TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC

PHÂN TÍCH NHẬN DẠNG MẪU

ĐỀ TÀI: PHÂN CỤM KHÁCH HÀNG

Giảng viên hướng dẫn: Vũ Ngọc Thanh Sang

Họ tên sinh viên: Lương Ngọc Minh Khuê

MSV: 3120410256

Thành phố Hồ Chí Minh, 12/2023

MỤC LỤC

DANH MỤC BẢNG BIỂU VÀ HÌNH ẢNH

I Giới thiệu:

1.1 Định nghĩa bài toán:

Bài toán được dùng để phân cụm khách hàng

1.2 Sự cần thiết của project:

Trong bối cảnh cạnh tranh ngày một khốc liệt của thị trường hiện nay thì việc nắm bắt được sở thích cũng như nhu cầu của khách hàng để từ đó, người bán hàng hay các doanh nghiệp có thể đề ra các chiến lược hiệu quả để tiếp cận từng bộ phận khách hàng là vô cùng quan trọng. Do mỗi khách hàng có thể có sở thích, độ tuổi, hoàn cảnh khác nhau nên nhu cầu mua hàng của họ cũng sẽ khác nhau. Do đó, ta phải tiến hành phân cụm khách hàng ra thành các cụm nhỏ có dựa trên đặc điểm chung để từ đó, với mỗi cụm khác nhau, ta có các chiến lược tiếp thị khác nhau, góp phần nâng cao doanh số bán ra. Đó cũng sự cần thiết cho project phân cụm khách hàng bằng nhận dạng mẫu ra đời.

II Cơ sở lý thuyết:

2.1 Kmean:

K-Means là một thuật toán phân cụm (clustering) trong lĩnh vực học máy và khai phá dữ liệu. Thuật toán này được sử dụng để phân nhóm các điểm dữ liệu vào các cụm sao cho các điểm trong cùng một cụm có sự tương đồng lớn, trong khi các cụm khác nhau có sự khác biệt cao.

Cách hoạt động của thuật toán K-Means như sau:

1. Chọn số lượng cụm (K): Đầu tiên, cần xác định số lượng cụm mà muốn thuật toán phân chia dữ liệu thành. Số lượng cụm này thường được ký hiệu là K.

2. Khởi tạo các điểm trung tâm cụm ban đầu: Chọn ngẫu nhiên K điểm từ dữ liệu làm điểm trung tâm của K cụm ban đầu.

3. Gán mỗi điểm dữ liệu vào cụm gần nhất: Dựa trên khoảng cách giữa mỗi điểm dữ liệu và các điểm trung tâm cụm, gán mỗi điểm vào cụm gần nhất.

4. Cập nhật điểm trung tâm cụm: Tính toán lại trung tâm cho mỗi cụm bằng cách lấy trung bình của tất cả các điểm thuộc cụm đó.

5. Lặp lại bước 3 và 4: Tiếp tục lặp lại quá trình gán và cập nhật cho đến khi không có sự thay đổi đáng kể nào trong việc gán các điểm vào cụm và cập nhật trung tâm cụm.

Thuật toán K-Means có thể giúp phân chia dữ liệu thành các cụm một cách tự động mà không yêu cầu sự giám sát, và nó thường được sử dụng trong nhiều ứng dụng như phân loại hình ảnh, phân loại văn bản, hoặc nhận dạng chủ đề trong dữ liệu.

2.2 PCA:

Principal Component Analysis (PCA) là một phương pháp thống kê được sử dụng để giảm chiều dữ liệu trong phân tích đa biến. Mục tiêu chính của PCA là tìm ra các thành phần chính (principal components) của dữ liệu, là các hướng trong không gian đặc trưng mà theo đó dữ liệu biến đổi nhiều nhất.

Quy trình PCA bắt đầu bằng việc tính ma trận hiệp phương sai của dữ liệu, đại diện cho mức độ biến động giữa các biến. Sau đó, PCA tìm các vectơ riêng và giá trị riêng của ma trận hiệp phương sai. Các vectơ riêng tương ứng với các giá trị riêng lớn nhất đại diện cho các hướng chính của biến động trong dữ liệu.

Các thành phần chính được sắp xếp theo thứ tự giảm dần của giá trị riêng, và chúng được sử dụng để tạo ra các biến mới (các thành phần chính) mà mỗi biến này là tổ hợp tuyến tính của các biến ban đầu. Khi chỉ sử dụng một số lượng ít các thành phần chính, ta có thể giảm số lượng chiều của dữ liệu mà vẫn giữ được phần lớn thông tin quan trọng.

2.3 Thang đo Silhouette:

Thang đo Silhouette (Silhouette Score) là một phương pháp đánh giá chất lượng của việc phân cụm trong phân tích phân cụm dữ liệu. Nó cung cấp một phản hồi số lượng về mức độ tách biệt giữa các cụm được tạo ra và đánh giá đồng nhất trong các cụm. Silhouette Score nằm trong khoảng từ -1 đến 1, và giá trị càng gần 1 thì phân cụm càng tốt.

2.4 Thang đo Elbow:

Thang đo "elbow" (còn được gọi là "phương thức elbow" hoặc "elbow method") là một kỹ thuật được sử dụng để chọn số lượng cụm tối ưu trong một thuật toán phân cụm, chẳng hạn như K-means clustering. Phương thức này giúp xác định điểm nơi một biểu đồ "đo sức căng" giảm dốc nhanh chóng, tương ứng với số lượng cụm tối ưu.

Quy trình thực hiện phương thức elbow như sau:

1. Huấn luyện thuật toán phân cụm với một loạt các giá trị k (số lượng cụm) khác nhau.
2. Đối với mỗi giá trị k, tính toán độ đo hiệu suất (ví dụ: bình phương khoảng cách từ điểm dữ liệu đến trung tâm cụm trong trường hợp K-means).
3. Vẽ biểu đồ các giá trị độ đo hiệu suất theo số lượng cụm k. Khi vẽ biểu đồ và quan sát nó, thì "elbow" thường xuất hiện như một điểm nơi đường cong có dấu hiệu giảm độ dốc nhanh chóng. Điểm này thường được coi là số lượng cụm tối ưu

Tuy nhiên, đôi khi biểu đồ không rõ ràng và việc chọn số lượng cụm tối ưu có thể phụ thuộc vào sự hiểu biết của người phân tích về dữ liệu cụ thể và mục tiêu phân cụm.

III Tổng quan về bộ dữ liệu:

3.1 Ý nghĩa các đặc trưng trong bộ dữ liệu:

ID	Mã ID của khách hàng
Year_Birth	Năm sinh của khách hàng
Education	Trình độ học vấn của khách hàng
Marital_Status	Tình trạng hôn nhân của khách hàng
Income	Thu nhập hàng năm của khách hàng
Kidhome	Số trẻ em trong gia đình
Teenhome	Số thanh thiếu niên trong gia đình
Dt_Customer	Thời gian khách hàng đăng kí vào công ty

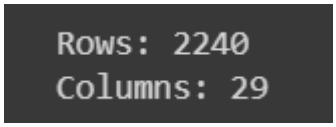
Recency	Số ngày kể từ lần mua cuối cùng của khách hàng
Complain	1 Nếu khách hàng phàn nàn trong 2 năm qua, 0 nếu ngược lại.
MntWines	Số tiền chi cho rượu vang trong 2 năm qua.
MntFruits	Số tiền chi mua trái cây trong 2 năm qua.
MntMeatProducts	Số tiền chi mua thịt trong 2 năm qua.
MntFishProducts	Số tiền chi mua cá trong 2 năm qua.
MntSweetProducts	Số tiền chi mua đồ ngọt trong 2 năm qua.
MntGoldProds	Số tiền chi cho vàng trong 2 năm qua.
NumDealsPurchases	Số lần mua hàng được giảm giá
AcceptedCmp1	1 nếu khách hàng chấp nhận chiến dịch ưu đãi đầu tiên, 0 nếu ngược lại.
AcceptedCmp2	1 nếu khách hàng chấp nhận chiến dịch ưu đãi thứ hai, 0 nếu ngược lại.
AcceptedCmp3	1 nếu khách hàng chấp nhận chiến dịch ưu đãi thứ ba, 0 nếu ngược lại.
AcceptedCmp4	1 nếu khách hàng chấp nhận chiến dịch ưu đãi thứ tư, 0 nếu ngược lại.
AcceptedCmp5	1 nếu khách hàng chấp nhận chiến dịch ưu đãi thứ năm, 0 nếu ngược lại.
Response	1 nếu khách hàng phản hồi lại các chiến dịch ưu đãi, 0 nếu ngược lại.
NumWebPurchases	Số lượt mua hàng thông qua trang web.

NumCatalogPurchases	Số lượt mua hàng thông qua danh mục.
NumStorePurchases	Số lượt mua hàng ở cửa hàng
NumWebVisitsMonth	Số lượt truy cập vào website của công ty tháng trước.

Mặc dù các bước thực hiện như nhau nhưng mua hàng qua danh mục và mua hàng qua web khác nhau ở chỗ mua hàng qua danh mục thì khách hàng có thể chọn sản phẩm trong một danh sách trực tuyến hoặc một danh sách được in giấy do công ty phát hành trong khi mua hàng trên web thì khách hàng chỉ có thể lựa chọn sản phẩm trên trang web. Đó cũng chính là điểm khác nhau cơ bản của 2 đặc trưng NumCatalogPurchases và NumWebPurchases.

3.2 Thống kê dữ liệu:

Bộ dữ liệu gồm có 2240 dòng, 29 cột.



```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
```

Có 25 đặc trưng có kiểu dữ liệu là int, 3 đặc trưng có kiểu dữ liệu là object, 1 đặc trưng có kiểu dữ liệu là float.

```
df.select_dtypes('object').nunique()
```

```
Education          5
Marital_Status     8
Dt_Customer      663
dtype: int64
```

Ba đặc trưng có kiểu dữ liệu là object gồm có: Education, Marital_Status, Dt_Customer.

Mô tả dữ liệu:

	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000
mean	5592.159821	1968.805804	52247.251354	0.444196	0.506250	49.109375	303.935714	26.302232	166.950000	37.525446	27.062946	44.021875
std	3246.662198	11.984069	25173.076661	0.538398	0.544538	28.962453	336.597393	39.773434	225.715373	54.628979	41.280498	52.167439
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000	24.000000	23.750000	1.000000	16.000000	3.000000	1.000000	9.000000
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	49.000000	173.500000	8.000000	67.000000	12.000000	8.000000	24.000000
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000	74.000000	504.250000	33.000000	232.000000	50.000000	33.000000	56.000000
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000	263.000000	362.000000

IV Phân tích dữ liệu:

Để phân tích dữ liệu, trước tiên, ta phải xử lý một số đặc trưng:

- Đầu tiên, ta cần tính được độ tuổi của khách hàng bằng cách lấy một mốc thời gian nhất định trừ cho năm sinh của khách.

  Ta lấy giá trị năm lớn nhất của đặc trưng Dt_Customer (Thời gian khách hàng đăng kí vào công ty) làm mốc thời gian để tính độ tuổi của khách

  Để làm được điều đó thì ta phải chuyển kiểu dữ liệu của đặc trưng Dt_Customer từ object sang datetime sau đó tìm giá trị max year.

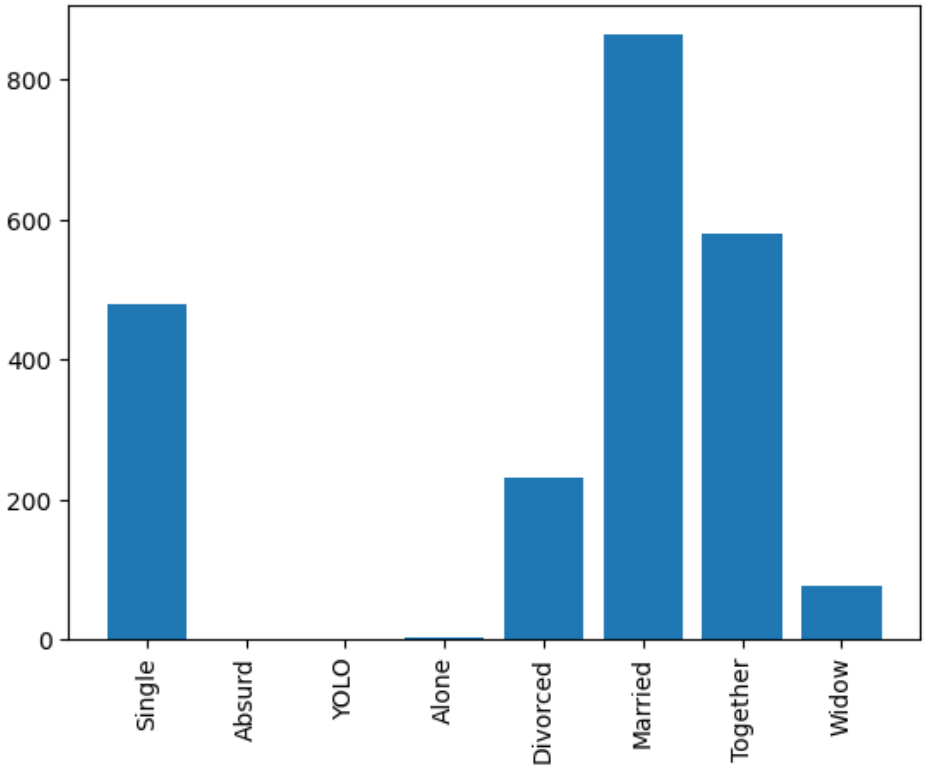
```
last_year = pd.to_datetime(df.Dt_Customer, format='%d-%m-%Y').dt.year.max()
print("Thời điểm gần nhất: ",last_year)
```

```
Thời điểm gần nhất:  2014
```

  Ta tìm được thời điểm gần nhất là 2014. Vì vậy, ta sẽ tính độ tuổi của khách hàng đến thời điểm này.

- Chuyển tên các đặc trưng về chữ thường để tiện hơn khi gọi tên.
- Đối với đặc trưng Marital_Status thì ta có các giá trị sau đây:

```
Married     0.385714
Together    0.258929
Single      0.214286
Divorced    0.103571
Widow       0.034375
Alone       0.001339
Absurd      0.000893
YOLO        0.000893
Name: Marital_Status, dtype: float64
```



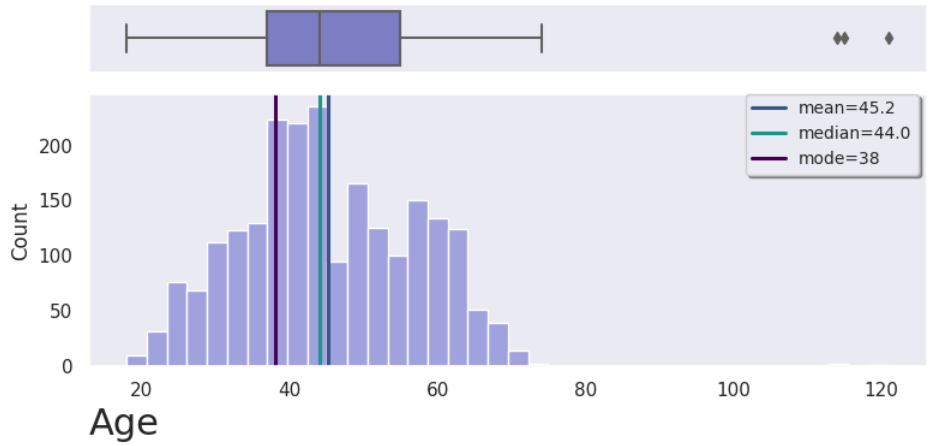
Hình 1.1: Biểu đồ thể hiện các giá trị của đặc trưng marital_status.

Do ‘Absurd’ mang nghĩa là ‘vô nghĩa’ nên ta có thể hiểu rằng trong quá trình thu thập dữ liệu có thể đã xảy ra sai sót nào đó. ‘YOLO’ ám chỉ một lối sống, phong cách sống nào đó, do đó không thể xếp vào tình trạng hôn nhân. Do các dòng này chiếm khá ít trong bộ dữ liệu, vì vậy, ta sẽ loại bỏ những dòng có marital_status = ‘Absurd’ hay marital_status = ‘YOLO’.

- Tạo đặc trưng partner dựa vào đặc trưng marital_status. Do có các trường hợp như Divorced và Window nên việc quy tình trạng hôn nhân về ‘kết hôn’ và ‘chưa kết hôn’ nó không khả quan bằng việc quy về ‘có bạn đời’ và ‘không có bạn đời’. Vì vậy, nếu marital_status = ‘Married’, ‘Together’ thì partner = ‘Partner’, các trường hợp còn lại là ‘No Partner’.
- Tạo đặc trưng mới là younghome bằng cách kết hợp hai đặc trưng là teenhome và kidhome.
- Tạo đặc trưng mới là mntotal biểu thị cho tổng số tiền mua sản phẩm bằng cách lấy các đặc trưng mntwines, mntfruits, mntmeatproducts, mntfishproducts, mntsweetproducts, mntgoldprods cộng lại.
- Tạo đặc trưng age_category để phân loại nhóm tuổi của khách hàng. Cụ thể, khách hàng dưới 30 tuổi sẽ xếp vào nhóm người trẻ tuổi, từ 30 đến 60 sẽ xếp vào nhóm trung niên, từ 60 trở lên sẽ vào nhóm người lớn tuổi.
- Tạo đặc trưng accepted để kiểm tra xem khách hàng có chấp nhận 1 trong các chiến dịch tiếp thị hay không bằng cách tạo 1 cột mới là mntaccepted bằng tổng của 5 chiến dịch tiếp thị. Nếu mntaccepted lớn hơn 0 tức là có ít nhất 1 chiến dịch tiếp thị được chấp nhận thì accepted sẽ mang giá trị là 1, ngược lại là 0.

```
df.columns = df.columns.str.lower()
couple = ['Married','Together']
df['partner'] = df['marital_status'].apply(lambda x: 'Partner' if x in couple else 'No Partner')
df['mnttotal'] = df.loc[:,'mntwines':'mntgoldprods'].sum(1)
df['mntaccepted'] = df.loc[:,'acceptedcmp1':'acceptedcmp5'].sum(1)
df['accepted'] = df['mntaccepted'].apply(lambda x: 1 if x > 0 else 0)
df['dt_customer'] = pd.to_datetime(df.dt_customer, format='%d-%m-%Y')
df['age'] = max(pd.to_datetime(df['dt_customer']).dt.year) - df['year_birth']
df['age_category'] = age_clusters
df['younghome']=df.kidhome+df.teenhome
```

4.1 Phân tích mối liên hệ giữa các đặc trưng:

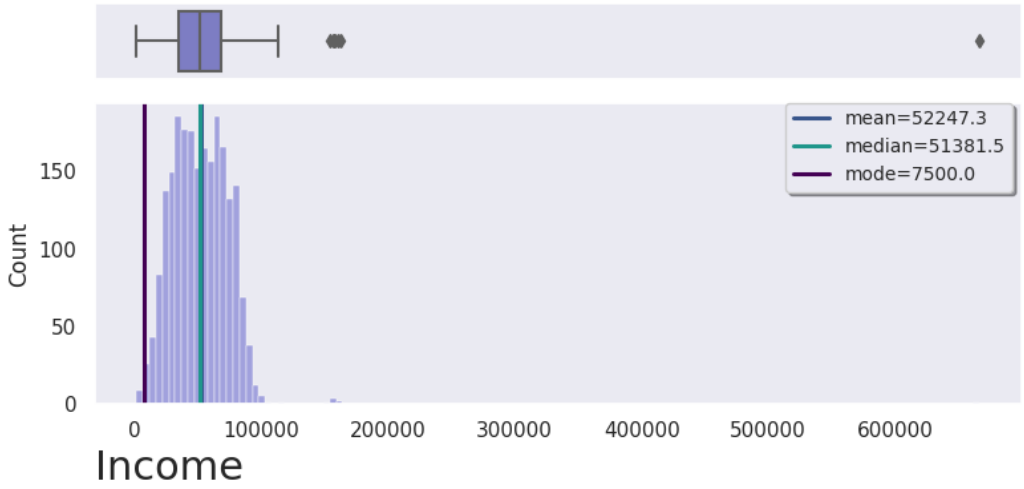


Hình 1.2 Biểu đồ thể hiện độ tuổi của khách hàng.

Phần lớn các đối tượng trong bộ dữ liệu nằm trong khoảng độ tuổi từ 30 đến 60. Điều này cho thấy đa phần các đối tượng trong bộ dữ liệu thuộc nhóm người trưởng thành.

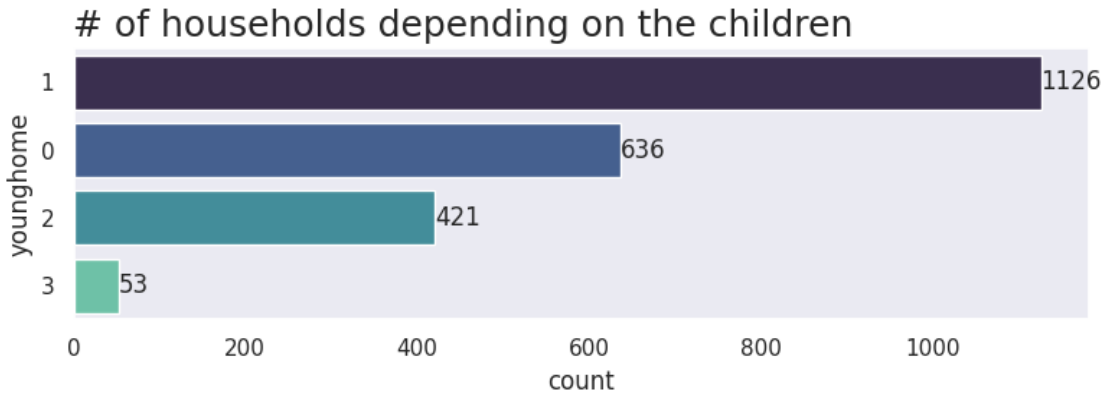
Dữ liệu phân bố tương đối bình thường nhưng biểu đồ bị lệch sang trái chủ yếu do ảnh hưởng của các giá trị ngoại lai. Ta có thể thấy rõ điều này ở biểu đồ boxplot phía trên, nơi tồn tại các giá trị ngoại lai cực đại.

Tiếp theo, ta sẽ xem xét sự phân bố của thu nhập của các đối tượng trong bộ dữ liệu:



Hình 1.3 Biểu đồ thể hiện sự phân bố của thu nhập.

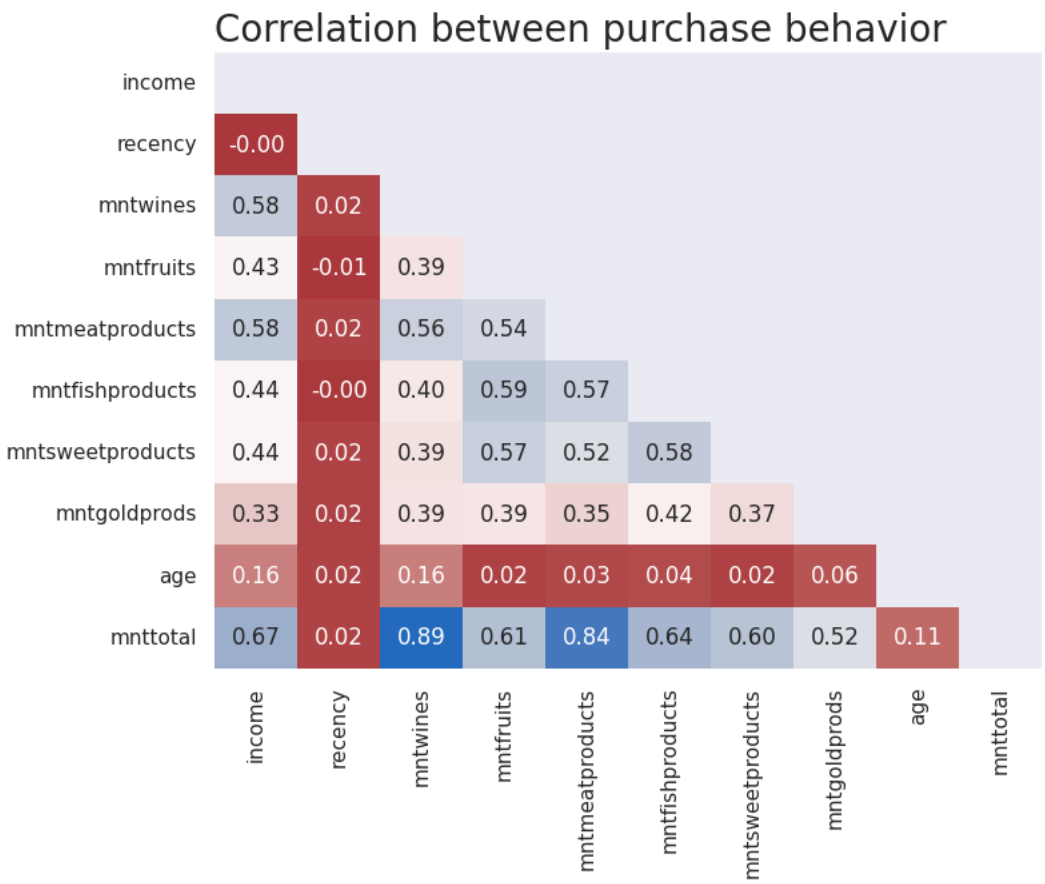
Từ biểu đồ, ta có thể thấy dữ liệu phân bố khá bình thường tuy nhiên biểu đồ cũng bị lệch trái do ảnh hưởng của các giá trị ngoại lai cụ thể là các giá trị ngoại lai cực đại mà ta có thể quan sát được thông qua boxplot.



Hình 1.4 Biểu đồ thể hiện sự phân bố số lượng người chưa trưởng thành trong mỗi hộ gia đình.

Phần lớn bộ dữ liệu có 1 hoặc không có những người chưa trưởng thành (thanh thiếu niên hoặc trẻ em) ở nhà. Từ đó, có thể nói số khách hàng có 1 con chiếm nhiều nhất, đồng thời, kết hợp với phân tích ở trên rằng phần lớn độ tuổi của khách hàng nằm trong khoảng 30 đến 60. Ta có thể suy đoán rằng với các khách hàng nằm trong độ tuổi từ 50 đến 60 có thể con cái đã trưởng thành và trở nên tự lập.

Với trường hợp này, ta cần mở rộng thêm bộ dữ liệu hướng đến đối tượng người trẻ tuổi để có thể đưa ra chiến lược hợp lý dành cho người già và cả người trẻ tuổi.



Hình 1.5 Biểu đồ thể hiện mối tương quan giữa các đặc trưng (1).

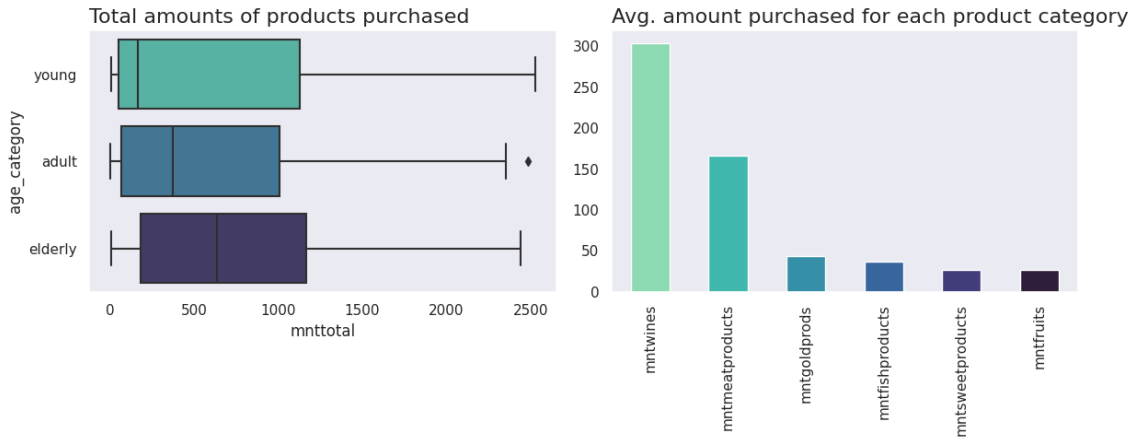
Qua biểu đồ, có thể thấy được thu nhập (income) và số tiền chi ra để mua thịt (mntmeatproducts) và rượu (mntwines) tương quan chặt chẽ với nhau với chỉ số tương quan là 0.58. Do 0.58>0 nên đây là tương quan thuận, có nghĩa là những cá nhân có thu nhập càng cao thì số tiền chi ra mua rượu và thịt càng nhiều và ngược lại. → Người có thu nhập cao chi nhiều tiền cho hai mặt hàng rượu và thịt.

Bên cạnh đó, số tiền chi ra để mua thịt (mntmeatproducts) và rượu (mntwines) cũng có chỉ số tương quan khá cao với tổng số tiền chi ra để mua sản phẩm(mnttotal) với chỉ số tương quan là 0.84 và 0.89. Điều này có nghĩa là những khách hàng mua thịt và rượu càng nhiều thì tổng số tiền chi tiêu để mua sản phẩm của họ càng lớn. Đây là một điều hợp lý do rượu và thịt là các loại sản phẩm có giá thành không rẻ đặc biệt là rượu.

Tổng số tiền chi ra để mua sản phẩm (mnttotal) và thu nhập (income) cũng có chỉ số tương quan lớn là 0.67 cho thấy các khách hàng thu nhập cao thì số tiền chi ra để mua sản phẩm cũng cao và ngược lại.

Ngoài ra, thì số tiền chi ra để mua trái cây (mntfruits) cũng tương quan chặt chẽ, tương quan thuận với số tiền chi ra để mua cá (mntfishproducts) và đồ ngọt(mntsweetproducts) với chỉ số tương quan là 0.59 và 0.57. Điều này cho thấy các khách hàng chi nhiều cho việc mua trái cây cũng có xu hướng mua cá và đồ ngọt.

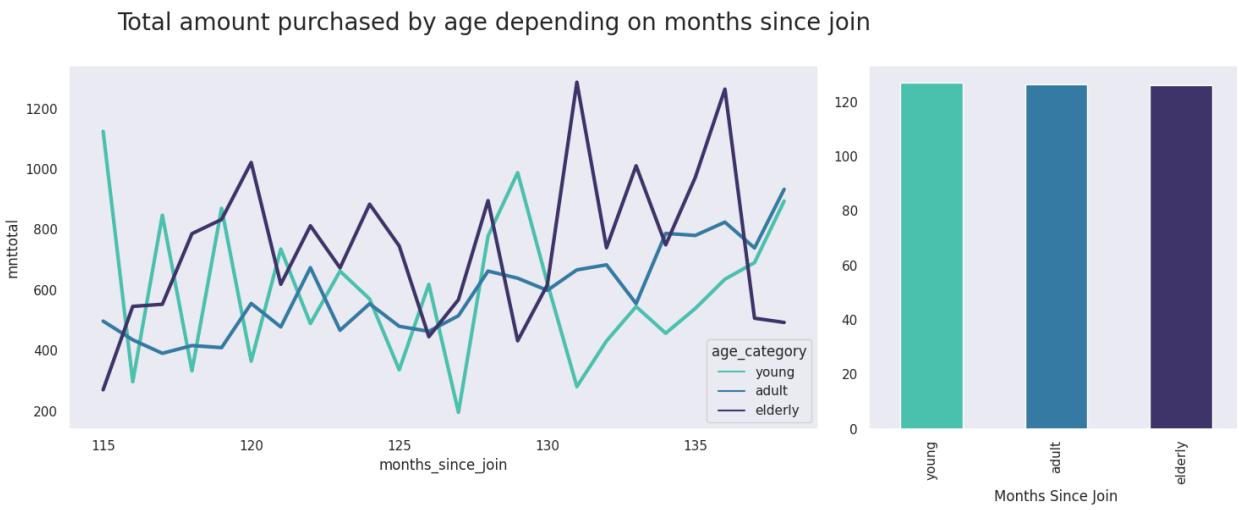
Từ đó, ta có thể đề ra chiến lược phát triển ưu đãi đi kém khi mua thịt và rượu, trái cây và đồ ngọt, trái cây và cá để khuyến khích chi tiêu.



Hình 1.6 Biểu đồ thể hiện các nhóm tuổi và số lượng mua của các mặt hàng.

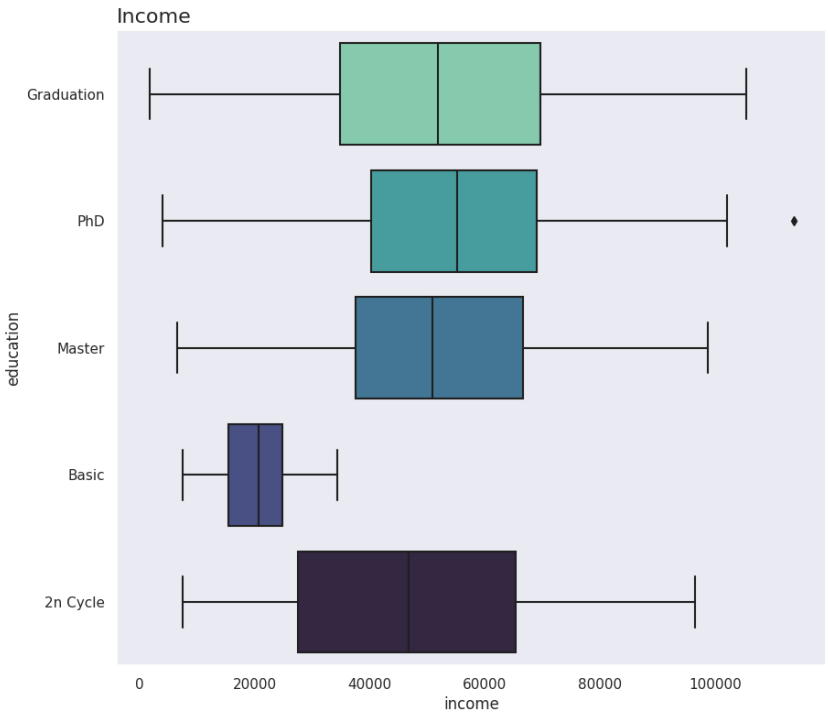
Qua biểu đồ "Total amounts of products purchased", ta thấy được người trẻ tuổi và người già có xu hướng mua nhiều sản phẩm hơn nhóm người lớn từ 30 đến dưới 60 tuổi.

Ở biểu đồ "Avg. amount purchased for each product category" ta thấy được số sản phẩm rượu và thịt được mua nhiều nhất trong số các sản phẩm còn lại.



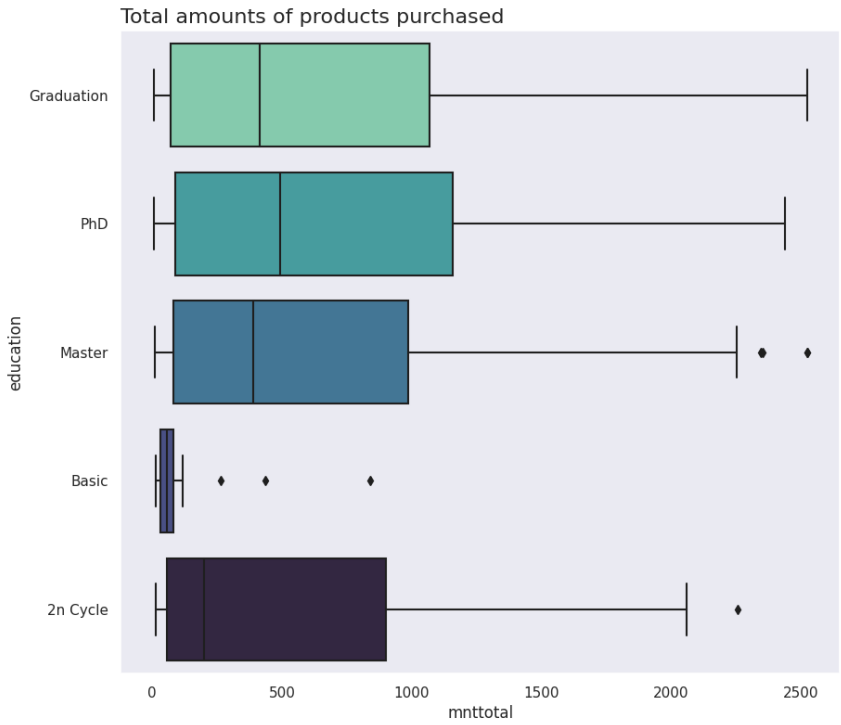
Hình 1.7: Biểu đồ thể hiện tổng số tiền mua sản phẩm của các nhóm tuổi dựa trên thời gian đăng kí vào công ty.

Thời gian (được tính theo số tháng) kể từ khi đăng ký với công ty (khách hàng thân thiết) dường như không phải là yếu tố chính ảnh hưởng đến lượng sản phẩm được mua. Rõ ràng là bất kể đã tham gia vào công ty được bao nhiêu tháng, người lớn vẫn luôn mua ít sản phẩm hơn người cao tuổi. Hơn nữa, có một xu hướng tăng lên có thể quan sát được, cho thấy rằng trung bình những cá nhân tham gia sau đã mua nhiều sản phẩm hơn.



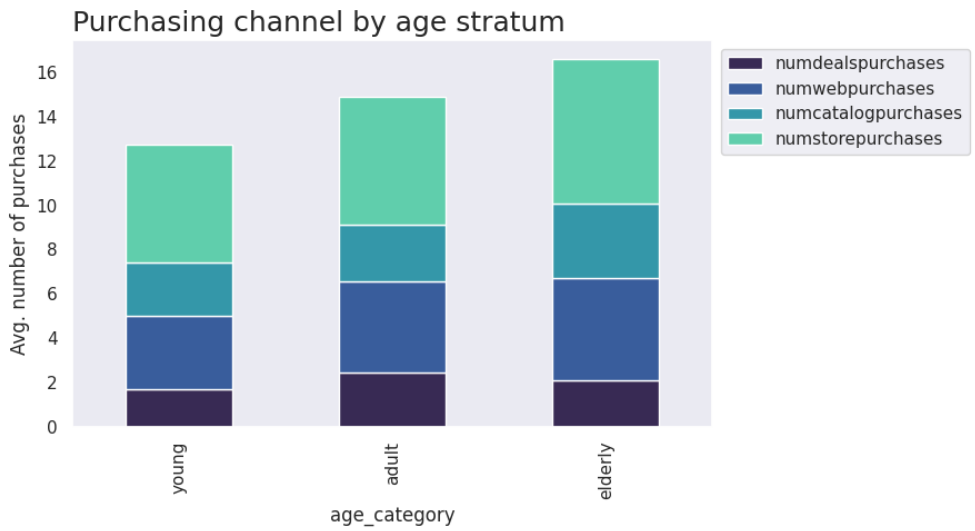
Hình 1.8: Biểu đồ thể hiện thu nhập dựa trên trình độ học vấn.

Qua biểu đồ, ta thấy được thu nhập của những khách hàng ở trình độ Basic thấp nhất trong số năm cấp bậc.



Hình 1.9: Biểu đồ thể hiện tổng số tiền chi ra để mua sản phẩm của từng trình độ học vấn.

Do thu nhập thấp nhất nên tổng số tiền mà nhóm khách hàng này chi ra đề mua sản phẩm cũng xếp thấp nhất so với các khách hàng ở các trình độ khác. Trong nhóm này có thể có một số khách hàng chưa trưởng thành vẫn còn đang đi học nên số tiền bỏ ra để chi tiêu cho các mặt hàng thấp.

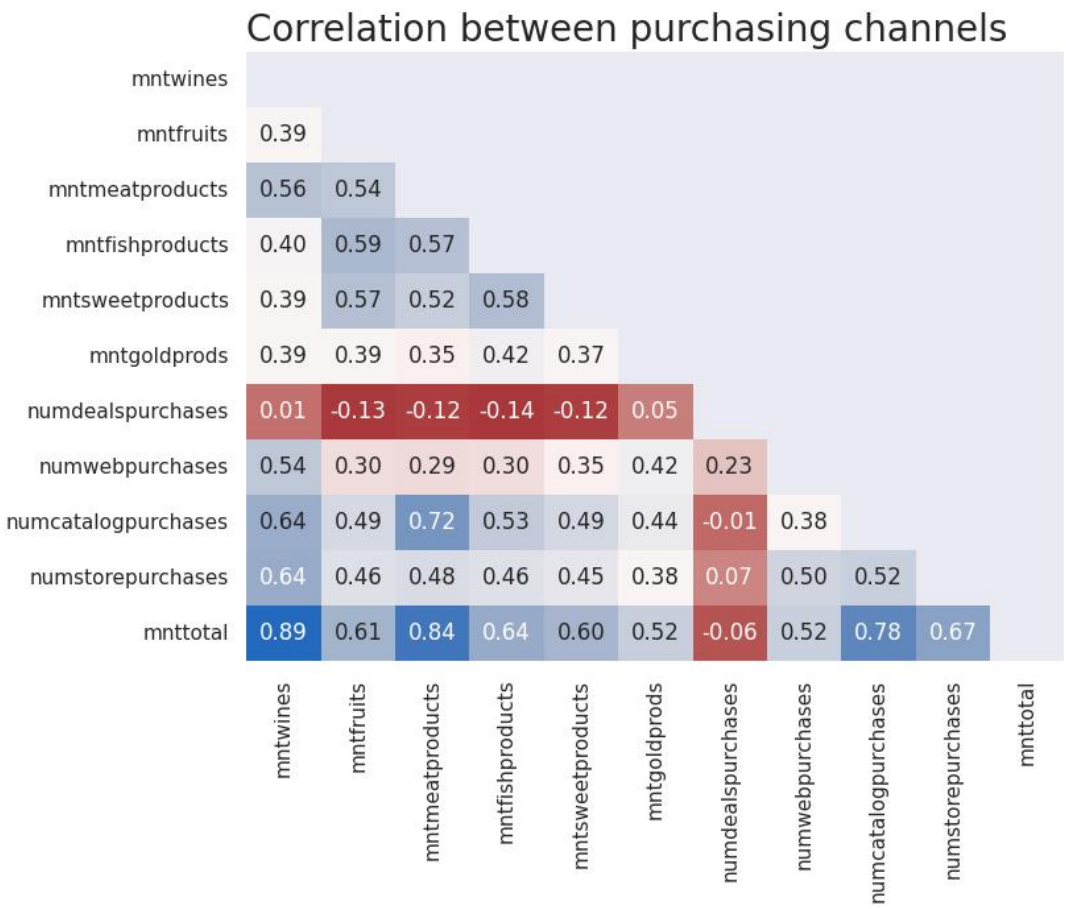


Hình 1.10: Biểu đồ thể hiện số lượt mua ở các kênh mua hàng và số lượt mua hàng giảm giá ở các nhóm tuổi.

Bất kể nhóm tuổi nào, việc mua hàng tại cửa hàng thực tế rõ ràng là kênh mua hàng được sử dụng phổ biến nhất.

Trái ngược với giả định trước đây, nhóm người lớn tuổi có mức trung bình cao nhất khi mua hàng trên web, ngoài ra, họ cũng có mức trung bình cao nhất khi mua hàng theo danh mục.

Các giao dịch mua hàng giảm giá có mức trung bình thấp nhất ở tất cả các độ tuổi. Trong đó, những người thuộc nhóm ‘adult’ có xu hướng mua hàng giảm giá nhiều hơn hai nhóm tuổi còn lại.



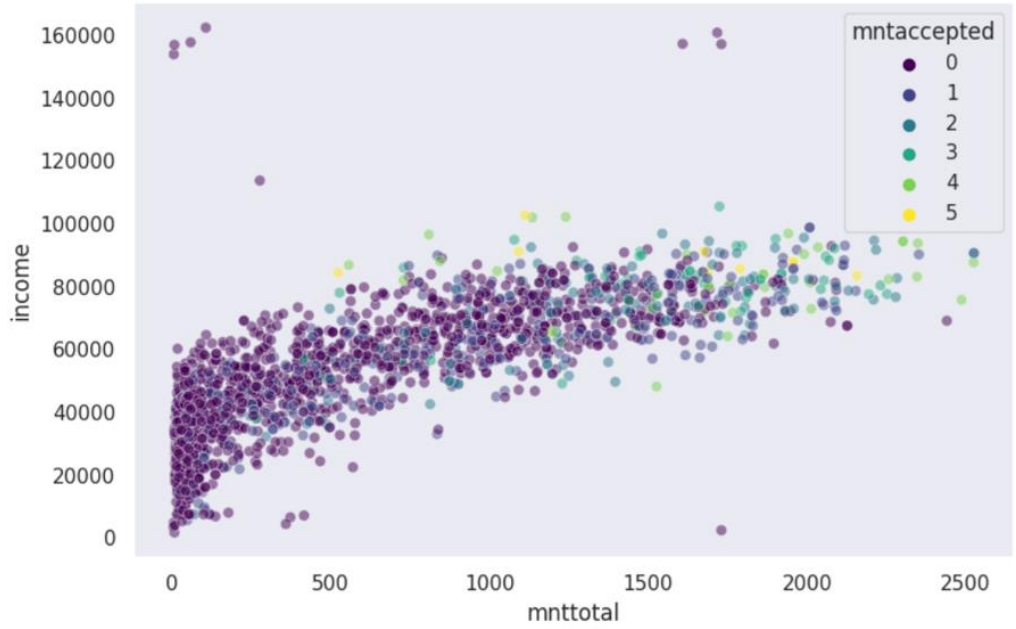
Hình 2.1: Biểu đồ thể hiện sự tương quan của các đặc trưng (2).

Qua biểu đồ, ta thấy được, số lượt mua hàng qua danh mục – numcatalogpurchases có chỉ số tương quan với rượu và thịt cao nhất so với các hình thức mua hàng còn lại.

Đồng thời, số lượt mua hàng thông qua danh mục cũng có chỉ số tương quan cao nhất đối với tổng số tiền chi ra để mua sản phẩm(mnttotal), chỉ số tương quan là 0.78.

Từ phân tích trên, có thể thấy khách hàng mua hàng mua thịt và rượu có xu hướng mua hàng theo danh mục. Vì vậy có thể đề ra các chiến lược tiếp thị, ưu đãi cho hạng mục thịt và rượu khi mua theo danh mục đồng thời đảm bảo ít nhiều sự hiện diện của hai loại sản phẩm này trên danh mục gửi đến khách hàng.

Ta cũng nên cân nhắc tuyển dụng nhân viên có hiểu biết về các loại thịt và rượu để có thể tiếp thị, cho khách hàng lời khuyên để kết hợp.

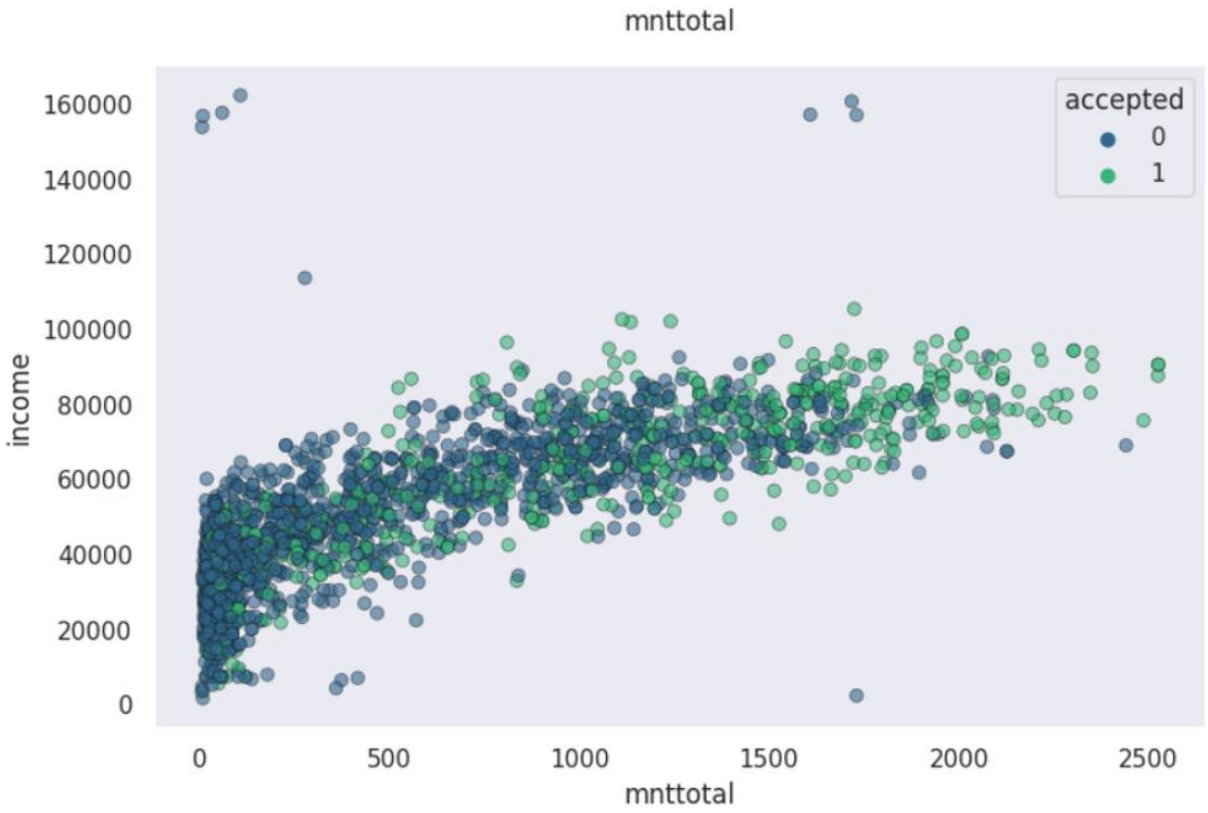


Hình 2.2: Biểu đồ thể hiện mối quan hệ giữa thu nhập và tổng số tiền chi tiêu để mua sản phẩm và độ chấp nhận của khách hàng đối với các chiến dịch tiếp thị.

Qua biểu đồ, ta thấy được các khách hàng có thu nhập càng cao thì tổng số tiền chi ra mua sản phẩm càng lớn và đa phần khách hàng không chấp nhận các chiến dịch tiếp thị.

Ở những khách hàng có tổng số tiền chi ra mua sản phẩm nằm ở mức thấp hoặc trung bình và bộ phận khách hàng thu nhập cao, chi nhiều tiền mua sản phẩm thì họ số ít chấp nhận chiến dịch tiếp thị 1 và 2 và đa phần không chấp nhận

Các chiến dịch tiếp thị được chấp nhận bởi những một số ít khách hàng có thu nhập cao và có tổng số tiền chi tiêu mua sản phẩm cao. Trong đó, chiến dịch 5 có độ chấp nhận thấp nhất trong số các chiến dịch.



Hình 2.3: Biểu đồ thể hiện mối quan hệ giữa thu nhập và tổng số tiền chi tiêu để mua sản phẩm và phản ứng của khách hàng đối với chiến dịch tiếp thị của công ty.

Với những khách hàng có thu nhập cao, mua nhiều sản phẩm thì mức độ chấp nhận các chiến dịch tiếp thị cũng cao hơn.

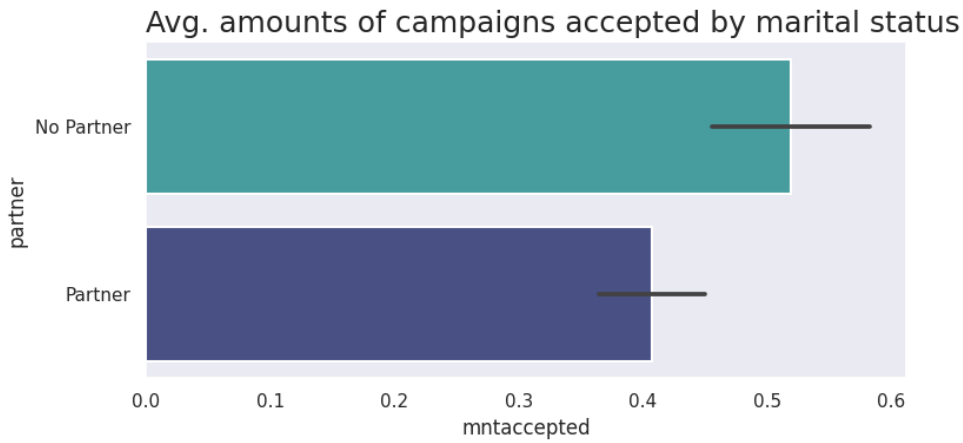
➔ Những khách hàng có thu nhập cao hơn cũng có xu hướng phản ứng tích cực hơn với các chiến dịch marketing.

Nên tập trung nỗ lực tiếp thị nhiều hơn vào những khách hàng có thu nhập cao hơn, vì họ vừa phản ứng nhanh hơn với các chiến dịch vừa có xu hướng chi tiêu nhiều hơn.

Bên cạnh đó, nên xem xét, điều chỉnh các chiến lược marketing hướng đến nhóm khách hàng có thu nhập trung bình hoặc thấp đồng thời, đưa ra các chiến lược để khuyến khích sự chấp nhận của khách hàng rộng hơn nhằm có khả năng thúc đẩy doanh số bán ra.

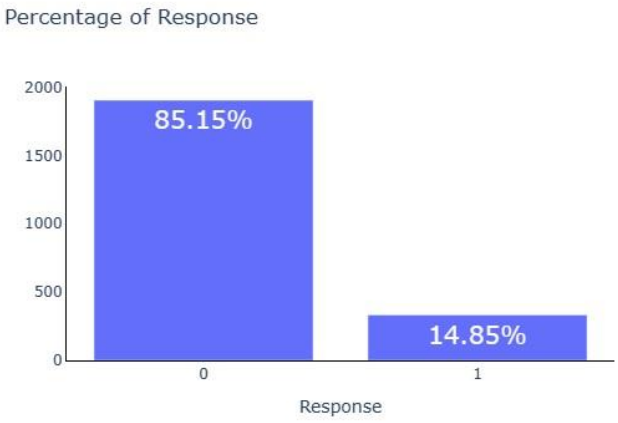
Định hướng lại kế hoạch tiếp thị do các chiến dịch càng về sau thì số lượng người chấp nhận càng ít.

Qua hai biểu đồ, ta thấy tồn tại các điểm có giá trị cực lớn phân bố lẻ tẻ, rời rạc. Đó chính là các giá trị ngoại lai, ta sẽ xử lý chúng sau.



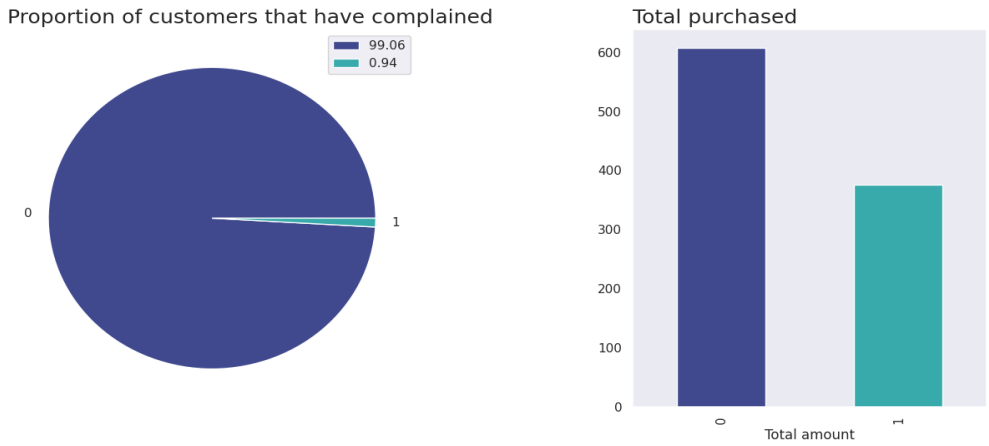
Hình 2.4: Biểu đồ thể hiện phản ứng của các khách hàng đối với các chiến dịch tiếp thị dựa vào tình trạng hôn nhân.

Qua biểu đồ, ta thấy được những khách hàng không có bạn đời dễ chấp nhận các chiến dịch tiếp thị hơn.



Hình 2.5: Biểu đồ thể hiện tỉ lệ phần trăm phản hồi của khách hàng.

Có 85.15% khách hàng không phản hồi và 14.85% khách hàng phản hồi.



Hình 2.6: Tỉ lệ phần trăm khách hàng phàn nàn và lượng mua của họ.

Ít hơn 1% khách hàng phàn nàn. Tuy chiếm thiểu số 1%, nhưng lượng sản phẩm mua của những khách hàng này lại hơn một nửa so với 99% khách hàng còn lại. Đây là một điều đáng lưu ý, có lẽ công ty hay nhà bán hàng cần lắng nghe ý kiến của khách hàng để có kế hoạch điều chỉnh hợp lý.

V Xử lý dữ liệu:

5.1 Xử lý dữ liệu ngoại lai:

Loại bỏ các giá trị ngoại lai bằng cách loại bỏ các giá trị lớn hơn phân vị 99% và nhỏ hơn phân vị 1% của bộ dữ liệu.

```
df.isnull().sum().sort_values(ascending=False)[:1]
```

```
Income    24
dtype: int64
```

Có 24 dòng có giá trị income = null, do số lượng chiếm khá ít trong bộ dữ liệu nên ta sẽ loại bỏ các dòng này.

5.2 Mã hóa dữ liệu:

Tạo 1 dataframe mới bằng cách copy data frame vừa mới xử lý. Ta sẽ tiến hành thao tác trên dataframe mới này.

Thay thế giá trị của đặc trưng partner bằng 0,1. ‘Partner’ = 1, ‘No Partner’ = 0.

Mã hóa các giá trị của đặc trưng Education như sau: ‘Basic’ = 1, ‘Graduation’ = 2, ‘2n Cycle’= 3, ‘Master’= 4, ‘PhD’= 5. Việc chọn thứ tự như trên dựa vào cấp bậc,

Basic tương đương với người đang học hoặc đã hoàn thành kiến thức phổ thông, Graduation tương đương với những người đã tốt nghiệp đại học, ‘2n Cycle’ tương đương với những người có bằng thạc sĩ theo hệ đào tạo của Ý, Master tương đương với bằng thạc sĩ, PhD tương đương với bằng tiến sĩ.

Mã hóa các giá trị của đặc trưng Age_category như sau: ‘young’ = 2, ‘adult’ = 2, ‘elderly’ = 3.

```
df['partner'] = df['partner'].replace({'Partner': 1, 'No Partner': 0})
df['age_category'] = df['age_category'].replace({'young': 1, 'adult': 2,'elderly': 3 })
df['education'] = df['education'].replace({'Basic': 1, 'Graduation': 2,'2n Cycle': 3,'Master': 4,'PhD': 5  })
```

5.3 Chọn lọc đặc trưng:

Do một số đặc trưng trở nên dư thừa trong việc phân loại nên ta sẽ loại bỏ 1 số đặc trưng

Ta thấy đặc trưng younghome là kết hợp của 2 đặc trưng kidhome và teenhome nên ta sẽ loại bỏ kidhome và teenhome chỉ giữ lại younghome.

Loại bỏ các đặc trưng id, dt_customer, z_cost contact, z_revenue, conplain, numwebvisitmonth, recency, response, year_birth, do các đặc trưng này không đóng góp nhiều cho việc phân cụm.

Bỏ đặc trưng age, giữ lại đặc trưng age_category.

Loại bỏ các đặc trưng marital_ status do đã có đặc trưng partner thay thế.’

Loại bỏ đặc trưng acceptedcmp1, acceptedcmp2, acceptedcmp3, acceptedcmp4, acceptedcmp5, mnaccepted, do đã có đặc trưng accepted thay thế.

```
feature_of_interest = 'mnttotal'
corr_matrix = df.corr()
correlations_with_feature = corr_matrix[feature_of_interest]
correlations_with_feature = correlations_with_feature[correlations_with_feature > 0.5]
print(f'Feature of Interest: {feature_of_interest}\n')
print('Correlation with Other Features:')
print(correlations_with_feature)
```

```
Feature of Interest: mnttotal

Correlation with Other Features:
income                 0.818571
mntwines               0.898939
mntfruits              0.621400
mntmeatproducts        0.858623
mntfishproducts        0.649828
mntsweetproducts       0.617456
mntgoldprods           0.537508
numwebpurchases        0.558831
numcatalogpurchases    0.804156
numstorepurchases      0.685918
mnttotal               1.000000
Name: mnttotal, dtype: float64
```

Ta thấy mnttotal có mối tương quan cao với mntwines, mntfruits, mntmeatproducts, mntfishproducts, mntsweetproducts, mntgoldprods. Đồng thời, mnttotal cũng do các đặc trưng trên kết hợp lại nên ta sẽ loại bỏ các đặc trưng trên, giữ lại mnttotal.

```
df.drop(['id','z_costcontact','z_revenue'], axis=1, inplace=True)
df.drop(['year_birth','marital_status','dt_customer','mntaccepted','complain','response','recency','age','numwebvisitsmonth'], axis=1, inplace=True)
df.drop(['acceptedcmp1','acceptedcmp2','acceptedcmp3','acceptedcmp4','acceptedcmp5'], axis=1, inplace=True)
df.drop(['kidhome','teenhome'], axis=1, inplace=True)
df.drop(['mntwines','mntfruits','mntmeatproducts','mntfishproducts','mntsweetproducts','mntgoldprods'], axis=1, inplace=True)
```

Việc mã hóa và chọn lọc đặc trưng do hàm modifi_column () thực hiện

```
def modify_column(df):
    df['partner'] = df['partner'].replace({'Partner': 1, 'No Partner': 0})
    df['age_category'] = df['age_category'].replace({'young': 1, 'adult': 2,'elderly': 3 })
    df['education'] = df['education'].replace({'Basic': 1, 'Graduation': 2,'2n Cycle': 3,'Master': 4,'PhD': 5  })
    df.drop(['id','z_costcontact','z_revenue'], axis=1, inplace=True)
    df.drop(['year_birth','marital_status','dt_customer','mntaccepted','complain','response','recency','age','numwebvisitsmonth'], axis=1, inplace=True)
    df.drop(['acceptedcmp1','acceptedcmp2','acceptedcmp3','acceptedcmp4','acceptedcmp5'], axis=1, inplace=True)
    df.drop(['kidhome','teenhome'], axis=1, inplace=True)
    df.drop(['mntwines','mntfruits','mntmeatproducts','mntfishproducts','mntsweetproducts','mntgoldprods'], axis=1, inplace=True)
    return df
```

Sau khi truyền dataframe vào hàm trên, ta có được dataframe mới đã qua xử lý còn lại 2212 dòng, 11 đặc trưng như sau:

```
df_copy_1 = df.copy()
modify_columns1(df_copy_1)
```

	education	income	numdealspurchases	numwebpurchases	numcatalogpurchases	numstorepurchases	partner	mnttotal	accepted	age_category	younghome
0	2	58138.0	3.0	8.0	10.0	4.0	0	1617.00	1	2	0
1	2	46344.0	2.0	1.0	1.0	2.0	0	27.00	0	2	2
2	2	71613.0	1.0	8.0	2.0	10.0	1	776.00	0	2	0
3	2	26646.0	2.0	2.0	0.0	4.0	1	53.00	0	1	1
4	5	58293.0	5.0	5.0	3.0	6.0	1	422.00	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...
2235	2	61223.0	2.0	9.0	3.0	4.0	1	1320.67	0	2	1
2236	5	64014.0	7.0	8.0	2.0	5.0	1	444.00	1	3	3
2237	2	56981.0	1.0	2.0	3.0	13.0	0	1241.00	1	2	0
2238	4	69245.0	2.0	6.0	5.0	10.0	1	843.00	0	2	1
2239	5	52869.0	3.0	3.0	1.0	4.0	1	172.00	1	2	2

2212 rows × 11 columns

5.4 Chuẩn hóa dữ liệu:

Áp dụng StandardScaler để chuẩn hóa dữ liệu, ta có được dataframe như sau:

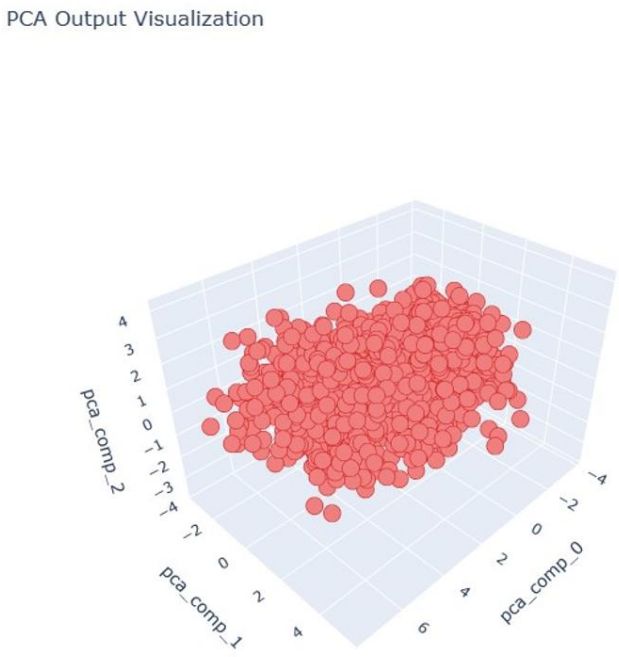
	education	income	numdealspurchases	numwebpurchases	numcatalogpurchases	numstorepurchases	partner	mnttotal	accepted	age_category	younghome
0	-0.818201	0.307895	0.393918	1.489477	2.667551	-0.562003	-1.352274	1.694808	1.633501	-0.010228	-1.265658
1	-0.818201	-0.260963	-0.165164	-1.159623	-0.593957	-1.181102	-1.352274	-0.963818	-0.612182	-0.010228	1.404476
2	-0.818201	0.957832	-0.724246	1.489477	-0.231567	1.295294	0.739495	0.288579	-0.612182	-0.010228	-1.265658
3	-0.818201	-1.211053	-0.165164	-0.781180	-0.956347	-0.562003	0.739495	-0.920344	-0.612182	-2.066897	0.069409
4	1.531251	0.315371	1.512082	0.354149	0.130822	0.057096	0.739495	-0.303342	-0.612182	-0.010228	0.069409
...	...	...	...	...	...	...	...	...	...	...	...
2207	-0.818201	0.456693	-0.165164	1.867920	0.130822	-0.562003	0.739495	1.199317	-0.612182	-0.010228	0.069409
2208	1.531251	0.591311	2.630247	1.489477	-0.231567	-0.252454	0.739495	-0.266556	1.633501	2.046442	2.739542
2209	-0.818201	0.252089	-0.724246	-0.781180	0.130822	2.223942	-1.352274	1.066101	1.633501	-0.010228	-1.265658
2210	0.748100	0.843616	-0.165164	0.732591	0.855602	1.295294	0.739495	0.400609	-0.612182	-0.010228	0.069409
2211	1.531251	0.053756	0.393918	-0.402737	-0.593957	-0.562003	0.739495	-0.721365	1.633501	-0.010228	1.404476

2212 rows × 11 columns

5.5 Trích xuất đặc trưng:

Áp dụng PCA để giảm số chiều của dữ liệu:

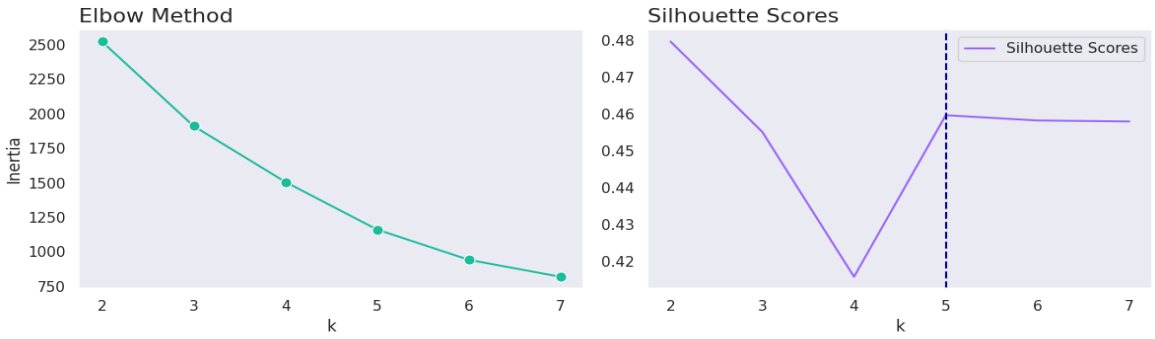
	pca_comp_0	pca_comp_1	pca_comp_2
0	4.496368	0.453069	1.899701
1	-2.657896	-0.604681	0.094160
2	1.873023	-0.289547	-0.946323
3	-2.628000	-1.022696	0.350576
4	-0.181613	0.784415	-1.095471



Hình 2.7: Dữ liệu sau khi đã giảm số chiều xuống còn 3.

VI Huấn luyện mô hình:

Với mô hình kmean, chọn số cụm từ 2 đến 7 và dùng thang đo Elbow và Silhouette.

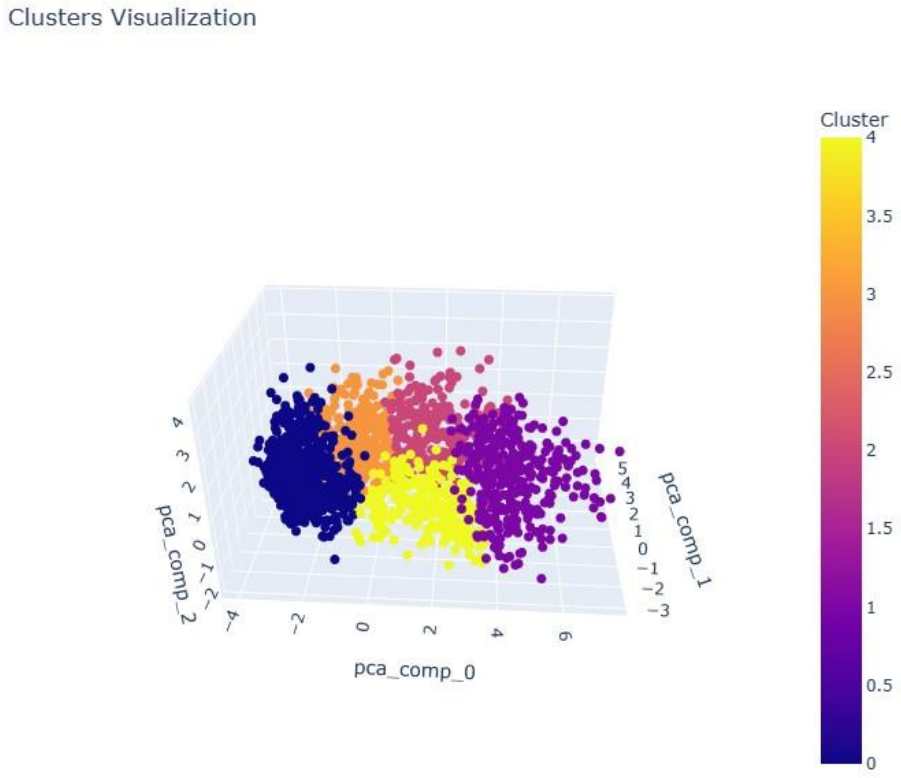


Hình 2.8: Biểu đồ thể hiện chỉ số Elbow và chỉ số Silhouette khi phân cụm từ 2 đến 7 cụm.

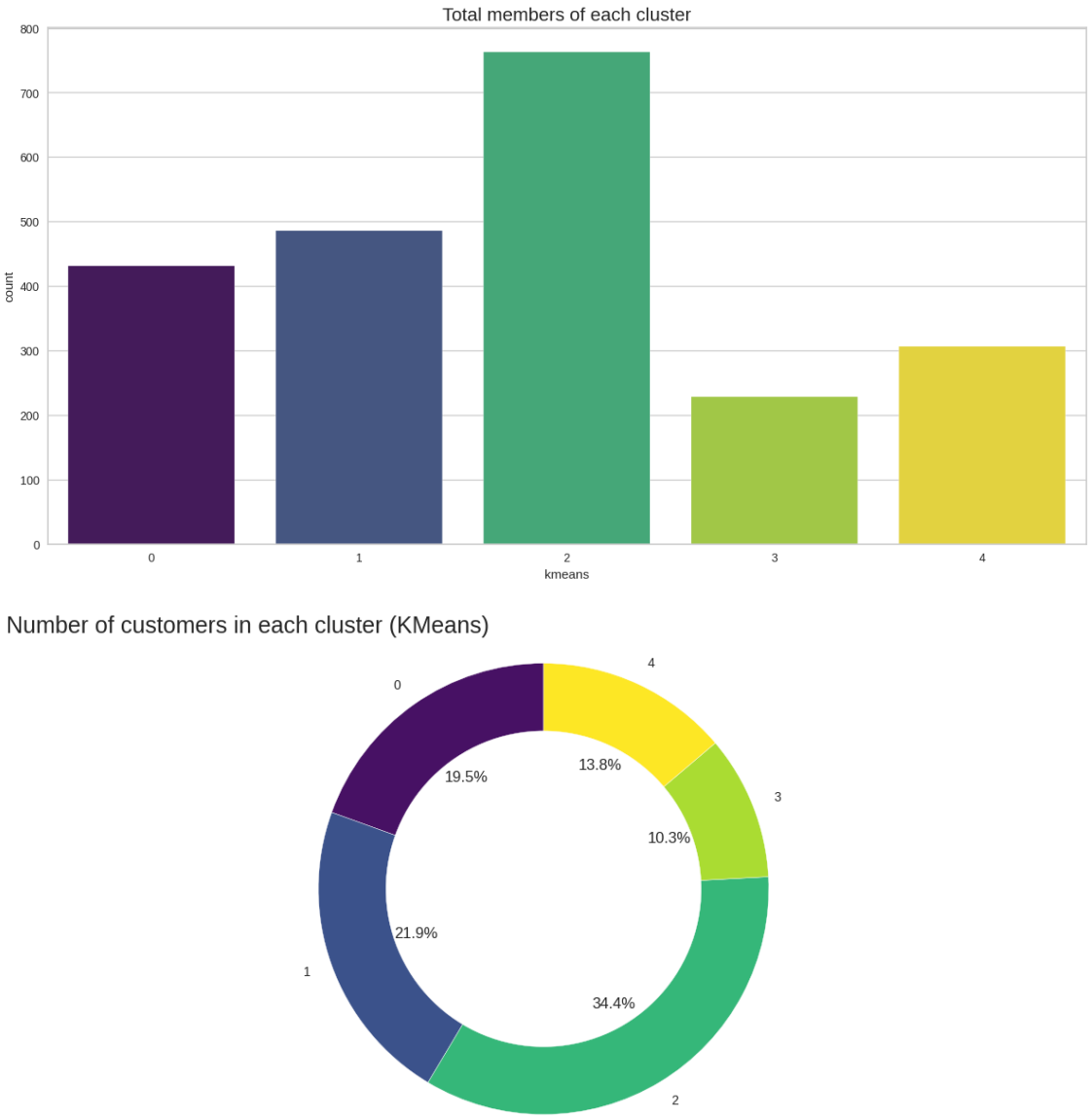
Qua biểu đồ, ta thấy được ở cụm số 3 và cụm số 5, biểu đồ có độ dốc giảm nhanh chóng hơn so với các cụm còn lại

Tuy nhiên, ở cụm số 3 có chỉ số Silhouette thấp hơn so với cụm 5 nên ta sẽ chọn 5 là điểm khuỷu tay và chọn số cụm là 5.

VII Nhận dạng mẫu:



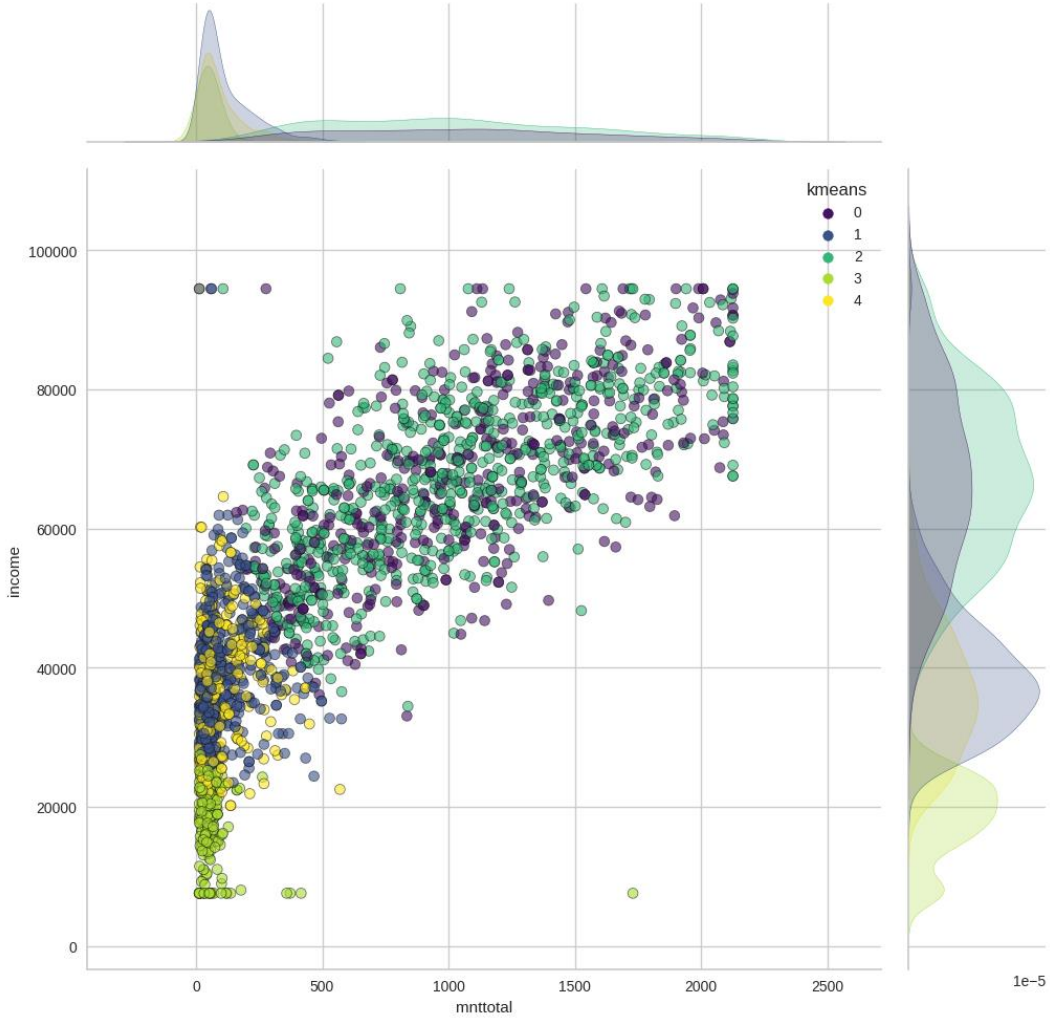
Hình 2.9: Trực quan hóa kết quả phân cụm trên không gian 3 chiều.



Hình 2.10: Biểu đồ thể hiện số lượng thành viên của từng cụm và tỉ lệ phần trăm mỗi cụm chiếm trong bộ dữ liệu.

Từ các biểu đồ trên, ta thấy được nhóm 2 chiếm nhiều nhất trong bộ dữ liệu với 34.4% kế đến là nhóm 1 và nhóm 0 với tỉ lệ phần trăm là 21.9% và 19.5%.

Nhóm 3 và 4 chiếm phần trăm khá ít so với 3 nhóm còn lại trong đó, nhóm 3 chiếm ít nhất với 10.3% và nhóm 4 chiếm 13.8% bộ dữ liệu.

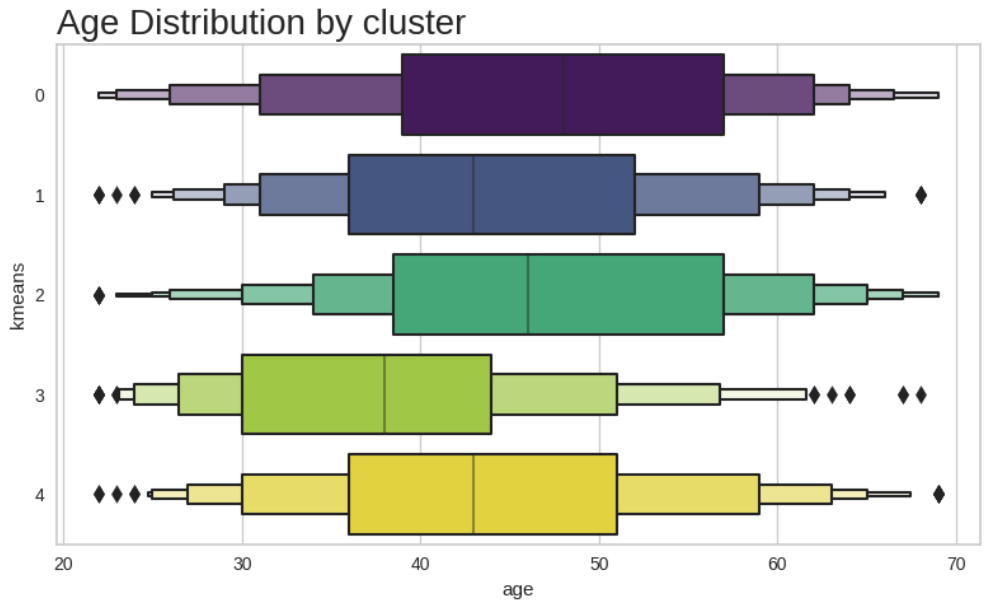


Hình 3.1: Biểu đồ thể hiện mối quan hệ giữa thu nhập và sức mua của khách hàng ở từng cụm.

Nhóm 0 và 2 dường như mua nhiều sản phẩm hơn và có thu nhập trung bình cao hơn các nhóm còn lại.

Nhóm 3 có thu nhập thấp và mua ít sản phẩm hơn.

Nhóm 1 và 4, mặc dù có thu nhập cao hơn nhưng thường không mua nhiều sản phẩm.

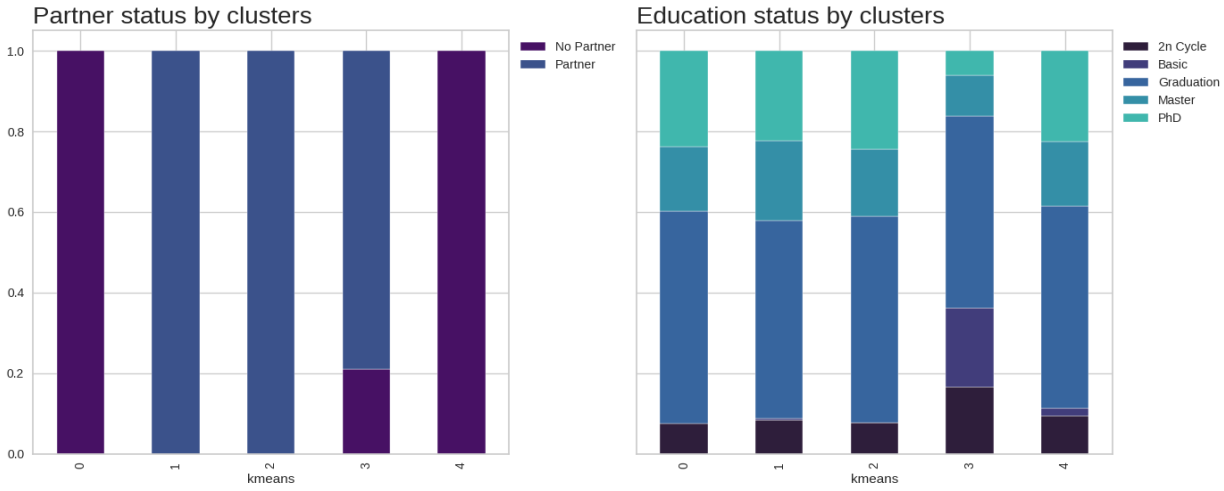


Hình 3.2: Biểu đồ thể hiện sự phân phối độ tuổi ở các cụm.

Nhóm 0 và 2 bao gồm những khách hàng lớn tuổi nhất, kết hợp với biểu đồ ở trên thì có thể thấy nhóm tuổi này có xu hướng có thu nhập cao hơn.

Nhóm 3 gồm những khách hàng nhỏ tuổi nhất.

Nhóm 1 và 4 chủ yếu bao gồm người lớn tuổi trung niên.



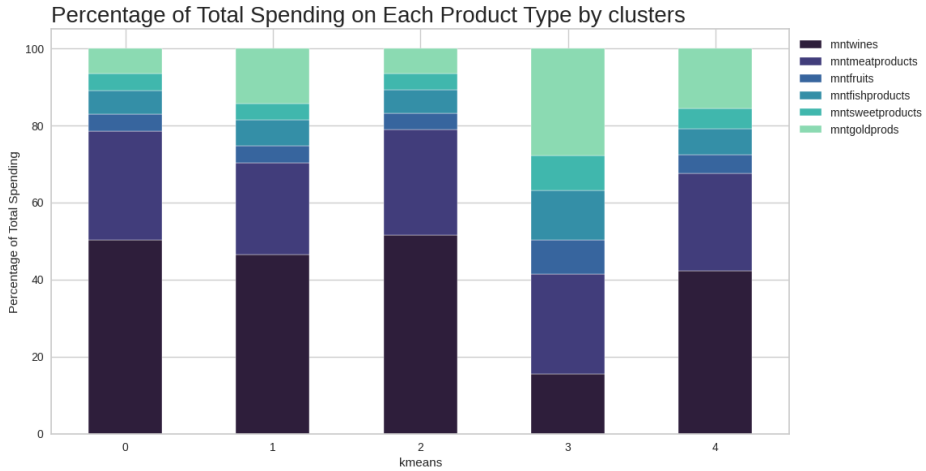
Hình 3.3: Biểu đồ thể hiện tình trạng hôn nhân và trình độ học vấn của các khách hàng ở các cụm.

Nhóm 0 và 4 hoàn toàn bao gồm các khách hàng không có bạn đời.

Nhóm 1 và 2 bao gồm hoàn toàn khách hàng có bạn đời.

Nhóm 3 chủ yếu là khách hàng có bạn đời

Trình độ học vấn phần lớn giống nhau ở tất cả các nhóm, ngoại trừ nhóm 3, nơi số lượng khách hàng chưa tốt nghiệp cao hơn đáng kể, có thể là do độ tuổi của nhóm tương đối trẻ hơn so với các nhóm còn lại.



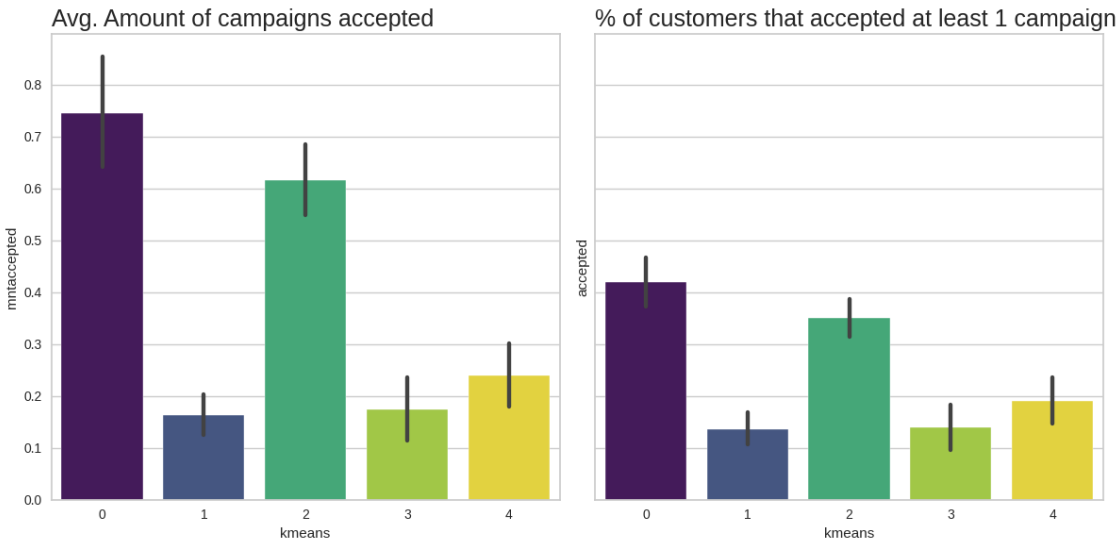
Hình 3.4 Biểu đồ thể hiện phần trăm chi tiêu cho từng mặt hàng ở các cụm.

Mô hình chi tiêu vẫn tương đối giống nhau ở tất cả các nhóm, ngoại trừ nhóm 3, chi tiêu ít hơn cho rượu vang nhưng phân bổ nhiều ngân sách hơn để trữ vàng.

Nhóm 0 và 2 thích chi tiêu nhiều hơn cho rượu và thịt hơn vàng.

Nhóm 1 và 4 thì chi tiêu nhiều cho rượu

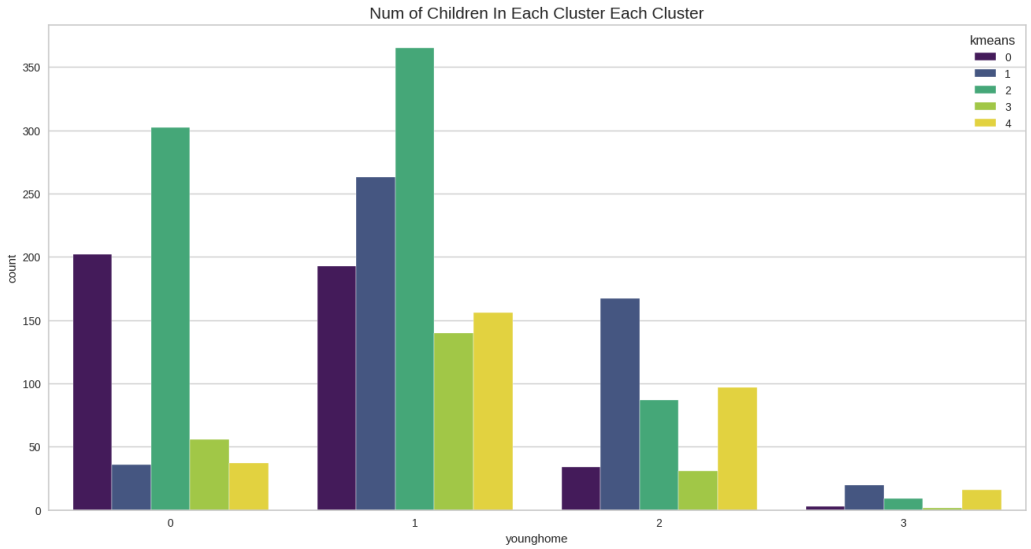
Điều này cho thấy xu hướng liên quan đến tuổi tác. Những cá nhân trẻ hơn có thể nghiêng về đầu tư vào vàng, trong khi những cá nhân lớn tuổi có xu hướng ưu tiên rượu, ở những khách hàng mua nhiều sản phẩm thì bên cạnh rượu họ cũng có xu hướng ưu tiên thịt.



Hình 3.5: Biểu đồ thể hiện mức độ chấp nhận của khách hàng đối với tiếp thị.

Nhóm 0 và 2 có xu hướng chấp nhận các chiến dịch tiếp thị cao hơn đáng kể.

Nhóm 1 có xu hướng chấp nhận các chiến dịch tiếp thị thấp nhất. Các nhóm còn lại không thể hiện sự khác biệt đáng chú ý về mặt này.

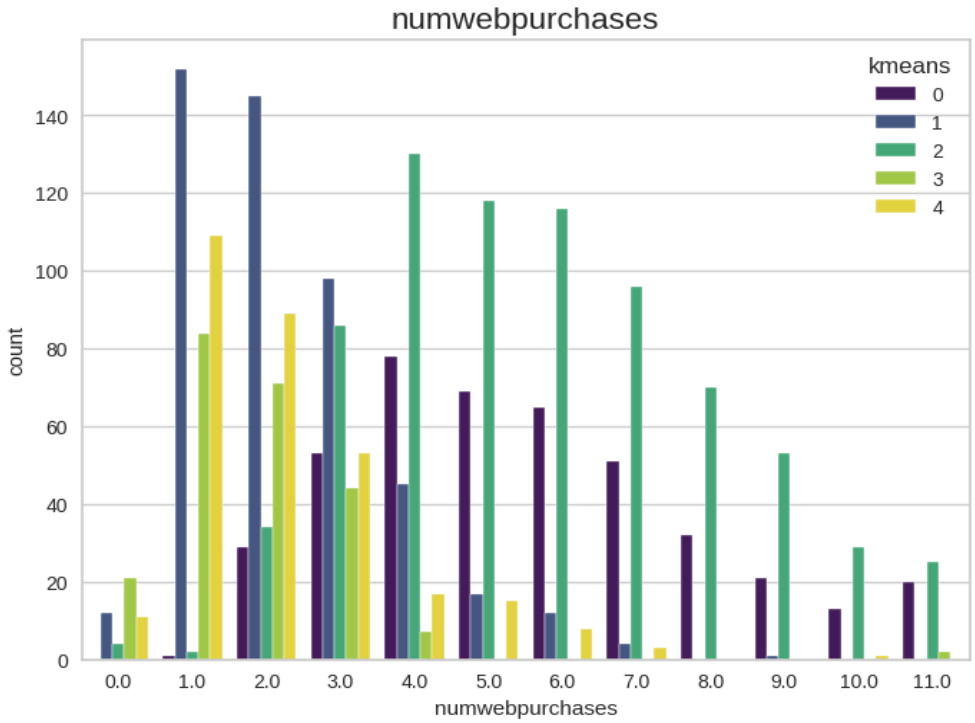


Hình 3.6: Biểu đồ thể hiện số con của khách hàng ở mỗi cụm.

Nhóm 0, nhóm 2 và nhóm 3 đa phần có 1 hoặc 0 con.

Nhóm 1 và nhóm 4 đa phần có 1 hoặc 2 con.

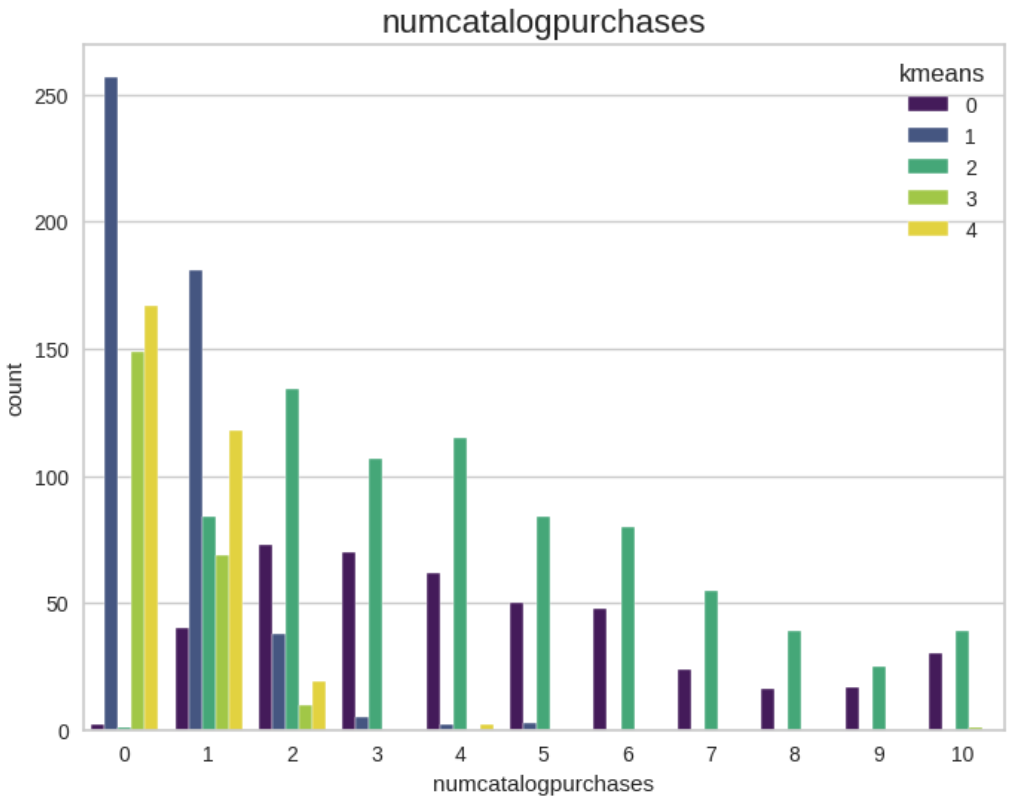
Các khách hàng có 3 con chiếm tỉ lệ khá ít trong các nhóm.



Hình 3.7: Biểu đồ thể hiện số lượt mua hàng trên web ở các cụm.

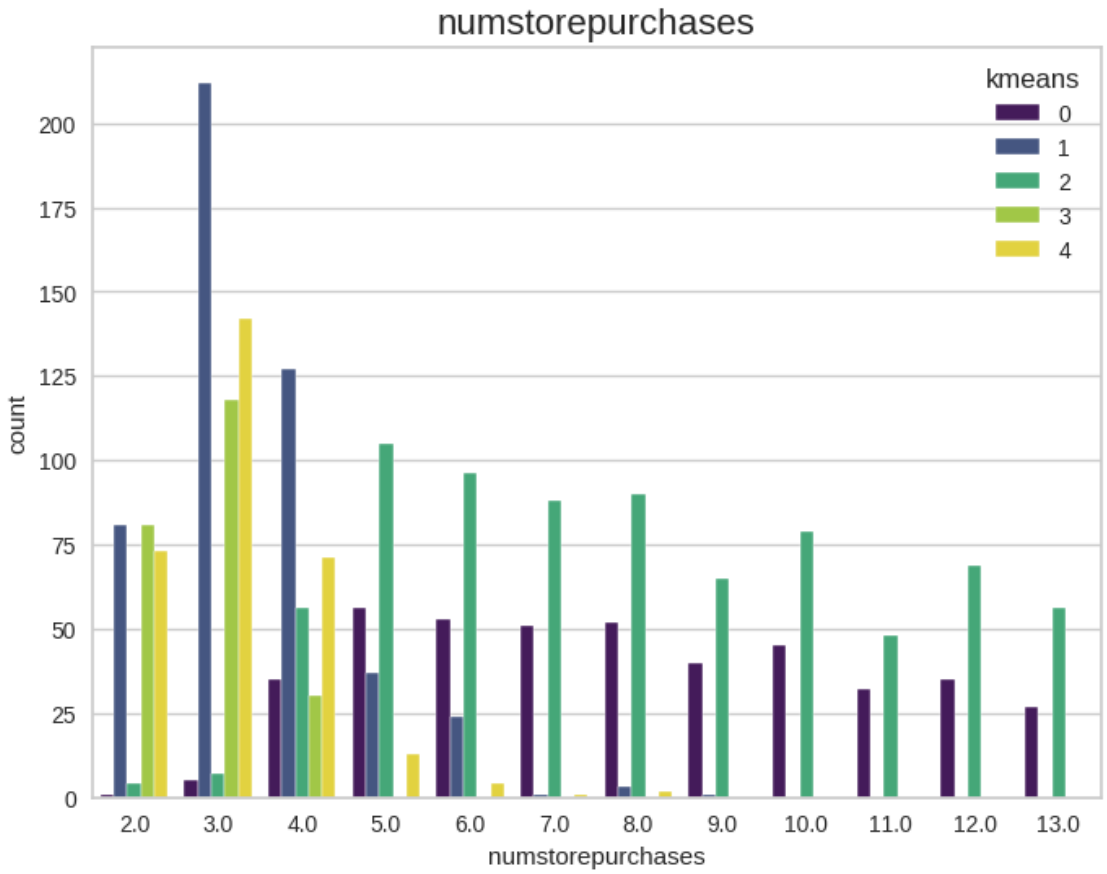
Nhóm 0 và 2 có số giao dịch mua hàng trên web cao hơn so với 3 cụm còn lại. Trong đó, nhóm 2 là cao nhất.

Đa phần các khách hàng trong nhóm 1 và 3 có số lượt mua hàng trên web dao động từ 1.0 đến 3.0.



Hình 3.8: Biểu đồ thể hiện số lượt mua hàng qua danh mục ở các cụm.

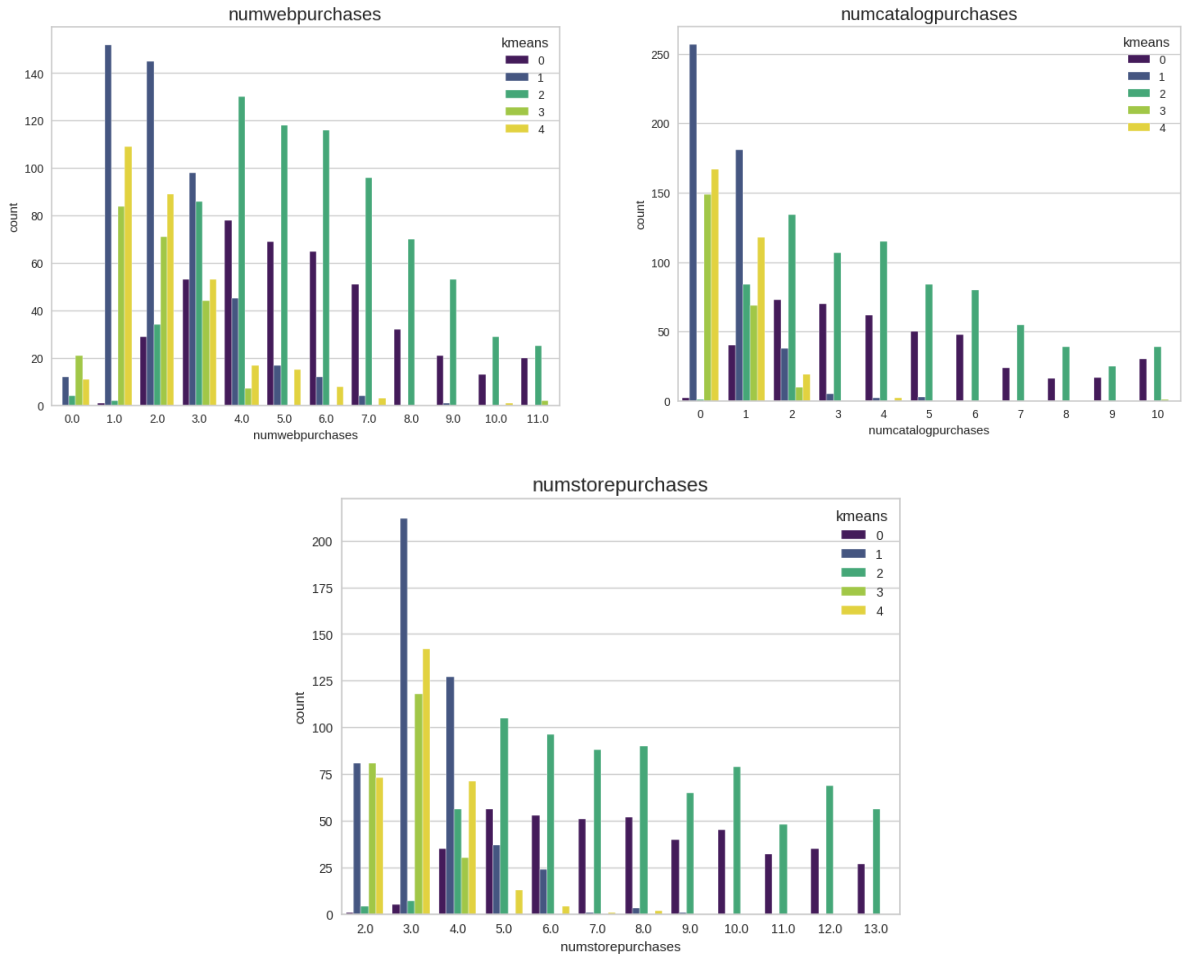
Đa phần các khách hàng ở nhóm 1, 3, 4 có số lượt mua hàng theo danh mục là 0. Nhóm 0, 2 có số lượt mua hàng theo danh mục dao động từ 1 đến 10. Cho thấy chỉ có các khách hàng nhóm 0 và 2 mua hàng qua danh mục.



Hình 3.9: Biểu đồ thể hiện số lượt mua hàng ở cửa hàng ở các cụm.

Đa phần các khách hàng trong nhóm 1, 3, 4 có số lượt mua hàng ở cửa hàng dao động từ 2.0 cho đến 4.0.

Riêng nhóm 0 và nhóm 2 thì phần lớn có số lượt mua hàng ở cửa hàng dao động từ 5.0 đến 13.0.



Hình 3.10: Biểu đồ thể hiện số lượt mua hàng thông qua web, danh mục và cửa hàng .

→Nhóm 1, 3, 4 có vẻ không ưa thích mua hàng trên danh mục, do đó, việc mua hàng trên danh mục đa phần được thực hiện bởi những khách hàng thuộc nhóm 0 và 2.

Nhóm 1, 3, 4 đa phần ưa thích mua hàng ở cửa hàng, web là kênh thanh toán còn lại.

Nhóm 1 và 2 có số lượt mua hàng nhiều nhất và mua hàng ở tất cả các kênh thanh toán tuy nhiên kênh thanh toán ưa thích là web.

Kết luận:

Nhóm 0:

Chiếm 19.5% bộ dữ liệu. Gồm những khách hàng lớn tuổi nhất, thu nhập cao, mua nhiều sản phẩm, không có bạn đời, đa phần các khách hàng có trình độ học vấn từ đại học trở lên, có xu hướng chấp nhận các chiến dịch tiếp thị đáng kể, đa phần có 1 con hoặc không có con.

Mặt hàng mua nhiều là rượu và thịt, mua hàng ở cửa hàng, web, danh mục, tuy nhiên kênh thanh toán yêu thích nhất là web.

Nhóm 1

Chiếm 21.9% bộ dữ liệu. Gồm các khách hàng tuổi trung niên, thu nhập cao nhưng không mua nhiều sản phẩm, có bạn đời, đa phần các khách hàng có trình độ học vấn từ đại học trở lên, có xu hướng ít chấp nhận các chiến dịch tiếp thị, đa phần có từ 1 đến 2 con.

Chi tiêu cho rượu nhiều nhất, mua hàng ở web và cửa hàng, có xu hướng ưa thích thanh toán ở cửa hàng nhất (số lượt mua hàng trên web phần lớn dao động từ 1.0 đến 3.0 còn số lượt mua hàng ở cửa hàng phần lớn dao động từ 2.0 đến 4.0.)

Nhóm 2:

Chiếm 34.4% bộ dữ liệu. Gồm những khách hàng lớn tuổi nhất, thu nhập cao, mua nhiều sản phẩm, đa phần các khách hàng có trình độ học vấn từ đại học trở lên, có xu hướng chấp nhận các chiến dịch tiếp thị đáng kể, có bạn đời nhưng đa phần chỉ có 1 hoặc không có con.

Chi tiêu cho rượu và thịt nhiều nhất, mua hàng ở cửa hàng, web, danh mục, kênh thanh toán yêu thích là web.

Nhóm 3:

Chiếm 10.3% bộ dữ liệu. Gồm những khách hàng nhỏ tuổi nhất, thu nhập không cao, sản phẩm mua cũng ít, đa phần các khách hàng có trình độ học vấn từ đại học trở lên, tuy nhiên, lượng khách hàng chưa tốt nghiệp nhiều hơn so với 4 nhóm còn lại, có xu hướng chấp nhận các chiến dịch tiếp thị trung bình, số lượng khách hàng có bạn đời chiếm phần lớn. Đa phần có 1 con hoặc không có con.

Phần lớn chi tiêu để trữ vàng, mua hàng ở cửa hàng, web, có xu hướng ưa thích thanh toán ở cửa hàng nhất (số lượt mua hàng trên web phần lớn dao động từ 1.0 đến 2.0 còn số lượt mua hàng ở cửa hàng phần lớn dao động từ 2.0 đến 3.0.)

Nhóm 4:

Chiếm 13.8% bộ dữ liệu. Gồm các khách hàng tuổi trung niên, thu nhập cao nhưng không mua nhiều sản phẩm, không có bạn đời, đa phần các khách hàng có trình độ học vấn từ đại học trở lên, có xu hướng chấp nhận các chiến dịch tiếp thị trung bình, đa phần có từ 1 đến 2 con. Có thể đoán được các khách hàng này thuộc nhóm ly hôn hoặc góa phụ mà đặc trưng marital_status của bộ dữ liệu gốc chưa qua xử lý đã đề cập.

Chi tiêu cho rượu và thịt nhiều nhất, mua hàng ở cửa hàng, web, có xu hướng ưa thích thanh toán ở cửa hàng nhất (số lượt mua hàng trên web phần lớn dao động từ 1.0 đến 3.0 còn số lượt mua hàng ở cửa hàng phần lớn dao động từ 2.0 đến 4.0.)