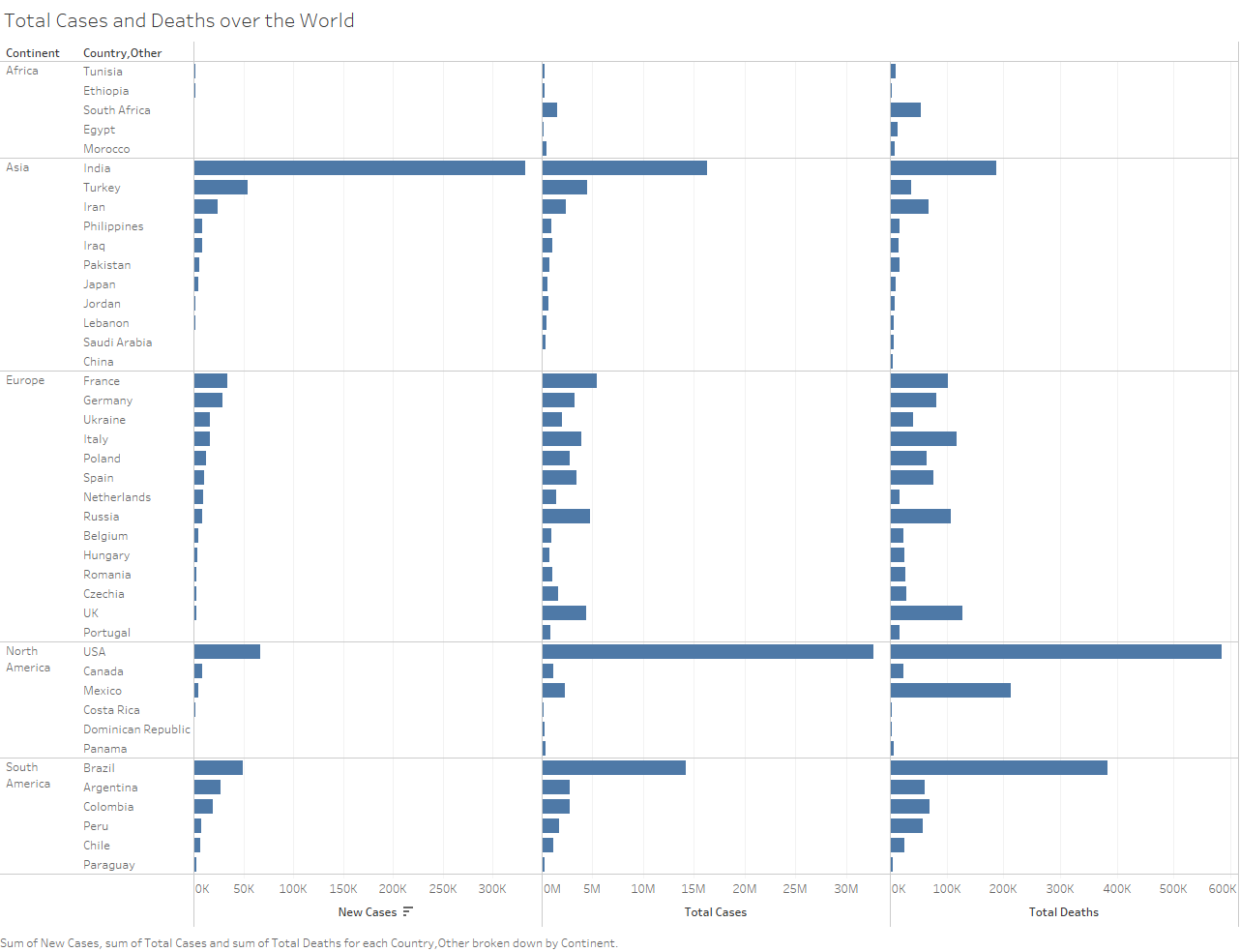
*Sử dụng các kỹ thuật được giới thiệu trong bài Manipulate View, Facet, Reduce, Embed để trình diễn trên Tableau với dữ liệu Woldometer. Giải thích việc chọn lựa và ý nghĩa mang lại.*

1. **Facet.**
   1. **Partition into Views.**

*Giải thích:* Chia nhỏ các biến dữ liệu trên nhiều ô con riêng lẻ và kết hợp các ô con đó thành một hình duy nhất.

*Ý nghĩa:* Giúp quan sát được nhiều biểu đồ hơn được sắp xếp trong một lưới duy nhất. Tách biệt các ô con giúp việc quan sát được rõ ràng, từng ô con đơn giản để theo dõi.

*Dữ liệu chọn lựa để trình diễn:* tổng số ca mắc mới, tổng số ca mắc và tổng ca tử vong theo lục địa của một số nước có ca mắc mới cao trong ngày 22-04-2021.



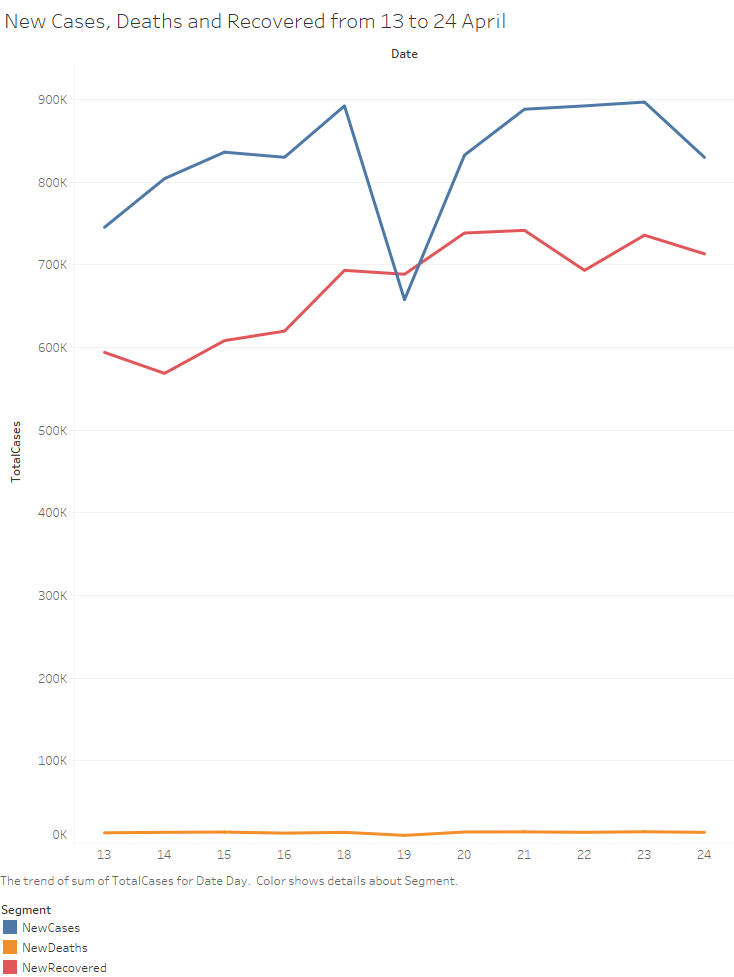
Hình . Số ca mới, Tổng số ca và Số ca chết qua từng nước trong các châu lục.

* 1. **Superimpose Layer**

Giải thích: Thuộc nhánh các kĩ thuật Embed. Là kết hợp các biến dữ liệu trên nhiều ô con riêng lẻ và chồng các ô con lên nhau thành các layer của một hình duy nhất.

*Ý nghĩa:* Tùy vào mục đích diễn đạt dữ liệu mà ta sử dụng phương pháp này. Nó sẽ giúp ta có nhiều thông tin hơn trên 1 biểu đồ. Tuy nhiên, với những biểu đồ có nhiều phần tử, phương pháp cũng đem lại sự phức tạp nhất định.

*Dữ liệu chọn lựa để trình diễn:* tổng số ca mắc mới, số ca phục hồi mới, số ca tử vong mới của toàn thế giới từ ngày 13 đến 24-04-2021.



Hình . Số ca mới, Số ca tử vong mới và Số ca hồi phục mới từ 13 đến 24 tháng 4.

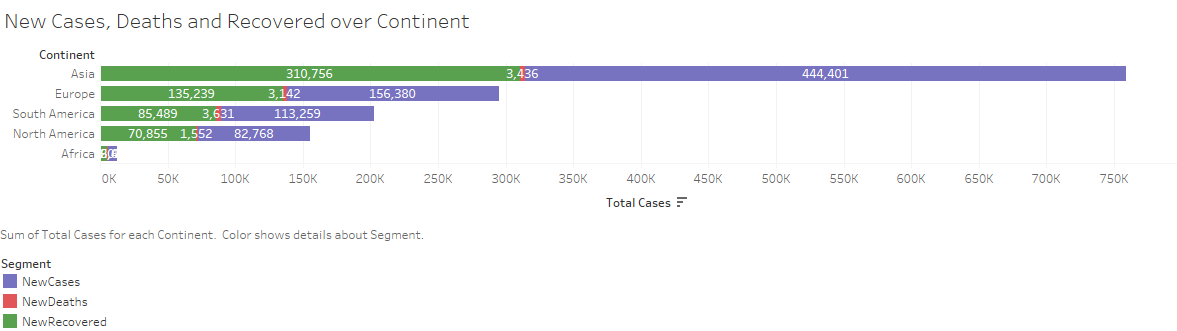
1. **Manipulate View.** 
   1. **Change View over Time**.

*Giải thích:* Bao gồm 2 kĩ thuật quan trọng là *LineUp* và *Animated Transistion.*

* *LineUp:* Sử dụng stacked bar để trực quan việc so sánh các thuộc tính thành phần.
* *Animated Transistion:* hiển thị hoạt ảnh 2D khi người xem tương tác bằng cách sắp xếp hoặc chọn lựa biến của stacked bar để thể hiện.

*Ý nghĩa:*

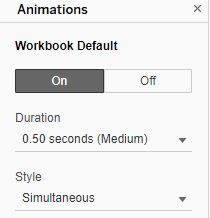
* Sử dụng *LineUp* giúp cho việc trực quan sẽ tổng quát, nhấn mạnh vào tổng số. Kết hợp với *Animated Transistion* sẽ giúp dễ so sánh sự thay đổi giữa các trường dữ liệu.
* Sử dụng *Animated Transistion* sẽ giúp ta thấy được sự liền mạch trong quá trình thay đổi thứ hạng khi ta chọn sắp xếp tăng dần hoặc giảm dần.



Hình . Số ca mới, Số ca hồi phục mới và Số ca chết mới theo châu lục.

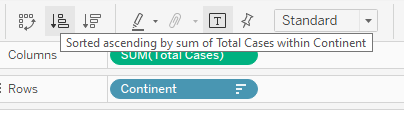
*Dữ liệu chọn lựa để trình diễn:* tổng số ca mắc mới, số ca phục hồi mới, số ca tử vong mới theo châu lục trong ngày 22-04-2021.

Để bật tính năng *Animated Transistion* trong Tableau, ta vào Format > Animations và bật On và chỉnh Duration ở mức Medium hoặc Slow để quan sát dễ hơn.



Hình . Bật animation trong Tableau.

Sau đó trên thanh menu ta chọn sắp xếp tăng dần hoặc giảm dần theo thuộc tính. Ta sẽ quan sát được chuyển động thay đổi giữa các thứ hạng.



Hình . Sắp xếp thuộc tính tăng/giảm trong Tableau.

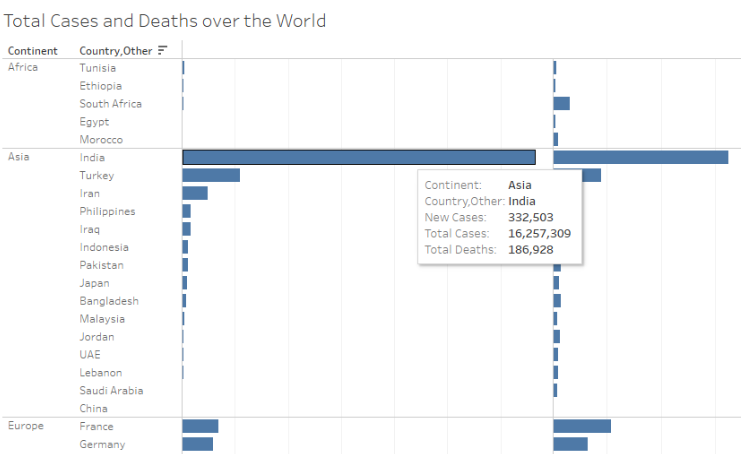
* 1. **Select Elements**.

*Giải thích:* Cho phép người dùng chọn một hoặc nhiều yếu tố quan tâm trong biểu đồ trực quan để hiển thị chi tiết. Kĩ thuật hay được dùng là *Highlighting* lên các yếu tố được chọn.

*Ý nghĩa:* giúp làm nổi bật, chú trọng vào biến được quan tâm trong biểu đồ trực quan.

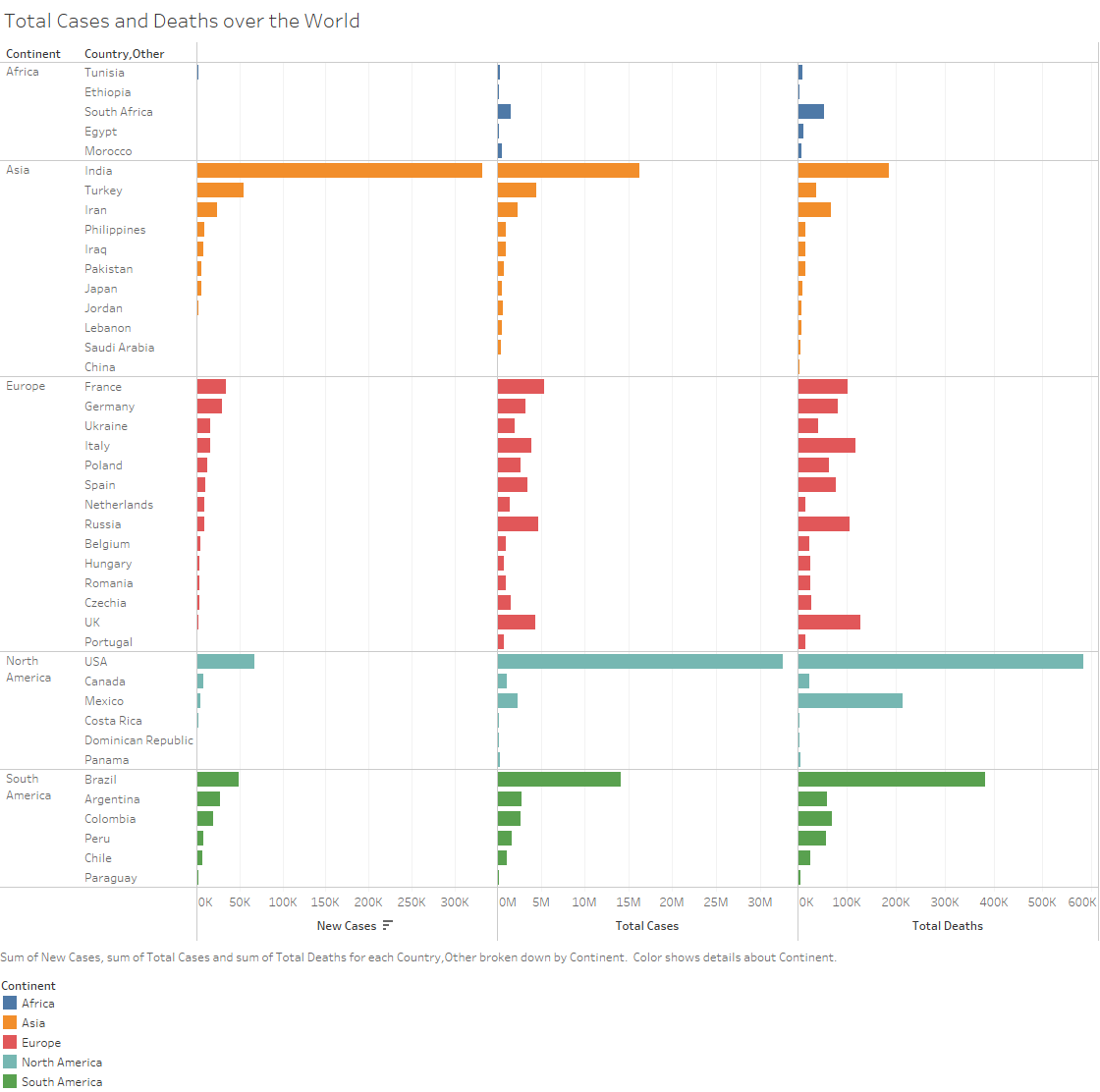
*Dữ liệu chọn lựa để trình diễn:* Sử dụng lại biểu đồ đã vẽ ở phần *Facet: Partition into Views*.

Trong Tableau, ta có thể hover lên 1 phần tử để xem thông tin chi tiết các thuộc tính khác của phần tử đó.



Hình . Hover lên item để xem chi tiết

Tùy vào nội dung muốn truyền tải. Ta có thể *Highlighting* dữ liệu bằng cách khác nhau. Phổ biến nhất là *Highlighting* bằng màu sắc, dải màu hoặc kích thước.

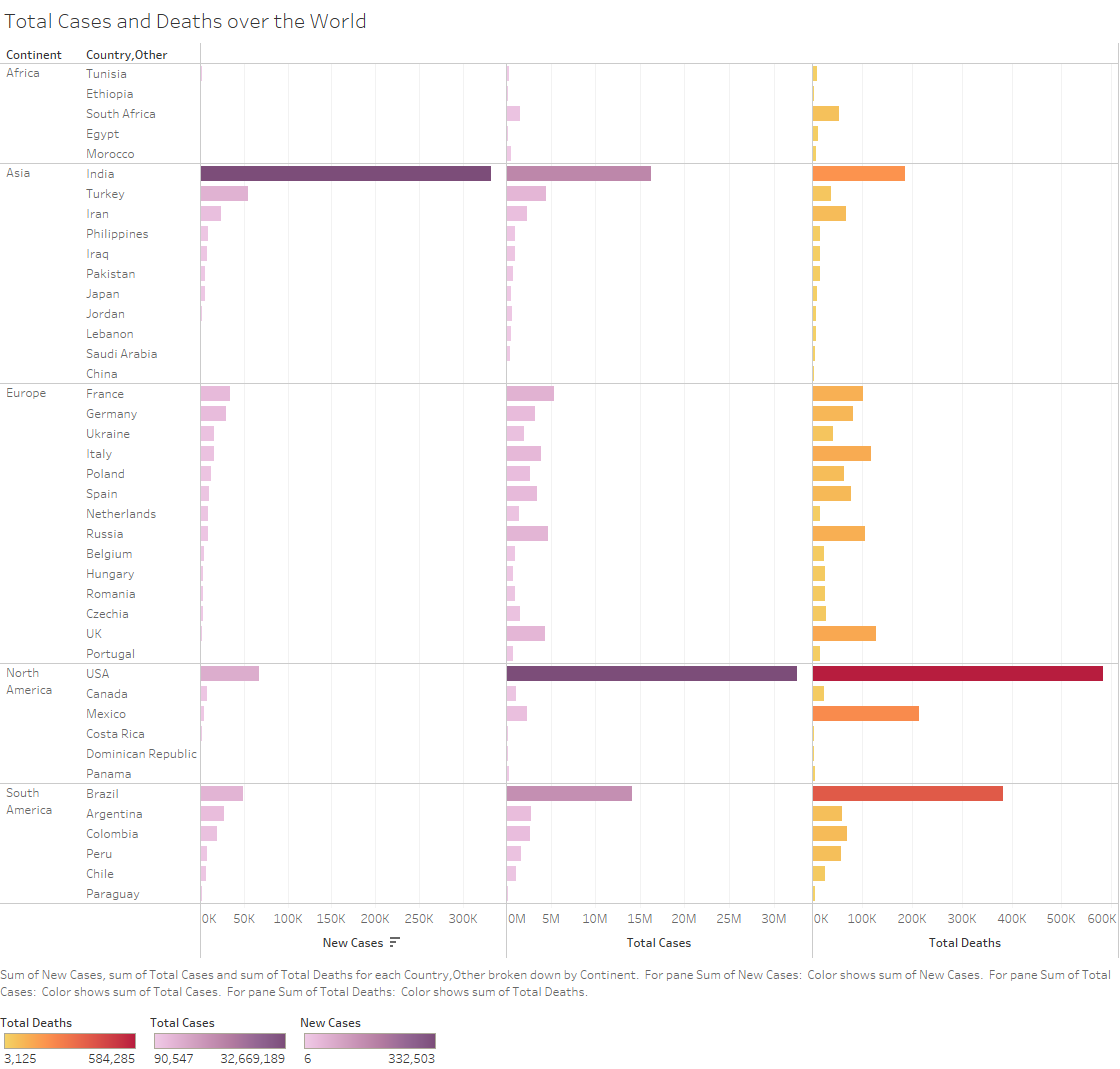


Hình . Số ca mới, Tổng số ca và Số ca chết qua từng nước trong các châu lục. Highlighting theo màu sắc.

Ở hình trên, ta phân loại dữ liệu trực quan bằng màu sắc theo từng *Continent*. Điều này giúp ta có thể nhấn mạnh và so sánh các biến còn lại dựa vào biến *Continent*.

Cụ thể ở đây, tập trung theo từng châu lục thì:

* Châu Phi có ít số ca mắc mới trong ngày nhất, các nước trong khu vực không vượt quá 2500 ca. Thậm chí ở Ai Cập, có chưa tới 200 ca mới.
* Sự phân bố các ca mắc ở Châu Mỹ là tương đối đều, khi mức độ của các ca mắc mới, tổng số ca, tổng ca tử vong có độ chênh lệch là không quá lớn. Có thể thấy, dịch bệnh lan đều giữa các nước trong khu vực.
* Trong khi đó, ở Châu Á thì Ấn Độ là nước có độ chênh lệch khủng khiếp so với phần còn lại.
* Tương tự ở Nam Mỹ là USA, còn Bắc Mỹ là Brazil. Xét trên tổng toàn thế giới thì cả 2 nước vẫn bị Ấn Độ bỏ xa.



Hình . Số ca mới, Tổng số ca và Số ca chết qua từng nước trong các châu lục. Highlighting theo dải màu.

Ở hình trên, ta phân loại dữ liệu trực quan bằng màu sắc từ đậm đến nhạt dựa vào biến mang tính định lượng. Điều này giúp ta thông qua dữ liệu có thể nhấn mạnh vào mức độ của các biến liên tục (*Total Case, New Cases* và *Total Deaths* là biến phụ thuộc vào các biến độc lập) hơn là các biến định danh (*Continent* và *Country*).

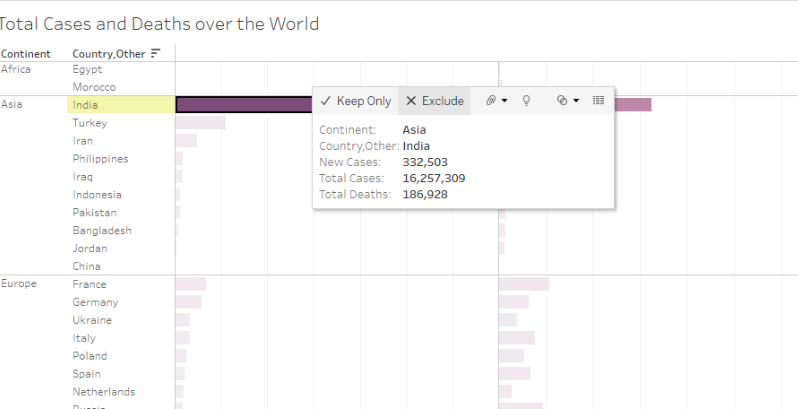
Cụ thể có thể thấy:

* Xét về số ca mắc mới thì Ấn Độ có mức độ lớn hơn hẳn.
* Còn xét về tổng số ca tử vong và tổng số ca mắc thì USA đang dẫn đầu, tiếp theo sau là Brazil rồi đến Ấn Độ.
  1. **Reducing Attributes.**

*Giải thích:* Lược bỏ một số thuộc tính, phần tử mà người dùng mong muốn. Gồm 2 kĩ thuật nhỏ là *Filter* và *Aggregate* sẽ được giới thiệu chi tiết sau.

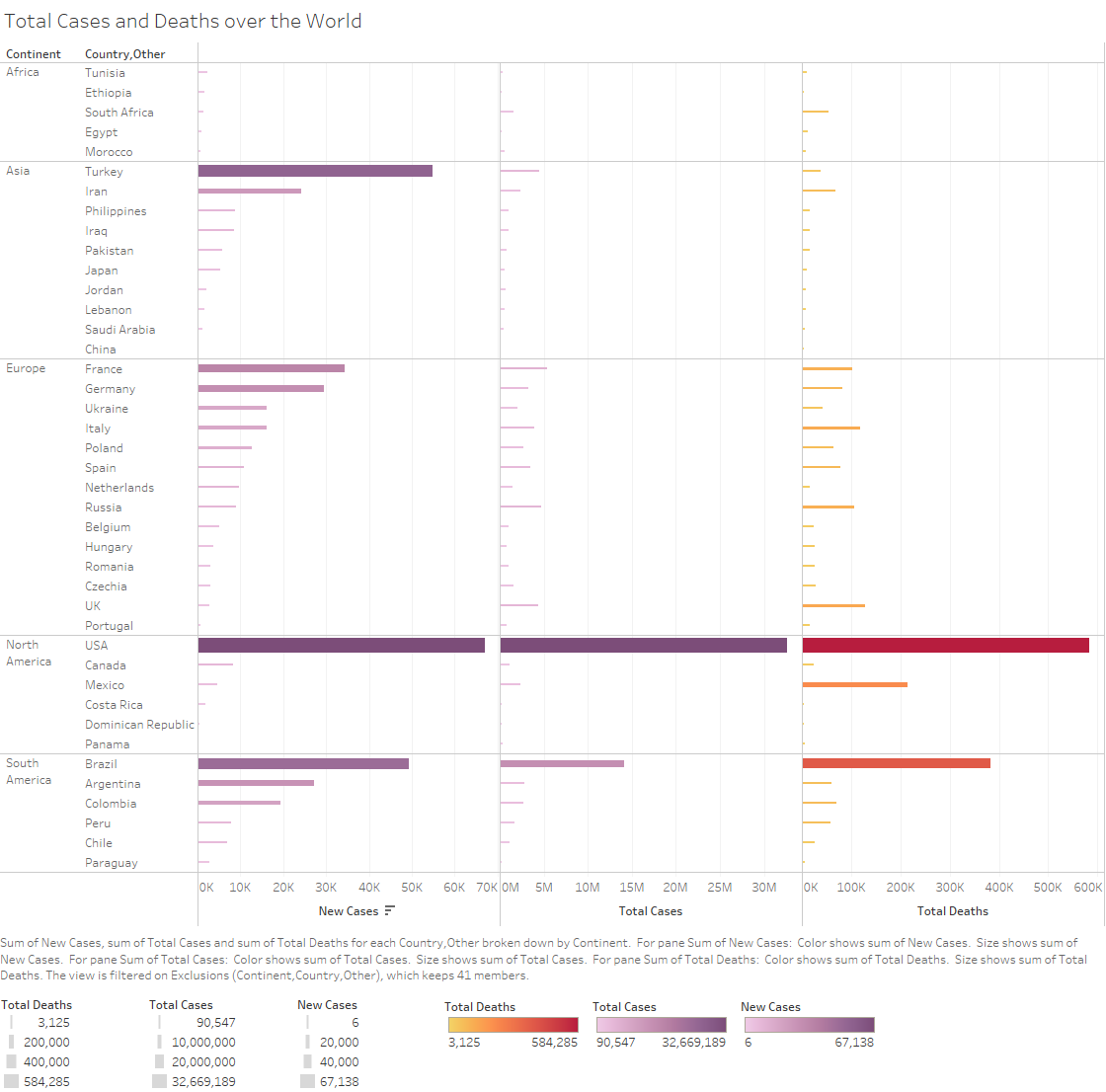
*Ý nghĩa:* Giúp lọc đi những thuộc tính không mong muốn để tập trung vào các thuộc tính quan trọng.

*Dữ liệu chọn lựa để trình diễn:* Sử dụng lại biểu đồ đã vẽ ở phần *Facet: Partition into Views*. Trên biểu đồ, do sự chênh lệch của India với các nước khác về mọi thuộc tính là quá lớn, nên ta có thể tạm lược bớt phần tử India để quan sát độ scale của các nước khác rõ hơn.



Hình . Chuột phải vào phần tử và chọn Exclude để lược bỏ (hay còn gọi là Filter)

Sau khi *Filter*, ta dùng kĩ thuật *Highlighting* theo cả dải màu và kích thước.



Hình . Số ca mới, Tổng số ca và Số ca chết qua từng nước trong các châu lục. Highlighting theo dải màu và kích thước.

Sau khi áp dụng các kĩ thuật, biểu đồ trở nên dễ quan sát hơn rất nhiều. Đặc biệt là dồn sự chú ý vào những quốc gia có vị trí xếp hạng cao nhất theo từng loại biến.

* 1. **Changing Viewpoint.**

*Ý nghĩa:* Ứng dụng với các loại biểu đồ 3D, ta thay đổi góc nhìn camera để có thể quan sát được nhiều chiều của dữ liệu, đặc biệt với biểu đồ scatter plot nhiều hơn 2 biến. Hiện tại, Tableau chưa hỗ trợ chức năng vẽ 3D nên ta không thực hiện được kĩ thuật này.

1. **Reduce.**
   1. **Filter.**

*Giải thích:* Loại bỏ những thuộc tính không mong muốn, lọc các loại biến theo các giá trị.

*Ý nghĩa:* Giúp hiển thị thêm/bớt các thành phần trong biểu đồ, đem lại hiệu quả hiển thị mong muốn.

|  |  |
| --- | --- |
| Hình . Kéo thuộc tính vào ô Filters. | Hình . Ô chọn miền giá trị. |

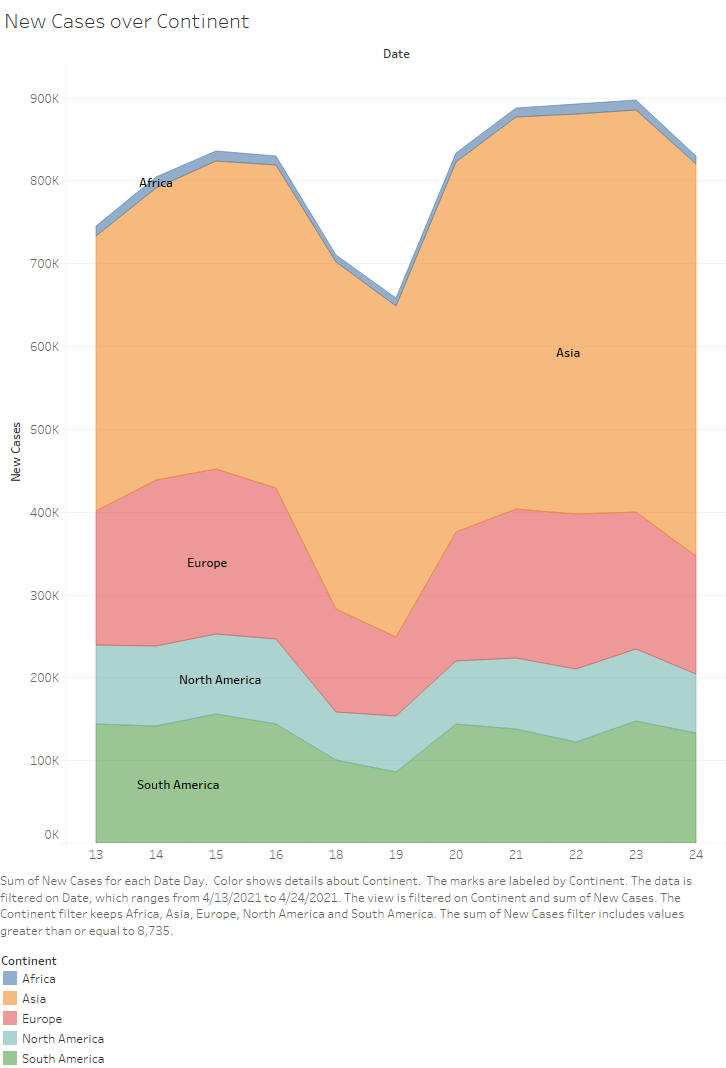
Để tạo filter, ta chỉ cần kéo thả thuộc tính mong muốn vào ô *Filters*. Sau đó tick chọn các miền giá trị có trong danh sách filter, rồi nhấn OK. Click chuột phải vào biến, chọn ‘Show Filter’ để hiển thị filter.

|  |  |
| --- | --- |
| Hình . Hiển thị filter. | Hình . Ô filter. |

Ô chọn filter hiển thị ở bên phải màn hình, ta có thể tick ô trống hoặc kéo thanh trượt để lọc theo từng biến. Nhấp chuột phải vào ô filter, Tableau cung cấp rất nhiều chọn lựa filter (list – danh sách, dropdown – hộp xổ xuống, silder – thanh trượt, v.v.).

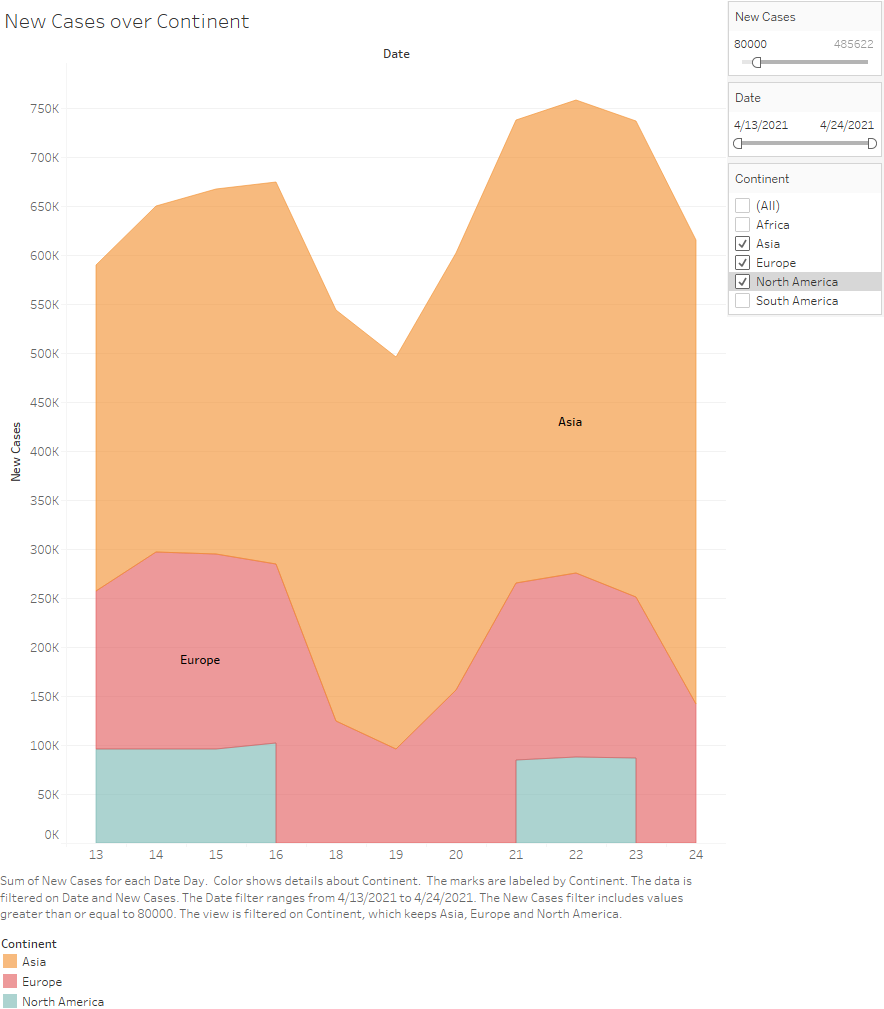
Phương pháp *Filter* kết hợp rất tốt với *Animated Transistion.* Giúp việc trực quan trở nên chuyên nghiệp và dynamic hơn.

*Dữ liệu chọn lựa để trình diễn:* tổng số ca mắc mới theo các lục địa từ ngày 13 đến 24-04-2021.



Hình . Số ca mắc mới của các lục địa từ 13 đến 24-04-2021.

Africa có số ca mới qua ít (<10,000 ca) so với các lục địa khác, nên ta sử dụng filter để tạm thời loại bỏ nó ra khỏi biểu đồ. Giả sử ta muốn xem các châu lục có số ca mới 80,000 ca ảnh hưởng thế nào tới tổng số ca mới trên thế giới, ta sử dụng filter cho New Cases.



Hình . Các lục địa có số ca mắc mới vượt 80,000 ca từ 13 đến 24-04-2021.

Từ biểu đồ trên, nếu xét theo tiêu chí chú trọng vào các châu lục có số ca mắc mới trên 80,000 ca thì có thể thấy South America không tham gia quá nhiều vào việc tăng số ca mới trong ngày cho thế giới. North America từ ngày 16 đến 21 cũng không có số ca mới vượt con số 80,000.

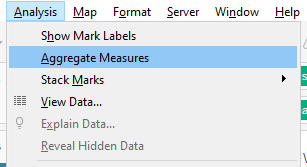
* 1. **Aggerate.**

*Giải thích*: Các phần tử được tổng hợp (*aggerate*) với nhau bằng tập hợp. Lúc này, một nhóm các phần tử được đại diện bởi một phần tử đại diện.

*Ý nghĩa*: Trong Tableau, khi thêm vào một *measure* (giá trị measure sẽ được Tableau hiển thị màu xanh lá cây), Tableau mặc định sẽ *aggerate* các giá trị của nó.

*Sum* (tổng), *Average* (trung bình) và *Median* (trung vị) là các *aggerate* thường gặp. Điều này giúp cho dữ liệu được thể hiện liền mạch hơn (ví dụ biểu đồ stack bar, line chart, v.v).

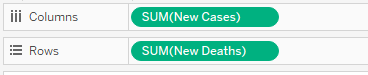
Tuy nhiên, khi ta muốn xem chi tiết sự phân bố của từng phần tử trong data. Đặc biệt khi ta muốn xem mối tương quan giữa 2 biến (ví dụ biểu đồ scatter plot). Thì ta sẽ phá bỏ *aggerate*. Biểu đồ sẽ plot từng dòng trong data thành các điểm nhỏ.



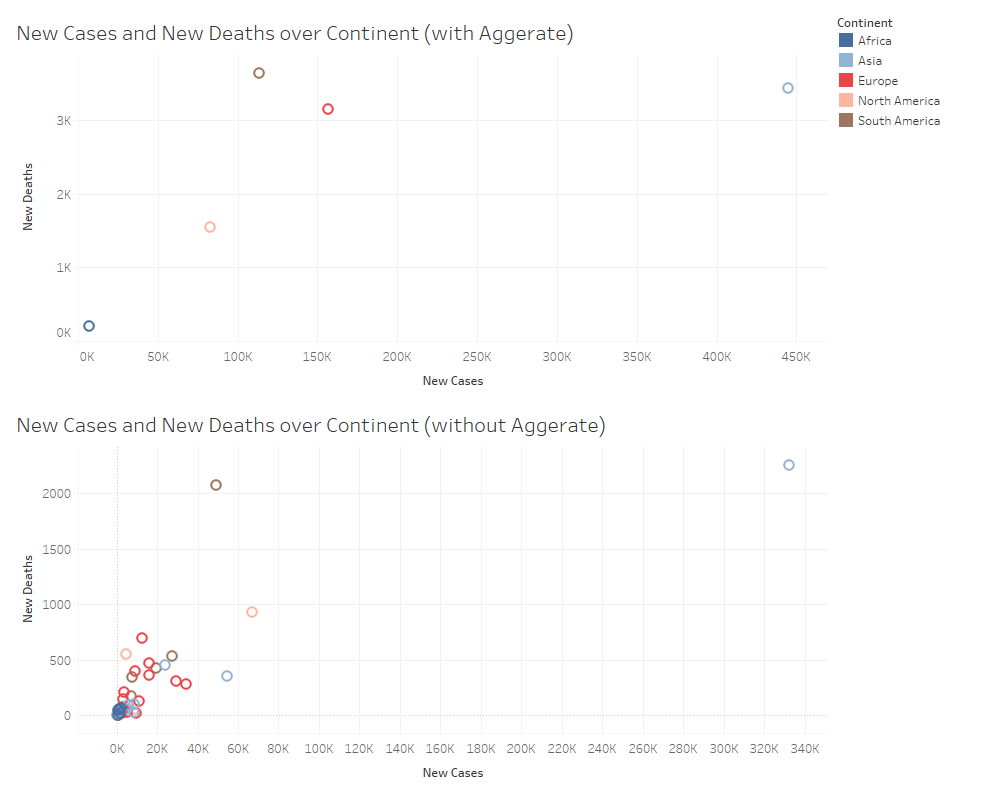
Hình . Vào Analysis > Aggregate Measures để tắt/bật aggerate cho biến.



Hình . Các biến measure khi không aggerate.



Hình . Các biến measure khi được Tableau aggerate. Ở đây sử dụng loại aggerate SUM.



Hình . Số ca mắc mới và Số ca tử vong mới qua các lục địa

*Dữ liệu chọn lựa để trình diễn:* Sốca mắc mới và số ca chết mới trong ngày 22-04-2021.

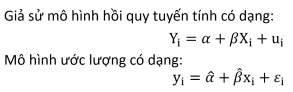
Đồ thị phía trên sử dụng aggerate nên mất đi tính chi tiết của dữ liệu, không phù hợp trong việc thể hiện mối tương quan giữa 2 loại biến, vì vậy nếu sử dụng stacked bar hoặc bar chart sẽ hợp lý hơn.

Còn ở đồ thị phía dưới, có thể thấy mối tương quan tích cực (x tăng thì y tăng) giữa 2 biến. Các điểm trên đồ thị nằm rải rác nhau chứng tỏ mối tương quan này là rất yếu. Ngoài ra, ta còn thấy điểm bất thường (outlier) ở South America và Asia.

*Chạy một số thuật toán học máy đơn giản để hiểu thêm về dữ liệu, thuật toán học máy được quyền sử dụng code có sẵn, ghi rõ nguồn gốc.*

1. **Hồi quy Bình phương nhỏ nhất (Ordinary Least Squares Regression).**

Đây là phương pháp hồi quy được sử dụng phổ biến nhất trong nghiên cứu. Dù cho trong một vài trường hợp các phương pháp hồi quy khác được ưa chuộng hơn, kết quả hồi quy bằng OLS vẫn được xem là kết quả tiêu chuẩn (benchmark).

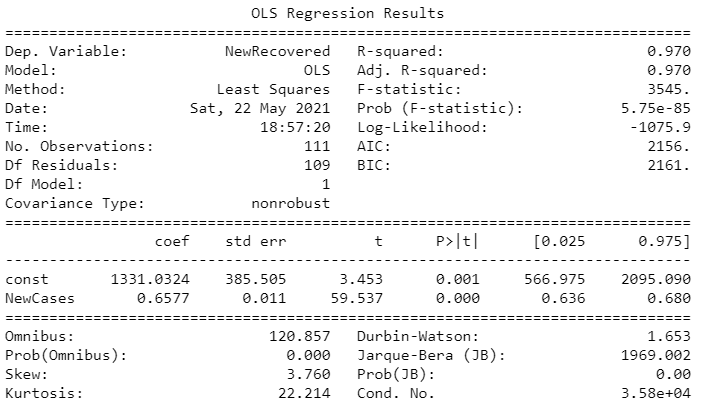


Y và X được sử dụng để thể hiện cho tất cả quan sát của một tổng thể (population), hệ số alpha và beta thể hiện các giá trị ước lượng. Phương pháp OLS sẽ lựa chọn các hệ số hồi quy alpha và beta sao cho bình phương sai số của mô hình ước lượng là nhỏ nhất.

Sử dụng dữ liệu vào ngày 23-04-2021.

|  |
| --- |
| from pandas import DataFrame  import statsmodels.api as sm  X = df['NewCases'] *# one variable for simple linear regression*  Y = df['NewRecovered']    X = sm.add\_constant(X) *# adding a constant*    model = sm.OLS(Y, X).fit()  predictions = model.predict(X)    print\_model = model.summary()  print(print\_model) |

Kết quả:



Từ kết quả trên ta có thể đưa ra kết luận:

* Biến NewCases có ý nghĩa đối với mô hình về mặt thống kê (p-value = 5.75e-85).
* Mô hình phù hợp tốt với dữ liệu quan sát về mặt thống kê (p-value = 5.75e-85).
* Biến NewCases có thể giải thích được 96.8% sự thay đổi của biến NewRecovered.
* Phương trình hồi quy:
  + NewRecovered = 1331,0324 + 0,970 \* NewCases

1. **Hồi quy tuyến tính (Linear Regression).**

Hồi quy tuyến tính là một phương pháp rất đơn giản nhưng đã được chứng minh được tính hữu ích cho một số lượng lớn các tình huống. Đây là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại.

x = df['NewDeaths']

y = df['NewCases']

from sklearn.model\_selection import train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.3)

from sklearn.linear\_model import LinearRegression

lr = LinearRegression()

lr.fit(np.array(x\_train).reshape(-1,1),np.array(y\_train).reshape(-1,1))

y\_pred=lr.predict(np.array(x\_test).reshape(-1,1))

from sklearn.metrics import mean\_squared\_error

mean\_squared\_error(x\_test,y\_pred)

Kết quả:



Ta thấy, giá trị MSE khá lớn chứng tỏ mức độ lỗi của mô hình còn lớn, nghĩa khác biệt lớn giữa giá trị thực tế và giá trị dự đoán. Cho nên mô hình dự đoán này không được đánh giá tốt để sử dụng.

**Tài liệu tham khảo:**

<https://datatofish.com/statsmodels-linear-regression/>

<https://www.youtube.com/watch?v=T_NRGfhZapE>

<https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3>

<https://stataguide.wordpress.com/2020/04/13/mo-hinh-hoi-quy-ols/>