


**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT
THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN**



BÁO CÁO ĐỒ ÁN CUỐI KỲ

**TÊN ĐỀ TÀI: NGHIÊN CỨU VỀ DÂN SỐ ƯỚC TÍNH HIỆN
TẠI CỦA CÁC NƯỚC TRÊN THẾ GIỚI**

 **GVHD : Quách Đình Hoàng**

 **Môn học: Nhập môn lập trình Python
cho phân tích**

 **Họ tên sinh viên :**

❖ **Nguyễn Thanh Bình - 20133025**

❖ **Nguyễn Nhật Triều - 20133102**

❖ **Đoàn Quốc Trung - 20133104**

Tp. Hồ Chí Minh, tháng 6 năm 2022

Mục lục:

1. Tóm tắt.....	3
2. Giới thiệu.....	3
3. Dữ liệu.....	4
4. Trực quan hóa dữ liệu.....	4
5. Mô hình hóa dữ liệu.....	7
6. Thực nghiệm, kết quả, thảo luận.....	8
7. Kết luận.....	12
8. Đóng góp.....	12
9. Tham khảo.....	12

1. Tóm tắt

- ❖ Vấn đề về dân số là một vấn đề gây nhức nhối cho nhiều nước trên thế giới. Nó gây ra nhiều vấn đề tiêu cực cho xã hội nếu không được quản lý kịp thời. Vì thế nhóm chúng tôi đã làm một bài phân tích dân số trên toàn thế giới dựa theo các số liệu đã thu thập được. Mục tiêu của bài là làm rõ các nguyên nhân có thể làm bùng nổ dân số, gia tăng tỷ lệ dân số già hóa,... Thông qua đó đưa ra các kết luận so sánh nhằm đánh giá chung về tình hình dân số thế giới cũng như một số biện pháp giúp ổn định tình hình dân số thế giới.
- ❖ Các vấn đề nghiên cứu :
 - ✚ Mật độ dân số khu vực Châu Âu nhỏ hơn Châu Phi .
 - ✚ Số người di cư của khu vực Bắc Mỹ lớn hơn Nam Mỹ.
 - ✚ Tỷ lệ sinh sản Châu Á nhỏ hơn Châu Âu .
- ❖ Các phương pháp nhóm sử dụng và kết quả :
 - ✚ Nhóm sử dụng phương pháp nghiên cứu quan sát kết hợp với phương pháp thu thập và xử lý số liệu. Đây là phương pháp được sử dụng trong trường hợp chúng ta chỉ quan sát, ghi lại số liệu, đồng thời trong số liệu có sự hỗn độn, dữ liệu chưa đáp ứng được cho quá trình nghiên cứu. Từ đó kết quả sẽ giúp khái quát đặc trưng tổng thể .
 - ✚ Kết quả :
 - ✓ Mật độ dân số khu vực Châu Âu lớn hơn Châu Phi
 - ✓ Số người di cư của khu vực Bắc Mỹ nhỏ hơn Nam Mỹ
 - ✓ Tỷ lệ sinh sản Châu Á lớn hơn Châu Âu

2. Giới thiệu

- ❖ Nhóm đã đưa ra những câu hỏi đó xuất phát từ bối cảnh tình hình dân số ở một số nước đang diễn biến một cách phức tạp và ngày càng khó lường. Cụ thể ở một số quốc gia, tình hình người dân vượt biên ở các quốc gia nghèo sang các quốc gia phát triển hơn đang ngày càng khó kiểm soát, ở một số quốc gia đang phát triển thì tỷ lệ sinh sản ngày một tăng cao, ở một số quốc gia phát triển thì hiện tượng mật độ dân số ngày càng đông ở các đô thị.
- ❖ Từ đó, chúng tôi xác định được dữ liệu đưa vào của bài toán là mật độ dân số, tổng số người di cư, tỷ lệ sinh sản. Sau đó, chúng tôi áp dụng các phương pháp xử lý số liệu và các phương pháp phân tích đánh giá để dự đoán được kết quả mà bài toán đã đưa ra như: mật độ dân số ở khu vực châu Âu lớn hơn châu Phi, tổng số người di cư ở Nam Mỹ lớn hơn Bắc Mỹ và tỷ lệ sinh ở châu Á lớn hơn châu Âu.

3. Dữ liệu

- ❖ Nguồn tài: <https://www.kaggle.com/datasets/anandhuh/countries-in-the-world-by-population-2022>
- ❖ Nguồn thu thập dữ liệu: <https://worldpopulationreview.com/>
- ❖ Cách thu thập dữ liệu : Thực hiện khảo sát dân số ở từng quốc gia , từng khu vực.
- ❖ Các đối tượng ở đây là các quốc gia trên thế giới . Bao gồm 11 cột (thuộc tính).
Các thuộc tính :
 - + Country/Other : Tên quốc gia và vùng lãnh thổ phụ thuộc.
 - + Population (2020) : Dân số năm 2020
 - + Yearly Change : Phần trăm thay đổi dân số hàng năm
 - + Net Change : Thay đổi thực về dân số
 - + Density (P/Km²) : Mật độ dân số (dân số trên km vuông)
 - + Land Area (Km²) : Diện tích đất của các quốc gia / vùng lãnh thổ phụ thuộc.
 - + Migrants (net) : Tổng số người di cư
 - + Fert. Rate : Tỷ lệ sinh
 - + Med. Age : Tuổi trung vị
 - + Urban Pop % : Phần trăm dân số thành thị
 - + World Share : Tỷ lệ dân số
- ❖ Nhóm thực hiện việc tiền xử lý dữ liệu . Phân chia các quốc gia trên thế giới thuộc những khu vực khác nhau . Chọn ngẫu nhiên các nước và xếp chúng theo các khu vực khác nhau .

4. Trực quan hóa dữ liệu (data visulization)

- Một số thống kê tóm tắt về các vấn đề dân số của các khu vực trên thế giới
 - o Khu vực Châu Âu (Europe)

Europe			
	Density (P/Km ²)	Migrants (net)	Med. Age
Mean	116.700000	64611.15000	41.800000
Median	99.500000	14600.00000	42.000000
Std	96.984046	132746.92855	3.302312

- o Khu vực Châu Phi (Africa)

Africal			
	Density (P/Km ²)	Migrants (net)	Med. Age
Mean	96.809524	-9046.857143	20.904762
Median	53.000000	-2000.000000	19.000000
Std	110.595488	58977.301998	5.566909

- Khu vực Bắc Mỹ (North American)

North American			
	Density (P/Km ²)	Migrants (net)	Med. Age
Mean	96.809524	-9046.857143	20.904762
Median	53.000000	-2000.000000	19.000000
Std	110.595488	58977.301998	5.566909

- Khu vực Nam Mỹ (South American)

South American			
	Density (P/Km ²)	Migrants (net)	Med. Age
Mean	96.809524	-9046.857143	20.904762
Median	53.000000	-2000.000000	19.000000
Std	110.595488	58977.301998	5.566909

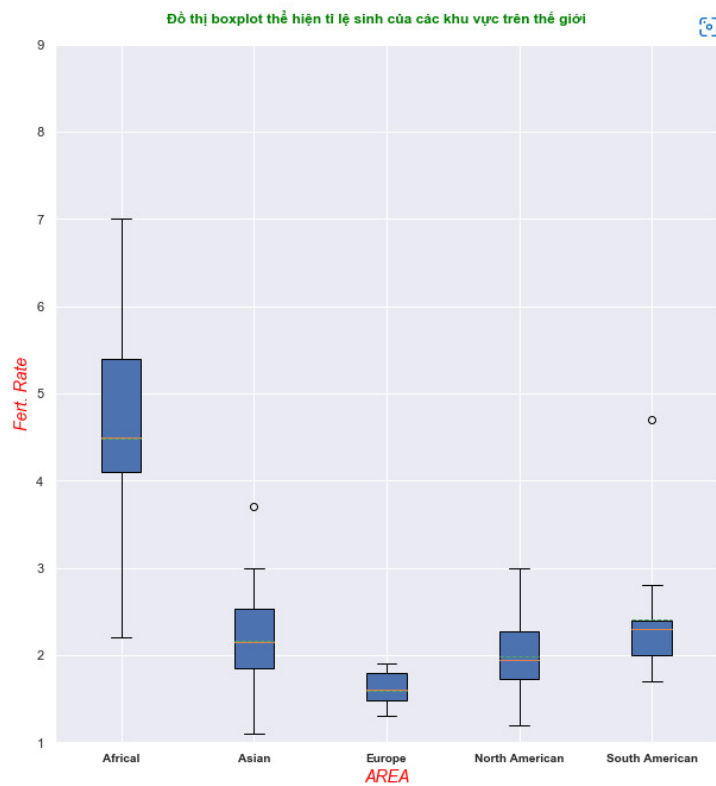
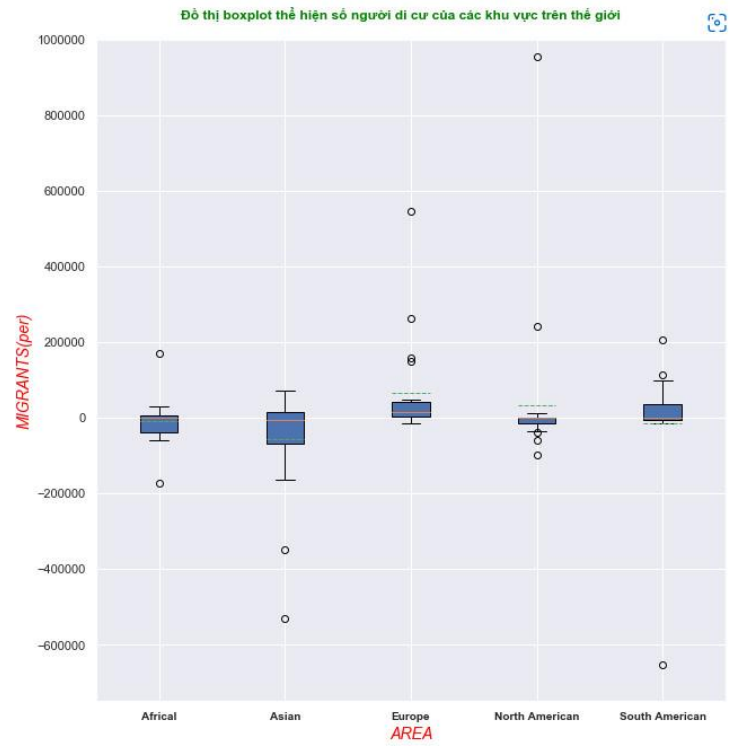
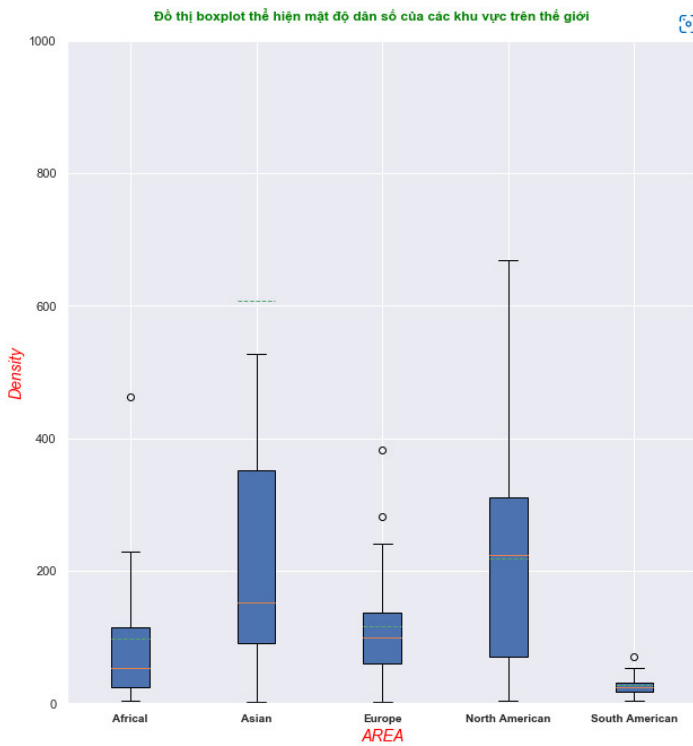
- Khu vực Châu Á(Asian)

Asian			
	Density (P/Km ²)	Migrants (net)	Med. Age
Mean	96.809524	-9046.857143	20.904762
Median	53.000000	-2000.000000	19.000000
Std	110.595488	58977.301998	5.566909

- Một số giải thích :

- Giá trị trung bình (Mean) : Là một trong những thước đo về xu hướng trung tâm của giá trị . Tính giá trị trung bình bằng cách : Tổng các giá trị chia cho số các giá trị .
- Giá trị trung vị (Median) : Là số nằm ở giữa trong một danh sách các số được sắp xếp tăng dần hoặc giảm dần và có thể mô tả nhiều hơn về tập dữ liệu so với giá trị trung bình, là một đại lượng thống kê mô tả dùng để đo mức độ phân tán của một tập dữ liệu .
- Độ lệch chuẩn : Là một đại lượng thống kê mô tả dùng để đo mức độ phân tán của một tập dữ liệu đã được lập thành bảng tần số.

- Biểu đồ trực quan hóa :



5. Mô hình hóa dữ liệu (data modeling)

Nhóm sử dụng phương pháp ước lượng khoảng để trong việc trả lời các câu hỏi đặt ra. Để tìm hiểu về phương pháp ước lượng khoảng, trước hết ta nên tìm hiểu về ước lượng điểm.

- Ước lượng điểm là ước lượng tham số chưa biết sử dụng một giá trị duy nhất dựa trên mẫu.
- Ước lượng khoảng bổ sung cho ước lượng điểm bằng cách cung cấp thông tin về sai số của ước lượng.

Ví dụ: ta tin tưởng 95% (95% confident) chiều cao trung bình của nam sinh ĐH SPKT là $X = 1.7 \pm 0.2$ m, tức trong khoảng (1.68, 1.72).

- Mức ý nghĩa alpha
 - Mức ý nghĩa alpha, (α), là một tiêu chí mà chúng ta sẽ sử dụng để quyết định có nên giữ lại hay loại bỏ giả thuyết đặt ra
 - Thông thường α được chọn là 0.05.
 - Khi ta đã chọn α , nếu sự khác biệt giữa thống kê trên mẫu và tham số của quần thể nhỏ hơn α , chúng ta có thể bác bỏ giả thuyết H_0 và kết luận rằng sự khác biệt này có lẽ không phải do tình cờ.
 - Khi ta bác bỏ giả thuyết H_0 , ta có thể sai (bác bỏ H_0 mặc dù nó đúng). Lỗi như vậy được gọi là lỗi loại 1.
 - Mức độ alpha, (α), đại diện cho tỷ lệ lỗi loại 1 mà chúng ta sẵn sàng chấp nhận trước khi tiến hành phân tích thống kê.
- Phân tích thống kê
 - Khi ta làm suy luận thống kê, ta muốn biết một hiện tượng mà ta quan sát được trên mẫu có đại diện cho một hiện tượng thực tế trên quần thể hay không.
 - Ta lập giả thuyết vô hiệu H_0 là không có sự khác biệt
 - Ta chọn một mức ý nghĩa α làm tiêu chuẩn để chấp nhận hay bác bỏ giả thuyết
 - Tính giá trị p (p-value).
- ✓ Nếu $p < \alpha$, ta bác bỏ giả thuyết H_0 và kết luận sự khác biệt nhiều khả năng không phải do tình cờ.
 - Khi bác bỏ H_0 ta có khả năng mắc sai lầm, đây là sai lầm loại 1.
- ✓ Nếu $p > \alpha$, ta không bác bỏ được H_0 và kết luận sự khác biệt nhiều khả năng là do tình cờ hoặc dữ liệu quan sát được là không đủ để chứng tỏ rằng có sự khác biệt.
- t-test cho giá trị trung bình
 - Tính t-test cho giá trị trung bình với python. Dùng thư viện scipy

```

from scipy.stats import stats
def getP_Value(list1,list2):
    res = stats.ttest_ind(list1, list2,
                          equal_var=True)
    display(res)

```

6. Thực nghiệm, kết quả, và thảo luận

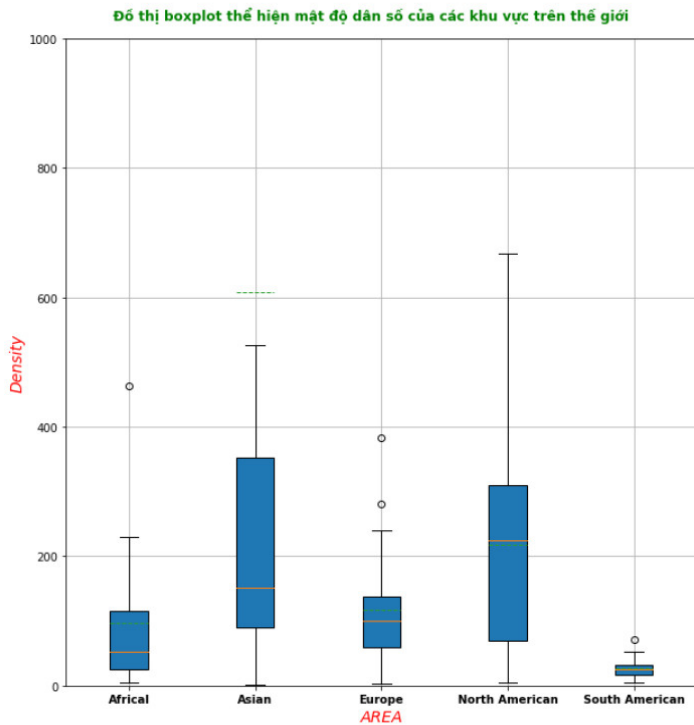
❖ Mô tả chi tiết về câu hỏi :

- ✚ Mật độ dân số khu vực Châu Âu lớn hơn Châu Phi .
- ✓ Mô tả : câu hỏi sẽ cho biết khu vực ở châu Âu sẽ lớn hoặc nhỏ hơn châu Phi, và nguyên nhân dẫn đến sự thay đổi như vậy và từ đó cho biết mật độ dân số có ảnh hưởng đến bùng nổ dân số hay không
- ✚ Số người di cư đến của khu vực Bắc Mỹ lớn hơn Nam Mỹ.
- ✓ Mô tả : câu hỏi sẽ cho biết số người di cư của khu vực Bắc Mỹ sẽ lớn hoặc nhỏ hơn Nam Mỹ, nguyên nhân tại sao dẫn đến sự di cư như vậy.
- ✚ Tỷ lệ sinh sản Châu Á nhỏ hơn Châu Âu .
- ✓ Mô tả : câu hỏi sẽ cho biết tỷ lệ sinh sản châu Á và châu Âu chênh lệch như thế nào. Từ đó ta có thể biết được liệu dân số có ảnh hưởng đến tỷ lệ sinh sản hay không.

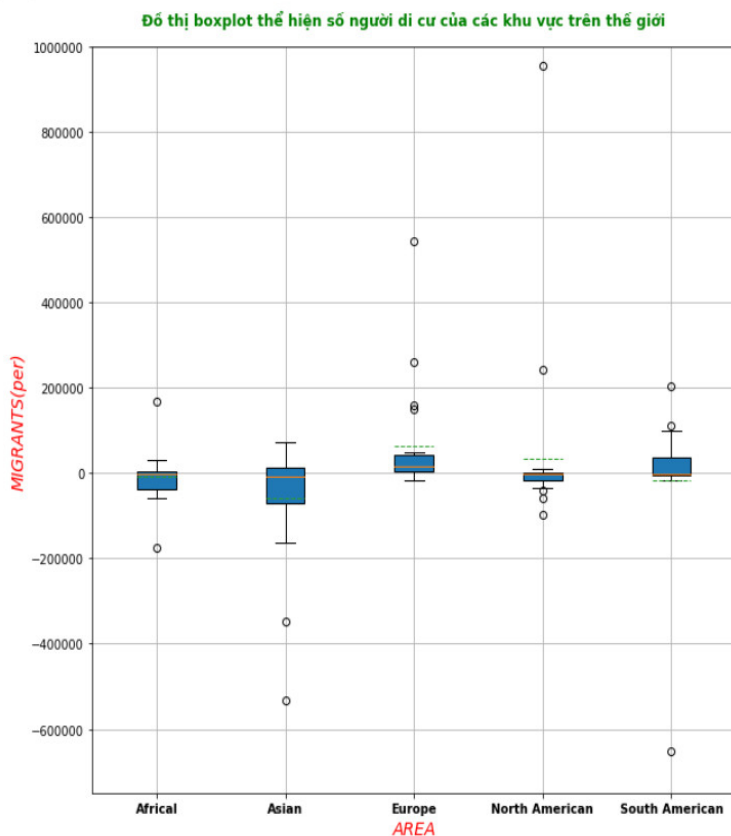
❖ Phương pháp:

- ✚ Nghiên cứu này là nghiên cứu quan sát, vì chúng ta chỉ quan sát và ghi lại tình trạng của các đặc điểm / đặc tính và sự kiện xảy ra trong nghiên cứu, và không ảnh hưởng đến đối tượng nghiên cứu.
- ✚ Phương pháp thu thập và xử lý số liệu. Đây là phương pháp được sử dụng trong trường hợp số liệu có sự hỗn độn, dữ liệu chưa đáp ứng được cho quá trình nghiên cứu. Từ đó kết quả sẽ giúp khái quát đặc trưng tổng thể .
- ❖ Thuật toán: ở đây vì sử dụng tập dữ liệu không có biến phân loại cụ thể nên chúng tôi không áp dụng các thuật toán mô hình hóa dữ liệu
- ❖ Các bảng được sử dụng như :
 - ✚ Density (P/Km²) : Mật độ dân số (dân số trên km vuông)
 - ✚ Migrants (net) : Tổng số người di cư
 - ✚ Fert. Rate : Tỷ lệ sinh
- ❖ Các biểu đồ sử dụng
 - ✚ Boxplot
 - ✚ Plot

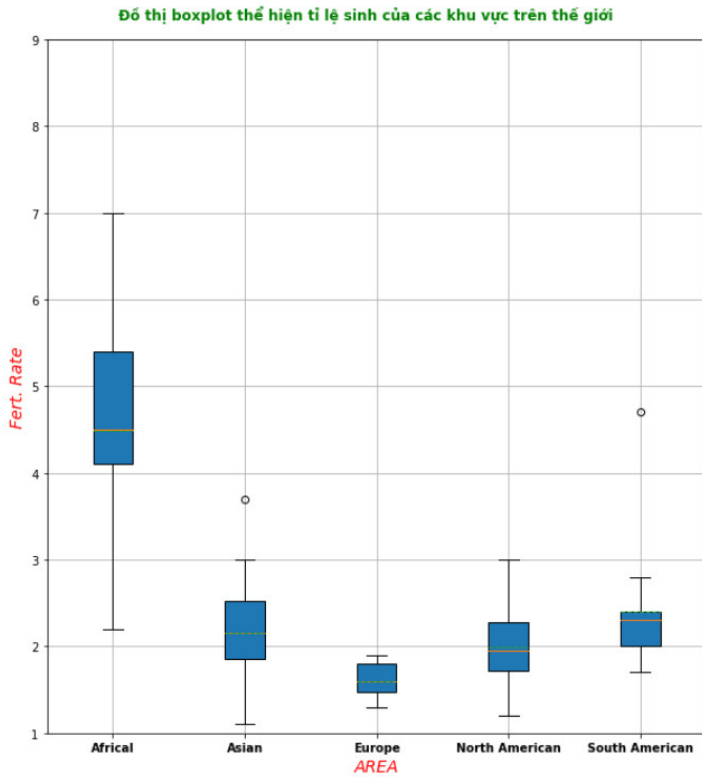
❖ Các biểu đồ trực quan và kết quả:



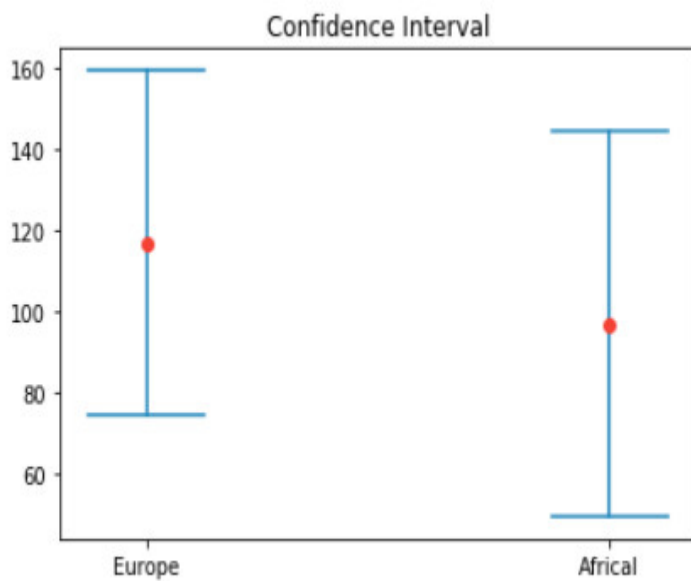
*Nhận xét:: Theo biểu đồ này , ta dễ dàng thấy được giá trị trung bình và trung vị cũng như khoảng giá trị ở từng khu vực . Theo đó , ta thấy giá trị trung bình và trung vị ở khu vực Bắc Mỹ cao nhất, giá trị trung bình và trung vị ở khu vực Nam Mỹ thấp nhất, đồng thời khoảng giá trị ở khu vực Nam Mĩ là thấp nhất.



*Nhận xét:: Theo biểu đồ này , ta dễ dàng thấy được giá trị trung bình và trung vị cũng như khoảng giá trị ở từng khu vực . Theo đó , ta thấy giá trị trung bình và trung vị ở khu vực Châu Âu cao nhất, giá trị trung bình và trung vị ở khu vực châu Á thấp nhất, đồng thời khoảng giá trị ở khu vực Bắc Mỹ là thấp nhất.



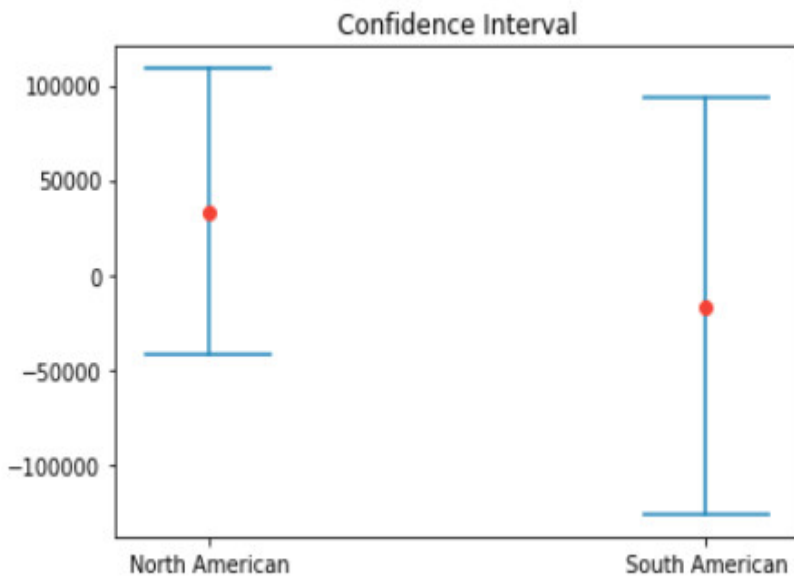
*Nhận xét:: Theo biểu đồ này , ta dễ dàng thấy được giá trị trung bình và trung vị cũng như khoảng giá trị ở từng khu vực . Theo đó , ta thấy giá trị trung bình và trung vị ở khu vực Châu Phi cao nhất và giá trị trung bình, giá trị trung vị cũng như khoảng giá trị ở khu vực Châu Âu là thấp nhất.



*Nhận xét:: - Với ước tính trung bình của 2 khu vực Châu Âu và Châu Phi và đồ thị tin cậy , ta tin tưởng rằng 95% mật độ dân số ở khu vực Châu Âu nằm trong khoảng (74,159) , khu vực Châu Phi nằm trong khoảng (50,144)

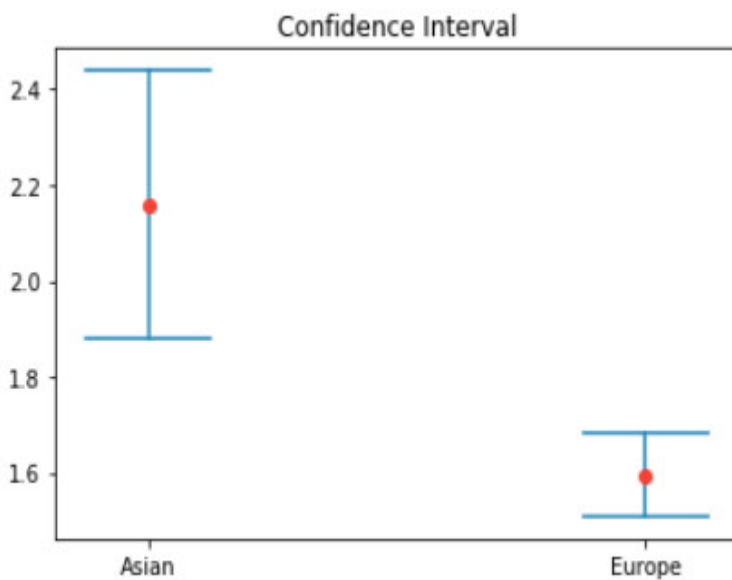
=> Các giá trị trung bình về mật độ dân số của Châu Âu nằm trong khoảng lớn hơn giá trị trung bình về mật độ dân số của Châu Phi

Biểu đồ thể hiện độ tin cậy về mật độ dân số của châu Âu và châu Phi



*Nhận xét:: Với ước tính trung bình của 2 khu vực Bắc Mỹ và Nam Mỹ và đồ thị tin cậy, ta tin tưởng rằng 95% số lượng người di cư ở khu vực Bắc Mỹ nằm trong khoảng $(-41885, 108721)$, khu vực Nam Mỹ nằm trong khoảng $(-126368, 93240)$ => Các giá trị trung bình về tổng số người di cư đến khu vực Bắc Mỹ nằm trong khoảng lớn hơn giá trị trung bình về tổng số người di cư đến khu vực Nam Mỹ.

Biểu đồ thể hiện độ tin cậy về người di cư của Bắc Mỹ và Nam Mỹ



*Nhận xét:: Qua kiểm định giả thuyết với tỷ lệ sinh sản ở khu vực châu Á và khu vực châu Âu, ta có giả thuyết ban đầu H_0 : tỷ lệ sinh sản ở khu vực châu Á lớn hơn khu vực châu Âu là đúng, điều này phù hợp với khoảng tin cậy đã tính ở trước đó.

=> Từ đó, ta có thể kết luận rằng tỷ lệ sinh sản ở khu vực châu Á lớn hơn khu vực châu Âu.

Biểu đồ thể hiện độ tin cậy về tỉ lệ sinh của châu Á và châu Âu

7. Kết luận

Các kết quả chính của nhóm :

- Qua kiểm định giả thuyết với mật độ dân số ở khu vực châu Âu và khu vực châu Phi, ta có giả thuyết ban đầu : mật độ dân số khu vực châu Phi nhỏ hơn ở châu Âu là đúng , giả thuyết thứ 2 : tổng số người di cư ở khu vực Bắc Mỹ nhỏ hơn khu vực Nam Mỹ là đúng , giả thuyết 3 : tỷ lệ sinh sản ở khu vực châu Á nhỏ hơn khu vực châu Âu là sai .
- Bởi vì vấn đề dân số là vấn đề toàn cầu , nên có thể tìm hiểu qua nhiều thông tin đại chúng . Với quy mô một bài phân tích nhỏ để đưa ra các câu hỏi và kiểm tra tính đúng sai thì sẽ không gây ra ngạc nhiên cho nhóm . Điều thú vị ở đây là có thể hiểu biết được thêm về các đặc tính dân số ở mỗi khu vực trên thế giới .
- Nếu có thời gian nhiều hơn , nhóm sẽ thu thập nhiều dữ liệu hơn và đưa ra nhiều câu hỏi hơn để tiến hành phân tích , thống kê . Việc có nhiều dữ liệu hơn thì chất lượng câu hỏi đặt ra cũng như câu trả lời sẽ có niềm tin hơn và tính đúng đắn cũng cao hơn .

8. Đóng góp

Tên thành viên – MSSV	Đóng góp
Nguyễn Thanh Bình - 20133025	Đưa ra ý tưởng phân tích dữ liệu . Làm slide báo cáo. Trực quan hóa dữ liệu .
Nguyễn Nhật Triều - 20133102	Trực quan hóa dữ liệu . Tìm hiểu các thư viện trong python dùng để tính toán . Phân công công việc .
Đoàn Quốc Trung - 20133104	Thu thập hình ảnh biểu đồ về dân số . Trực quan hóa dữ liệu . Tìm hiểu thuật toán .

9. Tham khảo

- ❖ Thư viện sử dụng : pandas, matplotlib, seaborn, numpy, scipy
- ❖ Code :
 - ✓ <https://www.kaggle.com/code/kharismabima/world-population-data-analysis>
 - ✓ <https://www.marsja.se/how-to-perform-a-two-sample-t-test-with-python-3-different-methods/>