BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN: ĐỒ ÁN 2

ÁP DỤNG KỸ THUẬT PHÂN CỤM ĐỂ PHÂN LOẠI KHÁCH HÀNG DỰA TRÊN HÀNH VI MUA SẮM TRỰC TUYẾN

Sinh viên thực hiện:

NGUYỄN ĐÌNH MẠNH	DHKL16A2HN	22174600037
PHẠM THỊ HÀ NAM	DHKL16A2HN	22174600009
NGUYỄN THỊ NHƯ	DHKL16A2HN	22174600047
TRẦN THỊ THU TRANG	DHKL16A2HN	22174600028
NGUYỄN MẠNH TIẾN	DHKL16A2HN	22174600066

Giáo viên giảng dạy: TH.S LÊ HẰNG ANH

Hà Nội, 04/2025

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN: ĐỒ ÁN 2

ÁP DỤNG KỸ THUẬT PHÂN CỤM ĐỂ PHÂN LOẠI KHÁCH HÀNG DỰA TRÊN HÀNH VI MUA SẮM TRỰC TUYẾN

Sinh viên thực hiện:

NGUYỄN ĐÌNH MẠNH	DHKL16A2HN	22174600037
PHẠM THỊ HÀ NAM	DHKL16A2HN	22174600009
NGUYỄN THỊ NHƯ	DHKL16A2HN	22174600047
TRẦN THỊ THU TRANG	DHKL16A2HN	22174600028
NGUYỄN MẠNH TIẾN	DHKL16A2HN	22174600066

Giáo viên giảng dạy: TH.S LÊ HẰNG ANH

Hà Nội, 04/2025

PHIẾU ĐĂNG KÝ ĐỀ TÀI

1. Tên đề tài: Áp dụng kỹ thuật phân cụm để phân loại khách hàng dựa trên hành vi mua sắm trực tuyến.

2. Thông tin nhóm sinh viên:

Sinh viên 1 (Nhóm trưởng):

- Họ và tên:Nguyễn Đình Mạnh
- Mã sinh viên: 22174600037
- Điện thoại: 0339691267
- Email: ndmanh.dhkl16a2hn@sv.uneti.edu.vn

Sinh viên 2:

- Họ và tên: Phạm Thị Hà Nam
- Mã sinh viên: 22174600009
- Điện thoại: 0988427151
- Email: pthnam.dhkl16a2hn@sv.uneit.edu.vn

Sinh viên 3:

- Họ và tên: Nguyễn Thị Như
- Mã sinh viên: 22174600047
- Điện thoại: 0869081635
- Email: ntnhu.dhkl16a2hn@sv.uneti.edu.vn

Sinh viên 4:

- Họ và tên: Trần Thị Thu Trang
- Mã sinh viên: 22174600028
- Điện thoại: 0355022517
- Email: ttttrang.dhkl16a2hn@sv.uneit.edu.vn

Sinh viên 5:

- Họ và tên: Nguyễn Mạnh Tiến
- Mã sinh viên: 22174600066
- Điện thoại: 0364181355
- Email: nmtien.dhkl16a2hn@sv.uneti.edu.vn

3. Tóm tắt nội dung đề tài:

Trong thời đại số hóa, việc thấu hiểu hành vi của khách hàng trên các nền tảng mua sắm trực tuyến là yếu tố quan trọng giúp doanh nghiệp cá nhân hóa trải nghiệm người dùng và nâng cao hiệu quả tiếp thị. Đề tài này hướng đến mục tiêu phân loại khách hàng dựa trên hành vi mua sắm, thông qua việc phân tích và khai thác dữ liệu thực tế từ các giao dịch thương mại điện tử.

Bộ dữ liệu được sử dụng là *"Online Retail II (phiên bản UCI)"*, bao gồm thông tin chi tiết về các đơn hàng được thực hiện trong năm 2011, với các trường dữ liệu như mã khách hàng, thời gian mua hàng, số lượng sản phẩm, đơn giá và quốc gia. Sau quá trình làm sạch và xử lý dữ liệu – loại bỏ giao dịch bị hủy, dữ liệu thiếu mã khách hàng,

các chỉ số hành vi như tần suất mua hàng, số tiền chi tiêu, thời gian mua gần nhất sẽ được trích xuất để mô tả rõ hơn về từng khách hàng.

Dựa trên những đặc trưng hành vi này, nhóm tiến hành phân chia tập khách hàng thành nhiều nhóm có hành vi tương đồng, giúp làm rõ sự khác biệt giữa các phân khúc như: nhóm chi tiêu cao, nhóm thường xuyên mua hàng, nhóm ít tương tác, v.v. Quá trình phân nhóm được đánh giá bằng các tiêu chí đo lường phù hợp, đảm bảo rằng mỗi nhóm được hình thành có tính đại diện cao và hỗ trợ hiệu quả cho việc ra quyết định trong kinh doanh.

Kết quả đề tài sẽ cung cấp một cái nhìn trực quan và hệ thống về đặc điểm khách hàng, từ đó đưa ra những đề xuất ứng dụng thực tiễn như cá nhân hóa chương trình khuyến mãi, chăm sóc khách hàng trọng điểm, hoặc thiết kế chiến dịch tiếp thị riêng cho từng nhóm cụ thể.

Ngày 11 tháng 4 năm 2025

Nhóm trưởng

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

1. Tên đề tài: Áp dụng kỹ thuật phân cụm để phân loại khách hàng dựa trên hành vi mua sắm trực tuyến.

2. Mục tiêu đề tài:

- Trích xuất các chỉ số hành vi tiêu dùng cốt lõi từ dữ liệu giao dịch, bao gồm:
  - Recency: Khoảng thời gian kể từ lần mua hàng gần nhất.
  - Frequency: Tần suất mua hàng.
  - Monetary: Tổng giá trị mua hàng.
- Xây dựng mô hình phân cụm khách hàng sử dụng các thuật toán như K-Means, Hierarchical Clustering, hoặc DBSCAN, nhằm chia khách hàng thành các nhóm có hành vi tương đồng.
- Phân tích đặc điểm từng nhóm khách hàng sau phân cụm để hiểu rõ chân dung tiêu dùng của từng phân khúc và đánh giá giá trị kinh tế tương ứng.
- Đề xuất các định hướng chiến lược tiếp thị và chăm sóc khách hàng phù hợp với từng nhóm như: duy trì nhóm khách hàng trung thành, tái kích hoạt nhóm khách hàng không còn mua sắm, và tối ưu chuyển đổi ở nhóm tiềm năng.

3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài:

Trong giai đoạn đầu, các doanh nghiệp chủ yếu phân khúc khách hàng dựa trên các yếu tố nhân khẩu học cơ bản như tuổi, giới tính và khu vực địa lý. Cách tiếp cận này tuy đơn giản nhưng tồn tại nhiều hạn chế do không phản ánh chính xác hành vi mua hàng thực tế. Theo nghiên cứu của Nielsen, phương pháp truyền thống thường gây lãng phí 30-40% ngân sách marketing do không tiếp cận đúng đối tượng mục tiêu.

Sự ra đời của mô hình RFM đã tạo nên bước đột phá trong phân tích khách hàng. Mô hình này đánh giá khách hàng dựa trên ba yếu tố then chốt: thời gian mua hàng gần nhất (Recency), tần suất mua hàng (Frequency) và giá trị chi tiêu (Monetary). Mỗi yếu tố được xếp hạng theo thang điểm từ 1-5, giúp xác định chính xác mức độ quan trọng của từng khách hàng.

Khi kết hợp với các thuật toán phân cụm hiện đại như K-Means, DBSCAN hay phân cấp, mô hình RFM càng phát huy hiệu quả. Các thuật toán này cho phép tự động phân nhóm khách hàng dựa trên đặc điểm hành vi, giúp doanh nghiệp xác định các phân khúc tiềm năng và có nguy cơ rời bỏ. Đặc biệt, K-Means được ưa chuộng nhờ khả năng xử lý nhanh với dữ liệu lớn.

Trên thực tế, nhiều tập đoàn lớn đã ứng dụng thành công phương pháp này. Amazon ghi nhận mức tăng 35% doanh thu nhờ cá nhân hóa khuyến mãi dựa trên RFM. Starbucks cải thiện 27% tỷ lệ giữ chân khách hàng sau khi triển khai hệ thống phân tích này. Nghiên cứu của MIT (2023) cũng chỉ ra doanh nghiệp áp dụng RFM kết hợp machine learning có tỷ lệ chuyển đổi cao hơn 40-60% so với phương pháp truyền thống.

Xu hướng hiện nay đang phát triển theo hướng tích hợp thêm các yếu tố hành vi và dữ liệu đa kênh. Nhiều doanh nghiệp đã bắt đầu áp dụng mô hình RFM-S tích hợp

dữ liệu mạng xã hội, hay hệ thống phân tích thời gian thực để nâng cao hiệu quả. Các giải pháp AI cũng được triển khai để dự đoán xu hướng và hành vi khách hàng chính xác hơn.

4. Nội dung đề tài:

Đề tài sử dụng bộ dữ liệu "Online Retail II (UCI)" gồm các giao dịch mua sắm trực tuyến của một công ty bán lẻ trong giai đoạn trong năm 2011, với mục tiêu phân nhóm khách hàng dựa trên hành vi mua hàng. Các bước thực hiện gồm:

Bước 1: Tiền xử lý và khám phá dữ liệu

- Kiểm tra và loại bỏ các dòng dữ liệu bị thiếu thông tin quan trọng.
- Xác định xem có tồn tại các giao dịch bị hủy hay không. Nếu có, cần xử lý phù hợp theo mục tiêu phân tích.
- Tính lại tổng giá trị đơn hàng cho mỗi dòng dữ liệu.
- Phân tích sơ bộ: Tổng số khách hàng, số lần mua hàng, sản phẩm bán chạy, quốc gia có nhiều giao dịch.
- Trực quan hóa dữ liệu để hiểu rõ cấu trúc và hành vi người dùng.

Bước 2: Tạo đặc trưng hành vi của khách hàng

- Tính các chỉ số RFM:
  - Recency: số ngày kể từ lần mua gần nhất
  - Frequency: số lần mua hàng
  - Monetary: tổng số tiền đã chi tiêu
- Bổ sung thêm một số chỉ số nâng cao:
  - Giá trị đơn hàng trung bình
  - Tỷ lệ đơn hàng bị hủy
- Chuẩn hóa dữ liệu để dễ áp dụng thuật toán phân cụm.

Bước 3: Phân nhóm khách hàng (Phân cụm)

- Sử dụng thuật toán K-Means để chia khách hàng thành các nhóm dựa trên hành vi.
- Xác định số cụm phù hợp bằng phương pháp Elbow hoặc Silhouette Score.
- Thử nghiệm thêm một vài thuật toán khác như Hierarchical Clustering hoặc DBSCAN để so sánh.
- Giảm chiều dữ liệu bằng PCA để dễ vẽ biểu đồ trực quan các nhóm khách hàng.

Bước 4: Phân tích kết quả

- Mô tả từng nhóm khách hàng sau khi phân cụm:
  - Nhóm chi tiêu cao
  - Nhóm mua hàng thường xuyên
  - Nhóm ít mua, có thể đã ngừng hoạt động
- Đánh giá ý nghĩa kinh doanh và cách khai thác từng nhóm.

Bước 5: Ứng dụng kết quả

- Đề xuất cách sử dụng phân cụm trong marketing:
  - Gửi ưu đãi cho nhóm mua nhiều

    - Khuyến khích nhóm mua ít quay lại
- Gợi ý ứng dụng vào hệ thống chăm sóc khách hàng hoặc quảng cáo.

5. Phương pháp thực hiện:

Đề tài sử dụng phương pháp khai phá dữ liệu và học máy không giám sát (unsupervised learning) để phân loại khách hàng dựa trên hành vi mua sắm. Dữ liệu đầu vào là bộ *"Online Retail II (UCI)",* trong đó nhóm thực hiện lựa chọn và xử lý các giao dịch trong năm 2011 để đảm bảo tính tập trung và đồng nhất theo thời gian.

Quy trình thực hiện cụ thể gồm các bước sau:

Bước 1: Tiền xử lý dữ liệu

- Lọc dữ liệu: chỉ giữ lại các bản ghi có thời gian thuộc năm 2011.
- Kiểm tra và xem xét các giao dịch bất thường (Các giao dịch bị hủy,….),các dòng có giá trị âm hoặc thiếu CustomerID.
- Tính toán các trường mới như: tổng số tiền mỗi đơn hàng (Quantity × UnitPrice).
- Mã hóa và chuẩn hóa dữ liệu để phù hợp với các thuật toán phân tích sau này.

Bước 2: Tạo tập đặc trưng hành vi

- Áp dụng mô hình RFM để xây dựng đặc trưng cho từng khách hàng:
    - Recency: Số ngày kể từ lần mua gần nhất đến ngày cuối cùng trong tập dữ liệu.
    - Frequency: Số lượng giao dịch đã thực hiện trong năm 2011.
    - Monetary: Tổng số tiền đã chi tiêu.
- Ngoài ra, có thể xây dựng thêm một số đặc trưng bổ sung như: giá trị đơn hàng trung bình, tỷ lệ đơn hàng bị hủy.

Bước 3: Phân cụm khách hàng

- Sử dụng thuật toán K-Means để thực hiện phân cụm khách hàng dựa trên tập đặc trưng RFM (sau chuẩn hóa).
- Dùng phương pháp Elbow và Silhouette Score để xác định số cụm phù hợp.
- Thử nghiệm thêm một số mô hình khác như Hierarchical Clustering hoặc DBSCAN để đối chiếu kết quả.

Bước 4: Đánh giá và phân tích

- Gắn nhãn cho từng nhóm khách hàng sau khi phân cụm (ví dụ: nhóm trung thành, nhóm tiềm năng, nhóm có rủi ro rời bỏ).
- Trực quan hóa kết quả bằng biểu đồ 2D sau khi giảm chiều dữ liệu bằng PCA.
- Phân tích ý nghĩa hành vi của từng nhóm và đề xuất hướng khai thác phù hợp.

Bước 5: Tổng hợp và đề xuất

- Tổng hợp kết quả:
    - Tổng hợp toàn bộ quy trình phân tích từ tiền xử lý, xây dựng đặc trưng RFM đến phân cụm khách hàng.

- Đánh giá hiệu quả mô hình phân cụm dựa trên các chỉ số như Elbow, Silhouette Score và kết quả trực quan PCA.
- Xác định rõ các nhóm khách hàng đặc trưng với hành vi tiêu dùng khác nhau.

- Đề xuất ứng dụng:
  - Ứng dụng kết quả phân cụm vào marketing cá nhân hóa, tối ưu chiến dịch quảng bá.
  - Ưu tiên chăm sóc khách hàng trung thành, giữ chân nhóm có nguy cơ rời bỏ.
  - Hỗ trợ ra quyết định trong quản lý quan hệ khách hàng và phân bổ nguồn lực.
  - Làm cơ sở cho việc phát triển hệ thống CRM thông minh hơn trong tương lai.

6. Phân công công việc (dự kiến)

STT	Họ và tên	Mã sinh viên	Nội dung công việc được phân công
1	Phạm Thị Hà Nam, Nguyễn Thị Như	22174600009, 22174600047	- Lọc dữ liệu năm 2011 - Loại bỏ dữ liệu thiếu/mã hóa đơn hủy - Tính tổng chi tiêu mỗi đơn - Trực quan dữ liệu sơ bộ (sản phẩm, quốc gia, tần suất mua)
2	Nguyễn Thị Như,Trần Thị Thu Trang	22174600047, 22174600028	- Tính RFM (Recency, Frequency, Monetary) - Tạo các chỉ số bổ sung nếu cần (giá trị đơn hàng TB, tỷ lệ hủy) - Chuẩn hóa dữ liệu để đưa vào mô hình
3	Nguyễn Đình Mạnh,Nguyễn Mạnh Tiến	22174600037, 22174600066	- Áp dụng K-Means, xác định số cụm tối ưu - Thử nghiệm các thuật toán bổ sung nếu cần (Hierarchical, DBSCAN) - Giảm chiều dữ liệu (PCA/t-SNE) - Vẽ biểu đồ cụm khách hàng
4	Nguyễn Đình Mạnh,Phạm Thị Hà Nam	22174600037, 22174600009	- Mô tả hành vi từng nhóm sau phân cụm - Đặt tên nhóm (khách trung thành, mới, rủi ro, v.v.) - Đề xuất cách khai thác mỗi nhóm trong thực tế

5	Nguyễn Mạnh Tiến,Trần Thị Thu Trang	22174600066, 22174600028	- Viết báo cáo dựa trên nội dung các phần - Thiết kế slide trình bày - Rà soát toàn bộ logic đề tài trước khi nộp/bảo vệ

7. Dự kiến kết quả đạt được:

Sau khi hoàn thành đề tài, nhóm dự kiến sẽ đạt được các kết quả chính sau:

- Bộ dữ liệu khách hàng đã được làm sạch và chuẩn hóa
  - Tập dữ liệu 2011 được xử lý loại bỏ lỗi, thiếu mã khách hàng, giao dịch hủy.
  - Các chỉ số như tổng chi tiêu, RFM và đặc trưng hành vi khách hàng được xây dựng rõ ràng, có thể sử dụng trong các mô hình phân tích khác.
- Mô hình phân cụm khách hàng hoạt động hiệu quả
  - Áp dụng thành công thuật toán K-Means để chia khách hàng thành các nhóm hành vi rõ ràng.
  - Số lượng cụm được xác định hợp lý thông qua các tiêu chí đánh giá (Elbow, Silhouette Score).
  - Kết quả phân cụm được minh họa trực quan bằng biểu đồ 2D sau khi giảm chiều dữ liệu.
- Phân tích và hiểu rõ đặc điểm từng nhóm khách hàng
  - Mỗi cụm khách hàng được phân tích dựa trên tần suất mua, mức chi tiêu, thời gian quay lại,…
  - Các nhóm như "khách hàng trung thành", "khách tiềm năng", "khách ít hoạt động" được xác định và gán nhãn cụ thể.
- Đề xuất ứng dụng phân cụm vào thực tế kinh doanh
  - Gợi ý các chiến lược tiếp thị riêng biệt cho từng nhóm khách hàng.
  - Khả năng tích hợp vào hệ thống CRM hoặc hỗ trợ quyết định trong các chiến dịch quảng bá – chăm sóc khách hàng.
- Báo cáo hoàn chỉnh và sản phẩm dữ liệu
  - Báo cáo học thuật trình bày toàn bộ quá trình thực hiện đề tài.
  - Tệp kết quả (dữ liệu phân cụm, mã nguồn Python, biểu đồ, phân tích nhóm) có thể dùng làm nền tảng cho các nghiên cứu hoặc ứng dụng tiếp theo

Ngày 11 tháng 4 năm 2025

Nhóm trưởng

MỞ ĐẦU

Trong thời đại số hóa hiện nay, thương mại điện tử đã trở thành một phần không thể thiếu trong cuộc sống hàng ngày của người tiêu dùng. Sự phát triển nhanh chóng của các nền tảng mua sắm trực tuyến đã tạo ra một môi trường cạnh tranh khốc liệt, nơi mà việc hiểu rõ hành vi và nhu cầu của khách hàng trở thành yếu tố quyết định cho sự thành công của doanh nghiệp. Hành vi mua sắm trực tuyến không chỉ đơn thuần là việc lựa chọn sản phẩm và thực hiện giao dịch; nó còn phản ánh những thói quen, sở thích và xu hướng tiêu dùng của từng cá nhân. Do đó, việc phân tích và phân loại khách hàng dựa trên hành vi mua sắm trực tuyến là một nhiệm vụ quan trọng và cần thiết.

Vậy làm sao để các doanh nghiệp có thể phân loại và hiểu rõ hơn về khách hàng của mình, từ đó đưa ra những chiến lược phù hợp nhằm nâng cao hiệu quả kinh doanh? Để giải quyết vấn đề này, việc áp dụng kỹ thuật phân cụm để phân loại khách hàng dựa trên hành vi mua sắm trực tuyến là một giải pháp hợp lý. Kỹ thuật này không chỉ giúp doanh nghiệp nhận diện các nhóm khách hàng khác nhau mà còn tối ưu hóa các chiến dịch quảng cáo và phát triển sản phẩm phù hợp với nhu cầu của từng phân khúc.

Vì vậy, nhóm đã thực hiện đề tài "ÁP DỤNG KỸ THUẬT PHÂN CỤM ĐỂ PHÂN LOẠI KHÁCH HÀNG DỰA TRÊN HÀNH VI MUA SẮM TRỰC TUYẾN". Nhóm tin rằng đây sẽ là một đề tài hữu ích, không chỉ cho các doanh nghiệp trong lĩnh vực thương mại điện tử mà còn cho những ai quan tâm đến việc phân tích dữ liệu và hiểu biết về hành vi tiêu dùng.

Trong quá trình thực hiện đề tài, nhóm đã nhận được nhiều sự chỉ bảo, giúp đỡ và những góp ý chân thành từ Cô Lê Hằng Anh. Nhóm em xin chân thành cảm ơn cô đã truyền đạt kiến thức và kinh nghiệm một cách tận tình và sâu sắc. Tuy nhiên, do hạn chế về mặt kiến thức và kinh nghiệm cũng như kỹ năng chưa cao, nên bài làm của nhóm chắc chắn còn nhiều thiếu sót. Nhóm rất mong nhận được sự góp ý chân thành của cô để nhóm có thể hoàn thiện tốt hơn.

Đồ án bao gồm các phần được phân chương như sau:

- Chương 1: Đặt vấn đề (giới thiệu)
- Chương 2: Cơ sở lý thuyết
- Chương 3: Thực nghiệm
- Chương 4: Kết quả đạt được
- Chương 5: Kết luận, ưu điểm, nhược điểm, hướng phát triển

MỤC LỤC

MỤC LỤC HÌNH VẼ

MỤC LỤC BẢNG

CHƯƠNG 1: ĐẶT VẤN ĐỀ

Thương mại điện tử đã trở thành một trong những động lực chính thúc đẩy tăng trưởng kinh tế toàn cầu trong thời đại công nghệ số, đánh dấu sự chuyển mình mạnh mẽ trong thói quen mua sắm và cách thức kinh doanh của các doanh nghiệp. Sự bùng nổ của các nền tảng mua sắm trực tuyến đã dẫn đến việc tích lũy một lượng dữ liệu giao dịch khổng lồ, chứa đựng thông tin chi tiết về hành vi, sở thích và nhu cầu của khách hàng — nguồn tài nguyên quý giá để tối ưu hóa chiến lược kinh doanh. Tuy nhiên, khai thác hiệu quả khối dữ liệu này nhằm hiểu rõ khách hàng và đưa ra các quyết định chiến lược vẫn là thách thức lớn, nhất là khi ngành bán lẻ trực tuyến đối mặt với cạnh tranh khốc liệt và yêu cầu ngày càng cao về cá nhân hóa dịch vụ. Vấn đề đặt ra là làm sao doanh nghiệp có thể phân tích và phân loại khách hàng một cách khoa học, từ đó xây dựng chiến lược marketing hiệu quả, tối ưu hóa nguồn lực và nâng cao trải nghiệm khách hàng trong môi trường thương mại điện tử đầy biến động.

Đề tài "Ứng dụng kỹ thuật phân cụm để phân loại khách hàng dựa trên hành vi mua sắm trực tuyến" được thực hiện nhằm giải quyết những vấn đề trên, với sự hỗ trợ của tập dữ liệu "Online Retail II" (phiên bản UCI). Tập dữ liệu này ghi nhận các giao dịch trong năm 2011 của một công ty bán lẻ trực tuyến tại Vương quốc Anh, cung cấp thông tin chi tiết về hóa đơn, mã sản phẩm, số lượng, giá cả, mã khách hàng và quốc gia, tạo nền tảng để phân tích sâu về hành vi tiêu dùng. Các doanh nghiệp thường gặp khó khăn trong việc nhận diện nhóm khách hàng tiềm năng, hiểu rõ thói quen mua sắm như tần suất giao dịch, giá trị chi tiêu hay thời gian mua sắm gần nhất, dẫn đến hạn chế trong việc triển khai các chiến lược tiếp thị phù hợp.

Kỹ thuật phân cụm — một phương pháp học máy không giám sát — được xem là giải pháp phù hợp để nhóm các khách hàng có đặc điểm tương đồng, từ đó hỗ trợ doanh nghiệp đưa ra quyết định kinh doanh chính xác. Mục tiêu của đề tài là phát triển một phương pháp phân cụm hiệu quả, giúp doanh nghiệp nhận diện các phân khúc khách hàng khác nhau, từ đó xây dựng chiến lược marketing cá nhân hóa, cải thiện dịch vụ và tăng cường sự hài lòng của khách hàng. Nghiên cứu không chỉ dừng lại ở việc phân loại khách hàng mà còn cung cấp những hiểu biết sâu sắc về xu hướng tiêu dùng, góp phần tối ưu hóa hoạt động kinh doanh và nâng cao hiệu quả cạnh tranh trên thị trường quốc tế.

Giải quyết vấn đề này có ý nghĩa quan trọng trong việc thúc đẩy sự phát triển bền vững của thương mại điện tử, đặc biệt khi doanh nghiệp cần thích nghi nhanh với sự thay đổi không ngừng của thị trường và nhu cầu đa dạng của người tiêu dùng. Trong bối cảnh dữ liệu ngày càng trở thành "vàng mỏ", ứng dụng kỹ thuật phân tích tiên tiến như phân cụm không chỉ giúp cải thiện hiệu suất kinh doanh mà còn tạo lợi thế cạnh tranh dài hạn, hướng tới mô hình kinh doanh thông minh và bền vững trong tương lai.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Tổng quan về học máy

2.1.1. Định nghĩa về học máy

Học máy (machine learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Các thuật toán học máy xây dựng một mô hình dựa trên dữ liệu mẫu, được gọi là dữ liệu huấn luyện, để đưa ra dự đoán hoặc quyết định mà không cần được lập trình chi tiết về việc đưa ra dự đoán hoặc quyết định này.

Tom Mitchell, giáo sư nổi tiếng của Đại học Carnegie Mellon University – CMU định nghĩa cụ thể và chuẩn mực hơn như sau: "Một chương trình máy tính CT được xem là học cách thực thi một lớp nhiệm vụ NV thông qua trải nghiệm KN, đối với thang đo năng lực NL nếu như dùng NL ta đo thấy năng lực thực thi của chương trình có tiến bộ sau khi trải qua KN" (máy đã học).[1]

2.1.2. Phân loại học máy

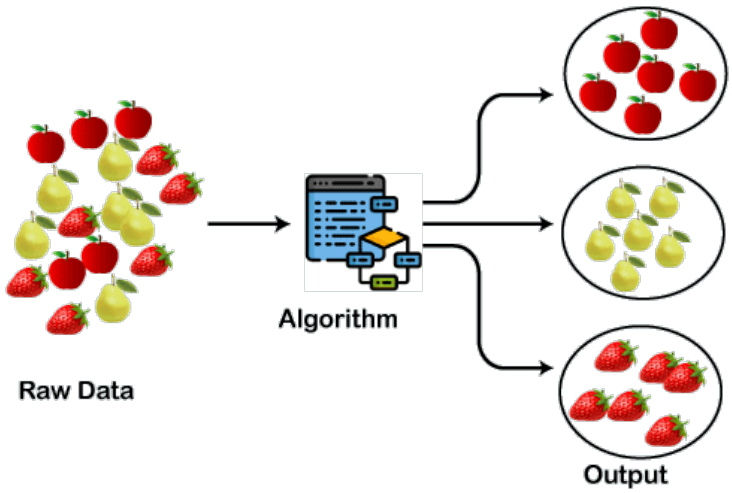
Machine Learning (học máy) là lĩnh vực rộng lớn với nhiều phương pháp và kỹ thuật khác nhau. Có nhiều cách để phân loại các thuật toán học máy, tuy nhiên, cách phân loại phổ biến và cơ bản nhất là dựa trên sự có mặt của nhãn dữ liệu trong quá trình học. Theo đó, machine learning thường được chia thành các loại:

- Học có giám sát (Supervised Learning): là phương pháp học mà trong đó dữ liệu đầu vào được gán nhãn đầy đủ. Mục tiêu là xây dựng một mô hình có khả năng dự đoán hoặc phân loại chính xác các dữ liệu mới dựa trên các mẫu đã học. Ví dụ: bài toán phân loại email thành thư rác hoặc không thư rác, dự đoán giá nhà dựa trên các đặc trưng như diện tích, vị trí,...
- Học không giám sát (Unsupervised Learning): Trong học không giám sát, dữ liệu đầu vào không có nhãn. Thuật toán sẽ tự động tìm kiếm các cấu trúc, mẫu hoặc nhóm ẩn trong dữ liệu. Phân cụm (clustering) là một ví dụ phổ biến của học không giám sát, giúp nhóm các đối tượng tương tự nhau lại với nhau.
- Học bán giám sát (Semi-supervised Learning): Kết hợp giữa dữ liệu có nhãn và không có nhãn để tận dụng tối đa thông tin, giúp cải thiện hiệu quả học khi dữ liệu có nhãn hạn chế.
- Học sâu (Deep Learning): Một nhánh của học có giám sát hoặc không giám sát, sử dụng các mạng nơ-ron nhiều lớp (deep neural networks) để học các đặc trưng phức tạp từ dữ liệu lớn.
- Học củng cố (Reinforcement Learning): Phương pháp học qua tương tác với môi trường, trong đó chương trình học cách đưa ra các hành động tối ưu dựa trên phản hồi (phần thưởng hoặc phạt).[2]

2.2. Phân cụm

2.2.1. Định nghĩa và mục tiêu

Trong học máy , các tác vụ được chia thành hai loại chính: học có giám sát , trong đó dữ liệu đi kèm với các nhãn rõ ràng và học không giám sát , trong đó dữ liệu không có các nhãn này. Phân cụm dữ liệu là một kỹ thuật phân tích các vấn đề học máy không giám sát để tìm ra các mẫu và đặc điểm ẩn trong dữ liệu. Đây là một phương pháp mạnh mẽ để nhận dạng mẫu, cung cấp những hiểu biết hữu ích về dữ liệu mà có thể không thấy rõ khi kiểm tra dữ liệu thô. Vào cuối quá trình phân cụm, tập dữ liệu được phân đoạn thành các cụm khác nhau. Mỗi nhóm chứa các điểm dữ liệu có đặc điểm tương tự, đảm bảo các cụm chứa các điểm dữ liệu khác biệt rõ rệt.[3]



Hình 2-1: Hình ảnh về thuật toán phân cụm

Phân cụm có thể được xem là bài toán tối ưu hóa, trong đó mục tiêu là tìm cách phân chia dữ liệu thành các cụm sao cho hàm mục tiêu được tối ưu. Hàm mục tiêu thường được thiết kế để tối thiểu hóa độ phân tán trong cụm, tức là giảm thiểu tổng khoảng cách hoặc sai số giữa các điểm dữ liệu với tâm cụm của chúng.

Về mặt thống kê, phân cụm giúp khám phá cấu trúc ẩn trong dữ liệu không nhãn thông qua việc phân tích các mẫu ẩn (latent patterns). Khi dữ liệu không có nhãn, phân cụm giúp phát hiện các nhóm tự nhiên, từ đó hiểu được cách dữ liệu được tổ chức, các mối quan hệ tiềm ẩn giữa các điểm dữ liệu.

Trong bối cảnh thương mại điện tử, phân cụm giúp phân loại khách hàng dựa trên các đặc trưng như tần suất mua hàng, giá trị đơn hàng, loại sản phẩm ưa thích, thời gian truy cập,... Qua đó, doanh nghiệp có thể xác định nhóm khách hàng tiềm năng, nhóm khách hàng trung thành hoặc nhóm khách hàng có xu hướng rời bỏ.[4]

2.2.2. Đặc điểm của phân cụm

Phân cụm dữ liệu (Clustering) là một kỹ thuật quan trọng trong học máy không giám sát, với mục tiêu nhóm các điểm dữ liệu tương tự nhau vào cùng một cụm. Các điểm trong cùng cụm có xu hướng giống nhau hơn so với các điểm thuộc cụm khác.

Phương pháp phân cụm được thực hiện khi dữ liệu chưa có cấu trúc định dạng rõ ràng về bảng dữ liệu

Phân cụm thuộc nhóm phương pháp học không giám sát (unsupervised learning) do số cụm dữ liệu không được biết trước. Nhiệm vụ chính là tìm ra và đo đạc sự khác biệt giữa các đối tượng dữ liệu.

Một phương pháp phân cụm tốt là phương pháp tạo ra các cụm có chất lượng cao:

- Độ tương đồng trong mỗi cụm cao, tức là giữa các điểm dữ liệu trong cụm có sự giống nhau.
- Độ tương tự giữa các cụm thấp, tức là giữa các cụm có sự khác biệt lớn.
- Số cụm dữ liệu không được biết trước, phụ thuộc vào dữ liệu cụ thể và mục tiêu phân tích.
- Có nhiều cách tiếp cận và kỹ thuật khác nhau, mỗi cách tiếp cận có thể phù hợp với từng loại dữ liệu và mục tiêu phân tích khác nhau.
- Các kỹ thuật phân cụm khác nhau thường mang lại kết quả khác nhau. Do đó, việc lựa chọn phương pháp phù hợp là một việc quan trọng trong quá trình phân cụm.

2.2.3. Nguyên tắc hoạt động

Quá trình phân cụm thường diễn ra theo một chu trình lặp, trong đó các điểm dữ liệu được liên tục gán vào các cụm phù hợp và các đặc trưng đại diện của cụm được cập nhật, nhằm dần dần hoàn thiện cấu trúc phân nhóm. Bên cạnh đó, việc lựa chọn các tham số phù hợp và đánh giá chất lượng cụm cũng đóng vai trò quan trọng để đảm bảo kết quả phân cụm phản ánh đúng bản chất của dữ liệu.

Các nguyên tắc hoạt động chính của phân cụm bao gồm các phương pháp đo khoảng cách, hàm mục tiêu, cách đánh giá chất lượng cụm, quy trình tính toán lặp và các tham số điều chỉnh phổ biến.

2.2.3.1. Độ đo khoảng cách

Để đánh giá mức độ tương đồng hay khác biệt giữa các điểm dữ liệu, phân cụm sử dụng các độ đo khoảng cách phổ biến như:

- Euclidean distance (Khoảng cách Euclid): Khoảng cách "thẳng" giữa hai điểm trong không gian đa chiều, thường dùng cho dữ liệu liên tục.

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - c_{jk})^2}$$

- Manhattan distance (Khoảng cách Manhattan): Tổng khoảng cách tuyệt đối theo từng chiều, phù hợp với dữ liệu có cấu trúc lưới hoặc khi muốn giảm ảnh hưởng của ngoại lệ.

$$d(x_i, c_j) = \sum_{k=1}^{n} |(x_{ik} - c_{jk}|$$

- Cosine similarity (Độ tương đồng cosine): Đo lường góc giữa hai vectơ, thường dùng cho dữ liệu văn bản hoặc dữ liệu có hướng.

$$cosine\ similarity(x_i, c_j) = \frac{x_i \cdot c_j}{\|x_i\|\|c_j\|}$$

2.2.3.2. Hàm mục tiêu

Hàm mục tiêu trong phân cụm thường liên quan đến việc tối ưu hóa tổng khoảng cách hoặc mật độ giữa các điểm trong cùng một cụm. Mục tiêu là giảm thiểu khoảng cách giữa các điểm trong cụm và tối đa hóa khoảng cách giữa các cụm khác nhau.

- Tối ưu hóa tổng khoảng cách: Ví dụ như thuật toán K-means cố gắng giảm thiểu tổng bình phương khoảng cách giữa các điểm và tâm cụm.
- Tối ưu hóa mật độ: Thuật toán như DBSCAN tập trung vào việc tìm các vùng có mật độ điểm cao và phân tách các vùng mật độ thấp.

2.2.3.3. Đánh giá chất lượng cụm

Đánh giá chất lượng cụm giúp xác định mức độ "tốt" của kết quả phân cụm, tức là các cụm có tính đồng nhất cao bên trong và phân biệt rõ ràng giữa các cụm. Hai chỉ số phổ biến nhất là Silhouette Score và Davies-Bouldin Index.

- Silhouette Score: Đo lường mức độ tương đồng của một điểm với cụm của nó so với các cụm khác. Giá trị gần 1 cho thấy điểm đó được phân cụm tốt.
- Davies-Bouldin Index: Đánh giá chất lượng cụm bằng cách tính tỷ lệ giữa khoảng cách giữa các cụm và độ phân tán của các cụm. Giá trị thấp hơn cho thấy chất lượng phân cụm tốt hơn.

2.2.3.4. Tính toán lặp

Quy trình phân cụm thường bao gồm các bước lặp đi lặp lại, trong đó các điểm dữ liệu được gán vào các cụm dựa trên khoảng cách và sau đó cập nhật các trung tâm cụm. Quá trình này tiếp tục cho đến khi không còn sự thay đổi đáng kể nào trong việc gán cụm.

- Gán điểm: Với mỗi điểm dữ liệu, tính khoảng cách đến các cụm hiện có (thường là đến centroid hoặc đại diện cụm). Gán điểm vào cụm có khoảng cách nhỏ nhất.
- Cập nhật cụm: Tính lại đặc trưng đại diện của cụm (ví dụ tâm cụm trong K-means) dựa trên các điểm vừa được gán.

Quy trình này lặp lại cho đến khi không còn sự thay đổi đáng kể hoặc đạt giới hạn số vòng lặp.

2.2.3.5. Tham số điều chỉnh

Mỗi thuật toán phân cụm đều có những tham số cần thiết để hoạt động hiệu quả. Việc chọn đúng giá trị tham số là rất quan trọng:

- Số cụm (k): Số lượng cụm mong muốn trong thuật toán như K-means.
- Bán kính mật độ (eps): Trong DBSCAN, xác định khoảng cách tối đa để điểm được xem là láng giềng.
- Ngưỡng liên kết: Trong các thuật toán phân cụm phân cấp, quyết định mức độ kết nối giữa các cụm.

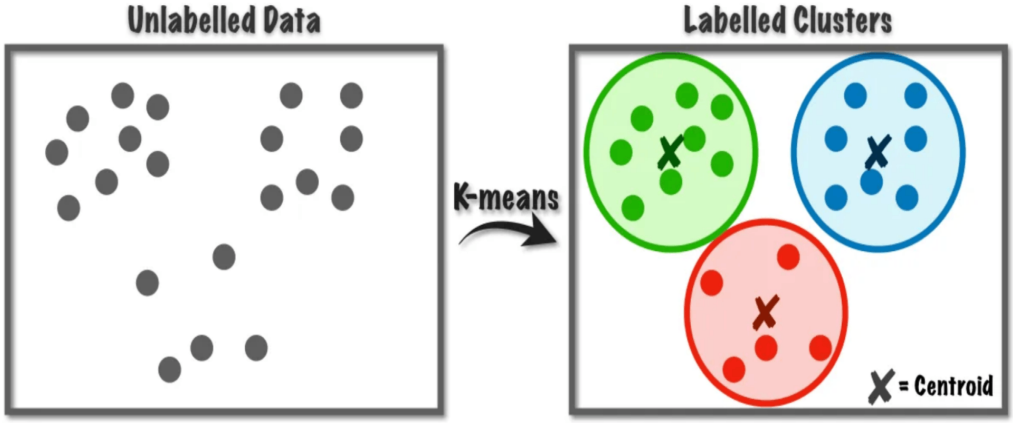
2.2.4. Các phương pháp phân cụm

2.2.4.1. K-Means

Thuật toán phân cụm K-means là một trong những phương pháp phổ biến và hiệu quả nhất trong lĩnh vực khai phá dữ liệu và học máy không giám sát. Mục tiêu chính của K-means là phân chia tập dữ liệu thành K cụm sao cho các điểm dữ liệu trong cùng một cụm có sự tương đồng cao nhất, đồng thời các cụm khác nhau càng khác biệt càng tốt. Thuật toán này hoạt động dựa trên việc tối đa hóa sự khác biệt giữa các cụm và đồng thời tối thiểu hóa sự phân tán trong từng cụm, giúp các điểm dữ liệu tương tự được nhóm lại gần nhau hơn trong không gian đặc trưng. Tuy nhiên, phương pháp này có chi phí tính toán khá cao, không thích hợp với các bộ dữ liệu lớn.[4]

Các bước thực hiện phân cụm K-means:

- Bước 1. Khởi tạo: Chọn ngẫu nhiên K điểm làm tâm cụm ban đầu.
- Bước 2. Gán điểm: Tính khoảng cách từ mỗi điểm dữ liệu đến các tâm cụm, gán điểm vào cụm gần nhất.
- Bước 3. Cập nhật tâm cụm: Tính lại vị trí trung bình (tâm) của các điểm trong mỗi cụm.
- Bước 4. Lặp lại: Lặp lại bước gán điểm và cập nhật tâm cụm cho đến khi tâm cụm không thay đổi hoặc đạt điều kiện dừng.
- Bước 5. Kết thúc: Khi thuật toán hội tụ, ta có các cụm dữ liệu ổn định.



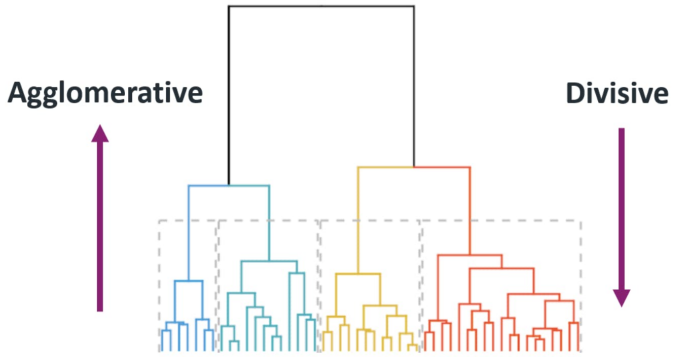
Hình 2-2: Thuật toán K-Means

2.2.4.2. Phân cụm phân cấp

Phân cụm phân cấp không yêu cầu xác định trước số cụm mà xây dựng một cấu trúc phân cấp các cụm thông qua quá trình hợp nhất hoặc phân chia. Phương pháp này cho phép người dùng quan sát cấu trúc dữ liệu ở nhiều mức độ khác nhau thông qua biểu đồ dendrogram. Có hai loại phương pháp này.[4]

- Kết tụ: Đây là phương pháp tiếp cận từ dưới lên trong đó mỗi quan sát được coi là một cụm riêng biệt lúc đầu và khi chúng ta di chuyển từ dưới lên trên, mỗi quan sát được hợp nhất thành từng cặp và các cặp được hợp nhất thành từng cụm.

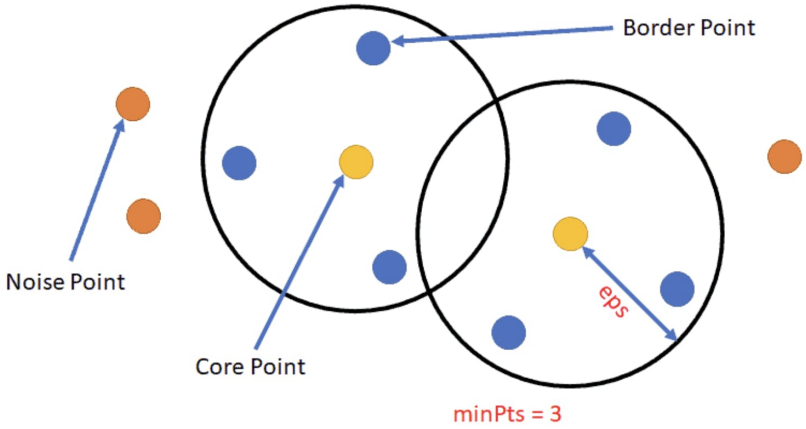
- Phân chia: Đây là phương pháp "từ trên xuống": tất cả các quan sát đều bắt đầu từ một cụm và việc phân chia được thực hiện đệ quy khi chúng ta di chuyển từ trên xuống dưới.



Hình 2-3: Phân cụm phân cấp

2.2.4.3. DBSCAN

DBSCAN là thuật toán phân cụm dựa trên mật độ điểm, không yêu cầu xác định trước số cụm. Thuật toán có khả năng phát hiện các cụm có hình dạng phức tạp và kích thước không đồng đều, đồng thời xử lý hiệu quả các điểm nhiễu (outliers). Đây là điểm mạnh giúp DBSCAN được ứng dụng rộng rãi trong các trường hợp dữ liệu thực tế có nhiều nhiễu hoặc cấu trúc phức tạp.[4]



Hình 2-4: Thuật toán DBSCAN

2.3. Thuật toán K-means Clustering

2.3.1. Thuật toán K-Means

K-means là một thuật toán học máy không giám sát, dùng để gán các điểm dữ liệu vào một trong K K K cụm dựa trên sự tương đồng về đặc điểm. Trong học máy không giám sát, dữ liệu không có nhãn phân loại (như trong học máy có giám sát). Thay vào đó, thuật toán tự động phát hiện các mẫu trong dữ liệu và nhóm các điểm có đặc điểm tương tự vào cùng một cụm.[5]

Ngoài K-means, còn có các thuật toán phân cụm khác như DBSCAN, Phân cụm Agglomerative, hay KNN. Tuy nhiên, K-means được ưa chuộng hơn nhờ tính đơn giản và hiệu quả.

Ý nghĩa của K: K đại diện cho số lượng cụm mà dữ liệu sẽ được chia thành.

Ví dụ:

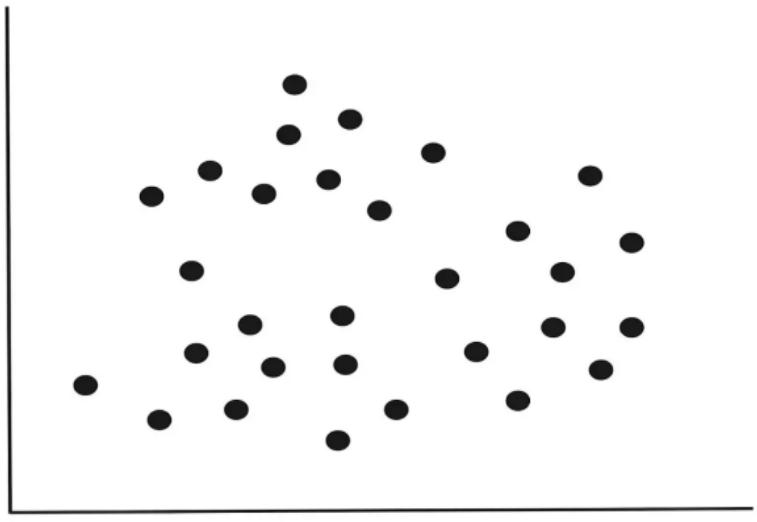
- Nếu K=3, dữ liệu được chia thành 3 cụm.
- Nếu K=5, dữ liệu được chia thành 5 cụm.

Việc chọn K phù hợp là yếu tố quan trọng, và chúng ta đã thảo luận về các phương pháp như khuỷu tay và hình bóng để xác định giá trị tối ưu trong phần trước.

2.3.2. Quy trình hoạt động của thuật toán K-Means

Thuật toán K-means là một phương pháp phân cụm dữ liệu, trong đó các điểm dữ liệu được chia thành các nhóm dựa trên sự tương đồng về đặc điểm. Quá trình bắt đầu bằng việc đặt ngẫu nhiên các điểm dữ liệu vào các nhóm, sau đó tính toán tâm điểm (centroid) của mỗi nhóm. Khoảng cách Euclid từ mỗi điểm dữ liệu đến các tâm điểm được đo lường, và nếu một điểm gần tâm điểm của nhóm khác hơn, nó sẽ được gán lại vào nhóm đó. Quá trình này lặp lại cho đến khi phương sai trong mỗi nhóm đạt mức tối thiểu, tức là các điểm trong cùng nhóm có đặc điểm càng giống nhau càng tốt.

Hãy hình dung một tập dữ liệu hai chiều, ví dụ như chiều cao và cân nặng của một nhóm người. Nếu thêm một biến như tuổi, tập dữ liệu sẽ trở thành không gian ba chiều, nhưng ở đây, chúng ta sẽ tập trung vào biểu đồ hai chiều dưới đây để minh họa.[5]



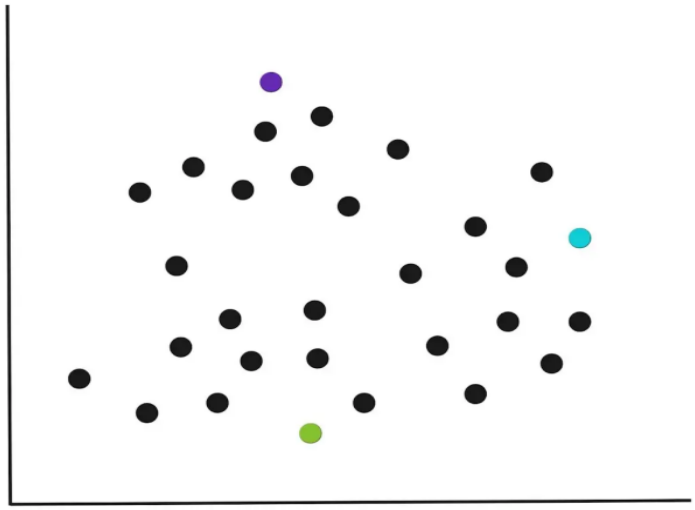
Hình 2-5: Biểu đồ dữ liệu ban đầu

2.3.2.1.Các bước thực hiện K-means

a) Bước 1: Khởi tạo

- Từ biểu đồ dữ liệu, ta có thể nhận thấy ba cụm tiềm năng. Khi xây dựng mô hình, ta chọn ngẫu nhiên K=3, nghĩa là dữ liệu sẽ được chia thành ba nhóm.

- Các tâm điểm ban đầu được chọn ngẫu nhiên, như trong hình dưới.



Hình 2-6: Khởi tạo tâm điểm

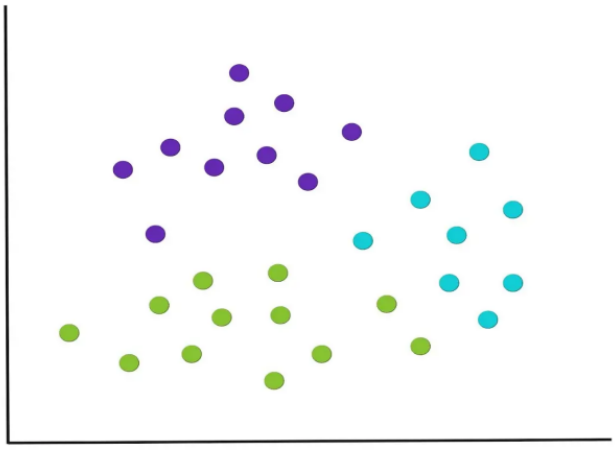
- Lưu ý: Có thể tự chọn số lượng cụm K, nhưng có cách tiếp cận tốt hơn để xác định K.

b) Bước 2: Gán điểm dữ liệu vào cụm

- Mỗi điểm dữ liệu được gán vào cụm có tâm điểm gần nhất, dựa trên khoảng cách Euclid:
- Công thức Euclid: Giả sử tính khoảng cách 2 điểm a và b

$$d = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

- Trong đó:
  - **d:** Khoảng cách Euclid giữa hai điểm a và b. Đây là độ dài đoạn thẳng nối hai điểm trong không gian 2 chiều.
  - x_a: Tọa độ X của điểm a trên trục X.
  - x_b: Tọa độ X của điểm b trên trục X.
  - y_a: Tọa độ Y của điểm a trên trục Y.
  - y_b: Tọa độ Y của điểm b trên trục Y.
- Công thức Euclid được dùng để tính khoảng cách của các điểm dữ liệu để xác định tâm điểm gần nhất



Hình 2-7: Phân cụm ban đầu

- Ngoài khoảng cách Euclid, các phép đo khác như khoảng cách Manhattan, tương quan Spearman, hoặc Pearson cũng có thể được sử dụng, nhưng Euclid và Manhattan là phổ biến nhất.

c) Bước 3: Cập nhật tâm điểm

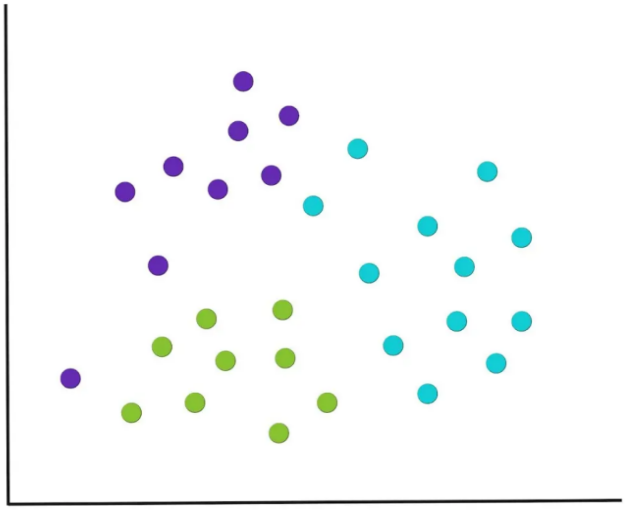
- Sau lần phân cụm đầu tiên, tâm điểm của mỗi cụm được tính lại dựa trên trung bình của các điểm trong cụm. Các điểm dữ liệu sau đó được gán lại vào cụm gần nhất.

$$x_{centroid} = \frac{\sum_{i=1}^{K} x_i}{K} \qquad y_{centroid} = \frac{\sum_{i=1}^{K} y_i}{K}$$

$$(x_i, y_i), i = 1 \ldots . K$$

- Trong đó:
  - $x_{centroid}$: Tọa độ X của tâm điểm (centroid) của một cụm trong không gian 2 chiều. Đây là giá trị trung bình của các tọa độ X của tất cả các điểm trong cụm.
  - $y_{centroid}$: Tọa độ Y của tâm điểm của cùng cụm đó. Đây là giá trị trung bình của các tọa độ Y của tất cả các điểm trong cụm.
  - $\frac{\sum_{i=1}^{K} x_i}{K}$: Tổng các tọa độ X của K điểm dữ liệu trong cụm x_i là tọa độ X của điểm thứ i trong cụm, và K là số điểm trong cụm.
  - $\frac{\sum_{i=1}^{K} y_i}{K}$: Tổng các tọa độ Y của K điểm dữ liệu trong cụm $y_i$là tọa độ Y của điểm thứ i trong cụm.
  - K : Số lượng điểm dữ liệu trong cụm (số điểm được sử dụng để tính trung bình). Lưu ý: Ở đây K không phải là số cụm tổng cộng mà là số điểm trong một cụm cụ thể.

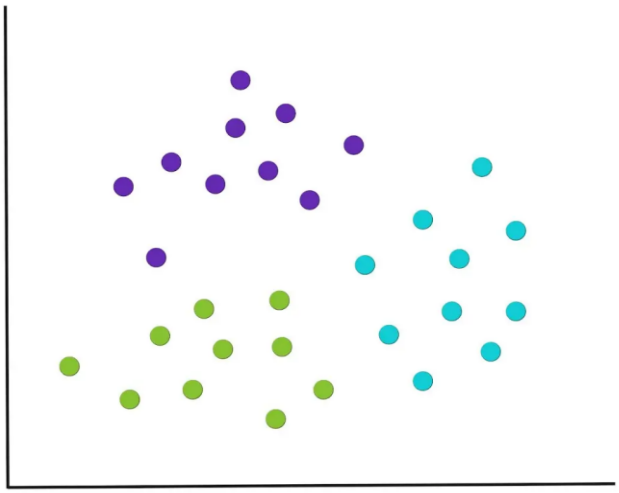
=> Vị trí các điểm gán cho centroid



Hình 2-8: Cụm sau khi cập nhật tâm điểm

d) Bước 4: Lặp lại

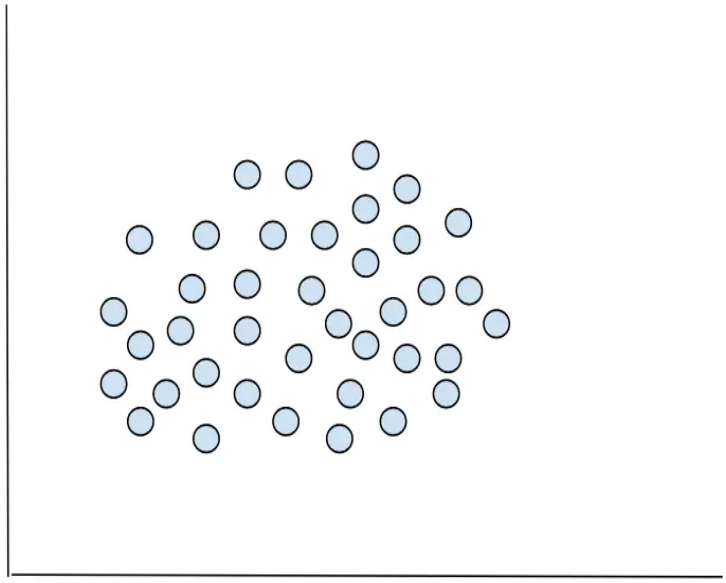
- Các bước 2 và 3 được lặp lại cho đến khi không còn điểm dữ liệu nào thay đổi cụm hoặc đạt đến số lần lặp tối đa. Kết quả cuối cùng là các cụm ổn định, như trong hình dưới.



Hình 2-9: Cụm cuối cùng

2.3.2.2.Lựa chọn số cụm K

Trong thực tế, dữ liệu thường phức tạp và không có ranh giới cụm rõ ràng.Với dữ liệu nhiều chiều hoặc khó trực quan hóa, việc xác định số cụm tối ưu trở nên thách thức. Hãy xem biểu đồ dưới đây:



Hình 2-10*: Dữ liệu phức tạp*

Từ biểu đồ này,Có thể xác định số cụm không? Rất khó. Vậy làm thế nào để chọn K phù hợp?

Có hai phương pháp phổ biến để tìm K tối ưu: phương pháp khuỷu tay (Elbow Method) và phương pháp hình bóng (Silhouette Method). Dưới đây là tóm tắt về cách chúng hoạt động.

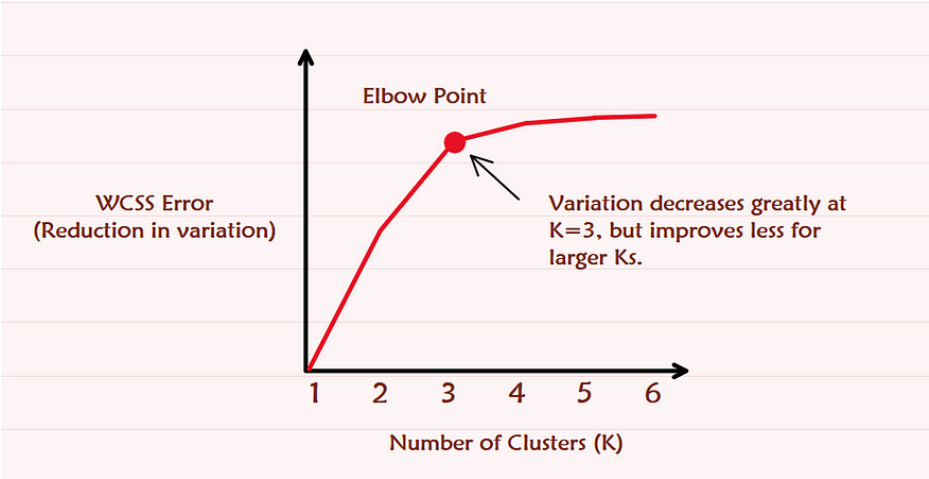
a) Phương pháp khuỷu tay((Elbow Method):

Phương pháp khuỷu tay là một kỹ thuật dùng để xác định số lượng cụm tối ưu (K) trong thuật toán K-means dựa trên tổng bình phương khoảng cách trong cụm, gọi là WCSS (Within-Cluster Sum of Squares). WCSS đo lường tổng phương sai của các điểm dữ liệu trong mỗi cụm, tức là tổng bình phương khoảng cách từ mỗi điểm đến tâm điểm của cụm mà nó thuộc về. Mục tiêu là tìm giá trị K sao cho WCSS nhỏ nhất, nghĩa là các điểm trong cùng cụm càng gần nhau càng tốt, nhưng không làm mô hình trở nên quá phức tạp.

$$\text{WCSS} = \sum_{\text{I=1}}^{\text{K}} \sum_{\text{x}\in\text{C}_\text{i}} ||x - \mu_\text{i}||^2$$

Trong đó:

- K: Số lượng cụm.
- C_i: Cụm thứ i
- x : Một điểm dữ liệu thuộc cụm C_i
- μ_i: Tâm điểm (centroid) của cụm C_i.
- $||x - \mu_i||^2$: Bình phương khoảng cách Euclid từ điểm x đến tâm điểm μi .
- $\sum_{x\in C_i} ||x - \mu_i||^2$: Tổng hợp khoảng cách bình phương cho tất cả các điểm trong cụm C_i.



Hình 2-11: Biểu đồ WCSS theo Số Lượng Cụm (Elbow Method)x`

Cách thực hiện:

- Chọn phạm vi giá trị K: Bắt đầu với một dãy giá trị K hợp lý, ví dụ từ 1 đến 10, tùy thuộc vào quy mô dữ liệu.
- Tính WCSS cho mỗi K :
  - Chạy thuật toán K-means với mỗi giá trị K.
  - Tính tâm điểm của mỗi cụm bằng trung bình tọa độ của các điểm trong cụm.
  - Tính tổng bình phương khoảng cách từ mỗi điểm đến tâm điểm của cụm mà nó thuộc về, rồi cộng lại để được WCSS.
- Vẽ biểu đồ WCSS theo K: Đặt K trên trục hoành (x-axis) và WCSS trên trục tung (y-axis). Biểu đồ thường cho thấy WCSS giảm khi K K K tăng, vì càng nhiều cụm thì khoảng cách trong cụm càng nhỏ.
- Xác định điểm khuỷu tay:
  - Điểm khuỷu tay là nơi đường cong WCSS bắt đầu phẳng dần, nghĩa là việc tăng K thêm không giảm đáng kể WCSS nữa.
  - Đây là giá trị K tối ưu, vì nó cân bằng giữa việc giảm phương sai trong cụm và tránh tạo ra quá nhiều cụm không cần thiết, giữ mô hình đơn giản.
- Phương pháp này giúp tìm số cụm cân bằng giữa độ chính xác và sự đơn giản của mô hình.

b) Phương pháp hình bóng (Silhouette Method)

Phương pháp này đánh giá chất lượng phân cụm bằng cách đo mức độ tương đồng của một điểm với cụm của nó so với các cụm khác. Kết quả là hệ số hình bóng, một chỉ số định lượng để chọn K tối ưu.

Cách thực hiện:

Bước 1: Chọn một dãy giá trị K , bắt đầu từ 2 (vì cần ít nhất 2 cụm để so sánh).

Bước 2: Với mỗi K K:

- Tính độ tương đồng cụm: Khoảng cách trung bình từ một điểm đến tất cả các điểm khác trong cùng cụm.

- Tính độ khác biệt cụm: Khoảng cách trung bình từ điểm đó đến tất cả các điểm trong cụm gần nhất khác.
- Tính hệ số hình bóng cho mỗi điểm theo công thức:

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

Trong đó:

- s(i) : Hệ số hình bóng của điểm dữ liệu thứ i .Giá trị này nằm trong khoảng từ -1 đến 1:
  - Gần 1: Điểm i rất giống với các điểm trong cùng cụm và khác biệt với các cụm khác (phân cụm tốt).
  - Gần 0: Điểm i nằm gần ranh giới giữa các cụm.
  - Gần -1: Điểm i có thể được gán sai cụm (phân cụm kém).
- $a(i)$: Khoảng cách trung bình từ điểm i đến tất cả các điểm khác trong cùng cụm với i. Giá trị này đo lường mức độ tương đồng của điểm i với cụm của nó (càng nhỏ càng tốt).
- $b(i)$: Khoảng cách trung bình từ điểm i đến tất cả các điểm trong cụm gần nhất khác (không phải cụm của i). Giá trị này đo lường mức độ khác biệt của điểm i với các cụm khác (càng lớn càng tốt).
- $b(i) - a(i)$: Hiệu giữa $b(i)$ và $a(i)$, thể hiện mức độ điểm i được phân cụm tốt (lớn hơn 0 là tốt, nhỏ hơn 0 là kém).
- $max\{a(i), b(i)\}$: Giá trị lớn nhất giữa $a(i)$ và $b(i)$, dùng để chuẩn hóa $s(i)$ về khoảng [-1, 1].

Bước 3: Tính hệ số hình bóng trung bình cho toàn bộ tập dữ liệu với mỗi K.

Bước 4: Chọn K có hệ số hình bóng trung bình cao nhất. Hệ số này dao động từ -1 (phân cụm kém) đến 1 (phân cụm tốt).

Phương pháp hình bóng đảm bảo các cụm vừa chặt chẽ (các điểm trong cụm giống nhau) vừa phân tách rõ rệt (các cụm khác nhau).[6]

2.3.3. Ưu và nhược điểm của thuật toán K-Means Clustering

Ưu điểm:

- Dễ hiểu và triển khai: K-Means là một thuật toán đơn giản, dễ áp dụng, phù hợp cho các bài toán phân cụm cơ bản, đặc biệt với người mới bắt đầu.
- Hiệu quả với dữ liệu lớn: Thuật toán xử lý tốt trên các tập dữ liệu lớn và có khả năng mở rộng khi số lượng cụm không quá nhiều.
- Không cần nhãn dữ liệu: Là thuật toán không giám sát, K-Means tự động phát hiện các mẫu trong dữ liệu chưa được gắn nhãn, linh hoạt áp dụng cho nhiều lĩnh vực khác nhau.

Nhược điểm:

- Phụ thuộc vào tâm điểm ban đầu: Việc chọn tâm điểm khởi tạo ngẫu nhiên có thể ảnh hưởng đến kết quả, dẫn đến phân cụm không tối ưu nếu chọn không tốt.
- Nhạy cảm với nhiễu và ngoại lai: Dữ liệu nhiễu hoặc điểm ngoại lai có thể làm sai lệch tâm điểm và ranh giới cụm, gây ảnh hưởng đến chất lượng phân cụm.
- Hạn chế về hình dạng cụm: K-Means giả định các cụm có hình dạng cầu lồi và kích thước tương đương, nên không hiệu quả với dữ liệu có cụm không lồi hoặc hình dạng phức tạp.

2.4 Phân tích RFM

2.4.1.Phân tích RFM: Phân đoạn khách hàng dựa trên hành vi giao dịch

Phân tích RFM (Recency, Frequency, Monetary) là một phương pháp phân đoạn khách hàng dựa trên dữ liệu giao dịch trong lịch sử, giúp doanh nghiệp hiểu rõ hành vi và giá trị của từng nhóm khách hàng. RFM tập trung vào ba chỉ số chính:

- Recency (Thời gian gần nhất): Khách hàng đã mua hàng gần đây nhất vào thời điểm nào? Khoảng thời gian từ lần giao dịch cuối cùng đến hiện tại là bao lâu? Khách hàng càng giao dịch gần đây thì càng có giá trị với doanh nghiệp, bởi họ có khả năng tiếp tục mua sắm cao hơn. Ngược lại, nếu khoảng thời gian này dài, doanh nghiệp có thể cần triển khai các chiến lược như khuyến mãi, upsell (bán thêm) hoặc cross-sell (bán chéo) để thu hút họ quay lại.
- Frequency (Tần suất): Khách hàng mua sắm thường xuyên đến mức nào? Tần suất giao dịch phản ánh mức độ trung thành của khách hàng với doanh nghiệp. Những khách hàng mua sắm thường xuyên (ví dụ: 3 lần/tháng) thường có giá trị cao hơn so với những khách hàng chỉ mua 1 lần/tháng. Tần suất giao dịch cao cũng cho thấy tiềm năng lớn hơn trong việc áp dụng các chương trình khuyến mãi hoặc xây dựng mối quan hệ lâu dài.
- Monetary (Giá trị tiền tệ): Tổng số tiền mà khách hàng đã chi tiêu là bao nhiêu? Giá trị này cho biết mức độ đóng góp tài chính của khách hàng trong một khoảng thời gian nhất định. Khách hàng chi tiêu nhiều thường mang lại doanh thu cao hơn, nhưng doanh nghiệp cũng cần xem xét tần suất và thời gian giao dịch để đánh giá đầy đủ giá trị của họ. Ví dụ, khách hàng chi tiêu lớn nhưng chỉ mua một lần có thể không bền vững bằng khách hàng chi tiêu đều đặn với giá trị trung bình.[7]

2.4.2. Làm thế nào để phân tích RFM cho phân khúc khách hàng hiệu quả?

Để phân tích RFM hiệu quả và tạo ra các phân khúc khách hàng có ý nghĩa, doanh nghiệp cần thực hiện các bước sau một cách chi tiết và có hệ thống:

Bước 1: Thu thập và làm sạch dữ liệu giao dịch

Đầu tiên, doanh nghiệp cần thu thập dữ liệu giao dịch từ hệ thống CRM, phần mềm bán hàng, hoặc cơ sở dữ liệu nội bộ. Dữ liệu cần bao gồm: thông tin khách hàng (ID, tên), ngày giao dịch, số lượng giao dịch, và tổng số tiền chi tiêu. Sau đó, làm sạch dữ liệu bằng cách loại bỏ các bản ghi trùng lặp, xử lý giá trị thiếu, và đảm bảo tính nhất quán (ví dụ: định dạng ngày tháng, đơn vị tiền tệ). Dữ liệu sạch giúp đảm bảo kết quả phân tích chính xác.

Bước 2: Tính toán các chỉ số RFM

Với mỗi khách hàng, tính ba chỉ số RFM:

- Recency: Lấy ngày hiện tại trừ đi ngày giao dịch gần nhất. Ví dụ, nếu ngày hiện tại là 05/05/2025 và khách hàng mua lần cuối vào 01/05/2025, Recency là 4 ngày.
- Frequency: Đếm tổng số lần giao dịch trong một khoảng thời gian nhất định (thường là 6 tháng hoặc 1 năm). Ví dụ, nếu khách hàng mua 10 lần trong 1 năm, Frequency là 10.
- Monetary: Cộng tổng số tiền chi tiêu trong cùng khoảng thời gian. Ví dụ, nếu khách hàng chi 5 triệu đồng trong 1 năm, Monetary là 5 triệu đồng.

Bước 3: Chấm điểm RFM

Chia mỗi chỉ số thành các khoảng và gán điểm (thường từ 1 đến 5) để chuẩn hóa dữ liệu:

- Sắp xếp khách hàng theo Recency (tăng dần), Frequency (giảm dần), và Monetary (giảm dần).
- Chia thành 5 nhóm bằng nhau (quintile). Ví dụ, với Recency: nhóm 20% khách hàng mua gần nhất được điểm 5, nhóm 20% tiếp theo được điểm 4, ..., nhóm xa nhất được điểm 1. Tương tự cho Frequency và Monetary: nhóm mua nhiều nhất/chi nhiều nhất được điểm 5, ít nhất được điểm 1.
- Kết quả: Mỗi khách hàng sẽ có bộ ba điểm RFM, ví dụ (5, 4, 3) – Recency 5, Frequency 4, Monetary 3.

Bước 4: Phân khúc khách hàng

Kết hợp điểm RFM để phân loại khách hàng thành các nhóm có ý nghĩa:

- Khách hàng trung thành: Điểm RFM cao, ví dụ (5, 5, 5) hoặc (4, 5, 4) – mua gần đây, thường xuyên, và chi tiêu lớn.
- Khách hàng tiềm năng: Điểm Recency và Frequency cao nhưng Monetary trung bình, ví dụ (5, 4, 3) – có thể khuyến khích chi tiêu nhiều hơn.
- Khách hàng nguy cơ rời bỏ: Điểm Recency thấp, ví dụ (1, 3, 2) – đã lâu không mua, cần chiến lược tái kích hoạt.
- Khách hàng mới: Recency cao nhưng Frequency và Monetary thấp, ví dụ (5, 1, 1) – mới mua lần đầu, cần nuôi dưỡng để tăng tần suất. Có thể tạo 8-11 phân khúc tùy vào mục tiêu kinh doanh, ví dụ thêm nhóm "khách hàng chi lớn nhưng không thường xuyên" (ví dụ: 3, 2, 5).

Bước 5: Phân tích và xây dựng chiến lược

Dựa trên các phân khúc, phân tích đặc điểm và hành vi của từng nhóm để xây dựng chiến lược phù hợp. Ví dụ, nhóm trung thành có thể được ưu tiên trong chương trình khách hàng thân thiết, trong khi nhóm nguy cơ rời bỏ cần ưu đãi hoặc khảo sát để hiểu lý do họ không quay lại. Sử dụng các công cụ như Excel, Python, hoặc phần mềm BI (Power BI, Tableau) để trực quan hóa và theo dõi hiệu quả phân khúc.

Bước 6: Theo dõi và điều chỉnh

Phân tích RFM không phải là một lần duy nhất. Doanh nghiệp nên định kỳ (hàng tháng hoặc hàng quý) cập nhật dữ liệu và tính toán lại RFM để phản ánh thay đổi trong hành vi khách hàng. Ví dụ, một khách hàng trung thành có thể trở thành "nguy cơ rời bỏ" nếu họ ngừng mua sắm, và doanh nghiệp cần điều chỉnh chiến lược kịp thời.[8]

2.4.3. Lợi ích của phân tích RFM

Phân tích RFM mang lại nhiều lợi ích thiết thực cho doanh nghiệp:

- Phân khúc khách hàng chính xác: RFM chia khách hàng thành các nhóm cụ thể dựa trên hành vi thực tế, giúp doanh nghiệp hiểu rõ từng nhóm.
- Tăng doanh thu thông qua cá nhân hóa: Triển khai chiến dịch upsell, cross-sell phù hợp, ví dụ khuyến khích nhóm tiềm năng chi tiêu nhiều hơn.
- Dự đoán hành vi khách hàng: Dữ liệu RFM giúp dự đoán khả năng mua sắm trong tương lai, từ đó tối ưu hóa chiến lược tiếp cận.
- Cải thiện tỷ lệ giữ chân và giá trị vòng đời khách hàng (CLTV): Tập trung vào nhóm giá trị cao để tăng tỷ lệ giữ chân và doanh thu bền vững.
- Tiết kiệm chi phí marketing: Tập trung nguồn lực vào các nhóm tiềm năng, giảm lãng phí ngân sách.

2.4.4. Vai trò của RFM khi phân cụm với K-Means

Phân tích RFM (Recency, Frequency, Monetary) đóng vai trò quan trọng khi kết hợp với thuật toán K-Means để phân cụm khách hàng, giúp doanh nghiệp hiểu sâu hơn về hành vi và giá trị của từng nhóm. Dưới đây là các vai trò cụ thể:

- Cung cấp dữ liệu đầu vào có ý nghĩa**:** RFM tạo ra ba chỉ số (Recency, Frequency, Monetary) từ dữ liệu giao dịch, cung cấp thông tin chi tiết về thời gian mua gần nhất, tần suất mua sắm, và giá trị chi tiêu. Khi sử dụng K-Means, các chỉ số này được sử dụng làm đặc trưng (features) để phân cụm khách hàng, thay vì chỉ dựa vào dữ liệu thô. Ví dụ, một khách hàng với Recency = 5 ngày, Frequency = 10 lần, Monetary = 5 triệu đồng sẽ là một vector đặc trưng (5, 10, 5), giúp K-Means phân loại chính xác hơn.
- Tăng hiệu quả phân cụm: K-Means phân chia dữ liệu thành K cụm dựa trên khoảng cách Euclid giữa các điểm. RFM chuẩn hóa các chỉ số (thường chấm điểm từ 1 đến 5) để đảm bảo chúng có cùng thang đo, tránh trường hợp Monetary (giá trị lớn) lấn át Recency và Frequency (giá trị nhỏ). Điều này giúp K-Means tạo ra các cụm cân bằng, phản ánh đúng hành vi khách hàng, như nhóm trung thành (điểm cao cả 3 chỉ số) hoặc nhóm ít hoạt động (Recency thấp).

- Hỗ trợ xác định số cụm tối ưu: Khi áp dụng phương pháp khuỷu tay (Elbow Method) với K-Means, RFM cung cấp dữ liệu để tính WCSS (Within-Cluster Sum of Squares). Các chỉ số RFM giúp đánh giá mức độ tương đồng trong cụm, từ đó xác định điểm khuỷu tay – nơi tăng thêm cụm không cải thiện đáng kể chất lượng phân cụm. Ví dụ, nếu WCSS giảm mạnh từ K=2 đến K=3 nhưng gần như không đổi từ K=4 trở đi, K=3 là lựa chọn tối ưu.
- Phân tích hành vi khách hàng sâu hơn: Sau khi K-Means phân cụm, RFM giúp giải thích ý nghĩa của từng cụm. Ví dụ, cụm có trung bình Recency thấp, Frequency cao, Monetary cao có thể là nhóm "khách hàng trung thành"; cụm có Recency cao, Frequency và Monetary thấp có thể là "khách hàng mới". Điều này hỗ trợ doanh nghiệp xây dựng chiến lược cụ thể, như ưu đãi cho nhóm trung thành hoặc khuyến khích nhóm mới tăng chi tiêu.
- Cải thiện tính linh hoạt và ứng dụng thực tiễn: RFM kết hợp với K-Means cho phép doanh nghiệp điều chỉnh số cụm (K) dựa trên mục tiêu kinh doanh, ví dụ: 3-5 cụm để quản lý dễ dàng. Kết quả phân cụm có thể được áp dụng trong marketing (email cá nhân hóa), giữ chân khách hàng, hoặc tối ưu hóa chiến lược bán hàng, tăng hiệu quả kinh doanh.[8]

2.4.5. Ứng dụng thực tiễn của RFM

Phân tích RFM có thể áp dụng trong nhiều khía cạnh kinh doanh:

- Email marketing cá nhân hóa: Gửi email phù hợp với từng nhóm, ví dụ ưu đãi cho nhóm nguy cơ rời bỏ, tăng tỷ lệ nhấp 20-30% qua các công cụ như MailChimp.
- Giới thiệu sản phẩm mới: Mời nhóm trung thành dùng thử sản phẩm để nhận phản hồi tích cực và tăng doanh số.
- Giữ chân khách hàng: Phát hiện nhóm nguy cơ rời bỏ để gửi ưu đãi, ví dụ "Nhận 20% giảm giá cho lần mua tiếp theo".
- Chương trình khách hàng thân thiết: Tặng điểm hoặc quà cho nhóm trung thành, ví dụ ly cà phê miễn phí sau 10 lần mua.
- Tối ưu hóa chiến lược bán hàng: Tập trung vào nhóm chi lớn nhưng không thường xuyên để tăng tần suất mua sắm.

CHƯƠNG 3: THỰC NGHIỆM

3.1. Chọn bộ dữ liệu

Bộ dữ liệu Online Retail II(2011) được sử dụng trong đồ án được trích xuất từ tập dữ liệu Online Retail II, thu thập bởi một công ty bán lẻ trực tuyến có trụ sở tại Vương quốc Anh. Dữ liệu này ghi nhận chi tiết các giao dịch mua bán trong năm 2011, bao gồm thông tin về hóa đơn, sản phẩm, khách hàng và quốc gia mua hàng.

Thông tin các trường dữ liệu:

- Invoice (object): Là mã số hóa đơn của mỗi giao dịch. Nếu mã bắt đầu bằng chữ "C", đó là hóa đơn bị hủy.
- StockCode (object): Mã định danh của sản phẩm được mua.
- Description (object): Tên hoặc mô tả sản phẩm.
- Quantity (int): Số lượng sản phẩm được mua trong mỗi giao dịch.
- InvoiceDate (datetime): Ngày và giờ giao dịch được thực hiện.
- Price (float): Giá đơn vị của sản phẩm, được tính bằng bảng Anh (GBP).
- Customer ID (float): Mã định danh khách hàng, có thể bị thiếu trong một số giao dịch.
- Country (object): Quốc gia nơi khách hàng sinh sống.

3.2. Chuẩn bị dữ liệu

3.2.1. Hiển thị thông tin cơ bản

Bảng 3-1: Hiển thị thông tin cơ bản

Column	Non-Null Count	Dtype
Invoice	499429 non-null	object
StockCode	499429 non-null	object
Description	498100 non-null	object
Quantity	499429 non-null	int64
InvoiceDate	499429 non-null	object
Price	499429 non-null	float64
Customer ID	379980 non-null	float64
Country	499429 non-null	object

Nhận xét:

Bộ dữ liệu bao gồm 499.429 dòng và 8 cột, trong đó:

- Invoid: Mã hóa đơn
- StockCode: Mã sản phẩm
- Description: Mô tả sản phẩm
- Quantity: Số lượng bán
- InvoiceDate: Ngày giao dịch
- Price: Giá mỗi đơn vị
- Customer ID: Mã khách hàng
- Country: Quốc gia

Kiểu dữ liệu của từng cột

- Kiểu chuỗi (object): Invoice, StockCode, Description, InvoiceDate và Country
- Kiểu số nguyên (int64): Quantity
- Kiểu số thực (float64): Price, Customer ID

3.2.2. Thống kê mô tả

Bảng 3-2: Thống kê mô tả

	Quantity	Price	Customer ID
count	499429.000000	499429.000000	379980.000000
mean	9.679500	4.481720	15271.305856
std	226.515501	92.194676	1710.632743
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13923.000000
50%	3.000000	2.080000	15116.000000
75%	10.000000	4.130000	16770.000000
max	80995.000000	38970.000000	18287.000000

Nhận xét:

- Quantity (Số lượng)
  - Mean: 9.08 | Std: 483.34 ⇒ Độ lệch chuẩn cực cao, nhiều giá trị bất thường.
  - Min: -74,215 | Max: 74,215 ⇒ Có giá trị âm và cực đại, có thể là do trả hàng hoặc lỗi nhập liệu.
  - Phân vị (25%, 50%, 75%): 1, 3, 9 ⇒ Phần lớn đơn hàng có số lượng nhỏ.
- Price (Giá):
  - Mean: 4.76 | Std: 109.33 ⇒ Độ lệch chuẩn rất cao so với trung bình, cho thấy có nhiều giá trị ngoại lệ (sản phẩm có giá rất cao so với phần lớn các mặt hàng khác).
  - Min: 0.00 | Max: 16,888.02 ⇒ Có giá trị bằng 0 (có thể là hàng khuyến mãi hoặc lỗi dữ liệu) và giá trị cực đại rất lớn, cần kiểm tra xem có phải lỗi nhập liệu hay hàng hóa đặc biệt.
  - Phân vị (25%, 50%, 75%): 1.25, 2.46, 4.13 ⇒ Phần lớn sản phẩm có giá thấp dưới 5 đơn vị, cho thấy đây là cửa hàng bán lẻ hàng hóa có giá rẻ là chủ yếu.
- Customer ID (Mã khách hàng)
  - Mean: 15215.64 | Std: 1789.32 ⇒ Không quan trọng lắm vì đây chỉ là số định danh, nhưng cho thấy các ID phân bố khá đều trong một khoảng nhất định.
  - Min: 12,346 | Max: 18283 ⇒ Dãy số ID có vẻ là do hệ thống sinh ra và liên tục, không có gì bất thường.
  - Phân vị (25%, 50%, 75%): 13694, 15046, 16904 ⇒ Customer ID phân bố đều, không nghiêng về một nhóm mã nhất định.

- Kết luận
  - Dữ liệu chứa nhiều giá trị ngoại lệ ở Quantity và Price, có thể do trả hàng hoặc lỗi nhập liệu.
  - Phần lớn sản phẩm có giá rẻ và đơn hàng có số lượng nhỏ, cho thấy đây là giao dịch bán lẻ.
  - Cần tiền xử lý dữ liệu trước khi phân tích sâu: loại bỏ giá trị âm/0 và xử lý dữ liệu thiếu.

3.3. Xử lí dữ liệu

Phần này mô tả quy trình xử lý dữ liệu để đảm bảo chất lượng và tính toàn vẹn của bộ dữ liệu trước khi thực hiện phân tích RFM (Recency, Frequency, Monetary) và phân cụm khách hàng. Các bước bao gồm xử lý giá trị thiếu, chuyển đổi kiểu dữ liệu, kiểm tra tính toàn vẹn, loại bỏ giao dịch không hợp lệ, xử lý giá trị âm, chuẩn hóa cột Country, và tạo cột OrderValue.

3.3.1. Xử lí giá trị thiếu

Kiểm tra dữ liệu bị thiếu bằng hàm isnull() và cho ra bảng kết quả sau:

Bảng 3-3: Dữ liệu thiếu

Invoice	0
StockCode	0
Description	1329
Quantity	0
InvoiceDate	0
Price	0
Customer ID	119449
Country	0
OrderValue	0

Nhận xét:

- Bộ dữ liệu có 2 cột bị thiếu:
  - Description: có 1329 giá trị thiếu
  - Customer ID: có 119449 giá trị
- Cách xử lí
  - Xóa các dòng thiếu CustomerID vì để tính RFM, cần nhóm dữ liệu theo CustomerID. Các giao dịch thiếu CustomerID không thể gán vào một khách hàng cụ thể, do đó không đóng góp vào phân tích RFM hoặc phân cụm.
  - Điền giá trị "Unknown" cho các giá trị thiếu trong cột Description vì các dòng này vẫn chứa thông tin quan trọng ở các cột khác (Customer ID, InvoiceDate, v.v.) và tỷ lệ thiếu rất nhỏ.
  - Kiểm tra sau xử lý: Sau khi áp dụng các phương pháp trên, bộ dữ liệu không còn giá trị thiếu ở bất kỳ cột nào.

3.3.2. Chuyển đổi kiểu dữ liệu

Để đảm bảo dữ liệu phù hợp với các phép toán và mục đích phân tích, các cột được chuyển đổi sang kiểu dữ liệu phù hợp:

- InvoiceDate**:**
  - Ban đầu: Kiểu object (chuỗi), không hỗ trợ các phép toán liên quan đến ngày giờ.
  - Chuyển đổi: Sang kiểu datetime để tính toán Recency (khoảng thời gian từ lần mua gần nhất đến thời điểm hiện tại).
  - Phương pháp: Sử dụng pd.to_datetime() trong Pandas.
- Customer ID:
  - Ban đầu: Kiểu float64 do sự hiện diện của giá trị thiếu (NaN).
  - Chuyển đổi: Sang kiểu string (chuỗi) và loại bỏ phần thập phân (".0") để đảm bảo định dạng danh nghĩa (mã khách hàng 5 chữ số).
  - Phương pháp: Sử dụng .astype(str) và xử lý chuỗi để loại bỏ ".0".
- Invoice và StockCode:
  - Ban đầu: Có thể ở dạng object hoặc int.
  - Chuyển đổi: Sang kiểu string để hỗ trợ kiểm tra định dạng danh nghĩa (ví dụ: kiểm tra ký tự "C" trong Invoice cho giao dịch bị hủy hoặc định dạng 5 chữ số của StockCode).
  - Phương pháp: Sử dụng .astype(str).

3.3.3. Kiểm tra tính toàn vẹn dữ liệu

Mục tiêu: Đảm bảo tính hợp lệ của các cột định danh chính: Invoice (mã hóa đơn), StockCode (mã sản phẩm), và Customer ID (mã khách hàng).

- Invoice có định dạng hợp lệ (6 chữ số hoặc "C" + 6 chữ số cho giao dịch bị hủy).
- StockCode không quá dài (>7 ký tự), không quá ngắn (<5 ký tự), và chứa chữ số/chữ cái.
- Customer ID có đúng 5 chữ số.

Kết quả kiểm tra StockCode: Phát hiện 1.805 mã StockCode không hợp lệ, bao gồm:

Bảng 3-4: Mã StockCode không hợp lệ

	StockCode
269	M
950	POST
1821	C2
1924	D
20027	BANK CHARGES
114714	PADS
275026	DOT
275027	CRUK

Nhận xét:

- Dựa trên danh sách "M", "POST", "C2", "D", "BANK CHARGES", "PADS", "DOT", "CRUK", đây là ý nghĩa của từng mã:
  - M (Manual): Giao dịch nhập tay, không phải sản phẩm.
  - POST (Postage): Phí bưu điện, không phải sản phẩm.
  - C2 (Carriage): Phí vận chuyển, không phải sản phẩm.
  - D (Discount): Giảm giá, không phải sản phẩm.
  - BANK CHARGES: Phí ngân hàng, không phải sản phẩm (đã phát hiện trước đó).
  - PADS (Pads to clear): Có thể là mã đặc biệt hoặc lỗi nhập liệu, không rõ ràng nhưng không giống mã sản phẩm.
  - DOT (DOTCOM Postage): Phí bưu điện trực tuyến, không phải sản phẩm.
  - CRUK (CRUK Commission): Hoa hồng hoặc phí liên quan đến tổ chức CRUK, không phải sản phẩm.
- Cách xử lí
  - Tất cả các mã này đều không phải sản phẩm thực tế và nên được loại bỏ để đảm bảo dữ liệu chỉ chứa các giao dịch mua sắm hợp lệ.

3.3.4. Loại bỏ giao dịch bị hủy

Lý do:

- Giao dịch bị hủy (Invoice bắt đầu bằng ký tự "C") thường liên quan đến trả hàng hoặc hủy đơn, không đóng góp vào các chỉ số RFM (Recency, Frequency, Monetary).
- Loại bỏ các giao dịch này đảm bảo chỉ giữ lại các giao dịch mua sắm hợp lệ.

Phương pháp:

- Lọc dữ liệu bằng điều kiện ~df['Invoice'].str.startswith('C') để loại bỏ các dòng có Invoice bắt đầu bằng "C".

Kết quả:

- Số dòng sau khi loại bỏ giao dịch bị hủy: 370.311 dòng

3.3.5. Xử lí giá trị âm và bằng 0

Lý do:

- Quantity âm hoặc bằng 0: Đại diện cho trả hàng, lỗi nhập liệu, hoặc giao dịch không hợp lệ, không phù hợp với phân tích RFM vì RFM yêu cầu giao dịch mua sắm với số lượng dương.
- Price âm hoặc bằng 0: Đại diện cho hoàn tiền, khuyến mãi, hoặc lỗi nhập liệu, không phản ánh giá trị sản phẩm thực tế.

Kiểm tra:

- Số dòng có Quantity âm hoặc bằng 0: 0 dòng.
- Số dòng có Price âm hoặc bằng 0: 30 dòng.

Cách xử lý:

- Loại bỏ tất cả các dòng có Price âm hoặc bằng 0 (30 dòng) vì chúng không hợp lệ cho phân tích.
- Phương pháp: Lọc dữ liệu bằng điều kiện df['Price'] > 0.

Kết quả:

- Số dòng sau khi loại bỏ giá trị âm và bằng 0: 370.281 dòng.

3.3.6. Thay thế tất cả giá trị không phải 'United Kingdom' thành 'Other'

Lý do:

- Bộ dữ liệu có số lượng giao dịch tập trung chủ yếu tại United Kingdom (330.073 giao dịch), trong khi các quốc gia khác chỉ chiếm một phần nhỏ (40.208 giao dịch).
- Để đơn giản hóa phân tích và tập trung vào thị trường chính, tất cả các giá trị không phải "United Kingdom" được thay thế bằng "Other".

Phương pháp:

- Sử dụng df.loc[df['Country'] != 'United Kingdom', 'Country'] = 'Other' để thay thế giá trị.

Kết quả:

- Phân bố cột Country sau khi chuẩn hóa: Country United Kingdom 330073, Other 40208.

3.3.7. Tạo cột OrderValue

Mục tiêu:

- Tạo cột OrderValue (giá trị đơn hàng) bằng cách nhân Quantity với Price để làm cơ sở tính Monetary (tổng giá trị mua sắm của mỗi khách hàng) trong phân tích RFM.

Phương pháp:

- Tạo cột mới: df['OrderValue'] = df['Quantity'] * df['Price'].

Bảng 3-5: Dữ liệu sau khi thêm cột OrderValue

Invoice	StockCode	Quantity	InvoiceDate	Price	Customer	ID Country	OrderValue
539993	22386	10	2011-01-04 10:00:00	1.95	13313	United Kingdom	19.5
539993	21499	25	2011-01-04 10:00:00	0.42	13313	United Kingdom	10.5
539993	21498	25	2011-01-04 10:00:00	0.42	13313	United Kingdom	10.5
539993	22379	5	2011-01-04 10:00:00	2.10	13313	United Kingdom	10.5
539993	20718	10	2011-01-04 10:00:00	1.25	13313	United Kingdom	12.5

Cột OrderValue là nền tảng để tính tổng giá trị giao dịch của từng khách hàng, từ đó xây dựng chỉ số Monetary trong RFM.

3.3.8. Kết luận

Sau các bước xử lý trên, bộ dữ liệu đã được làm sạch và chuẩn hóa:

- Không còn giá trị thiếu.
- Kiểu dữ liệu phù hợp với mục đích phân tích.
- Loại bỏ các giao dịch không hợp lệ (giao dịch bị hủy, mã StockCode không phải sản phẩm, Quantity/Price âm hoặc bằng 0).
- Cột Country được chuẩn hóa, tập trung vào thị trường chính (United Kingdom).
- Cột OrderValue được tạo để hỗ trợ tính toán RFM.

Bộ dữ liệu cuối cùng có 370.281 dòng, sẵn sàng cho các bước phân tích tiếp theo như tính toán RFM và phân cụm khách hàng.

3.4. Trực quan hóa dữ liệu

Để hiểu rõ hơn về đặc điểm của tập dữ liệu 'online_retail_2011_filtered' - giao dịch bán lẻ trực tuyến năm 2011, ta tiến hành trực quan hóa dữ liệu thông qua nhiều biểu đồ khác nhau. Các biểu đồ và phân tích sẽ giúp khám phá xu hướng giao dịch, hành vi mua hàng của khách hàng, đặc điểm sản phẩm bán ra, cũng như mối quan hệ giữa các yếu tố quan trọng như số lượng sản phẩm, đơn giá và tổng giá trị đơn hàng.

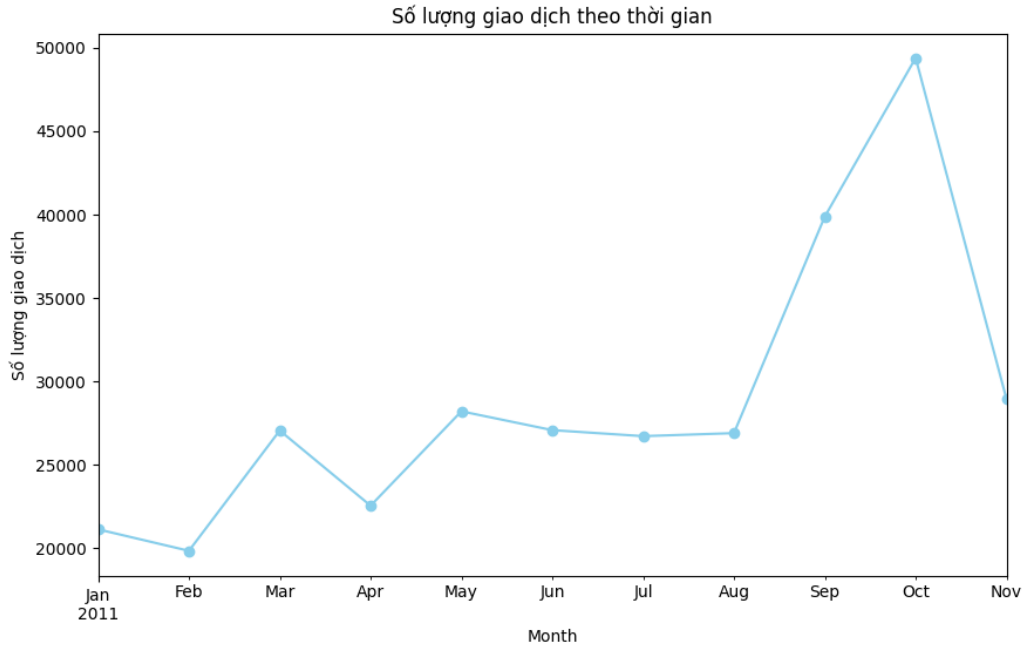
Các nội dung trực quan hóa bao gồm:

- Xu hướng số giao dịch theo thời gian: Phân tích sự biến động về số lượng giao dịch theo từng giai đoạn thời gian, nhằm phát hiện các mùa vụ hoặc xu hướng đặc biệt.
- Phân bố số giao dịch theo khách hàng: Khám phá hành vi mua sắm của khách hàng, đánh giá sự phân bố về tần suất mua hàng trong toàn bộ tập khách hàng.
- Phân phối số lượng sản phẩm bán ra: Phân tích mức độ phổ biến của các sản phẩm dựa trên tổng số lượng bán ra, nhằm nhận diện các sản phẩm chủ lực và đặc thù thị trường.
- Mối quan hệ giữa giá trị đơn hàng và giá sản phẩm: Nghiên cứu mối tương quan giữa tổng giá trị đơn hàng và đơn giá sản phẩm, từ đó hiểu rõ cách giá sản phẩm ảnh hưởng đến quy mô giao dịch.
- Tương quan giữa các yếu tố Quantity, Price, OrderValue: Đánh giá mối liên hệ giữa ba yếu tố chính trong giao dịch: số lượng sản phẩm mua, đơn giá, và tổng giá trị đơn hàng, nhằm phát hiện các xu hướng tiềm ẩn và hỗ trợ cho các phân tích sâu hơn.

3.4.1. Xu hướng số giao dịch theo thời gian

Phân tích số lượng giao dịch theo tháng giúp xác định các thời điểm cao điểm hoặc thấp điểm trong hoạt động mua sắm, từ đó cung cấp thông tin nền tảng cho việc phân cụm khách hàng. Kết quả của phân tích này sẽ được sử dụng như một đặc trưng bổ sung trong quá trình phân cụm, giúp phân loại khách hàng dựa trên tần suất và thời điểm mua sắm.

- Thư viện: Sử dụng thư viện pandas để xử lý dữ liệu và matplotlib.pyplot để vẽ biểu đồ.
- Nhóm theo tháng: Sử dụng dt.to_period('M') để trích xuất thông tin năm-tháng, tạo cột 'Month' với định dạng Y-M.
- Đếm giao dịch: Sử dụng groupby('Month').size() nhóm dữ liệu theo tháng và đếm số lượng giao dịch trong mỗi tháng, trả về một Series với chỉ số là các tháng và giá trị là số lượng giao dịch.
- Vẽ biểu đồ: Biểu đồ đường được tạo với kích thước 10x6 inch, sử dụng màu 'skyblue' và các điểm đánh dấu (marker='o'). Tiêu đề, nhãn trục x, y được thiết lập để tăng tính rõ ràng. Sử dụng xticks(rotation=0) đảm bảo các nhãn tháng hiển thị ngang.



Hình 3-1: Biểu đồ số lượng giao dịch theo tháng của năm 2011

Nhận xét:

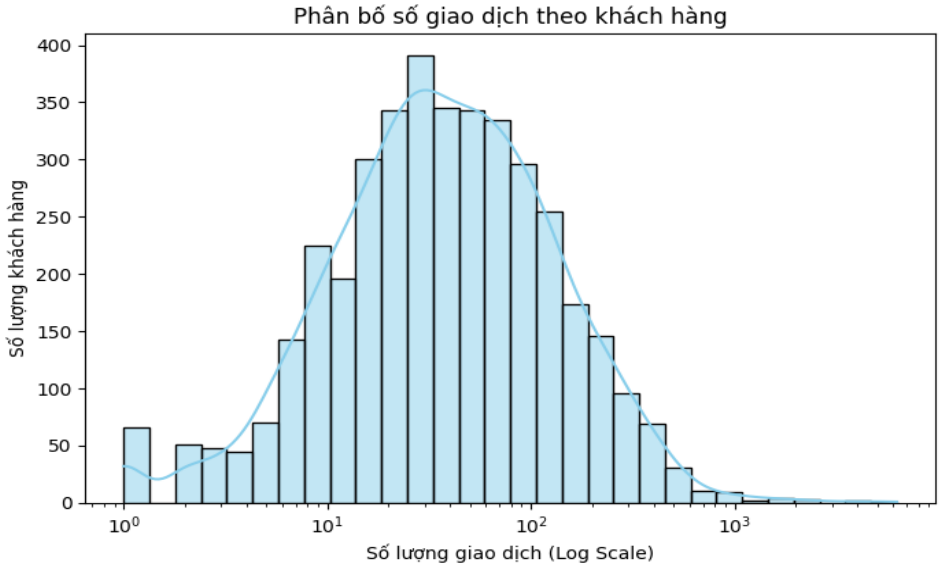
- Đầu năm (T1 - T4): Số giao dịch dao động nhẹ, ở mức từ khoảng 20.000 - 27.000 giao dịch. Có sự tăng nhẹ vào tháng 3, rồi giảm lại tháng 4.
- Giữa năm (T5 - T8):
  - Tháng 5 có sự tăng nhẹ (~28.000), sau đó giảm dần và ổn định quanh mức 27.000 giao dịch từ tháng 6 đến tháng 8.
- Cuối năm (T9 - T11):
  - Tăng mạnh: Từ tháng 9 bắt đầu tăng nhanh, từ khoảng 40.000 giao dịch, đến tháng 10 gần 50.000 và tháng 11 đạt đỉnh hơn 64.000 giao dịch.
  - Đây là giai đoạn có tốc độ tăng trưởng rất nhanh, có thể do nhu cầu cuối năm (ví dụ: mua sắm, lễ hội, khuyến mãi Black Friday, Giáng Sinh...).
- Tháng 12:
  - Số lượng giao dịch giảm mạnh đột ngột xuống dưới 20.000 giao dịch — thấp hơn cả đầu năm.

- Đây là điểm bất thường cần lưu ý. Nguyên nhân có thể:
    - Dữ liệu tháng 12 chưa đầy đủ (chỉ tính một phần tháng).
    - Hoặc đúng là có sự sụt giảm do đã "bùng nổ" ở tháng 11.

3.4.2. Phân bố số giao dịch theo khách hàng

Quá trình phân tích phân bố số giao dịch theo khách hàng, nhằm hiểu rõ mức độ tích cực của khách hàng trong hoạt động mua sắm. Phân tích này sử dụng biểu đồ histogram với thang logarit để trực quan hóa dữ liệu, giúp nhận diện các nhóm khách hàng dựa trên số lượng giao dịch.

- Thư viện:
    - pandas: Dùng để xử lý dữ liệu, cụ thể là đếm số giao dịch theo Customer ID.
    - seaborn: Dùng để vẽ histogram với đường mật độ kernel (KDE), cung cấp giao diện trực quan hóa chuyên nghiệp.
    - matplotlib.pyplot: Dùng để tùy chỉnh kích thước và các thuộc tính của biểu đồ.
- Đếm giao dịch: Sử dụng value_counts() trên cột 'Customer ID' đếm số lần xuất hiện của mỗi 'Customer ID', tương ứng với số lượng giao dịch của khách hàng đó.
- Vẽ histogram:
    - Hàm sns.histplot tạo histogram với dữ liệu là số lượng giao dịch của từng khách hàng.
    - Tham số log_scale=True áp dụng thang logarit cho trục x, giúp trực quan hóa dữ liệu có phân bố lệch.
    - kde=True thêm đường mật độ kernel để thể hiện hình dạng tổng quát của phân bố.
    - bins=30 chia dữ liệu thành 30 khoảng để hiển thị chi tiết.
    - color='skyblue' sử dụng màu xanh nhạt để tăng tính thẩm mỹ.



Hình 3-2: Biểu đồ phân phối số giao dịch theo khách hàng

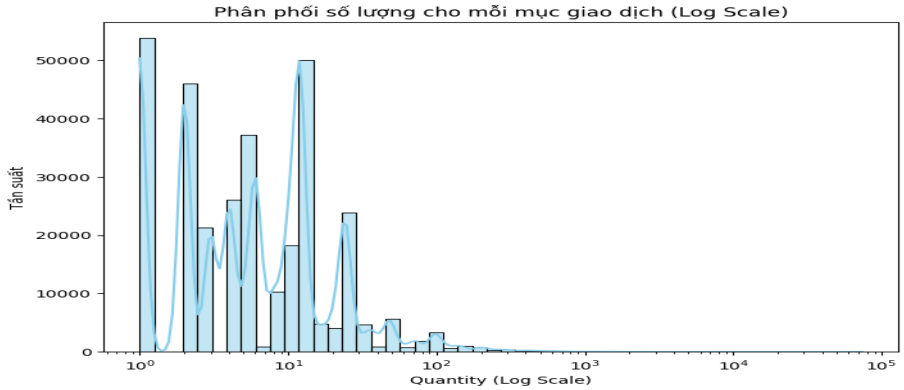
Nhận xét: Biểu đồ phân phối số giao dịch theo khách hàng, với trục x là số giao dịch (theo thang logarithmic) và trục y là số lượng khách hàng tương ứng.

- Đa số khách hàng có từ khoảng 10 đến 100 giao dịch.
  - Đây là vùng cao nhất trên biểu đồ (đỉnh khoảng 30–70 giao dịch).
  - Số lượng khách hàng đạt đỉnh gần 400 người ở khoảng 30-50 giao dịch.
- Phía bên phải:
  - Một lượng nhỏ khách hàng thực hiện hơn 1.000 giao dịch.
  - Số này giảm nhanh, cho thấy chỉ có rất ít khách hàng "cực kỳ trung thành" hoặc "hoạt động rất cao".
- Phía bên trái:
  - Cũng có một lượng kha khá khách chỉ thực hiện rất ít giao dịch (1–10 giao dịch).

3.4.3. Phân phối số lượng sản phẩm bán ra

Thông qua biểu đồ histogram với thang logarit ta phân tích được phân phối số lượng sản phẩm bán ra (số lượng sản phẩm trong mỗi mục giao dịch). Phân tích này giúp nhận diện các mẫu mua sắm, chẳng hạn như xu hướng mua số lượng lớn hay nhỏ, từ đó cung cấp thông tin hữu ích cho việc phân cụm khách hàng.

- Thư viện:
  - pandas: Dùng để truy cập và xử lý cột ‘Quantity’ trong dataframe.
  - seaborn: Dùng để vẽ histogram với đường mật độ kernel (KDE), cung cấp giao diện trực quan hóa chuyên nghiệp.
  - matplotlib.pyplot: Dùng để tùy chỉnh kích thước và các thuộc tính của biểu đồ.
- Vẽ histogram:
  - Hàm sns.histplot tạo histogram với dữ liệu từ cột ‘Quantity’.
  - Tham số log_scale=True áp dụng thang logarit cho trục x, giúp trực quan hóa dữ liệu có phân bố lệch.
  - kde=True thêm đường mật độ kernel để thể hiện hình dạng tổng quát của phân bố.
  - bins=50 chia dữ liệu thành 50 khoảng để hiển thị chi tiết các mức số lượng.
  - color='skyblue' sử dụng màu xanh nhạt để tăng tính thẩm mỹ.



Hình 3-3: Biểu đồ phân phối số lượng cho mỗi mục giao dịch

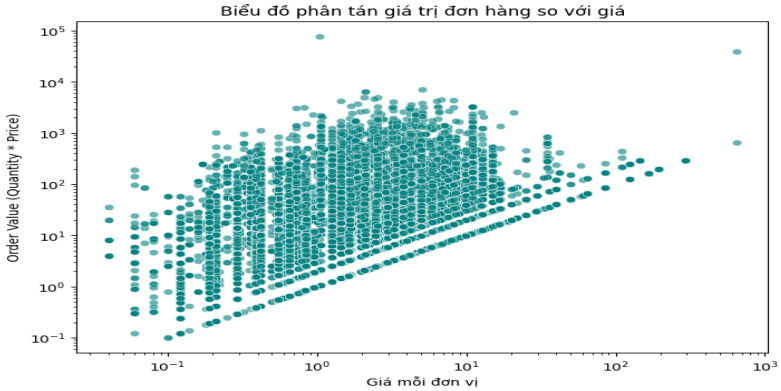
Nhận xét: Biểu đồ phân phối số lượng sản phẩm mỗi món giao dịch, với trục x là Quantity theo thang log và trục y là tần suất (Frequency). Các điểm nổi bật:

- Đỉnh lớn nhất: Số lượng sản phẩm là 1 món — chiếm tỷ lệ cao nhất (~65.000 lượt).
- Các đỉnh phụ khác:
  - Khoảng 2–3 sản phẩm, 5 sản phẩm, 10 sản phẩm và 20 sản phẩm.
  - Đây có thể là các "gói phổ biến" hoặc "mức mua quen thuộc" trong giao dịch.
- Từ sau khoảng 100 sản phẩm trở đi:
  - Số lượng giao dịch giảm rất mạnh.
  - Vẫn có một số lượng cực kỳ nhỏ giao dịch có số lượng sản phẩm lên đến hàng nghìn.

3.4.4. Mối quan hệ giữa các giá trị đơn hàng và giá sản phẩm

Phần này trình bày quá trình phân tích mối quan hệ giữa giá trị đơn hàng và giá sản phẩm thông qua biểu đồ phân tán (scatter plot) với thang logarit trên cả hai trục. Phân tích này giúp khám phá mức độ tương quan giữa hai biến và cung cấp thông tin hữu ích cho việc xây dựng các đặc trưng phân cụm.

- Thư viện:
  - pandas: Dùng để truy cập và xử lý cột Price và OrderValue trong dataframe.
  - seaborn: Dùng để vẽ biểu đồ phân tán với giao diện trực quan hóa chuyên nghiệp.
  - matplotlib.pyplot: Dùng để tùy chỉnh kích thước và các thuộc tính của biểu đồ.
- Vẽ scatter plot**:**
  - Hàm sns.scatterplot tạo biểu đồ phân tán với dữ liệu từ cột Price (trục x) và OrderValue (trục y).
  - Tham số alpha=0.6 thiết lập độ trong suốt cho các điểm, giúp dễ dàng quan sát các khu vực có mật độ điểm cao.
  - color='teal' sử dụng màu xanh lam để tăng tính thẩm mỹ.
  - plt.xscale('log') và plt.yscale('log') áp dụng thang logarit cho cả hai trục x và y, giúp trực quan hóa dữ liệu có phân bố lệch hoặc phạm vi giá trị lớn.



Hình 3-4: Biểu đồ phân tán giá trị đơn hàng so với giá

Nhận xét: Biểu đồ phân tán (scatter plot) thể hiện mối quan hệ giữa: Trục X: Giá mỗi đơn vị sản phẩm (Unit Price, theo thang log). Trục Y: Giá trị đơn hàng (Order Value = Quantity × Unit Price, cũng theo thang log).

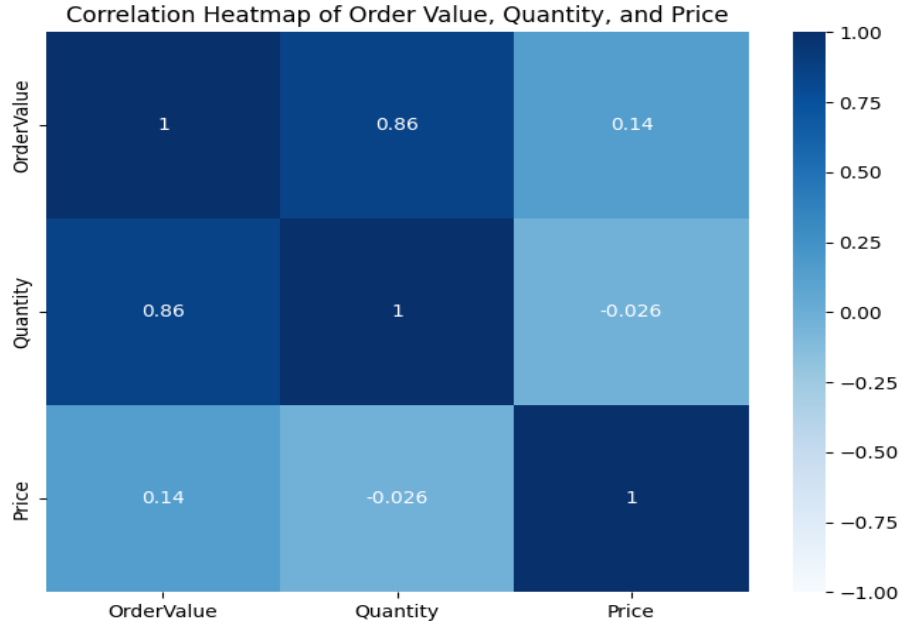
- Ý nghĩa xu hướng:
  - Khi giá mỗi đơn vị tăng, giá trị đơn hàng trung bình cũng tăng, nhưng không tuyến tính.
  - Các đơn hàng giá trị lớn có thể đến từ:
    - Đơn vị giá cao (sản phẩm đắt tiền) với số lượng vừa phải.
    - Hoặc sản phẩm giá thấp nhưng số lượng mua rất nhiều.
  - Đường xiên chéo phía dưới (các điểm xếp thành đường thẳng dọc theo cạnh dưới) đại diện cho những giao dịch với Quantity = 1:
    - Giá trị đơn hàng ≈ Giá mỗi đơn vị.
- Các cụm dữ liệu:
  - Tập trung nhiều nhất ở khoảng:
    - Giá đơn vị: từ 0.1 đến 10.
    - Giá trị đơn hàng: từ 1 đến 1000.
  - Ít giao dịch có giá trị cực cao (Order Value > 10,000).
  - Xuất hiện một số outlier (các điểm xa vùng tập trung), có thể là giao dịch cực lớn hoặc dữ liệu bất thường.

3.4.5. Tương quan giữa các yếu tố Quantity, Price, OrderValue

Trong quá trình trực quan hóa bộ dữ liệu này, phân tích tương quan giữa các yếu tố OrderValue (giá trị đơn hàng), Quantity (số lượng sản phẩm), và Price (giá mỗi đơn vị sản phẩm), nhằm hiểu rõ mức độ liên kết giữa các biến số này. Phần này trình bày quá trình phân tích tương quan thông qua ma trận tương quan và biểu đồ heatmap, giúp trực quan hóa mối quan hệ tuyến tính giữa ba biến, cung cấp thông tin hữu ích để lựa chọn các đặc trưng phù hợp cho thuật toán phân cụm và hiểu rõ hơn về hành vi mua sắm.

- Thư viện:
  - pandas: Dùng để chọn các cột số và tính ma trận tương quan.
  - seaborn: Dùng để vẽ biểu đồ heatmap với giao diện trực quan hóa chuyên nghiệp.
  - matplotlib.pyplot: Dùng để tùy chỉnh kích thước và các thuộc tính của biểu đồ.
- Tính tương quan:
  - Biến numeric_cols chỉ định ba cột số: OrderValue, Quantity, và Price.
  - corr() tính ma trận tương quan Pearson giữa các cột, trả về một dataframe với các giá trị từ -1 (tương quan nghịch hoàn hảo) đến 1 (tương quan thuận hoàn hảo).
- Vẽ heatmap**:**
  - Hàm sns.heatmap tạo biểu đồ heatmap từ ma trận tương quan.
  - Tham số annot=True hiển thị giá trị tương quan trên mỗi ô.

- cmap='Blues' sử dụng bảng màu xanh lam để tăng tính thẩm mỹ.
- vmin=-1, vmax=1, center=0 thiết lập phạm vi giá trị từ -1 đến 1, với màu trung tính tại 0, giúp dễ dàng nhận diện các mức tương quan âm, dương, hoặc gần 0.



Hình 3-5: Biểu đồ Tương quan giữa các yếu tố Quantity, Price, OrderValue

Nhận xét:

- Hệ số tương quan:
  - Quantity & TotalAmount
    - Hệ số tương quan (r): ≈ 0.59 – 0.70
    - Tương quan dương vừa. Số lượng sản phẩm tăng → tổng tiền thường tăng.
  - UnitPrice & TotalAmount
    - Hệ số tương quan (r): ≈ 0.40 – 0.60
    - Tương quan dương: Do giá cao góp phần tăng tổng tiền, nhưng yếu hơn số lượng.
  - Quantity & UnitPrice
    - Hệ số tương quan (r): ≈ -0.05 – -0.20
    - Tương quan âm yếu. Có thể do giá cao thì người ta mua ít (ngược lại), nhưng mối quan hệ này không mạnh.
- Nhận định:
  - TotalAmount phụ thuộc mạnh hơn vào Quantity (số lượng) hơn là Price (giá đơn vị).
  - Mối quan hệ giữa giá và số lượng không rõ ràng cho thấy có thể do sự đa dạng sản phẩm và hành vi người mua khác nhau.

3.5. Tính toán đặc trưng RFM

3.5.1. Giới thiệu về RFM

Phân tích RFM (Recency - Frequency - Monetary) là một phương pháp phân đoạn khách hàng dựa trên hành vi mua sắm, giúp doanh nghiệp hiểu rõ hơn về giá trị và tiềm năng của từng nhóm khách hàng. Phương pháp này được sử dụng rộng rãi trong quản lý quan hệ khách hàng (CRM) và marketing để tối ưu hóa chiến lược tiếp cận và tăng doanh thu.

- Recency (Thời gian gần nhất): Khoảng thời gian kể từ lần mua hàng gần nhất của khách hàng. Khách hàng mua hàng gần đây thường có khả năng quay lại cao hơn.
- Frequency (Tần suất): Số lần mua hàng của khách hàng trong một khoảng thời gian nhất định. Khách hàng mua sắm thường xuyên thường có mức độ trung thành cao.
- Monetary (Giá trị tiền tệ): Tổng số tiền khách hàng đã chi tiêu. Khách hàng chi tiêu nhiều thường mang lại giá trị lớn cho doanh nghiệp.

Mục tiêu của báo cáo này là trình bày cách áp dụng phân tích RFM để phân tích hành vi khách hàng, từ đó đưa ra các khuyến nghị chiến lược nhằm cải thiện hiệu quả kinh doanh.

3.5.2. Cách thực hiện

Cách thực hiện phân tích RFM (Recency - Frequency - Monetary):

- Recency:
  - Công thức: Số ngày từ lần mua cuối (InvoiceDate lớn nhất) của mỗi khách hàng đến ngày tham chiếu (thường là ngày cuối trong dữ liệu + 1 ngày).
  - Cách tính: Nhóm theo Customer ID, lấy max(InvoiceDate), trừ cho ngày tham chiếu.
- Frequency:
  - Công thức: Số hóa đơn duy nhất (Invoice) của mỗi khách hàng.
  - Cách tính: Nhóm theo Customer ID, đếm số Invoice duy nhất (unique).
- Monetary:
  - Công thức: Tổng của cột OderValue
  - Cách tính: thêm cột OderValue , nhóm theo Customer ID.

3.5.3. Thực nghiệm

3.5.3.1. Tính chỉ số RFM cho từng khách hàng

Tính toán các chỉ số RFM (Recency - Frequency - Monetary) nhằm phân tích hành vi khách hàng dựa trên dữ liệu giao dịch. Recency cho biết khách hàng mua hàng gần đây hay đã lâu chưa quay lại, từ đó đánh giá nguy cơ rời bỏ. Frequency phản ánh mức độ trung thành thông qua số lần mua hàng. Monetary thể hiện tiềm năng doanh thu dựa trên tổng chi tiêu của khách.

Thông qua RFM, doanh nghiệp có thể phân nhóm khách hàng, ưu tiên chăm sóc và thiết kế các chiến dịch marketing hiệu quả hơn.

Đầu tiên ta xác định ngày tham chiếu current-daten. Việc chọn ngày tham chiếu là ngày cuối + 1 đảm bảo Recency luôn là số dương và phản ánh chính xác khoảng thời gian kể từ lần mua cuối.

- df['InvoiceDate'].max() lấy ngày giao dịch lớn nhất (tức ngày gần nhất) trong tập dữ liệu.
- pd.Timedelta(days=1) thêm 1 ngày để tạo ngày tham chiếu (current_date), thường được dùng làm mốc để tính Recency.

Tiếp theo, tiến hành tính toán RFM:

- Nhóm dữ liệu: groupby('Customer ID') nhóm dữ liệu theo mỗi khách hàng để tính toán RFM.
- Recency: lambda x: (current_date - x.max()).days tính số ngày từ lần mua cuối (max(InvoiceDate)) đến current_date.
- Frequency: 'Invoice': 'nunique' đếm số hóa đơn duy nhất (Invoice) cho mỗi khách hàng, phản ánh tần suất mua sắm.
- Monetary: 'OrderValue': 'sum' tính tổng giá trị chi tiêu (OrderValue) của mỗi khách hàng.
- Đổi tên cột: rename đổi tên cột thành Recency, Frequency, Monetary cho rõ ràng.
- Đặt lại chỉ số: reset_index() chuyển Customer ID từ chỉ số thành cột thông thường.
- Kết quả: DataFrame rfm chứa các cột: Customer ID, Recency, Frequency, Monetary.

3.5.3.2. Tính điểm RFM

Hàm chấm điểm RFM phân loại khách hàng thành các nhóm khác nhau dựa trên điểm số R, F, M. Mỗi yếu tố được tính điểm từ 1 đến 5, với 1 điểm cho giá trị thấp nhất và 5 điểm cho giá trị cao nhất, cho phép doanh nghiệp dễ dàng xác định nhóm khách hàng nào có giá trị cao, nhóm nào cần được chăm sóc nhiều hơn, và nhóm nào có thể không còn quan tâm đến sản phẩm.

Chuyển các giá trị RFM thành điểm rời rạc(1-5) để phân khúc khách hàng. Chia dữ liệu thành:

- Recency:Gía trị thấp(mua gần đây) -> điểm cao(5),giá trị thấp(mua cách đây lâu) -> điểm thấp (1)
- Frequency: Gía trị cao(mua thường xuyên) -> điểm cao (5),ngược lại
- Monetary: chi tiêu nhiều -> điểm cao(5), chi tiêu thấp (1)

Hàm tính điểm Recency: Hàm recency_score gán điểm từ 1 đến 5 dựa trên giá trị recency (số ngày từ lần mua cuối).

- ≤ 20 ngày: 5 điểm (mua rất gần đây, khách hàng tích cực).
- ≤ 50 ngày: 4 điểm.
- ≤ 100 ngày: 3 điểm.
- ≤ 200 ngày: 2 điểm.

- 200 ngày: 1 điểm (mua từ lâu, khách hàng ít hoạt động).

Hàm tính điểm Frequency: Hàm frequency_score gán điểm từ 1 đến 5 dựa trên giá trị frequency (số hóa đơn duy nhất).

- 6 hóa đơn: 5 điểm (mua rất thường xuyên, khách hàng trung thành).
- ≥ 4 hóa đơn: 4 điểm.
- 3 hóa đơn: 3 điểm.
- 2 hóa đơn: 2 điểm.
- 1 hóa đơn: 1 điểm (mua rất ít, khách hàng không thường xuyên).

Hàm tính điểm Monetary: Hàm monetary_score gán điểm từ 1 đến 5 dựa trên giá trị monetary (tổng chi tiêu).

- 1000 (đơn vị tiền tệ): 5 điểm (chi tiêu rất cao, khách hàng giá trị cao).
- ≥ 500: 4 điểm.
- ≥ 250: 3 điểm.
- ≥ 100: 2 điểm.
- < 100: 1 điểm (chi tiêu thấp, khách hàng giá trị thấp).

Gán điểm cho từng khách hàng: Sử dụng .apply() để áp dụng từng hàm trên cho từng dữ liệu. Mỗi khách hàng sẽ có điểm riêng cho từng yếu tố: R, F, M.

- R_Score: Điểm cho cột Recency (dựa trên recency_score).
- F_Score: Điểm cho cột Frequency (dựa trên frequency_score).
- M_Score: Điểm cho cột Monetary (dựa trên monetary_score).

Tính điểm RFM tổng: Kết hợp ba điểm số (R_Score, F_Score, M_Score) thành một chuỗi ký tự (RFM_Score). Các điểm số được chuyển thành chuỗi (astype(str)) và ghép lại, tạo thành mã RFM.

Bảng 3-6: Kết quả tính RFM

Customer ID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Score
12346	326	1	77183.60	1	1	5	115
12347	2	6	3598.21	5	4	5	545
12348	75	3	784.44	3	3	4	334
12349	19	1	1457.55	5	1	5	515
12350	310	1	294.40	1	1	3	113

3.5.3.3. Xử lý ngoại lai

Bảng 3-7: Thống kê RFM sau chuẩn hóa

Thống kê	Recency	Frequency	Monetary
Count	4214.000000	4214.000000	4214.000000
Mean	-0.000000	0.000000	-0.000000
Std	1.000119	1.000119	1.000119
Min	-0.930569	-0.940541	-4.012506
25%	-0.753165	-0.940541	-0.675211
50%	-0.409445	-0.329807	-0.050155
75%	0.477576	0.439627	0.656938
Max	2.828181	5.886711	4.779892

Nhận xét:

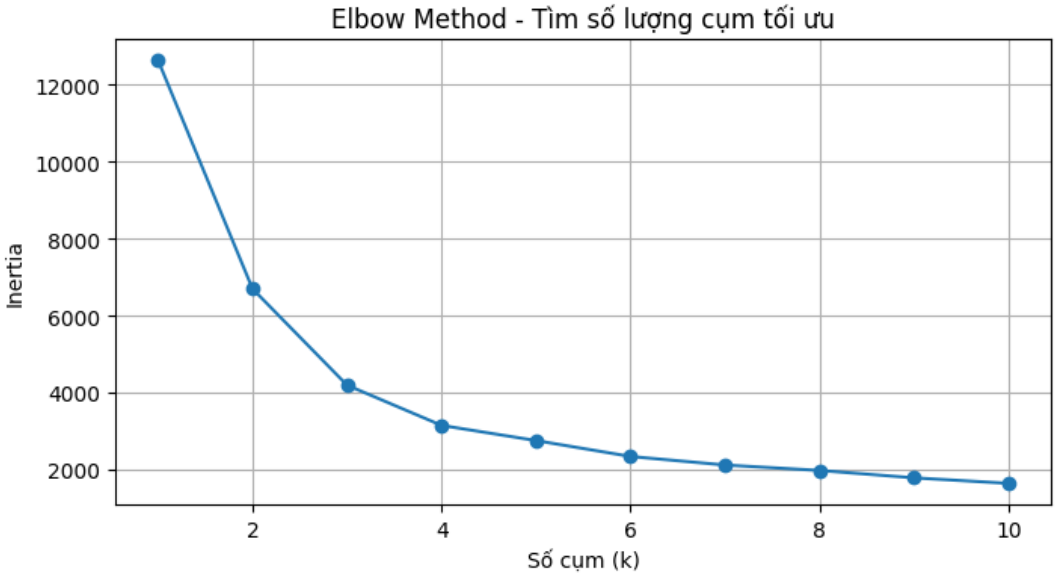
- Chuẩn hóa thành công:
  - Trung bình các cột ≈ 0, độ lệch chuẩn ≈ 1 → chuẩn hóa Z-score được thực hiện đúng.
- Phân phối lệch vẫn còn:
  - Frequency và Monetary có giá trị max lớn so với 75% → vẫn có khách hàng đặc biệt nhiều giao dịch hoặc chi tiêu cao → đây có thể là khách hàng VIP.
  - Monetary có giá trị min khá thấp (-4.01) → vẫn còn sự chênh lệch trong chi tiêu giữa các nhóm khách.
- Recency phân bố hợp lý hơn:
  - Dải giá trị không quá rộng (từ -0.93 đến 2.83).
  - Đây là tín hiệu tốt vì Recency thường ít bị skew hơn Monetary hay Frequency.

3.6. Xây dựng và Đánh giá mô hình

3.6.1. Sử dụng K-Means

Trong phương pháp Elbow:

- Chạy K-Means với nhiều giá trị *k* khác nhau và tính toán WCSS (Within-Cluster Sum of Squares) – tổng bình phương khoảng cách từ mỗi điểm đến tâm cụm của nó.
- Vẽ đồ thị biểu diễn WCSS theo từng giá trị *k*Khi số cụm tăng, WCSS sẽ giảm, nhưng sau một điểm nhất định, mức độ cải thiện sẽ giảm dần.
- Điểm gấp khúc (hay còn gọi là "khuỷu tay") trên đồ thị chính là nơi thích hợp để chọn *k*.



Hình 3-6: Biểu đồ Elbow Method - Tìm số lượng cụm tối ưu

- Giá trị Inertia (WCSS) giảm nhanh khi số cụm *k* tăng từ 1 đến 4.

- Sau k=3, đường cong bắt đầu "thoải" hơn, mức giảm không còn đáng kể.
- Do đó, điểm gấp khúc rõ ràng (elbow) nằm tại k=3. Đây là vị trí cân bằng giữa việc giảm WCSS và số lượng cụm, giúp phân cụm hiệu quả mà không gây dư thừa.

Silhouette Score là chỉ số đánh giá chất lượng phân cụm của thuật toán như K-Means, bằng cách đo lường:

- Mức độ gắn kết giữa một điểm và các điểm trong cùng cụm (gọi là a(i)),
- Mức độ tách biệt giữa điểm đó và các cụm khác (gọi là b(i)).

Công thức:

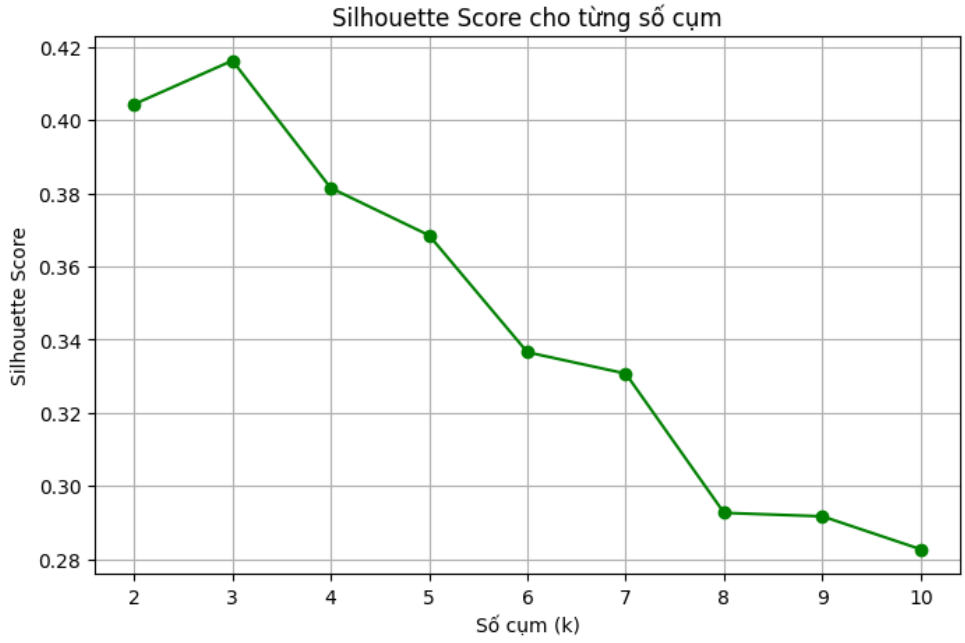
$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Giá trị Silhouette Score

- Gần 1 → điểm được phân cụm tốt.
- Gần 0 → điểm nằm ở ranh giới giữa các cụm.
- Nhỏ hơn 0 → điểm có thể bị phân vào sai cụm.

Áp dụng trong K-Means

- Sau khi chạy K-Means với số cụm k, tính Silhouette Score để đánh giá mức độ hiệu quả của việc phân cụm.
- So sánh Silhouette Score cho các giá trị k khác nhau → chọn k có điểm trung bình cao nhất → xác định số cụm tối ưu.



Hình 3-7: Biểu đồ Silhouette Score cho từng số cụm

Nhận xét:

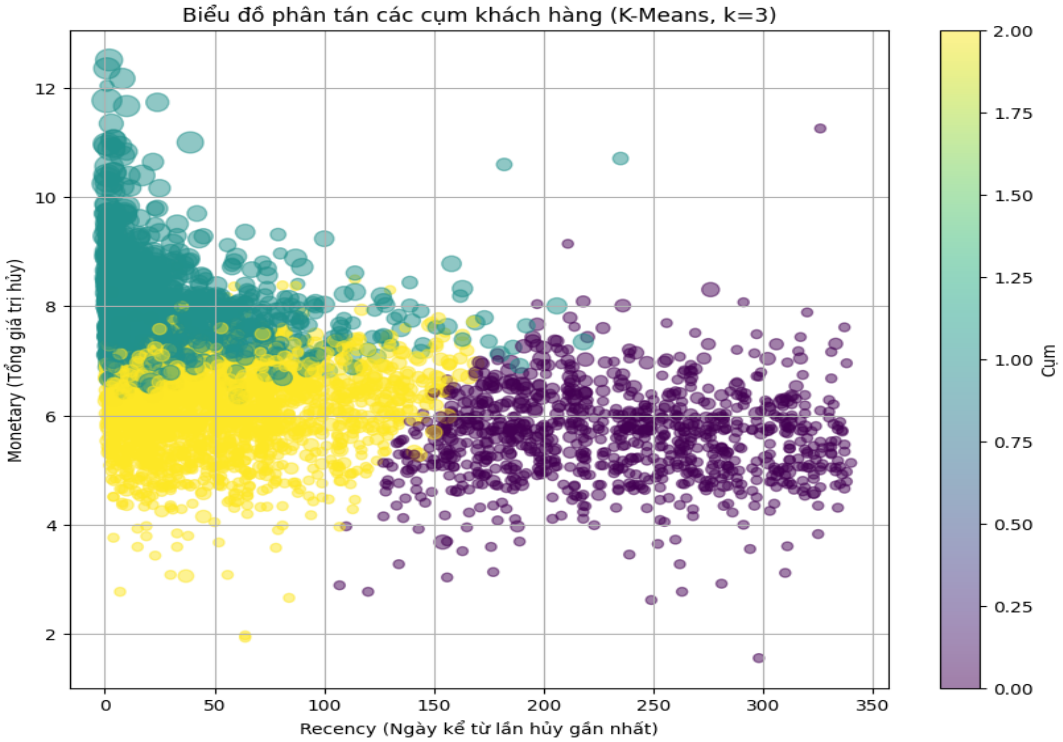
- Điểm Silhouette Score cao nhất tại k=3 và, cho thấy các điểm dữ liệu trong cùng một cụm có độ tương đồng cao và khác biệt rõ rệt so với các cụm khác.
- Kết hợp với chọn cụm tối ưu trong phương pháp Elbow thì cụm tối ưu được chọn k=3.

3.6.2. Huấn luyện mô hình trên dữ liệu đã chuẩn hóa

Silhouette Score: 0.42

Davies-Bouldin Index: 0.83

Calinski-Harabasz Index: 4230.78



Hình 3-8: Biểu đồ phân tán các cụm khách hàng

Nhận xét:

Có 3 cụm màu sắc (mỗi cụm đại diện một nhóm khách hàng):

Cụm 1 (màu xanh ngọc, Recency thấp, Monetary cao):

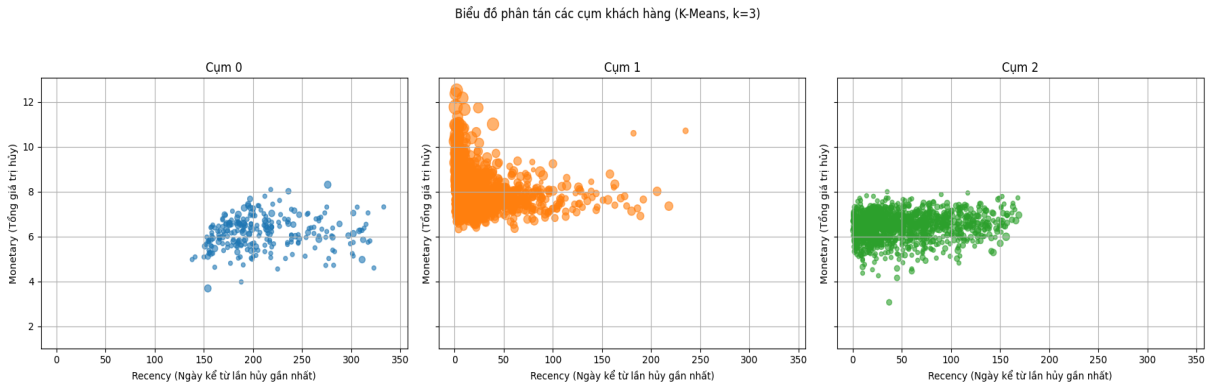
- Đây là nhóm khách hàng tốt nhất:
  - Giao dịch gần đây (Recency thấp),
  - Có giá trị giao dịch cao.
  - Có thể là nhóm trung thành, chi tiêu cao → nên giữ chân bằng các chương trình ưu đãi.

Cụm 2 (màu vàng, Recency vừa, Monetary trung bình):

- Nhóm khách hàng tầm trung:
  - Giao dịch không quá gần đây, giá trị chi tiêu ở mức trung bình.
  - Có thể đang có dấu hiệu rời bỏ, cần chăm sóc thêm để giữ lại.

Cụm 3 (màu tím, Recency cao, Monetary thấp):

- Nhóm khách hàng kém tiềm năng:
  - Đã lâu không giao dịch (Recency cao), chi tiêu thấp.
  - Có thể đã ngừng sử dụng dịch vụ → ít nên đầu tư nhiều vào nhóm này.



Hình 3-9: Biểu đồ phân tán các cụm khách hàng

Biểu đồ trình bày chi tiết 3 cụm khách hàng sau khi phân cụm bằng K-Means (k=3), theo hai chiều:

- Recency (R) – số ngày kể từ lần huỷ gần nhất.
- Monetary (M) – tổng giá trị huỷ (có thể là chi tiêu nếu đây là hệ thống thương mại).

Kích thước điểm có thể biểu diễn cho Frequency – tần suất giao dịch/hủy.

Cụm 0 (bên trái):

- Recency cao (dao động quanh 150–300 ngày): khách hàng đã lâu không quay lại.
- Monetary trung bình (4–8): từng có mức chi tiêu/hủy trung bình.
- Frequency thấp đến trung bình (bong bóng vừa).
- Nhóm này có thể là khách hàng đã bỏ đi hoặc rất lâu không tương tác, ít tiềm năng.

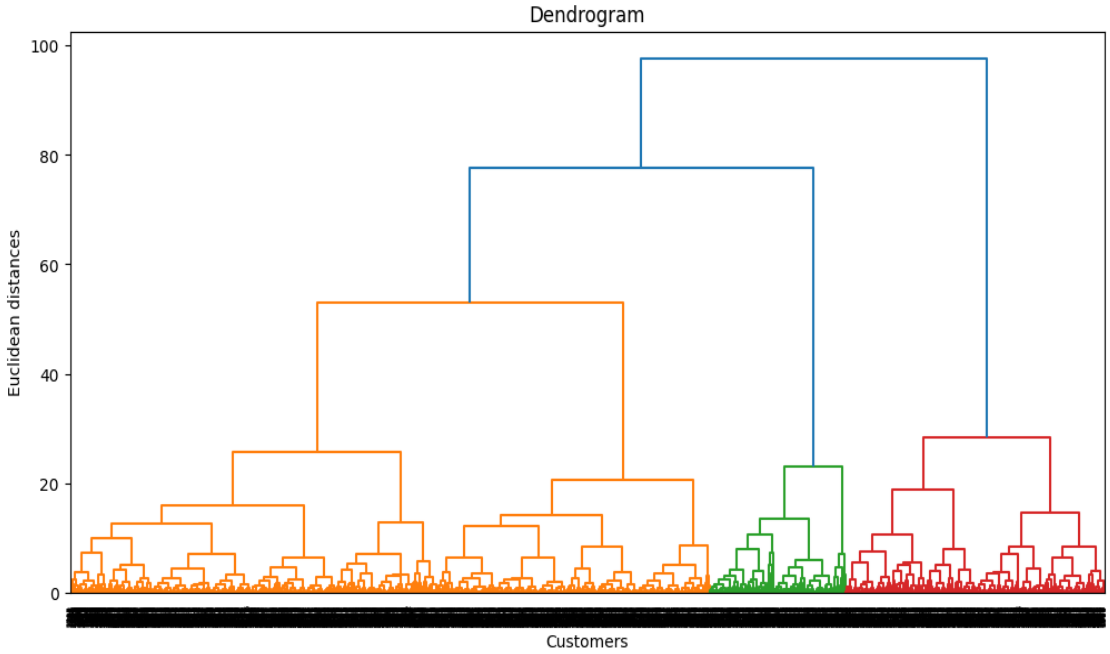
Cụm 1 (ở giữa):

- Recency thấp (gần đây): nhiều điểm tập trung ở 0–50 ngày.
- Monetary trải dài từ thấp đến rất cao (4–12).
- Frequency khá cao (nhiều bong bóng lớn).
- Đây là nhóm khách hàng giá trị cao và trung thành: mới giao dịch gần đây, tiêu nhiều, có thể là VIP. Nên chăm sóc kỹ, giữ chân bằng khuyến mãi hoặc chương trình thân thiết.

Cụm 2 (bên phải):

- Recency trung bình đến thấp (nằm trong khoảng 0–150): nghĩa là khá nhiều khách quay lại gần đây.
- Monetary thấp đến trung bình (dao động quanh 5–7).
- Frequency nhỏ hơn cụm 1 (bong bóng nhỏ hơn).
- Nhóm này là khách hàng phổ thông: thường xuyên quay lại nhưng chi tiêu không quá cao. Nên thúc đẩy họ mua nhiều hơn thông qua upsell hoặc bundle sản phẩm.

3.6.3. Vẽ dendrogram để xác định số cụm



Hình 3-10: Biểu đồ Dendrogram xác định số cụm

Một trong những mục đích chính của dendrogram là giúp chọn số cụm phù hợp (số lượng nhóm khách hàng hợp lý nhất):

Cách chọn số cụm bằng phương pháp "ngưỡng cắt" (threshold):

- Tìm đoạn dài nhất theo chiều dọc (khoảng cách lớn) mà không bị cắt ngang bởi đường ngang nào khác.
- Vẽ một đường ngang cắt qua vùng đó → đếm số đoạn giao nhau với các nhánh chính ⇒ số cụm hợp lý.

Trong biểu đồ:

- Nếu vẽ một đường ngang tại khoảng khoảng cách ~75, sẽ cắt qua 3 nhánh chính.
- Điều này xác nhận rằng việc chọn k = 3 cụm trong K-Means là hợp lý, vì dendrogram cũng chỉ ra điều tương tự.

3.7. Phân tích hành vi của các cụm khách hàng

Phần này tập trung phân tích hành vi của các cụm khách hàng dựa trên mô hình RFM (Recency, Frequency, Monetary) và giá trị vòng đời khách hàng (CLV). Mục tiêu là xác định đặc điểm, giá trị và tiềm năng của từng cụm để hỗ trợ xây dựng chiến lược kinh doanh, từ chăm sóc khách hàng trung thành đến kích hoạt lại các khách hàng không hoạt động.

3.7.1. Phân tích đặc điểm từng cụm

Dựa trên kết quả phân cụm, các chỉ số RFM trung bình, tối thiểu, tối đa và số lượng khách hàng trong mỗi cụm được trình bày trong bảng sau:

Bảng 3-8: Tóm tắt đặc điểm các cụm

Metric	Cluster 0	Cluster 1	Cluster 2
Recency_Mean	233.71	28.15	51.56
Recency_Min	107	1	1
Recency_Max	340	235	169
Frequency_Mean	0.82	2.12	1.03
Frequency_Min	0.69	1.10	0.69
Frequency_Max	1.95	5.23	2.08
Monetary_Mean	5.60	7.93	6.12
Monetary_Min	1.56	6.33	1.93
Monetary_Max	11.25	12.51	8.73
Count	937	1284	1993

Nhận xét:

- Cụm 0 có Recency rất cao (trung bình ~233 ngày, nghĩa là phần lớn khách hàng đã lâu chưa mua), tần suất mua rất thấp (0,82 lần) và giá trị chi tiêu trung bình thấp nhất (5,60).
    - Nhóm này gồm những khách hàng hiếm khi mua hàng và có thể đã mất dần sự quan tâm đến thương hiệu.Đây có thể xem là nhóm "Ngủ đông": họ “hiếm khi mua hàng, tần suất thấp và giá trị chi tiêu không đáng kể”
    - Nhóm này thường có nguy cơ chuyển sang không còn mua nữa nếu không được chú ý kịp thời.
- Cụm 1 có Recency rất thấp (trung bình 28 ngày), tần suất cao (2,12 lần) và giá trị chi tiêu lớn nhất (7,93). Điều này cho thấy nhóm này thường xuyên mua sắm và chi tiêu nhiều.
    - Có thể xem là nhóm “Khách hàng trung thành VIP”: họ mua hàng đều đặn, chi tiêu cao và có giao dịch gần đây nhất
    - Nhóm này là tài sản quan trọng của doanh nghiệp, cần được duy trì và nuôi dưỡng lâu dài.
- Cụm 2 có Recency trung bình khoảng 51 ngày, tần suất mua ~1,03 lần và giá trị chi tiêu trung bình 6,12. Nhóm này mua sắm ở mức trung bình – không quá thường xuyên nhưng cũng không quá lâu mới mua (có thể gồm khách mới hoặc khách mua lẻ tẻ).
    - Được phân loại là nhóm “Khách hàng có tiềm năng trung thành": họ mới mua gần đây nhưng tần suất chưa cao, chi tiêu ở mức trung bình và có khả năng tăng thêm trong tương lai
    - Nói cách khác, đây là khách hàng cần được nuôi dưỡng và khuyến khích tiếp tục mua để trở thành khách trung thành.

3.7.2. Số lượng khách hàng của từng cụm

Số lượng khách hàng trong mỗi cụm được thống kê dựa trên các chỉ số RFM trung bình, như sau:

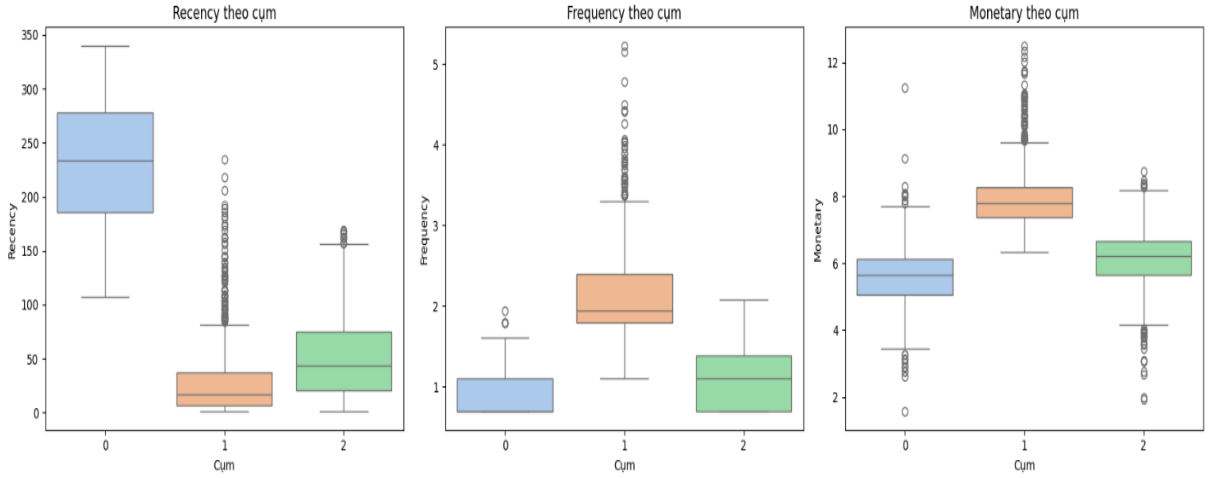
Bảng 3-9: Số lượng khách hàng trong mỗi cụm

Chỉ số	Cluster 0	Cluster 1	Cluster 2
Recency	233.71	28.15	51.56
Frequency	0.82	2.12	1.03
Monetary	5.60	7.93	6.12
Số lượng khách hàng	937	1284	1993

Nhận xét:

- Cụm 2 có số lượng khách hàng lớn nhất (1.993), chiếm phần lớn trong bộ dữ liệu. Điều này cho thấy đa số khách hàng thuộc nhóm tiềm năng, cần được đầu tư để tăng tần suất và giá trị mua sắm.
- Cụm 1 có 1.284 khách hàng, tuy chiếm số lượng ít hơn cụm 2 nhưng là nhóm giá trị cao nhất, cần được ưu tiên chăm sóc.
- Cụm 0 có số lượng ít nhất (937), cho thấy nhóm khách hàng không hoạt động chiếm tỷ lệ nhỏ, nhưng vẫn cần chú ý để kích hoạt lại hoặc đánh giá khả năng loại bỏ.

3.7.3. Biểu diễn biểu đồ trên các cụm



Hình 3.11. Biểu đồ biểu diễn trên các cụm

Nhận xét:

- Về Recency (số ngày kể từ lần giao dịch gần nhất), cụm 1 có giá trị thấp nhất, cho thấy các khách hàng trong nhóm này vừa mới tương tác gần đây – đây là những khách hàng rất tiềm năng và vẫn đang quan tâm tới sản phẩm hoặc dịch vụ. Cụm 2 có giá trị Recency trung bình, nghĩa là những khách hàng này đã từng tương tác, nhưng lần cuối cùng đã cách đây một thời gian tương đối. Trong khi đó, cụm 0 có Recency cao nhất – tức là nhóm này bao gồm những khách hàng đã rất lâu không còn giao dịch – khả năng cao là đã rời bỏ.
- Xét theo Frequency (tần suất mua hàng), cụm 1 tiếp tục nổi bật với số lần mua nhiều nhất, cho thấy đây là nhóm khách hàng trung thành và thường xuyên quay lại. Cụm 2 có số lần mua ở mức trung bình, còn cụm 0 lại một lần nữa thể

hiện đặc điểm tiêu cực khi tần suất mua hàng rất thấp, phần lớn chỉ mua một lần.

- Đối với Monetary (tổng giá trị giao dịch), cụm 1 tiếp tục là nhóm nổi bật nhất với mức chi tiêu cao nhất. Đây là những khách hàng không chỉ thường xuyên mua hàng mà còn có giá trị lớn – rất đáng được chăm sóc. Cụm 2 có giá trị Monetary ở mức trung bình, còn cụm 0 có chi tiêu không quá cao, phù hợp với xu hướng ít mua và lâu không quay lại.

3.7.4. Tính CLV của mỗi cụm

Khái niệm CLV:

- CLV (Customer Lifetime Value) là tổng lợi nhuận mà một khách hàng mang lại cho doanh nghiệp trong suốt thời gian họ tương tác với thương hiệu.
- Vai trò:
    - Xác định nhóm khách hàng cần ưu tiên chăm sóc (VIP, tiềm năng, ngủ đông).
    - Hỗ trợ phân bổ ngân sách marketing, xây dựng chiến lược giữ chân khách hàng.
    - Đánh giá giá trị dài hạn của từng nhóm khách hàng.

Công thức tính CLV:

- Công thức tổng quát:
  CLV = Giá trị trung bình mỗi lần mua × Tần suất mua hàng × Thời gian trung bình khách hàng gắn bó
- Trong trường hợp dữ liệu đã chuẩn hóa:
  CLV = Frequency × Monetary

Hình 3-11: Giá trị khách hàng (CLV) trung bình theo cụm

Kết quả CLV của từng cụm:

- Cụm 1 có CLV trung bình cao nhất, vượt xa các cụm còn lại, cho thấy đây là nhóm khách hàng có giá trị lớn nhất và tiềm năng sinh lợi cao, dù số lượng giao dịch có thể không nhiều nhưng giá trị mỗi lần giao dịch rất lớn và đều đặn.
- Cụm 2 có CLV trung bình ở mức khá, phản ánh nhóm này có tiềm năng tăng trưởng, đặc biệt vào các giai đoạn cao điểm trong năm – nếu được thúc đẩy đúng cách, nhóm này có thể trở thành nguồn sinh lợi quan trọng.
- Cụm 0 là nhóm có CLV trung bình thấp nhất, có thể là những khách hàng chỉ mua sắm nhỏ lẻ, không thường xuyên hoặc có hành vi giao dịch thiếu ổn định – cần được cân nhắc kỹ trong chiến lược duy trì hoặc loại bỏ.

Nhận xét:

- Cụm 1 là nhóm mang lại giá trị cao nhất, cần được ưu tiên chăm sóc để duy trì lợi nhuận dài hạn.
- Cụm 2 có tiềm năng tăng trưởng, đặc biệt nếu doanh nghiệp triển khai các chiến lược khuyến khích mua sắm thường xuyên hơn.
- Cụm 0 có giá trị thấp, cần cân nhắc chi phí để kích hoạt lại hoặc chấp nhận khả năng mất khách.

Bảng 3-10: Phân loại khách hàng theo cụm

Customer ID	Cluster	Nhóm khách hàng
12346	0	Khách tiềm năng
12347	1	Trung thành
12348	2	Ngủ quên
12349	2	Ngủ quên
12350	0	Khách tiềm năng

3.7.5. Phân tích thời gian mua sắm

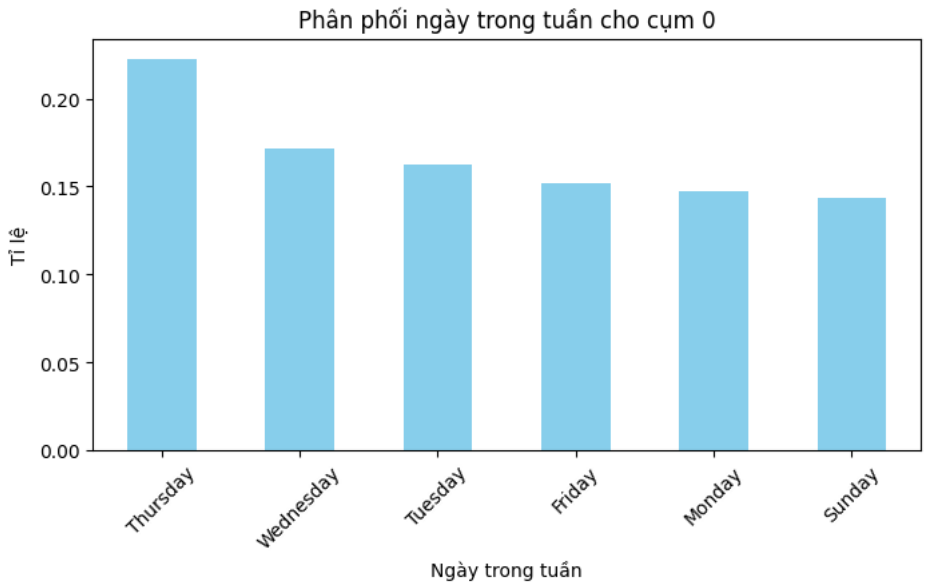
Phân tích thời gian mua sắm giúp xác định các xu hướng giao dịch theo ngày trong tuần, giờ trong ngày và tháng trong năm của từng cụm khách hàng. Thông tin này hỗ trợ tối ưu hóa các chiến dịch tiếp thị và chăm sóc khách hàng theo thời điểm khách hàng hoạt động mạnh nhất.

3.7.5.1. Cụm 0: Khách hàng ngủ đông

a) Phân bố ngày trong tuần:

Bảng 3-11: Phân bố ngày trong tuần của cụm 0

DayOfWeek	Tỷ lệ
Thursday	0.222798
Wednesday	0.171892
Tuesday	0.162327
Friday	0.151876
Monday	0.147536
Sunday	0.143570



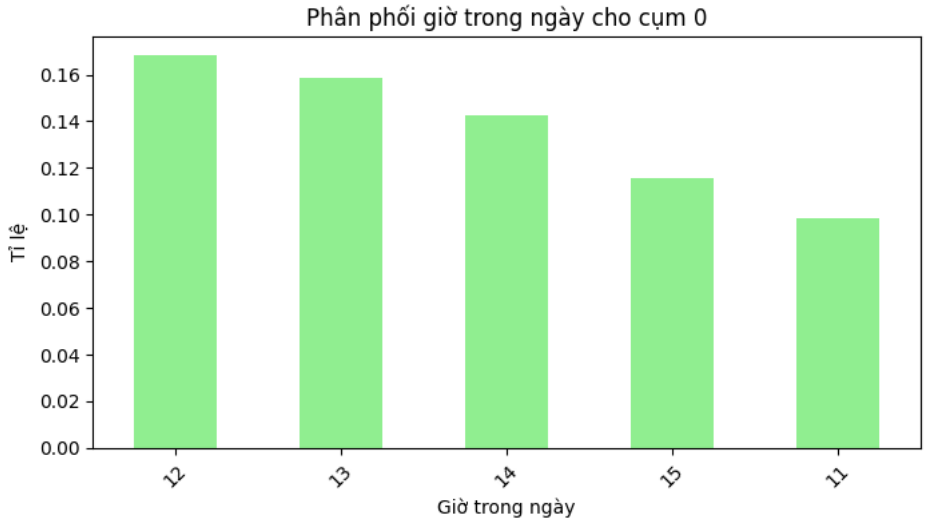
Hình 3-12: Phân phối ngày trong tuần cho cụm 0

Ngày trong tuần: Giao dịch tập trung vào giữa tuần, với Thứ Năm chiếm tỷ lệ cao nhất (22,28%), tiếp theo là Thứ Tư (17,19%) và Thứ Ba (16,23%). Các ngày còn lại (Thứ Sáu, Thứ Hai, Chủ Nhật) có tỷ lệ thấp hơn, dao động từ 14,36% đến 15,19%.

b) Phân bố giờ trong ngày

Bảng 3-12: Phân bố giờ trong tuần của cụm 0 (top5)

Giờ	Tỷ lệ
12	0.168113
13	0.158408
14	0.142591
15	0.115435
11	0.098171



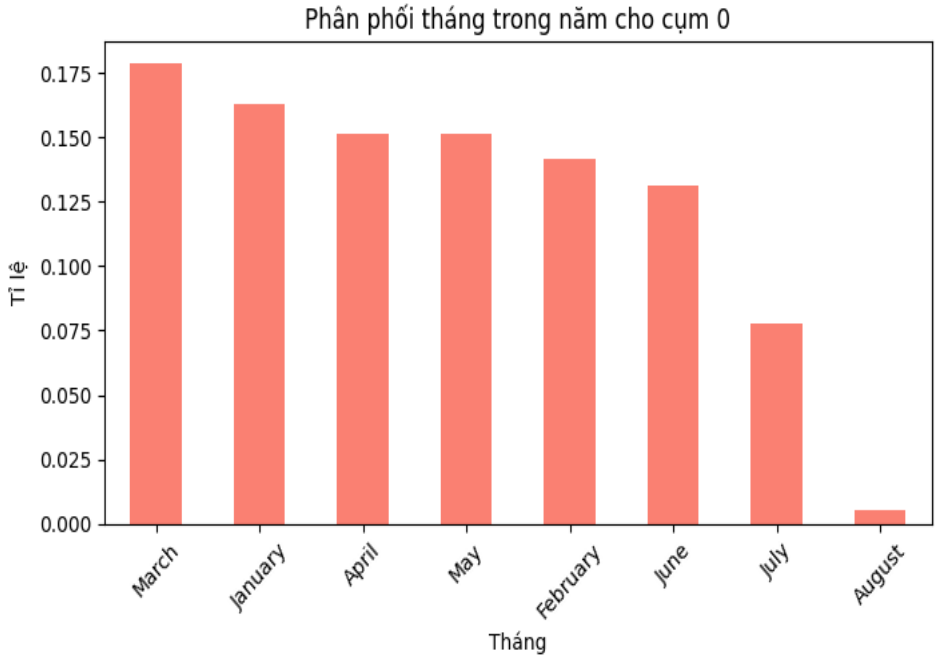
Hình 3-13: Phân phối giờ trong ngày cho cụm 0

Giờ trong ngày: Giao dịch chủ yếu diễn ra từ 12h-15h, với đỉnh điểm vào 12h (16,81%), tiếp theo là 13h (15,84%) và 14h (14,26%). Điều này cho thấy khách hàng trong cụm này có xu hướng mua sắm vào giờ nghỉ trưa hoặc đầu giờ chiều.

c) Phân bố tháng trong năm

Bảng 3-13: Phân bố tháng trong năm của cụm 0

Tháng	Tỷ lệ
March	0.178611
January	0.162747
April	0.151549
May	0.151129
February	0.141564
June	0.131299
July	0.078014
August	0.005086



Hình 3-14: Phân phối tháng trong năm cho cụm 0

Tháng trong năm: Giao dịch tập trung mạnh vào đầu năm (Tháng 1-5), với Tháng 3 chiếm tỷ lệ cao nhất (17,86%), tiếp theo là Tháng 1 (16,27%) và Tháng 4-5 (khoảng 15%). Hoạt động giảm mạnh từ Tháng 6 trở đi, đặc biệt gần như không có giao dịch vào Tháng 8 (0,51%).

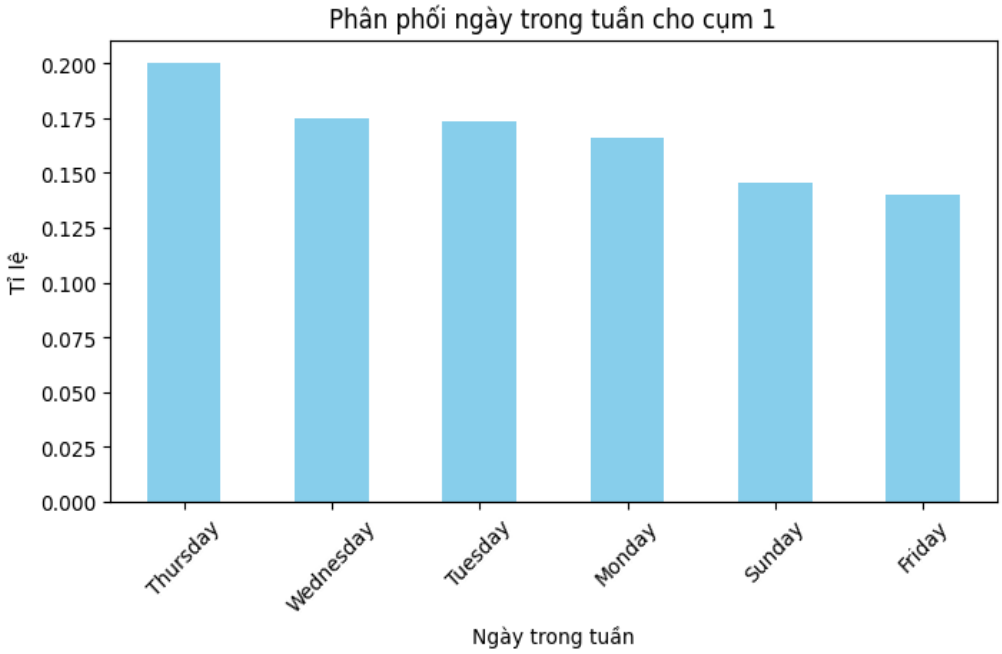
Kết luận: Cụm 0 có hành vi giao dịch không đều, với hoạt động mua sắm tập trung vào giữa tuần (đặc biệt Thứ Năm), khung giờ 12h-15h và giai đoạn đầu năm (đỉnh điểm Tháng 3). Sự sụt giảm mạnh vào giữa và cuối năm (đặc biệt Tháng 8) phản ánh đặc điểm "ngủ đông" của nhóm này.

3.7.5.2. Phân tích thời gian mua sắm cho cụm 1

a) Phân bố ngày trong tuần

Bảng 3-14: Phân bố tháng trong năm của cụm 1

DayOfWeek	Tỷ lệ
Thursday	0.200404
Wednesday	0.174960
Tuesday	0.173576
Monday	0.165768
Sunday	0.145524
Friday	0.139768



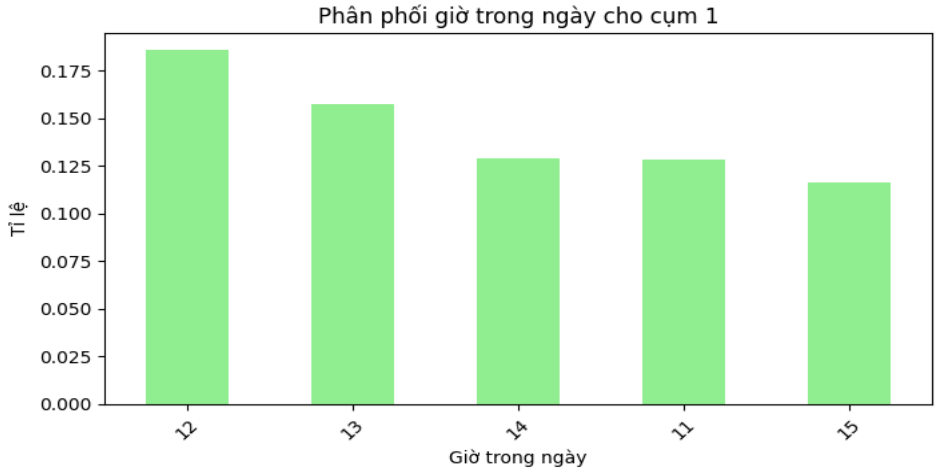
Hình 3-15: Phân phối ngày trong tuần cho cụm 1

Ngày trong tuần: Giao dịch phân bố đều hơn so với cụm 0, với Thứ Năm dẫn đầu (20,04%), tiếp theo là Thứ Tư (17,50%) và Thứ Ba (17,36%). Chủ Nhật và Thứ Sáu có tỷ lệ thấp hơn (14,55% và 13,98%).

b) Phân bố giờ trong ngày

Bảng 3-15: Phân bố giờ trong ngày của cụm 1(top 5)

Giờ	Tỷ lệ
12	0.185506
13	0.157511
14	0.128719
11	0.128105
15	0.116186



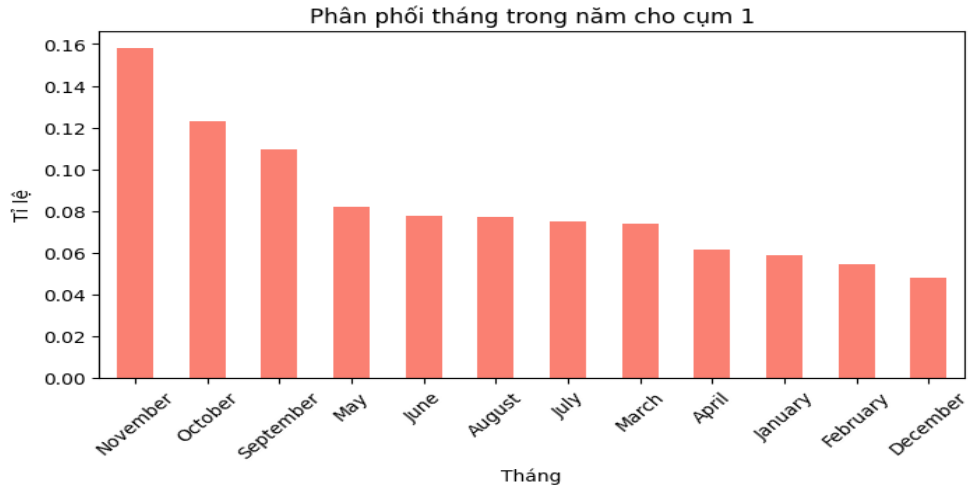
Hình 3-16: Phân phối giờ trong ngày cho cụm 1

Giờ trong ngày: Giao dịch tập trung từ 11h-15h, với đỉnh điểm vào 12h (18,55%), tiếp theo là 13h (15,75%) và 14h (12,87%). Điều này tương tự cụm 0, cho thấy thói quen mua sắm vào giờ nghỉ trưa.

c) Phân bố tháng trong năm:

Bảng 3-16: Phân bố tháng trong năm của cụm 1

Tháng	Tỷ lệ
November	0.158251
October	0.123223
September	0.109661
May	0.081772
June	0.077859
August	0.077444
July	0.074961
March	0.074095
April	0.061593
January	0.058831
February	0.054441
December	0.047870



Hình 3-17: Phân phối tháng trong năm cho cụm 1

Tháng trong năm: Giao dịch tăng mạnh vào cuối năm, với Tháng 11 chiếm tỷ lệ cao nhất (15,83%), tiếp theo là Tháng 10 (12,32%) và Tháng 9 (10,97%). Hoạt động giảm mạnh vào Tháng 12 (4,79%), có thể do kỳ nghỉ lễ cuối năm.

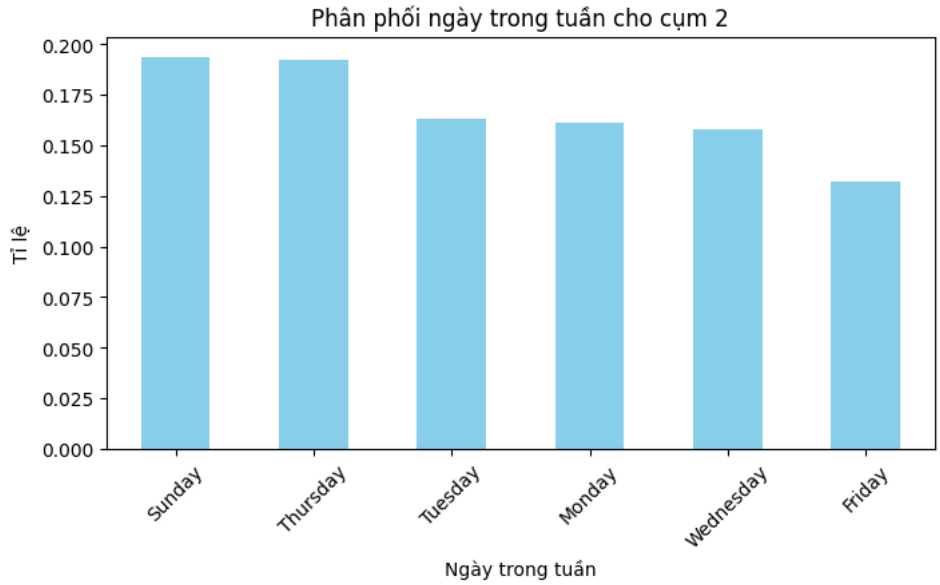
Kết luận: Cụm 1 có hành vi giao dịch đều đặn quanh năm, với xu hướng tăng mạnh vào cuối năm (đặc biệt Tháng 11). Giao dịch tập trung vào giữa tuần (Thứ Năm) và khung giờ 12h-15h, phản ánh thói quen mua sắm ổn định và chu kỳ rõ ràng của nhóm VIP trung thành.

3.7.5.3. Phân tích thời gian mua sắm cho cụm 2

a) Phân bố ngày trong tuần

Bảng 3-17: Phân bố ngày trong tuần của cụm 2

DayOfWeek	Tỷ lệ
Sunday	0.193750
Thursday	0.192088
Tuesday	0.162882
Monday	0.161185
Wednesday	0.157827
Friday	0.132267

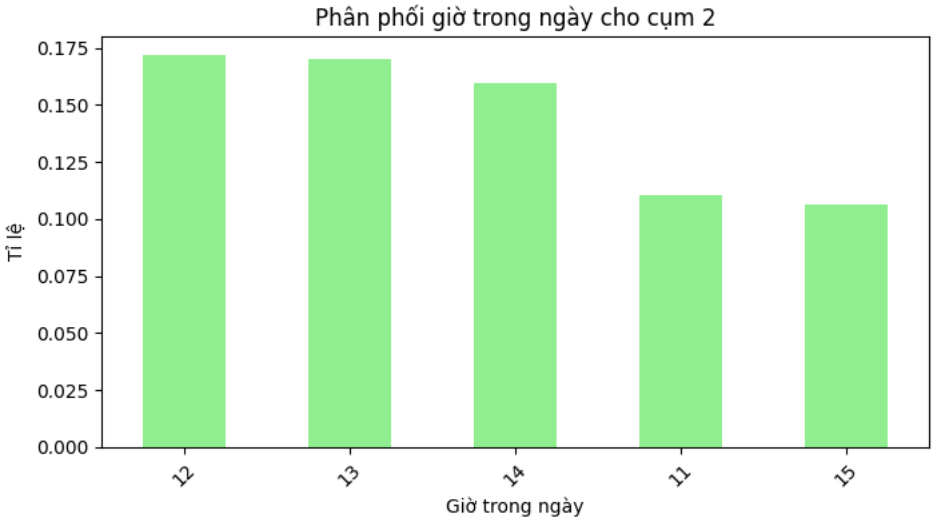


Hình 3-18: Phân phối ngày trong tuần cho cụm 2

b) Phân bố giờ trong ngày

Bảng 3-18: Phân bố giờ trong ngày của cụm 2 (top 5)

Giờ	Tỷ lệ
12	0.171594
13	0.170302
14	0.159616
11	0.110526
15	0.106441



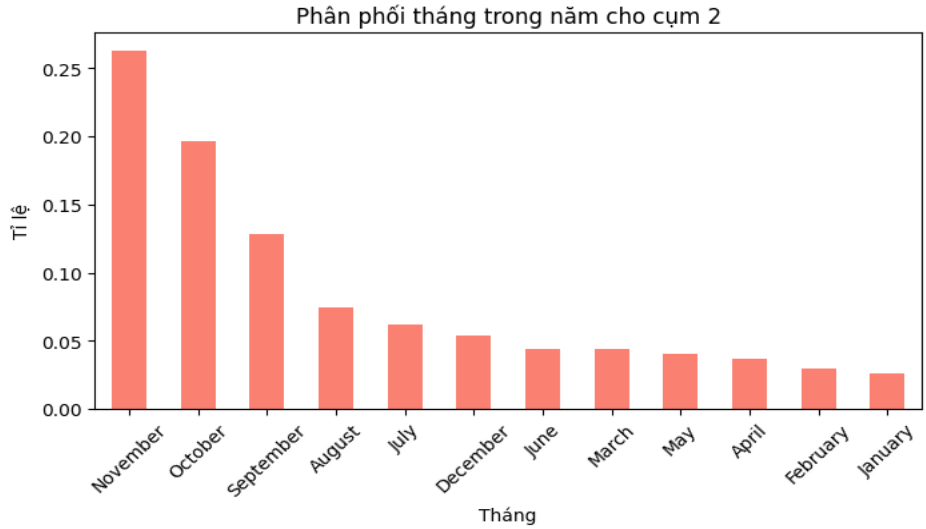
Hình 3-19: Phân phối giờ trong ngày cho cụm 2

Giờ trong ngày: Giao dịch tập trung từ 11h-15h, với đỉnh điểm vào 12h (17,16%) và 13h (17,03%), tương tự các cụm khác, nhưng có sự phân bố đều hơn giữa các giờ.

c) Phân bố tháng trong năm

Bảng 3-19: Phân bố tháng trong năm của cụm 2

Tháng	Tỷ lệ
November	0.263207
October	0.196854
September	0.128286
August	0.074846
July	0.062233
December	0.054132
June	0.044439
March	0.043851
May	0.040712
April	0.036431
February	0.029345
January	0.025664



Hình 3-20: Phân phối tháng trong năm cho cụm 2

Tháng trong năm: Giao dịch tăng mạnh vào cuối năm, với Tháng 11 chiếm tỷ lệ vượt trội (26,32%), tiếp theo là Tháng 10 (19,69%) và Tháng 9 (12,83%). Hoạt động giảm mạnh vào đầu năm (Tháng 1-5), đặc biệt Tháng 1 chỉ chiếm 2,57%.

Kết luận: Cụm 2 có hành vi giao dịch mạnh vào cuối tuần (Chủ Nhật) và giữa tuần (Thứ Năm), với khung giờ cao điểm 12h-13h. Giao dịch tăng đột biến vào cuối năm (đặc biệt Tháng 11), phản ánh xu hướng "bùng nổ theo mùa" của nhóm tiềm năng trung thành.

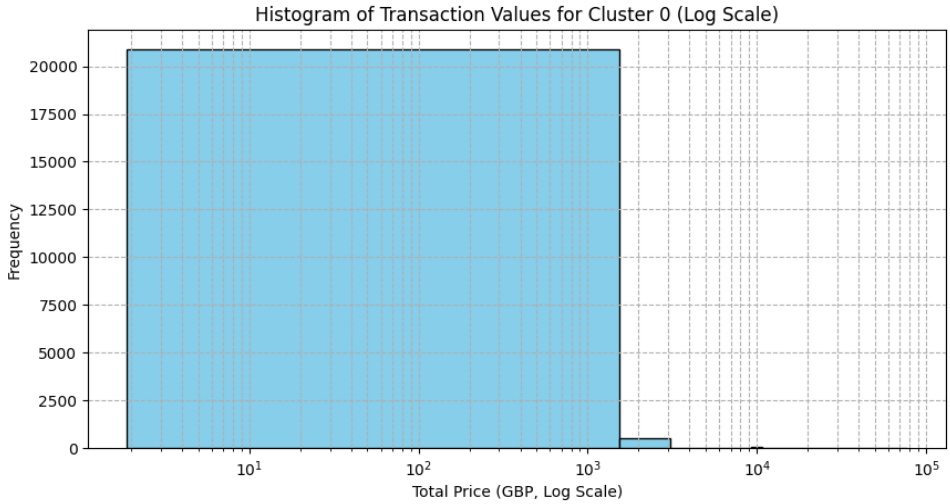
3.7.6. Tính giá trị giao dịch theo hóa đơn

Phân tích giá trị giao dịch theo hóa đơn giúp đánh giá mức chi tiêu của từng cụm khách hàng, bao gồm giá trị trung bình, trung vị, tỷ lệ giao dịch lớn (trên phân vị 75%) và phân bố giá trị giao dịch thông qua histogram log scale.

3.7.6.1 Phân tích giá trị giao dịch cho cụm 0

Phân tích:

- Giá trị trung bình mỗi giao dịch: 443.58 GBP
- Giá trị trung vị mỗi giao dịch: 308.58 GBP
- Tỷ lệ giao dịch lớn (trên phân vị 75%): 24.93%
- Ngưỡng giao dịch lớn (phân vị 75%): 510.73 GBP



Hình 3-21: Biểu đồ Histogram Giá trị Giao dịch của Cụm 0 (Thang Log)

Nhận xét:

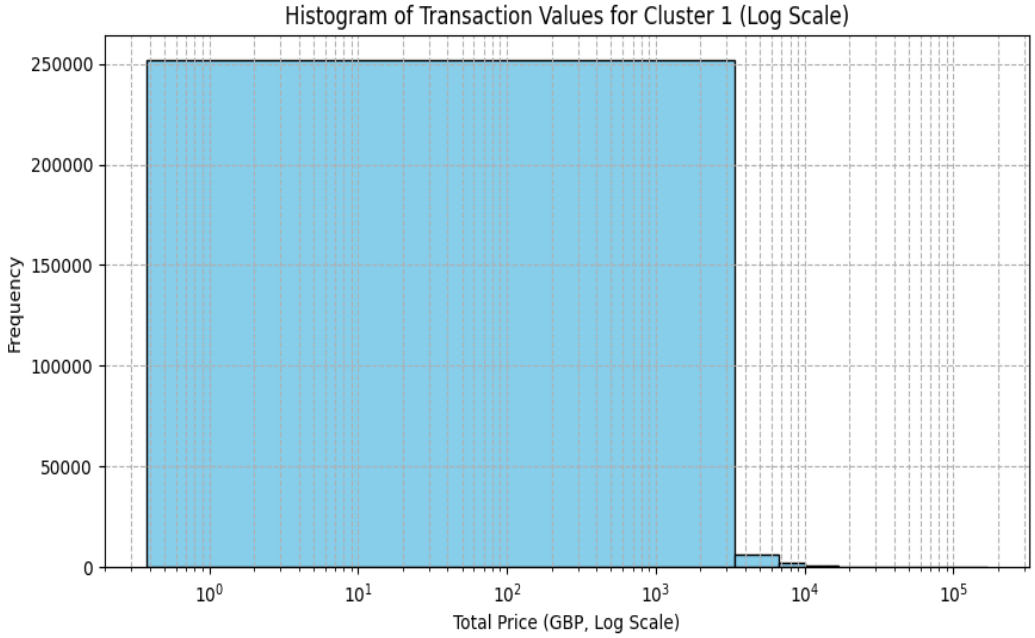
- Trung vị 308.58 GBP thấp hơn trung bình 443.58 GBP, cho thấy phân bố lệch phải với một số giao dịch giá trị cao kéo trung bình lên. Tỷ lệ 24.93% giao dịch lớn hơn 510.73 GBP xác nhận một phần đáng kể giao dịch ở mức cao. Histogram log scale cho thấy phần lớn giao dịch tập trung từ 1 GBP đến 1000 GBP, với mật độ cao ở 100 GBP và 1000 GBP, nhưng các giao dịch lớn hơn rất hiếm, có thể bị phân bố vào các bin lớn hơn (10^4 hoặc 10^5) với tần suất thấp.
- Giá trị giao dịch của nhóm này phân bố rộng từ 1 GBP đến 1000 GBP, với mật độ cao ở 100 GBP và 1000 GBP, chiếm phần lớn tần suất (trung vị 308.58

GBP). Tuy nhiên, 24.93% giao dịch lớn hơn 510.73 GBP, với trung bình 443.58 GBP, cho thấy sự hiện diện của các giao dịch giá trị cao, và giá trị tối đa có thể tồn tại nhưng cực kỳ hiếm, không nổi bật trên histogram log scale.

3.7.6.2. Phân tích giá trị giao dịch cho cụm 1

Phân tích:

- Giá trị trung bình mỗi giao dịch: 864.05 GBP
- Giá trị trung vị mỗi giao dịch: 441.10 GBP
- Tỷ lệ giao dịch lớn (trên phân vị 75%): 24.99%
- Ngưỡng giao dịch lớn (phân vị 75%): 785.79 GBP



Hình 3-22: Biểu đồ Histogram Giá trị Giao dịch của Cụm 1 (Thang Log)

Nhận xét:

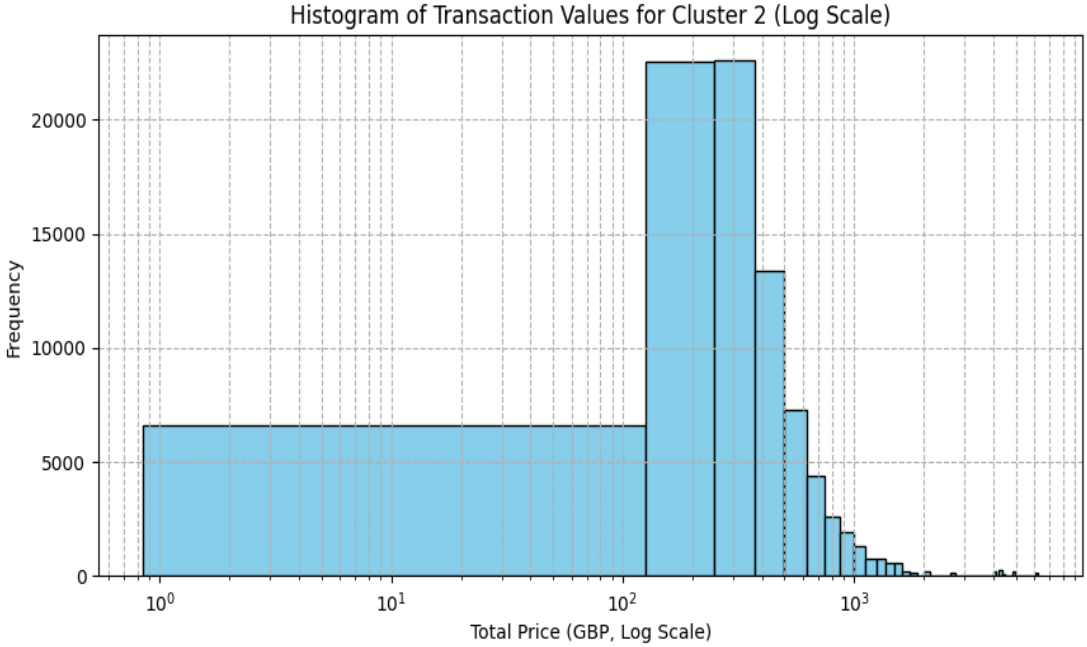
- Trung vị 441.10 GBP thấp hơn trung bình 864.05 GBP, cho thấy phân bố lệch phải với nhiều giao dịch giá trị cao kéo trung bình lên. Tỷ lệ 24.99% giao dịch lớn hơn 785.79 GBP cao nhất trong 3 cụm, cho thấy nhóm này có nhiều giao dịch giá trị lớn. Histogram log scale cho thấy phần lớn giao dịch tập trung gần 0 GBP, nhưng các giao dịch từ 10^4 đến 10^5 tồn tại, dù rất hiếm.
- Giá trị giao dịch của nhóm này chủ yếu tập trung gần 0 GBP (trung vị 441.10 GBP), nhưng trung bình cao (864.05 GBP) và 24.99% giao dịch trên 785.79 GBP cho thấy xu hướng chi tiêu lớn. Các giao dịch từ 10,000 GBP đến 175,000 GBP xuất hiện trên histogram log scale, nhưng số lượng rất ít, phản ánh tính chất ngoại lai.

3.7.6.3. Phân tích giá trị giao dịch cho cụm 2

Phân tích:

- Giá trị trung bình mỗi giao dịch: 446.28 GBP
- Giá trị trung vị mỗi giao dịch: 322.53 GBP

- Tỷ lệ giao dịch lớn (trên phân vị 75%): 24.93%
- Ngưỡng giao dịch lớn (phân vị 75%): 497.14 GBP



Hình 3-23: Biểu đồ Histogram Giá trị Giao dịch của Cụm 2 (Thang Log)

Nhận xét:

- Trung vị 322.53 GBP thấp hơn trung bình 446.28 GBP, cho thấy phân bố lệch phải với một số giao dịch giá trị cao kéo trung bình lên. Tỷ lệ 24.93% giao dịch lớn hơn 497.14 GBP cho thấy một phần đáng kể giao dịch ở mức cao. Histogram log scale cho thấy phân bố đều hơn, với tần suất đáng kể từ 10 GBP đến 1000 GBP, và các giao dịch lớn hơn rất hiếm.
- Giá trị giao dịch của nhóm này phân bố rộng từ 1 GBP đến 1000 GBP, với tần suất cao ở 100 GBP và 1000 GBP, chiếm phần lớn (trung vị 322.53 GBP). Trung bình 446.28 GBP và 24.93% giao dịch trên 497.14 GBP cho thấy khả năng chi tiêu cao, nhưng số lượng giao dịch lớn rất ít.

3.7.6.4. Đánh giá hành vi của từng cụm khách hàng

- Cụm 0: Khách hàng ngủ đông
  - Giao dịch của khách hàng tập trung chủ yếu vào giữa tuần, đặc biệt là Thứ Năm (chiếm 22.28%), và khung giờ từ 12h đến 15h, với đỉnh điểm vào lúc 12h (16.81%). Phần lớn các giao dịch có giá trị từ 1 GBP đến 1000 GBP, với trung vị là 308.58 GBP. Đáng chú ý, 24.93% giao dịch có giá trị cao hơn 510.73 GBP, thường diễn ra trong giai đoạn đầu năm, đặc biệt là Tháng 3 (17.86%).
  - Tuy nhiên, từ Tháng 6 trở đi, hoạt động giao dịch của nhóm này giảm mạnh, gần như "ngủ đông" với tỷ lệ cực thấp trong Tháng 8 (chỉ 0.51%). Điều này cho thấy nhóm khách hàng này có xu hướng “thức dậy” vào đầu

năm với cả các giao dịch nhỏ và một số giao dịch lớn đột biến (lên đến 80,000 GBP), sau đó dần rơi vào trạng thái không hoạt động vào giữa và cuối năm.
    - Điều này cho thấy Cụm 0 có hành vi giao dịch không đều, chỉ hoạt động mạnh vào đầu năm (Tháng 3) với phần lớn giao dịch từ 1 GBP đến 1000 GBP, nhưng cũng có 24.93% giao dịch lớn hơn 510.73 GBP, bao gồm các giao dịch giá trị cao hiếm hoi (tối đa 80,000 GBP). Sự tập trung vào Thứ Năm và giờ 12h-15h trong giai đoạn này phản ánh một mô hình hoạt động cố định nhưng hạn chế, phù hợp với đặc điểm "ngủ đông" trong phần lớn thời gian còn lại.
- Cụm 1: Khách hàng VIP trung thành
    - Giao dịch của nhóm này diễn ra đều đặn trong suốt tuần, với tâm điểm rơi vào Thứ Năm (20.04%) và khung giờ từ 12h đến 15h, đặc biệt là 12h trưa (18.55%), cho thấy thói quen giao dịch ổn định và có tính chu kỳ rõ ràng.
    - Điểm đáng chú ý là giao dịch tăng mạnh vào cuối năm, đặc biệt trong Tháng 11 (15.83%), gợi ý đây là thời điểm "cao trào mua sắm" của nhóm. Các giao dịch có giá trị trên 785.79 GBP chiếm 24.99%, với mức chi tiêu trung bình lên tới 864.05 GBP và có thể đạt đỉnh tới 175,000 GBP, thể hiện khả năng chi tiêu mạnh và có trọng điểm.
    - Nhóm này có thể được xem như "bùng nổ cuối năm", với hành vi chi tiêu cao, đều đặn, tập trung vào các thời điểm chiến lược – phản ánh sức mua lớn, ổn định và có mục tiêu rõ ràng.
    - Điều này cho thấy Cụm 1 có hành vi giao dịch ổn định quanh năm, với phần lớn giao dịch gần 0 GBP, nhưng tăng mạnh về giá trị (24.99% trên 785.79 GBP, tối đa 175,000 GBP) vào cuối năm (Tháng 11). Sự tập trung vào Thứ Năm và giờ 12h-15h, cùng với chi tiêu lớn vào thời điểm cao điểm, phản ánh đặc điểm trung thành và khả năng chi tiêu cao của nhóm "VIP".
- Cụm 2: Khách hàng có tiềm năng
    - Nhóm khách hàng này có xu hướng hoạt động mạnh vào cuối tuần (Chủ Nhật, 19.38%) và giữa tuần (Thứ Năm, 19.21%), đặc biệt tập trung trong khung giờ 12h-13h – thời điểm ghi nhận tỷ lệ giao dịch cao nhất (17.16% và 17.03%). Giá trị giao dịch chủ yếu nằm trong khoảng 1 – 1000 GBP, với trung vị 322.53 GBP, cho thấy nhóm này hoạt động với mức chi tiêu vừa phải.
    - Tuy nhiên, điều đáng chú ý là vào những giai đoạn cao điểm cuối năm (Tháng 11 chiếm 26.32%), nhóm này bắt đầu gia tăng giao dịch có giá trị lớn hơn 497.14 GBP (chiếm 24.93%), với mức tối đa đạt tới 6000 GBP, phản ánh tiềm năng chi tiêu đáng kể khi đúng thời điểm.

- Hành vi tiêu dùng này có thể được hiểu như một "sự trỗi dậy theo mùa", khi nhóm khách hàng này giữ mức chi tiêu ổn định trong năm nhưng sẵn sàng bùng nổ vào giai đoạn mua sắm cao điểm, thể hiện khả năng phát triển mạnh nếu được kích hoạt đúng cách.
- Điều này cho thấy Cụm 2 có hành vi giao dịch ổn định quanh năm, với phần lớn giao dịch từ 1 GBP đến 1000 GBP, nhưng tăng mạnh về giá trị (24.93% trên 497.14 GBP, tối đa 6000 GBP) vào cuối năm (Tháng 11). Sự tập trung vào Chủ Nhật và Thứ Năm, cùng với chi tiêu cao hơn vào thời điểm cao điểm, phản ánh tiềm năng phát triển chi tiêu của nhóm

3.7.6.4. Chiến lược CSKH và phương án xử lý

- Cụm 0: Khách hàng ngủ đông
  - Chiến lược CSKH:
    - Tập trung kích hoạt giao dịch vào đầu năm (Tháng 1-5, đặc biệt Tháng 3) khi nhóm này "thức dậy", bằng cách đẩy mạnh tiếp cận vào Thứ Năm và khung giờ 12h-15h, thời điểm họ hoạt động mạnh nhất. Tận dụng 24.93% giao dịch lớn hơn 510.73 GBP (trung bình 443.58 GBP) để khuyến khích các giao dịch giá trị cao, đặc biệt nhắm đến các giao dịch ngoại lai (tối đa 80,000 GBP) trong giai đoạn đầu năm.
  - Phương án xử lý:
    - Gửi thông báo hoặc ưu đãi đặc biệt vào đầu năm (Tháng 1-3), đặc biệt vào Thứ Năm từ 12h-15h, để khuyến khích giao dịch sớm, tránh "ngủ đông" kéo dài. Cung cấp ưu đãi cho các giao dịch lớn (trên 510.73 GBP) trong Tháng 3, nhằm tăng tần suất và giá trị giao dịch, đồng thời kéo dài thời gian hoạt động của nhóm sang các tháng sau (Tháng 6-8).
    - Để giải quyết tình trạng "ngủ đông" từ Tháng 6 trở đi (Tháng 8: 0.51%), thử nghiệm các chiến dịch kích hoạt nhỏ (ví dụ: ưu đãi thử nghiệm) vào giữa năm, nhắm vào Thứ Năm và giờ 12h-15h, nhằm đánh thức nhóm này.
- Cụm 1: Khách hàng VIP trung thành
  - Chiến lược CSKH:
    - Duy trì sự trung thành bằng cách tiếp cận thường xuyên vào Thứ Năm và khung giờ 12h-15h, thời điểm nhóm này giao dịch đều đặn quanh năm.
    - Tập trung đẩy mạnh giao dịch giá trị cao (24.99% trên 785.79 GBP, tối đa 175,000 GBP) vào cuối năm (Tháng 9-11, đặc biệt Tháng 11), khi nhóm này có xu hướng chi tiêu lớn nhất.
  - Phương án xử lý:

        - Cung cấp chương trình ưu đãi đặc biệt dành riêng cho "VIP" vào Thứ Năm từ 12h-15h, đặc biệt trong Tháng 11, để khuyến khích các giao dịch lớn (trên 785.79 GBP), tận dụng trung bình cao 864.05 GBP.
        - Gửi thông báo nhắc nhở hoặc ưu đãi cá nhân hóa quanh năm, đặc biệt vào Thứ Năm, để duy trì tần suất giao dịch và tăng giá trị giao dịch trung bình.
        - Để tăng giao dịch trong các tháng thấp điểm (Tháng 12: 4.79%), triển khai ưu đãi kéo dài từ Tháng 11 sang Tháng 12, nhắm đến các giao dịch giá trị cao, nhằm duy trì sự trung thành của nhóm.
- Cụm 2: Khách hàng có tiềm năng
    - Chiến lược CSKH:
        - Tận dụng tiềm năng tăng giá trị giao dịch (24.93% trên 497.14 GBP, tối đa 6000 GBP) bằng cách tập trung tiếp cận vào cuối tuần (Chủ Nhật, 19.38%) và giữa tuần (Thứ Năm, 19.21%), khung giờ 12h-13h, đặc biệt vào cuối năm (Tháng 11, 26.32%).
        - Khuyến khích giao dịch đều đặn quanh năm để tăng tần suất, đồng thời thúc đẩy các giao dịch lớn hơn vào thời điểm cao điểm.
    - Phương án xử lý:
        - Triển khai ưu đãi đặc biệt vào Chủ Nhật và Thứ Năm từ 12h-13h, đặc biệt trong Tháng 11, để khuyến khích các giao dịch lớn hơn 497.14 GBP, tận dụng tiềm năng chi tiêu cao (trung bình 446.28 GBP).
        - Gửi thông báo nhắc nhở hoặc ưu đãi nhỏ quanh năm, nhắm vào Chủ Nhật và Thứ Năm, để duy trì tần suất giao dịch và tăng giá trị giao dịch trung bình từ mức trung vị 322.53 GBP.
        - Để tăng giao dịch trong các tháng thấp điểm (Tháng 1-5, ví dụ Tháng 1: 2.57%), triển khai ưu đãi thử nghiệm vào đầu năm, nhắm đến khung giờ 12h-13h, nhằm kích thích nhóm này giao dịch thường xuyên hơn, từ đó khai thác tiềm năng phát triển.

CHƯƠNG 4: KẾT QUẢ ĐẠT ĐƯỢC

4.1. Mục tiêu và phạm vi của Project

Mục tiêu: Dự án hướng đến việc khám phá và phân tích hành vi mua sắm của khách hàng thông qua dữ liệu giao dịch trực tuyến trong năm 2011. Bằng cách áp dụng mô hình RFM (Recency - Tính mới của giao dịch, Frequency - Tần suất mua hàng, Monetary - Giá trị chi tiêu) kết hợp với thuật toán phân cụm K-Means, dự án nhằm mục tiêu phân loại khách hàng thành các nhóm đặc trưng. Phân tích đặc điểm từng nhóm khách hàng sau phân cụm để hiểu rõ chân dung tiêu dùng của từng phân khúc và đánh giá giá trị kinh tế tương ứng.

Phạm vi thực hiện: Dữ liệu được sử dụng là bộ dữ liệu Online Retail II (2011) của một công ty bán lẻ tại Anh, bao gồm hơn 499.000 giao dịch. Dự án chỉ tập trung vào các giao dịch hợp lệ sau khi đã tiến hành xử lý dữ liệu, bao gồm: loại bỏ các giao dịch bị hủy, các giá trị âm hoặc bất thường, và xử lý các giá trị thiếu. Phân tích chỉ áp dụng cho các khách hàng có mã khách hàng rõ ràng và giao dịch diễn ra tại thị trường chính là Anh Quốc.

4.2. Quy trình và phương pháp thực thiện.

Dự án được triển khai một cách hệ thống, bắt đầu từ việc nghiên cứu lý thuyết về phân tích RFM (Recency, Frequency, Monetary) và thuật toán phân cụm K-Means. Nhóm đã tổng hợp tài liệu để làm rõ nguyên lý hoạt động, các chỉ số đánh giá (Silhouette Score, Davies-Bouldin Index), cùng các kỹ thuật nâng cao như chuẩn hóa dữ liệu, Elbow Method và Dendrogram. Những khái niệm này được minh họa trực quan qua các ví dụ như phân nhóm khách hàng dựa trên hành vi mua sắm (Hình 3.8) và biểu đồ phân tán giá trị đơn hàng, giúp làm nổi bật mối quan hệ giữa các biến số.

Phần thực nghiệm tập trung vào ứng dụng phân tích RFM và K-Means trên bộ dữ liệu Online Retail II (2011), gồm 499.429 giao dịch với 8 đặc trưng (Invoice, StockCode, Quantity, …). Nhóm bắt đầu bằng việc làm sạch dữ liệu: loại bỏ 119.449 dòng thiếu Customer ID, xử lý giá trị âm/ngoại lệ (Price ≤ 0, Quantity ≤ 0), chuyển đổi kiểu dữ liệu (InvoiceDate sang datetime), và tạo cột OrderValue (Quantity × Price). Dữ liệu sau xử lý còn 370.281 giao dịch hợp lệ, được phân tích thông qua các biểu đồ histogram (Hình 3.2), scatter plo và heatmap để phát hiện xu hướng và tương quan giữa các biến.

Mô hình K-Means từ thư viện Scikit-learn được xây dựng trên 3 đặc trưng RFM đã chuẩn hóa (Recency, Frequency, Monetary). Dữ liệu được chia thành tập huấn luyện và kiểm tra, với tham số số cụm k được xác định bằng phương pháp Elbow và Silhouette Score, chọn k=3. Quá trình huấn luyện kết hợp StandardScaler để chuẩn hóa dữ liệu và Early Stopping để tránh overfitting.

Để đánh giá hiệu suất, sử dụng Silhouette Scor và Davies-Bouldin Index , đồng thời trực quan hóa qua biểu đồ phân tán và dendrogram. Phân tích sâu về hành vi khách hàng được thực hiện thông qua phân phối thời gian mua sắm và giá trị giao dịch, phản ánh rõ sự khác biệt giữa các nhóm.

4.3. Kết quả chính và phát hiện

Trong phần thực nghiệm, đã trình bày những kết quả chính thu được sau khi tiến hành xử lý, khám phá và phân tích dữ liệu từ bộ dữ liệu Online Retail II (2011). Mục tiêu của phần này là nhận diện xu hướng giao dịch, phân tích hành vi mua sắm đa chiều và phân cụm khách hàng dựa trên chỉ số RFM kết hợp thuật toán K-Means. Những phát hiện này không chỉ giúp khắc họa thói quen và giá trị của từng nhóm khách hàng, mà còn làm cơ sở cho việc đề xuất các chiến lược tiếp thị, chăm sóc và kích hoạt phù hợp.

- Tiền xử lý dữ liệu: Từ 499.429 giao dịch ban đầu, quá trình xử lý đã sàng lọc, giữ lại 370.281 dòng dữ liệu chất lượng cao sau khi loại bỏ các mã không phải sản phẩm (như POST, BANK CHARGES) và các giao dịch hủy, giá trị âm hoặc bằng không.
- Xu hướng giao dịch: Giao dịch đạt đỉnh vào Tháng 11 với hơn 64.000 lượt, phản ánh sức nóng của mùa mua sắm cuối năm, nhưng lại giảm đột ngột xuống dưới 20.000 lượt vào Tháng 12, đặt ra một ẩn số cần giải đáp.
- Hành vi khách hàng đa sắc thái:
    - Phần lớn khách hàng thực hiện từ 10 đến 100 giao dịch, trong khi một nhóm nhỏ đặc biệt trung thành vượt ngưỡng 1.000 giao dịch.
    - Số lượng sản phẩm bán ra tập trung chủ yếu ở mức 1-10 sản phẩm, với các đỉnh phụ tại 2, 5, 10 và 20 sản phẩm, thể hiện các mô hình mua sắm phổ biến.
- Phân cụm khách hàng:
    - Cụm 0 (Ngủ đông): Gồm 937 khách, với trung bình 233 ngày không giao dịch, tần suất thấp (0,82 lần), và giá trị chi tiêu khiêm tốn (5,60 đơn vị).
    - Cụm 1 (VIP trung thành): Bao gồm 1.284 khách, với chỉ 28 ngày kể từ lần mua gần nhất, tần suất cao (2,12 lần), và giá trị chi tiêu vượt trội (7,93 đơn vị).
    - Cụm 2 (Tiềm năng): Với 1.993 khách, có trung bình 51 ngày không giao dịch, tần suất vừa phải (1,03 lần), và giá trị chi tiêu trung bình (6,12 đơn vị).
- Giá trị vòng đời và thời gian giao dịch:
    - Cụm 1 dẫn đầu với CLV cao nhất, khẳng định tiềm năng sinh lời lớn; Cụm 2 cho thấy cơ hội phát triển; Cụm 0 có CLV thấp nhất, cần biện pháp kích hoạt.
    - Cụm 0 hoạt động sôi nổi vào Tháng 3, trong khi cụm 1 và 2 bứt phá vào Tháng 11. Giờ cao điểm là 12h-15h, với Thứ Năm là ngày giao dịch nổi bật nhất.
- Giá trị giao dịch đa dạng:
    - Cụm 1 ghi nhận trung bình 864,05 GBP mỗi giao dịch, với đỉnh cao lên tới 175.000 GBP.

- Cụm 0 và 2 dao động trung bình quanh 443-446 GBP, với 24,93-24,99% giao dịch vượt ngưỡng 497-785 GBP, chứng tỏ khả năng chi tiêu đáng kể trong một số trường hợp.

Qua các kết quả đạt được đã chỉ ra rằng phân khúc khách hàng giá trị cao cần được ưu tiên chăm sóc, nhóm tiềm năng có thể kích hoạt bằng khuyến mãi và nhóm ngủ đông cần chiến lược tái tương tác.

4.4. Thảo luận về ý nghĩa và đóng góp kết quả

Những phát hiện này không chỉ mang lại giá trị thực tiễn trong tối ưu hóa chiến lược chăm sóc khách hàng, mà còn đóng góp học thuật khi minh chứng hiệu quả của các kỹ thuật học máy trong bối cảnh bán lẻ.

Việc áp dụng RFM kết hợp thuật toán K-Means đã vẽ nên bức tranh toàn diện về ba phân khúc khách hàng nhóm *"VIP trung thành",* nhóm *"Tiềm năng"* và nhóm *"Ngủ đông"* giúp doanh nghiệp thấu hiểu sâu sắc hành vi mua sắm suốt năm 2011. Trên thực tế, các mốc thời gian "vàng" như Tháng 11, khung giờ 12h–15h vào các ngày trong tuần (đặc biệt là Thứ Năm) đã được xác định rõ để tối ưu hóa chiến dịch quảng cáo và ưu đãi. Từ đó, chiến lược chăm sóc khách hàng trở nên hiệu quả hơn: ưu tiên giữ chân khách VIP, kích hoạt lại nhóm tiềm năng bằng các chương trình cá nhân hóa, và khơi gợi sự quan tâm của nhóm ngủ đông qua các chương trình tái tương tác. Về mặt học thuật, nghiên cứu này chứng minh sự hiệu quả của việc kết hợp trực quan hóa dữ liệu, phân tích RFM và đánh giá phân cụm trong lĩnh vực bán lẻ, đồng thời mở ra hướng tiếp cận mới cho các ứng dụng học máy trong phân tích hành vi người tiêu dùng

4.5. Các hạn chế và đề xuất khắc phục

Mặc dù đã mang lại nhiều kết quả giá trị, vẫn tồn tại một số hạn chế cần khắc phục và định hướng cải tiến trong tương lai. Đầu tiên, dữ liệu tháng 12 đang bị thiếu hụt nghiêm trọng số lượt giao dịch giảm từ hơn 64.000 xuống dưới 20.000 khiến xu hướng cuối năm có thể không phản ánh đầy đủ thực tế. Thứ hai, vẫn còn những giá trị ngoại lai như đơn hàng với Quantity lên tới 80.995 hoặc Price 38.970 GBP, nhiều khả năng do lỗi nhập liệu, làm méo mó phân tích. Thứ ba, phạm vi nghiên cứu chỉ giới hạn trong năm 2011, chưa cho phép so sánh thay đổi hành vi theo chu kỳ nhiều năm. Cuối cùng, chỉ số Silhouette Score (0,42) cho thấy cấu trúc phân cụm chưa thực sự tối ưu.

Để khắc phục những hạn chế này, nghiên cứu có thể được mở rộng bằng cách bổ sung dữ liệu đầy đủ tháng 12 và cân nhắc bao gồm thêm các năm liền kề (2010, 2012) để theo dõi xu hướng dài hạn. Ngoại lệ nên được xử lý chặt chẽ hơn bằng các phương pháp IQR hoặc Z-score để loại bỏ các bản ghi bất thường. Ngoài ra, việc thử nghiệm các thuật toán phân cụm khác như DBSCAN, Hierarchical Clustering hoặc khảo sát thêm với số cụm lớn hơn (k = 4, 5) có thể cải thiện tính ổn định của kết quả. Cuối cùng, tích hợp các yếu tố ngoại cảnh chẳng hạn dữ liệu thời tiết, sự kiện kinh tế hay chương trình khuyến mãi sẽ giúp giải thích sâu hơn động lực mua sắm và nâng cao tính ứng dụng của nghiên cứu.

4.6. Kết luận chung

Phân tích RFM kết hợp K-Means đã thành công trong việc phân nhóm khách hàng dựa trên hành vi mua sắm, cung cấp cơ sở khoa học để tối ưu chiến lược CRM. Kết quả cho thấy rõ sự khác biệt giữa các nhóm khách hàng, từ đó đưa ra các chiến lược vào nhóm mang lại giá trị cao nhất. Dù còn nhiều hạn chế nhưng phần thực nghiệm này cũng giúp mở ra cho các hướng nghiên cứu sau này của các bài toán.

CHƯƠNG 5: KẾT LUẬN, ƯU ĐIỂM, NHƯỢC ĐIỂM, HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Trong thời đại thương mại điện tử phát triển mạnh mẽ, việc hiểu rõ hành vi tiêu dùng của khách hàng là yếu tố then chốt giúp doanh nghiệp nâng cao hiệu quả kinh doanh và duy trì lợi thế cạnh tranh. Nghiên cứu này đã áp dụng các kỹ thuật phân khúc khách hàng dựa trên hành vi mua sắm trực tuyến, sử dụng tập dữ liệu thực tế Online Retail 2011, cùng với các phương pháp phân tích như phân cụm K-Means và phân tích RFM (Recency - Frequency - Monetary).

Kết quả thu được cho thấy rằng khách hàng có thể được chia thành 3 nhóm chính:

- Cụm 0 – Khách hàng tiềm năng: là nhóm có mức chi tiêu chưa cao nhưng có tần suất truy cập/mua hàng gần đây, cho thấy tiềm năng chuyển đổi thành khách hàng thân thiết.
- Cụm 1 – Khách hàng trung thành, VIP: đây là nhóm mang lại giá trị cao nhất, mua hàng thường xuyên, gần đây và chi tiêu lớn.
- Cụm 2 – Khách hàng có giá trị tiềm ẩn: là nhóm có chi tiêu khá nhưng có thể đã ngưng mua hàng gần đây, cần chiến lược kích hoạt lại.

Bằng cách sử dụng các công cụ như biểu đồ Elbow, Silhouette Score (0.42), cùng với các chỉ số đánh giá như Davies-Bouldin Index và Calinski-Harabasz Index, nghiên cứu xác định số cụm tối ưu là k = 3, với hiệu suất phân cụm tương đối tốt. Đồng thời, phân tích giá trị vòng đời khách hàng (CLV) và hành vi mua theo thời gian đã cung cấp cái nhìn sâu sắc về xu hướng mua sắm, giúp doanh nghiệp hiểu rõ khi nào khách hàng có xu hướng chi tiêu nhiều nhất trong năm hoặc trong tuần, từ đó lập kế hoạch marketing chính xác hơn.

Tóm lại, nghiên cứu này chứng minh rằng việc kết hợp phân tích RFM và phân cụm dữ liệu là một hướng đi hiệu quả, hỗ trợ doanh nghiệp phân khúc khách hàng và cá nhân hóa chiến lược tiếp cận để tối đa hóa lợi nhuận.

5.2. Điểm mạnh của nghiên cứu

- Tính ứng dụng cao: Nghiên cứu sử dụng tập dữ liệu thực tế từ môi trường kinh doanh trực tuyến, đảm bảo tính thực tiễn và khả năng áp dụng cho các doanh nghiệp thương mại điện tử hiện nay.
- Kết hợp mô hình phân tích hiệu quả: Việc tích hợp phân cụm K-Means với phân tích RFM giúp nắm bắt tốt đặc điểm hành vi tiêu dùng, từ mức độ mua gần đây, tần suất giao dịch đến giá trị tiền chi tiêu.
- Xử lý dữ liệu chặt chẽ: Dữ liệu đầu vào được tiền xử lý cẩn thận: loại bỏ giá trị thiếu, bất thường, chuẩn hóa biến đầu vào – những bước cần thiết để đảm bảo mô hình hoạt động ổn định và chính xác.
- Hỗ trợ ra quyết định qua trực quan hóa: Các biểu đồ trực quan như biểu đồ tần suất, scatter plot, heatmap và biểu đồ thời gian mua sắm đã giúp làm nổi bật sự

khác biệt giữa các nhóm khách hàng và giúp nhà quản trị dễ dàng hiểu, phân tích.

- Phân tích CLV và hành vi thời gian: Việc đánh giá giá trị vòng đời khách hàng cùng với việc phân tích thói quen chi tiêu theo giờ, ngày, tháng... giúp doanh nghiệp xác định rõ khách hàng giá trị cao, từ đó tối ưu hóa chi phí quảng cáo và chăm sóc khách hàng.

5.3. Hạn chế của nghiên cứu

- Dữ liệu giới hạn về thời gian: Dữ liệu chỉ bao gồm các giao dịch trong năm 2011, do đó không phản ánh được xu hướng thay đổi hành vi khách hàng theo thời gian hoặc theo mùa.
- Thiếu thông tin bổ trợ: Bộ dữ liệu không bao gồm các yếu tố nhân khẩu học (tuổi, giới tính, nghề nghiệp) hoặc hành vi duyệt web, điều này làm giảm khả năng xây dựng hồ sơ khách hàng toàn diện.
- Phương pháp K-Means có nhược điểm: K-Means yêu cầu xác định trước số lượng cụm và giả định cụm có hình tròn đều (spherical clusters), khiến nó không phù hợp nếu dữ liệu có cấu trúc phức tạp hoặc chứa nhiều ngoại lệ.
- Điểm Silhouette chưa cao: Mặc dù mô hình đạt điểm Silhouette là 0.42 – ở mức chấp nhận được, nhưng chưa phản ánh được sự phân tách rõ ràng giữa các cụm.
- Thiếu kiểm chứng thực tế: Các chiến lược chăm sóc và tiếp cận khách hàng được đề xuất chưa được thử nghiệm trong môi trường kinh doanh thực tế, nên chưa thể đánh giá mức độ hiệu quả thực sự.

5.4. Hướng phát triển trong tương lai

Để nâng cao chất lượng nghiên cứu và mở rộng tính ứng dụng thực tế, một số hướng phát triển được đề xuất như sau:

- Mở rộng dữ liệu và tích hợp nhiều nguồn thông tin hơn: Sử dụng dữ liệu từ nhiều năm hoặc tích hợp các yếu tố như thông tin cá nhân, hành vi duyệt web, lịch sử phản hồi, để xây dựng mô hình hành vi khách hàng toàn diện hơn.
- Thử nghiệm các mô hình phân cụm khác: Các thuật toán như DBSCAN, Mean Shift hoặc phân cụm phân cấp (Hierarchical Clustering) có thể xử lý dữ liệu không đều và phát hiện cấu trúc phân cụm linh hoạt hơn mà không cần xác định trước số cụm.
- Ứng dụng học máy tiên tiến: Kết hợp các mô hình học máy như Random Forest, XGBoost, hoặc thậm chí học sâu (deep learning) để dự đoán hành vi mua hàng hoặc phân khúc khách hàng theo thời gian thực.
- Thử nghiệm chiến lược trong môi trường thực tế: Các chiến dịch chăm sóc khách hàng, tặng thưởng, cá nhân hóa ưu đãi nên được triển khai thử nghiệm A/B Testing để đánh giá hiệu quả thật sự và tối ưu hóa theo kết quả thực tế.
- Tích hợp mô hình vào hệ thống CRM hoặc Dashboard: Xây dựng công cụ hỗ trợ quản trị khách hàng, trực quan hóa phân khúc, theo dõi CLV, từ đó giúp đội

ngũ marketing hoặc chăm sóc khách hàng đưa ra hành động phù hợp theo từng nhóm.

TÀI LIỆU THAM KHẢO

[1]. T. M. Mitchell, *Machine Learning*, New York: McGraw-Hill, 1997
[2]. L. N. Khánh, "Bài 2: Phân nhóm các thuật toán Machine Learning," Machine Learning Cơ Bản, Dec. 27, 2016. [Online]. Available: https://machinelearningcoban.com/2016/12/27/categories/.
[3]. DataCamp, "Clustering in Machine Learning: 5 Essential Clustering Algorithms," DataCamp, [Online]. Available: https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms.
[4]. Analytics Vidhya, "4 Types of Distance Metrics in Machine Learning," *Analytics Vidhya*, Feb. 6, 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/.
[5]. https://blog.tomorrowmarketers.org/phan-tich-rfm-la-gi/
[6]. https://www.antsomi.com/vi/2024/02/29/rfm-model-co-hieu-qua-cho-viec-phan-tich-chan-dung-khach-hang/
[7]. https://machinelearningcoban.com/2017/01/01/kmeans/
[8]. https://phamdinhkhanh.github.io/deepai-book/ch_ml/KMeans.html