Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM

Khoa Công nghệ Thông tin

---o0o---

BÁO CÁO PROJECT 3
CLUSTERING

Môn: Khai thác dữ liệu và ứng dụng

Giảng viên hướng dẫn:

- **Lê Ngọc Thành**
- **Nguyễn Thái Vũ**

Nhóm 7:

- **19127326 - Vương Thành An**
- **19127281 - Trần Minh Thiện**
- **19127330 - Lê Tâm Anh**

MỤC LỤC

1/ Giới thiệu về nhóm và đánh giá kết quả:

Thông tin nhóm:

Nhóm 7			
STT	**MSSV**	**Họ và tên**	**Tỉ lệ công việc**
1	19127281	Trần Minh Thiện	34%
2	19127326	Vương Thành An	33%
3	19127330	Lê Tâm Anh	33%

Mức độ hoàn thành đồ án:

Yêu cầu	Hoàn thành
Khám phá dữ liệu	100%
Tiền xử lý và chuẩn bị dữ liệu	100%
Áp dụng thuật toán phân nhóm	100%
Trực quan kết quả phân nhóm	100%
Mức độ hoàn thành đồ án là 100%	

Mức độ hoàn thành công việc của các thành viên:

STT	Họ và tên	Hoàn thành
1	Trần Minh Thiện	100%
2	Vương Thành An	100%
3	Lê Tâm Anh	100%

2/ Mô tả đồ án và tập dữ liệu sử dụng:

- **Mô tả đồ án:** Trong đồ án này ta sẽ áp dụng các thuật toán phân nhóm, cụ thể ở đây ta sẽ sử dụng thuật toán k-means để phân nhóm cho các nhóm khách hàng dựa vào các thuộc tính được cung cấp trong tập dữ liệu để giúp cho các nhân viên có cái nhìn tốt hơn về các nhóm khách hàng cũng như giúp họ quản lý khách hàng hiệu quả hơn.

- **Tập dữ liệu sử dụng:**
  - Tập dữ liệu được sử dụng là Mall_Customers.csv được lấy từ kaggle qua link sau: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

- **Mô tả tập dữ liệu:** Tập dữ liệu gồm 200 dòng và 5 cột, trong đó không có dòng nào bị trùng. Các thuộc tính của năm cột:
  - CustomerID: mã số ID duy nhất của mỗi khách hàng.
  - Gender: Giới tính của khách hàng
  - Age: Độ tuổi của khách hàng
  - Annual Income(k$): thể hiện thu nhập hàng năm của khách hàng. (Đơn vị nghìn USD)
  - Spending Score(1-100): điểm tiêm dùng, được chấm dựa theo chi tiêu trong năm của khách hàng. (Đơn vị từ 1-100)

3/ Khám phá dữ liệu:

3.1/ Kiểu dữ liệu của các cột:

- In ra kiểu dữ liệu của các cột:

```
CustomerID                int64
Gender                   object
Age                       int64
Annual Income (k$)        int64
Spending Score (1-100)    int64
dtype: object
```

- 4 cột: CustomerId, Age, Annual Income và Spending Score có dạng số nguyên.

- Cột Gender có dạng object, sau khi tìm hiểu thì có kiểu dữ liệu thực là string với 2 giá trị chỉ giới tính là Male và Female.

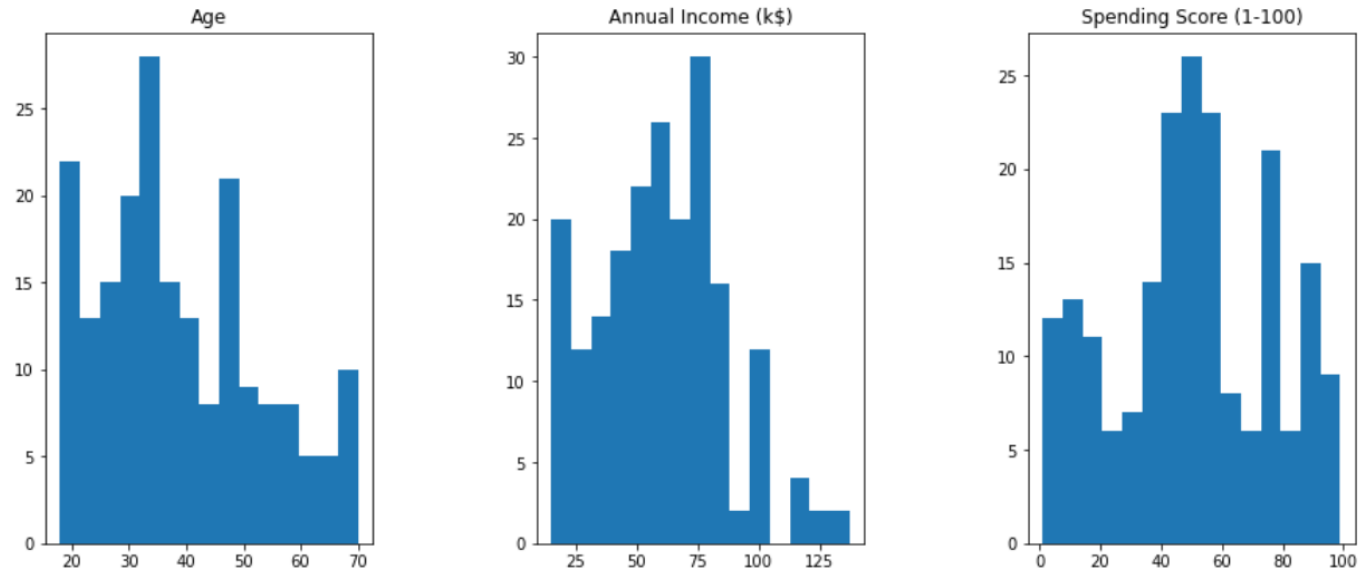
3.2/ Mô tả thống kê chi tiết:

- Bảng mô tả thống kê tổng quát của các cột dạng số (cột gender sẽ được giải thích sau):

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

- Nhìn chung tập dữ liệu không có gì bất thường, cũng không bị thiếu dữ liệu khi số lượng các cột đều bằng 200. Qua bảng này ta cũng biết được các giá trị max, min của các cột.

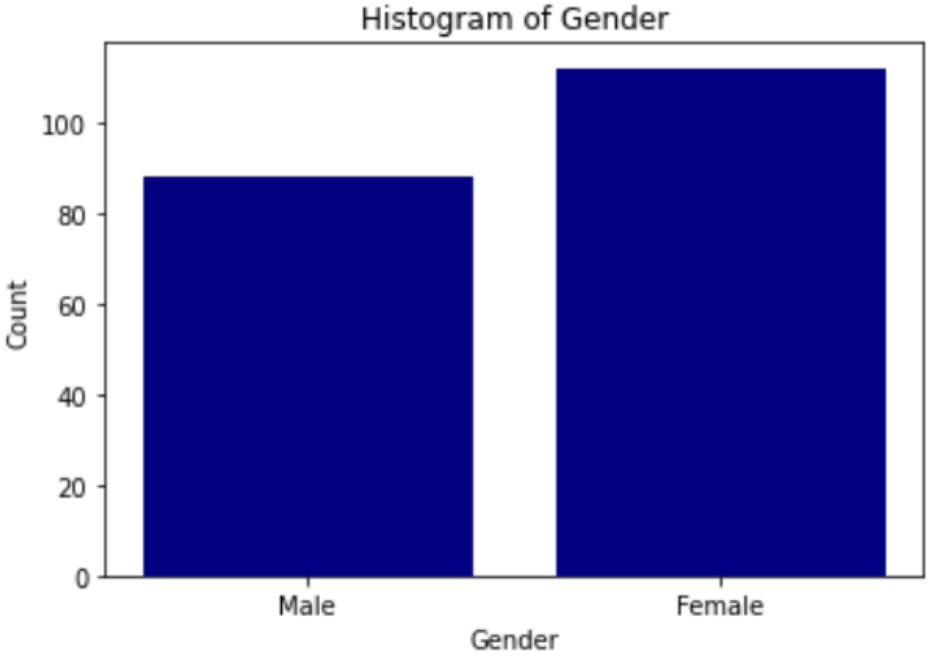
- Trực quan các biến phân bố các biến numeric bằng biểu đồ histogram:



- Nhận xét cho các biến numeric:
    - Cột Age:
        - ✓ Dữ liệu tập trung nhiều nhất ở khoảng từ 30 đến 40. + Các khoảng còn lại đều có số lượng điểm dữ liệu ở mức tương đối.
        - ✓ Không tìm thấy điểm outlier trong cột Age.
    - Cột Annual Income(k$):
        - ✓ Dữ liệu tập trung nhiều nhất ở khoảng từ 60 đến 80.
        - ✓ Ở đây có thể có 1 số điểm ngoại lai khi chỉ số Annual Income vượt qua 100.
    - Cột Spending Score:

- ✓ Các giá trị xuất hiện nhiều nhất là khoảng từ 40 đến 60.
- ✓ Nhìn chung không có điều gì bất thường, cũng như không có điểm outlier trong cột này.

- Trực quan biến có dạng categorical:

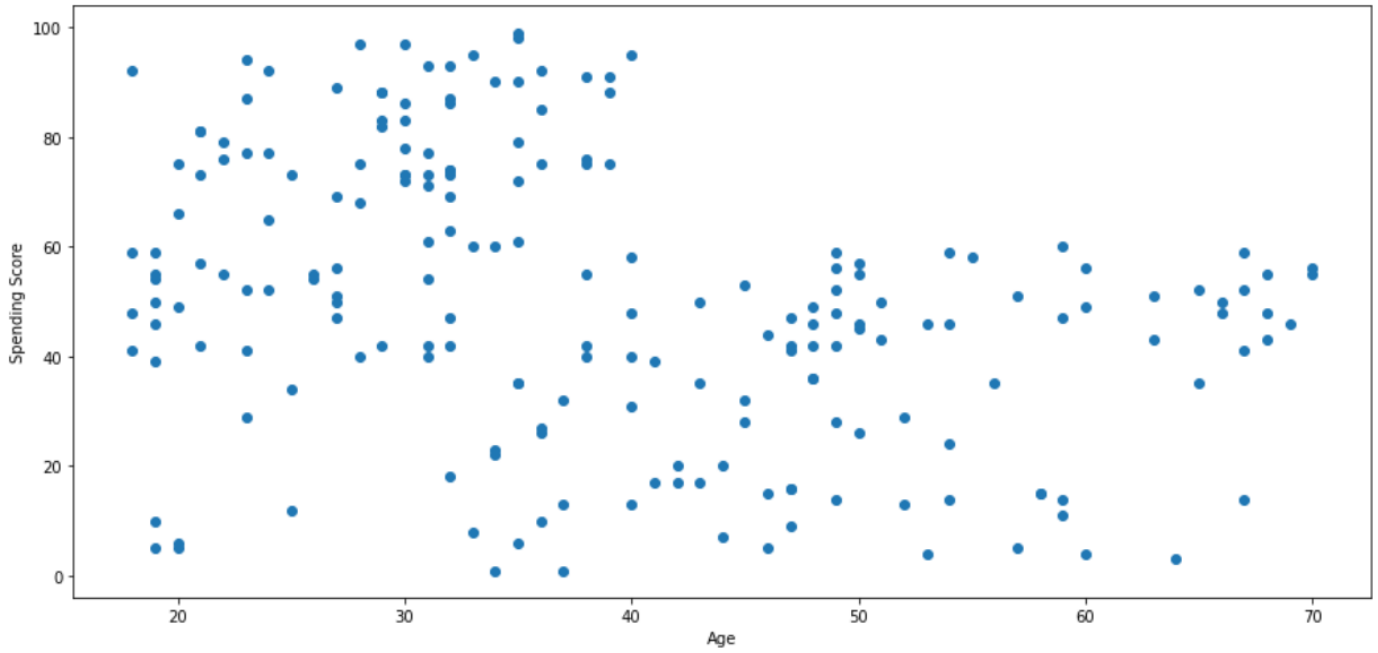


- Nhận xét cho cột Gender:
  - Số lượng giới tính nam là 88 và số lượng giới tính nữ 112.
  - Ở đây ta nhận thấy tỉ lệ giới tính nữ cao hơn một chút so với tính nam khoảng 20 người.

3.3/ Đặt câu hỏi về các cặp biến có thể phân nhóm

<u>Câu 1: Điểm tiêu dùng ở các nhóm khách hàng sử dụng dịch vụ thuộc các độ tuổi khác nhau có sự khác biệt gì với nhau hay không?</u>

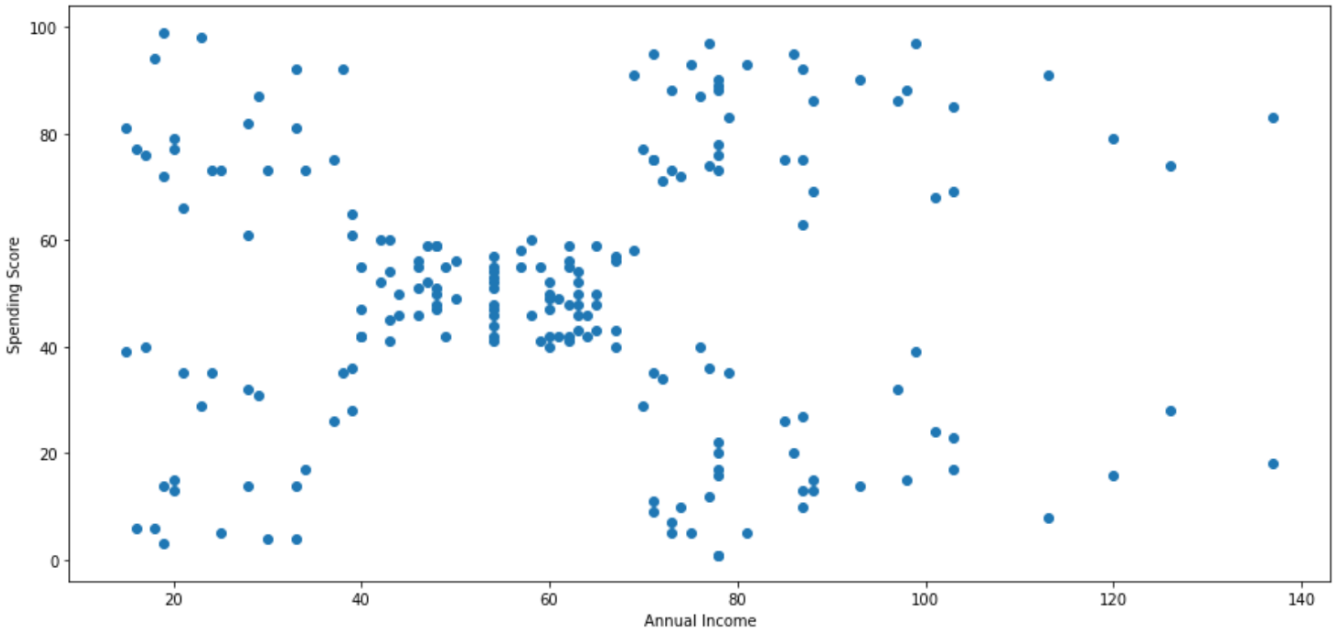
- Trực quan 2 biến bằng biểu đồ scatter:



- Nhận xét:
  - Mặc dù các nhóm không thể hiện rõ nhưng có thể phân nhóm được dựa trên 2 yếu tố này.
  - Không tìm thấy các điểm outlier do các điểm dữ liệu trải đều ở cả Age lẫn Spending Score.
  - Ta có thể dựa vào các nhóm này để giúp cho công ty có những chính sách ưu đãi cho từng nhóm khách hàng cụ thể nhằm khuyến khích khách hàng mua hàng nhiều hơn.
  - Với từng nhóm cụ thể ta có thể cử những nhóm quản lý khác nhau để có những nắm bắt nhu cầu tiêu dùng.

Câu 2: Liệu ta có thể chia các nhóm khách hàng theo thu nhập và điểm tiêu dùng hay không?

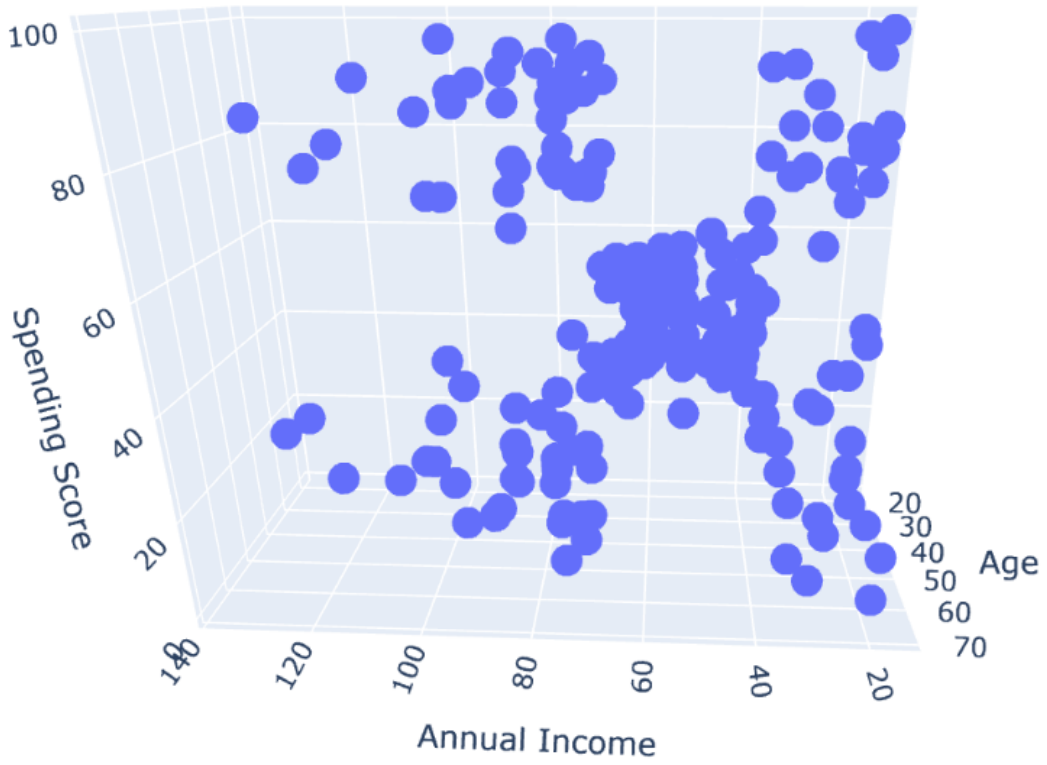
- Trực quan 2 biến Annual Income và Spending Score bằng biểu đồ scatter:



- Nhận xét:
    - Các nhóm thể hiện khá rõ trên biểu đồ và tất nhiên ta có thể áp dụng thuật toán k-means cho 2 biến này.
    - Với các nhóm sau khi chia được, ta có thể dựa vào thu nhập cũng như khả năng tiêu dùng của khách hàng để đề xuất các mặt hàng tương ứng cũng như có các chiến lược riêng để kích thích tiêu dùng.
    - Các nhân viên cũng có thể dễ dàng quan sát trong các nhóm khách hàng nhỏ để đưa ra các kế hoạch phát triển việc bán hàng trong tương lai.

Câu 3: Liệu có khả quan khi ta thực hiện chia các nhóm khách hàng dựa theo cả 3 tiêu chí Age, Annual Income và Spending Score?

- Trực quan 3 biến qua biểu Age, Annual Income và Spending Score bằng biểu đồ scatter 3D.



- Nhận xét:
    - Dựa vào biểu đồ ta có thể thấy có ít nhất 5 nhóm có thể chia nên ta có thể áp dụng thuật toán k-means cho 3 biến này.
    - Với việc gom nhóm theo cả 3 thuộc tính là tuổi, thu nhập và điểm tiêu thụ, ta có thể biết tổng quan và cụ thể các nhóm khách hàng mà công ty đang có ở cả 3 tieu chí.
    - Điều này sẽ cung cấp cho các nhân viên cái nhìn tổng quan về các nhóm khách hàng từ đó đưa ra chiến lược chung và lâu dài cho công ty.

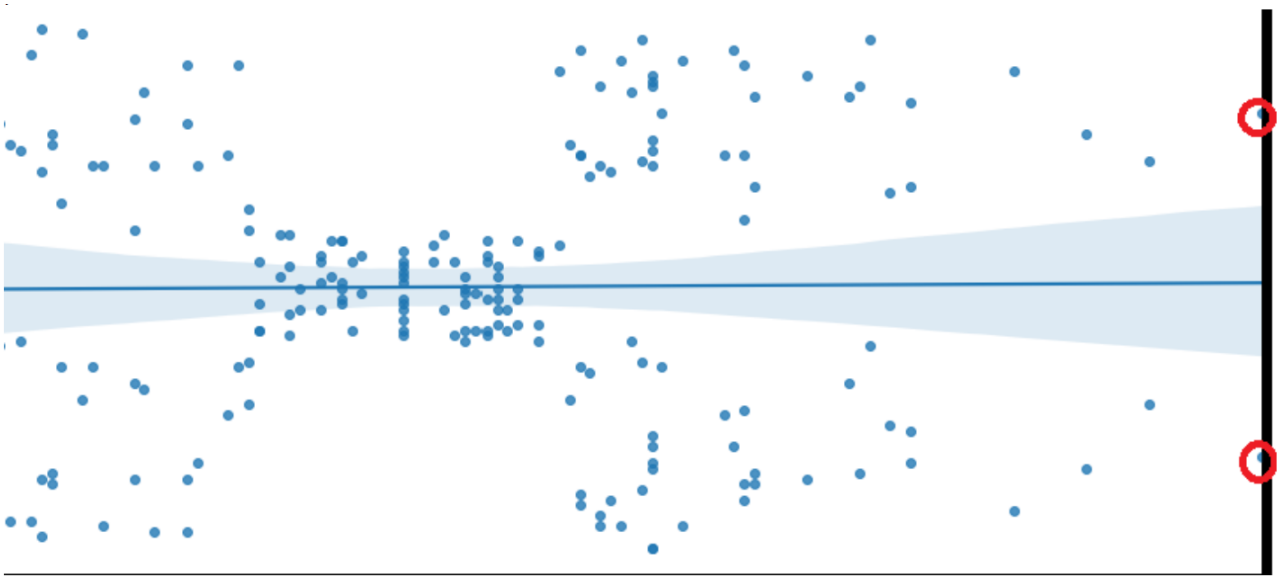
4/ Giải thích các thuật toán sử dụng trong đồ án

4.1/ Giải thích thuật toán k-means:

- Thuật toán k-means dùng để phân cụm các nhóm khách hàng dựa trên các thuộc tính hay nói cách khác là các cột dữ liệu mà ta đưa vào. Trong đồ án này, ta sẽ sử dụng thư việc sklearn để chạy thuật toán kmeans.
- Các bước thực hiện thuật toán:
  - **Bước 1:** chọn ngẫu nhiên k điểm dữ liệu làm điểm trung tâm (centroids).
  - **Bước 2:** tính khoảng cách (Euclidean), giữa tất cả các điểm dữ liệu và các điểm trung tâm.
  - **Bước 3:** gán mỗi điểm dữ liệu cho tâm gần nhất theo khoảng cách tìm được.
  - **Bước 4:** cập nhật vị trí trung tâm bằng cách lấy giá trị trung bình của các điểm trong mỗi nhóm cụm.
  - **Bước 5:** lặp lại từ **Bước 2** đến **Bước 4** cho đến khi các điểm trung tâm không thay đổi.

4.2/ Giải thích thuật toán loại bỏ Outlier:

- Sau khi biểu diễn dữ liệu, có vẻ ta thấy dữ liệu có vài điểm có giá trị Annual Income cao bất thường và nằm khá xa so với phần lớn dữ liệu nên ta sẽ loại bỏ các điểm dữ liệu này.



- Phương pháp loại bỏ: Ta sẽ sử dụng kỹ thuật "Inter-Quartile Range Method" để tìm ra các điểm dữ liệu ngoại lai.

- Inter-Quartile Range (được viết tắt là IQR). IQR được tính bằng công thức: IQR = Q3 - Q1

- Trong đó:
  - Q1: được gọi là 'first quartile' của dữ liệu. Nói cách khác, 25% dữ liệu, sẽ nhỏ hơn Q1.
  - Q3: được gọi là 'third quartile' của dữ liệu. Nói cách khác, 75% dữ liệu, sẽ nhỏ hơn Q3.

- Ta định nghĩa phạm vi quyết định (decision range), với công thức:
  - Giới hạn dưới: Q1 - 1.5 * IQR
  - Giới hạn trên: Q3 + 1.5 * IQR

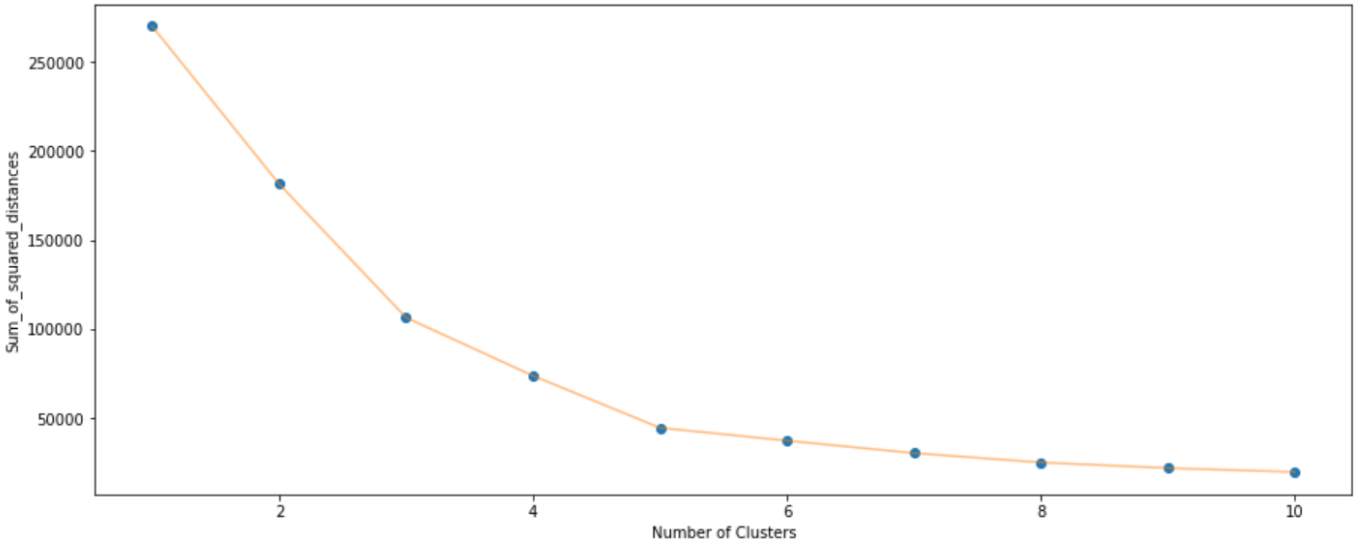
- Các điểm dữ liệu nằm ngoài giới hạn dưới và giới hạn trên, được xem là điểm dữ liệu ngoại lai.

4.3/ Giải thích thuật toán Elbow để tìm ra số cluster:

- Phương thức elbow chạy thuật toán K-means trên tập dữ liệu cho một phạm vi các giá trị k ở đây ta sẽ chạy từ 1 đến 10).

- Ta sẽ thực hiện thuật toán K-means với tất cả các giá trị từ 1 đến 10 của k. Với mỗi giá trị k, ta sẽ tính khoảng cách trung bình đến tâm cho tất cả các điểm dữ liệu.

- Vẽ đồ thị các điểm này ứng với giá trị k và tìm vị trí mà khoảng cách trung bình từ trung tâm giảm đột ngột (đó còn gọi là điểm "Khuỷu tay").



- Như hình ảnh trên: điểm tương ứng với 'Khuỷu tay' mà ta đã nói ở trên là k = 4. Cũng là số cluser tối ưu cho thuật toán.

4.4/ Giải thích thuật toán Silhoutte để tìm ra số cluster:

- Hệ số Silhoutte là thước đo mức độ tương tự của một điểm dữ liệu trong một cluster so với các cluster khác.

- Ta sẽ chọn một dải giá trị của k (ở đây ta sẽ chọn từ 2 đến 11).
- Sau đó vẽ đồ thị silhoutte cho mỗi giá trị của K.

- Phương trình tính toán hệ số Silhoutte cho một điểm dữ liệu:

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

- S(i) là hệ số Silhoutte của điểm dữ liệu i.

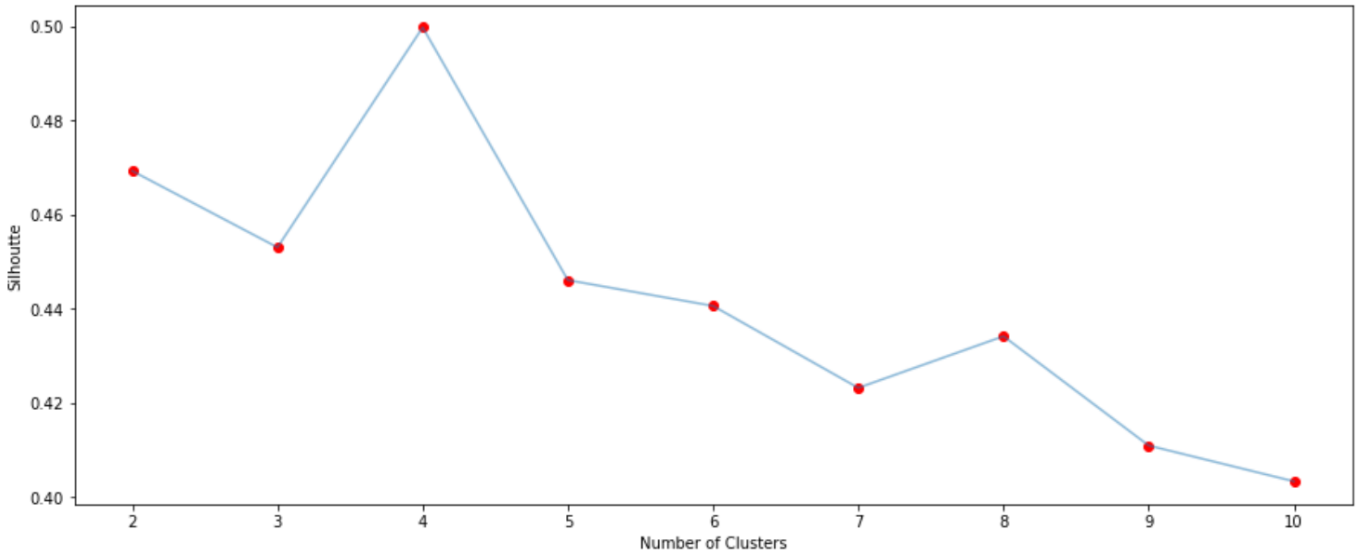
- a(i) là khoảng cách trung bình giữa i và tất cả các điểm dữ liệu khác trong cluster mà nó thuộc về.
- b(i) là khoảng cách trung bình từ i đến tất cả các cụm mà nó không thuộc về.

- Sau đó, ta sẽ tính toán trung bình silhouette cho mọi k:

$$\text{AverageSilhouette} = \text{mean}\{S(i)\}$$

- Trực quan các giá trị trung bình silhoutte và k. Sau đó chọn ra giá trị lớn nhất của silhoutte và chiếu xuống k, đó chính là số cluster tối ưu cần tìm.

- Ví dụ về trực quan các giá trị trung bình silhoutte:



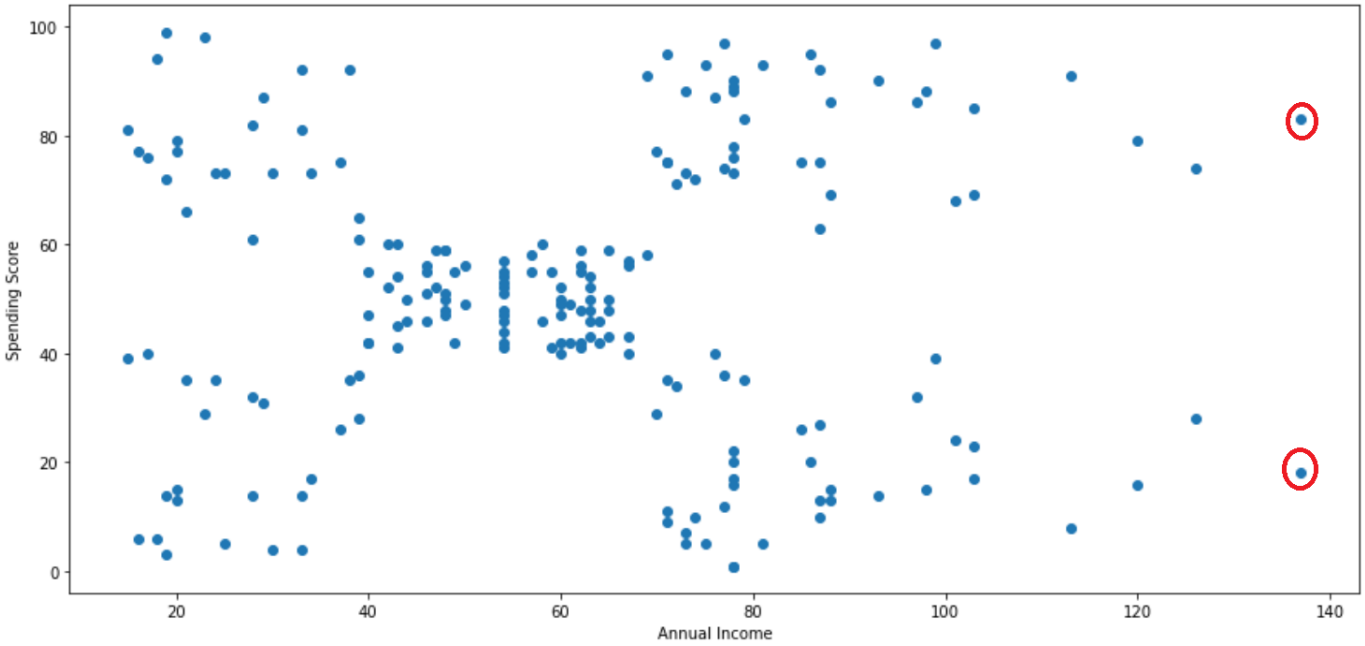
- Như vậy, ở đây 4 là số cluster cần tìm.

5/ Áp dụng thuật toán, trực quan kết quả và đưa ra kết luận:

5.1/ Tiền xử lý loại bỏ Outlier cho cột Annual Income:

- Như đã nói ở phần giải thích thuật toán loại bỏ Outlier. Ở cột Annual Income tồn tại một số điểm Outlier có thể làm ảnh hưởng đến mô hình nên ta cần loại bỏ trước khi thực hiện phân cụm.

- Sau khi trực quan tương quan giữa 2 biến Annual Income và Spending Score thì ta suy đoán có 2 điểm Outlier:



- Ta sẽ áp dụng thuật toán để loại bỏ Outlier và xem có đúng là 2 điểm này hay không:

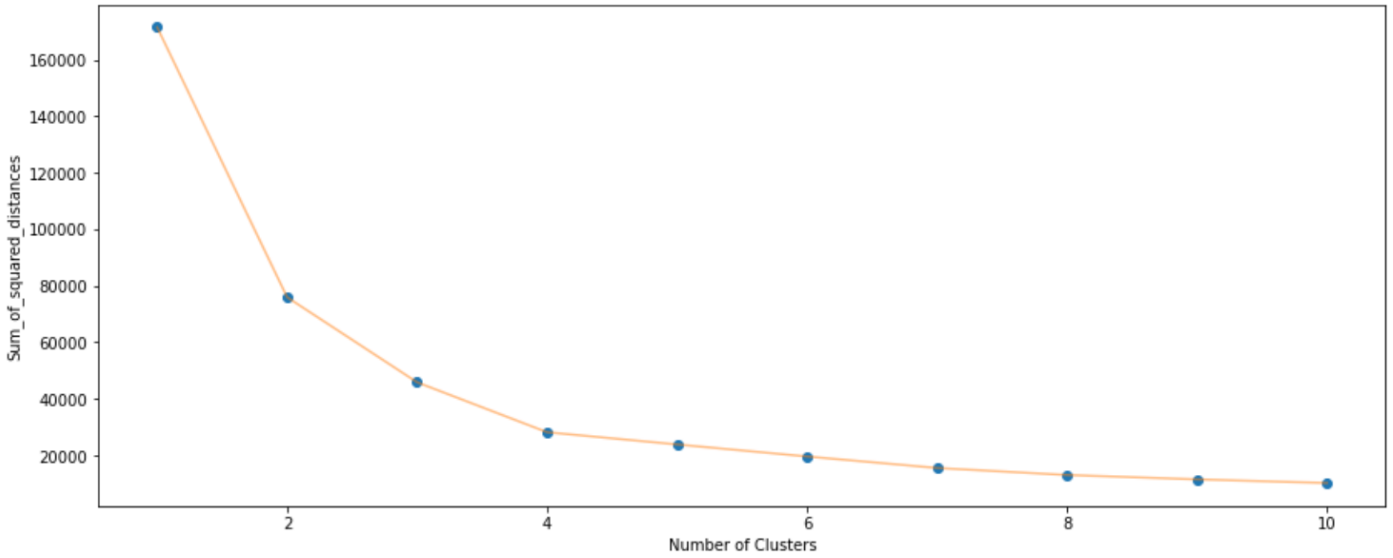
```
#Loại bỏ outlier ở cột Annual Income
q1_annual = round(np.quantile(mall_df['Annual Income (k$)'], 0.25))
q3_annual = round(np.quantile(mall_df['Annual Income (k$)'], 0.75))
iqr_annual = q3_annual - q1_annual

lower = q1_annual - 1.5 * iqr_annual
upper = q3_annual + 1.5 * iqr_annual
temp = mall_df.loc[(mall_df['Annual Income (k$)'] < lower) | (mall_df['Annual Income (k$)'] > upper)]
outlier_annual_list = list(temp.index)
print('Outlier index: ', outlier_annual_list)

second_df = mall_df.drop(outlier_annual_list)
print('Drop outlier completed!')
```

```
Outlier index:  [198, 199]
Drop outlier completed!
```

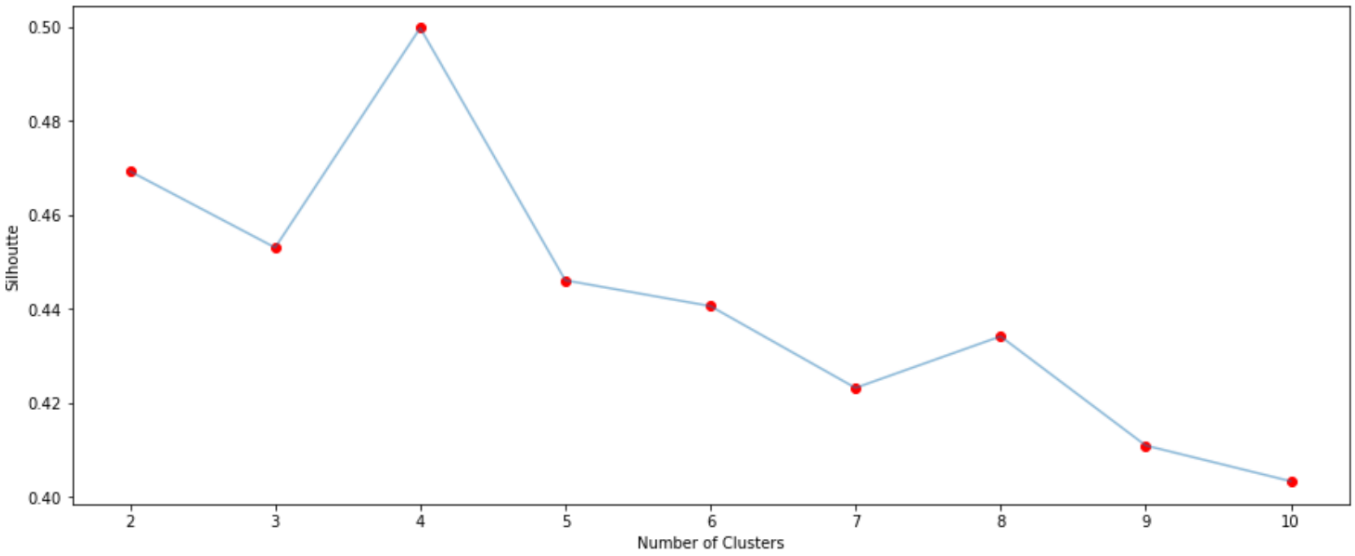
- Sau khi chạy thuật toán ta tìm ra 2 điểm outlier ở dòng thứ 198 và 199 nên ta sẽ loại bỏ 2 dòng này khỏi tập dữ liệu.

5.2/ Phân cụm dựa trên Age và Spending Score:

- Áp dụng thuật toán Elbow để tìm ra số cluster tối ưu:



- Áp dụng thuật toán Silhoutte để tìm ra số cluster tối ưu:

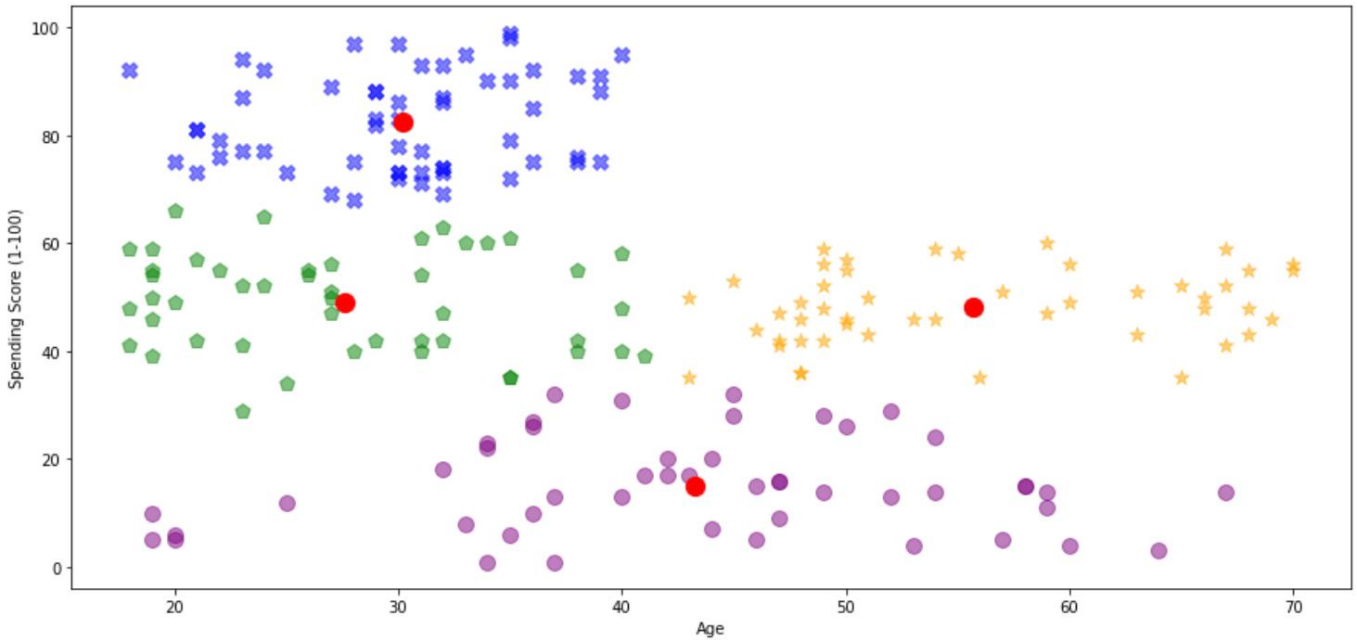


- Cả 2 thuật toán Elbow và Silhoutte đều chỉ ra số cluster tối ưu cho mô hình là 4.

- Gọi thư viện sklearn và áp dụng thuật toán kmeans với số cluster là 4. Dưới đây là phần giải thigh cách gọi thuật toán(trong các phần sau sẽ không giải thích lại):

```
# Áp dụng thuật toán k-means để phân nhóm
algorithm = (KMeans(n_clusters = 4 ,init='k-means++', n_init = 10 ,max_iter=300,
                    tol=0.0001,  random_state= 111  , algorithm='elkan') )
algorithm.fit(X1)
#labels1 là mảng chứa nhãn cluster của các điểm dữ liệu
labels1 = algorithm.labels_
#centroids là tọa độ tâm cluster
centroids1 = algorithm.cluster_centers_
```

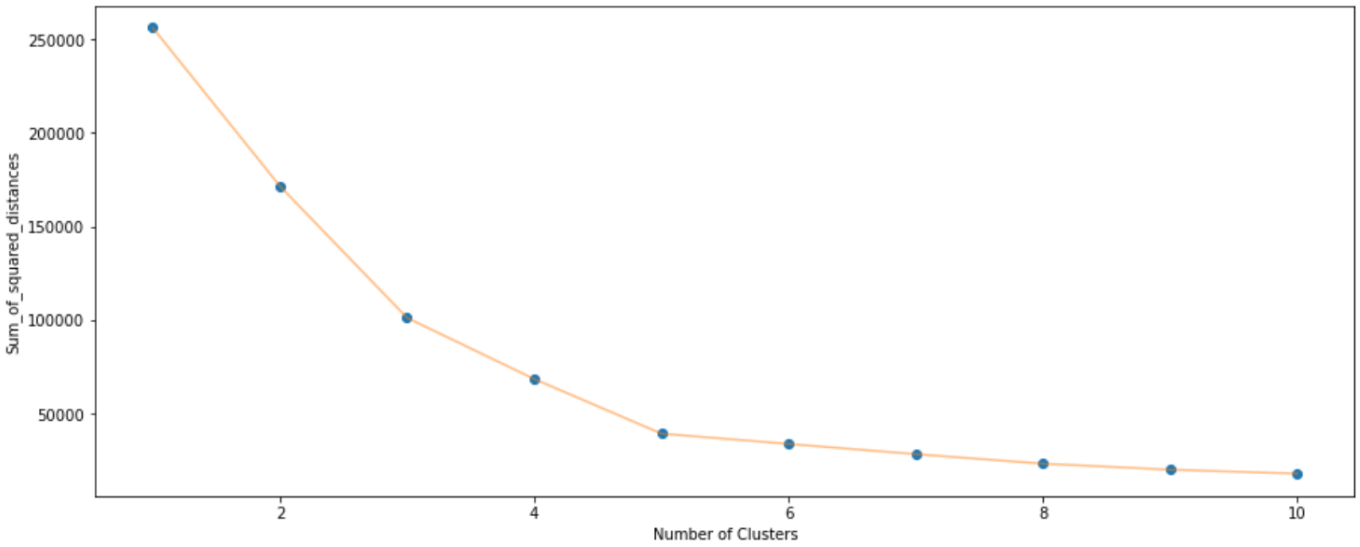
- Trực quan kết quả phân cụm với thuật toán kmeans:



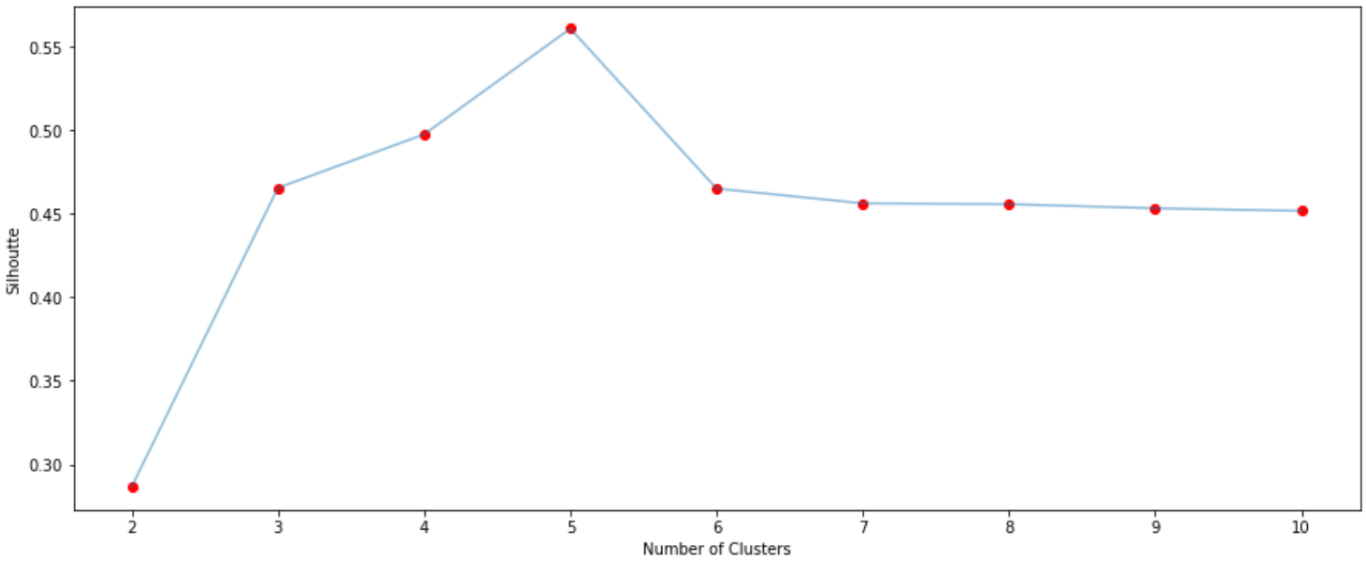
- **Kết luận:**
  - Nhóm thứ nhất: điểm tiêu dùng từ 0 đến 25 trải đều ở tất cả độ tuổi.
  - Điểm tiêu dùng từ 30 đến 65 thì có thể chia ra làm 2 nhóm:
    - ✓ Nhóm thứ hai có độ tuổi trẻ hơn từ 18 đến 40.
    - ✓ Nhóm thứ ba có độ tuổi trẻ hơn từ 45 đến 70.
  - Nhóm cuối cùng có điểm tiêu dùng cao nhất từ 70 đến 100 thì có độ tuổi còn khá nhỏ từ 18 đến 40.
  - Ta có thể dựa vào các nhóm này để giúp cho công ty có những chính sách ưu đãi cho từng nhóm khách hàng cụ thể nhằm khuyến khích khách hàng mua hàng nhiều hơn.
  - Với từng nhóm cụ thể ta có thể cử những nhóm quản lý khác nhau để có những nắm bắt nhu cầu tiêu dùng.

5.3/ Phân cụm dựa trên Annual Income và Spending Score:

- Áp dụng thuật toán Elbow để tìm ra số cluster tối ưu:

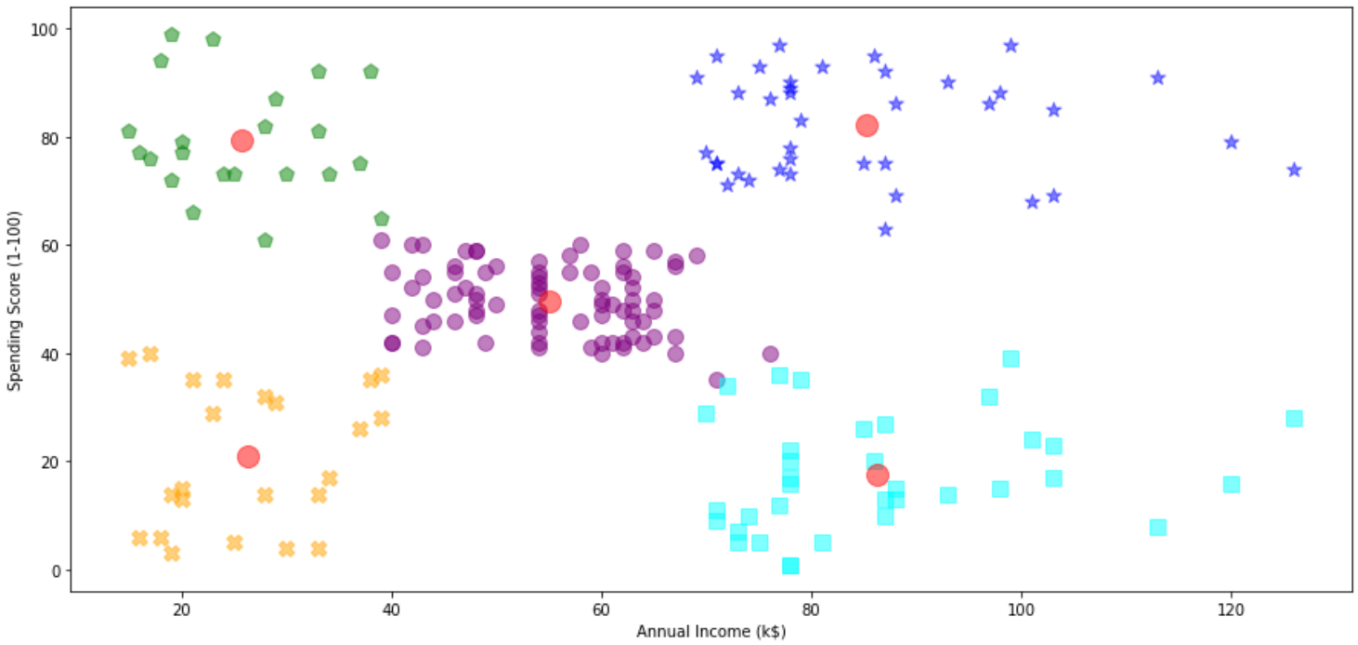


- Áp dụng thuật toán Silhoutte để tìm ra số cluster tối ưu:



- Cả 2 thuật toán Elbow và Silhoutte đều chỉ ra số cluster tối ưu cho mô hình là 5.

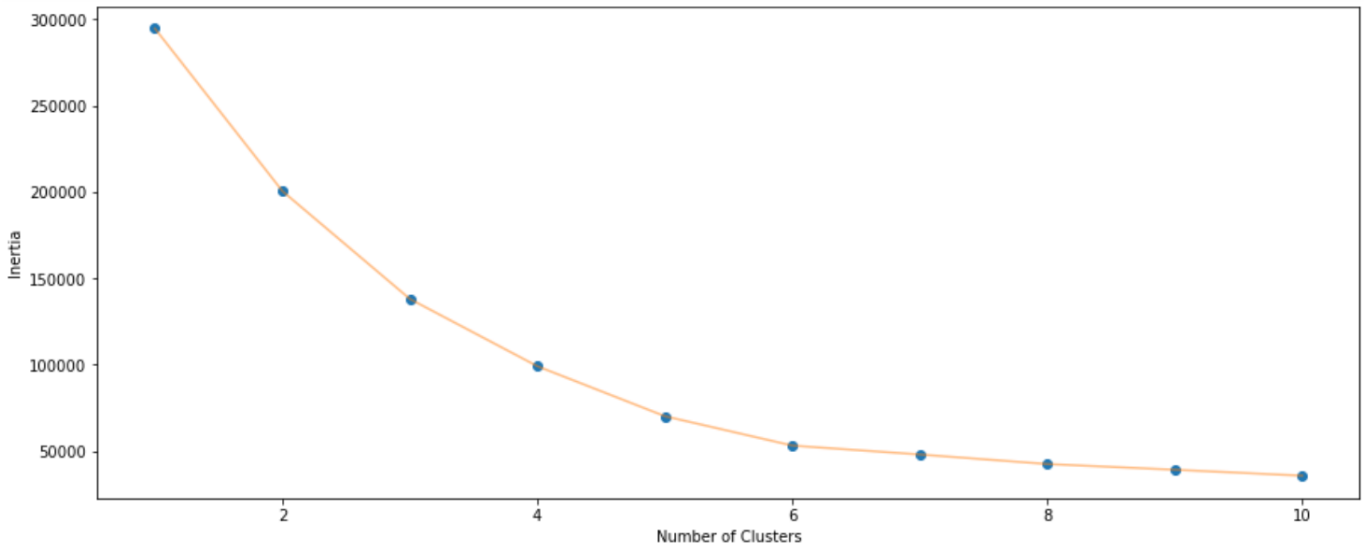
- Trực quan kết quả phân cụm với thuật toán kmeans:



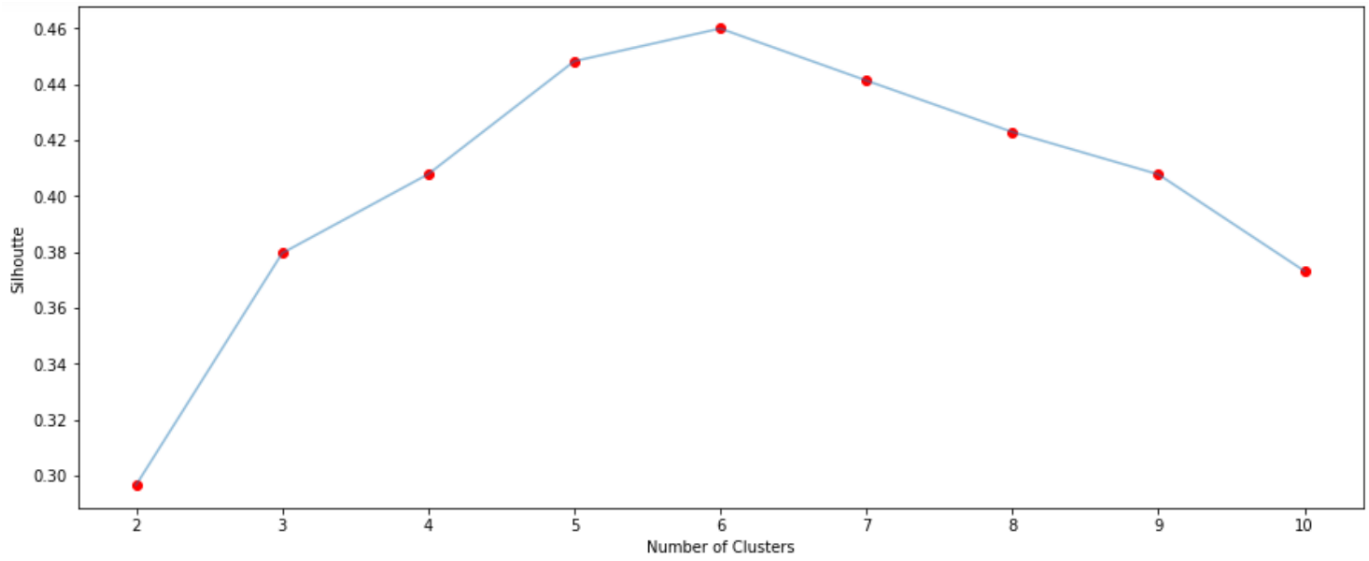
- **Kết luận:**
  - Hai nhóm đầu tiên có điểm tiêu dùng từ 0 đến 40:
    - ✓ Nhóm thứ nhất có thu nhập từ 15 đến 40.
    - ✓ Nhóm thứ hai có thu nhập từ khoảng 70 đến 140.
  - Nhóm thứ ba: thu nhập vào khoảng 40 đến 80 và điểm tiêu dùng ở mức từ 40 đến 60.
  - Hai nhóm cuối cùng có cùng điểm tiêu dùng từ 60 đến 100:
    - ✓ Nhóm thứ tư có thu nhập từ 14 đến 40.
    - ✓ Nhóm thứ năm có thu nhập từ 70 đến 140.
  - Với các nhóm sau khi chia được, ta có thể dựa vào thu nhập cũng như khả năng tiêu dùng của khách hàng để đề xuất các mặt hàng tương ứng cũng như có các chiến lược riêng để kích thích tiêu dùng.
  - Các nhân viên cũng có thể dễ dàng quan sát trong các nhóm khách hàng nhỏ để đưa ra các kế hoạch phát triển việc bán hàng trong tương lai.

5.4/ Phân cụm dựa trên Age, Annual Income và Spending Score:

- Áp dụng thuật toán Elbow để tìm ra số cluster tối ưu:

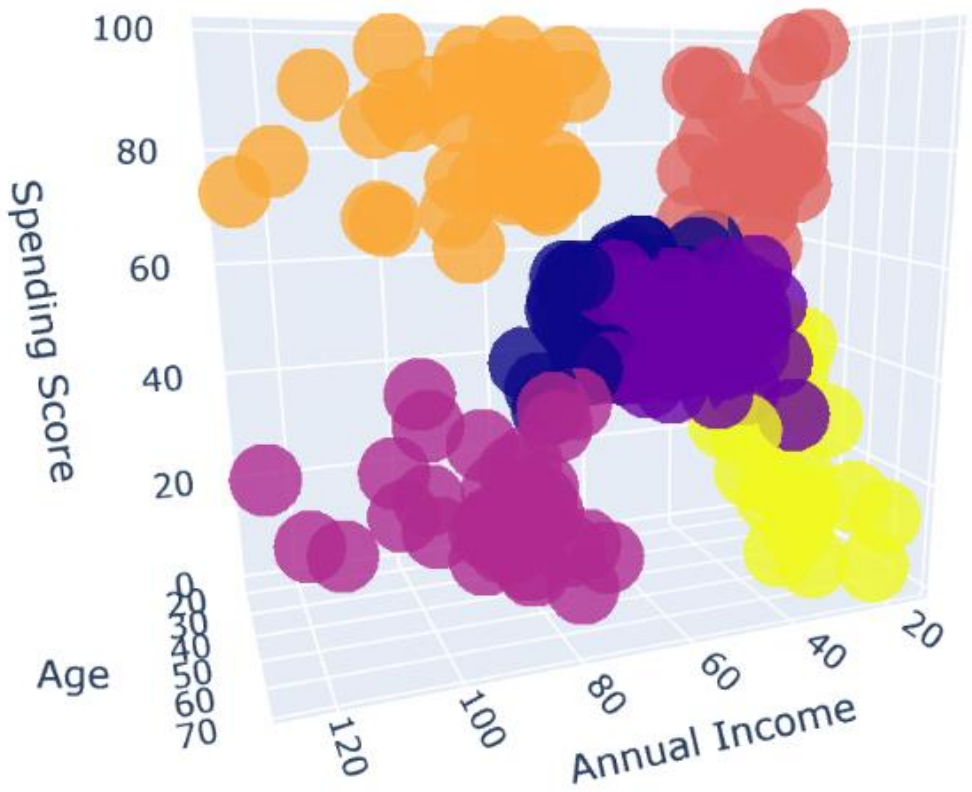


- Áp dụng thuật toán Silhoutte để tìm ra số cluster tối ưu:



- Cả 2 thuật toán Elbow và Silhoutte đều chỉ ra số cluster tối ưu cho mô hình là 6.

- Trực quan kết quả phân cụm với thuật toán kmeans:



- **Kết luận:**
  - Do trong không gian 3 chiều nên tương đối khó để nhận diện khoảng giá trị của các nhóm nên ở đây ta sẽ gọi tên các nhóm qua các điểm trung tâm:
    - ✓ **Nhóm 1:** Lấy trung tâm là độ tuổi 27, mức thu nhập là khoảng 56.57, và điểm tiêu thụ là 49.13.
    - ✓ **Nhóm 2:** Lấy trung tâm là độ tuổi 56, mức thu nhập là khoảng 53.37, và điểm tiêu thụ là 49.08.

- ✓ **Nhóm 3:** Lấy trung tâm là độ tuổi 42, mức thu nhập là khoảng 86.79, và điểm tiêu thụ là 17.26.
- ✓ **Nhóm 4:** Lấy trung tâm là độ tuổi 25, mức thu nhập là khoảng 25.72, và điểm tiêu thụ là 79.36.
- ✓ **Nhóm 5:** Lấy trung tâm là độ tuổi 33, mức thu nhập là khoảng 85.21, và điểm tiêu thụ là 82.11.
- ✓ **Nhóm 6:** Lấy trung tâm là độ tuổi 44, mức thu nhập là khoảng 25.14, và điểm tiêu thụ là 19.52.

- Với việc gom nhóm theo cả 3 thuộc tính là tuổi, thu nhập và điểm tiêu thụ, ta có thể biết tổng quan và cụ thể các nhóm khách hàng mà công ty đang có ở cả 3 tiêu chí.
- Điều này sẽ cung cấp cho các nhân viên cái nhìn tổng quan về các nhóm khách hàng từ đó đưa ra chiến lược chung và lâu dài cho công ty.