Nghiên cứu

Bài toán phân nhóm đối với khách hàng mua sắm tại siêu thị Coopextra Thủ Đức

Lê Hồng Diễn*, Nguyễn Phúc Sơn, Phạm Hoàng Uyên, Lê Văn Hinh

TÓM TẮT

Phân khúc khách hàng (customer segmentation) là quá trình phân nhóm khách hàng dựa trên các đặc điểm chung như hành vi, thói quen mua sắm và sử dụng dịch vụ của họ …để các công ty, doanh nghiệp có thể tiếp thị cho từng nhóm khách hàng một cách hiệu quả và phù hợp hơn. Phân khúc khách hàng giúp cho các nhà tiếp thị hiểu hơn về khách hàng cũng như đưa ra các mục tiêu, chiến lược và các phương thức tiếp thị cho các nhóm đối tượng khác nhau.Trong bài báo này, chúng tôi nghiên cứu bài toán phân khúc khách hàng thông qua các phương pháp phân cụm (clustering methods) trong thống kê và học máy không giám sát (unsupervised learning). Các thuật toán được dùng là K-means và Elbow vốn là các thuật toán nổi tiếng đã được ứng dụng thành công trong nhiều lĩnh vực như marketing, sinh học, thư viện, bảo hiểm, tài chính... Mục đích của việc phân cụm là tìm ra các phân khúc thị trường có ý nghĩa. Tuy nhiên, việc lựa chọn cũng như thay đổi các tham số của thuật toán để cho các thuật toán này trở nên hiệu quả trong việc tìm ra các phân khúc thị trường có ý nghĩa đó vẫn còn là một thách thức hiện nay. Trong bài báo này, chúng tôi đã tiến hành nghiên cứu triển khai cho một bộ dữ liệu khách hàng tại siêu thị CoopExtra Thủ Đức và đạt được một số phân khúc hữu dụng, hứa hẹn sẽ giúp việc chăm sóc, tiếp thị khách hàng hiệu quả hơn.

Từ khoá: phân khúc khách hàng, phân khúc thị trường, phương pháp phân cụm, thuật toán K-means, phương pháp Elbow

GIỚI THIỆU

Phân tích khách hàng là một nhánh cực kỳ quan trọng trong việc phân tích dữ liệu kinh doanh[1]. Tìm hiểu hành vi, ghi nhận thói quen mua sắm, nắm bắt sở thích khách hàng v.v... luôn được các doanh nghiệp đầu tư bài bản nhằm tạo ra lợi thế cạnh tranh lâu dài. Nhóm khách hàng của một công ty thường đa dạng về thành phần, khác nhau về độ tuổi v.v... từ đó dẫn đến tâm lý mua sắm rất khác nhau. Do đó, các doanh nghiệp thường phải phân chia khách hàng ra thành các nhóm có những đặc điểm tương tự nhau, từ đó đưa ra các chiến lược sản xuất, tiếp thị sản phẩm nhằm đáp ứng tốt hơn nhu cầu mua sắm, tăng doanh thu công ty. Có nhiều cách để phân chia hay phân cụm khách hàng. Trước đây, bộ phận marketing phân cụm chủ yếu dựa vào các thông tin truyền thống như:

- Nhân khẩu học (bao gồm độ tuổi, giới tính, thu nhập và giáo dục)
- Tâm lý học (như tầng lớp xã hội, lối sống và đặc điểm cá tính)
- Dữ liệu hành vi (bao gồm thói quen chi tiêu)
- Thông tin địa lý (thị trấn, quận, thành phố, tiểu bang, quốc gia cư trú).

Ngày nay, với các thành tựu của khoa học dữ liệu trong cuộc cách mạng công nghiệp 4.0, doanh nghiệp bắt đầu thu thập và xử lý dữ liệu khách hàng một cách bài bản và chi tiết hơn nhiều. Việc này giúp bộ phận chăm sóc, tiếp thị khách hàng có điều kiện hiểu sâu hơn hành vi mua sắm, thói quen, sở thích v.v...

Cấu trúc bài báo gồm các phần:

- Giới thiệu
- Phương pháp nghiên cứu
- Mô tả dữ liệu
- Các kết quả phân tích chính
- Thảo luận
- Kết luận

Đại học Kinh tế - Luật, Đại học Quốc gia Thành phố Hồ Chí Minh

Liên hệ

Lê Hồng Diễn, Đại học Kinh tế - Luật, Đại học Quốc gia Thành phố Hồ Chí Minh

Email: dienlh@uel.edu.vn

Lịch sử

- Ngày nhận: 12-12-2018
- Ngày chấp nhận: 22-01-2019
- Ngày đăng: 31-03-2019

DOI :

Check for updates

VNU-HCM Press

Trích dẫn bài báo này: Hồng Diễn L, Phúc Sơn N, Hoàng Uyên P, Văn Hinh L. **Bài toán phân nhóm đối với khách hàng mua sắm tại siêu thị Coopextra Thủ Đức**. *Sci. Tech. Dev. J. - Eco. Law Manag.;* 3(1):28-36.

PHƯƠNG PHÁP NGHIÊN CỨU

Phương pháp nghiên cứu chính của đề tài này là phương pháp phân cụm[2]. Phân cụm là một kĩ thuật Machine Learning phổ biến để phân tích dữ liệu được sử dụng trong nhiều lĩnh vực như marketing, y tế, sinh học…cũng như nghiên cứu kinh tế, tài chính.

Phân cụm là quá trình phân loại các điểm dữ liệu vào các nhóm cụ thể. Trong đó, các điểm dữ liệu trong cùng một nhóm phải có các thuộc tính tương tự (similar features) và ngược lại, các điểm trong các nhóm khác nhau phải có các thuộc tính không giống nhau (dissimilar features). Độ đo khoảng cách để đánh giá độ tương tự giữa các điểm dữ liệu.

Mục tiêu của phân cụm là tìm ra các nhóm dữ liệu tương đồng. Tuy nhiên, không có tiêu chí nào được xem là tốt nhất để đánh giá hiệu quả của phân cụm, điều này phụ thuộc vào mục đích của phân cụm.

Các phương pháp phân cụm có thể được chia thành hai loại cơ bản: phân cụm theo cấp bậc (Hierarchical clustering) và Partitional clustering. Hierarchical clustering tiến hành hợp nhất liên tiếp các cụm nhỏ thành các cụm lớn hơn hoặc bằng cách tách các cụm lớn thành các cụm nhỏ hơn. Partitional clustering là các phương pháp phân nhóm được sử dụng để phân loại các quan sát trong một tập dữ liệu thành nhiều nhóm dựa trên sự giống nhau của chúng. Các thuật toán yêu cầu người dùng chỉ định số lượng cụm được tạo. Trong bài báo này chúng tôi sử dụng phương pháp phân cụm phổ biến đó là phương pháp K-means[3].

Phân cụm K-means (MacQueen, 1967) là thuật toán học máy không được giám sát được sử dụng để phân nhóm các đối tượng đã cho vào k cụm, trong đó k được chỉ định trước. Trong phân cụm K-means, mỗi cụm được biểu diễn bằng tâm của nó (centroid) tương ứng với trung bình của các điểm được gán cho cụm[4].

Ý tưởng chính của thuật toán K-means là xác định các cụm sao cho total within-cluster variation là nhỏ nhất với định nghĩa total within-cluster variation như sau:

$$tot.withiness = \sum_{k=1}^{k} \mathrm{W}(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Trong đó, x_i là điểm dữ liệu thuộc cụm C_k, μ_k là giá trị trung bình của các điểm trong cụm C_k.

Thuật toán K-means có thể tóm tắt như sau

1. Chỉ định số lượng cụm k.
2. Chọn ngẫu nhiên k điểm từ tập dữ liệu làm trung tâm (centroids) cho k cụm.
3. Tính khoảng cách giữa các điểm đến k tâm (thường dùng khoảng cách Euclidean).
4. Nhóm các đối tượng vào nhóm gần nhất.
5. Xác định lại tâm mới cho các nhóm bằng cách tính giá trị trung bình cho các điểm dữ liệu trong các cụm tương ứng.
6. Thực hiện lại bước 3 cho đến khi không có sự thay đổi nhóm nào của các điểm dữ liệu

MÔ TẢ DỮ LIỆU

Bộ dữ liệu khách hàng thu thập được có 475 điểm dữ liệu từ các khách hàng mua sắm tại siêu thị CoopExtra quận Thủ Đức. Để có được bộ dữ liệu này, chúng tôi thực hiện thu hóa đơn mua hàng của 475 khách hàng. Sau đó thực hiện các thao tác tiền xử lý dữ liệu. Bộ dữ liệu bao gồm chi tiêu cho 1 lần mua sắm của khách hàng tại siêu thị trên các danh mục sản phẩm đa dạng. Số thuộc tính: 15. Đặc điểm của tập dữ liệu: Đa biến. Đặc tính thuộc tính: numeric và character.

Một mẫu dữ liệu (**Hình 1**) bao gồm các quan sát từ bộ dữ liệu trên được thực hiện bằng phần mềm R:

Chúng ta sẽ khai thác dữ liệu thông qua quan sát mô tả thống kê của tập dữ liệu để biết một số thông tin về từng thuộc tính và mối quan hệ giữa các thuộc tính như thế nào.

Hình 2 là bảng thống kê mô tả của bộ dữ liệu được thực hiện bằng hàm summary() trong R.

Nhìn vào biểu diễn Boxplot cho bộ dữ liệu (**Hình 3**) được vẽ bằng hàm boxplot() trong R, ta thấy mỗi tính năng có rất nhiều các điểm ngoại lệ.

Chúng ta lọc các outlier (**Hình 4**) bằng cách sử dụng khoảng cách Cook. Trong thống kê, khoảng cách Cook được dùng để xét ảnh hưởng của điểm dữ liệu khi thực hiện phân tích hồi quy bình phương nhỏ nhất. Khoảng cách này được đặt theo tên của nhà thống kê người Mỹ R. Dennis Cook, người đã đưa ra khái niệm này vào năm 1977.

Các outlier có thể làm ảnh hưởng đến độ chính xác của mô hình phân tích dự đoán. Tuy nhiên trong phân khúc khách hàng, nếu xóa bỏ các outlier thì chúng ta có thể bỏ lỡ nhiều thông tin hữu ích về khách hàng. Đây có thể là các khách hàng thuộc phân khúc tầm cao mang lại giá trị cho doanh nghiệp. Do đó, doanh nghiệp cần phân tích để có cách tiếp cận và dịch vụ chăm sóc khách hàng phù hợp.

CÁC KẾT QUẢ PHÂN TÍCH CHÍNH

Trong phần này chúng ta sẽ sử dụng hàm K-means trong ngôn ngữ lập trình R để phân khúc khách hàng thành các nhóm riêng biệt dựa trên thói quen mua hàng dựa vào tập dữ liệu trên. Thuật toán xác định được phân khúc hoặc cụm khách hàng có sự tương quan nào đó.

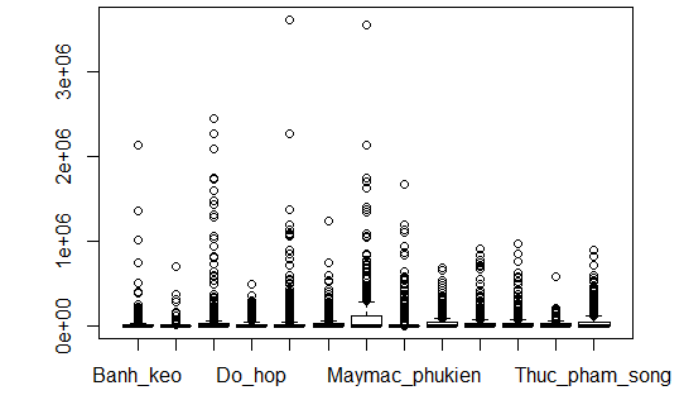
	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
1	1	Dong	54700	0	0	34000	0	12000	0	0	0.0	0	0	37000.0	0.0
2	2	Vang	0	0	35800	35600	0	12400	35000	0	53784.6	163400	325200	0.0	0.0
3	3	Khong	60900	0	1430100	0	105000	130700	359800	0	0.0	0	36000	0.0	0.0
4	4	Khong	0	0	0	0	5900	29800	0	0	0.0	0	0	0.0	19800.0
5	6	Bac	0	0	0	0	0	16500	0	0	0.0	37300	17000	0.0	0.0
6	7	Vang	0	0	202800	0	1134000	0	100000	0	0.0	0	0	0.0	0.0
7	8	Khong	0	0	0	0	7600	0	0	0	0.0	0	0	4000.0	0.0
8	9	Khong	0	0	0	0	0	0	0	58200	0.0	0	0	0.0	0.0
9	10	Khong	0	0	0	0	0	0	0	0	2730.0	0	16000	16000.0	0.0
10	11	Vang	26000	0	0	45500	0	29200	0	0	0.0	91000	54200	0.0	0.0

Showing 1 to 10 of 475 entries

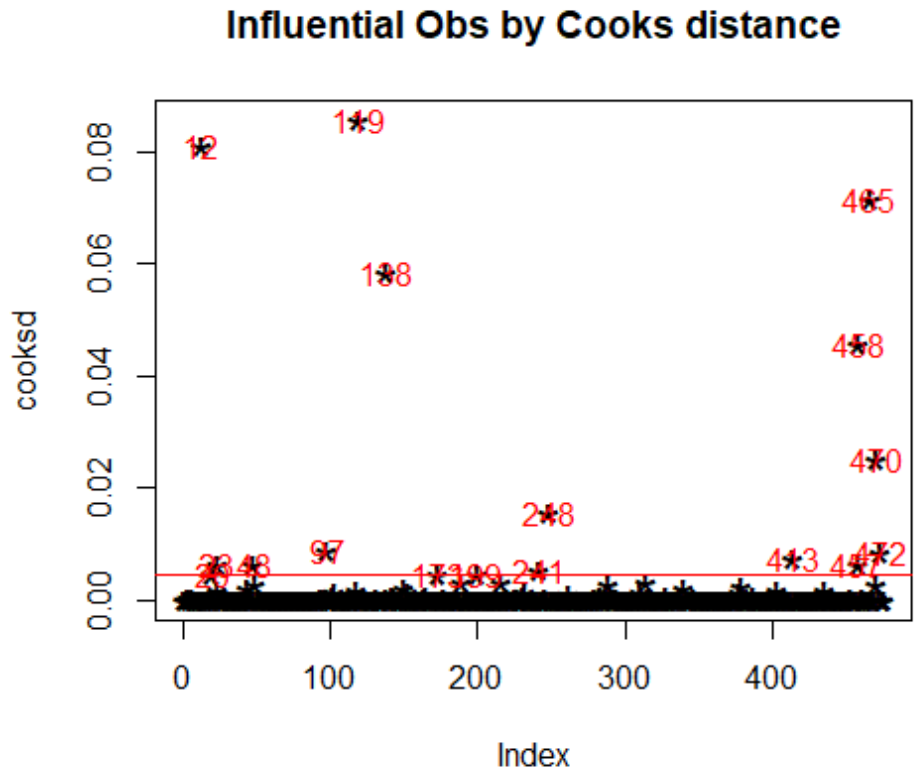
Hình 1: **Mẫu dữ liệu.**

```
   Banh_keo          Det_may       Do_dung_gia_dinh     Do_hop          Do_uong          Gia_vi
 Min.   :      0   Min.   :   0   Min.   :      0   Min.   :      0   Min.   :      0   Min.   :      0
 1st Qu.:      0   1st Qu.:   0   1st Qu.:      0   1st Qu.:      0   1st Qu.:      0   1st Qu.:      0
 Median :      0   Median :   0   Median :      0   Median :      0   Median :      0   Median :      0
 Mean   :  28230   Mean   :5836   Mean   :  85303   Mean   :  26109   Mean   :  63554   Mean   :  34392
 3rd Qu.:  13000   3rd Qu.:   0   3rd Qu.:  24600   3rd Qu.:  20138   3rd Qu.:  19150   3rd Qu.:  24300
 Max.   :2130600   Max.   :704000   Max.   :2451700   Max.   :496000   Max.   :3609000   Max.   :1236000
  Hoamypham_vs      Maymac_phukien     Rau_cu_qua          Sua          Thuc_pham       Thuc_pham_san
 Min.   :      0   Min.   :      0   Min.   :      0   Min.   :      0   Min.   :      0   Min.   :      0
 1st Qu.:      0   1st Qu.:      0   1st Qu.:      0   1st Qu.:      0   1st Qu.:      0   1st Qu.:      0
 Median :      0   Median :      0   Median :      0   Median :      0   Median :      0   Median :      0
 Mean   : 121745   Mean   :  43657   Mean   :  40945   Mean   :  43035   Mean   :  40360   Mean   :  17135
 3rd Qu.: 116250   3rd Qu.:      0   3rd Qu.:  36802   3rd Qu.:  27650   3rd Qu.:  27300   3rd Qu.:  21760
 Max.   :3553000   Max.   :1671500   Max.   :677747    Max.   :908600    Max.   :964300    Max.   :575000
 Thuc_pham_song
 Min.   :     0
 1st Qu.:     0
 Median :     0
 Mean   : 53412
 3rd Qu.: 44655
 Max.   :886724
```

Hình 2: **Thống kê mô tả của bộ dữ liệu.**



Hình 3: **Biểu diễn Boxplot.**



Hình 4: **Các outlier của bộ dữ liệu (Sử dụng hàm cooks.distance() trong R để vẽ).**

Trước tiên ta tiến hành tải bộ dữ liệu và chuẩn hóa bộ dữ liệu bằng hàm scale() trong R.

Thuật toán K-means chỉ định chọn số cụm k được tạo. Hiệu quả của thuật toán phụ thuộc vào việc chọn số cụm k. Vậy làm thế nào để xác định lượng cụm tối ưu trong tập dữ liệu phân tích? Hàm fviz_nbclust () [trong gói factoextra] cung cấp một giải pháp để ước tính số lượng cụm tối ưu. Và phương pháp sử dụng ở đây là phương pháp Elbow[2]. Dựa vào thuật toán phân cụm cho các giá trị k khác nhau, thường là từ 1 đến 10. Với mỗi k, tính total within-cluster sum of square (WSS). Sau đó vẽ đường cong WSS theo số cụm k. Vị trí uốn cong của đồ thị được xem là số cụm tối ưu.

Chúng ta thu được kết quả như **Hình 5**.

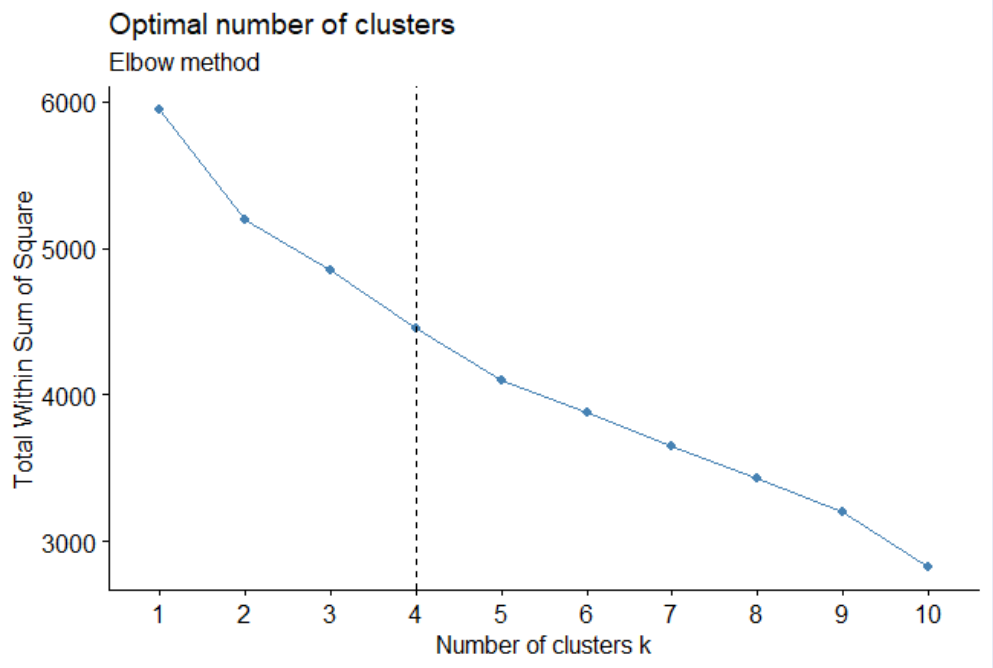
Phương pháp Elbow gợi ý cho chúng ta chọn cụm tối ưu là k=4. Thực ra chúng ta có thể chọn kết quả sai lệch 1 đơn vị, tức là k=3 hoặc k=5. Trong bài này chúng tôi chọn k=4. Sau đó, thực hiện phân cụm sử dụng thuật toán K-means với k=4 và thu được hình ảnh phân cụm như trong **Hình 6**).

Mỗi một màu tượng trưng cho một nhóm khách hàng có thể có chung một đặc điểm mua sắm nào đó. Chúng ta sẽ tìm hiểu và phân tích từng phân cụm để tìm ra đặc điểm chung của mỗi nhóm là gì.

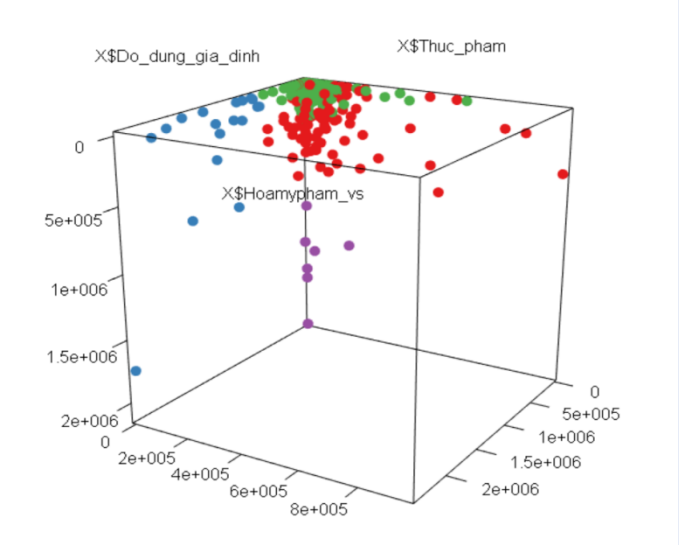
Trong phân cụm 1 bao gồm 7 khách hàng. Nhìn vào **Hình 7**, chúng ta nhận thấy rằng đa phần khách hàng trong phân cụm này mua sắm rất nhiều cho các mặt hàng hóa mỹ phẩm_vệ sinh, đặc biệt là các khách hàng số 3,6,7. Trong khi số tiền trung bình khách hàng chi trả cho hóa mỹ phẩm_vệ sinh trên toàn bộ dữ liệu chỉ là 121745 (VNĐ).Đây hầu hết là các khách hàng thuộc loại thẻ vàng.

Trong phân cụm 2 (**Hình 8**) có 18 khách hàng. Tất cả các khách hàng trong nhóm này đều chi tiêu rất nhiều vào các mặt hàng đồ dùng gia đình. Ngoài ra chúng ta còn khai thác thêm được một số thông tin đáng chú ý. Như khách hàng số 4 ngoài đồ dùng gia đình còn mua số lượng lớn mặt hàng hóa mỹ phẩm_vệ sinh. Hay như khách hàng số 3 còn mua sắm thêm nhiều các mặt hàng hóa mỹ phẩm_vệ sinh và may mặc_phụ kiện, khách hàng số 7, 8 còn chi rất nhiều cho sản phẩm đồ uống.

Trong phân cụm 3 (**Hình 9**) có 105 khách hàng. Nhìn vào bảng dữ liệu trong phân cụm này chúng ta thấy có



Hình 5: **Số cụm tối ưu (sử dụng Hàm fviz_nbclust () trong gói factoextra của R để vẽ).**



Hình 6: **Kết quả phân cụm với k=4.**

	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	97	Vang	0	0	102300	109500	95000	313400	1407000	0	184510.	0	59100	10000	0
2	198	Vang	0	0	0	0	45000	101800	1338600	0	0	31800	177500	0	0
3	280	Vang	0	0	0	0	0	0	1616700	0	0	0	0	0	0
4	360	Vang	0	0	0	0	0	168400	1051500	0	0	662400	0	0	0
5	411	Khong	20200	54000	27000	60200	0	0	1367300	262700	324605.	485600	0	15977.	0
6	536	Vang	0	0	0	0	0	0	1699000	0	0	0	0	0	0
7	681	Bac	0	0	0	0	0	0	2130800	0	0	0	0	0	0

Hình 7: **Dữ liệu của phân cụm 1.**

	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<db1>	<chr>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
1	3	Khong	60900	0	1430100	0	105000	130700	359800	0	0	0	36000	0	0
2	26	Dong	0	0	1031500	0	10500	8900	132700	29000	19327.	0	23500	0	25514.
3	49	Dong	33900	43800	1734800	23300	6200	19600	766700	1192900	80680	0	23800	197506.	60691.
4	93	Vang	0	0	2451700	0	0	0	1746000	0	0	0	0	0	0
5	95	Bac	0	0	1595200	9500	34200	68500	0	0	187549	52000	153500	52978.	64170
6	96	Bac	0	0	2258500	0	0	39900	35000	0	58698	0	48700	55800	110357.
7	117	Vang	29400	0	1024300	0	1069500	34900	0	0	0	0	0	0	0
8	118	Khong	0	54000	2086300	0	1365200	54000	0	0	0	0	50100	0	0
9	153	Bac	0	0	796900	0	0	332800	11900	0	482293.	35400	0	0	109335.
10	165	Vang	0	0	1737900	0	0	0	0	0	0	0	0	0	0
11	222	Khong	0	0	1029000	0	0	0	829900	0	0	0	0	0	0
12	224	Khong	0	376700	1469000	0	0	0	0	0	0	0	0	0	0
13	241	Dong	0	0	815000	0	0	0	0	0	0	0	0	0	0
14	255	Khong	0	0	815300	37200	0	36100	61900	0	243712.	0	24900	148800	333907
15	267	Bac	31900	146200	1311000	0	0	0	117200	563300	0	0	0	64160	0
16	277	Khong	0	0	1059000	0	0	0	0	0	0	0	99000	0	0
17	415	Khong	27800	0	1283800	0	7400	75000	193500	98000	0	161900	7200	0	353963
18	514	Bac	0	0	939000	0	0	0	0	0	0	0	0	0	0

Hình 8: **Dữ liệu của phân cụm 2.**

một số liên hệ giữa các khách hàng nhưng chưa thực sự rõ ràng. Do đó, chúng ta cần thực hiện phân cụm một lần nữa để tìm ra nhóm khách hàng cụ thể hơn.

Với các bước thực hiện phân cụm tương tự như trên cho dữ liệu của phân cụm 3, ta thu được 4 phân cụm tương ứng (**Hình 10**). Để tránh sự nhầm lẫn, chúng tôi kí hiệu các nhóm nhỏ trong phân cụm 3 này lần lượt là các nhóm 3.1, 3.2, 3.2, 3.4.

Nhóm đầu tiên được lọc ra có 8 khách hàng (**Hình 11**) thuộc nhóm chi tiêu nhiều cho sản phẩm đồ uống trong khoảng từ 548500 (VNĐ) đến 1192500 (VNĐ).

Nhóm 3.2 (**Hình 12**) có 16 khách hàng tập trung mua sắm trên mức trung bình cho các mặt hàng may mặc_phụ kiện trong khoảng từ 259000 (VNĐ) đến 1130000 (VNĐ).

Nhóm 3.3 (**Hình 13**) có 26 khách hàng đều chi tiêu trên mức trung bình cho các mặt hàng thực phẩm tươi sống. Chi tiêu trung bình của nhóm này vào mức 409172 (VNĐ).

Nhóm 3.4 (**Hình 14**) tập trung vào nhóm khách hàng mua các sản phẩm hóa mỹ phẩm_vệ sinh trong khoảng từ 253850 (VNĐ) đến 764800 (VNĐ). Nhóm này chi tiêu trên mức trung bình và ít hơn so với phân cụm 1. Có thể hiểu đây là nhóm phân khúc tầm trung và nhóm trong phân cụm 1 là phân khúc tầm cao hơn. Như vậy, sau khi phân tích phân cụm 3 chúng ta tìm ra được một số thông tin hữu ích về khách hàng.

Phân cụm 4 (**Hình 15**) là phân cụm có nhiều khách hàng nhất 328 khách hàng. Tuy nhiên nhìn vào bảng dữ liệu của phân cụm này, chúng ta không thấy mối liên hệ giữa các khách hàng. Và hầu hết các khách hàng chi tiêu cho các mặt hàng đều ở mức thấp. Đây có thể là hộ cá thể gia đình mua sắm không theo quy luật nào.

THẢO LUẬN

Để có dữ liệu phục vụ cho nghiên cứu này, nhóm nghiên cứu đã lên kế hoạch tổ chức và thu thập dữ liệu. Sau đó tiến hành phân tích dữ liệu bằng ngôn ngữ lập trình R. Trong bài báo này, thuật toán sử dụng phân cụm khách hàng là thuật toán K-means.Ưu điểm của thuật toán K-means là đơn giản và hiệu quả, có thể thực hiện trên bộ dữ liệu lớn. Định hướng nghiên cứu của nhóm trong tương lai là mở rộng nghiên cứu này bằng cách thêm vào bộ dữ liệu các biến mới và thực hiện thuật toán phân cụm khác như phân tích thành phần chính (PCA), phân cụm theo phân cấp hoặc thuật toán DBSCAN (Density-based spatial clustering of applications with noise)[5] để có những góc nhìn khác mà thuật toán K-means không nhìn thấy. Từ đó tìm ra những phân khúc khách hàng mới cụ thể và ý nghĩa hơn.

KẾT LUẬN

Tóm lại, qua quá trình phân tích và thử nghiệm bằng phương pháp Elbow nhóm nghiên cứu đã tìm ra được số phân cụm thích hợp là 4 cụm tương ứng với 4 phân khúc khách hàng khác nhau. Từ đó tìm được một số phân khúc có ý nghĩa như:

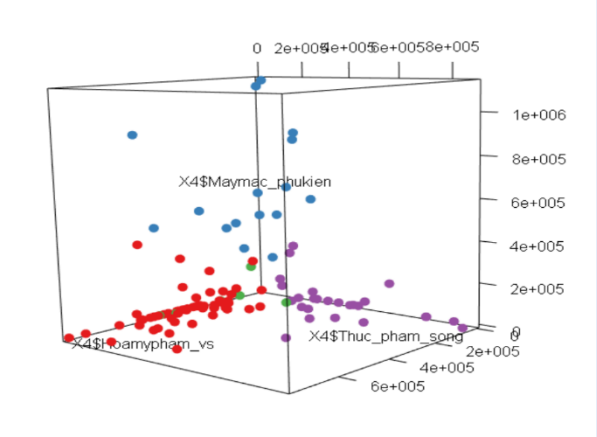
- Phân cụm 1 là những khách hàng tập trung vào mặt hàng hóa mỹ phẩm và vệ sinh.
- Phân cụm 2 tập trung vào mặt hàng đồ dùng gia đình. Đây đều là những khách hàng chi trả trên mức trung bình rất nhiều.
- Trong phân cụm 3, chúng ta cũng tìm được các phân khúc khách hàng cho nhóm đồ uống (nhóm 3.1), nhóm may mặc và phụ kiện (nhóm 3.2), nhóm thực phẩm sống (3.3), nhóm hóa mỹ phẩm và vệ sinh (nhóm 3.4, phân khúc này thấp hơn trong phân cụm 1).

Nghiên cứu phân khúc khách hàng là việc làm cần thiết đối với một công ty hay doanh nghiệp. Thông qua các phân khúc khách hàng trên phần nào giúp doanh nghiệp tìm hiểu, nắm bắt được hành vi mua sắm của khách hàng để có những giải pháp riêng, chiến lược quảng cáo, tiếp thị và dịch vụ chăm sóc khách hàng hiệu quả với sự khác biệt dù là nhỏ trong mỗi nhóm khách hàng.

	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<db1>	<chr>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
1	7	Vang	0	0	202800	0	1134000	0	100000	0	0	0	0	0	0
2	21	Vang	67800	0	0	218000	0	748600	549100	50700	0	175200	294500	0	35490
3	24	Bac	0	0	0	171380	0	112200	453800	0	0	546700	0	0	173162
4	27	Vang	0	0	0	0	93000	159400	63000	0	14732.	258900	168000	12000	203548
5	28	Khong	0	0	0	0	55500	60000	188500	0	61338	177600	152200	0	0
6	29	Vang	126500	0	163000	34500	0	114000	248700	0	0	224800	111500	21400	94700
7	30	Vang	0	0	35700	0	0	56500	157000	0	0	0	214200	24450	596688.
8	34	Vang	0	0	0	56600	0	192000	0	227000	81979.	0	169800	0	137176
9	43	Vang	0	316800	0	164700	430100	0	0	413900	0	0	166300	16000	0
10	45	Khong	23100	0	0	19900	0	0	587600	0	0	30000	0	0	0

... with 95 more rows

Hình 9: **Dữ liệu của phân cụm 3.**



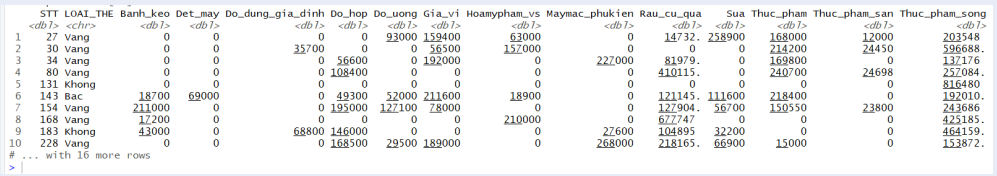
Hình 10: **Kết quả phân cụm của cụm 3.**

	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<db1>	<chr>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
1	7	Vang	0	0	202800	0	1134000	0	100000	0	0	0	0	0	0
2	124	Vang	0	0	0	0	793000	0	185000	0	0	326400	0	0	0
3	139	Bac	0	0	580900	106100	588000	0	388500	0	74600	25200	0	35376	23142
4	223	Khong	0	0	0	0	1062600	0	0	0	0	0	0	0	0
5	400	Khong	0	0	0	0	1192500	0	0	0	0	0	0	0	0
6	407	Vang	0	0	0	0	717800	0	436000	0	0	0	0	0	0
7	663	Vang	152800	0	0	120400	599800	0	45500	142000	94108.	284000	0	0	0
8	671	Vang	134500	0	327500	250200	548500	0	90500	0	55970.	41690	90900	0	206955.

Hình 11: **Dữ liệu của nhóm 3.1.**

	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<db1>	<chr>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
1	43	Vang	0	316800	0	164700	430100	0	0	413900	0	0	166300	16000	0
2	47	Vang	0	0	37600	62500	0	13800	647100	934800	96898.	21400	231200	0	181103.
3	106	Vang	0	0	0	161400	0	235400	261000	449000	39204	0	42800	0	168381
4	152	Khong	0	0	148500	0	512000	319100	0	1130000	13900	213600	0	0	24780
5	157	Vang	0	0	177000	0	0	0	36000	869900	0	0	123550	7900	200216
6	166	Vang	0	0	29000	17400	0	40000	0	828000	15346	139000	0	60000	160050
7	167	Khong	34900	0	20500	255300	171000	30000	261000	488800	0	12400	102300	44200	0
8	229	Vang	59100	0	733242.	131000	168000	0	442500	442000	59364	278000	307500	21620	0
9	271	Vang	0	0	152000	0	276500	52300	0	577000	98280.	0	0	33772	127900
10	279	Vang	103800	0	14800	0	106100	91200	147900	375000	0	0	0	0	0
11	357	Vang	70000	0	0	496000	33600	69800	336500	633500	82284.	11400	0	204000	347693
12	373	Khong	0	0	0	0	11400	0	0	1098500	0	0	0	0	0
13	413	Khong	0	0	588000	81000	0	0	284200	311800	87700	0	0	0	349288
14	433	Vang	0	0	0	73814	0	0	56000	444000	155924.	0	134548	67100	138374
15	504	Vang	119000	0	528500	0	134900	0	97600	259000	124322	56000	136700	55000	26000
16	667	Khong	65500	0	0	231800	51500	20600	39000	537000	0	0	0	0	275000

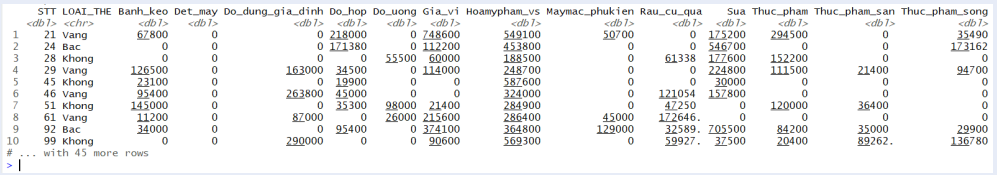
Hình 12: **Dữ liệu của nhóm 3.2.**



	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<db1>	<chr>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
1	27	Vang	0	0	0	0	93000	159400	63000	0	14732.	258900	168000	12000	203548
2	30	Vang	0	0	35700	0	0	56500	157000	0	0	0	214200	24450	596688.
3	34	Vang	0	0	0	56600	0	192000	0	227000	81979.	0	169800	0	137176
4	80	Vang	0	0	0	108400	0	0	0	0	410115.	0	240700	24698	257084.
5	131	Khong	0	0	0	0	0	0	0	0	0	0	0	0	816480
6	143	Bac	18700	69000	0	49300	52000	211600	18900	0	121145.	111600	218400	0	192010.
7	154	Vang	211000	0	0	195000	127100	78000	0	0	122904.	56700	150550	23800	243686
8	168	Vang	17200	0	0	0	0	0	210000	0	677747	0	0	0	425185.
9	183	Khong	43000	0	68800	146000	0	0	0	27600	104895	32200	0	0	464159.
10	228	Vang	0	0	0	168500	29500	189000	0	268000	218165.	66900	15000	0	153872.

... with 16 more rows

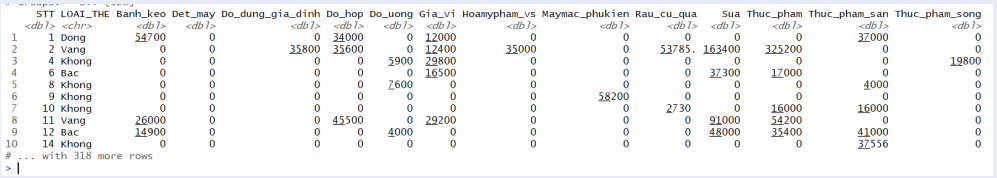
Hình 13: **Dữ liệu của nhóm 3.3.**



	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<db1>	<chr>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
1	21	Vang	67800	0	0	218000	0	748600	549100	50700	0	175200	294500	0	35490
2	24	Bac	0	0	0	171380	0	112200	453800	0	0	546700	0	0	173162
3	28	Khong	0	0	0	0	55500	60000	188500	0	61338	177600	152200	0	0
4	29	Vang	126500	0	163000	34500	0	114000	248700	0	0	224800	111500	21400	94700
5	45	Khong	23100	0	0	19900	0	0	587600	0	0	30000	0	0	0
6	46	Vang	95400	0	263800	45000	0	0	324000	0	121054	157800	0	0	0
7	51	Khong	145000	0	0	35300	98000	21400	284900	0	47250	0	120000	36400	0
8	61	Vang	11200	0	87000	0	26000	215600	286400	45000	172646.	0	0	0	0
9	92	Bac	34000	0	0	95400	0	374100	364800	129000	32589.	705500	84200	35000	29900
10	99	Khong	0	0	290000	0	0	90600	569300	0	59927.	37500	20400	89262.	136780

... with 45 more rows

Hình 14: **Dữ liệu của nhóm 3.4.**



	STT	LOAI_THE	Banh_keo	Det_may	Do_dung_gia_dinh	Do_hop	Do_uong	Gia_vi	Hoamypham_vs	Maymac_phukien	Rau_cu_qua	Sua	Thuc_pham	Thuc_pham_san	Thuc_pham_song
	<db1>	<chr>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>	<db1>
1	1	Dong	54700	0	0	34000	0	12000	0	0	0	0	0	37000	0
2	2	Vang	0	0	35800	35600	0	12400	35000	0	53785.	163400	325200	0	0
3	4	Khong	0	0	0	0	5900	29800	0	0	0	0	0	0	19800
4	6	Bac	0	0	0	0	0	16500	0	0	0	37300	17000	0	0
5	8	Khong	0	0	0	0	7600	0	0	0	0	0	0	4000	0
6	9	Khong	0	0	0	0	0	0	0	58200	0	0	0	0	0
7	10	Khong	0	0	0	0	0	0	0	0	2730	0	16000	16000	0
8	11	Vang	26000	0	0	45500	0	29200	0	0	0	91000	54200	0	0
9	12	Bac	14900	0	0	0	4000	0	0	0	0	48000	35400	41000	0
10	14	Khong	0	0	0	0	0	0	0	0	0	0	0	37556	0

... with 318 more rows

Hình 15: **Dữ liệu của phân cụm 4.**

DANH MỤC TỪ VIẾT TẮT

WSS: (Within-cluster Sum of Square) - Tổng biến thiên bình phương khoảng cách trong cụm
PCA: Phân tích thành phần chính
DBSCAN: (Density-based spatial clustering of applications with noise) -Phân cụm theo phân cấp hoặc thuật toán

TUYÊN BỐ VỀ XUNG ĐỘT LỢI ÍCH

Nhóm tác giả xin cam đoan rằng không có bất kì xung đột lợi ích nào trong công bố bài báo.

TUYÊN BỐ ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Lê Hồng Diễn và Nguyễn Phúc Sơn đã có đóng góp chính trong việc tiến hành xử lý, phân tích dữ liệu và viết bản thảo. Phạm Hoàng Uyên và Lê Văn Hinh đã có đóng góp chính trong quá trình tổ chức và thu thập dữ liệu.

CÁM ƠN

Nhóm tác giả chân thành cảm ơn sự hỗ trợ của đại sứ quán Ireland tại Hà Nội đã tài trợ kinh phi cho bài báo này.

TÀI LIỆU THAM KHẢO

1. Dolnicar S, Grn B, Leisch F. Market Segmentation. Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful. Springer; 2018. p. 11–22.
2. Kassambara A. Practical guide to cluster analysis in R: unsupervised machine learning. In: STHDA; 2017. .
3. Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu A, et al. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002;7:881–92.
4. Khan SS, Ahmad A. Ahmad AJPrl. Cluster center initialization algorithm for K-means clustering. Pattern Recognition Letters. 2004;25(11):1293–302.
5. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Ester M, Kriegel HP, Sander J, Xu X, editors. Proceedings of the SecondInternational Conference on Knowledge Discovery andData Mining (KDD-96). AAAI Press; 1996. p. 226–231.