**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**uuu**

**BÁO CÁO BÀI TẬP LỚN**

***ĐỀ TÀI:* NHỮNG KHÍA CẠNH KHÁC NHAU CỦA DOANH NGHIỆP**

**GVHD:** Nguyễn Văn Bảy

**NHÓM:** 10

**SINH VIÊN THỰC HIỆN:** Phan Quang Sang - 2151010318

Tô Thái Việt Quang - 2151010306

**LỚP:** DH21CS02

**TPHCM, 01/2024**

**MỤC LỤC**

[**1.** **Nguồn dữ liệu** 1](#_Toc157537880)

[**2.** **Các thư viện** 1](#_Toc157537881)

[**3.** **Đọc Dữ liệu** 1](#_Toc157537882)

[**3.1.** **Companies Dataset** 1](#_Toc157537883)

[**3.2.** **Continents Dataset** 2](#_Toc157537884)

[**4.** **Xử lý dữ liệu** 2](#_Toc157537885)

[**4.1.** **Companies Dataset** 2](#_Toc157537886)

[**4.2.** **Continents Dataset** 4](#_Toc157537887)

[**4.3.** **Kết hợp và xử lý 2 dataframe** 5](#_Toc157537888)

[**5.** **Trực quan hóa dữ liệu** 6](#_Toc157537889)

[ **Top 5 ngành nghề có tổng số nhân viên nhiều nhất trong công ty** 6](#_Toc157537890)

[ **Top 5 lục địa có tổng số nhân viên nhiều nhất trong công ty trước và sau năm 2000** 7](#_Toc157537891)

[ **Các công ty được thành lập trước và sau năm 2000** 8](#_Toc157537892)

[ **Top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước và sau năm 2000** 11](#_Toc157537893)

[ **Top 10 quốc gia xuất hiện nhiều nhất** 14](#_Toc157537894)

[ **Số lượng công ty công nghệ thông tin theo quốc gia** 14](#_Toc157537895)

[**6.** **Gom cụm** 16](#_Toc157537896)

[ **Sử dụng biểu đồ Elbow để tìm số cụm tối ưu có thể sử dụng** 16](#_Toc157537897)

[ **Ước tính nhân viên hiện tại theo dữ liệu “Current\_employee\_estimate”** 17](#_Toc157537898)

[ **Ước tính tổng số nhân viên hiện tại theo dữ liệu “Total\_employee\_estimate”** 18](#_Toc157537899)

[**7.** **K-Nearest Neighbors** 20](#_Toc157537900)

[**8.** **Cây quyết định** 22](#_Toc157537901)

[ **Biểu đồ hiển thị của cây quyết định và bộ luật của cây quyết định** 22](#_Toc157537902)

[**9.** **So sánh 2 Model** 22](#_Toc157537903)

**TÓM TẮT**

**Giải thích cơ bản về bộ dữ liệu doanh nghiệp**:

**7+ Million Company Dataset:**

Bộ dữ liệu của hơn 7 triệu công ty - bao gồm URL Linkedin của 237 quốc gia, tên miền, quy mô công ty từ 1-10.000+, địa điểm công ty, số lượng nhân viên. Bộ dữ liệu này chứa thông tin về hơn 7 triệu doanh nghiệp, cung cấp cái nhìn toàn diện về thị trường doanh nghiệp và sự đa dạng trong các ngành công nghiệp.

**Continent List for 2021 Olympics in Tokyo Dataset:**

Tập dữ liệu này cung cấp thông tin về các lục địa tham gia trong Thế vận hội Tokyo năm 2021. Điều này có thể hữu ích để đánh giá tầm ảnh hưởng quốc tế của các doanh nghiệp và nhóm nghiên cứu thị trường.

**==>** Bằng cách kết hợp cả hai nguồn dữ liệu trên, nhóm hy vọng sẽ có được cái nhìn đa chiều và chi tiết về hình ảnh toàn cảnh của doanh nghiệp. Tập dữ liệu này sẽ là nguồn cơ sở cho quá trình khai phá dữ liệu và phân tích chi tiết về các khía cạnh kinh doanh quan trọng. Nhóm sẽ tiếp tục quá trình nghiên cứu và phát triển để đảm bảo rằng thông tin thu thập là chính xác và có ý nghĩa cho mục tiêu nghiên cứu của chúng tôi.

**Thông tin về các thuộc tính của Data:**

**Name** - Cho biết tên của công ty

**Domain** - Tên trang website của công ty đó

**Year** **founded** - Cho biết năm thành lập của công ty

**Industry** - Cho biết ngành nghề tại công ty

**Size** **range** - Cho biết phạm vi kích thước của công ty

**Locality** - Cho biết địa điểm mà công ty đó đang hoạt động

**Country** - Cho biết đất nước của công ty

**Linkedin** **url** - Cho biết đường dẫn Linkedin của công ty

**Current** **employee** **estimate** - Ước tính nhân viên hiện tại tại công ty

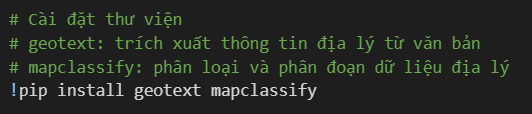
**Total** **employee** **estimate** - Ước tính tổng số nhân viên tại công ty

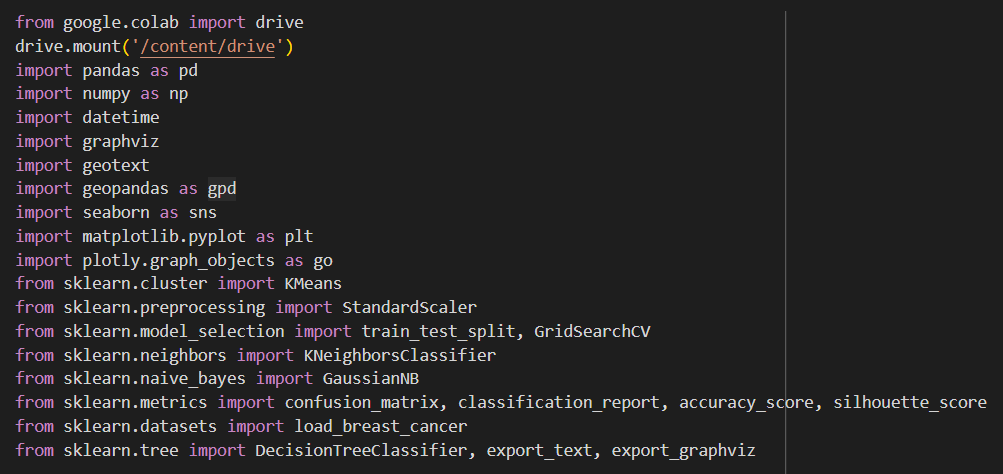
**Continent** - Cho biết lục địa của công ty

1. **Nguồn dữ liệu**

* Dữ liệu được lấy từ trang [kaggle](https://www.kaggle.com/), cụ thể [tại đây](https://www.kaggle.com/datasets/peopledatalabssf/free-7-million-company-dataset).

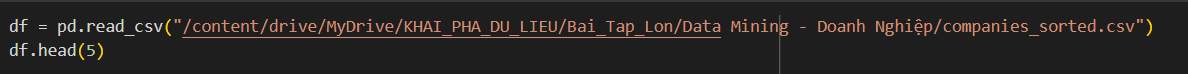
1. **Các thư viện**

****

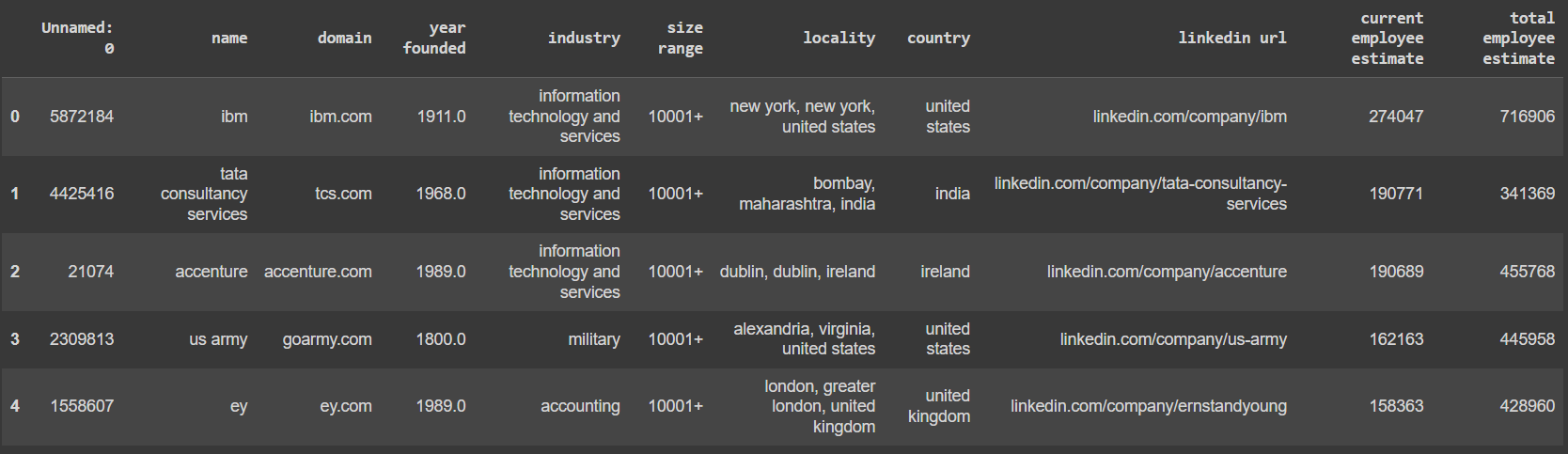
*Hình 2.1. Hình ảnh cài đặt thư viện vẽ bản đồ* ******

*Hình 2.2. Hình ảnh các thư viện*

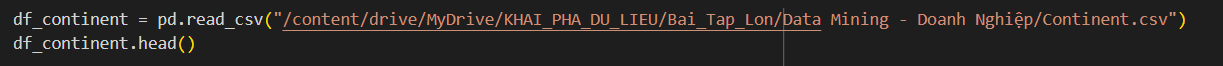
1. **Đọc Dữ liệu**
   1. **Companies Dataset**

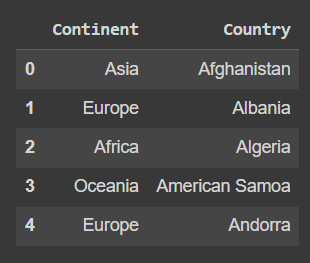
****

*Hình 3.1. Đọc dữ liệu từ file compaines\_sorted.csv*

*****Hình 3.2. Kết quả sau khi chạy file compaines\_sorted.csv*

* 1. **Continents Dataset**

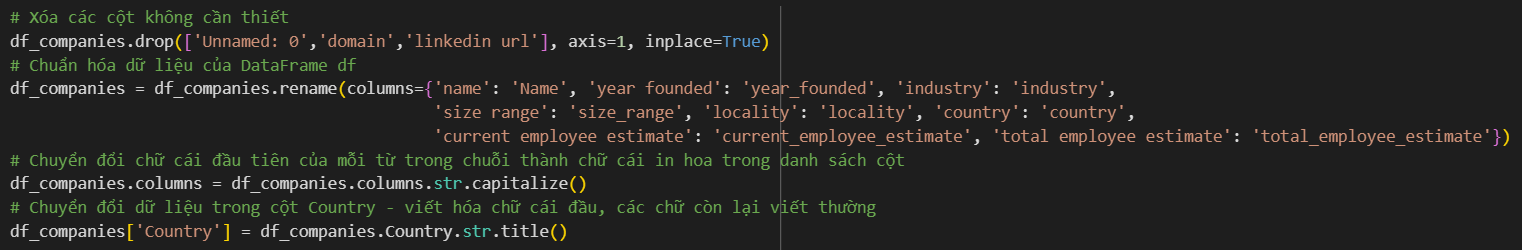
*Hình 3.3. Đọc dữ liệu từ file continent.csv*



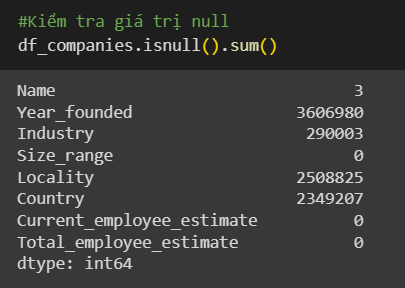
*Hình 3.4. Kết quả sau khi chạy file continent.csv*

1. **Xử lý dữ liệu**
   1. **Companies Dataset**

* Sau khi chạy thành công, xóa các cột không cần thiết. Chuẩn hóa lại các tên cột, chuyển đổi chữ cái đầu thành chữ viết hoa và các chữ còn lại viết thường (*Hình 4.1.1*)

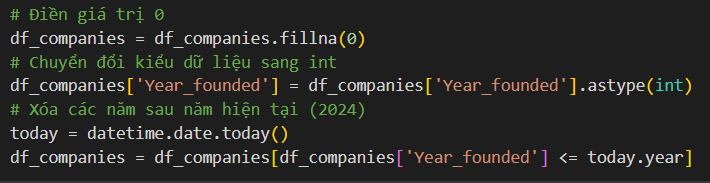
*Hình 4.1.1. Xóa các cột và xử lý chuỗi ký tự*

* Kế tiếp là kiểm tra giá trị null trong dataframe companies (*Hình 4.1.2*)



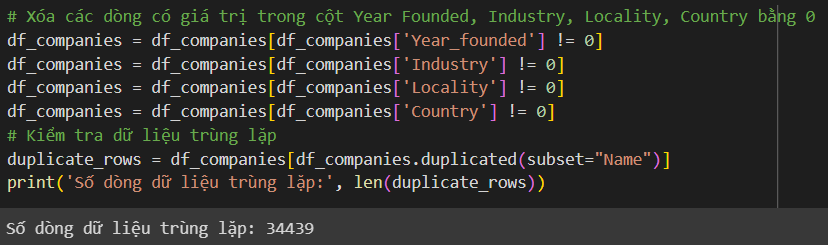
*Hình 4.1.2. Kiểm tra giá trị null trong Compaines*

* Sau khi kiểm tra có thấy giá trị null, điền giá trị 0 để các cột trong Compaines không còn null. Đồng thời chuyển đổi kiểu dữ liệu cột Year\_founder sang kiểu int. Xóa các năm sau năm hiện tại (*Hình 4.1.3*)

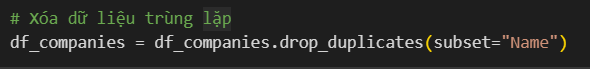


*Hình 4.1.3. Chuyển đổi các kiểu dữ liệu*

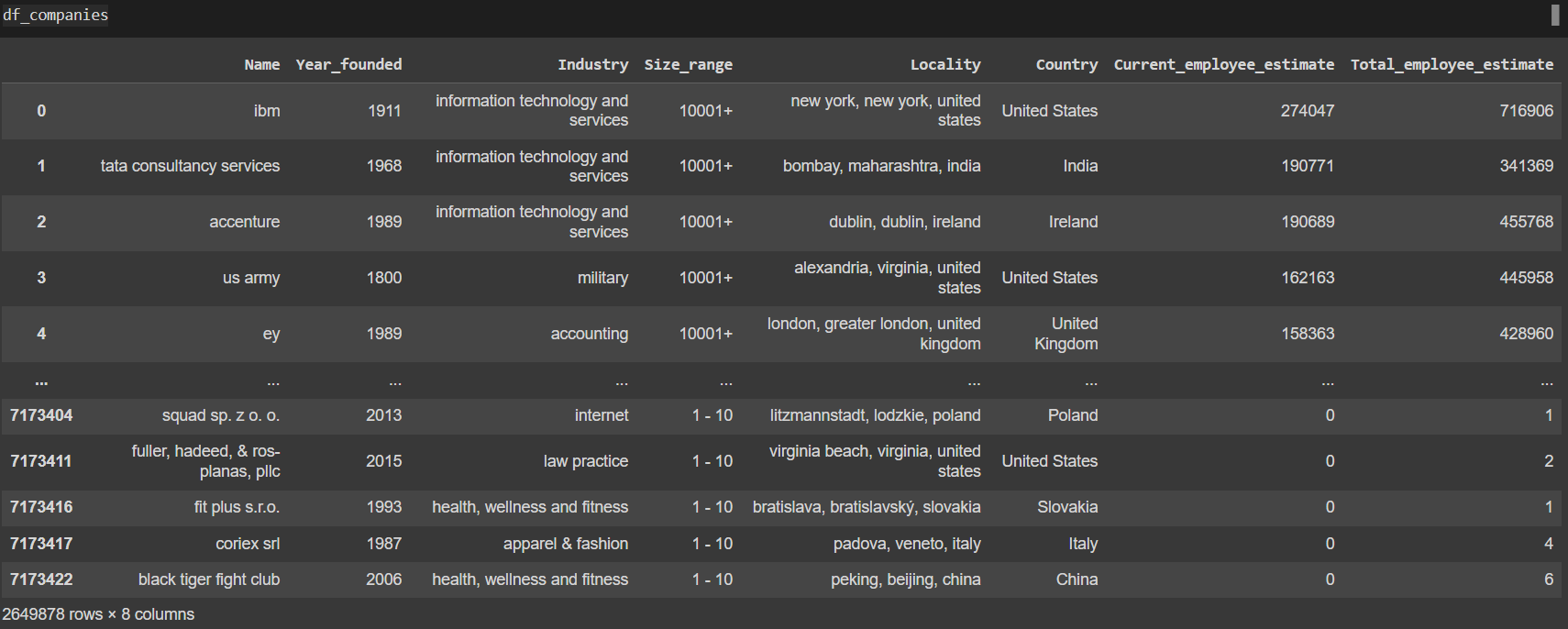
* Tiếp đến là xóa các cột có dòng dữ liệu có giá trị bằng 0, và kiểm tra-xóa dữ liệu trùng lặp của cột Name trong Compaines (*Hình 4.1.4* và *Hình 4.1.5*)



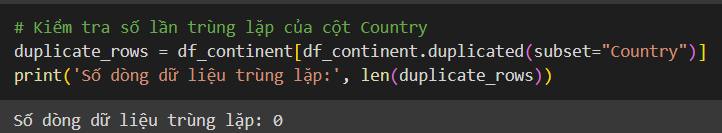
*Hình 4.1.4. Kiểm tra và xóa các dữ liệu trùng lặp*



*Hình 4.1.5. Xóa dữ liệu trùng lặp của cột Name*

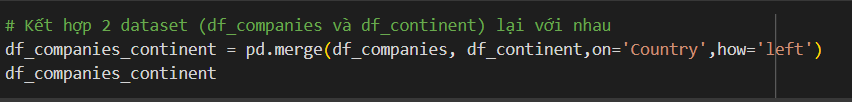
*Hình 4.1.6. Kết quả của Companies sau khi đã qua xử lý dữ liệu*

* 1. **Continents Dataset**
* Kiểm tra dữ liệu trùng lặp của cột Country trong dataframe Continent (*Hình 4.2.1*)

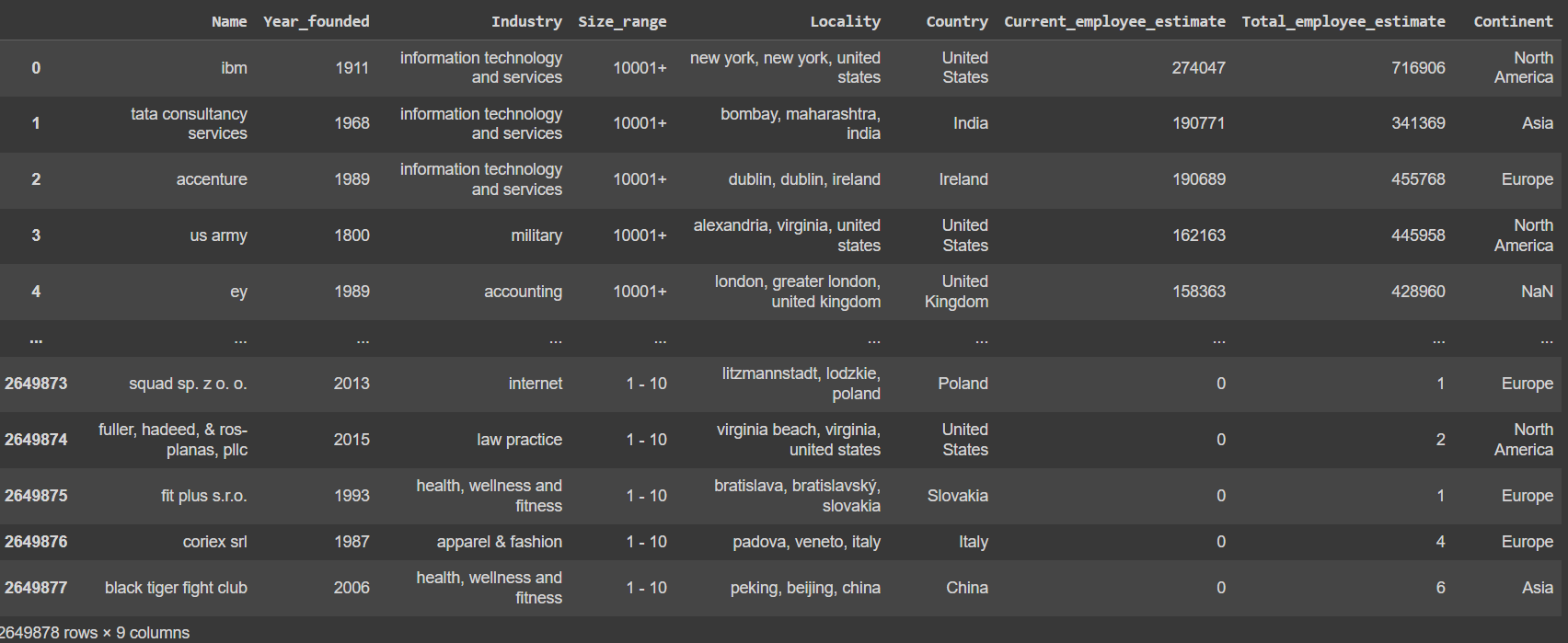


*Hình 4.2.1. Kiểm tra số lần trùng lặp của cột Country*

* 1. **Kết hợp và xử lý 2 dataframe**
* Kết hợp 2 dataframe lại với nhau, “how = ‘left’ là dồn hết tất cả cột của dataframe Continent vào dataframe Companies (*Hình 4.3.1*)

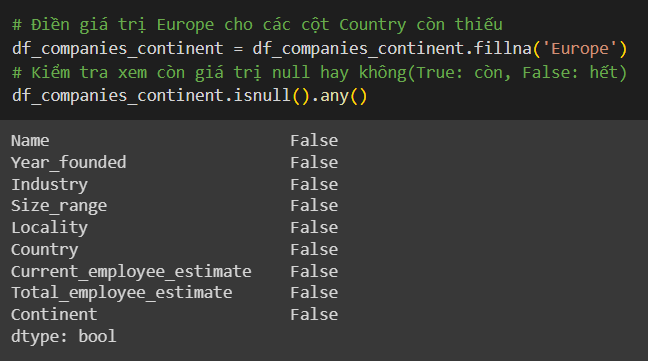


*Hình 4.3.1. Kết hợp dataframe Companies và Continent*



*Hình 4.3.2. Kết quả sau khi kết hợp hai dataframe*

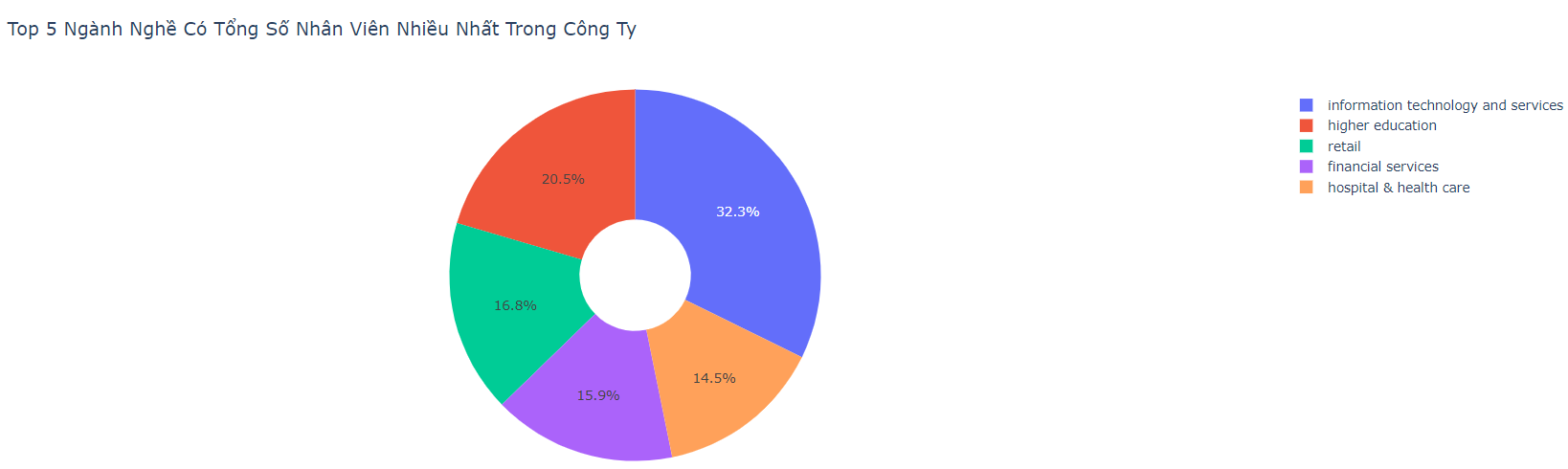
* Điền giá trị ‘Europe’ vào các cột Country còn thiếu, đồng thời kiểm tra xem sau khi kết hợp hai dataframe thì còn giá trị null nào nữa không (*Hình 4.3.3*)



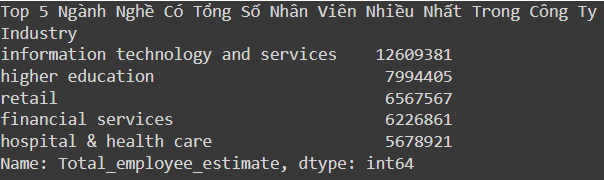
*Hình 4.3.3. Kiểm tra giá trị null của cột Country sau khi kết hợp*

1. **Trực quan hóa dữ liệu**

* **Top 5 ngành nghề có tổng số nhân viên nhiều nhất trong công ty**

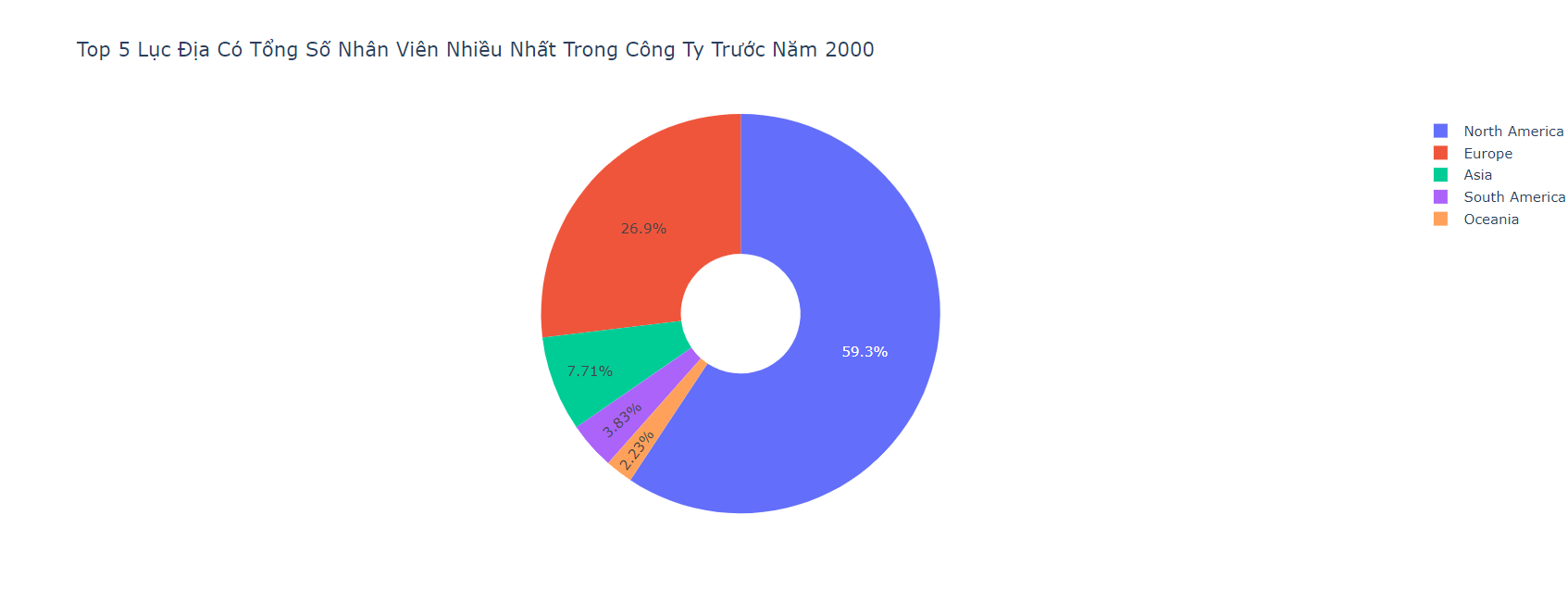
****

*Hình 5.1. Biểu đồ tròn thể hiện top 5 ngành nghề có tổng số nhân viên nhiều nhất trong công ty*

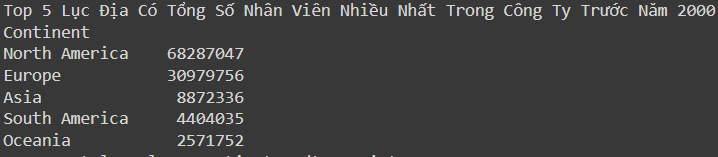


*Hình 5.2. Mô tả bằng chữ top 5 ngành nghề có tổng số nhân viên nhiều nhất trong công ty*

* Nhận xét: “Dịch vụ và công nghệ thông tin” vẫn đang rất hot và là ngành nghề được yêu thích với giới trẻ hiện nay. Chiếm đâu đó gần 1/4 các công ty ở châu Âu.
* **Top 5 lục địa có tổng số nhân viên nhiều nhất trong công ty trước và sau năm 2000**

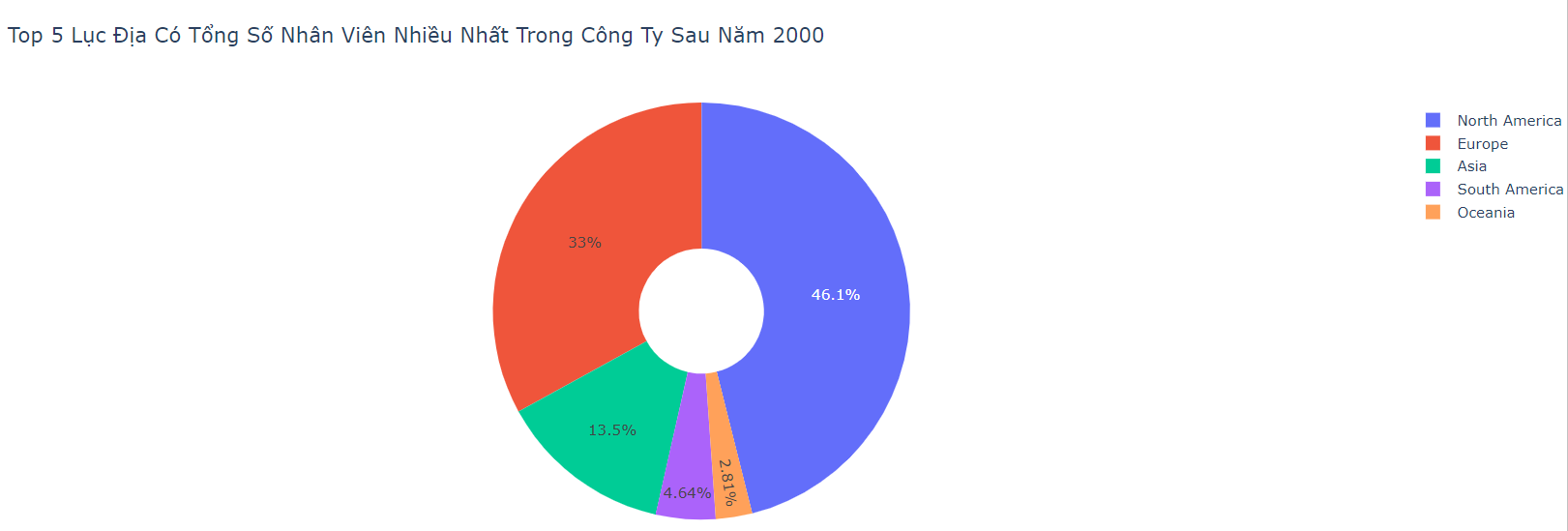
****

*Hình 5.3. Biểu đồ tròn thể hiện top 5 lục địa có tổng số nahn6 viên nhiều nhất trong công ty trước năm 2000*

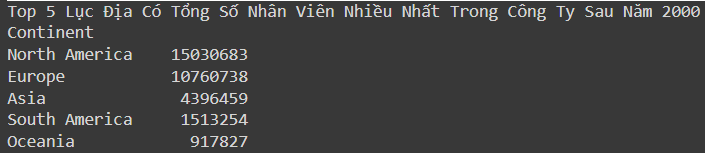
****

*Hình 5.4. Mô tả bằng chữ top 5 lục địa có tổng số nhân viên nhiều nhất trong công ty trước năm 2000*

* Nhận xét: Trước năm 2000, Bắc Mỹ đứng đầu trong 5 lục địa có tổng số nhân viên nhiều nhất trong công ty. Bên cạnh Bắc Mỹ là châu Âu đứng thứ 2

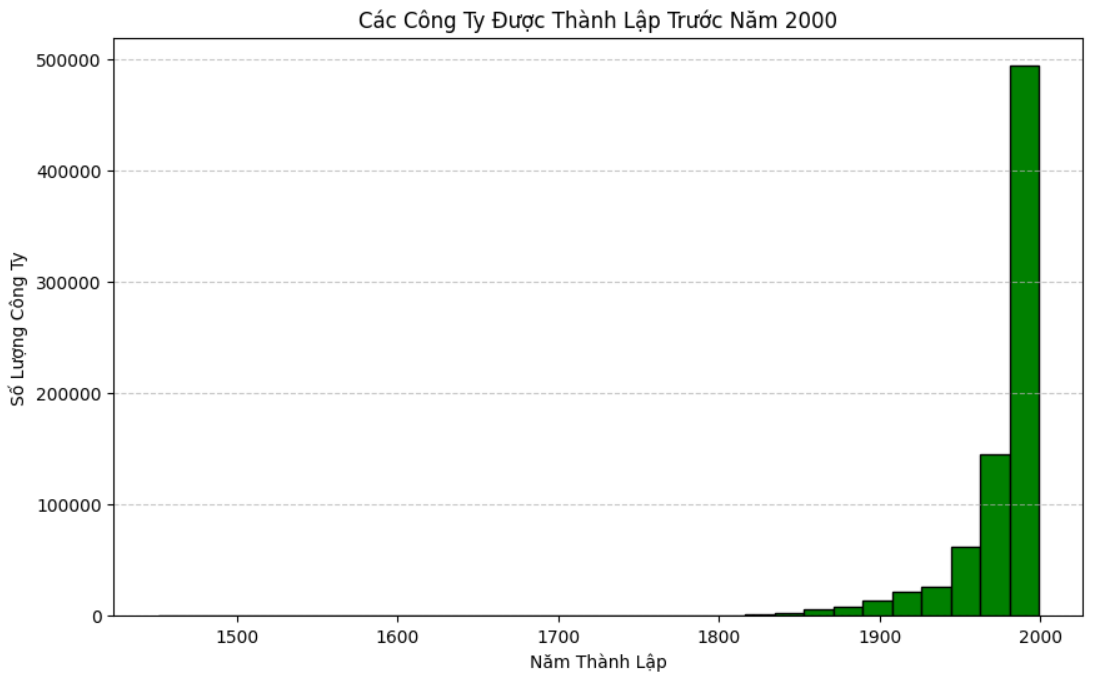


*Hình 5.5. Biểu đồ tròn thể hiện top 5 lục địa có tổng số nhân viên nhiều nhất trong công ty sau năm 2000*

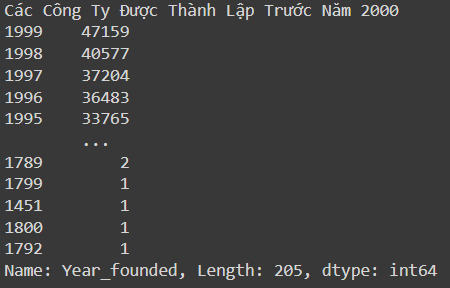


*Hình 5.6. Mô tả bằng chữ top 5 lục địa có tổng số nhân viên nhiều nhất trong công ty sau năm 2000*

* Sau năm 2000, không có nhiều biến đổi. Bắc Mỹ vẫn đứng đầu trong lục địa có tổng số nhân viên nhiều nhất.
* **Các công ty được thành lập trước và sau năm 2000**

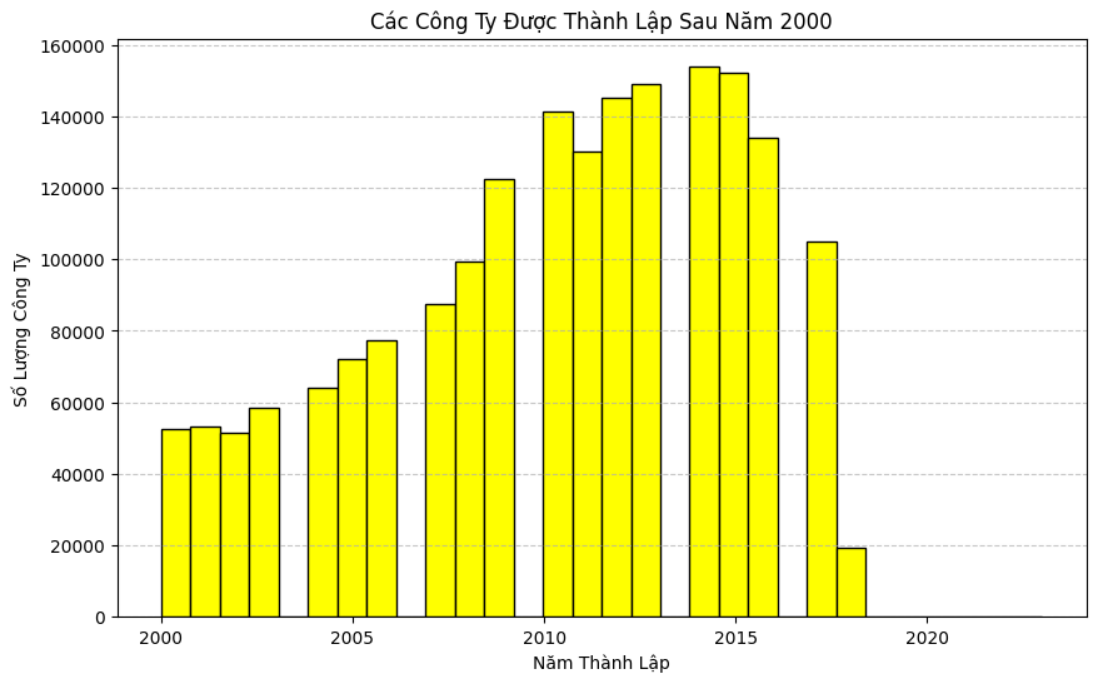


*Hình 5.7. Biểu đồ cột thể hiện các công ty được thành lập trước năm 2000*

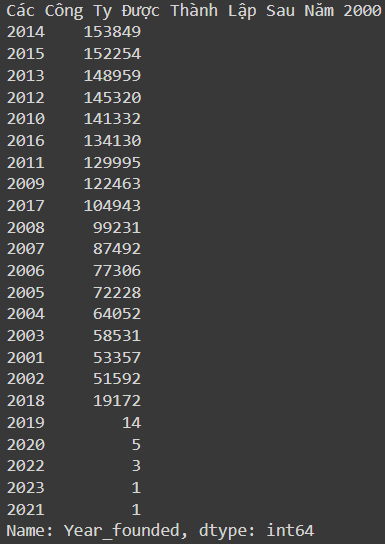


*Hình 5.8. Mô tả bằng chữ các công ty được thành lập trước năm 2000*

* Nhận xét: Các công ty dần dần thành lập từ rất lâu và đỉnh điểm năm 1999 là năm có nhiều công ty thành lập nhất trước năm 2000.

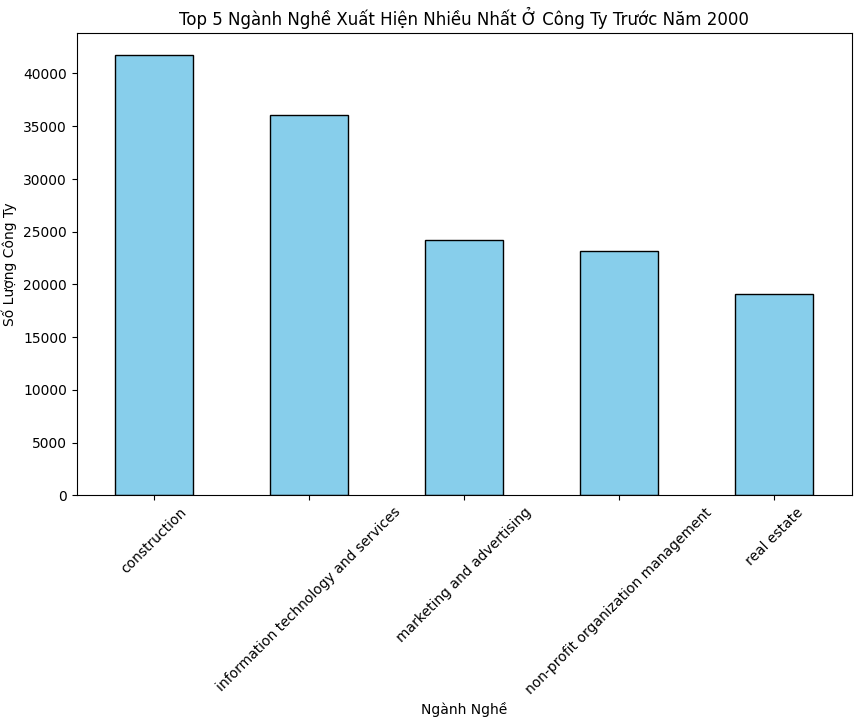


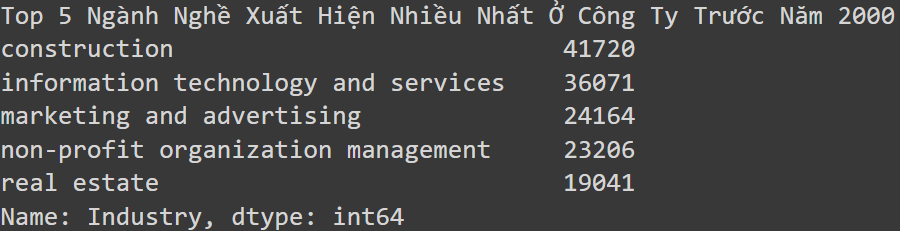
*Hình 5.9. Biểu đồ cột thể hiện các công ty được thành lập sau năm 2000*



*Hình 5.10. Mô tả bằng chữ các công ty được thành lập sau năm 2000*

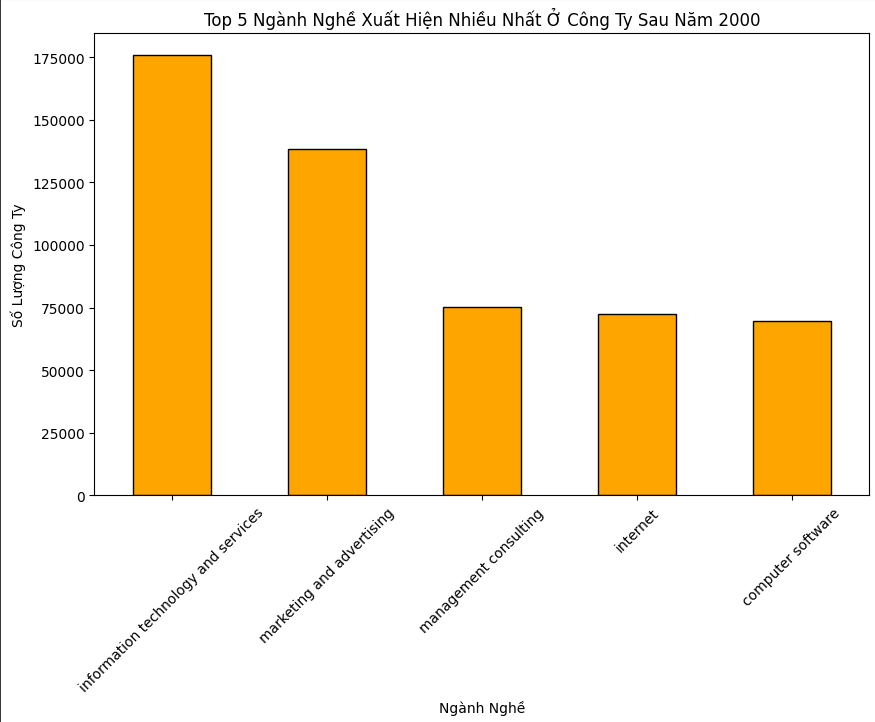
* Nhận xét: Sau năm 2000, năm có nhiều công ty được thành lập nhất là năm 2014. Vì sao không phải là năm 2023?. Vì ở thời điểm năm 2019, xảy ra đại dịch Covid-19 đã làm ảnh hưởng rất lớn đến các doanh nghiệp không chỉ ở châu Âu mà còn ở Việt Nam. Nhiều doanh nghiệp lớn đứng trước nguy cơ bị phá sản. Cho nên từ năm 2019-2023 rất ít công ty được thành lập (*Hình 5.5* và *Hình 5.6*)
* **Top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước và sau năm 2000**



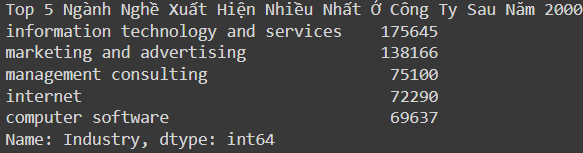
*Hình 5.11. Biểu đồ cột thể hiện top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước năm 2000 *

*Hình 5.12. Mô tả bằng chữ top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước năm 2000*

* Nhận xét: Trước năm 2000, ngành “dịch vụ và công nghệ thông tin” chưa được phát triển, hay vào đó là ngành “Xây dựng” đang chiếm tỉ lệ cao. Đó là ngành đã có từ rất lâu và cho đến hiện nay vẫn còn.

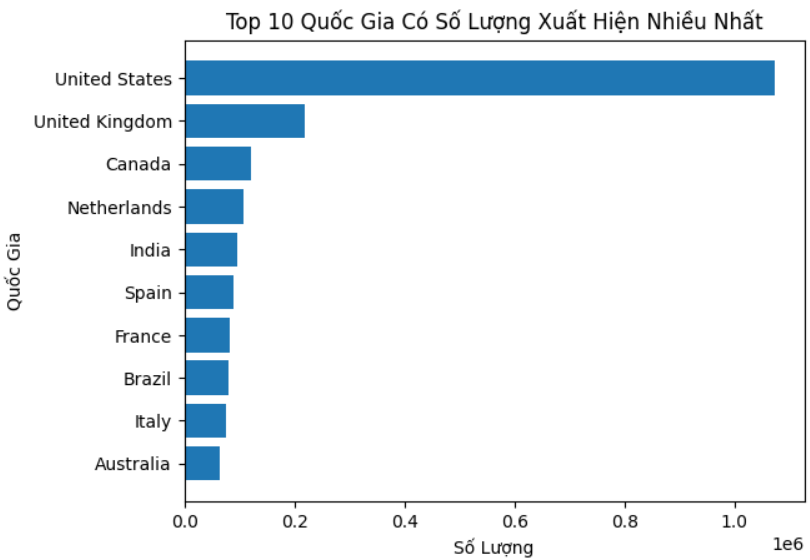


*Hình 5.13. Biểu đồ cột thể hiện top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước năm 2000*



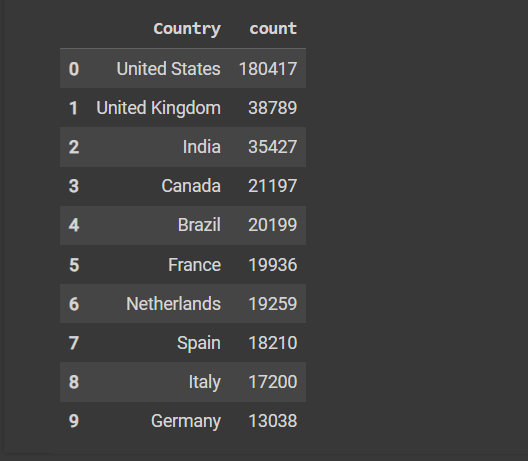
*Hình 5.14. Mô tả bằng chữ top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước năm 2000*

* Nhận xét: Sau năm 2000, ngành “Dịch vụ và công nghệ thông tin” phát triển mạnh mẽ. Con người biết đến và yêu thích ngành nghề này, đặc biệt đối với giới trẻ hiện nay. Việc tiếp xúc và sở hữu các công nghệ không còn xa lạ. Trong tương lai, ngành nghề này vẫn sẽ còn phát triển vượt bật.
* **Top 10 quốc gia xuất hiện nhiều nhất**

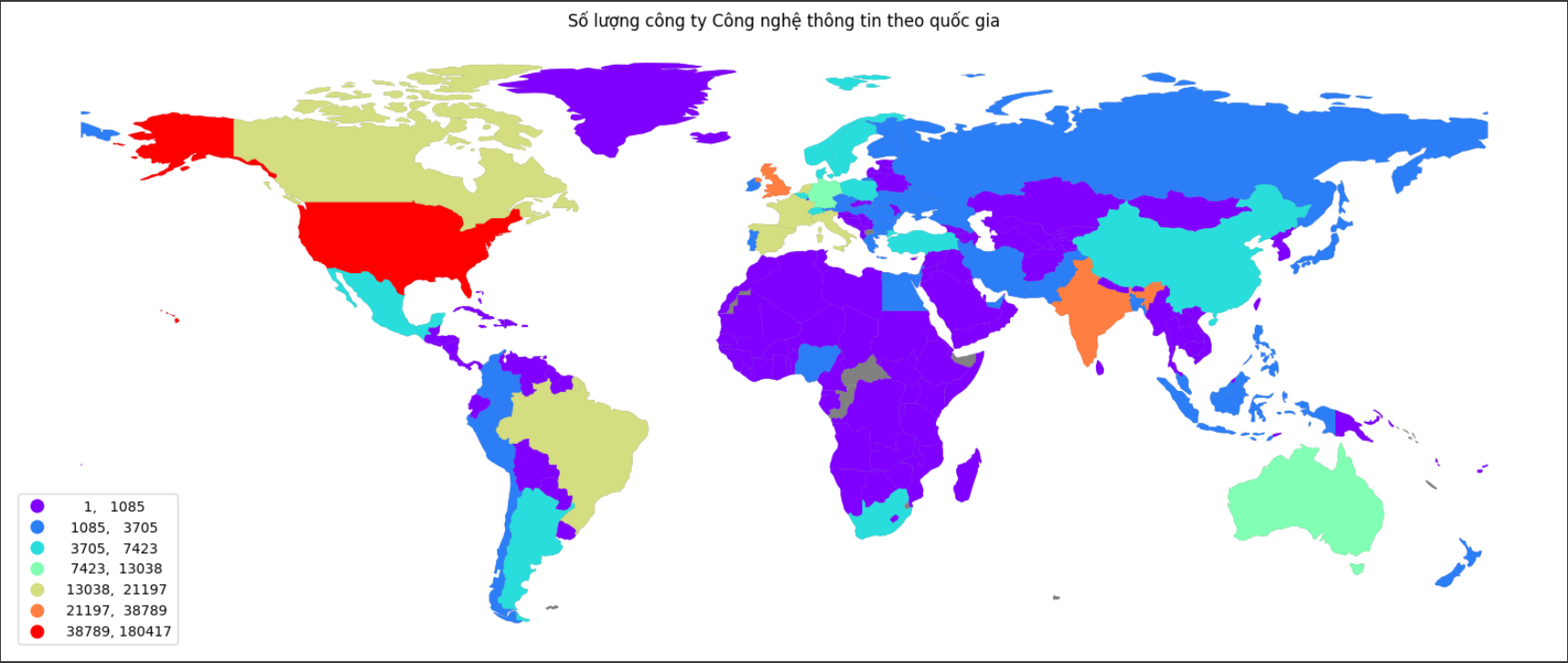
****

*Hình 5.15. Biểu đồ cột thể hiện top 10 quốc gia có số lượng xuất hiện nhiều nhất*

* Nhận xét: Dựa vào biểu đồ ta thấy được, Hoa Kỳ đứng đầu về quốc gia có số lượng xuất hiện nhiều nhất
* **Số lượng công ty công nghệ thông tin theo quốc gia**



*Hình 5.16. Mô tả 10 quốc gia có số lượng công ty Công nghệ thông tin nhiều nhất*

**

*Hình 5.17. Bản đồ thể hiện số lượng Công ty công nghệ thông tin theo quốc gia*

* Nhận xét: Nước Mỹ (Hoa Kỳ) nghiễm nhiên chiếm top đầu về số lượng công ty Công nghệ thông tin, đây là một quốc gia lớn mạnh về mọi mặt và về Công nghệ thông tin thì vẫn chiếm ưu thế cao hơn so với các nước gia. Một quốc gia đứng số 1 thế giới cho thấy được sự phát triển công nghệ vượt bậc của nước Mỹ là không nớc nào có thể sánh bằng.

1. **Gom cụm**

* **Sử dụng biểu đồ Elbow để tìm số cụm tối ưu có thể sử dụng**

A graph with a blue line

Description automatically generated

*Hình 6.1. Biểu đồ phương pháp Elbow cho dữ liệu ước tính tổng số nhân viên*

* Nhận xét:

+ Tổng phương sai: Tổng phương sai của tập dữ liệu giảm nhanh chóng khi số lượng cụm tăng từ 1 đến 3. Tuy nhiên, tổng phương sai giảm chậm hơn khi số lượng cụm tăng từ 3 đến 5. Điều này cho thấy rằng các cụm 3 và 4 có ý nghĩa hơn các cụm 1, 2 và 5

+ Đường cong: Đường cong phương pháp Elbow có dạng chữ "L". Điểm uốn của đường cong nằm ở giữa các cụm 3 và 4. Điều này cho thấy rằng số lượng cụm tối ưu là 4 hoặc 5.

+ Số lượng cụm: Số lượng cụm tối ưu có thể khác nhau tùy thuộc vào tập dữ liệu cụ thể và mục tiêu phân tích. Tuy nhiên, dựa trên biểu đồ này, số lượng cụm tối ưu có thể là 4 hoặc 5.

* **Ước tính nhân viên hiện tại theo dữ liệu “Current\_employee\_estimate”**

A graph with different colored lines

Description automatically generated

*Hình 6.2. Biểu đồ phân cụm K-means của dữ liệu “Current\_employee\_estimate”*

* Nhận xét: Theo hình, ta có thể thấy cụm 2 chiếm nhiều nhất, chiếm khoảng 40% tổng số điểm dữ liệu. Cụm 1 chiếm khoảng 20%, cụm 3 chiếm khoảng 25% và cụm 4 chiếm khoảng 15%. Cụm 2 chiếm nhiều nhất điều đó cho rằng có nhiều công ty có quy mô trung bình chiếm lượng nhân viên cao nhất ở hiện tại.
* **Ước tính tổng số nhân viên hiện tại theo dữ liệu “Total\_employee\_estimate”**

A graph with different colored dots

Description automatically generated

*Hình 6.4. Biểu đồ phân cụm K-means của dữ liệu “Total\_employee\_estimate”*

* Nhận xét: Theo hình, ta thấy không có quá nhiều sự khác biệt đối với dữ liệu “Current\_employee\_estimate” là mấy, nhưng ở đây là phân cụm về tổng số nhân viên cho nên trên hình thể hiện khá rõ ở cụm 2, đây cũng là cụm thể hiện cho công ty cho quy mô trung bình chiếm tỷ lệ nhìu nhất
* Biểu đồ phân cụm 2A diagram of a graph

  Description automatically generated with medium confidence

*Hình 6.4. Biểu đồ thể hiện các cụm thông qua việc in ra các tâm*

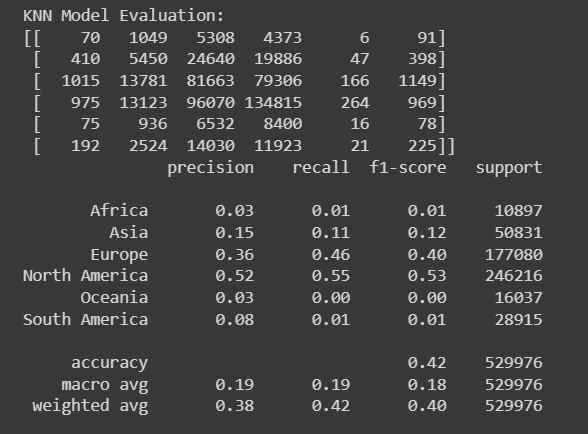
A red and green dotted diagram

Description automatically generated

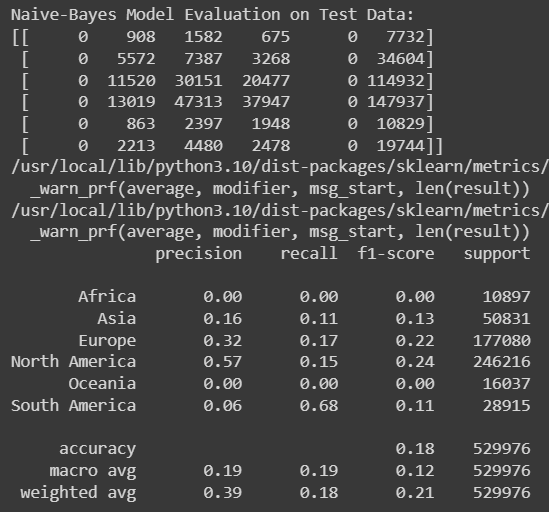
*Hình 6.5. Biểu đồ thể hiện hai thuộc tính “Current\_employee\_estimate” và “Total\_employee\_estimate”*

* Nhận xét: *Hình 6.4* cho thấy sử dụng biểu đồ elbow để có thể xác định tối ưu các cụm và sử dụng thuật toán k-means một cách hiệu quả. Hai điều này có thể thấy rõ ở *Hình 6.5* đây là biểu đồ thể hiện hai thuộc tính “Current” và “Total” cho thấy tập trung vào phân tích số lượng nhân viên trong các cụm, biểu đồ cho thấy có mối tương quan dương giữa hai biến. Điều này có nghĩa là khi giá trị của một biến tăng, giá trị của biến kia cũng có xu hướng tăng

1. **K-Nearest Neighbors**



*Hình 7.1. Kết quả của mô hình KNN Model*

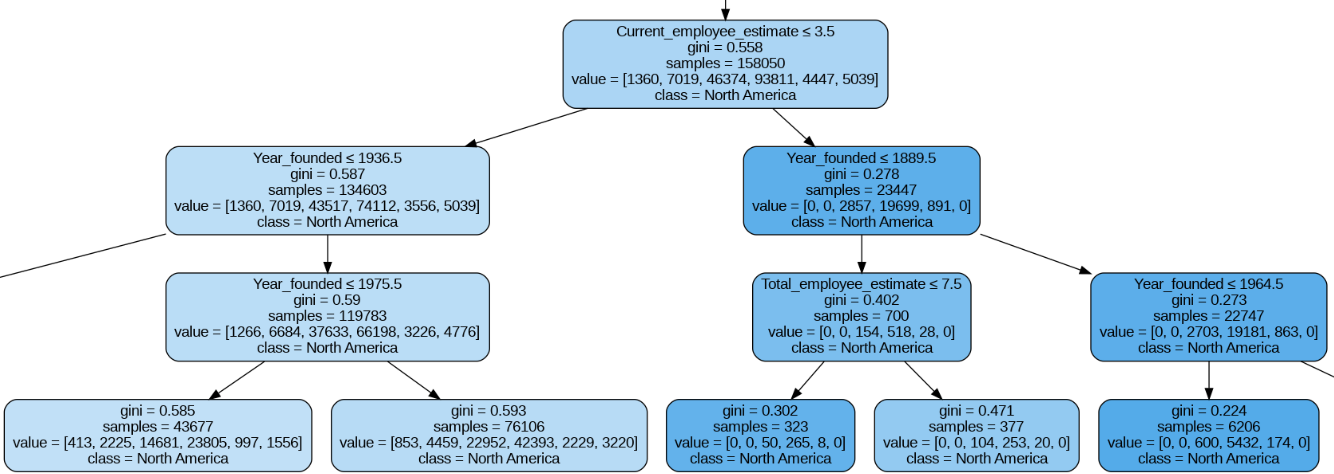


*Hình 7.2. Kết quả của mô hình Naive-Bayes Model*

* Nhận xét: Thời gian dự đoán của mô h ình Naïve-Bayes nhanh hơn mô hình K-Nearest Neighbors. Mô hình sử dụng để phân loại các công ty đến từ 5 châu lục, và có thể thấy độ chính xác lên đến 42% điều này có thể chứng minh được mô hình KNN đã có thể phân loại chính xác 42% các công ty có trong tập dữ liệu kiểm tra.

1. **Cây quyết định**

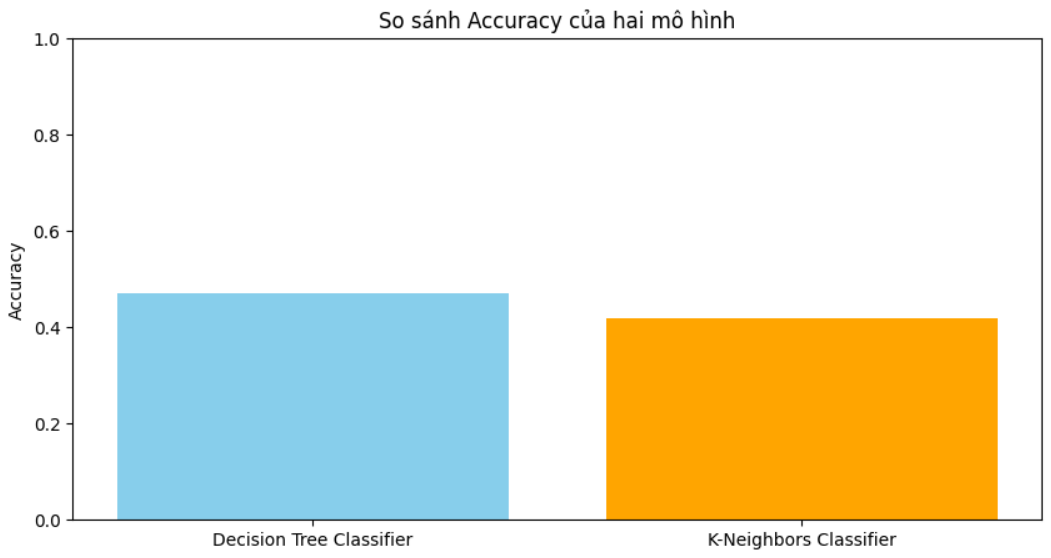
* **Biểu đồ hiển thị của cây quyết định và bộ luật của cây quyết định**

****

*Hình 8.1. Biểu đồ hiển thị của cây quyết định*

* Nhận xét: Qua biểu đồ hiển thị và bộ luật của cây quyết định có thể cho thấy được độ chính xác của mô hình trên là 47%, và bộ luật của cây quyết định cho thấy rằng các yếu tố quan trọng để xác định châu lục của một công ty là năm thành lập, quy mô và ngành nghề của công ty.
* <https://drive.google.com/file/d/1ZYTLQHxZwE41SwcRxxAnsQS9FyuFiLNC/view?usp=drive_link> đây là link *hình 8.1*.

1. **So sánh 2 Model**

****

*Hình 9.1. So sánh độ chính xác của hai mô hình*

* Nhận xét: Với bộ dữ liệu này, mô hình Decision Tree Classifier (Cây quyết định) hoạt động tốt hơn mô hình K-Neighors Classifier (KNN). Tuy nhiên thời gian dự đoán của mô hình Decision Tree Classifier lại nhanh hơn mô hình K-Neighors Classifier.