**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----- ������ -----**

****

**BÀI TẬP LỚN MÔN KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: PHÂN KHÚC KHÁCH HÀNG CỦA TRUNG TÂM MUA SẮM BẰNG PHƯƠNG PHÁP PHÂN CỤM K-MEANS**

**Giảng viên hướng dẫn: TS.TRẦN MẠNH TUẤN**

**Nhóm sinh viên thực hiện: Nhóm 2**

Thành viên nhóm:

1. Hoàng Ánh Nguyệt - 2151062839

2. Trương Thị Quỳnh Anh - 2151062713

3. Sái Thị Lan Hương - 2151060207

4. Nguyễn Thị Thu Thủy - 2151060256

**Hà Nội, ngày 20 tháng 10 năm 2024**

**MỤC LỤC**

**[I: Tổng quan về khai phá dữ liệu](#_Toc180311686)** [4](#_Toc180311686)

**[1.1. Khai phá tri thức và khai phá dữ liệu](#_Toc180311687)** [4](#_Toc180311687)

**[1.2. Quá trình khai phá tri thức](#_Toc180311688)** [4](#_Toc180311688)

**[1.3. Mô tả bài toán](#_Toc180311689)** [4](#_Toc180311689)

**[1.3.1. Tổng quan về bài toán](#_Toc180311690)** [4](#_Toc180311690)

**[II: Quy trình khai phá dữ liệu](#_Toc180311691)** [7](#_Toc180311691)

**[2.1. Tiền xử lý dữ liệu](#_Toc180311692)** [7](#_Toc180311692)

**[2.2. Làm sạch dữ liệu](#_Toc180311693)** [8](#_Toc180311693)

**[2.3. Tích hợp dữ liệu](#_Toc180311694)** [11](#_Toc180311694)

**[3.1.](#_Toc180311695)****[Khai phá dữ liệu bằng mô hình phân cụm](#_Toc180311695)** [11](#_Toc180311695)

**[3.1.1.](#_Toc180311696)****[Giới thiệu về phân cụm:](#_Toc180311696)** [11](#_Toc180311696)

**[3.1.2.](#_Toc180311697)****[Thuật toán phân cụm K-means:](#_Toc180311697)** [12](#_Toc180311697)

**BẢNG PHÂN CHIA CÔNG VIỆC**

|  |  |
| --- | --- |
| **Tên thành viên** | **Công việc** |
| Hoàng Ánh Nguyệt | Tìm hiểu thuật toán, chuẩn bị dữ liệu, code thuật toán, viết báo cáo. |
| Trương Thị Quỳnh Anh | Tìm hiểu thuật toán, chuẩn bị dữ liệu, code thuật toán, viết báo cáo. |
| Sái Thị Lan Hương | Tìm hiểu thuật toán, chuẩn bị dữ liệu, tiền xử lý dữ liệu, viết báo cáo. |
| Nguyễn Thị Thu Thủy | Tìm hiểu thuật toán, chuẩn bị dữ liệu, tiền xử lý dữ liệu, viết báo cáo. |

**MỞ ĐẦU**

Thời đại công nghệ 4.0 phát triển với tốc độ chóng mặt, các nguồn thông tin dữ liệu ngày càng đa dạng, phong phú và đặc biệt đang lớn lên rất nhiều. Vì vậy khai phá dữ liệu để xử lý các thông tin, chọn lọc thông tin là nhu cầu cần thiết, tăng năng suất và chất lượng cao trong quản lý với các lĩnh vực khác nhau như y tế, giao thông, giáo dục, …

Trong lĩnh vực mua sắm, việc hiểu rõ khách hàng, phân khúc, mua sắm của mỗi người có vai trò vô cùng quan trọng. Điều này có thể giúp các doanh nghiệp đưa ra các phương án marketing phù hợp với từng phân khúc khách hàng khác nhau đem lại lợi nhuận cao. Do đó, khai phá dữ liệu đã trở thành một công cụ mạnh mẽ để nghiên cứu và phân tích thông tin có ích từ dữ liệu về khách hàng.

Việc áp dụng kỹ thuật khai phá dữ liệu vào việc phân khúc khách hàng giúp tạo nên mô hình kinh doanh hiệu quả nhất. Vì thế, nhóm chúng em đã lựa chọn đề tài "Phân khúc khách hàng của trung tâm mua sắm bằng phương pháp phân cụm K-means"để làm báo cáo cho môn học này.

Trong quá trình thực hiện bài tập lớn, chúng em đã cố gắng tìm hiểu và tham khảo những tài liệu liên quan. Nhưng do thời gian và trình độ có hạn nên không tránh khỏi những hạn chế và thiếu sót. Chúng em rất mong nhận được các nhận xét và góp ý của thầy để hoàn thiện hơn kết quả thực nghiệm.

# **I: Tổng quan về khai phá dữ liệu**

## **1.1. Khai phá tri thức và khai phá dữ liệu**

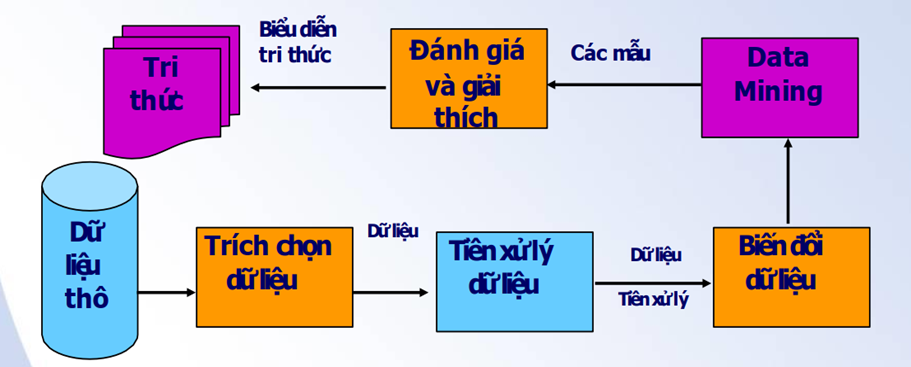
● Khai phá dữ liệu là một lĩnh vực nhằm tự động khai thác những thông tin tri thức đang tiềm ẩn trong dữ liệu và là bước chính của khai phá dữ liệu.

● Khai phá tri thức trong cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc mô hình trong dữ liệu và là mục tiêu chính của khai phá dữ liệu.

# **1.2. Quá trình khai phá tri thức**

● Khai phá dữ liệu là một lĩnh vực nhằm tự động khai thác những thông tin tri thức đang tiềm ẩn trong dữ liệu và là bước chính của khai phá dữ liệu.

● Khai phá tri thức trong cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc mô hình trong dữ liệu và là mục tiêu chính của khai phá dữ liệu.

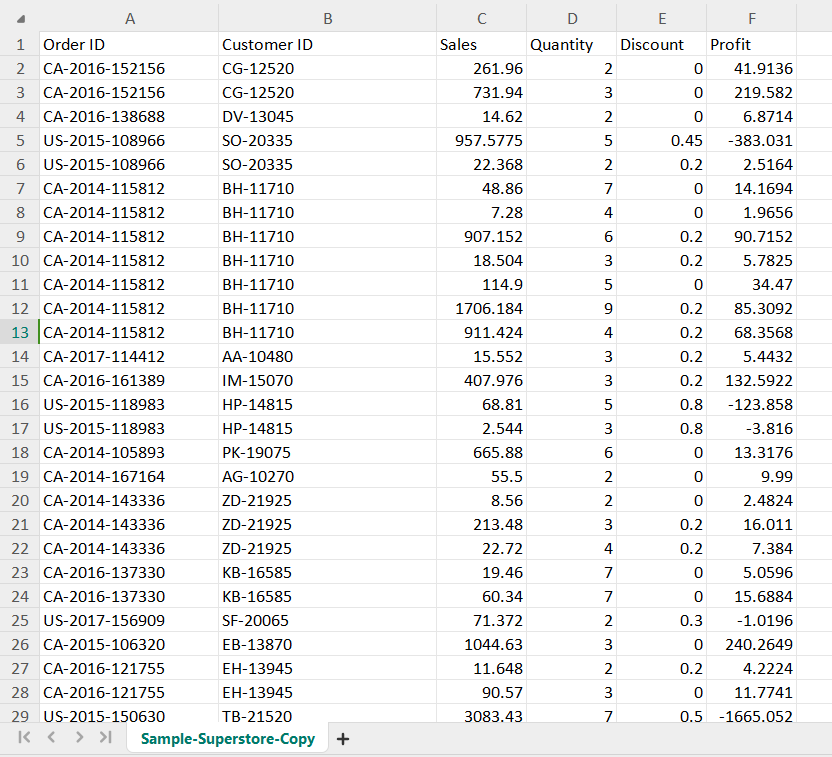
****

**Hình 1.2.1 Quá trình khai phá tri thức**

## **1.3. Mô tả bài toán**

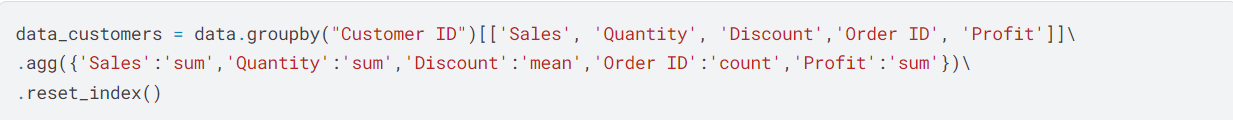
### **1.3.1. Tổng quan về bài toán**

* Tên bài toán: Phân khúc khách hàng của trung tâm mua sắm bằng phương pháp phân cụm K-means.
* Input:
  + Order ID: ID đơn hàng.
  + Customer ID: ID khách hàng.
  + Sales: Doanh thu bán hàng.
  + Quantity: Số lượng sản phẩm.
  + Discount: Phần trăm giảm giá.
  + Profit: Lợi nhuận.
* Output: Các cluster center.
* Dữ liệu bao gồm 9994 mẫu và 6 thuộc tính.

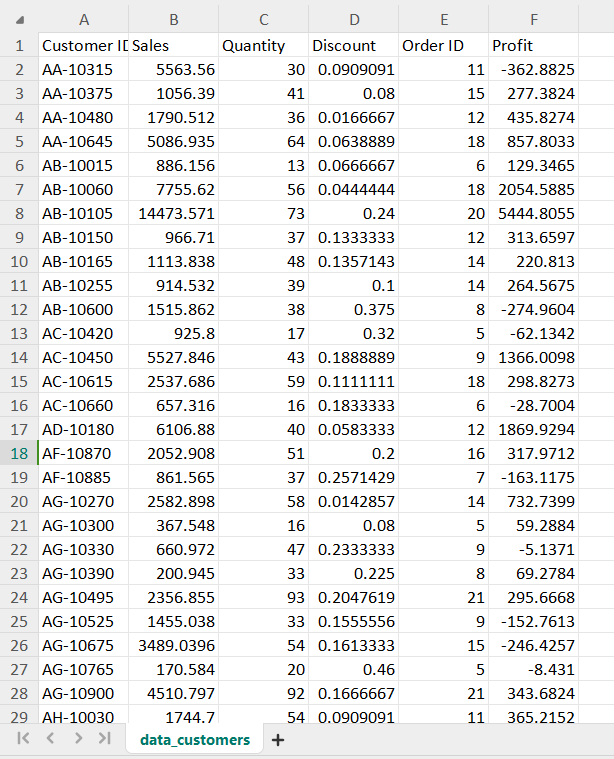


**Hình 1.3.1. Dữ liệu ban đầu**

* Để chuẩn bị cho quá trình phân cụm khách hàng, dữ liệu cần được gom nhóm theo Customer ID. Việc này giúp tổng hợp số đơn hàng mà mỗi khách hàng đã thực hiện mua sắm, từ đó có các thông tin chính xác về hành vi mua sắm của từng khách hàng. Cụ thể nhóm chúng em đã tiến hành các bước sau:
  + Nhóm dữ liệu theo **Customer ID:** các dòng có cùng Customer ID sẽ được gom thành một nhóm.
  + **Sales**: Sử dụng hàm sum() để tổng hợp các giá trị doanh số cho từng khách hàng, nhằm tính tổng doanh thu mà mỗi khách hàng đã đóng góp.
  + **Quantity**: Sử dụng hàm sum() để tổng hợp số lượng sản phẩm mà từng khách hàng đã mua, từ đó xác định tổng số lượng hàng hóa đã được bán cho họ.
  + **Profit**: Sử dụng hàm sum() để tính tổng lợi nhuận mà mỗi khách hàng đã mang lại, giúp đánh giá hiệu quả tài chính từ từng khách hàng.
  + **Discount**: Tính giá trị trung bình của chiết khấu mà mỗi khách hàng đã nhận được, giúp phân tích mức độ ưu đãi dành cho từng khách hàng.
  + **Order ID**: Đếm số lượng đơn hàng mà mỗi khách hàng đã mua, nhằm xác định tần suất giao dịch của họ.



* Dữ liệu sau khi gom nhóm khách hàng theo mã khách hàng. Dữ liệu bao gồm 793 mẫu và 6 cột thuộc tính.

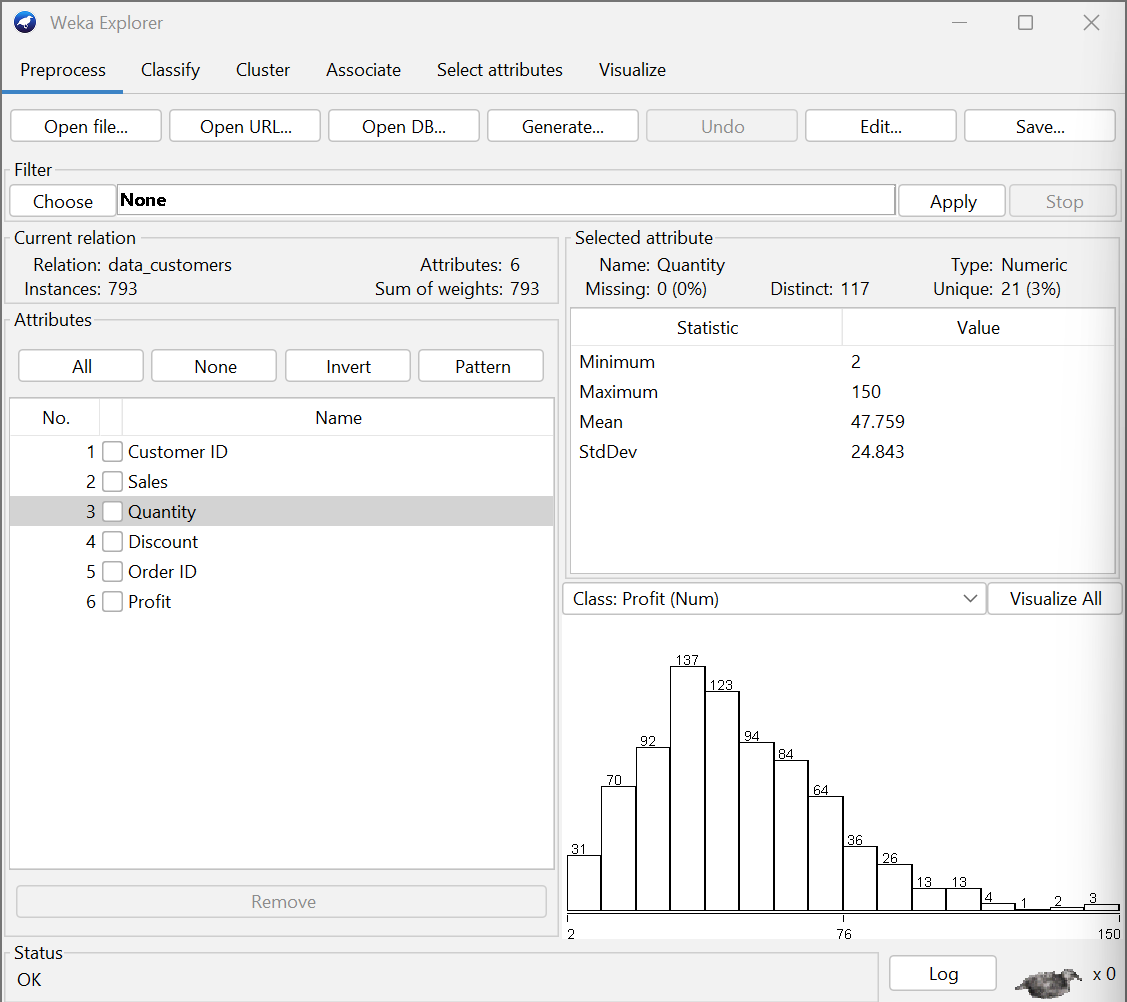


# **II: Quy trình khai phá dữ liệu**

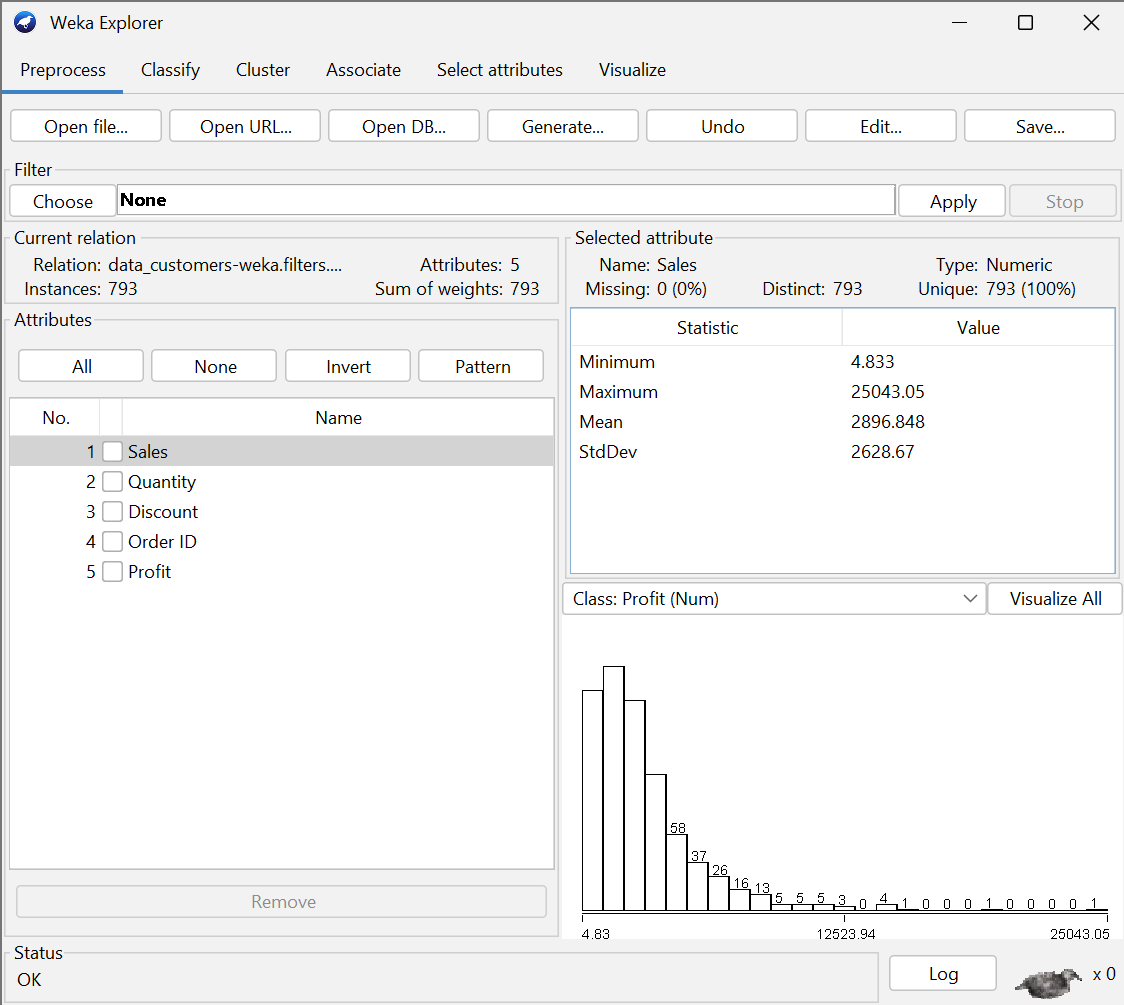
## **2.1. Tiền xử lý dữ liệu**

- Là quá trình xử lý dữ liệu thô/gốc nhằm cải thiện chất lượng dữ liệu và chất lượng của kết quả KPDL

* Đọc dữ liệu vào Weka:



* Loại bỏ thuộc tính dư thừa:



## **2.2. Làm sạch dữ liệu**

**●** Làm sạch dữ liệu là tiến hành xử lý các dữ liệu bị thiếu.

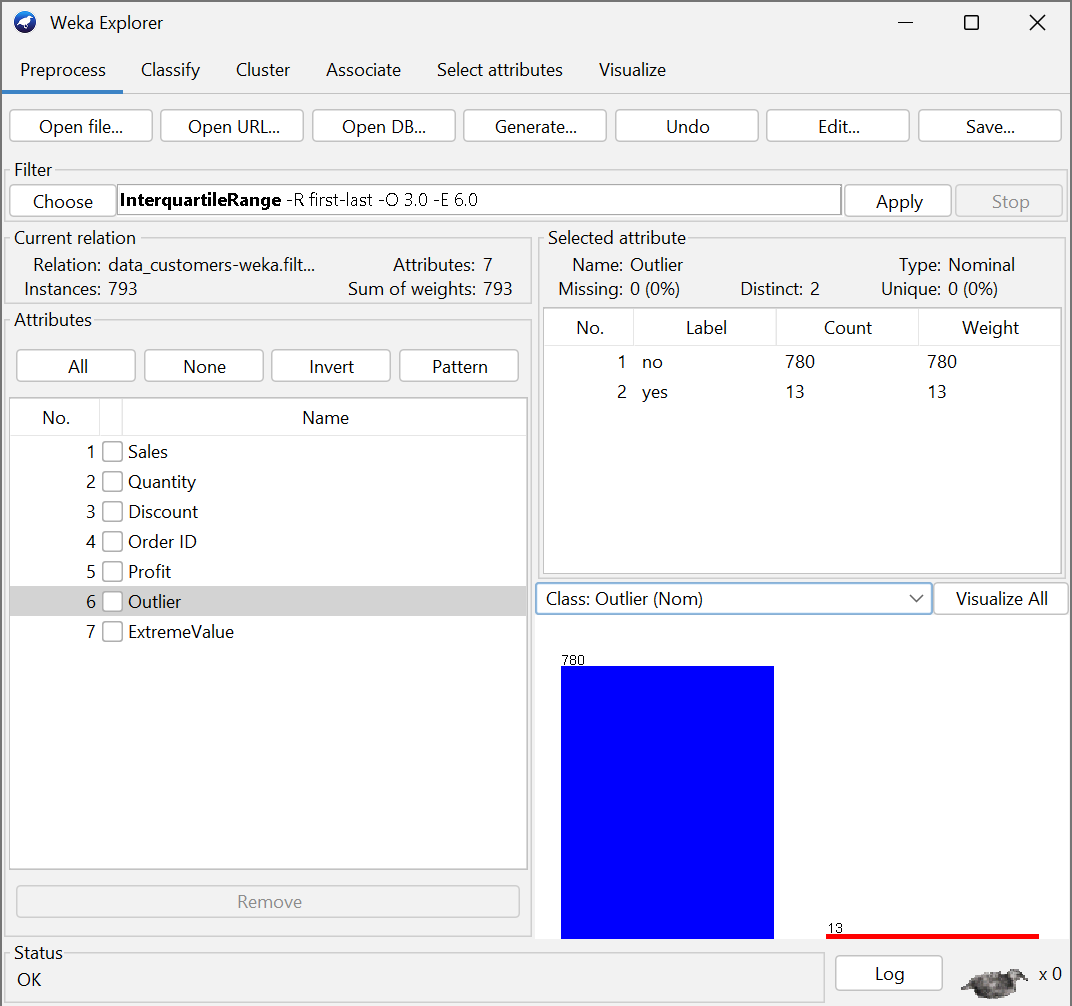
**●** Thực hiện:

(1) Tóm tắt dữ liệu: Nhận diện điểm chung của dữ liệu và sự hiện diện của nhiễu hoặc các phần tử kì dị (outliers).

- Thực hiện: Sử dụng bộ lọc InterquartileRange để phát hiện các giá trị ngoại lệ (outlier) và giá trị cực đoan (extreme values).

* Filter -> Unsupervised -> Attribute -> InterquartileRange -> Apply

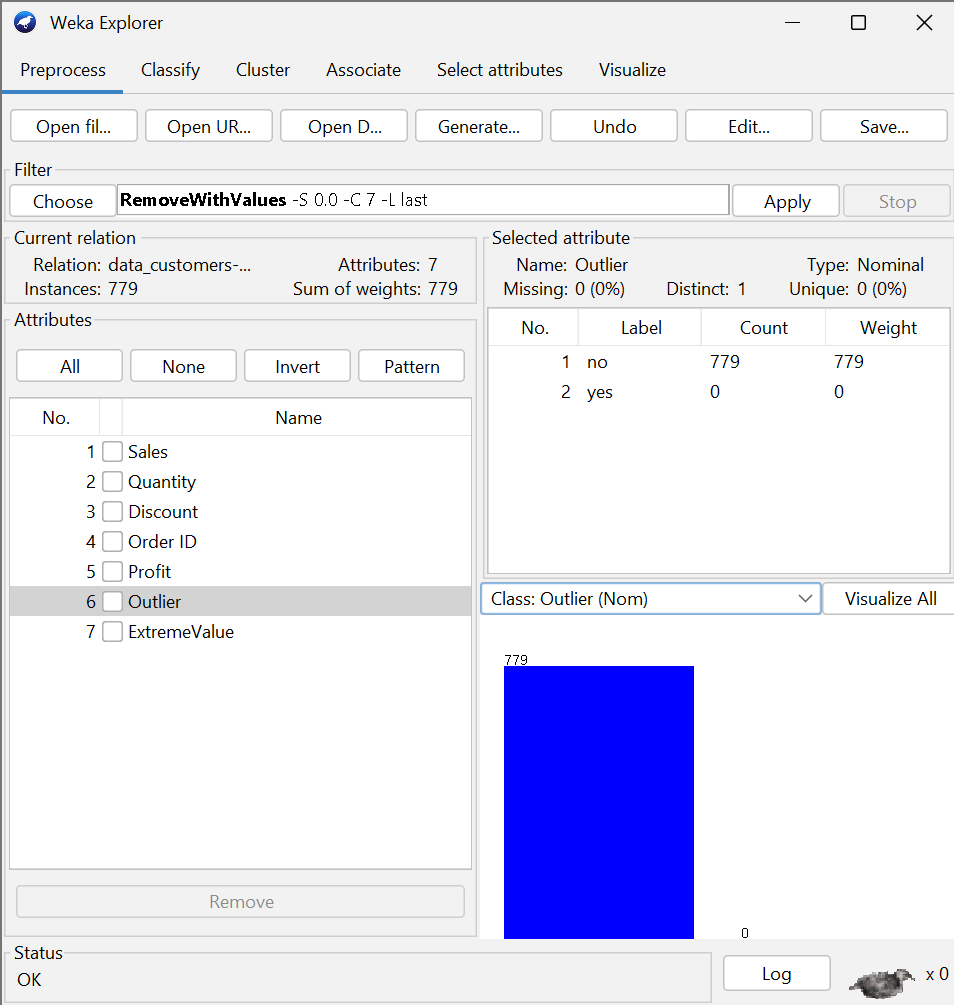
=> Phát hiện có ngoại lai và giá trị cực đoan

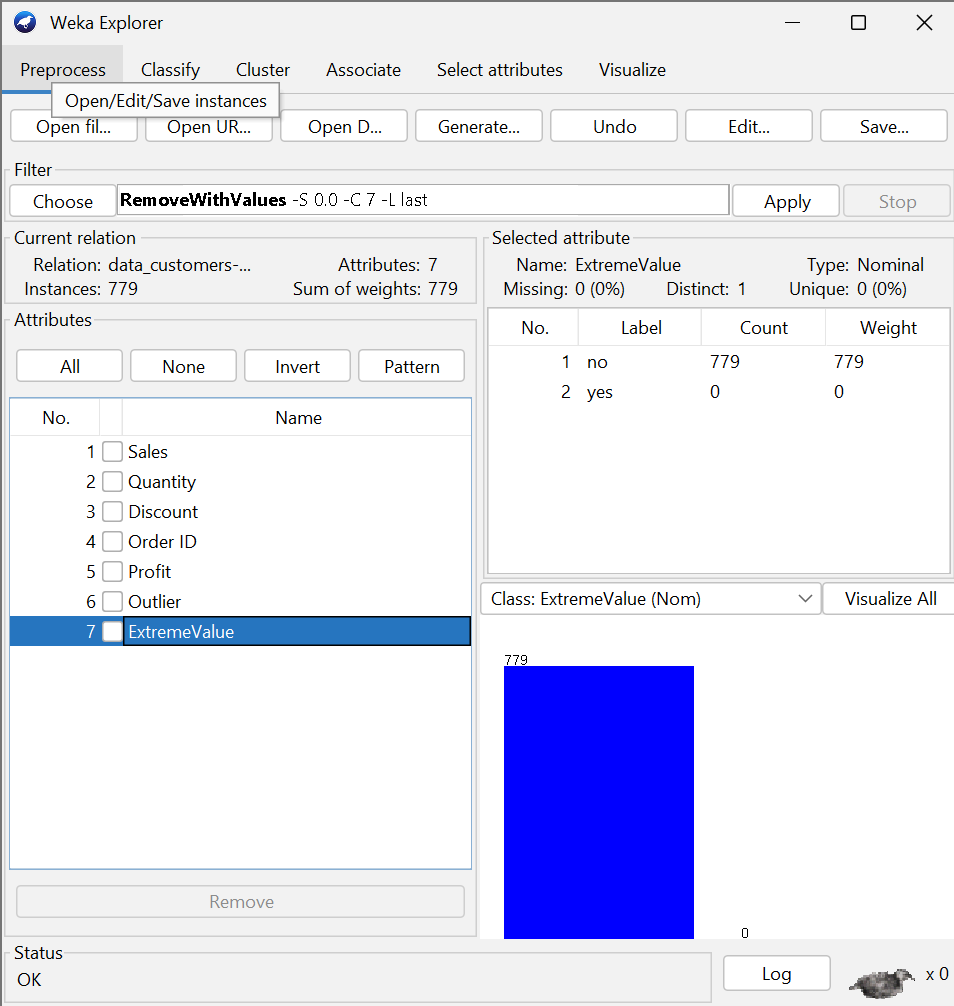


- Xử lý:

* Filter -> Unsupervised -> Instance -> RemoveWithValues -> Apply

=> Thu được kết quả sau:





(2) Xử lý dữ liệu bị thiếu (missing data):

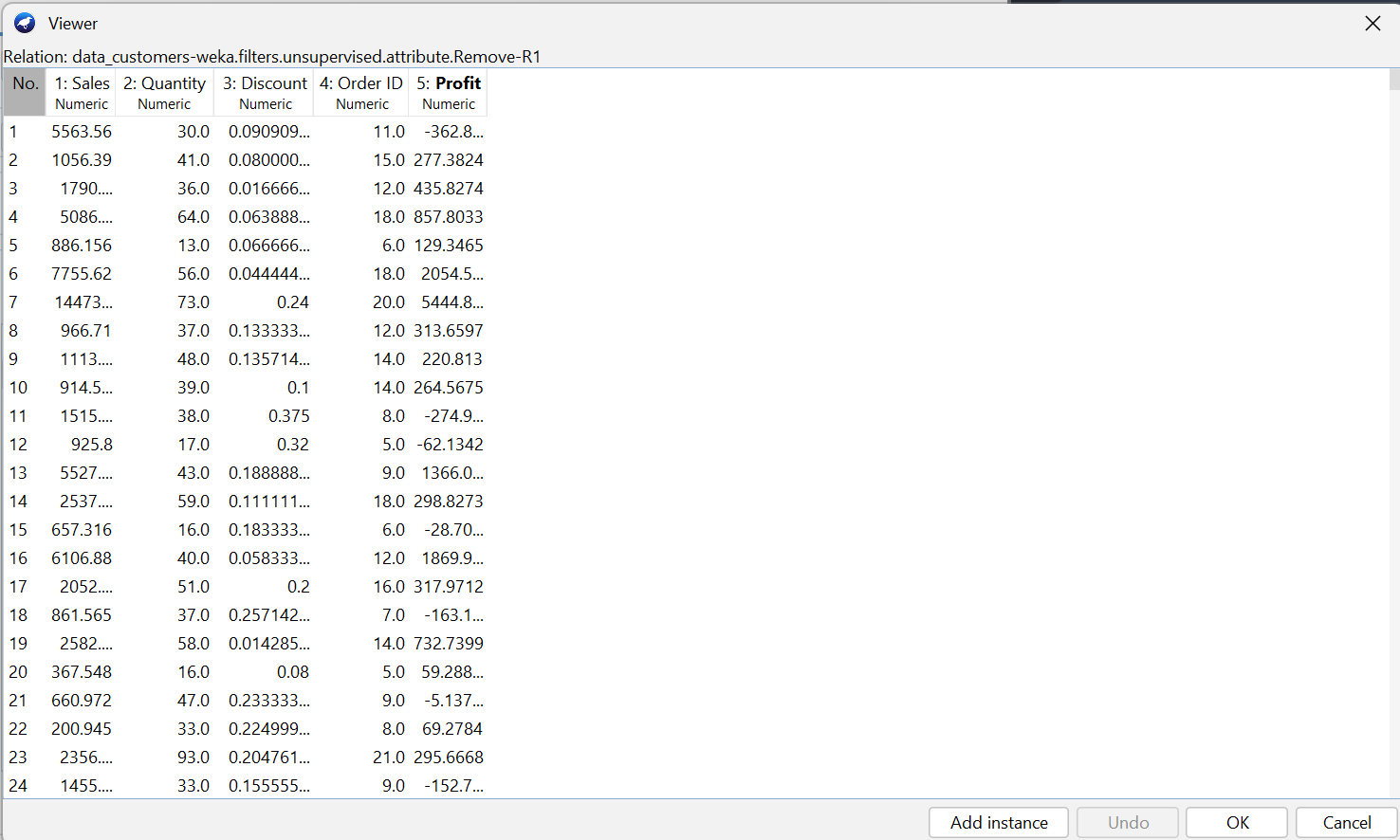
- Thực hiện: Trong cột có missing data.

- Các dữ liệu thiếu thay thế bằng giá trị trung bình của thuộc tính dùng bộ lọc ReplaceMissingValue.

=>Nhận thấy Không có cột nào bị missing data.

(3) Xử lý dữ liệu không nhất quán:

- Tập dữ liệu đã nhất quán.



## **2.3. Tích hợp dữ liệu**

- Là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn

sàng cho quá trình khai phá dữ liệu.

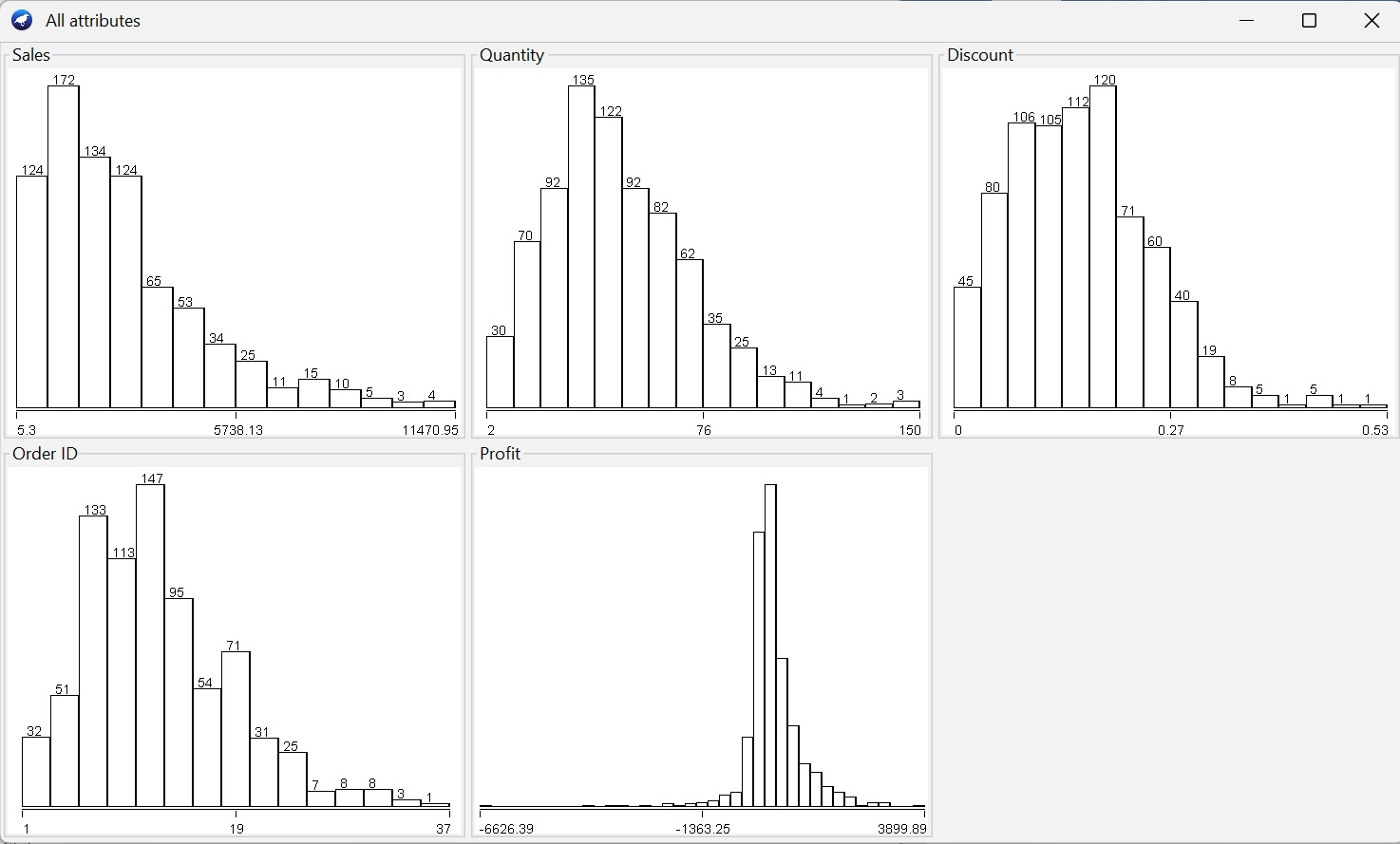
- Liên quan đến cấu trúc và tính không thuần nhất (heterogeneity) về ngữ nghĩa (semantics) của dữ liệu.

- Hỗ trợ việc giảm và tránh dư thừa và không nhất quán về dữ liệu -> cải thiện tính chính xác và tốc độ quá trình khai phá dữ liệu.

=> Do dữ liệu lấy từ 1 nguồn nên quy trình này bỏ qua.

**2.4. Biến đổi dữ liệu**

● Biến đổi dữ liệu là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu.

****

* 1. **Khai phá dữ liệu bằng mô hình phân cụm**
     1. **Giới thiệu về phân cụm:**

**-**  Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các quy trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

- Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, “natural clusters”, “useful” clusters, outlier detection.

* + 1. **Thuật toán phân cụm K-means:**

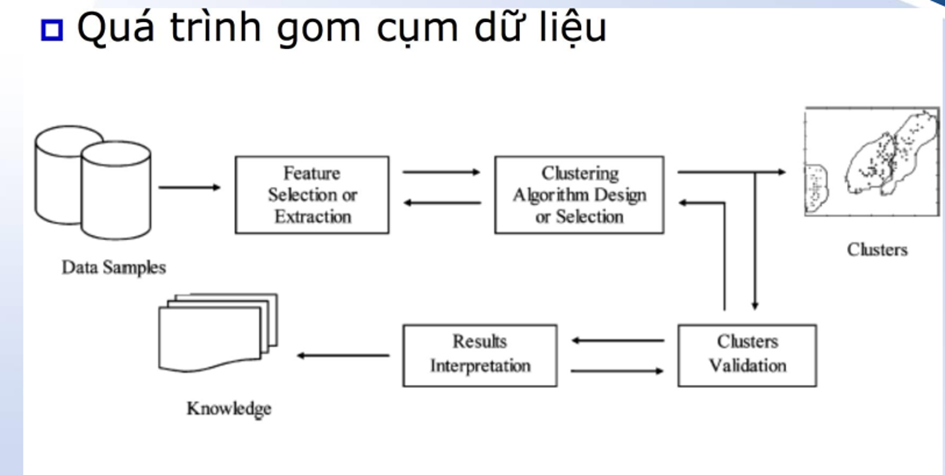
**-** Giới thiệu thuật toán K-means: là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid ) là nhỏ nhất.

- Thuật toán K-Means thực hiện qua các bước chính sau:

* Bước 1: Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
* Bước 2: Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)
* Bước 3: Nhóm các đối tượng vào nhóm gần nhất
* Bước 4: Xác định lại tâm mới cho các nhóm
* Bước 5: Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

- Các bước thực hiện thuật toán K-means trong weka

- Quá trình gom cụm dữ liệu



- Các bước thực hiện thuật toán K-means trong Weka:

Bước 1: Cho Weka đọc dữ liệu phân cụm đã khai phá ở trên.

Bước 2: Chon Cluster và chọn SimpleKMeans.

Bước 3: Chọn numClustes = 4

Bước 4: Tiến hành chạy thuật toán phân cụm.

● **Kết quả chạy thu được:**

=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation:     data\_customers-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.InterquartileRange-Rfirst-last-O3.0-E6.0-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C6-Llast-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-C7-Llast-weka.filters.unsupervised.attribute.Remove-R6-7

Instances:    779

Attributes:   5

              Sales

              Quantity

              Discount

              Order ID

              Profit

Test mode:    evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 31

Within cluster sum of squared errors: 38.05768981797041

Initial starting points (random):

Cluster 0: 2394.025, 55, 0.188462, 13, 28.7612

Cluster 1: 10310.88, 120, 0.071875, 32, 1393.5154

Cluster 2: 4805.344, 53, 0.092308, 13, 1308.5546

Cluster 3: 1252.8, 19, 0.233333, 6, 122.3612

Missing values globally replaced with mean/mode

Final cluster centroids:

                       Cluster#

Attribute   Full Data         0         1         2         3

              (779.0)   (164.0)   (101.0)   (268.0)   (246.0)

=============================================================

Sales       2705.8836 1400.5014  6264.781 3331.0822 1433.8545

Quantity      47.5045   29.7561   89.3069   58.6866   29.9919

Discount       0.1571     0.263    0.1577    0.1492    0.0947

Order ID      12.5353    7.9817   22.9802   15.3172     8.252

Profit       311.1495  -66.9957  764.9758  434.9266  242.0726

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      164 ( 21%)

1      101 ( 13%)

2      268 ( 34%)

3      246 ( 32%)

·**Giải thích kết quả:**

● ***Scheme***: Đây là thông số của thuật toán phân cụm SimpleKMeans. Các tham số bao gồm: cách khởi tạo (init), số lần chọn thử tối đa (max-candidates), chu kỳ cắt tỉa (periodic-pruning), mật độ tối thiểu (min-density), ngưỡng t1 và t2, số cụm (N), phương pháp tính khoảng cách (A), số lần lặp (I), số vị trí tạo cụm (num-slots) và hạt (S).

● ***Relation***: Đây là thông tin về quan hệ giữa các thuộc tính trong tập dữ liệu. Có 6 thuộc tính: Sales, Quantity, Discount, Order ID và Profit.

● ***Instances***: Số lượng các mẫu trong tập dữ liệu là 794.

● ***Clustering mode***l: Đây là kết quả của mô hình phân cụm trên toàn bộ tập dữ liệu huấn luyện.

● ***Number of iterations***: Số lần lặp để tối ưu hóa mô hình là 31.

● ***Within cluster sum of squared errors***: Tổng bình phương sai trong cụm, được sử dụng để đánh giá chất lượng phân cụm, là 38.05768981797041

● ***Initial starting points (random)***: Các điểm bắt đầu ban đầu của các cụm, được chọn ngẫu nhiên, được liệt kê cho từng cụm.

Cluster 0: 2394.025,55,0.188462,13,28.7612

Cluster 1: 10310.88,120,0.071875,32,1393.5154

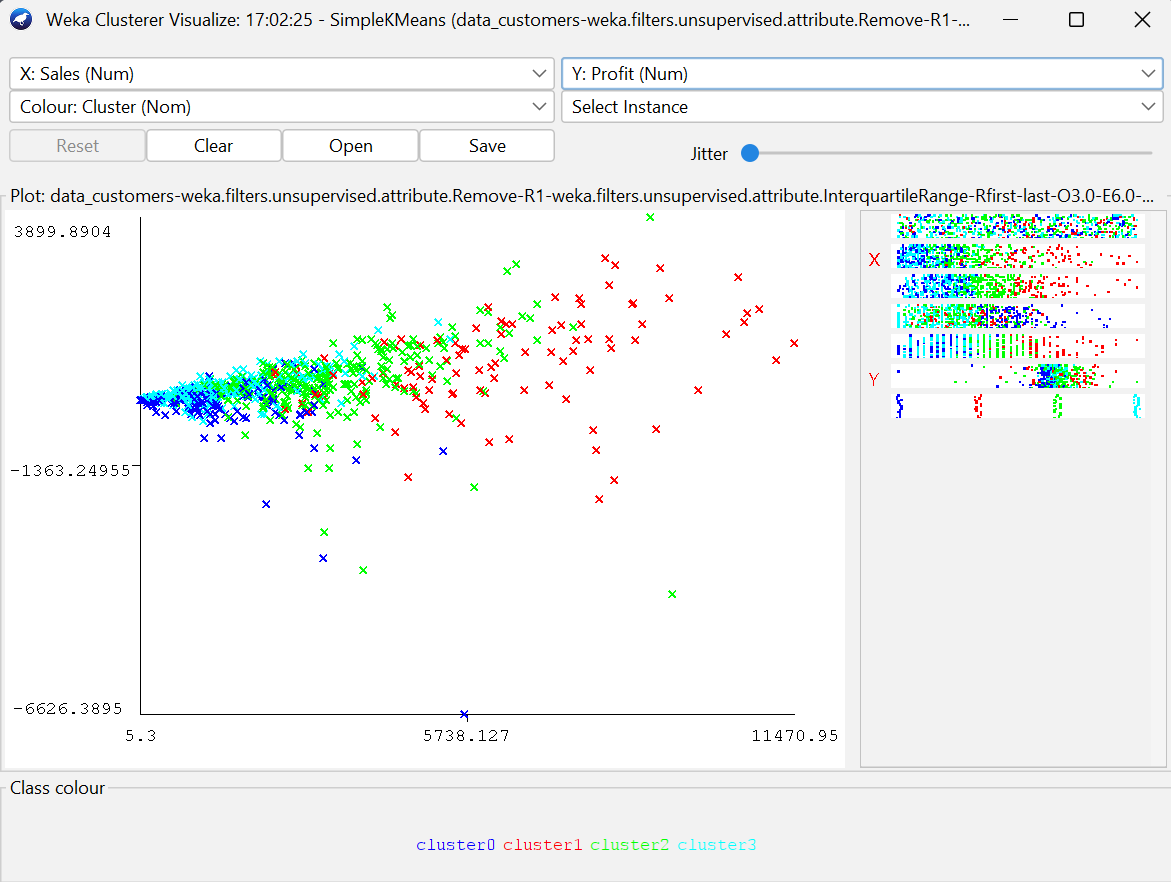
Cluster 2: 4805.344,53,0.092308,13,1308.5546

Cluster 3: 1252.8,19,0.233333,6,122.3612

● ***Final cluster centroids***: Đây là tọa độ trung tâm cuối cùng của mỗi cụm. Các thuộc tính của từng cụm được hiển thị, bao gồm Sales, Quantity, Discount, Order ID và Profit.

● ***Time taken to build model***: Thời gian để xây dựng mô hình trên toàn bộ tập dữ liệu huấn luyện là 0.03 giây.

* ***Model and evaluation on training set***: Đây là kết quả của mô hình phân cụm trên tập dữ liệu huấn luyện. Nó hiển thị số lượng mẫu trong mỗi cụm. Ví dụ: Cụm có 0 có 164 mẫu (chiếm 21% tổng số mẫu), cụm 1 có 101 mẫu (chiếm 13% tổng số mẫu), cụm 2 có 268 mẫu (chiếm 34% tổng số mẫu), cụm 3 có 246 mẫu (chiếm 32% tổng số mẫu)



* **Chạy thuật toán phân cụm K-Means với Python:**

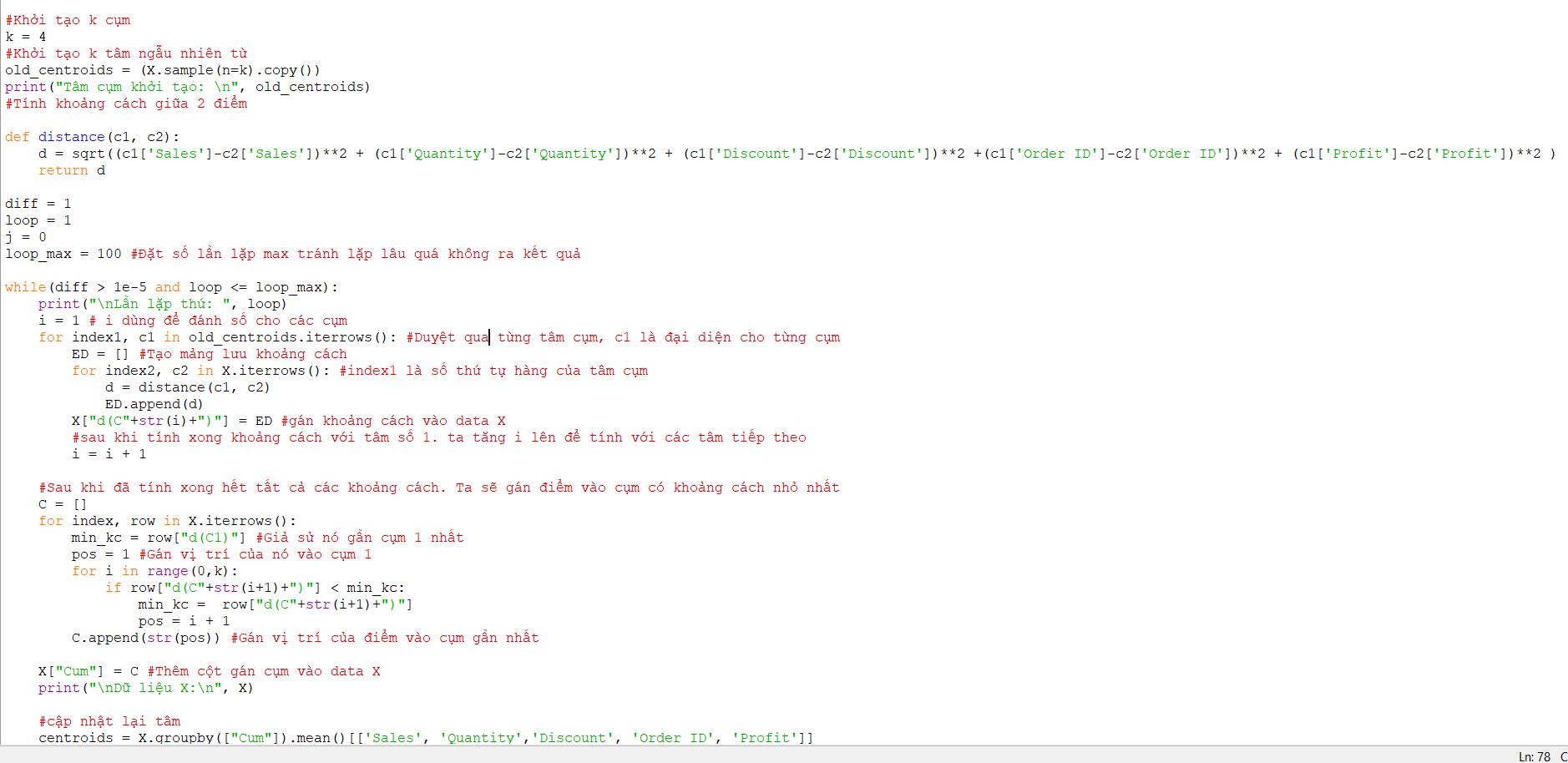
Bước 1: Đọc dữ liệu đã khai phá

Bước 2: Sử dụng mô hình phân cụm Kmean

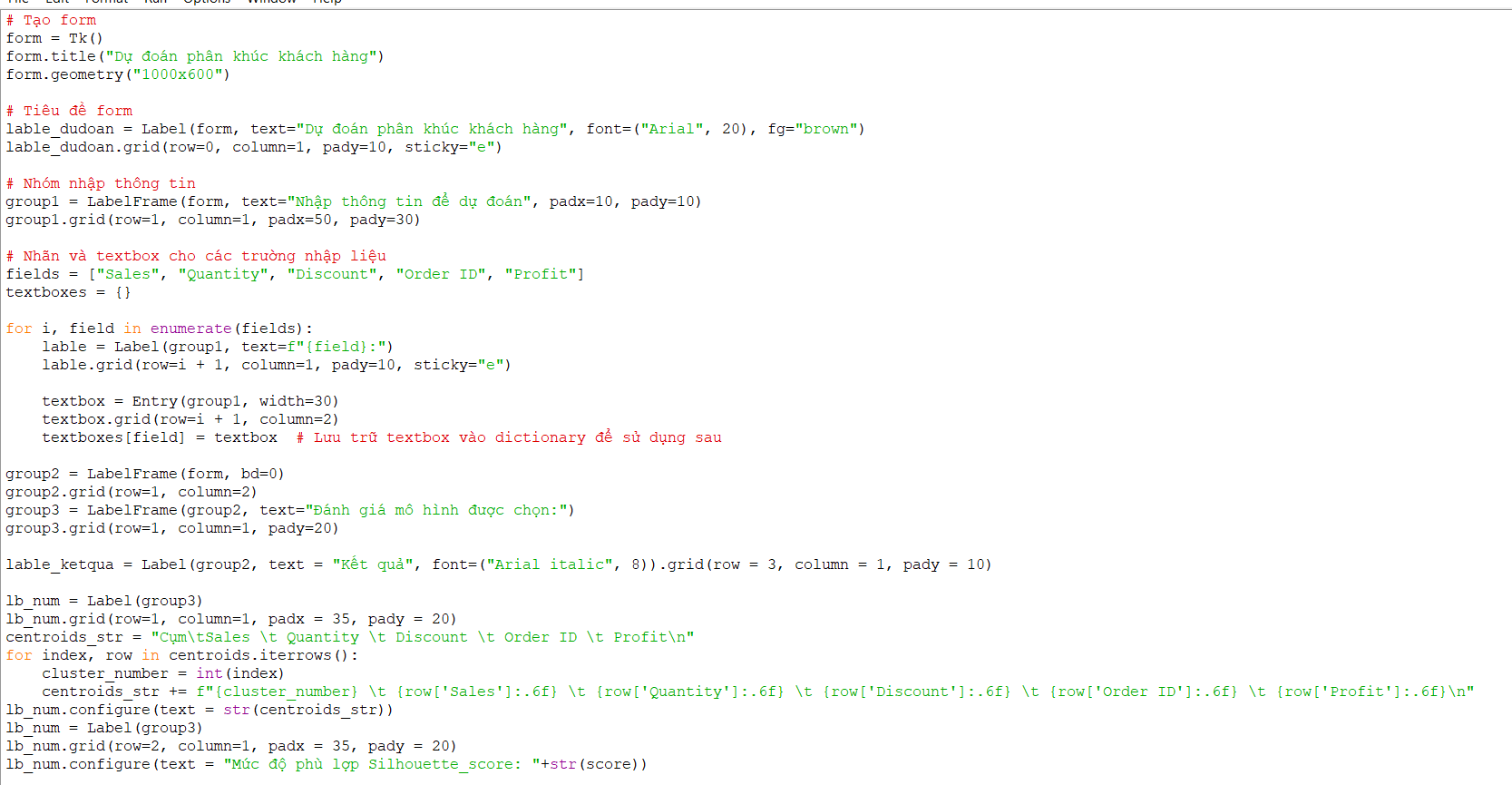
Bước 3: Chạy thuật toán và đọc dữ liệu

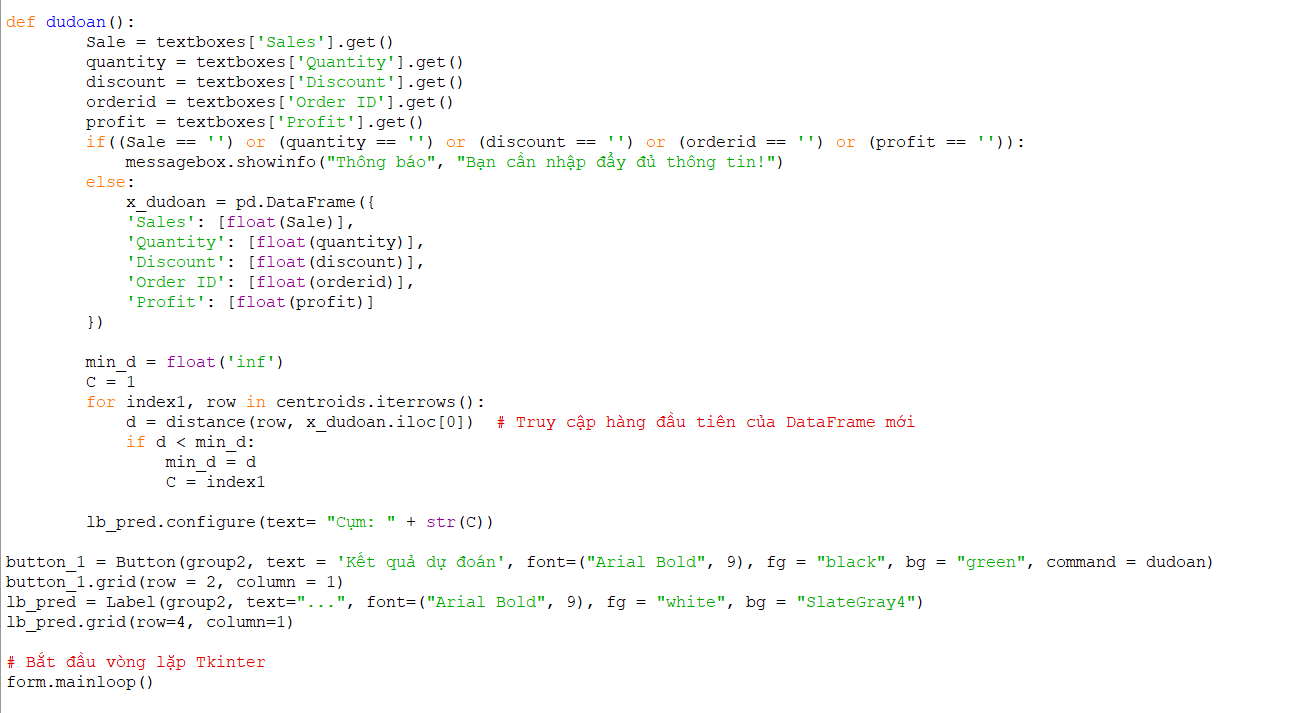
* **Chạy chương trình thực hiện thuật toán K-Means:**



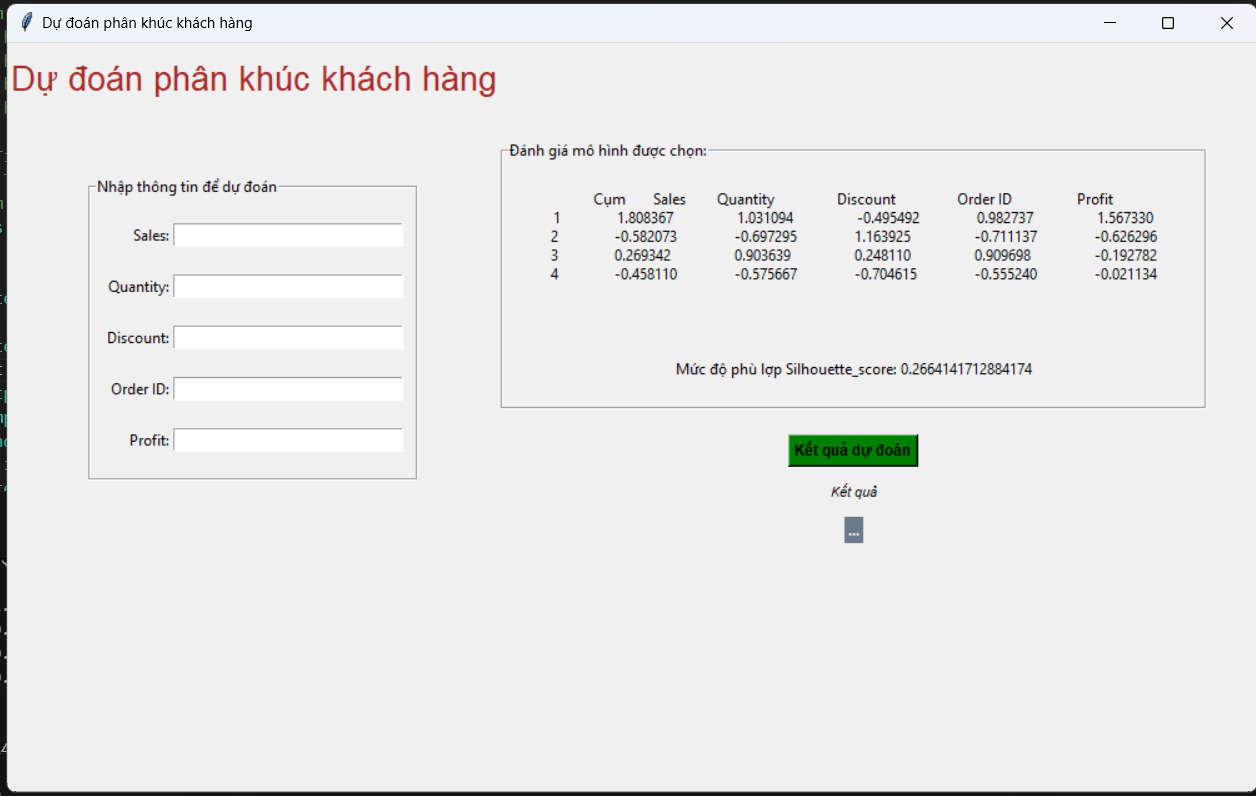








* **Kết quả chạy thu được:**



**Giải thích kết quả:**

- Thuật toán K-Means xác định 4 cụm 1, 2, 3, 4. Ta so sánh về tính tương đồng với các nhãn 1, 2, 3, 4 tự tạo trước trong tập train.

- Mẫu dự đoán được phân vào 4 cụm:

* + Cụm 1: Nhóm khách hàng tiềm năng: mua nhiều, nhận ít chiết khấu và mang lại lợi nhuận cao
  + Cụm 2: Khách hàng săn chiết khấu: mua số lượng thấp nhưng nhận được chiết khấu lớn => lỗ cho doanh nghiệp
  + Cụm 3: Khách hàng mua với số lượng tương đối lớn, chiết khấu thấp, tuy nhiên vẫn mang lại lợi nhuận âm nhẹ cho doanh nghiệp
  + Cụm 4: Khách hàng có doanh số thấp, không được chiết khấu và gần như không mang lại lợi nhuận cho doanh nghiệp

**Ý nghĩa của thuật toán K-Means:**

**- *Phân nhóm dữ liệu*:** K-means giúp tự động phát hiện cấu trúc và nhóm dữ liệu dựa trên sự tương đồng giữa các điểm. Điều này giúp trong việc tổ chức và tóm tắt thông tin từ tập dữ liệu lớn.

**- *Khám phá thông tin ẩn*:** Khi dữ liệu không có nhãn (unlabeled data), K-means có thể giúp tìm ra các cụm tự nhiên trong dữ liệu, từ đó có thể giúp hiểu sâu hơn về sự phân bố và mối quan hệ giữa các điểm dữ liệu.

**- *Nén dữ liệu*:** K-means có thể được sử dụng để biểu diễn dữ liệu dưới dạng một số lượng cụm nhỏ hơn, giúp giảm kích thước của dữ liệu mà vẫn giữ lại phần lớn thông tin quan trọng.

**- *Phát hiện ngoại lệ*:** Các điểm dữ liệu không rơi vào bất kỳ cụm nào hoặc rơi vào cụm rất xa có thể được xem xét là ngoại lệ. K-means có thể giúp xác định những điểm này.

**- *Phân cụm không giám sát (Unsupervised clustering):*** Khi có các điểm dữ liệu mới, K-means có thể được sử dụng để dự đoán cụm mà điểm mới thuộc về dựa trên các trung tâm cụm đã học từ dữ liệu huấn luyện.

**- *Tiền xử lý dữ liệu*:** K-means có thể được sử dụng để tiền xử lý dữ liệu bằng cách gán nhãn cụm cho từng điểm dữ liệu, từ đó tạo ra một biểu diễn mới của dữ liệu.

**- *Hạn chế*:** Nó cần phải biết trước số lượng cụm cần phân chia, và nó có thể bị ảnh hưởng bởi điểm khởi tạo ban đầu. Điều này có thể dẫn đến kết quả phân chia khác nhau nếu thuật toán được chạy nhiều lần với các điểm khởi tạo khác nhau.

**Kết luận****:**

- Phân lớp dữ liệu là lĩnh vực quan trọng trong khai phá dữ liệu, nó được ứng dụng trong nhiều ngành như y tế, kinh tế,… Với bài tập lớn này chúng em đã đạt được một số kết quả:

# + Hiểu về khai phá dữ liệu, bài toán phân cụm từ đó xây dựng được mô hình phân cụm để phân khúc khách hàng của trung tâm mua sắm.

# + Hiểu, biết cách xử lý các dữ liệu bị thiếu và chuẩn hóa dữ liệu.

# - Những hạn chế:

# + Do đây là thuật toán học không giám sát nên độ chính xác phụ thuộc vào dữ liệu mà chúng ta thu thập được và số dữ liệu nhiễu trong tập dữ liệu.

# + Tập dữ liệu này chỉ mới có tập trung để đánh giá việc phân khúc khách hàng của trung tâm mua sắm chứ chưa hỗ trợ để đánh giá những phần khác.

# - Trong quá trình thực hiện bài tập lớn, chúng em đã cố gắng tìm hiểu và tham khảo những tài liệu liên quan. Nhưng do thời gian và trình độ có hạn nên không tránh khỏi những hạn chế và thiếu sót. Chúng em rất mong nhận được các nhận xét và góp ý của thầy để hoàn thiện hơn kết quả thực nghiệm.

**Tài liệu tham khảo:**

**1. TS. Trần Mạnh Tuấn, Bài giảng Khai phá dữ liệu.**

**2. [https://machinelearningcoban.com](https://machinelearningcoban.com/)**

**3.https://www.kaggle.com/code/diaakotb/super-store-analysis-and-clustering/notebook#Customer-Behaviour-Clustering**