TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

MÔN HỌC: PHÂN TÍCH NHẬN DẠNG MẪU

ĐỀ TÀI: PHÂN CỤM KHÁCH HÀNG

Giảng viên hướng dẫn: Vũ Ngọc Thanh Sang

Sinh viên tham gia: Nguyễn Lê Đăng Khoa

MSSV: 3120410247

Ngày 21 tháng 12 năm 2023

Mục lục

LỜI NÓI ĐẦU

Trong bối cảnh kinh tế ngày nay, thị trường đang đối mặt với những thách thức ngày càng lớn, đặc biệt là khi môi trường kinh doanh trở nên phức tạp và đa dạng. Đối với các doanh nghiệp, việc hiểu rõ khách hàng không chỉ là một ưu tiên mà còn là chìa khóa để tối ưu hóa chiến lược kinh doanh và cung cấp trải nghiệm tốt nhất cho khách hàng. Một trong những thách thức lớn nhất đối mặt với doanh nghiệp ngày nay là làm thế nào để hiểu rõ và phục vụ một cách hiệu quả một thị trường ngày càng đa dạng. Khách hàng không chỉ đòi hỏi sự cá nhân hóa mà còn mong đợi được đối xử như là những cá nhân độc lập có nhu cầu và mong muốn riêng biệt.

Chính vì vậy, để có thể dễ dàng quản lý một khối lượng thông tin lớn như vậy cũng như là phải hiểu rõ và phân tích được các khách hàng tiềm năng. Thì hôm nay, em xin được phép giới thiệu về một kĩ thuật gom nhóm hữu ích đó là K-Means, kĩ thuật này sẽ được ứng dụng vào chủ đề đó là: "Phân cụm khách hàng". Việc áp dụng kĩ thuật này để giải quyết thì em đã thấy được sự hiệu quả của nó. Nội dung của đề tài sẽ gồm 3 chương như sau:

- **Chương I:** Giới thiệu các vấn đề khách hàng và những thông tin cần thiết.
- **Chương II:** Trình bày cơ sở lý thuyết của những phương pháp sẽ được áp dụng.
- **Chương III:** Xây dựng chương trình và đánh giá kết quả.

CHƯƠNG I: Tổng Quan

1.1. Giới thiệu

1.1.1. Nhu cầu thực tế

Hiện nay, việc phát triển kỹ thuật ngày càng mạnh mẽ cũng đồng nghĩa với việc nhu cầu lưu trữ lượng thông tin cũng sẽ không ngừng gia tăng theo. Việc ta có thể xử lý được khối thông tin này một cách hiệu quả cũng là một vấn đề nan giải. Để có thể sắp xếp và phân bổ lượng thông tin này một cách hợp lý thì ta sẽ cần đến sự trợ giúp của công nghệ, cũng như là là các thuật toán đã và đang được phát triển hiện nay.

Đối với hầu hết các ngành nghề, nó đều liên quan đến bộ cơ sở dữ liệu khổng lồ nhằm lưu trữ thông tin của các khách hàng và hiển hiển nhiên việc xử lý hết các dữ liệu với hiệu quả cao là 1 vấn đề khó khăn. Chính vì vậy, với việc hướng tới giải quyết vấn đề này cho các ngành nghề, em sẽ nhắm đến cung cấp cho người dùng một bộ dữ liệu đã xử lý hoàn chỉnh với các cụm thông tin khách hàng rõ ràng thông qua các hoạt động của khách hàng đã và đang sử dụng. Với chương trình này, ta sẽ không phải còn phải đối mặt với hàng tá dữ liệu bị trộn lẫn cũng như là nằm một cách rải rác, đồng thời còn có thể nắm rõ được từng cụm dữ liệu sẽ mang ý nghĩa thế nào rồi từ đó ta có thể quyết định được bước tiếp theo trong công việc.

1.1.2. Các ứng dụng trong thực tiễn hiện nay

Phân cụm là một kỹ thuật quan trọng và rộng rãi được sử dụng trong nhiều lĩnh vực khác nhau. Dưới đây là một số ứng dụng thực tiễn hiện nay của phân cụm:

+ Phân loại khách hàng: Để hiểu thị trường của mình, một công ty có thể phân cụm các khách hàng của mình thành các nhóm khác nhau dựa trên các yếu tố như độ tuổi, giới tính, thu nhập và sở thích mua sắm.

+ Phân loại tin tức: Một trang tin tức có thể sử dụng phân cụm để phân loại các tin tức của mình vào các nhóm khác nhau dựa trên chủ đề, thể loại hoặc độ phổ biến.

+ Phân loại đối tượng trong hình ảnh: Một hệ thống nhận diện hình ảnh có thể sử dụng phân cụm để phân loại các đối tượng trong hình ảnh thành các nhóm khác nhau, chẳng hạn như loại động vật hoặc loại thực phẩm.

Những ứng dụng trên chỉ là một số ví dụ và phân cụm có thể được áp dụng trong nhiều lĩnh vực khác nhau tùy thuộc vào nhu cầu cụ thể của ngành và doanh nghiệp.

1.2. Các hướng tiếp cận của bài toán

Bài toán phân cụm này cần phải xác định được hai vấn đề chính. Đầu tiên là xác định được thông tin đó là dạng thông tin gì, các đặc trưng cơ bản của nó là gì? Và thứ hai là cần áp dụng phương pháp nào để giải quyết dạng thông tin đó.

Có nhiều hướng tiếp cận để giải quyết bài toán phân cụm, tuy nhiên, các phương pháp phân cụm thường được chia thành 3 nhóm chính:

1. Phân cụm dựa trên khoảng cách: Phương pháp này đo khoảng cách giữa các điểm dữ liệu và phân loại chúng vào các nhóm dựa trên khoảng cách tương đối của chúng. Các phương pháp phổ biến trong nhóm này bao gồm K-Means và Phân cụm Hierarchical.
2. Phân cụm dựa trên tiêu chí xác suất: Phương pháp này sử dụng các mô hình xác suất để phân loại các điểm dữ liệu vào các nhóm. Các phương pháp phổ biến trong nhóm này bao gồm Gaussian Mixture Model (GMM) và Expectation-Maximization (EM).
3. Phân cụm dựa trên mô hình: Phương pháp này sử dụng mô hình học máy để phân loại các điểm dữ liệu vào các nhóm. Các phương pháp phổ biến trong nhóm này bao gồm phân cụm dựa trên SVM (Support Vector Machine) và phân cụm dựa trên mạng neural.

Mỗi phương pháp có ưu điểm và nhược điểm riêng, tùy thuộc vào đặc tính của dữ liệu và mục đích của việc phân cụm để lựa chọn phương pháp phù hợp.

1.3. Đề xuất hướng tiếp cận

Để giải quyết bài toán phân cụm này, hướng tiếp cận của em sẽ là áp dụng thuật toán K-Means để phân cụm. Và với thuật toán đó, em sẽ thực hiện phân cụm khách hàng dựa trên thông tin của khách hàng để đưa ra được các dự đoán cho từng nhóm khách hàng, để phù hợp cho việc maketing của trung tâm thương mại. Việc áp dụng sẽ được thực hiện với các bước như sau:

1. Chuẩn bị dữ liệu.
2. Tiền xử lý dữ liệu: Xử lý và làm sạch bộ dữ liệu, bao gồm việc xử lý các dữ liệu bị thiếu, các dữ liệu bị nhiễu và chuẩn hóa dữ liệu.
3. Phân tích giả định các nhóm khách hàng trước khi huấn luyện mô hình, cũng như là phác họa lên các thông tin của bộ dữ liệu để tìm ra giải pháp hợp lý nhất.

4. Tiến hành huấn luyện mô hình K-Means: Sử dụng thuật toán để phân cụm khách hàng. Mô hình sẽ xác định được các cụm khách hàng dựa trên các đặc trưng.
5. Đưa ra kết quả: Sau khi đã xác định được các cụm khách hàng, tiến hành phân tích các cụm khách hàng đó.

CHƯƠNG II: Cơ sở lý thuyết

2.1. Phân cụm

Phân cụm (clustering) là một phương pháp trong thống kê và máy học được sử dụng để nhóm các điểm dữ liệu có đặc trưng tương đồng lại với nhau thành các cụm hoặc nhóm. Mục tiêu của phân cụm là tạo ra các nhóm sao cho các điểm trong một nhóm gần giống nhau, trong khi giữ cho các nhóm có đặc trưng khác biệt với nhau.

Thuật toán phân cụm giúp hỗ trợ quá trình khám phá cấu trúc ẩn trong dữ liệu và phân loại dữ liệu thành các nhóm có ý nghĩa. Các phương pháp phân cụm thường dựa trên đo lường sự tương đồng hoặc khoảng cách giữa các điểm dữ liệu.

2.2. K-Means

K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát(tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau. [1]

2.3. Nguyên lý hoạt động của K-Means

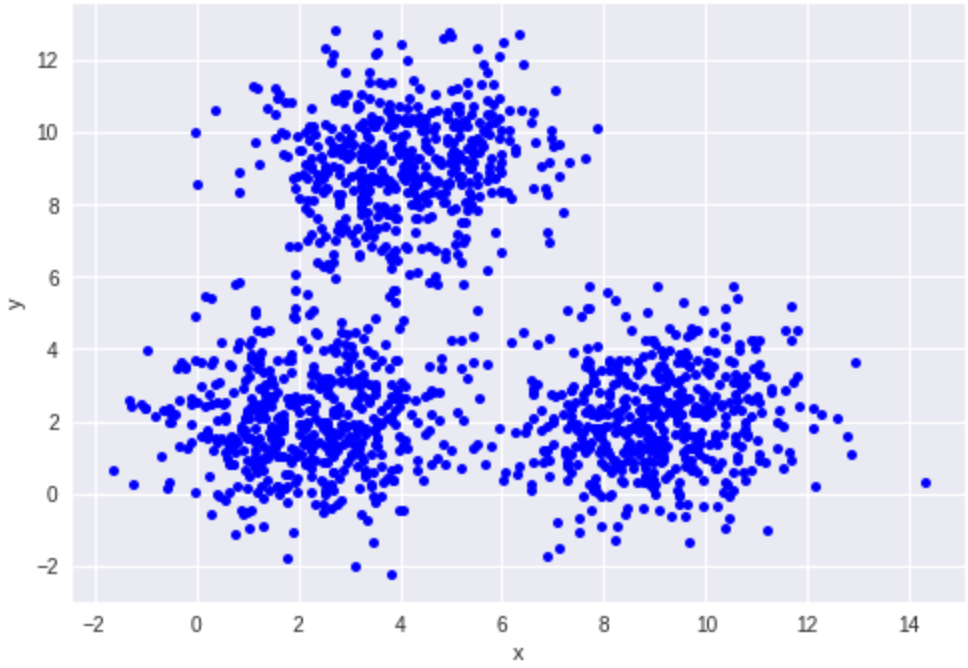
Thuật toán K-Means sẽ được thực hiện theo các bước sau:

1. Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu của chúng ta.
2. Với mỗi điểm dữ liệu trong bộ dữ liệu, tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.
3. Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.
4. Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

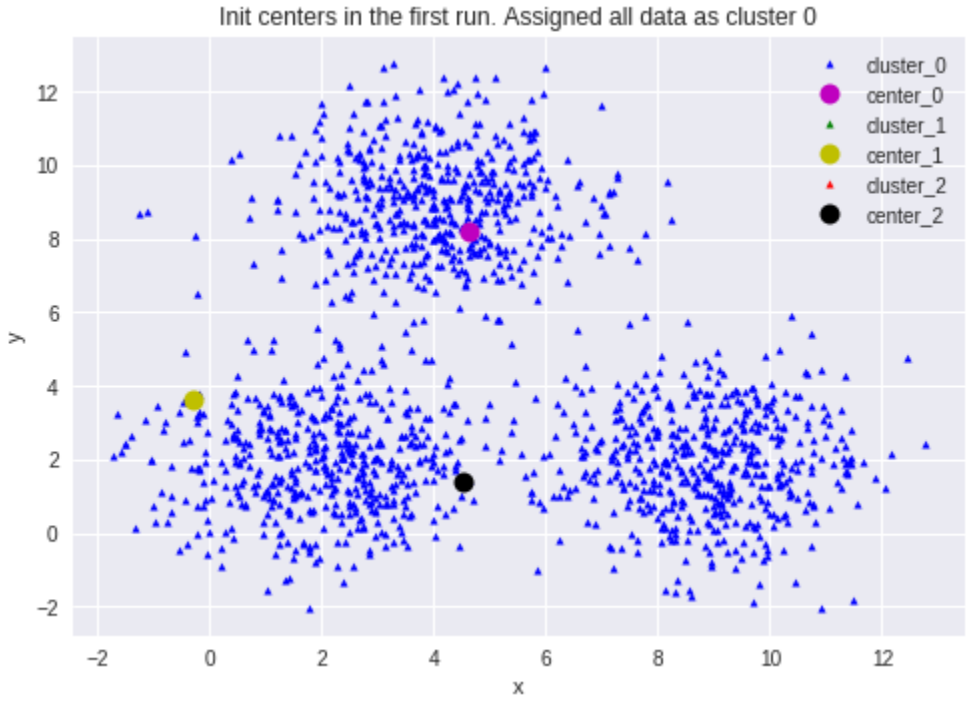
Trên thực tế, sẽ có 1 vài lưu ý cần phải giải quyết khi áp dụng thuật toán k-means, đó là:

Cách hoạt động của thuật toán K-Means như sau:

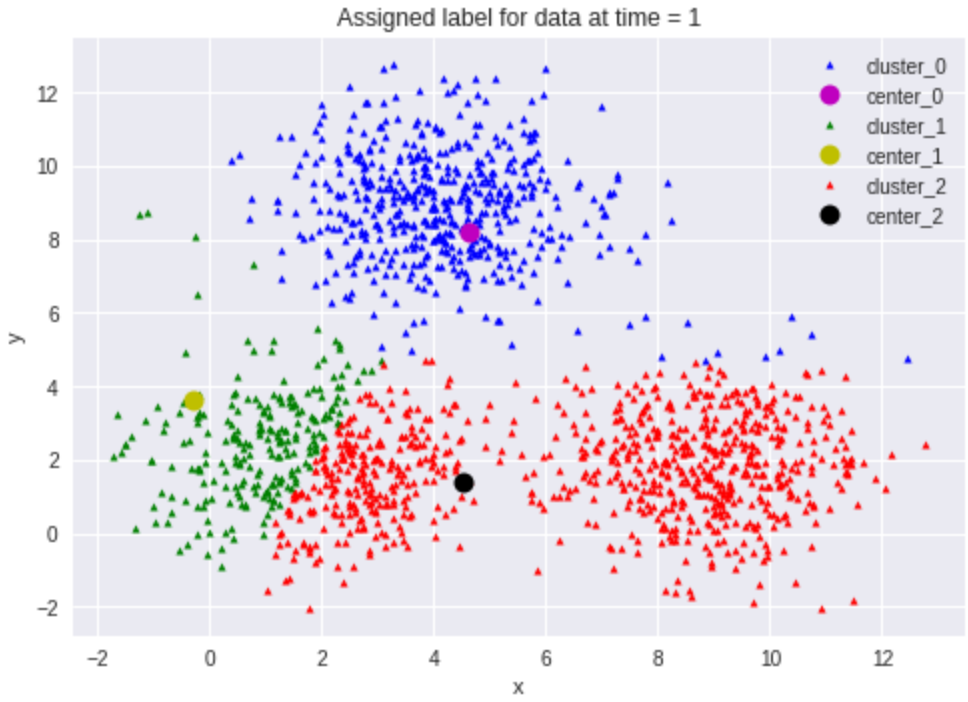
- Đầu tiên ta sẽ có bộ dữ liệu như thế này:



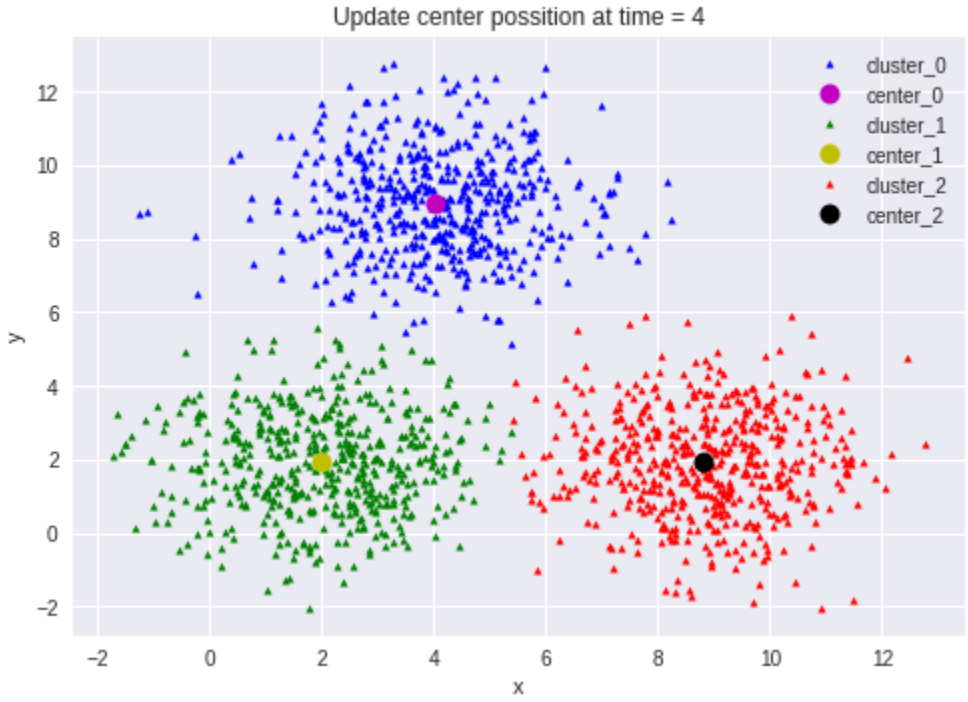
- Tiếp theo nó sẽ tiến hành tìm ra các tâm của các cụm:



- Sau đó sẽ tiến hành tô màu các các dữ liệu gần với cụm đó nhất thành các màu giống như màu của cụm, nó sẽ như sau:



- Và thuật toán sẽ cứ lập lại liên tiếp như vậy cho đến khi nào đã phân thành các cụm tách biệt nhất như sau:



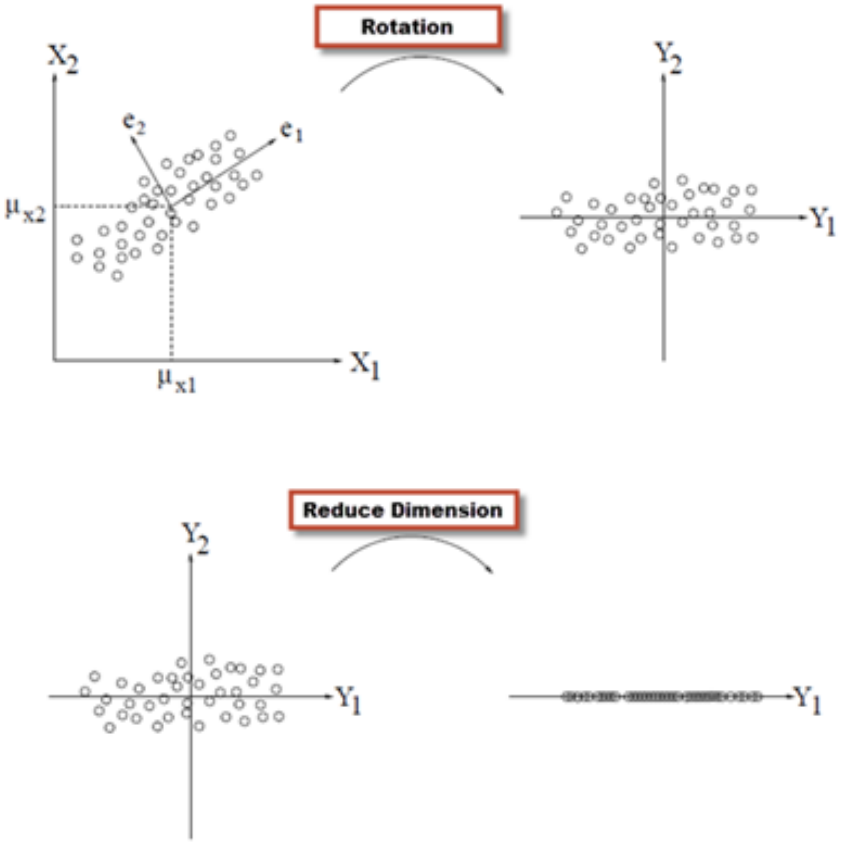
2.4. PCA – (Principal Component Analysis)

PCA là viết tắt của Principal Component Analysis. Ta dịch thô sang tiếng Việt là "Phân tích thành phần chính", tạm hiểu là ta sẽ phân tích dữ liệu và sau đó tìm ra các thành phần chính của dữ liệu để giữ lại các thành phần đó [2]. Như vậy, để có thể giải quyết được vấn đề trực quan hóa các cụm lên màn ảnh thì ta cần phải áp dụng PCA để có thể dễ dàng nhìn được các đặc trưng tương quan với nhau và thấy rõ được các cụm được phân biệt với nhau như thế nào.

Việc làm như trên sẽ giúp cho ta:

- Giảm chiều dữ liệu mà vẫn giữ được đặc trưng chính, chỉ mất đi "chút ít" đặc trưng.
- Tiết kiệm thời gian, chi phí tính toán
- Dễ dàng visualize dữ liệu hơn để giúp ta có cái nhìn trực quan hơn.

Ta sẽ có ví dụ minh họa về PCA như sau:



Ta thấy rằng dữ liệu trên trục mới đã giảm sự tương quan đáng kể (biến Y1 và Y2 gần như không tương quan), và sự thay đổi của dữ liệu phụ thuộc phần lớn vào biến Y1, ta có thể chỉ dung một biến Y1 để biểu diễn dữ liệu, điều này giúp ta giảm số chiều dữ

liệu mà không làm giảm quá nhiều “phương sai” của dữ liệu. Đây cũng chính là tư tưởng của phương pháp PCA.

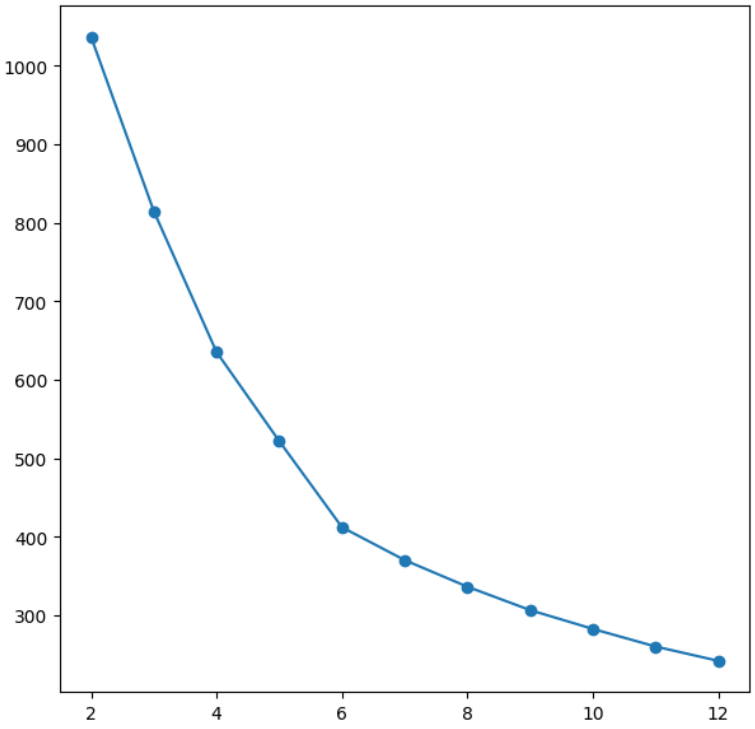
2.5. Phương pháp Elbow

Phương pháp Elbow là một kỹ thuật được sử dụng để chọn số lượng cụm tối ưu trong các thuật toán phân cụm, đặc biệt là trong trường hợp của K-means clustering. Mục tiêu của phương pháp này là tìm ra số lượng cụm (k) mà khi sử dụng, sự gia tăng trong hiệu suất của mô hình không đáng kể nữa, tạo ra một biểu đồ giống như "khuỷu tay" (elbow).

Quy trình sẽ được thực hiện như sau:

1. **Thực hiện phân cụm cho các giá trị k khác nhau**: Áp dụng thuật toán phân cụm (thường là K-means) với các giá trị k khác nhau và tính toán tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến trung tâm của cụm tương ứng.
2. **Vẽ biểu đồ Elbow**: Với mỗi giá trị k, tính tổng bình phương khoảng cách và vẽ một biểu đồ. Thông thường, bạn sẽ thấy một điểm trên biểu đồ mà nếu bạn kéo một đường từ nó đến gốc tọa độ (0,0), điểm này giống như "một phương gối" (elbow) trên cánh tay.
3. **Chọn số cụm tối ưu**: Số cụm tối ưu thường được chọn tại điểm "elbow" trên biểu đồ. Điểm này đại diện cho sự giảm đột ngột trong tổng bình phương khoảng cách, và từ đó, thêm cụm không còn mang lại sự cải thiện đáng kể.

Phương pháp Elbow sẽ có hình minh họa như sau:



Ở phần hoành độ sẽ tương ứng với cụm của các đốt. Và ta có thể thấy phần "khuỷu" sẽ nằm ở cụm số 6. Đồng nghĩa với việc ta có thể sử dụng k=6 và áp dụng cho việc phân cụm với k cụm.

2.6. Thang đó đánh giá phân cụm Silhoette

Silhouette là một phương pháp đo độ tách biệt của các nhóm trong phân tích cụm (clustering). Nó cung cấp một số đo lường cho việc xác định xem các điểm dữ liệu được gán vào một cụm cụ thể có tách biệt và tập trung hay không. Phương pháp này đo lường độ tách biệt của các nhóm bằng cách tính toán độ tương đồng giữa các điểm dữ liệu trong cùng một cụm và độ khác biệt giữa các cụm khác nhau [3].

Silhouette Score cho mỗi cụm dữ liệu sẽ được tính với công thức như sau:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Trong đó:

- s(i) là Silhouette Score của điểm dữ liệu i.
- a(i) là khoảng cách trung bình từ điểm i đến tất cả các điểm trong cùng 1 cụm.
- b(i) là khoảng cách trung bình từ điểm i đến tất cả các điểm trong cụm lân cận gần nhất (không phải cụm của nó).

Công thức trên tạo ra một giá trị từ -1 đến 1.

- Giá trị gần 1 cho thấy điểm i nằm gần trung tâm của cụm và xa trung tâm của các cụm khác, tức là phân loại tốt.
- Giá trị gần -1 cho thấy điểm i nằm gần trung tâm của cụm khác và xa trung tâm của cụm nó thuộc, tức là phân loại không tốt.
- Giá trị bằng 0 có thể xảy ra khi a(i) = b(i), cho thấy điểm i ở giữa giữa hai cụm.

CHƯƠNG III: Xây dựng chương trình và đánh giá kết quả

3.1. Giới thiệu chủ đề

Với phương pháp áp dụng K-Means vào việc phân cụm khách hàng, thì hôm nay ta sẽ áp dụng phương pháp đó vào bộ dữ liệu khách hàng của một trung tâm thương mại để nhằm đưa ra các cụm khách hàng riêng biệt với các đặc trưng nhất định. Với việc phân cụm này sẽ đưa ra được các phân tích đánh giá phù hợp để nhằm giúp cho việc maketing của trung tâm có phần chắc chắn và phù hợp hơn.

3.2. Mô tả dữ liệu

Tên thuộc tính	**Mô tả**
ID	ID của khách hàng.
Sex	Giới tính, gồm: + 0: Đàn ông. + 1: Phụ nữ.
Marital Status	Tình trạng hôn nhân, gồm: + 0: Độc thân. + 1: Đã kết hôn.
Age	Tuổi.
Education	Trình độ học vấn: + 0: Không xác định. + 1: Tốt nghiệp cấp 3. + 2: Trình độ đại học. + 3: Sau đại học.
Income	Thu nhập cá nhân
Occupation	Tình trạng nghề nghiệp: + 0: Thất nghiệp. + 1: Nhân viên thông thường. + 2: Quản lý / Tự làm chủ / Nhân viên / Cán bộ có trình độ cao.
Settlement size	Độ lớn thành phố đang ở: 0: Thành phố nhỏ. 1: Thành phố vừa. 2: Thành phố lớn.

3.3. Thống kê dữ liệu

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 4   Education        2000 non-null   int64
 5   Income           2000 non-null   int64
 6   Occupation       2000 non-null   int64
 7   Settlement size  2000 non-null   int64
dtypes: int64(8)
memory usage: 125.1 KB
```

Ở đây, ta có thể thấy ta sẽ có tổng cộng là 8 đặc trưng và 2000 dòng dữ liệu. Nhìn chung, bộ dữ liệu của ta sẽ chỉ gồm kiểu int64. Và các đặc trưng sẽ chỉ gồm 2 loại đó là dạng numberical – có 3 đặc trưng là ID, Income và Age. Còn loại còn lại là dạng categorical – có 5 đặc trưng là Sex, Marital status, Education, Occupation và Settlement size.

3.4. Xử lý dữ liệu

3.4.1. Tiền xử lý dữ liệu

- Xử lý dữ liệu null:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 4   Education        2000 non-null   int64
 5   Income           2000 non-null   int64
 6   Occupation       2000 non-null   int64
 7   Settlement size  2000 non-null   int64
dtypes: int64(8)
memory usage: 125.1 KB
```

Ở đây, nhìn chung thì bộ dữ liệu ta sẽ không có giá trị là null, cho nên ta không cần phải xử lý đầu vào tại đây.

- Xử lý các dữ liệu duplicate:

```
rows_before = len(df)
df.drop_duplicates(inplace=True)
rows_after = len(df)

print(f"Number of duplicate rows: {rows_before - rows_after}")
```

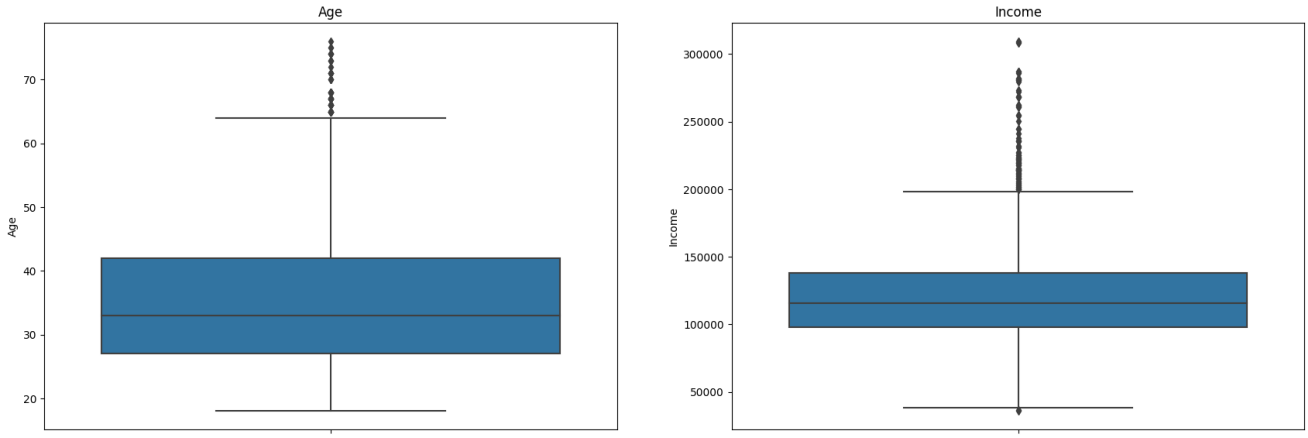
```
Number of duplicate rows: 0
```

Cũng như việc xử lý giá trị null, thì ở đây cung không tồn tại các dòng bị trùng lặp. Cho nên ta là khi ta thực hiện xóa các giá trị bị trùng lặp thì không hề thấy có dòng nào bị trùng để mà xóa.

- Xử lý các giá trị ngoại lai:

Về việc xử lý các giá trị nhiễu, ta sẽ chú trọng tới 2 đặc trưng chính để xử lý đó là "Age" và "Income".

Đầu tiên là trực quan hóa biểu đồ boxplot lên để xem xét:



Tiếp theo là tính các giá trị ngoại lai của 2 đặc trưng:

```
# find outlier in all columns
for i in df.select_dtypes(include=['float64','int64']).columns:
  max_thresold=df[i].quantile(0.95)
  min_thresold=df[i].quantile(0.05)
  cus_info=df[(df[i] < max_thresold) & (df[i] > min_thresold)].shape
  print(" outlier in ",i,"is" ,int(((df.shape[0]-cus_info[0])/df.shape[0])*100),"%")
```

```
 outlier in  Age is 11 %
 outlier in  Income is 10 %
```

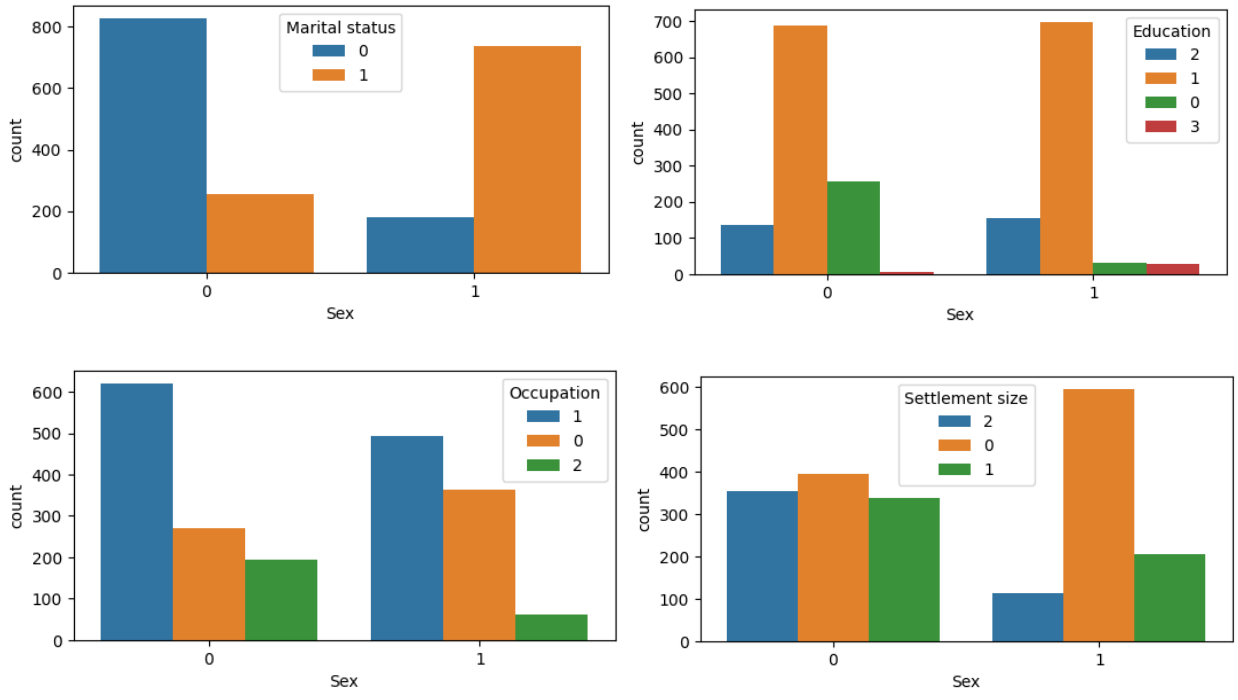
Có thể thấy khi trực quan lên ta sẽ thấy 1 số dữ liệu gây nhiễu: Ở đặc trưng “Age” sẽ có khoảng 11% giá trị ngoại lai và ở đặc trưng “Income” thì sẽ có khoảng 10% giá trị ngoại lai. Tuy nhiên, ta có thể phân tích được rằng:

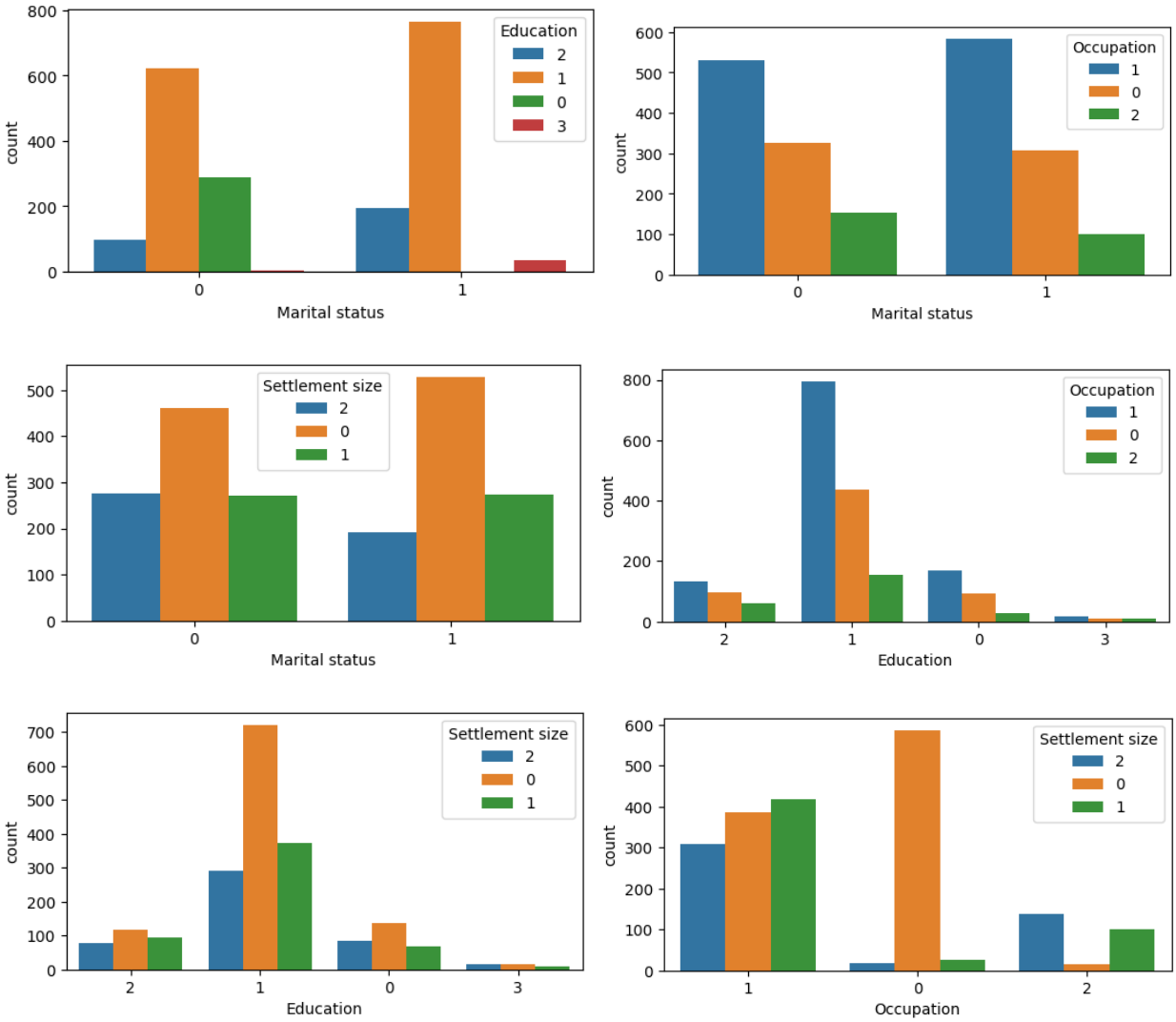
- Với dữ liệu của đặc trưng "Age", thì với các dạng trung tâm thương mại mà có những khách hàng có số tuổi lớn hơn các cận giới hạn của boxplot có thể là các ngoại lai đúng, vì các khách hàng vẫn sẽ có số tuổi cao hơn cận giới hạn đến trung tâm thương mại.
- Với dữ liệu của đặc trưng "Income", thì các ngoại lai này vẫn có thể đúng tại 1 số người có thu nhập thuộc dạng cao thì đối với cận giới hạn trên như vậy là chưa đủ, ngoài ra cũng sẽ có 1 số khách hàng thuộc dạng khó khăn cho nên các giá trị nằm ngoài cận giới hạn dưới vẫn có thể đúng. Từ đó có thể đúc kết được lại đối với các giá trị ngoại lai này, sau khi ta phân tích thì không thể loại bỏ được.

Thế nên, việc loại bỏ các ngoại lai này là không cần thiết. Vì có thể khi bỏ, nó sẽ không hợp lý cho bộ dữ liệu.

3.4.2. Trực quan hóa dữ liệu

- Ta sẽ có các hình như sau:





Ở đây, sau khi trực quan hóa các thuộc tính thuộc dạng phân loại thì ta sẽ có 1 số phân tích sơ bộ như sau:

+ Phụ nữ thì thường sẽ có xu hướng là đã kết hôn nhiều hơn so với đàn ông.

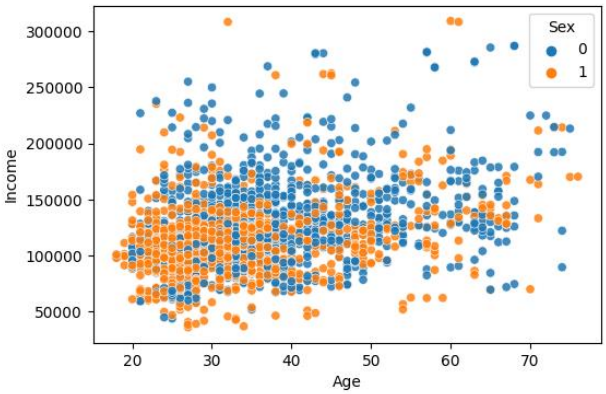
+ Với những người đã kết hôn thì thường sẽ sống ở các thành phố nhỏ.

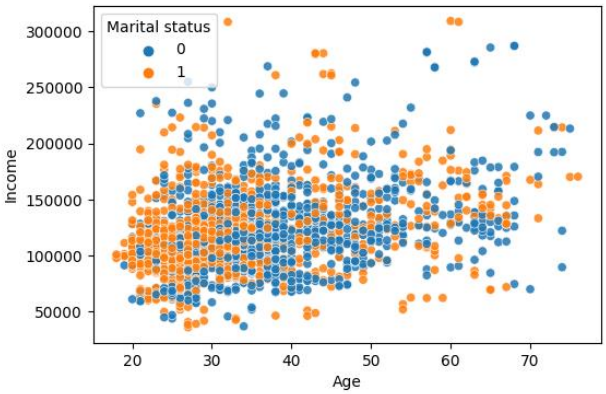
+ Những người với Occupation = ‘0’ (nghĩa là thất nghiệp) thì sẽ có xu hướng sống ở thành phố nhỏ.

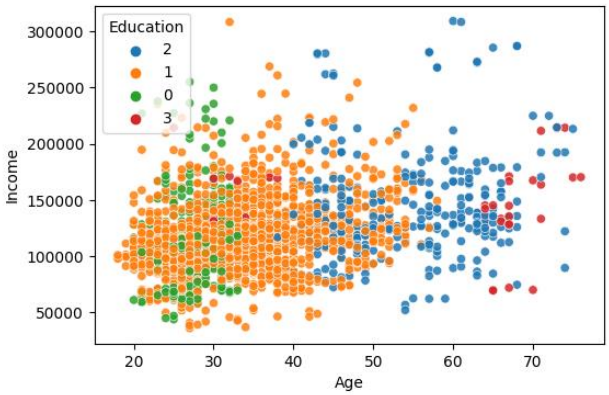
+ So về tỉ lệ có việc làm thì đàn ông thì sẽ có tỉ lệ việc làm cao hơn so với phụ nữ.

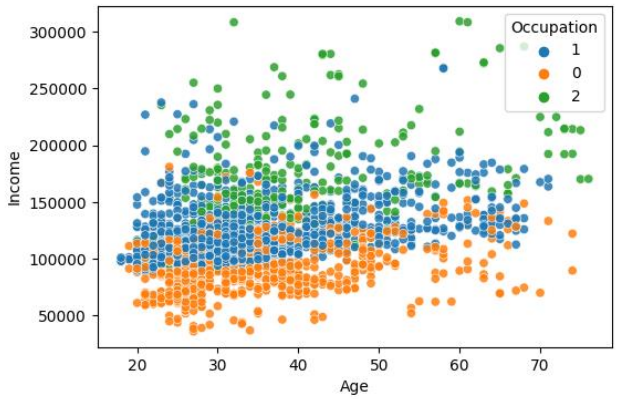
+ Tỉ lệ những người thất nghiệp sống ở thành phố nhỏ là rất cao.

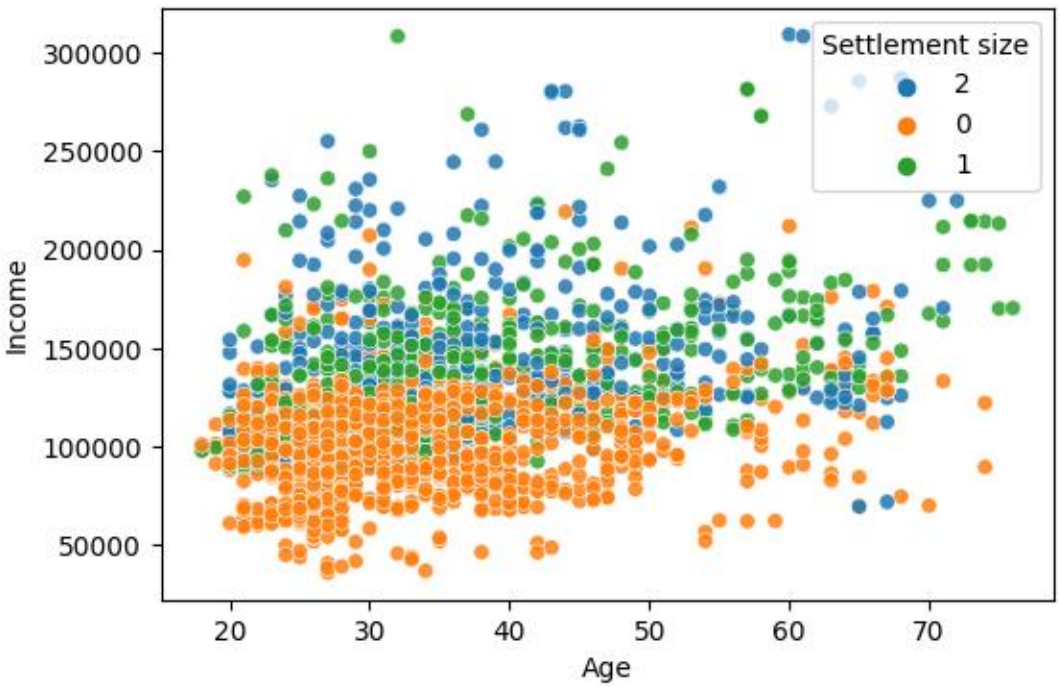
Ngoài ra, ta sẽ sử dụng biểu đồ phân tán để có thể tìm ra các đặc trưng categorical sẽ có mối liên quan tới các đặc trưng numberical như thế nào.











Ở đây, ta sẽ có 1 vài đánh giá như sau:

- Đối với trình độ học vấn, ta thấy được rằng đối với những người có trình độ học vấn cao, họ sẽ là những người có xu hướng có tuổi đời cao. Còn đối với những người trẻ thì sẽ tập trung chủ yếu ở độ tuổi từ 20 cho tới 50 tuổi.
- Ngoài ra, đối với trình độ học vấn ta còn có thể nhận thấy 1 điều rằng, với những người có trình độ học vấn cao, ta sẽ thấy rằng trình độ học vấn càng cao thì thu nhập cũng sẽ có chiều hướng cao theo.
- Đối với những người có công việc, ta sẽ thấy thu nhập được phân 1 cách rõ ràng, với những người là quản lý hoặc là những người doanh nhân thì thu nhập sẽ cao hơn đối với những người là nhân viên, và sẽ con hơn hẳn đối với những người đang thất nghiệp.

3.5. Huấn luyện mô hình

3.5.1. Phân phối chuẩn

Với 2 đặc trưng thuộc dạng "numberical" phía trên, hiện tại chúng đang bị lệch phải khá nhiều. Cho nên, ta sẽ tiến hành thử nghiệm đưa chúng về dạng ổn định hơn đó là đưa chúng về phân phối chuẩn để khi huấn luyện mô hình sẽ cho ra được kết quả khả quan hơn.

+ Đầu tiên sẽ là kiểm tra p-value của 2 đặc trưng.

```
normaltest_result_income = stats.normaltest(df['Income'])[1]
normaltest_result_age    = stats.normaltest(df['Age'])[1]

print(f'The p-value for the null hypothesis of the Income being Normally distributed is {normaltest_result_income}')
print(f'The p-value for the null hypothesis of the Age    being Normally distributed is {normaltest_result_age}')
```

```
The p-value for the null hypothesis of the Income being Normally distributed is 2.5009638879187734e-98
The p-value for the null hypothesis of the Age    being Normally distributed is 3.3428344869999675e-56
```

Sơ lược về hàm stats.normaltest: là một hàm trong module scipy.stats được sử dụng để kiểm tra tính chuẩn đoán của một mẫu dữ liệu. Điều này giúp xác định xem mẫu dữ liệu có tuân theo phân phối chuẩn (normal distribution) hay không.

Nếu p-value < mức ý nghĩa (thường là 0.05), ta có thể nói rằng dữ liệu hiện tại là không thuộc dạng phân phối chuẩn.

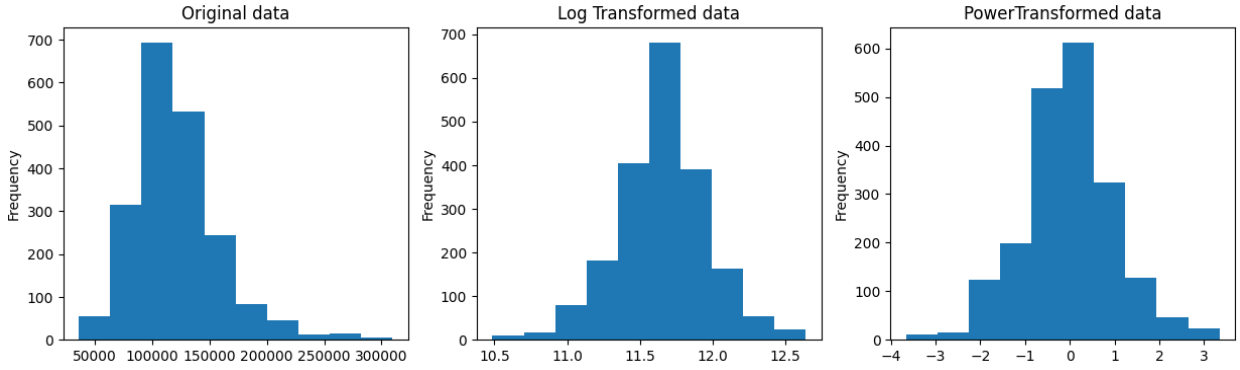
Ngược lại, nếu p-value >= mức ý nghĩa, thì ta có thể nói rằng dữ liệu đã có thể là dạng phân phối chuẩn.

Cho nên từ các kết luận trên ta có thể thấy được rằng:

Giá trị p-value ở đây lần lượt là: p-value(Income) < p-value(Age) < 0,05. Từ đây ta có thể khẳng định rằng là cả 2 đặc trưng đều chưa phải là dạng phân phối chuẩn, và ta sẽ phải đưa ra quyết định chuẩn hóa dữ liệu để đưa dữ liệu về dạng phân phối chuẩn.

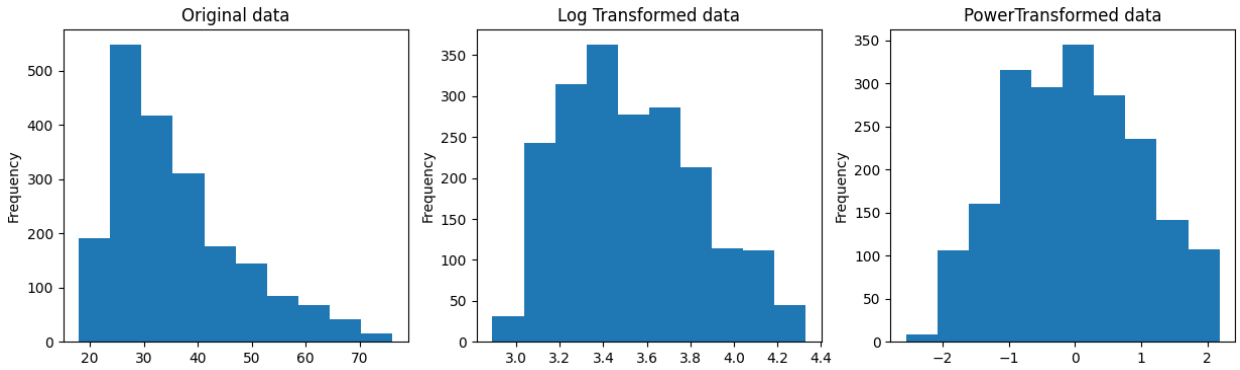
Bây giờ ta sẽ tiến hành đưa 2 đặc trưng về dạng phân phối chuẩn.

+ Đặc trưng “Income”:



	statistic	pvalue
Original data	449.473326	2.500964e-98
Log transform	32.357037	9.413664e-08
PowerTransformer	27.859212	8.921730e-07

+ Đặc trưng “Age”:



	statistic	pvalue
Original data	255.475892	3.342834e-56
Log transform	111.094201	7.519703e-25
PowerTransformer	161.196197	9.924088e-36

Sau khi ta đã thực hiện đưa dữ liệu của cả 2 đặc trưng về dạng phân phối chuẩn thì có thể thấy rằng dữ liệu sau khi áp dụng log transfermation hoặc là power transformation thì cung không thể đưa dữ liệu về dạng phân phối chuẩn. Mặc dù vẫn chưa là dạng phân phối chuẩn nhưng ta có thể thấy rằng nó đã được cải thiện 1 cách đáng kể so với dữ liệu đầu vào trước khi áp dụng phương pháp đưa về phân phối chuẩn. Chính vì vậy, ta có thể sử dụng dữ liệu đã được chuyển đổi này để thay thế cho bộ dữ liệu gốc.

Ở đây ta sẽ tiến hành drop cả 2 đặc trưng "Age" và "Income" của bộ dữ liệu gốc đi và thay thế bằng 2 đặc trưng "Age" và "Income" đã đuược chuyển đổi.

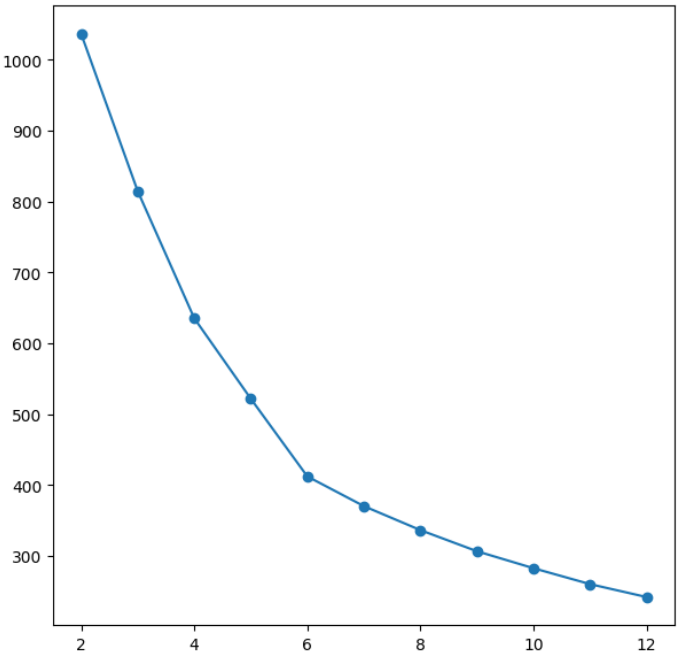
Bộ dữ liệu của ta sẽ có dạng như thế này:

	Sex	Marital status	Education	Occupation	Settlement size	transf_income	transf_age
0	0	0	2	1	2	0.239545	4.204693
1	1	1	1	1	2	0.869781	3.091042
2	0	0	1	0	0	-0.846121	3.891820
3	0	0	1	1	1	1.303766	3.806662
4	0	0	1	1	1	0.830970	3.970292
...	...	...	...	...	...	...	...
1995	1	0	1	0	0	0.209208	3.850148
1996	1	1	1	1	0	0.051991	3.295837
1997	0	0	0	0	0	-0.948374	3.433987
1998	1	1	1	0	0	-0.545364	3.178054
1999	0	0	0	0	0	-1.685848	3.218876

2000 rows × 7 columns

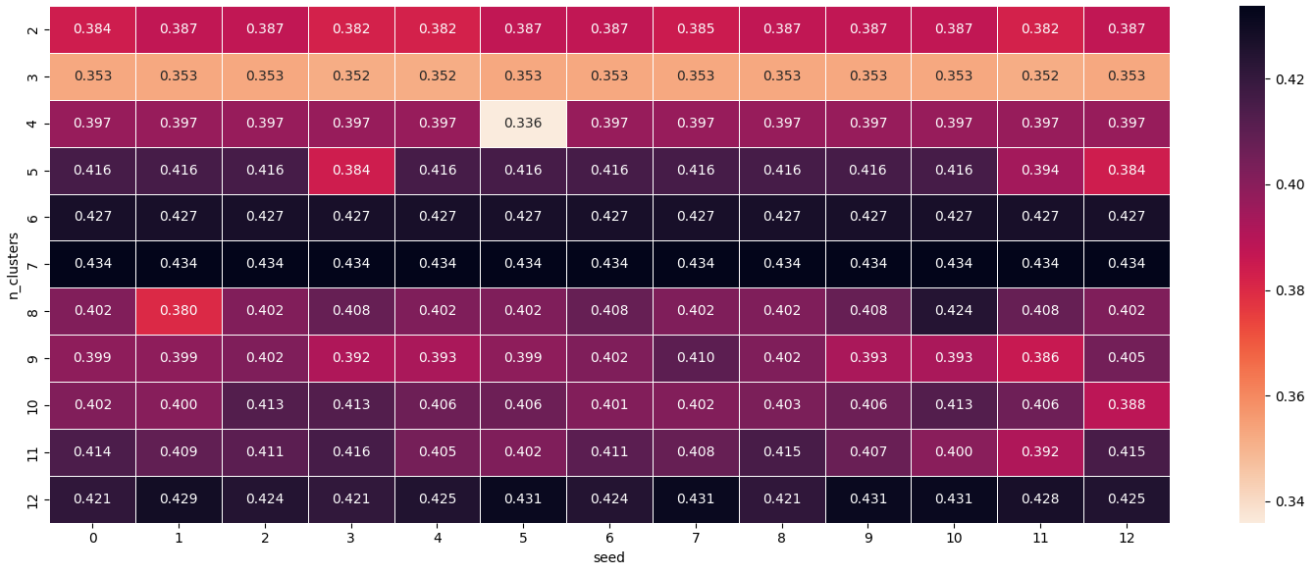
3.5.2. Phương pháp Elbow

Trước khi ta tiến hành áp dụng phương pháp Elbow, ta sẽ cần chuẩn hóa dữ liệu trước. Sau đó ta sẽ áp dụng Elbow và sẽ có được kết quả như sau:



Ở đây ta có thể thấy "khuỷu tay" sẽ nằm ở cụm 6, có nghĩa là ta sẽ nên chọn 6 cụm cho việc phân cụm K-Means. Tuy nhiên để chắc chắn hơn về việc lựa chọn, ta sẽ sử dụng silhouette score để đánh giá.

3.5.3. Silhouette Score



Ở đây, ta sẽ kết hợp giữa các cụm và các seed để có thể đưa ra được các điểm số ổn định sau mỗi lần đánh giá. Tại đây, có thể thấy ở cụm 6 và 7 sẽ có được điểm số khá ổn

định, phù hợp với số cụm Elbow đã hiển thị phía trên, các cụm tại đây sẽ có sự hơi tách biệt. Cho nên ta sẽ lựa chọn 6 hoặc 7 cụm để áp dụng cho phân cụm K-Means.

Ta sẽ chọn 6 cụm là số cụm sẽ truyền vào K-Means, thuật toán sẽ tiến hành các bước lặp để cho ra được các cụm với các dữ liệu tương ứng.

3.5.4. PCA (Principal Component Analysis)

Trước khi bắt đầu áp dụng thuật toán K-Means, ta sẽ điều chỉnh một vài chi tiết như sau:

```
[35] pca = PCA(n_components=3, random_state=42)
     X_pca = pca.fit_transform(X)

[36] X_pca_df = pd.DataFrame(data=X_pca, columns=['X1', 'X2', 'X3'])

[37] kmeans=KMeans(n_clusters=6, random_state=0).fit(X)
```

Ta sẽ tiến hành áp dụng từng dòng code như trên, nghĩa là ta sẽ chuyển đổi bộ dữ liệu hiện tại với 7 đặc trưng về 3 đặc trưng như ở dòng PCA, sau đó ta sẽ có các cột lần lượt là X1, X2, X3. Và cuối cùng là ta sẽ đưa vào K-Means và bắt đầu phân cụm với số cụm đầu vào là 6.

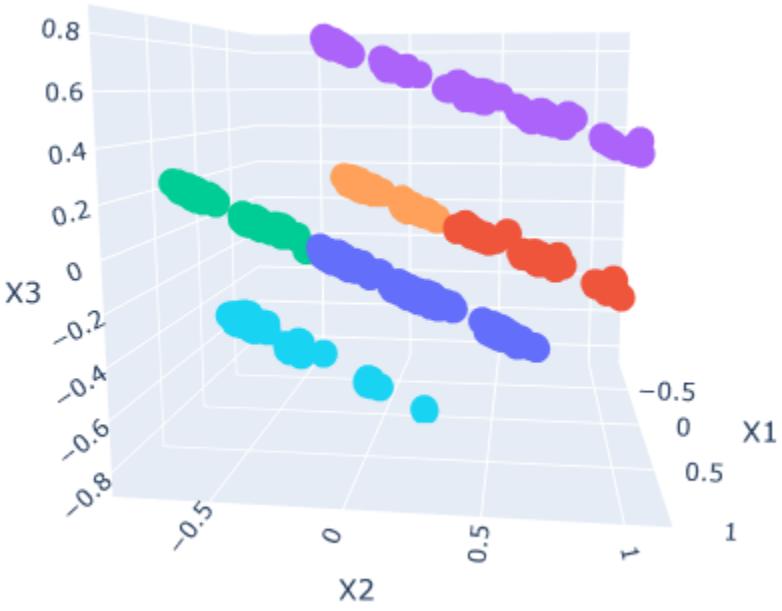
Sau khi đã phân cụm xong thì ta sẽ có một bộ dataframe như sau:

```
X_pca_df.head()
```

	X1	X2	X3	Labels
0	0.829143	0.350518	-0.162403	3
1	-0.531854	0.719504	-0.230299	2
2	0.489493	-0.693627	0.175269	0
3	0.712487	-0.027906	-0.049583	3
4	0.718334	-0.036815	-0.052850	3

3.5.5. Đánh giá kết quả phân cụm

Và kết quả phân cụm khi ta trực quan hóa nó lên sẽ như thế này:



Ta có thể quan sát thấy được cả 3 đặc trưng với 6 cụm có sự tách biệt rõ ràng với nhau. Bây giờ ta sẽ bắt đầu kết hợp các đặc trưng khác với PCA ta sẽ cho ra được bộ dataframe hoàn chỉnh với đầy đủ các đặc trưng gồm 7 đặc trưng ban đầu và sẽ có thêm 1 đặc trưng nữa là nhóm. Ta sẽ có bộ dataframe như sau:

```
results_df.head(10)
```

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	Labels
0	0	0	67	2	124670	1	2	3
1	1	1	22	1	150773	1	2	2
2	0	0	49	1	89210	0	0	0
3	0	0	45	1	171565	1	1	3
4	0	0	53	1	149031	1	1	3
5	0	0	35	1	144848	0	0	0
6	0	0	53	1	156495	1	1	3
7	0	0	35	1	193621	2	1	3
8	0	1	61	2	151591	0	0	4
9	0	1	28	1	174646	2	0	4

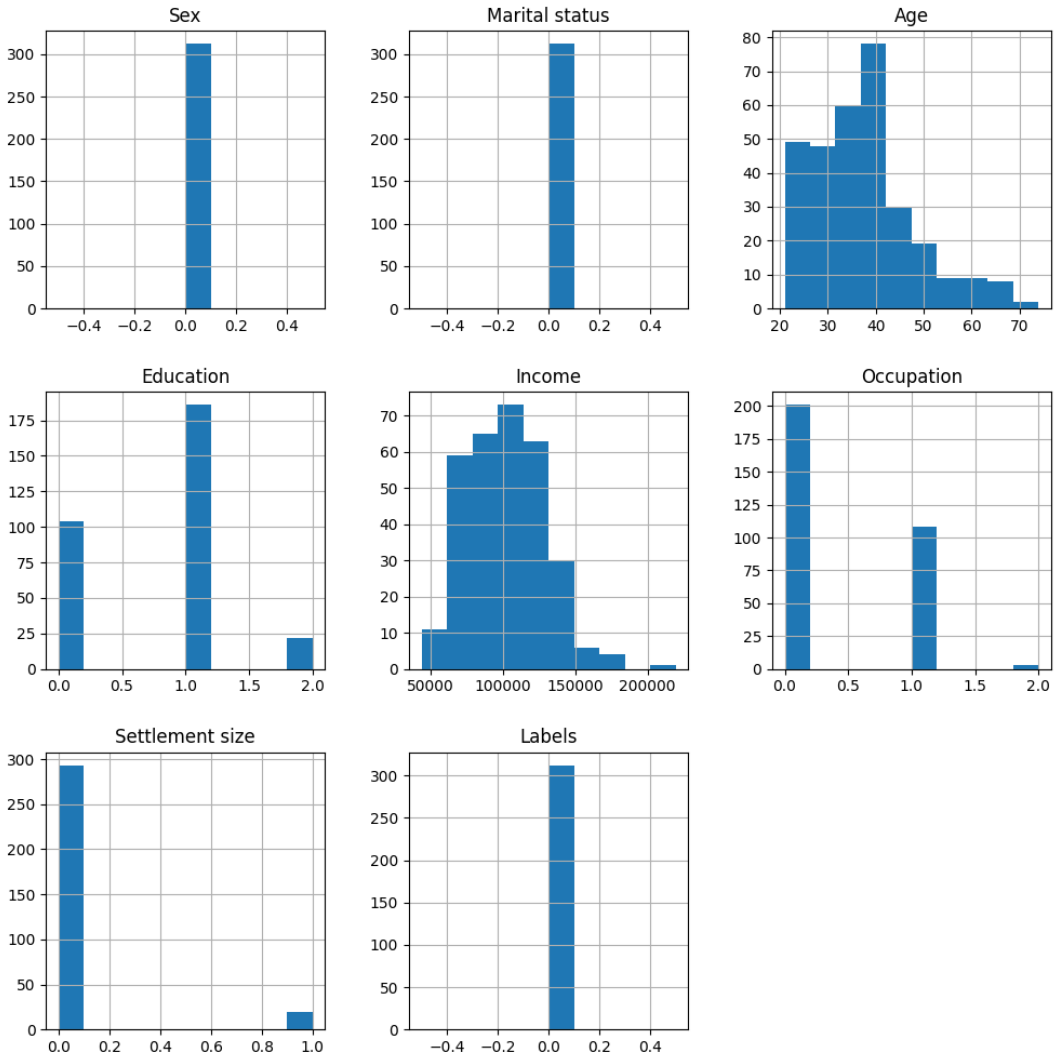
Và bước tiếp theo và cũng là bước cuối cùng đó là ta tiến hành phân tích dữ liệu của từng cụm.

3.6. Giải thích kết quả phân cụm

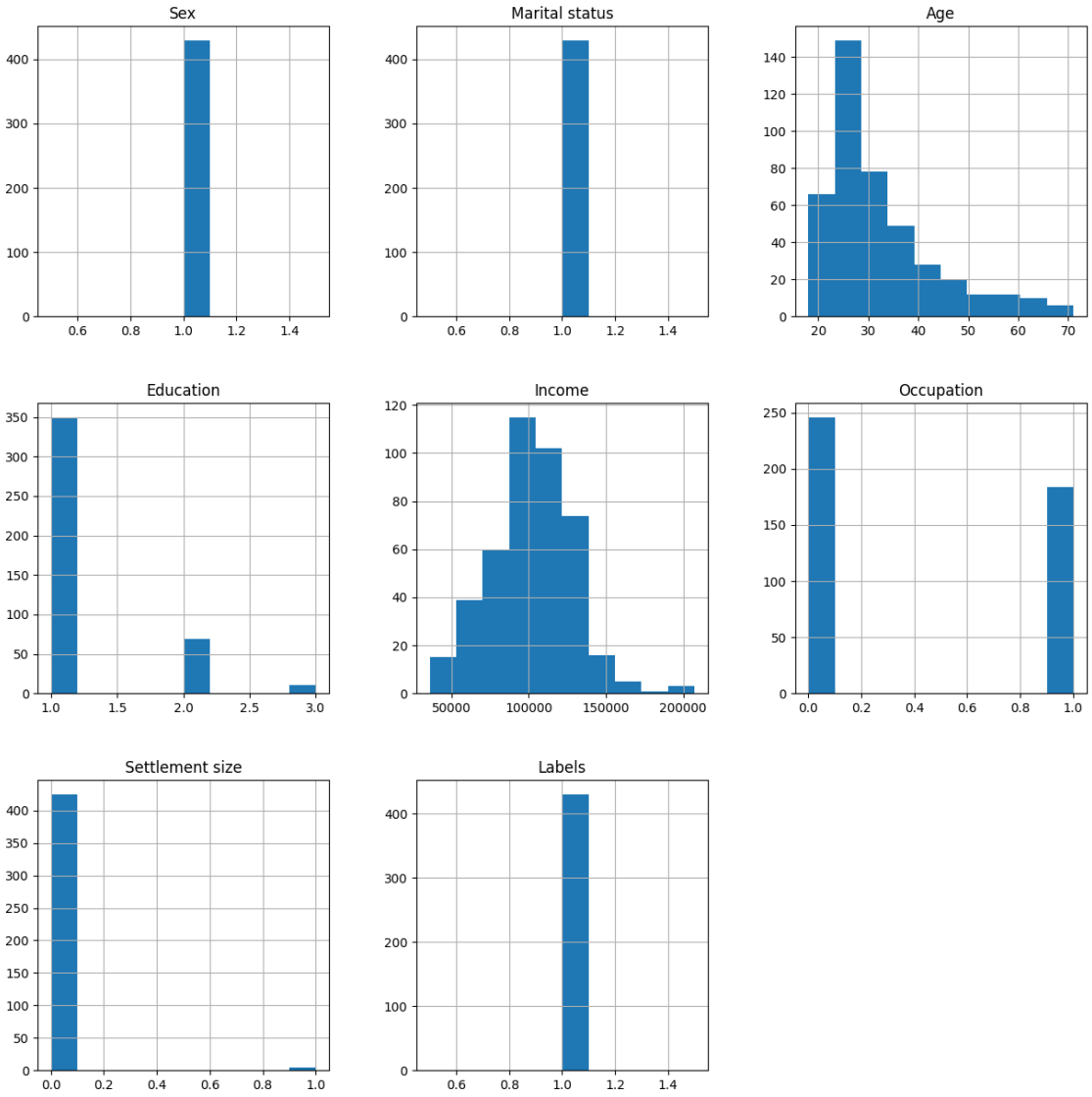
Đầu tiên ta sẽ có sử dụng biểu đồ cột để trực quan hóa toàn bộ các đặc trưng của cụm trước, rồi ta sẽ có 1 số nhận xét đối với các cụm và cuối cùng sẽ là nhận xét chung đối với toàn bộ cụm.

Ở đây, ta sẽ có lần lượt là 6 cụm với các biểu đồ cột của các đặc trưng như sau:

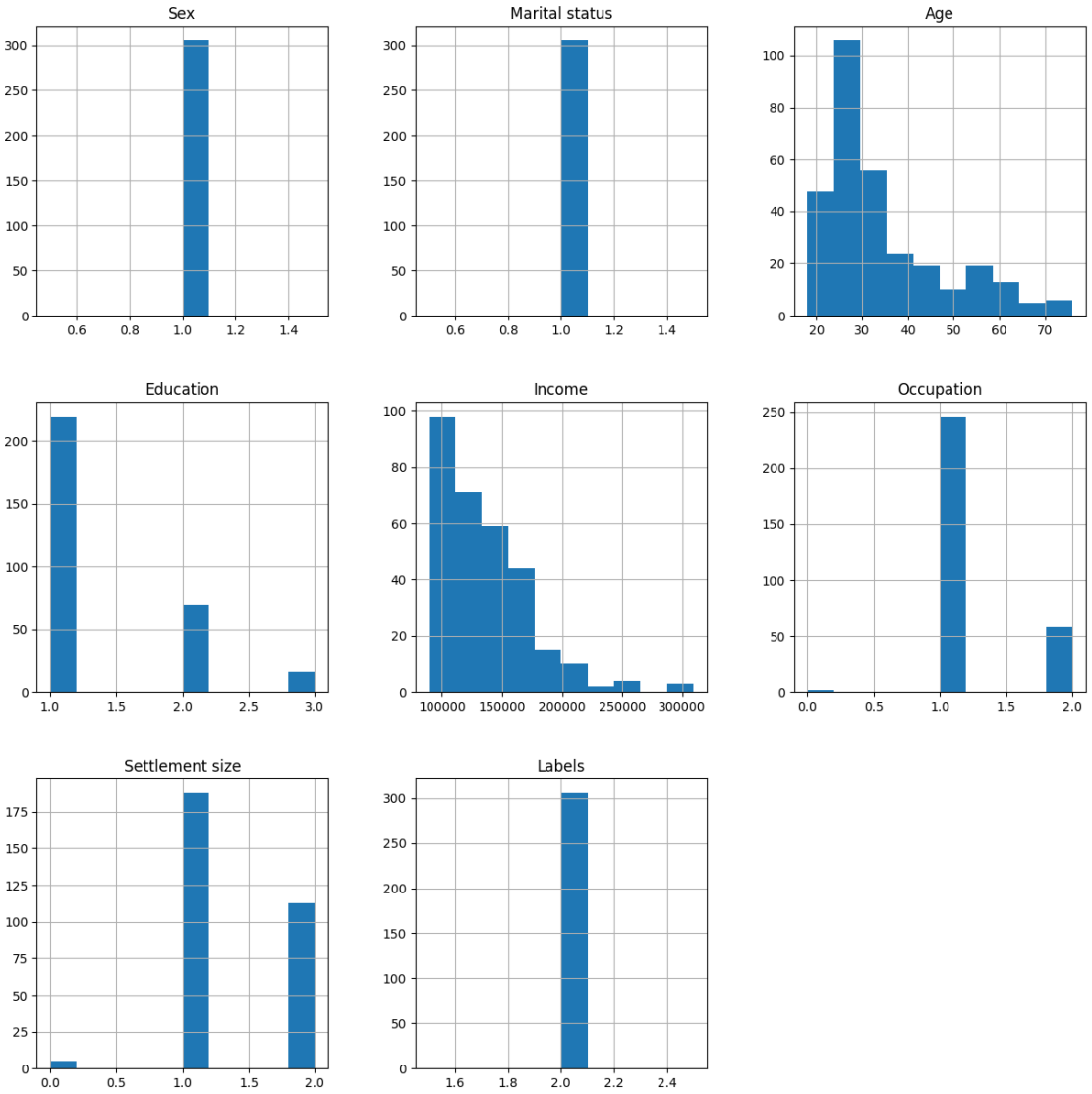
+ Cụm 0:



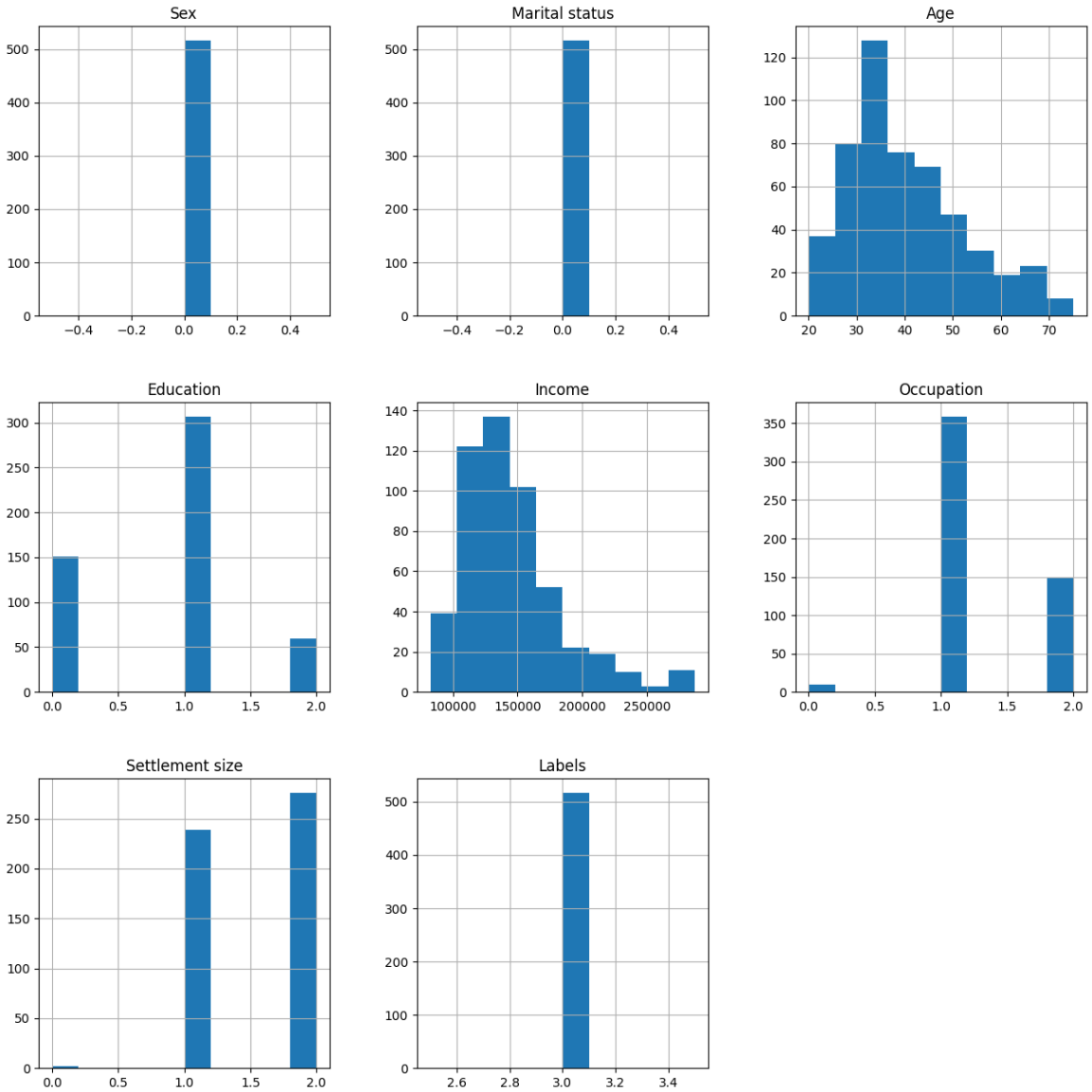
+ Cụm 1:


Sex
Marital status
Age
Education
Income
Occupation
Settlement size
Labels

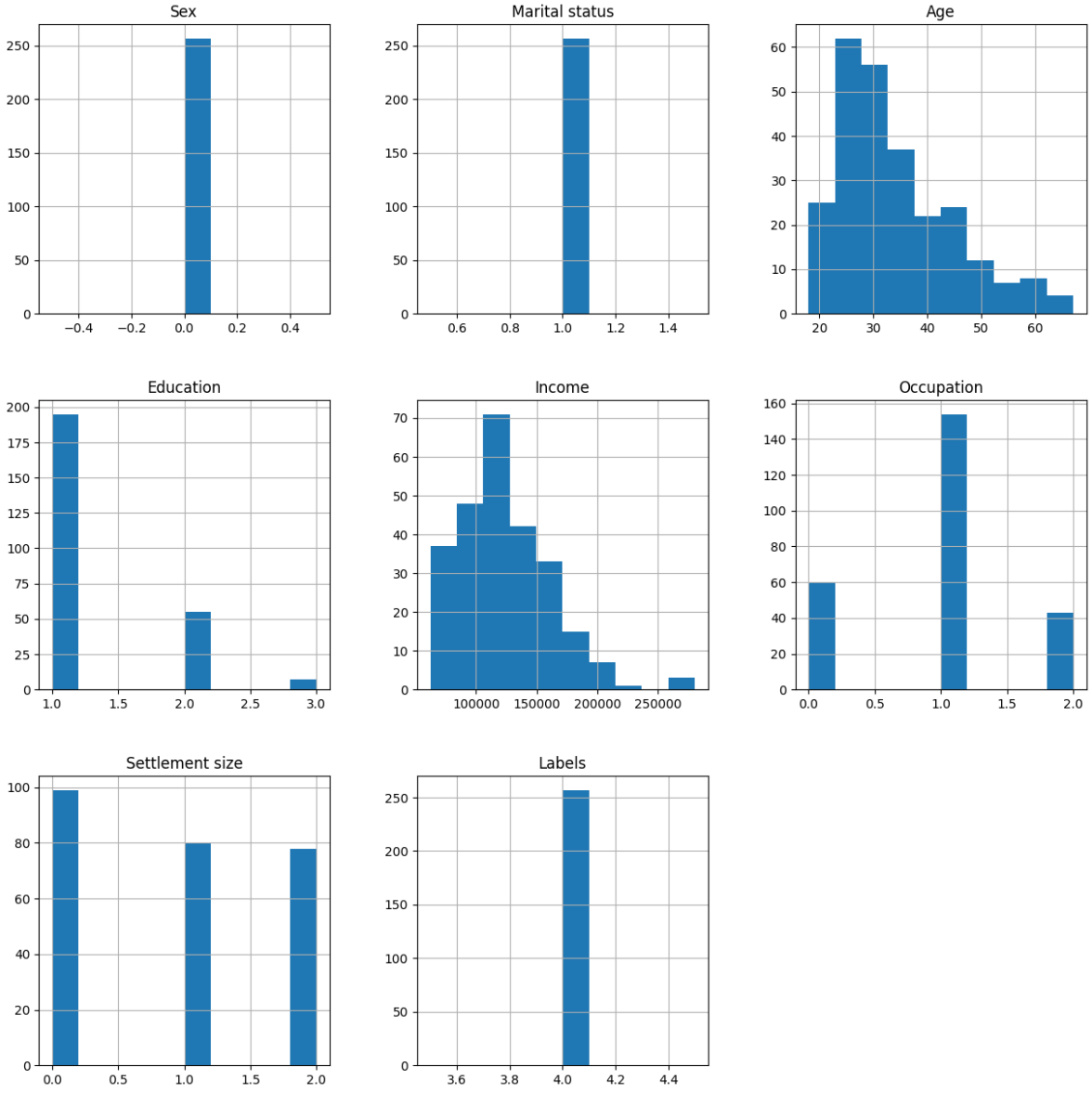
+ Cụm 2:


Sex
Marital status
Age
Education
Income
Occupation
Settlement size
Labels

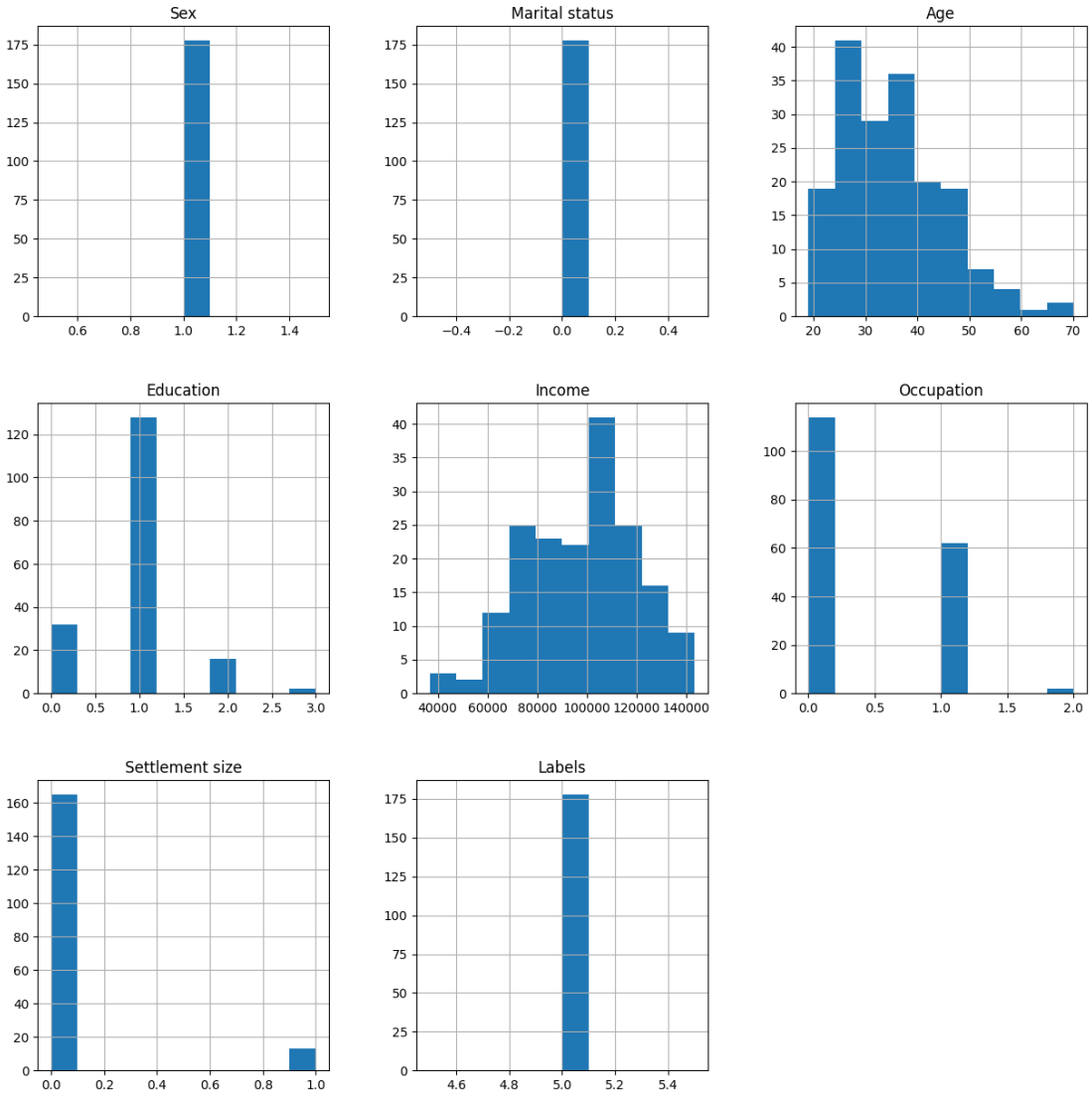
+ Cụm 3:


Sex
Marital status
Age
Education
Income
Occupation
Settlement size
Labels

+ Cụm 4:


Sex
Marital status
Age
Education
Income
Occupation
Settlement size
Labels

+ Cụm 5:



Ta sẽ có 1 số nhận xét như sau:

+ Cụm 0: Nhóm này là gồm những người đàn ông ở độ tuổi từ 20 tới 40 tuổi, họ thường là những người thất nghiệp và sống ở các thành phố nhỏ, trình độ học vấn của họ sẽ nằm ở mức trung học phổ thông hoặc là thấp hơn. Mức lương hàng năm của họ rơi vào khoảng độ từ 65000 cho tới 160000 và có xu hướng là những người chưa kết hôn.

+ Cụm 1: Nhóm này là gồm những người phụ nữ ở độ tuổi 20 cho tới 30, họ có trình độ học vấn ở mức tốt nghiệp trung học phổ thông và thu nhập của họ ở dạng trung bình. Họ có thể là những người thất nghiệp hoặc cũng có thể là những nhân viên bình thường. Ngoài ra sẽ chủ yếu sống ở các thành phố nhỏ.

+ Cụm 2: Nhóm này là gồm những người phụ nữ đã kết hôn, độ tuổi sẽ là khoảng từ 20 cho tới 30 là chủ yếu. Trình độ học vấn sẽ có xu hướng ở mức 1 nghĩa là học thường là những người đã tốt nghiệp trung học phổ thông và chỉ có một số ít là học cao và có trình độ sau đại học. Ngoài ra, Nghề nghiệp sẽ là nhân viên là chủ yếu, một số ít sẽ có công việc là quản lý, mức thu nhập của họ sẽ phần lớn nằm trong khoảng từ 1000000 cho tới 160000 và họ sẽ sống ở các thành phố vừa và lớn.

+ Cụm 3: Nhóm này là gồm những người đàn ông chưa kết hôn, độ tuổi sẽ là rơi vào khoảng từ 26 cho tới 48. Trình độ học vấn của họ sẽ nằm ở mức tốt nghiệp trung học phổ thông và thấp hơn là chủ yếu, tuy nhiên vẫn có một nhóm nằm ở mức trình độ đại học. Họ sẽ là những người có xu hướng công việc là nhân viên bình thường và bên cạnh đó cũng có một số lượng với công việc là làm quản lý. Thu nhập của họ sẽ ở mức trung bình là từ 100000 cho tới 160000, họ sẽ sống ở thành phố lớn và thành phố tầm trung là chủ yếu.

+ Cụm 4: Nhóm này là gồm những người đàn ông đã kết hôn, độ tuổi của họ sẽ nằm trung bình ở khoảng từ 23 cho tới 38. Trình độ học vấn của họ sẽ từ trung học phổ thông trở lên, họ có thể là những người nhân viên bình thường là chủ yếu, cũng có thể họ là những người quản lý hoặc thậm chí họ cũng có thể là những người thất nghiệp. Những khách hàng này sống rải đều ở cả 3 dạng thành phố cho nên không thể xác định được điều gì.

+ Cụm 5: Nhóm này là đại diện cho những người phụ nữ chưa kết hôn, họ sẽ là nhưng người có trình độ học vấn thấp là chủ yếu. Phần lớn họ là những người chỉ tốt nghiệp trung học phổ thông. Họ sống ở những thành phố nhỏ và là những người thất nghiệp hoặc là những nhân viên văn phòng bình thường.

Để có cái nhìn tổng quan hơn về các đặc trưng đối với các cụm thì ta sẽ dùng tâm của các cụm để có thể dễ dàng phân tích hơn. Và tâm của các cụm sẽ được trình bày như sau:

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
C0	3.885781e-16	1.332268e-15	0.245726	0.182692	0.030449	0.451144	0.484804
C1	1.000000e+00	1.000000e+00	0.403876	0.213953	0.005814	0.448891	0.375836
C2	1.000000e+00	1.000000e+00	0.444444	0.591503	0.676471	0.585119	0.405291
C3	2.553513e-15	1.776357e-15	0.274017	0.633462	0.764990	0.615822	0.524984
C4	-3.885781e-16	1.000000e+00	0.422827	0.466926	0.459144	0.530253	0.409658
C5	1.000000e+00	1.498801e-15	0.310861	0.185393	0.036517	0.433188	0.439125

+ Nhận xét chung: Sau khi nhìn sơ lược qua về 6 cụm ta sẽ thấy được rằng, bộ dữ liệu với các đặc trưng được phân khá sát với thực tế, đối với những người có trình độ học vấn cao thì họ sẽ là những người có thu nhập cao cũng như là sẽ sống tập trung chủ yếu ở các thành phố vừa và nhỏ. Những người có trình độ học vấn thấp hơn thì ngược lại, họ sẽ có thu nhập ít hơn và sẽ có xu hướng sống ở các thành phố nhỏ. Nếu xét theo các biểu đồ thì với đặc trưng "Age" thì những người ở cụm 1 và 2 sẽ có tuổi đời thường trẻ hơn so với những người ở cụm 0 và 3. Còn với đặc trưng "Income" thì theo bảng phân cụm thì cụm 2 3 và 4 là những cụm sẽ có mức thu nhập cao hơn hẳn so với cụm 0, 1 và 5, điều này đồng nghĩa là trình độ học vấn của cụm 2,3 và 4 sẽ cao hơn so với cụm 0, 1 và 5 như ta đã phân tích phía trên.

TÀI LIỆU THAM KHẢO

[1] N. V. Hiếu, "Luyện Code," 11 8 2018. [Online]. Available: https://blog.luyencode.net/thuat-toan-phan-cum-k-means/#gioi-thieu-ve-k-means. [Accessed 12 12 2023].

[2] N. c. Thắng, "Mì AI," 22 4 2021. [Online]. Available: https://www.miai.vn/2021/04/22/principal-component-analysis-pca-tuyet-chieu-giam-chieu-du-lieu/. [Accessed 18 12 2023].

[3] fixexp, "fixexpo," 26 5 2023. [Online]. Available: https://fixexpo.org/silhouette-la-gi/. [Accessed 19 12 2023].