**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH**

**TRƯỜNG CÔNG NGHỆ VÀ THIẾT KẾ**

🖎🕮✍



**ĐỒ ÁN MÔN HỌC**

**ĐỀ TÀI:**

PHÂN CỤM KHÁCH HÀNG THEO PHƯƠNG PHÁP RFM

**Học phần**: Máy Học

**Nhóm Sinh Viên**:

1. ĐOÀN VŨ MINH THANH
2. ĐẶNG CẨM TÚ
3. MAI TRẦN MỸ UYÊN
4. NGÔ THỊ HUYỀN

**Chuyên Ngành**: KHOA HỌC DỮ LIỆU

**Khóa**: K46

**Giảng Viên**: TS. Đặng Ngọc Hoàng Thành

**TP. Hồ Chí Minh, Ngày 22 tháng 04 năm 2023**

Lời cảm ơn

Để có thể hoàn thành được bài báo cáo cuối kỳ không chỉ có riêng sự cố gắng của các thành viên trong nhóm mà còn nhờ vào sự hỗ trợ của rất nhiều của thầy Đặng Ngọc Hoàng Thành. Chúng em xin được gửi lời cảm ơn chân thành đến thầy đã tận tình hướng dẫn chúng em về cách thức tiến hành đề tài nghiên cứu và qua các bài giảng trên lớp để có thể hoàn thành tốt bài báo cáo này

Trong quá trình phát triển đề tài do còn hạn chế về kiến thức cũng như kinh nghiệm vì vậy bài làm sẽ không tránh khỏi những sai sót. Em mong nhận được sự thông cảm và góp ý phê bình từ phía Thầy.

# 

Trân trọng.

# 

**MỤC LỤC NỘI DUNG**

[BẢNG ĐÁNH GIÁ CÔNG VIỆC 1](#_Toc135064036)

[CHƯƠNG 1. TỔNG QUAN 3](#_Toc135064037)

[1.1. Tổng Quan Về Bài Toán 3](#_Toc135064038)

[1.2. Lý Do Chọn Lựa Đề Tài 3](#_Toc135064039)

[CHƯƠNG 2. ĐỀ XUẤT MÔ HÌNH 5](#_Toc135064040)

[2.1. Bộ Dữ Liệu [7] 5](#_Toc135064041)

[2.2. Mô hình đề xuất 6](#_Toc135064042)

[CHƯƠNG 3. TRIỂN KHAI VÀ ĐÁNH GIÁ MÔ HÌNH [8] 10](#_Toc135064043)

[3.1. Tiền xử lý dữ liệu 10](#_Toc135064044)

[3.2. Phân tích khám phá dữ liệu (EDA - Exploratory Data Analysis) 11](#_Toc135064045)

[a. Số lượng sản phẩm trong mỗi đơn hàng: 11](#_Toc135064046)

[b. Số lượng đơn hàng của mỗi khách hàng: 16](#_Toc135064047)

[c. Tình hình mua sắm trực tuyến của khách hàng qua các thời điểm: 17](#_Toc135064048)

[d. Các mặt hàng phổ biến: 20](#_Toc135064049)

[e. Hình thức thanh toán phổ biến: 26](#_Toc135064050)

[3.3. Tính RFM 29](#_Toc135064051)

[3.4. Lấy mẫu và xác định các phần tử ngoại biên (outliers) 31](#_Toc135064052)

[3.5. Phân cụm dữ liệu 36](#_Toc135064053)

[a. Trường hợp 1: không loại bỏ các phần tử ngoại biên (outliers) 36](#_Toc135064054)

[b.Trường hợp 2: loại bỏ các phần tử ngoại biên (outliers) 48](#_Toc135064055)

[3.6. Mô tả đặc điểm các cụm 59](#_Toc135064056)

[CHƯƠNG 4. KẾT LUẬN 68](#_Toc135064057)

[4.1. Các kết quả đạt được 68](#_Toc135064058)

[4.2. Những Hạn Chế và Hướng Phát Triển 69](#_Toc135064059)

[4.2.1. Những hạn chế 69](#_Toc135064060)

[4.2.2. Hướng phát triển 70](#_Toc135064061)

[TÀI LIỆU THAM KHẢO 71](#_Toc135064062)

# MỤC LỤC HÌNH ẢNH

Hình 1. Lưu đồ cơ sở dữ liệu của bộ dữ liệu nghiên cứu 5

Hình 2. Mô hình đề xuất (1) 9

Hình 3. Mô hình đề xuất (2) 9

Hình 4. Kết quả bảng dữ liệu đã hợp nhất 10

Hình 5. bảng thống kê số lượng món hàng của mỗi mã đơn hàng 13

Hình 6. biểu đồ đường thể hiện số lượng đơn hàng qua thời gian 20

Hình 7. bảng kết quả tổng hợp số lượng đã bán và doanh thu trung bình của từng loại sản phẩm 21

Hình 8. Biểu đồ thể hiện số lượng và giá trị trung bình của các đơn hàng của từng loại sản phẩm 22

Hình 9. sản phẩm có lượng bán ra cao I 22

Hình 10. Sản phẩm có lượng bán ra cao II 23

Hình 11. Sản phẩm có lượng bán ra cao III 23

Hình 12. Sản phẩm có lượng bán ra cao IV 24

Hình 13. Sản phẩm có lượng bán ra cao V 24

Hình 14. Các thiết bị có lượng bán ra thấp I 25

Hình 15. Các thiết bị có lượng bán ra thấp II 25

Hình 16. Các thiết bị có lượng bán ra thấp III 25

Hình 17. Biểu đồ thể hiện hình thức thanh toán phổ biến và giá trị trung bình của các đơn hàng với từng hình thức thanh toán 28

Hình 18. Tình sử dụng hình thức thanh toán phiếu mua hàng - voucher 28

Hình 19. Tình hình sử dụng hình thức thanh toán thẻ tín dụng - credit card 29

Hình 20. Biểu đồ hộp của 3 biến R, F, M 31

Hình 21.Trường hợp 1 - Biểu đồ Elbow 39

Hình 22. Trường hợp 1 - Distorition score elbow để xác định số lượng cụm tối ưu 40

Hình 23. Trường hợp 1 - 1 phần của dendrogram 43

Hình 24. Trường hợp 2 - Distorition score elbow để xác định số lượng cụm tối ưu 52

Hình 25.Trường hợp 2 - Kết quả tính Silhouette Score của các số lượng cụm 53

Hình 26. Trường hợp 2 - Kết quả phân cụm bằng KMeans được thêm vào bảng RFM gốc 54

Hình 27. Trường hợp 2 - 1 phần của dendrogram 55

Hình 28. Trường hợp 2 - Kết quả phân cụm bằng HAC được thêm vào bảng RFM gốc 56

Hình 29. Kết quả phân cụm bằng mô hình AutoEncoder 58

Hình 30. Biểu đồ plot thể hiện tổng giá trị đơn hàng của các cụm (KMeans) 59

Hình 31. Biểu đồ thể hiện tần suất mua hàng của các cụm (KMeans) 60

Hình 32. Biểu đồ thể hiện khả năng mua hàng gần đây của khách hàng các cụm (KMeans) 61

Hình 33. Biểu đồ plot thể hiện tổng giá trị đơn hàng của các cụm (HAC) 62

Hình 34. Biểu đồ thể hiện tần suất mua hàng của các cụm (HAC) 63

Hình 35. Biểu đồ thể hiện khả năng mua hàng gần đây của khách hàng các cụm (HAC) 64

Hình 36. Biểu đồ plot thể hiện tổng giá trị đơn hàng của các cụm (AutoEncoder) 65

Hình 37. Biểu đồ plot thể hiện xu hướng tần số mua sắm trực tuyến của các cụm (AutoEncoder) 66

Hình 38. Biểu đồ plot thể hiện xu hướng thời gian từ lần mua hàng cuối của các cụm (AutoEncoder) 67

# MỤC LỤC BẢNG

Table 1. bảng kết quả tổng hợp số lượng món hàng của mỗi mã đơn hàng được sắp xếp theo chiều giảm dần 12

Table 2. bảng kết quả tổng hợp số lượng các loại sản phẩm của mỗi đơn hàng 14

Table 3. hợp nhất hai bảng tổng hợp số lượng sản phẩm của mỗi đơn hàng 15

Table 4. bảng thống kê số lượng loại sản phẩm của mỗi mã đơn hàng 16

Table 5. bảng thống kê số lượng khách hàng tương ứng với mỗi số lượng đơn hàng 17

Table 6. bảng kết quả tổng hợp số lượng đơn hàng trong các khoảng thời gian 19

Table 7. Bảng tạm chứa thông tin về thanh toán của từng mã đơn hàng 26

Table 8. Bảng kết quả tổng hợp số đơn hàng tương ứng với từng cặp hình thức thanh toán-giá trị đơn hàng 27

Table 9. danh sách các phần tử ngoại biên (IQR) 33

Table 10. danh sách các phần tử ngoại biên (Z-score) 35

Table 11. Trường hợp 1 - huấn luyện bảng dữ liệu và in bảng kết quả chuẩn hóa 37

Table 12. Trường hợp 1 - Kết quả phân cụm bằng KMeans được thêm vào bảng RFM gốc 42

Table 13. Trường hợp 1 - Kết quả phân cụm bằng HAC được thêm vào bảng RFM gốc 44

Table 14. Trường hợp 2 - Bảng dữ liệu RFM sau khi đã loại các phần tử ngoại biên - outliers 48

# BẢNG ĐÁNH GIÁ CÔNG VIỆC

|  |  |  |
| --- | --- | --- |
| **Thành viên** | **Công việc** | **Mức độ hoàn thành** |
| Ngô Thị Huyền | * TH2: chuẩn hóa dữ liệu, tìm số cụm, phân cụm học sâu AutoEncoder * Mô tả đặc điểm các cụm * Lý thuyết: KMeans, HAC, AutoEncoder * Hoàn thiện word báo cáo | * Đúng hạn * Có đóng góp ý kiến * Lắng nghe góp ý và có điều chỉnh hợp lý   ⇒ 100% |
| Đoàn Vũ Minh Thanh | * Trưởng nhóm * Phát hiện outliers * EDA * TH2: Phân cụm KMeans, HAC * Lý thuyết: đề xuất mô hình * Hoàn thiện word báo cáo * Hoàn chỉnh code | * Đúng hạn * Là người lên workflow, dàn bài báo cáo; Phân công việc rõ ràng, hợp lý * Có đóng góp ý kiến * Lắng nghe góp ý và có điều chỉnh hợp lý   ⇒ 100% |
| Đặng Thị Cẩm Tú | * Thu thập dữ liệu * Tiền xử lý dữ liệu * TH1: phân cụm KMeans, HAC * Tính RFM * Lý thuyết: RFM, Tiền xử lý dữ liệu, Mô tả bộ dữ liệu | * Đúng hạn * Có đóng góp ý kiến * Lắng nghe góp ý và có điều chỉnh hợp lý   ⇒ 100% |
| Mai Trần Mỹ Uyên | * Tiền xử lý dữ liệu * TH1: chuẩn hóa dữ liệu, tìm số cụm, phân cụm học sâu AutoEncoder * Đánh giá 2 trường hợp * Lý thuyết: Tổng quan; Elbow Curves, Silhouette Scores, các chỉ số đánh giá; Kết luận | * Đúng hạn * Có đóng góp ý kiến * Lắng nghe góp ý và có điều chỉnh hợp lý   ⇒ 100% |

# 

# CHƯƠNG 1. TỔNG QUAN

## 1.1. Tổng Quan Về Bài Toán

Từ những khách hàng tiêu dùng cho đến những khách hàng rời bỏ doanh nghiệp, tất cả những khách hàng đều có nhu cầu và mong muốn đa dạng. Doanh nghiệp mong muốn khách hàng chi tiêu nhiều hơn từ những chiến dịch tiếp thị chương trình, sản phẩm mới tới khách hàng theo những cách khác nhau. Tuy nhiên, câu hỏi đặt ra là làm thế nào để đưa ra được các chiến dịch tiếp thị phù hợp với những nhóm khách hàng đang có nhu cầu để từ đó tăng tỷ lệ phản hồi từ khách hàng và từ đó tăng doanh số bán hàng. **Bài toán đặt ra là làm thế nào để đánh giá giá trị của từng nhóm khách hàng một cách tương đối chính xác dựa trên hành vi giao dịch lịch sử của khách hàng.**

## 1.2. Lý Do Chọn Lựa Đề Tài

Theo nguyên lý pareto, 20% khách hàng sẽ mang lại 80% doanh số. Do đó doanh nghiệp cần hiểu rõ đặc điểm của từng nhóm khách hàng, chọn ra được nhóm khách hàng mục tiêu và có những chính sách quản trị khách hàng phù hợp, tối ưu nhất. Việc phân chia khách hàng thành các nhóm khác nhau dựa trên nhu cầu mua sắm sẽ giúp doanh nghiệp kinh doanh hiệu quả hơn, marketing đến đúng tập khách hàng để mang lại rất nhiều các lợi ích cho doanh nghiệp.

Có nhiều cách phân khúc khách hàng mà doanh nghiệp có thể áp dụng. Việc phát triển kinh doanh theo phân khúc khách hàng luôn mang lại lợi ích cho doanh nghiệp vì các khách hàng có hành vi tương đồng được gom lại để phát triển các sản phẩm, dịch vụ đi kèm phù hợp với nhu cầu của họ.

Mô hình RFM (Recency, Frequency, Monetary) là một mô hình phân tích khách hàng được sử dụng để đánh giá giá trị của khách hàng dựa trên ba yếu tố: thời gian gần nhất mua hàng (Recency), tần suất mua hàng (Frequency) và giá trị tổng cộng của các giao dịch (Monetary) sẽ giúp chúng ta giải quyết bài toán trên.

Các thuật toán phân cụm như K-means, HAC và Autoencoder có thể được sử dụng để phân loại khách hàng dựa trên các yếu tố RFM. Dưới đây là tác dụng của từng thuật toán:

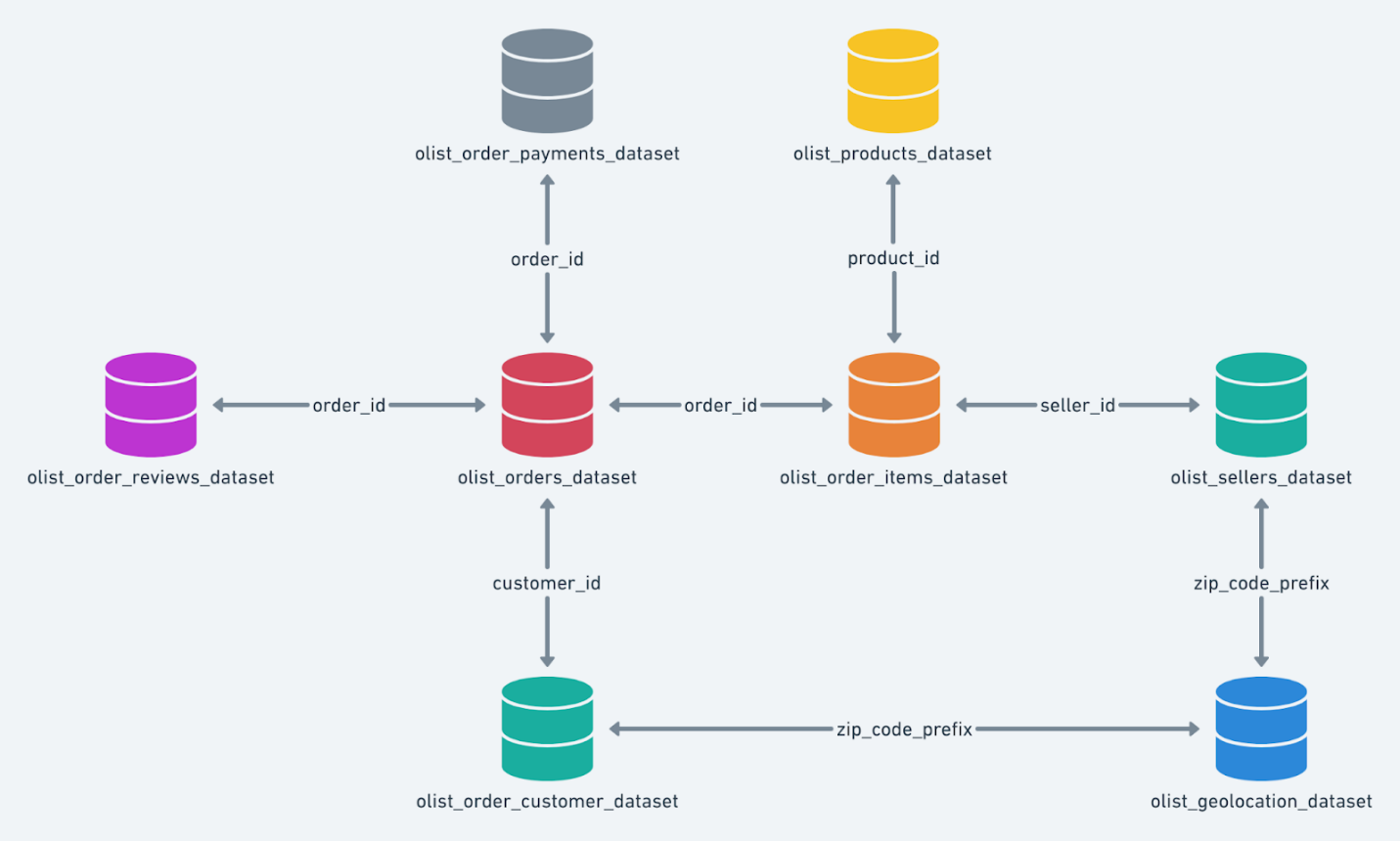
* K-means: K-means có thể được sử dụng để phân loại khách hàng thành các nhóm dựa trên các giá trị RFM của họ. Ví dụ, khách hàng có Recency thấp, Frequency cao và Monetary cao có thể được phân loại vào nhóm cao cấp, trong khi khách hàng có Recency cao, Frequency thấp và Monetary thấp có thể được phân loại vào nhóm thấp cấp.
* Hierarchical clustering (HAC)- phân cụm dữ liệu có thể được sử dụng để phân loại khách hàng thành các nhóm dựa trên các yếu tố RFM của họ, tương tự như K-means. Tuy nhiên, HAC cho phép xây dựng các cấu trúc phân cấp, giúp tạo ra các nhóm con và cha để phân tích dữ liệu một cách chi tiết hơn.
* Autoencoder là phương pháp học sâu có thể được sử dụng để giảm chiều dữ liệu RFM và tạo ra các biểu diễn tối thiểu của dữ liệu. Sau đó, các biểu diễn này có thể được sử dụng để phân loại khách hàng thành các nhóm dựa trên các đặc trưng RFM của họ.

Các phương pháp này có thể giúp doanh nghiệp tìm ra những khách hàng có giá trị cao nhất để tập trung vào việc phát triển mối quan hệ, tăng doanh số bán hàng và chiến lược marketing.

# CHƯƠNG 2. ĐỀ XUẤT MÔ HÌNH

## 2.1. Bộ Dữ Liệu [1]

Về bộ dữ liệu, bộ dữ liệu Olist này được thu thập từ Kaggle, bộ dữ liệu gồm 100k orders từ 2016 đến 2018 từ một trang thương mại điện từ ở Brazil. Và có 8+1 bộ dữ liệu được dùng để sử dụng và khai phá về E- commerce ở Brazil là như thế nào. Data model được mô tả bằng hình ảnh ở dưới và cách nó được tổ chức và liên kết với nhau như thế nào.



Hình 1. Lưu đồ cơ sở dữ liệu của bộ dữ liệu nghiên cứu

Dataset này có thể xem được theo nhiều khía cạnh cho từng mục đích khác nhau, từ thông tin đơn hàng ( gốm tình trạng đơn hàng, giá trị, pương thức thanh toán, sản phẩm,...) cho đến những review được viết bởi khách hàng.

Về chi tiết hơn ở mỗi dataset:

* Olist \_order\_payments\_dataset: chưa các thông tin đơn hàng, gồm các giá trị:
  + Order\_id:mã đơn hàng
  + Customer\_id: khóa để liên kết với dataset customers
  + Order\_status:
  + Order\_purchase\_timestamp:
  + Order\_approved\_at:
  + Order\_deliveried\_carrier\_date
  + Order\_deliveried\_customer\_date
  + order\_estimate\_deliverry\_time
* Olist\_product\_dataset: chưa thông tin về sản phẩm:
  + Product\_id
  + Product\_category\_name
  + Product\_name\_length
  + Product\_description/photo\_qty
  + Product weight/length/height
* Olist\_customer\_dataset:
  + Customer\_id: Khóa để liên kết với các dataset khách. Mội khách hàng chỉ có 1 id, nhưng 1 customer\_id có thể có nhiêu order
  + Và các thông tin khác về khachs hàng customer city/ state/zipcode
* Olist\_order\_items\_dataset: bảng chứa thông tin các sản phẩm của các đơn hàng
  + Order\_id: mã đơn hàng
  + Order\_item\_id
  + Product\_id: mã sản phẩm
  + Seller\_id: mã của người bán
  + Shipping\_limit\_date: ngày giới hạn vận chuyển để người bán chuẩn bị hàng
  + Price: giá của mỗi sản phẩm
  + Freight\_value:số lượng sản phẩm
* Olist\_order\_payment\_dataset: chứa thông tin giao dich của mỗi đơn hàng
* Olist\_order\_revirew\_dataset: chưa thông tin về đánh giá sản phẩm của khách hàng

## 2.2. Mô hình đề xuất

Sau khi đã nhập các tệp dữ liệu từ Kaggle, nhóm sẽ hợp nhất các bảng dữ liệu lại với nhau qua các cột khóa. Bảng dữ liệu hợp nhất mới này sẽ bao gồm thông tin đơn hàng, thông tin khách hàng, thông tin nhà bán hàng và hình thức thanh toán. Ở bước tiền xử lý dữ liệu, nhóm sẽ kiểm tra dữ liệu bị thiếu; nếu số lượng dữ liệu bị thiếu ít hơn 10% so với kích thước dữ liệu thì nhóm sẽ lược bỏ các dòng chứa dữ liệu bị thiếu, ngược lại, nhóm sẽ cân nhắc các phương pháp điền một giá trị đại diện cho các dữ liệu bị thiếu. Tiếp theo, nhóm sẽ lọc và lấy các đơn hàng có trạng thái là ‘delivered’ - đã giao hàng thành công. Nhóm sẽ tiến hàng trực quan hóa dữ liệu qua các mô hình để hiểu rõ thêm về đặc điểm, thông tin có được từ bộ dữ liệu này.

Nhóm sử dụng phương pháp RFM để tiến hành phân cụm khách hàng. Dữ liệu cần thiết của phương pháp này chính là ba chỉ số:

* R - recency: khoảng thời gian kể từ lần cuối cùng khách hàng thực hiện giao dịch/mua sắm trên trang thương mại điện tử
* F - frequency: tần số khách hàng mua sắm tại trang thương mại điện tử
* M - monetary: tổng số tiền khách hàng đã chi cho việc mua sắm các sản phẩm trên trang thương mại điện tử

Để tính R, nhóm sẽ tổng hợp thời gian mua sắm gần đây nhất của từng mã khách hàng, sau đó nhóm sẽ tính R bằng cách lấy mốc thời gian mà dữ liệu được thu thập (31/08/2018) trừ đi thời gian mua sắm gần nhất. Với chỉ số F, nhóm sử dụng tổng số lượng mã đơn hàng của từng mã khách hàng. Còn với chỉ số M, nhóm sẽ tính tổng giá trị các đơn hàng của từng mã khách hàng. Tổng hợp kết quả tính R, F, M cùng mã khách hàng vào một bảng dữ liệu mới để phục vụ cho việc phân cụm.

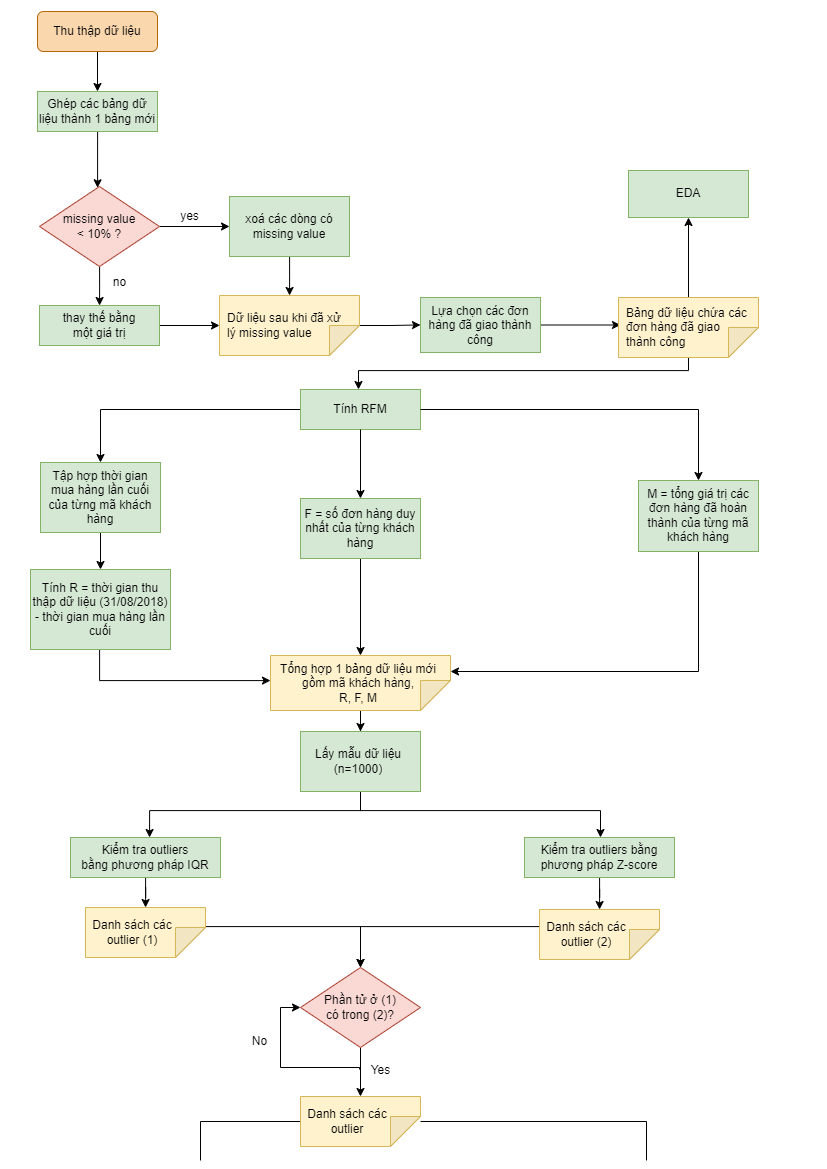
Vì kích thước của bộ dữ liệu gốc khá lớn (hơn 90,000 quan sát), và sự giới hạn về dung lượng của phần mềm chạy lệnh nên nhóm sẽ lấy mẫu ngẫu nhiên với kích thước 10000 phần tử. Sau đó, nhóm sử dụng hai phương pháp IQR và Z-score để xác định các phần tử ngoại biên (outliers). Mỗi phương pháp sẽ cho ra một danh sách các phần tử ngoại biên khác nhau. Nhóm sẽ tiến hành so sánh hai danh sách này, và những phần tử nào xuất hiện trong cả hai danh sách thì sẽ được coi là phần tử ngoại biên cần được xử lý, và chỉ số của các phần tử này sẽ được tổng hợp vào một danh sách.

Ở phần xử lý các phần tử ngoại biên, nhóm không chắc chắn được rằng những phần tử này thực sự gây nhiễu, làm giảm độ chính xác khi phân cụm, vì có thể những phần tử này là một cụm trong bộ dữ liệu, thế nên để chắc chắn, nhóm sẽ tiến hành phân cụm cho hai trường hợp:

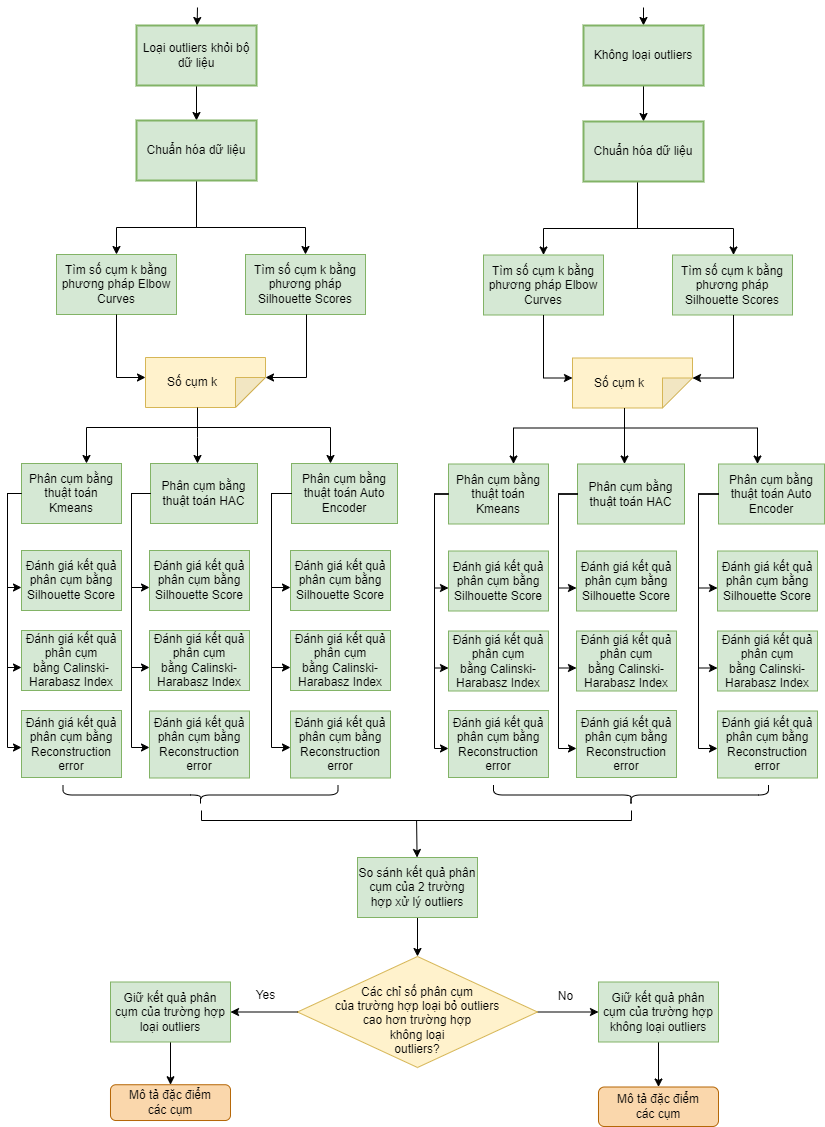
* Trường hợp 1: phân cụm với bộ dữ liệu chưa loại bỏ các phần tử ngoại biên
* Trường hợp 2: phân cụm với bộ dữ liệu đã loại bỏ phần tử ngoại biên

Mỗi trường hợp đều sẽ tiến hành các bước sau: chuẩn hóa dữ liệu, tìm số cụm k bằng phương pháp Elbow Curves và điểm số Silhouette, tiến hành phân cụm sử dụng ba thuật toán: KMeans, Hierarchical Agglomerative Clustering (HAC) và học sâu AutoEncoder; đánh giá các thuật toán phân cụm bằng các chỉ số Silhouette, Calinski-Harabasz và…

Các chỉ số đánh giá của các thuật toán trong cả hai trường hợp xử lý phần tử ngoại biên sẽ được so sánh với nhau, trường hợp nào có các chỉ số đánh giá cao hơn thì sẽ chọn cách xử lý phần tử ngoại biên đó và lấy kết quả phân cụm trong trường hợp đó làm kết quả phân cụm chính thức của bộ dữ liệu. Cuối cùng, nhóm thực hiện trực quan hóa các kết quả phân cụm để hiểu rõ hơn về đặc điểm khách hàng của từng cụm



Hình 2. Mô hình đề xuất (1)

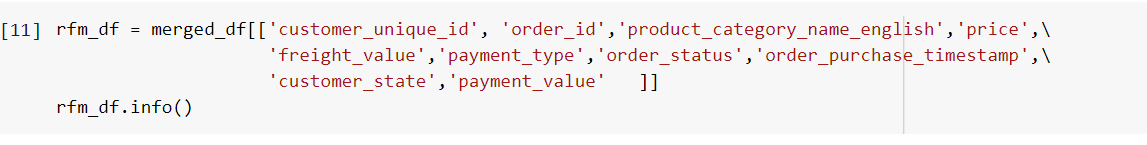


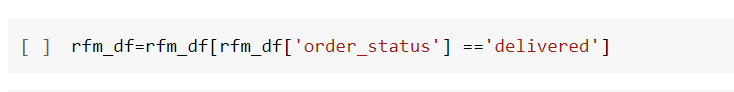
Hình 3. Mô hình đề xuất (2)

# CHƯƠNG 3. TRIỂN KHAI VÀ ĐÁNH GIÁ MÔ HÌNH [2]

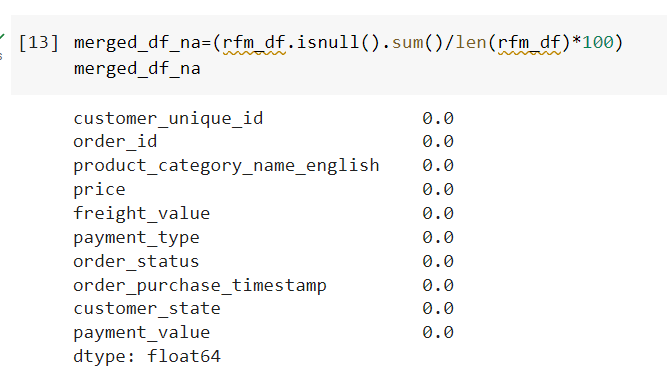
## 3.1. Tiền xử lý dữ liệu

Hợp nhất các dataset  và chọn các thuộc tính cần để phục vụ tính RFM. Các trường được chọn bao gồm: Customer Unique ID, Order  ID, Time of transaction, Payment\_value, Product Category Name English, Order\_status ( chọn các giao dịch đã delivered)



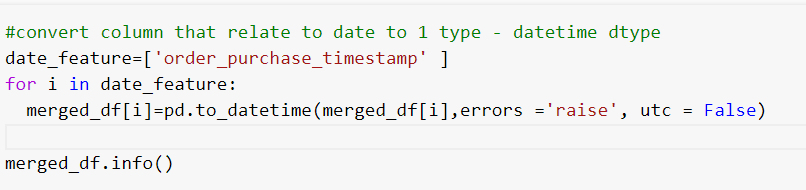


Kiểm tra các giá trị trong tập dữ liệu cho thấy bộ dữ liệu không có giá trị Null



Hình 4. Kết quả bảng dữ liệu đã hợp nhất

Nhận thấy có những thuộc tính về thời gian đang ở kiểu Object do vậy ta cần chuyển biến mang ý nghĩa thời gian thành kiểu Date time



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

### Số lượng sản phẩm trong mỗi đơn hàng:

Số lượng sản phẩm trong mỗi đơn hàng sẽ có hai loại: một là tổng số lượng món hàng (items) trong đơn hàng, hai là tổng số lượng các loại sản phẩm (unique product) trong đơn hàng. Ví dụ, một đơn hàng của khách hàng A có tổng cộng 5 sản phẩm, cụ thể gồm 2 sản phẩm X và 3 sản phẩm Y, vậy tổng số lượng món hàng của đơn hàng này là 5, nhưng tổng số lượng các loại sản phẩm của đơn hàng này là 2 (sản phẩm X và sản phẩm Y). Ở đây, nhóm sẽ phân tích cả 2 loại tổng số.

Với tổng số lượng món hàng trong mỗi đơn hàng, nhóm sử dụng hàm .groupby() trên cột mã đơn hàng ‘order\_id’ và đếm các đơn giá của những sản phẩm thuộc đơn hàng đó ‘price’.



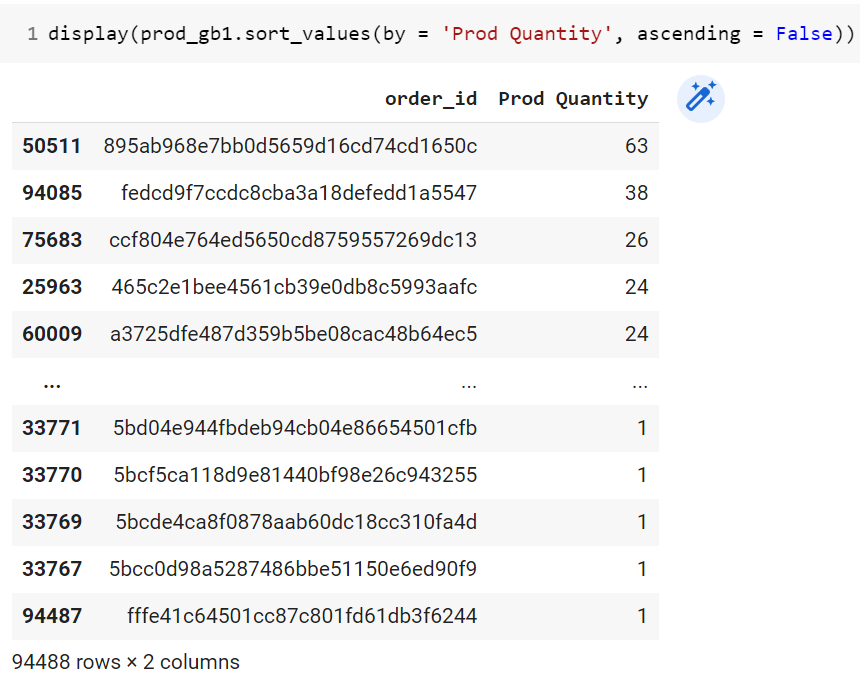
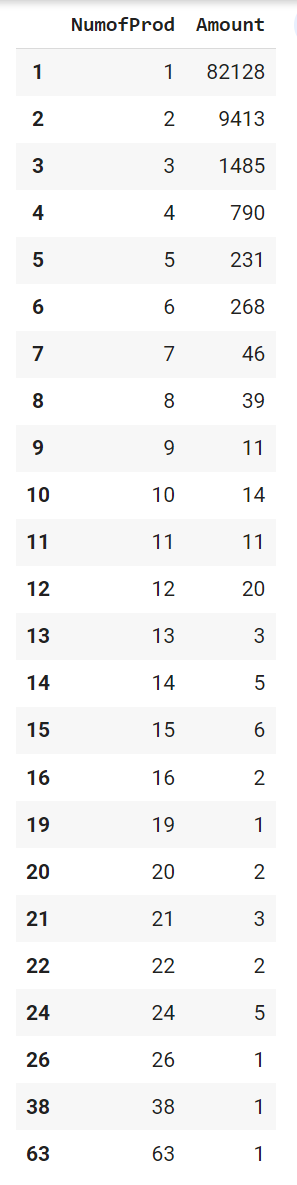
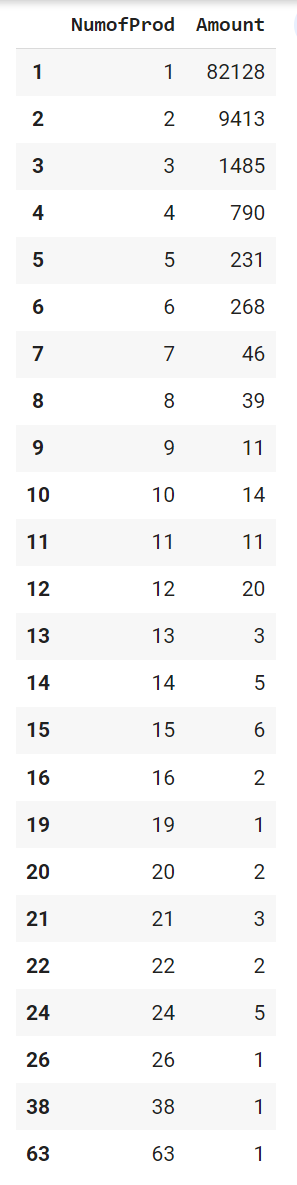


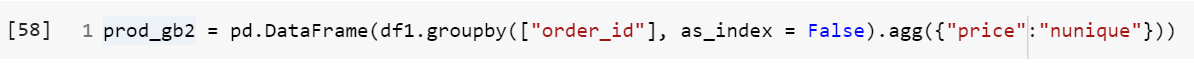
Table 1. bảng kết quả tổng hợp số lượng món hàng của mỗi mã đơn hàng được sắp xếp theo chiều giảm dần



Hình 5. bảng thống kê số lượng món hàng của mỗi mã đơn hàng

Các đơn hàng chứa chủ yếu từ 1 đến 6 món hàng, trong đó đơn hàng lẻ mua 1 sản phẩm là phổ biến nhất, chiếm khoảng 86,92% trong số các đơn hàng. Các đơn hàng lớn (>15 món) chỉ xuất hiện nhỏ giọt, không quá 10 đơn. Đơn hàng lớn nhất đã giao hàng thành công của trang thương mại điện tử này chứa 63 món hàng.

Tuy nhiên, liệu đơn hàng lớn nhất chứa 63 món hàng này có thực sự mua nhiều sản phẩm khác nhau hay chỉ gồm vài loại sản phẩm và mua với số lượng sỉ? Để giải đáp cho thắc mắc này, nhóm vẫn sử dụng hàm .groupby() trên cột mã hóa đơn ‘order\_id’, nhưng số lượng tổng hợp lần này là đếm số lượng các đơn giá duy nhất của từng mã đơn ‘price’ : ‘nunique’.



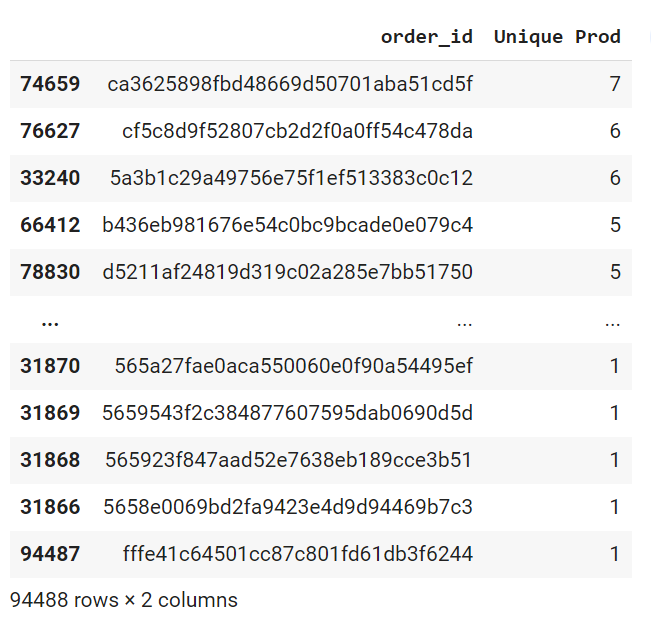


Table 2. bảng kết quả tổng hợp số lượng các loại sản phẩm của mỗi đơn hàng

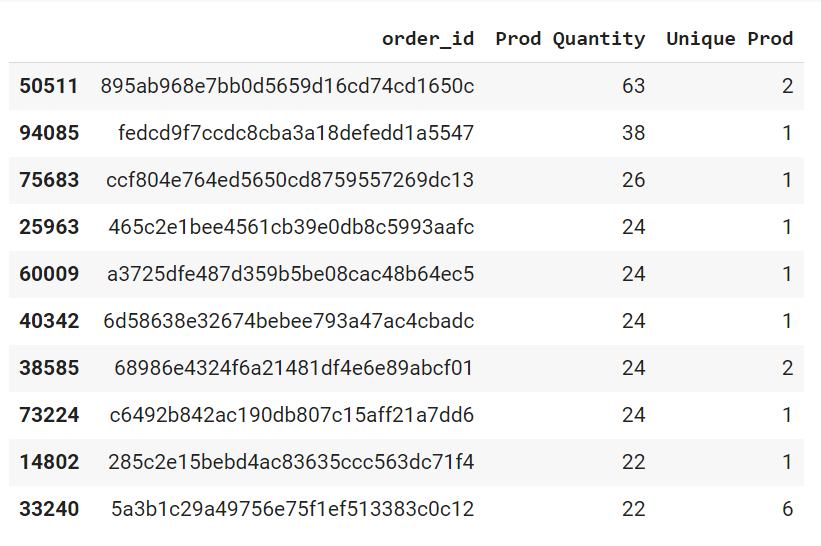


Table 3. hợp nhất hai bảng tổng hợp số lượng sản phẩm của mỗi đơn hàng

Nhìn vào bảng ta có thể thấy, không phải đơn hàng lớn thì độ đa dạng các loại sản phẩm sẽ cao. Như ở đơn hàng lớn nhất với 63 món hàng, số loại sản phẩm trong đơn hàng này thực chất chỉ có 2. Các đơn hàng tiếp theo cũng như vậy, tuy số lượng món hàng nhiều nhưng cũng chỉ gồm 1 đến 2 loại sản phẩm.

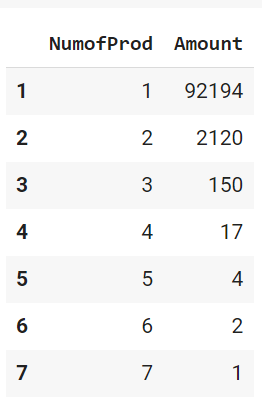
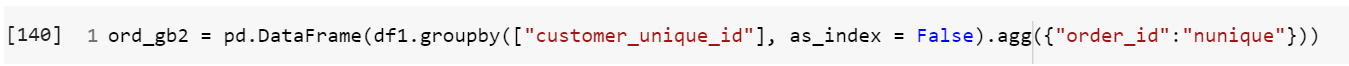


Table 4. bảng thống kê số lượng loại sản phẩm của mỗi mã đơn hàng

### Số lượng đơn hàng của mỗi khách hàng:

Để tính số lượng đơn hàng của mỗi khách hàng, nhóm sử dụng hàm .groupby() trên cột mã khách hàng “customer\_unique\_id” với số liệu cần tổng hợp là đếm số lượng các mã đơn hàng “order\_id” : “nunique”.



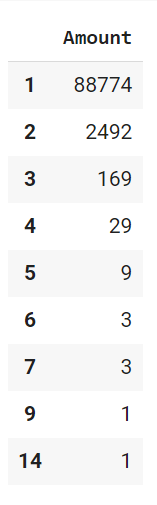


Table 5. bảng thống kê số lượng khách hàng tương ứng với mỗi số lượng đơn hàng

Các khách hàng đa số chỉ mua hàng 1 lần qua trang thương mại điện tử này, chứng minh qua con số hơn 97% (88,774 / 91481) khách hàng có 1 đơn hàng từ trang O-list. Điều này cho thấy khách hàng chưa có thói quen mua hàng trực tuyến qua trang thương mại điện tử, hoặc trang thương mại điện tử đang xem xét chưa đủ hấp dẫn để giữ chân khách hàng.

### Tình hình mua sắm trực tuyến của khách hàng qua các thời điểm:

Nhóm lọc những dòng trùng mã đơn hàng ‘order\_id’ để lấy thông tin gồm mã đơn hàng và thời gian đơn hàng được chốt trên hệ thống “order\_purchase\_timestamp” bằng hàm .drop\_duplicates() trên cột mã đơn hàng và giữ lại phần tử xuất hiện đầu tiên.



Nhóm sử dụng hàm .groupby() trên cột thời gian “order\_purchase\_timestamp” theo tháng-năm, và dữ liệu tổng hợp là đếm số đơn hàng thực hiện trong các khoảng thời gian.



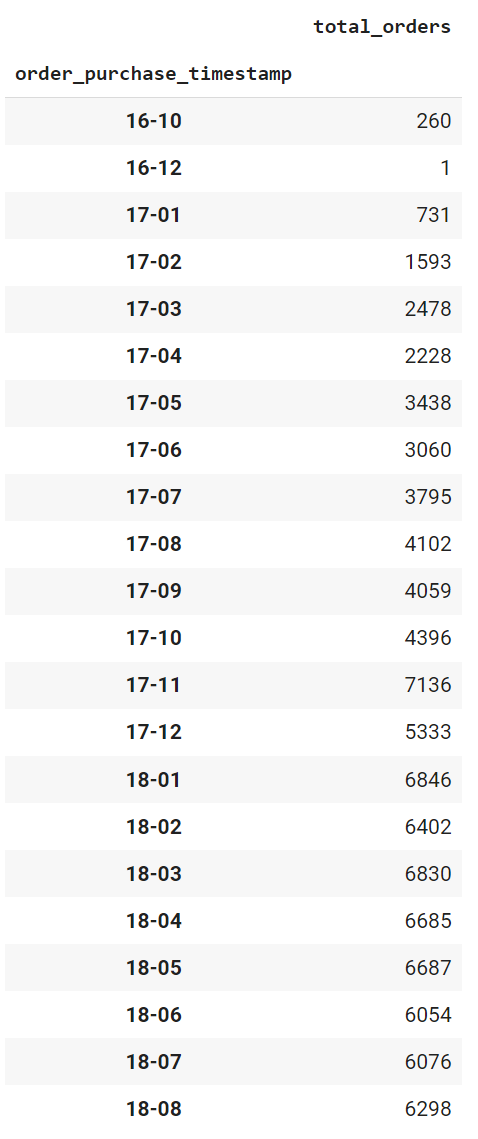
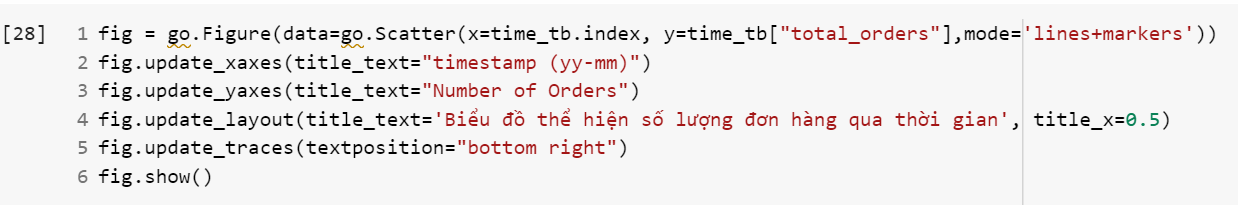
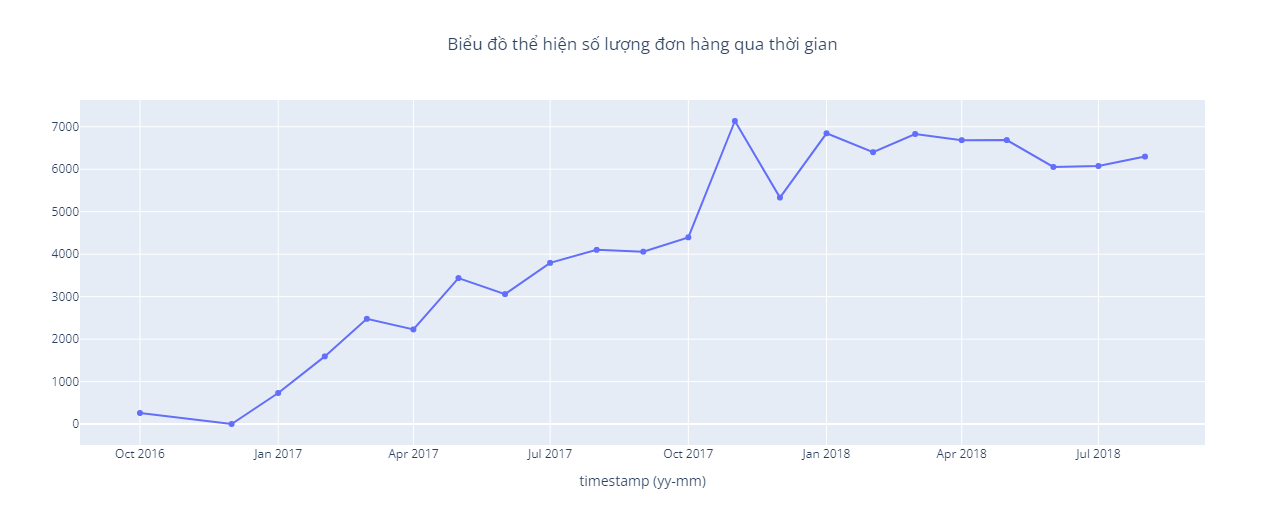


Table 6. bảng kết quả tổng hợp số lượng đơn hàng trong các khoảng thời gian

Để trực quan hóa bảng kết quả tổng hợp trên, nhóm sử dụng biểu đồ đường với trục hoành là các mốc thời gian, trục tung là số lượng đơn hàng tương ứng.





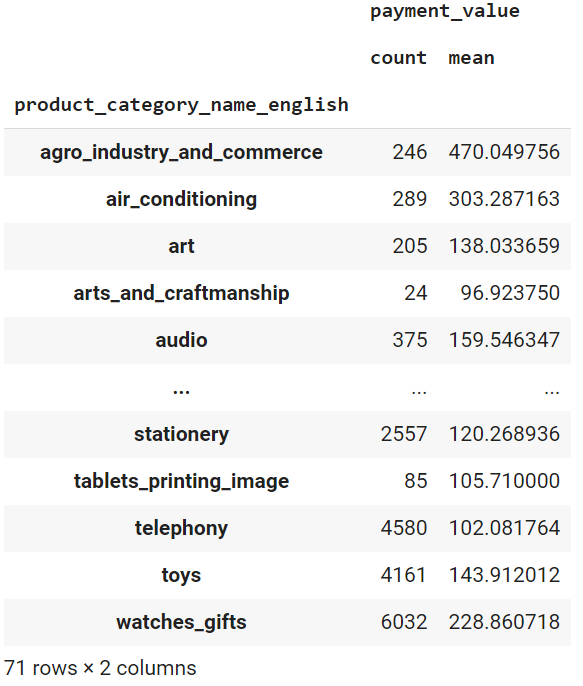
Hình 6. biểu đồ đường thể hiện số lượng đơn hàng qua thời gian

Biểu đồ cho thấy, số lượng đơn hàng tăng khá ổn định ở 3 tháng đầu năm 2017, và 4 tháng tiếp theo có sự giao động nhẹ nhưng vẫn giữ được đà tăng trưởng. Tới tháng 11 năm 2017 - thời điểm cuối năm, lễ hội và chuẩn bị cho năm mới - số lượng đơn hàng trực tuyến tăng bất ngờ và đạt đỉnh với con số 7136 đơn. Tuy rằng các tháng sau có những sự biến động không quá lớn nhưng nhìn chung, việc mua sắm trực tuyến đã dần trở nên phổ biến qua chiều tăng số lượng đơn hàng.

### Các mặt hàng phổ biến:

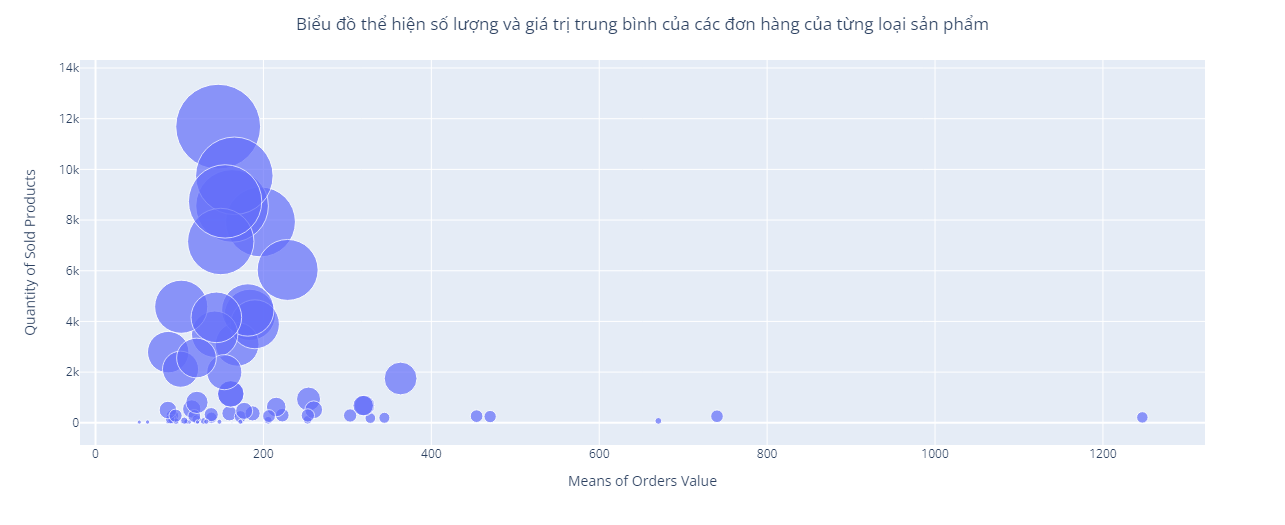
Nhóm sử dụng hàm .groupby() trên cột tên các loại sản phẩm “product\_category\_name\_english” với dữ liệu tổng hợp là số lượng đã bán và doanh thu trung bình của từng loại sản phẩm trên cột giá trị hóa đơn “payment\_value”.





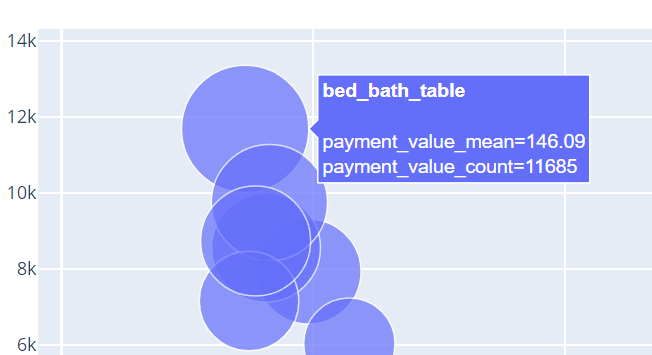
Hình 7. bảng kết quả tổng hợp số lượng đã bán và doanh thu trung bình của từng loại sản phẩm

Để trực quan hóa dữ liệu, nhóm sử dụng biểu đồ bong bóng với trục hoành là giá trị trung bình của đơn hàng từng loại sản phẩm, trục tung và độ lớn của bong bóng là số lượng sản phẩm đã bán được của từng loại sản phẩm.



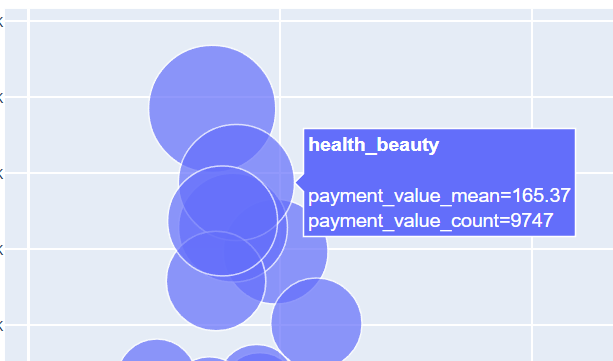
Hình 8. Biểu đồ thể hiện số lượng và giá trị trung bình của các đơn hàng của từng loại sản phẩm

5 nhóm sản phẩm có số lượng đã bán cao nhất lần lượt là các nhóm bàn ghế-nhà tắm-phòng ngủ, sức khỏe-làm đẹp, linh kiện máy tính, trang trí nội thất, thể thao giải trí. Đây đều là những nhóm sản phẩm thiết yếu, thân thuộc trong cuộc sống, đa dạng mẫu mã và lựa chọn, đồng thời giá trị cũng không quá cao, theo như trên biểu đồ có thể thấy cả 5 nhóm này đều có giá trị trung bình đơn hàng khá thấp (dao động khoảng 150-200 đơn vị tiền tệ), nên người dùng có thể mua sắm trực tuyến dễ dàng và thuận tiện.

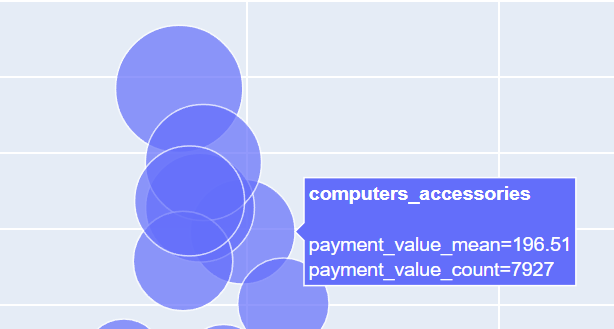


Hình 9. sản phẩm có lượng bán ra cao I

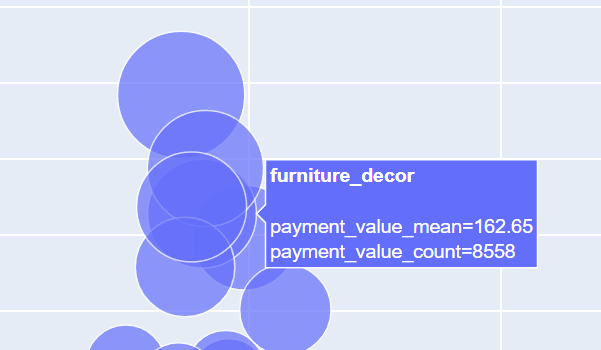
­­­



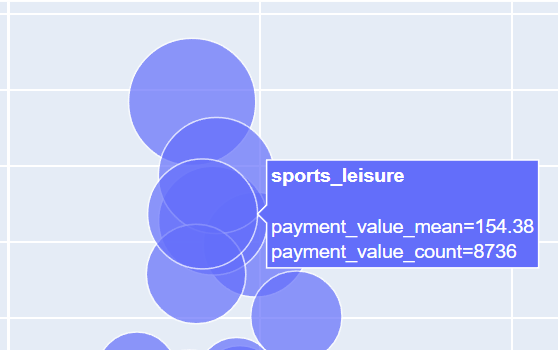
Hình 10. Sản phẩm có lượng bán ra cao II



Hình 11. Sản phẩm có lượng bán ra cao III



Hình 12. Sản phẩm có lượng bán ra cao IV

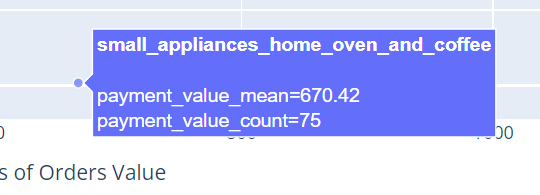


Hình 13. Sản phẩm có lượng bán ra cao V

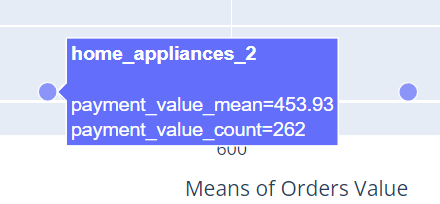
Các sản phẩm có giá trị cao hoặc thuộc nhóm sản phẩm phức tạp, khá kén chọn người dùng, có giá trị cao, cần nhiều sự cân nhắc trước khi mua như máy tính, máy pha cà phê, thiết bị điện gia dụng,... lại chưa thực sự đắt hàng.



Hình 14. Các thiết bị có lượng bán ra thấp I



Hình 15. Các thiết bị có lượng bán ra thấp II



Hình 16. Các thiết bị có lượng bán ra thấp III

### Hình thức thanh toán phổ biến:

Nhóm lập 1 bảng tạm gồm 3 cột: hình thức thanh toán, giá trị đơn hàng, mã đơn hàng, và lọc những dòng trùng mã đơn hàng ‘order\_id’ bằng hàm .drop\_duplicates(), giữ lại phần tử xuất hiện đầu tiên.



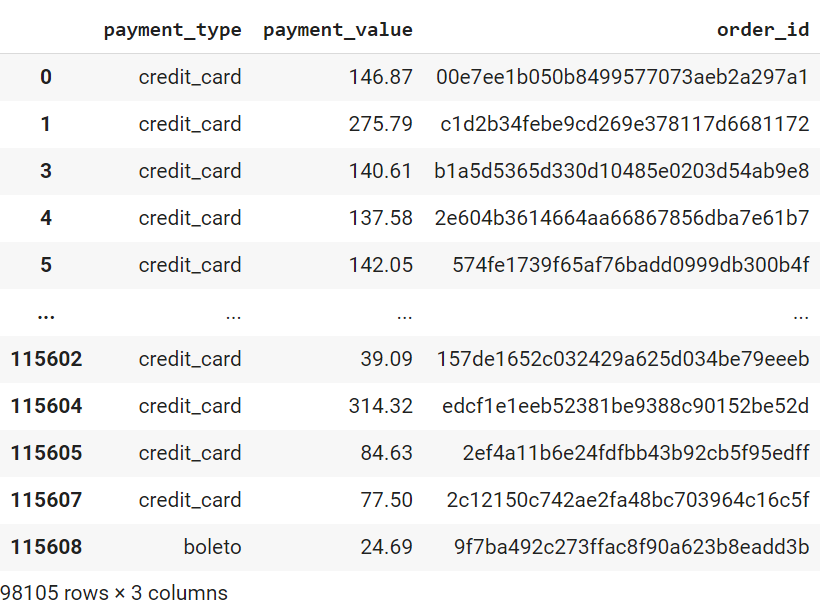


Table 7. Bảng tạm chứa thông tin về thanh toán của từng mã đơn hàng

Nhóm sử dụng hàm .groupby() trên 2 cột hình thức thanh toán và giá trị đơn hàng, với dữ liệu cần tổng hợp là số đơn hàng tương ứng với từng cặp hình thức thanh toán-giá trị đơn hàng.



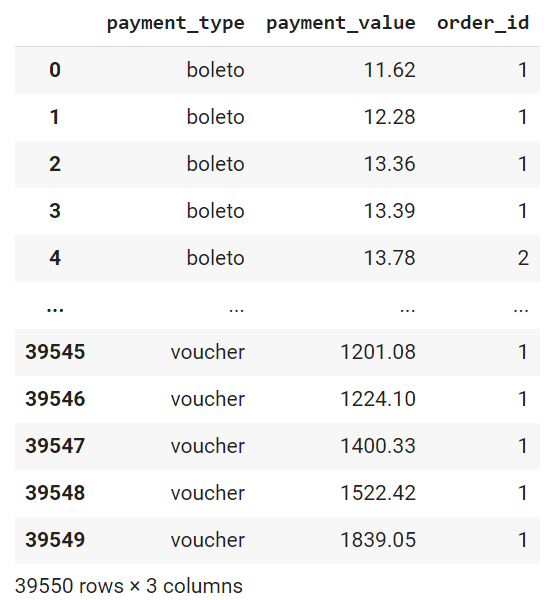
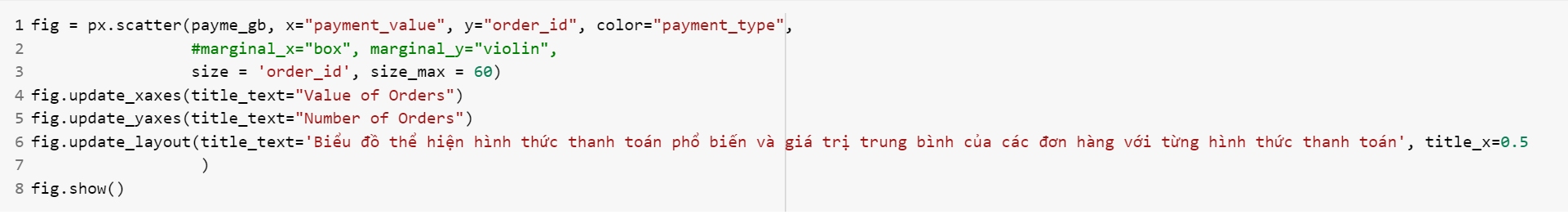
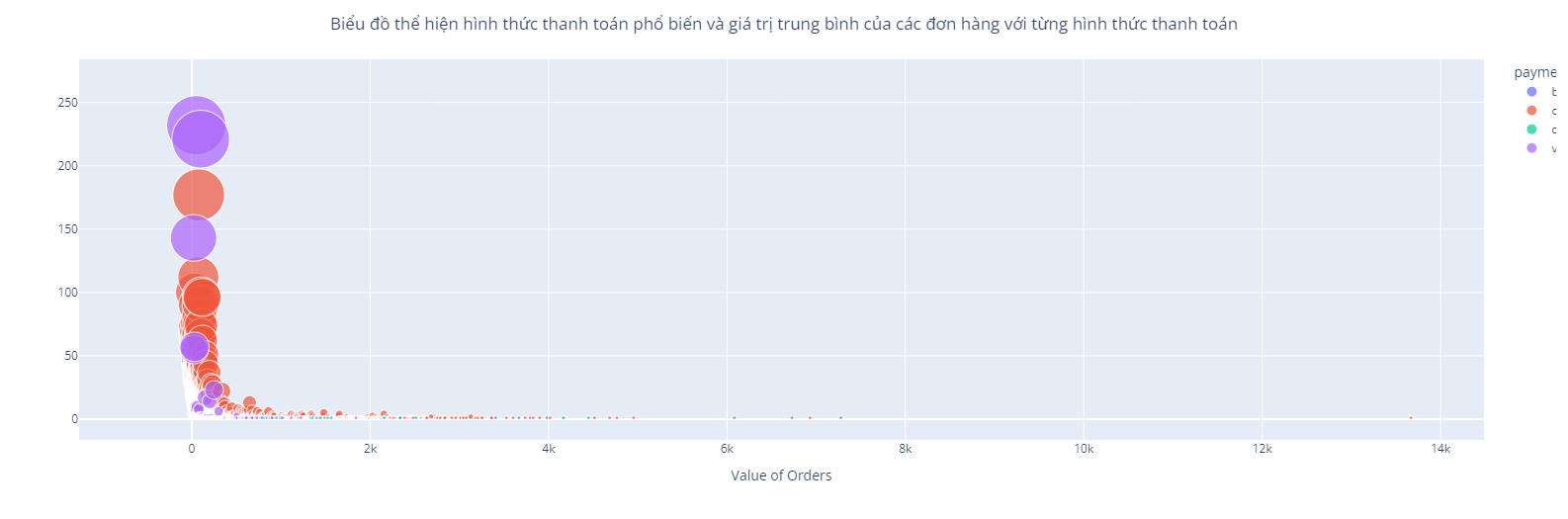


Table 8. Bảng kết quả tổng hợp số đơn hàng tương ứng với từng cặp hình thức thanh toán-giá trị đơn hàng

Nhóm sử dụng biểu đồ bong bóng để thể hiện hình thức thanh toán phổ biến và giá trị trung bình của các đơn hàng với từng hình thức thanh toán, với trục hoành là giá trị đơn hàng, trục tung và độ lớn bong bóng là số lượng đơn hàng, màu của các bong bóng là các hình thức thanh toán.

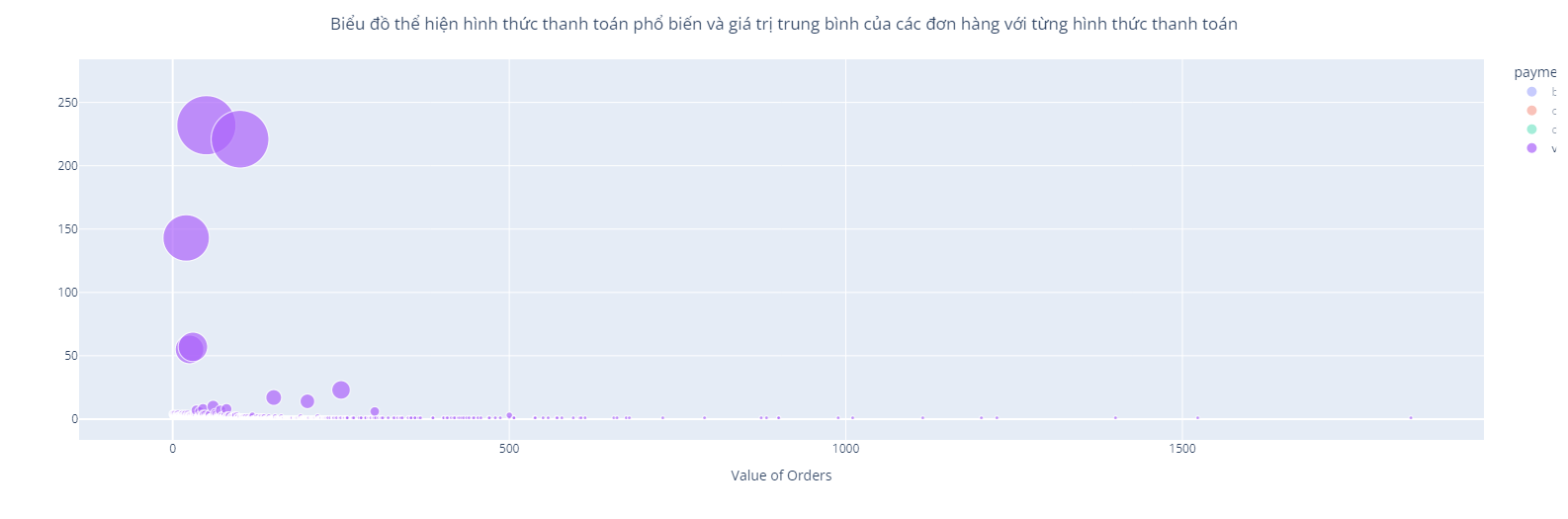




Hình 17. Biểu đồ thể hiện hình thức thanh toán phổ biến và giá trị trung bình của các đơn hàng với từng hình thức thanh toán

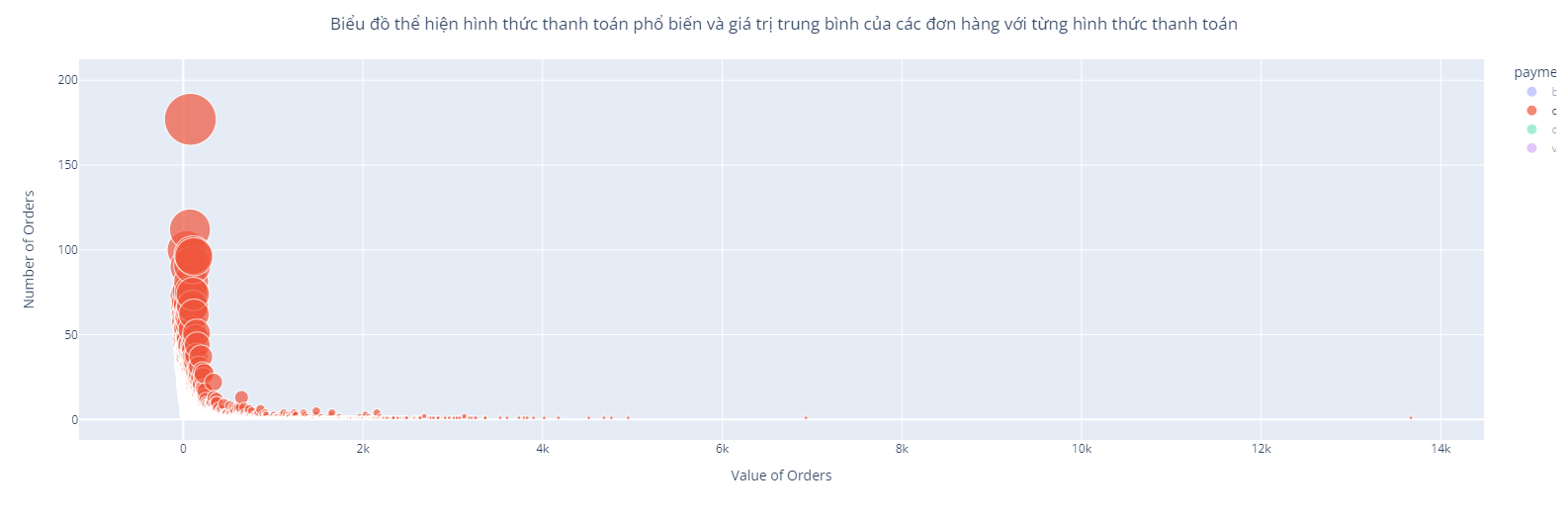
Có 4 hình thức thanh toán được sử dụng trên trang thương mại điện tử O-list, đó là: boleto - một dạng tiền mặt ở Brazil, thẻ ghi nợ (debit card), thẻ tín dụng (credit card), phiếu mua hàng (voucher). Hai hình thức thanh toán phổ biến nhất là bằng phiếu mua hàng (voucher) và thẻ tín dụng (credit card). Cụ thể:

* Hình thức thanh toán bằng phiếu mua hàng (voucher) được áp dụng nhiều cho các đơn hàng có giá trị nhỏ, dao động từ 0-100 đơn vị tiền tệ. Vì là phiếu mua hàng nên giá trị ưu đãi không cao, do đó đơn hàng cao nhất áp dụng hình thức thanh toán này có giá trị khoảng 1,500 đơn vị tiền tệ.



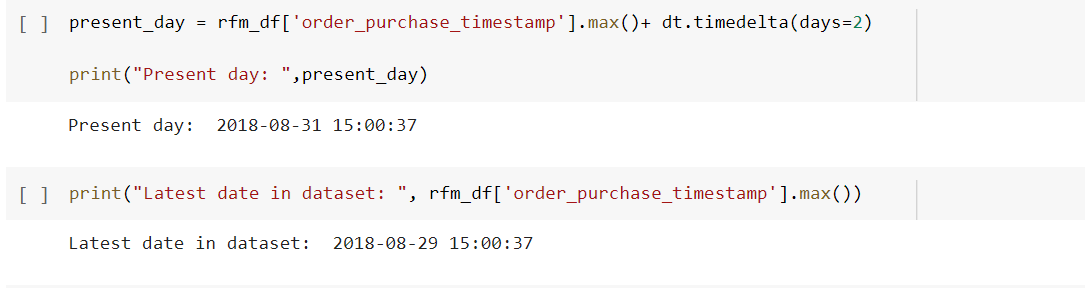
Hình 18. Tình sử dụng hình thức thanh toán phiếu mua hàng - voucher

* Hình thức thanh toán bằng thẻ tín dụng (credit card) phổ biến nhất, được áp dụng nhiều cho các hóa đơn có giá trị dao động từ 30-2,000 đơn vị tiền tệ

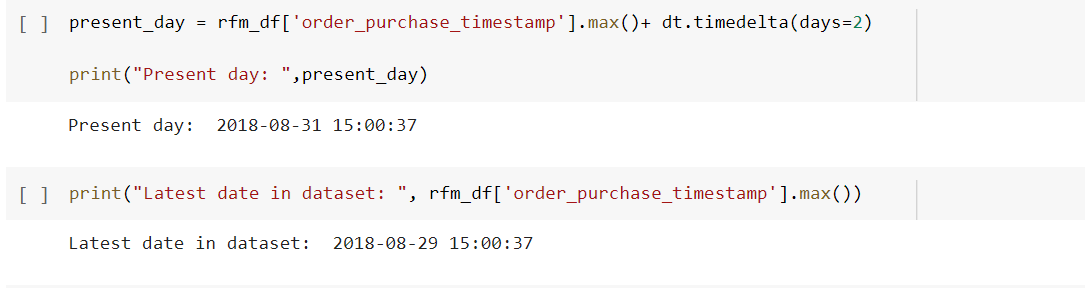


Hình 19. Tình hình sử dụng hình thức thanh toán thẻ tín dụng - credit card

## 3.3. Tính RFM



Ngày mua hàng cuối cùng



Ngày thu thập dữ liệu

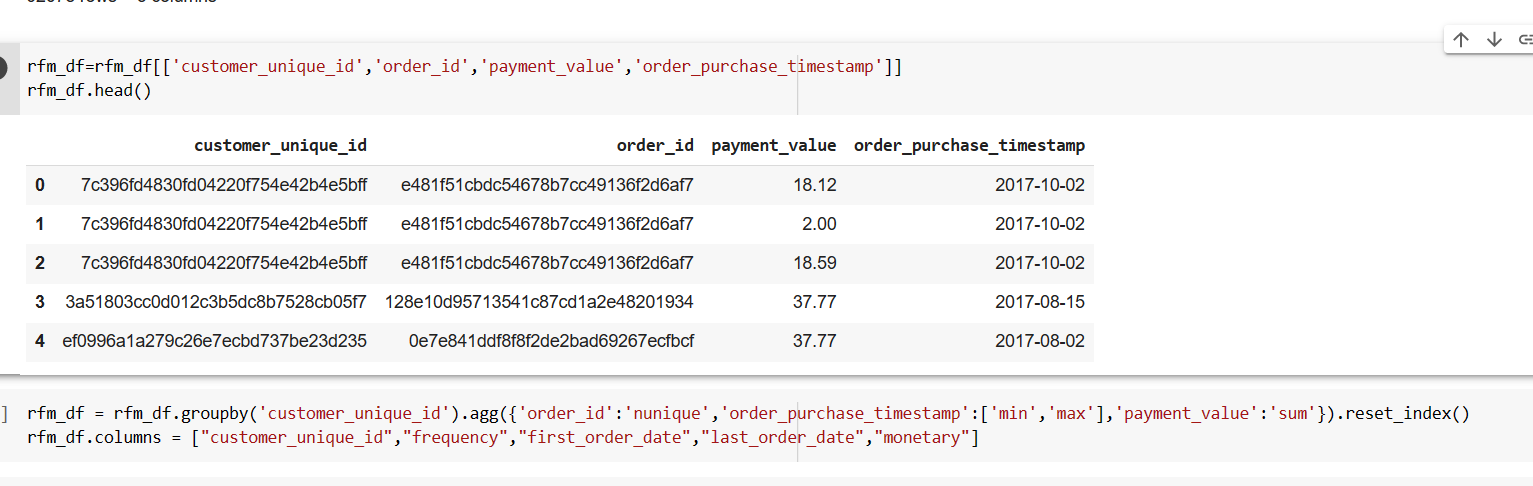
Các chỉ số RFM:

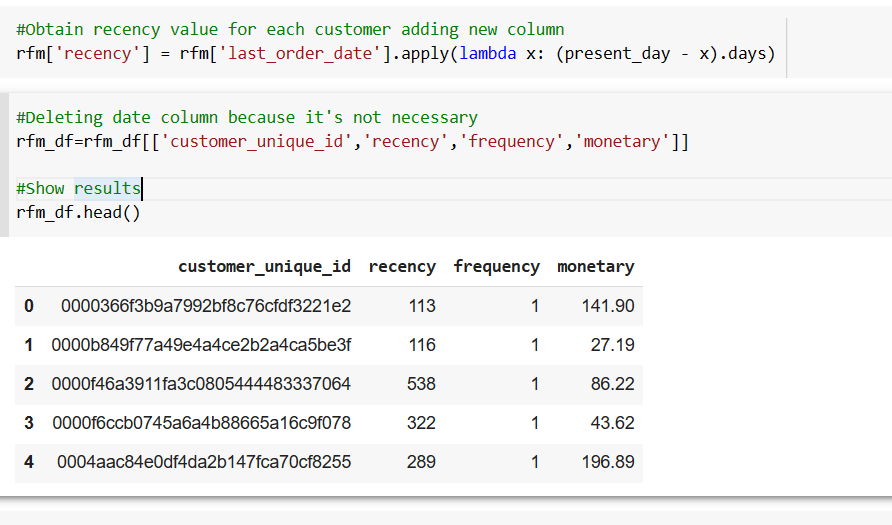
- Recency, có được bằng cách tính số ngày trôi qua từ ngày mua hàng cuối cùng của khách hàng đến thời điểm dữ liệu được thu thập (31-08-2018)

Dữ liệu lần cuối cùng cập nhật hệ thống là ngày 29/08 nhưng theo tập dữ liệu được thu thập vào ngày 31/08 nên chúng tôi giả định ngày hiện tại là ngày 31/08)

- Frequency tính số lượng đơn đặt hàng duy nhất. Sử dụng phương thức groupby để nhóm các đơn hàng theo customer\_unique\_id, sau đó sử dụng phương thức nunique để đếm số lượng đơn hàng duy nhất của mỗi khách hàng

- Monetary: Tổng số tiền giao dịch chi tiêu. Trong trường hợp dữ liệu này ta tính tổng giá trị thanh toán theo ID duy nhất của từng khách hàng



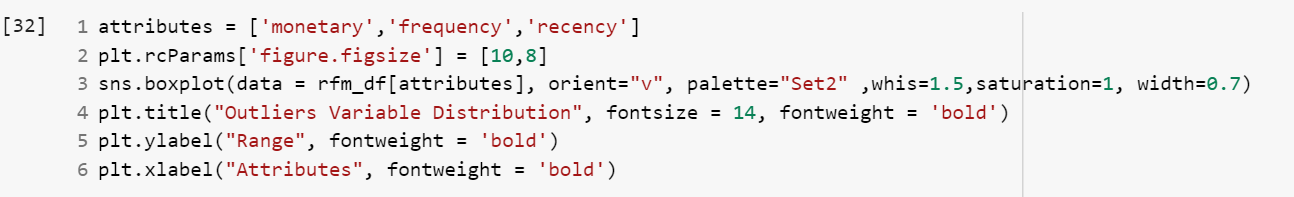


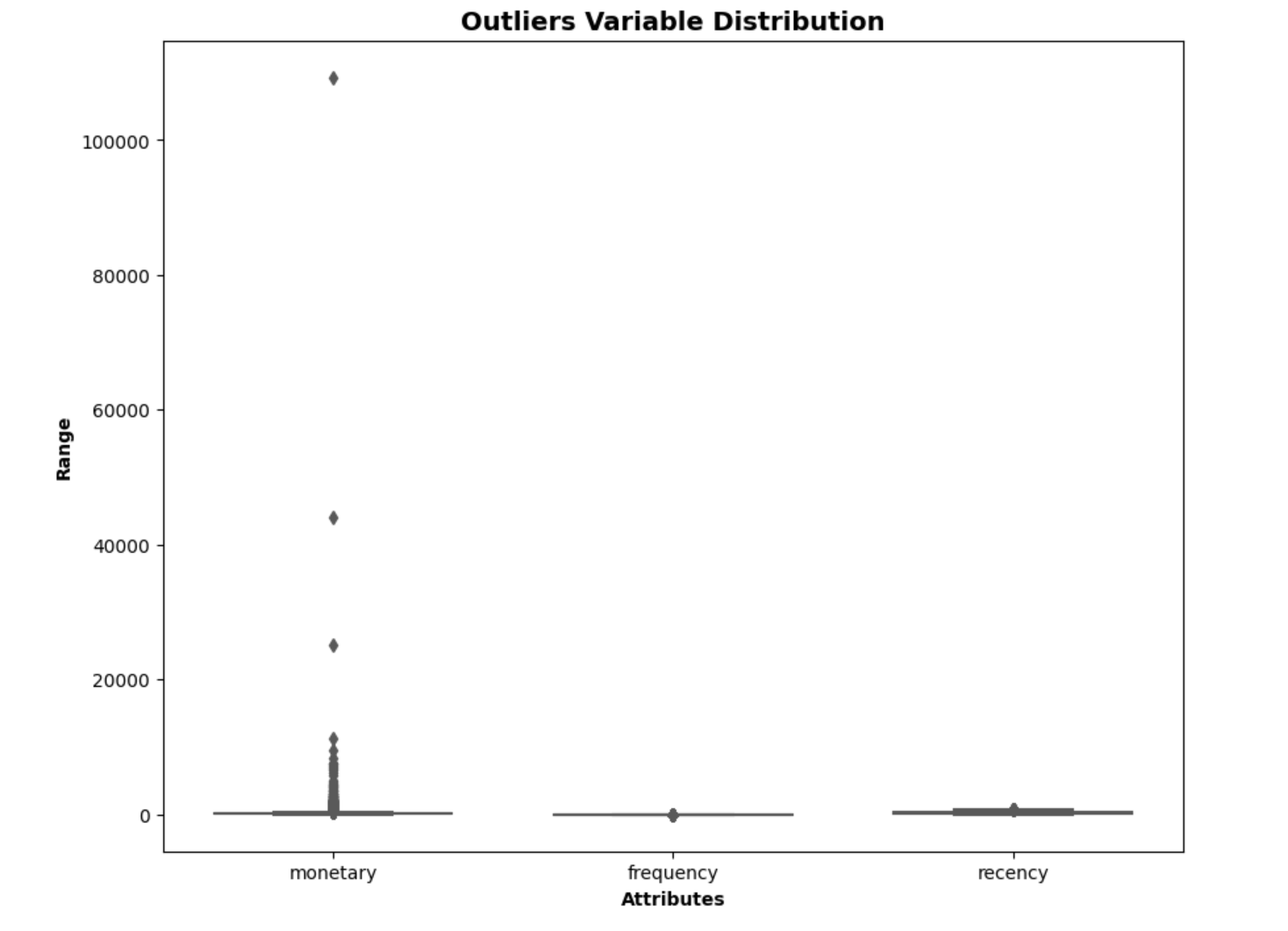
## 3.4. Lấy mẫu và xác định các phần tử ngoại biên (outliers)

Vì kích thước của bộ dữ liệu gốc khá lớn (hơn 90,000 quan sát), và sự hạn chế về dung lượng của Google Colab, nên nhóm sẽ lấy mẫu ngẫu nhiên với kích thước n=10,000 bằng hàm .sample().



Nhóm kiểm tra bộ dữ liệu sau khi lấy mẫu có chứa các phần tử ngoại biên (outliers) hay không bằng biểu đồ hộp (box plot).

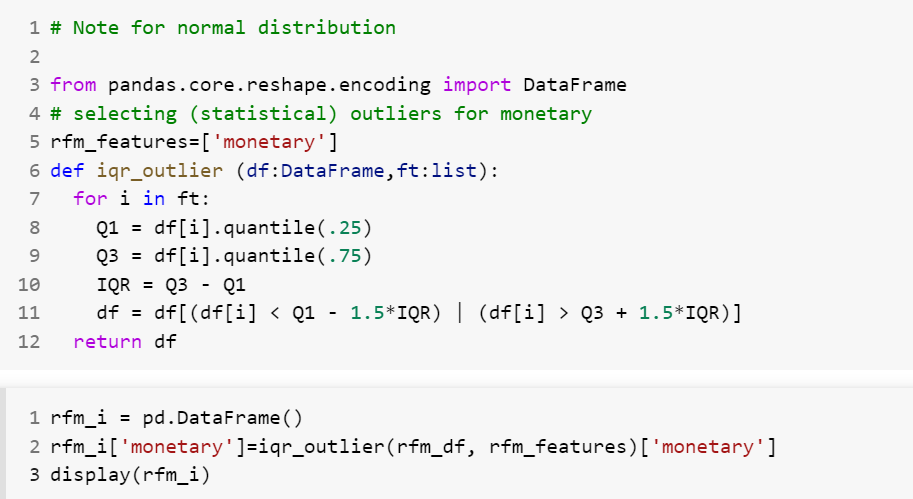




Hình 20. Biểu đồ hộp của 3 biến R, F, M

Nhìn vào biểu đồ hộp của 3 biến, ta có thể thấy rằng biến monetary chứa khá nhiều phần tử ngoại biên - outliers. Vì thế, nhóm sẽ áp dụng 2 phương pháp tìm kiếm phần tử ngoại biên - outliers trên cột ‘monetary’, đó là phương pháp IQR và Z-score.

Ở phương pháp IQR, nhóm xây dựng hàm iqr\_outlier() với 2 tham số đầu vào là bảng dữ liệu chứa các phần tử ngoại biên - outliers, và danh sách các thuộc tính cần kiểm tra outliers. Nhóm thiết lập vòng lặp for với từng thuộc tính: chỉ số Q1 là bách phân vị thứ 25 .quantile(.25), chỉ số Q3 là bách phân vị thứ 75 .quantile(.75), IQR = Q3 - Q1. Các phần tử nào có giá trị nhỏ hơn Q1 - 1.5\*IQR hoặc lớn hơn Q3 + 1.5\*IQR thì được xem là phần tử ngoại biên - outliers và thêm vào bảng dữ liệu outliers đã tạo trước đó. Áp dụng hàm iqr\_outlier() này lên cột ‘monetary’.



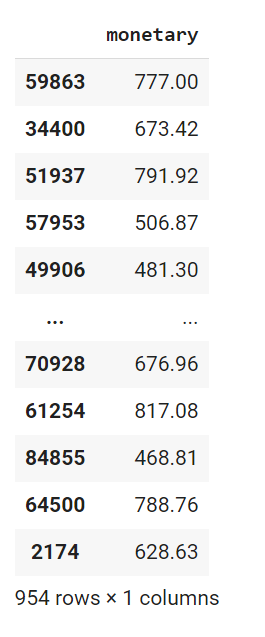
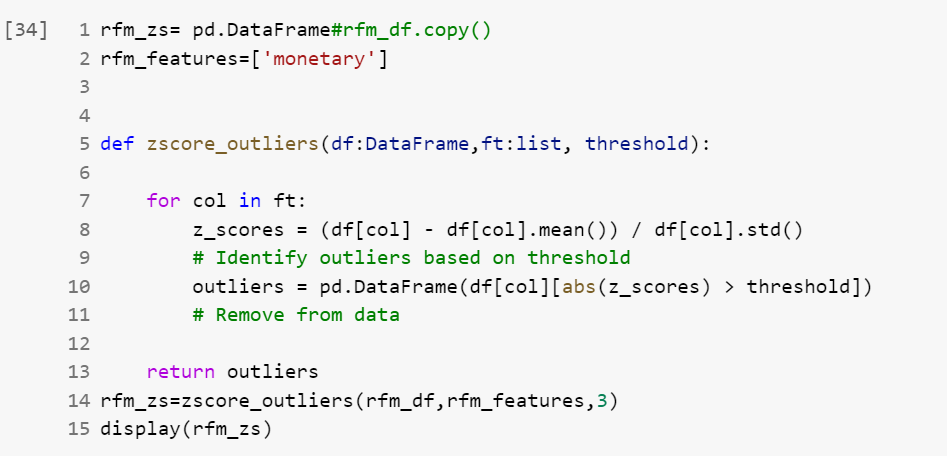


Table 9. danh sách các phần tử ngoại biên (IQR)

Theo phương pháp IQR, bộ dữ liệu mẫu hiện tại đang có 954 phần tử ngoại biên - outliers.

Với phương pháp Z-score, nhóm xây dựng hàm zscore\_outlier() với 3 tham số đầu vào là bảng dữ liệu lưu outliers, danh sách các thuộc tính cần kiểm tra outliers, và ngưỡng giá trị z-score. Nhóm thiết lập vòng lặp for cho từng thuộc tính: z\_score của mỗi phần tử được tính theo công thức (giá trị phần tử / trung bình cột) / độ lệch chuẩn của cột. Các phần tử có giá trị nguyên tuyệt đối của z\_score lớn hơn ngưỡng giá trị z\_score thì sẽ là outliers. Áp dụng hàm zscore\_outlier() lên cột ‘monetary’.



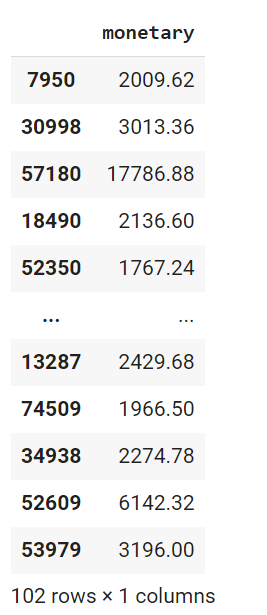
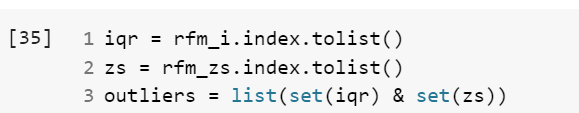


Table 10. danh sách các phần tử ngoại biên (Z-score)

Nhóm tiến hành so sánh 2 danh sách outliers của 2 phương pháp trên, và những phần tử nào xuất hiện trong cả 2 bảng thì nhóm sẽ chọn đó thực sự là những phần tử ngoại biên - outlier. Số thứ tự của các phần tử ngoại biên này sẽ được lưu vào một danh sách.



Sau khi lọc, bộ dữ liệu mẫu hiện tại có 102 phần tử ngoại biên. Tuy nhiên, nhóm không biết liệu các phần tử ngoại biên - outliers này có hại cho quá trình phân cụm hay không, nên để chắc chắn, nhóm sẽ phân cụm với 2 trường hợp:

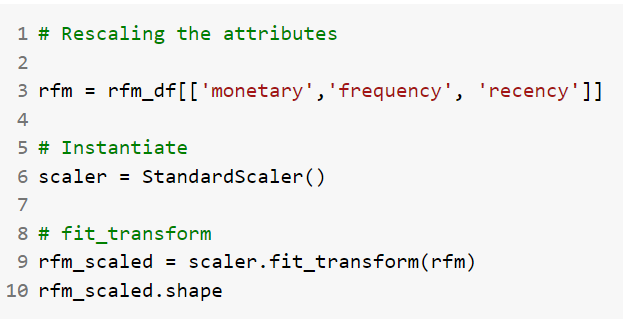
* Trường hợp 1: phân cụm nhưng không loại bỏ outliers
* Trường hợp 2: phân cụm với bộ dữ liệu đã loại bỏ outliers

Trường hợp nào có kết quả phân cụm tốt hơn sẽ được chọn.

## 3.5. Phân cụm dữ liệu

### a. Trường hợp 1: không loại bỏ các phần tử ngoại biên (outliers)

Nhóm chuẩn hóa dữ liệu bằng hàm .StandardScaler() cho 3 cột Recency, Frequency, Monetary.



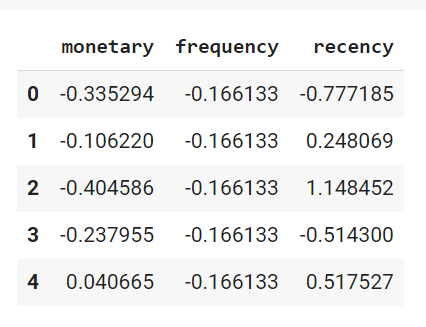
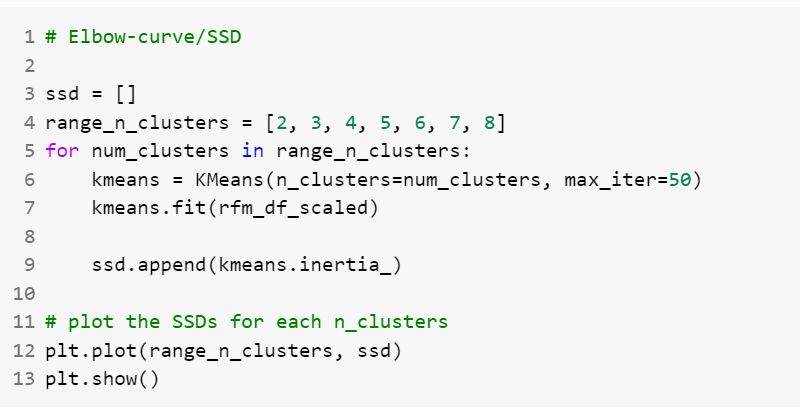
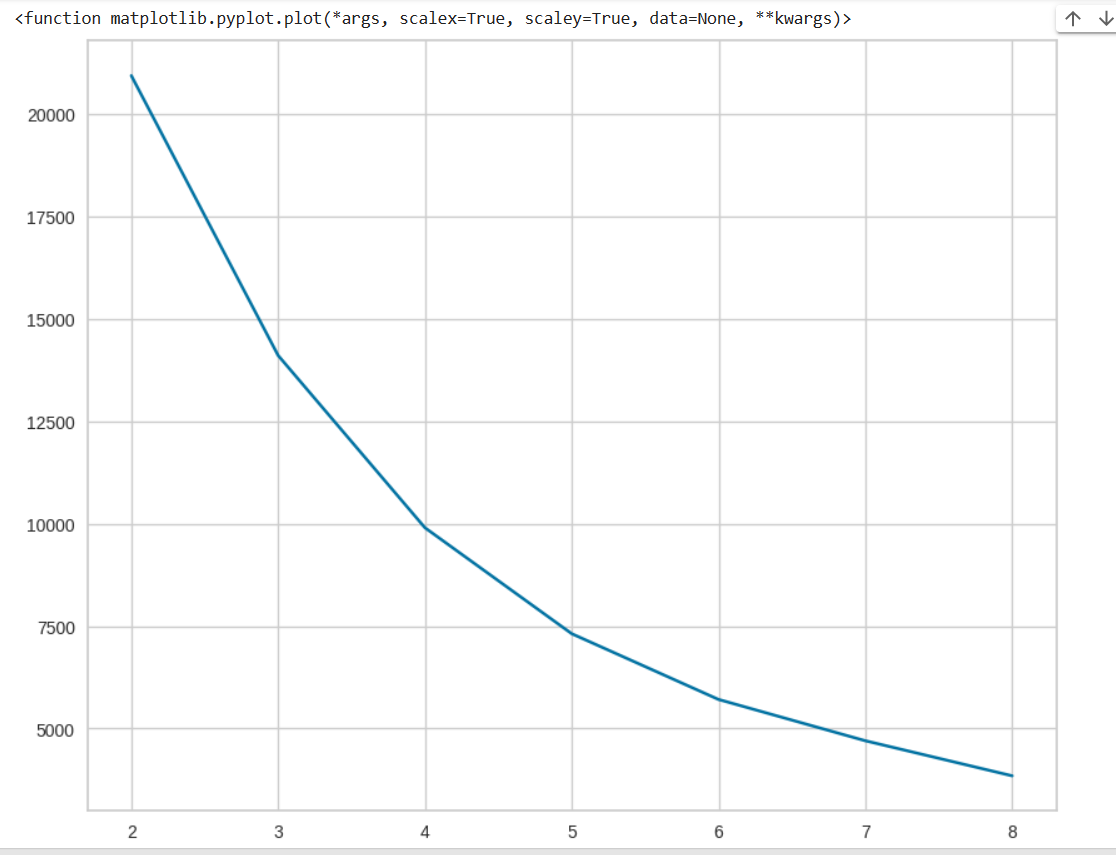


Table 11. Trường hợp 1 - huấn luyện bảng dữ liệu và in bảng kết quả chuẩn hóa

Một điều quan trọng trước khi tiến hành áp dụng các thuật toán phân cụm lên bộ dữ liệu đó là phải xác định được số cụm cần chia. Ở đây, nhóm kết hợp 2 phương pháp tìm số cụm k, đó là phương pháp khuỷu tay - Elbow Curve, và chỉ số Silhouette.

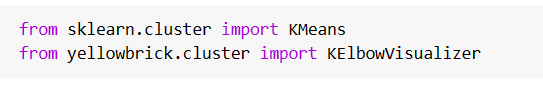
Với phương pháp khuỷu tay Elbow Curve, nhóm tạo 1 danh sách số lượng các cụm, từ 2 đến 8 cụm. Nhóm khởi tạo vòng lặp for: áp dụng thuật toán KMeans lên bộ dữ liệu đã chuẩn hóa, với mỗi số cụm k trong danh sách số lượng cụm đã tạo ở phía trên, sau đó ghi nhận chỉ số inertia, hay còn gọi là SSD (sum of squared distances) - tổng của căn bậc 2 khoảng cách giữa các phần tử đến điểm trung tâm của từng cụm mà phần tử đó thuộc về. Sau đó, nhóm trực quan hóa các chỉ số SSD tương ứng với từng số cụm qua biểu đồ đường .plot(). Từ biểu đồ này, nhóm có thể xác định số cụm k hợp lý tại điểm khuỷu tay - elbow point.

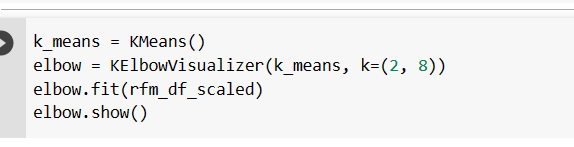


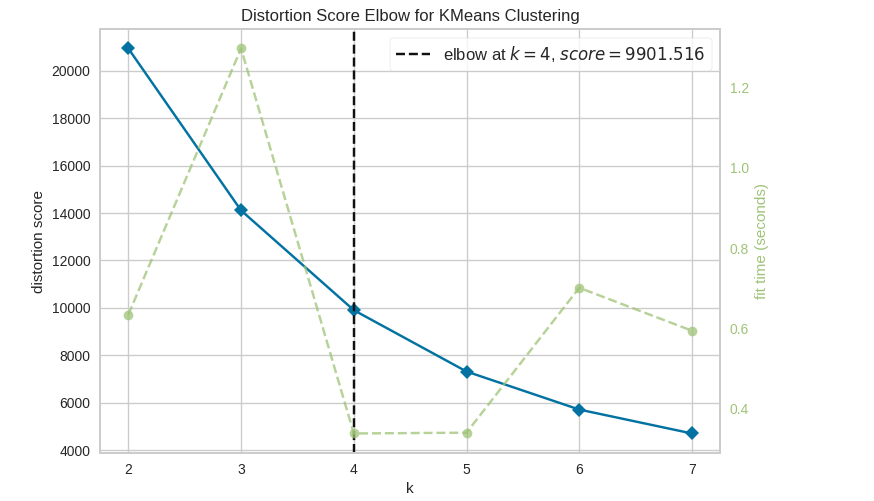


Hình 21.Trường hợp 1 - Biểu đồ Elbow

Dựa theo biểu đồ khuỷu trên thì khó để có thể xác định điểm có số cụm tối ưu là bao nhiêu. Do đó nhóm thực hiện thử cách khác là chọn số cụm tối ưu bằng biểu đồ xác định điểm khuỷu tay tương ứng với số cụm tối ưu cho mô hình KMeans từ thư viện Yellowbrick- một thư viện biểu diễn trực quan mở rộng tương tự Scikit-Learn giúp việc lựa chọn và điều chỉnh tham số cho mô hình dễ dàng hơn.



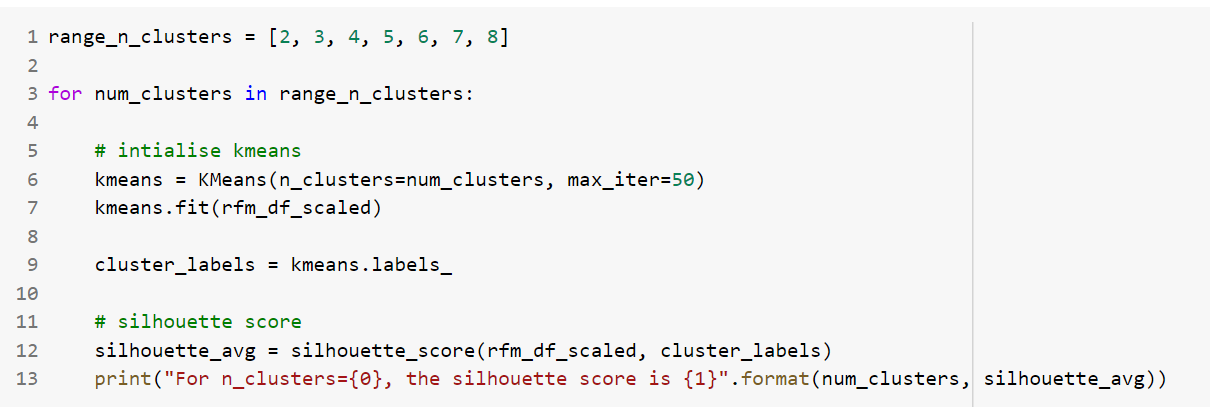


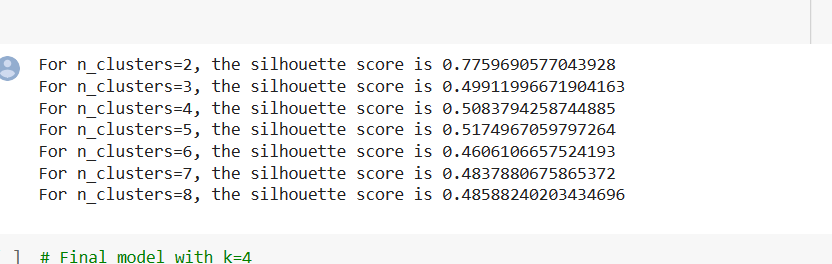


Hình 22. Trường hợp 1 - Distorition score elbow để xác định số lượng cụm tối ưu

Nhìn vào biểu đồ, ta thấy điểm khuỷu tay - Elbow point được cho là tối ưu ở vị trí k=4. Vì thế, số cụm lý tưởng cho việc phân cụm đối với bộ dữ liệu hiện tại nhóm nhận định là 4.

Để đảm bảo số cụm k nhóm chọn là hợp lý, nhóm sẽ xem xét thêm chỉ số Silhouette. Với mỗi số cụm cụ thể, sau khi đã thực hiện phân cụm trên bộ dữ liệu đã được chuẩn hóa, nhóm sử dụng hàm silhouette\_scores() tronng thư viện sklearn. Các số cụm có chỉ số Silhouette trên 0,5 được đánh giá là phân cụm tốt.



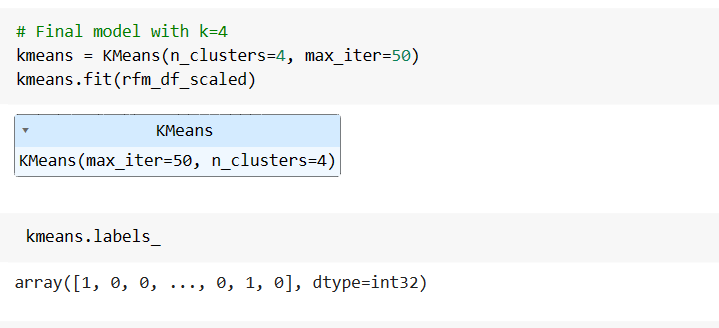


Tại số cụm k=4, chỉ số Silhouette là , một chỉ số Silhouette thể hiện việc phân cụm tốt. Do vậy, nhóm quyết định chọn số cụm k=4.

Sau khi đã xác định được số cụm, nhóm tiến hành phân cụm dữ liệu bằng ba thuật toán: K Means, HAC, học sâu AutoEncoder.

* + 1. KMeans:

Nhóm sử dụng hàm .Kmeans() trong thư viện sklearn.cluster với số cụm n\_clusters=4, số vòng lặp tối đa max\_iter=50. Kết quả phân cụm sẽ được thêm thành 1 cột mới tên là ‘clus\_kmeans’ trong bảng dữ liệu RFM.



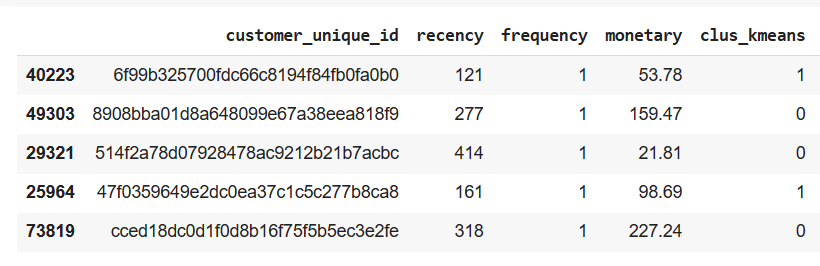
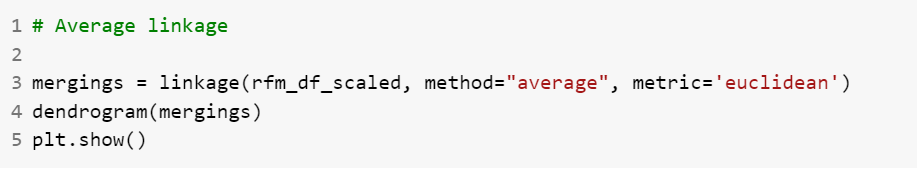
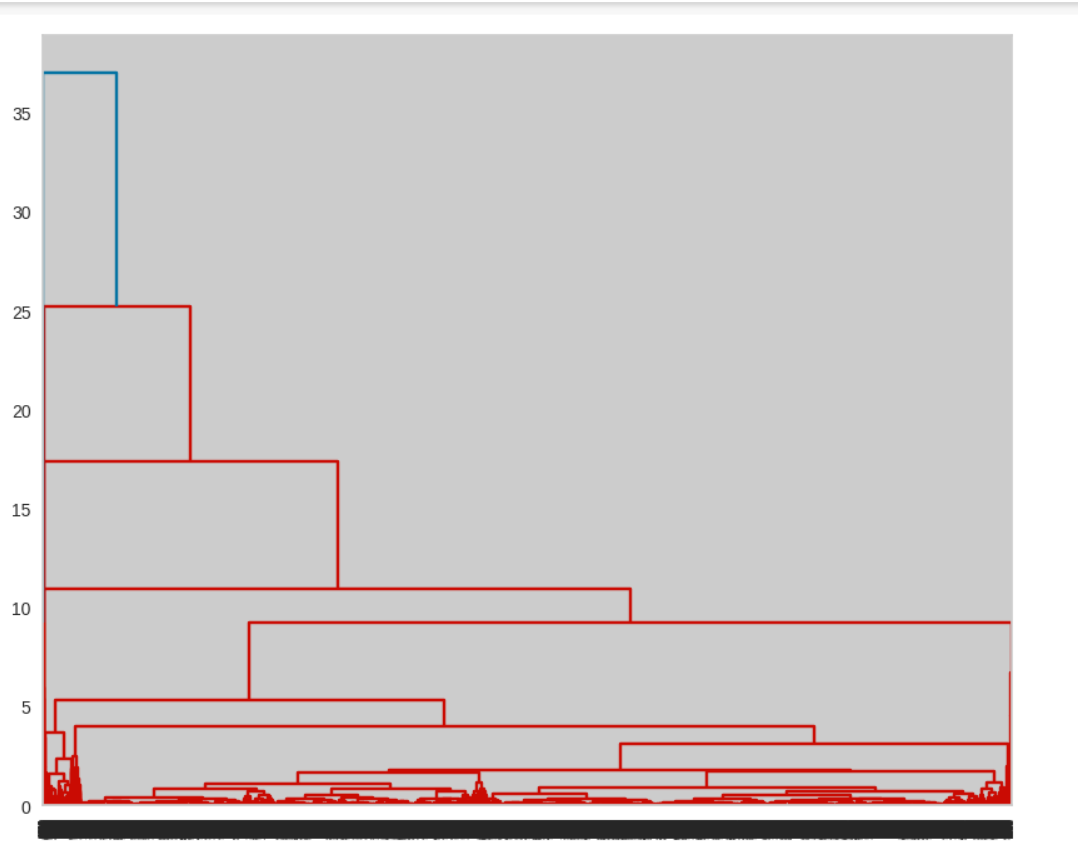


Table 12. Trường hợp 1 - Kết quả phân cụm bằng KMeans được thêm vào bảng RFM gốc

* + 1. HAC:

Ở phương pháp phân cụm HAC, nhóm tính khoảng cách giữa các cụm bằng hàm linkage() trong scipy.cluster.hierarchy, trong đó tham số method="average", metric='euclidean'. Từ dữ liệu tính toán khoảng cách vừa thực hiện, nhóm xây dựng dendrogram, trong đó các điểm dữ liệu và các cụm được đặt ở các nút của cây.





Hình 23. Trường hợp 1 - 1 phần của dendrogram

Sau đó, nhóm cắt cây bằng hàm cut\_tree() trong thư viện scipy.cluster.hierarchy, với số cụm bằng 4. Kết quả phân cụm sẽ được thêm thành 1 cột mới tên là ‘clus\_hac’ trong bảng dữ liệu RFM.



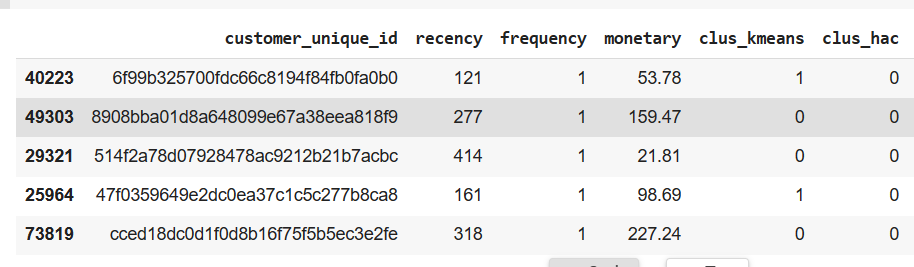
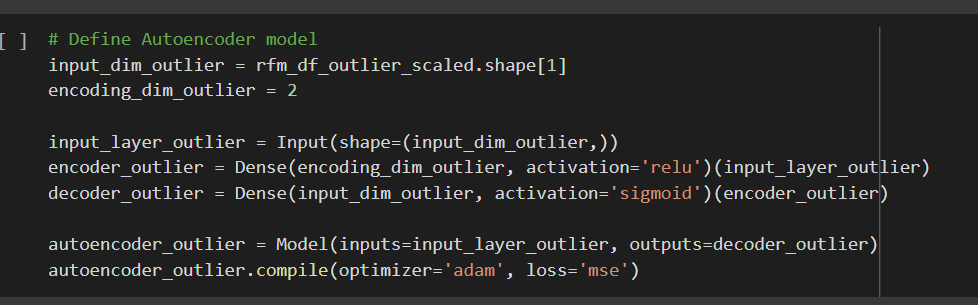


Table 13. Trường hợp 1 - Kết quả phân cụm bằng HAC được thêm vào bảng RFM gốc

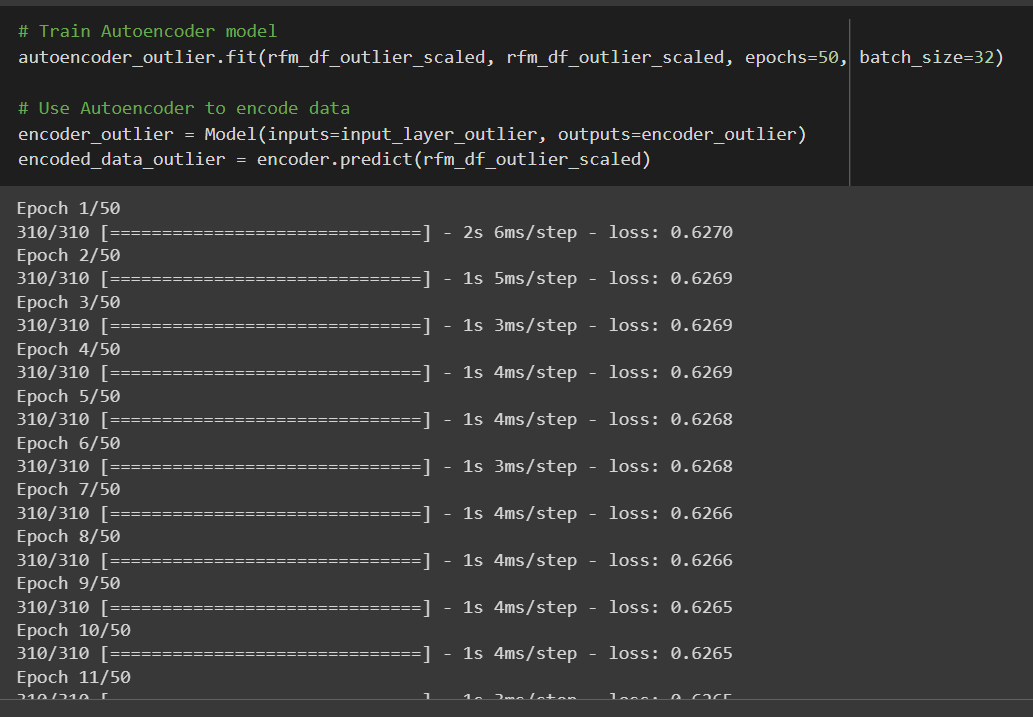
**Mô hình autoencoder**

Mô hình Autoencoder là một mô hình học không giám sát được sử dụng để giảm chiều dữ liệu. Nó bao gồm hai phần chính: một mạng lưới encoder để mã hóa dữ liệu và một mạng lưới decoder để giải mã lại dữ liệu. Mục đích của mô hình này là học cách giảm chiều dữ liệu sao cho dữ liệu giải mã có thể được tái tạo với độ chính xác cao.

  
Xây dựng mô hình autoencoder

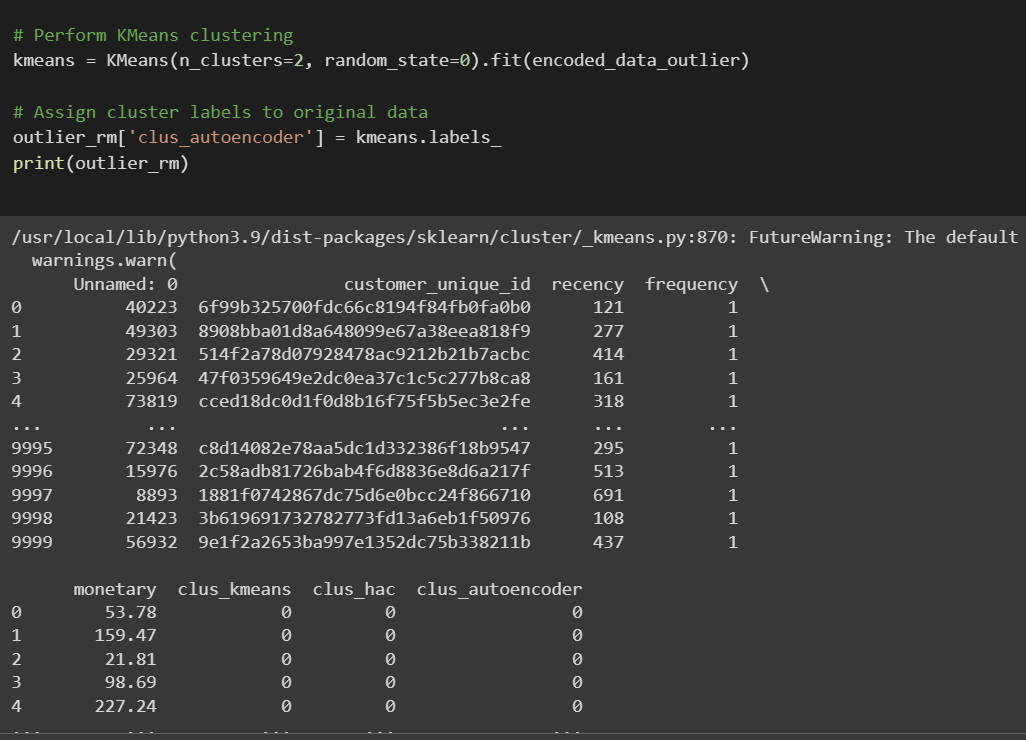
Chúng ta sẽ xây dựng mô hình Autoencoder gồm 3 lớp: lớp input, lớp ẩn và lớp output. Trong đoạn code trên, kích thước đầu vào của mô hình Autoencoder được định nghĩa bằng cách lấy số cột của tập dữ liệu được chuẩn hóa (rfm\_df\_outlier\_scaled). Kích thước mã hóa được đặt là 2, có nghĩa là mô hình sẽ học cách giảm chiều dữ liệu từ kích thước đầu vào xuống 2 chiều.

Lớp đầu vào (input\_layer\_outlier) được định nghĩa là một đối tượng Input với kích thước tương ứng với kích thước đầu vào của mô hình. Lớp mã hóa (encoder\_outlier) là một lớp Dense với kích thước bằng kích thước mã hóa và được kết nối với lớp đầu vào. Activation function được sử dụng cho lớp này là relu. Lớp giải mã (decoder\_outlier) cũng là một lớp Dense, nhưng với kích thước bằng kích thước đầu vào và được kết nối với lớp mã hóa. Activation function được sử dụng cho lớp này là sigmoid. Mô hình Autoencoder được biên dịch với hàm tối ưu adam và hàm mất mát MSE.



Huấn luyện mô hình autoencoder

Sau khi mô hình Autoencoder được định nghĩa và biên dịch, nó được huấn luyện trên tập dữ liệu được chuẩn hóa (rfm\_df\_outlier\_scaled). Mô hình được huấn luyện qua 50 epochs và với batch size là 32.

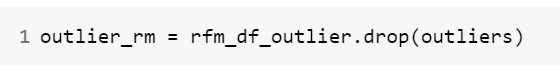


Sau khi mô hình Autoencoder được huấn luyện, nó được sử dụng để mã hóa dữ liệu ban đầu (rfm\_df\_outlier\_scaled). Kết quả mã hóa được sử dụng để thực hiện phân cụm bằng cách sử dụng thuật toán KMeans với số cụm là 4.

### 

### b.Trường hợp 2: loại bỏ các phần tử ngoại biên (outliers)

Nhóm tạo 1 bản sao bảng RFM và loại bỏ các phần tử ngoại biên (outliers) bằng hàm .drop() với tham số là danh sách các chỉ số của các phần tử cần được loại bỏ.



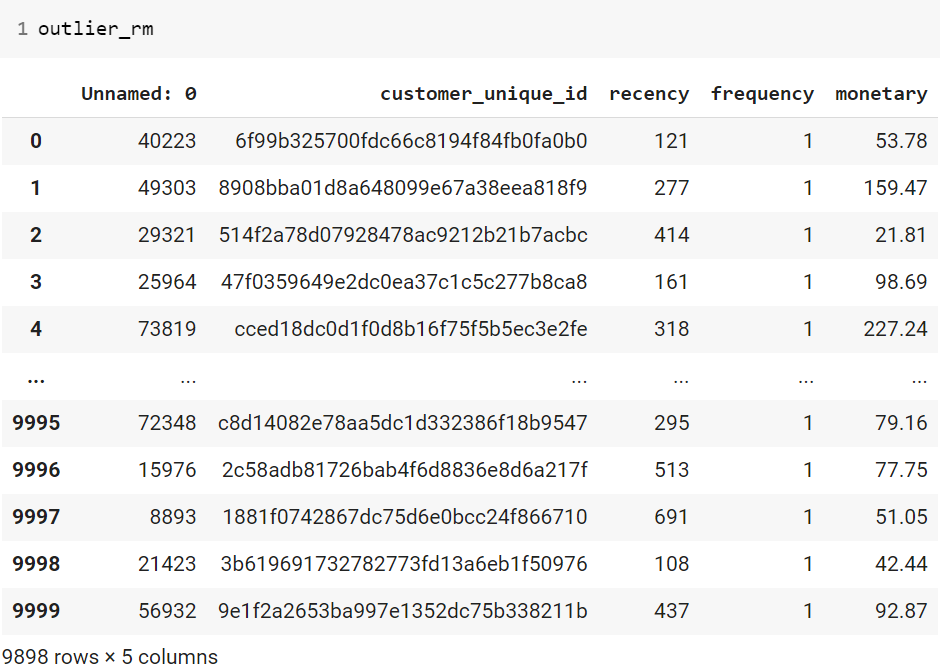
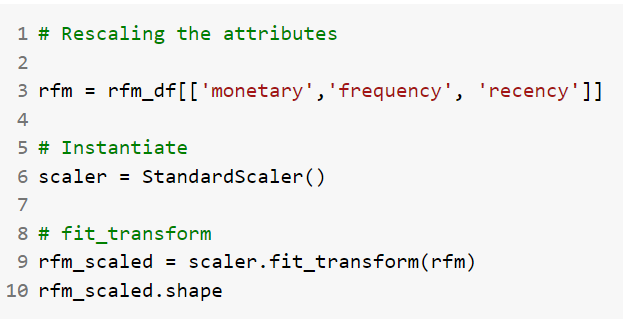
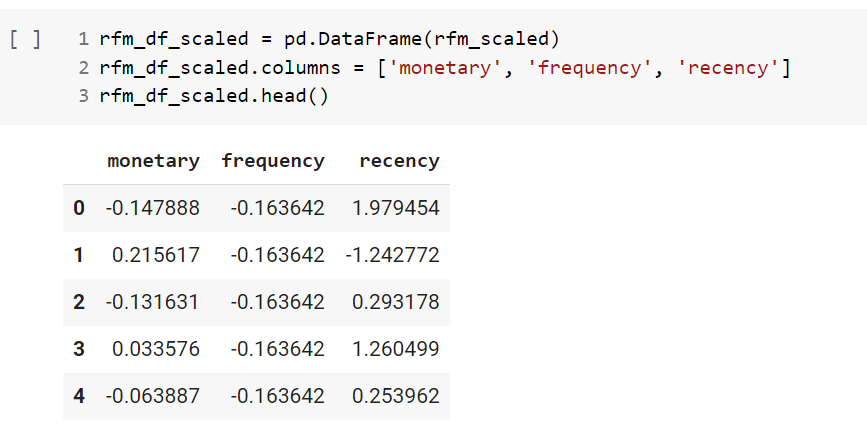


Table 14. Trường hợp 2 - Bảng dữ liệu RFM sau khi đã loại các phần tử ngoại biên - outliers

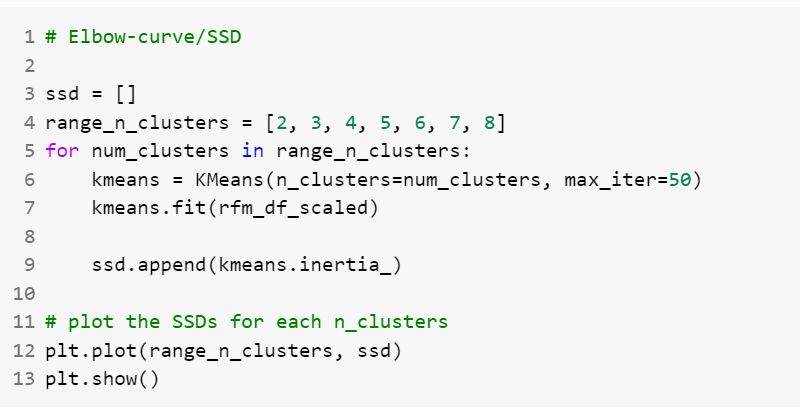
Tiếp theo, nhóm chuẩn hóa dữ liệu bằng hàm .StandardScaler() cho 3 cột Recency, Frequency, Monetary.

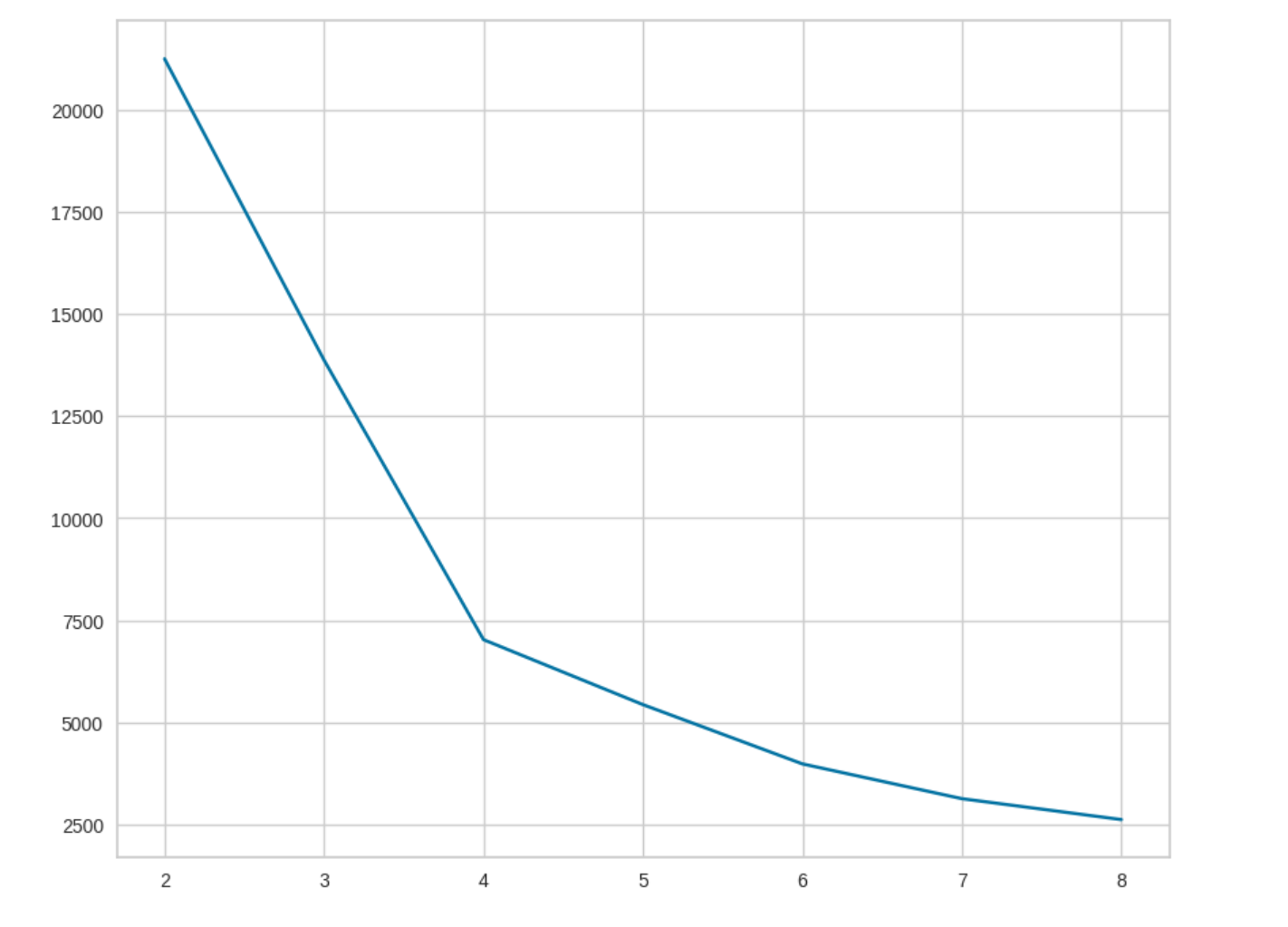


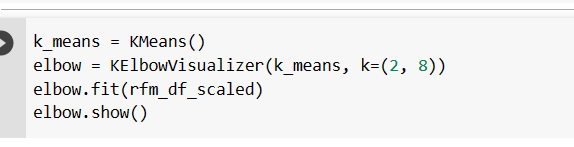


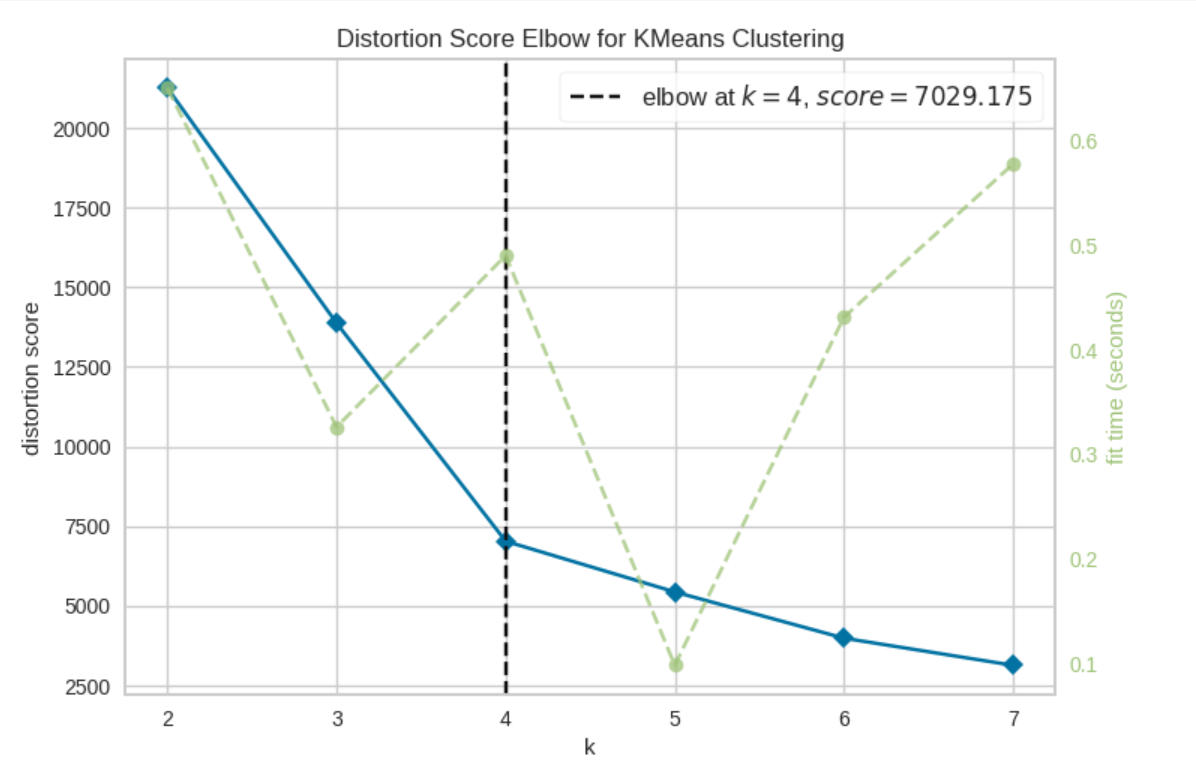
Một điều quan trọng trước khi tiến hành áp dụng các thuật toán phân cụm lên bộ dữ liệu đó là phải xác định được số cụm cần chia. Ở đây, nhóm kết hợp 2 phương pháp tìm số cụm k, đó là phương pháp khuỷu tay - Elbow Curve, và chỉ số Silhouette.

Với phương pháp khuỷu tay Elbow Curve, nhóm tạo 1 danh sách số lượng các cụm, từ 2 đến 8 cụm. Nhóm khởi tạo vòng lặp for: áp dụng thuật toán KMeans lên bộ dữ liệu đã chuẩn hóa, với mỗi số cụm k trong danh sách số lượng cụm đã tạo ở phía trên, sau đó ghi nhận chỉ số inertia, hay còn gọi là SSD (sum of squared distances) - tổng của căn bậc 2 khoảng cách giữa các phần tử đến điểm trung tâm của từng cụm mà phần tử đó thuộc về. Sau đó, nhóm trực quan hóa các chỉ số SSD tương ứng với từng số cụm qua biểu đồ đường .plot(). Từ biểu đồ này, nhóm có thể xác định số cụm k hợp lý tại điểm khuỷu tay - elbow point.





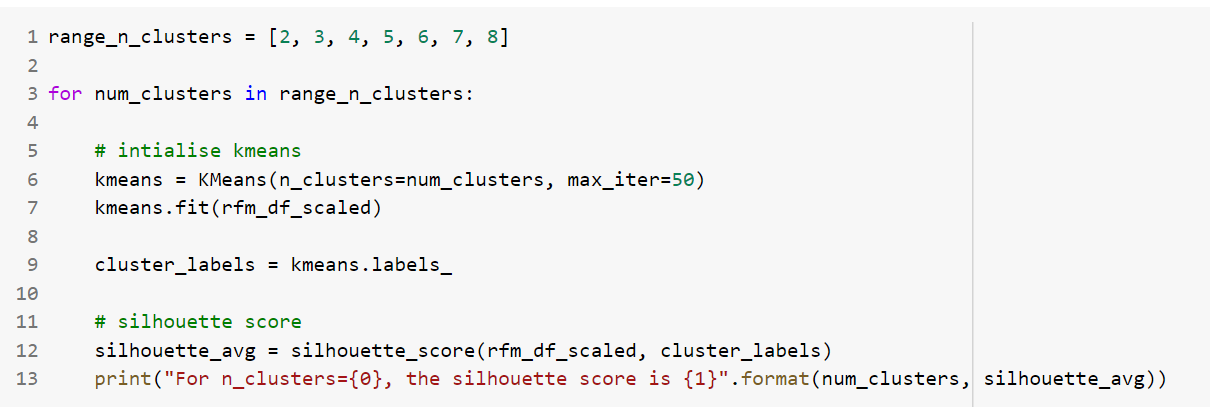


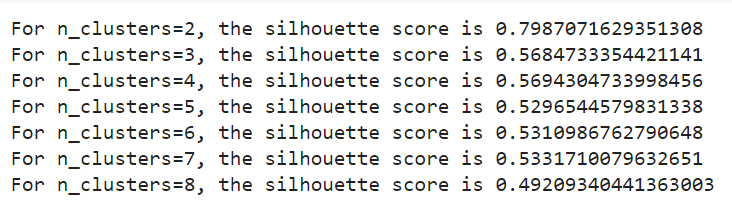


Hình 24. Trường hợp 2 - Distorition score elbow để xác định số lượng cụm tối ưu

Nhìn vào biểu đồ, ta thấy điểm khuỷu tay - Elbow point ở vị trí k=4. Vì thế, số cụm lý tưởng cho việc phân cụm đối với bộ dữ liệu hiện tại là 4.

Để đảm bảo số cụm k nhóm chọn là hợp lý, nhóm sẽ xem xét thêm chỉ số Silhouette. Với mỗi số cụm cụ thể, sau khi đã thực hiện phân cụm trên bộ dữ liệu đã được chuẩn hóa, nhóm sử dụng hàm silhouette\_scores() tronng thư viện sklearn. Các số cụm có chỉ số Silhouette trên 0,5 được đánh giá là phân cụm tốt.





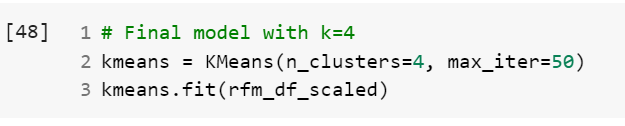
Hình 25.Trường hợp 2 - Kết quả tính Silhouette Score của các số lượng cụm

Tại số cụm k=4, chỉ số Silhouette là 0,569, một chỉ số Silhouette thể hiện việc phân cụm tốt. Do vậy, nhóm quyết định chọn số cụm k=4.

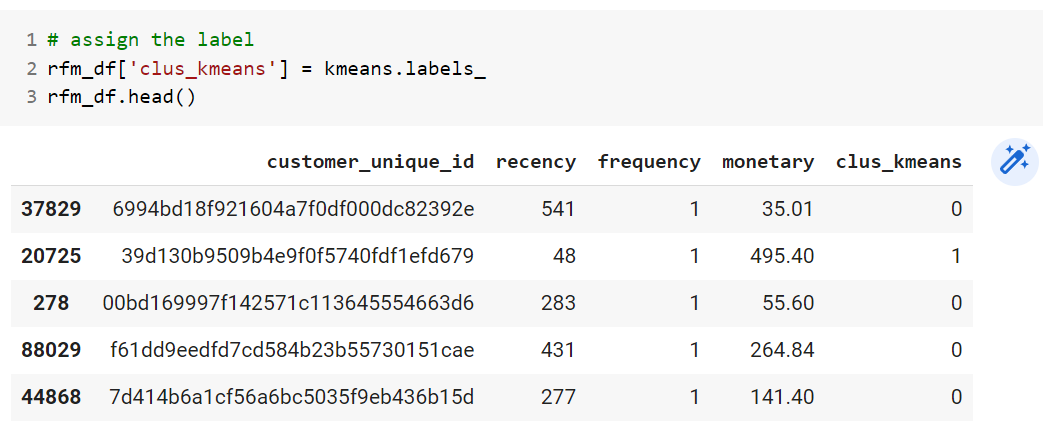
Sau khi đã xác định được số cụm, nhóm tiến hành phân cụm dữ liệu bằng ba thuật toán: KMeans, HAC, học sâu AutoEncoder.

**KMeans:**

Nhóm sử dụng hàm .Kmeans() trong thư viện sklearn.cluster với số cụm n\_clusters=4, số vòng lặp tối đa max\_iter=50. Kết quả phân cụm sẽ được thêm thành 1 cột mới tên là ‘clus\_kmeans’ trong bảng dữ liệu RFM.



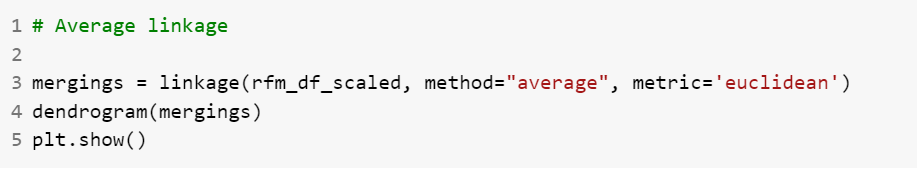
Áp dụng hàm .KMeans() lên bảng dữ liệu RFM đã chuẩn hóa với k=4)

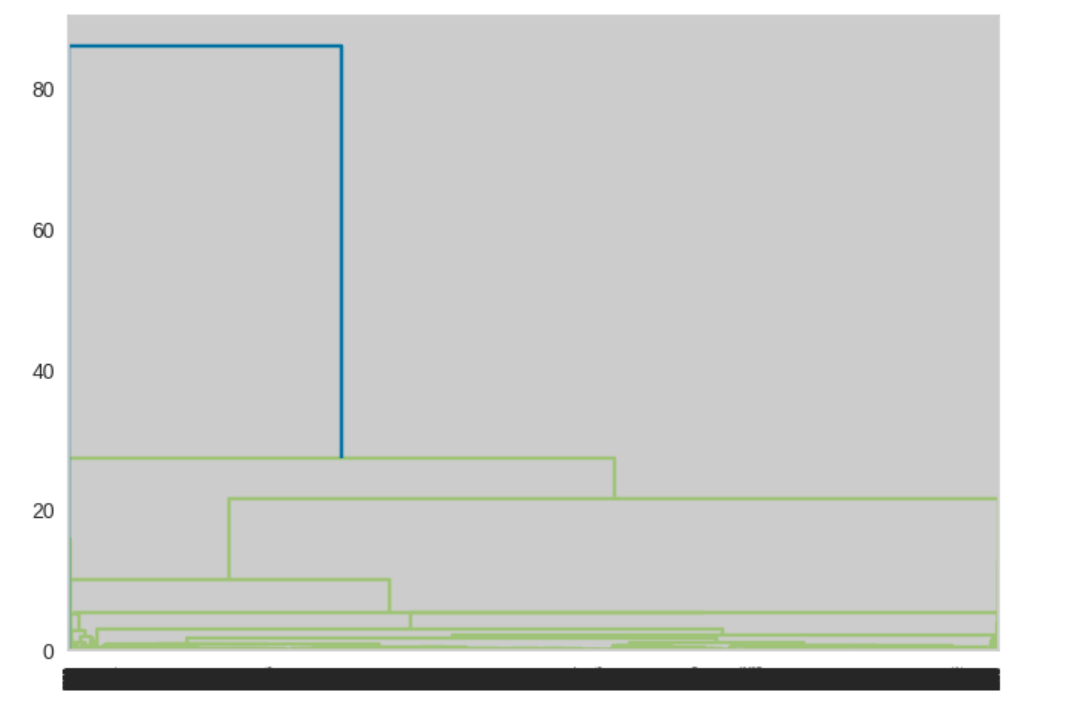


Hình 26. Trường hợp 2 - Kết quả phân cụm bằng KMeans được thêm vào bảng RFM gốc

**HAC:**

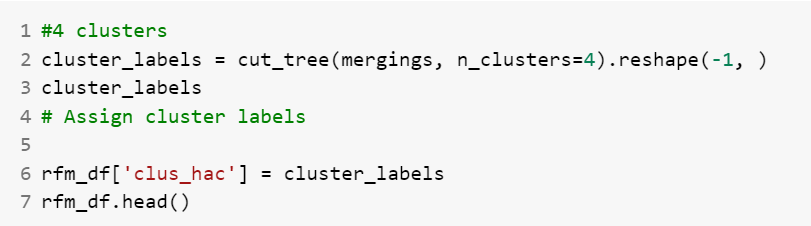
Ở phương pháp phân cụm HAC, nhóm tính khoảng cách giữa các cụm bằng hàm linkage() trong scipy.cluster.hierarchy, trong đó tham số method="average", metric='euclidean'. Từ dữ liệu tính toán khoảng cách vừa thực hiện, nhóm xây dựng dendrogram, trong đó các điểm dữ liệu và các cụm được đặt ở các nút của cây.

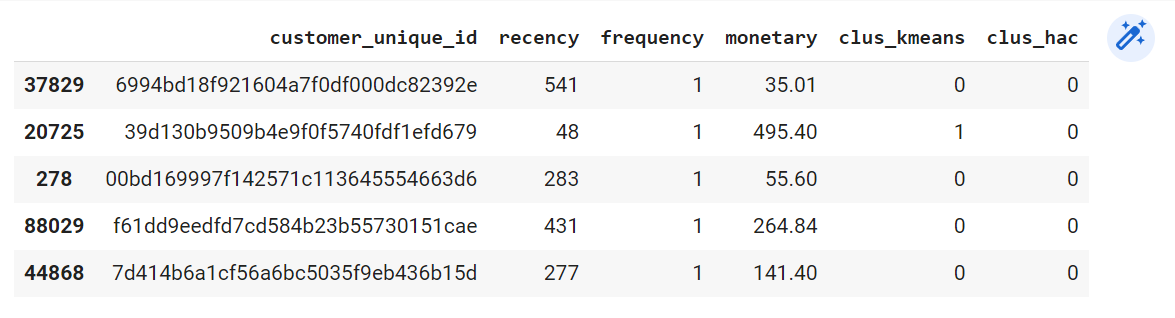




Hình 27. Trường hợp 2 - 1 phần của dendrogram

Sau đó, nhóm cắt cây bằng hàm cut\_tree() trong thư viện scipy.cluster.hierarchy, với số cụm bằng 4. Kết quả phân cụm sẽ được thêm thành 1 cột mới tên là ‘clus\_hac’ trong bảng dữ liệu RFM.

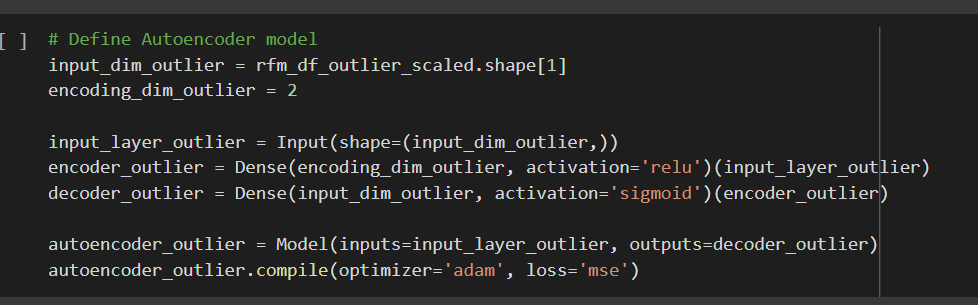




Hình 28. Trường hợp 2 - Kết quả phân cụm bằng HAC được thêm vào bảng RFM gốc

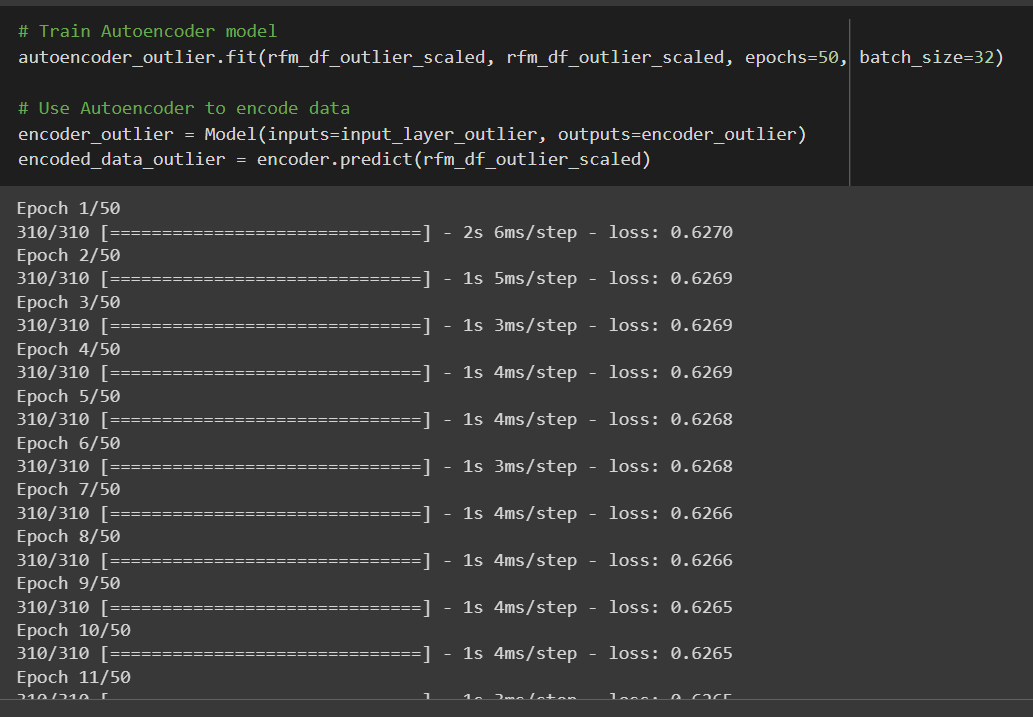
**Mô hình autoencoder**

Mô hình Autoencoder là một mô hình học không giám sát được sử dụng để giảm chiều dữ liệu. Nó bao gồm hai phần chính: một mạng lưới encoder để mã hóa dữ liệu và một mạng lưới decoder để giải mã lại dữ liệu. Mục đích của mô hình này là học cách giảm chiều dữ liệu sao cho dữ liệu giải mã có thể được tái tạo với độ chính xác cao.

  
Xây dựng mô hình autoencoder

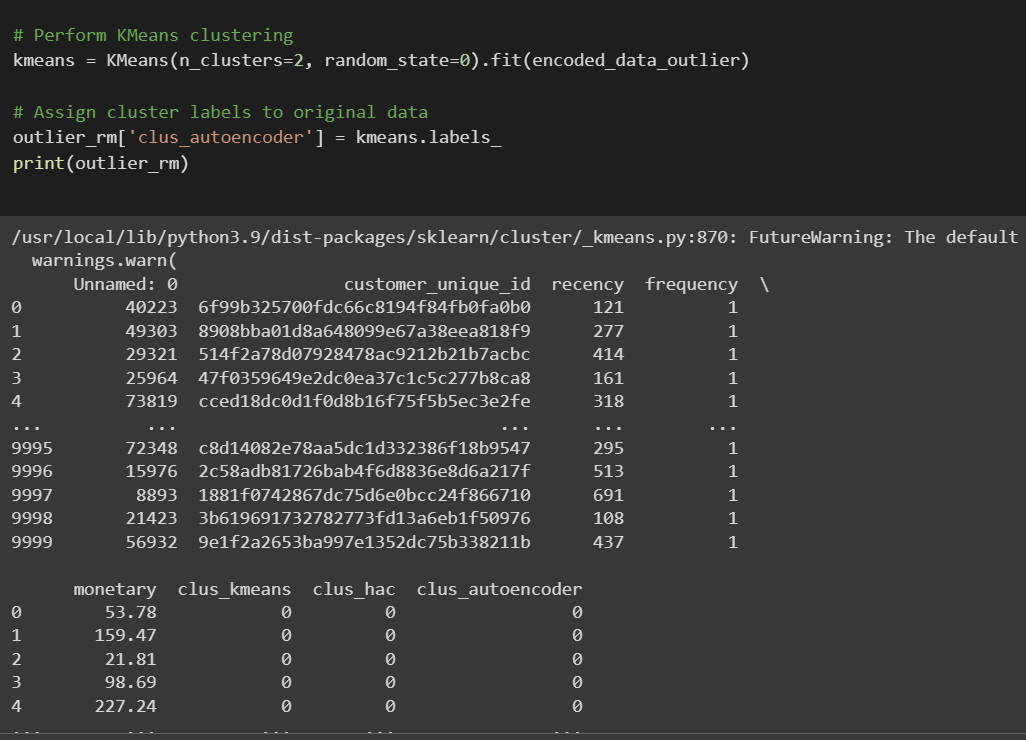
Chúng ta sẽ xây dựng mô hình Autoencoder gồm 3 lớp: lớp input, lớp ẩn và lớp output. Trong đoạn code trên, kích thước đầu vào của mô hình Autoencoder được định nghĩa bằng cách lấy số cột của tập dữ liệu được chuẩn hóa (rfm\_df\_outlier\_scaled). Kích thước mã hóa được đặt là 2, có nghĩa là mô hình sẽ học cách giảm chiều dữ liệu từ kích thước đầu vào xuống 2 chiều.

Lớp đầu vào (input\_layer\_outlier) được định nghĩa là một đối tượng Input với kích thước tương ứng với kích thước đầu vào của mô hình. Lớp mã hóa (encoder\_outlier) là một lớp Dense với kích thước bằng kích thước mã hóa và được kết nối với lớp đầu vào. Activation function được sử dụng cho lớp này là relu. Lớp giải mã (decoder\_outlier) cũng là một lớp Dense, nhưng với kích thước bằng kích thước đầu vào và được kết nối với lớp mã hóa. Activation function được sử dụng cho lớp này là sigmoid. Mô hình Autoencoder được biên dịch với hàm tối ưu adam và hàm mất mát MSE.



Huấn luyện mô hình autoencoder

Sau khi mô hình Autoencoder được định nghĩa và biên dịch, nó được huấn luyện trên tập dữ liệu được chuẩn hóa (rfm\_df\_outlier\_scaled). Mô hình được huấn luyện qua 50 epochs và với batch size là 32.

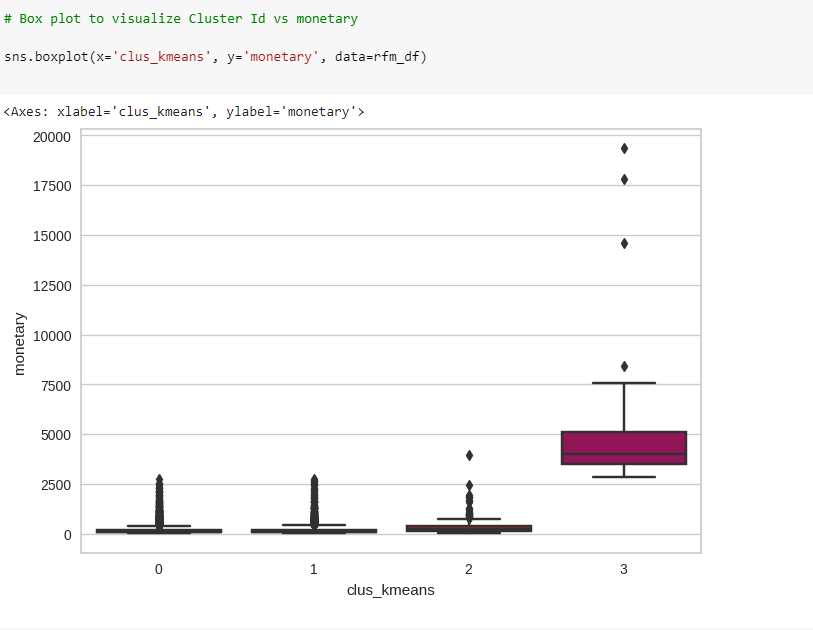


Hình 29. Kết quả phân cụm bằng mô hình AutoEncoder

Sau khi mô hình Autoencoder được huấn luyện, nó được sử dụng để mã hóa dữ liệu ban đầu (rfm\_df\_outlier\_scaled). Kết quả mã hóa được sử dụng để thực hiện phân cụm bằng cách sử dụng thuật toán KMeans với số cụm là 4.

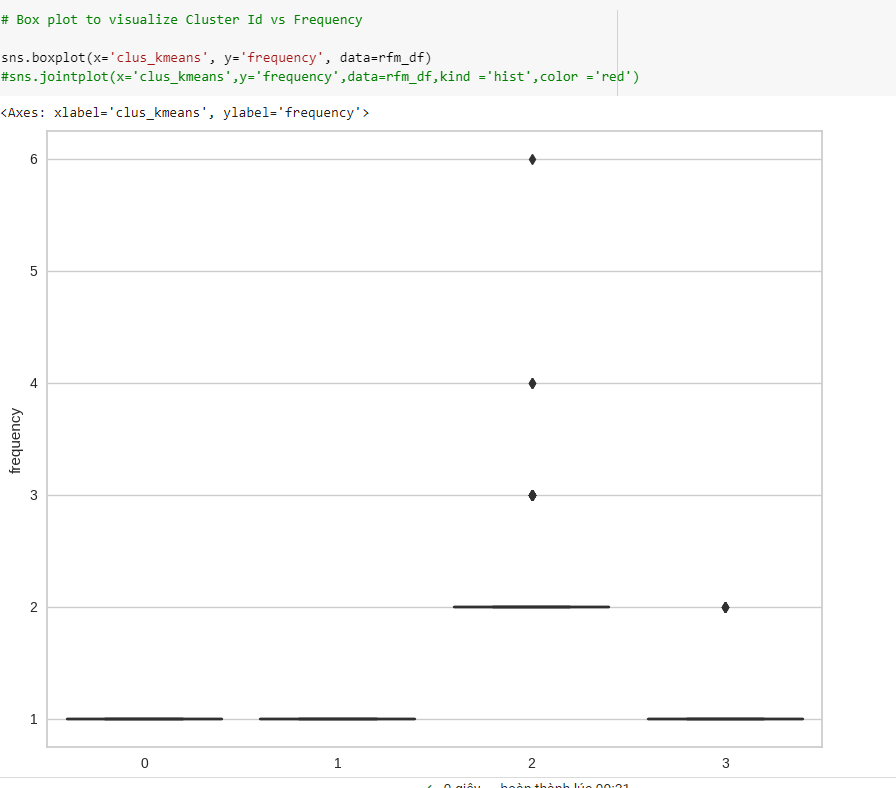
## 3.6. Mô tả đặc điểm các cụm

**Phân cụm K-Means**



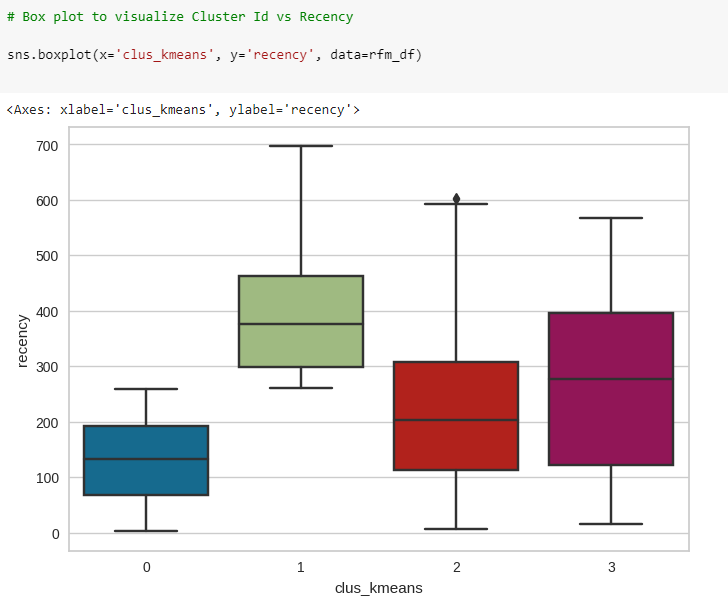
Hình 30. Biểu đồ plot thể hiện tổng giá trị đơn hàng của các cụm (KMeans)

Ở cụm 3, tổng số lượng tiền trung bình mà khách hàng đã bỏ ra để mua hàng (tổng giá trị đơn hàng) cao nhất so với các các cụm còn lại. Các cụm còn lại có tổng giá trị đơn hàng trung bình thấp hơn (nằm trong khoảng từ 0-3000$ bao gồm cả những giá trị đơn hàng thuộc ngoại lai). Thông thường, một người đã chi tiêu nhiều tiền hơn, thì người đó cũng sẽ chi nhiều hơn trong những lần mua hàng hóa/dịch vụ tiếp theo. Vì vậy trong trường hợp này ta có thể thấy được các khách hàng ở cụm 3 sẽ có xu hướng chi tiêu nhiều hơn so với khách hàng các cụm còn lại cho những lần mua hàng tiếp theo.



Hình 31. Biểu đồ thể hiện tần suất mua hàng của các cụm (KMeans)

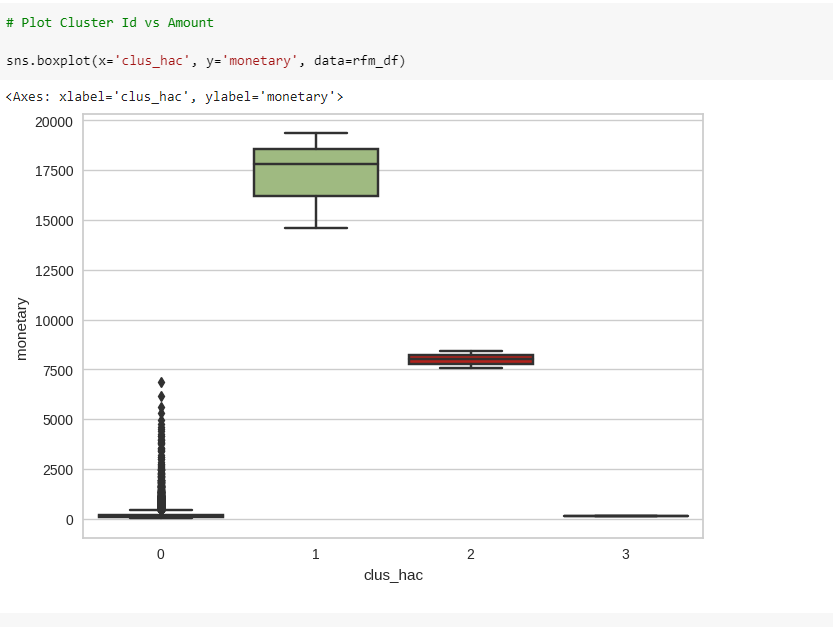
Các khách hàng ở cụm 2 có xu hướng thường xuyên mua hàng hơn so với các cụm còn lại, trung bình các khách hàng nằm trong cụm 0,1,3 chỉ đến mua hàng 1 lần, có khách 2 lần. Những khách hàng thường xuyên mua hàng sẽ là những đối tượng có nhiều khả năng nhất trong việc tiếp tục đến mua. Việc họ đến mua lần thứ 2,3,4,5 cho thấy, có thể họ thỏa mãn với hàng hóa của cửa hàng.



Hình 32. Biểu đồ thể hiện khả năng mua hàng gần đây của khách hàng các cụm (KMeans)

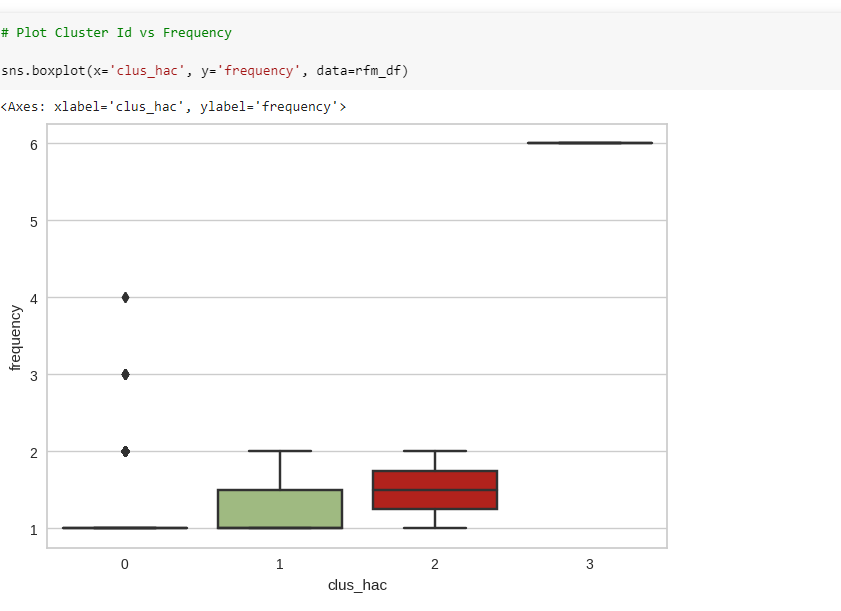
Các khách hàng nằm ở cụm 0 là những khách hàng đã mua hàng trong thời gian gần đây và các khách hàng ở cụm 1 là những khách hàng đã mua hàng 1 thời gian đã lâu, tính từ thời điểm lần cuối mua hàng cho đến thời điểm hiện tại. Từ kết quả này ta có thể suy ra được khách hàng càng mới mua sản phẩm thì càng có khả năng tiếp tục mua hàng, vì việc mua bán mới diễn ra nên khách hàng ít có khả năng tiếp xúc với các đối thủ cạnh tranh, trong khi nhu cầu của họ vẫn còn.

**Phân cụm phân cấp (Hierarchical clustering)**



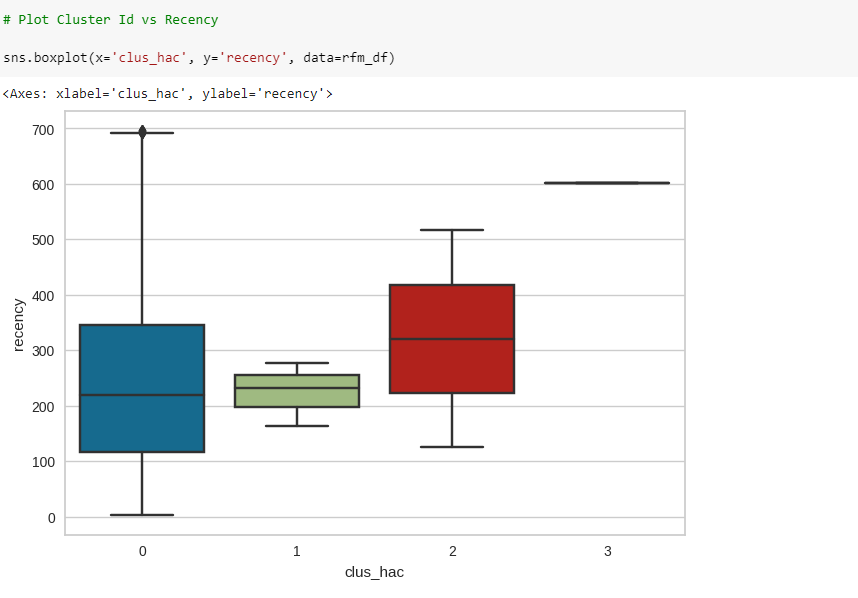
Hình 33. Biểu đồ plot thể hiện tổng giá trị đơn hàng của các cụm (HAC)

Tổng số lượng tiền trung bình mà khách hàng ở cụm 1 đã bỏ ra để mua hàng (tổng giá trị đơn hàng) cao hơn hẳn so với các các cụm còn lại. Các cụm còn lại có tổng giá trị đơn hàng trung bình thấp hơn. Thông thường, một người đã chi tiêu nhiều tiền hơn, thì người đó cũng sẽ chi nhiều hơn trong những lần mua hàng hóa/dịch vụ tiếp theo. Vì vậy trong trường hợp này ta có thể thấy được các khách hàng ở cụm 1 sẽ có xu hướng chi tiêu nhiều hơn so với khách hàng các cụm còn lại cho những lần mua hàng tiếp theo.



Hình 34. Biểu đồ thể hiện tần suất mua hàng của các cụm (HAC)

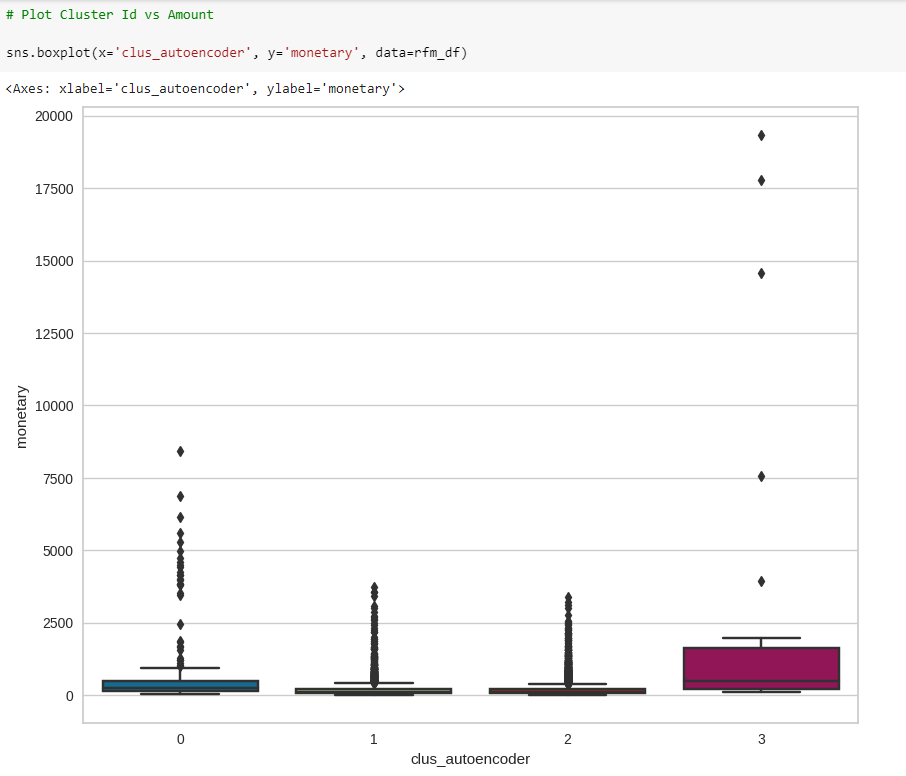
Các khách hàng ở cụm 3 có xu hướng thường xuyên mua hàng hơn so với các cụm còn lại, trung bình các khách hàng nằm trong cụm 0,1,2 chỉ đến mua hàng 1 lần, có khách 2 lần. Những khách hàng thường xuyên mua hàng sẽ là những đối tượng có nhiều khả năng nhất trong việc tiếp tục đến mua. Việc họ đến mua lần thứ 2 hoặc nhiều hơn cho thấy, có thể họ thỏa mãn với hàng hóa của cửa hàng.



Hình 35. Biểu đồ thể hiện khả năng mua hàng gần đây của khách hàng các cụm (HAC)

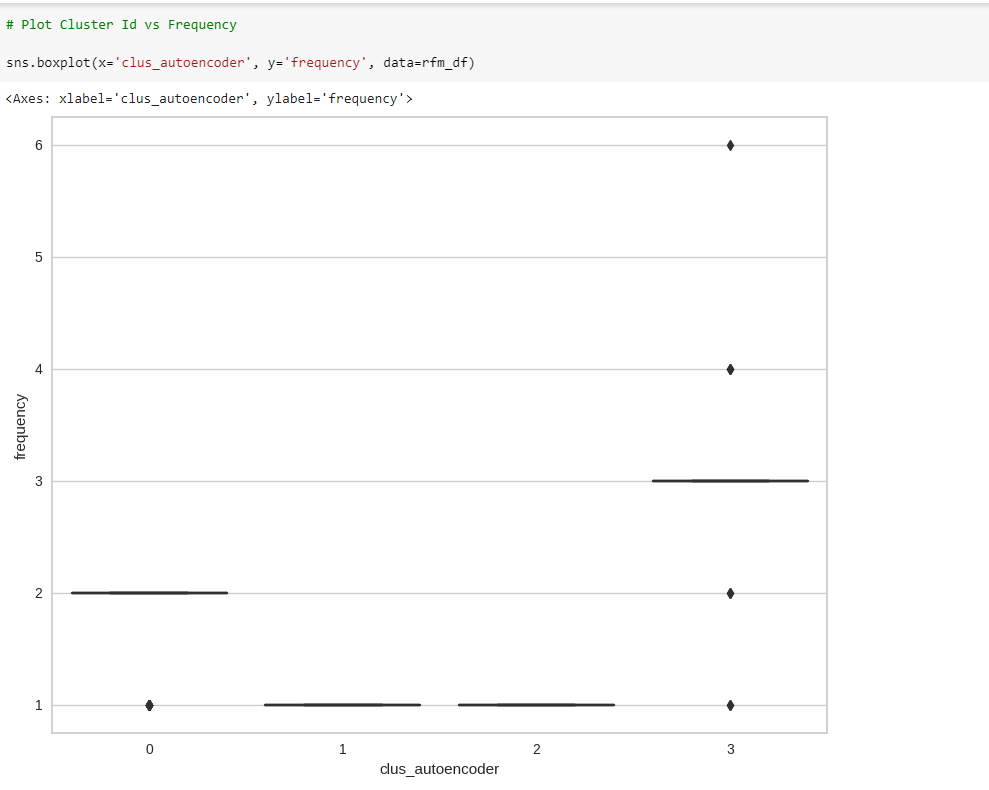
Các khách hàng nằm ở cụm 0,1 là những khách hàng đã mua hàng trong thời gian gần đây và các khách hàng ở cụm 3 là những khách hàng đã mua hàng 1 thời gian đã lâu, tính từ thời điểm lần cuối mua hàng cho đến thời điểm hiện tại. Từ kết quả này ta có thể suy ra được khách hàng càng mới mua sản phẩm thì càng có khả năng tiếp tục mua hàng, vì việc mua bán mới diễn ra nên khách hàng ít có khả năng tiếp xúc với các đối thủ cạnh tranh, trong khi nhu cầu của họ vẫn còn.

**Mô hình AutoEncoder**

****

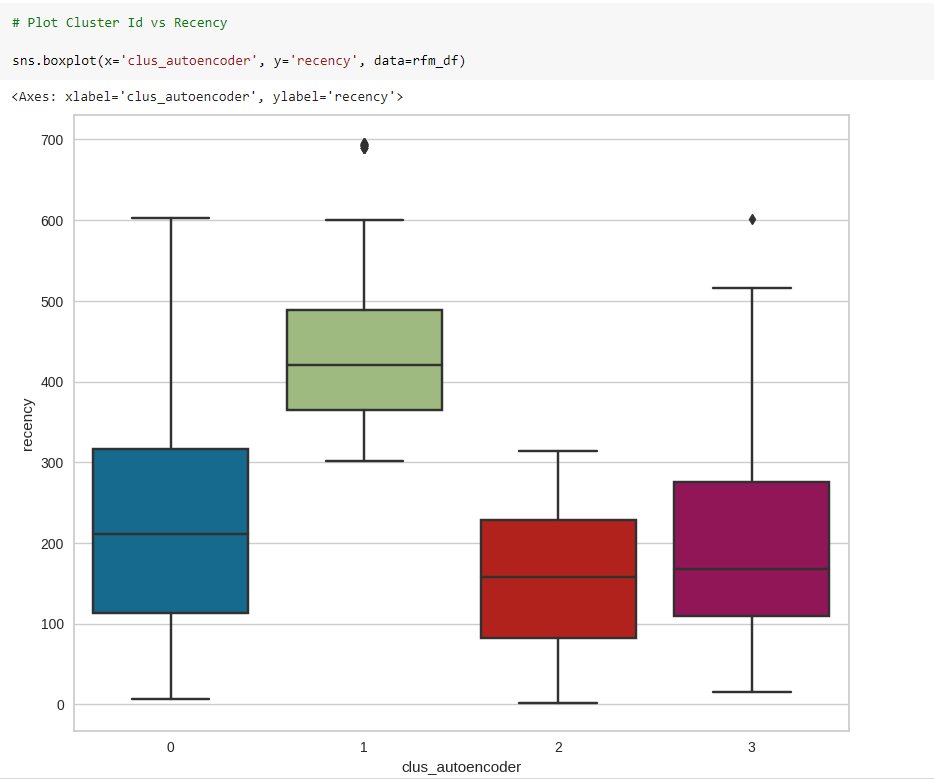
Hình 36. Biểu đồ plot thể hiện tổng giá trị đơn hàng của các cụm (AutoEncoder)

Tổng số lượng tiền trung bình mà khách hàng ở cụm 3 đã bỏ ra để mua hàng (tổng giá trị đơn hàng) cao hơn hẳn so với các các cụm còn lại. Các cụm còn lại có tổng giá trị đơn hàng trung bình thấp hơn. Tuy nhiên trong các cụm đều có nhiều giá trị ngoại lai, cho thấy có một số khách hàng chi tiêu cao hơn so với mức trung bình chung các khách hàng còn lại trong cụm. Thông thường, một người đã chi tiêu nhiều tiền hơn, thì người đó cũng sẽ chi nhiều hơn trong những lần mua hàng hóa/dịch vụ tiếp theo. Vì vậy trong trường hợp này ta có thể thấy được các khách hàng ở cụm 1 sẽ có xu hướng chi tiêu nhiều hơn so với khách hàng các cụm còn lại cho những lần mua hàng tiếp theo.



Hình 37. Biểu đồ plot thể hiện xu hướng tần số mua sắm trực tuyến của các cụm (AutoEncoder)

Các khách hàng ở cụm 3 có xu hướng thường xuyên mua hàng hơn so với các cụm còn lại trung bình tầm 3 lần có trường hợp ngoại lai gồm 4 lần, 2,1 lần. Các khách hàng nằm trong cụm 0,1,2 hầu hết chỉ đến mua hàng 1 lần, có khách 2 lần. Những khách hàng thường xuyên mua hàng sẽ là những đối tượng có nhiều khả năng nhất trong việc tiếp tục đến mua. Việc họ đến mua lần thứ 2 hoặc nhiều hơn cho thấy, có thể họ thỏa mãn với hàng hóa của cửa hàng.



Hình 38. Biểu đồ plot thể hiện xu hướng thời gian từ lần mua hàng cuối của các cụm (AutoEncoder)

Các khách hàng nằm ở cụm 2 là những khách hàng đã mua hàng trong thời gian gần đây và các khách hàng ở cụm 1 là những khách hàng đã mua hàng 1 thời gian đã lâu, tính từ thời điểm lần cuối mua hàng cho đến thời điểm hiện tại.

Từ kết quả này ta có thể suy ra được khách hàng càng mới mua sản phẩm thì càng có khả năng tiếp tục mua hàng, vì việc mua bán mới diễn ra nên khách hàng ít có khả năng tiếp xúc với các đối thủ cạnh tranh, trong khi nhu cầu của họ vẫn còn.

# CHƯƠNG 4. KẾT LUẬN

## 4.1. Các kết quả đạt được

Thông qua phân tích RFM và xây dựng mô hình phân cụm dữ liệu, nhóm chúng tôi đã phân loại khách hàng thành các nhóm dựa trên các giá trị RFM của họ. Trong bài này chúng tôi đã phân được 4 nhóm khách hàng với đặc điểm về Recency (Tần suất mua hàng gần đây), Frequency (Tần suất mua hàng) và Monetary (Giá trị đơn hàng trung bình) như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Thuật toán** | **Cụm 0** | **Cụm 1** | **Cụm 2** | **Cụm 3** |
| **K Means** | - Khách hàng đã mua hàng trong thời gian gần đây cao |  | - Xu hướng thường xuyên mua hàng hơn | - Tổng số lượng tiền trung bình mà khách hàng đã bỏ ra để mua hàng cao nhất so với các các cụm còn lại |
| **HAC** | - Khách hàng đã mua hàng trong thời gian gần đây cao | - Tổng giá trị đơn hàng cao hơn hẳn so với các các cụm còn lại.  -Khách hàng đã mua hàng trong thời gian gần đây cao |  | Xu hướng thường xuyên mua hàng hơn so với các cụm còn lại |
| **Auto encoder** |  |  | - Khách hàng đã mua hàng trong thời gian gần đây cao | - Tổng giá trị đơn hàng cao hơn hẳn so với các các cụm còn lại  - Xu hướng thường xuyên mua hàng hơn so với các cụm còn lại |

Bằng cách phân tích từng cụm khách hàng theo các thuật toán ở trên sẽ giúp cửa hàng có thể xây dựng chiến lược marketing phù hợp với từng nhóm khách hàng để tăng doanh số bán hàng và lợi nhuận, xác định những khách hàng có giá trị cao nhất và tập trung vào việc giữ chân và phát triển mối quan hệ với những khách hàng này.

* Đối với nhóm khách hàng thường xuyên mua hàng, hoặc mua với tổng số lương đơn hàng cao chúng tôi đề xuất cửa hàng cung cấp các ưu đãi đặc biệt để khuyến khích khách hàng tiếp tục sử dụng và tăng giá trị của đơn hàng.
* Riêng đối với nhóm khách hàng đã lâu không quay trở lại cửa hàng và có xu hướng rời bỏ nên tìm hiểu nguyên nhân và đưa ra các biện pháp nhằm để khuyến khích họ quay trở lại sử dụng sản phẩm hoặc dịch vụ của mình thì có thể tặng họ các voucher mua sắm, nhưng giảm giá trị của từng voucher lại, vì nếu không mua hàng lâu mà được tặng voucher giá trị bằng cả một đơn hàng thì sẽ giảm hứng thú của khách hàng đối với hệ thống (đợi đến khi có voucher mới mua hàng).

## 4.2. Những Hạn Chế và Hướng Phát Triển

### 4.2.1. Những hạn chế

* Chúng tôi lựa chọn thuật toán phân cụm để phân khúc khách hàng dựa trên phương pháp RFM là một trong những phương pháp khá phổ biến. Tuy nhiên giống như nhiều phương pháp phân cụm khác, phương pháp này vẫn tồn tại một vài những khuyết điểm.
* -Thiếu bối cảnh: khi phân cụm dựa vào RFM, chỉ tập trung vào Recency, Frequency và Monetary. Các yếu tố khác như sự trung thành, độ hài lòng của khách hàng, lịch sử mua hàng và đặc điểm cá nhân của khách hàng không được tính toán. Khi áp dụng phương pháp phân cụm trên RFM, việc lựa chọn số lượng cụm phù hợp có thể ảnh hưởng đến kết quả phân nhóm.
* Với K-means và HAC, việc xác định số lượng cụm (clusters) phù hợp có thể khó khăn và ảnh hưởng đến kết quả phân nhóm. Ngoài ra, hai phương pháp này không thể xử lý được dữ liệu nhiễu hoặc các đối tượng dữ liệu không thuộc vào bất kỳ cụm nào.
* Với Autoencoder, việc lựa chọn số lượng chiều của vector low-dimensional có thể ảnh hưởng đến kết quả phân nhóm. Ngoài ra, mô hình này có thể bị overfitting nếu không được huấn luyện đúng cách.
* Dữ liệu hạn chế: Chất lượng của các cụm phụ thuộc rất nhiều vào số lượng và chất lượng của dữ liệu có sẵn. Vì đây là tập dữ liệu có sẵn nên vẫn có giới hạn về lượng dữ liệu mà chúng tôi có thể khai thác để phân tích ra đúng với kết quả thực tế.

### 4.2.2. Hướng phát triển

* Tăng cường thu thập dữ liệu: Để đạt được kết quả tốt nhất, cần có đủ dữ liệu để phân tích. Việc thu thập dữ liệu chính xác và đầy đủ sẽ giúp tăng độ chính xác của mô hình.
* Tinh chỉnh tham số: Việc tinh chỉnh tham số của mô hình phân cụm RFM có thể giúp cải thiện các chỉ số đánh giá. Ví dụ, tăng hoặc giảm số lượng cụm có thể ảnh hưởng đến kết quả của các chỉ số.
* Sử dụng các thuật toán khác nhau: Có nhiều thuật toán khác nhau để phân cụm RFM, và mỗi thuật toán có những ưu điểm và hạn chế riêng. Việc sử dụng các thuật toán khác nhau có thể giúp tìm ra mô hình phù hợp nhất với dữ liệu.
* Tích hợp dữ liệu từ các nguồn khác nhau và sử dụng các phương pháp khai thác dữ liệu mới có thể giúp cải thiện độ chính xác và hiệu quả của mô hình.
* Áp dụng mô hình vào thực tế: Cuối cùng, sau khi đạt được kết quả tốt nhất, mô hình phân cụm RFM có thể được áp dụng vào thực tế để giúp các doanh nghiệp phân loại khách hàng và tối ưu hóa chiến lược kinh doanh của mình

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | A. Sionek, "Brazilian E-Commerce Public Dataset by Olist," 2018. [Online]. Available: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. |
| [2] | T. D. V. Minh, "MinhThanh2404/Customer-Segmentation-with-RFM-Method," GitHub, [Online]. Available: https://github.com/MinhThanh2404/Customer-Segmentation-with-RFM-Method. |