TRỰC QUAN HÓA DỮ LIỆU VỚI TABLEAU

Ngày 17 tháng 4 năm 2022

1 Giới thiệu thành viên

Họ và tên	Mã số sinh viên	Email
Kiều Vũ Minh Đức	18127080	18127080@student.hcmus.edu.vn
Nguyễn Công Anh Khoa	18127261	18127261@student.hcmus.edu.vn
Nguyễn Thái Tiên	19127575	19127575@student.hcmus.edu.vn

2 Bảng phân công công việc

Người Thực Hiện	Nhiệm Vụ	Mức Độ Hoàn Thành
Kiều Vũ Minh Đức	Giới thiệu về Tableau	100%
	Các tính năng của Tableau	100%
	2.1 Map	100%
	3.2 Line chart	100%
	4.2 Reduce	100%
Nguyễn Công Anh Khoa	2.2 Bar chart	100%
	2.3 Heat map	100%
	2.4 Packed bubble	100%
	3.1 Bar chart	100%
	3.4 Scatter Chart	100%
	6.5 Data Reduction	100%
Nguyễn Thái Tiến	4.1 Manipulate View	100%
	5.1 OLS	100%
	5.2 Hồi quy tuyến tính	100%
	5.3 ADF Test	100%
	5.4 Arima	100%

3 Tìm hiểu công cụ Tableau

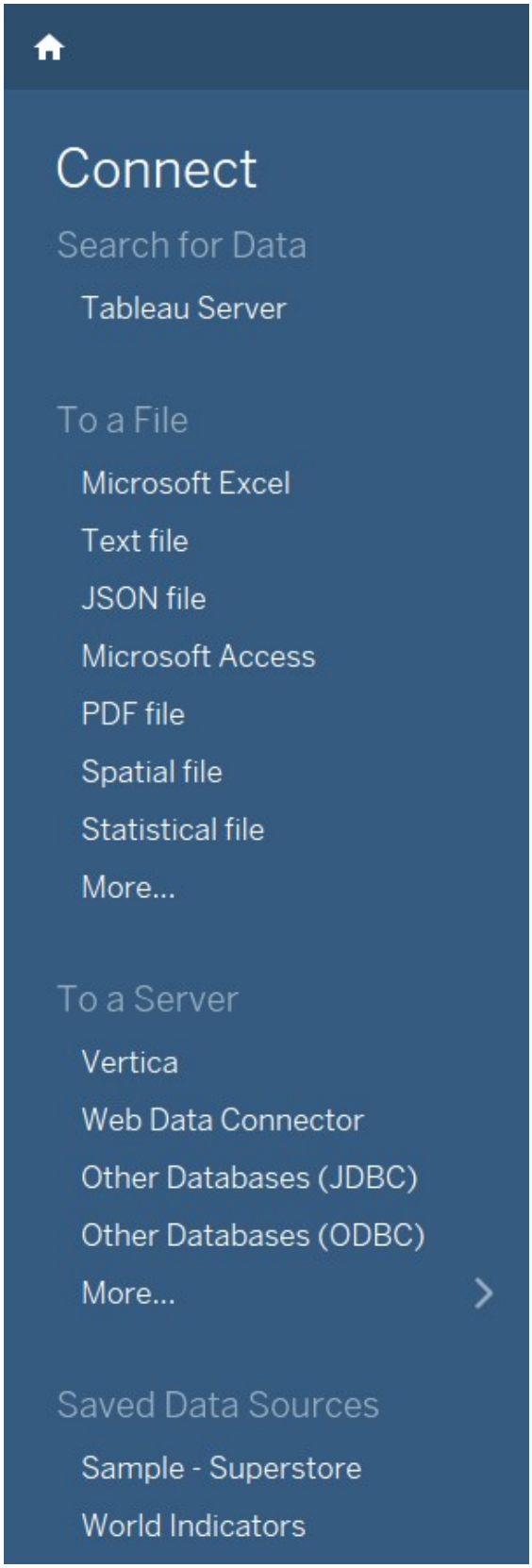
3.1 Giới thiệu về Tableau

Được phát triển vào năm 2003 là kết quả của dự án khoa học máy tính tại đại học Stanford nhằm mục đích cải thiện luồng của phân tích và thông qua trực quan để mọi người có thể tiếp cận dữ liệu . Là một công cụ phát triển nhanh chóng và mạnh mẽ trong lĩnh vực trực quan hóa dữ liệu . Tableau là business intelligence tool giúp phân tích và xây dựng nền tảng số cho doanh nghiệp . Nó bao gồm cả machine learning , thống kê , xử lí ngôn ngữ tự nhiên và chuẩn bị dữ liệu thông minh giúp tăng khả năng sáng tạo của con người trong phân tích .

3.2 Các tính năng hỗ trợ của Tableau

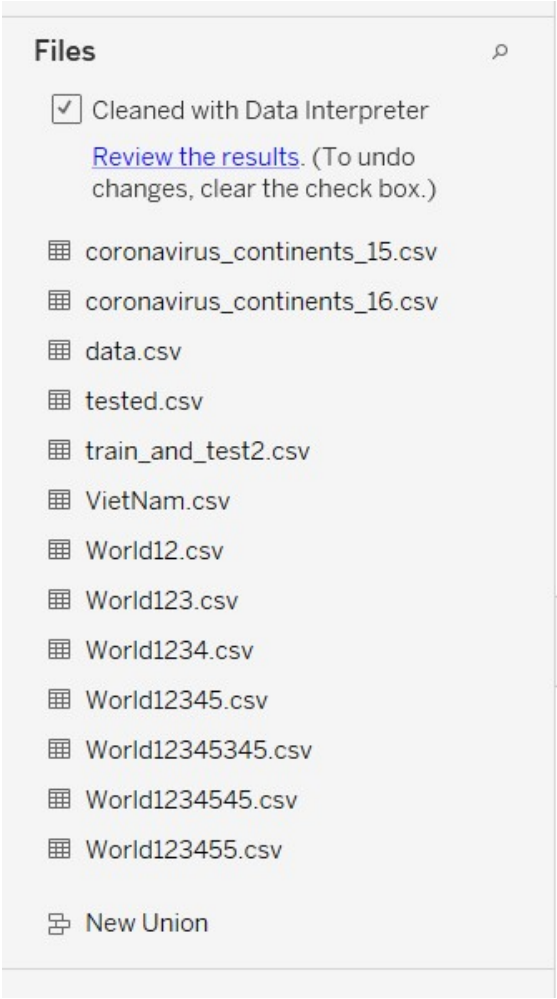
- Kết nối nhiều nguồn dữ liệu. Tableau cung cấp khả năng kết nối với nhiều nguồn dữ liệu:
  - Dữ liệu ở dạng tệp : Microsoft Excel , Text file , JSON file, ...

– Dữ liệu được lưu trữ trên server : Vertica , Web Data Connector , JDBC ,...
– Dữ liệu đã được lưu trữ.



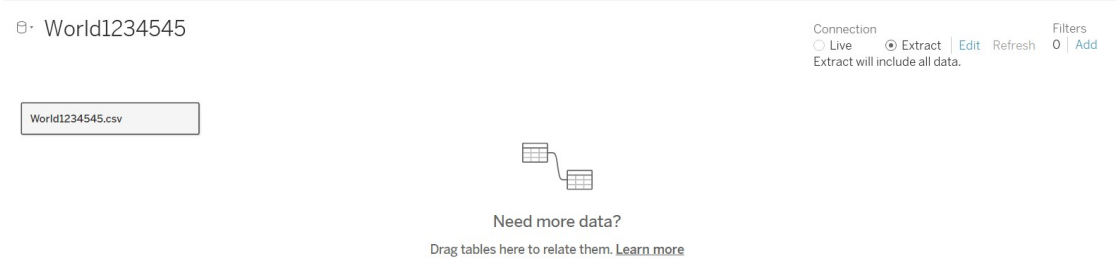
Hình 1: Tính năng kết nối nhiều nguồn khác nhau.

- Hỗ trợ công cụ để tiền xử lí dữ liệu cho việc phân tích.
- Data Interpreter có thể giúp làm sạch dữ liệu. Nó có thể phát hiện những thứ như tiêu đề, ghi chú, chân trang, ô trống, v.v. và bỏ qua chúng để xác định các trường và giá trị thực tế trong tập dữ liệu của bạn.



Hình 2: Data Interpreter.

- Lọc dữ liệu từ Data Source : để tạo một trích xuất từmột nguồn dữ liệu đã có sẵn các bộ lọc nguồn dữ liệu, các bộ lọc đó sẽ tự động được đề xuất làm bộ lọc trích xuất và sẽ xuất hiện trong hộp thoại Extract.



Hình 3: Lọc dữ liệu từ Data Source.

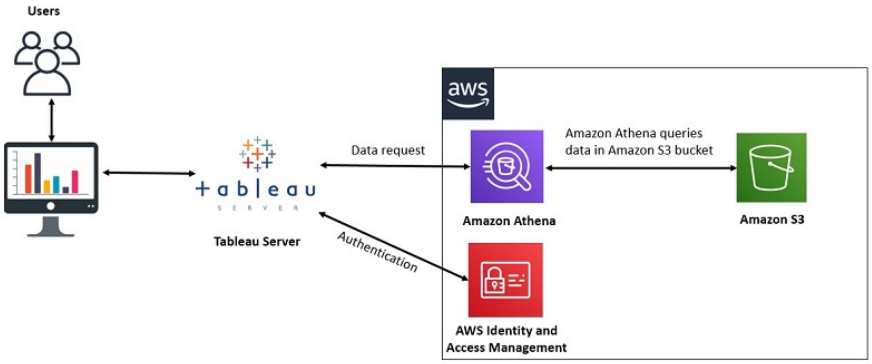
- Bằng các kéo thả đơn giản có thể dễ dàng tạo ra các phân tích dữ liệu. Tableau hỗ trợ thao tác kéo thả giúp người sử dụng có thể thao tác linh hoạt và dễ dàng hơn với các tác vụ.

Hình 4:
Thao
tác
kéo
thả.

- Chia sẻ các phân tích dữ liệu.
  Khi bạn hoặc công ty của bạn không sử dụng Tableau Server hoặc muốn học hỏi và chia sẻ những kết quả hoàn thành được cho người khác, Tableau Public là một lựa chọn tốt. Ngược lại , để đảm bảo sự bảo mật khi người được chỉ định mới được xem hoặc chỉnh sửa thì hãy chọn Tableau Server.

Hình 5: Chia sẽ dữ liệu.

- Linh động chọn môi trường triển khai.
  Được triển khai với on-premise, on-cloud với full cloud trên nền tảng của Tableau hoặc các public cloud của Amazon, Google hoặc Microsoft . Một nền tảng có khả năng tích hợp với Linux, Windows, AWS, Azure.



Hình 6: Linh động trong môi trường sử dụng.

- Vận hành tốt trên các thiết bị di động.
  Với sự phát triển mạnh mẽ Tableau hỗ trợ trên thiết bị di động cả IOS và Android giúp luôn có

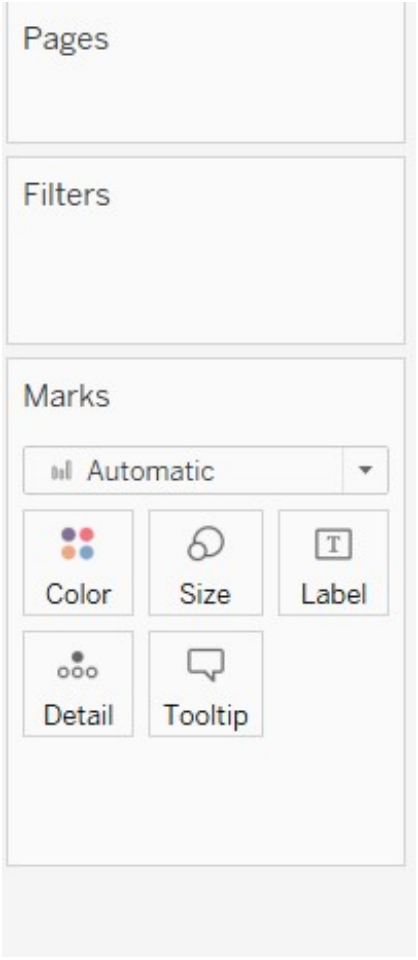
thông tin dữ liệu và phân tích mọi lúc mọi nơi .Ngay cả khi ngoại tuyến thì Tableau Mobile cho phép bạn tương tác với nội dung trên trang web của mình và khám phá thông tin chi tiết về dữ liệu.

Hình 7: Hỗ trợ nhiều thiết bị.

- Tạo các tương tác lọc , rút trích , drill-up , drill-down , hoặc các tham số xử lý dữ liệu ngay trên các biểu đồ.
  Tableau cung cấp nhiều loại biểu đồ như bar chart, packed bubbles , Gantt , are chart , ... giúp người dùng có nhiều góc nhìn trên tập dữ liệu.

Hình 8: Tương tác nhiều loại biểu đồ.

Bên cạnh đó để bổ trợ thêm ý nghĩa của đồ thị, Tableau cũng hỗ trợ nhiều filter khác nhau để hỗ trợ việc trực quan cho biểu đồ .

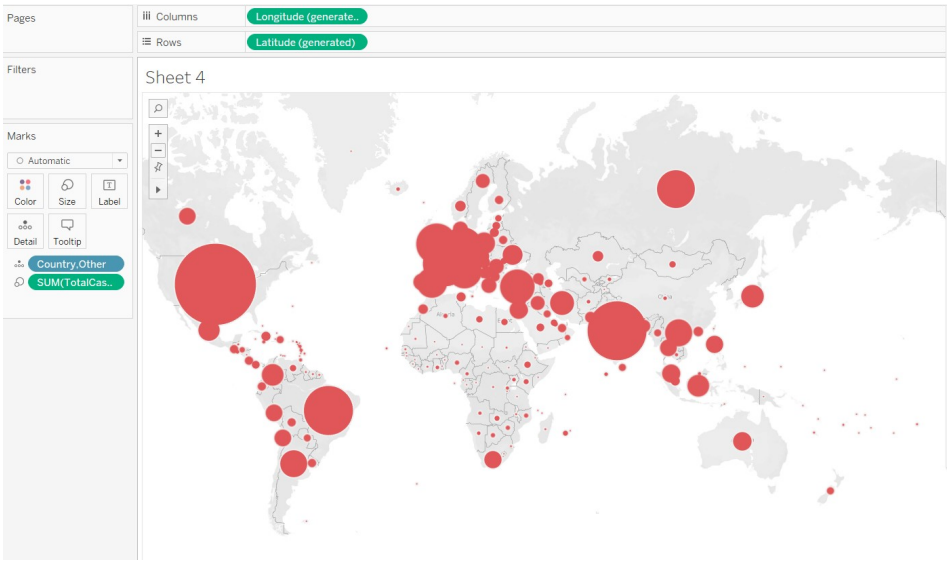


Hình 9: Các filter khác nhau hổ trợ cho trực quan biểu đồ.

4 Trực quan hóa dữ liệu Woldometer với Tableau

4.1 Map

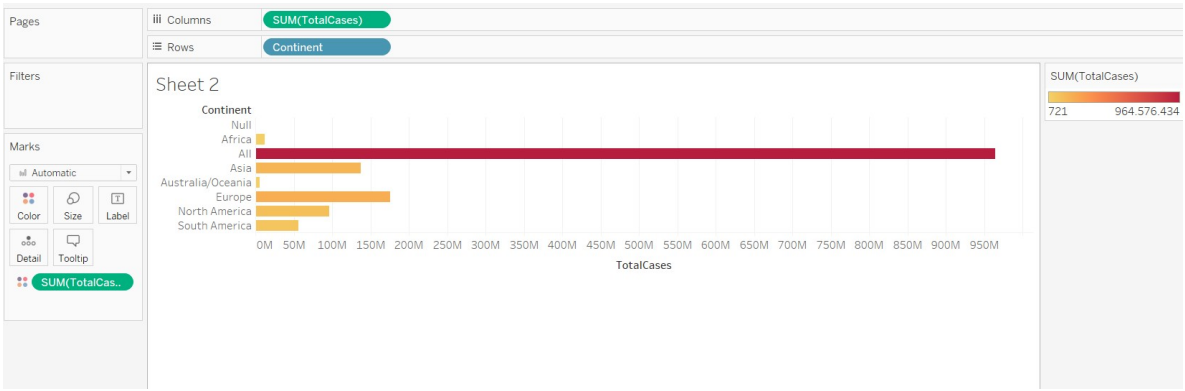
- Trực quan tổng số ca covid của từng quốc gia ngày 27-03. Sử dụng trường dữ liệu là Country,Other và TotalCases.
- Kích thước của mỗi điểm dữ liệu thuộc quốc gia sẽ to nhỏ tương ứng với số lượng ca nhiễm covid.Màu đỏ để thể hiện thêm độ nguy hiểm của dịch bệnh.
- Từ hình ta có thể dễ dàng nhận biết được khu vực có số ca nhiễm cao và thấp.
- Khu vực châu Âu có mật độ quốc gia có số ca nhiễm cao nhất với mức độ dày đặc . Mĩ , Ấn Độ và Brazil lần lượt là 3 quốc gia có tổng số nhiễm cao nhất.
- Các khu vực như Châu Phi , Nam Mĩ và Đông Nam Á có số ca ít hơn và phân bố rải rác.



Hình 10: Biểu đồ map trực quan hóa tổng số ca covid theo từng quốc gia ngày 27-03

4.2 Bar chart

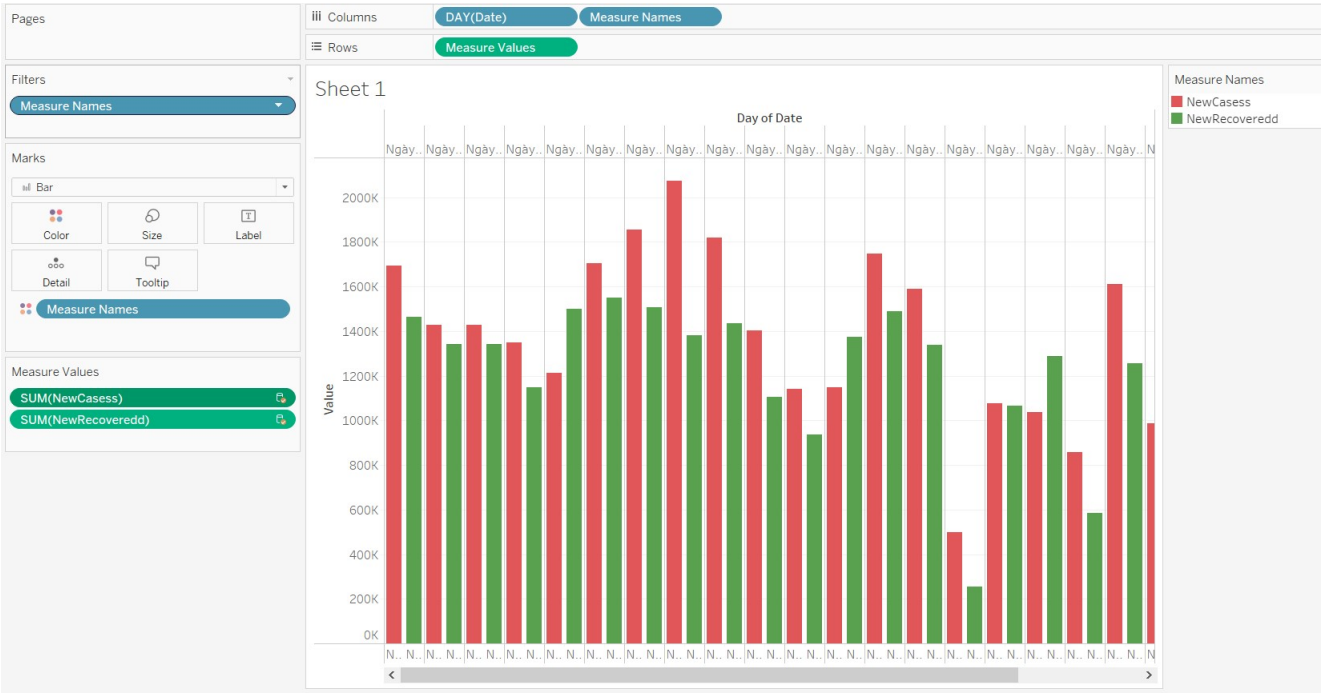
- Biểu đồ thể hiện tổng số ca nhiễm của 6 châu lục và cả thế giới của ngày 27-03 .Sử dụng trường dữ liệu là Continent và Totalcases.
- Cả thế giới đã vượt mốc 950 triệu ca.
- Châu Á và châu Âu có tổng số ca nhiễm cao nhất lần lượt là gần 150 triệu ca và khoảng 175 triệu ca.
- Qua biểu đồ ta thấy tình hình dịch bệnh của châu Úc và châu Phi đỡ căng thẳng hơn các châu lục còn lại với lần lượt là hơn 5 triệu ca và hơn 10 triệu ca.
- Trong khi đó Nam Mĩ có tổng số ca khoảng 55 triệu ca và Bắc Mĩ có hơn 96 triệu ca.
- Sử dụng màu sắc cho trường dữ liệu Totalcases để thể hiện mức độ ca dịch ở mỗi châu lục , tăng dần từ cam đến đỏ.



Hình 11: Biểu đồ cột thể hiện tổng số ca nhiễm theo châu lục ngày 27-03.

- Biểu đồ thể hiện tổng số ca nhiễm theo châu lục và cả thế giới của ngày 13-03. Trường dữ liệu cột là Continent, còn cột là tổng số ca nhiễm và tử vong.
- Africa và Australia là 2 châu lục có số ca nhiễm và tử vong thấp nhất.

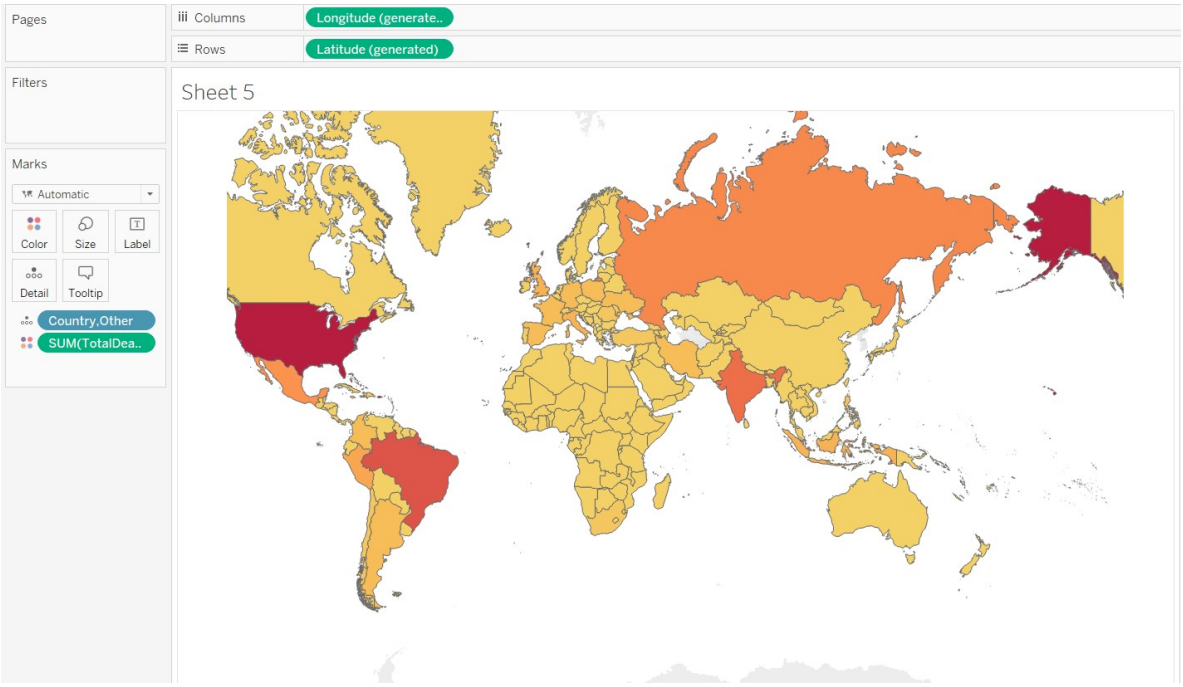
- Europe là châu lục có số ca nhiễm và tử vong cao nhất.
- Tuy Europe và Asia là 2 châu lục có số ca nhiễm nhiều hơn so với North America và South America, nhưng số ca tử vong của 4 châu lục này xêm xêm nhau.



Hình 12: Biểu đồ cột biểu hiện số ca nhiễm và tử vong theo châu lục ngày 13-03

4.3 Heat Map

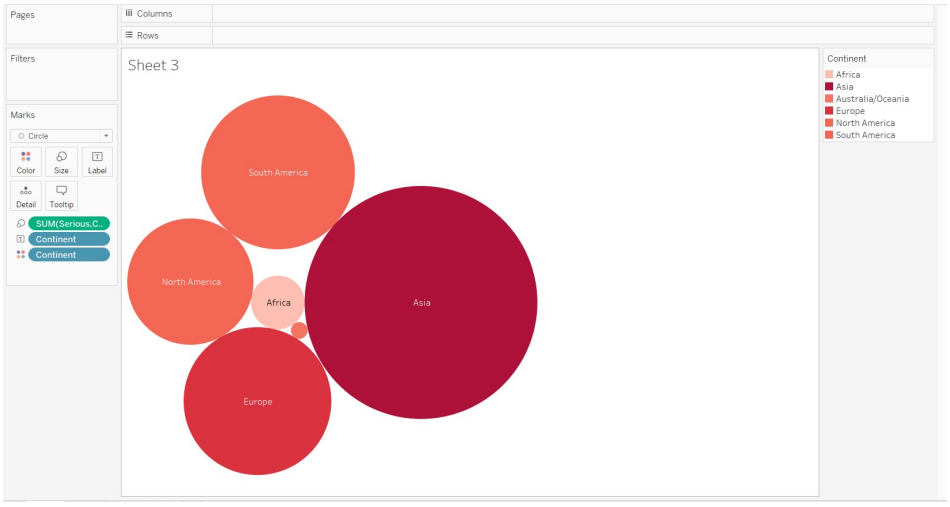
- Biểu đồ thể hiện tổng số ca chết của từng quốc gia ngày 27-03 . Sử dụng màu sắc từ vàng lạt đến đỏ đậm để thể hiện mức độ của mỗi quốc gia.
- Mĩ là quốc gia có tình hình ca chết căng thẳng nhất với màu đỏ đậm . Với HeatMap ta có thể quan sát ca nhiễm của mỗi quốc gia bằng cách rê chuột vào quốc gia đó.
- Tiếp theo đó là Brazil , Ấn Độ , Nga và Mexico lần lượt là 4 nước có tổng số ca chết trong ngày cao.
- Phần còn lại cao nhất ở mức 200 nghìn ca và phổ biến là màu vàng lạt với khoảng vài nghìn ca đến vài chục nghìn ca.



Hình 13: Biểu đồ heat map thể hiện số ca tử vong theo từng quốc gia ngày 27-03.

4.4 Packed bubble

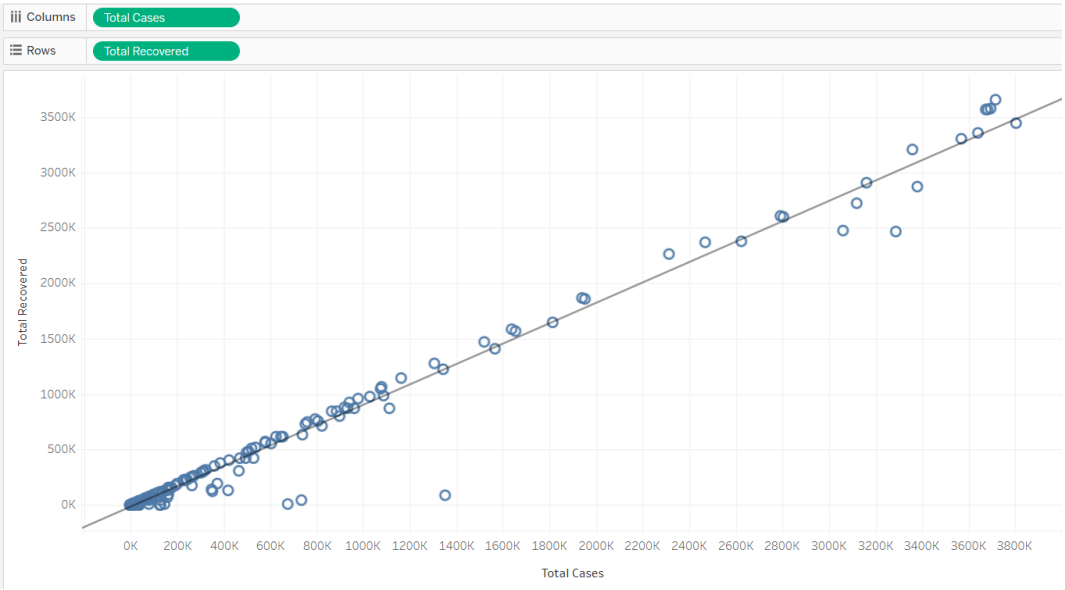
- Biểu đồ thể hiện số ca nghiêm trọng của các châu lục ngày 27-03 .Biểu đồ này được trực quan từ 2 trường dữ liệu là Continent và Serious,Critical .
- Từ kích thước của các hình tròn ta có thể so sánh được số ca nghiêm trọng giữa các châu lục với nhau .
- Châu Á có số ca cao hơn hẳn phần còn lại với khoảng 26 nghìn ca . Tiếp sau đó là Nam Mĩ và châu Âu có số ca nghi kịch lần lượt là khoảng 11.5 nghìn ca và 10.5 nghìn ca
- Châu Úc với số ca ít nhất chỉ khoảng hơn 100 ca .
- Việc sử dụng màu đỏ và lạt dần theo từng mức ca dịch thể hiện mức độ nghiêm trọng của số ca nghi hiểm trong ngày



Hình 14: Biểu đồ thể hiện số ca nghiêm trọng theo châu lục ngày 27-03.

4.5 Scatter Chart

- Scatter Chart thể hiện 2 trường dữ liệu với cột tổng số hồi phục, và dòng là tổng số ca nhiễm trong ngày 13-03.
- Tổng số hồi phục và tổng số ca nhiễm có quan hệ tương quan thuận mạnh

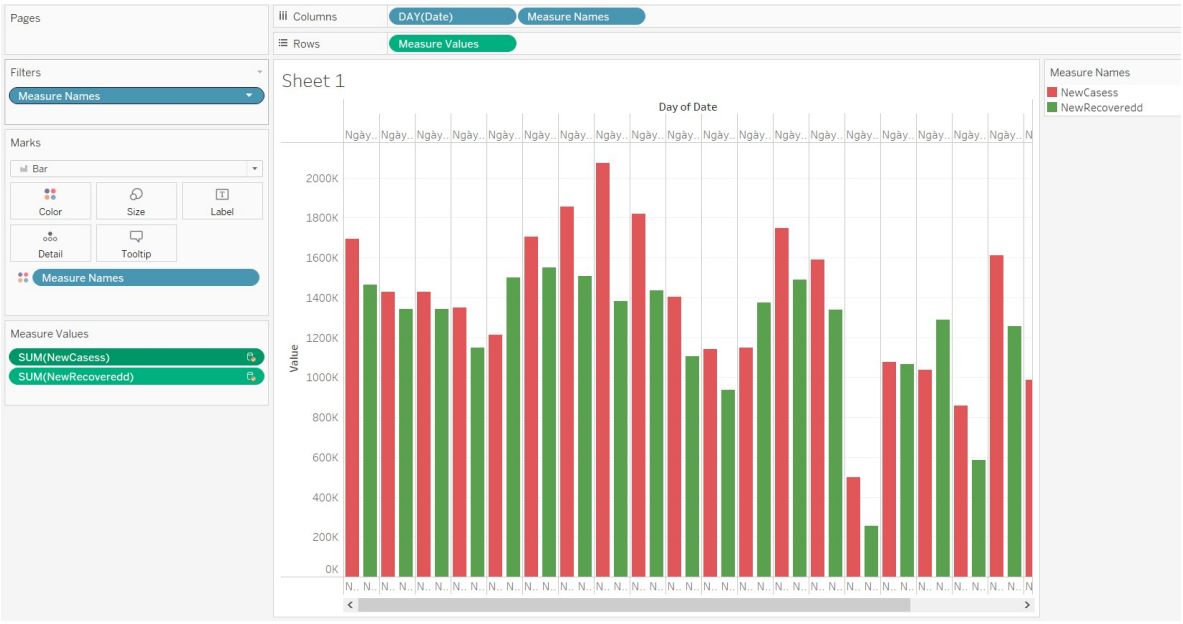


Hình 15: Biểu đồ Scatter thể hiện mối quan hệ giữa số ca nhiễm và số ca hồi phục

5 Thể hiện trực quan một số dữ liệu biến đối qua từng ngày

5.1 Bar chart

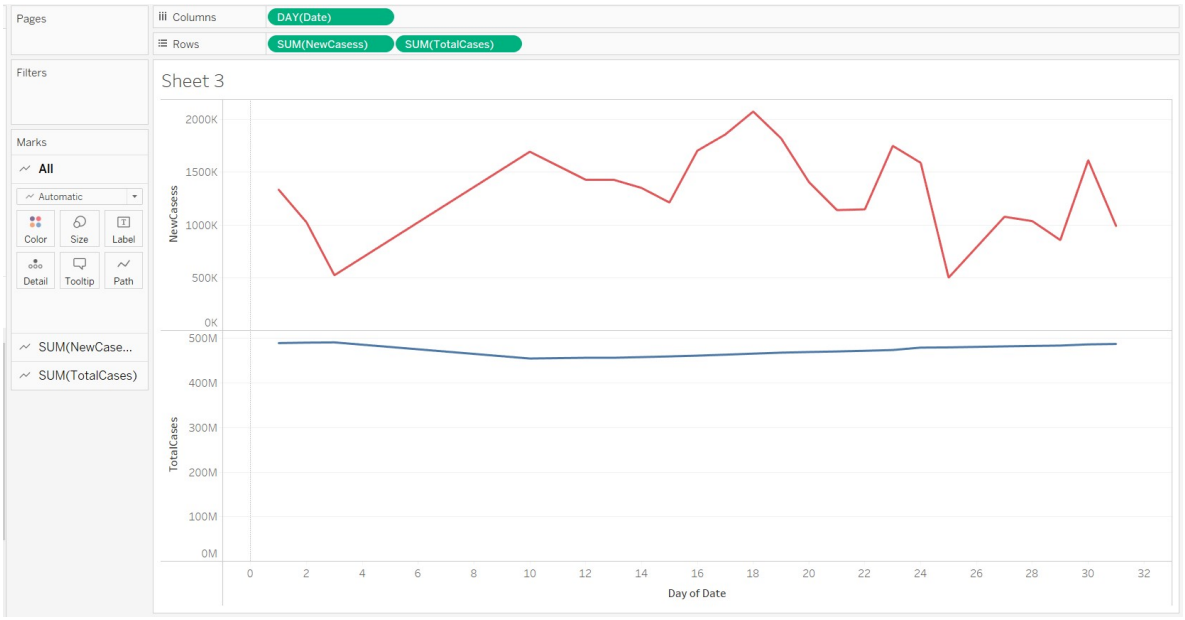
- Biểu đồ cột thể hiện số ca nhiễm và số ca phục hồi trong các ngày từ 10-03 đến 03-04 trên toàn thế giới. Sử dụng 2 trường dữ liệu là NewCases và NewRecovered .
- Từ biểu đồ ta có thể thấy rằng số ca nhiễm trong ngày hầu hết cao hơn số ca phục hồi .Việc sử dụng màu đỏ để biểu diễn cho trường dữ liệu số ca nhiễm mới thể hiện mức độ nghiêm trọng . Màu xanh biểu diễn cho trường dữ liệu số ca phục hồi thể hiện sự khả quan của tình hình dịch bệnh .
- Số ca nhiễm mới tăng mạnh từ ngày 16-03 đến ngày 19-03. Đỉnh điểm trong ngày 18-03 thì số ca nhiễm mới chạm mốc 2 triệu ca một ngày .
- Ngày 28-03 và ngày 02-04 là 2 ngày duy nhất mà thế giới có tín hiệu đáng mừng khi số ca phục hồi cao hơn số ca nhiễm mới.
- Ngày 25-03 là ngày có số ca nhiễm mới và số ca phục hồi thấp nhất trong tất cả các ngày .



Hình 16: Biểu đồ cột thể hiện số ca nhiễm và phục hồi từ ngày 10-03 đến 03-04.

5.2 Line Chart

- Line chart biểu diễn số ca nhiễm mới và tổng số ca nhiễm trên thế giới từ ngày 10-03 đến ngày 03-04 . Sử dụng 2 trường dữ liệu TotalCases và NewCases để trực quan line chart .
- Từ line chart màu đỏ phí trên của số ca nhiễm mới ta thấy sự biến động qua từng ngày chứ không có xu hướng nhất định .
- Số ca nhiễm mới được ghi nhận thấp nhất trong ngày là ngày thứ 3 và ngày thứ 25 với khoảng 500 nghìn ca .
- Trong khi đó line chart ở dưới thể hiện tổng số ca qua từng ngày thì ít biến động hơn . Miền giá trị của nó nằm trong khoảng từ 450 triệu ca đến hơn 500 triệu ca .

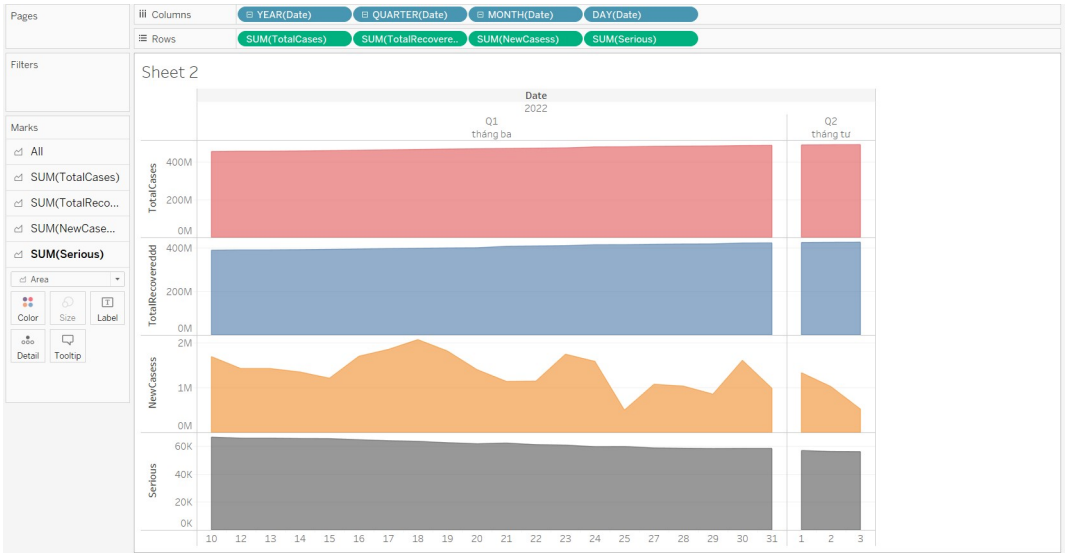


Hình 17: Biểu đồ đường biểu diễn số ca nhiễm mới và tổng số ca nhiễm trên thế giới từ 10-03 đến 03-04.

5.3 Area Chart

- Area Chart thể hiện 4 trường dữ liệu TotalCases , TotalRecovered , NewCases , Serious .

- Đồ thị được chia 2 phần tương ứng với tháng 3 và tháng 4 .
- Màu sắc của mỗi trường dữ liệu được thể hiện phù hợp với tính chất ý nghĩa của nó . Màu xanh thể hiện cho số ca phục hồi , màu xám để thể hiện tính chất không mấy sáng sủa của số ca nghi kịch và màu vàng , đỏ trực quan cho trường dữ liệu số ca nhiễm mới và tổng số ca .
- Ngoại trừ sự biến động theo ngày rõ rệt nhất của số ca nhiễm mới thì 3 trường dữ liệu còn lại thì có xu hướng tăng hoặc giảm .
- Quan sát ô dưới cùng của số ca nghi kịch thấy rằng nó giảm theo ngày mang lại sự tích cực trong tình hình dịch bệnh .



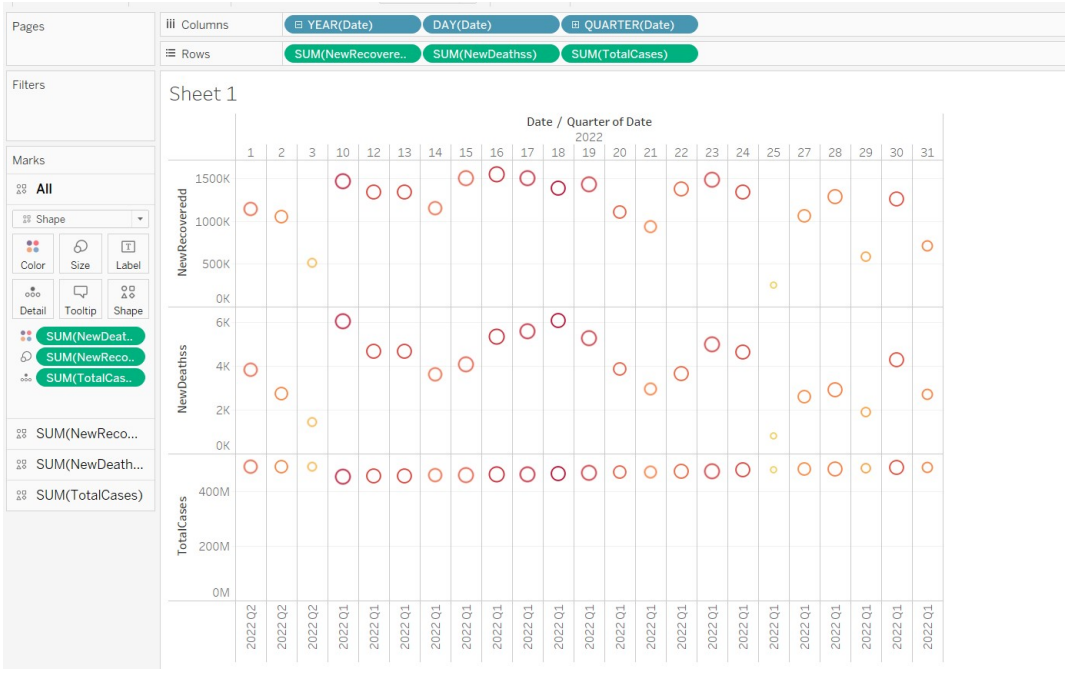
Hình 18: Biểu đồ area chart thể hiện tổng số ca nhiễn, phục hồi, ca mới và ca nghiêm trọng trong tháng 3 và 4.

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

6.1 Manipulate View

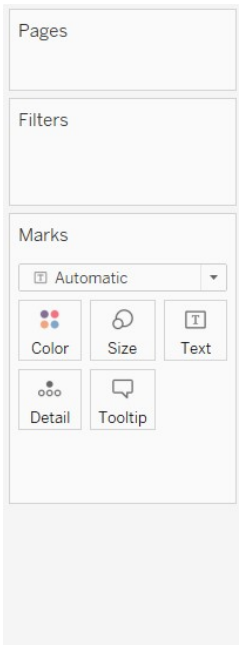
Sử dụng Change view over time của Tableau
. **Lý do lựa chọn :** Nhằm hiển thị nhiều trường dữ liệu và so sánh sự thay đổi của các trường dữ liệu đó theo thời gian . Hữu ích trong việc quan sát và đánh giá phân tích .

Ý nghĩa rút ra : Có thể quan sát được khái quát các trường dữ liệu NewDeaths , NewRecovered , Totalcases trên thế giới thay đổi theo ngày . Cùng với việc sử dụng màu sắc và kích cỡ của các scatter để biết được sự thay đổi cũng như mức độ tình hình dịch bệnh . Ví dụ số ca NewDeaths trong ngày cao nhất là ngày 10-3 .



Hình 19: 3 Biểu đồ Scatter thể hiện mối quan hệ giữa tổng số ca nhiễm , số ca mất mới và số ca hồi phục mới trên thế giới theo từng ngày .

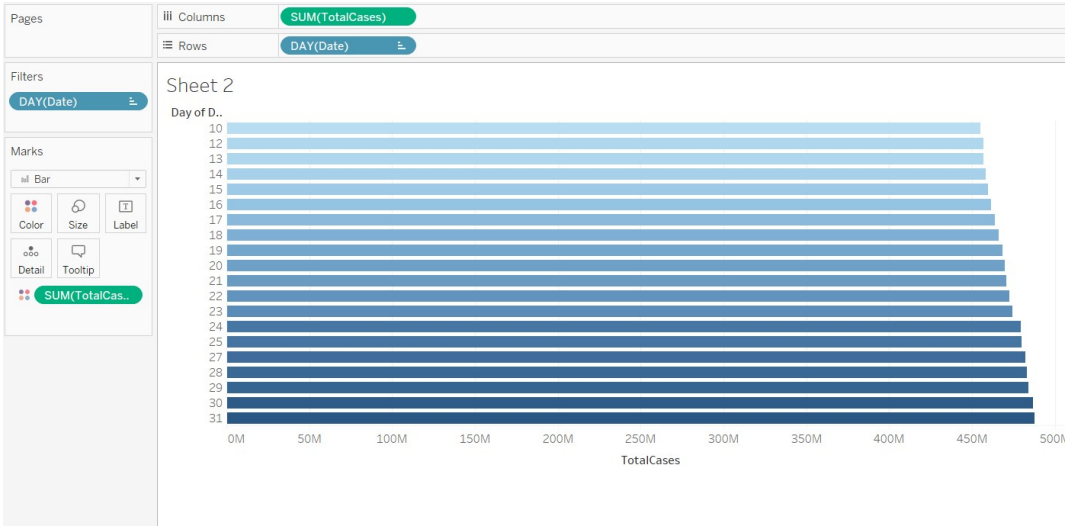
6.2 Reduce



Hình 20: Filter trong Tableau.

Tableau hỗ trợ kĩ thuật Reduce với chức năng Filter để lựa chọn , lược bỏ các giá trị dữ liệu để hỗ trợ việc trực quan tốt hơn và thể hiện được ý muốn của người dùng .
Lý do lựa chọn : Bằng những thao tác đơn giản như kéo thả và lựa chọn để lược bỏ đi những giá trị không mong muốn hiển thị trực quan .

Từ biểu đồ trên khi được lược đi dữ liệu ngày của tháng 4 chỉ hiển thị dữ liệu của những ngày tháng 3 . Sử dụng bar chart để trực quan tổng số ca trên thế giới trong những ngày tháng 3 . Độ đậm



Hình 21: Filter trong Tableau.

nhạt của màu sắc để thể hiện độ lớn của tổng số ca thay đổi theo ngày .

Ý nghĩa rút ra : Với những thao tác đơn giản được Tableau hỗ trợ có thể nhanh chóng lược đi những giá trị dữ liệu mình mong muốn . Qua hình có thể thấy số ca nhiễm trên thế giới tăng theo từng ngày khoảng từ 1 đến 2 triệu ca

7 Sử dụng thuật toán học máy

7.1 Bình phương nhỏ nhất thông thường (OLS - Ordinary Least) Squares

Là một phương pháp để ước lượng các tham số chưa biết trong mô hình hồi quy tuyến tính. Phương pháp này giảm thiểu tổng các khoảng cách theo chiều dọc bình phương giữa các câu trả lời được quan sát trong tập dữ liệu và các câu trả lời được dự đoán bằng phép gần đúng tuyến tính .

Trong phần này sẽ sử dụng thuật toán OLS để quan sát được mối quan hệ giữa 3 biến ActiveCases , Seriours và NewDeaths . Với biến phụ thuộc là NewDeaths .

- R-squared có thể là phép đo quan trọng nhất được tạo ra bởi bản tóm tắt này. R-squared là phép đo mức độ của biến độc lập được giải thích bằng những thay đổi trong các biến phụ thuộc của chúng ta. Tính theo tỷ lệ phần trăm, 0.955 có nghĩa là mô hình của chúng giải thích 95.5 % sự thay đổi trong biến NewCases .

- Adj. R-squared: cho thấy sự phù hợp của mô hình , quan sát hình bên dưới cho thấy các biến độc lập là giải thích 94.7% biến phụ thuộc , một kết quả khá ấn tượng .

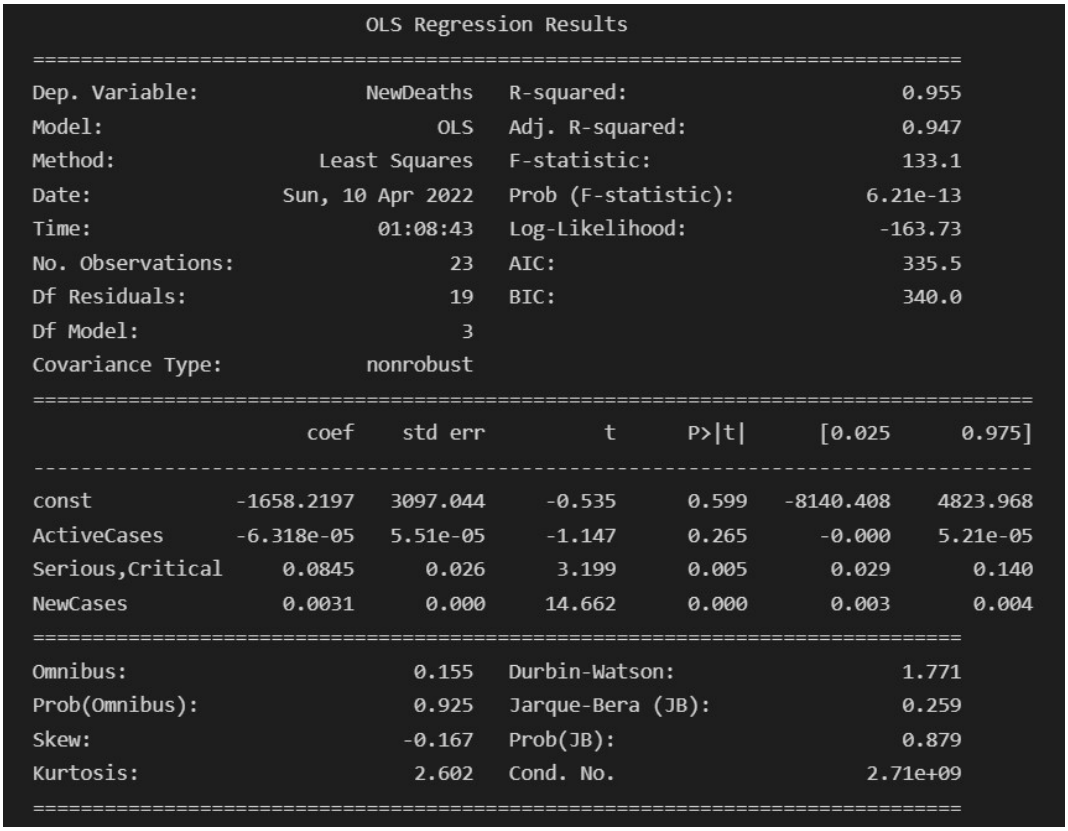
- Giá trị p-value là một trong những số liệu thống kê quan trọng nhất . Ví dụ p-value của ActiveCases là 0.265 cho biết 26.5% của biến ActiveCases không ảnh hưởng đến biến phụ thuộc NewDeaths.

- std err : ước lượng độ lệch chuẩn của hệ số , nó càng nhỏ thì mô hình có độ chính xác càng cao .

- coef : Đối với mỗi biến, nó là phép đo mức độ thay đổi của biến đó ảnh hưởng đến biến độc lập.

- 0,025 và 0,975 là phép đo giá trị của các hệ số trong phạm vi 95% dữ liệu hoặc trong hai độ lệch chuẩn . Bên ngoài những giá trị này thường xem là ngoại lệ .

- Const coef : Nếu hai biến độ lập đều bằng 0 thì biến phụ thuộc là 1 const.



OLS Regression Results

Dep. Variable:	NewDeaths	R-squared:	0.955
Model:	OLS	Adj. R-squared:	0.947
Method:	Least Squares	F-statistic:	133.1
Date:	Sun, 10 Apr 2022	Prob (F-statistic):	6.21e-13
Time:	01:08:43	Log-Likelihood:	-163.73
No. Observations:	23	AIC:	335.5
Df Residuals:	19	BIC:	340.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1658.2197	3097.044	-0.535	0.599	-8140.408	4823.968
ActiveCases	-6.318e-05	5.51e-05	-1.147	0.265	-0.000	5.21e-05
Serious,Critical	0.0845	0.026	3.199	0.005	0.029	0.140
NewCases	0.0031	0.000	14.662	0.000	0.003	0.004

Omnibus:	0.155	Durbin-Watson:	1.771
Prob(Omnibus):	0.925	Jarque-Bera (JB):	0.259
Skew:	-0.167	Prob(JB):	0.879
Kurtosis:	2.602	Cond. No.	2.71e+09

Hình 22: Kết quả thuật toán OLS.

7.2 Hồi quy tuyến tính

Hồi quy tuyến tính (Linear Regression) là một thuật toán học máy đơn giản , dễ hiểu và dễ cài đặt. Các bước mô hình hóa bằng Linear Regresion .

- Hai trường dữ liệu được sử dụng trong phần này là TotalCases và TotalDeaths .
- Dữ liệu được chia thành 80% và 20% lần lượt cho tác vụ train và test .
- Sử dụng mô hình có sẵn từ thư viện sklearn.linear_model và sử dụng sklearn.model_selection để chia tập dữ liệu .
- Quan sát 2 cột dự đoán và kết quả thực tế từ tập test ta thấy kết quả dự đoán không quá khác biệt với thực tế . Sự chênh lệch giữa giá trị dự đoán và giá trị thực tế là vài nghìn ca.

```
X = data['TotalCases'].to_numpy().reshape(-1,1)
Y = data['TotalDeaths'].to_numpy().reshape(-1,1)
```
[8] ✓ 0.1s

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.1, random_state=0)
```
[9] ✓ 0.1s

```
regressor = LinearRegression()
regressor.fit(X_train,y_train)
```
[10] ✓ 0.4s

```
LinearRegression()
```

```
predict =  regressor.predict(X_test)
df = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': predict.flatten()})
print(df)
```
[11] ✓ 0.2s

```
      Actual     Predicted
0  6116993.0  6.112983e+06
1  6100688.0  6.107725e+06
2  6173918.0  6.175782e+06
```

Hình 23: Thực nghiệm thuật toán hồi qui tuyến tính.

7.3 Augmented Dickey Fuller test (ADF Test)

ADF Test là một thử nghiệm thống kê phổ biến được sử dụng để kiểm tra một chuỗi thời gian có cố định hay không .Trong phần này ADF Test được sử dụng trên tập dữ liệu về tình hình dịch bệnh của cả thế giới từ ngày 10-3 đến ngày 3-4 .

```
1. ADF :  -0.3705221167534461
2. P-value :  0.9149054099216527
3. Num of Lags :  0
4. Num Of Observcations Used For ADF Regression And Critical Value Calculation :  22
5. Critical Values :
        1%  :  -3.769732625845229
        5%  :  -3.005425537190083
        10%  :  -2.6425009917355373
```

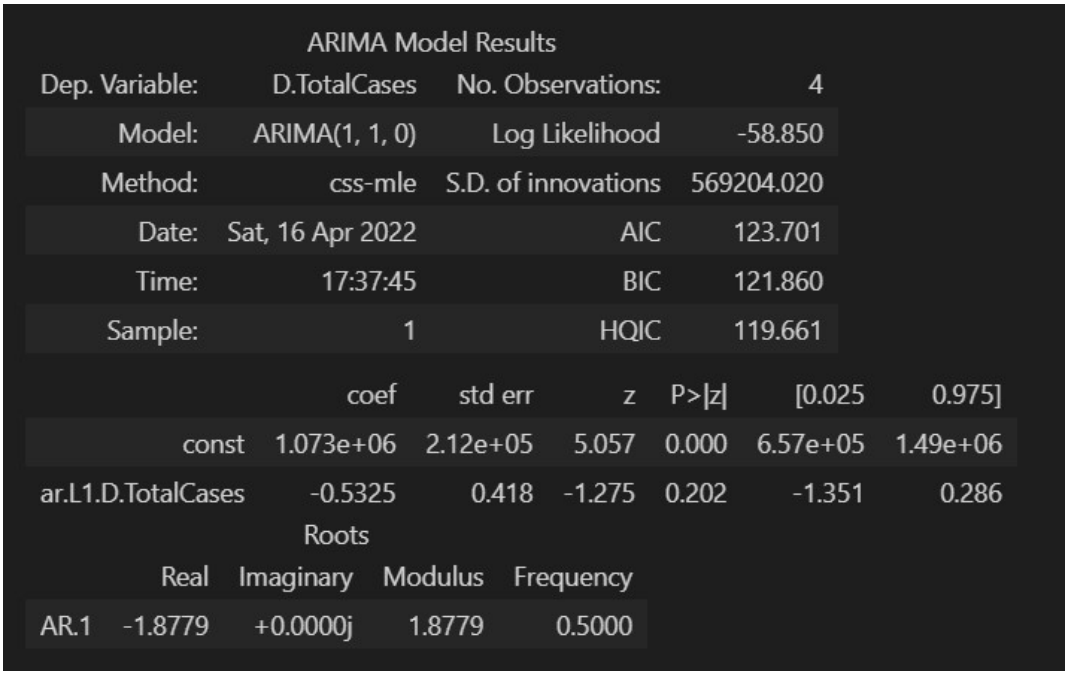
Hình 24: Kết quả thuật toán ADF.

Từ hình trên ta có thể thấy giá trị p-value = 0.915 lớn hơn mức ý nghĩa 0.05 và ADF Statistic lớn hơn tất cả Critical Values nên suy ra time series này là non-stationary.

7.4 Autoregressive Intergrated Moving Average (ARIMA)

Trong phần này sử dụng ARIMA để dự đoán được tổng số ca covid trong tương lai . Mô hình ARIMA sẽ biểu diễn phương trình hồi qui tuyến tính đa biến (multiple regression) với biến phụ thuộc ở đây sẽ là TotalCases .

- p là thứ tự của AR .
- q là thứ tự của MA.
- d là số chênh lệch cần thiết để làm cho time series stationary.
- Trong phần này sử dụng p = 1 , q = 1 và d = 0.



ARIMA Model Results

Dep. Variable:	D.TotalCases	No. Observations:	4
Model:	ARIMA(1, 1, 0)	Log Likelihood	-58.850
Method:	css-mle	S.D. of innovations	569204.020
Date:	Sat, 16 Apr 2022	AIC	123.701
Time:	17:37:45	BIC	121.860
Sample:	1	HQIC	119.661

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.073e+06	2.12e+05	5.057	0.000	6.57e+05	1.49e+06
ar.L1.D.TotalCases	-0.5325	0.418	-1.275	0.202	-1.351	0.286

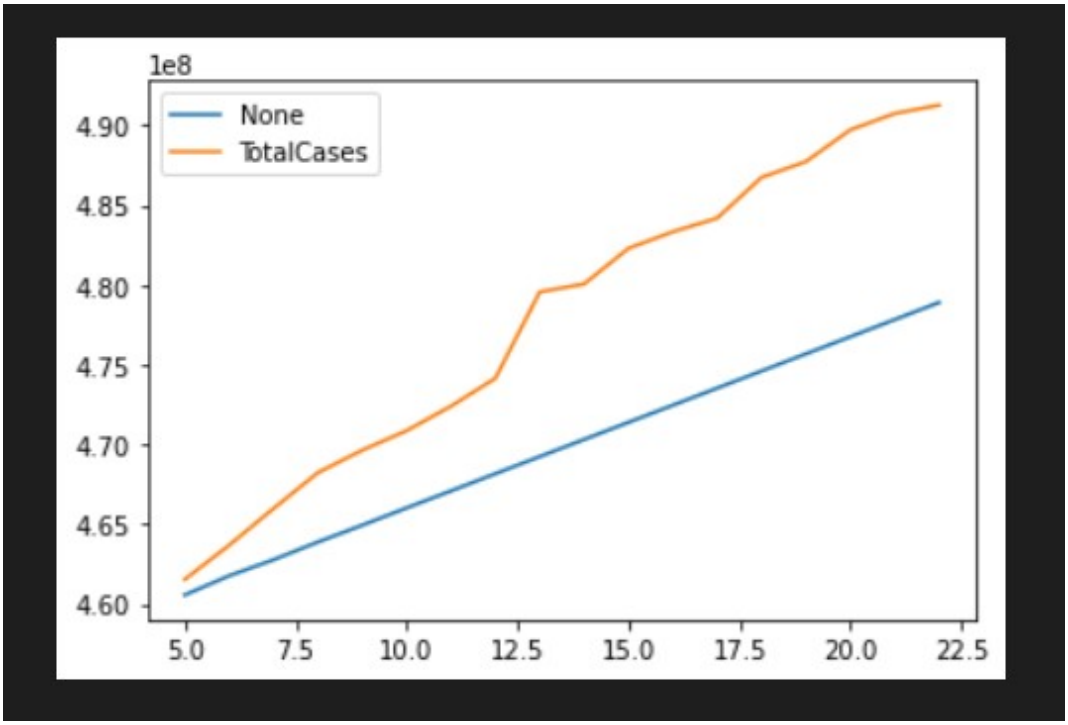
Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.8779	+0.0000j	1.8779	0.5000

Hình 25: Kết quả thuật toán ARIMA.

- Từ hình trên ta thấy p-value (p>|z|) < 0.05 có thể kết luận rằng coefficient có ý nghĩa thống kê .
- std err dùng để ước tính độ lỗi của giá trị dự đoán . Ở hình trên giá trị của nó là 2,12e+05 .

Mô hình ARIMA dự đoán tổng số ca covid trên thế giới trong tương lai



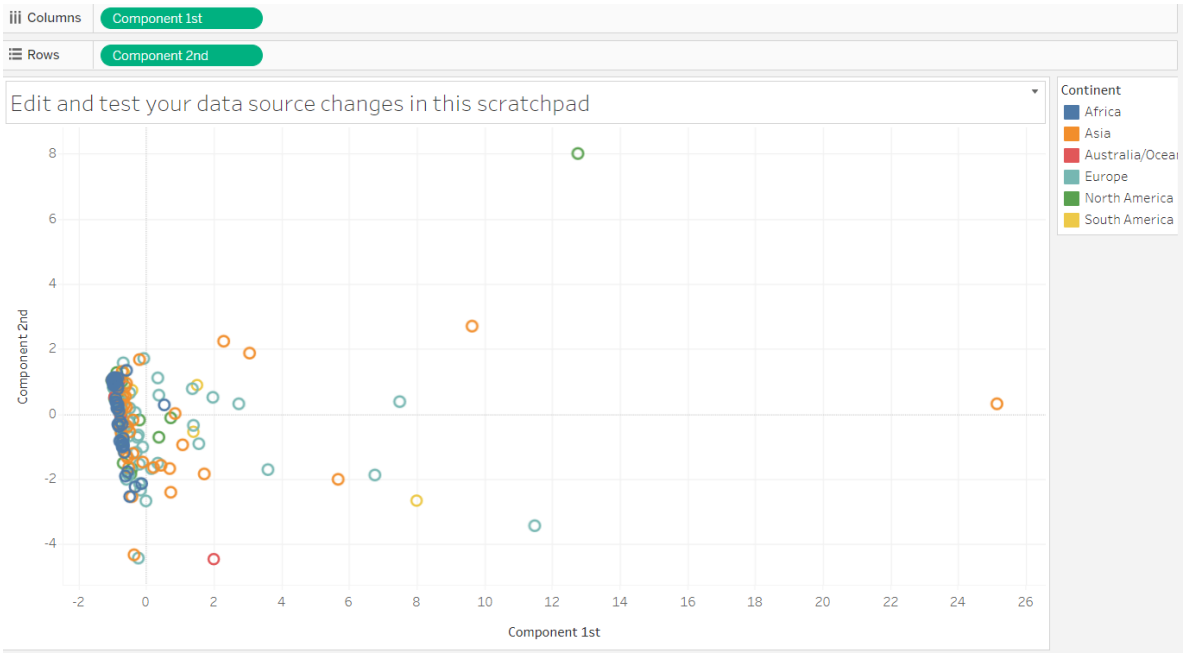
Hình 26: Mô hình ARIMA dự đoán số ca covid trên thế giới trong tương lai.

Từ đồ thị trên ta thấy kết quả mà ARIMA dự đoán cho kết quả không tốt lắm. Khoảng cách giữa kết quả thực tế và kết quả dự đoán xa dần theo thời gian

7.5 Data Reduction

Nhóm sử dụng 2 phương pháp giảm số chiều là PCA và LDA. Dữ liệu được sử dụng là dữ liệu vào ngày 1-4. Dữ liệu lấy cột Continent làm nhãn, và sử dụng các cột còn lại trừ cột Country để làm đặc trưng.

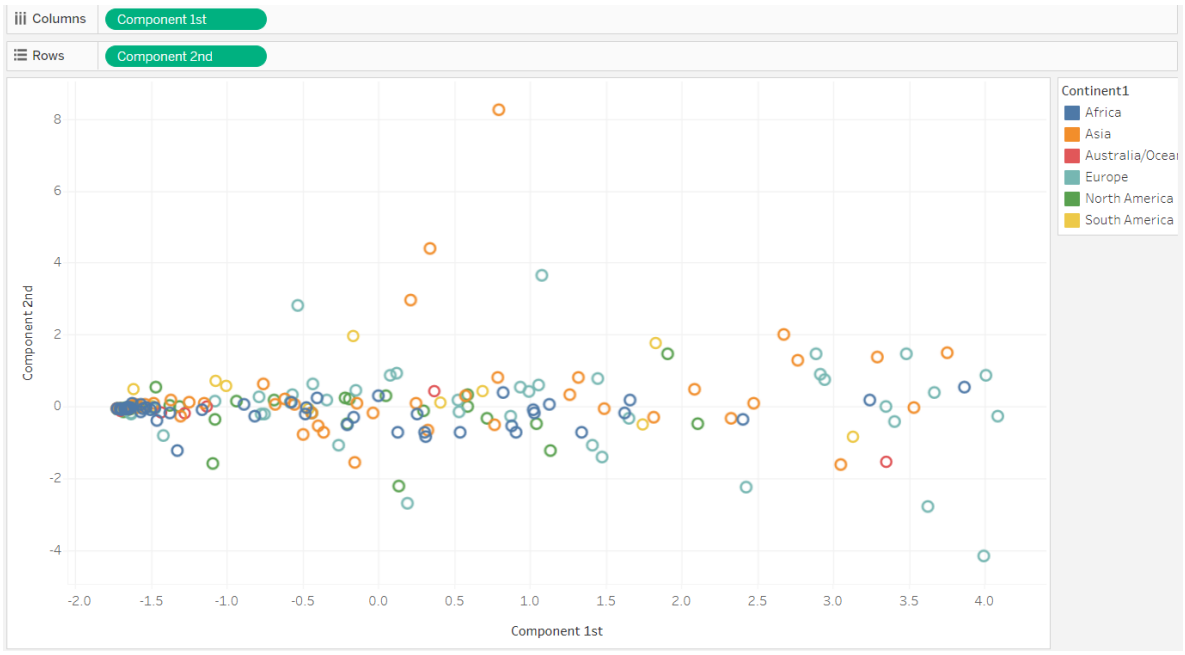
Kết quả thu được từ thuật toán PCA với 2 thành phần chính.



Hình 27: Kết quả thuật toán PCA

Từ kết quả cho thấy thuật toán PCA hoạt động không tốt, khi các điểm dữ liệu vẫn không tách nhau và các nhãn dữ liệu không tách biệt nhau.

Tiếp theo nhóm thử thuật toán LDA, kết quả thu được:



Hình 28: Kết quả thuật toán LDA

Từ kết quả cho thấy thuật toán LDA rải các điểm dữ liệu rời nhau hơn PCA, tuy nhiên các lớp không tụ lại với nhau và tách biệt.

Tài liệu

Tableau Mobile
Tableau Document
Principal Component Analysis
Linear Discriminant Analysis
Giới thiệu Tableau