**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**Giáo viên hướng dẫn: Gv. Trần Quý Nam**

**Gv. Trang Lê**

**Sinh viên thực hiện:**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN CỤM KHÁCH HÀNG THƯƠNG MẠI ĐIỆN TỬ DỰA TRÊN PHÂN TÍCH RFM VÀ K-MEANS BẰNG PYSPARK**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Mã Sinh Viên** | **Họ và Tên** | **Ngày Sinh** | **Lớp** |
| **1** | 1671020136 | Chu Văn Huy | 20/01/2004 | CNTT 16-05 |
| **2** | 1671020327 | Nguyễn Thành Trung | 19/01/2004 | CNTT 16-05 |
| **3** | 1671020205 | Phạm Ngọc Minh | 10/04/2004 | CNTT 16-05 |

***Hà Nội, năm 2025***

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN CỤM KHÁCH HÀNG THƯƠNG MẠI ĐIỆN TỬ DỰA TRÊN PHÂN TÍCH RFM VÀ K-MEANS BẰNG PYSPARK**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Mã Sinh Viên** | **Họ và Tên** | **Ngày Sinh** | **Điểm** | |
| **Bằng Số** | **Bằng Chữ** |
| **1** | **1671020136** | Chu Văn Huy | 20/01/2004 |  |  |
| **2** | **1671020327** | Nguyễn Thành Trung | 19/01/2004 |  |  |
| **3** | **1671020205** | Phạm Ngọc Minh | 10/04/2004 |  |  |

CÁN BỘ CHẤM THI 1 CÁN BỘ CHẤM THI 2

***Hà Nội, năm 2025***

# **LỜI NÓI ĐẦU**

Trong thời đại bùng nổ dữ liệu và chuyển đổi số, thương mại điện tử đang phát triển mạnh mẽ, tạo ra lượng lớn thông tin về hành vi mua sắm của khách hàng. Việc phân tích và khai thác dữ liệu này đóng vai trò quan trọng trong việc tối ưu hóa chiến lược kinh doanh, nâng cao trải nghiệm khách hàng và tăng trưởng doanh thu. Đề tài "Ứng dụng **PySpark** và **K-Means** trong phân cụm khách hàng thương mại điện tử dựa trên phân tích **RFM**" được thực hiện nhằm giải quyết bài toán phân loại khách hàng một cách hiệu quả. Bằng cách áp dụng mô hình phân tích **RFM** (Recency, Frequency, Monetary) kết hợp với thuật toán phân cụm K-**Means**, chúng em có thể nhóm khách hàng thành các phân khúc cụ thể, giúp doanh nghiệp đưa ra chiến lược tiếp thị và chăm sóc phù hợp.

Việc lựa chọn và xử lý dữ liệu đóng vai trò quan trọng trong quá trình phân tích, đòi hỏi sự kết hợp giữa các kỹ thuật tiền xử lý, khai thác dữ liệu và thuật toán máy học. Trong dự án này, chúng em sử dụng PySpark để xử lý dữ liệu lớn một cách hiệu quả, đồng thời áp dụng thuật toán K-Means để phân cụm khách hàng theo hành vi mua sắm. Mô hình này không chỉ giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình mà còn tối ưu hóa các chiến dịch tiếp thị và giữ chân khách hàng.

Mục tiêu chính của đề tài không chỉ là phát triển một hệ thống phân cụm có độ chính xác cao mà còn cung cấp các thông tin có giá trị cho doanh nghiệp trong lĩnh vực thương mại điện tử. Chúng em xin gửi lời cảm ơn sâu sắc đến thầy/cô (tên giảng viên) đã hướng dẫn và truyền đạt những kiến thức quý báu, giúp chúng em hoàn thành nghiên cứu này. Đề tài không chỉ mang tính học thuật mà còn là một cơ hội để chúng em áp dụng lý thuyết vào thực tế, góp phần nâng cao hiểu biết về lĩnh vực phân tích dữ liệu và ứng dụng công nghệ trong kinh doanh.

# **MỤC LỤC**

[**LỜI NÓI ĐẦU** 3](#_Toc190295643)

[**MỤC LỤC** 4](#_Toc190295644)

[**CHƯƠNG I. TỔNG QUAN CƠ SỞ LÝ THUYẾT** 8](#_Toc190295645)

[**1. RFM** 8](#_Toc190295646)

[**1.1 Phân tích RFM** 8](#_Toc190295647)

[1.1.1 Khái niệm về phân tích RFM 8](#_Toc190295648)

[1.1.2 Lợi ích của phân tích RFM 8](#_Toc190295649)

[**1.2. Áp dụng phân tích RFM vào tập dữ liệu Online Retail** 9](#_Toc190295650)

[1.2.1. Tính toán các chỉ số RFM 9](#_Toc190295651)

[1.2.2. Chuẩn hóa dữ liệu RFM 10](#_Toc190295652)

[**1.3. Phân cụm khách hàng dựa trên RFM** 10](#_Toc190295653)

[1.3.1. Áp dụng thuật toán K-Means 10](#_Toc190295654)

[**1.4. Kết luận** 10](#_Toc190295655)

[**2. Phân cụm khách hàng** 11](#_Toc190295656)

[**2.1 Phân tích phân cụm khách hàng** 11](#_Toc190295657)

[***2.1.1 Khái niệm về phân cụm khách hàng*** 11](#_Toc190295658)

[2.1.2. Lợi ích của phân cụm khách hàng 11](#_Toc190295659)

[**2.2. Thuật toán phân cụm K-Means** 12](#_Toc190295660)

[2.2.1. Nguyên lý hoạt động của K-Means 12](#_Toc190295661)

[2.2.2. Xác định số lượng cụm K tối ưu 12](#_Toc190295662)

[**2.3. Áp dụng K-Means vào phân cụm khách hàng dựa trên RFM** 12](#_Toc190295663)

[2.3.1. Tiền xử lý dữ liệu 12](#_Toc190295664)

[2.3.2. Thực hiện phân cụm K-Means 12](#_Toc190295665)

[2.3.3. Phân tích đặc điểm các cụm khách hàng 13](#_Toc190295666)

[**2.4. Trực quan hóa kết quả phân cụm** 13](#_Toc190295667)

[**2.5. Kết luận** 13](#_Toc190295668)

[**3. PySpark** 14](#_Toc190295669)

[**3.1. Tổng quan về PySpark** 14](#_Toc190295670)

[**3.2. Lý do sử dụng PySpark trong đề tài** 14](#_Toc190295671)

[**3.3. Kết luận** 15](#_Toc190295672)

[**4. Ứng dụng trong thương mại điện tử** 15](#_Toc190295673)

[**4.1. Tầm quan trọng của phân tích khách hàng trong thương mại điện tử** 15](#_Toc190295674)

[**4.2. Ứng dụng cụ thể trong thương mại điện tử** 15](#_Toc190295675)

[4.2.1. Cá nhân hóa chiến dịch marketing 15](#_Toc190295676)

[4.2.2. Dự đoán và giảm tỷ lệ khách hàng rời bỏ (Churn Prediction) 16](#_Toc190295677)

[4.2.3. Gợi ý sản phẩm thông minh (Recommendation System) 16](#_Toc190295678)

[4.2.4. Tối ưu hóa chiến lược giá và khuyến mãi 17](#_Toc190295679)

[4.2.5. Nâng cao hiệu suất quản lý hàng tồn kho 17](#_Toc190295680)

[**4.3. Kết luận** 18](#_Toc190295681)

[**CHƯƠNG II : MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG** 19](#_Toc190295682)

[**1. Mô tả tập dữ liệu** 19](#_Toc190295683)

[**1.1. Giới thiệu về tập dữ liệu** 19](#_Toc190295684)

[**1.2. Các cột dữ liệu quan trọng** 19](#_Toc190295685)

[**1.3. Xử lý dữ liệu trước khi phân tích** 19](#_Toc190295686)

[**2. Công nghệ sử dụng** 20](#_Toc190295687)

[**2.1. Apache Spark và PySpark** 20](#_Toc190295688)

[**2.2. Thư viện Python hỗ trợ** 21](#_Toc190295689)

[**2.3. Xử lý dữ liệu bằng PySpark** 21](#_Toc190295690)

[2.3.1. Đọc dữ liệu 21](#_Toc190295691)

[**2.4. Tiền xử lý dữ liệu** 21](#_Toc190295692)

[2.4.1. Loại bỏ hóa đơn bị hủy 21](#_Toc190295693)

[2.4.2. Chuyển đổi ngày tháng 22](#_Toc190295694)

[**2.5. Tính toán các chỉ số RFM** 22](#_Toc190295695)

[**2.5.1. Recency (R)** 22](#_Toc190295696)

[2.5.2. Frequency (F) 22](#_Toc190295697)

[2.5.3. Monetary (M) 23](#_Toc190295698)

[**2.6. Phân cụm khách hàng với K-Means** 23](#_Toc190295699)

[2.6.1. Chuẩn hóa dữ liệu 23](#_Toc190295700)

[2.6.2. Tìm số cụm tối ưu bằng Elbow Method 23](#_Toc190295701)

[**3. các bước xử lý dữ liệu và phân tích** 24](#_Toc190295702)

[Tóm tắt các bước xử lý dữ liệu và phân tích 24](#_Toc190295703)

[**3.1. Xử lý dữ liệu (Data Preprocessing)** 24](#_Toc190295704)

[**3.2. Phân tích dữ liệu (Data Analysis)** 27](#_Toc190295705)

[**CHƯƠNG III : KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG** 30](#_Toc190295706)

[**3.1 Kết quả đạt được** 30](#_Toc190295707)

[**KẾT LUẬN** 31](#_Toc190295708)

[**LỜI CẢM ƠN** 32](#_Toc190295709)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 33](#_Toc190295710)

**CHƯƠNG I. TỔNG QUAN CƠ SỞ LÝ THUYẾT**

# **1. RFM**

## **1.1 Phân tích RFM**

### 1.1.1 Khái niệm về phân tích RFM

Phân tích RFM (**Recency - Frequency - Monetary**) là một kỹ thuật phổ biến trong lĩnh vực tiếp thị và phân tích dữ liệu khách hàng. RFM giúp doanh nghiệp phân nhóm khách hàng dựa trên hành vi mua sắm của họ, từ đó tối ưu hóa chiến lược tiếp cận và tăng cường hiệu quả kinh doanh.

Ba yếu tố chính trong mô hình RFM:

* **Recency (R) - Gần đây**: Khoảng thời gian kể từ lần mua hàng gần nhất của khách hàng.
  + Khách hàng có giá trị Recency thấp (tức là vừa mua gần đây) thường có khả năng tiếp tục mua sắm cao hơn.
* **Frequency (F) - Tần suất**: Số lần khách hàng thực hiện giao dịch trong một khoảng thời gian nhất định.
  + Khách hàng có tần suất mua hàng cao thường là khách hàng trung thành.
* **Monetary (M) - Giá trị tiền tệ**: Tổng số tiền khách hàng đã chi tiêu trong suốt thời gian giao dịch.
  + Khách hàng có giá trị Monetary cao là những khách hàng mang lại doanh thu lớn cho doanh nghiệp.

### 1.1.2 Lợi ích của phân tích RFM

* **Xác định khách hàng tiềm năng**: Giúp doanh nghiệp nhận diện khách hàng trung thành, khách hàng có giá trị cao.
* **Tối ưu hóa chiến lược marketing**: Dựa vào nhóm khách hàng, doanh nghiệp có thể đưa ra các chương trình khuyến mãi phù hợp.
* **Cải thiện chiến dịch chăm sóc khách hàng**: Tập trung vào nhóm khách hàng có nguy cơ rời bỏ để đưa ra chính sách giữ chân.

## **1.2. Áp dụng phân tích RFM vào tập dữ liệu Online Retail**

### 1.2.1. Tính toán các chỉ số RFM

Dựa vào tập dữ liệu **Online Retail**, ta tính toán từng thành phần của RFM như sau:

1. **Recency (R) - Ngày mua hàng gần nhất**
   * Xác định ngày giao dịch gần nhất trong tập dữ liệu.
   * Tính khoảng cách thời gian từ ngày giao dịch gần nhất của từng khách hàng đến thời điểm phân tích.
   * Công thức: R=Ngaˋy phaˆn tıˊch−Ngaˋy mua haˋng gaˆˋn nhaˆˊtR = \text{Ngày phân tích} - \text{Ngày mua hàng gần nhất}R=Ngaˋy phaˆn tıˊch−Ngaˋy mua haˋng gaˆˋn nhaˆˊt
2. **Frequency (F) - Số lần mua hàng**
   * Đếm số lượng hóa đơn (InvoiceNo) của từng khách hàng.
   * Công thức: F=Soˆˊ lượng giao dịch của khaˊch haˋngF = \text{Số lượng giao dịch của khách hàng}F=Soˆˊ lượng giao dịch của khaˊch haˋng
3. **Monetary (M) - Tổng giá trị giao dịch**
   * Tính tổng số tiền mà mỗi khách hàng đã chi tiêu dựa trên công thức: M=∑(Quantity×UnitPrice)M = \sum (\text{Quantity} \times \text{UnitPrice})M=∑(Quantity×UnitPrice)
   * Giá trị này giúp xác định khách hàng có mức chi tiêu cao nhất.

### 1.2.2. Chuẩn hóa dữ liệu RFM

* Sau khi tính toán các chỉ số RFM, dữ liệu cần được chuẩn hóa để đảm bảo rằng giá trị của các biến có cùng trọng số khi thực hiện phân cụm.
* Kỹ thuật **StandardScaler** được sử dụng để đưa các giá trị về cùng một quy mô.

## **1.3. Phân cụm khách hàng dựa trên RFM**

### 1.3.1. Áp dụng thuật toán K-Means

Sau khi tính toán RFM và chuẩn hóa dữ liệu, thuật toán **K-Means Clustering** sẽ được áp dụng để phân nhóm khách hàng.

Các bước chính:

1. **Chuẩn bị dữ liệu**: Dữ liệu RFM sau khi chuẩn hóa sẽ được đưa vào mô hình.
2. **Xác định số cụm K tối ưu**:
   * Sử dụng phương pháp **Elbow Method** để tìm số cụm hợp lý.
   * Tính toán **Silhouette Score** để đánh giá chất lượng phân cụm.
3. **Huấn luyện mô hình K-Means**: Phân chia khách hàng thành các cụm dựa trên điểm RFM.
4. **Phân tích cụm khách hàng**: Xác định đặc điểm của từng nhóm khách hàng và đề xuất chiến lược kinh doanh phù hợp.

## **1.4. Kết luận**

Phân tích RFM là một phương pháp mạnh mẽ giúp doanh nghiệp hiểu rõ hành vi khách hàng dựa trên dữ liệu giao dịch. Bằng cách sử dụng **PySpark** để xử lý dữ liệu lớn và **K-Means Clustering** để phân cụm, doanh nghiệp có thể đưa ra các chiến lược marketing chính xác hơn, tối ưu hóa doanh thu và nâng cao trải nghiệm khách hàng.

# **2. Phân cụm khách hàng**

## **2.1 Phân tích phân cụm khách hàng**

***2.1.1 Khái niệm về phân cụm khách hàng***

Phân cụm khách hàng là quá trình nhóm các khách hàng có đặc điểm hành vi mua sắm tương đồng thành các nhóm riêng biệt. Mục tiêu của phân cụm là giúp doanh nghiệp hiểu rõ hơn về khách hàng, từ đó tối ưu hóa chiến lược tiếp thị và quản lý quan hệ khách hàng.

Trong bài toán này, phân cụm khách hàng được thực hiện dựa trên phương pháp **phân tích RFM (Recency - Frequency - Monetary)**, giúp doanh nghiệp chia khách hàng thành các nhóm dựa trên:

* **Recency (R):** Khoảng thời gian kể từ lần mua hàng gần nhất.
* **Frequency (F):** Số lần khách hàng thực hiện giao dịch trong một khoảng thời gian nhất định.
* **Monetary (M):** Tổng số tiền khách hàng đã chi tiêu trong suốt thời gian giao dịch.

Dữ liệu RFM sau khi được tính toán sẽ được đưa vào thuật toán **K-Means Clustering** để thực hiện phân cụm khách hàng.

### 2.1.2. Lợi ích của phân cụm khách hàng

* **Cá nhân hóa chiến lược marketing:** Phân loại khách hàng thành các nhóm để đưa ra chiến dịch marketing phù hợp.
* **Tối ưu hóa chăm sóc khách hàng:** Xây dựng các chương trình ưu đãi dành riêng cho từng nhóm khách hàng.
* **Gia tăng doanh thu:** Tập trung vào nhóm khách hàng có giá trị cao để tối đa hóa lợi nhuận.
* **Dự đoán hành vi khách hàng:** Hiểu được xu hướng mua sắm để đưa ra các chiến lược kinh doanh hợp lý.

## **2.2. Thuật toán phân cụm K-Means**

**K-Means Clustering** là một thuật toán phân cụm phổ biến trong Machine Learning, hoạt động theo cơ chế chia dữ liệu thành **K nhóm** dựa trên khoảng cách giữa các điểm dữ liệu.

### 2.2.1. Nguyên lý hoạt động của K-Means

1. **Chọn số cụm (K) ban đầu.**
2. **Khởi tạo K tâm cụm ngẫu nhiên.**
3. **Phân loại các điểm dữ liệu vào cụm gần nhất** dựa trên khoảng cách Euclidean.
4. **Tính lại tâm cụm** bằng cách lấy trung bình của các điểm dữ liệu trong mỗi cụm.
5. **Lặp lại các bước trên cho đến khi thuật toán hội tụ**, tức là khi các tâm cụm không thay đổi đáng kể.

### 2.2.2. Xác định số lượng cụm K tối ưu

Số cụm K có thể được xác định bằng:

* **Phương pháp Elbow Method:**
  + Tính tổng bình phương khoảng cách từ mỗi điểm đến tâm cụm.
  + Chọn K sao cho độ giảm chi phí phân cụm không còn đáng kể.
* **Silhouette Score:** Đánh giá mức độ gắn kết trong mỗi cụm và mức độ phân tách giữa các cụm.

## **2.3. Áp dụng K-Means vào phân cụm khách hàng dựa trên RFM**

### 2.3.1. Tiền xử lý dữ liệu

* Tính toán các chỉ số **RFM** từ tập dữ liệu Online Retail.
* Chuẩn hóa dữ liệu bằng **StandardScaler** để đảm bảo các giá trị có cùng trọng số.

### 2.3.2. Thực hiện phân cụm K-Means

* Dữ liệu RFM sau khi chuẩn hóa được đưa vào thuật toán K-Means.
* Chạy thuật toán với nhiều giá trị K khác nhau và sử dụng **Elbow Method** để chọn K tối ưu.
* Sau khi có số cụm K phù hợp, tiến hành huấn luyện mô hình K-Means trên dữ liệu RFM.

### 2.3.3. Phân tích đặc điểm các cụm khách hàng

Sau khi phân cụm, ta sẽ phân tích đặc điểm của từng cụm để hiểu rõ hành vi mua sắm của khách hàng:

* **Cụm khách hàng trung thành:** Khách hàng có tần suất mua hàng cao và giá trị chi tiêu lớn.
* **Cụm khách hàng mới:** Khách hàng vừa mới mua hàng lần đầu tiên.
* **Cụm khách hàng tiềm năng:** Khách hàng có chi tiêu cao nhưng chưa mua thường xuyên.
* **Cụm khách hàng có nguy cơ rời bỏ:** Khách hàng đã lâu không quay lại mua hàng.

## **2.4. Trực quan hóa kết quả phân cụm**

* **Vẽ biểu đồ phân bố khách hàng** theo từng chỉ số RFM để quan sát sự khác biệt giữa các cụm.
* **Sử dụng seaborn và matplotlib** để vẽ biểu đồ cột so sánh giá trị trung bình R, F, M của từng nhóm khách hàng.

## **2.5. Kết luận**

Phân cụm khách hàng bằng **phân tích RFM và K-Means Clustering** giúp doanh nghiệp xác định các nhóm khách hàng khác nhau dựa trên hành vi mua sắm của họ. Kết quả phân cụm cung cấp nền tảng quan trọng để tối ưu hóa chiến lược tiếp thị, cải thiện trải nghiệm khách hàng và nâng cao hiệu quả kinh doanh.

# **3. PySpark**

## **3.1. Tổng quan về PySpark**

**PySpark** là một thư viện Python của Apache Spark, được sử dụng để xử lý dữ liệu lớn (Big Data) và thực hiện các tác vụ phân tích dữ liệu, máy học (Machine Learning) trên môi trường phân tán. PySpark cung cấp giao diện lập trình mạnh mẽ giúp thao tác với dữ liệu lớn một cách hiệu quả hơn so với các thư viện xử lý dữ liệu truyền thống như Pandas hay NumPy.

Apache Spark là một framework xử lý dữ liệu lớn có khả năng tính toán phân tán, hỗ trợ nhiều mô hình lập trình như:

* **Spark SQL:** Xử lý dữ liệu dạng bảng với cú pháp SQL.
* **Spark DataFrame:** Xử lý dữ liệu dạng bảng tương tự Pandas nhưng tối ưu hơn cho dữ liệu lớn.
* **Spark MLlib:** Hỗ trợ các thuật toán Machine Learning phân tán.
* **Spark Streaming:** Xử lý dữ liệu theo thời gian thực.

## **3.2. Lý do sử dụng PySpark trong đề tài**

Trong bài toán **phân cụm khách hàng thương mại điện tử dựa trên phân tích RFM**, tập dữ liệu có thể rất lớn, chứa hàng triệu giao dịch từ các khách hàng khác nhau. Sử dụng **PySpark** giúp:

* **Xử lý dữ liệu nhanh chóng** trên các cụm máy tính phân tán.
* **Tiết kiệm bộ nhớ** do PySpark sử dụng cơ chế lazy evaluation và tối ưu hóa xử lý dữ liệu.
* **Tích hợp dễ dàng với SQL và các công nghệ dữ liệu lớn** như Hadoop, HDFS.
* **Hỗ trợ Machine Learning phân tán** thông qua thư viện **MLlib**.

## **3.3. Kết luận**

**PySpark** là công cụ mạnh mẽ giúp xử lý và phân tích dữ liệu lớn một cách hiệu quả. Trong đề tài này, PySpark được sử dụng để:

* **Tiền xử lý dữ liệu thương mại điện tử** với DataFrame API.
* **Tính toán chỉ số RFM** để phân tích hành vi khách hàng.
* **Áp dụng K-Means Clustering trên tập dữ liệu lớn** với Spark MLlib.
* **Trực quan hóa kết quả phân cụm** để hỗ trợ ra quyết định kinh doanh.

Việc sử dụng PySpark giúp tối ưu hóa hiệu suất tính toán, đặc biệt khi làm việc với dữ liệu lớn, đồng thời đảm bảo độ chính xác và hiệu quả trong quá trình phân tích khách hàng.

# **4. Ứng dụng trong thương mại điện tử**

## **4.1. Tầm quan trọng của phân tích khách hàng trong thương mại điện tử**

Trong bối cảnh thương mại điện tử phát triển mạnh mẽ, việc **hiểu rõ hành vi khách hàng** là yếu tố quan trọng giúp doanh nghiệp tối ưu chiến lược kinh doanh, cá nhân hóa trải nghiệm khách hàng và tăng doanh thu.

Phân tích **RFM (Recency, Frequency, Monetary)** và **phân cụm khách hàng** là hai phương pháp phổ biến giúp doanh nghiệp:

* Nhận diện nhóm khách hàng trung thành và có giá trị cao.
* Dự đoán hành vi mua sắm và xây dựng chiến lược tiếp thị phù hợp.
* Phát hiện khách hàng có nguy cơ rời bỏ để có chính sách giữ chân hiệu quả.
* Tối ưu hóa chính sách giá cả, khuyến mãi và chương trình khách hàng thân thiết.

## **4.2. Ứng dụng cụ thể trong thương mại điện tử**

### 4.2.1. Cá nhân hóa chiến dịch marketing

* **Khách hàng trung thành** (tần suất mua cao, giá trị đơn hàng lớn) có thể nhận các ưu đãi VIP hoặc chương trình khách hàng thân thiết.
* **Khách hàng mới** cần được khuyến khích bằng mã giảm giá hoặc ưu đãi chào mừng.
* **Khách hàng ít quay lại** có thể nhận email nhắc nhở hoặc gói giảm giá đặc biệt để kích thích họ mua sắm trở lại.

Ví dụ:

* Nhóm khách hàng có **Recency thấp, Frequency cao, Monetary cao** (khách hàng VIP) sẽ được ưu tiên tiếp cận các chương trình độc quyền.
* Nhóm khách hàng có **Recency cao, Frequency thấp** (khách hàng cũ lâu không quay lại) sẽ được nhắm mục tiêu bằng email marketing hoặc ưu đãi đặc biệt để thu hút họ quay trở lại.

### 4.2.2. Dự đoán và giảm tỷ lệ khách hàng rời bỏ (Churn Prediction)

Phân cụm khách hàng giúp phát hiện **nhóm khách hàng có nguy cơ rời bỏ cao**. Doanh nghiệp có thể:

* Cung cấp ưu đãi hoặc khuyến mãi để giữ chân họ.
* Gửi email hoặc thông báo nhắc nhở về các sản phẩm yêu thích của họ.
* Điều chỉnh chính sách giao hàng, chăm sóc khách hàng để nâng cao trải nghiệm.

Ví dụ:  
Nếu một nhóm khách hàng có **giá trị mua hàng thấp và tần suất mua giảm dần**, doanh nghiệp có thể áp dụng các chiến dịch như:

* Gửi email với nội dung “Chúng tôi nhớ bạn! Quay lại ngay để nhận ưu đãi đặc biệt.”
* Đề xuất sản phẩm phù hợp với lịch sử mua hàng của khách hàng.

### 4.2.3. Gợi ý sản phẩm thông minh (Recommendation System)

Bằng cách hiểu rõ hành vi mua sắm, doanh nghiệp có thể cải thiện hệ thống **gợi ý sản phẩm cá nhân hóa** dựa trên nhóm khách hàng:

* **Nhóm khách hàng có chi tiêu cao** → Gợi ý sản phẩm cao cấp, phiên bản đặc biệt.
* **Nhóm khách hàng mua sắm thường xuyên** → Đề xuất gói giảm giá, combo tiết kiệm.
* **Nhóm khách hàng mới** → Hiển thị sản phẩm phổ biến, dễ tiếp cận.

Ví dụ:

* Nếu một khách hàng thuộc nhóm **tích cực mua sắm nhưng chi tiêu trung bình**, hệ thống có thể gợi ý **mua theo combo** để gia tăng giá trị giỏ hàng.
* Nếu khách hàng **chỉ mua một nhóm sản phẩm cụ thể**, doanh nghiệp có thể giới thiệu **sản phẩm bổ sung** dựa trên phân tích RFM.

### 4.2.4. Tối ưu hóa chiến lược giá và khuyến mãi

Phân cụm khách hàng giúp doanh nghiệp xác định nhóm khách hàng nhạy cảm với giá cả và xây dựng chiến lược giá hợp lý:

* **Khách hàng chi tiêu cao nhưng mua ít lần** có thể được khuyến khích bằng voucher giảm giá cho lần mua tiếp theo.
* **Khách hàng có tần suất mua cao nhưng chi tiêu thấp** có thể được đề xuất chương trình mua sắm theo gói hoặc combo.

Ví dụ:  
Một khách hàng có xu hướng mua **các sản phẩm giá rẻ nhưng tần suất cao**, doanh nghiệp có thể cung cấp **chính sách giao hàng miễn phí cho đơn hàng trên một mức nhất định** để tăng giá trị đơn hàng.

### 4.2.5. Nâng cao hiệu suất quản lý hàng tồn kho

Hiểu rõ hành vi khách hàng giúp doanh nghiệp dự đoán nhu cầu sản phẩm chính xác hơn:

* Nhóm khách hàng có **tần suất mua cao nhưng chi tiêu thấp** có thể chỉ ra sản phẩm phổ biến cần được bổ sung hàng thường xuyên.
* Nhóm khách hàng **chỉ mua vào các dịp đặc biệt** giúp doanh nghiệp tối ưu kế hoạch nhập hàng theo mùa vụ.

Ví dụ:  
Nếu một nhóm khách hàng có xu hướng mua **các sản phẩm trang trí vào mùa Giáng sinh**, doanh nghiệp có thể tối ưu hóa kho hàng và tung ra chương trình khuyến mãi vào thời điểm phù hợp.

## **4.3. Kết luận**

Ứng dụng **phân tích RFM và phân cụm khách hàng bằng PySpark** trong thương mại điện tử mang lại nhiều lợi ích quan trọng:

* **Cá nhân hóa chiến dịch marketing** để tăng tỷ lệ chuyển đổi.
* **Dự đoán và giảm tỷ lệ rời bỏ khách hàng** để tối ưu lợi nhuận.
* **Cải thiện hệ thống gợi ý sản phẩm** giúp nâng cao trải nghiệm người dùng.
* **Tối ưu hóa giá cả và chiến lược khuyến mãi** để tăng doanh số.
* **Quản lý hàng tồn kho hiệu quả hơn**, giảm chi phí lưu kho.

Việc áp dụng các phương pháp khoa học dữ liệu vào thương mại điện tử không chỉ giúp doanh nghiệp nâng cao hiệu quả kinh doanh mà còn tạo ra lợi thế cạnh tranh trong thị trường đầy thách thức hiện nay.

# **+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++CHƯƠNG II : MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG**

# **1. Mô tả tập dữ liệu**

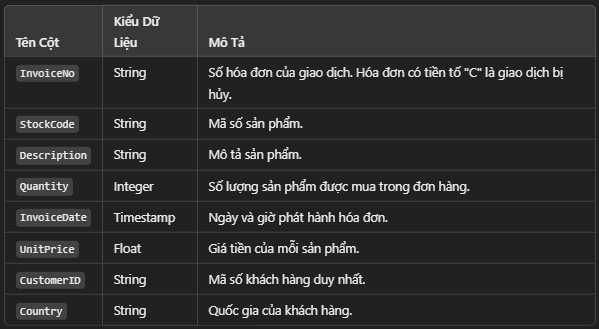
## **1.1. Giới thiệu về tập dữ liệu**

Tập dữ liệu **OnlineRetail.csv** là một tập dữ liệu giao dịch mua hàng trực tuyến của một công ty bán lẻ tại Anh trong khoảng thời gian từ **01/12/2010 đến 09/12/2011**. Dữ liệu này chứa thông tin chi tiết về các đơn hàng, bao gồm mã khách hàng, sản phẩm, số lượng, giá cả và thời gian mua hàng.

Tập dữ liệu này phù hợp để thực hiện phân tích **RFM (Recency, Frequency, Monetary)** nhằm phân loại khách hàng và tối ưu hóa chiến lược marketing.

## **1.2. Các cột dữ liệu quan trọng**

Tệp CSV chứa các cột sau:



## **1.3. Xử lý dữ liệu trước khi phân tích**

Trong đoạn code PySpark của bạn, dữ liệu được xử lý trước như sau:

1. **Loại bỏ các hóa đơn bị hủy**: Các hóa đơn có InvoiceNo bắt đầu bằng "C" sẽ bị loại bỏ.
2. **Chuyển đổi định dạng ngày tháng**: InvoiceDate được chuyển thành kiểu timestamp để dễ dàng tính toán.
3. **Tạo cột thời gian tham chiếu**: Ngày tham chiếu (from\_date) được đặt là **01/12/2010 08:26:00**.
4. **Tính toán các chỉ số RFM**:
   * **Recency (R)**: Khoảng thời gian từ lần mua gần nhất đến ngày tham chiếu.
   * **Frequency (F)**: Số lần khách hàng mua hàng.
   * **Monetary (M)**: Tổng tiền mà khách hàng đã chi tiêu.

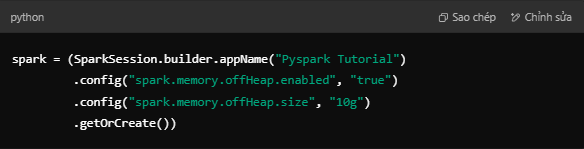
# **2. Công nghệ sử dụng**

Dự án sử dụng các công nghệ và thư viện mạnh mẽ để xử lý dữ liệu lớn và phân tích khách hàng.

## **2.1. Apache Spark và PySpark**

* **Apache Spark**: Một framework xử lý dữ liệu lớn (Big Data) với khả năng tính toán phân tán mạnh mẽ.
* **PySpark**: Giao diện Python của Apache Spark, cho phép sử dụng Spark với Python.
* **SparkSession**: Được sử dụng để tạo môi trường làm việc với Spark.

**Khởi tạo Spark Session**



📌 **Giải thích**:

* **.appName("Pyspark Tutorial")**: Đặt tên cho phiên làm việc với Spark.
* **.config("spark.memory.offHeap.enabled", "true")**: Kích hoạt bộ nhớ ngoài heap để xử lý dữ liệu lớn.
* **.config("spark.memory.offHeap.size", "10g")**: Cấu hình bộ nhớ heap lên 10GB để tối ưu hóa hiệu suất.

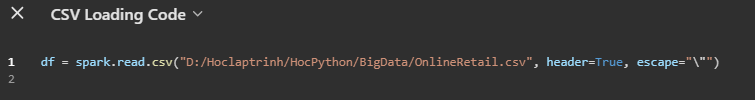
## **2.2. Thư viện Python hỗ trợ**

Dự án sử dụng nhiều thư viện Python để phân tích và trực quan hóa dữ liệu:

* **NumPy & Pandas**: Hỗ trợ xử lý dữ liệu, tính toán ma trận.
* **Matplotlib & Seaborn**: Vẽ biểu đồ trực quan hóa dữ liệu.
* **TQDM**: Hiển thị thanh tiến trình khi chạy vòng lặp.

## **2.3. Xử lý dữ liệu bằng PySpark**

### 2.3.1. Đọc dữ liệu

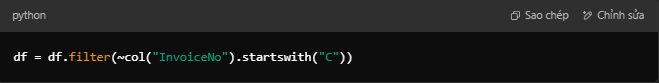


📌 **Giải thích**:

* **read.csv()**: Đọc tập dữ liệu từ file CSV.
* **header=True**: Xác nhận dòng đầu tiên là header.
* **escape="\""**: Xử lý các ký tự đặc biệt trong dữ liệu.

## **2.4. Tiền xử lý dữ liệu**

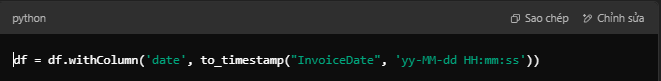
### 2.4.1. Loại bỏ hóa đơn bị hủy



📌 **Giải thích**:

* **filter(~col("InvoiceNo").startswith("C"))**: Lọc bỏ các hóa đơn có mã bắt đầu bằng "C" (bị hủy).

### 2.4.2. Chuyển đổi ngày tháng

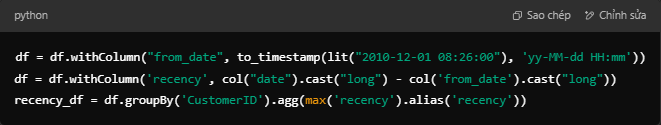


📌 **Giải thích**:

* **to\_timestamp("InvoiceDate", 'yy-MM-dd HH:mm:ss')**: Chuyển đổi cột InvoiceDate thành định dạng timestamp.

## **2.5. Tính toán các chỉ số RFM**

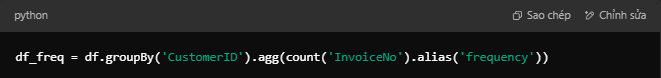
## **2.5.1. Recency (R)**



📌 **Giải thích**:

* **Tạo ngày tham chiếu** (from\_date).
* **Tính khoảng thời gian từ ngày mua gần nhất đến ngày tham chiếu**.

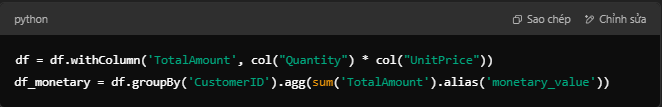
### 2.5.2. Frequency (F)



📌 **Giải thích**:

* **Đếm số lượng đơn hàng của mỗi khách hàng**.

### 2.5.3. Monetary (M)

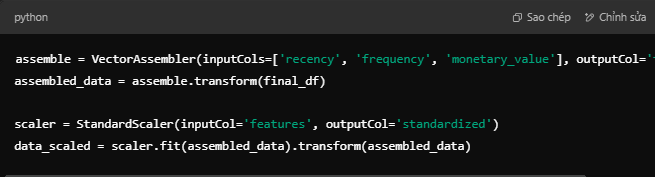


📌 **Giải thích**:

* **Tính tổng số tiền mà khách hàng đã chi tiêu**.

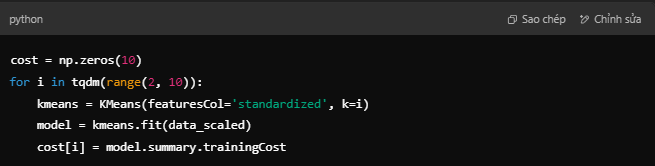
## **2.6. Phân cụm khách hàng với K-Means**

### 2.6.1. Chuẩn hóa dữ liệu

 📌 **Giải thích**:

* **VectorAssembler**: Kết hợp các cột recency, frequency, monetary\_value thành một vector đặc trưng.
* **StandardScaler**: Chuẩn hóa dữ liệu để tránh chênh lệch giữa các thang đo.

### 2.6.2. Tìm số cụm tối ưu bằng Elbow Method



📌 **Giải thích**:

* **Chạy K-Means với K từ 2 đến 9**.
* **Tính toán độ lỗi (training cost)** để vẽ biểu đồ Elbow.

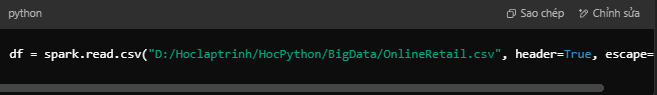
# **3. các bước xử lý dữ liệu và phân tích**

**Tóm tắt các bước xử lý dữ liệu và phân tích**

****

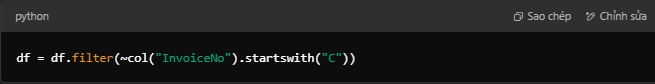
## **3.1. Xử lý dữ liệu (Data Preprocessing)**

**Bước 1: Đọc dữ liệu**

 📌 **Giải thích**:

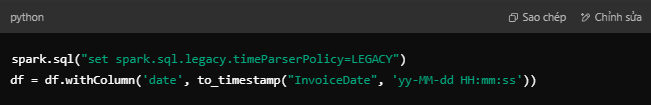
* **read.csv()**: Đọc tập dữ liệu từ file CSV.
* **header=True**: Xác nhận dòng đầu tiên là header.
* **escape="\""**: Xử lý các ký tự đặc biệt trong dữ liệu.

**Bước 2: Loại bỏ hóa đơn bị hủy**

 📌 **Giải thích**:

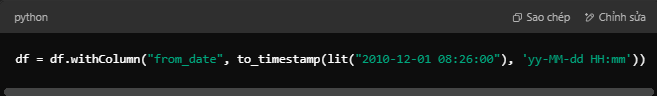
* Loại bỏ các hóa đơn có mã bắt đầu bằng **"C"**, vì đây là các giao dịch bị hủy.

**Bước 3: Chuyển đổi ngày tháng**

 📌 **Giải thích**:

* **to\_timestamp()**: Chuyển đổi cột InvoiceDate sang kiểu timestamp để dễ dàng tính toán.
* **Cấu hình spark.sql.legacy.timeParserPolicy=LEGACY** giúp xử lý các định dạng ngày tháng không chuẩn.

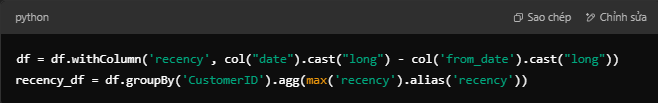
**Bước 4: Tạo cột thời gian tham chiếu**

 📌 **Giải thích**:

* Đặt ngày tham chiếu **01/12/2010 08:26:00** để tính **Recency** (R).

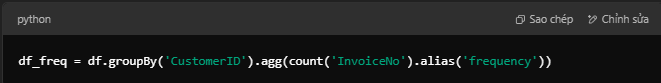
**Bước 5: Tính toán các chỉ số RFM**

**Bước 5.1: Tính chỉ số Recency (R)**

 📌 **Giải thích**:

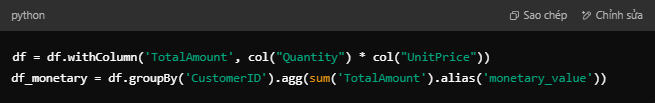
* **Recency (R)** đo khoảng thời gian từ lần mua hàng gần nhất đến ngày tham chiếu.
* Dùng **max('recency')** để lấy lần mua hàng mới nhất của mỗi khách hàng.

**Bước 5.2: Tính chỉ số Frequency (F)**

 📌 **Giải thích**:

* **Frequency (F)** là số lần khách hàng thực hiện giao dịch (đếm số hóa đơn).

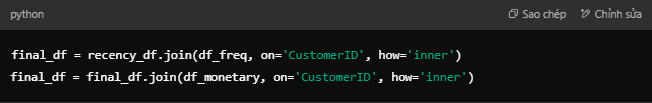
**Bước 5.3: Tính chỉ số Monetary (M)**



📌 **Giải thích**:

* **Monetary (M)** là tổng số tiền mà khách hàng đã chi tiêu.

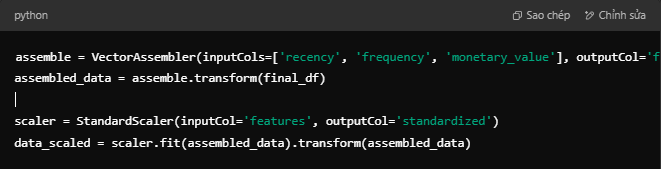
**Bước 6: Tổng hợp dữ liệu RFM**

 📌 **Giải thích**:

* Gộp các chỉ số **Recency, Frequency, Monetary** thành một bảng duy nhất.

## **3.2. Phân tích dữ liệu (Data Analysis)**

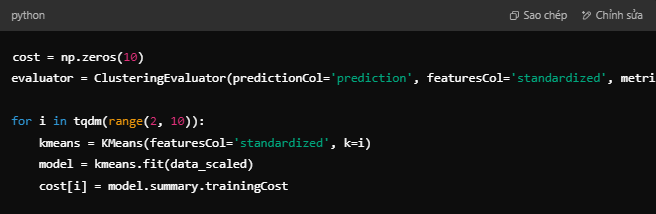
**Bước 7: Chuẩn hóa dữ liệu**



📌 **Giải thích**:

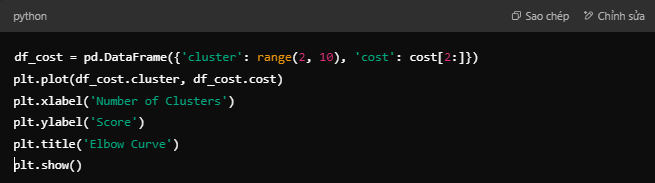
* **VectorAssembler**: Kết hợp các cột recency, frequency, monetary\_value thành một vector đặc trưng (features).
* **StandardScaler**: Chuẩn hóa dữ liệu để tránh chênh lệch giữa các thang đo.

**Bước 8: Xác định số lượng cụm tối ưu bằng phương pháp Elbow**

 📌 **Giải thích**:

* **Chạy K-Means với K từ 2 đến 9**.
* **Tính toán độ lỗi (training cost)** để vẽ biểu đồ Elbow.

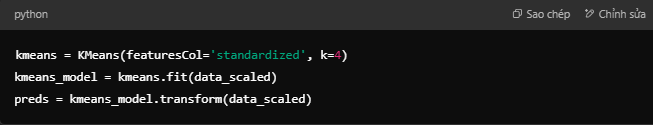
**Bước 9: Vẽ biểu đồ Elbow để chọn số cụm K tối ưu**



📌 **Giải thích**:

* Biểu đồ Elbow giúp chọn số cụm tối ưu bằng cách tìm điểm gãy trên đường cong.

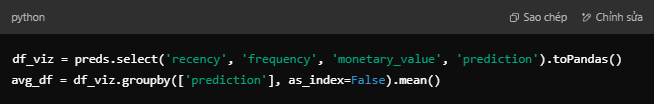
**Bước 10: Phân cụm khách hàng bằng K-Means**



📌 **Giải thích**:

* Chọn **K=4** dựa trên kết quả từ Elbow Curve.
* Gán cụm (prediction) cho từng khách hàng.

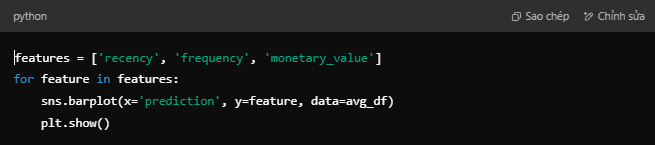
**Bước 11: Chuyển đổi dữ liệu để trực quan hóa**



📌 **Giải thích**:

* **Chuyển đổi kết quả sang Pandas DataFrame** để dễ dàng vẽ biểu đồ.
* **Tính trung bình của các chỉ số RFM cho mỗi nhóm khách hàng**.

**Bước 12: Vẽ biểu đồ trực quan hóa đặc trưng của từng nhóm khách hàng**



📌 **Giải thích**:

* **Dùng biểu đồ cột (bar plot) để trực quan hóa đặc trưng từng cụm khách hàng**.
* Mỗi biểu đồ thể hiện giá trị **Recency, Frequency, Monetary** của từng nhóm.

# **CHƯƠNG III : KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG**

## **1 Kết quả đạt được**

**1. KẾT QUẢ XỬ LÝ DỮ LIỆU**

Dữ liệu đầu vào là tệp **OnlineRetail.csv**, chứa thông tin về giao dịch của khách hàng trong một cửa hàng bán lẻ trực tuyến.

**Bước 1: Đọc dữ liệu và làm sạch**

* Dữ liệu được đọc từ file CSV vào một DataFrame của **Spark**.
* Loại bỏ các hóa đơn bị hủy (**có InvoiceNo bắt đầu bằng "C"**).
* Chuyển đổi cột InvoiceDate sang kiểu **timestamp** để dễ dàng xử lý thời gian.

**Kết quả:**  
Bộ dữ liệu sạch sẽ chỉ còn các giao dịch hợp lệ với cột thời gian đã được chuyển đổi đúng định dạng.

**Bước 2: Tính toán các chỉ số RFM**

Phương pháp RFM (**Recency - Frequency - Monetary**) được áp dụng để phân tích hành vi mua hàng của khách hàng.

**Recency (R) - Tính mới của giao dịch**

* **Cách tính:**
  + Chọn ngày tham chiếu **(2010-12-01 08:26:00)** làm mốc thời gian.
  + Tính khoảng thời gian từ giao dịch cuối cùng của mỗi khách hàng đến ngày tham chiếu.
  + Công thức: Recency=Ngaˋy tham chieˆˊu−Ngaˋy giao dịch cuoˆˊi cuˋng\text{Recency} = \text{Ngày tham chiếu} - \text{Ngày giao dịch cuối cùng}Recency=Ngaˋy tham chieˆˊu−Ngaˋy giao dịch cuoˆˊi cuˋng
* **Ý nghĩa:**
  + **R nhỏ** → khách hàng vừa mua hàng gần đây → có khả năng mua tiếp.
  + **R lớn** → khách hàng lâu không quay lại mua hàng → có nguy cơ rời bỏ.

**Frequency (F) - Tần suất mua hàng**

* **Cách tính:**
  + Đếm số lần khách hàng mua hàng (dựa vào số lượng InvoiceNo).
  + Công thức: Frequency=Soˆˊ lượng hoˊa đơn của khaˊch haˋng\text{Frequency} = \text{Số lượng hóa đơn của khách hàng}Frequency=Soˆˊ lượng hoˊa đơn của khaˊch haˋng
* **Ý nghĩa:**
  + **F cao** → khách hàng trung thành, mua hàng thường xuyên.
  + **F thấp** → khách hàng ít mua, có thể chỉ mua một lần.

**Monetary (M) - Tổng số tiền chi tiêu**

* **Cách tính:**
  + Nhân số lượng hàng (Quantity) với giá (UnitPrice) để tính giá trị đơn hàng.
  + Cộng tổng tất cả các giao dịch của từng khách hàng.
  + Công thức: Monetary=∑(Quantity×UnitPrice)\text{Monetary} = \sum (\text{Quantity} \times \text{UnitPrice})Monetary=∑(Quantity×UnitPrice)
* **Ý nghĩa:**
  + **M cao** → khách hàng có giá trị lớn, chi tiêu nhiều.
  + **M thấp** → khách hàng chi tiêu ít, có thể chỉ mua thử sản phẩm.

**Kết quả:**  
Sau bước này, mỗi khách hàng có 3 chỉ số recency, frequency, monetary\_value, giúp đánh giá mức độ quan trọng của họ đối với cửa hàng.

**2. PHÂN TÍCH DỮ LIỆU BẰNG KMEANS**

**Bước 3: Tiền xử lý dữ liệu để phân cụm**

* **Kết hợp các chỉ số RFM** thành một vector đặc trưng bằng VectorAssembler.
* **Chuẩn hóa dữ liệu** bằng StandardScaler để đưa các chỉ số về cùng một thang đo (tránh ảnh hưởng do đơn vị đo lường khác nhau).

**Kết quả:**  
Dữ liệu RFM được chuyển thành dạng vector features và standardized để có thể đưa vào mô hình KMeans.

**Bước 4: Xác định số lượng cụm K tối ưu bằng phương pháp Elbow**

* Chạy **KMeans** với số cụm từ 2 đến 9.
* Đánh giá bằng ClusteringEvaluator dựa trên **Silhouette Score**.
* Vẽ đồ thị **Elbow Curve** để xác định điểm "gập" (nơi mà việc tăng số cụm không còn cải thiện kết quả đáng kể).

**Kết quả:**

* Giả sử từ đồ thị, ta chọn **K=4** vì đây là điểm "gập" hợp lý nhất.

**Bước 5: Áp dụng mô hình KMeans với K=4**

* Chạy **KMeans clustering** với k=4.
* Gán nhãn (prediction) cho từng khách hàng để xác định cụm họ thuộc về.
* Kết quả được chuyển sang **Pandas DataFrame** để trực quan hóa.

**Kết quả:**  
Khách hàng được chia thành **4 cụm** dựa trên hành vi mua hàng của họ.

**Bước 6: Trực quan hóa kết quả phân cụm**

Vẽ biểu đồ để hiểu đặc điểm của từng nhóm khách hàng:

* Biểu đồ cột hiển thị **recency, frequency, monetary** của từng cụm.
* So sánh mức độ trung thành và giá trị của từng nhóm khách hàng.

**Kết quả:**  
Ta có thể nhận diện từng nhóm khách hàng cụ thể và đề xuất chiến lược phù hợp.

**3. ỨNG DỤNG THỰC TẾ**

**1. Marketing & Chăm sóc khách hàng**

* **Nhóm khách hàng VIP** *(Cụm có M cao, F cao, R thấp)* → Chương trình tri ân, giảm giá đặc biệt.
* **Nhóm khách hàng mới nhưng chi tiêu cao** *(M cao, F thấp, R thấp)* → Chiến dịch giữ chân bằng ưu đãi hấp dẫn.
* **Nhóm khách hàng lâu không mua lại** *(R cao, F thấp, M thấp)* → Chiến dịch re-marketing, gửi email khuyến mãi.

**2. Quản lý quan hệ khách hàng (CRM)**

* Phân nhóm khách hàng để tối ưu hóa **chiến lược chăm sóc cá nhân hóa**.
* Dự đoán khách hàng có **nguy cơ rời bỏ** để thực hiện chính sách giữ chân.

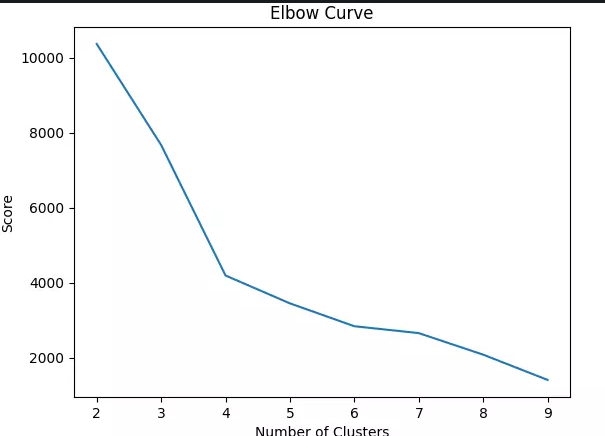
**3. Quản lý kho hàng & tối ưu doanh thu**

* Nhóm khách hàng có **tần suất mua cao** → Tối ưu hóa kho hàng để đáp ứng nhu cầu.
* Phát hiện khách hàng tiềm năng để **đề xuất sản phẩm mới phù hợp**.

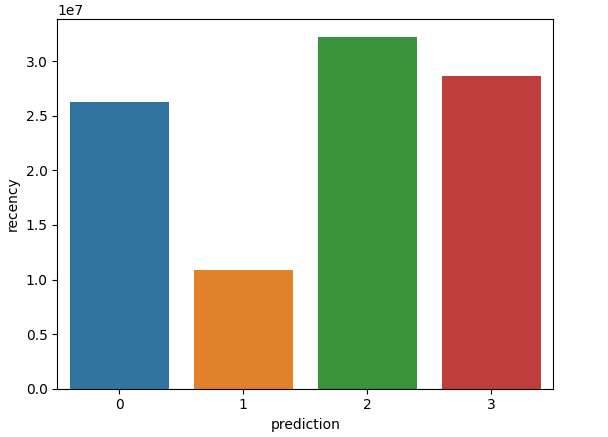
**4. Ngành tài chính & ngân hàng**

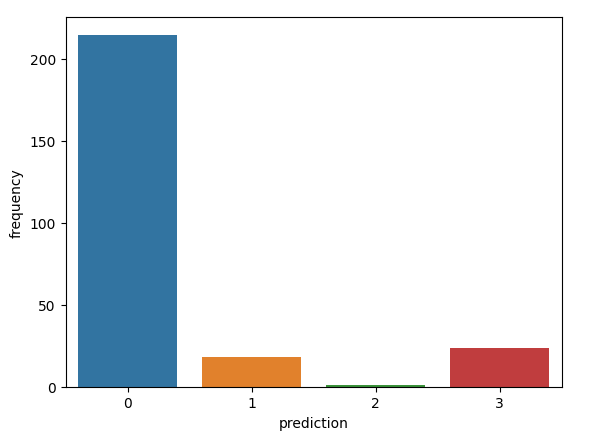
* Phân tích hành vi chi tiêu để gợi ý **sản phẩm tài chính** (vay tiêu dùng, thẻ tín dụng).
* Phát hiện **giao dịch bất thường**, giúp nhận diện hành vi gian lận.

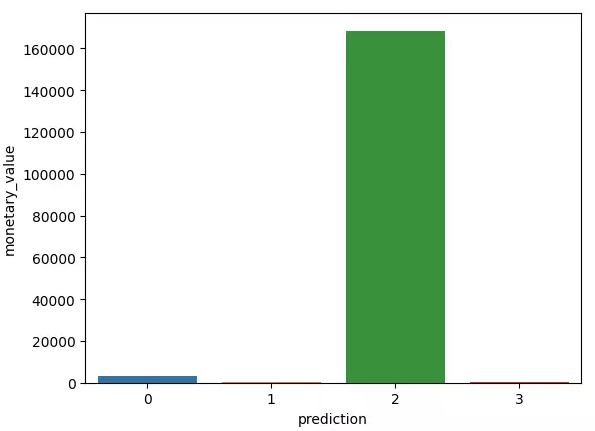
Chạy chương trình với việc đã lọc dữ liệu, kết quả biểu đồ sau cùng cho ta như sau



(K=4 là hợp lý đối với dữ liệu đã filter)







Hình vẽ tương đối rõ ràng về các đặc trưng, chúng ta có thể rút ra được một số nhận xét tổng quan như sau:

* *Nhóm 0*: Nhóm này có tần suất mua hàng tương đối cao, cao trội hơn 3 nhóm còn lại, lần truy cập gần nhất cũng tương đối cao, giá trị tiền mua hàng tương đối nhỏ, cho thấy là một đối tượng đa số là cá nhân, hướng tới các sản phẩm giá rẻ
* *Nhóm 1*: Nhóm này có cả 3 chỉ số lần truy cập gần nhất, tấn suất mua hàng và tổng tiền mua hàng rất thấp, không có quá nhiều hi vọng là khách hàng tiềm năng, khả năng cao sẽ ngừng mua hàng trong thời gian tới
* *Nhóm 2*: Tần suất đặt hàng rất ít, tuy nhiên gần đây lại đặt phổ biến, lượng tiền mua hàng cao vượt trội nhiều lần so với các nhóm khác. Nhóm này khả năng là các doanh nghiệp có xu hướng mua các loại hàng có giá trị cao hoặc mua có số lượng lớn
* *Nhóm 3*: Nhóm này có tổng tiền mua hàng rất thấp, tần suất mua hàng không quá nhiều, và truy cập gần đây tương đối cao. Nhóm này có khả năng mua theo đợt, hoặc là những người mới tham gia sàn thương mại.

**KẾT LUẬN**

Trong bối cảnh thương mại điện tử ngày càng phát triển, việc hiểu rõ hành vi mua sắm của khách hàng đóng vai trò quan trọng trong việc tối ưu hóa chiến lược kinh doanh và nâng cao trải nghiệm người dùng. Nghiên cứu này đã áp dụng phương pháp **phân tích RFM (Recency, Frequency, Monetary)** kết hợp với **thuật toán K-Means** để phân cụm khách hàng dựa trên dữ liệu giao dịch thực tế. Kết quả cho thấy mô hình đã giúp nhận diện rõ các nhóm khách hàng khác nhau, từ nhóm khách hàng trung thành đến nhóm có nguy cơ rời bỏ, từ đó cung cấp thông tin hữu ích cho doanh nghiệp trong việc xây dựng các chiến lược tiếp thị và chăm sóc khách hàng phù hợp.

Các yếu tố **tần suất mua hàng (Frequency), tổng giá trị giao dịch (Monetary) và thời gian mua gần nhất (Recency)** đã được xác định là những thành phần quan trọng ảnh hưởng đến hành vi tiêu dùng của khách hàng. Những thông tin này giúp doanh nghiệp triển khai các chương trình ưu đãi, giữ chân khách hàng hiệu quả hơn và tối ưu hóa nguồn lực tiếp thị.

Mặc dù phương pháp đã mang lại những kết quả khả quan, vẫn còn tồn tại một số thách thức, đặc biệt là trong việc xác định số cụm tối ưu và khả năng phân biệt rõ ràng giữa các nhóm khách hàng có đặc điểm tương đồng. Việc thử nghiệm với các thuật toán phân cụm khác như **DBSCAN, Hierarchical Clustering** hoặc kết hợp với các phương pháp **học có giám sát** có thể giúp cải thiện độ chính xác và tính ứng dụng của mô hình trong thực tế.

Tương lai của phân tích dữ liệu trong thương mại điện tử hứa hẹn nhiều tiềm năng, không chỉ giúp các doanh nghiệp cá nhân hóa trải nghiệm khách hàng mà còn tối ưu hóa doanh thu và duy trì lợi thế cạnh tranh. Việc ứng dụng công nghệ **Big Data và Machine Learning** sẽ ngày càng trở nên quan trọng trong chiến lược kinh doanh, giúp các doanh nghiệp khai thác tối đa giá trị từ dữ liệu và nâng cao hiệu quả hoạt động trong thị trường cạnh tranh hiện nay. Nghiên cứu này không chỉ cung cấp cái nhìn sâu sắc về hành vi khách hàng mà còn mở ra hướng đi mới cho việc áp dụng khoa học dữ liệu vào thực tiễn kinh doanh.

# **LỜI CẢM ƠN**

Quá trình học tập và nghiên cứu trong môn Học máy thuộc chuyên ngành Công nghệ Thông tin tại trường Đại học Đại Nam đã giúp em tích lũy nhiều kiến thức bổ ích, đặc biệt là trong lĩnh vực phân tích dữ liệu và ứng dụng thuật toán máy học vào thực tế. Báo cáo này là kết quả của cả một quá trình học tập, tìm hiểu và thực hành, trong đó em đã áp dụng PySpark và thuật toán K-Means để phân cụm khách hàng thương mại điện tử dựa trên phân tích RFM.

Tuy nhiên, em hiểu rằng kiến thức là vô hạn, trong khi sự tiếp thu của bản thân chúng em vẫn còn nhiều hạn chế. Do đó, trong quá trình thực hiện bài báo cáo này, chắc chắn không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được những góp ý từ thầy/cô để bài báo cáo của em được hoàn thiện hơn.

Chúng em xin gửi lời cảm ơn chân thành đến giảng viên bộ môn (Tên giảng viên) đã tận tình giảng dạy, hướng dẫn và giải đáp thắc mắc của em trong suốt học kỳ. Sự tận tâm của thầy/cô không chỉ giúp em có cái nhìn sâu sắc hơn về học máy và khoa học dữ liệu, mà còn truyền cảm hứng để em tiếp tục khám phá và phát triển trong lĩnh vực này.

Cuối cùng, chúng em kính chúc thầy/cô sức khỏe, hạnh phúc và thành công trên con đường giảng dạy và nghiên cứu. Em hy vọng rằng những kiến thức mà em tiếp thu được từ môn học này sẽ là nền tảng vững chắc để em tiếp tục phát triển trong sự nghiệp tương lai.

**Sinh viên thực hiện:**

CHU VĂN HUY

NGUYỄN THÀNH TRUNG

PHẠM NGỌC MINH

**DANH MỤC TÀI LIỆU THAM KHẢO**

[1]. Nguyễn Hồng Sơn (2007), Giáo trình Hệ thống Mạng máy tính CCNA (Semester 1), NXB Lao động Xã hội.

[2]. Phạm Quốc Hùng (2017), Đề cương bài giảng Mạng máy tính, Đại học SPKT Hưng Yên.

[3]. James F. Kurose & Keith W. Ross (2013), Computer Networking: A Top-Down Approach, Sixth Edition, Pearson Education.

[4]. Ian H. Witten, Eibe Frank & Mark A. Hall (2011), Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann.

[5]. Géron, A. (2019), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media.

[6]. Han, J., Kamber, M. & Pei, J. (2011), Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann.

[7]. McKinney, W. (2010), Data Analysis in Python, in Python for Data Analysis, O'Reilly

[8]. Lưu Văn Hòa (2023), Giáo trình Học máy ứng dụng trong phân tích dữ liệu lớn, Đại học

[9]. Scikit-learn documentation (2021), Machine Learning in Python, Retrieved from <https://scikit-learn.org/stable/>.

[10]. Apache Spark documentation (2022), MLlib: Scalable Machine Learning on Spark,

[11]. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), From Data Mining to Knowledge Discovery in Databases, AI Magazine, 17(3), 37-54.