TRƯỜNG ĐẠI HỌC
SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH
HCMC University of Technology and Education
HCMUTE

## BỘ MÔN: KHAI PHÁ DỮ LIỆU

***

# BÁO CÁO ĐỒ ÁN

---

# SUPERSTORE MINING REPORT

| | |
|---|---|
| Mã môn học | : **222DAMI330484** |
| Giáo viên hướng dẫn | : **ThS. Nguyễn Văn Thành** |
| Nhóm thực hiện đồ án | : **Nhóm 16** |

TP. Hồ Chí Minh, tháng 05 năm 2023

# DANH SÁCH THÀNH VIÊN NHÓM 16

***HỌC KỲ II NĂM HỌC 2022-2023***

**Tên đề tài:** Superstore Mining Report

| STT | HỌ TÊN | MSSV | TỶ LỆ CÔNG VIỆC |
|---|---|---|---|
| 1 | Lê Hoàng Minh | 20133068 | 100% |
| 2 | Võ Hữu Nghĩa | 20133074 | 100% |
| 3 | Nguyễn Ngọc Hoài | 20133043 | 100% |

Nhận xét của giáo viên:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

# MỤC LỤC

## I. Lời mở đầu và cảm ơn

Thương mại điện tử là một phương pháp hiện đại để quảng bá sản phẩm và dịch vụ dựa trên sự phát triển nhanh chóng của internet. Các cửa hàng trực tuyến như Shopee, Lazada, Amazon, eBay, … đã được thành lập và hàng ngàn cá nhân, tổ chức đã sử dụng những của hàng trực tuyến trên để tích hợp thương mại điện tử vào hoạt động kinh doanh của họ. Sự nhanh chóng và tiện lợi đã đưa thương mại điện tử thành sự ưu tiên và hoạt động từ người dùng internet. Từ đó, phân tích dựa trên cửa hàng điện tử để xác định được những sản phẩm bán chạy, sản phẩm tiềm năng, hành vi mua sắm của khách hàng là việc làm vô cùng quan trọng để phát triển kinh doanh.

Từ một tập dữ liệu lưu lại các hóa đơn của một cửa hàng trực tuyến, nhóm đã tiền xử lý dữ liệu, phân tích sơ bộ dựa trên thuật toán **Clustering (K-Means), Classification (Decision Trees) và Association Rule Mining** trên **Visual Studio 2019**. Sau đó, nhóm tiến hành vẽ biểu đồ, phân tích và đánh giá với công cụ hỗ trợ là **Power BI**. Từ đó, chúng ta có thể tìm được những sản phẩm tiềm năng, sản phẩm bán chạy và hành vi mua sắm của khách hàng của một cửa hàng trực tuyến ở Mỹ và có thể ứng dụng để phân tích và đánh giá cho những mục đích của riêng mình.

Nhóm xin gửi lời cảm ơn chân thành đến thầy **Nguyễn Văn Thành** – giảng viên lớp học '*Khai phá dữ liệu*' đã hỗ trợ và giúp đỡ chúng em trong suốt quá trình học tập. Nhóm chúng em xin chân thành cảm ơn!

## II. Nội dung

### 1. Giới thiệu dữ liệu – Superstore Dataset

Nguồn dữ liệu: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final

Tập dữ liệu **superstore** liên quan đến một cửa hàng trực tuyến có trụ sở tại **Mỹ**. Nó chứa dữ liệu từ năm 2014-2017 và mô tả các giao dịch được thực hiện trong cửa hàng trực tuyến trong những năm này. Cụ thể, tập dữ liệu chứa thông tin về ngày đặt hàng, vận chuyển, khách hàng và địa điểm của họ, các sản phẩm mà họ đã đặt, số lượng, tổng số tiền, chiếc khấu và lợi nhuận mà cửa hàng đã thu được từ những đơn hàng này.

Tập dữ liệu gồm có 9.994 dòng (rows), mỗi dòng là một giao dịch (transaction) và 21 cột (columns)

Mô tả:

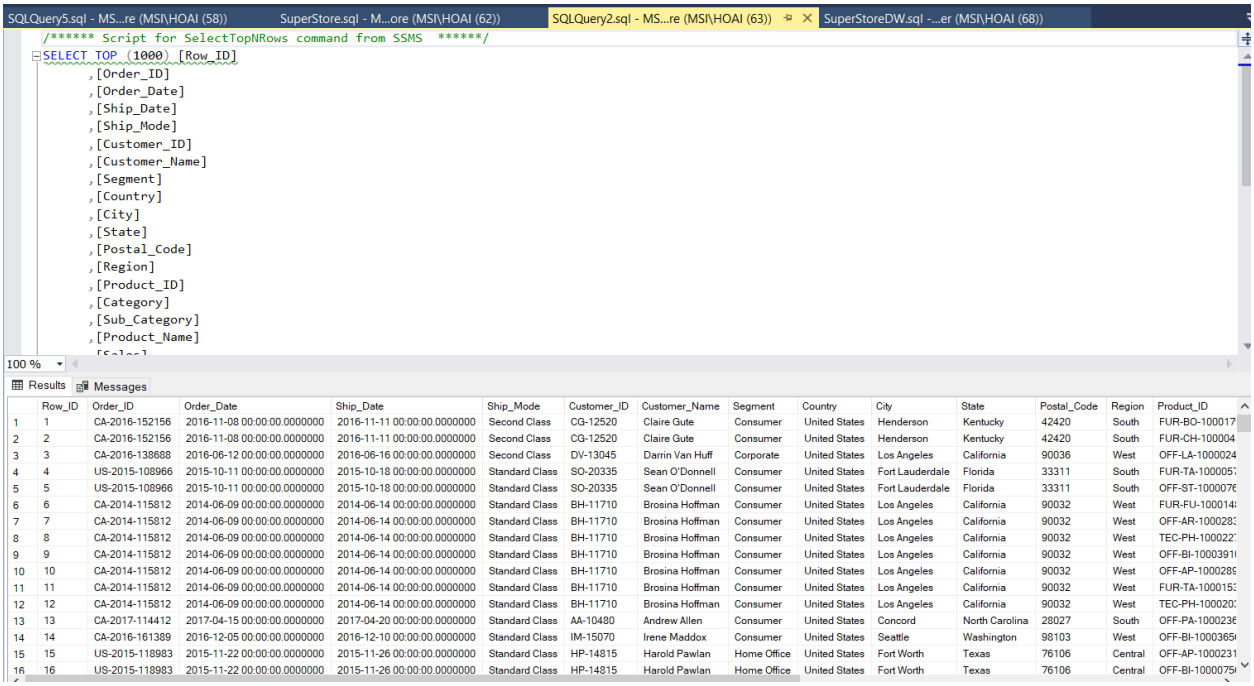
- Row ID: ID cho mỗi dòng
- Order ID: Order ID cho mỗi khách hàng
- Order Date: ngày đặt hàng của đơn hàng
- Ship Date: ngày giao hàng
- Ship Mode: hình thức giao hàng
- Customer ID : ID định danh cho mỗi khách hàng
- Customer Name: tên của khách hàng
- Segment: phân khúc khách hàng
- Country: quốc gia khách hàng sinh sống
- City: thành phố khách hàng sinh sống
- State: tiểu bang khách hàng sinh sống
- Postal code: mã vùng
- Region : vùng lãnh thổ khách hàng sinh sống

- Product ID: ID định danh của mỗi sản phẩm
- Category: thể loại chính
- Sub-category: các thể loại phụ của sản phẩm
- Product Name: tên sản phẩm
- Sales: số tiền thanh toán cho mỗi sản phẩm
- Quantity: số lượng đặt hàng cho mỗi sản phẩm
- Discount: giảm giá
- Profit: lợi nhuận

## 2. Phân tích trên Visual Studio 2019

### 2.1. Tiền xử lý dữ liệu

- Thêm cột '*Profitable*' dựa trên cột '*Profit*'. Nếu '*Profit*'<0 - '*Loss incurred*', '*Profit*'=0 – '*Zero*', '*Profit*'>0 và '*Profit*'<=300 – '*Normal profit*', '*Profit*'>300 – '*High profit*'
- Đưa dữ liệu vào SQL




| | Row_ID | Order_ID | Order_Date | Ship_Date | Ship_Mode | Customer_ID | Customer_Name | Segment | Country | City | State | Postal_Code | Region | Product_ID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | CA-2016-152156 | 2016-11-08 00:00:00.0000000 | 2016-11-11 00:00:00.0000000 | Second Class | CG-12520 | Claire Gute | Consumer | United States | Henderson | Kentucky | 42420 | South | FUR-BO-100017 |
| 2 | 2 | CA-2016-152156 | 2016-11-08 00:00:00.0000000 | 2016-11-11 00:00:00.0000000 | Second Class | CG-12520 | Claire Gute | Consumer | United States | Henderson | Kentucky | 42420 | South | FUR-CH-100004 |
| 3 | 3 | CA-2016-138688 | 2016-06-12 00:00:00.0000000 | 2016-06-16 00:00:00.0000000 | Second Class | DV-13045 | Darrin Van Huff | Corporate | United States | Los Angeles | California | 90036 | West | OFF-LA-1000024 |
| 4 | 4 | US-2015-108966 | 2015-10-11 00:00:00.0000000 | 2015-10-18 00:00:00.0000000 | Standard Class | SO-20335 | Sean O'Donnell | Consumer | United States | Fort Lauderdale | Florida | 33311 | South | FUR-TA-100005 |
| 5 | 5 | US-2015-108966 | 2015-10-11 00:00:00.0000000 | 2015-10-18 00:00:00.0000000 | Standard Class | SO-20335 | Sean O'Donnell | Consumer | United States | Fort Lauderdale | Florida | 33311 | South | OFF-ST-100007 |
| 6 | 6 | CA-2014-115812 | 2014-06-09 00:00:00.0000000 | 2014-06-14 00:00:00.0000000 | Standard Class | BH-11710 | Brosina Hoffman | Consumer | United States | Los Angeles | California | 90032 | West | FUR-FU-100014 |
| 7 | 7 | CA-2014-115812 | 2014-06-09 00:00:00.0000000 | 2014-06-14 00:00:00.0000000 | Standard Class | BH-11710 | Brosina Hoffman | Consumer | United States | Los Angeles | California | 90032 | West | OFF-AR-100028 |
| 8 | 8 | CA-2014-115812 | 2014-06-09 00:00:00.0000000 | 2014-06-14 00:00:00.0000000 | Standard Class | BH-11710 | Brosina Hoffman | Consumer | United States | Los Angeles | California | 90032 | West | TEC-PH-100022 |
| 9 | 9 | CA-2014-115812 | 2014-06-09 00:00:00.0000000 | 2014-06-14 00:00:00.0000000 | Standard Class | BH-11710 | Brosina Hoffman | Consumer | United States | Los Angeles | California | 90032 | West | OFF-BI-100039 |
| 10 | 10 | CA-2014-115812 | 2014-06-09 00:00:00.0000000 | 2014-06-14 00:00:00.0000000 | Standard Class | BH-11710 | Brosina Hoffman | Consumer | United States | Los Angeles | California | 90032 | West | OFF-AP-100028 |
| 11 | 11 | CA-2014-115812 | 2014-06-09 00:00:00.0000000 | 2014-06-14 00:00:00.0000000 | Standard Class | BH-11710 | Brosina Hoffman | Consumer | United States | Los Angeles | California | 90032 | West | FUR-TA-100015 |
| 12 | 12 | CA-2014-115812 | 2014-06-09 00:00:00.0000000 | 2014-06-14 00:00:00.0000000 | Standard Class | BH-11710 | Brosina Hoffman | Consumer | United States | Los Angeles | California | 90032 | West | TEC-PH-100020 |
| 13 | 13 | CA-2017-114412 | 2017-04-15 00:00:00.0000000 | 2017-04-20 00:00:00.0000000 | Standard Class | AA-10480 | Andrew Allen | Consumer | United States | Concord | North Carolina | 28027 | South | OFF-PA-100023 |
| 14 | 14 | CA-2016-161389 | 2016-12-05 00:00:00.0000000 | 2016-12-10 00:00:00.0000000 | Standard Class | IM-15070 | Irene Maddox | Consumer | United States | Seattle | Washington | 98103 | West | OFF-BI-100036 |
| 15 | 15 | US-2015-118983 | 2015-11-22 00:00:00.0000000 | 2015-11-26 00:00:00.0000000 | Standard Class | HP-14815 | Harold Pawlan | Home Office | United States | Fort Worth | Texas | 76106 | Central | OFF-AP-100023 |
| 16 | 16 | US-2015-118983 | 2015-11-22 00:00:00.0000000 | 2015-11-26 00:00:00.0000000 | Standard Class | HP-14815 | Harold Pawlan | Home Office | United States | Fort Worth | Texas | 76106 | Central | OFF-BI-100007 |

*Dữ liệu ban đầu chưa qua tiền xử lý*

- Loại bỏ các cột không cần thiết, thay đổi kiểu dữ liệu cho các cột

```
-- Column 'Country' take only the value 'United States' so this column is unnecessary. We have to drop it.
ALTER TABLE dbo.SuperstoreData DROP COLUMN Country;

-- Change Order_Date and Ship_Date to DATE.
ALTER TABLE dbo.SuperstoreData ALTER COLUMN Order_Date DATE;
ALTER TABLE dbo.SuperstoreData ALTER COLUMN Ship_Date DATE;

-- Change some column from string to float
ALTER TABLE dbo.SuperstoreData ALTER COLUMN Quantity INT;
ALTER TABLE dbo.SuperstoreData ALTER COLUMN Discount FLOAT;
ALTER TABLE dbo.SuperstoreData ALTER COLUMN Profit FLOAT;
```

| Customer_Name | Segment | City | State | Postal_Code | Region | Product_ID | Category | Sub_Category | Product_Name | Sales | Quantity | Discount | Profit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Claire Gute | Consumer | Henderson | Kentucky | 42420 | South | FUR-BO-10001798 | Furniture | Bookcases | Bush Somerset Collection Bookcase | 261.96 | 2 | 0 | 41.9136 |
| Claire Gute | Consumer | Henderson | Kentucky | 42420 | South | FUR-CH-10000454 | Furniture | Chairs | Hon Deluxe Fabric Upholstered Stacking Chairs, Ro... | 731.94 | 3 | 0 | 219.582 |
| Darrin Van Huff | Corporate | Los Angeles | California | 90036 | West | OFF-LA-10000240 | Office Supplies | Labels | Self-Adhesive Address Labels for Typewriters by Univ... | 14.62 | 2 | 0 | 6.8714 |
| Sean O'Donnell | Consumer | Fort Lauderdale | Florida | 33311 | South | FUR-TA-10000577 | Furniture | Tables | Bretford CR4500 Series Slim Rectangular Table | 957.5775 | 5 | 0.45 | -383.031 |
| Sean O'Donnell | Consumer | Fort Lauderdale | Florida | 33311 | South | OFF-ST-10000760 | Office Supplies | Storage | Eldon Fold 'N Roll Cart System | 22.368 | 2 | 0.2 | 2.5164 |
| Brosina Hoffman | Consumer | Los Angeles | California | 90032 | West | FUR-FU-10001487 | Furniture | Furnishings | Eldon Expressions Wood and Plastic Desk Accessor... | 48.86 | 7 | 0 | 14.1694 |
| Brosina Hoffman | Consumer | Los Angeles | California | 90032 | West | OFF-AR-10002833 | Office Supplies | Art | Newell 322 | 7.28 | 4 | 0 | 1.9656 |
| Brosina Hoffman | Consumer | Los Angeles | California | 90032 | West | TEC-PH-10002275 | Technology | Phones | Mitel 5320 IP Phone VoIP phone | 907.152 | 6 | 0.2 | 90.7152 |
| Brosina Hoffman | Consumer | Los Angeles | California | 90032 | West | OFF-BI-10003910 | Office Supplies | Binders | DXL Angle-View Binders with Locking Rings by Sam... | 18.504 | 3 | 0.2 | 5.7825 |
| Brosina Hoffman | Consumer | Los Angeles | California | 90032 | West | OFF-AP-10002892 | Office Supplies | Appliances | Belkin F5C206VTEL 6 Outlet Surge | 114.9 | 5 | 0 | 34.47 |
| Brosina Hoffman | Consumer | Los Angeles | California | 90032 | West | FUR-TA-10001539 | Furniture | Tables | Chromcraft Rectangular Conference Tables | 1706.184 | 9 | 0.2 | 85.3092 |
| Brosina Hoffman | Consumer | Los Angeles | California | 90032 | West | TEC-PH-10002033 | Technology | Phones | Konftel 250 Conference◆phone◆- Charcoal black | 911.424 | 4 | 0.2 | 68.3568 |
| Andrew Allen | Consumer | Concord | North Carolina | 28027 | South | OFF-PA-10002365 | Office Supplies | Paper | Xerox 1967 | 15.552 | 3 | 0.2 | 5.4432 |

*Dữ liệu đã qua tiền xử lý*

## 2.2. Product Mining (Clustering – Kmeans)

### 2.2.1. Thuật toán Kmeans trong kĩ thuật phân cụm (clustering)

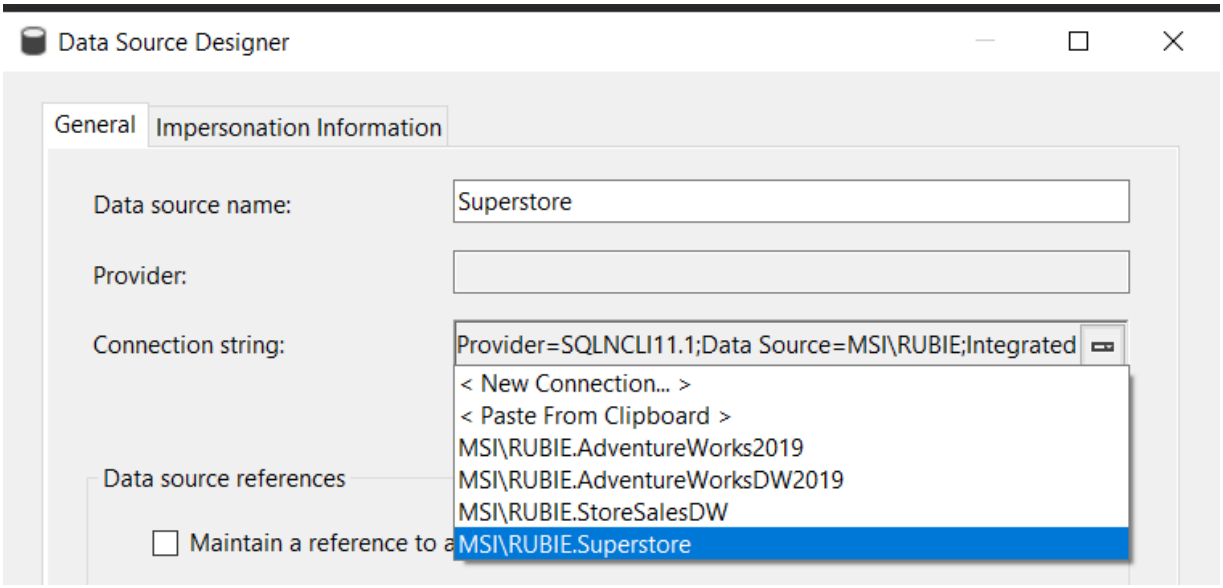
- Giải thuật phân cụm: giả sử ta có một tập dữ liệu và ta cần phải nhóm các dữ liệu có tính chất tương tự nhau vào các cụm khác nhau chưa biết trước. Một cách đơn giản để mô phỏng bài toán này là biểu diễn qua cái nhìn hình học. Các dữ liệu có thể coi là các điểm trong không gian và khoảnh cách giữa các điểm có thể được coi là thông số mức độ giống nhau của chúng. Hai điểm càng gần nhau thì càng giống nhau.
- Kmeans: là một phương pháp đơn giản và phổ biến trong kĩ thuật phân cụm
  - Bước 1: chọn k điểm bất kì làm điểm trung tâm.
  - Bước 2: nhóm dữ liệu vào một cụm có điểm trung tâm gần nhất với nó. Nếu các cụm sau khi nhóm không thay đổi so với trước khi nhóm thì ta dừng giải thuật.

- Bước 3: với mỗi cụm sau khi nhóm lại, ta cập nhật lại điểm trung tâm của chúng bằng cách lấy trung bình cộng. Sau đó, quay lại bước 2.

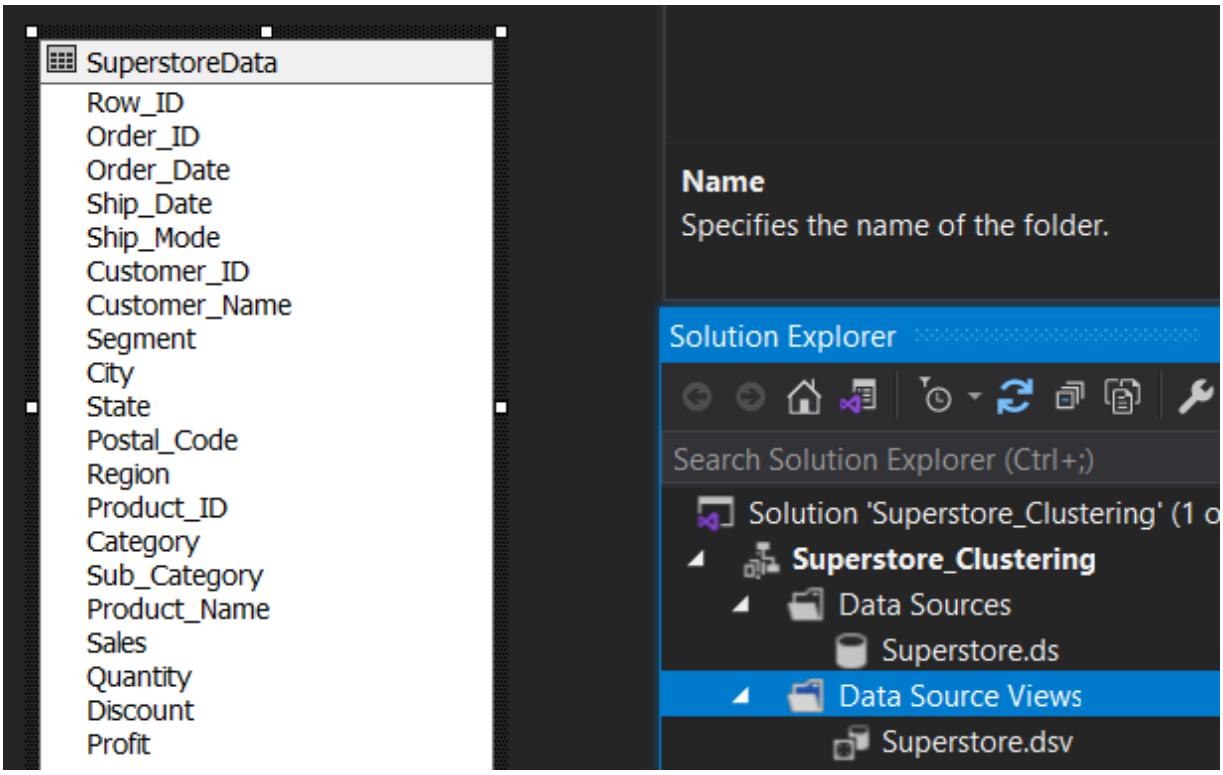
- Các hạn chế của giải thuật
  - Cần biết số nhóm trước: điều kiện đầu vào của giải thuật cần chỉ rõ giá trị của k, nhưng trong thực tế không phải lúc nào ta cũng biết trước được có bao nhiêu k cả. Dùng phương pháp elbow để xác định k hiệu quả nhất.
  - Khởi tạo ảnh hưởng tới chất lượng. Để cải thiện chất lượng thì cần chạy lại nhiều lần.
- Áp dụng vào trong bài toán: phân loại các sản phẩm (product) dựa trên lợi nhuận đem lại (profitable) và giảm giá (discount). Từ đó, ta thấy được những mặt hàng nào ứng với từng loại giảm giá nào thì đem lại lợi nhuận nhiều hay ít, từ đó đưa ra các chiến lược giảm giá hiệu quả hơn để tăng lợi nhuận cho việc kinh doanh.

### 2.2.2. Phân tích và đánh giá

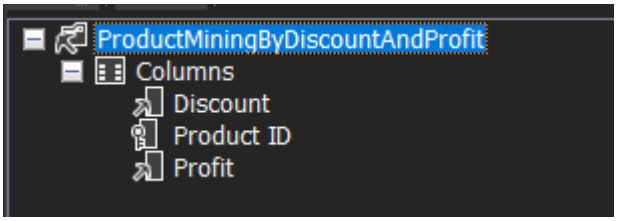
- Tạo Data Sources



- Tạo Data Source View

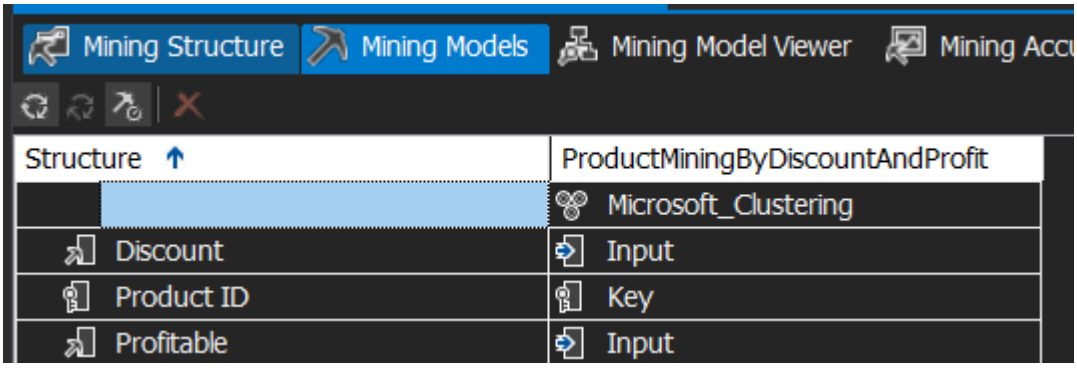


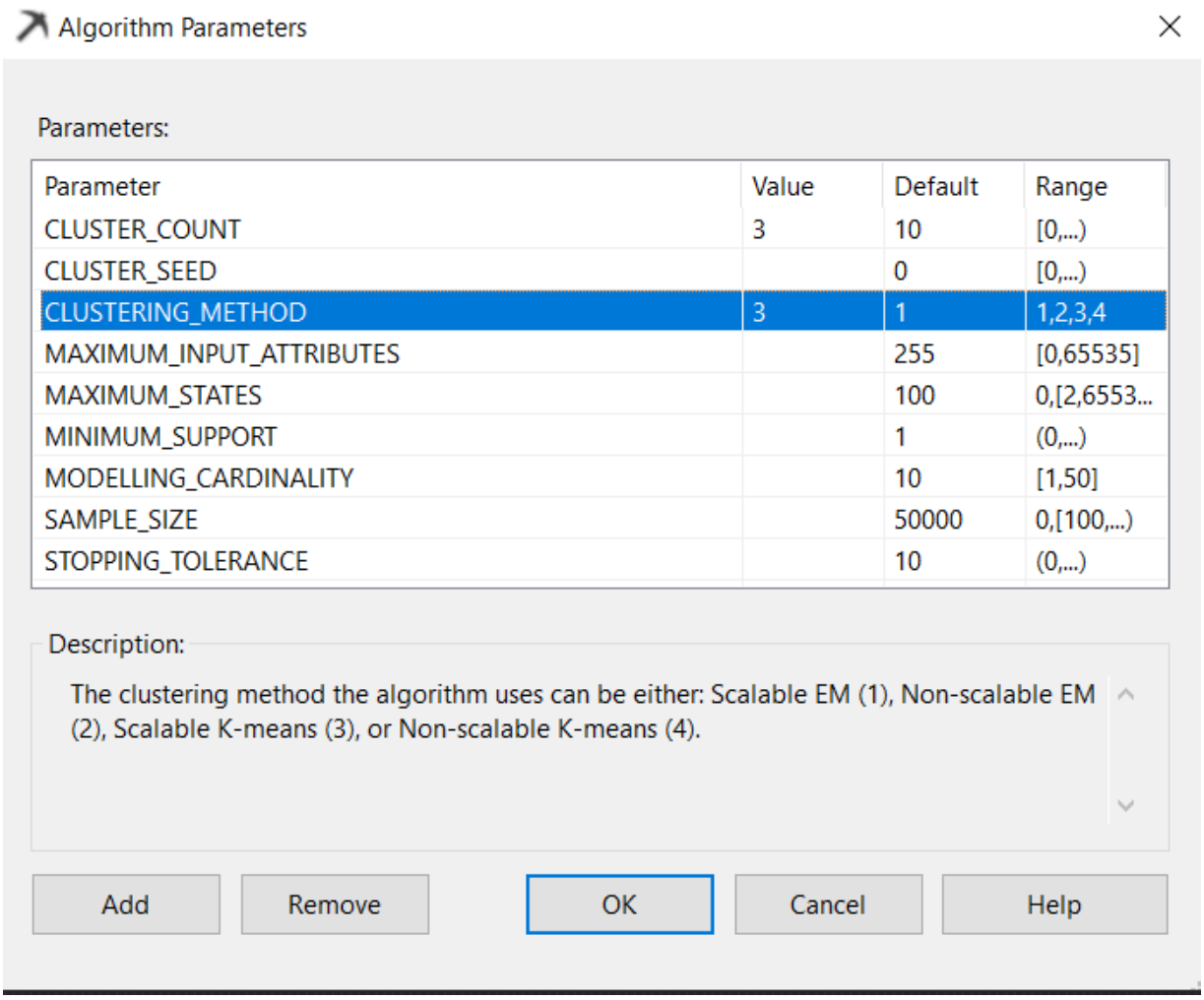

- Tạo Mining Structures




- Kết quả

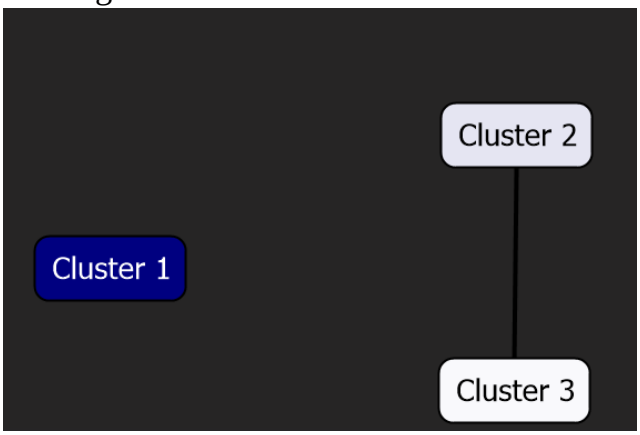
- Mining models: với mong muốn phân cụm sản phẩm để biết được sản phẩm tạo ra lợi nhuận cho cửa hàng theo từng phân cụm. Nhóm đã tạo ra ProductMining ByDiscountAndProfit, phân cụm theo từng ID sản phẩm dựa trên discount và profit. Với số lượng cụm là 3 và thuật toán sử dụng là scalable K-Means

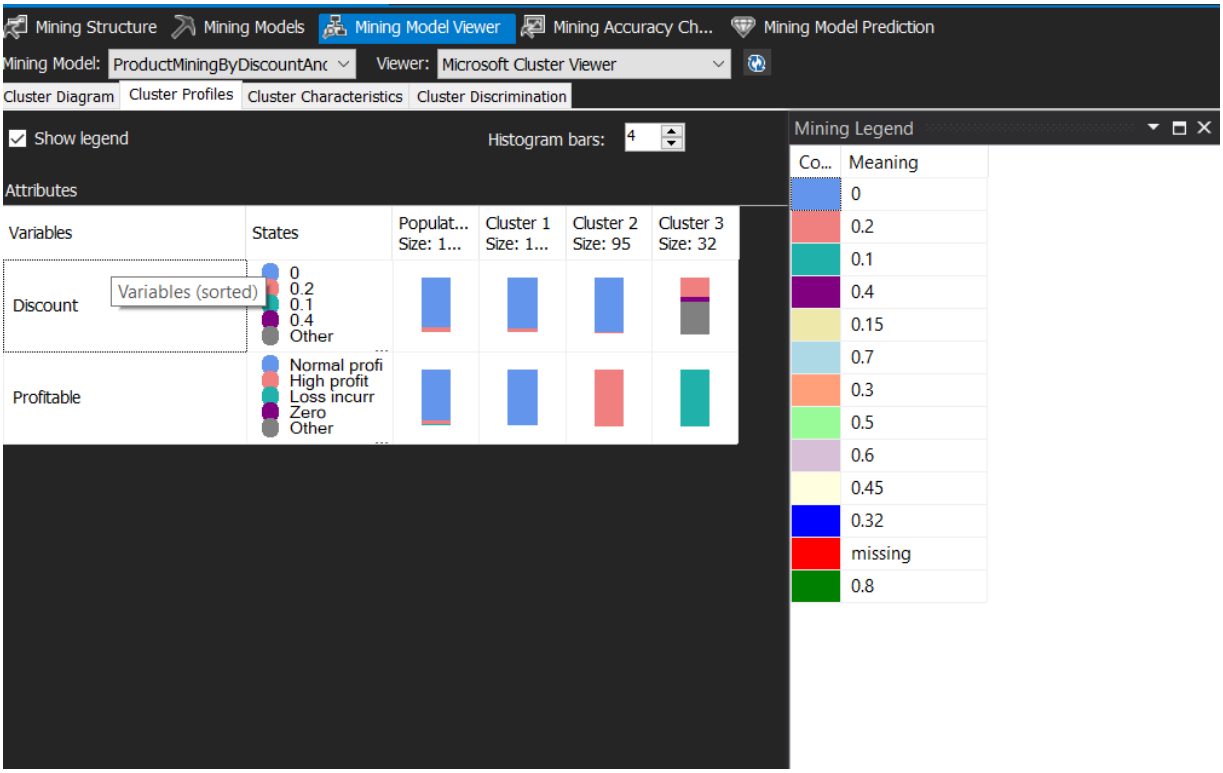



*Cấu hình số lượng cụm (3) và thuật toán xử lý (Kmeans)*

- Mining model viewer




*Chi tiết thông tin từng cụm*

- Sản phẩm được chia thành 3 cụm: Cluster1 chứa thông tin về các sản phẩm có lợi nhuận bình thường và không có lợi nhuận, cluster2 chứa thông tin về các sản phẩm có lợi nhuận cao và cluster3 chứa thông tin về các sản phẩm lợi nhuận âm.

- Nhận xét: Các sản phẩm đem lại lợi nhuận trung bình và cao thì có chỉ số giảm giá (discount) thấp. Những sản phẩm làm lỗ vốn hay không đem lại lợi nhuận thì có chỉ số giảm giá cao hơn.

### 2.3. Profitable (Classification – Decision Trees)

#### 2.3.1. Decision Trees in Classification

- Giải thuật Classification: là một phương pháp trong học máy để dự đoán và gán nhãn cho các mẫu dữ liệu vào các lớp hoặc nhãn đã được xác định trước. Nhiệm vụ của giải thuật này là học từ các mẫu dữ liệu huấn luyện có nhãn đã biết để xây dựng một mô hình hoặc hàm phân loại, sau đó áp dụng mô hình này để dự đoán nhãn cho các mẫu dữ liệu mới mà chưa có nhãn.
- Decision Trees: là một giải thuật phân loại trong học máy và trí tuệ nhân tạo. Nó dựa trên cấu trúc cây để đưa ra các quyết định phân loại dựa trên các thuộc tính của dữ liệu.
  - Bước 1: Chọn thuộc tính phân chia.
  - Bước 2: Phân chia dữ liệu.
  - Bước 3: Xử lý các nhánh con.
  - Bước 4: Xây dựng Decision Trees.
- Ưu điểm và hạn chế:
  - Ưu điểm:
    - Xử lý dữ liệu hỗn hợp và dữ liệu thiếu: có khả năng xử lý dữ liệu hỗn hợp, bao gồm cả dữ liệu rời rạc và dữ liệu số. Xử lý dữ liệu thiếu một cách tự nhiên và hiệu quả.
    - Tính tương tự và tính phân loại đa lớp: Cây quyết định có thể tính toán độ tương tự giữa các mẫu dữ liệu và hỗ trợ phân loại đa lớp.
  - Nhược điểm:

- Dễ bị overfitting: Cây quyết định có thể dễ bị overfitting khi cây quá phức tạp và quá fit với dữ liệu huấn luyện, dẫn đến hiệu suất kém trên dữ liệu mới.
- Nhạy cảm với biến đổi dữ liệu: Những biến đổi nhỏ trong dữ liệu có thể dẫn đến sự thay đổi lớn trong cấu trúc cây quyết định.
- Khó xử lý các quan hệ phức tạp và không tổng quát hóa tốt.

- Áp dụng vào bài toán: phân loại lợi nhuận (profitable) dựa vào doanh số bán ra của sản phẩm. Từ đó, ta sẽ xác định các điểm phân chia dữ liệu dựa trên giá trị của doanh số bán ra. Các quy tắc phân chia sẽ cho thấy mối quan hệ giữa mức độ doanh số và lợi nhuận.

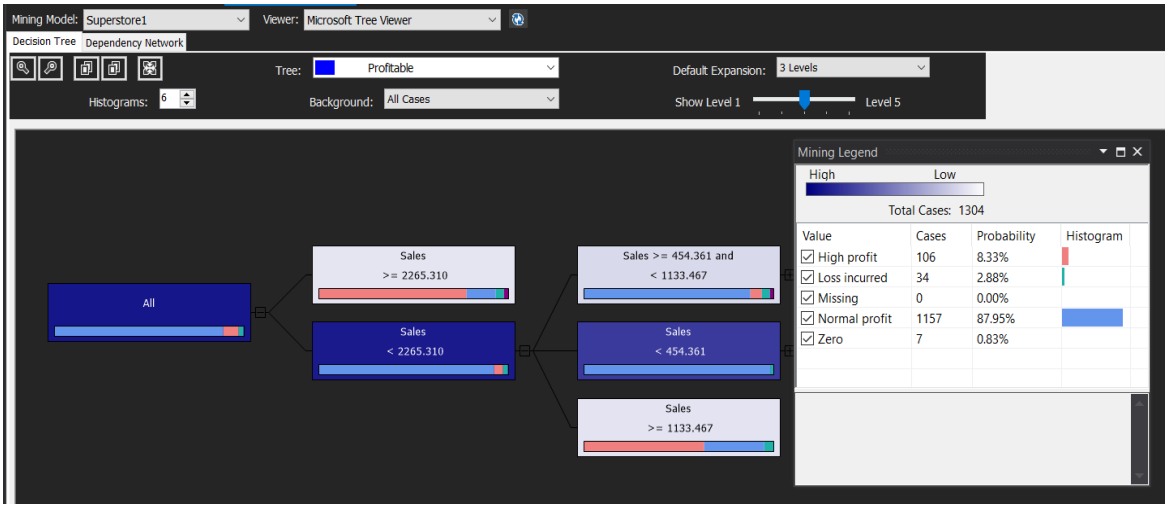
### 2.3.2. Phân tích và đánh giá

- Mining Structures

| Structure | Superstore1 |
|---|---|
| | Microsoft_Decision_Trees |
| Discount | Input |
| Product ID | Key |
| Profitable | PredictOnly |
| Quantity | Input |
| Sales | Input |

  - Input: discount (giảm giá), quantity (số lượng), sales (doanh số)
  - PredictOnly: profitable (khả năng lợi nhuận)
  - Key: product (sản phẩm)

- Kết quả




- Nhận xét: decision trees chia dữ liệu thành hai nhánh nhỏ.
  - Nhánh 1: Chứa giá trị doanh số (sales) >= $2265.310 cho mỗi sản phẩm. Nhánh này có doanh số bán ra cao và đem lại lợi nhuận khá nhiều cho cửa hàng superstore.
  - Nhánh 2: Chứa giá trị doanh số (sales) < $2265.310 cho mỗi sản phẩm. Nhánh này có doanh số bán ra thấp hơn so với nhanh 1 và đem lại lợi nhuận trung bình nhiều hơn cả. Trong nhánh này còn phân ra 3 nhánh nhỏ khác.

## 2.4. ShipMode (Association Rule)

### 2.4.1. Association Rule

- Giải thuật Association Rule: là một quy trình/phương pháp để phân loại dữ liệu vào các nhóm, lớp hoặc hạng mục khác nhau dựa trên các đặc trưng và thuộc tính của chúng. Mục tiêu của giải thuật này là xây dựng một mô hình hoặc một hàm quyết định có khả năng phân loại một mẫu dữ liệu mới thành một trong các lớp đã được định nghĩa trước.

- Association Rule: Apriori có thể nhìn vào quá khứ và khẳng định nếu một việc gì đó xảy ra thì sẽ có tỉ lệ bao nhiêu phần trăm sự việc tiếp theo xảy ra.
  - Bước 1: Duyệt toàn bộ cơ sở dữ liệu
  - Bước 2: Phân loại và kết hợp các thuộc tính xảy ra chung với nhau
  - Bước 3: So sánh và tiếp tục phân loại, kết hợp
  - Bước 4: Triển khai và sử dụng
- Ưu điểm và hạn chế:
  - Ưu điểm:
    - Tính tổng quát hóa tốt, linh hoạt và đa dạng
    - Nó có thể nhìn vào quá khứ để dự đoán tương lai, rất có ích cho các nhà kinh doanh cùng nhiều lĩnh vực khác
  - Nhược điểm:
    - Thời gian chạy có thể lâu do phải duyệt nhiều lần trong cơ sở dữ liệu
    - Độ phức tạp tính toán: tính toán phức tạp, đặc biệt khi số lượng thuộc tính hoặc mẫu dữ liệu lớn.
- Áp dụng vào bài toán: xác định xem các tiểu bang (State) nào thì hay đi kèm với loại giao hàng (Ship Mode) nào và lợi nhuận mà nó đem lại. Từ đó, ta có thể tối ưu hóa quá trình giao hàng, tập trung vào các tiểu bang và loại giao hàng có tiềm năng mang lại lợi nhuận cao, hoặc điều chỉnh chiến lược kinh doanh để tận dụng mối quan hệ giữa các yếu tố này và lợi nhuận.

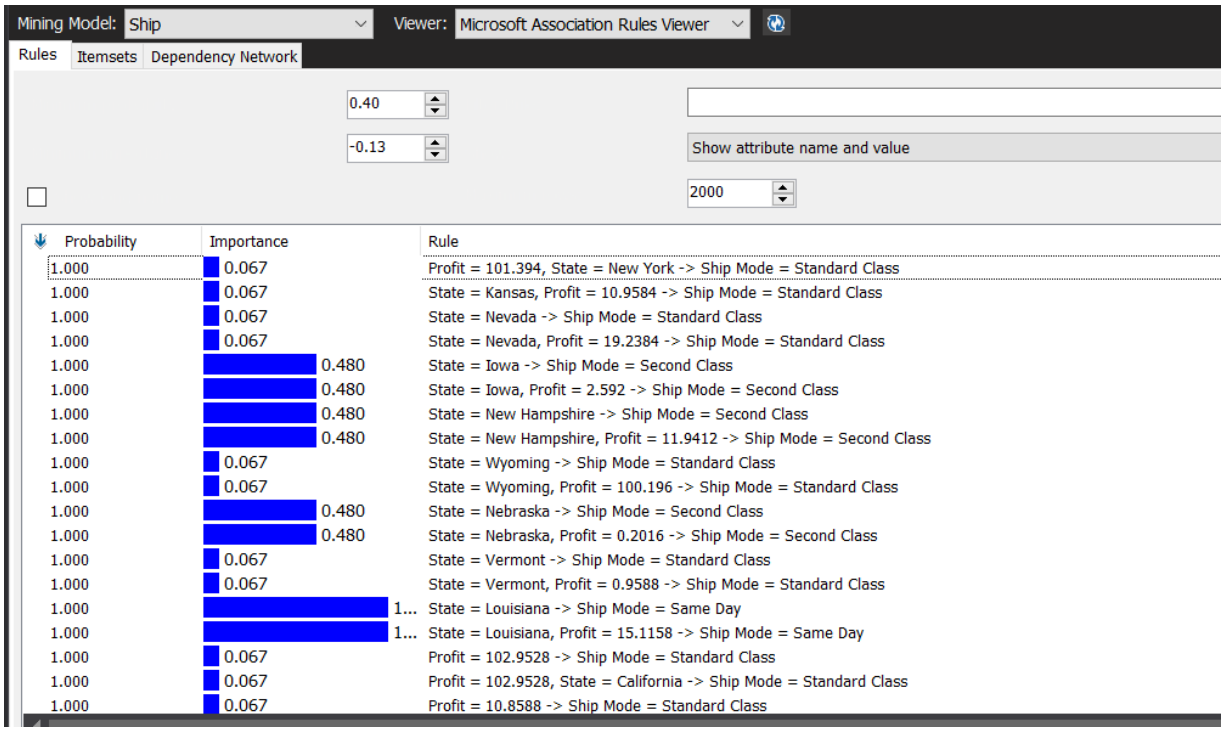
### 2.4.2. Phân tích và đánh giá

- Mining Structures

| Structure | Ship |
|---|---|
| | Microsoft_Association_Rules |
| Customer ID | Key |
| Profit | Input |
| Ship Mode | Predict |
| State | Input |

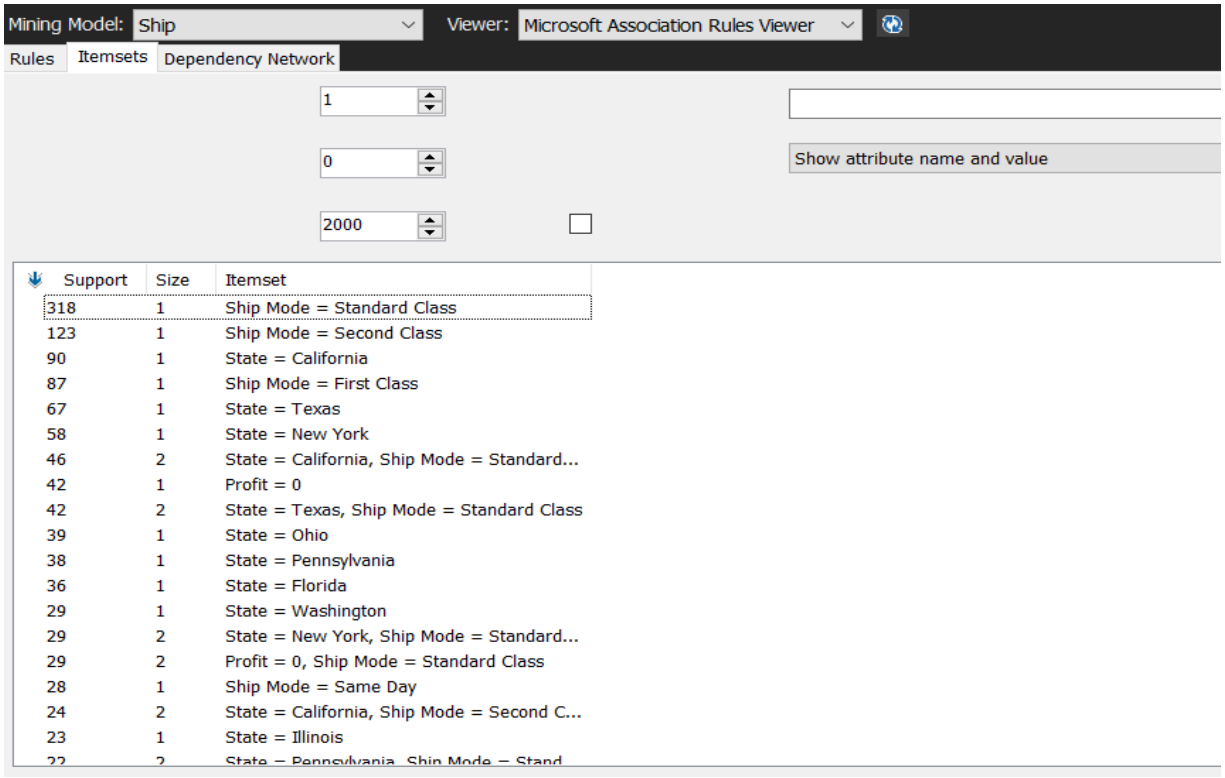
  - Input: profit (lợi nhuận), state (tiểu bang)
  - Predict: Ship Mode (chế độ giao hàng)
  - Key: customer (khách hàng)

- Kết quả:
  - Rules: Ta thấy được sự tương quan giữa profit, state và shipmode. Ở New York hay dùng loại giao hàng tiêu chuẩn (Standard Class) hay ở Louisiana thì loại giao hàng trong ngày (Same Day) là phổ biến nhất.



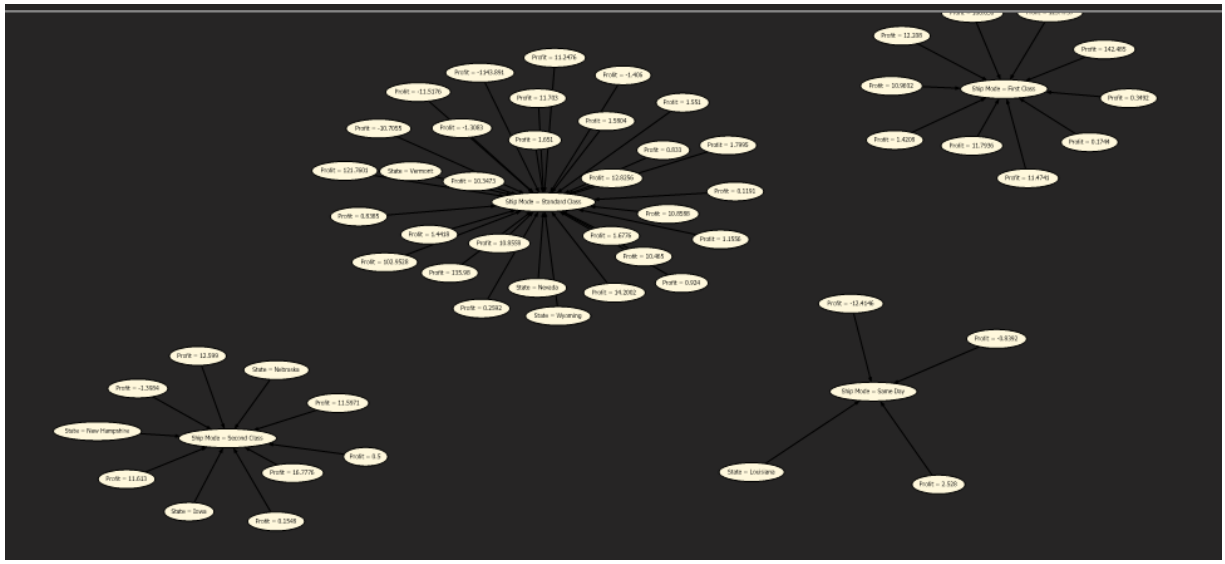

*Rules of superstore*

- Itemsets: Đây là các tập item của association rule. Tập chứa loại giao hàng tiêu chuẩn và giao hàng hạng hai là support cao hơn so với các loại còn lại.




| Support | Size | Itemset |
|---|---|---|
| 318 | 1 | Ship Mode = Standard Class |
| 123 | 1 | Ship Mode = Second Class |
| 90 | 1 | State = California |
| 87 | 1 | Ship Mode = First Class |
| 67 | 1 | State = Texas |
| 58 | 1 | State = New York |
| 46 | 2 | State = California, Ship Mode = Standard... |
| 42 | 1 | Profit = 0 |
| 42 | 2 | State = Texas, Ship Mode = Standard Class |
| 39 | 1 | State = Ohio |
| 38 | 1 | State = Pennsylvania |
| 36 | 1 | State = Florida |
| 29 | 1 | State = Washington |
| 29 | 2 | State = New York, Ship Mode = Standard... |
| 29 | 2 | Profit = 0, Ship Mode = Standard Class |
| 28 | 1 | Ship Mode = Same Day |
| 24 | 2 | State = California, Ship Mode = Second C... |
| 23 | 1 | State = Illinois |
| 22 | 2 | State = Pennsylvania, Ship Mode = Stand |

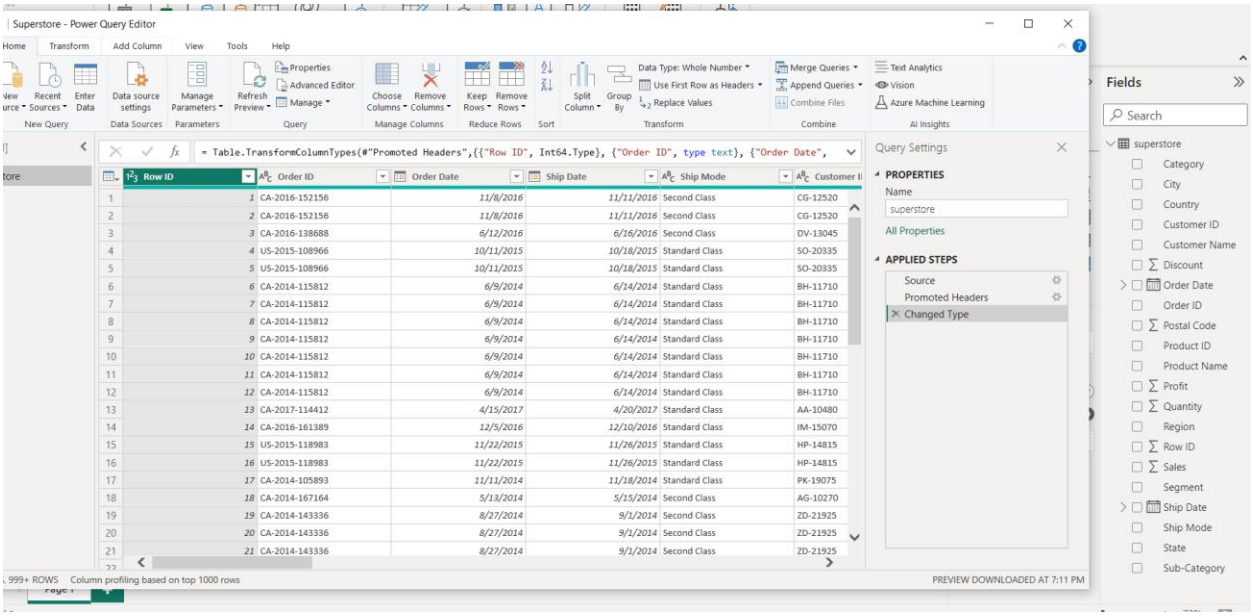
- Dependency Network: Chứa mạng lưới liên kết của profit, state và shipmode.



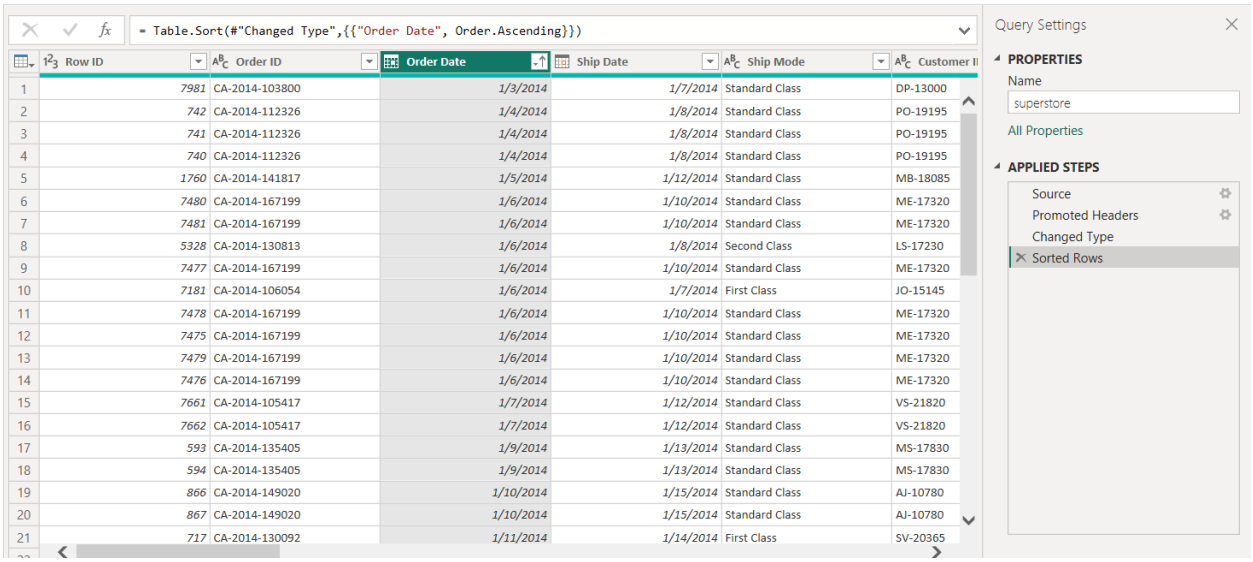
## 3. Phân tích trên Power BI

### 3.1. Tiền xử lý

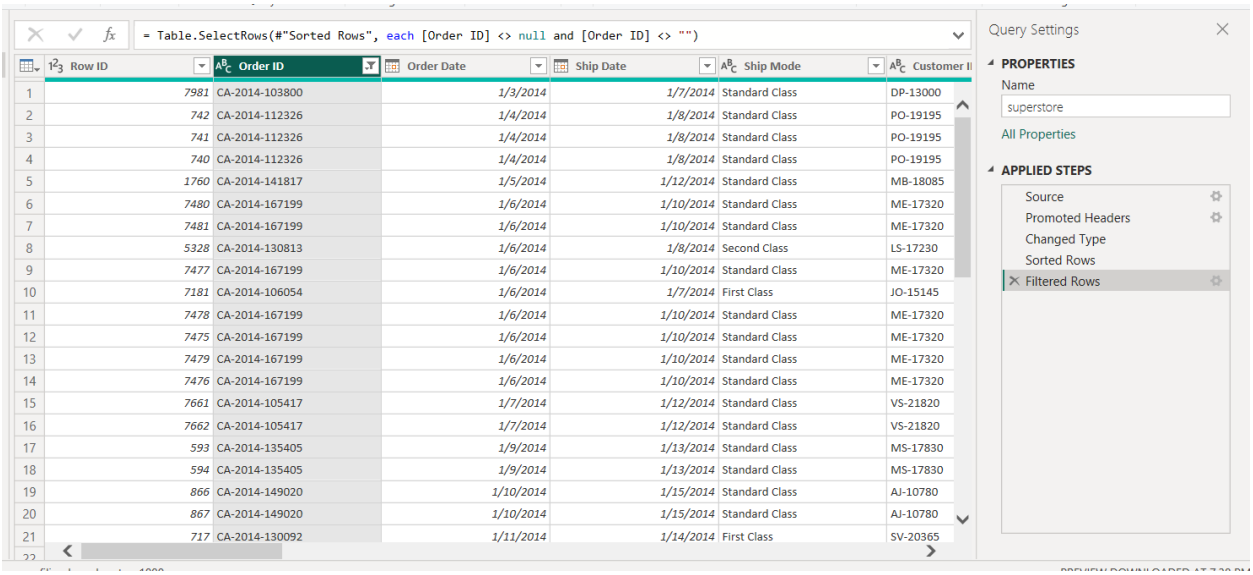
- Load dữ liệu




- Sắp xếp dữ liệu tăng dần theo Order Date

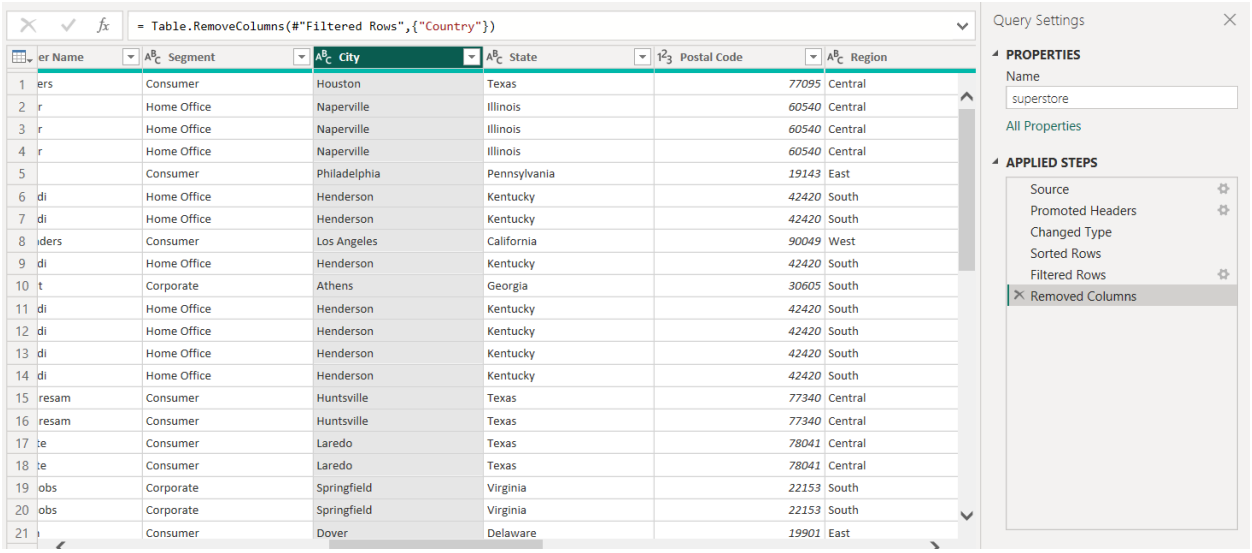


- Loại bỏ dữ liệu null theo Order ID



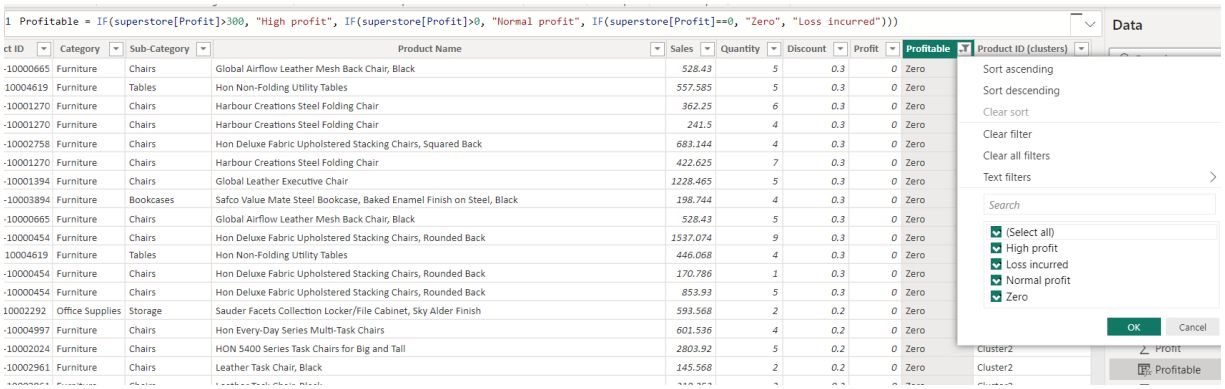

| | Row ID | Order ID | Order Date | Ship Date | Ship Mode | Customer I |
|---|---|---|---|---|---|---|
| 1 | 7981 | CA-2014-103800 | 1/3/2014 | 1/7/2014 | Standard Class | DP-13000 |
| 2 | 742 | CA-2014-112326 | 1/4/2014 | 1/8/2014 | Standard Class | PO-19195 |
| 3 | 741 | CA-2014-112326 | 1/4/2014 | 1/8/2014 | Standard Class | PO-19195 |
| 4 | 740 | CA-2014-112326 | 1/4/2014 | 1/8/2014 | Standard Class | PO-19195 |
| 5 | 1760 | CA-2014-141817 | 1/5/2014 | 1/12/2014 | Standard Class | MB-18085 |
| 6 | 7480 | CA-2014-167199 | 1/6/2014 | 1/10/2014 | Standard Class | ME-17320 |
| 7 | 7481 | CA-2014-167199 | 1/6/2014 | 1/10/2014 | Standard Class | ME-17320 |
| 8 | 5328 | CA-2014-130813 | 1/6/2014 | 1/8/2014 | Second Class | LS-17230 |
| 9 | 7477 | CA-2014-167199 | 1/6/2014 | 1/10/2014 | Standard Class | ME-17320 |
| 10 | 7181 | CA-2014-106054 | 1/6/2014 | 1/7/2014 | First Class | JO-15145 |
| 11 | 7478 | CA-2014-167199 | 1/6/2014 | 1/10/2014 | Standard Class | ME-17320 |
| 12 | 7475 | CA-2014-167199 | 1/6/2014 | 1/10/2014 | Standard Class | ME-17320 |
| 13 | 7479 | CA-2014-167199 | 1/6/2014 | 1/10/2014 | Standard Class | ME-17320 |
| 14 | 7476 | CA-2014-167199 | 1/6/2014 | 1/10/2014 | Standard Class | ME-17320 |
| 15 | 7661 | CA-2014-105417 | 1/7/2014 | 1/12/2014 | Standard Class | VS-21820 |
| 16 | 7662 | CA-2014-105417 | 1/7/2014 | 1/12/2014 | Standard Class | VS-21820 |
| 17 | 593 | CA-2014-135405 | 1/9/2014 | 1/13/2014 | Standard Class | MS-17830 |
| 18 | 594 | CA-2014-135405 | 1/9/2014 | 1/13/2014 | Standard Class | MS-17830 |
| 19 | 866 | CA-2014-149020 | 1/10/2014 | 1/15/2014 | Standard Class | AJ-10780 |
| 20 | 867 | CA-2014-149020 | 1/10/2014 | 1/15/2014 | Standard Class | AJ-10780 |
| 21 | 717 | CA-2014-130092 | 1/11/2014 | 1/14/2014 | First Class | SV-20365 |

- Loại bỏ cột ‘Country’

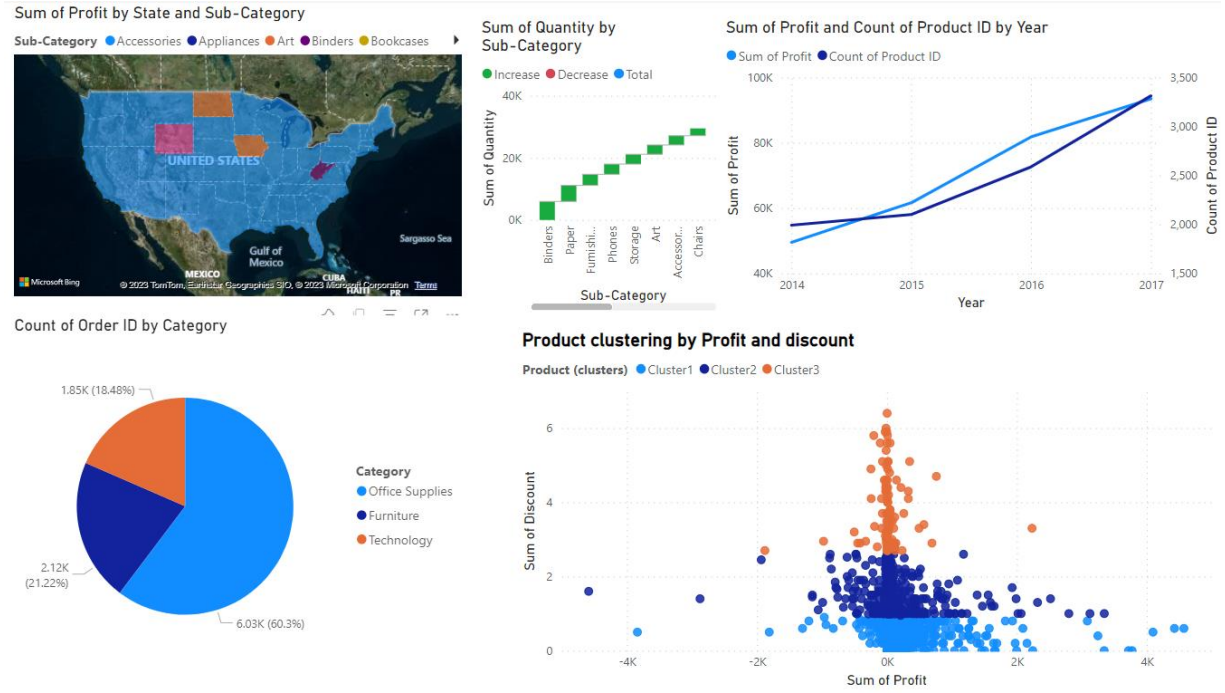



| | er Name | Segment | City | State | Postal Code | Region |
|---|---|---|---|---|---|---|
| 1 | ers | Consumer | Houston | Texas | 77095 | Central |
| 2 | r | Home Office | Naperville | Illinois | 60540 | Central |
| 3 | r | Home Office | Naperville | Illinois | 60540 | Central |
| 4 | r | Home Office | Naperville | Illinois | 60540 | Central |
| 5 | | Consumer | Philadelphia | Pennsylvania | 19143 | East |
| 6 | di | Home Office | Henderson | Kentucky | 42420 | South |
| 7 | di | Home Office | Henderson | Kentucky | 42420 | South |
| 8 | ders | Consumer | Los Angeles | California | 90049 | West |
| 9 | di | Home Office | Henderson | Kentucky | 42420 | South |
| 10 | t | Corporate | Athens | Georgia | 30605 | South |
| 11 | di | Home Office | Henderson | Kentucky | 42420 | South |
| 12 | di | Home Office | Henderson | Kentucky | 42420 | South |
| 13 | di | Home Office | Henderson | Kentucky | 42420 | South |
| 14 | di | Home Office | Henderson | Kentucky | 42420 | South |
| 15 | resam | Consumer | Huntsville | Texas | 77340 | Central |
| 16 | resam | Consumer | Huntsville | Texas | 77340 | Central |
| 17 | te | Consumer | Laredo | Texas | 78041 | Central |
| 18 | te | Consumer | Laredo | Texas | 78041 | Central |
| 19 | obs | Corporate | Springfield | Virginia | 22153 | South |
| 20 | obs | Corporate | Springfield | Virginia | 22153 | South |
| 21 | | Consumer | Dover | Delaware | 19901 | East |

- Thêm cột ‘Profitable’

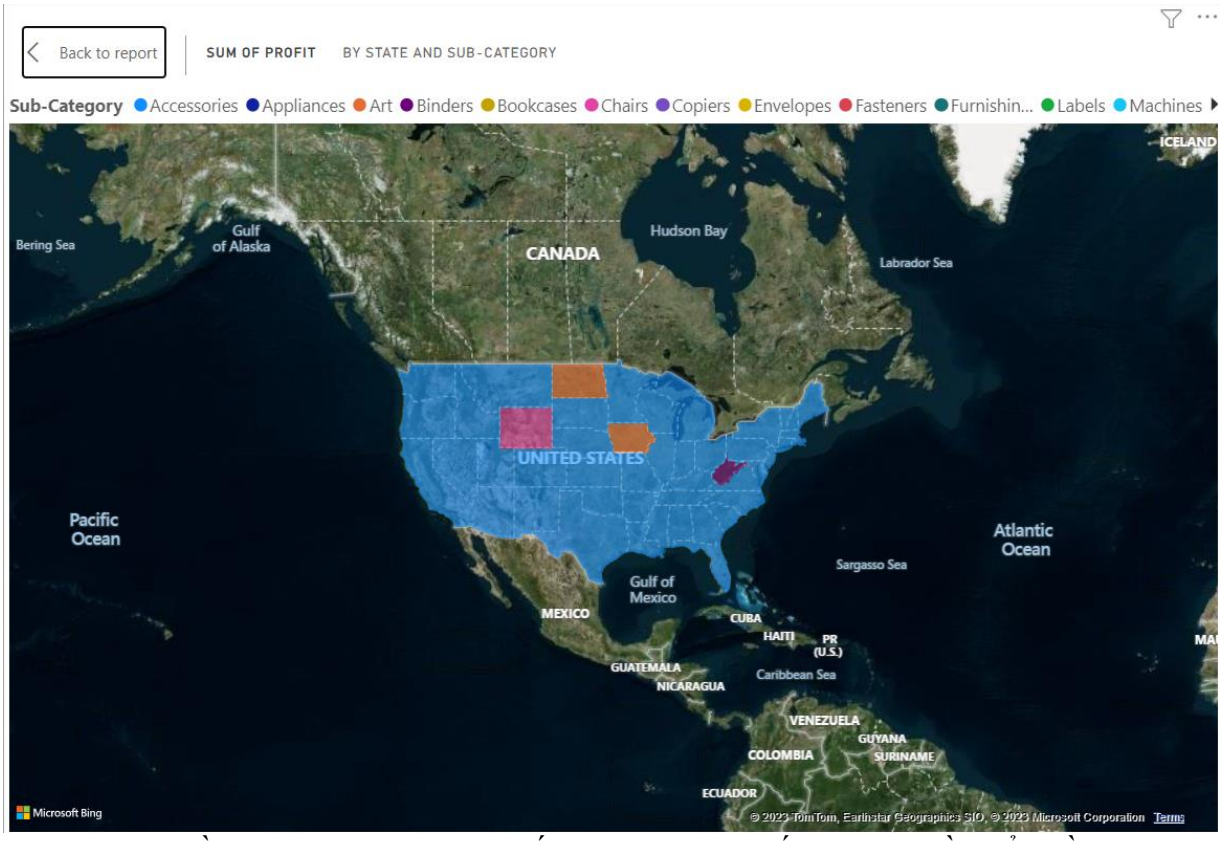



## 3.2. Product clustering




*Tổng quan của Product clustering*

### 3.2.1. Map – sum of Profit by State and Sub-Category



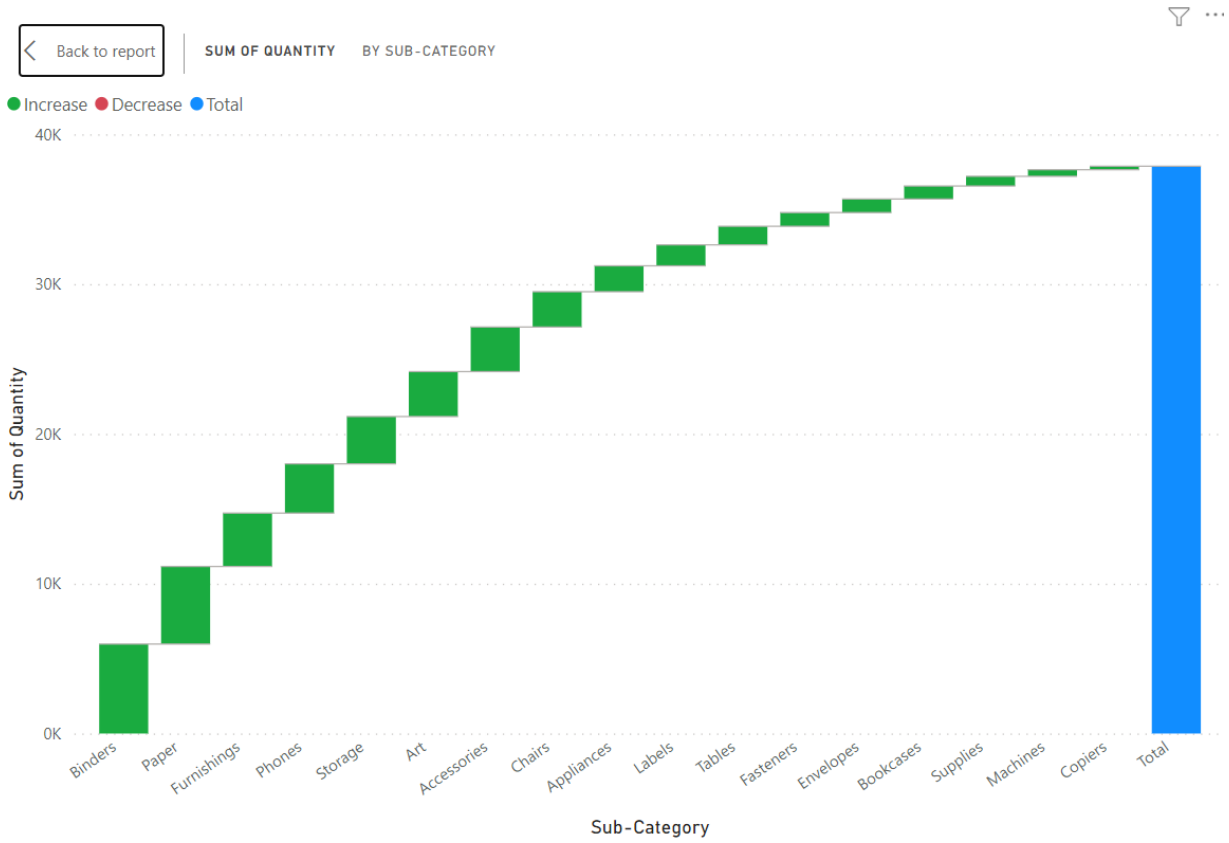

Bằng cách sử dụng màu sắc và các giá trị số trên bản đồ, biểu đồ này cho phép chúng ta so sánh mức độ lợi nhuận giữa các tiểu bang khác nhau. Nó có thể giúp chúng ta nhận ra các tiểu bang nào đóng góp nhiều lợi nhuận nhất hoặc ít lợi nhuận nhất đối với tổng thể, từ đó đưa ra các quyết định kinh doanh hoặc chiến lược tương ứng khác nhau vào các tiểu bang có lợi nhuận cao nhằm tối đa hóa hiệu quả kinh doanh.

Nhìn vào biểu đồ này cho ta thấy được lợi nhuận ‘Profit’ theo các ‘State’ và ‘Sub-category’ của sản phẩm. Tiểu bang nào đem lại lợi nhuận nhiều nhất và tiểu bang tiêu thụ từng loại sản phẩm riêng biệt nào.

Trong đó, bang IOWA và North Dakota chuộng các mặt hàng về Art hơn cả.

### 3.2.2. Waterfall chart – Sum of Quantity by Sub-Category

Biểu đồ được biểu diễn dưới dạng một loạt các thanh cột nằm dọc, trong đó mỗi cột đại diện cho một loại và độ dài của cột thể hiện giá trị số lượng tương ứng.

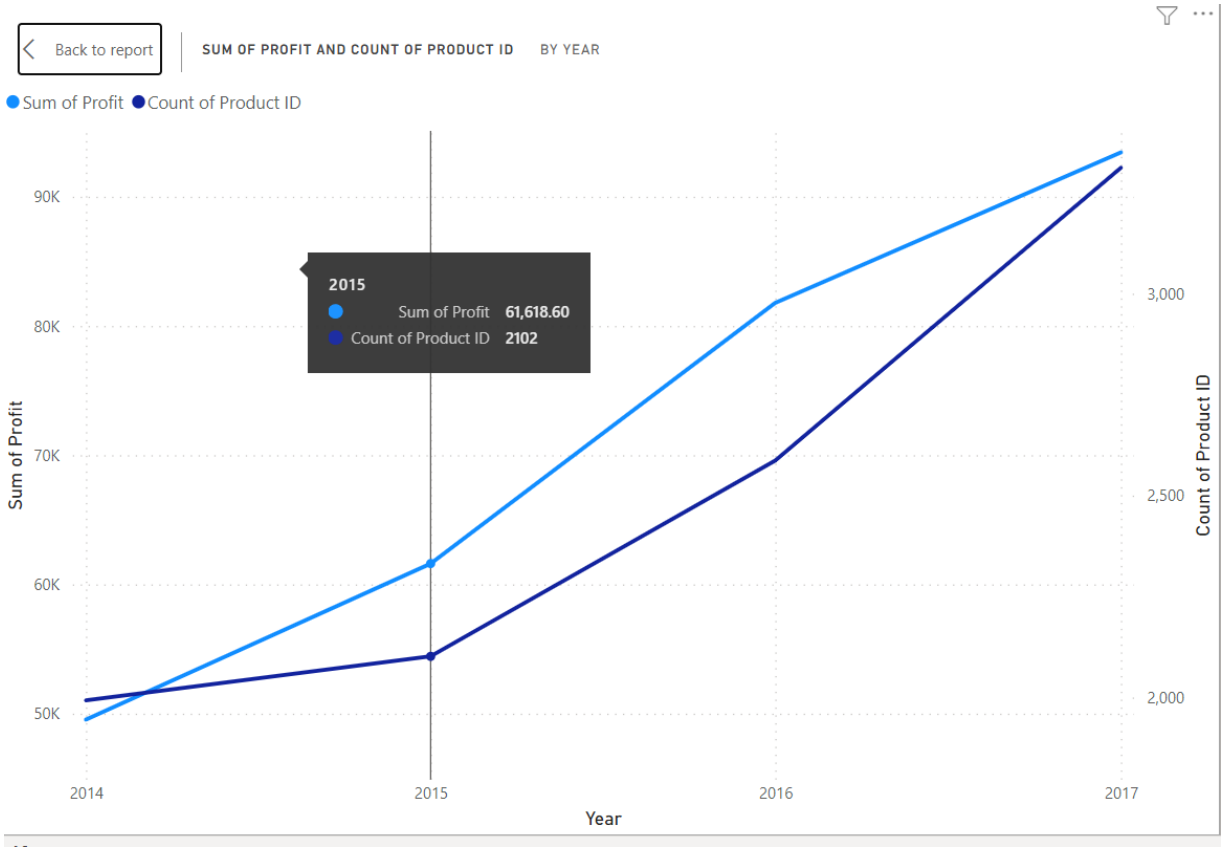



Nó giúp chúng ta nhìn thấy sự thay đổi trong tổng số lượng của từng loại và cách mỗi loại đóng góp vào sự thay đổi đó. Điều này có thể giúp chúng ta nhận ra những mục con tạo ra sự gia tăng hoặc giảm số lượng đáng kể và xác định nguyên nhân của sự thay đổi đó.

Từ đó đưa ra các quyết định kinh doanh liên quan đến quản lý và tối ưu hóa số lượng hàng hóa hoặc sản phẩm trong mỗi loại khác nhau.

Nhìn vào biểu đồ trên có thể thấy được số lượng các mặt hàng đã bán theo ‘Sub-Category’. Blinders là thể loại ‘Sub-Category’ đã bán ra với số lượng nhiều nhất. Ít nhất là Copiers.

### 3.2.3. Line chart – Sum of Profit and Count of Product by Year

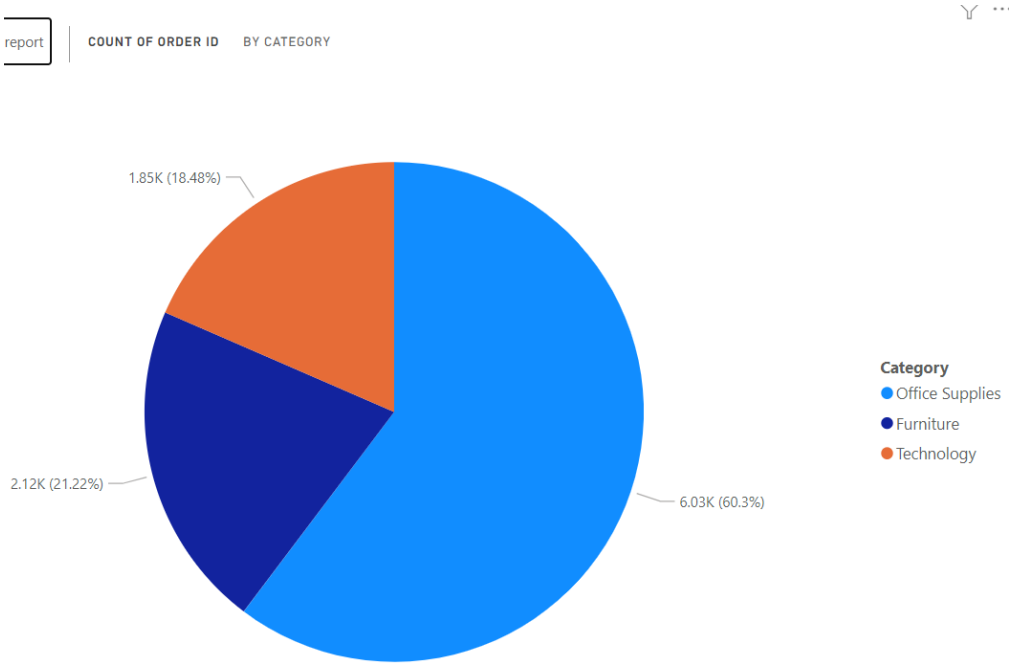



Bằng cách sử dụng đường cong nó cho phép chúng ta nhìn thấy sự biến đổi liên tục của tổng lợi nhuận và số lượng sản phẩm theo thời gian. Nếu đường cong tăng lên, điều này chỉ ra rằng tổng lợi nhuận và/hoặc số lượng sản phẩm đã tăng theo thời gian và ngược lại.

Biểu đồ Line giúp chúng ta nhận ra xu hướng dài hạn hoặc ngắn hạn cung cấp cái nhìn tổng quan về hiệu suất kinh doanh qua thời gian và có thể giúp chúng ta đưa ra các quyết định liên quan đến chiến lược kinh doanh và tối ưu hóa hiệu suất trong tương lai.

Qua biểu đồ trên, ta thấy được lợi nhuận ‘Profit’ và số lượng các loại ‘Product’ bán ra tăng trưởng trong giai đoạn 2014-2017.

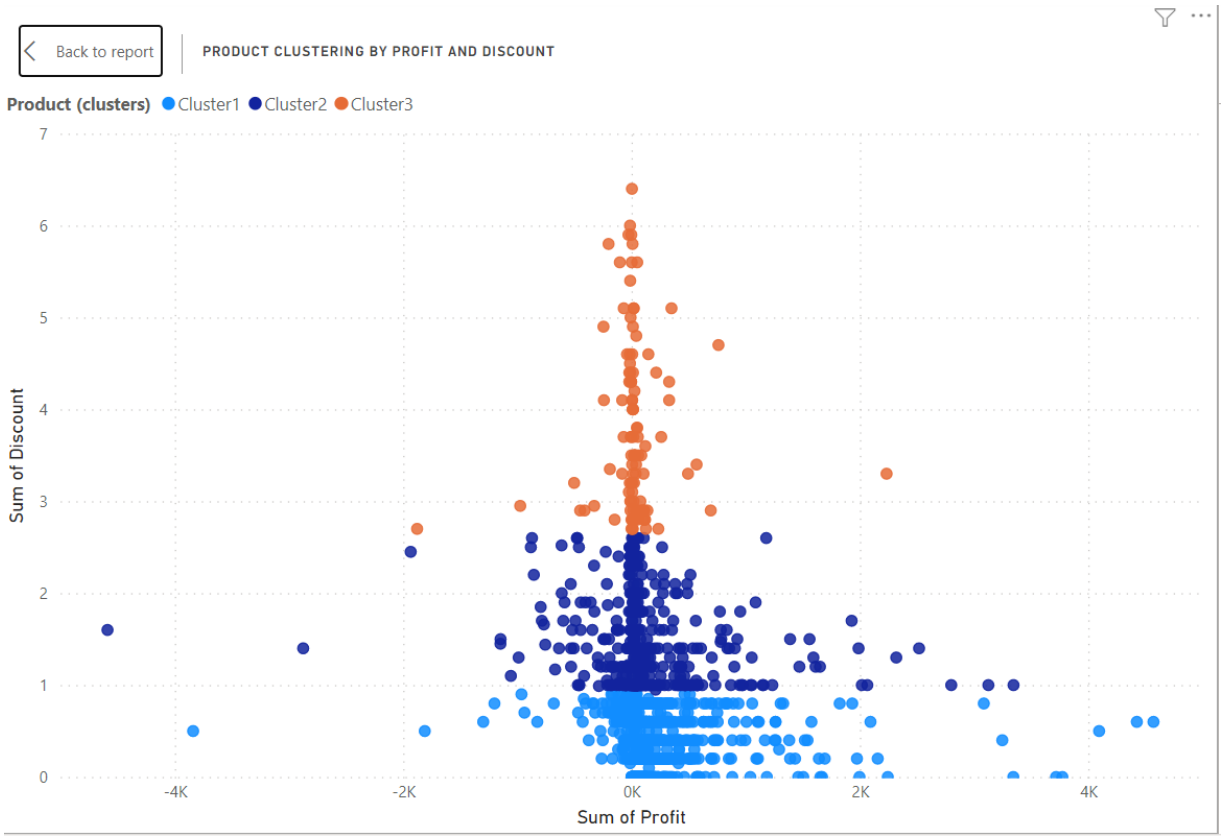
### 3.2.4. Pie chart – Count of Order by Category




Biểu đồ dùng hình tròn hiển thị các mục dưới dạng phần trăm giúp chúng ta nhìn thấy sự phân bố và quan trọng của mỗi mục đồng thời cũng cho phép chúng ta so sánh tỷ lệ đơn hàng giữa các mục khác nhau và nhận ra mục nào chiếm tỷ lệ lớn hơn hoặc nhỏ hơn trong số lượng đơn hàng. Giúp nhận định tốt các quyết định liên quan đến chiến lược kinh doanh và tiếp thị trong các danh mục đó.

Biểu đồ trên cho ta thấy tổng quan các đơn đặt hàng theo từng thể loại. Office Supplies là thể loại bán chạy nhất của cửa hàng trên.

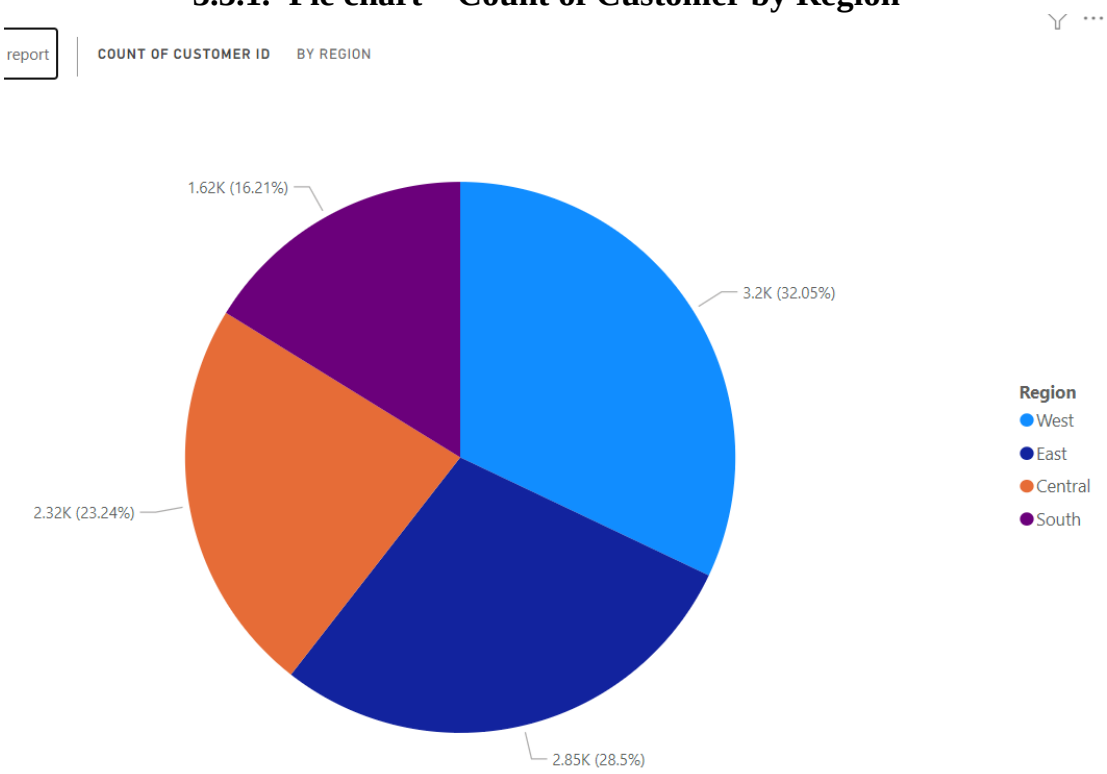
### 3.2.5. Scatter chart – Product clustering by Profit and Discount




Áp dụng phân cụm (clustering) sẵn có trong Power BI, chúng ta chia ra được 3 cụm sản phẩm. Cluster1 – sản phẩm có tổng Discount thấp (0-1) đem lại tổng lợi nhuận dương nhiều hơn âm và cũng là cụm có các sản phẩm đem lại tổng lợi nhuận cao nhất. Cluster2 – sản phẩm có tổng Discount trung bình (1-2.5) đem lại tổng lợi nhuận âm dương khá đều nhau. Cluster3 – sản phẩm có tổng Discount khá cao (2.5-6.5) nhưng lợi nhuận chỉ loanh quanh ở 0 (dường như là các sản phẩm xả hàng).

### 3.3. Customer clustering

#### 3.3.1. Pie chart – Count of Customer by Region




Biểu đồ này giúp chúng ta nhìn thấy phân bố và sự quan trọng của mỗi khu vực trong số lượng khách hàng. Nó cho phép chúng ta so sánh tỷ lệ khách hàng giữa các khu vực khác nhau và nhận ra khu vực nào có số lượng khách hàng lớn hơn hoặc nhỏ hơn.

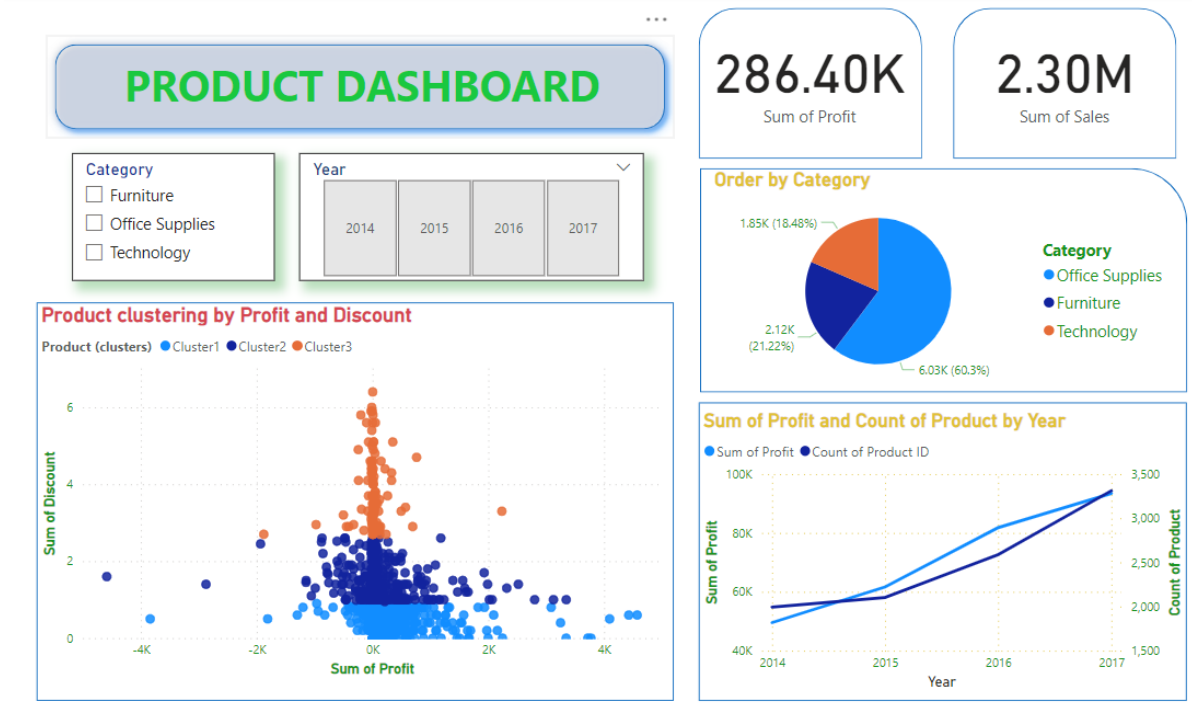
Biểu đồ Pie chart cung cấp cái nhìn tổng quan về phân phối khách hàng theo từng khu vực và có thể giúp chúng ta phân tích và đưa ra các quyết định liên quan đến chiến lược kinh doanh và tiếp thị trong các khu vực đó. Trên biểu đồ này cho ta thấy được số lượng khách hàng đến từng các khu vực khác nhau. Khách hàng đến từ Tây Mỹ chiếm số lượng lớn nhất và khách hàng Nam Mỹ chiếm số lượng ít nhất.

### 3.3.2. Scatter chart – Sum of Profit, Sum of Quantity and First **Category** by Customer

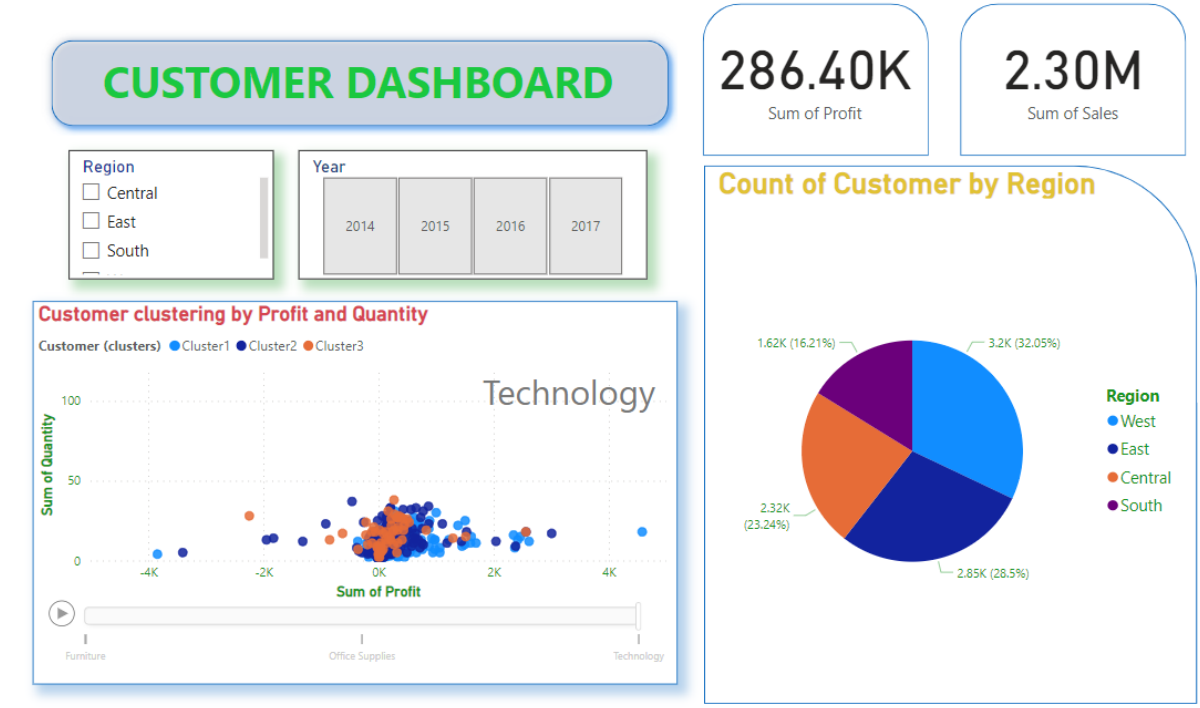

Áp dụng phân cụm (clustering) sẵn có trong Power BI, chúng ta chia ra được 3 cụm khách hàng. Cluster1 – khách hàng có tổng số lượng mua hàng ít (0-100) nhưng đem lại lợi nhuận nhiều nhất. Cluster2 – khách hàng có tổng số lượng mua hàng trung bình (25-125) đem lại lợi nhuận âm dương khá đều nhau. Cluster3 – khách hàng có tổng số lượng mua hàng khá cao (50-150) nhưng lợi nhuận đem lại không nhiều.

## 3.4. Mining dashboard

### 3.4.1. Dashboard: Product




### 3.4.2. Dashboard: Customer



Có thể tương tác theo với dashboard theo từng thể loại sản phẩm (category), năm (year) và vùng lãnh thổ (region).

## 4. Kết luận

Khai phá dữ liệu trên dữ liệu Superstore cho ta thấy được cái nhìn tổng quan về các mặt hàng bán chạy của cửa hàng điện tử, phân vùng khách hàng.

Các mặt hàng về công nghệ (Technology) ít có các chương trình giảm giá, ít đơn đặt hàng nhưng lợi nhuận đem về cho từng đơn hàng khá cao so với 2 loại mặt hàng còn lại (Office Supplies và Furniture). Các loại hình thức giao hàng cũng phản ánh được mức độ lợi nhuận đem lại cho từng hình thức. Từ đó, chúng ta có thể đưa ra các chiến lược kinh doanh, giao hàng phù hợp để cửa hàng superstore ngày càng phát triển và thịnh vượng.

Các khách hàng mua hàng trên cửa hàng điện tử phân phối khá đồng đều giữa các vùng với nhau (West, East, Central and South).

# TÀI LIỆU THAM KHẢO


1. Cluster Analysis in Power BI. InterationInsights. 2022. https://iterationinsights.com/article/cluster-analysis-in-power-bi/
2. Implement Clustering in Power BI. Deepika Singh. 2020. https://www.pluralsight.com/guides/implement-clustering-in-powerbi
3. Power BI Project End to End. Data Tutorials. 2023. https://www.youtube.com/watch?v=Hn9f13uoLAQ
4. Khóa học 'Khai phá dữ liệu'. Thầy Nguyễn Văn Thành. 2023.
5. Superstore Dataset. Vivek Chowdhury. 2022. https://www.kaggle.com/datasets/vivek468/superstore-dataset-final
6. Association Rule Mining in SQL Server. Dinesh Asanka. 2020. https://www.sqlshack.com/the-association-rule-mining-in-sql-server/
7. Microsolf Decision Trees in SQL Server. Dinesh Asanka. 2019. https://www.sqlshack.com/microsoft-decision-trees-in-sql-server/