NỘI DUNG: TỔNG QUAN VỀ BÀI TOÁN DOANH NGHIỆP

Bảng thông tin các trường

Feature	Data Type	Description
ID	int	Unique identifier for each record
Year_Of_Birth	float	Year of birth of the individual
Academic_Level	object	Academic level of the individual
Income	float	Income of the individual
Registration_Time	object	Time of registration
Recency	float	Recency of interaction
Liquor	float	Purchase frequency of liquor
Vegetables	float	Purchase frequency of vegetables
Pork	float	Purchase frequency of pork
Seafood	float	Purchase frequency of seafood
Candy	float	Purchase frequency of candy
Jewellery	float	Purchase frequency of jewellery
Num_Deals_Purchases	float	Number of deals or purchases made
Num_Web_Purchases	float	Number of purchases made via the web
Num_Catalog_Purchases	float	Number of purchases made via catalogues
Num_Store_Purchases	float	Number of purchases made in physical stores
Num_Web_Visits_Month	float	Number of visits to websites per month
Promo_30	float	Frequency of promotional events with a discount of 30%
Promo_40	float	Frequency of promotional events with a discount of 40%
Promo_50	float	Frequency of promotional events with a discount of 50%
Promo_10	float	Frequency of promotional events with a discount of 10%
Promo_20	float	Frequency of promotional events with a discount of 20%
Complain	float	Frequency of complaints
Gender	object	Gender of the individual
Phone	float	Frequency of phone usage
Phone_Number	float	Frequency of phone usage
Year_Register	float	Year of registration
Month_Register	float	Month of registration
Total_Purchase	float	Total purchase made
Living_With	object	Living arrangement of the individual
Payment_Method	object	Method of payment

1. Tổng quan yêu cầu của doanh nghiệp và hướng giải quyết

Bài toán doanh nghiệp yêu cầu **tìm ra phân nhóm khách hàng tiềm năng giúp doanh nghiệp điều chỉnh sản phẩm** của mình dựa trên khách hàng mục tiêu từ các đối tượng khách hàng khác nhau. Phân khúc khách hàng là việc phân chia khách hàng thành các nhóm phản ánh sự tương đồng giữa các khách hàng trong mỗi cụm.

Để giải quyết bài toán, nhóm DATA USP sẽ thực hiện:

Bước 1: Làm sạch dữ liệu và tiền xử lý dữ liệu

Bước 2: Xây dựng và đánh giá thuật toán KMeans

Bước 3: Phân cụm dữ liệu không được giám sát về hồ sơ của khách hàng từ cơ sở dữ liệu của công ty.

Bước 4: Tổng kết về phân khúc khách hàng

2. Khám phá dữ liệu

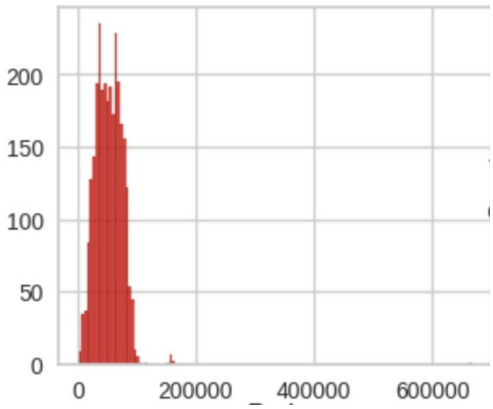
Bộ dữ liệu bao gồm **30 cột dữ liệu** ghi lại các khía cạnh đa dạng về hành vi cá nhân và thông tin nhân khẩu học. Các biến này có thể được sử dụng để phân nhóm khách hàng dựa trên thói quen mua sắm, sở thích và đặc điểm của họ. Việc phân nhóm này cho phép thực hiện các chiến lược tiếp thị phù hợp với nhu cầu và sở thích của từng phân đoạn, do đó nâng cao sự tương tác và hài lòng của khách hàng. Ngoài ra, các đặc điểm như thời gian tương tác gần nhất, tần suất khiếu nại và phương thức thanh toán có thể cung cấp thêm thông tin chi tiết về hành vi của khách hàng, hỗ trợ trong việc tinh chỉnh các chiến lược phân đoạn và tối ưu hóa hoạt động kinh doanh.

Chúng tôi phân loại dữ liệu thành 4 tập hợp chính, đó là 'Thông tin Khách hàng', 'Sản phẩm', 'Khuyến mãi', 'Địa điểm' như sau. Việc phân loại này chia dữ liệu thành các phần riêng biệt, góp phần vào việc giải thích và phân đoạn khách hàng sau này.

Bảng số lượng giá trị độc nhất mỗi trường

ID	2240
Year_Of_Birth	28
Academic_Level	5
Income	1974
Registration_Time	663
Recency	100
Liquor	776
Vegetables	158
Pork	558
Seafood	182

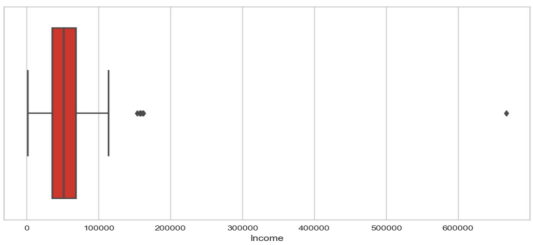
Biểu đồ tần suất của Income



Bảng giá trị được mã hóa nhãn

Column: Academic_Level Graduate: 0 Post Graduate: 1 Undergraduate: 2	Column: Marital_Status Married: 0 Single: 1 Unidentified: 2
Column: Payment_Method Card: 0 Cash: 1 Mobile: 2 Online: 3 Unknown: 4	Column: Gender Female: 0 Male: 1 Other: 2

Biểu đồ boxplot của Income



Bảng giá trị khi đã điều chỉnh quy mô

	Academic_Level	Income	Recency	Liquor
0	-0.953639	0.412716	1.479936	0.590205
1	0.877941	-0.515620	-1.284408	-0.864498
2	-0.953639	-0.963774	0.823404	-0.801898
3	0.877941	-0.071263	-0.593322	-0.503803
4	0.877941	-0.268106	1.203501	-0.816803

NỘI DUNG: LÀM SẠCH VÀ TIỀN XỬ LÝ DỮ LIỆU

3: Làm sạch và tiền xử lý dữ liệu

Trong bộ dataset trên, vấn đề đầu tiên mà nhóm kiểm tra đó chính là kiểm tra những giá trị độc nhất của mỗi cột, nhóm nhận thấy được là **cột "ID" có tận hơn 3000 dòng nhưng lại chỉ có 2240 giá trị độc nhất**, từ đó nhóm em nhận ra được dòng này bị nhân bản. Sau khi đọc dataset và kiểm tra những dòng bị nhân bản cột "ID" thì nhóm lại nhận ra được là có những thông tin chung một khách hàng bị nhân bản mà ở trường này có mà ở trường kia không có. Nhóm tiến hành kết hợp những thông tin khách hàng bị lặp lại đó và trả về cột "ID" là độc nhất.

Sau đó nhóm tiến hành kiểm tra những dữ liệu bị thiếu và những giá trị độc nhất của cột thì nhóm lại nhận thấy rằng có **rất nhiều dữ liệu đang bị thiếu**, đặc biệt **cột "Promo_40" lại còn xuất hiện thêm giá trị -1**. Điều này làm sai đi bản chất của cột vốn dĩ có giá trị 0 và 1. Nhóm tiến hành kiểm tra từng cột trong dataset và điền giá trị phù hợp vào những vào những vị trí còn thiếu thông qua việc xây dựng Histogram của từng cột để đánh giá và xử lí theo nhiều cách khác. Nhóm nhận thấy rằng tổng số đơn mua mỗi sản phẩm phải bằng tổng số đơn mua, nhóm tiến hành tính toán và điền giá trị còn thiếu vào những cột đó. Còn đối với riêng c**ột Promo_40 xuất hiện giá trị -1**, nhóm tiến hành thay thế giá trị -1 bằng giá trị trống và **xây dựng mô hình Random Forest** để điền giá trị 0 và 1 phù hợp vào.

Sau khi xử lí được các giá trị bị thiếu và giá trị bị sai ở trong bộ data trên, nhóm tiến hành **thay thế những giá trị** ở trong các cột "Living_With", "Academic_Level" thành những giá trị dễ hiểu hơn đồng thời đổi tên trường "Living_With" thành "Marital_Status". Bên cạnh đó nhóm còn xóa đi cột bị thừa là cột "Unnamed:0"

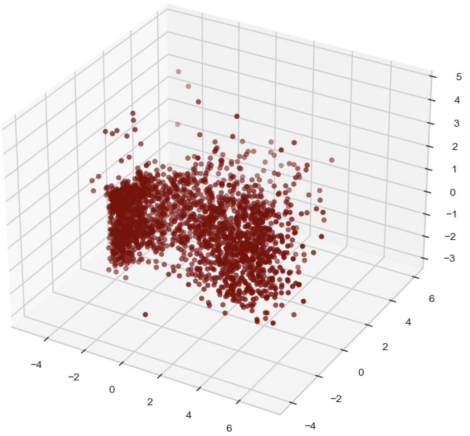
Sau khi xử lí được những lỗi sai trong bộ dataset, nhóm tiến hành **mã hóa nhãn** cho từng giá trị trong cột thành những giá trị phù hợp. Điều này giúp việc xử lí dữ liệu dễ dàng hơn đồng thời đưa dữ liệu thành quy mô phù hợp phục vụ cho việc **xây dựng thuật toán phân chia phân khúc khách hàng KMean Clustering** sau này.

Bên cạnh những vấn đề về giá trị ở trong cột, nhóm còn nhận thấy được những **vấn đề về giá trị ngoại lai** bằng cách **xây dựng biểu đồ Boxplot** cho 3 cột là "Income", "Total_Purchase" và **tạo cột mới** có tên là "Total_Spending" bằng tổng chi tiêu các sản phẩm phục vụ cho việc phân tích dữ liệu được dễ dàng hơn. Đồng thời qua biểu đồ Boxplot đó nhóm nhận thấy được có các giá trị ngoại lai xuất hiện, điều này gây ảnh hưởng đến việc xây dựng mô hình và giảm độ chính xác của thuật toán. Nhóm tiến hành loại bỏ các giá trị ngoại lai để thuật toán được xây dựng với độ chính xác cao hơn.

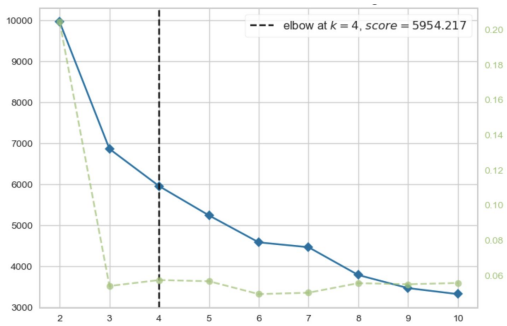
Sau khi xử lí xong bộ dữ liệu, nhóm tiến hành **điều chỉnh quy mô bộ dữ liệu** để quá trình xây dựng dữ liệu **không bị chênh lệch** quy mô quá lớn. Điều này đóng góp cho việc xây dựng thuật toán được hiệu quả cao hơn và với độ chính xác cao hơn. Tiếp đến nhóm tiến hành **xây dựng thuật toán Kmean Clustering** cho bộ dữ liệu đã được tiêu chuẩn hóa và **phân cụm dữ liệu** dựa trên thuật toán đã được xây dựng.

NỘI DUNG: XÂY DỰNG VÀ ĐÁNH GIÁ THUẬT TOÁN KMEANS

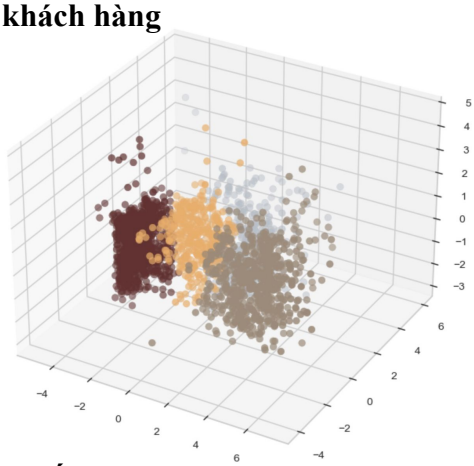
Phép chiếu dữ liệu 3D ở kích thước giảm



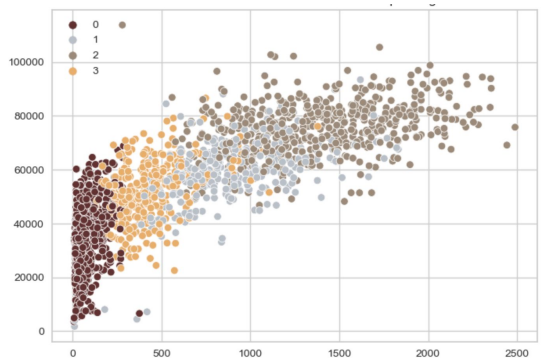
Điểm biến dạng của khuỷu tay cho phân cụm KMeans



Phép chiếu dữ liệu 3D từng cụm khách hàng



Mối quan hệ giữa doanh thu và chi tiêu của 4 nhóm khách hàng



4: Xây dựng và đánh giá thuật toán

Sau khi xử lý và tiêu chuẩn hóa được bộ dữ liệu, nhóm tiến hành **sử dụng PCA (Principal Component Analysis) để giảm chiều dữ liệu.** PCA giúp cho chúng ta **giảm chiều dữ liệu từ D xuống K** sao cho K < D và giữ lại những thành phần quan trọng nhất. Đây là phương pháp đi tìm một hệ cơ sở mới làm 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.

Nhóm sử dụng **PCA để giảm chiều dữ liệu xuống 3** và vẽ ra hình ảnh 3D của 3 nhóm được giữ lại để minh họa rõ hơn về phân bổ dữ liệu của mình.

Bước tiếp theo nhóm thực hiện đó là **sử dụng phương pháp Elbow để xác định được số cụm khách hàng** mình cần chia nhỏ là bao nhiêu. 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*). Sau khi sử dụng Elbow phân tích và xác định đươc là nhóm cần chia nhỏ thành 4 cụm khách hàng khác nhau thông qua biểu đồ đã được nhóm trực quan hóa. Chính vì biết được mình sẽ cần chia nhỏ khách hàng của mình thành bao nhiêu cụm dựa trên bộ dữ liệu đã được giảm chiều, nhóm đã **xây dựng được mô hình Kmean Clustering với số cụm chính xác và dễ phân biệt hơn giữa các cụm này.**

Ngay sau khi chọn được số cụm cho **thuật toán Kmean Clustering**, nhóm tiến chạy chạy thuật toán để phân chia rõ hơn những cụm khách hàng này đồng thời trực quan hóa nó lên để có thể dễ nhận biết và đánh giá được từng nhóm khác nhau. Biểu đồ này giúp chúng ta có thể thấy rõ được 4 nhóm đã được phân cụm và có cái nhìn tổng quan về 4 nhóm khách hàng. Do đó, việc chuẩn bị và phân tích sâu hơn về từng cụm khách hàng được nhóm làm cụ thể và chính xác hơn.

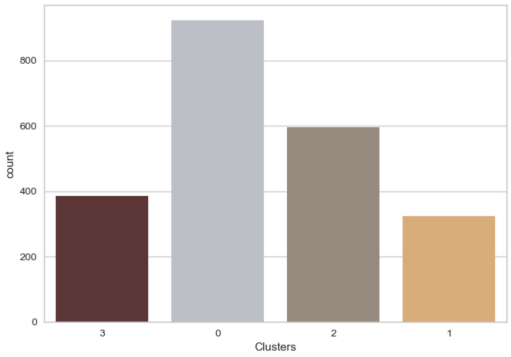
Việc phân cụm ra được từng cụm khác nhau và trực quan hóa nó lên không phải là kết thúc của việc xây dựng thuật toán Kmean. Để **đánh giá độ hiệu quả của thuật toán** phải dựa trên việc phân tích ra được gì, mỗi nhóm khách hàng có đặc điểm như thế nào, nhóm khách hàng này khác nhau với các nhóm khác như thế nào, nhóm này có đặc trưng gì mà những nhóm khác không có. Do đó nhóm tiến hành đi sâu vào từng nhóm khách hàng và vân tích theo nhiều tiêu chí khác nhau để nhóm có thể hiểu ra được nhóm khách hàng nào quan trọng, nhóm khách hàng nào tiềm năng và nhóm khách hàng nào mình nên tập trung vào để định hướng sản phẩm của mình.

Điều đầu tiên mà nhóm muốn đánh giá đó là về doanh thu và chi tiêu của từng nhóm và minh họa được mối quan hệ giữa hai yếu tố đó. Nhóm nhận ra được là nếu xét về thu nhập và chi tiêu. Nhóm có thể chia thành 4 nhóm khác nhau, đó chính là:

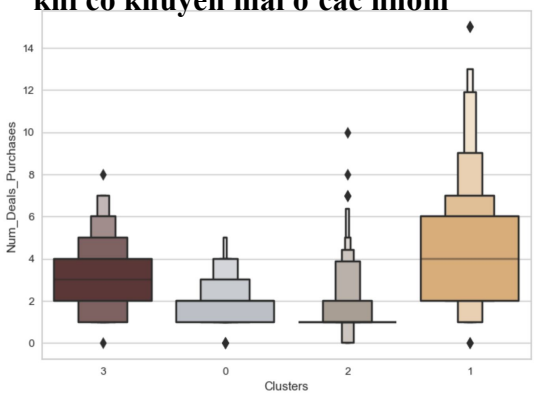
- **Nhóm 0:** Thu nhập thấp và chi tiêu thấp
- **Nhóm 1:** Thu nhập trung bình (cao hơn nhóm 1) và chi tiêu nhiều
- **Nhóm 2:** Thu nhập cao và chi tiêu rất nhiều
- **Nhóm 3:** Thu nhập trung bình và chi tiêu trung bình

NỘI DUNG: XÂY DỰNG VÀ ĐÁNH GIÁ THUẬT TOÁN KMEANS

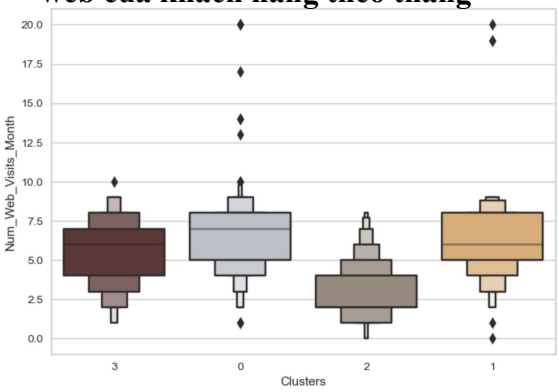
Biểu đồ phân phối số lượng khách hàng của mỗi nhóm



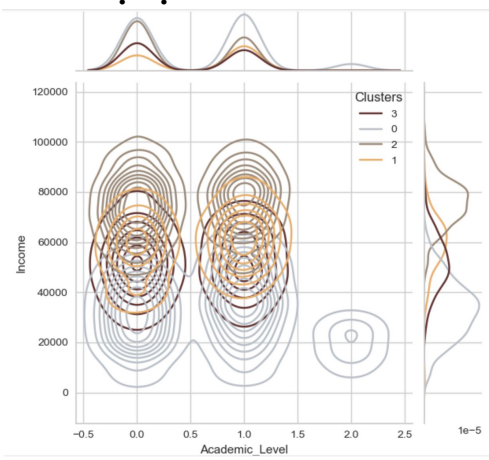
Biểu đồ hộp về phân phối lượng mua khi có khuyến mãi ở các nhóm



Biểu đồ hộp về phân phối số lượng lướt web của khách hàng theo tháng



Biểu đồ kết hợp giữa thu nhập và trình độ học vấn của 4 nhóm



Sau khi đánh giá về thu nhập và chi tiêu, nhóm tiến hành **nghiên cứu rõ hơn về từng nhóm khách hàng** thông qua rất nhiều yếu tố khác để có cái nhìn tổng quan hơn đồng thời việc phân cụm khách hàng được chi tiết hơn.

Nhóm tiến hành vẽ **bảng phân phối số lượng khách hàng** của mỗi nhóm và nhận thấy được **nhóm 0 là nhóm chiếm tỉ lệ cao nhất** với hơn 850 người, tiếp đến là nhóm 2 với khoảng 600 người, đứng thứ ba là nhóm 3 với khoảng gần 400 người và cuối cùng là nhóm 1 với khoảng hơn 300 người.

Dựa vào biểu đồ bên, nhóm nhận thấy được là g**iữa 4 nhóm có sự khác biệt rõ rệt về mức độ quan tâm đến deal** khi mà **nhóm 0 là nhóm có quan tâm đến deal nhưng mà không quá nhiều, nhóm 1 là nhóm quan tâm về deal nhiều nhất**. Tuy nhiên dựa vào bảng phân phối về số lượng khuyến mãi được khách hàng săn đón và chấp nhận khuyến mãi thì **rất ít sự hứng của khách hàng đối với chương trình khuyến mãi** của công ty khi mà tỉ lệ khách hàng không chấp nhận tham gia chương trình khuyến mãi chiếm tỉ lệ lớn so với những khách hàng tham gia vào chương trình khuyến mãi. Điều này yêu cầu doanh nghiệp phải **tập trung vào các chiến dịch quảng cáo** sản phẩm phù hợp với nhu cầu khách hàng hơn, mang lại cho họ sự hứng thú với các sản phẩm của mình thông qua các chương trình quảng cáo để tối ưu hóa doanh thu của công ty.

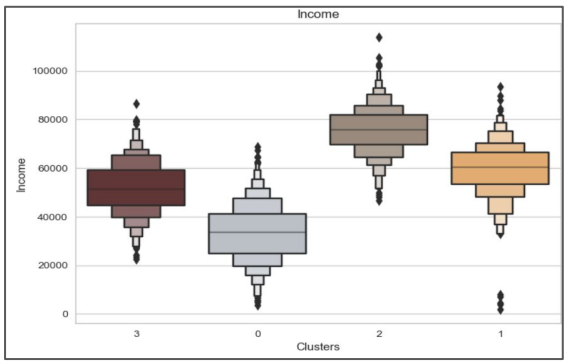
Bên cạnh những vấn đề về deal và khuyến mãi thì một vấn đề mà nhóm muốn chú ý đó chính là tỉ lệ lướt web ở các nhóm khách hàng mỗi tháng là một cái đáng được chú ý khi mà **nhóm có thu nhập cao nhất và chi tiêu nhiều nhất họ lại không lướt web nhiều** mà họ mua **chủ yếu trên các cửa hàng**, có thể họ là những người bận rộn nên họ mua nhiều ở các cửa hàng thuận đường để mang lại sự thuận tiện hơn cho họ. Trái ngược với **nhóm có thu nhập và chi tiêu thấp nhất họ lại thường hay lướt web**, họ là những người khá rảnh rỗi, họ có mong muốn mua sản phẩm của doanh nghiệp nhưng thu nhập lại thấp. **Nhóm 1 là nhóm sẵn sàng mua bất cứ chỗ nào** tuy nhiên họ lại không mua ở trên các sàn thương mại điện tử nhiều bằng những nơi khác, họ là những người quan tâm đến deal cho nên là ở các sàn thương mại điện tử không có nhiều deal để họ mua nhiều. Còn nhóm số 3 là nhóm khá trung lập về vị trí mua sản phẩm khi mà họ không có đặc biệt thích cũng không có ghét mua ở một chỗ nhất định, và họ là nhóm đối tượng rất thích lướt web để xem sản phẩm.

Ngoài những yếu tố về vấn đề doanh thu và vấn đề mua hàng, dựa vào biểu đồ kết hợp giữa nhóm còn nhận thấy sự khác nhau ở trong yếu tố về trình độ học vấn khi mà **nhóm thu nhập ít nhất và chi tiêu ít nhất có những người chưa có bằng cấp** trong khi đó ở các nhóm còn lại họ đều là những người đã tốt nghiệp, có bằng cấp hay thậm chí họ còn là những người thạc sĩ, tiến sĩ.

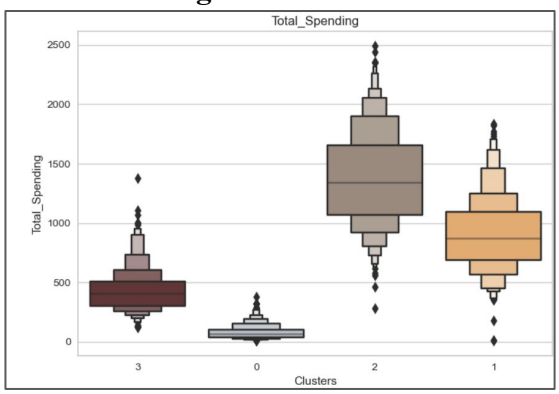
Những yếu tố đến đã giúp làm rõ hơn hình ảnh từng nhóm khách hàng và những đặc trưng của họ. Điều này góp phần cho việc xây dựng doanh nghiệp có thể chú trọng vào một số nhóm khách hàng nhất định và nên hạn chế kinh doanh vào những nhóm nào. Điều này sẽ được làm rõ ở phần sau đây.

NỘI DUNG: TỔNG KẾT PHÂN KHÚC KHÁCH HÀNG

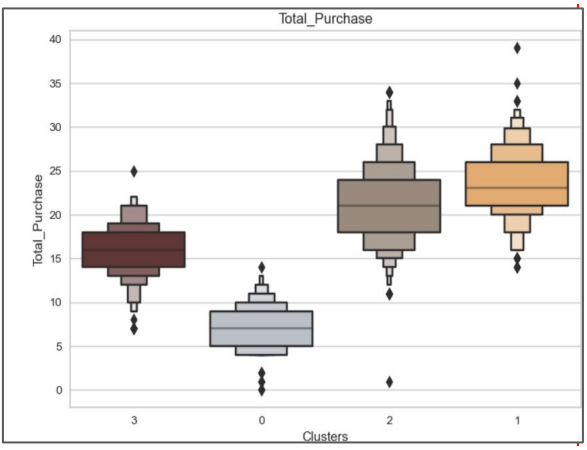
Biểu đồ so sánh thu nhập giữa các



Biểu đồ so sánh tổng giá trị chi tiêu giữa các nhóm



Biểu đồ so sánh số lượng giao dịch giữa các nhóm



5: Tổng kết phân khúc khách hàng

Do đó sau quá trình nghiên cứu cụ thể về từng nhóm khách hàng và những đặc điểm riêng của họ. Nhóm quyết định sẽ chia thành bốn phân khúc khách hàng chính, bao gồm: Cluster 0 **(nhóm vãng lai),** cluster 1 **(nhóm thân thiết),** cluster 2 **(nhóm quan trọng),** và cluster 3 **(nhóm tiềm năng).** Mỗi phân khúc sẽ được miêu tả dựa trên các đặc điểm thu nhập, hành vi mua sắm, sở thích và các thói quen truy cập website của họ.

Cluster 0 bao gồm các khách hàng có thu nhập thấp. Họ mua sắm ít và tiêu dùng ít, đồng thời có trình độ học vấn thấp. Nhóm này không mua sắm nhiều ở bất kỳ địa điểm nào và thường có thói quen lướt web. Họ không có xu hướng gắn bó với bất kỳ kênh mua sắm cụ thể nào, mà thường chỉ tham gia vào các giao dịch nhỏ lẻ và không thường xuyên. Chính vì vậy nhóm đặt cho họ là nhóm **Vãng lai.**

Cluster 1 bao gồm các khách hàng có thu nhập cao, thường xuyên giao dịch và tổng giá trị chi tiêu cao. Họ sẵn sàng mua sắm ở bất kỳ địa điểm nào và không ưa thích việc mua sắm qua catalog. Nhóm này cũng thường xuyên lướt web, nhưng không dành sự ưu tiên đặc biệt cho bất kỳ kênh mua sắm nào ngoài việc sẵn sàng mua bất cứ chỗ nào khi cần. Chính vì vậy nhóm đặt cho họ là nhóm **Thân thiết.**

Cluster 2 là những khách hàng có thu nhập rất cao, thực hiện nhiều giao dịch và có tổng giá trị chi tiêu rất cao. Họ là những người mua sắm nhiều nhất tại nhiều địa điểm, trừ việc mua sắm qua web không được họ ưa chuộng bằng. Nhóm này ít lướt web và thường tập trung vào các kênh mua sắm truyền thống và trực tiếp. Chính vì vậy nhóm đặt cho họ là nhóm **Quan trọng.**

Cluster 3 bao gồm các khách hàng có thu nhập trung bình, thực hiện giao dịch và có tổng giá trị chi tiêu trung bình. Họ quan tâm đến các chương trình khuyến mãi và deal tốt. Nhóm này có thái độ trung lập về địa điểm mua sắm, không có sự ưu tiên rõ ràng cho bất kỳ kênh nào, và thường xuyên lướt web. Chính vì vậy nhóm đặt cho họ là nhóm **Tiềm năng.**

Sau những thông tin mà nhóm đưa ra qua việc phân khúc khách hàng, doanh nghiệp sẽ có thêm thông tin để có thể tối ưu hóa chiến lược tiếp thị và dịch vụ, tăng cường tương tác cá nhân hóa và nâng cao trải nghiệm của khách hàng. Ngoài ra doanh nghiệp có thể cân nhắc tập trung đầu tư vào nhóm Thân thiết và Quan trọng để tăng tỉ lệ giữ chân khách hàng.