TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

HỌC PHẦN PHÂN TÍCH NHẬN DẠNG MẪU

CHỦ ĐỀ:

Xử lý dữ liệu và áp dụng giải thuật k-means cho bài toán gom nhóm khách hàng

Họ và tên sinh viên: **Nguyễn Châu Hiếu Duy**

Mã số sinh viên: **3120410092**

Mã học phần: **841453**

Học kỳ: **1**

Năm học: **2023-2024**

Người hướng dẫn: TS. Vũ Ngọc Thanh Sang

TP.Hồ Chí Minh, tháng 12 năm 2023

TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN
HỌC PHẦN PHÂN TÍCH NHẬN DẠNG MẪU

CHỦ ĐỀ:

Xử lý dữ liệu và áp dụng giải thuật k-means cho bài toán gom nhóm khách hàng

Họ và tên sinh viên: **Nguyễn Châu Hiếu Duy**
Mã số sinh viên: **3120410092**

Mã học phần: **841453**
Học kỳ: **1**
Năm học: **2023-2024**

Người hướng dẫn: TS. Vũ Ngọc Thanh Sang

TP.Hồ Chí Minh, tháng 12 năm 2023

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC BIỂU ĐỒ

LỜI MỞ ĐẦU:

Trong bối cảnh cuộc cách mạng số hiện nay, doanh nghiệp đối mặt với thách thức lớn trong việc định hình chiến lược tiếp thị và tương tác với khách hàng một cách linh hoạt và hiệu quả. Việc chuyển đổi số không chỉ là xu hướng mà còn là yếu tố quyết định sự thành công của doanh nghiệp. Khả năng hiểu rõ và gom nhóm khách hàng đang trở thành một phần quan trọng của chiến lược số hóa, giúp doanh nghiệp tối ưu hóa trải nghiệm khách hàng và tạo ra giá trị cao nhất từ dữ liệu có sẵn.

Gom nhóm khách hàng là một bài toán nổi bật trong lĩnh vực phân tích nhận dạng mẫu. Với đồ án này, em sẽ tập trung nghiên cứu về các phương pháp xử lý dữ liệu và áp dụng giải thuật k-means clustering để giải quyết vấn đề.

Với giá trị của việc phân loại khách hàng bằng K-means clustering, chúng ta có khả năng nhìn nhận được đặc điểm chung trong nhóm khách hàng, từ đó xây dựng chiến lược phân khúc hóa hiệu quả. Bằng cách này, doanh nghiệp có thể cá nhân hóa sản phẩm, dịch vụ, và chiến lược tiếp thị để phản ánh đúng nhu cầu và mong muốn của từng nhóm, tạo ra trải nghiệm khách hàng tích cực và tăng cường mối quan hệ khách hàng.

Mong rằng, qua đồ án này, em có thể trình bày một quy trình phân tích nhận dạng mẫu cho bài toán đã nêu cũng như chia sẻ và truyền đạt được ý nghĩa và tiềm năng ứng dụng của phân loại khách hàng trên bộ dữ liệu Marketing Campaign (kaggle.com).

Source code: PTNDM_SGU/PTNDM_project.ipynb at main · DuyAccel/PTNDM_SGU (github.com)

NỘI DUNG

1. Phân tích nhận dạng mẫu là gì?

1.1. Khái niệm

Phân tích nhận dạng mẫu (Pattern Recognition) là một lĩnh vực của trí tuệ nhân tạo và thống kê mà nói chung, tập trung vào quá trình nhận diện và phân loại các mô hình, cấu trúc, hay đặc điểm trong dữ liệu. Mục tiêu chính của phân tích nhận dạng mẫu là phát hiện các quy luật ẩn và thông tin có ý nghĩa từ dữ liệu đầu vào. Nói cách khác, nó có thể được xem là việc "cần thực hiện một tác động vào dữ liệu thô mà tác động cụ thể là gì sẽ tùy vào loại của dữ liệu đó" [1].

Phân tích nhận dạng mẫu có ứng dụng rộng rãi trong nhiều lĩnh vực như nhận dạng hình ảnh, nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, y học, và nhiều lĩnh vực khác nữa. Đối với việc phân loại khách hàng trong doanh nghiệp, phân tích nhận dạng mẫu có thể giúp xác định nhóm khách hàng tương tự và tối ưu hóa chiến lược tiếp thị.

1.2. Các thành phần chính

Sau đây là một số thành phần thường được đề cập đến trong các bài toán phân tích nhận dạng mẫu:

- *Mẫu (Pattern):* Một mẫu là một cấu trúc hoặc sự xuất hiện đặc biệt của thông tin trong dữ liệu. Đối với phân tích nhận dạng mẫu, mẫu có thể là hình dạng, biểu đồ, hoặc bất kỳ đặc điểm nào có thể được nhận diện và sử dụng để phân loại hay dự đoán.
- *Nhận dạng (Recognition):* Nhận dạng mẫu là quá trình xác định một mẫu cụ thể trong dữ liệu, dựa trên các quy luật hay đặc điểm đã được học từ các dữ liệu mẫu.
- *Gom cụm (Clustering):* Là quá trình tổ chức các điểm dữ liệu thành các nhóm hay cụm, sao cho các điểm trong cùng một cụm giống nhau và điểm giữa các cụm khác biệt nhau. Gom cụm không yêu cầu thông tin lớp hay nhãn trước đó.

- *Học máy (Machine Learning):* Phân tích nhận dạng mẫu thường liên quan đến việc sử dụng các phương pháp học máy để tự động học các quy luật và mô hình từ dữ liệu mẫu, sau đó áp dụng chúng cho dữ liệu mới.
- *Đặc trưng (Feature):* Là một phần hoặc thuộc tính của một mẫu, được sử dụng để mô tả và phân biệt giữa các mẫu khác nhau.
- *Sự Tổng quan (Generalization):* Khả năng của mô hình nhận diện mẫu có thể áp dụng những gì đã học từ dữ liệu huấn luyện vào dữ liệu mới mà nó chưa từng gặp.

2. Thấu hiểu dữ liệu

2.1. Mô tả đặc trưng

Sau đây là mô tả các đặc trưng được cung cấp bởi tác giả của bộ dữ liệu:

- `AcceptedCmp1` - 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- `AcceptedCmp2` - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- `AcceptedCmp3` - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- `AcceptedCmp4` - 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- `AcceptedCmp5` - 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- `Response` - 1 if customer accepted the offer in the last campaign, 0 otherwise
- `Complain` - 1 if customer complained in the last 2 years
- `DtCustomer` - date of customer's enrolment with the company
- `Education` - customer's level of education
- `Marital` - customer's marital status
- `Kidhome` - number of small children in customer's household
- `Teenhome` - number of teenagers in customer's household
- `Income` - customer's yearly household income
- `MntFishProducts` - amount spent on fish products in the last 2 years
- `MntMeatProducts` - amount spent on meat products in the last 2 years
- `MntFruits` - amount spent on fruits products in the last 2 years
- `MntSweetProducts` - amount spent on sweet products in the last 2 years
- `MntWines` - amount spent on wine products in the last 2 years
- `MntGoldProds` - amount spent on gold products in the last 2 years
- `NumDealsPurchases` - number of purchases made with discount
- `NumCatalogPurchases` - number of purchases made using catalogue
- `NumStorePurchases` - number of purchases made directly in stores
- `NumWebPurchases` - number of purchases made through company's web site
- `NumWebVisitsMonth` - number of visits to company's web site in the last month
- `Recency` - number of days since the last purchase
- `Year_Birth` - customer's year of birth

Bảng 1 : Mô tả các đặc trưng của dữ liệu.

2.2. Kiểm tra tính hợp lý của dữ liệu

Thông quan việc quan sát tổng thể dữ liệu, em phát hiện một số tính chất từ bộ dataset:

- Bộ dữ liệu bao gồm **2240 instances** và **29 đặc trưng**.
- 3 đặc trưng trong số chúng không được tác giả mô tả ở trên là **ID, Z_CostContact, Z_Revenue**.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   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
```

Bảng 2: Tổng quan bộ dữ liệu.

Để đánh giá cụ thể hơn về những đặc trưng ngoại lệ này, em đã thống kê số lượng giá trị đơn nhất trong chúng và nhận thấy:

- Đặc trưng **ID** biểu diễn các index của dữ liệu vì mỗi instance đều có giá trị khác nhau → đây là đặc trưng không cần thiết.
- Đặc trưng **Z_CostContact, Z_Revenue** chỉ có duy nhất 1 giá trị trong cả bộ dữ liệu → đây là đặc trưng không cần thiết.

→ Ta hoàn toàn có thể loại bỏ những đặc trưng trên vì chúng chẳng có ý nghĩa nào cho việc gom nhóm dữ liệu.

ID	2240
Z_CostContact	1
Z_Revenue	1

Bảng: Số lượng phần tử đơn nhất (unique) của các đặc trưng ngoại lai.

Sau khi bỏ đi những đặc trưng nêu trên, ta tiếp tục quan sát tổng quan những dữ liệu còn lại. Dựa vào việc thống kê số lượng phần tử đơn nhất, kết hợp cùng việc quan sát các instances cũng như mô tả dữ liệu của tác giả, ta nắm được tính chất, kiểu dữ liệu của những đặc trưng nhằm phục vụ mục đích xử lý sau này.

Year_Birth	59
Education	5
Marital_Status	8
Income	1974
Kidhome	3
Teenhome	3
Recency	100
MntWines	776
MntFruits	158
MntMeatProducts	558
MntFishProducts	182
MntSweetProducts	177
MntGoldProds	213
NumDealsPurchases	15
NumWebPurchases	15
NumCatalogPurchases	14
NumStorePurchases	14
NumWebVisitsMonth	16
AcceptedCmp3	2
AcceptedCmp4	2
AcceptedCmp5	2
AcceptedCmp1	2
AcceptedCmp2	2
Complain	2
Response	2

Bảng 3: Số phần tử đơn nhất còn lại của dữ liệu.

	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts
0	1957	Graduation	Single	58138.0	0	0	58	635	88	546
1	1954	Graduation	Single	46344.0	1	1	38	11	1	6
2	1965	Graduation	Together	71613.0	0	0	26	426	49	127
3	1984	Graduation	Together	26646.0	1	0	26	11	4	20
4	1981	PhD	Married	58293.0	1	0	94	173	43	118

	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
0	172	88	88	3	8	10	4
1	2	1	6	2	1	1	2
2	111	21	42	1	8	2	10
3	10	3	5	2	2	0	4
4	46	27	15	5	5	3	6

	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Response
0	7	0	0	0	0	0	0	1
1	5	0	0	0	0	0	0	0
2	4	0	0	0	0	0	0	0
3	6	0	0	0	0	0	0	0
4	5	0	0	0	0	0	0	0

Bảng 4: 5 dòng đầu tiên của bộ dữ liệu.

2.3. Xác định kiểu dữ liệu

Như đã đề cập ở trên, việc quan sát và đánh giá là một bước chuẩn bị để phục vụ xác định kiểu dữ liệu, những đặc trưng dưới 10 giá trị đơn nhất mà ta quan sát được rất phù hợp dưới vai trò là một đặc trưng phân loại (categorical). Một số chúng đã được mã hóa sẵn trong khi phần còn lại thì không. Các đặc trưng số học (numerical) chiếm nhiều hơn một chút trong bộ dữ liệu của chúng ta.

Việc đánh giá đặc trưng nào là phân loại/số học cũng có thể tùy thuộc vào cách nhìn nhận của người xử lý, sau đây là đánh giá của em:

Số lượng đặc trưng phân loại: 11
['Education', 'Marital_Status', 'Kidhome', 'Teenhome', 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'Complain', 'Response']

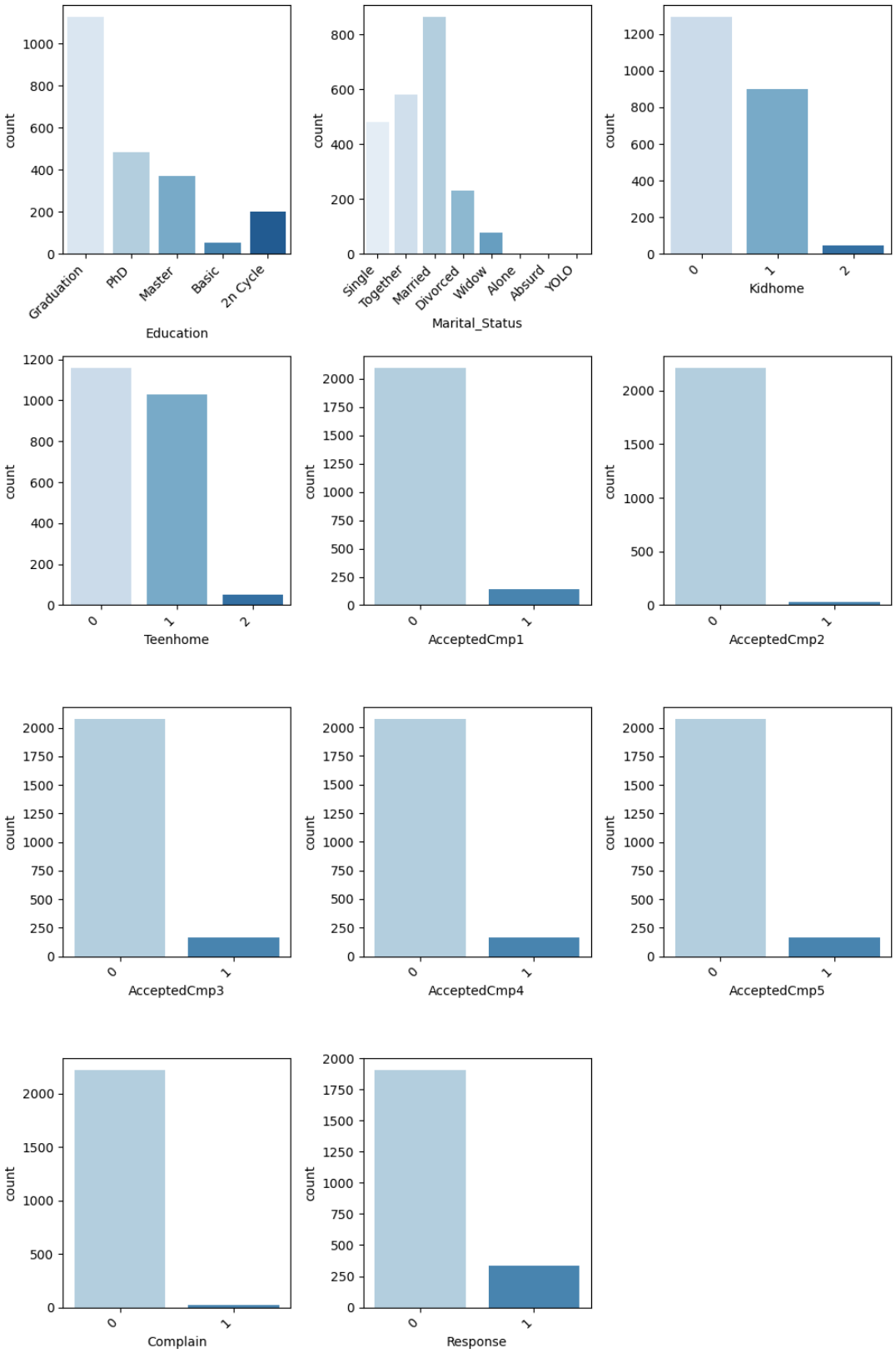
Số lượng đặc trưng số học: 14
['Year_Birth', 'Income', 'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth']

2.4. Phân phối của dữ liệu

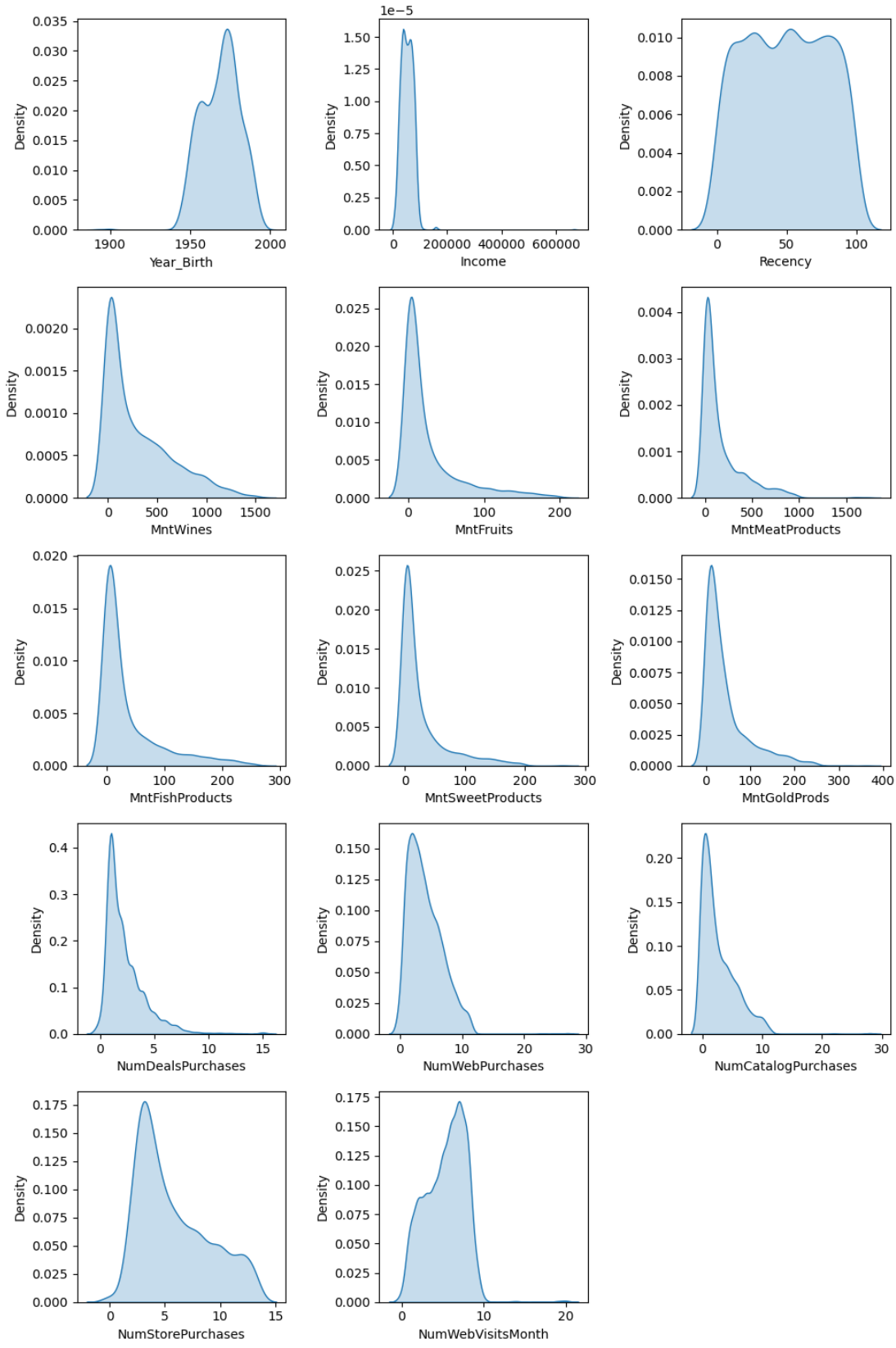
Một bước nữa để thấu hiểu dữ liệu chính là phát họa biểu đồ phân phối của từng đặc trưng. Việc phân phối dữ liệu của một đặc trưng là lệch hay cân bằng có thể là một yếu tố để hiểu hơn về tập khách hàng của công ty. Những đặc trưng bị lệch quá lớn về một phía (có thể do ngoại lệ/nhiễu) sẽ làm giảm hiệu suất của mô hình.

Việc phát họa biểu đồ phân phối của từng kiểu đặc trưng cũng có sự khác biệt, những đặc trưng phân loại với số lượng ít các giá trị unique sẽ sử dụng biểu đồ dạng cột để biểu diễn phân phối. Trong khi đặc trưng số học có dữ liệu thường là liên tục sẽ cần sử dụng những loại biểu đồ như histogram, kde-plot, boxplot để biểu diễn phân phối.

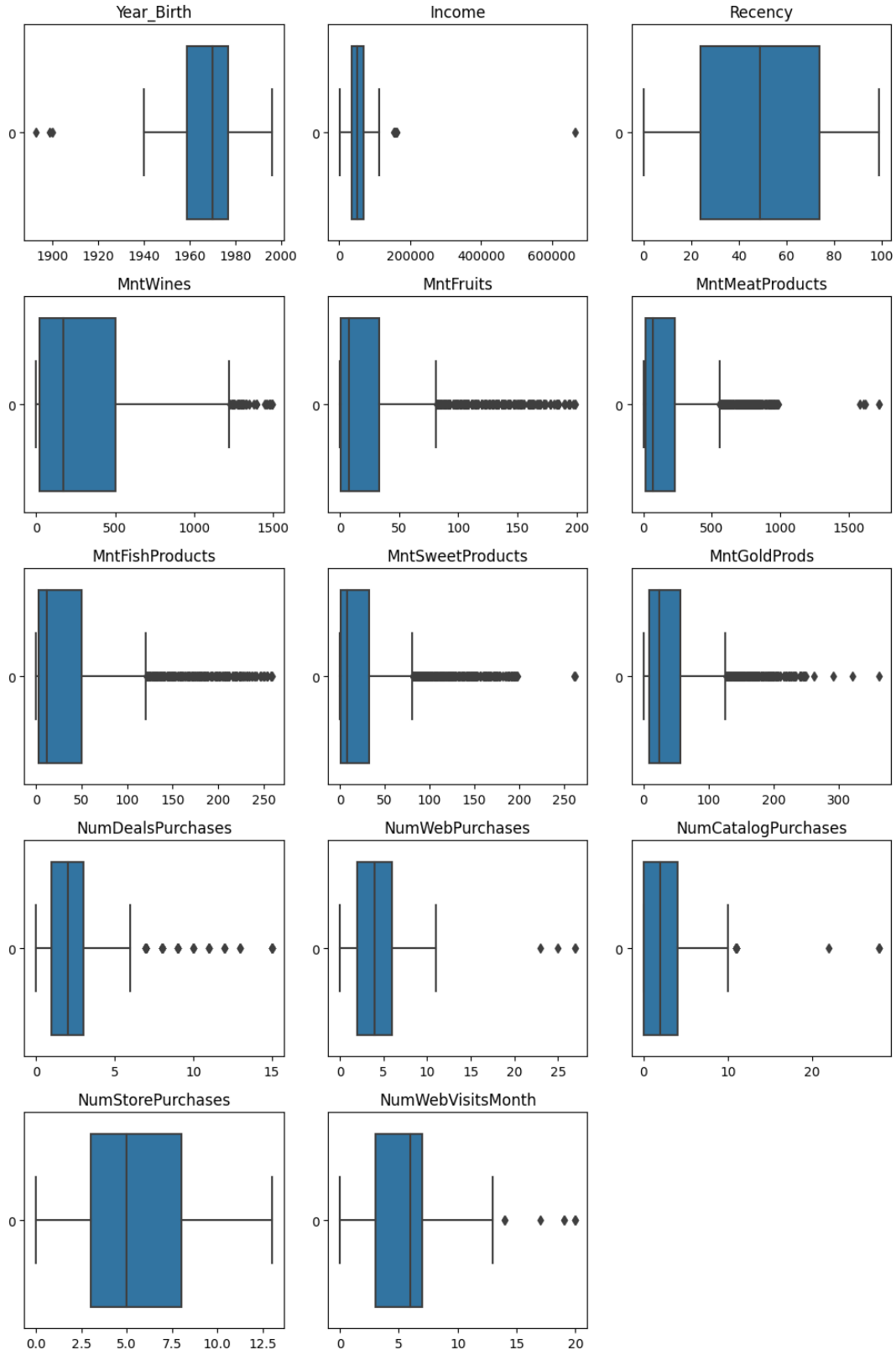
Đối với đồ án của mình, em sử dụng countplot cho các đặc trưng phân loại và kde-plot, box-plot cho công việc trực quan hóa phân phối dữ liệu các đặc trưng số học. Sở dĩ việc áp dụng 2 loại biểu đồ cho các đặc trưng số học là do em muốn quan sát rõ hơn về các giá trị ngoại lai của dữ liệu bằng box-plot, thứ mà kde-plot dù dễ hiểu hơn nhưng lại không thể hiện được.



Biểu đồ 1: Phân phối dữ liệu của các đặc trưng phân loại bằng count-plot.



Biểu đồ 2: Phân phối dữ liệu các đặc trưng số học bằng kde-plot.



Biểu đồ 3: Phân phối dữ liệu các đặc trưng số học bằng box-plot.

Ta có thể rút ra một số kết luận sau dựa trên việc quan sát phân phối của dữ liệu:

Categorical features:

- Các đặc trưng categorical có phân phối không đông đều.
- Có sự tương đồng trong phân phối cũng như ý nghĩa của 2 nhóm đặc trưng:
    - Kidhome, Teenhome.
    - AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Response.

Numerical features:

- Hầu hết có phân phối lệch trái nặng.
- Rất nhiều giá trị ngoại lai có thể nhìn thấy từ box-plot.
- Năm sinh (Year_Birth) nhỏ nhất là 1893 rất đáng nghi ngờ.

3. Khai phá dữ liệu

Thông qua các bước thấu hiểu dữ liệu, ta đã nắm được nhiều chi tiết quan trọng về dataset. Khai phá dữ liệu chính là quy trình tiếp theo nhằm tận dụng những tri thức đã có được mà làm giàu hơn giá trị của bộ dữ liệu.

3.1. Xử lý trùng lặp

Các giá trị trùng lặp không mang lại lợi ích cho giải thuật gom nhóm. Ngược lại nếu số lượng trùng lặp quá cao có thể làm sai lệch đối với một số mô hình gom cụm dựa vào khoảng cách như k-means. Vậy nên, trước khi đi xa hơn, ta nên loại bỏ những instance trùng lặp không mong muốn.

Sau khi loại bỏ trùng lặp, bộ dữ liệu của em còn lại 2058 instances.

3.2. Xử lý dữ liệu thiếu

Xử lý dữ liệu bị thiếu là một bước quan trọng khi thực hiện bài toán gom cụm (clustering). Dữ liệu bị thiếu có thể gây ra ảnh hưởng lớn đến kết quả của thuật toán gom cụm và dẫn đến các hiểu lầm hoặc biểu diễn không chính xác về cấu trúc thực sự của dữ liệu. Dưới đây là một số lý do chính:

- *Ảnh hưởng đến Tính Tương Đồng*: Dữ liệu bị thiếu có thể tạo ra độ chệch trong tính tương đồng giữa các điểm dữ liệu. Nếu một thuộc tính quan trọng bị thiếu, điều này có thể làm giảm khả năng đo lường sự tương đồng giữa các quan sát.
- *Ảnh Hưởng đến Tính Chất Topologique*: Các thuật toán gom cụm thường dựa vào các đặc trưng và cấu trúc topologique của dữ liệu để phân chia thành các cụm. Dữ liệu bị thiếu có thể làm suy giảm tính chất này và dẫn đến việc hình thành cụm không chính xác.
- *Nâng Cao Độ Chính Xác*: Xử lý dữ liệu bị thiếu giúp nâng cao độ chính xác của mô hình gom cụm. Khi một lượng lớn dữ liệu bị mất, khả năng phân loại đúng và mô hình hóa cụm sẽ giảm.
- *Tránh Nhiễu Dữ Liệu*: Dữ liệu bị thiếu có thể được coi như một dạng của nhiễu, và nếu không được xử lý, nó có thể làm suy giảm khả năng phân biệt giữa các cụm và tăng độ phức tạp của mô hình.
- *Đảm Bảo Tính Công Bằng*: Nếu dữ liệu bị thiếu không được xử lý, có thể dẫn đến việc mất mát thông tin quan trọng và tạo ra các đặc trưng không cân bằng trong quá trình gom cụm.

Có nhiều phương pháp để xử lý dữ liệu thiếu như loại bỏ instances/đặc trưng khiếm khuyết hoặc điền giá trị (Imputation).

Xét trên bộ dữ liệu của mình, chỉ duy nhất đặc trưng **Income** là thiếu 24 cell. Nhận thấy nó chỉ chiếm 1.17% tổng số 2058 instances của dữ liệu, em quyết định loại bỏ chúng.

3.3. Xử lý ngoại lai

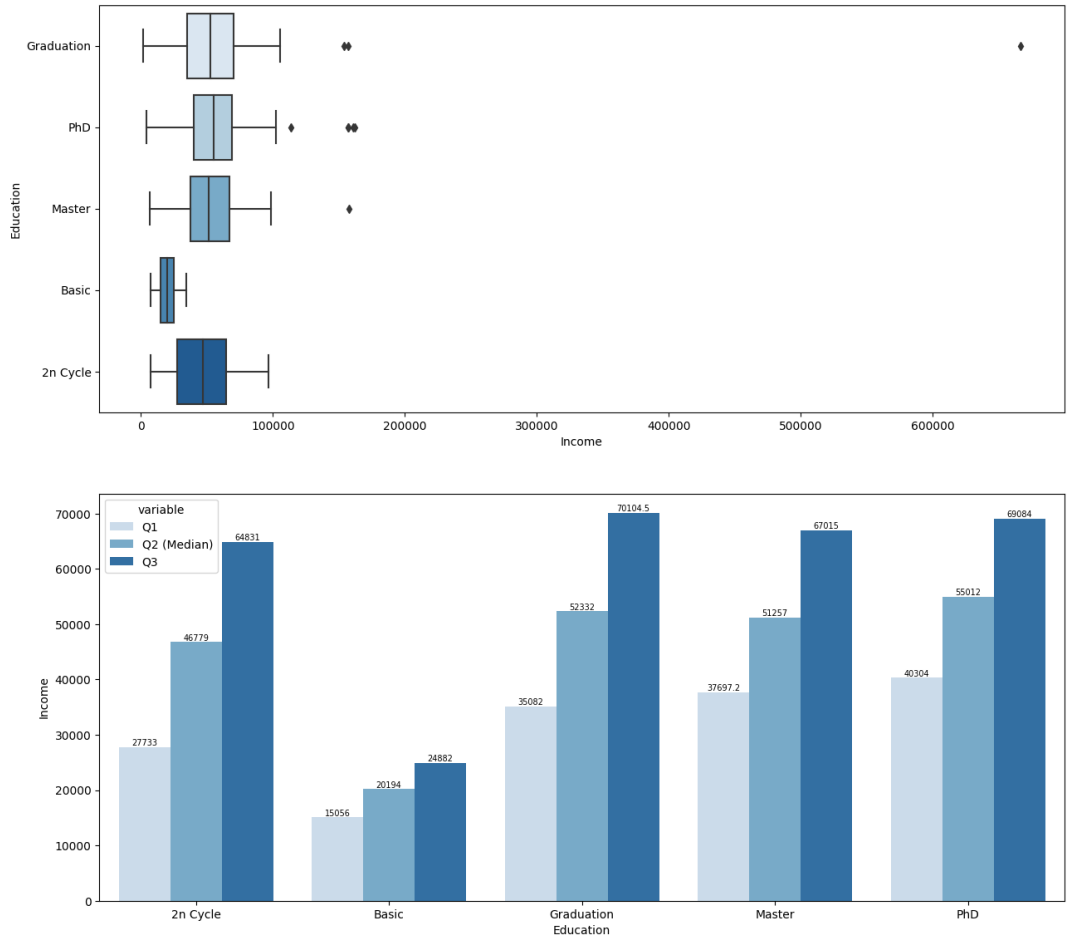
Ngoại lai (outlier) là một điểm dữ liệu hoặc một nhóm điểm dữ liệu có giá trị cách biệt lớn so với phần còn lại của tập dữ liệu. Các điểm ngoại lai có thể làm biến động và ảnh hưởng đáng kể đến kết quả của phân tích dữ liệu, bao gồm cả bài toán gom cụm mà thường là sẽ làm giảm hiểu suất và độ chính xác. Vậy nên, xử lý ngoại lai là một công việc rất quan trọng trong mọi bước khai phá, xử lý dữ liệu.

a) Income

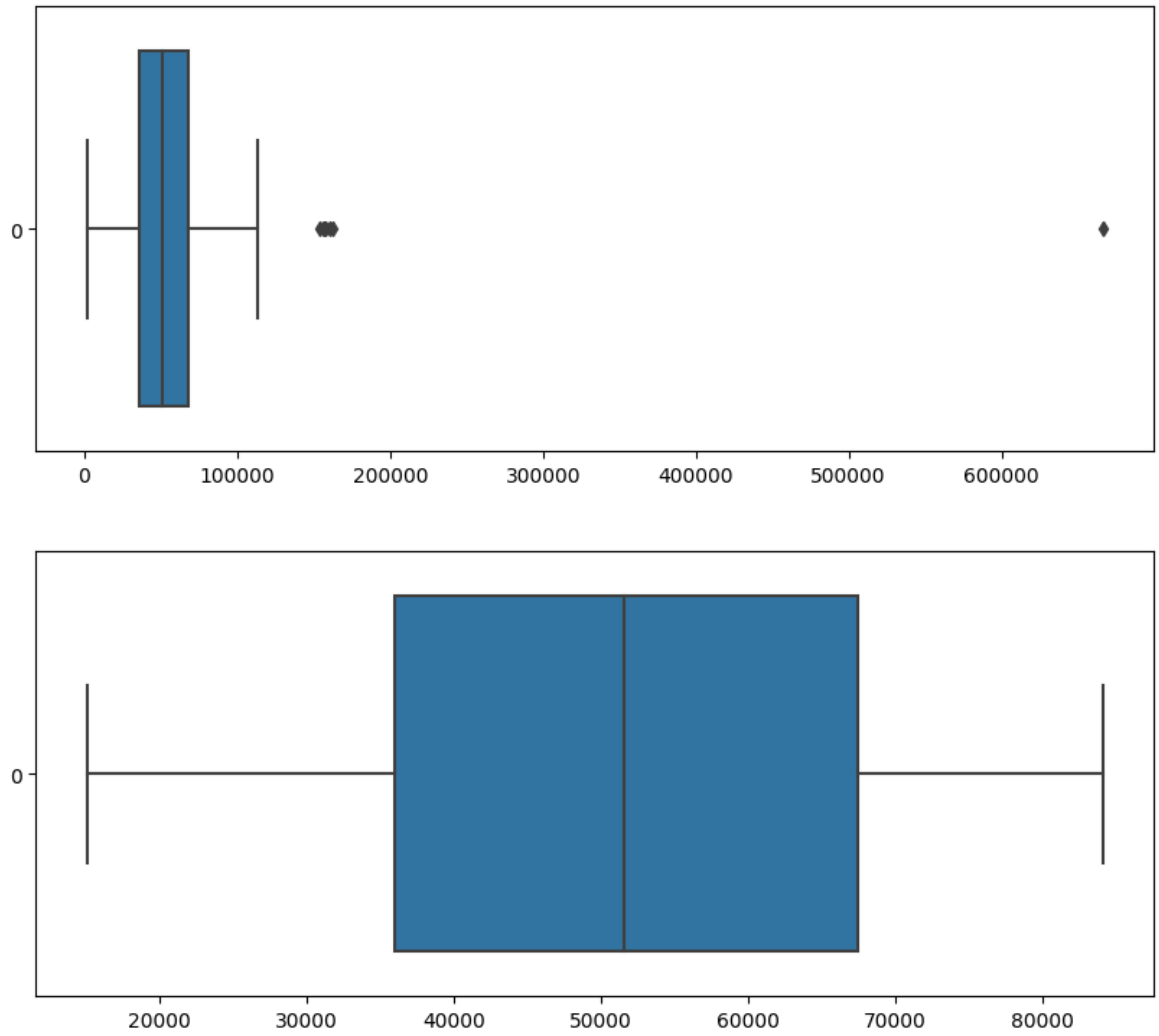
Dữ liệu về thu nhập của người dùng có độ lệch vô cùng lớn. Điều này có thể là hiển nhiên vì số người có thu nhập cao thường chiếm rất ít. Tuy nhiên, những instances với Income quá thấp cũng là đáng ngờ.

Để xử lý các ngoại lai trên đặc trưng này, em trực quan hóa mối quan hệ giữa thu nhập của khách hàng với học vị (**Education**) của họ. Sau đó, giới hạn thu nhập của khách hàng dựa vào học vị.

Cụ thể, giới hạn sẽ là từ *5th-percentile* đến *95th-percentile* của đặc trưng Income. Những khách hàng có thu nhập dưới mức này sẽ được gán giá trị thu nhập mới bằng với *25th-percentile* thu nhập của những người có cùng học vị, tương tự đối với những instances nằm trên giới hạn sẽ là *75th-percentile.*



Biểu đồ 4: Mối quan hệ và quantile của khách hàng dựa trên học vị.



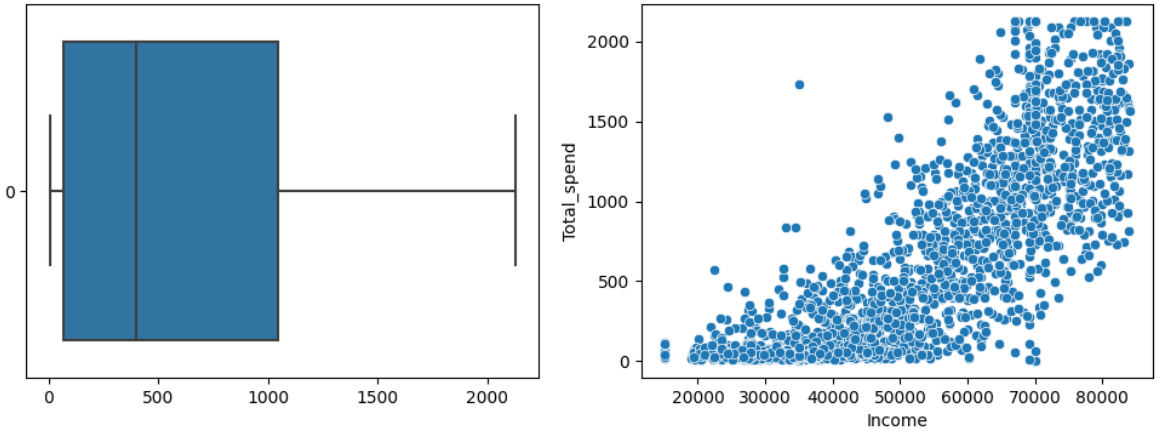
Biểu đồ 5: Phân phối của Income trước và sau khi xử lý ngoại lai.

b) Các đặc trưng về số lượng sản phẩm tiêu thụ

['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']

Những đặc trưng trình bày về số lượng sản phẩm mà khách hàng đã mua có phân phối tương tự nhau. Để đơn giản hóa bài toán, ta có thể gom chúng lại thành một đặc trưng duy nhất biểu diễn số lượng sản phẩm mà khách hàng đã mua, đặt tên là ***Total_spend.***

Sau khi gộp lại và giới hạn một điểm ngoại lệ về 99th-percentile, em đã thu được một đặc trưng mới có sự biến thiên khá thuận chiều với Income.



Biểu đồ 6: Phân phối của Total_spend và mối tương quan với Income.

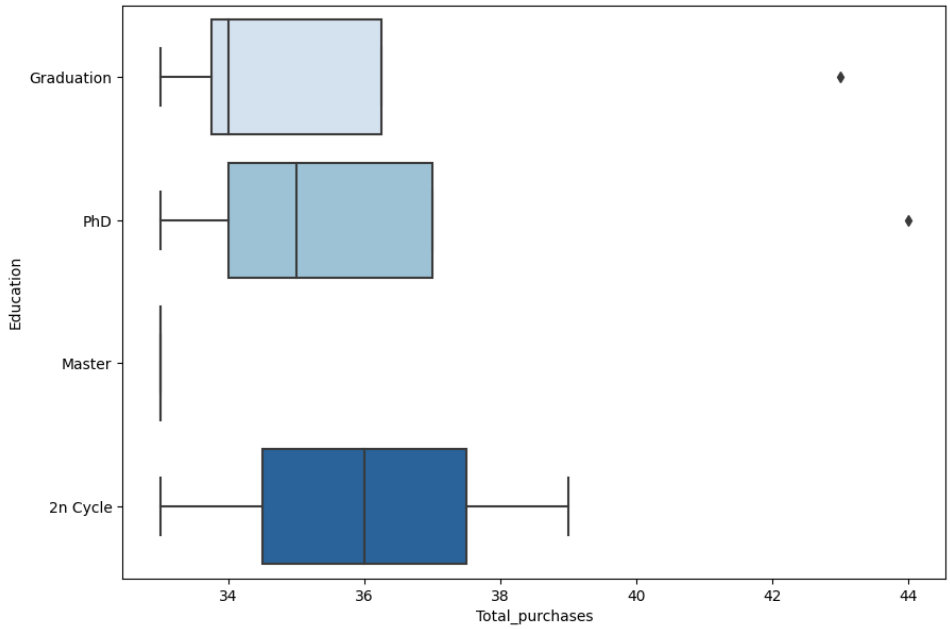
c) Các đặc trưng về số lượt mua hàng đã thực hiện.

['NumDealsPurchases', 'NumWebPurchases',
'NumCatalogPurchases', 'NumStorePurchases']

Những đặt trưng liên quan đến số lượt mua hàng cũng có thể được gom lại để thu nhỏ bài toán. Tuy nhiên, đặc trưng thể hiện tổng số lượt mua hàng ***Total_purchases*** có nhiều ngoại lệ hơn là trường hợp vừa rồi.

Vậy nên, em áp dụng phương pháp xử lý ngoại lai tương tự với trường hợp của Income sau khi quan sát phân phối dữ liệu của đặc trưng ***Total_purchases*** dựa trên **Education.**

Cụ thể, giới hạn điểm dữ liệu về đoạn [5th-percentile, 99th-percentile]. Các giá trị nằm dưới giới hạn này sẽ được thay thế bằng 5th-percentile, các điểm ở trên giới hạn sẽ được thay bằng *median* của các khách hàng có chung học vị.

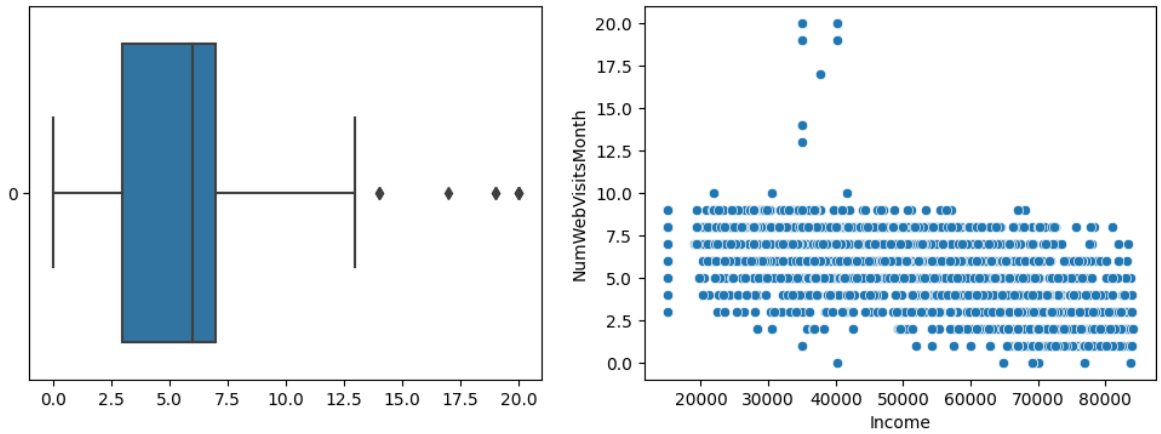


Biểu đồ 7: Phân phối dữ liệu của Total_purchases dựa trên Education.

d) NumWebVisitsMonth

Đặc trưng NumWebVisitsMonth có một số giá trị ngoại lai cần xử lý. Bằng việc biểu diễn mối tương quan giữa nó với thu nhập của khách hàng, em nhận xét rằng các giá trị ngoại lai nên được thay thế bằng *trung vị (median)* của đặc trưng.

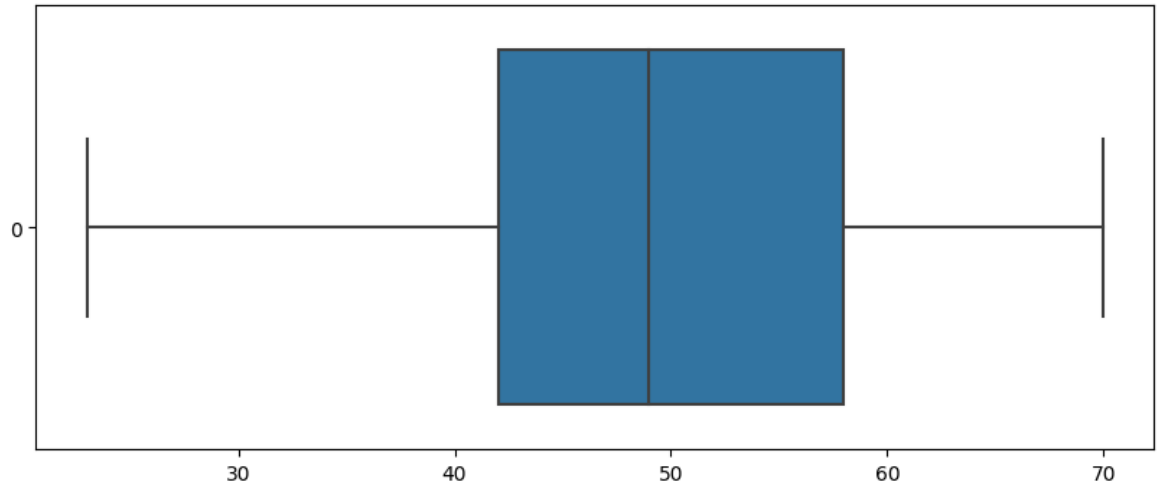
Median của NumWebVisitsMonth là 6.



Biểu đồ 8: Phân phối của NumWebVisitsMonth và mối tương quan với Income.

e) Year_Birth

Bộ dữ liệu được cập nhật lần cuối là từ năm **2019**, cho rằng đây là căn cứ để xác định độ tuổi của khác hàng thì những khách hàng sinh trước năm **1949** đã trên **70** tuổi. Đây có lẽ là một ngoại lai tự nhiên hoặc nhầm lẫn trong quá trình nhập liệu. Em lựa chọn phương án thay thế những giá trị nằm dưới ngưỡng **1949** bằng *median* của ***Year_Birth*** đồng thời chuyển đổi đặc trưng này thành tuổi tác: ***Age*** để dễ dàng hơn cho việc trình bày, diễn đạt ý nghĩa sau này.



Biểu đồ 9: Phân phối dữ liệu của đặc trưng Age.

4. Tiền xử lý dữ liệu

4.1. Mã hóa đặc trưng

Vì máy tính không thể hiểu được các chữ cái, ta cần mã hóa các đặc trưng phân loại về dạng số để đưa vào mô hình gom cụm.

a) Education

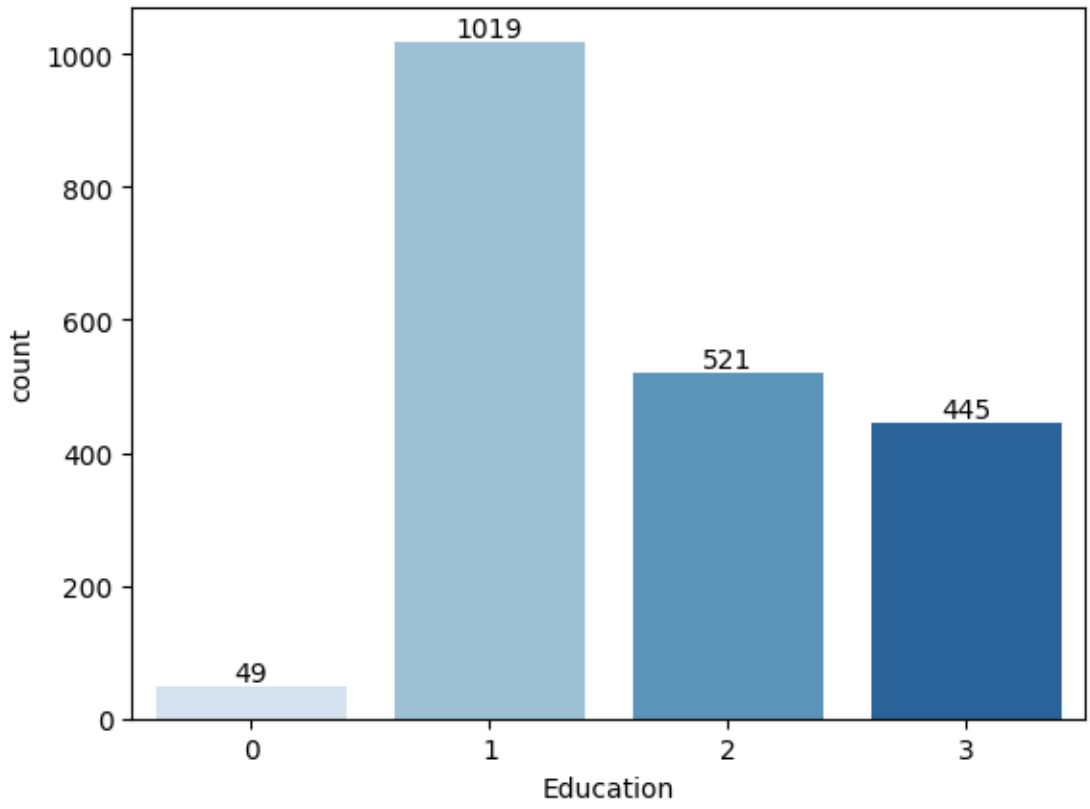
Các học vị có thể mã hóa kiểu thứ tự (ordinal-encoding) vì có sự khác biệt về cấp bậc giữa chúng. Cụ thể giá trị mã hóa sẽ như sau:

0 : ‘Basic’

1 : ‘Graduation’

2 : ‘Master’ và ‘2n Cycle’ (hai học vị này có giá trị tương đương nhau)

3 : ‘PhD’



Biểu đồ 10: Phân phối của Education sau mã hóa.

b) Marital_Status

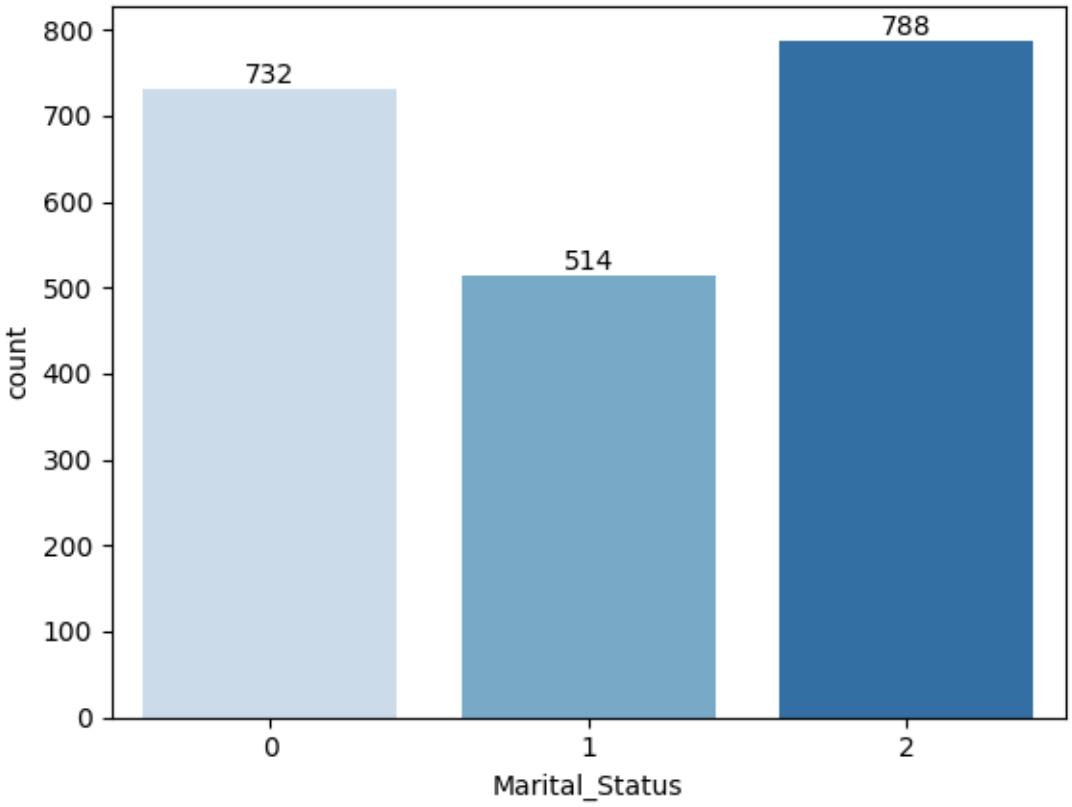
Tình trạng hôn nhân có thể được rút gọn lại theo từng mức độ:

0 : Đang một mình.

1 : Đang trong một mối quan hệ.

2 : Đã kết hôn.

Cụ thể thì các giá trị trong đặc trưng trên sẽ được mã hóa theo một dictionary như sau: *{'Alone' : **0**, 'YOLO' : **0**, 'Absurd' : **0**, 'Widow' : **0**, 'Divorced' : **0**, 'Single' : **0**, 'Together' : **1**, 'Married' : **2**}*



Biểu đồ 11: Phân phối của Marital_Status sau mã hóa.

c) Các đặc trưng AcceptedCampaign

['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'Response']

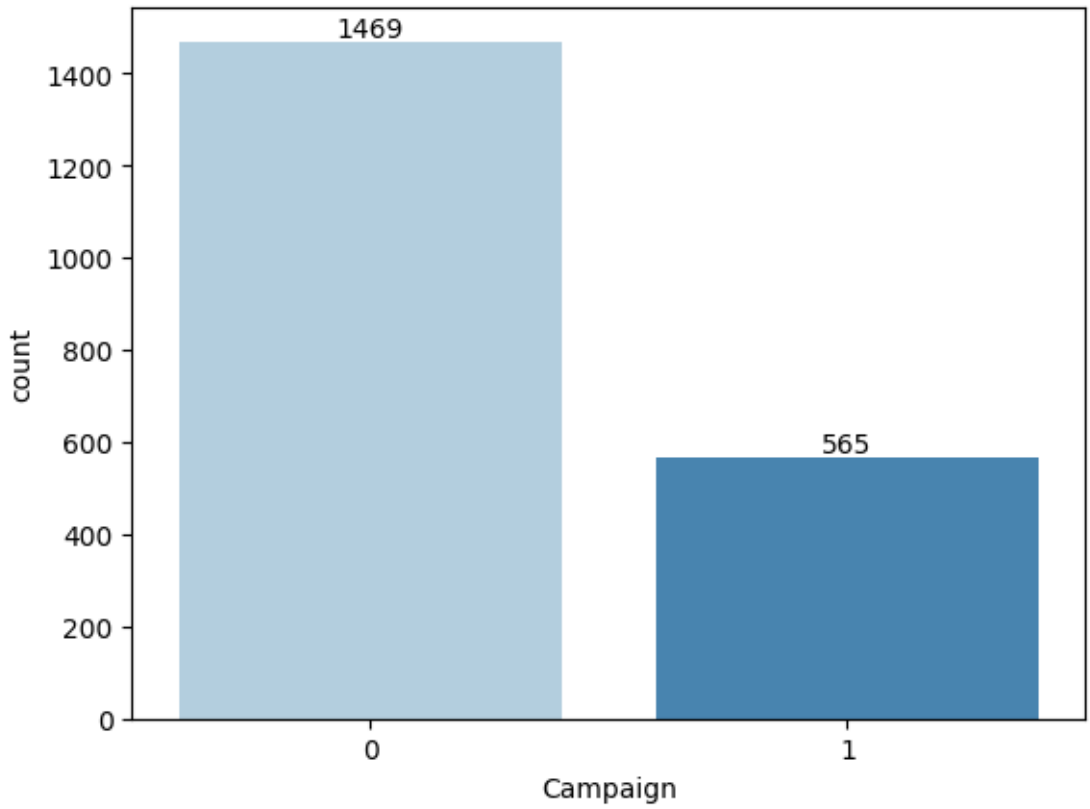
Ta có thể rút gọn các đặc trưng trên về một đặc trưng duy nhất vì chúng có ý nghĩa tương tự nhau. ***Campaign*** sẽ là đặc trưng mới đánh dấu cho ta biết khách hàng đã bao giờ mua mặt hàng nào trong các chiến dịch giảm giá trước đây của công ty không. Trong đó:

0 : Chưa từng

1 : Đã từng

	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	Response	Campaign
0	0	0	0	0	0	1	1
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	1	1
9	0	0	1	0	0	0	1

Bảng 5: Áp dụng toán tử OR để xây dựng đặc trưng Campaign.



Biểu đồ 12: Phân phối dữ liệu của đặc trưng Campaign.

d) Complain

Đây là một đặc trưng có phân phối siêu lệch (2014 | 20) đến mức không thể nhận ra số lượng instance có giá trị 1 trên countplot mà ta đã phát họa ở mục 2.4.

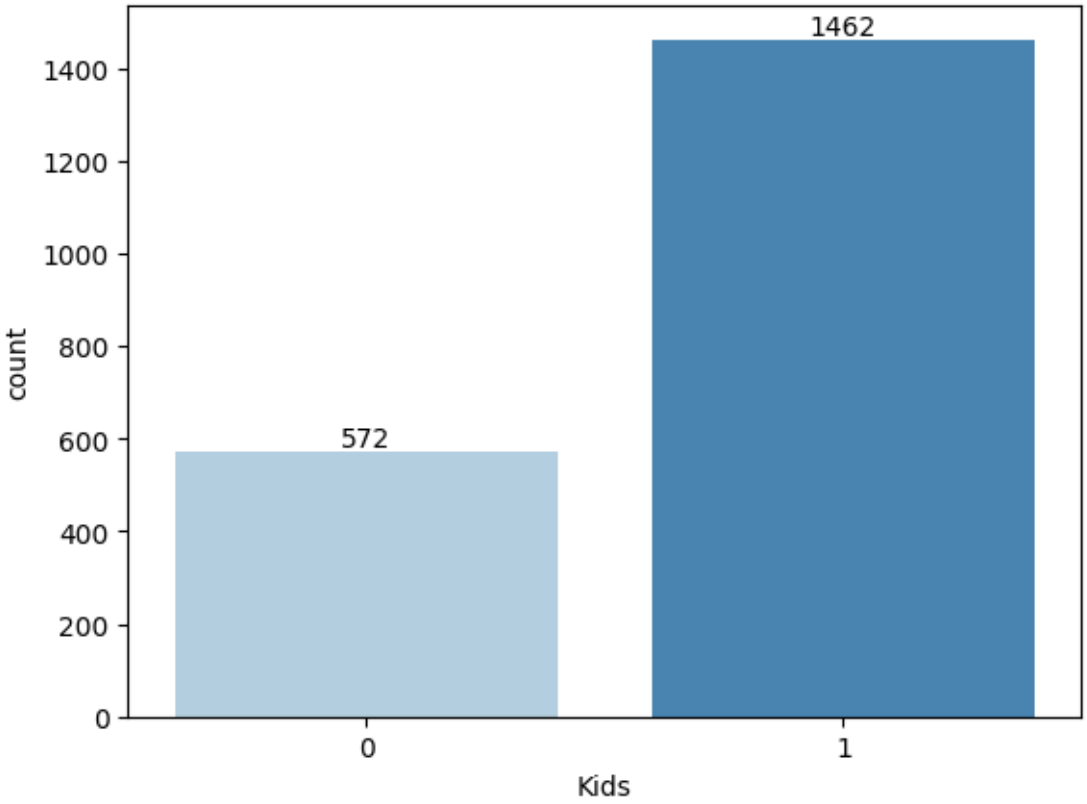
→Trong trường hợp này, em lựa chọn loại bỏ đặc trưng ***Complain*** vì nó có thể là một đặc trưng nhiễu trong bộ dữ liệu.

e) Các đặc trưng về con cái

Gộp những đặc trưng như ***'Kidhome', 'Teenhome'*** sẽ giúp giảm thiểu số chiều của dữ liệu vì chúng có ý nghĩa tương tự nhau. Đặc trưng mới, ***Kids*** sẽ biểu diễn cho ta biết khách hàng có con hay không.

0 : Không có con

1 : Có con



Biểu đồ 13: Phân phối dữ liệu của đặc trưng Kids.

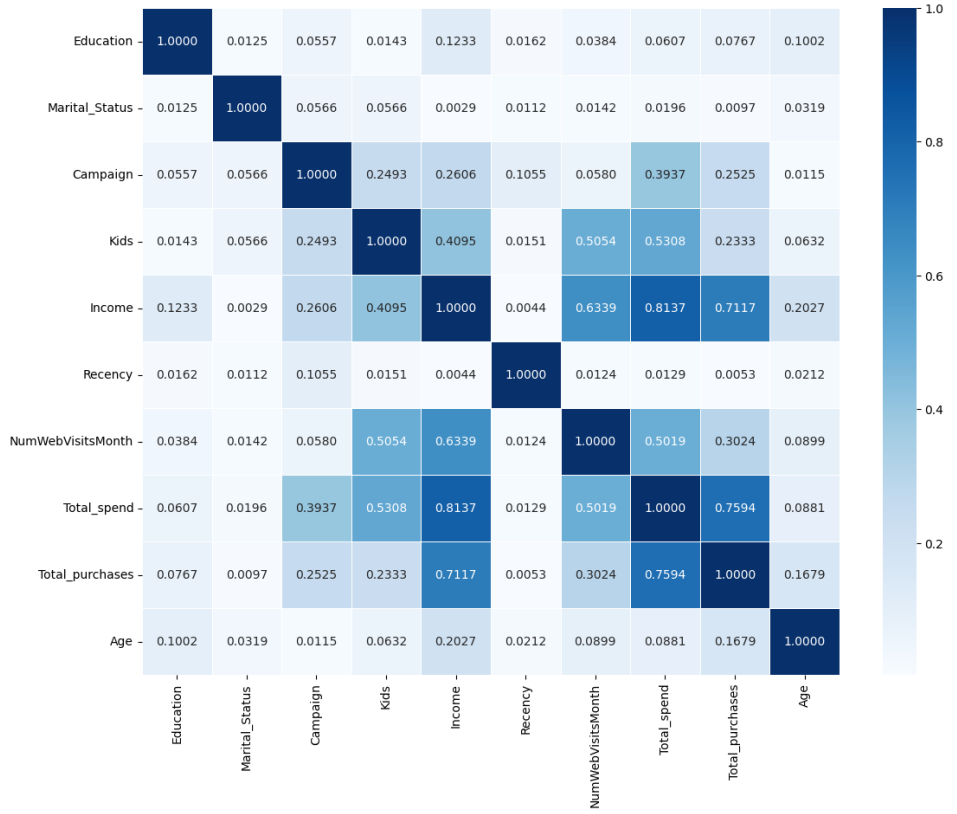
4.2. Trích xuất đặc trưng

Trích xuất đặc trưng (Feature Extraction) là quá trình chọn lọc hoặc biến đổi dữ liệu đầu vào thành một tập hợp nhỏ các đặc trưng quan trọng nhất để tạo ra mô hình hoặc phân tích dữ liệu. Mục tiêu của trích xuất đặc trưng là giảm kích thước (chiều) của dữ liệu và tăng cường khả năng hiểu và mô hình hóa.

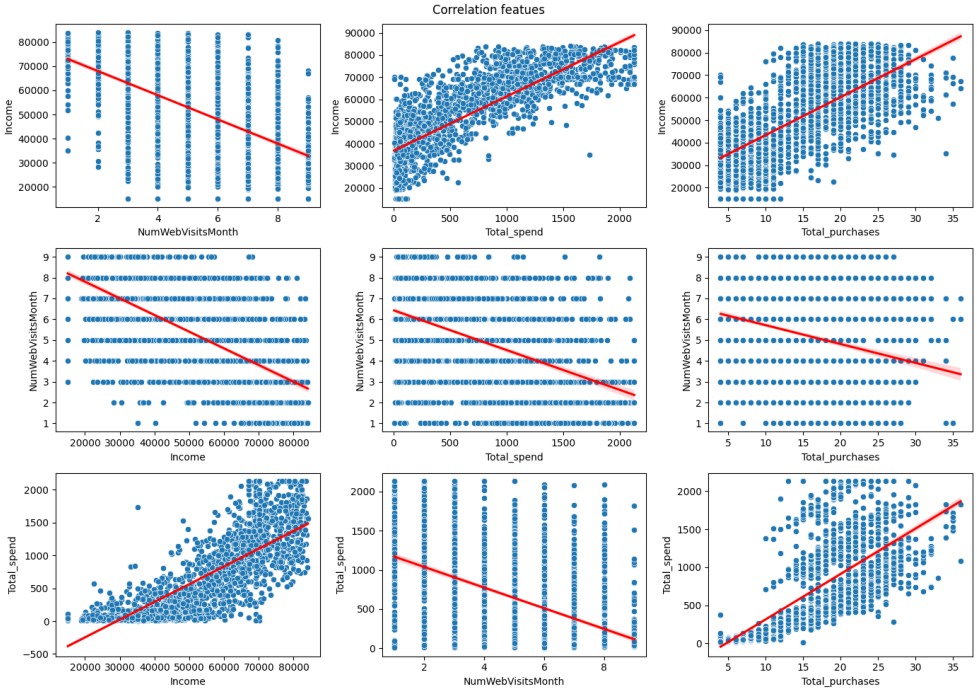
Ma trận tương quan

Ma trận tương quan biểu diễn mối tương quan của từng cặp đặc trưng khi giá trị của chúng biến thiên, đây chính là cơ sở cho phép ta thực hiện các thao tác trích xuất đặc trưng nhằm nén thông tin và giảm chiều dữ liệu. Thông qua ma trận tương quan em phát hiện 4 đặc trưng có tính tương quan đáng kể với nhau là:

['Income', 'NumWebVisitsMonth', 'Total_spend', 'Total_purchases']



Biểu đồ 14: Ma trận tương quan (lấy trị tuyệt đối) của dữ liệu.



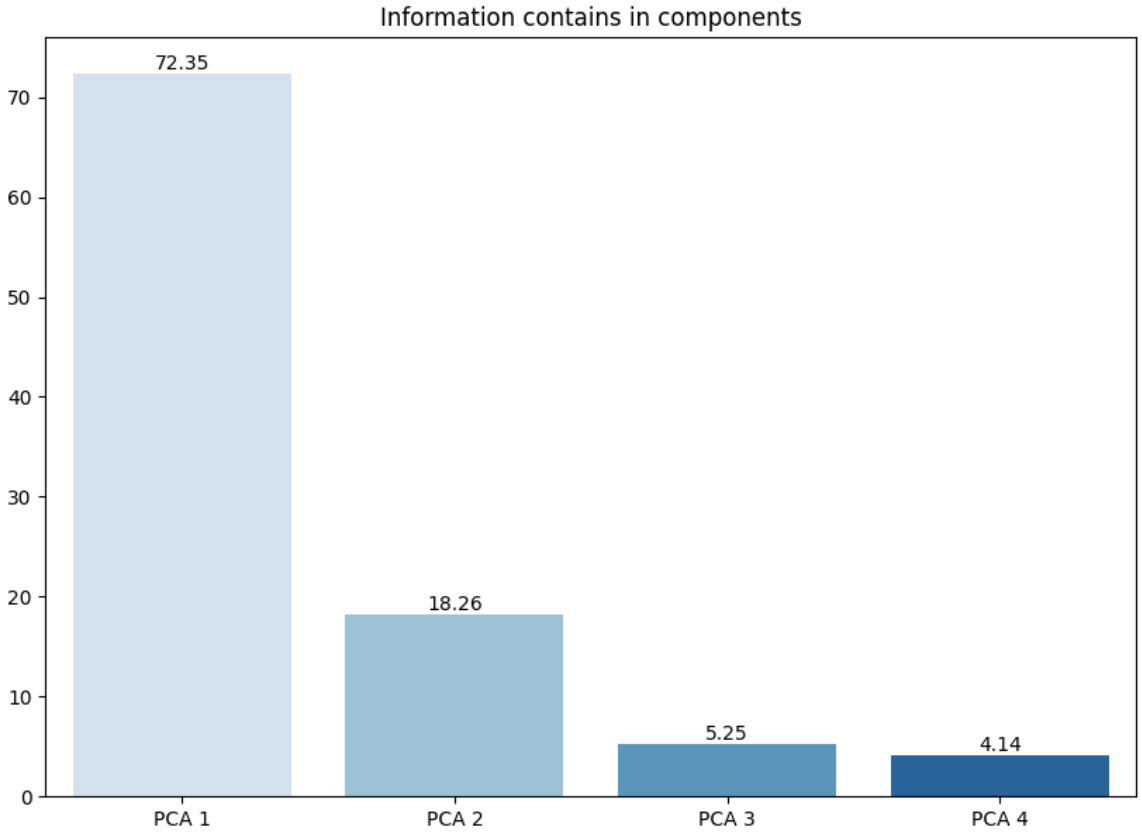
Biểu đồ 15: Biểu diễn sự biến thiên giữa các cặp đặc trưng tương quan cao.

Trích xuất đặc trưng với PCA.

Cách đơn giản nhất để giảm chiều dữ liệu từ D về K<Dlà chỉ giữ lại K phần tử *quan trọng nhất*. Tuy nhiên, việc làm này chắc chắn chưa phải tốt nhất vì chúng ta chưa biết xác định thành phần nào là quan trọng hơn. Hoặc trong trường hợp xấu nhất, lượng thông tin mà mỗi thành phần mang là như nhau, bỏ đi thành phần nào cũng dẫn đến việc mất một lượng thông tin lớn.

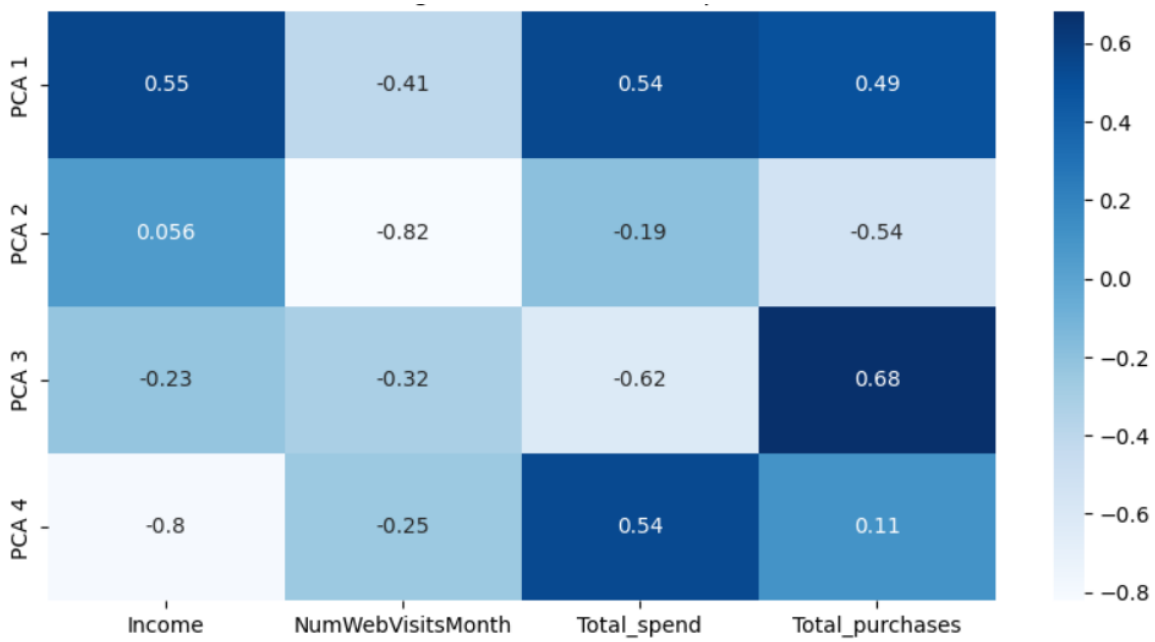
PCA chính là phương pháp đi tìm một hệ cơ sở mới sao cho thông tin của dữ liệu chủ yếu tập trung ở một vài toạ độ, phần còn lại chỉ mang một lượng nhỏ thông tin. Và để cho đơn giản trong tính toán, PCA sẽ tìm một hệ trực chuẩn để làm cơ sở mới [2].

Em sẽ áp dụng PCA cho các đặc trưng tương quan tốt nêu trên. Một điều cần lưu ý khi áp dụng kỹ thuật này là ta cần chuẩn hóa dữ liệu trước.



Biểu đồ 16: Lượng thông tin giữ lại từ các thành phần chính.

Như ta thấy, chỉ cần 2 thành phần chính PCA 1 và PCA 2, ta đã có hơn 90% thông tin từ 4 đặc trưng trước đó. Sử dụng chúng thay cho những đặc trưng cũ có thể giúp tăng cường hóa mô hình nhưng trước tiên ta cần phải xem xét ý nghĩa của chúng. Vậy nên, em sẽ giữ 2 thành phần chính này làm đặc trưng mới.

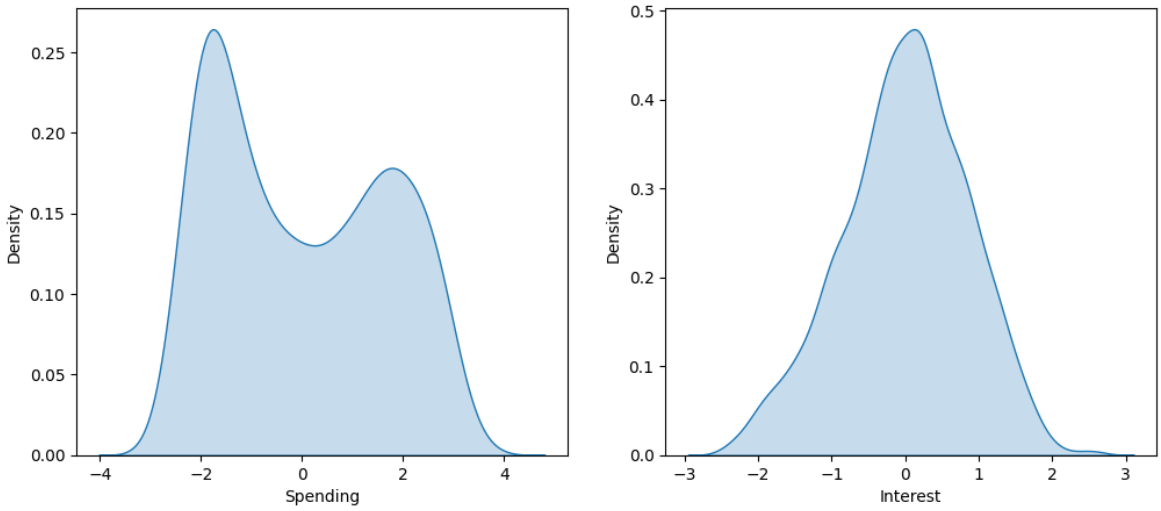


Biểu đồ 17: Trọng số của các đặc trưng tương ứng với những thành phần chính.

Thông qua phân tích biểu đồ trên, ta có thể nắm được ý nghĩa của các components. Cụ thể nếu xét đến PCA 1, PCA 2:

- PCA 1: Tỉ lệ thuận với *Income, Total_spend, Total_purchases* và tỉ lệ nghịch với *NumWebVisitsMonth* ở mức tương đối. → Thành phần này biểu diễn năng lực chi tiêu của khách hàng.
- PCA 2: Tỉ lệ nghịch mạnh mẽ với *NumWebVisitsMonth*, tỉ lệ thuận tương đối với *Total_purchases*. → Thành phần này là nghịch đảo của mức độ từ khách hàng đối với các sản phẩm.

→ Vậy ta có thể đặt tên cho những đặc trưng mới là *Spending* (PCA 1) và *Interest* (PCA 2).



Biểu đồ 18: Phân phối dữ liệu của 2 đặc trưng mới.

Vậy là sau các bước xử lý dữ liệu, ta thu được một bộ dữ liệu mới với số chiều (số đặc trưng) giảm xuống còn 8.

	Education	Marital_Status	Campaign	Kids	Recency	Age	Spending	Interest
0	1	0	1	0	0.315957	1.169405	1.454539	-1.634508
1	1	0	0	1	-0.374256	1.445682	-1.207527	0.891598
2	1	1	0	0	-0.788384	0.432663	1.384301	0.036669
3	1	1	0	1	-0.788384	-1.317097	-1.836136	0.320501
4	3	2	0	1	1.558341	-1.040819	0.350436	-0.112192

Bảng 6: Tổng quan cơ sở dữ liệu sau khi xử lý.

5. Gom nhóm khách hàng

5.1. Giải thuật gom cụm K-means

k-Means là một thuật toán rất đơn giản nhưng có rất nhiều ứng dụng trong thực tiễn. Một số ứng dụng của thuật toán này có thể kể đến như:

- Phân khúc khách hàng trong kinh doanh
- Phân tích gen trong y khoa
- Sử dụng trong các bài toán Image segmentation
- Nén hình ảnh.
- Phát hiện tế bào ung thư.
- Phát hiện bất thường (*anomaly detection*).

Trong thuật toán K-means chúng ta được cung cấp một tập dữ liệu đầu vào $\{x1,x2,\ldots,xn\}$, trong đó x là các instances và phân cụm chúng vào những nhóm dữ liệu có tính chất chung. Điểm đặc biệt của tập dữ liệu này là chúng hoàn toàn chưa được gán nhãn [3].

5.2. Xác định số cụm K tối ưu

Có nhiều phương pháp để xác định số lượng cụm tối ưu cho giải thuật K-means. Trong khuôn khổ đồ án này, em sử dụng 3 phương pháp đánh giá là *Elbow, Silhouette và Davis-Bouldin index.*

a) Phương pháp Elbow (Khuỷu tay):

Dựa trên biểu đồ Elbow, bạn thực hiện k-means với nhiều giá trị khác nhau của K và vẽ biểu đồ biểu diễn tổng bình phương khoảng cách (inertia hoặc distortion) theo K. Chọn giá trị K nào mà khi tăng giá trị K, giảm độ biến động (distortion) nhanh chóng và sau đó giảm chậm lại, tạo ra một "khuỷu tay" trên biểu đồ.

b) Phương pháp Silhouette:

Sử dụng phương pháp Silhouette để đánh giá chất lượng của cụm. Silhouette Score đo độ "tách biệt" giữa các cụm. Score cao nhất cho biết sự tách biệt giữa các cụm là tốt nhất. Chọn K có Silhouette Score lớn nhất.

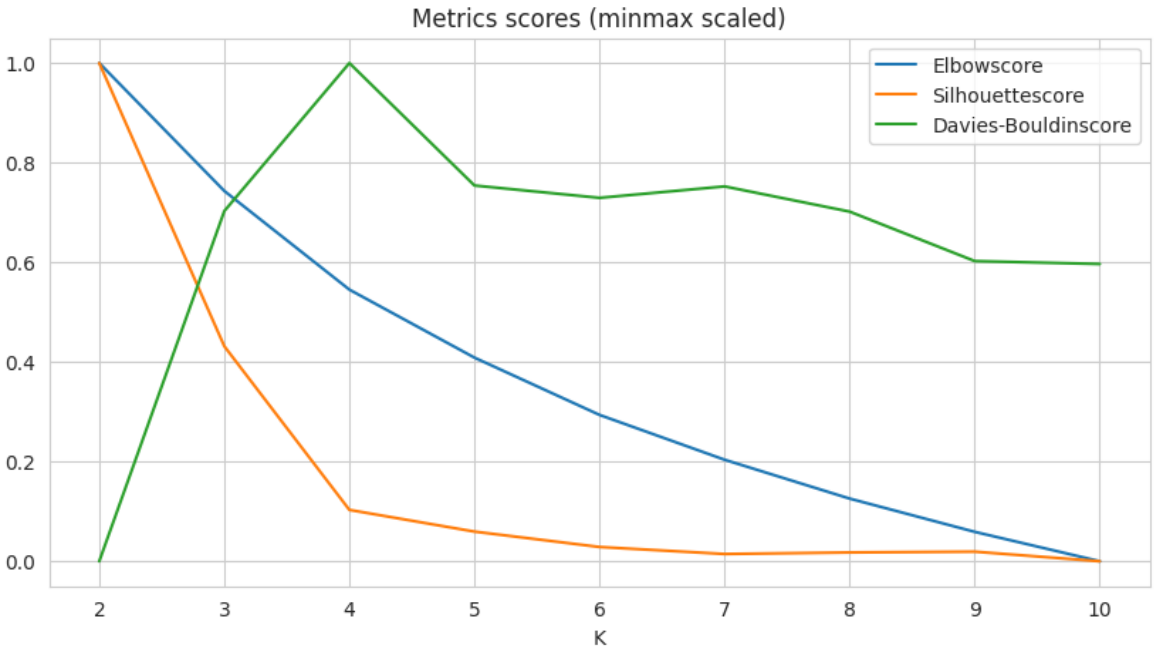
c) Phương pháp Davies-Bouldin Index:

Đo lường sự tách biệt giữa các cụm và đồng thời đảm bảo sự giảm thiểu sự chồng chéo giữa các cụm. Chọn K có Davies-Bouldin Index thấp nhất.

Qua thực nghiệm các giá trị K từ 2 đến 10, em xác định được số lượng cụm tốt nhất chính là K= 2.

K	Elbow	Silhouette	Davies-Bouldin
2	10362.763717	0.280327	1.430179
3	9351.339876	0.198834	1.890474
4	8571.190694	0.151748	2.084996
5	8033.529715	0.145503	1.924054
6	7580.573624	0.141076	1.907834
7	7226.049847	0.139060	1.922856
8	6918.818535	0.139502	1.889734
9	6655.894322	0.139727	1.824730
10	6423.380432	0.136970	1.820903

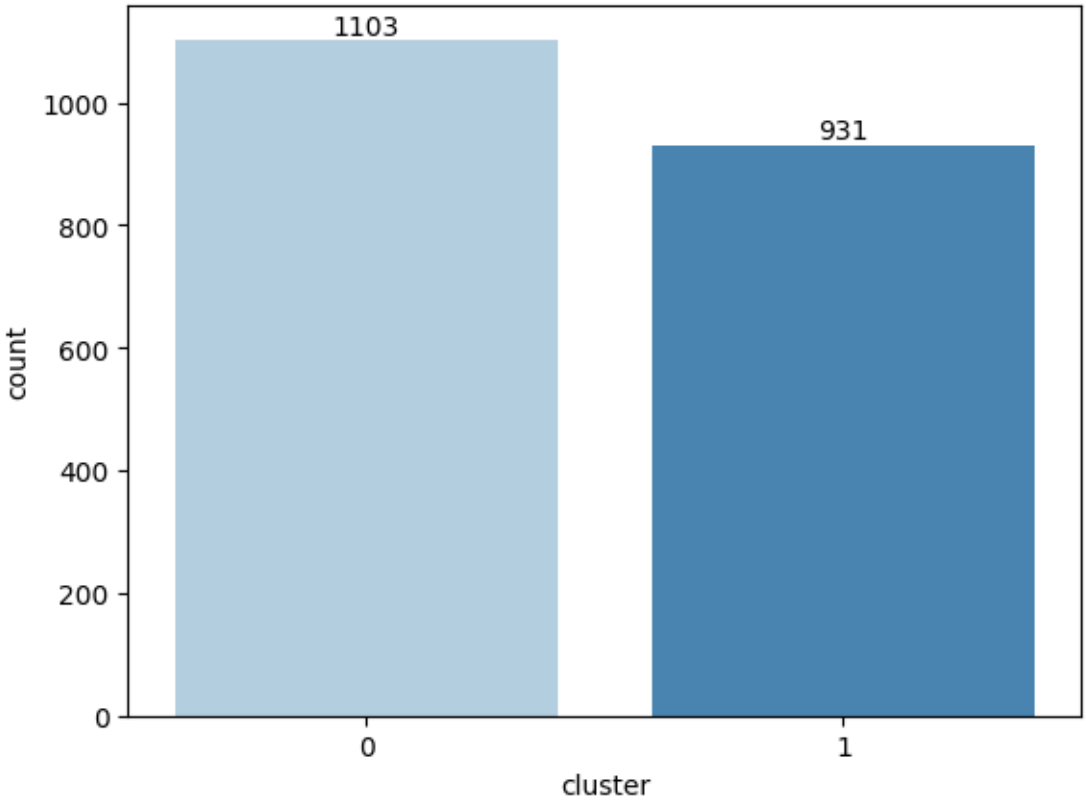
Bảng 7: Giá trị của những chỉ số đánh giá dựa trên K tương ứng.



Biểu đồ 19: Biến động của các chỉ số (đã chuẩn hóa về chung một giới hạn).

6. Biểu diễn mẫu

6.1. Phân phối của cụm



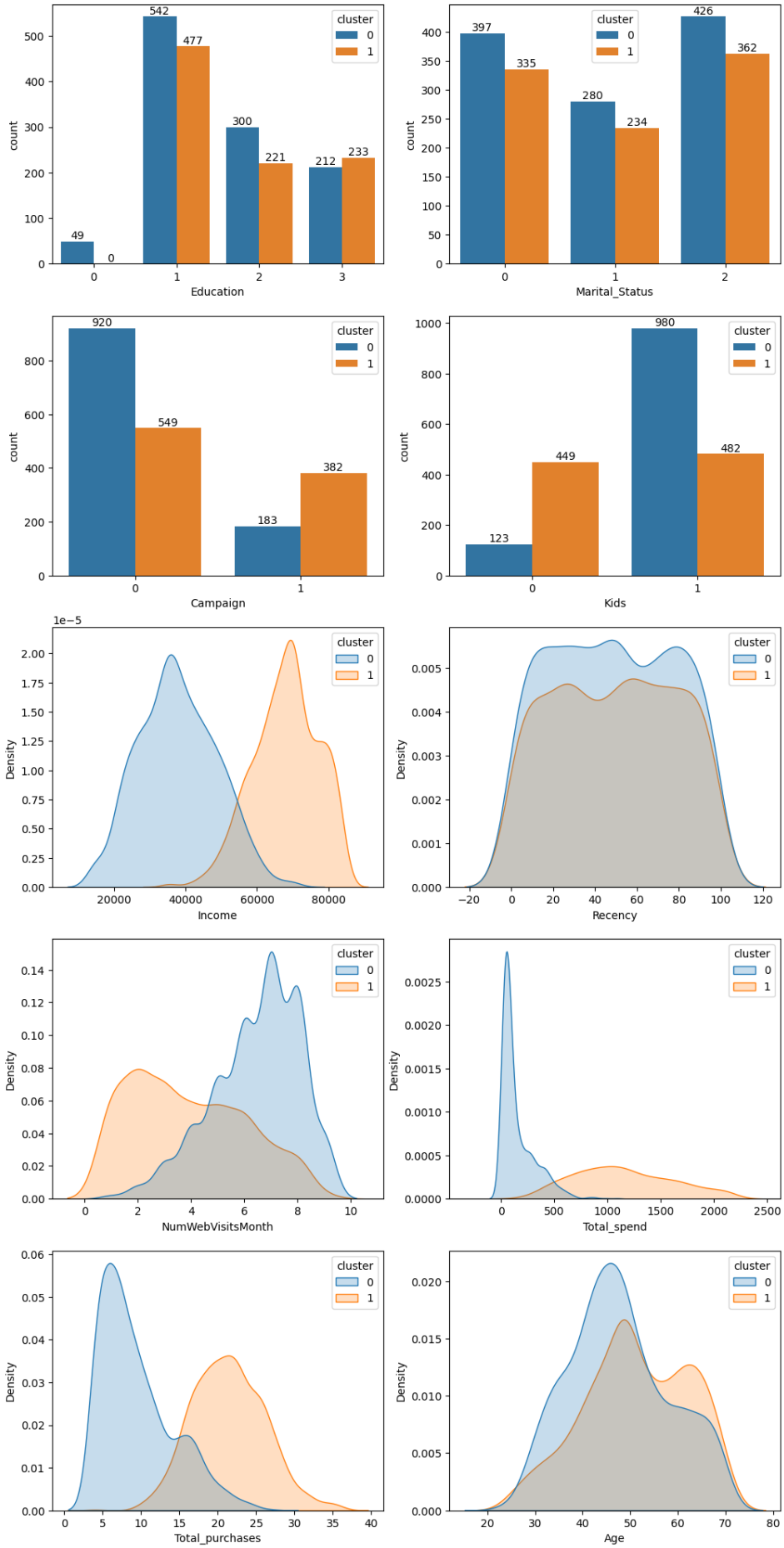
Biểu đồ 20: Phân phối dữ liệu trong các nhóm khách hàng.

Ta thấy, mô hình đã gom các khách hành thành 2 nhóm khác nhau với sự chệnh lệch không quá đáng kể. Để biết thêm về ý nghĩa của từng nhóm khách hàng, ta cần quan sát phân phối dữ liệu của từng đặc trưng khác.

6.2. Phân phối các đặc trưng.

Sau khi đã có nhãn của các cụm. Em sẽ tiến hành gán nhãn lại vào bộ dữ liệu trước khi trích xuất đặc trưng để phân tích của những đặc điểm của các nhóm khách hàng mà mô hình vừa tạo ra. Cụ thể, bộ dữ liệu lúc này có tổng cộng 10 đặc trưng (không kể nhãn của cụm) là:

['Education', 'Marital_Status', 'Campaign', 'Kids', 'Income', 'Recency', 'NumWeb VisitsMonth', 'Total_spend', 'Total_purchases', 'Age']



Biểu đồ 21: Phân phối dữ liệu trên từng đặc trưng của các cụm.

Quan sát sự phân bố dữ liệu, ta có thể rút ra một số nhận định:

- *Marital_Status* và *Recency* không có sự khác biệt đặc sắc nào giữa 2 cụm, có thể nói hai đặc trưng này không đóng góp nhiều cho quá trình gom nhóm khách hàng của mô hình.
- Học vị (*Education*) của 2 nhóm khách hàng không có nhiều khác biệt. Trừ trường hợp học vị thấp nhất (Basic) nằm hoàn toàn ở cụm 0.
- Nhóm khách hàng số 0 có phần lớn (khoảng 84% của nhóm) là những người chưa từng hưởng ứng một chiến dịch giảm giá nào của công ty. Trong khi sự khác biệt trên *Campaign* của nhóm còn lại là không đáng kể.
- Tương tự trên đặc trưng *Kids*, khi cụm 1 không chênh lệch nhiều thì tỉ lệ khách hàng có con trên cụm 0 lại gấp gần 8 lần so với những khách hàng không có con.
- *Income* là đặc trưng có sự tách biệt rõ ràng nhất về phân phối dữ liệu trong số các đặc trưng số học. Ta dễ thấy khách hàng thuộc cụm 0 có thu nhập ít hơn so với cụm 1.
- Những khách hàng được gán nhãn 0 cũng thường mua ít sản phẩn hơn (Total_spend) và thực hiện ít giao dịch hơn (Total_purchases) so với nhóm còn lại. Nhưng họ có xu hướng ghé qua website của công ty nhiều hơn (NumWebVisitsMonth).
- Đặc trưng *Age* không phản ánh nhiều về đặc điểm của từng cụm nhưng ta vẫn nhận thấy rằng nhóm khách hàng số 1 có tuổi thọ trung bình cao hơn một chút so với nhóm 0.

6.3. Giải thích mẫu

Dựa trên những nhận xét rút ra ở trên, một định nghĩa cho các nhóm khách hàng đã có thể được xây dựng như sau:

Cụm 0: Là những khách hàng có khả năng chi tiêu tiết kiệm.

Những người có trình độ học vấn ở mức cơ bản được xếp vào nhóm này. Phần nhiều khách hàng trong nhóm đã có con cái, thu nhập ở mức trung bình-

khá trở xuống. Không nhiều người sẵn sàng mua sắm trong cả những dịp giảm giá nên ít phát sinh giao dịch và mua ít sản phẩm hơn. Tuy nhiên, họ cũng là nhóm người hứng thú khá nhiều với các mặt hàng khi thường xuyên truy cập vào website của công ty hơn.

Cụm 1: Là những khách hàng sẵn sàng chi tiêu mạnh tay.

Nhóm người này sỡ hữu mức thu nhập trung bình cao. Mua sắm nhiều hơn, trung bình họ giao dịch nhiều gấp vài lần so với nhóm còn lại. Tổng sản phẩm mà một người trong đây người tiêu thụ có thể vượt xa so với con số lớn nhất ở cụm 0. Các khách hàng thuộc nhóm này phóng khoáng hơn khi trung bình không tốn quá nhiều lần vào website của công ty để tham khảo các sản phẩm dù về tuổi tác thì không cao hơn là mấy.

→ Xu hướng tiêu dùng dựa trên những đặc trưng của khách hàng đã được mô hình tự học hỏi và xây dựng làm 2 nhóm riêng biệt. Kết quả thu được là minh chứng cho sự mạnh mẽ của trí tuệ nhân tạo nói chung và học máy nói riêng trong thời đại công nghệ số hiện nay.

KẾT LUẬN

Trong quá trình thực hiện đồ án, em đã xây dựng và triển khai một quy trình phân tích dữ liệu khách hàng sử dụng phương pháp phân tích nhận dạng mẫu. Qua việc kết hợp các phương pháp trích xuất đặc trưng và thuật toán K-means clustering, chúng tôi đã thành công trong việc gom nhóm khách hàng dựa trên các đặc trưng quan trọng.

Mô hình đã xây dựng đã mang lại những cụm khách hàng có ý nghĩa và phản ánh đúng cấu trúc của dữ liệu. Quy trình phân tích này không chỉ cung cấp cái nhìn sâu sắc về hành vi của khách hàng mà còn giúp tối ưu hóa chiến lược tiếp thị và phục vụ khách hàng một cách hiệu quả.

Tuy nhiên, để phát triển đồ án trong tương lai, có một số hướng mà chúng tôi đề xuất. Đầu tiên, có thể cải thiện việc xử lý các đặc trưng bằng cách sử dụng các phương pháp tiền xử lý dữ liệu mạnh mẽ hơn để làm sạch và chuẩn hóa dữ liệu. Thêm vào đó, việc sử dụng các giải thuật gom nhóm phức tạp hơn, như Hierarchical Clustering hay DBSCAN, có thể mang lại những cụm có hình dạng và kích thước linh hoạt hơn, phản ánh đa dạng hóa hơn trong dữ liệu khách hàng.

Những cải tiến này sẽ giúp nâng cao hiệu suất và sức mạnh dự đoán của mô hình, đồng thời làm cho quy trình phân tích trở nên linh hoạt và thích ứng với sự biến động của dữ liệu khách hàng trong thời gian. Em hy vọng rằng đồ án này không chỉ mang lại những hiểu biết sâu sắc về khách hàng mà còn làm nền tảng cho những nghiên cứu và ứng dụng tiếp theo trong lĩnh vực phân tích dữ liệu và nhận dạng mẫu.

Cảm ơn sự quan tâm và hỗ trợ của giảng viên TS. Vũ Ngọc Thanh Sang trong quá trình thực hiện đồ án. Những lời tư vấn và tri thức của thầy là nguồn động lực vô cùng lớn giúp em hoàn thành nghiên cứu đồ án này.

___Hết___

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Richard O. Duda, Peter E. Hart, David G. Stork (2001) Pattern classification (2nd edition), Wiley, New York, ISBN 0-471-05669-3.

[2] Vũ Hữu Tiệp – machine learning cơ bản. Principal Component Analysis (phần ½), https://machinelearningcoban.com/2017/06/15/pca/.

[3] Phạm Đình Khánh, © Copyright 2021. Deep AI KhanhBlog, 13. k-Means Clustering, 13. k-Means Clustering — Deep AI KhanhBlog (phamdinhkhanh.github.io).