Trong báo cáo này, MANA team sẽ thực hiện phân tích chi tiết và phân loại khách hàng thành nhiều nhóm khác nhau để giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình và giúp họ dễ dàng sửa đổi sản phẩm hơn theo nhu cầu, hành vi và mối quan tâm cụ thể của các loại khách hàng khác nhau.

A. XỬ LÝ DỮ LIỆU

Data Cleaning:

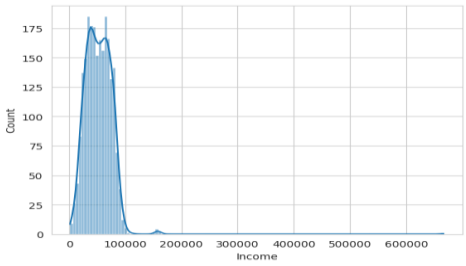
Xử lý CustomerIDs trùng lặp

► Ghi nhận bộ dữ liệu mới với 2240 bản ghi có ID duy nhất. (Giảm 829 dòng):
(1) Loại bỏ các ID trùng lặp
(2) Thực hiện điền dữ liệu khuyết thiếu bằng cách tìm kiếm giá trị đó không null từ các dòng khác
Kiểm tra trùng lặp

Kiểm tra dữ liệu khuyết thuyết

Kết hợp giá trị của Phone và Phone_number ► dữ liệu Phone_number không còn giá trị null
Với giả thuyết giá trị khuyết ở Phone thì sẽ có ở Phone_number
Xử lý dữ liệu bị thiếu cho Year_Register và Month_Register: Kiểm chứng giả thuyết, nếu không null thì Year_Register và Month_Register đúng bằng giá trị năm và tháng được tách ra từ Registration_Time ==> giá trị Năm và Tháng được lấy trực tiếp từ Registration_Time.

Data Transforming sau khi thực hiện Describe Data



Thực hiện loại bỏ ngoại lai đối với Income

- Ngoại lai: thu nhập lớn nhất được ghi nhận là 666.666 USD, trong khi lớn thứ hai là 162.397 USD ► xem xét 666.666 USD là ngoại lệ và loại bỏ nó.
- Giá trị khuyết thiếu: có 24 bản ghi bị thiếu ► Sử dụng kỹ thuật xác định K-Nearest Neighbors (KNN) để xử lý dữ liệu bị thiếu.

Biến đổi dữ liệu thành định dạng phù hợp từ float to integer để tối ưu:

- Kiểm tra và thực hiệ biến đổi các cột Year_Of_Birth, Recency, Liquor, Vegetables, Pork, Seafood, Candy, Jewellery, Num_Deals_Purchases, Num_Web_Purchases, Num_Catalog_Purchases, Num_Store_Purchases, Num_Web_Visits_Month, Complain, Phone_Number, Total_Purchase

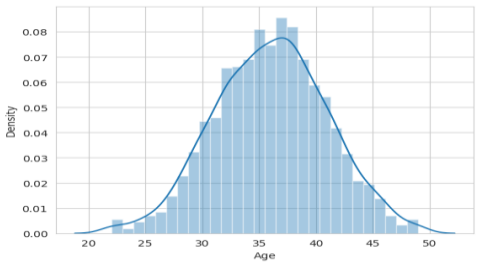
Phân tách Living_With with thành 2 cột: Marital_Status và Num_Children

Đổi 'Marital_Status' của 'Alone', 'YOLO', 'Absurd' thành Single vì tất cả đều chỉ sống 1 mình

Thay thế giá trị -1 bằng 0 trong Promo_40

Với giả thuyết khách hàng không dễ dàng chấp nhận chiến dịch thứ 4, vì trong các chiến dịch còn lại thì đa phần khách hàng không chấp nhận.

B. KHÁM PHÁ và GIẢI THÍCH DỮ LIỆU

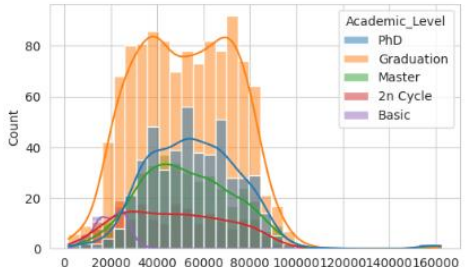


1. Khám phá dữ liệu nhân khẩu học

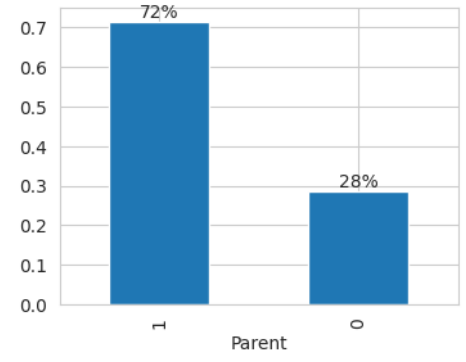
Tuổi của khách hàng phân phối chuẩn, chủ yếu là khách hàng có độ tuổi từ 32 đến 41, mỗi độ tuổi có trên 100 người.
Không có nhiều sự khác biệt giữa các giới tính, nam giới chiếm tỷ lệ lớn nhất 35%.

Biểu đồ thu nhập với các biến dân số:

- Nhóm có trình độ giáo dục cơ bản có thu nhập trung bình thấp hơn đáng kể so với 4 nhóm còn lại, và nhóm vòng đời thứ 2 cũng có thu nhập tương đối thấp.
- Khách hàng không có con cái có thu nhập cao hơn.
- Khách hàng nam có thu nhập trung bình cao hơn một chút so với khách hàng nữ hoặc các giới khác.
- Không có sự tương quan giữa độ tuổi và thu nhập.



Những người có trình độ PhD, Master và Graduation là những người kiếm tiền và chi tiêu nhiều nhất.
► Có thể được nhắm mục tiêu với các chiến dịch hàng hóa cao cấp hơn. Thu nhập phân phối trong khoảng từ 0 đến 100,000 USD. Tuy nhiên, phần tỉ lệ chi tiêu thấp hơn rất nhiều so với thu nhập.



2/3 số khách hàng đang sống chung với bạn đời (status = Married hoặc Together). 1/3 trong số họ đang sống một mình (status = Single, Divorced, or Widow)

Chủ yếu khách hàng là bố mẹ và có con (72%). Một nửa khách hàng có 1 con.

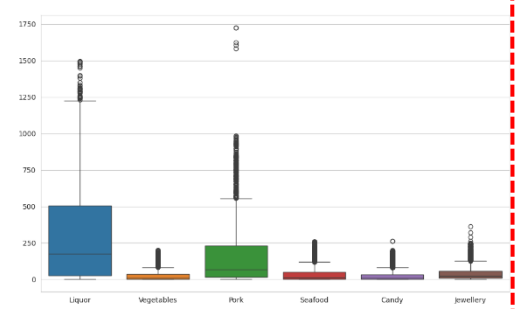
Tuy nhiên, sau khi phân tích sâu về Income và Spending của 2 nhóm này, nhóm không phải bố mẹ đã chi tiêu cao hơn hẳn dù chiếm số lượng ít hơn (30%).
Gap ratio in mean income between non-parents and parents: 41%

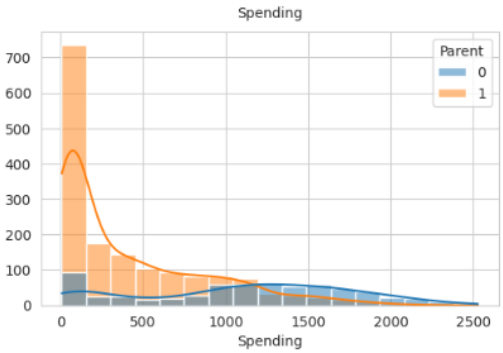
Gap ratio in mean spending between non-parents and parents: 170%

Ta có thể thấy rằng, nhóm khách hàng không có con cái chiếm ít hơn so với nhóm khách hàng có con cái (~30%). Tuy nhiên, nhóm này lại có chi tiêu cao hơn hẳn, Biểu đồ chi tiêu giữa 2 nhóm cho thấy nhóm này chi tiêu lệch phải, trong khi nhóm có con cái thì có chi tiêu lệch trái. Break down theo từng sản phẩm cho thấy, nhóm không có con cái có chi tiêu nhỉnh hơn rõ rệt ở tất cả sản phẩm. So sánh với income của hai nhóm này cho thấy, income của nhóm không có con cái vẫn có sự nhỉnh hơn so với nhóm có con cái, tuy nhiên gap của income nhỏ hơn rất nhiều so với gap của spending. Đề xuất tập trung vào các chiến dịch để tăng revenue từ nhóm này trong các chiến dịch sau này.

Kết luận này tương tự khi phân tích sâu vào các nhóm ngành hàng (Phân tích kĩ hơn ở mục 2)

2. Phân tích theo sản phẩm





Có một mối tương quan tích cực giữa thu nhập và chi tiêu ở tất cả các hạng mục, cho thấy rằng khi thu nhập tăng thì chi tiêu cũng tăng. Tuy nhiên, thu nhập cao hơn có xu hướng ảnh hưởng đến chi tiêu vào Rượu và Hải sản nhiều hơn. Team đề xuất tập trung tăng revenue từ nhóm khách hàng có thu nhập cao và tập trung vào 2 sản phẩm này.

Nhìn chung, sản phẩm rượu có mức chi tiêu cao hơn đáng kể so với các sản phẩm khác, tiếp theo là thịt lợn. Các sản phẩm còn lại có mức chi tiêu tương đối thấp hơn

Khách hàng là bố mẹ có xu hướng chi tiêu ít hơn.

Do đó, với mức thu nhập và chi tiêu của nhóm non-parents, đây được xem là nhóm tiềm năng, để thực hiện các bước tiếp theo phân loại, cũng như phân tích chiến lược marketing, promotion cho nhóm này.

C. ĐỀ XUẤT GIẢI PHÁP KĨ THUẬT

1. Feature Selection & Feature Engineering

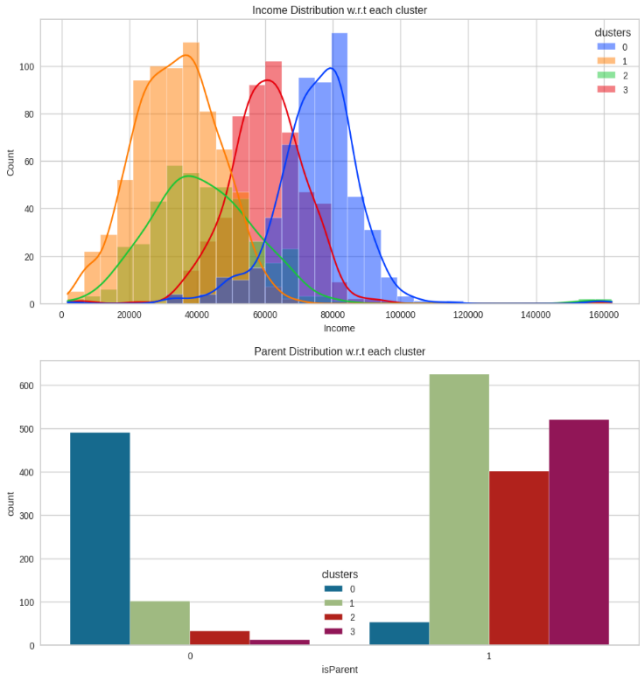
Đầu tiên, chúng ta sẽ tạo thêm một số biến mới như sau: Tuổi của khách hàng (dựa trên năm sinh), Là cha mẹ hay chưa (dựa vào số con trong gia đình), tổng số tiền đã chi tiêu kể từ khi đăng ký, thời gian với tư cách là khách hàng tính theo năm, tỷ lệ tổng tiêu dùng và thu nhập của khách hàng
Vì các cột 'Year_Of_Birth', 'Registration_Time' đều có các cột với lượng thông tin tương đương và cột 'Phone_Number' không có giá trị sử dụng quá lớn nên đây trở thành những cột thừa cần được bỏ đi.
Do chúng ta không thể đưa ra được thông tin gì từ bảng hệ số tương quan (heatmap) nên ta sẽ sử dụng hệ số phóng đại phương sai (VIF) để kiểm tra liệu các biến có xuất hiện đa cộng tuyến hay không. Kết quả cho thấy rằng các biến 'Amount_Spent', 'Total_Purchase' và 'Promo_50' chắc chắn xảy ra hiện tượng đa cộng tuyến do hệ số VIF bằng vô cùng nên ta sẽ bỏ các biến này.
Sau đó, ta sẽ chuyển các biến định tính ('Marital_Status', 'Academic_Level', 'Generation', 'Gender') thành các biến định lượng theo quy tắc sau:
Academic_Level: 2n Cycle = 0, Basic = 1, Graduation = 2, Master = 3, PhD = 4
Gender: Other = 0, Male = 1, Female = 2
Generation: Gen X = 0, Gen Y = 1, Gen Z = 2
Marital_Status: Married và Together bằng 2, trong khi các giá trị còn lại sẽ bằng 1
Cuối cùng, ta sẽ sử dụng phương pháp chuẩn hóa dữ liệu bằng hàm StandardScaler() từ thư viện sklearn.preprocessing để chuẩn hóa dữ liệu về cùng độ lớn với nhau.

2. Thuật toán

Có hai thuật toán được sử dụng trong bài toán này:

- Đầu tiên là thuật toán phân tích thành phần chính (Principal Component Analysis). Đây là phương pháp đơn giản nhất trong các thuật toán giảm chiều dữ liệu (Dimensionality Reduction) dựa trên một mô hình tuyến tính. Phương pháp này dựa trên quan sát rằng dữ liệu thường không phân bố ngẫu nhiên trong không gian mà thường phân bố gần các đường/mặt đặc biệt nào đó. PCA xem xét một trường hợp đặc biệt khi các mặt đặc biệt đó có dạng tuyến tính là các không gian con (subspace) (Vu, 2017).
- Sau khi giảm chiều dữ liệu, ta sẽ tiến hành phân cụm khách hàng dựa vào thuật toán K-Means Clustering. Mục đích cuối cùng của thuật toán phân nhóm này là: từ dữ liệu đầu vào và số lượng nhóm chúng ta muốn tìm, hãy chỉ ra center của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng (Vu, 2017).

3. Phương pháp thử nghiệm



Một vấn đề sẽ xảy ra với thuật toán K-Means Clustering là việc chọn đúng số cụm dữ liệu cần phân tách. Vì thế, có hai kỹ thuật được sử dụng để chọn đúng số cụm dữ liệu cần phân tách:
Phương pháp Elbow là một cách giúp ta lựa chọn được số lượng các cụm phù hợp dựa vào đồ thị trực quan hoá bằng cách nhìn vào sự suy giảm của hàm biến dạng và lựa chọn ra điểm khuỷu tay (elbow point) (13.1. Các bước của thuật toán k-Means Clustering — Deep AI KhanhBlog, no date).
Một phương pháp khác để có thể chọn đúng số lượng cụm là chỉ số Silhouette.
Cả hai phương pháp này đều có số cụm tương đương nhau là đều bằng 4. Sau khi chọn xong số cụm và tiến hành huấn luyện dữ liệu, ta sẽ tiến hành phân cụm dữ liệu dựa trên biến 'cluster' được tạo sau khi train xong. Tới đây, ta sẽ tiến hành phân cụm dữ liệu dựa trên các biến định tính/định lượng trong bộ dataset theo từng nhóm. Ví dụ như ảnh.

Reference:
Vu, T. (2017) Bài 27: Principal Component Analysis (phần 1/2). https://machinelearningcoban.com/2017/06/15/pca/#3-principal-component-analysis.
Vu, T. (2017) Bài 4: K-Means Clustering. https://machinelearningcoban.com/2017/01/01/kmeans/.
13.1. Các bước của thuật toán k-Means Clustering — Deep AI KhanhBlog (no date). https://phamdinhkhanh.github.io/deepai-book/ch_ml/KMeans.html#phuong-phap-elbow-trong-lua-chon-so-cum.