

**TRƯỜNG ĐẠI HỌC THỦY LỢI KHOA CÔNG NGHỆ THÔNG TIN**

**QUẢN LÝ DỮ LIỆU LỚN**

**Đề tài:**

**Phân cụm khách hàng trong shop thời trang bằng**

**phương pháp phân cụm Kmeans**

Giảng viên: Nguyễn Đắc Phương Thảo

Nhóm sinh viên thực hiện: Nhóm 6

1. Chu Đức Hoàng- 2251262604
2. Phạm Tiến Thành- 2251262645
3. Nguyễn Ngọc Tân- 2251262637
4. Mai Hoàng Tùng- 2251262656
5. Nguyễn Đình Thành- 2251262643

**HÀ NỘI, 2024**

[LỜI CẢM ƠN 3](#_Toc181449096)

[I. Mô tả bài toán 4](#_Toc181449097)

[1. Lý do chọn đề tài 4](#_Toc181449098)

[2. Tổng quan bài toán 4](#_Toc181449099)

[3. Quy trình thực hiện 5](#_Toc181449100)

[II. Triển khai thuật toán K-means 6](#_Toc181449101)

[1. Phân cụm 6](#_Toc181449102)

[2. Gán các cụm vào dữ liệu 12](#_Toc181449103)

[III. Phân tích dữ liệu theo cụm 15](#_Toc181449104)

[1.Biểu đồ cột đôi tỉ lệ nam nữ của từng cụm: 15](#_Toc181449105)

[2.Biểu đồ tương quan giữa mặt hàng và tuổi 15](#_Toc181449106)

[3.Biểu đồ đường kết hợp về tuổi: 18](#_Toc181449107)

[4.Biểu đồ cột về số lượng mặt hàng tại khu vực trong từng cụm: 19](#_Toc181449108)

[5.Biểu đồ phân tán giữa item, size, color: 19](#_Toc181449109)

[6.Biểu đồ top 10 mặt hàng bán chạy nhất: 23](#_Toc181449110)

[7.Biểu đồ cột tần suất mua hàng: 26](#_Toc181449111)

[8.Biểu đồ Heatmap thể hiện số lượng sản phẩm và màu sắc của chúng: 28](#_Toc181449112)

[IV. Tổng kết 32](#_Toc181449113)

[V. Phương pháp cải thiện 33](#_Toc181449114)

[KẾT LUẬN 35](#_Toc181449115)

# LỜI CẢM ƠN

Trong bối cảnh thị trường thời trang ngày càng phát triển và cạnh tranh, việc hiểu rõ hoạt động kinh doanh của các shop quần áo trở thành một yếu tố then chốt để đạt được thành công và thu hút khách hàng. Việc chọn và phân tích khách hàng của mình bằng phương pháp phân cụm K-means không chỉ nhằm hiểu sâu hơn về tình hình kinh doanh, mà còn giúp chúng em tối ưu hóa hoạt động quản lý và phát triển chiến lược để thu hút thêm khách hàng. Thông qua việc phân tích các yếu tố như đối tượng khách hàng, xu hướng mua sắm, hiệu quả chiến lược marketing, và các sản phẩm bán chạy, chúng em hy vọng sẽ nhận diện rõ hơn điểm mạnh và những điểm cần cải thiện. Từ đó, chúng em có thể điều chỉnh phương thức kinh doanh và mang lại trải nghiệm tốt nhất cho khách hàng của mình, đồng thời tạo ra lợi thế cạnh tranh trên thị trường.

Chúng em chân thành xin gửi lời cảm ơn đặc biệt đến cô Nguyễn Đắc Phương Thảo – người đã tận tình giảng dạy môn Quản lí dữ liệu lớn cho chúng em trong từng buổi học. Cô đã giúp chúng em trang bị kiến thức môn học và hơn cả là động lực tiếp tục trên con đường chinh phục công nghệ.

Bên cạnh những kết quả mà chúng em đạt được thì sẽ không khó tránh khỏi những thiếu sót trong quá trình làm đề tài vì thời gian không cho phép và chưa có kinh nghiệm thực tế. Chính vì vậy chúng em rất mong được sự cảm thông, chỉ bảo góp ý của cô. Những lời nhận xét, góp ý của cô chính là một bài học, kiến thức cho chúng em trên con đường sau này.

Chúng em xin chân thành cảm ơn!

## Mô tả bài toán

### Lý do chọn đề tài

Trong bối cảnh kinh doanh hiện đại, phân cụm khách hàng trở thành công cụ quan trọng giúp các cửa hàng quần áo hiểu sâu sắc nhu cầu và hành vi của từng nhóm khách hàng. Việc phân chia khách hàng thành các cụm dựa trên các đặc điểm tương đồng cho phép cửa hàng tối ưu hóa chiến lược tiếp thị, cá nhân hóa sản phẩm phù hợp với từng nhóm và nâng cao trải nghiệm mua sắm. Điều này không chỉ giúp tăng cường sự hài lòng mà còn thúc đẩy lòng trung thành và tối đa hóa giá trị dài hạn của khách hàng đối với cửa hàng.

### Tổng quan bài toán

Bài toán của chúng em tập trung vào việc phân tích danh sách dữ liệu khách hàng đã mua hàng trong cửa hàng, từ đó phân cụm khách hàng theo các khoảng giá. Phương pháp này giúp cửa hàng nhận diện các nhóm khách hàng có khả năng chi tiêu tương đồng, từ đó phát triển các chiến lược tiếp thị và gợi ý sản phẩm phù hợp với từng phân khúc.

Danh sách dữ liệu bao gồm 17 thuộc tính và 3900 dữ liệu



|  |  |
| --- | --- |
| Tên thuộc tính | Ý nghĩa |
| Customer | Khách hàng |
| Age | Tuổi |
| Gender | Giới tính |
| Item purchased | Mặt hàng đã mua |
| Category | Loại mặt hàng |
| Purchased amount (USD) | Số tiền đã mua |
| Location | Địa điểm |
| Size | Kích cỡ |
| Color | Màu sắc |
| Season | Mùa |
| Review rating | Đánh giá xếp hạng |
| Subscription status | Trạng thái đăng kí |
| Payment method | Phương thức thanh toán |
| Shipping type | Cách thức vận chuyển |
| Discount applied | Áp dụng giảm giá |
| Previous purchased | Lịch sử mua hàng |
| Frequency | Tần suất mua hàng |

### 3. Quy trình thực hiện

**3.1. Tiền xử lý dữ liệu**

Để phân cụm khách hàng dựa trên khoảng giá bằng thuật toán K-means, việc sử dụng hai cột *Purchased amount (USD)* và *Previous purchased* sẽ giúp tạo ra các nhóm khách hàng dựa trên mức độ chi tiêu và tần suất mua sắm trước đó.



**3.2. Phân cụm bằng thuật toán K-means**

**K-means** là một trong những thuật toán phân cụm (clustering) phổ biến nhất trong học máy không giám sát (unsupervised learning). Thuật toán này giúp chia một tập dữ liệu lớn thành các nhóm nhỏ hơn (cụm) dựa trên sự tương đồng giữa các điểm dữ liệu. Mục tiêu chính của K-means là tìm ra các cụm sao cho các điểm dữ liệu trong cùng một cụm có độ tương đồng cao nhất và các điểm dữ liệu thuộc các cụm khác nhau có độ tương đồng thấp nhất.

**Các bước của thuật toán K-means:**

* Bước 1:Tính khoảng cách giữa các điểm bằng khoảng cách Euclidean

Với x,y là các điểm dữ liệu cần xét. Mỗi điểm dữ liệu sẽ được gán vào cụm có tâm gần nó nhất.

* Bước 2:Khởi tạo tâm cụm:

Tâm cụm mới sẽ được tính bằng trung bình cộng của tất cả các điểm dữ liệu thuộc cụm đó. Giả sử cụm C có m điểm dữ liệu, tâm cụm mới c được tính như sau:

Trong đó:

* là điểm dữ liệu thứ i thuộc cụm C
* là tổng của tất cả các điểm dữ liệu trong cụm C
* Sau khi tìm được tâm cụm thì chúng em tiếp tục gán các điểm vào tâm
* Khi đã tìm được các điểm thuộc cụm nào chũng em lại phân chia dữ liệu thành các cụm riêng biệt.

**3.3. Phân tích các cụm và đưa ra phương pháp cải thiện**

Sau khi đã phân chia khách hàng thành 3 cụm, bước tiếp theo là tách riêng các cụm này để trực quan hóa các thuộc tính, qua đó phân tích chi tiết hành vi của từng nhóm khách hàng.

## II. Triển khai thuật toán K-means

### 1. Phân cụm

**1.1 PointWritable**

* Lưu trữ thuộc tính, số chiều và số lượng của các điểm

A black background with white text

Description automatically generated

* Khởi tạo đối tượng PointWritalbe

A screen shot of a computer program

Description automatically generated

* Tạo một bản sao của PointWritable để làm việc trên bản sao đó

A computer screen shot of a black background

Description automatically generated

* Gán các giá trị tọa độ, độ dài và thiết lập số lượng điểm

A computer screen with colorful text

Description automatically generated

* Đọc dữ liệu từ DataInput

A screen shot of a computer code

Description automatically generated

* Ghi dữ liệu vào DataOutput

A screen shot of a computer code

Description automatically generated

* Chuyển đổi điểm thành chuỗi để dễ dàng lưu trữ

A computer screen shot of text

Description automatically generated

* Cộng dồn các tọa độ và số lượng điểm

A screen shot of a computer code

Description automatically generated

* Tính khoảng cách Euclidean giữa 2 điểm

A screen shot of a computer code

Description automatically generated

* Tính trung bình các tọa độ của điểm

A computer code on a black background

Description automatically generated

**1.2. Mapper**

* Khởi tạo các biến

A screen shot of a computer

Description automatically generated

* Khởi tạo tâm cụm được lấy từ cấu hình Hadoop (được gọi 1 lần trước khi thực thi map)

A screen shot of a computer code

Description automatically generated

* Tính khoảng cách giữa điểm và tâm cụm rồi tìm tâm cụm có khoảng cách nhỏ nhất bằng phương thức calcDistance được định nghĩa trong PointWritable

A screen shot of a computer code

Description automatically generated

**1.3. Combiner**

* Tính tổng các thuộc tính của các điểm trong mỗi cụm trước khi truyền đến Reducer

A screen shot of a computer program

Description automatically generated

**1.4. Reducer**

* Tính trung bình tất cả giá trị của tâm cụm và ghi ra tâm cụm mới

A screen shot of a computer code

Description automatically generated

**1.5 Main**

* Nhập dữ liệu tâm cụm từ bàn phím

A screen shot of a computer program

Description automatically generated

* Lưu trữ các tâm cụm vào Configurationcủa Hadoop

A screen shot of a computer program

Description automatically generated

* Đọc tâm cụm mới từ đầu ra của Reducer và lưu chúng vào PointWritable

A screen shot of a computer program

Description automatically generated

* So sánh khoảng cách giữa các tâm cụm cũ và mới, nếu khoảng cách nhỏ hơn ngưỡng threshold thì vòng lặp dừng lại, nếu không thì tiếp tục chạy

A screen shot of a computer program

Description automatically generated

* Ghi kết quả phân cụm cuối cùng vào HDFS

A screen shot of a computer code

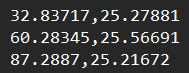
Description automatically generated

* Tạo một bản sao của mảng các điểm tâm cụm được truyền vào

A screen shot of a computer program

Description automatically generated

* **Kết quả:**



### 2. Gán các cụm vào dữ liệu

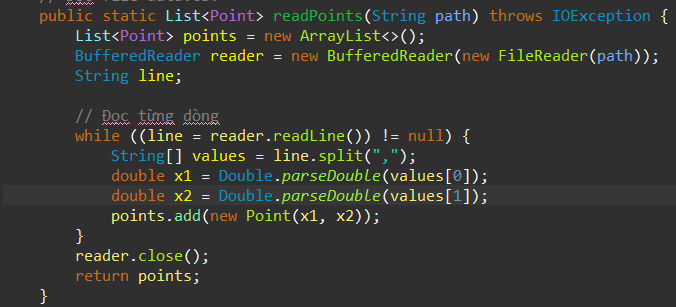
**2.1 Main**

* Đọc dữ liệu từ các file data

A screen shot of a computer code

Description automatically generated

* Đọc và chuyển đổi dữ liệu thô từ file thành các đối tượng Point



* Đọc tâm cụm từ file data

A screen shot of a computer code

Description automatically generated

* Tính toán và gán điểm vào tâm cụm

A screen shot of a computer program

Description automatically generated

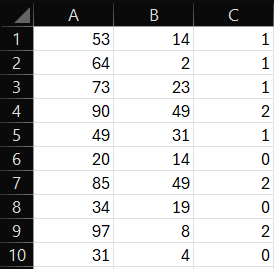
**2.2 Point**

* Tính khoảng cách giữa 2 điểm

A computer screen with text

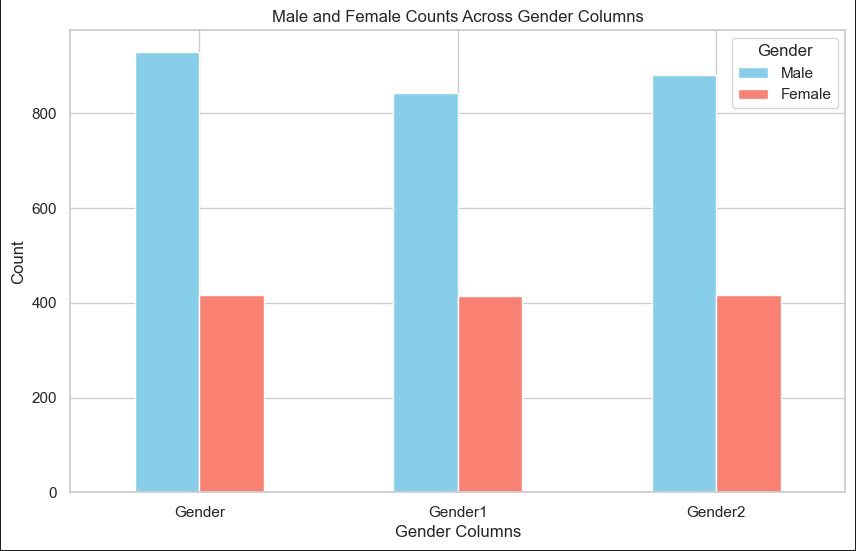
Description automatically generated

* **Kết quả:**



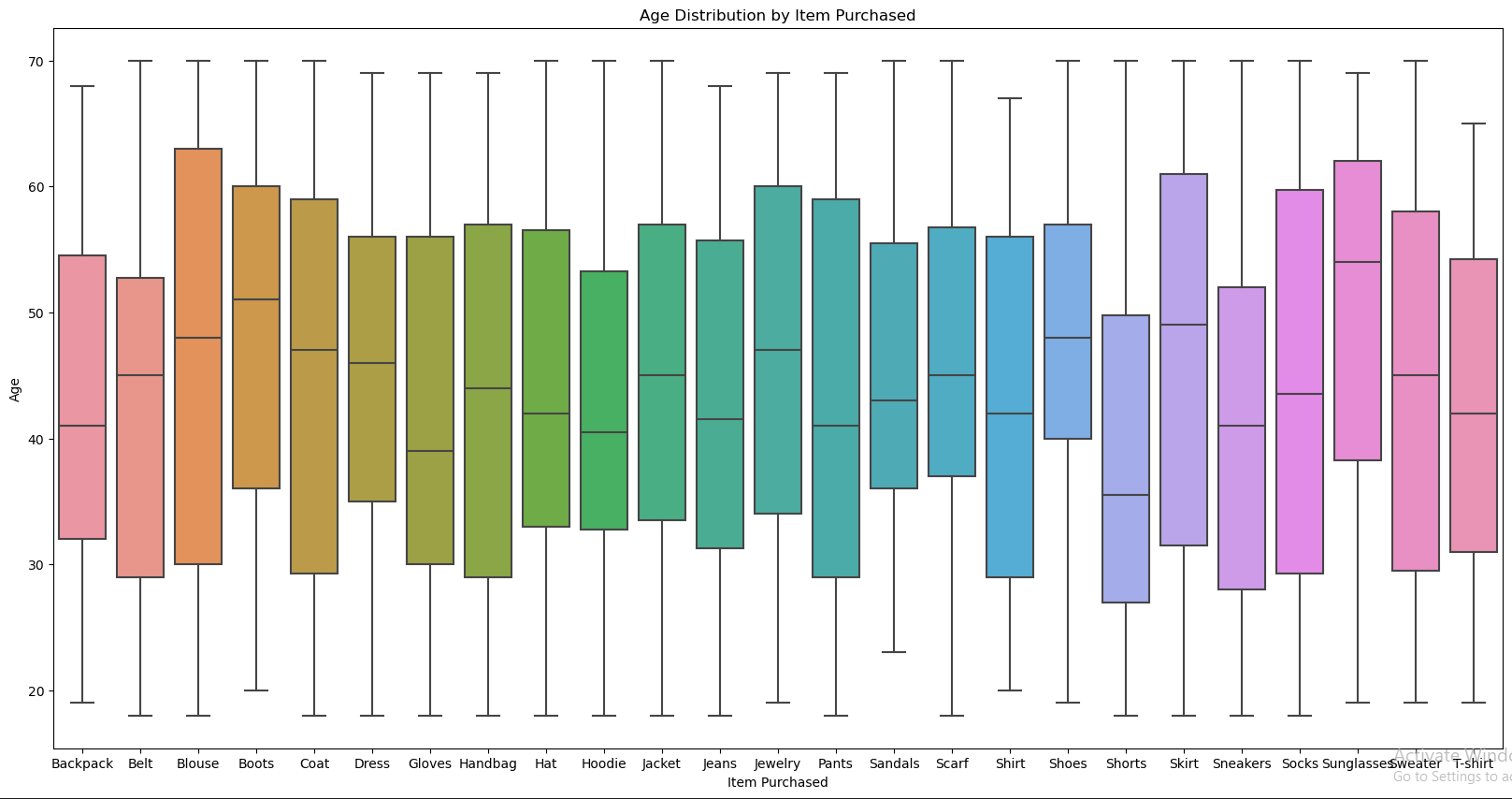
## III. Phân tích dữ liệu theo cụm

### 1.Biểu đồ cột đôi tỉ lệ nam nữ của từng cụm:



* **Số lượng nam luôn cao hơn nữ:** Ở cả ba cột, số lượng nam đều nhiều hơn số lượng nữ. Điều này cho thấy một sự chênh lệch giới tính đáng kể.
* **Sự khác biệt giữa các cột:** Mặc dù số lượng nam luôn cao hơn, nhưng mức độ chênh lệch giữa các cột có thể khác nhau. Ví dụ, sự chênh lệch giữa nam và nữ ở cột "Gender" có thể lớn hơn so với cột "Gender1".

### 2.Biểu đồ tương quan giữa mặt hàng và tuổi

* **Cụm 0**:

Biểu đồ hộp (box plot) ở cụm 1 hiển thị phân bố độ tuổi của khách hàng theo các loại mặt hàng mà họ đã mua

* **Độ tuổi trung bình**:

Nhiều loại mặt hàng có độ tuổi trung bình của khách hàng nằm trong khoảng từ 40 đến 60, với một số loại có độ tuổi trung bình thấp hơn hoặc cao hơn một chút.

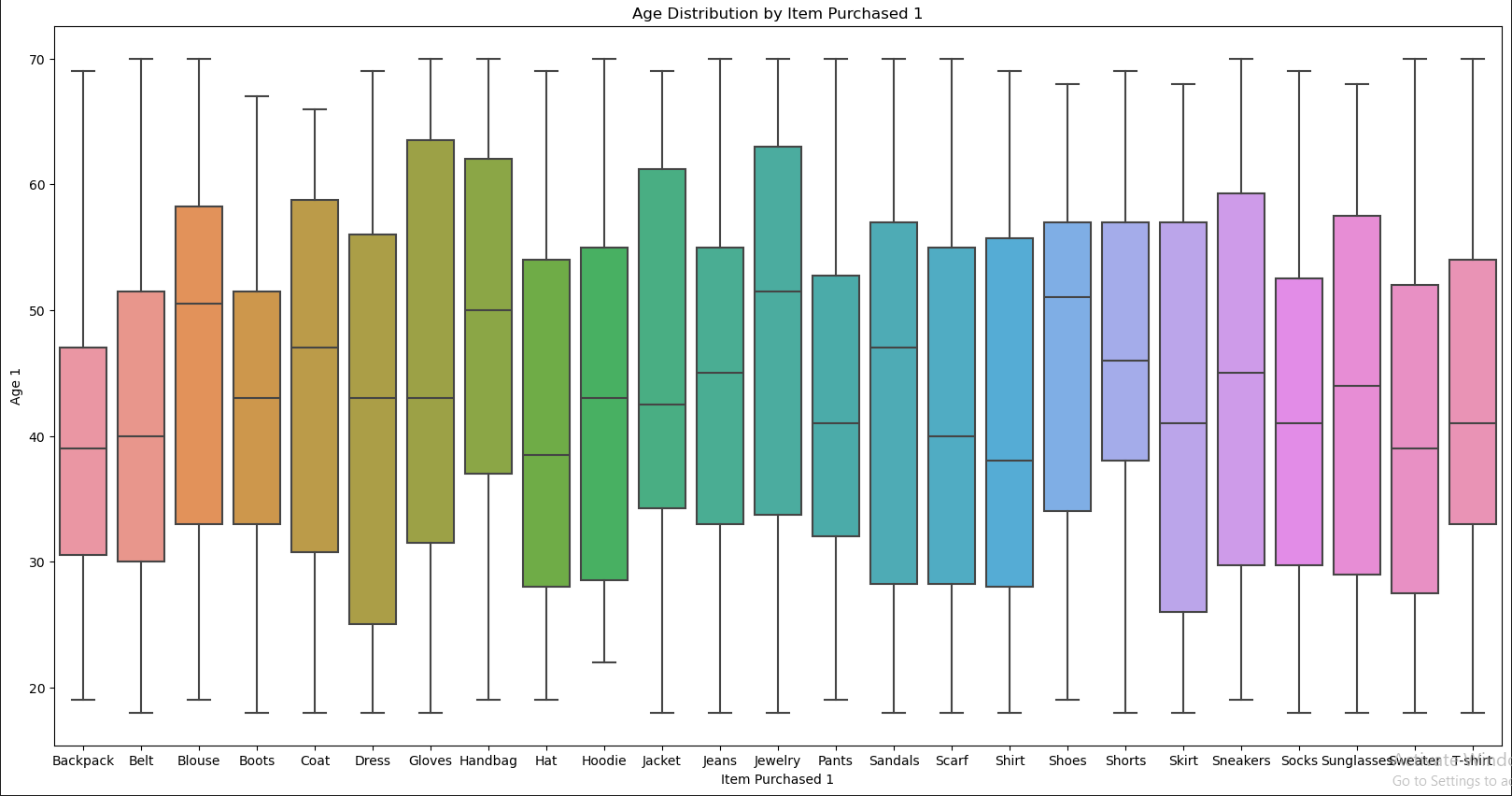
* **Sản phẩm với độ tuổi người mua trẻ**:

**T-shirt**, **Shorts**, và **Skirt** có độ tuổi trung bình của người mua thấp hơn, nằm trong khoảng 30-40 tuổi, những mặt hàng này có thể phổ biến hơn với đối tượng trẻ tuổi hơn

* **Sản phẩm với độ tuổi người mua cao**:

Các sản phẩm như **Blouse**, **Boots**, **Dress** và **Sweater** dường như thu hút nhiều người mua ở độ tuổi lớn hơn, với độ tuổi trung bình và tứ phân vị trên (upper quartile) nằm trong khoảng 50-60 tuổi.

* Các sản phẩm như **Sneakers** và **Hoodie** có độ phân tán lớn, cho thấy các mặt hàng này được mua bởi nhóm người có độ tuổi rất đa dạng.
* Một số sản phẩm khác, như **Skirt** và **Shorts**, có phân tán nhỏ hơn, điều này có thể gợi ý rằng chúng thu hút đối tượng tuổi nhất định hơn.
* Một số loại sản phẩm có các điểm dữ liệu ngoài (outliers) như **Jewelry** và **Dress**, cho thấy có những khách hàng có độ tuổi rất khác biệt so với trung bình khi mua các sản phẩm này.
* **Cụm 1:**



* **Phân phối độ tuổi chung**:

Độ tuổi người mua cho các mặt hàng dao động từ khoảng 20 đến 70 tuổi.

Độ tuổi trung bình cho hầu hết các mặt hàng nằm trong khoảng 40 đến 60, cho thấy cửa hàng phục vụ đa số khách hàng trong độ tuổi trung niên.

* **Nhóm sản phẩm thu hút khách hàng ở độ tuổi cao**:

Các sản phẩm như **Blouse**, **Boots**, **Coat**, **Dress**, và **Gloves** có độ tuổi trung bình cao hơn, nằm trong khoảng từ 50 đến 60 tuổi. Điều này cho thấy các mặt hàng này có thể phổ biến hơn với khách hàng lớn tuổi.

* **Nhóm sản phẩm thu hút khách hàng trẻ tuổi**:

**T-shirt**, **Shorts**, và **Skirt** có độ tuổi trung bình của người mua thấp hơn, trong khoảng từ 30 đến 40 tuổi. Những sản phẩm này có thể thu hút nhiều khách hàng trẻ hơn.

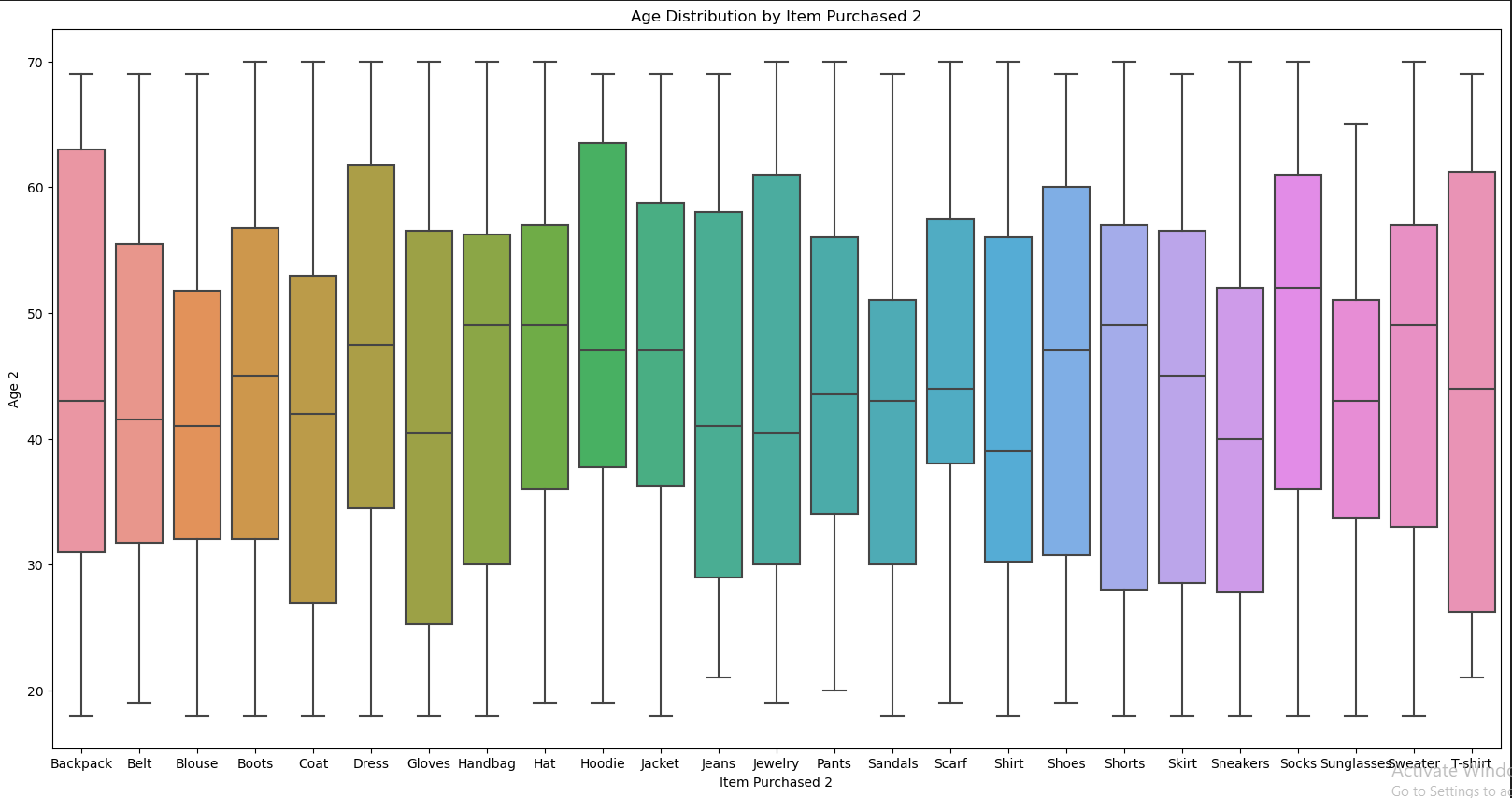
* **Phân tán độ tuổi người mua**:

**Sneakers**, **Hat**, và **Hoodie** có sự phân tán độ tuổi lớn, cho thấy các sản phẩm này được mua bởi các nhóm khách hàng ở nhiều độ tuổi khác nhau.

Các sản phẩm như **Skirt** và **Shorts** có phân tán nhỏ hơn, có thể là do chúng chủ yếu hấp dẫn một nhóm tuổi nhất định.

=> Một số sản phẩm có các điểm dữ liệu ngoài, chẳng hạn như **Jewelry** và **Dress**, cho thấy có một số ít khách hàng có độ tuổi khác biệt rõ rệt so với đa số khi mua các sản phẩm này.

* **Cụm 2:**



**+)Phân phối độ tuổi chung:**

* Độ tuổi người mua các mặt hàng dao động từ khoảng 20 đến 70 tuổi.
* Độ tuổi trung bình cho hầu hết các mặt hàng nằm trong khoảng 40 đến 60, cho thấy cửa hàng phục vụ phần lớn khách hàng trong độ tuổi trung niên.

**+) Nhóm sản phẩm thu hút khách hàng ở độ tuổi cao hơn:**

* Các sản phẩm như **Blouse, Boots, Coat, Dress, và Gloves** có độ tuổi trung bình cao hơn, từ 50 đến 60 tuổi. Điều này cho thấy các mặt hàng này phổ biến hơn với khách hàng lớn tuổi.

**+) Nhóm sản phẩm thu hút khách hàng trẻ tuổi:**

* Các sản phẩm như **T-shirt, Shorts, và Skirt** có độ tuổi trung bình người mua thấp hơn, từ 30 đến 40 tuổi, cho thấy chúng có thể thu hút nhiều khách hàng trẻ tuổi hơn.

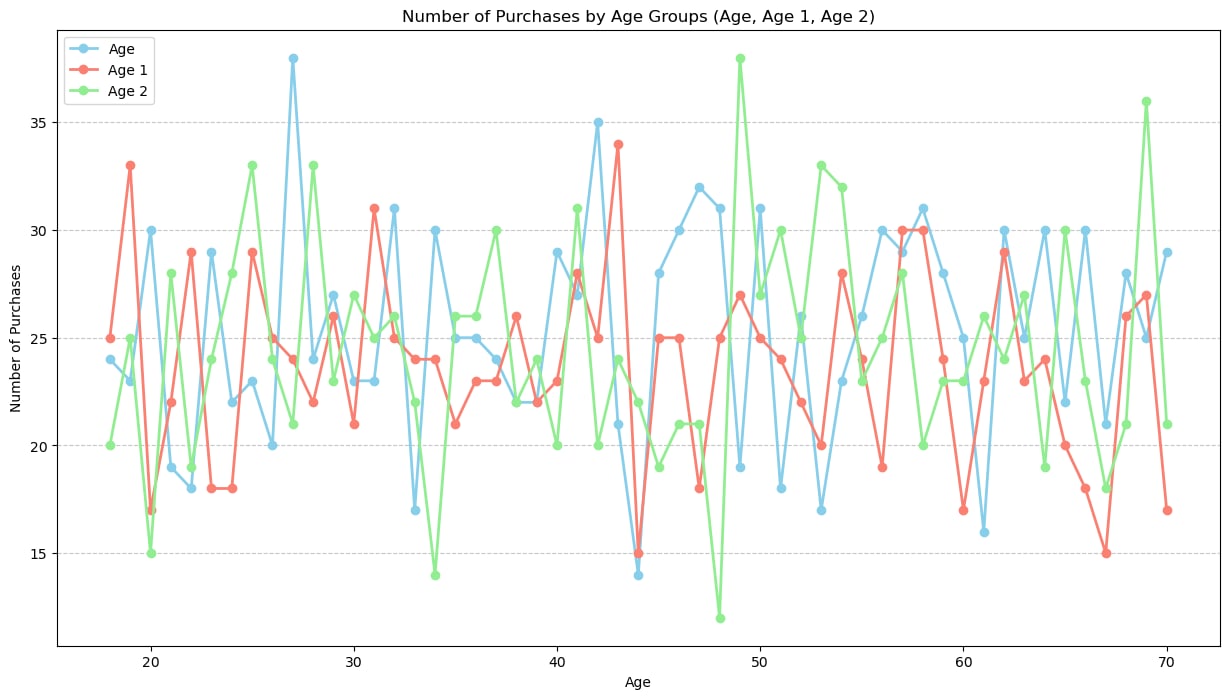
**+) Phân tán độ tuổi người mua:**

* **Sneakers, Hat, và Hoodie** có sự phân tán độ tuổi lớn, cho thấy các sản phẩm này thu hút nhiều nhóm tuổi khác nhau.
* **Skirt và Shorts** có phân tán nhỏ hơn, có thể vì chúng hấp dẫn một nhóm tuổi nhất định hơn.

+)Một số sản phẩm như **Jewelry và Dress** có các điểm dữ liệu ngoài, cho thấy một số ít khách hàng có độ tuổi khác biệt rõ rệt so với đa số khi mua các sản phẩm này.

### 

### 3.Biểu đồ đường kết hợp về tuổi:



* **Cụm 0:**

Độ tuổi mua hàng nhiều nhất trong khoảng 27-29 tuổi và thấp nhất trong khoảng 43-44 tuổi. Số lượng hàng đã mua chênh lệch ở mức cao.

* **Cụm 1:**

Độ tuổi mua hàng nhiều nhất trong khoảng 18-20 và 41-43 tuổi và thấp nhất trong khoảng 43-44 tuổi. Số lượng hàng đã mua chênh lệch mức trung bình.

* **Cụm 2:**

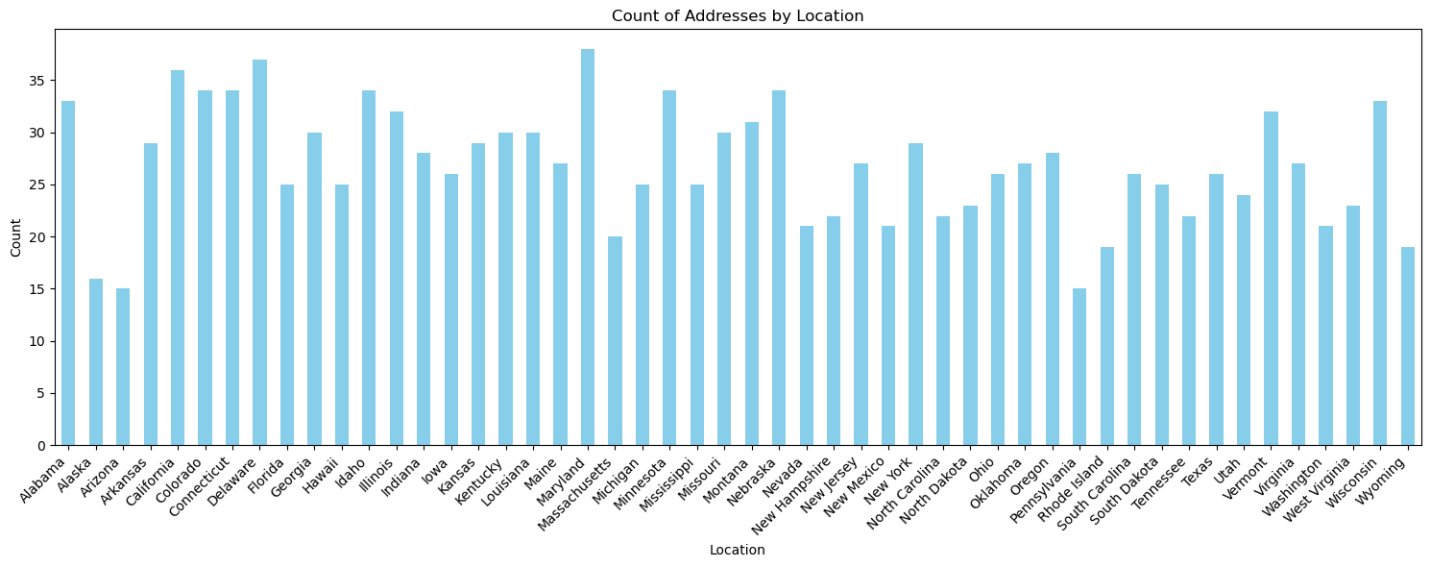
50 tuổi là độ tuổi mua hàng nhiều nhất và thấp nhất trong khoảng 48-49 tuổi. Số lượng mua hàng chênh lệch ở mức cao.

=> Nhận thấy tần suất mua hàng ở độ tuổi 49-50, 27-29 và 69-70 tuổi là cao nhất, ở độ tuổi khoảng 20,35,48 có xu hướng ít. Ngoài ra có một số độ tuổi đang có xu hướng mua hàng tang trong khoảng từ 18-20, 41-43 và giảm dần ở độ tuổi 31-33, 46-48.

### 

### 4.Biểu đồ cột về số lượng mặt hàng tại khu vực trong từng cụm:

* **Cụm 0:**



**+) Sự phân bổ địa chỉ trên các vị trí:**

* Số lượng địa chỉ được phân bổ khá đều trên các vị trí khác nhau.
* Hầu hết các vị trí có số lượng địa chỉ dao động từ 15 đến 30.

**+) Vị trí có số lượng địa chỉ cao:**

* Một số vị trí có số lượng địa chỉ cao hơn, trên 30 địa chỉ. Các vị trí này có thể là những khu vực đông dân cư hoặc các khu vực kinh doanh sầm uất hơn.
* Một số vị trí nổi bật có số lượng địa chỉ cao gồm **Greenfield, Los Angeles, New York City**, và **Nashville**.

**+) Vị trí có số lượng địa chỉ thấp:**

* Các vị trí như **Gatlinburg, Salem, Houston**, và **Pine Bluff** có số lượng địa chỉ thấp hơn, dưới 15 địa chỉ. Điều này có thể phản ánh đặc điểm dân cư hoặc hoạt động thương mại ít sôi động ở những khu vực này.

**+) Tính đồng đều trong phân phối địa chỉ:**

* Biểu đồ cho thấy sự đồng đều tương đối trong số lượng địa chỉ giữa các vị trí, không có vị trí nào quá chênh lệch đáng kể. Điều này có thể cho thấy cửa hàng có sự hiện diện đồng đều và phục vụ nhiều khu vực khác nhau.

**+) Kết luận:**

* Phân phối đồng đều và đa dạng trong số lượng địa chỉ theo vị trí có thể là một chiến lược của cửa hàng nhằm tiếp cận nhiều khách hàng ở các khu vực khác nhau. Các khu vực đông dân hoặc thành phố lớn có xu hướng có nhiều địa chỉ hơn, trong khi các khu vực ít sầm uất hơn có số lượng địa chỉ thấp hơn.
* **Cụm 1:**

### 

**Phân bố địa chỉ theo vị trí:**

* **Phân bố chung:**

Số lượng địa chỉ được phân bổ tương đối đều trên các vị trí khác nhau, với hầu hết các vị trí có từ 15 đến 30 địa chỉ.

* **Vị trí có số lượng địa chỉ cao:**

Một số vị trí có số lượng địa chỉ cao hơn, vượt qua 30 địa chỉ. Đây có thể là các khu vực có mật độ dân cư cao hoặc khu vực có nhiều hoạt động kinh doanh. Các vị trí nổi bật với số lượng địa chỉ cao bao gồm các bang như **Nebraska**, **New York**, và **Montana**.

* **Vị trí có số lượng địa chỉ thấp:**

Một số vị trí có số lượng địa chỉ thấp, dưới 15 địa chỉ. Điều này có thể phản ánh khu vực có dân cư ít hoặc ít hoạt động thương mại. Các vị trí với số lượng địa chỉ thấp bao gồm **Mississippi**, **Kansas**, và **Rhode Island**.

* **Tính đồng đều trong phân phối địa chỉ:**

Biểu đồ cho thấy sự đồng đều tương đối trong số lượng địa chỉ giữa các vị trí, không có vị trí nào có số lượng chênh lệch quá lớn. Điều này cho thấy cửa hàng có sự hiện diện đồng đều trên nhiều khu vực khác nhau.

* **Kết luận:**

Phân bố địa chỉ đều và đa dạng trên các vị trí có thể là một chiến lược nhằm tiếp cận khách hàng ở nhiều khu vực. Các bang có mật độ dân số cao hoặc thành phố lớn thường có nhiều địa chỉ hơn, trong khi các khu vực ít dân cư thì số lượng địa chỉ thấp hơn.

### 

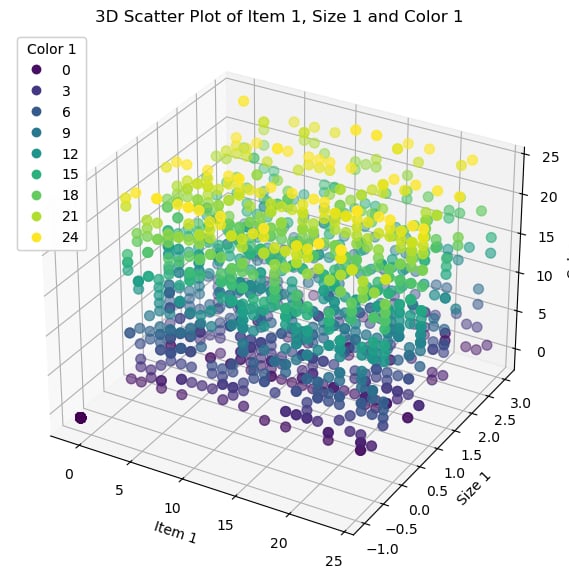
### 5.Biểu đồ phân tán giữa item, size, color:

* **Cụm 0:**



Các điểm dữ liệu phân bố tương đối đều trong không gian 3 chiều, cho thấy không có sự tập trung rõ ràng của dữ liệu ở một vùng cụ thể nào. Tuy nhiên, có thể thấy một số xu hướng phân bố nhất định:

* **Mối quan hệ giữa Item và Size:** Có vẻ như không có mối quan hệ tuyến tính rõ ràng giữa hai biến này. Các điểm dữ liệu phân bố khá ngẫu nhiên trên trục Item và Size.
* **Mối quan hệ giữa Color và các biến khác:** Màu sắc của các điểm dữ liệu thay đổi theo một gradient từ tím sang vàng, cho thấy có một mối liên hệ nào đó giữa màu sắc và các biến khác.
* **Cụm 1:**

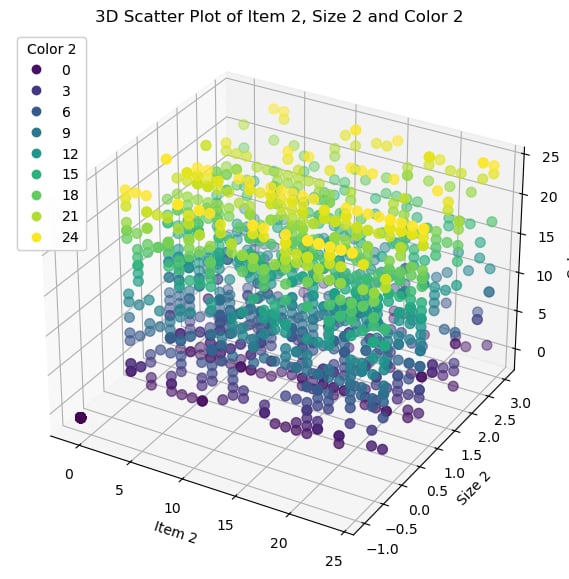


**Mối quan hệ giữa các biến:** Bằng cách quan sát sự phân bố của các điểm, chúng ta có thể nhận thấy các mối quan hệ giữa các biến :

+) Nếu các điểm có xu hướng tập trung thành các cụm riêng biệt, điều đó có thể cho thấy sự tồn tại của các nhóm dữ liệu khác nhau.

+) Nếu các điểm phân bố theo một mô hình nhất định (ví dụ: một đường thẳng hoặc một mặt phẳng), điều đó có thể chỉ ra một mối quan hệ tuyến tính hoặc phi tuyến tính giữa các biến.

* **Cụm 2:**



Biểu đồ 3D Scatter Plot này cung cấp một cách trực quan để thể hiện mối quan hệ giữa ba biến số: **Item 2**, **Size 2** và **Color 2**. Mỗi điểm trên biểu đồ đại diện cho một quan sát cụ thể, với vị trí của điểm xác định bởi giá trị của ba biến này.

**Mối quan hệ giữa các biến:**

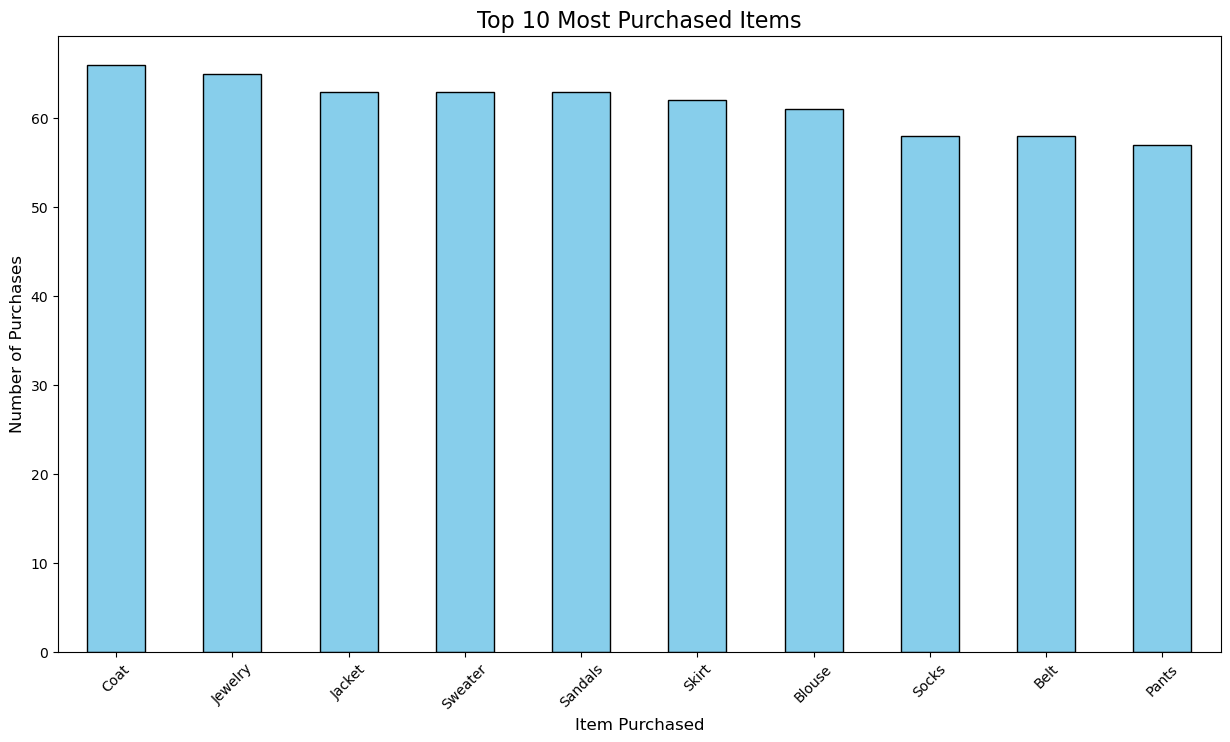
* **Tập trung của các điểm:** Nếu các điểm tập trung thành các cụm riêng biệt, điều đó cho thấy có thể có các nhóm dữ liệu khác nhau.
* **Hướng của các điểm:** Hướng của các điểm có thể cho thấy mối quan hệ giữa các biến. Ví dụ, nếu các điểm có xu hướng tăng theo cùng một hướng, điều đó có thể cho thấy một mối quan hệ dương.

Kết Luận :

* **Nhóm khách hàng:** Bằng cách quan sát các cụm điểm trên biểu đồ, chúng ta có thể xác định các nhóm khách hàng khác nhau dựa trên hành vi mua sắm của họ
* Chúng ta có thể thấy được mối quan hệ giữa giá cả, kích cỡ và tần suất mua hàng
* Từ đó ta có thể căn chỉnh số lượng sản xuất phù hợp của các mặt hàng, dựa theo size và màu sắc, đẩy mạnh sản xuất những sản phẩm có size và màu sắc có số lượng mua cao, căn chỉnh phù hợp những sản phẩm có size và màu sắc ít bán chạy hơn để không xảy ra việc bị tồn đọng hàng hóa

### 6.Biểu đồ top 10 mặt hàng bán chạy nhất:

* **Cụm 1:**



**Top 10 Most Purchased Items:** Tiêu đề cho biết biểu đồ này đang thể hiện 10 mặt hàng được mua nhiều nhất.

+) **Mặt hàng được mua nhiều nhất:** Áo khoác (Coat) là mặt hàng được khách hàng lựa chọn nhiều nhất.

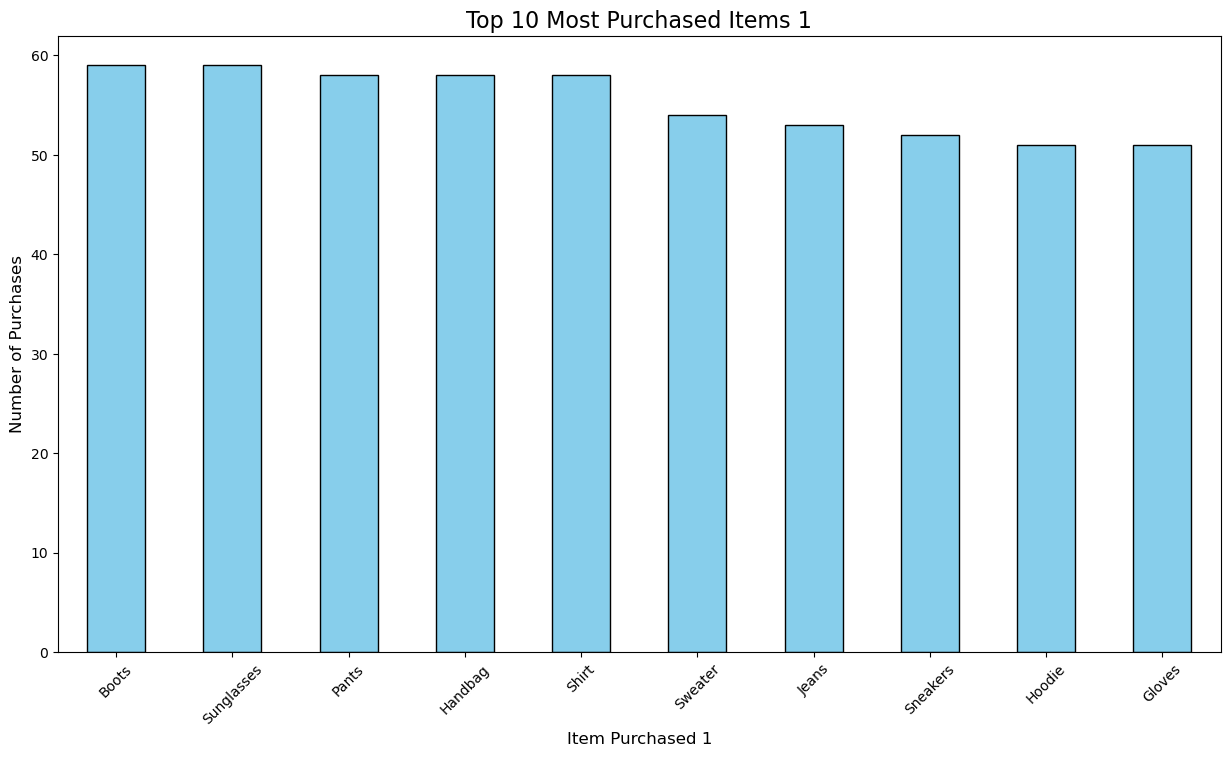
+) **Các mặt hàng phổ biến:** Ngoài áo khoác, đồ trang sức (Jewelry), áo khoác (Jacket), áo len (Sweater), dép (Sandals), váy (Skirt), áo sơ mi (Blouse), tất (Socks), thắt lưng (Belt) và quần (Pants) cũng là những mặt hàng được ưa chuộng.

=> Áo khoác Coat: Là mặt hàng bán chạy nhất, điều này có thể do thời tiết, xu hướng thời trang hoặc nhu cầu sử dụng của khách hàng, chúng ta nên đẩy mạnh sản xuất cũng như phát triển thêm các mẫu mã đa dạng của áo khoác, đồng thời đưa thêm nhiều chiến lược tiếp thị và quảng cáo đến với khách hàng để tăng doanh thu cho cửa hàng

Những mặt hàng khác như đồ trang sức (Jewelry), áo khoác (Jacket), áo len (Sweater), dép (Sandals), váy (Skirt), áo sơ mi (Blouse), tất (Socks), thắt lưng (Belt) và quần (Pants) cũng khá được ưa chuộng, chúng ta có thể sản xuất dàn trải theo số lượng hợp lí để có thể k quá dư thừa hàng tồn và có chiến lược marketing phù hợp để nâng cao số lượng bán

Việc đồ trang sức nằm trong top 10 cho thấy nhu cầu làm đẹp và trang sức bản thân của khách hàng khá lớn, chúng ta nên đẩy mạnh phát triển về mặt hàng này

* **Cụm 1:**



**Top 10 Most Purchased Items:** Tiêu đề cho biết biểu đồ này đang thể hiện 10 mặt hàng được mua nhiều nhất của cụm 1 :

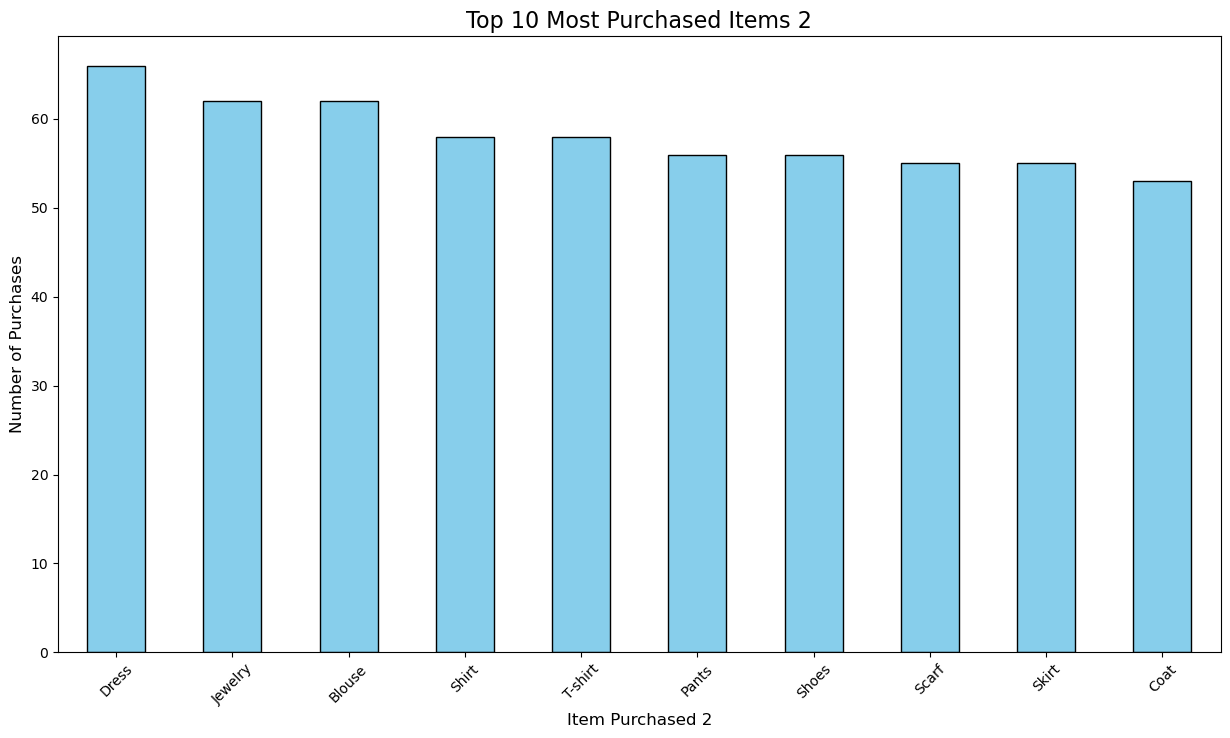
+) **Mặt hàng được mua nhiều nhất:** Giày bốt (Boots) là mặt hàng được khách hàng lựa chọn nhiều nhất.

+) Ngoài giày bốt, kính râm (Sunglasses), quần (Pants), túi xách (Handbag), áo sơ mi (Shirt), áo len (Sweater), quần jean (Jeans), giày thể thao (Sneakers), áo hoodie (Hoodie) và găng tay (Gloves) cũng là những mặt hàng được ưa chuộng.

=> Giày bốt là sản phầm giày có tính thời trang cao, việc khách hàng ưa chuộng sử dụng mua giày boots khiến chúng ta phải đẩy mạnh sản xuất chú trọng về chất lượng của từng sản phẩm điều đó sẽ giúp chúng ta có thể giữ chân được khách hàng cũ cũng như đưa sản phẩm tới nhiều khách hàng mới

Những sản phẩm khác như kính râm (Sunglasses), quần (Pants), túi xách (Handbag), áo sơ mi (Shirt), áo len (Sweater), quần jean (Jeans), giày thể thao (Sneakers), áo hoodie (Hoodie) và găng tay (Gloves) sẽ hướng tới những đối tượng chủ yếu là người trẻ ưa chuộng và yêu thích thời trang vì vậy chúng ta nên đẩy mạnh mục tiêu marketing để có thể đưa sản phẩm đến nhiều người cũng như đưa ra những chương trình khuyến mãi và ưu đãi giúp cho doanh số cao hơn

* **Cụm 2:**



**Top 10 Most Purchased Items:** Tiêu đề cho biết biểu đồ này đang thể hiện 10 mặt hàng được mua nhiều nhất của cụm 2 :

+) **Váy (Dress):** Là sản phẩm bán chạy nhất, cho thấy nhu cầu về trang phục nữ cao cấp hoặc dành cho các dịp đặc biệt.

**+) Trang sức (Jewelry):** Việc trang sức nằm trong top đầu cho thấy xu hướng làm đẹp và trang sức bản thân của khách hàng khá lớn.

**+) Áo sơ mi (Blouse), Áo thun (T-shirt):** Đây là những sản phẩm cơ bản trong tủ đồ của nhiều người, việc chúng nằm trong top 10 cho thấy nhu cầu về trang phục hàng ngày là rất lớn.

**+) Quần (Pants), Giày (Shoes):** Là những item không thể thiếu trong bất kỳ bộ trang phục nào, việc chúng có mặt trong top 10 cho thấy nhu cầu về các sản phẩm này là ổn định.

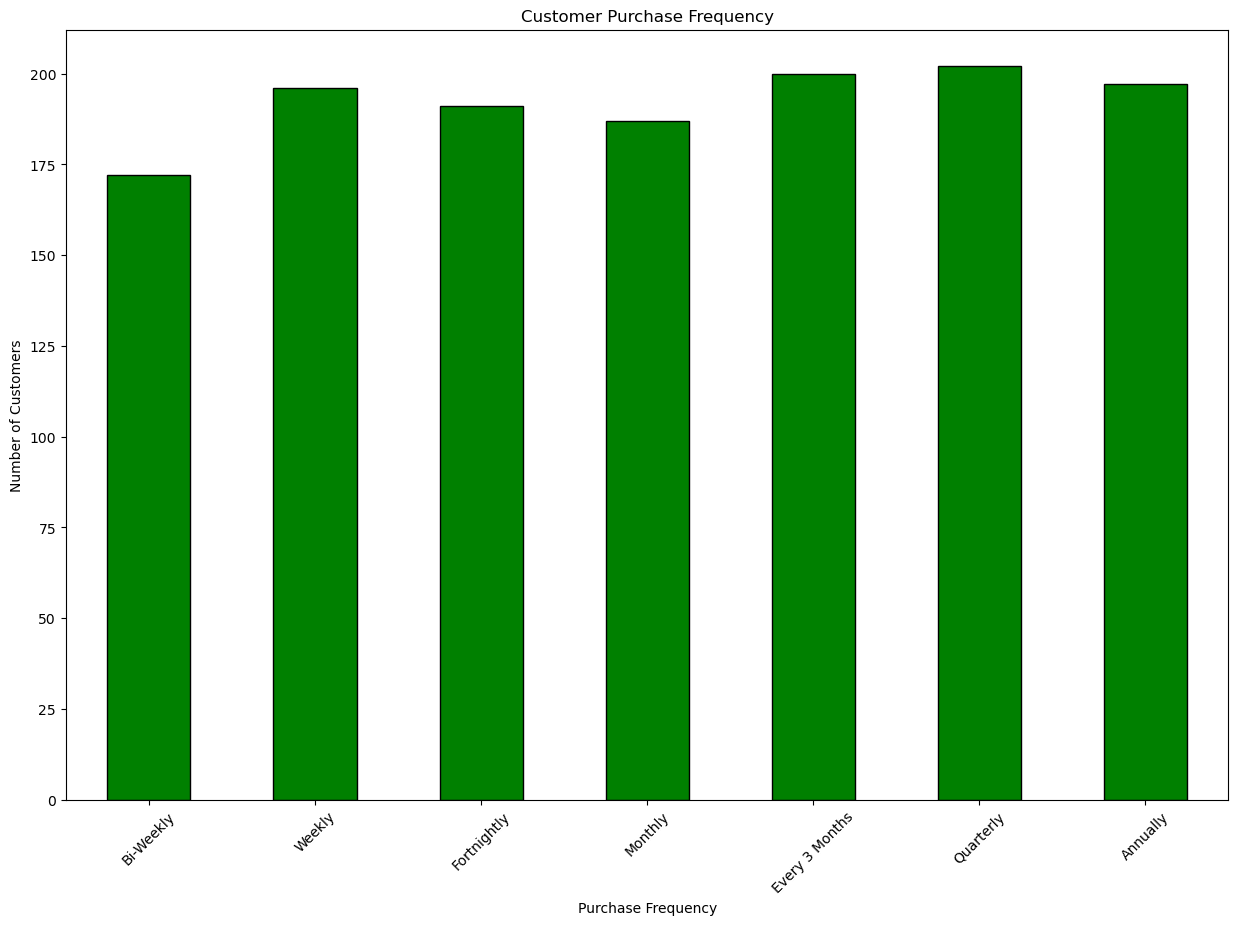
**+) Khăn quàng cổ (Scarf), Váy (Skirt), Áo khoác (Coat):** Những sản phẩm này thường được mua theo mùa hoặc để phối đồ, việc chúng có mặt trong top 10 cho thấy sự đa dạng trong sở thích mua sắm của khách hàng.

=> Chúng ta có thể thấy việc sản phẩm Váy bán chạy nhất cho thấy đa phần khách hàng của nhóm này là khách hàng nữ vì vậy nên đẩy mạnh thêm các sản phẩm đi kèm như phụ kiện hoặc quần áo váy, trang sức những món được ưa chuộng bởi phái nữ

Các sản phẩm khác như Áo sơ mi (Blouse), Áo thun (T-shirt), Quần (Pants), Giày (Shoes), Khăn quàng cổ (Scarf), Áo khoác (Coat) là những sản phẩm thường dc mua để phối đồ và kết hợp vì vậy cần có chiến lược và kế hoạch sản xuất hợp lí để có thể tăng doanh thu hợp lí

### 7.Biểu đồ cột tần suất mua hàng:

* **Cụm 0:**



* **Tần suất mua hàng phổ biến:**

Biểu đồ cho thấy số lượng khách hàng có sự phân bố khá đồng đều giữa các mức tần suất mua hàng khác nhau. Tất cả các tần suất đều có số lượng khách hàng dao động quanh mức 180 đến 200.

* **Các nhóm tần suất mua hàng cao:**

Nhóm khách hàng mua hàng hàng tuần và mỗi 3 tháng là những nhóm có số lượng khách hàng cao nhất, gần với 200 khách hàng. Điều này cho thấy các mức tần suất này có thể là lựa chọn phổ biến cho một số lượng lớn khách hàng, phù hợp với nhu cầu thường xuyên hoặc định kỳ của họ.

* **Các nhóm tần suất mua hàng thấp hơn:**

Nhóm khách hàng mua hàng hai tuần một lần (bi-weekly) có số lượng khách hàng thấp hơn một chút so với các nhóm còn lại, nhưng không có sự chênh lệch đáng kể.

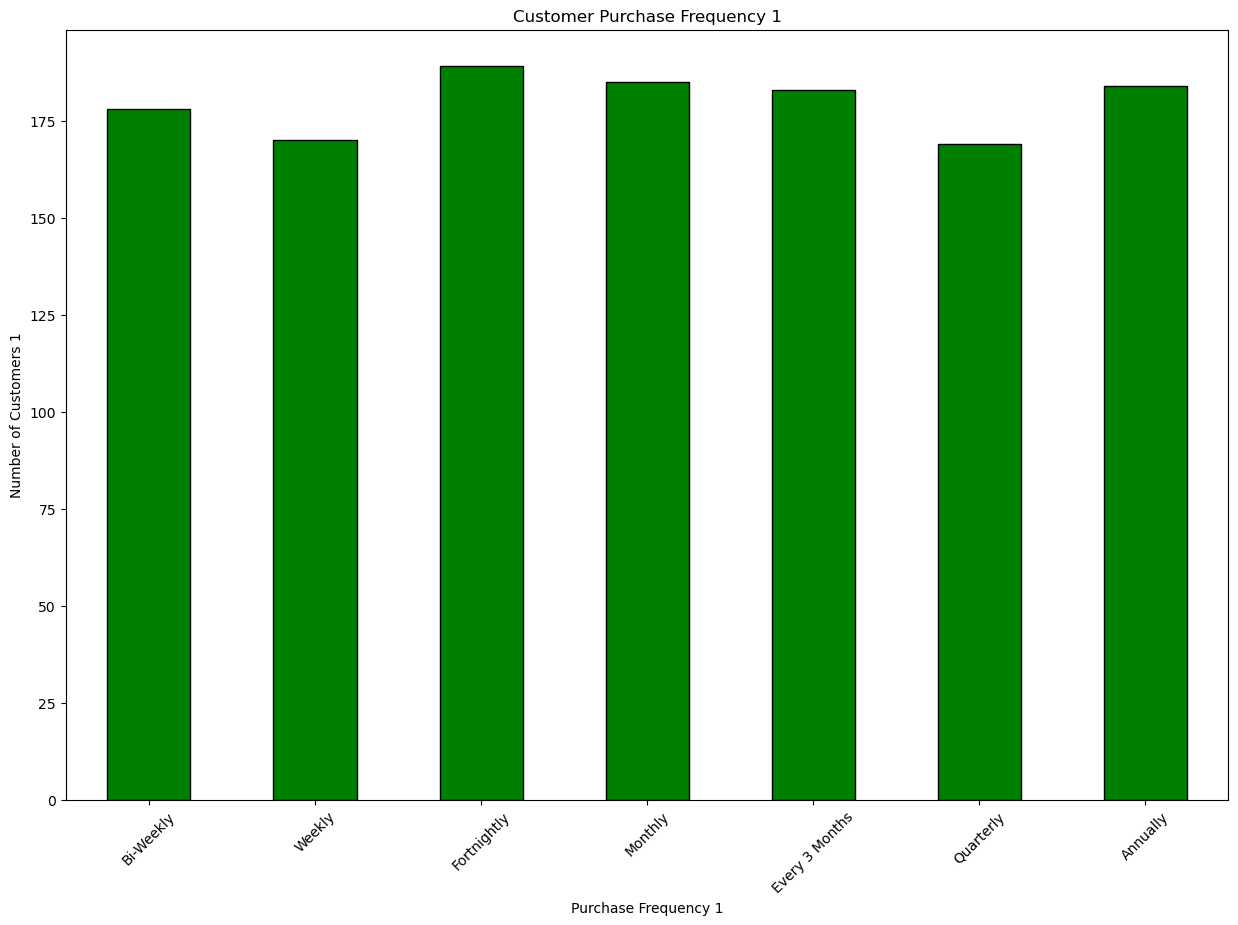
* **Sự đồng đều trong tần suất mua hàng:**

Biểu đồ cho thấy sự đồng đều tương đối trong số lượng khách hàng giữa các nhóm tần suất, không có nhóm nào quá chênh lệch. Điều này có thể phản ánh nhu cầu đa dạng của khách hàng, khi mỗi nhóm có sở thích và tần suất mua sắm khác nhau nhưng vẫn giữ một mức ổn định.

* **Kết luận:**

Tần suất mua hàng của khách hàng được phân bố khá đồng đều, cho thấy cửa hàng có khả năng phục vụ nhiều nhu cầu khác nhau, từ những khách hàng mua hàng thường xuyên đến những người mua theo tần suất lâu dài hơn. Điều này có thể là một lợi thế trong việc tối ưu hóa chiến lược bán hàng để đáp ứng đa dạng nhu cầu của khách hàng.

* **Cụm 1:**



* **Tần suất mua hàng phổ biến:**

Biểu đồ thể hiện sự phân bố khá đồng đều giữa các mức tần suất mua hàng khác nhau. Các mức tần suất đều có số lượng khách hàng dao động trong khoảng 180 đến 200, cho thấy không có mức nào vượt trội hay quá thấp.

* **Các nhóm tần suất mua hàng cao:**

Nhóm khách hàng mua hàng hàng tuần và hàng tháng có số lượng cao nhất, gần đạt mức 200 khách hàng. Điều này có thể phản ánh rằng các tần suất này là lựa chọn phổ biến cho nhu cầu mua hàng đều đặn của một bộ phận lớn khách hàng.

* **Các nhóm tần suất mua hàng thấp hơn:**

Nhóm khách hàng mua hàng hai tuần một lần (bi-weekly) có số lượng khách hàng thấp hơn một chút so với các nhóm còn lại, nhưng không có sự chênh lệch đáng kể, cho thấy sự đa dạng trong lựa chọn tần suất mua hàng của khách hàng.

* **Sự đồng đều trong tần suất mua hàng:**

Biểu đồ phản ánh sự đồng đều tương đối về số lượng khách hàng giữa các mức tần suất khác nhau, không có nhóm nào quá chênh lệch. Điều này cho thấy rằng nhu cầu của khách hàng rất đa dạng, từ nhu cầu mua hàng thường xuyên đến mua hàng theo định kỳ dài hơn.

* **Kết luận:**

Tần suất mua hàng của khách hàng được phân bổ đồng đều giữa các mức khác nhau, cho thấy cửa hàng có khả năng đáp ứng đa dạng nhu cầu của khách hàng. Việc phục vụ nhiều tần suất mua hàng có thể là một lợi thế giúp cửa hàng tối ưu hóa chiến lược bán hàng, đáp ứng từ nhu cầu thường xuyên đến nhu cầu định kỳ của khách hàng.

* **Cụm 2:**

### 

* **Tần suất mua hàng phổ biến:**

Số lượng khách hàng phân bố tương đối đồng đều giữa các mức tần suất, với một số mức như "Bi-Weekly" và "Every 3 Months" có số lượng khách hàng cao nhất, gần chạm mức 200.

* **Các nhóm tần suất mua hàng cao:**

**Bi-Weekly** và **Every 3 Months** là hai mức tần suất phổ biến nhất, với số lượng khách hàng cao nhất, gần 200. Điều này cho thấy các mức này đáp ứng nhu cầu mua hàng định kỳ của một lượng lớn khách hàng.

* **Các nhóm tần suất mua hàng thấp hơn:**

**Fortnightly** có số lượng khách hàng thấp hơn so với các nhóm còn lại, khoảng dưới 150 khách hàng. Điều này có thể gợi ý rằng tần suất mua hàng này không phải là lựa chọn phổ biến đối với phần lớn khách hàng.

* **Sự đồng đều trong tần suất mua hàng:**

Số lượng khách hàng tương đối ổn định giữa các nhóm tần suất khác nhau. Điều này cho thấy khách hàng có nhu cầu mua sắm đa dạng, không bị tập trung vào một vài nhóm cụ thể, và mỗi nhóm đều duy trì được một số lượng khách hàng ổn định.

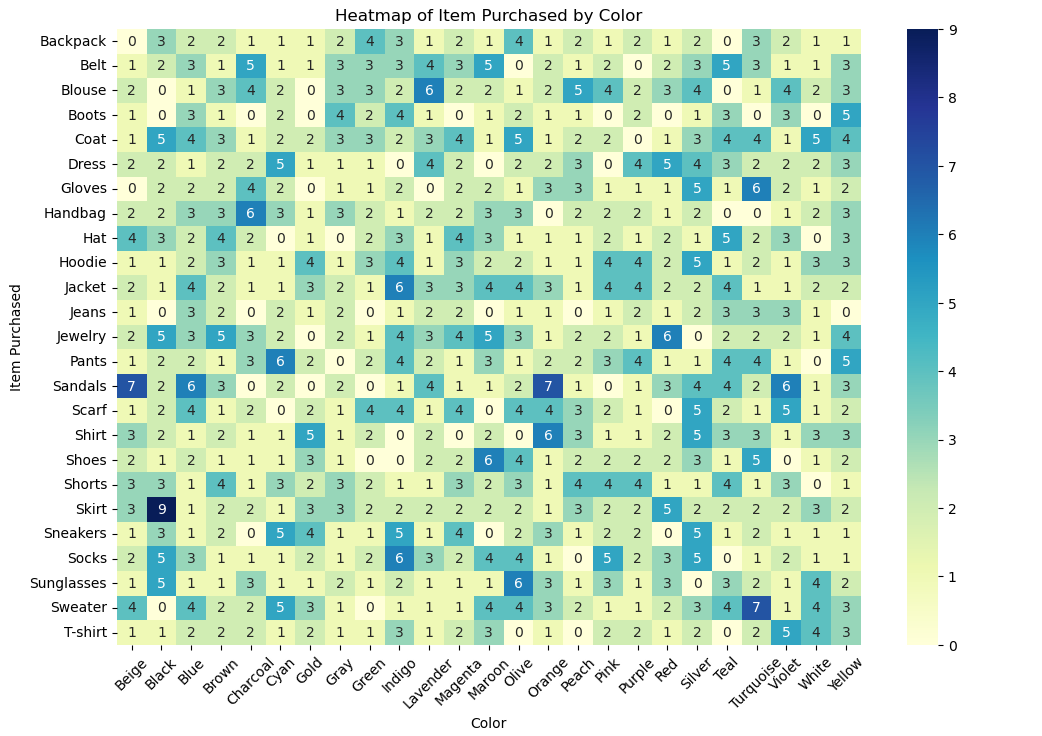
* **Kết luận:**

Tần suất mua hàng của khách hàng khá đồng đều, với một số mức phổ biến như "Bi-Weekly" và "Every 3 Months". Điều này phản ánh rằng cửa hàng có thể phục vụ tốt các nhu cầu đa dạng của khách hàng, từ mua hàng thường xuyên đến mua hàng định kỳ dài hơn.

=>Biểu đồ cho thấy số lượng khách quay lại cửa hàng ở mức cao và đồng đều trong mỗi chu kỳ. Tuy nhiên số lượng khách mua hàng ở chu kỳ “Annually”còn cao vì thế cần cải thiện chất lượng dịch vụ, chăm sóc khách hàng.

### 8.Biểu đồ Heatmap thể hiện số lượng sản phẩm và màu sắc của chúng:

* **Cụm 0:**

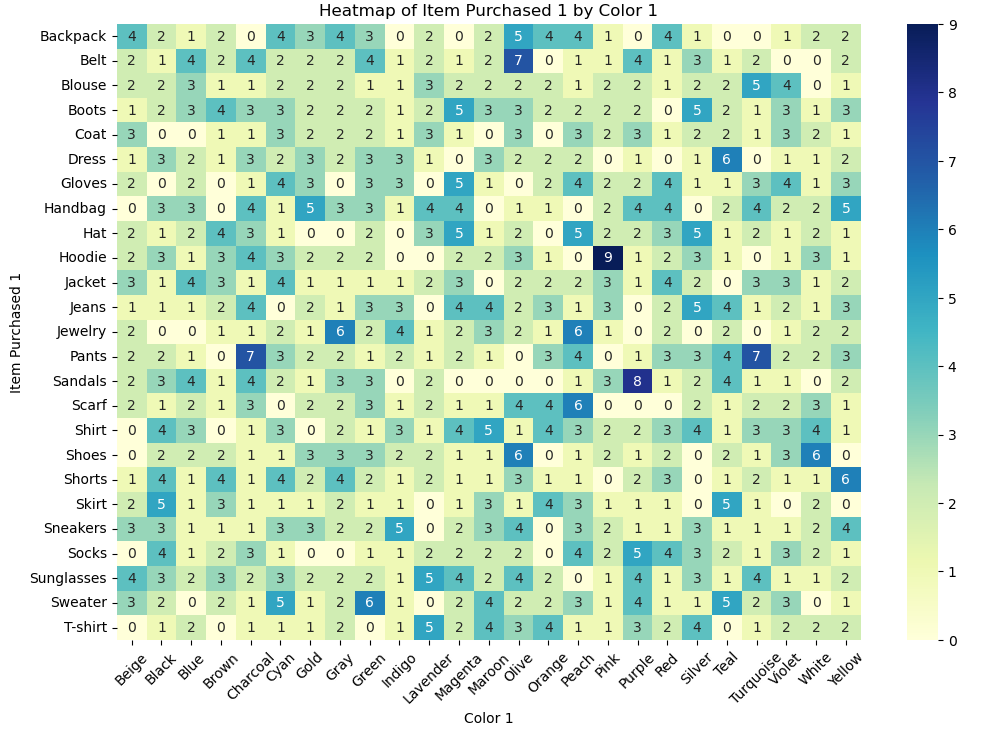


Các sản phẩm và màu sắc của chúng phổ biển:

* **Shirt (áo sơ mi)**: Có tần suất mua cao ở các màu như **Blue**, **White**, và **Black**, có thể là do đây là các màu phổ biến cho áo sơ mi, thường được dùng trong nhiều dịp khác nhau.
* **Pants (quần)**: Các màu **Black**, **Gray**, và **Navy** cũng có tần suất cao, phù hợp với nhu cầu thông thường của khách hàng cho trang phục hàng ngày hoặc công sở.
* **Sneakers (giày thể thao)**: Có màu sắc đa dạng, với tần suất cao ở các màu **White** và **Black**, điều này phù hợp với xu hướng giày thể thao màu cơ bản dễ phối.

Các sản phẩm và màu sắc của chúng ít phổ biến:

* **Áo sơ mi (Shirt):** Ít được mua với màu xanh lá cây, xanh lam nhạt (indigo).
* **Quần short (Shorts):** Ít được mua với màu tím (purple).
* **Váy (Skirt):** Ít được mua với màu vàng.
* **Áo len (Sweater):** Ít được mua với màu hồng (pink).
* **Áo thun (T-shirt):** Ít được mua với màu cam (orange).
* **Giày thể thao (Sneakers):** Ít được mua với màu trắng.
* **Găng tay (Gloves):** Ít được mua với nhiều màu sắc khác nhau, ngoại trừ đen.
* **Cụm 1:**

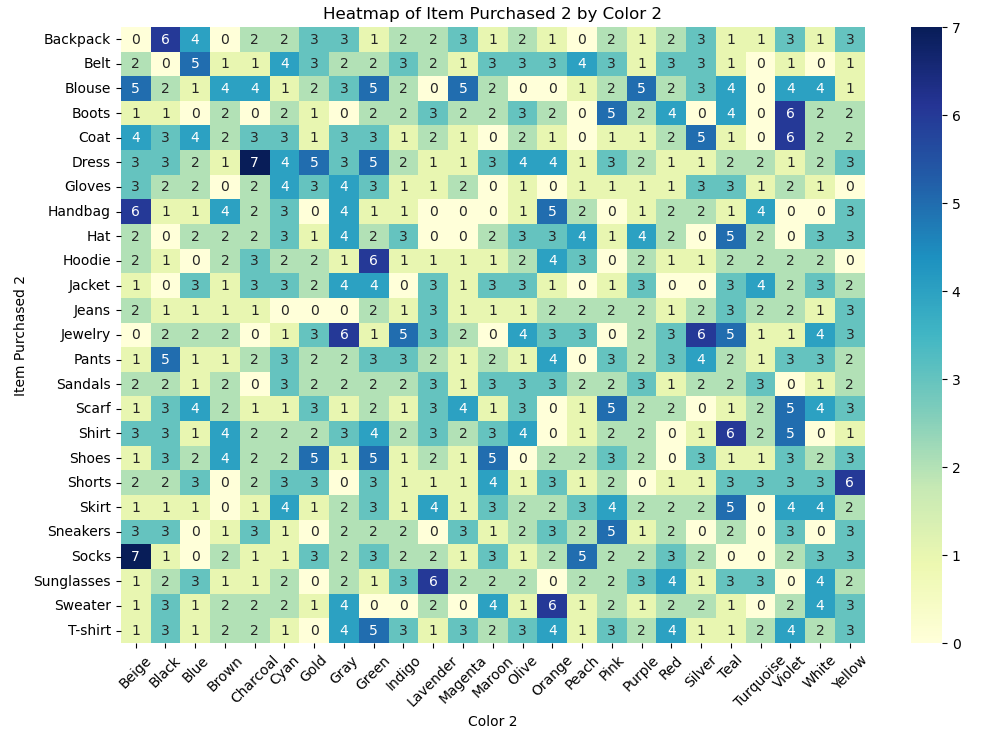


Các sản phẩm và màu sắc của chúng phổ biển:

* **Backpack (Ba lô)**: Các màu phổ biến nhất là **Navy**, **Blue**, và **Charcoal**. Điều này cho thấy khách hàng thích mua ba lô với các màu sắc trung tính hoặc tối màu, có lẽ vì chúng dễ phối hợp và phù hợp với nhiều hoàn cảnh.
* **Shirt (Áo sơ mi)**: Sơ mi được mua nhiều nhất ở màu **White**, **Blue**, và **Gray**. Đây đều là các màu sắc phổ biến cho áo sơ mi, dễ mặc trong các dịp khác nhau, từ công việc cho đến các dịp không chính thức.
* **Sneakers (Giày thể thao)**: Các màu **White** và **Black** là lựa chọn hàng đầu, phản ánh xu hướng chọn các màu cơ bản dễ phối đồ trong các dòng giày thể thao.

Các sản phẩm và màu sắc của chúng ít phổ biến:

* **Scarf (Khăn quàng)**: Ít được mua ở các màu sắc sáng như **Yellow** và **Purple**, nhưng lại được chọn với các màu trung tính như **Gray** và **Navy**. Điều này có thể do sở thích sử dụng khăn quàng với các màu trầm dễ kết hợp vào mùa lạnh.
* **Jewelry (Trang sức)**: Các màu sắc được chọn cho trang sức khá đa dạng, nhưng không tập trung nhiều vào bất kỳ màu nào cụ thể, có lẽ vì trang sức thường được phối theo phong cách cá nhân hơn là xu hướng màu sắc chung.
* **Hat (Mũ)** và **Gloves (Găng tay)**: Có ít sự xuất hiện của các màu sặc sỡ như **Pink**, **Orange**, và **Yellow**, cho thấy các phụ kiện này có xu hướng được ưa chuộng hơn ở các màu sắc trung tính hoặc tối.
* **Cụm 2:**



Các sản phẩm và màu sắc của chúng phổ biển:

* **Áo khoác (Coat):** Được mua với nhiều màu sắc khác nhau, đặc biệt là màu xanh dương đậm (blue).
* **Váy (Dress):** Được mua với nhiều màu sắc, cho thấy sự đa dạng trong lựa chọn của khách hàng.
* **Quần (Pants):** Cũng được mua với nhiều màu sắc khác nhau.

Các sản phẩm và màu sắc của chúng ít phổ biến:

* **Áo sơ mi (Shirt):** Ít được mua với màu xanh lá cây, xanh lam nhạt (indigo).
* **Quần short (Shorts):** Ít được mua với màu tím (purple).
* **Váy (Skirt):** Ít được mua với màu vàng.
* **Áo len (Sweater):** Ít được mua với màu hồng (pink).
* **Áo thun (T-shirt):** Ít được mua với màu cam (orange).
* **Giày thể thao (Sneakers):** Ít được mua với màu trắng.
* **Găng tay (Gloves):** Ít được mua với nhiều màu sắc khác nhau, ngoại trừ màu đen.

=> Thông qua biểu đồ, chúng ta có thể dễ dàng nắm bắt được xu hướng mua sắm của khách hàng về màu sắc, từ đó tối ưu hóa việc quản lý hàng tồn kho và lên kế hoạch sản xuất.

## IV. Tổng kết

Sau khi phân tích bằng các biểu đồ chúng em thấy được đặc trưng của từng cụm như sau:

* **Cụm 0:**
  + Là lượng khách hàng có chi tiêu ở mức thấp khoảng trung bình là 32,83 $, nhóm khách hàng này chủ yếu là những người có độ tuổi trung bình và cao tuổi với độ tuổi từ 40 – 60 tuổi, họ chủ yếu mua những mặt hàng có tính linh hoạt và độ bền cao như Blouse, Boots hay Sweater,…
  + Họ mua với tuần suất trung bình rải rác theo từng tuần từng tháng ở mức đều nhau và chủ yếu sẽ tập trung mua theo hàng tuần và đa phần nhóm khách hàng này có giới tính là nam

* **Cụm 1:**
  + Là lượng khách hàng có chi tiêu ở mức trung bình cao với mức trung bình là 60,28 $, nhóm khách hàng này chủ yếu là người có độ tuổi phân bố đồng đều dao động từ 20 – 70 tuổi.
  + Chiếm số lượng lớn ở độ tuổi 40 – 60 cho thấy khách hàng chủ yếu ở cụm này sẽ nằm trong độ tuổi trung niên, với những mặt hàng chủ yếu được mua là Giày bốt (Boots) với lượt mua cao nhất thể hiện tính ưa thích thời trang của nhóm khách hang này.
  + Những sản phẩm như kính râm (Sunglasses), quần (Pants), túi xách (Handbag), áo sơ mi (Shirt), áo len (Sweater), quần jean (Jeans), giày thể thao (Sneakers), áo hoodie (Hoodie) và găng tay (Gloves) cũng được ưa chuộng bởi tính linh hoạt và dễ dàng sử dụng vào đời thường.
  + Họ mua hàng với tuần suất trung bình thấp hơn so với Cụm 0 nên chúng ta cần có xu hướng kích cầu mua sắm và gia tang mức độ tiêu dùng của khách hàng, đa phần khách hàng của Cụm này sẽ thuộc giới tính nam.
* Vì vậy nên căn chỉnh đẩy mạnh những mặt hàng có xu hướng tập trung về chất lượng và độ bền cao, ngoài ra cũng phát triển thêm những mặt hàng của phái nữ để gia tang doanh số.

* **Cụm 2:**
  + Là khách hàng có chi tiêu ở mức cao với mức tiêu là 87,28 $, nhóm khách hàng này có độ tuổi phân bố khoảng 20 đến 70 tuổi.
  + Cũng như Cụm 1 sẽ chiếm phần lớn độ tuổi 40 -60, chủ yếu những sản phẩm được ưa chuộng là váy cũng như trang sức và phụ kiện là chủ yếu điều đó cho ta thấy được số lượng mua của tệp khách hàng nữ sẽ chiếm nhiều ở cụm này so với 2 cụm khác.
  + Số lượng mua theo hàng tuần và hàng tháng cũng ở mức cao cho ta thấy nhu cầu của họ trong việc làm đẹp là khá lớn.
* Vì lẽ đó chúng ta nên đẩy mạnh sản xuất them nhiều mẫu mã đa dạng, cũng như đẩy mạnh và đâu tư vào truyền thông hình ảnh và marketing để nâng cao doanh số.

## V. Phương pháp cải thiện

* Nhìn chung chúng ta có thể thấy được doanh số của cửa hàng vẫn ở mức ổn và đều đặn, các sản phẩm đều được bán chạy và không bị xảy ra quá nhiều tình trạng dư đọng hàng tồn và thua lỗ
* Để có thể nâng cao doanh thu một cách tốt hơn, chúng ta cần đưa ra một số phương pháp sau khi đã phân tích hành vi tiêu dùng của khách hàng :

+) Ở khâu sản xuất và nhập hàng, chúng ta nên chủ yếu nhập số lượng nhiều các sản phẩm chủ đạo và bán chạy điển hình như Blouse hay Boots và Skirt, những sản phẩm này sẽ đóng góp chính vào lợi nhuận của cửa hàng.

+) Đồng thời phải liên tục nghiên cứu và cải thiện chất lượng sản phẩm cũng như phát triển them về mẫu mã đa dạng để có thể cạnh tranh và thu hút khách hàng.

+) Với những sản phẩm có lượng bán ít hơn, chúng ta nên hạn chế sản xuất ồ ạt mà hãy sản xuất những sản phẩm tập trung vào chất lượng và kiểu dáng thực sự được ưa thích.

+) Cuối cùng là lược bỏ những sản phẩm có số lượng bán quá thấp và chưa mang lại nhiều lợi nhuận cho cửa hàng để tránh lỗ.

* Ở khâu dịch vụ:

+) Chúng ta phải biết trong kinh doanh, thái độ và dịch vụ chính là thứ vô cùng quan trọng, điều sẽ khiến cho tỷ lệ quay lại của khách hàng ở mức cao hay thấp, điều đó sẽ ảnh hưởng trực tiếp đến doanh thu của cửa hàng.

+) Cần có chiến lược đào tạo cách cư xử, cách bán hàng sao cho thật chuyên nghiệp ở cả 2 mảng online và offline.

+) Tối ưu hóa sự thuận tiện cho khách hàng từ khâu vận chuyển hay thanh toán và cả khâu bảo hành đối với những sản phẩm lỗi.

* Ở khâu marketing và truyền thông:

+) Chúng ta nên làm tốt nhất ở khâu hình ảnh, tập trung xây dựng thương hiệu trên 3 kênh chính là FaceBook, Instagram và TikTok.

+) Tuyển dụng thêm nhiều nhân viên giỏi có kĩ năng chuyên môn để nâng cao lượt tương tác của cửa hàng.

+) Thực hiện các chiến dịch quảng cáo và chiến dịch khuyến mại để kích cầu tối đa nhu cầu mua sắm của khách hàng, đồng thời trên 2 kênh Shoppe và Tiktokshop, cần chú trọng giải quyết về phần đánh giá đặc biệt là những đánh giá xấu.

**KẾT LUẬN**

Sau khi hoàn thành đề tài “**Phân cụm khách hàng trong shop thời trang bằng phương pháp Kmeans”,** nhóm em đã đạt được kết quả sau:

* Tìm hiểu tổng quan về quản lý dữ liệu lớn, phương pháp Kmeans để phân cụm khách hàng
* Thông qua quá trình thực hiện đề tài nhóm đã học tập và có cơ hội làm quen với framework Hadoop, nâng cao kỹ năng lập trình, xử lý và trực quan hóa dữ liệu
* Nhóm chúng em đã được trau dồi thêm kiến thức và kỹ năng về các phương pháp thu thập, lưu trữ, xử lý và phân tích dữ liệu lớn

Trong quá trình hoàn thành đề tài, dù nhóm em đã cố gắng tìm hiểu và thực hiện bài toán nhưng vẫn không thể tránh những thiếu sót, khiếm khuyết. Do vậy, rất mong nhận được sự đóng góp ý kiến của thầy cô để chúng em rút kinh nghiệm, cải thiện kỹ năng và tích lũy kiến thức trong môn học.