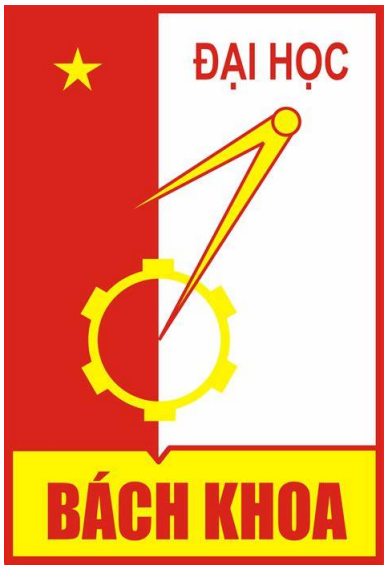
ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN-TIN

——————o0o——————



ĐỒ ÁN II

XÂY DỰNG HỆ THỐNG PHÂN TÍCH DỮ LIỆU

Chuyên ngành: Hệ thống thông tin quản lý

Bộ môn: Toán Tin

Giảng viên hướng dẫn: ThS. Lê Kim Thư _____

Chữ kí GVHD

Sinh viên thực hiện: Nguyễn Thị Quỳnh

Mã số sinh viên: 20206301

HÀ NỘI, 01/2024

Nhận xét của giảng viên hướng dẫn

1. Mục tiêu và nội dung của đồ án

(a) Mục tiêu:

(b) Nội dung:

2. Kết quả đạt được

3. Ý thức làm việc của sinh viên:

Hà Nội, ngày ... tháng 1 năm 2024

Giảng viên hướng dẫn

ThS. LÊ KIM THƯ

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

CHƯƠNG 1. MỞ ĐẦU

1.1 Tổng quan đề tài

1.1.1 Lí do chọn đề tài

Dữ liệu đang trở thành một tài sản quan trọng đối với các tổ chức và cá nhân. Theo một báo cáo của Gartner, trong năm 2022, lượng dữ liệu được tạo ra trên toàn thế giới sẽ đạt 85 zettabyte[1]. Phân tích dữ liệu là cách tiếp cận tốt nhất để hiểu hơn về khách hàng, các đối thủ cạnh tranh và nắm bắt được thị trường.

Các công nghệ và kĩ thuật đang dần được tìm tòi và cập nhật không ngừng, mở ra những năng mới trong phân tích dữ liệu. Chính vì vậy lĩnh vực phân tích dữ liệu trở nên quan trọng và đầy tiềm năng để nghiên cứu và phát triển.

Trong môi trường kinh doanh này, một thách thức được đặt ra với các doanh nghiệp bán lẻ để hiểu rõ hơn về khách hàng, nhu cầu đối với các loại mặt hàng và xác định cơ hội mới trong thị trường đầy cạnh tranh và biến động.

Chính vì vậy quản lý dữ liệu người tiêu dùng được quan tâm hàng đầu, một hệ thống phân tích dữ liệu bán lẻ trở thành yếu tố quan trọng giúp doanh nghiệp thích ứng và phát triển đóng vai trò quyết định trong sự thành công của doanh nghiệp.

1.1.2 Mục tiêu và phạm vi đề tài

Xây dựng hệ thống quản lí và lưu trữ dữ liệu theo mô hình cơ sở dữ liệu đa chiều, đưa ra các phân tích về dữ liệu qua các báo cáo.

Hệ thống được xấy dựng phục vụ cho thị trường bán lẻ, cụ thể tập trung và một doanh nghiệp bán lẻ.

1.1.3 Bố cục đồ án

Phần còn lại của báo cáo được tổ chức như sau:

- Chương 2: Cơ sở lý thuyết - trình bày các khái niệm, công cụ sử dụng trong quá trình làm việc.
- Chương 3: Phân tích thiết kế hệ thống - phân tích các qúa trình của hệ thống tiến tới thiết kế hệ thống.
- Chương 4: Xây dựng chương trình - xây dựng các chương trình làm việc với dữ liệu của hệ thống
- Chương 5: Tổng kết - trình bày kết luận và hướng phát triển của đồ án.

1.2 Lời cảm ơn

Lời đầu tiên, em xin được gửi lời cảm ơn đến thầy cô, anh chị em, bạn bè xung quanh đã hỗ trợ, giúp đỡ em trong suốt quá trình làm Đồ án.

Tiếp đến, cho phép em được gửi lời cảm ơn đến các thầy cô khoa Toán Tin, Đại học Bách khoa Hà Nội đã truyền đạt cho em những kiến thức, kinh nghiệm để em hoàn thiện bản thân hơn mỗi ngày.

Đặc biệt, em xin gửi lời cảm ơn đến ThS. Lê Kim Thư, người đã tận tâm hướng dẫn, chỉ bảo em trong suốt thời gian qua.

Em xin chân thành cảm ơn!!

Hà Nội, ngày ... tháng 1 năm 2024
Tác giả đồ án

Nguyễn Thị Quỳnh

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Các khái niệm cơ bản

2.1.1 Phân tích dữ liệu

Phân tích dữ liệu là quá trình khám phá và sử dụng dữ liệu để đưa ra thông tin và kiến thức có giá trị[2].

Khái niệm này nhấn mạnh hai khía cạnh chính của phân tích dữ liệu:

- Khám phá dữ liệu: Phân tích dữ liệu bao gồm việc khám phá dữ liệu để tìm hiểu về dữ liệu đó. Điều này có thể bao gồm việc xác định các xu hướng, mẫu và mối quan hệ trong dữ liệu.
- Sử dụng dữ liệu: Phân tích dữ liệu cũng bao gồm việc sử dụng dữ liệu để đưa ra thông tin và kiến thức có giá trị. Điều này có thể bao gồm việc đưa ra dự đoán, giải quyết vấn đề hoặc hiểu rõ hơn về một vấn đề

Các kỹ thuật phân tích dữ liệu có thể bao gồm phân tích thống kê, học máy (machine learning), khai phá dữ liệu (data mining) và các phương pháp khác.

2.1.2 Phân tích nghiệp vụ

Phân tích nghiệp vụ là một quá trình khám phá, hiểu và giao tiếp các yêu cầu của các hệ thống kinh doanh[3]

- Khám phá nhu cầu: Hiểu nhu cầu của các bên liên quan.
- Phân tích nhu cầu: Xác định các yêu cầu của hệ thống.
- Giao tiếp nhu cầu: Mô tả các yêu cầu của hệ thống cho các bên liên quan.

Phân tích nghiệp vụ thường tập trung vào việc thiết kế, tối ưu hóa và quản lý quy trình làm việc, và có thể kết hợp với phân tích dữ liệu để hỗ trợ quyết định kinh doanh.

2.1.3 Kinh doanh thông minh

Kinh doanh thông minh (BI) là một tập hợp các công nghệ, quy trình và chiến lược được sử dụng để thu thập, phân tích và trình bày dữ liệu kinh doanh để hỗ trợ ra quyết định[4].

Nó liên quan đến việc sử dụng trí tuệ nhân tạo (AI), học máy và dữ liệu để tạo ra giá trị và cải thiện khả năng dự đoán, quyết định và tương tác với khách hàng.

Kinh doanh thông minh thường kết hợp cả phân tích dữ liệu và phân tích nghiệp vụ để tạo ra hệ thống thông minh giúp doanh nghiệp hiểu và phản ứng nhanh hơn với thay đổi trong môi trường kinh doanh.

2.2 Các công cụ sử dụng

2.2.1 Ngôn ngữ lập trình Python

Python là một ngôn ngữ lập trình phổ biến trong lĩnh vực khoa học dữ liệu và phân tích dữ liệu.

Vai trò:

- Python cung cấp các thư viện mạnh mẽ cho phân tích dữ liệu, như Pandas, NumPy, và Matplotlib, giúp bạn thao tác dữ liệu, tính toán thống kê, và biểu đồ hóa dữ liệu.
- Có thể sử dụng Python để thực hiện phân tích dữ liệu một cách tùy chỉnh, xây dựng các mô hình học máy, và giải quyết các vấn đề đặc thù của dự án phân tích.

2.2.2 MySQL (Structured Query Language)

SQL là một ngôn ngữ dùng để truy vấn và quản lý cơ sở dữ liệu quan hệ.

Vai trò:

- SQL được sử dụng để trích xuất dữ liệu từ cơ sở dữ liệu, thực hiện các truy vấn phức tạp, và kết hợp dữ liệu từ nhiều bảng.
- Các câu lệnh SQL giúp làm sạch, biến đổi, và tổng hợp dữ liệu trước khi nó được sử dụng cho phân tích. SQL là công cụ quan trọng trong việc làm việc với cơ sở dữ liệu lớn.

2.2.3 Power BI

Power BI là một công cụ Business Intelligence của Microsoft, giúp kết nối, biểu đồ hóa, và hiển thị dữ liệu một cách trực quan.

Vai trò:

- Power BI cho phép bạn kết nối đến nhiều nguồn dữ liệu, biểu đồ hóa dữ liệu dễ dàng và tạo báo cáo tương tác.
- Nó cung cấp khả năng tự động hóa quá trình phân tích và cung cấp thông báo kinh doanh cùng với khả năng tạo các trang tổng hợp dữ liệu.

2.3 Thông tin về bộ dữ liệu

2.3.1 Bộ dữ liệu

Bộ dữ liệu Dunnhumby - The Complete Journey chứa các giao dịch của các hộ gia đình thường xuyên mua sắm tại một nhà bán lẻ trên nhiều cửa hàng khác nhau. Nó chứa tất cả các khoản mua hàng của mỗi hộ gia đình. Đi kèm với đó là thông tin về nhân khẩu học của từng hộ gia đình, thông tin về các mặt hàng, các chiến dịch, mã giảm giá cũng như các phương thức quảng cáo đều được ghi lại[5].

Bộ dữ liệu có kích thước 847 MB gồm 8 bảng tương ứng với 46 trường dữ liệu:

1. Bảng campaign_desc(540 B)

Bảng campaign_desc chứa thông tin về 30 chiến dịch được tổ chức. Các chiến dịch này có thể trùng thời gian với nhau.

- DESCRIPTION: Loại chiến dịch
- CAMPAIGN: Chiến dịch
- START_DAY: Ngày bắt đầu chiến dịch

- END_DAY: Ngày kết thúc chiến dịch

2. Bảng campaign_table(95.87 kB)

Bảng campaign_table chứa thông tin về các hộ gia đình tham gia vào các chiến dịch cụ thể.

- DESCRIPTION: Loại chiến dịch
- HOUSEHOLD_KEY: Mã hộ gia đình
- CAMPAIGN: Chiến dịch

3. Bảng causal_data(695.86 MB)

Bảng causal_data chứa thông tin về các lượt tiếp thị trên các vị trí trong cửa hàng và hình quảng cáo của một vài sản phẩm theo từng tuần ứng với mỗi cửa hàng.

- PRODUCT_ID: Mã sản phẩm
- STORE_ID: Mã cửa hàng
- WEEK_NO: Tuần
- DISPLAY: Vị trí trong cửa hàng
- MAILER: Vị trí trên quảng cáo

4. Bảng coupon (2.82 MB)

Bảng coupon chứa thông tin về các mã giảm giá của một vài sản phẩm được áp dụng ứng với từng chiến dịch. Một sản phẩm có thể sử đụng được một hoặc nhiều mã giảm giá. Các mã giảm giá ứng với từng sản phẩm có thể lặp lại trong các chiến dịch khác nhau.

- COUPON_UPC: Mã giảm giá
- PRODUCT_ID: Mã sản phẩm được áp dụng
- CAMPAIGN: Chiến dịch

5. Bảng coupon_redempt (54.11 kB)

Bảng coupon_redempt chứa thông tin chi tiết về việc sử dụng mã giảm giá trong từng chiến dịch của từng hộ gia đình.

- hOUSEHOLD_KEY: Mã hộ gia đình
- DAY: Ngày sử dụng mã giảm giá
- COUPON_UPC: Mã giảm giá được sử dụng
- CAMPAIGN: Chiến dịch

6. Bảng hh_demographic (44.35 kB)

Bảng hh_demographic chứa thông tin tiết về nhân khẩu học của từng hộ gia đình.

- AGE_DESC: Độ tuổi

- MARITAL_STATUS_CODE: Tình trạng hôn nhân
- INCOME_DESC: Mức thu nhập
- HOMEOWNER_DESC: Loại hộ gia đình
- HH_COMP_DESC: Thành phần hộ gia đình
- HOUSEHOLD_SIZE_DESC: Quy mô hộ gia đình
- KID_CATEGORY_DESC: Số lượng trẻ em
- HOUSEHOLD_KEY: Mã định danh hộ gia đình

7. Bảng product (6.43 MB)

Bảng product chứ thông tin chi tiết về các sản phẩm được bày bán trên hệ thống.

- PRODUCT_ID: Mã định danh sản phẩm
- MANUFACTURER: Mã nhà sản xuất
- DEPARTMENT: Chủng loại sản phẩm
- BRAND: Thương hiệu
- COMMODITY_DESC: Nhóm sản phẩm, phân loại thấp hơn của chủng loại.
- SUB_COMMODITY_DESC: Sản phẩm
- CURR_SIZE_OF_PRODUCT: Kích cỡ sản phẩm

8. Bảng transaction (141.74 MB)

Bảng transaction ghi lại thông tin chi tiết về các giao dịch của từng hộ gia đình trên từng cửa hàng.

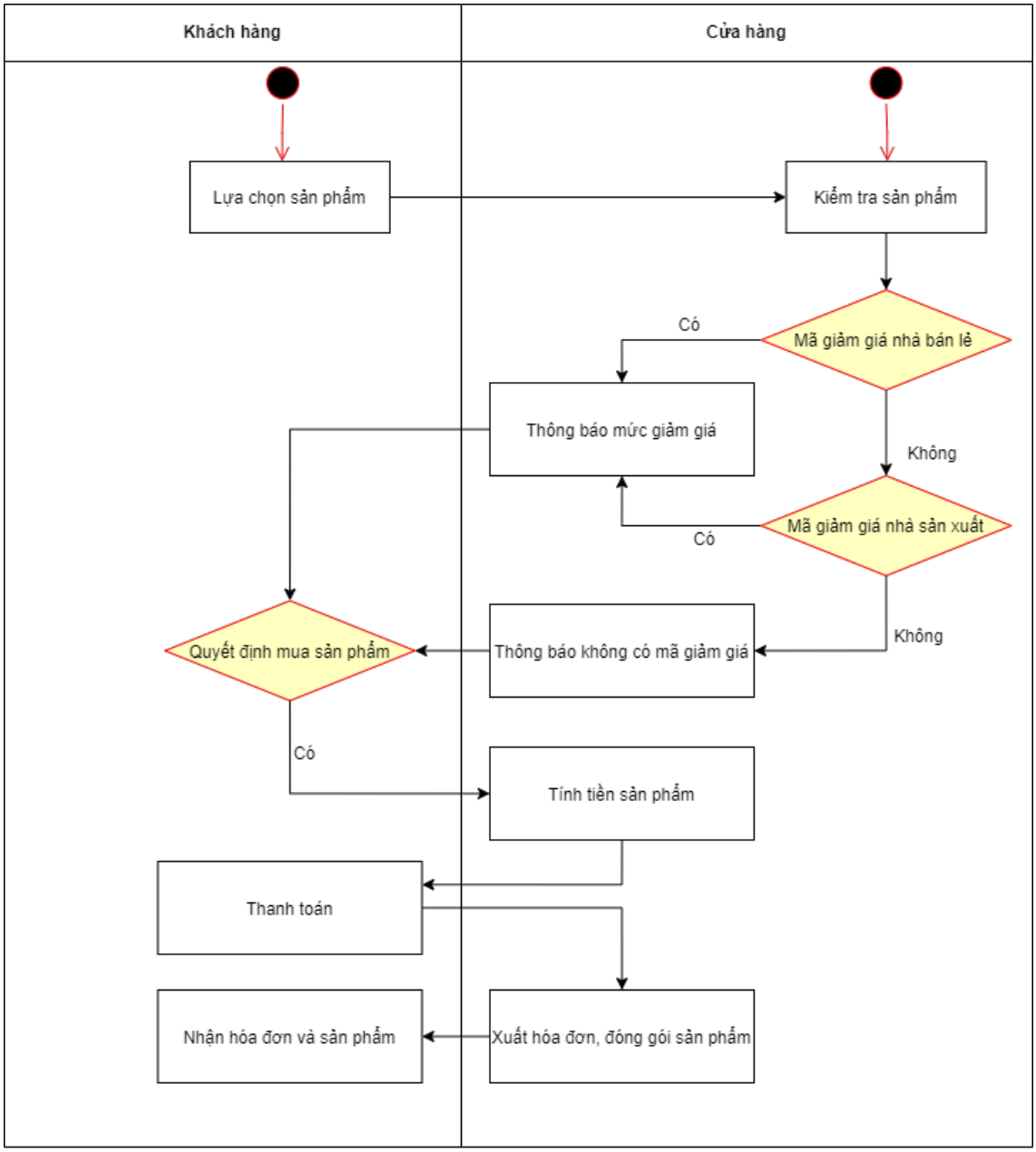
Một lần mua sắm của khách hàng ứng với một giỏ hàng gồm nhiều giao dịch, mỗi giao dịch được tính dựa trên một loại sản phẩm với số lượng bất kì của khách hàng mua.

Một giao dịch có thể sử dụng nhiều mã giảm giá của một hay nhiều chiến dịch.

- HOUSEHOLD_KEY: Mã hộ gia đình
- BASKET_ID: Mã giỏ hàng
- DAY: Ngày giao dịch
- PRODUCT_ID: Mã sản phẩm
- QUANTITY: Số lượng
- SALES_VALUE: Giá tiền khách hàng thanh toán
- STORE_ID: Mã cửa hàng
- RETAIL_DISC: Giảm giá nhà bán lẻ
- TRANS_TIME: Thời gian giao dịch
- WEEK_NO: Tuần

- COUPON_DISC: Giảm giá của nhà sản xuất
- COUPON_MATCH_DISC: Giảm giá nhà bán lẻ khớp với nhà sản xuất

2.3.2 Quy trình nghiệp vụ



Hình 2.1: Quy trình nghiệp vụ

Trong quy trình trên gồm 2 đối tượng là khách hành và nhân viên cửa hàng:

- Khách hàng thực hiện các hoạt động lựa chọn sản phẩm rồi hành thanh toán nếu quyết định mua hàng sau đó nhận hóa đơn để hoàn tất.
- Nhân viên cửa hàng thực hiện kiểm tra mã giảm giá của nhà bán lẻ và nhà sản xuất, tiến hành tính tiền khi khách hàng đồng ý và xuất hóa đơn khi hoàn thành.

CHƯƠNG 3. PHÂN TÍCH THIẾT KẾ HỆ THỐNG

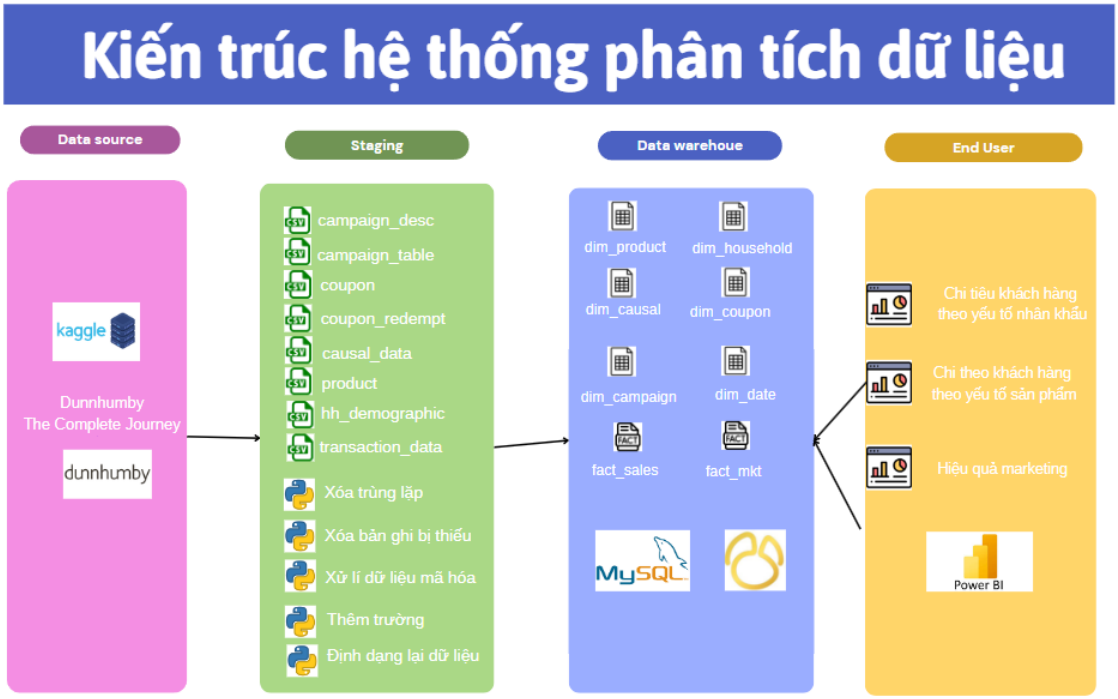
3.1 Kiến trúc hệ thống

Thiết kế hệ thống phân tích dữ liệu là quá trình xác định các thành phần và giao diện của một hệ thống phân tích dữ liệu, cũng như các mối quan hệ giữa chúng[6].

Thiết kế kiến trúc hệ thống là một bước quan trọng và cần thiết trong quá trình xây dựng hệ thống thông tin[7].

Thiết kế hệ thống mang lại các lợi ích sau:

- Định hình cấu trúc hệ thống sẽ xây dựng
- Dễ dàng quản lí hệ thống
- Phân chia công việc phù hợp cho các bộ phận.



Hình 3.1: Kiến trúc hệ thống phân tích dữ liệu

Hệ thống phân tích dữ liệu được thiết kế có kiến trúc 4 lớp như sau:

Data Source: Nguồn gốc của dữ liệu phân tích. Dữ liệu được sử dụng trong hệ thống được lấy từ website Kaggke: "Dumnhumby - The Complete Journey"

Staging: Vùng tập kết dữ liệu. Trong phần này sử dụng ngôn ngữ lập trình Python với thư viện Pandas để tiến hành xử lí, làm sạch dữ liệu.

Data Warehouse: Khu vực tổ chức, lưu trữ dữ liệu theo cấu trúc các bảng dim - fact. Sử dụng ngôn ngữ truy vấn MySQL để thiết kế, tổ chức dim - fact và lưu trữ dữ liệu.

End Use: Khu vực khai thác dữ liệu vào sử dụng với các mục đích tạo báo cáo, phân

tích. Phần mềm Power BI hỗ trợ trực quan hóa dữ liệu bằng các biểu đồ tạo thành các báo cáo dựa trên phân tích.

3.2 Xử lí dữ liệu

3.2.1 Quy trình xử lí dữ liệu

Xử lý dữ liệu (Extract - Transform - Load) viết tắt là ETL. ETL là quá trình thu thập, biến đổi và lưu trữ, quản lý dữ liệu[8]

Quá trình ETL, dữ liệu:

Hình 3.2: Quá trình ETL dữ liệu

Trong đó:

- E-Extract: Đọc dữ liệu từ nguồn cơ sở dữ liệu.
- T-Transform: Chuyển đổi, biến đổi dữ liệu để phù hợp với cấu trúc và định dạng của hệ thống đích, bao gồm việc làm sạch dữ liệu, chuẩn hóa định dạng. Một số thao tác trong chuyển đổi dữ liệu như: data cleansing, chuẩn hóa, loại bỏ trùng lặp, sắp xếp,...
- L-Load: Dữ liệu đã được biến đổi được tải vào hệ thống đích, thường là một kho dữ liệu hoặc hệ thống quản lý cơ sở dữ liệu. Quá trình tải có thể được thực hiện một lần (tải toàn bộ) hoặc theo định kỳ (tải tăng phần).

ETL là một bộ phận không thể thiếu trong quản lý dữ liệu, giúp tổ chức dữ liệu một cách hiệu quả, chuẩn bị dữ liệu cho phân tích, và hỗ trợ quyết định dựa trên dữ liệu:

- Tập trung Dữ liệu
- Cải Thiện Chất Lượng Dữ liệu
- Hiệu Quả về Thời Gian
- Hỗ Trợ Quyết Định Kinh Doanh
- Bảo Mật và Tuân Thủ Pháp Luật

3.2.2 Xử lí dữ liệu trên bộ dữ liệu

Trong phần xử lí dữ liệu này, em sử dụng thư viện pandas của ngôn ngữ lập trình Python trong phần mềm Colab Notebook, Visual Code để tiến hành các hoạt động ETL

như :

- Xóa dữ liệu sai (null/trống)
- Định dạng lại kiểu dữ liệu
- Mã hóa dữ liệu thành nội dung có nghĩa
- Thêm các trường thuộc tính cần thiết.

Cụ thể hơn, các quá trình xử lí được thực hiện như sau:

1. Import dữ liệu

Trong phần này ta đẩy dữ liệu vào phần mềm, sử dụng thư viện pandas để đọc các file.

2. Lấy ra thông tin dữ liệu

index	Table	Number of Rows	Number of Columns	Null Values	Data Types	Memory Usage	duplicate of rows	blank values
0	df_campaign_desc	30	4	0	object,int64,int64,int64	1088	0	{'DESCRIPTION': 0, 'CAMPAIGN': 0, 'START_DAY': 0, 'END_DAY': 0}
1	df_campaign_table	7208	3	0	object,int64,int64	173120	0	{'DESCRIPTION': 0, 'household_key': 0, 'CAMPAIGN': 0}
2	df_causal_data	36786524	5	0	int64,int64,int64,object,object	1471461088	0	{'PRODUCT_ID': 0, 'STORE_ID': 0, 'WEEK_NO': 0, 'display': 0, 'mailer': 0}
3	df_coupon	124548	3	0	int64,int64,int64	2989280	5164	{'COUPON_UPC': 0, 'PRODUCT_ID': 0, 'CAMPAIGN': 0}
4	df_coupon_redempt	2318	4	0	int64,int64,int64,int64	74304	0	{'household_key': 0, 'DAY': 0, 'COUPON_UPC': 0, 'CAMPAIGN': 0}
5	df_hh_demographic	801	8	0	object,object,object,object,object,object,object,int64	51392	0	{'AGE_DESC': 0, 'MARITAL_STATUS_CODE': 0, 'INCOME_DESC': 0, 'HOMEOWNER_DESC': 0, 'HH_COMP_DESC': 0, 'HOUSEHOLD_SIZE_DESC': 0, 'KID_CATEGORY_DESC': 0, 'household_key': 0}
6	df_product	92353	7	0	int64,int64,object,object,object,object,object	5171896	0	{'PRODUCT_ID': 0, 'MANUFACTURER': 0, 'DEPARTMENT': 15, 'BRAND': 0, 'COMMODITY_DESC': 15, 'SUB_COMMODITY_DESC': 15, 'CURR_SIZE_OF_PRODUCT': 30607}
7	df_transaction	2595732	12	0	int64,int64,int64,int64,int64,float64,int64,float64,int64,int64,float64,float64	249190400	0	{'household_key': 0, 'BASKET_ID': 0, 'DAY': 0, 'PRODUCT_ID': 0, 'QUANTITY': 0, 'SALES_VALUE': 0, 'STORE_ID': 0, 'RETAIL_DISC': 0, 'TRANS_TIME': 0, 'WEEK_NO': 0, 'COUPON_DISC': 0, 'COUPON_MATCH_DISC': 0}

Hình 3.3: Thông tin dữ liệu trước khi xử lí

Khảo sát sơ bộ về dữ liệu ta thấy xuất hiện các vấn đề cần xử lí: Bảng coupon (mã giảm giá) có chứa 5164 bản ghi bị trùng lặp Bảng product (sản phẩm) ở các trường thuộc tính 'DEPARTMENT', 'COMMODITY_DESC','SUB_COMMODITY_DESC' có 15 giá trị trống, ở thuộc tính 'CURR_SIZE_OF_PRODUCT' có 30607 giá trị trống

3. ETL bảng product

- Xóa các dòng có giá trị trống trong các cột 'DEPARTMENT', 'COMMODITY_DESC', 'SUB_COMMODITY_DESC'
- Điền giá trị 'Unknown' vào cột CURR_SIZE_OF_PRODUCT cho các dòng có giá trị trống

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 92353 entries, 0 to 92352
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   PRODUCT_ID            92353 non-null  int64
 1   MANUFACTURER          92353 non-null  int64
 2   DEPARTMENT            92353 non-null  object
 3   BRAND                 92353 non-null  object
 4   COMMODITY_DESC        92353 non-null  object
 5   SUB_COMMODITY_DESC    92353 non-null  object
 6   CURR_SIZE_OF_PRODUCT  92353 non-null  object
dtypes: int64(2), object(5)
memory usage: 4.9+ MB
```

Hình 3.4: Thông tin bảng product trước xử lí

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 92338 entries, 0 to 92352
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   PRODUCT_ID            92338 non-null  int64
 1   MANUFACTURER          92338 non-null  int64
 2   DEPARTMENT            92338 non-null  object
 3   BRAND                 92338 non-null  object
 4   COMMODITY_DESC        92338 non-null  object
 5   SUB_COMMODITY_DESC    92338 non-null  object
 6   CURR_SIZE_OF_PRODUCT  92338 non-null  object
dtypes: int64(2), object(5)
memory usage: 5.6+ MB
```

Hình 3.5: Thông tin bảng product sau xử lí

	PRODUCT_ID	MANUFACTURER	DEPARTMENT	BRAND	COMMODITY_DESC	SUB_COMMODITY_DESC	CURR_SIZE_OF_PRODUCT
0	25671	2	GROCERY	National	FRZN ICE	ICE - CRUSHED/CUBED	22 LB
1	26081	2	MISC. TRANS.	National	NO COMMODITY DESCRIPTION	NO SUBCOMMODITY DESCRIPTION	
2	26093	69	PASTRY	Private	BREAD	BREAD:ITALIAN/FRENCH	
3	26190	69	GROCERY	Private	FRUIT - SHELF STABLE	APPLE SAUCE	50 OZ
4	26355	69	GROCERY	Private	COOKIES/CONES	SPECIALTY COOKIES	14 OZ

Hình 3.6: Dữ liệu bảng product trước khi xử lí

	PRODUCT_ID	MANUFACTURER	DEPARTMENT	BRAND	COMMODITY_DESC	SUB_COMMODITY_DESC	CURR_SIZE_OF_PRODUCT	PRODUCT_CATEGORY
0	25671	2	GROCERY	National	FRZN ICE	ICE - CRUSHED/CUBED	22 LB	Food
1	26081	2	MISC. TRANS.	National	NO COMMODITY DESCRIPTION	NO SUBCOMMODITY DESCRIPTION	Unknown	Services
2	26093	69	PASTRY	Private	BREAD	BREAD:ITALIAN/FRENCH	Unknown	Food
3	26190	69	GROCERY	Private	FRUIT - SHELF STABLE	APPLE SAUCE	50 OZ	Food
4	26355	69	GROCERY	Private	COOKIES/CONES	SPECIALTY COOKIES	14 OZ	Food

Hình 3.7: Dữ liệu bảng product sau khi xử lí

4. ETL bảng "hh_demographic"

- Thay thế các giá trị A, B, U bằng Married, Single, Unknown

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 801 entries, 0 to 800
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   AGE_DESC             801 non-null    object
 1   MARITAL_STATUS_CODE  801 non-null    object
 2   INCOME_DESC          801 non-null    object
 3   HOMEOWNER_DESC       801 non-null    object
 4   HH_COMP_DESC         801 non-null    object
 5   HOUSEHOLD_SIZE_DESC  801 non-null    object
 6   KID_CATEGORY_DESC    801 non-null    object
 7   household_key        801 non-null    int64
dtypes: int64(1), object(7)
memory usage: 50.2+ KB
```

Hình 3.8: Thông tin bảng hh_demographic trước xử lí

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 801 entries, 0 to 800
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   AGE_DESC             801 non-null    object
 1   MARITAL_STATUS_CODE  801 non-null    object
 2   INCOME_DESC          801 non-null    object
 3   HOMEOWNER_DESC       801 non-null    object
 4   HH_COMP_DESC         801 non-null    object
 5   HOUSEHOLD_SIZE_DESC  801 non-null    object
 6   KID_CATEGORY_DESC    801 non-null    object
 7   household_key        801 non-null    int64
dtypes: int64(1), object(7)
memory usage: 50.2+ KB
```

Hình 3.9: Thông tin bảng hh_demographic sau xử lí

	AGE_DESC	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC	household_key
0	65+	A	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	1
1	45-54	A	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	7
2	25-34	U	25-34K	Unknown	2 Adults Kids	3	1	8
3	25-34	U	75-99K	Homeowner	2 Adults Kids	4	2	13
4	45-54	B	50-74K	Homeowner	Single Female	1	None/Unknown	16

Hình 3.10: Dữ liệu bảng hh_demographic trước khi xử lí

	AGE_DESC	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC	household_key	Income_Level
0	65+	Married	35-49K	Homeowner	2 Adults No Kids	2	None/Unknown	1	Low
1	45-54	Married	50-74K	Homeowner	2 Adults No Kids	2	None/Unknown	7	Medium
2	25-34	Unknown	25-34K	Unknown	2 Adults Kids	3	1	8	Low
3	25-34	Unknown	75-99K	Homeowner	2 Adults Kids	4	2	13	Medium
4	45-54	Single	50-74K	Homeowner	Single Female	1	None/Unknown	16	Medium

Hình 3.11: Dữ liệu bảng hh_demographic sau khi xử lí

5. ETL bảng coupon

- Xóa giá trị trùng lặp
- Xóa bản ghi có dữ liệu ở cột CAMPAIGN mà không có trong bảng df_campain_desc
- Xóa bản ghi có dữ liệu ở cột PRODUCT_ID mà không có trong bảng df_product

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 124548 entries, 0 to 124547
Data columns (total 3 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   COUPON_UPC   124548 non-null  int64
 1   PRODUCT_ID   124548 non-null  int64
 2   CAMPAIGN     124548 non-null  int64
dtypes: int64(3)
memory usage: 2.9 MB
```

Hình 3.12: Thông tin bảng coupon trước xử lí

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 119384 entries, 0 to 124547
Data columns (total 3 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   COUPON_UPC   119384 non-null  int64
 1   PRODUCT_ID   119384 non-null  int64
 2   CAMPAIGN     119384 non-null  int64
dtypes: int64(3)
memory usage: 3.6 MB
```

Hình 3.13: Thông tin bảng coupon sau xử lí

	COUPON_UPC	PRODUCT_ID	CAMPAIGN
0	10000089061	27160	4
1	10000089064	27754	9
2	10000089073	28897	12
3	51800009050	28919	28
4	52100000076	28929	25

Hình 3.14: Dữ liệu bảng coupon trước xử lí

	COUPON_UPC	PRODUCT_ID	CAMPAIGN
0	10000089061	27160	4
1	10000089064	27754	9
2	10000089073	28897	12
3	51800009050	28919	28
4	52100000076	28929	25

Hình 3.15: Dữ liệu bảng coupon sau khi xử lí

6. ETL bảng causal_data

- Xóa các bản ghi có dữ liệu ở cột PRODUCT_ID mà không có trong bảng df_product
- Mã hóa các giá trị ở cột mailer
- Mã hóa các giá trị ở cột display

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36786524 entries, 0 to 36786523
Data columns (total 5 columns):
 #   Column      Dtype
---  ------      -----
 0   PRODUCT_ID  int64
 1   STORE_ID    int64
 2   WEEK_NO     int64
 3   display     object
 4   mailer      object
dtypes: int64(3), object(2)
memory usage: 1.4+ GB
```

Hình 3.16: Thông tin bảng causal trước xử lí

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 36786524 entries, 0 to 36786523
Data columns (total 5 columns):
 #   Column      Dtype
---  ------      -----
 0   PRODUCT_ID  int64
 1   STORE_ID    int64
 2   WEEK_NO     int64
 3   display     object
 4   mailer      object
dtypes: int64(3), object(2)
memory usage: 1.6+ GB
```

Hình 3.17: Thông tin bảng causal sau xử lí

	PRODUCT_ID	STORE_ID	WEEK_NO	display	mailer
0	26190	286	70	0	A
1	26190	288	70	0	A
2	26190	289	70	0	A
3	26190	292	70	0	A
4	26190	293	70	0	A

Hình 3.18: Dữ liệu bảng causal trước xử lí

	PRODUCT_ID	STORE_ID	WEEK_NO	display	mailer
0	26190	286	70	display	InteriorFeature
1	26190	288	70	display	InteriorFeature
2	26190	289	70	display	InteriorFeature
3	26190	292	70	display	InteriorFeature
4	26190	293	70	display	InteriorFeature

Hình 3.19: Dữ liệu bảng causal sau khi xử lí

7. ETL bảng campaign_table

- Xóa các bản ghi có dữ liệu ở cột CAMPAIGN mà không có trong bảng campaign_desc
- Xóa các bản ghi có dữ liệu ở cột household_key mà không có trong bảng df_hh_demographic

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7208 entries, 0 to 7207
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   DESCRIPTION    7208 non-null   object
 1   household_key  7208 non-null   int64
 2   CAMPAIGN       7208 non-null   int64
dtypes: int64(2), object(1)
memory usage: 169.1+ KB
```

Hình 3.20: Thông tin bảng campaign_table trước xử lí

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4213 entries, 0 to 7207
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   DESCRIPTION    4213 non-null   object
 1   household_key  4213 non-null   int64
 2   CAMPAIGN       4213 non-null   int64
dtypes: int64(2), object(1)
memory usage: 131.7+ KB
```

Hình 3.21: Thông tin bảng campaign_table sau xử lí

	DESCRIPTION	household_key	CAMPAIGN
0	TypeA	17	26
1	TypeA	27	26
2	TypeA	212	26
3	TypeA	208	26
4	TypeA	192	26

Hình 3.22: Dữ liệu bảng campaign_table trước xử lí

	DESCRIPTION	household_key	CAMPAIGN
0	TypeA	17	26
1	TypeA	27	26
2	TypeA	212	26
3	TypeA	208	26
4	TypeA	192	26

Hình 3.23: Dữ liệu bảng campaign_table sau khi xử lí

8. ETL bảng coupon_redempt

- Xóa các bản ghi có dữ liệu ở cột household_key mà không có trong bảng df_hh_demographic
- Xóa các bản ghi có dữ liệu ở cột CAMPAIGN mà không có trong bảng df_campaign_desc
- Xóa các bản ghi có dữ liệu ở cột COUPON_UPC mà không có trong bảng df_coupon

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2318 entries, 0 to 2317
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   household_key  2318 non-null   int64
 1   DAY            2318 non-null   int64
 2   COUPON_UPC     2318 non-null   int64
 3   CAMPAIGN       2318 non-null   int64
dtypes: int64(4)
memory usage: 72.6 KB
```

Hình 3.24: Thông tin bảng coupon_redempt trước xử lí

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1856 entries, 0 to 2314
Data columns (total 4 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   household_key  1856 non-null   int64
 1   DAY            1856 non-null   int64
 2   COUPON_UPC     1856 non-null   int64
 3   CAMPAIGN       1856 non-null   int64
dtypes: int64(4)
memory usage: 72.5 KB
```

Hình 3.25: Thông tin bảng coupon_redempt sau xử lí

	household_key	DAY	COUPON_UPC	CAMPAIGN
0	1	421	10000085364	8
1	1	421	51700010076	8
2	1	427	54200000033	8
3	1	597	10000085476	18
4	1	597	54200029176	18

Hình 3.26: Dữ liệu bảng coupon_redempt trước khi xử lí

	household_key	DAY	COUPON_UPC	CAMPAIGN
0	1	421	10000085364	8
1	1	421	51700010076	8
2	1	427	54200000033	8
3	1	597	10000085476	18
4	1	597	54200029176	18

Hình 3.27: Dữ liệu bảng coupon_redempt sau khi xử lí

9. ETL bảng transaction_data

- Xóa bản ghi có dữ liệu ở cột household_key mà không có trong bảng df_hh_demographic
- Xóa bản ghi có giá trị ở cột PRODUCT_ID thuộc bảng transaction mà không có trong bảng product
- Thêm cột DAY_OF_WEEK dựa vào cột DAY

- Chuyển đổi giá trị cột TRANS_TIME thành thời gian

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2595732 entries, 0 to 2595731
Data columns (total 12 columns):
 #   Column             Dtype
---  ------             -----
 0   household_key      int64
 1   BASKET_ID          int64
 2   DAY                int64
 3   PRODUCT_ID         int64
 4   QUANTITY           int64
 5   SALES_VALUE        float64
 6   STORE_ID           int64
 7   RETAIL_DISC        float64
 8   TRANS_TIME         int64
 9   WEEK_NO            int64
 10  COUPON_DISC        float64
 11  COUPON_MATCH_DISC  float64
dtypes: float64(4), int64(8)
memory usage: 237.6 MB
```

Hình 3.28: Thông tin bảng transaction trước xử lí

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1422306 entries, 11 to 2595706
Data columns (total 14 columns):
 #   Column             Non-Null Count    Dtype
---  ------             --------------    -----
 0   household_key      1422306 non-null  int64
 1   BASKET_ID          1422306 non-null  int64
 2   DAY                1422306 non-null  int64
 3   PRODUCT_ID         1422306 non-null  int64
 4   QUANTITY           1422306 non-null  int64
 5   SALES_VALUE        1422306 non-null  float64
 6   STORE_ID           1422306 non-null  int64
 7   RETAIL_DISC        1422306 non-null  float64
 8   TRANS_TIME         1422306 non-null  int64
 9   WEEK_NO            1422306 non-null  int64
 10  COUPON_DISC        1422306 non-null  float64
 11  COUPON_MATCH_DISC  1422306 non-null  float64
 12  DAY_OF_WEEK        1422306 non-null  object
 13  TIME               1422306 non-null  object
dtypes: float64(4), int64(8), object(2)
memory usage: 162.8+ MB
```

Hình 3.29: Thông tin bảng transaction sau xử lí

	household_key	BASKET_ID	DAY	PRODUCT_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	TRANS_TIME	WEEK_NO	COUPON_DISC	COUPON_MATCH_DISC
0	2375	26984851472	1	1004906	1	1.39	364	-0.60	1631	1	0.0	0.0
1	2375	26984851472	1	1033142	1	0.82	364	0.00	1631	1	0.0	0.0
2	2375	26984851472	1	1036325	1	0.99	364	-0.30	1631	1	0.0	0.0
3	2375	26984851472	1	1082185	1	1.21	364	0.00	1631	1	0.0	0.0
4	2375	26984851472	1	8160430	1	1.50	364	-0.39	1631	1	0.0	0.0

Hình 3.30: Dữ liệu bảng transaction trước khi xử lí

	household_key	BASKET_ID	DAY	PRODUCT_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	TRANS_TIME	WEEK_NO	COUPON_DISC	COUPON_MATCH_DISC	DAY_OF_WEEK	TIME
11	1364	26984896261	1	842930	1	2.19	31742	0.00	1520	1	0.0	0.0	Wednesday	15:20
12	1364	26984896261	1	897044	1	2.99	31742	-0.40	1520	1	0.0	0.0	Wednesday	15:20
13	1364	26984896261	1	920955	1	3.09	31742	0.00	1520	1	0.0	0.0	Wednesday	15:20
14	1364	26984896261	1	937406	1	2.50	31742	-0.99	1520	1	0.0	0.0	Wednesday	15:20
15	1364	26984896261	1	981760	1	0.60	31742	-0.79	1520	1	0.0	0.0	Wednesday	15:20

Hình 3.31: Dữ liệu bảng transaction sau khi xử lí

10. Thông tin dữ liệu sau khi ETL

index	Table	Number of Rows	Number of Columns	Null Values	Data Types	Memory Usage	duplicate of rows	blank values
0	df_campaign_desc	30	4	0	object,int64,int64,int64	1088	0	{'DESCRIPTION': 0, 'CAMPAIGN': 0, 'START_DAY': 0, 'END_DAY': 0}
1	df_campaign_table	4213	3	0	object,int64,int64	134816	0	{'DESCRIPTION': 0, 'household_key': 0, 'CAMPAIGN': 0}
2	df_causal_data	36786524	5	0	int64,int64,int64,object,object	1765753152	0	{'PRODUCT_ID': 0, 'STORE_ID': 0, 'WEEK_NO': 0, 'display': 0, 'mailer': 0}
3	df_coupon	119384	3	0	int64,int64,int64	3820288	0	{'COUPON_UPC': 0, 'PRODUCT_ID': 0, 'CAMPAIGN': 0}
4	df_coupon_redempt	1856	4	0	int64,int64,int64,int64	74240	0	{'household_key': 0, 'DAY': 0, 'COUPON_UPC': 0, 'CAMPAIGN': 0}
5	df_hh_demographic	801	8	0	object,object,object,object,object,object,object,int64	51392	0	{'AGE_DESC': 0, 'MARITAL_STATUS_CODE': 0, 'INCOME_DESC': 0, 'HOMEOWNER_DESC': 0, 'HH_COMP_DESC': 0, 'HOUSEHOLD_SIZE_DESC': 0, 'KID_CATEGORY_DESC': 0, 'household_key': 0}
6	df_product	92338	7	0	int64,int64,object,object,object,object,object	5909632	0	{'PRODUCT_ID': 0, 'MANUFACTURER': 0, 'DEPARTMENT': 0, 'BRAND': 0, 'COMMODITY_DESC': 0, 'SUB_COMMODITY_DESC': 0, 'CURR_SIZE_OF_PRODUCT': 0}
7	df_transaction	1422306	14	0	int64,int64,int64,int64,int64,float64,int64,float64,int64,int64,float64,float64,object,object	170676720	0	{'household_key': 0, 'BASKET_ID': 0, 'DAY': 0, 'PRODUCT_ID': 0, 'QUANTITY': 0, 'SALES_VALUE': 0, 'STORE_ID': 0, 'RETAIL_DISC': 0, 'TRANS_TIME': 0, 'WEEK_NO': 0, 'COUPON_DISC': 0, 'COUPON_MATCH_DISC': 0, 'DAY_OF_WEEK': 0, 'TIME': 0}

Hình 3.32: Thông tin dữ liệu sau khi xử

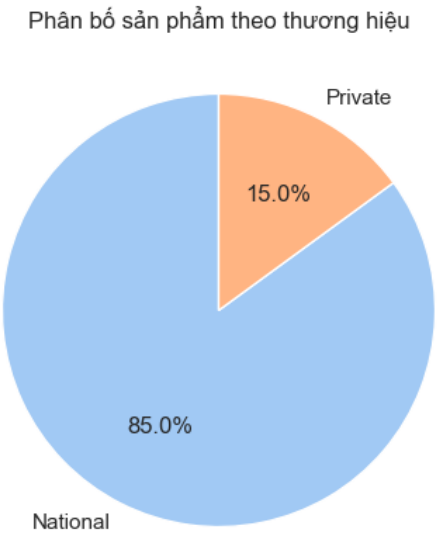
Sau quá trình xử lí, dữ liệu đã không còn các giá tị trống, các giá trị thừa, một vài thuộc tính đã được định dạng lại kiểu dữ liệu và mã hóa thành các giá trị có nghĩa.

Sau đó dữ liệu được xuất thành các file csv để sử dụng cho các bước sau.

3.3 Khám phá dữ liệu

3.3.1 Phân tích các phân phối và đặc trưng của dữ liệu

- Sản phẩm



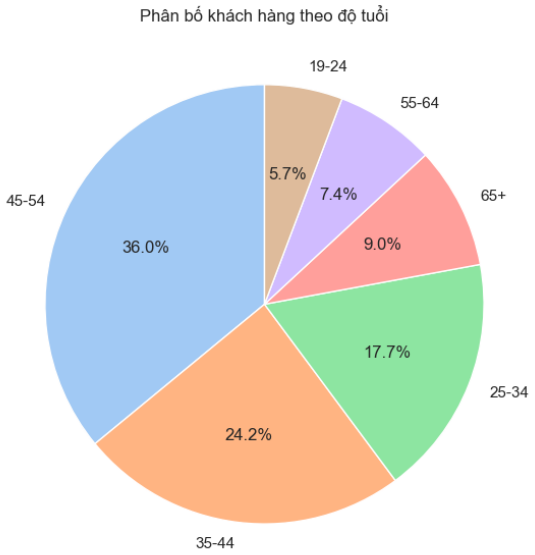
Hình 3.33: Phân bố sản phẩm theo thương hiệu

	Thương hiệu	Số Lượng
0	National	78537
1	Private	13816

Hình 3.34: Thống kê số lượng sản phẩm theo thương hiệu

Trong tổng số 92353 sản phẩm được bày bán, sản phẩm có thương hiệu từ nước ngoài chiếm số lượng lớn, còn lại phần nhỏ là sản phẩm có thương hiệu trong nước. Như vậy có thể thấy rằng số lượng được bày bán chủ yếu được nhập khẩu từ nước ngoài.

- Khách hàng



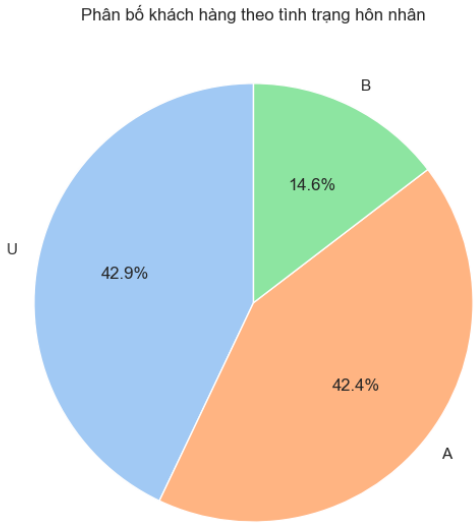
Hình 3.35: Phân bố khách hàng theo độ tuổi

	Độ Tuổi	Số Lượng
0	45-54	288
1	35-44	194
2	25-34	142
3	65+	72
4	55-64	59
5	19-24	46

Hình 3.36: Số lượng khách hàng theo độ tuổi

Theo dõi các khách hàng dựa trên 6 độ tuổi, phần lớn lượng khách hàng ở độ tuổi trung niên từ 35 đến 64 tuổi , sau đó ít hơn 1 chút là các khách hàng trẻ có độ tuổi từ 25 đến 34 tuổi.

Các nhóm khách hàng ngoài độ tuổi trên chiếm phần nhỏ, chiếm ít nhất là độ tuổi 19 - 24, độ tuổi chủ yếu độc thân chưa lập gia đình. Nhận xét sơ bộ thấy số lượng khách hàng chủ yếu là những người trung niên lớn tuổi, có gia đình.

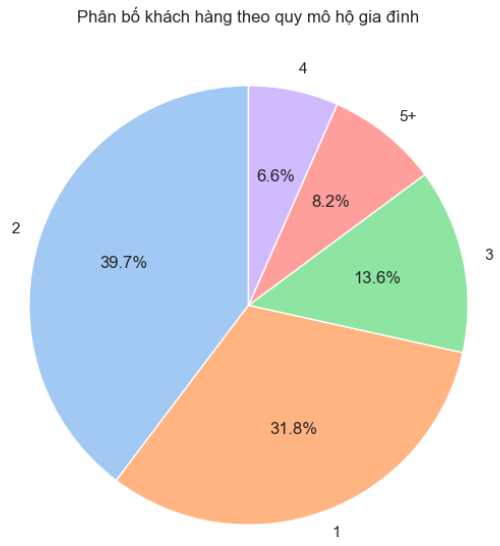


Hình 3.37: Tỉ trọng khách hàng theo tình trạng hôn nhân

	Tình trạng hôn nhân	Số Lượng
0	U	344
1	A	340
2	B	117

Hình 3.38: Số lượng khách hàng theo tình trạng hôn nhân

Bỏ qua trường hợp không xác định được tình trạng hôn nhân của khách hàng, tỉ lệ khách hàng đã lập gia đình chiếm phần lớn. Điều này hợp lí với phân bố khách hàng theo độ tuổi đã kể trên.



Hình 3.39: Tỉ trọng khách hàng theo quy mô gia đình

	Quy mô hộ gia đình	Số Lượng
0	2	318
1	1	255
2	3	109
3	5+	66
4	4	53

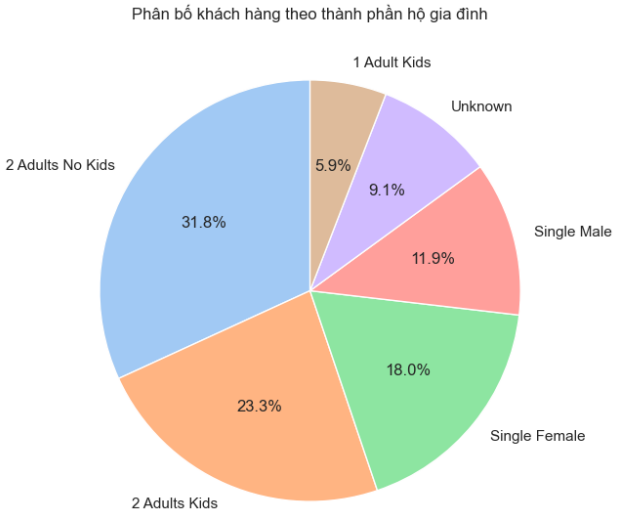
Hình 3.40: Số lượng khách hàng theo quy mô gia đình

Phần lớn gia đình của khách hàng có số lượng thành viên là 2 người, cùng với các phân tích trên có thể là các khách hàng ở độ tuổi trung niên đã lập gia đình và chưa có con.

Sau đó là các hộ gia đình có 1 người - điều này không có mâu thuẫn khi tỉ lệ độc thân chiếm khá nhỏ.

Vì số lượng khách hàng có tình trạng hôn nhân không xác định khá lớn hoặc có thể đã kết hôn nhưng li hôn.

Chiếm tỉ trọng nhỏ là các hộ gia đình có số lượng thành viên lớn hơn.



Hình 3.41: Tỉ trọng khách hàng theo thành phần gia đình

	Thành phần hộ gia đình	Số Lượng
0	2 Adults No Kids	255
1	2 Adults Kids	187
2	Single Female	144
3	Single Male	95
4	Unknown	73
5	1 Adult Kids	47

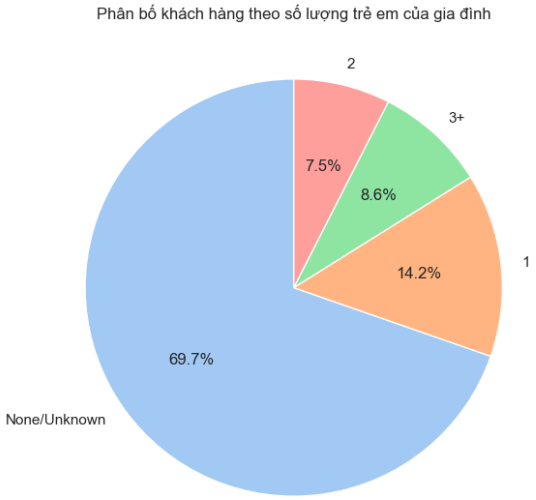
Hình 3.42: Số lượng khách hàng theo thành phần gia đình

Phân tích sâu hơn vào thành phần gia đình của khách hàng, ta thấy gia đình khách hàng

chủ yếu gồm 2 người lớn, đúng với nhận xét chủ yếu hộ gia đình là các gia đình trung niên.

Như đã nhìn thấy ở trên: các hộ gia đình có 1 người cũng chiếm phần không nhỏ, ở biểu đồ này ta thấy được tỉ trọng nam, nữ độc thân chiếm phần lớn nhưng tình trạng hôn nhân độc thân lại chiếm số nhỏ.

Điều này khẳng định thêm việc tỉ trọng khách hàng đã kết hôn nhưng li dị là không ít.



Hình 3.43: Tỉ trọng khách hàng theo số lượng trẻ em trong gia đình

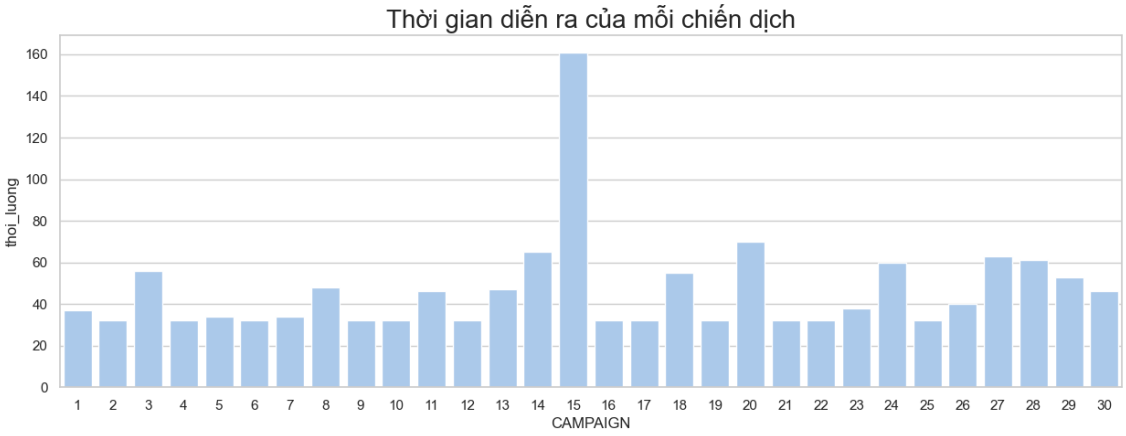
	Số lượng trẻ em	Số Lượng
0	None/Unknown	558
1	1	114
2	3+	69
3	2	60

Hình 3.44: Số lượng khách hàng theo số lượng trẻ em trong gia đình

Số lượng trẻ em trong gia đình khách hàng chủ yếu là không có hoặc không xác định.

Phần còn lại đa phần là nhóm gia đình có 1 trẻ em. Chiếm tổng cộng gần bằng nhóm gia đình có 1 trẻ em là các gia đình có 2 hoặc hơn 3 trẻ em.

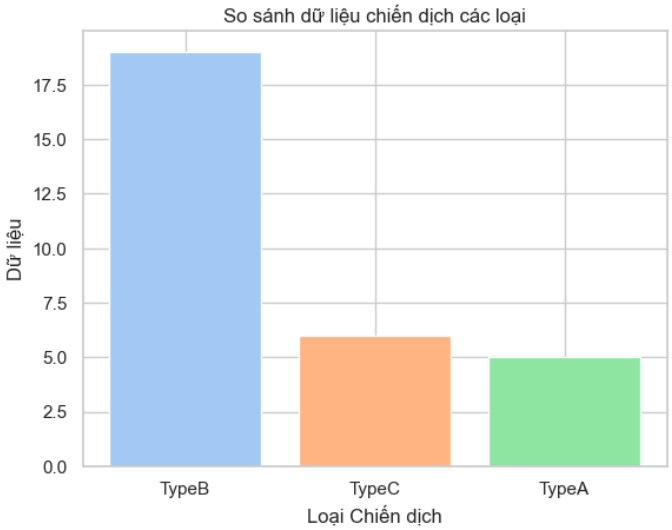
- Chiến dịch



Hình 3.45: Thời lượng kéo dài của các chiến dịch

Theo biểu đồ so sánh thời lượng của 30 chiến dịch diễn ra, đa phần các chiến dịch kéo dài trong khoảng từ 30 đến 45 ngày.

Một vài chiến dịch kéo dài hơn đến 60 ngày và nổi trộ nhất là chiến dịch 15, đây cũng là chiến dịch duy nhất kéo dài đến 160 ngày.



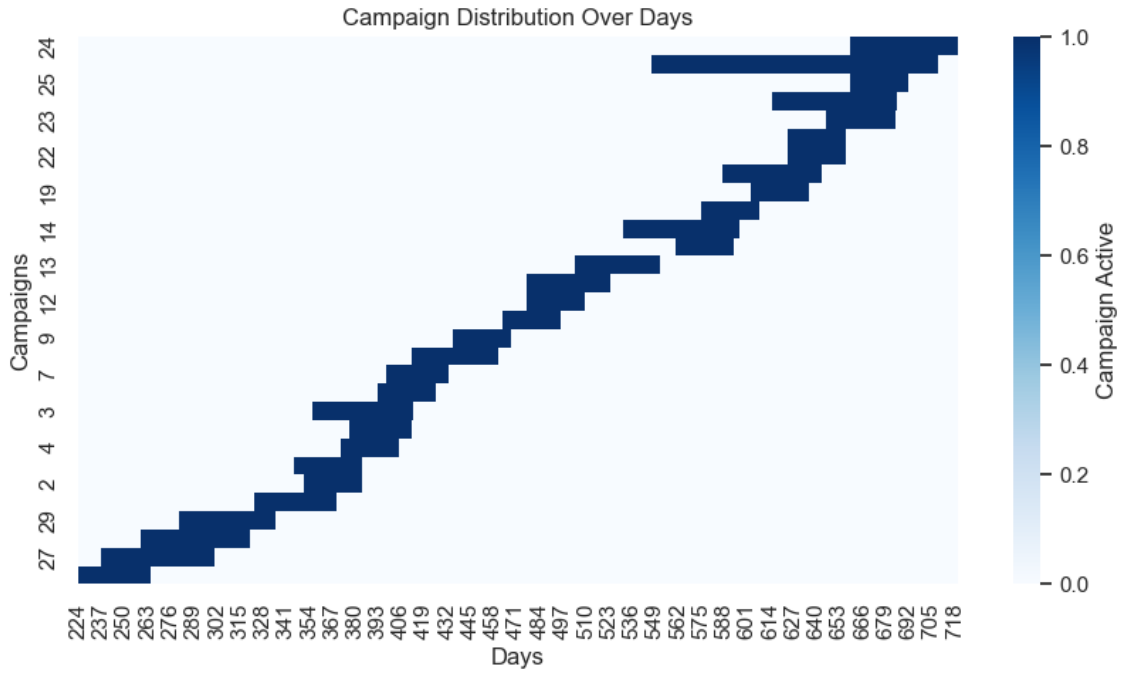
Hình 3.46: Phân phối chiến dịch theo loại

	Loại chiến dịch	Số Lượng
0	TypeB	19
1	TypeC	6
2	TypeA	5

Hình 3.47: Số lượng chiến dịch theo loại

Các chiến dịch diễn ra được phân nhóm thành 3 loại chiến dịch A, B, C.

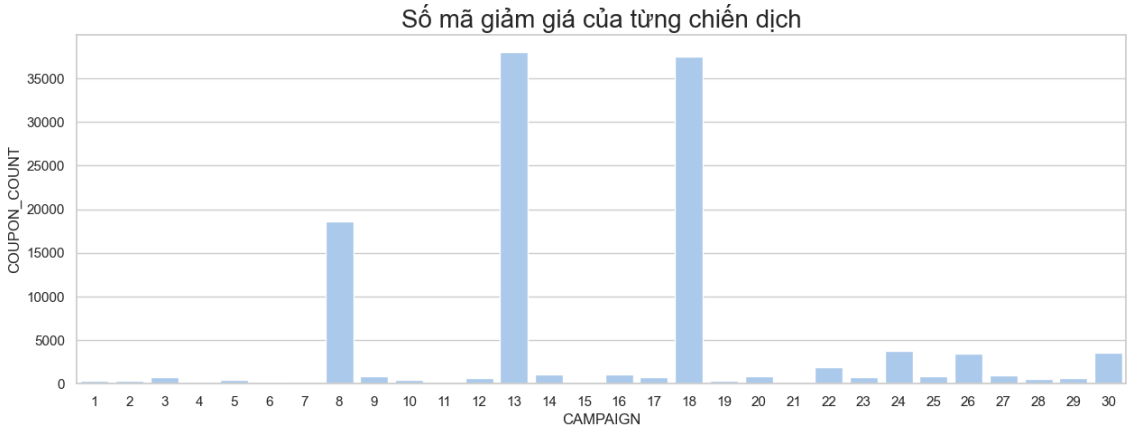
Trong đó nhóm chiến dịch loại B chiếm đến gần 2/3 tổng số chiến dịch. Còn lại chiếm số ít và xấp xỉ nhau là chiến dịch loại A và loại C.



Hình 3.48: Phân bố chiến dịch theo thời gian

Theo dõi phân bố tổ chứ các chiến dịch theo thời gian thấy rằng các chiến dịch diễn ra nối tiếp nhau, gối lên nhau kéo dài từ ngày 214 đến khi dừng theo dõi.

- Mã giảm giá



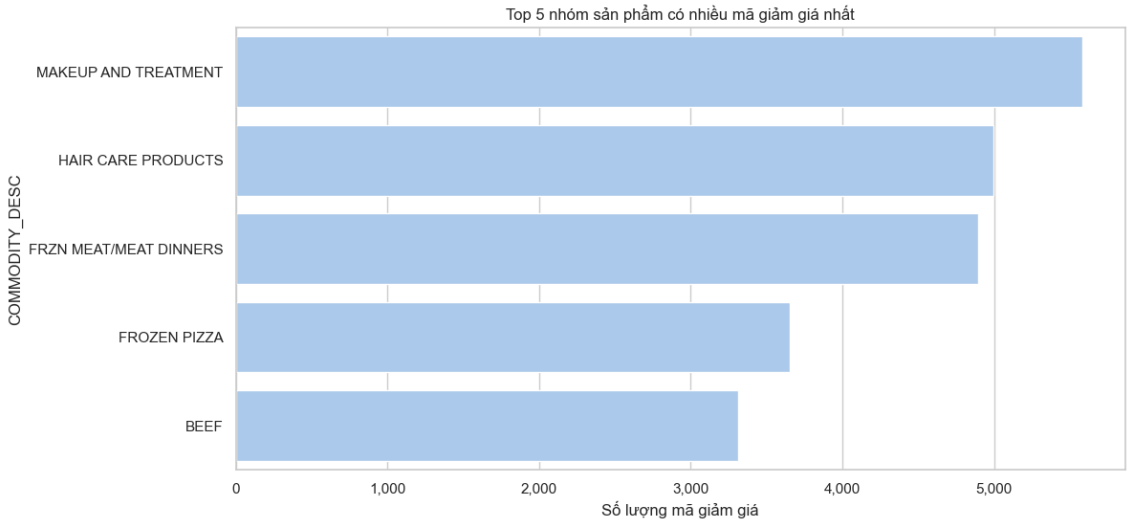
Hình 3.49: Số lượng mã giảm giá của từng chiến dịch

Đi kèm với việc tổ chứ các chiến dịch là việc tung ra các mã giảm giá đi kèm với các sản phẩm.

Các mã giảm giá của từng chiến dịch chủ yếu có số lượng trong khoảng dưới 2000 mã giảm giá.

Như theo dõi bên trên thời lượng của các chiến dịch 8, 13, 18 chỉ nằm trong tầm trung nhưng ở biểu đồ này số lượng mã giảm giá của các chiến dịch này rất nhiều.

Trong khi đó, chiến dịch 15 có thời lượng đến 160 ngày nhưng số lượng mã giảm giá chiếm số nhỏ. Đây có thể là chiến lược của tổ chức để truyền thông dài ngày nhưng vẫn đảm bảo được lợi nhuận.

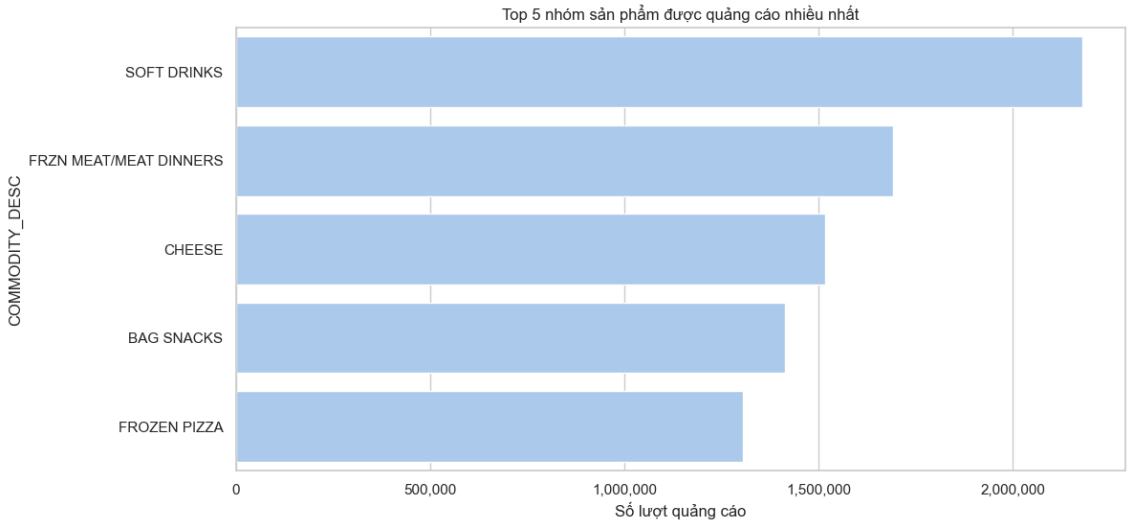


Hình 3.50: Top 5 nhóm sản phẩm có nhiều mã giảm giá nhất

Trong số những mã giảm giá được tung ra thì rất nhiều sản phẩm được đính kèm sử dụng mã giảm giá nhiều lần.

Trong đó, đứng đầu là nhóm các sản phẩm mỹ phẩm và chăm sóc sắc đẹp.

- Tiếp thị

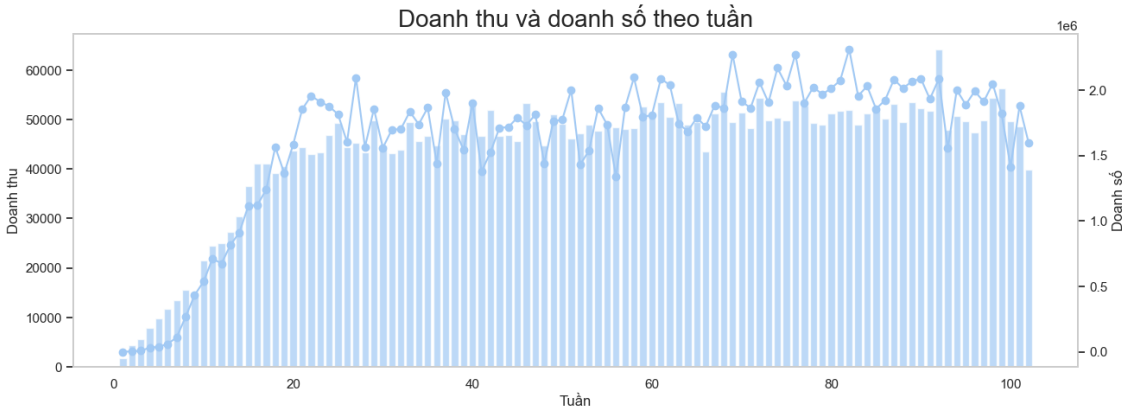


Hình 3.51: Top 5 nhóm sản phẩm được tiếp thị nhiều nhất

Dù nhóm sản phẩm 'FRZN MEAT/MEAT DINNERS' chỉ xếp thứ 3 lượt tung mã giảm giá nhưng luôn được ưu tiên trong việc trưng bày và quảng cáo qua qua thư.

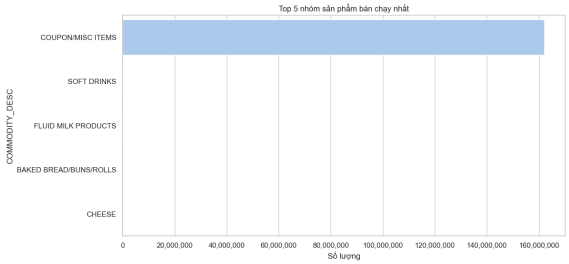
Tuy nhiên trong 5 nhóm sản phẩm được tiếp thị nhiều nhất chủ yếu là các thức ăn nhanh.

- Doanh số và doanh thu

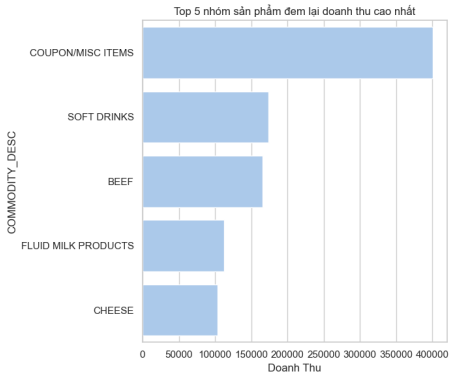


Hình 3.52: Tăng trưởng doanh số, doanh thu theo tuần

Theo từng tuần, trong khoảng 15 tuần đầu doanh số và doanh thu khá thấp cho đến tuần 20 doanh số và doanh thu có chút biến động nhưng khá đều nhau.



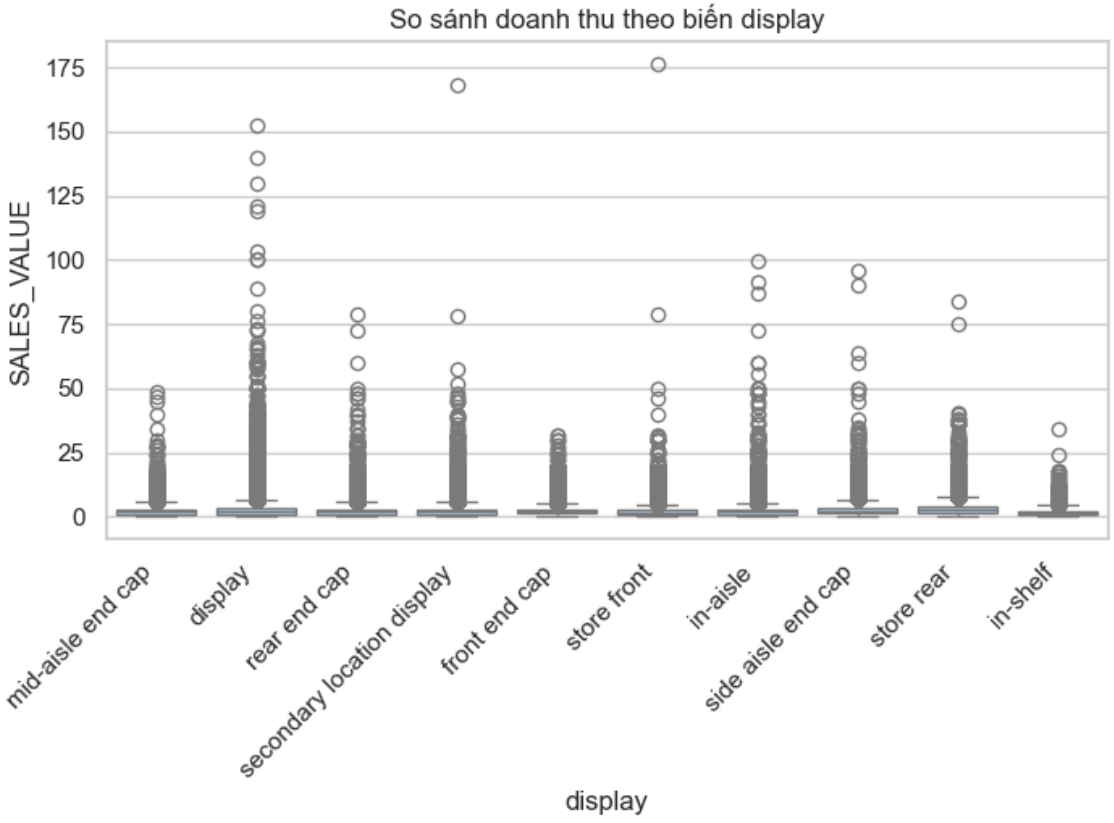
Hình 3.53: Top 5 nhóm sản phẩm bán chạy nhất



Hình 3.54: Top 5 nhóm sản phẩm có doanh thu cao nhất

Đứng đầu về lượt bán cũng như doanh thu là nhóm sản phẩm COUPON/MISC ITEMS với con số áp đảo.

Tuy nhiên, so về mặt số lượng bán ra của các nhóm sản phẩm còn lại không đáng kể với mặt hàng trên nhưng doanh thu của các mặt hàng này khá cao khi doanh thu từng nhóm gần bằng 1/2 doanh số nhóm COUPON/MISC ITEMS.



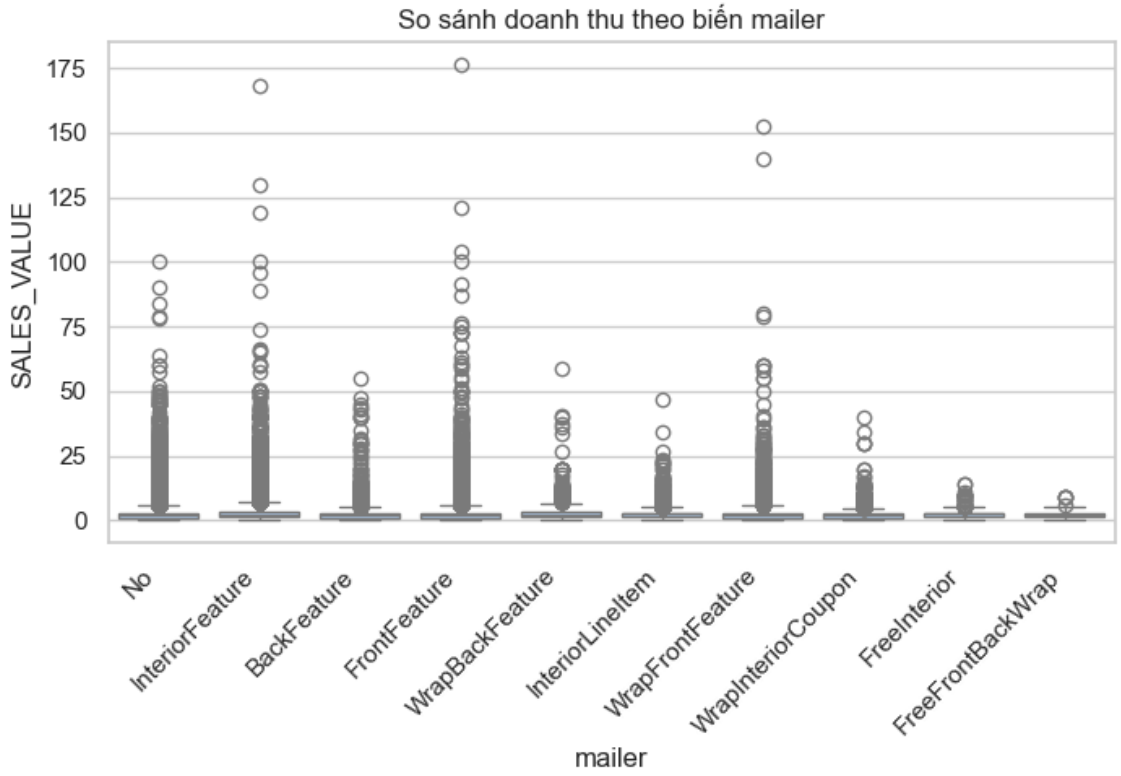
Hình 3.55: Doanh thu dựa trên vị trí trưng bày

Biểu đồ trên thể hiện sự phân phối doanh thu dựa trên các vị trí trưng bày:

- Có nhiều điểm nằm ngoài boxplot, đây có thể là các giá trị ngoại lai, đây chính là các giao dịch có doanh thu bán hàng cao bất thường so với mặt bằng chung các giao

dịch.

- Tại các vị trí "display"và "secondary location display"có giá trị trung vị và phạm vi doanh thu cao hơn các vị trí khác. Điều này cho thấy sản phẩm trưng bày trên vị trí khu trưng bày chính và phụ đem lại doanh thu cao hơn.
- Vị trí có trung vị thấp hơn là "front end cap"và "in-shelf". Hai vị trí này khuất tầm mắt nên doanh thu thấp hơn so với các vị trí khác.
- Có một số giá trị quá xa trung vị, đặc biệt là với các vị trí như "in-aisle"và "store rear", chúng có các giao dịch với doanh thu bán hàng đặc biệt cao so với mặt bằng chung với cùng vị trí đó.



Hình 3.56: Doanh thu dựa trên vị trí của quảng cáo qua thư

Biểu đồ trên cho thấy sự phân phối doanh thu theo vị trí ở quảng cáo qua thư:

- Các vị trí quảng cáo như "FrontFeature"và "interiorFeature"có trung vị cao hơn hẳn so với các vị trí khác. Điều này cho thấy vị trí này đem lại hiệu quả doanh thu hơn các vị trí khác.
- Trong khi đó, các loại quảng cáo như "FreeFrontBackWrap"và "FreeFontBackWrap"có vị và phạm vi giá trị bán hàng thấp hơn thậm chí xấp xỉ 0, điều này có thể chỉ ra rằng chúng ít hiệu quả hơn trong việc tạo ra doanh thu.
- Có một vài giá tị ngoại lai trong một số loại quảng cáo, đặc biệt là "FrontFeature", "interiorFeature"và "WrapFontFeature"cho thấy có những giao dịch có giá trị bán hàng cao bất thường trong những vị trí quảng cáo này.

Từ việc phân tích các phân phối của các thuộc tính ta tính được ra các đặc trưng sau:

Doanh Thu Trung Bình	32
Số Lượng Sản Phẩm Trung Bình	1170
Số Lượng Mã Giảm Giá Trung Bình Trong 1 Chiến Dịch	3979
Thời Gian Kéo Dài Chiến Dịch Trung Bình	47

Hình 3.35: Đặc trưng trung bình của dữ liệu

3.3.2 Phân tích tương quan

	COUPON_DISC	COUPON_MATCH_DISC	RETAIL_DISC	QUANTITY	SALES_VALUE
count	1422306.000	1422306.000	1422306.000	1422306.000	1422306.000
mean	-0.015	-0.004	-0.537	115.262	3.162
std	0.179	0.044	1.227	1254.316	4.255
min	-55.930	-4.050	-130.020	0.000	0.000
25%	0.000	0.000	-0.690	1.000	1.290
50%	0.000	0.000	-0.010	1.000	2.100
75%	0.000	0.000	0.000	1.000	3.490
max	0.000	0.000	0.790	89638.000	840.000

Hình 3.58: Bảng phân tính thống kê mô tả về số lượng, doanh thu và các khoản giảm giá

Từ bảng thống kê trên ta biết được các giá trị về mặt thống kê như số lượng, tổng, trung bình, trung vị,... của các thuộc tính định lượng cần quan tâm như QUANTITY, SALES_VALUE, RETAIL_DISC, COUPON_DISC, COUPON_MATCH_DISC.

- COUPON_DISC và COUPON_MATCH_DISC:

Hai biến này có giá trị trung bình tiến tới 0(-0.015,-0.004), độ lệch chuẩn khá nhỏ. Vì vậy xét trên tổng thể thì giá trị giảm giá từ mã giảm giá và hoàn trả mã giảm giá của nhà sản xuất không lớn.

- RETAIL_DISC:

Biến này có giá trị trung bình (-0.537), có nghĩa là giảm giá bán lẻ được áp dụng gần như trung bình trên tất các giao dịch.

Độ lệch chuẩn của biến này khá lớn (1.227), cho thấy được sự biến động rõ ràng giữa các lần giảm giá của nhà bán lẻ.

Thậm chí có giá trị nhỏ nhất rất nhỏ (-130.020). Điều này cho thấy có những giao dịch, nhà bán lẻ giảm giá rất nhiều.

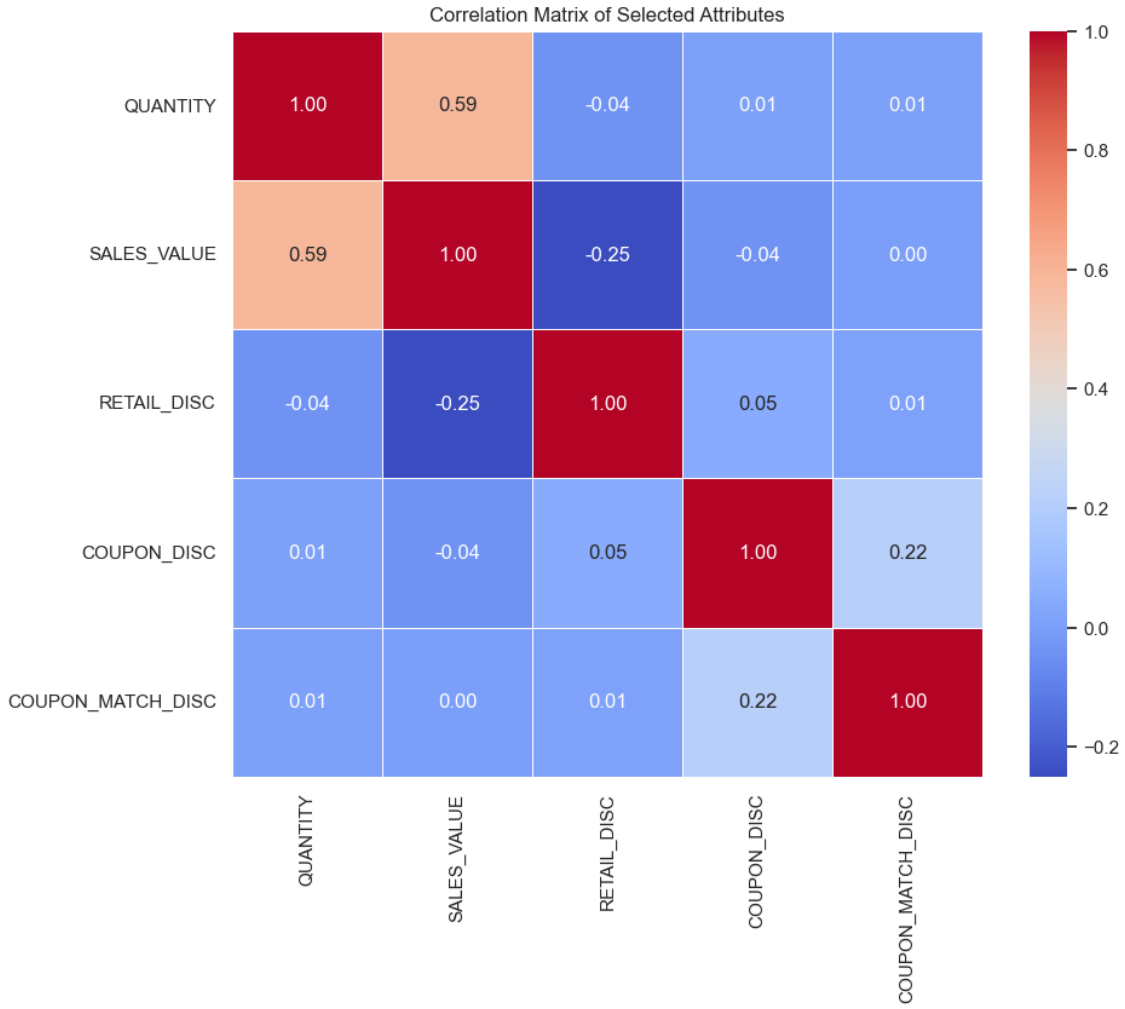
- QUANTITY:

Xét thấy trung bình số lượng sản phẩm bán ra trung bình trong một giao dịch là 115.262 nhưng độ lệch chuẩn lại rất lớn (1254.316), có nghĩa là có sự chênh lệch rất lớn giữa số lượng sản phẩm được mua trong các giao dịch.

Ngoài ra, giá trị lớn nhất rất lớn (89638.000), tức là có giao dịch mua với số lượng rất lớn. Đây có thể là các giao dịch bán buôn.

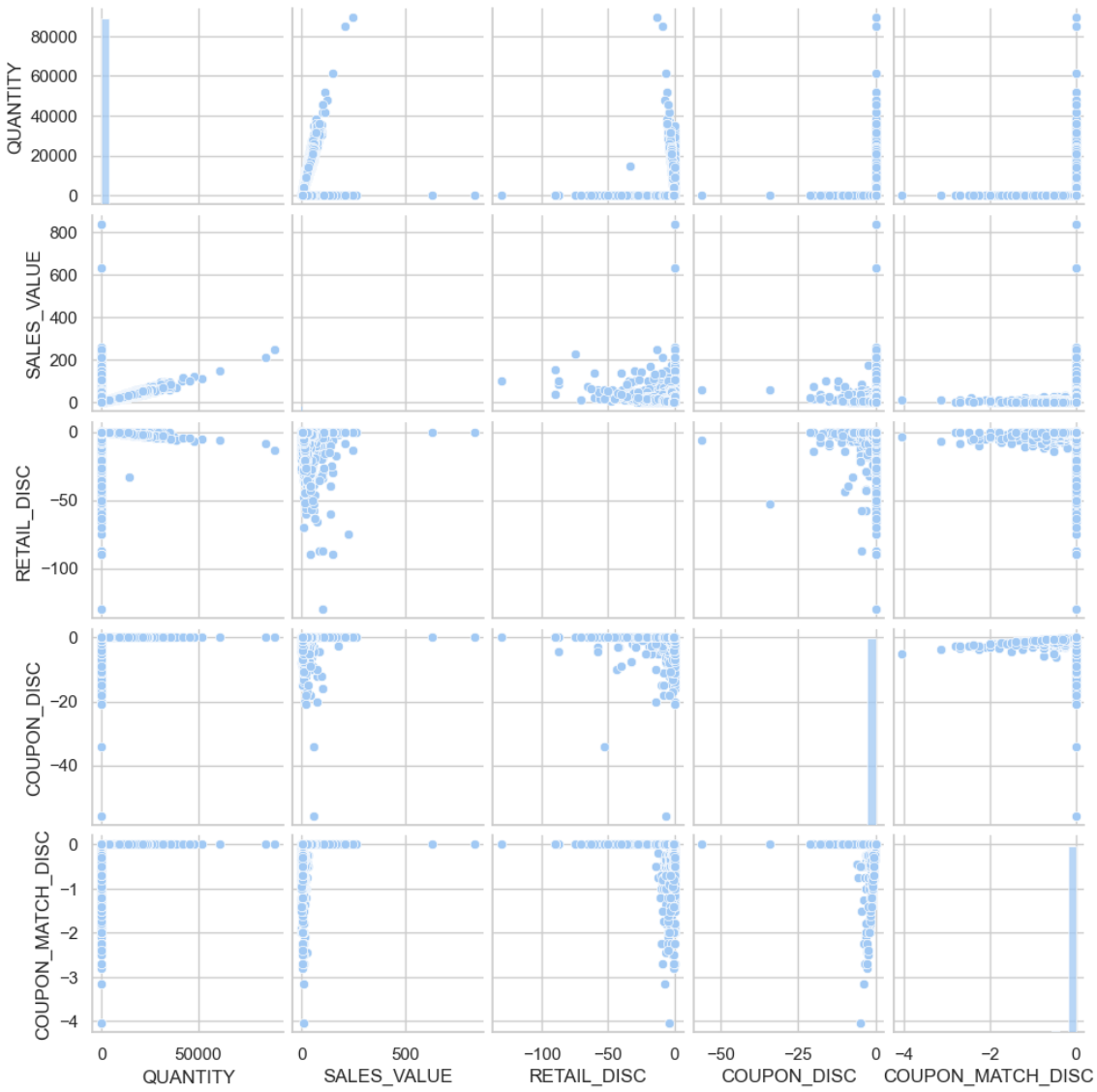
- SALES_VALUE:

Doanh thu bán hàng có độ lệch chuẩn tương đối lớn (4.255) với giá trị tối đa là 840.000, thể hiện sự biến động lớn về doanh thu giữa các giao dịch, có thể do các giao dịch với số lượng lớn đã kể trên



Hình 3.59: Ma trận tương quan giữa số lượng sản phẩm, giá trị và các khoản giảm giá

Dựa vào ma trận tương quan giữa các biến ta thấy nổi bật là tương quan giữa QUANTITY và SALES_VALUE ngoài ra còn có tương quan âm giữa QUANTITY và RETAIL_DISC.



Hình 3.60: Tương quan giữa lượng, doanh thu và các khoản giảm giá

Cụ thể hơn ta thấy:

- Giữa QUANTITY và SALES_VALUE, các điểm tập trung thành đường thẳng tuyến tính đi lên theo chiều dương. Điều này hoàn toàn hợp lí với thực tế, khi sản phẩm được mua với số lượng lớn thì giá trị doanh thu từ sản phẩm cũng tăng lên.
- Giữa QUANTITY và RETAIL_DISC, các điểm tập trung thành đường theo hệ số âm, điều này giải thích việc khi mua hàng hóa với số lượng nhiều thì nhà bán lẻ cũng giảm giá nhiều hơn.
- Giữa COUPON_DISC và COUPON_MATCH_DISC các điểm tập trung thành đường thẳng không rõ ràng, điều này phản ánh một phần sự hoàn trả giảm giá của nhà sản xuất có liên quan đến các giảm giá của nhà sản xuất.
- Giữa SALES_VALUE, QUANTITY với COUPON_DISC, COUPON_MATCH_DISC không có sự tương quan nào. Có nghĩa là sự giảm giá của nhà sản xuất và sự hoàn trả giảm giá của nhà sản xuất không hề phụ thuộc vào số lượng hay doanh thu, điều

này hoàn toàn đúng khi các mã giảm giá của nhà sản xuất được tung ra với các sản phẩm cố định.

3.4 Mô hình dữ liệu

Mô hình dữ liệu là mô hình xác định các thông tin, dữ liệu, liên kết cấu thành nên các hình ảnh nhằm trình bày tổng thể về dữ liệu của hệ thống.

Mô hình dữ liệu giúp các tổ chức khai thác thông tin dễ dàng dựa trên các mô hình được trình bày. Qua đó, các doanh nghiệp sẽ có cách nhìn trực diện về dữ liệu và có kế hoạch thúc đẩy doanh thu cho doanh nghiệp.

Mô hình dữ liệu gồm ba loại, mỗi loại đảm nhận một vai trò riêng trong việc xây dựng cơ sở dữ liệu của hệ thống:

1. Mô hình dữ liệu khái niệm
2. Xây dựng mô hình dữ liệu logic
3. Xây dựng mô hình dữ liệu vật lí

3.4.1 Mô hình dữ liệu khái niệm

Mô hình dữ liệu khái niệm tập trung vào việc mô tả các khái niệm, mối quan hệ và luồng thông tin chính trong hệ thống.

1. Phân tích các chiều dim - fact

Dữ liệu được tổ chức theo các Dim - Fact rõ ràng giúp nâng cao hiệu năng truy xuất và xử lí dữ liệu, trong đó:

- Dimension table (dim) là bảng dữ liệu được thiết kế dựa trên các khía cạnh phân tích của từng chủ điểm. Mỗi khía cạnh thể hiện một thuộc tính của dữ liệu
- Fact table (fact) là bảng dữ liệu được thiết kế dựa trên các chủ điểm phân tích, giúp lưu trữ thông tin chi tiết về các sự kiện xảy ra trong hệ thống. Bảng fact có các khóa ngoại tham chiếu đến các bảng dim.

1. Hệ thống chiều khái niệm

- Hệ thống chiều khái niệm nhóm gia đình

Hệ thống chiều khái niệm theo nhóm hộ gia đình

6 bản ghi	3 bản ghi	12 bản ghi	5 bản ghi	6 bản ghi	5 bản ghi	4 bản ghi
Độ tuổi	**Tình trạng hôn nhân**	**Mức thu nhập**	**Nhóm hộ gia đình**	**Thành phần hộ gia đình**	**Quy mô hộ gia đình**	**Số lượng trẻ em**
65+	Married	Under 15K	Homeowner	2 Adults No Kids	2	None/Unknown
45-54	Unknown	15-24K	Unknown	2 Adults Kids	3	1
25-34	Single	25-34K	Renter	Single Female	4	2
35-44		35-49K	Probable Renter	Unknown	1	3+
19-24		50-74K	Probable Owner	Single Male	5	
55-64		75-99K		1 Adult Kids		
		100-124K				
		125-149K				
		150-174K				
		175-199K				
		200-249K				
		250K+				

Hình 3.61: Chiều khái niệm nhóm hộ gia đình

- Hệ thống chiều khái niệm nhóm sản phẩm

Hệ thống chiều khái niệm theo nhóm sản phẩm

2 bản ghi	44 bản ghi	308 bản ghi	2383 bản ghi
Thương hiệu	**Chủng loại**	**Nhóm sản phẩm**	**Sản phẩm**
National	GROCERY	FRZN ICE	ICE - CRUSHED/CUBED
Private	MISC. TRANS.	NO COMMODITY DESCRIPTION	NO SUBCOMMODITY DESCRIPTION
	PASTRY	BREAD	BREAD:ITALIAN/FRENCH
	DRUG GM	FRUIT - SHELF STABLE	APPLE SAUCE
	MEAT-PCKGD	COOKIES/CONES	SPECIALTY COOKIES
	SEAFOOD-PCKGD	SPICES & EXTRACTS	SPICES & SEASONINGS
	PRODUCE	VITAMINS	TRAY PACK/CHOC CHIP COOKIES
	NUTRITION	BREAKFAST SWEETS	VITAMIN - MINERALS
	DELI	PNT BTR/JELLY/JAMS	SW GDS: SW ROLLS/DAN
	COSMETICS	ICE CREAM/MILK/SHERBTS	HONEY
	MEAT	MAGAZINE	TRADITIONAL
	FLORAL	AIR CARE	TV/MOVIE-MAGAZINE
	TRAVEL & LEISUR	CHEESE	AIR CARE - AEROSOLS
	SEAFOOD	SHORTENING/OIL	STRING CHEESE
	MISC SALES TRAN	COFFEE	VEGETABLE/SALAD OIL
	SALAD BAR	DIETARY AID PRODUCTS	INSTANT DECAF FLVR COFFEE W/ S
	KIOSK-GAS	PAPER HOUSEWARES	DIET CNTRL LIQS NUTRITIONAL
	ELECT &PLUMBING	BAKED BREAD/BUNS/ROLLS	PAPER AND FOAM DRINKING CUPS

Hình 3.62: Chiều khái niệm nhóm sản phẩm

- Hệ thống chiều khái niệm nhóm chiến dịch

Hệ thống chiều khái niệm theo nhóm chiến dịch

3 bản ghi	30 bản ghi	503 bản ghi
Loại chiến dịch	**Chiến dịch**	**Phiếu giảm giá**
TypeB	24	10000085364
TypeC	15	51700010076
TypeA	25	54200000033
	20	10000085476
	23	54200029176
	21	53600000078
	22	53700048182
	18	52370020076
	19	53600050042
	17	53600000082
	14	51600070033
	16	52500060033
	13	54154889076
	11	10000085427
	12	52740022050

Hình 3.63: Chiều khái niệm nhóm chiến dịch

- Hệ thống chiều khái niệm nhóm quảng cáo

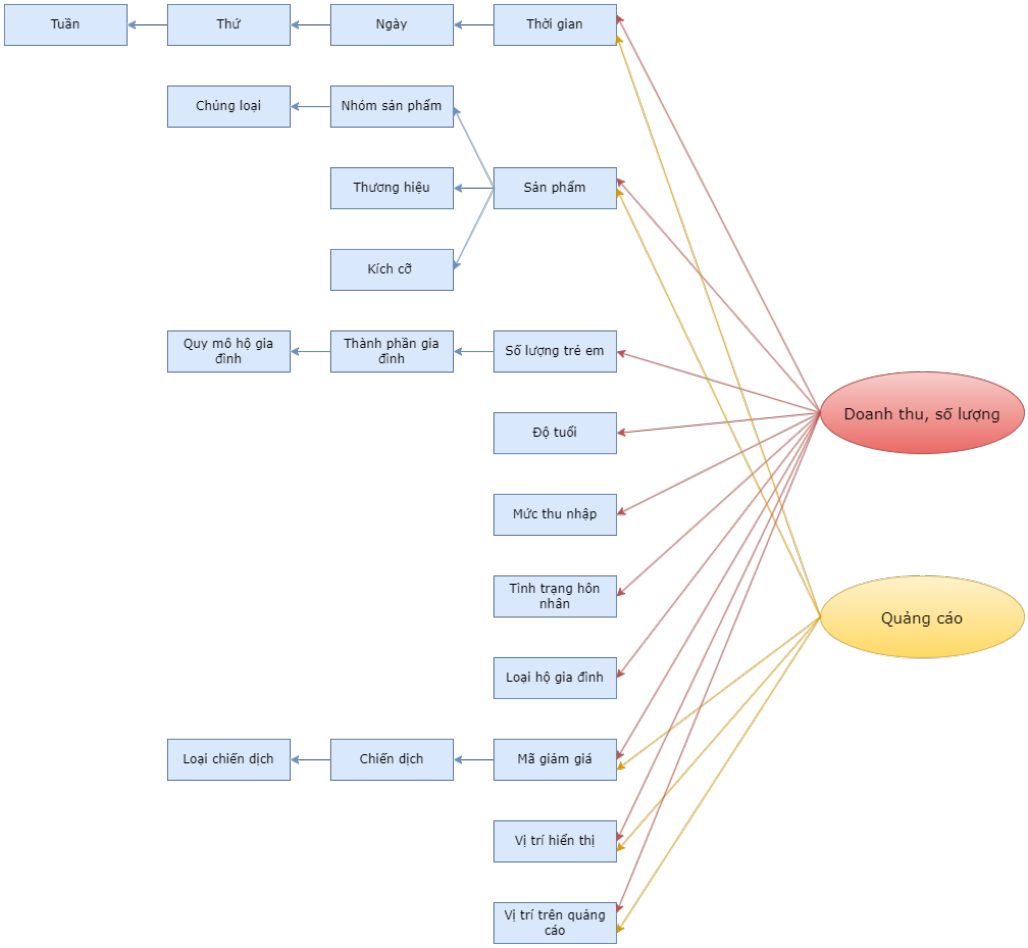
Hệ thống chiều khái niệm theo nhóm quảng cáo

10 bản ghi	11 bản ghi
Vị trí hiển thị, trưng bày	**Vị trí trên quảng cáo**
display	No
store front	InteriorFeature
store rear	InteriorLineItem
front end cap	FrontFeature
mid-aisle end cap	BackFeature
rear end cap	WrapFrontFeature
side aisle end cap	WrapInteriorCoupon
in-aisle	WrapBackFeature
secondary location display	InteriorCoupon
in-shelf	FreeInterior
	FreeFrontBackWrap

Hình 3.64: Chiều khái niệm nhóm quảng cáo

3.4.2 Mô hình dữ liệu logic

Mô hình dữ liệu logic tập trung vào mô tả các thực thể, các thuộc tính và mối quan hệ giữa các thực thể.



Hình 3.65: Mô hình dữ liệu logic

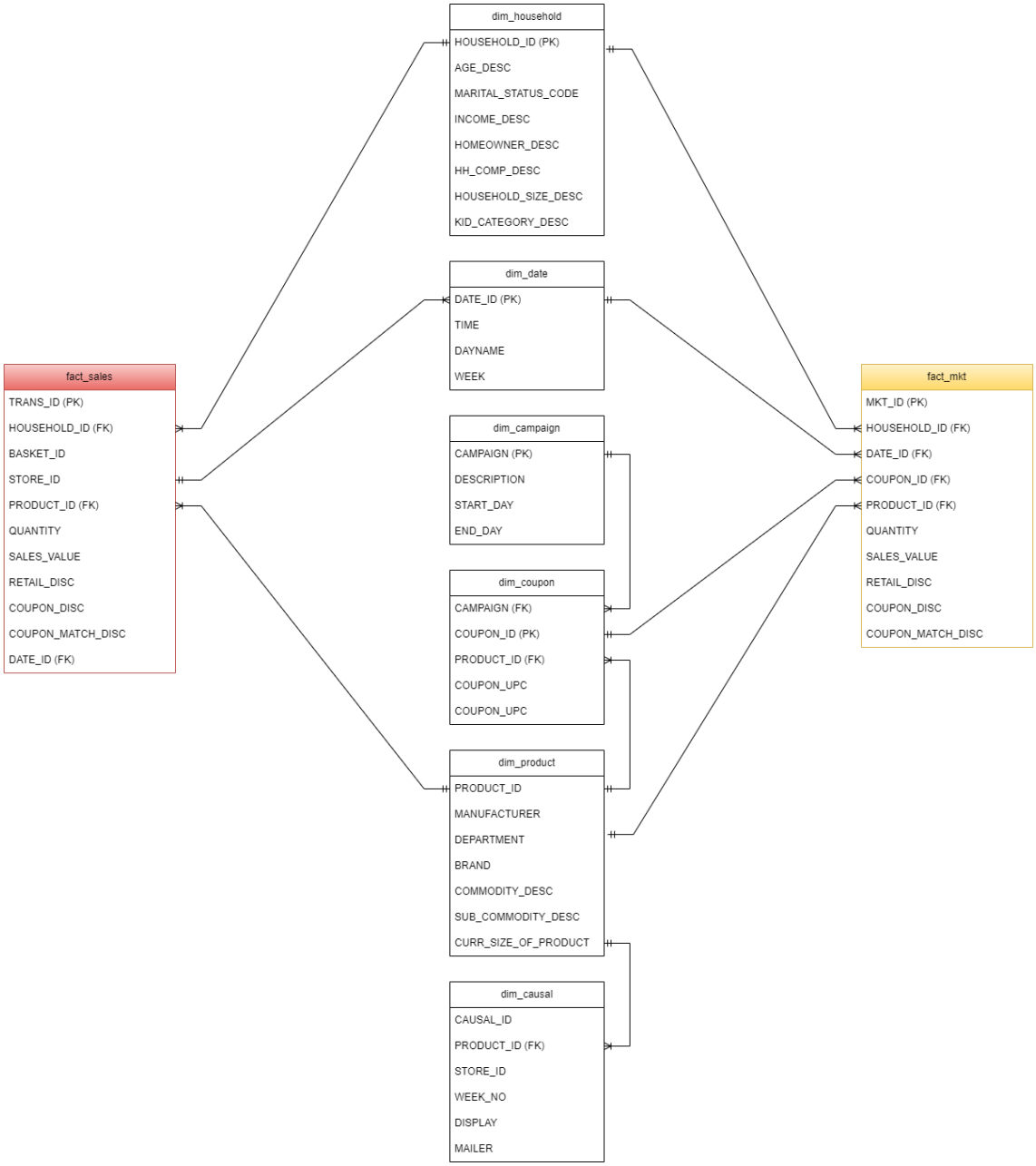
3.4.3 Mô hình dữ liệu vật lí

Mô hình dữ liệu vật lí tập trung mô tả các đối tượng vật lý như bảng, cột, chỉ mục và quan hệ giữa chúng.

Trong mô hình này gồm 6 bảng dimension và 2 bảng fact được thiết kế theo sơ đồ dữ liệu hình bông tuyết.

Trong sơ đồ dữ liệu hình bông tuyết, mỗi fact liên kết với các dimension liên quan, những dimension đó có thể liên kết với những dimension khác, phân nhánh thành các cấp độ.

Sử dụng sơ đồ dữ liệu hình bông tuyết giúp tối ưu hiệu xuất truy vấn, giảm thiểu được lưu trữ dữ liệu. Đồng thời, hệ thống sử dụng sơ đồ dữ liệu này có khả năng mở rộng mối quan hệ qua lại giữa các dimension tạo nên cấu trúc phân cấp dữ liệu.



Hình 3.66: Mô hình dữ liệu vật lí

3.5 Thiết kế hệ thống kho dữ liệu

Các bảng trong cơ sở dữ liệu của kho dữ liệu (Data Warehouse) được thiết kế thành các bảng dim - fact như sau :

Tên bảng	Tên trường	Kiểu dữ liệu	Mô tả
dim_product	PRODUCT_ID	int	Mã sản phẩm
	MANUFACTURER	int	Mã sản xuất
	DEPARTMENT	varchar(255)	Chủng loại sản phẩm
	BRAND	varchar(255)	Thương hiệu
	COMMODITY_DESC	varchar(255)	Nhóm sản phẩm

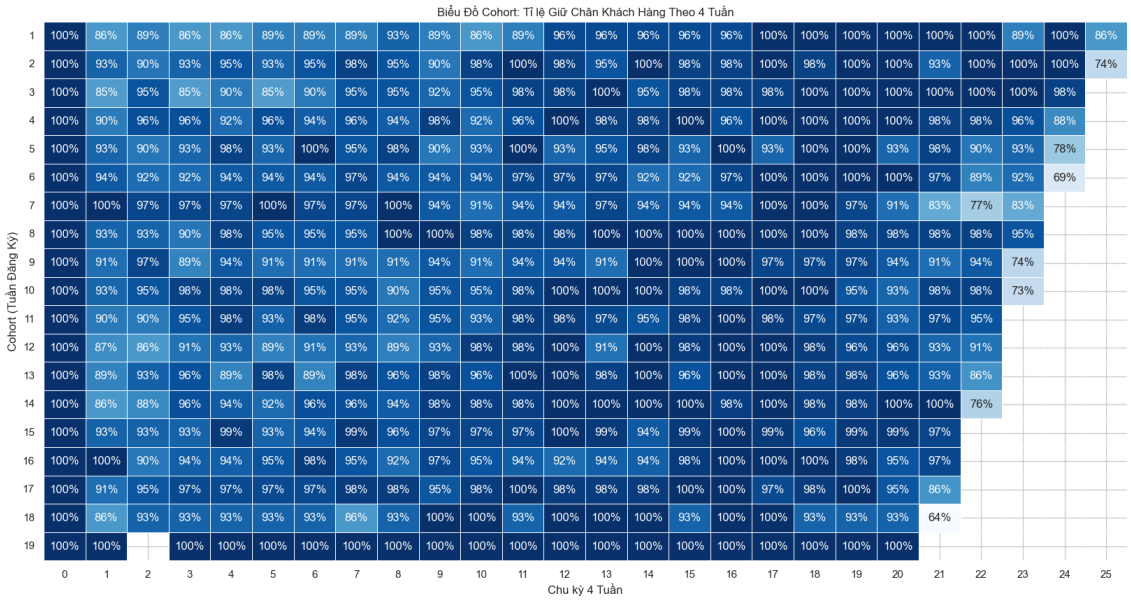
	SUB_COMMODITY_DESC	varchar(255)	Tên sản phẩm
	CURR_SIZE_OF_PRODUCT	varchar(255)	Kích cỡ sản phẩm
dim_household	HOUSEHOLD_ID (PK)	int	Mã hộ gia đình
	AGE_DESC	varchar(255)	Độ tuổi
	MARITAL_STATUS_CODE	varchar(255)	Tình trạng hôn nhân
	INCOME_DESC	varchar(255)	Mức thu nhập
	HOMEOWNER_DESC	varchar(255)	Loại hộ gia
	HH_COMP_DESC	varchar(255)	Thành phần gia đình
	HOUSEHOLD_SIZE_DESC	varchar(255)	Quy mô hộ gia đình
	KID_CATEGORY_DESC	varchar(255)	Số lượng trẻ em
dim_campaign	CAMPAIGN (PK)	int	Mã chiến dịch
	DESCRIPTION	varchar(255)	Loại chiến dịch
	START_DAY	int	Ngày bắt đầu chiến dịch
	END_DAY	int	Ngày kết thúc chiến dịch
dim_coupon	COUPON_ID (PK)	int	Mã định danh phiếu giảm giá
	COUPON_UPC	bigint	Mã giảm giá
	PRODUCT_ID (FK)	int	Mã sản phẩm được giảm giá
	CAMPAIGN (FK)	int	Mã chiến dịch tung mã giảm giá
dim_causal	CAUSAL_ID (PK)	int	Mã lượt hiển thị
	PRODUCT_ID (FK)	int	Mã sản phẩm được hiển thị
	STORE_ID	int	Mã cửa hàng
	WEEK_NO	int	Tuần
	DISPLAY	varchar(255)	Vị trí hiển thị
	MAILER	varchar(255)	Vị trí trên quảng cáo
dim_date	DATE_ID (PK)	int	Mã thời gian
	TIME	int	Thời gian(giờ)
	DAY	int	Ngày
	DAYNAME	varchar(255)	Thứ
	WEEK	int	Tuần

fact_sales	TRANS_ID (PK)	int	Mã giao dịch
	HOUSEHOLD_ID (FK)	int	Mã hộ gia đình
	BASKET_ID	bigbigint	Mã giỏ hàng
	PRODUCT_ID (FK)	int	Mã sản phẩm
	QUANTITY	int	Số lượng
	SALES_VALUE	decimal(10,2)	Số tiền khách hàng phải trả
	RETAIL_DISC	decimal(10,2)	Giảm giá của nhà bán lẻ
	COUPON_DISC	decimal(10,2)	Giá giá của nhà sản xuất
	COUPON_MATCH_DISC	decimal(10,2)	Giảm giá nhà bán lẻ khớp với nhà sản xuất
	DATE_ID (FK)	int	Mã thời gian
fact_mkt	MKT_ID (PK)	int	Mã giao dịch sử dụng mã giảm giá
	HOUSEHOLD_ID (FK)	int	Mã hộ gia đình
	PRODUCT_ID (FK)	int	Mã sản phẩm
	DATE_ID (FK)	int	Mã thời gian
	COUPON_ID (FK)	int	Mã giảm giá
	QUANTITY	int	Số lượng
	SALES_VALUE	decimal(10,2)	Số tiền khách hàng phải trả
	RETAIL_DISC	decimal(10,2)	Giảm giá của nhà bán lẻ
	COUPON_DISC	decimal(10,2)	Giảm giá của nhà sản xuất
	COUPON_MATCH_DISC	decimal(10,2)	Giảm giá nhà bán lẻ khớp với nhà sản xuất

Bảng 3.1: Các bảng dim - fact trong hệ thống OLAP

3.6 Khai phá dữ liệu

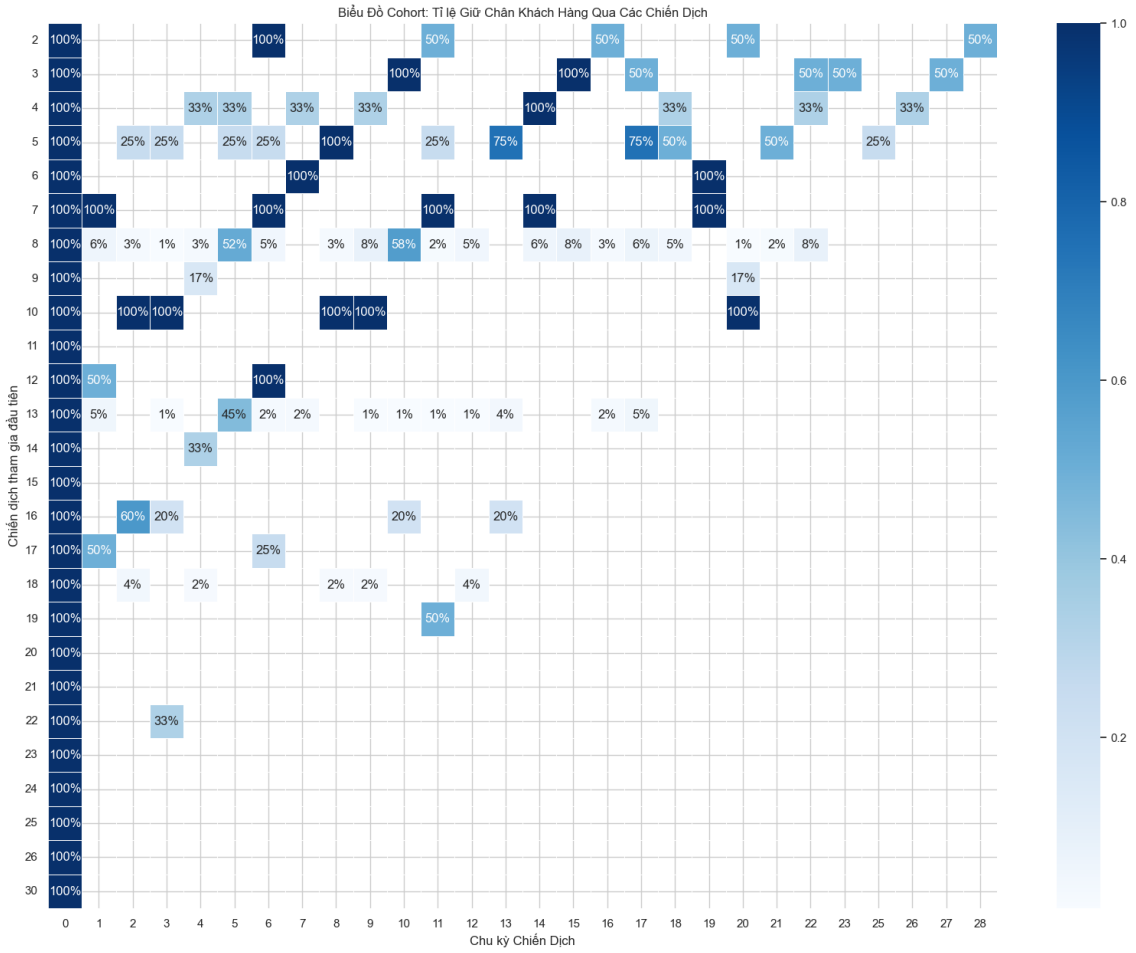
3.6.1 Tình hình kinh doanh



Hình 3.67: Theo dõi tỉ lệ rời bỏ của khách hàng sau chu kì 4 tuần

Trong theo dõi chu kì 4 tuần một, nhận thấy rằng tỉ lệ rời bỏ của nửa đầu các tuần lớn hơn nửa gần cuối. Tuy nhiên tỉ lệ này chưa đáng quan ngại.

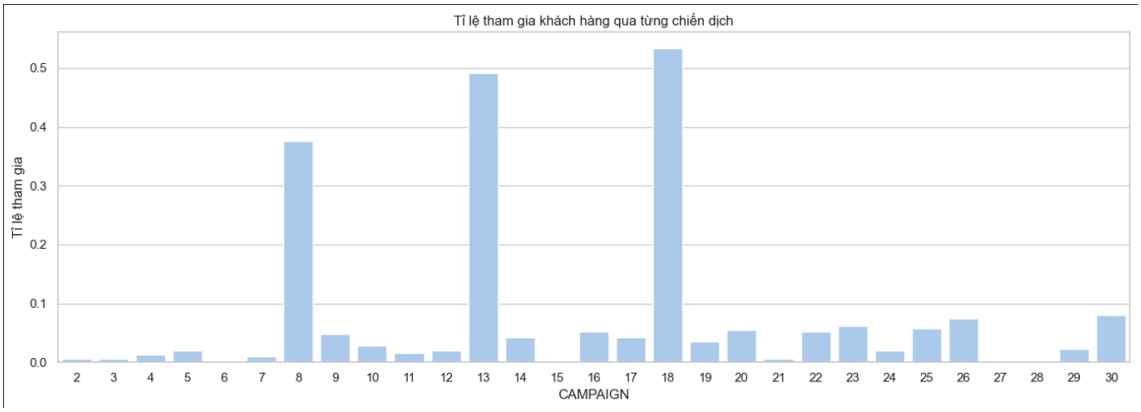
Mặc dù các tuần cuối có tỉ lệ khách hàng mua lại cao nhưng đến cuối thì tỉ lệ này giảm đáng kể. Tỉ lệ này dường như chạm mức báo động khi tỉ lệ rời bỏ chiếm 30 đến 40 %.



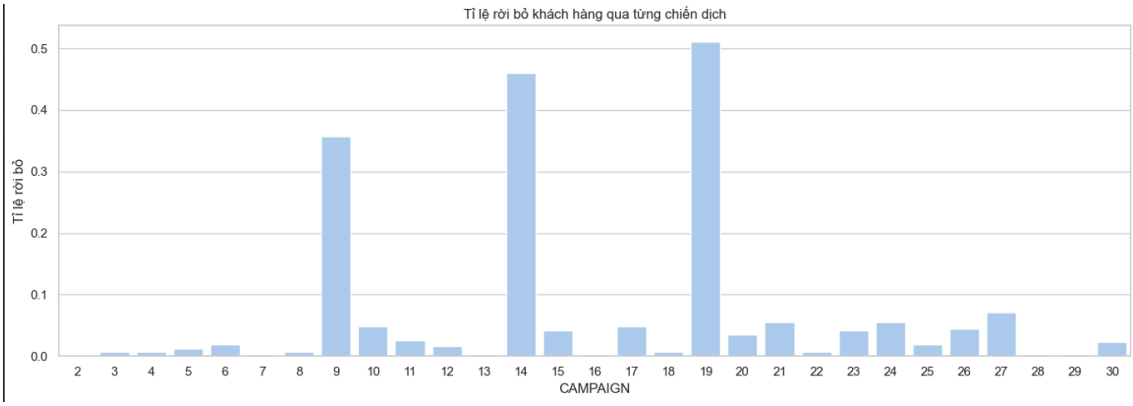
Hình 3.68: Theo dõi sự rời bỏ của khách hàng theo các chiến dịch

Theo dõi sự rời bỏ của các khách hàng qua các chiến dịch không được rõ ràng do thời gian giữa các chiến dịch bị trùng nhau, thêm việc các mã giảm giá ứng với các sản phẩm có khả năng lặp lại ở các chiến dịch khác nhau dẫn đến sự theo dõi này trở lên phức tạp hơn.

Tuy nhiên nhìn chung thấy tỉ lệ khách hàng tiếp tục tham gia chiến dịch của các chiến dịch là không đều. Nguyên nhân có thể do sự chênh lệch về thời lượng chiến dịch cũng như số lượng mã giảm giá của các chiến dịch.



Hình 3.69: Tỉ lệ khách hàng tiếp tục tham gia các chiến dịch



Hình 3.70: Tỉ lệ khách hàng rời bỏ so với chiến dịch trước

Cụ thể hơn, theo dõi tỉ lệ khách hàng tiếp tục tham gia vào các chiến dịch thêm phần khẳng định về sự không đồng đều, chênh lệch lớn giữa các chiến dịch.

Có thể thấy có những chiến dịch tỉ lệ khách hàng tiếp tục tham gia là rất thấp , nhưng có những chiến dịch như 8, 13, 18 thu hút khách hàng quay trở lại với tỉ lệ lớn.

Chính vì vậy việc đột ngột quay lại nhưng không giữ chân được khách hàng ở chiến dịch tiếp theo dẫn tới việc sau mỗi chiến dịch đó là sự tuột dốc về tỉ lệ tiếp tục tham gia của khách hàng. Điều đó tạo nên sự không đồng đều giữa các chiến dịch.

Đặt ra một câu hỏi cho tổ chức về việc tổ chức các chiến dịch sao cho giữ được lượng khách hàng tham gia ổn định và ngày càng cao.

3.6.2 Xu hướng mua các sản phẩm cùng nhau

Trong quá trình kinh doanh, để hiệu suất lợi nhuận cao nhất cần hiểu được nhu cầu khách hàng, họ có thói quen mua các sản phẩm nào cùng nhau trong cùng 1 giỏ hàng. Từ đó thực hiện các các sắp xếp, ưu đãi với các cặp sản phẩm.

	PRODUCT_ID	MANUFACTURER	DEPARTMENT	BRAND	COMMODITY_DESC	SUB_COMMODITY_DESC	CURR_SIZE_OF_PRODUCT
0	25671	2	GROCERY	National	FRZN ICE	ICE - CRUSHED/CUBED	22 LB
1	26081	2	MISC. TRANS.	National	NO COMMODITY DESCRIPTION	NO SUBCOMMODITY DESCRIPTION	
2	26093	69	PASTRY	Private	BREAD	BREAD:ITALIAN/FRENCH	
3	26190	69	GROCERY	Private	FRUIT - SHELF STABLE	APPLE SAUCE	50 OZ
4	26355	69	GROCERY	Private	COOKIES/CONES	SPECIALTY COOKIES	14 OZ

Hình 3.71: Dữ liệu bảng sản phẩm

	household_key	BASKET_ID	DAY	PRODUCT_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	TRANS_TIME	WEEK_NO	COUPON_DISC	COUPON_MAT
0	2375	26984851472	1	1004906	1	1.39	364	-0.60	1631	1	0.0	
1	2375	26984851472	1	1033142	1	0.82	364	0.00	1631	1	0.0	
2	2375	26984851472	1	1036325	1	0.99	364	-0.30	1631	1	0.0	
3	2375	26984851472	1	1082185	1	1.21	364	0.00	1631	1	0.0	
4	2375	26984851472	1	8160430	1	1.50	364	-0.39	1631	1	0.0	

Hình 3.72: Dữ liệu bảng giao dịch

Từ bảng transaction và product ta tiến hành ghép hai bảng rồi lọc ra các cột cần thiết, thu được bảng tổng hợp giao dịch các sản phẩm như sau:

	BASKET_ID	DEPARTMENT	COMMODITY_DESC	SUB_COMMODITY_DESC
0	29046618323	GROCERY	FRZN ICE	ICE - CRUSHED/CUBED
1	30707611686	GROCERY	FRZN ICE	ICE - CRUSHED/CUBED
2	33046710871	GROCERY	FRZN ICE	ICE - CRUSHED/CUBED
3	30760265177	MISC. TRANS.	NO COMMODITY DESCRIPTION	NO SUBCOMMODITY DESCRIPTION
4	33783848749	PASTRY	BREAD	BREAD:ITALIAN/FRENCH

Hình 3.73: Bảng tổng hợp giao dịch của các sản phẩm

Sau đó duyệt qua từng giỏ hàng để đếm số lần xuất hiện của từng cặp hai nhóm sản phẩm. Trong quá trình này, mỗi nhóm sản phẩm trong mỗi giỏ hàng là duy nhất và sắp xếp theo thứ tự của bảng chữ cái. Điều này đảm bảo các cặp nhóm sản phẩm chỉ được đến một lần và không lặp lại.

Ta thu được top 10 cặp nhóm sản phẩm hay xuất hiện cùng nhau cùng với số lần xuất hiện:

	Product 1	Product 2	Count
4093	BAKED BREAD/BUNS/ROLLS	FLUID MILK PRODUCTS	31010
26675	FLUID MILK PRODUCTS	SOFT DRINKS	25452
13038	CHEESE	FLUID MILK PRODUCTS	24584
4239	BAKED BREAD/BUNS/ROLLS	SOFT DRINKS	24519
4022	BAKED BREAD/BUNS/ROLLS	CHEESE	23937
3951	BAG SNACKS	SOFT DRINKS	20573
3700	BAG SNACKS	BAKED BREAD/BUNS/ROLLS	19643
13184	CHEESE	SOFT DRINKS	18886
3995	BAKED BREAD/BUNS/ROLLS	BEEF	18822
3807	BAG SNACKS	FLUID MILK PRODUCTS	18356

Hình 3.74: Top 10 cặp nhóm sản phẩm hay xuất hiện cùng nhau trong giỏ hàng

Nhìn vào các nhóm sản phẩm này có thể thấy sữa là mặt hàng được mua nhiều nhất cùng với các loại sản phẩm khác.

Sữa được mua kèm với các mặt hàng như bánh mì, bim bim và các sản phẩm thức uống nhanh khác.

Sau đó là bim bim được xuất hiện nhiều với các đồ uống nhanh.

Việc các nhóm sản phẩm xuất hiện cùng nhau nhiều lần thì khả năng cao sẽ xuất hiện ở top các sản phẩm hay được mua nhất. Em tiến hành lọc ra top 20 nhóm sản phẩm được mua nhiều nhất

```
COMMODITY_DESC
SOFT DRINKS                    117532
FLUID MILK PRODUCTS             85630
BAKED BREAD/BUNS/ROLLS          83232
CHEESE                          74885
BAG SNACKS                      67190
FRZN MEAT/MEAT DINNERS          56064
BEEF                            48726
SOUP                            46135
YOGURT                          44697
FROZEN PIZZA                    43362
VEGETABLES - SHELF STABLE       41612
COLD CEREAL                     37870
CANDY - CHECKLANE               35556
LUNCHMEAT                       35162
TROPICAL FRUIT                  34442
CANDY - PACKAGED                34177
REFRGRATD JUICES/DRNKS          31469
CRACKERS/MISC BKD FD            29723
CANNED JUICES                   29149
EGGS                            28420
Name: count, dtype: int64
```

Hình 3.75: Top 20 sản phẩm bán chạy nhất

Ta thấy rằng trong số 20 nhóm sản phẩm bán chạy nhất thì 5 nhóm sản phẩm đầu tiên chính là 5 nhóm sản phẩm hay xuất hiện cùng nhau theo từng cặp.

Tìm hiểu sau hơn, em đi tính tỉ lệ số lần mua hàng cùng nhau trên tổng số lần mua hàng của nhóm sản phẩm. Trong quá trình này những cặp nhóm sản phẩm được mua cùng nhau dưới 100 lần sẽ bị bỏ để tránh tỉ lệ ảo

Product 1	Product 2	Count	Sales_x	Sales_y	Relative Count 1	Relative Count 2	Relative Count	Percentage of Times Secondary Bought with Primary
BAKED BREAD/BUNS/ROLLS	HOT DOGS	7759	60311	10998	0.128650	0.705492	0.705492	70.549191
COLD CEREAL	FLUID MILK PRODUCTS	16861	25166	69278	0.669991	0.243382	0.669991	66.999126
BAKED BREAD/BUNS/ROLLS	DELI MEATS	10859	60311	17122	0.180050	0.634213	0.634213	63.421329
CHEESES	DELI MEATS	7134	11331	17122	0.629600	0.416657	0.629600	62.960021
ICE CREAM/MILK/SHERBTS	SYRUPS/TOPPINGS	662	20090	1064	0.032952	0.622180	0.622180	62.218045
BAKED BREAD/BUNS/ROLLS	MEAT - SHELF STABLE	8573	60311	13805	0.142147	0.621007	0.621007	62.100688
EGGS	FLUID MILK PRODUCTS	17274	27886	69278	0.619451	0.249343	0.619451	61.945062
COCOA MIXES	FLUID MILK PRODUCTS	2362	3826	69278	0.617355	0.034095	0.617355	61.735494
BAKED BREAD/BUNS/ROLLS	DINNER MXS:DRY	8144	60311	13309	0.135033	0.611917	0.611917	61.191675
FLUID MILK PRODUCTS	MOLASSES/SYRUP/PANCAKE MIXS	2732	69278	4514	0.039435	0.605228	0.605228	60.522818

Hình 3.76: Tỉ lệ mua hàng cùng nhau trên tổng số

Product 1	Product 2	Percentage of Times Secondary Bought with Primary
BAKED BREAD/BUNS/ROLLS	HOT DOGS	70.54919076195671
COLD CEREAL	FLUID MILK PRODUCTS	66.99912580465708
BAKED BREAD/BUNS/ROLLS	DELI MEATS	63.42132928396216
CHEESES	DELI MEATS	62.96002118083135
ICE CREAM/MILK/SHERBTS	SYRUPS/TOPPINGS	62.21804511278195
BAKED BREAD/BUNS/ROLLS	MEAT - SHELF STABLE	62.10068815646505
EGGS	FLUID MILK PRODUCTS	61.94506203829879
COCOA MIXES	FLUID MILK PRODUCTS	61.735493988499734
BAKED BREAD/BUNS/ROLLS	DINNER MXS:DRY	61.1916748065219
FLUID MILK PRODUCTS	MOLASSES/SYRUP/PANCAKE MIXS	60.52281789986708

Hình 3.77: Tỉ lệ mua cùng nhau của 10 nhóm sản phẩm

Nhận thấy rằng khoảng 71% các nhóm sản phẩm như bánh mì được mua cùng với hot dog, đây là hai sản phẩm hay được kết hợp trong bữa ăn nhanh, bữa sáng

67% khách hàng khi nhóm các sản phẩm ngũ cốc sẽ mua cùng với các loại sữa. Trứng cũng được mua cùng sữa tới 62%.

Điều này là tiền đề cho việc hiểu hành vi mua sắm của khách hàng để thực hiện việc trưng bày sắp xếp, nhóm các cặp sản phẩm cùng nhau hướng tới sự quan tâm của khách hàng đối với các sản phẩm có liên hệ với nhau.

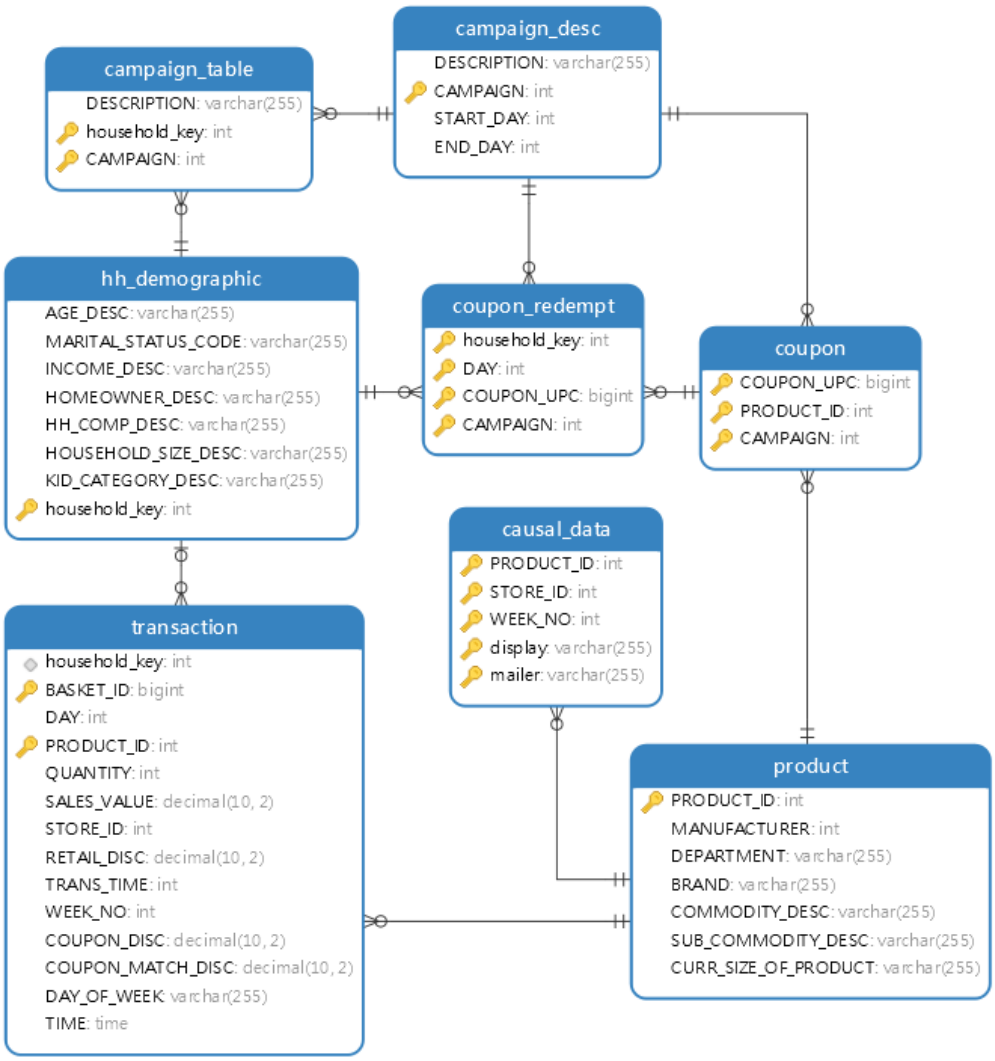
CHƯƠNG 4. XÂY DỰNG CHƯƠNG TRÌNH

Trong phần này, em sử dụng ngôn ngữ truy vấn MySQL với các câu lệnh truy vấn, tạo bảng tạo kết nối để xây dựng hệ thống cơ sở dữ liệu phần lưu trữ Staging và cơ sở dữ liệu OLAP.

4.1 Xây dựng tầng tập kết dữ liệu (Staging)

4.1.1 Tạo database và các bảng dữ liệu trong khu vực Staging

Tạo database "Sales_Staging", sau đó sử dụng các câu lệnh tạo bảng để lưu trữ dữ liệu.



Hình 4.1: ERD Staging

4.1.2 Đưa dữ liệu vào khu vực Staging

Dùng câu lệnh để tải các file dữ liệu đã xử lí vào các bảng đã tạo

Open Table | Design Table | New Table | Delete Table | Import Wizard | Export Wizard

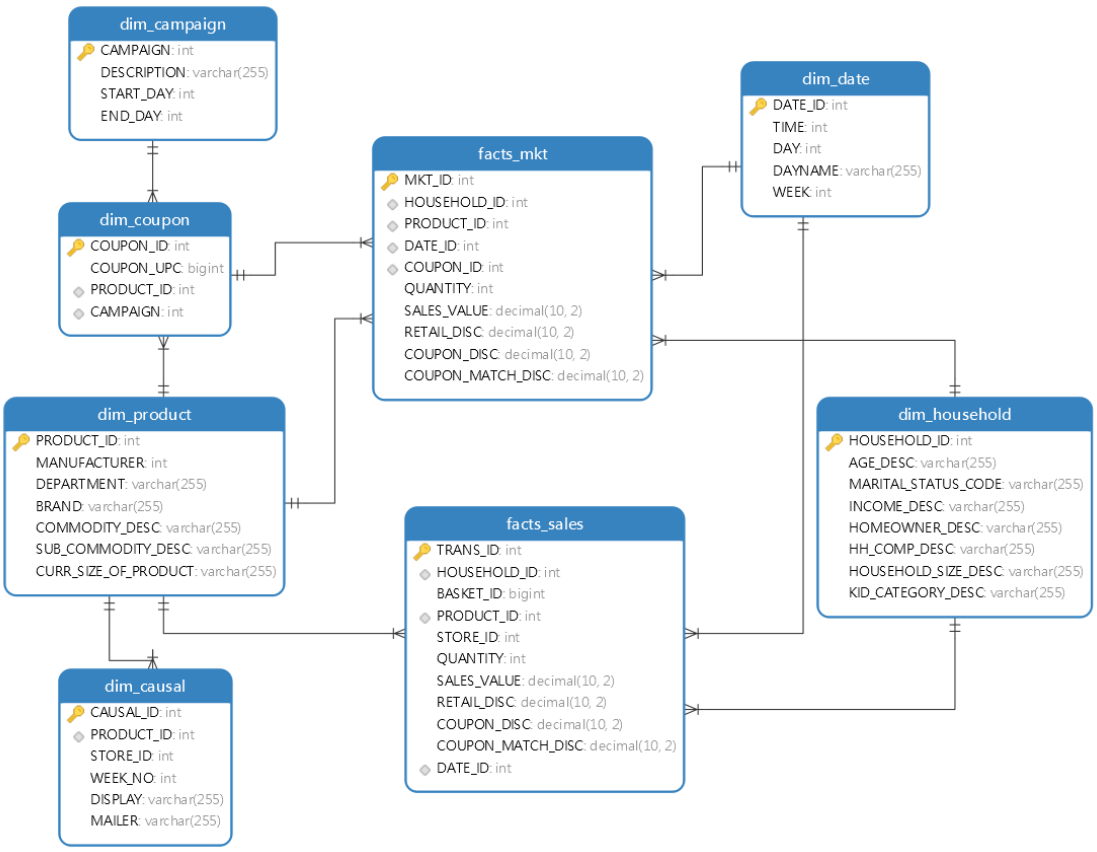
Name	Auto Increment Value	Modified Date	Data Length	Engine	Rows
campaign_desc	0		16 KB	InnoDB	30
campaign_table	0		272 KB	InnoDB	4213
causal_data	0		2816000 KB	InnoDB	351858...
coupon	0		8720 KB	InnoDB	122572
coupon_redempt	0		112 KB	InnoDB	1856
hh_demographic	0		96 KB	InnoDB	801
product	0		8720 KB	InnoDB	91226
transaction	0		131808 KB	InnoDB	1420965

Hình 4.2: Dữ liệu khu vực Staging

4.2 Xây dựng kho dữ liệu

Tương tự, tạo database "Sales_OLAP và các bảng dim - fact của cơ sở dữ liệu OLAP, thiết lập khóa chính, khóa ngoại của các bảng để tham chiếu đến bảng khác.

4.2.1 Tạo database OLAP, các bảng Dimension và Fact



Hình 4.3: Hệ thống OLAP

4.2.2 Đổ dữ liệu từ Staging vào OLAP

Sử dụng câu lệnh chèn, chọn để đổ các trường dữ liệu cần thiết vào các bảng trong database "Sales_OLAP đã tạo trước đó.

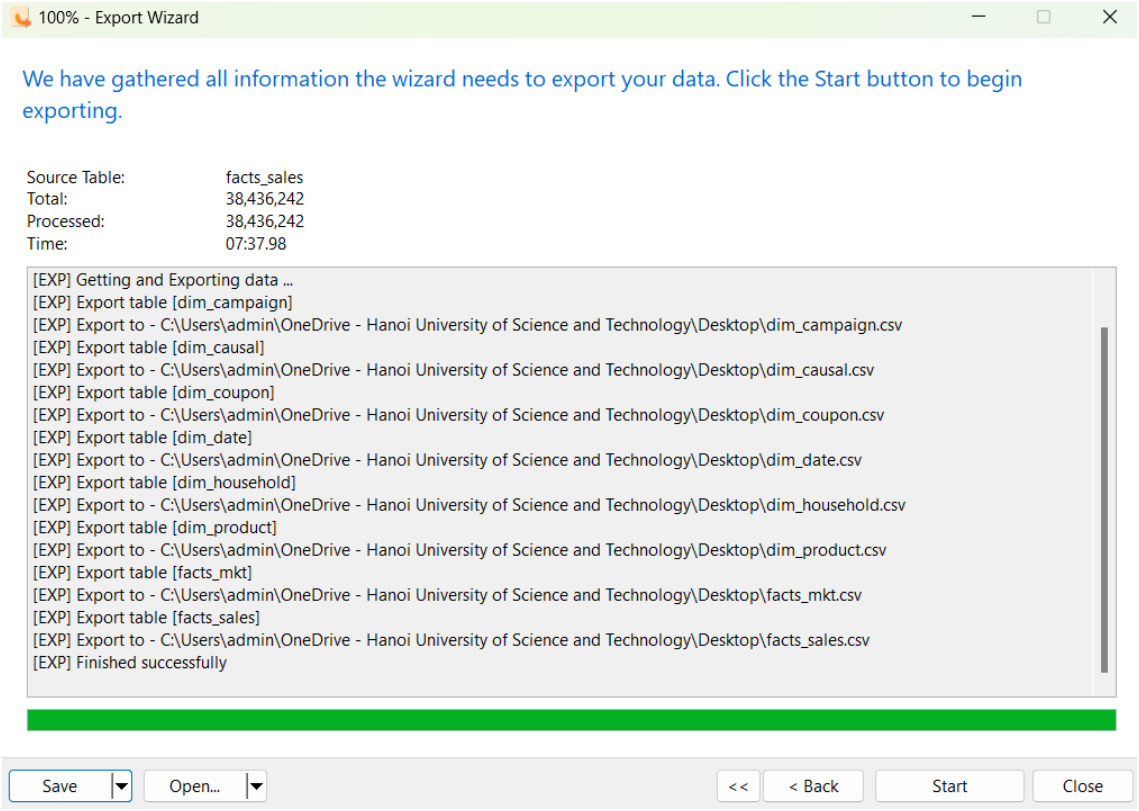
Open Table | Design Table | New Table | Delete Table | Import Wizard | Export Wizard

Name	Auto Increment Value	Modified Date	Data Length	Engine	Rows
dim_campaign	0		16 KB	InnoDB	30
dim_causal	73027380		2163712 KB	InnoDB	347628...
dim_coupon	119385		5648 KB	InnoDB	119424
dim_date	13727		1552 KB	InnoDB	13822
dim_household	0		96 KB	InnoDB	801
dim_product	0		8720 KB	InnoDB	90538
facts_mkt	1134		112 KB	InnoDB	1133
facts_sales	1422307		103040 KB	InnoDB	1416168

Hình 4.4: Dữ liệu OLAP

4.2.3 Xuất dữ liệu hệ thống OLAP

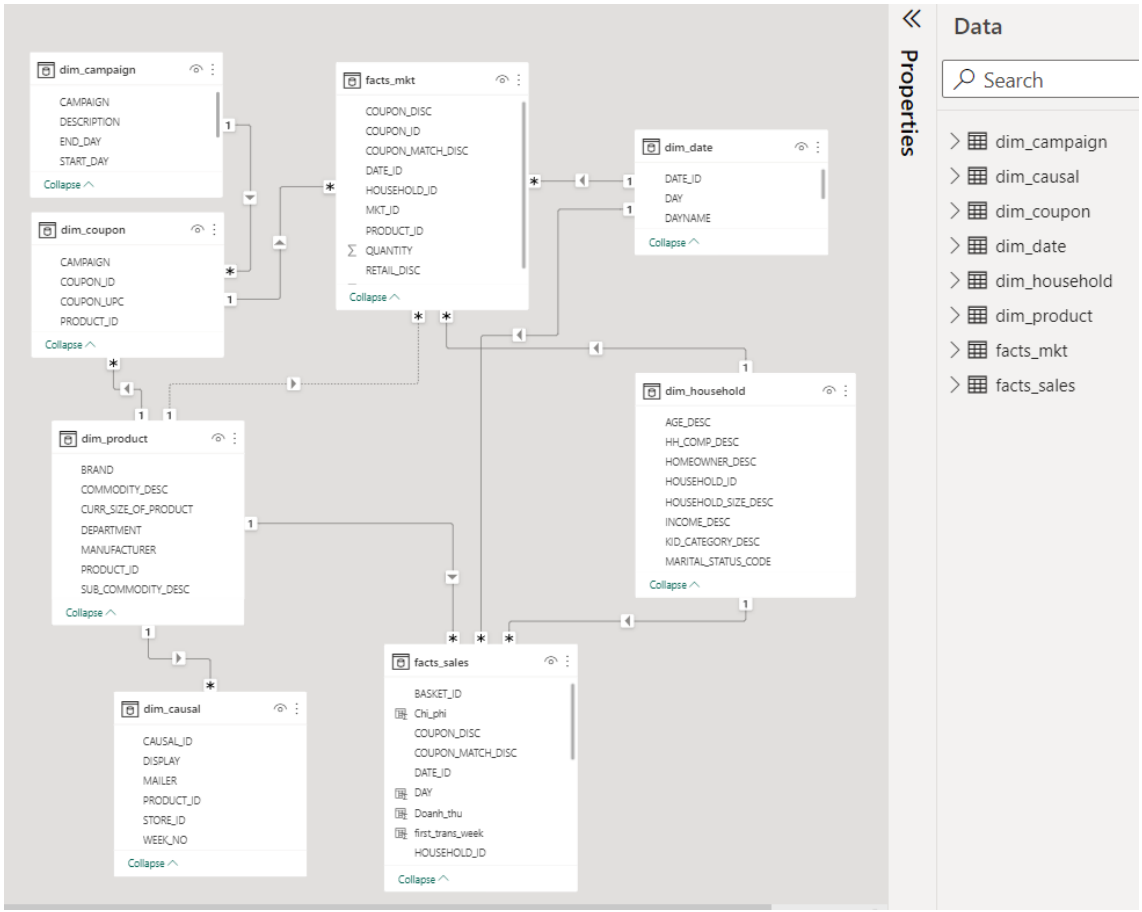
Dữ liệu đã được tổ chức theo hệ thống đa chiều OLAP ta tiến hành xuất các file dữ liệu dưới dạng csv.



Hình 4.5: Xuất dữ liệu hệ thống OLAP

4.2.4 Tải dữ liệu vào phần mềm Power BI

Sau khi có các file được tổ chức theo mô hình dữ liệu đa chiều OLAP, ta tiến hành tải dữ liệu vào phần mềm Power BI để trực quan hóa dữ liệu bằng các biểu đồ.

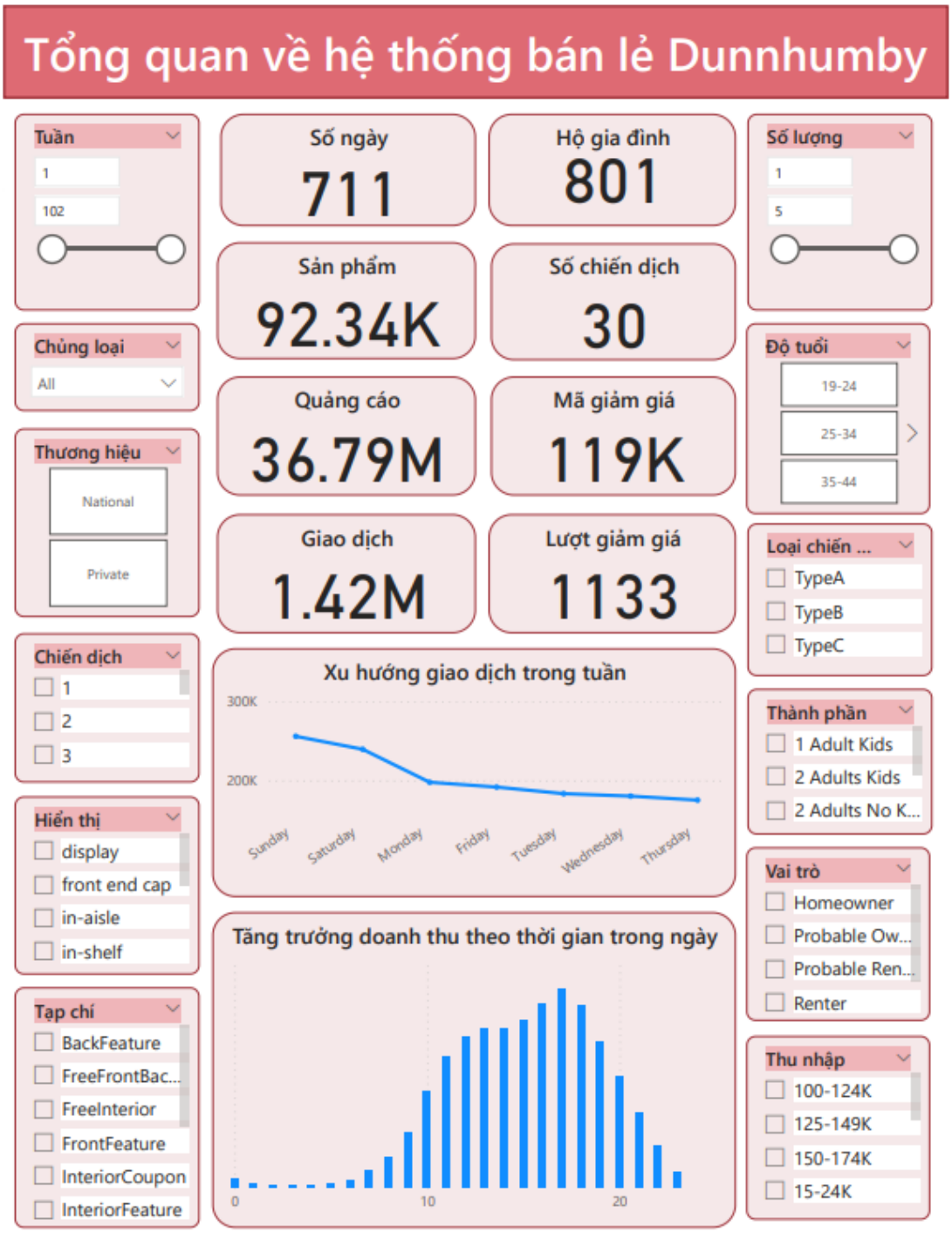


Hình 4.6: Tải dữ liệu vào Power BI

4.3 Xây dựng dashboad

Trong phần này em sử dụng các biểu đồ của phần mềm power BI trực quan hóa dữ liệu thuận tiên cho việc tiến hành phân tích các chủ điểm theo nhiều chiều.

4.3.1 Tổng quan

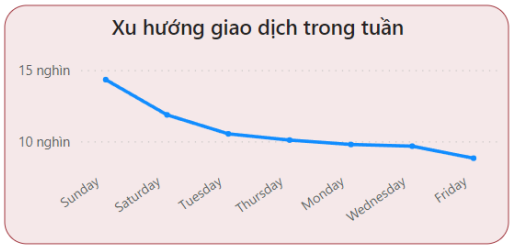


Hình 4.7: Dashboard tổng quan

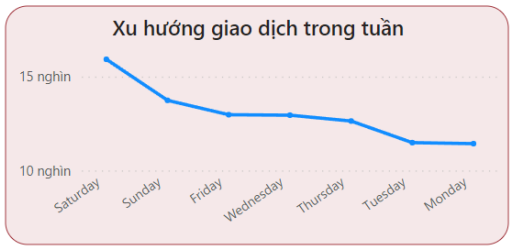
Từ dashboard tổng quan ta thấy được rằng:

- Về thời gian mua hàng trong tuần:

Những người thuộc nhóm độ tuổi dưới 55 thường mua hàng nhiều vào Chủ nhật và Thứ bảy, nhóm còn lại mua hàng nhiều vào Thứ sáu và Thứ bảy.



Hình 4.8: Nhóm tuổi dưới 55 mua hàng trong tuần

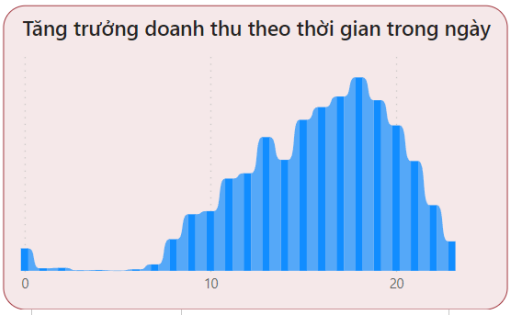


Hình 4.9: Nhóm tuổi trên 55 mua hàng trong tuần

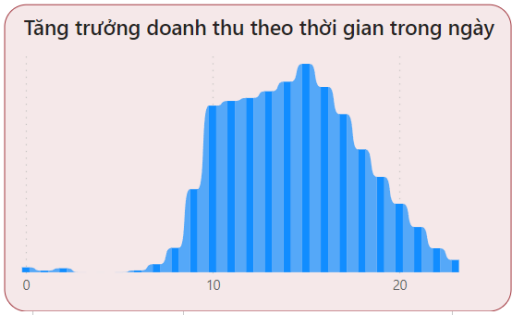
Điều đó thể hiện rằng những người trẻ thường có thói quen mua sắm nhiều vào Thứ bảy và Chủ nhật khi họ không phải đi làm. Còn nhưng người lớn tuổi có thói quen mua hàng vào Thứ sáu và Thứ bảy. Có thể do thói quen chuẩn bị và cẩn thận của những người lớn tuổi, hoặc họ mua để chuẩn bị cho gia đình, con cái ăn tiệc vào Chủ nhật.

- Về thời gian mua hàng trong ngày:

Những người thuộc nhóm tuổi dưới 25 có thói quen mua sắm trong khoảng 16 giờ đến 22 giờ.
Trong đó nhóm tuổi 25 - 45 có thói quen mua hàng vào 14 giờ đến 20 giờ.
Còn lại tuổi trên 45 có thói quen mua hàng vào 10 giờ đến 16 giờ.



Hình 4.10: Nhóm người trẻ mua hàng trong ngày

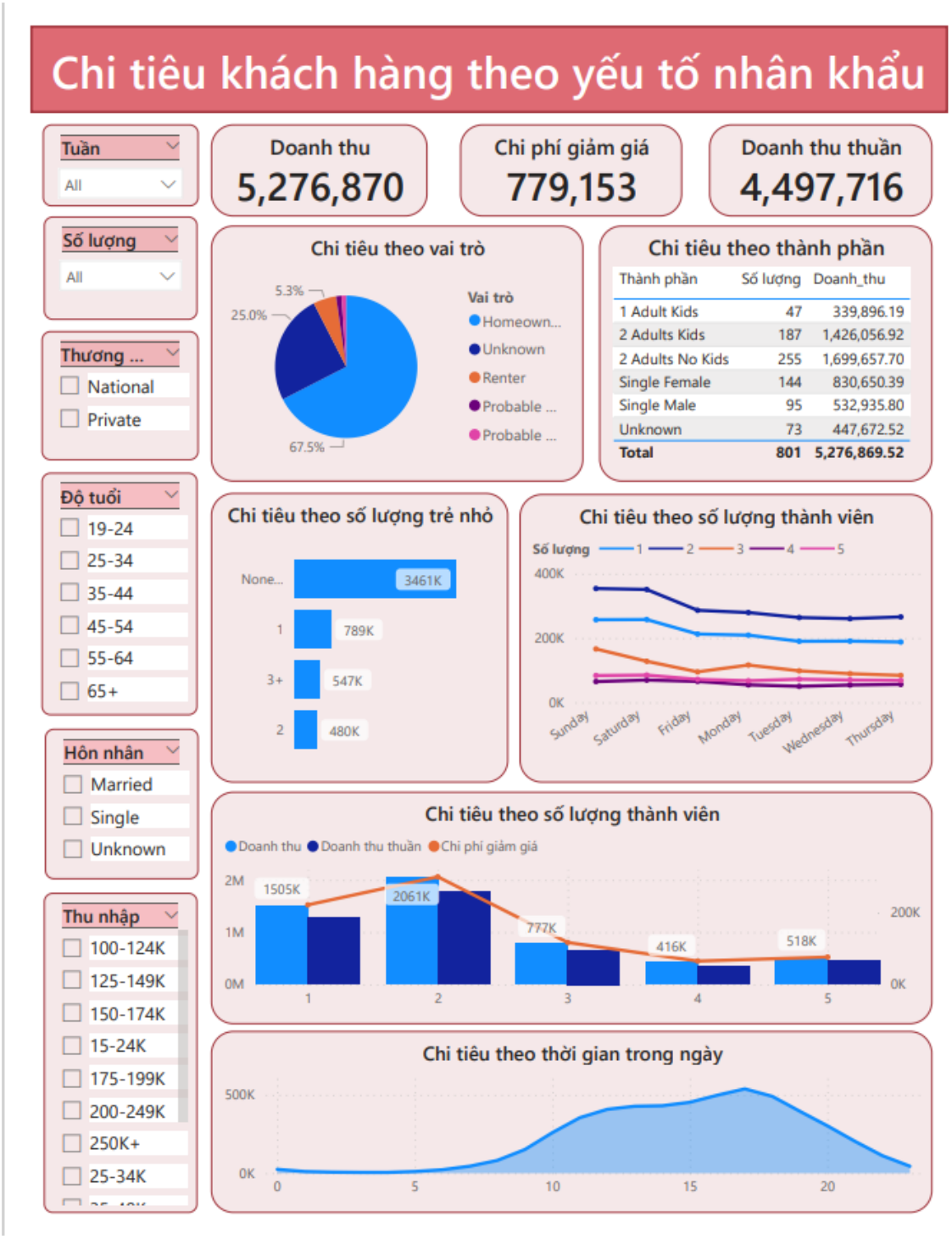


Hình 4.11: Nhóm người lớn tuổi mua hàng trong ngày

Nhận xét rằng những người trẻ, độc thân có thói quen mua hàng vào buổi tối muộn,họ năng động về đêm hơn những người lớn tuổi nhưng những người trung niên và lớn tuổi lại có thói quen mua hàng sớm trong ngày và thường vào các khoảng giờ nấu ăn.

4.3.2 Phân tích chi tiêu khách hàng theo yếu tố nhân khẩu học

1. Theo thành phần hộ gia đình



Hình 4.12: Dashboard phân tích chi tiêu khách hàng theo thành phần hộ gia đình

Dashboard cho thấy thành hộ gia đình ảnh hưởng đến nhu cầu mua sắm của khách hàng:

- Những người phụ nữ độc thân sẽ chi tiêu nhiều hơn nhưng người nam độc thân.
- Những gia đình có 2 người lớn chi tiêu nhiều hơn gia đình 2 người lớn có trẻ nhỏ.

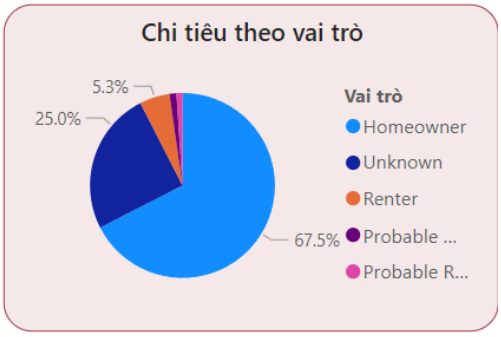
Có thể lí giải cho sự việc trên như sau:
Vì phụ nữ thường có thói quen thích mua sắm nhiều hơn nam giới.
Do các gia đình đã có con nhỏ chi tiêu tiết kiệm hơn so với các gia đình vợ chồng chưa có con.

Chi tiêu theo thành phần

Thành phần	Số lượng	Doanh_thu
1 Adult Kids	47	339,896.19
2 Adults Kids	187	1,426,056.92
2 Adults No Kids	255	1,699,657.70
Single Female	144	830,650.39
Single Male	95	532,935.80
Unknown	73	447,672.52
Tổng	**801**	**5,276,869.52**

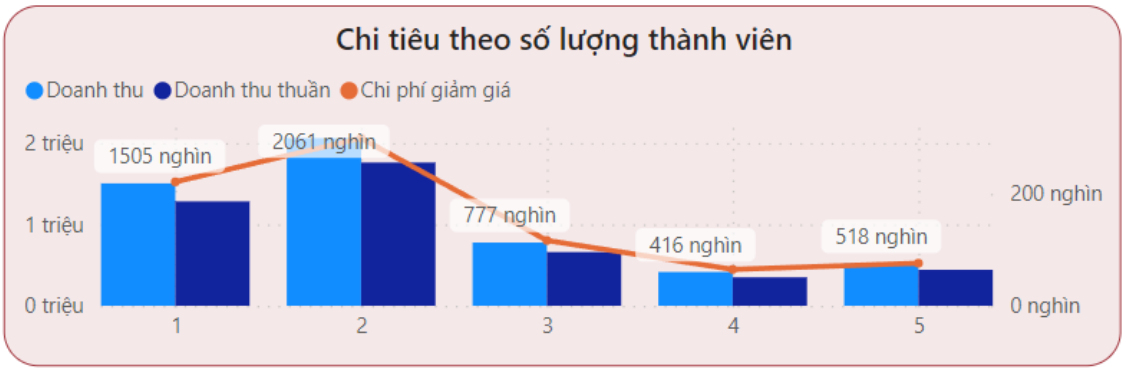
Hình 3.35: Chi tiêu theo thành phần hộ gia đình

- Những người thuê nhà mua sắm ít hơn những người là chủ nhà . Họ chủ yếu mua các sản phẩm có xuất xứ trong nước.
- Trong khi đó những người chủ nhà là thành phần chính lựa chọn các sản phẩm nước ngoài.



Hình 3.35: Chi tiêu theo thành phần hộ gia đình

Các gia đình chỉ có dưới 2 thành viên chi tiêu cho việc mua sắm nhiều gấp 2 lần tổng chi tiêu của những gia đình có hơn 2 thành viên.



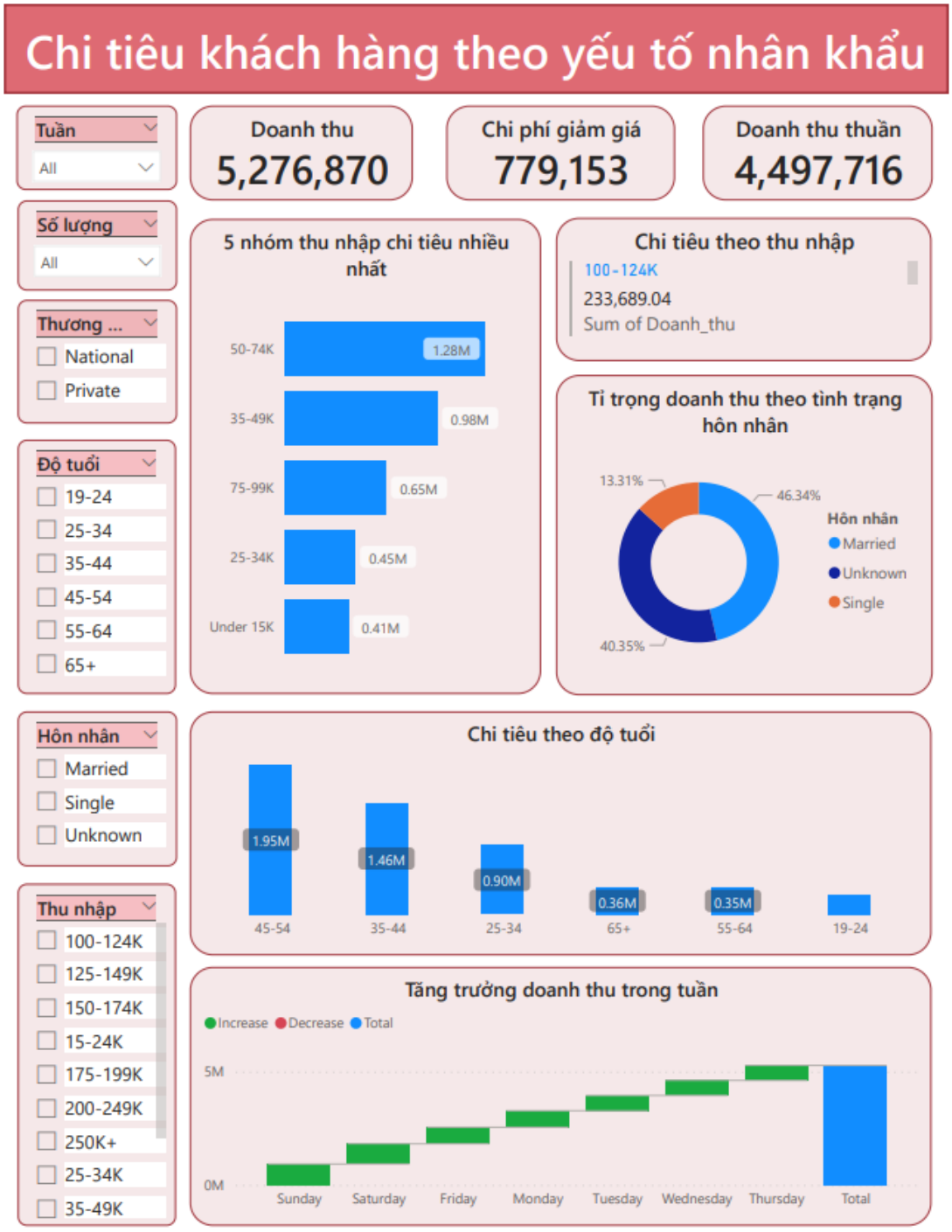
Hình 4.15: Chi tiêu của các hộ gia đình theo số lượng thành viên

Nguyên nhân có thể là vì các hộ gia đình dưới 2 thành viên là các khách hàng độc thân hoặc cặp vợ chồng nên họ không bị eo hẹp kinh tế, sẵn sàng chi tiêu với số tiền lớn. Còn nhưng gia đình nhiều thành viên phải cân nhắc hơn trong việc chi tiêu của cả gia đình.
Bên cạnh đó:

- Các gia đình có trẻ em chi tiêu nhiều về các sản phẩm như trứng, sữa bột, rau củ, bánh kẹo, thuốc...

- Các sản phẩm như mỹ phẩm, sản phẩm làm đẹp được tiêu thụ phần lớn bởi nhưng người phụ nữ độc thân và những gia đình chưa có trẻ nhỏ.
- Các sản phẩm như báo, thiệp được mua nhiều bởi các khách hàng thuộc nhóm thuê nhà trong khi đó các sản phẩm gia dụng, vật liệu xây dựng được mua chủ yếu bởi chủ nhà.

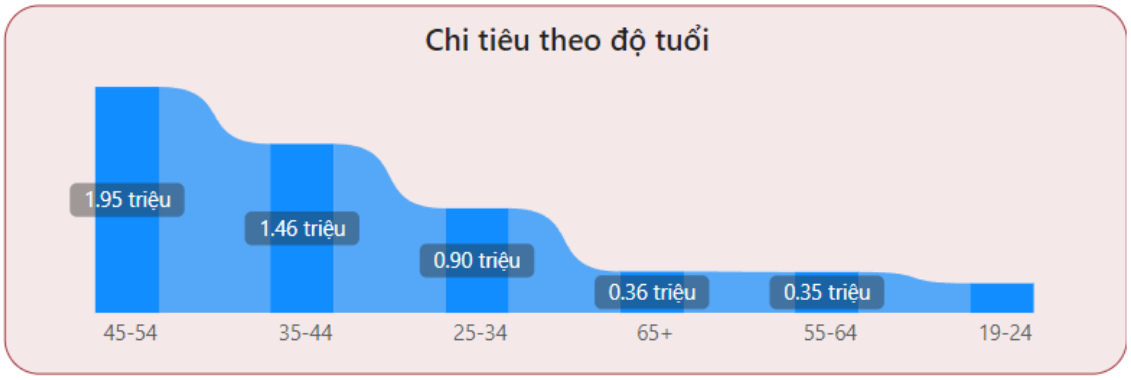
2. Theo các đặc điểm hộ gia đình



Hình 4.16: Dashboard phân tích chi tiêu khác hàng theo đặc điểm hộ gia đình

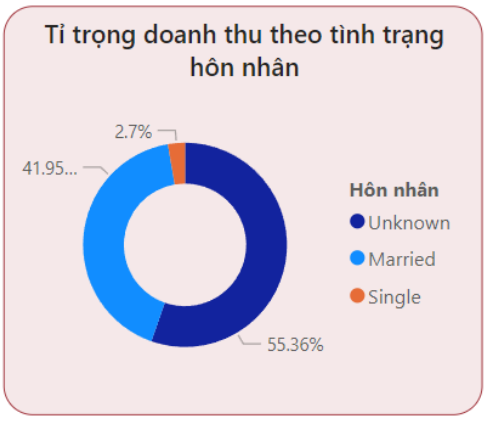
- Chủ yếu chi tiêu nhiều ở các nhóm thu nhập từ thấp đến trung bình.
- Độ tuổi 25 đến 55 chi tiêu nhiều gấp đôi các nhóm tuổi còn lại.

Có thể giải thích rằng do các mức thu nhập cao thường có nhiều lựa chọn hơn trong việc mua sắm, nhóm tuổi khác ngoài khoảng 22 đến 55 là nhóm những người trẻ và người lớn tuổi, nhu cầu mua sắm cũng như kinh tế ít hơn các nhóm tuổi khác.



Hình 4.17: Chi tiêu theo độ tuổi

- Các sản phẩm như bánh kẹo, đồ ăn nhanh chủ yếu được tiêu thụ ở nhóm đã kết hôn.
- Các dịch vụ giải trí, xem phim, chụp ảnh được tiêu thụ chính bởi nhóm tuổi 19 đến 35 tuổi.



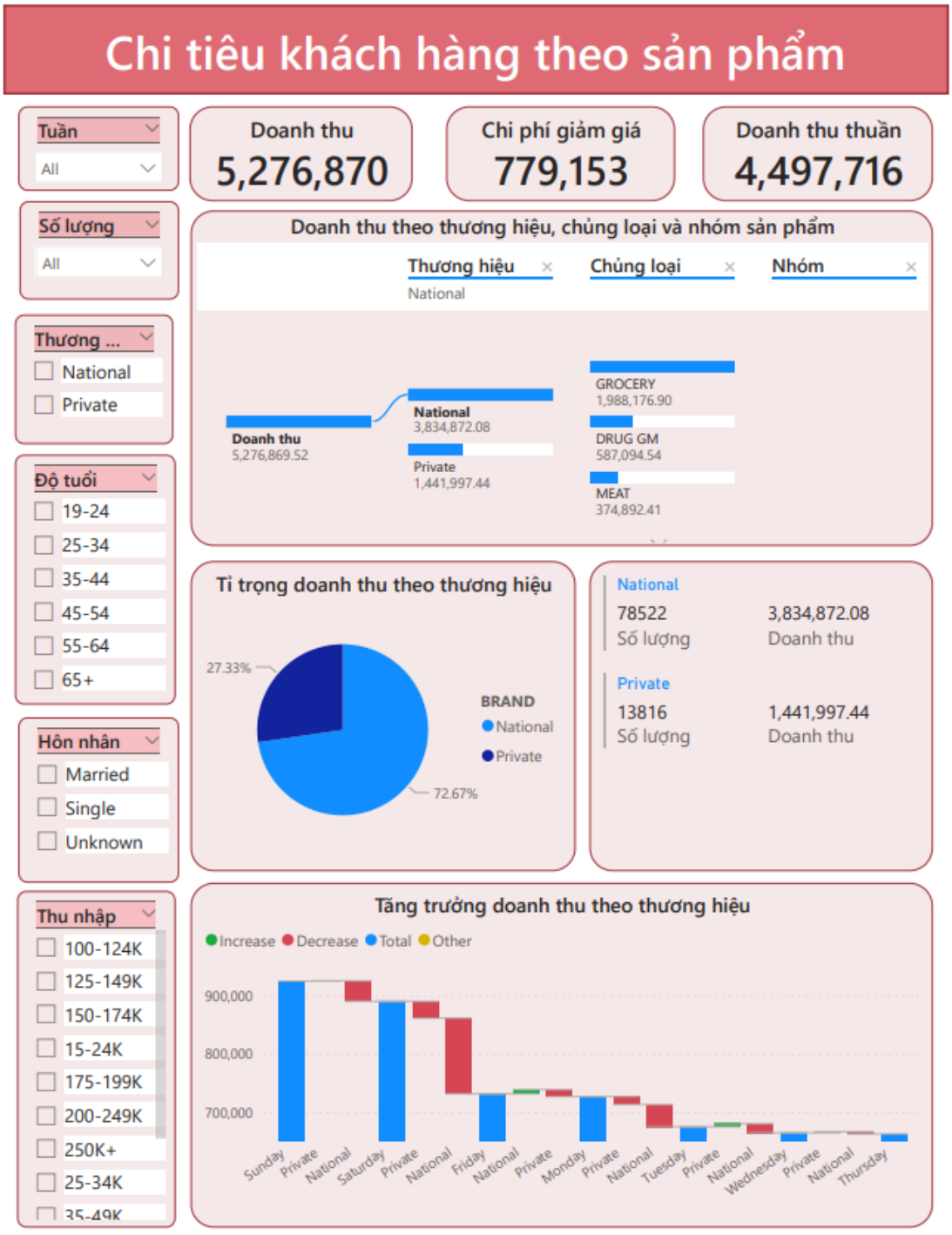
Hình 3.35: Chi tiêu các sản phẩm bánh kẹo, đồ ngọt theo tình trạng hôn nhân

Bên cạnh đó:

- Các sản phẩm nhóm du lịch, giải trí chủ yếu được chi tiêu bởi những khách hàng ở mức thu nhập trung bình từ 50 đến 100K.
- Trong khi nhóm khách hàng đã kết hôn là nhóm chủ yếu chi tiêu cho các dịch vụ nhà hàng, tổ chức tiệc.

4.3.3 Phân tích chi tiêu khách hàng theo yếu tố sản phẩm

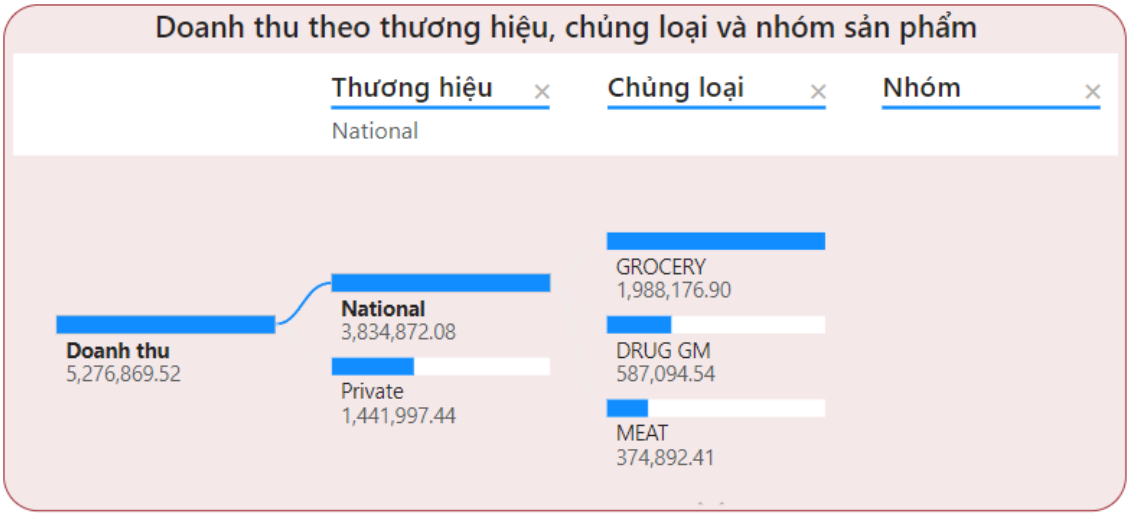
1. Theo thương hiệu sản phẩm



Hình 4.19: Dashboard phân tích chi tiêu theo thương hiệu sản phẩm

Từ dashboard trên nhận thấy rằng:

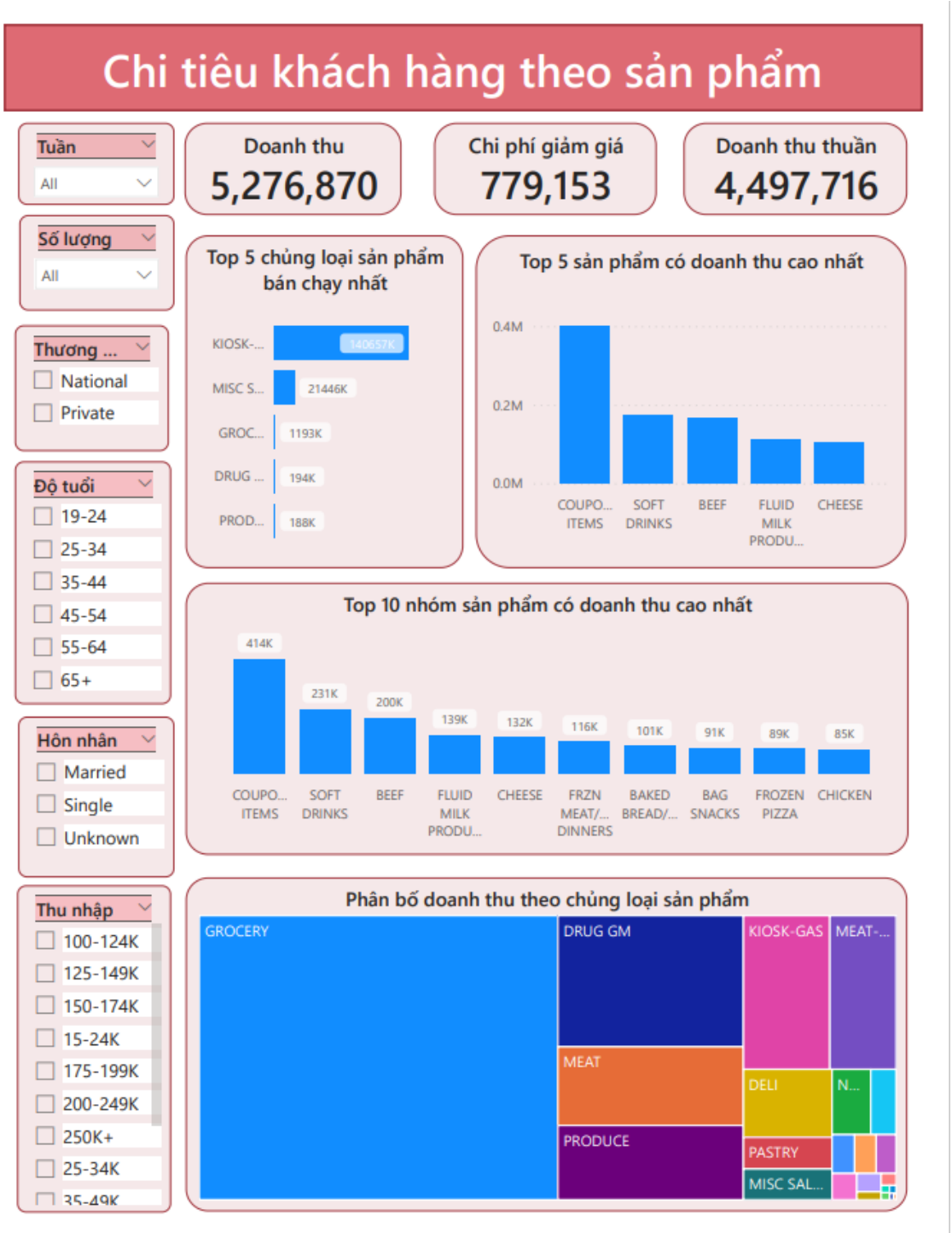
- Doanh thu từ các sản phẩm nhập khẩu cao gấp 3 lần doanh thu của các sản phẩm có xuất xứ trong nước. Sự chênh lệch này càng rõ hơn ở các nhóm tuổi từ 25 đến 45 tuổi.



Hình 4.20: Doanh thu theo thương hiệu,chủng loại và nhóm sản phẩm

- Chi tiêu cho sản phẩm nước ngoài tỉ lệ thuận với mức thu nhập, tức là những người có mức thu nhập cao sẽ chi tiêu cho các sản phẩm nước ngoài hơn những người ở các mức thu nhập thấp hơn.

2. Theo chủng loại, nhóm sản phẩm



Hình 4.21: Dashboard phân tích chi tiêu theo chủng loại, nhóm sản phẩm

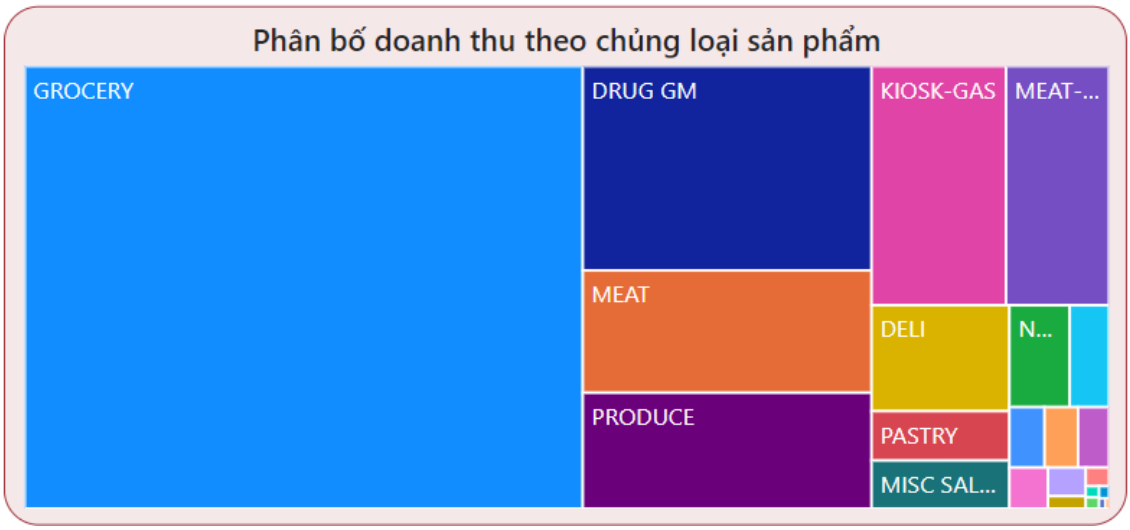
Nhóm sản phẩm dịch vụ, đồ uống nhanh, sữa và bim bim là những mặt hàng đứng đầu về số lượt bán và doanh thu.

Đặc biệt nhóm sản phẩm nhiên liệu, khí đốt ga là nhóm sản phẩm bán chạy nhất, chênh lệch lớn với các sản phẩm còn lại trong top.

- Những hộ gia đình có quy mô nhỏ và vừa 1 đến 2 thành viên sử dụng nhiều khí gas hơn.
- Các khách hàng đã kết hôn sử dụng khi gas nhiều gấp 4 lần các khách hàng độc thân.

Điều này có thể giải thích do những người độc thân không ở nhà nhiều, thường xuyên ăn ngoài nên không sử dụng nhiên liệu.

Hình 3.35: Top 5 chủng loại bán chạy nhất



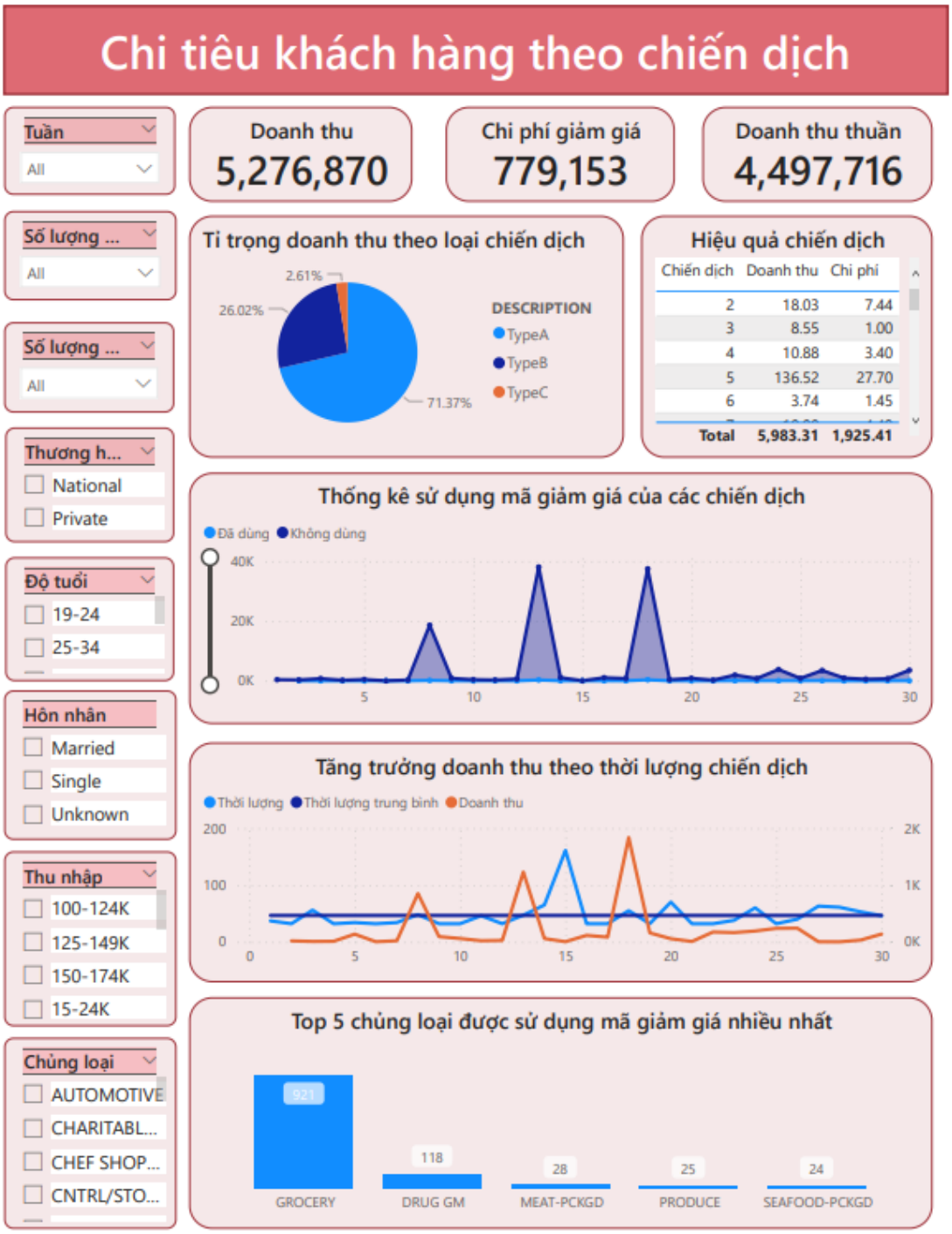
Hình 4.23: Phân bố doanh thu theo chủng loại

Khí gas là chủng loại bán chạy nhất, xếp thứ ba là GROCERY tuy nhiên doanh thu khí gas lại xếp sau GROCERY và hai chủng loại khác. Lí do có thể là vì khí gas là nhiên liệu dùng hằng ngày, giá thành rẻ nên dù xếp đầu về doanh số nhưng doanh thu vẫn đứng sau các chủng loại có giá thành cao.

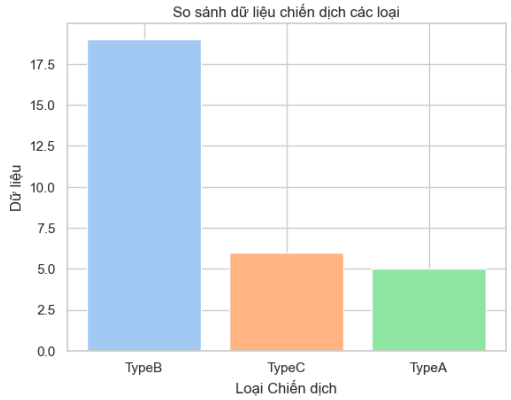
Nhóm tuổi dưới 45 có nhu cầu cao về các mặt hàng như dịch vụ, nhà hàng, đồ uống có cồn. Trong khi đó nhóm tuổi trên 45 thường chi tiêu nhiều cho các loại thực phẩm chức năng, sữa, các loại hạt, ngũ cốc và cháo.

Giải thích cho sự khác biệt này là do nhóm tuổi trung niên thường xuyên tổ chức tiệc, vui chơi, ăn uống đồ ăn nhanh còn người lớn tuổi sử dụng nhiều thực phẩm chức năng, sữa, các loại hạt để đảm bảo sức khỏe và ăn các thức ăn dễ ăn như cháo, soup.

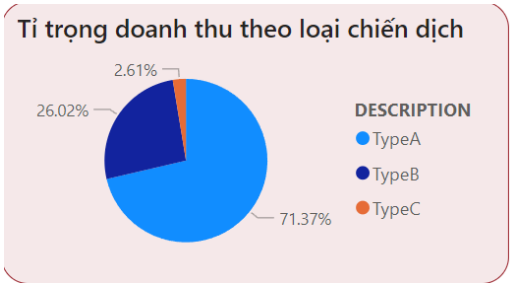
4.3.4 Phân tích chi tiêu khách hàng theo chiến dịch



Hình 4.24: Dashboard phân tích chi tiêu theo các chiến dịch



Hình 4.25: Phân bố chiến dịch theo loại chiến dịch



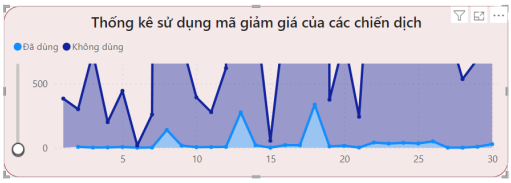
Hình 4.26: Tỉ trọng doanh thu theo loại chiến dịch

Tỉ trong doanh thu theo các loại chiến dịch có sự phân hóa rõ ràng, chiếm đến 2/3 doanh thu là các chiến dịch loại A, sau đó là chiến dịch loại B 26% và thấp nhất là chiến dịch loại C 3%.

Trong quá trình tổ chức các chiến dịch, chiến dịch loại A chỉ chiếm 1/6 số lượng chiến dịch nhưng doanh thu mang lại đến 2/3. Chiến dịch loại B chiếm đến 2/3 tổng số chiến dịch nhưng kết quả doanh thu chỉ bằng 1/3 chiến dịch loại A.

Điều này cho thấy chiến dịch loại A rất được ưa chuộng bởi khách hàng, doanh nghiệp cần định hình lại cách tổ chức cũng như số lượng chiến dịch các loại để đạt hiệu quả cao nhất trong việc thu hút khách hàng tham gia.

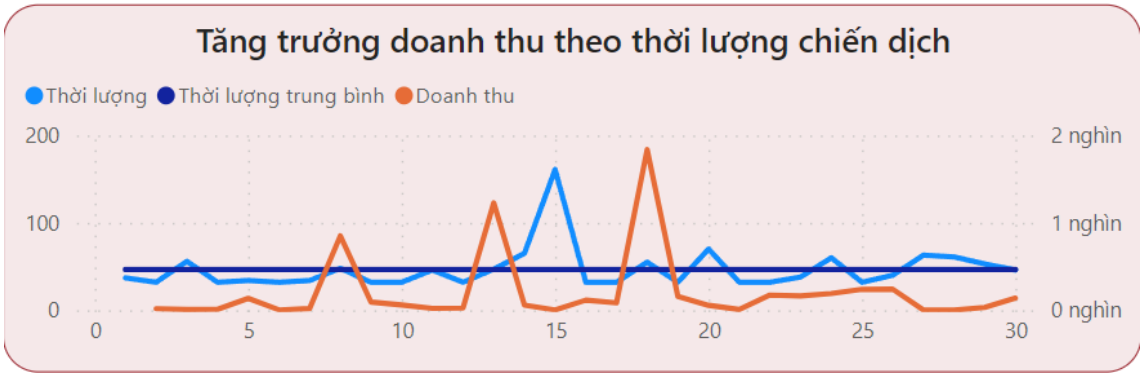
Hình 4.27: Số lượng mã giảm giá được tung ra và sử dụng



Hình 4.28: Mức độ sử dụng các mã giảm giá được tung ra

Các mã giảm giá được tung ra từ các chiến dịch luôn không được sử dụng hết.

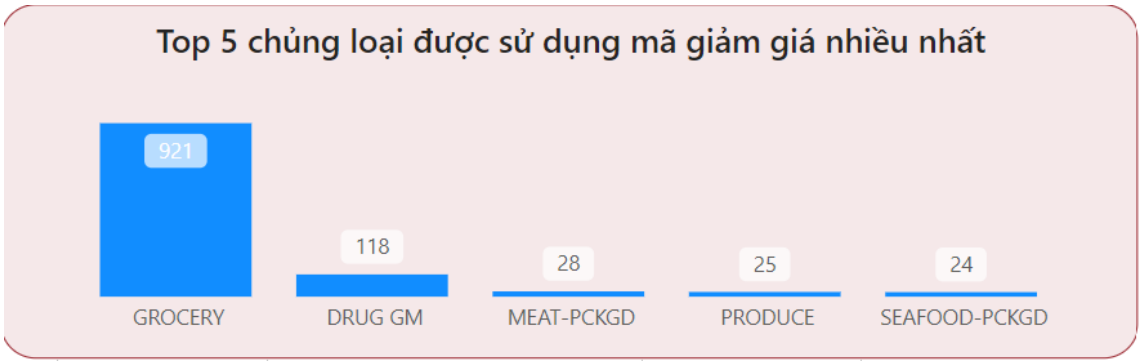
Tuy nhiên số lượng các mã giảm giá được tung ra từ các chiến dịch khá đồng đều ngoại trừ ba chiến dịch 8, 13, 18 có số lượng mã giảm giá tăng vượt mức bình thường. Đồng thời, số lượng mã giảm giá được sử dụng cũng tăng theo.



Hình 4.29: Tăng trưởng doanh thu theo thời lượng chiến dịch

Như trên đã đánh giá, các chiến dịch 8, 13, 18 là các chiến dịch đem lại doanh thu cao nhất, đứng đầu là chiến dịch thứ 18. Tìm hiểu cụ thể hơn về chiến dịch này, đây là chiến dịch loại A đem về doanh thu 1.8 triệu chiếm gần 2/3 tổng doanh thu của 30 chiến dịch.

Thời lượng kéo dài của các chiến dịch phân bố khá đồng đều ngoại trừ chiến dịch 15 được tổ chức kéo dài đến 161 ngày, tuy nhiên số lượng mã giảm giá tung ra chỉ tương đương với các chiến dịch bình thường. Điều này giải thích một phần lí do doanh thu chiến dịch này thấp dù được kéo dài khá lâu.



Hình 4.30: Top 5 chủng loại sản phẩm được sử dụng nhiều mã giảm giá nhất

Mặc dù cùng năm trong top 5 chủng loại được sử dụng mã giảm giá nhiều nhất nhưng có sự chênh lệch lớn giữa chủng loại đứng đầu GROCERY và các chủng loại khác trong top 5.

GROCERY là chủng loại luôn nằm trong top những sản phẩm bán chạy và đứng đầu về doanh thu, bởi nó bao gồm các mặt hàng nhu yếu phẩm cần thiết. Số lượng mã giảm giá được dùng cho chủng loại này trong các chiến dịch cũng luôn đứng đầu. Đặc biệt trong chiến dịch 18, số lượng mã giảm giá dùng cho chủng loại này chiếm đến gần 1/4 tổng số lần sử dụng cho 30 chiến dịch.

Xét theo mức độ thu nhập, phần lớn những khách hàng sử dụng mã giảm giá có mức thu nhập trung bình từ 35K - 70K và hầu như không xuất hiện ở nhóm có thu nhập cao trên 200K, điều này đúng với nhận định những người có mức thu nhập vừa sẽ quan tâm việc giảm giá để tiết kiệm chi phí mua hàng hơn.

Báo cáo cũng chỉ ra rằng các mã giảm giá được sử dụng chủ yếu bởi nhóm khách hàng có độ tuổi từ 35 - 54. Những khách hàng trong gia đình có số lượng từ 1 đến 2 thành viên hay đã kết hôn đóng góp phần lớn trong số lượng mã giảm giá được sử dụng.

4.3.5 Kết luận

Từ quá trình phân tích dữ liệu hệ thống, thông qua việc phân tích chi tiêu dựa trên đặc điểm của các yếu tố sản phẩm, khách hàng và các chiến dịch markerting em có những đề xuất sau:

1. Về mặt khách hàng: Thường xuyên thu thập, phân tích dữ liệu khách hàng, quan tâm đến các nhóm gia đình chiếm tỉ trọng doanh thu lớn để phục vụ cho việc đề xuất sản phẩm, phát triển chiến lược nhắm đến hành vi mua hàng phụ thuộc vào đặc điểm nhân khẩu của khách hàng.
2. Về mặt sản phẩm: Lựa chọn phân phối thương hiệu sản phẩm, sắp xếp các cặp sản phẩm xuất hiện cùng nhau hợp lí để nắm được hành vi mua sắm của khách hàng với từng chủng loại, nhóm sản phẩm. Phân nhóm các sản phẩm chính, phụ, cùng nhau dựa trên sự quan tâm của khách hàng khi mua một sản phẩm nào đó.
3. Về mặt chiến dịch marketing: Nghiên cứu sự phân bổ hợp lí về thời lượng, số lượng mã giảm giá trong từng loại chiến dịch. Đồng thời dựa trên hành vi sử dụng mã giảm giá của khách hàng đưa ra những mã giảm giá ứng với từng loại sản phẩm phù hợp với nhu cầu người tiêu dùng.

Trên đây là những ý kiến của cá nhân em rút ra từ quá trình phân tích. Như vậy, các doanh nghiệp nên có những biện pháp tốt và phù hợp để tối ưu chi phí, nâng cao hiệu quả, thích nghi với sự biến động thất thường của thị trường kinh doanh hiện nay.

CHƯƠNG 5. TỔNG KẾT

5.1 Kết luận

- Trong quá trình thực hiện đồ án em đã đạt được các kết quả sau:
- Nắm được quy trình phân tích dữ liệu ứng dụng vào kinh doanh thông minh
- Thiết kế và xây dựng được hệ thống lưu trữ quản lí dữ liệu đa chiều
- Xây dựng được hệ thống các báo cáo phân tích

Do thời gian và kiến thức còn hạn chế nên hệ thống của em vẫn còn nhiều thiếu sót:

- Chưa tự động được các quy trình của hệ thống
- Phạm vi ứng dụng nhỏ và hẹp
- Các tính năng còn hạn chế

Em rất mong nhận được sự góp ý của quý thầy cô và bạn đọc để đề tài này được hoàn thiện hơn.

5.2 Hướng phát triển

Hướng phát triển trong thời gian tới của đề tài có thể như sau:

- Hoàn thiện thêm các chức năng
- Tự động hóa quy trình hệ thống
- Mở rộng để ứng dụng rộng rãi hơn

TÀI LIỆU THAM KHẢO

[1] M. Warrilow, V. President **and** R. Director, *Worldwide Data and Analytics Forecast*. Gartner, 2022-07-25.

[2] F. Provost **and** T. Fawcett, *Data Science for Business*. O'Reilly Media, 2022, **pages** 1–7.

[3] I. I. of Business Analysis, *Business Analysis Body of Knowledge (BABOK Guide)*. Wiley, 2022, **pages** 1–7.

[4] R. Sherman, *Business Intelligence Guidebook*. Morgan Kaufmann, 2014, **pages** 1–7.

[5] K. Gupta **and** V. Rao, *The Complete Journey*. 2018.

[6] M. Kleppmann, *Designing Data-Intensive Applications*. O'Reilly Media, 2017, **pages** 2–5.

[7] N. V. Ba, *Phân tích và thiết kế hệ thống thông tin*. Nhà xuất bản Đại học quốc gia Hà Nội, 2006.

[8] N. D. Tú, *Slide Kho dữ liệu và kinh doanh thông minh*. 2023.