TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN MÔN HỌC
KHAI PHÁ DỮ LIỆU

ĐỀ TÀI:
Phân tích, khai phá dữ kiệu tập dữ liệu doanh nghiệp

Sinh viên thực hiện:

Đặng Minh Phúc

Lớp: **ITEC3417**

Giảng viên hướng dẫn: **Nguyễn Trung Hậu**

Tháng 11 năm 2023

Mục lục

Lời nói đầu **3**

1. Bảng Phân Công **3**

2.Giới thiệu về nguồn dữ liệu, phương pháp thu thập **3**

3. Tiền xử lý dữ liệu: **3**

3.1 Phân tích tập dữ liệu 3

3.2. Xử lý dữ liệu 4

3.3 Biến đổi dữ liệu 4

4.Mô tả tập dữ liệu sau khi xử lý **4**

5. Phương pháp khai phá dữ liệu **5**

5.1 Trực quan hóa, biểu diễn dữ liệu 5

5.2 Mô hình hóa 5

6. Kết quả thực nghiệm của mô hình và đối sánh với các lập luận đã đưa ra. **5**

6.1 Trực quan hóa, biểu diễn dữ liệu 5

6.2 Mô hình hóa 7

7. Kết luận **9**

Lời nói đầu

Dữ liệu doanh nghiệp là dữ liệu vô cùng lớn với nhiều khía cạnh khai thác. Nhóm chúng tôi chọn sử dụng tập dữ liệu này với mong muốn tìm hiểu thêm về các ngành, lịch sử phát triển và lý do đằng sau sự phát triển đó

1. Bảng Phân Công

MSSV	Tên	Nhiệm vụ
2151010290	Đặng Minh Phúc	- Tiền xử lý dữ liệu thống kê, báo cáo
2151013020	Lê Trung Hậu	- Chọn mô hình khai phá phù hợp - Tổng hợp dữ liệu - Phương pháp đánh giá mô hình
1951052217	Hồ Sỹ Quang Trung	- Tìm kiếm, gộp, làm sạch data và mô tả tập dữ liệu sau khi xử lý.

2.Giới thiệu về nguồn dữ liệu, phương pháp thu thập

Đầu tiên, nhóm bắt đầu quá trình thu thập dữ liệu từ các nền tảng cung cấp tập dữ liệu mở như Kaggle, Gov, Goggle dataset,... Mục tiêu của nhóm là những thông tin dữ liệu quan đến nhiều khía cạnh khác nhau của doanh nghiệp để tiến hành khai phá và phân tích. Để đạt được điều này, nhóm đã tiến hành quá trình đối chiếu toàn diện trên nhiều nguồn thông tin khác nhau để đúc kết ra một bộ dữ liệu chuyên sâu về doanh nghiệp.

Qua quá trình này, nhóm đã tìm hiểu và tổng hợp được một tập dữ liệu đa dạng và đầy đủ về các khía cạnh của doanh nghiệp.

Các tập dữ liệu mà trong quá trình nhóm tìm hiểu, thu thập được bao gồm:

+ Bản ghi (record):
    - 7+ Million Company (link)
    - Continent List for 2021 Olympics in Tokyo Dataset (link)

3. Tiền xử lý dữ liệu:

3.1 Phân tích tập dữ liệu

Sau khi hoàn tất quá trình thu thập dữ liệu, cần xác định những thông tin trong tập dữ liệu đã được làm sạch chưa. Nếu dữ liệu chưa được làm sạch sẽ ảnh hưởng đến độ tin cậy của dữ liệu dẫn đến các quyết định không chính xác và gây khó khăn trong quá trình khai phá.

Một số vấn đề nhóm gặp phải khi chưa làm sạch dữ liệu trong tập dữ liệu **7+ Million Company**:

+ Dữ liệu không hoàn chỉnh (incomplete): Thiếu các giá trị thuộc tính:
    + Tổng giá trị bị mất: 8,512,762
        + Trong đó:
            + year_founded: 3510217 (50.11%)
            + industry: 277286 (3.96%)
            + country: 2440634 (34.84%)
            + locality: 2284625 (32.62%)
+ Nhiễu/Lỗi (noise/error):
    + Các giá trị year_founded lớn hơn năm hiện tại: 7 (2029, 2027, 2103, 2025, 2025, 2025, 2025)

Giải pháp:

- Tiến hành lọc từng cụm dữ liệu: xóa các đối tượng thiếu giá trị, lọc các giá trị nhiễu,…

Một số vấn đề nhóm gặp phải khi chưa làm sạch dữ liệu trong tập dữ liệu **Continent List for 2021 Olympics in Tokyo Dataset**:

+ Dữ liệu không trùng khớp (mismatched): một số giá trị với cùng tính chất nhưng khác tên gây thiếu sót trong gộp dữ liệu bên cạnh một số thuộc tính thiếu so với tập dữ liệu cần gộp

Giải pháp:

- Tiến hành chuẩn hóa tên các thuộc tính, bổ sung các giá trị còn thiếu,…

3.2. Xử lý dữ liệu

- Đối với tập dữ liệu '7+ Million Company':
    + Tiến thành xóa trống các đối tượng chứa thuộc tính nhiễu, thiếu nhưng không làm thay đổi tính chất, sự cân bằng của tập dữ liệu gốc.
- Đối với tập dữ liệu ''Continent List for 2021 Olympics in Tokyo Dataset':
    + Tiến hành đối chiếu các thuộc tính không trùng giữa tập dữ liệu này và '7+ Million Company', tiến hành tinh chỉnh tên giá trị, thêm các giá trị bị thiếu, loại bỏ 1 số giá trị không mang nhiều ý nghĩa để giảm nhẹ chi phí tính toán mô hình mà không làm ảnh hướng dữ liệu chung.

3.3 Biến đổi dữ liệu

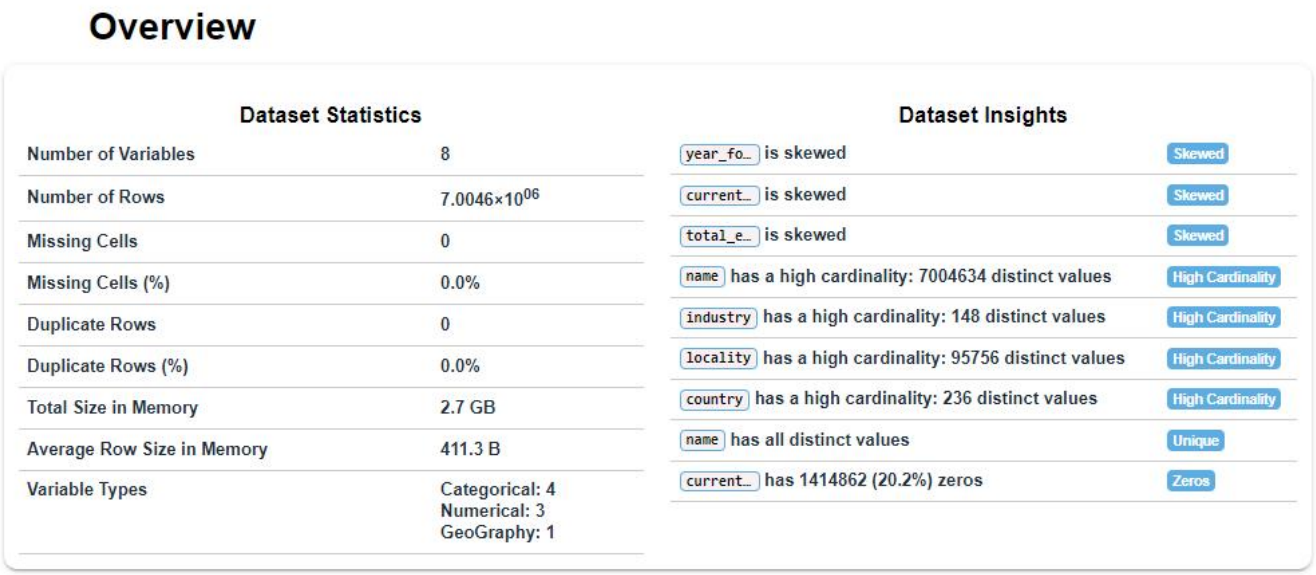
Biến đổi dữ liệu là việc chuyển toàn bộ tập giá trị của một thuộc tính sang một tập giá trị thay thế sao cho giá trị cũ tương ứng giá trị mới.

Trong bài tập lớn này nhóm chúng tôi sử dụng các phương pháp biến đổi sau:

-Làm trơn (smoothing): Loại bỏ nhiễu/lỗi khỏi tập dữ liệu.

-Chuẩn hóa (normalization): Đưa các giá trị về một khoảng được chỉ định.

4.Mô tả tập dữ liệu sau khi xử lý



Overview

Dataset Statistics

Number of Variables	8
Number of Rows	7.0046×10^{06}
Missing Cells	0
Missing Cells (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	2.7 GB
Average Row Size in Memory	411.3 B
Variable Types	Categorical: 4 Numerical: 3 GeoGraphy: 1

Dataset Insights

year_fo… is skewed	Skewed
current… is skewed	Skewed
total_e… is skewed	Skewed
name has a high cardinality: 7004634 distinct values	High Cardinality
industry has a high cardinality: 148 distinct values	High Cardinality
locality has a high cardinality: 95756 distinct values	High Cardinality
country has a high cardinality: 236 distinct values	High Cardinality
name has all distinct values	Unique
current… has 1414862 (20.2%) zeros	Zeros

- Tổng quan tập dữ liệu sau tiền xử lý:
  - 7046000 đối tượng
  - 8 thuộc tính
  - Số giá trị bị mất 0 (0%)
  - Các cột giá trị bị lệch: ‘year_founded’, ’ current_employee_estimate’, ‘ total_employee_estimate’
  - Các cột có giá trị 0:
    - current_employee_estimate’: 1414862 (20.2%)

5. Phương pháp khai phá dữ liệu

5.1 Trực quan hóa, biểu diễn dữ liệu

Chúng tôi sử dụng các biểu đồ cũng như các bảng để thể hiện các xu hướng dữ liệu, sự phân bổ để tìm ra các xu hướng cũng như giải thích, chứng minh cho cá xu hướng đó

5.2 Mô hình hóa

Chúng tôi sử dụng mô hình gom cụm cùng thuật toán K-means cho tập dữ liệu này vì đây là mô hình khá phổ biến cũng như dễ áp dụng để tìm ra các đặc trưng trong các cụm dữ liệu.

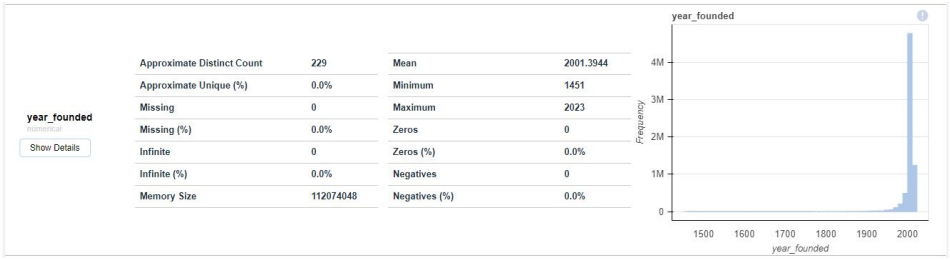
Vấn đề đặt ra cho việc gom cụm dữ liệu là chọn các k clusters phù hợp. Có nhiều cách tiếp cận để tìm k cluster như: Elbow method, AffinityMatrix, SilhoutteScore,… và dựa vào kinh nghiệm cá nhân. Trong tập dữ liệu này chúng tôi tiếp cận với Elbow method là một trong những cách phổ biến nhất để tối ưu k clusters. Chúng tôi lặp lại các giá trị của k từ 1 đến 11 và tính toán độ biến dạng hoặc quán tính cho từng giá trị của k trong phạm vi đã cho. Trong đó Quán tính là tổng bình phương khoảng cách của các mẫu đến tâm cụm gần nhất của chúng. Để xác định số cụm tối ưu, chúng ta phải chọn giá trị k tại “khuỷu tay”, tức là điểm mà sau đó độ méo/quán tính bắt đầu giảm theo kiểu tuyến tính.

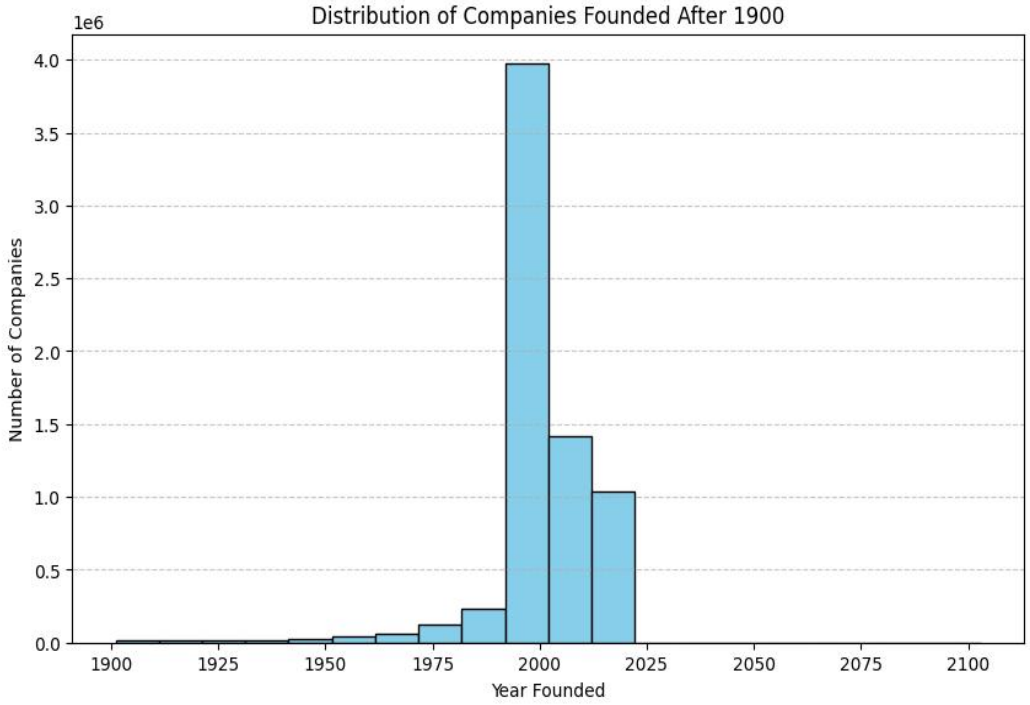
6. Kết quả thực nghiệm của mô hình và đối sánh với các lập luận đã đưa ra.

6.1 Trực quan hóa, biểu diễn dữ liệu

Tổng quan:

- Thuộc tính năm thành lập:

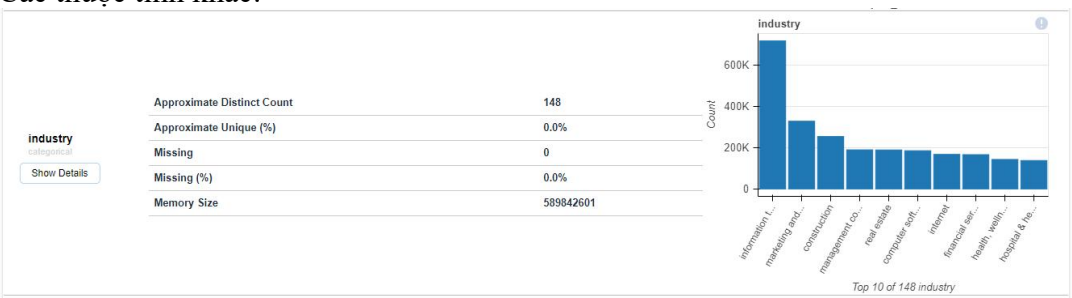


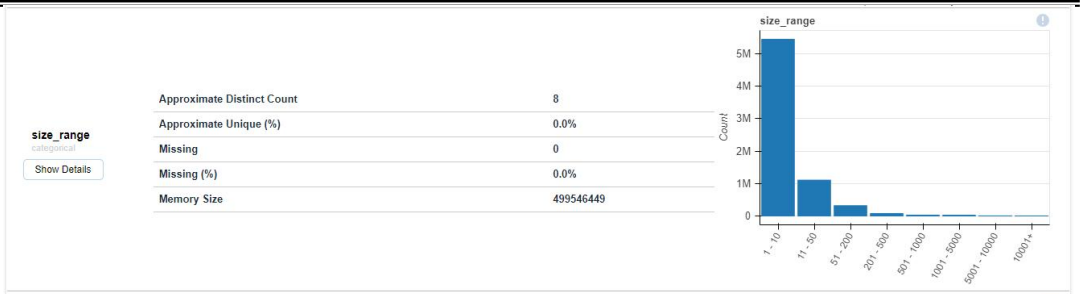


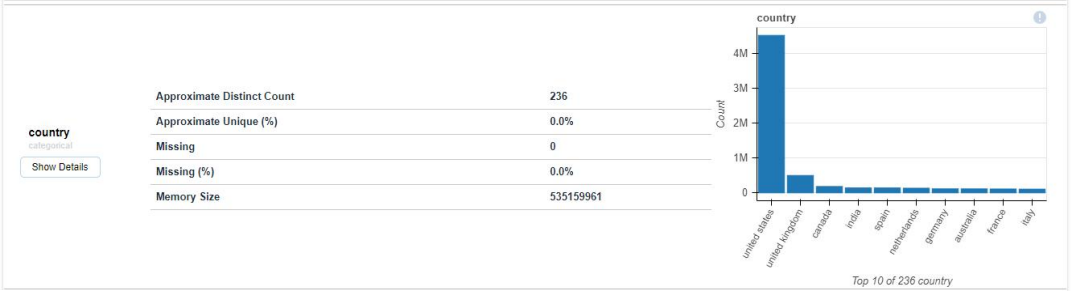
Biểu đồ phân bố năm thành lập sau 1900

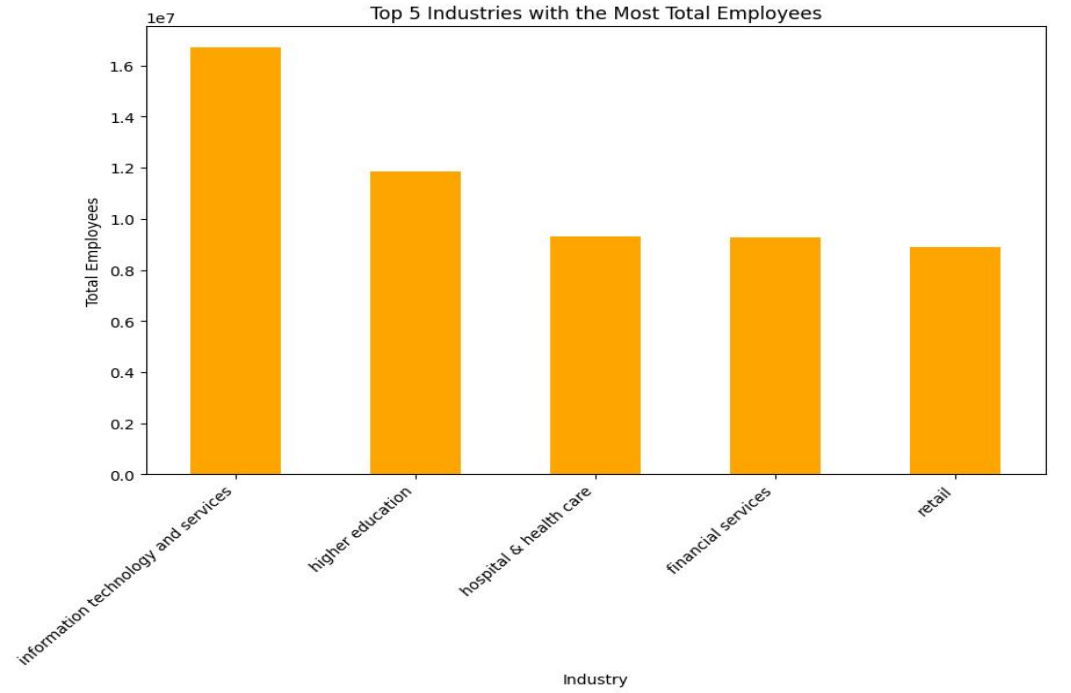
⇨ Phần lớn các công ty thành lập vào những năm 2000 với giá trị Mean rơi vào giá trị 2001. Trong đó công ty thành lập cũ nhất vào năm 1451 và mới nhất vào 2023

- Các thuộc tính khác:







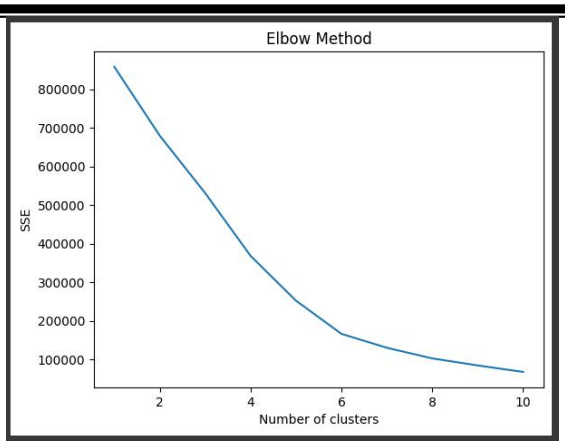


Ngành công nghệ thông tin chiếm đa số, phần lớn công ty thành lập có quy mô nhỏ và Mỹ là quốc gia với nhiều công ty nhất.

- Phần lớn các công ty CNTT ra đời vào những năm 2000 cùng với sự phát triển nhanh của mạng internet.
    - ⇨ Xu hướng phát triển các công ty đang theo hướng ngành CNTT trong thời đại phát triển nhanh của các công nghệ hiện đại

6.2 Mô hình hóa

- Sau khi thực thi thuật toán chúng tôi tìm ra số cụm tối ưu cho tập dữ liệu này là 6



- Sau đó chúng tôi tiến hành gom cụm bằng thuật toán K-means
  Kết quả sau khi gom cụm như sau:

```
Cluster 1
['501 - 1000']

Cluster 2
['201 - 500']

Cluster 3
['1001 - 5000']

Cluster 4
['10001+']

Cluster 5
['5001 - 10000']

Cluster 6
['10001+']
```

```
Cluster 1
['north america' 'oceania' 'europe' 'south america' 'asia' 'africa' nan]

Cluster 2
['north america' 'oceania' 'europe' 'asia' 'south america' 'africa' nan]

Cluster 3
['north america' 'europe' 'oceania' 'asia' 'south america' 'africa' nan]

Cluster 4
['north america' 'europe' 'asia' 'south america' 'oceania' 'africa']

Cluster 5
['north america' 'oceania' 'europe' 'asia' 'africa' 'south america']

Cluster 6
['north america' 'asia' 'europe']
```

```
Cluster 1
[500 499 498 497 496 495 494 493 492 491 490 489 488 487 486 485 484 483
 482 481 480 479 478 477 476 475 474 473 472 471 470 469 468 467 466 465
 464 463 462 461 460 459 458 457 456 455 454 453 452 451 450 449 448 447
 446 445 444 443 442 441 440 439 438 437 436 435 434 433 432 431 430 429
 428 427 426 425 424 423 422 421 420 419 418 417 416 415 414 413 412 411
 410 409 408 407 406 405 404 403 402 401 400 399 398 397 396 395 394 393
 392 391 390 389 388 387 386 385 384 383 382 381 380 379 378 377 376 375
 374 373 372 371 370 369 368 367 366 365 364 363 362 361 360 359 358 357
 356 355 354 353 352 351 350 349 348 347 346 345 344 343 342 341 340 339
 338 337 336 335 334 333 332 331 330 329 328 327 326 325 324 323 322 321
 320 319 318 317 316 315 314 313 312 311 310 309 308 307 306 305 304 303
 302 301 300 299 298 297 296 295 294 293 292 291 290 289 288 287 286 285
 284 283 282 281 280 279 278 277 276 275 274 273 272 271 270 269 268 267
 266 265 264 263 262 261 260 259 258 257 256 255 254 253 252 251 250 249
 248 247 246 245 244 243 242 241 240 239 238 237 236 235 234 233 232 231
 230 229 228 227 226 225 224 223 222 221 220 219 218 217 216 215 214 213
 212 211 210 209 208 207 206 205 204 203 202 201 200 199 198 197 196 195
 194 193 192 191 190 189 188 187 186 185 184 183 182 181 180 179 178 177
 176 175 174 173 172 171 170 169 168 167]

Cluster 2
[250 249 248 247 246 245 244 243 242 241 240 239 238 237 236 235 234 233
 232 231 230 229 228 227 226 225 224 223 222 221 220 219 218 217 216 215
 214 213 212 211 210 209 208 207 206 205 204 203 202 201 200 199 198 197
 196 195 194 193 192 191 190 189 188 187 186 185 184 183 182 181 180 179
 178 177 176 175 174 173 172 171 170 169 168 167 166 165 164 163 162 161
 160 159 158 157 156 155 154 153 152 151 150 149 148 147 146 145 144 143
 142 141 140 139 138 137 136 135 134 133 132 131 130 129 128 127 126 125
 124 123 122 121 120 119 118 117 116 115 114 113 112 111 110 109 108 107
 106 105 104 103 102 101]

Cluster 3
[2500 2499 2498 ...  336  335  334]

Cluster 4
[51441 48806 47434 ...  3344  3339  3337]

Cluster 5
[4995 4994 4993 ... 1670 1668 1667]

Cluster 6
[274047 190771 190689 162163 158363 127952 122031 120753 116196 115188
 113997 111372 109532 104752 104112 101482  97357  95234  94458  93247
  90095  87381  85090  84327  84218  84179  78261  75640  75109  74357
  68233  67692  67564  67261  66632  65839  65335  64046  62685  61803
  61638  61040  60602  60324  59993  59712  59588  58819  58538  57720
  57255  55945  54117  51721  50715  50303  46523  44799  42043  38917]
```

⇨ Các cụm phân chia theo quy mô công ty. Trong đó đặc biệt với cụm 6 gồm các công ty với quy mô lớn và số lượng nhân viên đông đảo nhất tập trung ở các quốc gia thuộc châu Âu, châu Á và châu Mỹ. Trong khi đó các cụm còn lại với quy mô thấp hơn phân bổ đều cho các châu lục

7. Kết luận

Sau khi thực hiện khai phá nhóm rút ra các kết quả, hạn chế cũng như các đề xuất phát triển như sau:

- Về kết quả: Nhóm đã dùng các kỹ thuật để tiến hành khai phá dữ liệu, tìm hiểu được các góc nhìn của tập dữ liệu, là bước đầu cho các công việc trong việc khai thác tri thức
- Về mặt hạn chế:
  - Nhóm còn hạn chế trong kinh nghiệm đọc các dữ liệu sau phân tích, các kết quả sau phân tích thường được thể hiện rõ qua trực quan mà chưa có kinh nghiệm tron khai thác sâu dữ liệu
  - Do hạn chế về kiến thức nên nhóm chỉ thực hiện 1 mô hình, vì không đảm bảo tối ưu và kết quả khi áp dụng nhiều mô hình
  - Xử lý dữ liệu chưa thật sự tốt với các kết quả sau gom cụm còn lọt nhiều giá trị nhiễu gây ảnh hướng kết quả chung
- Về mặt đề xuất phát triển:
  - Nhóm có thể dành thêm thời gian đầu tư tham khảo các phương pháp khai phá khác, học hỏi thêm kinh nghiệm về dữ liệu

- Tối ưu, cải tiến xử lý dữ liệu sâu hơnn để đảm bảo không lọt các giá trị không đáp ứng yêu cầu của mô hình nhằm đạt kết quả cao hơn
- Áp dụng nhiều loại mô hình nhiều các tối ưu mô hình hơn để có thêm nhiều tri thức, góc nhìn sâu hơn khi gặp các tập dữ liệu phức tạp khác.

Phục lục:

Github: link