**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC KINH TẾ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH**



**ĐỒ ÁN MÔN HỌC**  
BIỂU DIỄN TRỰC QUAN DỮ LIỆU

***Đề tài:***

***Biểu diễn trực quan dữ liệu kinh doanh của cửa hàng bán lẻ trực tuyến****.*

**Thành viên**:Nguyễn Phúc Hải

Lê Đức Dũng

Nguyễn Văn Hoàng Dũng

Võ Ngọc Dung

Nguyễn Quỳnh Khánh Hà

**Giảng viên**: TS. Nguyễn An Tế

*Thành phố Hồ Chí Minh , ngày 18 tháng 12 năm 2022*

MỤC LỤC

[MỤC LỤC HÌNH ẢNH 3](#_Toc122211301)

[LỜI CẢM ƠN 4](#_Toc122211302)

[Chương I: Tổng Quan Đề Tài 1](#_Toc122211303)

[1.     Giới thiệu đề tài 1](#_Toc122211304)

[2.     Mục tiêu nghiên cứu 1](#_Toc122211305)

[3.     Phương pháp nghiên cứu: 1](#_Toc122211306)

[4.     Tài nguyên sử dụng: 2](#_Toc122211307)

[Chương II: Tổng Quan Bộ Dữ Liệu 2](#_Toc122211308)

[1.     Tổng quan bộ dữ liệu thu thập 2](#_Toc122211309)

[2.     Các thuộc tính của bộ dữ liệu 2](#_Toc122211310)

[Chương III: Tiền xử lý dữ liệu 4](#_Toc122211311)

[1.     Exploratory Data Analysis (EDA) 4](#_Toc122211312)

[a.     Tổng quan bộ dữ liệu nguyên bản 4](#_Toc122211313)

[b. Biểu diễn dữ liệu nguyên bản: 6](#_Toc122211314)

[2. Chỉnh dạng dữ liệu 9](#_Toc122211315)

[a.     Thêm cột 9](#_Toc122211316)

[b.     Xoá cột: 10](#_Toc122211317)

[3.     Xử lý dữ liệu bị thiếu (Missing Values) 10](#_Toc122211318)

[4.  Xử lý outliers của các biến định lượng 12](#_Toc122211319)

[a.     Biến Totalsale 12](#_Toc122211320)

[b.   Biến Quantity 16](#_Toc122211321)

[5.   Xử lý biến Description 19](#_Toc122211322)

[Chương IV: Phân Tích Dữ Liệu 20](#_Toc122211323)

[1. Time Series: 20](#_Toc122211324)

[2. FP-Growth: 25](#_Toc122211325)

[Chương V: Machine Learning 33](#_Toc122211326)

[1. Cluster Category by Product Name (Description) - K-means: 33](#_Toc122211327)

[2. Customer Segmentation by Country and Behavior - K-Prototypes 42](#_Toc122211328)

[CHƯƠNG VI: KIỂM ĐỊNH VÀ GIẢM CHIỀU DỮ LIỆU 47](#_Toc122211329)

[1. Kiểm Định 47](#_Toc122211330)

[1.1. Chi-square test 47](#_Toc122211331)

[1.2. Kiểm định Levene 50](#_Toc122211332)

[1.3. Kiểm định Shapiro 52](#_Toc122211333)

[2. Giảm Chiều Dữ Liệu 53](#_Toc122211334)

[CHƯƠNG VI: KẾT LUẬN 57](#_Toc122211335)

[Tài Liệu Tham Khảo 58](#_Toc122211336)

[Bảng phân công: 59](#_Toc122211337)

MỤC LỤC HÌNH ẢNH

[Hình 1: Biểu đồ Scatter plot tương quan giữa Quantity và các biến khác. 7](#_Toc122199251)

[Hình 2: Heat map biểu diễn tương quan giữa các cột dữ liệu 8](#_Toc122199252)

[Hình 3: Scatter plot biểu diễn số lượng hàng bán ra ở mỗi quốc gia. 9](#_Toc122199253)

[Hình 4: Box plot của biến Totalsale trước khi xử lý Outliers. 12](#_Toc122199254)

[Hình 5: Box plot của biến Totalsale sau khi xử lý Outliers. 15](#_Toc122199255)

[Hình 6: Box plot của biến Quantity trước khi xử lý Outliers. 16](#_Toc122199256)

[Hình 7: Box plot của biến Quantity sau khi xử lý Outliers. 19](#_Toc122199257)

[Hình 8: Biểu đồ thanh thể hiện số lượng đặt hàng theo mùa 21](#_Toc122199258)

[Hình 9: Biểu đồ thanh thể hiện số lượng đơn hàng theo tháng. 22](#_Toc122199259)

[Hình 10: Biểu đồ thanh biểu diễn tỷ lệ đặt hàng của năm 2010 và 2011. 22](#_Toc122199260)

[Hình 11: Biểu đồ đường thể hiện tổng doanh thu theo tháng của năm 2010 và 2011. 23](#_Toc122199261)

[Hình 12: Biểu đồ thanh kết hợp với đường biểu diễn tổng doanh thu và tổng lượng hàng bán ra theo các tháng của năm 2011. 24](#_Toc122199262)

[Hình 13: Biểu đồ thanh ngang biểu diễn số lượng sản phẩm bán ra của 5 mặt hàng bán chạy nhất năm 2011. 24](#_Toc122199263)

[Hình 14: Biểu đồ đường biểu diễn doanh thu bán hàng của 5 mặt hàng bán chạy nhất năm 2011 theo các tháng chẵn. 25](#_Toc122199264)

[Hình 15: Biểu đồ thanh dọc thể hiện doanh số bán hàng theo mùa của 4 mặt hàng. 33](#_Toc122199265)

[Hình 16 36](#_Toc122199266)

[Hình 17: Biểu đồ cây thể hiện các nhóm sản phẩm. 39](#_Toc122199267)

[Hình 18: Biểu đồ thanh dọc biểu diễn tổng số lượng hàng bán ra của 5 loại hàng bán chạy nhất. 40](#_Toc122199268)

[Hình 19: Biểu đồ thanh dọc biểu diễn tổng số đơn đặt hàng cho 5 loại hàng bán chạy nhất. 41](#_Toc122199269)

[Hình 20: Biểu đồ thanh dọc biểu diễn tổng doanh thu của 5 loại hàng bán chạy nhất mang lại. 41](#_Toc122199270)

[Hình 21: Biểu đồ Paired Plot phân cụm theo Country và Customer Behaviour. 54](#_Toc122199271)

[Hình 22: Biểu đồ phân tán sau khi giảm chiều dữ liệu. 56](#_Toc122199272)

**LỜI CẢM ƠN**

Hiện nay, trong thời đại 4.0, mặc dù đã có sự xuất hiện của trí tuệ nhân tạo giúp hỗ trợ cho công việc phân tích kết quả kinh doanh nhằm dự báo, đưa ra các quyết định kinh doanh giúp tối ưu hóa doanh thu cho doanh nghiệp nhưng sự cần thiết của những người làm công việc phân tích dữ liệu vẫn là rất cần thiết đối với các doanh nghiệp. Ngoài công việc khai thác các thông tin, các vấn đề từ những dữ liệu kinh doanh, người làm phân tích dữ liệu còn phải biết thể hiện, truyền tải các thông tin đó một cách đầy đủ, chính xác, và dễ hiểu, dễ nhận thấy cho các đối tượng mình muốn truyền tải, các nhà đầu tư, các doanh nghiệp, khách hàng. Chính vì vậy, nhóm đã quyết định chọn bộ dữ liệu “Online Retails Sale Dataset” làm bộ dữ liệu để làm báo cáo cho môn học “Biểu diễn trực quan dữ liệu”. Mục đích của việc phân tích bộ dữ liệu này là để tìm ra các mặt hàng được bán chạy theo combo, thời gian; tìm ra các khách hàng tiềm năng để từ đó có thể trực quan hóa, giúp các nhà đầu tư, khách hàng nhìn nhận, thấy được đúng các vấn đề, các insight, các thông tin mà những người phân tích dữ liệu, chúng em, muốn truyền tải đến.

Trong quá trình làm đồ án môn học vẫn còn các hạn chế, sai sót, chưa tối ưu hóa về mặt kiến thức, kỹ thuật. Nhóm chúng em mong sẽ nhận được sự phản hồi, nhận xét của thầy cô giảng viên hướng dẫn để cải thiện các điểm này.

Đặc biệt nhóm xin được gửi lời cảm ơn đến thầy Nguyễn An Tế, giảng viên hướng dẫn của học phần “Biểu diễn trực quan dữ liệu” này. Thầy đã giúp đỡ, hướng dẫn, cung cấp các tài liệu, kiến thức, kỹ năng cần thiết để nhóm em có thể hoàn thành báo cáo Đồ án kết thúc môn học này.

Chúng em xin chân thành cảm ơn thầy.

Thay mặt nhóm sinh viên thực hiện đồ án,

Hải,

Nguyễn Phúc Hải

# Chương I: Tổng Quan Đề Tài

## 1.     Giới thiệu đề tài

Ngày nay, mạng Internet là một trong những công cụ cần thiết trong cuộc sống hiện đại, là nền tảng cho sự truyền tải và trao đổi thông tin trên toàn cầu. Cùng với sự phát triển nhảy vọt của công nghệ thông tin, Internet đang dần chiếm giữ vai trò quan trọng trong mọi mặt của đời sống, giúp con người làm việc với độ chính xác cao, quản lý và tổ chức công việc hiệu quả, cũng như nhanh chóng cập nhật thông tin một cách chính xác.

Vì vậy, thương mại điện tử ngày càng phát triển và đem lại bước đột phá mới cho công tác quản lý bán hàng. Nó giúp doanh nghiệp dễ dàng nắm bắt thông tin, dữ liệu về người tiêu dùng, hàng hóa và các đơn đặt hàng một cách nhanh chóng. Cùng với đó là sự phát triển của việc phân tích dữ liệu khách hàng, giúp cho quản lý doanh nghiệp có cái nhìn tổng quát về tình hình bán hàng của doanh nghiệp, khai thác được những thông tin hữu ích từ bộ dữ liệu mả họ đang có để từ đó đưa ra những chính sách bán hàng hiệu quả.

## 2.     Mục tiêu nghiên cứu

Phân tích bộ dữ liệu bán hàng giúp doanh nghiệp tìm ra các sản phẩm có khả năng thường xuyên được bán cùng nhau, chuỗi thời gian bán hàng tiềm năng, phân khúc khách hàng theo vị trí địa lý và hành vi mua hàng. Từ đó giúp đưa ra các chiến lược quảng cáo và phát triển sản phẩm phù hợp với từng vị trí địa lý, thúc đẩy khả năng bán chéo giữa các sản phẩm và các chiến dịch quảng bá theo mùa.

Hình ảnh hoá các dữ liệu nhằm dễ dàng đưa ra các so sánh trực quan, tính toán tỷ trọng, nhận biết trend, phát hiện outlier, nhận diện đặc điểm phân phối của biến tốt hơn.

## 3.     Phương pháp nghiên cứu:

* EDA: Sử dụng các biểu đồ vẽ nhằm tương quan cũng như làm rõ mục đích nghiên cứu đề tài, sự liên kết với nhau giữa các biến.
* FP-Growth: Sử dụng FP-Growth để tìm ra các mặt hàng thường được bán chung với nhau (trong cùng 1 hóa đơn) và khoảng thời gian mà các mặt hàng được bán chạy nhất giúp tối ưu hóa lượng hàng được bán ra.
* Các loại biểu đồ: Sử dụng các loại biểu đồ chuyên dụng và phù hợp với mục đích trực quan hoá các dữ liệu, giúp người đọc báo cáo dễ dàng quan sát và đánh giá.
* Kiểm định Chi-Squared: Kiểm định tính độc lập giữa 2 biến phân loại, xác định xem liệu có mối liên hệ giữa 2 biến phân loại hay không
* Kiểm định ANOVA: một kỹ thuật thống kê tham số được sử dụng để phân tích sự khác nhau giữa giá trị trung bình của các biến phụ thuộc với nhau, thay vì chỉ so sánh các đối tượng trong một nhóm nghiên cứu, phân tích ANOVA giúp so sánh trong phạm vi rộng hơn, giữa hai hoặc nhiều nhóm đối tượng.
* SVD: kĩ thuật giảm chiều dữ liệu dựa trên kĩ thuật phép chiếu các dữ liệu lên một chiều khác, lợi dụng sự tương quan giữa các chiều dữ liệu để giảm chiều biểu diễn dữ liệu mà không gây ra quá nhiều sai số

## 4.     Tài nguyên sử dụng:

-       Ngôn ngữ lập trình: Python.

-       Bộ dữ liệu “Online Retails Sale Data” được lấy từ Kaggle.

# Chương II: Tổng Quan Bộ Dữ Liệu

## 1.     Tổng quan bộ dữ liệu thu thập

-       Bộ dữ liệu “Online Retails Sale Data” chứa các giao dịch từ ngày 12/01/2010 đến 12/09/2011 của một công ty bán lẻ trực tuyến có trụ sở tại UK.

Bao gồm 10 thuộc tính, số dòng của bộ dữ liệu là 541909 dòng

## 2.     Các thuộc tính của bộ dữ liệu

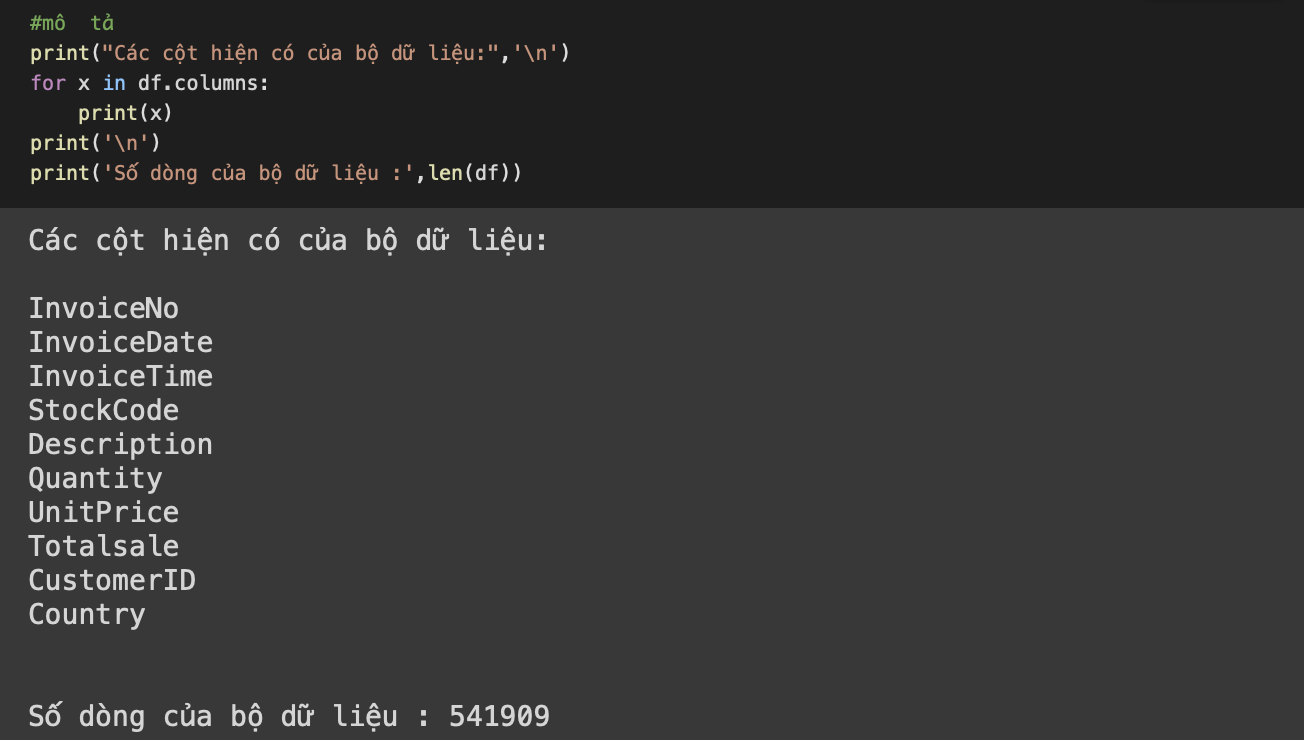
|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Mô tả** | **Ghi chú** |
| InvoiceNo | Mã đơn hàng cho mỗi order thành công | Bao gồm 25900 mã đơn hàng khác nhau |
| InvoiceDate | Ngày đặt hàng | Từ ngày 12 tháng 1 năm 2010 đến ngày 10 tháng 12 năm 2011 |
| InvoiceTime | Thời gian đặt hàng | Phụ thuộc vào vị trí địa lí của từng khách hàng |
| StockCode | Mã sản phẩm | Bao gồm 4070 mã sản phẩm khác nhau |
| Description | Mô tả sản phẩm | Bao gồm 4224 mô tả, trong đó có các sản phẩm bị lỗi, hư hỏng, bị trả về, thất lạc |
| Quantity | Số lượng mỗi sản phẩm | Số lượng của mỗi đơn vị sản phẩm được bán ra trong 1 đơn hàng. |
| UnitPrice | Giá bán mỗi đơn vị sản phẩm | Giá của 1 sản phẩm, mỗi sản phẩm sẽ có mức giá khác nhau. |
| Totalsale | Doanh thu từ mỗi sản phẩm trong 1 đơn hàng. | Totalsale = Quantity \* UnitPrice |
| CustomerID | Mã khách hàng | Mỗi khách hàng sẽ được cấp 1 mã duy nhất. |
| Country | Quốc gia của khách đặt hàng | Căn cứ vào vị trí địa lý của khách hàng. |

# Chương III: Tiền xử lý dữ liệu

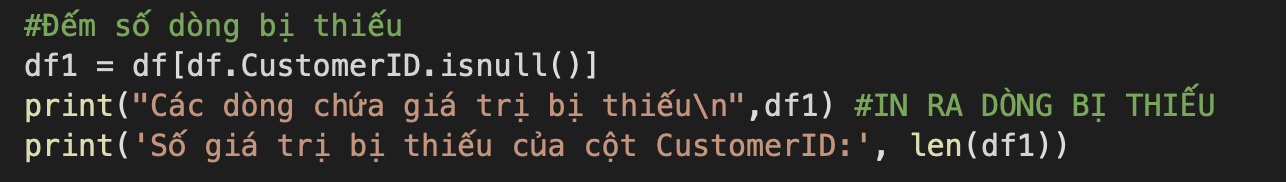
## 1.     Exploratory Data Analysis (EDA)

### a.     Tổng quan bộ dữ liệu nguyên bản

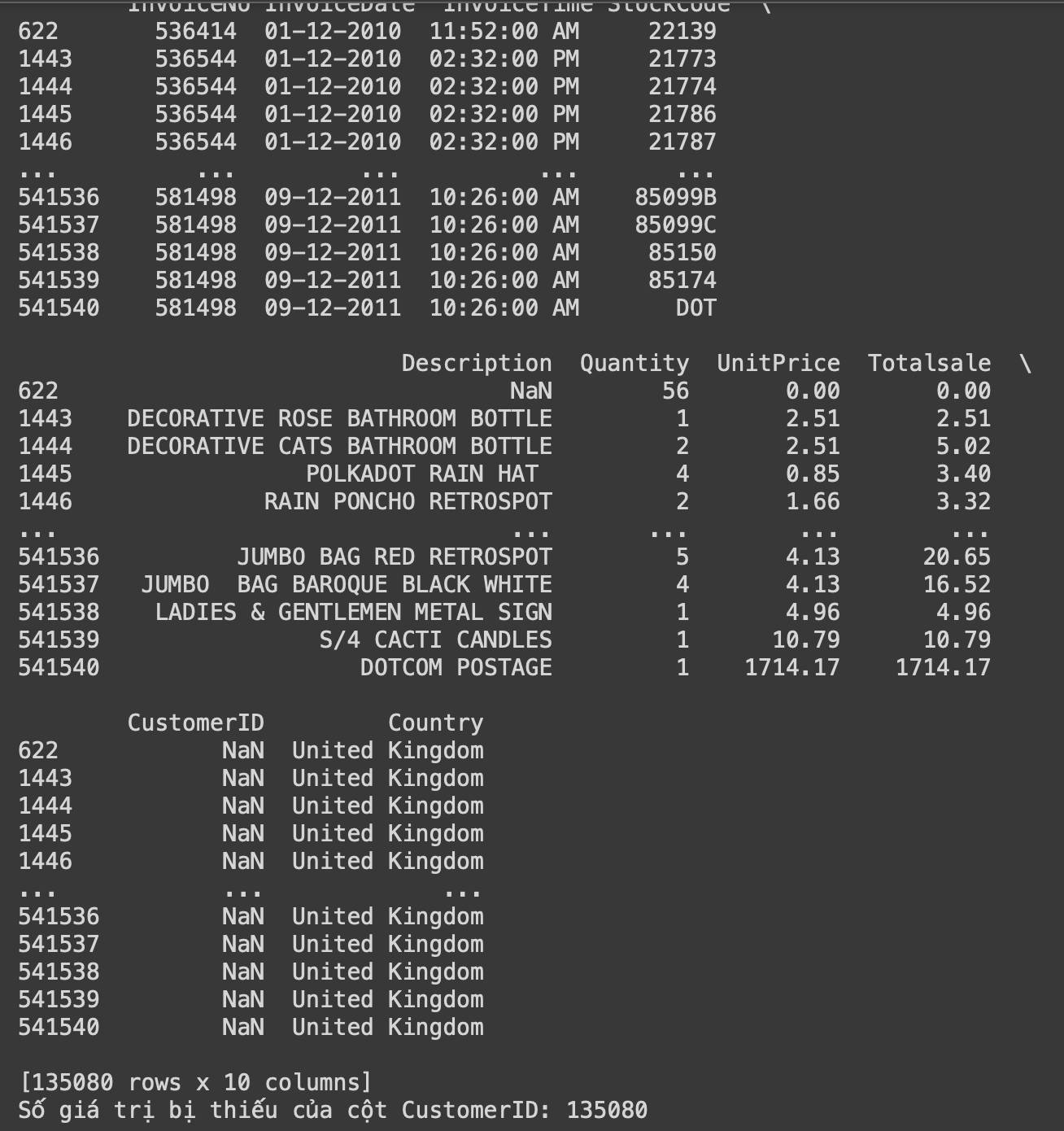
* Để thăm dò bộ dữ liệu, ta cần biết được tổng quan các thông tin về: số dòng, số cột, có tồn tại giá trị bị thiếu hay không, nếu có thì ở dòng nào, thuộc cột nào và chiếm bao nhiêu phần trăm của bộ dữ liệu.
* Xem số dòng, số cột hiện có của bộ dữ liệu nguyên bản để nắm được các thông tin sơ lược trước khi tiến hành tiền xử lý:



-       Tiến hành kiểm tra các dòng chứa giá trị bị thiếu:

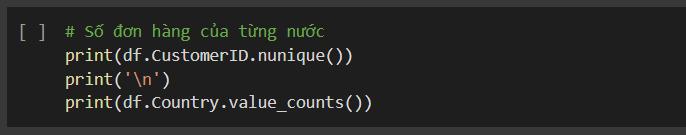


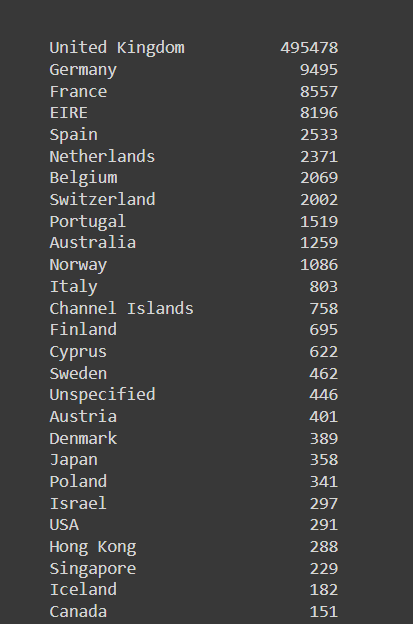
* Kết quả trả về:



*Nhận xét: Bộ dữ liệu trên có 10 cột, số dòng dữ liệu là**541909, tồn tại cột CustomerID với số giá trị bị thiếu là 135080, tức xấp xỉ 25% bộ dữ liệu tồn tại giá trị bị thiếu ở cột này. Vì vậy ta sẽ xử lý các giá trị thiếu này ở bước sau.*

* Để khám phá dữ liệu, chúng ta sẽ đưa ra số đơn hàng của từng nước, để tiện trong việc chọn các phân cụm clustering sau này

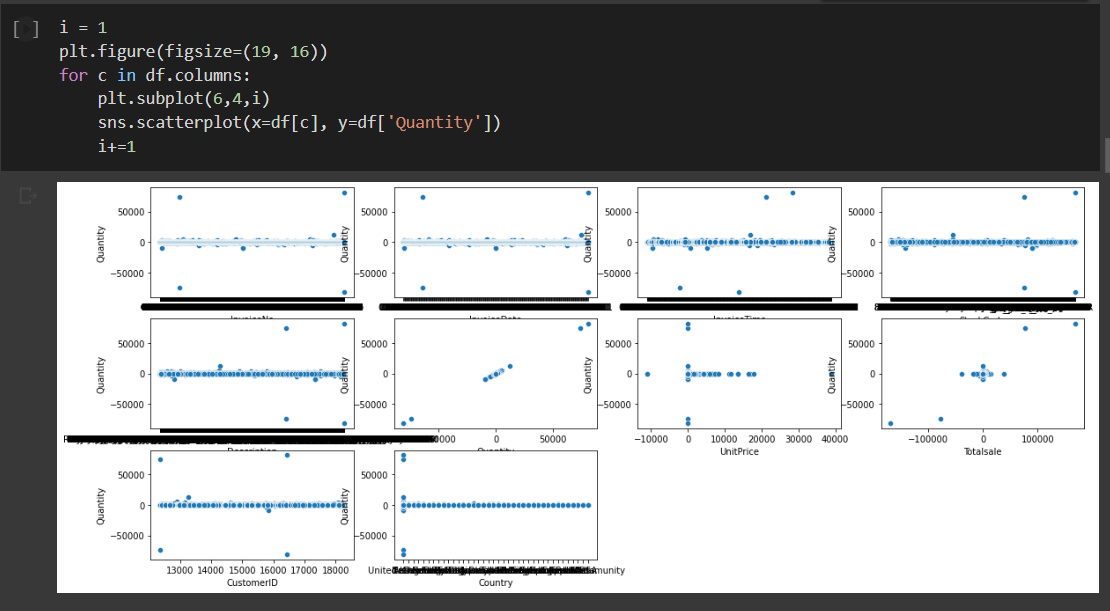
******

******

*-       Nhận thấy United Kingdom chiếm rất nhiều trong số đơn hàng (> 90%/ tổng số quốc gia), điều này giúp ta định hướng rằng, chúng ta sẽ tạo nhãn có thuộc tính country thành 2 loại : United Kingdom và các quốc gia khác.*

### b. Biểu diễn dữ liệu nguyên bản:

* Tiếp theo, việc phác hoạ biểu đồ tương quan giữa các biến làm nổi bật target của vấn đề. Qua biểu đồ cũng như các công thức liên quan, nhận thấy target của bộ dữ liệu này có thể sử dụng được biến Quantity, Totalsale cũng như UnitPrice như nhau



Hình 1: Biểu đồ Scatter plot tương quan giữa Quantity và các biến khác.

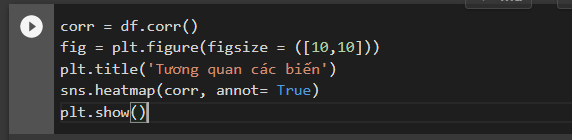
**Nhận xét**: Sử dụng Scatter diagram để biểu diễn mối tương quan giữa biến Quantity so với các biến còn lại là Totalsale, UnitPrice, Country, CustomerID.

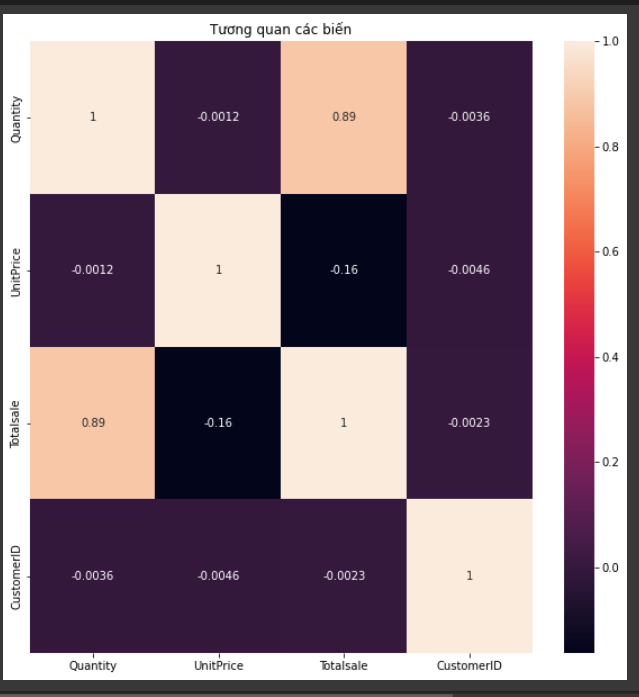
Ngoài ra, nhóm còn sử dụng biểu đồ nhiệt (Heat map) để xem xét sự tương quan giữa các cột dữ liệu. Ô nào có màu sắc có cường độ ánh sáng càng mạnh sẽ mang giá trị càng lớn, ngược lại, màu sắc có cường độ ánh sáng càng nhạt sẽ mang giá trị nhỏ hơn.

* Hệ số tương quan có giá trị âm cho thấy hai biến có mối quan hệ nghịch biến hoặc tương quan âm (nghịch biến tuyệt đối khi giá trị bằng -1)
* Hệ số tương quan có giá trị dương cho thấy mối quan hệ đồng biến hoặc tương quan dương (đồng biến tuyệt đối khi giá trị bằng 1)
* Tương quan bằng 0 cho hai biến độc lập với nhau.

**Đánh giá biểu đồ**: Do vấn đề khách quan (bộ dữ liệu có nhiều dòng dữ liệu) nên cột x của các biểu đồ bị “đen đặc”, các tên cột y bị dính vào nhau gây khó nhìn.

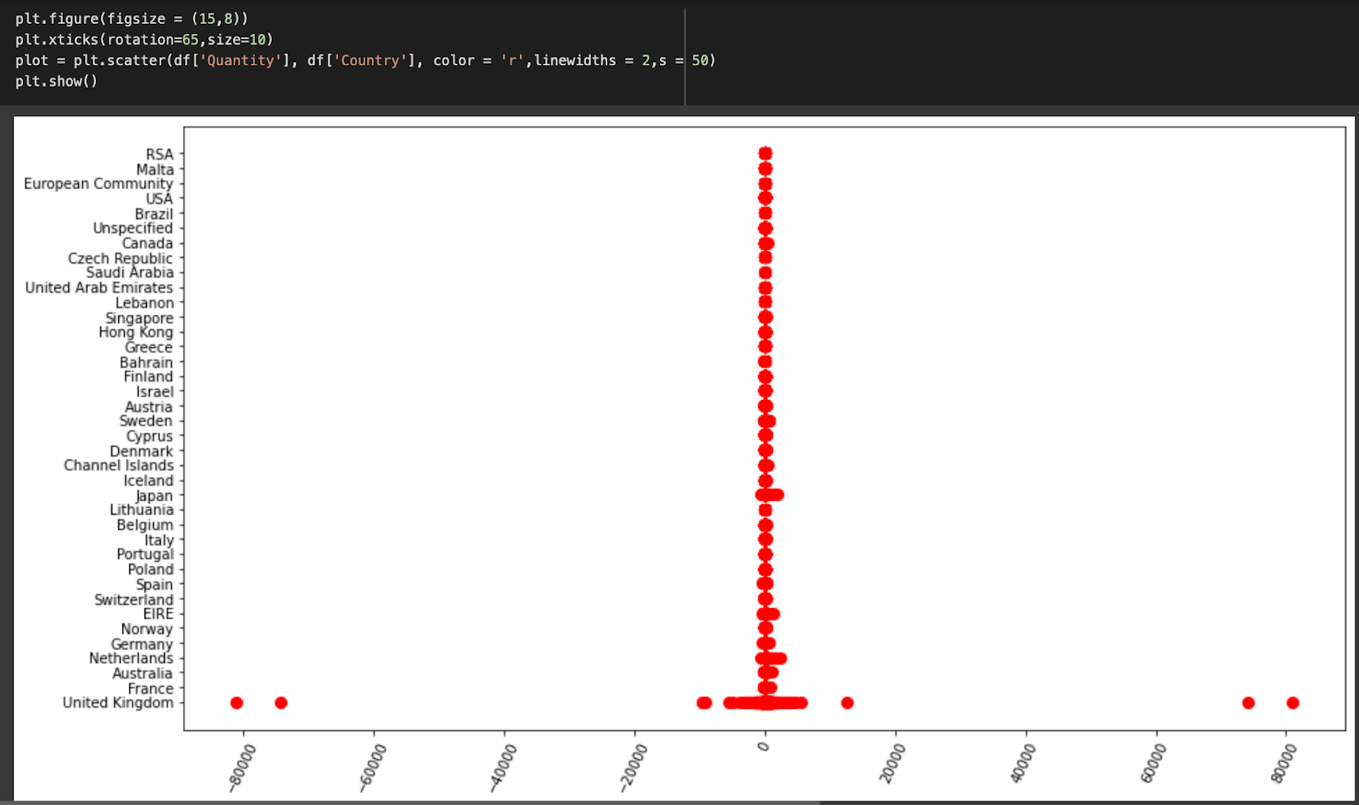
**Cải thiện**: Thay đổi chiều biểu diễn của subplot từ 6, 4 thành 4, 3 đã giúp cải thiện được vấn đề về tên cột bị dính vào nhau tuy là vẫn chưa cải thiện được vấn đề khách quan nhưng nhìn chung các biểu đồ đã trở nên dễ nhìn hơn





Hình 2: Heat map biểu diễn tương quan giữa các cột dữ liệu

* Kiểm tra xem bộ dữ liệu có tồn tại Outliers :

**

Hình 3: Scatter plot biểu diễn số lượng hàng bán ra ở mỗi quốc gia.

**Nhận xét***:* Sử dụng biểu đồ điểm (Scatter plot) để biểu diễn sự phân phối của biến Quantity theo biến Country. Ở đây ta thấy được số lượng hàng bán ra ở các quốc gia khác United Kingdom chỉ nằm trong khoảng từ 0 đến 20000, trong khi đó ở United Kingdom thì có giá trị vượt ngoài 80000, vậy có thể tồn tại giá trị Outliers cần được loại bỏ.

**Đánh giá biểu đồ**: Biểu đồ phù hợp để thể hiện, trực quan hóa được thông tin về việc có các giá trị outlier giúp người xem dễ nhận biết, màu sắc dễ nhìn. Các giá trị của các cột x, y tương đối nhìn được ở văn bản thường

## 2. Chỉnh dạng dữ liệu

### a.     Thêm cột

* Thêm vào cột Month:

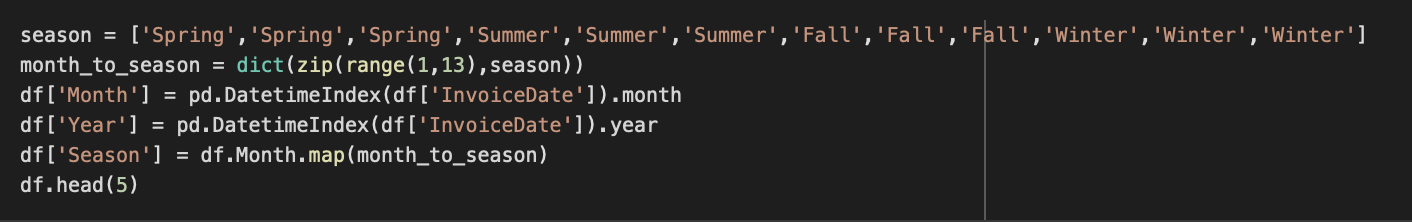
+ Lý do: Với mục tiêu phân loại các hóa đơn theo mùa, ta cần sắp xếp thời gian của các hoá đơn theo tháng (thay vì sắp xếp theo ngày) để thuận tiện cho việc phân loại. Nếu sắp xếp theo ngày thì không có ý nghĩa phân tích nhiều như tháng vì các ngày sẽ có sự trùng lặp rất nhiều.

* Thêm vào cột Year:

+ Lý do: Bộ dữ liệu bao gồm thời gian bán hàng của năm 2010 và 2011, ta cần phân loại hoá đơn dựa theo năm để phục vụ cho việc so sánh 2 năm, từ đó biết được xu hướng mua hàng của người mua tập trung vào loại sản phẩm nào.

* Thêm vào cột Season:

+ Lý do: Để phục vụ cho mục tiêu nghiên cứu ở mục 2: tìm ra các sản phẩm bán chạy theo mùa. Ta cần thêm vào bộ dữ liệu cột Season nhằm sắp xếp các hoá đơn có tháng 1/2/3 vào mùa xuân, tháng 4/5/6 vào mùa hè, tháng 7/8/9 vào mùa thu và các tháng 10/11/12 vào mùa đông. Như vậy, ở phần phân tích dữ liệu sẽ dễ dàng thực hiện hơn.

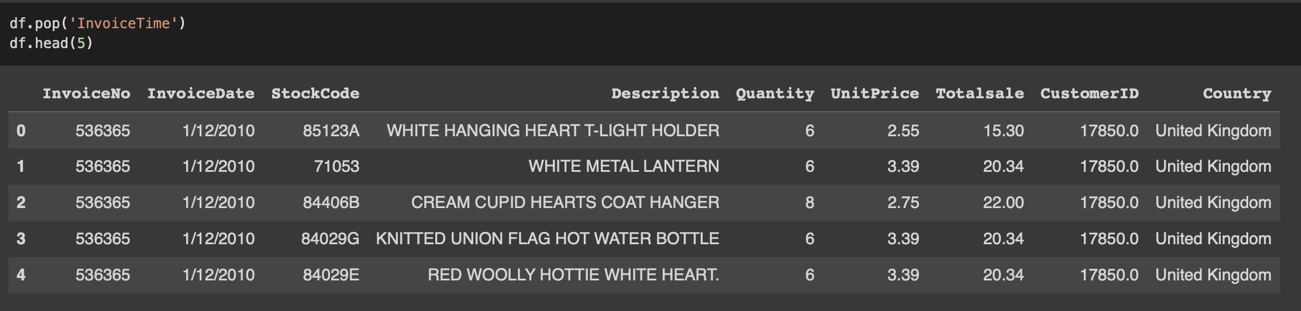




### b.     Xoá cột:

* Loại bỏ cột InvoiceTime.

+ Lý do: Dữ liệu thời gian của tệp dữ liệu này không nhất quán giữa các khách hàng. Phụ thuộc vào mỗi vị trí địa lý khác nhau nên múi giờ đặt hàng sẽ có sự chênh lệch. Vì vậy, để khách quan và bao quát hơn, nhóm chỉ sử dụng dữ liệu dạng ngày của cột InvoiceDate

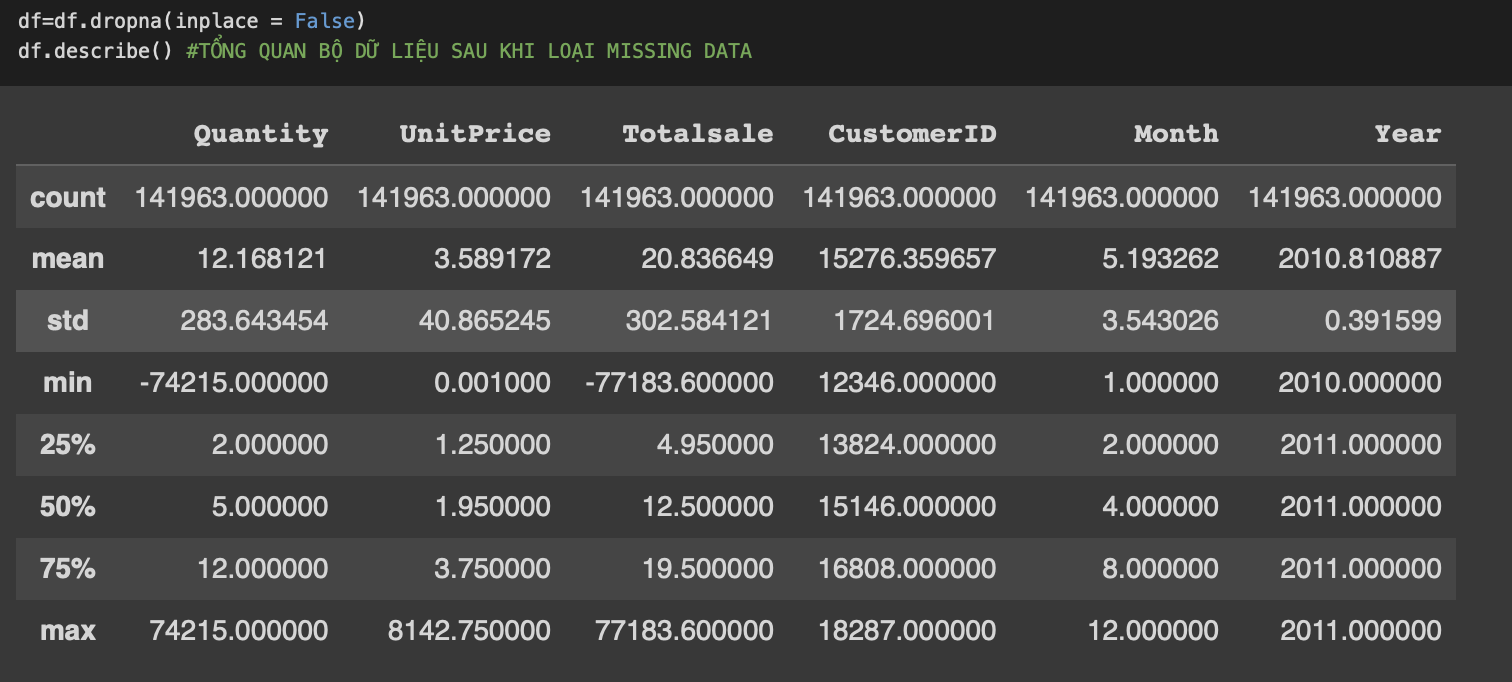


## 3.     Xử lý dữ liệu bị thiếu (Missing Values)

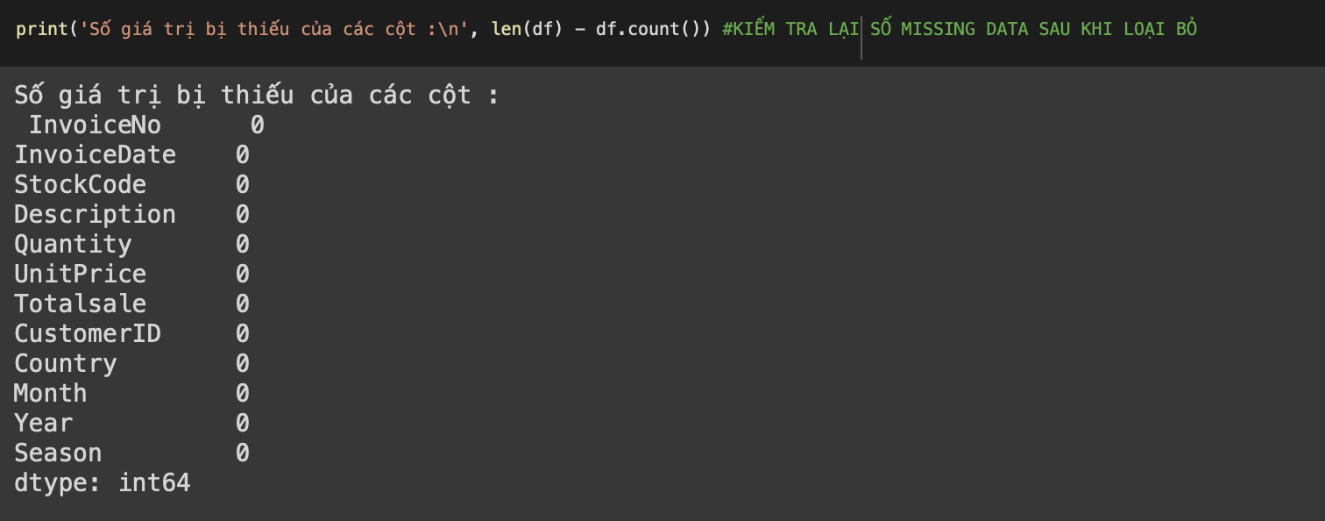
* Tiến hành kiểm tra số lượng dòng chứa giá trị bị thiếu của mỗi cột:



* Quan sát từ kết quả, ta thấy rằng chỉ có cột CustomerID là còn tồn lại missing values. Ta tiến hành loại bỏ luôn các dòng này ra khỏi bộ dữ liệu:



* Kiểm tra lại bộ dữ liệu sau khi xử lý:



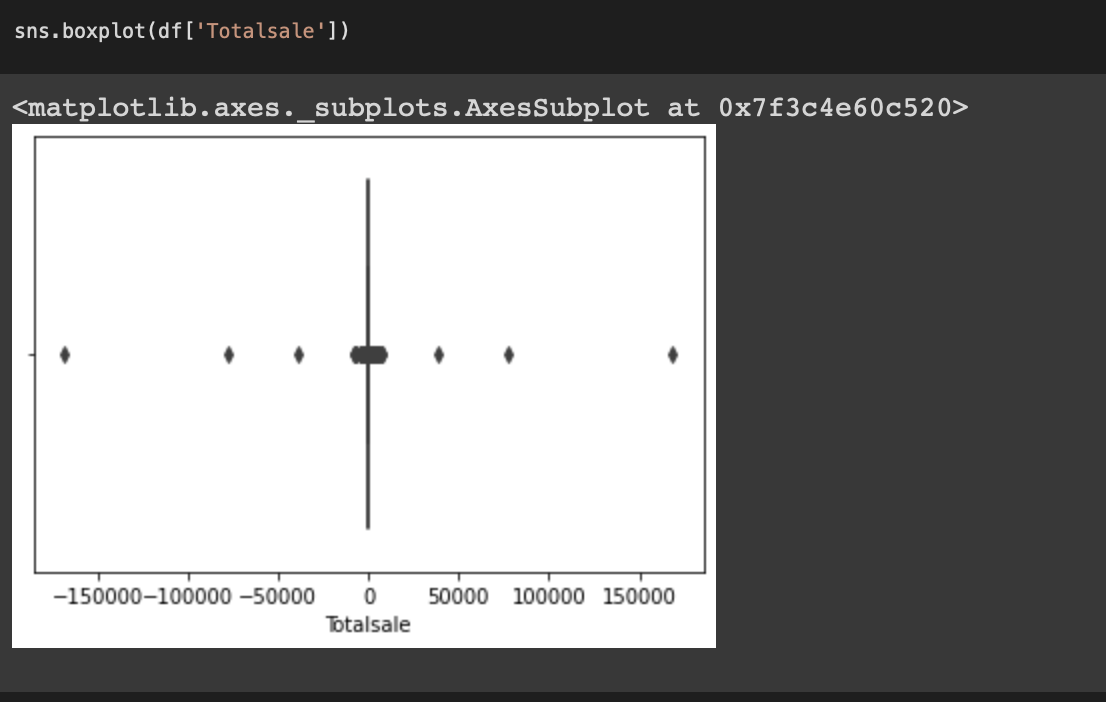
* Nhận xét: Sau khi đã bỏ đi các dòng chứa giá trị bị thiếu của cột CustomerID, bộ dữ liệu đã không còn missing values ở bất kỳ cột nào.



## 4.  Xử lý outliers của các biến định lượng

### a.     Biến Totalsale

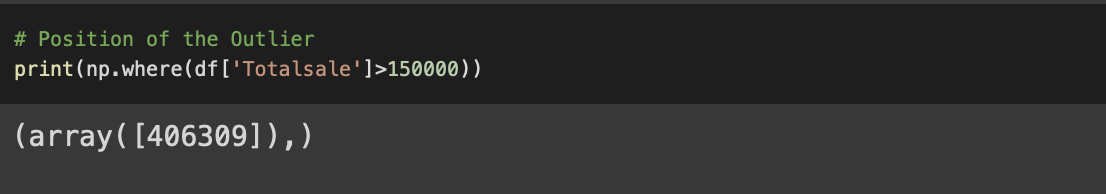
* Để quan sát trực quan dữ liệu của biến Totalsale, ta sẽ sử dụng biểu đồ hộp Boxplot để thể hiện các phân phối dữ liệu, nghĩa là giúp chúng ta biết được độ dàn trải của các điểm dữ liệu như thế nào, dữ liệu có đối xứng không, phân bố rộng hay hẹp, giá trị nhỏ nhất, lớn nhất và các điểm ngoại lệ.



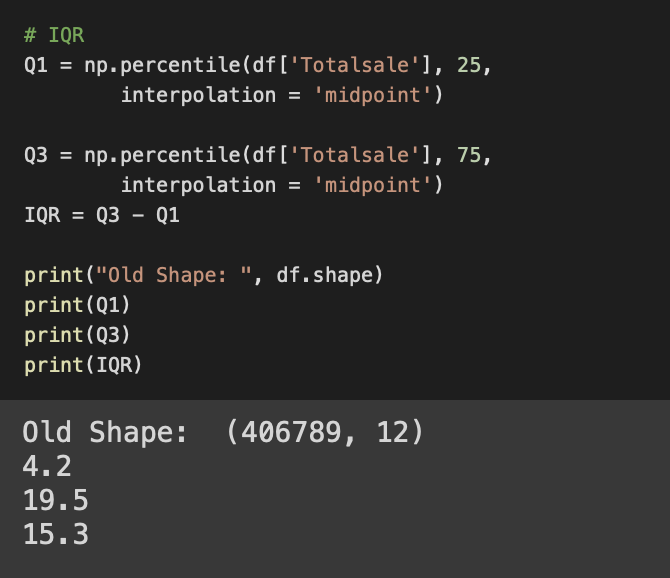
Hình 4: Box plot của biến Totalsale trước khi xử lý Outliers.

**Nhận xét**: Dựa vào Boxplot phía trên, ta thấy giá trị Max = 150000 và Min = -150000 lần lượt là giá trị lớn nhất và bé nhất để một phần tử được xem là không phải Outliers. Vậy để xác định được Outliers của biến Total Sale, ta chỉ cần tìm các phần tử vượt quá 2 giá trị này.

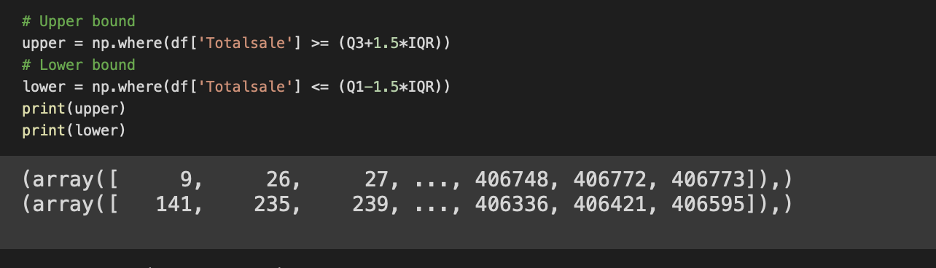
**Đánh giá biểu đồ**: Biểu đồ phù hợp để thể hiện, trực quan hóa được thông tin về việc có các giá trị outlier giúp người xem dễ nhận biết, màu sắc dễ nhìn. Các giá trị của các dòng thể hiện giá trị của biểu đồ boxplot tương đối nhìn được ở văn bản thường. Một phần do dữ liệu dày đặc nên không nhìn thấy được “hộp” rõ ràng mà chỉ nhìn được 1 đường thẳng.



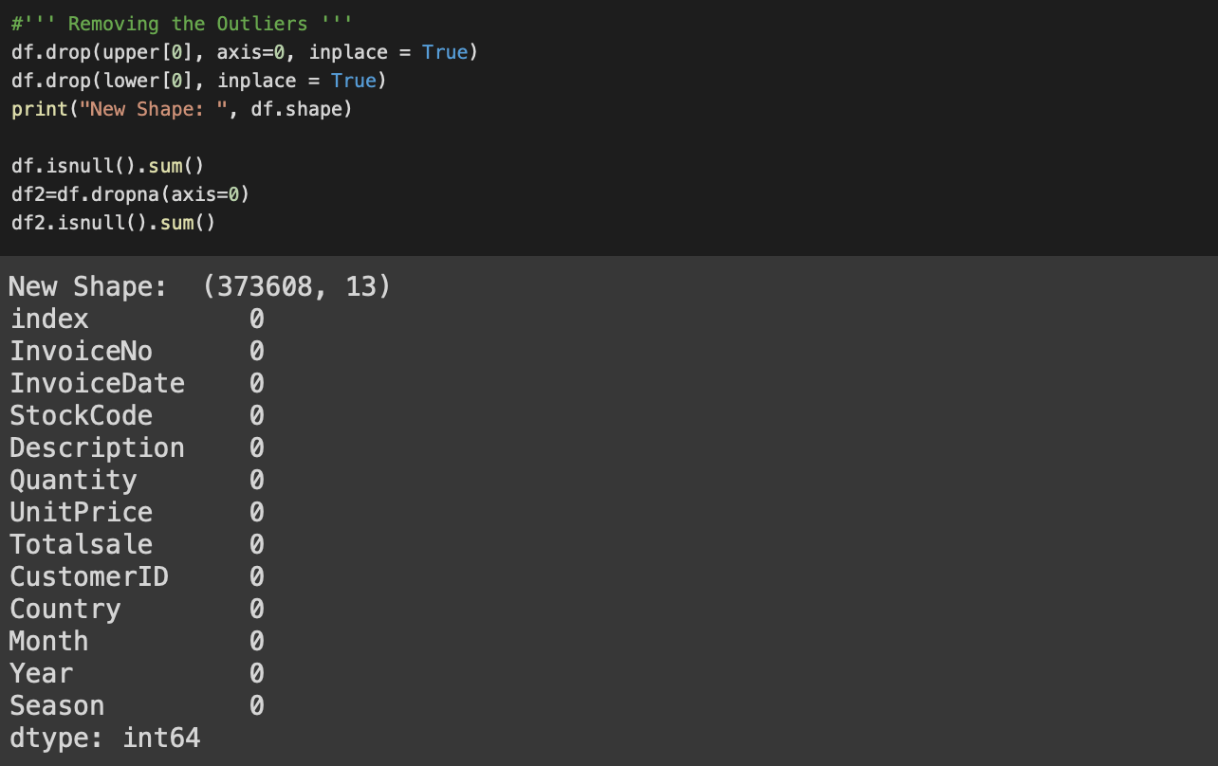
* Xác định giá trị của các tứ phân vị nhằm kiểm soát tác động của những giá trị ngoại lệ (Outliers) ở 2 đầu mút. Ở đây ta chỉ cần xác định giá trị của tứ phân vị thứ 1 và 3:



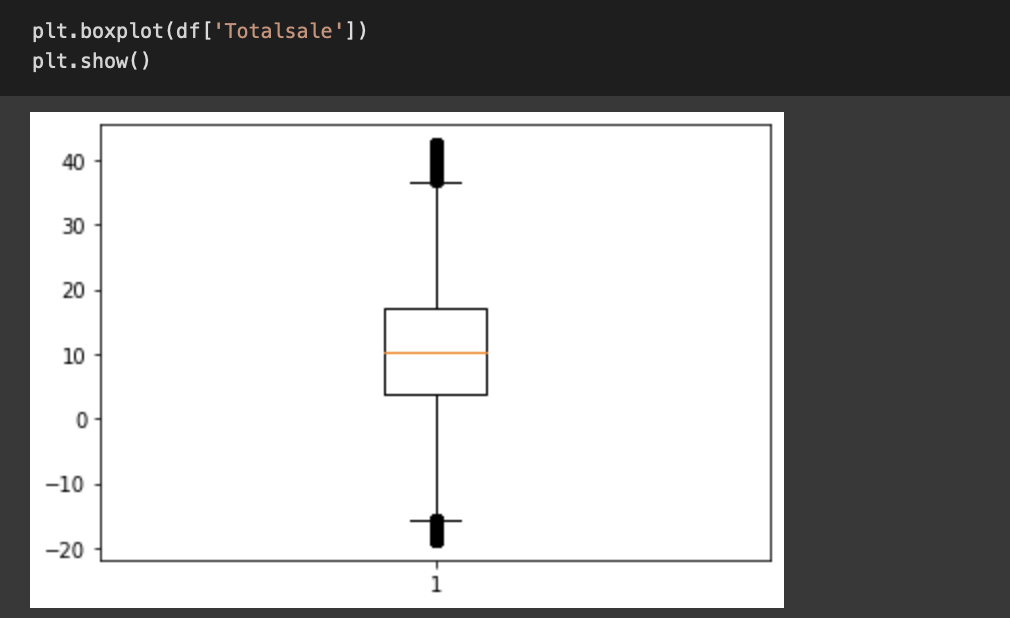
* Sau khi đã có Q1 và Q3, ta tiếp tục xác định giá trị biên của biến Total Sale, nếu giá trị của 1 phần tử bất kì nằm trong khoảng lower và upper thì gọi là expected data, ngược lại gọi là Outliers. Tạo 2 mảng chứa các index của phần tử là Outliers.



* Sau khi đã có index của các outliers, tiến hành loại bỏ outliers dựa trên index được lưu trong 2 array upper và lower.



* Sau khi đã xử lý các outliers, sử dụng Boxplot để quan sát bộ dữ liệu. Nếu phát hiện còn tồn tại Outliers thì ta tiến hành xử lý tiếp, ngược lại, kiểm tra bộ dữ liệu sau khi đã xử lý.

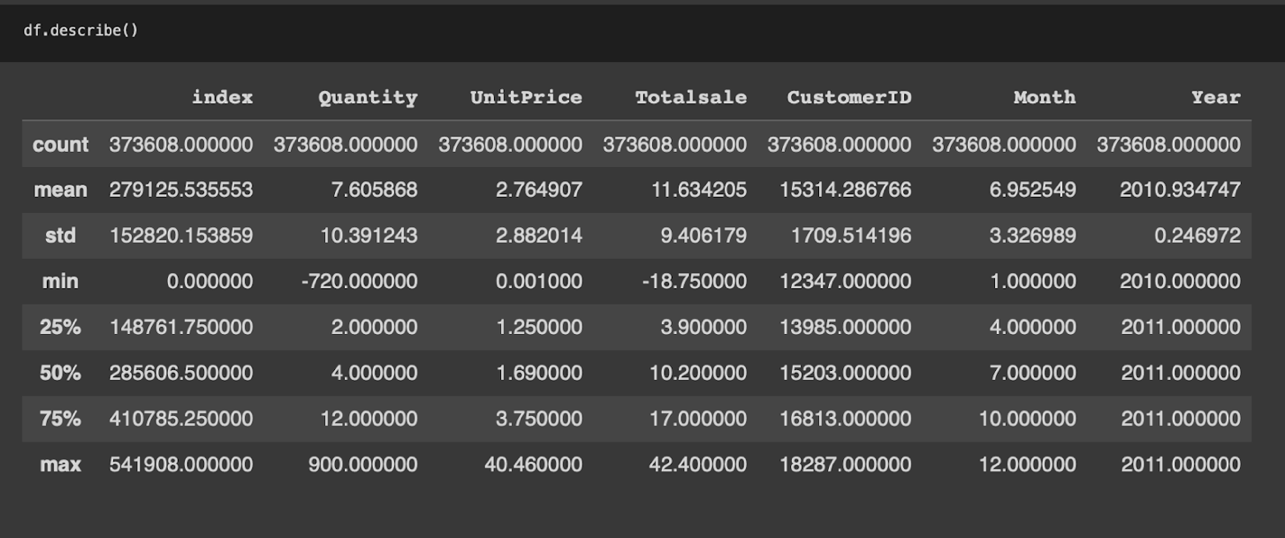


Hình 5: Box plot của biến Totalsale sau khi xử lý Outliers.

* Nhận xét: Sau khi loại bỏ các Outliers , cột Totalsale có sự thay đổi trong các giá trị như sau: Q1 = 3.9, Q2 = 10.2, Q3 = 17.0, Min = -18.75 và Max = 42. Quan sát biểu đồ cũng không còn thấy các giá trị nào nằm ngoài vùng expected data. Vậy ta kết luận rằng các Outliers đã được loại bỏ hoàn toàn. vậy ta sẽ xem tổng quan bộ dữ liệu đã thay đổi ra sao sau khi loại đi Outliers

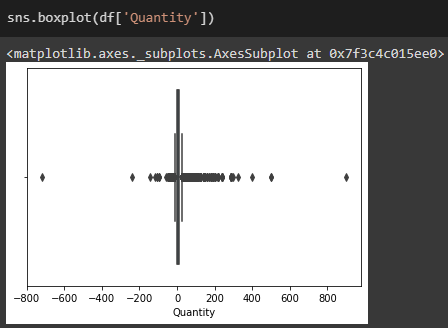
**Đánh giá biểu đồ**

* Biểu đồ tốt, thể hiện, trực quan hóa được thông tin về việc có các giá trị outlier giúp người xem dễ nhận biết, màu sắc dễ nhìn. Các giá trị của các dòng thể hiện giá trị của biểu đồ boxplot nhìn rõ ở văn bản thường.



### b.   Biến Quantity

* Sử dụng biểu đồ hộp để xem các giá trị của Quantity phân phối như thế nào. Ở đây ta thấy chủ yếu dữ liệu dàn trải trong khoảng từ 0-300 và xuất hiện outliers ở 2 đầu mút -800 và 800. Vì vậy ta sẽ tiến hành loại bỏ các giá trị outliers này để tăng độ chính xác cho cho các mô hình phân tích phía sau:

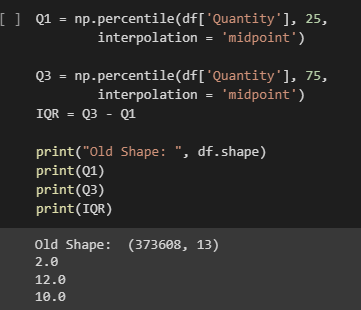


Hình 6: Box plot của biến Quantity trước khi xử lý Outliers.

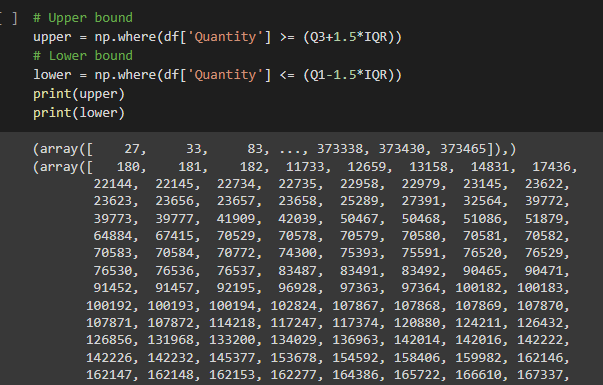
* Tìm Min, Max và các mô tả cơ bản về biến Quantity. Ở đây tồn tại giá trị âm là vì có các sản phẩm bị lỗi, hư hỏng và được hoàn về.



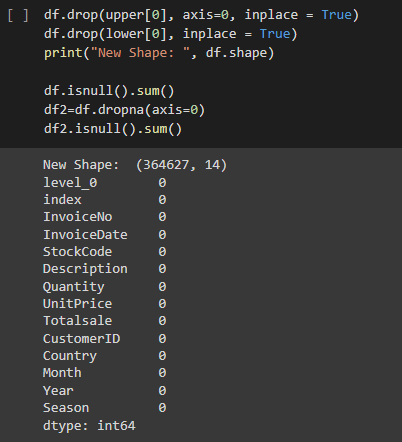
* Xác định giá trị của các tứ phân vị nhằm kiểm soát tác động của những giá trị ngoại lệ (Outliers) ở 2 đầu mút. Ở đây ta chỉ cần xác định giá trị của IQR, tứ phân vị thứ 1 và 3.



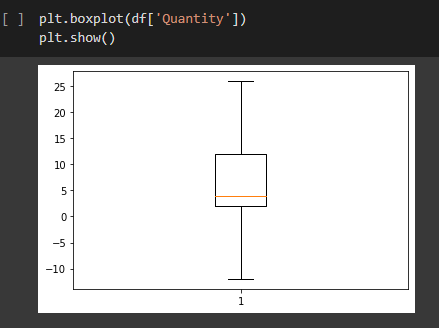
* Sau khi đã có Q1 và Q3, ta tiếp tục xác định giá trị biên của biến Total Sale, nếu giá trị của 1 phần tử bất kì nằm trong khoảng lower và upper thì gọi là expected data, ngược lại gọi là outliers. Tạo 2 mảng chứa các index của phần tử là outliers.



* Sau khi đã có index của các outliers, tiến hành loại bỏ outliers dựa trên index được lưu trong 2 array upper và lower.



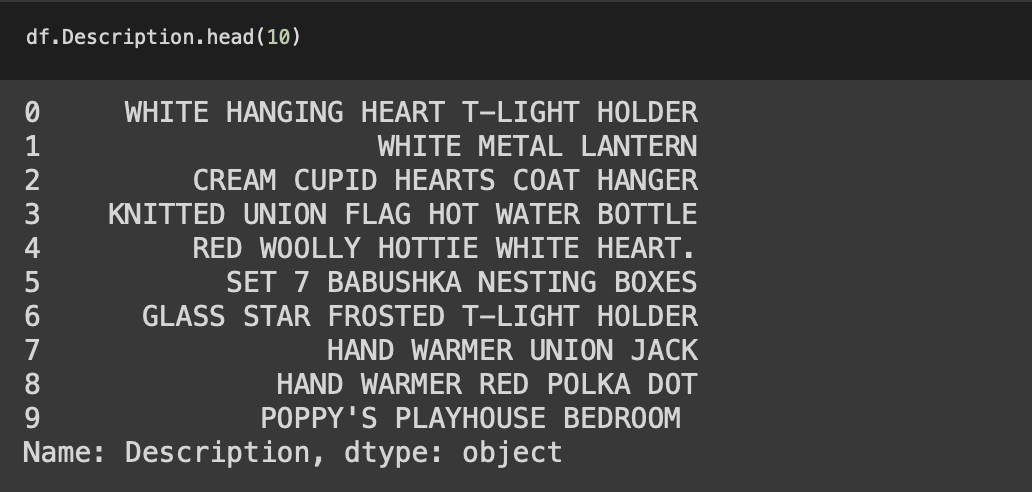
* Sau khi đã xử lý các outliers, sử dụng Boxplot để quan sát bộ dữ liệu. Nếu phát hiện còn tồn tại Outliers thì ta tiến hành xử lý tiếp, ngược lại, kiểm tra bộ dữ liệu sau khi đã xử lý.



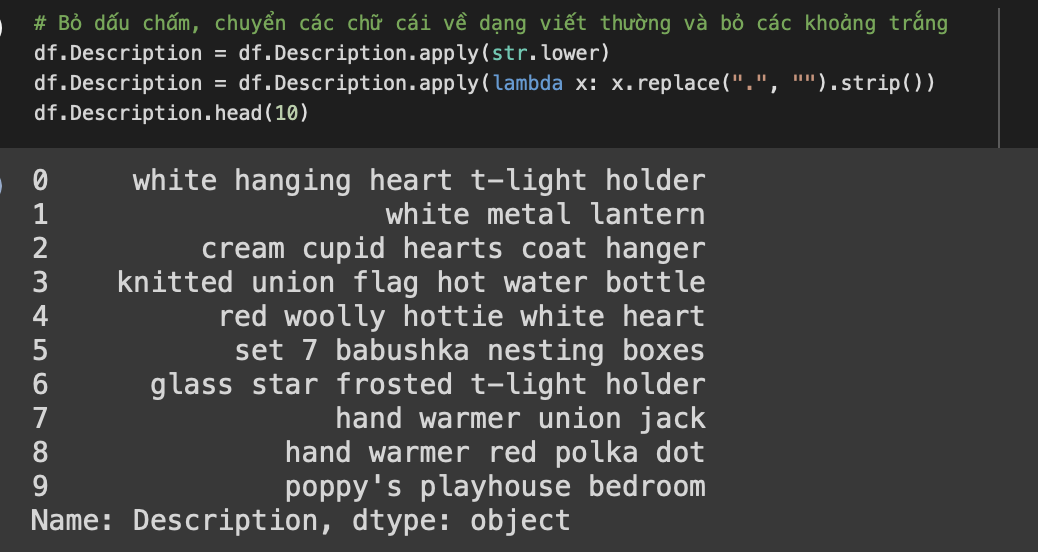
Hình 7: Box plot của biến Quantity sau khi xử lý Outliers.

## 5.   Xử lý biến Description

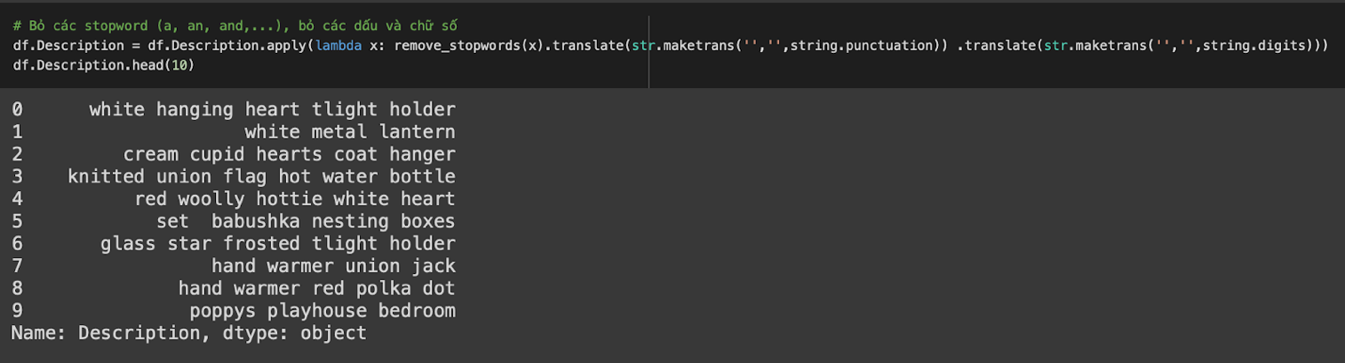
* Biến Description bao gồm các mô tả sơ lược về loại hàng hóa mà người dùng đã mua, nên biến này sẽ có dạng chuỗi ký tự nhưng chưa thống nhất về cấu trúc. Chính vì vậy, ta cần cập nhật lại biến này sao cho ở các dòng, biến Description sẽ có cấu trúc giống nhau để phục vụ cho việc phân tích ở phần sau.
* Đầu tiên, ta cần xem sơ lược biến Description thông qua 10 dòng đầu của bộ dữ liệu:



* Quan sát thấy các kí tự đang ở dạng chữ in hoa nên ta tiến hành chuyển về chữ thường và loại luôn các khoảng trắng (nếu có):



* Bên cạnh đó, một vài chuỗi ký tự vẫn còn bao gồm số, các ký tự (-,’)  và các stopword (a, an, and,..) nên ta tiến hành loại bỏ.



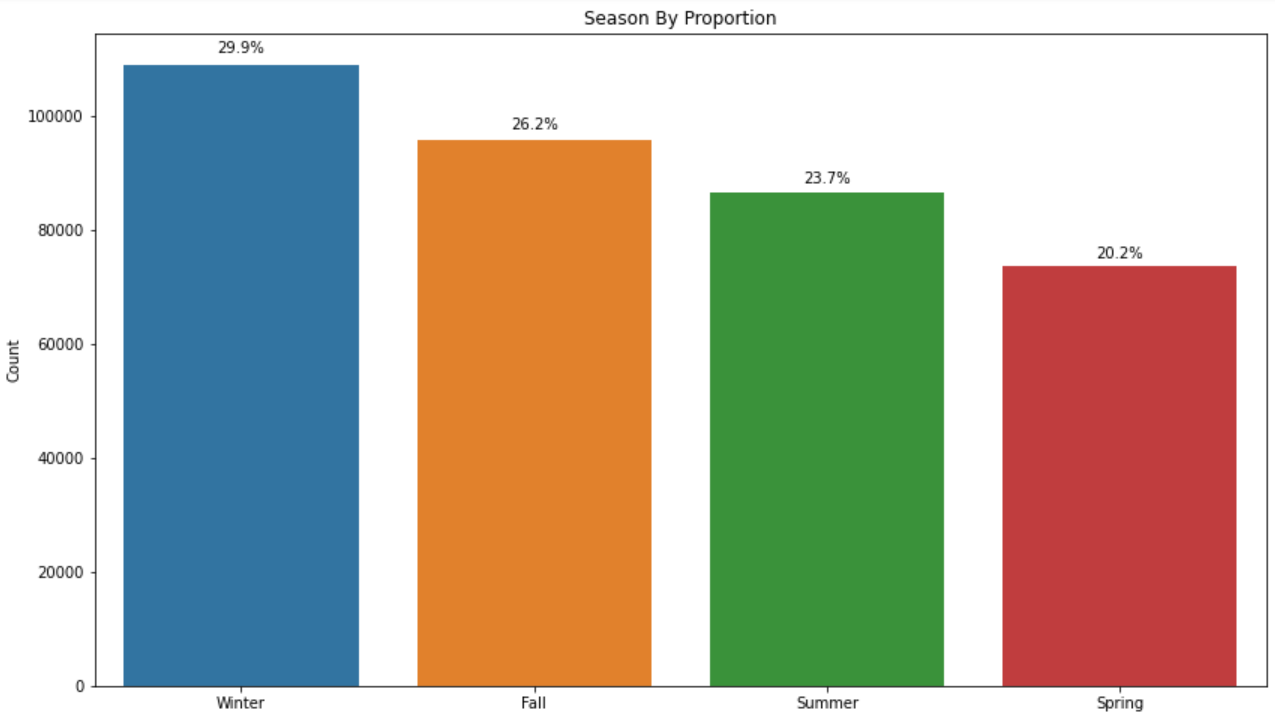
# Chương IV: Phân Tích Dữ Liệu

## Time Series:

Trong xã hội hiện đại, khi lượng thông tin càng ngày càng tăng, việc phân tích dữ liệu time series trở nên cực kỳ quan trọng, nhưng cho dù có nhiều thuật toán hay mô hình trí tuệ nhân tạo sinh ra, việc dự đoán tương lai vẫn thất bại do sự thiếu ổn định trong chuỗi thời gian.

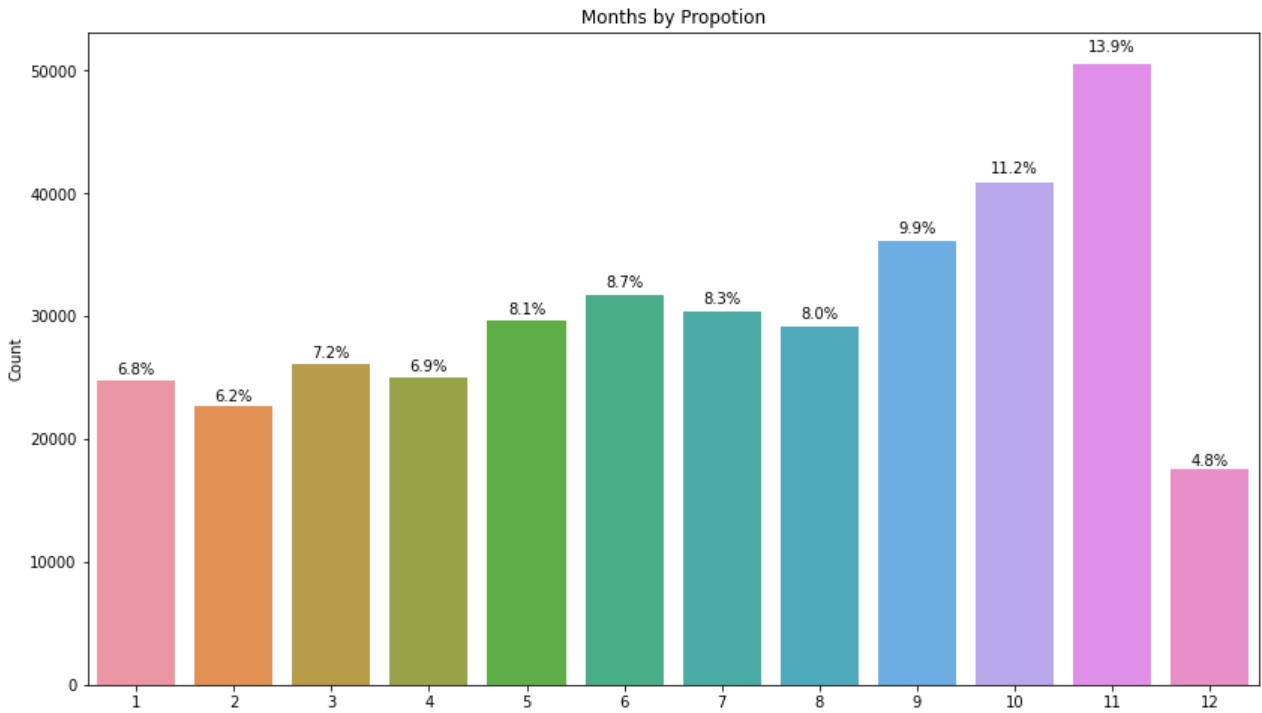
Bằng cách phân tích dữ liệu theo thời gian giúp chúng ta có một cái nhìn tổng quan về tình hình bán hàng của doanh nghiệp, nhận biết tốt hơn về xu hướng của dữ liệu và kiểm tra xem liệu có tính mùa vụ trong bộ dữ liệu không nhằm hỗ trợ đưa ra các quyết định bán hàng trong tương lai.

Sử dụng biểu đồ thanh dọc thể hiện doanh thu của 4 mùa, với 4 màu sắc khác nhau. Doanh nghiệp có xu hướng được đặt hàng tăng dần từ đầu năm đến cuối năm khi lượng đặt hàng cao nhất là vào mùa đông (~110,000 đơn hàng) và lượng đặt thấp nhất là vào mùa xuân (~78,000 đơn hàng).



Hình 8: Biểu đồ thanh thể hiện số lượng đặt hàng theo mùa

Để có một cái nhìn tốt hơn, ta hãy nhìn vào lượng đặt theo tháng. Vẫn tiếp tục sử dụng biểu đồ thanh dọc với 12 màu đại diện cho 12 tháng trong năm, mặc dù xu hướng đặt hàng vẫn tăng dần từ đầu năm đến cuối năm và đạt đỉnh vào tháng 11 (~50,000 đơn hàng), tuy nhiên  sau đó lại giảm mạnh ngay lập tức vào tháng 12 (~18,000 đơn hàng), thấp nhất trong tất cả các tháng. Vậy nếu chỉ sử dụng biểu đồ theo mùa ở trên thì người đọc báo cáo sẽ không thể biết được sự sụt giảm mạnh đơn hàng ở tháng 12.



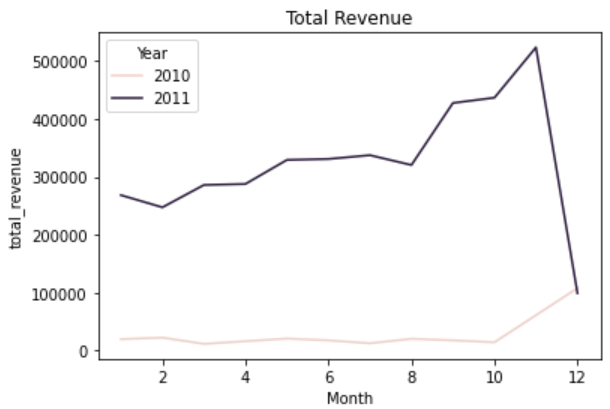
Hình 9: *Biểu đồ thanh thể hiện số lượng đơn hàng theo tháng.*

Bên cạnh đó, ta cũng có thể dễ dàng nhìn thấy lượng đơn hàng được đặt trong năm 2010 không đáng kể so với năm 2011 (6.5% đơn hàng của năm 2010 so với 93.5% đơn hàng của năm 2011 trên tổng gần 500,000 đơn hàng).



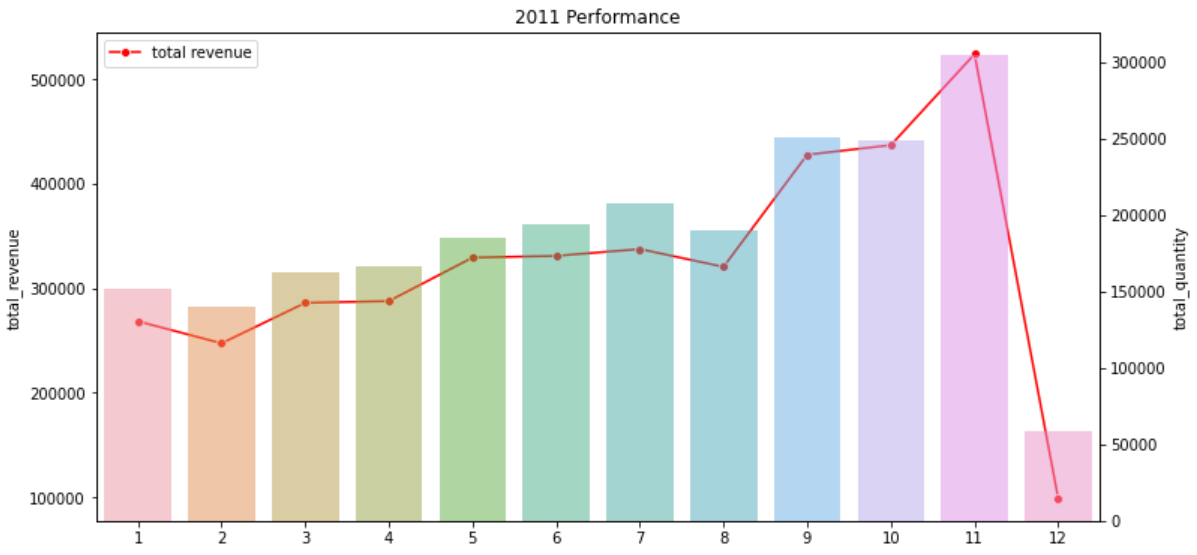
Hình 10: Biểu đồ thanh biểu diễn tỷ lệ đặt hàng của năm 2010 và 2011.

Chính vì lượng đơn hàng vượt trội trong năm 2011 so với năm 2010 đã giúp cho tổng doanh thu theo tháng của năm 2011 đều cao hơn năm 2010, riêng chỉ có khoảng thời gian tháng 12, doanh thu của năm 2011 giảm mạnh, khác với năm 2010 khi vào thời điểm cuối năm là lúc doanh thu nảy bật.



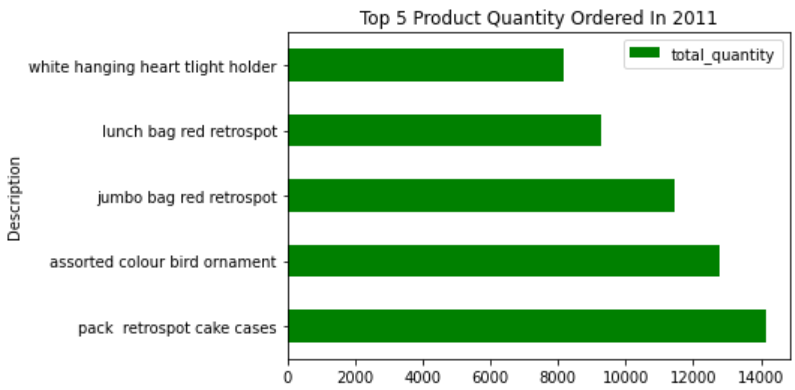
Hình 11: Biểu đồ đường thể hiện tổng doanh thu theo tháng của năm 2010 và 2011.

Ta có thể thấy ảnh hưởng của năm 2011 nhiều hơn đáng kể so với năm 2010, vậy nên ta sẽ cùng nhau có một cái nhìn rõ hơn vào năm 2011. Xu hướng tăng của lượng hàng và doanh thu song hành với nhau, Vậy ta có thể kết luận rằng, giá sản phẩm không tác động nhiều đến doanh thu như số lượng sản phẩm bán ra.



Hình 12: Biểu đồ thanh kết hợp với đường biểu diễn tổng doanh thu và tổng lượng hàng bán ra theo các tháng của năm 2011.

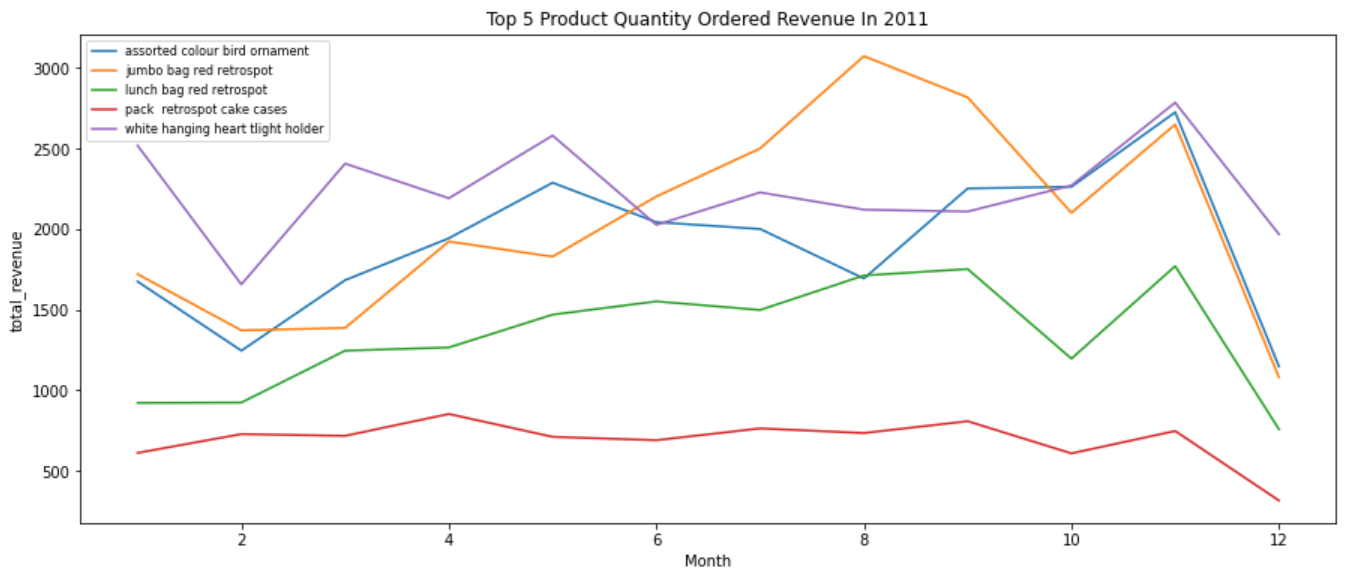
Tiếp đến ta cùng xem 5 sản phẩm có lượng hàng được bán ra cao nhất trong năm 2011. Đứng đầu danh sách là sản phẩm *pack retrospot cake cases* với số lượng ~ 14,000 sản phẩm được bán ra trong năm 2011.



Hình 13: Biểu đồ thanh ngang biểu diễn số lượng sản phẩm bán ra của 5 mặt hàng bán chạy nhất năm 2011.

Qua biểu đồ thanh ngang trên cho biết rằng :  tuy *pack retrospot cake cases* là sản phẩm có số lượng bán hàng cao nhất nhưng doanh thu mang lại cho doanh nghiệp không đáng kể so với 4 sản phẩm còn lại khi doanh thu trung bình mỗi tháng chỉ rơi vào khoảng $600 - $700, chiếm một phần rất nhỏ khi doanh thu trung bình tháng của doanh nghiệp rơi vào khoảng $300,000 - $450,000.

Để so sánh tổng doanh thu của 5 mặt hàng bán chạy nhất năm 2011, nhóm sử dụng biểu đồ đường với 5 màu sắc khác nhau. Tuy nhiên chỉ so sánh các tháng 2,4,6,8,10 và 12 để biểu đồ dễ quan sát hơn. Có thể thấy được “jumbo bag red retrospot” có doanh thu đạt đỉnh vào tháng 8/2011 và là mặt hàng đạt được doanh thu cao nhất trong tất cả 5 mặt hàng tại 1 thời điểm.



Hình 14: Biểu đồ đường biểu diễn doanh thu bán hàng của 5 mặt hàng bán chạy nhất năm 2011 theo các tháng chẵn.

**Đánh giá**: Các barchart, linechart được thể hiện rất tốt trong phần này, các màu, các số liệu, tên biểu đồ, tên cột dễ nhìn

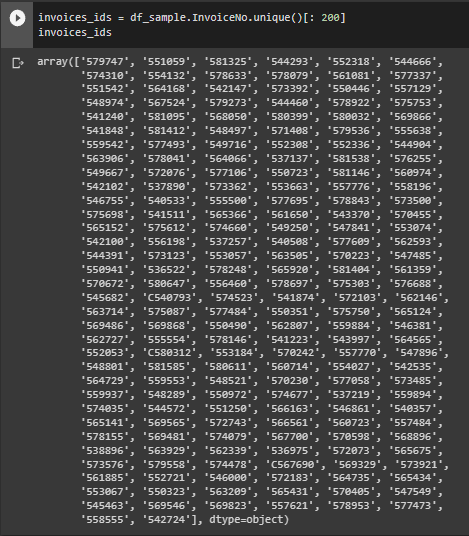
## FP-Growth:

Mục tiêu: tìm ra các mặt hàng thường được bán chạy với nhau và mùa được bán chạy nhất để kết hợp sản xuất, quảng cáo giúp tăng lượng hàng được bán ra từ đó giúp tăng đáng kể doanh thu. Sử dụng FP- Growth để tìm ra các mặt hàng thường được bán chung với nhau (trong cùng một hoá đơn) và khoảng thời gian mà các mặt hàng được bán chạy nhất giúp tối ưu hoá lượng hàng được bán ra.

Bước 1: Lấy mẫu ngẫu nhiên 350000 dòng (lấy random\_state = 8 nhằm giữ nguyên kết quả cho mục đích nghiên cứu)

****

Bước 2: Lấy 200 dữ liệu InvoiceNo đầu nhằm giảm sự phức tạp của thuật toán FP-Growth. Lưu vào array có tên là invoice\_ids.

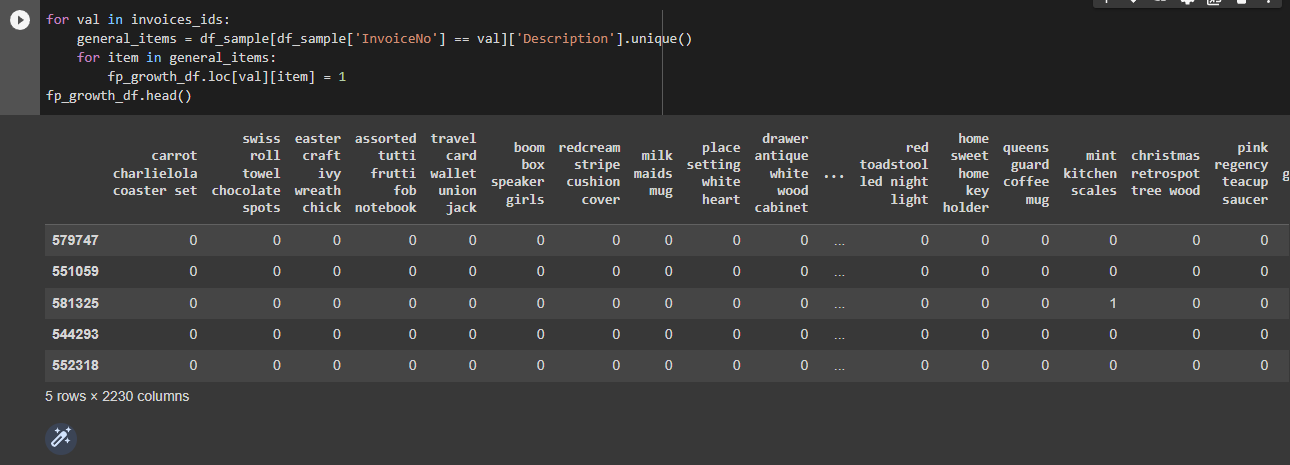
****

Bước 3: Tạo ra dataframe với dòng là các dữ liệu “InvoiceNo” và các cột là tên các sản phẩm (“Description”). Dựa trên array invoice\_ids, ta lấy các “Description” rồi đẩy dữ liệu vào một array khác có tên là all\_items. Các cột “Description” được tạo ra dựa vào việc ghép các “Description” đã được đưa array all\_items (ghép thành 1 set dữ liệu). Đưa hết toàn bộ dữ liệu của dataframe về 0.

Đây là bước chuẩn bị dataframe để tiếp đến đánh dấu các “Description” có trong “InvoiceNo”.

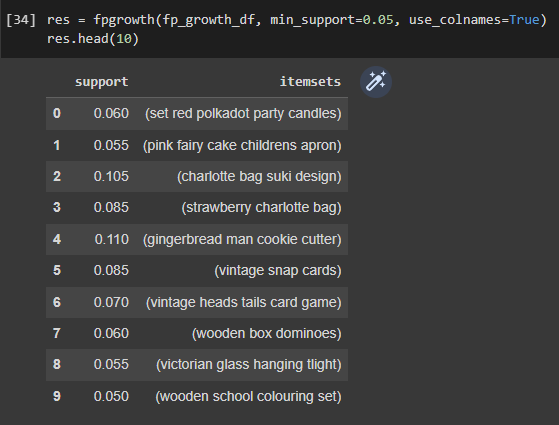
****

**Bước 4:** Đánh dấu các “Description” nếu nó có tồn tại trong “InvoiceNo” tương ứng.Đánh dấu bằng giá trị “1” đối với các “Description” có tồn tại trong “InvoiceNo” được xét.

****

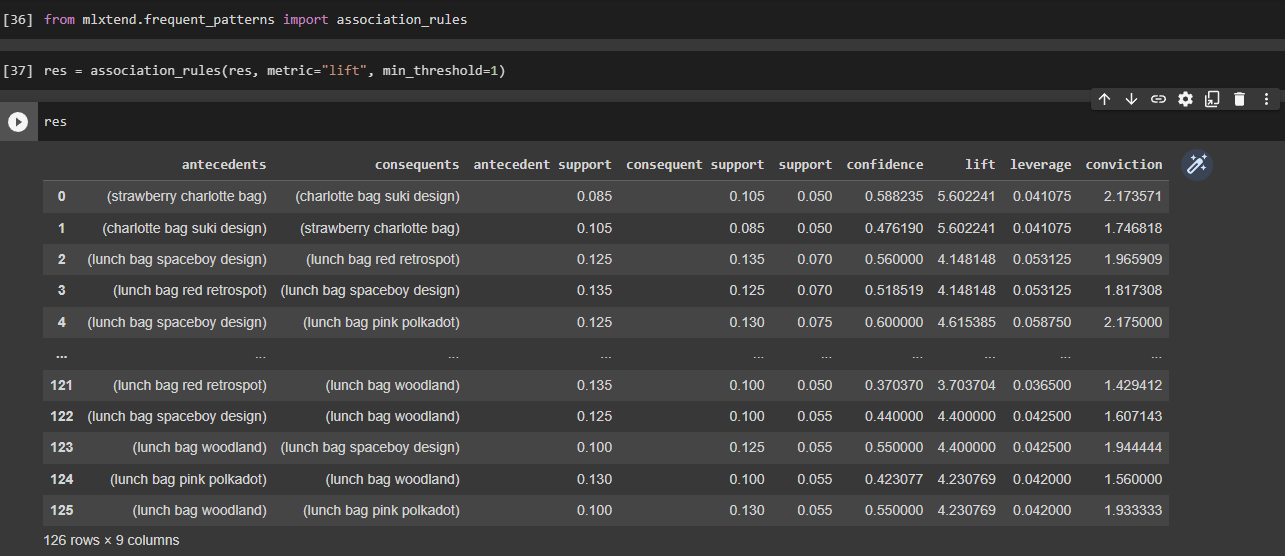
Bước 5:Sử dụng hàm fpgrowth để tính toán độ support của các items, lấy min\_support là 0.05.

**mlxtend.frequent\_patterns.fpgrowth(df, min\_support = 0.5, use\_colnames = False, max\_len = None, verbose = 0)**

****

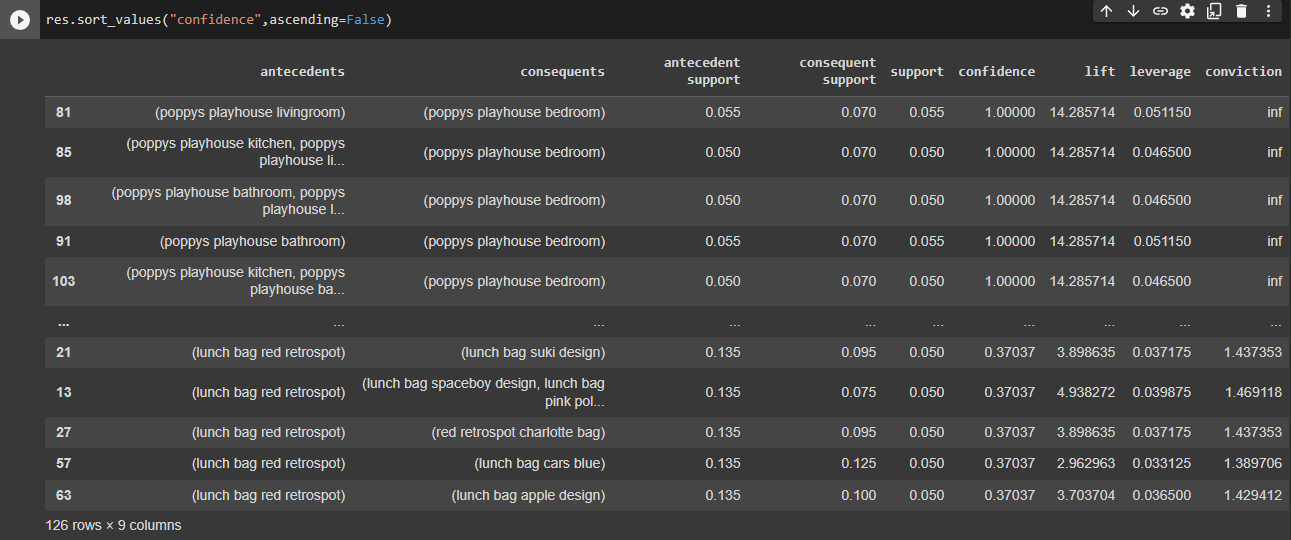
Bước 6:Lọc ra các luật kết hợp có độ đo lift 1

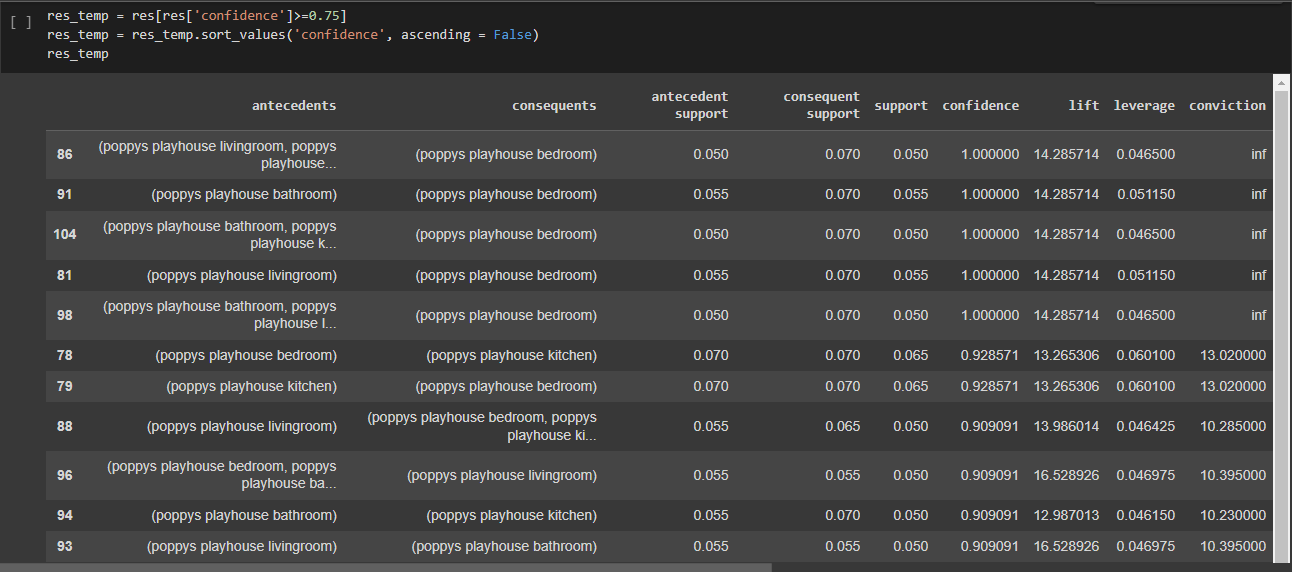
**mlxtend.frequent\_patterns.association\_rules(df, metric = “confidence”, min\_threshold = 0.8, support\_only = False)**

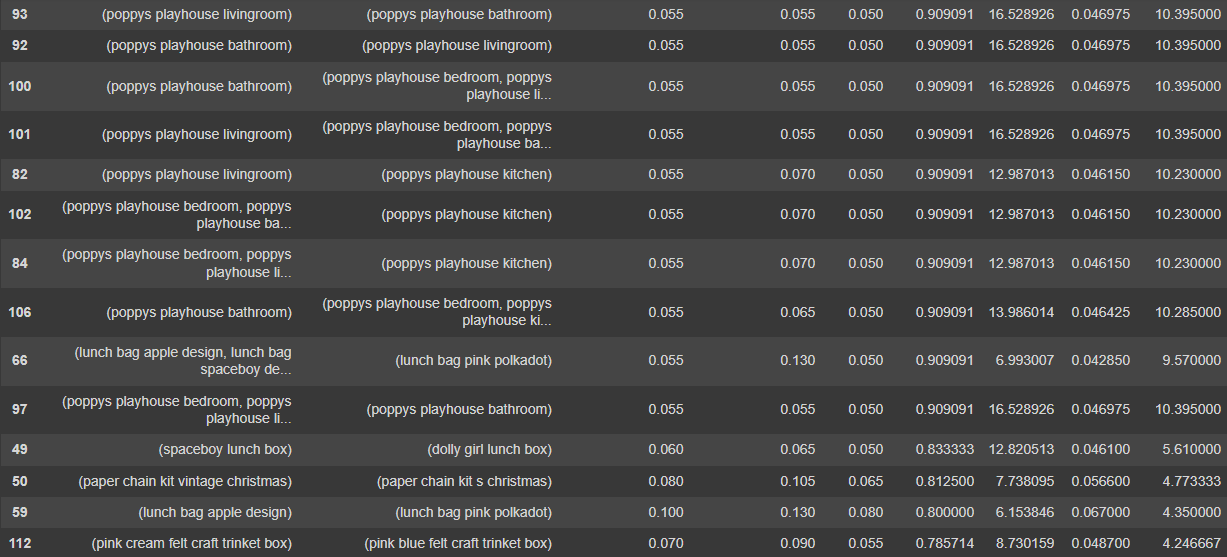
****

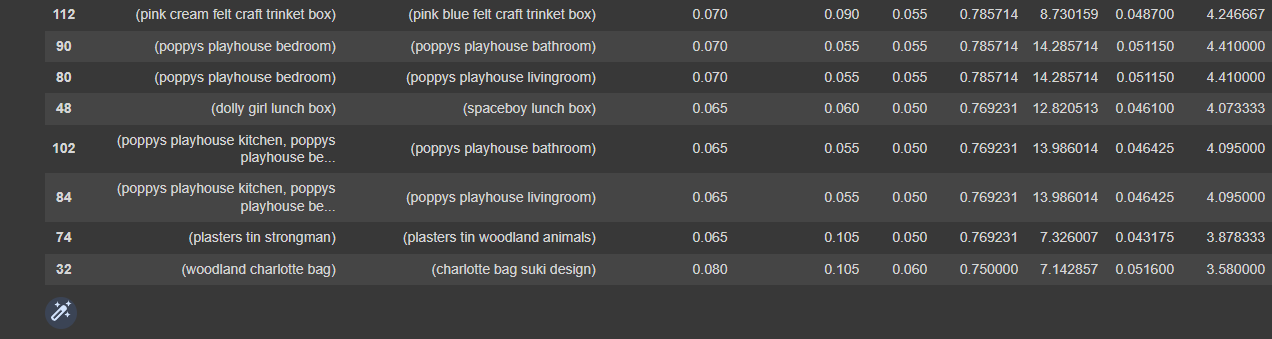
**Bước 7:** Sắp xếp các luật kết hợp giảm dần theo độ “confidence” và lọc ra các luật kết hợp có “confidence” 0.75.

*Nhận xét*: Ta quan sát thấy được rằng, đa phần các sản phẩm thường đi chung với nhau là “poppys playhouse livingroom”, “poppys playhouse bedroom”, “poppys playhouse bathroom”, “poppys playhouse kitchen”. Đặc biệt là 5 luật kết hợp đầu tiên có độ “confidence” và “conviction” tuyệt đối (confidence = 1, conviction = inf). Chính vì vậy, ta cần tập trung chú ý đến các mặt hàng này.

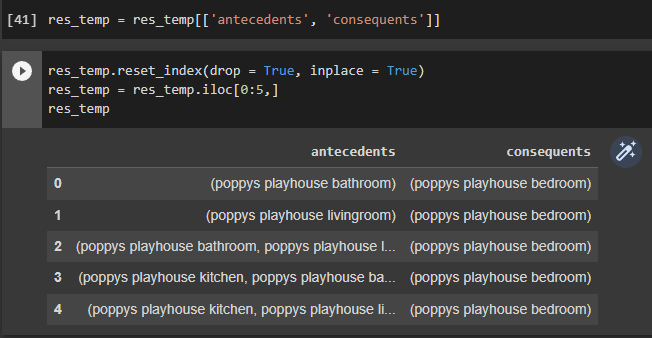
****

****

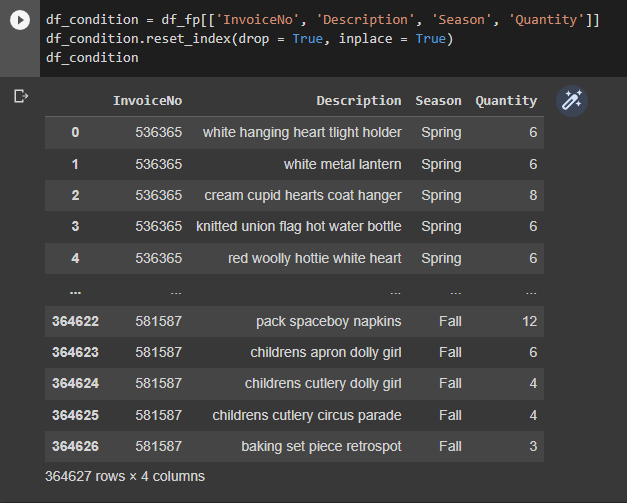
****

****

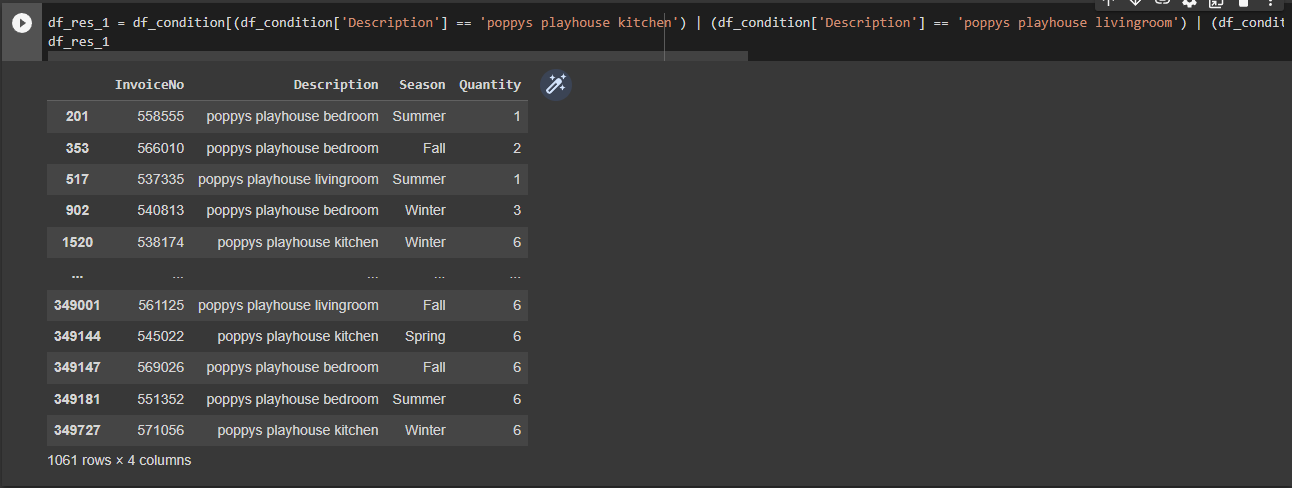
Bước 8:Lọc ra 5 luật kết hợp đầu tiên (5 luật kết hợp với confidence = 1 và conviction = inf)

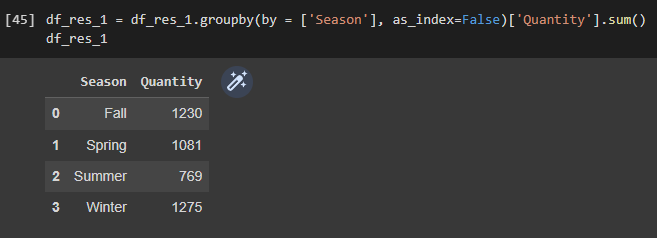
****

Bước 9**:** Lấy bộ dữ liệu gốc chưa lấy mẫu ngẫu nhiên với các cột “InvoiceNo”, “Description”, “Season”, “Quantity” để phục vụ cho mục đích so sánh đối chiếu lượng hàng được bán ra của các mặt hàng được chú ý.

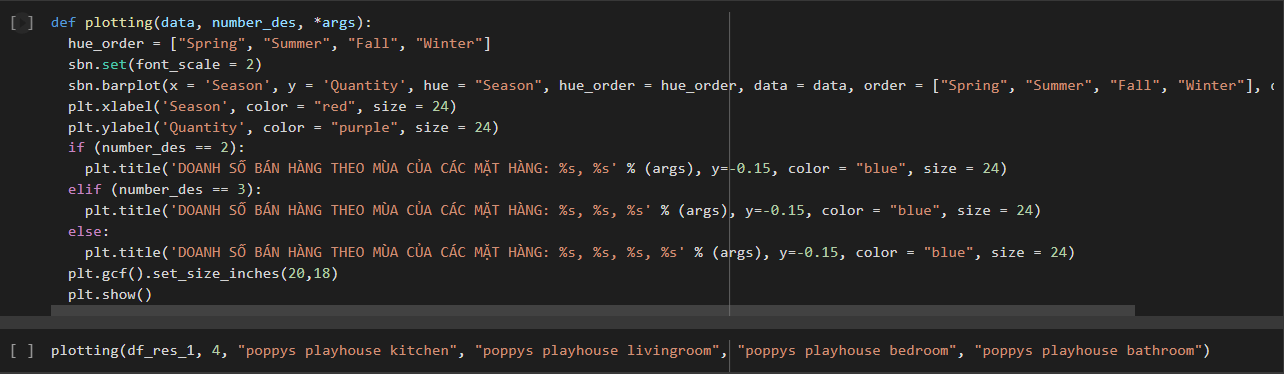
****

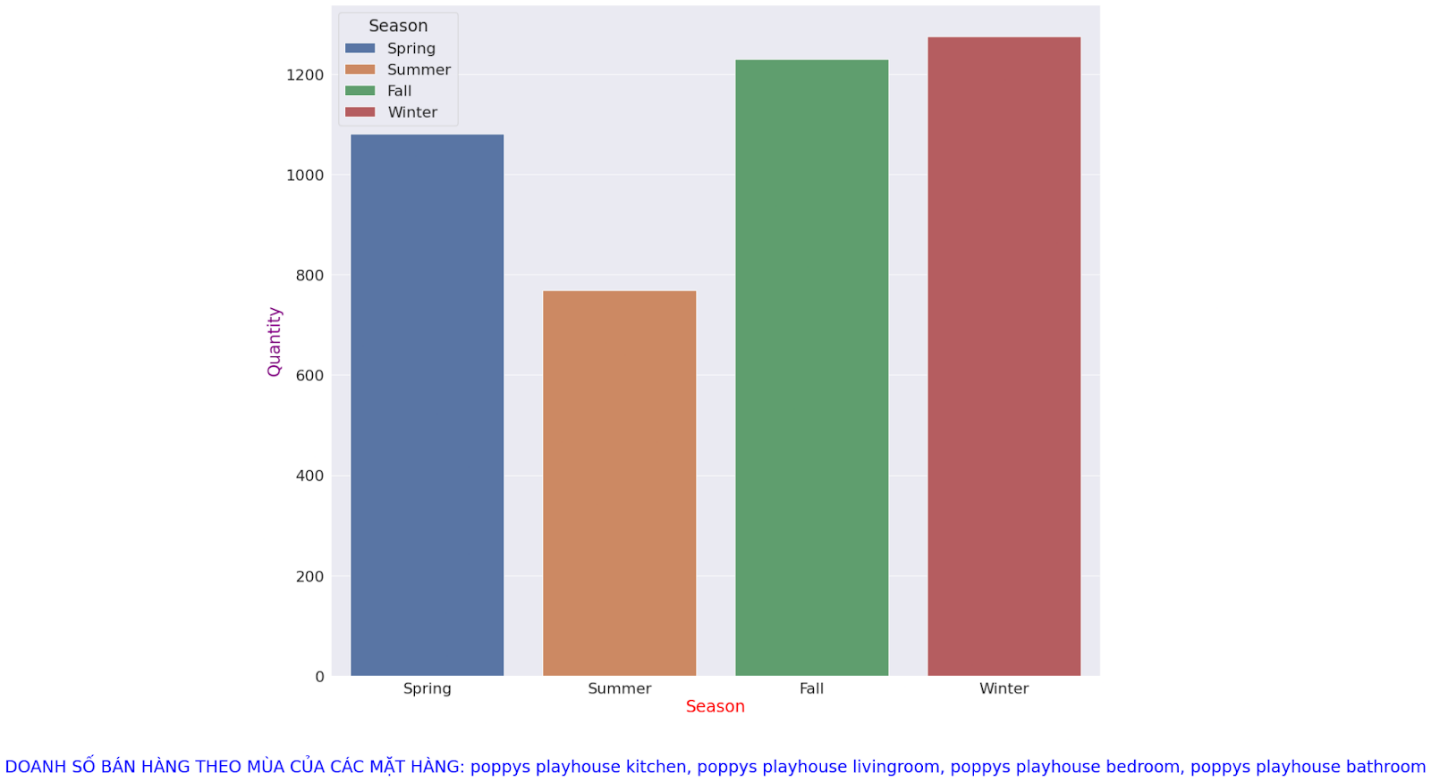
Bước 10**:** Thống kê số lượng mặt hàng “poppys playhouse kitchen”, “poppys playhouse livingroom”, “poppys playhouse bedroom” và “poppys playhouse bathroom” theo từng mùa. Ta không cần lọc theo InvoiceNo vì nếu các item được bán nhiều và đi chung với nhau ở mùa nào thì mùa đó sẽ có số lượng hàng bán ra nhiều hơn với các mặt hàng này nên việc đó là không cần thiết.

****

****

**Bước 11:** Vẽ biểu đồ số lượng hàng “**poppys playhouse kitchen”,** “**poppys playhouse livingroom**”, “**poppys playhouse bedroom**” và “**poppys playhouse bathroom**” đã được bán theo mùa.

****

****

Hình 15: Biểu đồ thanh dọc thể hiện doanh số bán hàng theo mùa của 4 mặt hàng.

**Nhận xét**: dùng biểu đồ cột (Bar plot), ta thấy được các mặt hàng “poppys playhouse kitchen”, “poppys playhouse bedroom”, “poppys playhouse livingroom” và “poppys playhouse bathroom” được bán nhiều nhất ở 2 mùa Thu (Fall) và Đông (Winter). Chính vì vậy, ta sẽ cần có các chính sách giảm giá, khuyến mãi bán theo combo các mặt hàng này để thúc đẩy lượng hàng được bán ra; gia tăng sản xuất các mặt hàng này, gia tăng quảng cáo để kích sức bán lên cao nhất có thể.

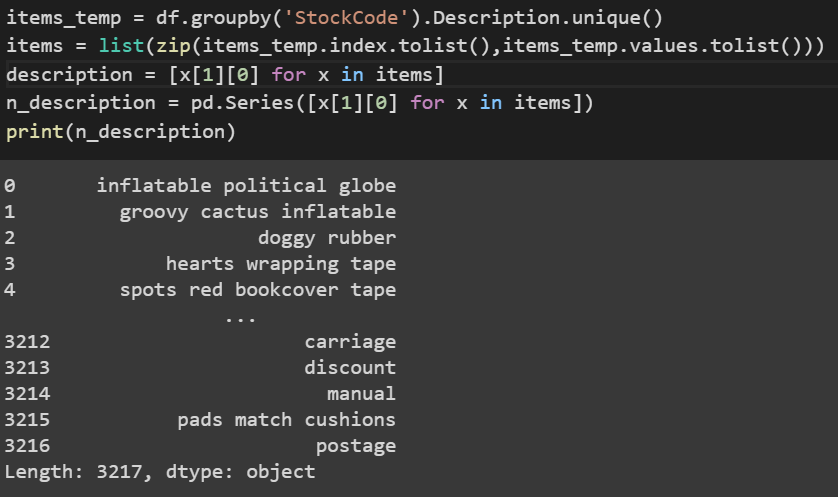
**Đánh giá**: Tiêu đề của biểu đồ quá dài gây mất cân đối về mặt thẩm mỹ. Màu sắc dễ nhìn, thông tin được truyền tải đầy đủ cho người xem.

# Chương V: Machine Learning

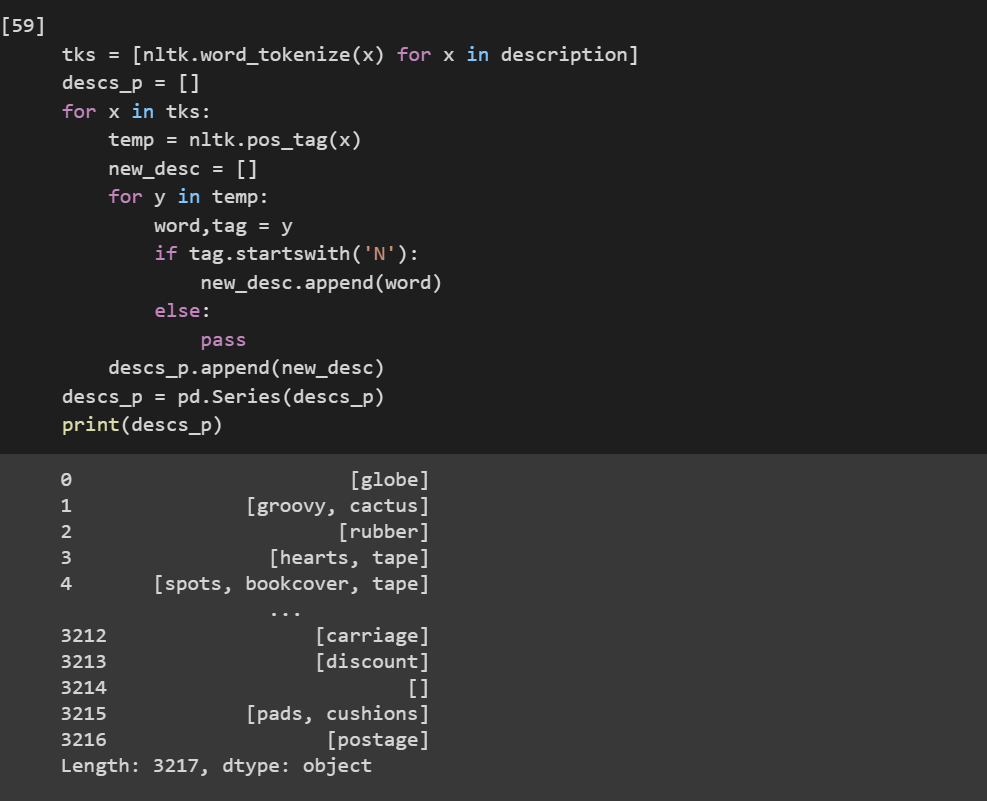
## Cluster Category by Product Name (Description) - K-means:

* 1. **Mục đích:**

Bộ dữ liệu có 3,217 mã sản phẩm (stock code) khác nhau tuy nhiên nhiều sản phẩm lại có tên (Description) tương tự nhau và có thể gộp chung lại thành một nhóm sản phẩm. Việc gom những sản phẩm tương tự nhau lại thành một nhóm giúp chúng ta có được cái nhìn tổng quát hơn về một mặt hàng được bán chạy thay vì từng sản phẩm riêng biệt. Bên cạnh đó, trực quan hóa dữ liệu cho 3,217 đầu sản phẩm cũng không hiệu quả nếu so với những nhóm sản phẩm chung hơn.



* 1. **Xử lý dữ liệu Description:**
* Dù đã được xử lý tốt ở bước tiền xử lý, chúng ta vẫn cần phải căn chỉnh lại dữ liệu **Description** cho phù hợp với với việc phân cụm.
* Đầu tiên, chúng ta cần phải tokenize từng vùng dữ liệu trong **Description** (tên của từng sản phẩm) để dễ dàng loại bỏ những từ sẽ gây nhiễu đến việc phân cụm. Tích hợp trong bước tokenize, chúng ta cũng sẽ loại bỏ những động từ, tính từ, trạng từ, …, chỉ giữ lại danh từ, phù hợp cho việc phân cụm.

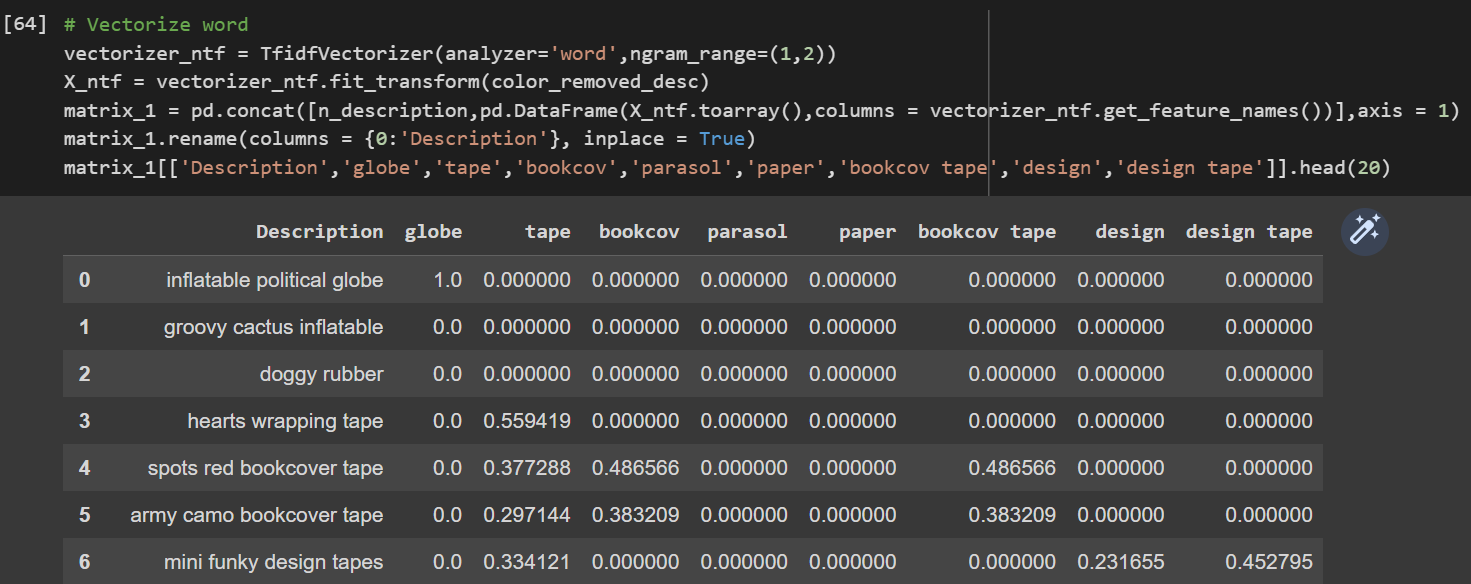


* Tiếp theo, Ta đưa các từ trên về dạng từ cán (ví dụ programming, programmer đưa về dạng từ cán là program) đồng thời loại bỏ các danh từ chỉ màu sắc vì có khả năng sẽ tồn tại và gây nhiễu trong quá trình chạy phân cụm.



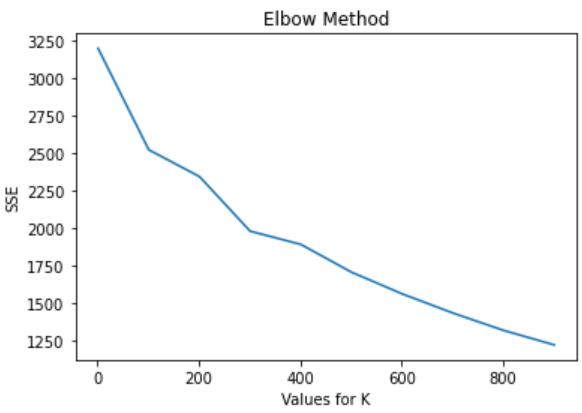
* 1. **Vector hóa**

Sau khi làm cho những dòng dữ liệu trở nên sạch sẽ hơn, ta tiến hành chuyển các dữ liệu dạng chuỗi của ta thành dữ liệu dạng số. Cụ thể hơn, trong bài này ta sẽ sử dụng phương pháp **TF - IDF (term frequency–inverse document frequency)** để gắn trọng số cho mỗi từ trong 1 tên sản phẩm. **TF - IDF** là một phương pháp đánh giá độ quan trọng của một từ hoặc một cụm từ trong tập văn bản khi văn bản đó nằm trong một tập hợp các văn bản.Phương pháp này được xác định dựa trên tần suất của từ trong văn bản và tần suất của văn bản chứa từ đó trong tập hợp các văn bản. Như ví dụ bên dưới ta có thể thấy, áp dụng phương pháp **Tf - IDF** để gắn trọng số cho mỗi từ trong tên sản phẩm, với sản phẩm có tên *“spots red bookcover tape”*  thì từ bookcover có trọng số ~ 0.486 và từ tape có trọng số ~ 0.377. Với giá trị trọng số càng cao, từ đó sẽ mang ý nghĩa cao hơn trong văn bản. Bên cạnh đó, trong bài này cũng có sử dụng n-gram level để thứ tự của các từ cũng được xem xét.



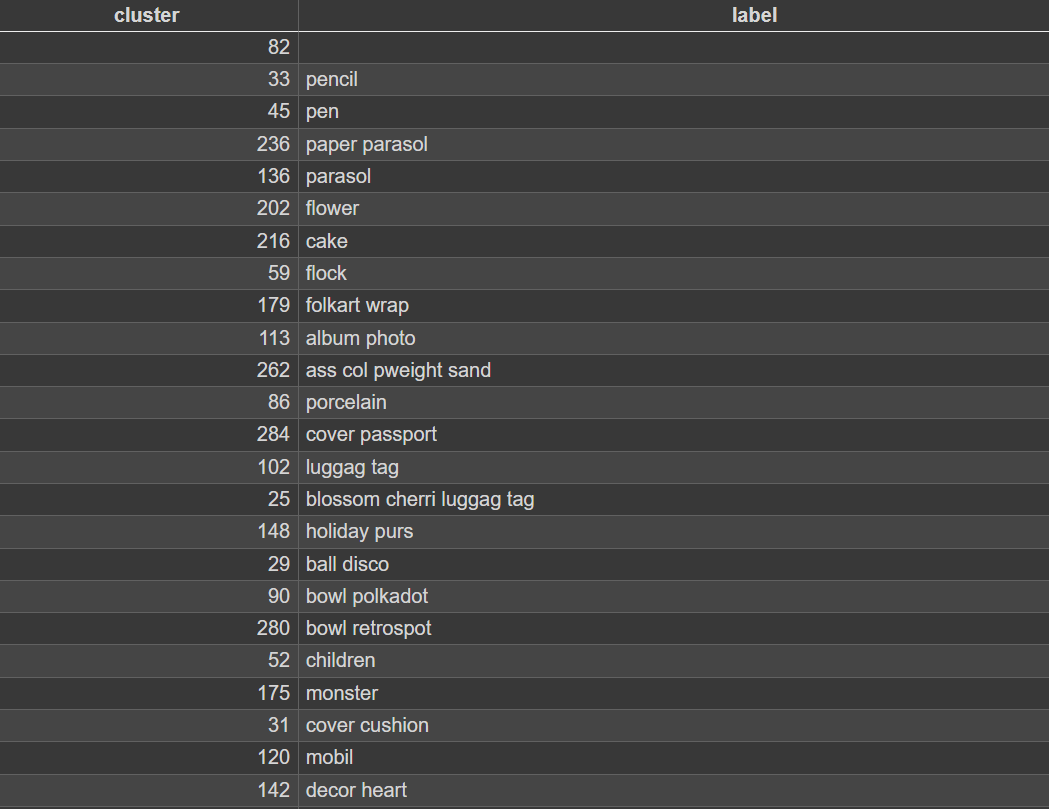
* 1. **Phân cụm với K-means:**

Sau khi đã có ma trận trọng số các từ, ta tiến hành phân cụm với k-means. Đầu tiên sử dụng phương pháp elbow để tìm được số cụm tối ưu nhất. Phương pháp elbow sử dụng độ tương quan nghịch giữa SSE (Sum of Square Error) và số cụm. Ở đây, ta sử dụng biểu đồ đường để có thể dễ dàng nhìn thấy sự tương quan này, Và theo biểu đồ, 300 là số cụm tối ưu.

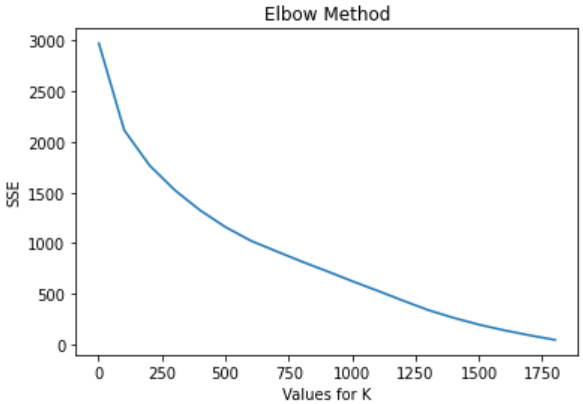


Hình 16

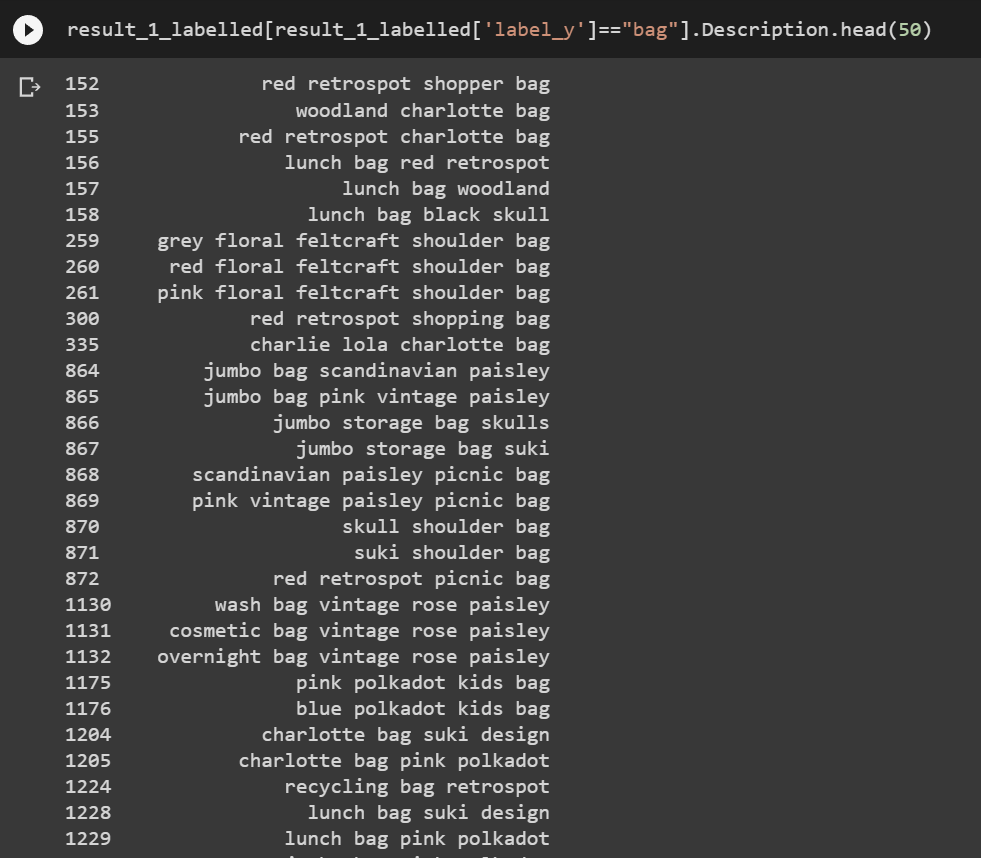
Sau khi được phân thành 300 cụm, ta tiến hành dán nhãn cho mỗi cụm để xem liệu các cụm được đặt tên hợp lý với các sản phẩm trong cụm hay không Sau khi kiểm tra các cụm được dán nhãn, ta dễ dàng nhìn thấy các nhãn không hợp lý như *Children* hay cùng một loại bowl thì được phân thành 2 cụm *bowl polkadot* và *bowl retrospot.* Bước này, ta cần thủ công loại bỏ những từ không phù để các cụm được nhóm lại hiệu quả hơn.



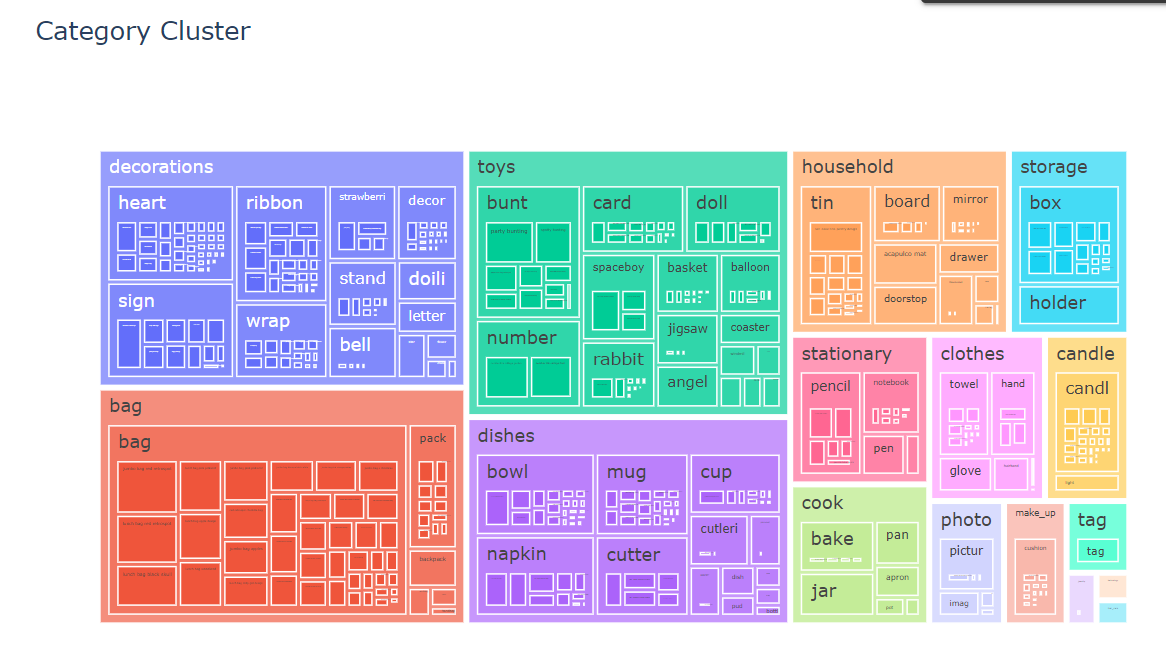
Sau khi loại bỏ các từ bất hợp lý, ta tiến hành vectorize lại phân cụm lại với bộ dữ liệu mới và thấy rằng điểm SSE chung đã giảm. Tại K = 300, điểm SSE ở bộ dữ liệu cũ rơi vào khoảng ~ 2,000 còn ở bộ dữ liệu đã được lọc từ không hợp lí chỉ còn khoảng ~ 1500.



* Ta kiểm tra xem các sản phẩm được đưa vào trong 1 cụm có hợp lý hay không.Với cụm được dán nhãn lag *bag*, thì các sản phẩm đều là *bag*, một tín hiệu đáng mừng.

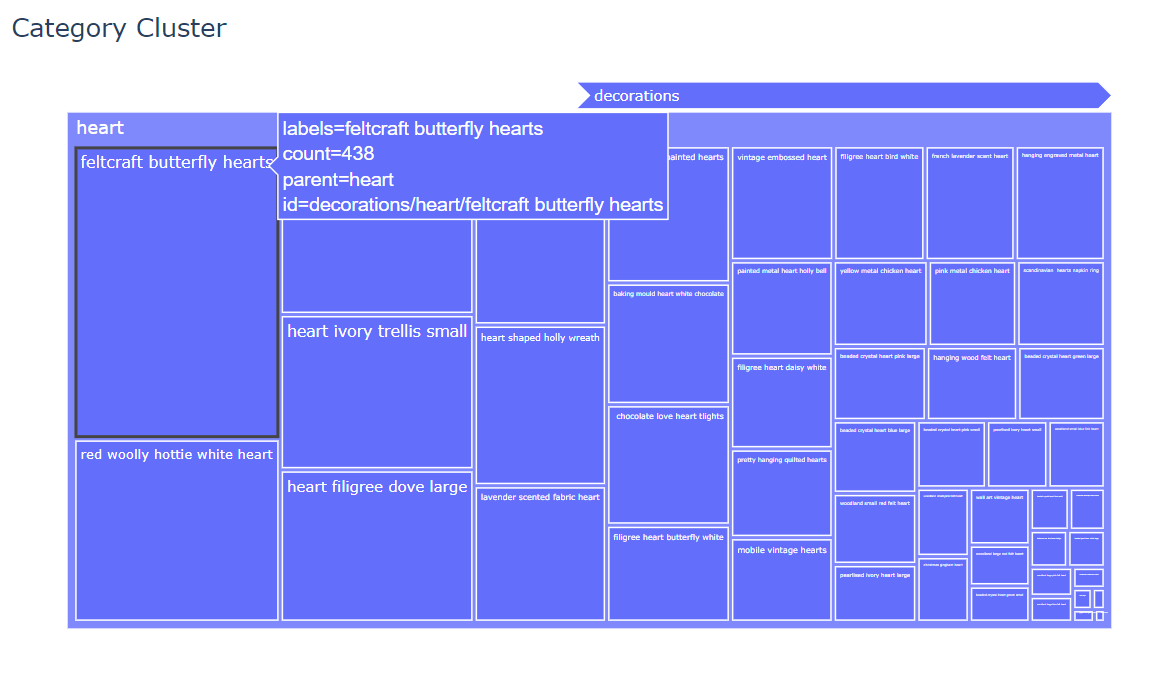


* Thông thường, để trực quan hóa kết quả của việc phân cụm, ta thường sử dụng biểu đồ phân tán. Nhưng trong trường hợp này, chúng ta có đến 300 cụm, tương ứng với 300 categories khác nhau, vì vậy, dùng biểu đồ phân tán sẽ không phù hợp, thay vào đó, biểu đồ cây là loại biểu đồ thể hiện rất tốt ở công đoạn trực quan hóa này.

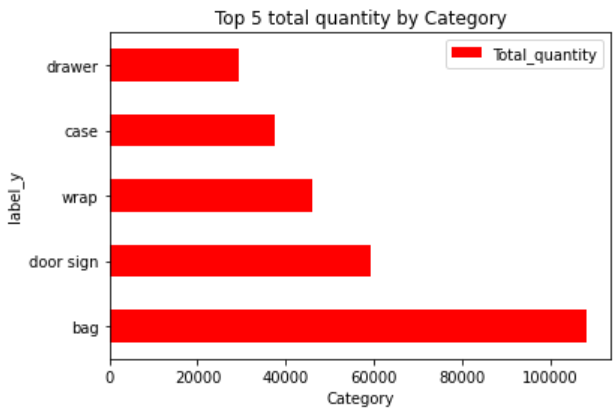


Hình 17: Biểu đồ cây thể hiện các nhóm sản phẩm.

* Ta cũng có thể nhấp vào từng lớp để xem các sản phẩm nằm trong đó:



* Cuối cùng, sau khi phân những sản phẩm thành những nhóm, ta sử dụng những nhóm này cho mục đích phân tích dữ liệu bằng cách trực quan hóa.Ta có thể dễ dàng nhìn thấy *bag* đang là loại sản phẩm được bán nhiều nhất với số lượng đặt hàng ~ 45,000 (Một insight khác khi ta sử dụng category để phân tích vì sản phẩm bán nhiều nhất lại là *pack retrospot cake case* như đã đề cập ở trên không thuộc nhóm bag)



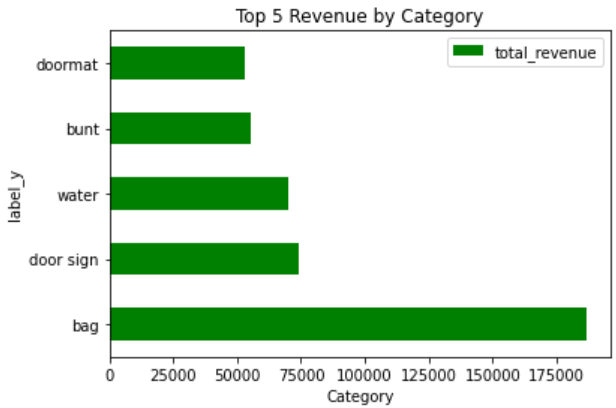
Hình 18: Biểu đồ thanh dọc biểu diễn tổng số lượng hàng bán ra của 5 loại hàng bán chạy nhất.

* Bên cạnh đó ta cũng có thể biết được loạt mặt hàng nào được đặt hàng nhiều nhất. Ta có thể thấy sự khác nhau ở top 3,4,5 giữa số lượng đặt và số lần đặt theo từng nhóm.



Hình 19: Biểu đồ thanh dọc biểu diễn tổng số đơn đặt hàng cho 5 loại hàng bán chạy nhất.

* Và cuối cùng ta cũng có thể thấy rằng *bag* là dòng sản phẩm bán chạy nhất và cũng là dòng sản phẩm mang lại doanh thu cao nhất cho doanh nghiệp với revenue ~ 80,000



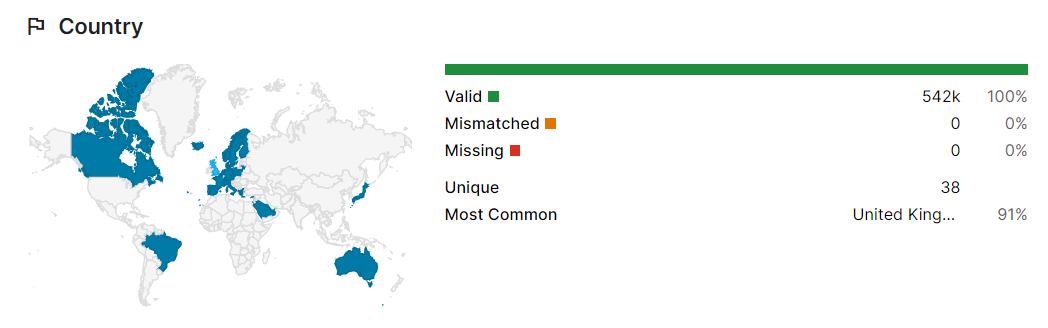
Hình 20: Biểu đồ thanh dọc biểu diễn tổng doanh thu của 5 loại hàng bán chạy nhất mang lại.

**Đánh giá biểu đồ**:

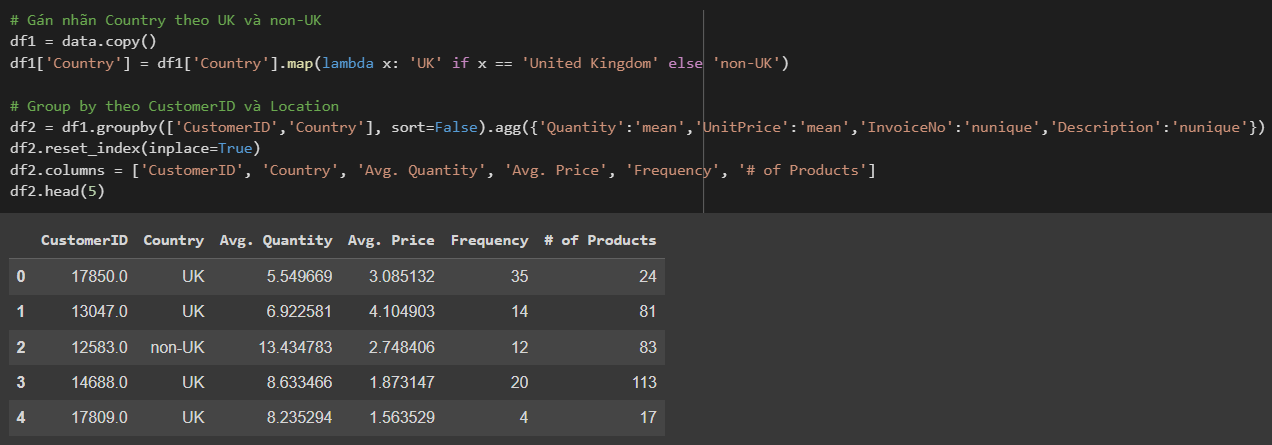
* 3 biểu đồ barchart đã dùng ở phần này thể hiện tốt các thông tin cần truyền tải đến người xem, màu sắc dễ nhìn.
* Đặc biệt là treemap có thể nhấp vào được từng ô dữ liệu để xem kĩ các phần dữ liệu được biểu diễn

## Customer Segmentation by Country and Behavior - K-Prototypes

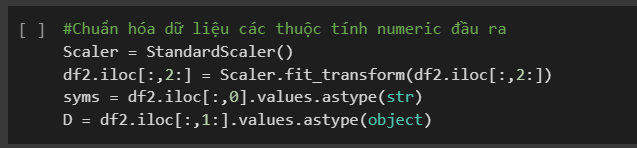
Tập dữ liệu có phần lớn khách hàng ở UK (91%) và phần nhỏ thuộc các quốc gia khác (9% còn lại). Vì vậy ta sẽ tiến hành phân cụm khách hàng khách hàng theo vị trí địa lý và hành vi mua hàng, xem xét những gì khách hàng mua, mức giá, tần suất và số lượng họ mua hàng để từ đó có thể đưa ra các chiến dịch tiếp thị phù hợp với từng nhóm khách hàng.



* Đầu tiên ta sẽ gán nhãn thuộc tính country thành UK (United Kingdom) và non-UK (các quốc gia khác):



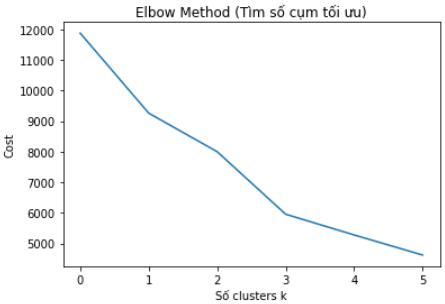
* Ta tiến hành chuẩn hóa dữ liệu cho các thuộc tính numeric bằng cách scale dữ liệu về một phân bố, trong đó giá trị trung bình của các quan sát bằng 0 và độ lệch chuẩn bằng 1, giúp giảm thiểu dữ liệu dư thừa, cải thiện quá trình phân cụm.



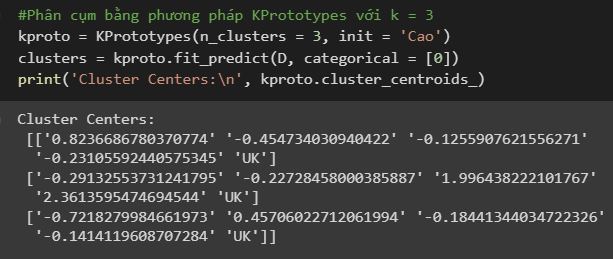
* Sử dụng phương pháp K-Prototypes - thuật toán cải tiến của K-Means để làm việc với các tập dữ liệu hỗn hợp (thuộc tính số và thuộc tính phân lớp)
* Ta tiến hành tìm số phân cụm tối ưu k bằng cách chạy vòng lặp với số cụm k từ 2 đến 8 để tìm ra số k phân cụm tối ưu bằng phương pháp Elbow.



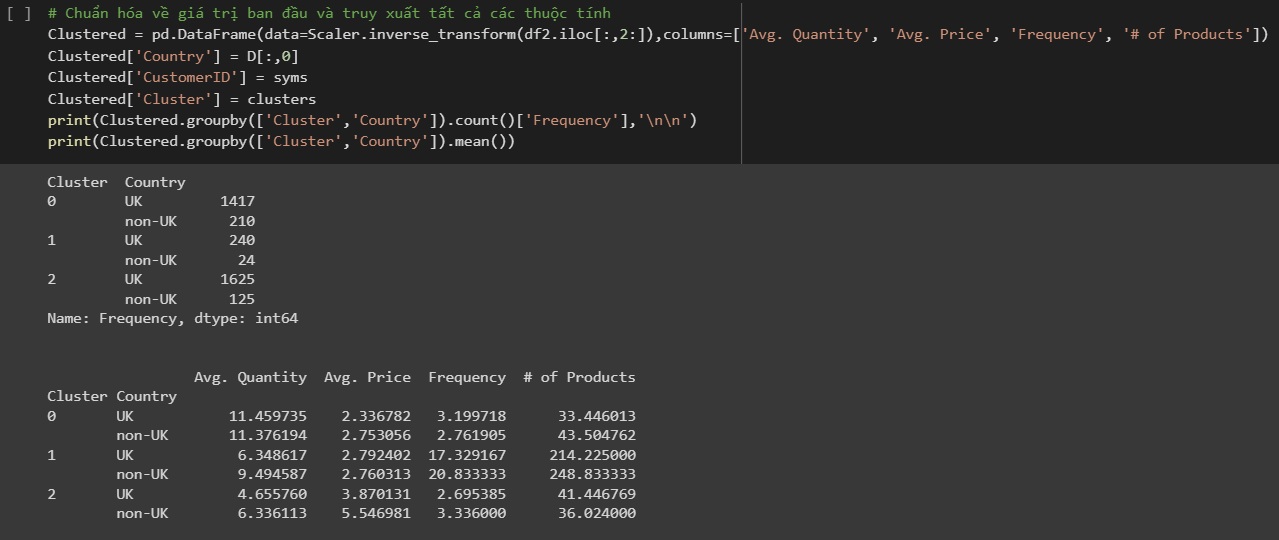




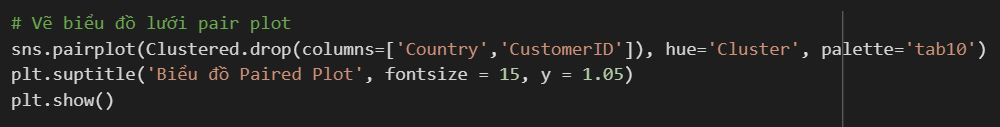
* Ta tiến hành phân cụm bằng phương pháp KPrototypes với k = 3

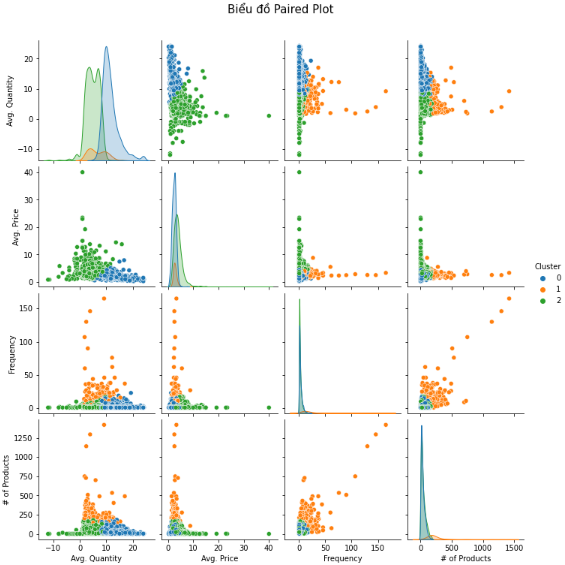


* Sau đó chuẩn hóa dữ liệu về giá trị ban đầu và truy xuất các thuộc tính



* Vẽ biểu đồ lưới pairplot thể hiện mối tương quan giữa các chiều dữ liệu (số lượng khách hàng, đơn giá, số lần mua hàng và mô tả sản phẩm theo vị trí địa lý) theo từng cặp với nhau





Hình 21:Biểu đồ lưới Pair plot phân loại cụm khách hàng.

**Nhận xét:** Ta có thể thấy cách các cụm được phân phối theo từng biến trong biểu đồ phân tán.

* Trong cụm đầu tiên, khách hàng đã mua sản phẩm với số lượng lớn và mức giá thấp, các sản phẩm không đa dạng, tần suất khách hàng trở lại cũng khá thấp. Khách hàng chủ yếu nằm ngoài UK
* Cụm thứ hai đại diện cho những khách hàng mua sản phẩm với số lượng nhỏ, mức giá thấp nhưng đa dạng nhiều loại sản phẩm và họ có xu hướng thường xuyên quay trở lại mua hàng. Phần lớn khách hàng (cả ở UK và không ở UK) đều thuộc cụm này.
* Cụm thứ ba bao gồm những khách hàng đã mua những sản phẩm với số lượng thấp, mức giá cao và thỉnh thoảng mua hàng. Vì vậy những khách hàng mua hàng 1 lần có nhiều khả năng thuộc phân khúc này. Đồng thời phần lớn khách hàng (cả ở UK và không ở UK) đều thuộc cụm này

**Đánh giá biểu đồ**:

* Biểu đồ có màu sắc dễ nhìn, tuy nhiên chưa thực sự trực quan. Mặc dù là thể hiện được một phần thông tin phân cụm qua scatterplot nhưng nhìn chung vẫn khó truyền tải được đầy đủ ý của người phân tích biểu đồ. Một phần do có quá nhiều thuộc tính trong quá trình phân cụm tạo nên vấn đề này. Giải pháp có khả năng cải thiện được vấn đề này là việc giảm chiều dữ liệu, giúp trực quan hóa biểu đồ, thông tin truyền tải.

# CHƯƠNG VI: KIỂM ĐỊNH VÀ GIẢM CHIỀU DỮ LIỆU

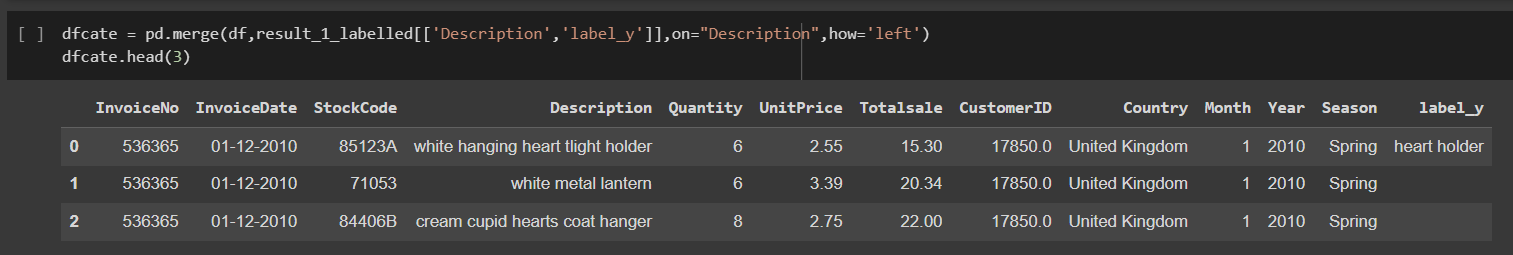
## Kiểm Định

### Chi-square test

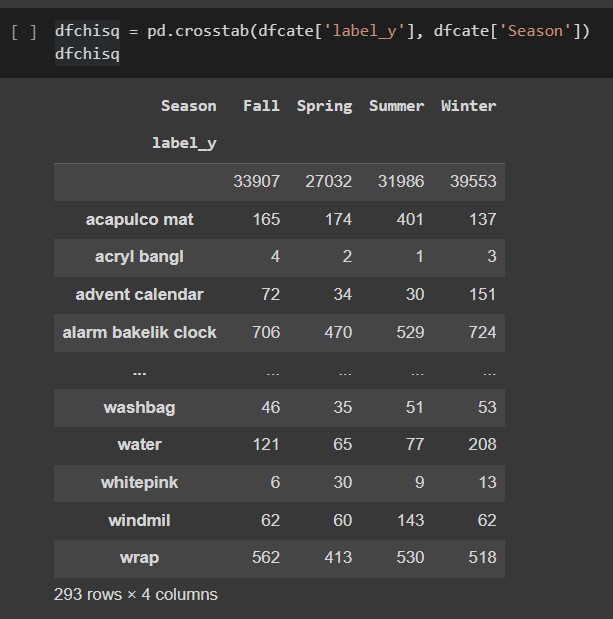
Kiểm định chi bình phương (hay kiểm định Chi-square) là phép thử về tính độc lập được dùng để xác định xem liệu có mối liên hệ giữa các biến phân loại hay không.

Ở đây bộ dữ liệu có các biến phân loại là Season (mùa), Category (loại sản phẩm có được từ clustering Description), và Country (quốc gia).

1. **Kiểm định tính độc lập giữa biến Season và Category**

**Bước 1:** Đầu tiên ta sẽ merge cột label\_y có được từ phần phân cụm vào dataframe ****

**Bước 2:** Tạo bảng chéo crosstab giữa biến label\_y và Season

****

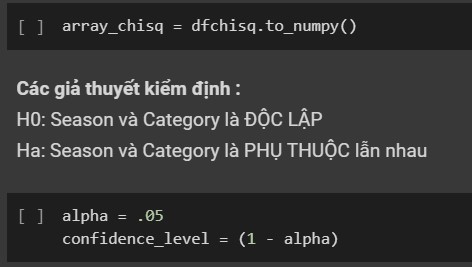
**Bước 3:** Chuyển đổi bảng crosstab thành array.

Xác định giả thuyết kiểm định:

Ho: Season và Category là ĐỘC LẬP

Ha: Season và Category là PHỤ THUỘC lẫn nhau.

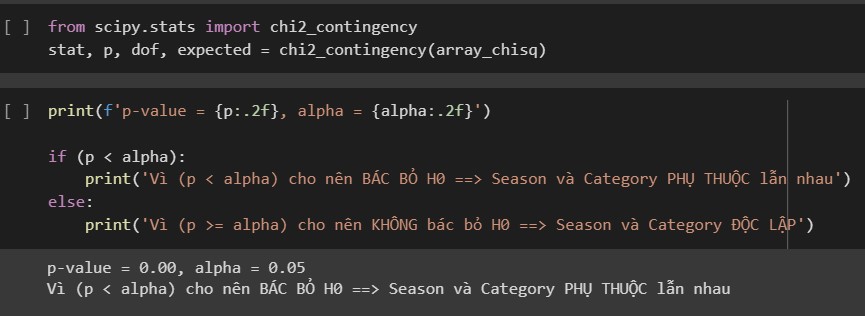
Với alpha = 0.05 và độ tin cậy = 95%

****

**Bước 4:** Tiến hành kiểm định Chi-square và kết luận theo phương pháp p-value (trị số p).

Với điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

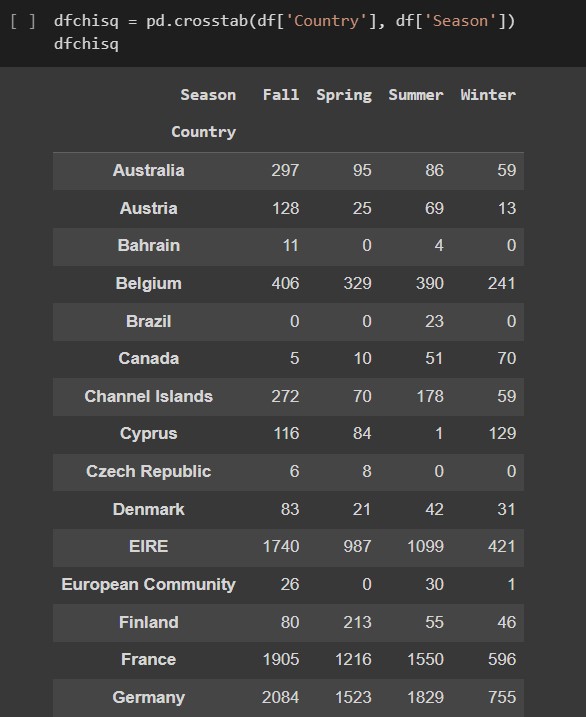
Ở đây p-value xấp xỉ bằng 0 và nhỏ hơn alpha = 0.05. Vì vậy ta sẽ bác bỏ Ho và đưa ra được kết luận Season và Category phụ thuộc lẫn nhau

****

**b) Kiểm định tính độc lập giữa biến Season và Country**

Với yêu cầu kiểm định tính độc lập giữa biến Season và Country ta cũng sẽ thực hiện tương tự các bước như trên.

**Bước 1:** Tạo bảng chéo crosstab giữa biến Country và Season



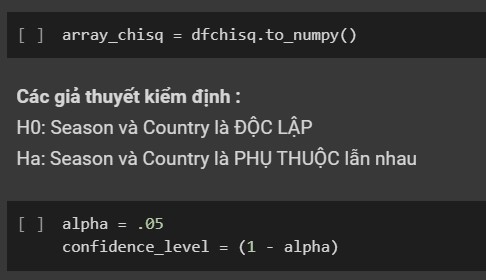
**Bước 2:** Chuyển đổi bảng crosstab thành array.

Xác định giả thuyết kiểm định:

Ho: Season và Country là ĐỘC LẬP

Ha: Season và Country là PHỤ THUỘC lẫn nhau.

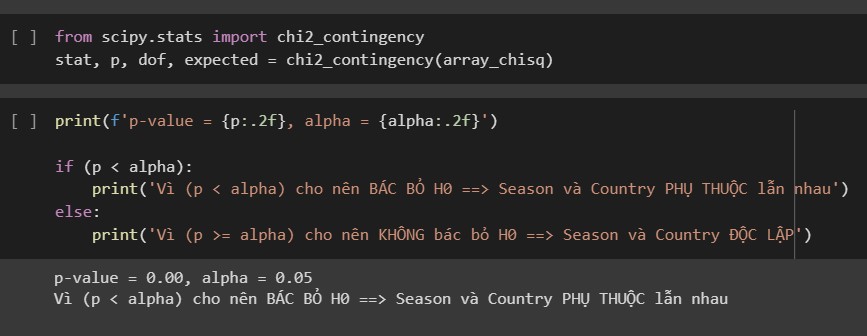
Với alpha = 0.05 và độ tin cậy = 95%



**Bước 3:** Tiến hành kiểm định Chi-square và kết luận theo phương pháp p-value (trị số p).

Với điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

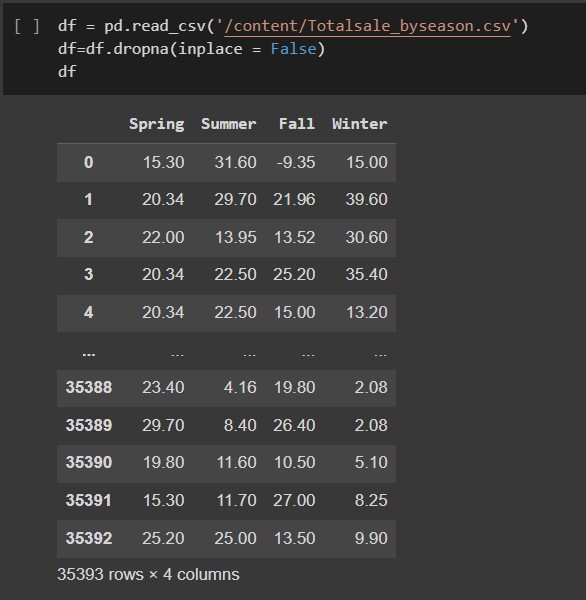
Ở đây p-value xấp xỉ bằng 0 và nhỏ hơn alpha = 0.05. Vì vậy ta sẽ bác bỏ Ho và đưa ra được kết luận Season và Country phụ thuộc lẫn nhau



### Kiểm định Levene

Phép kiểm định Levene giúp kiểm định sự đồng nhất phương sai của các nhóm giá trị biến định tính trước khi tiến hành kiểm định One-way ANOVA.

**Bước 1:** Tiến hành load bộ dữ liệu Totalsale\_byseason.csv và loại bỏ những cột chứa missing value.

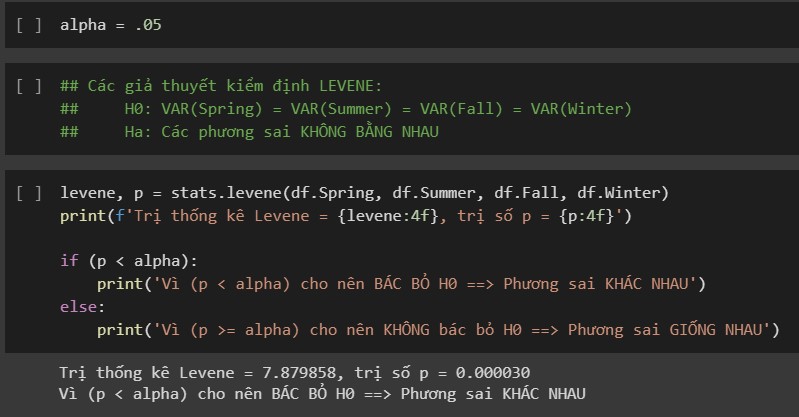


**Bước 2:** Xác định giả thuyết kiểm định:

H0: Var(Spring) = Var(Summer) = Var(Fall) = Var(Winter)

Ha: Các phương sai không bằng nhau

Với alpha = 0.05 và độ tin cậy 95%



**Bước 3:** Tiến hành kiểm định Levene và kết luận theo phương pháp p-value (trị số p).

Với điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

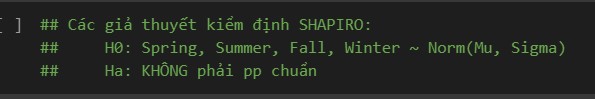
Ở đây p-value xấp xỉ bằng 0 và nhỏ hơn alpha = 0.05. Vì vậy ta sẽ bác bỏ Ho và đưa ra được kết luận các phương sai của Totalsale theo mùa Spring, Summer, Fall, Winter là khác nhau.

### Kiểm định Shapiro

**Bước 1:** Xác định giả thuyết kiểm định:

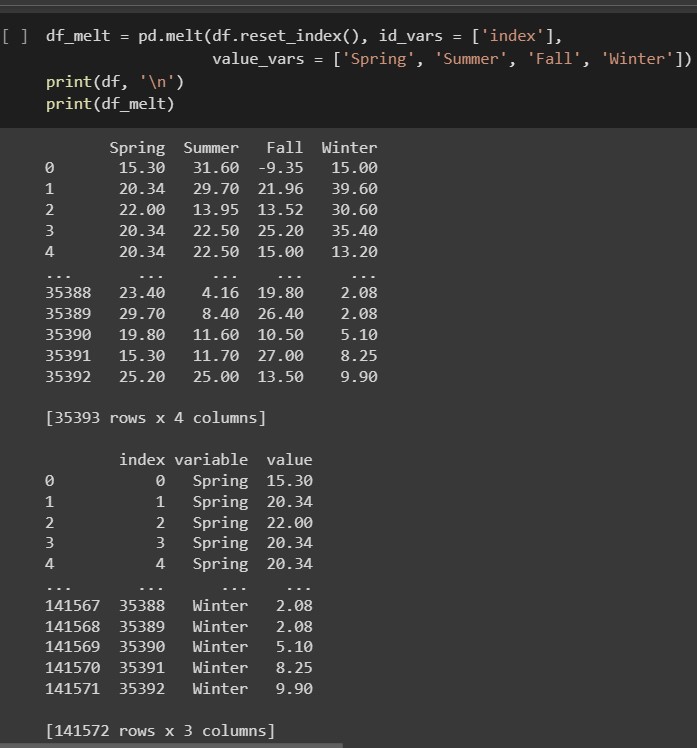
H0: Spring, Summer, Fall, Winter có phân phối chuẩn

Ha: Các mẫu không có phân phối chuẩn



**Bước 2:**

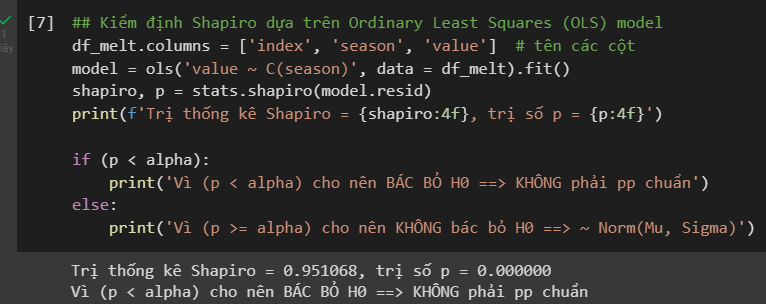
Vì quy định của statsmodels khi kiểm định Shapiro yêu cầu tên cột là feature, không được phép là giá trị, vì vậy ta tiến hành chuyển dạng dữ liệu về dạng cơ sở dữ liệu (features – values) bằng hàm melt để chuyển tên cột là biến và đưa giá trị xuống dưới.



**Bước 3:** Kiểm định Shapiro dựa trên Ordinary Least Squares (OLS) model và kết luận dựa trên phương pháp p-value.

Với điều kiện nếu p-value < alpha, ta sẽ bác bỏ giả thuyết Ho, và nếu p-value >= alpha, ta không thể bác bỏ giả thuyết Ho.

Ở đây p-value xấp xỉ bằng 0 và nhỏ hơn alpha = 0.05. Vì vậy ta sẽ bác bỏ Ho và đưa ra được kết luận Totalsale theo từng mùa không có phân phối chuẩn



Từ kết quả của kiểm định Levene và Shapiro, ta nhận thấy phân phối của Totalsale theo từng Season không cùng phương sai và không có phân phối chuẩn. Vì vậy ta không thể thực hiện kiểm định ANOVA về sự độc lập giữa Totalsale và Season

## Giảm Chiều Dữ Liệu

SVD (Singular Value Decomposition) là một trong những phương pháp làm giảm chiều dữ liệu giúp giảm sức ép lên các thuật toán, giảm độ nặng của tập dữ liệu cũng như giúp dễ hình dung, dễ trực quan hóa dữ liệu hơn mà không tạo ra sai số quá lớn. Ở đây, nhóm áp dụng SVD để giảm chiều dữ liệu với mục đích giảm sự phức tạp, khó quan sát của các biểu đồ biểu diễn phân cụm dữ liệu. Cụ thể là sau khi phân cụm dữ liệu theo nhóm khách hàng và hành vi (Customer Segmentation by Country and Customer Behaviour - K-prototypes), nhóm nhận ra được sự khó quan sát, dữ liệu chưa thực sự đạt được ý nghĩa của việc “trực quan hóa dữ liệu” nên quyết định giảm chiều dữ liệu để dễ quan sát và hiểu được ý nghĩa, mối tương quan giữa các cụm dữ liệu.

Đây là biểu đồ paired plot sau khi phân cụm ban đầu:

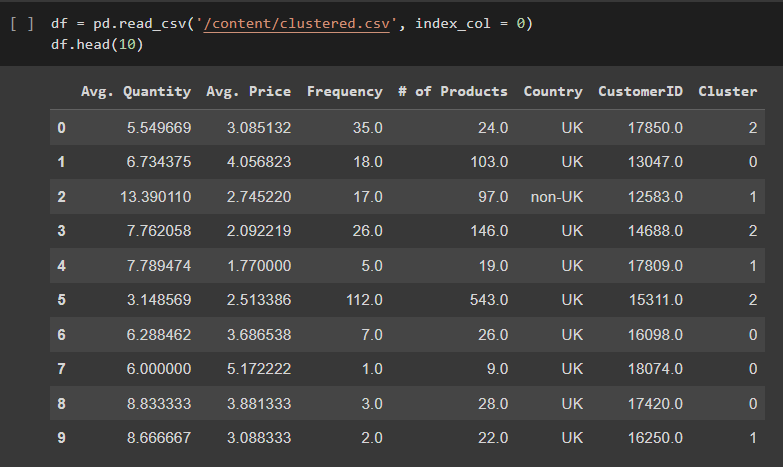


Hình 21: Biểu đồ Paired Plot phân cụm theo Country và Customer Behaviour.

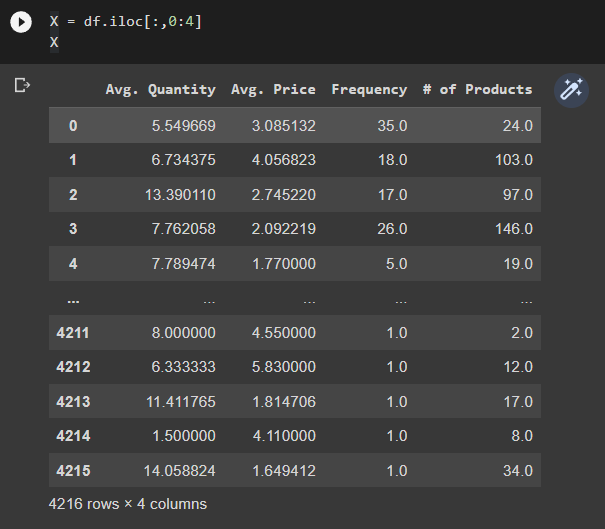
Các biểu đồ trên tuy đã thể hiện được mối tương quan, sự phân cụm của dữ liệu nhưng bù lại, phải cần 16 biểu đồ khác nhau kết hợp và kèm theo đó sẽ mất thời gian để hiểu được toàn bộ ý nghĩa, sự tương quan của các cụm dữ liệu.

Thực hiện giảm chiều dữ liệu:

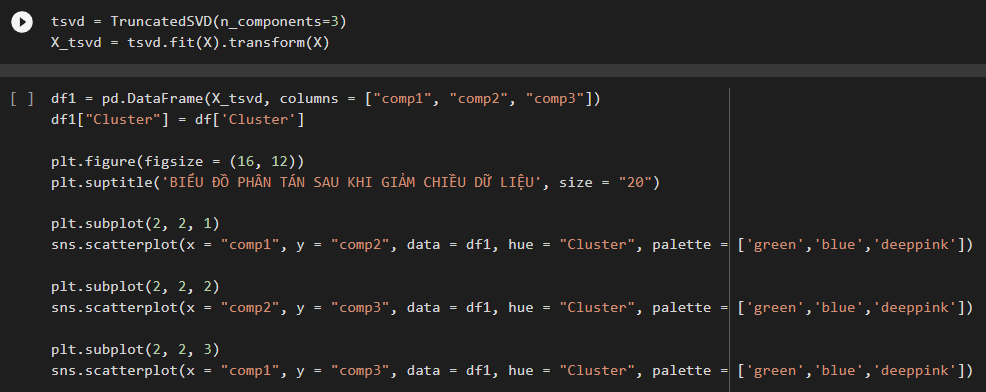
**Bước 1:** Đọc dữ liệu sau phân cụm

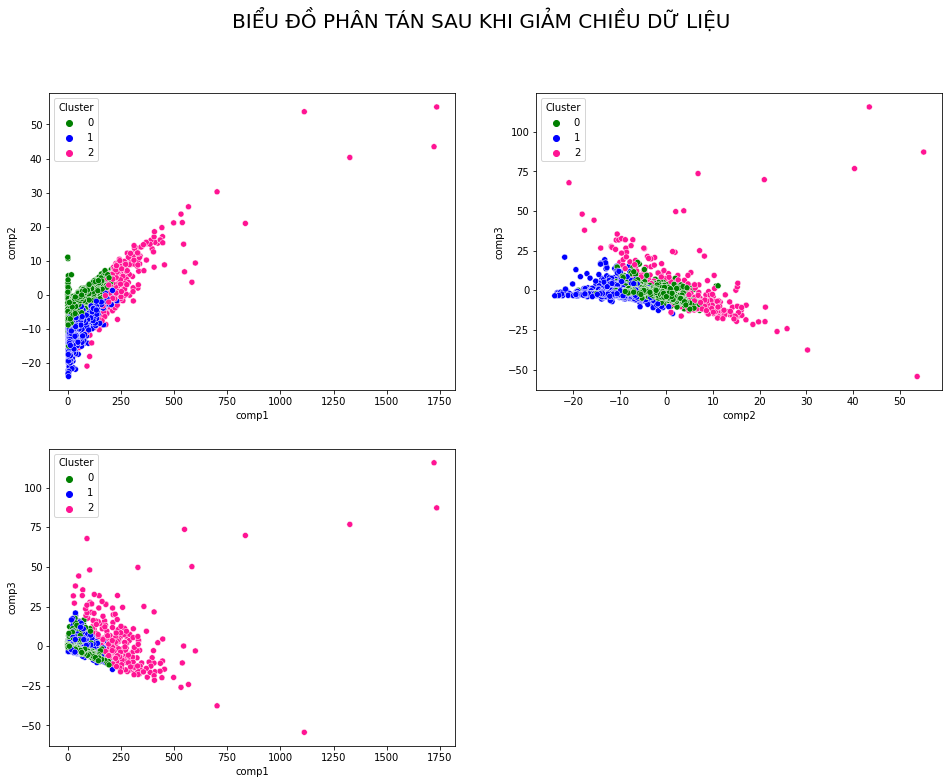
****

**Bước 2:** Bỏ cột CustomerID, Country do không cần thiết vì cột Cluster đã phân cụm theo nhóm khách hàng và hành vi khách hàng. Cột Cluster sẽ được tách riêng cho bước sau

****

**Bước 3:** Sử dụng hàm truncatedSVD giảm chiều dữ liệu từ 4 xuống 3 chiều, sau đó dùng phép chiếu để lấy tọa độ trong không gian dữ liệu mới. Tạo ra dataframe mới với 3 cột thuộc tính đặt tên là “comp1”, “comp2”,  “comp3” (3 thuộc tính sau khi giảm chiều dữ liệu sử dụng phương pháp SVD). Ghép cột Cluster vào dataframe mới. Rồi tiến hành vẽ biểu đồ phân tán thể hiện sự tương quan, phân cụm giữa 3 cột thuộc tính:

****

****

Hình 22: Biểu đồ phân tán sau khi giảm chiều dữ liệu.

**Nhận xét:** sau khi giảm chiều dữ liệu, ta có thể quan sát, dễ hình dung được sự tương quan, sự phân cụm của các dữ liệu hơn. Và nếu để ý kĩ hơn ta thấy được rằng có 1 số sự tương đồng nhất định giữa 3 biểu đồ phân tán sau giảm chiều dữ liệu với các biểu đồ phân tán trước khi giảm chiều dữ liệu. Điều này cho ta thấy rằng, việc giảm chiều dữ liệu tuy vẫn có các sai số nhất định, nhưng vẫn trong phạm vi cho phép, vẫn có sự phân cụm gần như ban đầu nhưng giảm được số biểu đồ cần thể hiện, giảm được sự khó khăn để hiểu được mối tương quan, sự phân cụm dữ liệu.

**Đánh giá biểu đồ**: biểu đồ nhìn chung có màu sắc dễ nhìn, các thông tin như tên biểu đồ, chú thích, tên chiều dữ liệu dễ nhìn. Tuy nhiên, cả 3 biểu đồ dùng chung 1 chú thích nhưng chưa tối ưu hóa được bằng việc gom cả 3 chú thích đó làm một chú thích chung thống nhất cho subplot.

# CHƯƠNG VI: KẾT LUẬN

Với đề tài “Biểu diễn trực quan dữ liệu”, nhóm đã sử dụng bộ *Phân tích dữ liệu kinh doanh của cửa hàng bán lẻ trực tuyến,* trực thuộc đồ án của môn “Biểu diễn trực quan dữ liệu” để tiến hành biểu diễn trực quan các dữ liệu và phát triển đề tài theo hướng trực quan hoá hơn, dễ tiếp nhận thông tin hơn, nhóm cũng đã áp dụng được các nguyên tắc thiết kế biểu đồ, sử dụng các biểu đồ đúng mục đích.

Việc sử dụng biểu đồ để trực quan hóa dữ liệu cũng giúp người phân tích và những người ra quyết định dễ dàng hiểu được ý nghĩa của dữ liệu để thúc đẩy các quyết định trong kinh doanh và quản lý. Phương thức phổ biến nhất làm cho dữ liệu trở nên trực quan là thông qua biểu đồ. Ở đây, nhóm đã sử dụng các loại biểu đồ phổ biến như biểu đồ thanh, biểu đồ đường, biểu đồ điểm, biểu đồ cây. Ngoài ra, nhóm còn sử dụng các nguyên tắc về màu sắc biểu đồ, màu sắc theo thang định tính ***(qualitative color scale)*** được dùng để thể hiện sự phân biệt các nhóm rời rạc không có trật tự nội tại, chẳng hạn như các quốc gia khác nhau hoặc các nhà sản xuất khác nhau của một sản phẩm nhất định, thang màu tuần tự ***(sequential color scale)*** được sử dụng để biểu diễn các giá trị dữ liệu định lượng. Tất cả yếu tố trên đều nhằm mục đích phục vụ cho việc trực quan hoá bộ dữ liệu, nhằm hỗ trợ người đọc báo cáo thuận tiện hơn trong việc đưa ra các quyết định, là một sự phát triển so với đồ án “Lập trình phân tích dữ liệu”.

Tuy nhiên, đồ án vẫn còn một số điểm chưa thực hiện được do các vấn đề khách quan và chủ quan. Với bộ dữ liệu trên, việc áp dụng phương pháp Principal Component Analysis (PCA) để giảm chiều dữ liệu là không khả thi vì bộ dữ liệu chưa được dán nhãn phân lớp dữ liệu. Ngoài ra, nhóm vẫn còn một số điểm chưa cải thiện, tối ưu hóa ở một số biểu đồ, ví dụ như “Biểu đồ phân tán sau khi giảm chiều dữ liệu”, mặc dù cả 3 biểu đồ có cùng 1 bảng chú thích các màu được biểu diễn trong biểu đồ nhưng vẫn chưa tổng hợp ra thành 1 bảng chung cho 3 biểu đồ.

# Tài Liệu Tham Khảo

[1] Empowering Business with Effective Insights, *“Online Retails Sale Dataset”,* 2022. Địa chỉ: <https://www.kaggle.com/datasets/rohitmahulkar/online-retails-sale-dataset>

[2] Big Data Uni, “Tìm hiểu về Time Series Phần 1”, 2021.

Địa chỉ: <https://bigdatauni.com/tin-tuc/tim-hieu-ve-time-series-phan-tich-chuoi-thoi-gian-p-1.html>

[3] Kaggle, “*Beginner Friendly EDA + FP-Growth Algorithm”,* 2022.

Địa chỉ: <https://www.kaggle.com/code/ahmedhammad01/beginner-friendly-eda-fp-growth-algorithm/notebook>

[4] Mlxtend, “ Apriori”.

Địa chỉ: [https://rasbt.github.io/mlxtend/api\_subpackages/mlxtend.frequent\_patterns/ - association\_rules](https://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/#association_rules)

[5] Bài giảng học phần Khai phá dữ liệu, *“Chương 4: Luật kết hợp”*, TS.Nguyễn An Tế, khoa Công nghệ thông tin kinh doanh, trường Công nghệ và Thiết kế UEH, 2022.

[6] Lily Wu, “Clustering Product Names with Python - Part 1”, 2021.

Địa chỉ:  <https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fclustering-product-names-with-python-part-1-f9418f8705c8>

[7] Lily Wu, “Clustering Product Names with Python - Part 2”, 2021.

Địa chỉ: <https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fclustering-product-names-with-python-part-1-f9418f8705c8>

[8] Dương Phạm, “*TF-IDF( term frequency – inverse document frequency)”,*2016.

Địa chỉ: <https://viblo.asia/p/tf-idf-term-frequency-inverse-document-frequency-JQVkVZgKkyd>

[9] Kaggle, “*Store Customers Clustering Analysis*”, 2021.

Địa chỉ: <https://www.kaggle.com/code/miguelfzzz/store-customers-clustering-analysis>

[10] Github, “Data analysis online retail transactions”, 2019.

Địa chỉ: <https://github.com/amir-hojjati/Data-Analysis-Online-Retail-Transactions>

[11] Các bài giảng học phần Biểu diễn trực quan dữ liệu, TS. Nguyễn An Tế, khoa Công nghệ thông tin trong kinh doanh, trường Công nghệ và Thiết kế UEH, 2022.

# Bảng phân công:

|  |  |  |
| --- | --- | --- |
| Thành viên | Phân công | Đánh giá |
| Nguyễn Phúc Hải | FP-Growth, Giảm chiều dữ liệu, Đánh giá-cải thiện biểu đồ, Tiền xử lý dữ liệu, Báo cáo phần Word. | 20% |
| Nguyễn Văn Hoàng Dũng | K Means, Timeseries, Tiền xử lý dữ liệu, Trực quan hóa K means, Cải thiện biểu đồ, Báo cáo phần Word. | 20% |
| Lê Đức Dũng | Khám phá dữ liệu (EDA), Timeseries, Đánh giá-nhận xét biểu đồ, Báo cáo phần Word. | 20% |
| Võ Ngọc Dung | Tiền xử lý dữ liệu, Đánh giá-nhận xét biểu đồ, Chỉnh sửa, hoàn thiện - Báo cáo phần Word. | 20% |
| Nguyễn Quỳnh Khánh Hà | K-prototypes, Kiểm định, Chỉnh sửa, hoàn thiện - Báo cáo phần Word . | 20% |