**TRƯỜNG ĐẠI HỌC KINH TẾ TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH**



*GVHD:* TS. Huỳnh Văn Đức

*Lớp:* BI001

*Khóa:* K43

***MSSV:***

31171024411

|  |
| --- |
|  |
|  |

**NHÓM 4 THỰC HIỆN:**

|  |
| --- |
| Lâm Thủy Liên |
| Hoàng Thị Hồng |
| Lê Thị Diễm Vi |
| Lê Trần Thu Thủy |
| Ngô Phương Thảo Nguyên |
| Nguyễn Thị Hoàng Kim |

***Người thực hiện:***

Lê Anh Quốc

|  |
| --- |
|  |
|  |

**NHÓM 4 THỰC HIỆN:**

|  |
| --- |
| Lâm Thủy Liên |
| Hoàng Thị Hồng |
| Lê Thị Diễm Vi |
| Lê Trần Thu Thủy |
| Ngô Phương Thảo Nguyên |
| Nguyễn Thị Hoàng Kim |



**ĐỒ ÁN CUỐI KHÓA**

**MÔN: BUSINESS INTELLIGENCE**

**Đề tài:** Covid-19 Situation with Power BI and R



Mục lục

[Lời nói đầu 3](#_Toc44244025)

[I. TỔNG QUAN 4](#_Toc44244026)

[1. Đặc tả 4](#_Toc44244027)

[1.1. Nguồn dữ liệu: 4](#_Toc44244028)

[1.2. Các câu hỏi phân tích: 4](#_Toc44244029)

[1.3. Phác thảo báo cáo: 5](#_Toc44244030)

[1.4. Phác thảo mô hình DW 5](#_Toc44244031)

[2. Mục tiêu 5](#_Toc44244032)

[3. Phương pháp 6](#_Toc44244033)

[II. QUÁ TRÌNH ETL 6](#_Toc44244034)

[III. MÔ HÌNH DATA WAREHOUSE 7](#_Toc44244035)

[1. Fact table 9](#_Toc44244036)

[2. Measures Table (Những Measures tự thêm vào) 9](#_Toc44244037)

[3. Dim Table 10](#_Toc44244038)

[3.1. Location 10](#_Toc44244039)

[3.2. Calendar 10](#_Toc44244040)

[3.3. Flu US (Bảng bổ trợ) 10](#_Toc44244041)

[IV. DASHBOARD COVID-19 SITUATION 10](#_Toc44244042)

[V. CÁC VISUALIZATION TRẢ LỜI CHO NHỮNG CÂU HỎI ĐÃ ĐẶT RA 12](#_Toc44244043)

[1. Nên đưa ra những biện pháp kịp thời cho những quốc gia nào có số ca nhiễm ở mức báo động? 13](#_Toc44244044)

[2. So sánh tỷ lệ tử vong của dịch Covid – 19 và dịch cúm mùa ở Mỹ về độ nguy hiểm. 14](#_Toc44244045)

[3. Phân cụm những bang của Mỹ có diễn biến phức tạp với Covid 19 dựa trên số ca nhiễm và tỷ lệ tử vong của ngày gần nhất. 16](#_Toc44244046)

[VI. TÀI LIỆU THAM KHẢO 17](#_Toc44244047)

# Lời nói đầu

*Tình hình dịch bệnh Covid – 19 diễn biến phức tạp trên toàn cầu trong đó có Việt Nam chúng ta. Nhằm chung tay đẩy lùi và vượt qua dịch bệnh, thôi thúc em làm một project liên quan đến vấn đề thực tiễn hiện tại (tình hình Covid – 19) với đầy sự quyết tâm.*

*Covid – 19 Situation with Power BI là bảng báo cáo về tình hình sơ lược của đại dịch toàn cầu Covid – 19 với sự hỗ trợ của công cụ Microsoft Power BI.*

*Dữ liệu của bài project lấy từ nguồn dữ liệu chuẩn và đáng tin cậy vì vậy độ chính xác của dữ liệu cao (Đại học John Hoskin).*

*Qua bài nghiên cứu này cũng giúp em học được nhiều điều mới mẻ, thú vị và ý nghĩa.*

*Xin chân thành cảm ơn*

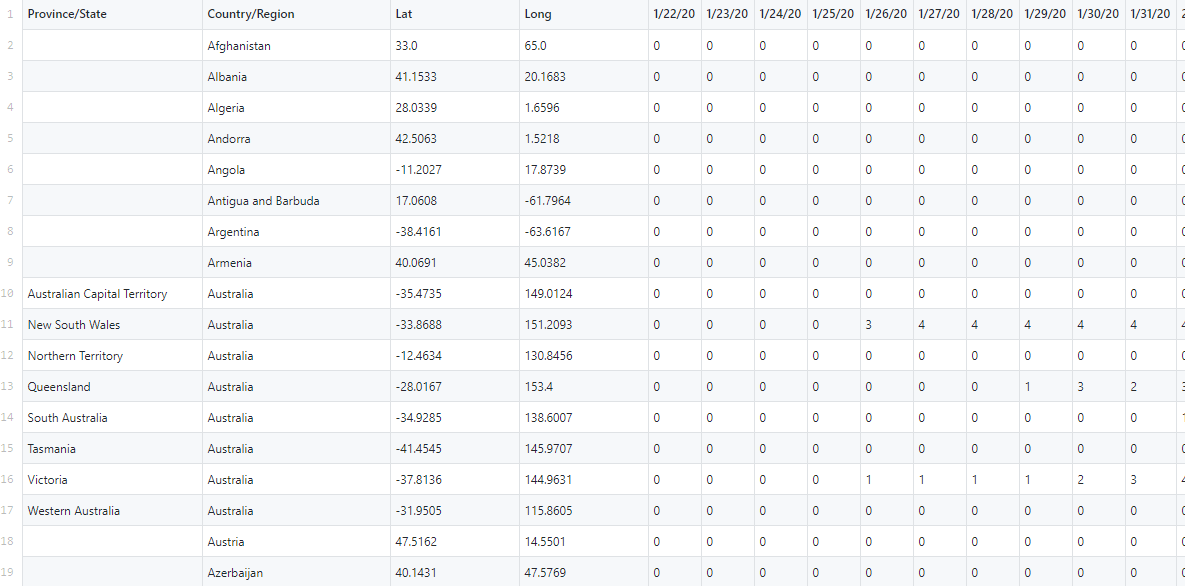
# TỔNG QUAN

## Đặc tả

### Nguồn dữ liệu:

* Dữ liệu về Covid-19 của Đại Học Johns Hopkins <https://github.com/search?q=covid>

<https://www.kaggle.com/search?q=covid+19>



*Ví dụ về dữ liệu*

* Dữ liệu timeseries về các quốc gia, vùng lãnh thổ, kinh độ, vĩ độ, số ca nhiễm, ca chết, ca hồi phục theo từng ngày.

### Các câu hỏi phân tích:

* Nên đưa ra những biện pháp kịp thời cho những quốc gia nào có số ca nhiễm ở mức báo động?
* So sánh tỷ lệ tử vong của dịch Covid – 19 và dịch cúm mùa ở Mỹ về độ nguy hiểm.
* Phân cụm những bang ở Mỹ có mức độ ảnh hưởng Covid dựa trên tỷ lệ tử vong và số ca nhiễm.

### Phác thảo báo cáo:

* Báo cáo thể hiện được các thông tin thiết yếu về dịch bệnh Covid-19 ở thời điểm hiện tại.
* Báo cáo được thể hiện dưới dạng dashboard theo dõi.

### Phác thảo mô hình DW

* Dữ liệu được chuẩn hóa dưới mô hình Data Warehouse - Star schema gồm 1 bảng fact và các bảng dim xung quanh:
* Bảng fact:
* Master: Country/Region, Date, Location, Status, Value
* Bảng dim:
* Calendar: Date, Day, Day Name, Month, Month Name, Quarter, Week of Month, Year.
* Location: Location, Country/Region, Province/State, Lat, Long
* Measures Table: Cases, Deaths, Recovered, Active, Increase\_Case, Increase\_Death, …

*(Mô hình DW cụ thể sẽ được nói rõ trong phần II)*

## Mục tiêu

* Sử dụng Power BI để tạo dashboard đơn giản thể hiện được thông tin cần thiết về dịch bệnh Covid-19 bao gồm:
* Tổng số ca nhiễm, ca tử vong, ca hồi phục, ca hiện tại trên toàn cầu.
* Thể hiện xu hướng tăng ca nhiễm theo từng ngày toàn cầu.
* Số ca nhiễm, ca tử vong, ca phục hồi ở 1 số quốc gia nổi bật.
* Số lượng quốc gia, vùng lãnh thổ có ca nhiễm là bao nhiêu?
* Bản đồ nhiễm trên toàn cầu.
* Kết hợp các loại biểu đồ
* Card: Để thể hiện tổng số ca nhiễm, ca tử vong, ca phục hồi, ca hiện tại và số quốc gia ảnh hưởng
* Line Chart: Thể hiện xu hướng của dịch bệnh qua timeseries.
* Clustered Column Chart: Thể hiện tỷ lệ tử vong và tỷ lệ hồi phục của từng quốc gia.
* Multi Row Card: Thể hiện số ca nhiễm, ca tử vong của từng quốc gia.
* Map: Thể hiện số ca nhiểm toàn cầu.
* Slicer: Lọc quốc gia, lọc theo thời gian.
* Kết hợp các hàm DAX: Hàm CALCULATE, VAR, IF … ELSE, COUNT, MAX, SUM và một số hàm về thời gian.
* Tương tác người dùng: Người dùng có thể tương tác bằng cách chọn Quốc gia và Khoảng thời gian mà họ muốn xem các thông tin trên.

## Phương pháp

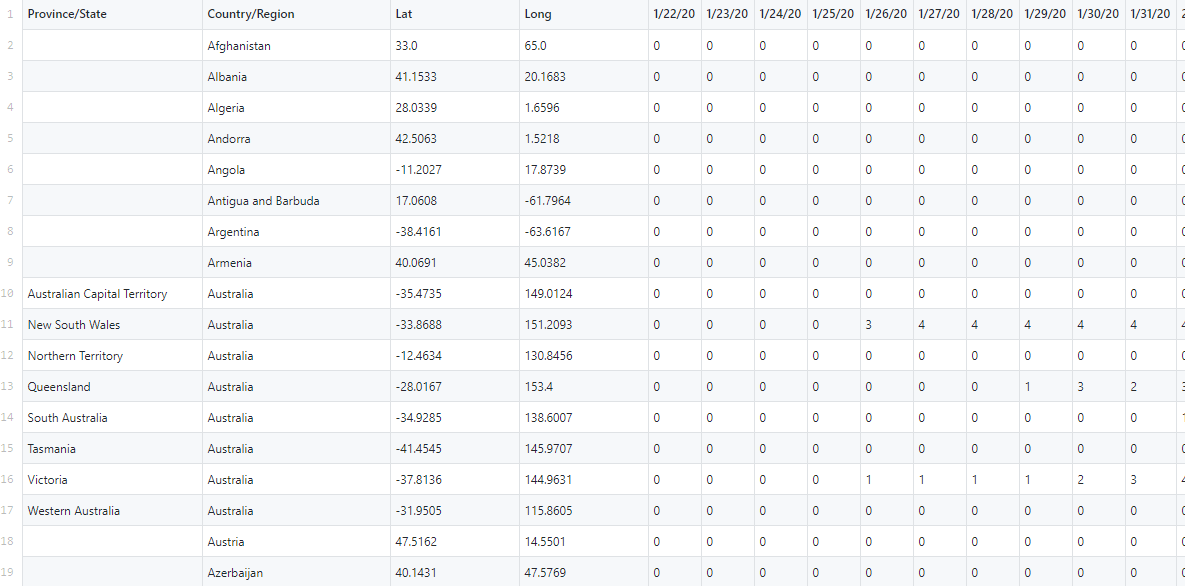
* Mô hình Data Warehouse: Mô hình hình sao với 1 bảng fact (chứa các thông tin chính) và 2 bảng dim (location và calendar) cùng 2 bảng bộ trợ (Measure table và Flu US)
* Mô hình Data Mining: K-means

Ý tưởng chính của thuật toán k-means là tìm cách phân nhóm các đối tượng (objects) đã cho vào k cụm (k là số các cụm được xác định trước, k là số nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.

Thuật toán k-means áp dụng cho các đối tượng được biểu diễn bởi các điểm trong không gian vectơ d chiều U = {xi | i = 1, … , N}, với xi ∈ ℜ biểu thị đối tượng (hay điểm dữ liệu) thứ i.

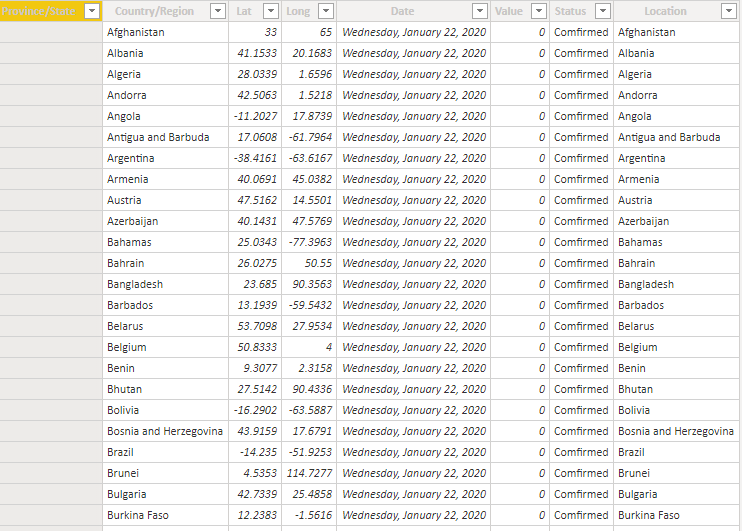
# QUÁ TRÌNH ETL

Từ dữ liệu raw được lấy trực tiếp web (Get data -> Web) có 3 file dưới dạng như hình bên dưới:



Dữ liệu raw ban đầu

Sau khi get data, bắt đầu transform dữ liệu raw (chuyển những cột ngày tháng với số liệu thành 1 cột time và thêm cột status ứng với tình trạng của mỗi file (Comfirmed, Deaths, Recovered). Sau đó ta được như hình bên dưới:



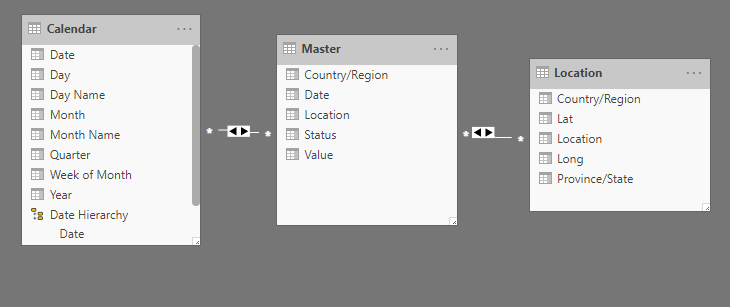
Tiếp đến là ghép 3 file như trên lại thì được 1 table tổng quát về Covid19 mang tên Master (Có thể xem thêm ở phần bên dưới).

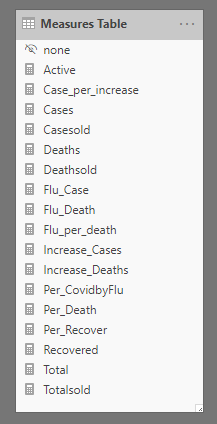
Bước tiếp theo là tạo khóa và tạo các bảng có liên quan với mô hình bao gồm 2 bảng (Time và Location)

Bước cuối cùng là load dữ liệu vào.

# MÔ HÌNH DATA WAREHOUSE

Mô hình Data Warehouse: Star Schema





*Như đã trình bày ở phần Tổng quan, cấu trúc của Data Warehouse gồm 1 bảng fact và 2 bảng dimension. Thêm vào đó, việc tạo ra 1 bảng Measures Table riêng biệt (các measures mới không để trong bảng fact) nhằm để kiểm soát, quản lý số lượng measures được thêm vào chặt chẽ, rõ ràng và tốt hơn.*

## Fact table

|  |  |
| --- | --- |
| Tên thuộc tính | Chú thích |
| Country/Region | Tên quốc gia |
| Date | Mốc thời gian theo ngày tháng năm |
| Location | Bang, tỉnh của từng quốc gia |
| Status | Tình trạng (Comfirmed, Deaths, Recovered) |
| Value | Số ca theo từng mốc thời gian |

## Measures Table (Những Measures tự thêm vào)

|  |  |  |
| --- | --- | --- |
| Tên measure | Câu lệnh DAX | Chú thích |
| Total | Total =  var newdate = MAX(Master[Date])  return CALCULATE(SUM(Master[Value]), Master[Date] = newdate) | Tổng số ca (nhiễm, tử vong, hồi phục) ở ngày gần nhất |
| Totalold | Totalsold =  var olddate = MAX(Master[Date]) - 1  return CALCULATE(SUM(Master[Value]), Master[Date] = olddate) | Tổng số ca (nhiễm, tử vong, hồi phục) ở ngày trước đó |
| Cases | Cases = CALCULATE([Total],  Master[Status]="Comfirmed") | Tổng số ca nhiễm ở ngày gần nhất |
| Deaths | Deaths = CALCULATE([Total],  Master[Status]="Deaths") | Tổng số ca tử vong ở ngày gần nhất |
| Recovered | Recovered = CALCULATE([Total],  Master[Status]="Recovered") | Tổng số ca phục hồi ở ngày gần nhất |
| Active | Active = [Cases] - [Deaths] - [Recovered] | Tổng số ca còn đang nhiễm ở ngày gần nhất |
| Per\_Death | Per\_Death = [Deaths]/[Cases] | Tỷ lệ tử vong |
| Per\_Recovered | Per\_Recover = [Recovered]/[Cases] | Tỷ lệ phục hồi |
| Casesold | Casesold = CALCULATE([Totalsold],  Master[Status]="Comfirmed") | Tổng số ca nhiễm của ngày trước đó |
| Deathsold | Deathsold = CALCULATE([Totalsold],  Master[Status]="Deaths") | Tổng số ca tử vong của ngày trước đó |
| Increase\_Cases | Increase\_Cases = [Cases] - [Casesold] | Số ca nhiễm tăng so với ngày trước đó |
| Increase\_Deaths | Increase\_Deaths = [Deaths] - [Deathsold] | Số ca tử vong tăng so với ngày trước đó |
| Case\_per\_increase | Case\_per\_increase = [Increase\_Cases]/[Casesold] | Tỷ lệ tăng ca nhiễm so với ngày trước đó |
| Flu\_Case | Flu\_Case = SUM('Flu US'[Case]) | Tổng số ca nhiễm do cúm ở Mỹ |
| Flu\_Death | Flu\_Death = SUM('Flu US'[Death]) | Tổng số ca tử vong do cúm ở Mỹ |
| Flu\_per\_death | Flu\_per\_death = [Flu\_Death]/[Flu\_Case] | Tỷ lệ tử vong do cúm ở Mỹ |
| Per\_CovidFlu | Per\_CovidbyFlu = CALCULATE([Per\_Death],  Master[Country/Region]="US")  /[Flu\_per\_death] | Số lần khi so sánh tỷ lệ tử vong giữa dịch Covid19 và dịch cúm ở Mỹ. |

## Dim Table

### Location

|  |  |
| --- | --- |
| Tên thuộc tính | Chú thích |
| Location | Bang, tỉnh của từng quốc gia |
| Country/Region | Tên quốc gia |
| Province/State | Tỉnh, bang |
| Lat | Vĩ độ |
| Long | Kinh độ |

### Calendar

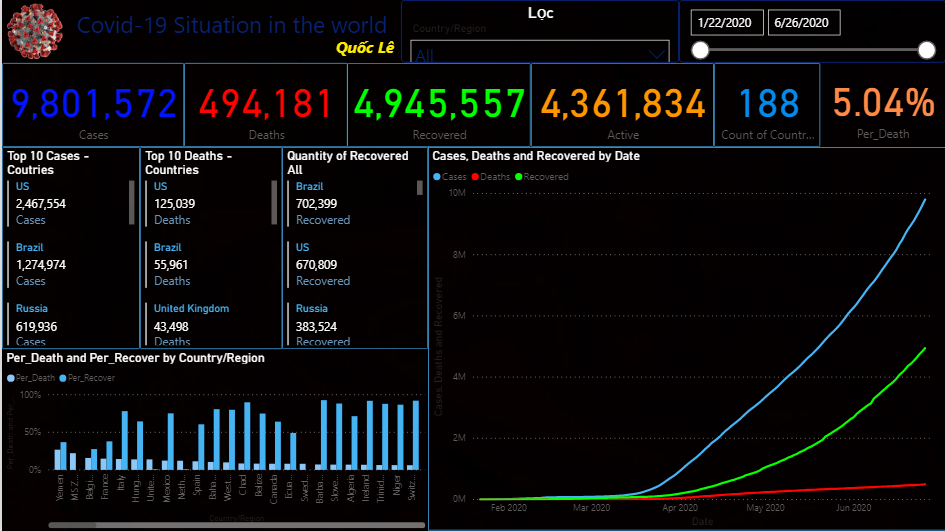
|  |  |
| --- | --- |
| Tên thuộc tính | Chú thích |
| Date | Mốc thời gian theo ngày tháng năm |
| Day | Ngày |
| Day Name | Thứ |
| Month | Tháng |
| Month Name | Tên tháng |
| Quarter | Quý |
| Week of Month | Tuần trong tháng |
| Year | Năm |

### Flu US (Bảng bổ trợ)

|  |  |
| --- | --- |
| Tên thuộc tính | Chú thích |
| Nhóm tuổi | Nhóm tuổi của người bị cúm |
| Case | Số ca nhiễm cúm theo độ tuổi ở Mỹ |
| Death | Số ca tử vong do cúm theo độ tuổi ở Mỹ |
| Recover | Số ca phục hồi do cúm theo độ tuổi ở Mỹ |

# DASHBOARD COVID-19 SITUATION

Dữ liệu của dashboard được cập nhật theo thời gian và có độ trễ là 23h

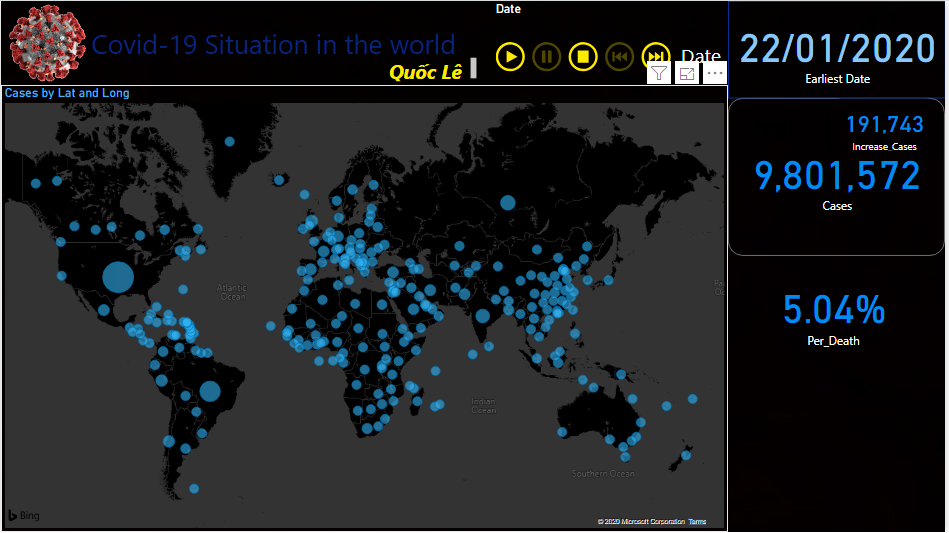


*Dashboard 1*

Bao gồm các biểu đồ:

* Card: Thể hiện tổng số ca nhiễm, ca tử vong, ca phục hồi, ca hiện tại, số quốc gia, và tỷ lệ tử vong trên toàn cầu.
* Multi-Row Card: Thể hiện top 10 quốc gia có số ca nhiễm, ca tử vong và phục hồi cao nhất.
* Cluster-Column Chart: Thể hiện tỷ về tử vong và tỷ lệ phục hồi của từng quốc gia.
* Line chart: Thể hiện tổng quan diễn biến theo timeseries.

Người dùng có thể tương tác với dashboard bằng cách Chọn vào 2 slicer ở trên (lọc theo quốc gia và theo chuỗi thời gian). Mặc khác, người dùng có thể tương tác bằng cách chọn 1 điểm trên Line chart tương ứng với 1 mốc thời gian nào đó, kết quả trả về các biểu đồ còn lại sẽ hiển thị tương ứng dữ liệu tại thời điểm đó.



*Dashboard 2*

Bao gồm các biểu đồ:

* Card: Thể hiện tổng số ca nhiễm, số ca nhiễm tăng mỗi ngày, thời gian và tỷ lệ tử vong trên toàn cầu.
* Map: Thể hiện số ca nhiễm của từng bang, từng quốc gia theo từng ngày và quy mô.

Người dùng có thể tương tác với dashboard bằng cách bấm Play để bắt đầu xem quá trình diễn ra của dịch bệnh Covid-19 trong suốt thời gian vừa qua tại thời điểm bắt đầu được lấy là ngày 22/01/2020 cho đến thời điểm hiện tại.

*Note: Có thể xem kỹ và trải nghiệm ở file Microsoft Power BI được đính kèm.*

# CÁC VISUALIZATION TRẢ LỜI CHO NHỮNG CÂU HỎI ĐÃ ĐẶT RA

Trả lời các câu hỏi phân tích đã đề ra từ trước***.***

***Lưu ý: Dữ liệu trả lời cho những câu hỏi bên dưới được lấy cứng ngày 28/04/2020. Riêng dữ liệu của Dashboard là dữ liệu động update mỗi ngày.***

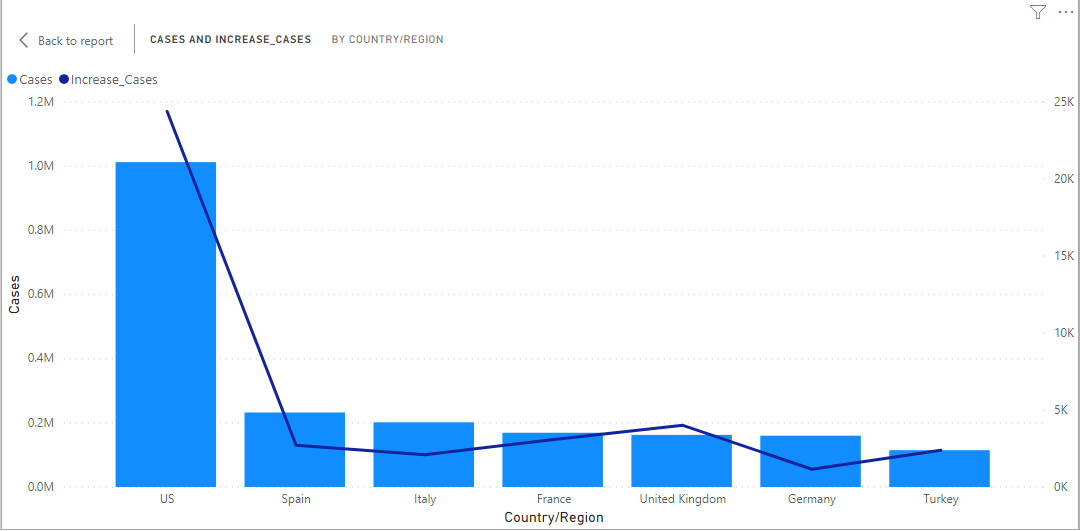
## Nên đưa ra những biện pháp kịp thời cho những quốc gia nào có số ca nhiễm ở mức báo động?

Để trả lời câu hỏi này, cần phải đưa ra mức báo động của dịch bệnh Covid-19 trong thời điểm này. Và dưới đây là mẫu mức độ báo động do đánh giá từ phía Việt Nam.

|  |  |
| --- | --- |
| Criteria (Cases) | Priority |
| > 100000 | 1 |
| 10000 - 99999 | 2 |
| 1000 - 9999 | 3 |
| 100 - 999 | 4 |
| 1 - 99 | 5 |
| 0 | 6 |

Nhìn vào bảng trên có thể thấy được mức ca nhiễm và độ ưu tiên của các quốc gia bị nhiễm bệnh. Trong bài nghiên cứu này, về mặc chủ quan sẽ lấy mức độ ưu tiên đầu tiên tức số ca nhiễm vượt quá 100000 ca và mức độ tăng mỗi ngày trên 1000 ca mỗi ngày là những quốc gia đáng báo động cho toàn cầu.

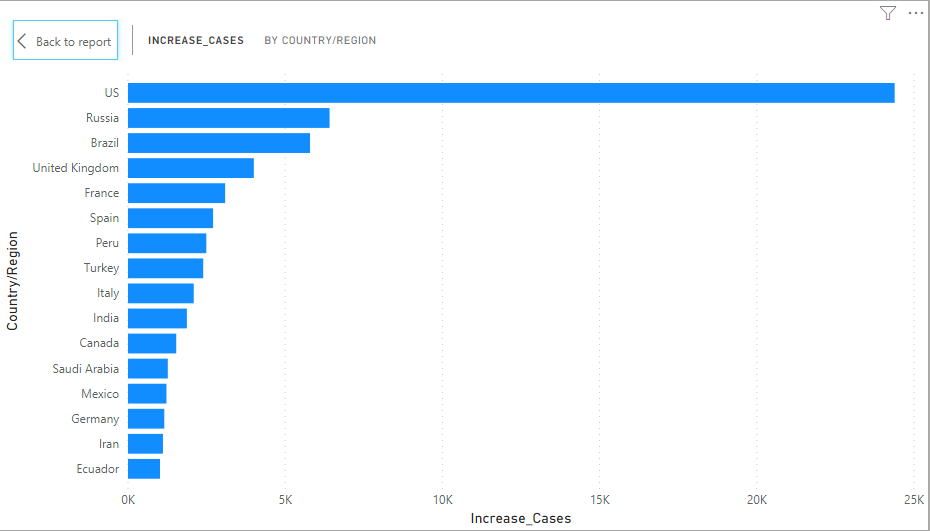
Với phân tích trên, nhìn vào biểu đồ 1.1 Cases and Increase\_cases by Country/Region, có thể thấy được có 7 quốc gia đang có mức độ tăng trên 1000 ca mỗi ngày và tổng số ca nhiễm hiện tại vượt quá 100000 ca.



Biểu đồ sử dụng: 2 measures (Cases, Increase\_Cases) và 1 dimension (Country/Region) với bộ lọc Cases >= 100000 và Increase\_Cases >= 1000.

*Với dạng biểu đồ này, người dùng có thể click vào từng phần (từng quốc gia) để thấy rõ số ca tăng mỗi ngày và tổng ca nhiễm của quốc gia đó.*

Ngoài ra, những quốc gia dù tổng số ca nhiễm chưa cao nhưng về mức độ tăng mỗi ngày ở mức đáng lưu ý cũng nên được quan tâm và đưa ra biện pháp cần thiết khẩn cấp. Và biểu đồ 1.2 dưới đây thể hiện được những quốc gia có số ca tăng trên 1000 ca mỗi ngày.

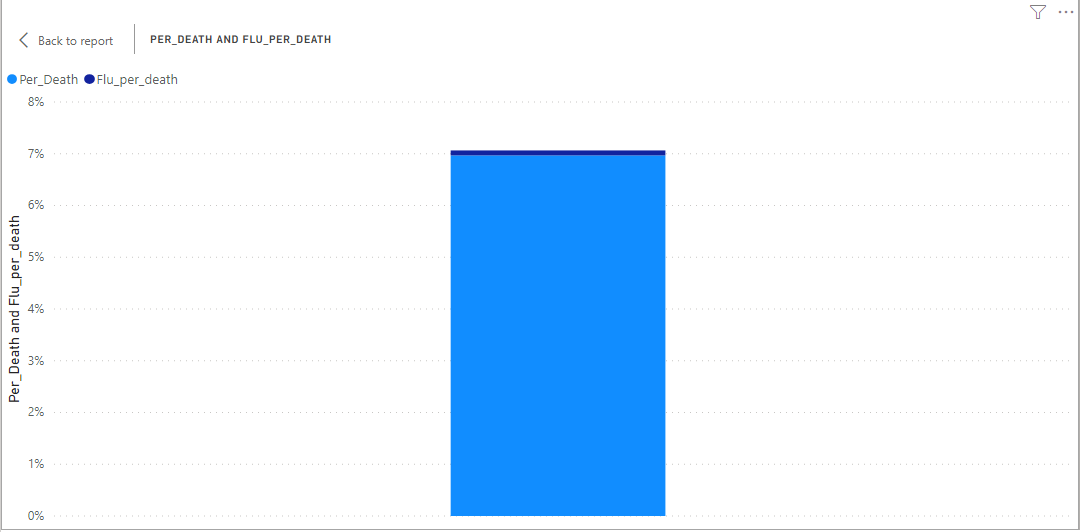


Biểu đồ sử dụng: 1 measures (Increase\_Cases) và 1 dimension (Country/Region) với bộ lọc Increase\_Cases >= 1000.

*Với dạng biểu đồ này, người dùng có thể click vào từng phần (từng quốc gia) để thấy rõ số ca tăng mỗi ngày của quốc gia đó.*

## So sánh tỷ lệ tử vong của dịch Covid – 19 và dịch cúm mùa ở Mỹ về độ nguy hiểm.

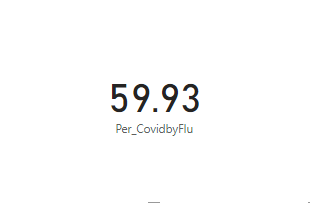
Theo ước tính của CDC thì dịch cúm mùa ở Mỹ khiến khoảng 35,520,882 người mắc và khoảng 34,158 người tử vong do cúm chiếm tỷ lệ gần 0.1% (*nguồn bên dưới*). Nhưng đối với dịch bệnh Covid-19 lại mang đến hậu quả nghiêm trọng hơn cả thể. Biểu đồ 2.1 Per\_Death and Flu\_per\_death dưới đây sẽ so sánh tỷ lệ tử vong của 2 dịch bệnh.



Biểu đồ sử dụng 2 measures (Per\_Death và Flu\_per\_death)

Theo biểu đồ có thể thấy được tỷ lệ tử vong do cúm mùa ở Mỹ chỉ chiếm vỏn vẹn 0.1% trong khi đó đối với dịch bệnh Covid-19 là 6.97%, gấp khoảng 60 lần dịch cúm, cho thấy được mức độ nguy hiểm của dịch bệnh Covid-19 ở Mỹ tại thời điểm hiện tại.

*Với dạng biểu đồ này, người dùng có thể click vào từng phần (theo màu) để thấy rõ số phần trăm.*

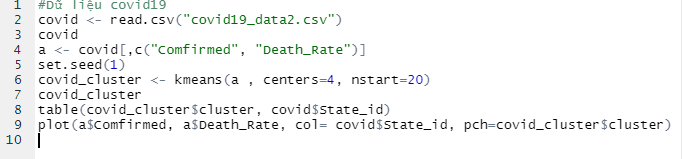


Biểu đồ sử dụng 1 dimension (Per\_CovidbyFlu)

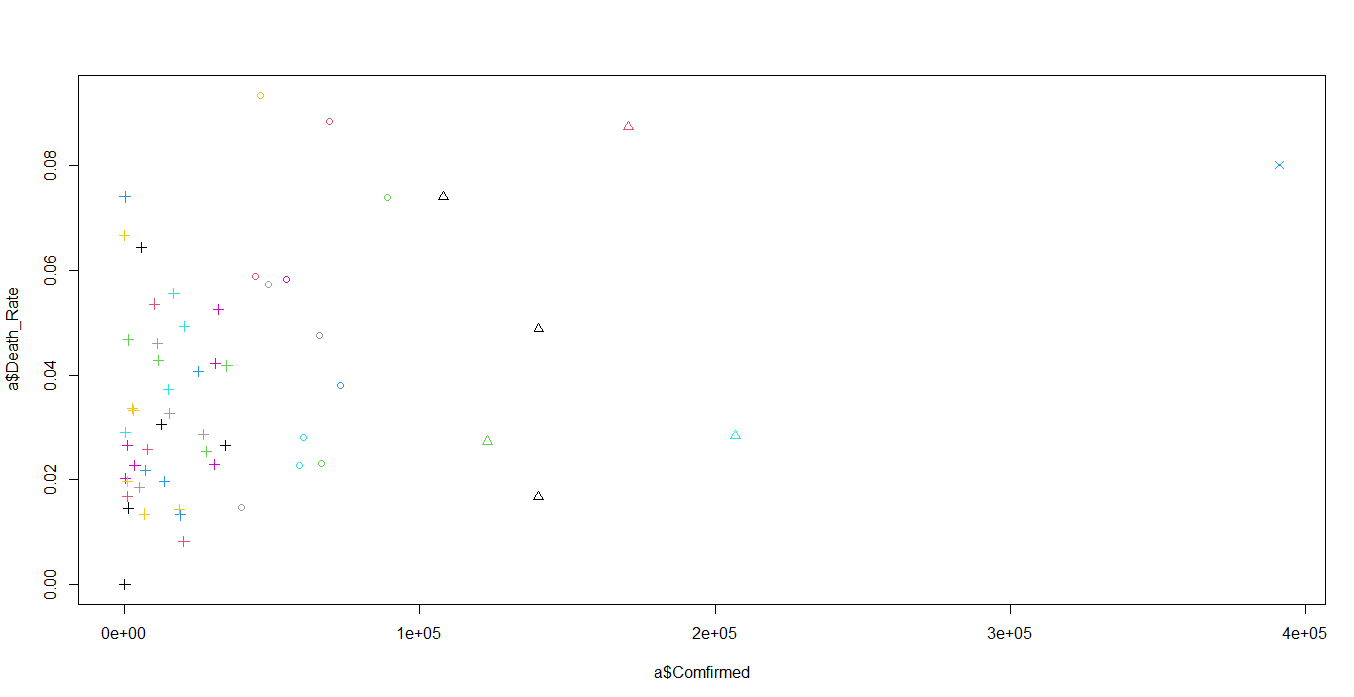
Thêm vào đó, WHO và CDC ước tính có khoảng 291,000 đến 646,000 người tử vong trên toàn cầu do cúm và con số chính xác tại thời điểm hiện tại là 168,902 người tử vong do cúm, trong khi đó số ca tử vong do Covid-19 trên toàn cầu đã vượt ngưỡng 220,000 ca tử vong (vào ngày 28/04/2020). Cho thấy được mức độ nguy hiểm và tình trạng khẩn cấp dịch bệnh trên toàn cầu.

## Phân cụm những bang của Mỹ có diễn biến phức tạp với Covid 19 dựa trên số ca nhiễm và tỷ lệ tử vong của ngày gần nhất.

***Lưu ý: Riêng dữ liệu trả lời cho câu hỏi này được được update ngày 26/06/2020.***



*Code R*



*Biều đồ khi chạy code*

Sử dụng 1 dimension: State\_Id, 2 measure: Số ca nhiễm của bang và tỷ lệ tử vong của bang.

Việc sử dụng mô hình phân cụm clustering thuật toán k-means với k = 4 (trong trường hợp của câu hỏi này) mang lại kết quả tương đối. Nhằm chia ra các cụm bang có mức độ nhiễm dịch bệnh và tỷ lệ tử vong của từng bang đó nhằm đánh giá mức độ ảnh hưởng và nguy hiểm của dịch bệnh đồng thời giúp đưa ra những biện pháp kịp thời để ngăn chặn, đẩy lùi dịch bệnh và giảm thiểu thương vong ở mức thấp nhất (chẳng hạn như đóng cửa bang, phong tỏa, …)

# TÀI LIỆU THAM KHẢO

Technical Analysis With R – Ko Chiu Yu

<https://www.worldometers.info/>

<https://www.who.int/mediacentre/news/statements/2017/flu/en/>

<https://www.cdc.gov/media/releases/2017/p1213-flu-death-estimate.html>

<https://www.cdc.gov/flu/about/burden/past-seasons.html>

<https://www.kaggle.com/search?q=clustering+covid19+with+r+kernelLanguage%3AR>

<https://systems.jhu.edu/>

<https://www.thelancet.com/pdfs/journals/laninf/PIIS1473-3099(20)30120-1.pdf>