TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

MÔN HỌC: KHAI PHÁ DỮ LIỆU

KHAI THÁC DỮ LIỆU NHẰM XÁC ĐỊNH CÁC PHÂN KHÚC KHÁCH HÀNG TRONG DOANH NGHIỆP BÁN LẺ

NHÓM 10

Danh sách nhóm:

K204061394 Nguyễn Thị Bảo Hà (nhóm trưởng)

K204061392 Cù Thị Mỹ Duy

K204061389 Nguyễn Mỹ Dung

K204061390 Nguyễn Thị Mỹ Dung

K204060282 Nguyễn Thị Cẩm Giang

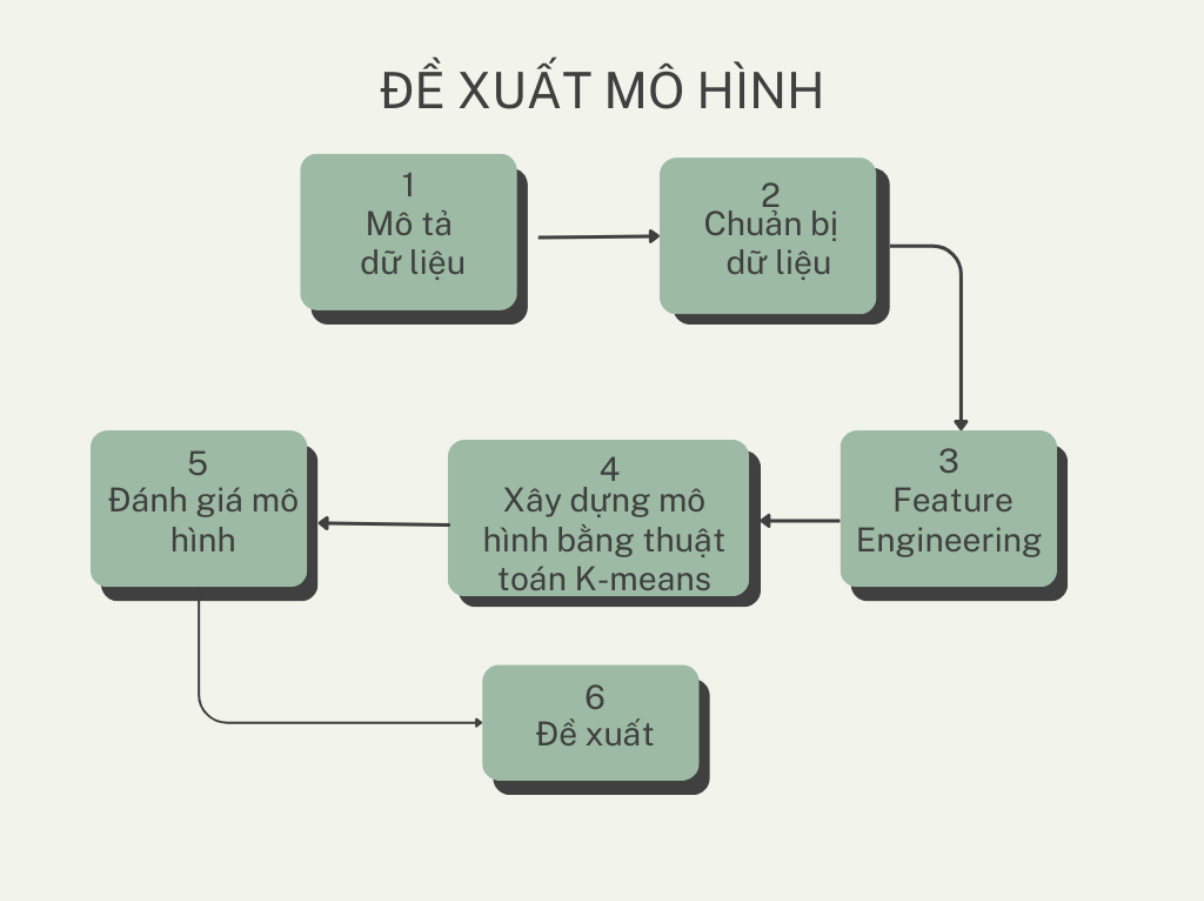
Giảng viên hướng dẫn:

Tiến sĩ: Nguyễn Thôn Dã

Tp Hồ Chí Minh, 2023

Mục lục

A. Đề xuất mô hình



1. Mô tả dữ liệu

Đầu tiên, chúng ta cần tìm ra tập dữ liệu Kaggle tập trung vào các phân khúc khách hàng khác nhau trong các giao dịch của một doanh nghiệp bán lẻ. Ở phần này, nhóm sẽ đưa ra mô tả chung về tập dữ liệu, đề cập đến các chi tiết như số hàng và cột cũng như xác định các cột có nhãn mục tiêu. Ngoài ra, nhóm cũng sẽ xác định và hình dung mối quan hệ giữa các biến độc lập và phụ thuộc.

2. Chuẩn bị dữ liệu

Giai đoạn này tập trung vào việc chuẩn bị tập dữ liệu thu được để phân tích và áp dụng các mô hình học máy bằng cách làm sạch và chuyển đổi nó. Các bước cơ bản của giai đoạn được liệt kê dưới đây:

Để đảm bảo rằng tập dữ liệu được tổ chức tốt, không có lỗi và được cấu trúc chính xác để làm đầu vào cho các thuật toán học máy, trước tiên chúng tôi áp dụng Xử lý các giá

trị bị thiếu, Xử lý các giá trị ngoại lệ, Chuyển đổi bao gồm Chuẩn hóa và Tiêu chuẩn hóa cho tập dữ liệu.

Sau khi chuẩn bị dữ liệu, chúng tôi áp dụng thuật toán K-Means để phân cụm khách hàng khi có bộ dữ liệu đầu ra chất lượng cao.

3. Feature Engineering

Phân tích RFM (Recency, Frequency, Monetary): mô hình phân tích và phân khúc khách hàng theo các đặc điểm hành vi tiêu dùng dựa trên các giao dịch.
Phân tích các biến để hiểu rõ, loại bỏ các giá trị trong biến bị sai lệch, qua đó xác định mức độ quan trọng của các tính năng.
Bước này cho phép đào tạo và đánh giá hiệu quả các mô hình học máy, đồng thời cải thiện hiệu suất và khả năng diễn giải của mô hình

4. Phân cụm K Means

Sử dụng các phương pháp: Elbow, Shihouse để xác định k phù hợp.
Sử dụng thuật toán K Means gọi từ thư viện sklearn để xác định khách hàng thuộc phân khúc nào.

5. Đánh giá chất lượng phân cụm

Bước này rất cần thiết để xác nhận tính hiệu quả của việc phân cụm và đảm bảo rằng các cụm được mạch lạc và tách biệt tốt. Các số liệu đánh giá và kỹ thuật trực quan mà nhóm sử dụng: 3D Visualization of Top PCs, Cluster Distribution Visualization, Evaluation Metrics.

6. Đề xuất

Khi doanh nghiệp biết được khách hàng thuộc phân khúc nào, doanh nghiệp sẽ có hướng tiếp cận phù hợp hơn, cải thiện vòng đời của khách hàng, tăng sự tin cậy và trung thành thông qua các tương tác hợp lý.

B. Kết quả đánh giá và phân tích

1. Thiết lập và khởi tạo

1.1 Nhập các thư viện cần thiết

Trước hết, nhập tất cả các thư viện cần thiết mà chúng tôi sẽ sử dụng trong suốt dự án bao gồm các thư viện để thao tác dữ liệu, trực quan hóa dữ liệu và các thư viện khác dựa trên nhu cầu cụ thể của dự án.

1.2 Tải tập dữ liệu

Mô tả tập dữ liệu:

Biến	**Mô tả**
InvoiceNo	Mã đại diện cho mỗi giao dịch duy nhất. Nếu mã này bắt đầu bằng chữ cái 'c' thì nó biểu thị việc hủy.
StockCode	Mã được gán duy nhất cho từng sản phẩm riêng biệt.
Description	Mô tả của từng sản phẩm.
Quantity	Số lượng đơn vị của một sản phẩm trong một giao dịch.
InvoiceDate	Ngày và thời gian của giao dịch.
UnitPrice	Đơn giá của sản phẩm bằng đồng bảng Anh.
CustomerID	Mã định danh được gán duy nhất cho mỗi khách hàng.
Country	Quốc gia của khách hàng.

2. Phân tích dữ liệu ban đầu

2.1 Tổng quan về Tập dữ liệu

Cấu trúc và các loại cột dữ liệu:

RangeIndex: 541909 entries, 0 to 541908

Data columns (total 8 columns):

#	Column	Non-Null Count	Dtype

--- ------ -------------- -----

0 InvoiceNo 541909 non-null object

1 StockCode 541909 non-null object

2 Description 540455 non-null object

3 Quantity 541909 non-null int64

4 InvoiceDate 541909 non-null object

5 UnitPrice 541909 non-null float64

6 CustomerID 406829 non-null float64

7 Country 541909 non-null object

Từ tổng quan sơ bộ, có vẻ như thiếu giá trị trong cột Description và CustomerID cần được giải quyết. Cột InvoiceDate đã có định dạng ngày giờ, điều này sẽ tạo điều kiện thuận lợi cho việc phân tích chuỗi thời gian sâu hơn. Chúng tôi cũng quan sát thấy rằng một khách hàng có thể có nhiều giao dịch được suy ra từ CustomerID lặp lại ở các hàng đầu tiên.

2.2 Thống kê tóm tắt

Thống kê tóm tắt cho các biến số:

	count	mean	std	min	25%	50%	75%	max
Quantity	541909.0	9.552250	218.081158	-80995.00	1.00	3.00	10.00	80995.0
UnitPrice	541909.0	4.611114	96.759853	-11062.06	1.25	2.08	4.13	38970.0
CustomerID	406829.0	15287.690570	1713.600303	12346.00	13953.00	15152.00	16791.00	18287.0

- Quantity:
  + Số lượng sản phẩm trung bình trong một giao dịch là khoảng 9,55.
  + Số lượng có phạm vi rộng, với giá trị tối thiểu là -80995 và giá trị tối đa là 80995. Giá trị âm cho biết các đơn hàng bị trả lại hoặc bị hủy, cần được xử lý thích hợp.
  + Độ lệch chuẩn khá lớn, cho thấy sự phân tán đáng kể trong dữ liệu. Sự hiện diện của các ngoại lệ được biểu thị bằng sự khác biệt lớn giữa giá trị phân vị tối đa và 75.
- UnitPrice:
  + Đơn giá trung bình của các sản phẩm là khoảng 4,61.

+ Đơn giá cũng hiển thị trong phạm vi rộng, từ -11062,06 đến 38970, điều này cho thấy có sai sót hoặc nhiễu trong dữ liệu vì giá âm không có ý nghĩa gì.
+ Tương tự như cột Số lượng, sự hiện diện của các giá trị ngoại lệ được biểu thị bằng sự chênh lệch lớn giữa giá trị phân vị tối đa và giá trị phân vị thứ 75.

- CustomerID:
  + Có 406829 mục nhập không rỗng, biểu thị các giá trị bị thiếu trong tập dữ liệu cần được giải quyết.
  + ID khách hàng nằm trong khoảng từ 12346 đến 18287, giúp xác định các khách hàng duy nhất.

Thống kê tóm tắt cho các biến phân loại

	count	unique	top	freq
InvoiceNo	541909	25900	573585	1114
StockCode	541909	4070	85123A	2313
Description	540455	4223	WHITE HANGING HEART T-LIGHT HOLDER	2369
InvoiceDate	541909	23260	10/31/2011 14:41	1114
Country	541909	38	United Kingdom	495478

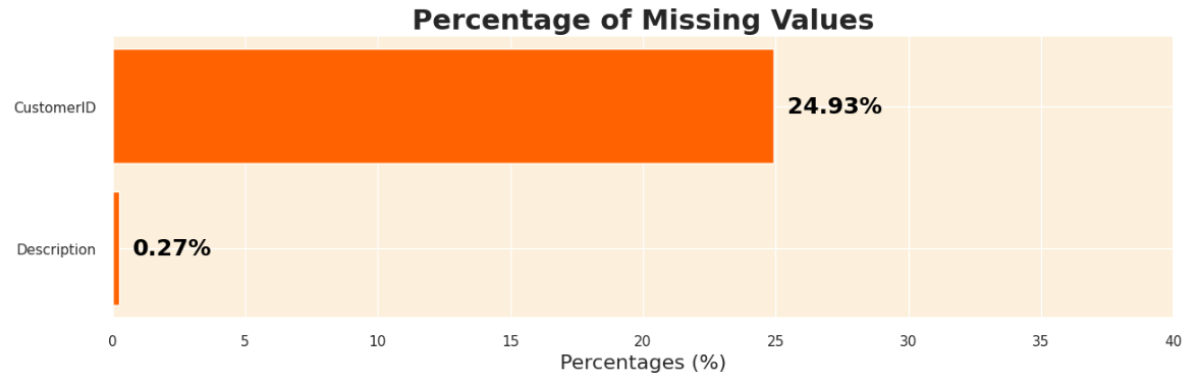
- InvoiceNo:
  + Có 25900 số hóa đơn duy nhất, biểu thị 25900 giao dịch riêng biệt.
  + Số hóa đơn thường xuyên nhất là 573585, xuất hiện 1114 lần, có thể đại diện cho một giao dịch lớn hoặc một đơn đặt hàng có nhiều mặt hàng.
- StockCode:
  + Có 4070 mã chứng khoán duy nhất đại diện cho các sản phẩm khác nhau.
  + Mã chứng khoán thường xuyên nhất là 85123A, xuất hiện 2313 lần trong tập dữ liệu.
- Description:
  + Có 4223 mô tả sản phẩm độc đáo.
  + Mô tả sản phẩm thường xuyên nhất là 'WHITE HANGING HEART T-LIGHT HOLDER', xuất hiện 2369 lần.
  + Có một số giá trị bị thiếu trong cột này cần được xử lý.

- Country:
    + Các giao dịch đến từ 38 quốc gia khác nhau, với phần lớn giao dịch (khoảng 91,4%) có nguồn gốc từ Vương quốc Anh.

3. Làm sạch và chuyển đổi dữ liệu

Bước này bao gồm quá trình làm sạch và chuyển đổi toàn diện để tinh chỉnh tập dữ liệu. Nó bao gồm việc giải quyết các giá trị còn thiếu, loại bỏ các mục trùng lặp, sửa các điểm bất thường trong mã và mô tả sản phẩm cũng như các điều chỉnh cần thiết khác để chuẩn bị dữ liệu cho phân tích và lập mô hình chuyên sâu.

3.1 Xử lý các giá trị bị thiếu



Tỷ lệ phần trăm các giá trị còn thiếu

- CustomerID (thiếu 24.93% giá trị)
- Description (thiếu 0,27% giá trị)

Bằng cách xóa các hàng có giá trị bị thiếu trong cột CustomerID và Mô tả, chúng tôi mong muốn xây dựng một tập dữ liệu rõ ràng hơn và đáng tin cậy hơn, điều này rất cần thiết để đạt được khả năng phân cụm chính xác và tạo ra một hệ thống đề xuất hiệu quả.

3.2 Xử lý trùng lặp

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
494	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	12/1/2010 11:45	1.25	17908.0	United Kingdom
517	536409	21866	UNION JACK FLAG LUGGAGE TAG	1	12/1/2010 11:45	1.25	17908.0	United Kingdom
485	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	12/1/2010 11:45	4.95	17908.0	United Kingdom
539	536409	22111	SCOTTIE DOG HOT WATER BOTTLE	1	12/1/2010 11:45	4.95	17908.0	United Kingdom
489	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	12/1/2010 11:45	2.10	17908.0	United Kingdom
527	536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	12/1/2010 11:45	2.10	17908.0	United Kingdom
521	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	12/1/2010 11:45	2.95	17908.0	United Kingdom
537	536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	12/1/2010 11:45	2.95	17908.0	United Kingdom
578	536412	21448	12 DAISY PEGS IN WOOD BOX	1	12/1/2010 11:49	1.65	17920.0	United Kingdom
598	536412	21448	12 DAISY PEGS IN WOOD BOX	1	12/1/2010 11:49	1.65	17920.0	United Kingdom

Trong dự án này, sự hiện diện của các hàng hoàn toàn giống nhau, bao gồm cả thời gian giao dịch giống hệt nhau, cho thấy rằng đây có thể là lỗi ghi dữ liệu chứ không phải là các giao dịch lặp lại thực sự. Việc giữ các hàng trùng lặp này có thể gây ra nhiễu và tiềm ẩn sự thiếu chính xác trong hệ thống phân cụm và đề xuất.

Do đó, tôi sẽ xóa những hàng trùng lặp hoàn toàn giống hệt này khỏi tập dữ liệu. Việc xóa các hàng này sẽ giúp đạt được tập dữ liệu rõ ràng hơn, từ đó sẽ hỗ trợ xây dựng các cụm khách hàng chính xác hơn dựa trên hành vi mua hàng riêng biệt của họ. Hơn nữa, nó sẽ giúp tạo ra một hệ thống đề xuất chính xác hơn bằng cách xác định chính xác các sản phẩm được mua nhiều nhất.

3.3 Xử lý các giao dịch bị hủy

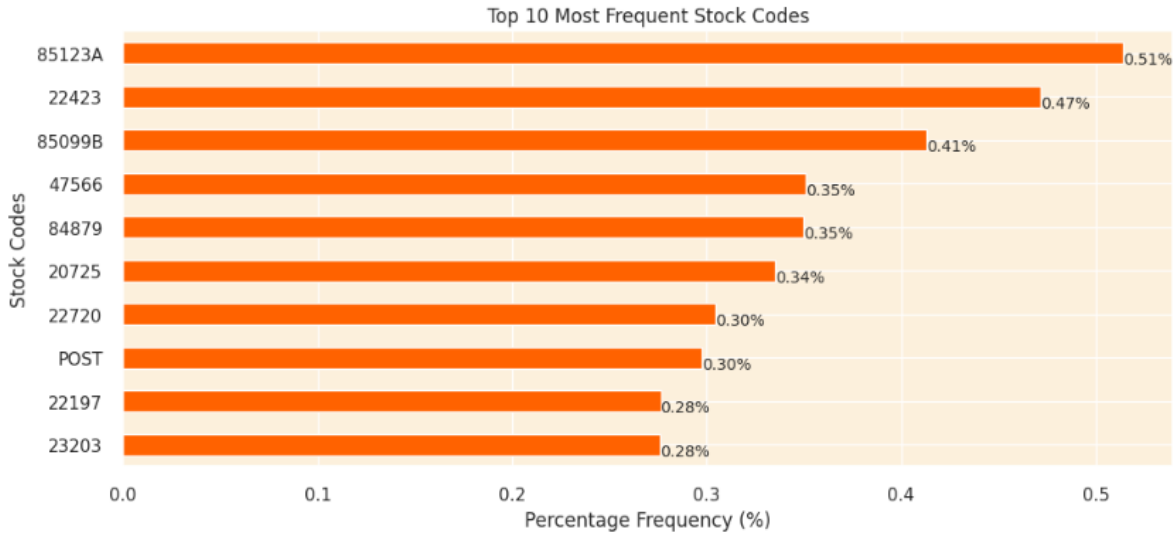
	Quantity	UnitPrice
count	8872.000000	8872.000000
mean	-30.774910	18.899512
std	1172.249902	445.190864
min	-80995.000000	0.010000
25%	-6.000000	1.450000
50%	-2.000000	2.950000
75%	-1.000000	4.950000
max	-1.000000	38970.000000

- Tất cả số lượng trong các giao dịch bị hủy đều âm, cho thấy đây thực sự là những đơn hàng đã bị hủy.

- Cột UnitPrice có độ chênh lệch đáng kể, cho thấy rất nhiều loại sản phẩm, từ giá trị thấp đến giá trị cao, đều nằm trong số giao dịch bị hủy.

3.4 Sửa lỗi bất thường của StockCode

- Số lượng mã chứng khoán duy nhất trong tập dữ liệu là: 3684
- Top 10 mã chứng khoán thường xuyên sử dụng nhất:



Suy luận về StockCode:

+ Đa dạng sản phẩm: Tập dữ liệu chứa 3684 mã chứng khoán duy nhất, cho biết rất nhiều loại sản phẩm có sẵn trong cửa hàng bán lẻ trực tuyến. Sự đa dạng này có khả năng dẫn đến việc xác định các nhóm khách hàng riêng biệt với sở thích dành cho các loại sản phẩm khác nhau.
+ Các mặt hàng phổ biến: Việc xem xét kỹ hơn 10 mã chứng khoán thường xuyên nhất có thể cung cấp thông tin chi tiết về các sản phẩm hoặc danh mục phổ biến được khách hàng thường xuyên mua.

- Số lượng giá trị của tần số ký tự số trong mã chứng khoán duy nhất:
  + 5 3676
  + 0 7
  + 1 1
- Đầu ra chỉ ra như sau:

+ Phần lớn các mã chứng khoán duy nhất (3676 trên 3684) chứa chính xác 5 ký tự số, dường như đây là định dạng tiêu chuẩn để thể hiện mã sản phẩm trong tập dữ liệu này.
+ Có một số điểm bất thường: 7 mã chứng khoán không chứa ký tự số và 1 mã chứng khoán chỉ chứa 1 ký tự số. Những điều này rõ ràng khác với định dạng tiêu chuẩn và cần điều tra thêm để hiểu bản chất của chúng cũng như liệu chúng có đại diện cho các giao dịch sản phẩm hợp lệ hay không.

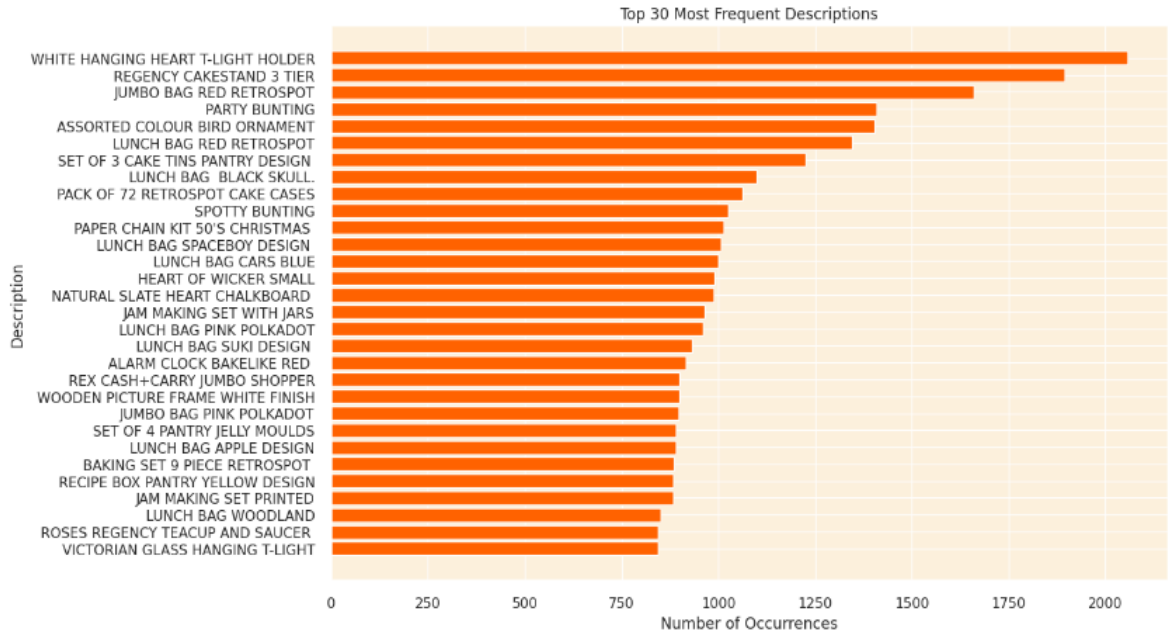
- Mã chứng khoán bất thường: POST, D, C2, M, BANK CHARGES, PADS, DOT, CRUK
- Tỷ lệ bản ghi có mã chứng khoán bất thường trong tập dữ liệu là: 0,48%

Dựa trên phân tích, chúng tôi thấy rằng một tỷ lệ rất nhỏ các bản ghi, 0,48%, có mã chứng khoán bất thường, khác với định dạng điển hình được quan sát thấy trong phần lớn dữ liệu. Ngoài ra, những mã bất thường này chỉ là một phần nhỏ trong số tất cả các mã chứng khoán duy nhất (chỉ 8 trên 3684). Chúng dường như đại diện cho các giao dịch phi sản phẩm như 'BANK CHARGES', 'POST' nên việc đưa chúng vào phân tích có thể dẫn đến gây nhiễu và bóp méo hệ thống phân cụm và đề xuất.

Do đó, lọc và loại bỏ các hàng có mã chứng khoán bất thường này khỏi tập dữ liệu trước khi tiến hành phân tích sâu hơn và phát triển mô hình.

3.5 Cột mô tả làm sạch

Đầu tiên, tôi sẽ tính số lần xuất hiện của từng mô tả duy nhất trong tập dữ liệu. Sau đó, tôi sẽ vẽ ra 30 mô tả hàng đầu. Hình ảnh trực quan này sẽ cung cấp cái nhìn rõ ràng về các mô tả xuất hiện cao nhất trong tập dữ liệu:



- Các mô tả thường xuyên nhất nói chung là đồ gia dụng, đặc biệt là những đồ dùng liên quan đến đồ dùng nhà bếp, túi đựng đồ ăn trưa và đồ trang trí.
- Điều thú vị là tất cả các mô tả đều được viết hoa, có thể là định dạng chuẩn để nhập mô tả sản phẩm vào cơ sở dữ liệu. Tuy nhiên, xét đến sự không nhất quán và bất thường gặp phải trong tập dữ liệu cho đến nay, cần thận trọng kiểm tra xem có mô tả nào được nhập bằng chữ thường hay kết hợp nhiều kiểu chữ hay không.
- Các mô tả duy nhất chứa các ký tự chữ thường là:
    + BAG 500g SWIRLY MARBLES
    + POLYESTER FILLER PAD 45x45cm
    + POLYESTER FILLER PAD 45x30cm
    + POLYESTER FILLER PAD 40x40cm
    + FRENCH BLUE METAL DOOR SIGN No
    + BAG 250g SWIRLY MARBLES
    + BAG 125g SWIRLY MARBLES
    + 3 TRADITIONAl BISCUIT CUTTERS SET
    + NUMBER TILE COTTAGE GARDEN No
    + FOLK ART GREETING CARD,pack/12
    + ESSENTIAL BALM 3.5g TIN IN ENVELOPE
    + POLYESTER FILLER PAD 65CMx65CM

+ NUMBER TILE VINTAGE FONT No
+ POLYESTER FILLER PAD 30CMx30CM
+ POLYESTER FILLER PAD 60x40cm
+ FLOWERS HANDBAG blue and orange
+ Next Day Carriage
+ THE KING GIFT BAG 25x24x12cm
+ High Resolution Image

- Suy luận:

Khi xem xét các mô tả có chứa ký tự chữ thường, rõ ràng là một số mục không phải là mô tả sản phẩm, chẳng hạn như 'Next Day Carriage' và 'High Resolution Image'. Những mục này dường như không liên quan đến sản phẩm thực tế và có thể thể hiện các loại thông tin hoặc chi tiết dịch vụ khác.

- Chiến lược:
  + Bước 1: Xóa các hàng có mô tả chứa thông tin liên quan đến dịch vụ như 'Next Day Carriage' và 'High Resolution Image', vì những hàng này không đại diện cho sản phẩm thực tế và sẽ không đóng góp vào hệ thống phân cụm và đề xuất mà chúng tôi hướng tới xây dựng.
  + Bước 2: Đối với các mô tả còn lại bằng chữ thường, hãy chuẩn hóa văn bản thành chữ hoa để duy trì tính đồng nhất trên toàn tập dữ liệu. Điều này cũng sẽ giúp giảm nguy cơ có các mục trùng lặp với các kiểu chữ khác nhau.

Bằng cách triển khai chiến lược trên, chúng tôi có thể nâng cao chất lượng của tập dữ liệu, làm cho nó phù hợp hơn với các giai đoạn phân tích và lập mô hình trong dự án của chúng tôi.

3.6 Xử lý UnitPrice bằng 0

Ở bước này, đầu tiên tôi sẽ xem mô tả thống kê của cột UnitPrice:

count 399606.000000
mean 2.904957
std 4.448796
min 0.000000

25% 1.250000

50% 1.950000

75% 3.750000

max 649.500000

Name: UnitPrice, dtype: float64

- Sự suy luận:

Giá trị đơn giá tối thiểu bằng không. Điều này cho thấy rằng có một số giao dịch có đơn giá bằng 0, có khả năng chỉ ra một mặt hàng miễn phí hoặc có lỗi nhập dữ liệu. Để hiểu bản chất của chúng, điều cần thiết là phải điều tra thêm các giao dịch có đơn giá bằng 0 này. Một phân tích chi tiết về mô tả sản phẩm có liên quan đến đơn giá bằng 0 sẽ được tiến hành để xác định xem chúng có tuân thủ một mẫu cụ thể hay không:

	Quantity
count	33.000000
mean	420.515152
std	2176.713608
min	1.000000
25%	2.000000
50%	11.000000
75%	36.000000
max	12540.000000

- Suy luận về UnitPrice:
    + Các giao dịch có đơn giá bằng 0 tương đối ít về số lượng (33 giao dịch).
    + Các giao dịch này có sự thay đổi lớn về số lượng mặt hàng liên quan, dao động từ 1 đến 12540, với độ lệch chuẩn đáng kể.

Việc đưa các giao dịch này vào phân tích phân cụm có thể gây ra nhiễu và có thể làm sai lệch các mẫu hành vi của khách hàng được xác định bởi thuật toán phân cụm.

- Chiến lược:

Với số lượng nhỏ các giao dịch này và khả năng gây nhiễu trong phân tích dữ liệu, chiến lược nên là loại bỏ các giao dịch này khỏi tập dữ liệu. Điều này sẽ giúp duy

trì một tập dữ liệu rõ ràng và nhất quán hơn, điều này rất cần thiết để xây dựng một hệ thống đề xuất và mô hình phân cụm chính xác và đáng tin cậy.

3.7 Xử lý ngoại lai

Trong phân cụm K-means, thuật toán nhạy cảm đến quy mô dữ liệu và sự hiện diện của các ngoại lệ, vì chúng có thể ảnh hưởng đáng kể đến vị trí của các trung tâm cụm và gây ra việc gán cụm không chính xác. Tuy nhiên, khi xem xét ngữ cảnh của dự án với mục tiêu cuối cùng là hiểu hành vi và sở thích của khách hàng thông qua phân cụm K-means, thì có lẽ nên xem xét việc giải quyết vấn đề về ngoại lệ sau giai đoạn kỹ thuật tính năng. Trong giai đoạn này, chúng ta tạo tập dữ liệu với khách hàng là trung tâm, và dữ liệu này có tính chất giao dịch. Loại bỏ các giá trị ngoại lệ có thể dẫn đến việc mất thông tin quan trọng có thể cần cho việc phân khúc khách hàng sau này. Vì vậy, chúng ta quyết định hoãn việc xử lý ngoại lệ và tiến hành bước tiếp theo ngay lập tức.

Sử dụng 2 phương pháp xử lý:

- IQR:
- IsolationForest:

4. Trích chọn đặc trưng (Feature Engineering):

Feature Engineering là quá trình chuyển đổi tập dữ liệu thô ban đầu thành tập các thuộc tính (features) có thể giúp biểu diễn tập dữ liệu ban đầu tốt hơn, tạo điều kiện để giải quyết các bài toán dễ dàng hơn, giúp tương thích với từng mô hình dự đoán cụ thể, cũng như cải thiện độ chính xác của mô hình dự đoán hiện tại.

4.1 Đặc điểm RFM

- RFM là một phương pháp được sử dụng để phân tích giá trị khách hàng và phân khúc cơ sở khách hàng.
- Recency (R): Số liệu cho biết khách hàng đã mua hàng gần đây như thế nào.
- Frequency (F): Số liệu biểu thị tần suất khách hàng thực hiện mua hàng trong một khoảng thời gian nhất định.
- Monetary (M): Số liệu thể hiện tổng số tiền khách hàng đã chi tiêu trong một khoảng thời gian nhất định.

⇒ Giúp hiểu được hành vi và sở thích mua hàng của khách hàng, điều này đóng đóng vai trò quan trọng trong các chiến lược tiếp thị và tạo ra hệ thống đề xuất.

4.1.1 Recency

Days Since Last Purchase: Số ngày đã trôi qua kể từ lần mua hàng cuối cùng của khách hàng.

	CustomerID	Days_Since_Last_Purchase
0	12346.0	325
1	12347.0	2
2	12348.0	75
3	12349.0	18
4	12350.0	310

4.1.2 Frequency

Total Transactions: Tổng số giao dịch được thực hiện bởi một khách hàng.

Total Products Purchased: Biết tổng số lượng sản phẩm được khách hàng mua trong tất cả các giao dịch.

	CustomerID	Days_Since_Last_Purchase	Total_Transactions	Total_Products_Purchased
0	12346.0	325	2	0
1	12347.0	2	7	2458
2	12348.0	75	4	2332
3	12349.0	18	1	630
4	12350.0	310	1	196

4.1.3 Monetary

Total Spend: Tổng số tiền mà mỗi khách hàng đã chi tiêu. Được tính bằng tích **UnitPrice** và **Quantity** cho tất cả các giao dịch được thực hiện bởi khách hàng.

Average Transaction Value: Cho biết giá trị trung bình của một giao dịch được thực hiện bởi khách hàng. Được tính bằng **Total Spend** / **Total Transactions** cho mỗi khách hàng.

	CustomerID	Days_Since_Last_Purchase	Total_Transactions	Total_Products_Purchased	Total_Spend	Average_Transaction_Value
0	12346.0	325	2	0	0.00	0.000000
1	12347.0	2	7	2458	4310.00	615.714286
2	12348.0	75	4	2332	1437.24	359.310000
3	12349.0	18	1	630	1457.55	1457.550000
4	12350.0	310	1	196	294.40	294.400000

4.2 Sự đa dạng sản phẩm

Tìm hiểu sự đa dạng trong hàng vi mua sản phẩm của khách hàng.

Unique Products Purchased: Thể hiện số lượng sản phẩm riêng biệt được khách hàng mua. Giá trị cao cho thấy khách hàng có sở thích đa dạng, mua nhiều loại sản phẩm, trong khi giá trị thấp hơn có thể cho biết sở thích tập trung hoặc cụ thể.

	CustomerID	Days_Since_Last_Purchase	Total_Transactions	Total_Products_Purchased	Total_Spend	Average_Transaction_Value	Unique_Products_Purchased
0	12346.0	325	2	0	0.00	0.000000	1
1	12347.0	2	7	2458	4310.00	615.714286	103
2	12348.0	75	4	2332	1437.24	359.310000	21
3	12349.0	18	1	630	1457.55	1457.550000	72
4	12350.0	310	1	196	294.40	294.400000	16

4.3 Phân tích biến: Behavior

Nắm bắt các mô hình và hành vi mua sắm của khách hàng.

Average Days Between Purchases: Thể hiện số ngày trung bình mà khách hàng chờ đợi được khi thực hiện một giao dịch mua hàng khác.

Favorite Shopping Day: Là ngày trong tuần mà khách hàng mua sắm nhiều nhất.

Favorite Shopping Hour: Là giờ trong ngày mà khách hàng mua sắm nhiều nhất.

	CustomerID	Days_Since_Last_Purchase	Total_Transactions	Total_Products_Purchased	Total_Spend	Average_Transaction_Value	Unique_Products_Purchased	Average_Days_Between_Purchases	Day_Of_Week	Hour
0	12346.0	325	2	0	0.00	0.000000	1	0.000000	1	10
1	12347.0	2	7	2458	4310.00	615.714286	103	2.016575	1	14
2	12348.0	75	4	2332	1437.24	359.310000	21	10.884615	3	19
3	12349.0	18	1	630	1457.55	1457.550000	72	0.000000	0	9
4	12350.0	310	1	196	294.40	294.400000	16	0.000000	2	16

4.4 Phân tích biến Geographic

Địa điểm địa lý phản ánh vị trí địa lý của khách hàng.

Country: Xác định quốc gia nơi mỗi khách hàng sinh sống. Các khu vực khác nhau có thể có sở thích và hành vi mua hàng khác nhau.

```
[ ]  df['Country'].value_counts(normalize=True).head()
```

```
United Kingdom    0.890971
Germany           0.022722
France            0.020402
EIRE              0.018440
Spain             0.006162
Name: Country, dtype: float64
```

4.5 Phân tích về sự hủy bỏ đơn hàng: Cancellation

Hủy đơn hàng của khách hàng dựa vào:

Cancellation Frequency (Tần suất hủy): Biểu thị tổng số giao dịch mà khách hàng đã hủy.

Cancellation Rate (Tỷ lệ hủy): Thể hiện tỷ lệ giao dịch mà khách hàng đã hủy trong số tất cả các giao dịch của họ.

Cancellation_Frequency	Cancellation_Rate
1.0	0.5
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0

4.6 Phân tích xu hướng tiêu dùng:

Tính thời vụ và xu hướng trong hành vi mua hàng của khách hàng:

Monthly Spending Mean: Số tiền trung bình mà khách hàng chi tiêu hàng tháng. Đánh giá được thói quen chi tiêu chung của từng khách hàng.

Monthly Spending Std: Cho biết sự thay đổi trong chi tiêu hàng tháng của khách hàng.

Spending Trend: Phản ánh xu hướng chi tiêu của khách hàng theo thời gian, được tính bằng độ dốc của đường xu hướng tuyến tính phù hợp với dữ liệu chi tiêu của họ.

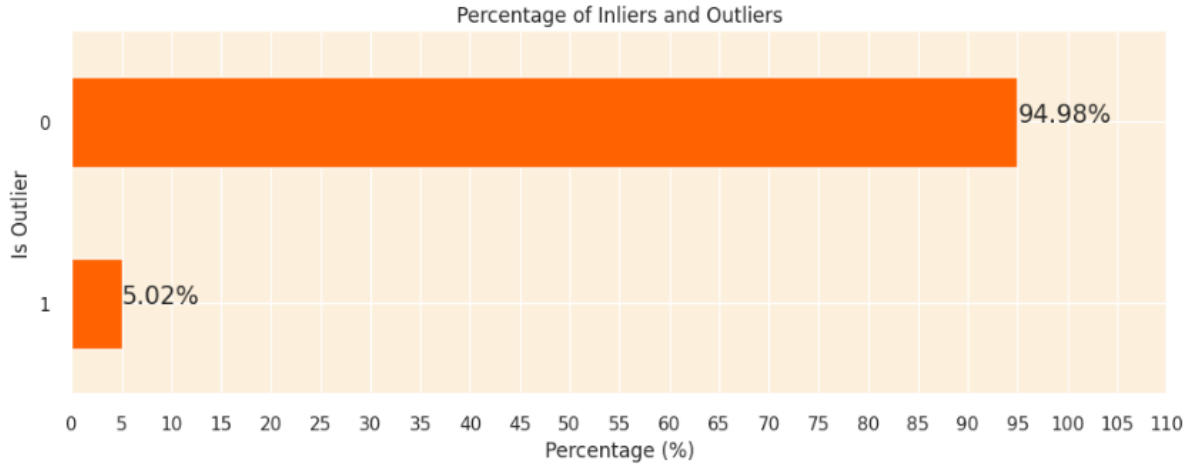
Monthly_Spending_Mean	Monthly_Spending_Std	Spending_Trend
0.0	0.0	0.0
615.714286	341.070789	4.486071
359.31	203.875689	-100.884
1457.55	0.0	0.0
294.4	0.0	0.0
316.3525	134.700629	9.351
89.0	0.0	0.0
1079.4	0.0	0.0
459.4	0.0	0.0
829.143333	991.462585	-944.635

4.7 Xác định và xử lý ngoại lai

Xác định và xử lý các ngoại lai trong tập dữ liệu.

Outlier_Scores	Is_Outlier
1	0
1	0
1	0
1	0
1	0

⇒ Áp dụng thuật toán Isolation Forest, xác định được các ngoại lệ và đánh dấu chúng trong cột mới tên Is Outlier.



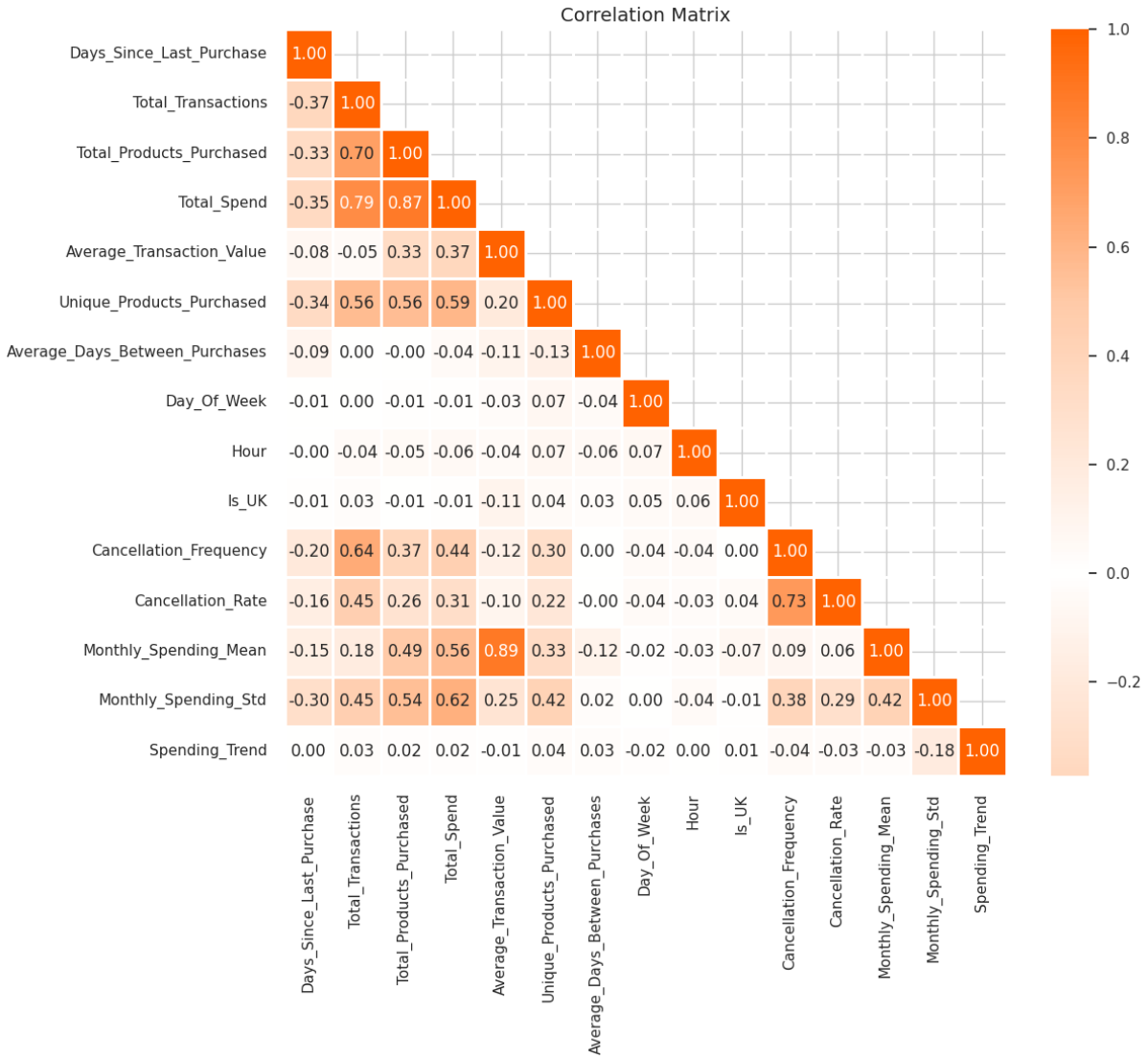
4.8 Phân tích tương quan

Trước khi tiến hành phân cụm KMeans, điều cần thiết là phải kiểm tra mối tương quan giữa các tính năng trong tập dữ liệu.

Sử dụng Kỹ thuật PCA giúp vô hiệu tác động của đa cộng tuyến bằng cách cách chuyển đổi các đặc điểm tương quan thành một tập hợp các biến mới không tương quan.

Giúp nâng cao chất lượng các cụm được thành và làm cho quá trình phân cụm hiệu quả hơn về mặt tính toán.

PCA (Principal Component Analysis) là phương pháp tuyến tính thông dụng nhất để giảm chiều dữ liệu. PCA thực hiện một phép biến đổi tuyến tính, chuyển dữ liệu sang không gian thấp chiều hơn. Khi thực hiện biến đổi, PCA tối đa hóa variance dữ liệu ở không gian thấp chiều.



⇒ Heatmap, thấy được các cặp biến có độ tương quan cao:

- Monthly_Spending_Mean and Average_Transaction_Value
- Total_Spend and Total_Products_Purchased
- Total_Transactions and Total_Spend
- Cancellation_Rate and Cancellation_Frequency
- Total_Transactions and Total_Products_Purchased

Những mối tương quan cao này cho thấy các biến này có mối quan hệ chặt chẽ với nhau, hàm ý mức độ đa cộng tuyến.

4.9 Chuẩn hóa dữ liệu

Trước khi tiến hành phân cụm và giảm kích thước, điều bắt buộc chuẩn hóa dữ liệu. Bước này có tầm quan trọng đáng kể, đặc biệt là khi áp dụng thuật toán k-means và sử dụng phương pháp PCA:

- Phân cụm k-means: Kmeans phụ thuộc rất nhiều vào khoảng cách của các điểm dữ liệu để tạo chúng thành các cụm, ví dụ các điểm dữ liệu không có tỷ lệ tương tự, hay có giá trị lớn ảnh hưởng không tương xứng đến kết quả phân cụm, các điểm dữ liệu này có khả năng dẫn đến việc phân nhóm không chính xác.
- PCA(Principal component analysis): Đây là thuật toán học máy không giám sát, làm giảm kích thước của tập dữ liệu trong khi vẫn giữ lại nhiều thông tin nhất có thể. Để làm điều này, thuật toán tạo một tập hợp các tính năng mới từ tập hợp các tính năng hiện có.

Để đảm bảo ảnh hưởng cân bằng đến mô hình và các mẫu ảnh hưởng cao trong dữ liệu, chúng ta sẽ bắt đầu chuẩn hóa dữ liệu, nghĩa là chuyển đổi các đặc điểm để có giá trị trung bình bằng 0 và độ lệch chuẩn là 1. Tuy nhiên, không phải tất cả các đặc điểm yêu cầu chuẩn hóa. Dưới đây là các trường hợp ngoại lệ và lý do tại sao chúng bị

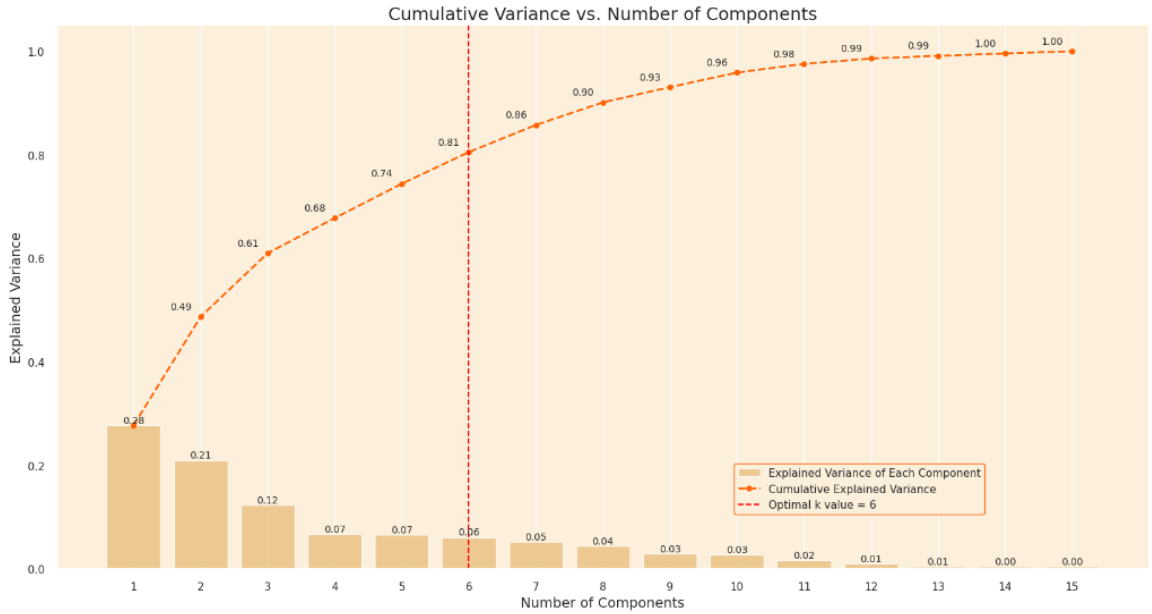
```
# List of columns that don't need to be scaled
columns_to_exclude = ['CustomerID', 'Is_UK', 'Day_Of_Week']
```

- CustomerID: chỉ là mã định danh cho khách hàng và không chứa bất kì thông tin nào để phân cụm
- Is_UK: là feature nhị phân cho biết khách hàng có đến từ Vương Quốc Anh hay không, vì nó chỉ nhận giá trị là 0 và 1, nên việc chia tỉ lệ sẽ không tạo ra bất kì sự khác biệt
- Day_of_week: 1 tính năng phân loại bằng số nguyên (1-7), nên chia tỷ lệ không cần thiết.

4.10 Giảm chiều dữ liệu

Việc giảm chiều giúp giảm hiện tượng đa cộng tuyến, phân cụm tốt hơn với K-means, giảm dữ liệu nhiễu, trực quan hóa dữ liệu, cải thiện hiệu quả.

Áp dụng PCA trên tất cả các component có sẵn và vẽ biểu đồ phương sai:



Ở đây chúng ta có thể quan sát thấy: number of component =1 thì explained variance=0,28, number of component=1 và number of component =2 là 0,49.

Từ biểu đồ, chúng ta có thể thấy rằng mức tăng của phương sai tích lũy bắt đầu chậm lại ở number of components =6 (chiếm khoảng 81% phương sai). Do đó việc giữ lại 6 number of components đầu là một lựa chọn cân bằng.

	PC1	PC2	PC3	PC4	PC5	PC6
CustomerID						
12346.0	-2.186469	-1.705370	-1.576745	1.008187	-0.411803	-1.658012
12347.0	3.290264	-1.387375	1.923310	-0.930990	-0.010591	0.873150
12348.0	0.584684	0.585019	0.664727	-0.655411	-0.470280	2.306657
12349.0	1.791116	-2.695652	5.850040	0.853418	0.677111	-1.520098
12350.0	-1.997139	-0.542639	0.578781	0.183682	-1.484838	0.062672

Sau đó trích xuất các hệ số tương ứng với từng các biến để hiểu rõ hơn về phép biến đổi PCA:

	PC1	PC2	PC3	PC4	PC5	PC6
Days_Since_Last_Purchase	-0.217859	-0.013986	0.067660	0.273430	-0.240968	-0.373059
Total_Transactions	0.380301	0.014759	-0.259180	-0.138165	-0.017356	-0.028257
Total_Products_Purchased	0.401425	0.007365	0.069133	-0.134806	0.057476	-0.013373
Total_Spend	0.431260	0.010159	0.065165	-0.092047	0.025202	-0.036947
Average_Transaction_Value	0.176225	-0.015544	0.589050	0.114307	0.021847	-0.101738
Unique_Products_Purchased	0.324992	0.063346	0.014010	-0.230502	-0.193981	0.124604
Average_Days_Between_Purchases	-0.022600	-0.036007	-0.127341	-0.160627	0.753462	0.211787
Day_Of_Week	-0.026572	0.994650	-0.006591	0.028870	0.058359	-0.060799
Hour	-0.024259	0.056388	-0.002019	-0.226832	-0.528881	0.621915
Is_UK	-0.001014	0.007435	-0.018378	-0.013419	-0.005353	0.014384
Cancellation_Frequency	0.287102	-0.018576	-0.400697	0.225923	-0.100595	-0.168050
Cancellation_Rate	0.229885	-0.022616	-0.381347	0.290702	-0.126048	-0.216073
Monthly_Spending_Mean	0.274127	-0.005116	0.498142	0.134989	-0.004123	-0.101941
Monthly_Spending_Std	0.334168	0.014557	0.036156	0.218369	0.128494	0.193648
Spending_Trend	-0.014574	-0.014845	-0.002987	-0.730466	-0.078739	-0.523692

5. Phân cụm K-means

K-Means là một thuật toán học máy không giám sát, phân cụm dữ liệu thành một số nhóm (K) được chỉ định bằng cách giảm thiểu tổng bình phương trong cụm (WCSS).

Để xác định số cụm tối ưu cho việc phân khúc khách hàng, chúng ta có 2 phương pháp phổ biến: elbow, silhouette

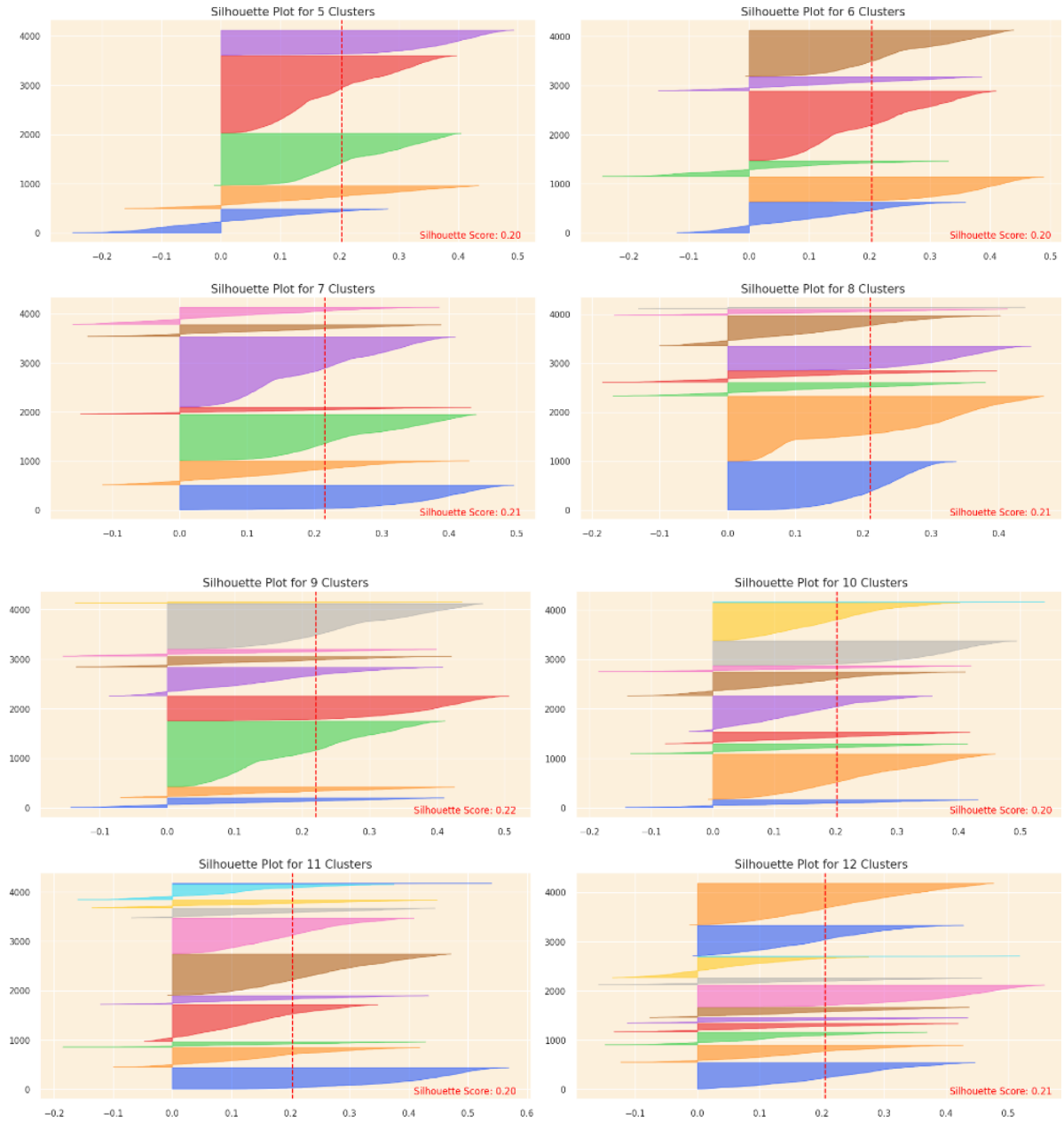
5.1. Phương pháp elbow

Phương pháp Elbow là một kỹ thuật để xác định số cụm lý tưởng trong một tập dữ liệu, tạo ra các cụm cho các giá trị khác nhau của k. Thuật toán k-mean tính tổng bình phương khoảng cách giữa mỗi điểm dữ liệu và trọng tâm cụm được chỉ định của nó, được gọi là quán tính hoặc điểm WCSS. Bằng cách vẽ điểm quán tính theo giá trị k, chúng ta tạo ra một biểu đồ thường thể hiện hình dạng khuỷu tay, do đó có tên là "Phương pháp khuỷu tay". Điểm khuỷu tay biểu thị giá trị k trong đó mức giảm quán tính đạt được khi tăng k trở nên không đáng kể, biểu thị điểm dừng tối ưu cho số lượng cụm.

Sử dụng Thư viện Yellow Brick. Trong phần này, chúng ta sẽ sử dụng thư viện Yellow Brick để tạo điều kiện thuận lợi cho việc triển khai phương pháp Elbow, chúng ta nhận thấy rằng k tối ưu được đề xuất là k=5

5.2. Phương pháp Silhouette

Phương pháp Silhouette là một cách tiếp cận để tìm số cụm tối ưu trong tập dữ liệu bằng cách đánh giá tính nhất quán trong các cụm và sự tách biệt của chúng với các cụm khác. Nó tính toán hệ số hình bóng cho từng điểm dữ liệu, đo lường mức độ tương tự của một điểm với cụm của chính nó so với các cụm khác.



Dựa trên các hướng dẫn ở trên và sau khi xem xét cẩn thận các đồ thị hình bóng, rõ ràng rằng việc chọn (k = 3) là lựa chọn tốt hơn. Lựa chọn này mang lại cho chúng tôi các cụm được kết hợp đồng đều hơn và được xác định rõ ràng hơn, giúp giải pháp phân cụm chính xác và đáng tin cậy hơn.

5.3. Mô hình phân cụm k-means:

Trong bước này, chúng ta sẽ áp dụng thuật toán phân cụm K-mean để phân nhóm khách hàng thành các cụm khác nhau dựa trên hành vi mua hàng của họ và các đặc điểm khác, sử dụng số lượng cụm tối ưu được xác định ở bước trước.

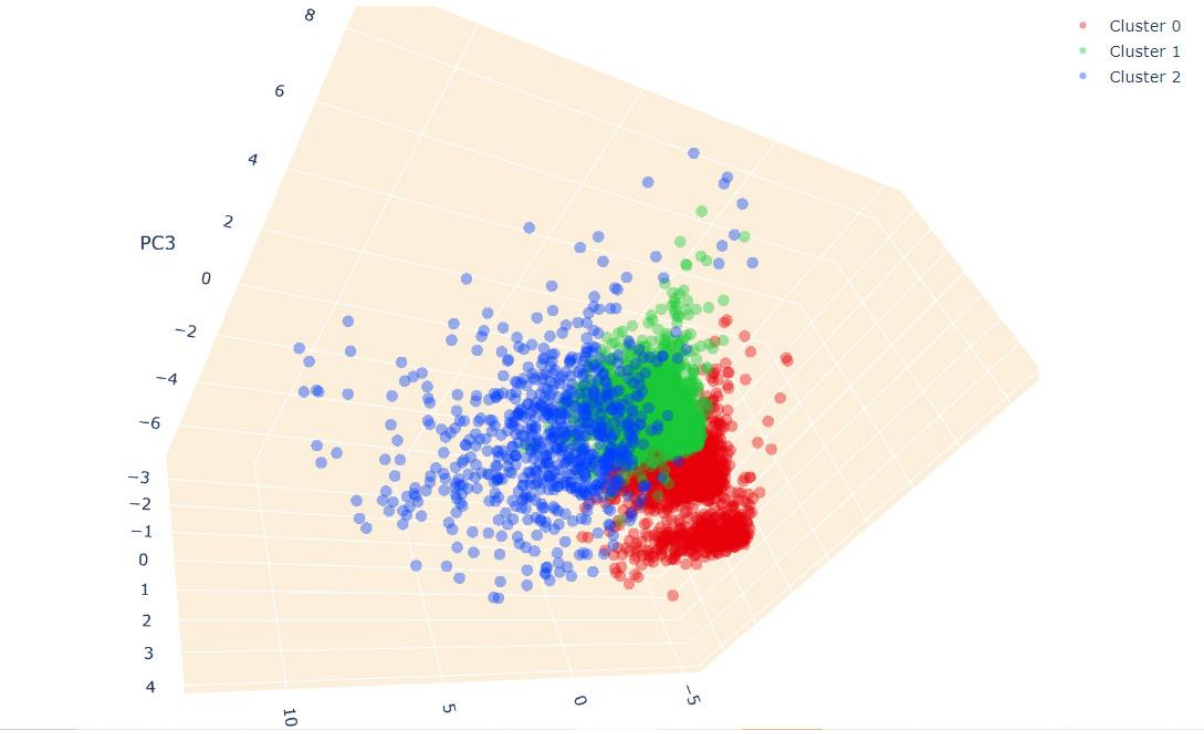
Điều quan trọng cần lưu ý là thuật toán K-mean có thể gán các nhãn khác nhau cho các cụm trong mỗi lần chạy. Để giải quyết vấn đề này, chúng ta sẽ thực hiện thêm một bước để hoán đổi nhãn dựa trên tần suất mẫu trong mỗi cụm, đảm bảo việc gán nhãn nhất quán trên các lần chạy khác nhau. Chúng ta sẽ có kết quả như sau:

	CustomerID	Days_Since_Last_Purchase	Total_Transactions	Total_Products_Purchased	Total_Spend	Average_Transaction_Value	Unique_Products_Purchased	Average_Days_Between_Purchases
0	12346.0	325	2	0	0.0	0.0	1	0.0
1	12347.0	2	7	2458	4310.0	615.714286	103	2.016575
2	12348.0	75	4	2332	1437.24	359.31	21	10.884615
3	12349.0	18	1	630	1457.55	1457.55	72	0.0
4	12350.0	310	1	196	294.4	294.4	16	0.0

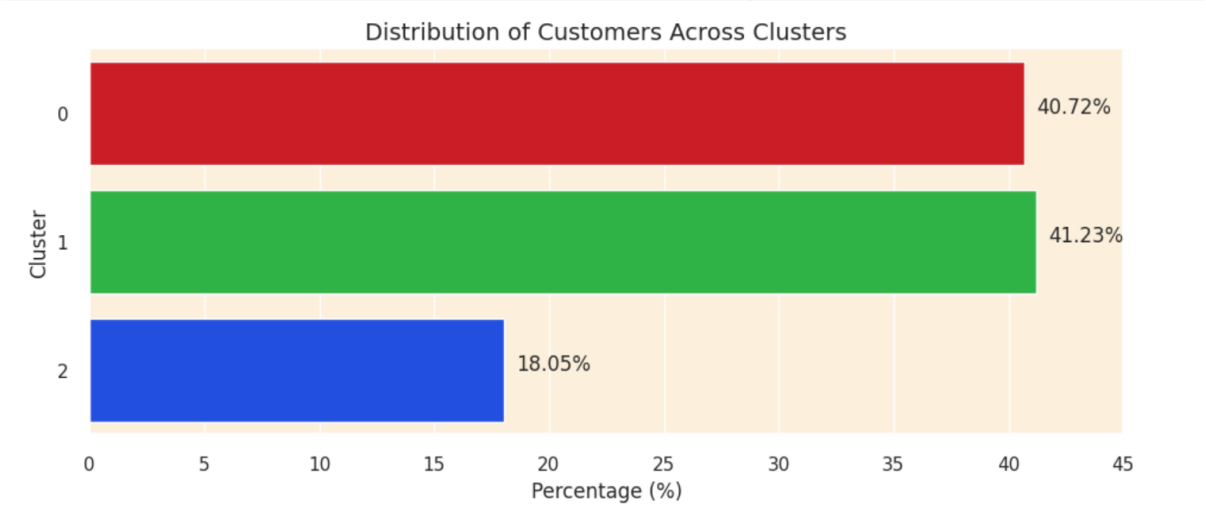
6. Đánh giá chất lượng phân cụm

Sau khi xác định số cụm tối ưu, chuyển sang bước đánh giá để đánh giá chất lượng của các cụm được hình thành. Bước này rất cần thiết để xác nhận tính hiệu quả của việc phân cụm và đảm bảo rằng các cụm được mạch lạc và tách biệt tốt. Các số liệu đánh giá và kỹ thuật trực quan mà nhóm sử dụng được nêu dưới đây:

- **3D Visualization of Top PCs**: là một quá trình tạo nội dung đồ họa bằng phần mềm 3D hiện đại. Nếu giải thích thuật ngữ này bằng những từ đơn giản thì đó là cách tái tạo một đối tượng hoặc một số đối tượng với sự trợ giúp của máy tính bằng cách tạo chế độ xem 360°.



- **Cluster Distribution Visualization**



Sự phân bố khách hàng giữa các cụm, như được mô tả bằng biểu đồ thanh, cho thấy sự phân bổ khá cân bằng với cụm 0 và 1 nắm giữ khoảng 41% số khách hàng mỗi cụm và cụm 2 chứa khoảng 18% số khách hàng.

Sự phân bổ cân bằng này cho thấy rằng quá trình phân cụm của nhóm phần lớn đã thành công trong việc xác định các mẫu có ý nghĩa trong dữ liệu, thay vì chỉ phân nhóm nhiễu hoặc các ngoại lệ. Nó ngụ ý rằng mỗi cụm đại diện cho một phân khúc cơ sở khách hàng đáng kể và riêng biệt, từ đó cung cấp những hiểu biết sâu sắc có giá trị cho các chiến lược kinh doanh trong tương lai.

Hơn nữa, thực tế là không có cụm nào chứa tỷ lệ phần trăm khách hàng rất nhỏ, đảm bảo với nhóm rằng mỗi cụm đều có ý nghĩa quan trọng và không chỉ đại diện cho các ngoại lệ hoặc nhiễu trong dữ liệu. Thiết lập này cho phép hiểu biết và phân tích nhiều sắc thái hơn về các phân khúc khách hàng khác nhau, tạo điều kiện cho việc ra quyết định hiệu quả và sáng suốt.

- **Evaluation Metrics**

Metric	Value
Number of Observations	4067
Silhouette Score	0.236271370227799 26
Calinski Harabasz Score	1257.1794962921572
Davies Bouldin Score	1.368405537299602

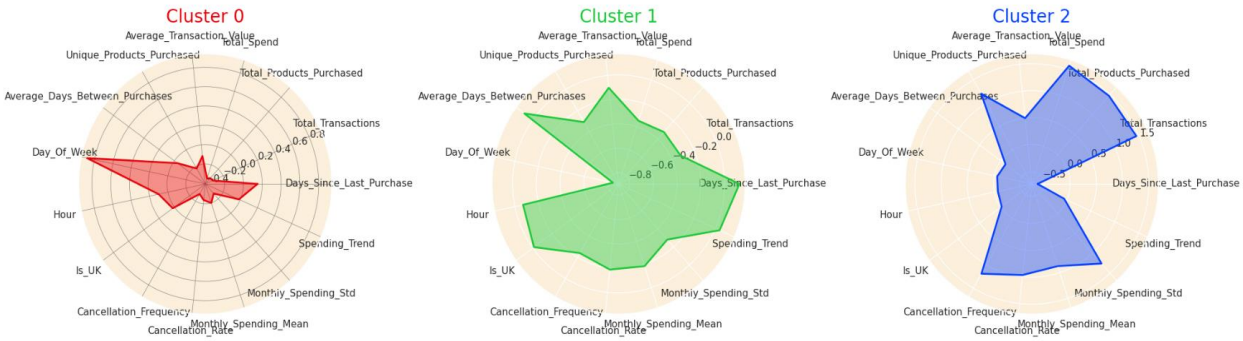
Điểm Silhouette xấp xỉ 0,236, mặc dù không gần bằng 1, nhưng vẫn cho thấy mức độ tách biệt khá lớn giữa các cụm. Nó gợi ý rằng các cụm có phần khác biệt nhưng có thể có sự chồng chéo nhẹ giữa chúng. Nói chung, điểm gần 1 sẽ là lý tưởng, biểu thị các cụm khác biệt và tách biệt rõ ràng hơn.

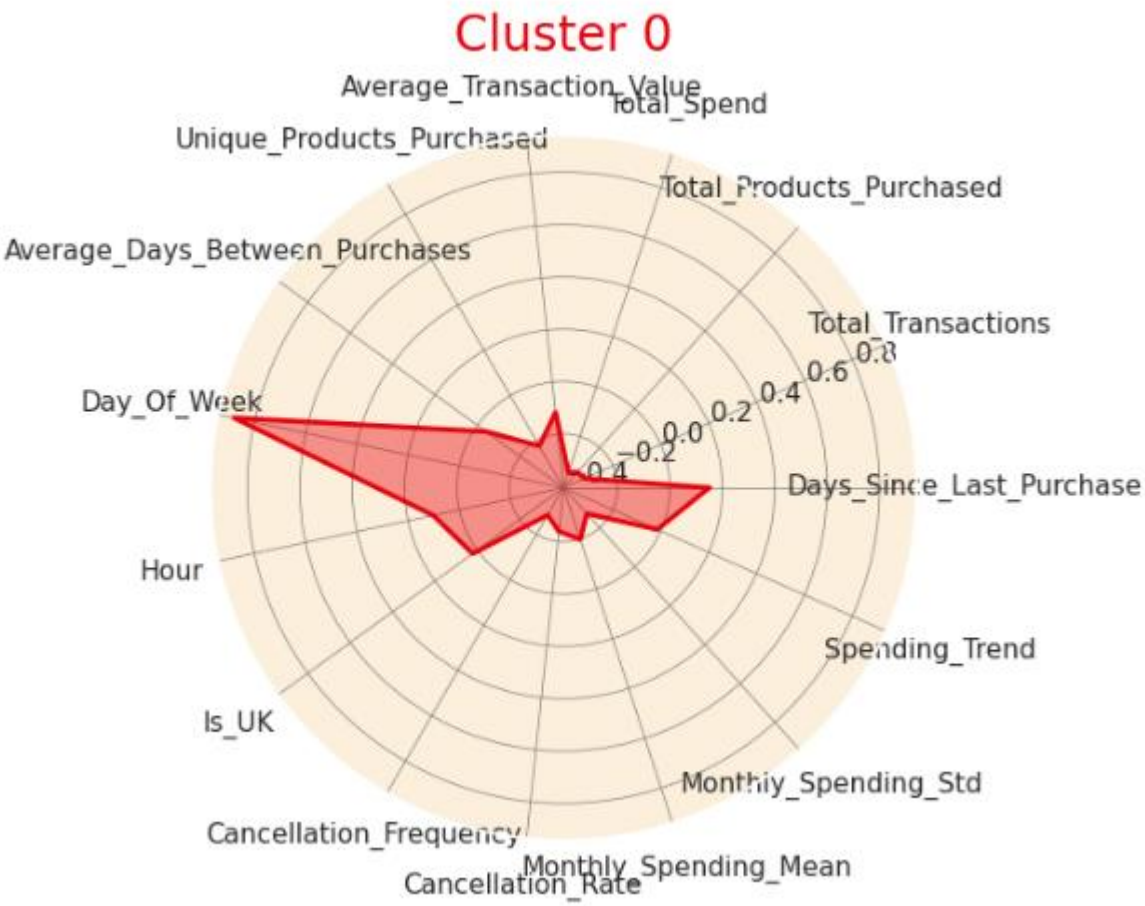
Điểm Calinski Harabasz là 1257,17, cao đáng kể, cho thấy các cụm được xác định rõ ràng. Điểm cao hơn trong số liệu này thường báo hiệu các định nghĩa cụm tốt hơn, do đó ngụ ý rằng việc phân cụm của chúng tôi đã tìm được cấu trúc quan trọng trong dữ liệu.

Điểm Davies Bouldin là 1,37 là điểm số hợp lý, cho thấy mức độ tương đồng vừa phải giữa mỗi cụm và cụm tương tự nhất. Điểm thấp hơn thường tốt hơn vì nó cho thấy ít sự tương đồng hơn giữa các cụm và do đó, điểm số của chúng tôi ở đây cho thấy sự tách biệt hợp lý giữa các cụm.

Tóm lại, các số liệu cho thấy việc phân cụm có chất lượng tốt, với các cụm được xác định rõ ràng và khá tách biệt. Tuy nhiên, vẫn có thể cần tối ưu hóa hơn nữa để tăng cường phân tách và định nghĩa cụm, có thể bằng cách thử các thuật toán phân cụm và giảm kích thước khác.

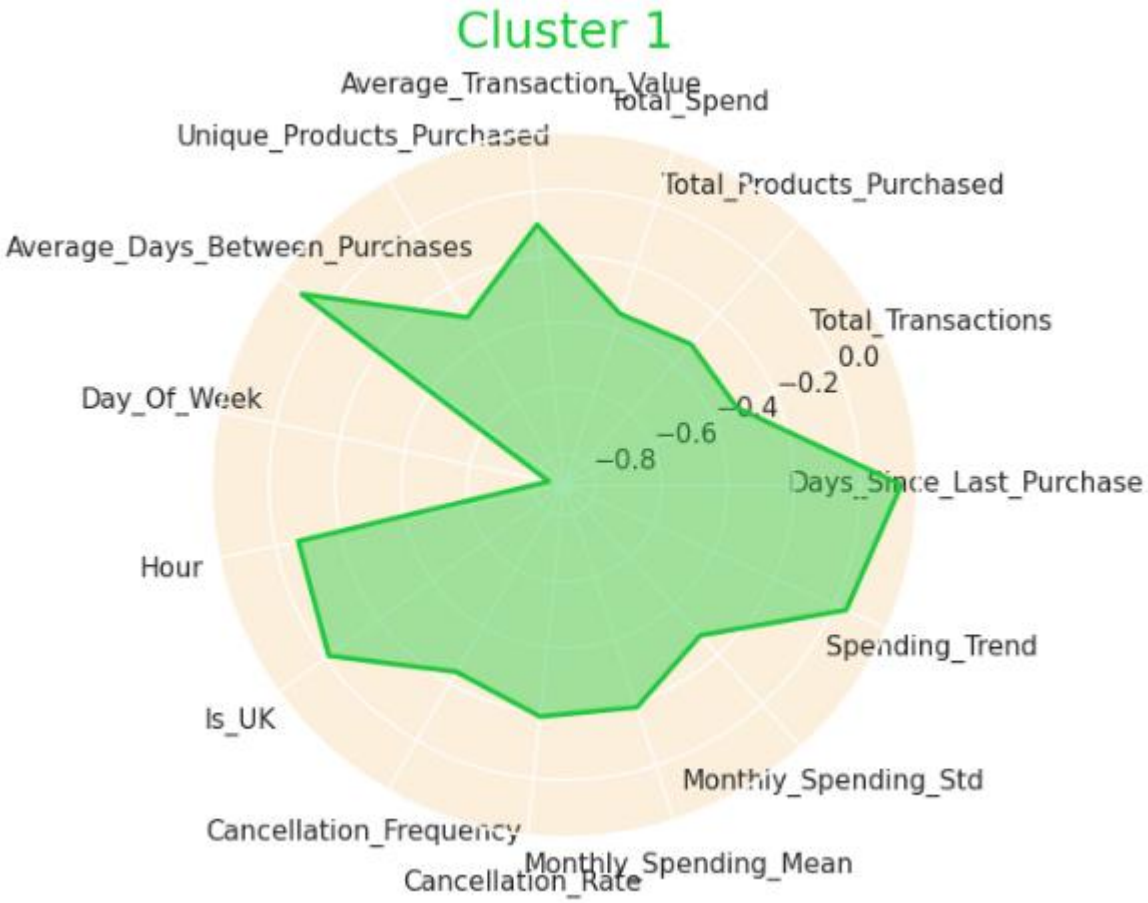
7. Cluster Analysis and Profiling (Phân tích và lập hồ sơ cụm)



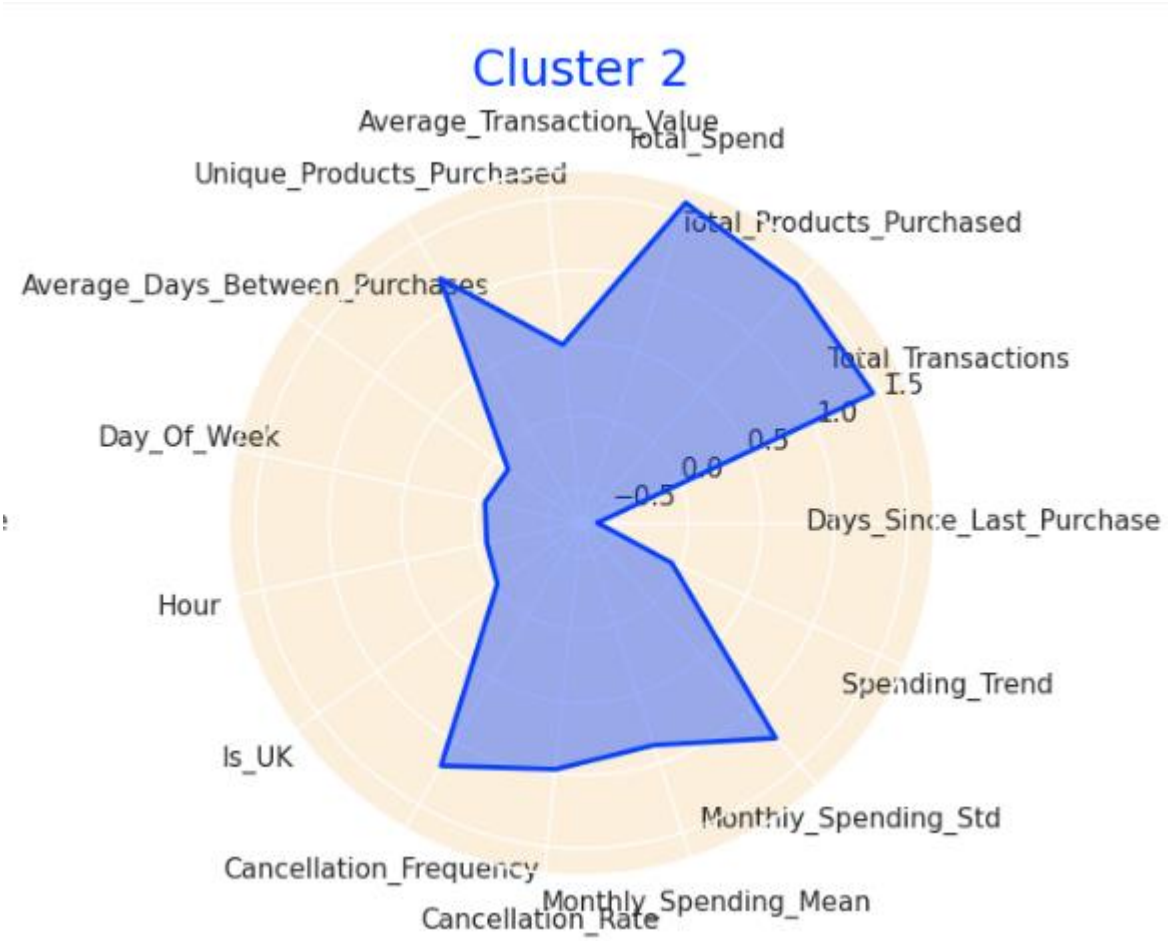


Khách hàng trong nhóm này có xu hướng chi tiêu ít hơn, với số lượng giao dịch và sản phẩm mua ít hơn. Họ có xu hướng mua sắm nhẹ vào cuối tuần, được biểu thị bằng giá trị Day_of_week rất cao. Xu hướng chi tiêu của họ tương đối ổn định nhưng ở mức thấp hơn và họ có mức chênh lệch chi tiêu hàng tháng thấp (Monthly_Spending_Mean thấp). Những khách hàng này chưa thực hiện nhiều lần hủy,

cho thấy tần suất và tỷ lệ hủy thấp. Giá trị giao dịch trung bình ở mức thấp hơn, cho thấy rằng khi mua sắm, họ có xu hướng chi tiêu ít hơn cho mỗi giao dịch.

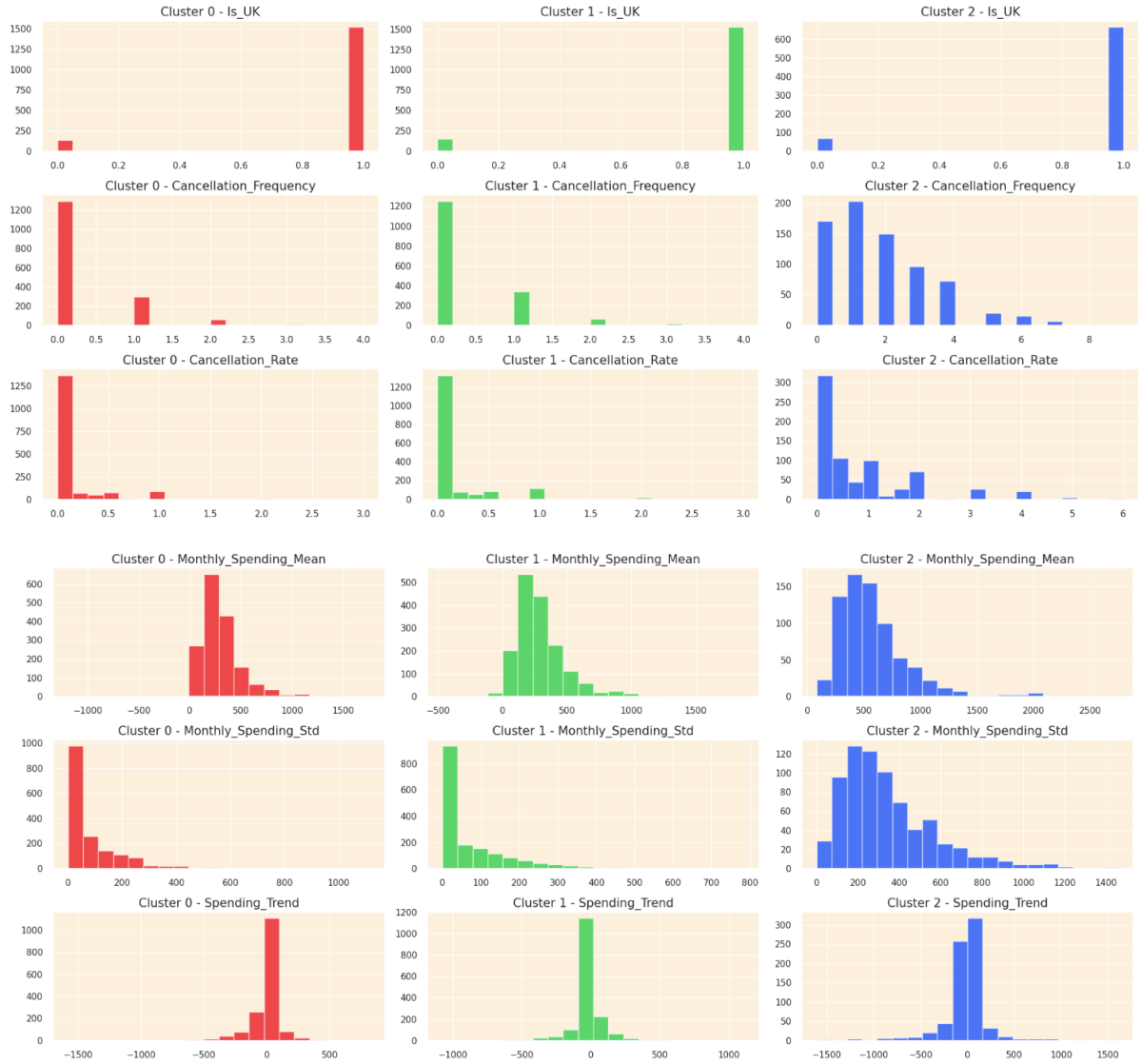


Khách hàng trong cụm này có mức chi tiêu vừa phải nhưng giao dịch của họ không thường xuyên, được biểu thị bằng Số ngày_Since_Last_Purchase và Average_Days_Between_Purchases cao. Họ có xu hướng chi tiêu rất cao, cho thấy mức chi tiêu của họ ngày càng tăng theo thời gian. Những khách hàng này thích mua sắm vào cuối ngày, được biểu thị bằng giá trị Giờ cao và họ chủ yếu cư trú ở Vương quốc Anh. Họ có xu hướng hủy một số lượng giao dịch vừa phải, với tần suất và tỷ lệ hủy trung bình. Giá trị giao dịch trung bình của họ tương đối cao, nghĩa là khi mua sắm, họ có xu hướng mua những món hàng có giá trị lớn.



Khách hàng trong nhóm này là những người chi tiêu nhiều với tổng chi tiêu rất cao và họ mua nhiều loại sản phẩm độc đáo. Họ tham gia vào các giao dịch thường xuyên nhưng cũng có tần suất và tỷ lệ hủy cao. Những khách hàng này có thời gian trung bình giữa các lần mua hàng rất thấp và họ có xu hướng mua sắm sớm trong ngày (giá trị Hour thấp). Chi tiêu hàng tháng của họ có mức độ biến động cao, cho thấy mô hình chi tiêu của họ có thể khó dự đoán hơn so với các nhóm khác. Mặc dù chi tiêu cao nhưng họ có xu hướng chi tiêu thấp, cho thấy mức chi tiêu cao của họ có thể giảm theo thời gian


Cluster 0 - Days_Since_Last_Purchase
Cluster 1 - Days_Since_Last_Purchase
Cluster 2 - Days_Since_Last_Purchase
Cluster 0 - Total_Transactions
Cluster 1 - Total_Transactions
Cluster 2 - Total_Transactions
Cluster 0 - Total_Products_Purchased
Cluster 1 - Total_Products_Purchased
Cluster 2 - Total_Products_Purchased
Cluster 0 - Total_Spend
Cluster 1 - Total_Spend
Cluster 2 - Total_Spend
Cluster 0 - Average_Transaction_Value
Cluster 1 - Average_Transaction_Value
Cluster 2 - Average_Transaction_Value
Cluster 0 - Unique_Products_Purchased
Cluster 1 - Unique_Products_Purchased
Cluster 2 - Unique_Products_Purchased
Cluster 0 - Average_Days_Between_Purchases
Cluster 1 - Average_Days_Between_Purchases
Cluster 2 - Average_Days_Between_Purchases
Cluster 0 - Day_Of_Week
Cluster 1 - Day_Of_Week
Cluster 2 - Day_Of_Week
Cluster 0 - Hour
Cluster 1 - Hour
Cluster 2 - Hour



Cụm 0 - Người mua sắm thông thường cuối tuần:

Khách hàng trong cụm này thường mua sắm ít thường xuyên hơn và chi tiêu ít tiền hơn so với các cụm khác.

Họ thường có số lượng giao dịch ít hơn và mua ít sản phẩm hơn.

Những khách hàng này có sở thích mua sắm vào cuối tuần, có thể mua sắm thông thường hoặc mua sắm qua cửa sổ.

Thói quen chi tiêu của họ khá ổn định theo thời gian, ít biến động trong chi tiêu hàng tháng. Họ hiếm khi hủy giao dịch, điều này cho thấy hành vi mua sắm quyết đoán hơn.

Khi họ mua sắm, mức chi tiêu cho mỗi giao dịch của họ có xu hướng thấp hơn so với các nhóm khác.

Cụm 1 - Những người chi tiêu lớn không thường xuyên:

Khách hàng trong nhóm này không mua sắm thường xuyên nhưng có xu hướng chi tiêu một khoản đáng kể khi mua sắm nhiều loại sản phẩm.

Chi tiêu của họ ngày càng tăng, cho thấy sự quan tâm hoặc đầu tư vào việc mua sắm của họ ngày càng tăng.

Họ thích mua sắm muộn hơn trong ngày, có thể sau giờ làm việc và chủ yếu ở Anh. Họ có xu hướng hủy giao dịch vừa phải, điều này có thể là do họ chi tiêu nhiều hơn:

Có lẽ họ sẽ xem xét lại việc mua hàng của mình thường xuyên hơn.

Việc mua hàng của họ nhìn chung rất lớn, cho thấy họ ưa chuộng các sản phẩm chất lượng hoặc cao cấp.

Cụm 2 - Những người mua sắm bộc trực:

Khách hàng trong nhóm này có đặc điểm là có thói quen chi tiêu cao. Họ có xu hướng mua nhiều loại sản phẩm độc đáo và tham gia vào nhiều giao dịch.

Mặc dù chi tiêu cao nhưng họ có xu hướng hủy một phần đáng kể các giao dịch của mình, điều này có thể cho thấy hành vi mua sắm bốc đồng.

Họ thường mua sắm vào đầu giờ trong ngày, có thể tìm thời gian trước những công việc hàng ngày hoặc tận dụng những ưu đãi sớm.

Mô hình chi tiêu của họ khá thay đổi, với mức chi tiêu hàng tháng biến động cao, cho thấy mô hình mua sắm ít dự đoán hơn.

Điều thú vị là xu hướng chi tiêu của họ đang giảm nhẹ, điều này có thể báo hiệu sự thay đổi trong thói quen mua sắm của họ trong tương lai.

Vậy khách hàng trong cụm 1 được xác định là nhóm khách hàng tiềm năng nhất

Danh sách mã số khách hàng: CustomerID các khách hàng trong cluster1:

Potential - List Customer of Cluter 2

```
[ ] customer_data_cluster1 = customer_data_pca[customer_data_pca['cluster'] == 1]
```

```
[ ] customer_data_cluster1.head(100)
```

CustomerID	PC1	PC2	PC3	PC4	PC5	PC6	cluster
12346.0	-2.186469	-1.705370	-1.576745	1.008187	-0.411803	-1.658012	1
12349.0	1.791116	-2.695652	5.850040	0.853418	0.677111	-1.520098	1
12350.0	-1.997139	-0.542639	0.578781	0.183682	-1.484838	0.062672	1
12352.0	0.428268	-1.482771	-0.758378	-0.593492	-0.376960	0.652029	1
12355.0	-1.341366	-2.608400	1.298483	0.321563	-0.651027	-0.416435	1
...	...	...	...	...	...	...	...
12660.0	-1.455941	-2.562414	0.071896	-0.441032	-0.186869	0.550211	1
12665.0	-2.249715	-0.764065	-1.205439	1.023124	0.435688	-2.085599	1
12666.0	-2.930939	-0.609039	-2.544623	0.567758	-1.374108	-0.572359	1
12667.0	-0.920493	-0.477833	1.195162	-0.539562	-1.125481	1.163320	1
12669.0	1.454352	-2.501351	2.590624	3.350262	0.702247	1.347363	1

100 rows × 7 columns

8 .Đề xuất

CustomerID	Rec1_StockCode	Rec1_Description	Rec2_StockCode	Rec2_Description	Rec3_StockCode	Rec3_Description
15746.0	84879	ASSORTED COLOUR BIRD ORNAMENT	15036	ASSORTED COLOURS SILK FAN	85123A	WHITE HANGING HEART T-LIGHT HOLDER
15728.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	15036	ASSORTED COLOURS SILK FAN
17459.0	18007	ESSENTIAL BALM 3.5G TIN IN ENVELOPE	84879	ASSORTED COLOUR BIRD ORNAMENT	17003	BROCADE RING PURSE
17415.0	18007	ESSENTIAL BALM 3.5G TIN IN ENVELOPE	84879	ASSORTED COLOUR BIRD ORNAMENT	17003	BROCADE RING PURSE
15339.0	18007	ESSENTIAL BALM 3.5G TIN IN ENVELOPE	84879	ASSORTED COLOUR BIRD ORNAMENT	17003	BROCADE RING PURSE
14335.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	15036	ASSORTED COLOURS SILK FAN
15367.0	22616	PACK OF 12 LONDON TISSUES	84879	ASSORTED COLOUR BIRD ORNAMENT	16014	SMALL CHINESE STYLE SCISSOR
17604.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	15036	ASSORTED COLOURS SILK FAN
17828.0	22616	PACK OF 12 LONDON TISSUES	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	85099B	JUMBO BAG RED RETROSPOT
13229.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	84879	ASSORTED COLOUR BIRD ORNAMENT	15036	ASSORTED COLOURS SILK FAN

Nhóm tập trung vào 95% nhóm khách hàng cốt lõi, nhóm phân tích dữ liệu khách hàng đã được làm sạch để xác định các sản phẩm bán chạy nhất trong mỗi cụm. Tận dụng thông tin này, hệ thống sẽ đưa ra các đề xuất được cá nhân hóa, đề xuất ba sản phẩm phổ biến nhất trong cụm của họ mà họ chưa mua. Điều này không chỉ tạo điều kiện thuận lợi cho các chiến lược tiếp thị có mục tiêu mà còn làm phong phú thêm trải nghiệm mua sắm cá nhân, có khả năng thúc đẩy doanh số bán hàng. Đối với nhóm ngoại

lệ, cách tiếp cận cơ bản có thể là đề xuất các sản phẩm ngẫu nhiên làm điểm khởi đầu để thu hút họ.

9. Mở rộng: K Means trong SPMF của Philippe-fournier-Viger

9.1 Input của dữ liệu

K-Means lấy đầu vào là một tập hợp các phiên bản có tên và chứa một hoặc nhiều giá trị kép, tham số K (số nguyên dương >=1) cho biết số lượng cụm được tạo và hàm khoảng cách.

Định dạng tệp đầu vào của K-Means là một tệp văn bản chứa một số phiên bản.

Các dòng đầu tiên (tùy chọn) chỉ định tên của các thuộc tính được sử dụng để mô tả các thể hiện. Trong ví dụ này, hai thuộc tính sẽ được sử dụng, có tên là X và Y. Nhưng lưu ý rằng có thể sử dụng nhiều hơn hai thuộc tính. Mỗi thuộc tính được chỉ định trên một dòng riêng biệt bằng từ khóa "@ATTRIBUTEDEF=", theo sau là tên thuộc tính

Sau đó, mỗi trường hợp được mô tả bằng hai dòng. Dòng đầu tiên (không bắt buộc) chứa chuỗi "@NAME=" theo sau là tên của phiên bản. Sau đó, dòng thứ hai cung cấp danh sách các giá trị kép được phân tách bằng dấu cách đơn.

9.2 Output

Định dạng tệp đầu ra được xác định như sau. Một vài dòng đầu tiên cho biết tên thuộc tính. Mỗi thuộc tính được chỉ định trên một dòng riêng biệt với từ khóa "ATTRIBUTEDEF=" theo sau là tên thuộc tính (một chuỗi). Sau đó, danh sách các cụm được chỉ định. Mỗi cụm được chỉ định trên một dòng riêng biệt, liệt kê các phiên bản có trong cụm. Một phiên bản là một tên theo sau là danh sách các giá trị kép được phân tách bằng " " và giữa các ký tự "[" và "]".

Các cụm được tìm thấy bằng thuật toán có thể được xem trực quan bằng cách sử dụng "Trình xem cụm" được cung cấp trong SPMF. Nếu bạn đang sử dụng giao diện đồ họa của SPMF, hãy nhấp vào hộp kiểm "Trình xem cụm" trước khi nhấn nút "Chạy thuật toán". Kết quả sẽ được hiển thị trong Cluster Viewer.

C. Kết luận

1. Tóm tắt kết quả

Tóm lại, kỹ thuật khai thác dữ liệu thường được sử dụng trong kinh doanh để phân loại khách hàng thuộc các phân khúc nhất định.

Các doanh nghiệp nên sử dụng sức mạnh của máy học biết được đặc điểm, dự báo tiêu dùng cho từng nhóm nhằm duy trì tính cạnh tranh trên thị trường. Sau đó, họ có thể tạo ra các chiến lược tiếp thị thành công dựa trên kết quả dự đoán của mình.

Có thể tìm thấy các phân khúc khách hàng có tỷ lệ rời bỏ cao nhất bằng cách áp dụng phân cụm K-Means để phân chia nhóm khách hàng. Điều này cho phép các doanh nghiệp tập trung nỗ lực vào các lĩnh vực thị trường cụ thể này. Ví dụ, đối với nhóm khách hàng thường xuyên mua vào cuối tuần, có thể thực hiện các đợt giảm giá vào thời gian này để thu hút họ.

2. Hạn chế

Tuy nhiên, nghiên cứu này có những hạn chế nhất định. Chúng bao gồm kích thước nhỏ của tập dữ liệu và thực tế là mỗi hàng đại diện cho một giao dịch riêng biệt mà không có thông tin cấp độ chi tiết về vị trí cửa hàng. Kết quả là, nó ngăn cản việc điều tra kỹ lưỡng hành vi của khách hàng theo khu vực. Ngoài ra, K-Means là kỹ thuật phân cụm duy nhất được sử dụng.

3. Hướng phát triển

Để phát triển hệ thống đề xuất tập trung vào việc giữ chân những khách hàng quan trọng, điều cần thiết là phải thử nghiệm các bộ dữ liệu lớn hơn bao gồm các khía cạnh liên quan đến sản phẩm và dịch vụ. Hơn nữa, nên nghiên cứu các kỹ thuật học máy bổ sung như SVM, CatBoost và các phương pháp phân cụm thay thế. Làm như vậy sẽ đưa ra đánh giá đầy đủ về hiệu suất của mô hình và nâng cao hiểu biết của chúng ta về khả năng dự đoán.

Tài liệu tham khảo

Fournier-Viger, P. (2022) *SPMF: A java open-source data mining library*. Available at: https://www.philippe-fournier-viger.com/spmf/ (Accessed: 22 October 2023).

Aliaymansoliman (2023a) *Online_Retail-Customer Segmentation, Kaggle*. Available at: https://www.kaggle.com/code/aliaymansoliman/online-retail-customer-segmentation (Accessed: 22 October 2023).

(No date) *Pearson*. Available at: https://www.pearsoned.ca/highered/showcase/kotler/pdf/9780132473958_ch01.pdf (Accessed: 22 October 2023).

Fontanella, C. (2021) *What is a high-value customer? [+5 ways to identify them], HubSpot Blog*. Available at: https://blog.hubspot.com/service/high-value-customer (Accessed: 22 October 2023).