

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

-----♪♪♪-----



**BÁO CÁO LAB 02
TRỰC QUAN HÓA DỮ LIỆU VỚI TABLEAU**

Bộ môn: Trực quan hóa dữ liệu
Giảng viên hướng dẫn: Lê Ngọc Thành

2020 - 2021

MỤC LỤC

I. THÔNG TIN NHÓM VÀ MỨC ĐỘ HOÀN THÀNH ĐỒ ÁN	2
1. Thông tin nhóm.....	2
2. Mức độ hoàn thành đồ án	2
II. BÁO CÁO ĐỒ ÁN.....	3
A. LÝ THUYẾT: TÌM HIỂU CÔNG CỤ TABLEAU.....	3
1. Giới thiệu về Tableau.....	3
2. Các tính năng hỗ trợ của Tableau.....	4
B. THỰC HÀNH.....	19
B.1. TRỰC QUAN BẰNG TABLEAU	19
1. Country, New_Cases	19
2. Country, New_Deaths	23
3. Country, Total_Cases	27
4. Active_Cases, Tests/1M_pop	30
5. Country, New_Deaths, Active_Cases	32
6. Total_Cases, New_Deaths, New_Cases.....	36
7. Country, Active_Cases, Serious_Critical	38
8. Country, Total_Cases, Total_Recovered	41
9. Region, Tot_Cases/1M_pop, Tests/1M_pop, Deaths/1M_pop	45
10. Country, Total_Cases, Total_Deaths, Total_Recovered, Active_Cases	47
11. Day, New_Cases	51
12. Day, New_Deaths	52
13. Total Cases, Day, Regions.....	54
14. Active_Cases, Day	56
B.2. CHẠY THUẬT TOÁN HỌC MÁY	57
1. Linear Regression	57
2. SVM.....	57
III. TÀI LIỆU THAM KHẢO	58

I. THÔNG TIN NHÓM VÀ MỨC ĐỘ HOÀN THÀNH ĐỒ ÁN**1. Thông tin nhóm**

Họ và tên	MSSV	Công việc	Mức độ hoàn thành
Đinh Phan Kim Ngân	18120476	Sử dụng Tableau để trực quan dữ liệu; giải thích, rút ra ý nghĩa sau mỗi dữ liệu trực quan	100%
Lê Hoàng Phương Nhi	18120496	Tìm hiểu Tableau; sử dụng Tableau để trực quan dữ liệu; giải thích, rút ra ý nghĩa sau mỗi dữ liệu trực quan; chạy Model dự đoán	100%
Nguyễn Thị Hồng Nhung	18120498	Sử dụng Tableau để trực quan dữ liệu; giải thích, rút ra ý nghĩa sau mỗi dữ liệu trực quan	100%
Nguyễn Thành Phát	18120501	Sử dụng Tableau để trực quan dữ liệu; giải thích, rút ra ý nghĩa sau mỗi dữ liệu trực quan	100%
Lê Thị Như Quỳnh	18120530	Tìm hiểu Tableau; sử dụng Tableau để trực quan dữ liệu; giải thích, rút ra ý nghĩa sau mỗi dữ liệu trực quan	100%

2. Mức độ hoàn thành đồ án

Các tiêu chí	Mức độ hoàn thành
Tìm hiểu về Tableau	100%
Áp dụng Tableau để trực quan dữ liệu	100%
Rút ra ý nghĩa hợp lý sau mỗi dữ liệu được trực quan	100%
Báo cáo trình bày bối cảnh và định dạng hợp lý, rõ ràng	100%

II. BÁO CÁO ĐỒ ÁN

A. LÝ THUYẾT: TÌM HIỂU CÔNG CỤ TABLEAU

1. Giới thiệu về Tableau

a. *Sơ lược về Tableau*

- Phần mềm Tableau là một trong những công cụ data visualization đang phát triển nhanh nhất hiện đang được sử dụng trong ngành BI (Business Intelligence). Data visualization chuyển đổi bảng dữ liệu thô thành định dạng dễ hiểu mà không cần nhiều kỹ thuật và kiến thức mã hóa.
- Tableau là phần mềm hỗ trợ phân tích và trực quan hóa dữ liệu, cung cấp môi trường có thể phân tích, xử lý và tổng hợp dữ liệu dưới dạng hình ảnh, biểu đồ trực quan thông qua các thao tác kéo thả; từ đó tối ưu thời gian so sánh, tổng kết, đánh giá và đưa ra những quyết định về các dữ liệu từ đơn giản đến phức tạp.
- Các sản phẩm chính của Tableau:
 - Tableau Prep: dùng để kết nối với các data source, thiết lập các lọc dữ liệu và các công thức để chuẩn bị cho Tableau Creator để phân tích dữ liệu.
 - Tableau Creator: dùng để kết nối với các data source, thiết kế Tableau Data Mart cùng với việc xây dựng các phân tích, dashboard quản trị.
 - Tableau Explorer: dùng để kết nối vào Tableau Server, sử dụng các Tableau Data Mart để xây dựng các phân tích, các dashboard cho bộ phận hoặc cho cá nhân và có thể chia sẻ.
 - Tableau Viewer: dùng cho các business user mà không cần phải thực hiện các phân tích hoặc xây dựng các dashboard mà chỉ sử dụng các phân tích và dashboard của đồng nghiệp.

b. *Lợi ích, ưu điểm của Tableau*

- Dễ sử dụng
- Dễ dàng xây dựng các Dashboard và Phân tích rất đẹp
- Đưa dữ liệu tới tay của người vận hành và họ tự phục vụ
- Xây dựng môi trường làm việc dựa trên dữ liệu và phân tích
- Luôn có dữ liệu và phân tích mọi lúc mọi nơi
- Sử dụng cho tất cả phòng ban, mọi nhân viên và mọi ngành nghề của doanh nghiệp
- Tốc độ xử lý dữ liệu cực nhanh với công nghệ in-memory
- Khả năng mở rộng theo độ lớn của dữ liệu và phức tạp của quản trị
- Cộng tác, chia sẻ và bảo mật
- Kết nối và làm việc với nhiều loại dữ liệu cùng lúc
- Đáp ứng với các công nghệ Big Data, AI và khả năng tích hợp cao

2. Các tính năng hỗ trợ của Tableau

- Tùy chọn sử dụng nguồn dữ liệu:

Tableau cung cấp vô số tùy chọn nguồn dữ liệu mà ta có thể kết nối và trích xuất dữ liệu. Các nguồn dữ liệu có thể là các file dữ liệu cố định, bảng tính, các cơ sở dữ liệu, big data cho đến các dữ liệu lưu trữ trên không gian đám mây đều sẵn có ở Tableau

➤ To a File:

Có thể chọn dữ liệu với các file: Microsoft Excel, Text file, JSON file, Microsoft Access, PDF file, Spatial file, Statistical file

Ví dụ chọn dữ liệu là file Microsoft Excel: Ở màn hình Connect → Microsoft Excel → Chọn file excel muốn mở

Sau khi chọn xong dữ liệu:

F1	F2	F3	F4	F5	F6
MAKE THE STORY UN...	null	null	null	null	null
null	DATA TO GRAPH	null	null	null	null
null	null	2,016.00	2,017.00	2,018.00	2,019.00
null	Credit Card	92.10	89.40	94.00	97.10
null	Home Equity Line of C...	5.60	12.70	47.50	65.00
null	Fixed-Term Loan	48.20	53.80	61.90	53.70
null	Total	58.90	47.60	78.40	75.60
null	L Credit Card	100.00	100.00	100.00	100.00
null	100.00	100.00	100.00	100.00	100.00

➤ To a server:

Có thể chọn dữ liệu từ các server:

Search			
Action Matrix	Google Drive	OData	Teradata OLAP Connector
Action Vector	Google Sheets	OneDrive	TIBCO Data Virtualization
Amazon Athena	Hortonworks Hadoop Hive	Oracle	Vertica
Amazon Aurora	IBM BigInsights	Oracle Eloqua	Web Data Connector
Amazon EMR Hadoop Hive	IBM DB2	Oracle Essbase	
Amazon Redshift	IBM PDA (Netezza)	Pivotal Greenplum Database	Other Databases (JDBC)
Anaplan	Intuit QuickBooks Online	PostgreSQL	Other Databases (ODBC)
Apache Drill	Intuit QuickBooks Online (9.3-2018.1)	Presto	
Aster Database	Kognitio	Progress OpenEdge	
Azure SQL Data Warehouse	MapR Hadoop Hive	Salesforce	
Box	MariaDB	SAP HANA	
Cloudera Hadoop	Marketo	SAP NetWeaver Business Warehouse	
Denodo	MarkLogic	SAP Sybase ASE	
Dropbox	MemSQL	SAP Sybase IQ	
Exasol	Microsoft Analysis Services	ServiceNow ITSM	
Firebird	Microsoft PowerPivot	SharePoint Lists	
Google Ads	Microsoft SQL Server	Snowflake	
Google Analytics	MonetDB	Spark SQL	
Google BigQuery	MongoDB BI Connector	Splunk	
Google Cloud SQL	MySQL	Teradata	

Ví dụ chọn dữ liệu từ Server Google Sheets: Ở màn hình Connect ở mục To a Server → More → Google Sheets → Chọn tài khoản Google Sheets và cho phép Tableau kết nối dữ liệu → Chọn dữ liệu muốn mở
Sau khi chọn xong dữ liệu:

Họ Tên	Ngày Sinh	Số báo danh	Điểm thi
NGUYỄN THỊ GIANG	10/10/1999	16008953	Toán: 4.40 Ngữ văn: ...
KIỀU THỊ THANH TÂM	08/01/1999	16007411	Toán: 8.60 Ngữ văn: ...
NGUYỄN HẢI HUY	28/12/1999	16001646	Toán: 8.60 Ngữ văn: ...
NGUYỄN THỊ LAN ANH	14/09/1999	16001087	Toán: 7.00 Ngữ văn: ...
NGUYỄN THỊ TRANG	22/09/1999	16004855	Toán: 7.80 Ngữ văn: ...
NGUYỄN THỊ MINH P...	08/01/1999	16009103	Toán: 6.00 Ngữ văn: ...
PHÙNG THỊ TÂN	17/08/1999	16002239	Toán: 7.20 Ngữ văn: ...
ĐỖ KIỀU ANH	20/10/1999	16001078	Toán: 8.20 Ngữ văn: ...
NGUYỄN THỊ LÃM ANH	20/02/1996	16000226	Toán: 1.60 Ngữ văn: ...

- Cung cấp công cụ để làm sạch và chuẩn bị dữ liệu cho việc phân tích:
 - Xóa các cột trùng nhau:

Trong file orthers_west, thì có các cột bắt đầu Right_..., có vẻ như các cột này trùng lắp với các cột còn lại
Để xóa các cột này thì bỏ dấu tích ở các cột này đi

Type	Field Name	Original Field Name	Changes	Preview
Abc	Right_Ship Mode	Right_Ship Mode		Standard Class
Abc	Right_Customer...	Right_Customer ID		JM-15655
Abc	Right_Customer...	Right_Customer Name		Jim Mitchum
Abc	Right_Segment	Right_Segment		Corporate
Abc	Right_Country	Right_Country		United States
Abc	Right_City	Right_City		Glendale
Abc	Right_State2	Right_State2		Arizona
#	Right_Postal Co...	Right_Postal Code		85,301
Abc	Right_Region	Right_Region		West
Abc	Right_Product ID	Right_Product ID		OFF-FA-10000611, OFF-PA-10004733
Abc	Right_Category	Right_Category		Office Supplies
Abc	Right_Sub-Cate...	Right_Sub-Category		Fasteners, Paper
Abc	Right_Prof...	Right_Sub-Category		Binder Clips by OIC, Things To Do Today Spiral Book
#	Right_Sales	Right_Sales		2,368, 19,008
#	Right_Quantity	Right_Quantity		2, 3
#	Right_Discount	Right_Discount		0,2
#	Right_Profit	Right_Profit		0,8288, 6,8904
Abc	State	State		AZ

➤ Thay đổi tên cột:

Nhận thấy cột State, dữ liệu của nó là dạng viết tắt, trong khi các file còn lại ghi rõ tên của bang. Mà ta thấy cột Right_State2 trùng với cột State, vậy xóa cột State, đổi tên cột Right_State2 thành State

Tableau Prep Builder - Flow1* - Trial expires in 14 days

File Edit Flow Server Help

Input

Changes (21)

Orders_West 41 fields Filter Values...

Fields selected: 21 of 41

Type	Field Name	Original Field Name	Changes	Preview
Abc	Right_Ship Mode	Right_Ship Mode		Standard Class
Abc	Right_Customer...	Right_Customer ID		JM-15655
Abc	Right_Customer...	Right_Customer Name		Jim Mitchum
Abc	Right_Segment	Right_Segment		Corporate
Abc	Right_Country	Right_Country		United States
Abc	Right_City	Right_City		Glendale
Abc	state	Right_State2		Arizona
#	Right_Pos_state	Right_Postal Code		85301
Abc	Right_Region	Right_Region		West
Abc	Right_Product ID	Right_Product ID		OFF-FA-10000611, OFF-PA-10004733
Abc	Right_Category	Right_Category		Office Supplies
Abc	Right_Sub-Cate...	Right_Sub-Category		Fasteners, Paper
Abc	Right_Product ...	Right_Product Name		Binder Clips by OIC, Things To Do Today Spiral Book
#	Right_Sales	Right_Sales		2,368, 19,008
#	Right_Quantity	Right_Quantity		2, 3
#	Right_Discount	Right_Discount		0.2
#	Right_Profit	Right_Profit		0.8288, 6,8904
Abc	State	State		AZ

- Thêm một cột mới:
- Thêm một cột Region vào file Orders_central

Tableau Prep Builder - Flow1* - Trial expires in 14 days

File Edit Flow Server Help

orders_south... View and clean data

Orders_Central Clean 1

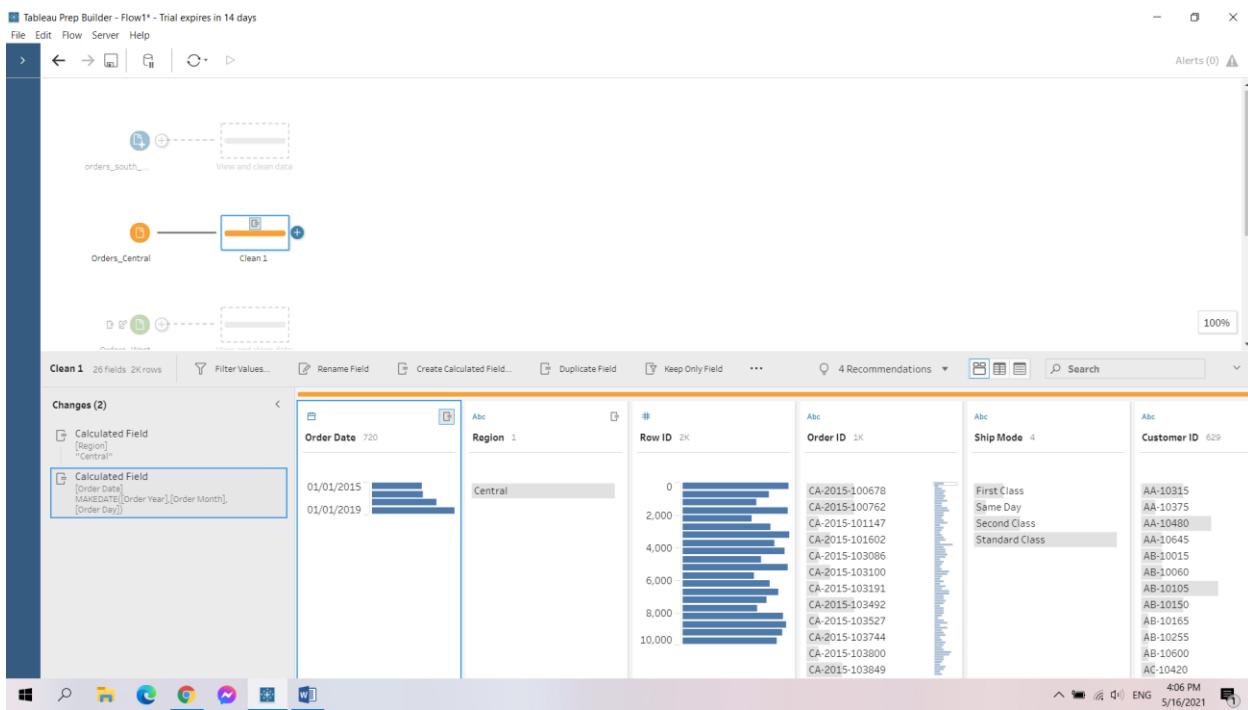
Clean 1 25 fields 2K rows Filter Values... Create Calculated Field...

Changes (1)

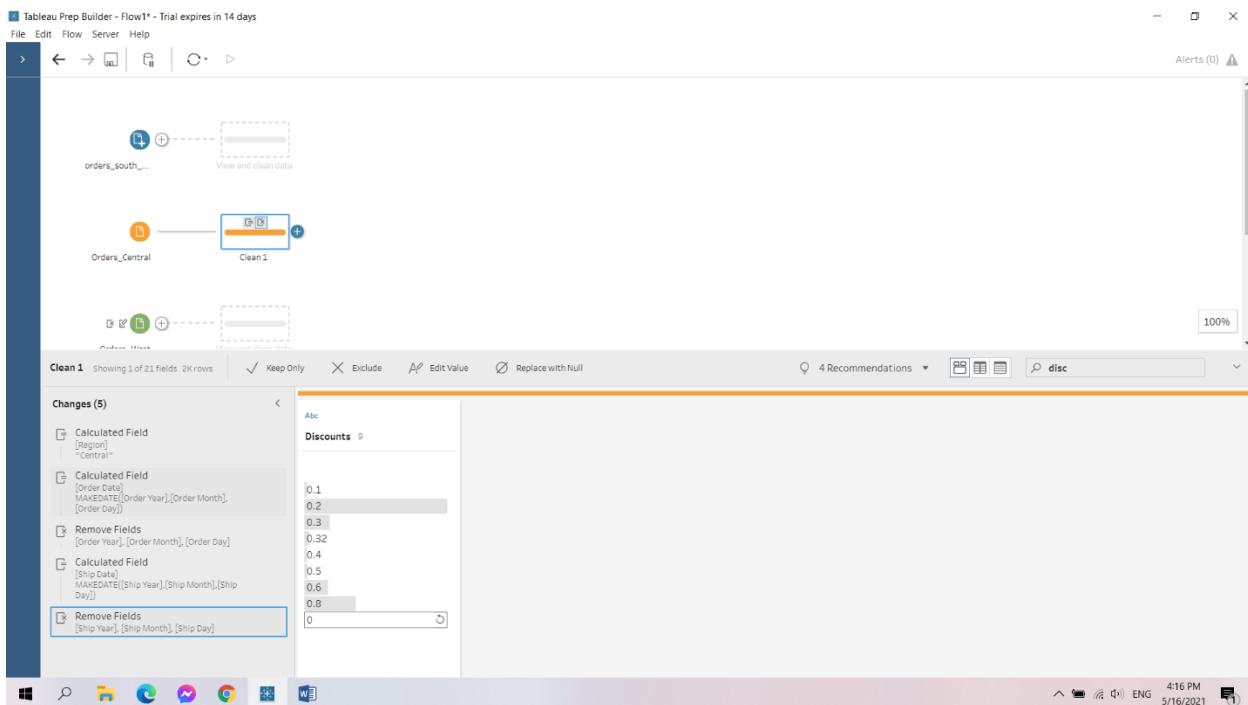
Calculated Field [Region] "Central"

Region	Row ID	Order ID	Ship Mode	Customer ID	Customer Name
Central	0	CA-2015-100678	First Class	AA-10315	Aaron Bergman
Central	2,000	CA-2015-100762	Same Day	AA-10375	Aaron Smalley
Central	4,000	CA-2015-101147	Second Class	AA-10480	Adam Bellavance
Central	6,000	CA-2015-101602	Standard Class	AA-10645	Adam Hart
Central	8,000	CA-2015-103086		AB-10015	Adam Shillingburg
Central	10,000	CA-2015-103100		AB-10060	Adrian Barton
Central		CA-2015-103191		AB-10105	Adrian Hane
Central		CA-2015-103492		AB-10150	Aimee Bixby
Central		CA-2015-103527		AB-10165	Alan Barnes
Central		CA-2015-103744		AB-10255	Alan Dominguez
Central		CA-2015-103800		AB-10600	Alan Haines
Central		CA-2015-103849		AC-10420	Alan Hwang

- Một cột được thêm mới được tính toán từ các cột tồn tại trong file:

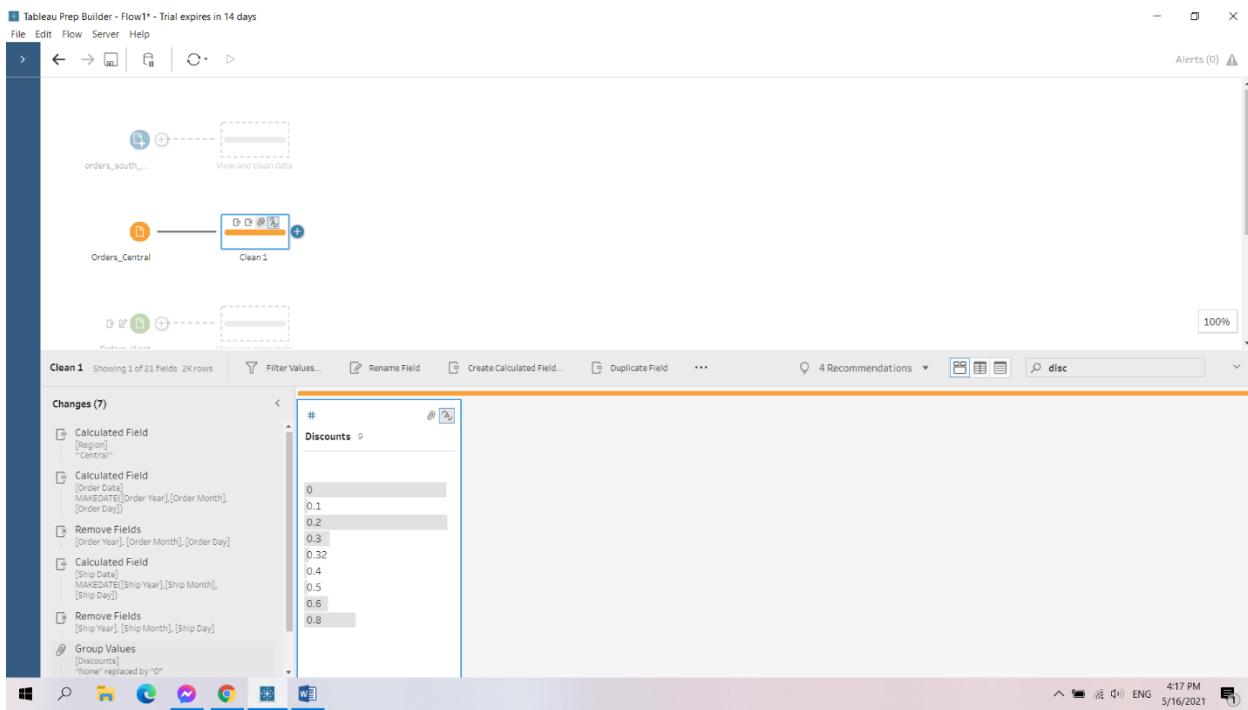


- Thay đổi giá trị:
Cột discounts, đổi giá trị none thành 0

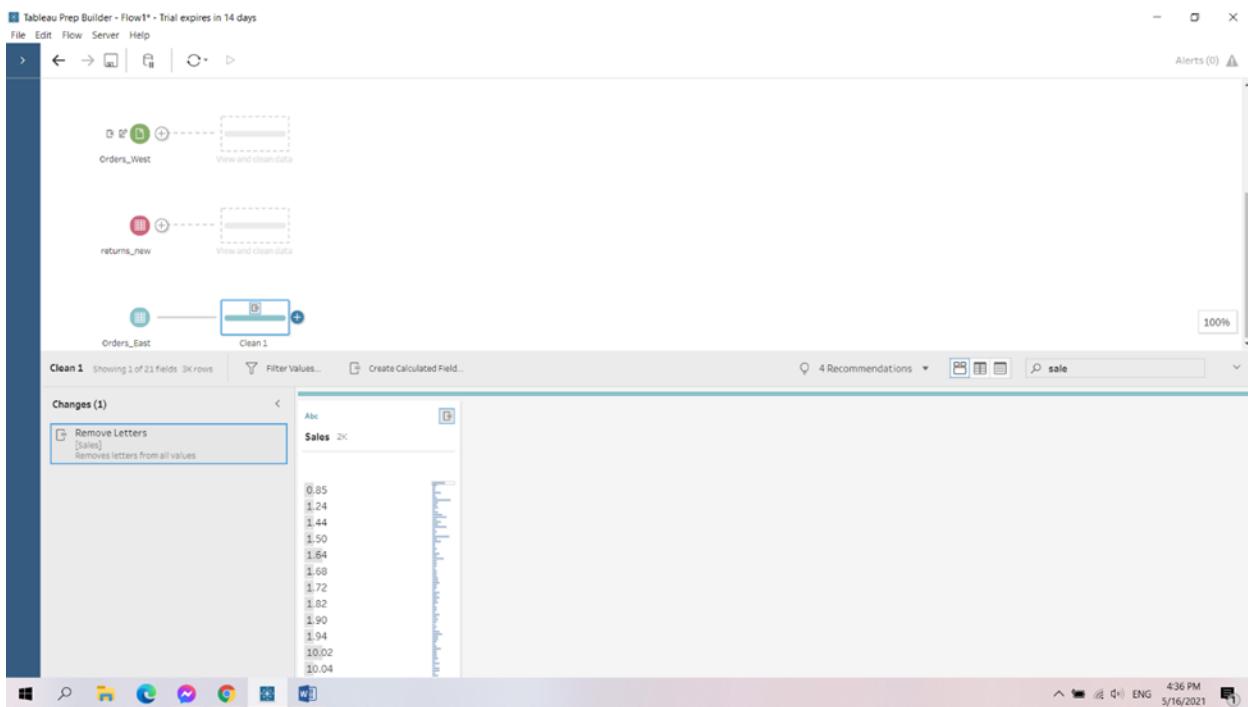


- Thay đổi kiểu dữ liệu của một cột:

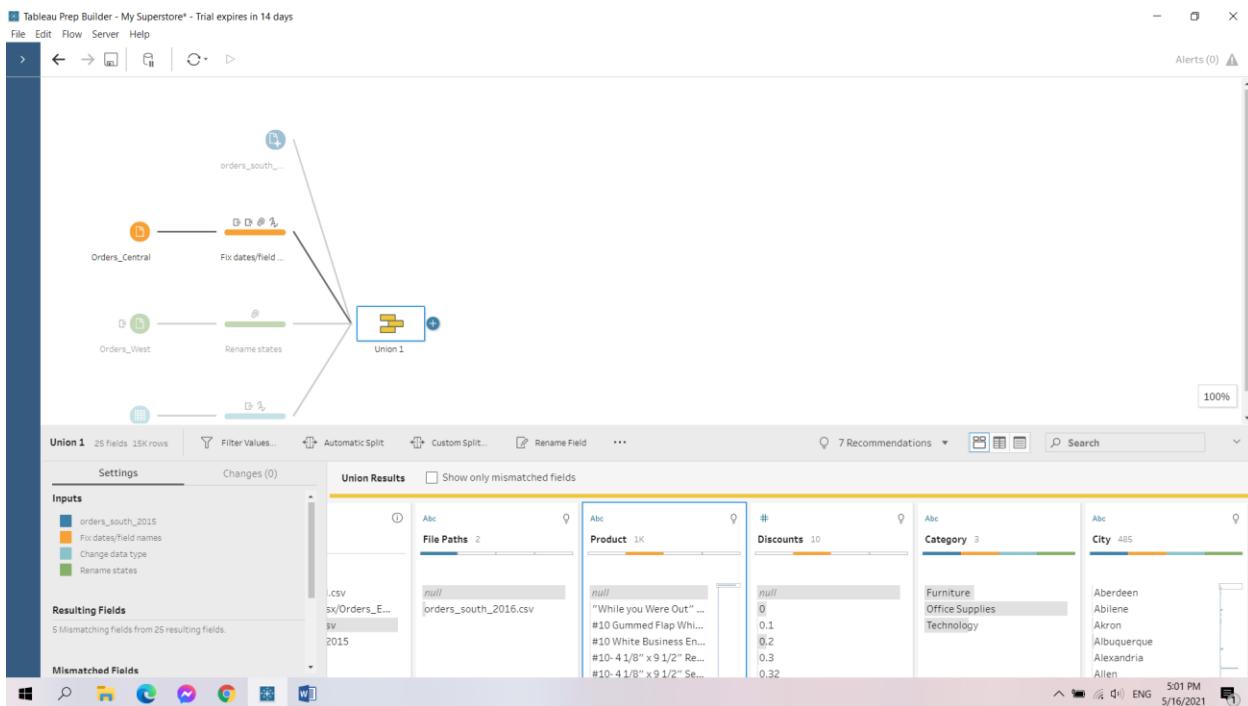
Cột discounts đổi sang dạng numeric



- Xóa kí tự chữ trong giá trị của cột:
Trong file orders_East, cột Sales ban đầu các giá trị có dạng bắt đầu là USD.. ví dụ USD11
Xóa kí tự USD chỉ lấy mỗi số



➤ Kết hợp nhiều file dữ liệu với nhau:



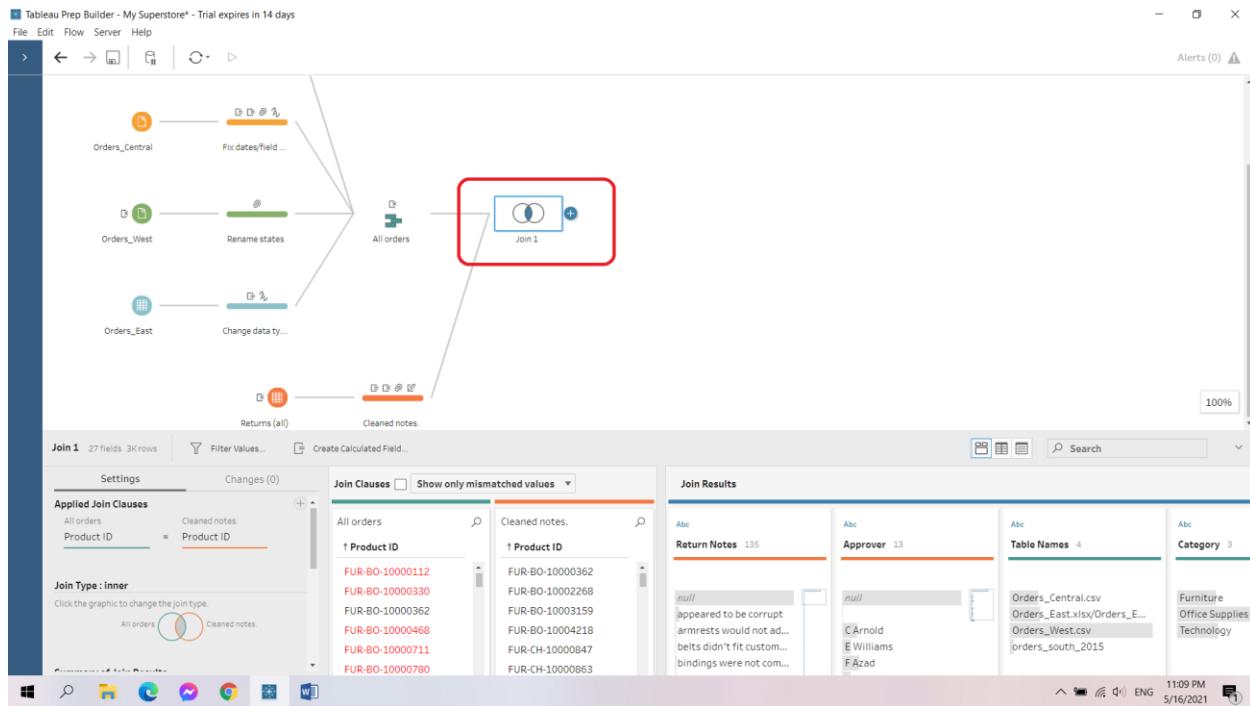
➤ Khi kết hợp nhiều file khác nhau, có nhiều file:

The screenshot shows the Tableau Prep Builder interface. A flow starts with 'Orders_West' and ends with 'All orders'. A step labeled 'Rename states' is highlighted. The 'Union Results' pane displays four tables: 'Table Names', 'Category', 'City', and 'Country'. The 'Customer' table is expanded, showing rows for Aaron Bergman, Aaron Hawkins, etc. A preview of the combined data is shown at the bottom.

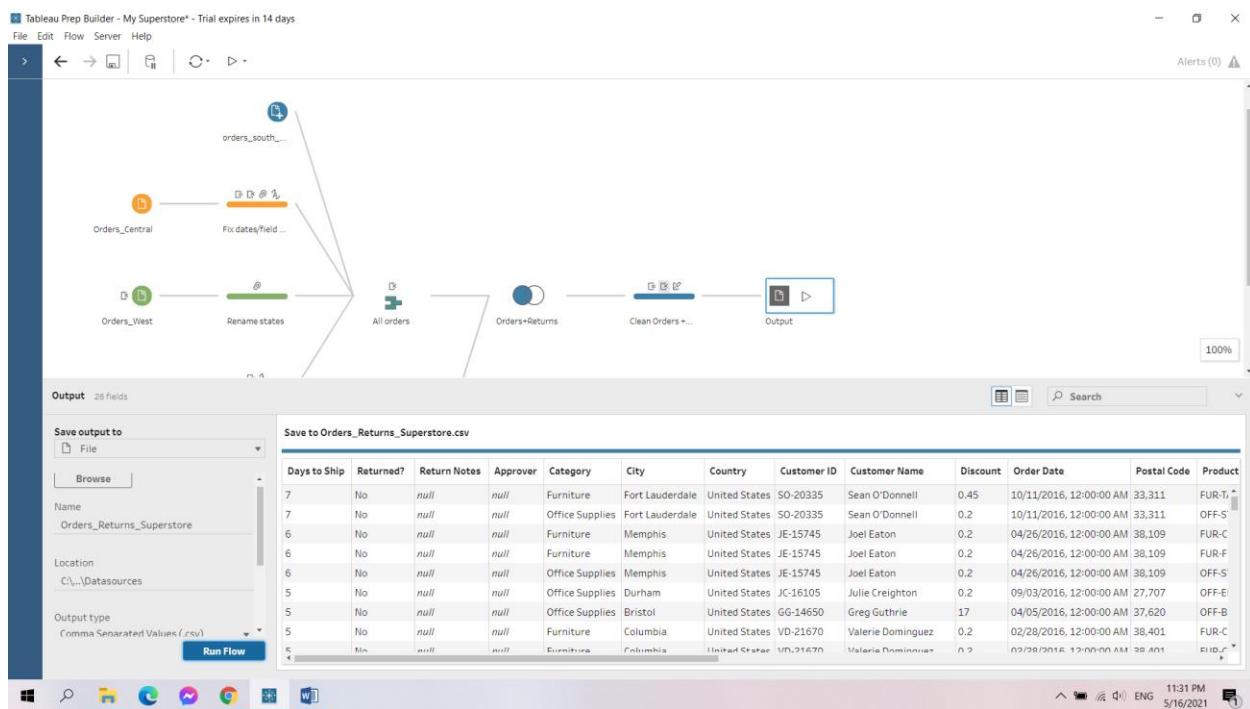
- Các giá trị trong cột tuy mang ý nghĩa là cùng một giá trị nhưng nhập khác nhau:
Trong cột Approver file return (alls) có các giá trị trong cột như C Arnold, C Arnold, c. arnold, C.. Arnold, C/ Arnold đều chỉ một giá trị, dùng tableau để gom cụm các giá trị đó thành một

The screenshot shows the Tableau Prep Builder interface. A flow starts with 'Orders_East' and ends with 'Clean 2'. The 'Clean 2' step is highlighted. The 'Changes (6)' pane shows various cleaning steps like Trim Spaces, Calculated Field, and Rename Field. The 'Approver' column in the 'Return Notes' table is selected, and a context menu is open, showing options like Filter, Clean, Group Values, Split Values, and Common Characters.

➤ Join các database, với điều kiện hai database có ít nhất một cột chung:

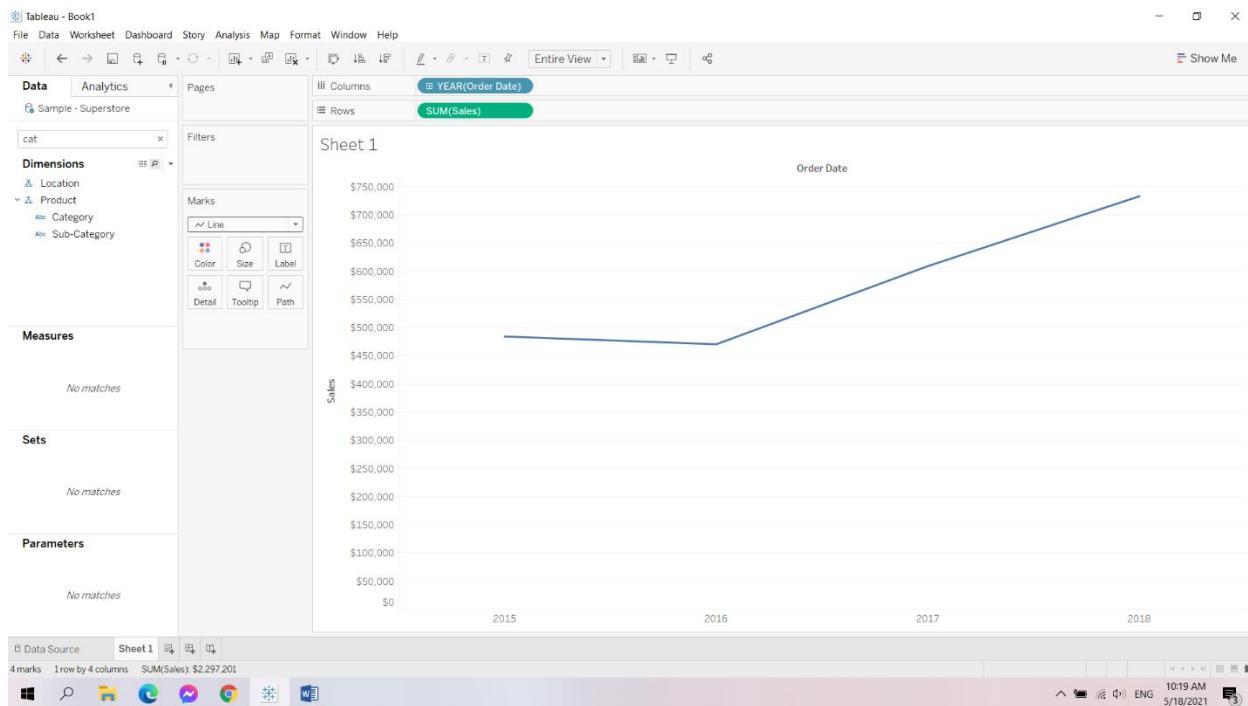


➤ Sau khi tiền xử lí có thể xuất ra một file khác:

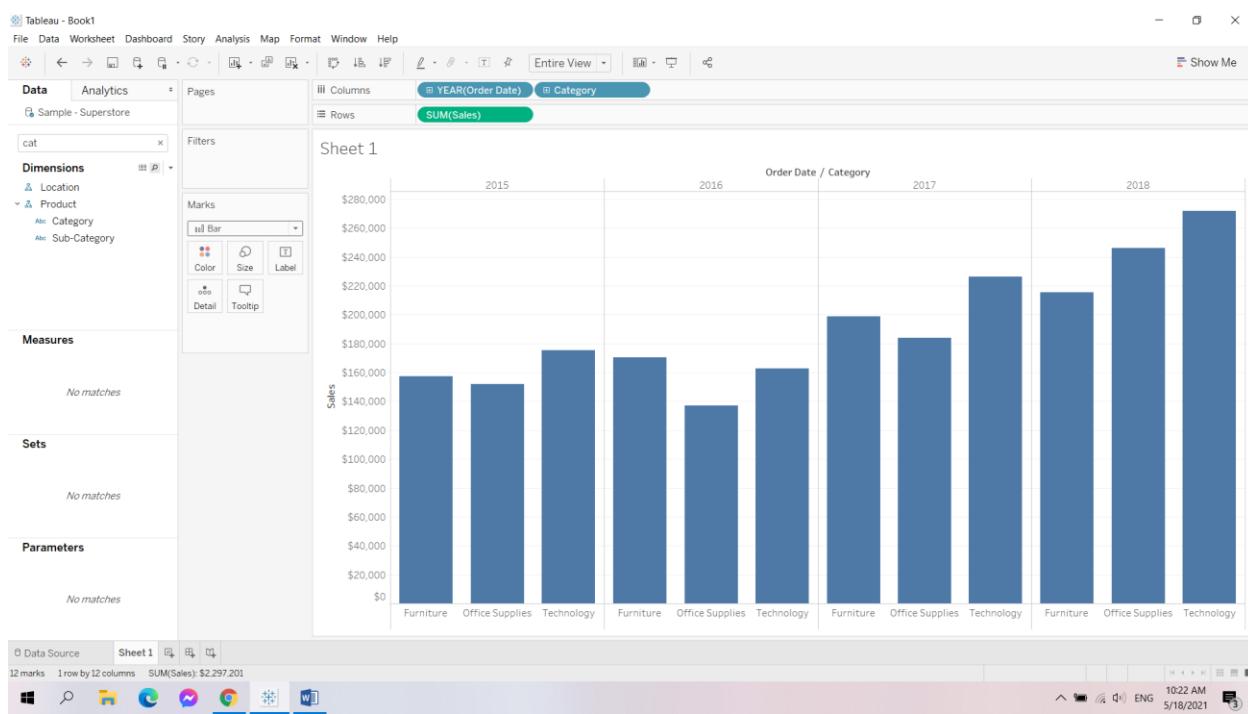


- Dễ dàng tạo các biểu đồ bằng cách kéo thả đơn giản

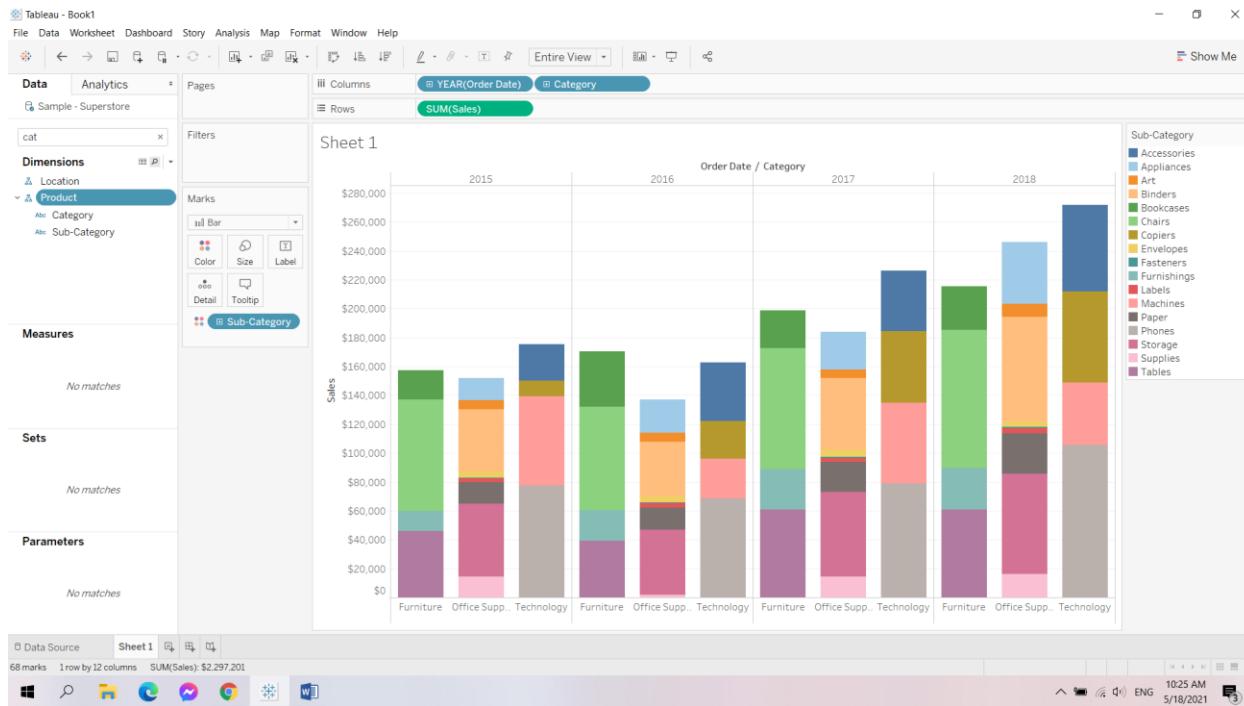
➤ Tạo những chart đơn giản:



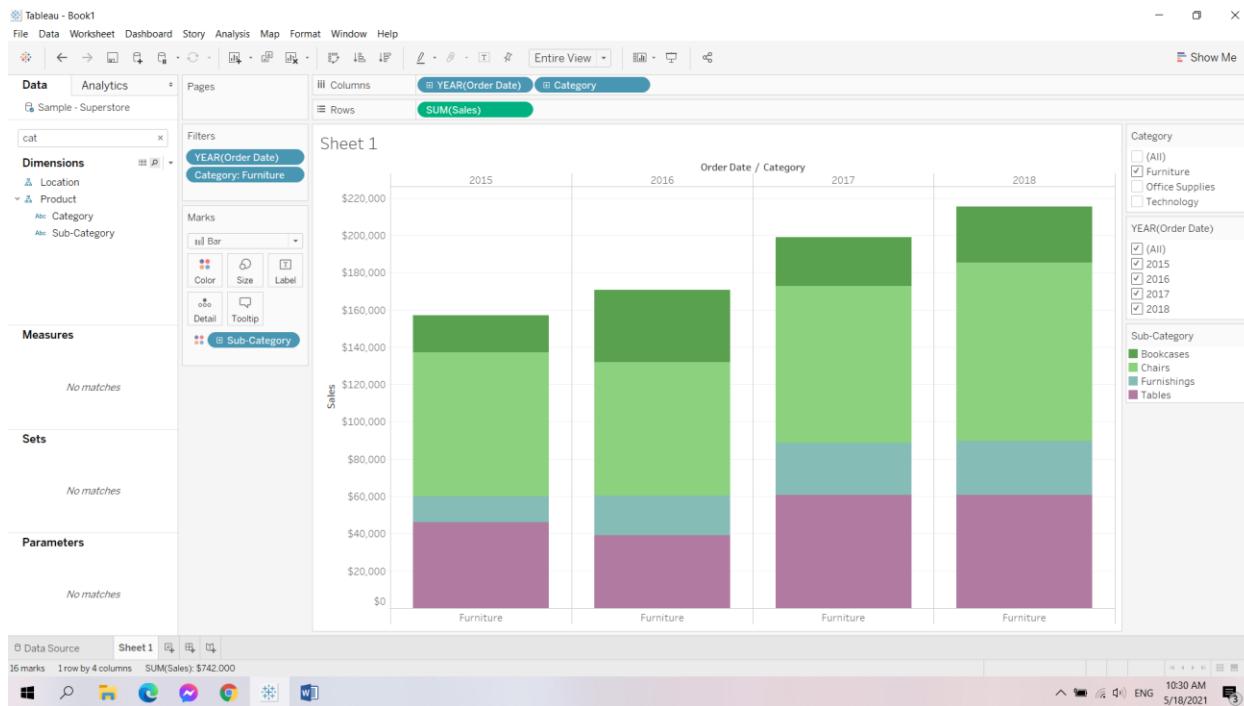
➤ Tạo chart khi kết hợp nhiều hơn 3 thuộc tính trở lên:



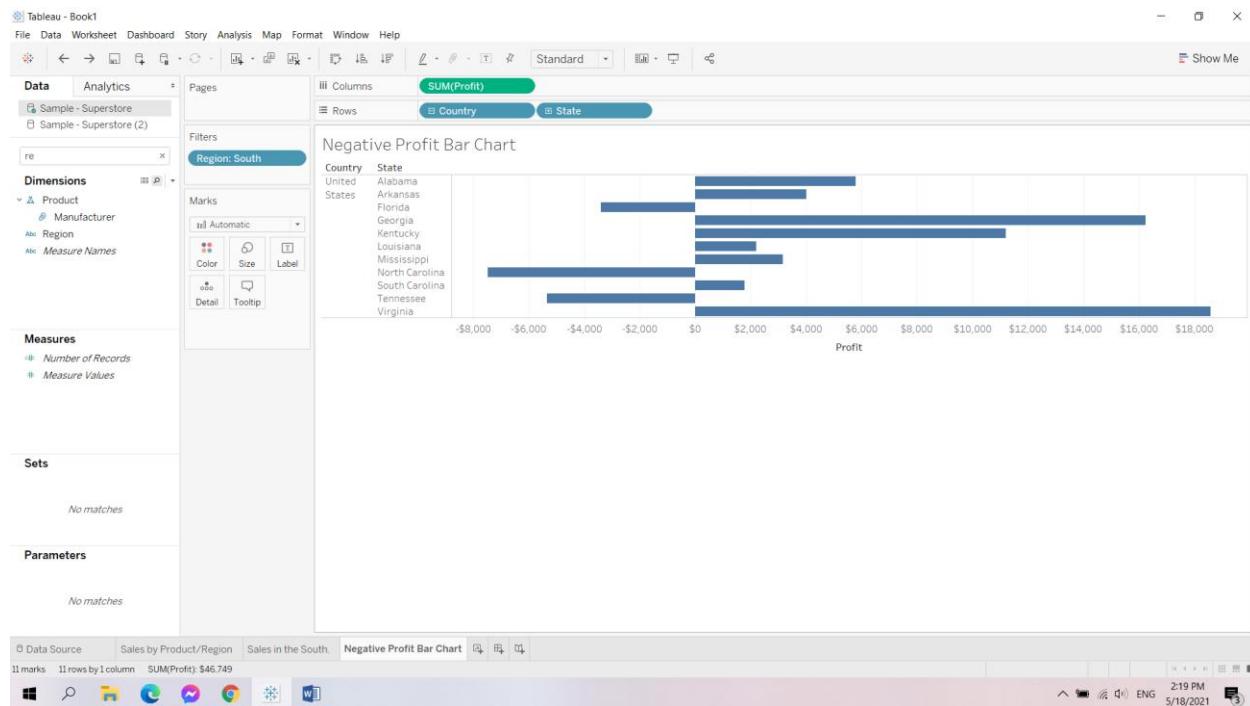
➤ Tạo chart mà biết đóng góp từng giá trị trong cột:



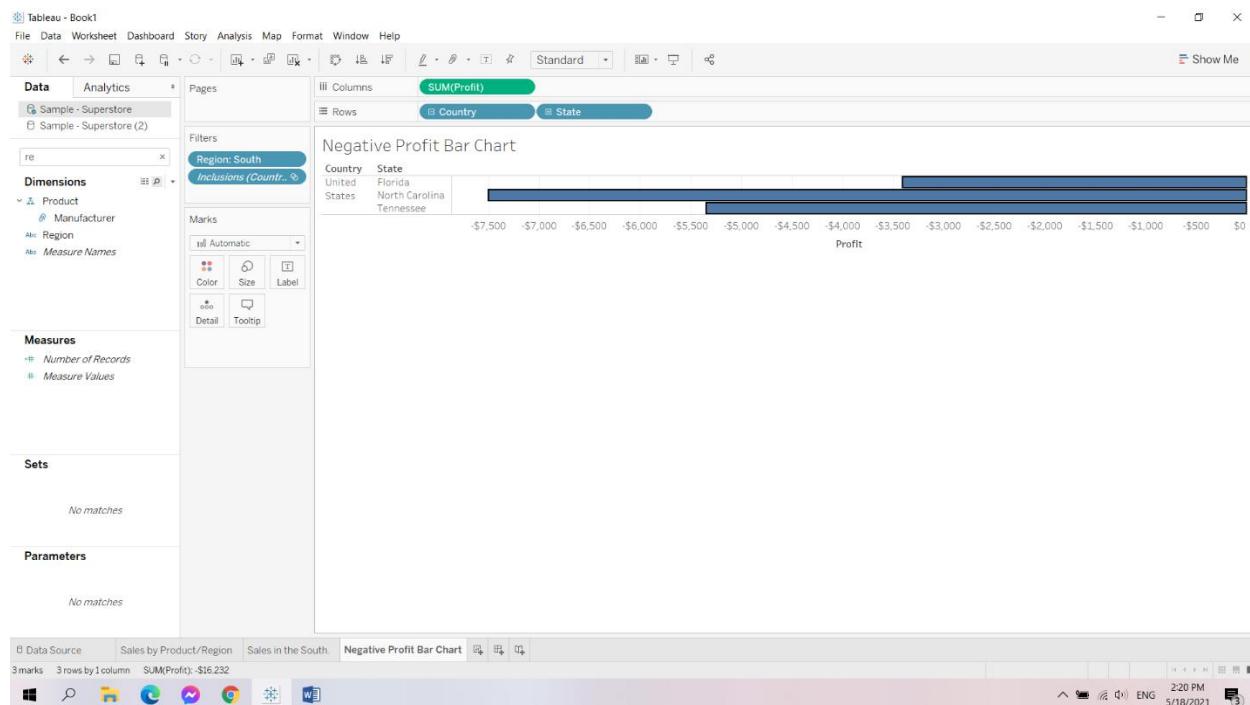
➤ Tạo chart mà chỉ thể hiện những giá trị người dùng mong muốn, không cần phải thể hiện hết các giá trị:



- Có thể tùy chọn giữ lại những cột, dòng..trong chart:
Chart đầy đủ, muốn giữ lại những cột mang giá trị âm



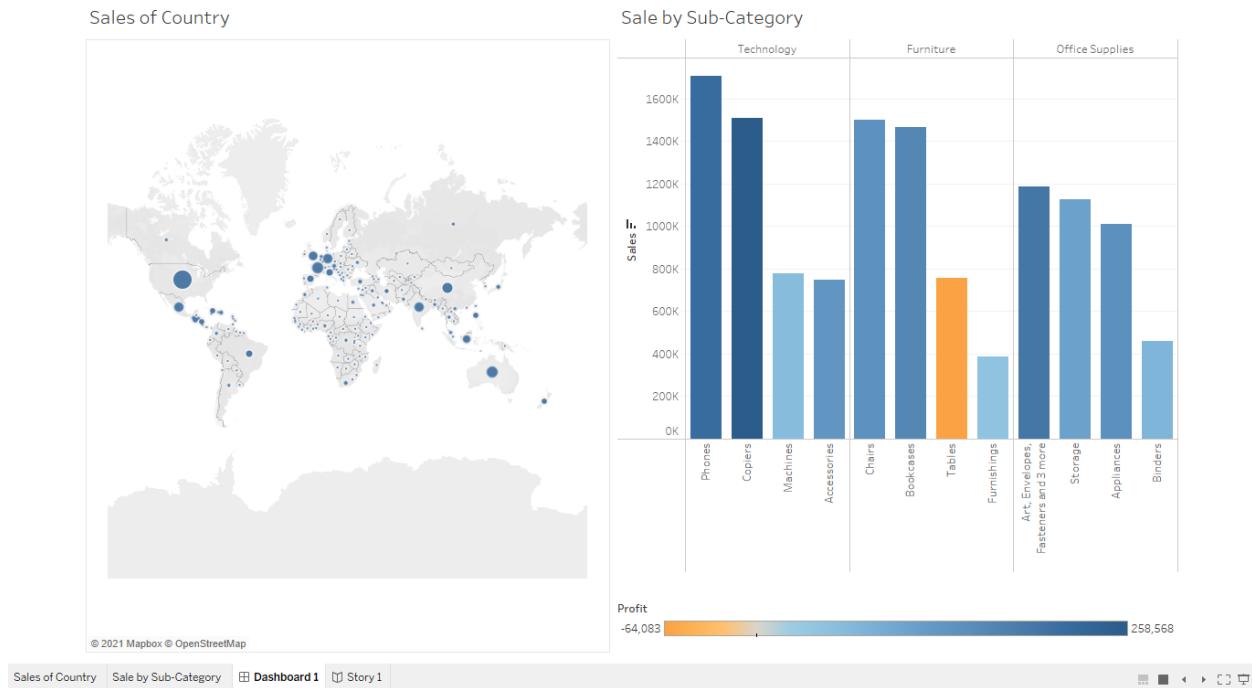
Sau khi thao tác:



- Dễ dàng tạo ra các Dashboard quản trị và các Story của dữ liệu:

➤ Dashboard:

Sau khi tạo ra được 2 biểu đồ ‘Sales of Country’ và ‘Sale by Sub-Category’. Chọn ‘New Dashboard’ và tiến hành kéo thả các biểu đồ theo ý muốn, thực hiện các thao tác chỉnh sửa theo mục đích thì ta sẽ tạo được Dashboard mà mình muốn



➤ Story:

Sau khi tạo ra được 2 biểu đồ ‘Sales of Country’ và ‘Sale by Sub-Category’. Chọn ‘New Story’ và tiến hành kéo thả các biểu đồ theo ý muốn, thực hiện các thao tác chỉnh sửa theo mục đích thì ta sẽ tạo được Story mà mình muốn

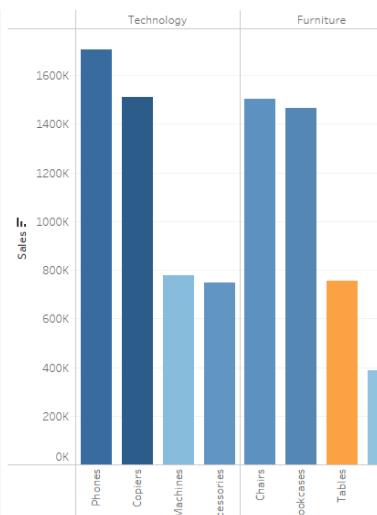
Story 1

< 1 2 >

Sales of Country



Sale by Sub-Category



Sales of Country Sale by Sub-Category Dashboard 1 Story 1



- Tạo bản đồ:

Tableau tích hợp sẵn nhiều thông tin trên bản đồ như thành phố, mã bưu chính, ranh giới hành chính,... Điều này khiến cho các bản đồ được tạo lập trên Tableau thường rất chi tiết và đầy đủ thông tin. Nhiều loại bản đồ khác nhau sẵn có trên Tableau như Bản đồ nhiệt, Bản đồ phân luồng, Bản đồ chuyên biệt, Bản đồ phân bố điểm,...

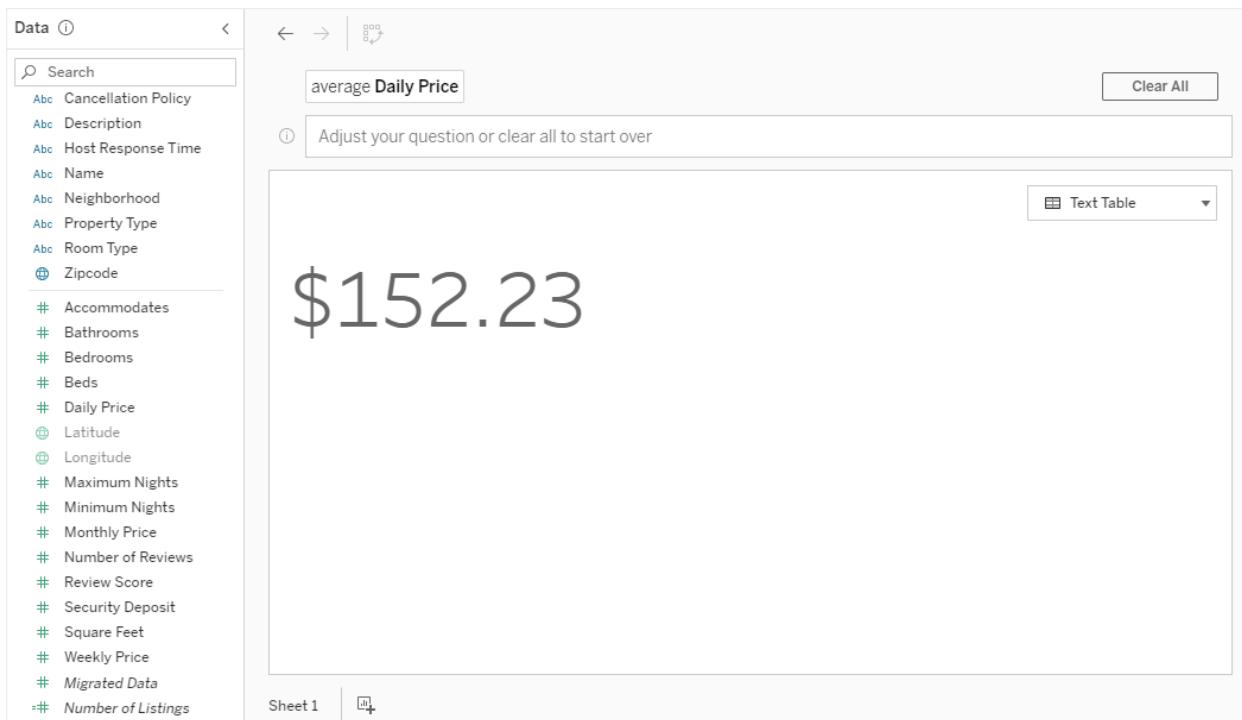
Ví dụ vẽ bản đồ thể hiện lượng Sales của các quốc gia trên thế giới:



- Ask Data:

Với tính năng này, ta chỉ cần gõ ra một truy vấn và Tableau sẽ hiển thị mọi đáp án phù hợp nhất. Các đáp án không chỉ ở dưới dạng văn bản mà còn là các hình ảnh trực quan. Tính năng này giúp ta dễ tiếp cận và tìm hiểu sâu hơn về dữ liệu cũng như thu được insight hay các quy luật mới

Ví dụ sử dụng tính năng Ask Data trên Seattle Airbnb data và câu hỏi là: average Daily Price



B. THỰC HÀNH

B.1. TRỰC QUAN BẰNG TABLEAU

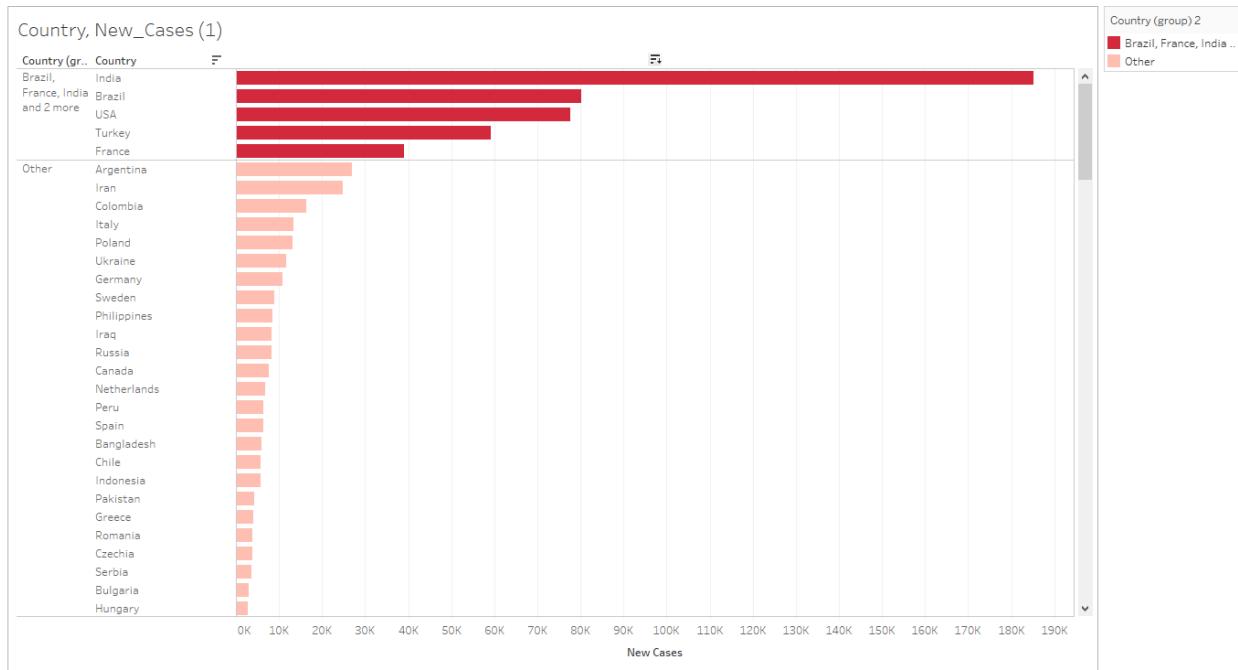
1. Country, New Cases

a. Lý do chọn các trường dữ liệu

Nhận xét được tình hình dịch bệnh ở một số quốc gia

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 1.1. Biểu đồ thể hiện tỉ lệ số ca nhiễm mới của các quốc gia: India, Brazil, USA, Turkey, France và Others (các quốc gia còn lại)



Hình 1.2. Biểu đồ thể hiện tỉ lệ số ca nhiễm mới của các quốc gia: India, Brazil, USA, Turkey, France và Others (các quốc gia còn lại)

- ✓ Tính phù hợp của biểu đồ:
 - Biểu đồ *hình 1.1* cho được cái nhìn trực quan tỉ lệ số ca nhiễm mới của các quốc gia từ đó dễ rút ra được tình hình dịch bệnh ở các quốc gia nhưng khó

so sánh được tổng số ca nhiễm của các quốc gia trong top 5 cao nhất so với tổng số ca nhiễm của các quốc gia còn lại.

- Biểu đồ *hình 1.2* cho được cái nhìn trực quan tỉ lệ số ca nhiễm mới của các quốc gia từ đó dễ rút ra được tình hình dịch bệnh ở các quốc gia, vì độ chênh lệch của dữ liệu là không quá nhỏ nên không quá khó trong việc nhận ra được sự chênh lệch giữa các phần và nhờ việc phân vùng ra nên ta có thể so sánh được tổng số ca nhiễm của các quốc gia trong top 5 cao nhất so với tổng số ca nhiễm của các quốc gia còn lại.

➔ Tập trung nhận xét bằng trực quan của *hình 1.2*

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Các quốc gia có số lượng ca nhiễm mới trong ngày chiếm tỉ lệ cao nhất, đáng báo động là: India, Brazil, USA, Turkey, France
- Tình hình dịch bệnh ở Ấn Độ đang cực kì nghiêm trọng, qua biểu đồ ta thấy được chỉ riêng Ấn Độ đã chiếm gần 25% số ca nhiễm mới trên thế giới
- Tình hình dịch bệnh ở Brazil (10.8%) và USA (10.4%) cũng ở mức báo động
- Chỉ riêng 4 quốc gia dẫn đầu về số ca nhiễm mới là: India, Brazil, USA và Turkey (54%) đã nhiều hơn tổng số ca nhiễm của các quốc gia còn lại

✓ Ý nghĩa:

Từ biểu đồ ta thấy được các quốc gia có tình hình dịch bệnh đáng báo động từ đó thúc đẩy các quốc gia ở các vùng và khu vực lân cận nâng cao kiểm soát và tăng cường phòng chống dịch.

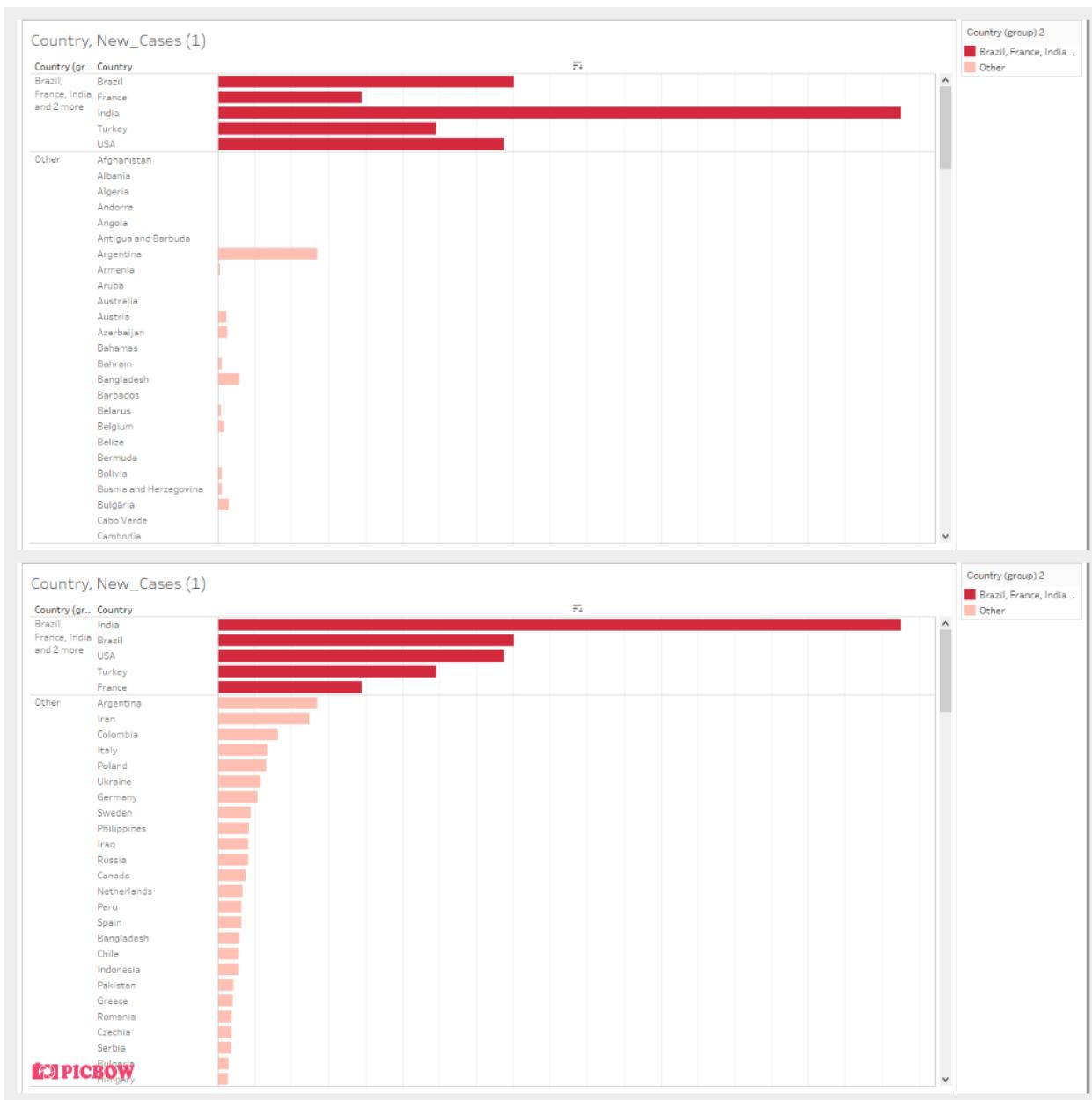
d. Sử dụng màu sắc để thể hiện dữ liệu

- Ở biểu đồ *hình 1.1*: Chọn màu đỏ để thể hiện cho các cột vì để thể hiện ý nghĩa của thuộc tính 'New_Cases' tức là số ca nhiễm mới trong ngày – một điều đáng báo động. Chọn màu đỏ đậm cho top 5 quốc gia cao nhất và màu đỏ nhạt cho các quốc gia còn lại để phân biệt và nhấn mạnh được mức độ nguy cấp của nhóm 5 quốc gia cao nhất.
- Ở biểu đồ *hình 1.2*: Chọn dải màu đỏ để thể hiện dữ liệu vì để phù hợp với ý nghĩa của thuộc tính 'New_Cases' tức là số ca nhiễm mới trong ngày – một điều đáng báo động. Và màu đỏ sẽ nhạt dần theo số ca nhiễm mới, điều này giúp ta dễ dàng nhận ra được mức độ nghiêm trọng của dịch bệnh ở các quốc gia.

e. Sử dụng các kỹ thuật đã học

✓ Manipulate View:

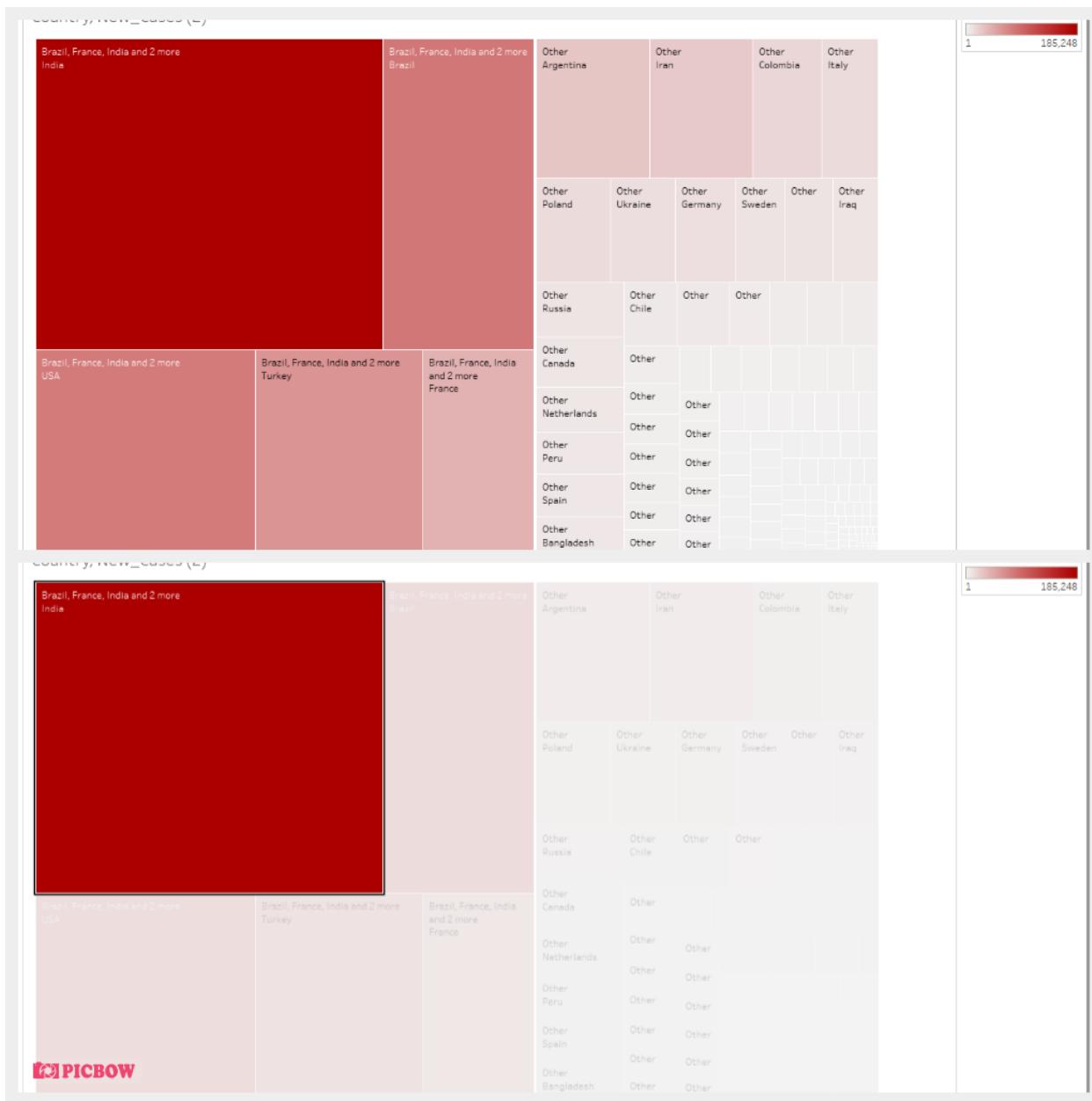
- Sort dữ liệu theo thuộc tính 'New_Cases'



Trước và sau khi sort dữ liệu theo thuộc tính New_Cases

Có thể thấy trước khi sort dữ liệu theo thuộc tính 'New_Cases' biểu đồ rất lộn xộn, khó xem xét. Việc lựa chọn sort dữ liệu đã giúp cho biểu đồ trực quan hơn, dễ dàng thấy được quốc gia có số ca nhiễm mới cao nhất, thấp nhất, dễ so sánh số ca nhiễm mới của các quốc gia.

- Highlight dữ liệu



Trước và sau khi highlight India

Dùng highlight để nhấn mạnh vào Ấn Độ - quốc gia có số ca nhiễm mới cao nhất, đáng báo động nhất.

2. Country, New Deaths

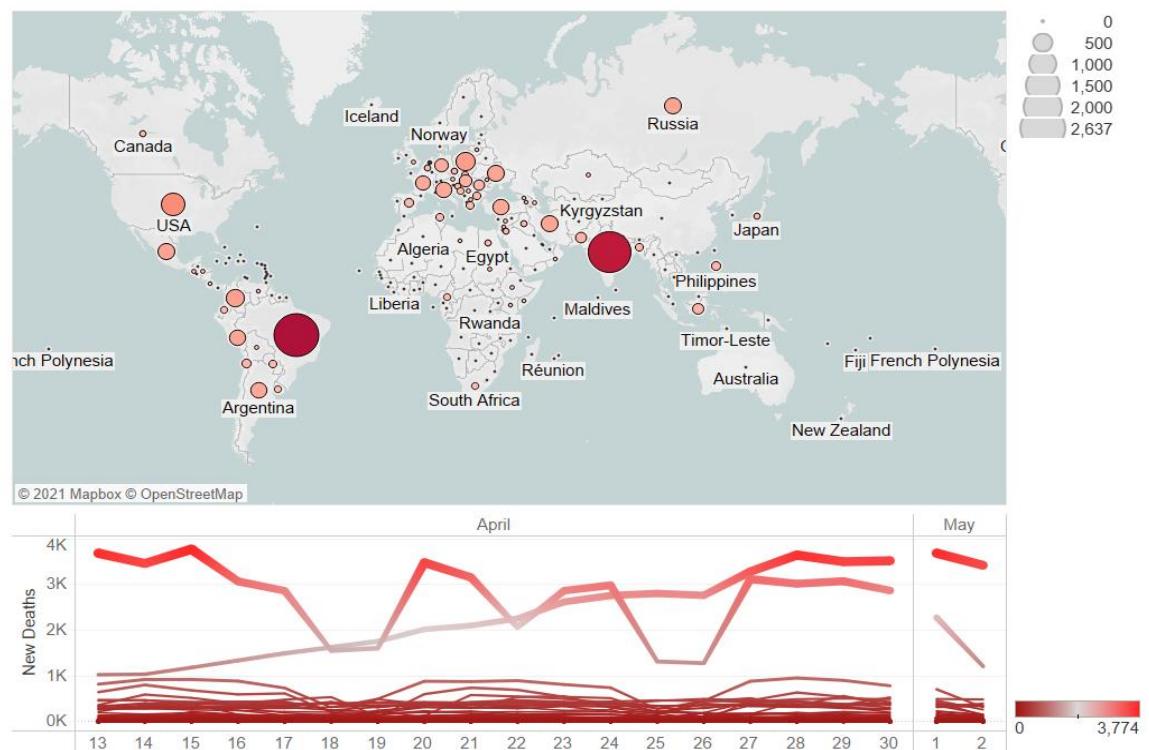
a. Lý do chọn các trường dữ liệu

Quan sát, phân tích được tình hình tổng quan của dịch bệnh trên Thế giới.

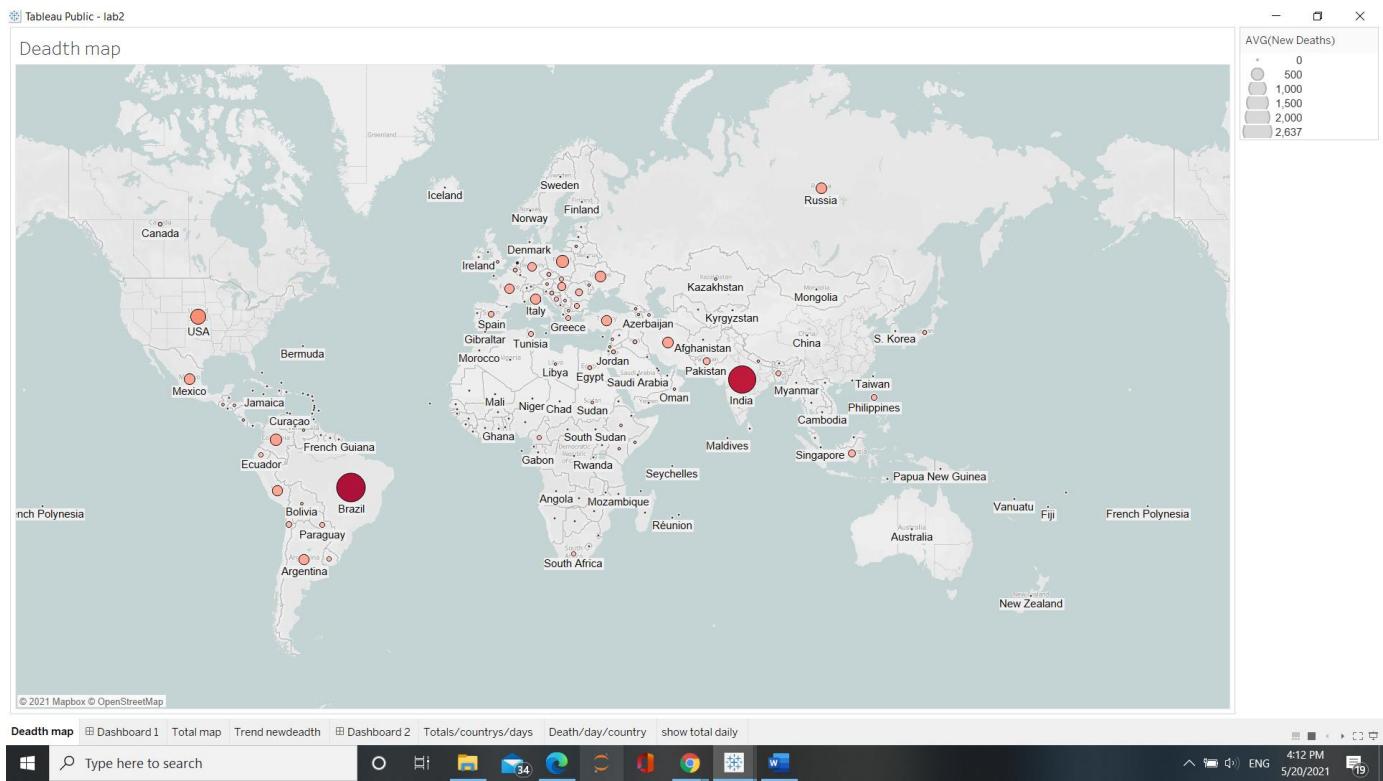
b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:

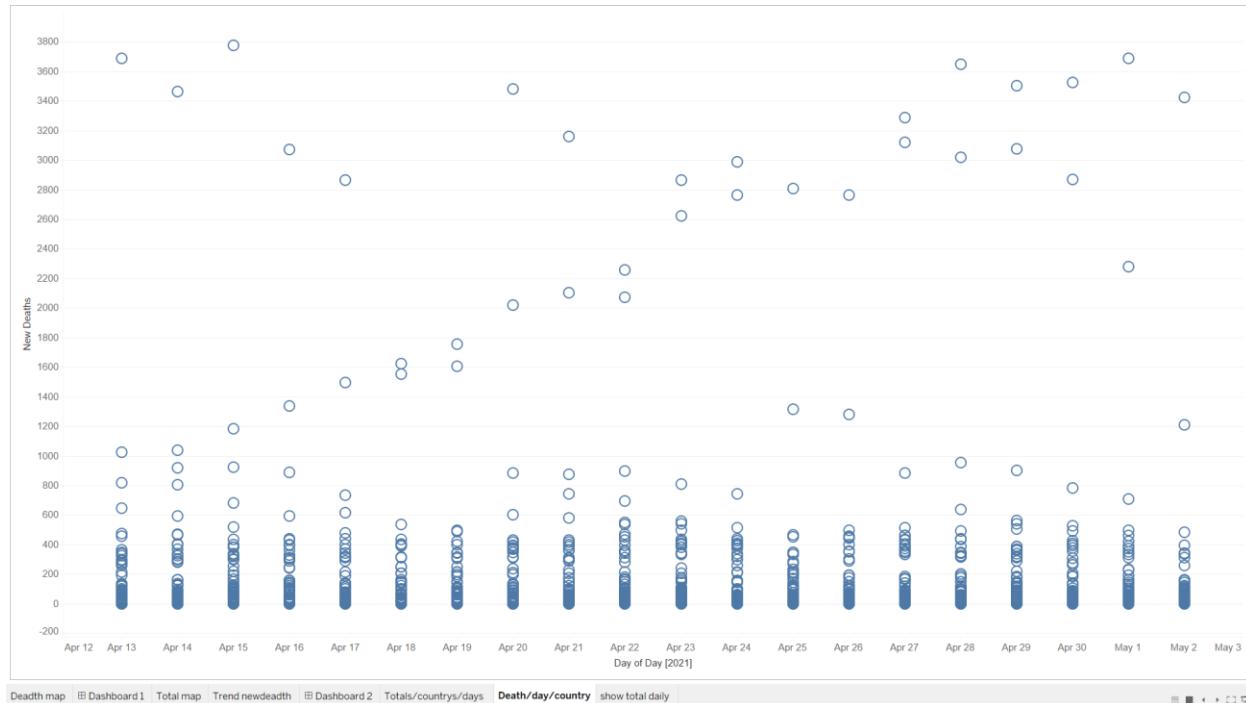
Global NewDeaths Case



Hình 2.1.Dashboard thể hiện số ca nhiễm mới của các quốc gia trên TG



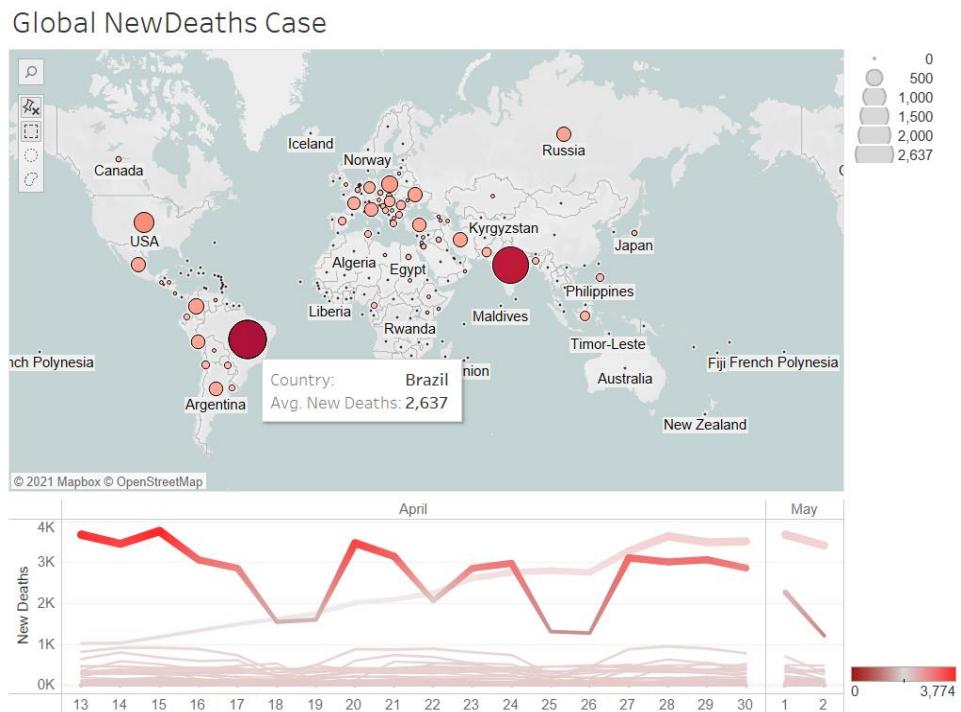
Hình 2.2.1.Biểu đồ thể phân bố TB số ca chết mới trên toàn cầu từ ngày 13/04-02/05/2021



Hình 2.2.2.Biểu đồ thể phân bố số ca nhiễm mới trên toàn cầu từ ngày 13/04-02/05/2021

- ✓ Tính phù hợp của biểu đồ:

- Biểu đồ *hình 2.2.1(map)* cho được cái nhìn trực quan sự phân bố về số ca chết mới của các quốc gia từ đó dễ rút ra được tình hình tổng quát dịch bệnh trên thế giới và khu vực, quốc gia => Quan sát/so sánh được với tình hình các quốc gia lân cận/trong vùng.
- Biểu đồ *hình 2.2.1* cho ta thấy được những quốc gia nào có tỉ lệ ca chết vì dịch bệnh cao trong 20 ngày qua.
- Biểu đồ *hình 2.2.2* thể hiện được sự phân bố của trung bình số ca chết mỗi quốc gia thường tập trung vào khoảng giá trị nào ->rút ra xu hướng chung. => Từ 2 biểu đồ 2.2.1 và 2.2.2 ta có được dashboard 2.2 thể hiện đầy đủ sự phân bố/xu hướng chung toàn cầu và các quốc gia có ca chết cao nhất trong những ngày qua.



c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Các quốc gia có Trung bình số lượng ca chết mới trong 20 ngày cao vượt trội, đáng báo động là: India, Brazil.

- Nhìn chung trong 20 ngày qua, số ca chết mỗi ngày của 1 quốc gia thường <1000 ca và có xu hướng ít biến động, nhiều quốc gia không có số ca chết mới trong nhiều ngày.

=>Tình hình chung: ổn định và kiểm soát được.

- Duy chỉ có 2 quốc gia là India và Brazil là số ca chết mỗi ngày biến động liên tục và cao vượt trội.
- India có xu hướng tăng vọt trong suốt 20 ngày qua
 - Tình hình nguy cấp.
- Brazil biến động mạnh qua mỗi ngày => Tình hình dịch bệnh khó kiểm soát và chưa ổn định.

✓ Ý nghĩa:

Từ biểu đồ ta thấy được các quốc gia có tình hình dịch bệnh đáng báo động số ca chết tăng mạnh so với tình hình chung thế giới.

d. Sử dụng màu sắc để thể hiện dữ liệu

- Ở biểu đồ *hình 2.2.1*: Chọn dải màu đỏ để thể hiện cho các cột vì để thể hiện ý nghĩa của thuộc tính ‘New_death’ tức là số ca chết mới trong ngày – một điều đáng báo động. và sử dụng dải màu để thể hiện sự thay đổi mỗi ngày.
Và dễ nhận thấy khi thể hiện sự phân bố trên map và khi highlight biến động số ca chết mới qua mỗi ngày của 1 quốc gia nào.

e. Sử dụng các kỹ thuật đã học

- ✓ Manipulate View
- ✓ Facet
- ✓ Reduce

Sử dụng kết hợp các kỹ thuật trên giúp viewer dễ tương tác và trực quan:

- Người sử dụng dễ dàng xem các thông tin của 1 quốc gia ngay trên map khi Hover chuột trên map (Change view over time), reduce thuộc tính/thông tin cần thiết cho người quan sát thấy (thể hiện nội dung chính cần quan tâm là newdeath cases).
- Xem chi tiết xu hướng hay sự biến động newdeaths qua mỗi ngày của 1 quốc gia cụ thể(facet across multive view) khi select quốc gia trên map.
- So sánh trực quan với các quốc gia khác và rút ra được điểm khác bằng cách hightlight.

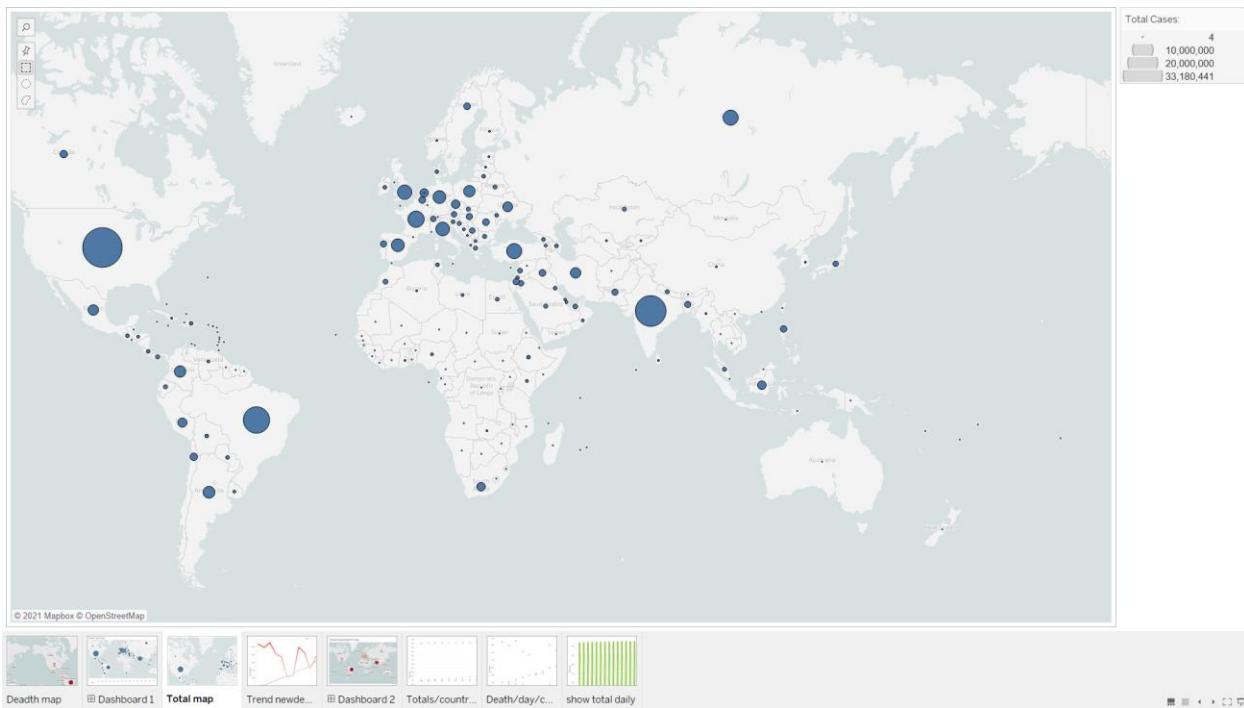
3. Country Total Cases

a. Lý do chọn các trường dữ liệu

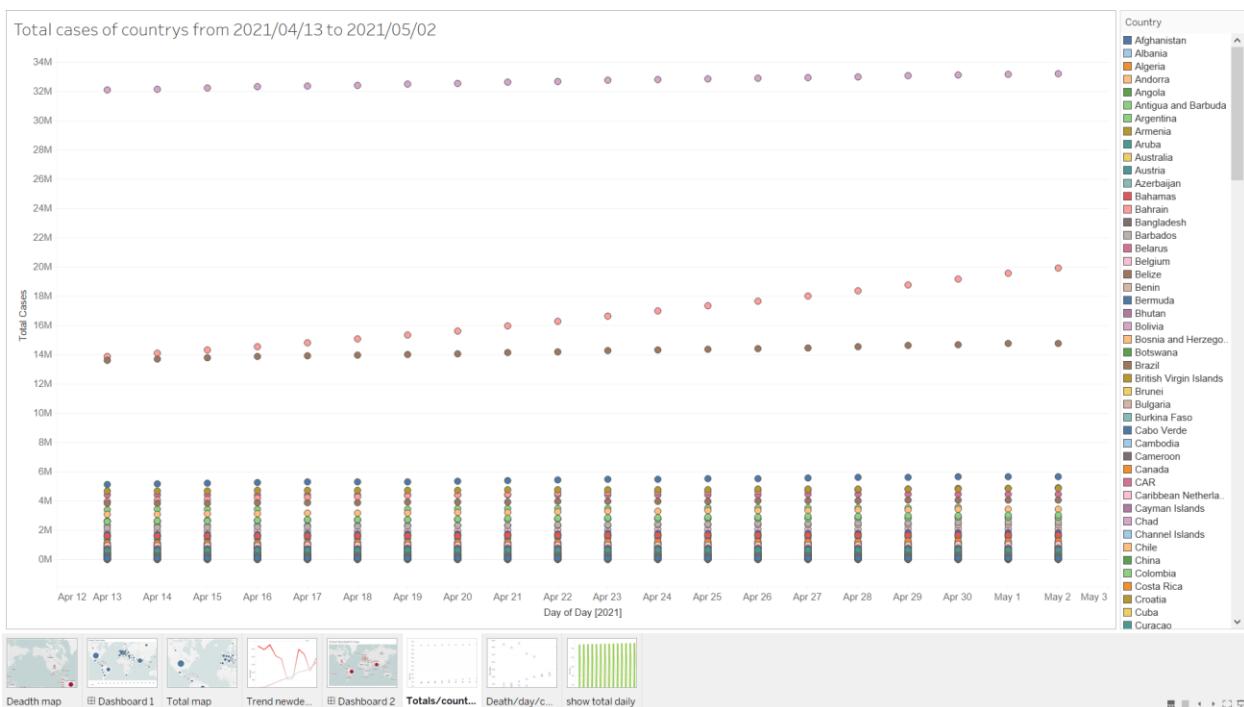
Quan sát sự phân bố ,tình hình kiểm soát dịch bệnh toàn cầu và các quốc gia trong suốt đại dịch và,tình hình kiểm soát trong 20 ngày vừa qua.

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



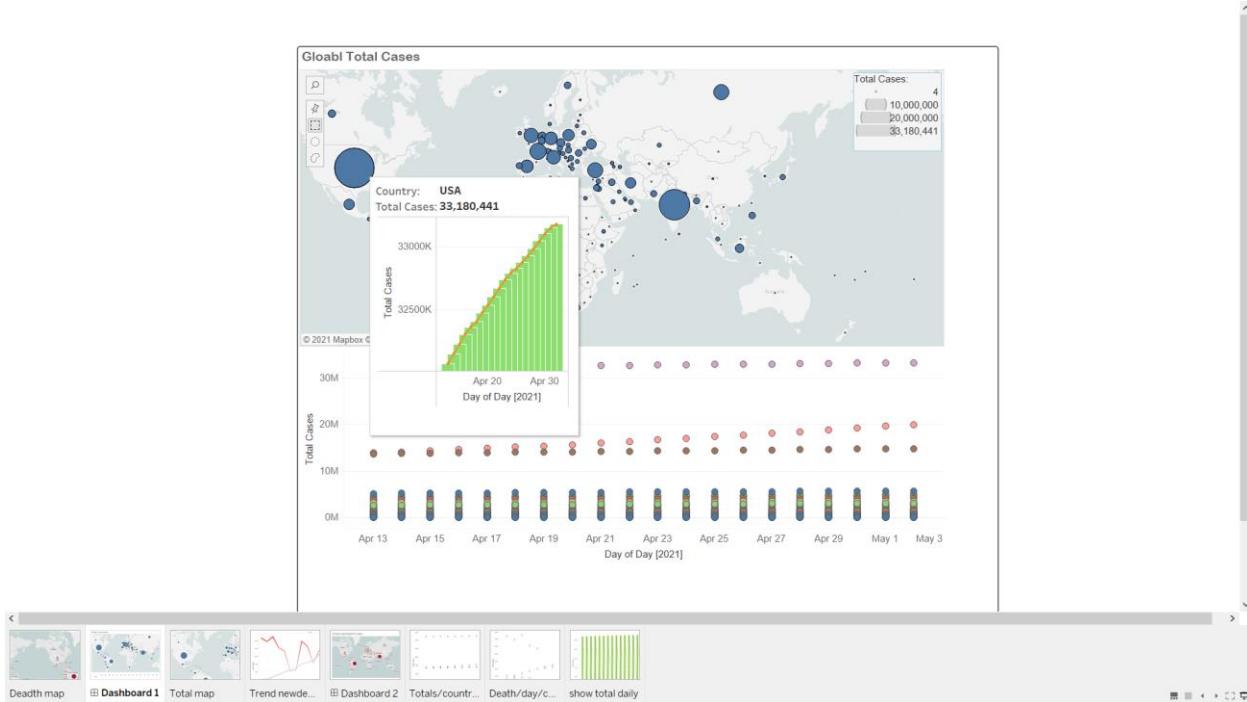
Hình 3.1. Biểu đồ thể hiện sự phân bố tổng số ca nhiễm trên toàn thế giới đến 02/05/2021



Hình 3.2. Biểu đồ thể sự tăng tổng ca nhiễm qua mỗi ngày của mỗi quốc gia.

✓ Tính phù hợp của biểu đồ:

- Biểu đồ *hình 3.1* cho thấy được sự phân bố số ca nhiễm của các quốc gia/ khu vực đến hiện tại .
- Biểu đồ *hình 3.2* trực quan được tổng số ca. nhiễm của mỗi quốc gia trên thế giới đa số nằm trong khoảng bao nhiêu
⇒ Dasboard 4.0 vừa thể hiện được cả 2 điều trên.



c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Quốc gia có tổng số ca nhiễm tính đến ngày 02/05 cao nhất là USA.
- Có thể chia các quốc gia thành 3 nhóm theo tổng số ca nhiễm hiện tại:
 - Nhóm các quốc gia có số ca nhiễm dưới 10 triệu ca
 - Nhóm các quốc gia có số ca nhiễm từ 10 triệu ca – 30 triệu ca
 - Nhóm các quốc gia có số ca nhiễm trên 30 triệu ca
- Nhìn chung các quốc gia trên thế giới đều có tổng số ca nhiễm tính đến thời gian xét là dưới 10 triệu ca
⇒ Tình hình chung thế giới
- Duy chỉ có Mỹ là quốc gia cao vượt trội(33 triệu ca), ngay sau đó là India(19,9 triệu) và Brazil(14,7 triệu ca) là cao hơn nhiều so với xu hướng chung.
- Tổng số ca nhiễm của các quốc gia trên thế giới trong 20 ngày qua là tương đối ổn định => xu hướng chung.

- Riêng India có xu hướng tăng mạnh (có độ dốc cao hơn so với các nước còn lại)
⇒ Cần kiểm soát lại tình hình.
- ✓ Ý nghĩa:
Tình hình dịch bệnh thế giới 20 ngày qua tương đối ổn định.
Mỹ vẫn là quốc gia có tổng số ca nhiễm cao nhất.

d. Sử dụng màu sắc để thể hiện dữ liệu

Thể hiện dải màu khác nhau cho Country để giúp view thấy rõ sự khác biệt của Mỹ so với các quốc gia còn lại.

e. Sử dụng các kỹ thuật đã học

- ✓ Manipulate View
- ✓ Facet
- ✓ Reduce

Sử dụng kết hợp các kỹ thuật trên giúp viewer dễ tương tác và trực quan:

- Người sử dụng dễ dàng xem các thông tin của 1 quốc gia ngay trên map khiHover chuột trên map (Change view over time), reduce thuộc tính/thông tin cần thiết cho người quan sát.
- Xem chi tiết xu hướng hay sự biến động tổng số ca nhiễm qua mỗi ngày của 1 quốc gia cụ thể (facet across multive view) khi select quốc gia trên map.
- So sánh trực quan với các quốc gia khác và rút ra được điểm khác bằng cách highlight.

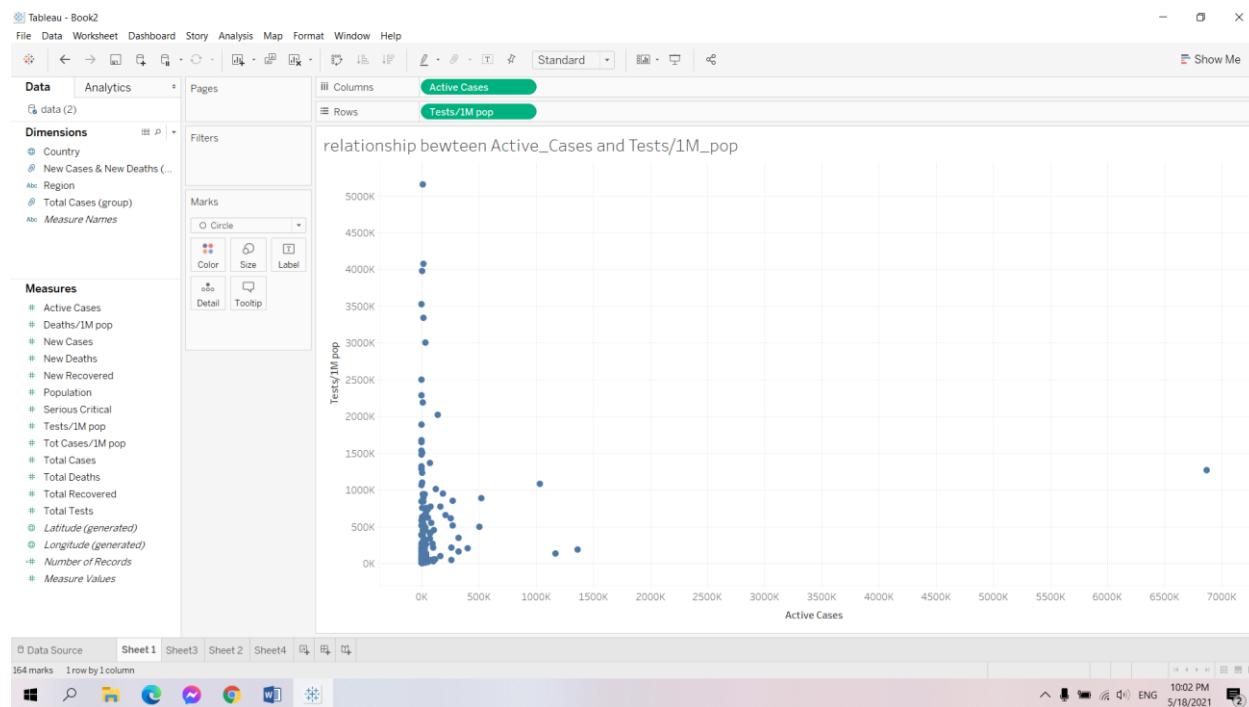
4. Active Cases, Tests/1M pop

a. Lý do chọn các trường dữ liệu

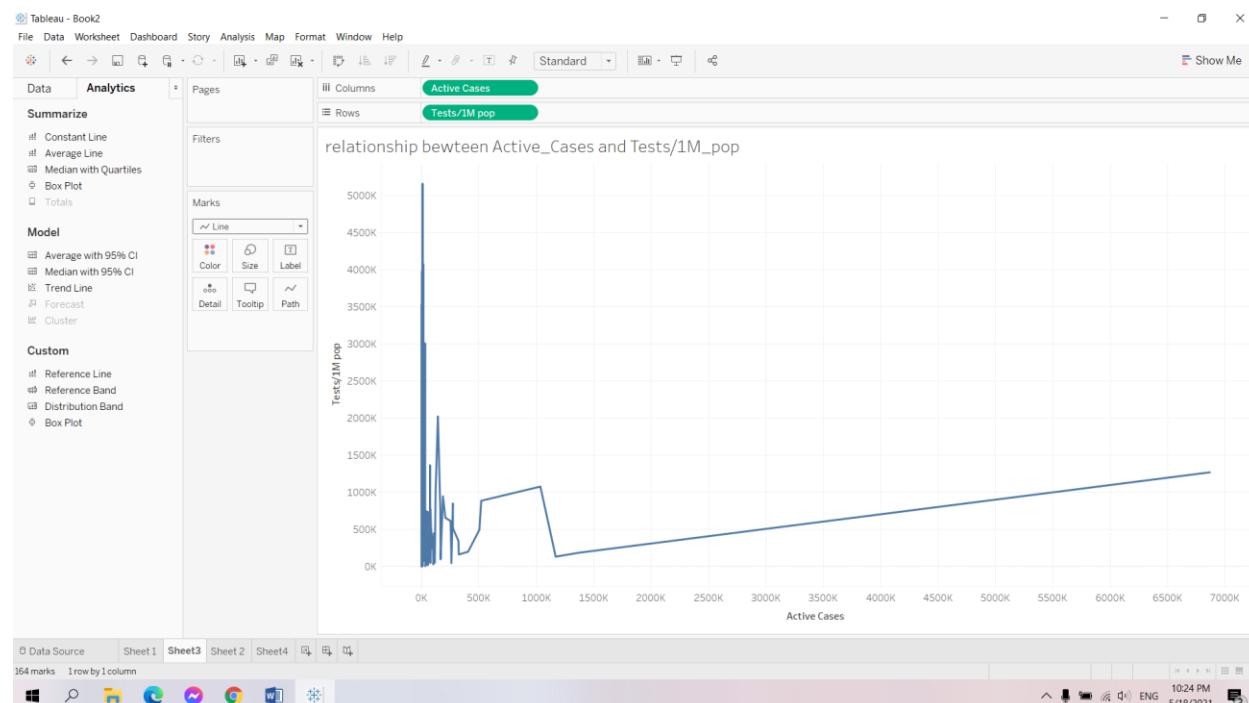
Nhận xét mối quan hệ của những ca đang mắc với số lần xét nghiệm/triệu người.

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 4.1. Biểu đồ thể hiện mối quan hệ của những ca đang mắc với số lần xét nghiệm trên triệu người



Hình 4.2. Biểu đồ thể hiện mối quan hệ của những ca đang mắc với số lần xét nghiệm trên triệu người

- ✓ Tính phù hợp của biểu đồ:
- ⇒ Chọn biểu đồ **hình 4.1**.

- Trực quan được mối quan hệ giữa các trường dữ liệu được chọn.
- Biểu đồ dễ nhìn, dễ thấy được sự tăng giảm của trường này có ảnh hưởng thế nào đối với trường còn lại.
- Vì giá trị là rời rạc nên việc chọn biểu đồ *hình 4.2* không phù hợp

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Từ biểu đồ có thể thấy các quốc gia mà có số ca đang mắc lớn thì có tỉ lệ test/triệu người bé.
- Từ biểu đồ có thể thấy các quốc gia có tỉ lệ test/triệu người cao thì số ca đang mắc bé.
- Từ biểu đồ có thể thấy các nước có tỉ lệ test/triệu người thấp thì số ca mắc bé và ngược lại. Có thể các nước này kiểm soát dịch khá tốt, phần nào cách ly được người đang mắc bệnh với cộng đồng.

✓ Ý nghĩa:

Từ biểu đồ thấy tầm quan trọng của việc xét nghiệm để tìm ra kết quả. Việc có nhiều ca xét nghiệm sẽ càng giúp phần nào cho việc hạn chế lây lan dịch bệnh. Tình hình dịch bệnh càng căng thẳng nếu người dân có bệnh mà không đến cơ sở y tế để xét nghiệm mà tự ý điều trị ở nhà. Ngoài ra có thể nhiều quốc gia không đủ khả năng để làm nhiều xét nghiệm.

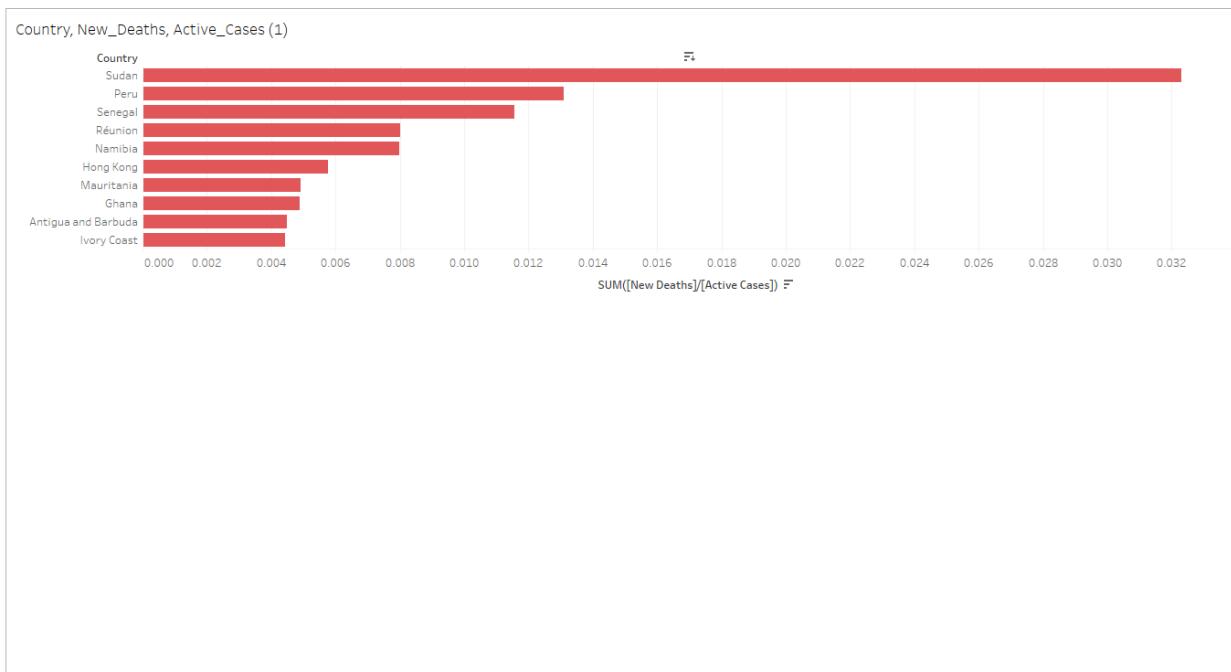
5. Country, New Deaths, Active Cases

a. Lý do chọn các trường dữ liệu

Quan sát được các quốc gia có tỉ lệ ca tử vong mới/tổng số ca đang nhiễm cao qua đó thấy được tình hình kiểm soát dịch bệnh và khả năng y tế của các quốc gia này

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

✓ Trực quan dữ liệu:



Hình 5.1. Biểu đồ thể hiện top 10 quốc gia có tỉ lệ số ca tử vong mới/số ca còn nhiễm hiện tại



Hình 5.2. Biểu đồ thể hiện top 10 quốc gia có tỉ lệ số ca tử vong mới/số ca còn nhiễm hiện tại

✓ Tính phù hợp của biểu đồ:

- Biểu đồ *hình 5.1* trực quan được top 10 các quốc gia có tỉ lệ số ca tử vong mới/số ca còn nhiễm cao, biểu đồ dẽ nhìn, thể hiện được rõ ràng số liệu của từng quốc gia.

- Biểu đồ *hình 5.2* trực quan được top 10 các quốc gia có tỉ lệ số ca tử vong mới/số ca còn nhiễm cao, nhưng vì ở đây dữ liệu chênh lệch khá nhỏ nên các ô có thể bị sát nhau gây khó khăn trong việc quan sát.
➔ Tập trung nhận xét bằng trực quan của *hình 5.2*

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Các quốc gia có số lượng ca tử vong mới/số lượng ca còn nhiễm trong ngày chiếm tỉ lệ cao nhất, đáng báo động là: Sudan, Peru, Senegal, Réunion, Namibia, Hong Kong, Mauritania, Ghana, Antigua and Barbuda, Ivory Coast
- Ta có thể thấy các quốc gia trong top chủ yếu ở khu vực Châu Phi, nó phản ánh đúng thực tế rằng khả năng y tế ở Châu Phi còn thấp

✓ Ý nghĩa:

Từ biểu đồ ta thấy được các quốc gia có số lượng ca tử vong mới/số lượng ca còn nhiễm trong ngày chiếm tỉ lệ cao chủ yếu là ở khu vực Châu Phi vì vậy ở khu vực này cần tăng cường, nâng cao khả năng y tế để đối phó tốt hơn với dịch bệnh

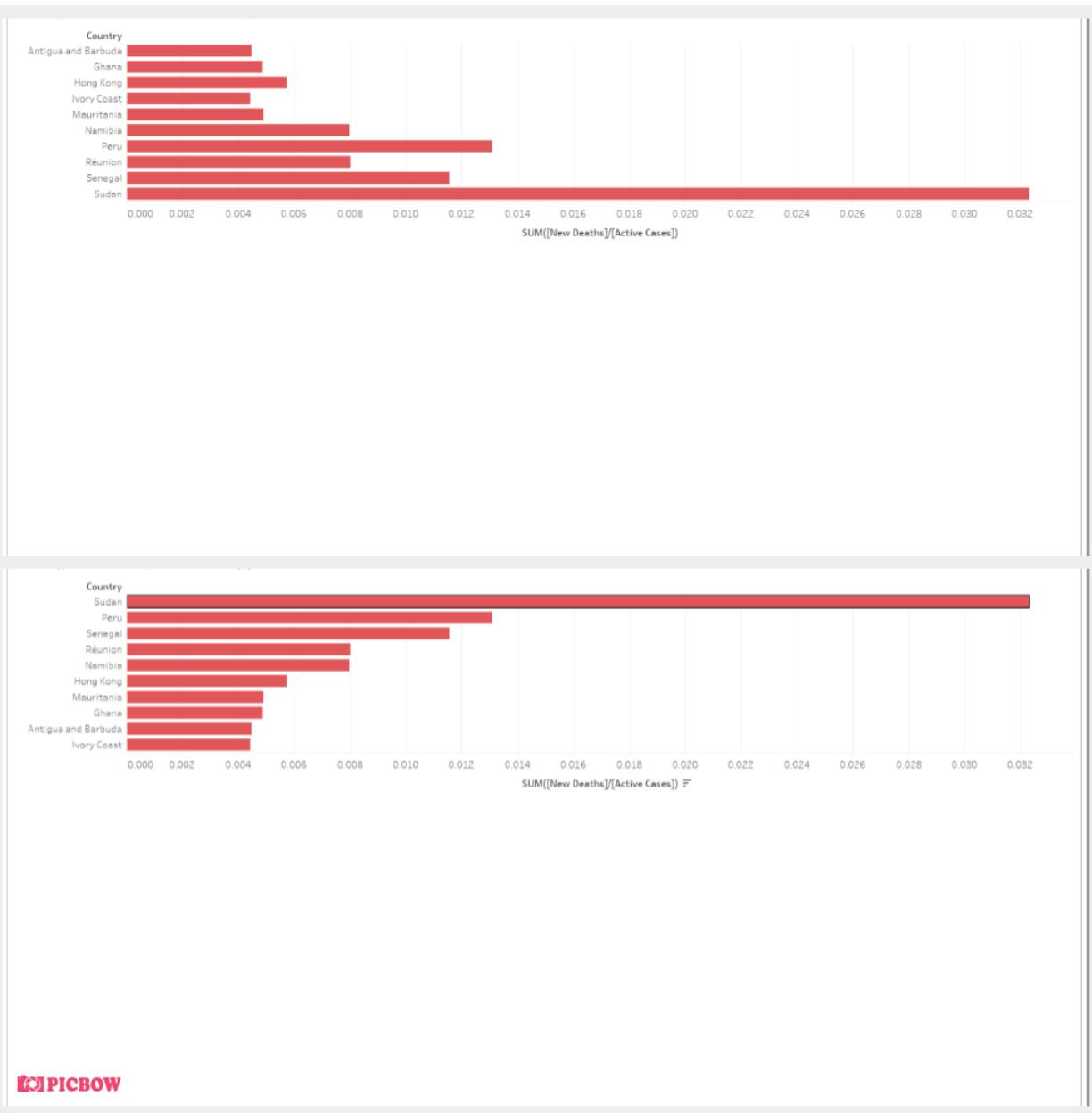
d. Sử dụng màu sắc để thể hiện dữ liệu

- ✓ Ở biểu đồ *hình 5.1*: Sử dụng màu đỏ cho các cột để thể hiện đúng ý nghĩa mà biểu đồ muốn đề cập đó là số lượng ca tử vong mới/số lượng ca còn nhiễm trong ngày – một điều đáng báo động.

e. Sử dụng các kỹ thuật đã học

✓ Manipulate View:

Sort dữ liệu theo thuộc tính 'New_Deaths/Active_Cases'

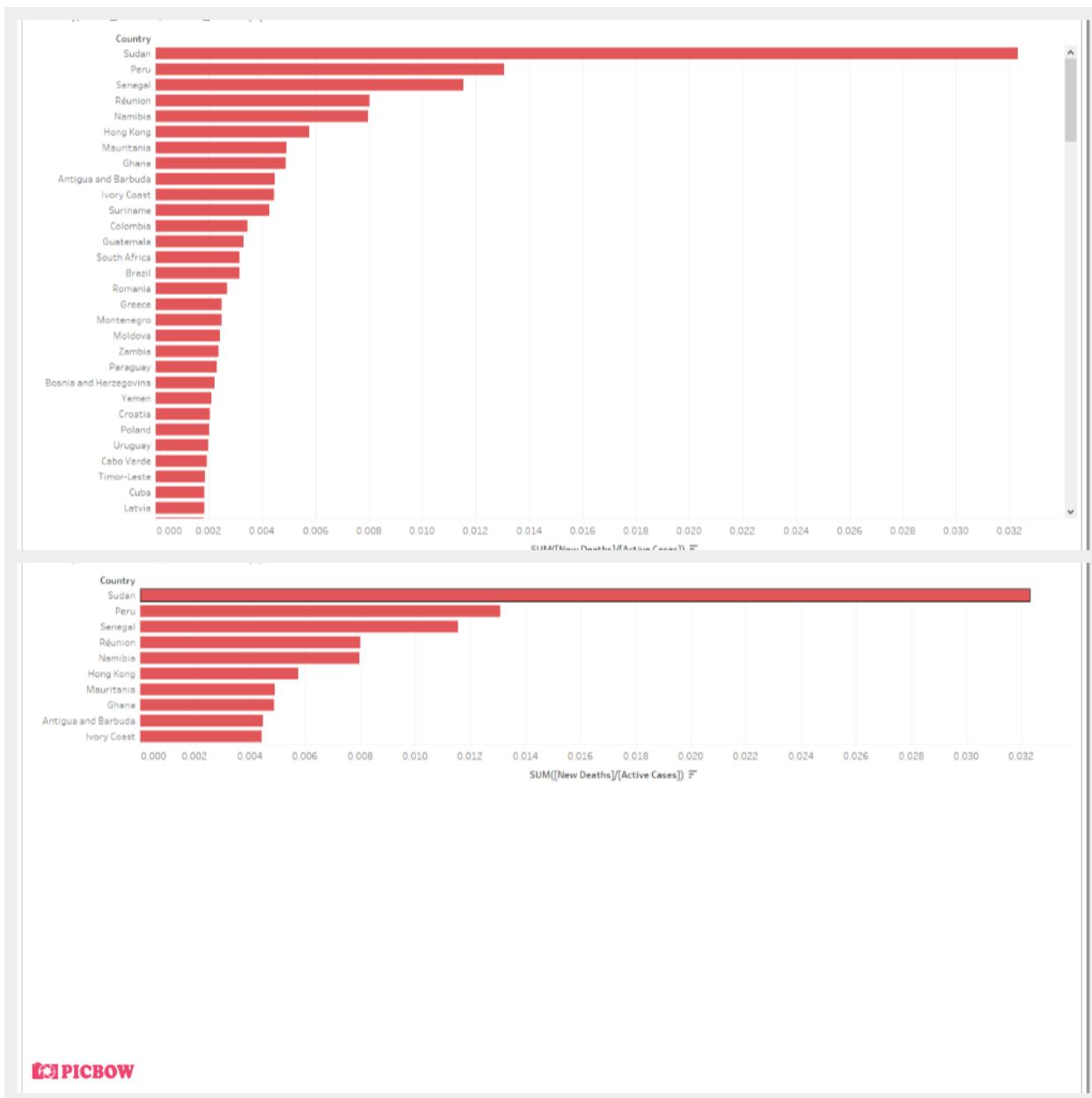


Trước và sau khi sort dữ liệu theo thuộc tính New_Deaths/Active_Cases

Có thể thấy trước khi sort dữ liệu theo thuộc tính 'New_Deaths/Active_Cases' biểu đồ rất lộn xộn, khó xem xét. Việc lựa chọn sort dữ liệu đã giúp cho biểu đồ trực quan hơn, dễ dàng thấy được quốc gia có số lượng ca tử vong mới/số lượng ca còn nhiễm trong ngày cao nhất, thấp nhất, dễ so sánh số lượng ca tử vong mới/số lượng ca còn nhiễm trong ngày của các quốc gia.

✓ Reduce:

Vì dữ liệu có quá nhiều quốc gia nên nếu để yên như vậy trực quan thì việc quan sát hơi bất cập và ta cũng không cần phải quan sát nhiều quốc gia như vậy. Do đó, sử dụng filter để giảm số lượng quốc gia xuống, chỉ lấy top 10.



Biểu đồ trước và sau khi sử dụng filter để giảm số quốc gia

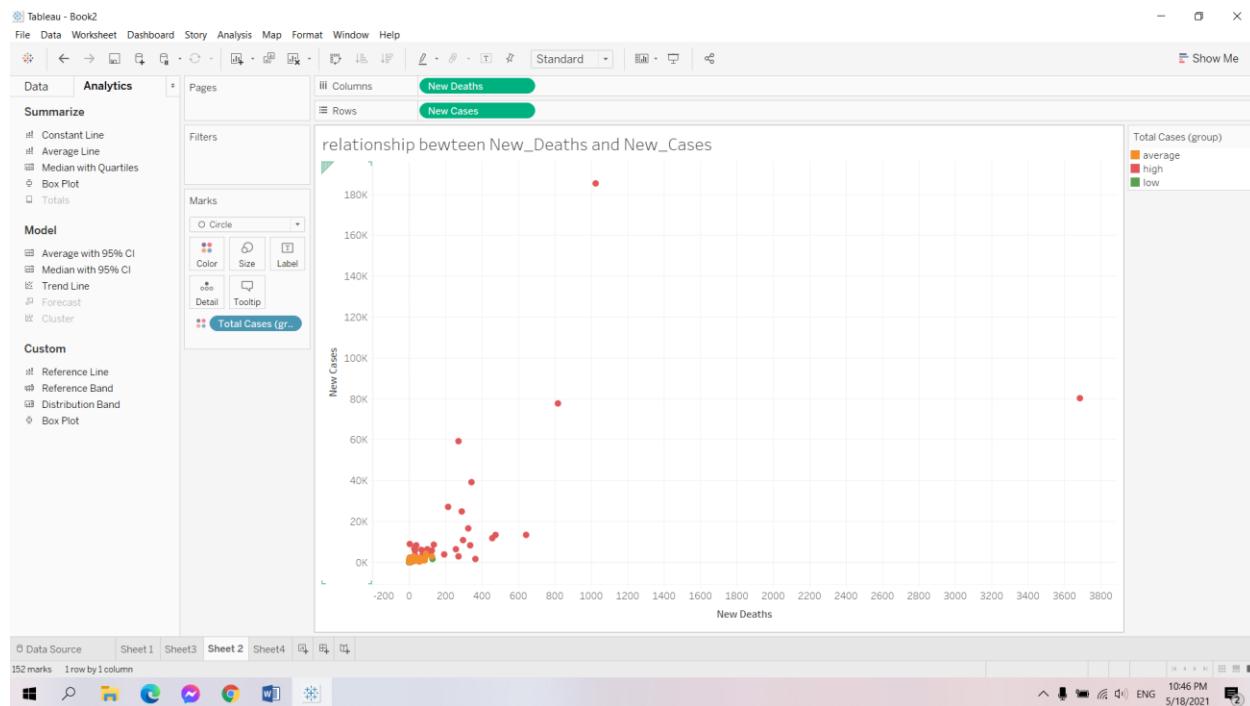
6. Total Cases, New Deaths, New Cases

a. Lý do chọn các trường dữ liệu

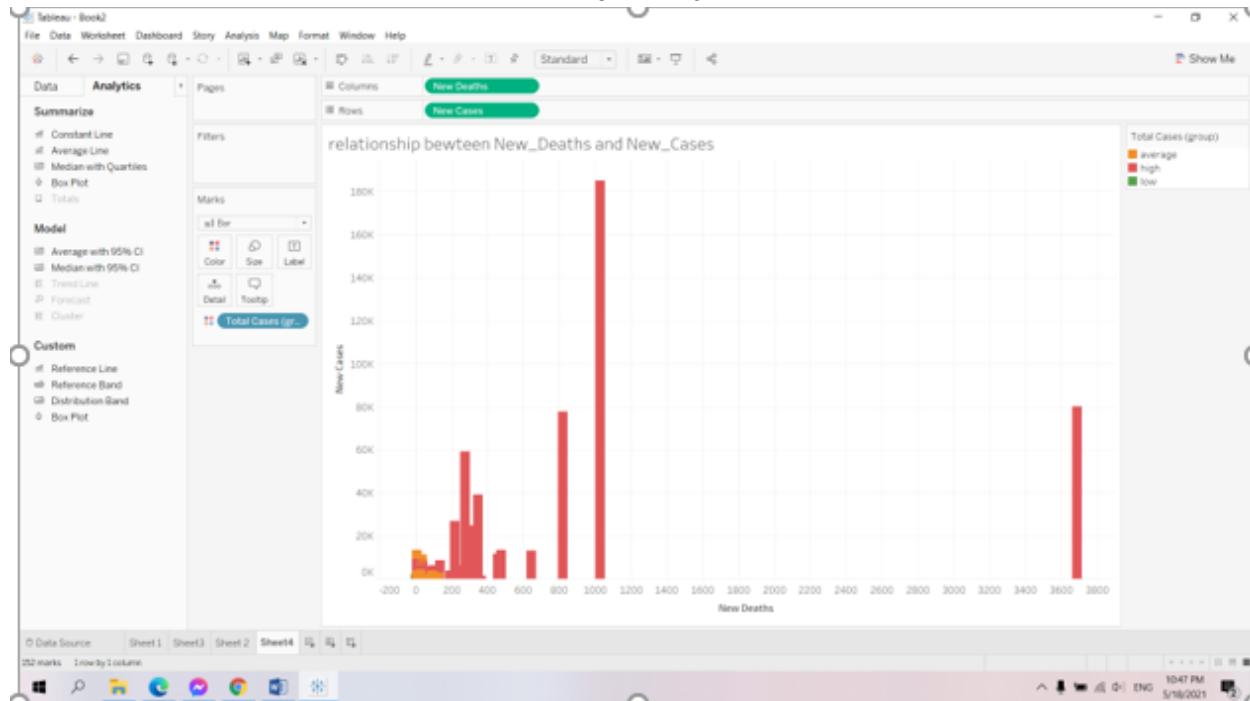
Nhận xét mối quan hệ số người tử vong với số trường hợp mắc mới theo từng nhóm nước được xác định theo tổng ca mắc, nhóm nước có ca mắc lớn từ 500.000 ca trở lên, nhóm nước trung bình có số ca mắc từ 100.000 đến 500.000 ca trở lên, nhóm nước thấp có số ca mắc dưới 100.000 ca.

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 6.1. Biểu đồ thể hiện mối quan hệ của những ca mắc mới và những ca mới tử vong theo từng nhóm nước



Hình 6.2. Biểu đồ thể hiện mối quan hệ của những ca mắc mới và những ca mới tử vong theo từng nhóm nước

- ✓ Tính phù hợp của biểu đồ:
 - Trực quan được mối quan hệ của trường dữ liệu được chọn.
 - Biểu đồ dễ nhìn, dễ thấy được sự tăng giảm của trường này có ảnh hưởng thế nào đối với trường còn lại.
 - Dễ nhìn thấy những từng nhóm nước có tình hình ca mới tử vong và những ca mới mắc như thế nào.

c. Nhận xét và rút ra ý nghĩa

- ✓ Nhận xét:
 - Trong nhóm nước có tổng ca mắc cao, ta thấy được:
 - Một số nước có ca mắc mới và ca tử vong mới đều cao, có vẻ đây là nhóm nước đang trong tình trạng kiểm soát dịch gặp khó khăn.
 - Một số nước có ca mắc mới và ca tử vong mới đều thấp, có vẻ đây là nhóm nước từng bị bùng phát dịch, nhưng hiện tại kiểm soát dịch tốt.
 - Trong nhóm nước có tổng ca mắc trung bình và thấp, có ca mắc mới và ca tử vong mới khá thấp so với thế giới, có thể đoán một là kiểm soát dịch tốt, hoặc đang trong thời kì đầu của bùng dịch.
- ✓ Ý nghĩa:

Có thể thấy rõ ràng các nước có tổng ca mắc cao, có số ca mắc mới và ca tử vong đều cao có thể đang trong giai đoạn bùng dịch, chính vì vậy cần có những chiến lược để ngăn chặn bệnh dịch, ngoài ra người dân nên chung tay chống dịch cùng chính phủ. Các nước đang kiểm soát được dịch thì cũng nên đề cao cảnh giác.

d. Sử dụng màu sắc để thể hiện dữ liệu

- ✓ Chọn màu đỏ cho những nhóm nước có Total_Cases cao để thể hiện tính chất nghiêm trọng của nhóm nước này.
- ✓ Chọn màu cam cho những nhóm nước có Total_Cases trung bình để thể hiện tính chất ít nghiêm trọng hơn các nhóm nước có Total_Cases cao, nhưng vẫn thể hiện được nhóm nước này có Total_Cases lớn.
- ✓ Chọn màu xanh cho những nhóm nước có Total_Cases thấp để thể hiện tính chất ít nghiêm trọng nhất trong 3 nhóm nước.

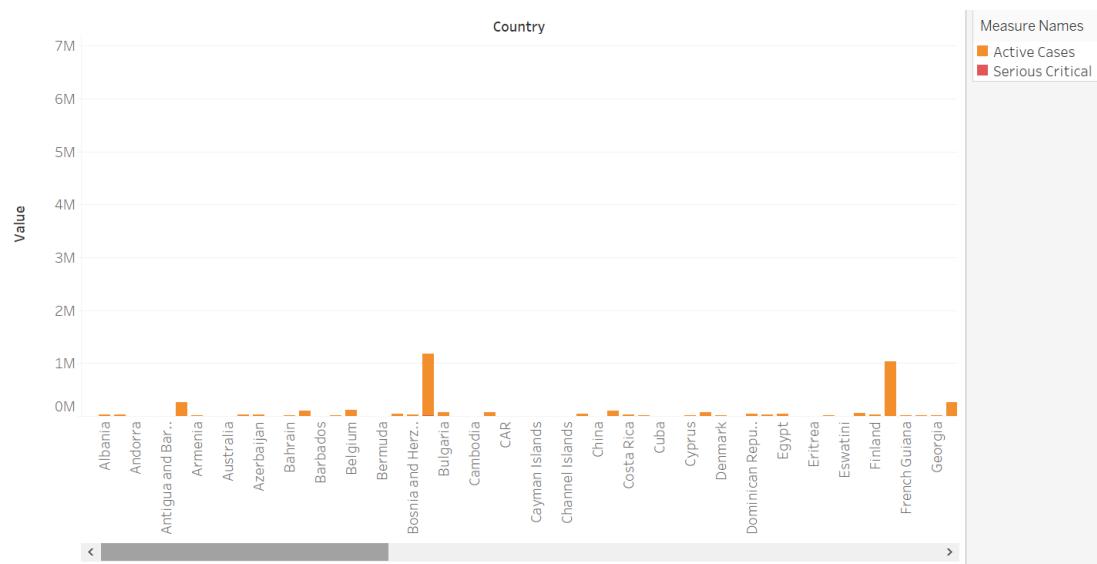
7. Country, Active Cases, Serious Critical

a. Lý do chọn các trường dữ liệu

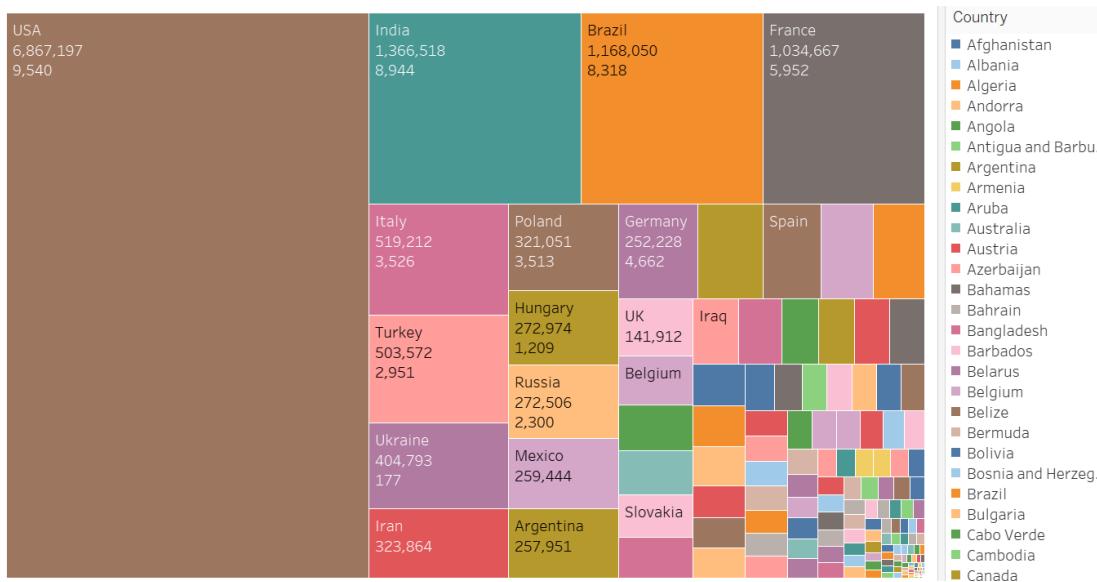
Xem xét được mức độ nghiêm trọng của bệnh dịch thông qua tỷ lệ ca nghiêm trọng trong tổng số ca đang trong quá trình điều trị của các quốc gia.

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 7.1. Tỷ lệ ca nghiêm trọng trong tổng số ca đang trong quá trình điều trị của các quốc gia ngày 13/4/2021



Hình 7.2. Tỷ lệ ca nghiêm trọng trong tổng số ca đang trong quá trình điều trị của các quốc gia ngày 13/4/2021

✓ Tính phù hợp của biểu đồ:

- Ở hình 7.1: So sánh được đồng thời dữ liệu tổng số ca đang trong quá trình điều trị và số ca nghiêm trọng thông qua việc đặt dữ liệu ca nghiêm trọng ở dưới cùng
- Ở hình 7.2: So sánh được đồng thời dữ liệu tổng số ca đang trong quá trình điều trị và số ca nghiêm trọng thông qua việc đặt dữ liệu ca nghiêm trọng ở dưới cùng. Tuy nhiên dữ liệu này độ chênh lệch giữa các miền khá là nhỏ nên hình 7.2 khó có thể phân biệt được.

➔ Tập trung nhận xét bằng trực quan của *hình 7.1*

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

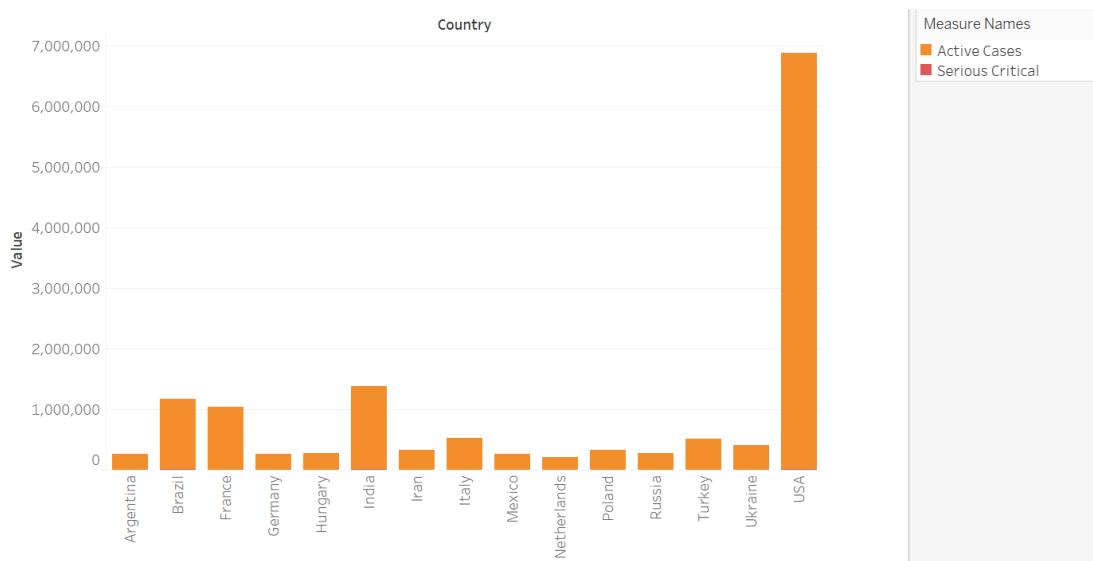
- Top các quốc gia có số lượng ca đang trong quá trình điều trị (**Active_Cases**) cao: lần lượt là USA, India, Brazil, France, Italy. Cao nhất là USA với 6867197 ca
- Top các quốc gia có tỷ lệ ca nghiêm trọng (**Serious_Critical**) cao nhất: lần lượt là USA, India, Brazil, France và Italy
- Các quốc gia không có ca nghiêm trọng/tỷ lệ ca nghiêm trọng thấp: Cayman Islands, Taiwan, Sao Tome and Principe (không có ca nghiêm trọng nào); Channel Islands, Sint Maarten (chỉ có 1 ca nghiêm trọng)
- Biểu đồ phỏng to 10000x cho thấy ca nghiêm trọng và ca đang trong quá trình điều trị của các quốc gia lần lượt là 9540% và 624290%

✓ Ý nghĩa: Thấy được mức độ nghiêm trọng của tình hình dịch bệnh thông qua số liệu ca nhiễm, ca nghiêm trọng và độ chênh lệch dữ liệu giữa các quốc gia từ đó nâng cao kiểm soát và đổi phó đối với tình hình dịch bệnh chưa có dấu hiệu thuyên giảm hiện nay.

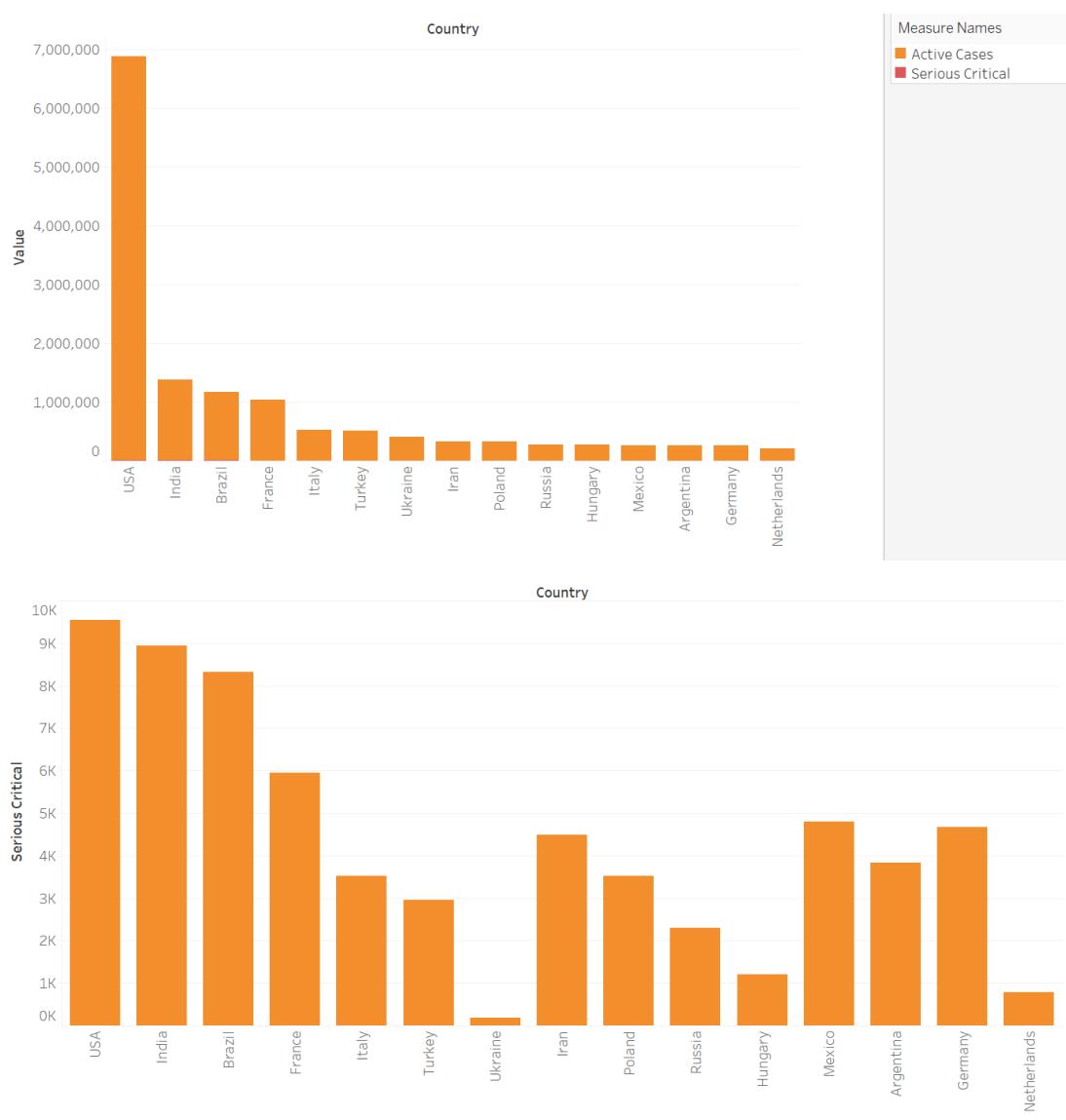
e. Sử dụng các kỹ thuật đã học

✓ Reduce: Active_Cases > 200 000

Vì dữ liệu có quá nhiều quốc gia dẫn đến các cột dữ liệu quá bé nên việc quan sát khá khó khăn và chúng ta cũng không cần phải quan sát quá nhiều quốc gia như vậy. Do đó, sử dụng filter để giảm số lượng quốc gia xuống, chỉ lấy top 15 quốc gia có số ca đang điều trị cao nhất.



✓ Manipulate view



(Biểu đồ này là do tách Serious_Critical của 15 nước trên ra)

Việc lựa chọn sort dữ liệu đã làm cho biểu đồ trực quan hơn, ta có thể dễ dàng so sánh số ca đang trong quá trình điều trị và tỷ lệ ca nghiêm trọng trong số ca đang điều trị của các quốc gia, dễ dàng thấy được quốc gia nào đang có tỷ lệ ca nghiêm trọng trong số ca đang điều trị cao nhất, thấp nhất.

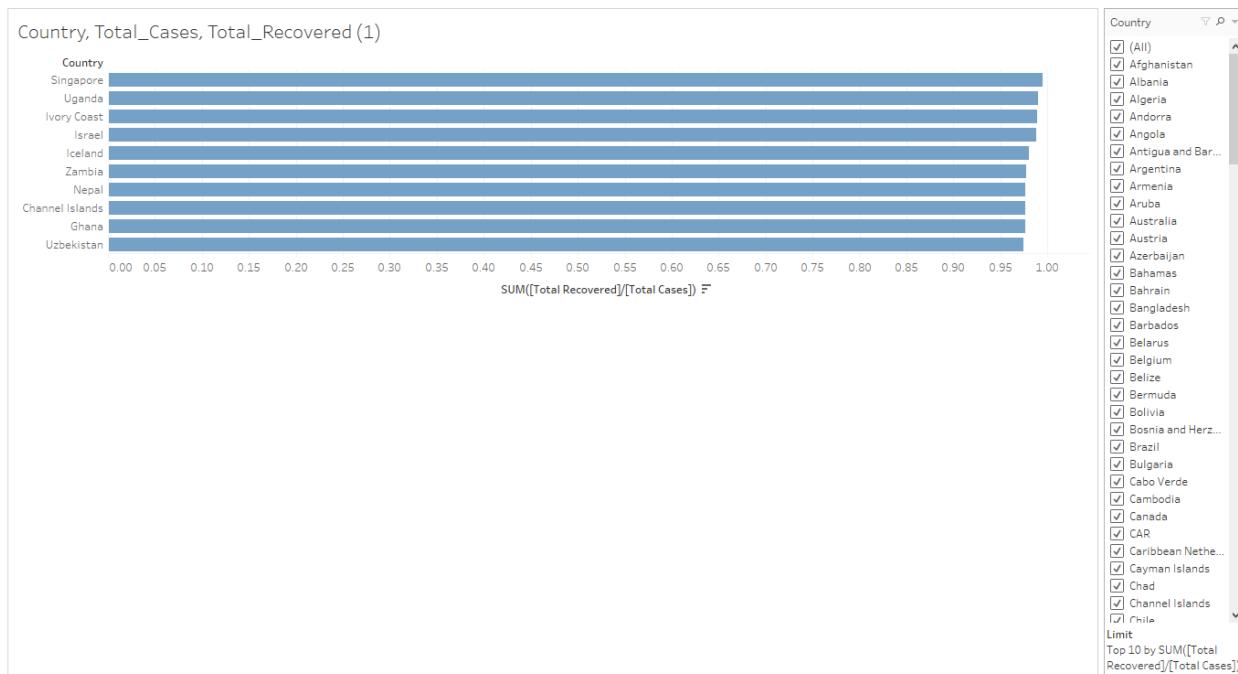
8. Country, Total Cases, Total Recovered

a. Lý do chọn các trường dữ liệu

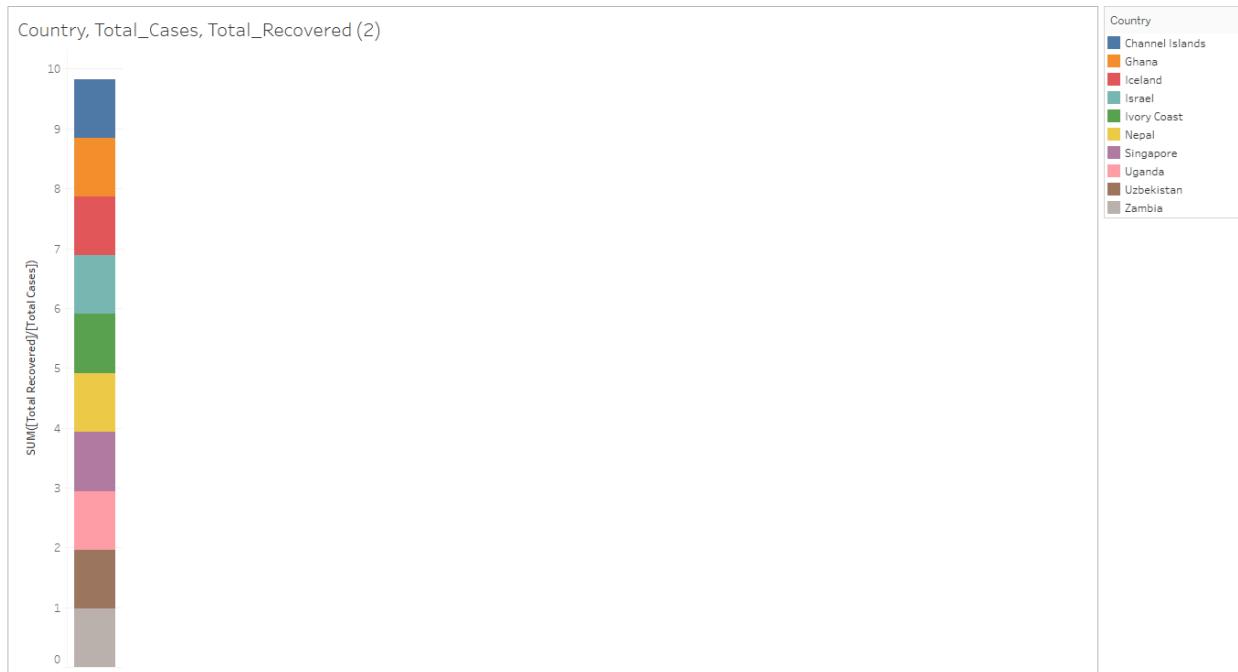
Quan sát được các quốc gia có tổng số ca hồi phục/tổng số ca nhiễm cao qua đó thấy được tình hình kiểm soát dịch bệnh và khả năng y tế của các quốc gia này

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 8.1. Biểu đồ thể hiện top 10 quốc gia có tổng số ca hồi phục/tổng số ca nhiễm cao nhất thế giới



Hình 8.2. Biểu đồ thể hiện top 10 quốc gia có tổng số ca hồi phục/tổng số ca nhiễm cao nhất thế giới

- ✓ Tính phù hợp của biểu đồ:
 - Biểu đồ *hình 8.1* trực quan được top 10 các quốc gia có tổng số ca hồi phục/tổng số ca nhiễm cao, biểu đồ dễ nhìn, thể hiện được rõ ràng số liệu của từng quốc gia.

- Biểu đồ *hình 8.2* trực quan được top 10 các quốc gia có tổng số ca hồi phục/tổng số ca nhiễm cao, biểu đồ dễ nhìn, thể hiện được rõ ràng số liệu của từng quốc gia, tuy nhiên vì dữ liệu chênh nhau rất nhỏ nên khi vẽ cột chồng như vậy nên khó thấy được sự khác nhau về số liệu của các nước.
➔ Tập trung nhận xét bằng trực quan của *hình 8.1*

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Các quốc gia có tổng số ca hồi phục/tổng số ca nhiễm chiếm tỉ lệ cao nhất là: Singapore, Uganda, Ivory Coast, Israel, Iceland, Zambia, Nepal, Channel Islands, Ghana, Uzbekistan
- Các quốc gia được thể hiện trong biểu đồ này đã kiểm soát dịch bệnh khá tốt, nên tiếp tục duy trì

✓ Ý nghĩa:

Từ biểu đồ ta thấy được các quốc gia có tổng số ca hồi phục/tổng số ca nhiễm chiếm tỉ lệ cao vì vậy chúng tôi được các quốc gia này đã kiểm soát dịch bệnh khá tốt, khả năng y tế cũng ở mức độ ổn, các quốc gia khác có thể học hỏi các biện pháp phòng chống từ các quốc gia này

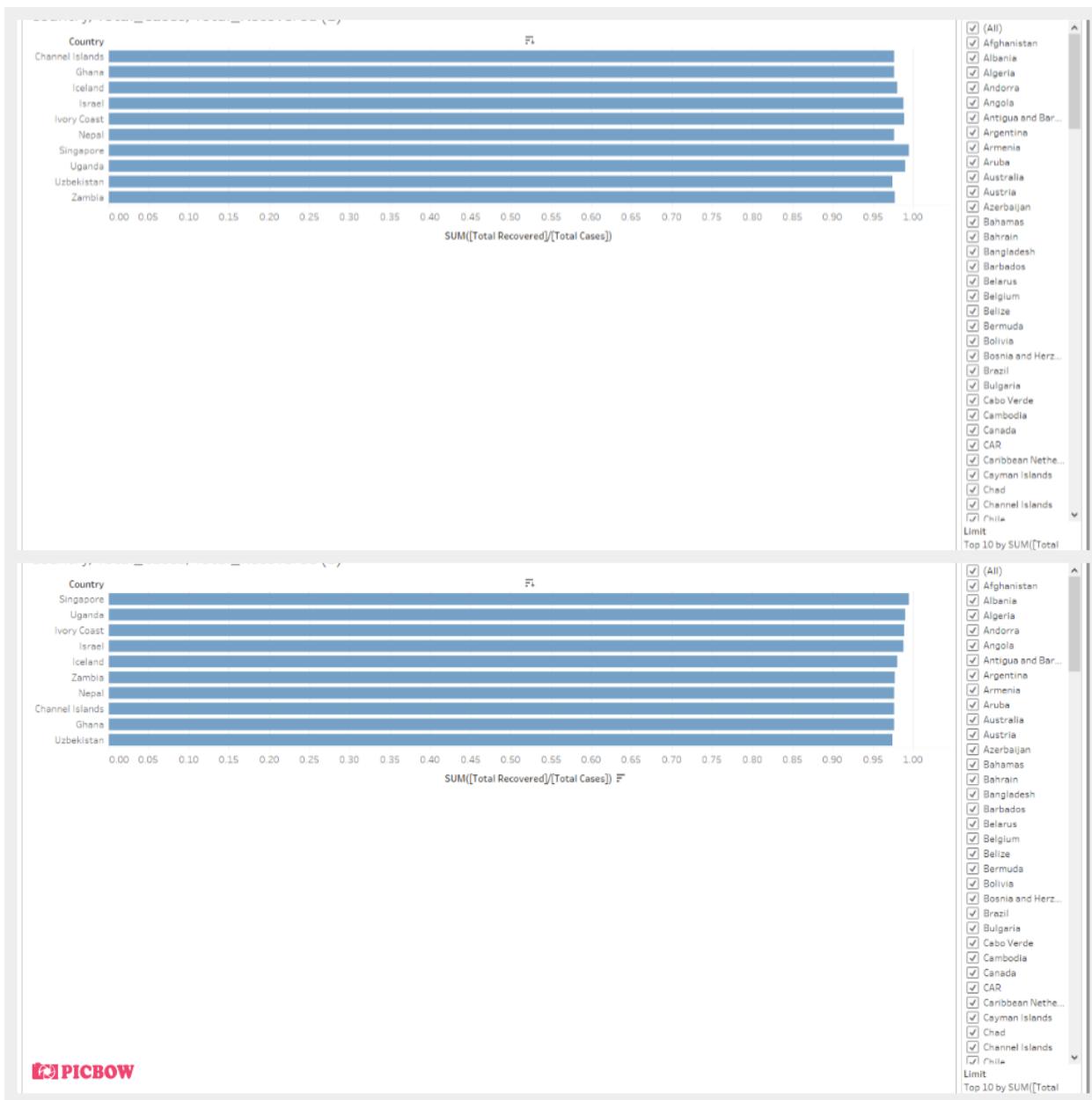
d. Sử dụng màu sắc để thể hiện dữ liệu

- ✓ Ở biểu đồ *hình 8.1*: Sử dụng màu xanh cho các cột để thể hiện đúng ý nghĩa mà biểu đồ muốn đề cập đó là tổng số ca hồi phục/tổng số ca nhiễm – một điều rất đáng mừng.

e. Sử dụng các kỹ thuật đã học

✓ Manipulate View:

Sort dữ liệu theo thuộc tính 'Total_Recovered/Total_Cases'



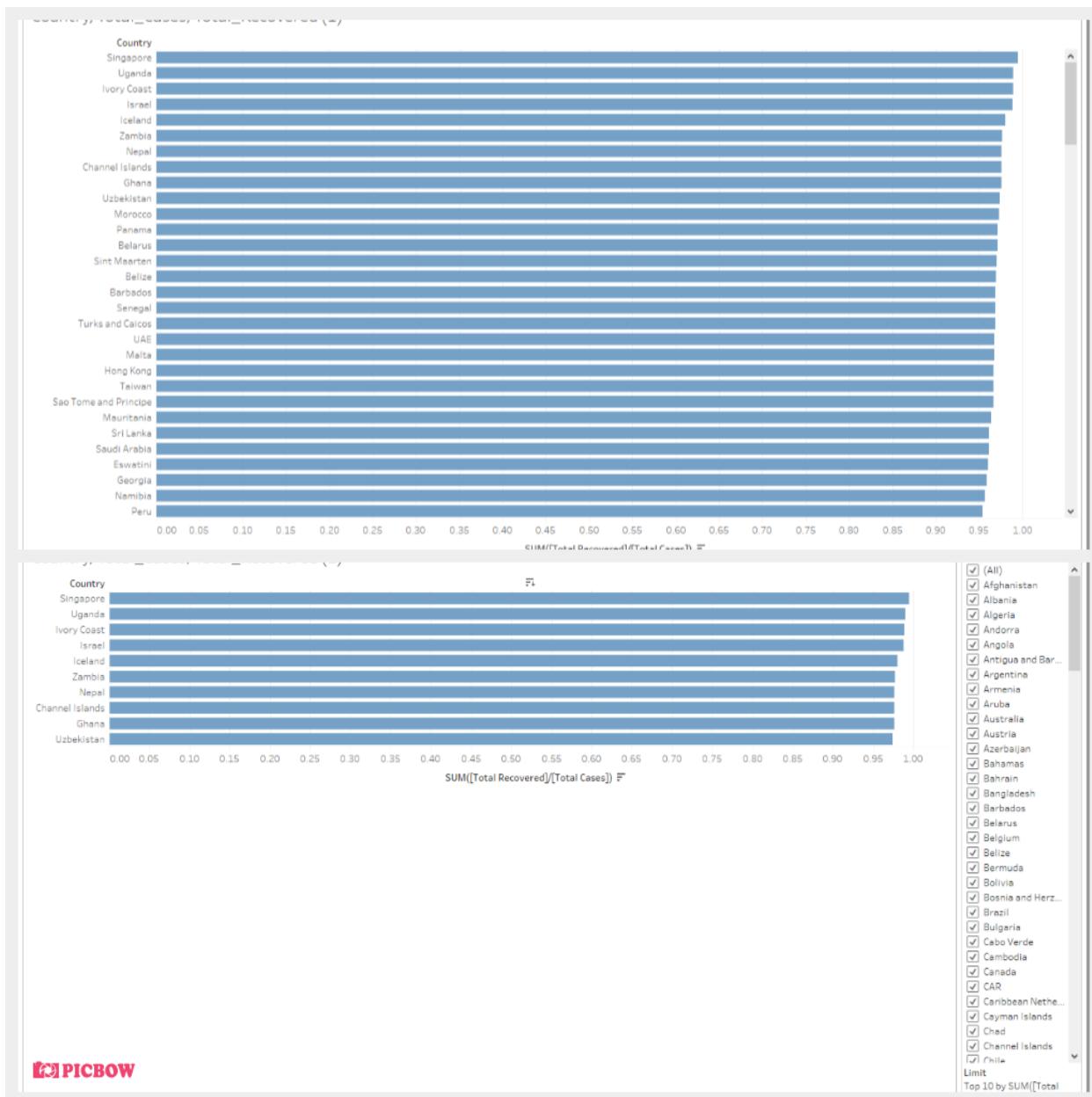
PICBOW

Trước và sau khi sort dữ liệu theo thuộc tính Total_Recovered/Total_Cases

Có thể thấy trước khi sort dữ liệu theo thuộc tính 'Total_Recovered/Total_Cases' biểu đồ rất lộn xộn, khó xem xét. Việc lựa chọn sort dữ liệu đã giúp cho biểu đồ trực quan hơn, dễ dàng thấy được quốc gia có số lượng ca tử vong mới/số lượng ca còn nhiễm trong ngày cao nhất, thấp nhất, dễ so sánh số lượng ca tử vong mới/số lượng ca còn nhiễm trong ngày của các quốc gia.

✓ Reduce:

Vì dữ liệu có quá nhiều quốc gia nên nếu để yên như vậy trực quan thì việc quan sát hơi bất cập và ta cũng không cần phải quan sát nhiều quốc gia như vậy. Do đó, sử dụng filter để giảm số lượng quốc gia xuống, chỉ lấy top 10.



PICBOW

Biểu đồ trước và sau khi sử dụng filter để giảm số quốc gia

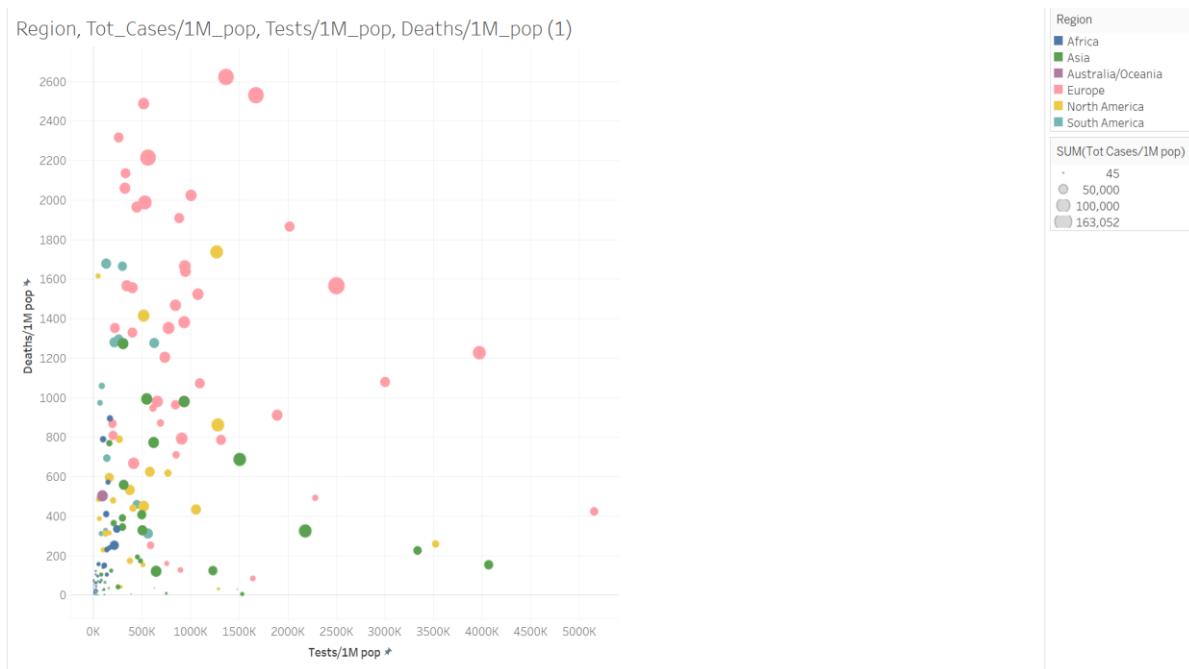
9. Region, Tot Cases/1M pop, Tests/1M pop, Deaths/1M pop

a. Lý do chọn các trường dữ liệu

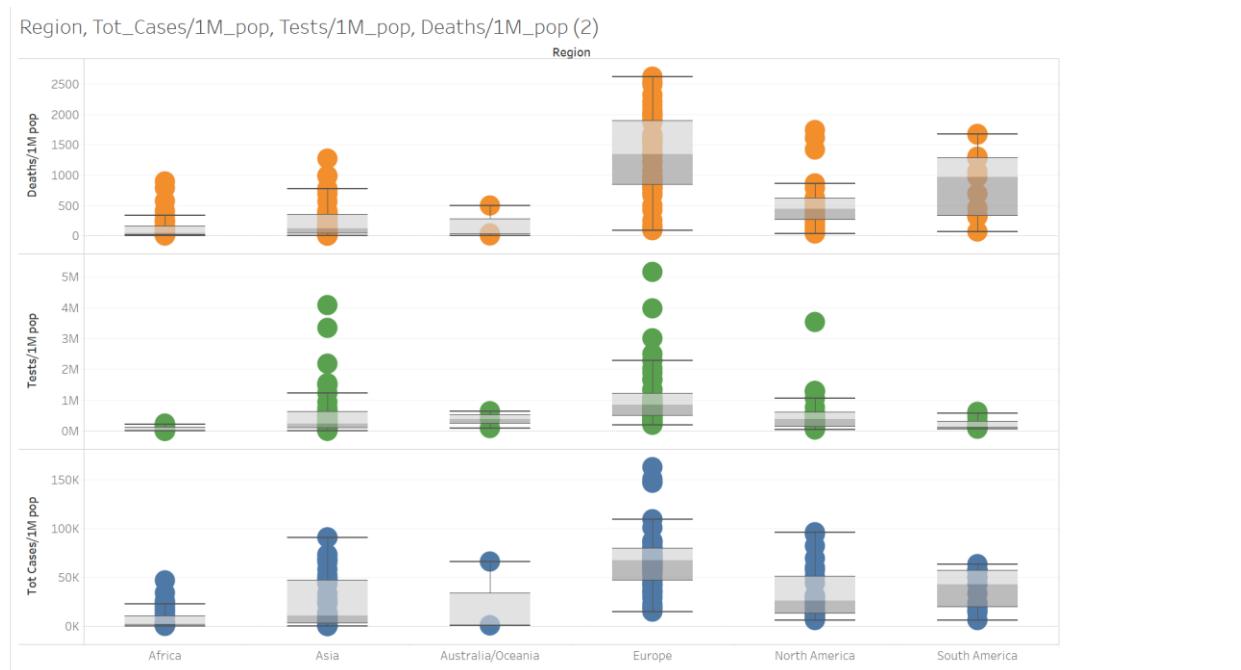
Nhận xét được tương quan giữa số xét nghiệm/triệu người, số lượng người chết/triệu người và tổng số ca/triệu người.

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 9.1. Biểu đồ phân phối điểm thể hiện tổng số ca/triệu người, số xét nghiệm/triệu người và số ca tử vong/triệu người của các quốc gia thuộc các khu vực trên thế giới trong ngày 13-4-2021



Hình 9.2. Biểu đồ phân phối điểm và vùng trung thể hiện tổng số ca/triệu người, số xét nghiệm/triệu người và số ca tử vong/triệu người của các quốc gia thuộc các khu vực trên thế giới trong ngày 13-4-2021

Ngoài cột x thể hiện dữ liệu Test/1M_pop và cột y thể hiện dữ liệu Deaths/1M_pop, ta có độ lớn của hình tròn thể hiện dữ liệu Tot_Cases/1M_pop.

✓ Tính phù hợp:

- Biểu đồ phân phối điểm (*Hình 9.1*) cho được cái nhìn tổng quan về mối tương quan giữa 3 dữ liệu là Tot_Cases/1M_pop, Deaths/1M_pop và Test/1M_pop. Thấy được độ tập trung của mối tương quan của các khu vực, dễ đưa ra đánh giá chung, so sánh và nhận thấy được các quốc gia có số liệu khác biệt rõ rệt.
- Biểu đồ phân phối và vùng tập trung (*Hình 9.2*) cho được cái nhìn tổng quan về sự tập trung của các dữ liệu (Tot_Cases/1M_pop, Deaths/1M_pop, Test/1M_pop) của từng khu vực, dễ dàng so sánh các dữ liệu của các khu vực thuộc cùng một trường. Tuy nhiên khó nhận thấy được tương quan của các dữ liệu và khó so sánh được vì chúng được vẽ trên tỷ lệ khác nhau tùy thuộc vào mỗi trường.

➔ Tập trung nhận xét bằng trực quan của *Hình 9.1*

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Khu vực có tổng ca/triệu người (Tot_Cases/1M_pop) cao nhất tập trung vào các quốc gia thuộc khu vực Europe, sau đó là Asia và North America
- Khu vực có số ca chết/triệu người (Deaths/1M_pop) cao nhất tập trung vào các quốc gia thuộc khu vực Europe, sau đó là North America và South America
- Khu vực có số xét nghiệm/triệu người (Test/1M_pop) cao nhất tập trung vào các quốc gia thuộc khu vực Europe, sau đó là Asia và North America
- Các quốc gia có số lượng người chết/triệu người (Deaths/1M_pop) cao (>1500 ca) tập trung ở phần các quốc gia có số xét nghiệm/triệu (Test/1M_pop) người thấp
- Các quốc gia có số lượng xét nghiệm/triệu người (Test/1M_pop) (>3 triệu ca) cao đều có số người chết/triệu người (Deaths/1M_pop) thuộc mức thấp/trung bình

✓ Ý nghĩa:

Từ các nhận xét và hình dáng đi xuống của biểu đồ, dễ thấy mối quan hệ giữa số xét nghiệm/triệu người và số lượng người chết/triệu người là tương quan nghịch. Từ đó thấy được tầm quan trọng của việc đầu tư vào xét nghiệm để phát hiện sớm và đưa vào điều trị, hạn chế tình trạng chữa trị muộn hoặc không phát hiện mắc bệnh dẫn đến tử vong.

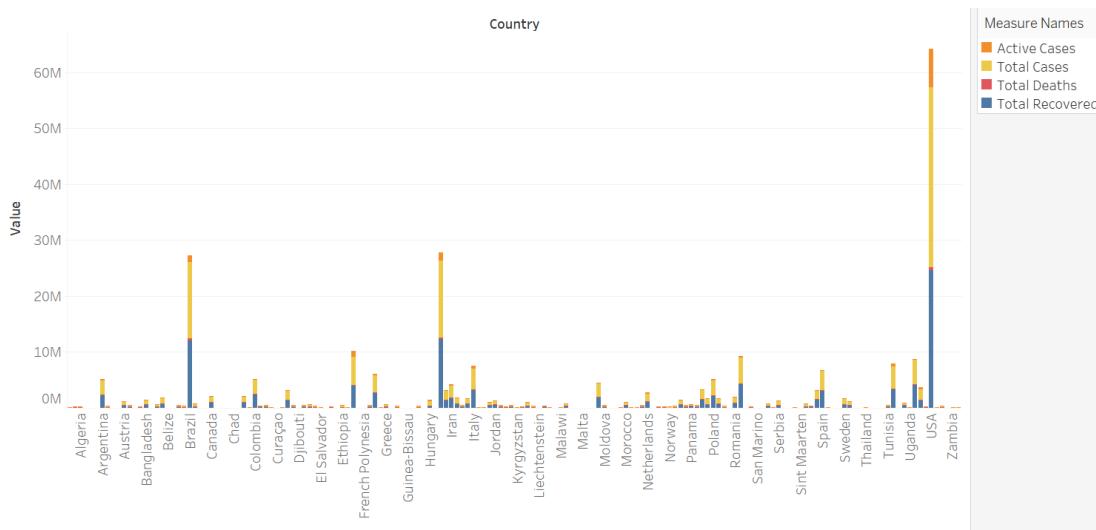
10. Country, Total Cases, Total Deaths, Total Recovered, Active Cases

a. Lý do chọn các trường dữ liệu

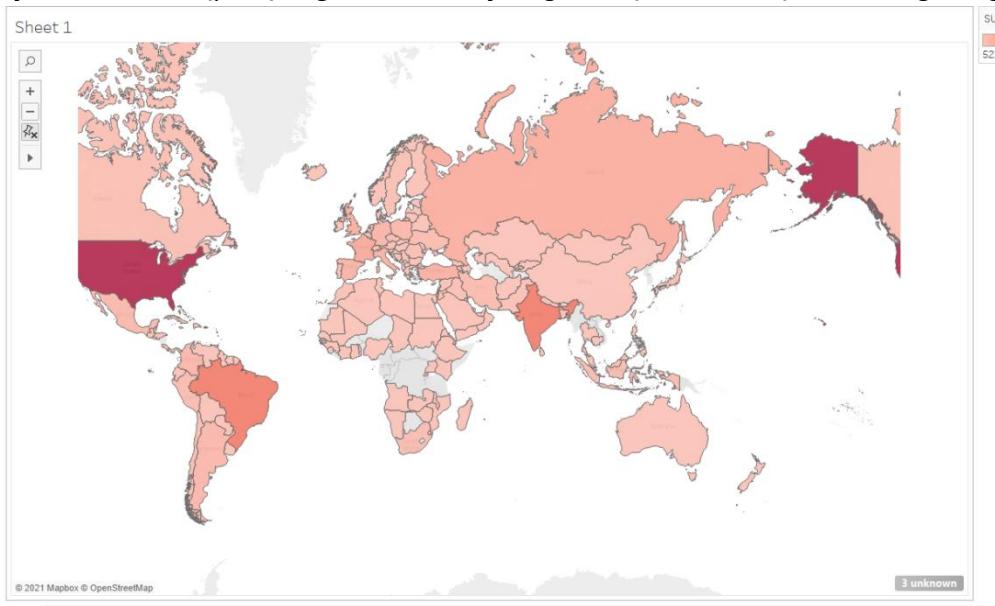
Xem xét được khả năng điều trị của các quốc gia thông qua tỷ lệ của số ca chết, đã được điều trị và đang trong quá trình điều trị trong tổng số ca.

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

✓ Trực quan dữ liệu:



Hình 10.1. Tổng số ca chết/ triệu người, tổng số ca đã được điều trị/ triệu người, tổng số ca đang trong quá trình điều trị/ triệu người của các quốc gia thuộc các khu vực trên thế giới ngày 13/4/2021



Hình 10.2. Tổng số ca chết/ triệu người, tổng số ca đã được điều trị/ triệu người, tổng số ca đang trong quá trình điều trị/ triệu người của các quốc gia thuộc các khu vực trên thế giới ngày 13/4/2021

✓ Tính phù hợp của biểu đồ:

- Biểu đồ **hình 10.1**: Cho được cái nhìn tổng quan đầy đủ về các trường dữ liệu cần thể hiện và các tỷ lệ tổng số ca chết/ tổng số ca, tổng số ca đang được điều trị/ tổng số ca và tổng số ca đã được điều trị/ tổng số ca. Đồng thời, dễ dàng nhận thấy được tiến triển dịch của mỗi khu vực và đưa ra được so sánh giữa chúng.
- Biểu đồ **hình 10.2**: Cho được cái nhìn trực quan về các quốc gia cũng như về tổng số ca của từng quốc gia trên các khu vực trên thế giới, đồng thời dễ dàng quan sát, so sánh chúng với các quốc gia và khu vực lân cận. Tuy nhiên, vì ở

đây cần phải thể hiện nhiều trường dữ liệu khác nhau nên biểu đồ này lại không thể cho cái nhìn tổng quan đầy đủ về các trường dữ liệu cần thể hiện.

➔ Tập trung nhận xét bằng trực quan của *Hình 10.1*

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Top các quốc gia có tỷ lệ ca tử vong/tổng số ca cao: lần lượt là Brazil (~2,64%), Antigua and Barbuda (~2.58%), Russia (~2.22%), France (~1.95%), USA (~1.8%).
- Top các quốc gia có tỷ lệ ca được điều trị thành công/tổng số ca cao: lần lượt là Taiwan (~96,7%), Cayman Island (~95.22%), Caribbean Netherlands (~92.79%), Russia (~91.93%).
- Top các quốc gia có tỷ lệ ca đang trong quá trình điều trị/tổng số ca cao: lần lượt là Timor-Leste (~49.17%), USA (~21.41%), France (~20.26%), Antigua and Barbuda (~18.57%).

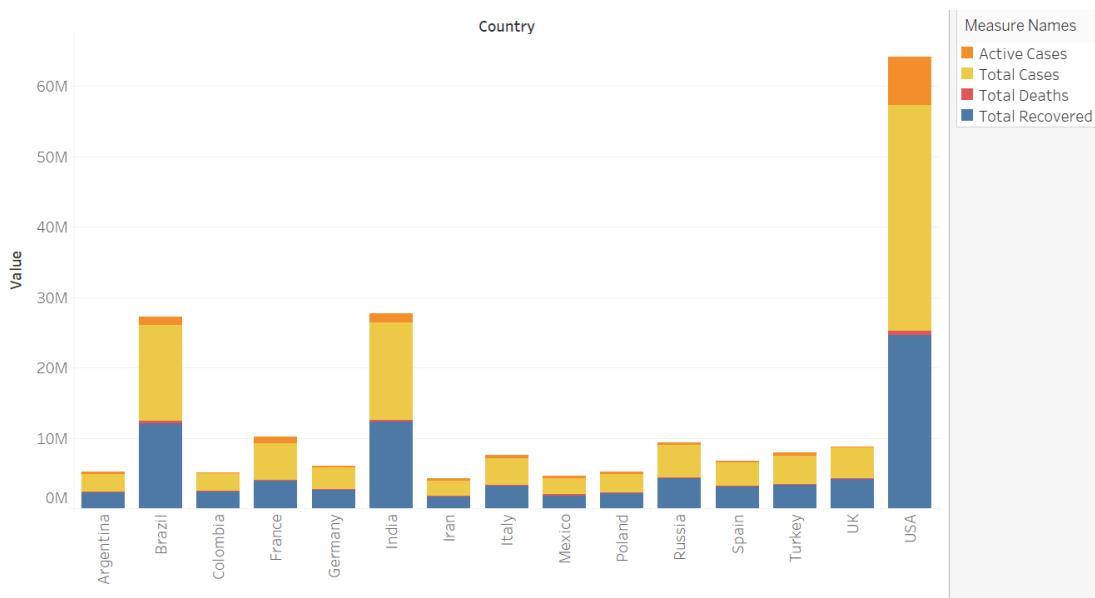
✓ Ý nghĩa: Thấy được sự tương quan thuận giữa các thuộc tính Total_Recovered và Active_Cases từ đó suy ra tăng cường biện pháp phòng ngừa và phát hiện sớm sẽ tăng số lượng người được phục hồi.

d. Sử dụng màu sắc để thể hiện dữ liệu

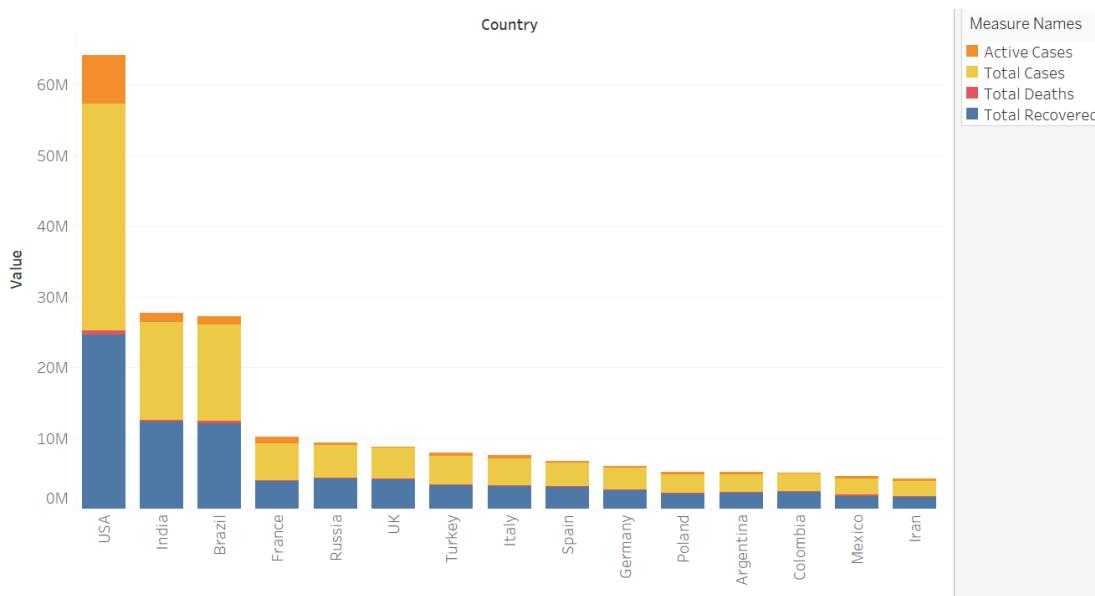
- ✓ Ở *hình 10.1*, chọn màu đỏ cho thuộc tính Total_Deaths, màu cam cho thuộc tính Active_Cases, màu vàng cho thuộc tính Total_Cases, màu xanh nước biển cho thuộc tính Total_Recovered vì để thể hiện ý nghĩa của các thuộc tính giảm dần theo mức nguy hiểm: từ màu đỏ (Total_Deaths) đến màu cam (Active_Cases) và màu xanh nước biển (Total_Recovered).
- ✓ Ở *hình 10.2*, chọn dải màu đỏ để thể hiện dữ liệu vì thuộc tính Total_Cases là tổng số ca nhiễm của một quốc gia, nó là một điều đáng báo động. Và màu đỏ sẽ nhạt dần theo tổng số ca nhiễm, điều này sẽ giúp ta dễ dàng nhận ra được mức độ nghiêm trọng của dịch bệnh ở các quốc gia.

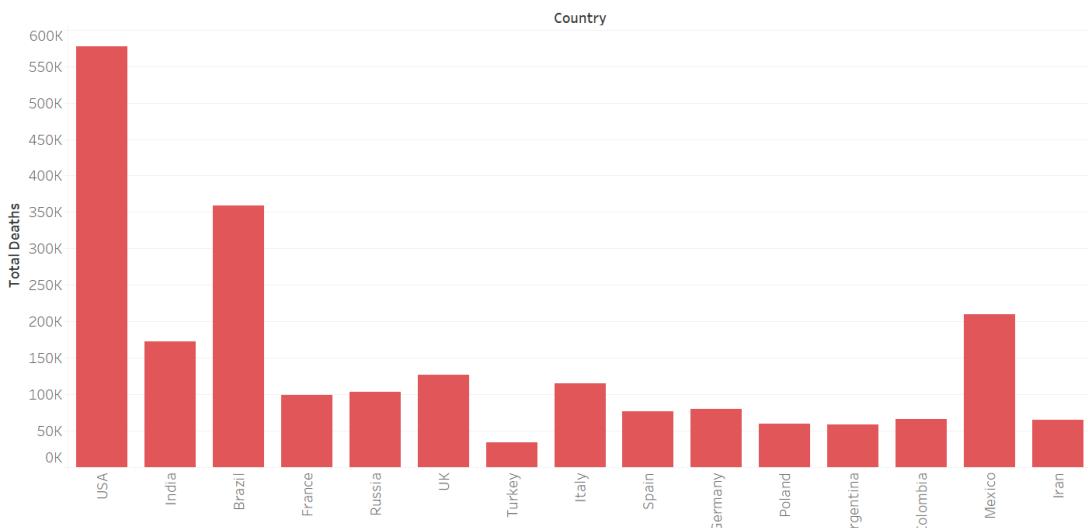
e. Sử dụng các kỹ thuật đã học

- ✓ Reduce: Total_Cases > 2 000 000
Vì dữ liệu có quá nhiều quốc gia dẫn đến các cột dữ liệu quá bé nên việc quan sát khá khó khăn và chúng ta cũng không cần phải quan sát quá nhiều quốc gia như vậy. Do đó, sử dụng filter để giảm số lượng quốc gia xuống, chỉ lấy top 15 quốc gia có tổng số ca nhiễm cao nhất.



✓ Manipulate view:





(Biểu đồ này là do tách Total_Deaths của 15 nước trên ra)

Việc lựa chọn sort dữ liệu đã làm cho biểu đồ trực quan hơn, ta có thể dễ dàng so sánh tổng số ca chết, tổng số ca đã được điều trị và tổng số ca đang trong quá trình điều trị của các quốc gia, dễ dàng thấy được quốc gia nào đang có tổng số ca chết, tổng số ca đã được điều trị và tổng số ca đang điều trị cao nhất, thấp nhất.

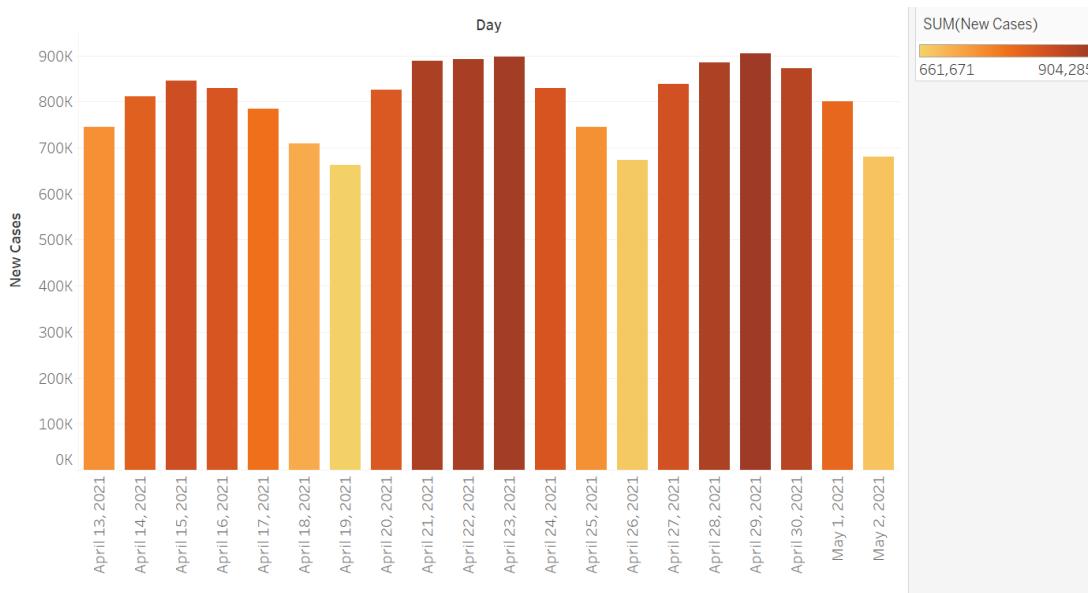
11. Day, New Cases

a. Lý do chọn các trường dữ liệu

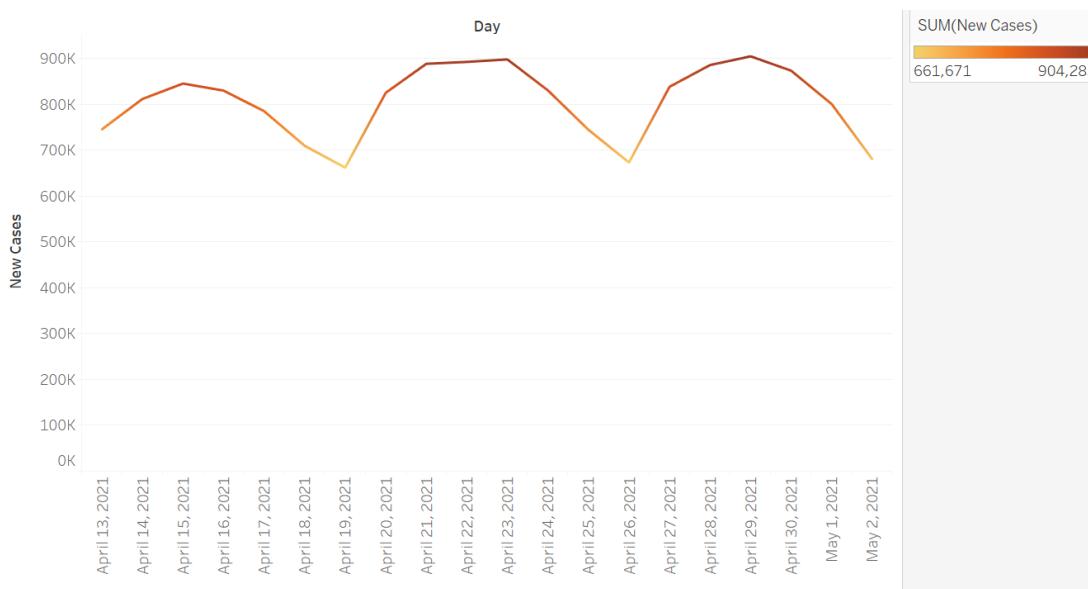
Xem xét được tình hình số ca nhiễm mới trên toàn thế giới qua thời gian (13/4/2021 - 2/5/2021).

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 11.1. Biểu đồ thể hiện số ca nhiễm mới qua thời gian (13/4/2021 - 2/5/2021)



Hình 11.2. Biểu đồ thể hiện số ca nhiễm mới qua thời gian (13/4/2021 - 2/5/2021)

- ✓ Tính phù hợp của biểu đồ:

Cả 2 biểu đồ *hình 11.1* và *hình 11.2* đều thể hiện được sự biến động của số ca nhiễm mới qua từng ngày trên toàn thế giới. Tuy nhiên, biểu đồ *hình 11.1* sẽ thể hiện được rõ ràng dữ liệu và dải phân bố của từng ngày hơn.

- ➔ Tập trung nhận xét bằng trực quan của *Hình 11.1*

c. Nhận xét và rút ra ý nghĩa

- ✓ Nhận xét:

- Từ ngày 13/4 đến 2/5, số ca nhiễm mới trên thế giới đã giảm từ 745 043 ca xuống còn 680 233 ca (giảm 64 810 ca).
- Trong khoảng thời gian này, số ca nhiễm mới tăng, giảm không liên tục: giai đoạn tăng mạnh nhất là 20/3 - 23/3 và 27/3 - 30/3, trong đó ngày 29/3 có số ca nhiễm mới cao nhất lên đến 904 285 ca. Ngày 18/3, 19/3 là 2 ngày có số ca nhiễm mới ít nhất, trong đó ngày 19/3 là ít nhất với 661 671 ca.

- ✓ Ý nghĩa:

Qua biểu đồ, ta có thể thấy tình hình kiểm soát dịch vẫn còn nhiều vấn đề vì cứ đang giảm thì lại tăng lên đột ngột, từ đó cho thấy tình hình dịch bệnh vẫn đang ở mức đáng báo động và cần có biện pháp kiểm soát cũng như tăng cường phòng chống dịch bệnh.

d. Sử dụng màu sắc để thể hiện dữ liệu

Chọn dải màu đỏ để thể hiện dữ liệu vì thuộc tính New_Cases là số ca nhiễm mới trong một ngày, nó là một điều đáng báo động. Và màu đỏ sẽ nhạt dần theo số ca nhiễm mới, điều này sẽ giúp ta dễ dàng nhận ra được mức độ nghiêm trọng của dịch bệnh trên toàn thế giới.

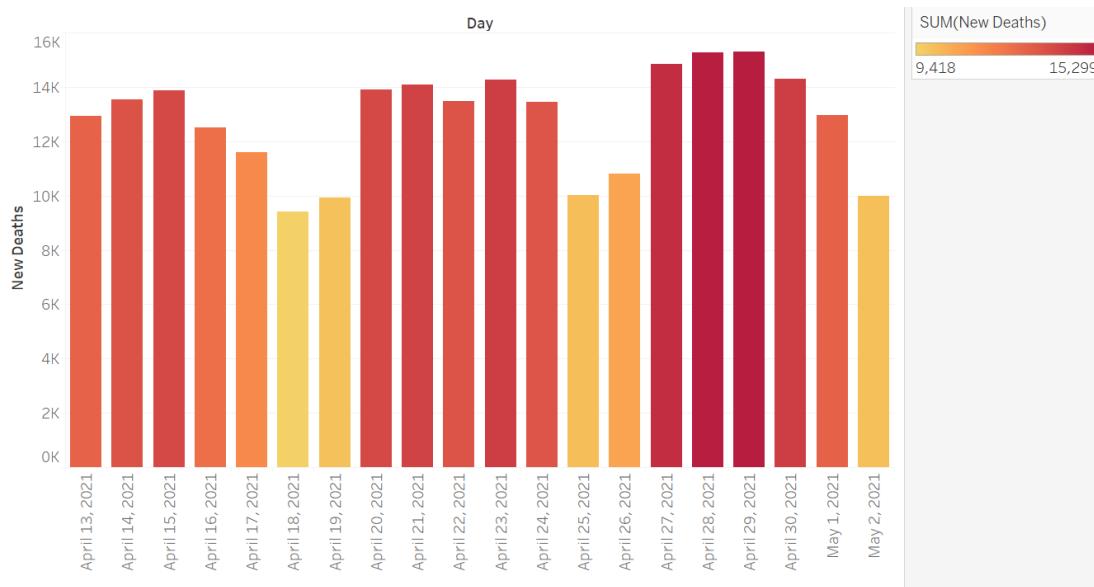
12. Day, New Deaths

a. Lý do chọn các trường dữ liệu

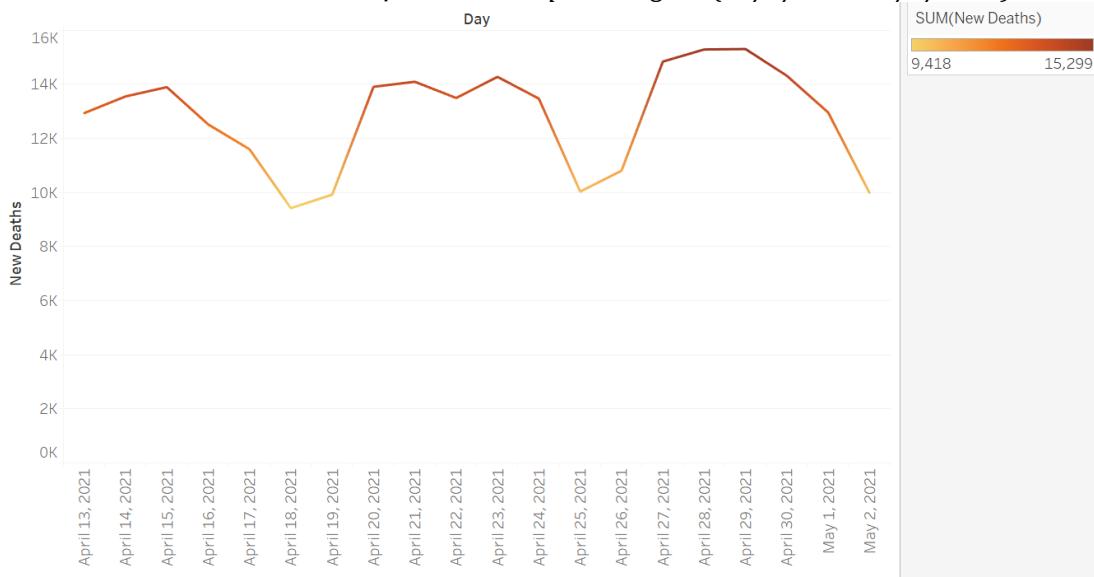
Xem xét được tình hình số ca chết mới trên toàn thế giới qua thời gian (13/4/2021 - 2/5/2021).

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 12.1. Biểu đồ thể hiện số ca chết qua thời gian (13/4/2021 - 2/5/2021)



Hình 12.2. Biểu đồ thể hiện số ca chết qua thời gian (13/4/2021 - 2/5/2021)

- ✓ Tính phù hợp của biểu đồ:

Cả 2 biểu đồ **hình 12.1** và **hình 12.2** đều thể hiện được sự biến động của số ca nhiễm mới qua từng ngày trên toàn thế giới. Tuy nhiên, biểu đồ **hình 12.1** sẽ thể hiện được rõ ràng dữ liệu và dài phân bố của từng ngày hơn.

- ➔ Tập trung nhận xét bằng trực quan của **Hình 12.1**

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Từ ngày 13/4 đến 2/5, số ca chết giảm từ 12 930 ca xuống còn 9 981 ca (giảm 2949 ca).
- Trong khoảng thời gian này, số ca chết cũng tăng, giảm không liên tục: giai đoạn có số ca chết cao nhất là 20/3 - 24/3 và 27/3 - 30/3, trong đó ngày 29/3 có số ca chết cao nhất lên đến 15 299 ca. Ngày 18 và 19/3 là 2 ngày có số ca chết ít nhất, trong đó 18/3 là ít nhất với 9 418 ca.

✓ Ý nghĩa:

Qua biểu đồ, số ca chết tuy có giảm nhưng không đáng kể, đa số chúng đều còn đang ở mức cao, đáng báo động. Vì vậy, cần tăng cường và nâng cao khả năng y tế để đối phó với dịch bệnh tốt hơn.

d. Sử dụng màu sắc để thể hiện dữ liệu

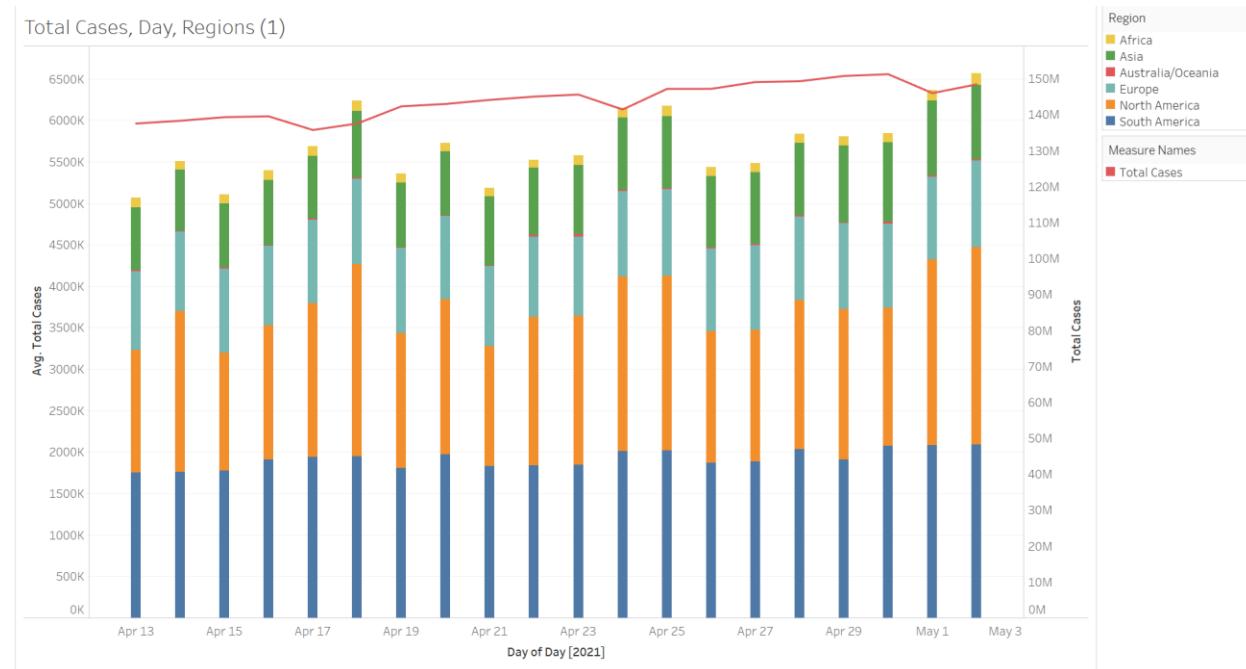
Chọn dải màu đỏ để thể hiện dữ liệu vì thuộc tính New_Deaths là số ca chết mới trong một ngày, nó là một điều đáng báo động. Và màu đỏ sẽ nhạt dần theo số ca nhiễm chết mới, điều này sẽ giúp ta dễ dàng nhận ra được mức độ nghiêm trọng của dịch bệnh trên toàn thế giới.

13. Total Cases, Day, Regions**a. Lý do chọn các trường dữ liệu**

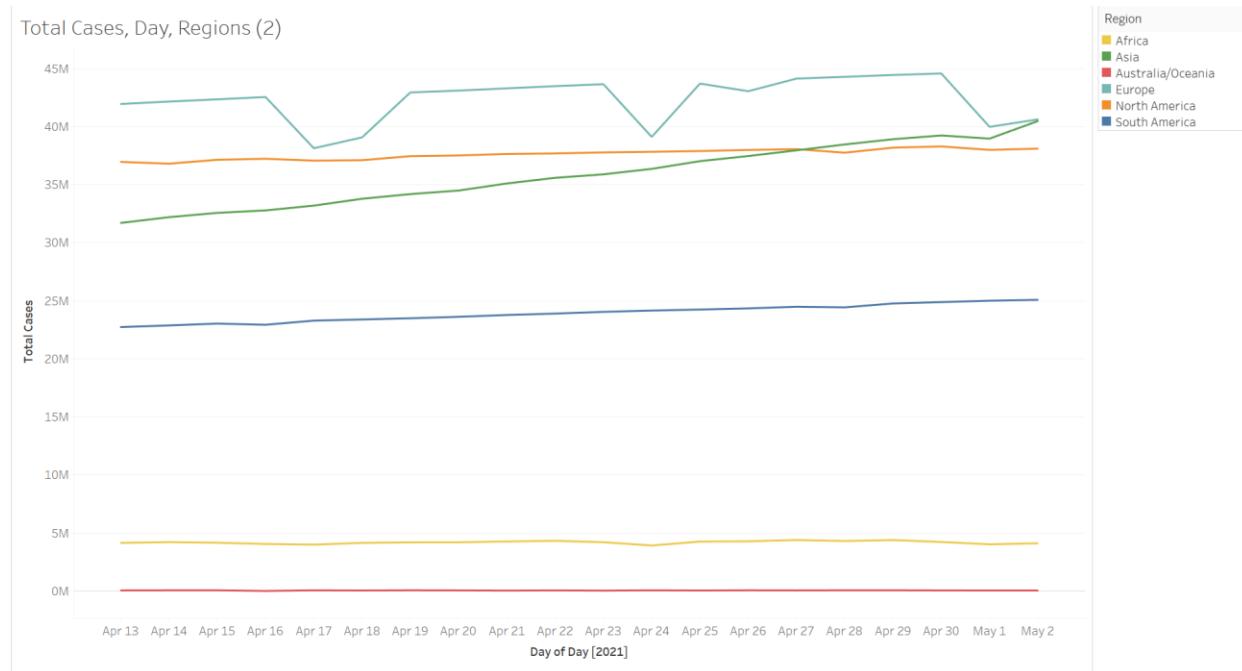
Nhận xét tổng số ca chia theo vùng thay đổi theo thời gian (13/4/2021 – 3/5/2021).

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

✓ Trực quan:



Hình 13.1. Biểu đồ cột chồng và đường trung bình thể hiện tổng số ca chia theo các vùng thay đổi theo thời gian (13/4/2021 – 3/5/2021)



Hình 13.2. Biểu đồ đường thể hiện tổng số ca chia theo các vùng thay đổi theo thời gian (13/4/2021 – 3/5/2021)

✓ Tính phù hợp:

- Biểu đồ cột chồng và biểu đồ đường (Hình 13.1) cho được cái nhìn tổng quan về tổng số ca và sự biến đổi theo thời gian thống kê của **toàn thế giới**, tỷ lệ tổng số ca của các khu vực trên toàn thế giới. Tuy nhiên vì bộ dữ liệu lớn nên khó nhận xét và so sánh giữa các khu vực có số liệu gần nhau như South America và North America, Europe và Asia.
 - Biểu đồ đường (Hình 13.2) cho được cái nhìn tổng quan về tổng số ca được ghi nhận và sự biến đổi theo thời gian thống kê của **từng khu vực**. Đồng thời dễ dàng nhận thấy được tiến triển dịch của mỗi khu vực và đưa ra được so sánh giữa chúng.
- ➔ Tập trung nhận xét bằng trực quan của Hình 13.2

c. Nhận xét và rút ra ý nghĩa

✓ Nhận xét:

- Khu vực có tổng ca (Total Cases) cao nhất tập trung vào các quốc gia thuộc khu vực Europe, North America và Asia

- Tổng số ca của các quốc gia thuộc các khu vực không có sự thay đổi lớn theo thời gian thống kê, khu vực có tổng số ca (Total Cases) có xu hướng tăng theo thời gian thống kê là North America và Asian.
 - Khu vực có sự thay đổi lớn và bất thường nhất là Europe. Sự thay đổi của tổng số ca là tương đối đột ngột ($>10M$ ca) và có sự lặp lại.
- ✓ Ý nghĩa:

Từ các nhận xét và sự lặp lại các thay đổi có thể thấy tình hình kiểm soát dịch còn nhiều vấn đề khi cứ khoảng 4 ngày lại có sự giảm và tăng đột ngột của tổng số ca, đặc biệt là các nước thuộc khu vực Europe khu vực có sự thay đổi lớn và bất thường nhất. Từ đó cần nâng cao kiểm soát dịch không những đặc biệt đối với các nước thuộc khu vực này mà còn các vùng lân cận, nhìn nhận vấn đề và rút kinh nghiệm đối với các nước.

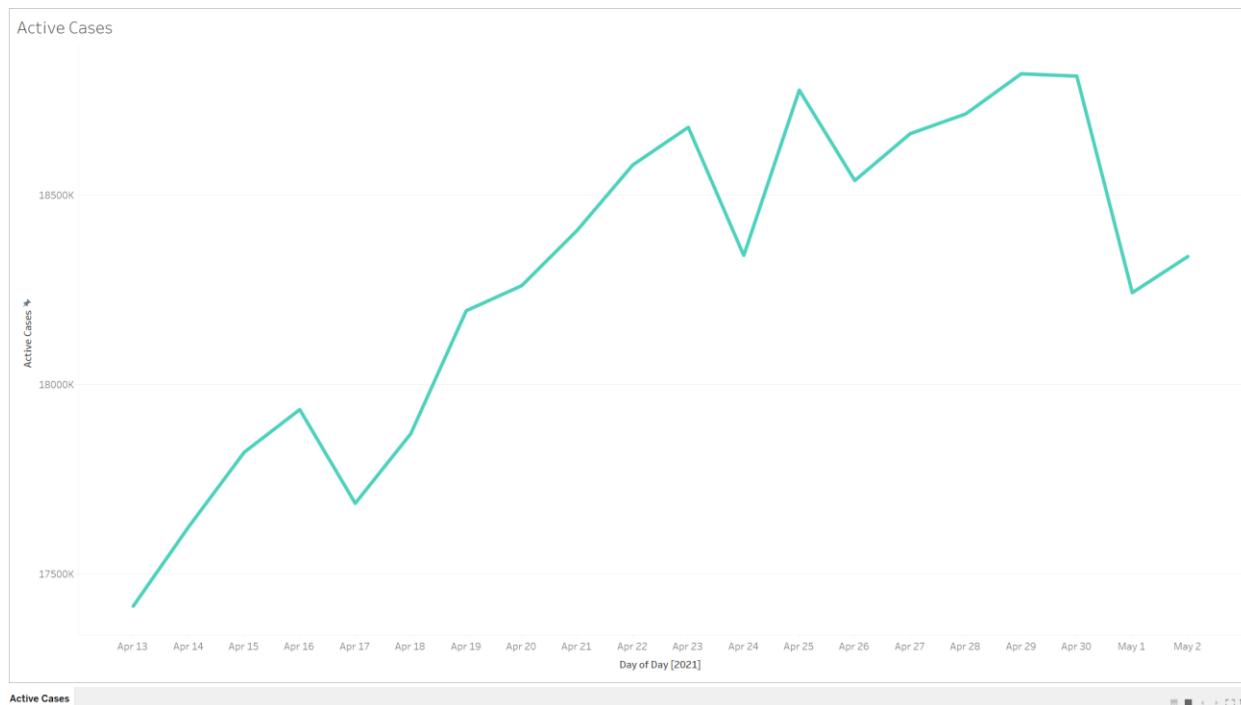
14. Active Cases, Day

a. Lý do chọn các trường dữ liệu

Quan sát được được tình hình, xu hướng dịch bệnh chung trên toàn cầu từ 13/04 – 02/05/2021.

b. Trực quan dữ liệu và giải thích tính phù hợp của biểu đồ đối với tính chất trường dữ liệu

- ✓ Trực quan dữ liệu:



Hình 14.1. Biểu đồ thể hiện sự thay đổi tình hình dịch bệnh toàn cầu thông qua active case qua mỗi ngày.

- ✓ Tính phù hợp của biểu đồ:
 - Biểu đồ *hình 14.1* trực quan rõ sự biến động từng ngày của toàn cầu, thấy được thực tại kiểm soát bệnh dịch mỗi ngày như thế nào thông qua số người đang bị nhiễm.

c. Nhận xét và rút ra ý nghĩa

- ✓ Nhận xét:
 - Số ca active bắt đầu tăng nhanh từ ngày 17/04 (độ dốc cao), giảm/tăng đột ngột (hay bất ổn) từ ngày 23 – 26/04, các ngày sau đó dần ổn định trở lại và bắt đầu có xu hướng giảm.
- ✓ Ý nghĩa:
 - Tình hình dịch bệnh toàn cầu từng ngày.

d. Sử dụng màu sắc để thể hiện dữ liệu

- ✓ Ở biểu đồ *hình 14.1*: Sử dụng màu xanh lục kết hợp line để thể hiện đúng ý nghĩa mà biểu đồ muốn đề cập đó là sự thay đổi, biến động mỗi ngày trên toàn cầu qua tổng số người đang nhiễm bệnh.

e. Sử dụng các kỹ thuật đã học

- ✓ Manipulate View
- ✓ Reduce

Sử dụng kết hợp các kỹ thuật trên giúp viewer dễ tương tác và trực quan:

Người sử dụng dễ dàng xem thông tin tổng số ca active trên toàn cầu trong 1 ngày ngay trên view khi Hover chuột (Change view over time), Reduce thuộc tính/thông tin cần thiết cho người quan sát thấy (ngày, tổng số ca active trong ngày).

B.2. CHẠY THUẬT TOÁN HỌC MÁY

Sử dụng các thuật toán học máy để dự đoán số ca nhiễm mới **New_Cases** qua các yếu tố khác (được trình bày ở file Model.ipynb trong folder Code, trước khi chạy file này thầy copy tất cả file ở folder Data vào folder Code đã ạ).

1. Linear Regression

Sử dụng mô hình Linear Regression của thư viện Sklearn với các tham số mặc định.
Kết quả của mô hình trên tập Train và tập Validation:

```
Train score: 0.917399159577572
Validation score: 0.9364471181538294
```

2. SVM

Sử dụng mô hình SVM.SVR của thư viện Sklearn với các tham số $C = [0.2, 0.5, 0.8, 1.0]$ và $kernel = ['linear', 'poly', 'rbf', 'sigmoid']$ để tìm mô hình SVR tốt nhất.

Kết quả của mô hình SVR tốt nhất trên tập Validation:

0.8654004575021833

- ➔ Như vậy, mô hình **Linear Regression** cho kết quả trên tập Validation cao hơn vì vậy lấy mô hình này train cho toàn tập train + validation và đánh giá kết quả trên tập test. Kết quả trên tập test:

0.9148098882375496

III. TÀI LIỆU THAM KHẢO

<https://bsdinsight.com/phan-mem-tableau/>

<https://nordiccoder.com/blog/tableau-data-visualization-cho-nguo-i-moi-bat-dau/>

<https://www.gimasys.com/tin-tuc/10-tinh-nang-manh-me-cua-tableau-giup-doanh-nghiep-hoat-dong-vuot-troi>

<https://www.tableau.com/products/new-features/ask-data>