**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM**

Sinh viên thực hiện: 18120408 - Trần Ngọc Lan Khanh

18120401 – Mai Khánh Huyền

18120379 – Võ Thị Hiếu

18120378 – Trần Văn Hiếu

GV phụ trách: Lê Ngọc Thành

TRỰC QUAN HÓA DỮ LIỆU

HỌC KỲ Ii – NĂM HỌC 2020-2021

**Trực quan hóa dữ liệu**

**BẢNG THÔNG TIN CHI TIẾT NHÓM**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ tên** | **Email** |
| 18120408 | Trần Ngọc Lan Khanh | 18120408@student.hcmus.edu.vn |
| 18120401 | Mai Khánh Huyền | 18120401@student.hcmus.edu.vn |
| 18120379 | Võ Thị Hiếu | [18120379@student.hcmus.edu.vn](mailto:18120379@student.hcmus.edu.vn) |
| 18120378 | Trần Văn Hiếu | [18120378@student.hcmus.edu.vn](mailto:18120378@student.hcmus.edu.vn) |

**YÊU CẦU ĐỒ ÁN- BÀI TẬP**

# **Yêu cầu của Đồ án/Bài tập**

Sử dụng Tableau để trực quan hóa dữ liệu

# **Kết quả**

1. **Lý thuyết**
2. **Giới thiệu về Tableau**

Được thành lập vào năm 2003 là kết quả của một dự án khoa học máy tính tại Standford nhằm mục đích là một công cụ để phân tích dữ liệu, giúp cho người sử dụng có thể dễ dàng trực quan hóa dữ liệu, thấy được ý nghĩa của dữ liệu, khai thác dữ liệu một cách dễ dàng mà không cần các kỹ năng công nghệ thông tin nâng cao

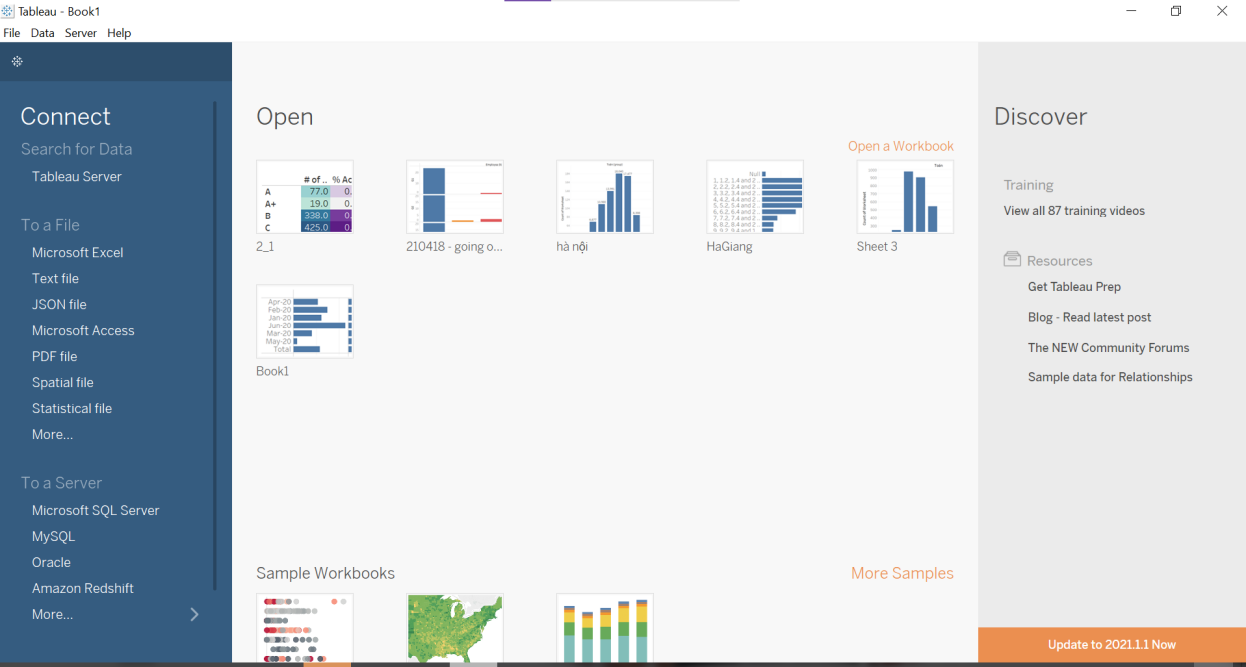
1. **Các tính năng hỗ trợ của Tableau**
2. Kết nối với dữ liệu

Tableau phếp kết nối với:

* Dữ liệu được lưu trữ dưới dạng file: Microsoft Excel, PDF, json, text,...
* Dữ liệu được lưu trữ như một server: Microsoft SQL Server, MySQL, Oracle,...

Có thể dễ dàng kết nối với các dữ liệu này thông qua giao diện đầu tiên của Tableau

Trong bài này, ta sẽ thực hiện với dataset Sample-Superstore có sẵn trong Tableau



Ảnh có chứa bàn

Mô tả được tạo tự động

1. Vẽ biểu đồ

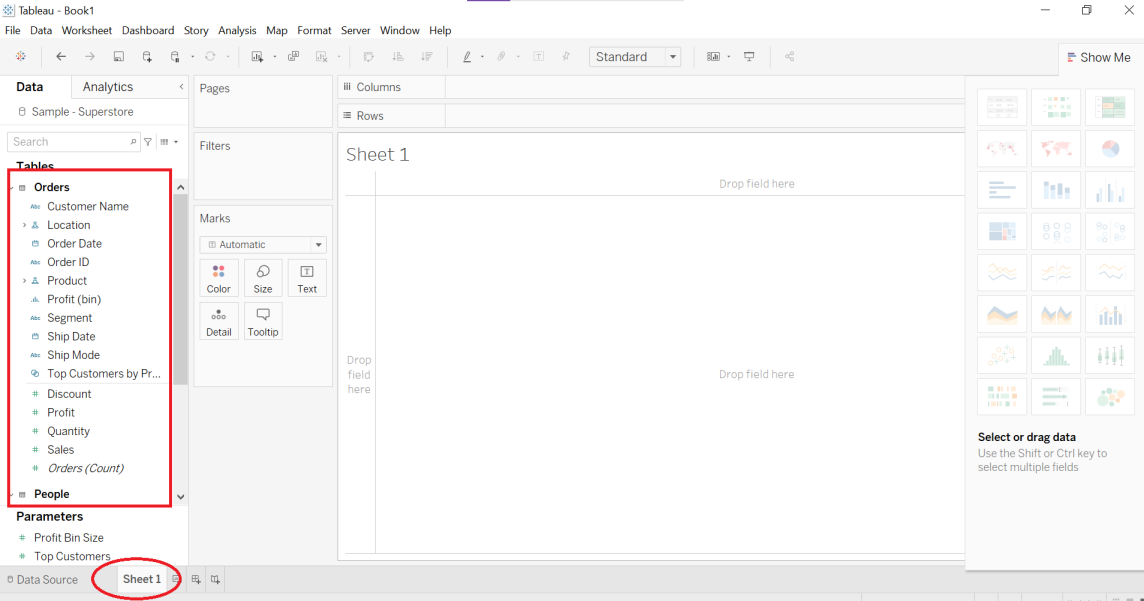
* Dimension và Measure
* Demension: dữ liệu định tính (tên, ngày). Theo mặc định, Tableau tự động phân loại dữ liệu chứa thông tin định tính hoặc phân loại dưới dạng thứ nguyên.

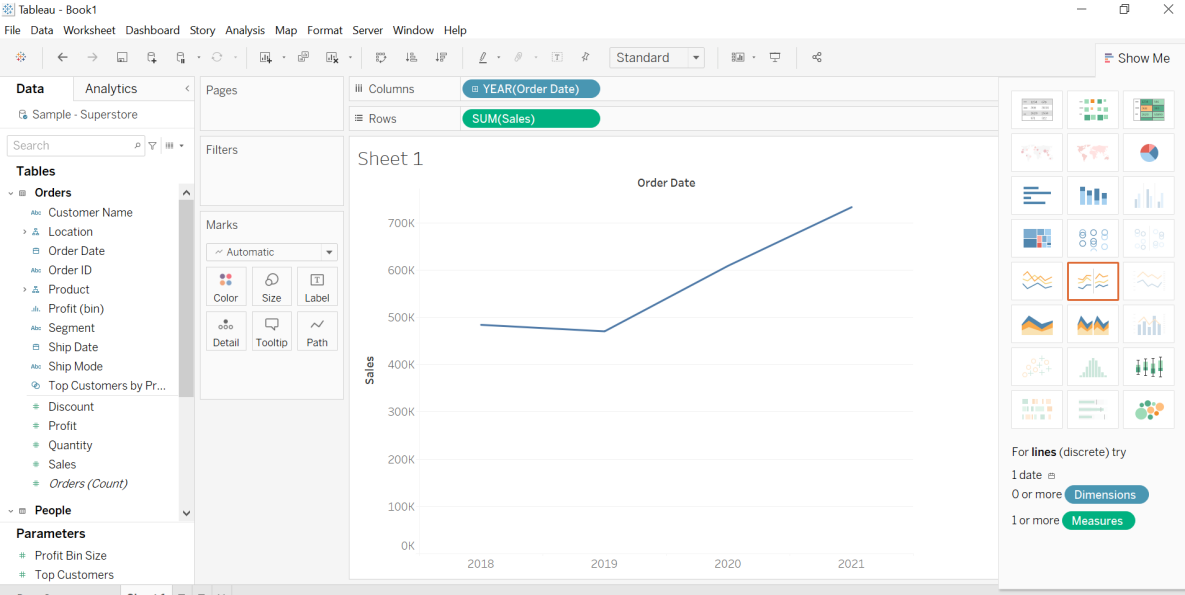
Ví dụ: Bất kì trường nào có chứa văn bản hoặc ngày. Các trường này thường xuất hiện dưới dạng tiêu đề cột cho các hàng dữ liệu như tên khách hàng hoặc ngày đặt hàng, và cũng xác định mức độ chi tiết hiển thị trong chế độ xem

* Measure: dữ liệu số định lượng. Theo mặc định, Tableau coi bất kỳ trường nào chứa loại dữ liệu này làm thước đo

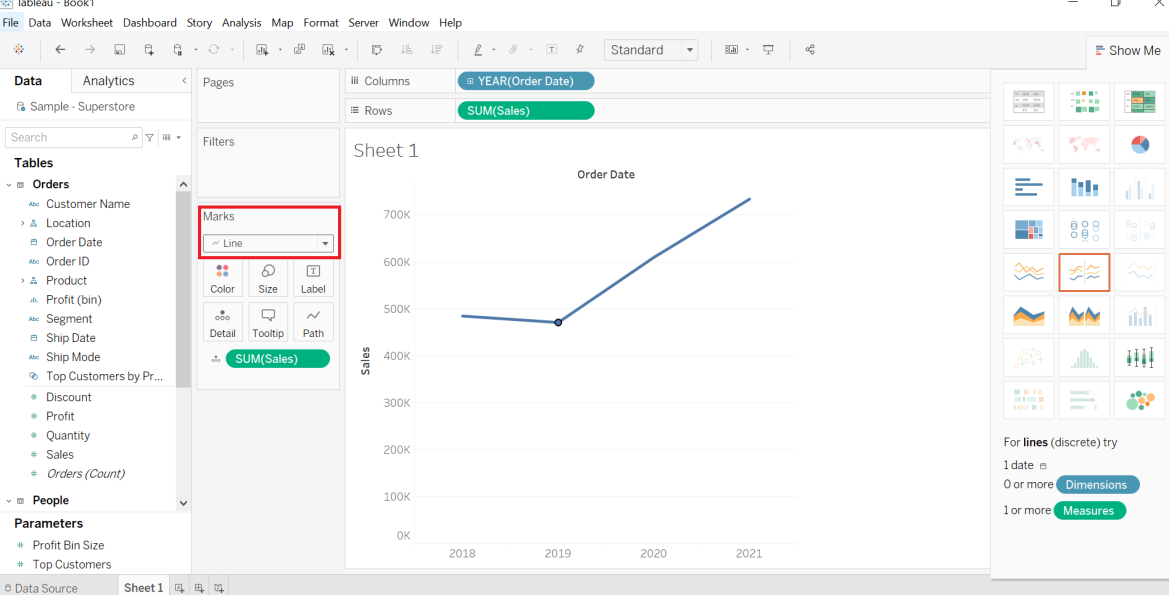
Ví dụ: giao dịch bán hàng hoặc lợi nhuận. Dữ liệu được phân loại là thước đo có thể được tổng hợp dựa trên một thứ nguyên nhất định, ví dụ: tổng doanh số bán hàng (Measure) theo khu vực (Dimension).

* Để vẽ biểu đồ, đi đến các sheet, kéo thả dữ liệu mong muốn và Rows shelf và Columns shelf. Ví dụ như kéo Order Date vào Columns Shelf và Sales vào Rows shelf, Tableau sẽ tự động chọn loại biểu đồ phù hợp với kiểu dữ liệu đã chọn

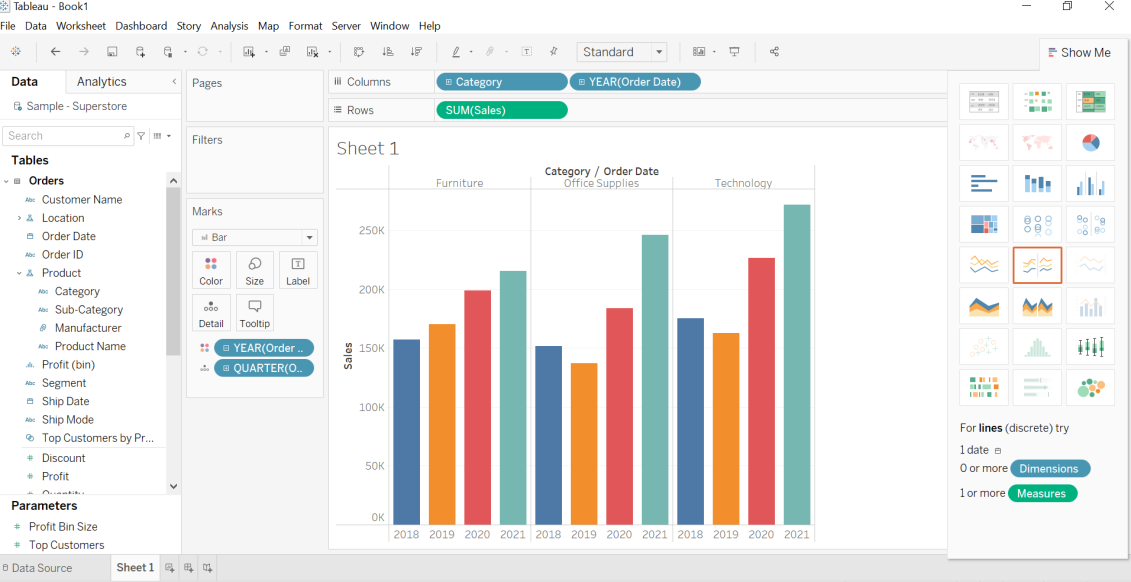




* Ngoài ta, người dùng cũng có thể chọn loại biểu đồ đễ biểu diễn dữ liệu. Tableau hỗ trợ nhiều loại biểu đồ khác nhau, được gọi là loại Mark

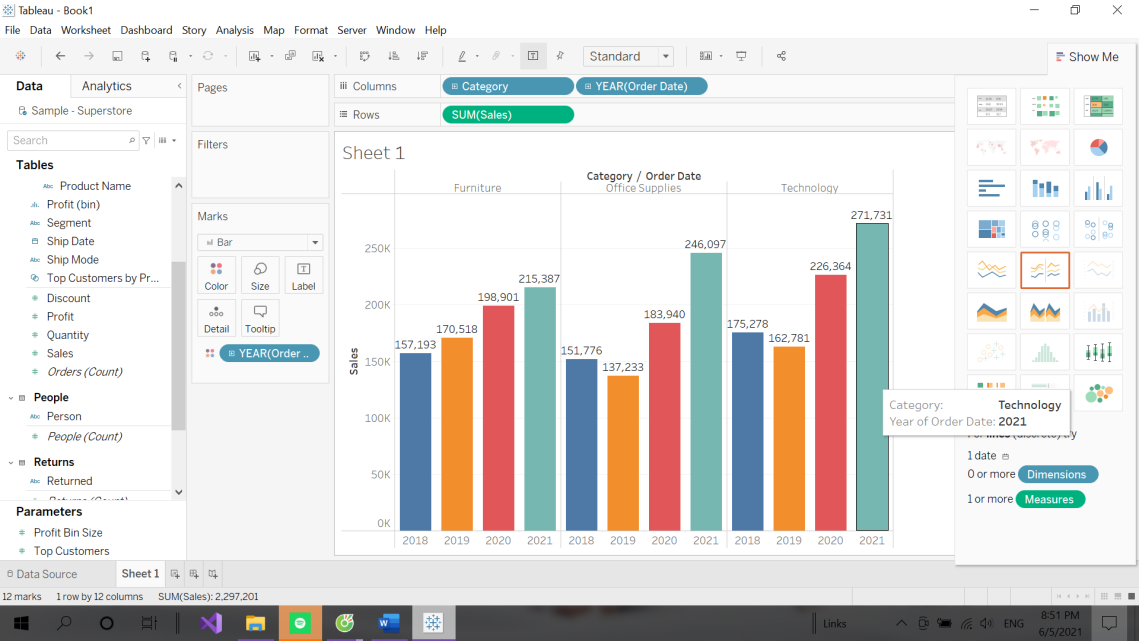


* Biểu đồ bên trên thể hiện sự phát triển của lượng đặt hàng, nhưng để phân tích sâu hơn, mặt hàng nào được đặt hàng nhiều nhất, ta có thể kết hợp thêm Category vào Columns shelf



Điều này tạo ra một bar chart và hiển thị cho bạn doanh số bán hàng tổng thể cho từng danh mục sản phẩm theo năm.

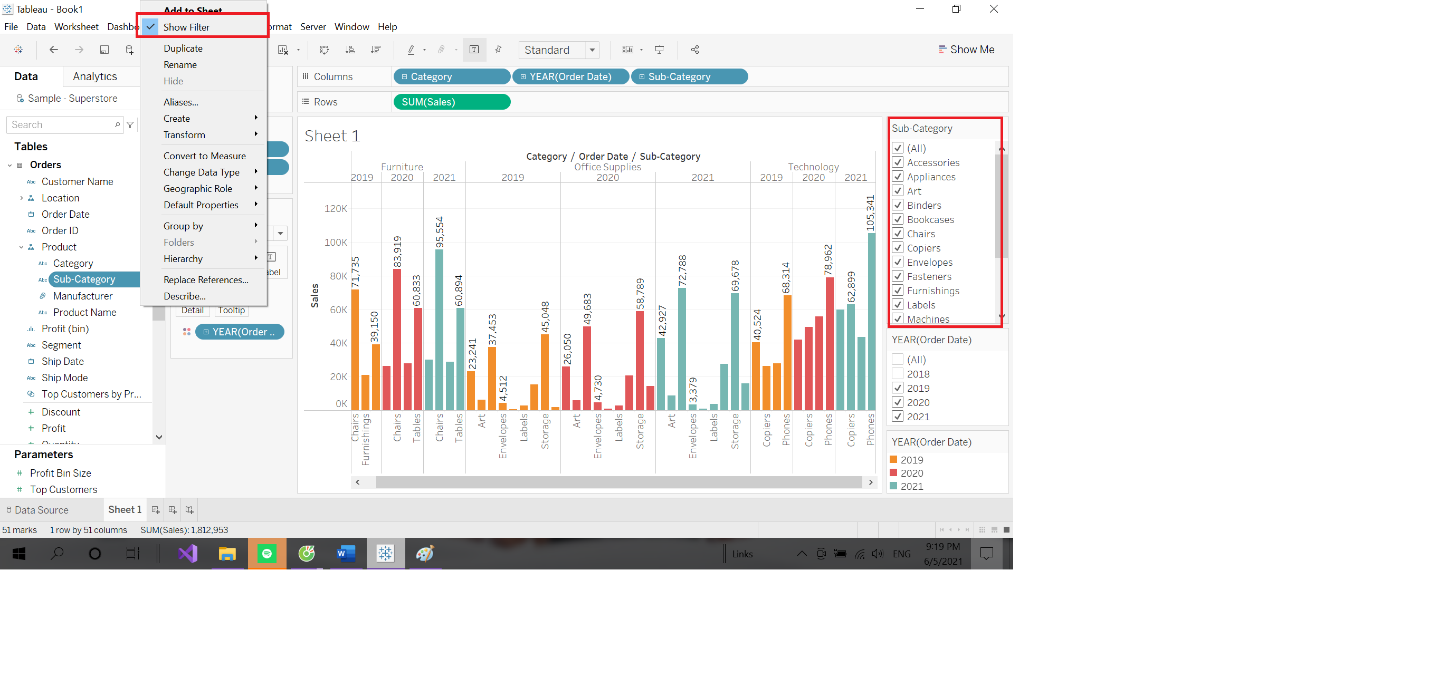
* Để xem thông tin về từng điểm dữ liệu, di chuột qua một trong các thanh để hiển thị tooltip
* Để thêm thông tin điểm dữ liệu dưới dạng nhãn, hãy nhấp vào Show Mark Labels trên toolbar



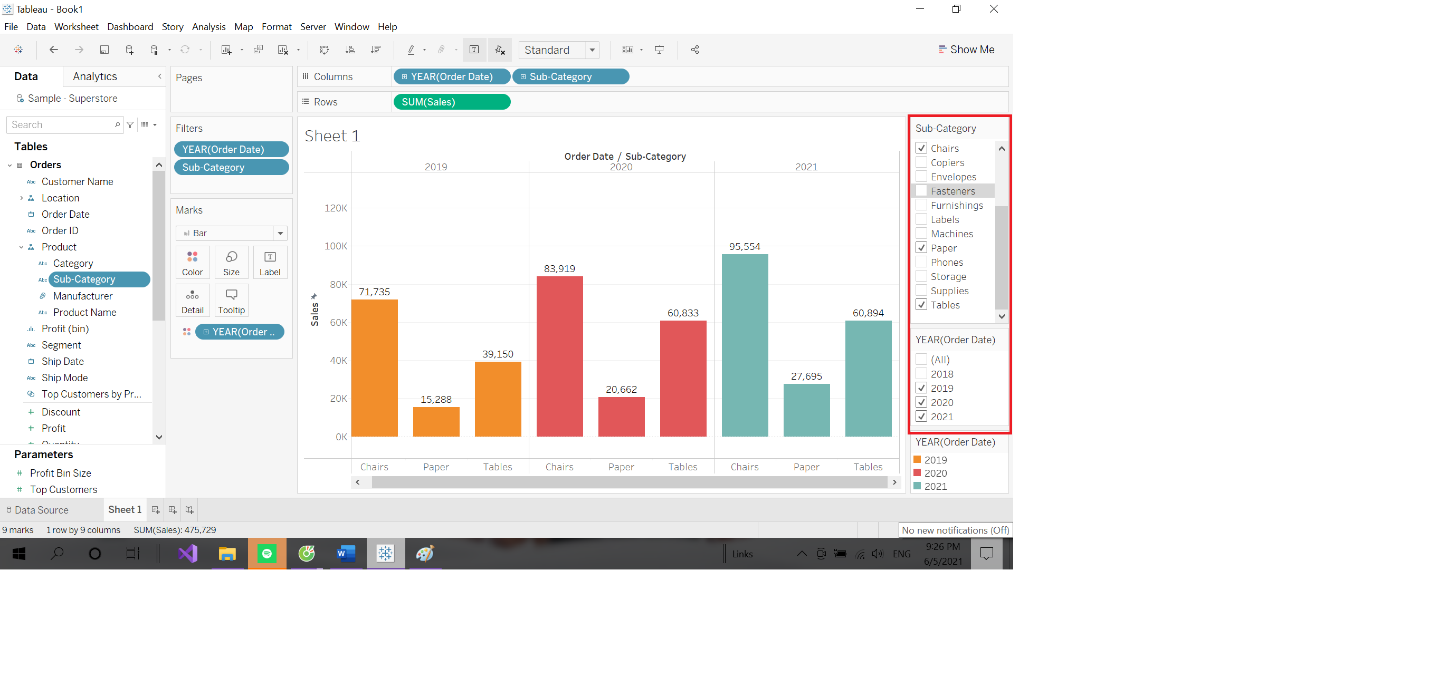
* Có thể đào sâu vào dữ liệu hơn như kết hợp thêm Sub-Category, Profit, Location…. Tùy vào mục đích phân tích dữ liệu nhắm đến.

1. Xử lý kết quả dữ liệu

* Khi trên biểu đồ có quá nhiều thể hiện, và ta chỉ cần tập trung vào một vài trong số đó. Tableau cho phép sử dụng bộ lọc và màu sắc để tập trung vào các chi tiết quan tâm đến.
* Để thực hiện, trong phần Data pane, nhấn chuột phải vào trường dữ liệu chọn  **Show Filter**
* Các bộ lọc được thêm vào phía bên phải của trang giao diện theo thứ tự đã chọn chúng. Chọn các dữ liệu mà ta muốn quan sát ở trong phần này.



VD: Chỉ quan tâm đến các mặt hàng Chairs, Paper, Table vào các năm 2019, 2020, 2021



* Để thêm màu sắc vào biểu đồ theo ý muốn, ta có thể kéo các trường dữ liệu vào phần Color của Marks card. Có thể thay đổi màu sắc theo ý muốn có thấy nhấp chuột trái vào phần Colors, và chọn Edit Colors.

VD: Sub-Category được phân loại theo màu sắc để dễ dàng quan sát sự khác biệt giữa các mặt hàng.



1. Khám phá dữ liệu về mặt địa lý

* Để trực quan hóa các dữ liệu dạng địa lý (lượng bán hàng theo các tỉnh, dân số của mỗi quốc gia), có thể sử dụng map views trong tableau.
* Các dữ liệu dạng Dimension địa lý, chỉ cần nhấp đúp chuột vào vào trường dữ liệu, map view sẽ xuất hiện.

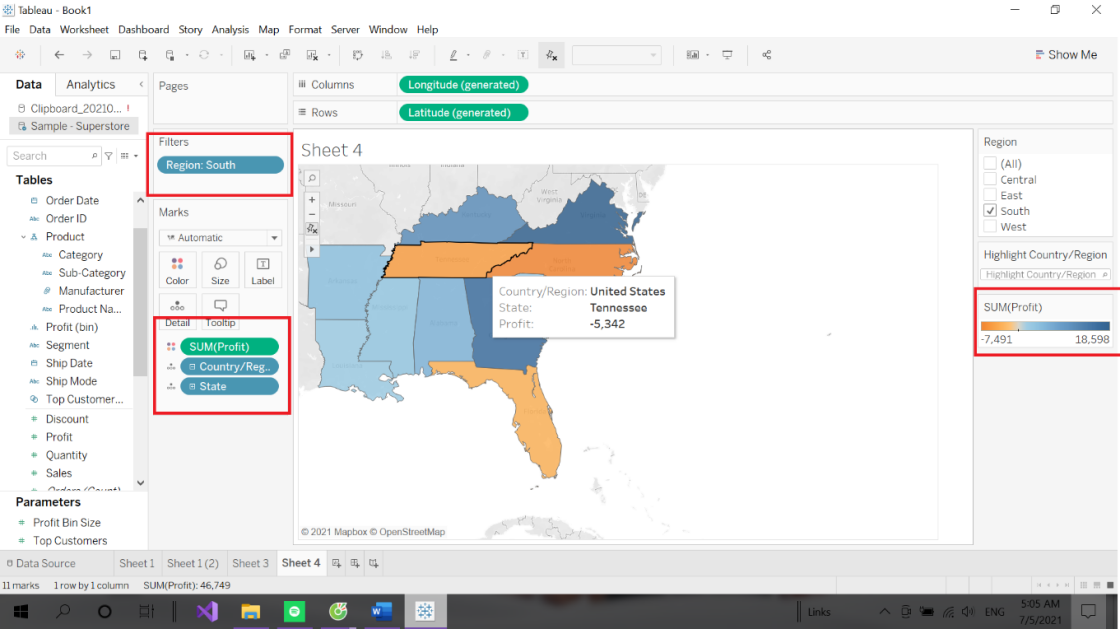
VD: Map view thể hiện các bang của Mỹ.

Ảnh có chứa bản đồ

Mô tả được tạo tự động

* Có thể sử dụng chức năng Filter để lọc ra các vùng cần thiết so sánh, sử dụng màu sắc để thể các mối quan hệ về số lượng.

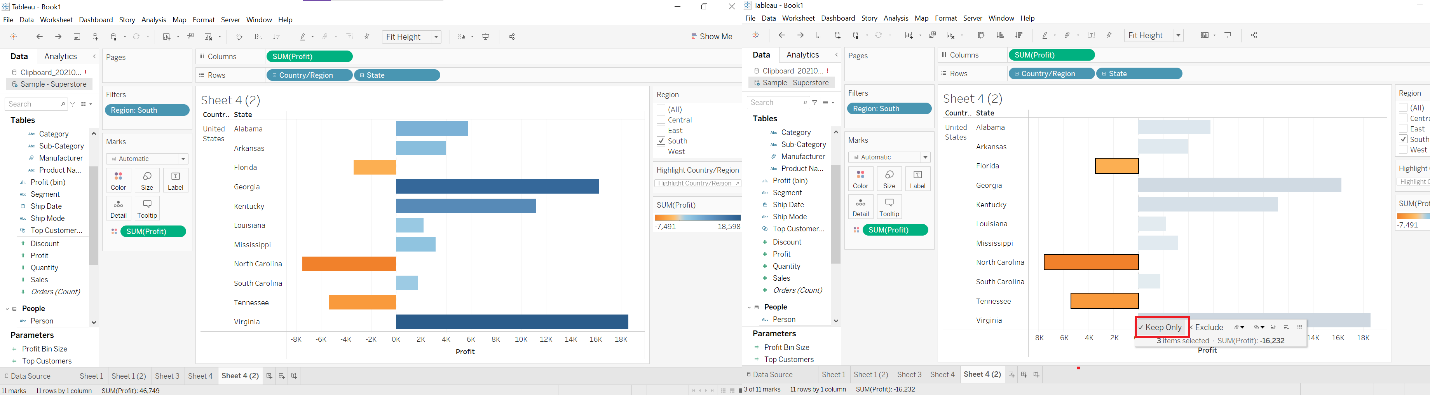
VD: Map view thể hiện lợi nhuận bán hàng của các bang ở miền nam, màu sắc của mỗi bang sẽ dựa vào số lượng bán ở bang đó, tương quan với nhau qua độ đậm nhạt của màu sắc. Ở đây ta có thể thấy các bang có lợi nhuận âm biểu diễn màu cam, lợi nhuận dương biểu diễn màu xanh.



1. Đi sâu vào chi tiết

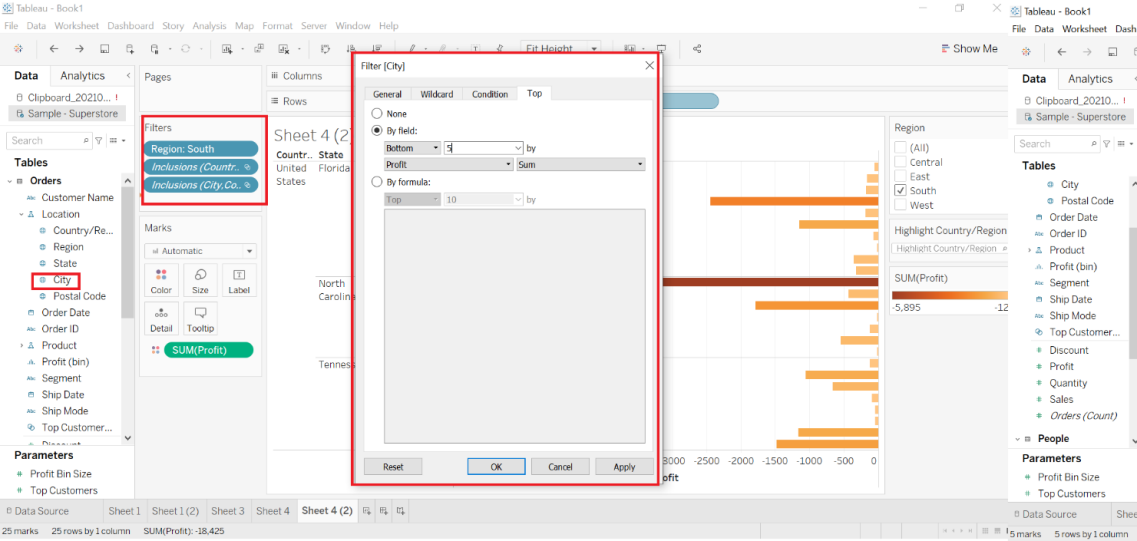
* Tùy vào mục đích, có thể lựa chọn các loại biểu đồ khác nhau để biểu diễn theo nhiều cách nhằm phân tích dữ liệu 1 cách tổng quát, kỹ lưỡng hơn.

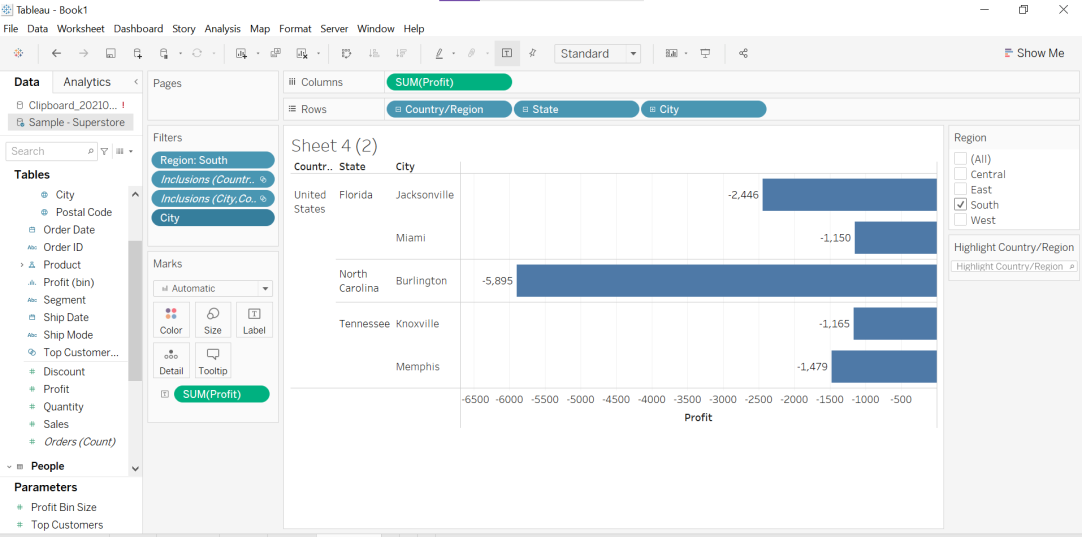
VD: Để biểu diễn rõ chi tiết lợi nhuận có thể lựa chọn horizontal bars, giúp dễ dàng quan sát được sự chênh lệch lợi nhuận của các bang.



* Có thể lựa chọn các cột mong muốn quan sát, kéo thả con trỏ đến những cột muốn quan sát, sau đó trên trên tooltip lựa chọn Kêp Only. Ngoài ta có thể chọn Exclude để loại bỏ những cột không muốn quan sát
* Top N Filter để đặt giới hạn cho số lượng label được hiển thị trong chế độ xem. Trong phần Data pane, kéo trường dữ liệu muốn lọc vào Filters shelf, trong Filter dialog box, chọn tab Top, and sau đó thực hiện lọc dữ liệu theo ý muốn

VD: Lọc top 5 thành phố có lợi nhuận thấp nhất:



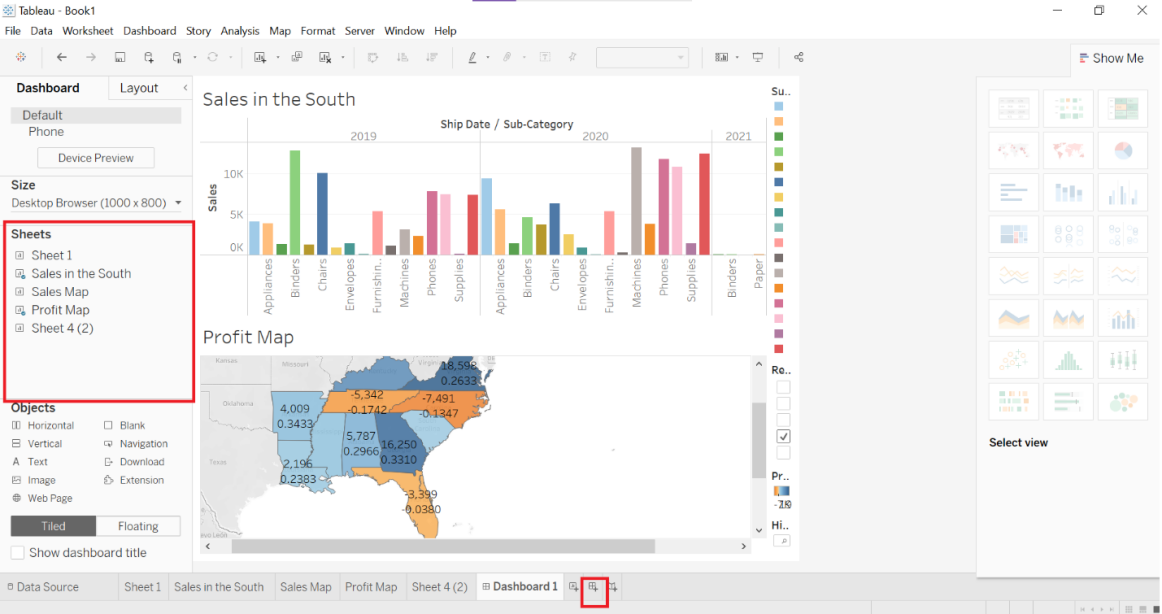


1. Dashboard

* Để tổng hợp tất cả những biểu đồ muốn quan sát với nhau vào trong 1 worksheets, và muốn các biểu đồ muốn tương tác với nhau có thể dụng Dashboard.
* Tạo 1 Dashboard:

+ Click vào New dashboard.

+ Trong phần Dashboard pane, kéo các biểu đồ muốn thể hiện vào phần dashboard trống.



* Các thao tác xử lý thay đổi trên biểu đồ gốc, biểu đồ thể hiện ở Dashboard cũng có thể thay đổi theo.
* Có thể thực hiện thay đổi ngay ở Dashboard, click chuột phải vào phần muốn thay đổi, thực hiện thay đổi trên tooltip.

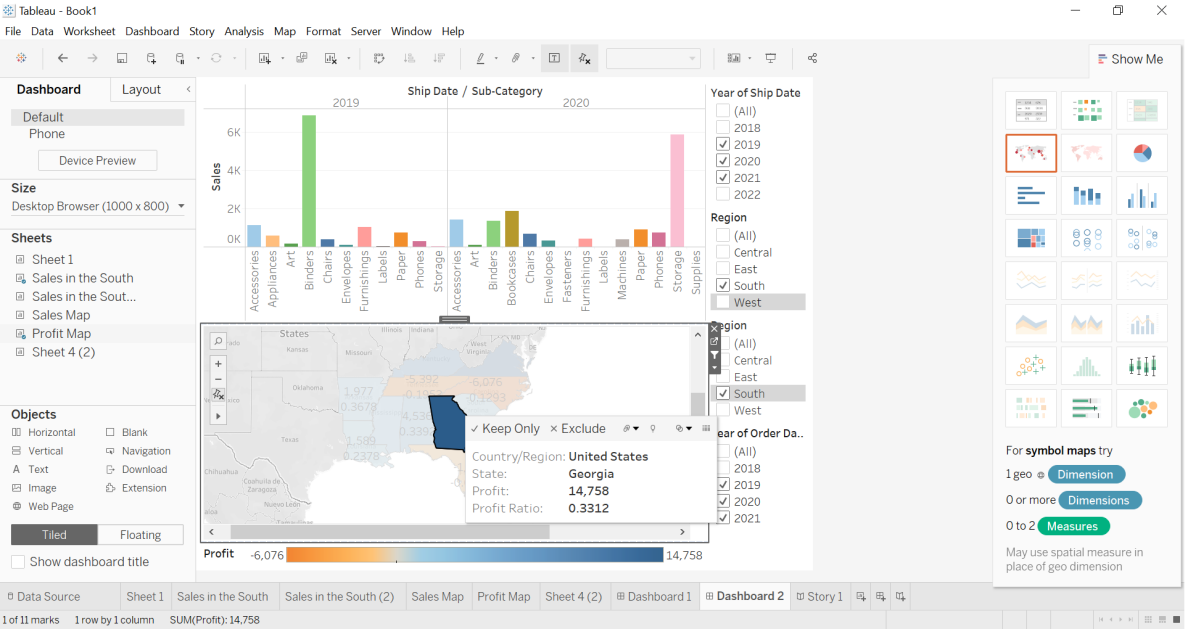
VD: Dashboard biểu đồ số lượng bán ra và map thể hiện lợi nhuận thu được ở các bang miền Nam sau khi chỉnh lại, có thể tương tác ngay trên dashboard bằng các lựa chọn năm, miền.

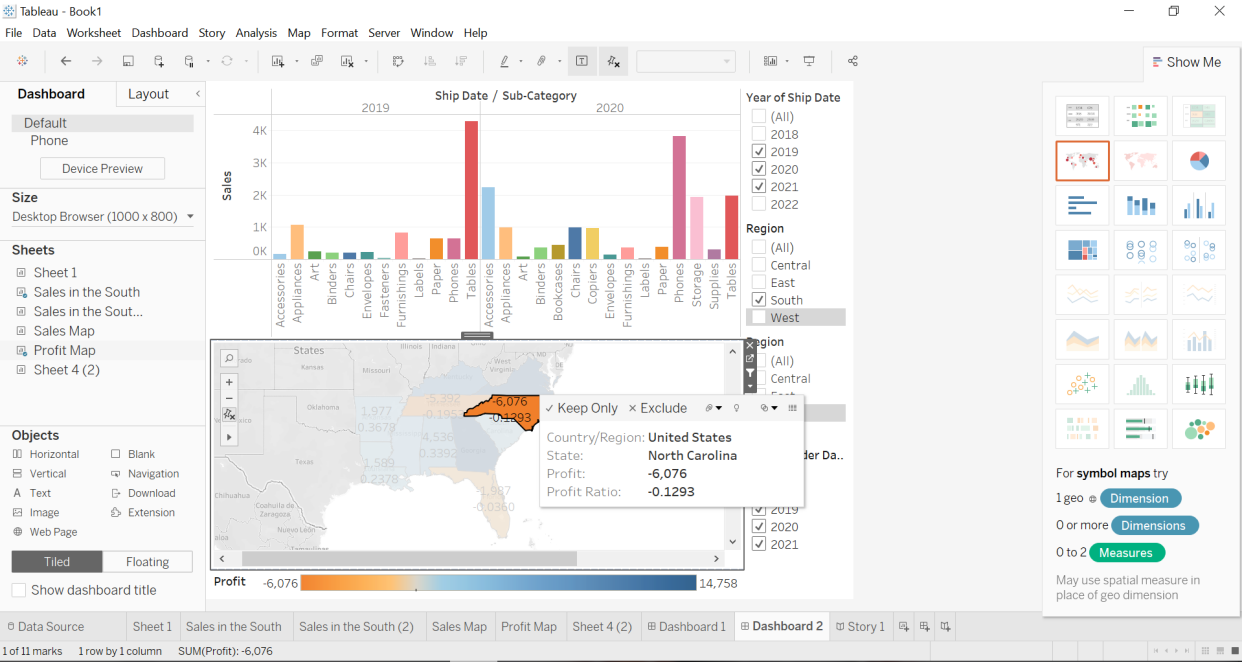
Ảnh có chứa văn bản, công cụ viết, văn phòng phẩm

Mô tả được tạo tự động

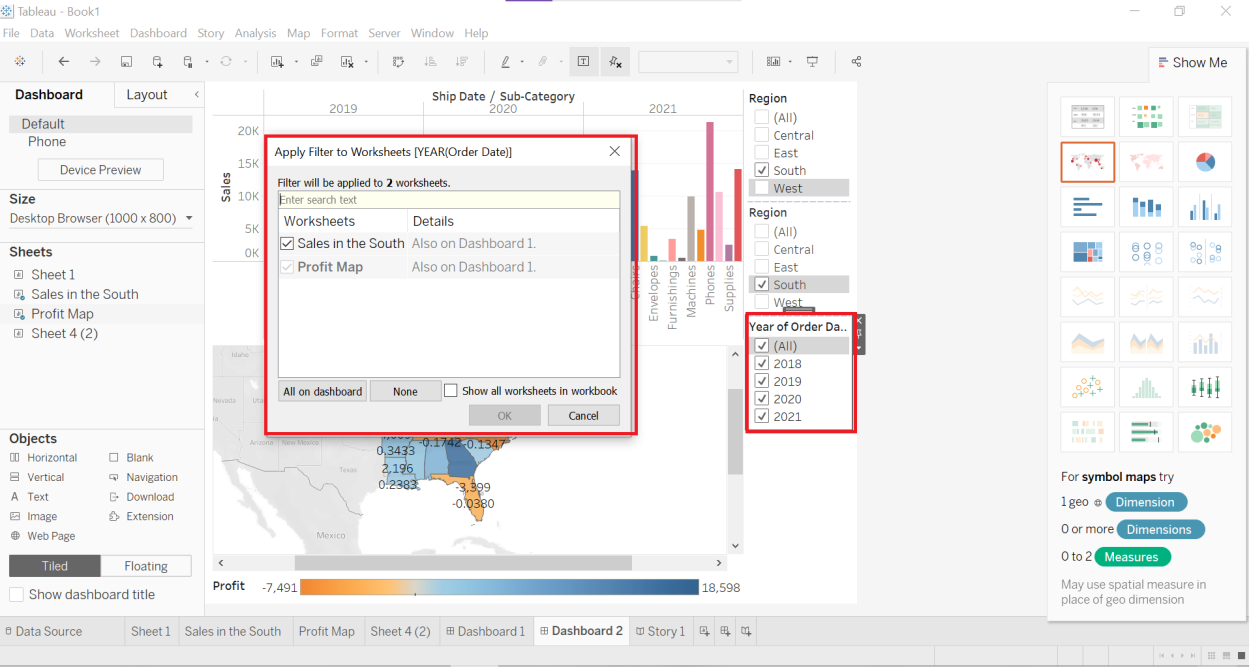
* Để có thể tương tác giữa 2 biểu đồ, chọn **Use as filter** ở góc trên cùng bên trái của biểu đồ.

VD: số lượng bán hàng của bang Georgia khi click vào vị trí của bang trên bản đồ



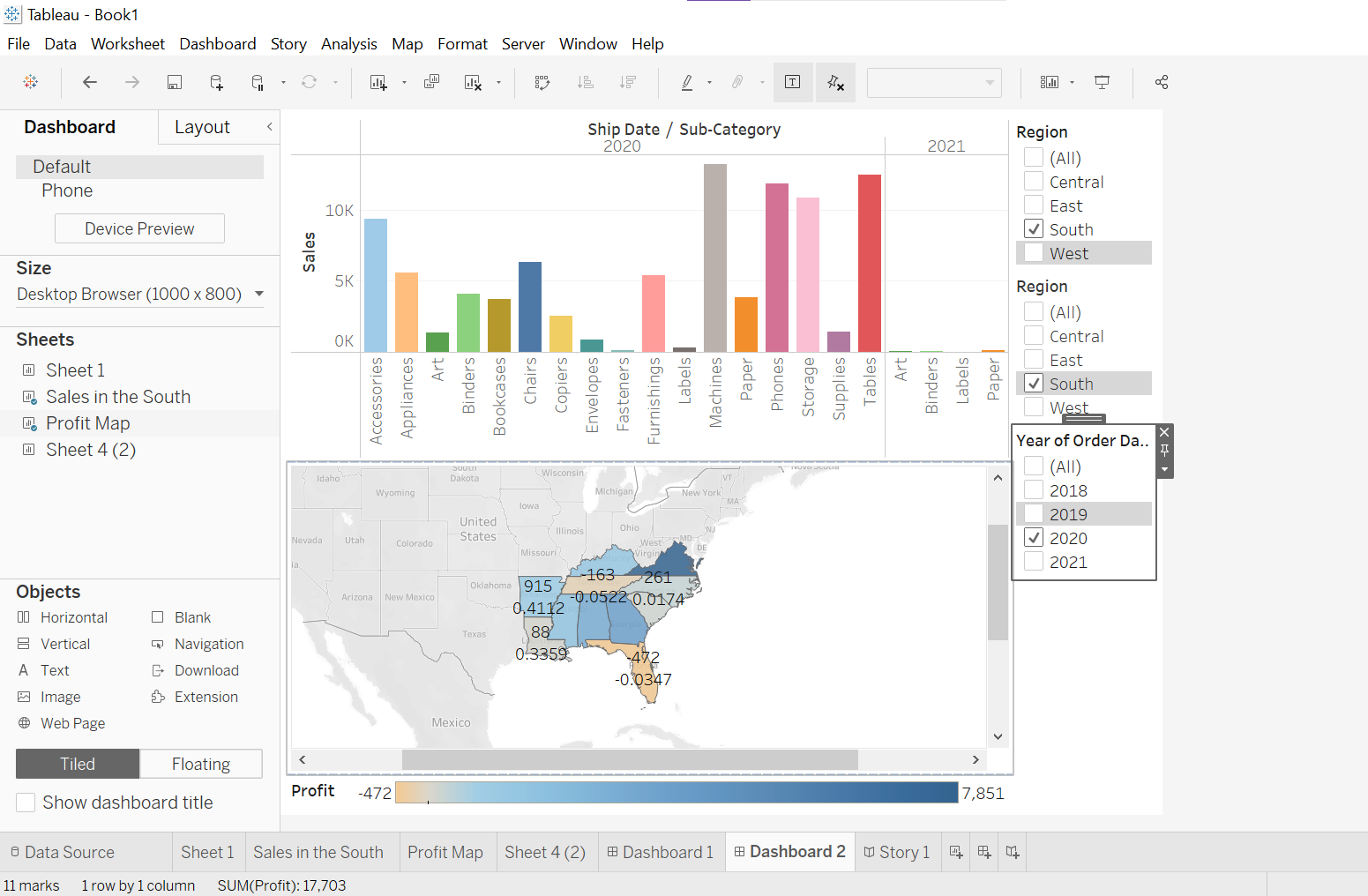


* Để thấy được thay đổi qua các năm của cả 2 biểu đồ, nhaaso chuột phải vào filter thời gian, chọn Apply to Worksheets -> Selected Worksheets. Trong Apply Filter to Worksheets dialog, chọn  All in dashboard, và chọn OK.



* Sau đó, tùy chỉnh thời gian muốn quan sát

VD: Dashboard biểu đồ số lượng bán ra và map thể hiện lợi nhuận thu được ở các bang miền nam vào năm 2020.



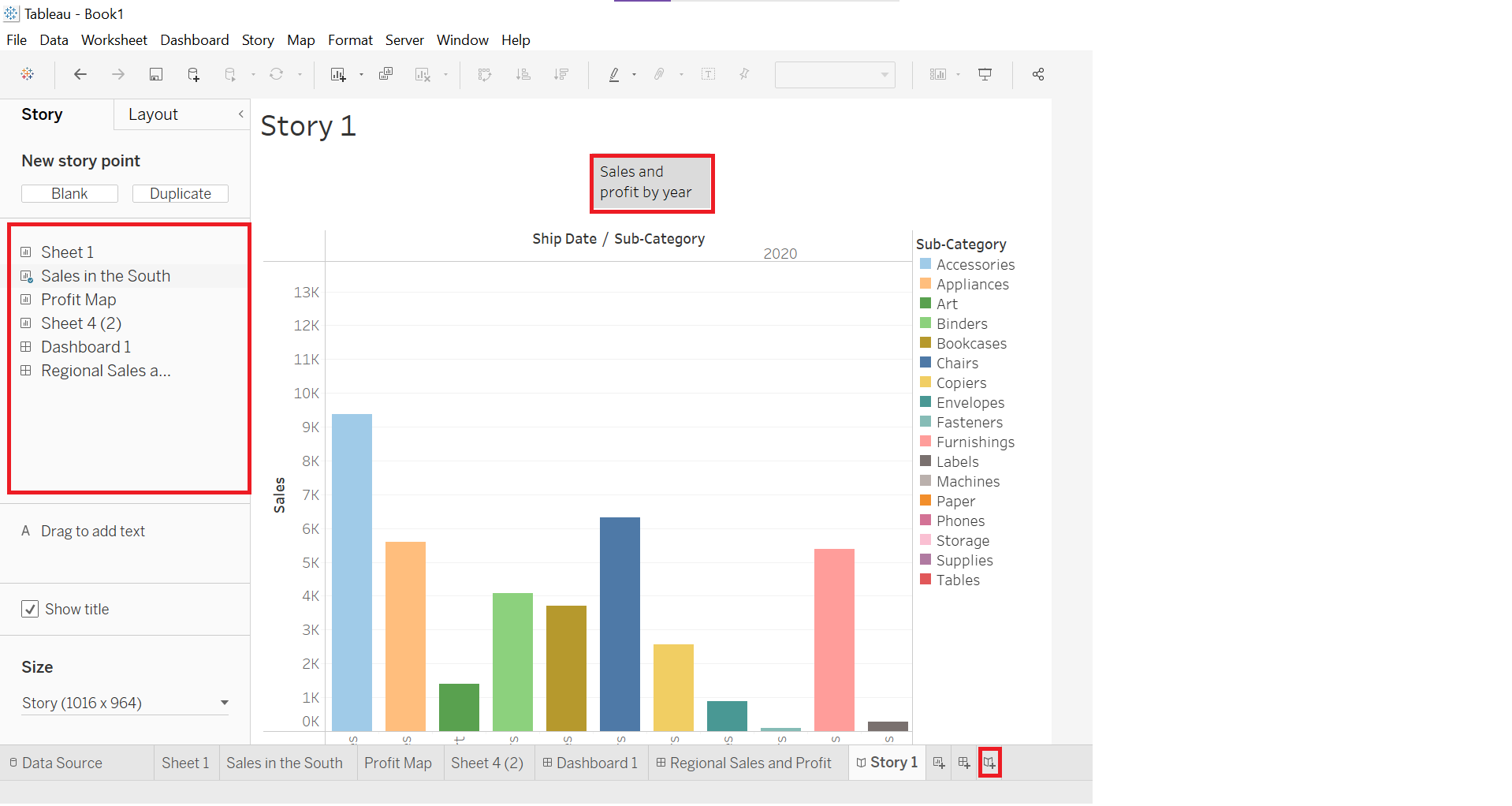
1. Story

* Là một chuỗi các hình ảnh phối hợp với nhau để truyền tải thông tin. Có thể tạo các story để kể một câu chuyện dữ liệu, cung cấp bối cảnh, chứng minh các quyết định liên quan đến kết quả như thế nào, hoặc chỉ đơn giản là đưa ra một trường hợp thuyết phục.
* Để tạo 1 story mới:

+ chọn **New story.**

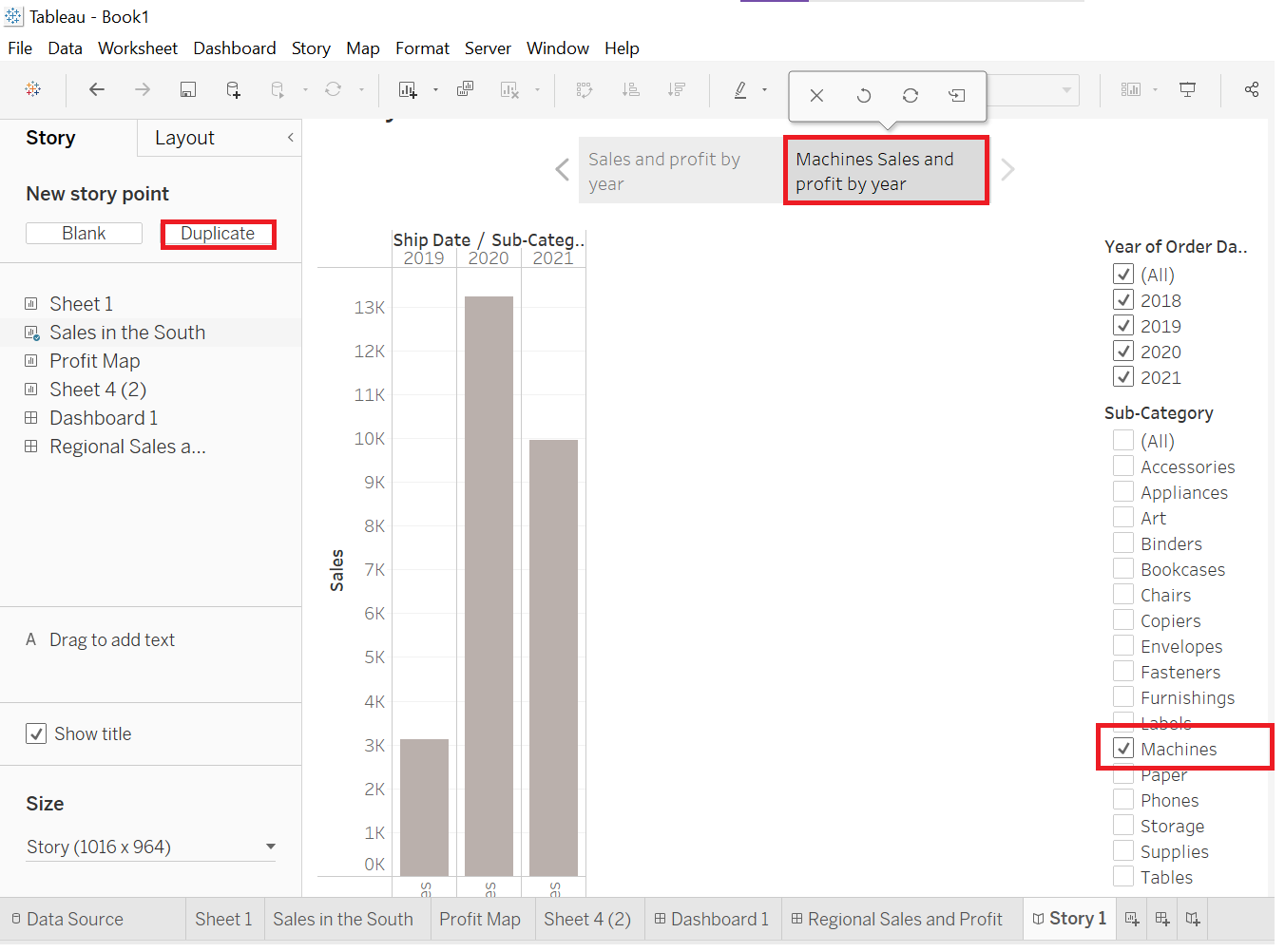
+ Trong Story pane kéo thả các biểu đồ hoặc dashboard mong muốn.

+ Thêm chú thích bằng cách chỉnh sửa văn bản trong hộp màu xám phía trên worksheet.

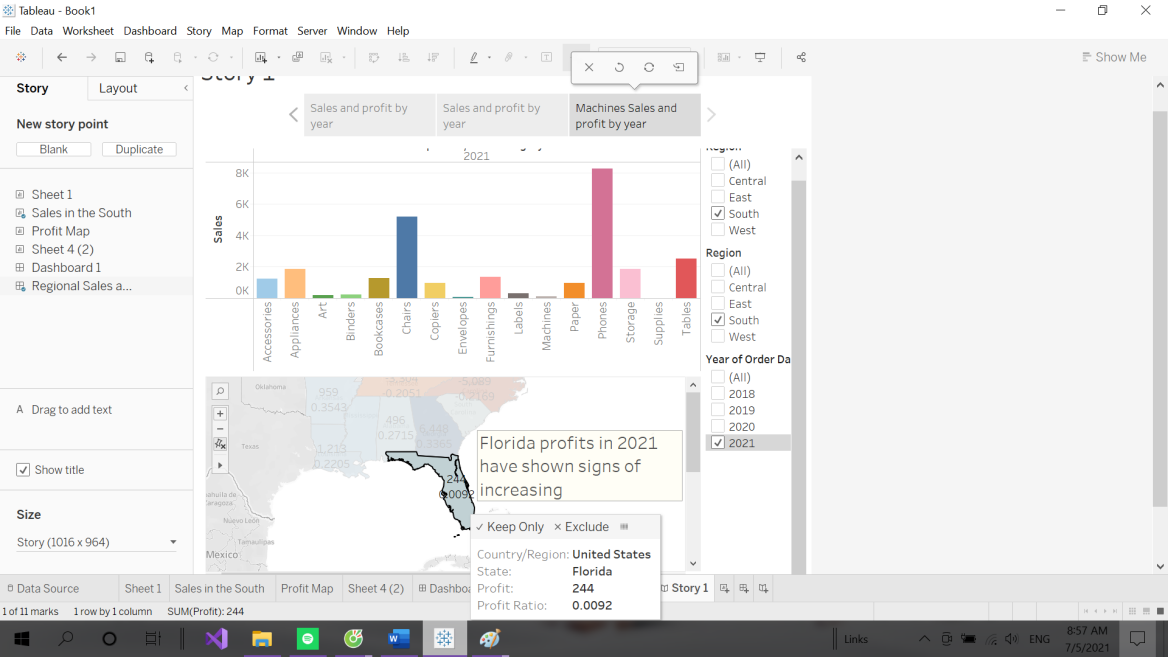


* Để thể hiện kĩ hơn trong biểu đtory panel, chọn Duplicate sau đó thực hiện các thao tác phân tích chi tiết biểu đồ như các bước trên (Filter, Top N Filter)

VD: xác định doanh số và lợi nhuận của Machines theo năm



* Nếu muốn ghi chú vào các ý muốn thể hiện. Trong pane ở phía trái, kéo Drag to add text vào vị trí thích hợp



1. **Tiền xử lý dữ liệu:**

Sử dụng thư viện Pandas, Numpy để đọc và sử lý dữ liệu

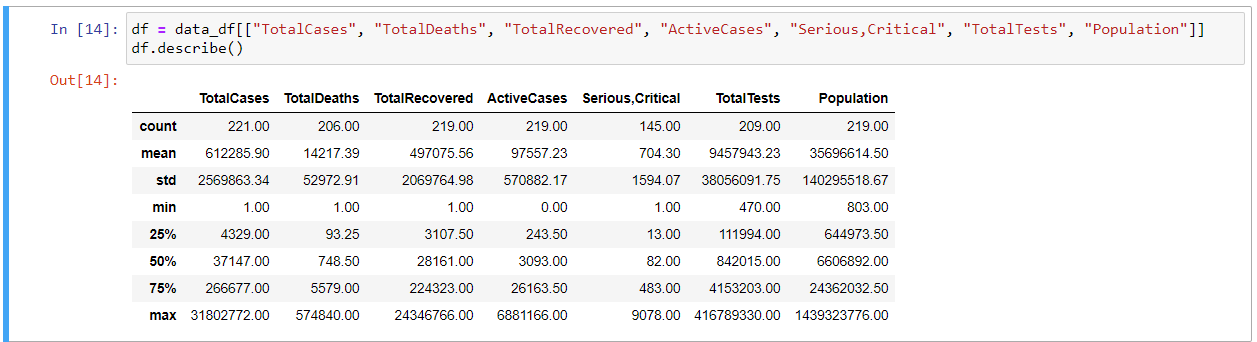
Chuyển dữ liệu từ Object thành numeric trừ cột Date và Country,Others

Để đảm bảo tính nhất quán và khách quan, không xử lý dữ liệu có giá trị NaN

1. **Phân tích và trực quan dữ liệu:**
2. P**hân tích một sô trường dữ liệu**
3. Các trường dữ liệu được chọn:

Chọn các thuộc tính TotalCases, TotalDeaths, TotalRecovered, ActiveCases, Serious,Critical, TotalTests, Population

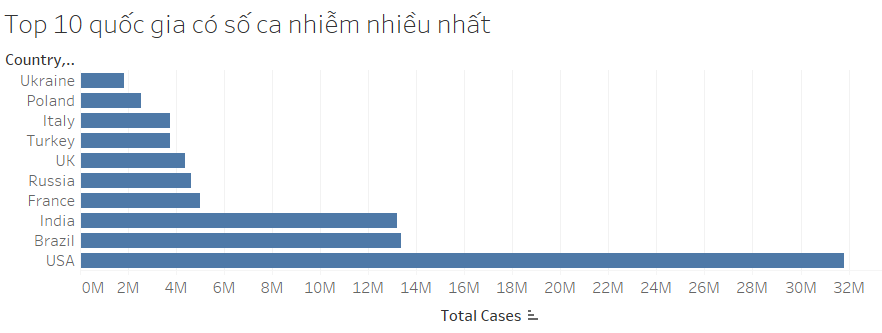
Tính chất các thuộc tính đã chọn:

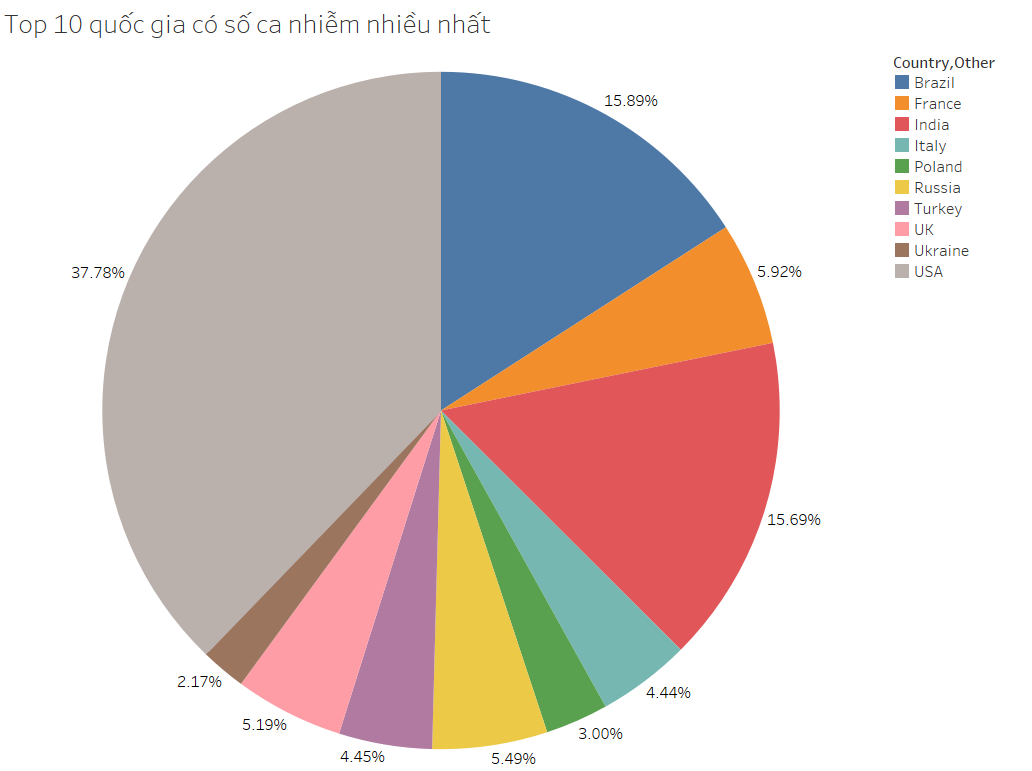


1. Phân tích các trường dữ liệu đã chọn

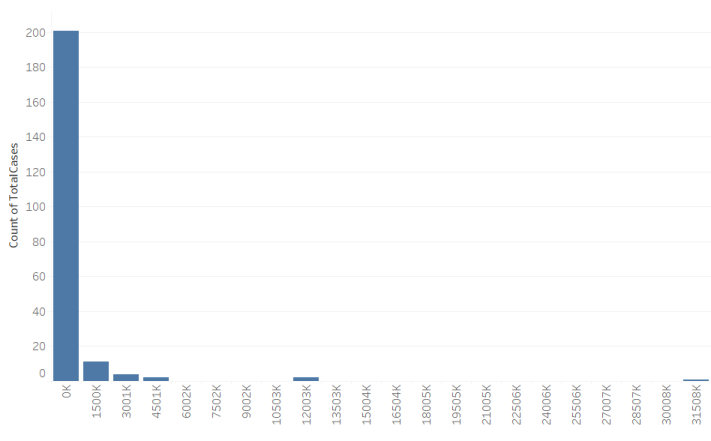
* TotalCases

Top 10 quốc gia có số ca nhiễm nhiều nhất:





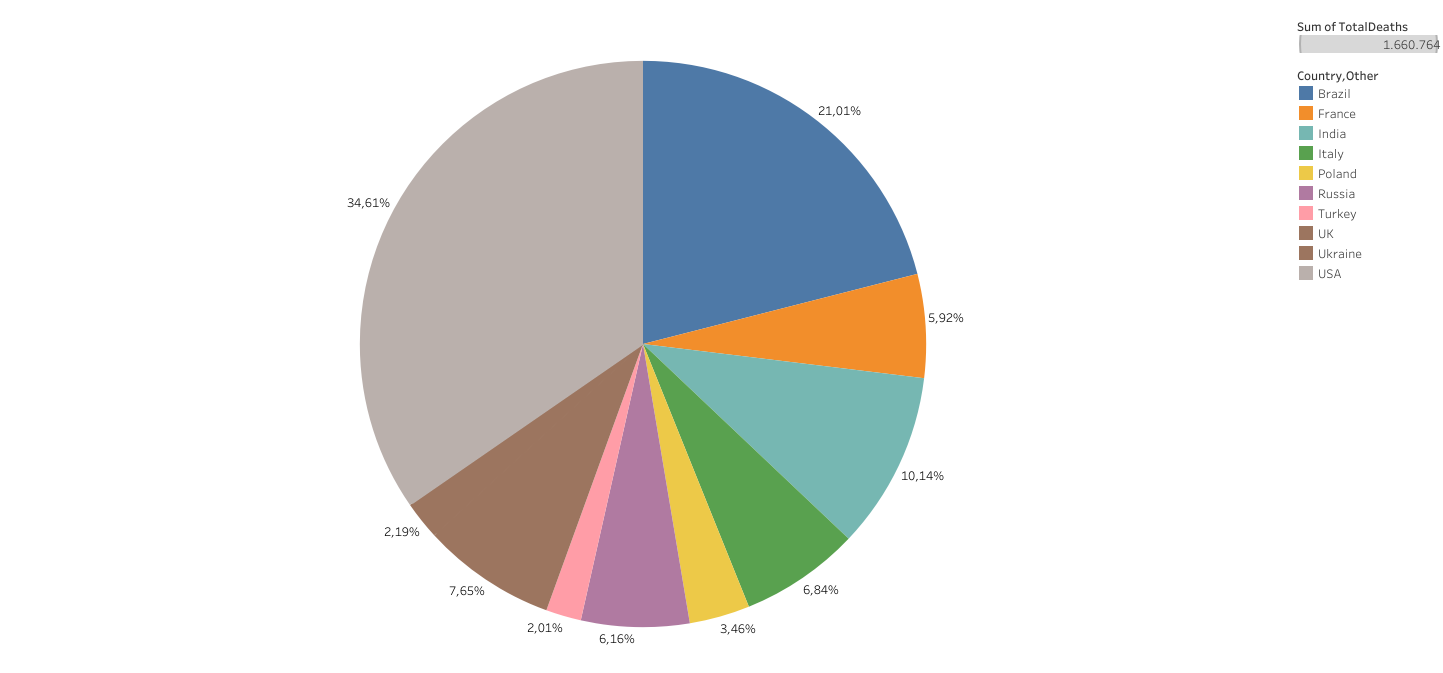
Biểu đồ thể hiện sự phân bổ về số ca nhiễm ở mỗi quốc gia:



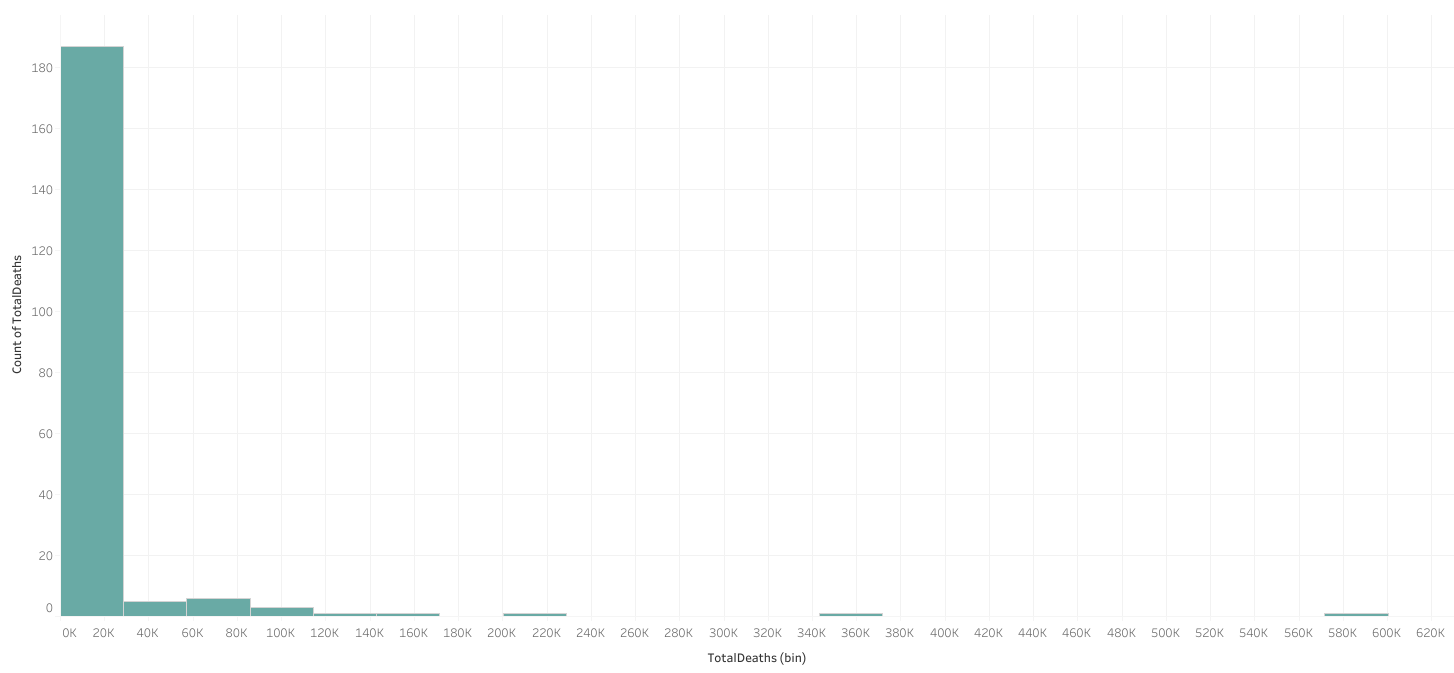
* Ta được các thông tin:
  + Hoa Kỳ là quốc gia có số ca nhiễm bệnh cao nhất
  + Phần lớn các quốc gia có số ca nhiễm ít hơn 5 triệu
  + 3 nước có số ca nhiễm lớn hơn 5 triệu là: Hoa Kỳ (hơn 30 triệu ca), Brazil (hơn 13 triệu ca), Ấn Độ (hơn 13 triệu ca)
  + Hơn
  + Sự chênh lệch về số ca nhiễm của các quốc gia có sự khác biệt lớn theo từng nhóm ( Hoa Kỳ - Brazil, Ấn Độ - phần còn lại)
* TotalDeaths

Số ca tử vong ở 10 quốc gia có số ca nhiễm lớn nhất:



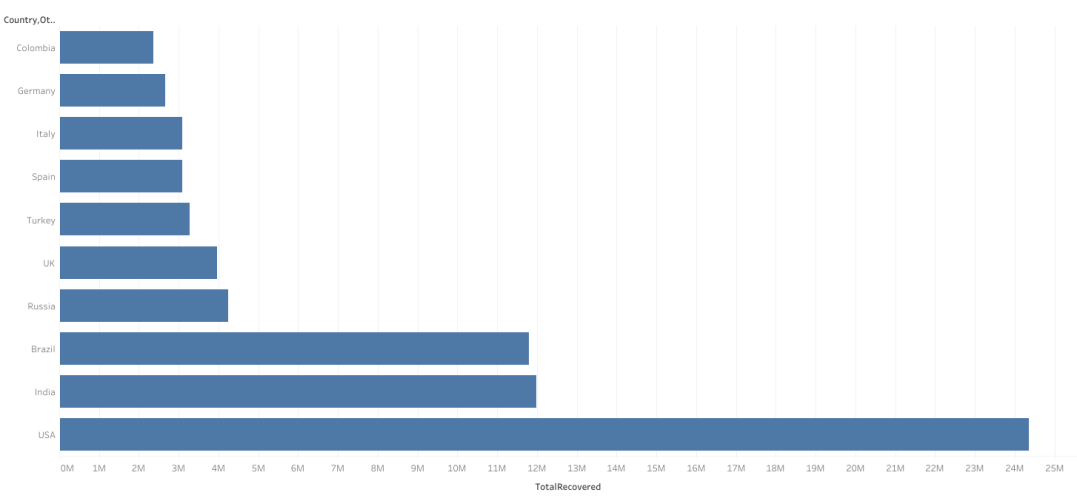


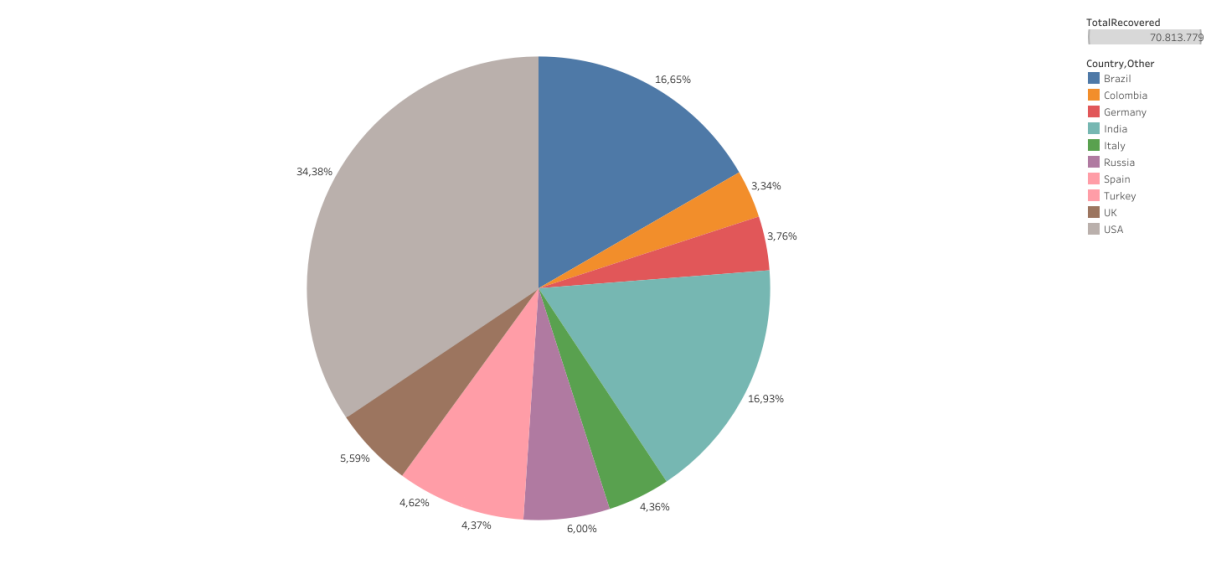
Số phân bố về số lượng quốc gia theo số ca tử vong:



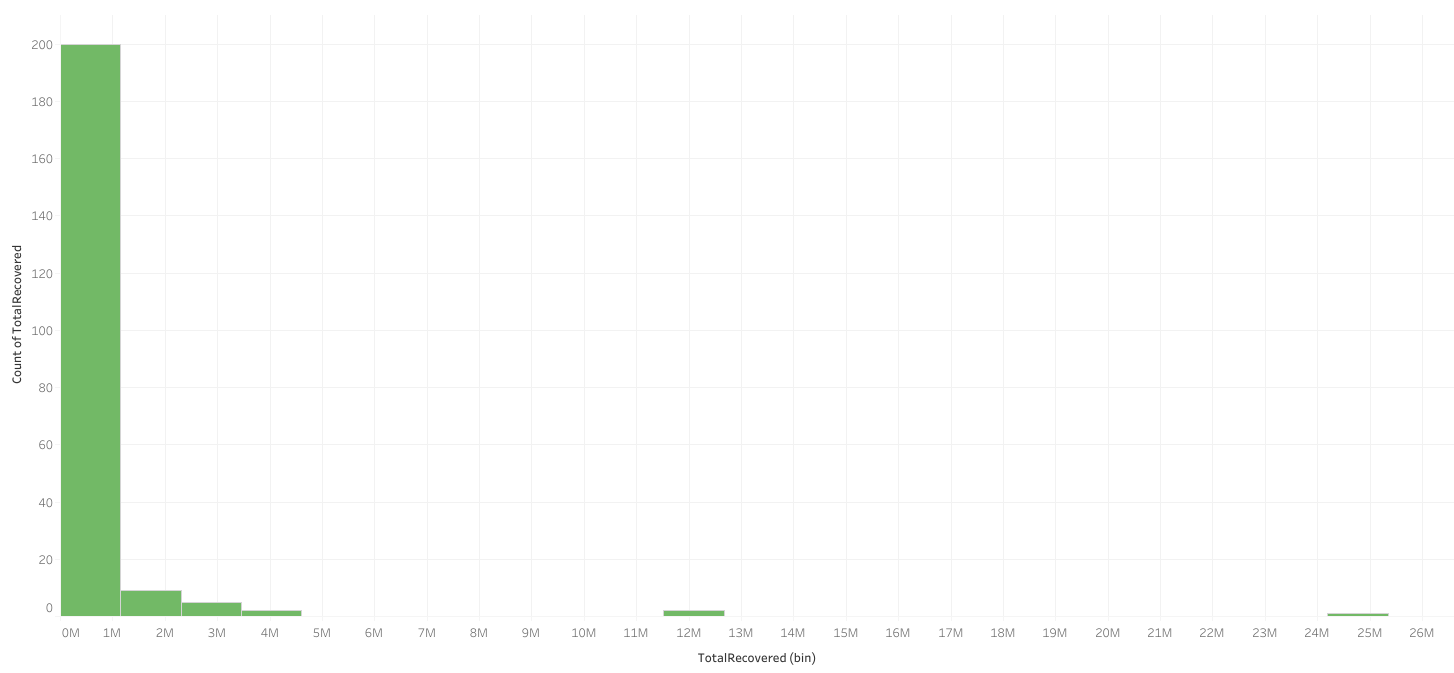
* Kết luận:
  + Hoa Kỳ dẫn đầu về số người tử vong (trong số 10 nước được xét)
  + Phần lớn các quốc gia có số ca tử vong dưới 50 000
  + Số ca tử vong của các quốc gia có sự chênh lệch lớn nhưng không phân chia thành các nhóm rõ ràng
* TotalRecovered

Tổng số ca hồi phục:



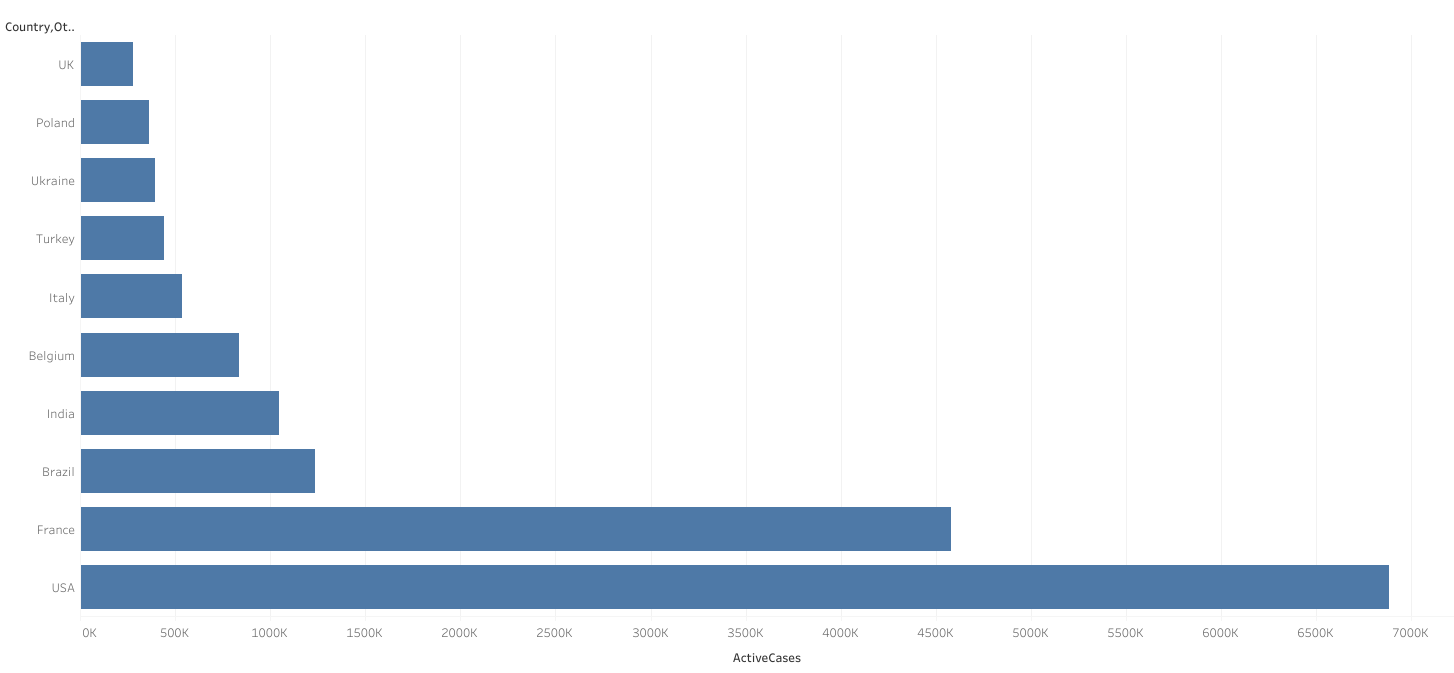


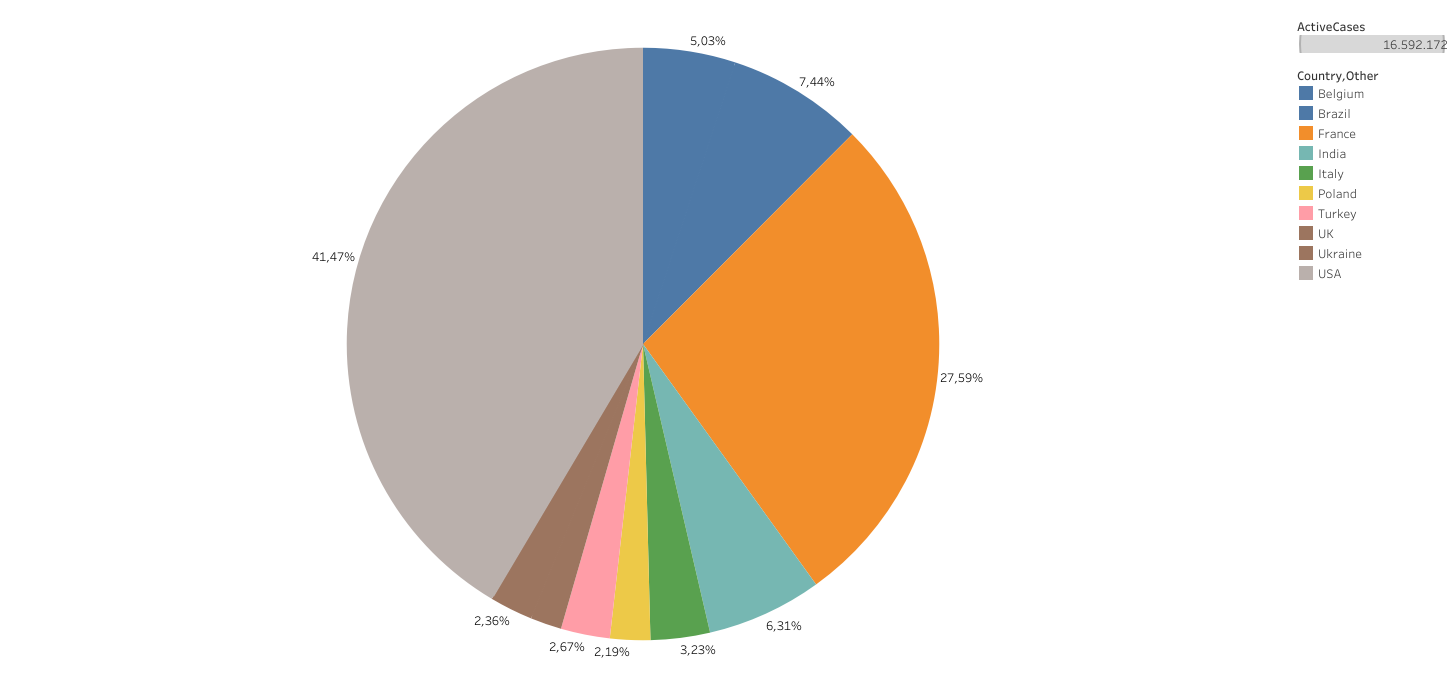
Sự phân bổ về số lượng quốc gia theo số ca hồi phục:



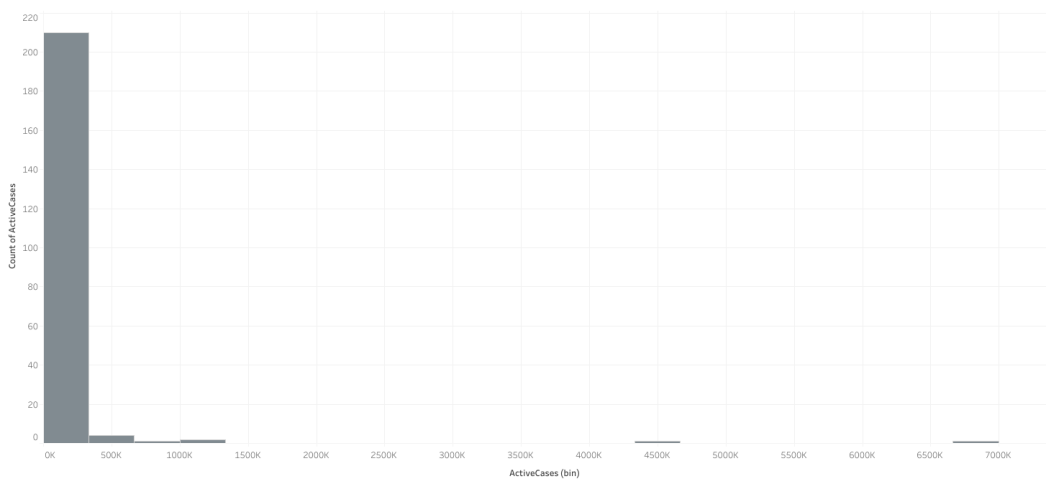
* Kết luận:
  + Hoa Kỳ có số ca hồi phục nhiều nhất
  + Phần lớn quốc gia có số ca hồi phục dưới 5 triệu
  + Số ca hồi phục có sự chênh lệch lớn và phân thành từng nhóm (Hoa Kỳ - Brazil, Ấn Độ - còn lại)
* ActiveCases

Số ca bệnh tại thời điểm phân tích



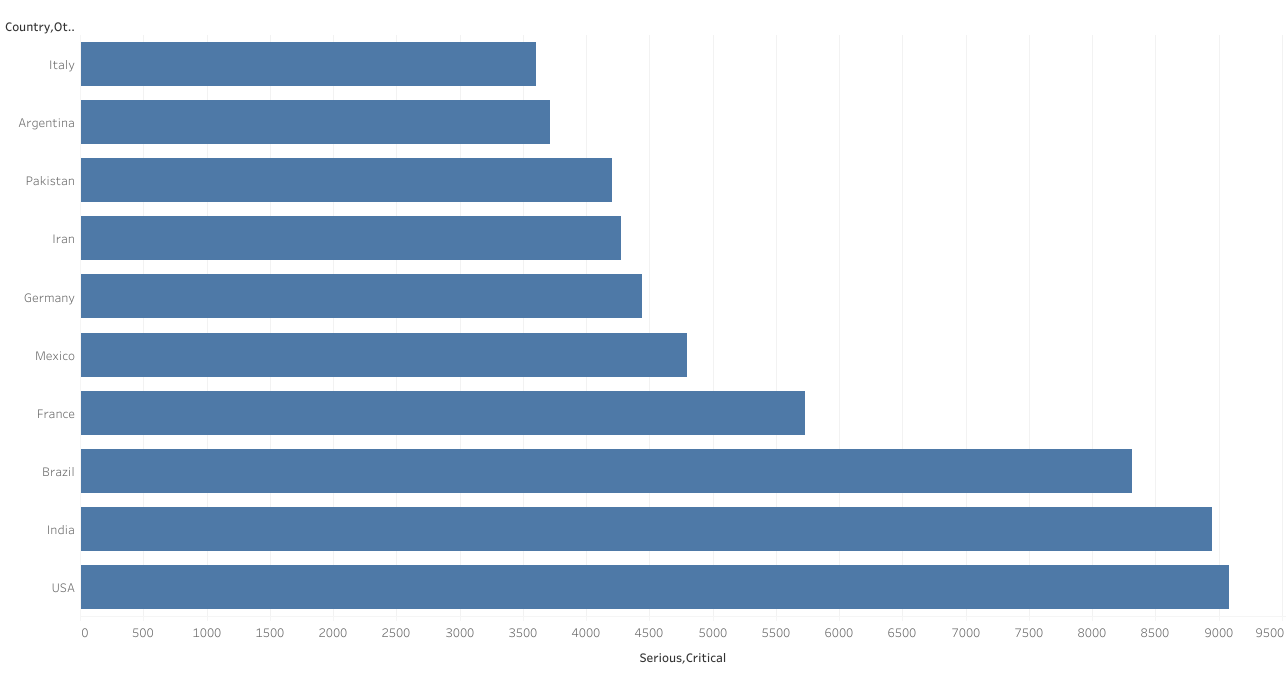


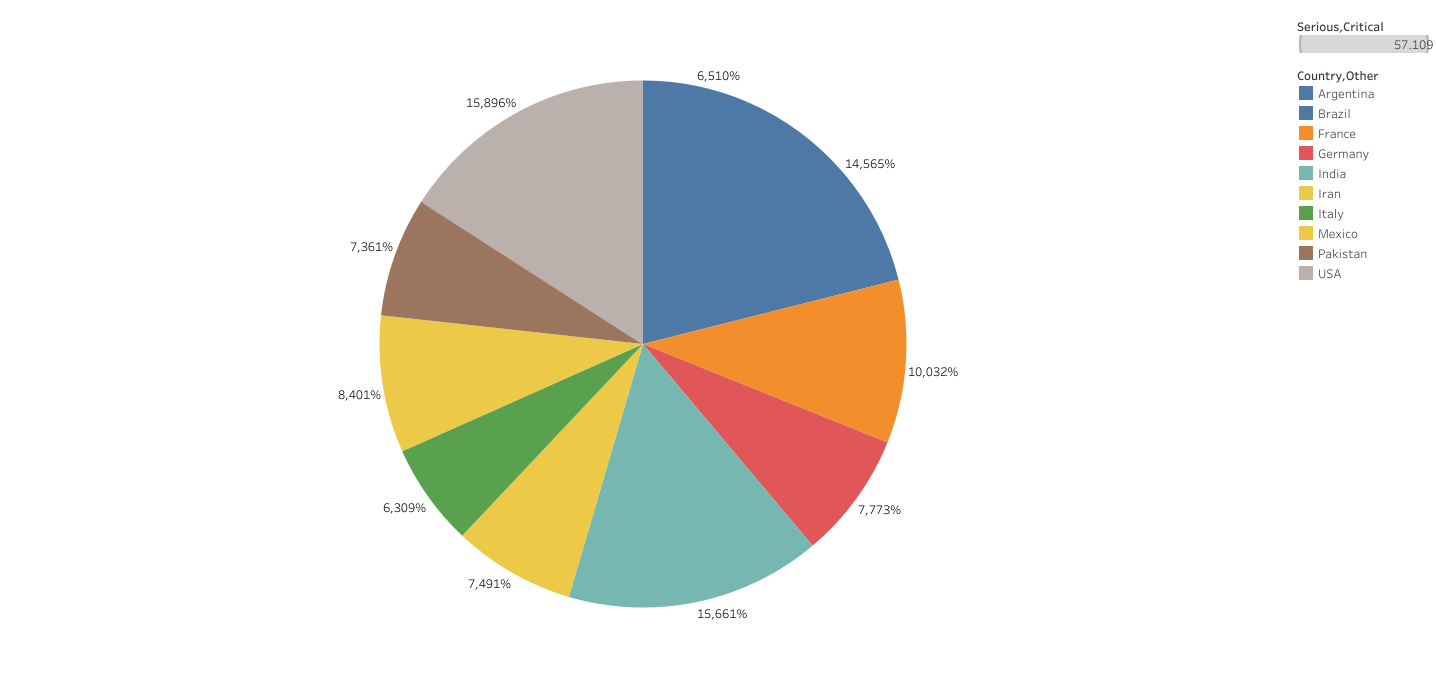
Sự phân bổ về số lượng quốc gia theo số ca dương tính:



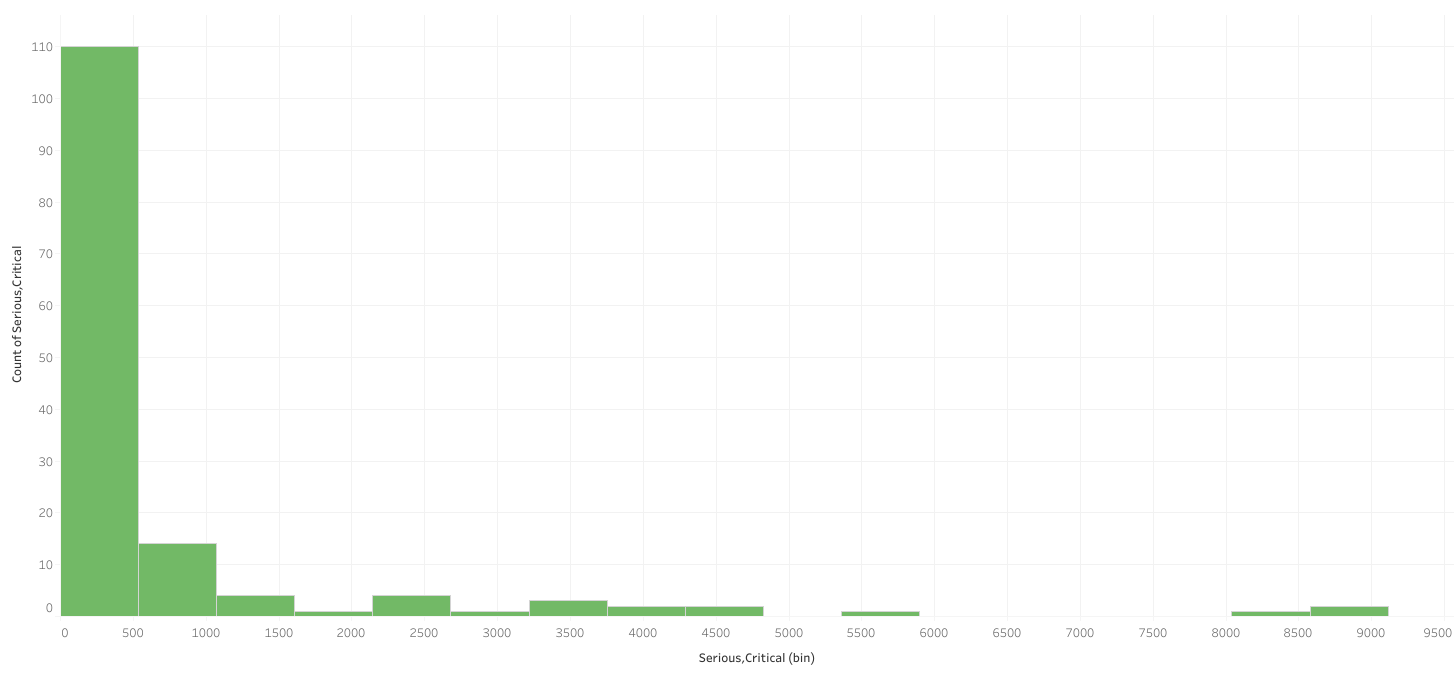
* Kết luận:
  + Hoa Kỳ có số ca dương tính nhiều nhất
  + Phần lớn quốc gia có số ca hồi phục dưới 1 triệu
  + Số ca dương tính có sự chênh lệch lớn giữa các quốc gia
* Serious,Critical

Số ca nghiêm trọng tại thời điểm phân tích



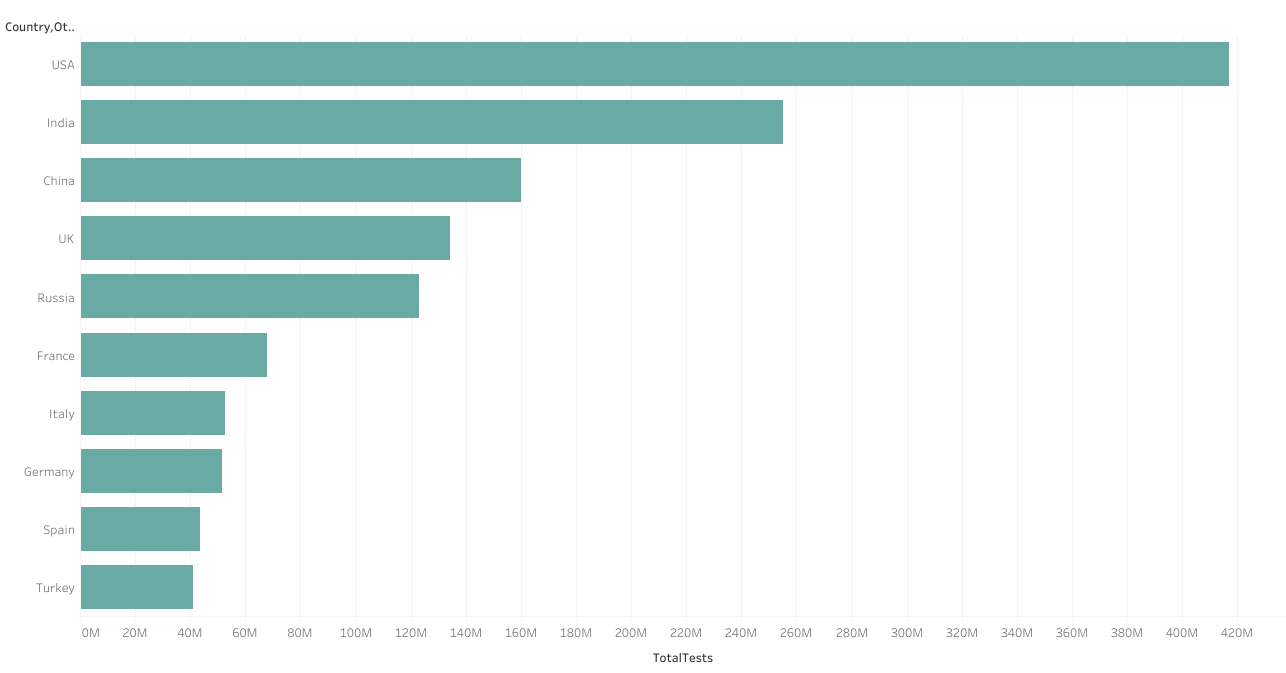


Sự phân bố về số lượng quốc gia theo số ca nghiêm trọng:



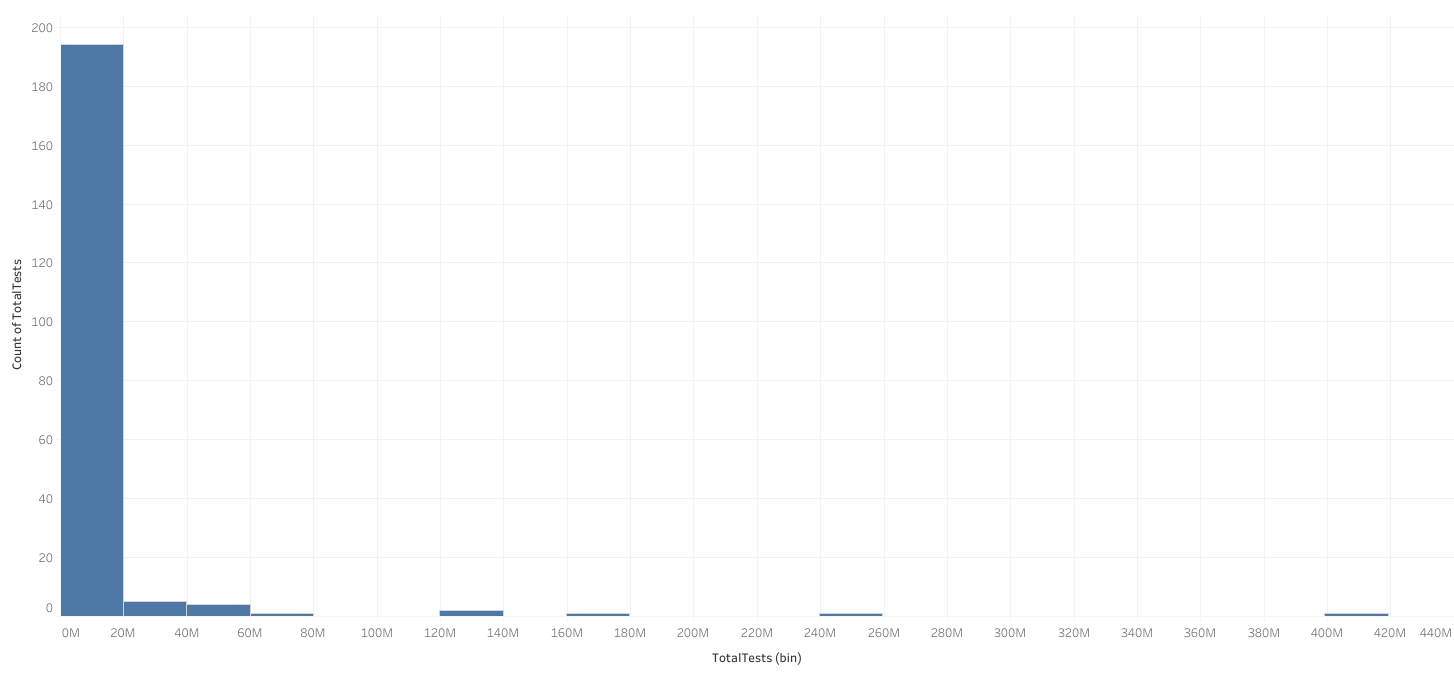
* Nhận xét:
  + Hoa kỳ có số ca nghiêm trọng nhiều nhất
  + Phần lớn quốc gia có số ca nghiêm trong ít hơn 1000
  + Số ca dương tính có sự chênh lệch không quá lớn giữa các quốc gia.
* TotalTests

Tổng số ca được xét nghiệm:





Sự phân bổ về số quốc gia theo số ca được xét nghiệm



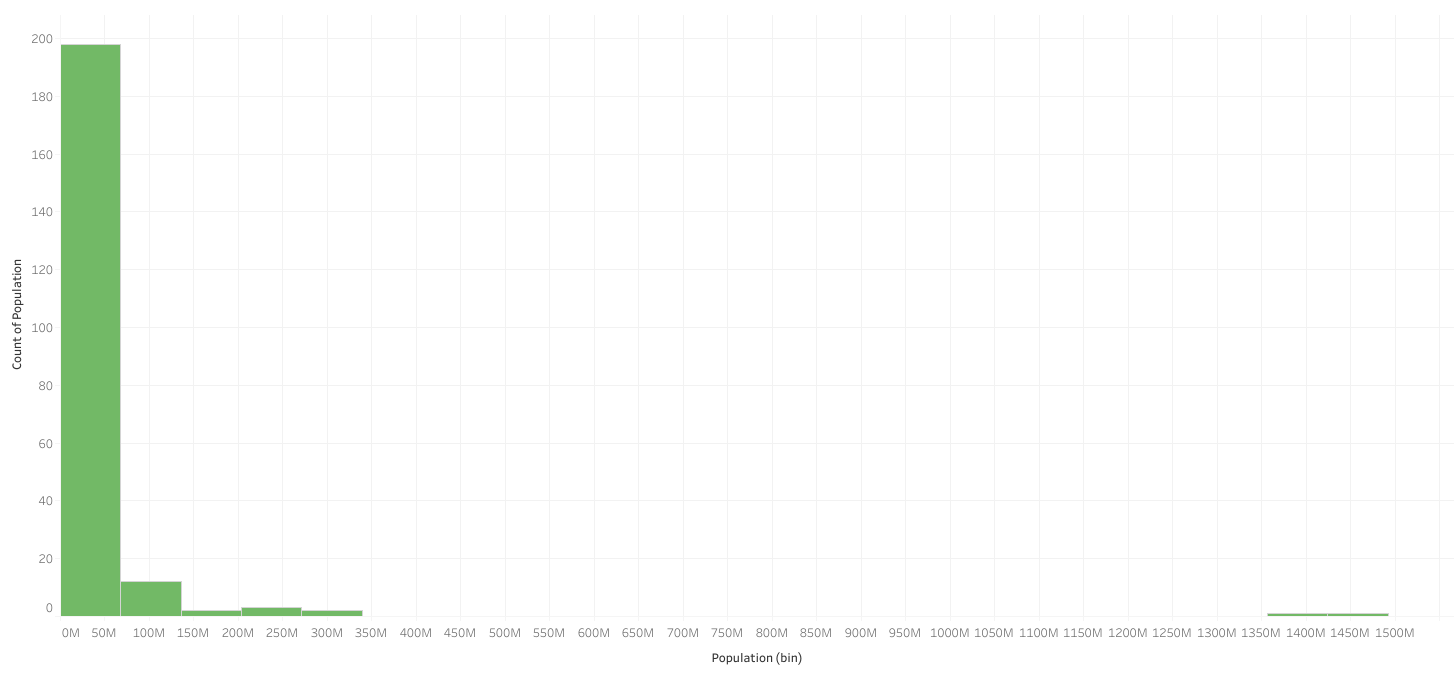
* Nhận xét:
  + Hoa kỳ có số ca xét nghiệm nhiều nhất
  + Phần lớn quốc gia có số ca xét nghiệm trong ít hơn 100 triệu
  + Số ca xét nghiệm của từng quốc gia có sự chênh lệch lớn
* Population

Dân số của các quốc gia:





Sự phân bổ về số quốc gia theo dân số:



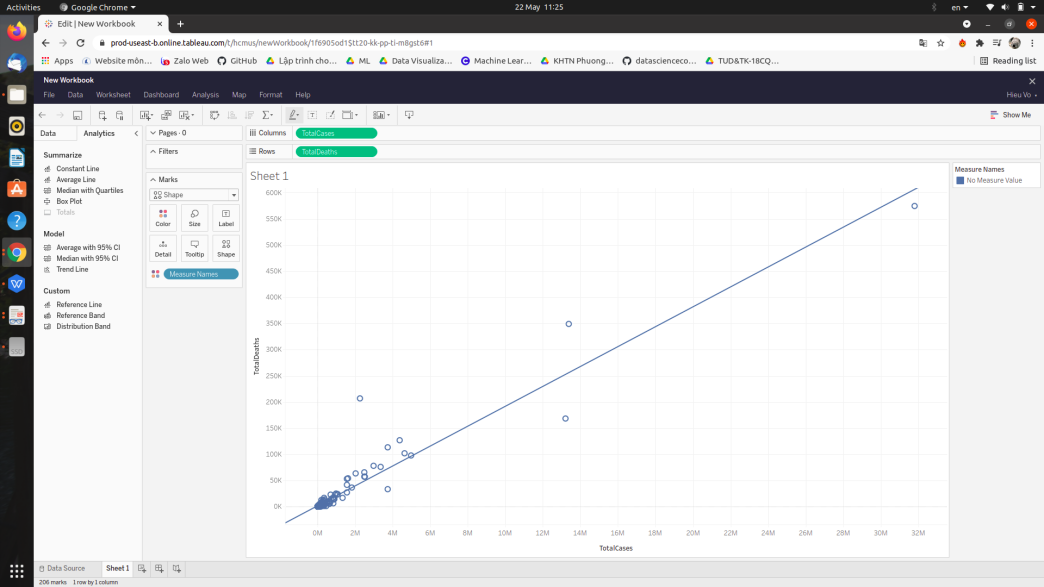
* Nhận xét:
  + Trung Quốc và Ấn Độ có dân số lớn nhất
  + Phần lớn quốc gia có dân số ít hơn 200 triệu

1. **Phân tích các cặp thuộc tính đã chọn**
2. Các cặp thuộc tính đã chọn

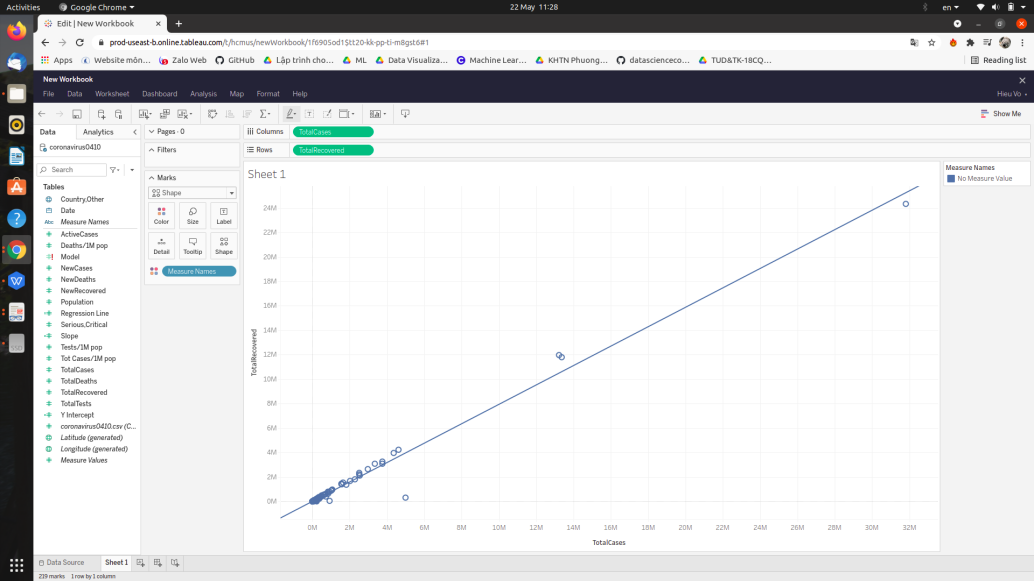
* TotalCases – TotalDeaths
* TotalCases – TotalRecovered
* TotalDeaths – TotalRecovered
* ActiveCases - TotalCases
* TotalCases – TotalTests
* TotalRecovered – TotalTests

1. Phân tích mối quan hệ

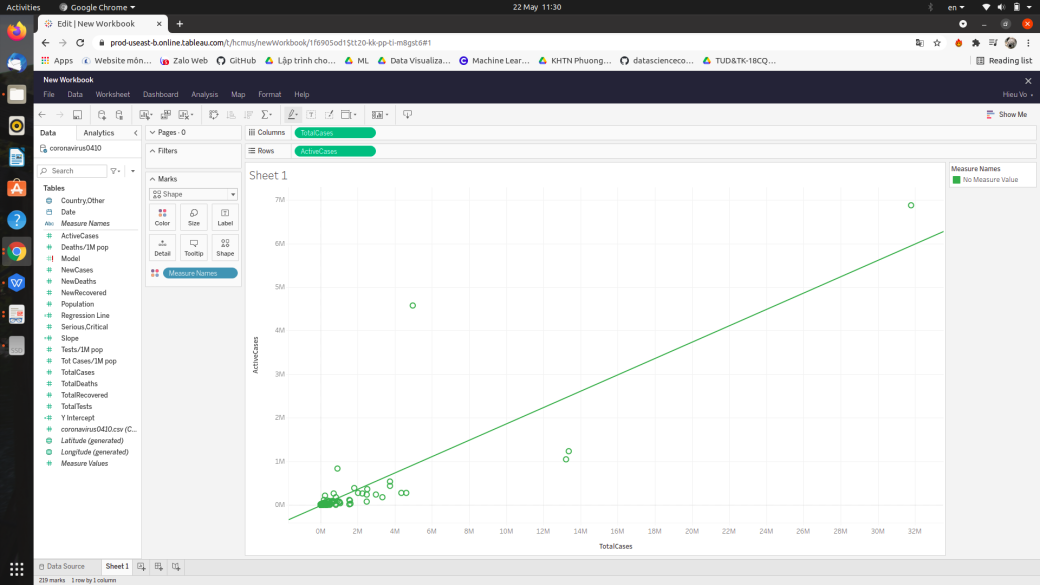
* TotalCases và TotalDeaths



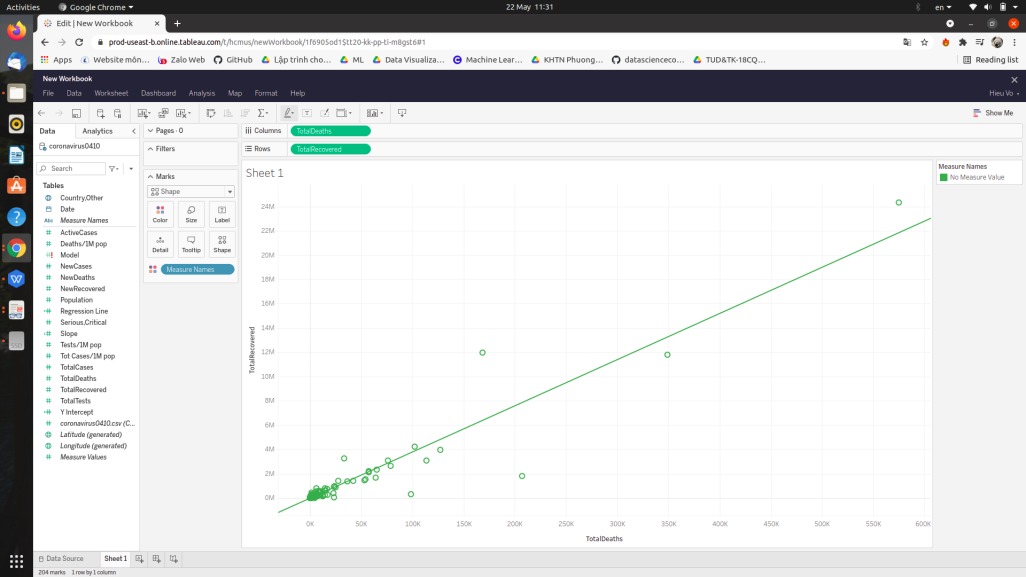
* TotalCases – TotalRecovered



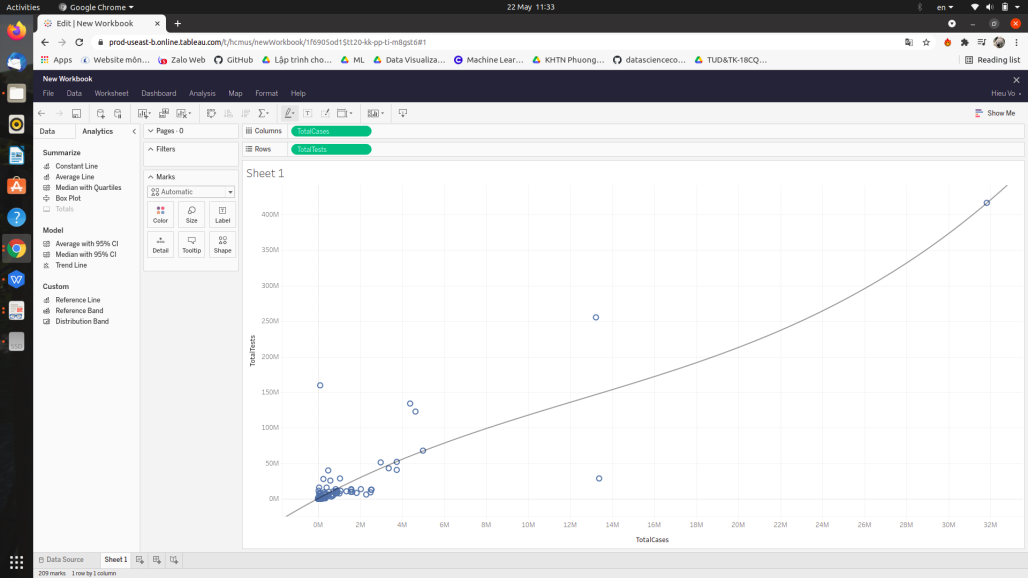
* ToalCases – ActiveCases



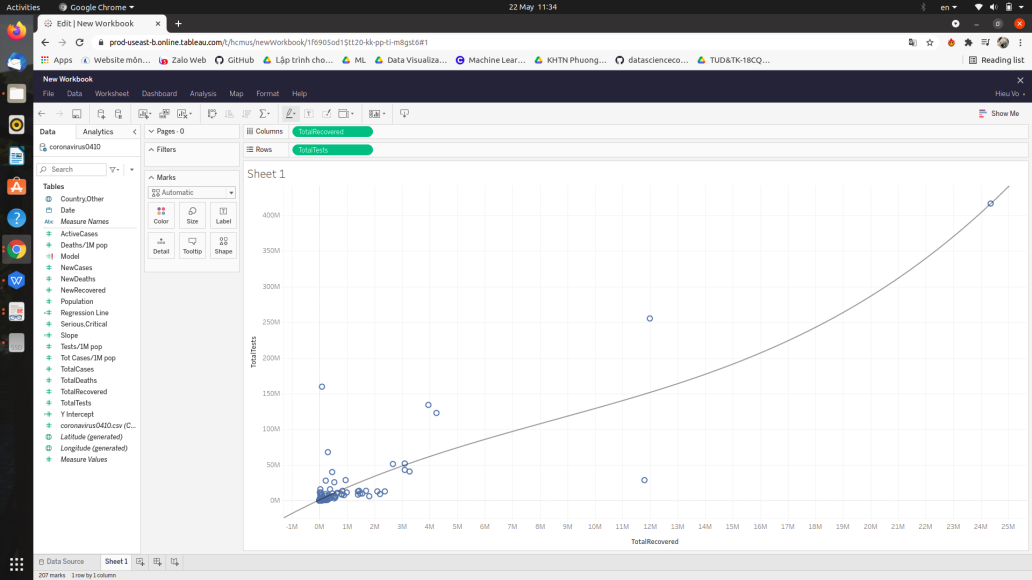
* TotalDeaths – TotalRecovered



* TotalCases – TotalTests



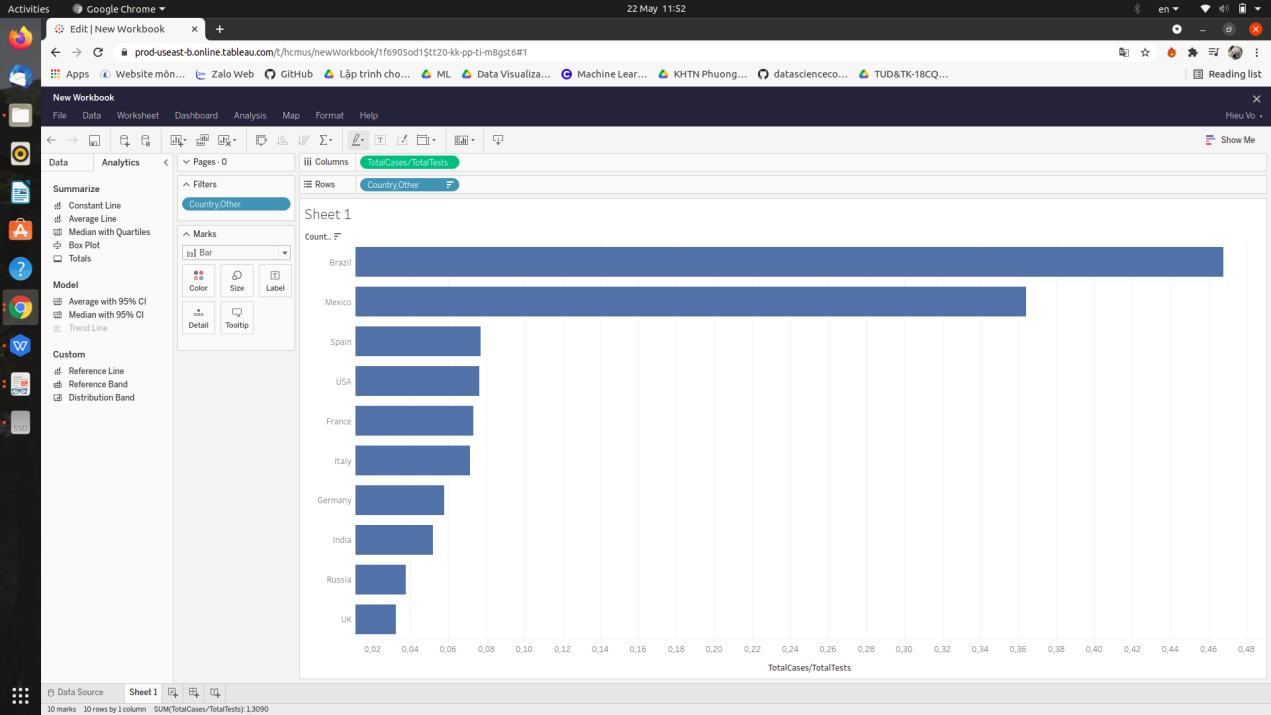
* TotalRecovered – TotalTest:



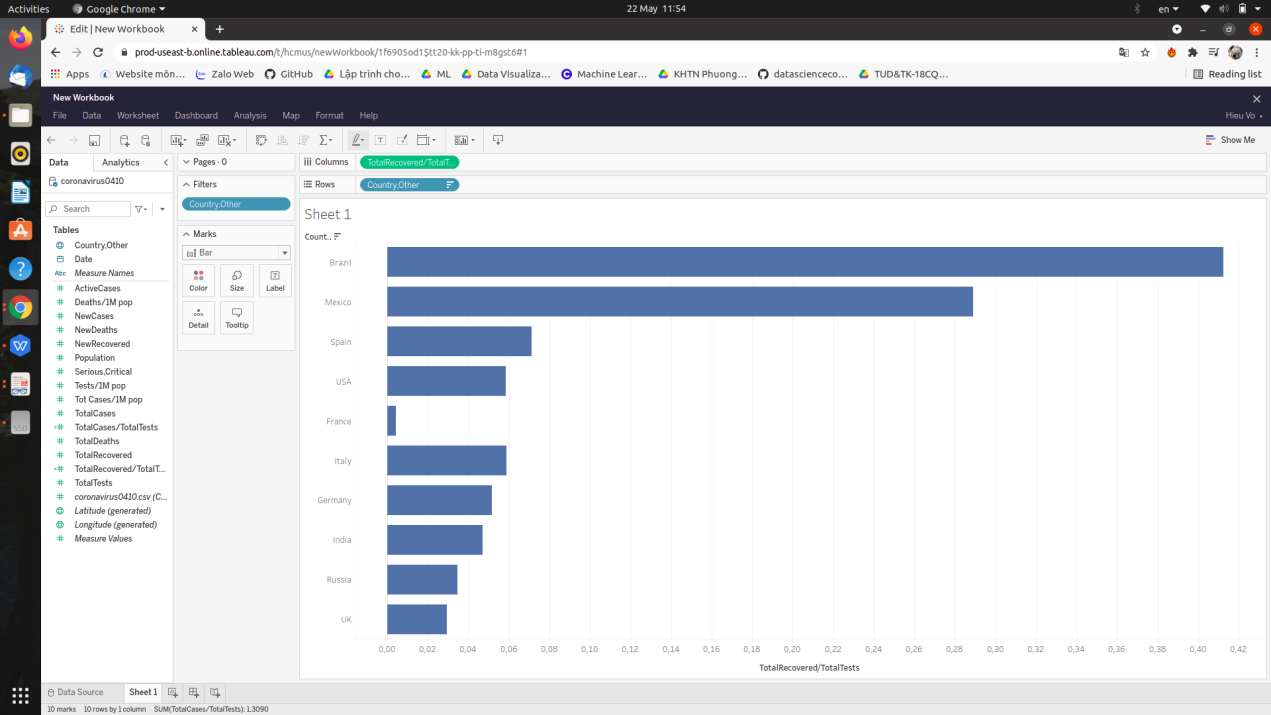
***Tất cả các cặp thuộc tính kiệt kê ở trên đều có mối quan hệ tuyến tính hoàn hảo/ tương đối hoàn hảo với nhau***

1. Phân tích các cặp dữ liệu khác:

* Tỉ lệ TotalCases/TotalTests giữa một số nước:



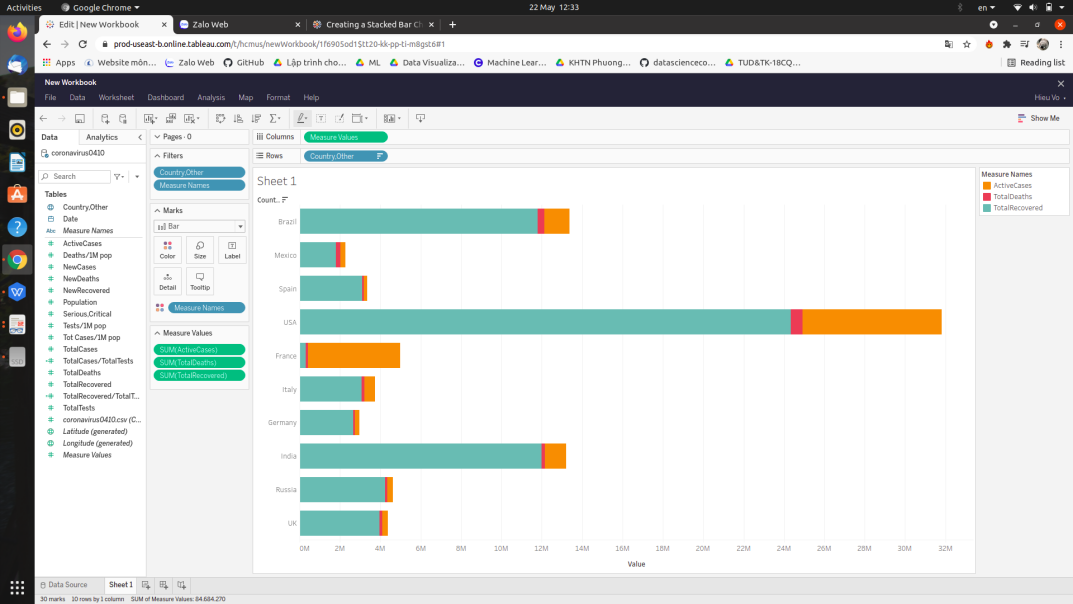
* Nhận xét:
* Tỉ lệ TotalCases/TotalTests của Mexico(>0.3) và Brazil(>0.4) là có cao vướt bậc so với các quốc gia khác
* Các quốc gia khác đều có tỉ lệ TotalCases/TotalTests gần như bằng nhau (0.05-0.1)
* Điều này cho thấy tỉ lệ dương tính trên số ca xét nghiệm COVID-19 ở Mexico và Brazil rất cao, điều này nghĩa là do chính phủ không thể nắm rõ và có đủ khả năng xử lý dịch bệnh của 2 nước Mỹ Latinh này.
* Tỉ lệ TotalRecovered/TotalCases giữa các nước:



* Nhận xét:
* Tỉ lệ TotalRecovered/TotalCases của Pháp là rất thấp với các nước còn lại (<0.1)
* Các quốc gia còn lại đều có TotalRecovered/TotalCases khá là cao (~0.8)
* Điều này cho thấy việc kiểm soát địch bệnh của Pháp rất kém, bệnh nhân khi mắc Covid-19 ít có cơ hồi phục hồi.

Tuy nhiên có dấu hiện đáng mừng khi tỉ lệ phục hồi bệnh của các nước đều cao.

* Số lượng TotalDeaths,TotalRecovered, ActiveCases giữa các nước:



* Nhận xét:
* Mỹ là quốc gia có tổng số ca mắc Covid-19 rất cao, cao vượt bậc so với các quốc gia khác.
* Mỹ, Brazil, Mexico là các quốc gia có số ca tử vong cao nhất, 3 nước này cũng lần lượt 3 nước có số ca mắc Covid cao nhất thế giới.

Số lượng ca đang mắc Covid-19 tại thời điểm hiện tại của Mỹ và Pháp là cao nhất. Tuy nhiên, điều đáng lo ngại là Pháp là nước có tỉ lệ phục hồi thấp.

***Nguồn tham khảo***

https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-home.htm