**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

---------------o0o---------------

LAB 1: DATA RELATIONSHIP

**Giáo viên phụ trách:** Lê Ngọc Thành

**Nhóm thực hiện:** Nhóm 1

20127023 - Trần Minh Hiệp

20127058 - Lê Hồ Bảo Nhật

20127088 - Nguyễn Thiện Hoàng Trí

20127288 - Nguyễn Nam Phương

20127603 - Ninh Trần Hoàng Quân

**Lớp: 20TQHDL**

TP.Hồ Chí Minh, tháng 1 năm 2023

**MỤC LỤC**

[**1. Tổng quan thành viên :** 3](#_Toc129974087)

[**2. Tổng quan đồ án:** 4](#_Toc129974088)

[**2.1. Giới thiệu đồ án:** 4](#_Toc129974089)

[**2.2. Yêu cầu đồ án:** 5](#_Toc129974090)

[**2.3. Đánh giá:** 6](#_Toc129974091)

[**3. Phân công công việc :** 7](#_Toc129974092)

[**4. Mức độ hoàn thành tổng thể của mỗi yêu cầu:** 8](#_Toc129974093)

[**5. Mức độ hoàn thành của từng thành viên:** 9](#_Toc129974094)

[**6. Môi trường cài đặt:** 10](#_Toc129974095)

[**7. Các hàm được cài đặt:** 10](#_Toc129974096)

[**8. Thu thập và tiền xử lý dữ liệu:** 11](#_Toc129974097)

[**8.1. Thu thập dữ liệu:** 11](#_Toc129974098)

[**8.2. Xử lý dữ liệu:** 11](#_Toc129974099)

[**9. Trực quan hoá:** 12](#_Toc129974100)

[**9.1. Vẽ biểu đồ heatmap sau khi đã xử lí để quan sát sự tương quan giữa các cột còn lại với nhau:** 12](#_Toc129974101)

[**9.2. Vẽ biểu đồ Radarchart thể hiện mối tương quan của tổng số ca, ca tử vong, ca hồi phục, ca điều trị, số ca tử vong / tổng ca.:** 13](#_Toc129974102)

[**9.3. Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca nhiễm bệnh và tổng số ca xét nghiệm:** 14](#_Toc129974103)

[**9.4. Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca và tổng số dân.:** 15](#_Toc129974104)

[**9.5. Vẽ biểu đồ Scatterplot để xem có mối tương quan của số ca hồi phục và số ca điều trị không:** 16](#_Toc129974105)

[**9.6. Vẽ biểu đồ Barchart để thể hiện mối quan hệ giữa của tổng ca tử vong, ca hồi phục, ca điều trị giữa các khu vực:** 17](#_Toc129974106)

[**9.7. Vẽ biểu đồ Histogram để thể hiện sự phân bố của các quốc gia về tổng ca và số ca tử vong bằng histogram:** 18](#_Toc129974107)

[**9.8. Vẽ biểu đồ Donutchart để thể hiện mối quan hệ giữa của tổng số ca, ca tử vong, ca hồi phục, ca điều trị giữa các khu vực.:** 19](#_Toc129974108)

[**9.9. Vẽ biểu đồ Boxplot thể hiện sự chênh lệch ca nhiễm khu vực, sau đó dung kiểm định thống kê để xem sự chênh lệch ca nhiễm khu vực:** 21](#_Toc129974109)

[**10. Đánh giá đóng góp:** 22](#_Toc129974110)

[**11. Tham khảo:** 22](#_Toc129974111)

# **1. Tổng quan thành viên :**

- Mã nhóm: nhóm 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | MSSV | HỌ TÊN | EMAIL | VAI TRÒ |
| 1 | 20127023 | Trần Minh Hiệp | 20127023@student.hcmus.edu.vn | Member |
| 2 | 20127058 | Lê Hồ Bảo Nhật | 20127058@student.hcmus.edu.vn | Leader |
|  | 20127088 | Nguyễn Thiện Hoàng Trí | 20127088@student.hcmus.edu.vn | Member |
| 3 | 20127288 | Nguyễn Nam Phương | 20127288@student.hcmus.edu.vn | Member |
| 4 | 20127603 | Ninh Trần Hoàng Quân | 20127603@student.hcmus.edu.vn | Member |

# **2. Tổng quan đồ án:**

## **2.1. Giới thiệu đồ án:**

Graphical user interface, application, table

Description automatically generated

Từ khoảng cuối năm 2019 và đầu năm 2020, một bệnh dịch hạch lan tràn khủng khiếp trên toàn thế giới. Mỗi ngày có hàng ngàn người bị nhiễm và hàng chục đến hàng trăm người chết. Tổ chức Worldometer (www.worldometers.info) đã thu thập dữ liệu thống kê từ nhiều nguồn và từ nhiều quốc gia báo cáo hàng ngày để tổng hợp thành một bảng trong Hình 1. Trong trang web, tổ chức Worldmeter cũng thực hiện vẽ biểu đồ để cho thấy sự thay đổi trực quan tình hình diễn biến dịch bệnh. Tuy nhiên chúng ta tạm thời không sử dụng nó.

Bạn và nhóm của mình được quốc gia giao trọng trách để tìm hiểu dữ liệu này như giữa các trường dữ liệu có mối quan hệ gì không, liệu có bất thường trong dữ liệu hay không như báo cáo quốc gia khác với dữ liệu tổng hợp, sự bất bình thường trong việc nhảy số liệu, ….Nhiệm vụ này đòi hỏi nhiều kiến thức liên quan như Trực quan hóa dữ liệu, Phân tích dữ liệu thông minh, Học máy, … Tuy nhiên, nhóm bạn là biệt đội chuyên làm trực quan hóa dữ liệu nên chúng ta sẽ tập trung hướng này trước.

## **2.2. Yêu cầu đồ án:**

Cụ thể trong lab này, nhóm được yêu cầu thực hiện các nhiệm vụ sau:

* Thu thập số liệu thống kê từng ngày từ trang Worldmeter.
* Nhóm sinh viên (NSV) có thể chọn làm trên 1 ngày xác định. Do trang Worldometer chỉ thể hiện ngày hôm nay và ngày hôm qua nên nhóm cần thu thập nhiều ngày để có thể thực hiện tiếp cho các bài lab tiếp theo.
* NSV có thể thủ công để chép dữ liệu và lưu trữ vào định dạng chuẩn .CSV hoặc sử dụng code để lấy dữ liệu (khuyến khích)
* NSV có thể tiền xử lý dữ liệu trước khi chuyển sang pha tiếp theo nhưng cần báo cáo vấn đề này trong mục Tiền xử lý dữ liệu. Dữ liệu gốc và dữ liệu đã điều chỉnh cần lưu lại và nộp kèm trong bài nộp.
* Sử dụng nhận xét, code/thuật toán để thể hiện trực quan các mối quan hệ giữa các trường dữ liệu
* NSV thảo luận và chọn ra các trường dữ liệu để thể hiện trực quan bằng các loại biểu đồ đã học.
* Việc chọn biểu đồ cần giải thích tính phù hợp với tính chất trường dữ liệu. Có thể sử dụng nhiều hơn 1 loại biểu đồ cho trường dữ liệu nhưng cần giải thích lí do.
* Việc thể hiện quan hệ phải tích hơp dần dần nghĩa là từ đơn giản đến phức tạp, từ một trường đơn đến quan hệ giữa nhiều trường, …
* Ngoài quan hệ độc lập, NSV xem xét liệu trong dữ liệu có quan hệ nhân quả không (cause-effect). Ví dụ: liệu có thể có mối quan hệ giữa tỉ lệ ca nhiễm tăng với số ca chết không, … Cần chứng minh thông qua các phép trực quan dữ liệu.
* NSV không cần phải làm hết tất cả các quan hệ nhưng nhiều nhất có thể và phủ được nhiều loại biểu đồ đã học.
* NSV giữ lại các dữ liệu để có thể thực hiện tiếp cho các bài sau.

## **2.3. Đánh giá:**

- Các tiêu chí đánh giá:

1. Thu thập và tiền xử lý dữ liệu (5%)

2. Chọn lựa, giải thích, trực quan các trường và các mối quan hệ giữa chúng (50%)

3. Rút ra ý nghĩa hợp lý sau mỗi dữ liệu được trực quan (20%)

4. Xem xét trên nhiều quan hệ, nhiều góc nhìn khác nhau (10%)

5. Báo cáo trình bày bố cục và định dạng hợp lý, rõ ràng (15%)

Lưu ý: nếu số quan hệ quá ít thì sẽ xem xét giảm tỉ lệ ở mức 2 và 3.

# **3. Phân công công việc :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Y ÊU CẦU | NỘI DUNG | NGƯỜI THỰC HIỆN | GHI CHÚ |
| 1 | Thu thập dữ liệu. | Thu thập dữ liệu từ web. | Lê Hồ Bảo Nhật |  |
| Thu thập dữ liệu từ web. | Ninh Trần Hoàng Quân |  |
| 2 | Tiền xử lý dữ liệu. | Kiểm tra số dòng, số cột, các dòng trùng. | Nguyễn Thiện Hoàng Trí |  |
| Kiểm tra kiểu dữ liệu số và loại. | Trần Minh Hiệp |  |
| 3 | Trực quan và nhận xét. | Vẽ biểu đồ heatmap sau khi đã xử lí để quan sát sự tương quan giữa các cột còn lại với nhau. | Ninh Trần Hoàng Quân |  |
| Vẽ biểu đồ Radarchart thể hiện mối tương quan của tổng số ca, ca tử vong, ca hồi phục, ca điều trị, số ca tử vong / tổng ca. | Lê Hồ Bảo Nhật |  |
| Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca nhiễm bệnh và tổng số ca xét nghiệm. | Nguyễn Thiện Hoàng Trí |  |
| Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca và tổng số dân. | Nguyễn Thiện Hoàng Trí |  |
| Vẽ biểu đồ Scatterplot để xem có mối tương quan của số ca hồi phục và số ca điều trị không. | Nguyễn Thiện Hoàng Trí |  |
| Vẽ biểu đồ Barchart để thể hiện mối quan hệ giữa tử vong, ca hồi phục, ca điều trị các khu vực. | Nguyễn Nam Phương |  |
| Vẽ biểu đồ Histogram để thể hiện sự phân bố của các quốc gia về tổng ca và số ca tử vong bằng histogram. | Nguyễn Nam Phương |  |
| Vẽ biểu đồ Donutchart để thể hiện mối quan hệ giữa tổng số ca, ca tử vong, ca hồi phục, ca điều trị giữa các khu vực. | Nguyễn Nam Phương |  |
| Vẽ biểu đồ Boxplot thể hiện sự chênh lệch ca nhiễm khu vực, sau đó dùng kiểm định thống kê để xem sự chênh lệch ca nhiễm khu vực. | Trần Minh Hiệp |  |
| 4 | Báo cáo. | Viết báo cáo. | Nguyễn Nam Phương |  |

# **4. Mức độ hoàn thành tổng thể của mỗi yêu cầu:**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | NỘI DUNG | MỨC ĐỘ | KHÓ KHĂN |
| 1 | Thu thập và tiền xử lý dữ liệu (5%) | Hoàn thành. | Không có. |
| 2 | Chọn lựa, giải thích, trực quan các trường và các mối quan hệ giữa chúng (50%) | Hoàn thành. | Không có. |
| 3 | Rút ra ý nghĩa hợp lý sau mỗi dữ liệu được trực quan (20%) | Hoàn thành. | Không có. |
| 4 | Xem xét trên nhiều quan hệ, nhiều góc nhìn khác nhau (10%) | Hoàn thành. | Không có. |
| 5 | Báo cáo trình bày bố cục và định dạng hợp lý, rõ ràng (15%) | Hoàn thành. | Không có. |

# **5. Mức độ hoàn thành của từng thành viên:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | YÊU CẦU | NỘI DUNG | NGƯỜI THỰC HIỆN | MỨC ĐỘ |
| 1 | Thu thập dữ liệu. | Thu thập dữ liệu từ web. | Lê Hồ Bảo Nhật | 100% |
| Thu thập dữ liệu từ web. | Ninh Trần Hoàng Quân | 100% |
| 2 | Tiền xử lý dữ liệu. | Kiểm tra số dòng, số cột, các dòng trùng. | Nguyễn Thiện Hoàng Trí | 100% |
| Kiểm tra kiểu dữ liệu số và loại. | Trần Minh Hiệp | 100% |
| 3 | Trực quan và nhận xét. | Vẽ biểu đồ heatmap sau khi đã xử lí để quan sát sự tương quan giữa các cột còn lại với nhau. | Ninh Trần Hoàng Quân | 100% |
| Vẽ biểu đồ Radarchart thể hiện mối tương quan của tổng số ca, ca tử vong, ca hồi phục, ca điều trị, số ca tử vong / tổng ca. | Lê Hồ Bảo Nhật | 100% |
| Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca và tổng số ca xét nghiệm. | Nguyễn Thiện Hoàng Trí | 100% |
| Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca và tổng số dân. | Nguyễn Thiện Hoàng Trí | 100% |
| Vẽ biểu đồ Scatterplot để xem có mối tương quan của số ca hồi phục và số ca điều trị không. | Nguyễn Thiện Hoàng Trí | 100% |
| Vẽ biểu đồ Barchart để thể hiện mối quan hệ giữa của tổng ca tử vong, ca hồi phục, ca điều trị giữa các khu vực. | Nguyễn Nam Phương | 100% |
| Vẽ biểu đồ Histogram để thể hiện sự phân bố của các quốc gia về tổng ca và số ca tử vong bằng histogram. | Nguyễn Nam Phương | 100% |
| Vẽ biểu đồ Donutchart để thể hiện mối quan hệ giữa của tổng số ca, ca tử vong, ca hồi phục, ca điều trị giữa các khu vực. | Nguyễn Nam Phương | 100% |
| Vẽ biểu đồ Boxplot thể hiện sự chênh lệch ca nhiễm khu vực, sau đó dùng kiểm định thống kê để xem sự chênh lệch ca nhiễm khu vực. | Trần Minh Hiệp | 100% |
| 4 | Báo cáo. | Viết báo cáo. | Nguyễn Nam Phương | 100% |

# **6. Môi trường cài đặt:**

Ngôn ngữ lập trình: Python.

Môi trường lập trình: Jupyter Notebook.

IDE: Visual Studio Code.

Thư viện được sử dụng:

+ Numpy

+ Pandas

+ Seaborn

+ Matplotlib

# **7. Các hàm được cài đặt:**

- Hàm processing:

+ Chức năng:

+ Input:

+ Output:

# **8. Thu thập và tiền xử lý dữ liệu:**

## **8.1. Thu thập dữ liệu:**

## **8.2. Xử lý dữ liệu:**

- Tại đây sẽ thực hiện việc đọc dữ liệu.

- Sau đó sẽ xem dữ liệu này có bao nhiêu dòng và bao nhiêu cột.

- Ý nghĩa từng dòng của data.

- Sau đó kiểm tra các dòng có bị trùng hay không.

- Nêu ý nghĩa của từng cột của data.

- Kiểm tra kiểu dữ liệu, xem sự phân phối của các kiểu dữ liệu rồi mới quyết định có cần phải xóa bỏ đi những cột không cần thiết.

Dữ liệu sau khi xử lý:

Graphical user interface

Description automatically generated with low confidence

# **9. Trực quan hoá:**

## **9.1. Vẽ biểu đồ heatmap sau khi đã xử lí để quan sát sự tương quan giữa các cột còn lại với nhau:**

- Trường dữ liệu:

- Loại biểu đồ:

- Ý tưởng:

- Lí do sử dụng:

- Biểu đồ:

- Nhận xét:

## **9.2. Vẽ biểu đồ Radarchart thể hiện mối tương quan của tổng số ca, ca tử vong, ca hồi phục, ca điều trị, số ca tử vong / tổng ca.:**

- Trường dữ liệu:

- Loại biểu đồ:

- Ý tưởng:

- Lí do sử dụng:

- Biểu đồ:

- Nhận xét:

## **9.3. Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca nhiễm bệnh và tổng số ca xét nghiệm:**

- Trường dữ liệu:

- Loại biểu đồ:

- Ý tưởng:

- Lí do sử dụng:

- Biểu đồ:

- Nhận xét:

## **9.4. Vẽ biểu đồ Scatterplot thể hiện mối tương quan của tổng số ca và tổng số dân.:**

- Trường dữ liệu:

- Loại biểu đồ:

- Ý tưởng:

- Lí do sử dụng:

- Biểu đồ:

- Nhận xét:

## **9.5. Vẽ biểu đồ Scatterplot để xem có mối tương quan của số ca hồi phục và số ca điều trị không:**

- Trường dữ liệu:

- Loại biểu đồ:

- Ý tưởng:

- Lí do sử dụng:

- Biểu đồ:

- Nhận xét:

## **9.6. Vẽ biểu đồ Barchart để thể hiện mối quan hệ giữa của tổng ca tử vong, ca hồi phục, ca điều trị giữa các khu vực:**

- Trường dữ liệu: totaldeaths, total recovered và activecases.

- Loại biểu đồ: Barchart.

- Ý tưởng: Ta sẽ sử dụng các dòng cần thiết như là tổng theo từng khu vực, sau đó sẽ lấy những trường cần thiết như đã nói trên, rồi gom nhóm theo khu vực. Sau đó thực hiện việc scale dữ liệu để có thể dễ dàng nhìn rõ từng cột của từng khu vực (Nếu không thực hiện thì các cột của total recovered và activecases sẽ rất khso nhìn thấy và so sánh).

- Lí do sử dụng: Có thể dễ dàng so sánh mỗi trường dữ liệu của các khu vực.

Chart, bar chart

Description automatically generated- Biểu đồ:

- Nhận xét:

+ Nhìn chung thì các khu vực có các ca hồi phục cao hơn những ca tử vong và ca đang điều trị. Trong đó có 2 khu vực cao khá tương đương nhau đó là Asia và Europe.

+ Số ca đang điều trị cũng ít hơn, số ca tử vong của các khu vực cũng ít dần.

+ Điều này là minh chứng của việc chính phủ đã có những biện pháp rất tốt để ngăn chặn dịch bệnh và cũng nói lên rằng điều kiện y tế cũng rất phát triển, cứu giúp kịp thời.

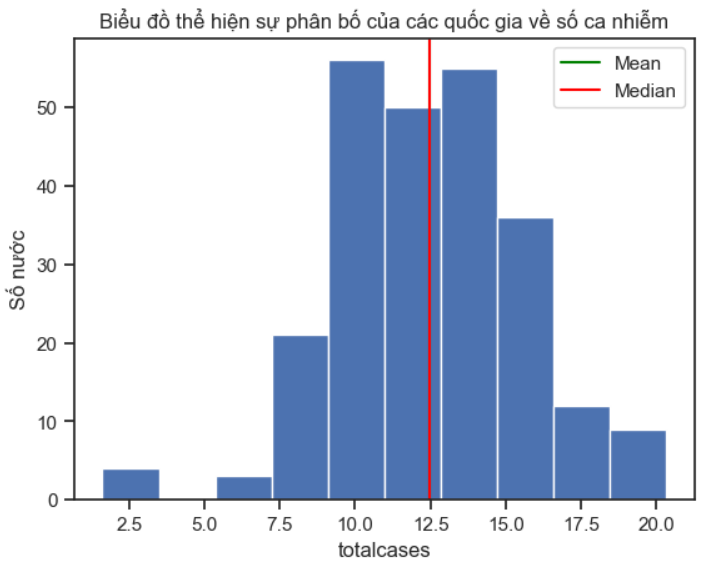
## **9.7. Vẽ biểu đồ Histogram để thể hiện sự phân bố của các quốc gia về tổng ca và số ca tử vong bằng histogram:**

- Trường dữ liệu: totalcases và totaldeaths.

- Loại biểu đồ: Histogram.

- Ý tưởng: Ta sẽ sử dụng 2 trường cần thiết như đã nêu rồi sau đó vẽ biểu đồ thể hiện sự phân bố của các quốc gia dựa vào các trường đó.

- Lí do sử dụng: Ở đây ta sẽ xem sự phân bố các nước dựa trên mỗi trường dữ liệu đã nêu.

- Biểu đồ:

Chart, histogram

Description automatically generated

- Nhận xét:

Đối với biểu đồ bên trái:

+ Sự phân bố ca nhiễm khá đều.

+ Ta thấy mean và median trùng nhau, đó chính là điều minh chứng cho ý trên.

+ Khoảng 12.5 thì có gần 50 nước như thế.

Đối với biểu đồ bên phải:

+ Sự phân bố ca tử vong cũng khá đều.

+ Ta thấy đường mean và median khá gần nhau, tức là số ca tử vong (đã được scale) là 8 thì có gần 50 nước.

## **9.8. Vẽ biểu đồ Donutchart để thể hiện mối quan hệ giữa của tổng số ca, ca tử vong, ca hồi phục, ca điều trị giữa các khu vực.:**

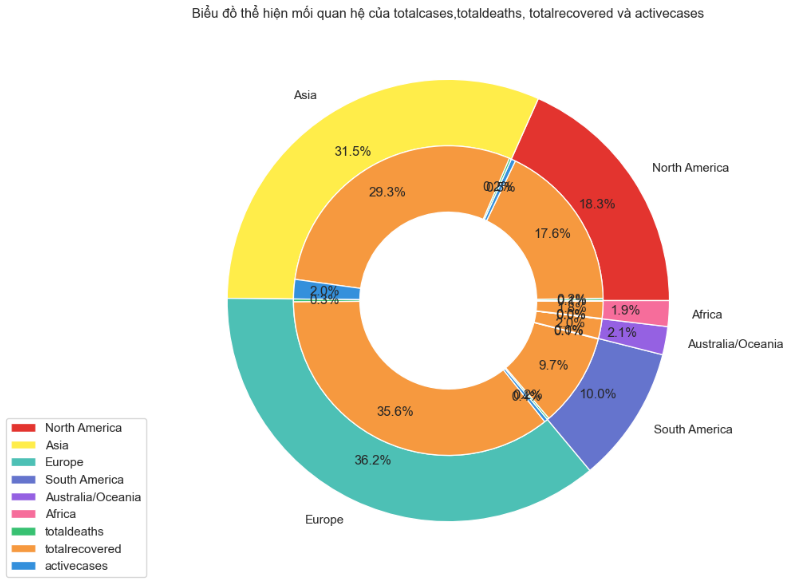
- Trường dữ liệu: totalcases, totaldeaths, totalrecovered và activecases.

- Loại biểu đồ: Donutchart.

- Ý tưởng: Ta sẽ sử dụng các dòng cần thiết như là tổng theo từng khu vực, sau đó sẽ lấy những trường cần thiết như đã nói trên, rồi gom nhóm theo khu vực. Rồi trực quan dữ liệu đó theo 2 biểu đồ:

+ Biểu đồ thứ nhất: Vòng tròn bên ngoài thể hiển phần trăm của từng khu vực đối với tổng ca; Vòng tròn bên trong sẽ là phần trăm lần lượt của số ca tử vong, số ca hồi phục và số ca điều trị của từng khu vực.

+ Biểu đồ thứ hai: Sẽ gồm 3 donutchart, mỗi donutchart sẽ đại diện của các trường dữ liệu total deaths, totalrecoverd và activecases và phần trăm của từng khu vực trong các trường đó.

- Lí do sử dụng: Ta đang xem xét mối quan hệ của từng khu vực so với tổng thể nên việc sử dụng biểu đồ donutchart sẽ rất là hợp lý.

- Biểu đồ:



- Nhận xét:

Đối với biểu đồ đầu tiên:

+ Ở vòng tròn thứ nhất, tức là vòng ngoài cùng, đó là totalcases của các khu vực. Trong đó các khu vực Asia, Europe chiếm tỉ trọng khá cao, cụ thể lần lượt là 31.5% và 36.2%, đồng nghĩa là các nước trong khu vực này sẽ nhiều tổng ca bao gồm ca nhiễm, ca hồi phục và ca tử vong. Tiếp theo, các khu vực chiếm tỉ trọng tương đối là South America và North America. Cuối cùng là các nước ở khu vực Africa và Australia/Oceania, chiểm tỉ trọng thấp nhất trong biểu, cũng có thể xem là tổng số ca ít nhất so với các khu vực khác.

+ Để có thể hiểu rõ hơn từng khu vực, ta có thể xem vòng tròn thứ hai, tức là vòng tròn trong cùng, mỗi khu vực sẽ bao gồm totaldeaths, totalrecovered, activecases để thể tỉ trọng trong từng khu vực. Như ta thấy thì số ca hồi phục chiểm nhiều hơn so với số ca nhiễm và số ca tử vong. Cũng chính vì điều đó đã cho ta thấy được y tế được cải thiện rất nhiều so với những năm trước, Chính phủ đã quan tâm tới sức khoẻ của người dân nhiều hơn hoặc là đã tìm ra các phương pháp cứu chữa để làm giảm số ca tử vong.

Đối với biểu đồ thứ hai:

+ Đối với totaldeaths,totalrecovered thì các khu vực NorthAmerica, Asia và Europe chiếm tỉ trọng khá lớn.

+ Đối với activecases thì Asia lại chiếm tỉ trọng lớn nhất.

## **9.9. Vẽ biểu đồ Boxplot thể hiện sự chênh lệch ca nhiễm khu vực, sau đó dung kiểm định thống kê để xem sự chênh lệch ca nhiễm khu vực:**

- Trường dữ liệu:

- Loại biểu đồ:

- Ý tưởng:

- Lí do sử dụng:

- Biểu đồ:

- Nhận xét:

# **10. Đánh giá đóng góp:**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | HỌ VÀ TÊN | MỨC ĐỘ ĐÓNG GÓP (%) | GHI CHÚ |
| 1 | Trần Minh Hiệp | 100% |  |
| 2 | Lê Hồ Bảo Nhật | 100% |  |
| 3 | Nguyễn Thiện Hoàng Trí | 100% |  |
| 4 | Nguyễn Nam Phương | 100% |  |
| 5 | Ninh Trần Hoàng Quân | 100% |  |

# **11. Tham khảo:**