A white background with blue and orange text

AI-generated content may be incorrect.

**MỤC LỤC**

[Chương 1: Giới thiệu về khoa học dữ liệu và đề tài 4](#_Toc199080505)

[1.1. Tổng quan về khoa học dữ liệu 4](#_Toc199080506)

[1.2. Giới thiệu đề tài nghiên cứu 5](#_Toc199080507)

[Chương 2: Tổng quan chương trình sử dụng và các phương pháp áp dụng 5](#_Toc199080508)

[2.1. Phần mềm và công cụ sử dụng 5](#_Toc199080509)

[2.1.1. Python 5](#_Toc199080510)

[2.1.2. Power BI 6](#_Toc199080511)

[2.2. Các thuật toán và phương pháp sử dụng 8](#_Toc199080512)

[2.2.1. RFM Analysis (Recency, Frequency, Monetary) 8](#_Toc199080513)

[2.2.2. K-Means Clustering (phân khúc khách hàng) 12](#_Toc199080514)

[2.2.3. Recommendation System cơ bản (dựa trên nhóm RFM) 14](#_Toc199080515)

[2.2.4. Các dashboard và thành phần được sử dụng trong Power BI 15](#_Toc199080516)

[Chương 3: Ứng dụng các phương pháp vào bài toán cụ thể 16](#_Toc199080517)

[3.1. Tiền xử lý dữ liệu 16](#_Toc199080518)

[3.1.1. Mô tả dữ liệu 16](#_Toc199080519)

[3.1.2. Tiền xử lý dữ liệu 21](#_Toc199080520)

[3.2. Xây dựng Dashboard với Power BI: 22](#_Toc199080521)

[3.3. Phân tích và phân khúc khách hàng bằng RFM + K-Means 29](#_Toc199080522)

[3.4. Xây dựng hệ thống gợi ý sản phẩm 33](#_Toc199080523)

[Chương 4: Kết quả mô hình 34](#_Toc199080524)

[4.1. Kết quả dashboard trực quan hóa 34](#_Toc199080525)

[4.1.1. Kết quả Dashboard Sales Analysis 34](#_Toc199080526)

[4.1.2. Kết quả Dashboard Product Analysis 36](#_Toc199080527)

[4.1.3. Kết quả Dashboard Customer Analysis 37](#_Toc199080528)

[4.2. Kết quả phân khúc khách hàng (mô tả các nhóm RFM) 39](#_Toc199080529)

[4.3. Kết quả hệ thống gợi ý sản phẩm 41](#_Toc199080530)

[4.4. Đánh giá và đề xuất cải tiến 42](#_Toc199080531)

**DANH MỤC HÌNH ẢNH**

[Hình 1: Khả năng kết nối dữ liệu của Power BI 7](#_Toc199080549)

[Hình 2: Phương thức trực quan hóa trong Power BI 7](#_Toc199080550)

[Hình 3: Ví dụ về phân khúc khách hàng 11](#_Toc199080551)

[Hình 4: Bốn loại chính của hệ thống gợi ý 14](#_Toc199080552)

[Hình 5. Model view của bộ dữ liệu sau khi xử lý bằng Power Query 22](#_Toc199080553)

[Hình 6. Dashboard Sales Analysis 25](#_Toc199080554)

[Hình 7: Dashboard Product Analysis 27](#_Toc199080555)

[Hình 8: Dashboard Customer Analysis 28](#_Toc199080556)

[Hình 9: Phân phối khách hàng mua gần đây 29](#_Toc199080557)

[Hình 10: Phân phối tần suất mua 30](#_Toc199080558)

[Hình 11: Phân phối số tiền chi tiêu của khách hàng 30](#_Toc199080559)

[Hình 12: Box Plot dữ liệu RFM sau khi scale 31](#_Toc199080560)

[Hình 13: Điểm sillhouette phân cụm Kmeans 32](#_Toc199080561)

[Hình 14: Box Plot chỉ số RFM theo các cụm 39](#_Toc199080562)

[Hình 15: Biểu đồ điểm của các chỉ số RFM theo các cụm 40](#_Toc199080563)

[Hình 16: Kết quả danh sách sản phẩm gợi ý 42](#_Toc199080564)

**DANH MỤC BẢNG BIỂU**

[Bảng 1: Cấu trúc bảng distribution\_centers 16](#_Toc199080565)

[Bảng 2: Cấu trúc bảng events 17](#_Toc199080566)

[Bảng 3: Cấu trúc bảng inventory\_items 18](#_Toc199080567)

[Bảng 4: Cấu trúc bảng order\_items 19](#_Toc199080568)

[Bảng 5: Cấu trúc bảng orders 19](#_Toc199080569)

[Bảng 6: Cấu trúc bảng products 20](#_Toc199080570)

[Bảng 7: Cấu trúc bảng users 21](#_Toc199080571)

[Bảng 8: bảng Pareto Analysis 23](#_Toc199080572)

[Bảng 9: Bảng tổng hợp các chỉ số 24](#_Toc199080573)

[Bảng 10: Kết quả Dashboard Sales Analysis 34](#_Toc199080574)

[Bảng 11: Doanh thu theo giới tính 35](#_Toc199080575)

[Bảng 12: Doanh thu theo sản phẩm 36](#_Toc199080576)

[Bảng 13: Số đơn hàng và doanh thu trung bình mỗi khách hang f 37](#_Toc199080577)

[Bảng 14: Thời gian, thời điểm mua hàng 38](#_Toc199080578)

# Chương 1: Giới thiệu về khoa học dữ liệu và đề tài

## 1.1. Tổng quan về khoa học dữ liệu

Trong thời đại dữ liệu số phát triển mạnh mẽ, khoa học dữ liệu (Data Science) đã trở thành một lĩnh vực quan trọng, đóng vai trò then chốt trong việc hỗ trợ ra quyết định dựa trên dữ liệu. Khoa học dữ liệu là sự kết hợp liên ngành giữa thống kê, học máy, khai phá dữ liệu (data mining), và hiểu biết nghiệp vụ nhằm khám phá các mẫu ẩn và tạo ra giá trị từ dữ liệu thô.

Theo Dhar (2013), khoa học dữ liệu là "sự mở rộng của phân tích dữ liệu truyền thống, tích hợp các kỹ thuật học máy và xử lý dữ liệu lớn để giải quyết các vấn đề có độ phức tạp cao"Bên cạnh việc xây dựng mô hình dự đoán, các nhà khoa học dữ liệu còn có nhiệm vụ trực quan hóa dữ liệu, trình bày kết quả một cách dễ hiểu và hỗ trợ người ra quyết định.

Một quy trình khoa học dữ liệu tiêu biểu bao gồm các bước: thu thập dữ liệu, xử lý và làm sạch dữ liệu, phân tích thăm dò (exploratory data analysis), xây dựng mô hình, đánh giá mô hình và triển khai. Theo Provost & Fawcett (2013), điều cốt lõi trong khoa học dữ liệu không chỉ nằm ở thuật toán, mà còn ở khả năng áp dụng chúng vào các bài toán thực tiễn để hỗ trợ ra quyết định kinh doanh hiệu quả.

Trong bối cảnh doanh nghiệp, khoa học dữ liệu giúp phân tích hành vi khách hàng, đo lường hiệu quả hoạt động, tối ưu hóa quy trình kinh doanh và dự đoán xu hướng thị trường. Một nghiên cứu của McKinsey (Manyika et al., 2011) chỉ ra rằng việc áp dụng phân tích dữ liệu quy mô lớn có thể giúp doanh nghiệp tăng năng suất, cải thiện dịch vụ khách hàng và nâng cao khả năng cạnh tranh.

Sự phát triển của các công cụ và nền tảng như Python, R, SQL, Power BI, Tableau... cũng góp phần thúc đẩy việc ứng dụng khoa học dữ liệu trong thực tiễn, đặc biệt trong lĩnh vực Business Intelligence – nơi dữ liệu được chuyển hóa thành thông tin hữu ích cho việc ra quyết định chiến lược.

## 1.2. Giới thiệu đề tài nghiên cứu

Đề tài nghiên cứu tập trung vào việc xây dựng một hệ thống phân tích và hỗ trợ ra quyết định trong kinh doanh thông qua ba mục tiêu chính.

* Xây dựng một dashboard trực quan nhằm theo dõi doanh thu, sản phẩm được tiêu thụ và hành vi khách hàng.
* Thực hiện phân khúc khách hàng dựa trên mô hình RFM (Recency, Frequency, Monetary) để xác định giá trị và mức độ tương tác của từng nhóm khách hàng.
* Phát triển hệ thống gợi ý sản phẩm dựa trên đặc điểm của từng nhóm khách hàng, nhằm cá nhân hóa trải nghiệm mua sắm và gia tăng doanh số bán hàng.

Phạm vi nghiên cứu sử dụng bộ dữ liệu công khai từ TheLook — một trang thương mại điện tử mô phỏng do nhóm Looker phát triển. Dữ liệu bao gồm thông tin về khách hàng, sản phẩm, đơn hàng, vận chuyển, sự kiện web và chiến dịch tiếp thị.

# Chương 2: Tổng quan chương trình sử dụng và các phương pháp áp dụng

## 2.1. Phần mềm và công cụ sử dụng

### 2.1.1. Python

Python đang trở thành công cụ không thể thiếu trong kho vũ khí của các nhà phân tích dữ liệu, nhà phân tích kinh doanh cũng như các kỹ sư dữ liệu. Python có lợi thế hơn so với Excel khi không bị giới hạn về số dòng và hiệu suất, Python có thể xử lý dữ liệu hàng triệu dòng với tốc độ cao, đồng thời dễ dàng tích hợp với các nền tảng lưu trữ đám mây, cơ sở dữ liệu lớn và các API dịch vụ.

Python nổi bật với kho thư viện phong phú và mạnh mẽ như:

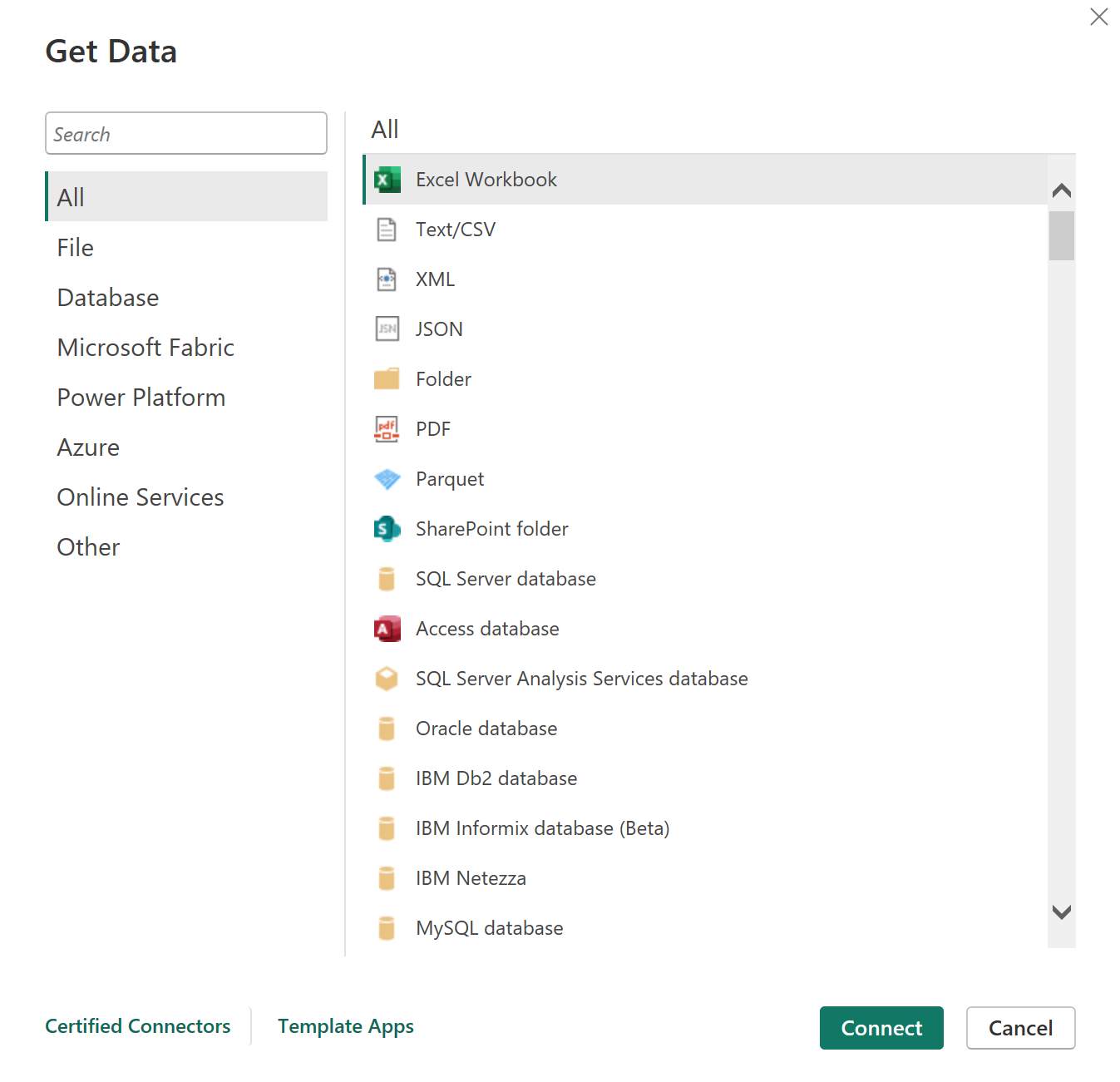
* pandas – để xử lý và phân tích dữ liệu dạng bảng,
* matplotlib, seaborn, plotly – để trực quan hóa dữ liệu sinh động và tùy biến,
* scikit-learn – để triển khai mô hình học máy,
* requests, beautifulsoup4, sqlalchemy – để kết nối và trích xuất dữ liệu từ nhiều nguồn khác nhau như Web, API, SQL,...

Hơn nữa, Python dễ dàng được nhiều đối tượng tiếp cận khi có cú pháp đơn giản, cộng đồng người dùng rộng lớn và tài liệu đa dạng, và đồng thời vẫn đáp ứng được các yêu cầu chuyên sâu của chuyên gia. Nhờ khả năng xử lý mạnh mẽ, tính mở rộng cao và linh hoạt, Python đang là lựa chọn hàng đầu của các tổ chức khi xây dựng hệ thống phân tích và xử lý dữ liệu quy mô lớn.

### 2.1.2. Power BI

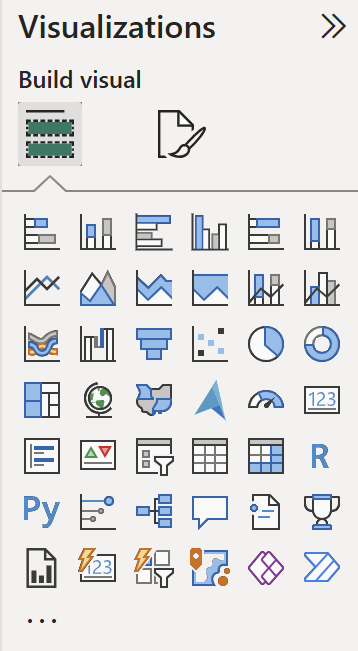
Với nguồn dữ liệu ngày càng khổng lồ phải xử lý như hiện nay, Excel đã dần không còn là một lựa chọn ưa thích của các công ty trong việc trực quan hóa dữ liệu. Excel bị giới hạn trong tầm khoảng 1 triệu dòng, nếu dữ liệu đạt tới độ lớn này, Excel dễ bị chậm hoặc treo máy. Với những người làm trong vị trí DA, BA hay các vị trí khác cần xử lý và trực quan hóa một lượng dữ liệu lớn, họ mong muốn có sự xử lý nhanh gọn, đơn giản, trải nghiệm liền mạch khi truyền tải dữ liệu từ nhiều nguồn và Cloud khác nhau và dữ liệu luôn được cập nhật theo thời gian thực. Tất cả những điều kiện này hiện đã có mặt trong Power BI.

Khi nói tới một ứng dụng có thể kết nối tới gần như mọi nguồn dữ liệu, chắc hẳn Power BI sẽ là cái tên được nhớ tới đầu tiên. Power BI hỗ trợ kết nối tới hơn 100 nguồn dữ liệu khác nhau, từ những nguồn phổ biến như Excel, CSV, SQL,... tới các nền tảng dịch vụ như Google Analytics, Salesforce, Azure, SharePoint,... Đặc biệt hơn, khi dữ liệu trong nguồn đó được cập nhật hay thay đổi, bảng báo cáo trên Power BI cũng sẽ được cập nhật và thay đổi theo, đảm bảo thông tin luôn chính xác.



Hình 1: Khả năng kết nối dữ liệu của Power BI

Power BI cung cấp hàng chục loại biểu đồ và công cụ trực quan như biểu đồ cột, đường, bản đồ, biểu đồ dạng donut, gauge, KPI,… kèm khả năng tương tác như lọc động (slicer), drill-down (đi sâu vào chi tiết), và cross-filtering (lọc chéo giữa các biểu đồ). Giao diện kéo thả thân thiện giúp người dùng dễ dàng thiết kế các báo cáo sinh động mà không cần kỹ năng lập trình.



Hình 2: Phương thức trực quan hóa trong Power BI

Power BI hỗ trợ các tính năng bảo mật nâng cao như Row-Level Security (RLS) – cho phép chỉ định dữ liệu mà mỗi người dùng có thể xem. Bên cạnh đó, việc tích hợp với Azure Active Directory giúp kiểm soát đăng nhập, phân quyền truy cập và bảo vệ dữ liệu nhạy cảm trong tổ chức một cách chặt chẽ và linh hoạt.

Một trong những điểm mạnh là lý do mà nhiều công ty lựa chọn Power BI chính là khả năng chia sẻ báo cáo dễ dàng và nhanh gọn thông qua Power BI Service. Người dùng Người dùng có thể xuất bản báo cáo lên đám mây, chia sẻ với đồng nghiệp hoặc cấp quản lý, phân quyền truy cập theo vai trò.

## 2.2. Các thuật toán và phương pháp sử dụng

### 2.2.1. RFM Analysis (Recency, Frequency, Monetary)

Phân tích RFM là một công cụ mạnh mẽ để phân khúc khách hàng, dựa trên dữ liệu giao dịch lịch sử nhằm đánh giá hành vi tiêu dùng thông qua ba chỉ số chính: Recency (Thời gian gần nhất), Frequency (Tần suất) và Monetary (Giá trị tiền tệ). Phương pháp này giúp doanh nghiệp nhận diện các nhóm khách hàng có giá trị khác nhau, từ đó xây dựng chiến lược tiếp thị phù hợp nhằm tăng giá trị vòng đời khách hàng (CLTV), nâng cao tỷ lệ giữ chân và tối ưu hóa chi phí tiếp thị.

#### 2.2.1.1. Các Chỉ Số RFM

Các Chỉ Số RFM Phân tích RFM sử dụng ba chỉ số để đánh giá hành vi khách hàng

* Recency (R): Đo lường khoảng thời gian từ giao dịch gần nhất của khách hàng đến thời điểm hiện tại. Khoảng thời gian dài hơn cho thấy nguy cơ khách hàng rời bỏ (churn) cao, đòi hỏi doanh nghiệp nỗ lực tái thu hút. Ngược lại, khoảng thời gian ngắn thường liên quan đến cơ hội lớn cho các hoạt động upsell (bán sản phẩm cao cấp hơn) và cross-sell (bán sản phẩm bổ sung).
* Frequency (F): Thể hiện số lần giao dịch của khách hàng trong một khoảng thời gian xác định. Tần suất cao phản ánh mức độ trung thành và khả năng phản hồi tích cực với các chiến dịch tiếp thị. Khách hàng mua sắm thường xuyên (hàng tuần hoặc hàng tháng) có tiềm năng upsell cao hơn so với những người giao dịch thưa thớt (ví dụ, ba tháng một lần).
* Monetary (M): Tổng số tiền khách hàng đã chi tiêu cho thương hiệu trong một khoảng thời gian nhất định. Chỉ số này phản ánh khả năng tài chính và mức độ tin tưởng của khách hàng. Khi kết hợp với tần suất, Monetary giúp tính toán chi tiêu trung bình mỗi giao dịch, từ đó xác định nhóm khách hàng có xu hướng chi tiêu cao hoặc tiềm năng mua các sản phẩm giá trị lớn.

#### 2.2.1.2. Quy Trình Phân Tích RFM

Quy Trình Phân Tích RFM Phân tích RFM được thực hiện qua các bước sau

Bước 1: Tính Giá Trị RFM Dựa trên dữ liệu giao dịch, các giá trị RFM được tính như sau:

* Recency: Khoảng thời gian từ giao dịch gần nhất đến thời điểm hiện tại, tính bằng cách lấy ngày hiện tại trừ ngày giao dịch cuối cùng.
* Frequency: Tổng số lần giao dịch trong khoảng thời gian xác định, thường được chuẩn hóa bằng cách chia cho khoảng thời gian từ giao dịch đầu tiên đến giao dịch cuối cùng.
* Monetary: Tổng giá trị chi tiêu của khách hàng, tính bằng cách cộng tất cả giá trị giao dịch trong lịch sử mua hàng. Dữ liệu đầu vào bao gồm thông tin định danh khách hàng (ID, email hoặc tên) và chi tiết giao dịch (ngày giao dịch, giá trị giao dịch).

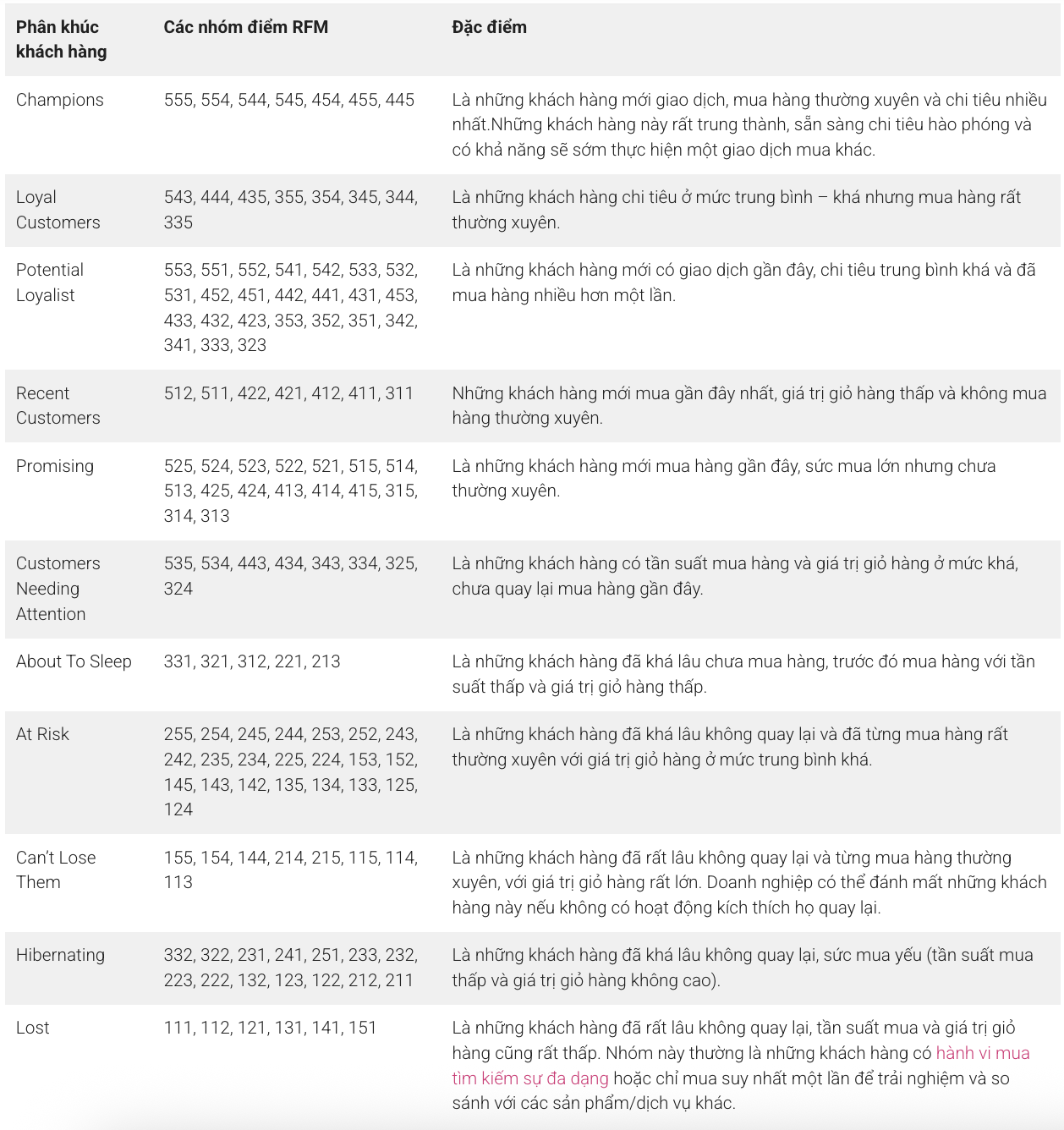
Bước 2: Gán Điểm Số RFM Sau khi tính được các giá trị RFM, chúng được phân chia thành thang điểm từ 1 đến 5 (1 là thấp nhất, 5 là cao nhất) bằng một trong ba phương pháp:

* Phân khoảng đơn giản: Các giá trị RFM được chia thành các khoảng dựa trên đặc điểm ngành và hành vi khách hàng. Ví dụ, với Recency: 5 điểm cho giao dịch trong 24 giờ, 4 điểm cho 3 ngày, 3 điểm cho 30 ngày, 2 điểm cho 180 ngày, 1 điểm cho hơn 180 ngày.
* Quintiles: Dữ liệu RFM được sắp xếp giảm dần và chia thành 5 nhóm, mỗi nhóm chứa 20% khách hàng. Ví dụ, 20% khách hàng có Recency gần nhất được gán 5 điểm, 20% tiếp theo được 4 điểm, v.v. Phương pháp này không yêu cầu hiểu biết sâu về hành vi tiêu dùng và phù hợp với nhiều ngành.
* Phân cụm (Clustering): Sử dụng thuật toán học máy không giám sát như k-means để tự động phân nhóm khách hàng dựa trên giá trị RFM. Số lượng cụm tối ưu được xác định bằng các phương pháp như elbow hoặc silhouette. Phương pháp này mang lại độ chính xác cao hơn bằng cách sử dụng giá trị RFM trực tiếp thay vì điểm số.

Bước 3: Phân khúc khách hàng từ phân tích RFM

Trong phân tích RFM, mỗi chỉ số Recency (R), Frequency (F) và Monetary (M) được gán điểm từ 1 đến 5, tạo ra 5x5x5 = 125 tổ hợp giá trị. Tuy nhiên, để trực quan hóa và phân tích hiệu quả, việc sử dụng biểu đồ 3D cho ba chỉ số này không thực sự tiện lợi. Thay vào đó, một biểu đồ hai chiều được sử dụng, trong đó Frequency và Monetary được kết hợp thành một trục (sức mua, với tổng điểm từ 0 đến 10) và Recency biểu thị mức độ tương tác trên trục còn lại (từ 1 đến 5). Cách tiếp cận này giảm số tổ hợp từ 125 xuống còn 50, giúp dữ liệu dễ hiểu và dễ mô tả hơn. Ví dụ, với một tạp chí, độc giả đăng ký hàng tháng (100 đô la/tháng, tổng 1200 đô la trong 12 tháng) có Monetary = 1200, Frequency = 12, trong khi độc giả đăng ký cả năm (1200 đô la, thanh toán một lần) cũng có Monetary = 1200 nhưng Frequency = 1. Dù sức mua tương đương, sự khác biệt về tần suất giao dịch cho thấy việc kết hợp Frequency và Monetary là hợp lý để đánh giá hành vi khách hàng.

Để đơn giản hóa hơn nữa, 50 tổ hợp này được gom thành 11 nhóm khách hàng có đặc điểm tương đồng, chẳng hạn như Champions (giá trị cao, giao dịch thường xuyên, tương tác gần đây), Loyal Customers (giao dịch thường xuyên nhưng chi tiêu chưa cao), hoặc At Risk (từng tích cực nhưng không còn quay lại). Các nhóm này giúp doanh nghiệp dễ dàng xác định chiến lược tiếp thị phù hợp, từ giữ chân khách hàng trung thành đến tái thu hút những người có nguy cơ rời bỏ.



Hình 3: Ví dụ về phân khúc khách hàng

Bước 4: Trực quan hóa phân tích RFM với công cụ visualize

Bước 5: Ứng Dụng Phân Tích RFM Dựa trên các phân khúc, doanh nghiệp xây dựng chiến lược tiếp thị cá nhân hóa cho từng nhóm khách hàng:

* Champions: Đây là nhóm khách hàng giá trị cao nhất, mua sắm thường xuyên, chi tiêu lớn và có giao dịch gần đây. Doanh nghiệp nên tập trung vào chương trình khách hàng thân thiết, ưu đãi cá nhân hóa và trải nghiệm vượt trội. Ví dụ, chương trình VinID kết hợp thẻ thành viên và ứng dụng di động để cung cấp tích điểm, đổi thưởng và đặc quyền theo hạng mức. Thay vì giảm giá, doanh nghiệp có thể đề xuất sản phẩm cao cấp hoặc gói sản phẩm để tăng giá trị đơn hàng.
* Loyal Customers và Potential Loyalists: Nhóm này giao dịch thường xuyên nhưng giá trị đơn hàng chưa cao. Doanh nghiệp có thể khuyến khích chi tiêu lớn hơn qua các chương trình thưởng (quà tặng khi đạt ngưỡng chi tiêu) hoặc mời tham gia chương trình giới thiệu bạn bè để trở thành đại sứ thương hiệu.
* Recent Customers và Promising: Nhóm khách hàng mới, đang đánh giá trải nghiệm. Doanh nghiệp nên gửi lời cảm ơn sau đơn hàng đầu tiên, cung cấp mã giảm giá cho lần mua tiếp theo và gợi ý sản phẩm phù hợp hoặc combo giá trị cao để tăng tần suất giao dịch.
* Customers Needing Attention và At Risk: Nhóm này từng tương tác tích cực nhưng gần đây không quay lại. Doanh nghiệp cần khảo sát nguyên nhân (chất lượng sản phẩm, dịch vụ hoặc nhu cầu thay đổi) và cung cấp ưu đãi giá trị gia tăng (sản phẩm cải tiến, dịch vụ nâng cấp) để khôi phục niềm tin.
* About to Sleep, Hibernating và Lost: Nhóm khách hàng có tần suất và giá trị thấp, đã lâu không tương tác. Doanh nghiệp cần chiến dịch tái tương tác mạnh mẽ như retargeting, email cá nhân hóa, nhắn tin qua mạng xã hội hoặc gọi điện trực tiếp. Ưu đãi giới hạn thời gian (phiếu giảm giá, quà tặng) có thể kích thích giao dịch sớm, tránh nguy cơ thương hiệu bị lãng quên.

#### 2.2.1.3. ý nghĩa và ứng dụng của phân tích RFM.

Phân tích RFM là một phương pháp hiệu quả để phân khúc khách hàng dựa trên hành vi giao dịch, cho phép doanh nghiệp tối ưu hóa chiến lược tiếp thị bằng cách nhắm mục tiêu vào các nhóm cụ thể, như duy trì lòng trung thành của nhóm Champions qua chương trình tích điểm như VinID hoặc tái thu hút nhóm At Risk bằng chiến dịch retargeting với ưu đãi giới hạn. Bằng cách cá nhân hóa thông điệp, kết hợp các phương pháp phân tích hiện đại như quintiles hoặc clustering với dữ liệu chất lượng cao, RFM không chỉ nâng cao giá trị vòng đời khách hàng (CLTV), tỷ lệ giữ chân và giảm chi phí tiếp thị kém hiệu quả, mà còn mang lại độ chính xác và tính ứng dụng vượt trội trong nhiều ngành, từ bán lẻ đến dịch vụ.

### 2.2.2. K-Means Clustering (phân khúc khách hàng)

#### 2.2.2.1. Học máy không giám sát

Unsupervised Learning là học không giám sát vì chúng ta không biết được đầu ra (outcome hay nhãn) trông như thế nào mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân cụm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán. Hiểu theo một cách khác, Unsupervised learning là khi chúng ta chỉ có dữ liệu đầu vào X mà không biết nhãn Y tương ứng. Những thuật toán loại này được gọi là Unsupervised learning vì không giống như Supervised learning, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Cụm không giám sát được đặt tên theo nghĩa này.

Một trong những bài toán cụ thể của Unsupervised learning chính là Clustering (phân cụm). Clustering là một bài toán phân cụm toàn bộ dữ liệu X thành các cụm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm. Ví dụ: phân cụm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng.

#### 2.2.2.2. Thuật toán phân cụm K-means

Thuật toán K-means clustering (phân cụm K-means) chính là một trong những thuật toán cơ bản nhất trong Unsupervised learning. Trong thuật toán K-means clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Đối với thuật toán K-Means Clustering, mỗi cụm dữ liệu được đặc trưng bởi một trọng tâm (centroid). Trọng tâm của cụm được xác định bởi trung bình của tọa độ toàn bộ các điểm dữ liệu nằm trong cụm. Nhãn của mỗi điểm sẽ được xác định bởi trọng tâm có khoảng cách Euclidean gần nhất đến quan sát đó.

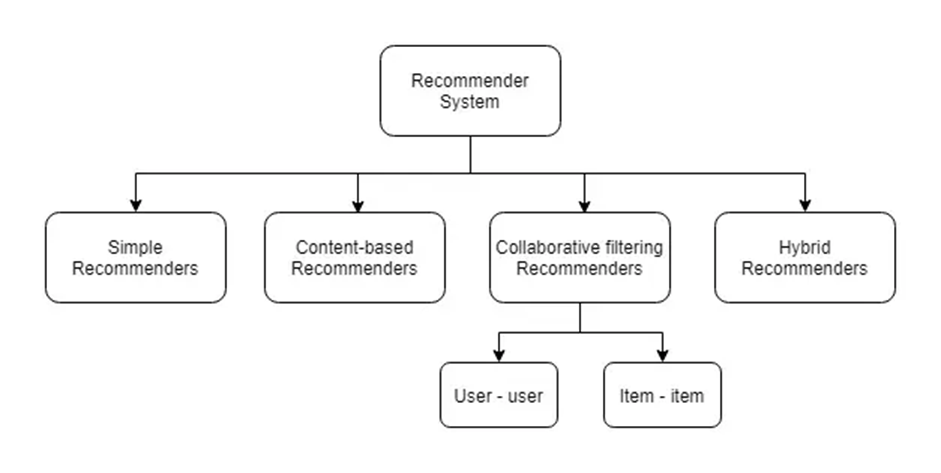
Cụ thể hơn thuật toán K-Means hoạt động như sau:

* Bước 1: Khởi tạo ngẫu nhiên k trọng tâm {u1, u2, u3,...}
* Bước 2: Với mọi điểm dữ liệu sẽ được gán nhãn theo trọng tâm gần nhất. Trong đó, khoảng cách từ điểm dữ liệu đến trọng tâm sẽ được tính bằng khoảng cách Euclid.
* Bước 3: Cập nhật tọa độ của các trọng tâm mới bằng cách tính trung bình tọa độ của các điểm dữ liệu trong cụm tương ứng.
* Bước 4: Lặp lại bước 2 cho đến khi nhãn của các điểm dữ liệu không thay đổi nữa thì thuật toán kết thúc.

### 2.2.3. Recommendation System cơ bản (dựa trên nhóm RFM)

Hệ thống gợi ý (Recommend System) là một thuật toán trí tuệ nhân tạo (AI), thường liên quan đến học máy, sử dụng Big data để gợi ý hoặc đề xuất các sản phẩm tương tự cho người tiêu dùng. Các gợi ý này có thể dựa trên nhiều tiêu chí khác nhau, bao gồm các lần mua trước, lịch sử tìm kiếm, thông tin nhân khẩu học, các đặc trưng của sản phẩm và các yếu tố khác. Hệ thống gợi ý rất hữu ích vì chúng giúp người dùng khám phá những sản phẩm và dịch vụ phù hợp với nhu cầu của họ mà họ khó có thể tìm thấy nếu không có sự trợ giúp.

Hệ thống gợi ý có thể chia thành 4 loại chính:



Hình 4: Bốn loại chính của hệ thống gợi ý

Trong khuôn khổ bài luận này, nhóm sẽ tập trung vào thiết kế hệ thống gợi ý thuộc loại Hybrid-Recommenders kết hợp giữa Collaborative Filtering Recommenders (CF) và Content-based Recommenders (CB). Hệ thống này hoạt động bằng cách sử dụng Content-based Recommenders đối với các khách hàng lần đầu mua hàng tại cửa hàng, thuật toán hoạt động bằng cách lấy các đặc điểm của sản phẩm để đề xuất những sản phẩm tương tự. Đối với các khách hàng đã từng mua hàng tại cửa hàng trước đó, hệ thống sẽ sử dụng Collaborative Filtering, tận dụng thông tin hành vi từ tất cả lịch sử mua sắm của các khách hàng để cố gắn suy luận ra hành vi của khách hàng hiện tại.

### 2.2.4. Các dashboard và thành phần được sử dụng trong Power BI

Power Query - một công cụ trong Power BI Desktop được sử dụng để chuyển đổi và chuẩn bị dữ liệu. Công cụ này giúp chúng ta kết nối, định hình và chuyển đổi dữ liệu từ nhiều nguồn để có được đúng lượng dữ liệu cần thiết theo nhu cầu của người dùng. Khi chúng ta tải dữ liệu lên từ nhiều nguồn khác nhau, dữ liệu đã tải lên có thể sẽ không ở định dạng mong muốn. Chuyển đổi dữ liệu giúp dọn dẹp và sắp xếp dữ liệu bằng cách xóa các hàng không cần thiết, tách các cột và thay đổi định dạng trước khi tải vào Power BI để phân tích.  
DAX (Data Analysis Expressions) - ngôn ngữ công thức và truy vấn được thiết kế để hoạt động với các mô hình dữ liệu dạng bảng và chủ yếu được sử dụng để đơn giản hóa các tác vụ phân tích và tính toán dữ liệu, được dùng phổ biến nhất trong Power BI. Nó cung cấp cho người dùng khả năng tạo các phép tính phức tạp, xác định số liệu tùy chỉnh và thực hiện các thao tác dữ liệu phức tạp. DAX có nhiều chức năng mạnh mẽ mà Excel không có.

Power BI Visualizations - là chức năng nổi bật nhất khi sử dụng Power BI. Đây là nơi chứa tất cả các biểu đồ, nơi biến những bảng biểu và con số khô khan trở nên sinh động, đầy màu sắc và đặc biệt là dễ dàng truyền tải trực quan nguồn thông tin mà người dùng muốn biết. Power BI Visualizations chứa các biểu đồ gồm:

* Nhóm biểu đồ cột và thanh: Clustered column chart (Biểu đồ cột cụm), Stacked column chart (Biểu đồ cột xếp chồng), 100% Stacked column chart (Biểu đồ cột xếp chồng 100%), Clustered bar chart (Biểu đồ thanh ngang cụm), Stacked bar chart (Biểu đồ thanh ngang xếp chồng), 100% Stacked bar chart (Biểu đồ thanh ngang xếp chồng 100%).
* Nhóm biểu đồ đường và vùng: Line chart (Biểu đồ đường), Area chart (Biểu đồ vùng), Stacked area chart (Biểu đồ vùng xếp chồng), Line and stacked column chart (Biểu đồ kết hợp đường và cột xếp chồng), Line and clustered column chart (Biểu đồ kết hợp đường và cột cụm).
* Nhóm biểu đồ phân tích: Waterfall chart (Biểu đồ thác nước), Funnel chart (Biểu đồ phễu), Scatter chart/Scatter plot (Biểu đồ phân tán), Pie chart (Biểu đồ tròn), Donut chart (Biểu đồ bánh vòng).
* Nhóm bảng và ma trận: Table (Bảng dữ liệu), Matrix (Ma trận dữ liệu).
* Nhóm bản đồ: Map (Bản đồ cơ bản), Filled map (Bản đồ tô màu theo vùng), Azure map visual (Bản đồ Azure).
* Các biểu đồ khác: Card (Thẻ thông tin đơn lẻ), Multi-row card (Thẻ thông tin nhiều hàng), Gauge (Đồng hồ đo), KPI (Chỉ số hiệu suất chính), Decomposition tree (Cây phân tích), Treemap (Biểu đồ cây)

# Chương 3: Ứng dụng các phương pháp vào bài toán cụ thể

## 3.1. Tiền xử lý dữ liệu

### 3.1.1. Mô tả dữ liệu

Bảng mô tả sơ bộ dữ liệu, gồm có 7 bộ dữ liệu:

* **Bảng distribution\_centers**

Mục đích: Lưu thông tin về các trung tâm phân phối hàng hóa.

|  |  |
| --- | --- |
| id | Mã định danh trung tâm phân phối |
| name | Tên trung tâm phân phối |
| latitude | Vĩ độ |
| longitude | Kinh độ |

Bảng 1: Cấu trúc bảng distribution\_centers

* **Bảng events**

Mục đích: Ghi lại các hành vi và tương tác của người dùng trên nền tảng (như xem sản phẩm, thêm vào giỏ hàng, mua hàng,...).

|  |  |
| --- | --- |
| id | Mã định danh sự kiện |
| user\_id | Mã người dùng |
| sequence\_number | Số thứ tự sự kiện |
| session\_id | Mã phiên truy cập |
| created\_at | Thời gian diễn ra sự kiện |
| ip\_address | Địa chỉ IP |
| city | Thành phố |
| state | Bang / Tỉnh |
| postal\_code | Mã bưu chính |
| browser | Trình duyệt |
| traffic\_source | Nguồn truy cập |
| uri | Đường dẫn URI |
| event\_type | Loại sự kiện |

Bảng 2: Cấu trúc bảng events

* **Bảng inventory\_items**

Mục đích: Quản lý từng đơn vị hàng hóa trong kho (mức độ chi tiết tới từng sản phẩm cụ thể)

|  |  |
| --- | --- |
| id | Mã định danh hàng tồn kho |
| product\_id | Mã sản phẩm |
| created\_at | Thời gian nhập kho |
| sold\_at | Thời gian bán ra |
| cost | Giá vốn |
| product\_category | Danh mục sản phẩm |
| product\_name | Tên sản phẩm |
| product\_brand | Thương hiệu sản phẩm |
| product\_retail\_price | Giá bán lẻ sản phẩm |
| product\_department | Bộ phận / phòng ban sản phẩm |
| product\_sku | Mã SKU sản phẩm |
| product\_distribution\_center\_id | Mã trung tâm phân phối sản phẩm |

Bảng 3: Cấu trúc bảng inventory\_items

* **Bảng order\_items**

Mục đích: Lưu thông tin từng mặt hàng trong mỗi đơn hàng cụ thể.

|  |  |
| --- | --- |
| id | Mã định danh mục hàng trong đơn |
| order\_id | Mã đơn hàng |
| user\_id | Mã người dùng |
| product\_id | Mã sản phẩm |
| inventory\_item\_id | Mã hàng tồn kho |
| status | Trạng thái đơn hàng |
| created\_at | Thời gian tạo đơn |
| shipped\_at | Thời gian giao hàng |
| delivered\_at | Thời gian giao thành công |
| returned\_at | Thời gian hoàn trả |

Bảng 4: Cấu trúc bảng order\_items

* **Bảng orders**

Mục đích: Ghi lại thông tin chung của các đơn hàng được tạo bởi người dùng.

|  |  |
| --- | --- |
| order\_id | Mã đơn hàng |
| user\_id | Mã người dùng |
| status | Trạng thái đơn hàng |
| gender | Giới tính người dùng |
| created\_at | Thời gian tạo đơn |
| returned\_at | Thời gian hoàn trả |
| shipped\_at | Thời gian giao hàng |
| delivered\_at | Thời gian giao thành công |
| num\_of\_item | Số lượng mặt hàng trong đơn |

Bảng 5: Cấu trúc bảng orders

* **Bảng products**

Mục đích: Chứa thông tin cơ bản về sản phẩm (danh mục, thương hiệu, giá,...).

|  |  |
| --- | --- |
| id | Mã định danh sản phẩm |
| cost | Giá vốn sản phẩm |
| category | Danh mục sản phẩm |
| name | Tên sản phẩm |
| brand | Thương hiệu |
| retail\_price | Giá bán lẻ |
| department | Bộ phận / phòng ban |
| sku | Mã SKU |
| distribution\_center\_id | Mã trung tâm phân phối |

Bảng 6: Cấu trúc bảng products

* **Bảng users**

Mục đích: Quản lý thông tin người dùng đăng ký trên nền tảng.

|  |  |
| --- | --- |
| id | Mã người dùng |
| first\_name | Tên |
| last\_name | Họ |
| email | Địa chỉ email |
| age | Tuổi |
| gender | Giới tính |
| state | Bang / Tỉnh |
| street\_address | Địa chỉ đường |
| postal\_code | Mã bưu chính |
| city | Thành phố |
| country | Quốc gia |
| latitude | Vĩ độ |
| longitude | Kinh độ |
| traffic\_source | Nguồn truy cập |
| created\_at | Thời gian tạo tài khoản |

Bảng 7: Cấu trúc bảng users

### 3.1.2. Tiền xử lý dữ liệu

Source code: https://colab.research.google.com/drive/1CfgMdS6NtalQ11uT-0mFjDFCpeZPb1oE

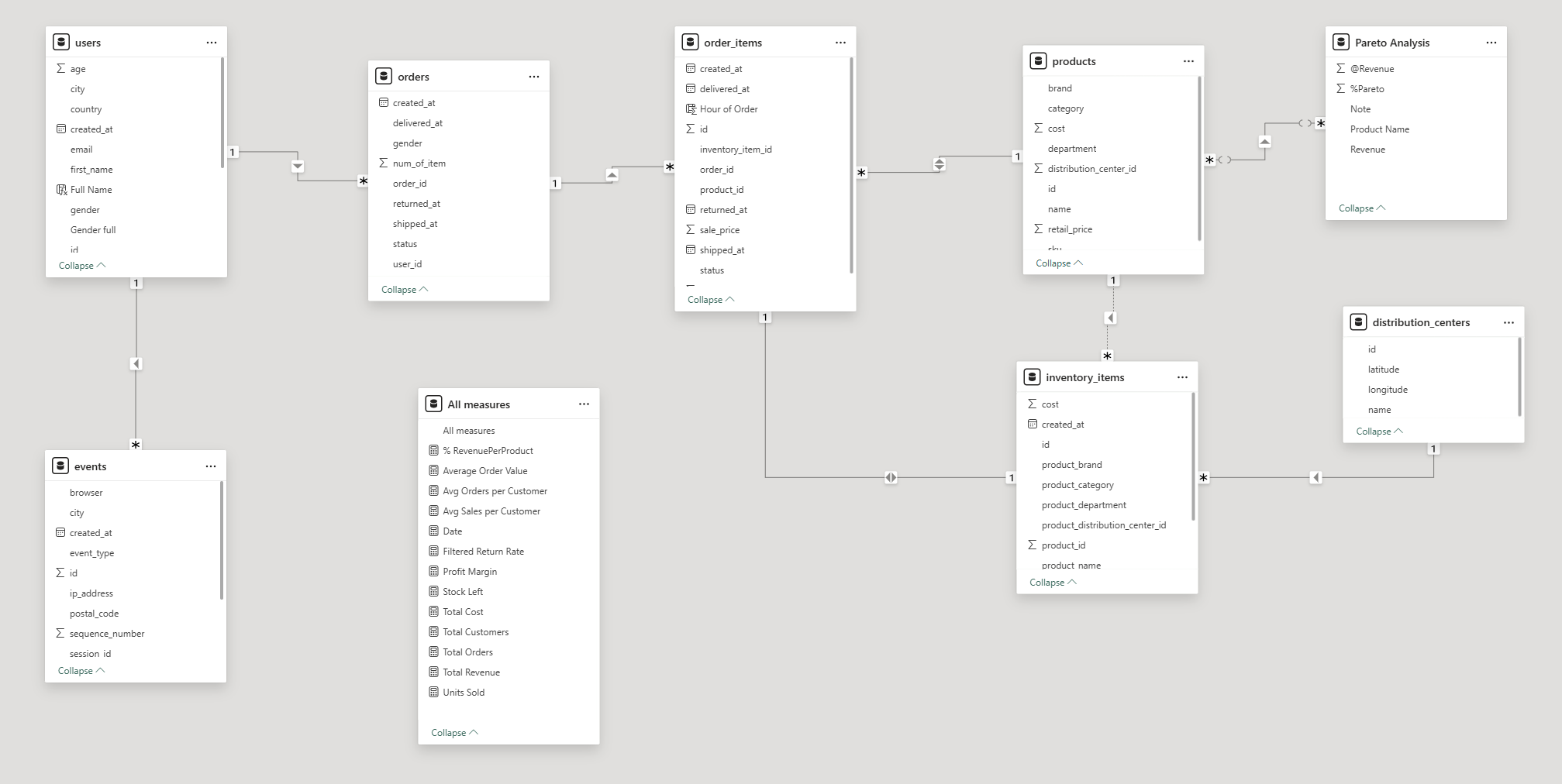
Mô tả:

Nhóm chỉ hướng tới việc xử lý các phần dữ liệu bị trống, còn lại giữ nguyên nhằm đảm bảo tính toàn vẹn của bộ dữ liệu.

* Khi kiểm tra bảng user, ta điền các giá trị trống (null) thành unknown.
* Khi kiểm tra bảng events, các giá trị trống trong cột city chiếm số nhỏ, cũng điền unknown.
* Khi kiểm tra bảng events, các giá trị trống trong cột city chiếm tỷ lệ nhỏ, cũng được điền là "unknown".
* Ngoài ra, ta nhận thấy số user\_id trong bảng user không khớp với bảng orders — điều này có thể do nhiều người tạo tài khoản nhưng không thực hiện mua hàng, điều này là bình thường trong thời đại ngày nay.
* Đặc biệt Trong bảng orders, các dòng có shipped\_at là NULL cũng đồng thời có cả hai cột return và deliver là NULL. Điều này cho thấy các đơn hàng này đã bị hủy ngay sau khi đặt, trước khi được giao đi.
* Ngoài ra, với những dòng có deliver là NULL thì return cũng luôn NULL. Điều này nghĩa là các đơn hàng này đã bị hủy trong quá trình giao hàng, trước khi đến tay người mua.
* Khi kiểm tra bảng df\_inventory, các dòng có sold\_at là NULL được hiểu là hàng tồn kho. Các giá trị thiếu ở product\_name và product\_brand được điền là "unknown".
* Khi kiểm tra bảng df\_products, có 2 dòng thiếu product\_name và 24 dòng thiếu brand, tất cả đều được điền là "unknown".

## 3.2. Xây dựng Dashboard với Power BI:

Sau khi mô tả và hiểu rõ bộ dữ liệu, nhóm tiến hành làm sạch và chuyển đổi dữ liệu bằng Power Query để chuẩn bị cho bước vẽ dashboard - trực quan hóa dữ liệu bằng Power BI. Đây là bộ dataset cụ thể sau khi được xử lý bằng Power Query:



Hình 5. Model view của bộ dữ liệu sau khi xử lý bằng Power Query

Ở dataset sau khi đã được xử lý trên, ta có thể nhận thấy được 2 table mới đó là *Pareto Analysis* và *All measures*. Mỗi bảng mới được tạo đều đóng vai trò rất quan trọng, cung cấp thêm nhiều thông tin hữu ích để dễ dàng hơn trong việc trực quan hóa dữ liệu trong dashboard.

Table mới *Pareto Analysis* được tạo ra dựa trên nguyên lý Pareto (80/20). Mục đích của table này chính là để xác định được các sản phẩm chiếm phần lớn doanh thu trong dataset (có thể hiểu là xác định top 20% sản phẩm giúp tạo ra 80% doanh thu). Các thuộc tính trong table bao gồm:

|  |  |
| --- | --- |
| Product Name | Tên sản phẩm |
| Revenue | Doanh thu |
| %Pareto | Tỷ lệ doanh thu tích lũy, dùng để xác định ngưỡng 80% |
| Note | Phân loại sản phẩm đó theo nguyên lý Pareto, ví dụ: “Top 20%”, “Others”, “Remaining 80%”… |

Bảng 8: bảng Pareto Analysis

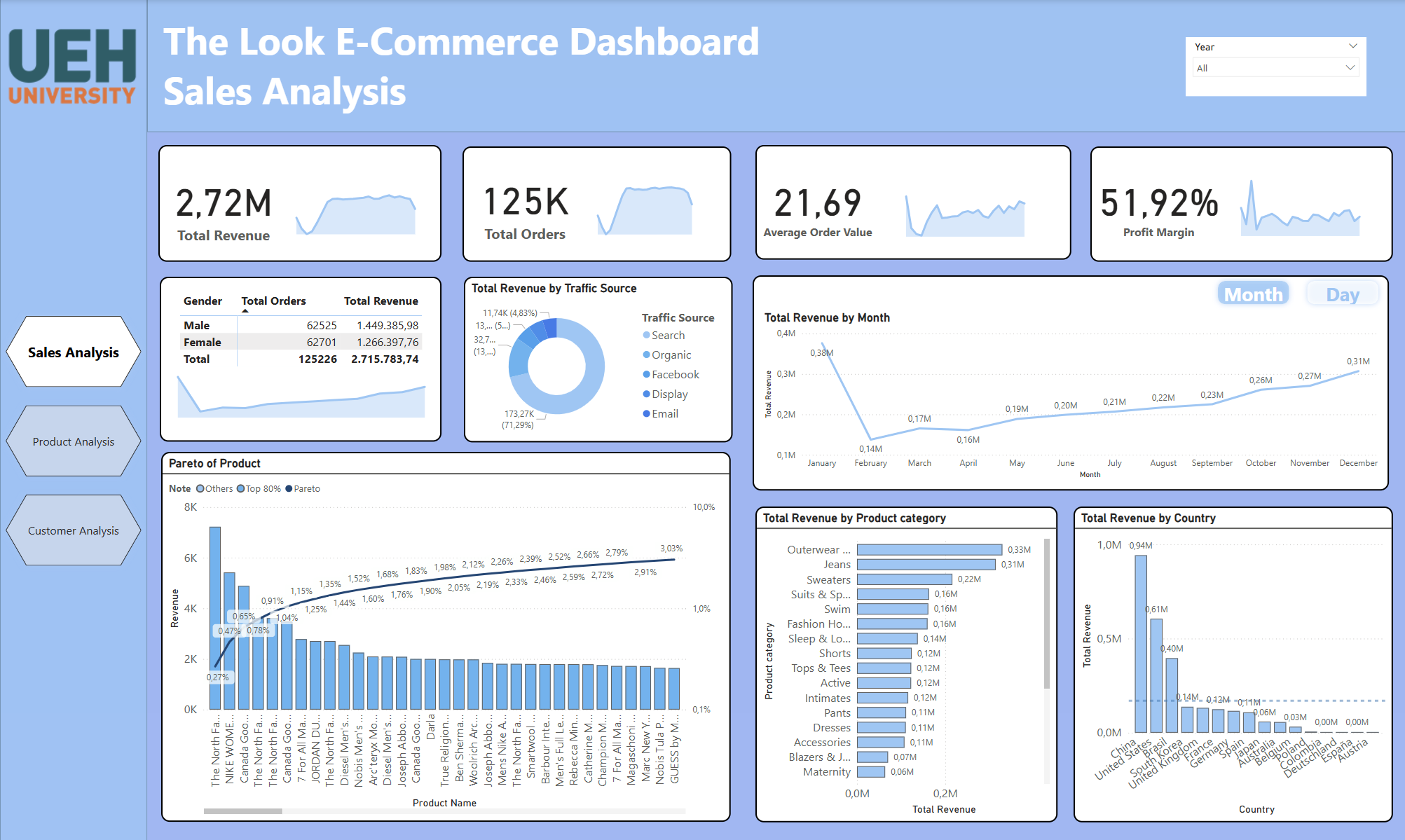
Đây là một table riêng không phụ thuộc vào các table chính (không liên kết với các table khác), được tạo ra để liệt kê các measures (chỉ số) cần thiết sẽ sử dụng trong dashboard. Table này giúp cho các table khác gọn gàng hơn, dễ dàng lọc hoặc tạo slicer theo thời gian. Table này bao gồm các measures:

|  |  |
| --- | --- |
| % RevenuePerProduct | % doanh thu của từng sản phẩm |
| Average Order Value | Giá trị trung bình của một đơn hàng |
| Avg Orders per Customer | Số đơn trung bình của mỗi khách hàng |
| Avg Sales per Customer | Doanh thu trung bình trên mỗi khách hàng |
| Date | Trường thời gian dùng để lọc, tạo slicer thời gian |
| Filtered Return Rate | Tỷ lệ hoàn trả hàng đã lọc |
| Profit Margin | Biên lợi nhuận |
| Stock Left | Hàng tồn kho còn lại |
| Total Cost | Tổng chi phí |
| Total Customers | Tổng số lượng khách hàng |
| Total Orders | Tổng số đơn hàng |
| Total Revenue | Tổng doanh thu |
| Units Sold | Tổng số sản phẩm đã bán |

Bảng 9: Bảng tổng hợp các chỉ số

Để trực quan hóa dữ liệu lên dashboard được đầy đủ và rõ ràng, nhóm quyết định chia nhỏ nội dung theo các nhóm nhỏ tương ứng với 3 sheet dashboard: Sales Analysis, Product Analysis, Customer Analysis. Giống như tên gọi của nó, từng sheet là một dashboard riêng, mỗi dashboard sẽ phân tích chuyên sâu từng khía cạnh của bộ dữ liệu và tập trung theo lần lượt các mảng sale, sản phẩm, khách hàng.

* *Sales Analysis*: Tập trung vào doanh thu, đơn hàng, tỷ suất lợi nhuận, hiệu suất theo kênh bán hàng, thời gian, giới tính, v.v.
* *Product Analysis*: Phân tích chi tiết theo sản phẩm, danh mục sản phẩm, sản phẩm theo Pareto, tồn kho, v.v.
* *Customer Analysis*: Phân tích hành vi khách hàng, tỷ lệ quay lại, giá trị vòng đời khách hàng,...



Hình 6. Dashboard Sales Analysis

*Sales Analysis* tập trung vào cái nhìn tổng quát về tình hình kinh doanh thông qua bộ dữ liệu. Có thể hiểu *Sales Analysis* đóng vai trò như một bức tranh toàn cảnh mô tả về tình hình kinh doanh của doanh nghiệp. Nhìn chung, dashboard có 10 charts chính chia thành các nhóm biểu đồ.

Nhóm biểu đồ Cards thể hiện 4 KPI chính. Mục đích của các biểu đồ này chính là giúp người xem nắm được kết quả và hiệu suất kinh doanh tổng thể của doanh nghiệp từ bộ dữ liệu. Các KPI bao gồm:

* *Total Revenue:* Tổng doanh thu từ tất cả các đơn hàng.
* *Total Orders:* Tổng số lượng đơn hàng đã được đặt.
* *Average Order Value:* Giá trị trung bình mỗi đơn hàng.
* *Profit Margin :* Biên lợi nhuận, phản ánh khả năng sinh lời.

Tiếp theo là biểu đồ phân tích tổng lượng đơn hàng và doanh thu theo giới tính. Biểu đồ này so sánh sự đóng góp vào tổng đơn hàng và doanh thu của nam và nữ nhằm mục đích phân tính hành vi tiêu dùng theo từng giới, từ đó có thể có cách tiếp cận phù hợp hơn với từng giới.

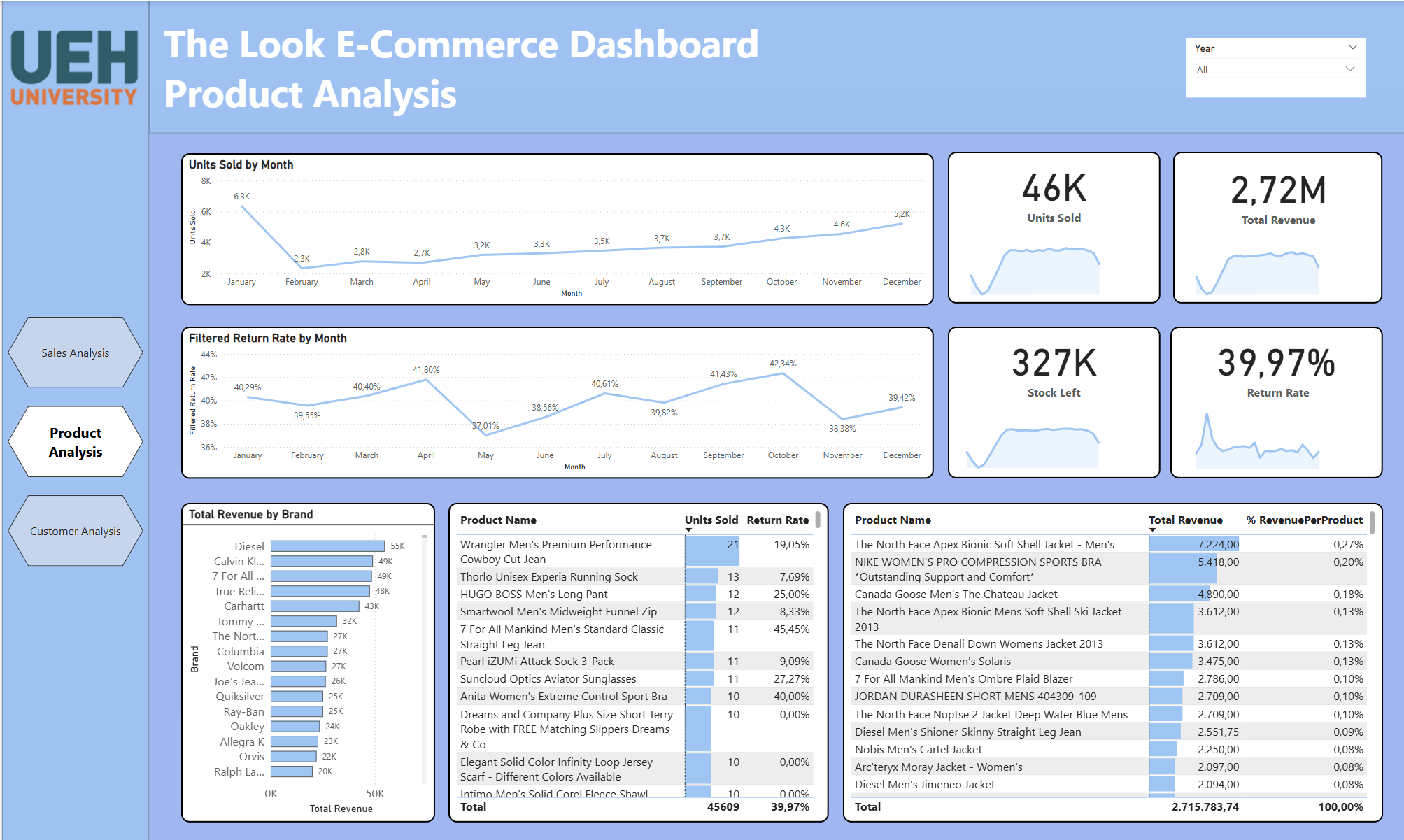
Biểu đồ tròn *Total Revenue by Traffic Source* thể hiện tổng doanh thu theo từng nguồn truy cập khác nhau của khách hàng với 5 phương thức chính: Search, Organic, Facebook, Display và Email. Biểu đồ giúp xác định các kênh marketing mang lại hiệu quả cao, từ đó tập trung markting vào đúng đối tượng hơn, tối ưu chi phí quảng cáo.

Biểu đồ đường *Total Revenue by Month* thể hiện doanh thu theo từng tháng trong năm với trục hoành thể hiện các tháng và trục tung là tổng doanh thu tháng đó. *Total Revenue by Month* giúp theo dõi xu hướng biến động doanh thu theo thời gian, từ đó nhận biết được khoảng thời gian đó là cao điểm hay thấp điểm. Ngoài ra, biểu đồ này còn có thể biểu diễn doanh thu theo từng ngày trong một tháng bằng cách sử dụng cú pháp (ctrl + click vào chữ *Day* ở góc trên bên phải của biểu đồ).

*Pareto of Product* có sự kết hợp giữa biểu đồ cột và đường cong Pareto. Pareto là một đường cong được vẽ trên biểu đồ, biểu diễn giá trị phần trăm tích lũy của các yếu tố từ trái sang phải trên trục X. Nó giúp người đọc nhìn thấy tỷ lệ phần trăm tác động của các yếu tố này so với tổng thể. Đường cong tích lũy thường bắt đầu từ 0% ở cột đầu tiên và tăng dần lên đến 100% ở cột cuối cùng. Biểu đồ hiển thị tỷ lệ doanh thu do các sản phẩm mang lại, từ đó xác định nhóm sản phẩm nào là chủ lực giúp tạo ra phần lớn doanh thu.

Biểu đồ thanh ngang *Total Revenue by Product category* biểu thị doanh thu theo từng ngành hàng hay danh mục sản phẩm, ví dụ như Outerwear, Jeans, Sweater,... Biểu đồ giúp người xem hình dung hiệu quả kinh doanh của từng dòng sản phẩm khác nhau của doanh nghiệp.

Cuối cùng là biểu đồ cột *Total Revenue by Country* - doanh thu theo quốc gia, nhằm xác định thị trường tiềm năng cho doanh nghiệp, từ đó điều chỉnh nhân lực và chiến lược kinh doanh theo từng khu vực cụ thể.



Hình 7: Dashboard Product Analysis

Nếu như *Sales Analysis* là bức tranh toàn cảnh về doanh nghiệp thì *Product Analysis* sẽ cung cấp cái nhìn chuyên sâu hơn về hiệu suất bán hàng theo từng sản phẩm của doanh nghiệp. *Product Analysis* có 9 thành phần được chia thành các nhóm thể hiện các nội dung cụ thể riêng.

Giống như *Sales Analysis*, *Product Analysis* cũng có 4 Cards thể hiện 4 KPI chính:

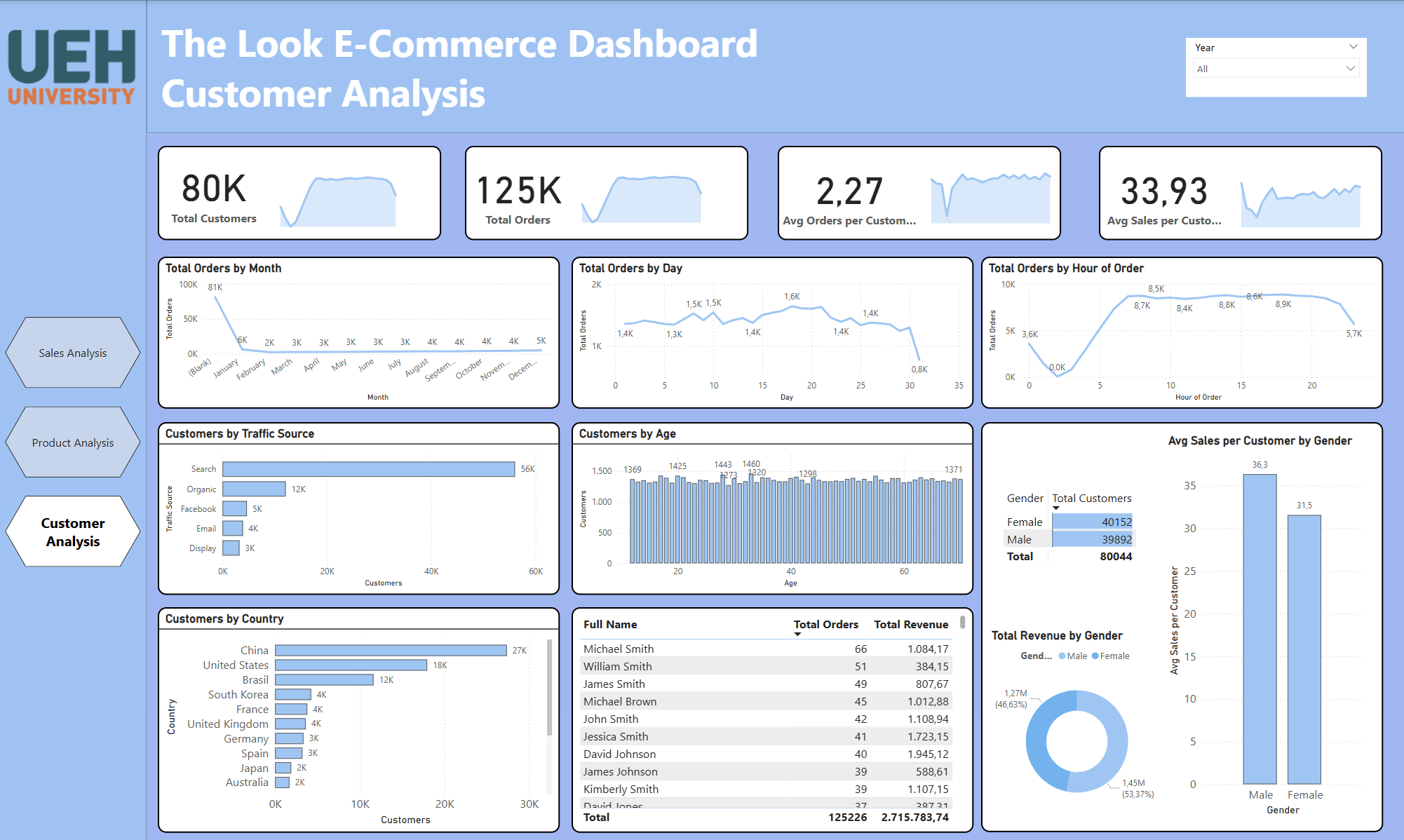
* *Units Sold:* Tổng số sản phẩm đã bán.
* *Total Revenue:* Tổng doanh thu từ các sản phẩm.
* *Stock Left:* Tồn kho còn lại.
* *Return Rate:* Tỷ lệ sản phẩm bị trả lại.

Nhóm biểu đồ đường *Units Sold by Month* và *Filtered Return Rate by Month* lần lượt mô tả diễn biến số lượng sản phẩm bán được theo tháng và tỷ lệ trả hàng theo tháng.

Biểu đồ thanh nganh *Total Revenue by Brand* hiển thị doanh thu theo từng thương hiệu cụ thể nhằm đánh giá hiệu suất kinh doanh theo thương hiệu, phục vụ cho việc tập trung đầu tư vào thương hiệu có doanh thu cao.

Hai tables trong dashboard gồm:

* Bảng thống kê chi tiết số lượng bán và tỷ lệ hoàn trả theo sản phẩm. Tỷ lệ này được tính bằng Tổng số đơn hoàn lại (*'Returned'*) / Tổng số đơn hoàn thành (*'Complete'*) của sản phẩm. Mức phần trăm này càng cao cho thấy sản phẩm này có đơn hoàn trả nhiều hơn so với đơn hoàn thành, có thể hiểu sản phẩm này đang tồn tại nhiều nhược điểm khiến khách hàng không hài lòng. Và ngược lại, mức phần trăm thấp tức số đơn hoàn sản phẩm này ít hơn số đơn hoàn thành.
* Bảng liệt kê sản phẩm theo sự đóng góp vào doanh thu và tỷ lệ doanh thu mà mỗi sản phẩm mang lại (%RevenuePerProduct).



Hình 8: Dashboard Customer Analysis

Cuối cùng là dashboard *Customer Analysis*, dashboard này tập trung vào phân tích đặc điểm và hành vi tiêu dùng của khách hàng. Từ đó doanh nghiệp hiểu rõ hơn về người tiêu dùng và đưa ra chiến lược phù hợp để tăng trưởng doanh thu cũng như nâng cao trải nghiệm khách hàng.

Tương tự như 2 dashboard trước, *Customer Analysis* cũng có 4 Cards thể hiện 4 KPI chính để đo lường và thống kê khách hàng:

* *Total Customers*: Tổng số khách hàng đã mua hàng
* *Total Orders*: Tổng số đơn hàng đặt bởi tất cả khách hàng
* *Avg Orders per Customer*: Số đơn hàng trung bình mà mỗi khách hàng thực hiện
* *Avg Sales per Customer*: Doanh thu trung bình tạo ra từ mỗi khách hàng

Nhóm biểu đồ đường thể hiện xu hướng thay đổi tổng lượng đơn hàng theo từng khoảng thời gian khác nhau:

* *Total Orders by Month*: Phân bổ đơn hàng theo từng tháng trong năm.
* *Total Orders by Day*: Theo dõi số đơn hàng theo từng ngày trong tháng.
* *Total Orders by Hour of Order*: Thời gian khách hàng thường xuyên đặt hàng trong ngày.

Nhóm biểu đồ cột và thanh ngang thể hiện sự phân bố của lượng khách hàng theo từng thang đo khác nhau:

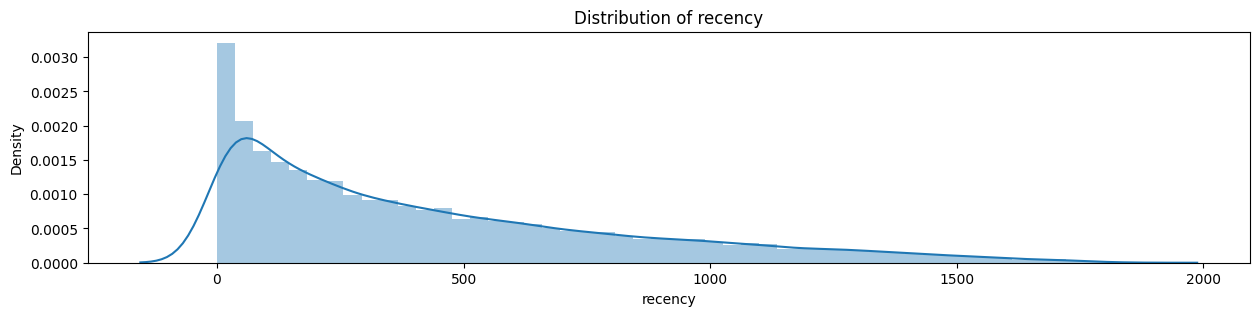
* *Customers by Traffic Source*: Số lượng khách hàng đến từ các nguồn truy cập khác nhau (Search, Organic, Facebook, Display và Email)
* *Customers by Age*: Phân bố khách hàng theo độ tuổi
* *Customers by Country*: Số lượng khách hàng theo từng quốc gia

Nhóm biểu đồ thể hiện sự phân bố của các chỉ số theo từng giới:

* Bảng thể hiện số lượng khách hàng của từng giới
* *Avg Sales per Customer by Gender*: Biểu đồ cột thể hiện sự chênh lệch chỉ số sales trung bình giữa nam và nữ
* *Total Revenue by Gender*: Biểu đồ Donut thể hiện tổng doanh thu bán hàng theo từng giới tính.

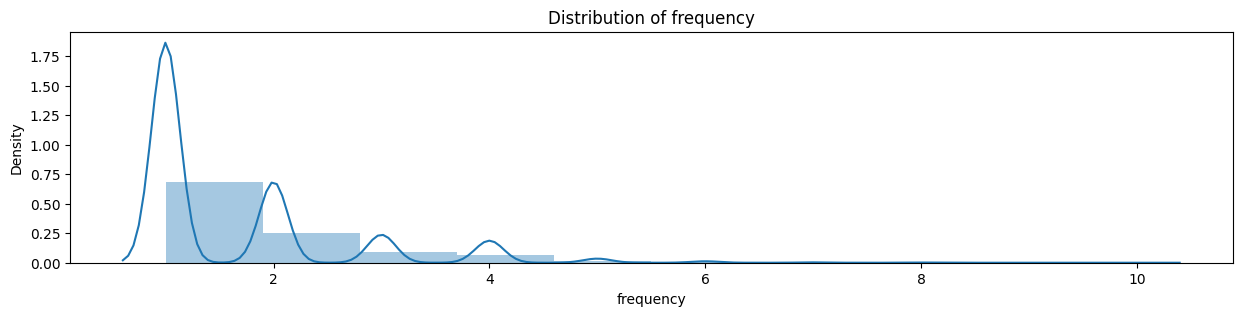
Với mỗi sheet dashboard đều có filter lọc theo từng năm (từ 2019 tới 2024) ở góc phía trên bên phải của dashboard, giúp cho người dùng dễ dàng theo dõi tình hình biến động của bộ data theo từng năm.

## 3.3. Phân tích và phân khúc khách hàng bằng RFM + K-Means



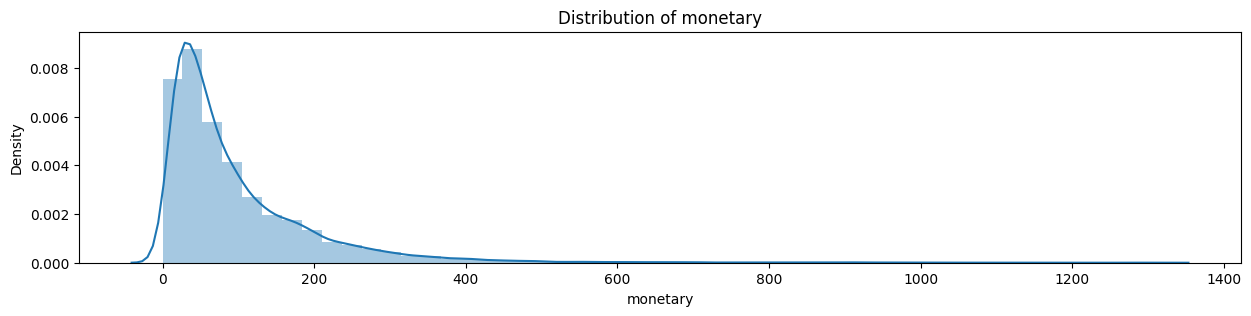
Hình 9: Phân phối khách hàng mua gần đây

Biểu đồ cho thấy phần lớn khách hàng có giá trị recency thấp, tức là họ vừa mua hàng gần đây (phổ biến trong khoảng 0–100 ngày). Mật độ tập trung cao nhất nằm gần mốc 0, sau đó giảm dần theo hình dạng phân phối lệch phải. Điều này cho thấy doanh nghiệp có nhiều khách hàng đang tương tác gần đây, là nhóm khách hàng tiềm năng cần được duy trì sự chú ý và chăm sóc. Tuy nhiên, cũng tồn tại một phần khách hàng có recency rất cao (trên 500 đến hơn 1000 ngày), nghĩa là họ đã lâu không quay lại. Đối với nhóm khách hàng này, doanh nghiệp nên xây dựng những chiến dịch gợi nhớ khách hàng quay lại, gửi email marketing hoặc các giảm giá đặc biệt cho nhóm này để kích thích người dùng sử dụng lại các sản phẩm và dịch vụ.



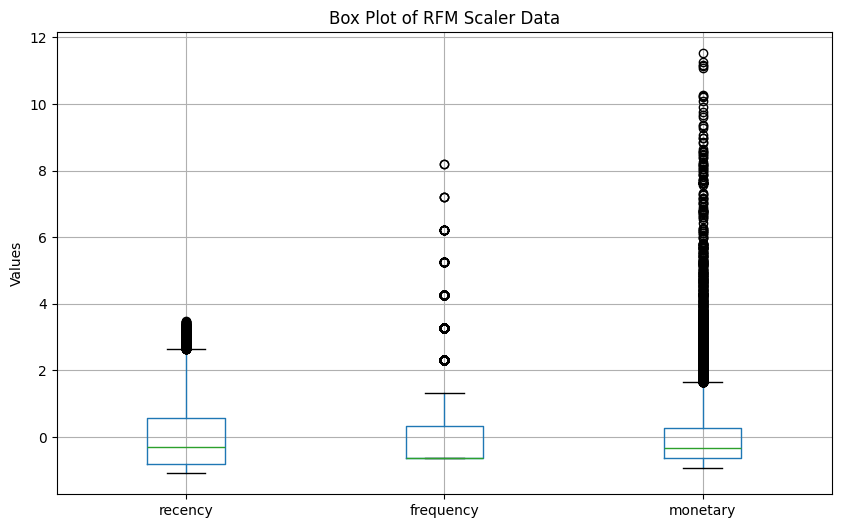
Hình 10: Phân phối tần suất mua

Tần suất mua hàng có dạng phân phối rời rạc và lệch phải rõ rệt. Đa phần khách hàng chỉ mua từ 1 đến 2 lần, trong đó giá trị frequency = 1 chiếm tỷ trọng cao nhất, với mật độ lên đến gần 2. Điều này phản ánh hành vi một lần mua rồi ngưng – một vấn đề phổ biến trong thương mại điện tử. Doanh nghiệp nên thiết kế các chương trình như tích điểm, giảm giá cho lần mua kế tiếp hoặc giới thiệu sản phẩm tương tự để gia tăng tần suất mua hàng. Ngoài ra, có những "đỉnh nhỏ" tại các giá trị 2, 3, 4,... thể hiện các nhóm khách hàng trung thành hơn – đây là nhóm nên được ưu tiên giữ chân bằng chính sách hậu mãi hoặc chương trình khách hàng thân thiết.



Hình 11: Phân phối số tiền chi tiêu của khách hàng

Phân phối giá trị chi tiêu cũng có dạng lệch phải. Phần lớn khách hàng chi tiêu ở mức rất thấp (0–200), trong khi một nhóm nhỏ khách hàng có mức chi tiêu lớn hơn (trên 200). Mật độ cao nhất nằm ở khoảng từ 0 đến 50, sau đó giảm dần. Những khách hàng có monetary cao nên được xác định rõ để tập trung chăm sóc, cá nhân hóa trải nghiệm, hoặc mời tham gia các chương trình VIP. Trong khi đó, phần lớn khách hàng chi tiêu thấp có thể được khai thác thêm qua khuyến mãi, combo sản phẩm, hoặc upsell.

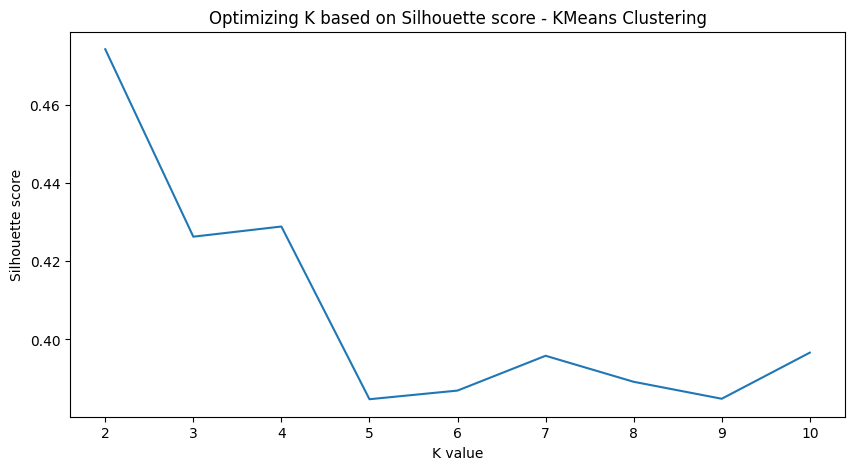


Hình 12: Box Plot dữ liệu RFM sau khi scale

Nhìn chung, cả ba biến đều cho thấy một sự phân bố không đối xứng, nghiêng mạnh về phía các giá trị thấp, với một "đuôi" dài của các giá trị cao hơn thể hiện qua sự xuất hiện dày đặc của các điểm ngoại lai.

Đối với biến "recency", phần lớn dữ liệu tập trung ở các giá trị rất thấp (gần 0), cho thấy sau khi chuẩn hóa, đa số khách hàng có thời gian từ lần mua cuối cùng là không đáng kể. Tuy nhiên, một vài điểm ngoại lai nằm phía trên phần râu cho thấy có một nhóm nhỏ khách hàng đã không mua sắm trong một khoảng thời gian dài hơn đáng kể so với mặt bằng chung. Tương tự, biến "frequency" (tần suất mua hàng) cũng hiển thị một hộp hẹp tập trung gần 0, chỉ ra rằng phần lớn khách hàng có tần suất mua thấp. Điểm đáng chú ý là sự xuất hiện của rất nhiều giá trị ngoại lai trải dài từ 2 đến hơn 8, minh chứng cho một nhóm khách hàng nhỏ nhưng có tần suất mua cực kỳ cao. Cuối cùng, biến "monetary" (giá trị chi tiêu) cũng cho thấy sự tập trung dày đặc của dữ liệu ở các giá trị thấp, nhưng lại sở hữu số lượng và phạm vi các giá trị ngoại lai lớn nhất, kéo dài từ 2 đến gần 12. Điều này khẳng định rằng có một nhóm khách hàng "ưu tú" chiếm tỷ lệ nhỏ nhưng đóng góp doanh thu vượt trội so với đại đa số.

Sự lệch dương rõ rệt và lượng lớn các giá trị ngoại lai trong cả ba biến, ngay cả sau khi chuẩn hóa, cho thấy sự chênh lệch lớn giữa nhóm khách hàng thông thường và nhóm khách hàng có giá trị cao.



Hình 13: Điểm sillhouette phân cụm Kmeans

Dựa trên biểu đồ Silhouette score, điểm cao nhất đạt được tại K=2, cho thấy đây là số lượng cụm mang lại sự phân tách rõ ràng nhất cho dữ liệu. Tuy nhiên, việc lựa chọn số lượng cụm tối ưu không chỉ dựa vào việc tìm kiếm điểm Silhouette cao nhất đơn thuần mà còn cần xem xét đến mục tiêu kinh doanh hoặc sự cần thiết của việc phân loại khách hàng chi tiết hơn.

Đối với các doanh nghiệp, hành trình khách hàng thường được chia thành 5 giai đoạn: nhận thức, cân nhắc, mua hàng, sử dụng và duy trì, đồng nghĩa với việc sẽ có những chương trình chăm sóc khách hàng hướng đến các đối tượng khác nhau. Do đó, để phù hợp hơn với bài toán trong doanh nghiệp, nhóm quyết định chọn chia dữ liệu thành 4 cụm để triển khai phân tích cho các nội dung sau.

## 3.4. Xây dựng hệ thống gợi ý sản phẩm

Hệ thống đề xuất đóng vai trò quan trọng trong việc nâng cao trải nghiệm của khách hàng và thúc đẩy hiệu suất kinh doanh bằng cách cá nhân hóa các đề xuất về sản phẩm hoặc nội dung. Hai phương pháp tiếp cận truyền thống được sử dụng rộng rãi là lọc cộng tác (lọc cộng tác) và lọc dựa trên nội dung (lọc dựa trên nội dung). CF tận dụng sở thích và hành vi của những người dùng tương tự để tạo ra các đề xuất nhưng thường bị giới hạn bởi sự thưa thớt dữ liệu và các vấn đề khởi động nguội (Su & Khoshgoftaar, 2009). Mặt khác, CBF tập trung vào các tính năng mục và hồ sơ người dùng để đề xuất các mục tương tự, điều này có thể dẫn đến quá chuyên môn hóa và hạn chế tính mới lạ (Lops, De Gemmis & Semeraro, 2011). Để khắc phục những hạn chế này, các nhà nghiên cứu đã khám phá các hệ thống đề xuất lai kết hợp nhiều kỹ thuật. Các mô hình lai này nhằm mục đích tận dụng điểm mạnh của các phương pháp riêng lẻ đồng thời giảm thiểu điểm yếu của chúng, mang lại các đề xuất chính xác và mạnh mẽ hơn (Burke, 2002).

Trong bối cảnh kinh doanh như thương mại điện tử và dịch vụ kỹ thuật số, các hệ thống lai đặc biệt có lợi vì chúng có thể xử lý nhiều nguồn dữ liệu khác nhau—chẳng hạn như lịch sử duyệt web, mẫu mua hàng và siêu dữ liệu sản phẩm—để đưa ra các khuyến nghị được cá nhân hóa giúp cải thiện sự tương tác và sự hài lòng của khách hàng (Linden, Smith, & York, 2003). Hơn nữa, những tiến bộ gần đây trong học sâu đã làm phong phú thêm các phương pháp lai bằng cách cho phép mô hình hóa các mối quan hệ phức tạp, phi tuyến tính trong tương tác giữa người dùng và mục (Zhang, Yao, Sun, & Tay, 2019).

Trong đồ án này, nhóm xây dựng hệ thống gợi ý sản phẩm cho khách hàng dựa trên mô hình lai giữa lọc cộng tác (Collaborative Filtering - CF), lọc theo nội dung người dùng (Content-Based - CB) và độ phổ biến sản phẩm trong từng cụm người dùng (cluster-based popularity).

Quy trình thực hiện gồm các bước chính sau:

* **Phân cụm người dùng**: Dữ liệu người dùng được gán nhãn *cluster\_id* nhằm nhóm các khách hàng có hành vi tương đồng.
* **Tạo ma trận người dùng - sản phẩm**: Đếm số lượng sản phẩm đã mua để phục vụ cho CF trong từng cụm.
* **Xử lý đặc trưng người dùng**: Chuẩn hóa độ tuổi và biến đổi các thuộc tính phân loại như giới tính, nguồn truy cập thành dạng one-hot để áp dụng cho mô hình CB.
* **Tính độ phổ biến sản phẩm trong cụm**: Xác định sản phẩm được ưa chuộng nhất trong từng cluster dựa trên tổng số lượng đã mua.
* **Tìm người dùng tương tự**:
  + Dựa trên lịch sử mua hàng (CF).
  + Dựa trên đặc điểm người dùng (CB).
  + Kết hợp điểm tương đồng từ hai nguồn trên (trọng số có thể điều chỉnh qua hệ số alpha).
* **Tổng hợp sản phẩm từ người dùng tương tự** và **điểm phổ biến sản phẩm trong cụm** để tính điểm gợi ý cuối cùng.
* **Lọc bỏ các sản phẩm đã mua** và **gợi ý danh sách sản phẩm mới phù hợp nhất** cho người dùng.

# Chương 4: Kết quả mô hình

## 4.1. Kết quả dashboard trực quan hóa

### 4.1.1. Kết quả Dashboard Sales Analysis

|  |  |
| --- | --- |
| Total Revenue (Tổng doanh thu từ tất cả các đơn hàng) | 2.72M |
| Total Orders (Tổng số lượng đơn hàng được đặt) | 125K |
| Average Order Value (Giá trị trung bình mỗi đơn hàng) | 21.69 |
| Profit Margin (Biên lợi nhuận) | 51.92% |

Bảng 10: Kết quả Dashboard Sales Analysis

Đây là 4 chỉ số quan trọng nhất, có khả năng đánh giá tốt nhất cho hiệu suất kinh doanh hiện tại của một doanh nghiệp.

|  |  |  |
| --- | --- | --- |
| Gender | Total Orders | Total Revenue |
| Male | 62,525 | 1.449.385,98 |
| Female | 62,701 | 1.266.397,76 |
| Tổng | 125,226 | 2.715.783,74 |

Bảng 11: Doanh thu theo giới tính

Số đơn hàng từ nam và nữ gần như là tương đương nhau, tuy nhiên đơn hàng từ nam mang lại lượng doanh thu cao hơn.

Biểu đồ *Total Revenue by Traffic Source* cho thấy phương thức *Search* chiếm tới 71,29% trong tất cả các nguồn truy cập, đây chính là kênh đóng vai trò chính và nên được đầu tư SEO để tạo nên hiệu quả ngày càng tốt hơn.

Doanh thu theo tháng từ biểu đồ *Total Revenue by Month* biểu thị điểm cao nhất là tháng 1 (0.33M), những tháng sau đó giảm mạnh nhưng lại tiếp tục tăng dần trở lại từ tháng 5 và không dừng lại cho tới hết tháng 12 (~0.31M). Vậy khoảng thời gian kinh doanh tốt nhất là vào khoảng cuối năm cho tới tháng 1 năm sau. Đồng thời cần tìm hiểu vì sao nửa đầu năm lại có xu hướng giảm mạnh như vậy, từ đó tìm biện pháp để cải thiện.

*Pareto Analysis of Product* cho thấy chỉ có một số sản phẩm top đầu (ví dụ: *The North Face, Nike, etc.*) chiếm phần lớn doanh thu, các sản phẩm còn lại mang tới rất ít giá trị doanh thu. Từ đó cho thấy nên tập trung vào nhóm sản phẩm top 20% để tối đa hóa lợi nhuận.

Những danh mục hàng có doanh thu cao nhất trong biểu đồ *Total Revenue by Product Category* có thể kể đến như *Outerwear* (0.33M), *Jeans* (0.31M), *Sweaters* (0.22M),... Với những danh mục top nên được tăng số lượng mặt hàng lên và đẩy sales ở những danh mục này, nhưng danh mục hàng với doanh thu quá ít có thể cân nhắc để tối ưu lại.

Ba thị trường có doanh thu cao nổi trội nhất theo biểu đồ *Total Revenue by Country* đó là *China* (0.94M), *United States* (0.61M) và *Brasil* (0.4M). Các đề xuất có thể được đưa ra ví dụ như mở rộng thị phần ở các thị trường top và cân nhắc lại việc đầu tư vào những thị trường quá yếu.

### 4.1.2. Kết quả Dashboard Product Analysis

|  |  |
| --- | --- |
| Units Sold (Tổng số sản phẩm đã bán) | 46K |
| Total Revenue (Tổng doanh thu) | 2,72 triệu USD |
| Stock Left (Số lượng tồn kho) | 327K sản phẩm |
| Return Rate (Tỷ lệ hoàn trả) | 39,97% |

Bảng 12: Doanh thu theo sản phẩm

Số lượng tồn kho hiện tại xấp xỉ 327K sản phẩm, chiếm tỷ lệ lớn so với tổng số sản phẩm đã bán (~46K). Điều này có thể dẫn đến lãng phí lưu kho nếu các sản phẩm không được tiêu thụ hiệu quả. Doanh nghiệp nên phân tích thêm tồn kho theo danh mục để có chiến lược xả hàng hợp lý. Tỷ lệ hoàn trả gần 40% là con số khá cao, điều này cảnh báo doanh nghiệp cần xem xét kỹ lưỡng chất lượng sản phẩm, quy trình đóng gói, vận chuyển hoặc chính sách đổi trả.

*Units Sold by Month:* Doanh số cao nhất đạt được vào tháng 1 (6,3K sản phẩm), sau đó giảm mạnh trong quý I, dần ổn định và tăng trở lại từ tháng 6 đến tháng 12, đạt 5,2K sản phẩm vào tháng 12. Điều này cho thấy chu kỳ tiêu dùng tương tự như ở dashboard doanh thu: cao điểm mua sắm vào đầu và cuối năm.

*Filtered Return Rate by Month:* Tỷ lệ hoàn trả dao động quanh mức 38–42%, cao nhất vào tháng 10 (42,34%) và thấp nhất vào tháng 5 (37,01%). Tỷ lệ hoàn trả tăng cao vào cuối năm có thể do số lượng đơn hàng tăng đột biến, dẫn đến sai sót nhiều hơn trong giao hàng hoặc kỳ vọng của khách hàng không được đáp ứng.

Biểu đồ *Total Revenue by Brand* cho thấy các thương hiệu mang lại doanh thu cao nhất ví dụ như *Diesel* (~55K USD), *Calvin Klein* (~49K USD), *7 For All Mankind* (~49K USD),... Vậy các thương hiệu này nên được doanh nghiệp ưu tiên đẩy mạnh quảng bá và phân phối.

Top sản phẩm có doanh thu cao nhất như *The North Face Apex Bionic Soft Shell Jacket - Men's* (7.224 USD) hay *NIKE WOMEN'S PRO COMPRESSION SPORTS BRA* (5.418 USD)... là những sản phẩm chủ lực, nên được đẩy mạnh trong sale và hàng tồn kho, đồng thời thúc đẩy quảng cáo.

Các sản phẩm top chỉ chiếm khoảng 0,1–0,27% tổng doanh thu, điều này phản ánh mức độ phân tán doanh thu giữa nhiều sản phẩm, tuy nhiên cũng nhấn mạnh vai trò then chốt của nhóm sản phẩm top đầu. Các sản phẩm này nên được xem xét về chất lượng, hình ảnh mô tả, kích cỡ, hoặc hướng dẫn sử dụng để giảm tỷ lệ trả hàng và cải thiện trải nghiệm khách hàng.

### 4.1.3. Kết quả Dashboard Customer Analysis

|  |  |
| --- | --- |
| Total Customers (Tổng lượng khách hàng) | 80K |
| Total Orders (Tổng số đơn hàng) | 125K |
| Avg Orders per Customer (Số đơn hàng trung bình trên mỗi khách hàng) | 2,27 |
| Avg Sales per Customer (Giá trị trung bình mỗi khách hàng) | 33,93 USD |

Bảng 13: Số đơn hàng và doanh thu trung bình mỗi khách hang f

Các chỉ số này phản ánh mức độ tương tác và giá trị trung bình của khách hàng đối với doanh nghiệp. Mỗi khách hàng thực hiện trung bình hơn 2 đơn hàng, mang lại doanh thu khoảng 34 USD, cho thấy mức độ quay lại tương đối ổn định.

Tổng số khách hàng nữ: 40.152

Tổng số khách hàng nam: 39.892

Lượng khách hàng giữa hai giới gần như tương đương, tuy nhiên:

* Doanh thu từ khách hàng nam: 1,45 triệu USD (53,37%)
* Doanh thu từ khách hàng nữ: 1,27 triệu USD (46,63%)
* Giá trị trung bình mỗi khách hàng nam: 36,3 USD
* Giá trị trung bình mỗi khách hàng nữ: 31,5 USD

Có thể thấy khách hàng nam mang lại doanh thu cao hơn và giá trị tiêu dùng trung bình lớn hơn, là phân khúc tiềm năng để doanh nghiệp tăng cường tập trung khai thác.

|  |  |
| --- | --- |
| Total Orders by Month | Tháng 1 là thời điểm bùng nổ với 81.000 đơn hàng. Từ tháng 2 trở đi, số đơn hàng mỗi tháng dao động quanh mức 3K–5K. |
| Total Orders by Day | Lượng đơn hàng phân bố khá đều theo ngày trong tháng, dao động quanh mức 1.300–1.600 đơn/ngày, có xu hướng giảm nhẹ về cuối tháng. |
| Total Orders by Hour of Order | Khung giờ cao điểm đặt hàng nằm trong khoảng 8h–16h, đặc biệt là 10h–14h với đỉnh điểm là 8.900 đơn vào lúc 14h. Đây là những khung giờ nên đẩy mạnh chiến dịch quảng bá, ưu đãi và chăm sóc khách hàng. |

Bảng 14: Thời gian, thời điểm mua hàng

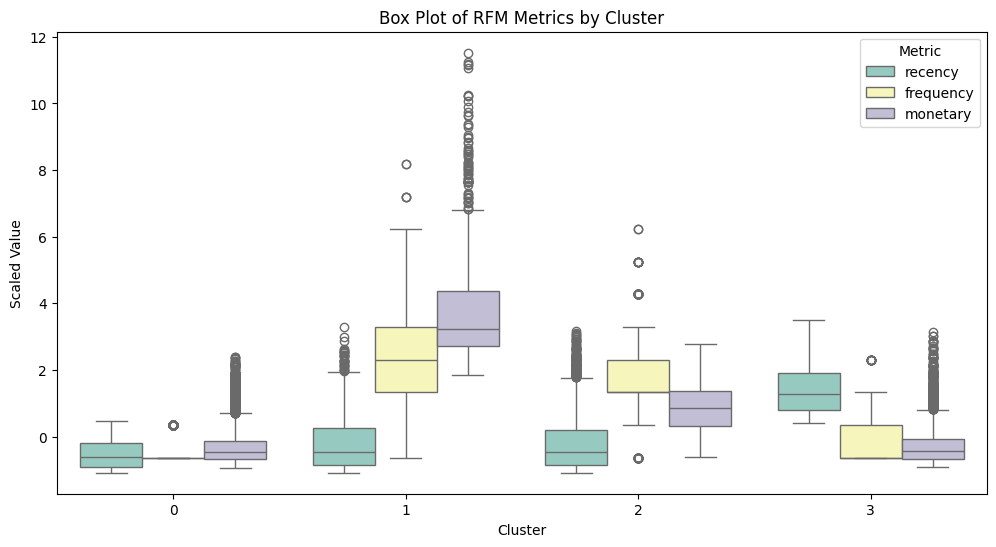
Tương tự như phân tích ở Dashboard *Sales Analysis*, biểu đồ *Customers by Traffic Source* cho thấy *Search* là nguồn truy cập mang lại khách hàng nhiều nhất, khẳng định tầm quan trọng của đầu tư vào *SEO, Google Ads* và chiến lược từ khóa.

Theo biểu đồ *Customers by Age*, khách hàng chủ yếu tập trung ở độ tuổi 20–40, mỗi độ tuổi có trung bình 1.300–1.500 khách. Đặc biệt các nhóm tuổi 25–30 chiếm tỉ lệ cao nhất (~1.450 khách/tuổi), cho thấy đây là nhóm nhân khẩu học trọng tâm cần nhắm tới trong các chiến lược marketing, sản phẩm và chăm sóc khách hàng.

Từ biểu đồ *Customers by Country*, có thể thấy Trung Quốc, Mỹ, Brasil là những thị trường lớn nhất về lượng khách hàng. Các thị trường này cần được ưu tiên triển khai chương trình giữ chân và phát triển khách hàng trung thành, đồng thời mở rộng tệp khách hàng mới tại các quốc gia có tiềm năng như Hàn Quốc, Pháp, Anh,...

Bảng tổng lượng đơn hàng và doanh thu của từng khách hàng giúp doanh nghiệp nhận diện được nhóm khách hàng tiềm năng. Nhóm khách hàng này nên được doanh nghiệp đặc biệt chú trọng giữ chân thông qua các chương trình VIP, khách hàng thân thiết, hoặc ưu đãi cá nhân hóa.

## 4.2. Kết quả phân khúc khách hàng (mô tả các nhóm RFM)



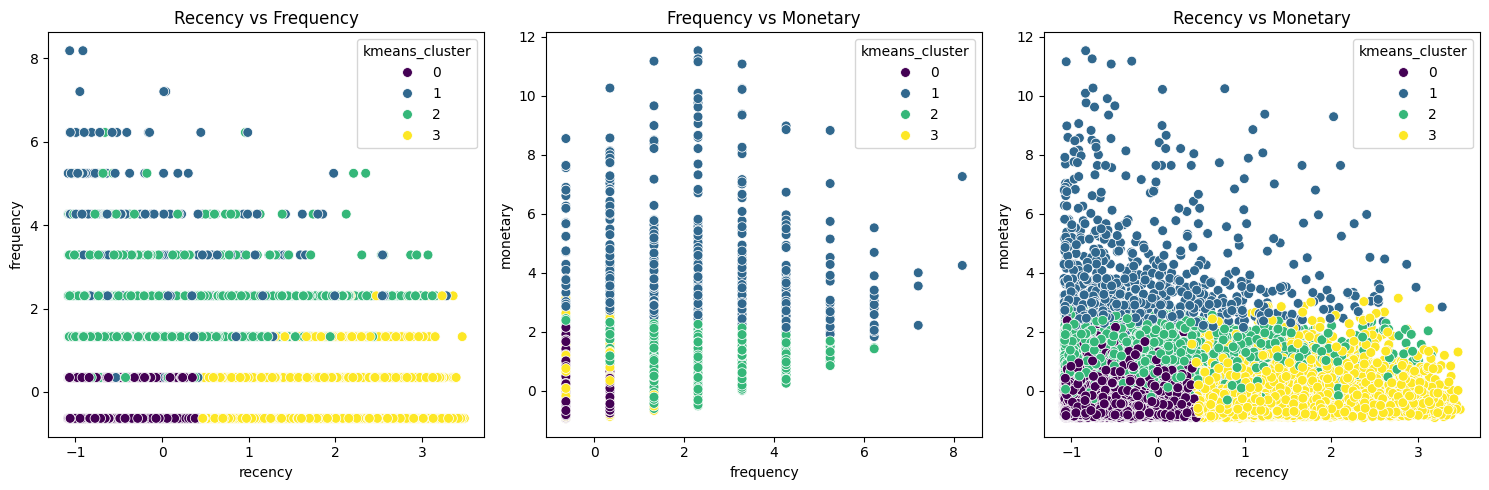
Hình 14: Box Plot chỉ số RFM theo các cụm

Cụm 0 đại diện cho **nhóm khách hàng ít hoạt động** nhất. Các chỉ số Recency, Frequency, và Monetary của cụm này đều tập trung ở các giá trị rất thấp (gần 0), với các hộp hẹp và rất ít giá trị ngoại lai. Cụm này tập trung vào những khách hàng đã lâu không mua hàng, có tần suất mua sắm cực kỳ thấp và giá trị chi tiêu không đáng kể. Nhóm này có thể bao gồm những khách hàng đã ngừng tương tác hoàn toàn với doanh nghiệp hoặc không còn quan tâm đến sản phẩm/dịch vụ.

Cụm 1 là **nhóm khách hàng có giá trị cao và hoạt động tích cực**. Mặc dù giá trị Recency của họ có thể không phải là thấp nhất (tức là không phải là những người mua gần đây nhất), nhưng các chỉ số Frequency và Monetary của cụm này lại đạt mức cao nhất trong số các cụm, với các hộp rộng hơn và đặc biệt là sự hiện diện dày đặc của các giá trị ngoại lai ở phía trên. Điều này chứng tỏ đây là những khách hàng thường xuyên mua sắm và chi tiêu rất hào phóng, có thể được xem là khách hàng khách hàng trung thành, mang lại doanh thu đáng kể cho doanh nghiệp.

Cụm 2 là một **nhóm khách hàng trung gian**, đang có xu hướng giảm tương tác. Biến Recency của cụm này có giá trị trung vị thấp hơn (cho thấy đã lâu hơn kể từ lần mua gần nhất), trong khi Frequency và Monetary nằm ở mức trung bình. Điều này gợi ý rằng đây là những khách hàng từng hoạt động nhưng có thể đang dần ít mua sắm hơn hoặc có nguy cơ rời bỏ. Việc theo dõi sát sao và có các chiến lược giữ chân kịp thời là cần thiết để ngăn chặn họ trở thành khách hàng ở Cụm 0.

Cuối cùng, Cụm 3 là **nhóm khách hàng mới** hoặc những người mua một lần với giá trị thấp. Điểm nổi bật của cụm này là giá trị Recency cao nhất (tức là họ là những người mua gần đây nhất). Tuy nhiên, cả Frequency và Monetary lại rất thấp, với các hộp hẹp và tập trung gần 0. Điều này chỉ ra rằng dù họ mới mua hàng, nhưng tần suất và giá trị chi tiêu của họ không đáng kể. Đây là nhóm khách hàng có tiềm năng để phát triển, cần có các chiến lược để khuyến khích họ mua lại, tăng tần suất và giá trị chi tiêu để trở thành khách hàng có giá trị hơn trong tương lai.



Hình 15: Biểu đồ điểm của các chỉ số RFM theo các cụm

**Cụm 0: Nhóm Khách hàng rời bỏ**

* Trên biểu đồ 1, các điểm Cụm 0 tập trung lại ở góc dưới bên trái.
* Trên biểu đồ 2, chúng nằm ở vùng có tần suất và chi tiêu thấp nhất.
* Và trên biểu đồ 3, chúng cũng cố định ở vùng Recency thấp và Monetary thấp. Điều này nhất quán cho thấy đây là nhóm khách hàng đã ngừng tương tác hoàn toàn với doanh nghiệp, hoặc đã lâu không mua hàng và không có giá trị chi tiêu đáng kể. Họ có thể là những khách hàng đã "ngủ đông" hoặc đã mất.

**Cụm 1: Nhóm Khách hàng Trung Thành**

* Trên biểu đồ 2, các điểm của Cụm 1 chiếm lĩnh vùng có Frequency cao và Monetary cao nhất. Điều này khẳng định họ là những người mua sắm thường xuyên và chi tiêu rất lớn.
* Trên biểu đồ 1 và 3 Cụm 1 phân tán rộng hơn về Recency nhưng vẫn duy trì đặc điểm Frequency và Monetary cao. Điều này cho thấy trong nhóm này có cả những khách hàng vẫn đang hoạt động mạnh mẽ và những khách hàng từng “trung thành” nhưng đã lâu không mua lại. Đây là nhóm khách hàng cốt lõi mang lại doanh thu chính.

**Cụm 2: Nhóm Khách hàng có nguy cơ rời bỏ**

* Trên tất cả các biểu đồ, các điểm của Cụm 2 thường xuất hiện ở vùng có giá trị Recency, Frequency, và Monetary ở mức trung bình.
* Đây có thể là những khách hàng đã từng hoạt động tích cực nhưng đang có dấu hiệu giảm dần tương tác, hoặc là những khách hàng mua sắm không thường xuyên với giá trị chi tiêu vừa phải.

**Cụm 3: Nhóm Khách hàng Mới**

* Trên biểu đồ 1 các điểm Cụm 3 tập trung ở góc trên bên trái, với Recency cao và Frequency thấp.
* Trên biểu đồ 2, chúng nằm cùng với Cụm 0 ở vùng có tần suất và chi tiêu thấp.
* Và trên biểu đồ 3, Cụm 3 nằm ở vùng Recency cao và Monetary thấp. Điều này cho thấy đây là những khách hàng vừa mới thực hiện giao dịch gần đây, nhưng tần suất và giá trị chi tiêu của họ còn rất hạn chế. Họ có thể là khách hàng mới hoàn toàn hoặc những người chỉ mua sản phẩm/dịch vụ một lần với giá trị thấp.

## 4.3. Kết quả hệ thống gợi ý sản phẩm

Kết quả của hệ thống gợi ý là danh sách sản phẩm phù hợp với người dùng đầu vào

|  |
| --- |
| result\_df = show\_recommendations(457, users\_df=df\_users, orders\_df=df\_orders, products\_df=df\_products, top\_n=5)  print(result\_df) |

Đoạn code trên in ra những sản phẩm được đề xuất với người dùng có id 457. Kết quả danh sách đề xuất như sau

Ảnh có chứa văn bản, biên lai, Phông chữ, màu trắng

Nội dung do AI tạo ra có thể không chính xác.

Hình 16: Kết quả danh sách sản phẩm gợi ý

## 4.4. Đánh giá và đề xuất cải tiến

Hiện tại, hệ thống gợi ý xây dựng dựa trên dữ liệu lịch sử người dùng và chưa xét đến tình trạng tồn kho thực tế. Điều này có thể dẫn đến việc đề xuất những sản phẩm đã hết hàng, làm giảm trải nghiệm khách hàng và hiệu quả kinh doanh. Trong các nghiên cứu tương lai, hệ thống nên được tích hợp với dữ liệu từ hệ thống quản lý kho (Inventory Management System) để đảm bảo chỉ gợi ý những sản phẩm còn hàng tại thời điểm tương tác. Ngoài ra, có thể áp dụng chiến lược ưu tiên gợi ý các mặt hàng có tồn kho cao nhằm hỗ trợ giải phóng hàng hóa tồn đọng, từ đó nâng cao hiệu suất quản lý kho và tối ưu doanh thu.

Bên cạnh đó, việc tích hợp hệ thống gợi ý với nền tảng trực quan hóa dữ liệu như Power BI là một hướng mở rộng thiết thực. Sự tích hợp này cho phép các nhà quản lý dễ dàng giám sát hiệu suất hệ thống gợi ý, theo dõi hành vi mua sắm của từng nhóm khách hàng, và đưa ra quyết định kinh doanh kịp thời. Ngoài ra, các dashboard tùy chỉnh trong Power BI có thể được thiết kế để hiển thị sản phẩm đề xuất theo khu vực, nhóm tuổi hoặc phân khúc khách hàng, qua đó hỗ trợ các chiến dịch tiếp thị cá nhân hóa hiệu quả hơn.

Cuối cùng, một hướng nghiên cứu tiềm năng là mở rộng hệ thống theo hướng gợi ý thời gian thực (real-time recommendation). Trong tương lai, hệ thống có thể được tích hợp với các luồng dữ liệu sự kiện như lượt xem sản phẩm, hành vi cuộn, thêm vào giỏ hàng hoặc bỏ sản phẩm – tất cả được xử lý theo thời gian thực nhằm đưa ra gợi ý tức thì. Điều này sẽ giúp tăng khả năng tương tác, giữ chân người dùng và nâng cao tỷ lệ chuyển đổi mua hàng.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64–73.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data, 1(1), 51–59.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331–370.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1), 76–80.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook* (pp. 73–105). Springer.

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009, 421425.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1), 1–38.

Keser, M. (n.d.). Looker eCommerce BigQuery Dataset [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/mustafakeser4/looker-ecommerce-bigquery-dataset?select=users.csv>