BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

◆◆◆

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

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, cụ thể tại đây.

2. Các thư viện

```
# Cài đặt thư viện
# geotext: trích xuất thông tin địa lý từ văn bản
# mapclassify: phân loại và phân đoạn dữ liệu địa lý
!pip install geotext mapclassify
```

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

```
from google.colab import drive
drive.mount('/content/drive')
import pandas as pd
import numpy as np
import datetime
import graphviz
import geotext
import geopandas as gpd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.graph_objects as go
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, silhouette_score
from sklearn.datasets import load_breast_cancer
from sklearn.tree import DecisionTreeClassifier, export_text, export_graphviz
```

Hình 2.2. Hình ảnh các thư viện

3. Đọc Dữ liệu

3.1. Companies Dataset

```
df = pd.read_csv("/content/drive/MyDrive/KHAI_PHA_DU_LIEU/Bai_Tap_Lon/Data Mining - Doanh Nghiệp/companies_sorted.csv")
df.head(5)
```

Hình 3.1. Đọc dữ liệu từ file compaines_sorted.csv

	Unnamed: 0	name	domain	year founded	industry	size range	locality	country	linkedin url	current employee estimate	total employee estimate
0	5872184	ibm	ibm.com	1911.0	information technology and services	10001+	new york, new york, united states	united states	linkedin.com/company/ibm	274047	716906
1	4425416	tata consultancy services	tcs.com	1968.0	information technology and services	10001+	bombay, maharashtra, india	india	linkedin.com/company/tata-consultancy-services	190771	341369
2	21074	accenture	accenture.com	1989.0	information technology and services	10001+	dublin, dublin, ireland	ireland	linkedin.com/company/accenture	190689	455768
3	2309813	us army	goarmy.com	1800.0	military	10001+	alexandria, virginia, united states	united states	linkedin.com/company/us-army	162163	445958
4	1558607	ey	ey.com	1989.0	accounting	10001+	london, greater london, united kingdom	united kingdom	linkedin.com/company/ernstandyoung	158363	428960

Hình 3.2. Kết quả sau khi chạy file compaines_sorted.csv

3.2. Continents Dataset

```
df_continent = pd.read_csv("/content/drive/MyDrive/KHAI_PHA_DU_LIEU/Bai_Tap_Lon/Data Mining - Doanh Nghiệp/Continent.csv")
df_continent.head()
```

Hình 3.3. Đọc dữ liệu từ file continent.csv

	Continent	Country
0	Asia	Afghanistan
1	Europe	Albania
2	Africa	Algeria
3	Oceania	American Samoa
4	Europe	Andorra

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

4. Xử lý dữ liệu

4.1. 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*)

```
# Xóa các cột không cần thiết
df_companies.drop(['Unnamed: 0','domain','linkedin url'], axis=1, inplace=True)
# Chuẩn hóa dữ liệu của DataFrame df
df_companies = df_companies.rename(columns={'name': 'Name', 'year founded': 'year_founded', 'industry': 'industry',
                                            'size range': 'size_range', 'locality': 'locality', 'country': 'country',
                                            'current employee estimate': 'current_employee_estimate', 'total employee estimate': 'total_employee_estimate'})
# Chuyển đổi chữ cái đầu tiên của mỗi từ trong chuỗi thành chữ cái in hoa trong danh sách cột
df_companies.columns = df_companies.columns.str.capitalize()
# Chuyển đổi dữ liệu trong cột Country - viết hóa chữ cái đầu, các chữ còn lại viết thường
df_companies['Country'] = df_companies.Country.str.title()
```

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*)

```
#Kiểm tra giá trị null
df_companies.isnull().sum()
```

```
Name                          3
Year_founded            3606980
Industry                 290003
Size_range                    0
Locality                2508825
Country                 2349207
Current_employee_estimate     0
Total_employee_estimate       0
dtype: int64
```

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*)

```
# Điền giá trị 0
df_companies = df_companies.fillna(0)
# Chuyển đổi kiểu dữ liệu sang int
df_companies['Year_founded'] = df_companies['Year_founded'].astype(int)
# Xóa các năm sau năm hiện tại (2024)
today = datetime.date.today()
df_companies = df_companies[df_companies['Year_founded'] <= today.year]
```

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*)

```
# Xóa các dòng có giá trị trong cột Year Founded, Industry, Locality, Country bằng 0
df_companies = df_companies[df_companies['Year_founded'] != 0]
df_companies = df_companies[df_companies['Industry'] != 0]
df_companies = df_companies[df_companies['Locality'] != 0]
df_companies = df_companies[df_companies['Country'] != 0]
# Kiểm tra dữ liệu trùng lặp
duplicate_rows = df_companies[df_companies.duplicated(subset="Name")]
print('Số dòng dữ liệu trùng lặp:', len(duplicate_rows))
```

```
Số dòng dữ liệu trùng lặp: 34439
```

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

```
# Xóa dữ liệu trùng lặp
df_companies = df_companies.drop_duplicates(subset="Name")
```

Hình 4.1.5. Xóa dữ liệu trùng lặp của cột Name

df_companies

	Name	Year_founded	Industry	Size_range	Locality	Country	Current_employee_estimate	Total_employee_estimate
0	ibm	1911	information technology and services	10001+	new york, new york, united states	United States	274047	716906
1	tata consultancy services	1968	information technology and services	10001+	bombay, maharashtra, india	India	190771	341369
2	accenture	1989	information technology and services	10001+	dublin, dublin, ireland	Ireland	190689	455768
3	us army	1800	military	10001+	alexandria, virginia, united states	United States	162163	445958
4	ey	1989	accounting	10001+	london, greater london, united kingdom	United Kingdom	158363	428960
...	...	...	...	...	...	...	...	...
7173404	squad sp. z o. o.	2013	internet	1 - 10	litzmannstadt, lodzkie, poland	Poland	0	1
7173411	fuller, hadeed, & ros-planas, pllc	2015	law practice	1 - 10	virginia beach, virginia, united states	United States	0	2
7173416	fit plus s.r.o.	1993	health, wellness and fitness	1 - 10	bratislava, bratislavský, slovakia	Slovakia	0	1
7173417	coriex srl	1987	apparel & fashion	1 - 10	padova, veneto, italy	Italy	0	4
7173422	black tiger fight club	2006	health, wellness and fitness	1 - 10	peking, beijing, china	China	0	6

2649878 rows × 8 columns

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

4.2. Continents Dataset

- Kiểm tra dữ liệu trùng lặp của cột Country trong dataframe Continent (*Hình 4.2.1*)

```
# Kiểm tra số lần trùng lặp của cột Country
duplicate_rows = df_continent[df_continent.duplicated(subset="Country")]
print('Số dòng dữ liệu trùng lặp:', len(duplicate_rows))

Số dòng dữ liệu trùng lặp: 0
```

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

4.3. 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*)

```
# Kết hợp 2 dataset (df_companies và df_continent) lại với nhau
df_companies_continent = pd.merge(df_companies, df_continent,on='Country',how='left')
df_companies_continent
```

Hình 4.3.1. Kết hợp dataframe Companies và Continent

	Name	Year_founded	Industry	Size_range	Locality	Country	Current_employee_estimate	Total_employee_estimate	Continent
0	ibm	1911	information technology and services	10001+	new york, new york, united states	United States	274047	716906	North America
1	tata consultancy services	1968	information technology and services	10001+	bombay, maharashtra, india	India	190771	341369	Asia
2	accenture	1989	information technology and services	10001+	dublin, dublin, ireland	Ireland	190689	455768	Europe
3	us army	1800	military	10001+	alexandria, virginia, united states	United States	162163	445958	North America
4	ey	1989	accounting	10001+	london, greater london, united kingdom	United Kingdom	158363	428960	NaN
...	...	...	...	...	...	...	...	...	...
2649873	squad sp. z o. o.	2013	internet	1 - 10	litzmannstadt, lodzkie, poland	Poland	0	1	Europe
2649874	fuller, hadeed, & ros-planas, plic	2015	law practice	1 - 10	virginia beach, virginia, united states	United States	0	2	North America
2649875	fit plus s.r.o.	1993	health, wellness and fitness	1 - 10	bratislava, bratislavský, slovakia	Slovakia	0	1	Europe
2649876	coriex srl	1987	apparel & fashion	1 - 10	padova, veneto, italy	Italy	0	4	Europe
2649877	black tiger fight club	2006	health, wellness and fitness	1 - 10	peking, beijing, china	China	0	6	Asia

2649878 rows × 9 columns

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*)

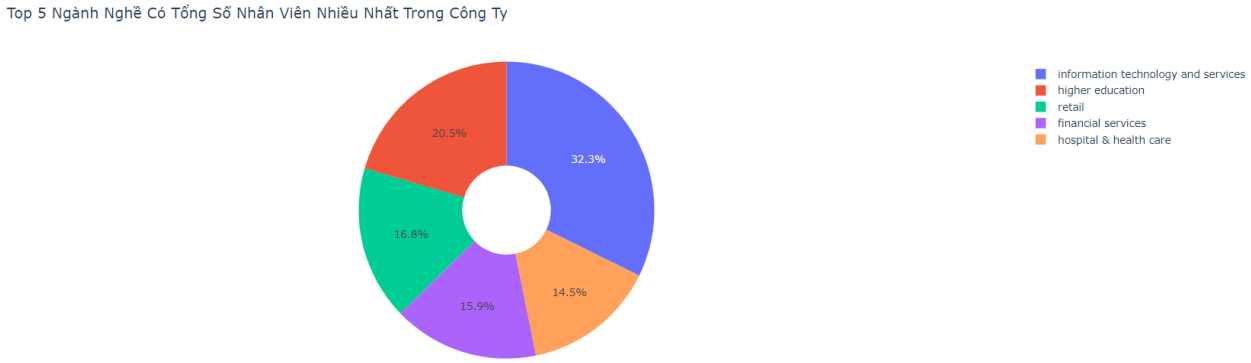
```
# Điền giá trị Europe cho các cột Country còn thiếu
df_companies_continent = df_companies_continent.fillna('Europe')
# Kiểm tra xem còn giá trị null hay không(True: còn, False: hết)
df_companies_continent.isnull().any()
```

```
Name                        False
Year_founded                False
Industry                    False
Size_range                  False
Locality                    False
Country                     False
Current_employee_estimate   False
Total_employee_estimate     False
Continent                   False
dtype: bool
```

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

5. Trực quan hóa dữ liệu

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



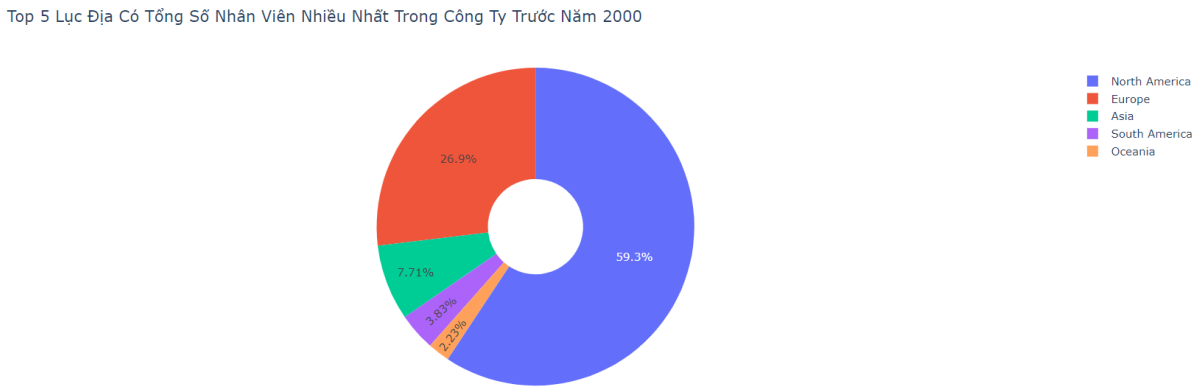
Hình 5.1. Biểu đồ tròn thể hiện top 5 ngành nghề có tổng số nhân viên nhiều nhất trong công ty

```
Top 5 Ngành Nghề Có Tổng Số Nhân Viên Nhiều Nhất Trong Công Ty
Industry
information technology and services    12609381
higher education                        7994405
retail                                  6567567
financial services                      6226861
hospital & health care                  5678921
Name: Total_employee_estimate, dtype: int64
```

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**

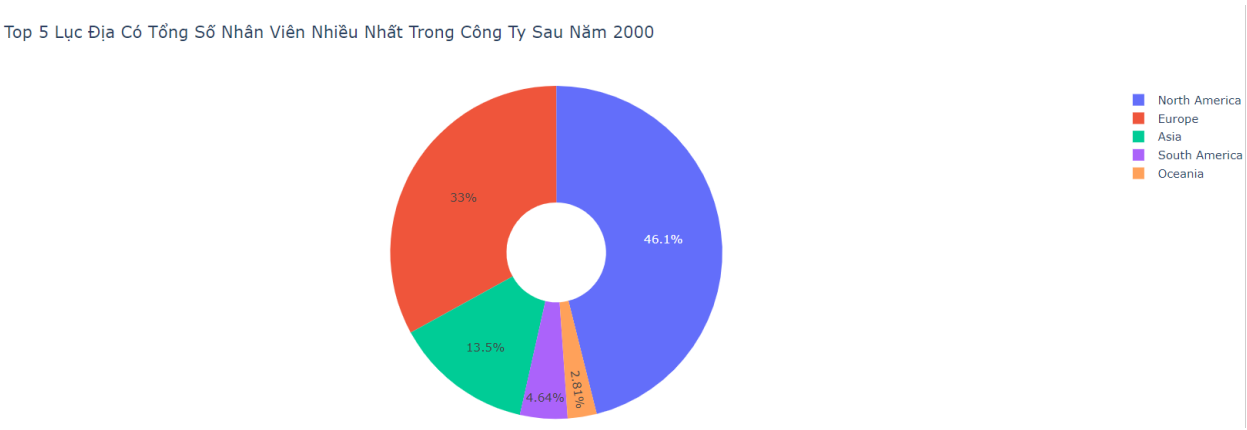


Hình 5.3. Biểu đồ tròn thể hiện top 5 lục địa có tổng số nahn6 viên nhiều nhất trong công ty trước năm 2000

```
Top 5 Lục Địa Có Tổng Số Nhân Viên Nhiều Nhất Trong Công Ty Trước Năm 2000
Continent
North America    68287047
Europe           30979756
Asia              8872336
South America     4404035
Oceania           2571752
```

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



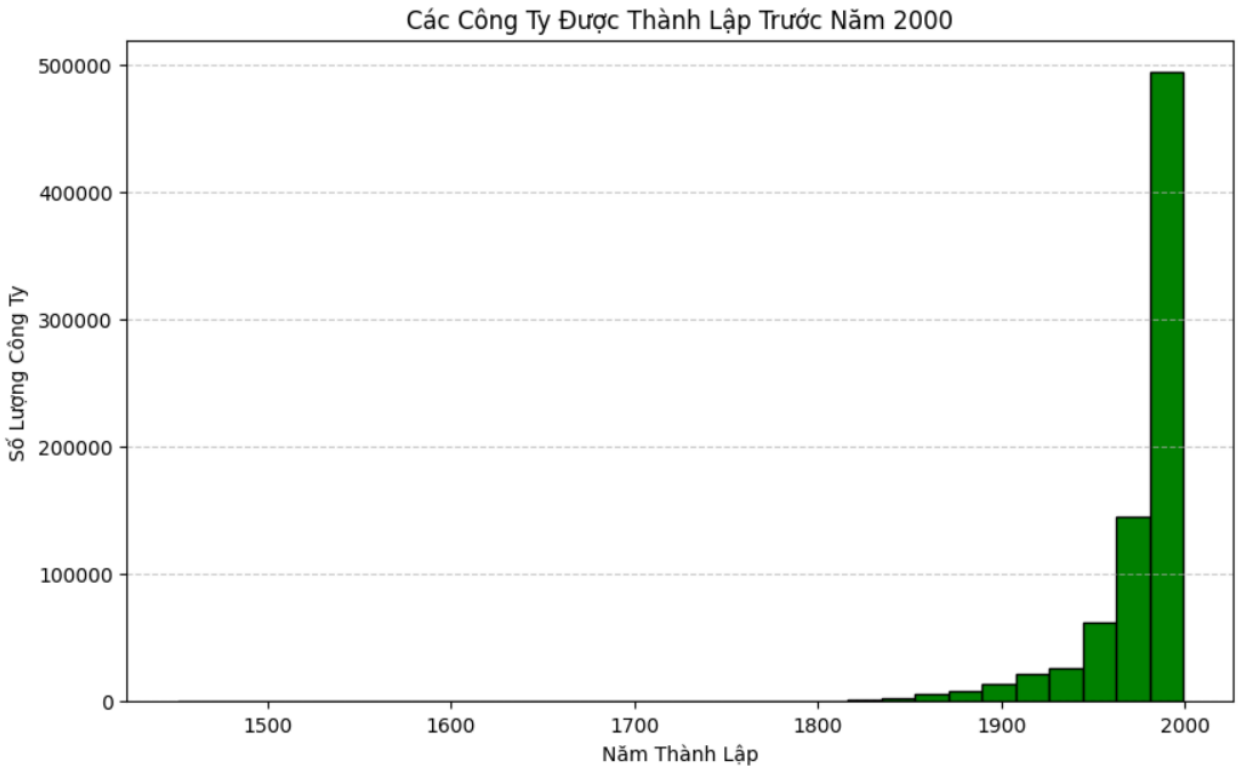
Hình 5.5. Biểu đồ tròn thể hiện top 5 lục địa có tổng số nhân viên nhiều nhất trong công ty sau năm 2000

```
Top 5 Lục Địa Có Tổng Số Nhân Viên Nhiều Nhất Trong Công Ty Sau Năm 2000
Continent
North America    15030683
Europe           10760738
Asia              4396459
South America     1513254
Oceania            917827
```

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**

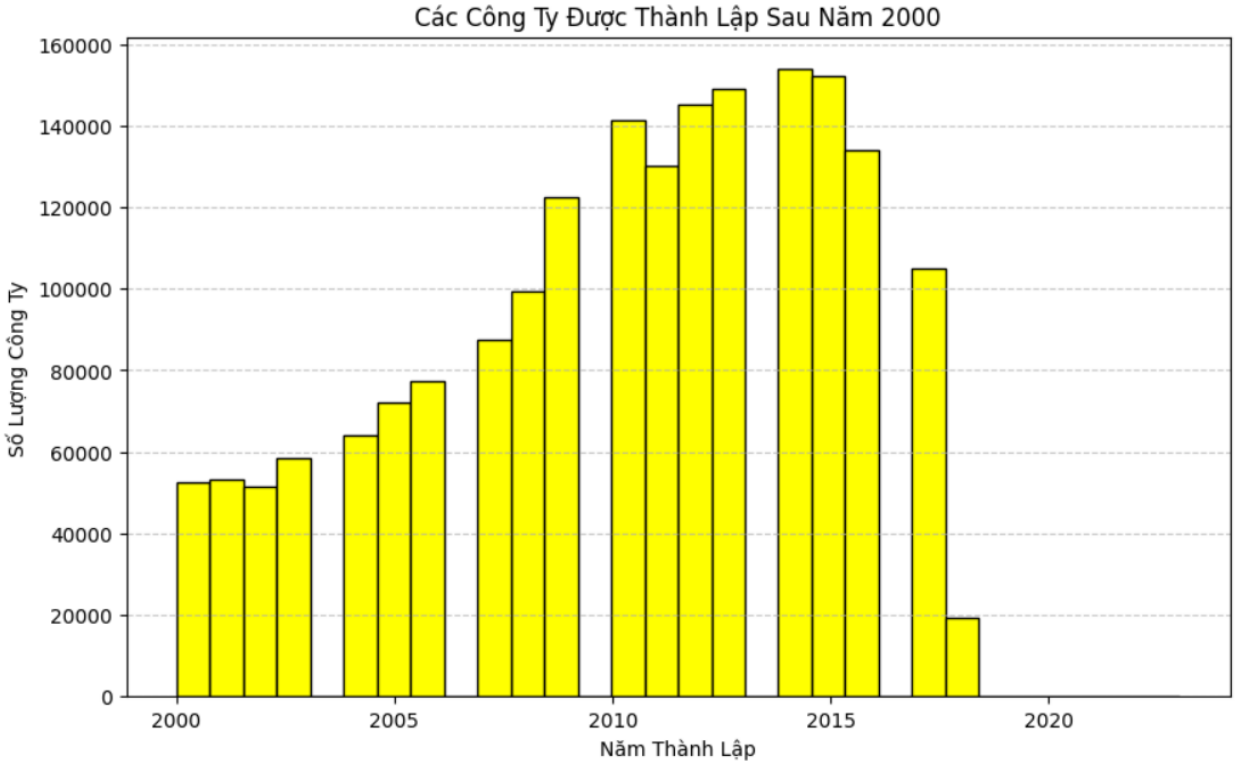


Hình 5.7. Biểu đồ cột thể hiện các công ty được thành lập trước năm 2000

```
Các Công Ty Được Thành Lập Trước Năm 2000
1999    47159
1998    40577
1997    37204
1996    36483
1995    33765
        ...
1789        2
1799        1
1451        1
1800        1
1792        1
Name: Year_founded, Length: 205, dtype: int64
```

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.



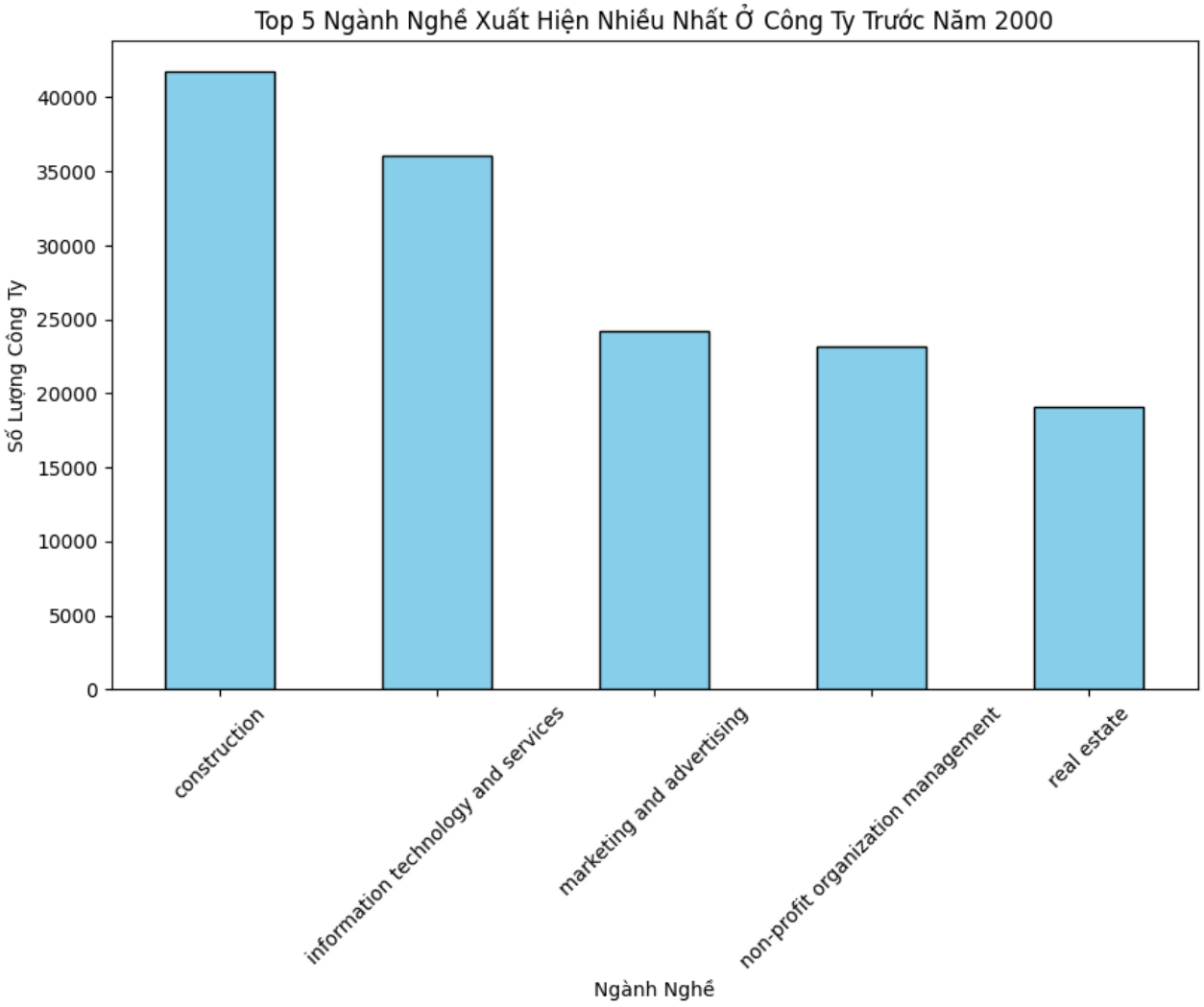
Hình 5.9. Biểu đồ cột thể hiện các công ty được thành lập sau năm 2000

```
Các Công Ty Được Thành Lập Sau Năm 2000
2014    153849
2015    152254
2013    148959
2012    145320
2010    141332
2016    134130
2011    129995
2009    122463
2017    104943
2008     99231
2007     87492
2006     77306
2005     72228
2004     64052
2003     58531
2001     53357
2002     51592
2018     19172
2019        14
2020         5
2022         3
2023         1
2021         1
Name: Year_founded, dtype: int64
```

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**

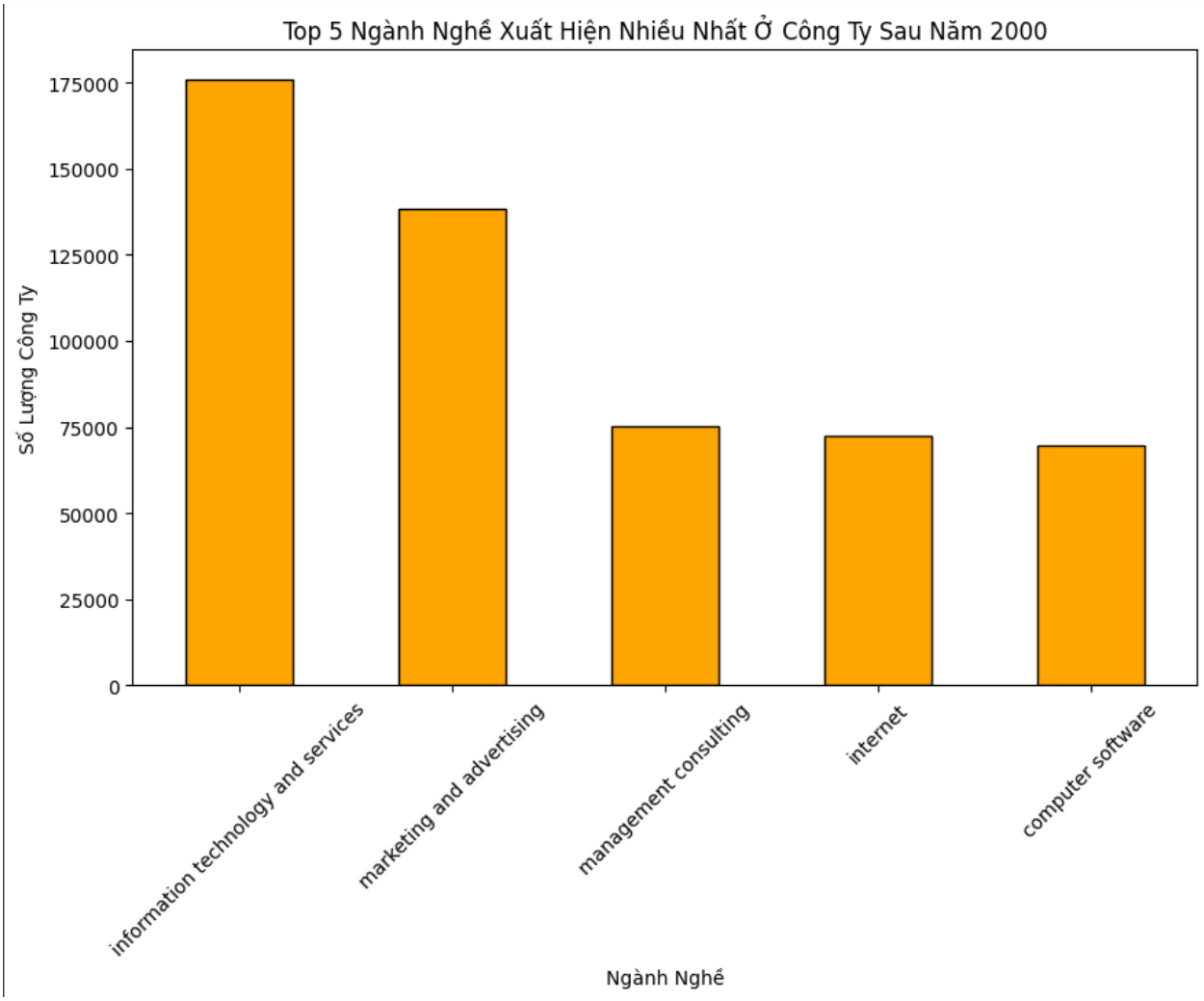


Hình 5.11. Biểu đồ cột thể hiện top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước năm 2000

```
Top 5 Ngành Nghề Xuất Hiện Nhiều Nhất Ở Công Ty Trước Năm 2000
construction                                41720
information technology and services         36071
marketing and advertising                   24164
non-profit organization management          23206
real estate                                 19041
Name: Industry, dtype: int64
```

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.



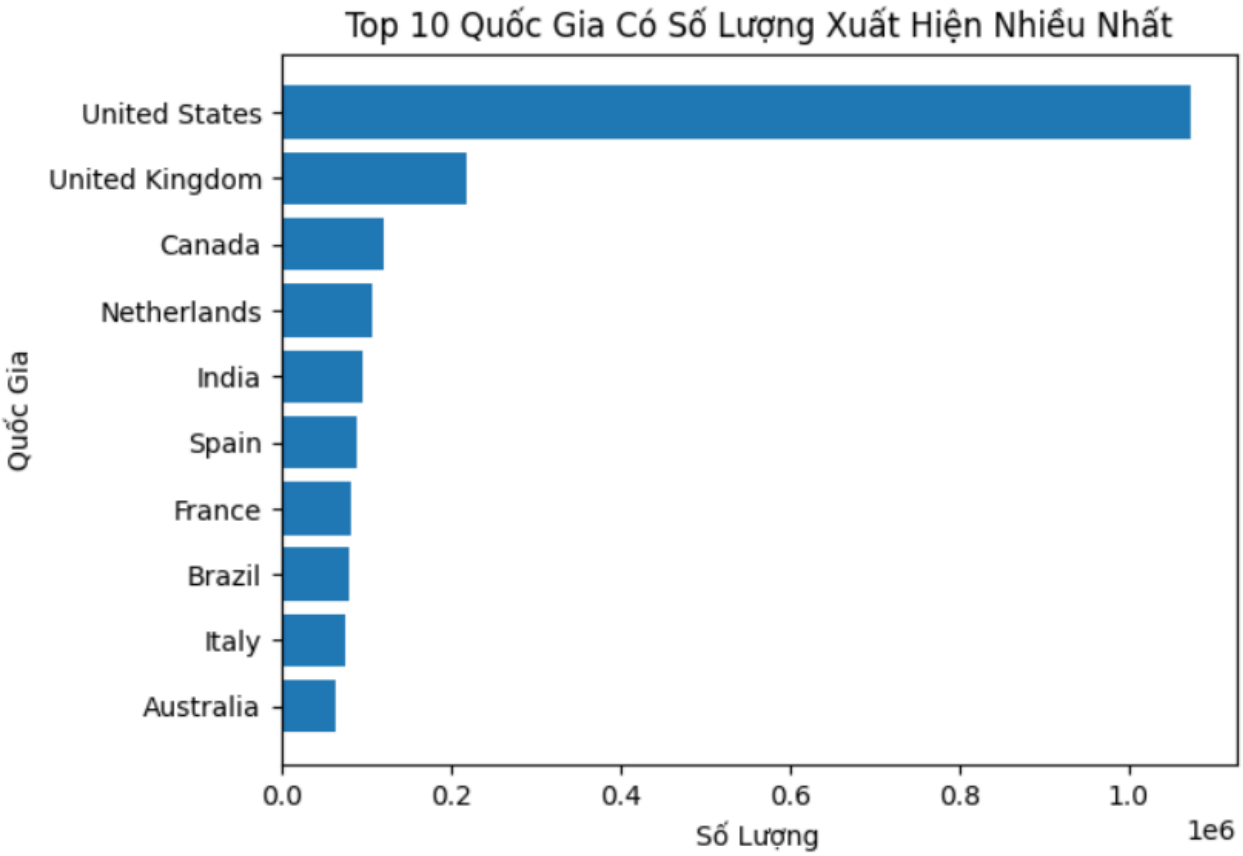
Hình 5.13. Biểu đồ cột thể hiện top 5 ngành nghề xuất hiện nhiều nhất ở công ty trước năm 2000

```
Top 5 Ngành Nghề Xuất Hiện Nhiều Nhất Ở Công Ty Sau Năm 2000
information technology and services    175645
marketing and advertising              138166
management consulting                   75100
internet                                72290
computer software                       69637
Name: Industry, dtype: int64
```

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**



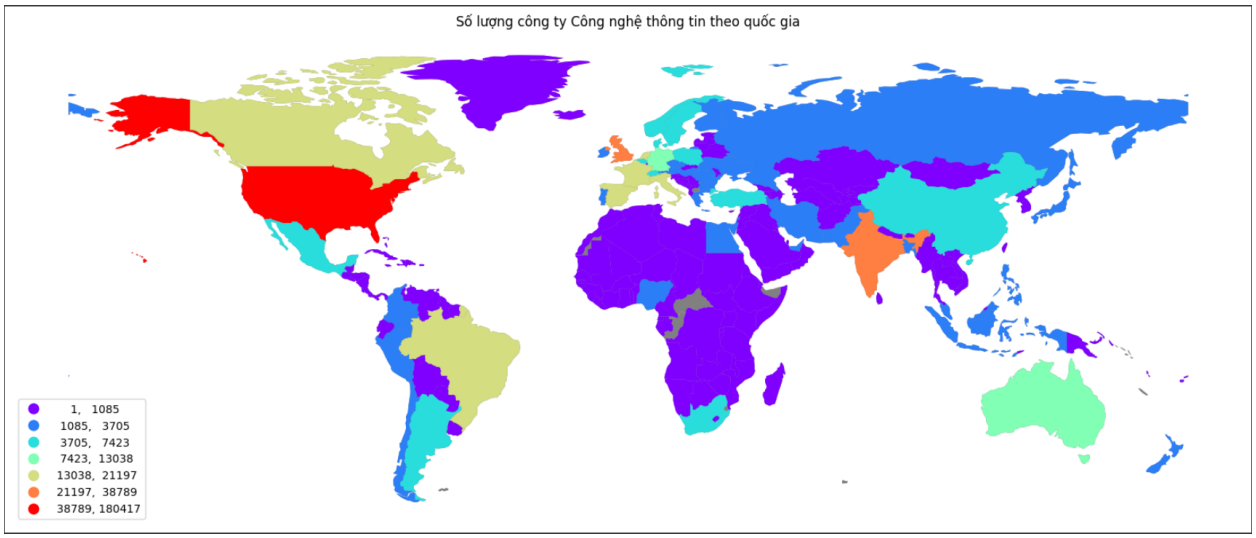
Hình 5.15. Biểu đồ cột thể hiện top 10 quốc gia có số lượng xuất hiện nhiều nhất

⇨ 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**

	Country	count
0	United States	180417
1	United Kingdom	38789
2	India	35427
3	Canada	21197
4	Brazil	20199
5	France	19936
6	Netherlands	19259
7	Spain	18210
8	Italy	17200
9	Germany	13038

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



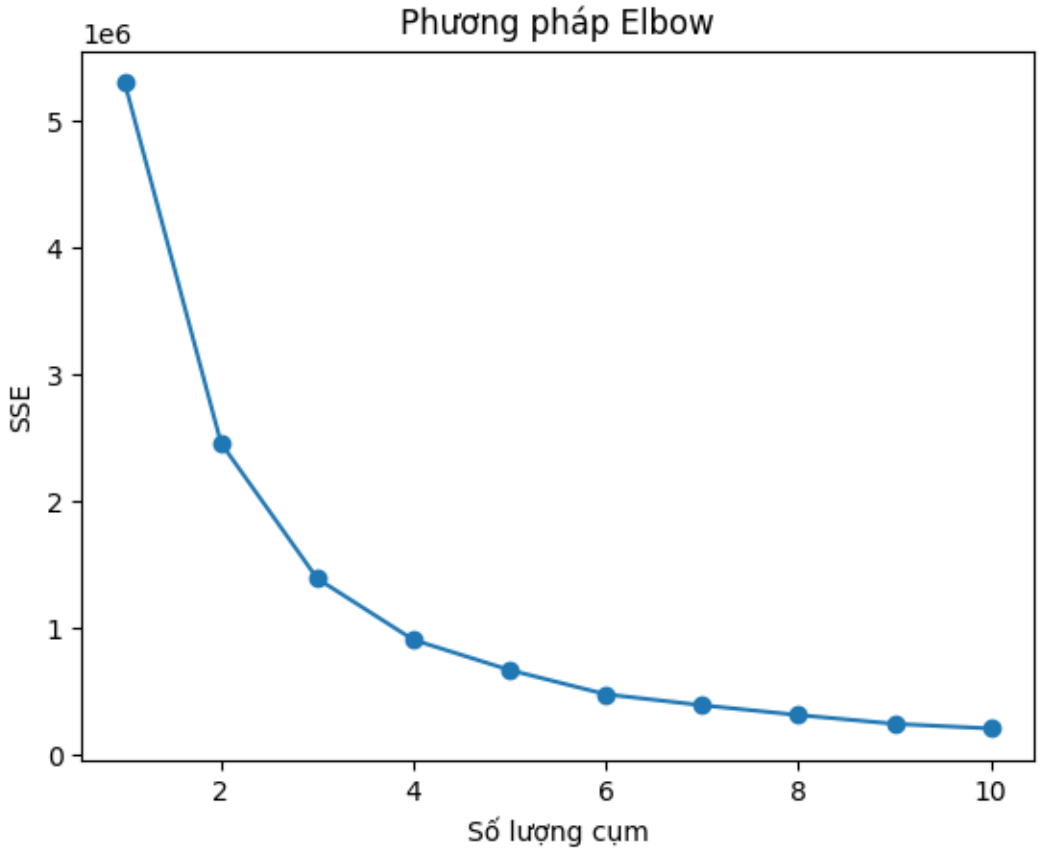
Hình 5.17. Bản đồ thể hiện số lượng Công ty công nghệ thông tin theo quốc gia

⇨ 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.

6. Gom cụm

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



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

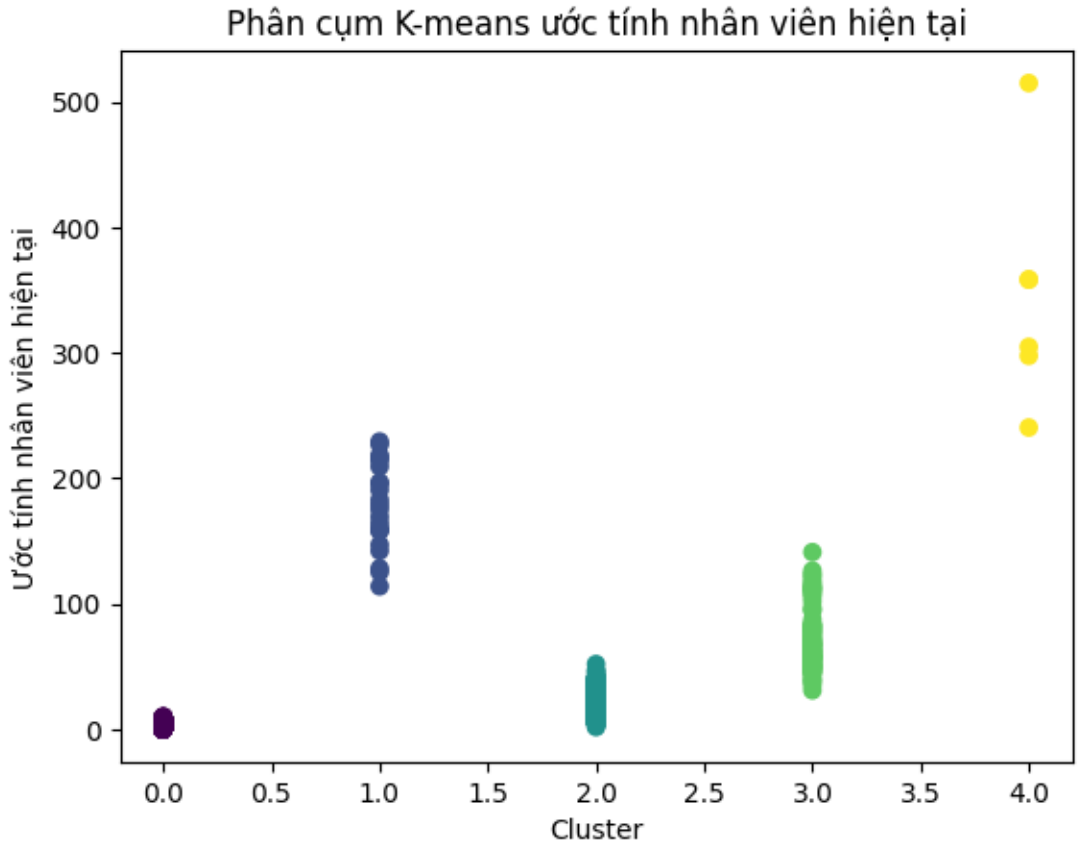
⇨ Nhận xét:

+ 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"**

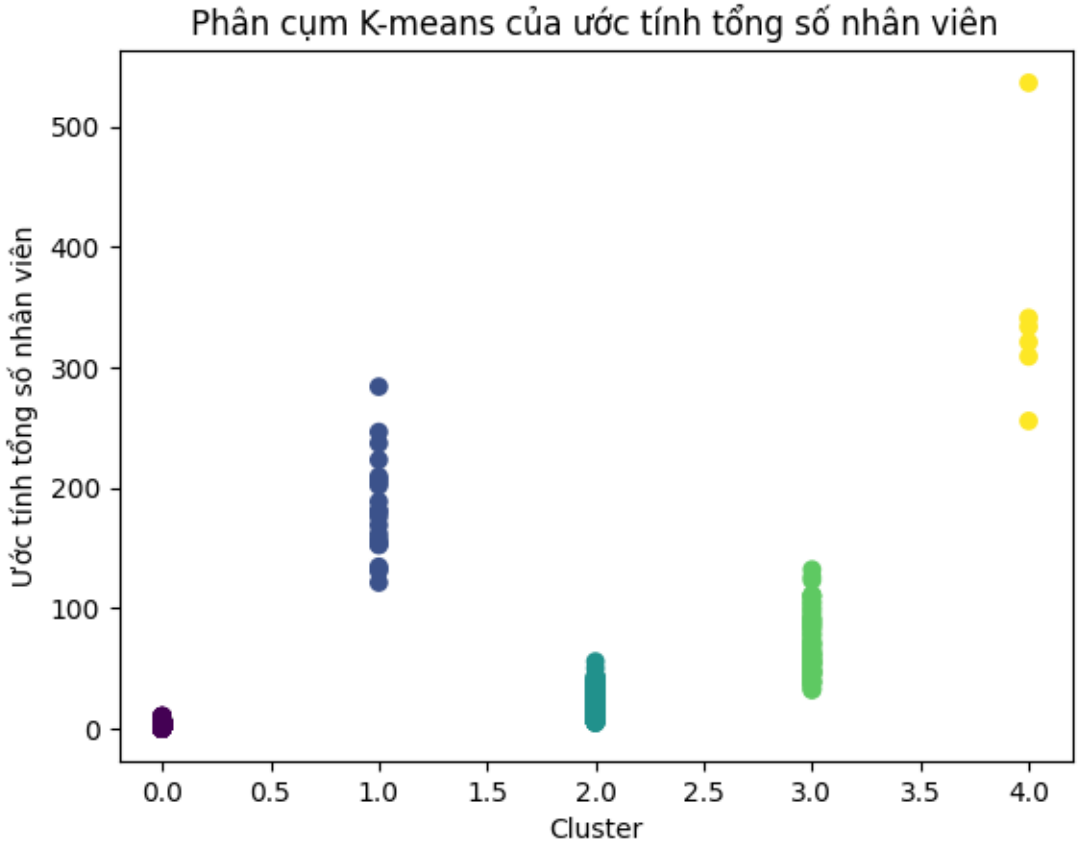


Hình 6.2. Biểu đồ phân cụm K-means của dữ liệu "Current_employee_estimate"

⇨ 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.

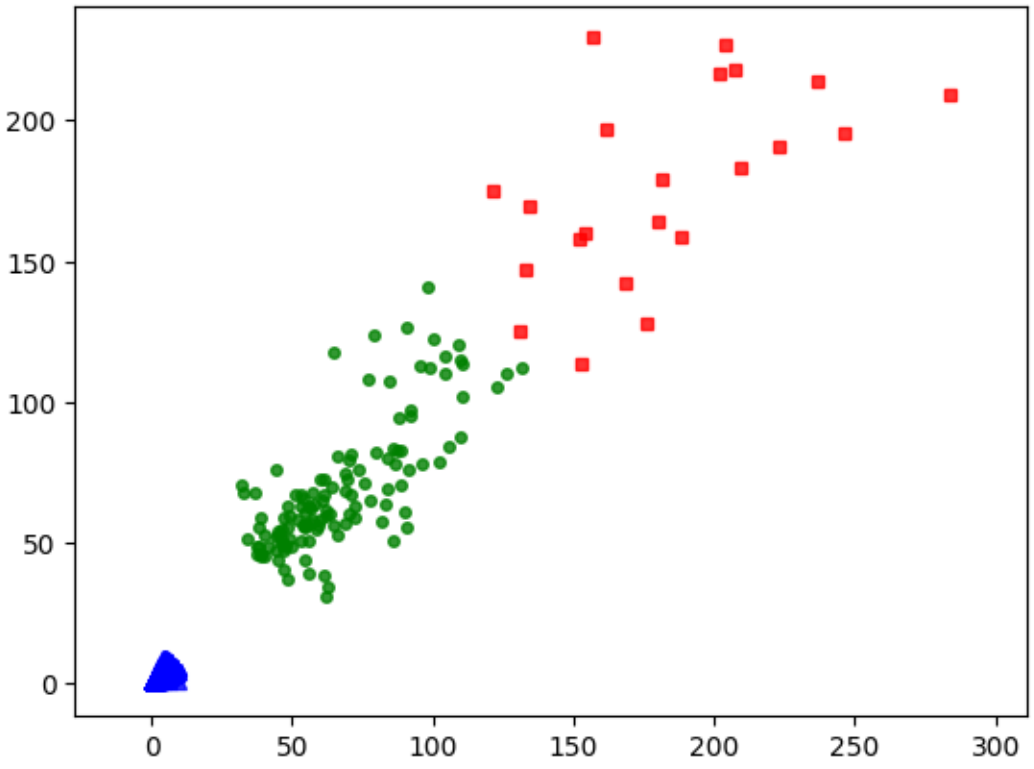
- **Ước tính tổng số nhân viên hiện tại theo dữ liệu "Total_employee_estimate"**



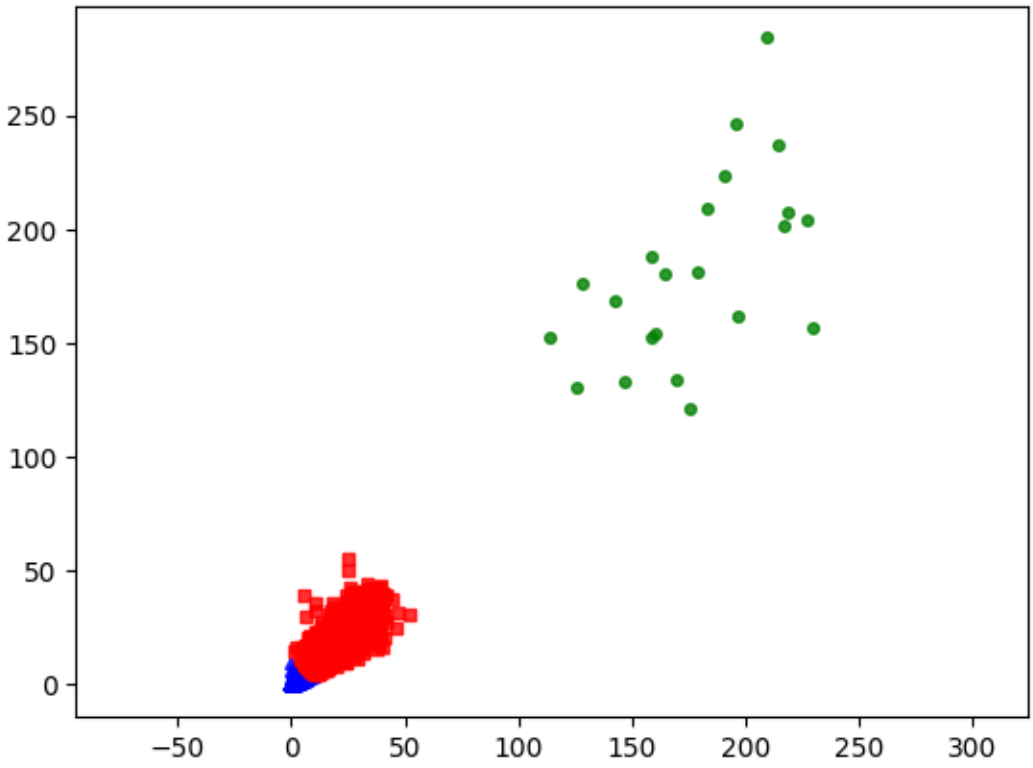
Hình 6.4. Biểu đồ phân cụm K-means của dữ liệu "Total_employee_estimate"

⇨ 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 2



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



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

7. K-Nearest Neighbors

```
KNN Model Evaluation:
[[    70   1049   5308   4373      6     91]
 [   410   5450  24640  19886     47    398]
 [  1015  13781  81663  79306    166   1149]
 [   975  13123  96070 134815    264    969]
 [    75    936   6532   8400     16     78]
 [   192   2524  14030  11923     21    225]]
               precision    recall  f1-score   support

       Africa       0.03      0.01      0.01     10897
         Asia       0.15      0.11      0.12     50831
       Europe       0.36      0.46      0.40    177080
North America       0.52      0.55      0.53    246216
      Oceania       0.03      0.00      0.00     16037
South America       0.08      0.01      0.01     28915

     accuracy                           0.42    529976
    macro avg       0.19      0.19      0.18    529976
 weighted avg       0.38      0.42      0.40    529976
```

Hình 7.1. Kết quả của mô hình KNN Model

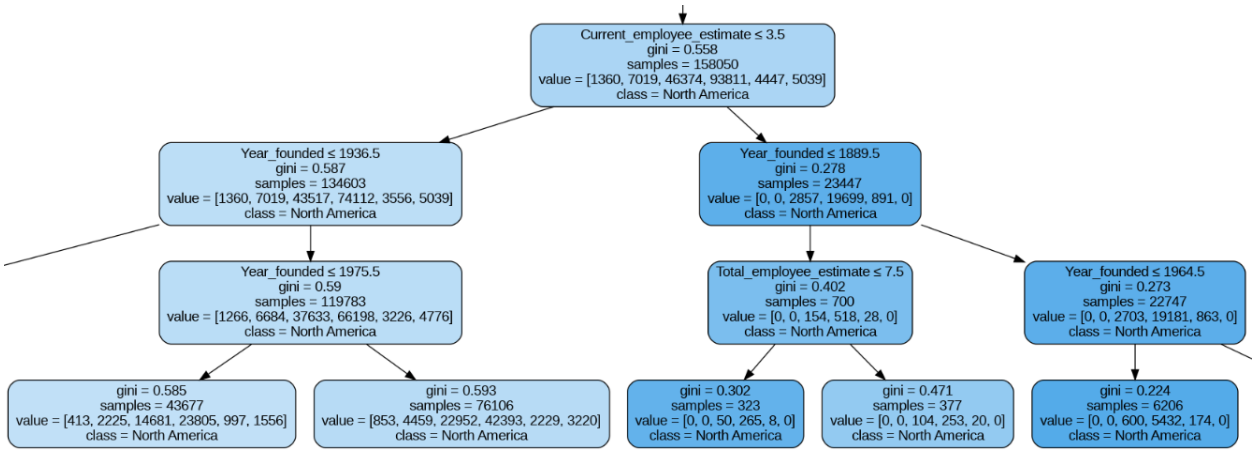
```
Naive-Bayes Model Evaluation on Test Data:
[[     0    908   1582    675      0   7732]
 [     0   5572   7387   3268      0  34604]
 [     0  11520  30151  20477      0 114932]
 [     0  13019  47313  37947      0 147937]
 [     0    863   2397   1948      0  10829]
 [     0   2213   4480   2478      0  19744]]
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/
  _warn_prf(average, modifier, msg_start, len(result))
               precision    recall  f1-score   support

       Africa       0.00      0.00      0.00     10897
         Asia       0.16      0.11      0.13     50831
       Europe       0.32      0.17      0.22    177080
North America       0.57      0.15      0.24    246216
      Oceania       0.00      0.00      0.00     16037
South America       0.06      0.68      0.11     28915

     accuracy                           0.18    529976
    macro avg       0.19      0.19      0.12    529976
 weighted avg       0.39      0.18      0.21    529976
```

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.

8. Cây quyết định

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

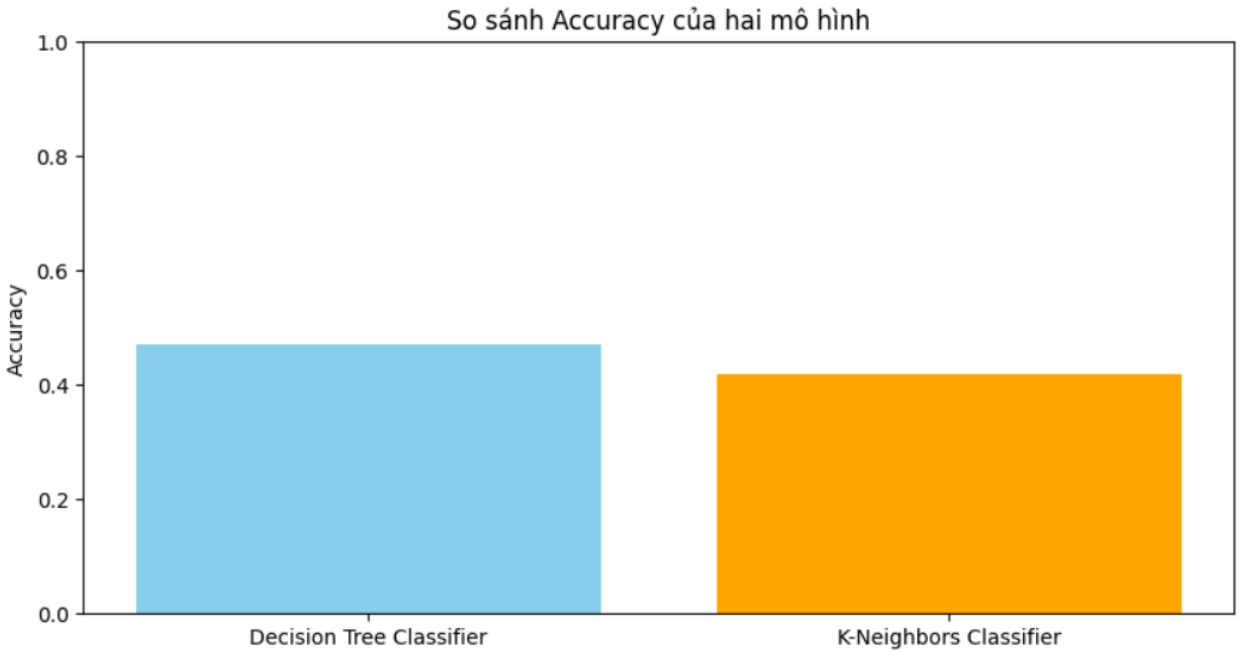


Hình 8.1. Biểu đồ hiển thị của cây quyết định

⇨ 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*.

9. 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.