ĐẠI HỌC ĐÀ NẴNG

TRƯỜNG ĐẠI HỌC KINH TẾ

A white shield with blue green orange and red text

AI-generated content may be incorrect.

**BÁO CÁO DỰ ÁN NHÓM**

**HỌC PHẦN PHÂN TÍCH DỮ LIỆU BẰNG PYTHON**

**CHỦ ĐỀ: CRM - HYPOTHESIS TEST**

Mã học phần : MIS3041

Giảng viên: Lê Diên Tuấn

Nhóm: 10

Thành viên thực hiện: Trương Thị Mỹ Hồng - 0385276173

Nguyễn Thị Thanh Tâm

Nguyễn Thị Thúy

Nguyễn Thị Mai Anh

Nguyễn Thị Khánh Linh

*Đà Nẵng, ngày 03 tháng 05 năm 2025*

Tỷ lệ đóng góp của các thành viên:

|  |  |
| --- | --- |
| Tên thành viên | % Đóng góp |
| Trương Thị Mỹ Hồng | 100% |
| Nguyễn Thị Thúy | 100% |
| Nguyễn Thị Mai Anh | 100% |
| Nguyễn Thị Thanh Tâm | 100% |
| Nguyễn Thị Khánh Linh | 100% |

Mục lục

[1 Giới thiệu tổng quan 1](#_Toc196764132)

[1.1 Giới thiệu chung 1](#_Toc196764133)

[1.2 Mục tiêu và ứng dụng của bài báo cáo 2](#_Toc196764134)

[1.2.1 Mục tiêu 2](#_Toc196764135)

[1.2.2 Ứng dụng thực tiễn 2](#_Toc196764136)

[1.3 Đối tượng và phạm vi của bài báo cáo 3](#_Toc196764137)

[1.3.1 Đối tượng của bài báo cáo 3](#_Toc196764138)

[1.3.2 Phạm vi nghiên cứu 3](#_Toc196764139)

[1.4 Phương pháp thực hiện bài báo cáo 3](#_Toc196764143)

[2 Cơ sở lý thuyết 3](#_Toc196764144)

[2.1 Quản trị quan hệ khách hàng CRM 3](#_Toc196764145)

[2.2 Tổng quan về phân cụm khách hàng 6](#_Toc196764146)

[2.2.1 Định nghĩa phân cụm khách hàng 6](#_Toc196764147)

[2.2.2 Các phương pháp phân cụm phổ biến 6](#_Toc196764148)

[2.3 Phương pháp RFM trong phân cum khách hàng 7](#_Toc196764149)

[2.3.1 Định nghĩa Recency, Frequency, Monetary 7](#_Toc196764150)

[2.3.2 Cách tính điểm RFM từ dữ liệu khách hàng 9](#_Toc196764151)

[2.3.3 Ứng dụng của RFM trong phân tích dữ liệu khách hàng 11](#_Toc196764156)

[2.4 Thuật toán K-Means 12](#_Toc196764157)

[2.4.1 Nguyên lý hoạt động của K-Means 12](#_Toc196764158)

[2.4.2 Phương pháp Elbow để xác định số cụm tối ưu 13](#_Toc196764162)

[2.4.3 Ưu và nhược điểm của thuật toán K-Means 14](#_Toc196764163)

[2.5 Kiểm định sự khác biệt giữa các cụm 15](#_Toc196764166)

[2.5.1 Tổng quan về kiểm định giả thuyết trong phân tích cụm 15](#_Toc196764167)

[2.5.2 Các phương pháp kiểm định 16](#_Toc196764168)

[3 Kết quả thực nghiệm 18](#_Toc196764169)

[3.1 Xây dựng framework 18](#_Toc196764170)

[3.2 Chọn dữ liệu (Dataset) 20](#_Toc196764173)

[3.2.1 Giới thiệu về dữ liệu 20](#_Toc196764174)

[3.2.2 Mô tả dữ liệu 21](#_Toc196764175)

[3.3 Chuẩn bị dữ liệu 22](#_Toc196764176)

[3.3.1 Triển khai tiền xử lý dữ liệu 22](#_Toc196764177)

[3.3.2 Triển khai tính toán các chỉ số RFM 29](#_Toc196764178)

[3.3.3 Chuẩn hóa chỉ số RFM bằng phương pháp z-score 32](#_Toc196764185)

[3.3.4 Triển khai trực quan hóa dữ liệu RFM 35](#_Toc196764190)

[3.3.5 Loại bỏ ngoại lệ bằng phương pháp IQR 36](#_Toc196764191)

[3.4 Phân cụm khách hàng 39](#_Toc196764192)

[3.4.1 Triển khai xác định số cụm tối ưu bằng phương pháp Elbow 39](#_Toc196764193)

[3.4.2 Triển khai thuật toán K-Means để phân cụm khách hàng 41](#_Toc196764194)

[3.4.3 Triển khai đánh giá kết quả phân cụm bằng trực quan hóa dữ liệu 44](#_Toc196764195)

[3.5 Kiểm định sự khác biệt giữa các cụm 47](#_Toc196764196)

[3.5.1 Thực hiện kiểm định ANOVA 47](#_Toc196764197)

[3.5.2 Triển khai Tukey HSD 47](#_Toc196764198)

[3.5.3 Phân tích cụm 53](#_Toc196764203)

[3.5.4 Kết luận các cụm khách hàng 58](#_Toc196764204)

[3.6 Phân tích kết quả 59](#_Toc196764206)

[4 Kết luận 61](#_Toc196764209)

Danh mục hình ảnh

[Hình 1 Các bước tính RFM 9](#_Toc196764710)

[Hình 2: Minh họa thuật toán k-mean 12](#_Toc196764711)

[Hình 3: Framework phân cụm và phân tích khách hàng 18](#_Toc196764712)

[Hình 4: Dữ liệu mua hàng 21](#_Toc196764713)

[Hình 5: Dữ liệu trước khi tiến hành tiền xử lí 22](#_Toc196764714)

[Hình 6: Thực hiện đọc file dữ liệu 22](#_Toc196764715)

[Hình 7: Tổng quát dữ liệu 23](#_Toc196764716)

[Hình 8: Thực hiện in 5 dòng đầu tiên của dữ liệu 24](#_Toc196764717)

[Hình 9: 5 dòng đầu tiên của dữ liệu 24](#_Toc196764718)

[Hình 10: Thực hiện xem tổng quan và thống kê mô tả của dữ liệu 25](#_Toc196764719)

[Hình 11: Thống kê mô tả tổng quan 25](#_Toc196764720)

[Hình 12: Thực hiện kiểm tra các giá trị null 25](#_Toc196764721)

[Hình 13: Thực hiện kiểm tra dữ liệu không hợp lệ 26](#_Toc196764722)

[Hình 14: Kết quả sau khi kiểm tra tính hợp lệ 26](#_Toc196764723)

[Hình 15: Giữ lại các thông tin quan trọng cho quá trình tính toán RFM 27](#_Toc196764724)

[Hình 16: Thực hiện lưu dữ liệu đã làm sạch 27](#_Toc196764725)

[Hình 17: Hình ảnh dữ liệu sau khi làm sạch 28](#_Toc196764726)

[Hình 18: Dữ liệu sau khi làm sạch 29](#_Toc196764727)

[Hình 19: thực hiện chuyển đổi cột ngày giao dịch sang định dạng datetime và xác định ngày tham chiếu 30](#_Toc196764728)

[Hình 20: Thực hiện tính toán các chỉ số RFM 31](#_Toc196764729)

[Hình 21: Kết quả sau khi tính RFM 31](#_Toc196764730)

[Hình 22: Thực hiện đảo chiều Recency 33](#_Toc196764731)

[Hình 23: Thực hiện chuẩn hóa Z-score 33](#_Toc196764732)

[Hình 24: Kết quả sau khi chuẩn hóa Z-score 34](#_Toc196764733)

[Hình 25: Boxplot thể hiện phân bố các giá trị RFM sau khi chuẩn hóa 35](#_Toc196764734)

[Hình 26 : Thực hiện loại bỏ ngoại lệ 37](#_Toc196764735)

[Hình 27: Kết quả sau khi loại bỏ ngoại lệ 37](#_Toc196764736)

[Hình 28: Boxplot của Recency, Frequency và Monetary sau khi loại bỏ ngoại lệ bằng phương pháp IQR 38](#_Toc196764737)

[Hình 29: Thực hiện xác định số cụm tối ưu bằng phương pháp Elbow 39](#_Toc196764738)

[Hình 30: Biểu đồ Elbow để xác định số cụm tối ưu 40](#_Toc196764739)

[Hình 31: Triển khai phân cụm với k =4 41](#_Toc196764740)

[Hình 32: Lưu các cụm vào các file riêng 41](#_Toc196764741)

[Hình 33: Khách hàng cụm 0 42](#_Toc196764742)

[Hình 34: Hình ảnh cụm 1 43](#_Toc196764743)

[Hình 35: Hình ảnh khách hàng cụm 2 43](#_Toc196764744)

[Hình 36: Hình ảnh khách hàng cụm 3 44](#_Toc196764745)

[Hình 37: Phân cụm khách hàng bằng K-Means với k=4 45](#_Toc196764746)

[Hình 38: Trung bình RFM của từng cụm khách hàng 46](#_Toc196764747)

[Hình 39: Thực hiện kiểm định Tukey HSD 48](#_Toc196764748)

[Hình 40:  *HSD - Recency* 49](#_Toc196764749)

[Hình 41: *HSD - Frequency* 51](#_Toc196764750)

[Hình 42: *HSD - Monetary* 52](#_Toc196764751)

[Hình 43: Vẽ biểu đồ phân phối cụm 0 53](#_Toc196764752)

[Hình 44: Biểu đồ phân phối cụm 0 53](#_Toc196764753)

[Hình 45: Vẽ biểu đồ phân phối cụm 1 54](#_Toc196764754)

[Hình 46: Biểu đồ phân phối cụm 1 55](#_Toc196764755)

[Hình 47: Vẽ biểu đồ phân phối cụm 2 56](#_Toc196764756)

[Hình 48: Biểu đồ phân phối cụm 2 56](#_Toc196764757)

[Hình 49: Vẽ biểu đồ phân phối cụm 3 57](#_Toc196764758)

[Hình 50: Biểu đồ phân phối cụm 3 58](#_Toc196764759)

# Giới thiệu tổng quan

## Giới thiệu chung

- Trong môi trường kinh doanh hiện đại, việc đưa ra quyết định dựa trên dữ liệu đang ngày càng trở nên quan trọng nhằm giảm thiểu rủi ro và tối ưu hóa hiệu quả hoạt động. Trong đó, kiểm định sự khác biệt giữa các nhóm khách hàng đóng vai trò then chốt trong việc xác định chiến lược tiếp cận phù hợp, giúp doanh nghiệp phân bổ nguồn lực một cách hợp lý và chính xác hơn. Với tầm quan trọng đó, đề tài “Kiểm định sự khác biệt giữa các cụm khách hàng” được thực hiện dựa trên bộ dữ liệu hành vi mua sắm trực tuyến của khách hàng, nhằm khám phá đặc điểm tiêu dùng và hành vi mua hàng thực tế của người tiêu dùng.  
- Nghiên cứu này tập trung vào việc phân tích hành vi khách hàng của doanh nghiệp thông qua dữ liệu bán hàng thực tế, với mục tiêu chính là phân khúc khách hàng dựa trên mô hình RFM (Recency - Tần suất mua gần nhất, Frequency - Tần suất mua, và Monetary -Giá trị mua). Phân tích được thực hiện bằng cách ứng dụng các thuật toán học máy không giám sát, cụ thể là K-Means, nhằm tìm ra các nhóm khách hàng có đặc điểm tiêu dùng tương đồng. Việc xác định các cụm khách hàng như vậy giúp doanh nghiệp hiểu rõ hơn về hành vi của từng nhóm đối tượng, từ đó xây dựng các chiến lược tiếp thị cá nhân hóa và phù hợp với từng phân khúc. (Li, n.d.)

- Việc phân khúc khách hàng đóng vai trò then chốt trong việc tối ưu hóa chiến lược kinh doanh tổng thể. Cụ thể, nó hỗ trợ doanh nghiệp trong việc ra quyết định về cách phân bổ nguồn lực tiếp thị, xây dựng chính sách khuyến mãi, quản lý mối quan hệ khách hàng, và dự báo nhu cầu tiêu dùng. Thay vì áp dụng một chiến lược chung cho toàn bộ tập khách hàng, doanh nghiệp có thể thiết kế những chương trình chuyên biệt cho từng nhóm, nhằm nâng cao trải nghiệm cá nhân hóa và tăng mức độ hài lòng của khách hàng. Điều này không chỉ thúc đẩy doanh số bán hàng mà còn giúp giữ chân khách hàng trung thành và cải thiện khả năng cạnh tranh trên thị trường. Ngoài ra, nghiên cứu còn mang lại ý nghĩa thiết thực về mặt chiến lược lâu dài. Việc hiểu rõ từng phân khúc khách hàng dựa trên hành vi thực tế giúp doanh nghiệp nhận diện các nhóm khách hàng có giá trị cao, từ đó tập trung đầu tư nguồn lực vào các nhóm mang lại lợi nhuận lớn. Đồng thời, những khách hàng có trải nghiệm tích cực sẽ có xu hướng quay lại và giới thiệu thêm khách hàng mới, tạo nên hiệu ứng lan tỏa tích cực. Nghiên cứu này không chỉ mang tính ứng dụng trong ngắn hạn mà còn hỗ trợ doanh nghiệp trong việc phát triển bền vững, thông qua các quyết định kinh doanh dựa trên dữ liệu (data-driven decision making).

## Mục tiêu và ứng dụng của bài báo cáo

### Mục tiêu - Khi dữ liệu khách hàng ngày càng trở nên phong phú và phức tạp, việc khai thác giá trị tiềm ẩn từ dữ liệu không còn dừng lại ở việc mô tả hành vi, mà tiến xa hơn đến việc khám phá và kiểm chứng những mối quan hệ ẩn sâu trong đó. Trong bối cảnh đó, kiểm định sự khác biệt giữa các cụm khách hàng trở thành một bước quan trọng nhằm đánh giá độ chính xác và tính ứng dụng thực tiễn của các phân khúc đã được xây dựng. Trong toàn bộ quá trình này, phân khúc khách hàng đóng vai trò then chốt, giúp doanh nghiệp chia nhỏ thị trường thành các nhóm có đặc điểm, nhu cầu và hành vi tiêu dùng tương đồng — từ đó tạo tiền đề cho việc triển khai các chiến lược tiếp cận hiệu quả và phù hợp hơn với từng đối tượng (Ideascale, n.d.)

### Ứng dụng thực tiễn

Phân cụm khách hàng không chỉ là một bước trong quy trình phân tích dữ liệu, mà còn mang lại giá trị thực tiễn rõ rệt trong hoạt động kinh doanh và quản trị khách hàng. Việc chia khách hàng thành các nhóm có hành vi tiêu dùng tương đồng giúp doanh nghiệp hiểu rõ đặc điểm của từng phân khúc, từ đó thiết kế các chiến lược tiếp thị, chăm sóc và bán hàng phù hợp với từng nhóm đối tượng.  
- Trên thực tế, doanh nghiệp có thể áp dụng kết quả phân cụm để tối ưu hóa việc phân bổ ngân sách marketing, thay vì áp dụng một chiến dịch đại trà, sẽ tập trung nguồn lực vào những nhóm khách hàng tiềm năng hoặc có giá trị cao. Điều này không chỉ giúp gia tăng hiệu quả tiếp cận và chuyển đổi, mà còn nâng cao trải nghiệm cá nhân hóa, khiến khách hàng cảm thấy được quan tâm và thấu hiểu, từ đó tăng mức độ hài lòng và trung thành.  
- Bên cạnh đó, phân cụm khách hàng còn hỗ trợ dự báo nhu cầu tiêu dùng, giúp doanh nghiệp chuẩn bị nguồn hàng và điều chỉnh kế hoạch bán hàng một cách chủ động. Các nhóm khách hàng có xu hướng rời đi cũng có thể được phát hiện sớm, từ đó đưa ra các chiến lược giữ chân kịp thời. Ngoài ra, việc hiểu rõ các cụm khách hàng còn giúp xác định được những phân khúc có tỷ suất sinh lợi cao, là cơ sở để doanh nghiệp định hướng đầu tư dài hạn và phát triển bền vững.  
- Việc ứng dụng phân cụm khách hàng trong thực tiễn mang đến nhiều lợi ích thiết thực cho doanh nghiệp, từ việc xây dựng chiến lược tiếp cận phù hợp, tối ưu nguồn lực tiếp thị, đến nâng cao chất lượng chăm sóc khách hàng. Thông qua cách tiếp cận dựa trên dữ liệu, doanh nghiệp không chỉ cải thiện hiệu quả hoạt động kinh doanh hiện tại mà còn tạo nền tảng vững chắc cho sự phát triển bền vững trong tương lai. (TopOnSeek, n.d.)

## Đối tượng và phạm vi của bài báo cáo

### Đối tượng của bài báo cáo Đối tượng nghiên cứu của bài báo cáo là quá trình phân cụm dữ liệu khách hàng và sản phẩm từ tập dữ liệu Superstore – một tập dữ liệu thương mại điện tử phổ biến, chứa thông tin về đơn hàng, khách hàng, sản phẩm, khu vực địa lý, doanh thu và lợi nhuận.

### Phạm vi nghiên cứu

### Việc áp dụng các kỹ thuật phân cụm dữ liệu để tìm ra các nhóm khách hàng hoặc sản phẩm có đặc điểm tương đồng.

### Xác định số cụm tối ưu dựa trên các chỉ số đánh giá nội tại của mô hình.

### Kiểm định độ tin cậy của phân cụm bằng các phương pháp thống kê.

## Phương pháp thực hiện bài báo cáo - Phương pháp nghiên cứu lý thuyết: Thu thập và tổng hợp thông tin từ các tài liệu học thuật, sách chuyên ngành và các blog trực tuyến uy tín. Mục tiêu là củng cố nền tảng kiến thức và cung cấp cơ sở lý luận cần thiết phục vụ cho quá trình nghiên cứu đề tài. - Phương pháp nghiên cứu thực nghiệm: Thực hiện các thử nghiệm liên tục nhằm điều chỉnh và hoàn thiện quy trình phân khúc khách hàng sao cho phù hợp với nhu cầu thực tiễn. Việc này giúp đảm bảo tính linh hoạt và khả năng thích ứng của mô hình trong các tình huống cụ thể - Kết hợp giữa lý thuyết và thực nghiệm: Tích hợp các cơ sở lý luận với dữ liệu thực tiễn để làm rõ các nguyên tắc phân khúc khách hàng. Đồng thời, phương pháp này cũng giúp kiểm chứng và hiệu chỉnh quy trình, đảm bảo độ chính xác cao trong việc xác định và chia nhóm các cụm khách hàng mục tiêu.

# Cơ sở lý thuyết

## Quản trị quan hệ khách hàng CRM

CRM là viết tắt của cụm từ Customer Relationship Management, nghĩa là quản lý quan hệ khách hàng. CRM là sự kết hợp giữa các chiến lược và công nghệ để cải thiện, xây dựng các mối quan hệ với khách hàng hiện tại cũng như khách hàng tiềm năng.

[CRM](https://bizfly.vn/techblog/crm-la-gi.html) giúp các doanh nghiệp thu thập, quản lý và phân tích thông tin về khách hàng từ nhiều nguồn khác nhau như cuộc gọi điện thoại, email, trang web, mạng xã hội và giao dịch mua bán. Bằng cách sử dụng hệ thống CRM (CRM system), các tổ chức có thể hiểu rõ hơn về nhu cầu và mong muốn của khách hàng, từ đó tăng cường khả năng phục vụ và tạo ra trải nghiệm tốt hơn cho khách hàng.

Các tính năng chính của CRM bao gồm quản lý thông tin khách hàng, quản lý hoạt động bán hàng, quản lý dịch vụ hỗ trợ khách hàng, quản lý chiến dịch Marketing và phân tích dữ liệu khách hàng. Ngoài ra, CRM cũng có thể tích hợp với các hệ thống khác như hệ thống quản lý khách hàng, hệ thống quản lý doanh nghiệp và hệ thống tự động hóa [Marketing](https://www.pace.edu.vn/tin-kho-tri-thuc/marketing-la-gi) để tối ưu hóa quy trình kinh doanh và tăng cường hiệu suất làm việc.

Lợi ích của hệ thống CRM với doanh nghiệp

Hệ thống quản lý mối quan hệ khách hàng (CRM) mang lại nhiều lợi ích quan trọng cho doanh nghiệp:

* Quản lý thông tin khách hàng

Hệ thống CRM cho phép doanh nghiệp thu thập, tổ chức và quản lý thông tin về khách hàng một cách hiệu quả. Các thông tin như tên, địa chỉ, lịch sử mua hàng, tương tác trước đó với doanh nghiệp, cùng các thông tin khác có thể được lưu trữ và truy cập dễ dàng. Giúp cải thiện quá trình phục vụ khách hàng, cung cấp thông tin hữu ích về khách hàng, từ đó có những chiến lược cải thiện, làm hài lòng khách hàng.

* Tăng cường tương tác khách hàng

CRM cung cấp các tính năng, công cụ cho phép doanh nghiệp tương tác với khách hàng một cách hiệu quả. Nhờ vào hệ thống CRM, doanh nghiệp có thể nhanh chóng gửi các [email Marketing](https://www.pace.edu.vn/tin-kho-tri-thuc/email-marketing-la-gi), thông báo hoặc những thông tin cập nhật tới khách hàng một cách tự động và cá nhân hóa, dựa trên thông tin sẵn có của khách hàng trong hệ thống. Giúp [đội ngũ Marketing](https://www.pace.edu.vn/tin-kho-tri-thuc/phong-marketing) tạo ra các chiến dịch tiếp thị hiệu quả hơn và tăng cường tương tác với khách hàng.

* Cải thiện chăm sóc khách hàng

Với hệ thống CRM, công cụ này cho phép doanh nghiệp theo dõi và quản lý tương tác với khách hàng theo thời gian thực. Bằng cách hiểu rõ hơn về nhu cầu và mong muốn của khách hàng theo từng giai đoạn, doanh nghiệp có thể cung cấp dịch vụ tốt hơn, cải thiện chất lượng chăm sóc khách hàng và tạo ra trải nghiệm tuyệt vời hơn cho khách hàng.

* Tối ưu hóa quy trình bán hàng

CRM giúp tối ưu hóa quy trình bán hàng bằng cách theo dõi tiến trình bán hàng, từ việc tìm kiếm khách hàng tiềm năng, liên hệ, đến quản lý các giao dịch và theo dõi sau bán hàng. Giúp tăng cường hiệu suất bán hàng, đẩy nhanh quá trình chốt hợp đồng và nâng cao tỷ lệ chuyển đổi khách hàng.

* Phân tích và báo cáo

Hệ thống CRM cung cấp các công cụ phân tích và báo cáo để đánh giá hiệu suất kinh doanh, hiểu rõ hơn về hành vi khách hàng. Thông qua việc phân tích dữ liệu khách hàng, doanh nghiệp có thể nhận ra các xu hướng, mẫu hành vi và triển khai các chiến lược kinh doanh phù hợp.

* Xây dựng quan hệ khách hàng lâu dài

Hệ thống CRM giúp doanh nghiệp xây dựng mối quan hệ lâu dài với khách hàng. Bằng cách cung cấp dịch vụ và hỗ trợ tốt, tạo ra trải nghiệm khách hàng tích cực và theo dõi sự hài lòng của khách hàng thường xuyên, doanh nghiệp có thể tạo ra một cộng đồng khách hàng trung thành và thu hút khách hàng mới.

* Kết nối đội ngũ

Khi các đội ngũ hoạt động riêng biệt và không chia sẻ thông tin khách hàng theo thời gian thực, điều này có thể trở thành một rào cản lớn, ảnh hưởng đến sự phát triển của tổ chức. Sự tách biệt này có thể gây khó khăn trong việc phục vụ nhu cầu của khách hàng và không đảm bảo một trải nghiệm tốt cho họ. Do đó, có một hệ thống CRM giúp các bộ phận/ phòng ban có thể nắm bắt thông tin nhanh, kịp thời hơn, từ đó gắn kết đội ngũ với tinh thần [teamwork](https://www.pace.edu.vn/tin-kho-tri-thuc/teamwork-la-gi) cao.

* Kiểm tra mục nhập dữ liệu không chính xác

Sử dụng dữ liệu kém chất lượng có thể dẫn đến những quyết định sai lầm đáng lo ngại. Để tránh tình trạng này, việc kiểm tra tính chính xác của dữ liệu trước khi nhập vào hệ thống là rất quan trọng. Áp dụng một quy trình xác thực và tính năng nhập dữ liệu tự động, doanh nghiệp có thể ngăn chặn luồng thông tin không chính xác trước khi gây rối và gây hậu quả khôn lường.

## Tổng quan về phân cụm khách hàng

### Định nghĩa phân cụm khách hàng

Phân cụm khách hàng là quá trình phân chia khách hàng thành nhiều cụm/nhóm có chung sự tương đồng theo những tiêu chí như giới tính, tuổi tác, sở thích, thu nhập và thói quen chi tiêu, hành vi mua sắm,… để doanh nghiệp có phương thức tiếp thị hiệu quả. Khi thực hiện được phân cụm khách hàng giúp đơn vị giải quyết đúng các yêu cầu của từng khách hàng, giúp tăng lợi nhuận, giữ chân các khách hàng quan trọng, cũng như thực hiện các chiến dịch, chiến lược marketing hiệu quả hơn. (Thịnh, n.d.)

### Các phương pháp phân cụm phổ biến (SlideShare, n.d.)

Phân cụm dựa trên phân vùng (Partitioning Clustering): Chia tập dữ liệu thành k cụm sao cho tối ưu một tiêu chí nhất định, thường là giảm khoảng cách trong cụm và tăng khoảng cách giữa các cụm. Phương pháp phổ biến nhất là K-Means, sử dụng trung bình cộng làm đại diện cụm, và K-Medoids, thay vì dùng trung bình, chọn một điểm thực tế làm trung tâm cụm để giảm ảnh hưởng của ngoại lệ.

Phân cụm phân cấp (Hierarchical Clustering): Xây dựng cấu trúc cây phân cấp của dữ liệu dựa trên một ma trận khoảng cách. Có hai hướng chính: Agglomerative (gộp cụm) – bắt đầu từ các điểm riêng lẻ và hợp nhất dần thành cụm lớn hơn; Divisive (chia cụm) – bắt đầu từ một cụm lớn và chia nhỏ dần. Kết quả được biểu diễn dưới dạng dendrogram giúp người dùng có thể chọn mức phân cụm phù hợp.

Phân cụm dựa trên mật độ (Density-based Clustering): Nhóm các điểm dữ liệu dựa vào vùng có mật độ cao, giúp xác định các cụm có hình dạng bất kỳ và loại bỏ nhiễu. DBSCAN là phương pháp phổ biến nhất, trong đó cụm được hình thành từ các điểm lõi có đủ số lượng điểm lân cận trong một bán kính nhất định. OPTICS mở rộng DBSCAN để xử lý dữ liệu có mật độ thay đổi.

Phân cụm dựa trên mô hình (Model-based Clustering): Giả định rằng dữ liệu tuân theo một số mô hình thống kê, phổ biến nhất là Gaussian Mixture Model (GMM), trong đó mỗi cụm được biểu diễn bằng một phân phối Gaussian. Phương pháp này thường sử dụng thuật toán kỳ vọng tối đa (Expectation-Maximization - EM) để tìm cụm tốt nhất.

Phân cụm dựa trên lưới (Grid-based Clustering): Chia không gian dữ liệu thành các ô lưới nhỏ, sau đó nhóm các ô có mật độ cao để xác định cụm. Cách tiếp cận này giúp giảm độ phức tạp tính toán, đặc biệt với dữ liệu lớn. Các thuật toán tiêu biểu gồm STING, sử dụng cấu trúc cây phân cấp để tóm gọn thông tin cụm, và CLIQUE, kết hợp phương pháp dựa trên lưới với phân cụm theo không gian con để xử lý dữ liệu nhiều chiều.

## Phương pháp RFM trong phân cum khách hàng

### Định nghĩa Recency, Frequency, Monetary (Bizfly, n.d.)

Recency (Thời gian mua hàng gần nhất)

* Recency trong mô hình RFM là thước đo thời gian từ lần mua hàng gần nhất của khách hàng. Đây là một yếu tố quan trọng để đánh giá mức độ tương tác gần nhất của khách hàng với doanh nghiệp.
* Một khách hàng được xem là có giá trị cao hơn nếu lần mua hàng gần nhất của họ diễn ra gần đây hơn. Ví dụ, một khách hàng mới mua hàng trong tháng trước sẽ có mức Recency cao hơn so với một khách hàng mua hàng cách đây một năm. Qua đó doanh nghiệp có thể dễ dàng tập trung vào việc tương tác và tiếp cận các khách hàng theo cách phù hợp, để tăng cường quan hệ và gia tăng giá trị từ mỗi khách hàng.
* Recency phân loại khách hàng thành các nhóm khác nhau như:
* Khách hàng mới: Lần mua hàng gần nhất xảy ra gần đây và có thể được coi là tiềm năng cho việc tương tác và tiếp thị tiếp theo.
* Khách hàng thường xuyên: Lần mua hàng gần nhất xảy ra không quá lâu và đánh dấu một mức độ tương tác liên tục với doanh nghiệp.
* Khách hàng không hoạt động: Lần mua hàng gần nhất xảy ra rất lâu và khách hàng không có hoạt động mới. Nhóm này thường ít tương tác hoặc không mấy quan tâm đến doanh nghiệp.

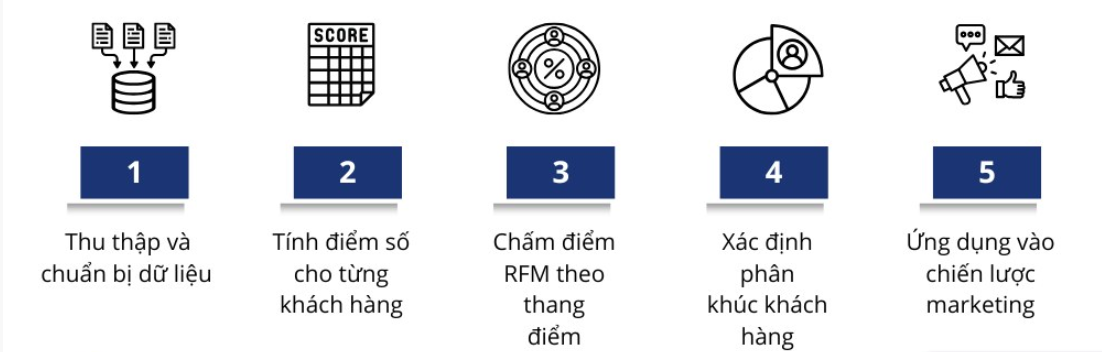
Frequency (Tần suất mua hàng)

* Frequency đo số lần mua hàng của khách trong một khoảng thời gian nhất định, đánh giá mức độ tương tác lặp lại của khách hàng với doanh nghiệp. Một khách hàng được coi là quan trọng và có giá trị cao nếu họ mua hàng thường xuyên hơn. Ví dụ, một khách hàng mua hàng mỗi tháng sẽ có mức Frequency cao hơn so với một khách hàng chỉ mua hàng một lần trong năm.
* Frequency trong mô hình RFM được sử dụng để phân loại khách hàng thành các nhóm khác nhau như:
* Khách hàng thường xuyên: Mua hàng với tần suất cao trong một khoảng thời gian nhất định. Đây là nhóm khách hàng quan trọng và có tiềm năng mang lại doanh thu lớn.
* Khách hàng định kỳ: Mua hàng với tần suất ổn định trong một khoảng thời gian nhất định, không quá thường xuyên. Đây là nhóm khách hàng có thể duy trì sự ổn định.
* Khách hàng ngẫu nhiên: Mua hàng không đều đặn, không có tần suất cụ thể. Đây là nhóm khách hàng khá khó dự đoán và cần quan tâm để tăng cường tương tác.
* Khách hàng không thường xuyên: Mua hàng rất ít lần hoặc không mua hàng trong một thời gian dài. Với nhóm khách hàng không phải ưu tiên tiếp thị, và có thể yêu cầu nỗ lực đặc biệt để thúc đẩy họ tăng tần suất mua hàng.
* Một tần suất mua hàng lý tưởng tương đương với [khách hàng quay lại mua hàng đều đặn](https://bizfly.vn/techblog/lam-the-nao-de-khach-hang-quay-lai.html) là yếu tố quan trọng trong chiến lược kinh doanh. Khách hàng có thói quen mua sắm định kỳ không chỉ tạo ra nguồn doanh thu ổn định mà còn chứng tỏ sự hài lòng và độ trung thành của họ đối với thương hiệu hoặc cửa hàng. Việc duy trì một mức tần suất mua sắm ổn định từ phía khách hàng giúp xây dựng một mối quan hệ lâu dài và tăng cường [giá trị khách hàng](https://bizfly.vn/techblog/gia-tri-khach-hang.html).

Monetary (Giá trị tiền mỗi lần mua hàng)

* Trong mô hình RFM, Monetary đo giá trị tiền hoặc giá trị đơn hàng mà khách hàng đã chi tiêu cho mỗi lần mua hàng, thể hiện mức độ giá trị mà khách hàng mang lại cho doanh nghiệp. Giá trị tiền được đo bằng tổng số tiền mà khách hàng đã chi trả hoặc tổng giá trị đơn hàng mỗi lần giao dịch.
* Monetary phân loại khách hàng thành các nhóm khác nhau:
* Khách hàng có giá trị cao: Chi tiêu nhiều tiền trong mỗi lần mua hàng, đóng góp lớn vào doanh thu của doanh nghiệp. Đây là nhóm khách hàng quan trọng và cần được tạo điều kiện thuận lợi để duy trì mức chi tiêu cao.
* Khách hàng có giá trị trung bình: Chi tiêu một số tiền vừa phải trong mỗi lần mua hàng. Đây là nhóm khách hàng có tiềm năng phát triển và có thể tăng giá trị bằng cách tăng cường tương tác và tiếp thị.
* Khách hàng có giá trị thấp: Chi tiêu ít tiền trong mỗi lần mua hàng. Đây là nhóm khách hàng có thể không mang lại lợi nhuận lớn cho doanh nghiệp và có thể cần nỗ lực để tăng cường giá trị mỗi giao dịch.

### Cách tính điểm RFM từ dữ liệu khách hàng



Hình Các bước tính RFM

Hình: Các bước tính RFM

### Bước 1: Thu thập và chuẩn bị dữ liệu giao dịch

* Trước tiên cần thu thập dữ liệu lịch sử mua hàng của khách hàng từ hệ thống CRM hoặc cơ sở dữ liệu bán hàng. Dữ liệu cần có các thông tin sau:
* Customer ID: Mã định danh của từng khách hàng.
* Transaction Date: Ngày thực hiện giao dịch.
* Transaction Amount: Giá trị đơn hàng.

### Bước 2: Tính toán điểm số RFM cho từng khách hàng

* Recency (R) – Gần đây
* Đo lường khoảng thời gian từ lần mua gần nhất đến ngày phân tích.
* Công thức: R = Ngày phân tích – Ngày giao dịch gần nhất
* Giả định: Khách hàng mua gần đây có xu hướng quay lại cao hơn.
* Frequency (F) – Tần suất mua hàng
* Tính tổng số lần khách hàng đã mua hàng trong khoảng thời gian nhất định.
* Giả định: Khách hàng mua hàng nhiều lần có giá trị cao hơn khách mua hàng ít lần.
* Monetary (M) – Tổng chi tiêu
* Tổng số tiền mà khách hàng đã chi tiêu.
* Công thức: M = ∑ Giá trị các đơn hàng của khách hàng
* Giả định: Khách hàng chi tiêu nhiều có giá trị cao hơn.

### Bước 3: Chấm điểm RFM theo thang điểm

* Sau khi tính toán các giá trị R, F, M, bạn cần phân nhóm khách hàng theo thang điểm (thường là 1 – 5).
* Chia nhỏ dữ liệu: Sử dụng phương pháp quintile (phân vị 20%) để chia khách hàng thành 5 nhóm cho mỗi tiêu chí (1 là thấp nhất, 5 là cao nhất).

Ví dụ xếp hạng:

* Recency:
* 5 điểm: Khách hàng mua hàng trong 7 ngày qua.
* 1 điểm: Khách hàng mua hàng hơn 6 tháng trước.
* Frequency:
* 5 điểm: Mua hàng 10+ lần.
* 1 điểm: Chỉ mua 1 lần.
* Monetary:
* 5 điểm: Tổng chi tiêu trên 10 triệu VND.
* 1 điểm: Tổng chi tiêu dưới 500 nghìn VND.

Bước 4: Xác định phân khúc khách hàng dựa trên RFM

* Sau khi chấm điểm, mỗi khách hàng sẽ có một bộ ba số điểm RFM (ví dụ: R = 5, F = 3, M = 4). Dựa vào điểm số này, ta chia khách hàng thành các nhóm khác nhau như:
* Champions (5-5-5): Khách hàng mua gần đây, thường xuyên và chi tiêu cao → Cần giữ chân bằng ưu đãi VIP.
* Loyal Customers (4-5-4): Khách hàng trung thành, phản hồi tích cực với chương trình khuyến mãi → Bán thêm sản phẩm cao cấp.
* At Risk (1-4-5): Từng mua nhiều nhưng đã lâu không quay lại → Gửi email cá nhân hóa để tái kích hoạt.
* Lost (1-1-1): Khách hàng ít tương tác, chi tiêu thấp → Không ưu tiên, có thể loại bỏ khỏi chiến dịch chính.

### Bước 5: Ứng dụng RFM vào chiến lược marketing

* Sau khi phân loại khách hàng, doanh nghiệp có thể sử dụng RFM để tối ưu hóa chiến lược tiếp thị như:
* Email marketing cá nhân hóa: Gửi thông điệp phù hợp với từng nhóm khách hàng.
* Chương trình khách hàng thân thiết: Tạo ưu đãi đặc biệt cho nhóm Loyal Customers & Champions.
* Remarketing & Retargeting Ads: Nhắm mục tiêu quảng cáo đến nhóm khách hàng có giá trị cao nhưng có nguy cơ rời bỏ.
* Ưu đãi độc quyền & Flash Sale: Dành cho nhóm khách hàng sắp rời đi để kéo họ quay lại.

### Ứng dụng của RFM trong phân tích dữ liệu khách hàng

Phân loại khách hàng nhanh chóng, dễ hiểu: RFM sử dụng ba tiêu chí đơn giản nhưng hiệu quả để phân nhóm khách hàng. Không cần công nghệ phức tạp, doanh nghiệp có thể nhanh chóng xác định nhóm khách hàng trung thành, khách hàng tiềm năng và khách hàng có nguy cơ rời bỏ.

Tối ưu chiến lược tiếp thị: Dựa trên phân loại khách hàng, doanh nghiệp có thể cá nhân hóa các chiến dịch marketing như ưu đãi cho khách hàng trung thành, nhắc nhở khách hàng lâu chưa mua sắm hoặc khuyến khích khách hàng tiềm năng tăng tần suất mua hàng.

Cải thiện tỷ lệ giữ chân khách hàng: Bằng cách tập trung vào nhóm khách hàng có giá trị cao và tái kích hoạt những khách hàng ít tương tác, doanh nghiệp có thể tăng tỷ lệ giữ chân, giảm chi phí tìm kiếm khách hàng mới.

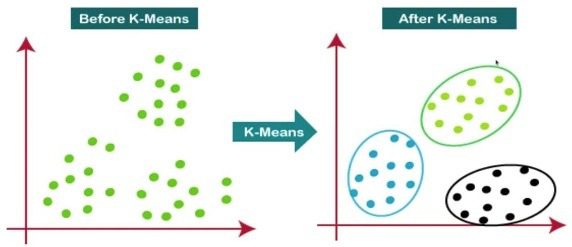
Tối ưu chi phí và nâng cao doanh thu: RFM giúp doanh nghiệp đầu tư hiệu quả hơn vào các chiến dịch tiếp thị. Thay vì chi tiêu dàn trải, doanh nghiệp có thể tập trung vào nhóm khách hàng có khả năng mang lại doanh thu cao nhất.

## Thuật toán K-Means

### Nguyên lý hoạt động của K-Means

K-Means Clustering là một thuật toán phân cụm phổ biến, thường được sử dụng trong các ứng dụng học máy (Machine Learning) và khai phá dữ liệu (Data Mining). Mục tiêu chính của K-Means là chia một tập hợp các điểm dữ liệu thành k cụm (clusters) sao cho các điểm trong cùng một cụm có sự tương đồng cao nhất với nhau và khác biệt tối đa với các cụm khác.

Trong thuật toán K-Means, mỗi cụm được đại diện bởi một centroid, là giá trị trung bình của các điểm dữ liệu trong cụm đó. Thuật toán sẽ liên tục điều chỉnh các centroid và tái phân cụm các điểm dữ liệu cho đến khi đạt được sự hội tụ (convergence).



Hình : Minh họa thuật toán k-mean

### Bước 1: Chọn số cụm K và khởi tạo centroid ban đầu

Xác định số lượng cụm K cần phân chia và chọn ngẫu nhiên K điểm dữ liệu để làm trung tâm ban đầu (centroid) cho mỗi cụm. Việc lựa chọn K có thể dựa vào kinh nghiệm hoặc áp dụng các phương pháp như Elbow để tìm ra giá trị phù hợp

 Bước 2: Gán từng điểm dữ liệu vào cụm gần nhất

Tính khoảng cách từ mỗi điểm dữ liệu đến các centroid hiện tại bằng một tiêu chí nhất định, thường là khoảng cách Euclidean. Mỗi điểm sẽ được gán vào cụm có centroid gần nhất, tạo thành các nhóm dữ liệu tạm thời dựa trên vị trí tương đối của chúng.

### Bước 3: Cập nhật centroid cho mỗi cụm

Sau khi tất cả các điểm đã được gán vào các cụm, tính lại vị trí centroid mới cho từng cụm. Centroid mới được xác định bằng cách lấy trung bình cộng của tất cả các điểm thuộc cùng một cụm. Các centroid này sẽ là đại diện mới cho mỗi cụm trong vòng lặp tiếp theo.

### Bước 4: Lặp lại quá trình cho đến khi các cụm ổn định

Tiếp tục lặp lại bước 2 và bước 3 với các centroid mới. Thuật toán dừng lại khi các điểm không còn thay đổi cụm. Khi đó, các cụm được xem là ổn định và kết quả phân nhóm được giữ nguyên.

### Phương pháp Elbow để xác định số cụm tối ưu

Trong thuật toán k-Means thì chúng ta cần phải xác định trước số cụm. Câu hỏi đặt ra là đâu là số lượng cụm cần phân chia tốt nhất đối với một bộ dữ liệu cụ thể? Phương pháp Elbow là một cách giúp ta lựa chọn được số lượng các cụm phù hợp dựa vào đồ thị trực quan hoá bằng cách nhìn vào sự suy giảm của hàm biến dạng và lựa chọn ra điểm khuỷu tay (elbow point).

Điểm khuỷu tay là điểm mà ở đó tốc độ suy giảm của hàm biến dạng sẽ thay đổi nhiều nhất. Tức là kể từ sau vị trí này thì gia tăng thêm số lượng cụm cũng không giúp hàm biến dạng giảm đáng kể.

WSS (Within-Cluster Sum of Squares) và BSS (Between-Cluster Sum of Squares) là hai chỉ số quan trọng liên quan đến việc xác định số lượng cụm tối ưu (K) trong thuật toán K-Means

### Ưu và nhược điểm của thuật toán K-Means

### Ưu điểm của K-Means Clustering

#### Đơn giản và dễ hiểu: K-Means là một trong những thuật toán phân cụm dễ hiểu nhất, với quy trình hoạt động rõ ràng và dễ triển khai.

#### Hiệu quả: Thuật toán này có thể xử lý một lượng lớn dữ liệu với chi phí tính toán thấp, đặc biệt là khi số lượng cụm k nhỏ.

* Khả năng mở rộng: K-Means có thể mở rộng tốt với dữ liệu lớn, nhờ vào tính chất tuyến tính của nó. Nó có thể được áp dụng trên các tập dữ liệu lớn với hàng triệu điểm dữ liệu.
* Linh hoạt: Có thể áp dụng K-Means cho nhiều loại dữ liệu khác nhau, bao gồm cả dữ liệu số và dữ liệu danh mục (categorical).

### Nhược điểm của K-Means Clustering

#### Số cụm phải được xác định trước: Một trong những hạn chế lớn nhất của K-Means là yêu cầu người dùng phải xác định số lượng cụm k trước khi chạy thuật toán. Điều này có thể khó khăn khi không biết trước số cụm tối ưu.

#### Nhạy cảm với vị trí khởi tạo centroid: Kết quả của K-Means phụ thuộc rất nhiều vào việc khởi tạo centroid ban đầu. Khởi tạo kém có thể dẫn đến hội tụ tại một cực trị cục bộ không tối ưu.

#### Chỉ nhận diện các cụm hình cầu: K-Means hoạt động tốt với các cụm có hình dạng gần như cầu và kích thước tương đồng. Nó không thể xử lý tốt các cụm có hình dạng phi tuyến hoặc không đồng đều.

#### Nhạy cảm với nhiễu: K-Means rất nhạy cảm với các điểm dữ liệu ngoại lai (outliers), vì các outliers có thể kéo centroid ra khỏi vị trí trung tâm của cụm thực sự.

#### Khi nào nên sử dụng K-Means Clustering

K-Means Clustering là một lựa chọn tốt trong các trường hợp sau:

#### Dữ liệu có cấu trúc đơn giản: Khi dữ liệu có cấu trúc phân cụm đơn giản với các cụm hình cầu hoặc gần hình cầu, K-Means là một công cụ mạnh mẽ.

#### Số lượng cụm được biết trước: Khi bạn đã biết hoặc có thể ước lượng chính xác số cụm k, K-Means có thể nhanh chóng phân nhóm dữ liệu.

#### Phân tích dữ liệu sơ bộ

* Khi cần phân tích dữ liệu sơ bộ để khám phá các mẫu ẩn, K-Means có thể cung cấp cái nhìn nhanh chóng và rõ ràng về cấu trúc của dữ liệu.

#### Yêu cầu thời gian tính toán thấp

* Khi thời gian và tài nguyên tính toán hạn chế, K-Means cung cấp giải pháp nhanh chóng với hiệu suất tính toán cao.

## Kiểm định sự khác biệt giữa các cụm

### Tổng quan về kiểm định giả thuyết trong phân tích cụm

Kiểm định giả thuyết là một kỹ thuật thống kê nhằm đưa ra quyết định hoặc kết luận về một quần thể thông qua việc phân tích dữ liệu mẫu, dựa trên một giả định ban đầu (giả thuyết không – H₀) và một giả thuyết thay thế (H₁), bao gồm hai phương pháp phổ biến là T-test và F-test.

* Kiểm định T là một thống kê suy luận được sử dụng để xác định xem có sự khác biệt đáng kể giữa các giá trị trung bình của hai nhóm và mối quan hệ của chúng như thế nào. Kiểm định T được sử dụng khi các tập dữ liệu tuân theo phân phối chuẩn và có phương sai chưa biết, như tập dữ liệu được ghi lại từ việc tung đồng xu 100 lần. Kiểm định T là một kiểm định được sử dụng để kiểm định giả thuyết trong thống kê và sử dụng thống kê T, giá trị phân phối T và bậc tự do để xác định ý nghĩa thống kê. (Kenton, 2023)
* Kiểm định F(F-test) có thể được định nghĩa là một kiểm định sử dụng thống kê kiểm định f để kiểm tra xem phương sai của hai mẫu (hoặc quần thể) có bằng cùng một giá trị hay không. Để tiến hành kiểm định f, quần thể phải tuân theo phân phối f và các mẫu phải là các biến cố độc lập. Khi tiến hành kiểm định giả thuyết, nếu kết quả của kiểm định f có ý nghĩa thống kê thì có thể bác bỏ giả thuyết không, nếu không thì không thể bác bỏ. (Team, n.d.)
* HSD Tukey (hay quy trình Tukey ), còn được gọi là bài kiểm tra Tukey's Honest Significant Difference, là một bài kiểm tra hậu kiểm dựa trên phân phối phạm vi studentized. Kiểm định Anova  có thể cho bạn biết kết quả của bạn có đáng kể hay không , nhưng nó sẽ không cho bạn biết chính xác những khác biệt đó nằm ở đâu. Sau khi bạn chạy ANOVA và tìm thấy kết quả đáng kể, sau đó bạn có thể chạy Tukey's HSD để tìm ra các giá trị trung bình của các nhóm cụ thể nào (so với nhau) là khác nhau. Bài kiểm tra so sánh tất cả các cặp giá trị trung bình có thể có. (Glen, n.d.)

### Các phương pháp kiểm định

#### Nguyên lý của kiểm định ANOVA (Luanvan1080, n.d.)

1. Đặt giả thuyết kiểm định:

* Giả thuyết không (H₀): Các nhóm có trung bình bằng nhau, tức là không có sự khác biệt đáng kể giữa các nhóm.
* Giả thuyết đối (H₁): Ít nhất một nhóm có trung bình khác biệt so với các nhóm còn lại.

1. Kiểm định giả thuyết Anova

* Phân tích phương sai
* ANOVA đo lường sự khác biệt giữa các cụm bằng cách so sánh phương sai giữa các nhóm (Between-group variance) và phương sai trong từng nhóm (Within-group variance).
* Nếu phương sai giữa các nhóm lớn hơn đáng kể so với phương sai trong nhóm, điều đó chứng tỏ các nhóm có sự khác biệt thực sự.

#### Giá trị F (F-statistic) và quyết định kiểm định

* Chỉ số F (F-statistic) được tính bằng tỉ số giữa phương sai giữa các nhóm và phương sai trong nhóm:​
* Nếu F lớn, nghĩa là sự khác biệt giữa các nhóm lớn hơn sự biến động ngẫu nhiên trong nhóm.
* Nếu F nhỏ, sự khác biệt giữa các nhóm có thể do ngẫu nhiên.
* Quy tắc bác bỏ H₀:

Nếu p-value < mức ý nghĩa (thường là 0.05), ta bác bỏ H₀ và kết luận có ít nhất một nhóm khác biệt đáng kể.

Ngược lại, nếu p-value ≥ 0.05, không đủ bằng chứng để bác bỏ H₀, nghĩa là các nhóm có trung bình tương đương nhau.

#### Ứng dụng của kiểm định ANOVA trong phân tích cụm (Exchange., n.d.)

Trong phân tích cụm, kiểm định ANOVA (Analysis of Variance) được sử dụng như một công cụ hỗ trợ nhằm đánh giá mức độ khác biệt giữa các cụm dựa trên các biến định lượng. Sau khi thực hiện quá trình phân cụm (chẳng hạn như bằng thuật toán K-means hoặc phân cụm phân cấp), ANOVA được áp dụng để kiểm tra giả thuyết rằng các cụm có sự khác biệt có ý nghĩa thống kê về giá trị trung bình của từng biến.

Cụ thể, ANOVA hỗ trợ trong ba khía cạnh chính. Thứ nhất, kiểm định ANOVA giúp xác định liệu các cụm được phân chia có khác biệt đáng kể về mặt thống kê hay không đối với từng biến đầu vào, thông qua giá trị p-value. Nếu p-value nhỏ hơn mức ý nghĩa (thường là 0.05), ta có thể kết luận rằng tồn tại sự khác biệt trung bình đáng kể giữa các cụm đối với biến đang xét.

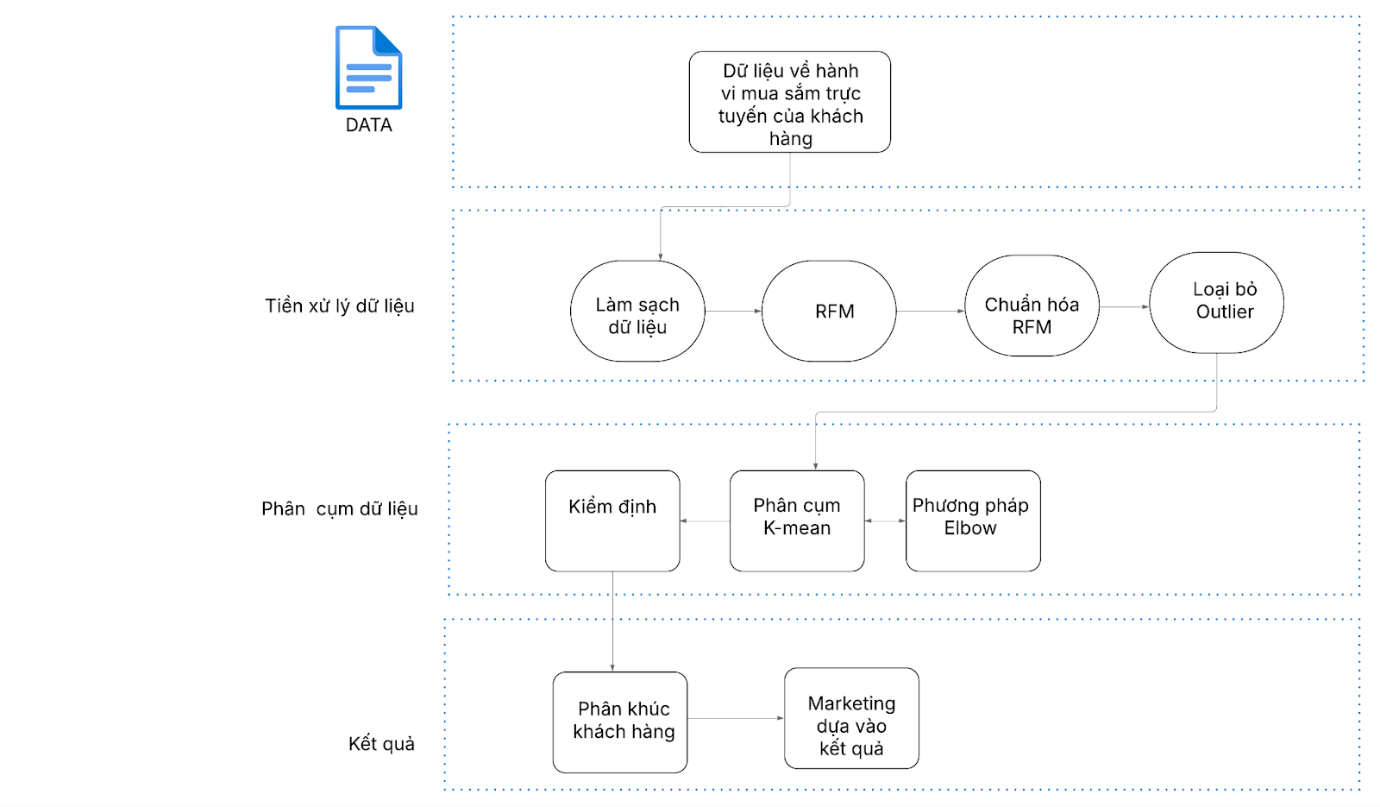
Thứ hai, ANOVA được sử dụng để lựa chọn các biến có vai trò nổi bật trong việc phân biệt giữa các cụm. Những biến có giá trị thống kê F lớn và p-value nhỏ thường là những yếu tố đặc trưng, giúp mô tả hoặc giải thích bản chất của từng cụm.

Thứ ba, kiểm định ANOVA đóng vai trò trong việc đánh giá chất lượng của kết quả phân cụm. Nếu các biến cho thấy sự khác biệt rõ ràng giữa các cụm, điều đó cho thấy mô hình phân cụm có hiệu quả trong việc phân loại dữ liệu. Ngược lại, nếu phần lớn các biến không có sự khác biệt đáng kể, thì có thể cần xem xét lại số lượng cụm hoặc thuật toán phân cụm đã sử dụng.

Tóm lại, mặc dù kiểm định ANOVA không trực tiếp tham gia vào quá trình phân cụm, nhưng nó là một công cụ phân tích hậu kiểm quan trọng, giúp đánh giá độ rõ ràng của các cụm và hỗ trợ diễn giải kết quả một cách có cơ sở thống kê.

# Kết quả thực nghiệm

## Xây dựng framework



Hình : Framework phân cụm và phân tích khách hàng

* Quy trình phân cụm khách hàng được xây dựng nhằm mục tiêu phân loại các nhóm khách hàng có hành vi tiêu dùng tương đồng, từ đó đề xuất các chiến lược marketing phù hợp với từng nhóm. Toàn bộ quá trình được chia thành bốn giai đoạn chính: thu thập và chuẩn bị dữ liệu, tính toán và xử lý đặc trưng, phân cụm dữ liệu, và phân tích – ứng dụng kết quả.

### Giai đoạn 1: Thu thập và chuẩn bị dữ liệu

#### Nguồn dữ liệu

Dữ liệu đầu vào là thông tin liên quan đến hành vi mua sắm trực tuyến của khách hàng. Cụ thể, dữ liệu  bao gồm các trường : mã đơn hàng, mã khách hàng, ngày đặt hàng, số lượng mua, tổng giá trị đơn hàng,… Dữ liệu này được trích xuất từ hệ thống thương mại điện tử hoặc cơ sở dữ liệu bán hàng.

#### Làm sạch dữ liệu

Dữ liệu thực tế thường tồn tại nhiều vấn đề như thiếu sót, dư thừa hoặc không thống nhất định dạng. Do đó, việc làm sạch là bước quan trọng đầu tiên nhằm đảm bảo độ chính xác và nhất quán cho toàn bộ quy trình xử lý sau này. Một số thao tác tiêu biểu bao gồm:

* Loại bỏ các dòng dữ liệu bị trống hoặc không chứa thông tin có giá trị.
* Chuẩn hóa tên cột để tránh lỗi khi xử lý (ví dụ loại bỏ khoảng trắng đầu/cuối).
* Loại bỏ các cột dư thừa không có giá trị phân tích (ví dụ: cột index thừa do sinh ra trong quá trình lưu/đọc dữ liệu).

### Giai đoạn 2: Tính toán và xử lý đặc trưng

#### Tính toán RFM

Sau khi dữ liệu đã được làm sạch, bước tiếp theo là xây dựng các đặc trưng định lượng hành vi của khách hàng thông qua mô hình RFM:

* Recency (R): Khoảng thời gian kể từ lần mua hàng gần nhất đến thời điểm phân tích.
* Frequency (F): Tổng số lần khách hàng đã thực hiện mua hàng.
* Monetary (M): Tổng số tiền mà khách hàng đã chi tiêu trong toàn bộ quá trình mua sắm.

Ba chỉ số này giúp phản ánh toàn diện mức độ tương tác và giá trị của khách hàng đối với doanh nghiệp.

#### Chuẩn hóa RFM

Do các giá trị RFM nằm trên những thang đo khác nhau nên cần được chuẩn hóa để đảm bảo rằng không có yếu tố nào chi phối quá trình phân cụm. Việc chuẩn hóa giúp đảm bảo tính công bằng giữa các đặc trưng, từ đó cải thiện hiệu quả và độ chính xác của mô hình phân cụm.

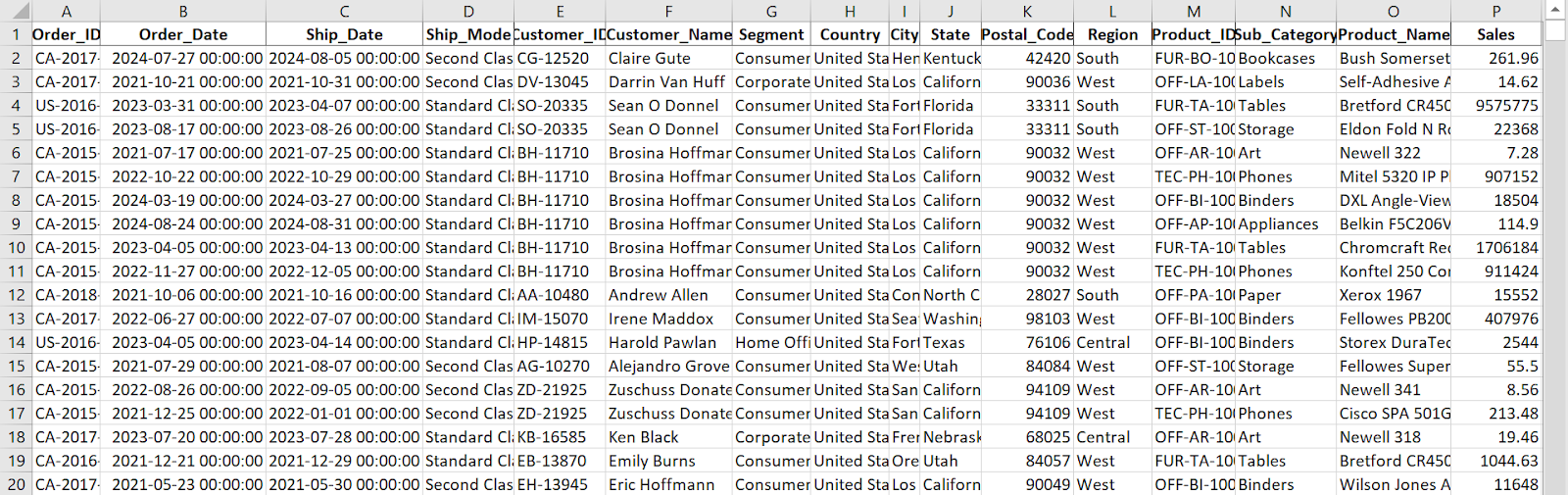
#### Loại bỏ ngoại lệ (outliers)

Một số khách hàng có hành vi mua hàng quá khác biệt (quá thường xuyên, quá ít hoặc chi tiêu quá lớn) có thể gây nhiễu cho kết quả phân cụm. Do đó, bước loại bỏ ngoại lệ được thực hiện nhằm loại bỏ những điểm dữ liệu không đại diện cho xu hướng chung. Phương pháp loại bỏ phổ biến là dựa vào khoảng tứ phân vị (IQR) để xác định và loại bỏ các giá trị nằm ngoài phạm vi hợp lý.

## Chọn dữ liệu (Dataset)

### Giới thiệu về dữ liệu

* Tệp dữ liệu Superstore chứa thông tin về 1.000 giao dịch mua hàng từ 391 khách hàng tại Hoa Kỳ, diễn ra trong khoảng thời gian từ năm 2021 đến 2024. Dữ liệu bao gồm 16 cột, mô tả chi tiết về đơn hàng, khách hàng, sản phẩm và thông tin vận chuyển.
* Cụ thể, mỗi đơn hàng có mã Order\_ID, ngày đặt hàng Order\_Date, ngày giao hàng dự kiến Ship\_Date, và phương thức vận chuyển Ship\_Mode (bao gồm Standard Class, Second Class, First Class, Same Day).
* Khách hàng được nhận diện qua Customer\_ID và Customer\_Name, đồng thời được phân loại theo Segment gồm Consumer, Corporate và Home Office.
* Đơn hàng được giao tại nhiều thành phố và bang khác nhau, với thông tin địa lý như City, State, Postal\_Code, và thuộc một trong bốn khu vực chính: West, East, Central, South. Các phương thức vận chuyển phổ biến là Standard Class và Second Class, phản ánh xu hướng giao hàng trong giai đoạn này.
* Về sản phẩm, dữ liệu ghi nhận Product\_ID, tên sản phẩm Product\_Name, danh mục sản phẩm Sub\_Category như Đồ nội thất, Văn phòng phẩm, Thiết bị công nghệ, cùng với doanh thu tương ứng Sales.
* Tóm lại, tệp dữ liệu này cung cấp một cái nhìn tổng quan về hành vi mua hàng, phân khúc khách hàng và hiệu suất bán hàng của doanh nghiệp.



Hình : Dữ liệu mua hàng

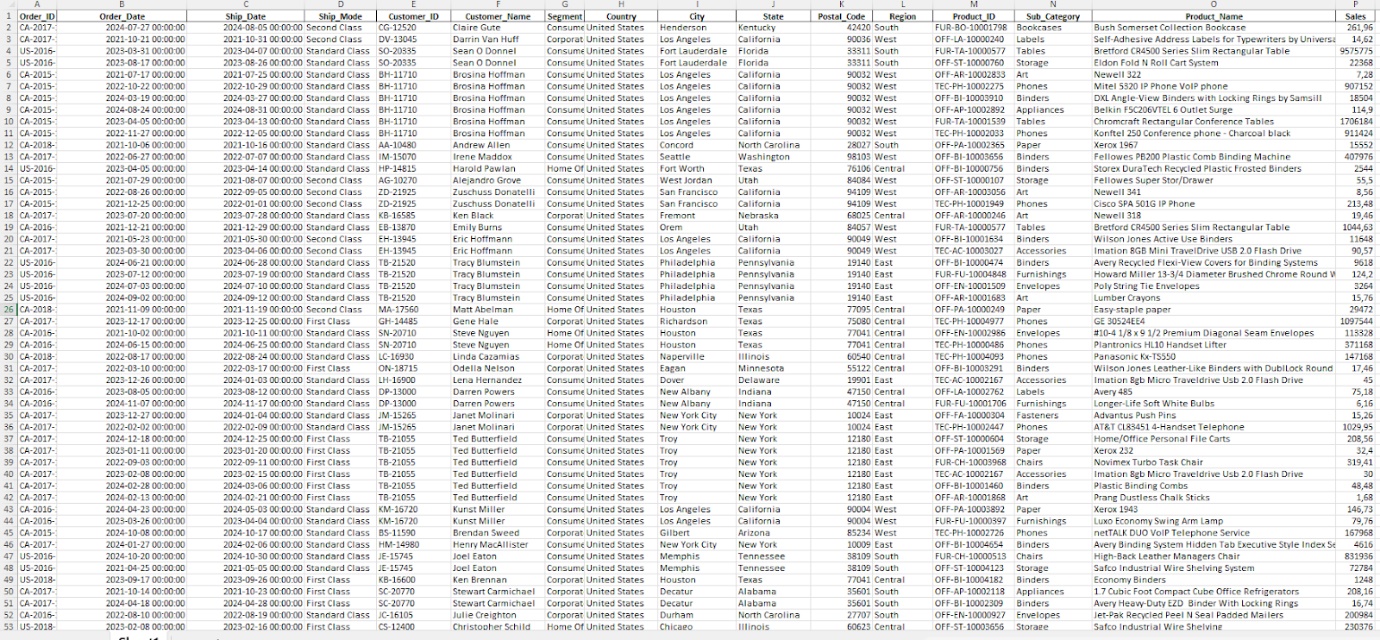
### Mô tả dữ liệu

Bộ dữ liệu Superstore có 16 trường và 1000 bản ghi, bao gồm những trường dưới đây:

* Order\_ID: Mã đơn hàng – định danh duy nhất cho mỗi đơn hàng.
* Order\_Date: Ngày đặt hàng – thể hiện thời điểm khách hàng thực hiện đặt đơn hàng.
* Ship\_Date: Ngày giao hàng – ngày đơn hàng được giao cho khách.
* Ship\_Mode: Phương thức vận chuyển đơn hàng
* Customer\_ID: Mã khách hàng – định danh duy nhất cho mỗi khách hàng.
* Customer\_Name: Tên khách hàng – tên đầy đủ của người mua hàng.
* Segment: Phân khúc khách hàng – thể hiện  các loại khách hàng
* Country: Quốc gia cư trú của khách hàng.
* City: Thành phố cư trú của khách hàng.
* State: Bang cư trú của khách hàng.
* Postal\_Code: Mã bưu chính – mã vùng địa lý của khách hàng.
* Region: Khu vực sinh sống của khách hàng – chia khu vực theo vùng (West, East, Central, South).
* Product\_ID: Mã sản phẩm – định danh duy nhất của mỗi mặt hàng.
* Sub\_Category: Danh mục con của sản phẩm được đặt hàng
* Product\_Name: Tên sản phẩm – mô tả chi tiết của sản phẩm.
* Sales: Doanh thu của sản phẩm.

## Chuẩn bị dữ liệu

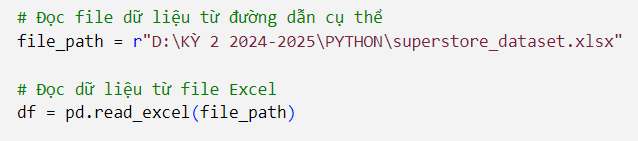
### Triển khai tiền xử lý dữ liệu



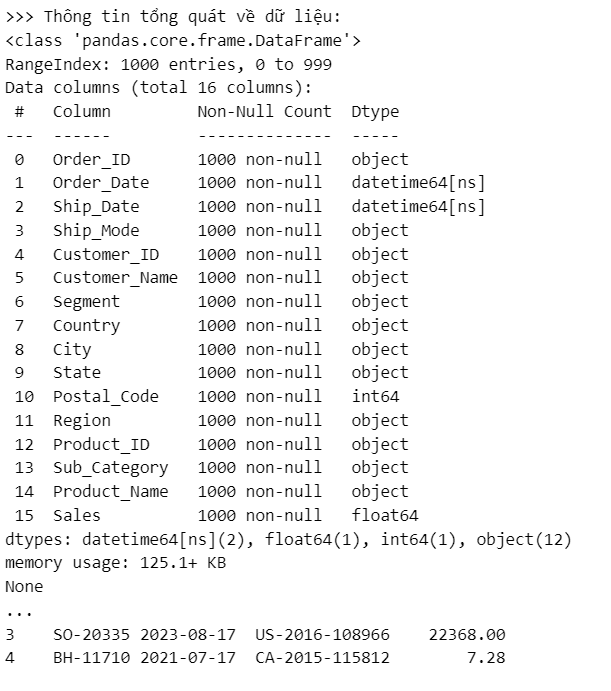
Hình : Dữ liệu trước khi tiến hành tiền xử lí

1. Khám phá dữ liệu

* Đọc file dữ liệu:

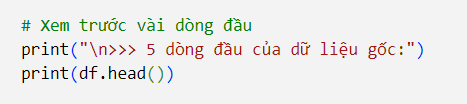


Hình : Thực hiện đọc file dữ liệu

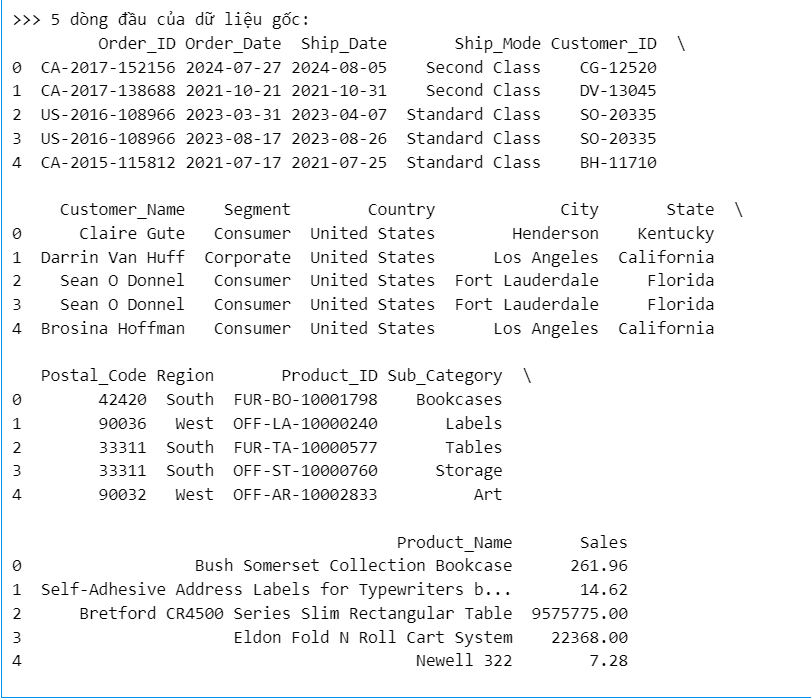


Hình : Tổng quát dữ liệu

* Xem 5 dòng đầu tiên của dữ liệu

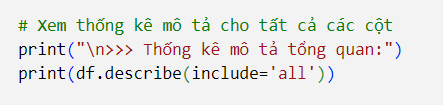


Hình : Thực hiện in 5 dòng đầu tiên của dữ liệu

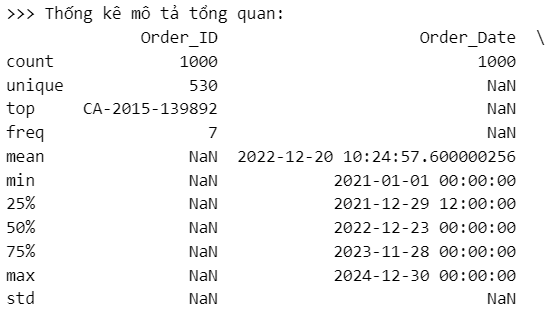
-

Hình : 5 dòng đầu tiên của dữ liệu

* Xem tổng quan và thống kê mô tả của dữ liệu



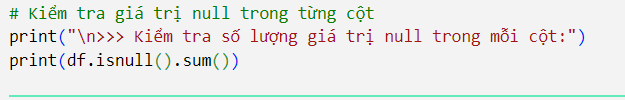
Hình : Thực hiện xem tổng quan và thống kê mô tả của dữ liệu



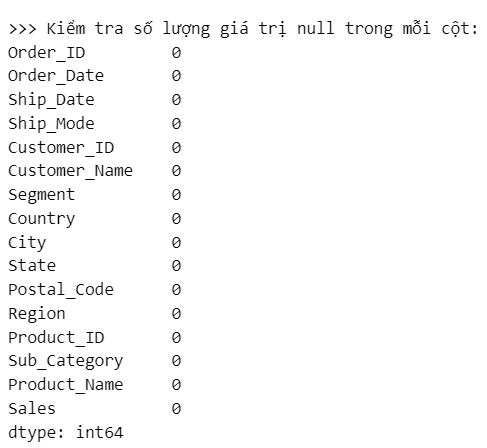
Hình : Thống kê mô tả tổng quan

2. Làm sạch dữ liệu:

- Kiểm tra giá trị null

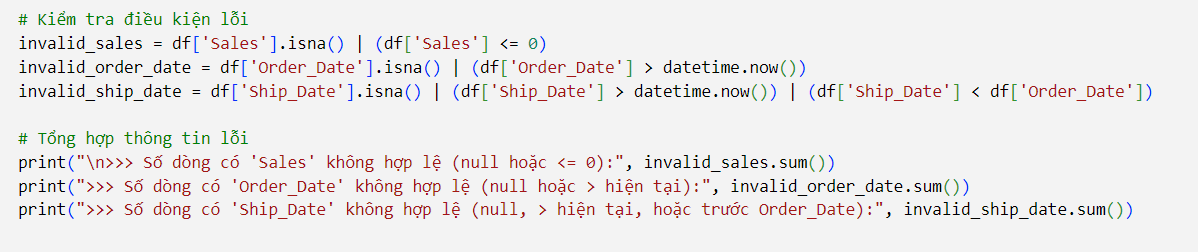


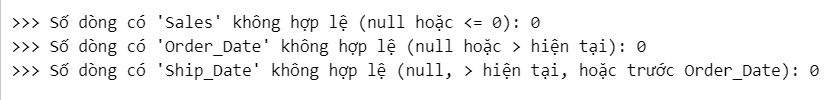
Hình : Thực hiện kiểm tra các giá trị null



Kết quả: Không có giá trị null nào trong dữ liệu

* Kiểm tra dữ liệu không hợp lệ:



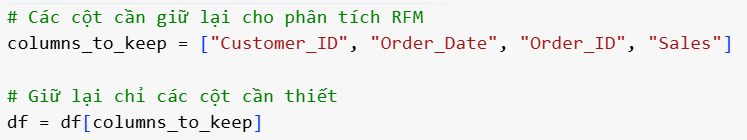
Hình : Thực hiện kiểm tra dữ liệu không hợp lệ  


Hình : Kết quả sau khi kiểm tra tính hợp lệ

Kết quả: Không có dữ liệu không hợp lệ

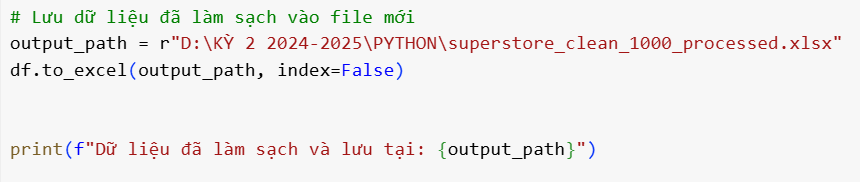
* Loại bỏ các cột không cần thiết để tính toán

Sau khi làm sạch các cột dư thừa, bước tiếp theo là lựa chọn các cột dữ liệu thực sự cần thiết phục vụ cho phân tích RFM. Các chỉ số RFM chỉ cần một số thông tin trọng yếu như mã khách hàng, ngày giao dịch, mã đơn hàng và giá trị đơn hàng.

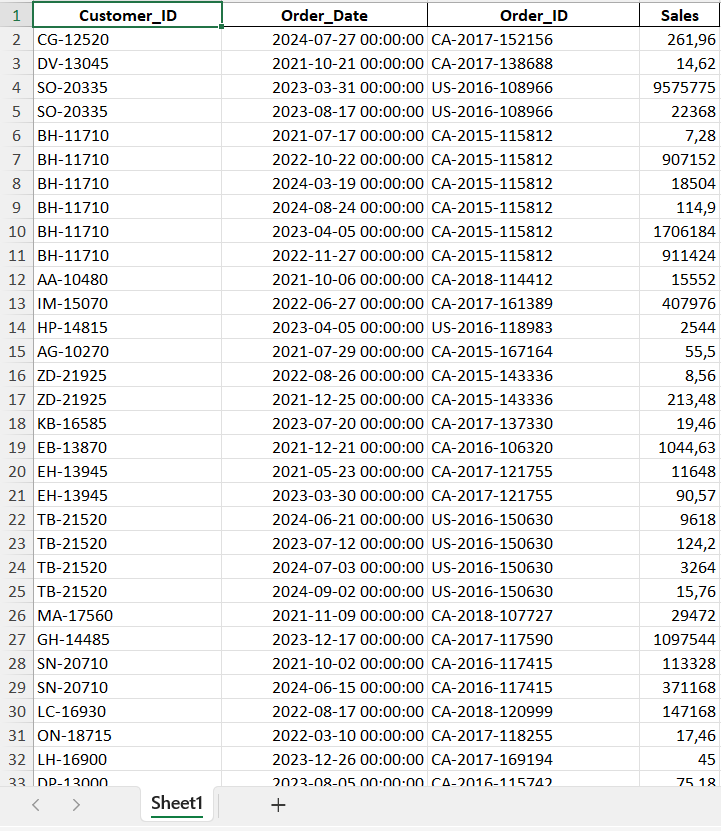


Hình : Giữ lại các thông tin quan trọng cho quá trình tính toán RFM

* In và xuất dữ liệu đã tiền xử lý ra file csv



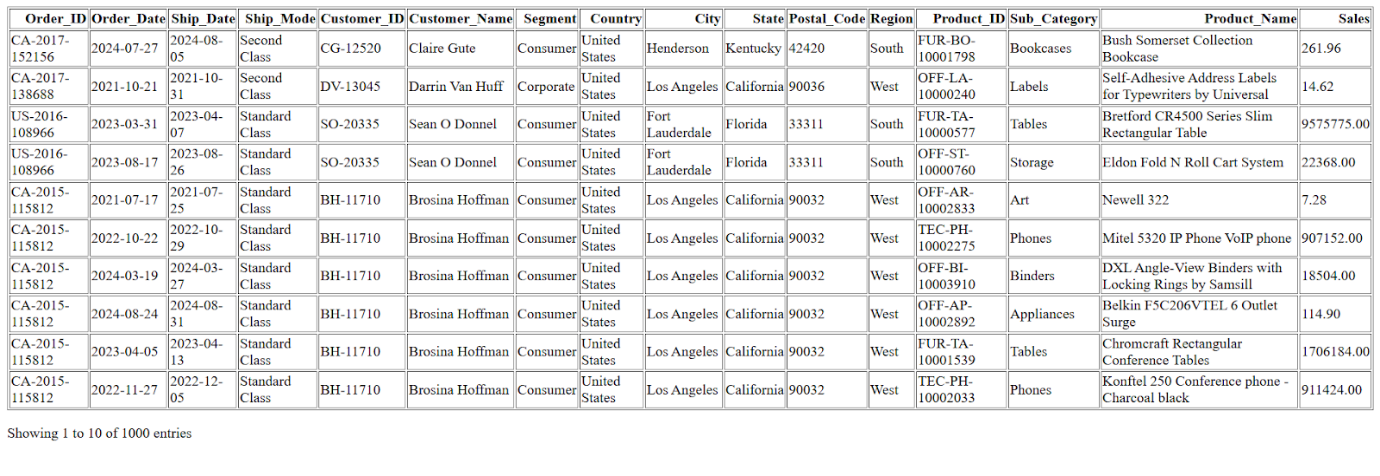
Hình : Thực hiện lưu dữ liệu đã làm sạch



Hình : Hình ảnh dữ liệu sau khi làm sạch

Kết quả:

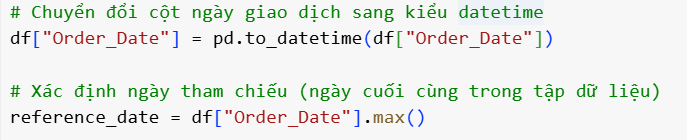
* Dữ liệu đã gọn gàng và nhất quán, chỉ còn lại các cột cần thiết: Customer\_ID, Order\_Date, Order\_ID, và Sales, phù hợp để thực hiện phân tích RFM.
* Các dòng trống và thông tin không cần thiết đã được loại bỏ, giúp tăng hiệu quả xử lý và giảm rủi ro khi phân tích.
* Cột tên đã được chuẩn hóa, tránh lỗi khi truy cập dữ liệu (do khoảng trắng dư thừa trong tên cột).
* Dữ liệu sạch giúp dễ dàng tổng hợp thông tin theo từng khách hàng, từ đó tính toán Recency, Frequency và Monetary một cách chính xác hơn.



Hình : Dữ liệu sau khi làm sạch

### Triển khai tính toán các chỉ số RFM

* Dùng thư viện pandas đây là thư viện phân tích dữ liệu mạnh mẽ nhất trong Python, đặc biệt hữu ích khi làm việc với dữ liệu dạng bảng như file Excel, CSV hoặc cơ sở dữ liệu. Trong phân tích RFM, pandas giúp làm những việc sau một cách dễ dàng và nhanh chóng.
* Bước đầu tiên trong phân tích RFM là chuẩn hóa dữ liệu thời gian. Việc chuyển đổi cột ngày giao dịch sang định dạng datetime giúp dễ dàng tính toán thời gian giữa các lần mua hàng. Đồng thời, cần xác định ngày tham chiếu – thường là ngày cuối cùng trong tập dữ liệu – để làm cơ sở so sánh cho các chỉ số RFM.



Hình : thực hiện chuyển đổi cột ngày giao dịch sang định dạng datetime và xác định ngày tham chiếu

* Tiếp theo là thực hiện tính toán chỉ số RFM gồm:

### Recency (Gần nhất)

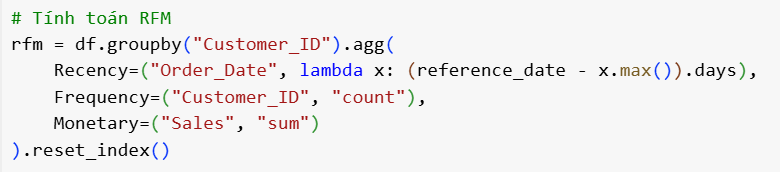
* Mục đích: Đo lường mức độ gần đây nhất mà khách hàng thực hiện mua hàng.
* Cột dữ liệu sử dụng: Order\_Date
* Cách làm:
  + Với mỗi khách hàng, ta lấy ngày đặt hàng mới nhất mà họ từng thực hiện.
  + Sau đó, ta tính khoảng thời gian (số ngày) từ ngày đặt hàng mới nhất đó đến ngày cuối cùng có trong dữ liệu (gọi là ngày tham chiếu).
  + Kết quả cho biết: đã bao nhiêu ngày kể từ lần mua hàng cuối cùng của khách hàng đó.

### Frequency (Tần suất mua hàng)

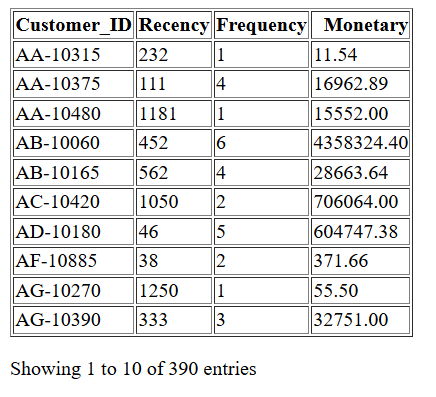
* Mục đích: Đo lường mức độ thường xuyên mà khách hàng mua hàng.
* Cột dữ liệu sử dụng: Customer\_ID
* Cách làm:
  + Vì mỗi dòng trong bảng dữ liệu là một đơn hàng, nên ta đếm xem khách hàng đó xuất hiện bao nhiêu lần trong dữ liệu.
  + Số lần xuất hiện tương đương với số đơn hàng mà khách đó đã thực hiện.
  + Kết quả cho biết: khách hàng đã mua hàng bao nhiêu lần.

### Monetary (Tổng chi tiêu)

* Mục đích: Đo lường tổng số tiền mà khách hàng đã chi trả cho các đơn hàng.
* Cột dữ liệu sử dụng: Sales
* Cách làm:
  + Với mỗi khách hàng, ta cộng lại toàn bộ giá trị trong cột Sales tương ứng với các đơn hàng của họ.
  + Sales chính là cột thể hiện doanh thu từ từng đơn hàng.
  + Kết quả cho biết: khách hàng đã chi tiêu tổng cộng bao nhiêu tiền



Hình : Thực hiện tính toán các chỉ số RFM



Hình : Kết quả sau khi tính RFM

Nhận xét RFM:

### Recency (R):

* Giá trị dao động từ 38 đến 1250 ngày.
* Một số khách hàng mua hàng rất gần thời điểm tham chiếu (ví dụ: AF-10885: 38 ngày).
* Một số khách hàng đã rất lâu không mua hàng (ví dụ: AG-10270: 1250 ngày).

→ Độ phân tán cao, cho thấy có cả khách hàng mới và cũ trong tập dữ liệu.

### Frequency (F):

* Dao động từ 1 đến 6 lần mua.
* Phần lớn khách hàng có tần suất thấp (nhiều khách hàng chỉ mua 1 lần).
* Một số khách hàng nổi bật với tần suất mua cao, ví dụ: AB-10060 (6 lần), AD-10180 (5 lần).

→ Phân bố lệch, có thể tồn tại một số khách hàng trung thành, trong khi đa số mua không thường xuyên.

### Monetary (M):

* Dao động rất lớn: từ 11.54 đến hơn 4 triệu (ví dụ AB-10060: 4,358,324.40).
* Một số khách hàng tạo ra doanh thu rất cao → khách hàng VIP tiềm năng.
* Sự chênh lệch lớn có thể khiến chỉ số Monetary chi phối toàn bộ mô hình nếu không chuẩn hóa.

### Chuẩn hóa chỉ số RFM bằng phương pháp z-score

Trong thực tế, các chỉ số Recency, Frequency, và Monetary thường có quy mô giá trị rất khác nhau:

* Recency: thường nằm trong khoảng vài chục đến vài trăm ngày.
* Frequency: có thể chỉ từ 1–10 lần mua.
* Monetary: có thể dao động từ vài trăm đến hàng ngàn đơn vị tiền tệ.

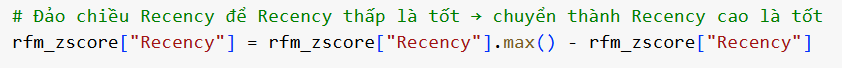
Nếu ta sử dụng trực tiếp các giá trị này để phân tích (ví dụ như phân nhóm khách hàng), thì chỉ số có giá trị lớn (như Monetary) sẽ lấn át hai chỉ số còn lại. Để tránh điều này, ta cần chuẩn hóa các giá trị về cùng một thang đo.

Đầu tiên, sử dụng các thư viện như Pandas để thao tác với dữ liệu dạng bảng, và StandardScaler từ thư viện sklearn để thực hiện chuẩn hóa dữ liệu

### Các bước chuẩn hóa RFM bằng Z-score

#### Bước 1: Đảo chiều Recency

* Trong RFM, Recency càng thấp càng tốt (nghĩa là khách mua hàng gần đây).
* Nhưng Z-score giả định rằng giá trị càng lớn càng "tốt", nên ta đảo chiều Recency:  
   → Lấy giá trị lớn nhất trong cột Recency trừ đi từng giá trị → khi đó, khách nào mua gần đây sẽ có giá trị cao hơn.



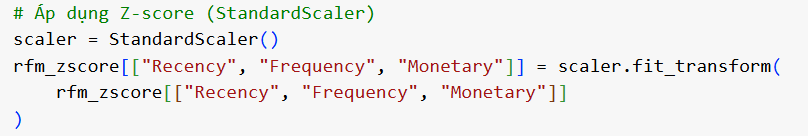
Hình : Thực hiện đảo chiều Recency

#### Bước 2: Áp dụng Z-score

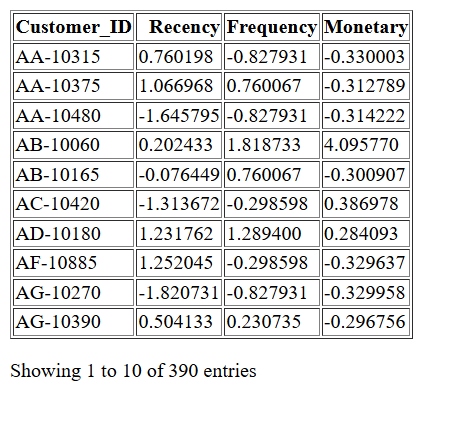
Áp dụng công thức chuẩn hóa Z-score cho từng chỉ số:

* Recency (đã đảo)
* Frequency
* Monetary

Kết quả: 3 cột chuẩn hóa có giá trị Z-score tương ứng.



Hình : Thực hiện chuẩn hóa Z-score



Hình : Kết quả sau khi chuẩn hóa Z-score

Nhận xét RFM:

### Recency (R):

* Recency chuẩn hóa nằm trong khoảng từ -1.82 đến +1.25.
* Các giá trị âm thể hiện khách hàng mua hàng gần đây (tốt).
* Các giá trị dương thể hiện khách hàng lâu rồi không mua (không tốt).

Ví dụ:

* AG-10270: Recency = -1.82 → khách hàng rất cũ (mua đã lâu).
* AF-10885: Recency = +1.25 → khách hàng rất mới (vừa mới mua hàng gần đây).

### Frequency (F):

* Từ khoảng -0.82 đến +1.81.
* Frequency càng cao (giá trị z càng dương) → khách hàng mua hàng nhiều lần.
* Frequency thấp (giá trị z âm) → khách hàng ít giao dịch.

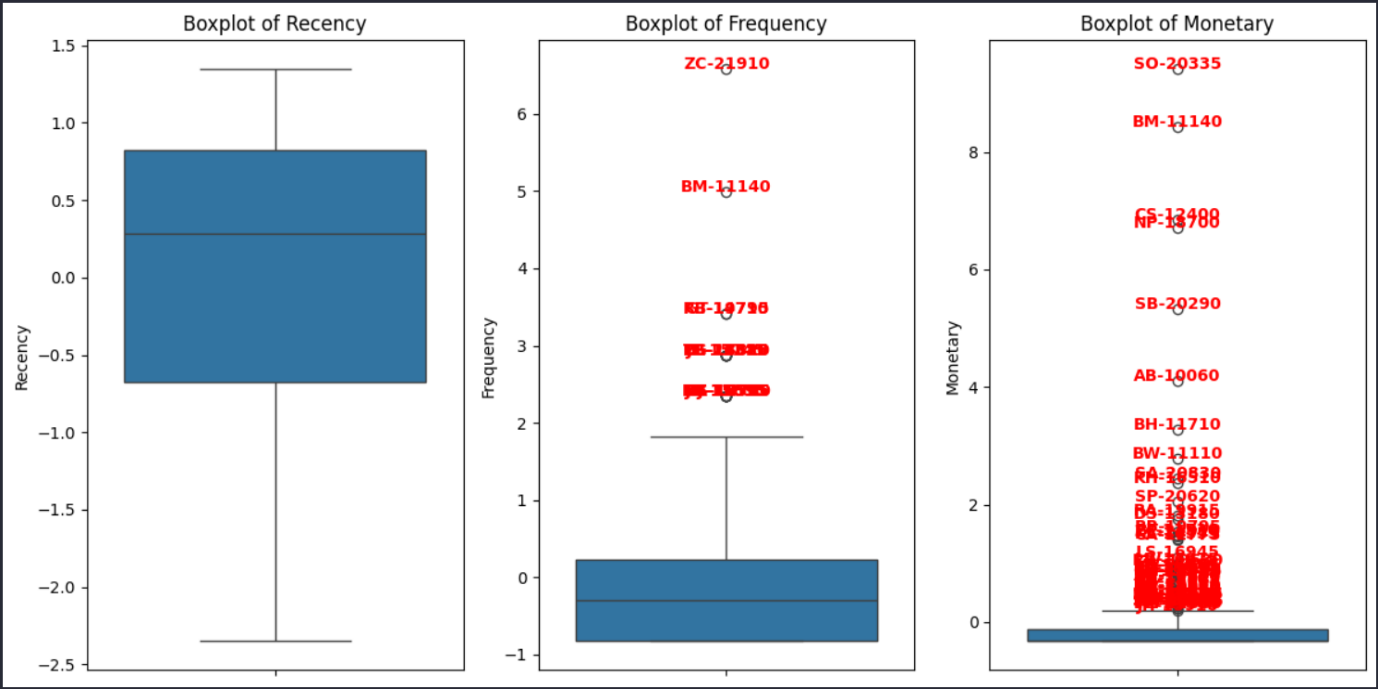
### Monetary (M):

* Từ khoảng -0.33 đến +4.09.
* Chỉ số này biến động mạnh nhất (trước chuẩn hóa có sự chênh lệch cực lớn).
* Giá trị dương cao → khách hàng chi tiêu rất nhiều (VIP).
* Giá trị âm nhẹ → khách hàng chi tiêu thấp.

Tóm lại:

* Recency có sự phân tán vừa phải → phản ánh rõ ràng khoảng thời gian tương tác gần đây của từng khách hàng.
* Frequency hơi lệch trái → phần lớn khách hàng chỉ mua 1-2 lần.
* Monetary có phân bố lệch phải rõ rệt, do một số khách hàng chi tiêu cực kỳ cao tạo ra độ lệch.

### Triển khai trực quan hóa dữ liệu RFM

 Hình : Boxplot thể hiện phân bố các giá trị RFM sau khi chuẩn hóa

Nhận xét:

Biểu đồ boxplot trên thể hiện phân bố của ba chỉ số Recency, Frequency và Monetary sau khi được chuẩn hóa bằng z-score. Đây là ba yếu tố then chốt trong mô hình RFM, được sử dụng để mô tả hành vi mua sắm của khách hàng.

* Recency: Phân bố khá đối xứng, không xuất hiện giá trị ngoại lệ rõ rệt. Điều này cho thấy phần lớn khách hàng có thời gian mua hàng gần đây không chênh lệch quá lớn so với trung bình.
* Frequency: Xuất hiện một số lượng đáng kể các điểm ngoại lệ (outliers) ở phía trên – biểu thị một số khách hàng mua hàng nhiều hơn hẳn so với phần còn lại. Những khách hàng này có thể là khách hàng trung thành, hoặc cũng có thể là các tài khoản bất thường cần được xem xét thêm.
* Monetary: Có nhiều giá trị ngoại lệ, thể hiện sự phân hóa mạnh mẽ về mức độ chi tiêu giữa các khách hàng. Một số khách hàng chi tiêu vượt trội so với phần lớn còn lại – có thể là nhóm khách hàng cao cấp, hoặc cũng là đối tượng dễ gây nhiễu cho mô hình phân cụm nếu không xử lý thích hợp.

Biểu đồ cho thấy hai chỉ số Frequency và Monetary có nhiều ngoại lệ, trong khi Recency khá ổn định. Điều này đặt ra yêu cầu loại bỏ hoặc xử lý các ngoại lệ trước khi tiến hành phân cụm để tránh ảnh hưởng đến kết quả mô hình. Ngoài ra, sự chênh lệch đáng kể trong hành vi chi tiêu và tần suất mua sắm cũng là cơ sở tốt để kỳ vọng mô hình phân cụm có thể tách biệt rõ các nhóm khách hàng khác nhau.

### Loại bỏ ngoại lệ bằng phương pháp IQR

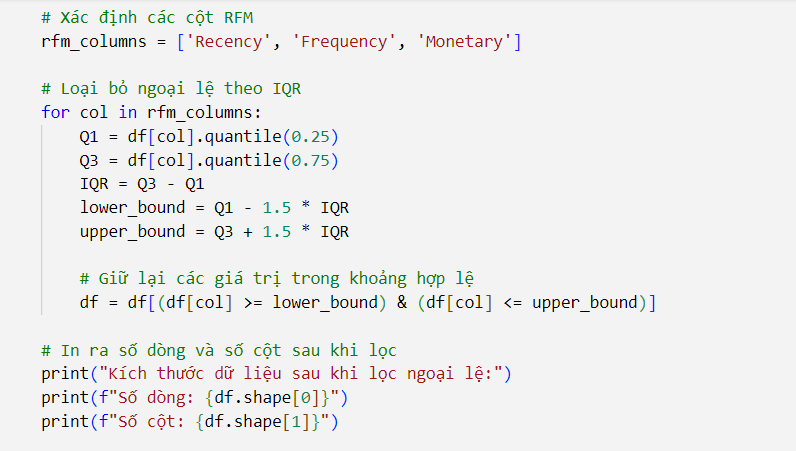
Cách thực hiện:

* Tính giá trị trung bình (mean) và độ lệch chuẩn (standard deviation) của cột Sales.  
   Việc này giúp xác định trung tâm phân phối dữ liệu và đo lường mức độ phân tán của các giá trị xung quanh trung tâm.
* Xác định khoảng giá trị hợp lý bằng công thức:

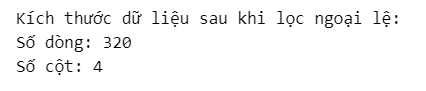
+ Giới hạn dưới (lower) = mean - 3 \* std

+ Giới hạn trên (upper) = mean + 3 \* std  
 Khoảng này thường bao gồm hầu hết các giá trị nếu dữ liệu phân phối gần chuẩn.

* Lọc dữ liệu:  
   Giữ lại các dòng có giá trị Sales nằm trong khoảng từ lower đến upper.  
   Loại bỏ các giá trị quá thấp hoặc quá cao (được xem là ngoại lệ).
* Kiểm tra kết quả:  
   In ra số dòng và số cột của dữ liệu sau khi lọc để biết số lượng dữ liệu còn lại.

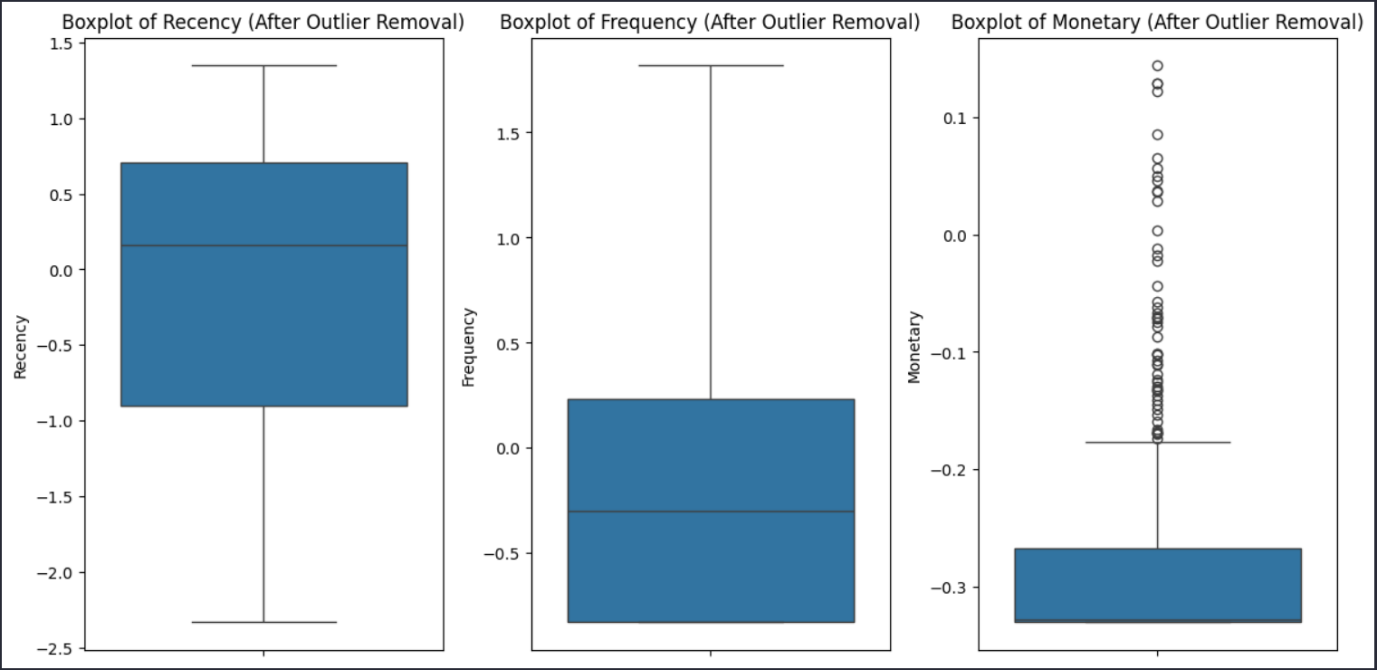


Hình : Thực hiện loại bỏ ngoại lệ



Hình : Kết quả sau khi loại bỏ ngoại lệ

Nhận xét: Sau khi loại bỏ các ngoại lệ (outliers) trong 3 cột Recency, Frequency, và Monetary, thì còn 320 bản ghi (dòng) dữ liệu hợp lệ.

 Hình : Boxplot của Recency, Frequency và Monetary sau khi loại bỏ ngoại lệ bằng phương pháp IQR

Nhận xét:

Biểu đồ boxplot trên thể hiện sự phân phối của ba chỉ số Recency, Frequency, và Monetary sau khi loại bỏ ngoại lệ bằng phương pháp IQR. Có thể thấy rằng việc xử lý này đã giúp dữ liệu trở nên ổn định và ít bị ảnh hưởng bởi các giá trị cực đoan.

* Với Recency, các giá trị đã được thu gọn trong khoảng hợp lý, không còn xuất hiện những điểm quá thấp. Phân phối tương đối đối xứng, trung vị nằm gần giữa hộp, cho thấy dữ liệu không bị lệch nhiều.
* Đối với Frequency, những khách hàng có tần suất mua hàng bất thường (quá cao) đã được loại bỏ, làm cho phân phối trở nên chặt chẽ hơn, giảm sự phân tán và giúp mô hình dễ phân biệt nhóm khách hàng trung thành và không trung thành hơn.
* Riêng Monetary, mặc dù vẫn xuất hiện một vài điểm nằm ngoài (outliers nhẹ), nhưng nhìn chung phần lõi của dữ liệu đã được rút gọn đáng kể. Điều này phản ánh thực tế rằng một số ít khách hàng vẫn có hành vi chi tiêu vượt trội.

Tổng thể, việc loại bỏ ngoại lệ đã giúp làm sạch dữ liệu RFM, giảm nhiễu và tăng độ tin cậy cho các phân tích tiếp theo, đặc biệt là trong quá trình phân cụm khách hàng.

## Phân cụm khách hàng

### Triển khai xác định số cụm tối ưu bằng phương pháp Elbow

A screenshot of a computer code

AI-generated content may be incorrect.

Hình : Thực hiện xác định số cụm tối ưu bằng phương pháp Elbow

A graph with a line

AI-generated content may be incorrect.

Hình : Biểu đồ Elbow để xác định số cụm tối ưu

Biểu đồ cho thấy sự thay đổi của WCSS (Within-Cluster Sum of Squares) theo số cụm k từ 1 đến 10. Quan sát thấy rằng khi số cụm tăng từ 1 đến 4, giá trị WCSS giảm nhanh chóng, chứng tỏ rằng việc tăng số cụm trong giai đoạn này giúp cải thiện rõ rệt độ phân tán trong cụm – các điểm dữ liệu gần tâm cụm hơn, và mô hình phân cụm trở nên hiệu quả hơn.

Tuy nhiên, từ k = 4 trở đi, tốc độ giảm của WCSS bắt đầu chậm lại rõ rệt. Điều này thể hiện rằng việc tiếp tục tăng số cụm sau điểm này không còn mang lại nhiều lợi ích đáng kể trong việc cải thiện độ chính xác của phân cụm. Điểm này được gọi là điểm khuỷu tay (elbow point) – chính là thời điểm mô hình đạt được sự cân bằng giữa hiệu quả và độ phức tạp.

Vì vậy, theo biểu đồ, k = 4 được xem là số cụm tối ưu. Việc chọn số cụm này sẽ giúp mô hình K-Means tạo ra các cụm rõ ràng và hợp lý, đồng thời tránh hiện tượng overfitting khi chia dữ liệu thành quá nhiều cụm không cần thiết.

### Triển khai thuật toán K-Means để phân cụm khách hàng

Bước 1: Phân cụm K-Means với k=4

A screenshot of a computer code

AI-generated content may be incorrect.

Hình : Triển khai phân cụm với k =4

Bước 2: Lưu từng cụm dữ liệu vào file riêng biệt

A white background with black text

AI-generated content may be incorrect.

Hình : Lưu các cụm vào các file riêng

A table of numbers with black text

AI-generated content may be incorrect.

Hình : Khách hàng cụm 0

Khách hàng trong cụm 0 có giá trị Recency và Frequency khá tốt, cho thấy họ đã giao dịch gần đây và thường xuyên. Tuy nhiên, giá trị Monetary âm cho thấy mức chi tiêu của họ không cao. Đây là nhóm khách hàng tiềm năng cần được khuyến khích đầu tư thêm để tăng giá trị chi tiêu.

A table of numbers with black text

AI-generated content may be incorrect.

Hình : Hình ảnh cụm 1

Các khách hàng trong cụm 1 có Recency và Frequency thấp, với tất cả tần suất giao dịch đều giống nhau. Điều này cho thấy họ đã lâu không mua hàng và có thể cần một chiến dịch kích thích quay lại. Mức chi tiêu và giá trị âm trong Monetary chỉ ra rằng nhóm này có thể chưa tạo ra nhiều giá trị cho doanh nghiệp.

A table of numbers with black text

AI-generated content may be incorrect.

Hình : Hình ảnh khách hàng cụm 2

Cụm 2 có khách hàng với Recency dương, cho thấy họ vẫn đang hoạt động, nhưng Frequency lại thấp, cho thấy tần suất giao dịch không cao. Giá trị Monetary âm cho thấy các khách hàng này không chi tiêu nhiều. Điều này thể hiện cơ hội để khởi động các chương trình khuyến mãi nhằm khuyến khích họ thực hiện nhiều giao dịch hơn.

A table with numbers and text

AI-generated content may be incorrect.

Hình : Hình ảnh khách hàng cụm 3

Khách hàng trong cụm 3 có Recency cũng như Frequency thấp, với giá trị tần suất đồng nhất. Giá trị Monetary âm trong tất cả các trường hợp cho thấy tình trạng chi tiêu thấp và tiềm năng doanh thu không cao. Nhóm khách hàng này có thể cần được phân tích thêm để hiểu lý do không giao dịch và tìm kiếm cách thu hút họ trở lại.

### Triển khai đánh giá kết quả phân cụm bằng trực quan hóa dữ liệu

* Vẽ biểu đồ phân tán từng cụm theo chỉ số RFM

A graph of colored dots

AI-generated content may be incorrect.

Hình : Phân cụm khách hàng bằng K-Means với k=4

Biểu đồ thể hiện kết quả phân cụm khách hàng theo mô hình RFM (Recency, Frequency, Monetary) trong không gian 3 chiều với số cụm k=4, cho thấy sự phân chia rõ ràng giữa các nhóm khách hàng. Các điểm dữ liệu được chia thành 4 cụm với màu sắc khác nhau, thể hiện sự khác biệt về hành vi mua sắm.

Cụm màu xanh đậm chủ yếu gồm những khách hàng mua hàng gần đây, thường xuyên và có chi tiêu cao – đây có thể là nhóm khách hàng trung thành, giá trị cao. Ngược lại, cụm màu tím và xanh lá tập trung ở khu vực có Recency cao (tức đã lâu không mua hàng), thể hiện nhóm khách hàng không còn tương tác hoặc có nguy cơ rời bỏ. Cụm màu vàng phân bố rộng, có thể là nhóm khách hàng có hành vi đa dạng, khó dự đoán hơn.

Nhìn chung, việc chọn k=4 là hợp lý vì biểu đồ thể hiện rõ sự phân tách giữa các cụm, giúp doanh nghiệp dễ dàng nhận diện các nhóm khách hàng khác nhau để từ đó xây dựng các chiến lược tiếp thị, chăm sóc phù hợp.

* Vẽ biểu đồ nhiệt thể hiện chỉ số trung bình RFM của từng cụm khách hàng

A screenshot of a graph

AI-generated content may be incorrect.

Hình : Trung bình RFM của từng cụm khách hàng

* Biểu đồ thể hiện giá trị trung bình đã chuẩn hóa của ba chỉ số Recency, Frequency và Monetary cho từng cụm khách hàng sau khi phân cụm bằng K-Means (k=4). Mỗi hàng đại diện cho một cụm khách hàng, còn mỗi cột tương ứng với một đặc trưng trong mô hình RFM.

+ Cụm 0 có Recency = 0.71, Frequency = 1.09, thể hiện nhóm khách hàng trung thành và mua hàng thường xuyên, gần đây, dù giá trị chi tiêu (Monetary = -0.26) không quá cao. Đây có thể là khách hàng tích cực nhưng giá trị đơn hàng chưa cao.

+ Cụm 1 có Recency = -1.33, Frequency = -0.68, Monetary = -0.32, là nhóm khách hàng đã lâu không quay lại, ít mua hàng và giá trị mua thấp. Đây là nhóm khách hàng rời rạc hoặc không còn hoạt động, cần chiến lược tái kích hoạt.

+ Cụm 2 có Recency = 0.45, nhưng Frequency và Monetary đều âm, cho thấy khách hàng có thể vừa quay lại gần đây nhưng tần suất và chi tiêu không cao – có thể là khách hàng mới hoặc chưa gắn bó.

+ Cụm 3 có các giá trị RFM dao động quanh 0, thể hiện nhóm khách hàng trung bình, hành vi chưa quá nổi bật, có thể cần theo dõi thêm hoặc thúc đẩy để nâng cao giá trị.

## Kiểm định sự khác biệt giữa các cụm

### Thực hiện kiểm định ANOVA

ANOVA One Way là một trong những phương pháp phân tích trong kiểm định giả thuyết dùng để so sánh các giá trị trung bình của nhiều nhóm biến độc lập để xác định giữa các nhóm dữ liệu có sự khác biệt hay không, việc kiểm định ANOVA One Way được thực hiện dựa trên một biến độc lập và một biến phụ thuộc.

Biến phụ thuộc (Dependent variable): Đây là những chỉ số được sử dụng để đánh giá sự khác biệt về giá trị trung bình giữa các nhóm khách hàng bao gồm:

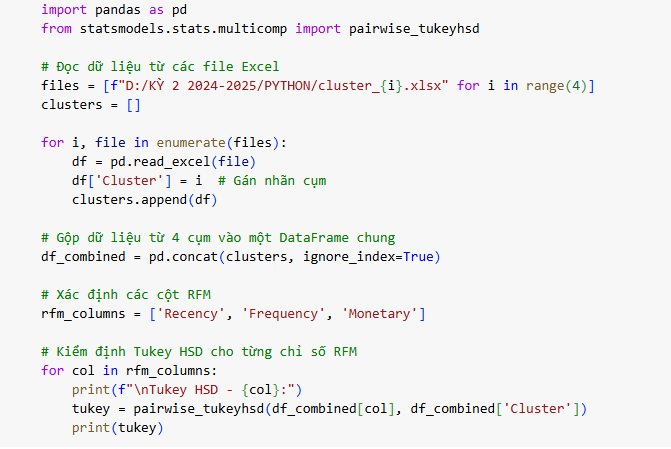
* Recency: Khoảng thời gian kể từ lần mua hàng gần nhất của khách hàng.
* Frequency: Tần suất mua hàng của khách hàng trong một khoảng thời gian xác định.
* Monetary: Tổng số tiền mà khách hàng đã chi tiêu trong một khoảng thời gian nhất định.

Biến độc lập (Independent variable): Đây là yếu tố phân nhóm mà dựa vào đó để so sánh các chỉ số kể trên. Trong trường hợp này, biến độc lập chính là các nhóm khách hàng (Clusters) được phân loại từ trước.

Phương pháp kiểm định ANOVA được áp dụng nhằm xác định liệu có sự khác biệt có ý nghĩa thống kê về giá trị trung bình của các chỉ số Recency, Frequency và Monetary giữa các nhóm khách hàng khác nhau hay không.

### Triển khai Tukey HSD

Kiểm định Tukey HSD ("sự khác biệt có ý nghĩa trung thực" hoặc "sự khác biệt có ý nghĩa trung thực") là một công cụ thống kê được sử dụng để xác định xem mối quan hệ giữa hai tập dữ liệu có ý nghĩa thống kê hay không – tức là liệu có khả năng cao rằng sự thay đổi về số quan sát được ở một giá trị có liên quan nhân quả đến sự thay đổi quan sát được ở một giá trị khác hay không. Nói cách khác, kiểm định Tukey là một cách để kiểm định một giả thuyết thực nghiệm.

Cách thực hiện  


Hình : Thực hiện kiểm định Tukey HSD

### Bước 1: Nhập thư viện cần thiết:

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

* pairwise\_tukeyhsd: Là hàm chính để thực hiện kiểm định Tukey HSD, giúp xác định cặp nhóm nào có sự khác biệt có ý nghĩa thống kê. Đây là thư viện hỗ trợ các kiểm định thống kê trong Python, nhưng không trực tiếp sử dụng trong mã này.
* pandas (import ở đầu): Dùng để xử lý dữ liệu.

### Bước 2: Dữ liệu đầu vào:

files = [‘ ’]: Cột này chứa dữ liệu về Recency, Frequency, và Monetary

* Đọc 4 file chứa dữ liệu từng cụm (Cluster 0 → Cluster 3) đã được phân nhóm bằng KMeans.

df['Cluster'] = i

* Gán nhãn cụm (Cluster) cho từng DataFrame.

df\_combined = pd.concat(clusters, ignore\_index=True)

* Gộp toàn bộ cụm lại thành một bảng duy nhất để tiện phân tích.

### Bước 3: Thực hiện kiểm định Tukey HSD:

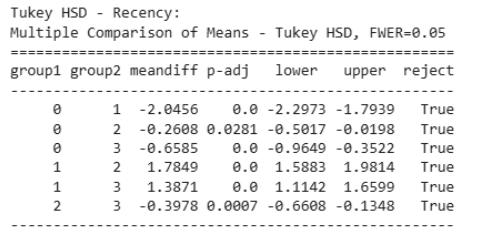
tukey = pairwise\_tukeyhsd(df\_combined['Recency'], df\_combined['Cluster'])

* So sánh giá trị Recency giữa các nhóm cụm khác nhau.
* Có thể thay 'Recency' bằng 'Frequency' hoặc 'Monetary' để kiểm tra từng chỉ số.

### Bước 4: Kết quả trả về:

print(tukey): Kết quả kiểm định sẽ được in ra màn hình dưới dạng bảng. Kết quả này sẽ chỉ ra các cặp nhóm có sự khác biệt có ý nghĩa thống kê về Recency.

Kết quả:



Hình :  *HSD - Recency*

Nhận xét:  
0 vs 1: Sự khác biệt trung bình là -2.0456, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là có sự khác biệt đáng kể giữa nhóm 0 và nhóm 1.

0 vs 2: Sự khác biệt trung bình là -0.2608, p-adj = 0.0281, cũng có ý nghĩa thống kê (reject null hypothesis), mặc dù mức độ khác biệt không lớn như so với nhóm 1.

0 vs 3: Sự khác biệt trung bình là -0.6585, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), cho thấy sự khác biệt rõ rệt giữa nhóm 0 và nhóm 3.

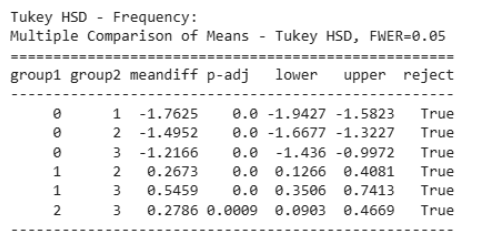
1 vs 2: Sự khác biệt trung bình là 1.7849, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), cho thấy nhóm 1 có giá trị trung bình cao hơn nhóm 2.

1 vs 3: Sự khác biệt trung bình là 1.3871, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), cho thấy nhóm 1 cũng có giá trị trung bình cao hơn nhóm 3.

2 vs 3: Sự khác biệt trung bình là -0.3978, p-adj = 0.0007, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 2 có giá trị trung bình thấp hơn nhóm 3.

→ Nhận xét chung:

1. Nhóm 0 so với các nhóm khác:
   * Nhóm 0 có sự khác biệt rõ rệt với cả ba nhóm còn lại (1, 2, 3) với giá trị p-adj rất nhỏ (0.0000 cho nhóm 1 và nhóm 3, 0.0281 cho nhóm 2). Điều này cho thấy nhóm 0 có sự khác biệt lớn và có ý nghĩa thống kê so với các nhóm khác.
   * Cụ thể, nhóm 0 có giá trị trung bình thấp hơn nhóm 1 (-2.0456), nhóm 2 (-0.2608), và nhóm 3 (-0.6585).
2. Nhóm 1 so với các nhóm còn lại:
   * Nhóm 1 có sự khác biệt rõ rệt và có ý nghĩa thống kê với cả nhóm 2 và nhóm 3, với p-adj = 0.0000 trong cả hai trường hợp. Điều này chỉ ra rằng nhóm 1 có giá trị trung bình cao hơn các nhóm còn lại.
   * Cụ thể, sự khác biệt giữa nhóm 1 và nhóm 2 là 1.7849, và giữa nhóm 1 và nhóm 3 là 1.3871.
3. Nhóm 2 và nhóm 3:
   * Sự khác biệt giữa nhóm 2 và nhóm 3 là có ý nghĩa thống kê (p-adj = 0.0007), nhưng mức độ khác biệt không lớn như giữa các nhóm khác. Nhóm 2 có giá trị trung bình thấp hơn nhóm 3 với sự khác biệt là -0.3978.

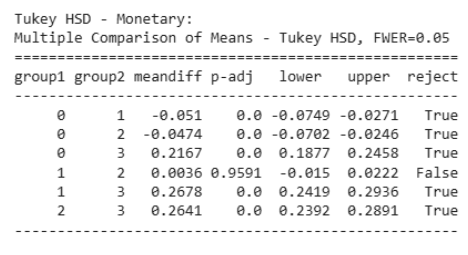


Hình : *HSD - Frequency*

Nhận xét:

1. 0 vs 1: Sự khác biệt trung bình là -1.7625, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), cho thấy nhóm 0 có giá trị tần suất thấp hơn nhóm 1.
2. 0 vs 2: Sự khác biệt trung bình là -1.4952, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), cho thấy nhóm 0 có giá trị tần suất thấp hơn nhóm 2.
3. 0 vs 3: Sự khác biệt trung bình là -1.2166, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), cho thấy nhóm 0 có giá trị tần suất thấp hơn nhóm 3.
4. 1 vs 2: Sự khác biệt trung bình là 0.2673, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 1 có giá trị tần suất cao hơn nhóm 2.
5. 1 vs 3: Sự khác biệt trung bình là 0.5459, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 1 có giá trị tần suất cao hơn nhóm 3.
6. 2 vs 3: Sự khác biệt trung bình là 0.2786, p-adj = 0.0009, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 2 có giá trị tần suất thấp hơn nhóm 3.  
   → Nhận xét chung

* Nhóm 0 có giá trị tần suất thấp hơn rõ rệt so với các nhóm khác (1, 2, 3).
* Nhóm 1 có giá trị tần suất cao hơn các nhóm 2 và 3.
* Nhóm 2 và nhóm 3 có sự khác biệt nhỏ, nhưng vẫn có ý nghĩa thống kê, với nhóm 2 có giá trị tần suất thấp hơn nhóm 3



Hình : *HSD - Monetary*

Nhận xét

1. 0 vs 1: Sự khác biệt trung bình là -0.0510, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 0 có giá trị tiền tệ thấp hơn nhóm 1.
2. 0 vs 2: Sự khác biệt trung bình là -0.0474, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 0 có giá trị tiền tệ thấp hơn nhóm 2.
3. 0 vs 3: Sự khác biệt trung bình là 0.2167, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 0 có giá trị tiền tệ thấp hơn nhóm 3.
4. 1 vs 2: Sự khác biệt trung bình là 0.0036, p-adj = 0.9591, không có ý nghĩa thống kê (do p-adj > 0.05), nên không bác bỏ giả thuyết không có sự khác biệt. Điều này cho thấy nhóm 1 và nhóm 2 không có sự khác biệt đáng kể về giá trị tiền tệ.
5. 1 vs 3: Sự khác biệt trung bình là 0.2678, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 1 có giá trị tiền tệ thấp hơn nhóm 3.
6. 2 vs 3: Sự khác biệt trung bình là 0.2641, p-adj = 0.0000, có ý nghĩa thống kê (reject null hypothesis), tức là nhóm 2 có giá trị tiền tệ thấp hơn nhóm 3.  
   → Nhận xét chung

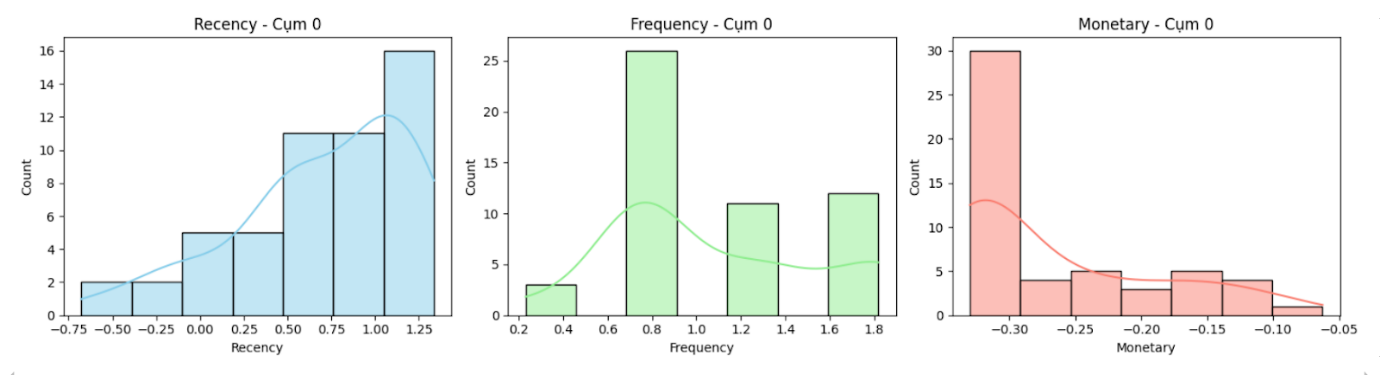
* Nhóm 0 có giá trị tiền tệ thấp hơn so với các nhóm 1 và 3, nhưng cao hơn nhóm 2.
* Nhóm 1 và nhóm 2 không có sự khác biệt đáng kể về giá trị tiền tệ (p-adj = 0.9591).
* Nhóm 1 có giá trị tiền tệ thấp hơn nhóm 3.
* Nhóm 2 có giá trị tiền tệ thấp hơn nhóm 3.

### Phân tích cụm

* 1. Cụm 0



Hình : Vẽ biểu đồ phân phối cụm 0



Hình : Biểu đồ phân phối cụm 0

* Nhận xét:  
  Phân bố Recency trong cụm 0
* Phân bố lệch trái (bị nghiêng về bên phải).
* Phần lớn khách hàng có giá trị Recency cao (nghĩa là họ mới tương tác gần đây).
* Điều này gợi ý rằng cụm này chứa nhiều khách hàng mới hoạt động hoặc gần đây vừa có giao dịch.

Frequency - Cụm 0

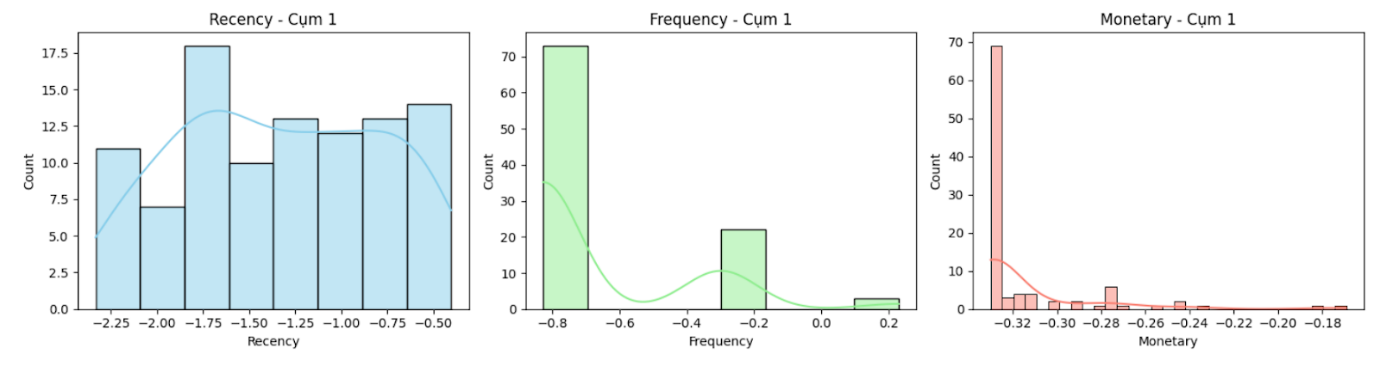
* Có vẻ gần như phân bố chuẩn lệch phải nhẹ hoặc phân bố đa đỉnh (có nhiều đỉnh).
* Phần lớn khách hàng có tần suất tương tác ở mức trung bình - thấp (dao động từ 0.8 đến 1.0).
* Cho thấy nhóm khách này không quá thường xuyên mua hàng, nhưng vẫn có tần suất tương đối ổn định.

Monetary - Cụm 0

* Phân bố lệch phải rất mạnh (đa số tập trung ở bên trái, giá trị thấp).
* Khách hàng trong cụm này có xu hướng chi tiêu ít tiền hơn so với nhóm khác
* Đây có thể là những khách hàng mới, hoặc ít giá trị về mặt doanh thu.
  1. Cụm 1:



Hình : Vẽ biểu đồ phân phối cụm 1



Hình : Biểu đồ phân phối cụm 1

Nhận xét:

Recency – Cụm 1

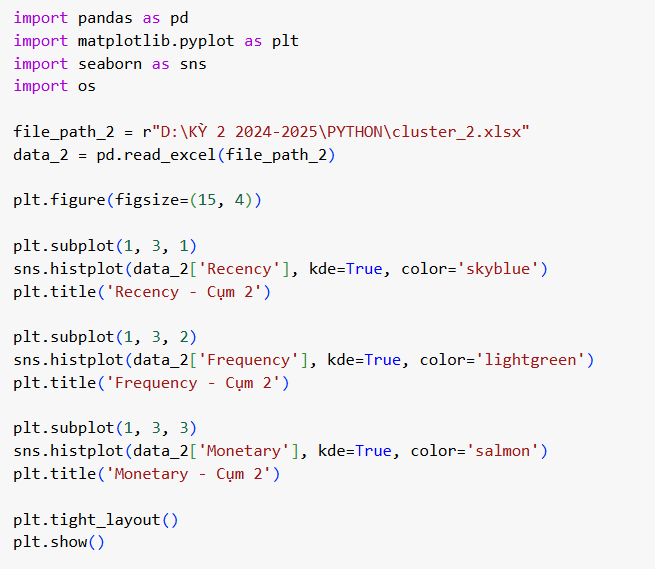
* Phân bố Recency gần như đồng đều nhưng hơi lệch phải nhẹ, phần lớn giá trị nằm ở mức thấp.
* Phần lớn khách hàng có điểm Recency thấp (nghĩa là đã mua hàng cách đây không quá lâu).
* Gợi ý đây là nhóm khách đã từng tương tác gần đây nhưng không quá mới, có thể đang tạm thời ngưng mua.

Frequency – Cụm 1

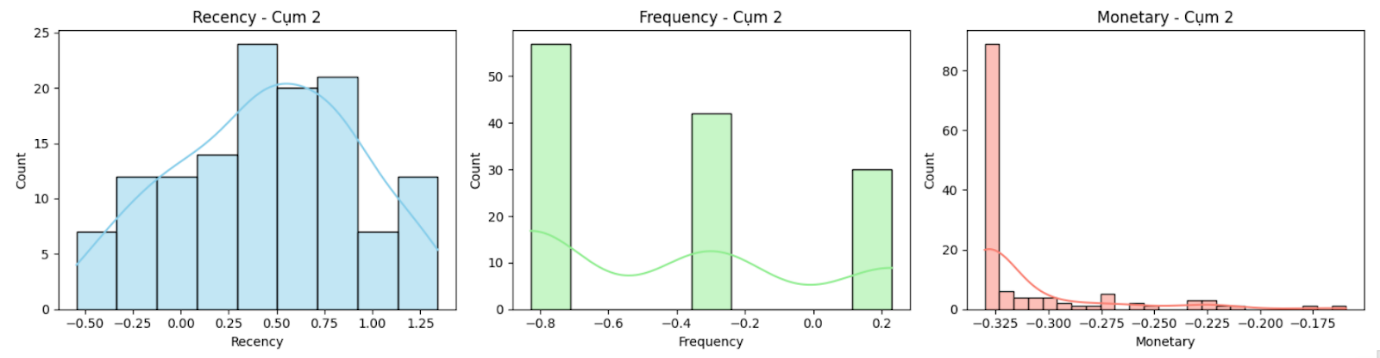
* Phân bố lệch trái rõ rệt, tập trung chủ yếu ở vùng Frequency thấp.
* Hầu hết khách hàng chỉ mua 1–2 lần hoặc tần suất rất thấp.
* Cho thấy cụm này là nhóm khách chưa có sự trung thành, mới mua thử hoặc không quay lại thường xuyên.

Monetary – Cụm 1

* Phân bố lệch phải rất mạnh, phần lớn tập trung ở giá trị thấp.
* Đa số khách hàng chỉ chi tiêu ở mức tối thiểu, rất ít người có giá trị mua cao.
* Đây có thể là nhóm khách ít đóng góp doanh thu, cần được khuyến khích chi tiêu thêm qua ưu đãi phù hợp.
  1. Cum 2:



Hình : Vẽ biểu đồ phân phối cụm 2



Hình : Biểu đồ phân phối cụm 2

Nhận xét:

Recency - Cụm 2

* Phân bố gần chuẩn, hơi lệch phải nhẹ (có xu hướng nghiêng về bên trái, nhiều giá trị nhỏ hơn trung bình).
* Phần lớn khách hàng có giá trị Recency từ khoảng 0.25 đến 0.75, tức là họ không tương tác quá gần đây, nhưng cũng không quá lâu.
* Điều này cho thấy cụm này gồm những khách hàng tương tác ở mức độ trung bình, không phải nhóm mua hàng gần nhất, nhưng vẫn đang hoạt động.

Frequency - Cụm 2

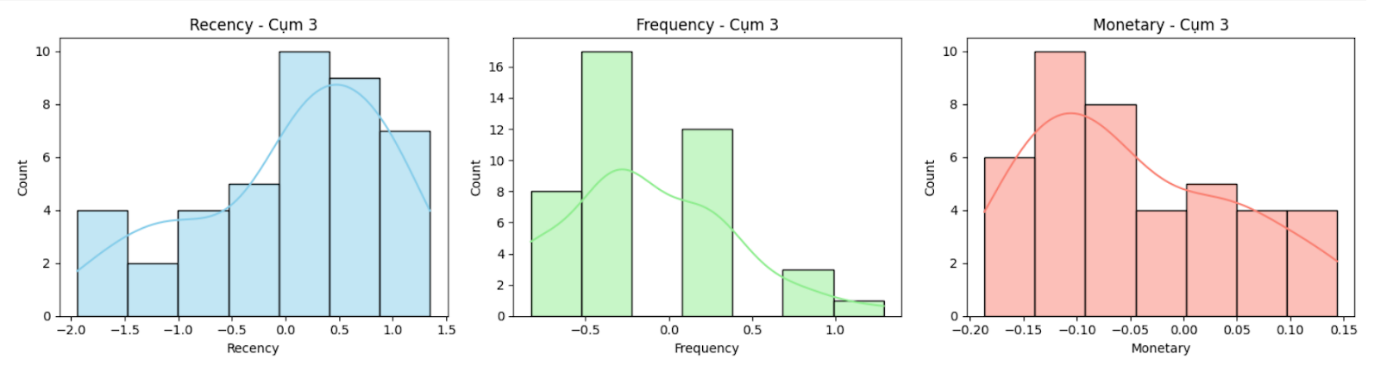
* Phân bố không đều, đa đỉnh rõ rệt, có 3 cụm giá trị riêng biệt.
* Có một đỉnh lớn ở vùng -0.8, tức là nhiều khách hàng có tần suất mua hàng rất thấp.
* Hai đỉnh nhỏ hơn ở khoảng -0.3 và 0.2, thể hiện một số khách hàng vẫn mua khá thường xuyên.
* Nhìn chung, cụm này gồm đa số khách hàng mua ít, chỉ một phần nhỏ là có tần suất mua đều đặn.

Monetary - Cụm 2

* Phân bố lệch phải cực mạnh (rất nhiều giá trị nhỏ tập trung ở bên trái).
* Gần như tất cả khách hàng trong cụm này có giá trị chi tiêu thấp hơn trung bình.
* Điều này cho thấy cụm này gồm những khách hàng có giá trị doanh thu thấp
  1. Cụm 3:



Hình : Vẽ biểu đồ phân phối cụm 3



Hình : Biểu đồ phân phối cụm 3

Nhận xét:

 Recency - Cụm 3

* Phân bố nghiêng về bên trái, tức là nhiều khách hàng có giá trị Recency gần 0 hoặc dương.
* Ý nghĩa: Những khách hàng này mua hàng gần đây, tức là vẫn còn tương tác gần với doanh nghiệp.
* Tuy có một số ít khách hàng có Recency âm (tức là không tương tác gần đây), nhưng phần lớn vẫn cho thấy sự gần gũi về mặt thời gian.

Frequency - Cụm 3

* Phân bố nghiêng phải, tập trung nhiều ở phía âm.
* Ý nghĩa: Khách hàng trong cụm này ít khi mua hàng, tần suất mua sắm không cao.
* Chỉ có một số rất ít khách hàng có Frequency dương, tức là mua hàng thường xuyên hơn.

Monetary - Cụm 3

* Phân bố lệch trái mạnh, với đa số các giá trị nằm ở phía âm.
* Ý nghĩa: Khách hàng trong cụm này chi tiêu ít, có giá trị mua sắm không cao.
* Chỉ một vài cá nhân chi tiêu tương đối tốt, nhưng không đại diện cho toàn cụm.

### Kết luận các cụm khách hàng

#### Cụm 0: Khách hàng mới tiềm năng (Recency thấp, Frequency thấp, Monetary trung bình)

* Đặc điểm: Mới mua gần đây, mua ít lần, chi tiêu ở mức trung bình.

#### Cụm 1: Khách hàng cũ không còn chi tiêu (Recency cao, Frequency cao, Monetary thấp nhất)

* Đặc điểm: Mua nhiều lần trong quá khứ nhưng đã ngừng giao dịch gần đây, chi tiêu ít.

Cụm 2: Khách Hàng Trung Lập (Recency trung bình, Frequency trung bình, Monetary thấp)

Đặc điểm:

* Mua hàng thỉnh thoảng, không quá thường xuyên
* Không phải là khách hàng mới, nhưng cũng không bỏ đi lâu
* Chi tiêu thấp, không đóng góp nhiều vào doanh thu

### Cụm 3: Khách Hàng VIP Trung Thành

Đặc điểm :

* Recency (Thời gian gần nhất mua hàng): Thấp (mới mua gần đây) → Rất tích cực.
* Frequency (Tần suất mua hàng): Cao → Mua thường xuyên, gắn bó với thương hiệu.
* Monetary (Giá trị chi tiêu): Cao nhất → Đóng góp doanh thu lớn.

## Phân tích kết quả

Dựa trên kết quả phân cụm, doanh nghiệp có thể cá nhân hóa chiến lược marketing cho từng phân khúc khách hàng, từ đó nâng cao hiệu quả kinh doanh một cách tối ưu. Cụ thể như sau:

#### Cụm 0: Khách hàng mới tiềm năng (Recency thấp, Frequency thấp, Monetary trung bình)

Đặc điểm: Mới mua gần đây, mua ít lần, chi tiêu ở mức trung bình.

Chiến lược marketing:

* Gửi email chào mừng kèm mã giảm giá cho lần mua tiếp theo, mục tiêu là giữ khách hàng quay lại sớm sau lần mua đầu tiên, tạo thành thói quen mua sắm.
* Remarketing bằng quảng cáo sản phẩm tương tự đã mua giúp giữ thương hiệu trong tâm trí khách hàng và thúc đẩy hành động mua tiếp
* Tạo các chương trình khách hàng thân thiết để khuyến khích mua lại tăng khả năng gắn bó dài hạn với thương hiệu.

#### Cụm 1: Khách hàng cũ không còn chi tiêu (Recency cao, Frequency cao, Monetary thấp nhất)

Đặc điểm: Mua nhiều lần trong quá khứ nhưng đã ngừng giao dịch gần đây, chi tiêu ít.

Chiến lược marketing:

* Gửi email chăm sóc lại với ưu đãi đặc biệt như “Mua lại sau 3 tháng – Giảm 30%” tạo cảm giác được quan tâm, thúc đẩy hành động trở lại.
* Khảo sát lý do ngừng mua để điều chỉnh sản phẩm/dịch vụ phù hợp giúp doanh nghiệp hiểu rõ nguyên nhân mất khách và điều chỉnh sản phẩm/dịch vụ.
* Tạo cảm giác cấp bách (limited time offer) để kích thích hành động, ví dụ như nếu khách từng thích danh mục sản phẩm nào, hãy đẩy mạnh quảng cáo đúng nhóm sản phẩm đó. Điều đó giúp tăng tỷ lệ click và tỷ lệ chuyển đổi bằng cách đưa đúng nội dung.

Cụm 2: Khách Hàng Trung Lập (Recency trung bình, Frequency trung bình, Monetary thấp)

#### Đặc điểm nhóm: Mua hàng nhiều lần trước đây nhưng đã ngừng giao dịch gần đây và chi tiêu thấp.

### Chiến Lược Cụ Thể:

* Chiến lược email chăm sóc với ưu đãi "Mua lại sau 3 tháng - Giảm 30%" nhằm kích thích khách hàng quay trở lại sau thời gian không tương tác. Ưu đãi có thời hạn tạo cảm giác cấp bách, đồng thời thể hiện sự quan tâm của doanh nghiệp, giúp khách hàng cảm thấy được trân trọng.
* Khảo sát lý do ngừng mua hàng được thực hiện để thấu hiểu nguyên nhân khách hàng giảm tương tác. Thông tin thu thập sẽ giúp doanh nghiệp điều chỉnh sản phẩm/dịch vụ phù hợp hơn với nhu cầu thực tế, từ đó cải thiện trải nghiệm và gia tăng cơ hội giữ chân khách hàng.
* Chiến dịch quảng cáo nhắm đúng nhóm sản phẩm khách từng quan tâm tập trung vào các mặt hàng khách đã xem hoặc mua trước đây. Bằng cách cá nhân hóa nội dung và kèm ưu đãi có thời hạn, chiến dịch này tối ưu tỷ lệ click và chuyển đổi, đồng thời củng cố nhận diện thương hiệu trong tâm trí khách hàng.

### Cụm 3: Khách Hàng VIP Trung Thành

Đặc điểm nhóm: Mua hàng thường xuyên và chi tiêu cao, thể hiện sự trung thành với thương hiệu.

Chiến lược chính:

* Chương trình ưu đãi đặc quyền (giảm giá 15-20%, quà tặng cao cấp) được thiết kế để duy trì mức độ trung thành và nâng cao giá trị vòng đời khách hàng (LTV). Những ưu đãi độc quyền giúp khách VIP cảm nhận được sự khác biệt, từ đó giảm thiểu nguy cơ chuyển sang đối thủ cạnh tranh.
* Chiến dịch giới thiệu bạn bè (thưởng 10-15% giá trị đơn hàng) khuyến khích khách hàng VIP trở thành đại sứ thương hiệu. Bằng cách tận dụng mạng lưới quan hệ của nhóm khách hàng giá trị này, doanh nghiệp có thể mở rộng cơ sở khách hàng chất lượng với chi phí acquisition thấp hơn so với các kênh truyền thống.
* Dịch vụ hỗ trợ 24/7 qua hotline riêng được triển khai nhằm nâng cao trải nghiệm khách hàng VIP. Với dịch vụ chăm sóc cá nhân hóa và ưu tiên, doanh nghiệp củng cố mối quan hệ gắn bó lâu dài, đồng thời gia tăng sự hài lòng và niềm tin của nhóm khách hàng mang lại giá trị cao nhất cho doanh nghiệp.

# Kết luận

Dự án phân tích dữ liệu khách hàng đã được hoàn thành với những kết quả tích cực, mang lại nhiều giá trị thực tiễn cho hoạt động marketing và chăm sóc khách hàng của doanh nghiệp. Nhóm đã áp dụng mô hình RFM (Recency - Frequency - Monetary) kết hợp với thuật toán phân cụm K-Means để phân loại khách hàng thành bốn nhóm dựa trên ba tiêu chí chính: thời gian mua hàng gần nhất, tần suất mua sắm và giá trị chi tiêu. Mỗi nhóm khách hàng thể hiện những đặc điểm hành vi riêng biệt, từ đó giúp doanh nghiệp có cơ sở để xây dựng các chiến lược tiếp cận phù hợp.

Quy trình thực hiện được triển khai một cách bài bản và khoa học. Dữ liệu được thu thập từ nguồn Superstore, sau đó được làm sạch, xử lý và chuẩn hóa. Nhóm tiến hành tính toán các chỉ số RFM, chuẩn hóa các thang đo và thực hiện phân cụm bằng thuật toán K-Means. Số lượng cụm được lựa chọn dựa trên phương pháp Elbow nhằm đảm bảo mô hình đạt được sự cân bằng giữa độ chính xác và mức độ đơn giản. Sau khi phân nhóm, các kiểm định thống kê ANOVA và Tukey HSD được sử dụng để đánh giá sự khác biệt giữa các cụm, đảm bảo tính tin cậy của kết quả.

Dựa trên đặc điểm từng nhóm, nhóm nghiên cứu đã đề xuất các hướng tiếp thị phù hợp. Nhóm khách hàng thân thiết nên được ưu đãi để giữ chân, khách hàng mới cần được thu hút bằng khuyến mãi, và nhóm ít tương tác có thể được tái kích hoạt thông qua các chiến dịch chăm sóc cá nhân hóa.

Tổng thể, việc kết hợp mô hình RFM và thuật toán K-Means đã cho thấy hiệu quả trong việc phân khúc khách hàng, đồng thời cung cấp nền tảng vững chắc để doanh nghiệp phát triển các chiến lược tiếp thị mục tiêu một cách hiệu quả hơn.

Tài liệu tham khảo

Bizfly. (n.d.). *Phân tích RFM là gì? Phân loại khách hàng theo mô hình RFM*. Retrieved from Bizfly: https://bizfly.vn/techblog/phan-tich-rfm-la-gi-phan-loai-khach-hang-theo-mo-hinh-rfm.html

Exchange., S. (n.d.). *Appropriateness of ANOVA after K-means cluster analysis*. Retrieved from Cross Validated.: https://stats.stackexchange.com/questions/116294/appropriateness-of-anova-after-k-means-cluster-analysis

Glen, S. (n.d.). *Tukey’s Test for Honest Significant Difference (HSD)*. Retrieved from Statistics How To: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/post-hoc/tukey-test-honest-significant-difference/

Ideascale. (n.d.). *Phân khúc khách hàng là gì?* Retrieved from Ideascale: https://ideascale.com/vi/viblog/phan-khuc-khach-hang-la-gi/

Investopedia. (n.d.). *T-test*. Retrieved from Investopedia.: https://www.investopedia.com/terms/t/t-test.asp

Kenton, W. (2023, March 29). *T-Test: What It Is With Multiple Formulas and When To Use Them*. Retrieved from Investopedia: https://www.investopedia.com/terms/t/t-test.asp

Li, X. &. (n.d.). *A Study of Customer Segmentation Based on RFM Analysis and K-Means*. Retrieved from ResearchGate: https://www.researchgate.net/publication/377039251\_A\_Study\_of\_Customer\_Segmentation\_Based\_on\_RFM\_Analysis\_and\_K-Means

Luanvan1080. (n.d.). *Phân tích phương sai một chiều (One-Way ANOVA) là gì?* Retrieved from luanvan1080.com: https://luanvan1080.com/anova-one-way.html

Phượng, N. T. (n.d.). *Ứng dụng phân tích RFM trong phân khúc khách hàng tại công ty cổ phần sữa Việt Nam (Vinamilk)*. Retrieved from Thư viện số UEH: https://digital.lib.ueh.edu.vn/handle/UEH/63165

SlideShare. (n.d.). *Phân cụm dữ liệu trong Data Mining*. Retrieved from SlideShare: https://fr.slideshare.net/slideshow/phn-cm-d-liu-trong-datamingpdf/258056391

Team, C. (n.d.). *F-Test*. Retrieved from Cuemath: https://www.cuemath.com/data/f-test/

TopOnSeek. (n.d.). *Phân khúc khách hàng: Các nội dung cơ bản.* Retrieved from TopOnSeek.: https://www.toponseek.com/blogs/phan-khuc-khach-hang/

Thịnh, N. V. (n.d.). *Phân cụm khách hàng của các siêu thị điện máy tại TP. Hồ Chí Minh bằng thuật toán K-Means*. Retrieved from Tạp chí Khoa học Đại học Mở TP. HCM – CN Khoa học Tự nhiên và Công nghệ (VJOL): https://vjol.info.vn/index.php/tcidhmodiachat/article/view/55494

Link Youtube: <https://youtu.be/tVheWXuf3tg>