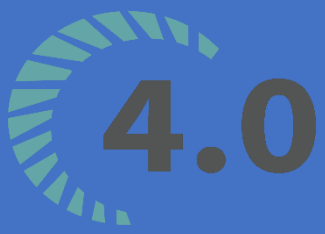
ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM

MÔN TRỰC QUAN HÓA DỮ LIỆU

LAB2: TRỰC QUAN HÓA DỮ LIỆU VỚI TABLEAU



GV hướng dẫn: Bùi Tiến Lên, Lê Ngọc Thành

TRỰC QUAN HÓA DỮ LIỆU

HỌC KỲ II – NĂM HỌC 2021-2022

BẢNG THÔNG TIN CHI TIẾT NHÓM

Số lượng:	**3**		
MSSV	**Họ tên**	**Email**	**Điện thoại**
19120559	Hà Duy Lãm	19120559@student.hcmus.edu.vn	0948311305
19120545	Lê Ngọc Khoa	19120545@student.hcmus.edu.vn	0337175835
19120677	Nguyễn Diệp Minh Tiến	19120677@student.hcmus.edu.vn	0939437078

Bảng phân công & đánh giá hoàn thành công việc			
Công việc thực hiện	**Người thực hiện**	**Mức độ hoàn thành**	**Đánh giá của nhóm**
Các kĩ thuật sử dụng và thuật toán học máy	Hà Duy Lãm	100%	10/10
Trực quan dữ liệu với tableau	Lê Ngọc Khoa	100%	10/10
Tìm hiểu tableau và làm báo cáo	Nguyễn Diệp Minh Tiến	100%	10/10

MỤC LỤC

I. Tìm hiểu Tableau

1. Giới thiệu Tableau

- Tableau là phần mềm hỗ trợ phân tích (analyze) và trực quan (Visualization) hóa dữ liệu, được dùng nhiều trong ngành công nghiệp kinh doanh thông minh - BI (Business Intelligence). Cũng giống như Excel, Tableau giúp tổng hợp các dữ liệu nhưng ở một cấp độ cao hơn khi chuyển những liệu này từ các dãy số thành những hình ảnh, biểu đồ trực quan một cách nhanh chóng và hiệu quả.
- Tableau là một giải pháp quản lý dữ liệu của doanh nghiệp, bạn không cần có kiến thức xử lý dữ liệu dưới dạng đồ họa, kiến thức lập trình, và chuyên môn tổng hợp vẫn có thể lấy dữ liệu cần thiết từ nguồn dữ liệu.
- Quá trình phân tích dữ liệu diễn ra nhanh chóng, không bị gián đoạn và cập nhật thường xuyên ngay khi có sự thay đổi dữ liệu gốc.
- Tableau hỗ trợ cả mọi đối tượng người dùng, bằng cách cài đặt trên PC, bạn có thể nhanh chóng xây dựng một môi trường phân tích dựa trên dữ liệu hoàn hảo, hiển thị đẹp mắt và trực quan.

2. Tầm quan trọng của Tableau

- Phân tích và trực quan hóa dữ liệu là những bước quan trọng trong ngành khoa học dữ liệu và đặc biệt là các ứng dụng của nó trong kinh doanh, có ý nghĩa trong việc theo dõi tình hình phát triển, lợi nhuận, doanh thu, phân tích đối thủ, đánh giá các chiến lược marketing. Việc thể hiện kết quả của các bảng phân tích này dưới dạng trực quan đem đến cái nhìn chi tiết, trực quan và dễ chịu hơn, không chỉ là đưa ra các phép so sánh, tổng hợp thông tin quan trọng mà còn giúp dự báo các xu hướng thị trường, hỗ trợ đưa ra các quyết định tối ưu nhất.
- Thực tế, việc tổng hợp dữ liệu thủ công là công việc rất khó khăn và tốn sức do phần lớn dữ liệu thô đều được lưu trữ dưới dạng bảng (excel sheets, google sheets,…) với vô số trường và bản ghi dữ liệu. Các công cụ như Tableau giúp tự động hóa việc phân tích dữ liệu với tốc độ rất nhanh chóng và trực quan, đáp ứng nhu cầu của các doanh nghiệp cần đưa ra đánh giá và quyết định kinh doanh với tốc độ cao.

3. Tableau có những tính năng gì?

- Tableau Dashboard

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

- Cộng tác và chia sẻ

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

- Dữ liệu trực tiếp và In-memory

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

- Kết nối đến nguồn cấp dữ liệu

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

Kết nối đến nguồn cấp dữ liệu Tableau

- Trực quan hóa nâng cao dưới dạng biểu đồ

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

- Trình diễn thông tin trên bản đồ

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

- Bảo mật mạnh

Hệ thống phân quyền và xác thực có sẵn giúp Tableau giảm thiểu nguy cơ mất mát dữ liệu. Ngoài ra, công cụ còn cho phép tự sử dụng các giao thức bảo mật khác từ môi trường desktop như Active Directory, Kerberos,…

- Xem trên thiết bị di động

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

- Truy vấn dữ liệu

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

- Phân tích, dự đoán

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

4. Tableau bao gồm những gì?

Các sản phẩm của Tableau

- Tableau Prep

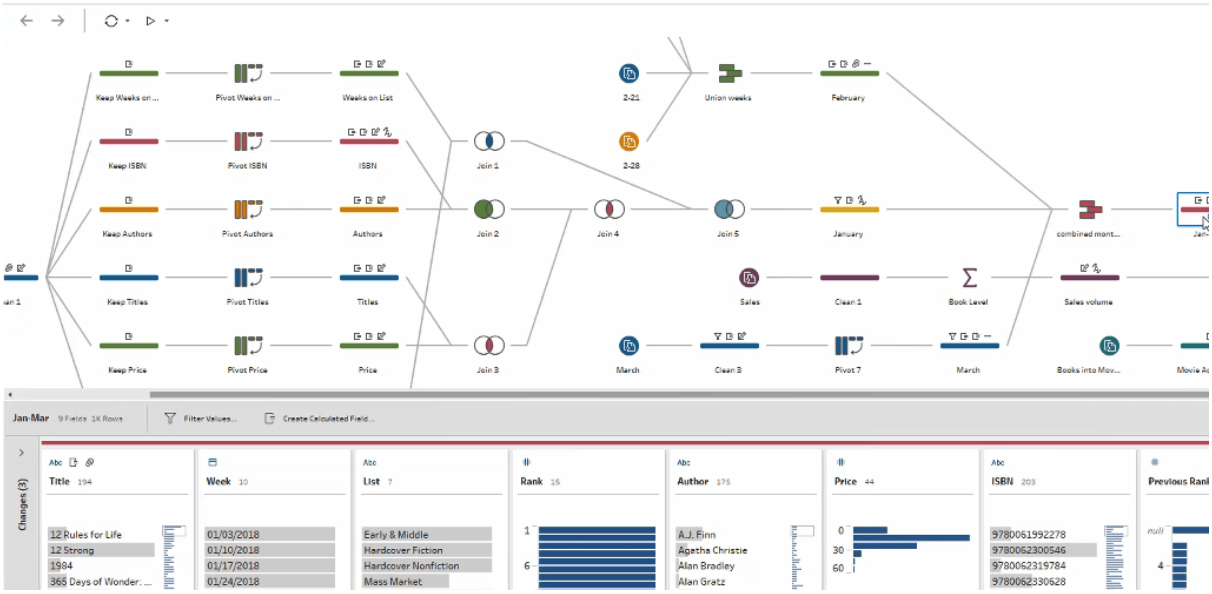


Tableau Prep là nơi để chuẩn bị dữ liệu

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

Cụ thể, ứng dụng giúp người dùng doanh nghiệp và nhà phân tích định hình dữ liệu nhanh chóng. Cho phép thực hiện các truy vấn, kết hợp và làm sạch dữ liệu cực kì đơn giản và tiện lợi. Sử dụng Tableau Prep giúp dữ liệu có tổ chức, rõ ràng, dễ quản lý hơn. Hiện có hai công cụ là Tableau Prep Builder để xây dựng luồng dữ liệu và Tableau Prep Conductor để quản lý các luồng.

- Tableau Desktop

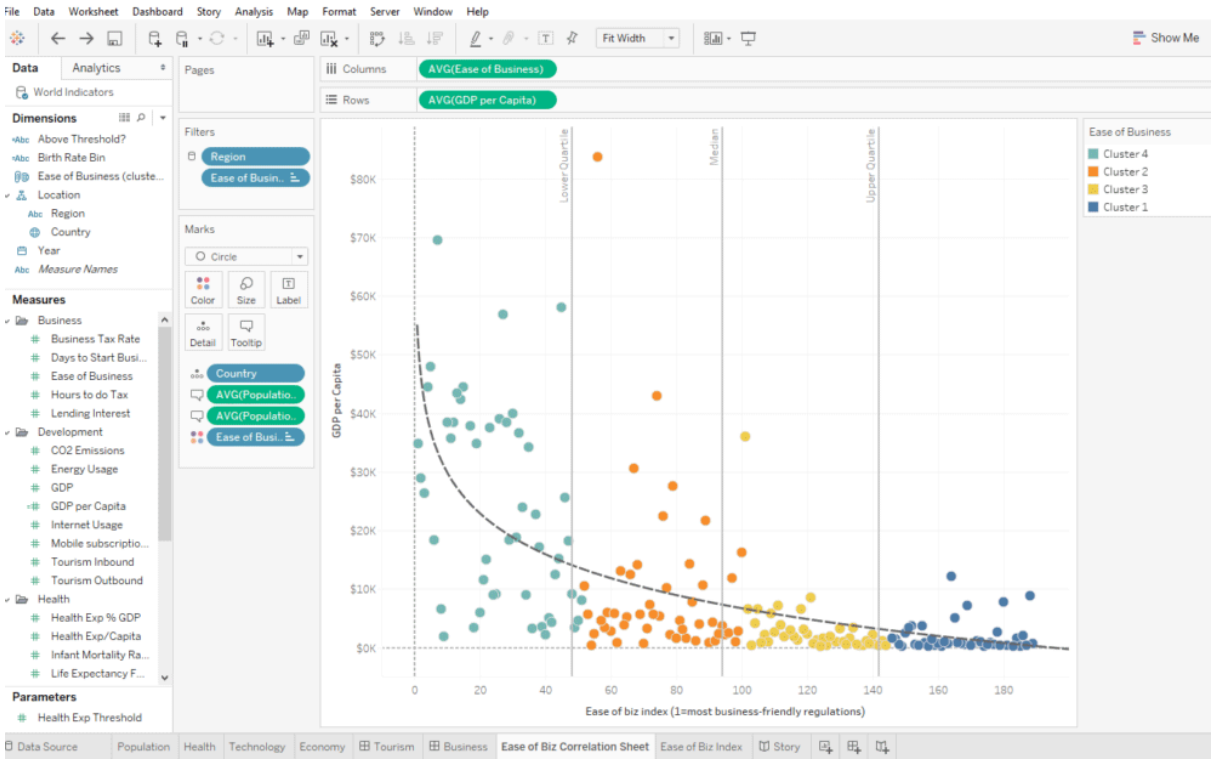


Tableau Desktop là nơi thực hiện các phân tích và trực quan dữ liệu

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

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

Tableau Desktop có thể chia thành 2 loại:

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

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

- Tableau Online

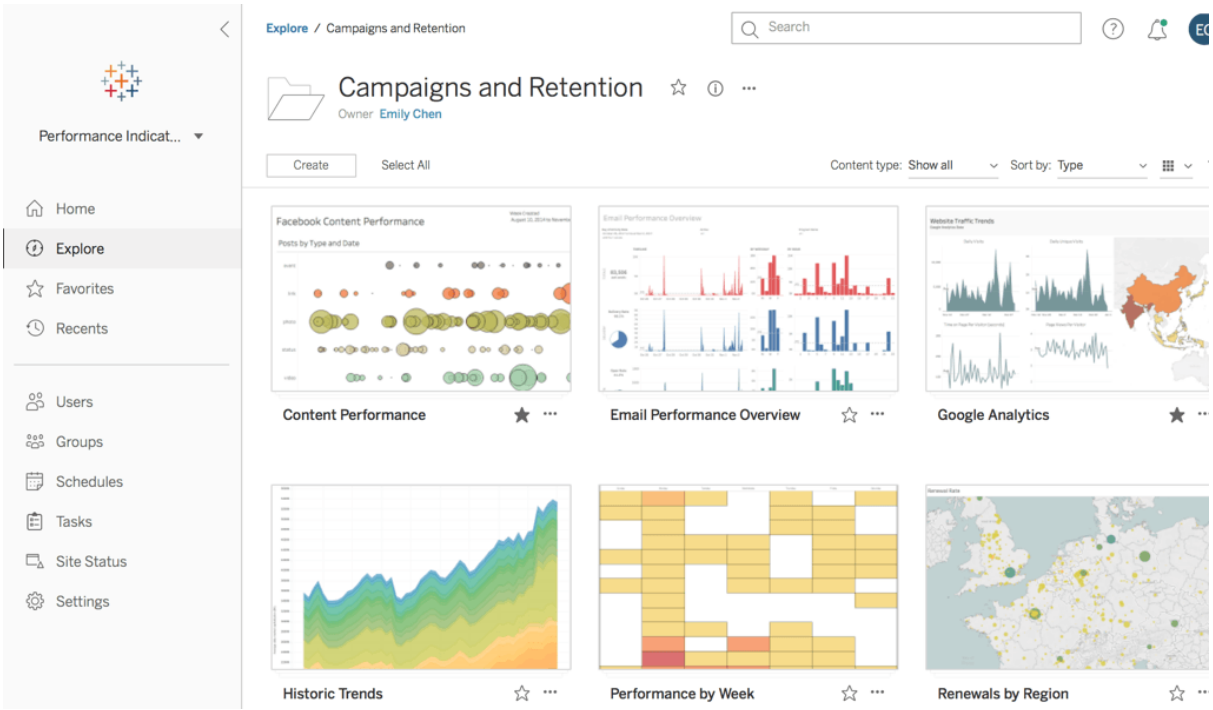


Tableau Online là dịch vụ hoàn toàn miễn phí

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

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

- Tableau Server

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

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

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

- Tableau Public

Nếu đồng ý chia sẻ những dữ liệu của mình, Tableau Public là phiên bản có mức giá hấp dẫn hơn cho người dùng. Dữ liệu phân tích sẽ không thể lưu trữ trên máy cá nhân mà sẽ được tải lên đám mây công cộng của Tableau mà bất cứ ai cũng có quyền truy cập.

- Tableau Reader

Với những người chỉ có nhu cầu xem các báo cáo mà không trực tiếp tham gia phân tích xử lý thì Tableau Reader chính là công cụ dành cho họ. Với việc lược bỏ các tính năng phức tạp, Tableau Reader là một công cụ nhẹ nhàng nhưng vẫn đảm bảo hiển thị được tất cả định dạng báo cáo của Tableau.

5. Vì sao nên sử dụng Tableau Data Visualization?

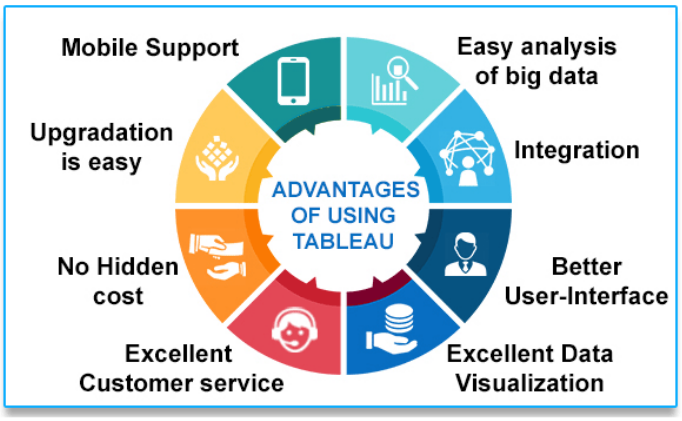


Tableau nhiều năm liên tục thuộc TOP công cụ phân tích phổ biến

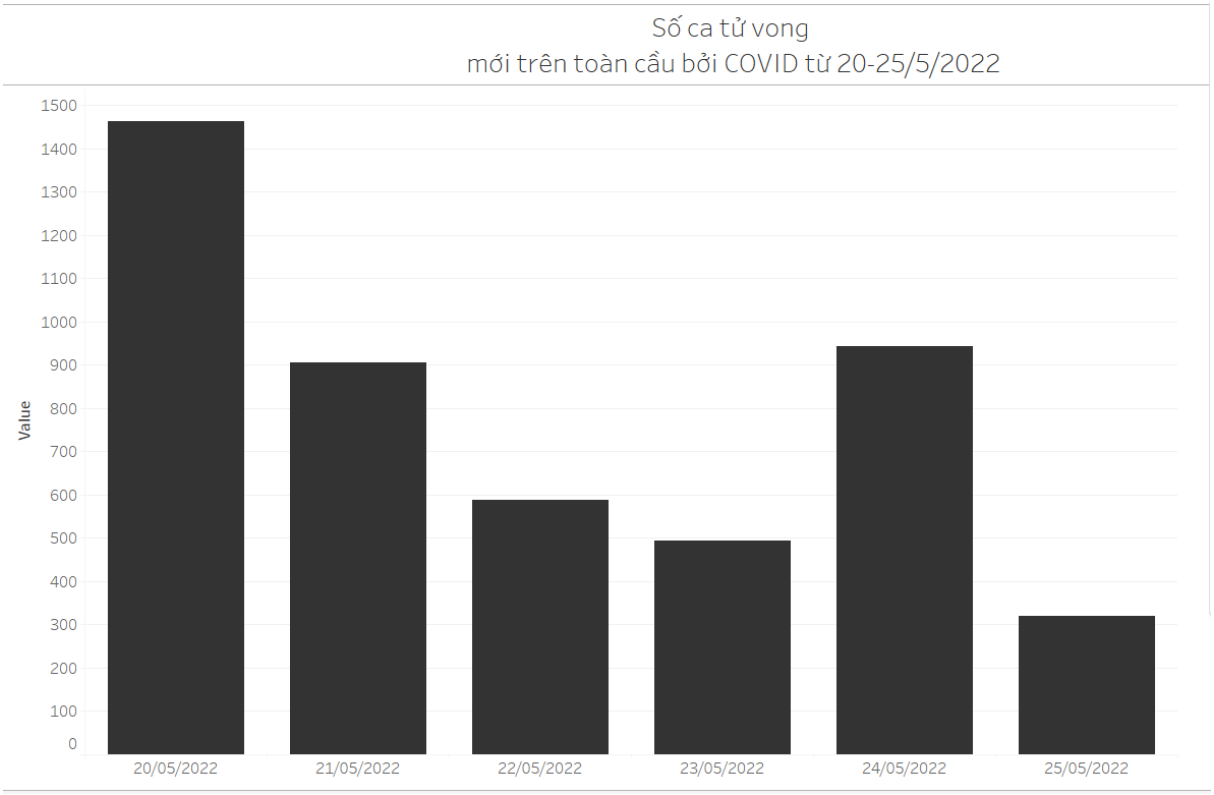
Có thể thấy Tableau mang đến nhiều ứng dụng phục vụ nhu cầu phân tích dữ liệu. Từ khâu chuẩn bị, tổ chức đến phân tích, khám phá, trực quan, chia sẻ… Hơn thế nữa, tất cả những thao tác này phần lớn đều thực hiện bằng thao tác kéo thả đơn giản.

Nếu so sánh giữa Excel và Tableau Data Visualization có thể thấy rất nhiều cải tiến. Đây có lẽ cũng là phần nào lý do mà ngày nay những công cụ phân tích và trực quan dữ liệu như Tableau được khuyến khích sử dụng hơn so với Excel.

II. Trực quan dữ liệu với Tableau

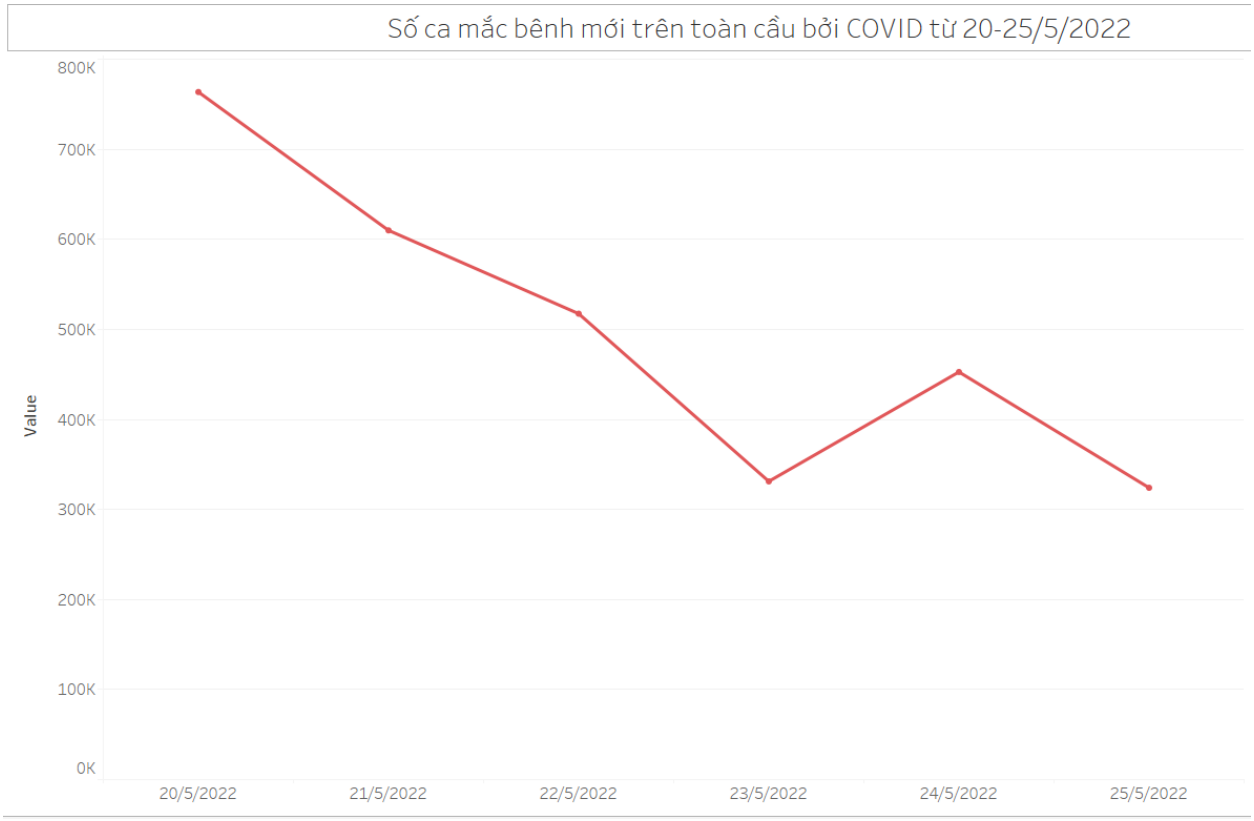
1. Thể hiện trực quan một số dữ liệu biến đổi qua từng ngày và ý nghĩa của chúng:

a. Hozizontal bars:



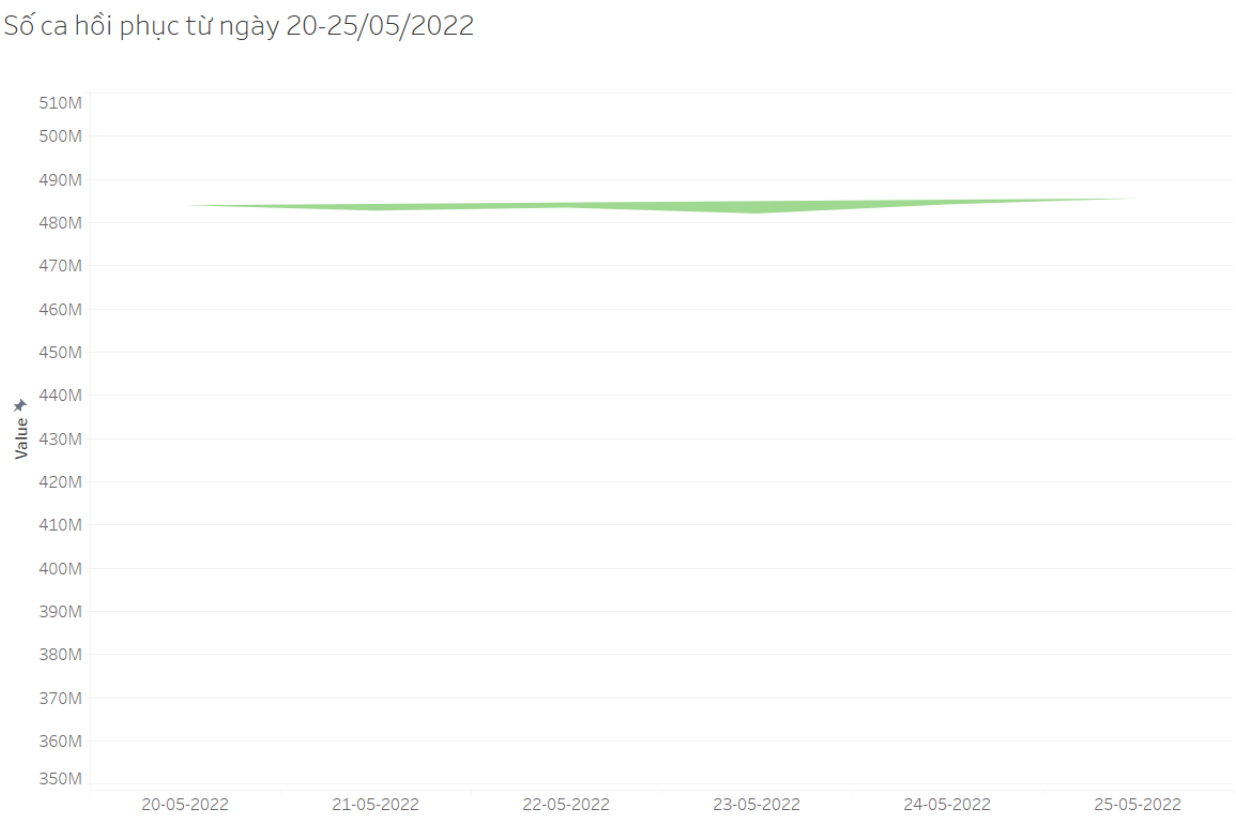
- Để trực quan biểu đồ trên, nhóm lựa chọn trường dữ liệu là New Deaths từ ngày 20-25/05/2022 để thể tình hình của các ca tử vong qua từng ngày của dịch bệnh trên toàn cầu.
- Qua biểu đồ này ta có thể nhìn thấy một cách rõ ràng về sự viến động số lượng các ca tử vong. Khi rê chuột vào vị trí các cột, ta có thể xem chi tiết thông số của số lượng ca tử vong của ngày.
- **Nhận xét về biểu đồ:** Từ biểu đồ trên ta có thể thấy số lượng ca tử vong mới có xu hướng giảm mạnh và liên tục. Tuy nhiên từ cột ngày 24/05/2022 ta có thể thấy được sự bất thường trong số ca tử vong của tường ngày, dịch vẫn đang diễn biến phức tạp ở một số nơi. Chúng ta không nên chủ quan.
- Việc lựa chọn màu đen để trực quan vì đây là trường dữ liệu mang lại cảm giác tiêu cực, khi nhìn vào dữ liệu sẽ dễ cho thấy đúng tính chất của dữ liệu.

b. Lines chart:



- Để trực quan biểu đồ trên, nhóm lựa chọn trường dữ liệu là New Cases từ ngày 20 – 25/05/2022 để thể hiện mức độ tăng trưởng của dịch bệnh trên toàn cầu.
- Qua biểu đồ này ta có thể dễ dàng quan sát được sự tăng hay giảm của số lượng ca mắc mới qua từng ngày trong giai đoạn 20 – 25/05/2022. Khi rê chuột vào từng đoạn của lines ta có thể thấy chi tiết về số lượng ca mắc mới cụ thể của từng ngày.
- **Nhận xét về biểu đồ:** Qua biểu đồ này ta có thể thấy số lượng ca mắc mới đang có xu hướng giảm mạnh từ ngày 20 – 23, nhưng có sự tăng bất thường vào ngày 24. Vì vậy ta có thể thấy tình hình dịch bệnh vẫn còn khá phức tạp nên chúng ta không nên chủ quan.
- Khi chúng ta so sánh biểu đồ này với biểu đồ Hozizontal bars ở trên chúng ta có thể thấy số ca mắc mới và số ca tử vong có xu hướng khá giống nhau, đều giảm mạnh và cũng có điểm bất thường vào ngày 24/05/2022.
- Ở đây nhóm lựa chọn màu đỏ cho màu sắc thể hiện của line để thể hiện sự cảnh báo đáng quan tâm và lưu ý nhất của việc kiểm soát dịch bệnh.

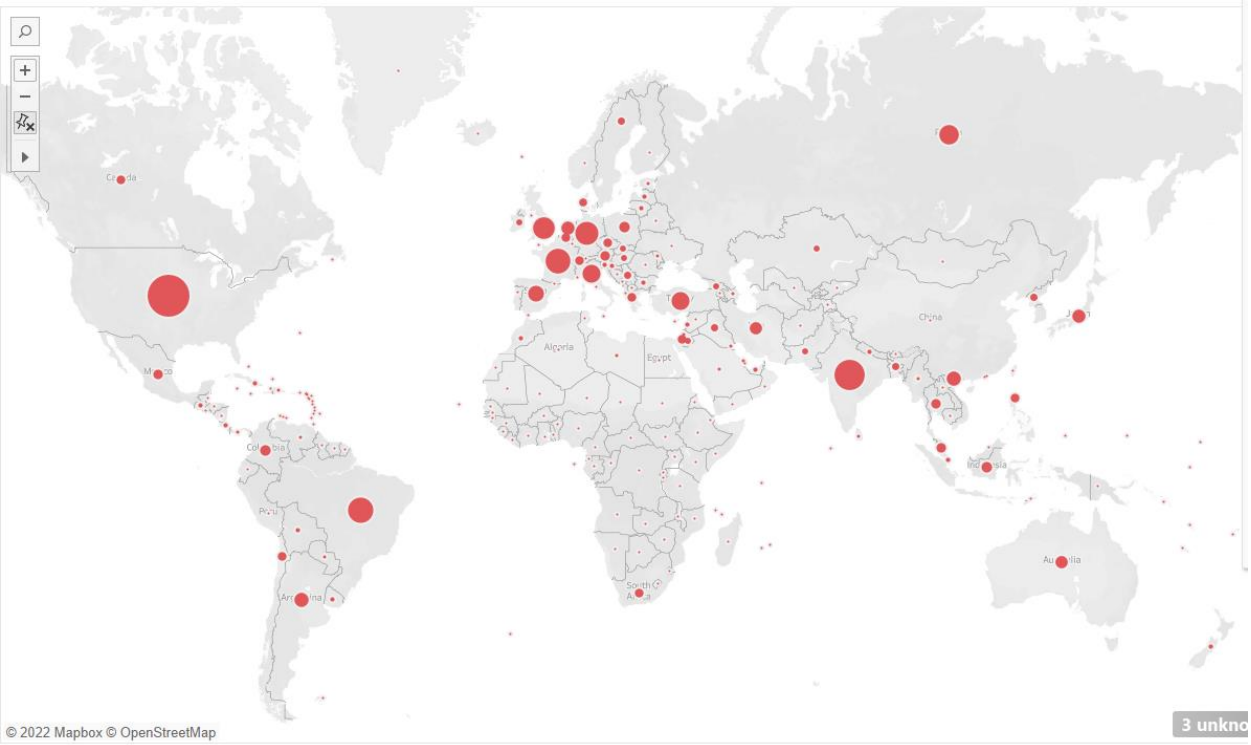
c. Polygon:



- Để trực quan biểu đồ trên, nhóm lựa chọn trường dữ liệu là Total Recoverds từ ngày 20-25/05/2022 để thể hiện tình hình hồi phục dịch bệnh trên toàn cầu.
- Ở đây ta chỉ xét đến giá trị của sự phục hồi qua từng ngày nên việc sử dụng biểu đồ này cũng sẽ đáp ứng được nhu cầu của chúng ta thông qua độ lớn của vùng hiển thị. Ta cũng có thể rê chuột vào vùng hiển thị để biết chính xác hơn về con số của sự phục hồi.
- **Nhận xét về biểu đồ:** Như ta thấy trên biểu đồ thì tình hình hồi phục trong giai đoạn từ ngày 20-25/05/2022 đều có sự tương đồng, riêng giai đoạn ngày 23 thì chỉ số hồi phục tăng vượt trội so với các ngày còn lại trong giai đoạn thông qua việc độ to của cùng hiển thị bị phình ra.
- Việc lựa chọn màu xanh để hiển thị cho tình hình phục hồi trong cơn dịch bệnh sẽ mang lại cảm giác lạc quan và tích cực cho người xem biểu đồ. Tông màu mát, dịu sẽ đem lại cảm giác thoải mái, dễ chịu cho người xem.

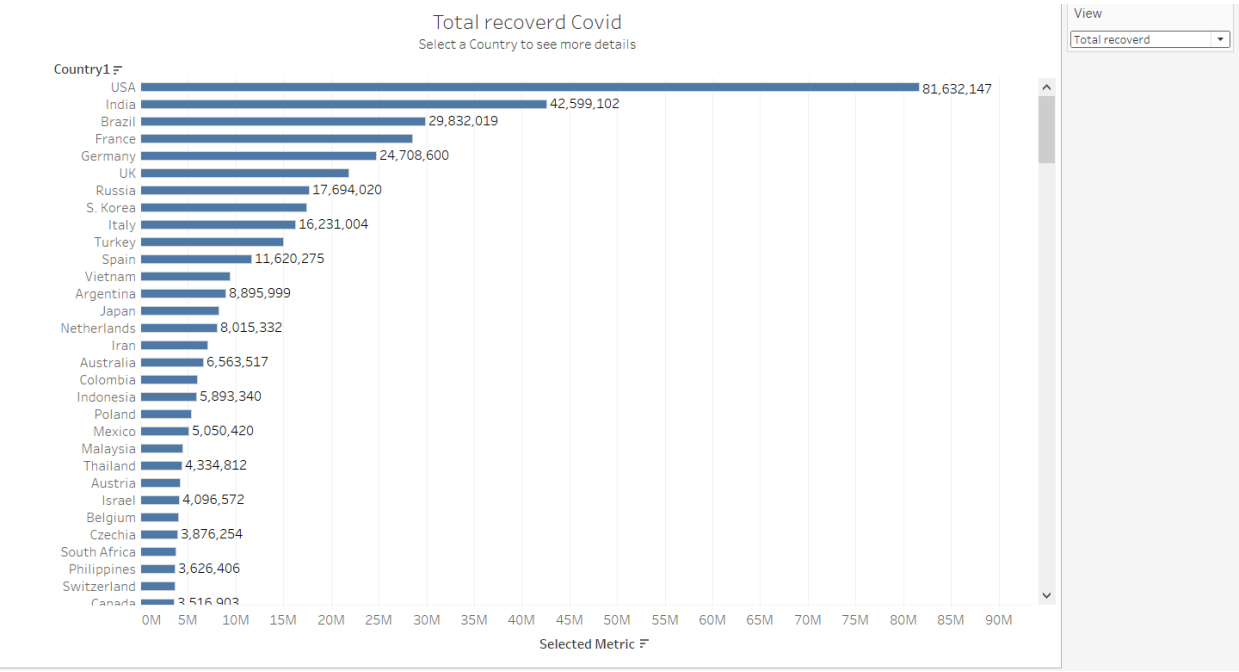
2. Trực quan một số loại biểu đồ với Tableau:

a. Symbol Maps:



- Để trực quan biểu đồ trên, nhóm lựa chọn 2 trường dữ liệu là Country và New Case để thể hiện thông tin số lượng ca nhiễm mới của tất cả các quốc gia trên thế giới.
- Nhìn qua biểu đồ ta có thể dễ dàng nhận ra vùng lãnh thổ nào đang có tình hình dịch bệnh diễn biến phức tạp qua kích cỡ của các chấm tròn và mật độ của chúng.
- Chẳng hạn các chấm tròn ở trên ta có thể thấy, Mỹ, Nga, Ấn Độ, Brazil đang là những quốc qua có lượng ca nhiễm tăng mạnh nhất, các chấm tròn phân bố không tập trung và không đồng đều, cho ta thấy tình hình lây lan dịch bệnh còn khá nghiêm trọng.
- Ngoài ra biểu đồ này còn có tính năng tìm kiếm tên quốc gia bằng các nhập tên quốc gia đó vào mục tìm kiếm với icon là cái kính lúp.
- Ở đây nhóm lựa chọn màu đỏ cho màu sắc thể hiện của line để thể hiện sự cảnh báo đáng quan tâm và lưu ý nhất của việc kiểm soát dịch bệnh.

b. Bar chart:



- Với loại biểu đồ này ta sẽ so sánh được chi tiết dữ liệu giữa các nước khi label chi tiết của từng nước sẽ hiện lên ngay bên cạnh cột dữ liệu của mình
- Như hình trên ta có thể thấy tổng ca hồi phục của Mỹ là cao nhất với 81.632.147 ca.
- Phía bên phải có bộ lọc, chúng ta cũng có thể lựa chọn các trường mình quan tâm để xem các dữ liệu chúng ta cần, biểu đồ theo các trường dữ liệu:

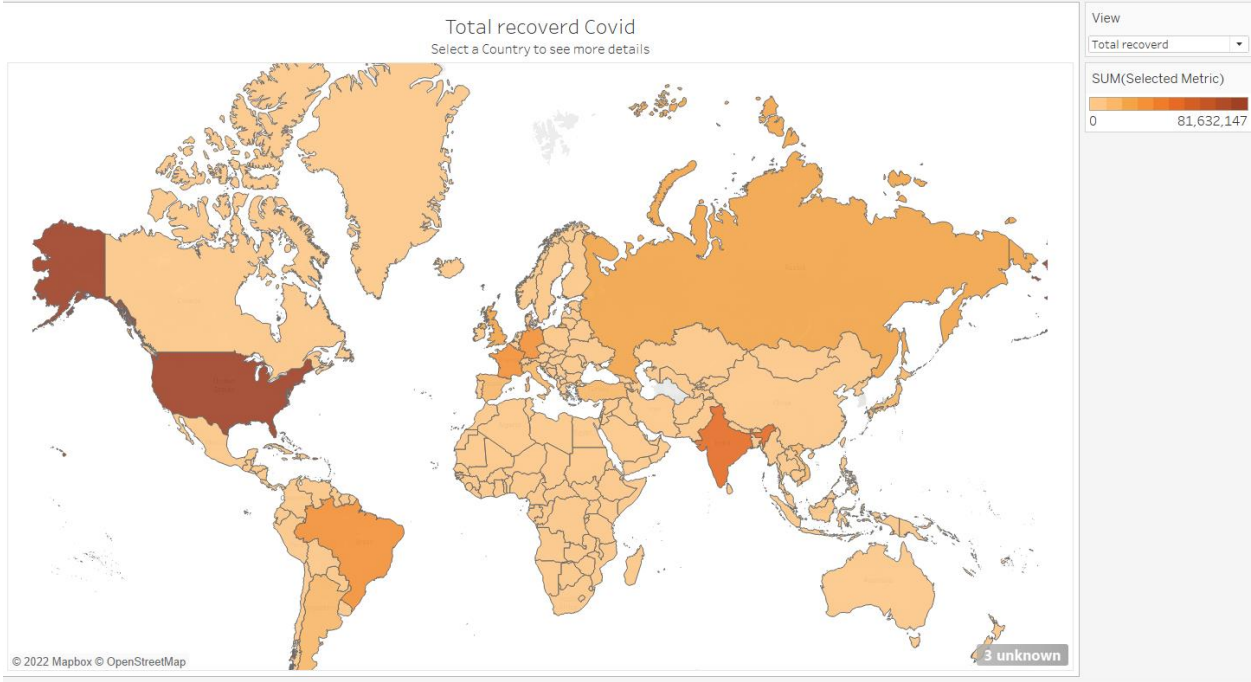
  + Total recoverd

  + Total death

  + Active case

  + Total case

c. Global Heatmap:



- Với biểu đồ trên, ta có thể quan sát được số lượng ca nhiễm ứng với từng quốc gia bằng cách di chuyển chuột vào vùng quốc gia tương ứng.
- Qua biểu đồ trên ta có thể nhìn được một cách khái quát toàn cảnh tình trạng nhiễm bệnh mới của các quốc gia trên toàn thế giới. Biết được quốc gia nào vẫn đang trong tầm nguy hiểm, quốc gia nào đã không còn ca nhiễm mới. Từ đó, ta có thể kết luận được tình trạng dịch bệnh diễn biến trên toàn cầu như thế nào.
- Phía bên phải có bộ lọc, chúng ta cũng có thể lựa chọn các trường mình quan tâm để xem các dữ liệu chúng ta cần, biểu đồ theo các trường dữ liệu:

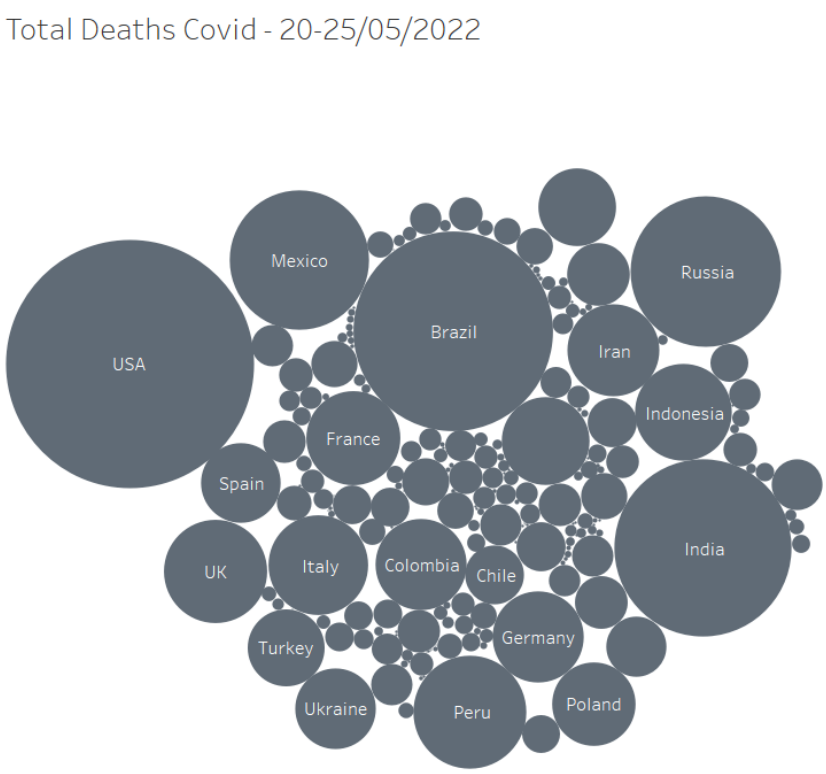
  + Total recoverd

  + Total death

  + Active case

  + Total case

d. Packed Bubbles:

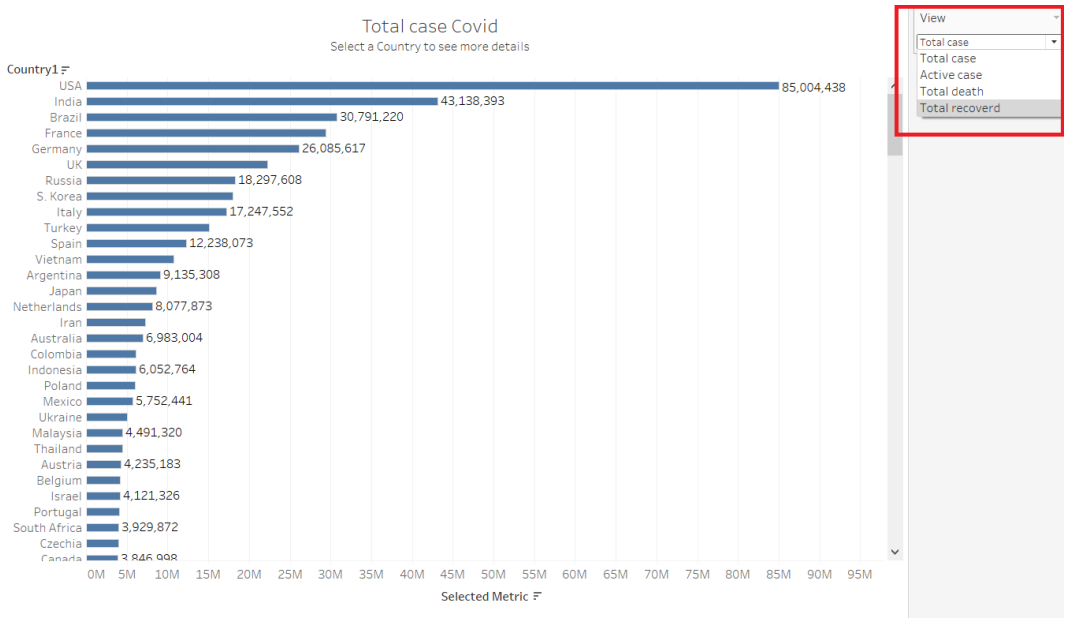


- Để trực quan biểu đồ trên, nhóm lựa chọn hai trường dữ liệu là Country, Total Deaths để thể mức độ biểu hiện tiêu cực dịch bệnh trên toàn cầu.
- Qua đây, ta dễ dàng thấy các quốc gia mà đang chịu ảnh hưởng tiêu cực nặng nề nhất của dịch bên thông qua kích cỡ của các hình tròn, khi lê chuột vào các hình tròn thì nó sẽ hiển thị lên chỉ số của tổng ca mất bệnh và tổng ca chết kèm theo tên quốc gia mà hình tròn đó hiển thị.
- Qua biểu đồ ta thấy Mỹ, Brazil, Ấn Độ, Nga đang là các quốc gia có tình hình dịch bệnh nghiêm trọng nhất trên thế giới.
- Việc lựa chọn mày xám cho biểu đồ này để thể hiện sự tiêu cực, không sáng sủa của tình hình dịch bệnh đang diễn ra trên toàn cầu.
- Việc lựa chọn màu đen để trực quan vì đây là trường dữ liệu mang lại cảm giác tiêu cực, khi nhìn vào dữ liệu sẽ dễ cho thấy đúng tính chất của dữ liệu.

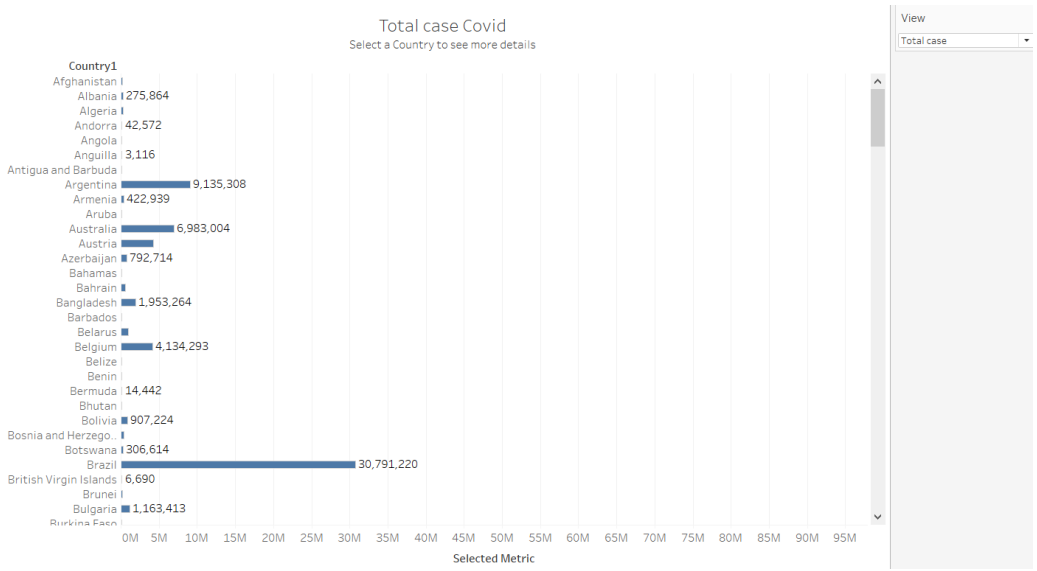
III. Sử dụng các kỹ thuật được giới thiệu trong bài Manipulate View, Facet, Reduce, Embed để trình diễn trên Tableau với dữ liệu Woldometer:

1. Manipulate View:

- **Lý do lựa chọn:** Do dữ liệu quá nhiều để có thể visualization. Do đó đây là một khả năng hữu ích để có thể chọn ra những gì thích hợp nhất để hiển thị tùy thuộc vào nhu cầu của người sử dụng.

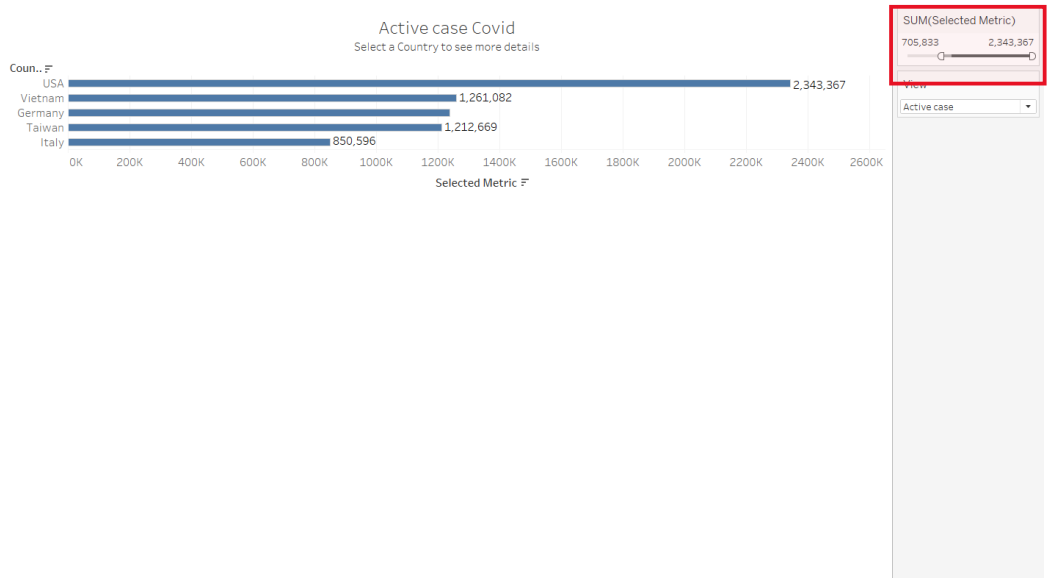


- **Ý nghĩa mang lại:** Ta có thể chọn các kiểu xem phù hợp theo nhu cầu ví dụ ta muốn xem tổng số ca mắc thì chọn total case, tổng số ca mất thì chọn total death,… ngoài ra ta còn chọn được kiểu xem của biểu đồ bar chart trên được sắp xếp theo từ lớn đến bé hoặc theo bảng chữ cái.



2. Reduce:

- Một trong những kỹ thuật để Reduce là Filter. Trong tableau có hỗ trợ Filters để giả số lượng items cũng như lựa chọn các atribute mà người dùng mong muốn.
- **Lý do lựa chọn:** Những chức năng này khá dễ dàng và phù hợp với người mới bắt đầu, thao tác nhanh nhưng vẫn đem lại hiệu quả hiển thị như mong muốn



- **Ý nghĩa mang lại:** Chẳng hạn khi ta muốn xem những quốc gia có lượng khỏi bệnh trên 700 000 thì ta kéo thanh trượt đến 700 00. Từ đó ta thấy được các quốc gia khắc phục tình hình dịch bệnh tốt.

VI. Áp dụng một số thuật toán máy học:

1. Bình phương nhỏ nhất thông thường (Ordinary Least Squares - OLS):

- Đây là phương pháp được sử dụng rộng rãi nhất để ước lượng các tham số trong phương trình hồi quy. Để tối thiểu hoá tổng bình phương của các khoảng cách theo phương thẳng đứng giữa số liệu thu thập được và đường (hay mặt) hồi quy

```
from pandas import DataFrame
import statsmodels.api as sm

df = DataFrame(dataset,columns=['New Cases','Total Deaths','New Deaths','Total Recoverd','Active Cases','Serious Cases'])
X = df[['New Cases','Active Cases']]
Y = df['Serious Cases']
X = sm.add_constant(X)
model = sm.OLS(Y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Serious Cases   R-squared:                       0.094
Model:                            OLS   Adj. R-squared:                  0.086
Method:                 Least Squares   F-statistic:                     11.67
Date:                Sat, 28 May 2022   Prob (F-statistic):           1.51e-05
Time:                        11:50:20   Log-Likelihood:                -1818.0
No. Observations:                 229   AIC:                             3642.
Df Residuals:                     226   BIC:                             3652.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        107.0680     46.868      2.284      0.023      14.714     199.422
New Cases     -0.0033      0.004     -0.823      0.411      -0.011       0.005
Active Cases   0.0010      0.000      4.570      0.000       0.001       0.001
==============================================================================
Omnibus:                      384.058   Durbin-Watson:                   1.791
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            77713.907
Skew:                           8.594   Prob(JB):                         0.00
Kurtosis:                      91.596   Cond. No.                     2.53e+05
==============================================================================
```

- Mục tiêu ở đây là tìm mối quan hệ phụ thuộc Serious Cases và hai biến New Cases và Active Cases
- Một số thanh phần quan trọng trong kết quả:
    - **Adjusted R-squared:** Phản ánh sự phù hợp của mô hình. Các giá trị bình phương R nằm trong khoảng từ 0 đến 1, trong đó giá trị cao hơn thường biểu thị mức độ phù hợp tốt hơn, giả sử các điều kiện nhất định được đáp ứng.
    - **Const coefficient:** Là Y-intercept. Điều đó có nghĩa là nếu cả hai hệ số New Case và Active Case đều bằng 0, thì sản lượng dự kiến (nghĩa là, Y) sẽ bằng với hệ số const
    - **New Case coefficient**: biểu thị sự thay đổi của đầu ra Y do thay đổi một đơn vị New Case (mọi thứ khác được giữ cố định)

- **Active Case coefficient**: Đại diện cho sự thay đổi của sản lượng Y do sự thay đổi của một đơn vị trong tỷ lệ thất nghiệp (mọi thứ khác được giữ cố định).
- **std err:** phản ánh mức độ chính xác của các hệ số. Nó càng thấp, mức độ chính xác càng cao
- **P >|t|:** Là giá trị p-value. Giá trị p nhỏ hơn 0,05 được coi là có ý nghĩa thống kê
- **Confidence Interval:** Đại diện cho phạm vi mà hệ số của chúng tôi có khả năng giảm (với khả năng là 95%).

2. Phép phân tích thành phần chính (PCA)

- Phép phân tích thành phần chính là một thuật toán thống kê sử dụng phép biến đổi trực giao để biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn nhằm tối ưu hóa việc thể hiện sự biến thiên của dữ liệu.

PCA

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df = DataFrame(dataset,columns=['New Cases','Total Deaths','New Deaths','Total Recoverd','Active Cases','Serious Cases'])
scaler.fit(df)
scaled_data = scaler.transform(df)
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
pca.fit(scaled_data)
x_pca = pca.transform(scaled_data)
print(scaled_data.shape)
print(x_pca.shape)
```

```
(229, 6)
(229, 2)
```

- Đầu tiên, chúng ta cần xử lý trước dữ liệu, tức là chúng ta cần chia tỷ lệ dữ liệu sao cho mỗi tính năng có phương sai đơn vị và không có tác động lớn hơn dữ liệu kia.
- Chúng ta sẽ chỉ định số lượng thành phần là 2. Sau đó chúng ta có thể chuyển đổi dữ liệu này thành 2 thành phần chính đầu tiên.
- Khi kiểm tra kích thước của dữ liệu trước và sau PCA ta thấy nó giảm từ 6 thành 2

3. Hồi quy tuyến tính (Linear regression)

- Phân tích hồi quy tuyến tính là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X. Mô hình hóa sử dụng hàm tuyến tính. Các tham số của mô hình được ước lượng từ dữ liệu.

```
import seaborn as seabornInstance
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
%matplotlib inline
X = dataset['Active Cases'].values.reshape(-1,1)
y = dataset['Serious Cases'].values.reshape(-1,1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
regressor = LinearRegression()
regressor.fit(X_train, y_train)
print(regressor.intercept_)
print(regressor.coef_)
y_pred = regressor.predict(X_test)
df = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': y_pred.flatten()})
df
```

```
[112.99859279]
[[0.00096412]]
```

	Actual	Predicted
0	6	116.422199
1	0	112.998593
2	0	113.003413
3	4	123.364859
4	0	112.998593
5	0	114.017672
6	0	113.154781
7	73	119.295290
8	0	112.998593
9	0	116.388454
10	6	113.931865
11	0	112.998593
12	0	112.998593
13	291	916.159599
14	0	113.014983
15	0	113.987785
16	2	113.212628
17	188	329.369337
18	30	116.183096

- Mô hình hồi quy tuyến tính về cơ bản tìm thấy giá trị tốt nhất cho phần chặn và độ dốc, dẫn đến một dòng phù hợp nhất với dữ liệu. Để xem giá trị của phần chặn và độ dốc được tính toán bằng thuật toán hồi quy tuyến tính cho tập dữ liệu.
- Kết quả phải là khoảng 112.9985 và 0.0009 tương ứng. Điều này có nghĩa là cứ một đơn vị thay đổi Active Case, sự thay đổi Serious Case là khoảng 0,0009%
- Sau đó đưa ra dự đoán về dữ liệu thử nghiệm, so sánh các giá trị đầu ra thực tế cho X_test với các giá trị dự đoán

VII. Tài liệu tham khảo:

- Linear Regression in Python using Statsmodels - Data to Fish
- Principal Component Analysis (PCA) with Python | DataScience+ (datascienceplus.com)
- TamaraChp11-manipulate.pdf (ohio-state.edu)