**TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG**

**KHOA THỐNG KÊ – TIN HỌC**

∙∙∙∙∙🙞⯎🙜∙∙∙∙



**BÁO CÁO BÀI TẬP NHÓM**

**MÔN: PHÂN TÍCH DỮ LIỆU BẰNG PYTHON**

**ỨNG DỤNG MÔ HÌNH RFM KẾT HỢP THUẬT TOÁN K-MEANS TRONG PHÂN KHÚC KHÁCH HÀNG THEO HÀNH VI MUA SẮM PHỤC VỤ CHIẾN LƯỢC CRM**

Giảng viên hướng dẫn : TS. Lê Diên Tuấn

Nhóm : 5

Sinh viên thực hiện : Nguyễn Võ Ngọc Nhi (21,6%)

(Số di động: 0934816654)

Nguyễn Thị Thanh Bình (19,8%)

Nguyễn Trương Thùy Dương (19,8%)

Nguyễn Thị Huỳnh Nga (19,4%)

Lê Cẩm Tú (19,4%)

*Đà Nẵng, ngày 20/4/2025*

**MỤC LỤC**

[CHƯƠNG 1 TỔNG QUAN ĐỀ TÀI 7](#_Toc196408907)

[1.1 Giới thiệu tổng quan về chủ đề: 7](#_Toc196408908)

[CHƯƠNG 2 XÂY DỰNG CƠ SỞ LÝ THUYẾT 8](#_Toc196408909)

[2.1 Quản lý quan hệ khách hàng (CRM - Customer Relationship Management): 8](#_Toc196408910)

[2.2 Mô hình RFM trong phân khúc khách hàng: 8](#_Toc196408911)

[2.3 Bài toán phân cụm K-means: 9](#_Toc196408912)

[2.4 EDA (Exploratory Data Analysis - Phân tích khám phá dữ liệu): 10](#_Toc196408913)

[2.5 Tiền xử lý dữ liệu: 11](#_Toc196408914)

[CHƯƠNG 3 KẾT QUẢ MINH CHỨNG ĐƯỢC TRIỂN KHAI BẰNG PYTHON 12](#_Toc196408915)

[3.1 Xây dựng tổng thể (Framework): 12](#_Toc196408916)

[3.1.1 Thu thập dữ liệu 12](#_Toc196408917)

[3.1.2 Khám phá dữ liệu 12](#_Toc196408918)

[3.1.3 Tiền xử lý dữ liệu: 12](#_Toc196408919)

[3.1.3.1 Xử lý dữ liệu không nhất quán 12](#_Toc196408920)

[3.1.3.2 Tính toán giá trị RFM 12](#_Toc196408921)

[3.1.3.3 Xử lý dữ liệu ngoại lai (outlier): 12](#_Toc196408922)

[3.1.3.4 Chuẩn hóa dữ liệu: 12](#_Toc196408923)

[3.1.4 Xây dựng mô hình: 13](#_Toc196408924)

[3.1.4.1 Xác định số cụm 13](#_Toc196408925)

[3.1.4.2 Phân cụm theo thuật toán K-means 13](#_Toc196408926)

[3.1.4.3 Trực quan hóa kết quả 13](#_Toc196408927)

[3.1.5 Đề xuất: Đề xuất chiến lược CRM phù hợp cho từng phân khúc 13](#_Toc196408928)

[3.2 Chọn dữ liệu (Dataset) 13](#_Toc196408929)

[3.2.1 Tổng quan dữ liệu: 13](#_Toc196408930)

[3.2.2 Mô tả dữ liệu: 14](#_Toc196408931)

[3.3 Triển khai bằng Python: 15](#_Toc196408932)

[3.3.1 Khám phá dữ liệu ban đầu: 15](#_Toc196408933)

[3.3.1.1 Đọc dữ liệu: 15](#_Toc196408934)

[3.3.1.2 Tìm hiểu tổng quan dữ liệu: 15](#_Toc196408936)

[3.3.1.3 Kiểm tra các giá trị trùng lặp: 16](#_Toc196408937)

[3.3.1.4 Kiểm tra các giá trị nhiễu: 17](#_Toc196408938)

[3.3.1.5 Biểu diễn dữ liệu gốc: 18](#_Toc196408939)

[3.3.2 Tiền xử lý dữ liệu 23](#_Toc196408940)

[3.3.2.1 Loại bỏ cột không cần thiết: 23](#_Toc196408941)

[3.3.2.2 Xử lý dữ liệu không nhất quán: 24](#_Toc196408942)

[3.3.2.3 Tính toán giá trị RFM: 24](#_Toc196408943)

[3.3.2.4 Phân tích đặc trưng hành vi: 26](#_Toc196408944)

[3.3.2.5 Phân tích tương quan: 31](#_Toc196408945)

[3.3.2.6 Xử lý các giá trị ngoại lai: 32](#_Toc196408946)

[3.3.2.7 Chuẩn hóa dữ liệu: 34](#_Toc196408947)

[3.3.3 Xây dựng mô hình: 36](#_Toc196408948)

[3.3.3.1 Xác định K tối ưu bằng Elbow Method: 36](#_Toc196408949)

[3.3.3.2 Phân cụm bằng thuật toán K-means: **37**](#_Toc196408950)

[3.4 Đề xuất chiến lược CRM phù hợp với từng phân khúc khách hàng: 41](#_Toc196408951)

[CHƯƠNG 4 KẾT LUẬN ĐỀ TÀI 43](#_Toc196408952)

**LỜI MỞ ĐẦU**

Trong thời đại 4.0, khi trí tuệ nhân tạo và công nghệ dữ liệu phát triển mạnh mẽ, việc phân tích dữ liệu không chỉ còn là một lợi thế mà đã trở thành yếu tố thiết yếu giúp doanh nghiệp dự báo xu hướng, đưa ra quyết định chính xác và tối ưu hóa hiệu quả kinh doanh. Dù các công cụ thông minh ngày càng phổ biến, vai trò của con người trong việc hiểu, xử lý và khai thác dữ liệu vẫn giữ vai trò quan trọng.

Xuất phát từ nhận thức đó, nhóm đã lựa chọn bộ dữ liệu “Store Sales Forecasting Dataset” trên nền tảng Kaggle để thực hiện báo cáo cho học phần Phân tích dữ liệu bằng Python. Mục tiêu của báo cáo là phân tích hành vi tiêu dùng dựa trên dữ liệu giao dịch, từ đó phân nhóm khách hàng theo đặc điểm tương đồng bằng cách ứng dụng mô hình RFM kết hợp thuật toán phân cụm K-Means. Kết quả phân tích kỳ vọng sẽ giúp khám phá các đặc điểm tiềm ẩn trong dữ liệu, hỗ trợ doanh nghiệp trong việc ra quyết định chiến lược phù hợp hơn với từng phân khúc khách hàng

**DANH MỤC HÌNH ẢNH**

[Hình 1 Mô phỏng quy trình nghiên cứu (Framework) 7](#_Toc196084851)

[Hình 2 Kết quả của việc tìm hiểu tổng quan bộ dữ liệu 10](#_Toc196084852)

[Hình 3 Kết quả kiểm tra các giá trị trùng lặp 11](#_Toc196084853)

[Hình 4 Kết quả kiểm tra các giá trị nhiễu 12](#_Toc196084854)

[Hình 5 Biểu đồ cột biểu diễn tần suất xuất hiện của từng giá trị trong các cột có kiểu dữ liệu object 14](#_Toc196084855)

[Hình 6 Biểu đồ tròn thể hiện tỷ lệ phân bố của các giá trị - Phân khúc khách hàng theo hành vi mua sắm (Segment) 15](#_Toc196084856)

[Hình 7 Biểu đồ tròn thể hiện tỷ lệ phân bố của các giá trị - Phương thức giao hàng (Ship Mode) 16](#_Toc196084857)

[Hình 8 Biểu đồ tròn thể hiện tỷ lệ phân bố của các giá trị - Vùng miền (Region) 17](#_Toc196084858)

[Hình 9 Kết quả của việc loại bỏ các cột không cần thiết 18](#_Toc196084859)

[Hình 10 Kết quả xử lý dữ liệu không nhất quán 18](#_Toc196084860)

[Hình 11 Kết quả tính toán các giá trị R,F,M 20](#_Toc196084861)

[Hình 12 Biểu đồ Histogram thể hiện đặc trưng hành vi mua sắm của khách hàng theo phân bố của các chỉ số R, F, M 21](#_Toc196084862)

[Hình 13 Biểu đồ Box plot để kiểm tra sự tồn tại của các giá trị ngoại lệ trong ba đặc trưng R, F, M 23](#_Toc196084863)

[Hình 14 Biểu đồ phân tán thể hiện các cụm khách hàng có đặc điểm hành vi tương đồng 24](#_Toc196084864)

[Hình 15 Biểu đồ Heatmap biểu diễn trực quan ma trận tương quan 26](#_Toc196084865)

[Hình 16 Kết quả dùng IQR để phát hiện và xử lý các giá trị ngoại lai 28](#_Toc196084866)

[Hình 17 Kết quả chuẩn hóa dữ liệu R, F, M theo Z-score Normalization 29](#_Toc196084867)

[Hình 18 Biểu đồ thể hiện mật độ phân phối theo Z-score cho R, F, M 29](#_Toc196084868)

[Hình 19 Đồ thị thể hiện kết quả của phương pháp Elbow 31](#_Toc196084869)

[Hình 20 Biểu đồ phân tán thể hiện kết quả phân cụm khách hàng 32](#_Toc196084870)

[Hình 21 Biểu đồ 3D thể hiện kết quả phân cụm khách hàng 34](#_Toc196084871)

**TỔNG QUAN BÀI BÁO CÁO**

1. **Mục tiêu bài báo cáo:**

* Hiểu và vận dụng các kiến thức về quản trị quan hệ khách hàng (CRM), mô hình RFM và thuật toán phân cụm K-Means trong phân tích dữ liệu.
* Xử lý, trực quan hóa và phân tích dữ liệu khách hàng từ tập dữ liệu thực tế.
* Thực hiện các bước xử lý dữ liệu giao dịch, tính toán chỉ số RFM và phân cụm khách hàng dựa trên hành vi mua sắm.
* Trực quan hóa kết quả phân tích để hỗ trợ đề xuất các chiến lược CRM phù hợp với từng phân khúc khách hàng nhằm nâng cao hiệu quả kinh doanh.

1. **Đối tượng, phạm vi báo cáo:**

* **Đối tượng nghiên cứu:** Hành vi tiêu dùng của khách hàng trong lĩnh vực bán lẻ trực tuyến, thông qua dữ liệu giao dịch trong quá khứ.
* **Phạm vi nghiên cứu:**
  + Đề tài được thực hiện trên phạm vi tập dữ liệu giao dịch khách hàng của một chuỗi cửa hàng bán lẻ trực tuyến có trụ sở tại Hoa Kỳ, với các đơn hàng được ghi nhận từ ngày 03/09/2016 đến 14/01/2017.
  + Các giải thuật phân cụm và xác định số cụm trong khai phá dữ liệu

1. **Phương pháp thực hiện báo cáo:**

Báo cáo được thực hiện dựa trên sự kết hợp giữa nghiên cứu lý thuyết và phân tích dữ liệu thực nghiệm, cụ thể như sau:

* Phương pháp nghiên cứu lý thuyết: Nhóm tiến hành thu thập, tìm hiểu và hệ thống hóa các lý thuyết liên quan đến quản trị quan hệ khách hàng (CRM), mô hình RFM, và thuật toán phân cụm K-Means thông qua các tài liệu chuyên ngành, sách, báo cáo khoa học và các nguồn học thuật đáng tin cậy. Mục tiêu là xây dựng nền tảng lý luận vững chắc để ứng dụng vào phân tích dữ liệu.
* Phương pháp phân tích định lượng: Báo cáo sử dụng phương pháp phân tích định lượng để xử lý và khai thác dữ liệu giao dịch từ bộ dữ liệu thực tế. Thông qua các chỉ số định lượng như Recency, Frequency, Monetary và thuật toán học máy không giám sát K-Means, nhóm thực hiện việc phân cụm khách hàng dựa trên hành vi mua sắm.
* Phương pháp mô hình hóa và trực quan hóa dữ liệu: Dữ liệu sau khi xử lý được mô hình hóa dưới dạng các cụm khách hàng nhằm phân biệt các đặc điểm hành vi. Kết quả được trực quan hóa bằng các biểu đồ nhằm hỗ trợ phân tích và đưa ra định hướng chiến lược CRM phù hợp.
* Phương pháp thực nghiệm: Quá trình xử lý và phân tích dữ liệu được thực hiện bằng ngôn ngữ Python với sự hỗ trợ của các thư viện như Pandas, Matplotlib, Seaborn, Scikit-learn,... Đây là phương pháp thực nghiệm giúp kiểm chứng khả năng ứng dụng lý thuyết vào thực tế.

# TỔNG QUAN ĐỀ TÀI

## Giới thiệu tổng quan về chủ đề:

Quản trị quan hệ khách hàng (Customer Relationship Management – CRM) là một chiến lược kinh doanh quan trọng, giúp doanh nghiệp xây dựng, duy trì và phát triển mối quan hệ lâu dài với khách hàng. Trong bối cảnh thị trường cạnh tranh gay gắt và hành vi tiêu dùng thay đổi liên tục, việc hiểu rõ nhu cầu, hành vi và giá trị của khách hàng là yếu tố then chốt giúp doanh nghiệp giữ chân khách hàng và gia tăng lợi nhuận

Theo tổng quan nghiên cứu (CRM-Literature Review), các phương pháp quản trị quan hệ khách hàng ngày nay không chỉ dừng lại ở việc lưu trữ dữ liệu, mà đã chuyển sang hướng khai thác và phân tích dữ liệu để dự đoán hành vi, đánh giá giá trị vòng đời khách hàng và cá nhân hóa chiến lược tiếp cận. Nhiều nghiên cứu cũng khẳng định vai trò của mô hình RFM (Recency - Frequency - Monetary) trong việc phân khúc khách hàng hiệu quả, cùng với sự hỗ trợ của các thuật toán học máy như K-Means, Hierarchical Clustering trong quá trình phân cụm khách hàng [1]. Nếu không ứng dụng hiệu quả các kỹ thuật phân tích dữ liệu trong CRM, doanh nghiệp có thể gặp khó khăn trong việc giữ chân khách hàng, dễ triển khai các chiến dịch marketing kém hiệu quả, từ đó làm sụt giảm doanh thu và đánh mất lợi thế cạnh tranh. [1]

Xuất phát từ định hướng đó, đề tài “Ứng dụng mô hình RFM kết hợp thuật toán K-Means trong phân khúc khách hàng theo hành vi mua sắm phục vụ chiến lược CRM” tập trung vào khai thác dữ liệu bán hàng từ tập *“Store Sales Forecasting Dataset”* được công bố trên nền tảng Kaggle. Thay vì chỉ dừng lại ở việc chấm điểm hoặc chia nhóm theo ngưỡng cố định như cách làm truyền thống, đề tài nghiên cứu kết hợp RFM với thuật toán phân cụm K-Means – một phương pháp học máy không giám sát – để tự động phân nhóm khách hàng có hành vi tương đồng. Sự kết hợp này giúp tự động phân cụm dựa trên dữ liệu thực tế, tăng tính khách quan và phản ánh chính xác hành vi người tiêu dùng. Kết quả nghiên cứu kỳ vọng sẽ hỗ trợ doanh nghiệp trong việc xây dựng chiến lược chăm sóc khách hàng hiệu quả hơn, từ đó nâng cao mức độ hài lòng và giữ chân khách hàng trong dài hạn

Chủ đề nghiên cứu về việc ứng dụng các kỹ thuật phân tích dữ liệu trong quản lý quan hệ khách hàng (CRM) có mối liên hệ chặt chẽ với học phần Phân tích dữ liệu. Trong suốt quá trình thực hiện, các bước như tiền xử lý dữ liệu, chuẩn hóa, trích xuất đặc trưng, và áp dụng các thuật toán học máy không giám sát như K-Means được sử dụng nhằm phân nhóm khách hàng theo hành vi. Đây là những nội dung cốt lõi trong học phần, giúp hiểu cách khai thác dữ liệu lớn phục vụ mục tiêu kinh doanh. Ngoài ra, việc phân tích và trực quan hóa dữ liệu CRM còn giúp nhóm hiểu sâu hơn về lý thuyết, đồng thời rèn luyện khả năng vận dụng kiến thức vào các tình huống thực tế trong môi trường kinh doanh. Việc triển khai các kỹ thuật phân tích dữ liệu trong CRM không chỉ hỗ trợ cá nhân hóa sản phẩm, dịch vụ mà còn là minh chứng rõ ràng cho vai trò quan trọng của phân tích dữ liệu trong việc nâng cao hiệu quả hoạt động doanh nghiệp. Chính vì vậy, việc thực hành với các phương pháp phân tích dữ liệu trong bối cảnh CRM cũng góp phần rèn luyện tư duy phân tích và khả năng áp dụng kiến thức vào môi trường doanh nghiệp thực tế, từ đó nâng cao năng lực chuyên môn và kỹ năng giải quyết vấn đề bằng dữ liệu.

# XÂY DỰNG CƠ SỞ LÝ THUYẾT

## Quản lý quan hệ khách hàng (CRM - Customer Relationship Management):

Quản lý quan hệ khách hàng (CRM – Customer Relationship Management) là một phương pháp giúp các doanh nghiệp thu thập, phân tích và quản lý các tương tác của khách hàng một cách có hệ thống và hiệu quả, các thông tin của khách hàng từ đó hình thành mối quan hệ trong kinh doanh.

CRM thường được phân loại thành ba nhóm chính**: CRM vận hành (Operational CRM)** – hỗ trợ tự động hóa các quy trình bán hàng, tiếp thị và chăm sóc khách hàng; **CRM phân tích (Analytical CRM)** – tập trung vào khai thác dữ liệu khách hàng nhằm phục vụ việc ra quyết định và **CRM cộng tác (Collaborative CRM)** – thúc đẩy sự phối hợp giữa các bộ phận trong và ngoài tổ chức để nâng cao trải nghiệm khách hàng

Nhiệm vụ của quản lý quan hệ khách hàng (CRM) trong hệ thống doanh nghiệp:

* Đối với khách hàng:
  + Cá nhân hóa trải nghiệm khách hàng: CRM giúp doanh nghiệp lưu trữ thông tin chi tiết về khách hàng, bao gồm lịch sử mua hàng, sở thích và phản hồi, từ đó cung cấp dịch vụ và sản phẩm phù hợp, gia tăng sự hài lòng và trung thành của khách hàng.
  + Tăng cường tương tác và chăm sóc khách hàng: Hệ thống CRM cho phép doanh nghiệp tương tác với khách hàng qua nhiều kênh khác nhau như email, điện thoại và mạng xã hội, đảm bảo phản hồi nhanh chóng và hiệu quả, giúp xây dựng mối quan hệ bền vững.
* Đối với doanh nghiệp:
  + CRM giúp doanh nghiệp quản lý, lắng nghe ý kiến và bảo mật thông tin khách hàng, từ đó mà dễ dàng quản lý tình hình kinh doanh và phát triển của doanh nghiệp.
  + CRM giúp doanh nghiệp quảng bá sản phẩm, thương hiệu nhanh chóng, dễ dàng và tiết kiệm chi phí, từ đó tăng doanh số bán hàng, giảm thiểu rủi ro, chi phí, tối ưu hóa hiệu quả của nhân viên kinh doanh, tập trung vào việc quản lý khách hàng một cách hiệu quả nhất.

## Mô hình RFM trong phân khúc khách hàng:

RFM là một phương pháp được sử dụng để phân loại và quản lý khách hàng. Nó thường được sử dụng trong marketing cơ sở dữ liệu (dựa trên phân tích dữ liệu hành vi của khách hàng để triển khai các chiến lược tiếp thị phù hợp) và marketing trực tiếp bằng cách sử dụng 3 yếu tố RFM để chia khách hàng thành các nhóm riêng biệt, mà từ đó xác định được chiến lược tiếp thị phù hợp cho từng nhóm khách hàng.

Theo nguyên lý Pareto “20% khách hàng sẽ mang lại 80% doanh số”, do đó doanh nghiệp cần xác định được những khách hàng quan trọng nhất để chăm sóc đặc biệt. RFM định lượng giá trị của một khách hàng dựa trên 3 yếu tố chính:

* Recency: Khoảng thời gian mua hàng gần đây nhất là bao lâu. Cho biết khách hàng có đang thực sự hoạt động gần thời điểm đánh giá. Chỉ số này càng lớn càng cho thấy xu hướng rời bỏ của khách hàng càng cao.
* Frequency: Tần suất mua hàng của khách hàng. Nếu khách hàng mua càng nhiều đơn thì giá trị về doanh số mang lại cho công ty càng cao và tất nhiên giá trị của họ càng lớn.
* Monetary: Là số tiền chi tiêu của khách hàng. Đây là yếu tố trực quan nhất ảnh hưởng tới doanh số. Hay nói cách khác, doanh nghiệp quan tâm nhất là khách hàng đã dành bao nhiêu tiền để mua sắm sản phẩm của công ty? Monetary sẽ tác động trực tiếp tới doanh thu và bị tác động gián tiếp thông qua 2 yếu tố còn lại là Recency và Frequency.

Phân tích RFM giúp xác định các nhóm khách hàng quan trọng bằng cách phân loại khách hàng dựa trên hành vi và lịch sử giao dịch của họ. Thông qua việc phân tích ba chỉ số chính, doanh nghiệp có thể xác định các nhóm khách hàng mua hàng thường xuyên, chi tiêu nhiều hoặc các nhóm khách hàng có khả năng quay lại và tiếp tục mua hàng, mang lại hiệu quả tốt hơn trong việc tạo ra chiến lược tiếp thị và phục vụ khách hàng.

## Bài toán phân cụm K-means:

Phân cụm (Clustering) là một kỹ thuật thuộc phương pháp Unsupervised Learning (học không giám sát) trong khai thác dữ liệu nhằm chia tập hợp các đối tượng thành từng nhóm sao cho các đối tượng trong cùng một nhóm (hay cụm) có mức độ tương đồng cao với nhau, đồng thời khác biệt rõ rệt so với các đối tượng thuộc các nhóm khác. Mục tiêu của phân cụm là phát hiện ra cấu trúc tiềm ẩn hoặc các mẫu phân bố tự nhiên trong dữ liệu mà không cần đến thông tin gán nhãn trước đó

Các kỹ thuật Clustering được chia thành 2 nhóm là phân hoạch (Partitional clustering) và phân cụm phân cấp (Hierarchical clustering):

* **Hierarchical (Phân cấp):** Các thuật toán phân cụm phân cấp xây dựng một cấu trúc phân cấp các cụm, thường được biểu diễn dưới dạng cây (dendrogram). Có hai cách tiếp cận chính:
  + **Single Link (Liên kết đơn):** Khoảng cách giữa hai cụm được định nghĩa là khoảng cách ngắn nhất giữa bất kỳ hai điểm nào trong hai cụm đó.
  + **Complete Link (Liên kết đầy đủ):** Khoảng cách giữa hai cụm được định nghĩa là khoảng cách lớn nhất giữa bất kỳ hai điểm nào trong hai cụm đó.

→ Phân cụm phân cấp cho ra kết quả dễ hiểu, giải thuật đơn giản, không cần tham số đầu vào. Tuy nhiên phương pháp này có tốc độ chậm, không thích hợp cho các dữ liệu lớn, phức tạp. Không thể xử lý được đối với các dữ liệu bị thiếu, nhạy cảm với nhiễu.

* **Partitional (Phân hoạch):** Các thuật toán phân cụm phân hoạch chia dữ liệu thành một số lượng cụm không chồng lấp nhau. Có thể kể đến các kỹ thuật như :
  + **Square Error:** Dựa trên tiêu chí tối thiểu hóa tổng bình phương khoảng cách từ mỗi điểm dữ liệu đến tâm cụm tương ứng. K-means là thuật toán phổ biến nhất trong nhóm này
  + **Graph Theoretic** **(Lý thuyết đồ thị)**: Dữ liệu được biểu diễn dưới dạng một đồ thị, phân cụm được thực hiện bằng cách cắt đồ thị thành các phần sao cho các điểm trong cùng một phần có kết nối mạnh hơn so với các điểm ở phần khác.
  + **Mixture Models** **(GMM)**: Giả định dữ liệu là tổng hợp của nhiều phân phối xác suất. Thuật toán **Expectation-Maximization (EM)** dùng để ước lượng tham số mô hình
  + **Mode Seeking (Tìm điểm mật độ cao)**: Phân cụm dựa trên việc tìm các điểm có mật độ cao (mode) và các điểm dữ liệu khác sẽ được gán vào cụm có mode gần nhất. Thuật toán phổ biến: **Mean Shift, DBSCAN**, có khả năng phát hiện cụm với hình dạng bất kỳ

Thuật toán K-means (K-means clustering) là một thuật toán rất phổ biến trong bài toán phân cụm, được sử dụng để chia tập dữ liệu thành các nhóm (cụm) dựa trên sự tương tự hoặc liên quan giữa các điểm dữ liệu. Trong phân cụm K-means, ta không biết tính chất của từng điểm dữ liệu mà thay vào đó, chúng tìm kiếm sự tương tự dựa trên đặc trưng hoặc thuộc tính của các điểm dữ liệu và sử dụng chúng để tạo ra các một số nhóm.

Thuật toán K-means có thể tóm tắt như sau:

1. **Khởi tạo:** Bắt đầu bằng cách chọn ngẫu nhiên K điểm từ tập dữ liệu. Các điểm này sẽ hoạt động như tâm cụm ban đầu. (thường là sử dụng phương pháp Elbow để tìm K)
2. **Gán cụm:** Đối với mỗi điểm dữ liệu trong tập dữ liệu, hãy tính khoảng cách giữa điểm đó và mỗi K tâm. Gán điểm dữ liệu cho cụm có tâm gần nhất với nó. Bước này thực sự tạo thành K cụm.
3. **Cập nhật trọng tâm:** Sau khi tất cả các điểm dữ liệu đã được gán cho các cụm, hãy tính toán lại trọng tâm của các cụm bằng cách lấy giá trị trung bình của tất cả các điểm dữ liệu được gán cho mỗi cụm.
4. **Lặp lại:** Lặp lại các bước 2 và 3 cho đến khi hội tụ. Hội tụ xảy ra khi các tâm không còn thay đổi đáng kể nữa hoặc khi đạt đến số lần lặp lại được chỉ định.
5. **Kết quả cuối cùng:** Khi đạt được sự hội tụ, thuật toán sẽ đưa ra trọng tâm cụm cuối cùng và chỉ định từng điểm dữ liệu cho một cụm.

## EDA (Exploratory Data Analysis - Phân tích khám phá dữ liệu):

Phân tích khám phá dữ liệu (EDA) là một quy trình quan trọng trong phân tích dữ liệu. Ở giai đoạn này, tập dữ liệu sẽ đi qua một số kỹ thuật để hiểu rõ và diễn giải đặc điểm của tập dữ liệu trước khi tiến hành phân tích.

Các kỹ thuật phân tích trong EDA được chia thành ba nhóm chính:

1. **Phân tích đơn biến**: Tập trung vào một biến duy nhất để hiểu rõ phân bố và đặc điểm của biến đó.
   * **Histogram (Biểu đồ phân phối)**: Thể hiện tần suất của các giá trị số.
   * **Boxplot (Biểu đồ hộp)**: Thể hiện phân bố dữ liệu, trung vị, và các giá trị ngoại lệ.
   * **Bar Chart (Biểu đồ cột)**: So sánh giá trị của các nhóm dữ liệu phân loại.
   * **Pie Chart (Biểu đồ tròn)**: Hiển thị tỷ lệ phần trăm của từng danh mục trong tổng thể.
2. **Phân tích hai biến**: Phân tích mối liên hệ giữa hai biến để xác định xu hướng hoặc mức độ tương quan.
   * **Scatter Plot (Biểu đồ phân tán)**: Hiển thị mối quan hệ giữa hai biến thông qua các điểm trên mặt phẳng tọa độ.
   * **Correlation Chart (Biểu đồ tương quan)**: Thể hiện mức độ liên quan và chiều hướng (tích cực hoặc tiêu cực) giữa hai biến.
   * **Regression Plot (Biểu đồ hồi quy)**: Dùng để ước lượng mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc.
3. **Phân tích đa biến**: Phân tích đồng thời nhiều biến để nhận diện mối liên hệ phức tạp và đơn giản hóa dữ liệu.
   * **Multivariate Distribution Plot (Biểu đồ phân phối đa biến)**: Hiển thị mật độ phân phối và mối liên hệ giữa nhiều biến cùng lúc.
   * **Scatterplot Matrix (Biểu đồ phân tán ma trận)**: Trình bày toàn bộ các cặp tương quan giữa nhiều biến trong một ma trận biểu đồ.
   * **Boxplot Matrix (Biểu đồ hộp đa biến)**: So sánh sự phân phối của nhiều biến theo nhóm hoặc điều kiện khác nhau.

EDA giúp hiểu rõ cấu trúc và mối quan hệ giữa các trường dữ liệu, phát hiện và xử lý giá trị thiếu, dữ liệu lỗi, cũng như xác định mối tương quan giữa các biến. EDA còn hỗ trợ xây dựng cơ sở dữ liệu có cấu trúc rõ ràng và chuẩn bị dữ liệu sạch, phù hợp cho các bước phân tích hoặc áp dụng thuật toán học máy sau này.

## Tiền xử lý dữ liệu:

Tiền xử lý dữ liệu là quá trình xử lý dữ liệu thô hoặc ban đầu nhằm nâng cao chất lượng dữ liệu và từ đó cải thiện độ chính xác và hiệu quả của các thuật toán phân tích hoặc khai phá dữ liệu​. Dữ liệu trong thực tế thường không đầy đủ, có nhiễu, không nhất quán, có thể dư thừa hoặc không cần thiết.

Các kỹ thuật tiền xử lý dữ liệu bao gồm:

* **Làm sạch dữ liệu (data cleaning)** là quá trình loại bỏ nhiễu, xử lý dữ liệu thiếu và chỉnh sửa những điểm không nhất quán trong tập dữ liệu
* **Tích hợp dữ liệu (Data integration)** là kỹ thuật dùng để kết hợp dữ liệu từ nhiều nguồn khác nhau thành một tập dữ liệu thống nhất. Quá trình này cũng giúp phát hiện và xử lý dữ liệu trùng lặp, hoặc các thuộc tính biểu diễn khác nhau cho cùng một khái niệm.
* **Biến đổi dữ liệu (Data Transformation)** là bước biến đổi dữ liệu thô thành dạng dễ xử lý hơn, bao gồm chuẩn hóa (normalization) – đưa dữ liệu về cùng một định dạng.
* **Thu giảm dữ liệu (Data Reduction)** giúp giảm kích thước dữ liệu mà vẫn giữ lại thông tin quan trọng. Điều này có thể thực hiện thông qua việc chọn lọc thuộc tính quan trọng, loại bỏ thuộc tính dư thừa, gom cụm dữ liệu, hoặc sử dụng kỹ thuật nén và lấy mẫu.

Những kỹ thuật trên không chỉ giúp cải thiện chất lượng dữ liệu đầu vào mà còn đóng vai trò quan trọng trong việc đảm bảo độ chính xác và hiệu quả cho các thuật toán phân tích và khai phá dữ liệu.

# KẾT QUẢ MINH CHỨNG ĐƯỢC TRIỂN KHAI BẰNG PYTHON

## Xây dựng tổng thể (Framework):

### Thu thập dữ liệu: Thực hiện thu thập tập dữ liệu bán hàng trực tuyến trên nền tảng Kaggle

### Khám phá dữ liệu: Thực hiện khám phá dữ liệu để có được cái nhìn tổng quan về bộ dữ liệu đồng thời hỗ trợ phát hiện các mối quan hệ giữa các biến số, nhận diện các giá trị ngoại lai và hiểu cách thức phân bố của các giá trị

### Tiền xử lý dữ liệu:

#### Xử lý dữ liệu không nhất quán: Chuẩn hóa định dạng ngày tháng (Order Date, Ship Date) về kiểu dữ liệu ngày tháng và hiển thị theo định dạng DD/MM/YYYY

#### Tính toán giá trị RFM: Dựa trên bảng đã tổng hợp theo từng **Customer ID tính toán các giá trị RFM:**

* **Recency (R - Gần đây nhất)**: Khoảng thời gian từ lần mua gần nhất đến hiện tại
* **Frequency (F - Tần suất):** Số lần mua hàng trong khoảng thời gian nhất định
* **Monetary (M - Giá trị tiền tệ):** Tổng chi tiêu của khách hàng

#### Xử lý dữ liệu ngoại lai (outlier):

Áp dụng **phương pháp IQR** để phát hiện và loại bỏ các giá trị ngoại lai trong các chỉ số R, F, M, từ đó giúp mô hình K-means không bị nhiễu bởi các điểm quá chênh lệch. Phương pháp IQR được lựa chọn để xử lý ngoại lai vì IQR không phụ thuộc vào phân phối của dữ liệu. Trong thực tế, các giá trị R, F, M thường có phân phối lệch, đặc biệt là Frequency và Monetary bị ảnh hưởng bởi số ít khách hàng mua hàng nhiều hoặc chi tiêu lớn. Việc sử dụng IQR giúp phát hiện và loại bỏ các điểm dữ liệu nằm ngoài khoảng [Q1 - 1.5×IQR, Q3 + 1.5×IQR], từ đó giảm thiểu ảnh hưởng của các giá trị cực đoan đến quá trình phân cụm. Điều này đảm bảo dữ liệu đầu vào cho thuật toán K-means ổn định và đáng tin cậy hơn

#### Chuẩn hóa dữ liệu:

Vì các thuộc tính R, F, M có đơn vị đo lường khác nhau (ngày, số lần mua, tiền tệ), nếu không chuẩn hóa, thuật toán K-Means sẽ bị chi phối bởi những biến có giá trị lớn (đặc biệt là Monetary). Sau khi loại bỏ các giá trị ngoại lai bằng phương pháp IQR, dữ liệu RFM trở nên “sạch” hơn, nhưng vẫn tồn tại sự chênh lệch về quy mô và phân phối không chuẩn giữa các thuộc tính.

Trong trường hợp này, Z-score normalization được lựa chọn thay vì Min-Max scaling vì:

* **Z-score** đưa dữ liệu về phân phối với trung bình bằng 0 và độ lệch chuẩn bằng 1, giúp cân bằng ảnh hưởng giữa các biến và phù hợp với các thuật toán dựa trên khoảng cách như K-means.
* **Min-Max scaling** dễ bị ảnh hưởng bởi giá trị biên (ngay cả sau khi loại bỏ outlier), làm giảm khả năng phân biệt giữa các điểm dữ liệu nếu khoảng giá trị còn lại bị thu hẹp.

Do đó, lựa chọn phương pháp Z-score để xử lý dữ liệu RFM trước khi tiến hành phân cụm bằng thuật toán K-means.

### Xây dựng mô hình:

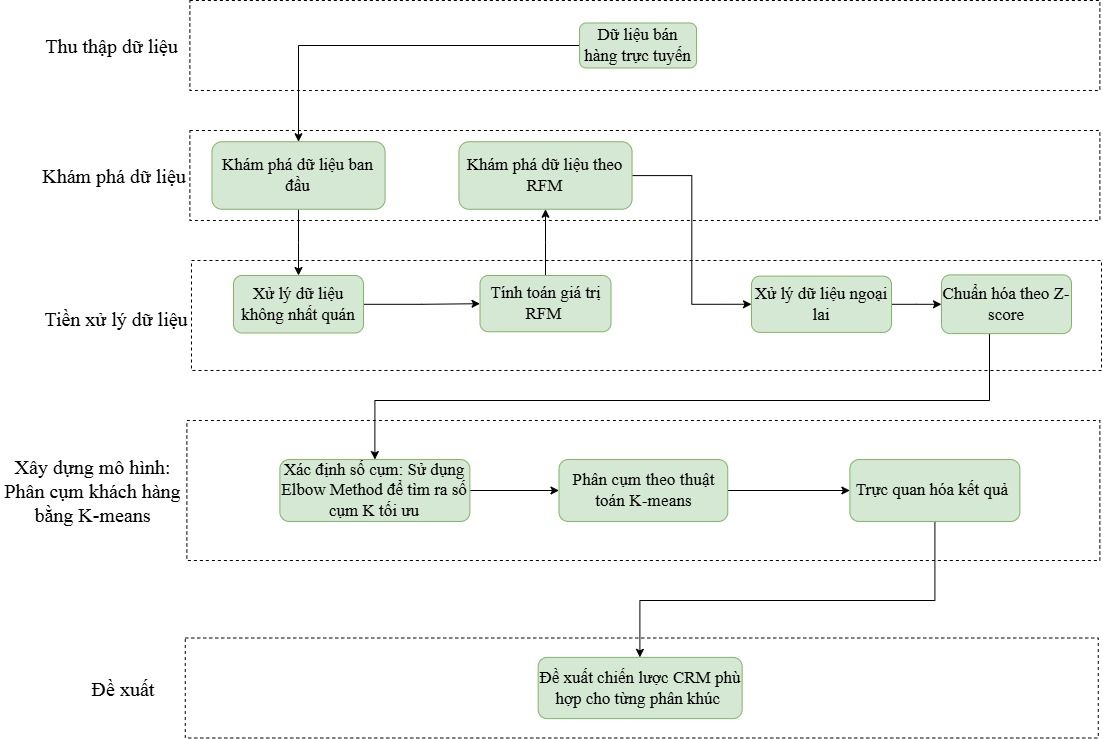
#### Xác định số cụm: Sử dụng Elbow Method để tìm ra số cụm k tối ưu: Elbow Method giúp xác định điểm mà sự cải thiện trong tổng phương sai giảm dần khi tăng số cụm

#### Phân cụm theo thuật toán K-means: Áp dụng K-means clustering trên dữ liệu chuẩn hóa RFM để phân nhóm khách hàng. Kết quả là các khách hàng sẽ được gán vào các cụm dựa trên hành vi mua sắm của họ

#### Trực quan hóa kết quả: Sử dụng các biểu đồ để trực quan hóa phân bố các chỉ số R, F, M cho từng cụm, giúp dễ dàng so sánh các phân khúc khách hàng

### Đề xuất: Đề xuất chiến lược CRM phù hợp cho từng phân khúc

Mô phỏng quy trình nghiên cứu (Framework):



Hình 1 Mô phỏng quy trình nghiên cứu (Framework)

## Chọn dữ liệu (Dataset)

### Tổng quan dữ liệu:

Là tập dữ liệu xuyên quốc gia chứa tất cả các giao dịch xảy ra từ 3/9/2016 đến 1/14/2017 cho một cửa hàng trực tuyến và có trụ sở tại United States.

Nguồn dữ liệu được lấy từ trang Kaggle: https://www.kaggle.com/datasets/tanayatipre/store-sales-forecasting-dataset

Tập dữ liệu được chuyển sang định dạng Excel: Tệp đính kèm

### Mô tả dữ liệu:

Dữ liệu gồm 2121 dòng và 21 cột, trong đó:

* **Row ID:** Mã định danh tuần tự cho mỗi hàng
* **Order ID:** Mã định danh duy nhất cho mỗi đơn đặt hàng. Có định dạng XX-YYYY-123456 trong đó: XX đại diện Mã khu vực (ký hiệu gồm 2 chữ cái); YYYY đại diện Năm đặt hàng; 123456 đại diện Mã số đơn hàng duy nhất
* **Order Date:** Ngày đặt hàng, được tổ chức theo định dạng MM/DD/YYYY
* **Ship Date:** Ngày giao hàng thực tế cho đơn hàng, được tổ chức theo định dạng MM/DD/YYYY
* **Ship Mode:** Phương thức vận chuyển cho đơn hàng.
  + Standard class (Giao hàng tiêu chuẩn): Giao hàng mất trung bình gần 5 ngày, lâu nhất trong các phương thức
  + Second class (Giao hàng tiết kiệm): Giao hàng trung bình khoảng 3 ngày
  + First class (Giao hàng nhanh): Giao hàng trong khoảng 2 ngày
  + Same day (Giao hàng trong ngày): Giao hàng gần như ngay lập tức (trong vòng 1 ngày)
* **Customer ID:** Mã định danh duy nhất cho mỗi khách hàng, có dạng XX-1234(XX-viết tắt chữ cái đầu tên khách hàng; 1234 đại diện cho mã số khách hàng duy nhất)
* **Customer Name:** Tên khách hàng
* **Segment:** Phân khúc khách hàng, bao gồm: Consumer (Người tiêu dùng), Corporate (Doanh nghiệp), Home Office (Văn phòng tại nhà)
* **Country:** Quốc gia nơi nhận hàng của khách hàng, chỉ có United States (Hoa Kỳ)
* **City:** Thành phố nơi nhận hàng của khách hàng
* **State:** Bang nơi nhận hàng của khách hàng (ví dụ: California, Texas, New York, v.v.)
* **Postal Code:** Mã bưu chính nơi nhận hàng của khách hàng
* **Region:** Khu vực nơi nhận hàng của khách hàng, trong đó: South (Nam), West (Tây), East (Đông), Central (Trung tâm)
* **Product ID:** Mã định danh duy nhất cho mỗi sản phẩm có dạng XXX-XX-123456 trong đó: XXX đại diện Loại sản phẩm (FUR, OFF, TEC...); YY đại diện Danh mục con (FU, ST, AP...), 1000xxxx đại diện Mã số sản phẩm duy nhất
* **Category:** Danh mục sản phẩm, ở đây cửa hàng chỉ có dữ liệu về danh mục Furniture (Nội thất)
* **Sub\_Category:** Phân loại sản phẩm theo danh mục phụ (chi tiết cho danh mục chính), bao gồm: Bookcases (Tủ sách), Chairs (Ghế), Tables (Bàn), Furnishings (Đồ trang trí nội thất)
* **Product Name:** Tên sản phẩm
* **Sales:** Tổng số tiền bán hàng cho đơn đặt hàng
* **Quantity:** Số lượng sản phẩm được bán trong đơn hàng
* **Discount:** Giảm giá áp dụng cho đơn hàng
* **Profit:** Lợi nhuận được tạo ra từ đơn hàng

## Triển khai bằng Python:

### Khám phá dữ liệu ban đầu:

Trong giai đoạn đầu của quá trình phân tích dữ liệu, nhóm đã thực hiện khám phá dữ liệu (EDA) nhằm hiểu rõ hơn về cấu trúc, đặc điểm và chất lượng của bộ dữ liệu thu thập được. Bộ dữ liệu được sử dụng trong đề tài là một tệp Excel có tên *“stores\_sales\_forecasting\_dataset.xlsx”*, chứa thông tin về các giao dịch bán hàng tại một chuỗi cửa hàng:

#### Đọc dữ liệu:

#### Đoạn mã sau được sử dụng để đọc dữ liệu từ tệp Excel:

#1. EDA - Khám phá dữ liệu

import pandas as pd

file\_path = "C:\\Users\\hp\\OneDrive - The Danang University of Economics\\HocTap\\NĂM 3\\KÌ 2\\PHÂN TÍCH DỮ LIỆU BẰNG PYTHON\\BT nhóm\\NHÓM 5\_48K21.2\\stores\_sales\_forecasting\_dataset.xlsx"

# Đọc dữ liệu

df = pd.read\_excel(file\_path)

#### Tìm hiểu tổng quan dữ liệu:

Để hiểu rõ hơn về cấu trúc và đặc điểm của bộ dữ liệu, nhóm đã thực hiện tìm hiểu một số thông tin tổng quan bao gồm:

#Tìm hiểu về dữ liệu

#Tìm hiểu các thông tin cơ bản về dữ liệu

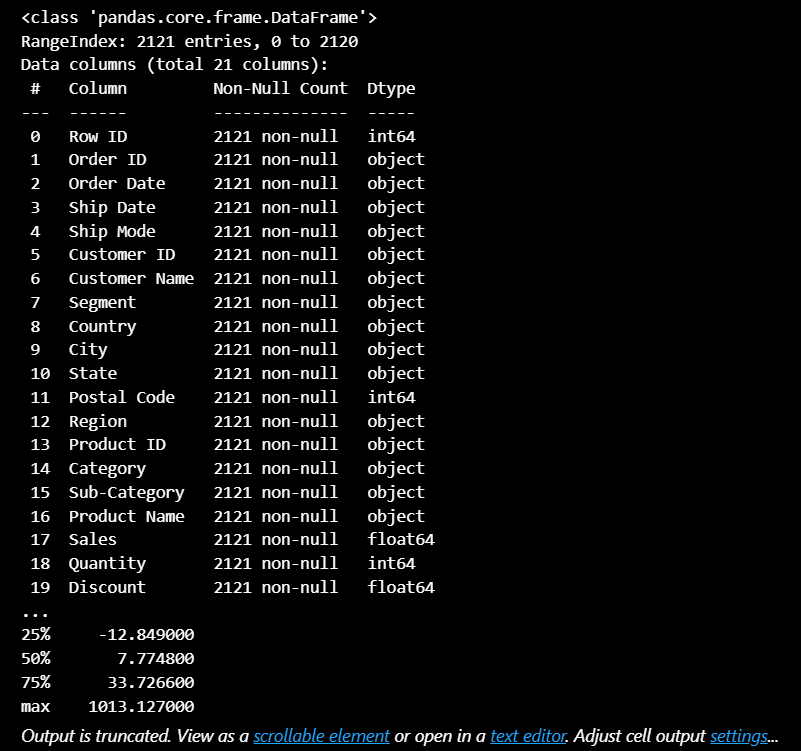
print(df.info())

print(df.shape)

print(df.head())

print(df.describe())

Thu được kết quả:



Hình 2 Kết quả của việc tìm hiểu tổng quan bộ dữ liệu

Từ kết quả nhóm nhận thấy được rằng: Bộ dữ liệu có tổng cộng 2121 dòng và 21 cột, với đầy đủ giá trị không bị thiếu trong bất kỳ cột nào. Các kiểu dữ liệu bao gồm:

* int64: Các cột định lượng như Row ID, Postal Code, Quantity
* float64: Các cột có giá trị thực như Sales, Discount, Profit
* object: Các cột chứa chuỗi như Order ID, Customer ID, City, Product Name,…
* Order Date và Ship Date hiện tại đang ở dạng object, cần được chuyển sang kiểu datetime để phục vụ xử lý thời gian trong bước tính Recency

#### Kiểm tra các giá trị trùng lặp:

Một trong những bước quan trọng trong quá trình làm sạch dữ liệu là kiểm tra các dòng bị trùng lặp hoàn toàn, nhằm đảm bảo chất lượng dữ liệu đầu vào cho mô hình phân tích.

Nhóm đã sử dụng đoạn mã sau để phát hiện các dòng trùng lặp trong bộ dữ liệu:

#Kiểm tra dữ liệu trùng lặp

duplicate\_rows = df[df.duplicated()]

num\_duplicates = len(duplicate\_rows)

if num\_duplicates > 0:

    print("Có", num\_duplicates, "dòng bị trùng lặp:")

    print(duplicate\_rows)

else:

    print("Không có dòng dữ liệu trùng lặp")

Cho ra được kết quả không có dòng dữ liệu trùng lặp trong bộ dữ liệu:

https://lh7-rt.googleusercontent.com/docsz/AD_4nXeMb5ECXa54RiSOuBua5u4uL5DUqSWLjreWKt3i2G7OLmf4UuqjPj7XYva5y8JKIL5i4JCbTdhfxh0pa94mxLPapR1bLdvcapF9K2xNtHxlsHFDYjtYFOCNlq1e15w_nIubq5_4fg?key=WgFyMmuxBNdgN9oFYsQ0unEC

Hình 3 Kết quả kiểm tra các giá trị trùng lặp

#### Kiểm tra các giá trị nhiễu:

Tiếp theo, chúng ta kiểm tra các giá trị nhiễu để giúp tăng độ tin cậy của dữ liệu. Nếu phát hiện các giá trị bất thường, chúng ta sẽ cần có các bước xử lý phù hợp để đảm bảo chất lượng cho quá trình phân tích sau này.

#Kiểm tra các giá trị nhiễu

# 1. Kiểm tra số lượng âm

if (df['Quantity'] < 0).any():

    print("Có dòng có số lượng âm.")

    print(df[df['Quantity'] < 0])

else:

    print("Không có dòng nào có số lượng âm")

# 2. Kiểm tra doanh thu âm

if (df['Sales'] < 0).any():

    print("Có dòng có giá trị doanh thu âm")

    print(df[df['Sales'] < 0])

else:

    print("Không có dòng nào có doanh thu âm")

# 3. Kiểm tra cột Discount

# a. Kiểm tra Discount âm

if (df['Discount'] < 0).any():

    print("Có dòng có Discount âm:")

    print(df[df['Discount'] < 0])

else:

    print("Không có dòng nào có Discount âm")

# b. Kiểm tra Discount > 1 (tức là > 100%)

if (df['Discount'] > 1).any():

    print("Có dòng có Discount lớn hơn 100%:")

    print(df[df['Discount'] > 1])

else:

    print("Không có dòng nào có Discount lớn hơn 100%")

# 4. Kiểm tra ngày giao hàng sớm hơn ngày đặt hàng

# Chuyển cột ngày sang kiểu datetime

df['Order Date'] = pd.to\_datetime(df['Order Date'])

df['Ship Date'] = pd.to\_datetime(df['Ship Date'])

# Kiểm tra ngày giao hàng sớm hơn ngày đặt hàng

condition = df['Ship Date'] < df['Order Date']

if condition.any():

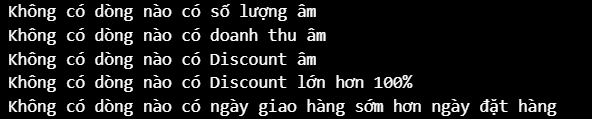
    print("Có dòng có ngày giao hàng sớm hơn ngày đặt hàng:")

    print(df[condition])

else:

    print("Không có dòng nào có ngày giao hàng sớm hơn ngày đặt hàng")

Kết quả cho thấy được:



Hình 4 Kết quả kiểm tra các giá trị nhiễu

#### Biểu diễn dữ liệu gốc:

* 1. Biểu đồ cột (Bar chart):

Để trực quan hóa và hiểu rõ hơn về các thuộc tính phân loại trong bộ dữ liệu, nhóm đã sử dụng biểu đồ **cột** nhằm biểu diễn **tần suất xuất hiện** của từng giá trị trong các cột có kiểu dữ liệu object:

Đoạn code phục vụ:

# Biểu diễn dữ liệu gốc

# Vẽ biểu đồ cột

import matplotlib.pyplot as plt

import seaborn as sns

import math

# Lấy danh sách các cột kiểu object

object\_cols = df.select\_dtypes(include='object').columns

num\_plots = len(object\_cols)

# Đếm số lượng cột kiểu object

print("Số lượng cột kiểu object:", num\_plots)

# Tính số hàng và cột phù hợp

cols = 2

rows = math.ceil(num\_plots / cols)

# Vẽ biểu đồ cho từng cột kiểu object

plt.figure(figsize=(15, 5 \* rows))

for i, col in enumerate(object\_cols, 1):

    plt.subplot(rows, cols, i)

    plt.xticks(rotation=65, size=10)

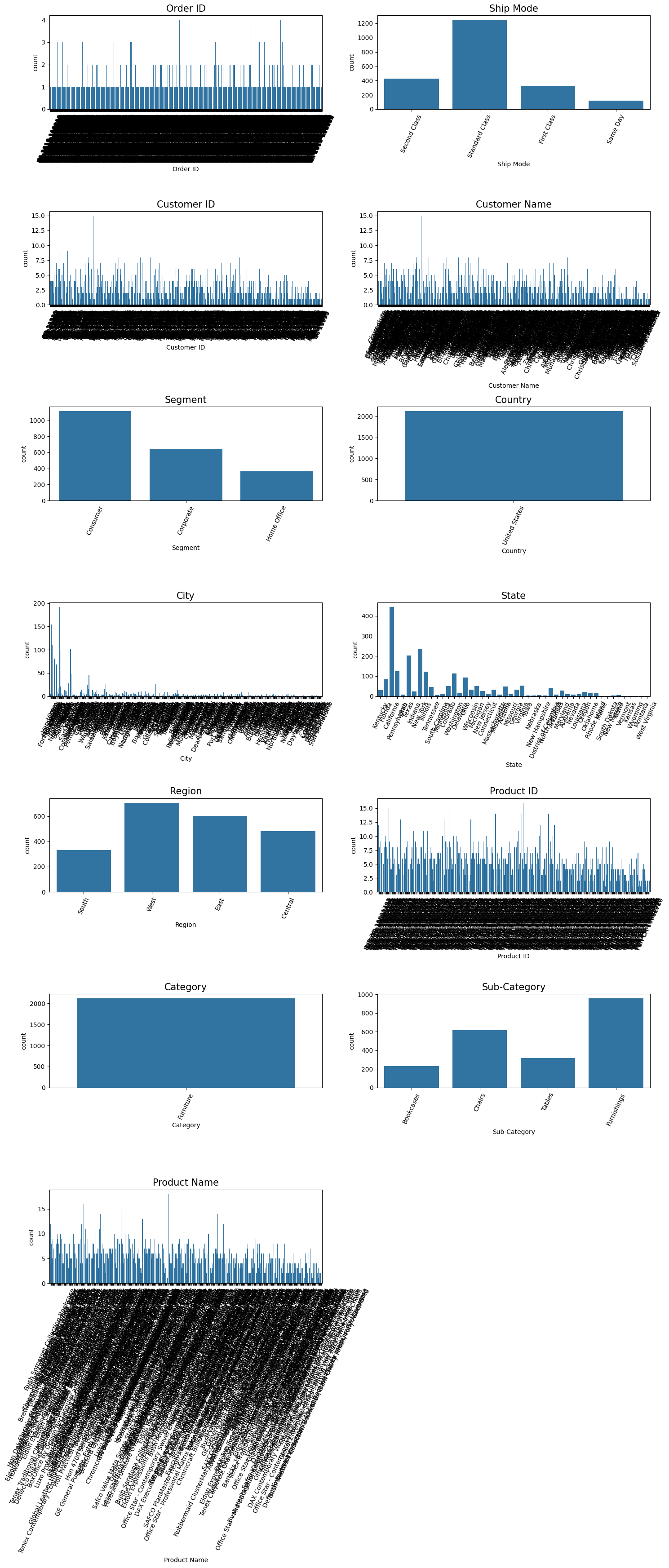
    sns.countplot(data=df, x=col)

    plt.title(col, size=15)

plt.tight\_layout()

plt.show()

Kết quả:



Hình 5 Biểu đồ cột biểu diễn tần suất xuất hiện của từng giá trị trong các cột có kiểu dữ liệu object

Qua quan sát, có thể thấy phần lớn các cột đều có sự phân bố giá trị khá đa dạng. Tuy nhiên, một số cột như **Country** và **Category** chỉ chứa một giá trị duy nhất, cho thấy không có sự khác biệt về quốc gia và danh mục sản phẩm ở cấp độ tổng thể trong bộ dữ liệu. Các cột còn lại như Ship Mode, Segment, Region, State, City, Product Name,... thể hiện sự đa dạng rõ rệt, phản ánh tính phong phú của hành vi và đặc điểm khách hàng.

* 1. Biểu đồ tròn (Pie chart):

Biểu đồ tròn được sử dụng để thể hiện **tỷ lệ phân bố** của các giá trị giúp khám phá, hiểu rõ đặc điểm hành vi mua sắm ban đầu của khách hàng để định hướng cho việc xây dựng chiến lược CRM phù hợp với từng nhóm khách hàng

Đoạn code:

#Vẽ biểu đồ tròn cho cột 'Segment'

import matplotlib.pyplot as plt

# Đếm số lượng khách hàng theo phân khúc (Segment)

segment\_counts = df['Segment'].value\_counts()

# Vẽ biểu đồ

plt.figure(figsize=(6, 6))

plt.pie(segment\_counts, labels=segment\_counts.index, autopct='%1.1f%%', startangle=90)

plt.title('Phân khúc khách hàng theo hành vi mua sắm (Segment)')

plt.axis('equal')

plt.show()

#Vẽ biểu đồ tròn cho cột 'Ship Mode'

# Đếm số lượng khách hàng theo phương thức giao hàng (Ship Mode)

ship\_mode\_counts = df['Ship Mode'].value\_counts()

# Vẽ biểu đồ

plt.figure(figsize=(6, 6))

plt.pie(ship\_mode\_counts, labels=ship\_mode\_counts.index, autopct='%1.1f%%', startangle=90)

plt.title('Phương thức giao hàng (Ship Mode)')

plt.axis('equal')

plt.show()

#Vẽ biểu đồ tròn cho cột 'Region'

# Đếm số lượng khách hàng theo vùng miền (Region)

region\_counts = df['Region'].value\_counts()

# Vẽ biểu đồ

plt.figure(figsize=(6, 6))

plt.pie(region\_counts, labels=region\_counts.index, autopct='%1.1f%%', startangle=90)

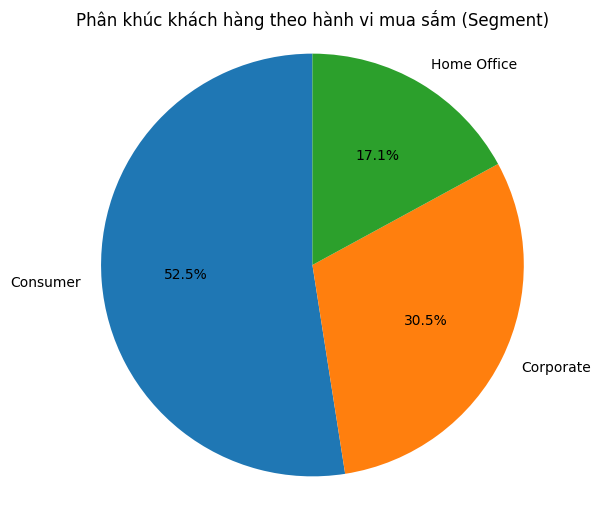
plt.title('Vùng miền (Region)')

plt.axis('equal')

plt.show()

Kết quả:

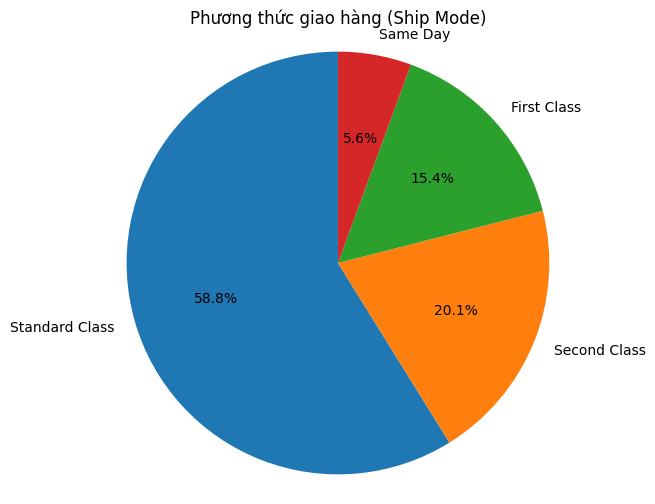
* + 1. **Phân khúc khách hàng theo hành vi mua sắm (Segment):**



Hình 6 Biểu đồ tròn thể hiện tỷ lệ phân bố của các giá trị - Phân khúc khách hàng theo hành vi mua sắm (Segment)

Nhận xét: Consumer (Tiêu dùng cá nhân) chiếm tỷ trọng cao nhất (52.5%), tiếp theo là Corporate (Doanh nghiệp) (30.5%) và ít nhất là Home Office (17.1%). Từ đó cho thấy chiến lược CRM nên tập trung chính vào nhóm khách hàng cá nhân, do đây là nhóm lớn nhất

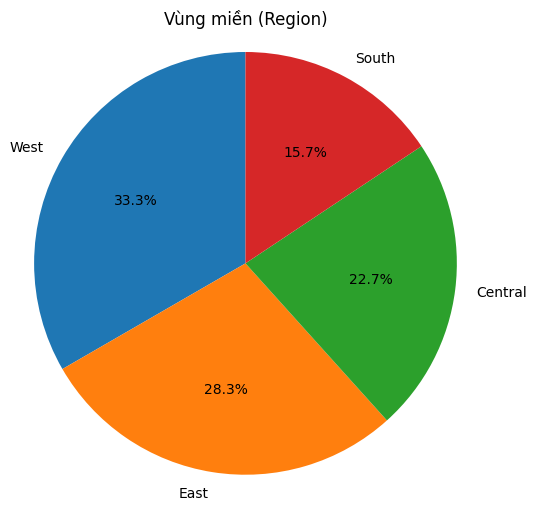
* + 1. **Phương thức giao hàng (Ship Mode):**



Hình 7 Biểu đồ tròn thể hiện tỷ lệ phân bố của các giá trị - Phương thức giao hàng (Ship Mode)

Nhận xét: Standard Class được sử dụng nhiều nhất (58.8%), sau đó là Second Class (20.1%) và First Class (15.4%). Same Day rất ít (5.6%). Biểu đồ cho thấy phần lớn khách hàng ưu tiên hình thức giao hàng tiết kiệm chi phí, ít quan tâm đến tốc độ

* + 1. **Vùng miền (Region):**



Hình 8 Biểu đồ tròn thể hiện tỷ lệ phân bố của các giá trị - Vùng miền (Region)

Nhận xét: West là vùng có nhiều giao dịch nhất (33.3%), theo sau là East (28.3%), Central (22.7%) và ít nhất là South (15.7%). Dữ liệu có sự phân bố khu vực không đồng đều, có thể ảnh hưởng đến nhu cầu và hành vi mua sắm theo địa lý.

### **Tiền xử lý dữ liệu**

#### Loại bỏ cột không cần thiết:

Thông qua biểu đồ cột trong bước Khám phá dữ liệu, nhóm nhận thấy cột Country và Category chỉ chứa một giá trị duy nhất, không đóng góp vào việc phân biệt khách hàng đồng thời cột Row ID chỉ là số thứ tự dòng, cũng không mang giá trị phân tích nên nhóm thực hiện loại bỏ các cột này nhằm mục đích **tập trung phân tích vào các thuộc tính có giá trị**, tránh làm nhiễu mô hình và tăng hiệu suất xử lý cho mô hình

# Tiền xử lý dữ liệu

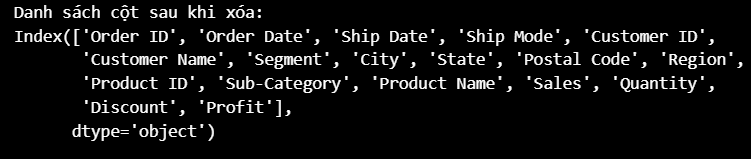
# Xóa cột: RowID, Country, Category

df.drop(['Row ID', 'Country', 'Category'], axis=1, inplace=True)

print("Danh sách cột sau khi xóa:")

print(df.columns)

Kết quả in ra các cột sau khi thực hiện xóa:



Hình 9 Kết quả của việc loại bỏ các cột không cần thiết

#### Xử lý dữ liệu không nhất quán:

Trong tập dữ liệu gốc, hai cột *Order Date* và *Ship Date* đang có kiểu dữ liệu dạng chuỗi (*object*). Để thuận tiện cho việc phân tích, đặc biệt là khi tính chỉ số **Recency** trong mô hình RFM (khoảng thời gian kể từ lần mua hàng gần nhất), nhóm tiến hành chuyển đổi hai cột này sang kiểu dữ liệu **ngày – giờ (***datetime***)** theo định dạng gốc là *MM/DD/YYYY*. Sau khi chuyển đổi, dữ liệu ngày được hiển thị lại ở định dạng *DD/MM/YYYY* để dễ quan sát.

# Xử lý dữ liệu không nhất quán (chuyển về định dạng datetime)

# Chuyển các cột ngày về kiểu datetime theo định dạng gốc MM/DD/YYYY

df['Order Date'] = pd.to\_datetime(df['Order Date'], errors='coerce', dayfirst=False)

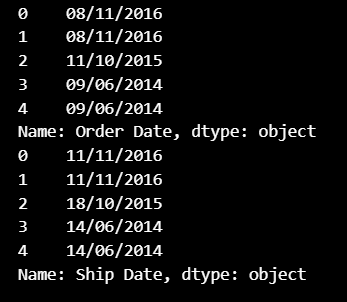
df['Ship Date'] = pd.to\_datetime(df['Ship Date'], errors='coerce', dayfirst=False)

# Hiển thị định dạng ngày dưới dạng dd/mm/yyyy

print(df['Order Date'].dt.strftime('%d/%m/%Y').head())

print(df['Ship Date'].dt.strftime('%d/%m/%Y').head())

Kết quả:



Hình 10 Kết quả xử lý dữ liệu không nhất quán

#### Tính toán giá trị RFM:

Sau khi đã xử lý dữ liệu ngày tháng, nhóm tiến hành tính toán bộ chỉ số RFM – một kỹ thuật phổ biến trong phân tích hành vi khách hàng, giúp đo lường giá trị và mức độ tương tác của từng khách hàng với doanh nghiệp. Cụ thể:

* **Recency (R):** Khoảng thời gian (tính bằng ngày) kể từ lần mua hàng gần nhất đến thời điểm hiện tại.
* **Frequency (F):** Số lần khách hàng đã thực hiện mua hàng.
* **Monetary (M):** Tổng giá trị đơn hàng (doanh thu) mà khách hàng đã chi tiêu

Để tính toán các chỉ số này, nhóm đã sử dụng đoạn mã sau để gộp nhóm theo **Customer ID** và tính toán giá trị **Recency**, **Frequency** và **Monetary** cho từng khách hàng:

# Tính toán giá trị RFM

from datetime import datetime

# Lấy ngày tính toán là ngày hiện tại

tt\_date = pd.to\_datetime(datetime.today())

# Gộp nhóm theo CustomerID

rfm = df.groupby('Customer ID').agg({

    'Order Date': lambda x: (tt\_date - x.max()).days,  # Recency

    'Customer ID': 'count',                             # Frequency

    'Sales': 'sum'})                                      # Monetary

# Đổi tên cột

rfm.columns = ['Recency', 'Frequency', 'Monetary']

df = df.merge(rfm, on='Customer ID', how='left')

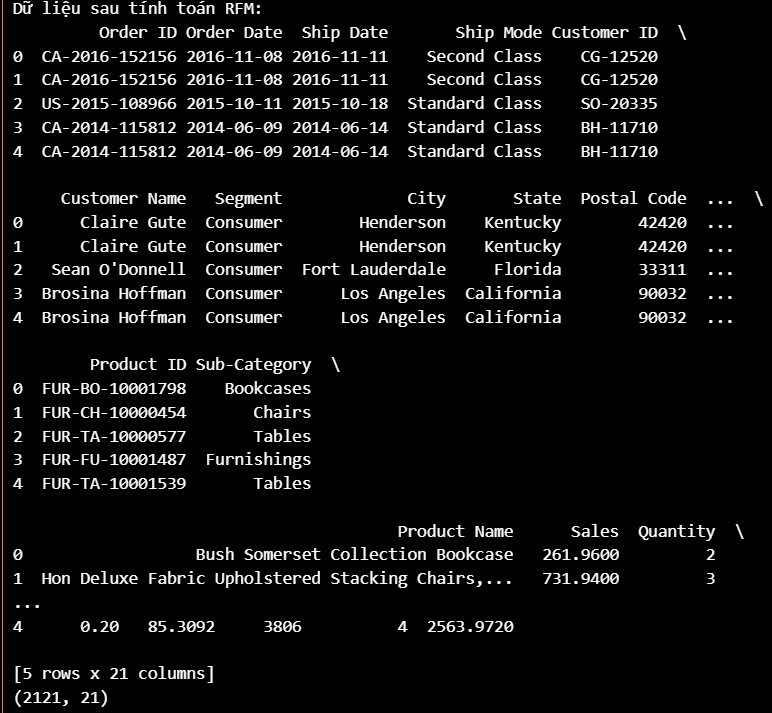
# Hiển thị kết quả

print("Dữ liệu sau tính toán RFM:")

print(df.head())

print(df.shape)

Sau khi thực hiện, dữ liệu thu được sẽ bao gồm ba chỉ số R, F, M, cung cấp thông tin đầy đủ về mức độ tương tác và giá trị của từng khách hàng. Đây sẽ là cơ sở quan trọng cho bước phân tích và phân khúc khách hàng trong chiến lược CRM tiếp theo



Hình 11 Kết quả tính toán các giá trị R,F,M

#### **Phân tích đặc trưng hành vi:**

* 1. Biểu đồ Histogram:

Để hiểu rõ hơn về đặc trưng hành vi mua sắm của khách hàng, nhóm đã sử dụng biểu đồ Histogram để phân tích sự phân bố của các chỉ số **Recency**, **Frequency**, và **Monetary**. Mỗi biểu đồ này giúp nhóm hình dung rõ hơn về đặc điểm và xu hướng của khách hàng trong ba khía cạnh quan trọng:

* **Recency**: Phân bố khoảng thời gian kể từ lần mua hàng gần nhất đến thời điểm hiện tại. Việc này giúp xác định liệu khách hàng có đang duy trì sự quan tâm và tương tác với doanh nghiệp hay không.
* **Frequency**: Phân bố số lần mua hàng của khách hàng, giúp đánh giá mức độ trung thành và sự tái mua hàng.
* **Monetary**: Phân bố tổng số tiền khách hàng chi tiêu, cung cấp cái nhìn về giá trị khách hàng đối với doanh nghiệp.

Nhóm đã thực hiện việc vẽ biểu đồ histogram cho ba chỉ số này, qua đó cung cấp cái nhìn sâu sắc về phân bố dữ liệu của từng nhóm khách hàng

#Phân tích đặc trưng hành vi

# Vẽ biểu đồ Histogram cho cột Recency, Frequency, Monetary

import matplotlib.pyplot as plt

plt.figure(figsize=(18, 5))

# Recency

plt.subplot(1, 3, 1)

plt.hist(df['Recency'], bins=30, color='skyblue', edgecolor='black')

plt.title('Phân bố Recency')

plt.xlabel('Recency (days)')

plt.ylabel('Số lượng KH')

# Frequency

plt.subplot(1, 3, 2)

plt.hist(df['Frequency'], bins=30, color='lightgreen', edgecolor='black')

plt.title('Phân bố Frequency')

plt.xlabel('Số lần mua')

# Monetary

plt.subplot(1, 3, 3)

plt.hist(df['Monetary'], bins=30, color='salmon', edgecolor='black')

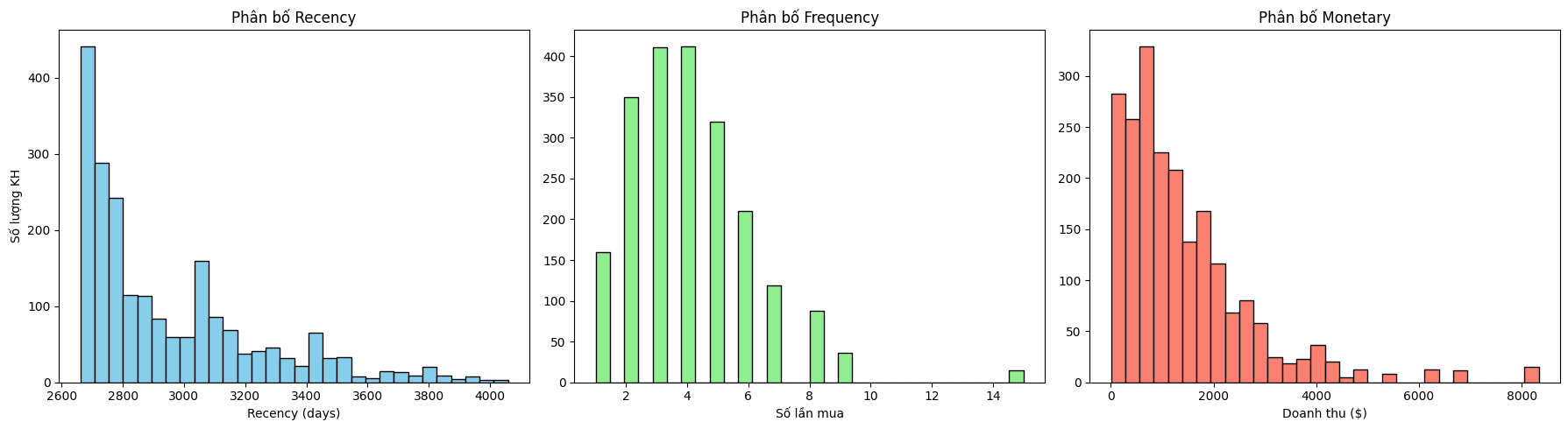
plt.title('Phân bố Monetary')

plt.xlabel('Doanh thu ($)')

plt.tight\_layout()

plt.show()

Kết quả từ các biểu đồ histogram này giúp nhóm có cái nhìn rõ ràng về phân bổ hành vi khách hàng trong ba chỉ số quan trọng. Điều này có ý nghĩa quan trọng trong việc phân loại khách hàng, phát hiện nhóm khách hàng có mức độ tương tác cao, và từ đó xây dựng các chiến lược CRM hiệu quả hơn.



Hình 12 Biểu đồ Histogram thể hiện đặc trưng hành vi mua sắm của khách hàng theo phân bố của các chỉ số R, F, M

Nhóm nhận thấy rằng:

* Biểu đồ histogram của chỉ số Recency cho thấy phần lớn khách hàng có thời gian mua hàng gần đây, tập trung nhiều ở khoảng từ 0 đến 3000 ngày. Số lượng khách hàng giảm dần khi giá trị Recency tăng, phản ánh rằng doanh nghiệp có tệp khách hàng chủ yếu tương tác gần đây và một nhóm nhỏ đã lâu không quay lại. Việc nhận diện rõ đặc điểm phân bố lệch phải này giúp hình dung được sự chênh lệch trong hành vi mua sắm theo thời gian, từ đó tạo cơ sở để phân tách hiệu quả giữa nhóm khách hàng mới – trung bình – không còn tương tác trong quá trình phân cụm K-Means
* Biểu đồ Frequency thể hiện rằng đa số khách hàng chỉ mua hàng từ 2 đến 5 lần, trong khi số lượng khách hàng có tần suất mua hàng cao hơn (trên 7 lần) là rất ít. Phân bố lệch phải này cho thấy sự không đồng đều rõ rệt trong mức độ trung thành của khách hàng. Thông tin này có ý nghĩa quan trọng trong việc định hướng phân cụm, vì có thể phân biệt được nhóm khách hàng chỉ mua một vài lần với nhóm có hành vi mua lặp lại nhiều – một yếu tố then chốt trong chiến lược chăm sóc khách hàng và giữ chân họ
* Biểu đồ Monetary cho thấy phần lớn khách hàng mang lại doanh thu ở mức thấp, trong khi một số ít khách hàng tạo ra giá trị rất cao. Sự phân bố lệch phải thể hiện tính không cân xứng trong giá trị mà từng nhóm khách hàng đóng góp cho doanh nghiệp. Nhận biết đặc điểm này từ sớm giúp xác định được sự tồn tại của nhóm khách hàng VIP – nhóm mục tiêu cần được chăm sóc chuyên sâu trong chiến lược CRM. Đồng thời, đặc trưng này cũng là một yếu tố hỗ trợ phân tách rõ ràng các cụm khách hàng theo khả năng chi tiêu khi áp dụng thuật toán K-Means.
  1. Biểu đồ Box plot:

Để kiểm tra sự tồn tại của các giá trị ngoại lệ trong ba đặc trưng Recency, Frequency và Monetary, nhóm đã sử dụng biểu đồ Boxplot. Đây là một công cụ trực quan giúp thể hiện sự phân bố dữ liệu, khoảng tứ phân vị (IQR) và dễ dàng nhận diện các điểm dữ liệu nằm ngoài phạm vi bình thường – hay còn gọi là giá trị ngoại lệ (outliers).

* Biểu đồ Boxplot của Recency cho biết có tồn tại một số khách hàng có thời gian không quay lại rất dài, thể hiện nhóm khách hàng tiềm ẩn rủi ro rời bỏ.
* Biểu đồ Boxplot của Frequency chỉ ra có sự chênh lệch lớn giữa nhóm khách mua hàng ít và nhóm mua hàng thường xuyên – đây là cơ sở để phân nhóm khách hàng thân thiết.
* Biểu đồ Boxplot của Monetary cũng thể hiện sự phân tán lớn, cho thấy một số khách hàng mang lại giá trị doanh thu vượt trội, đáng chú ý trong chiến lược chăm sóc đặc biệt.

Nhóm đã sử dụng đoạn mã sau để trực quan hóa và kiểm tra các đặc trưng:

# Vẽ biểu đồ Boxplot cho cột Recency, Frequency, Monetary

plt.figure(figsize=(18, 5))

plt.subplot(1, 3, 1)

plt.boxplot(df['Recency'], vert=False)

plt.title('Boxplot Recency')

plt.xlabel('Recency (ngày)')

plt.subplot(1, 3, 2)

plt.boxplot(df['Frequency'], vert=False)

plt.title('Boxplot Frequency')

plt.xlabel('Số lần mua')

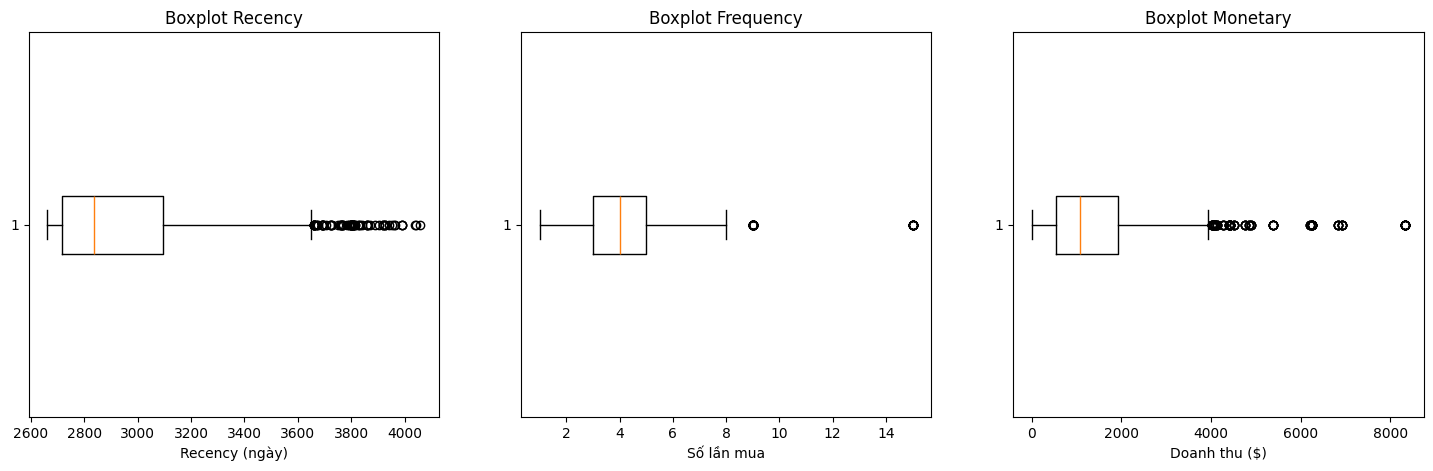
plt.subplot(1, 3, 3)

plt.boxplot(df['Monetary'], vert=False)

plt.title('Boxplot Monetary')

plt.xlabel('Doanh thu ($)')

Thu được kết quả như sau:



Hình 13 Biểu đồ Box plot để kiểm tra sự tồn tại của các giá trị ngoại lệ trong ba đặc trưng R, F, M

Biểu đồ boxplot cho ba chỉ số RFM cho thấy rõ sự tồn tại của nhiều giá trị ngoại lệ, đặc biệt ở các chỉ số Recency và Monetary. Cụ thể, Recency có nhiều điểm nằm ngoài khoảng giá trị phổ biến, phản ánh sự chênh lệch lớn về thời gian tương tác gần nhất giữa các khách hàng. Frequency phân bố khá tập trung, nhưng vẫn xuất hiện một số khách hàng có tần suất mua cao đột biến. Monetary cho thấy sự phân tán mạnh, với nhiều khách hàng chi tiêu vượt xa phần lớn còn lại. Việc nhận diện các ngoại lệ và mức độ phân tán như vậy giúp hiểu sâu hơn về đặc điểm dữ liệu, từ đó hỗ trợ tiền xử lý phù hợp và đảm bảo mô hình phân cụm K-Means hoạt động hiệu quả, không bị ảnh hưởng bởi các giá trị cực đoan

* 1. Biểu đồ phân tán:

Để hiểu rõ hơn mối quan hệ giữa các đặc trưng Recency, Frequency và Monetary, nhóm đã sử dụng biểu đồ phân tán (scatter plot), trong đó:

* Trục hoành biểu diễn **Recency** – số ngày kể từ lần mua hàng gần nhất,
* Trục tung biểu diễn **Frequency** – số lần mua hàng của khách,
* Kích thước và màu sắc của từng điểm thể hiện giá trị **Monetary** – tổng doanh thu mà khách hàng mang lại.

Biểu đồ phân tán giúp dễ dàng nhận ra các cụm khách hàng có đặc điểm hành vi tương đồng. Nhóm đã sử dụng đoạn mã sau để tạo biểu đồ:

#Vẽ biểu đồ phân tán (Scatter plot) cho Recency và Frequency, Monetary

import matplotlib.pyplot as plt

import seaborn as sns

# Thiết lập kích thước và font

plt.figure(figsize=(15, 8))

plt.xticks(rotation=65, size=10)

# Vẽ biểu đồ scatter: màu sắc và kích thước dựa trên Monetary

plot = plt.scatter(

    df['Recency'],

    df['Frequency'],

    c=df['Monetary'],

    s=df['Monetary'] / 10,

    alpha=0.5,

    cmap='Reds'

)

plt.colorbar(plot, label='Monetary ($)')

# Nhãn và tiêu đề

plt.xlabel('Recency (Ngày gần nhất mua hàng)', size=14)

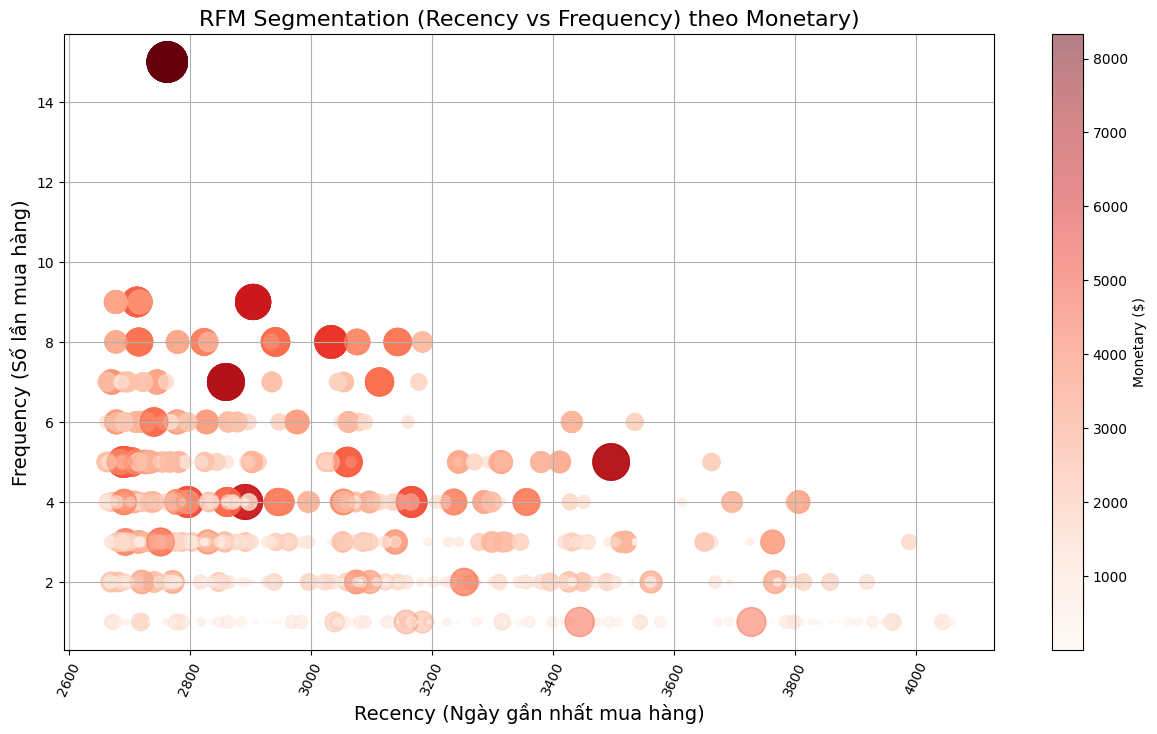
plt.ylabel('Frequency (Số lần mua hàng)', size=14)

plt.title('RFM Segmentation (Recency vs Frequency) theo Monetary)', size=16)

plt.grid(True)

plt.show()

Thu được kết quả:



Hình 14 Biểu đồ phân tán thể hiện các cụm khách hàng có đặc điểm hành vi tương đồng

Cụ thể, các khách hàng có Recency thấp (tức mua gần đây), Frequency cao (mua nhiều lần) và Monetary lớn (bong bóng to, màu đậm) tập trung rõ rệt ở góc dưới bên trái. Đây là những khách hàng có giá trị cao và hoạt động tích cực, rất tiềm năng để duy trì và tăng cường chăm sóc trong chiến lược CRM.

Ngược lại, ở vùng Recency cao (mua từ rất lâu), đa số các bong bóng có kích thước nhỏ và màu nhạt, phản ánh nhóm khách hàng ít tương tác, chi tiêu thấp – có thể cần xem xét các chiến dịch kích hoạt lại hoặc loại bỏ khỏi đối tượng ưu tiên.

Sự phân tán rõ ràng của dữ liệu theo ba chiều R, F, M trong biểu đồ này góp phần hỗ trợ trực quan trong việc phân khúc khách hàng nhằm xây dựng chiến lược CRM phù hợp.

#### Phân tích tương quan:

Để khám phá mối quan hệ giữa các biến định lượng trong tập dữ liệu, nhóm đã sử dụng biểu đồ Heatmap nhằm biểu diễn trực quan ma trận tương quan. Biểu đồ này giúp dễ dàng xác định được mức độ tương quan giữa các đặc trưng như Recency, Frequency, Monetary, Sales, Quantity hay Discount.

Cụ thể:

* Màu sắc của biểu đồ thể hiện mức độ tương quan: đỏ thể hiện tương quan dương mạnh, xanh thể hiện tương quan âm.
* Các ô được ghi chú bằng hệ số tương quan Pearson, cho thấy mức độ tuyến tính giữa các cặp biến.

Nhóm đã sử dụng đoạn mã sau để thực hiện biểu đồ:

#Vẽ biểu đồ Heatmap

# Bước 1: Tính toán ma trận tương quan (chỉ lấy các cột số)

corr = df.select\_dtypes(include='number').drop(columns=['Postal Code'], errors='ignore').corr()

# Bước 2: Vẽ heatmap tương quan

plt.figure(figsize=(10, 10))

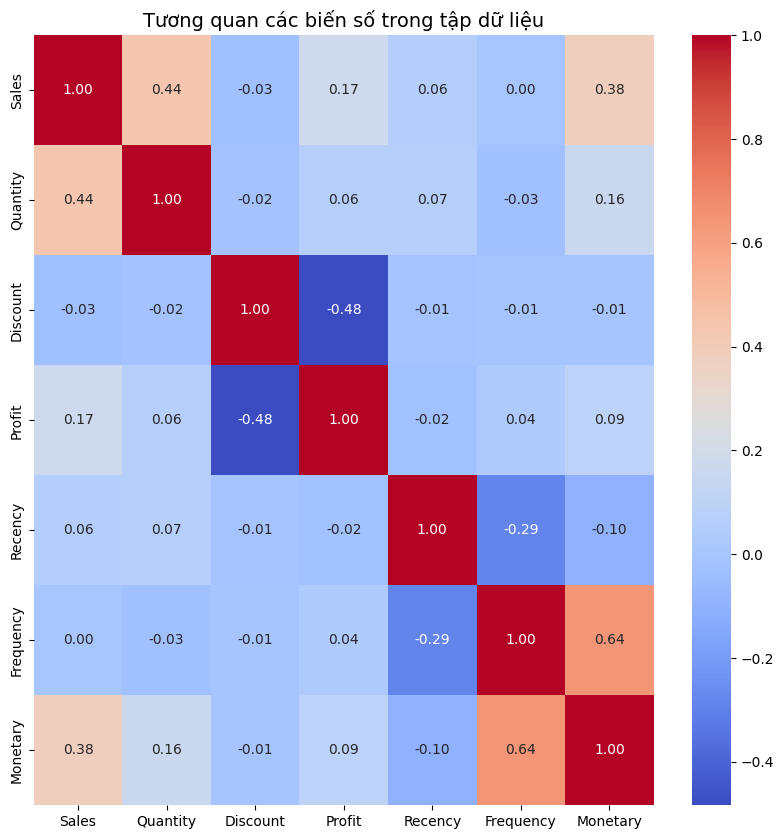
plt.title('Tương quan các biến số trong tập dữ liệu', fontsize=14)

sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")

plt.show()

Thông qua Heatmap, nhóm có thể nhận ra rằng:

* Monetary có tương quan dương khá cao với Sales (vì bản chất đều liên quan đến doanh thu),
* Discount có tương quan âm nhẹ với Sales, cho thấy việc giảm giá có thể ảnh hưởng đến doanh thu nhưng không quá rõ rệt,
* Recency và Frequency có tương quan thấp, chứng minh rằng thời gian mua gần nhất không hẳn ảnh hưởng nhiều đến tần suất mua hàng



Hình 15 Biểu đồ Heatmap biểu diễn trực quan ma trận tương quan

Từ đây, nhóm nhận thấy tâm điểm nằm ở khu vực tương quan giữa Frequency và Monetaryvới giá trị 0.64. Điều này cho thấy có mối quan hệ tích cực và khá mạnh giữa tần suất mua hàng và giá trị tiền tệ mà khách hàng chi tiêu → Khách hàng mua hàng thường xuyên hơn cũng có xu hướng chi tiêu nhiều tiền hơn

#### Xử lý các giá trị ngoại lai:

Trong quá trình tiền xử lý dữ liệu, việc phát hiện và loại bỏ các giá trị ngoại lai là một bước quan trọng nhằm đảm bảo tính chính xác cho các phân tích tiếp theo, đặc biệt là các thuật toán phân cụm như K-Means vốn nhạy cảm với dữ liệu nhiễu.

Nhóm đã sử dụng phương pháp IQR (Interquartile Range – khoảng tứ phân vị) để phát hiện các giá trị ngoại lai trong 3 đặc trưng quan trọng: Recency, Frequency và Monetary. Cụ thể:

* Tính Q1 (tứ phân vị thứ nhất), Q3 (tứ phân vị thứ ba) và IQR = Q3 - Q1.
* Xác định ngưỡng dưới và ngưỡng trên:
* Lower bound = Q1 - 1.5 \* IQR
* Upper bound = Q3 + 1.5 \* IQR
* Các giá trị nằm ngoài khoảng này được xem là ngoại lai.

Sau khi phát hiện, nhóm tiến hành loại bỏ những dòng dữ liệu có ít nhất một giá trị ngoại lai ở một trong ba đặc trưng kể trên để đảm bảo tính đại diện cho tập khách hàng trong phân tích sau này.

#Xử lý các giá trị ngoại lai

# Dùng IQR để phát hiện ngoại lai

columns = ['Recency', 'Frequency', 'Monetary']

# Khởi tạo DataFrame lưu thông tin ngoại lai

outlier\_info = pd.DataFrame(index=df.index)

# Phát hiện ngoại lai và xóa các dòng có ngoại lai

for col in columns:

    Q1 = df[col].quantile(0.25)  # Q1 (tứ phân vị thứ nhất)

    Q3 = df[col].quantile(0.75)  # Q3 (tứ phân vị thứ ba)

    IQR = Q3 - Q1  # Khoảng cách IQR

    # Tính các ngưỡng dưới và trên dựa trên IQR

    lower\_bound = Q1 - 1.5 \* IQR

    upper\_bound = Q3 + 1.5 \* IQR

    # Đánh dấu ngoại lai

    outlier\_info[col + '\_isoutlier'] = (df[col] < lower\_bound) | (df[col] > upper\_bound)

    # In thông tin về ngoại lai

    print(f"Cột '{col}':")

    print(f" - Q1: {Q1:.2f}")

    print(f" - Q3: {Q3:.2f}")

    print(f" - IQR: {IQR:.2f}")

    print(f" - Ngưỡng dưới: {lower\_bound:.2f}")

    print(f" - Ngưỡng trên: {upper\_bound:.2f}")

    print(f" - Số lượng ngoại lai: {outlier\_info[col + '\_isoutlier'].sum()}")

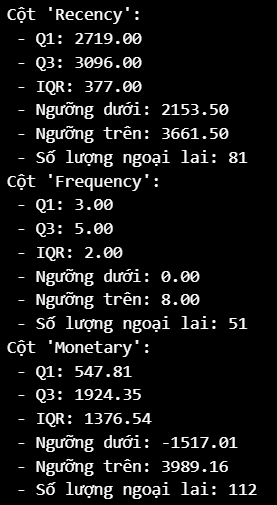
# Lọc ra các dòng có ít nhất 1 giá trị ngoại lai

outliers\_detected = outlier\_info.any(axis=1)

# Xóa ngoại lai

df = df[~outliers\_detected]

Kết quả:



Hình 16 Kết quả dùng IQR để phát hiện và xử lý các giá trị ngoại lai

#### Chuẩn hóa dữ liệu:

Để đảm bảo các đặc trưng **Recency, Frequency** và **Monetary** có cùng đơn vị và quy mô, nhóm đã tiến hành **chuẩn hóa dữ liệu bằng phương pháp Z-score**. Việc này đặc biệt quan trọng khi áp dụng thuật toán phân cụm **K-Means**, vốn sử dụng khoảng cách (Euclidean Distance) nên rất nhạy cảm với sự chênh lệch về đơn vị đo.

# Chuẩn hóa dữ liệu

import seaborn as sns

from scipy import stats

rfm\_columns = ['Recency', 'Frequency', 'Monetary']

# Chuẩn hóa bằng Z-score cho các cột RFM

df\_zscore = df[rfm\_columns].apply(lambda x: stats.zscore(x, nan\_policy='omit'))

print(df\_zscore.values)

# Vẽ biểu đồ mật độ phân phối theo Z-score

plt.figure(figsize=(10, 6))

for col in df\_zscore.columns:

    sns.kdeplot(df\_zscore[col], label=col, linewidth=2)

plt.title("Biểu đồ mật độ phân phối theo Z-score cho R,F,M")

plt.xlabel("Z-score")

plt.ylabel("Mật độ")

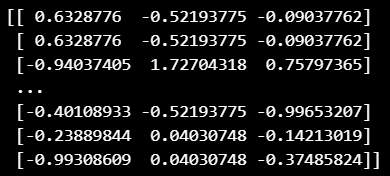
plt.legend()

plt.grid(True)

plt.tight\_layout()

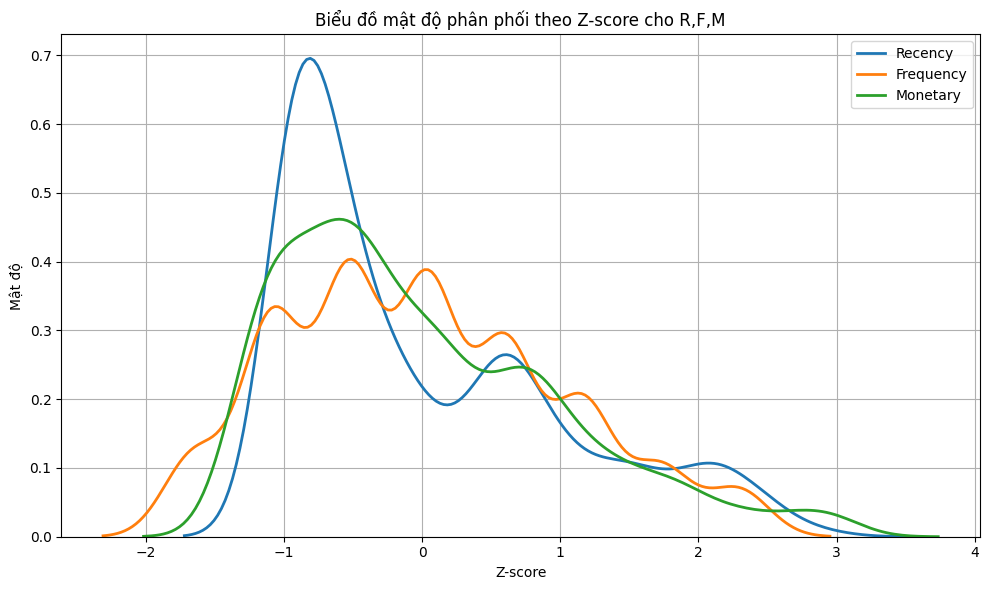
plt.show()

Kết quả sau khi thực hiện chuẩn hóa:



Hình 17 Kết quả chuẩn hóa dữ liệu R, F, M theo Z-score Normalization

Sau khi chuẩn hóa, nhóm cũng đã trực quan hóa phân phối của các đặc trưng RFM dưới dạng biểu đồ mật độ để kiểm tra sự lan tỏa và xu hướng phân phối dữ liệu sau chuẩn hóa:



Hình 18 Biểu đồ thể hiện mật độ phân phối theo Z-score cho R, F, M

Biểu đồ cho thấy cả ba đặc trưng RFM sau khi chuẩn hóa bằng Z-score đều có xu hướng phân phối không hoàn toàn chuẩn và xuất hiện độ lệch nhất định:

* Recency (đường màu xanh): Phân phối lệch trái, tập trung nhiều ở giá trị Z nhỏ hơn 0, nghĩa là phần lớn khách hàng có thời gian mua hàng gần đây hơn trung bình. Mật độ đạt đỉnh quanh Z ≈ -0.7 và giảm nhanh về phía Z dương.
* Frequency (đường màu cam): Phân phối có phần phân tán rộng, với nhiều đỉnh nhỏ cho thấy sự đa dạng trong số lần mua của khách hàng. Một bộ phận nhỏ khách hàng có Frequency cao vượt trội (Z > 2), nhưng phần lớn vẫn tập trung quanh trung bình (Z ≈ 0).
* Monetary (đường màu xanh lá): Phân phối có hình dạng khá tương đồng với Frequency, cho thấy sự phân hóa rõ rệt về giá trị chi tiêu giữa các khách hàng. Phần lớn khách hàng chi tiêu dưới mức trung bình, nhưng vẫn có nhóm chi tiêu cao (Z > 2).

### Xây dựng mô hình:

Sau khi chuẩn hóa dữ liệu bằng Z-score, nhóm tiến hành **phân cụm khách hàng** dựa trên các đặc trưng hành vi RFM (Recency, Frequency, Monetary) bằng thuật toán **K-Means**. K-Means là một trong những thuật toán phân cụm phổ biến, được sử dụng rộng rãi trong bài toán phân khúc khách hàng

#### Xác định K tối ưu bằng Elbow Method:

Nhóm đã áp dụng phương pháp Elbow (Khuỷu tay) để tìm ra giá trị k phù hợp nhất cho mô hình K-Means.

Phương pháp Elbow hoạt động dựa trên tiêu chí WCSS (Within-Cluster Sum of Squares), tức là tổng bình phương khoảng cách từ các điểm dữ liệu đến tâm cụm của chúng. Khi tăng số cụm, WCSS sẽ giảm. Tuy nhiên, sau một điểm nhất định, tốc độ giảm của WCSS chậm lại, tạo thành hình khuỷu tay trên đồ thị. Đó chính là điểm thích hợp để chọn số lượng cụm

# Elbow để xác định số K

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

import seaborn as sns

wcss = []  # within-cluster sum of squares

# Thử K từ 1 đến 10

for k in range(1, 11):

    kmeans = KMeans(n\_clusters=k, init='k-means++', random\_state=42)

    kmeans.fit(df\_zscore)

    wcss.append(kmeans.inertia\_)

# Vẽ biểu đồ Elbow

plt.figure(figsize=(8, 5))

plt.plot(range(1, 11), wcss, marker='o')

plt.title('Elbow Method xác định số cụm tối ưu')

plt.xlabel('Số cụm (k)')

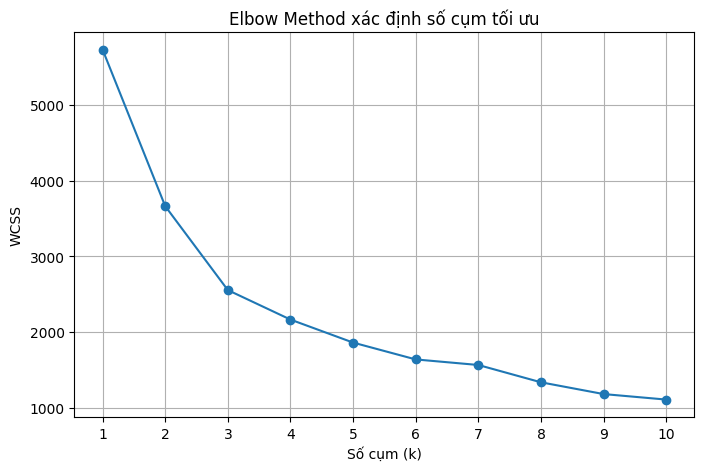
plt.ylabel('WCSS')

plt.xticks(range(1, 11))

plt.grid(True)

plt.show()

Quan sát biểu đồ:



Hình 19 Đồ thị thể hiện kết quả của phương pháp Elbow

Quan sát biểu đồ, nhóm xác định được điểm gấp khúc (elbow) xuất hiện rõ ràng tại **K = 3**, do đó, mô hình K-Means sẽ được huấn luyện với số cụm là 3 ở bước tiếp theo. Đây là số lượng cụm hợp lý để phân chia khách hàng thành các nhóm hành vi tương đồng mà không gây phân mảnh quá mức hoặc gom cụm thiếu chính xác

#### Phân cụm bằng thuật toán K-means:

Sau khi xác định được số cụm tối ưu là 3 bằng phương pháp Elbow, nhóm tiến hành áp dụng thuật toán K-Means để phân cụm khách hàng dựa trên các giá trị đã chuẩn hóa của bộ chỉ số RFM (Recency, Frequency, Monetary). Ở đây, nhóm không gán trọng số cho bất kỳ đặc trưng nào mà giữ nguyên tầm quan trọng ngang nhau giữa ba chỉ số R, F, M để giữ nguyên tính khách quan và giúp phản ánh tổng quát hành vi tiêu dùng thông qua các đặc trưng phân cụm. Mỗi khách hàng được gán nhãn cụm tương ứng dựa trên đặc điểm hành vi mua hàng.

# Phân cụm KMeans

from sklearn.cluster import KMeans

import seaborn as sns

import matplotlib.pyplot as plt

# Khởi tạo mô hình KMeans với 3 cụm

kmeans = KMeans(n\_clusters=3, init='k-means++', random\_state=42)

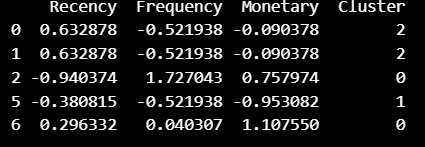
kmeans.fit(df\_zscore)

# Gán nhãn cụm vào df\_score

df\_zscore['Cluster'] = kmeans.labels\_

print(df\_zscore.head())

Thu được kết quả các cụm như sau:



Nhằm trực quan hóa kết quả phân cụm, nhóm đã sử dụng biểu đồ phân tán với hai trục đại diện cho Recency và Monetary, trong đó màu sắc biểu thị nhãn cụm. Biểu đồ giúp quan sát rõ sự phân tách giữa các nhóm khách hàng theo các đặc trưng hành vi gần nhất và giá trị chi tiêu.

#Trực quan hóa kết quả phân cụm bằng biểu đồ 2D

# Định nghĩa màu sắc

cluster\_colors = {0: 'red', 1: 'green', 2: 'blue'}

# Trực quan hóa kết quả

plt.figure(figsize=(8, 5))

sns.scatterplot(

    data=df\_zscore,

    x='Recency',

    y='Monetary',

    hue='Cluster',  # Phân biệt theo cụm

    palette=cluster\_colors,  # Sử dụng màu sắc đồng nhất cho các cụm

    s=50  # Kích thước điểm

)

plt.title('Phân cụm khách hàng theo Recency & Monetary (Z-score)')

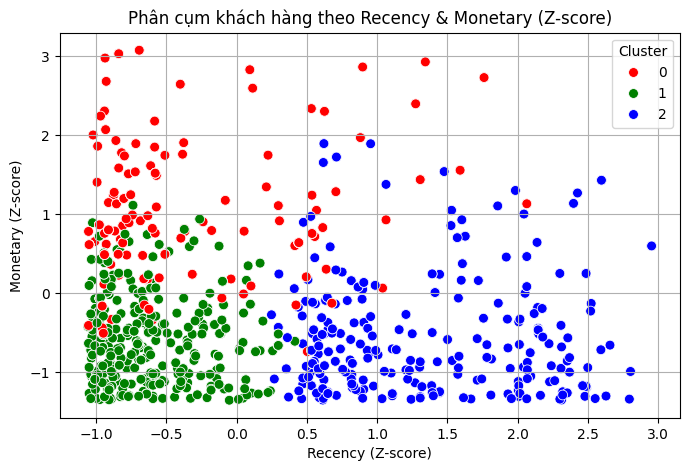
plt.xlabel('Recency (Z-score)')

plt.ylabel('Monetary (Z-score)')

plt.grid(True)

plt.show()

Kết quả thu được:



Hình 20 Biểu đồ phân tán thể hiện kết quả phân cụm khách hàng

Để trực quan hóa kết quả phân cụm toàn diện hơn, nhóm tiếp tục sử dụng biểu đồ phân tán 3 chiều (3D scatter plot), thể hiện mối quan hệ giữa ba thành phần chính của mô hình RFM: Recency, Frequency và Monetary. Mỗi điểm dữ liệu đại diện cho một khách hàng và được tô màu theo cụm mà họ thuộc vào.

Việc trực quan hóa theo không gian 3 chiều giúp nhóm nghiên cứu đánh giá được mức độ phân tách giữa các cụm, kiểm tra sự phân bố của khách hàng trong từng nhóm cũng như độ rõ nét của biên cụm sau khi áp dụng thuật toán K-Means

from mpl\_toolkits.mplot3d import Axes3D

import matplotlib.pyplot as plt

# Gán nhãn cụm

df\_zscore['Cluster'] = kmeans.labels\_

# Tạo figure 3D

fig = plt.figure(figsize=(10, 7))

ax = fig.add\_subplot(111, projection='3d')

# Vẽ từng cụm bằng màu sắc khác nhau

colors = ['red', 'green', 'blue']

for cluster in df\_zscore['Cluster'].unique():

    cluster\_data = df\_zscore[df\_zscore['Cluster'] == cluster]

    ax.scatter(

        cluster\_data['Recency'],

        cluster\_data['Frequency'],

        cluster\_data['Monetary'],

        s=50,

        c=colors[cluster],

        label=f'Cụm {cluster}',

        alpha=0.7

    )

ax.set\_title('Phân cụm khách hàng theo RFM (Biểu đồ 3D)')

ax.set\_xlabel('Recency (Z-score)')

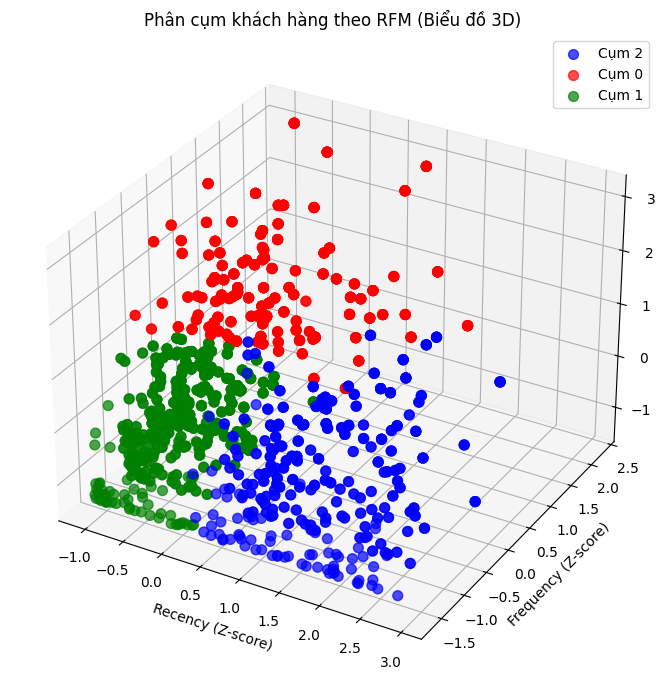
ax.set\_ylabel('Frequency (Z-score)')

ax.set\_zlabel('Monetary (Z-score)')

ax.legend()

plt.tight\_layout()

plt.show()



Hình 21 Biểu đồ 3D thể hiện kết quả phân cụm khách hàng

Sau khi tiến hành phân cụm khách hàng dựa trên ba đặc trưng hành vi RFM, nhóm đã xác định được ba cụm khách hàng có đặc điểm khác nhau. Cụm 0 được xác định là nhóm khách hàng VIP với các đặc trưng nổi bật như mua hàng gần đây (Recency thấp), tần suất mua cao (Frequency cao) và giá trị đơn hàng lớn (Monetary cao). Đây là nhóm khách hàng mang lại giá trị cao cho doanh nghiệp và cần được ưu tiên chăm sóc, giữ chân thông qua các chính sách ưu đãi đặc biệt hoặc chương trình khách hàng thân thiết.

Ngược lại, cụm 2 thể hiện nhóm khách hàng có nguy cơ rời bỏ khi sở hữu Recency cao (đã lâu không mua hàng), tần suất mua thấp và giá trị đơn hàng không đáng kể. Nhóm này cần được quan tâm thông qua các chiến lược kích hoạt lại, chẳng hạn như email cá nhân hóa, mã giảm giá hoặc các chiến dịch remarketing.

Cuối cùng, cụm 1 đại diện cho nhóm khách hàng trung bình hoặc tiềm năng, với các chỉ số RFM ở mức vừa phải. Dù chưa thực sự nổi bật nhưng nhóm này có khả năng phát triển thành khách hàng trung thành nếu được chăm sóc tốt, thông qua các hoạt động như giới thiệu sản phẩm mới, tư vấn cá nhân hóa, hoặc chương trình khuyến mãi phù hợp. Từ kết quả phân cụm, nhóm đã rút ra được các định hướng tiếp cận phù hợp cho từng nhóm khách hàng nhằm tối ưu hóa hiệu quả marketing và gia tăng giá trị vòng đời khách hàng (CLV).

## Đề xuất chiến lược CRM phù hợp với từng phân khúc khách hàng:

Dựa trên kết quả phân cụm bằng thuật toán K-means, nhóm đã xác định được ba nhóm khách hàng tiêu biểu với đặc điểm hành vi khác nhau: cụm 0 là nhóm khách hàng VIP, cụm 1 là nhóm khách hàng trung bình hoặc tiềm năng, và cụm 2 là nhóm khách hàng có nguy cơ rời bỏ. Từ đó, các chiến lược hành động cụ thể được đề xuất nhằm tối ưu hóa giá trị vòng đời khách hàng và tăng hiệu quả trong hoạt động chăm sóc, giữ chân cũng như phát triển khách hàng.

Đối với cụm 0 – nhóm khách hàng VIP, đây là những khách hàng có giá trị cao nhất vì họ mua hàng thường xuyên, với giá trị lớn và vẫn đang tiếp tục giao dịch gần đây. Doanh nghiệp cần tập trung giữ chân nhóm này bằng mọi giá thông qua việc triển khai các chương trình khách hàng thân thiết mang tính cá nhân hóa cao. Bên cạnh đó, thay vì tập trung vào các chương trình giảm giá đại trà, doanh nghiệp nên tăng giá trị vòng đời khách hàng (CLV) bằng cách đề xuất các sản phẩm cao cấp, các gói combo có giá trị cao dựa trên hành vi mua sắm trong quá khứ.

Cụm 1 là nhóm khách hàng trung bình hoặc tiềm năng. Họ có tần suất mua hàng khá ổn định nhưng giá trị giỏ hàng vẫn ở mức trung bình. Đối với nhóm này, chiến lược phù hợp là khuyến khích họ gia tăng giá trị đơn hàng thông qua các ưu đãi gắn với ngưỡng chi tiêu (ví dụ: tặng quà cho đơn hàng trên mức trung bình). Ngoài ra, các chương trình giới thiệu bạn bè hoặc thúc đẩy khách hàng trở thành người ủng hộ thương hiệu cũng rất hữu ích. Với những khách hàng mới chỉ mua một lần gần đây, doanh nghiệp cần xây dựng trải nghiệm ban đầu tích cực thông qua gửi lời cảm ơn, khảo sát trải nghiệm, tặng voucher chào mừng và đề xuất các sản phẩm liên quan để tăng khả năng quay lại và trung thành.

Cụm 2 là nhóm khách hàng có nguy cơ rời bỏ. Họ từng có tần suất và giá trị mua hàng tốt nhưng không còn quay lại gần đây. Với nhóm này, thay vì áp dụng các chương trình ưu đãi đại trà, doanh nghiệp nên tìm hiểu nguyên nhân khách hàng rời đi bằng cách khảo sát, từ đó có hướng điều chỉnh phù hợp về sản phẩm, dịch vụ hoặc hình thức chăm sóc. Việc kết nối lại qua các kênh cá nhân hóa như email, mạng xã hội hoặc gọi điện cũng có thể giúp tái kích hoạt nhu cầu. Ngoài ra, các chiến dịch tiếp thị lại (retargeting) hoặc ưu đãi độc quyền ngắn hạn sẽ giúp thu hút sự chú ý và khuyến khích khách hàng thực hiện giao dịch mới trong thời gian sớm nhất.

Những chiến lược này không chỉ góp phần tăng hiệu quả tiếp thị, bán hàng mà còn giúp doanh nghiệp tối ưu hóa nguồn lực chăm sóc khách hàng dựa trên đặc điểm hành vi của từng nhóm cụ thể.

# KẾT LUẬN ĐỀ TÀI

Bài báo cáo đã trình bày quá trình ứng dụng mô hình RFM kết hợp với thuật toán phân cụm K-Means nhằm phân khúc khách hàng theo hành vi mua sắm, dựa trên tập dữ liệu Store Sales Forecasting Dataset. Nhóm đã tiến hành xử lý, chuẩn hóa và phân tích dữ liệu, từ đó xây dựng mô hình RFM giúp lượng hóa hành vi khách hàng dựa trên ba yếu tố: lần mua hàng gần nhất (Recency), tần suất mua hàng (Frequency) và tổng giá trị giao dịch (Monetary). Sau đó, thuật toán K-Means được áp dụng để phân cụm khách hàng thành các nhóm hành vi tiêu dùng đặc trưng, với số cụm tối ưu được xác định bằng phương pháp Elbow nhằm đảm bảo tính hợp lý trong phân khúc. Việc trực quan hóa dữ liệu bằng các công cụ như histogram, boxplot, heatmap và scatter plot giúp làm nổi bật đặc điểm của từng nhóm, hỗ trợ việc phân tích và diễn giải kết quả rõ ràng hơn.

Nghiên cứu này khẳng định vai trò cốt lõi của việc phân khúc khách hàng trong chiến lược CRM hiện đại, khi mà hiểu biết sâu sắc về hành vi người tiêu dùng cho phép doanh nghiệp tối ưu hóa các hoạt động tương tác khách hàng và nâng cao giá trị vòng đời (CLV).

Kết quả phù hợp với định hướng nghiên cứu, trong đó cho thấy mô hình RFM kết hợp với phân cụm là một công cụ hiệu quả trong CRM phân tích (Analytical CRM), giúp doanh nghiệp ra quyết định dựa trên dữ liệu (data-driven CRM)

Việc kết hợp mô hình RFM với thuật toán K-Means đã mang lại cái nhìn toàn diện và sâu sắc hơn về hành vi tiêu dùng của từng nhóm khách hàng. Qua đó, doanh nghiệp có thể dễ dàng xác định các phân khúc khách hàng mục tiêu để từ đó xây dựng các chiến lược tiếp thị phù hợp hơn như cá nhân hóa trải nghiệm, duy trì sự trung thành của khách hàng hiện tại hoặc thu hút nhóm khách hàng tiềm năng. Đặc biệt, phương pháp này giúp nhận diện rõ nhóm khách hàng có giá trị cao để tập trung đầu tư phát triển.

Kết quả phân tích cũng cho phép doanh nghiệp phân loại khách hàng thành các nhóm như VIP, tiềm năng, không năng động..., từ đó triển khai các chiến lược CRM phù hợp với từng nhóm. Quá trình triển khai mô hình không chỉ hỗ trợ ra quyết định trong kinh doanh mà còn giúp nhóm nâng cao kỹ năng xử lý và phân tích dữ liệu, hiểu rõ hơn về cách ứng dụng học máy không giám sát trong việc tối ưu hóa hoạt động marketing và chăm sóc khách hàng. Quá trình thực hiện dự án không chỉ mang lại giá trị thực tiễn mà còn giúp nhóm củng cố kiến thức về phân tích dữ liệu và học máy không giám sát trong bối cảnh kinh doanh hiện đại.

**LỜI CẢM ƠN**

Trong suốt quá trình nghiên cứu và hoàn thiện đề tài “Ứng dụng mô hình RFM kết hợp thuật toán K-means trong phân khúc khách hàng theo hành vi mua sắm phục vụ chiến lược CRM”, chúng em đã nhận được sự quan tâm, hỗ trợ và đồng hành quý báu từ nhiều cá nhân và tổ chức.

Trước hết, chúng em xin bày tỏ lòng biết ơn sâu sắc đến T.S Lê Diên Tuấn – giảng viên hướng dẫn, người đã không chỉ truyền đạt kiến thức chuyên môn một cách tâm huyết, mà còn đưa ra những định hướng rõ ràng, các nhận xét mang tính thực tiễn và sự hỗ trợ kịp thời trong suốt quá trình triển khai đề tài. Những đóng góp từ Thầy là nền tảng quan trọng giúp chúng em hoàn thiện phương pháp tiếp cận, khai thác hiệu quả dữ liệu, và xây dựng mô hình có tính ứng dụng cao trong bối cảnh thực tiễn.

Chúng em cũng xin chân thành cảm ơn Khoa Thống kê - Tin học, Trường Đại học Kinh tế – Đại học Đà Nẵng vì đã tạo điều kiện học tập, nghiên cứu thuận lợi cả về cơ sở vật chất, môi trường học thuật và tài nguyên nghiên cứu. Những nguồn lực này đã góp phần không nhỏ giúp chúng em hoàn thành đề tài đúng tiến độ và định hướng ban đầu.

Cuối cùng, chúng em gửi lời tri ân đến bạn bè, người thân và các cộng sự đã hỗ trợ về mặt tinh thần, kỹ thuật và góp ý khách quan trong suốt quá trình làm việc.

Dù đã cố gắng hoàn thiện với tinh thần trách nhiệm cao, song do hạn chế về thời gian, kinh nghiệm thực tiễn và phạm vi tiếp cận, đề tài vẫn còn những điểm cần tiếp tục cải tiến. Chúng em rất mong nhận được những ý kiến đóng góp quý báu từ Thầy để phát triển nghiên cứu này theo hướng ứng dụng thực tiễn sâu rộng hơn.

Trân trọng cảm ơn!

**Phần trình bày báo cáo:** [**https://youtu.be/Otfrg1i3j00**](https://youtu.be/Otfrg1i3j00)

**TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | Okky Putra Barus; Celine Nathasya; Jefri Junifer Pangaribuan, "The Implementation of RFM Analysis to Customer Profiling Using K-Means Clustering," 30 10 2022. [Online]. Available: https://iieta.org/journals/mmep/paper/10.18280/mmep.100135. [Accessed 14 12 2022]. |
| [2] | I. T. a. M. Science, "Data Analytics in CRM Processes: A Literature Review," 2017. |
| [3] | T. L. D. Tuấn, Bài giảng Phân tích dữ liệu bằng Python, 2025. |
| [4] | P. Đ. Khánh, "Model RFM phân khúc khách hàng," 8 11 2019. [Online]. Available: https://phamdinhkhanh.github.io/2019/11/08/RFMModel.html. |
| [5] | A. Vidhya, "K-Means Clustering Algorithm," 10 5 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/. |
| [6] | T. Marketers, "Exploratory Data Analysis (EDA)," 27 4 2023. [Online]. Available: https://blog.tomorrowmarketers.org/exploratory-data-analysis-la-gi/. |
| [7] | M. AMIS, "CRM," 17 4 2025. [Online]. Available: https://amis.misa.vn/15274/he-thong-crm/. |