

NỘI DUNG PHÂN TÍCH

I. TỔNG QUAN NGHIỆP VỤ:

1. **Giới thiệu:** Bộ dữ liệu trên được thu thập và tổng hợp từ thông tin khách hàng đăng ký với công ty trong khoảng thời gian từ 30/7/2021 đến 29/6/2023. Bộ dữ liệu bao gồm các thông tin cá nhân, chỉ tiêu hành vi mua hàng.
2. **Cấu trúc bộ dữ liệu:** Bộ dữ liệu bao gồm tất cả gồm 3069 bản ghi (bao gồm 8 trường category và 23 trường numeric), được thu thập và tổng hợp trong 2 năm. Trong đó, các trường được chia theo 4 mục chính:
 - People: gồm 7 trường về thông tin cá nhân, đặc điểm nhân khẩu học. Đáng chú ý là trường ID định danh khách hàng.
 - Products: gồm 6 trường tương ứng với số tiền chi tiêu cho 6 loại mặt hàng trong 2 năm qua
 - Promotion: gồm 6 trường với thể hiện tổng số lượt mua hàng có giảm giá và có hay không tham gia các giảm giá 10% đến 50%.
 - Place: gồm 5 trường thể hiện thông tin về tổng số lượt mua hàng và xác định số lượt qua 3 kênh: website, cửa hàng, catalog; cùng với số lượt truy cập website.

II. LÀM SẠCH DỮ LIỆU:

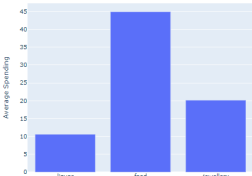
Sau khi xem xét tất các các biến có trong dataset, chúng tôi sẽ xử lý bộ dữ liệu trên thông qua 6 bước chính. Tất các các bước đã được mô tả một cách chi tiết như sau:

Step	Action	Variable Name	Explanation
1	Chỉnh sửa kiểu dữ liệu	Registration_Time	Chuyển dữ liệu dạng object sang dạng datetime
2	Xử lý dữ liệu ngoại lai và trùng lặp		Sử dụng phương pháp Z-Score xử lý những dữ liệu ngoại lai và loại bỏ những ID trùng lặp
3	Xóa biến không cần thiết	Phone_Number	Xóa những biến không có ý nghĩa trong việc phân tích dữ liệu
		Phone	
		Year_Register	
		Month_Register	
4	Xử lý dữ liệu bị thiếu	Payment_Method	Thay thế những dữ liệu bị thiếu trong cột bằng giá trị 'Other'
		Income	Thay thế dữ liệu bị thiếu bằng giá trị trung vị
5	Tạo biến mới	Marital_Status	Tạo các biến mới từ biến Living_With và xóa biến sau khi tách.
		Children	
		Year_Month	Trích xuất dữ liệu các biến mới từ biến Registration_Time
		Year	
		Month	
		Day	
		Age	Tính tuổi của khách hàng từ biến Year_Of_Birth

NỘI DUNG PHÂN TÍCH

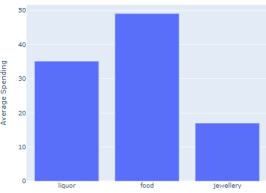
6	Đổi tên biến	Tất cả các biến	Đổi tên biến nhằm mục đích thuận tiện hơn trong quá trình xử lý
---	--------------	-----------------	---

Average Spending in <20K Income Segment



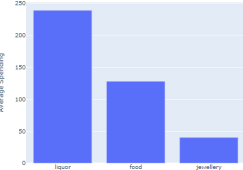
Hình 2.1

Average Spending in 20-40K Income Segment



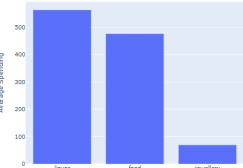
Hình 2.2

Average Spending in 40-60K Income Segment



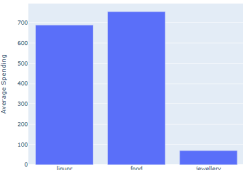
Hình 2.3

Average Spending in 60-80K Income Segment



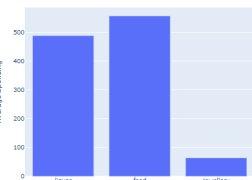
Hình 2.4

Average Spending in >80K Income Segment



Hình 2.5

Average Spending for People who do not have Child



Hình 2.6

III. Phân tích EDA

1. Tổng quan khách hàng phổ hiện tại

Qua phân tích distribution, chúng tôi nhận thấy tệp khách hàng hiện tại có những đặc điểm phổ biến như sau:

- Giới tính: Nam, nữ hoặc khác.
- Độ tuổi: từ 22 đến 49 tuổi
- Thu nhập chủ yếu trong khoảng từ \$10,000 - \$90,000
- Bao gồm 5 trình độ học vấn: Graduation, 2n Cycle, Basic, Master, PhD
- Tình trạng hôn nhân chủ yếu: Divorced, Married, Single, Together, sống chung với từ 0 đến 4 con.
- Mua hàng thường xuyên qua 3 kênh: Website, Category, Store.
- Mặt hàng thường mua: Liquor (rượu), Vegetables, Pork, Seafood, Candy và Jewellery.

2. Phân tích EDA

a. Phân tích theo thu nhập

Đầu tiên, chúng tôi chia Products thành 3 nhóm: Liquor, Food (Vegetables, Pork, Seafood, Candy) và Jewellery.

Tiếp theo, chúng tôi phân chia thu nhập ra làm 5 nhóm:

Nhóm 1: Thu nhập dưới \$20,000/ năm (Hình 2.1)

⇒ Những người thu nhập dưới \$20,000 thường dùng phần lớn số tiền để mua Food (Vegetables, Pork, Seafood, Candy). Lý do là vì với thu nhập thấp họ chỉ có thể chi trả cho những nhu cầu thiết yếu như thức ăn.

Nhóm 2: Thu nhập \geq \$20,000 và $<$ \$40,000 (Hình 2.2)

⇒ Những người thu nhập dưới \$40,000 thường dùng phần lớn số tiền để mua Food (Vegetables, Pork, Seafood, Candy). Ngoài ra ta có thể thấy họ chi trả cho Liquor gấp 3.5 lần Nhóm 1. Lý do là vì với thu nhập cao hơn Nhóm 1, họ có thể chi tiêu cho Liquor, bên cạnh Food.

Nhóm 3: Thu nhập \geq \$40,000 và $<$ \$60,000 (Hình 2.3)

⇒ Khác với 2 nhóm trước, số tiền Nhóm 3 chi trả cho Liquor cao hơn Food, điều đó chứng tỏ rằng nhóm người này đã sẵn sàng chi trả cho những thú vui của họ vì những nhu cầu thiết yếu như Food không còn là trở ngại.

Nhóm 4: Thu nhập \geq \$60,000 và $<$ \$80,000 (Hình 2.4)

⇒ Giống với Nhóm 3

Nhóm 5: Thu nhập \geq \$80,000 (Hình 2.5)

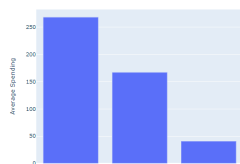
⇒ Nhóm người này với thu nhập cao, họ chi tiêu cho Liquor và Food gần như là bằng nhau. Tuy vậy, mức chi tiêu vẫn cao hơn so với các nhóm còn lại.

b. Phân tích theo số con trong gia đình

Chúng tôi chia khách hàng làm 2 nhóm: nhóm có con và nhóm không có con.

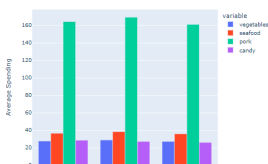
- Nhóm không có con thường là những người có thu nhập cao (20K - 60K), và vì sống một mình và không có gánh nặng tài chính nên họ thường sẽ sống hưởng thụ và chi tiêu cho những thú vui của mình nhiều hơn (số tiền chi cho Liquor

NỘI DUNG PHÂN TÍCH



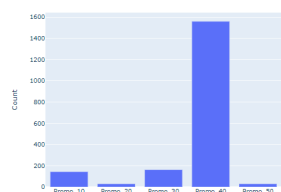
Hình 2.7

Correlation between Age Group and Spending on Food Categories



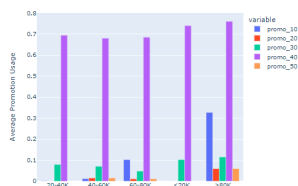
Hình 2.8

Number of Purchase by Promotion



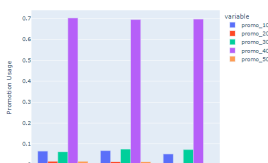
Hình 2.9

Correlation between Income Group and Promotion Usage



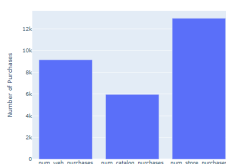
Hình 2.10

Promotion Usage by Age Group



Hình 2.11

Number of Purchases through Different Channels



Hình 12.12

nhiều hơn chỉ cho Food) (Hình 2.6).

- Nhóm có con, thường là những người có thu nhập thấp (dưới 20K), do đó họ mức chi tiêu trung bình của họ thấp hơn và chỉ chi tiêu phần lớn cho những mặt hàng thực phẩm thiết yếu (Food) (Hình 2.7).

c. Phân tích theo độ tuổi

Chúng tôi chia khách hàng thành 3 nhóm tuổi: dưới 30 tuổi (nhóm 1), từ 31 đến 40 tuổi (nhóm 2) và trên 40 tuổi (nhóm 3).

Chúng tôi thấy rằng, nhóm 1 chi tiêu ít nhất cho rau củ và hải sản nhưng lại có chi tiêu tương đối cao cho kẹo. Nhóm 2 có mô hình chi tiêu cân bằng cho tất cả các loại thực phẩm. Nhóm 3 chi tiêu nhiều nhất cho rau củ và hải sản, cho thấy sự ưu tiên cho các lựa chọn thực phẩm lành mạnh hơn trong khi ít chi tiêu cho kẹo (Hình 2.8).

d. Phân tích theo promotion

- Trong 5 chiến dịch giảm giá, chiến dịch giảm giá 40% thu hút được nhiều khách hàng nhất và mang doanh thu về nhiều nhất (Hình 2.9).
- Những người thuộc Nhóm thu nhập thấp (1 và 2) không quan tâm đến chương trình Promo_10, Promo_20 và Promo_50. Những người thuộc nhóm có thu nhập cao (4 và 5) có mức sử dụng khuyến mãi trung bình là thấp nhất do những người có thu nhập cao sẽ ít bị ảnh hưởng bởi các chương trình khuyến mãi (Hình 2.10).
- Ngoài ra, người thuộc nhóm tuổi 3 sẽ có độ nhạy cảm về việc giảm giá hơn, sau đó đến nhóm tuổi 2 và 1, lý do là vì càng lớn tuổi, họ sẽ càng quan tâm và tham gia vào các chương trình giảm giá hơn (Hình 2.11).

e. Phân tích theo kênh mua hàng

- Người tiêu dùng có xu hướng mua hàng tại cửa hàng nhiều nhất cho thấy sự ưa thích của họ đối việc mua sắm tại cửa hàng. Mua hàng trực tuyến cũng khá đáng kể cho thấy sự ưa chuộng mạnh mẽ đối với thương mại điện tử. Mua sắm qua catalog là phương thức ít được ưa chuộng nhất trong số các kênh mua hàng (Hình 2.12).

3. Chân dung khách hàng mục tiêu

Từ những phân tích trên, chúng tôi rút ra một số đặc điểm của tệp khách hàng mục tiêu như sau:

***Demographic:** Bất kể là nam hay nữ, độ tuổi là 31 đến 40, thu nhập nằm ở tầm trung bình cao (40K - 80K), đã tốt nghiệp, không có con, là người chủ động, có quyền quyết định về chi tiêu và mua sản phẩm.

***Hành vi:**

- Chi tiêu cho các mặt hàng Liquor và Food là ngang nhau.
- Họ sử dụng kênh mua hàng truyền thống tại cửa hàng là chủ yếu.
- Mức độ nhạy cảm với các chương trình giảm giá là trung bình bởi vì bên cạnh việc giảm giá, họ sẽ coi trọng chất lượng sản phẩm và không sẵn sàng chi trả cho những sản phẩm rẻ mà kém chất lượng.

***Nhu cầu:** Họ không có con nên nhu cầu chi tiêu nhiều hơn mà không cần phải lo về gánh nặng tài chính.

NỘI DUNG PHÂN TÍCH

IV. PHÂN CỤM KHÁCH HÀNG

1.1. Feature Selection:

- Recency: Số ngày từ lần mua hàng cuối cùng của khách hàng.
- Frequency: Tổng số lượt mua hàng của mỗi khách hàng
- Monetary: Tổng số tiền mỗi khách hàng chi trả
- Income: Mức thu nhập trung bình của mỗi khách hàng
- Children: Số con trong gia đình

⇒ Dựa vào 3 yếu tố trên có thể xác định được hành vi mua hàng của khách hàng từ đó dự đoán hành vi tương lai. Ví dụ như recency để xác định khách hàng nào có nguy cơ rời bỏ, frequency thể hiện việc họ có phải khách hàng thường xuyên hay không, Monetary để xác định sức mua của khách hàng và đây là tệp main revenue drivers .

Ngoài ra, qua phân tích EDA cho thấy có 2 yếu tố ảnh hưởng lớn đến hành vi tiêu dùng của khách hàng là Income và Children. Người có thu nhập cao và có ít con thường chi trả cho liquor và food ngang nhau. Trong khi đó, người có thu nhập bình quân thì thường chi tiền gấp đôi cho liquor khi so với nhu yếu phẩm là food.

1.2. Feature engineering:

- Frequency= Total_purchase
- Monetary= Liquor+Vegetables+Pork+Seafood+Candy+Jewellery

2. Thuật toán: có hai model: RFM, và RFM kết hợp K means clustering

Flow thuật toán:

1. RFM: xác định cụm khách hàng dựa trên 3 yếu tố RFM để chỉ ra được 5 nhóm khách hàng nổi bật: Best customers, Loyal customers, Big spenders, Almost Lost, Lost customers, và Lost Cheap customers. Phân nhóm dựa trên thang điểm của các yếu tố. Tuy nhiên số lượng khách hàng thuộc các nhóm trên khá ít so với tổng số lượng khách hàng nên không chọn phương pháp này.

2. RFM kết hợp K Means clustering:

- RFM để xác định mức độ mua hàng gần đây, thường xuyên, và lượng chi tiền cùng 2 yếu tố income và children
- K mean clustering: Chia các khách hàng thành những clusters với những đặc điểm giống nhau để có chiến lược phù hợp cho từng clusters

3. Phương pháp thử nghiệm:


- Đối với phương pháp RFM + Kmeans Clustering, nhóm sẽ sử dụng Elbow method (dựa trên tổng khoảng cách bình phương từ mỗi điểm dữ liệu đến centroid) để chọn ra số k nhóm mà từ điểm đó trở đi thì việc chia thêm nhóm không cải thiện mô hình nhiều
- Trực quan hoá khi k=3,4,5 ⇒ Thấy rằng việc chia thành 4 nhóm sẽ được hiệu quả phân tách nhóm nhưng vẫn đảm bảo được tệp khách hàng không bị chia quá nhỏ dẫn đến việc khó tập trung phân bổ nguồn lực vào các chiến lược tiếp thị.

	id	recency	frequency	monetary
831	1001.0	37.0	31.0	1105.0
1132	1002.0	92.0	21.0	738.0
301	1005.0	65.0	27.0	1318.0
1101	1006.0	12.0	7.0	67.0
2888	1007.0	55.0	33.0	1665.0

Hình 4.1. Yếu tố RFM

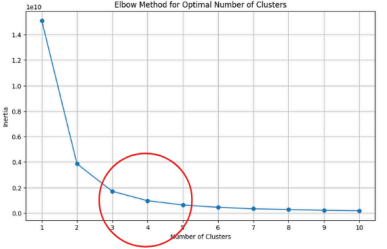
```
data_rfm['revenue']= data_rfm['liquor']
+ data_rfm['vegetables'] + data_rfm['pork']
+data_rfm['seafood']+data_rfm['candy']
+data_rfm['jewellery']
```

Hình 4.2. Feature engineering

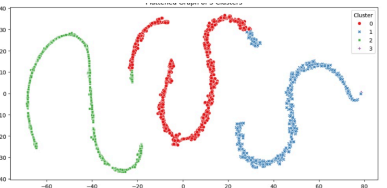


Best Customers: 51
Loyal Customers: 512
Big Spenders: 560
Almost Lost: 10
Lost Customers: 103
Lost Cheap Customers: 129

Hình 4.3. Phân nhóm RFM



Hình 4.4. Elbow method cho Kmeans clustering



Hình 4.5. Phân 4 cụm sử dụng K means clustering

NỘI DUNG PHÂN TÍCH

4. Kết quả phân cụm

Nhìn chung, các nhóm đều có mức recency ngang nhau, tuy nhiên về giá trị mang lại, sở thích tiêu dùng và tần suất lại có sự khác biệt như sau

Nhóm	Đặt Tên	Đặc điểm
0	Potential Customers	-Thu nhập: trung bình -Tỉ lệ con: nhiều con nhất so với mặt bằng chung -Tần suất: số lần mua hàng nhiều thứ 2 trong các nhóm - Giá trị: đứng thứ 2, 25%/tổng - Xu hướng chi tiêu: Chi tiêu cho rượu rất nhiều (gấp 2 lần mức chi thức ăn bình thường)
1	Main Revenue Drivers/Champions	-Thu nhập: cao -Tỉ lệ con: ít nhất - Tần suất: số lần mua hàng nhiều nhất trong các nhóm - Giá trị: chiếm 65% tổng thu - Xu hướng chi tiêu: chi cho rượu và thức ăn ngang nhau
2	Needing Attention	-Thu nhập: thấp -Tỉ lệ con: khá nhiều - Tần suất: số lần mua hàng thấp - Giá trị: nhỏ, chiếm khoảng 5%/tổng - Xu hướng chi tiêu: chi nhiều nhất cho thức ăn (nhu yếu phẩm), hạn chế các mặt hàng khác
3		Không đáng kể về số lượng



Cluster	income	children	recency	frequency
0	52344.65	1.23	49.72	16.13
1	76967.84	0.44	49.02	20.86
2	28331.94	1.12	48.56	7.90
3	66666.00	1.00	23.00	11.00

Hình 4.6. Kết quả tổng hợp phân cụm K means clustering



Cluster	liquor	food	candy	jewellery
0	290.65	146.85	18.19	45.89
1	616.96	539.38	60.20	70.17
2	31.33	42.68	6.06	17.74
3	10.00	42.00	1.00	12.00

Hình 4.7. Kết quả tổng hợp phân cụm K means clustering

5. Đề xuất: Tập trung nhóm: Main Revenue Driver/Champions

Sau khi phân tích EDA và RFM, chúng tôi nhận thấy rằng chân dung khách hàng của phần EDA khá tương tự với nhóm khách hàng 1 sau khi phân cụm RFM. Từ đó, chúng tôi đưa ra một số chiến lược như sau:

Insights	Nhóm khách hàng chi tiêu nhiều và thường xuyên, thu nhập cao, tỷ lệ con ít, chi tiêu cho rượu và thức ăn ngang nhau.
Mục tiêu	Tập trung đẩy mạnh tần suất mua hàng và xây dựng lòng trung thành
Chiến lược	- Tiến hành các chương trình ưu đãi cho khách hàng làm thẻ thành viên, tăng cường after-sale services. - Đẩy mạnh kết nối với khách hàng (vd thông qua mail marketing, personal contacts....) - Thu thập thông tin, tìm hiểu tại sao khách hàng không còn mua nữa

NỘI DUNG PHÂN TÍCH

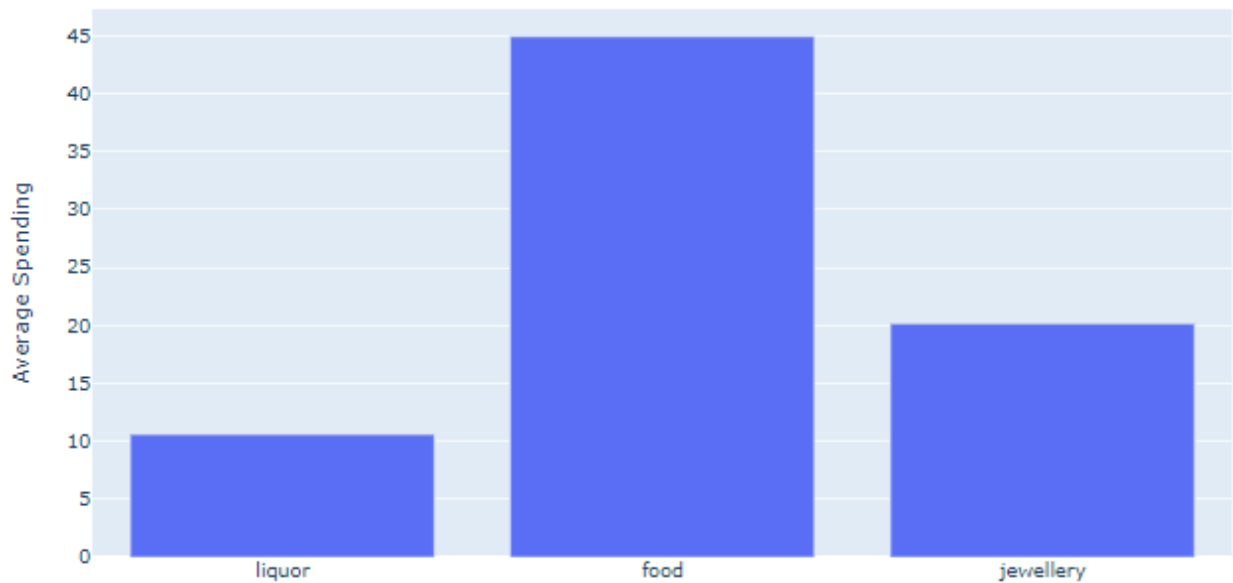
REFERENCES

1. Tomorrow Marketer (2023). Phân tích RFM là gì và các bước phân khúc khách hàng theo RFM. [online] Truy cập tại: <https://blog.tomorrowmarketers.org/phan-tich-rfm-la-gi/> . (Truy cập ngày 25/05/2024)
2. Medium (2020). RFM Analysis. [online] Truy cập tại: <https://medium.com/@denizcansuturan/rfm-analysis-3930a9a5238> . (Truy cập ngày 25/05/2024)

NỘI DUNG PHÂN TÍCH

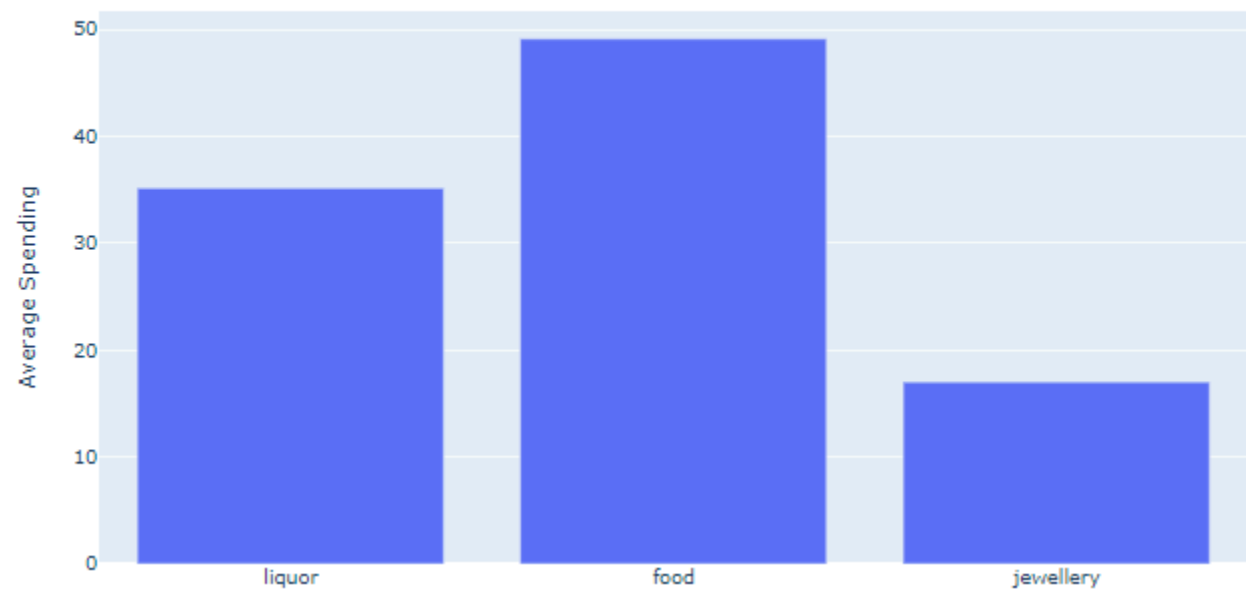
APPENDIX

Average Spending in <20K Income Segment



Appendix 2.1: Chi tiêu trung bình của nhóm người có thu nhập dưới 20K/năm

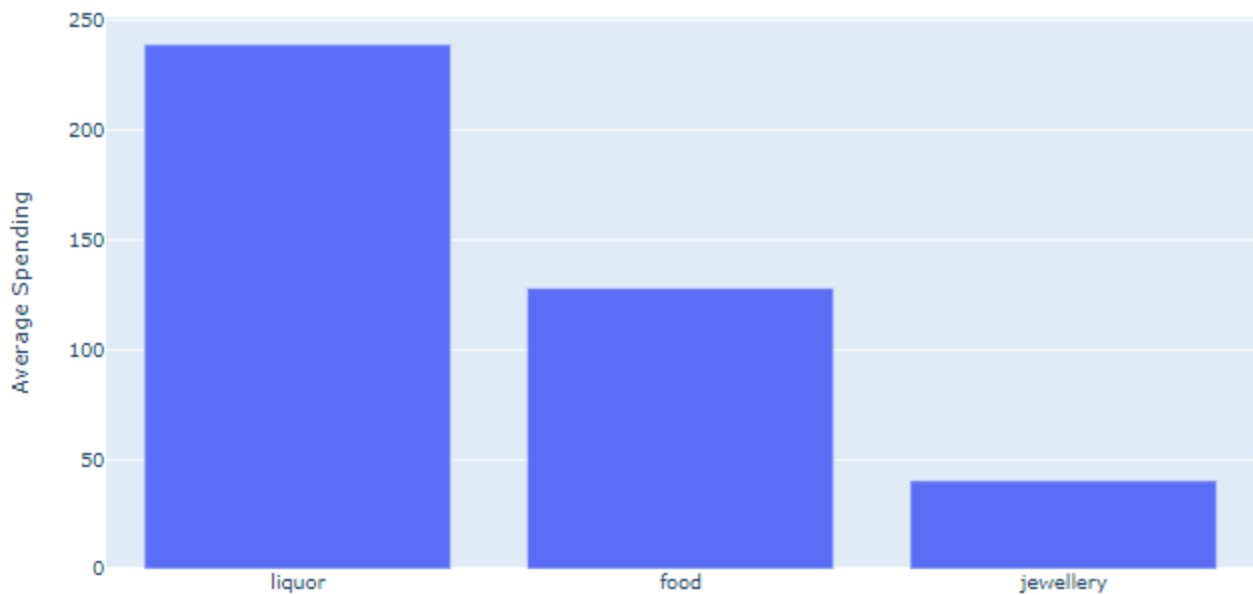
Average Spending in 20-40K Income Segment



Appendix 2.2: Chi tiêu trung bình của nhóm người có thu nhập từ 20-40K/năm

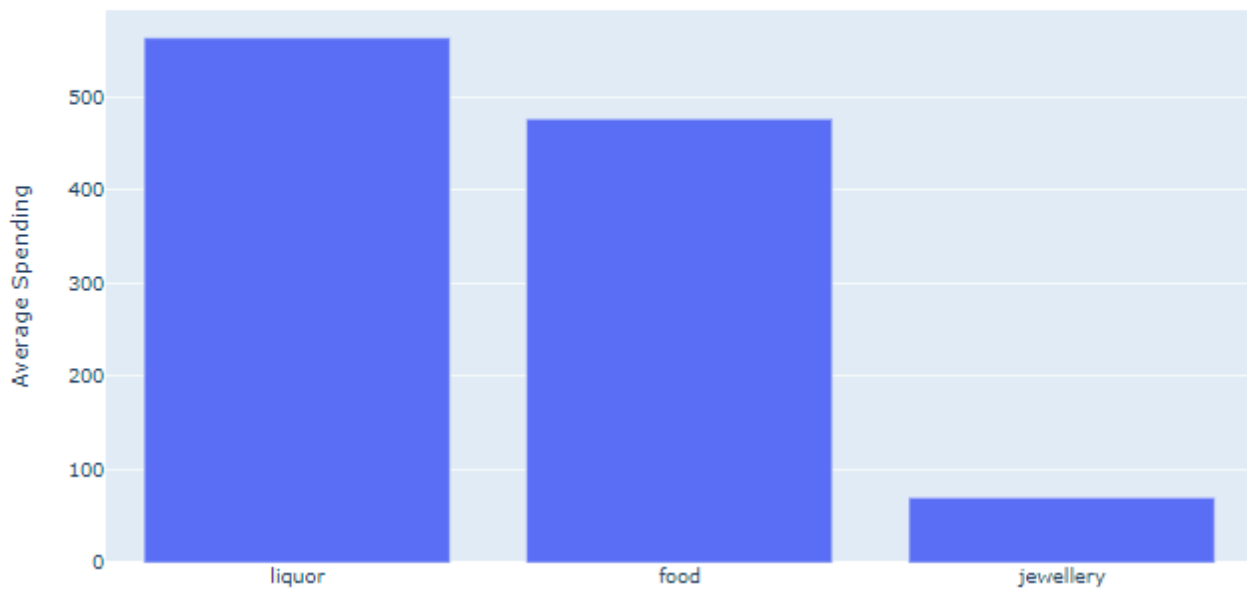
NỘI DUNG PHÂN TÍCH

Average Spending in 40-60K Income Segment



Appendix 2.3: Chi tiêu trung bình của nhóm người có thu nhập từ 40-60K/năm

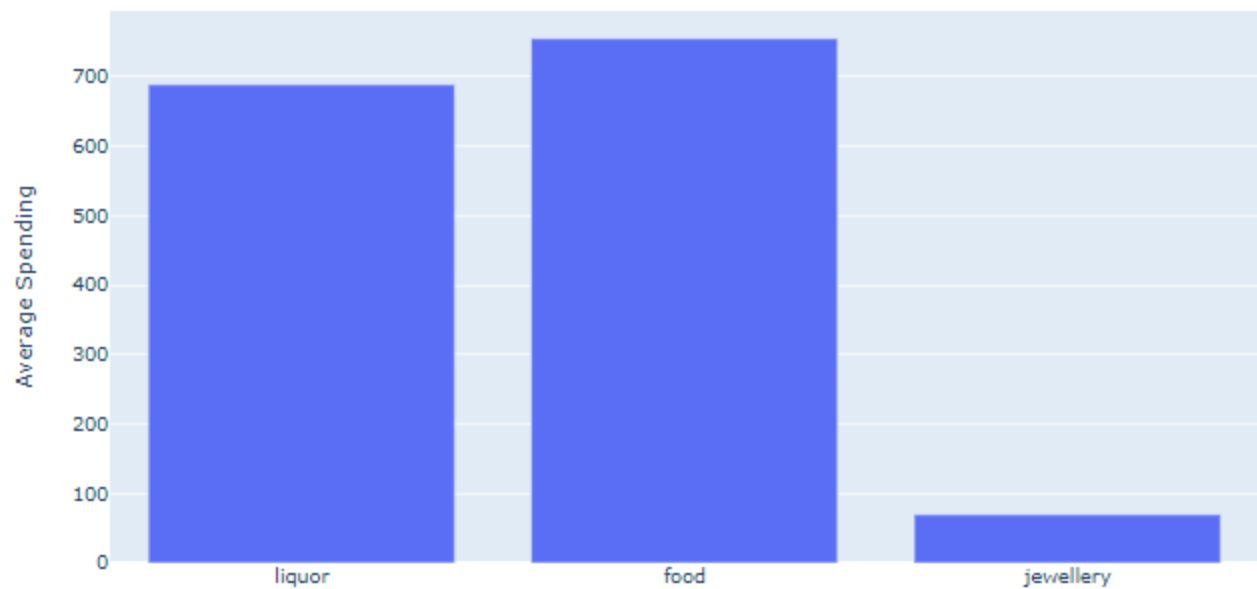
Average Spending in 60-80K Income Segment



Appendix 2.4: Chi tiêu trung bình của nhóm người có thu nhập từ 60-80K/năm

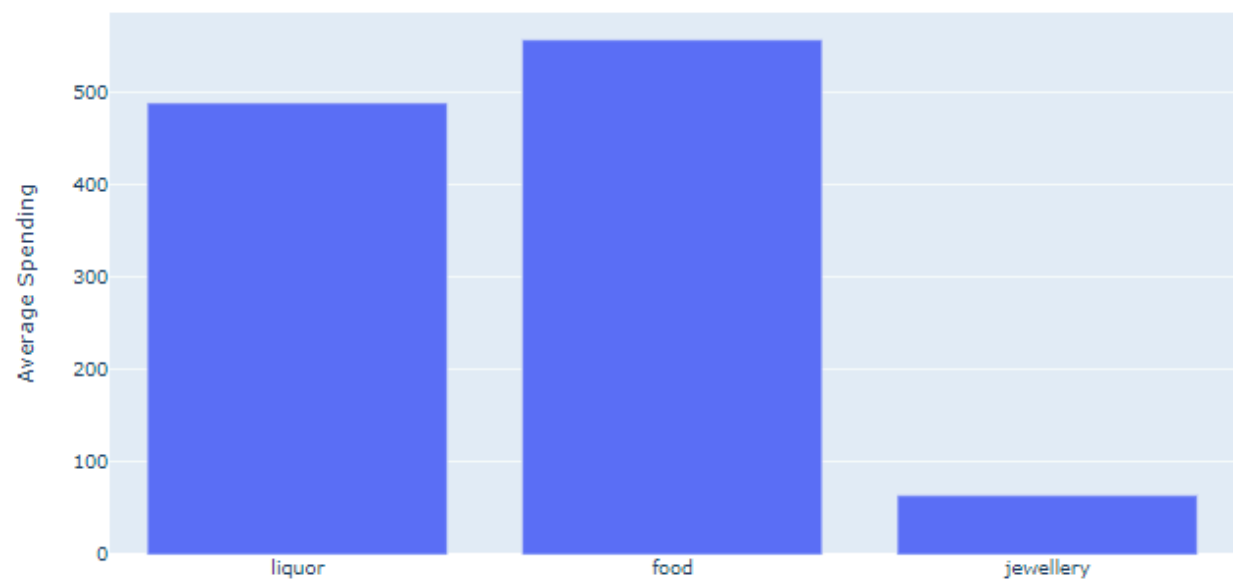
NỘI DUNG PHÂN TÍCH

Average Spending in >80K Income Segment



Appendix 2.5: Chi tiêu trung bình của nhóm người có thu nhập trên 80K/năm

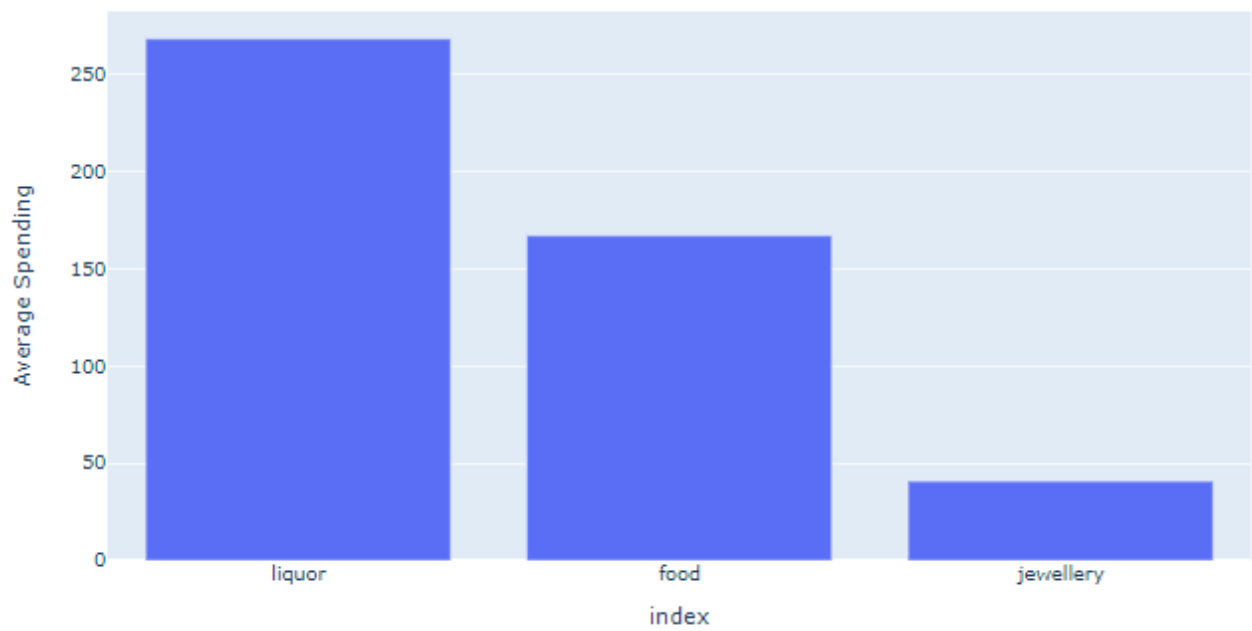
Average Spending for People who do not have Child



Appendix 2.6: Chi tiêu trung bình của những người chưa có con

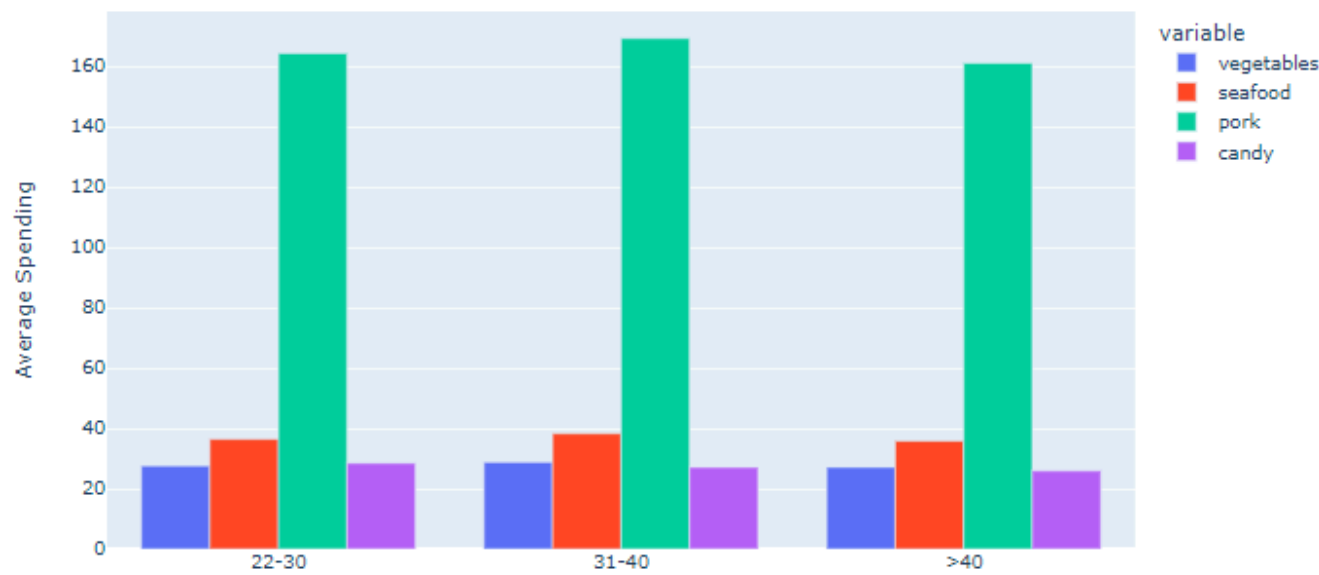
NỘI DUNG PHÂN TÍCH

Average Spending of People who have Children



Appendix 2.7: Chi tiêu trung bình của những người có con

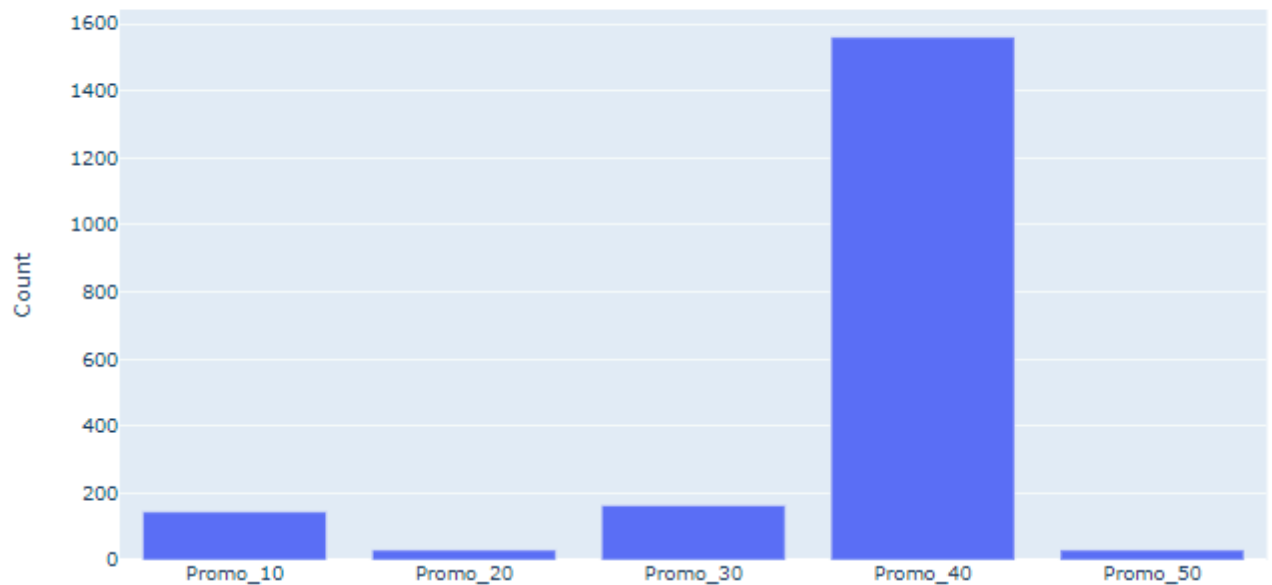
Correlation between Age Group and Spending on Food Categories



Appendix 2.8: Tương quan giữa từng nhóm tuổi và loại thực phẩm

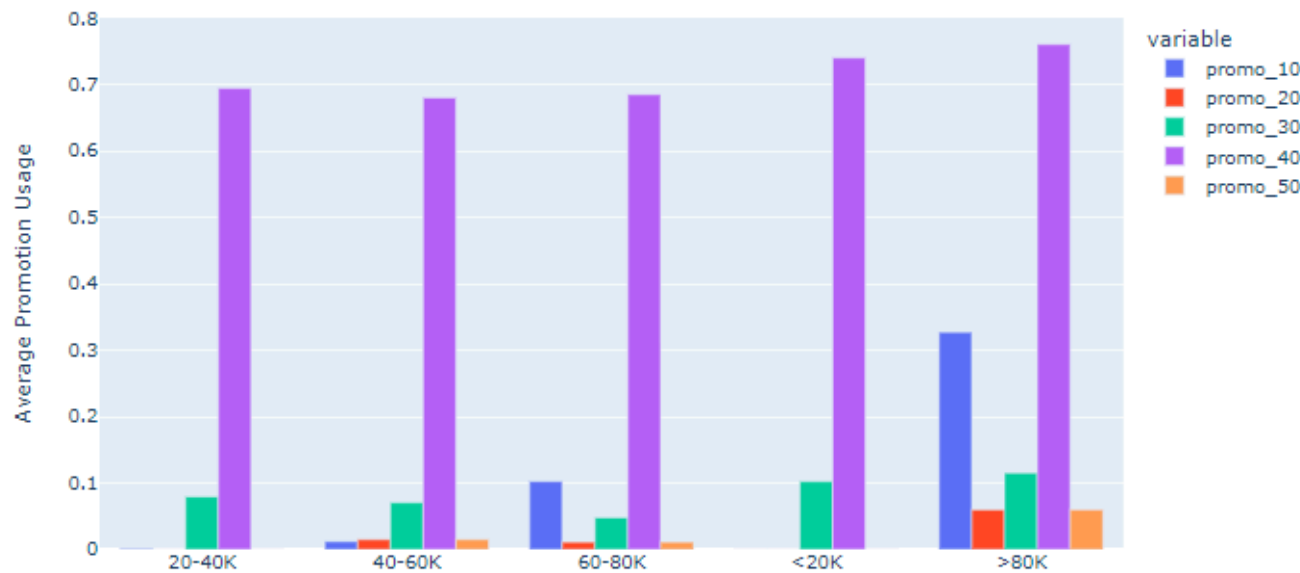
NỘI DUNG PHÂN TÍCH

Number of Purchases using Promotion



Appendix 2.9: Số lượng giao dịch sử dụng khuyến mãi

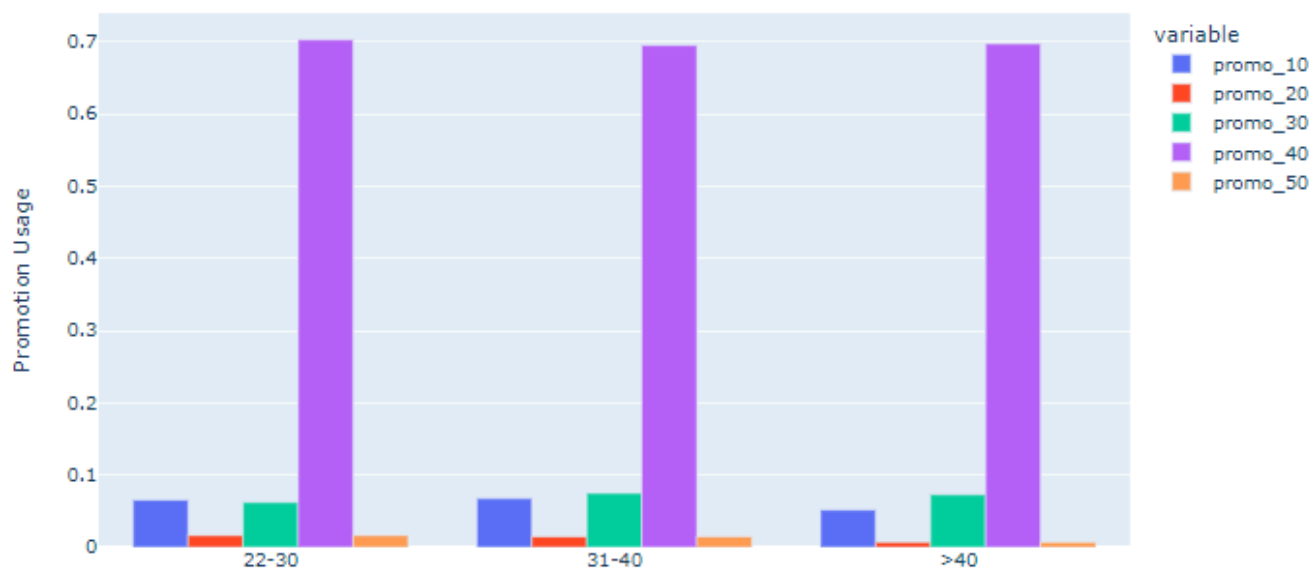
Correlation between Income Group and Promotion Usage



Appendix 2.10: Tương quan giữa nhóm thu nhập và lượt sử dụng khuyến mãi

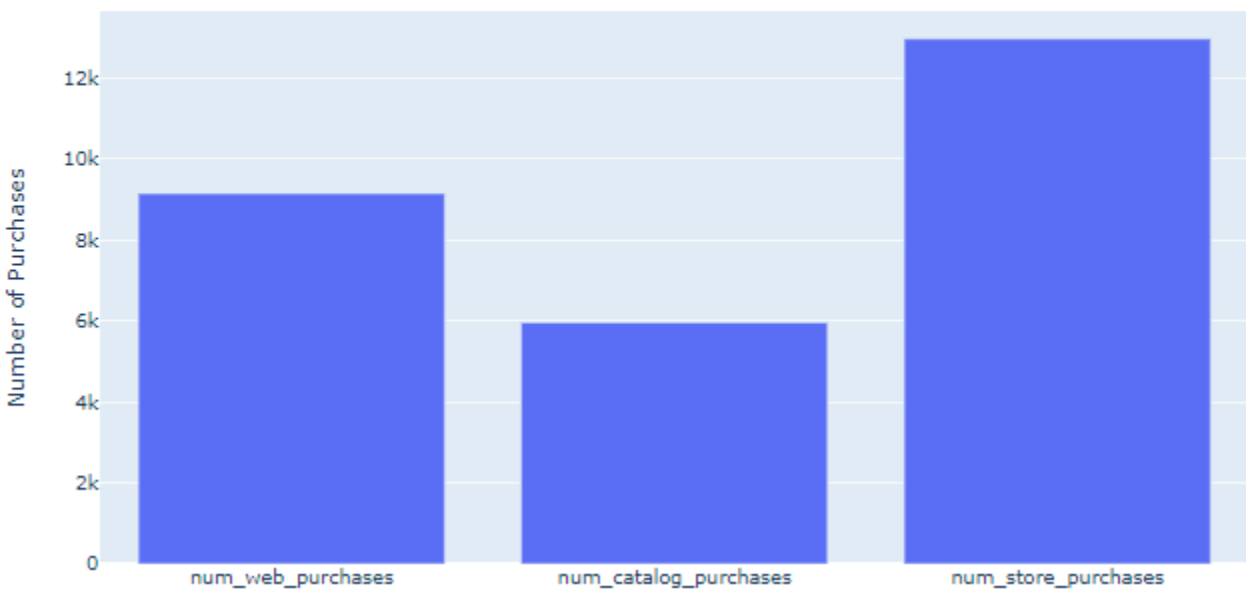
NỘI DUNG PHÂN TÍCH

Promotion Usage by Age Group



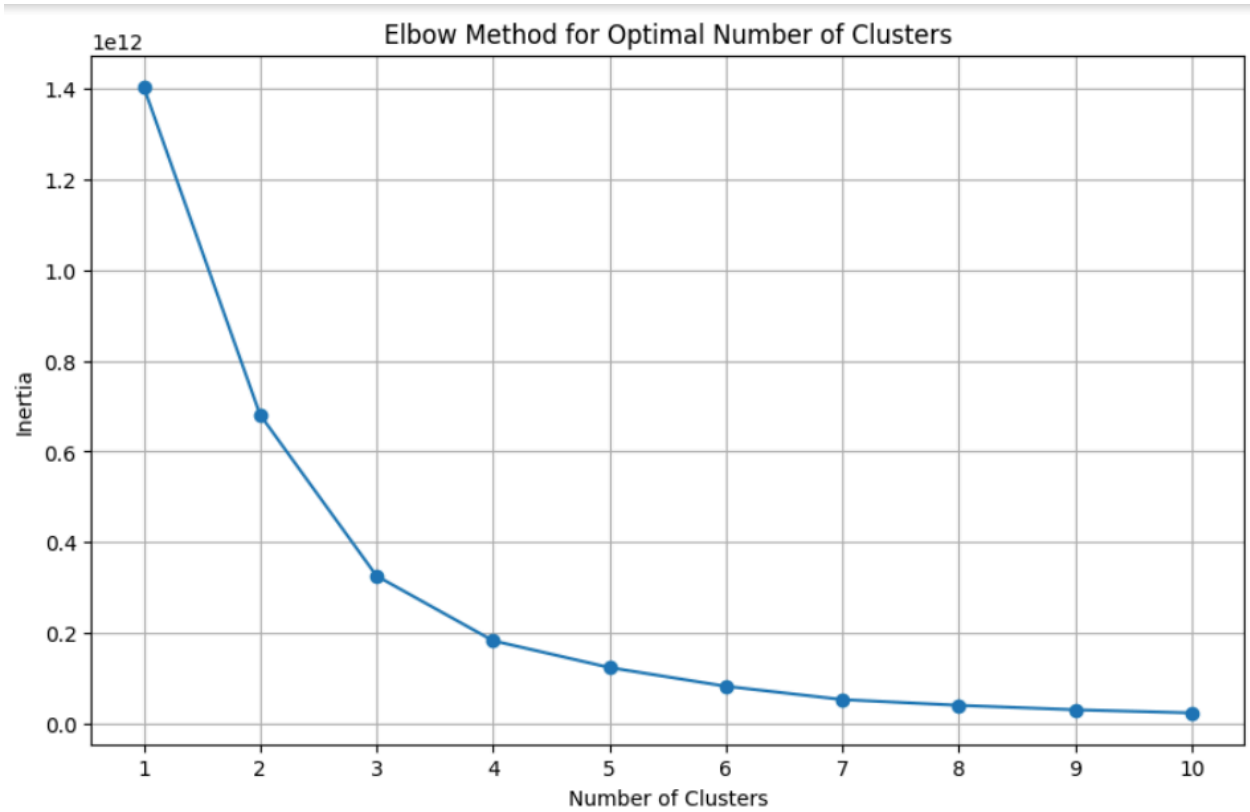
Appendix 2.11: Lượt sử dụng từng loại khuyến mãi theo từng nhóm tuổi

Number of Purchases through Different Channels

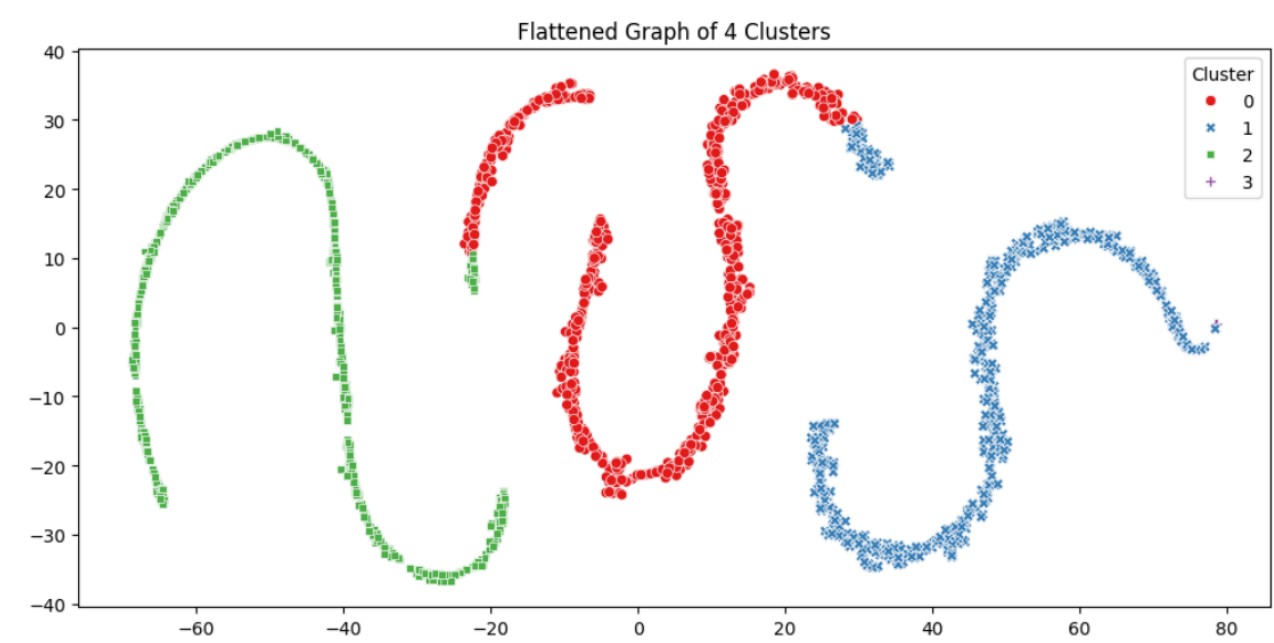


Appendix 2.12: Số lượng giao dịch qua từng kênh mua hàng khác nhau

NỘI DUNG PHÂN TÍCH



Appendix 4.1. Kết quả sử dụng Elbow method nhằm tìm số k nhóm tối ưu



Appendix 4.2. Phân nhóm khách hàng 4 cụm sử dụng K-means clustering

NỘI DUNG PHÂN TÍCH



	income	children	recency	frequency	liquor	food	candy	jewellery
Cluster								
0	52344.65	1.23	49.72	16.13	290.65	146.85	18.19	45.89
1	76967.84	0.44	49.02	20.86	616.96	539.38	60.20	70.17
2	28331.94	1.12	48.56	7.90	31.33	42.68	6.06	17.74
3	666666.00	1.00	23.00	11.00	10.00	42.00	1.00	12.00

Appendix 4.3. Kết quả tổng hợp phân nhóm khách hàng thành 4 cụm sử dụng K-mean clustering