**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

A picture containing logo

Description automatically generated

**BÁO CÁO THỰC HÀNH 1 – TRỰC QUAN HÓA DỮ LIỆU**

**MỐI QUAN HỆ DỮ LIỆU**

|  |  |  |
| --- | --- | --- |
| **THÀNH VIÊN NHÓM:** | 20127323 - | Võ Nhật Tân |
|  | 20127447 - | Ngô Đức Bảo |
|  | 20127275 - | Lê Nguyễn Nhật Phú |
|  | 20127681 - | Nguyễn Thiên Phúc |
|  |  | | |
| **GVHD:** | Bùi Tiến Lên | | |
|  | Lê Ngọc Thành | | |

**MỤC LỤC**

[I. THÔNG TIN NHÓM 3](#_Toc130052168)

[II. MỨC ĐỘ HOÀN THÀNH TỔNG THỂ CÁC YÊU CẦU 3](#_Toc130052169)

[III. MỨC ĐỘ HOÀN THÀNH CỦA TỪNG THÀNH VIÊN 3](#_Toc130052170)

[IV. MÔ TẢ QUÁ TRÌNH PHÂN TÍCH: 3](#_Toc130052171)

[1. Thu nhập dữ liệu 3](#_Toc130052172)

[2. Tiền xử lý dữ liệu: 3](#_Toc130052173)

[3. Thông tin mỗi thuộc tính: 3](#_Toc130052174)

[4. Phân tích dữ liệu: 4](#_Toc130052175)

[4.1. Tình hình thế giới hiện nay 4](#_Toc130052176)

[4.2. Tình hình giữa các châu lục 4](#_Toc130052177)

[4.3. Tình hình dịch bệnh ở các quốc gia có tổng ca nhiễm cao nhất 4](#_Toc130052178)

[4.4. Tình hình chống dịch ở những quốc gia đang có dịch bệnh căng thẳng như thế nào? 5](#_Toc130052179)

[4.5. Tình hình dịch bệnh của 5 quốc gia có tỉ lệ tử vong cao nhất 7](#_Toc130052180)

# THÔNG TIN NHÓM

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ và tên** | **MSSV** |
| 20127323 | Võ Nhật Tân | 20127323 |
| 20127447 | Ngô Đức Bảo | 20127447 |
| 20127275 | Lê Nguyễn Nhật Phú | 20127275 |
| 20127681 | Nguyễn Thiên Phúc | 20127681 |

# MỨC ĐỘ HOÀN THÀNH TỔNG THỂ CÁC YÊU CẦU

|  |  |  |
| --- | --- | --- |
| **STT** | **Yêu cầu** | **Mức độ hoàn thành** |
| 1 | Thu thập số liệu thống kê từng ngày từ trang Worldmeter | 100% |
| 2 | Sử dụng nhận xét, code/thuật toán để thể hiện trực quan các mối quan hệ giữa các trường dữ liệu | 100% |

# MỨC ĐỘ HOÀN THÀNH CỦA TỪNG THÀNH VIÊN

|  |  |  |  |
| --- | --- | --- | --- |
| **MSSV** | **Họ và tên** | **Công việc** | **Mức độ hoàn thành** |
| 20127323 | Võ Nhật Tân | - Thu nhập, tiền xử lý dữ liệu  - Phân tích xu hướng của thế giới và châu lục | 100% |
| 20127275 | Lê Nguyễn Nhật Phú | - Tình hình dịch bệnh ở các quốc gia có tổng ca nhiễm cao nhất | 100% |
| 20127447 | Ngô Đức Bảo | - Tình hình chống dịch ở những quốc gia đang có dịch bệnh căng thẳng | 100% |
| 20127681 | Nguyễn Thiên Phúc | Tình hình dịch bệnh của 5 quốc gia có tỉ lệ tử vong cao nhất | 100% |

# MÔ TẢ QUÁ TRÌNH PHÂN TÍCH:

## Thu nhập dữ liệu

* Nhóm em sử dụng thư viện “selenium” để lấy dữ liệu từ web [https://www.worldometers.info/coronavirus/#main\_table](https://www.worldometers.info/coronavirus/%23main_table)
* Sau đó sử dụng “BeautifulSoup” để decode đoạn mã html và lấy thông tin cần thiết
* Trong qua trình decode html, chúng em tiền xử lý những giá trị “NA” thành “-” để tránh gây nhầm lẫn trong quá trình phân tích.
* Các file sau khi trích xuất và xử lý xong sẽ được lưu vào thư mục “raw\_data”, mỗi file sẽ tương ứng với mỗi ngày lấy thông tin.
* Nhóm em lấy dữ liệu từ ngày 26/03/2023 – 09/03/2023

## Tiền xử lý dữ liệu:

* Đối với các dataset trong thư mục “raw\_data” nhóm tiến hành kiểm tra từng file riêng biệt sau đó kiểm tra dữ liệu giữa các dataset với nhau
* Các bước để xử lý dữ liệu bao gồm:

+ Chuẩn hóa tên cột

+ Xóa cột Serious,Critical (vì dữ liệu chỉ thể hiện số bệnh nhân điều trị ở Intensive Care Unit (ICU), nếu có báo cáo. Vì vậy, nhóm quyết định không quan tâm đến cột này)

+ Tiền xử lý các giá trị thiếu, giá trị lỗi

+ Kiểm tra kiểu dữ liệu mỗi cột

+ Kiểm tra tính chính xác của một số cột

+ Lưu thành file sau khi tiền xử lý

* Sau khi tiến hành kiểm tra, các file đều có cùng một cấu trúc, các quốc gia của dataset đều như nhau trên mỗi dataset
* Các file sau khi tiền xử lý xong sẽ được lưu trong thư mục “data”

## Thông tin mỗi thuộc tính:

|  |  |
| --- | --- |
| **Tên thuộc tính** | **Ý nghĩa** |
| Country,Other | Quốc gia, khác |
| TotalCases | Tổng cộng dồn các ca nhiễm đã được phát hiện hoặc ghi nhận là dương tính, hoặc có thể là những ca nghi nhiễm |
| NewCases | Tổng số ca nhiễm mới trong ngày |
| TotalDeaths | Tổng cộng dồn số ca tử vong đã được ghi nhận |
| NewDeaths | Tổng số ca nhiễm tử vong mới trong ngày |
| TotalRecovered | Tổng cộng dồn số ca hồi phục đã được ghi nhận |
| NewRecovered | Tổng số ca hồi phục mới trong ngày |
| ActiveCases | Số lượng người hiện đã được phát hiện hay xác nhận là nhiễm virus |
| Tot Cases/1M pop | Tổng cộng dồn số ca nhiễm đã được ghi nhận trên một triệu dân số |
| Deaths/1M pop | Tổng cộng dồn số ca nhiễm chết người đã được ghi nhận trên một triệu dân số |
| TotalTests | Tổng cộng dồn số lượng xét nghiệm kiểm tra Covid-19 |
| Tests/ 1M pop | Tổng cộng dồn số lượng xét nghiệm kiểm tra Covid-19 trên một triệu dân số |
| Population | Dân số của một quốc gia |
| Continent | Châu lục |
| DateCollected | Ngày thu thập thông tin |

## Phân tích dữ liệu:

### Tình hình thế giới hiện nay

#### Mô tả

* + Đầu tiên, nhóm khảo sát tình hình dịch bệnh trên toàn thế giới để có cái nhìn tổng quan nhất về dịch bệnh

#### Các biểu đồ và nhận xét

Chart, line chart

Description automatically generated

* + **Nhận xét:**
    - Tốc độ tăng trưởng các ca trên toàn thế giới đang có xu hướng tăng chậm lại. Nguyên nhân chính là do tổng số ca tử vong và tổng số ca hồi phục tăng tương đối nhanh
    - Điều đó ta thấy tình hình Covid trên toàn thế giới được kiểm soát nhờ vào sự gia tăng số ca hồi phục. Tuy nhiên, số lượng ca tử vong ước tính vẫn còn tăng cao.

### Tình hình giữa các châu lục

#### Mô tả

* + Nhóm em cần quan tâm đến 3 vấn đề chính:

+ Xác định tình trạng hiện tại của mỗi châu lục

+ Xác định mức độ lây lan lan và có mối quan hệ với tình trạng hiện tại như thế nào?

+ Xác định số ca mắc ở mỗi châu lục nào chiếm đa số, đồng thời thể hiện so sánh tỷ lệ hồi phục và tỷ lệ tử vong giữa các châu lục dựa theo số ca nhiễm

#### Các biểu đồ và nhận xét

##### Tình trạng hiện tại của mỗi châu lục

A picture containing chart

Description automatically generated

* + - **Nhận xét:**
      * Tổng số trường hợp nhiễm ghi nhận ở châu Âu và Châu Á chiếm đa số
      * Châu Đại Dương và châu Phi cũng ghi nhận số ca ít nhất.

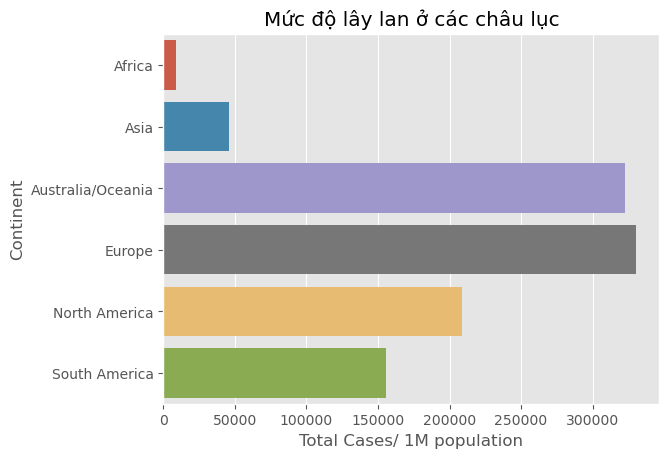
Chart, map

Description automatically generated

* + - **Nhận xét:**

- Tuy nhiên số ca hiện tại được ghi nhận ở Châu Âu và Bắc Mỹ tương đối thấp, có thể nói ở giai đoạn đầu số ca ghi nhận ở hai châu lục này khá cao tuy nhiên nhờ sự nỗ lực đã làm giảm số ca trong giai đoạn sau này.

##### Mức độ lây lan ở các châu lục



* + - **Nhận xét:**
      * Tốc độ lây lan ở Châu Âu và Châu Đại Dương rất nhanh trong khi ở châu Phi ghi nhận khá chậm.
      * Châu Đại Dương tuy ghi nhận số ca ít nhưng tỷ lệ lây lan khá cao.
      * Đặc biệt ở châu Âu, nếu tiếp tục mức độ lây lan cao có thể dẫn đến tình trạng khó kiểm soát số lượng ca nhiễm.

##### Tỷ lệ số ca nhiễm, số ca tử vong (trên số ca nhiễm) và số ca hồi phục (trên số ca nhiễm) giữa các châu lục.

Chart, pie chart

Description automatically generated

* + - **Nhận xét:**
* Các nước ở Châu Đại Dương, Châu Phi có tỉ lệ hồi phục cao: Châu đại dương chiếm 2.1% tổng số ca nhiễm trong đó tỷ lệ tử vong khoảng 2.7%, tỷ lệ ca hồi phục khoảng 17.6%, tương tự đối với Châu Phi
* Các quốc gia khác đang cho thấy tỷ lệ số ca tử vong cao, đặc biệt là Nam Mỹ khi tổng số ca chỉ chiếm 10% nhưng tỷ lệ tử vong chiếm đến 28.3% tổng số ca tử vong trong khi số ca phục hồi chỉ đạt 17.2%.

### Tình hình dịch bệnh ở các quốc gia có tổng ca nhiễm cao nhất

#### Mô tả

- Để xem xét tình hình dịch bệnh ở các quốc gia hiện nay, ta khảo sát các ca tử vong mới, các ca nhiễm mới và các ca hồi phục mới những ngày gần đây như thế nào ở các quốc gia, đồng thời so sánh tốc độ lây lan giữa các quốc gia

#### Các biểu đồ và nhận xét

##### Tổng số ca và tốc độ lây lan của 5 quốc gia có tổng số ca lớn nhất

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

\* **Nhận xét:**

**-** Pháp và Đức có tốc độ lây lan khá cao dù tổng số ca chỉ bằng nửa tổng số ca của Mỹ.

**-** Ấn Độ có tổng số ca đứng thứ 2 nhưng tốc độ nhiễm lại khá thấp.

##### Các ca tử vong mới, các ca nhiễm mới và các ca hồi phục mới những ngày gần đây như thế nào ở các quốc gia?

Chart, line chart

Description automatically generated

* **Nhận xét:**

- Các ca mắc mới và tử vong mới đang giảm và duy trì ở mức thấp, trong khi các ca phục hồi mới đang gia tăng và ổn định ở các quốc gia có nhiều ca nhiễm nhất. Tuy nhiên số ca tử vong đang có dấu hiệu tăng

### Tình hình chống dịch ở những quốc gia đang có dịch bệnh căng thẳng như thế nào?

#### Mô tả

* + Để trả lời câu hỏi này, ta cần làm rõ 3 ý:
    - Thế nào là quốc gia đang có dịch bệnh căng thẳng? Đâu là những quốc gia có tình hình dịch bệnh căng thẳng?
      * **Mục đích:** xác định quốc gia có tình hình dịch bệnh căng thẳng
    - Số ca nhiễm mới của những quốc gia đó như thế nào?
      * **Mục đích:** xác định tình hình dịch bệnh của các nước đó
    - Số ca hồi phục và số ca chết mới ở những quốc gia đó như thế nào?
      * **Mục đích:** xác định tình hình công tác chống dịch của nước đó

#### Các biểu đồ và nhận xét

##### Thế nào là quốc gia đang có dịch bệnh căng thẳng? Đâu là những quốc gia có tình hình dịch bệnh căng thẳng?

* Ở đây em định nghĩa quốc gia đang có dịch bệnh căng thẳng là quốc gia có tổng số lượng ca nhiễm trong khoảng thời gian thu thập thông tin là nhiều nhất.

Chart, bar chart

Description automatically generated

\* **Nhận xét:**

**-** Có sự chênh lệch về số ca nhiễm mới ở 6 quốc gia đầu tiên và những quốc gia còn lại. Đặc biệt, 6 quốc gia đầu tiên có tổng số lượng ca nhiễm mới rất lớn (hơn 100000 ca nhiễm mới) 🡺 Xem xét 6 quốc gia này.

##### Số ca nhiễm mới của những quốc gia đó như thế nào?

Graphical user interface, chart

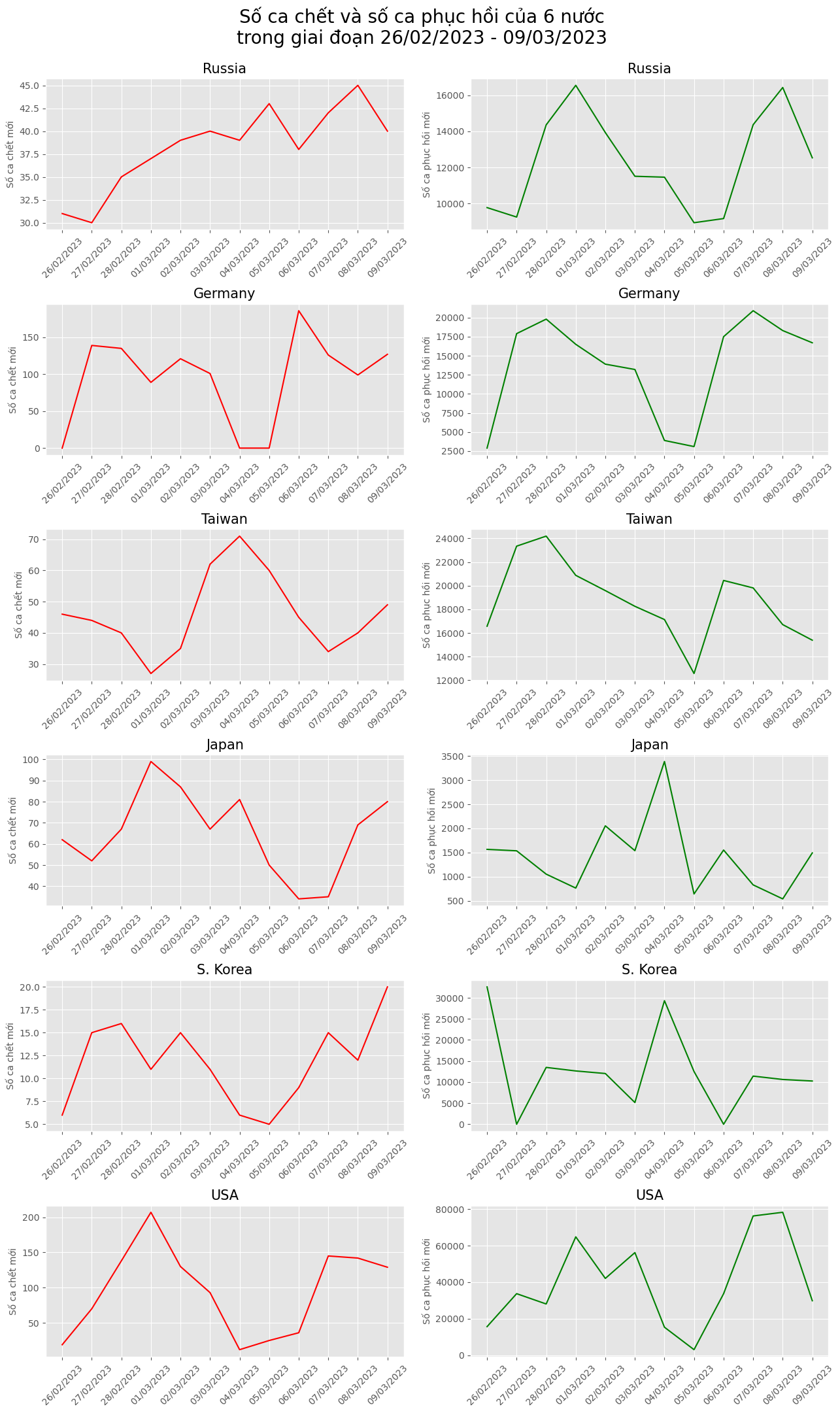
Description automatically generated

\* **Nhận xét:**

**-** Ở Đức vừa có vẻ như vừa trải qua một đợt bùng phát dịch lớn, số lượng ca nhiễm từ 0 à hơn 30000 chỉ sau 1 đêm. Tuy nhiên tình hình đã có tiến triển tốt khi số ca mắc ít đi dần theo từng ngày.

**-** Ở Hàn Quốc, tình hình dịch vẫn đang kéo dài và có dấu hiệu căng thẳng hơn trong khi 4 quốc gia còn lại đang có dấu hiệu suy giảm.

##### Số ca hồi phục và số ca chết mới ở những quốc gia đó như thế nào?



**\* Nhận xét:**

**-** Số lượng ca chết mới ở Đài Loan đang có dấu hiệu tăng dần trong khi số ca hồi phục có dấu hiệu giảm hoặc đi ngang theo thời gian.

**-** Ở các nước còn lại thì số ca chết và số ca hồi phục đi ngang theo thời gian.

#### Tổng kết

* Tình hình chống dịch ở Đức vẫn đang duy trì ở mức khá tốt do đã có sự suy giảm rõ rệt trong số ca nhiễm mới mỗi ngày.
* Nhìn chung tuy tình hình dịch bệnh căng thẳng nhưng các nước vẫn làm tốt công tác phòng chống dịch. Riêng Hàn Quốc đang có dấu hiệu tệ đi khi số ca chết tăng nhưng số ca hồi phục lại giảm cùng với số ca nhiễm ở mỗi ngày đang có dấu hiệu tăng lên.

### Tình hình dịch bệnh của 5 quốc gia có tỉ lệ tử vong cao nhất

#### Mô tả

* + Các vấn đề cần quan tâm:

+ Xác định xem số ca nhiễm mới có xu hướng ra sao, có cùng thay đổi chung với số ca tử vong hay không?

+ Xác định tình hình dịch bệnh của 5 nước cũng như xem xét khả năng hồi phục có thay đổi như thế nào. Đồng thời cũng thể hiện sự tương quan giữa số ca tử vong so với số ca hồi phục

#### Các biểu đồ và nhận xét

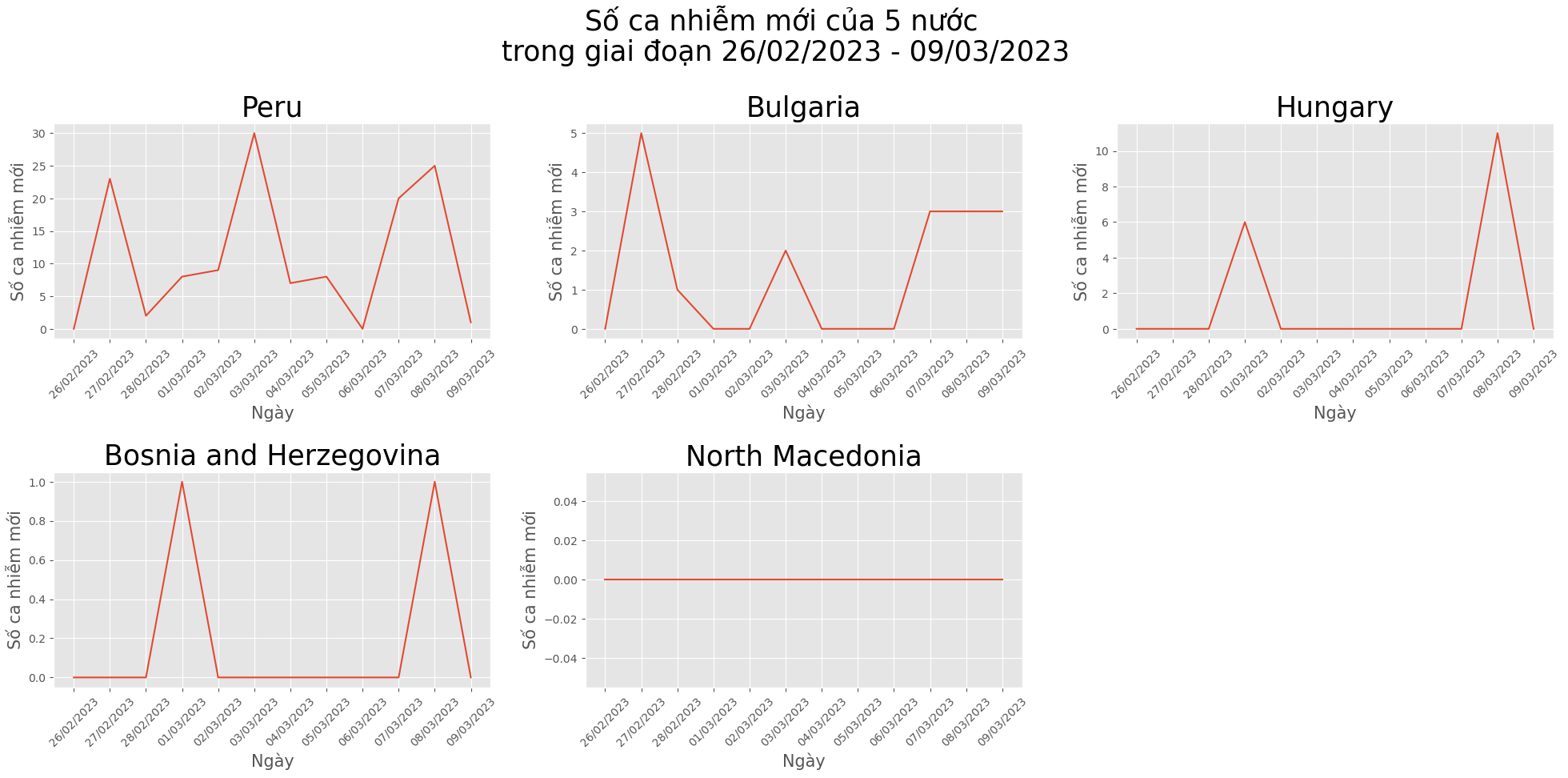
##### Tổng số ca tử vong

Chart, bar chart

Description automatically generated

* + - **Nhận xét:**
* Nhìn chung, tỷ lệ tử vong (xét trên 1 triệu dân số) của 5 nước đều cao, trong đó Peru dẫn đầu với tổng số ca tử vong là 6000 người chết/1M

##### Tổng số ca nhiễm mới



**\* Nhận xét:**

* + - * Số ca nhiễm mới của 2 nước Peru và Bulgaria biến động khá lớn, trong khi đó nước North Macedonia không ghi nhận thêm ca nhiễm mới nào
      * Nước Hungary có số ca nhiễm mới đạt đỉnh điểm vào 2 ngày 1/3/2023 và 8/3/2023, khu vực Bosnia and Herzegovina có số ca nhiễm mới đạt đỉnh điểm vào 2 ngày 28/2/2023 và 8/3/2023

##### Tỷ lệ số ca tử vong và số ca hồi phục

**A picture containing graphical user interface

Description automatically generated**

* + - **Nhận xét:**

+ Nhìn chung, số ca tử vong của 5 nước có tỉ lệ tử vong cao nhất có xu hướng giảm

+ Về số ca tử vong mới, 2 nước Peru và Bulgaria có số ca tử vong mới biến động khá lớn và đạt giá trị khá cao so với các nước còn lại. Trong khi đó North Macedonia không ghi nhận thêm ca tử vong mới nào. Nước Hungary và khu vực Bosnia và Herzegovina có số ca tử vong mới đạt đỉnh điểm vào 2 ngày 1/3/2023 và 8/3/2023

+ Về số ca hồi phục mới, 2 nước Peru và Bulgaria cũng có số ca hồi phục mới biến động mạnh và đạt giá trị khá cao so với các nước còn lại. Trong khi đó North Macedonia và khu vực Bosnia và Herzegovina không ghi nhận thêm ca hồi phục mới nào. Nước Hungary có số ca hồi phục mới đạt đỉnh điểm vào 2 ngày 1/3/2023 và 8/3/2023

##### Tổng số ca test

Chart

Description automatically generated

* + - **Nhận xét:**

**+** Vì số ca tử vong ở nước Peru biến động mạnh và đạt giá trị cao, do đó chính phủ đã tăng cường kiểm soát dịch bệnh bằng việc thực hiện việc test, từ biểu đồ cho thấy số ca test tăng mạnh, cao nhất là 1,13 triệu ca/ngày. Bên cạnh đó, nước Bulgaria cũng có tỷ lệ thực hiện test khá cao

**+** Ngoài ra, tỷ lệ ca test ở các nước Hungary, Bosnia and Herzegovina, North Macedonia cũng đạt mức cao nhưng không tăng dần theo thời gian

#### Tổng kết

* + - * Từ biểu đồ tỷ lệ số ca hồi phục và số ca tử vong, ta thấy nước Hungary có số ca tử vong mới đạt đỉnh điểm vào 2 ngày 1/3/2023 và ngày 8/3/2023, và trong 2 ngày này, số ca hồi phục ở quốc gia này cũng đạt đỉnh điểm, điều đó cho thấy công tác phòng chống dịch tại Hungary được tăng cường mạnh vào 2 mốc thời gian này
      * Cũng vào mốc thời gian này, số ca tử vong ở khu vực Bosnia and Herzegovina cũng tăng đột biến, nhưng số ca hồi phục lại không được ghi nhận thêm
      * Tại quốc gia Peru, số ca nhiễm mới và số ca tử vong đồng thời tăng đột biến vào ngày 3/3/2023 cho thấy tình hình dịch bệnh ở quốc gia này đang diễn biến phức tạp vào thời gian này.