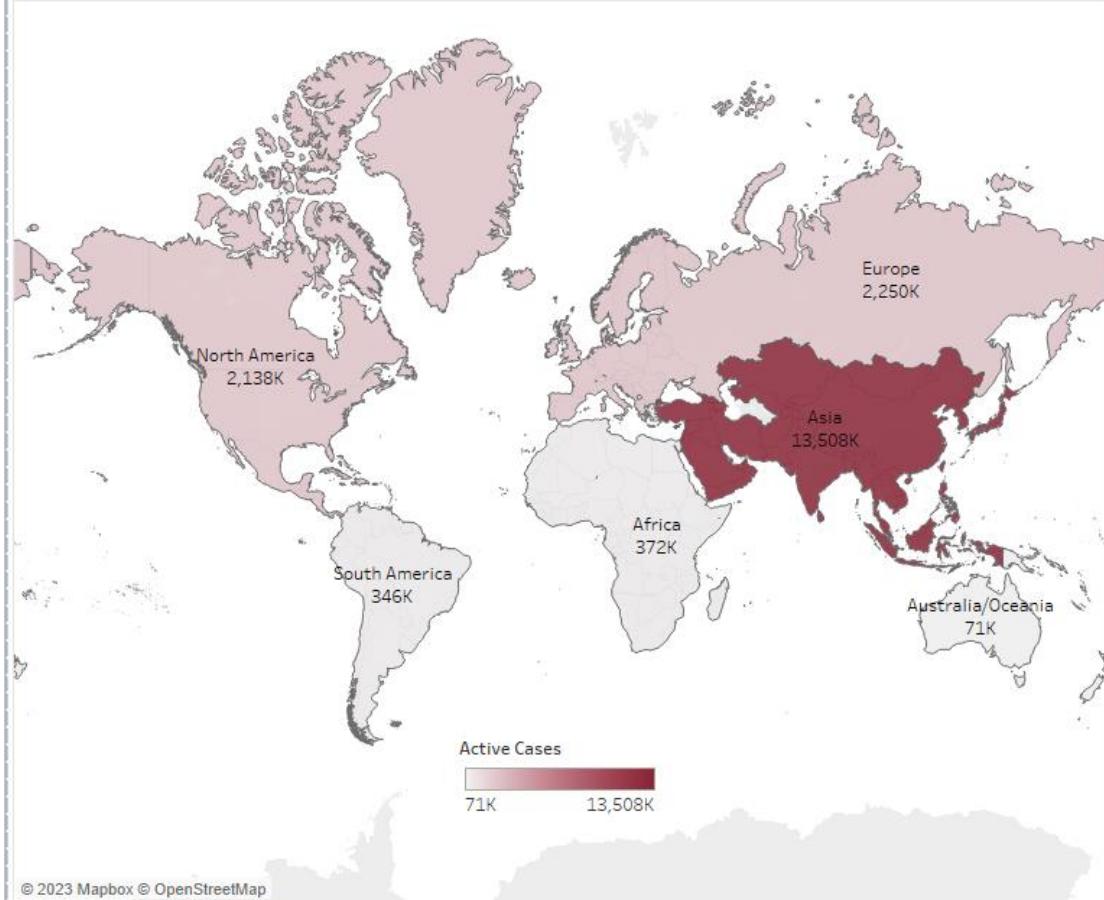
**Summary**

Field: [Continent]  
 Selection: Excluded 1 of 7 values  
 Wildcard: All  
 Condition: None  
 Limit: None

Reset      OK      Cancel      Apply

b. Tổng số ca nhiễm chưa hồi phục của các châu lục

active cases by continents on 17/03/2023



- **Nhận xét về dữ liệu:**

- Từ biểu đồ trước ta nhận thấy tổng số ca nhiễm ở Châu Âu và Bắc Mỹ cao, nhưng từ biểu đồ trên thì số ca nhiễm hiện tại đã giảm nhiều

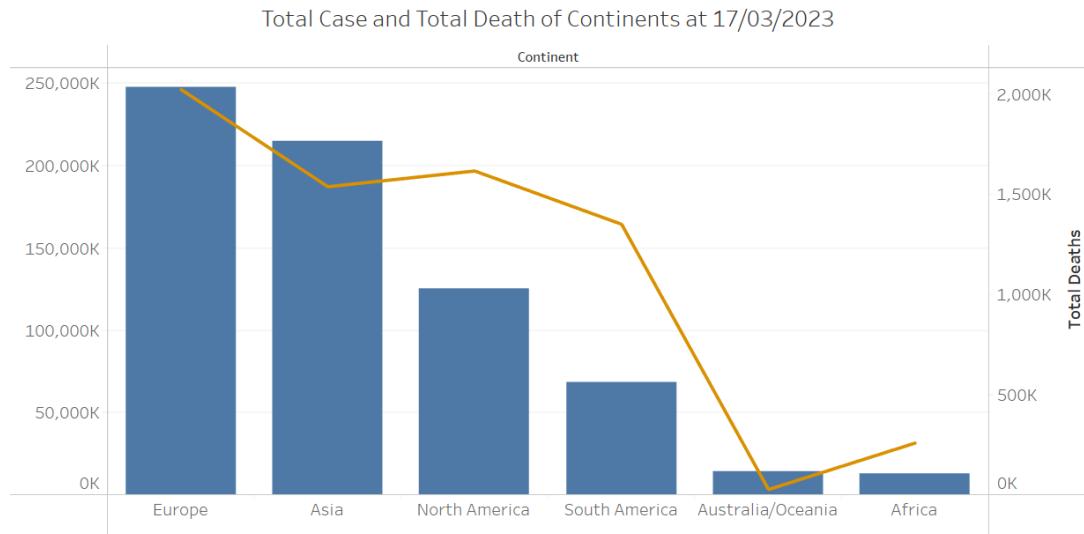
### 4.3. Số ca mắc cao thì sẽ dẫn đến số ca tử vong cao?

#### 4.3.1. Tiếp cận vấn đề

Trực quan đồng thời tổng số ca nhiễm và tổng số người chết, từ đó quan sát xem hai thuộc tính này có đồng biến/nghịch biến hay không. Sau đó ta sẽ xem xét đến mức độ quan tâm của châu lục đó đối với dịch bệnh qua tổng số ca xét nghiệm. Cuối cùng đánh giá khả năng chống lại dịch bệnh qua tổng số ca hồi phục.

#### 4.3.2. Trực quan bằng biểu đồ tableau

- a. Tổng số ca nhiễm và ca tử vong tính đến ngày 17/03/2023



- **Nhận xét dữ liệu:**

- Nhìn chung, những châu lục có ca mắc cao đều có số lượng tử vong cao. Nhưng có hai cặp châu lục có vẻ như không bình thường. Số ca mắc ở Châu Á cao hơn Bắc Mỹ nhưng số người tử vong lại thấp hơn. Tương tự Châu Đại Dương và Châu Phi.
- Biểu đồ tiếp theo sẽ phân tích kỹ hơn về hai cặp châu lục này

- **Nhận xét tính trực quan:**

- Sử dụng bar chart kết hợp với line chart để nhận thấy có quan hệ nhân quả dễ dàng giữa các thuộc tính về số lượng. Tuy nhiên cách làm này sẽ khó khăn đối với quan hệ về phân lớp.
- Sử dụng hai màu sáng khác nhau giữa bar và line để dễ dàng quan sát.
- Đối với hai châu lục có số lượng ca mắc ngang nhau và cần so sánh thì thêm số lượng ở đầu mỗi cột

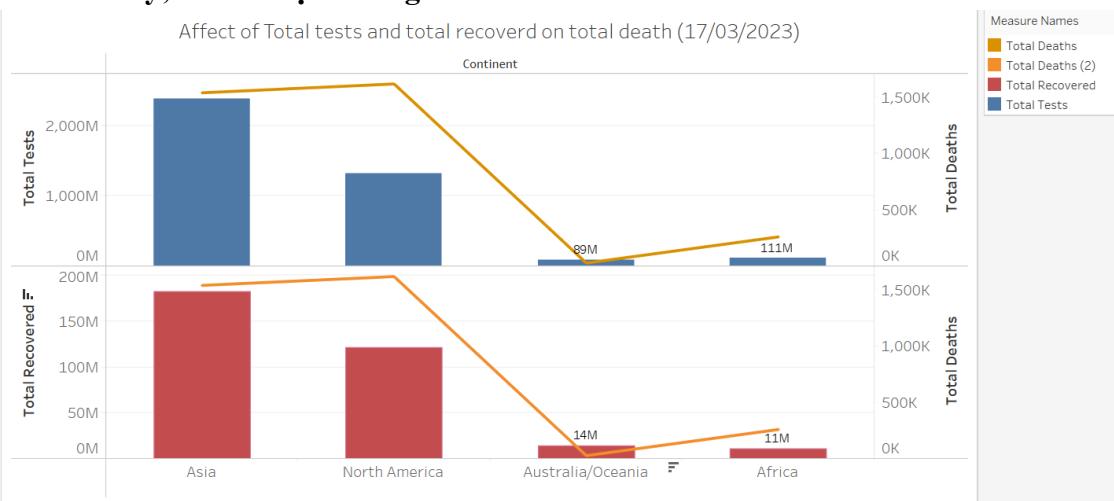
#### Các bước thực hiện trực quan biểu đồ:

B1: Thả thuộc tính *Continent* vào vùng Columns, thả thuộc tính *TotalCases*, *TotalDeaths* vào vùng Rows. Chọn *TotalDeaths* làm dual axis.

B2: Chọn *TotalCases* làm bar chart và *TotalDeaths* làm line chart.

B3: Dùng Filters để lọc ngày và loại “All” khỏi thuộc tính *Continent*

**b. Quan hệ số ca test, số ca hồi phục với số ca tử vong của hai cặp châu lục Châu Á và Bắc Mỹ; Châu Đại Dương và Châu Phi**



- **Nhận xét dữ liệu:**

- Ở cặp đầu tiên: Châu Á và Bắc Mỹ, chúng ta có thể cho rằng tổng số ca xét nghiệm(TotalTests) cao hơn tổng số ca phục hồi(TotalRecovered) thì tổng số ca tử vong sẽ thấp hơn. Ở cặp thứ hai: Châu Phi và Châu Đại Dương, tổng số ca xét nghiệm dường như không ủng hộ lý thuyết trước đó nhưng tổng số ca hồi phục thì có. Trong cả hai cặp thì châu lục có tổng số ca hồi phục cao hơn luôn có số ca tử vong thấp hơn
- Chúng ta có thể kết luận rằng tổng số ca nhiễm cao chưa chắc đã dẫn đến tổng số ca tử vong cao. Tổng số ca tử vong có mối quan hệ cả tổng số ca nhiễm và tổng số ca hồi phục. Tổng số ca nhiễm phản ánh mức độ nghiêm trọng của dịch bệnh, Tổng số ca phục hồi phản ánh trình độ y tế của lục địa đó

- Nhận xét trực quan:**

- Giống như biểu đồ trước, sử dụng bar chart kết hợp với line chart để thấy được quan hệ tăng giảm của hai thuộc tính.
- Vì hai cột bên phải của mỗi biểu đồ có chiều cao chênh lệch nhỏ, khó phân biệt bằng mắt thường, nên nhóm quyết định hiển thị label thẻ hiện giá trị của từng cột.
- Sử dụng màu khác nhau cho các cột của hai biểu đồ để phân biệt số ca xét nghiệm và số ca hồi phục. Hai đường line sẽ có màu giống nhau tượng trưng số ca tử vong

**Các bước trực quan:**

B1: Thả *Continent* vào vùng Columns. Ở vùng Row sử dụng *TotalDeaths* làm dual cho *TotalTests* và *TotalRecovered*.



B2: Chọn mark type là bar chart cho *TotalTests* và *TotalRecovered*, line chart cho *TotalDeaths*

B3: Chuột phải vào hai cột cuối bên phải chọn Always Show ở Marl Label

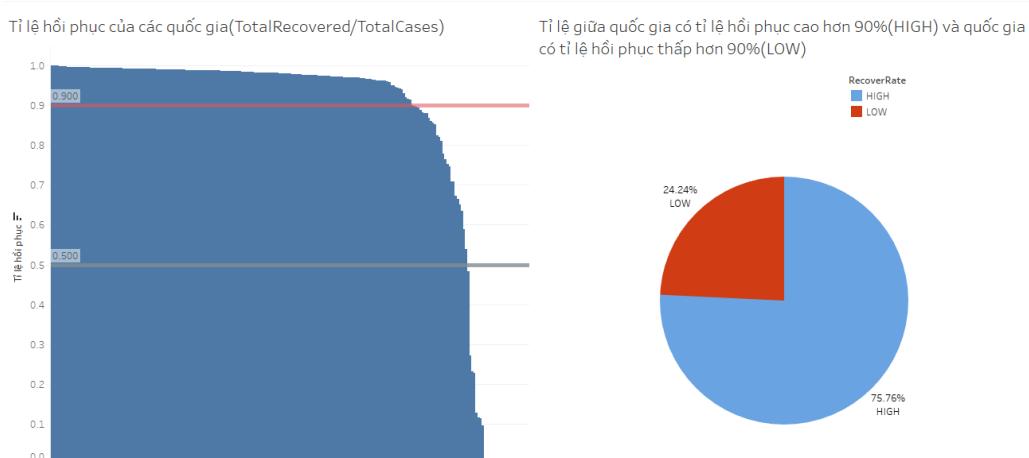
## 4.4. Tốc độ khỏi bệnh của Covid-19 là bao nhiêu? Những quốc gia nào đang có tỷ lệ khỏi bệnh cao nhất?

### 4.4.1. Tiếp cận vấn đề

- Thể hiện sự phân bố của tốc độ khỏi bệnh các quốc gia. Tốc độ khỏi bệnh được nhóm quy định =  $TotalRecovered/TotalTests$ . Tìm những quốc gia có tỷ lệ khỏi bệnh cao nhất và phân tích sâu hơn về các quốc gia này

### 4.4.2. Trực quan bằng biểu đồ tableau

#### a. Tỉ lệ hồi phục của các quốc gia và tỉ lệ quốc gia có tỉ lệ hồi phục cao hơn 90%



- **Nhận xét dữ liệu:**

- Biểu đồ bên trái cho thấy được phân bố tỉ lệ hồi phục của các quốc gia. Có rất nhiều quốc gia có tỉ lệ hồi phục trên 50%. Bên cạnh đó qua biểu đồ thứ hai, có tới hơn 75% các quốc gia có tỷ lệ hồi phục trên 90%. Điều này rất đáng mừng, cho thấy được tình hình dịch bệnh của thế giới đang được kiểm soát

- **Nhận xét trực quan:**

- Biểu đồ bên trái cho có dạng giống như histogram, cho thấy phân bố tỉ lệ hồi phục giữa các quốc gia. Hai đường line kẻ ngang qua biểu đồ chính là các mức độ cần xét, giúp việc quan sát dễ dàng hơn.
- Để làm rõ hơn thì pie chart bên trái cho thấy tỉ lệ phần trăm của hai nhóm. Màu sắc của pie chart cũng được sử dụng đúng với ý nghĩa

**Các bước trực quan:**

Biểu đồ bar bên trái

B1: Thả thuộc tính *Country, Other* vào vùng columns.

B2: Tính tỉ lệ TotalRecovered/TotalTests ở vùng rows

SUM([[Total Recovered]]/[Total Cases]])

B3: Vào **Analytics** → **Reference line** để tạo hai đường line với constant tùy chỉnh

B4: Dùng Filters để lọc ngày và quốc gia “null”

Biểu đồ pie bên phải

B1: Tạo caculator field **RecoverRate** để trả về HIGH, LOW tương ứng với các quốc gia có tỉ lệ hồi phục cao hơn 90% và thấp hơn 90%



B2: Tạo caculator field **ShowRecoverRate** hiển thị phần trăm của hai nhóm

ShowRecoverRate

X

Totals summarize values from Table (across).

COUNT ([RecoverRate]) / TOTAL(COUNT ([RecoverRate]))

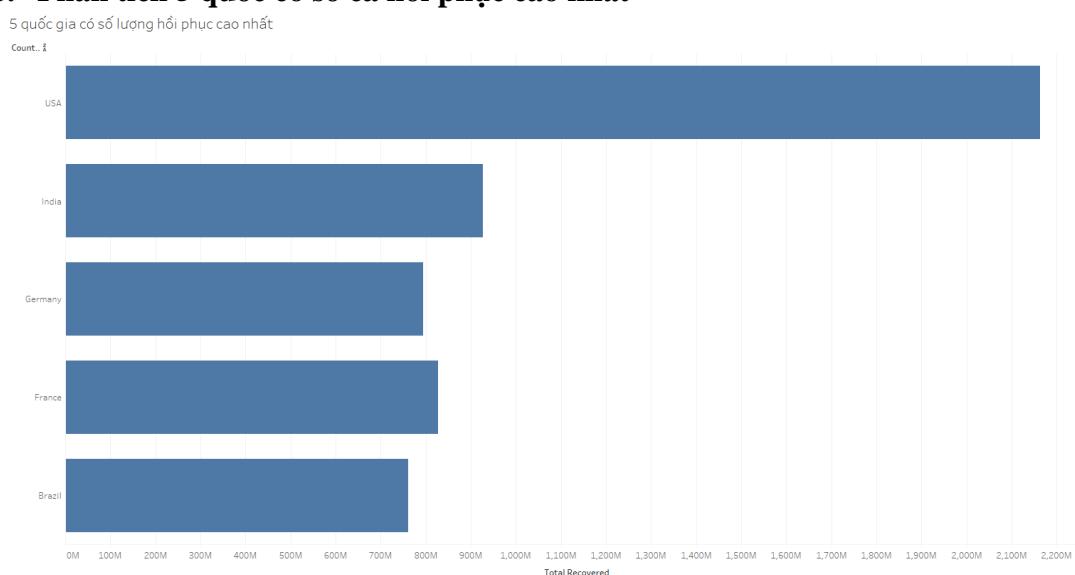
B3: Thả RecoverRate vào color mark và chinh mark type sang pie

B4: Thả TotalRecovered vào Angle, đổi Measure sang Count

B5: Thả ShowRecoverRate và RecoverRate vào Angle

Cuối cùng, thêm hai biểu đồ vào dashboard

### b. Phân tích 5 quốc gia có số ca hồi phục cao nhất

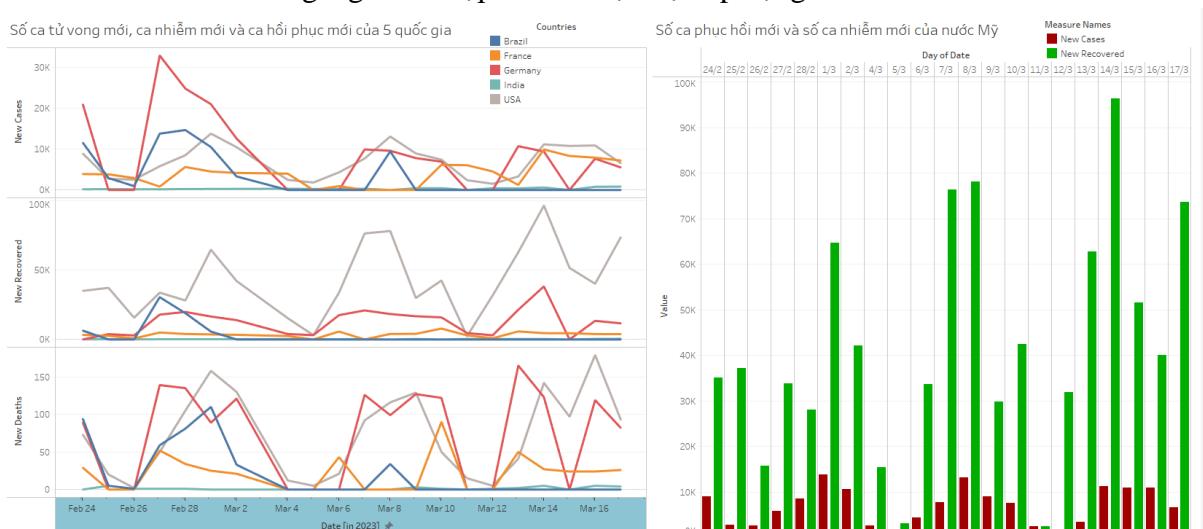


- **Nhận xét dữ liệu:**

- Mỹ đứng đầu danh sách với hơn 100 triệu ca hồi phục. Có thể thấy Mỹ có số ca hồi phục chênh lệch so với các nước Ấn Độ, Đức, Pháp, Brazil.

- **Nhận xét tính trực quan:**

- Bar chart nằm ngang thích hợp để thể hiện độ xếp hạng



- **Nhận xét dữ liệu:**

- Nhìn vào biểu đồ bên trái:

- Các ca nhiễm và tử vong mới đang giảm và duy trì ở mức thấp. Các ca hồi phục đang tăng ổn định trên những quốc gia có nhiều ca nhiễm
- Số tử vong ở Mỹ và Đức trong những ngày qua có sự gia tăng lớn
- Số ca phục hồi ở Mỹ cao hơn các nước khác, có lẽ vì họ có cơ sở y tế tốt. Để nhìn rõ hơn chúng ta sẽ cùng nhìn kỹ hơn vào nước Mỹ ở biểu đồ bên phải
- Nhìn vào biểu đồ bên phải
- Số ca phục hồi mới luôn cao hơn số ca nhiễm mới. Chứng tỏ y tế của Mỹ rất tốt, có khả năng chữa trị covid cao.
- Đặc biệt là những ngày cuối (gần đây nhất) số ca hồi phục tăng mạnh hơn. Cho thấy Mỹ đang rất cố gắng đẩy lùi dịch bệnh
- **Nhận xét tính trực quan:**
  - Ở biểu đồ bên trái
  - Sử dụng line chart để thể hiện sự thay đổi theo thời gian từ đó có cái nhìn về quan hệ giữa các thuộc tính
  - Các line có các màu sắc khác nhau tương ứng với từng quốc gia\
  - Ở biểu đồ bên phải
  - Dùng dual bar chart để thể hiện được sự chênh lệch giữa hai lớp theo thời gian
  - Sử dụng hai màu khác nhau cho 2 cột để thấy được sự khác biệt. Màu xanh tượng trưng cho sự hồi phục , màu đỏ mang tính nghiêm trọng nên dùng biểu thị cho số ca nhiễm mới.

#### Các bước trực quan:

Ở biểu đồ bên trái

B1: Thả Date vào vùng columns, NewCases, NewRecovered, NewDeaths vào vùng Rows.

B2: Dùng Filter để lọc ra 5 quốc gia: Mỹ, Ấn Độ, Đức, Pháp và Brazil

Ở biểu đồ bên phải

B1: Thả Date vào vùng columns, NewCases, NewRecovered vào vùng Rows.

B2: Chọn biểu đồ dual bar chart trên Tag **Show Me**

B3: Dùng Filters để chọn quốc gia Mỹ

## 4.5. So sánh tình hình dịch bệnh và khả năng kiểm soát bệnh ở giữa các nước phát triển và các nước đang phát triển

### 4.5.1. Tiếp cận vấn đề.

Đối với 2 nhóm đối tượng nước phát triển và nước đang phát triển, nhóm em chọn các nước tiêu biểu cho 2 nhóm này như sau:

- Nhóm nước phát triển gồm 7 nước: Israel, USA, UK, France, Germany, S. Korea, Japan.
- Nhóm nước đang phát triển gồm 7 nước: Vietnam, China, Russia, India, Nepal, Egypt.

#### Chuẩn bị dữ liệu:

Nhóm em trực quan dữ liệu ngày gần nhất của dữ liệu thu thập được. Ngoài ra còn có thể chọn ngày khác để trực quan. Trong Tableau, ta kéo thả field Date vào Filter Tag. Ta áp dụng Filter Date này cho toàn bộ các sheet. Bằng cách right click vào filter Date > apply to worksheets > selected worksheets > all .

Đối với các nhóm nước phát triển cũng tương tự.



Để có thể so sánh được tình hình dịch bệnh giữa hai nhóm, nhóm em cần tìm hiểu được tình hình dịch bệnh của từng nhóm nước.

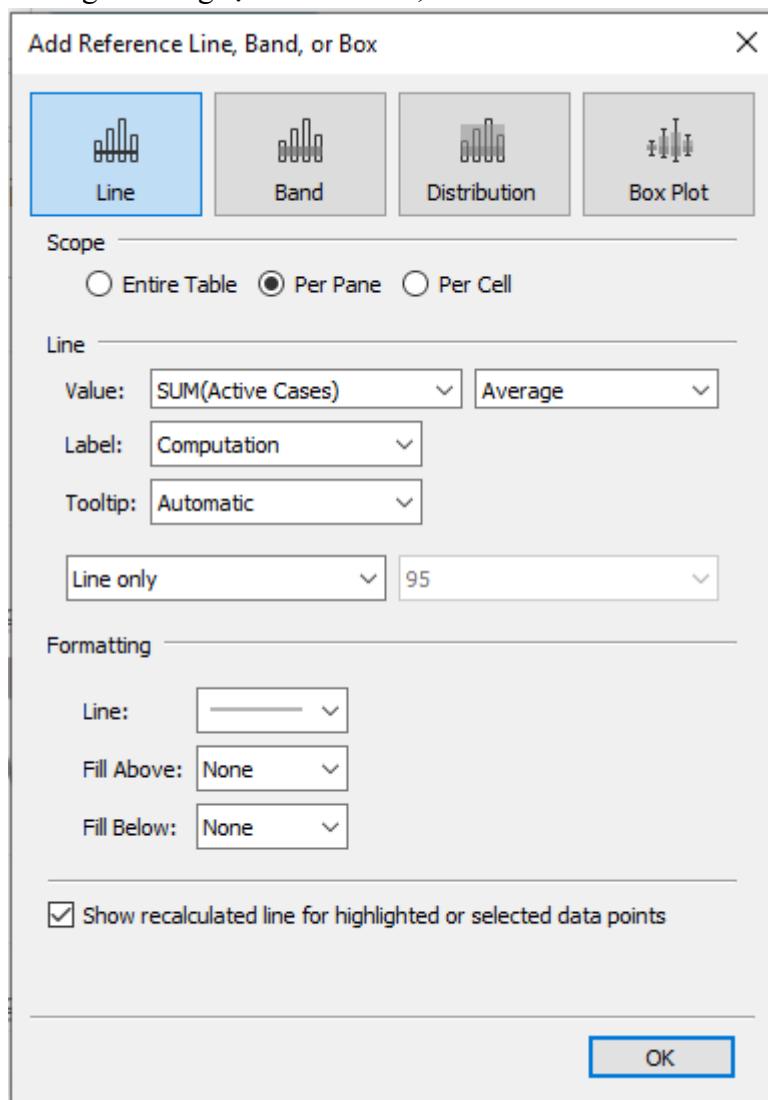
#### 4.5.2. Trực quan bằng biểu đồ Tableau

##### a) Tình hình dịch bệnh ở 2 nhóm phát triển và đang phát triển.

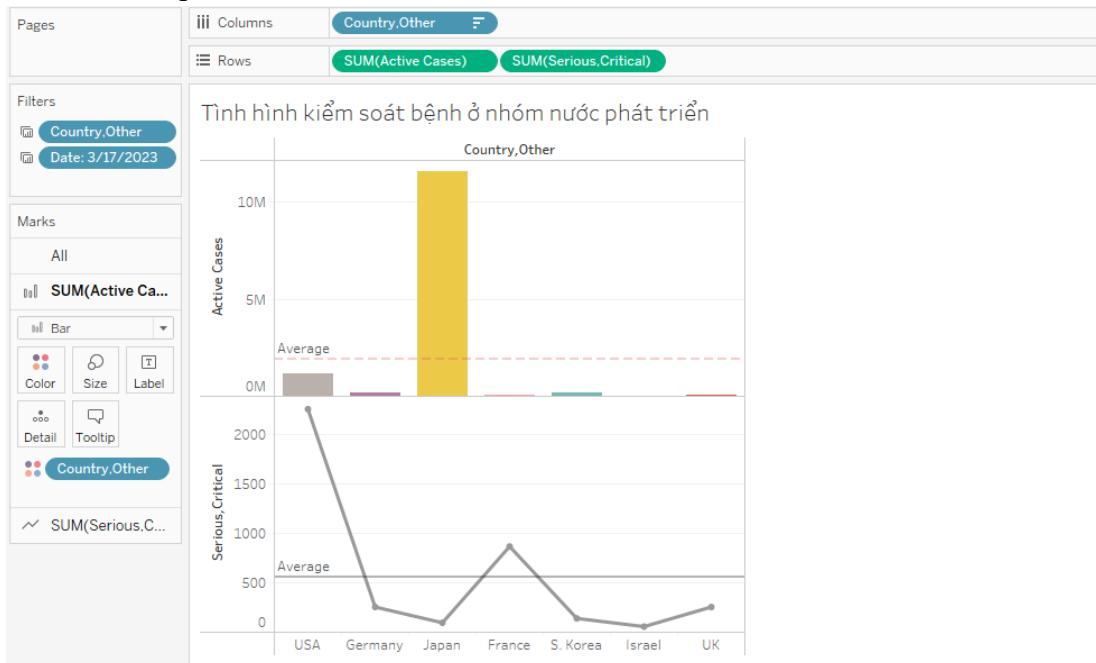
B1: Trực quan hóa nhóm nước phát triển và nhóm nước đang phát triển.. Sử dụng biểu đồ cột (bar chart) kết hợp với biểu đồ đường (Line chart).

Trực quan biểu đồ cột nhóm em làm như sau:

- Chọn filter các nước phát triển trong phần Chuẩn bị dữ liệu.
- Kéo thả field “Active Cases” (trường biểu hiện số ca đang mắc tại mỗi quốc gia) và “Serious,Critical” vào Rows. Có thể chọn các Aggregation method khác như AVG() hoặc SUM().
- Kéo thả field “Country,Other” vào Columns.
- Chuột phải vào “Active Cases” axis, chọn Add Reference Line, tại ô Value chọn Average. Tương tự với “Serious,Critical” axis.



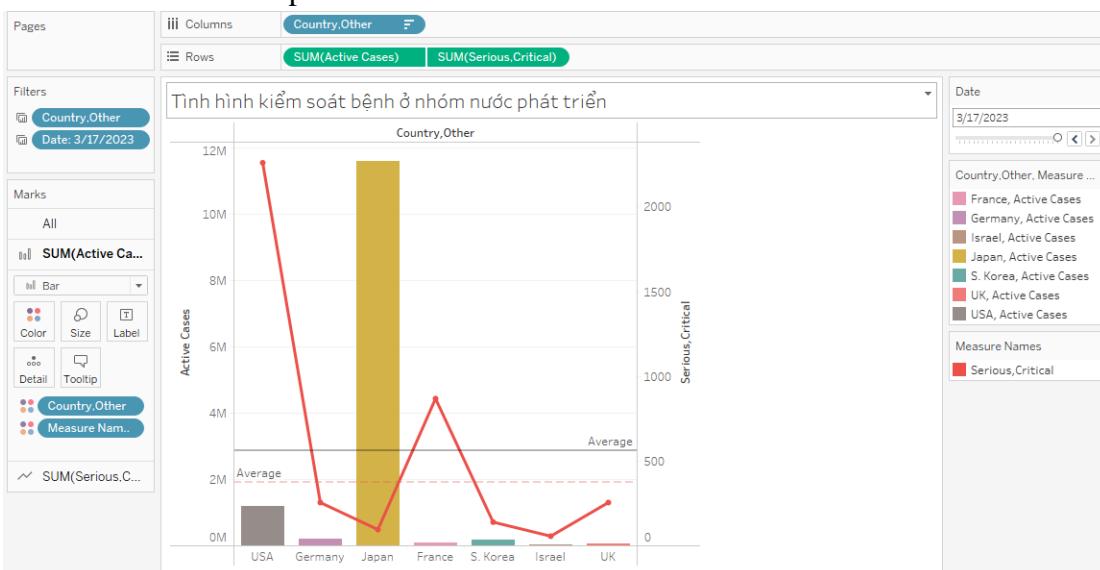
- Tùy chọn màu sắc cho các nước, em thả field “Country,Other” vào Color trong Marks Tag. Trong SUM(Serious,Critical) chọn tùy chọn Line. Nhóm em thu được kết quả như sau:



- Right-Click vào “Serious,Critical” axis, chọn Dual Axis. Tùy chọn lại màu sắc cho Line sang màu đỏ, vì em nghĩ nó sẽ làm nổi bật được số ca bệnh nặng.
- Đối với nhóm nước đang phát triển, nhóm em cũng làm tương tự.

Kết quả:

#### Nhóm nước phát triển:



#### Nhóm nước đang phát triển:



Đối với loại biểu đồ này, nhóm em cảm thấy chưa biếu thị được mối quan hệ rõ ràng giữa số ca tiên triển nặng và số ca đang mắc. Nên nhóm em đề xuất thêm 2 biếu đồ boxplot và bar chart về tỷ lệ số ca tiên triển nặng / số ca đang mắc.

## B2: Thiết kế Bar Chart và Boxplot:

Bar chart:

- Tạo một Calculated Field đơn giản về tỷ lệ trên:
- Kéo thả “Country,Other” vào Columns đã được lọc nhóm nước phát triển hoặc đang phát triển. Trong Rows nhóm em để Calculated Field vừa tạo ở trên có dạng AGG(Rate Serious). Tùy chọn màu sắc cho “Country,Other” để giống với Barchart+LineChart sao cho đồng màu nhau.

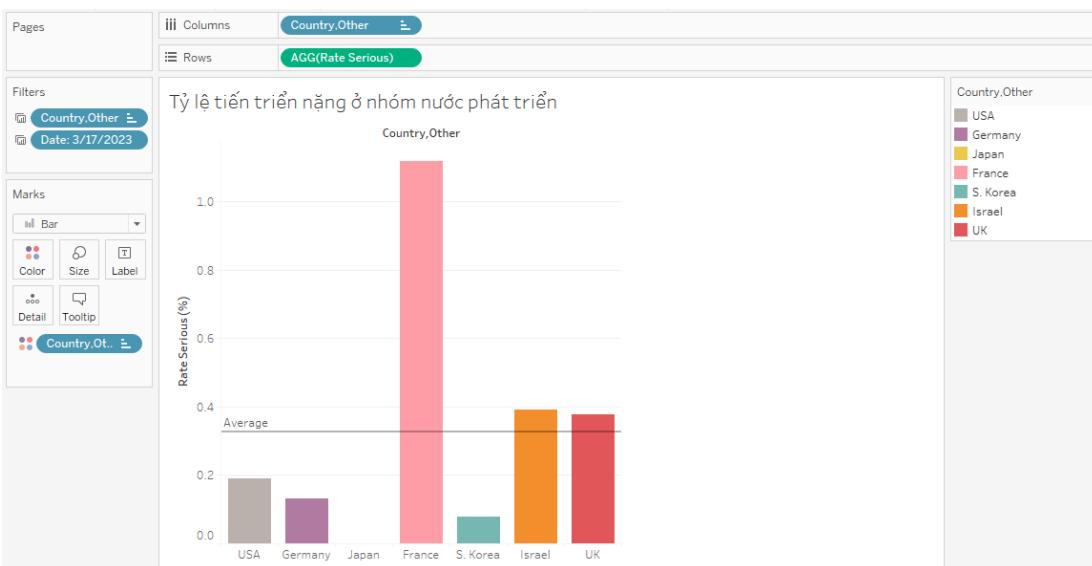
The calculation is valid.      7 Dependencies      Apply      OK

**ABS (number)**  
Returns the absolute value of the given number.  
Example: ABS (-7) = 7

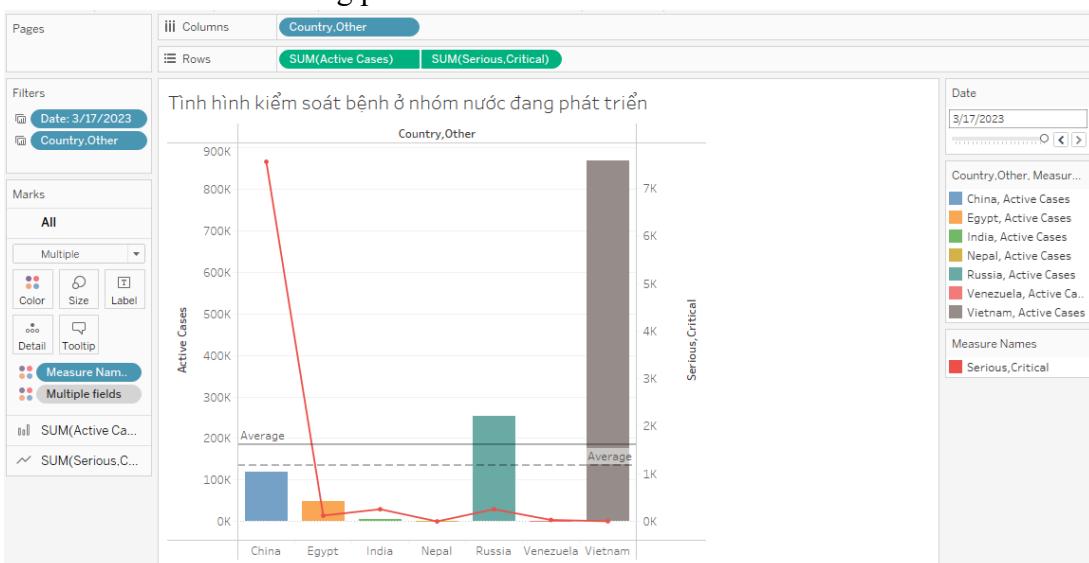
- Thêm Average Reference Line: chuột phải vào axis > Add Reference Line > Tại ô value chọn “Average”.

Kết quả:

Nhóm nước phát triển:



Nhóm nước đang phát triển:

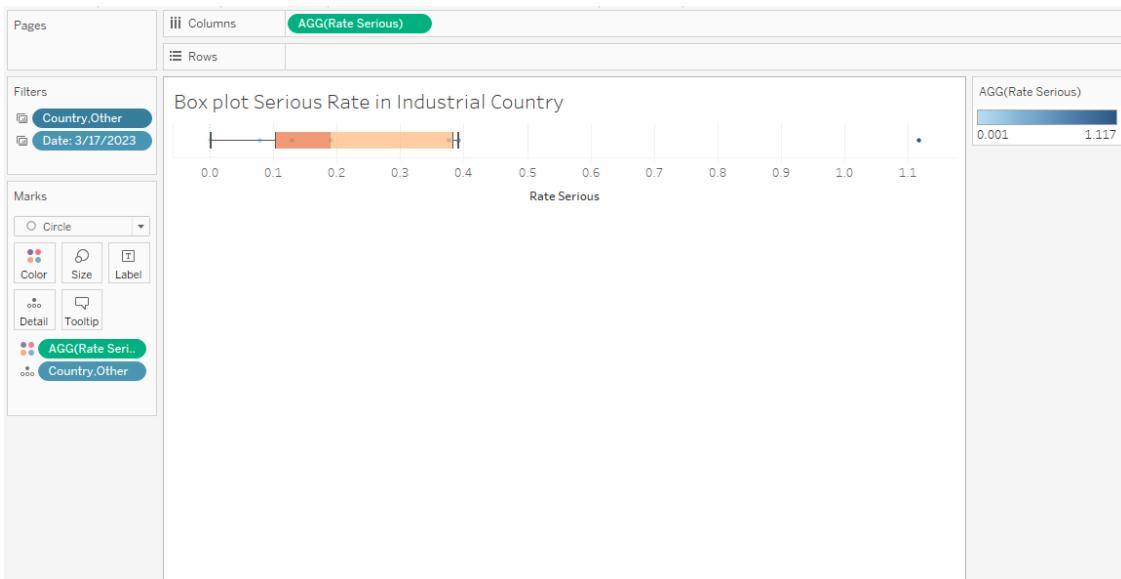


Boxplot: Thể hiện sự phân phối của tỷ lệ tiến triển nặng của Covid ở nhóm nước phát triển và đang phát triển.

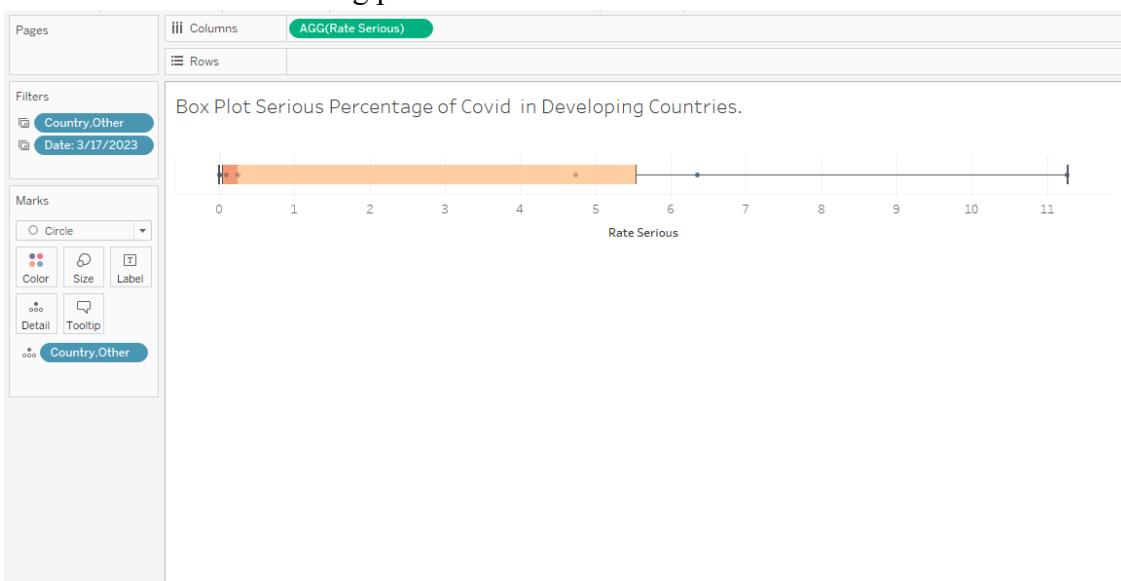
- Kéo thả Country,Other đã được chọn nhóm nước vào Column và AGG(Rate Serious) vào Rows.
- Trong tag Show me: Chọn Box plot.
- Chọn màu sắc: chuột phải vào boxplot vừa tạo > Edit... > Trong ô Fill chọn màu Orage.

Kết quả:

Nhóm nước phát triển:

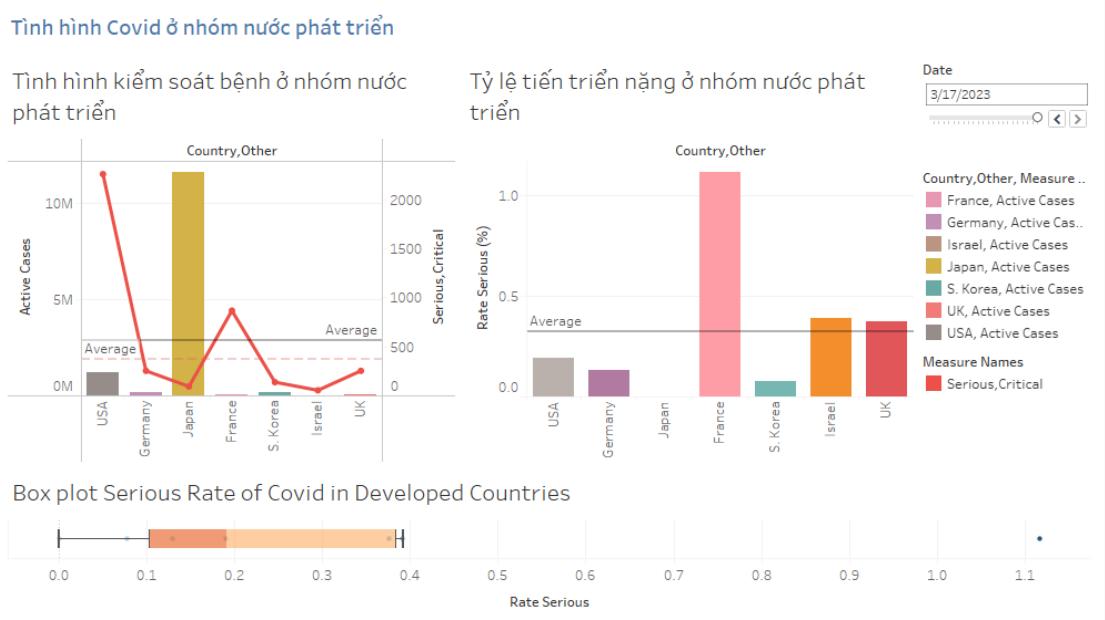


Nhóm nước đang phát triển:



### B3: Thiết kế Dashboard cho nhóm nước phát triển và đang phát triển.

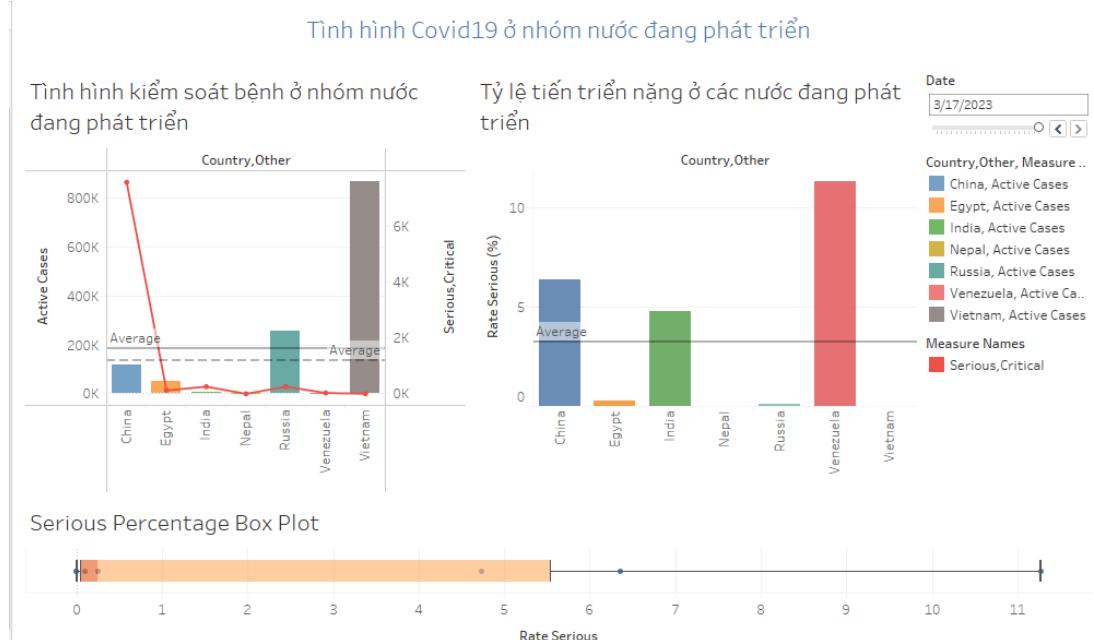
- Tạo dashboard, sau đó kéo thả các worksheet ta vừa làm ở trên vào dashboard.
- Nhóm nước phát triển:



## Nhận xét:

- Ta có thể thấy nhóm nước đang có nhiều ca bệnh nhất ở nước phát triển là Nhật bản, Ngược lại có Israel là thấp nhất.
- Số ca bệnh nặng nhiều nhất ở Mỹ và thấp nhất cũng ở Israel.
- Về tỷ lệ kiểm soát bệnh (tỷ lệ bệnh trở nặng) thông qua Rate Serious, đất nước có tỷ lệ bệnh nhân tiến triển nặng cao nhất là ở Pháp. Nhưng chỉ ở mức 1.1%. Thông qua boxplot có thể coi là 1 outlier.
- Nhìn chung qua boxplot, thấy tình hình kiểm soát bệnh ở các nước phát triển rất tốt. Duy trì ở mức 0.1-0.4%.
- Về màu sắc, các màu tương phản sáng tốt, tránh được tính đơn điệu.

Nhóm nước đang phát triển:



## Nhận xét:

- Ta có thể thấy nhóm nước đang có nhiều ca bệnh nhất ở nước đang phát triển là Việt Nam, ngược lại có Nepal là thấp nhất.
  - Số ca bệnh nặng nhiều nhất ở China và thấp nhất cũng ở Vietnam và Nepal.
  - Có thể thấy Nepal kiểm soát dịch bệnh cũng rất tốt cả về số ca mắc lẫn kỹ thuật điều trị.
  - Về tỷ lệ kiểm soát bệnh (tỷ lệ bệnh trở nặng) thông qua Rate Serious, đất nước có tỷ lệ bệnh nhân tiến triển nặng cao nhất là ở Venezuela. Ở mức ngạc nhiên khoảng 11%. Mặc dù có thể thấy ở số ca bệnh nặng và ca mắc đều thấp. Điều này cho thấy khả năng điều trị của Venezuela không ổn.
  - Nhìn chung qua boxplot, thấy tình hình kiểm soát bệnh ở các nước phát triển tốt. Nhưng không bằng các nước phát triển. Qua boxplot thấy dao động từ 0-6%.
- Về màu sắc, các màu tương phản sáng tốt, tránh được tính đơn điệu.

### b) So sánh tình hình dịch bệnh và khả năng kiểm soát bệnh ở nhóm nước phát triển và đang phát triển:

Sử dụng Parallel System và Pie Chart.

- Thiết kế Parallel: sử dụng các đường tượng trưng cho các nước (14 nước) và các yếu tố đánh giá dịch bệnh: “NewCases”, “NewDeaths”, “NewRecovered”, “Serious,Critical” và “ActiveCases” ở ngày gần nhất (17/03/2023). Sử dụng “Country,Other” và “Date” làm Filter.

B1: Chuẩn hóa dữ liệu theo min-max scale. Thêm các Calculated Field của các field liệt kê ở trên. Ví dụ với “ActiveCases”:



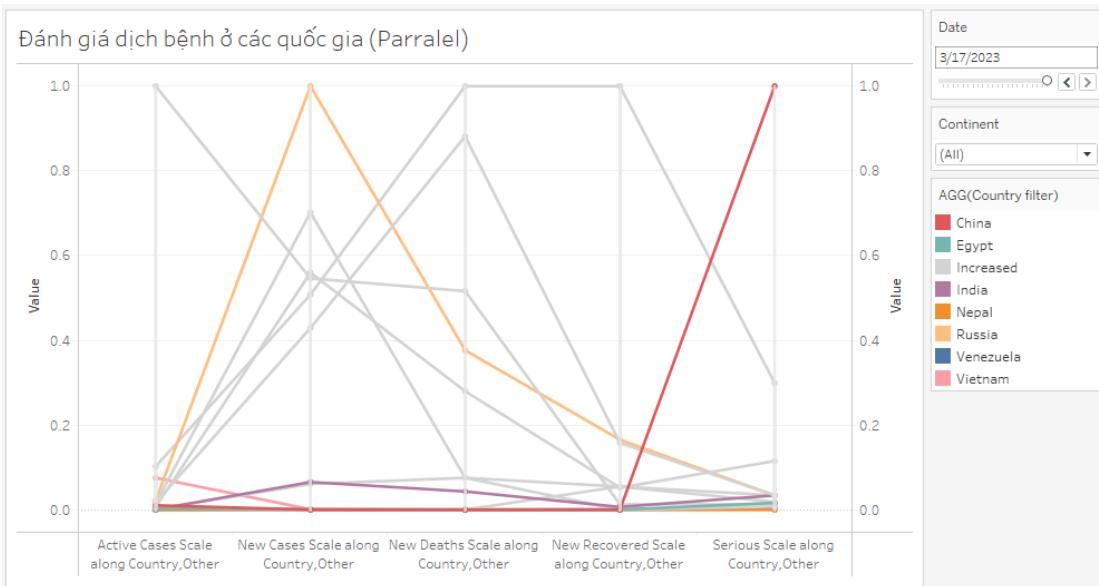
B2: Kéo thả Measure Values vào Rows, “Country,Other” vào Detail. Xóa đi các Measure Values không cần thiết (Chỉ chừa các Calculated Fields đã tạo ở trên). Sau đó duplicate Measures vừa tạo. Thêm Measures Name vào Columns.

B3: Chuột phải vào mỗi Calculated Fields ở Measure Values Tag > Chọn Compute Using > Country, Other.

B4: Tạo một calculated field để lọc màu cho các nước đang phát triển.



B5: Duplicate Measure Values ở Rows. Với Measure Values thứ nhất, thả Country filter vừa tạo ở B4 vào Color trong Marks Tag, chọn màu cho “IncreasedCountries” thành màu chìm và các nước khác màu nỗi; chọn Line với mark này. Với Measure Values thứ 2, kéo thả lại “Country,Other” vào path. Tùy chọn màu sắc giống với đường Grid và size nỗi hơn so với Grid. Right-click measure thứ 2 và chọn Dual Axis. Ta được kết quả như sau.



B6: Ta có thể thêm Continents vào Filter để so sánh dịch bệnh giữa các châu lục.

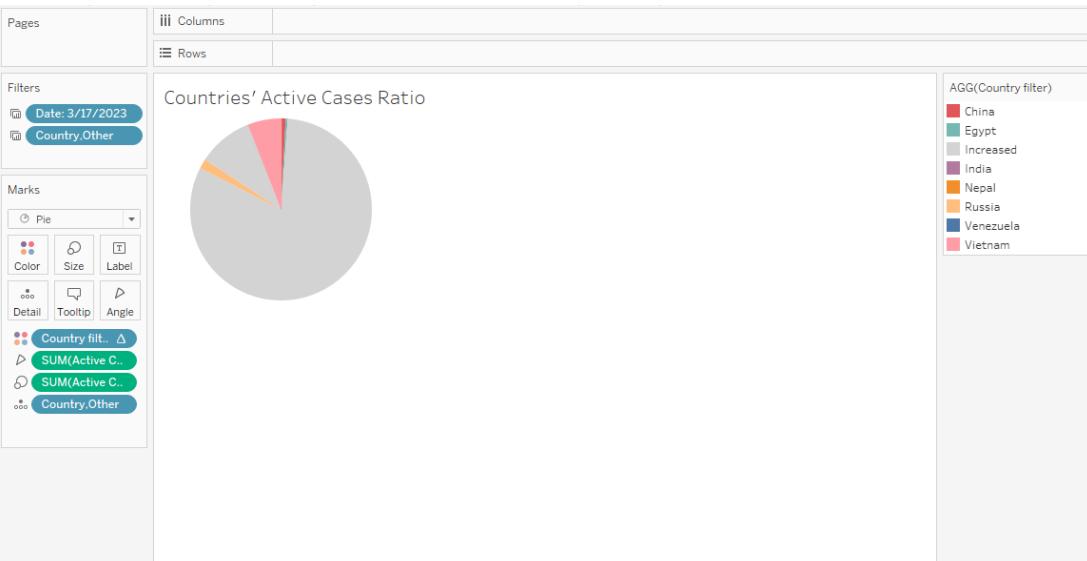
- Thiết kế Pie chart: Để thấy rõ hơn tỷ trọng giữa các nước phát triển và các nước đang phát triển.

B1: Kéo thả “Country,Other” vào Columns, “Active Cases” vào Rows.

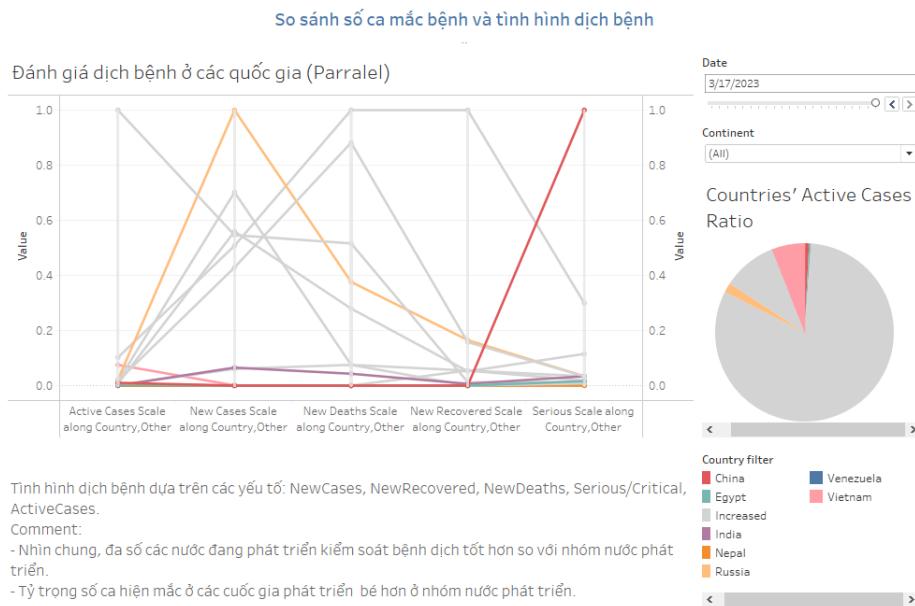
B2. Trong Show me Tag chọn Pie Chart.

B3. Thả “Country,Other” lại vào Detail trong Marks Tag, Country Filter tạo ở trên vào Color trong Marks Tag.

Kết quả:



Thiết kế Dashboard dựa trên 2 worksheets vừa tạo:



### Nhận xét:

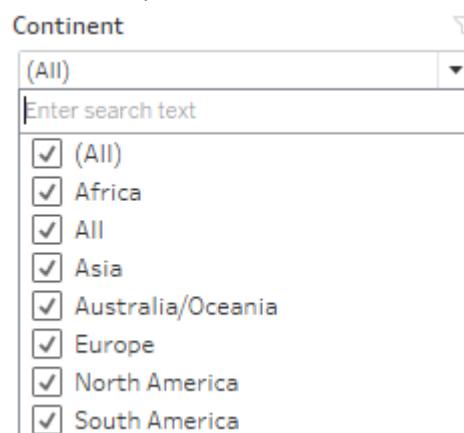
- Nhìn chung, đa số các nước đang phát triển kiểm soát bệnh dịch tốt hơn so với nhóm nước phát triển.
  - Đối với Parallel, các nước đang phát triển đều nằm dưới các nước phát triển (đường xám).
  - Tỷ trọng số ca hiện mắc ở các quốc gia phát triển bé hơn ở nhóm nước phát triển.
  - Vì chúng ta ở Việt Nam nên nhóm em quyết định áp dụng màu chìm cho những nước phát triển, và áp dụng những màu nổi bật cho nhóm đang phát triển. Đồng thời tránh tính đơn điệu cho biểu đồ và nhìn đỡ rối hơn so với trước.

#### 4.5.3. Kỹ thuật sử dụng:

- **Manipulate view:** Người quan sát có thể xem các view khác nhau bằng cách xem các ngày khác nhau (**Change view over time**) => Đánh giá trực quan hơn khi xem dữ liệu của nhiều ngày + thứ hạng thay đổi của các nước.



- **Facet:** Xem các biểu đồ khác nhau để hiểu rõ hơn về dữ liệu được trực quan, một biểu đồ có thể sẽ không trình bày rõ tình hình dịch bệnh nên cần các biểu đồ khác.
- **Reduce:** Sử dụng Filter cho biểu đồ, ví dụ như Continent để giảm số nước trên biểu đồ và xem cụ thể các nước ở mỗi châu lục.



## 4.6. Tốc độ lây lan của virus Covid-19 có khác biệt giữa các khu vực không? Những khu vực nào đang ghi nhận tốc độ lây lan cao nhất?

### 4.6.1. Tiếp cận vấn đề

Khảo sát số ca nhiễm mới tính đến thời điểm gần nhất của bộ dữ liệu (17/03/2023). Từ đó có được cái nhìn mới nhất về tình hình dịch Covid-19 trong thời gian gần đây.

### 4.6.2. Trực quan bằng Tableau

#### a) Số ca nhiễm Covid-19 trên thế giới thời gian gần đây

Để thực hiện trực quan, sử dụng biểu đồ Filled Map trong Tableau

B1:

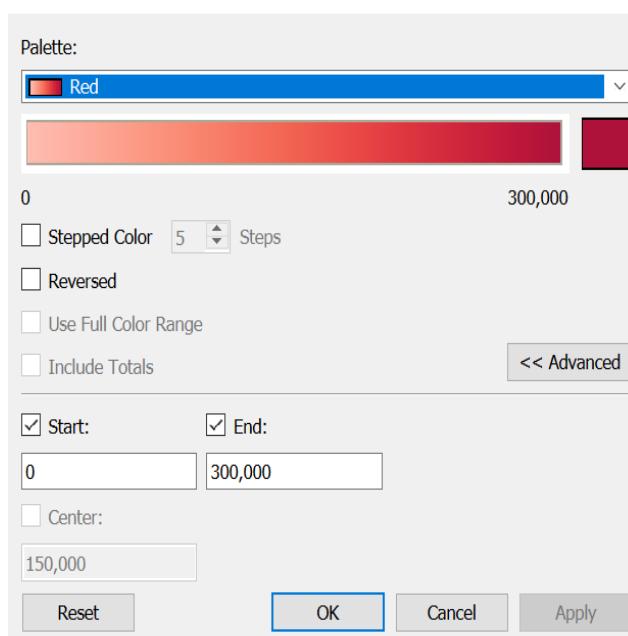
- Kéo thả field “NewCases” vào Columns, tính Sum cho field “New Cases” (Measure: Sum).

SUM(New Cases)

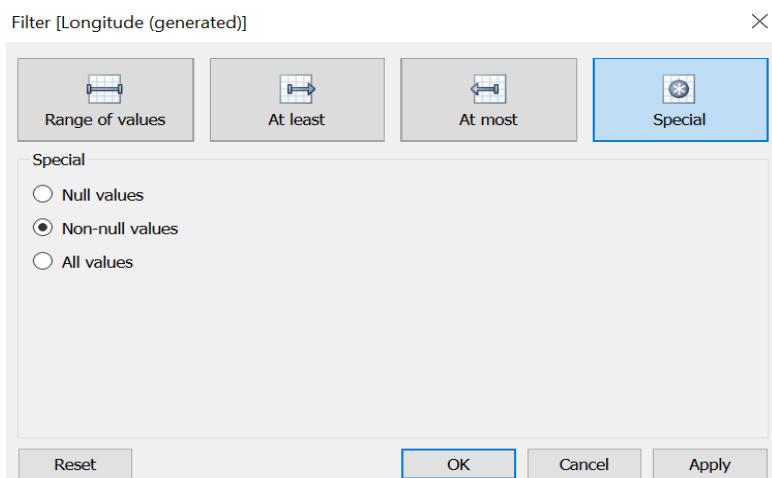
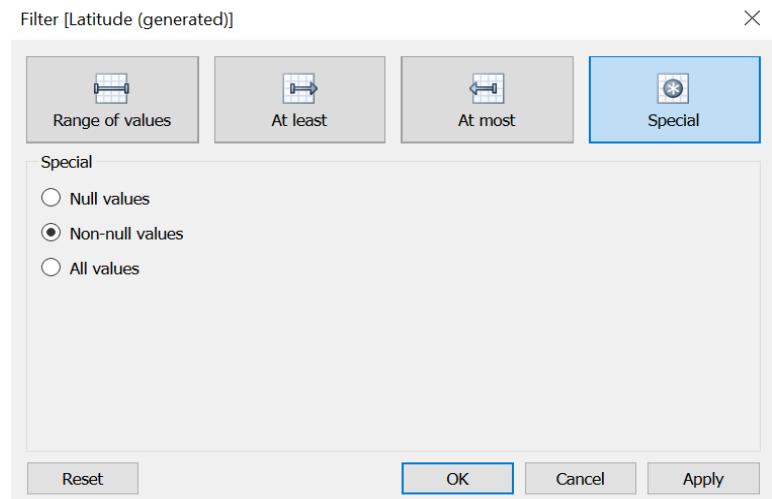
- Filter color trong field “NewCases”, chỉnh sửa màu sắc cho biểu đồ thành màu đỏ, với giá trị bắt đầu của dải màu là 0 và kết thúc là 300000

Edit Colors [New Cases]

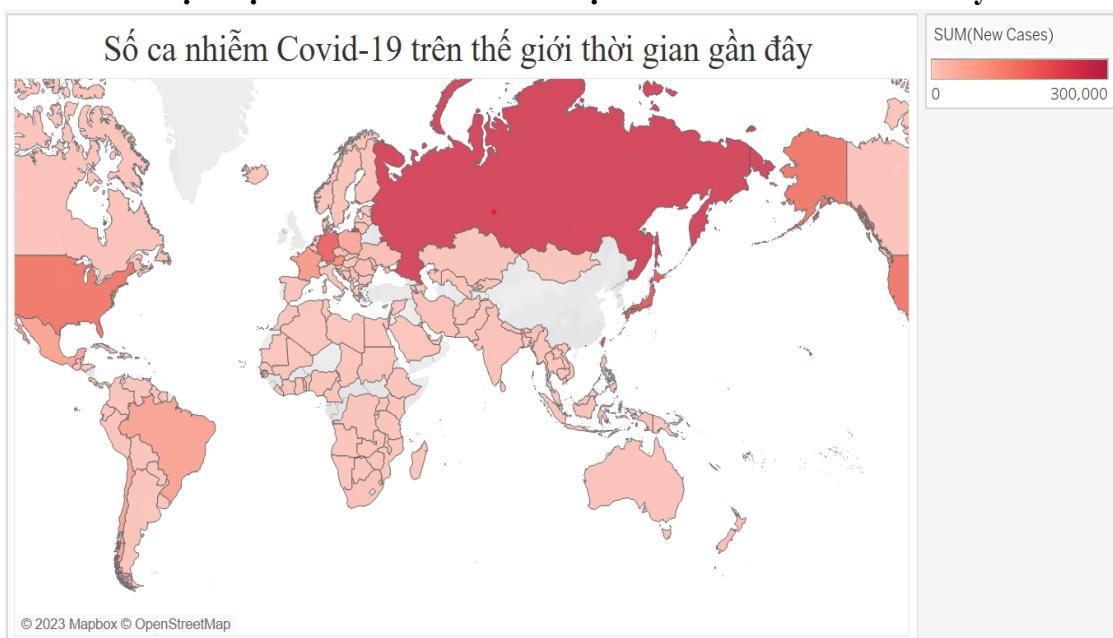
X



B2: Kéo thả field “Country/Other” vào Rows, lúc này sẽ tự động xuất hiện kinh độ và vĩ độ của các quốc gia và 2 trường này có filter là sử dụng giá trị Non-null



Sau khi thực hiện các bước trên ta sẽ được biểu đồ như hình dưới đây:



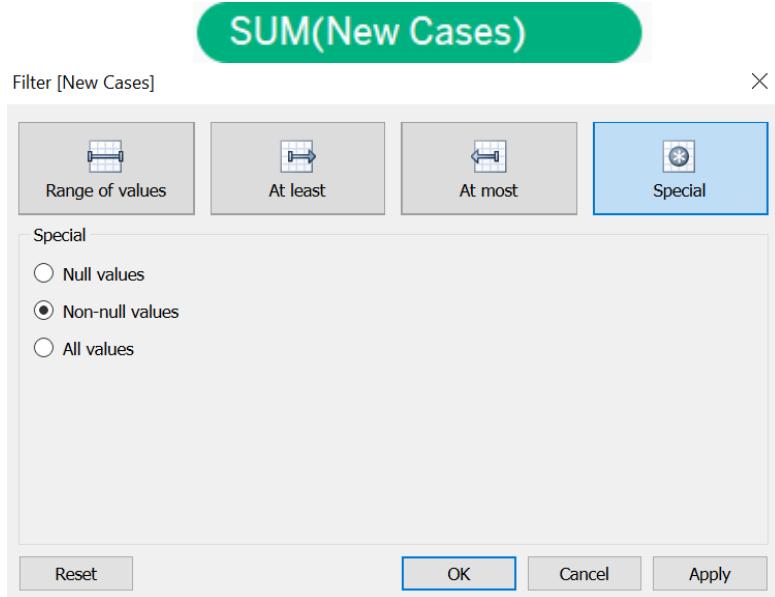
### Nhận xét:

- Việc sử dụng biểu đồ filled map, để người dùng có thể nhìn trực tiếp số ca nhiễm trên bản đồ thế giới và quốc gia tương ứng.
- Màu sắc được sử dụng là màu đỏ, một màu phổ biến trong hệ màu và màu đỏ sẽ làm nổi bật biểu đồ.

- b) Tốc độ lây lan của dịch Covid-19 theo thời gian đối với từng khu vực trên thế giới

Để thực hiện trực quan, sử dụng Line chart trong Tableau

B1: Kéo thả field “NewCases” vào Columns, tính Sum và filter chỉ sử dụng các giá trị Non-null



B2: Kéo thả field “Continent” vào Rows, chọn các khu vực cần thể hiện và chọn màu sắc cho các khu vực .

Filter [Continent] X

General	Wildcard	Condition	Top																
<input checked="" type="radio"/> Select from list <input type="radio"/> Custom value list <input type="radio"/> Use all <span style="float: right;">☰</span>																			
<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <input style="width: 100%; height: 25px; border: none; margin-bottom: 5px;" type="text"/> <span style="font-size: 10px; color: #ccc;">Enter search text</span> </div> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td><input type="checkbox"/></td><td>Null</td></tr> <tr><td><input checked="" type="checkbox"/></td><td>Africa</td></tr> <tr><td><input type="checkbox"/></td><td>All</td></tr> <tr><td><input checked="" type="checkbox"/></td><td>Asia</td></tr> <tr><td><input checked="" type="checkbox"/></td><td>Australia/Oceania</td></tr> <tr><td><input checked="" type="checkbox"/></td><td>Europe</td></tr> <tr><td><input checked="" type="checkbox"/></td><td>North America</td></tr> <tr><td><input checked="" type="checkbox"/></td><td>South America</td></tr> </table>				<input type="checkbox"/>	Null	<input checked="" type="checkbox"/>	Africa	<input type="checkbox"/>	All	<input checked="" type="checkbox"/>	Asia	<input checked="" type="checkbox"/>	Australia/Oceania	<input checked="" type="checkbox"/>	Europe	<input checked="" type="checkbox"/>	North America	<input checked="" type="checkbox"/>	South America
<input type="checkbox"/>	Null																		
<input checked="" type="checkbox"/>	Africa																		
<input type="checkbox"/>	All																		
<input checked="" type="checkbox"/>	Asia																		
<input checked="" type="checkbox"/>	Australia/Oceania																		
<input checked="" type="checkbox"/>	Europe																		
<input checked="" type="checkbox"/>	North America																		
<input checked="" type="checkbox"/>	South America																		
<input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px; margin-right: 10px;" type="button" value="All"/> <input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px;" type="button" value="None"/>		<input type="checkbox"/> Exclude																	
<b>Summary</b> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 5px;">       Field: [Continent]        Selection: Selected 6 of 8 values        Wildcard: All        Condition: None        Limit: None     </div>																			
<input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px;" type="button" value="Reset"/>		<input style="width: 100px; height: 25px; border: 1px solid #0070C0; border-radius: 5px; background-color: #0070C0; color: white; font-size: 10px;" type="button" value="OK"/>	<input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px;" type="button" value="Cancel"/>																

Edit Colors [Continent] X

<b>Select Data Item:</b> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;"> <input type="checkbox"/> Africa  <input type="checkbox"/> Asia  <input type="checkbox"/> Australia/Oceania  <input type="checkbox"/> Europe  <input type="checkbox"/> North America  <input type="checkbox"/> South America       </div>	<b>Select Color Palette:</b> <div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 10px;">       Automatic <span style="float: right;">▼</span> </div> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 50%; height: 50px; background-color: #4F81BD;"></td><td style="width: 50%; height: 50px; background-color: #E6A239;"></td></tr> <tr><td style="width: 50%; height: 50px; background-color: #FF8C00;"></td><td style="width: 50%; height: 50px; background-color: #9370DB;"></td></tr> <tr><td style="width: 50%; height: 50px; background-color: #E63366;"></td><td style="width: 50%; height: 50px; background-color: #FFB6C1;"></td></tr> <tr><td style="width: 50%; height: 50px; background-color: #4FC3F7;"></td><td style="width: 50%; height: 50px; background-color: #A08060;"></td></tr> <tr><td style="width: 50%; height: 50px; background-color: #2ECC71;"></td><td style="width: 50%; height: 50px; background-color: #BDBDBD;"></td></tr> </table> <div style="border: 1px solid #ccc; padding: 5px; text-align: center; margin-top: 10px;"> <input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px;" type="button" value="Assign Palette"/> </div>										
<input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px;" type="button" value="Reset"/>											
<input style="width: 100px; height: 25px; border: 1px solid #0070C0; border-radius: 5px; background-color: #0070C0; color: white; font-size: 10px;" type="button" value="OK"/>											
<input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px;" type="button" value="Cancel"/>											
<input style="width: 100px; height: 25px; border: 1px solid #ccc; border-radius: 5px; background-color: #f0f0f0; font-size: 10px;" type="button" value="Apply"/>											

B3: Kéo thả field “Day” vào Columns, chỉnh thành định dạng Day (theo ngày tháng năm) và chọn khoảng thời gian (từ 24/2-17/3)

ISO-8601 Week-Based

Year	2015
Quarter	Q2
Month	May
Day	8
More	▶

Year	2015
Quarter	Q2 2015
Month	May 2015
Week Number	Week 5, 2015
Day	May 8, 2015
More	▶

Exact Date

Attribute

Measure	▶
---------	---

Discrete

Continuous	●
------------	---

**Edit in Shelf**

Remove

Filter [Day of Date] X

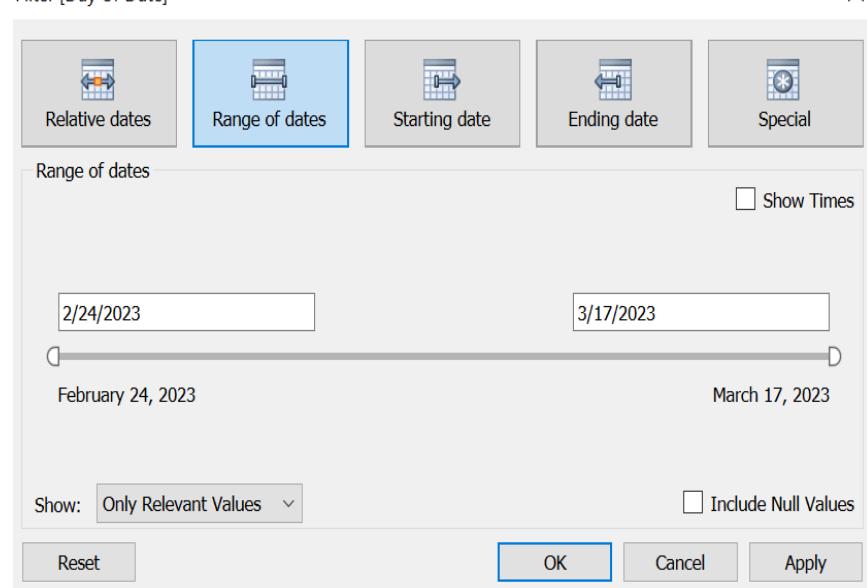
Range of dates  Show Times

2/24/2023 3/17/2023

February 24, 2023 March 17, 2023

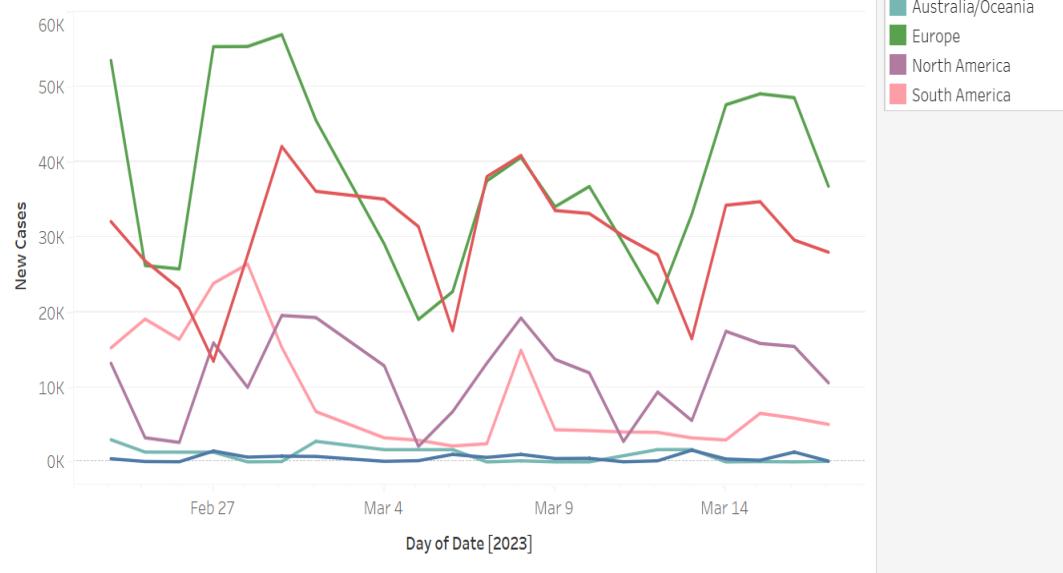
Show: Only Relevant Values  Include Null Values

OK Cancel Apply



Sau khi thực hiện các bước trên ta sẽ được biểu đồ dưới đây:

## Tốc độ lây lan của dịch Covid-19 theo thời gian đối với từng khu vực trên thế giới



### Nhận xét:

- Việc sử dụng biểu đồ Line chart giúp ta hiểu được xu hướng và sự thay đổi của các ca nhiễm mới trong khoảng thời gian gần đây trên các khu vực địa lý khác nhau.
- Màu sắc được sử dụng là các màu cơ bản, mỗi khu vực tương ứng với một màu sắc khác nhau.

### c) Xây dựng Dashboard từ 2 biểu đồ trên

B1: Kéo thả 2 sheets thể hiện 2 biểu đồ vào dashboard.

B2: Tùy chỉnh vị trí và kích cỡ bằng cách Floating các biểu đồ

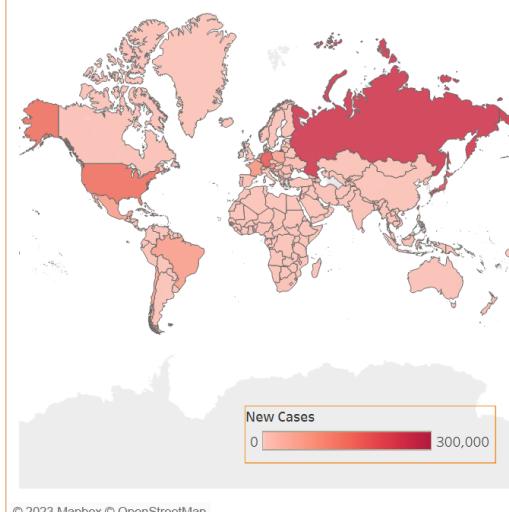
B3: Thêm nhận xét (add text) sau khi quan sát biểu đồ, có thể tùy chỉnh vị trí và kích cỡ bằng Floating và edit text.

### Objects

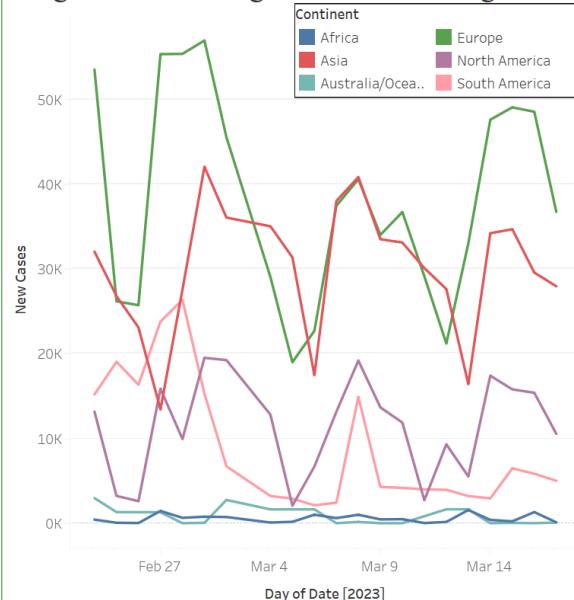
- Horizontal Container
- Vertical Container
- Text
- Extension
- Ask Data
- Data Story
- Image

Sau khi thực hiện các bước trên ta sẽ được Dashboard dưới đây:

Số ca nhiễm Covid-19 trên thế giới thời gian gần đây



Tốc độ lây lan của dịch Covid-19 theo thời gian đối với từng khu vực trên thế giới



### Nhận xét:

- Nhìn vào biểu đồ có thể thấy các quốc gia có màu sắc đỏ đậm hơn là các quốc gia có số ca nhiễm Covid-19 trong thời gian gần đây là nhiều nhất, ví dụ như Nga, Đài Loan, Nhật Bản, Đức, Mỹ.
- Tốc độ lây lan ở mỗi khu vực trên thế giới là khác nhau, trong đó châu Âu và châu Á hiện là hai khu vực có số ca nhiễm mới hàng ngày cao nhất. Điều này có thể do hai khu vực này là nơi sinh sống của nhiều quốc gia và đông dân cư trên thế giới.
- Ngoài ra, Bắc Mỹ cũng có tốc độ lây lan cao, với Hoa Kỳ là một trong những quốc gia trên thế giới có số ca mắc mới được xác nhận mỗi ngày cao nhất.

#### 3.6.3 Các kỹ thuật được sử dụng

- Manipulate view:** Người xem có thể thay đổi các view khác nhau  
VD: Xem trên biểu đồ Filled Map, Nước Nga có bao nhiêu ca nhiễm bệnh mới trong thời gian gần đây.

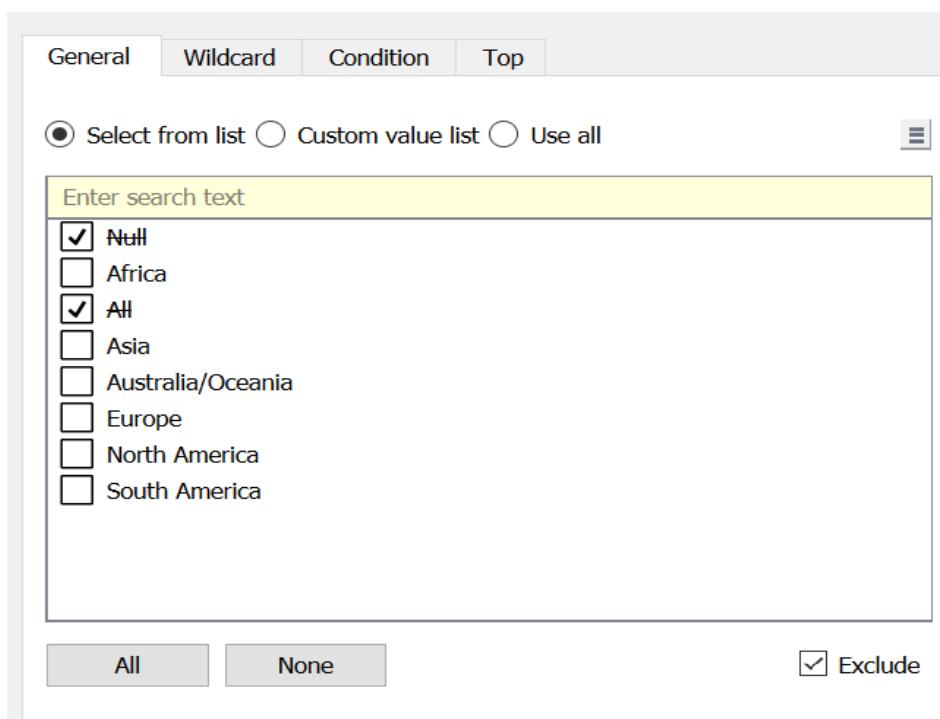


- **Facet:** Xem các biểu đồ khác nhau để hiểu rõ hơn về dữ liệu được trực quan, một biểu đồ có thể sẽ không trình bày rõ tình hình dịch bệnh nên cần các biểu đồ khác.
- **Reduce:** Sử dụng filter để lọc bỏ và chỉ chọn những giá trị cần thiết để thực hiện trực quan

**VD:** Không thể hiện trực quan các giá trị của “NewCases” cho **Null** và **World** trong filter “Continent”.

Filter [Continent]

X



## 4.7. Có mối tương quan nào giữa mật độ dân số của một quốc gia và số ca nhiễm COVID-19 của quốc gia đó không?

### 4.7.1 Tiếp cận vấn đề

Khảo sát tổng số ca nhiễm và dân số các nước tính đến thời điểm gần nhất của bộ dữ liệu (17/03/2023). Từ đó tìm được mối liên hệ giữa 2 thuộc tính và có được cái nhìn toàn diện hơn về dịch Covid-19.

### 4.7.2 Trực quan bằng Tableau

#### a) Top 10 quốc gia có tổng số ca nhiễm Covid-19 cao nhất thế giới

Để thực hiện trực quan, sử dụng biểu đồ Bar chart trong Tableau

B1: Kéo thả field “TotalCases” vào Columns, tính sum cho thuộc tính.

SUM(Total Cases)

B2: Kéo thả field “Country/Other” vào Rows , chọn filter lấy **top 10** quốc gia và loại bỏ các giá trị thuộc về **null** và **world** .

Filter [Country,Other] X

General    Wildcard    Condition    Top

None  
 By field:

Top  10  by  
Total Cases  Sum

Filter [Country,Other] X

General    Wildcard    Condition    Top

Select from list  Custom value list  Use all

Enter Text to Search or Add

Null  
World

B3: Filter cho “Day” lấy các giá trị ngày 17/3, chọn màu và kích thước của biểu đồ .

Filter [Day of Date] X

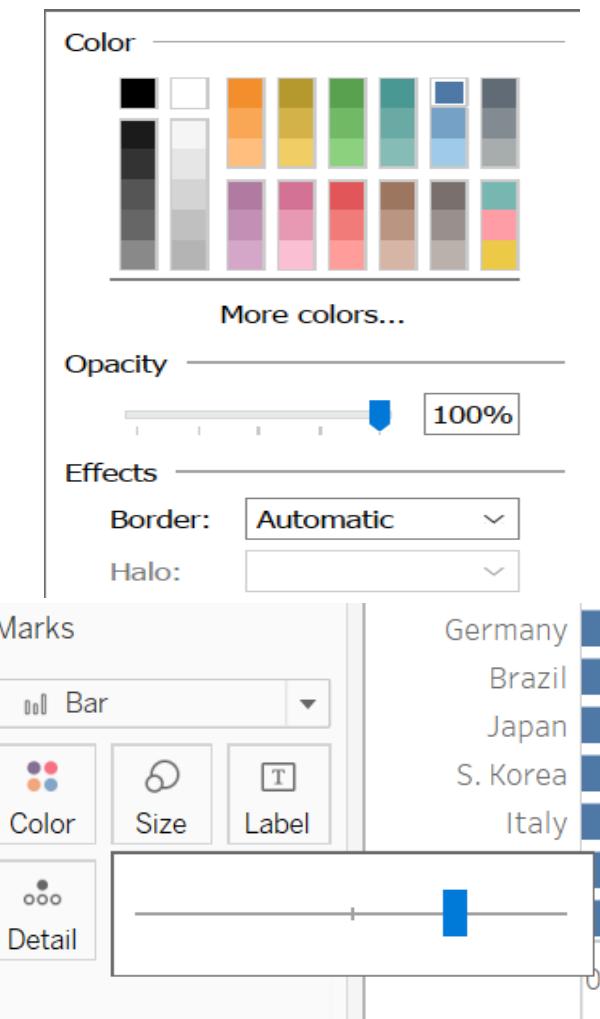
General    Condition    Top

Select from list  Custom value list  Use all

Enter Text to Add

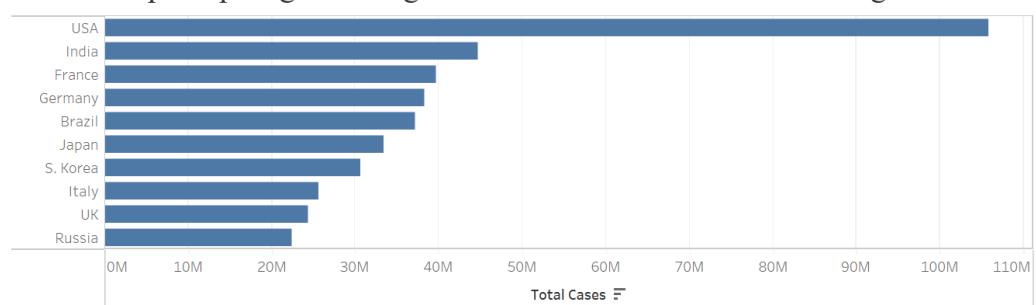
March 17, 2023

Include all values when empty  Exclude



Sau khi thực hiện các bước trên ta sẽ được biểu đồ như dưới đây:

Top 10 quốc gia có tổng số ca nhiễm Covid-19 cao nhất thế giới



### Nhận xét:

- Việc sử dụng biểu đồ Bar chart để thấy được sự chênh lệch về tổng số ca nhiễm giữa các quốc gia, thanh dài hơn thể hiện quốc gia đó có tổng số ca nhiễm nhiều hơn và ngược lại.
- Màu xanh được sử dụng là một màu dễ nhìn và là màu phổ biến trong bảng màu.

b) **Mối tương quan giữa mật độ dân số và tổng số ca nhiễm Covid-19**

Để thực hiện trực quan, sử dụng Scatter plot trong Tableau

B1: Kéo thả field “TotCase/1M pop” vào Columns và tính log.

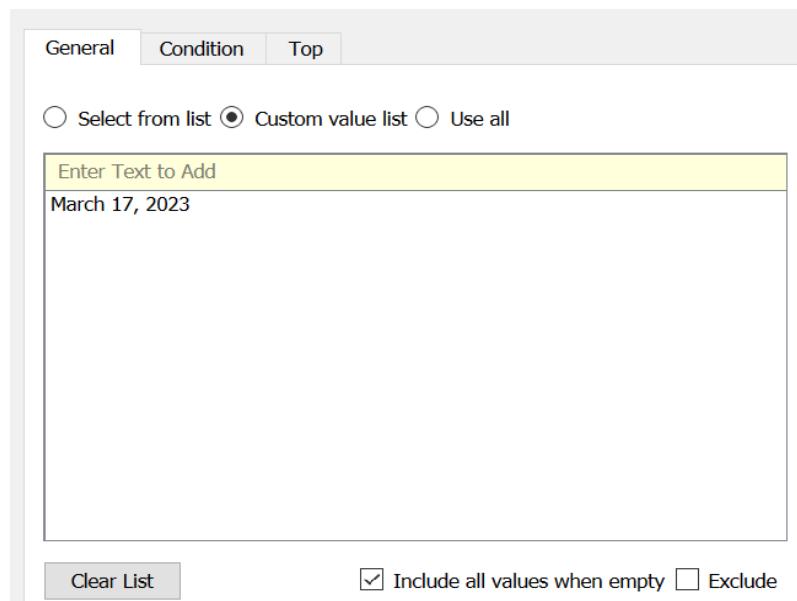
B2: Kéo thả field “Population” vào Rows và lấy log.



B3: Chọn filter “Day” ngày 17/3 và chọn filter “Country/Other” bỏ đi giá trị **null** và **world**

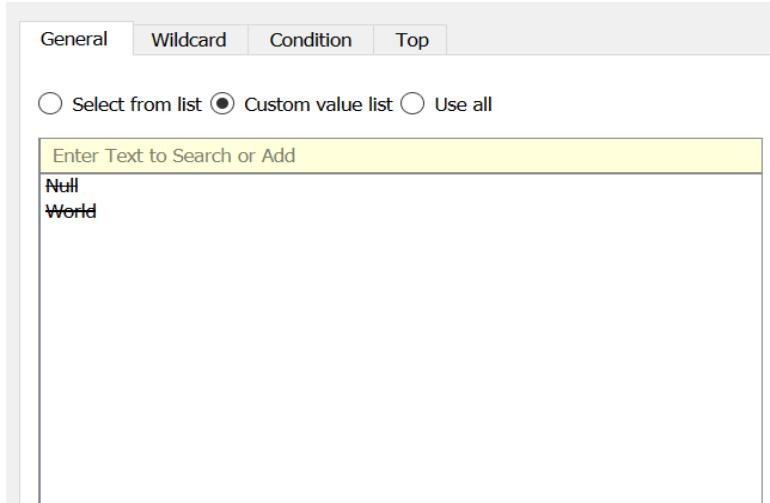
Filter [Day of Date]

X

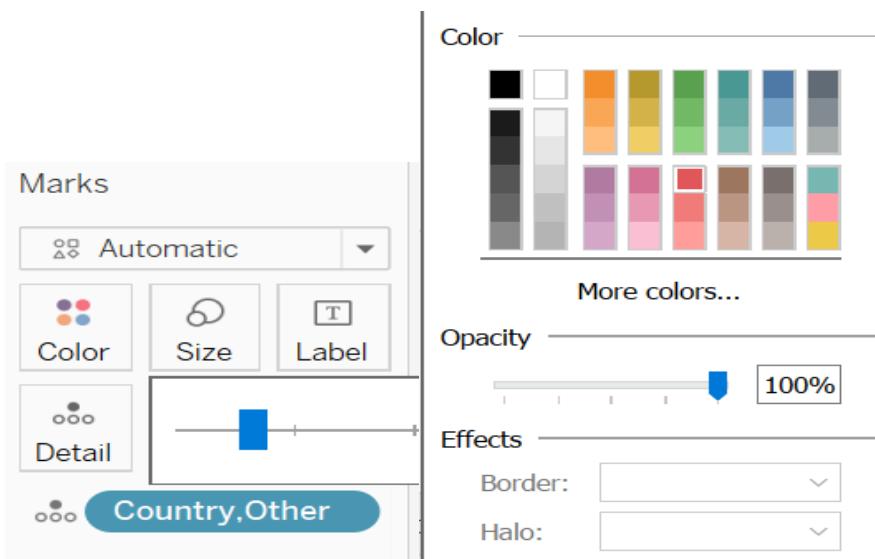


Filter [Country,Other]

X

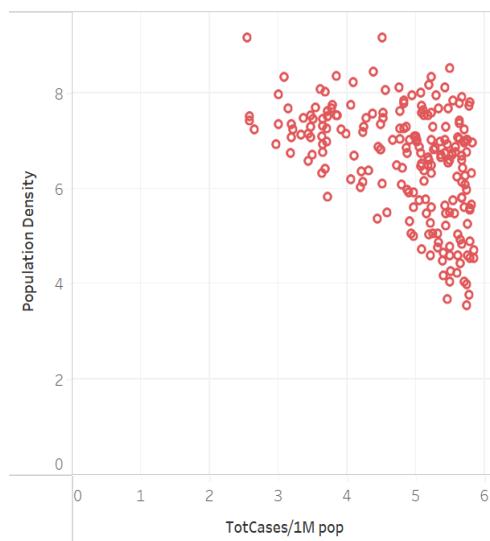


B4: Chính sửa kích thước của các điểm dữ liệu trên biểu đồ và sử dụng màu đỏ làm màu thể hiện



Sau khi thực hiện các bước trên ta sẽ được biểu đồ dưới đây:

Mối tương quan giữa mật độ dân số và tổng số ca nhiễm Covid-19



#### Nhận xét:

- Việc sử dụng biểu đồ scatter plot để thể hiện mối tương quan chính xác và rõ ràng hơn giữa 2 biến là “Population” và “TotalCases/1M pop”
- Màu đỏ được sử dụng là một màu phổ biến trong bảng màu, làm nổi bật hình ảnh trực quan.

#### c) Xây dựng Dashboard từ 2 biểu đồ trên

B1: Kéo thả 2 sheets thể hiện 2 biểu đồ vào dashboard.

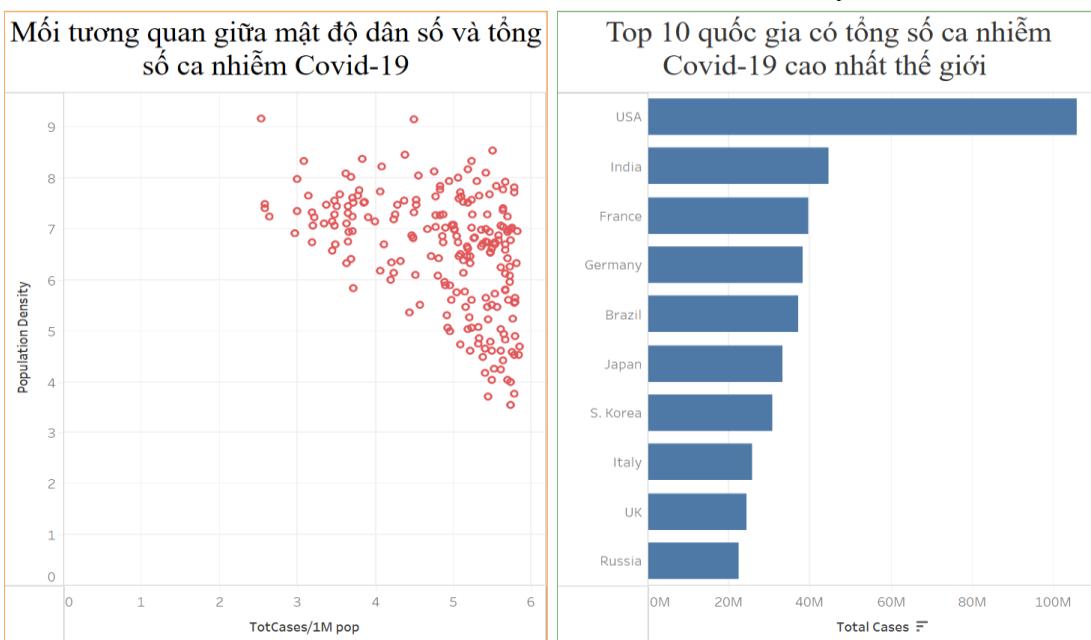
B2: Tùy chỉnh vị trí và kích cỡ bằng cách Floating các biểu đồ

B3: Thêm nhận xét (add text) sau khi quan sát biểu đồ, có thể tùy chỉnh vị trí và kích cỡ bằng Floating và edit text.

## Objects

- Horizontal Container
- Vertical Container
- Text
- Extension
- Ask Data
- Data Story
- Image

Sau khi thực hiện các bước trên ta sẽ được Dashboard dưới đây:



### Nhận xét:

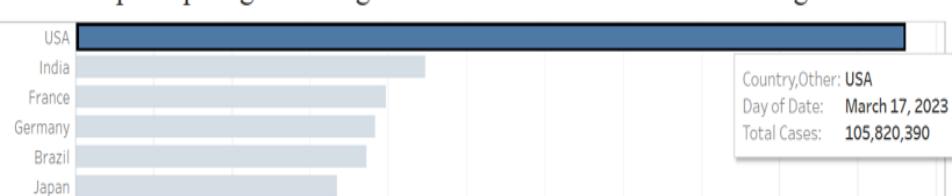
- Có thể có mối tương quan giữa mật độ dân số và số ca mắc COVID-19 ở một quốc gia. Với mật độ dân số cao, có thể có nhiều sự tiếp xúc gần gũi hơn giữa các cá nhân, dẫn đến khả năng lây truyền vi-rút cao hơn. Tuy nhiên, sự khác biệt này còn phụ thuộc vào các biện pháp được thực hiện để ngăn chặn sự lây lan của dịch bệnh và khả năng kiểm soát của chính phủ.
- Ví dụ, Ấn Độ có mật độ dân số cao và hiện đang có số ca mắc COVID-19 rất lớn. Tuy nhiên, các quốc gia khác có mật độ dân số cao như UK đang có số ca nhiễm thấp hơn đáng kể.

### 4.7.3 Các kĩ thuật được sử dụng

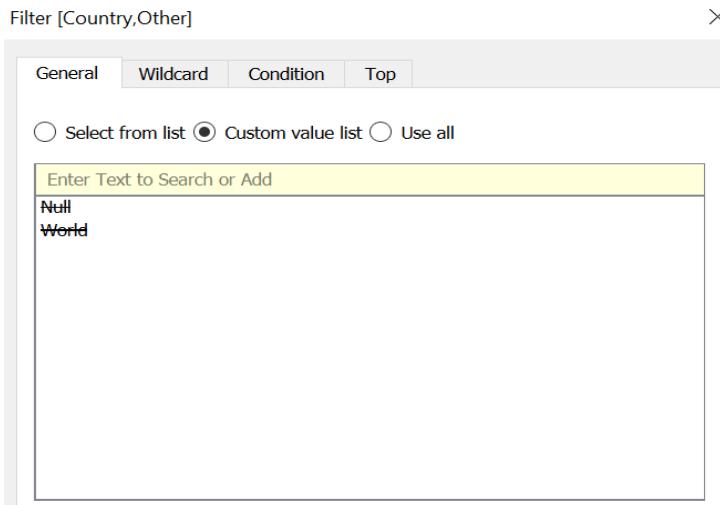
- **Manipulate view:** Người xem có thể thay đổi các view khác nhau

VD: Xem trên biểu đồ Bar chart, USA có tổng bao nhiêu ca nhiễm bệnh tính đến thời điểm hiện tại.

Top 10 quốc gia có tổng số ca nhiễm Covid-19 cao nhất thế giới



- **Facet:** Xem các biểu đồ khác nhau để hiểu rõ hơn về dữ liệu được trực quan, một biểu đồ có thể sẽ không trình bày rõ tình hình dịch bệnh nên cần các biểu đồ khác.
- **Reduce:** Sử dụng filter để lọc bỏ và và chỉ chọn những giá trị cần thiết để thực hiện trực quan  
VD: Không thể hiện trực quan các giá trị của “TotalCases” cho **Null** và **World** trong filter “Country/Other”.



## 5. Áp dụng một số kỹ thuật máy học để hiểu rõ hơn về dữ liệu.

### 5.1. Tiếp cận vấn đề.

Để hiểu rõ hơn về dữ liệu, nhóm em áp dụng một số thuật toán học máy đơn giản như Regression Tree (Decision Tree) và Linear Regression để dự đoán số ca nhiễm mới (NewCases), đồng thời đánh giá hiệu quả của mô hình.

Các yếu tố (biến) dùng để dự đoán NewCases: TotalCases, TotalDeaths, NewDeaths, TotalRecovered và NewRecovered. Cơ sở lựa chọn các biến này dựa trên tương quan dương (correlation coefficient  $> 0$ ) mà nhóm em đã phân tích ở Lab01.

Yêu cầu:

- Cài đặt và sử dụng tabpy (trình bày ở mục 1), đã kết nối với Tableau.
- Ngôn ngữ python, các thư viện numpy, pandas, sklearn, tabpy,...

Về Dữ Liệu:

Dữ liệu về dịch bệnh Coronavirus từ ngày 24/02/2023 đến ngày 17/03/2023. Do đó, để chuẩn bị và đánh giá Model, nhóm em chia Data thành 2 bộ Dataset:

- Train.csv: Dùng để train cho mô hình, từ ngày 24/02/2023  $\Rightarrow$  10/03/2023.
- Test.csv: Dùng để dự đoán số ca nhiễm mới, từ ngày 11/03/2023  $\Rightarrow$  17/03/2023.
- Sử dụng dataset "train.csv" và chia thành 2 bộ nhỏ hơn train set và validation set, Để train mô hình và kiểm tra để đánh giá mô hình. Sau đó áp dụng cho bộ dữ liệu test set trong "test.csv". Được trình bày trong file "preprocess.ipynb".

Sử dụng dataset "train.csv" và chia thành 2 bộ nhỏ hơn train set và validation set, Để train mô hình và kiểm tra để đánh giá mô hình. Sau đó áp dụng cho bộ dữ liệu test set trong "test.csv". Được trình bày trong file "preprocess.ipynb"

Xử lý dữ liệu và hàm mô hình: file "preprocess.ipynb" và "model.py". Sau khi xử lý dữ liệu trong file "preprocess.ipynb", kết quả dự đoán của mô hình được lưu vào file "result.csv".

Strategy:

Sử dụng R-square để đánh giá mô hình. Đối với mô hình có R-square < 0.5 cho thấy mô hình đó không hiệu quả và phù hợp với dữ liệu. Mô hình có R-square càng cao thì khả năng dự đoán chính xác hơn (mô hình phù hợp hơn).

### a. Cài đặt và nạp dữ liệu:

Sau khi thiết kế mô hình (trình bày ở mục a và b), ta cần deploy để nạp mô hình vào tabpy.

```
client = Client('http://localhost:9004/')
client.deploy('New_Cases_Covid_prediction_Linear',
              Linear,
              'Returns prediction of New cases for Covid-19.'
              , override = True)
client.deploy('New_Cases_Covid_prediction_Tree',
              RegressionTree,
              'Returns prediction of New cases for Covid-19.'
              , override = True)
```

Trong terminal: chạy syntax “python model.py”.

Sau khi deploy thành công, trên server local của tabpy như sau:

```
Deployed Models:
{
  "New_Cases_Covid_prediction_Linear": {
    "description": "Returns prediction of New cases for Covid-19.",
    "type": "model",
    "version": 11,
    "dependencies": [],
    "target": null,
    "creation_time": 1681443763,
    "last_modified_time": 1681571919,
    "schema": null,
    "docstring": "-- no docstring found in query function --"
  },
  "New_Cases_Covid_prediction_Tree": {
    "description": "Returns prediction of New cases for Covid-19.",
    "type": "model",
    "version": 11,
    "dependencies": [],
    "target": null,
    "creation_time": 1681443766,
    "last_modified_time": 1681571920,
    "schema": null,
    "docstring": "-- no docstring found in query function --"
  }
}
```

**Useful links:**

- [TabPy Documentation](#)
- [TabPy Source Code](#)
- [TabPy PyPi](#)

### b. Linear Regression.

Có dạng biểu thức:

$$y = \theta_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

Cài đặt trong file model.py:

```
def Linear(_arg1, _arg2, _arg3, _arg4, _arg5):
    X = pd.DataFrame({
        'TotalCases': _arg1,
        'TotalRecovered': _arg2,
        'TotalDeaths': _arg3,
        'NewDeaths': _arg4,
        'NewRecovered': _arg5,
    })
    model = LinearRegression()
    model.fit(train_X, train_Y)

    return model.predict(X).reshape(-1).astype(int).tolist()
```

### c. Regression Tree.

Với dữ liệu đầu ra dự đoán có kiểu numerical, nhóm em nghĩ tới có thể áp dụng Regression Tree (một dạng của Decision Tree).

Cài đặt trong file model.py:

```
def RegressionTree(_arg1, _arg2, _arg3, _arg4, _arg5):
    X = pd.DataFrame({
        'TotalCases': _arg1,
        'TotalRecovered': _arg2,
        'TotalDeaths': _arg3,
        'NewDeaths': _arg4,
        'NewRecovered': _arg5,
    })
    model = DecisionTreeRegressor(random_state=0)
    model.fit(train_X, train_Y)

    return model.predict(X).reshape(-1).astype(int).tolist()
```

## 5.2. Trực quan hóa trong Tableau.

Vì dữ liệu được train đến ngày 10/03/2023 nên sẽ không dùng để trực quan và nhóm em trực quan dữ liệu trên bộ test. Connect Data trong file test.csv và result.csv với nhau.

The screenshot shows the Tableau Data Source interface. On the left, under 'Connections', there is a connection named 'test' (Text file). Under 'Files', there are three files: 'corr.csv', 'result.csv', and 'test.csv'. A connection line connects 'test.csv' to 'result.csv'. At the top right, there are 'Connection' (Live or Extract) and 'Filters' (0 | Add) options. Below the connections, a preview of the data is shown in a grid format. The columns are labeled: '#', 'test.csv', '#', 'test.csv', '#', 'test.csv', '#', 'test.csv', '#', 'test.csv', '#', 'test.csv'. The rows show data for countries: China, USA, India, France, and Germany. The preview also indicates there are 16 fields and 1603 rows.

Tạo Calculated Field cho mô hình Linear và Decision Tree:

- Mô hình Linear:

The screenshot shows the 'Linear Prediction' calculated field dialog in Tableau. The field name is 'Linear Prediction' and it is connected to the 'test' connection. The description says 'Results are computed along Table (across)'. The script is defined as follows:

```
SCRIPT_REAL(
    return tabpy.query('New_Cases_Covid_prediction_Linear', _arg1,_arg2,_arg3,_arg4,_arg5)['response']
),
ATTR([Total Cases]), ATTR([Total Recovered]), ATTR([Total Deaths]),
ATTR([New Deaths]), ATTR([New Recovered])
)
```

At the bottom, it says 'The calculation is valid.' and shows '4 Dependencies'. There are 'Apply' and 'OK' buttons.

- Mô hình Regression Tree:

Regression Tree Prediction  X

Results are computed along Table (across).

```
SCRIPT_REAL("
    return tabpy.query('New_Cases_Covid_prediction_Tree', _arg1,_arg2,_arg3,_arg4,_arg5)['response']
",
ATTR([Total Cases]), ATTR([Total Recovered]), ATTR([Total Deaths]),
ATTR([New Deaths]), ATTR([New Recovered])
)
```

The calculation is valid.

Default Table Calculation

4 Dependencies

Tạo Calculated Field để tính R-square cho 2 mô hình sử dụng:

- Mô hình Linear:

R2Linear  X

Results are computed along Table (across).

```
SCRIPT_REAL("
    from sklearn.metrics import r2_score
    return r2_score(_arg2, _arg1)
",
[Linear Prediction], ATTR([New Cases]))
```

The calculation is valid.

Default Table Calculation

3 Dependencies

- Mô hình Regression Tree:

R2Tree  X

Results are computed along Table (across).

```
SCRIPT_REAL("
    from sklearn.metrics import r2_score
    return r2_score(_arg2, _arg1)
",
[Regression Tree Prediction], ATTR([New Cases]))
```

The calculation is valid.

Default Table Calculation

3 Dependencies

Thêm các Field “Country,Other” và “Date” vào Columns, “R2 Linear” và “R2 Tree” vào Rows. Sau đó thêm “Date” vào Filter. Ta có biểu đồ đơn giản để xem các giá trị R2, thay đổi theo các ngày dự đoán khác nhau:

Basic Look Prediction		Date
R2 Square Linear	R2 Square Tree	
0.4878	0.5572	<input checked="" type="checkbox"/> (All)
<		<input checked="" type="checkbox"/> 3/11/2023
		<input checked="" type="checkbox"/> 3/12/2023
		<input checked="" type="checkbox"/> 3/13/2023
		<input checked="" type="checkbox"/> 3/14/2023
		<input checked="" type="checkbox"/> 3/15/2023
		<input checked="" type="checkbox"/> 3/16/2023
		<input checked="" type="checkbox"/> 3/17/2023

- Nhận xét:

- Dựa trên giá trị R2, có thể thấy mô hình Linear không phù hợp, còn với Regression Tree có thể sử dụng được nhưng không đem lại hiệu quả cao (underfitting).
- Nhưng nếu dự đoán dựa trên các ngày khác nhau và ít hơn, nhóm em thấy mô hình Regression Tree có hiệu quả tốt. Ví dụ nếu chỉ dự đoán cho ngày 11/03/2023:

Basic Look Prediction		Date
R2 Square Linear	R2 Square Tree	
0.5891	0.8769	<input checked="" type="checkbox"/> (All)
<		<input checked="" type="checkbox"/> 3/11/2023
		<input type="checkbox"/> 3/12/2023
		<input type="checkbox"/> 3/13/2023
		<input type="checkbox"/> 3/14/2023
		<input type="checkbox"/> 3/15/2023
		<input type="checkbox"/> 3/16/2023
		<input type="checkbox"/> 3/17/2023

Ngoài ra ta có thể xem trước kết quả dự đoán ở trong file “result.csv” như sau:

- Pivot data trong Data Source, chọn cột R2 predicted Tree và R2 predicted Linear, right click và chọn pivot (ảnh trái) ta được cột mới (ảnh phải):

The screenshot shows the Power BI interface with two tables. On the left, a table has columns '# result.csv' and 'R2 predicted Tree' with values 0.557153 repeated five times. A context menu is open over the second row of the 'R2 predicted Tree' column, with options like 'Rename', 'Copy Values', 'Hide', 'Create Calculated Field...', 'Pivot', and 'Merge Mismatched Fields'. The 'Pivot' option is highlighted. On the right, another table shows the results of the pivot operation. It has columns '# Pivot' and 'R2 Calculated' with values 0.4878093 and 0.5571534 respectively. Below it, the original table is shown again with columns '# Pivot' and 'R2 values' containing the same data.

#	result.csv	R2 predicted Tree
		0.557153
		0.557153
		0.557153
		0.557153
		0.557153

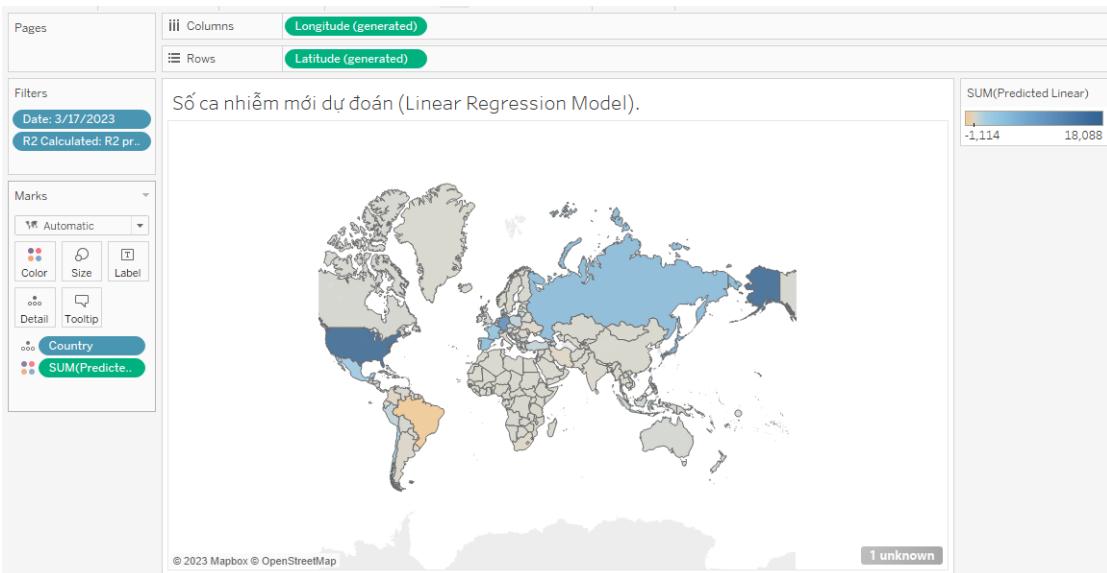
#	Pivot	R2 Calculated
	R2 Calculated	0.4878093
	R2 predicted Tree	0.5571534
	R2 predicted Linear	0.4878093
	R2 predicted Tree	0.5571534
	R2 predicted Linear	0.4878093

- Thả “R2 Calculated” vào Columns, Color trong Marks Tag; và “R2 Values” vào Rows.



Để xem thêm kết quả dự đoán của mỗi mô hình, ta sẽ xem trên dữ liệu trong “results.csv”, thả “Country,Other” vào Rows và Date vào Filter, ta chỉ xem ngày gần nhất (17/03/2023). Trong Show me chọn maps. Ta có kết quả:

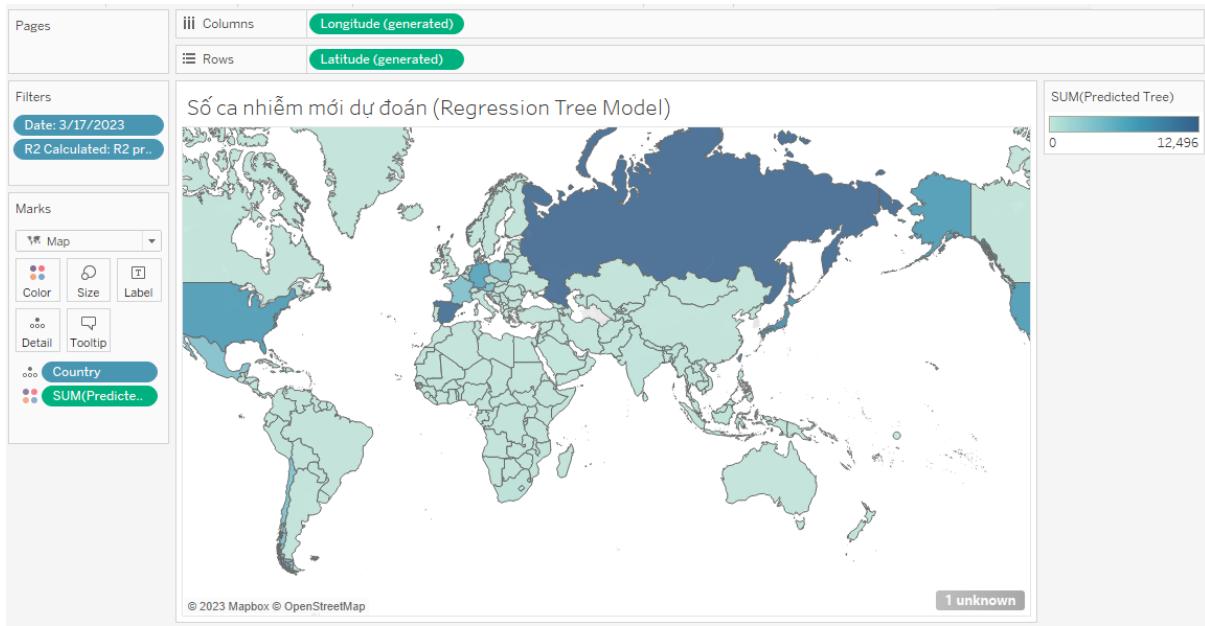
- Mô hình Linear:



#### Nhận xét:

- Nhóm em thấy với Linear Regression, mô hình dự đoán được giá trị âm (màu cam) là giá trị không hợp lệ.
- Với giá trị  $R^2 = 0.48 < 0.5$ , có thể kết luận mô hình Linear không phù hợp trong việc dự đoán NewCases.

- Mô hình Regression Tree:



### Nhận xét:

- Nhóm em thấy với Regression Tree, mô hình dự đoán kết quả tốt hơn và không có giá trị âm.
- Với giá trị  $R^2 = 0.56 > 0.5$ , có thể kết luận mô hình Regression Tree có thể sử dụng được, nhưng không đem lại hiệu quả cao (underfitting).

## 6. References

- [Tabpy](#)
- [Regression Tree - GeeksforGeeks](#)
- [Fundamentals of Visualization with Tableau - University of California - Coursera](#)