ĐẠI HỌC KINH TẾ TP. HỒ CHÍ MINH

BÁO CÁO ĐỒ ÁN HỌC PHẦN
KHOA HỌC DỮ LIỆU

Đề tài

PHÂN CỤM KHÁCH HÀNG TỪ DỮ LIỆU CỬA HÀNG BÁN LẺ SỬ DỤNG MÔ HÌNH K-MEAN

Giảng viên hướng dẫn: TS.GVC Nguyễn Quốc Hùng

Nhóm thực hiện: 05
Thái Hoài An (Nhóm trưởng)

TP. Hồ Chí Minh, tháng 5 năm 2026

LỜI MỞ ĐẦU

Trước hết, nhóm xin chân thành cảm ơn thầy Nguyễn Quốc Hùng đã hướng dẫn, góp ý và hỗ trợ nhóm trong suốt quá trình thực hiện đồ án học phần. Những định hướng của thầy đã giúp nhóm từng bước hoàn thiện workflow xử lý dữ liệu, cách trình bày báo cáo và định hướng diễn giải kết quả theo hướng rõ ràng, có ý nghĩa thực tiễn hơn.

Trong bối cảnh doanh nghiệp ngày càng sở hữu lượng dữ liệu giao dịch lớn, bài toán khai thác dữ liệu để hiểu khách hàng trở thành một nhu cầu có ý nghĩa thực tiễn rõ rệt. Thay vì áp dụng một chiến lược tiếp thị chung cho toàn bộ khách hàng, doanh nghiệp cần nhận diện các nhóm khách hàng có hành vi tương đồng để xây dựng chương trình chăm sóc, giữ chân và bán hàng phù hợp hơn. Đó chính là lý do bài toán phân cụm khách hàng ngày càng được quan tâm trong thực tế kinh doanh.

Báo cáo này trình bày quá trình thực hiện đề tài *Phân cụm khách hàng từ dữ liệu cửa hàng bán lẻ sử dụng mô hình K-Mean*. Nhóm sử dụng bộ dữ liệu giao dịch *Online Retail* để thực hiện các bước tiền xử lý, làm sạch dữ liệu, phân tích khám phá, xây dựng đặc trưng ở cấp khách hàng và áp dụng mô hình phân cụm K-Means. Về công cụ triển khai, Excel được sử dụng ở giai đoạn kiểm tra dữ liệu thô và xử lý giá trị thiếu; Orange được sử dụng làm môi trường trực quan để tạo workflow tiền xử lý, tổng hợp feature và mô hình hóa không cần lập trình.

Điểm nhấn của đề tài nằm ở việc chuyển đổi dữ liệu giao dịch thô thành dữ liệu ở cấp khách hàng, từ đó làm nổi bật hành vi mua sắm theo nhiều khía cạnh như khối lượng, giá trị, tần suất, mức độ đa dạng sản phẩm và đặc điểm mỗi giao dịch. Trên cơ sở đó, mô hình K-Means được sử dụng để chia khách hàng thành các cụm có ý nghĩa quản trị, hỗ trợ doanh nghiệp đề xuất chiến lược marketing theo từng phân khúc.

BẢNG PHÂN CÔNG CÁC THÀNH VIÊN

Bảng 1: Phân công công việc trong nhóm

Thành viên	**Nội dung phụ trách**
Thái Hoài An (Nhóm trưởng)	Phụ trách điều phối tiến độ chung, kiểm tra dữ liệu thô, thống kê mô tả, phân tích khám phá dữ liệu, xây dựng workflow Orange, feature engineering, mô hình K-Means, diễn giải cụm khách hàng và hoàn thiện nội dung báo cáo.

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Ý nghĩa
EDA	Exploratory Data Analysis, phân tích khám phá dữ liệu.
RFM	Recency – Frequency – Monetary, bộ ba chỉ số mô tả mức độ gần đây, tần suất và giá trị chi tiêu của khách hàng.
K-Means	Thuật toán phân cụm dựa trên khoảng cách đến tâm cụm.
PCA	Principal Component Analysis, phương pháp phân tích thành phần chính.
UI	User Interface, giao diện người dùng.
UK	United Kingdom.
CSV	Comma-Separated Values, định dạng tập tin dữ liệu dạng văn bản có dấu phân cách.

MỤC LỤC

LỜI MỞ ĐẦU i

BẢNG PHÂN CÔNG CÁC THÀNH VIÊN ii

DANH MỤC TỪ VIẾT TẮT iii

1 GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU VÀ GIỚI THIỆU ĐỀ TÀI 1

1.1 Giới thiệu về khoa học dữ liệu . . . 1

1.1.1 Tổng quan về dữ liệu . . . 1

1.1.2 Tổng quan về khoa học dữ liệu . . . 1

1.2 Giới thiệu đề tài . . . 2

1.2.1 Lý do chọn đề tài . . . 2

1.2.2 Mục tiêu nghiên cứu . . . 2

1.2.3 Câu hỏi nghiên cứu . . . 2

1.2.4 Đối tượng và phạm vi nghiên cứu . . . 3

1.2.5 Cấu trúc báo cáo . . . 3

2 TỔNG QUAN VỀ CHƯƠNG TRÌNH SỬ DỤNG VÀ CÁC PHƯƠNG PHÁP SỬ DỤNG 4

2.1 Các phương pháp khai thác dữ liệu bằng Excel . . . 4

2.1.1 Tổng quan về Excel . . . 4

2.1.2 Phương pháp thống kê mô tả trong Excel . . . 4

2.1.3 Vai trò của Excel trong pipeline hiện tại . . . 5

2.2 Phần mềm Orange . . . 5

2.2.1 Tổng quan về phần mềm Orange . . . 5

2.2.2 Khung phương pháp luận đang sử dụng . . . 5

2.2.3 Các widget được sử dụng trong đề tài . . . 6

2.3 Các phương pháp sử dụng trong đề tài . . . 7

2.3.1 Mô hình RFM như một lớp tham chiếu . . . 7

2.3.2 Chuẩn hóa dữ liệu trước khi phân cụm . . . 7

2.3.3 Loại bỏ outlier bằng Local Outlier Factor . . . 7

2.3.4 Thuật toán K-Means . . . 8

2.3.5 Đánh giá số cụm bằng Silhouette . . . 8

2.3.6 Trực quan hóa bằng t-SNE 9

3 ỨNG DỤNG PHƯƠNG PHÁP VÀO BÀI TOÁN THỰC TẾ 10

3.1 Giới thiệu bộ dữ liệu 10

3.2 Tiền xử lý dữ liệu 11

3.2.1 Xử lý giá trị bị thiếu 11

3.2.2 Loại bỏ các giá trị không hợp lệ 13

3.2.3 Xử lý dữ liệu trùng lặp 14

3.2.4 Tạo biến mới 15

3.2.5 Kết quả sau tiền xử lý 15

3.3 Phân tích khám phá dữ liệu 16

3.3.1 Doanh thu theo thời gian 16

3.3.2 Hành vi mua sắm theo thời gian 17

3.3.3 Phân tích sản phẩm và khách hàng 18

3.3.4 Phân tích RFM 19

3.3.5 Nhận xét chung về dữ liệu 19

4 XÂY DỰNG MÔ HÌNH PHÂN CỤM KHÁCH HÀNG 20

4.1 Tạo đặc trưng 20

4.1.1 Tại sao cần mở rộng thêm đặc trưng 20

4.1.2 Bộ đặc trưng ở cấp khách hàng 20

4.1.3 Thực hiện trong Orange 22

4.1.4 Chuẩn hóa và loại bỏ outlier 25

4.2 Mô hình phân cụm khách hàng 28

4.2.1 Thiết lập thực nghiệm 28

4.2.2 So sánh lựa chọn số cụm giữa k=3 và k=4 28

4.2.3 Trực quan hóa bằng t-SNE 30

4.2.4 Diễn giải kết quả với k=3 31

4.2.5 Diễn giải kết quả với k=4 33

4.2.6 Lựa chọn cấu hình cuối cùng 35

5 ĐỀ XUẤT PHƯƠNG PHÁP VÀ HƯỚNG MỞ RỘNG 37

5.1 Định hướng sử dụng kết quả phân cụm trong thực tế 37

5.1.1 Nhóm khách hàng quy mô nhỏ 37

5.1.2 Nhóm khách mua đa dạng 37

5.1.3 Nhóm mua thường xuyên, số lượng lớn 38

5.2 Giá trị bổ sung của phương án k=4 38

5.3 Hoàn thiện workflow trong Orange 39

5.4 Mở rộng theo hướng dashboard và tự động hóa 39

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN **40**

TÀI LIỆU THAM KHẢO **41**

DANH MỤC HÌNH ẢNH

2.1 Khung phương pháp luận phản ánh đúng pipeline hiện tại của đề tài 6

3.1 Minh họa một phần dữ liệu gốc Online Retail trong Excel 11
3.2 Kết quả thống kê dữ liệu thiếu trong Excel 11
3.3 Lọc các dòng thiếu `CustomerID` trong Excel 12
3.4 Dữ liệu sau khi loại missing và nhập vào Orange 12
3.5 Top 10 quốc gia có số lượng giao dịch nhiều nhất 13
3.6 Lọc dữ liệu hợp lệ bằng widget `Select Rows` 14
3.7 Loại bỏ dữ liệu trùng lặp bằng widget `Unique` 14
3.8 Tạo biến `TotalPrice` bằng widget `Formula` 15
3.9 Kết quả dữ liệu sau khi hoàn thành tiền xử lý 16
3.10 Biểu đồ doanh thu theo ngày và theo tháng 17
3.11 Heatmap tần suất mua hàng theo giờ và ngày trong tuần 18
3.12 Top 10 sản phẩm bán chạy trong bộ dữ liệu 18
3.13 Phân phối của các chỉ số RFM 19

4.1 Tạo nhóm feature cơ bản bằng `Group By` theo `CustomerID` 22
4.2 Đổi tên các feature bằng widget `Edit Domain` 22
4.3 Tạo feature theo hóa đơn trong Orange 23
4.4 Tạo feature theo sản phẩm trong Orange 24
4.5 Hợp nhất các bảng đặc trưng bằng `Merge Data` 24
4.6 Phân phối của các feature trước khi biến đổi hoặc chuẩn hóa 25
4.7 Thiết lập widget `Normalize` để chuẩn hóa bộ 16 feature trong Orange 26
4.8 Phân phối của các feature sau khi chuẩn hóa 27
4.9 Thiết lập widget `Outliers` bằng LOF trước khi chạy K-Means 28
4.10 Diễn biến silhouette theo số cụm và tỷ trọng các cụm ở hai phương án được giữ lại 29
4.11 Trực quan hóa phân cụm bằng t-SNE với k=3 30
4.12 Trực quan hóa phân cụm bằng t-SNE với k=4 31
4.13 Biểu đồ radar hồ sơ cụm khách hàng khi k=3 32
4.14 Biểu đồ radar hồ sơ cụm khách hàng khi k=4 34

DANH MỤC BẢNG BIỂU

1 Phân công công việc trong nhóm ii

3.1 Mô tả các biến trong tập dữ liệu Online Retail 10

4.1 Bộ 16 đặc trưng khách hàng sử dụng cho phân cụm 21
4.2 Tóm tắt ý nghĩa kinh doanh của các cụm khi k=3 33
4.3 Tóm tắt ý nghĩa kinh doanh của các cụm khi k=4 35

CHƯƠNG 1

GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU VÀ GIỚI THIỆU ĐỀ TÀI

1.1. Giới thiệu về khoa học dữ liệu

1.1.1. Tổng quan về dữ liệu

Dữ liệu là tập hợp các quan sát phản ánh một hiện tượng, một hoạt động hay một đối tượng trong thực tế. Trong môi trường kinh doanh, dữ liệu có thể xuất hiện dưới nhiều dạng như dữ liệu giao dịch bán hàng, dữ liệu khách hàng, dữ liệu sản phẩm, dữ liệu hành vi truy cập và dữ liệu phản hồi từ thị trường. Giá trị thực sự của dữ liệu không nằm ở việc lưu trữ số lượng lớn bản ghi, mà nằm ở khả năng biến dữ liệu thành thông tin phục vụ ra quyết định.

Đối với bài toán bán lẻ, mỗi giao dịch mua hàng đều mang theo nhiều tín hiệu quan trọng như thời điểm mua, số lượng mua, mức giá, giá trị hóa đơn và danh mục sản phẩm. Khi các tín hiệu này được tổng hợp ở cấp khách hàng, doanh nghiệp có thể trả lời những câu hỏi thiết thực như khách hàng nào mua thường xuyên, khách hàng nào đóng góp doanh thu cao, hay nhóm nào có xu hướng mua đa dạng sản phẩm hơn.

1.1.2. Tổng quan về khoa học dữ liệu

Khoa học dữ liệu là lĩnh vực kết hợp giữa thống kê, công nghệ thông tin và kiến thức miền ứng dụng nhằm khai thác dữ liệu để tạo ra tri thức có ích [9]. Quy trình làm việc trong một bài toán khoa học dữ liệu thường bao gồm các bước: hiểu bài toán, thu thập dữ liệu, tiền xử lý dữ liệu, phân tích khám phá, xây dựng mô hình, đánh giá kết quả và triển khai ứng dụng.

Trong đề tài này, bài toán được tiếp cận theo hướng *học không giám sát* vì dữ liệu không có sẵn nhãn phân khúc khách hàng. Thay vì dự đoán một biến đầu ra cụ thể, mô hình sẽ tìm kiếm cấu trúc tiềm ẩn trong dữ liệu để nhóm các khách hàng có hành vi tương đồng thành các cụm. Đây là một dạng phân tích có tính khám phá cao, rất phù hợp với bối cảnh doanh nghiệp muốn hiểu khách hàng từ dữ liệu lịch sử giao dịch [5].

1.2. Giới thiệu đề tài

1.2.1. Lý do chọn đề tài

Trong môi trường bán lẻ, khách hàng không có hành vi mua sắm đồng nhất. Có khách hàng mua thường xuyên nhưng giá trị đơn hàng thấp, có khách hàng mua ít lần nhưng mỗi lần chi tiêu lớn, và cũng có nhóm khách hàng trải rộng trên nhiều loại sản phẩm khác nhau. Nếu doanh nghiệp áp dụng một chiến lược duy nhất cho toàn bộ tập khách hàng, hiệu quả marketing thường không cao và ngân sách dễ bị phân tán.

Vì vậy, nhóm lựa chọn đề tài phân cụm khách hàng nhằm hướng tới hai mục tiêu. Thứ nhất, đề tài giúp vận dụng kiến thức khoa học dữ liệu vào một bài toán có ý nghĩa kinh doanh rõ rệt. Thứ hai, đề tài phù hợp với định hướng học phần khi sử dụng Excel và Orange để xây dựng toàn bộ quy trình phân tích theo hướng trực quan, dễ kiểm chứng và thuận tiện khi trình bày kết quả.

1.2.2. Mục tiêu nghiên cứu

Đề tài được thực hiện với các mục tiêu chính sau:

- Làm sạch bộ dữ liệu giao dịch bán lẻ để tạo được tập dữ liệu nhất quán, đáng tin cậy cho quá trình phân tích.
- Khai thác các đặc trưng hành vi khách hàng từ dữ liệu giao dịch, không dừng ở bộ chỉ số RFM truyền thống.
- Áp dụng thuật toán K-Means để phân cụm khách hàng và trực quan hóa kết quả trong không gian giảm chiều.
- Đề xuất định hướng ứng dụng các cụm khách hàng vào hoạt động marketing và quản trị quan hệ khách hàng.

1.2.3. Câu hỏi nghiên cứu

Đề tài tập trung trả lời các câu hỏi sau:

- Dữ liệu giao dịch bán lẻ cần được xử lý như thế nào để đủ chất lượng cho bài toán phân cụm khách hàng?
- Những đặc trưng nào ở cấp khách hàng có khả năng phản ánh hành vi mua sắm đa chiều tốt hơn bộ chỉ số RFM cơ bản?
- Số cụm khách hàng phù hợp có thể được xác định bằng những tiêu chí nào?

- Mỗi cụm khách hàng mang đặc điểm kinh doanh gì và có thể gắn với những hành động quản trị nào?

1.2.4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là dữ liệu giao dịch bán lẻ trực tuyến trong tập dữ liệu *Online Retail*. Phạm vi xử lý tập trung vào các giao dịch hợp lệ tại thị trường *United Kingdom*, bởi đây là quốc gia chiếm tỷ trọng lớn nhất trong dữ liệu và có tính đồng nhất tốt hơn khi phân tích hành vi tiêu dùng.

Về công cụ, Excel được dùng cho các bước kiểm tra dữ liệu thô và xử lý giá trị thiếu; Orange được dùng để xây dựng workflow tiền xử lý, tổng hợp đặc trưng, trực quan hóa và mô hình hóa. Phần lập trình chỉ được sử dụng ở mức bổ trợ cho các biểu đồ đã có sẵn trong workspace.

1.2.5. Cấu trúc báo cáo

Báo cáo gồm năm chương chính. Chương 1 giới thiệu bối cảnh khoa học dữ liệu và lý do chọn đề tài. Chương 2 trình bày tổng quan về Excel, Orange, thuật toán K-Means và các phương pháp đánh giá số cụm. Chương 3 mô tả bộ dữ liệu, quá trình tiền xử lý và phân tích khám phá dữ liệu. Chương 4 trình bày cách xây dựng đặc trưng khách hàng, quy trình phân cụm và khung diễn giải kết quả. Chương 5 tổng hợp các đề xuất nhằm hoàn thiện mô hình và mở rộng ứng dụng trong thực tế. Phần cuối cùng là kết luận và hướng phát triển tiếp theo.

CHƯƠNG 2

TỔNG QUAN VỀ CHƯƠNG TRÌNH SỬ DỤNG VÀ CÁC PHƯƠNG PHÁP SỬ DỤNG

2.1. Các phương pháp khai thác dữ liệu bằng Excel

2.1.1. Tổng quan về Excel

Microsoft Excel là công cụ quen thuộc trong xử lý dữ liệu bảng nhờ khả năng nhập liệu, lọc dữ liệu, kiểm tra nhanh các cột và tính toán thống kê mô tả cơ bản. Trong phạm vi đề tài này, Excel không đóng vai trò là môi trường huấn luyện mô hình, mà được dùng như lớp kiểm định ban đầu nhằm bảo đảm dữ liệu đầu vào đủ sạch trước khi đưa sang Orange.

Cách tiếp cận này phù hợp với mục tiêu học phần. Thay vì đi thẳng vào mô hình, nhóm ưu tiên kiểm soát chất lượng dữ liệu thô, xác định số lượng bản ghi thiếu, rà soát các dòng không thể định danh khách hàng và chuẩn bị một tệp trung gian có cấu trúc rõ ràng. Điều này giúp toàn bộ pipeline phía sau trong Orange ổn định hơn và cũng làm cho phần trình bày báo cáo trở nên dễ kiểm chứng hơn.

2.1.2. Phương pháp thống kê mô tả trong Excel

Excel hỗ trợ tốt cho những thao tác tiền xử lý có tính trực quan cao. Cụ thể, nhóm sử dụng Excel để:

- đếm số lượng giá trị bị thiếu bằng các hàm thống kê cơ bản;
- lọc các dòng thiếu `CustomerID` trước khi chuyển sang Orange;
- quan sát nhanh dữ liệu gốc theo cột, kiểu dữ liệu và mức độ hợp lệ;
- đối chiếu thủ công một số quan sát bất thường trước khi quyết định quy tắc làm sạch.

Ưu điểm của Excel trong giai đoạn đầu là dễ quan sát từng dòng dữ liệu, từ đó giúp giải thích rõ lý do vì sao một số bản ghi bị loại bỏ. Đối với báo cáo học phần, đây là điểm quan trọng vì người đọc có thể theo dõi mạch xử lý từ dữ liệu thô đến dữ liệu sẵn sàng cho mô hình mà không bị "nhảy bước".

2.1.3. Vai trò của Excel trong pipeline hiện tại

Trong pipeline hiện tại, Excel đảm nhiệm ba vai trò chính. Thứ nhất, Excel giúp phát hiện các dòng thiếu khóa `CustomerID`, là trường bắt buộc để tổng hợp hành vi ở cấp khách hàng. Thứ hai, Excel hỗ trợ tạo một tệp dữ liệu trung gian đã loại bỏ missing value quan trọng, nhờ đó Orange chỉ còn tập trung vào các bước lọc nghiệp vụ, tạo đặc trưng và mô hình hóa. Thứ ba, Excel đóng vai trò như công cụ đối chiếu, giúp xác nhận lại một số thống kê quan trọng như số dòng còn lại sau làm sạch hoặc cấu trúc của bộ dữ liệu trước khi nhóm triển khai phân cụm.

2.2. Phần mềm Orange

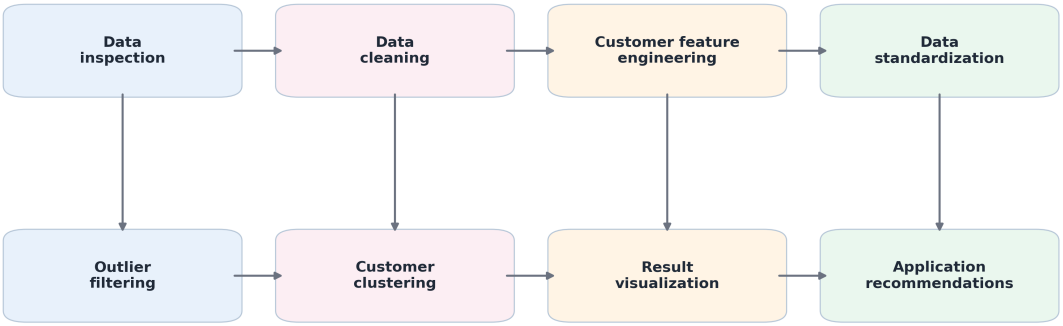
2.2.1. Tổng quan về phần mềm Orange

Orange là một môi trường khai phá dữ liệu trực quan theo hướng kéo thả widget, cho phép người dùng xây dựng quy trình phân tích mà không cần viết mã nguồn dài [3]. Thay vì lập trình toàn bộ pipeline, người dùng có thể kết nối tuần tự các widget như `File`, `Select Rows`, `Unique`, `Formula`, `Group By`, `Merge Data`, `Normalize`, `Outliers`, `k-Means` và `t-SNE`.

Đối với đề tài này, Orange đặc biệt phù hợp vì hai lý do. Thứ nhất, bản chất bài toán là khám phá dữ liệu và phân cụm khách hàng, nên một workflow trực quan giúp nhóm theo dõi rõ dữ liệu thay đổi như thế nào sau từng bước. Thứ hai, Orange cho phép lưu toàn bộ quy trình dưới dạng file `.ows`, nhờ đó bài làm có thể tái hiện lại và kiểm tra từng thao tác một cách minh bạch.

2.2.2. Khung phương pháp luận đang sử dụng

So với bài tham khảo chính, pipeline hiện tại của nhóm có một điểm điều chỉnh quan trọng: sau bước chuẩn hóa, nhóm loại outlier bằng Local Outlier Factor trước khi chạy K-Means, và dùng t-SNE để trực quan hóa thay cho PCA. Việc cập nhật này xuất phát từ thực nghiệm trên dữ liệu thực tế, khi các quan sát cực trị có xu hướng kéo lệch tâm cụm và làm xuất hiện những cụm rất nhỏ, khó diễn giải nếu đưa thẳng vào K-Means.



Hình 2.1: Khung phương pháp luận phản ánh đúng pipeline hiện tại của đề tài

Khung trên cho thấy quy trình được triển khai thành tám chặng liên tiếp: kiểm tra dữ liệu trong Excel, tiền xử lý trong Orange, tạo bộ 16 feature ở cấp khách hàng, chuẩn hóa, loại outlier, phân cụm, trực quan hóa và cuối cùng là gắn kết quả phân cụm với hành động kinh doanh. Cách tổ chức này giúp báo cáo không chỉ mô tả kỹ thuật, mà còn thể hiện được luồng chuyển đổi từ dữ liệu giao dịch sang tri thức quản trị.

2.2.3. Các widget được sử dụng trong đề tài

Nhóm sử dụng các widget chính sau:

- `File`: nhập dữ liệu từ Excel hoặc định dạng `.tab`;
- `Data Table`: quan sát nhanh số dòng, số cột và nội dung dữ liệu;
- `Select Rows`: lọc các giao dịch hợp lệ theo điều kiện;
- `Unique`: kiểm tra và loại bỏ bản ghi trùng lặp hoàn toàn;
- `Formula`: tạo biến mới `TotalPrice`;
- `Group By`: tổng hợp dữ liệu lên cấp khách hàng, hóa đơn hoặc sản phẩm;
- `Edit Domain`: đổi tên các cột đặc trưng để thuận tiện trình bày;
- `Merge Data`: hợp nhất nhiều bảng đặc trưng theo khóa `CustomerID`;
- `Normalize`: đưa các feature về cùng thang đo;
- `Outliers`: phát hiện và loại bỏ các khách hàng cực trị;

- `k-Means`: phân cụm khách hàng;
- `t-SNE`: trực quan hóa hình học của cụm trong không gian hai chiều.

Điểm mạnh của Orange trong bài toán này không chỉ nằm ở số lượng widget, mà nằm ở khả năng làm lộ rõ logic phân tích. Người đọc có thể nhìn vào workflow để biết chính xác dữ liệu đã đi qua những bước nào, được biến đổi ra sao và mô hình cuối cùng đang nhận đầu vào là gì.

2.3. Các phương pháp sử dụng trong đề tài

2.3.1. Mô hình RFM như một lớp tham chiếu

RFM là bộ ba chỉ số kinh điển trong phân tích khách hàng, bao gồm *Recency*, *Frequency* và *Monetary* [10]. Trong nhiều nghiên cứu về hành vi mua sắm, RFM thường được dùng như lớp mô tả đầu tiên để tóm tắt mức độ gần đây, tần suất và giá trị chi tiêu của khách hàng. Trên chính dữ liệu Online Retail, Chen và cộng sự cũng sử dụng RFM làm nền tảng để phân khúc khách hàng bằng K-Means [2].

Trong đề tài này, RFM được xem là lớp tham chiếu quan trọng, nhưng không phải là điểm dừng. Nhóm chỉ dùng RFM để quan sát phân phối và nắm nhanh cấu trúc hành vi mua sắm tổng quát. Sau đó, nhóm mở rộng sang bộ 16 feature ở cấp khách hàng để phản ánh thêm chiều sâu về số lượng mua, độ đa dạng sản phẩm, hành vi theo hóa đơn và mức độ lặp lại theo từng mã hàng.

2.3.2. Chuẩn hóa dữ liệu trước khi phân cụm

Sau khi tổng hợp từ dữ liệu giao dịch sang cấp khách hàng, các đặc trưng có quy mô rất khác nhau. Một số biến như `Sum_TotalPrice` hoặc `Count_Invoice` có thể lớn hơn rất nhiều lần so với các biến trung bình theo sản phẩm hoặc theo hóa đơn. Nếu đưa trực tiếp vào K-Means, những biến có biên độ lớn sẽ chi phối khoảng cách Euclid và làm mô hình thiên lệch [5].

Vì lý do đó, nhóm chuẩn hóa dữ liệu về cùng thang đo bằng Z-score trước khi phân cụm. Bản chất của bước này là đưa các biến về trung tâm xấp xỉ 0 và độ lệch chuẩn xấp xỉ 1, giúp khoảng cách tính trong K-Means phản ánh đồng đều hơn trên toàn bộ không gian đặc trưng. Đây là bước đặc biệt quan trọng khi làm việc với các bộ feature được tổng hợp từ nhiều logic khác nhau như trong bài toán phân khúc khách hàng.

2.3.3. Loại bỏ outlier bằng Local Outlier Factor

Local Outlier Factor (LOF) là phương pháp phát hiện ngoại lệ dựa trên mật độ cục bộ của điểm dữ liệu so với các láng giềng gần nhất [7]. Trực giác của LOF là: nếu một khách hàng nằm trong vùng có mật độ thấp hơn đáng kể so với các điểm lân cận, khách hàng đó có thể là ngoại lệ.

Nhóm sử dụng LOF trước K-Means vì mô hình tâm cụm rất nhạy với các quan sát cực trị. Trên dữ liệu bán lẻ, nếu giữ nguyên toàn bộ các điểm có giá trị quá cao về số lượng mua hoặc doanh thu, K-Means dễ sinh ra một hoặc vài cụm rất nhỏ chỉ để "ôm" các điểm cực trị này. Điều đó làm giảm tính ổn định của các cụm còn lại và gây khó khăn khi diễn giải trên góc độ kinh doanh.

Tuy nhiên, loại outlier trong bài toán khách hàng không có nghĩa là mặc định xem mọi khách hàng giá trị cao là không quan trọng. Trong phạm vi bài này, LOF được dùng với mục tiêu kỹ thuật là làm cho mặt bằng phân cụm tổng quát rõ hơn; còn các cụm cực nhỏ sau đó vẫn được phân tích như một tín hiệu nghiệp vụ riêng trong phương án $k = 4$.

2.3.4. Thuật toán K-Means

K-Means là thuật toán phân cụm phân hoạch cổ điển, được giới thiệu bởi MacQueen [4]. Ý tưởng cốt lõi của thuật toán là chia tập dữ liệu thành k nhóm sao cho tổng khoảng cách từ các điểm đến tâm cụm tương ứng là nhỏ nhất. Thuật toán lặp lại hai bước chính: gán từng điểm vào tâm cụm gần nhất và cập nhật lại tâm cụm bằng trung bình của các điểm trong cụm đó.

Theo tổng quan của Jain, Murty và Flynn, K-Means là một trong những phương pháp phổ biến nhất trong bài toán phân cụm dữ liệu số nhờ tính đơn giản, tốc độ xử lý nhanh và khả năng mở rộng tốt [5]. Với đề tài này, K-Means phù hợp vì toàn bộ bộ feature sau chuẩn hóa đều là biến liên tục và mục tiêu của bài toán là tìm ra những nhóm khách hàng có hành vi đủ đồng nhất để phục vụ diễn giải.

Hạn chế chính của K-Means là phải chọn trước số cụm và nhạy với dữ liệu ngoại lệ. Vì vậy, pipeline hiện tại mới kết hợp cả chuẩn hóa, loại outlier và đánh giá số cụm thay vì chạy K-Means trực tiếp trên dữ liệu gốc.

2.3.5. Đánh giá số cụm bằng Silhouette

Silhouette là chỉ số đánh giá chất lượng phân cụm dựa trên đồng thời hai yếu tố: độ chặt bên trong cụm và mức độ tách biệt với cụm gần nhất [6]. Với mỗi điểm dữ liệu, hệ số silhouette được tính từ khoảng cách nội cụm trung bình và khoảng cách tới cụm lân cận gần nhất. Giá trị càng gần 1 thì điểm đó càng "hợp" với cụm hiện tại; giá trị gần 0 cho thấy vùng chồng lấn; còn giá trị âm gợi ý rằng điểm có thể đang bị gán sai cụm.

Trong bài này, silhouette không được dùng một cách cơ học để chọn số cụm. Nhóm dùng nó như chỉ số định lượng nền, sau đó kết hợp thêm tiêu chí về kích thước cụm và khả năng diễn giải kinh doanh. Đây là lý do mà hai phương án $k = 3$ và $k = 4$ đều được giữ lại để so sánh song song trong phần thực nghiệm.

2.3.6. Trực quan hóa bằng t-SNE

t-SNE là kỹ thuật giảm chiều phi tuyến được phát triển nhằm trực quan hóa dữ liệu nhiều chiều trong không gian hai hoặc ba chiều [8]. Khác với PCA thiên về bảo toàn phương sai toàn cục, t-SNE ưu tiên giữ lại cấu trúc lân cận cục bộ, nhờ đó thường cho hình trực quan các nhóm dữ liệu rõ hơn trong các bài toán khám phá.

Trong đề tài này, t-SNE chỉ được dùng cho mục tiêu minh họa và diễn giải hình học của các cụm sau khi K-Means đã gán nhãn. Điều cần nhấn mạnh là nhóm không dùng tọa độ t-SNE làm dữ liệu đầu vào để huấn luyện K-Means. Nói cách khác, K-Means vẫn chạy trên bộ 16 feature đã chuẩn hóa; t-SNE chỉ là lớp nhìn trực quan giúp người đọc dễ quan sát hơn mức độ cô lập hay chồng lấn giữa các cụm.

CHƯƠNG 3

ỨNG DỤNG PHƯƠNG PHÁP VÀO BÀI TOÁN THỰC TẾ

3.1. Giới thiệu bộ dữ liệu

Dự án sử dụng tập dữ liệu *Online Retail*, ghi nhận các giao dịch mua hàng của một cửa hàng bán lẻ trực tuyến trong giai đoạn từ ngày 01/12/2010 đến ngày 09/12/2011 [1]. Bộ dữ liệu gốc gồm $541{,}909$ dòng và 8 cột, phản ánh thông tin về hóa đơn, mã sản phẩm, mô tả sản phẩm, số lượng, thời gian giao dịch, đơn giá, mã khách hàng và quốc gia. Theo mô tả của UCI, doanh nghiệp trong bộ dữ liệu chủ yếu bán các mặt hàng quà tặng và có một tỷ trọng đáng kể khách hàng thuộc nhóm mua sỉ, do đó đây là nguồn dữ liệu phù hợp cho bài toán phân khúc khách hàng [1, 2].

Bảng 3.1: Mô tả các biến trong tập dữ liệu Online Retail

STT	Tên cột	Kiểu dữ liệu	Mô tả
1	InvoiceNo	Categorical	Mã hóa đơn giao dịch.
2	StockCode	Categorical	Mã sản phẩm.
3	Description	Text	Tên sản phẩm.
4	Quantity	Numeric	Số lượng sản phẩm được mua.
5	InvoiceDate	Datetime	Thời điểm phát sinh giao dịch.
6	UnitPrice	Numeric	Giá của mỗi đơn vị sản phẩm.
7	CustomerID	Categorical	Mã định danh khách hàng.
8	Country	Categorical	Quốc gia của khách hàng.

	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/10 08:26	2,55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	1/12/10 08:26	3,39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/10 08:26	2,75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/10 08:26	3,39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/10 08:26	3,39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	1/12/10 08:26	7,65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	1/12/10 08:26	4,25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	1/12/10 08:28	1,85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	1/12/10 08:28	1,85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	1/12/10 08:34	1,69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	1/12/10 08:34	2,1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	1/12/10 08:34	2,1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	1/12/10 08:34	3,75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	1/12/10 08:34	1,65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	1/12/10 08:34	4,25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	1/12/10 08:34	4,95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	1/12/10 08:34	9,95	13047	United Kingdom
19	536367	21754	HOME BUILDING BLOCK WORD	3	1/12/10 08:34	5,95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BLOCK WORD	3	1/12/10 08:34	5,95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH METAL HEART	4	1/12/10 08:34	7,95	13047	United Kingdom
22	536367	48187	DOORMAT NEW ENGLAND	4	1/12/10 08:34	7,95	13047	United Kingdom
23	536368	22960	JAM MAKING SET WITH JARS	6	1/12/10 08:34	4,25	13047	United Kingdom
24	536368	22913	RED COAT RACK PARIS FASHION	3	1/12/10 08:34	4,95	13047	United Kingdom
25	536368	22912	YELLOW COAT RACK PARIS FASHION	3	1/12/10 08:34	4,95	13047	United Kingdom
26	536368	22914	BLUE COAT RACK PARIS FASHION	3	1/12/10 08:34	4,95	13047	United Kingdom
27	536369	21756	BATH BUILDING BLOCK WORD	3	1/12/10 08:35	5,95	13047	United Kingdom
28	536370	22728	ALARM CLOCK BAKELIKE PINK	24	1/12/10 08:45	3,75	12583	France
29	536370	22727	ALARM CLOCK BAKELIKE RED	24	1/12/10 08:45	3,75	12583	France
30	536370	22726	ALARM CLOCK BAKELIKE GREEN	12	1/12/10 08:45	3,75	12583	France

Hình 3.1: Minh họa một phần dữ liệu gốc Online Retail trong Excel

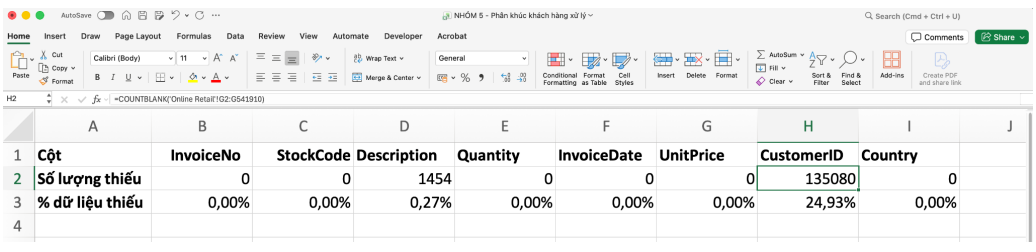
Quan sát dữ liệu mẫu cho thấy một hóa đơn có thể bao gồm nhiều dòng sản phẩm khác nhau, một khách hàng có thể xuất hiện ở nhiều giao dịch và dữ liệu đến từ nhiều quốc gia. Các đặc điểm này cho thấy nếu muốn phân cụm khách hàng, dữ liệu cần được tổng hợp lại ở cấp độ khách hàng thay vì sử dụng trực tiếp từng dòng giao dịch. Đây cũng là cách tiếp cận đã được Chen và cộng sự áp dụng khi xây dựng các biến phân tích trên dữ liệu Online Retail [2].

3.2. Tiền xử lý dữ liệu

3.2.1. Xử lý giá trị bị thiếu

Việc xử lý dữ liệu thiếu được thực hiện trước tiên vì đây là bước nền tảng đảm bảo các giao dịch có đủ thông tin cần thiết cho quá trình phân tích sau đó. Kết quả kiểm tra bằng Excel cho thấy:

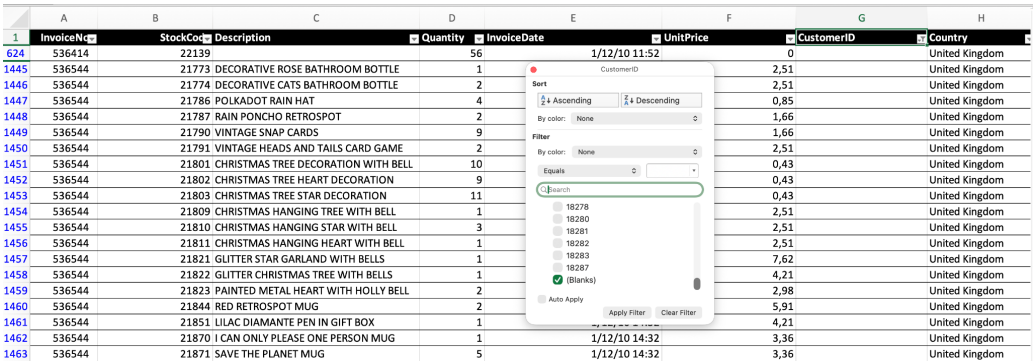
- cột `CustomerID` có 135,080 giá trị bị thiếu, chiếm 24.93%;
- cột `Description` có 1,454 giá trị bị thiếu, chiếm 0.27%;
- các cột còn lại không xuất hiện dữ liệu trống.



	A	B	C	D	E	F	G	H	I	J
1	Cột	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	
2	Số lượng thiếu	0	0	1454	0	0	0	135080	0	
3	% dữ liệu thiếu	0,00%	0,00%	0,27%	0,00%	0,00%	0,00%	24,93%	0,00%	
4										

Hình 3.2: Kết quả thống kê dữ liệu thiếu trong Excel

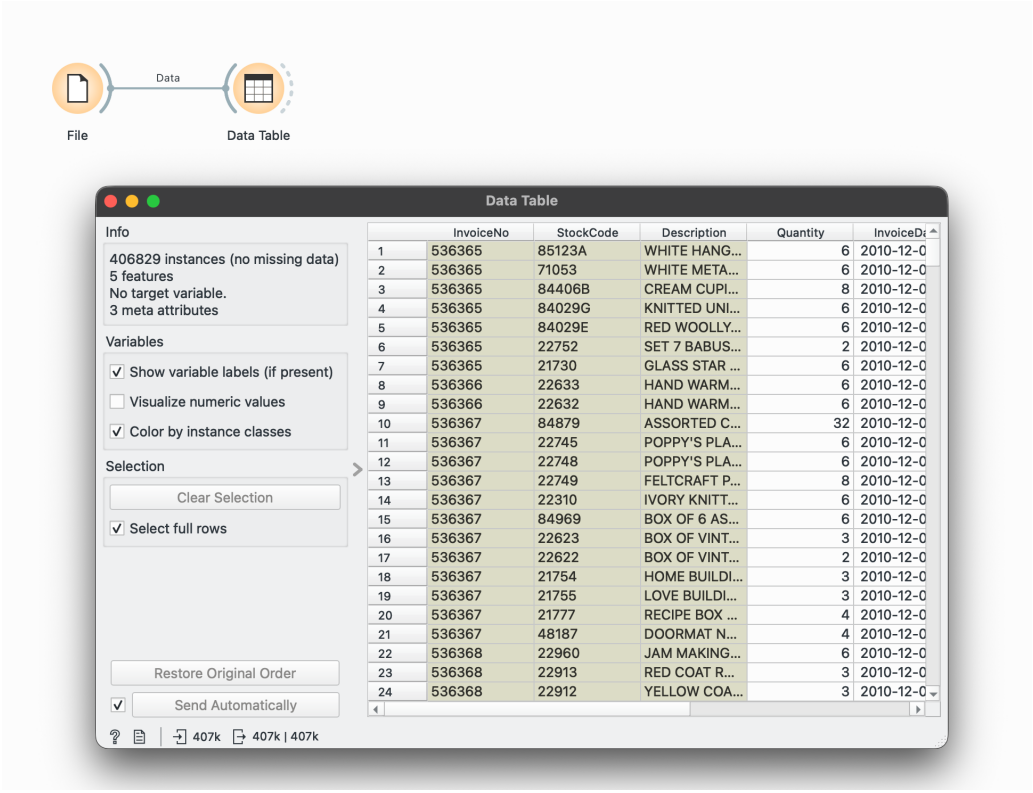
Trong bài toán phân cụm khách hàng, `CustomerID` là trường khóa để liên kết giao dịch với từng khách hàng. Những dòng bị thiếu trường này không thể được dùng để xây dựng feature ở cấp khách hàng, vì vậy nhóm loại bỏ toàn bộ các bản ghi thiếu `CustomerID`. Việc thực hiện được hỗ trợ bằng bộ lọc của Excel nhằm chọn chính xác các dòng rỗng trước khi xóa.



	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCod	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
624	536414	22139		56	1/12/10 11:52	0		United Kingdom
1445	536544	21773	DECORATIVE ROSE BATHROOM BOTTLE	1		2,51		United Kingdom
1446	536544	21774	DECORATIVE CATS BATHROOM BOTTLE	2		2,51		United Kingdom
1447	536544	21786	POLKADOT RAIN HAT	4		0,85		United Kingdom
1448	536544	21787	RAIN PONCHO RETROSPOT	2		1,66		United Kingdom
1449	536544	21790	VINTAGE SNAP CARDS	9		1,66		United Kingdom
1450	536544	21791	VINTAGE HEADS AND TAILS CARD GAME	2		2,51		United Kingdom
1451	536544	21801	CHRISTMAS TREE DECORATION WITH BELL	10		0,43		United Kingdom
1452	536544	21802	CHRISTMAS TREE HEART DECORATION	9		0,43		United Kingdom
1453	536544	21803	CHRISTMAS TREE STAR DECORATION	11		0,43		United Kingdom
1454	536544	21809	CHRISTMAS HANGING TREE WITH BELL	1		2,51		United Kingdom
1455	536544	21810	CHRISTMAS HANGING STAR WITH BELL	3		2,51		United Kingdom
1456	536544	21811	CHRISTMAS HANGING HEART WITH BELL	1		2,51		United Kingdom
1457	536544	21821	GLITTER STAR GARLAND WITH BELLS	1		7,62		United Kingdom
1458	536544	21822	GLITTER CHRISTMAS TREE WITH BELLS	1		4,21		United Kingdom
1459	536544	21823	PAINTED METAL HEART WITH HOLLY BELL	2		2,98		United Kingdom
1460	536544	21844	RED RETROSPOT MUG	2		5,91		United Kingdom
1461	536544	21851	LILAC DIAMANTE PEN IN GIFT BOX	1		4,21		United Kingdom
1462	536544	21870	I CAN ONLY PLEASE ONE PERSON MUG	1	1/12/10 14:32	3,36		United Kingdom
1463	536544	21871	SAVE THE PLANET MUG	5	1/12/10 14:32	3,36		United Kingdom

Hình 3.3: Lọc các dòng thiếu `CustomerID` trong Excel

Sau khi loại các dòng thiếu `CustomerID`, các giá trị thiếu ở cột `Description` cũng biến mất. Điều này cho thấy phần lớn dữ liệu thiếu tập trung trong cùng một nhóm quan sát không hợp lệ. Tệp dữ liệu sau bước này được chuyển vào Orange để tiếp tục xử lý.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDa
1	536365	85123A	WHITE HANG...	6	2010-12-0
2	536365	71053	WHITE META...	6	2010-12-0
3	536365	84406B	CREAM CUPI...	8	2010-12-0
4	536365	84029G	KNITTED UNI...	6	2010-12-0
5	536365	84029E	RED WOOLLY...	6	2010-12-0
6	536365	22752	SET 7 BABUS...	2	2010-12-0
7	536365	21730	GLASS STAR ...	6	2010-12-0
8	536366	22633	HAND WARM...	6	2010-12-0
9	536366	22632	HAND WARM...	6	2010-12-0
10	536367	84879	ASSORTED C...	32	2010-12-0
11	536367	22745	POPPY'S PLA...	6	2010-12-0
12	536367	22748	POPPY'S PLA...	6	2010-12-0
13	536367	22749	FELTCRAFT P...	8	2010-12-0
14	536367	22310	IVORY KNITT...	6	2010-12-0
15	536367	84969	BOX OF 6 AS...	6	2010-12-0
16	536367	22623	BOX OF VINT...	3	2010-12-0
17	536367	22622	BOX OF VINT...	2	2010-12-0
18	536367	21754	HOME BUILDI...	3	2010-12-0
19	536367	21755	LOVE BUILDI...	3	2010-12-0
20	536367	21777	RECIPE BOX ...	4	2010-12-0
21	536367	48187	DOORMAT N...	4	2010-12-0
22	536368	22960	JAM MAKING...	6	2010-12-0
23	536368	22913	RED COAT R...	3	2010-12-0
24	536368	22912	YELLOW COA...	3	2010-12-0

Hình 3.4: Dữ liệu sau khi loại missing và nhập vào Orange

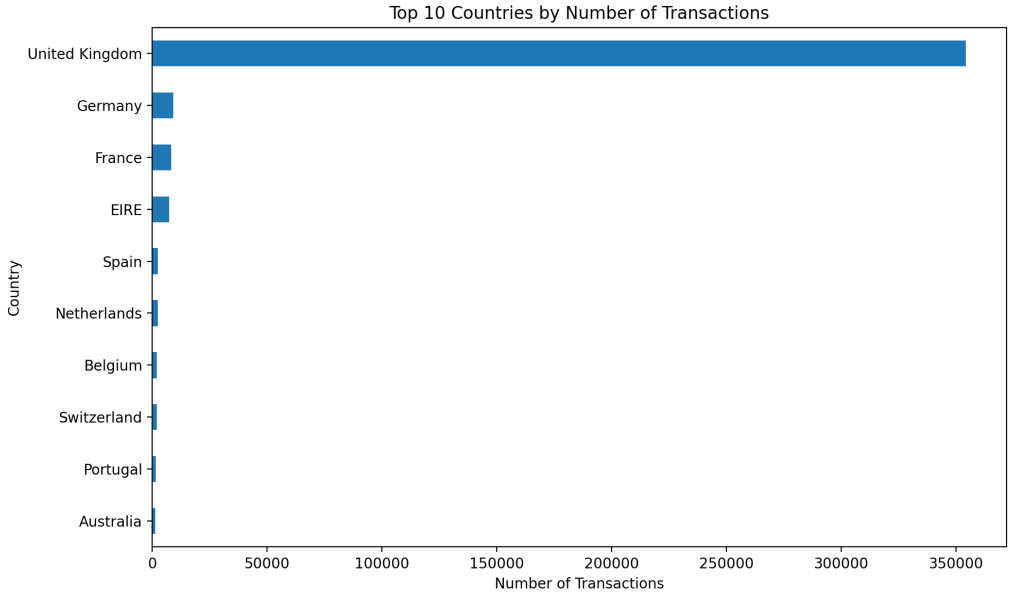
Kết quả kiểm tra trong Orange cho thấy dữ liệu còn lại 406,829 quan sát và không còn missing value.

3.2.2. Loại bỏ các giá trị không hợp lệ

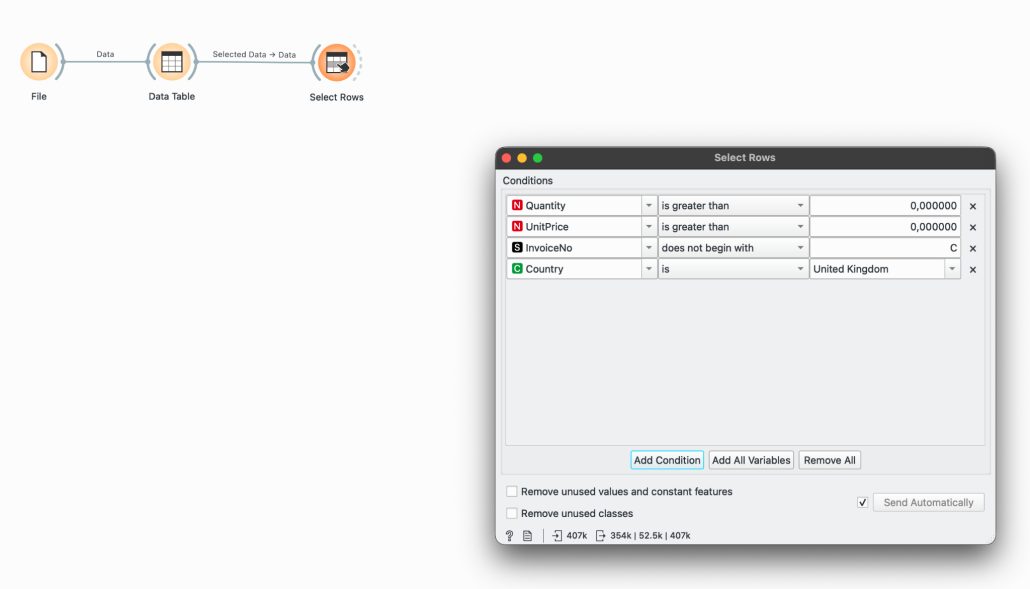
Sau khi xử lý missing value, nhóm tiếp tục loại bỏ các giao dịch không phản ánh hành vi mua hàng thực tế. Các điều kiện lọc được áp dụng gồm:

- `Quantity > 0`;
- `UnitPrice > 0`;
- `InvoiceNo` không bắt đầu bằng ký tự `C`;
- chỉ giữ lại các giao dịch tại `United Kingdom`.

Lý do của bước lọc theo quốc gia là dữ liệu từ United Kingdom chiếm tỷ trọng áp đảo, trong khi mỗi quốc gia có thể có đặc điểm tiêu dùng khác nhau. Giữ lại duy nhất thị trường UK giúp bài toán đồng nhất hơn và giảm nhiễu khi phân cụm.



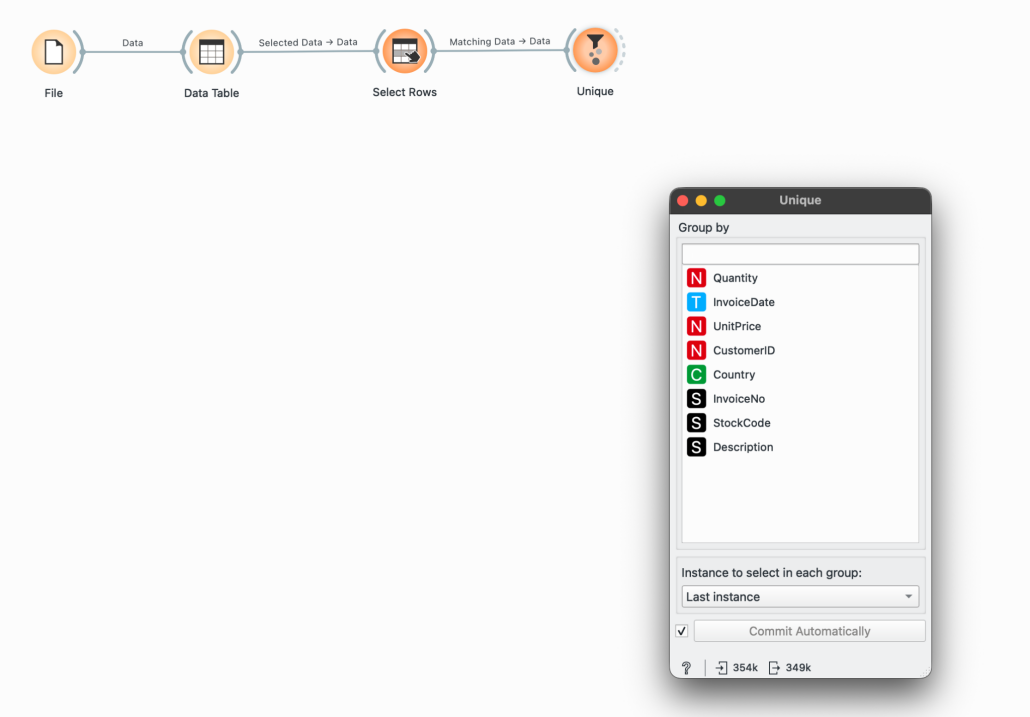
Hình 3.5: Top 10 quốc gia có số lượng giao dịch nhiều nhất



Hình 3.6: Lọc dữ liệu hợp lệ bằng widget `Select Rows`

3.2.3. Xử lý dữ liệu trùng lặp

Dữ liệu trùng lặp là hiện tượng hai hay nhiều dòng có cùng giá trị trên toàn bộ biến. Nếu giữ lại các dòng này, một số thống kê đếm và tổng sẽ bị sai lệch. Vì vậy, nhóm sử dụng widget `Unique` trong Orange để kiểm tra và loại bỏ các bản ghi trùng lặp hoàn toàn.



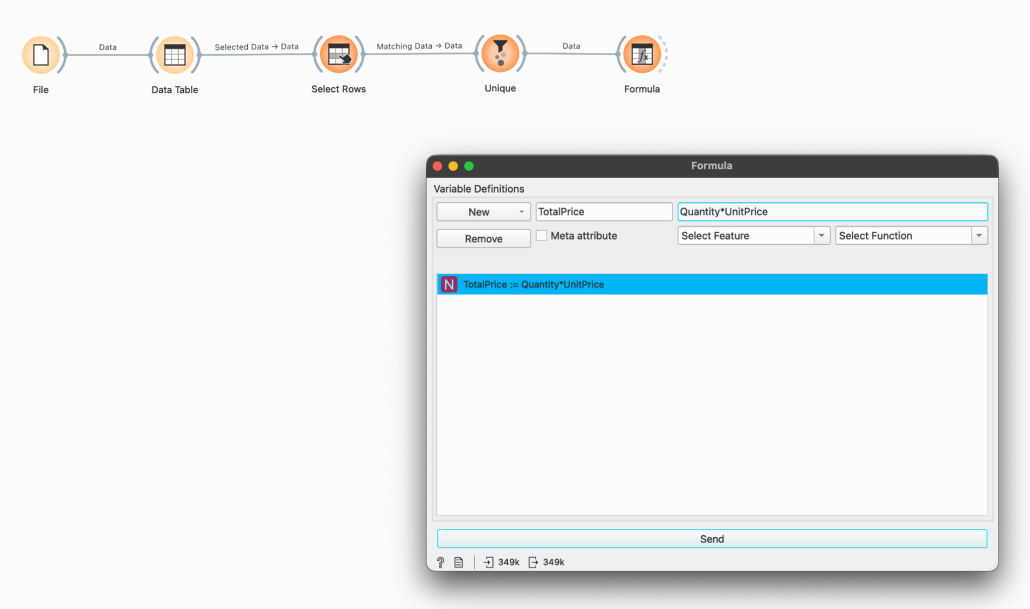
Hình 3.7: Loại bỏ dữ liệu trùng lặp bằng widget `Unique`

3.2.4. Tạo biến mới

Để phục vụ cho việc phân tích chi tiêu, nhóm tạo biến mới:

$$\texttt{TotalPrice} = \texttt{Quantity} \times \texttt{UnitPrice}$$

Biến này phản ánh tổng giá trị của mỗi dòng giao dịch và là cơ sở để tính các chỉ số liên quan tới doanh thu cũng như giá trị khách hàng. Trong Orange, bước này được thực hiện bằng widget `Formula` hoặc `Feature Constructor`.

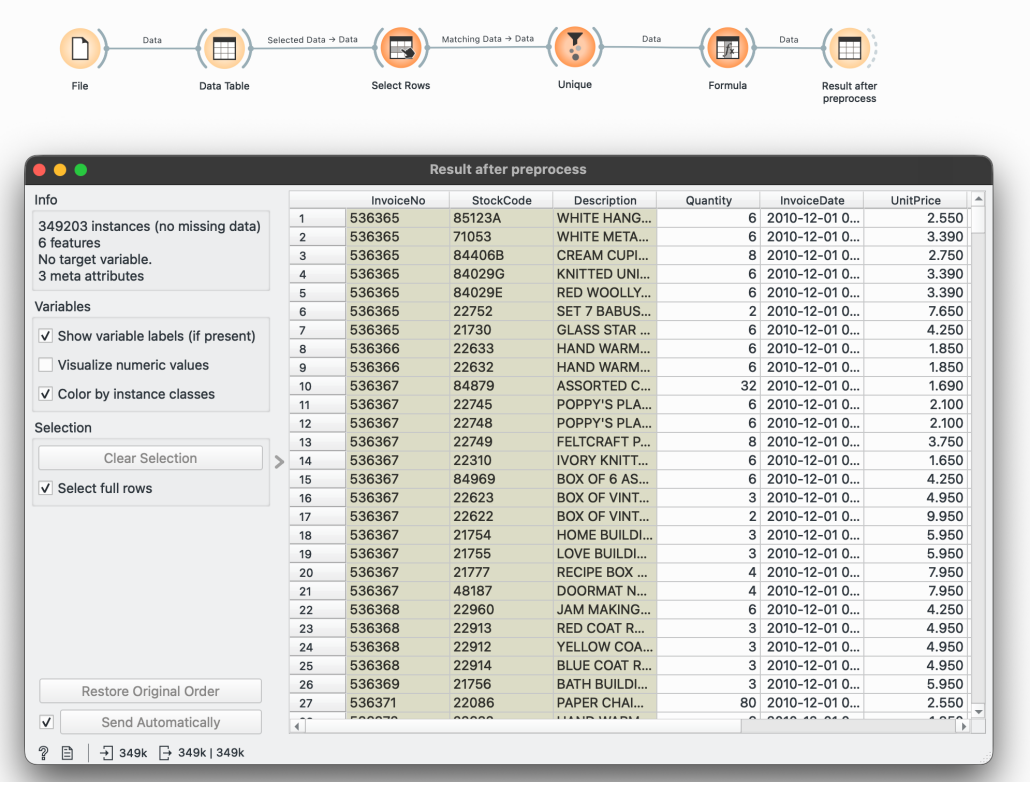


Hình 3.8: Tạo biến `TotalPrice` bằng widget `Formula`

3.2.5. Kết quả sau tiền xử lý

Sau khi hoàn thành các bước làm sạch dữ liệu, bộ dữ liệu sử dụng cho các phân tích tiếp theo có các đặc điểm chính:

- 349,203 bản ghi hợp lệ;
- 3,920 khách hàng;
- 3,665 sản phẩm;
- 16,646 hóa đơn.



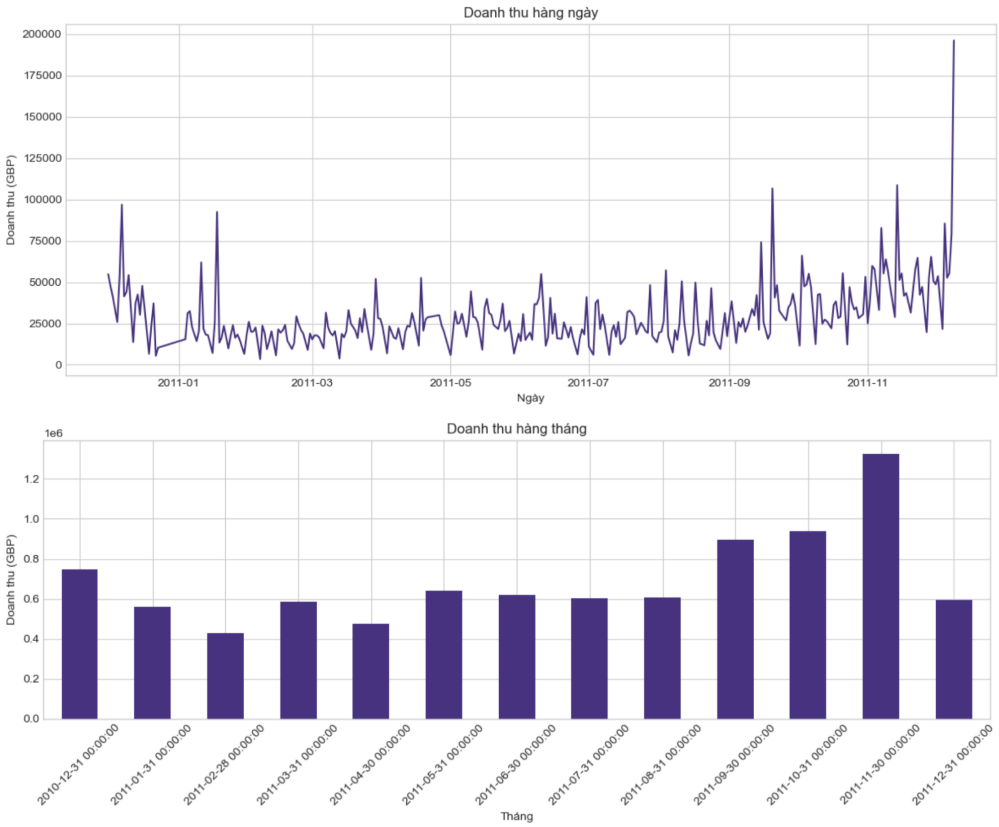
Hình 3.9: Kết quả dữ liệu sau khi hoàn thành tiền xử lý

Tập dữ liệu lúc này không còn missing value, không chứa giao dịch hủy, không có giá trị số lượng và đơn giá không hợp lệ, đồng thời đã được bổ sung biến `TotalPrice`. Đây là cơ sở quan trọng để chuyển sang phân tích khám phá dữ liệu và xây dựng feature ở cấp khách hàng.

3.3. Phân tích khám phá dữ liệu

3.3.1. Doanh thu theo thời gian

Phân tích doanh thu theo thời gian cho phép nhận diện xu hướng tăng trưởng và yếu tố mùa vụ trong hoạt động kinh doanh. Khi quan sát ở cấp ngày, doanh thu biến động mạnh với nhiều đỉnh bất thường, đặc biệt tập trung vào giai đoạn cuối năm. Khi tổng hợp theo tháng, xu hướng tăng trưởng trở nên rõ ràng hơn và cho thấy doanh thu cao điểm rơi vào các tháng cuối năm, phù hợp với bối cảnh mua sắm dịp lễ.

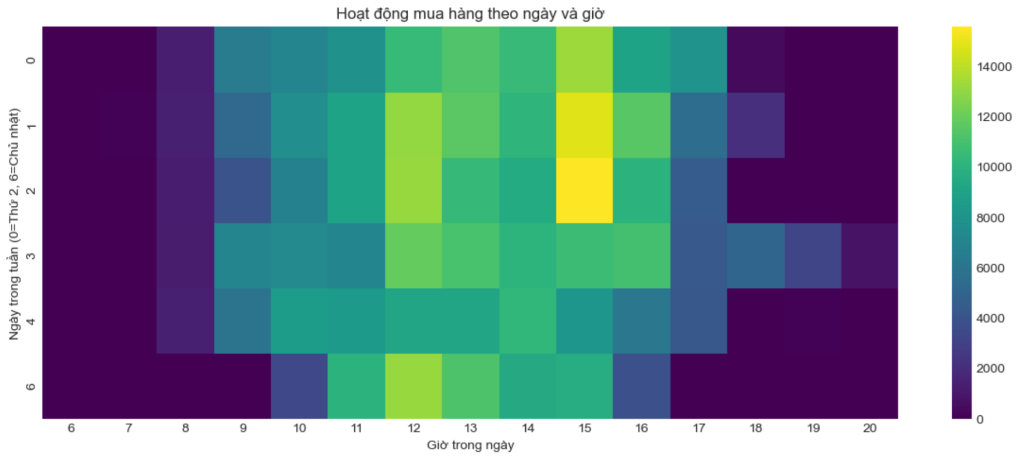


Hình 3.10: Biểu đồ doanh thu theo ngày và theo tháng

Kết quả này cho thấy dữ liệu bán lẻ mang tính mùa vụ rõ rệt. Điều đó không chỉ có ý nghĩa mô tả mà còn hữu ích trong việc hiểu hành vi mua sắm của các nhóm khách hàng khi giải thích kết quả phân cụm sau này.

3.3.2. Hành vi mua sắm theo thời gian

Ngoài doanh thu, việc xem xét thời điểm mua hàng trong ngày và trong tuần giúp nhận diện khung giờ hoạt động chính của khách hàng. Dựa trên heatmap, hoạt động mua sắm thường tập trung vào giờ hành chính, đặc biệt là khoảng từ 10 giờ đến 16 giờ, và mạnh hơn vào các ngày đầu tuần tới giữa tuần so với cuối tuần.

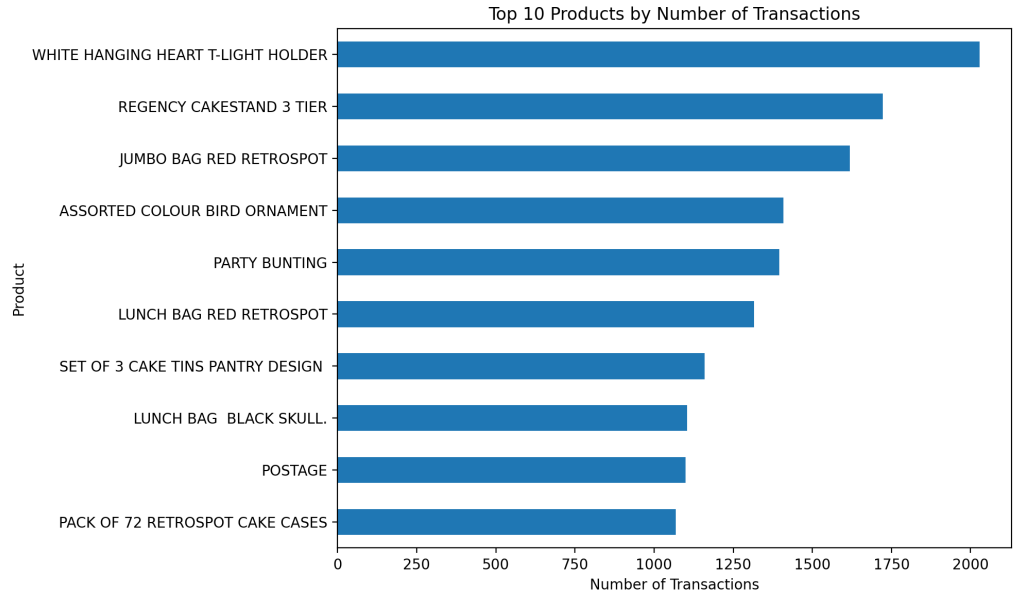


Hình 3.11: Heatmap tần suất mua hàng theo giờ và ngày trong tuần

Từ góc nhìn quản trị, phát hiện này gợi ý rằng các chương trình truyền thông, email marketing hoặc thông báo ưu đãi nên được triển khai gần các khung giờ có xác suất tương tác cao để gia tăng hiệu quả tiếp cận.

3.3.3. Phân tích sản phẩm và khách hàng

Danh mục sản phẩm bán chạy phản ánh nhu cầu phổ biến của thị trường. Các sản phẩm xuất hiện trong nhóm dẫn đầu phần lớn thuộc nhóm quà tặng, đồ trang trí và vật dụng gia đình nhỏ. Đây là những mặt hàng có tính thẩm mỹ cao và thường được mua lặp lại trong các dịp lễ hoặc mua sắm bổ sung.



Hình 3.12: Top 10 sản phẩm bán chạy trong bộ dữ liệu

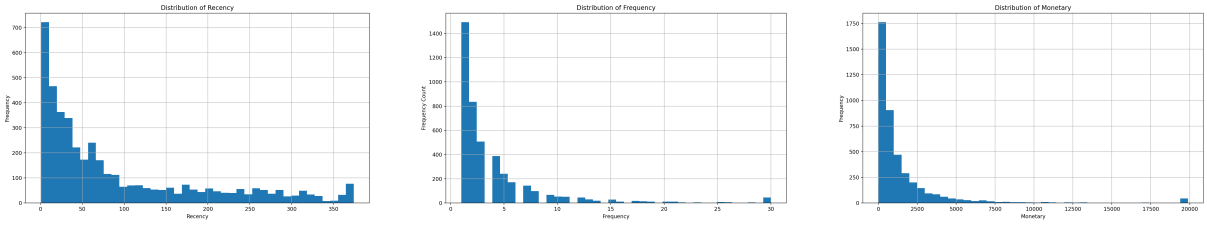
Việc xác định được những sản phẩm nổi bật giúp doanh nghiệp hiểu rõ hơn về cấu trúc

danh mục bán hàng, đồng thời hỗ trợ diễn giải hành vi ở từng cụm khách hàng. Chẳng hạn, một cụm khách hàng mua đa dạng sản phẩm có thể thiên về nhóm quà tặng và vật phẩm trang trí hơn các nhóm khác.

3.3.4. Phân tích RFM

Trước khi đi vào mô hình phân cụm, nhóm kiểm tra phân phối của ba chỉ số tham chiếu RFM. Cả ba chỉ số đều có phân phối lệch phải rõ rệt:

- **Recency**: phần lớn khách hàng có giá trị thấp, cho thấy vẫn có nhiều khách hàng hoạt động gần đây;
- **Frequency**: đa số khách hàng chỉ mua một vài lần, trong khi một nhóm nhỏ giao dịch rất thường xuyên;
- **Monetary**: doanh thu tập trung ở một nhóm khách hàng có giá trị cao.



(a) Recency (b) Frequency (c) Monetary

Hình 3.13: Phân phối của các chỉ số RFM

Sự lệch phải mạnh của RFM cho thấy bài toán có đặc trưng quen thuộc của dữ liệu bán lẻ: một số ít khách hàng đóng góp phần lớn giá trị và tần suất giao dịch. Chính vì thế, nếu chỉ dùng RFM thì mô hình dễ bỏ sót những khía cạnh quan trọng khác như độ đa dạng sản phẩm, hành vi theo từng hóa đơn và mức giá khách hàng thường lựa chọn.

3.3.5. Nhận xét chung về dữ liệu

Từ các kết quả tiền xử lý và EDA, có thể thấy bộ dữ liệu sau khi làm sạch đã đủ chất lượng để chuyển sang giai đoạn xây dựng feature. Dữ liệu phản ánh khá rõ đặc trưng của một doanh nghiệp bán lẻ: doanh thu tăng mạnh theo mùa vụ, khách hàng mua hàng chủ yếu trong giờ hành chính và hành vi chi tiêu phân hóa mạnh giữa các cá nhân.

Những kết quả này là nền tảng quan trọng cho bước kế tiếp. Thay vì phân cụm trực tiếp trên từng giao dịch, nhóm sẽ tổng hợp dữ liệu ở cấp khách hàng để mô hình học được các dạng hành vi có ý nghĩa hơn đối với bài toán quản trị quan hệ khách hàng.

CHƯƠNG 4

XÂY DỰNG MÔ HÌNH PHÂN CỤM KHÁCH HÀNG

4.1. Tạo đặc trưng

4.1.1. Tại sao cần mở rộng thêm đặc trưng

Nếu chỉ sử dụng bộ chỉ số RFM, mô hình sẽ nhìn khách hàng chủ yếu qua ba khía cạnh là mức độ gần đây, tần suất và tổng giá trị chi tiêu. Cách tiếp cận này có ưu điểm là ngắn gọn, dễ tính toán và phù hợp cho các bài toán phân khúc ở mức cơ bản. Tuy nhiên, khi áp dụng trực tiếp vào dữ liệu bán lẻ nhiều dòng giao dịch, RFM vẫn chưa phản ánh đầy đủ cấu trúc hành vi mua sắm.

Thứ nhất, RFM không cho thấy độ đa dạng sản phẩm mà khách hàng đã mua. Hai khách hàng có cùng tổng chi tiêu có thể khác nhau rất lớn nếu một người chỉ mua lặp lại một vài mã hàng còn người kia mua trải rộng trên nhiều loại sản phẩm. Thứ hai, RFM không mô tả được cấu trúc của từng hóa đơn, chẳng hạn mỗi lần mua khách hàng thường chọn bao nhiêu loại hàng hay giá trị trung bình của một hóa đơn là bao nhiêu. Thứ ba, RFM khó tách được nhóm mua ít lần nhưng mỗi lần mua nhiều với nhóm mua đều hơn, lặp lại hơn và có chiều sâu tương tác cao hơn.

Vì vậy, nhóm mở rộng dữ liệu đầu vào thành bộ feature ở cấp khách hàng nhằm mô tả hành vi mua sắm trên nhiều chiều hơn. Cách làm này giúp K-Means có nhiều tín hiệu hơn để tách các nhóm khách hàng thực sự khác biệt về hành vi, thay vì chỉ khác nhau ở tổng chi tiêu cuối cùng.

4.1.2. Bộ đặc trưng ở cấp khách hàng

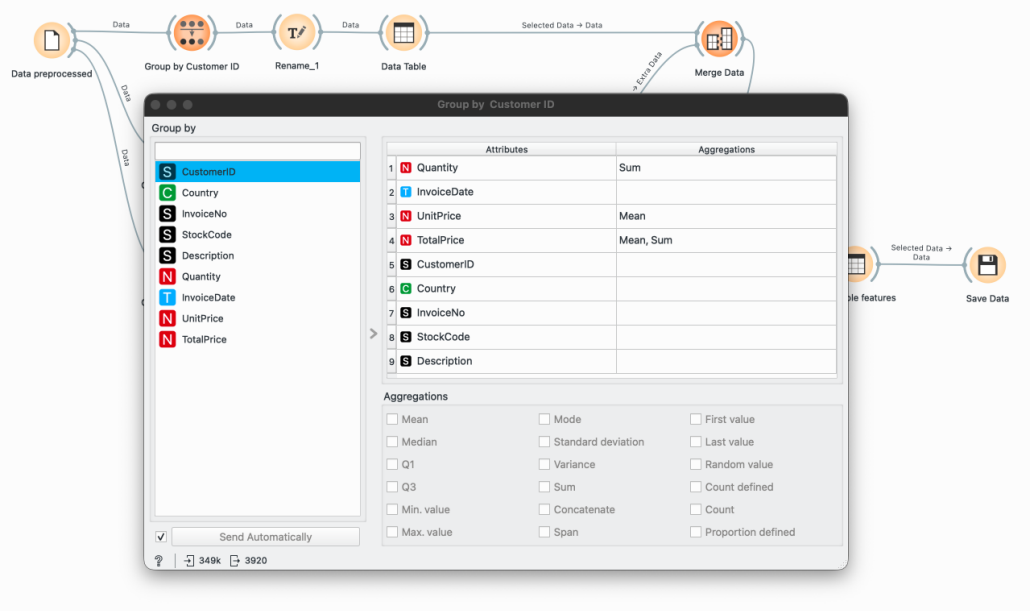
Theo tài liệu tham khảo chính và sau khi đối chiếu lại với dữ liệu thực tế, bộ đặc trưng mục tiêu của đề tài gồm 16 feature, chia thành bốn nhóm lớn: khối lượng và giá trị mua hàng, hành vi theo hóa đơn, hành vi theo sản phẩm và mức độ lặp lại của từng sản phẩm. Bảng dưới đây tóm tắt vai trò của từng feature.

Bảng 4.1: Bộ 16 đặc trưng khách hàng sử dụng cho phân cụm

STT	Tên feature	Nhóm	Ý nghĩa
1	Sum_Quantity	Tổng quan	Tổng số lượng sản phẩm khách hàng đã mua.
2	Mean_UnitPrice	Tổng quan	Mức giá trung bình của các sản phẩm đã chọn.
3	Mean_TotalPrice	Tổng quan	Giá trị trung bình của một dòng giao dịch.
4	Sum_TotalPrice	Tổng quan	Tổng chi tiêu của khách hàng.
5	Count_Invoice	Tổng quan	Số hóa đơn duy nhất của khách hàng.
6	Count_Stock	Tổng quan	Số mã sản phẩm duy nhất khách hàng từng mua.
7	Mean_InvoiceCountPerStock	Theo sản phẩm	Tần suất xuất hiện trung bình của một sản phẩm trong lịch sử mua của khách hàng.
8	Mean_StockCountPerInvoice	Theo hóa đơn	Số loại sản phẩm trung bình trong một hóa đơn.
9	Mean_UnitPriceMeanPerInvoice	Theo hóa đơn	Giá đơn vị trung bình xét trên từng hóa đơn.
10	Mean_QuantitySumPerInvoice	Theo hóa đơn	Số lượng trung bình trong một hóa đơn.
11	Mean_TotalPriceMeanPerInvoice	Theo hóa đơn	Giá trị trung bình của từng dòng giao dịch trong một hóa đơn.
12	Mean_TotalPriceSumPerInvoice	Theo hóa đơn	Tổng giá trị trung bình của một hóa đơn.
13	Mean_UnitPriceMeanPerStock	Theo sản phẩm	Mức giá trung bình của từng loại sản phẩm mà khách hàng mua.
14	Mean_QuantitySumPerStock	Theo sản phẩm	Số lượng trung bình đã mua trên mỗi sản phẩm.
15	Mean_TotalPriceMeanPerStock	Theo sản phẩm	Chi tiêu trung bình trên mỗi sản phẩm.
16	Mean_TotalPriceSumPerStock	Theo sản phẩm	Tổng chi tiêu trung bình theo từng loại sản phẩm.

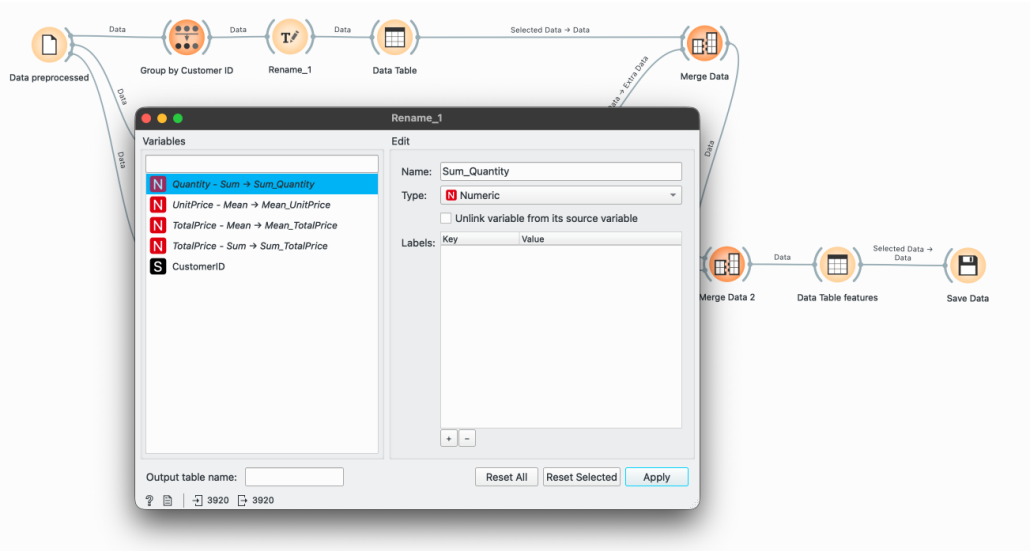
4.1.3. Thực hiện trong Orange

Quy trình tạo feature được triển khai bằng nhiều nhánh `Group By`. Mỗi nhánh bắt đầu từ dữ liệu giao dịch đã được làm sạch, sau đó tổng hợp lên cấp khách hàng theo mục tiêu khác nhau. Đây là bước biến đổi quan trọng nhất của đề tài, vì nó chuyển bài toán từ không gian "dòng giao dịch" sang không gian "hành vi khách hàng".



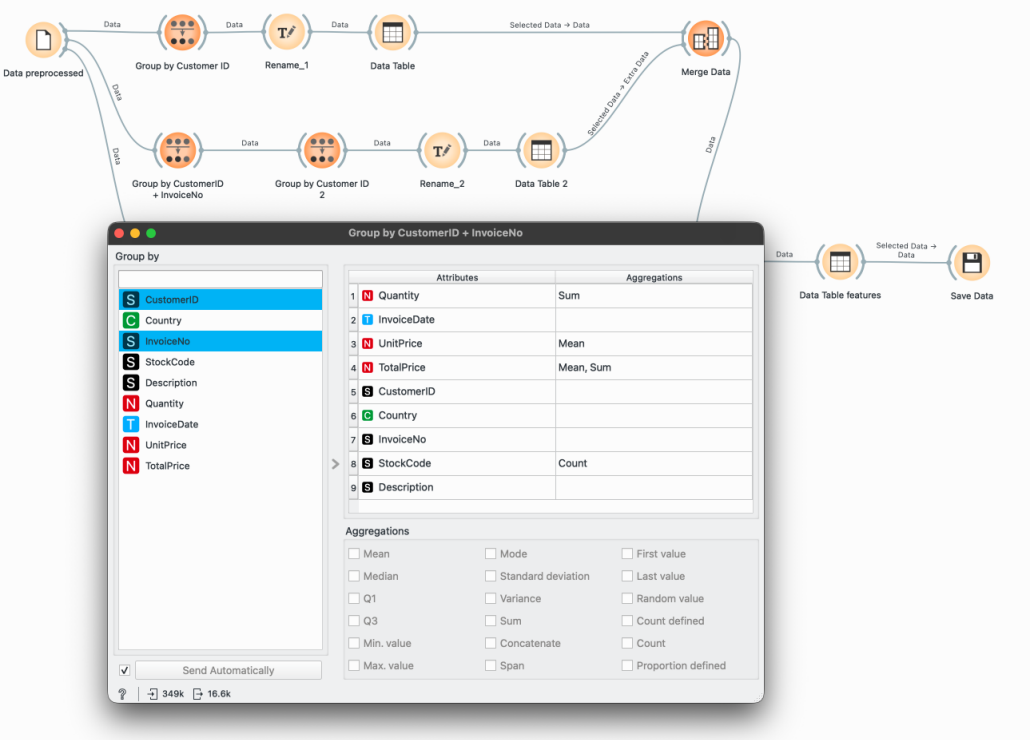
Hình 4.1: Tạo nhóm feature cơ bản bằng `Group By` theo `CustomerID`

Ở nhánh thứ nhất, nhóm gom theo `CustomerID` để tạo ra các chỉ số cơ bản như tổng số lượng, giá đơn vị trung bình, tổng chi tiêu, số hóa đơn và số mã hàng đã mua. Đây là lớp đặc trưng nền, phản ánh quy mô tiêu dùng tổng quát của mỗi khách hàng.



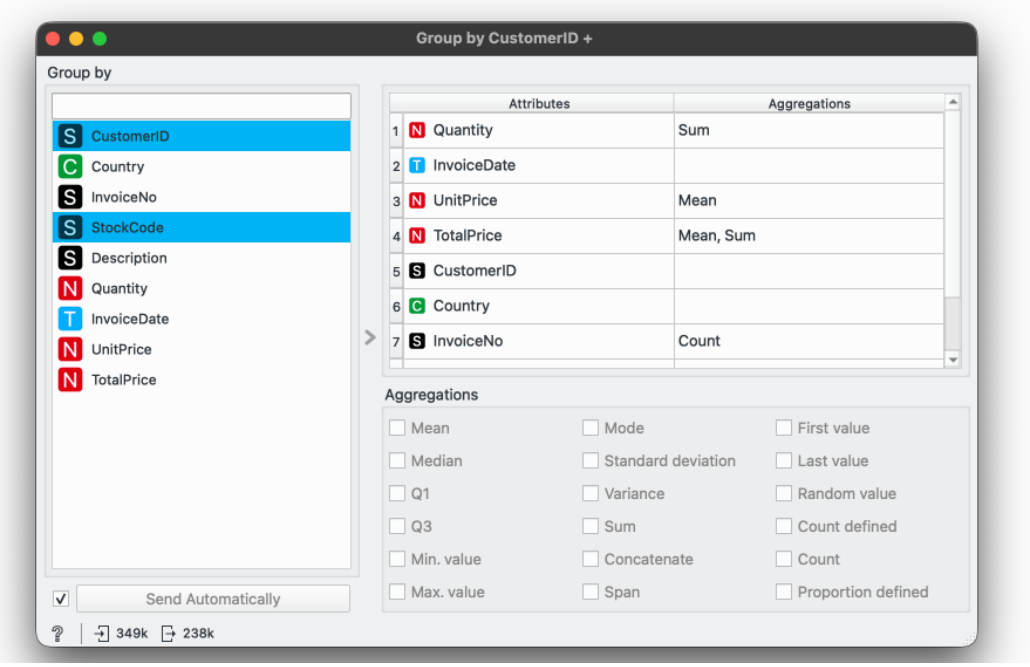
Hình 4.2: Đổi tên các feature bằng widget `Edit Domain`

Sau khi tổng hợp, nhóm sử dụng `Edit Domain` để đổi tên các cột thành tên có ý nghĩa hơn cho báo cáo. Việc chuẩn hóa tên cột ngay trong Orange giúp workflow về sau dễ đọc hơn, đồng thời tránh nhầm lẫn giữa các chỉ số tổng, trung bình theo hóa đơn và trung bình theo sản phẩm.



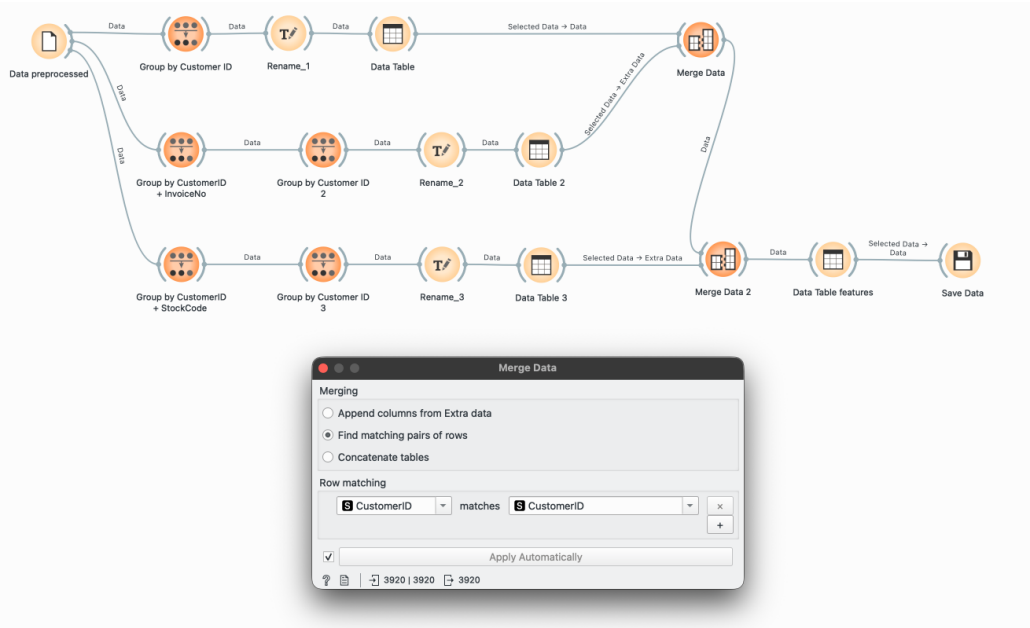
Hình 4.3: Tạo feature theo hóa đơn trong Orange

Nhánh thứ hai nhóm theo cặp `CustomerID + InvoiceNo`. Bước này nhằm mô tả hành vi trong từng hóa đơn, chẳng hạn trung bình mỗi hóa đơn có bao nhiêu loại sản phẩm, bao nhiêu sản phẩm và giá trị trung bình của một lần mua. Khi tổng hợp ngược trở lại cấp khách hàng, các feature này giúp mô hình nhìn được "hình dạng" giao dịch chứ không chỉ nhìn tổng chi tiêu.



Hình 4.4: Tạo feature theo sản phẩm trong Orange

Nhánh thứ ba nhóm theo cặp `CustomerID + StockCode`. Mục tiêu là mô tả sở thích và mức độ lặp lại theo từng mã sản phẩm. Đây là lớp thông tin rất quan trọng để phân biệt nhóm mua đa dạng với nhóm mua lặp lại mạnh trên một tập sản phẩm hẹp.

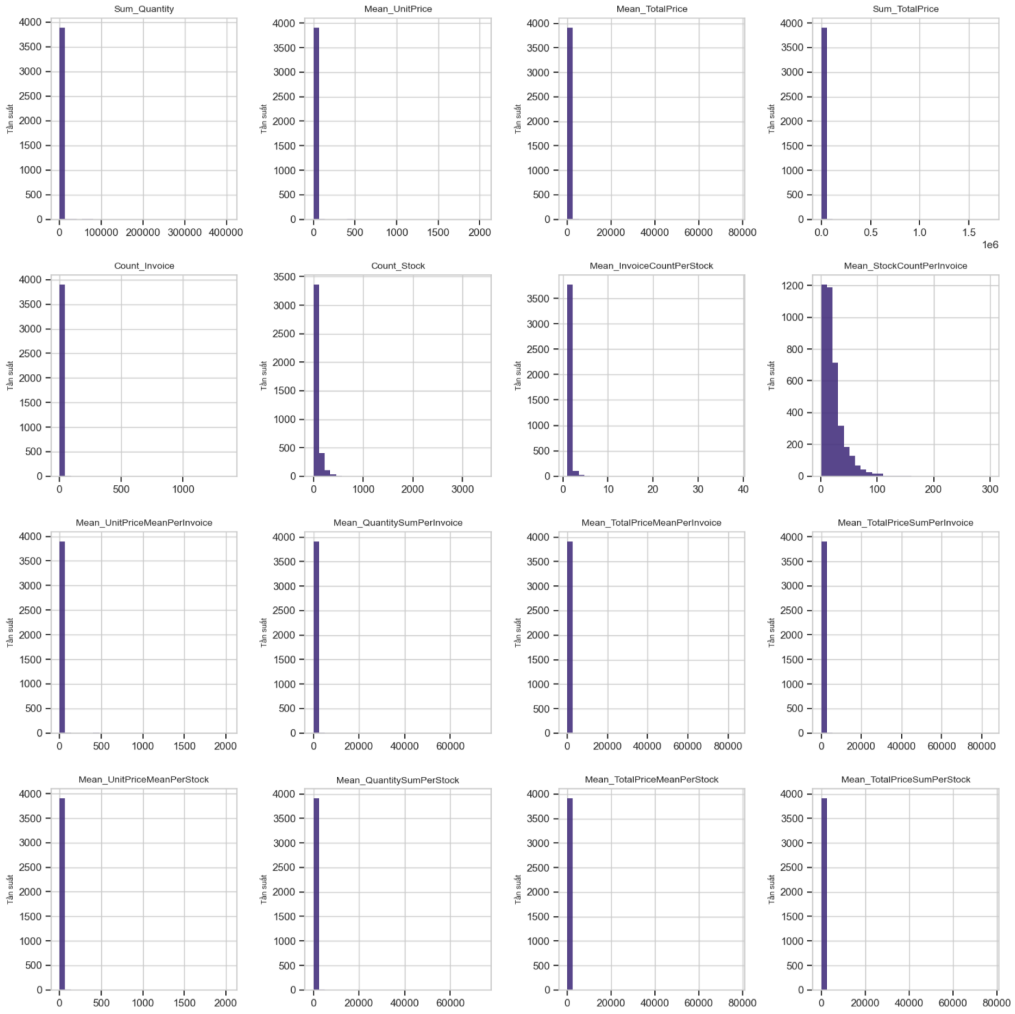


Hình 4.5: Hợp nhất các bảng đặc trưng bằng `Merge Data`

Cuối cùng, các bảng feature trung gian được hợp nhất bằng `Merge Data` theo khóa `CustomerID`. Kết quả đầu ra là một bảng dữ liệu ở cấp khách hàng, trong đó mỗi hàng tương ứng với một khách hàng duy nhất và mỗi cột là một đặc trưng hành vi. So với cách dùng trực tiếp RFM, bộ dữ liệu đầu ra giữ lại được nhiều tín hiệu giàu ý nghĩa hơn cho bài toán phân cụm.

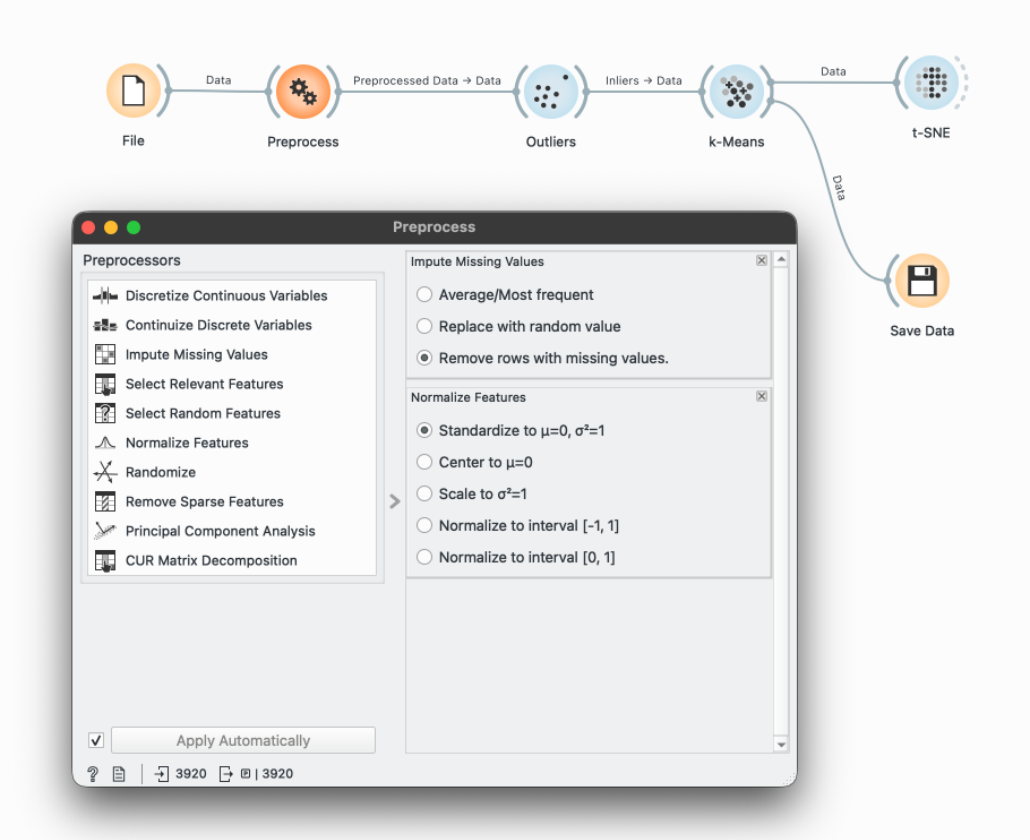
4.1.4. Chuẩn hóa và loại bỏ outlier

Do dữ liệu bán lẻ thường có phân phối lệch phải mạnh, các đặc trưng sau khi tổng hợp cũng có xu hướng tập trung ở vùng thấp và kéo dài đuôi sang bên phải. Điều này khiến K-Means dễ bị chi phối bởi một số khách hàng có giá trị cực lớn nếu nhóm không chuẩn hóa dữ liệu trước. Vì vậy, nhóm quan sát phân phối của feature bằng histogram rồi chuẩn hóa về cùng thang đo trước khi mô hình hóa.



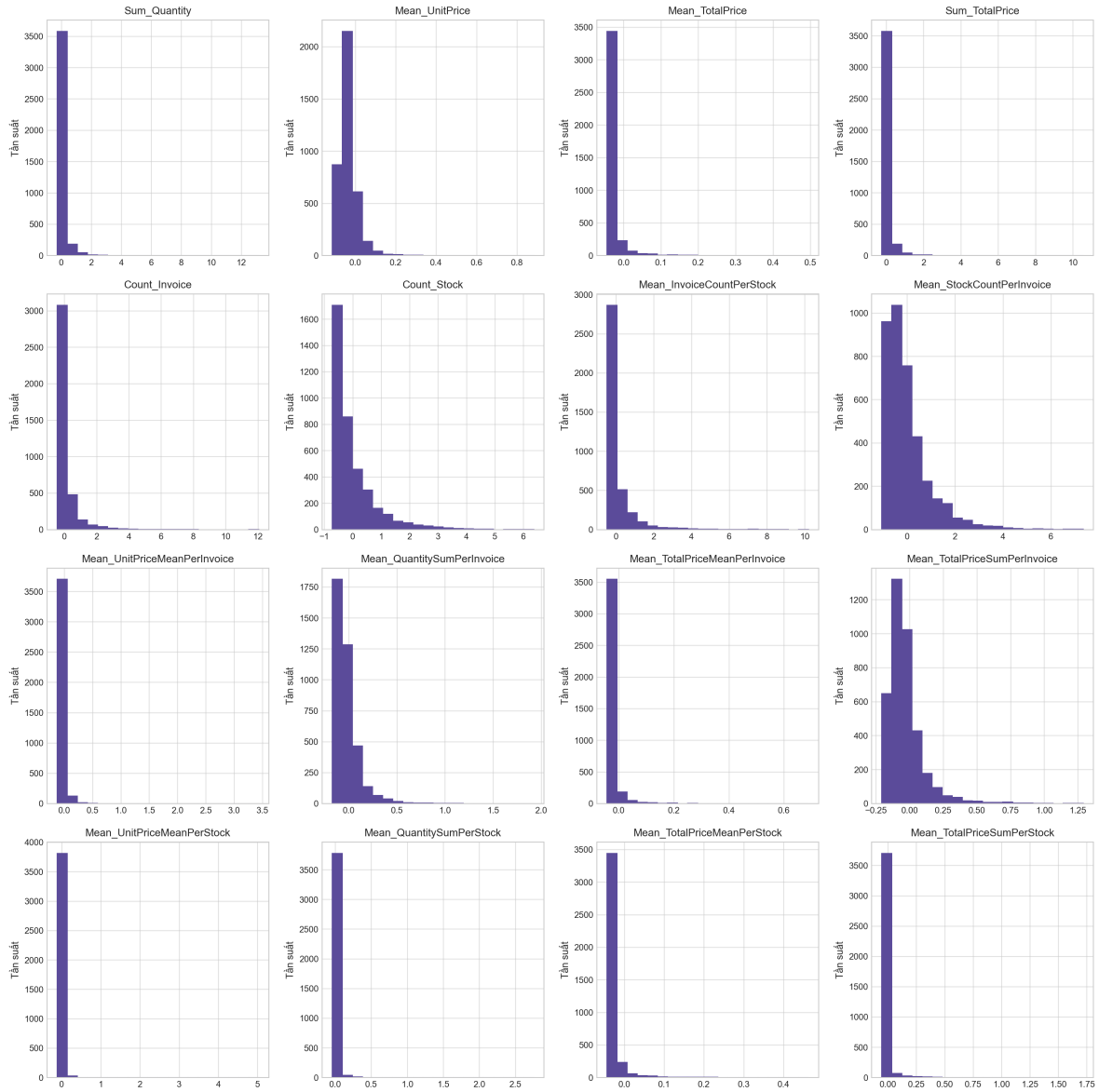
Hình 4.6: Phân phối của các feature trước khi biến đổi hoặc chuẩn hóa

Trong Orange, bước này được thực hiện bằng widget `Normalize`. Sau khi nối bộ dữ liệu 16 feature vào widget, nhóm chuẩn hóa toàn bộ các biến số trước khi đưa sang bước phát hiện outlier và phân cụm. Việc chuẩn hóa giúp các feature về cùng mặt bằng, tránh trường hợp các biến có biên độ lớn như tổng chi tiêu hoặc tổng số lượng lấn át các biến còn lại.



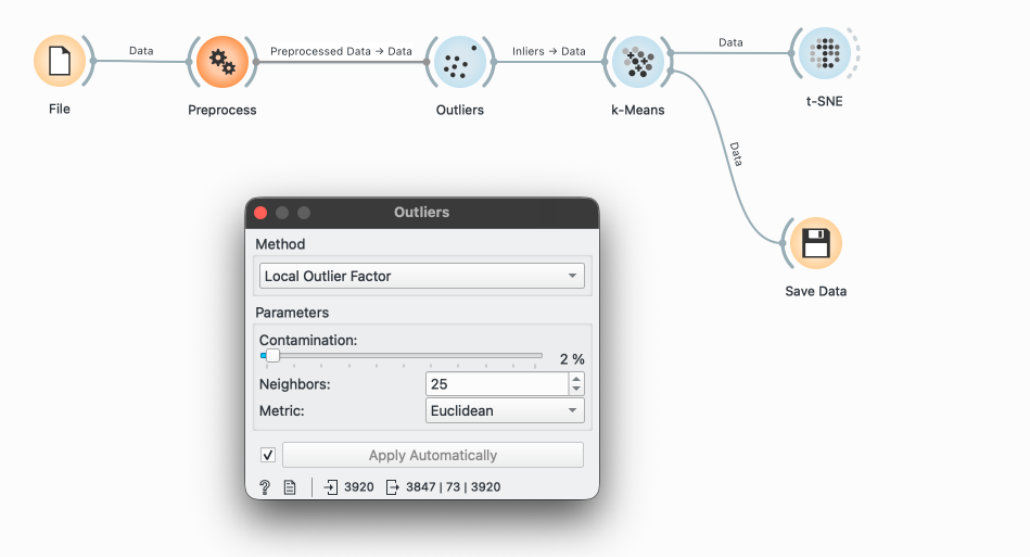
Hình 4.7: Thiết lập widget `Normalize` để chuẩn hóa bộ 16 feature trong Orange

Sau chuẩn hóa, phần lớn đặc trưng đã được đưa về vùng quanh 0, giúp việc tính khoảng cách Euclid giữa các khách hàng trở nên cân bằng hơn. Hình dưới đây cho thấy phân phối sau chuẩn hóa vẫn còn một số đuôi kéo dài, nhưng mức độ chênh lệch giữa các biến đã giảm rõ rệt.



Hình 4.8: Phân phối của các feature sau khi chuẩn hóa

Sau bước chuẩn hóa, nhóm đưa dữ liệu qua widget `Outliers` và chọn phương pháp LOF để nhận diện các khách hàng quá khác biệt so với mặt bằng chung. Trong workflow này, Orange tạo ra hai nhánh kết quả: một nhánh chứa toàn bộ dữ liệu đã gắn nhãn outlier để kiểm tra, và một nhánh `Inliers` dùng làm đầu vào chính cho K-Means.



Hình 4.9: Thiết lập widget `Outliers` bằng LOF trước khi chạy K-Means

Quyết định thêm bước loại outlier xuất phát từ thực nghiệm trên chính dữ liệu của đề tài: nếu giữ nguyên toàn bộ khách hàng cực trị, K-Means có xu hướng tách ra những cụm rất nhỏ chỉ để bao lấy các điểm xa tâm, làm kết quả chung kém ổn định hơn. Khi đưa qua LOF trước, cấu trúc phân khúc tổng quát trở nên rõ hơn và việc diễn giải các cụm lớn cũng nhất quán hơn.

4.2. Mô hình phân cụm khách hàng

4.2.1. Thiết lập thực nghiệm

Sau khi có bộ dữ liệu khách hàng ở dạng số đã chuẩn hóa và đã loại outlier, nhóm tiến hành mô hình hóa bằng K-Means. Trong phạm vi bài này, hai cấu hình $k = 3$ và $k = 4$ được giữ lại để so sánh song song. Mỗi cấu hình được đánh giá qua ba lớp thông tin:

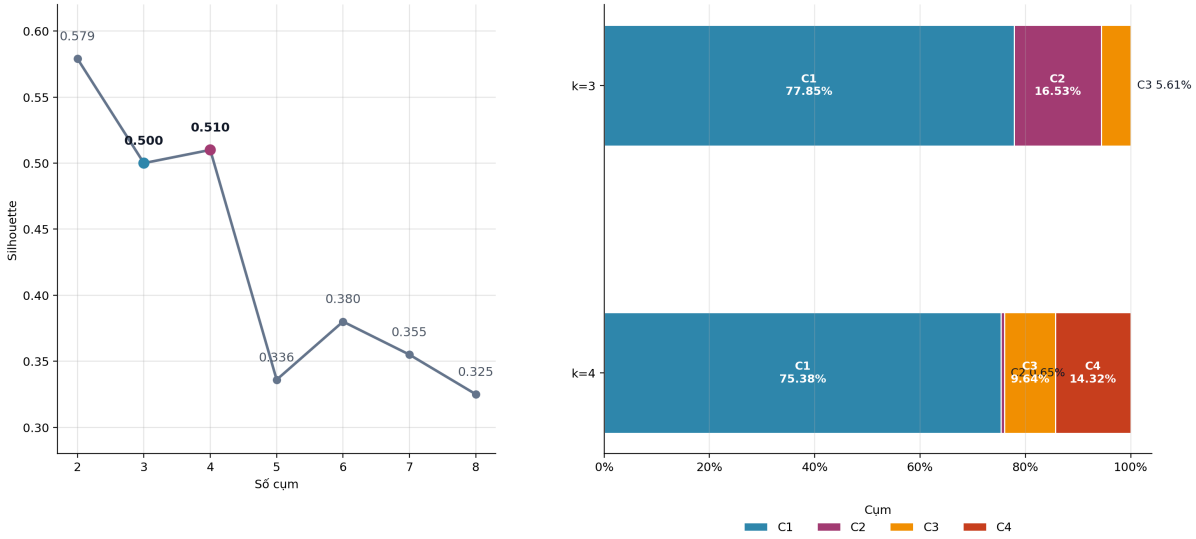
- chất lượng phân cụm định lượng thông qua silhouette trung bình;
- mức độ cân bằng quy mô giữa các cụm;
- khả năng diễn giải ý nghĩa kinh doanh của từng nhóm khách hàng.

Điều này có nghĩa là việc chọn số cụm không chỉ dựa trên một chỉ số duy nhất. Nếu một phương án cho silhouette nhỉnh hơn nhưng làm xuất hiện một cụm quá nhỏ và khó hành động, phương án đó chưa chắc đã là lựa chọn tốt hơn về mặt quản trị.

4.2.2. So sánh lựa chọn số cụm giữa k=3 và k=4

Trong Orange, nhóm sử dụng widget `k-Means` ở chế độ tối ưu số cụm để kiểm tra nhiều phương án ứng viên. Kết quả dò từ $k = 2$ đến $k = 8$ cho silhouette trung bình lần lượt là 0.579, 0.500,

0.510, 0.336, 0.380, 0.355 và 0.325. Mặc dù $k = 2$ cho giá trị cao nhất, cách chia này chỉ tạo ra hai nhóm quá rộng và chưa đủ chiều sâu để phục vụ diễn giải hành vi mua sắm. Vì vậy, nhóm giữ lại hai cấu hình $k = 3$ và $k = 4$ để phân tích sâu hơn.



Hình 4.10: Diễn biến silhouette theo số cụm và tỷ trọng các cụm ở hai phương án được giữ lại

Kết quả cho thấy:

- $k = 3$ là phương án đầu tiên giữ được silhouette ở mức 0.500 và đồng thời tạo ra ba nhóm khách hàng có thể diễn giải rõ về mặt nghiệp vụ;
- $k = 4$ có silhouette nhỉnh hơn một chút, đạt 0.510, nhưng sự cải thiện này không lớn;
- từ $k = 5$ trở đi, silhouette giảm khá mạnh xuống vùng 0.325 đến 0.380, cho thấy cấu trúc phân cụm bắt đầu kém ổn định hơn.

Xét về tỷ trọng cụm, phương án $k = 3$ cho cấu trúc 77.85% – 16.53% – 5.61%, tức là vẫn có một cụm lớn chi phối nhưng hai cụm còn lại đủ rõ để mô tả các nhóm hành vi khác nhau. Trong khi đó, phương án $k = 4$ cho cấu trúc 75.38% – 0.65% – 9.64% – 14.32%, nghĩa là mô hình đã tách riêng một cụm rất nhỏ chỉ gồm 25 khách hàng. Điều này có ích nếu mục tiêu là bóc tách nhóm khách hàng cực trị, nhưng lại làm phương án $k = 4$ kém cân bằng hơn khi dùng làm mô hình phân khúc chính.

Từ góc nhìn thực hành trong Orange, kết quả này cho thấy workflow hiện tại đủ ổn định để tái chạy nhiều lần với các cấu hình số cụm khác nhau. Nhóm có thể giữ nguyên toàn bộ pipeline tiền xử lý và chỉ điều chỉnh tham số của widget `k-Means` để kiểm tra nhanh các kịch bản phân khúc trước khi chọn ra phương án cuối cùng.

4.2.3. Trực quan hóa bằng t-SNE

Trong pipeline hiện tại, nhóm dùng t-SNE để trực quan hóa hình học của cụm sau khi K-Means đã gán nhãn. Khác với PCA thiên về phương sai toàn cục, t-SNE giữ cấu trúc lân cận cục bộ tốt hơn nên phù hợp hơn khi mục tiêu là quan sát mức độ tập trung và giao thoa giữa các nhóm khách hàng trong không gian hai chiều.

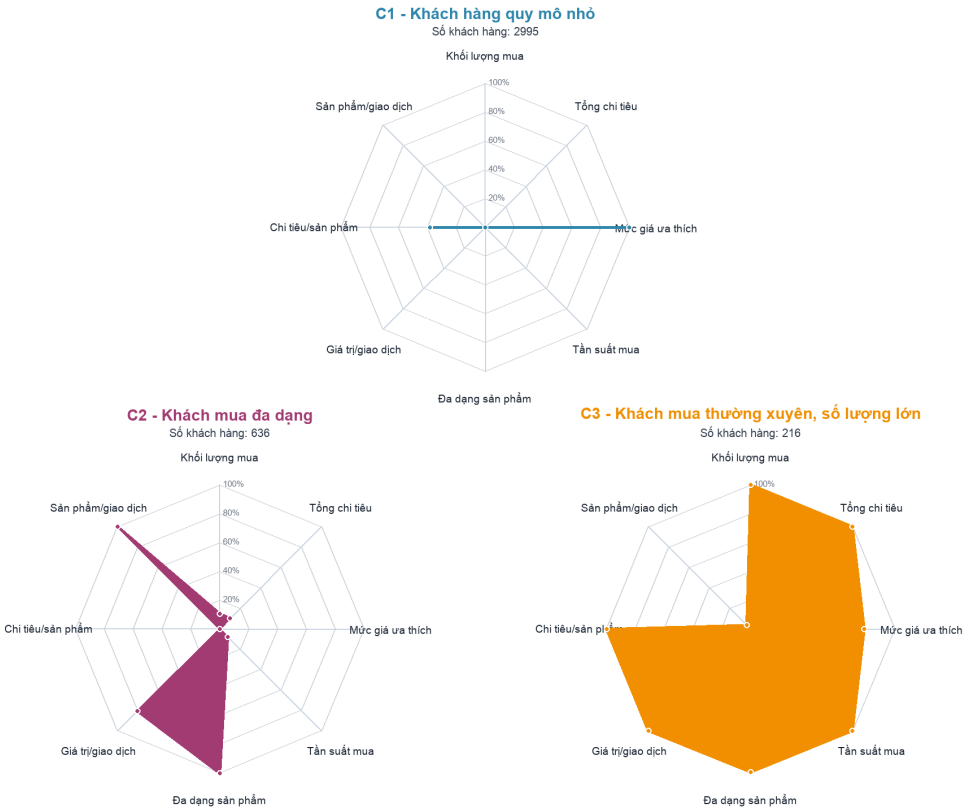
Hình 4.11: Trực quan hóa phân cụm bằng t-SNE với k=3

Hình 4.12: Trực quan hóa phân cụm bằng t-SNE với k=4

Quan sát hai hình trên cho thấy $k = 3$ tạo ra ba vùng hành vi tương đối rõ, trong đó nhóm khách hàng quy mô nhỏ chiếm phần lớn diện tích hình học. Khi chuyển sang $k = 4$, một phần trong nhóm khách hàng có giá trị cao được tách riêng thành một cụm rất nhỏ. Đây là lý do silhouette tăng nhẹ, nhưng cũng là nguyên nhân khiến phương án $k = 4$ khó dùng làm mô hình phân khúc chính nếu mục tiêu là ổn định và dễ triển khai.

4.2.4. Diễn giải kết quả với k=3

Phương án $k = 3$ cho ba cụm với hình dạng khá rõ trên t-SNE và radar chart. Đây là phương án nhóm chọn làm kết quả phân khúc chính vì vừa đủ chi tiết để tạo hành động, vừa tránh được việc sinh ra những cụm quá nhỏ. Trong Orange, sau khi có nhãn cụm từ `k-Means`, nhóm xuất dữ liệu ra bảng dữ liệu có kèm cột `Cluster`, rồi tổng hợp tiếp để tạo hồ sơ trung bình cho từng cụm.



Hình 4.13: Biểu đồ radar hồ sơ cụm khách hàng khi k=3

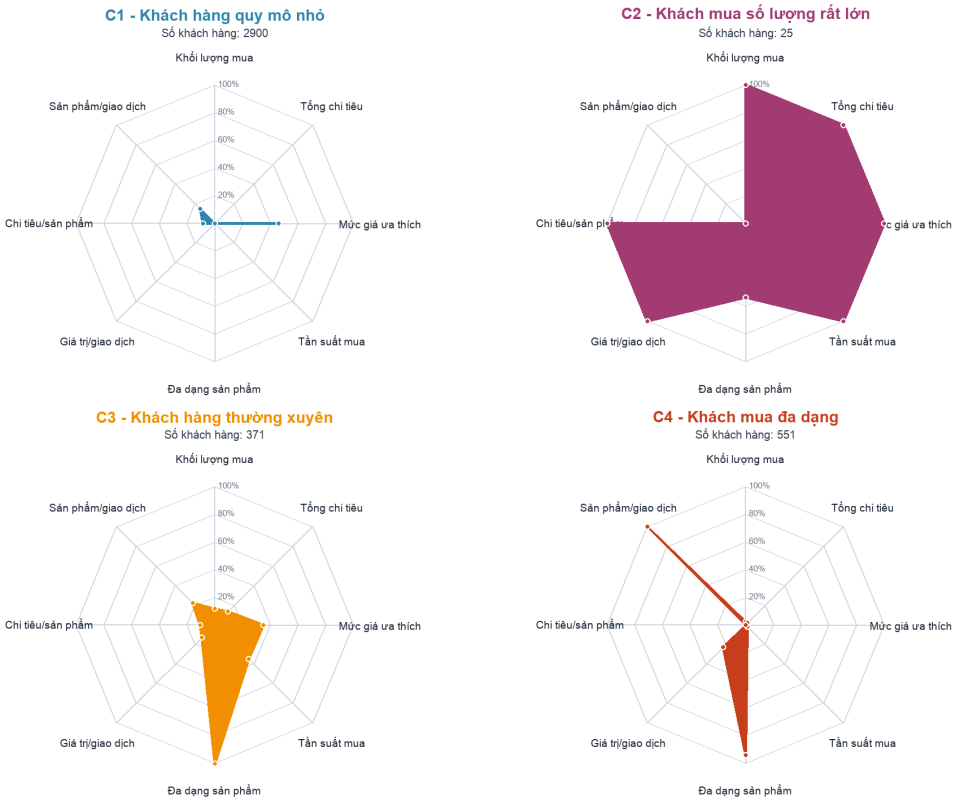
Bảng 4.2: Tóm tắt ý nghĩa kinh doanh của các cụm khi k=3

Cụm	Tỷ trọng	Đặc trưng nổi bật	Diễn giải
C1	77.85%	Hầu hết feature chính đều thấp hơn trung bình, đặc biệt là số hóa đơn và số mã sản phẩm.	Đây là nhóm khách hàng quy mô nhỏ, mua ít và tương tác chưa sâu. Nhóm này phù hợp với các chiến dịch kích hoạt lại, bundle giá thấp và ưu đãi chuyển đổi lần mua tiếp theo.
C2	16.53%	Độ đa dạng sản phẩm và số sản phẩm trong mỗi hóa đơn cao hơn rõ rệt, nhưng tổng chi tiêu chỉ quanh mức trung bình.	Đây là nhóm khách mua đa dạng. Họ có xu hướng khám phá danh mục rộng hơn, phù hợp với chiến dịch cross-sell, gợi ý sản phẩm bổ sung và giới thiệu bộ sưu tập mới.
C3	5.61%	Tổng số lượng, tổng chi tiêu, số hóa đơn và mức độ lặp lại theo sản phẩm đều cao nhất. Giá đơn vị không nổi bật.	Đây là nhóm mua thường xuyên với khối lượng lớn, gần với nhóm khách mua sỉ hoặc khách hàng có giá trị vòng đời cao. Điểm cần lưu ý là sức mạnh của cụm này đến từ tần suất và số lượng, chứ không phải từ mức giá đơn vị cao.

Điểm đáng chú ý trong phương án $k = 3$ là cụm C3 không phải nhóm "premium" theo nghĩa truyền thống. `Mean_UnitPrice` của nhóm này chỉ gần mức trung bình, trong khi tổng chi tiêu cao đến từ việc mua nhiều hơn và mua lặp lại nhiều hơn. Điều đó cho thấy nếu doanh nghiệp chỉ nhìn vào doanh thu, rất dễ gắn nhãn sai cụm này thành khách hàng cao cấp, trong khi về bản chất đây là nhóm khách hàng giá trị cao nhờ sức mua và tần suất.

4.2.5. Diễn giải kết quả với k=4

Phương án $k = 4$ tách thêm một cụm rất nhỏ từ nhóm khách hàng giá trị cao. Vì vậy, phương án này hữu ích khi doanh nghiệp muốn quan sát riêng các khách hàng cực trị hoặc khách mua sỉ nổi bật thay vì gộp họ vào một phân khúc lớn hơn. Về mặt thao tác, workflow Orange của phương án này gần giống với $k = 3$, chỉ khác ở tham số số cụm trong widget `k-Means`; phần trực quan và tổng hợp kết quả vẫn giữ nguyên logic.



Hình 4.14: Biểu đồ radar hồ sơ cụm khách hàng khi k=4

Bảng 4.3: Tóm tắt ý nghĩa kinh doanh của các cụm khi k=4

Cụm	Tỷ trọng	Đặc trưng nổi bật	Diễn giải
C1	75.38%	Giữ vai trò nhóm nền với các chỉ số phần lớn dưới trung bình.	Đây vẫn là nhóm khách hàng quy mô nhỏ, tương tự C1 ở phương án $k = 3$.
C2	0.65%	Tổng số lượng, tổng chi tiêu và số hóa đơn rất cao, vượt trội hẳn so với các cụm còn lại.	Đây là nhóm khách mua số lượng rất lớn, có thể xem như cụm cực trị hoặc khách mua sỉ nổi bật. Cụm này có giá trị giám sát cao nhưng quá nhỏ để dùng làm phân khúc đại trà.
C3	9.64%	Tần suất và tổng chi tiêu cao hơn trung bình, nhưng chưa cực trị như C2.	Đây là nhóm khách hàng thường xuyên, có thể là lớp "trung gian" giữa khách phổ thông và khách mua sỉ rất mạnh. Nhóm này phù hợp cho chiến lược nuôi dưỡng lên hạng khách hàng giá trị cao.
C4	14.32%	Độ đa dạng sản phẩm và số sản phẩm trong mỗi hóa đơn là nổi bật nhất.	Đây là nhóm khách mua đa dạng, có khuynh hướng mua theo danh mục rộng và có khả năng phản hồi tốt với cross-sell hoặc giới thiệu sản phẩm mới.

Nhìn trên góc độ phân tích, $k = 4$ cho phép doanh nghiệp nhìn thấy thêm một lớp cấu trúc mà $k = 3$ đã gộp lại. Cụm C2 gồm 25 khách hàng là minh chứng rõ nhất: đây là nhóm rất hiếm nhưng có tác động lớn tới tổng doanh thu. Trong môi trường kinh doanh thực tế, những khách hàng như vậy thường đáng được theo dõi riêng bằng chương trình chăm sóc hoặc chính sách giá đặc thù.

4.2.6. Lựa chọn cấu hình cuối cùng

Sau khi đối chiếu cả silhouette, quy mô cụm, hình t-SNE và radar chart, nhóm chọn $k = 3$ là phương án phân khúc chính của báo cáo. Lý do là:

- điểm silhouette của $k = 3$ đã ở mức tốt và không chênh nhiều so với $k = 4$;
- ba cụm của $k = 3$ có quy mô đủ lớn để chuyển thành hành động marketing rõ ràng;
- cách diễn giải của $k = 3$ ổn định hơn, dễ trình bày hơn và phù hợp với mục tiêu học phần.

Tuy nhiên, nhóm vẫn giữ $k = 4$ như một lớp phân tích bổ sung. Trong bối cảnh doanh nghiệp muốn bóc tách riêng nhóm khách hàng cực trị hoặc khách mua sỉ giá trị rất cao, $k = 4$

lại là phương án hữu ích hơn. Nói cách khác, hai cấu hình không loại trừ nhau: $k = 3$ phù hợp cho phân khúc vận hành tổng quát, còn $k = 4$ phù hợp cho phân tích đào sâu và quản trị nhóm khách hàng hiếm nhưng quan trọng.

CHƯƠNG 5

ĐỀ XUẤT PHƯƠNG PHÁP VÀ HƯỚNG MỞ RỘNG

5.1. Định hướng sử dụng kết quả phân cụm trong thực tế

Giá trị lớn nhất của bài toán phân cụm không nằm ở việc gắn nhãn `C1`, `C2`, `C3` hay `C4`, mà nằm ở khả năng chuyển các cụm này thành hành động quản trị cụ thể. Nếu không đi tiếp đến bước này, kết quả mô hình chỉ dừng ở mức mô tả dữ liệu. Vì vậy, trên cơ sở phương án $k = 3$ được chọn làm cấu hình chính, nhóm đề xuất một số hướng ứng dụng thực tế như sau.

5.1.1. Nhóm khách hàng quy mô nhỏ

Đây là nhóm chiếm tỷ trọng lớn nhất trong dữ liệu nhưng phần lớn có giá trị giao dịch và độ sâu tương tác thấp. Về bản chất, đây là nhóm nền của hệ khách hàng: họ đông nhưng chưa đóng góp mạnh về doanh thu hoặc mức độ mua lặp lại.

Đối với nhóm này, mục tiêu không nên là bán sản phẩm giá trị cao ngay lập tức, mà nên tập trung vào việc tăng số lần mua lặp lại và nâng dần độ phong phú của giỏ hàng. Một số hành động phù hợp gồm:

- gửi mã giảm giá cho đơn hàng kế tiếp nhằm rút ngắn chu kỳ mua;
- xây dựng các combo giá thấp hoặc gói sản phẩm phổ biến để kích thích mua nhiều hơn trong một hóa đơn;
- triển khai email hoặc thông báo tái kích hoạt đối với khách hàng có dấu hiệu không quay lại sau một khoảng thời gian nhất định.

5.1.2. Nhóm khách mua đa dạng

Nhóm này không nhất thiết là nhóm chi tiêu cao nhất, nhưng nổi bật ở số loại sản phẩm đã mua và số sản phẩm xuất hiện trong mỗi hóa đơn. Điều đó cho thấy họ có xu hướng khám phá danh mục, dễ phản hồi với các đề xuất sản phẩm mới và có tiềm năng cross-sell tốt.

Với nhóm khách hàng này, doanh nghiệp nên ưu tiên các chiến lược:

- gợi ý sản phẩm bổ sung dựa trên lịch sử danh mục đã mua;

- triển khai bộ sưu tập theo chủ đề hoặc chiến dịch giới thiệu sản phẩm mới;
- dùng nội dung cá nhân hóa theo danh mục quan tâm thay vì chỉ giảm giá đại trà.

Nhìn từ góc độ quản trị, đây là nhóm rất đáng đầu tư cho hoạt động tăng chiều sâu giỏ hàng, vì họ đã cho thấy xu hướng mở rộng sở thích chứ không bị bó hẹp ở một vài mã hàng nhất định.

5.1.3. Nhóm mua thường xuyên, số lượng lớn

Ở phương án $k = 3$, đây là nhóm khách hàng có giá trị cao nhất nhờ tần suất mua và khối lượng mua lớn. Điểm đáng chú ý là mức giá đơn vị của họ không vượt trội, nên doanh thu của nhóm này đến từ hành vi mua đều, mua lặp lại và mua nhiều hơn, chứ không đến từ việc chọn các sản phẩm đắt tiền.

Điều đó gợi ý rằng chiến lược phù hợp với nhóm này không phải là khuyến mãi ngắn hạn đơn lẻ, mà là:

- chính sách chiết khấu theo bậc hoặc theo giá trị lũy kế;
- ưu đãi vận chuyển, ưu đãi thành viên hoặc hỗ trợ đặt hàng định kỳ;
- chương trình chăm sóc riêng cho khách hàng doanh nghiệp hoặc khách mua sỉ nếu doanh nghiệp muốn phát triển theo chiều B2B.

Về mặt kinh doanh, đây là nhóm cần được bảo vệ nhiều nhất vì chỉ cần một tỷ lệ nhỏ rời bỏ cũng có thể tạo ra tác động lớn tới doanh thu tổng thể.

5.2. Giá trị bổ sung của phương án k=4

Mặc dù nhóm chọn $k = 3$ là phương án phân khúc chính, kết quả $k = 4$ vẫn mang lại một góc nhìn quan trọng: nó tách riêng được một cụm rất nhỏ gồm các khách hàng cực trị về số lượng mua và tổng chi tiêu. Trong thực tế, đây có thể là:

- khách mua sỉ có chu kỳ nhập hàng riêng;
- khách hàng doanh nghiệp đặt đơn khối lượng lớn;
- hoặc một nhóm khách hàng VIP đóng góp doanh thu bất thường.

Nhóm này không phù hợp để xem như phân khúc đại trà, nhưng rất phù hợp để quản trị theo cơ chế "watchlist". Doanh nghiệp có thể theo dõi riêng nhóm này theo tháng hoặc theo quý, kiểm tra biến động đơn hàng và xây dựng chính sách phục vụ ưu tiên nếu thấy đây là nhóm có giá trị chiến lược.

5.3. Hoàn thiện workflow trong Orange

Qua quá trình triển khai, có thể thấy giá trị lớn nhất của đề tài không chỉ nằm ở kết quả phân cụm cuối cùng mà còn nằm ở cách tổ chức quy trình phân tích dữ liệu một cách minh bạch. Vì vậy, một hướng phát triển hợp lý là chuẩn hóa hoàn toàn workflow Orange theo bộ 16 feature ở cấp khách hàng, kèm theo bước `Normalize`, `Outliers`, `k-Means` và `t-SNE`. Khi workflow được hoàn thiện và lưu lại dưới dạng `.ows`, nhóm có thể tái sử dụng cho các kỳ dữ liệu mới mà không cần xây dựng lại từ đầu.

Trong thực tế, phần workflow này có thể được vận hành theo một chuỗi thao tác cố định. Trước hết, người dùng chỉ cần thay file đầu vào ở widget `File`. Sau đó, Orange sẽ tự chạy lại các bước tạo feature, chuẩn hóa, loại outlier và phân cụm. Khi workflow đã ổn định, việc cập nhật kết quả cho kỳ dữ liệu mới sẽ nhanh hơn rất nhiều so với cách xử lý thủ công từng bước rời rạc.

Ngoài ra, nhóm có thể chuẩn hóa thêm phần đầu ra bằng cách luôn xuất ra ba nhóm kết quả sau mỗi lần chạy: bảng khách hàng kèm nhãn cụm, bảng trung bình feature theo cụm và ảnh radar/t-SNE. Khi đó, quá trình cập nhật báo cáo hoặc dashboard sẽ nhanh hơn nhiều. Đây cũng là lợi thế rõ nhất của Orange trong bối cảnh học phần: người thực hiện có thể vừa kiểm tra được trực quan từng bước, vừa duy trì được một pipeline đủ chặt để lặp lại nhiều lần.

5.4. Mở rộng theo hướng dashboard và tự động hóa

Ở giai đoạn tiếp theo, nhóm có thể phát triển một dashboard trực quan để theo dõi phân khúc khách hàng theo thời gian, đồng thời tự động hóa bước nhập dữ liệu mới và cập nhật phân cụm định kỳ. Một hệ thống như vậy sẽ hỗ trợ nhà quản lý quan sát nhanh tỷ trọng từng cụm, xu hướng dịch chuyển giữa các cụm và mức đóng góp doanh thu của từng phân khúc.

Nếu doanh nghiệp có thêm dữ liệu về sản phẩm, mùa vụ hoặc phản hồi khách hàng, nhóm có thể kết hợp phân cụm với các hướng phân tích khác như:

- phân tích giỏ hàng để tìm các cặp sản phẩm hay được mua cùng nhau;
- phân tích churn để dự báo nguy cơ rời bỏ của từng phân khúc;
- dự báo giá trị vòng đời khách hàng để ưu tiên ngân sách chăm sóc cho đúng nhóm;
- so sánh động học cụm theo thời gian nhằm phát hiện khách hàng đang dịch chuyển từ nhóm giá trị thấp sang nhóm giá trị cao hoặc ngược lại.

Những hướng mở rộng này giúp chuyển đề tài từ một mô hình học phần sang một khung hỗ trợ ra quyết định có khả năng ứng dụng cao hơn trong bối cảnh doanh nghiệp thực tế.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Đề tài *Phân cụm khách hàng từ dữ liệu cửa hàng bán lẻ sử dụng mô hình K-Mean* đã xây dựng được một quy trình phân tích tương đối đầy đủ, bắt đầu từ kiểm tra dữ liệu thô, xử lý missing value trong Excel, làm sạch và tổng hợp dữ liệu trong Orange, đến bước xây dựng đặc trưng ở cấp khách hàng và chuẩn bị dữ liệu cho mô hình K-Means. Quá trình này cho thấy dữ liệu giao dịch bán lẻ chứa nhiều tín hiệu quan trọng về hành vi mua sắm, nhưng cần được tổ chức lại một cách hợp lý thì mới có thể khai thác hiệu quả cho bài toán phân cụm.

Thông qua phân tích khám phá dữ liệu, nhóm đã nhận diện được một số đặc điểm nổi bật của dữ liệu như tính mùa vụ trong doanh thu, xu hướng mua hàng tập trung vào giờ hành chính và sự phân hóa mạnh về tần suất cũng như giá trị chi tiêu giữa các khách hàng. Đây là những kết quả quan trọng để làm nền cho bước mô hình hóa. Đồng thời, việc mở rộng bộ feature vượt ra ngoài RFM cho phép mô tả khách hàng đa chiều hơn, phù hợp hơn với mục tiêu phân khúc phục vụ quản trị.

Kết quả thực nghiệm cho thấy phương án $k = 3$ phù hợp hơn để sử dụng làm cấu hình phân khúc chính vì tạo được ba nhóm khách hàng có quy mô hợp lý, dễ diễn giải và thuận tiện cho việc chuyển hóa thành hành động marketing. Đồng thời, phương án $k = 4$ vẫn có giá trị như một lớp phân tích bổ sung nhằm tách riêng nhóm khách hàng cực trị hoặc nhóm mua sỉ nổi bật.

Trong giai đoạn tiếp theo, nhóm có thể tiếp tục mở rộng đề tài theo hướng theo dõi sự dịch chuyển cụm theo thời gian, kết nối workflow Orange với dashboard trực quan và bổ sung thêm các lớp phân tích như dự báo churn hoặc giá trị vòng đời khách hàng. Khi đó, mô hình không chỉ dừng ở mức phân cụm mô tả mà còn có thể trở thành công cụ hỗ trợ ra quyết định thực tế trong môi trường bán lẻ.

TÀI LIỆU THAM KHẢO

[1] Chen, D. (2015). *Online Retail* [Dataset]. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5BW33.

[2] Chen, D., Sain, S. L. and Guo, K. (2012). *Data Mining for the Online Retail Industry: A Case Study of RFM Model-Based Customer Segmentation Using Data Mining*. Journal of Database Marketing & Customer Strategy Management, 19(3), 197–208. DOI: https://doi.org/10.1057/dbm.2012.17.

[3] Demšar, J., Curk, T., Erjavec, A. et al. (2013). *Orange: Data Mining Toolbox in Python*. Journal of Machine Learning Research, 14, 2349–2353. https://jmlr.org/papers/v14/demsar13a.html.

[4] MacQueen, J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297. https://digicoll.lib.berkeley.edu/record/113015.

[5] Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). *Data Clustering: A Review*. ACM Computing Surveys, 31(3), 264–323. DOI: https://doi.org/10.1145/331499.331504.

[6] Rousseeuw, P. J. (1987). *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*. Journal of Computational and Applied Mathematics, 20, 53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

[7] Breunig, M. M., Kriegel, H.-P., Ng, R. T. and Sander, J. (2000). *LOF: Identifying Density-Based Local Outliers*. SIGMOD Record, 29(2), 93–104. https://sigmodrecord.org/2000/06/07/lof-identifying-density-based-local-outliers/.

[8] van der Maaten, L. and Hinton, G. (2008). *Visualizing Data Using t-SNE*. Journal of Machine Learning Research, 9, 2579–2605. https://jmlr.org/papers/v9/vandermaaten08a.html.

[9] Han, J., Kamber, M. and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

[10] Hughes, A. M. (1994). *Strategic Database Marketing*. Probus Publishing.