**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN MÔN HỌC**

**MÔN: PHÂN TÍCH DỮ LIỆU**

**Giảng viên hướng dẫn: Phan Thành Huấn**

**Thực hiện: Lã Trần Minh Anh – 3120410028**

**Phan Anh - 3120410037**

**TP. HCM tháng 12 / 2023**

**BẢNG ĐÁNH GIÁ MỨC ĐỘ THAM GIA**

|  |  |  |
| --- | --- | --- |
| Danh sách thành viên | MSSV | Mức độ tham gia |
| Lã Trần Minh Anh | 3120410028 | 100% |
| Phan Anh | 3120410037 | 100% |

**Chủ đề của đồ án:**

***Mua sắm trực tuyến:*** *Phân tích các mẫu mua hàng trực tuyến để tìm hiểu các luật kết hợp giữa các mặt hàng.*

*- Dữ liệu về lịch sử mua hàng: sử dụng để xác định các mặt hàng thường được mua cùng nhau, các mặt hàng thường được mua trước hoặc sau một mặt hàng khác, và các mặt hàng thường được mua bởi các khách hàng có chung sở thích.*

*- Dữ liệu về dữ liệu vị trí: sử dụng để xác định các mặt hàng thường được mua ở cùng một địa điểm, các mặt hàng thường được mua bởi các khách hàng ở cùng một khu vực, và các mặt hàng thường được mua vào cùng một thời điểm.*

*- Dữ liệu về dữ liệu thanh toán: sử dụng để xác định các mặt hàng thường được mua bởi cùng một khách hàng, các mặt hàng thường được mua cùng với một phương thức thanh toán cụ thể, và các mặt hàng thường được mua vào cùng một thời điểm trong năm.*

MỤC LỤC

[CHƯƠNG I: KHÁI QUÁT ĐỒ ÁN 1](#_Toc153407307)

[1. Tổng quan về Mua sắm trực tuyến và luật kết hợp giữa các mặt hàng: 1](#_Toc153407308)

[1.1. Khái niệm mua sắm trực tuyến: 1](#_Toc153407309)

[1.2. Luật kết hợp: 1](#_Toc153407310)

[2. Mô tả đồ án 2](#_Toc153407311)

[3. Mô tả tập dữ liệu: 2](#_Toc153407312)

[4. Các thư viện Python được sử dụng để phục vụ việc phân tích: 3](#_Toc153407313)

[CHƯƠNG II: LỌC DỮ LIỆU VÀ TRỰC QUAN HÓA 6](#_Toc153407314)

[1. Phân tích tổng quan 6](#_Toc153407315)

[1.1. Tiền xử lý dữ liệu: 6](#_Toc153407316)

[1.2. Doanh số bán hàng theo tháng: 9](#_Toc153407317)

[1.3. Top 10 các sản phẩm được mua nhiều nhất: 11](#_Toc153407318)

[1.4. Top 10 các sản phẩm được mua ít nhất: 11](#_Toc153407319)

[1.5. Khu vực mua hàng phổ biến: 12](#_Toc153407320)

[1.6. Số Lượng Giao Dịch Qua Các Mùa Ở Uk 14](#_Toc153407321)

[2. Biểu đồ boxplot cho cột TotalValue: 15](#_Toc153407322)

[3. Sử dụng thuật toán KMeans để phân cụm (cluster) các khách hàng: 17](#_Toc153407323)

[3.1. Lý thuyết: 17](#_Toc153407324)

[3.2. Phân tích RFM (Recency, Frequency, Monetary) trên dữ liệu khách hàng 18](#_Toc153407325)

[4. Phân tích phân khúc: 22](#_Toc153407326)

[5. Hồi quy logistic: 26](#_Toc153407327)

[6. Tập phổ biến: 27](#_Toc153407328)

[CHƯƠNG III: TỔNG KẾT 31](#_Toc153407329)

[1. Kết luận: 31](#_Toc153407330)

[2. Hạn chế: 31](#_Toc153407331)

[3. Hướng phát triển: 31](#_Toc153407332)

[TÀI LIỆU THAM KHẢO 32](#_Toc153407333)

# CHƯƠNG I: KHÁI QUÁT ĐỒ ÁN

1. **Tổng quan về Mua sắm trực tuyến và luật kết hợp giữa các mặt hàng:**
   1. **Khái niệm mua sắm trực tuyến:**

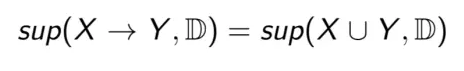
"Mua sắm trực tuyến" là việc mua các sản phẩm và dịch vụ thông qua internet, thay vì phải đến cửa hàng truyền thống để mua hàng. Đây là một hình thức mua sắm tiện lợi và phổ biến ngày nay, cho phép người tiêu dùng chọn mua hàng hóa từ các trang web bán lẻ và đặt mua trực tuyến.

* 1. **Luật kết hợp:**

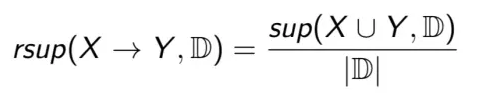
Luật kết hợp là mối quan hệ giữa các tập thuộc tính trong cơ sở dữ liệu. Luật kết hợp là phương tiện hữu ích để khám phá các mối liên kết trong dữ liệu.  
Một luật kết hợp là một mệnh đề kéo theo có dạng X -> Y, trong đó X, Y ⊆ I, thỏa mãn điều kiện X giao Y = rỗng. Các tập hợp X và Y được gọi là các tập hợp thuộc tính (itemset). Tập X gọi là nguyên nhân, tập Y gọi là hệ quả. Có 2 độ đo quan trọng đối với luật kết hợp: Độ hỗ trợ (support) và độ tin cậy (confidence), được định nghĩa như phần dưới đây.

**Độ hỗ trợ**

Độ hỗ trợ của một luật kết hợp X -> Y là tỷ lệ giữa số lượng các bản ghi chứa tập hợp X -> Y, so với tổng số các bản ghi trong D - Ký hiệu supp(X -> Y).



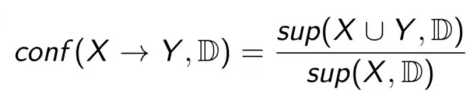
Độ hỗ trợ tương đối của luật X->Y trong cở sở dữ liệu D kí hiệu là rsup(X->Y, D) là số phần trăm các giao dịch trong D chứa cả X và Y.



Nếu độ hỗ trợ của một kết kết hợp X -> Y là 30% thì có nghĩa là 30% tổng số bản ghi chứa X hợp Y. Như vậy độ hỗ trợ mang ý nghĩa thống kê của luật.

**Độ tin cậy**

Độ tin cậy (confidence) của luật X → Y trong D, ký hiệu conf (X → Y , D), là tỉ lệ giữa số giao dịch chứa cả X và Y trên số giao dịch chỉ chứa X.

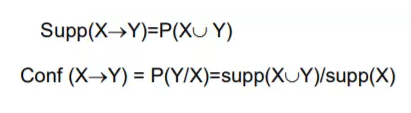


Ký hiệu độ tin cậy của một luật r là conf(r). Ta có 0 <= conf(r) <= 1

Độ hỗ trợ và độ tin cậy có xác xuất như sau:

Độ hỗ trợ là xác xuất trong giao dịch chứa cả X và Y.

Độ tin cậy là xác xuất có điều kiện mà một giao dịch trong D chứa Y trong khi đã chứa X (bản chất vẫn là mức độ in cậy của luật).



**Kết luận**

* Luật X -> Y được gọi là phổ biến nếu sup(X->Y, D) >= minsup (minsup do người dùng định nghĩa)
* Luật X-> Y được gọi là mạnh nếu độ tin cậy của nó lớn hơn hoặc bằng một ngưỡng minconf do người dùng định nghĩa: conf (X->Y) >= minconf

1. **Mô tả đồ án**

Trong bài báo cáo dữ liệu đồ án của mình, chúng em sẽ tiến hành phân tích các dữ liệu về lịch sử mua hàng, thanh toán,vị trí các khách hàng cùng mua loại sản phẩm đó, các sản phẩm được mua thường xuyên để tìm hiểu các luật kết hợp giữa các mặt hàng với nhau.

1. **Mô tả tập dữ liệu:**

Nhóm em sẽ tiến hành phân tích trên tập dữ liệu Online\_retail.csv gồm 8 cột và 541910 dòng. Dữ liệu là lịch sử mua hàng trực tuyến của các nước từ 12/2010 tới tháng 12/2011

***Thông tin các thuộc tính:***

InvoiceNo: Số hóa đơn.

StockCode: Mã sản phẩm (mặt hàng)

Description: Tên sản phẩm (mặt hàng).

Quantity: Số lượng của mỗi sản phẩm (mặt hàng) trong mỗi giao dịch.

InvoiceDate: Ngày và giờ hóa đơn.

UnitPrice: Đơn giá.

CustomerID: Số khách hàng.

Country: Tên quốc gia.

Trong những thuộc tính kể trên, nhóm chúng em sẽ chọn ra những thuộc tính cần thiết và tiến hành tiền xử lý dữ liệu để phục vụ cho việc phân tích của mình như sau:

StockCode, Description, Quantity: Phân loại các mặt hàng sản phẩm được mua thường xuyên nhất, biết được tên sản phẩm và số lượng sản phẩm được mua nhiều nhất.

InvoiceDate: Thuộc tính này để biết được vào các thời điểm nào thì mặc hàng nào được khách hàng mua nhiều nhất cũng như việc ảnh hưởng của món hàng đến khách hàng.

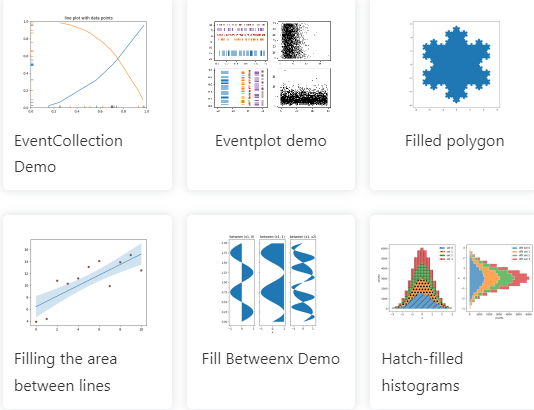
UnitPrice: thuộc tính này biết được giá trị mà đơn hàng mang lại nhằm giúp tìm ra được sản phẩm mua nhiều nhất.

CustomerID, country: 2 thuộc tính này nhằm biết vị trí của khách hàng từ đó phân loại khu vực nào sẽ có sản phẩm được mua nhiều theo hình như online.

1. **Các thư viện Python được sử dụng để phục vụ việc phân tích:**

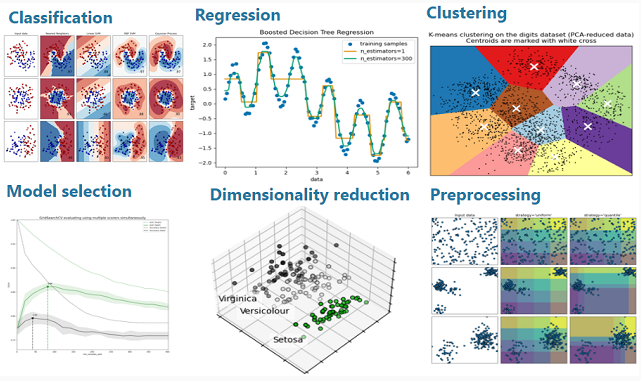
- Pandas: thư viện này được sử dụng để phục vụ việc tiền xử lý các dữ liệu trong file qua việc chọn các thuộc tính cần thiết, lọc các dữ liệu theo điều kiện, tính toán các số liệu ở các cột, thêm các cột tính toán,…

- Matplotlib: thư viện này được sử dụng để vẽ nên các biểu đồ từ các số liệu đã được tính toán giúp cho việc thể hiện dữ liệu trở nên trực quan và dễ hiểu hơn.



- NumPy (Numerical Python) : thư viện mở được sử dụng để xử lý các dữ liệu số, tính toán khoa học và toán học trong Python. NumPy chứa các cấu trúc dữ liệu ma trận (matrix data) và mảng đa chiều (multidimensional array), có thể thực hiện nhiều phép toán khác nhau trên mảng.

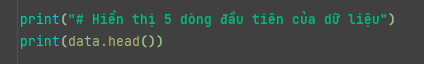
- Scikit-learn: xử lý hầu hết các kỹ thuật học máy (machine learning) phổ biến, với các chức năng bao gồm: Hồi quy (Regression: Linear and Logistic Regression); Phân loại (Classification: K-Nearest Neighbors); Phân cụm (Clustering: K-Means, K-Means++); Lựa chọn mô hình (Model selection); Tiền xử lý (Preprocessing: Min-Max Normalization)..

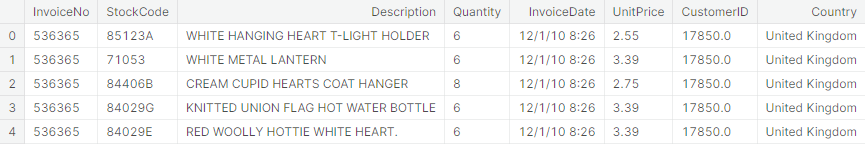


# CHƯƠNG II: LỌC DỮ LIỆU VÀ TRỰC QUAN HÓA

* 1. **Phân tích tổng quan**
  2. **Tiền xử lý dữ liệu:**

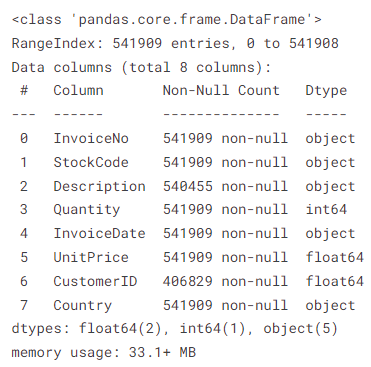
**# Hiển thị 5 dòng đầu tiên của dữ liệu**

****



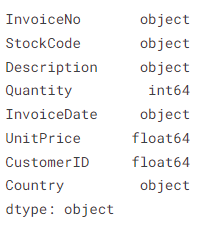
**# Thông tin tổng quan về dữ liệu:**

****

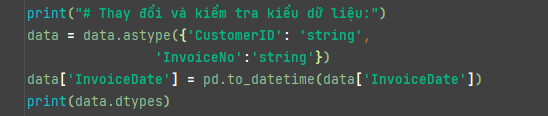
****

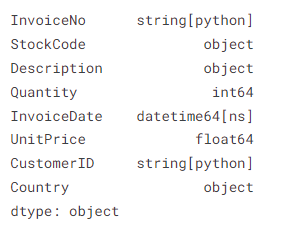
**# Kiểm tra kiểu dữ liệu của từng cột:**

****

****

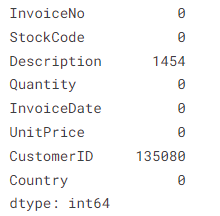
**# Thay đổi và kiểm tra kiểu dữ liệu:**

****

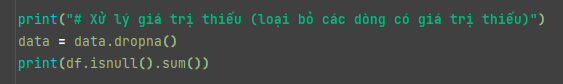
****

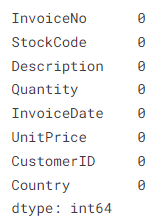
**# Kiểm tra các giá trị thiếu:**

****

****

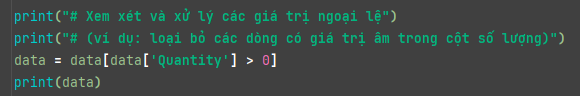
**# Loại bỏ các giá trị Null và check lại:**

****

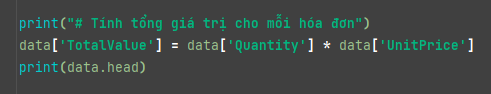
****

**# Xem xét và xử lý các giá trị ngoại lệ**

**# (ví dụ: loại bỏ các dòng có giá trị âm trong cột số lượng)**

****

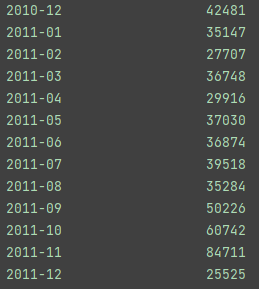
**# Tính tổng giá trị cho mỗi hóa đơn**

****

****

* 1. **Doanh số bán hàng theo tháng:**

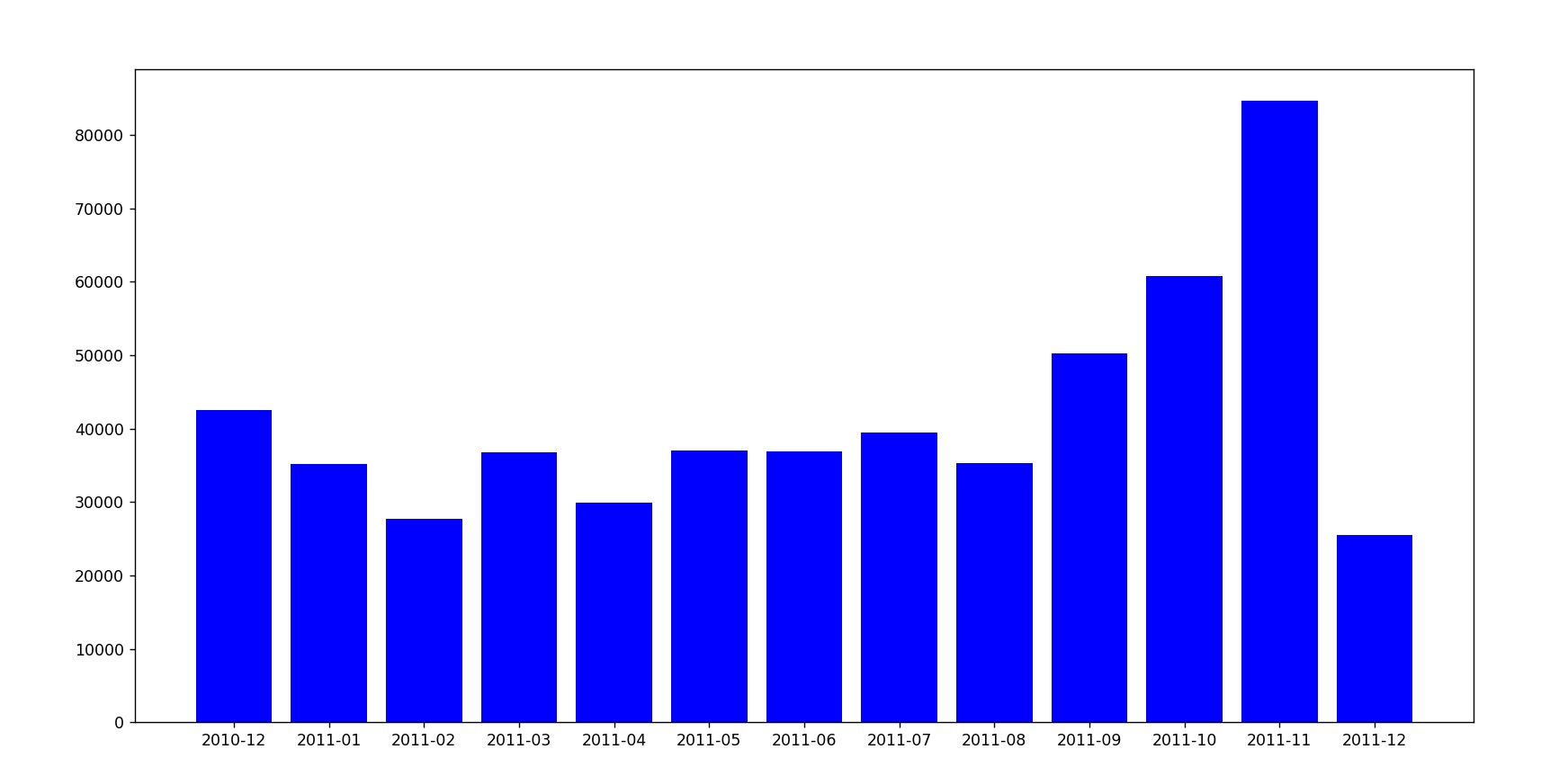
Doanh số bán hàng thu được là:



Từ biểu đồ dưới đây, doanh thu các tháng có sự chênh lệch không nhiều và tháng có doanh thu nhiều nhất là tháng 11/2011, với lượng đơn hàng là 84711 đơn.

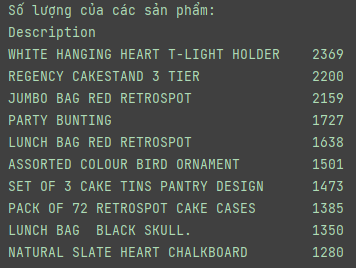
Từ biểu đồ cột, có thể nhận xét doanh số bán hàng như sau:

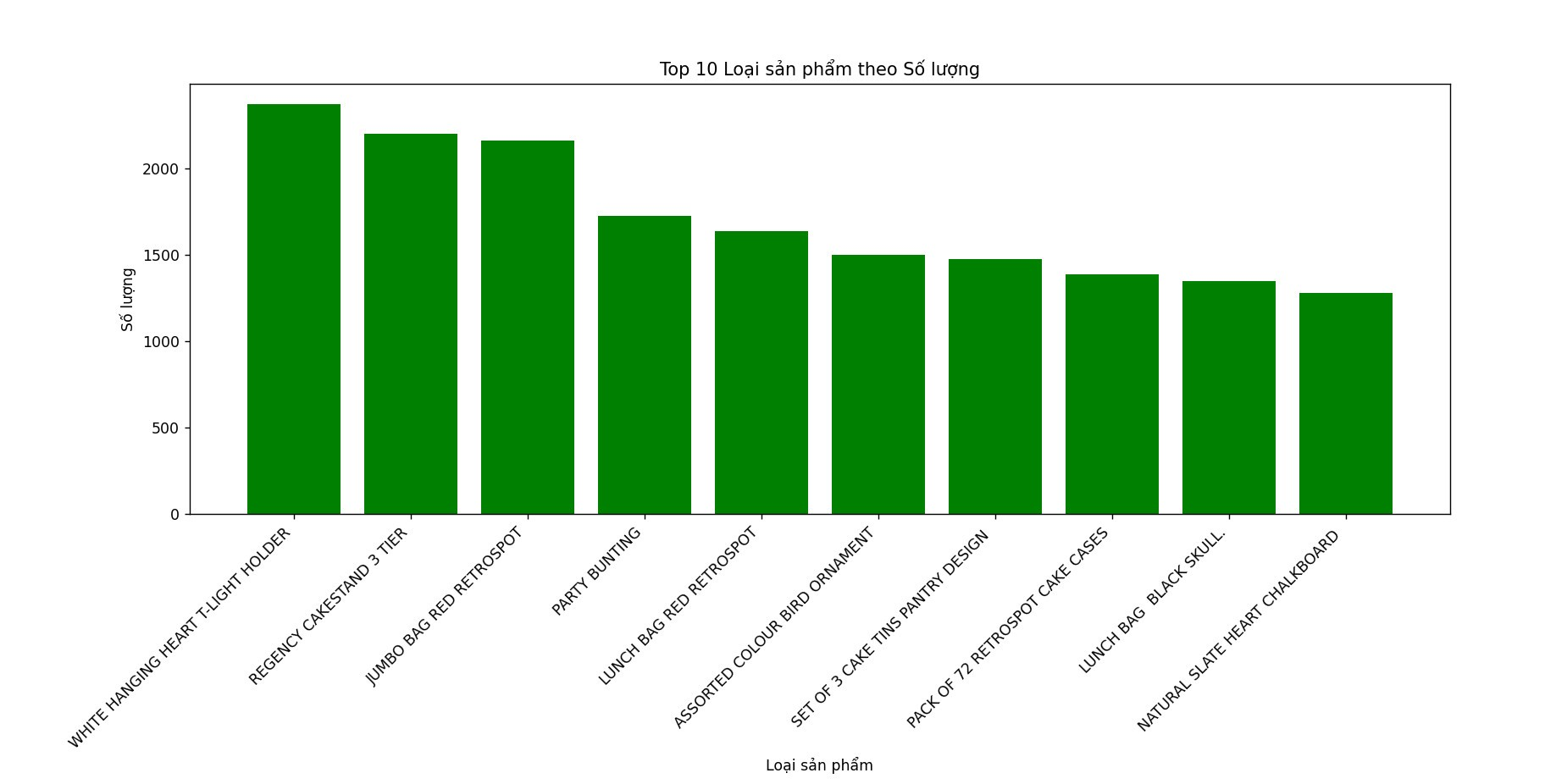
Giai đoạn từ 12/2010 tới 8/2011 luôn ổn định và tăng dần vào tháng 9 trở đi, đạt doanh thu gấp đôi vào tháng 11/2011 và ít dần vào cuối năm.

****

* 1. **Top 10 các sản phẩm được mua nhiều nhất:**

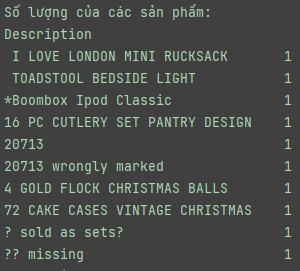
Dữ liệu lọc được:



****

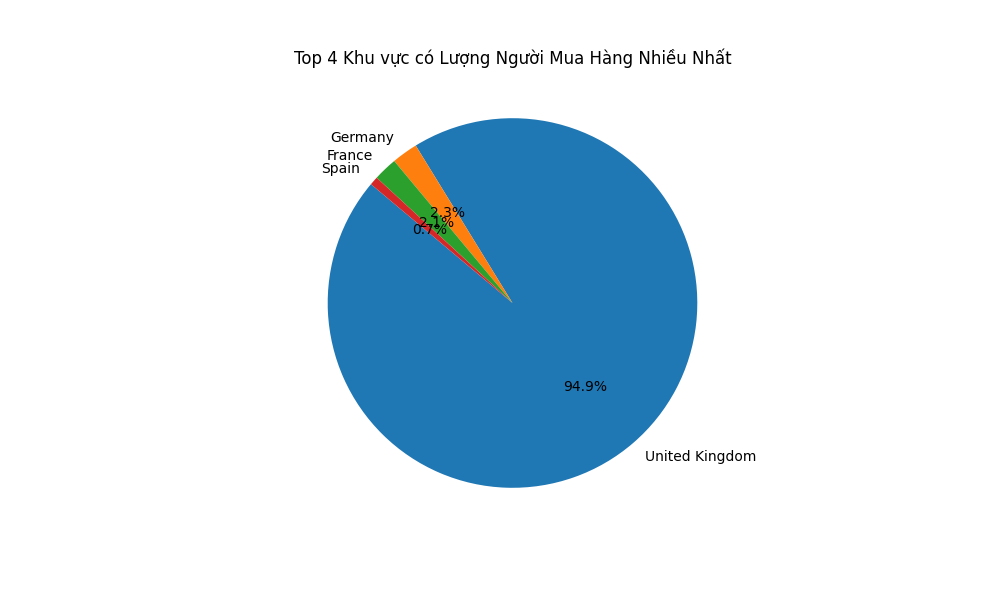
Biểu đồ này thể hiện được sản phẩm được mua nhiều nhất là White Hanging Heart T-light Holder

* 1. **Top 10 các sản phẩm được mua ít nhất:**

****

* 1. **Khu vực mua hàng phổ biến:**

Theo như file dữ liệu, khu vực có lượng khách hàng nhiều nhất là tại Anh chiếm 94.9% lượng người mua hàng trên khắp thế giới.

****

Bảng dưới đây là các khu vực mua hàng

Country

United Kingdom 16649

Germany 457

France 389

EIRE 260

Belgium 98

Netherlands 95

Spain 90

Australia 57

Portugal 57

Switzerland 51

Finland 41

Italy 38

Norway 36

Sweden 36

Channel Islands 26

Poland 19

Japan 19

Denmark 18

Austria 17

Cyprus 16

Unspecified 8

Iceland 7

Singapore 7

Canada 6

Malta 5

Israel 5

Greece 5

USA 5

Lithuania 4

European Community 4

United Arab Emirates 3

Czech Republic 2

Bahrain 2

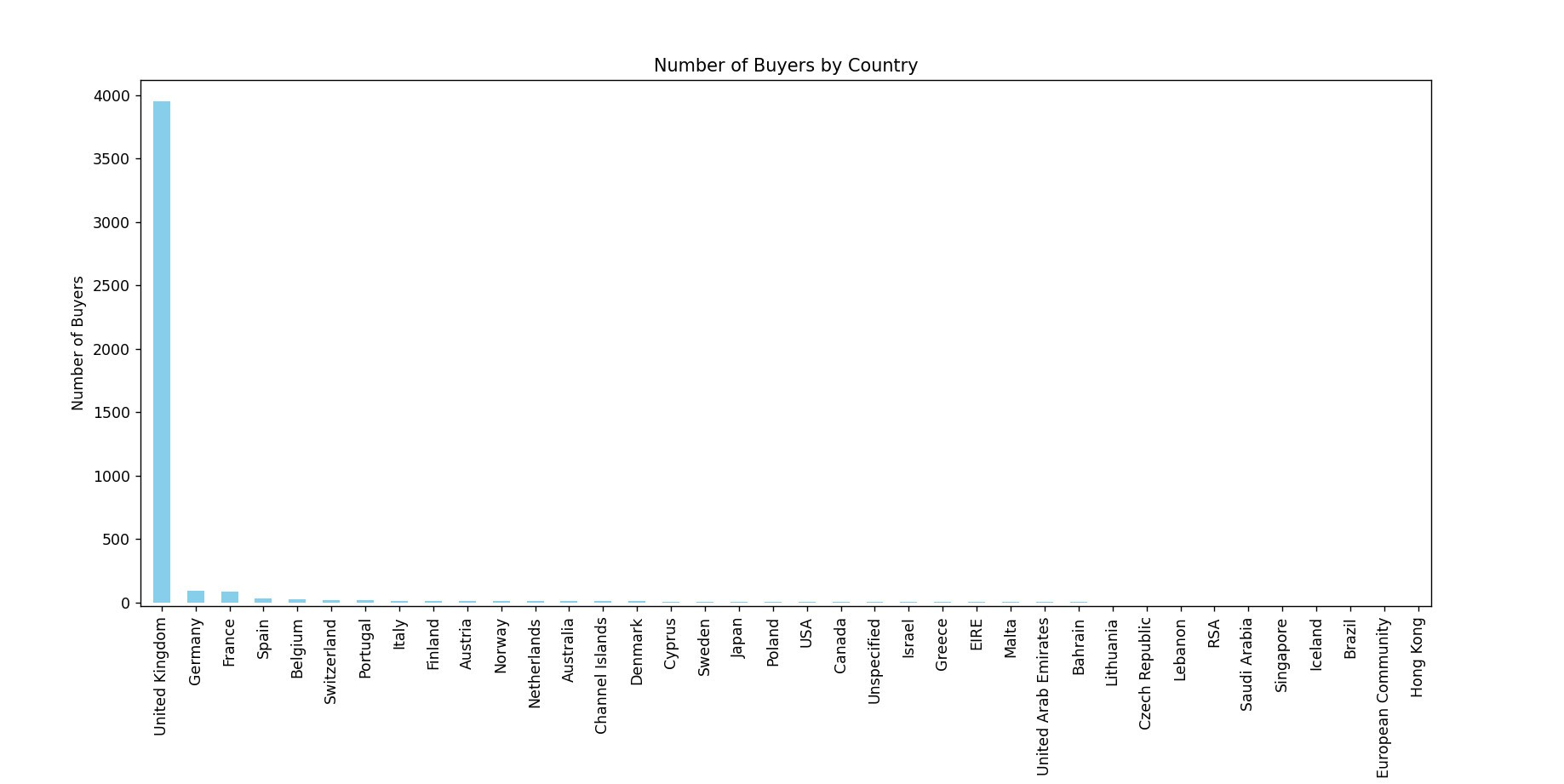
Lebanon 1

RSA 1

Saudi Arabia 1

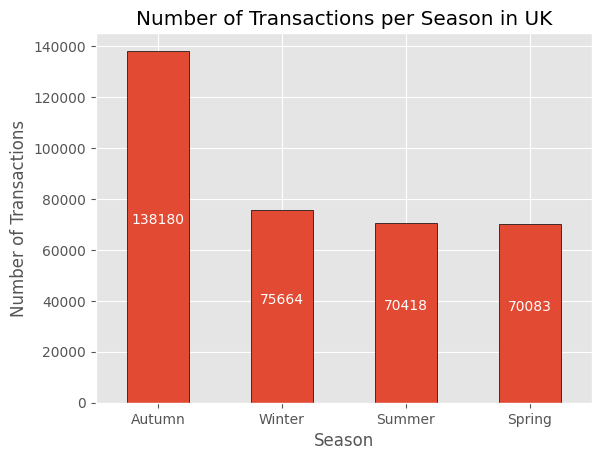
Brazil 1

Name: InvoiceNo, dtype: int64

****

* 1. **Số Lượng Giao Dịch Qua Các Mùa Ở Uk**

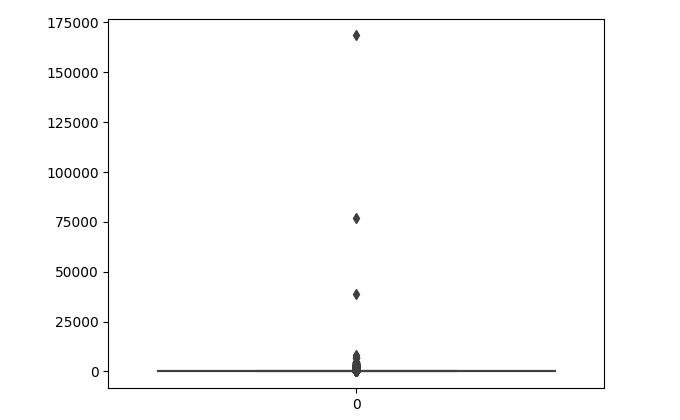
Mùa có lượng giao dịch nhiều nhất là mùa thu, với lượng người mua sắm trực tuyến là 138180 gd.

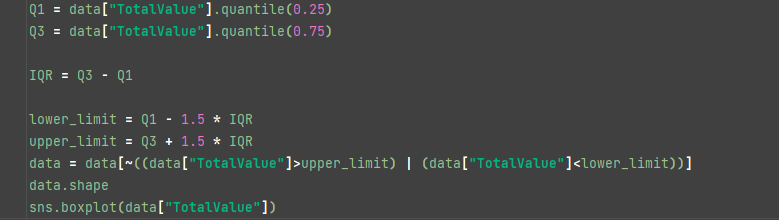


1. **Biểu đồ boxplot cho cột TotalValue:**

Biểu đồ boxplot là một dạng biểu đồ trực quan hóa dữ liệu giúp chúng ta hiểu được phân phối dữ liệu, bao gồm các giá trị trung bình, giá trị trung vị, giá trị tứ phân vị thứ nhất và thứ ba, cũng như các giá trị ngoại lai.

****

****

****

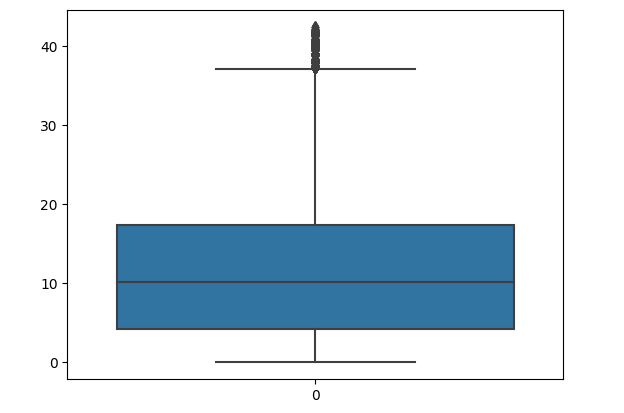
Bọn em sử dụng phương pháp IQR (Interquartile Range) để loại bỏ các điểm dữ liệu nằm ngoài khoảng giới hạn

* Tính toán các giá trị tứ phân vị thứ nhất (Q1) và thứ ba (Q3) của dữ liệu trong cột "TotalValue" của DataFrame.
* Tính toán giá trị IQR (Interquartile Range).
* Tính toán giá trị giới hạn dưới (lower\_limit) và giá trị giới hạn trên (upper\_limit) của các giá trị bình thường.

Số hàng và số cột của DataFrame sau khi loại bỏ các outlier



Boxplot cho cột "Total" trong DataFrame sau khi loại bỏ các outlier.



1. **Sử dụng thuật toán KMeans để phân cụm (cluster) các khách hàng:**
   1. **Lý thuyết:**

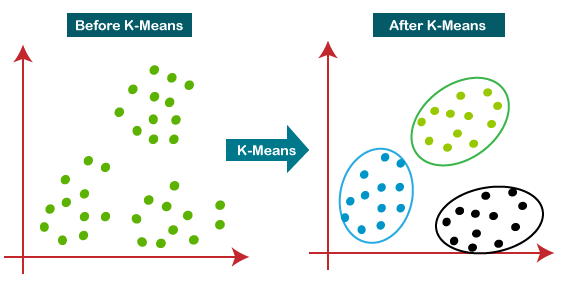
Phân cụm dữ liệu là quá trình gom cụm/ nhóm các đối tượng/dữ liệu có đặc điểm tương đồng vào các nhóm/ cụm tương ứng. Trong đó, tương đồng giữa những phần tử trong cùng cụm; khác biệt với với những phần tử trong các cụm khác.

**- Đánh giá mô hình phân cụm:**

+ Ưu điểm của phân cụm phân lớp là không phải xác định trước số lượng cụm điều này khá vượt trội so với K-Means. Tuy nhiên, nó không hoạt động tốt với lượng dữ liệu khổng lồ.

+ Thuật toán phân cụm phân lớp có thể được sử dụng để xác định, dự đoán số cụm trước khi thực hiện thuật toán K-Means.

🟎Thuật toán K-Means là một thuật toán học không giám sát, được sử dụng để giải quyết các vấn đề phân cụm trong học máy hoặc khoa học dữ liệu. Nó hoạt động bằng cách chia tập dữ liệu không được gán nhãn thành k nhóm khác nhau sao cho mỗi dữ liệu chỉ thuộc về một nhóm có các thuộc tính tương tự.



*Mô hình Thuật toán K-Means*

* 1. **Phân tích RFM (Recency, Frequency, Monetary) trên dữ liệu khách hàng**

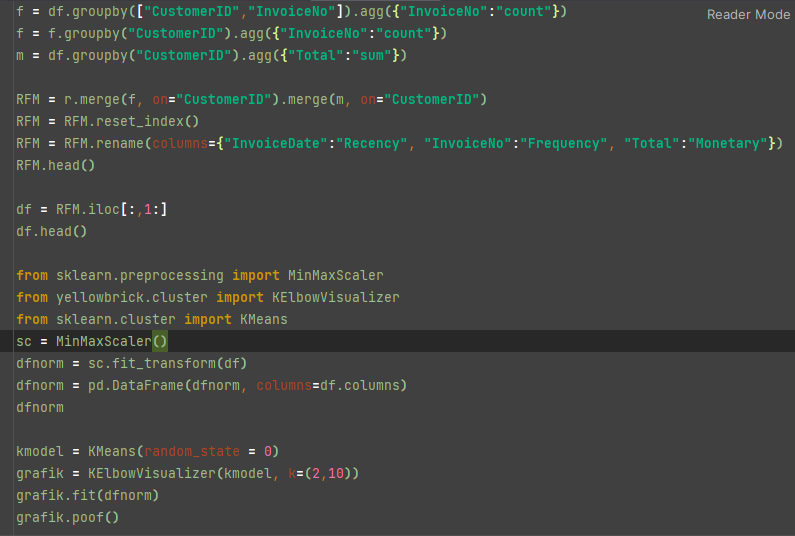
**Recency (R):** thời gian kể từ ngày giao dịch gần nhất của mỗi khách hàng đến thời điểm cuối cùng trong dữ liệu.

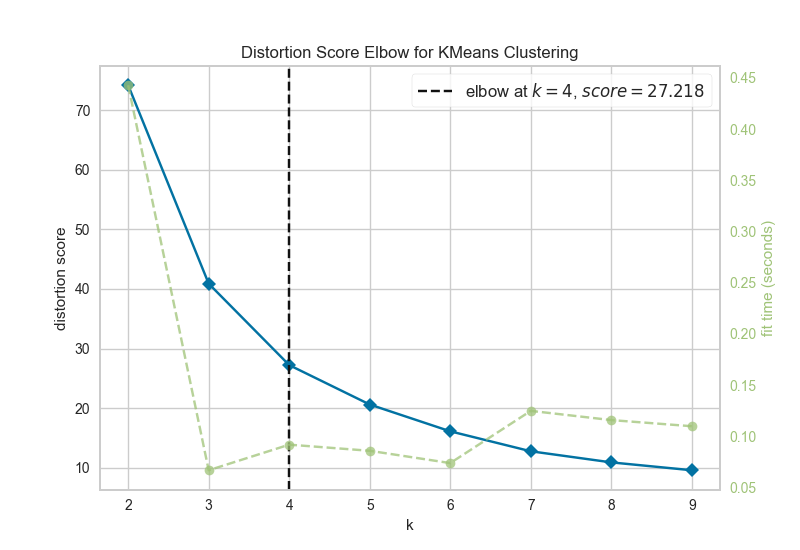
**Frequency (F):** số lần mua hàng cho mỗi đơn hàng của mỗi khách hàng và sau đó tính tổng số đơn hàng mà mỗi khách hàng đã thực hiện.

**Monetary (M):** tổng giá trị tiền đã chi tiêu của mỗi khách hàng.

**Tạo DataFrame RFM:** Gộp các thông tin về Recency, Frequency, và Monetary thành một DataFrame mới có tên là **RFM**.

**Chuẩn hóa dữ liệu và Ước lượng số cụm (clusters) tối ưu bằng phương pháp Elbow:** Chuẩn hóa dữ liệu sử dụng MinMaxScaler, sau đó sử dụng thuật toán KMeans để phân cụm khách hàng. Phương thức Elbow Visualizer được sử dụng để ước lượng số cụm tối ưu dựa trên biểu đồ Elbow.

****

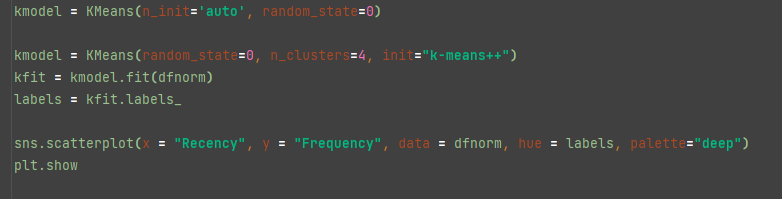
****

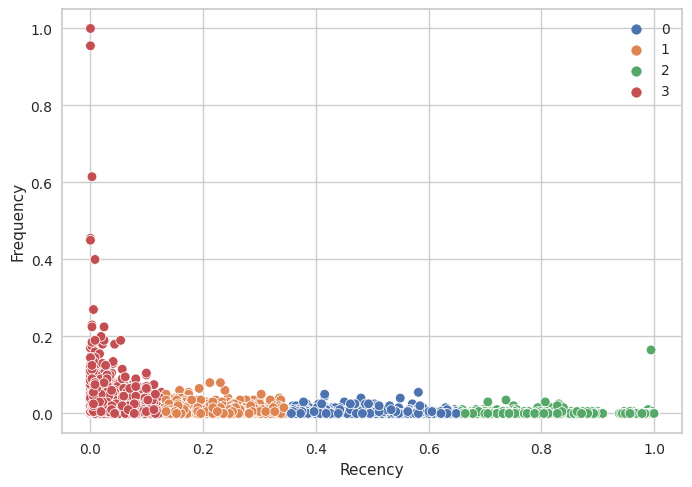
Nhận xét:

Biểu đồ K-means cho thấy rằng dữ liệu được chia thành 4 cụm. Cụm 1 chứa các điểm dữ liệu có giá trị Recency thấp, Frequency thấp và Monetary thấp. Cụm 2 chứa các điểm dữ liệu có giá trị Recency trung bình, Frequency trung bình và Monetary trung bình. Cụm 3 chứa các điểm dữ liệu có giá trị Recency cao, Frequency cao và Monetary cao. Cụm 4 chứa các điểm dữ liệu nằm ở giữa các cụm còn lại.

Có thể thấy rằng, các cụm này có xu hướng phân bố theo hình xoáy ốc, với các cụm có giá trị Recency thấp nằm ở trung tâm và các cụm có giá trị Recency cao nằm ở ngoài cùng. Điều này cho thấy rằng, các khách hàng có giá trị Recency thấp có xu hướng có giá trị Frequency và Monetary thấp hơn các khách hàng có giá trị Recency cao.

Ngoài ra, có thể thấy rằng, các cụm 1 và 4 có xu hướng phân bố theo hình elip. Điều này cho thấy rằng, các khách hàng trong các cụm này có xu hướng có giá trị Frequency và Monetary có sự chênh lệch lớn.





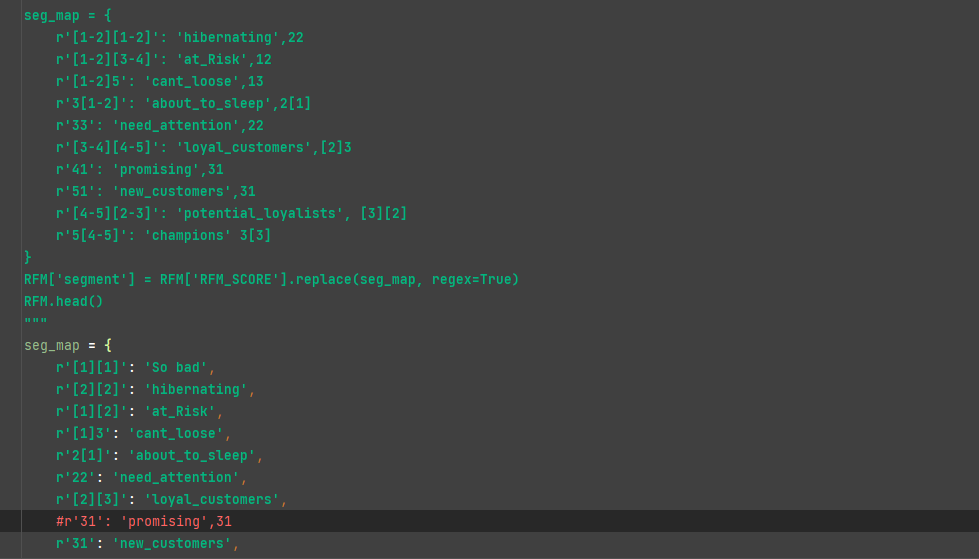
Biểu đồ phân cụm dữ liệu khách hàng:

* Cụm 1: Khách hàng mới hoặc khách hàng không hoạt động.
* Cụm 2: Khách hàng trung bình.
* Cụm 3: Khách hàng trung thành.
* Cụm 4: Khách hàng giá trị cao.
* Cụm 1: Đây là nhóm khách hàng ít hoạt động nhất. Họ đã không mua hàng trong một thời gian dài và có giá trị mua hàng thấp. Doanh nghiệp có thể cần phải nỗ lực hơn để thu hút nhóm khách hàng này.
* Cụm 2: Đây là nhóm khách hàng trung bình. Họ mua hàng với tần suất vừa phải và có giá trị mua hàng trung bình. Doanh nghiệp có thể tiếp tục giữ chân nhóm khách hàng này bằng cách cung cấp các ưu đãi hấp dẫn.
* Cụm 3: Đây là nhóm khách hàng hoạt động tích cực nhất. Họ mua hàng thường xuyên và có giá trị mua hàng cao. Doanh nghiệp có thể tập trung vào nhóm khách hàng này bằng cách cung cấp các sản phẩm và dịch vụ đáp ứng nhu cầu của họ.
* Cụm 4: Đây là nhóm khách hàng có giá trị mua hàng trung bình nhưng có sự chênh lệch lớn về tần suất mua hàng. Doanh nghiệp có thể phân tích thêm dữ liệu của nhóm khách hàng này để xác định các yếu tố ảnh hưởng đến tần suất mua hàng của họ.

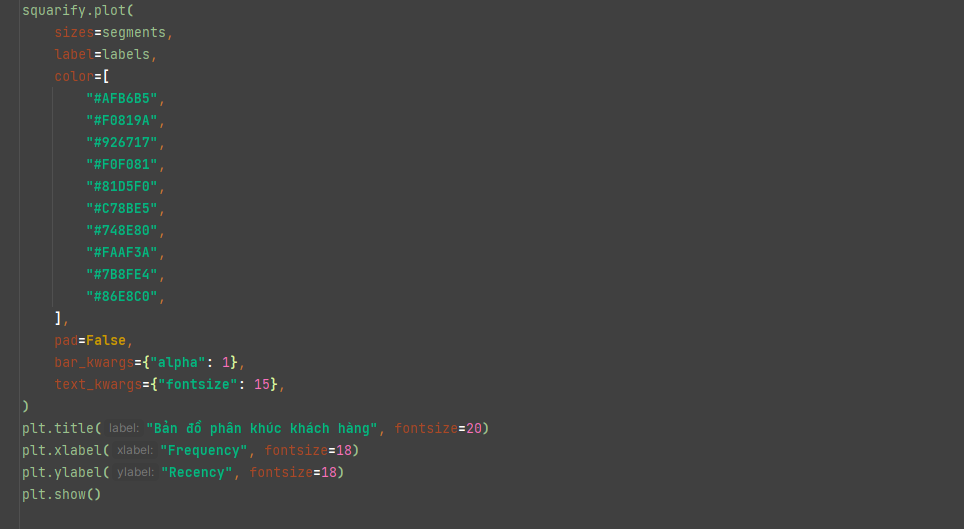
Nhìn chung, biểu đồ K-means cho thấy rằng, dữ liệu có thể được phân chia thành 4 cụm với các đặc điểm khác nhau. Điều này có thể giúp các doanh nghiệp hiểu rõ hơn về khách hàng của mình và đưa ra các chiến lược tiếp thị phù hợp.

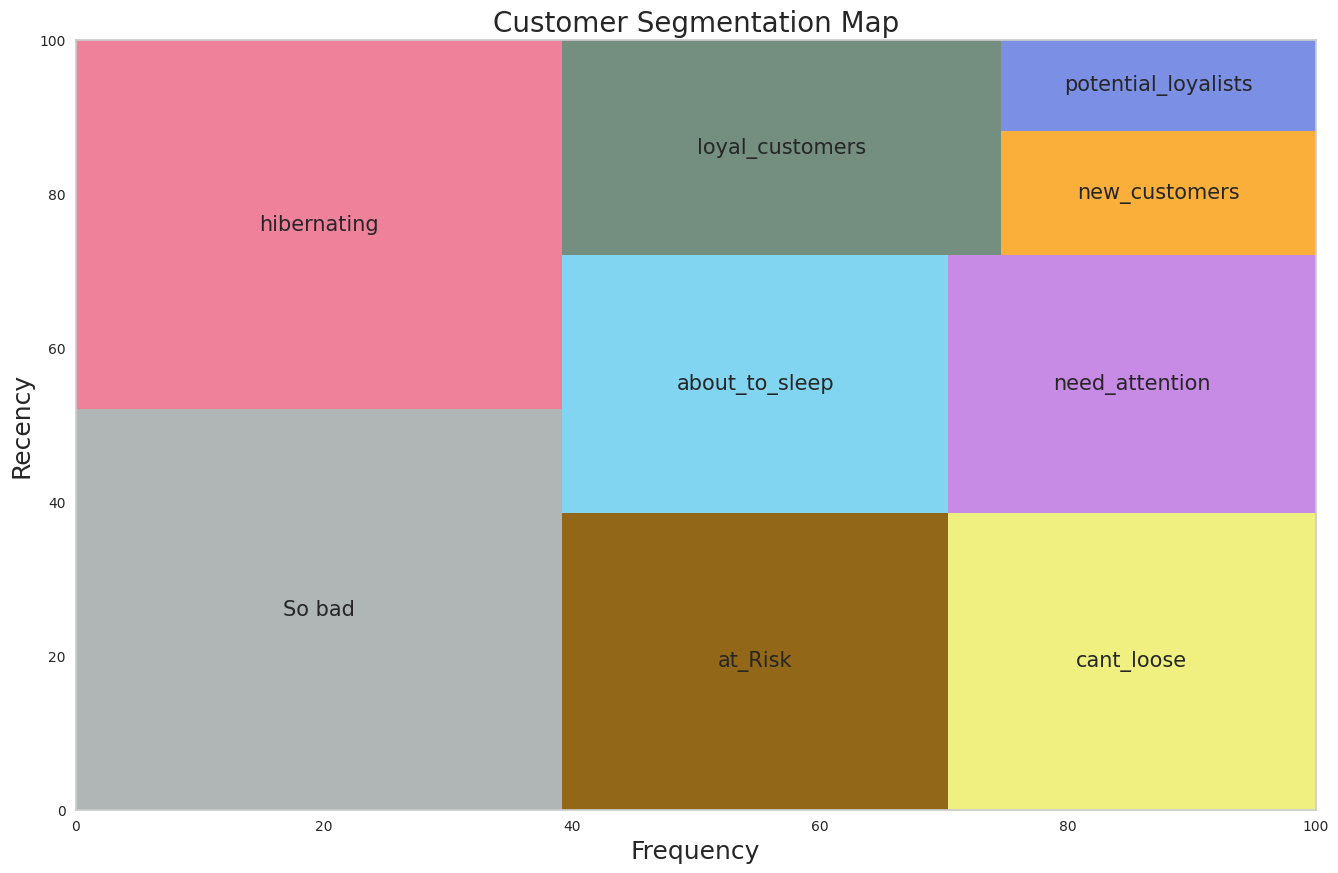
1. **Phân tích phân khúc:**









****

Biểu đồ phân tích phân khúc khách hàng dựa trên 3 yếu tố:

Recency: Khoảng thời gian kể từ lần mua hàng gần nhất của khách hàng.

Frequency: Số lần mua hàng của khách hàng trong một khoảng thời gian nhất định.

Monetary: Tổng giá trị của tất cả các giao dịch mua hàng của khách hàng.

Gồm các cụm:

**Cant\_loose** → Loại khách hàng thường xuyên mua sắm nhưng đã lâu không mua sắm và không nên để mất.

**At\_risk** → Tầng lớp khách hàng mua sắm tương đối thường xuyên nhưng đã lâu không mua sắm. ngủ đông → Khách hàng đang ngủ theo đúng nghĩa đen, vừa mua sắm không thường xuyên vừa đã lâu không mua sắm.

**About to sleep** → Nhóm khách hàng không thường xuyên mua sắm và đã lâu không mua sắm; họ đang hướng tới giấc ngủ.

**Need attention** → Đây là nhóm khách hàng nằm giữa đồ thị RF (33%), tiến tới nhóm rủi ro nếu không được giải quyết.

**New customers** → Nhóm khách hàng chưa mua sắm thường xuyên (có thể một lần) và mới mua sắm trong thời gian ngắn thì được coi là khách hàng mới.

**Promissing** → Khách hàng mới là phiên bản cũ của khách hàng mới :) họ là những khách hàng không mua sắm thường xuyên và chỉ mua sắm trong thời gian ngắn.

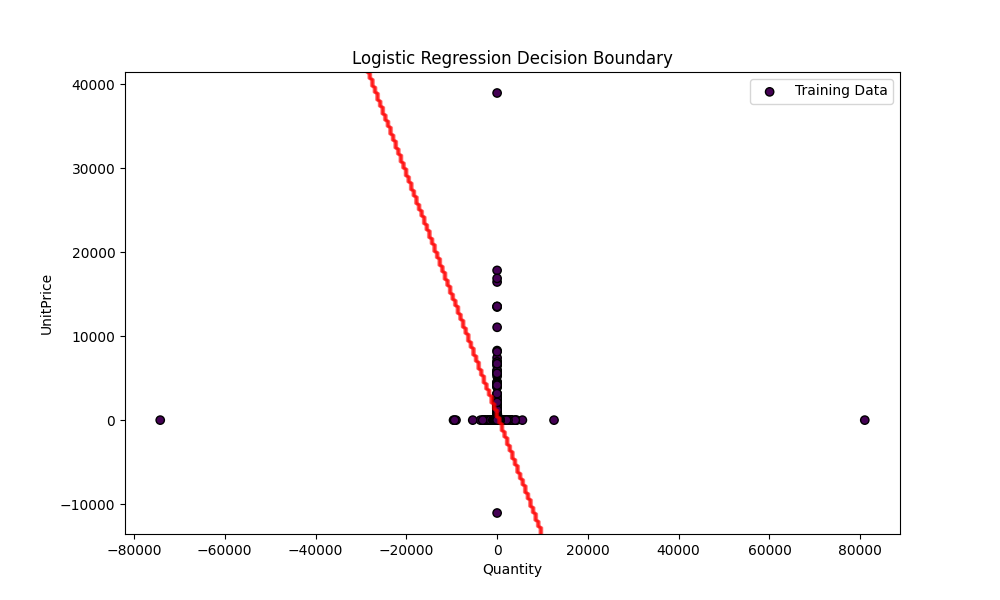
**Potential** **loyalists** → Những khách hàng mua sắm thường xuyên và chưa được lâu kể từ lần mua hàng cuối cùng của họ.

**Loyal** **customers** → Khách hàng mua sắm rất thường xuyên và đã một thời gian ngắn kể từ lần mua hàng cuối cùng của họ.

**Champions**  → Họ là nhà vô địch của chúng tôi, viên ngọc quý trên vương miện của chúng tôi! Những khách hàng mua sắm rất thường xuyên và thực hiện lần mua hàng cuối cùng trong một khoảng thời gian rất ngắn.

1. **Hồi quy logistic:**

Sử dụng mô hình Logistic Regression để dự đoán một biến phân loại (PurchaseArea) dựa trên hai biến độc lập (Quantity và UnitPrice)



*Ranh giới quyết định hồi quy logictic*

1. **Tập phổ biến:**

Sử dụng thuật toán Apriori để khai thác các tập phổ biến có trong dữ liệu:

**Cột 'support':**

Cột này đại diện cho hỗ trợ của mỗi tập mục. Hỗ trợ là tỷ lệ của số lượng hóa đơn chứa tập mục đó trên tổng số hóa đơn.

**Cột 'itemsets':**

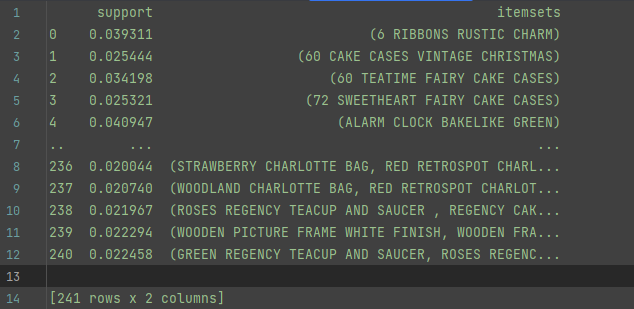
* Cột này chứa tập hợp các mục (sản phẩm) thuộc tập phổ biến. Mỗi dòng đại diện cho một tập hỗ trợ có hỗ trợ lớn hơn ngưỡng tối thiểu (**min\_support**).

Em chọn ngưỡng min\_sup = 0.02

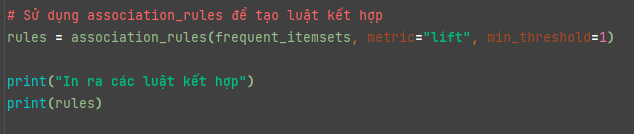
Sau khi chạy tập dữ liệu thì chúng ta thấy được:

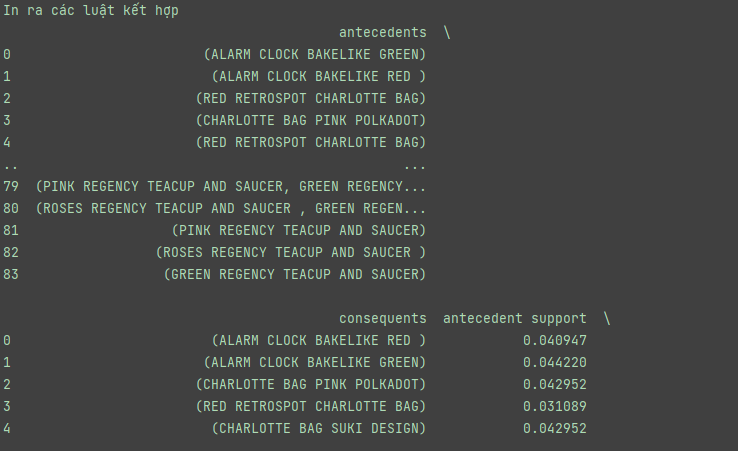
Dòng đầu tiên có hỗ trợ là 0.039311 và chứa một mục là '6 RIBBONS RUSTIC CHARM'. Điều này thể hiện là khoảng 3.93% hóa đơn chứa mục '6 RIBBONS RUSTIC CHARM'.

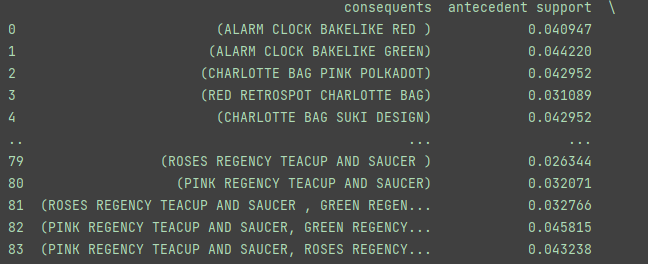
Dòng thứ hai có hỗ trợ là 0.025444 và chứa một mục là '60 CAKE CASES VINTAGE CHRISTMAS'. Điều này có thể được hiểu là khoảng 2.54% hóa đơn chứa mục '60 CAKE CASES VINTAGE CHRISTMAS'.

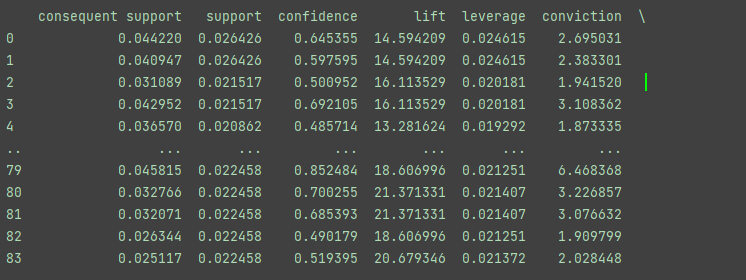
****

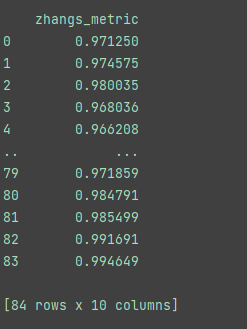
**Luật kết hợp**











# CHƯƠNG III: TỔNG KẾT

* 1. **Kết luận:**

Doanh nghiệp cần xây dựng các chiến lược tiếp thị phù hợp với từng nhóm khách hàng. Các chiến lược này cần tập trung vào việc đáp ứng nhu cầu và mong muốn của từng nhóm khách hàng để nâng cao nhu cầu mua sắm của khách hàng

Doanh nghiệp cần phân tích dữ liệu RFM để hiểu rõ hơn về từng nhóm khách hàng. Thông tin này sẽ giúp doanh nghiệp đưa ra các chiến lược phù hợp.

Khai thác các tập phổ biến cũng là kĩ thuật tính toán để phát triển cũng như xem xét nhóm sản phẩm nào mang lại lợi ích cao cho doanh nghiệp. Các tổ chức có thể sử dụng nó để cải thiện hiệu quả và năng suất của mình.

* 1. **Hạn chế:**

Chúng em sử dụng tập dữ liệu còn hạn chế về mặt thời gian nên không có sự so sánh lâu dài cũng như tìm hướng phát triển nhiều hơn, phạm vi nghiên cứu hạn hẹp còn nhiều mô hình chưa áp dụng để đánh giá các mối kết hợp giữa các mặt hàng với nhau. Kiến thức chuyên ngành còn nhiều thiếu sót trong nghiên cứu.

* 1. **Hướng phát triển:**

Mở rộng dữ liệu phân tích. Thêm các thuộc tính để nâng cao khả năng nghiên cứu.

Sử dụng nhiều mô hình phức tạp hơn để đánh giá và phân tích.

Ứng dụng các thông tin thực tế để đánh giá khách quan hơn.

# TÀI LIỆU THAM KHẢO

Slide bài giảng môn phân tích dữ liệu các chương

<https://www.kaggle.com/datasets/tunguz/online-retail/code>

<https://machinelearningcoban.com/2017/01/01/kmeans/>

<https://machinelearningcoban.com/tabml_book/ap/visualization.html>

<https://viblo.asia/p/khai-pha-mau-pho-bien-va-luat-ket-hop-gGJ59QAa5X2>

do-an-khoa-hoc-du-lieu-thay-nguyen-manh-tuan.pdf