1. **CƠ SỞ LÝ THUYẾT**

**Data Visualization – Trực quan hóa dữ liệu**

Data Visualization là phương pháp biểu diễn dữ liệu thành hình ảnh dạng sơ đồ, đường, cột, … giúp dữ liệu dễ hiểu, dễ tiếp cận và sử dụng hơn.

Dữ liệu trong Data Visualization chia 2 loại sử dụng kết hợp với nhau:

* **Categorical** (Phân loại): Mô tả bản chất của dữ liệu. VD: “Tên”, “Tuổi”. Bao gồm cả dữ liệu định tính (không phải dữ liệu số).
* **Quantitative** (Định lượng): Dữ liệu dạng số.

Để biểu diễn dữ liệu phù hợp, ta cần phải dựa trên các tiêu chí:

* Những đặc trưng nào cần được biểu diễn?
* Giá trị của đặc trưng đó là rời rạc hay liên tục? Bởi mỗi loại giá trị sẽ có cách biểu diễn khác nhau.
* Có bao nhiêu dòng dữ liệu cần biểu diễn? Các biểu đồ khác nhau sẽ phù hợp với số lượng dữ liệu nhiều hay ít.
* Các giá trị có phụ thuộc thời gian, phân chia theo nhóm?
* Mình muốn truyền đạt tới người dùng điều gì? VD: sự so sánh của dữ liệu, nhận diện xu hướng dữ liệu, …

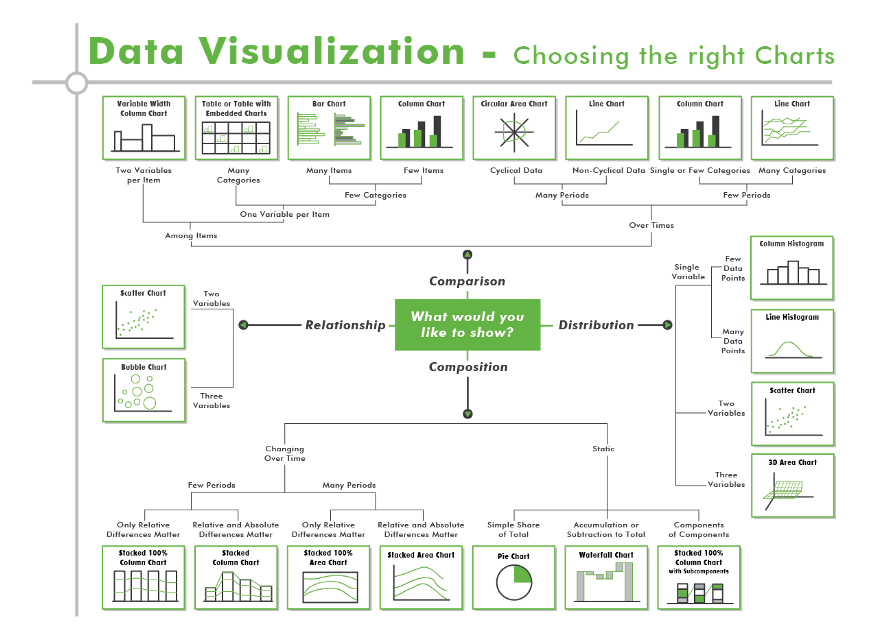
Bên cạnh đó để lựa chọn loại biểu đồ phù hợp cũng cần phải xét những tiêu chí sau:

* **Có cần so sánh dữ liệu**? Các loại chart phù hợp: Column, Mekko, Bar, Pie, Line, Scatterplot, Bullet, …
* **Có cần biểu diễn sự kết hợp của các đặc trưng**? Các loại chart phù hợp: Pie, Stacked Bar, Mekko, Stacked Column, Area, Waterfall, …
* **Có cần hiểu phân phối của dữ liệu**? Các biểu đồ dạng phân phối giúp hiểu thêm về các phạm vi, xu hướng thông thường và miền giá trị của dữ liệu. Các chart phù hợp: Scatterplot, Mekko, Line, Column, Bar.
* **Có cần phân tích xu hướng dữ liệu**? Các chart sau giúp biết thêm thông tin về cách một tập dữ liệu được thực hiện trong một khoảng thời gian cụ thể: Line, Dual-Axis Line, Column.
* **Có cần xem mối quan hệ giữa các tập dữ liệu**? Các chart này có thể show ra những ảnh hưởng tích cực hoặc tiêu cực của thuộc tính này tới các thuộc tính khác: Scatterplot, Bubble, Line.

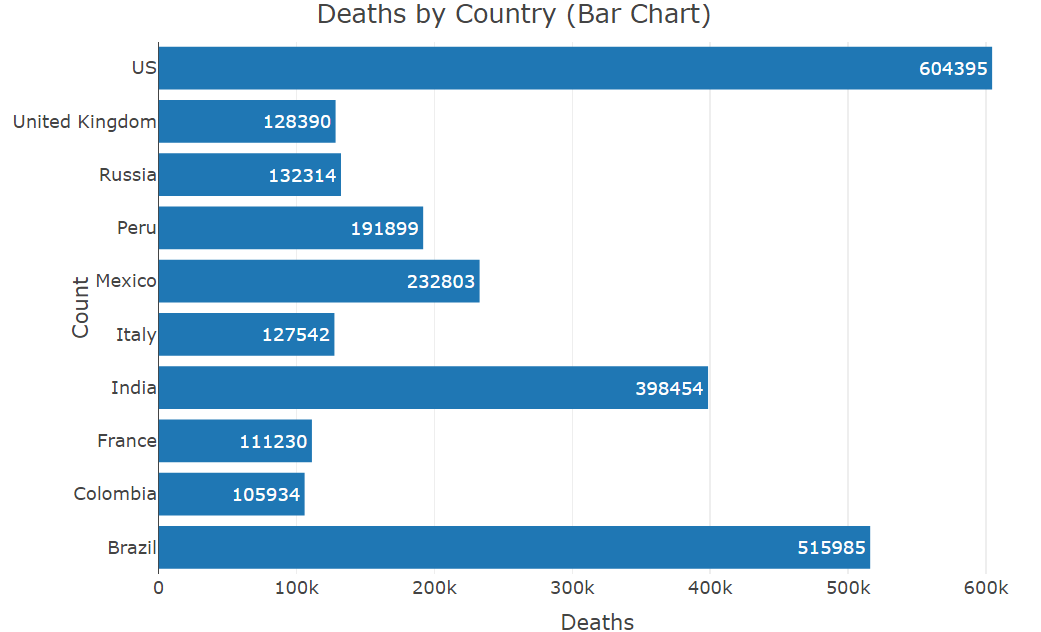
Có 4 nhóm biểu đồ chính:

* **Comparison**: Thể hiện sự so sánh giữa 1 hoặc nhiều tập dữ liệu.
* **Relationship**: Thể hiện mối quan hệ hoặc tương quan giữa 2 hoặc nhiều giá trị.
* **Composition**: Thể hiện những phần trong tổng thể và thay đổi theo thời gian.
* **Distribution**: Thể hiện phân phối theo thời gian, giúp nhìn thấy được xu hướng và phạm vi dữ liệu.

Sơ đồ sau gợi ý các dạng biểu đồ và lựa chọn như thế nào. Tùy thuộc vào nhu cầu chúng ta có thể kết hợp các dạng biểu đồ và thể hiện chúng cho phù hợp.



1. **TIẾN HÀNH TRỰC QUAN HÓA DỮ LIỆU BẰNG NGÔN NGỮ LẬP TRÌNH R**
   1. **Bar Chart**



Mô tả đồ thị/biểu đồ:

- Trục x thể hiện số ca tử vong (theo đơn vị hàng nghìn ~ k ), Trục y theo các nước

- Số ứng với mỗi cột thể hiện chính xác số ca tử vong mỗi nước

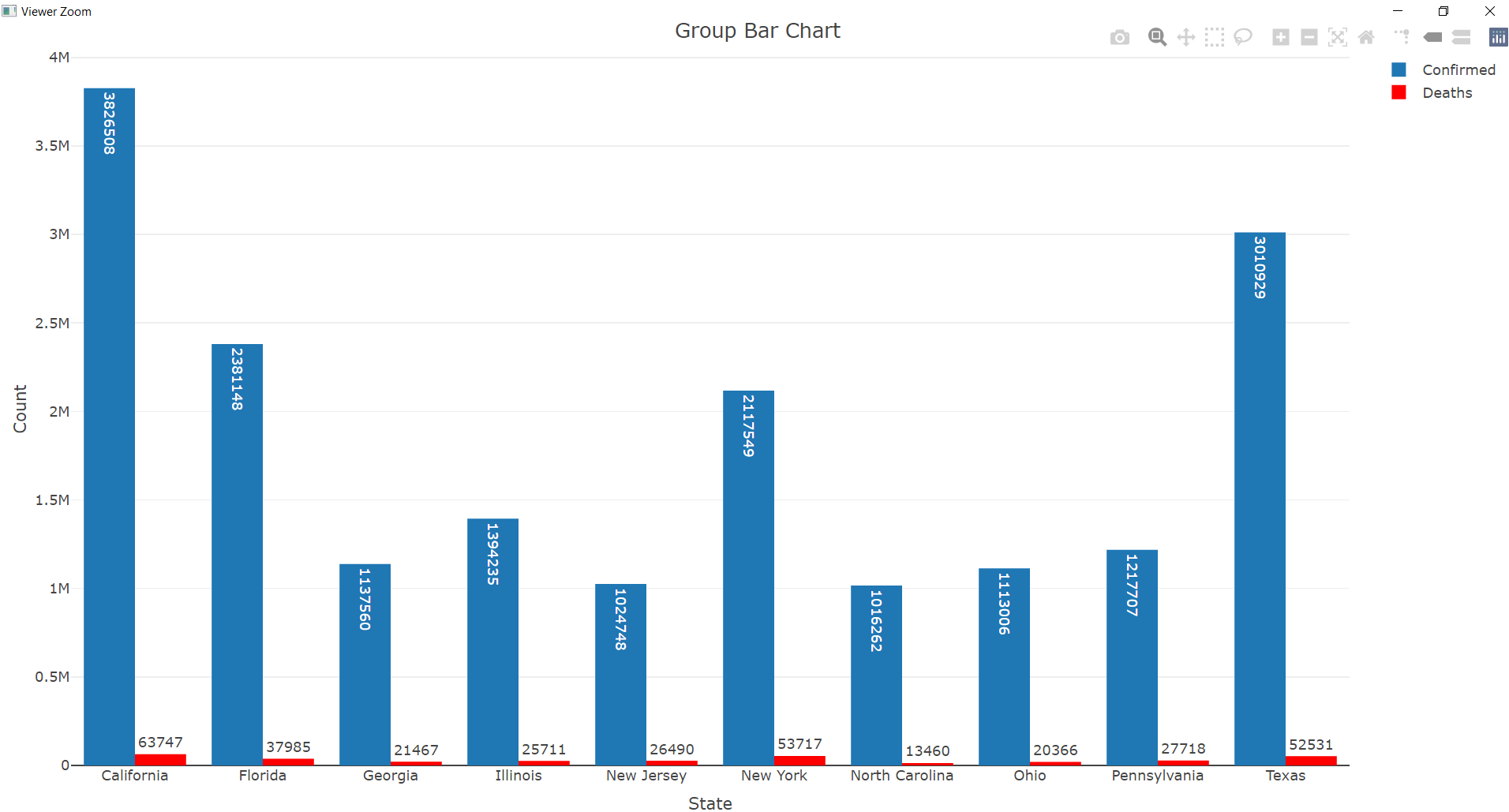
- Trong 10 nước trên, số ca tử vong chủ yếu vào khoảng 10.000

- Mỹ có số ca tử vong do covid19 ghi nhận được nhiều nhất (604.395 ca)

- Các nước: Brazil, Ấn Độ, Mỹ có số ca tử vong cao (khoảng trên 40.000)

Link:

* 1. **Group Bar Chart**

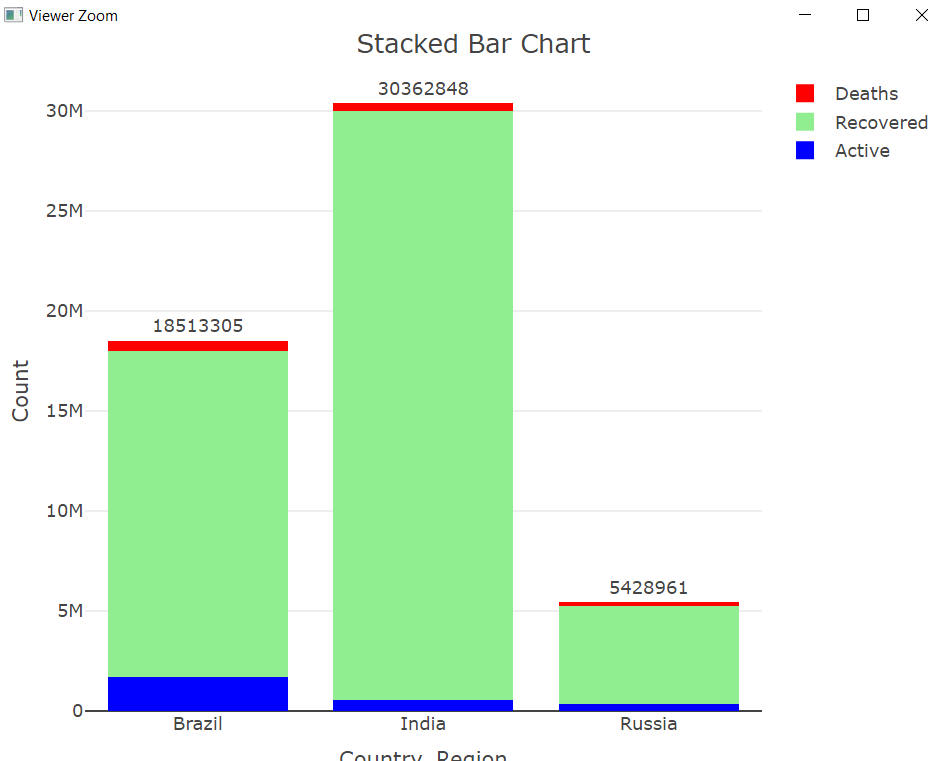


Mô tả đồ thị/biểu đồ:

- Trục y thể hiện số ca (theo đơn vị hàng triệu ~ M ), Trục x theo các Bang của Mỹ  
- Trong các Bang trên, California là bang có số ca nhiễm nhiều nhất (3826508 ca)  
- Trong các Bang trên, North Carolina là bang có số ca tử vong thấp nhất. (13460 ca)  
- Tỉ lệ tử vong được tính bằng (Deaths/Confirmed\*100):  
 => Bang có tỉ lệ tử vong thấp nhất là: North Carolina (1.324462%)  
 =>Bang có tỉ lệ tử vong cao nhất là: New Jersey (2.585026%)   
- Bình quân số ca nhiễm ở các bang trên vào khoảng 10.000 ca

Link:

* 1. **Stacked Bar Chart**

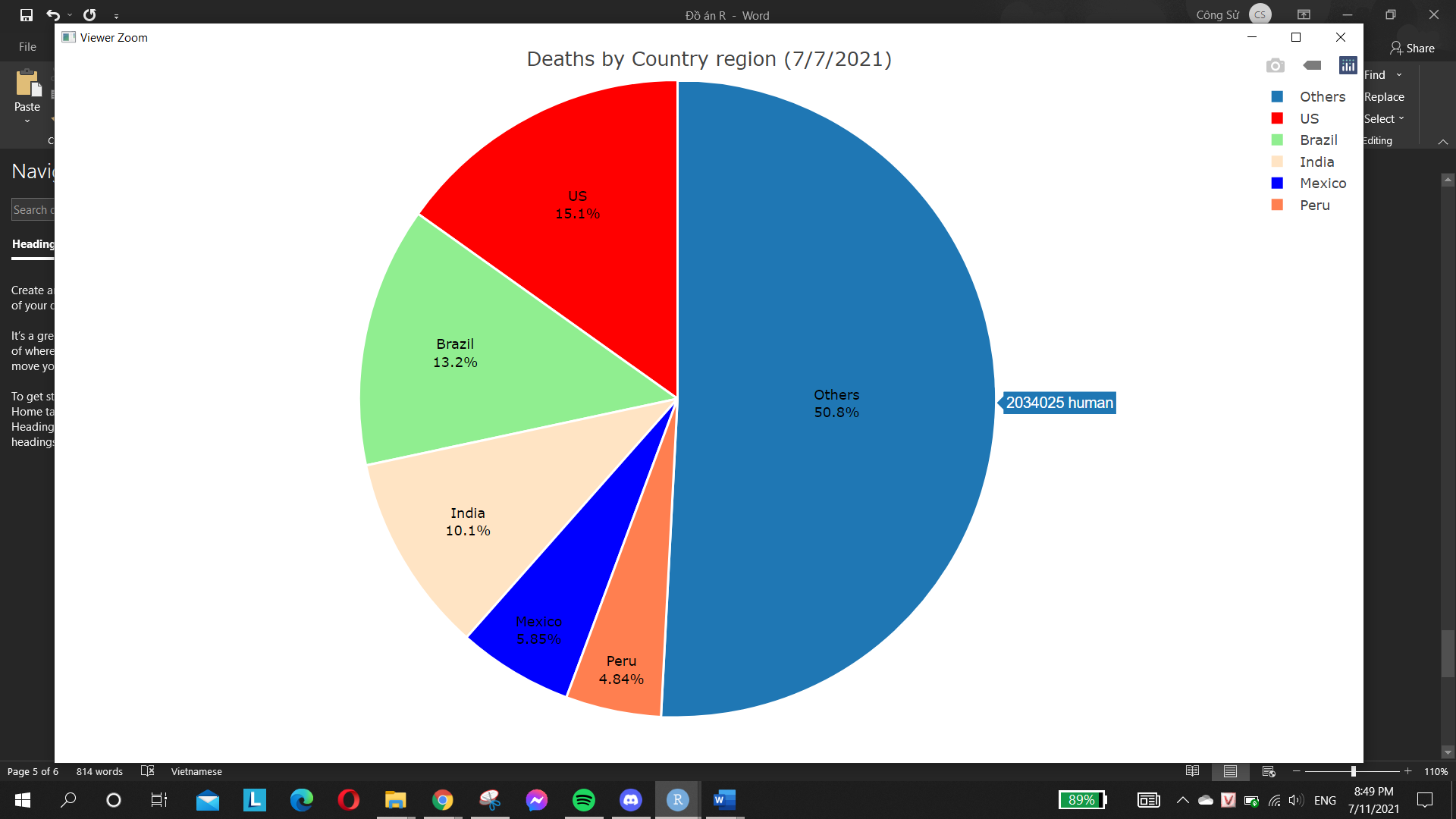


Mô tả đồ thị/biểu đồ:

- Trục y thể hiện số ca (theo đơn vị hàng triệu ~ M ), Trục x theo các nước.  
- Confirmed = Deaths +Active +Recovered  
- Trong 3 nước trên, Ấn Độ có số ca nhiễm nhiều nhất (Cột dài nhất)  
- Tỉ lệ tử vong ở 3 nước là: (Deaths/Confirmed\*100):  
 +Brazil: 2.787%  
 +Ấn Độ: 1.312%  
 +Nga: 2.437188  
- Tỉ lệ hồi phục ở 3 nước là: (Recovered/Confirmed\*100)  
 +Brazil: 87.924%  
 +Ấn Độ: 96.918%  
 +Nga: 90.738%

Link:

* 1. **Pie Chart**

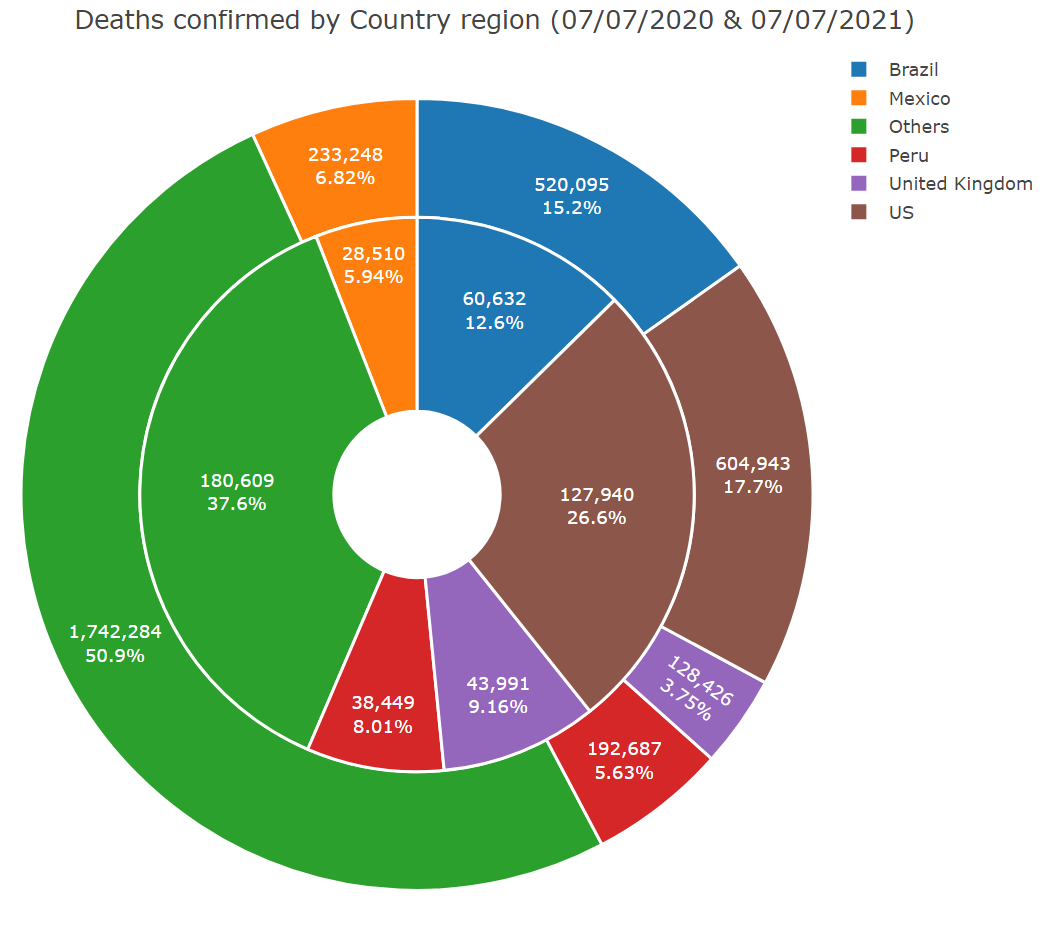


Mô tả đồ thị/biểu đồ:

- Biểu đồ Tròn (PieChart) sử dụng biểu diễn tỉ lệ phần trăm của các nước so với thế giới (Tính đến ngày 07/07/2021, theo tổng số ca tử vong được ghi nhận)  
- Ứng với mỗi nước, có tỉ lệ % kèm theo  
- Biểu đồ thể hiện có số ca tử vong lớn nhấn trên thế giới, chiếm 15,1%  
- Các nước còn lại chiếm 50,8% ứng với 2.034.025 ca   
=> Tổng số ca tử vong ghi nhận được trên toàn thế giới là: 2.034.025/50.8\*100 = 4003986 (ca)  
- 5 nước: Mỹ, Brazil, Ấn Độ, Mexico, Peru có số ca tử vong được ghi nhận chiếm xấp xỉ 1 nửa so với số ca ghi nhận toàn cầu.  
 -

Link:

* 1. **Doughnut Chart**

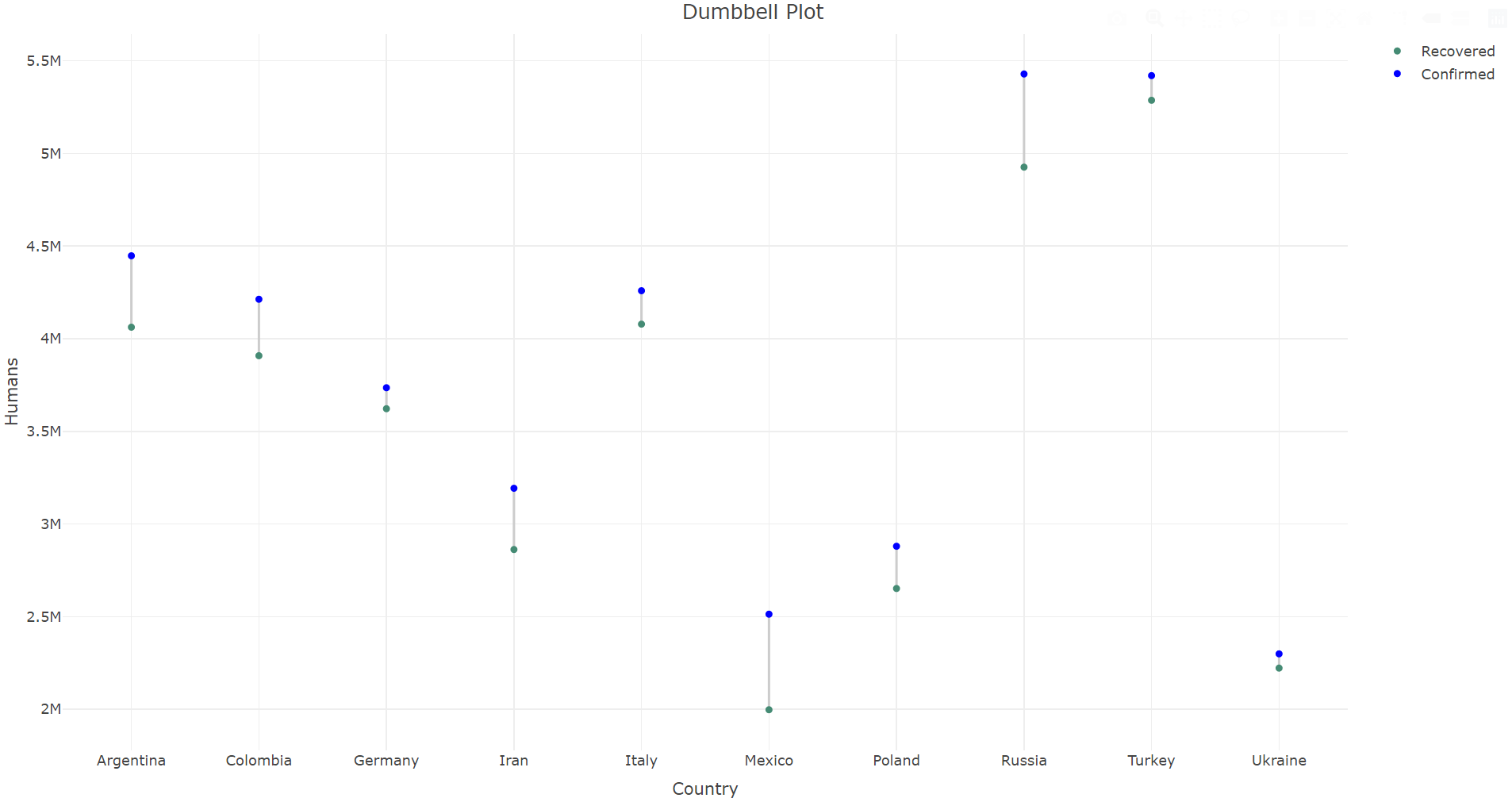


Mô tả đồ thị/biểu đồ:

- Biểu đồ Doughnut Chart (StackedPieChart) sử dụng biểu diễn tỉ lệ phần trăm của các nước so với thế giới giữa năm 2020 (tính đến ngày 07/07/2021, theo tổng số ca tử vong được ghi nhận) và năm 2021 (tính đến ngày 07/07/2021, theo tổng số ca tử vong được ghi nhận).  
- Ứng với mỗi nước, có tỉ lệ % và số ca kèm theo.  
- Phần bên trong biểu diễn năm 2020, phần ngoài biểu diễn năm 2021.  
- Cả năm 2020 và 2021, Mỹ đều có số lượng ca tử vong lớn nhất (chiếm tỉ lệ nhiều nhất) .  
- Số ca tử vong ở Mỹ sau 1 năm tăng lên gần gấp 6 lần. -

Link:

* 1. **Dumbbell Plots**

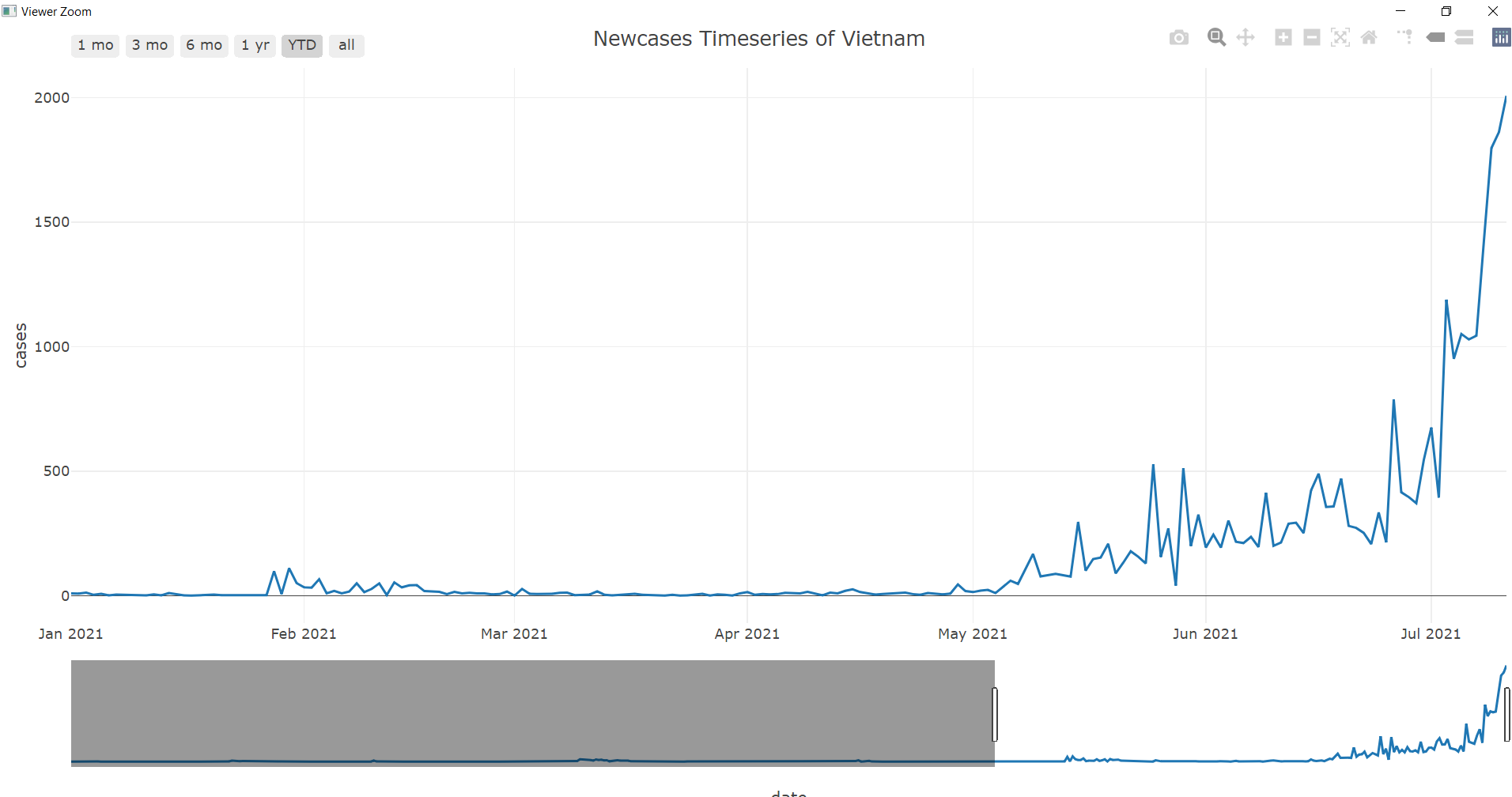


Mô tả đồ thị/biểu đồ:

- Biểu đồ Dumbbell Plots sử dụng thể hiện giá trị “Số ca nhiễm” (Chấm xanh lam) và “Số ca hồi phục” (chấm xanh lục) của 1 số nước.  
- Số ca tử vong/đang điều trị = Số ca nhiễm – Số ca hồi phục  
- Mexico , Nga và Argentina có sự chênh lệch lớn giữa số ca mắc và hồi phục.  
- Thổ Nhĩ Kỳ có số ca nhiễm ghi nhận cao, đồng thời có lượng hồi phục lớn -> Có phương pháp điều trị tốt.  
- Tỉ lệ chửa khỏi bệnh ở Mexico thấp

Link:

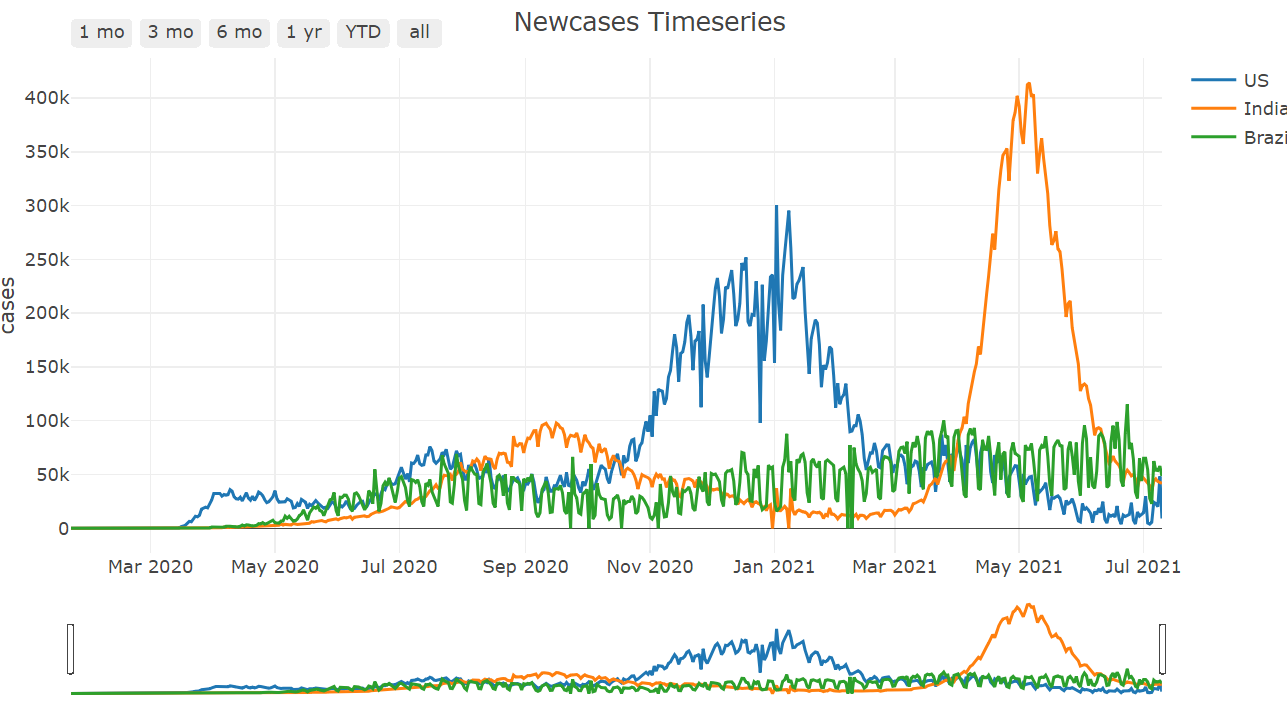
* 1. **Single line plot:**



Mô tả đồ thị/biểu đồ:

- Biểu đồ Line plot biểu diễn số ca ghi nhận mỗi ngày theo thời gian của Việt Nam  
- Từ tháng 1/2021 đến đầu tháng 2/2021 không có ca mắc mới nào, tháng 2 Việt Nam có đợt bùng phát dịch nhỏ. (số ca ghi nhận mỗi ngày dưới 100 ca)  
- Từ tháng 3/2021 đến đầu tháng 5/2021 ghi nhận rất ít ca mới.  
- Kể từ tháng 5/2021, Việt Nam bắt bùng phát đợt dịch, số ca nhiễm mỗi ngày tăng đáng kể so với tháng trước.  
- Tính đến nay số ca ghi nhận trong ngày tối đa lên đến 2000ca  
- Số ca ghi nhận mỗi ngày của Việt Nam tăng mạnh và có xu hướng tiếp tục tăng

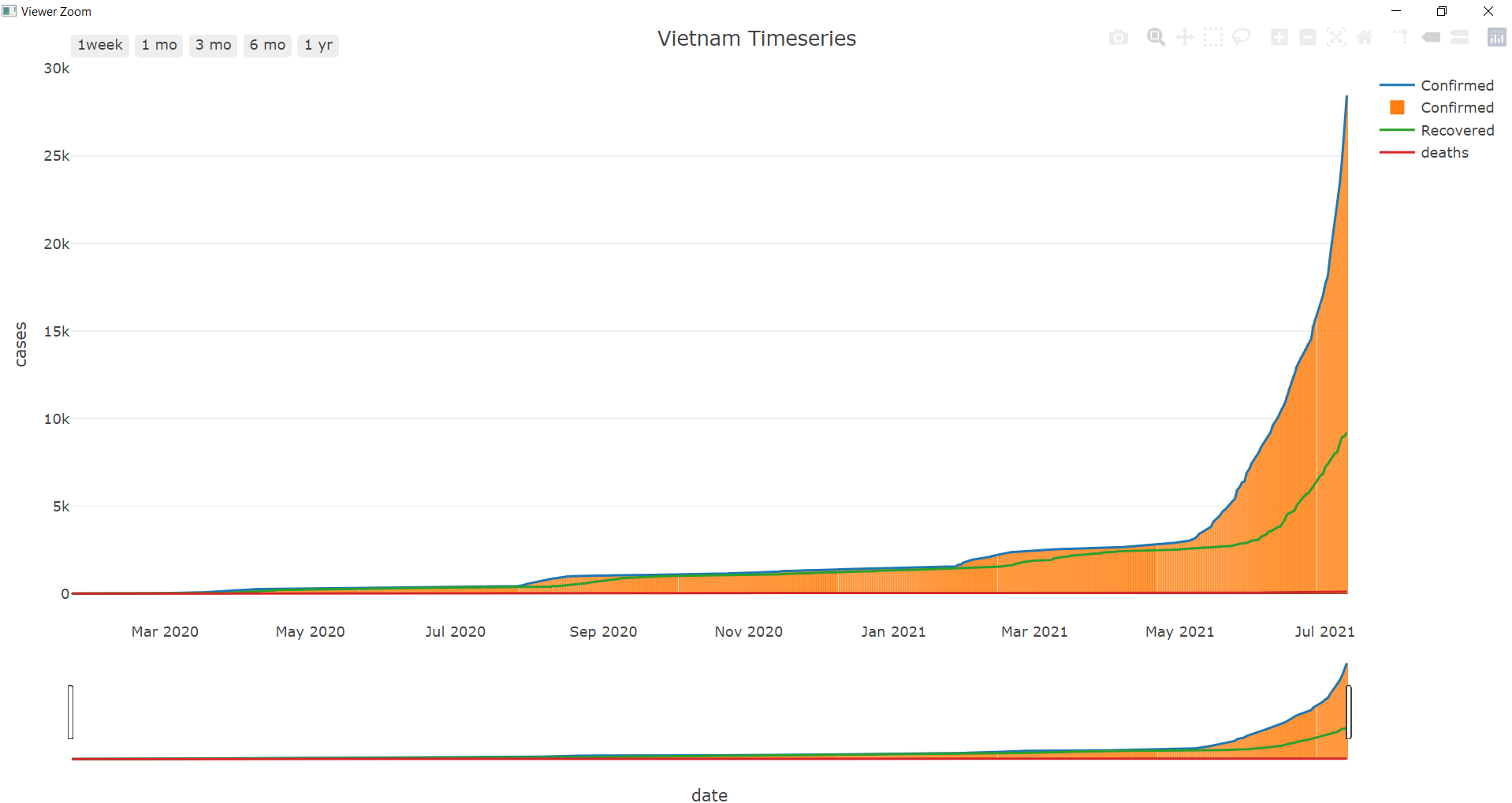
* 1. **MultiLines plot (Time Series) – Newcases**



Mô tả đồ thị/biểu đồ:

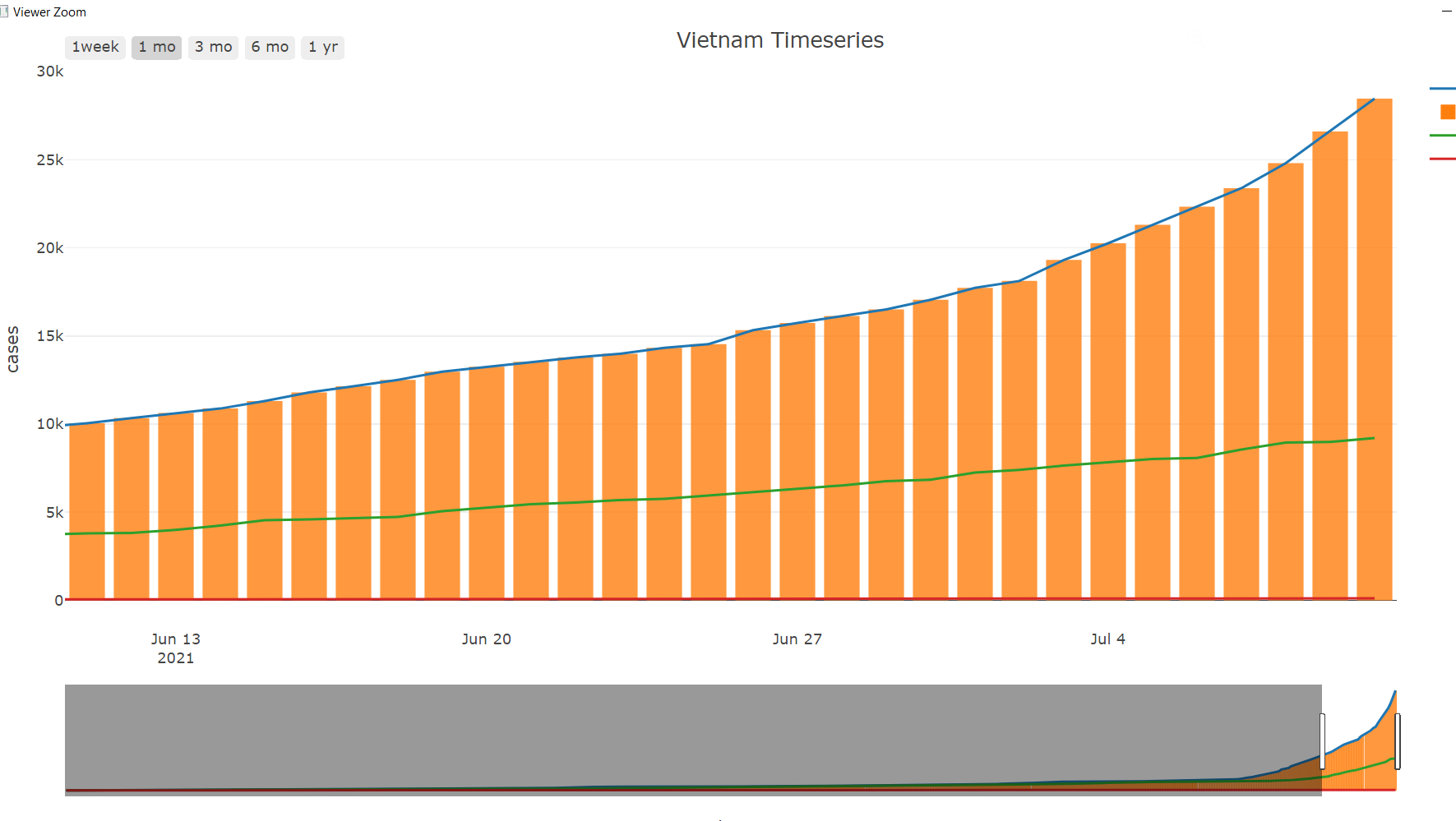
- Biểu đồ MultiLineplot biểu diễn số ca ghi nhận mỗi ngày theo thời gian của các nước. (Mỹ, Ấn Độ, Brazil)  
- Ấn độ có 2 lần bùng phát dịch, lần 1 (tháng 8-tháng 10/2020) và đợt 2 (tháng 5/2021)  
- Đỉnh dịch đợt 2 của Ấn Độ vào tháng 5/2021, với số ca mắc tối đa trong ngày : trên 400.000 ca  
- Đỉnh dịch của Mỹ vào khoảng tháng 12/2020 đến tháng 1/2021 (thời điểm lễ giáng sinh), khoảng thời gian này trung bình mỗi ngày có trên 100.000 ca  
- Tuy Brazil không có đỉnh điểm bùng phát dịch, nhưng bình quân số ca ghi nhận mỗi ngày đều dao động trong khoảng 50.000 ca.

* 1. **Bar Chart with line (Time Series) – VN**

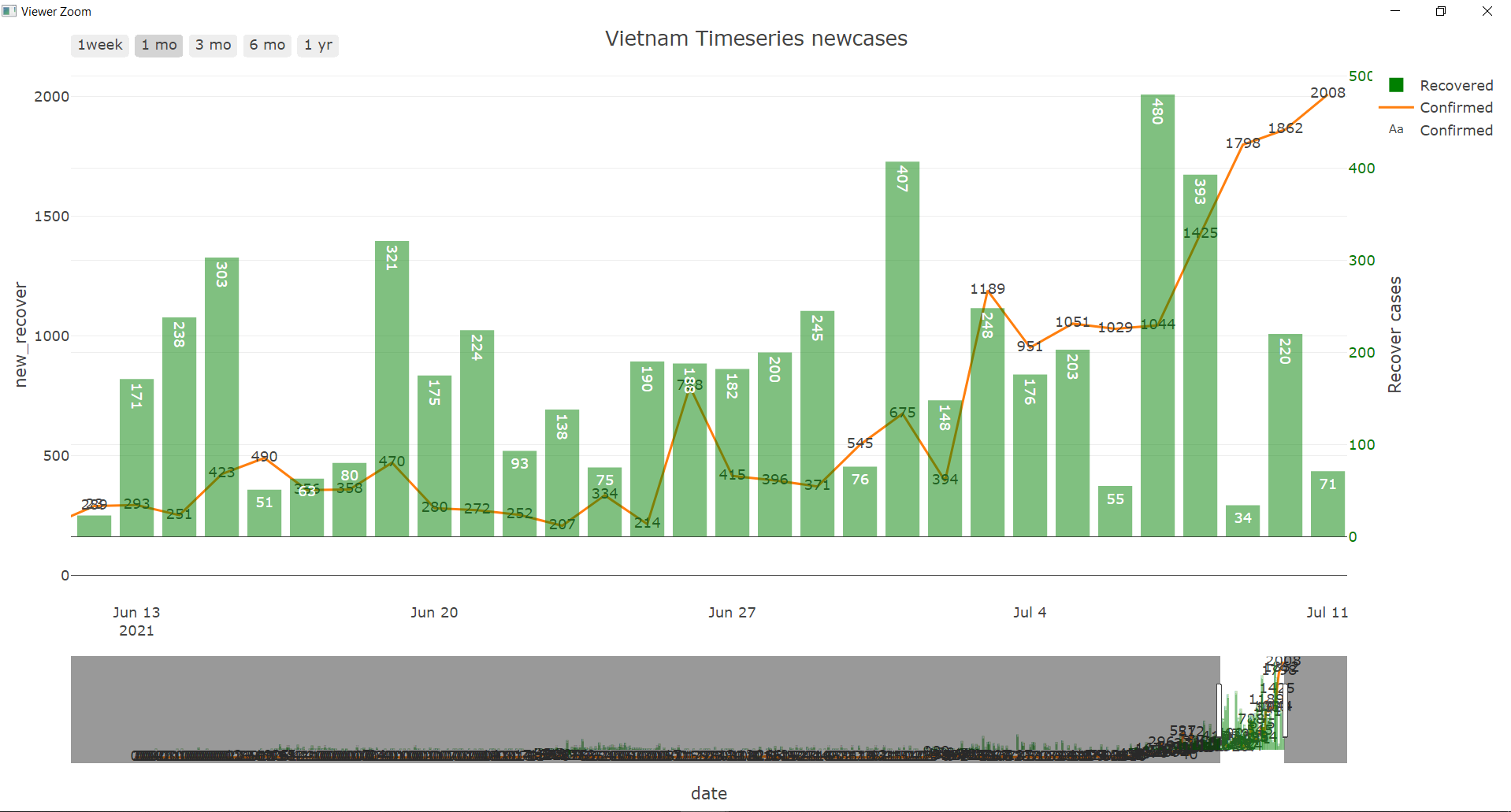


Mô tả đồ thị/biểu đồ:

- Biểu đồ biểu diễn “Số ca mắc ghi nhận được” bằng line (màu xanh lục) và phần cột (màu cam); Số ca hồi phục là đường xanh lục còn số ca tử vong là đường màu đỏ.  
- Có thể thấy phần “Tử vong” rất nhỏ -> số ca tử vong tại Việt Nam rất thấp  
- Tính đến tháng 5/2021, “Số ca mắc” và “Số ca hồi phục” tại Việt Nam tương đối bằng nhau  
=> Việt Nam rất tốt trong công tác phòng và chữa.  
- Bắt đầu vào khoảng giữa Tháng 5/2021 , Việt Nam bắt đầu bùng phát dịch, số ca nhiễm mới mỗi ngày có xu hướng tăng mạnh, và vẫn tiếp tục tăng.  
- Tính đến hiện nay (tháng 7/2021) số ca nhiễm ở Việt Nam đã xấp xỉ 30.000 ca.



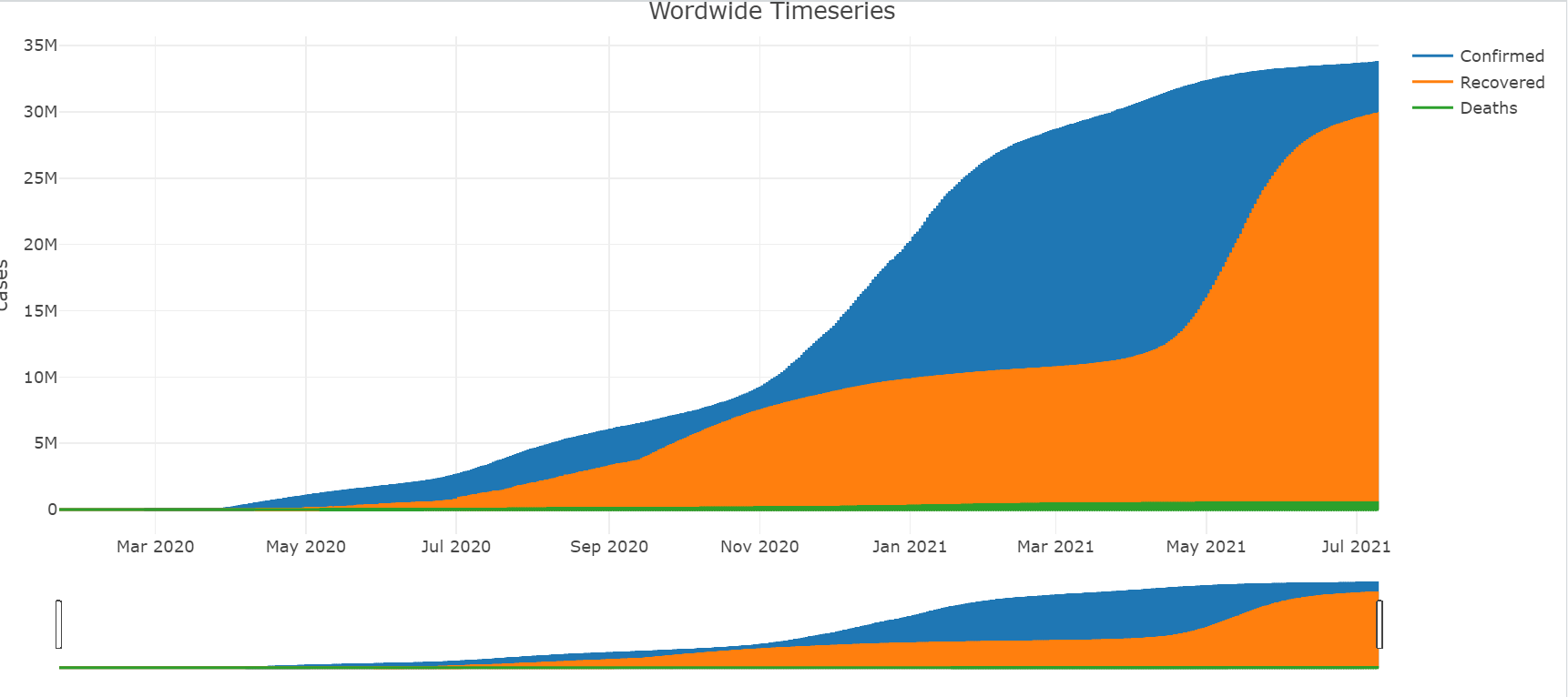
* 1. **Bar chart with line 2 axis**



Mô tả đồ thị/biểu đồ:

- Biểu đồ biểu diễn “Số ca mắc ghi nhận trong ngày” bằng line (màu cam); “Số ca hồi phục trong ngày” là cột xanh lục  
- Số ca mắc có xu hướng tăng lên   
- Số ca mắc cao nhất ghi nhận được trong 1 ngày là :2008 ca (ngày 11/7/2021)  
- Trong 1 tháng vừa qua, số ca chữa khỏi trong ngày cao nhất là 480 ca (07/07/2021)  
- Nhìn chung số ca nhiễm ghi nhận thêm mỗi ngày đều trên 200 ca

* 1. **Area Chart**

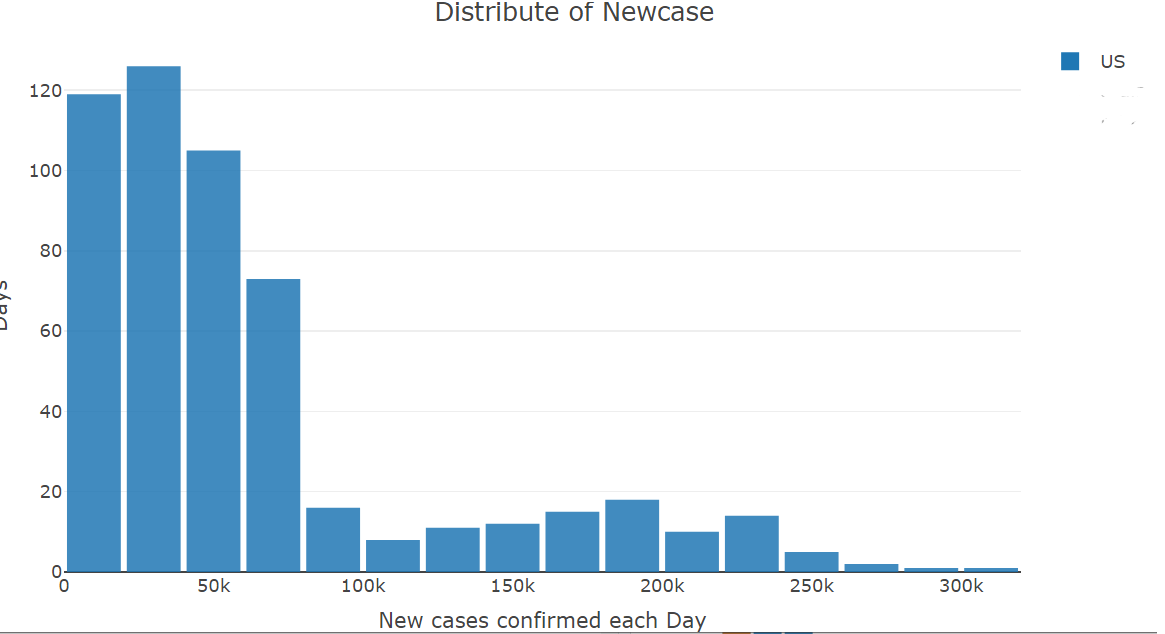


Mô tả đồ thị/biểu đồ:

- Biểu đồ Are Chart sử dụng biểu diễn “Số ca mắc”, “Số ca hồi phục”, “Số ca tử vong” theo thời gian (sử dụng Time Series Dataset)  
- Có thể thấy phần “Tử vong” chiếm diện tích rất nhỏ -> tỷ lệ tử vong thấp  
- Bắt đầu từ Tháng 9/202 Phần diện tích màu cam chiếm gần bằng phần xanh  
=> Lúc này thế giới đã tìm ra cách chữa covid  
- Bắt đầu vào khoảng giữa Tháng 11/2020 và tháng 1/2021, số ca nhiễm tăng mạnh  
=> Đây là khoảng thời gian lễ giáng sinh (Đồng thời là thời điểm bùng dịch mạnh ở Mỹ và các qgia châu Âu)  
- Cũng ở tháng 5/2021 -> Ấn độ bùng dịch (lễ hội) -> Số ca nhiễm >> Số ca phục hồi.

Link:

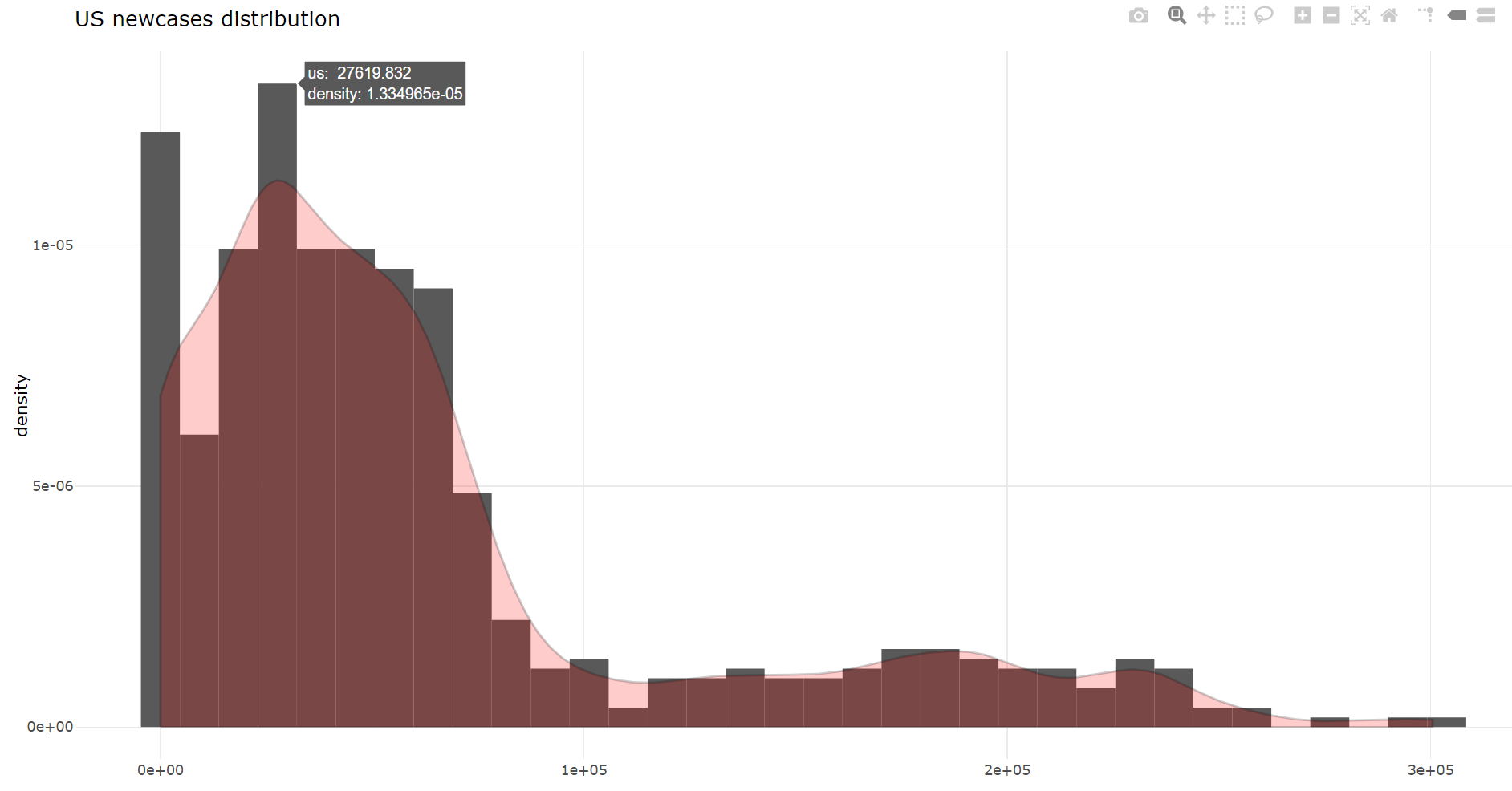
* 1. **Histogram**



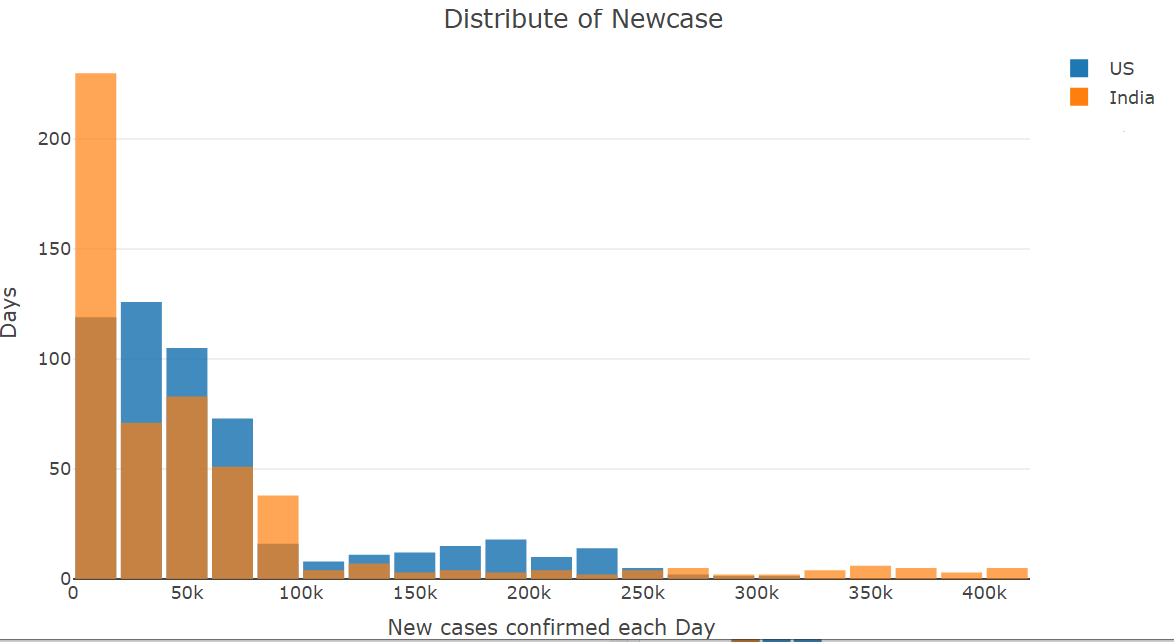
Mô tả đồ thị/biểu đồ:

- Biểu đồ Histogram thể hiện sự phân bố của dữ liệu (Số ca nhiễm mới mỗi ngày) của Mỹ  
- Trục x thể hiện số ca ghi nhận được trong 1 ngày (đơn vị hàng nghìn – k), Trục y là số lượng ngày  
- Xấp xỉ 120 ngày nước Mỹ không co thêm ca mắc mới  
- Số ca mắc ghi nhận được cao kỉ lục trong 1 ngày: 300.000 ca  
- Chủ yếu số ca mắc ở Mỹ phân bố khoảng từ 0-50.000 ca

Link:

 **# lồng histogram vào density plot**

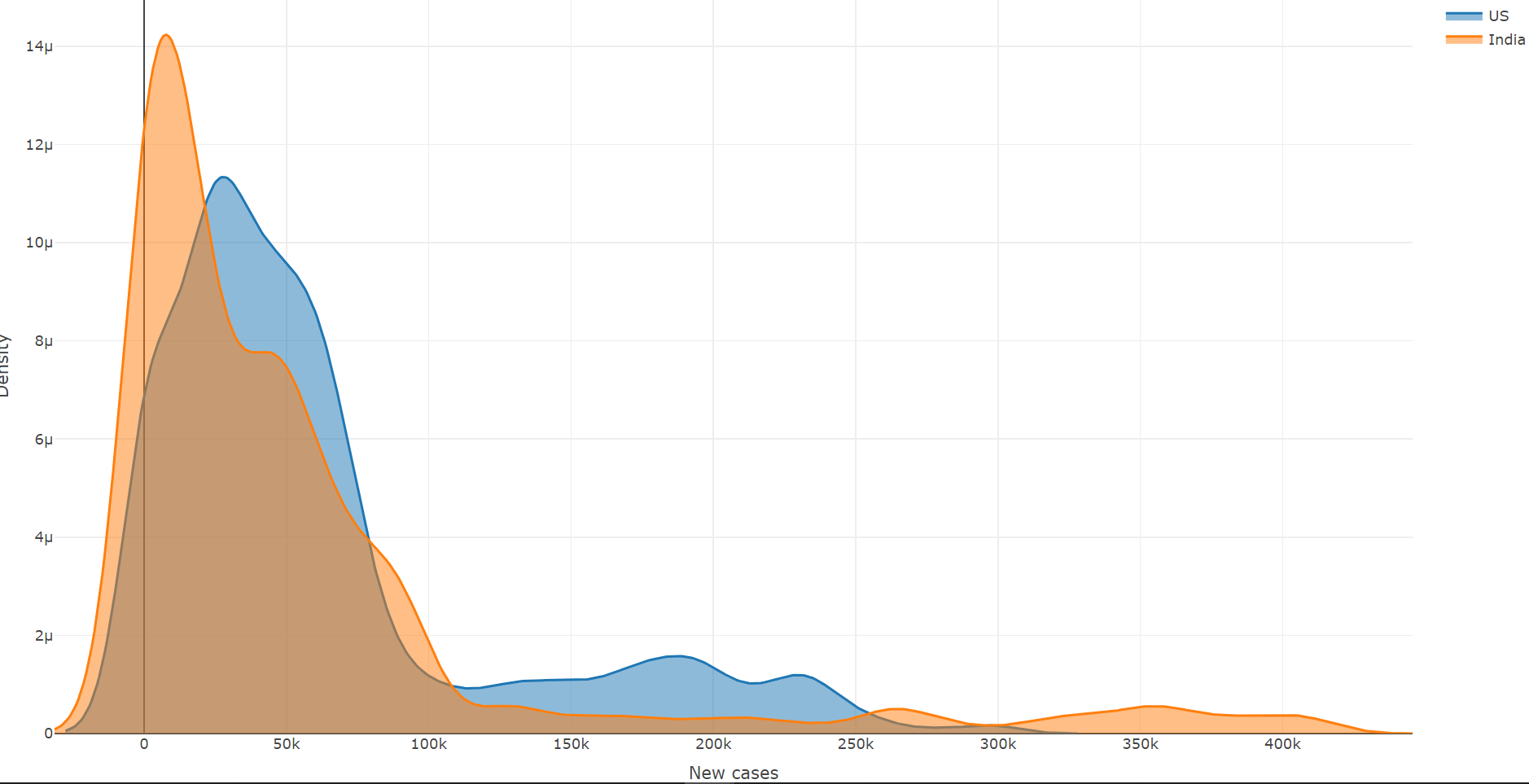
* 1. **Multiple Histogram**



Mô tả đồ thị/biểu đồ:

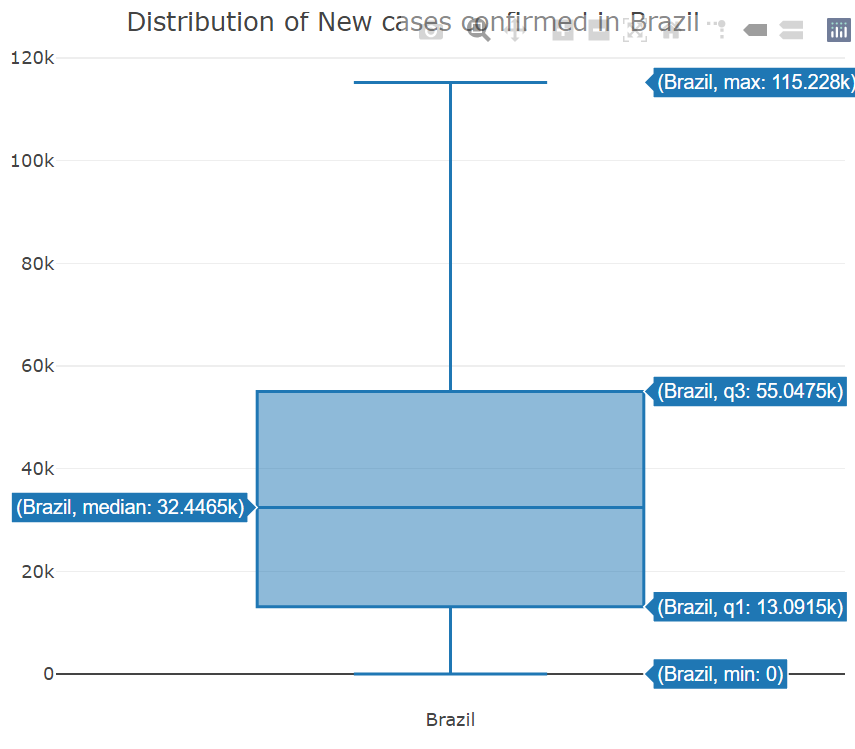
- Biểu đồ Histogram thể hiện sự phân bố của dữ liệu (Số ca nhiễm mới mỗi ngày) của Mỹ (màu xanh), và Ấn Độ (Màu cam)  
- Số ngày không ghi nhận ca mắc mới của Ấn Độ > Mĩ (hơn 200 ngày không có ghi nhận ca nhiễm mới)  
- Số ca mắc ghi nhận được cao kỉ lục trong 1 ngày ở Ấn Độ lơn hơn Mỹ (hơn 400.000 ca trong 1 ngày)  
- Số ca mắc mỗi ngày ở cả Mỹ và Ấn độ phân bố vào khoảng từ 50.000 ca  
- Có thể nói số ca nhiễm ở Mỹ đều hơn Ấn Độ (tần suất cao hơn, còn Ấn Độ có số ca tăng đột biến)

* 1. **Density Chart**



Mô tả đồ thị/biểu đồ:  
- Biểu đồ thể hiện mật độ của số lượng ca mắc bệnh ở Mỹ và Ấn Độ  
- Cột x thể hiện số ca nhiễm bệnh đã được ghi nhận ở các bang của US.  
- Cột y thể hiện tần suất  
- Mật độ số lượng ca mắc bệnh ở Mỹ chủ yếu vào khoảng 30.000 , còn của Ấn Độ vào khoảng 10.000  
- Giá trị cao nhất ghi nhận được của Mỹ là 300.000 còn của Ấn Độ vượt 400.000

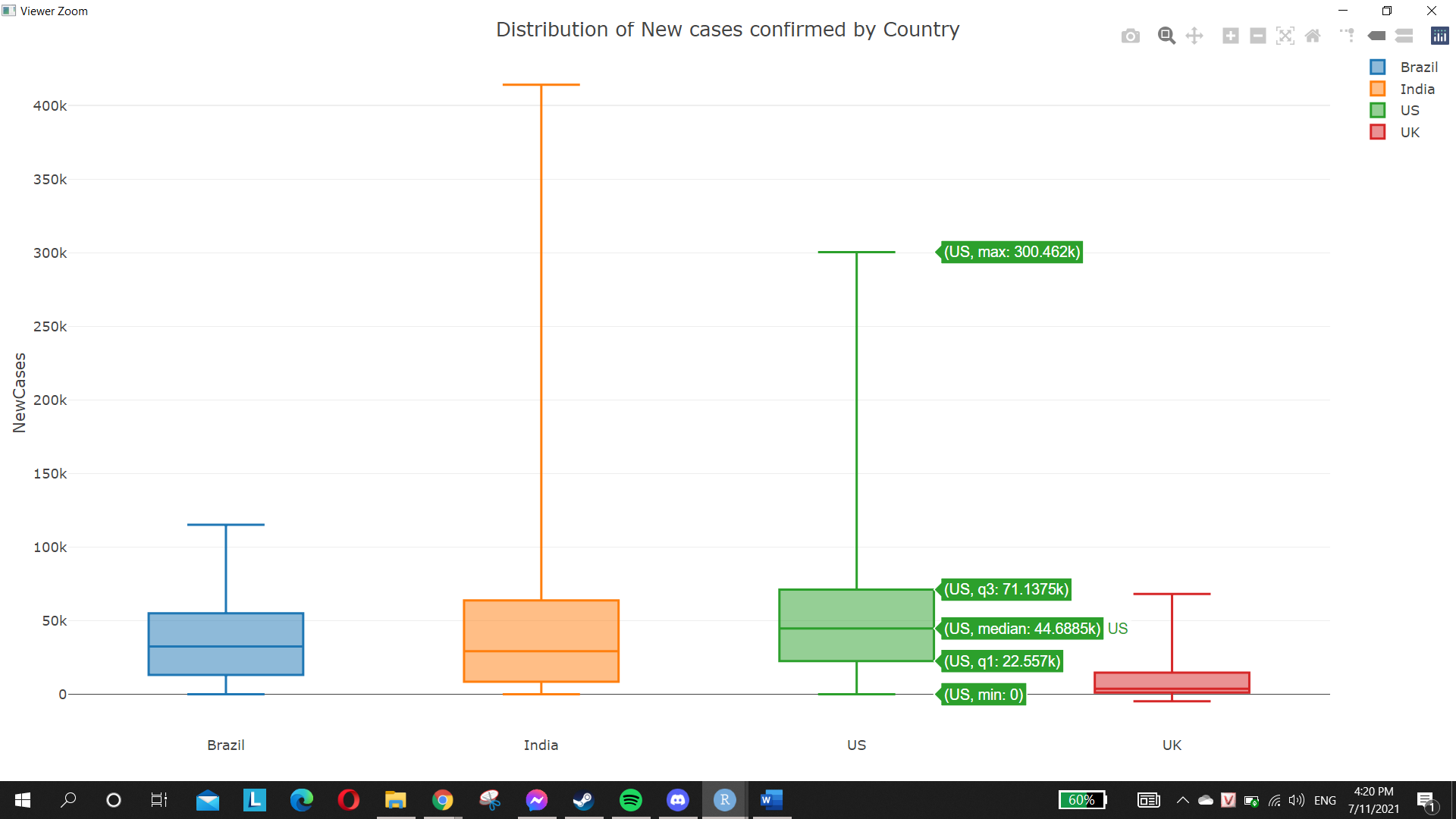
* 1. **Boxplot**



Mô tả đồ thị/biểu đồ:

- Boxplot thể hiện sự phân bố của dữ liệu (Số ca mắc ghi nhận mỗi ngày tại Brazil)  
 - Giá trị trung vị (Median) : 32.446 ca -> Trung bình mỗi ngày có khoảng 32.000 ca mắc tại Brazil  
 - Số ca nhiễm tối đa ghi nhận được trong 1 ngày là 115.228 ca  
 - Khoảng số phân tử = Q3 - Q1 = 55.047 – 13.091 = 41.956  
 - Nhìn chung, số ca mắc ở Brazil có sự dao động lớn, không ổn định, kiểm soát dịch kém.

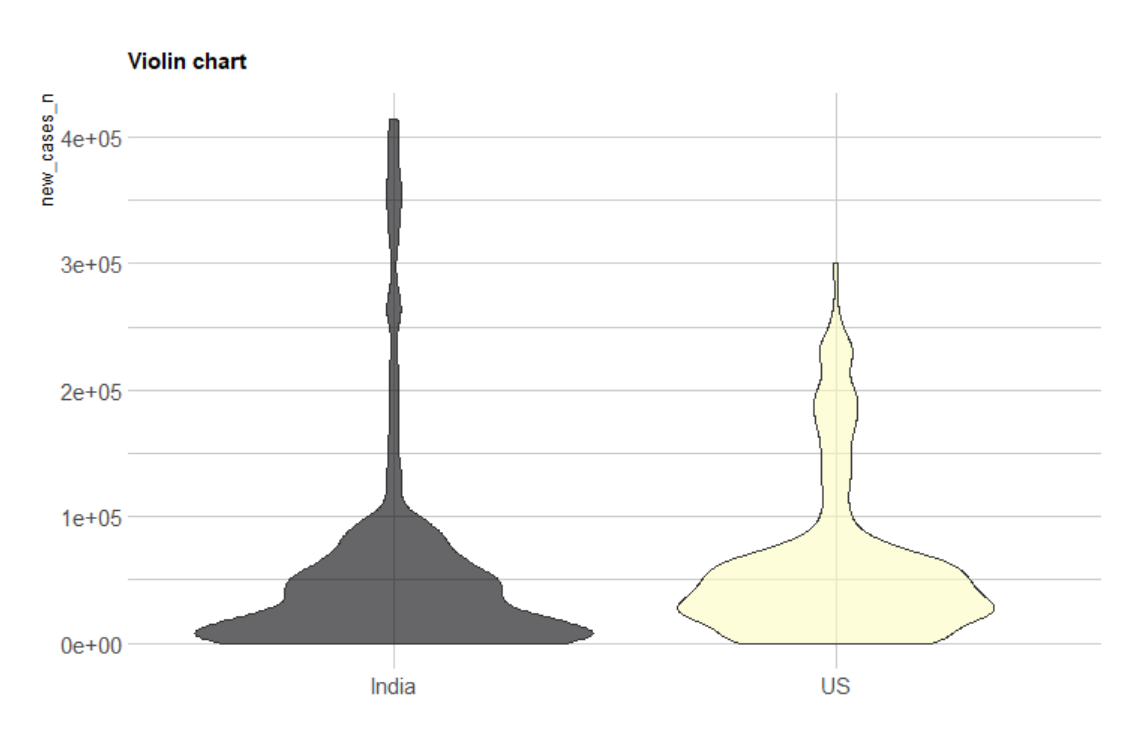
* 1. **Multiple Boxplot**



Mô tả đồ thị/biểu đồ:

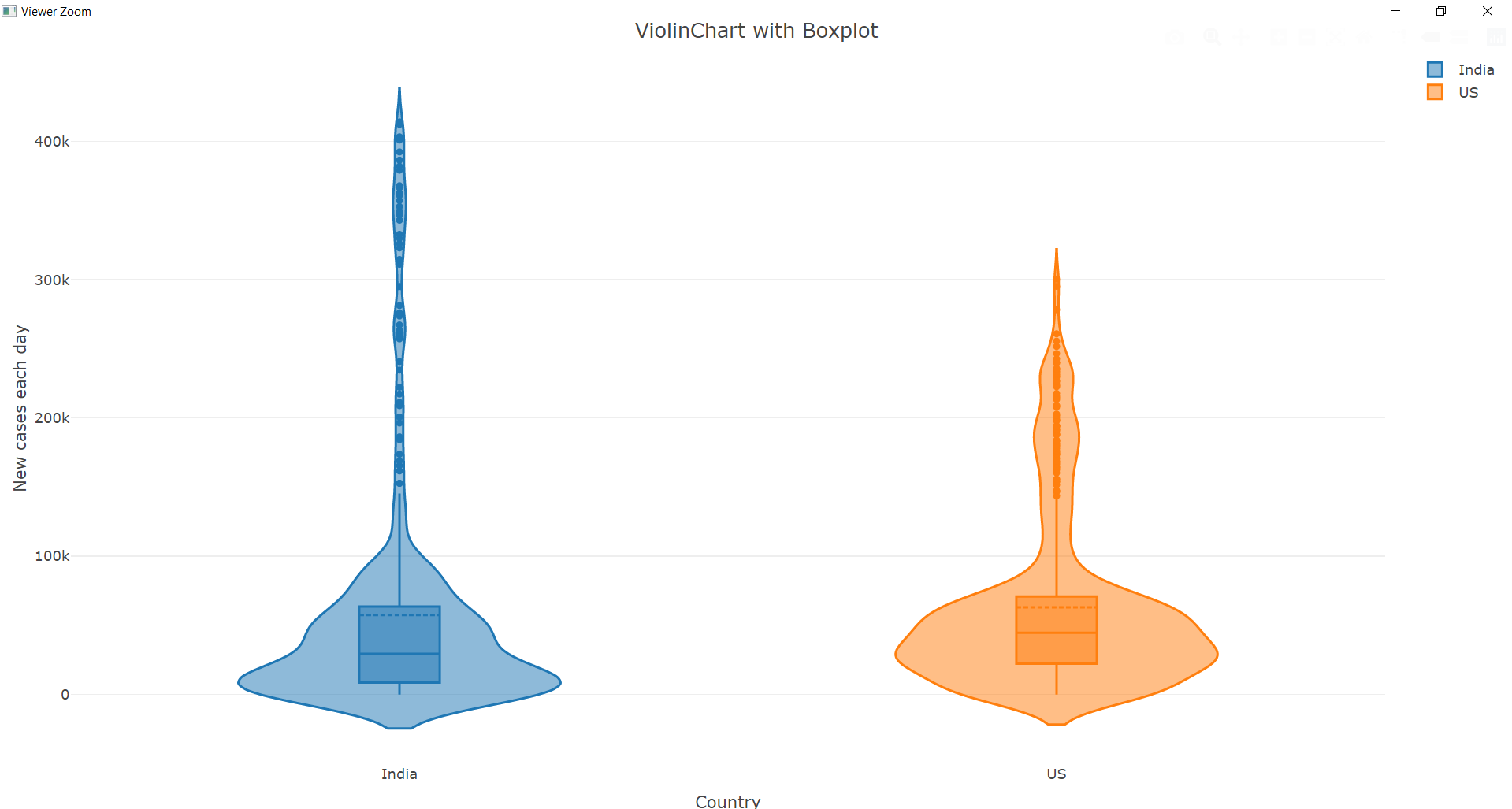
- Boxplot thể hiện sự phân bố của dữ liệu (Số ca mắc ghi nhận mỗi ngày tại 4 nước: Brazil, Ấn Độ, Mỹ và Anh)  
 - Phân bố của dữ liệu chủ yếu tập trung ở phần box (phần hộp)  
=> Các nước có số ca mắc trung bình khoảng 0-50.000 ca ghi nhận mỗi ngày  
 - Số ca nhiễm môĩ ngày của Anh có xu hướng ít hơn Brazil, Ấn Độ và Mỹ  
 - Ngoài ra đỉnh mỗi box thể hiện số ca mắc tối đa ghi nhận được trong 1 ngày của từng nước, Ấn độ là nước ghi nhận được số ca mắc trong ngày cao nhất. Ấn Độ là nước có số ca ghi nhận tăng đột biến (outliner), vượt 400.000 ca  
 - Số ca nhiễm ở Brazil phân bố đều hơn các nước còn lại (ít giá trị outliner)

* 1. **Violin Chart**

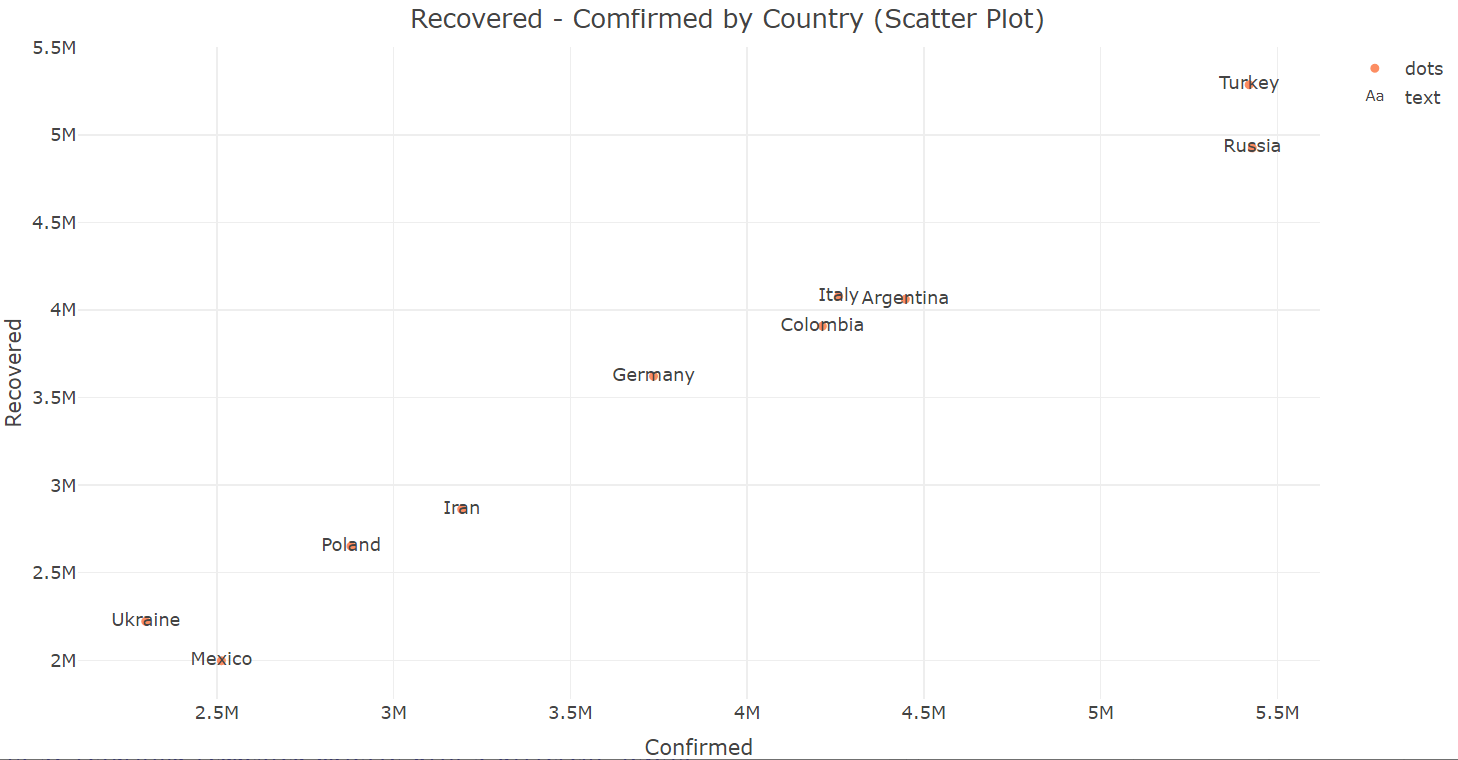


Mô tả đồ thị/biểu đồ:

- Biểu đồ Violin thể hiện sự phân bố của lượng ca nhiễm mới ở Ấn Độ và Mỹ.  
- Phần bị phình ra ở mỗi hình thể hiện nơi tập trung dữ liệu.  
- Ở Ấn chủ yếu mỗi ngày tăng khoảng 10.000 ca.   
- Còn ở Mỹ chủ yếu mỗi ngày tăng khoảng 25.000 ca.  
- Ấn Độ có ngày tăng đột biến (lên đến 400.000 ca mới mỗi ngày)

  
# Kết hợp Violin Chart và BoxPlot

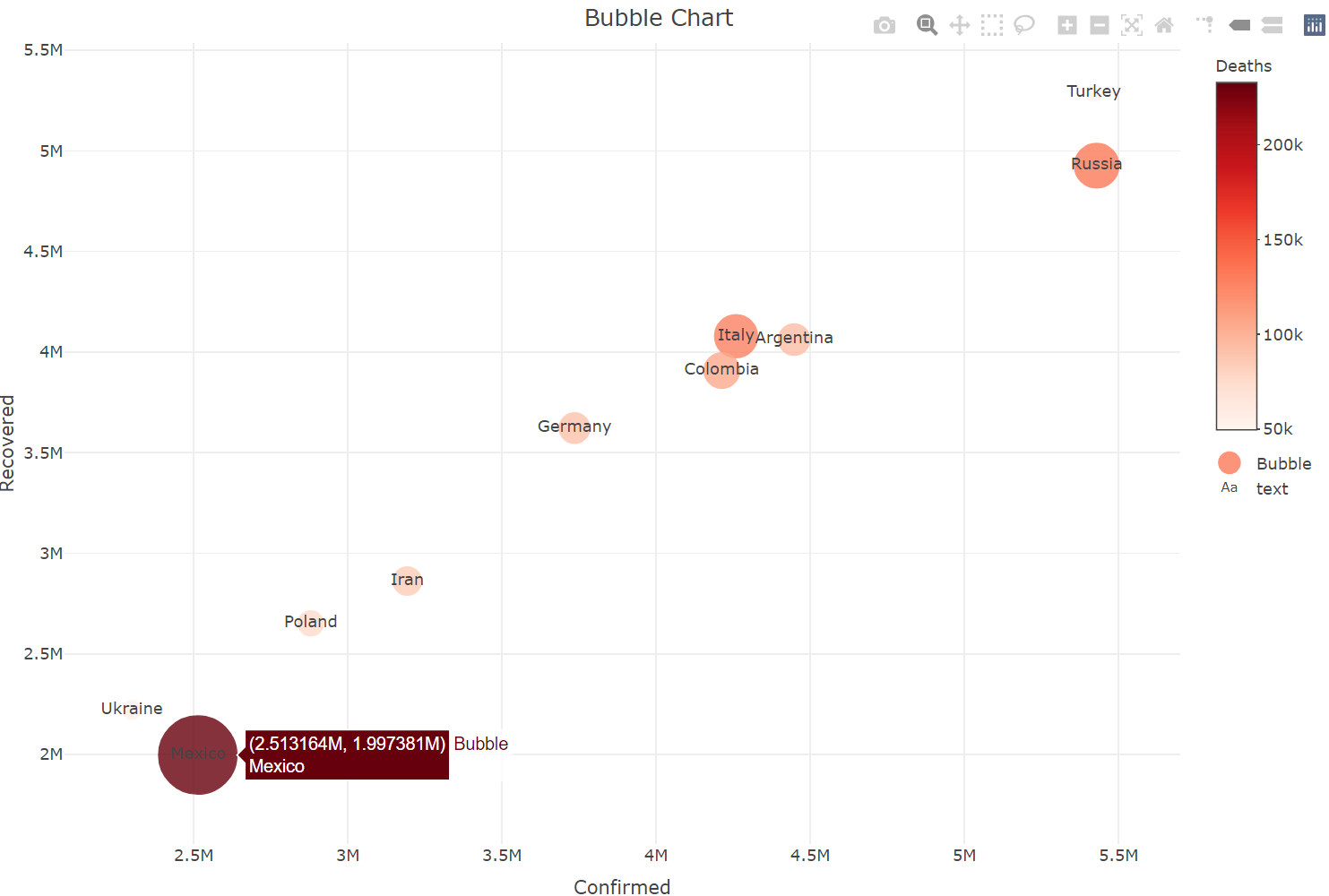
* 1. **Scatter Plot**



Mô tả đồ thị/biểu đồ:

- Biểu đồ Scatter Plots thể hiện mối quan hệ giữa “Số ca nhiễm” và “Số ca hồi phục” của 1 số nước.  
- Trục x ứng với số ca nhiễm đã ghi nhận (đơn vị hàng triệu), trục y ứng với số ca hồi phục.(đơn vị hàng triệu)  
- Trong các nước được biểu diễn, Thổ Nhĩ Kì là nước có số ca nhiễm cao (trên 5 triệu ca) và cũng là nước có số ca hồi phục cao nhất => Thổ Nhĩ Kì có phương pháp chưa trị tốt  
- Hầu như các nước đều có tỉ lệ hồi phục gần bằng số ca nhiễm   
- Riêng Mexico có sự chênh lệch lớn giữa số ca nhiễm và số ca hồi phục.

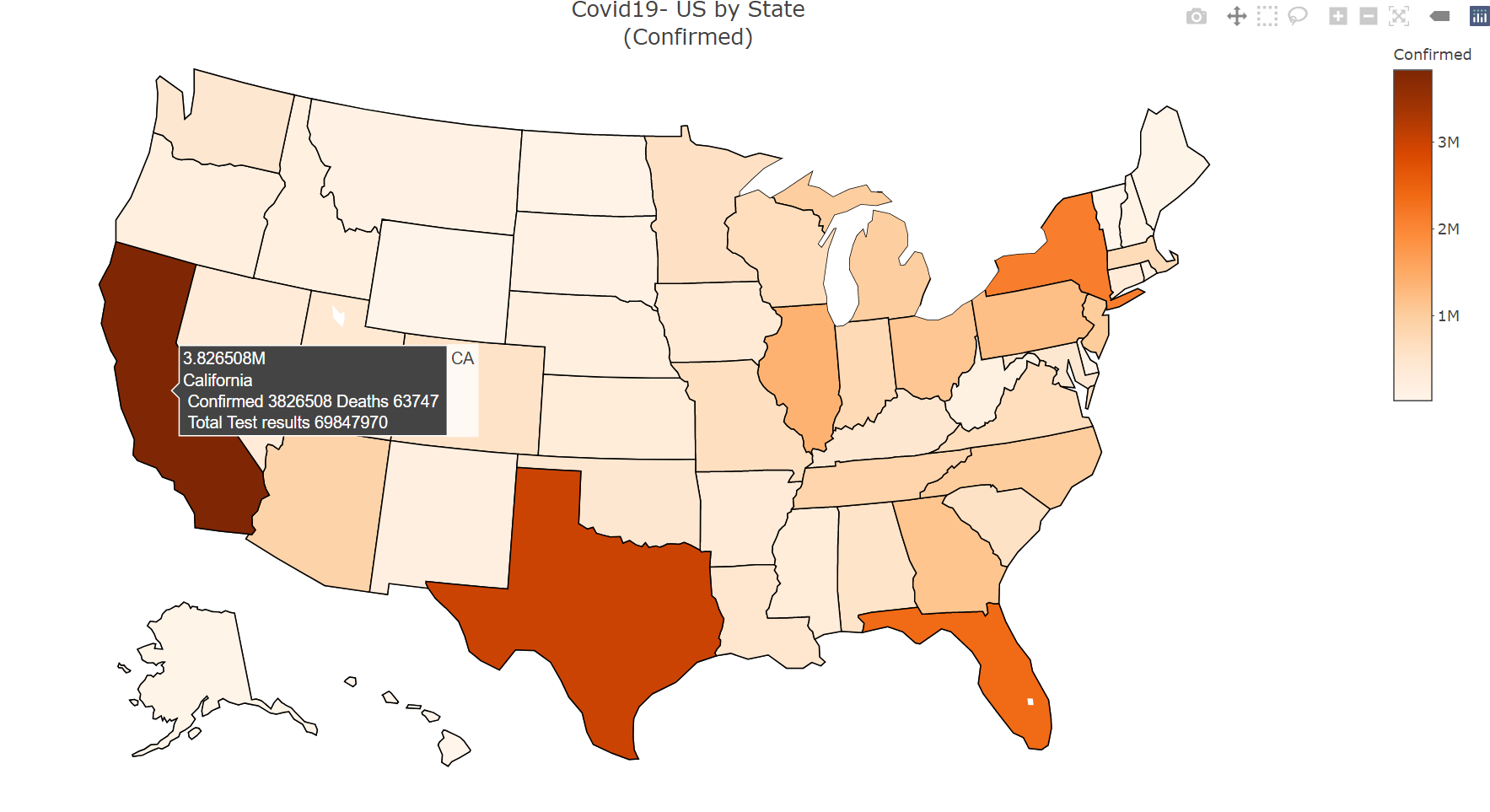
* 1. **Bubble Chart**



Mô tả đồ thị/biểu đồ:

- Biểu đồ Bubble Chart thể hiện mối quan hệ giữa “Số ca nhiễm” ,“Số ca hồi phục” và “Số ca tử vong” của 1 số nước.  
- Trục x ứng với số ca nhiễm đã ghi nhận (đơn vị hàng triệu), trục y ứng với số ca hồi phục.(đơn vị hàng triệu), Dải màu thể hiện “số ca tử vong”  
- Số lượng ca tử vong càng nhiều, kích thước của hình tròn càng lớn.  
- Trong các nước được biểu diễn, Mexico có số ca tử vong lớn nhất (trên 200.000 ca)  
- Hầu như các nước được biểu diễn có số ca tử vong khoảng dưới 100.000ca

* 1. **Choropleth Map**



Mô tả đồ thị/biểu đồ:

- Biểu đồ Choropleth Map thể hiện số ca nhiễm theo từng bang của nước Mỹ.   
- Màu sắc thể hiện số lượng người chết ở mỗi bang được quy định ở thanh bên phải biểu đồ.  
- Hầu hết các bang có khoảng dưới 2 triệu ca nheiemx đã ghi nhận.  
- Các Bang có số ca nhiễm lớn chủ yếu là những bang có các thành phố lớn phía Nam và nổi tiếng (Cali, NewYork, Texas, Florida)  
- California là Bang có số ca nhiễm lớn nhất. (trên 3.8 triệu ca)