Khám phá dữ liệu:

1. Tổng quan về bộ dữ liệu:

* Kích thước: 847.05MB
* Gồm 8 bảng:
* hh\_demographic: mô tả thông tin nhân khẩu học
* transaction\_data: mô tả thông tin sản phẩm được mua bởi mỗi hộ gia đình trong nghiên cứu
* campaign\_table: danh sách các chiến dịch mỗi nhà nhận được
* campaign\_desc: thời gian một chiến dịch diễn ra
* product: thông tin tất cả các sản phẩm
* coupon: mã giảm giá được gửi cho các hộ gia đình trong các chiến dịch
* coupon\_redempt: xác định các coupon đã được đổi
* casual\_data:cho thông tin sản phẩm được giới thiệu trong bưu phẩm hang tuần hay được trưng bày tại cửa hàng
* Đối tượng nghiên cứu gồm có 2500 hộ gia đình
* Kéo dài trong 711 ngày
* Đã diễn ra 276484 giao dịch trên 582 cửa hàng khác nhau.
* Tổng số tiền đã giao dịch là hơn 8tr$, với hơn 9000 mặt hàng khác nhau

1. Các thông tin thống kê về dữ liệu:
2. Thông tin nhân khẩu học khách hàng:

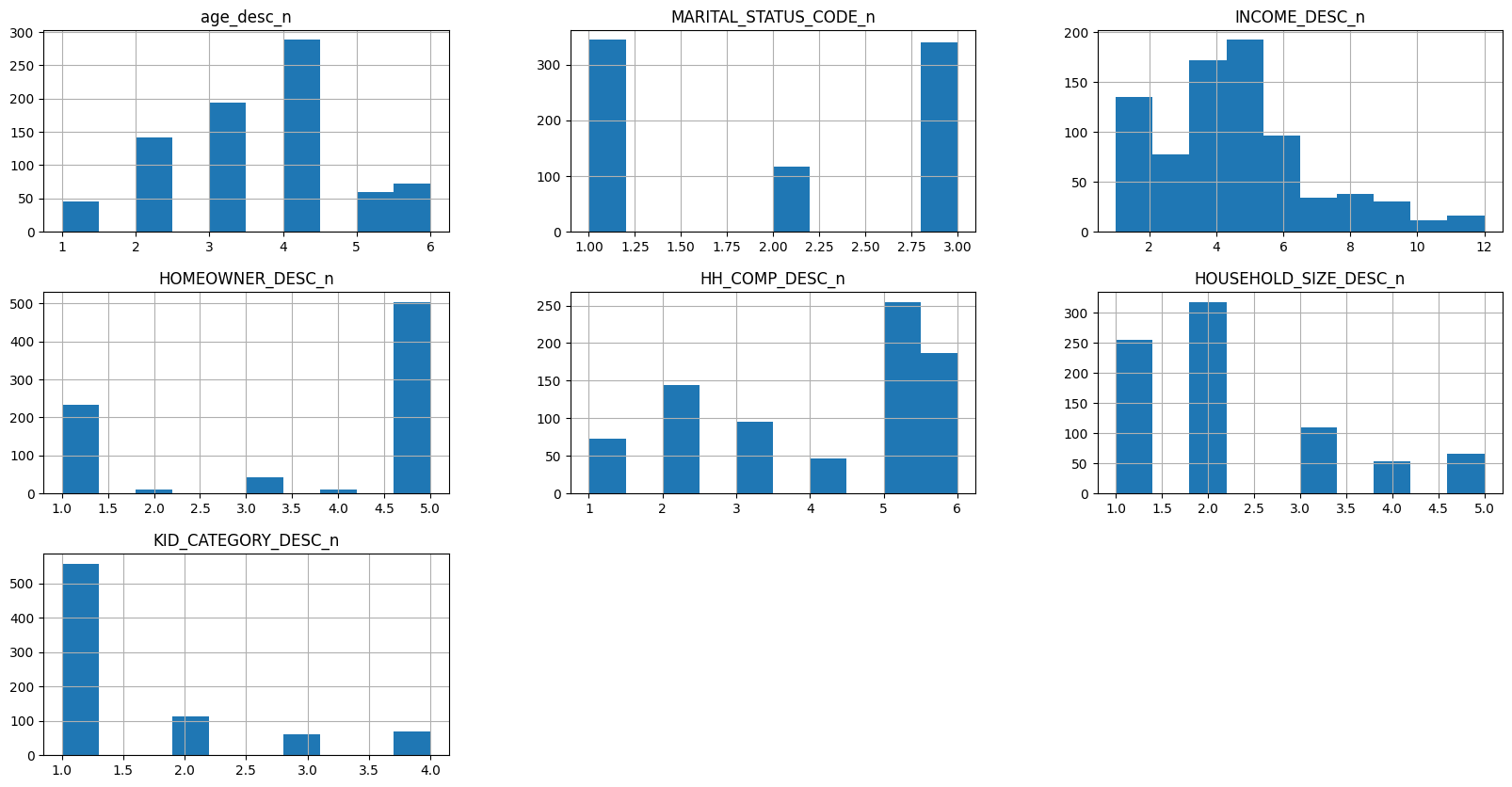
Khi mua hàng, dữ liệu về nhân khẩu học của khách hàng được thu thâp để giúp nhà bán lẻ hiểu rõ hơn về khách hàng của mình. Trên thực tế, có nhiều khách hàng không chấp nhận cũng cấp thông tin, cũng như cung cấp thông tin thiếu. Dữ liệu này được thể hiện ở bảng hh\_demographic. Tuy nhiên, dựa trên những dữ liệu đầy đủ của các hộ gia đình thu thập được, ta thực hiện mapping:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age\_desc | Martial\_status | Incom\_desc | Homeowner\_desc | Hh\_comp\_desc | Hh\_size | Kid\_category | Giá trị |
| 19-24 | U | Under15K | Unknown | Unknown | 1 | None/Unknown | 1 |
| 25-34 | B | 15-24K | Probable Renter | Single Female | 2 | 1 | 2 |
| 35-44 | A | 25-34K | Renter | Single Male | 3 | 2 | 3 |
| 45-54 |  | 35-49K | Probable Owner | 1 Adult Kids | 4 | 3+ | 4 |
| 55-64 |  | 50-74K | Homeowner | 2 Adults No Kids | 5+ |  | 5 |
| 65+ |  | 75-99K |  | 2 Adults Kids |  |  | 6 |
|  |  | 100-124K |  |  |  |  | 7 |
|  |  | 125-149K |  |  |  |  | 8 |
|  |  | 150-174K |  |  |  |  | 9 |
|  |  | 175-199K |  |  |  |  | 10 |
|  |  | 200-249K |  |  |  |  | 11 |
|  |  | 250K+ |  |  |  |  | 12 |

Theo đó, ta lập một bảng thống kê đơn giản:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Độ tuổi | Tình trạng hôn nhân | Mức thu nhập | Trạng thái sở hữu nhà | Thành phần hộ gia đình | Số lượng thành viên | Số lượng trẻ con trong nhà |
| count | 801 | 801 | 801 | 801 | 801 | 801 | 801 |
| mean | 3.48439 | 1.995006 | 4.710362 | 3.676654182 | 4.0337079 | 2.1972534 | 1.5505618 |
| std | 1.27968 | 0.924649 | 2.2646203 | 1.80390321 | 1.7066377 | 1.193123 | 0.95668846 |
| min | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 25% | 3 | 1 | 3 | 1 | 2 | 1 | 1 |
| 50% | 4 | 2 | 5 | 5 | 5 | 2 | 1 |
| 75% | 4 | 3 | 6 | 5 | 5 | 3 | 2 |
| max | 6 | 3 | 12 | 5 | 6 | 5 | 4 |

Để dễ hình dung các thông tin nhân khẩu học của khách hàng, ta có thể biểu diễn những dữ liệu này dưới dạng histogram:



Nhìn vào histogram, ta có thể nhận thấy, trong số 801 hộ gia đình cung cấp thông tin đầy đủ về nhân khẩu học của họ:

* Phần lớn khách hàng ở trong độ tuổi 35-54. Trong đó độ tuổi từ 45-54 chiếm phần lớn (độ tuổi trung niên, có thu nhập cao và nhiều thời gian)
* Tỉ lệ khách hàng đã kết hôn nhiều hơn, phần lớn là chủ sở hữu nhà
* Phần lớn khách hàng có thu nhập dưới 100.000$
* Phần lớn khách hàng tới từ hộ gia đình chỉ có 1-2 người

1. Thông tin giao dịch:

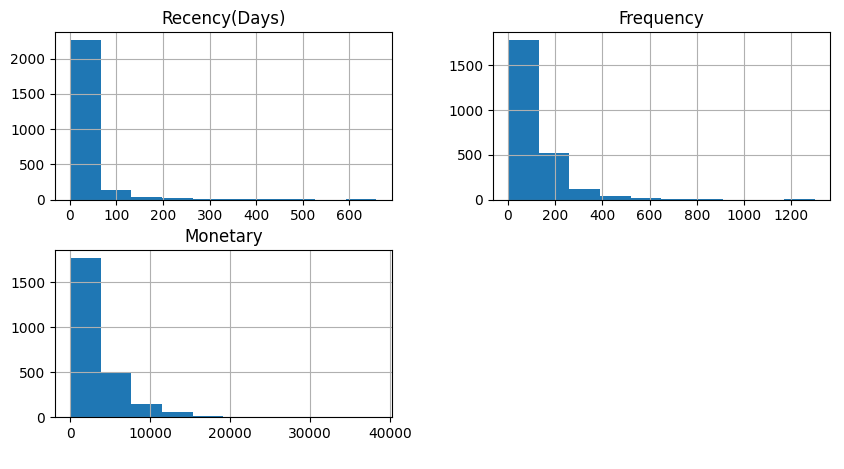
Sau khi tìm hiểu về khách hàng, điều quan trọng hơn cả là các thông tin về giao dịch đã thực hiện. Các thông tin này được nhà bản lẻ thu thập ở mỗi giao dịch và trong bộ dữ liệu này được lưu ở bảng transaction\_data. Thông tin này không chỉ cho biết các xu hướng mua hàng, nhu cầu của khách hàng… để nhà bán lẻ đưa ra chính sách nhập hàng/sản xuất hợp lý.

Thống kê chung về dữ liệu giao dịch, ta thấy được:

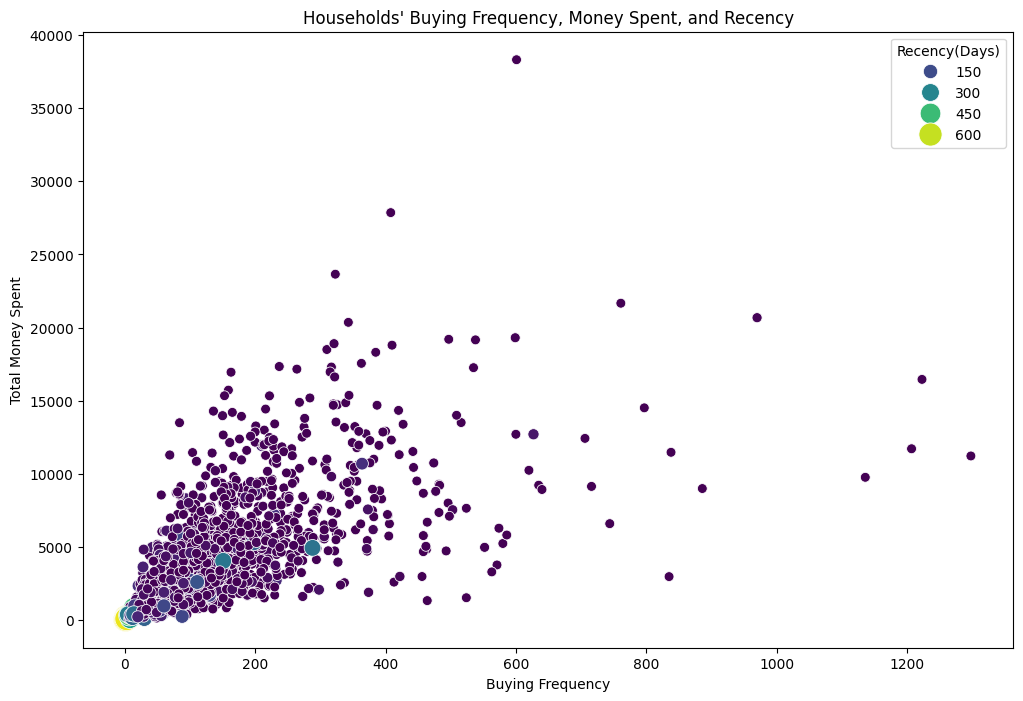
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | DAY | PRODUCT\_ID | QUANTITY | SALES\_VALUE |
| count | 2595732 | 2595732 | 2595732 | 2595732 |
| mean | 388.7562218 | 2891435.16 | 100.4285581 | 3.104119794 |
| std | 189.7209953 | 3837403.689 | 1153.436211 | 4.182274119 |
| min | 1 | 25671 | 0 | 0 |
| 25% | 229 | 917459 | 1 | 1.29 |
| 50% | 390 | 1028816 | 1 | 2 |
| 75% | 553 | 1133018 | 1 | 3.49 |
| max | 711 | 18316298 | 89638 | 840 |
|  |  |  |  |  |
| RETAIL\_DISC | TRANS\_TIME | WEEK\_NO | COUPON\_DISC | COUPON\_MATCH\_DISC |
| 2595732 | 2595732 | 2595732 | 2595732 | 2595732 |
| -0.53871 | 1561.586139 | 56.22149898 | -0.016416001 | -0.002918564 |
| 1.249191 | 399.8378199 | 27.1022279 | 0.21684099 | 0.039690043 |
| -180 | 0 | 1 | -55.93 | -7.7 |
| -0.69 | 1308 | 33 | 0 | 0 |
| -0.01 | 1613 | 56 | 0 | 0 |
| 0 | 1843 | 80 | 0 | 0 |
| 3.99 | 2359 | 102 | 0 | 0 |
|  |  |  |  |  |

Dựa trên 2.595.732 giao dịch, kết hợp cùng thông tin nhân khẩu học khách hàng, ta thu được những thông tin rất quan trọng về khách hàng như tần suất mua hàng, số tiền đã chi tiêu, lần mua hàng gần nhất. Thông tin này được thể hiện ở bảng dưới đây:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Recency(Days) | Frequency | Monetary |
| count | 2500 | 2500 | 2500 |
| mean | 26.576 | 110.2168 | 3222.9852 |
| std | 62.79167255 | 115.27154 | 3349.0261 |
| min | 1 | 1 | 8.17 |
| 25% | 2 | 38 | 970.74 |
| 50% | 7 | 78 | 2157.75 |
| 75% | 21 | 142 | 4413.32 |
| max | 658 | 1298 | 38319.79 |



Như vậy, trong 276484 giao dịch đã được thực hiện, giá trị trung binh một giỏ hàng khách hàng mua trong một lần là 29.14$. Trung bình một hộ gia đình thực hiện giao dịch 90.2 lần trong 2 năm, 560.6 sản phẩm.

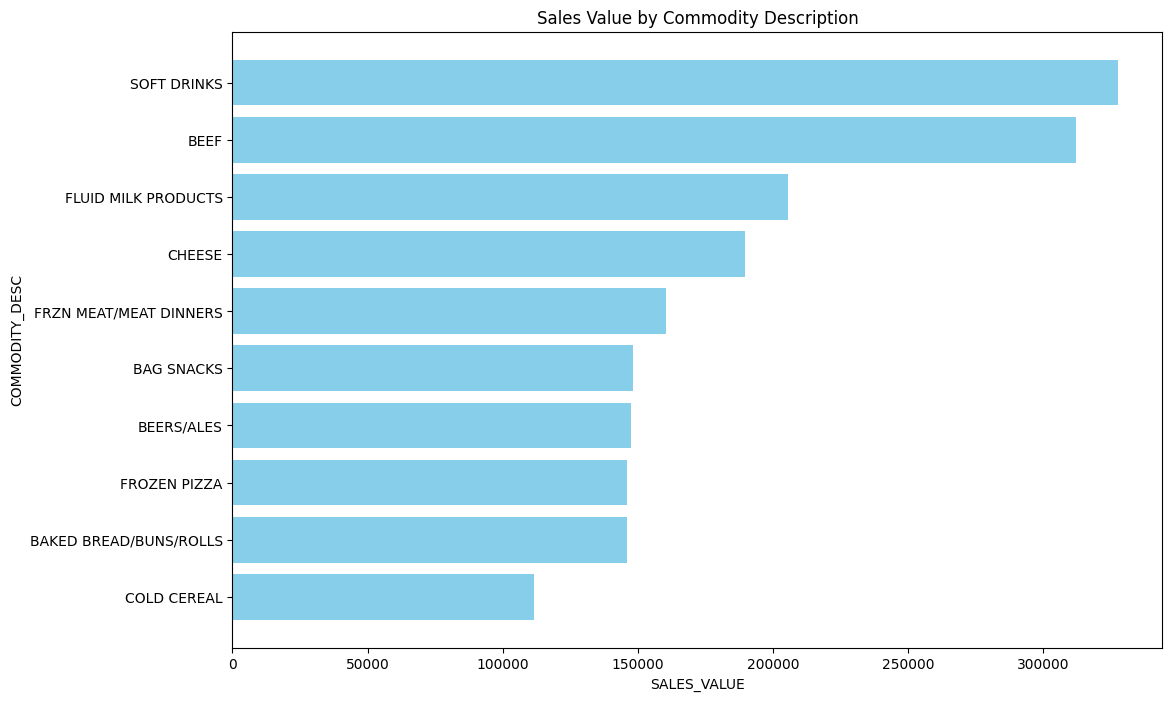


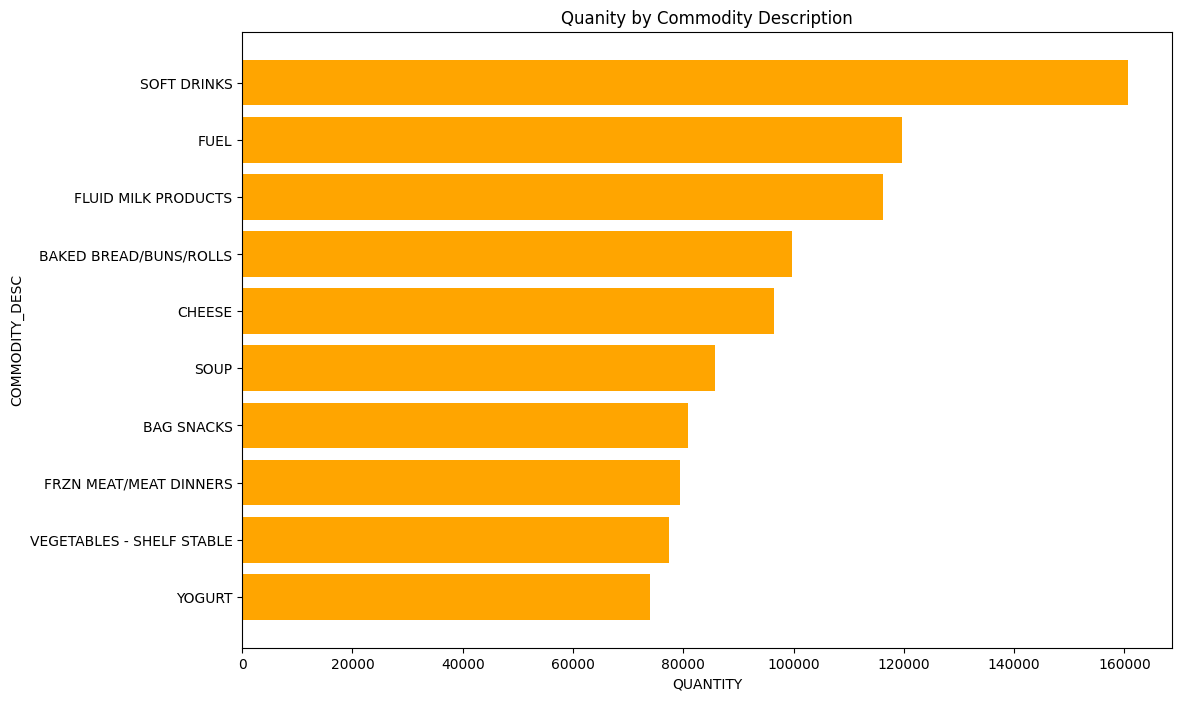
DỰa vào scatter plot thể hiện quan hệ giữa tần suất mua hàng, số tiền đã mua, cũng như ngày mua gần nhất. Ta có thể nhận ra phần lớn khách hàng đều có lần mua hàng gần nhất trong khoảng 150 ngày. Nhìn chung tần suất mua hàng lớn dẫn đến tổng số tiền chi tiêu nhiều hơn. Những khách hàng có ngày mua gần nhất quá xa như 300 ngày, 400 ngày có thể được xác định là đã chuyển đi và không còn tham gia mua hàng nữa.

Yếu tố nhà bán lẻ quan tâm đến nhất có lẽ là doanh số. Dựa vào thông tin giao dịch, ta có thể thu được thống kê về doanh thu/số lượng hàng bán ra hàng ngày:

|  |  |  |
| --- | --- | --- |
|  | **SALES\_VALUE** | **QUANTITY** |
| count | 156376 | 156376 |
| mean | 47.43 | 22.17 |
| std | 69.63 | 141.02 |
| min | 0 | 0 |
| 25% | 9.98 | 3 |
| 50% | 25.85 | 9 |
| 75% | 56.72 | 25 |
| max | 1300.29 | 30093 |

Theo đó, một biểu đồ thanh để thể hiện top các sản phẩm bán chạy nhất trong ngày ( biểu đồ màu cam)/ mang lại doanh thu cao nhất (biểu đồ màu xanh) sẽ rất có giá trị nhận biết các mặt hàng mang lại lợi nhuận ngắn hạn tốt và thường xuyên:



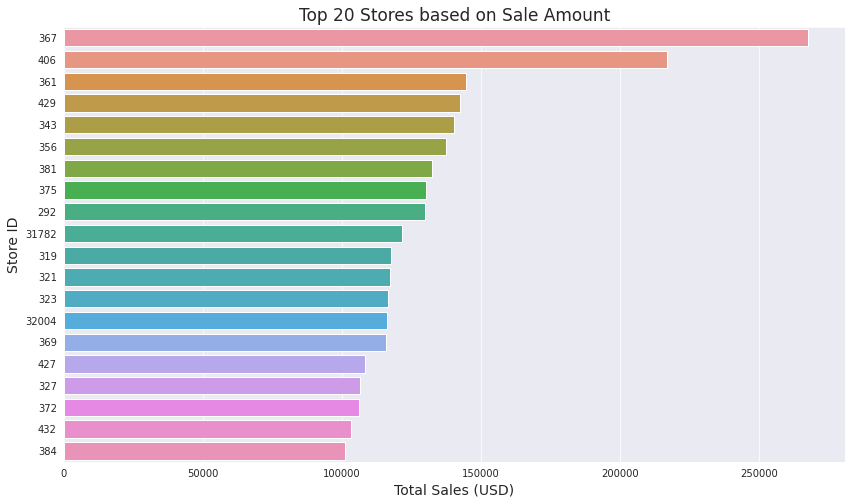


Ta có thể thấy mặt hàng nước giải khát (softdrink) là mặt hàng bán chạy nhất hàng ngày, cũng là mặt hàng mang lại doanh thu cao nhất hàng ngày.

Mặt hàng nhiên liệu dù bán chạy thứ hai, nhưng lại không phải mặt hàng mang lại doanh thu cao nhất, rõ ràng ta có thể thấy mặt hàng này còn chịu nhiều ảnh hưởng bới các chính sách phân phối của chính phủ.

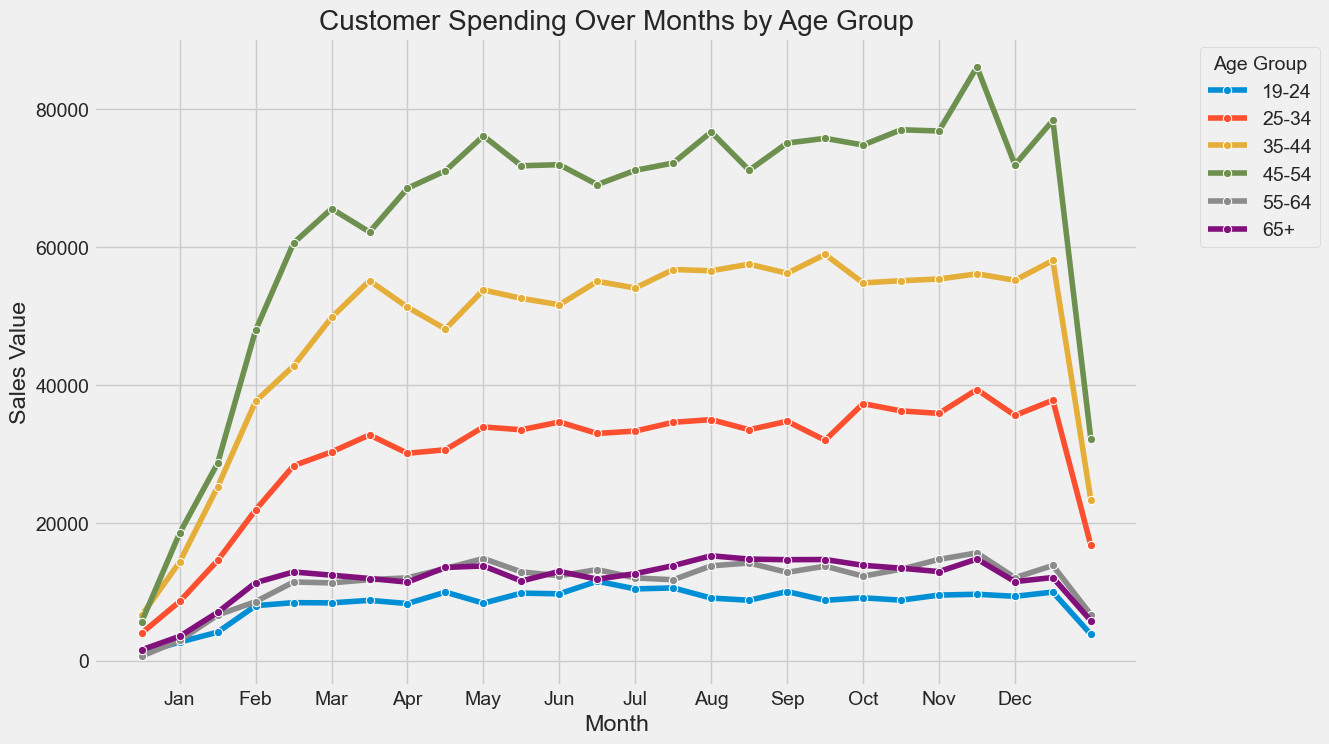
Ngoài nhiên liệu, phần lớn các sản phẩm bán chạy hàng ngày khác đều nằm trong danh mục thực phẩm. Những sản phẩm tuy không nằm trong top bán chạy như thịt bò có vẻ thuộc dạng sản phẩm cao cấp và mang lại biên lợi nhuận cao dẫu số lượng bán ra không phải quá lớn so với các mặt hàng thực phẩm khác.

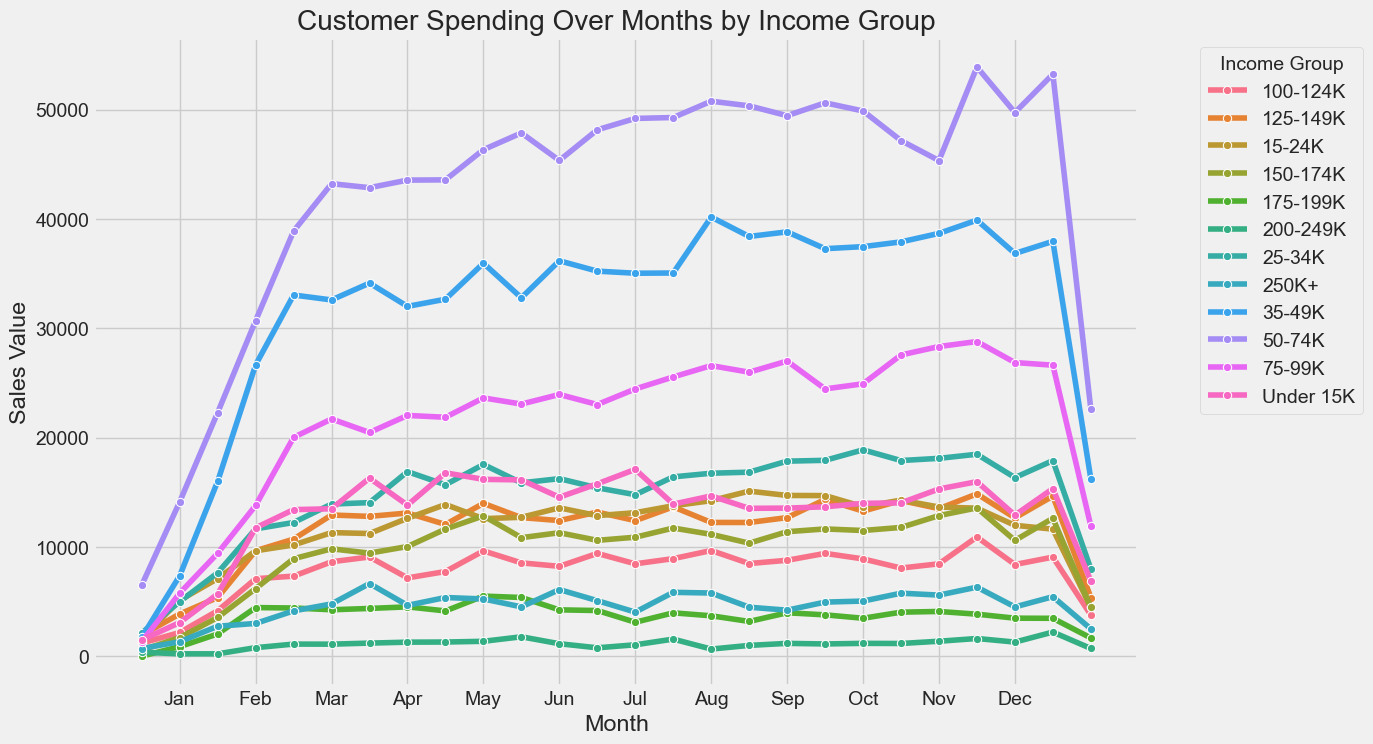
Ta cũng lập biểu đồ thanh để thể hiện top 20 cửa hàng dựa trên doanh thu và số lượng hàng bán ra.



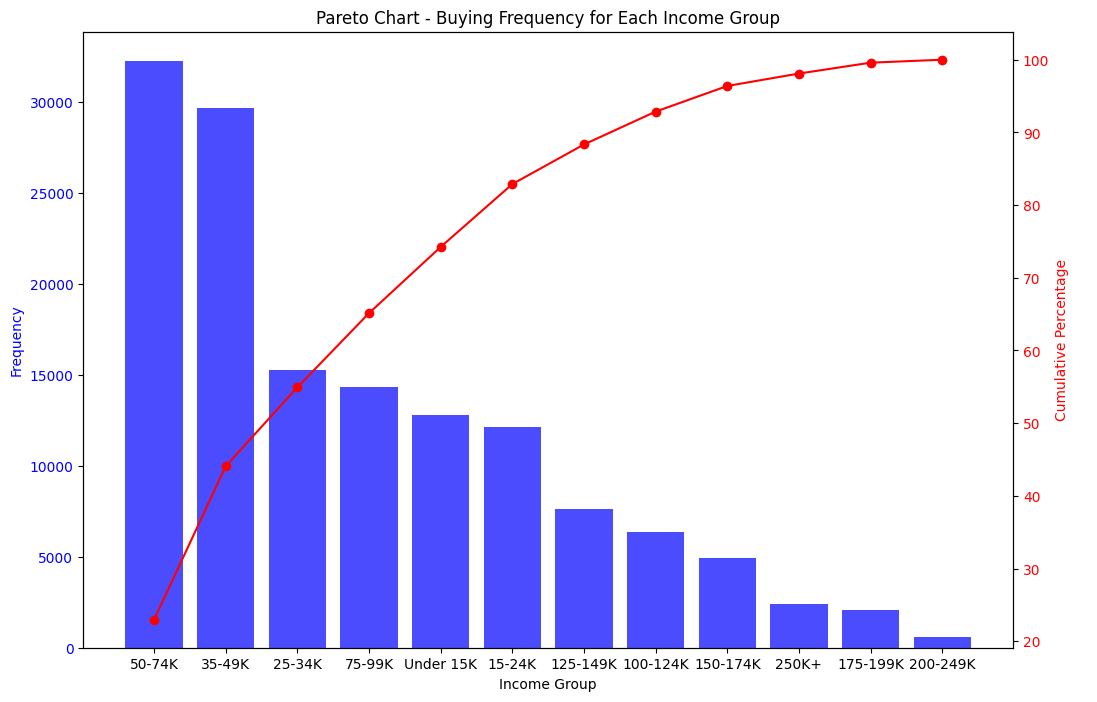
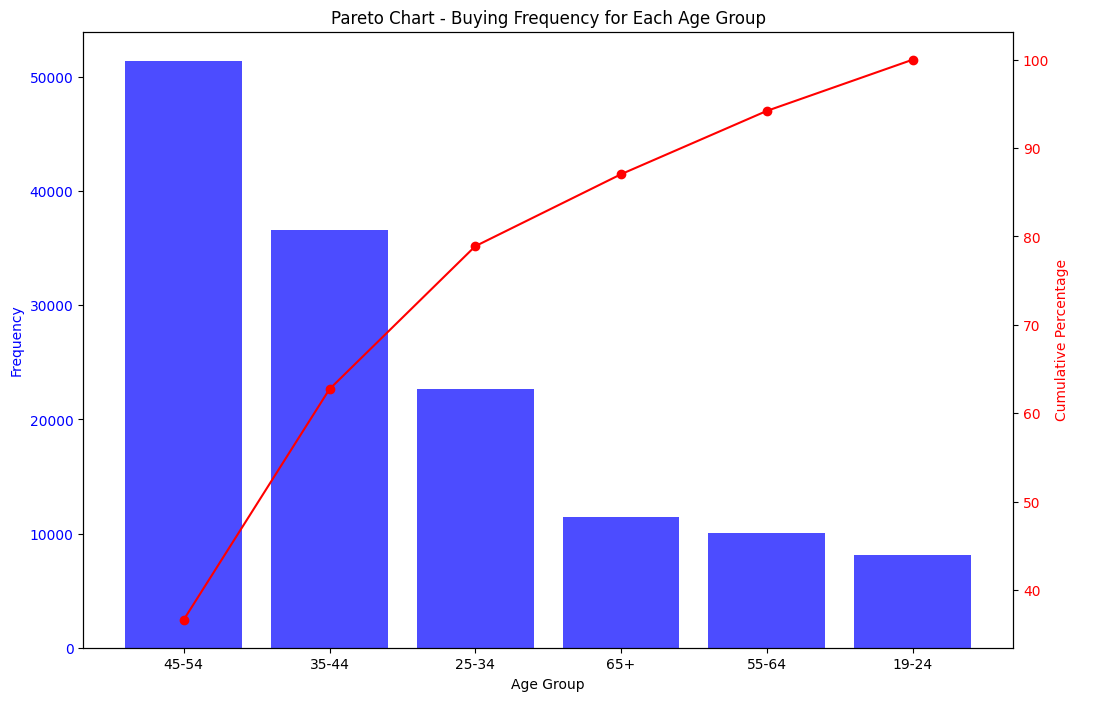
Theo đó, 2 cửa hàng mã 367 và 406 là 2 cửa hàng bán lẻ mang lại doanh thu vượt trội so với những cửa hàng ngược lại. Ta có thể lấy 2 cửa hàng này làm mẫu và điều tra thêm để tìm hiểu lý do mang lại sự khác biệt này từ đó cải thiện doanh số của các cửa hàng khác.

Một giả thuyết được đặt ra là độ tuổi và mức thu nhập có ảnh hưởng đến thói quen mua hàng của khách hàng. Ta lập biểu đồ đường để có thể nhìn thấy mức chi tiêu của từng nhóm tuổi/ mức thu nhập khác nhau để kiểm định điều này.



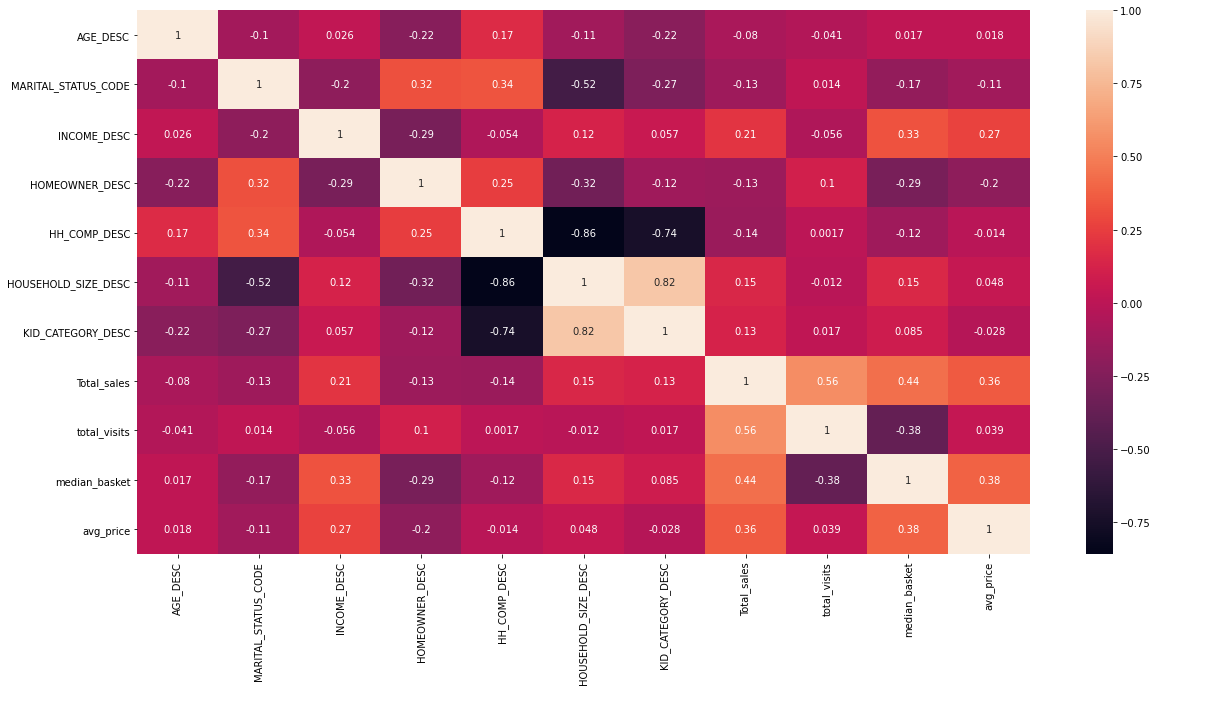


Ta có thể nhìn thấy cả 2 yếu tố này đều có ảnh hưởng đến tổng mức độ chi tiêu, thể hiện ở sự chênh lệch giữ các đường trong cả 2 biểu đồ độ tuổi (ở trên) và thu nhập (ở dưới). Tuy nhiên, ta cần tính tới một yếu tố đó là phân bố số lượng khách hàng nằm trong từng nhóm này. Vì một nhóm có thể không có thói quen chi tiêu nhiều nhưng chiếm số lượng lớn thì vẫn có thể trở thành nhóm khách hàng có tổng số tiền chi tiêu lớn.



Nhìn vào pareto chart thể hiện tần suất mua hàng của khách hàng phân biệt theo 2 cách, đó là dựa vào độ tuổi và mức thu nhập, ta cũng có thể thấy nhìn chung tần suất mua hàng của từng nhóm ảnh hưởng lớn đến tổng số tiền đã chi. Riêng với mức thu nhập 25-34K, dù đứng thứ 3 trong tần suất mua hàng, nhóm này lại không phải nhóm đứng thứ 3 trong tổng mức chi tiêu hàng tháng.

Tổng kết lại thông tin nhân khẩu học và giao dịch để tìm những điểm tương quan, ta có thể thu được một ma trận tương quan như sau:



Dựa vào ma trận tương quan này, rõ ràng yếu tố ảnh hưởng lớn nhất đến thói quen mua hàng của khách hàng là mức thu nhập. Yếu tố độ tuổi không có mức độ tương quan đáng kể. Vậy nên việc biểu đồ chi tiêu ở trên thể hiện sư chênh lệch trong chi tiêu giữa các nhóm tuổi là do sự chênh lệch về số lượng khách hàng ở mỗi nhóm tuổi có khác biệt lớn.

Đi kèm với đó, hộ gia đình có số lượng thành viên nhiều hơn cũng có xu hướng chi tiêu nhiều hơn. Ngoài ra, ta cũng thấy được nếu số lần đến cửa hàng nhiều hơn thì tỉ lệ chi tiêu cũng cao hơn, vậy nên nhà bán lẻ nên có những biện pháp kích thích khách hàng ghé thăm cửa hàng nhiều hơn.

1. Thông tin các campaign và coupon đi kèm:

Có 30 campaign khác nhau, ứng với 3 dạng campaign A,B,C. Các chiến dịch là việc gửi coupon có thời hạn tới khách hàng với mong muốn khuyến khích khách hàng mua hàng.

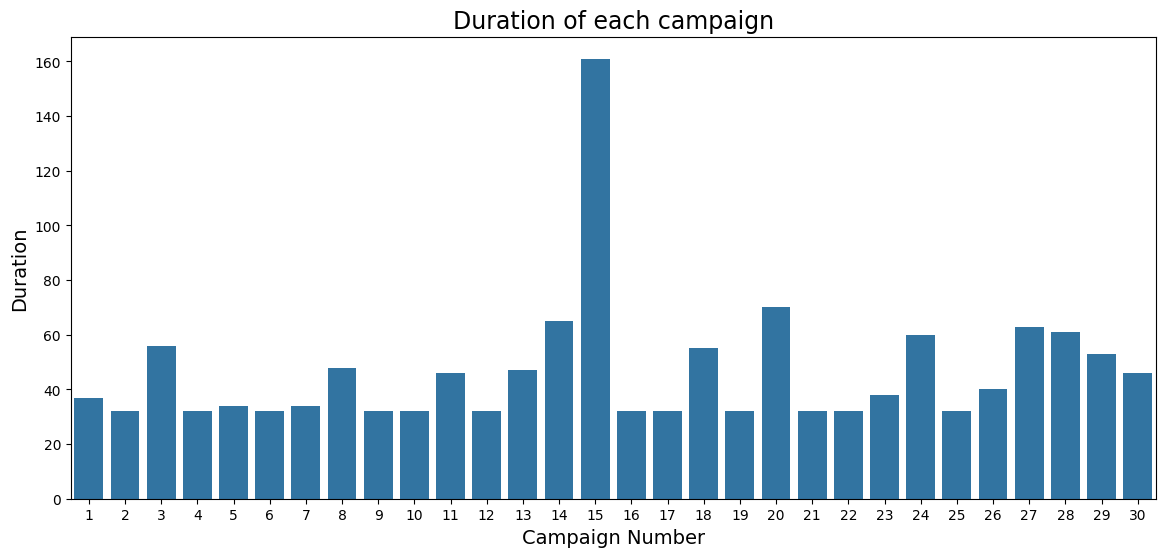
Trong số 2500 hộ gia đình được nghiên cứu,có 1584 hộ gia đình tham gia một campaign, thời gian trung bình diễn ra một campaign là 46.6 ngày

Trong tổng số các hộ gia đình nhận được coupon, chỉ có 434 hộ sử dụng coupon

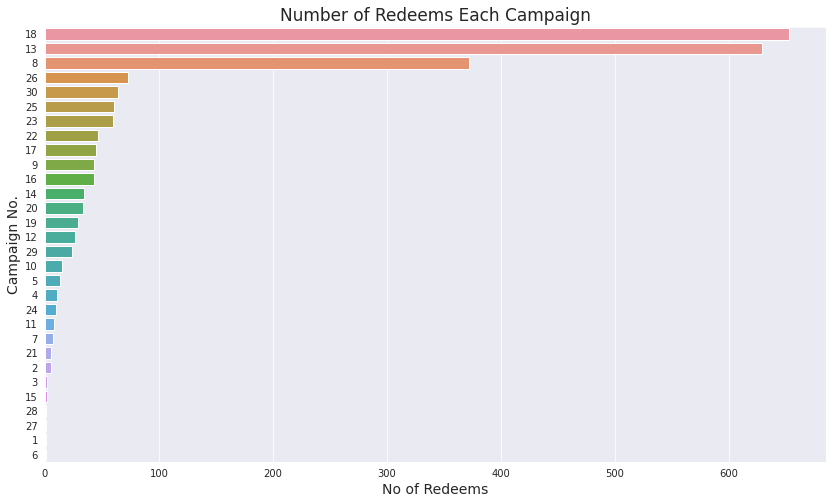
Tổng kết chung về 3 loại campaign được thể hiện ở bảng thống kê sau:

|  |  |  |  |
| --- | --- | --- | --- |
|  | DESCRIPTION | Number of Campaigns | avg\_duration |
| 0 | TypeA | 5 | 47.2 |
| 1 | TypeB | 19 | 37.631579 |
| 2 | TypeC | 6 | 74.5 |

Nghiên cứu về từng campaign để xem xét mức độ hiệu quả của chúng. Đầu tiên ta nhìn vào chiều dài của mỗi campaign, được thể hiện qua biểu đồ cột sau.

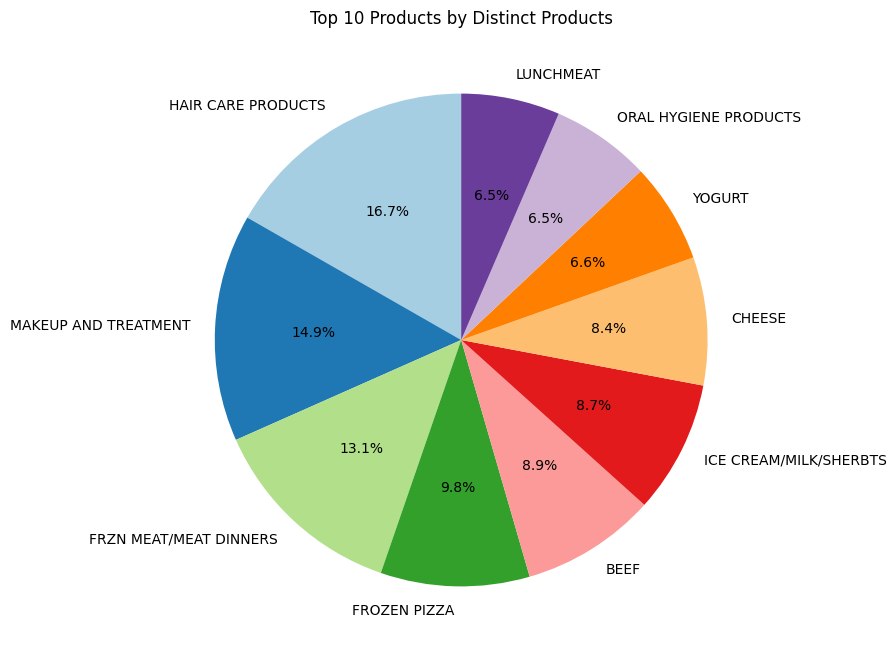


Các campaign có phần lớn có chiều dài gần nhau, campaign 15 có chiều dài nổi trội hơn cả. TIếp theo, ta thống kê lại số lượng coupon đã dược đổi tại mỗi campaign:



Trong số các coupon phát ra ở mỗi campaign, campaign 18, 13 và 8 có tỉ lệ được sử dụng cao nhất. Rõ ràng các campaign này chưa các coupon có khả năng kích cầu tốt, nà bán lẻ cần thực hiện điều tra sâu hơn để hiểu rõ và đưa ra các chiến dịch quảng bá trong tương lai.

Theo dõi tỉ lệ những sản phẩm được mua sử dụng coupon trong các campaign, ta thu được 10 sản phẩm thu hút nhất với khách hàng khi được đi kèm mã giảm giá là:



Theo biểu đồ trên, sản phẩm được mua nhiều khi có chương trình khuyến mãi vẫn là thực phẩm như thịt bò, pizza đông lạnh, sữa, phô mai, sữa chua. Kèm với đó là sản phẩm chăm sóc tóc và trang điểm. Trong tương lai, thông tin này có thể được sử dụng để phát triển các chiến dịch truyền thông tốt hơn.