

**TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN
KHOA TOÁN KINH TẾ**

-----**-----



BÀI TẬP LỚN
MÔN DATA VISUALIZATION

Đề tài: Phân khúc khách hàng – Xu hướng người tiêu dùng.

Sinh viên: Bùi Thị Mai Lương – 11193191

Trần Thế Phong – 11194149

Nguyễn Bá Đăng Khôi - 11192597

Lớp : DSEB

Mở đầu

Nghiên cứu thị trường luôn là bước quan trọng để mỗi công ty có thể hiểu được thị trường, hiểu được khách hàng họ đang hướng đến. Trục quan hóa dữ liệu là phương pháp quan trọng để hoàn thành những dự án nghiên cứu. Vì vậy với bài tập lớn này, nhóm 8 dùng EDA là phương pháp chính để hoàn thành ý tưởng nghiên cứu. Nhóm 8 dùng tình huống của doanh nghiệp làm đề tài hoàn thành bài báo cáo.

Tình huống: Một công ty bán thực phẩm đã mở được 5 năm. Trong 3 năm đầu họ nỗ lực và mang lại doanh thu khá tốt cho công ty; nhưng trong vòng 2 năm tiếp theo lợi nhuận thu lại không được như mong muốn do lượng khách hàng ít đi mặc dù đã chạy một số chương trình marketing do đối thủ cạnh tranh ngày một nhiều. Do đó họ làm nghiên cứu thị trường. Và nhóm 8 làm một phần trong báo cáo tổng thể của nghiên cứu thị trường đó: nghiên cứu khách hàng thành viên trong vòng 2 năm qua của công ty và xu hướng tiêu dùng của họ.

- Bối cảnh (S): Đối thủ cạnh tranh nhiều.
- Vấn đề (C): Khách hàng ít hơn và lượng tiêu thụ sản phẩm ít hơn.
- Câu hỏi (Q): Làm sao để có thể thu hút lại khách hàng được như trước.
- Trả lời (A):
 - + Tìm ra phân khúc khách hàng hợp lý.
 - + Xu hướng tiêu dùng khách hàng.
 - + Chiến dịch marketing hiệu quả.

Báo cáo chi tiết

Nghiên cứu thị trường : Phân khúc khách hàng đi mua thực phẩm

Phần 1 : Thông tin nghiên cứu.

Phần 2: Mục tiêu nghiên cứu.

Phần 3 : Báo cáo chi tiết.

Phần 4: Kiến nghị dựa trên kết quả nghiên cứu.

Phần 1: Thông tin nghiên cứu

1. Thông tin khảo sát.

- Mẫu khảo sát: 30/07/2017 – 30/06/2019.
- Tổng mẫu nghiên cứu: 2240.
- Độ tuổi: Trên 20 tuổi.
- Điều kiện nghiên cứu: Những khách hàng đăng ký thành viên.
- Mục tiêu nghiên cứu: Phân khúc khách hàng và xu hướng mua hàng của khách hàng.

2. Bộ dữ liệu

Khách hàng

- ID: Mã định danh khách hàng.
- Year_Birth: Năm sinh khách hàng.
- Education: Trình độ học vấn của khách hàng.
- Marital_Status: Tình trạng hôn nhân khách hàng.
- Income: Thu nhập hộ gia đình hàng năm của khách hàng.
- Kidhome: Số trẻ em trong hộ gia đình của khách hàng.
- Teenhome: Số thanh thiếu niên trong hộ gia đình của khách hàng.
- Dt_Customer: Ngày khách hàng đăng ký thành viên của công ty.
- Recency: Số ngày kể từ lần mua hàng cuối cùng của khách hàng.
- Complain: 1 nếu khách hàng phàn nàn trong 2 năm qua, 0 nếu khách hàng không phàn nàn.

Sản phẩm

- MntWines: Số tiền chi cho rượu trong 2 năm.
- MntFruits: Số tiền chi cho hoa quả trong 2 năm.
- MntMeatProducts: Số tiền chi cho thịt trong 2 năm.
- MntFishProducts: Số tiền chi cho cá trong 2 năm.
- MntSweetProducts: Số tiền chi cho đồ ngọt trong 2 năm.
- MntFrozenCommodity: Số tiền chi cho đồ đông lạnh trong 2 năm.

Quảng bá

- NumDealsPurchases: Số lần mua hàng được giảm giá
- AcceptedCmp1: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch đầu tiên, 0 nếu không chấp nhận.
- AcceptedCmp2: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ hai, 0 nếu không chấp nhận.
- AcceptedCmp3: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ ba, 0 nếu không chấp nhận.
- AcceptedCmp4: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ tư, 0 nếu không chấp nhận.
- AcceptedCmp5: 1 nếu khách hàng chấp nhận ưu đãi trong chiến dịch thứ năm, 0 nếu không chấp nhận.
- Response: 1 nếu khách hàng chấp nhận đề nghị trong chiến dịch cuối cùng, 0 nếu không.

Địa điểm

- NumWebPurchases: Số lần mua hàng được thực hiện thông qua trang web của công ty.
- NumCatalogPurchases: Số lần khách hàng mua hàng qua Catalog.
- NumStorePurchases: Số lần khách hàng mua trực tiếp tại cửa hàng.
- NumWebVisitsMonth: Trung bình số lượt truy cập vào trang web của công ty mỗi tháng.

Phần 2: Mục tiêu nghiên cứu

1. Phân khúc khách hàng.
2. Nhu cầu mua hàng của khách hàng.
3. Hình thức khách hàng mua hàng của mỗi phân khúc.

Phần 3: Báo cáo chi tiết

1. Tìm phân khúc khách hàng.

1.1. Làm sạch dữ liệu.

1.2. Xử lý dữ liệu.

1.3. Dựa vào thông tin phân loại khách hàng.

2. Xu hướng khách hàng.

2.1. Phân tích sản phẩm.

2.2. Tác động chiến dịch marketing đến khách hàng.

2.3. Xu hướng mua sản phẩm của khách hàng.

1. Phân khúc khách hàng.

1.1. Làm sạch dữ liệu.

1.1.1. Tìm số lượng giá trị duy nhất có trong mỗi cột.

```
#Tìm số lượng giá trị duy nhất có trong mỗi cột  
data.nunique()
```

ID	2240
Year_Birth	59
Education	5
Marital_Status	8
Income	1973
Kidhome	3
Teenhome	3
Dt_Customer	663
Recency	100
MntWines	776
MntFruits	158
MntMeatProducts	558
MntFishProducts	182
MntSweetProducts	177
MntFrozenCommodity	213
NumDealsPurchases	15
NumWebPurchases	15
NumCatalogPurchases	14
NumStorePurchases	14
NumWebVisitsMonth	16
AcceptedCmp3	2
AcceptedCmp4	2
AcceptedCmp5	2
AcceptedCmp1	2
AcceptedCmp2	2
Complain	2
Z_CostContact	1
Z_Revenue	1
Response	2

Trong các cột dữ liệu, có biến “Z_CostContact” và “Z_Revenue” có cùng giá trị trong tất cả các hàng. Do đó hai biến này sẽ không ảnh hưởng đến việc xây dựng mô hình tìm phân khúc khách hàng.

1.1.2. Kiểm tra điểm dữ liệu trống.

```
data.isna().any()
```

ID	False
Year_Birth	False
Education	False
Marital_Status	False
Income	True
Kidhome	False
Teenhome	False
Dt_Customer	False
Recency	False
MntWines	False
MntFruits	False
MntMeatProducts	False
MntFishProducts	False
MntSweetProducts	False
MntFrozenCommodity	False
NumDealsPurchases	False
NumWebPurchases	False
NumCatalogPurchases	False
NumStorePurchases	False
NumWebVisitsMonth	False
AcceptedCmp3	False
AcceptedCmp4	False
AcceptedCmp5	False
AcceptedCmp1	False
AcceptedCmp2	False
Complain	False
Response	False

Nhận thấy, ở “Income” có điểm dữ liệu bị trống. Cần loại bỏ tất cả điểm dữ liệu trống.

1.1.3. Kiểm tra sự tương quan giữa các biến.

```
# ma trận tương quan
corrmat= data.corr()
plt.figure(figsize=(20,20))
sns.heatmap(corrmat,annot=True, cmap=cmap, center=0)
```



Vì không có giá trị $|\text{cor}| > 0.8$ nên không có hiện tượng tự tương quan giữa các biến với nhau.

Vậy không cần bỏ bất kì biến nào.

1.1.4. Tạo features mới và bỏ một số features không cần thiết để giảm kích thước và độ phức tạp.

- Tính số tuổi của khách hàng theo năm sinh của khách hàng đã đăng ký. Tạo biến “Age”.
- Tính tổng số tiền khách hàng đã chi tiêu các sản phẩm khác nhau trong khoảng thời gian 2 năm. Tạo biến “Spent”.
- Tạo biến chỉ ra hoàn cảnh sống khách hàng. Tạo biến : “Living_With” với hai giá trị “Alone” và “Partner”.
- Tính số lượng trẻ em có trong cách hộ gia đình, bao gồm: trẻ em (“Kidhome”) và thiếu niên (“Teenhome”). Tạo biến: “Children”
- Tính số lượng người có trong hộ gia đình của khách hàng. Tạo biến : “Family_Size”.
- Tạo ba danh mục trong “Education” bằng cách đơn giản hóa số lượng giá trị.

- Xóa bỏ một số biến không cần thiết để giảm kích thước và độ phức tạp của mô hình.

```
# Thiết kế features
# Tuổi tác hiện tại của khách hàng
data["Age"] = 2021-data[["Year_Birth"]]

# Tổng chi tiêu cho các mặt hàng khác nhau
data["Spent"] = data["MntWines"]+ data["MntFruits"]+ data["MntMeatProducts"]+
                data["MntFishProducts"]+data["MntSweetProducts"]+ data["MntGoldProds"]

# Xuất phát hoàn cảnh sống theo tình trạng hôn nhân
data["Living_With"]=data["Marital_Status"].replace({"Married":"Partner", "Together":"Partner",
                                                    "Absurd":"Alone", "Widow":"Alone", "YOLO":"Alone",
                                                    "Divorced":"Alone", "Single":"Alone",})

# Tính năng cho biết tổng số trẻ em sống trong hộ gia đình
data["Children"]=data["Kidhome"]+data["Teenhome"]

# Tính năng cho tổng số thành viên trong gia đình
data["Family_Size"] = data["Living_With"].replace({"Alone": 1, "Partner":2})+ data["Children"]

# Tính năng liên quan đến tình phụ tử (parenthood)
data["Is_Parent"] = np.where(data.Children> 0, 1, 0)

# Phân chia trình độ học vấn theo ba nhóm
data["Education"]=data["Education"].replace({"Basic":"Undergraduate", "2n Cycle":"Graduate",
                                              "Graduation":"Graduate", "Master":"Postgraduate", "PhD":"Postgraduate"})

# For clarity (Cho rõ ràng)
data=data.rename(columns={"MntWines": "wines", "MntFruits":"Fruits", "MntMeatProducts":"Meat",
                        "MntFishProducts":"Fish", "MntSweetProducts":"Sweets", "MntGoldProds":"Gold"})

# Xóa một số cột không cần thiết để giảm kích thước và độ phức tạp của mô hình.

col_del = ["AcceptedCmp1" , "AcceptedCmp2", "AcceptedCmp3" , "AcceptedCmp4", "AcceptedCmp5",
           "Response", "NumWebVisitsMonth", "NumWebPurchases", "NumCatalogPurchases", "NumStorePurchases",
           "NumDealsPurchases" , "Kidhome", "Teenhome"]
data=data.drop(columns=col_del,axis=1)
data.head()

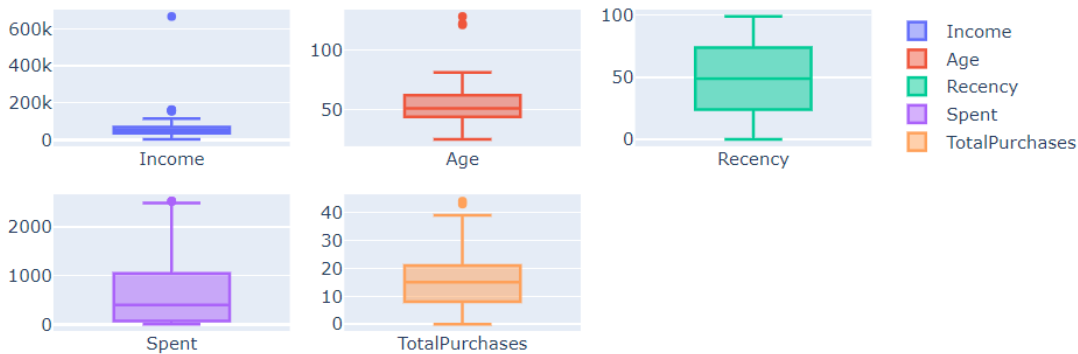
# Bỏ một số tính năng thừa
to_drop = ["Marital_Status", "Dt_Customer", "Year_Birth", "ID"]
data = data.drop(to_drop, axis=1)
```

1.1.5. Check outliers.

```
trace0 = go.Box(y=data['Income'],name='Income')
trace1 = go.Box(y=data['Age'],name='Age')
trace2 = go.Box(y=data['Recency'],name='Recency')
trace3 = go.Box(y=data['Spent'],name='Spent')
trace4 = go.Box(y=data['TotalPurchases'],name='TotalPurchases')
fig = make_subplots(rows=2, cols=3)
fig.append_trace(trace0, row = 1, col = 1)
fig.append_trace(trace1, row = 1, col = 2)
fig.append_trace(trace2, row = 1, col = 3)
fig.append_trace(trace3, row = 2, col = 1)
fig.append_trace(trace4, row = 2, col = 2)
fig.update_layout(width=800, height=400, title='Box Plot to check for outliers')

fig.show()
```

Box Plot to check for outliers



⇒ Có một vài outliers ở biến “Income”, “Age” và “TotalPurchases”.
Nguyên nhân: - Khách hàng ghi sai thông tin.
- Nhân viên ghi sai thông tin khách hàng.

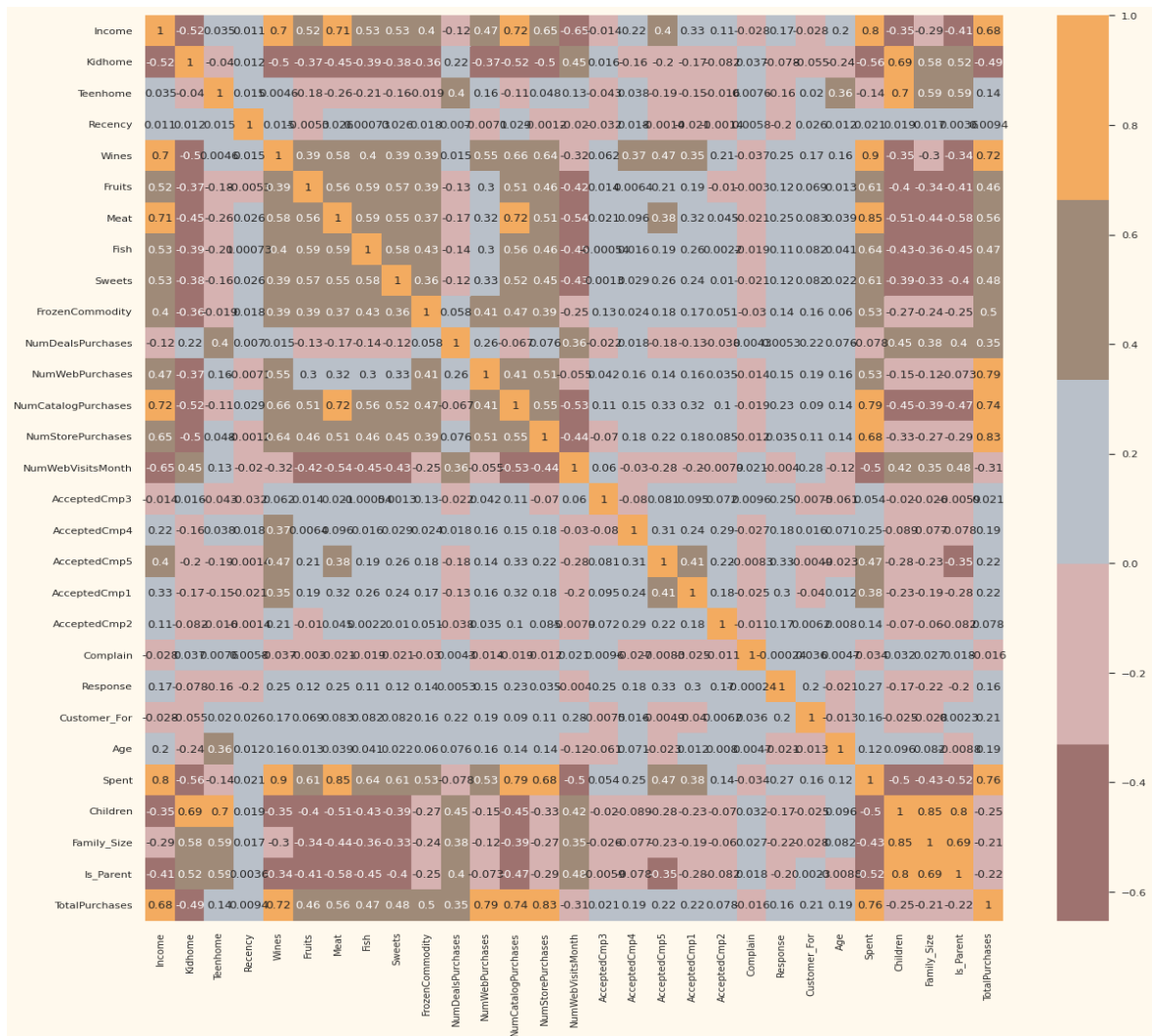
Cần loại bỏ những outlier đó để xây dựng mô hình chính xác.

```
# Loại bỏ các yếu tố ngoại lai bằng cách đặt giới hạn về tuổi, thu nhập, tổng số lần mua.
data = data[(data["Age"]<80)]
data = data[(data["Income"]<160000)]
data = data[(data["TotalPurchases"]<40)]
```

Sau khi loại bỏ những điểm outliers, tổng số điểm dữ liệu còn lại 2207.

1.1.6. Kiểm tra hiện tượng tự tương quan sau khi thay đổi feature trong dữ liệu.

```
# ma trận tương quan
corrmatrix = data.corr()
plt.figure(figsize=(20,20))
sns.heatmap(corrmatrix,annot=True, cmap=cmap, center=0)
```



Có một vài biến có lượng tương quan khá cao đến nhau ($|\text{cor}| > 0.8$), cần loại bỏ khỏi dữ liệu khi tiến hành xây dựng mô hình: “Meat”, “Wine”, “Family_Size”.

1.2. Xử lý dữ liệu.

Trong phần này, thực hiện xử lý trước dữ liệu để thực hiện các hoạt động phân cụm.

- Mã hóa nhãn cho các đối tượng phân loại.
- Chia tỷ lệ các biến bằng cách sử dụng bộ chia tỷ lệ tiêu chuẩn.
- Tạo khung dữ liệu tập hợp con để giảm kích thước.

1.2.1. Mã hóa nhãn cho các biến phân loại.

- Những biến phân loại trong dữ liệu.

```
#Những biến phân loại trong dữ liệu
s = (data.dtypes == 'object')
object_cols = list(s[s].index)

print("Biến phân loại trong bộ dữ liệu:", object_cols)
```

Có hai biến phân loại trong bộ dữ liệu : “ Education” và “ Living_With”.

- Mã hóa các đối tượng

```
#Label Encoding the object dtypes.
LE=LabelEncoder()
for i in object_cols:
    data[i]=data[[i]].apply(LE.fit_transform)
```

1.2.2. Scaling

```
# Tạo bản sao dữ liệu
ds = data.copy()

# Scaling
scaler = StandardScaler().fit_transform(ds.values)

scaled_ds = pd.DataFrame(scaler, index= ds.index,columns= ds.columns )
print("Tất cả các tính năng hiện đã được scale")
```

1.2.3. Giảm kích thước.

- Trong dữ liệu này có nhiều yếu tố trên cơ sở đó thực hiện phân loại cuối cùng.
 - Các yếu tố này cơ bản là thuộc tính hoặc tính năng. Số lượng tính năng càng cao thì càng khó làm việc. Nhiều tính năng trong số này có tương quan và do đó không cần thiết trong việc xử lý mô hình.
- ⇒ Cần giảm kích thước, tức loại bỏ một số tính năng đã chọn trước khi đưa qua bộ phân loại. Giảm thứ nguyên là quá trình giảm số lượng các biến ngẫu nhiên đang được xem xét, bằng cách thu được một tập hợp các biến chính.
- Phân tích thành phần chính (PCA) là một kỹ thuật để giảm kích thước của tập dữ liệu đó, tăng khả năng diễn giải nhưng đồng thời giảm thiểu mất mát thông tin.
 - Các bước trong phần này:
 - + Giảm kích thước (PCA).
 - + Vẽ khung dữ liệu đã giảm.
 - Giảm kích thước (PCA): Trong dự án này sẽ giảm chiều xuống còn 3.

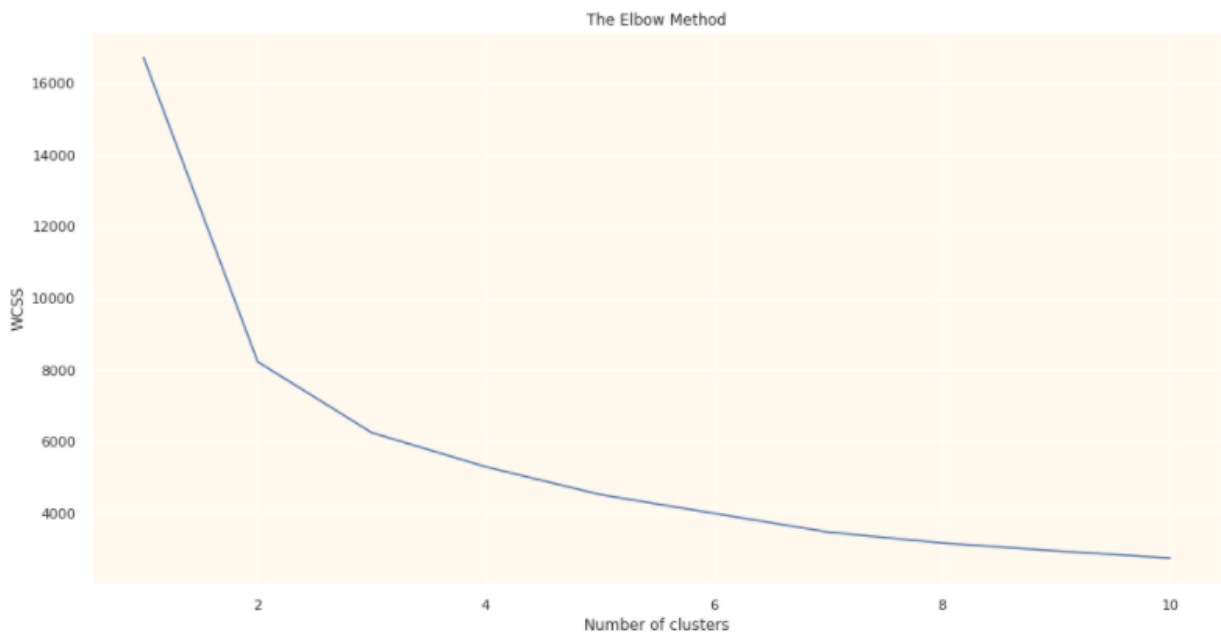
```
#Initiating PCA to reduce dimentions aka features to 3
pca = PCA(n_components=3)
pca.fit(scaled_ds)
PCA_ds = pd.DataFrame(pca.transform(scaled_ds), columns=([ "col1", "col2", "col3" ]))
PCA_ds.describe().T
```

	count	mean	std	min	25%	50%	75%	max
col1	2216.0	1.302608e-17	2.008417	-4.686471	-1.280422	-0.430781	1.835423	5.180277
col2	2216.0	2.945899e-17	1.342758	-3.263396	-1.183774	0.210313	0.976040	9.575968
col3	2216.0	2.655317e-17	1.096019	-3.131832	-1.140100	0.415669	0.891113	3.584629

1.2.4. Tìm K-Means

- **Phương pháp Elbow:**

```
wcss=[]  
for i in range (1,11):  
    kmeans=KMeans(n_clusters=i,init='k-means++',random_state=42)  
    kmeans.fit(PCA_ds)  
    wcss.append(kmeans.inertia_)  
plt.figure(figsize=(16,8))  
plt.plot(range(1,11),wcss, 'bx-')  
plt.title('The Elbow Method')  
plt.xlabel('Number of clusters')  
plt.ylabel('WCSS')  
plt.show()
```



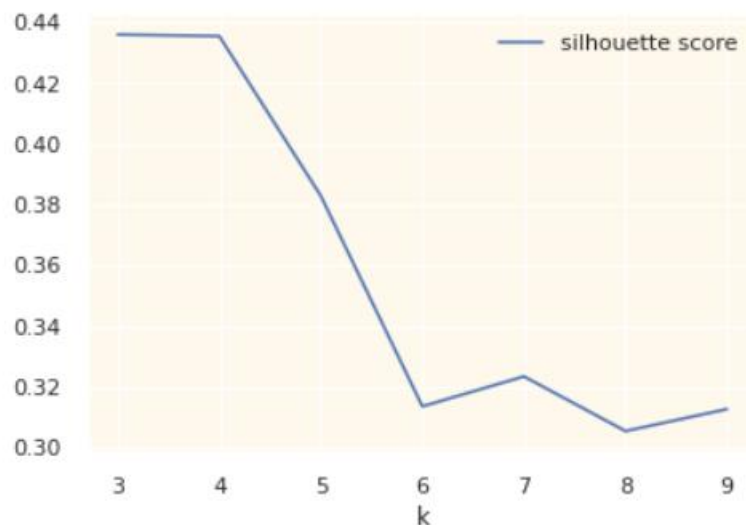
⇒ Nhìn vào biểu đồ thấy không được rõ ràng để chọn K bằng phương pháp Elbow.

- *Phương pháp Silhouette Score*

```
X = scaled_ds.values
from sklearn.metrics import silhouette_score
k_values= []
silhouette_score_values = []

for i in range(3,10):
    kmeans = KMeans(n_clusters = i)
    y_pred = kmeans.fit_predict(PCA_ds)
    k_values.append(i)
    silhouette_score_values.append(silhouette_score(PCA_ds,y_pred))

result = pd.DataFrame({'k': k_values, 'silhouette score': silhouette_score_values})
result.plot.line(x='k', y='silhouette score')
plt.show()
```



```
from sklearn.metrics import silhouette_score
kmeans = KMeans(n_clusters=4)
y_pred = kmeans.fit_predict(PCA_ds)
PCA_ds["clusters"] = y_pred
print(silhouette_score(PCA_ds,y_pred))
```

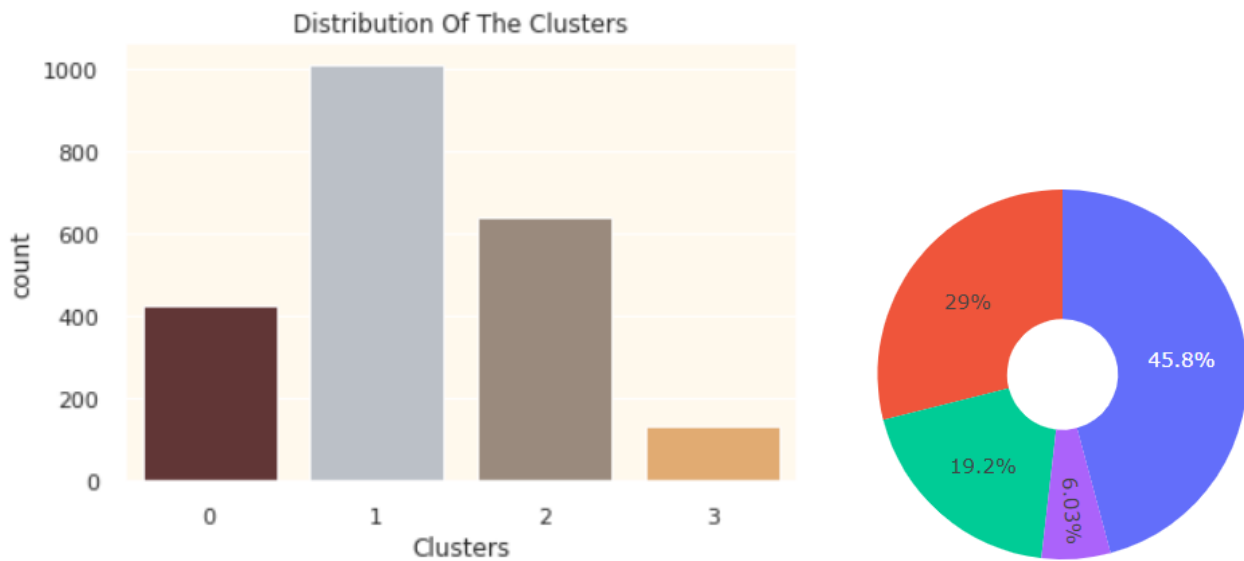
0.48604542512618226

Ở phương pháp này, silhouette score có giá trị trong khoảng từ [-1;1].

Khi silhouette score càng gần 1 thì sự phân cụm càng rõ ràng, và các cụm cách xa nhau.

Dựa vào phương pháp Silhouette Score chọn k = 4.

⇒ Dữ liệu khách hàng được phân thành 4 cụm.



Từ biểu đồ, công ty có phần lớn là khách hàng cụm 1, chiếm 45.8% số lượng khách hàng. Cụm 0: 19.2%, Cụm 2: 29%, Cụm 3: 6.03%

1.3. Dựa vào thông tin phân loại khách hàng

1.3.1. Dựa vào thông tin Chi tiêu và Thu nhập.

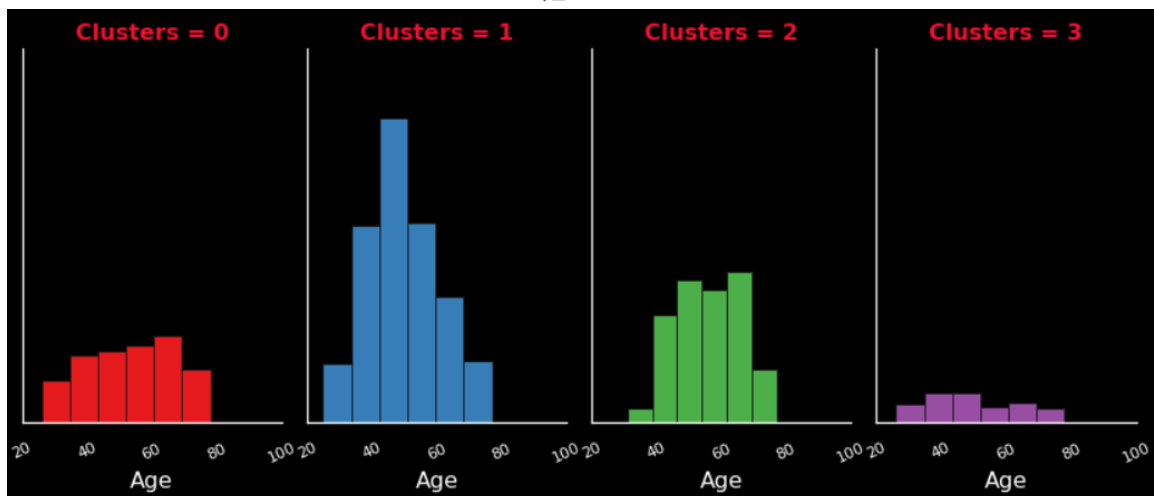
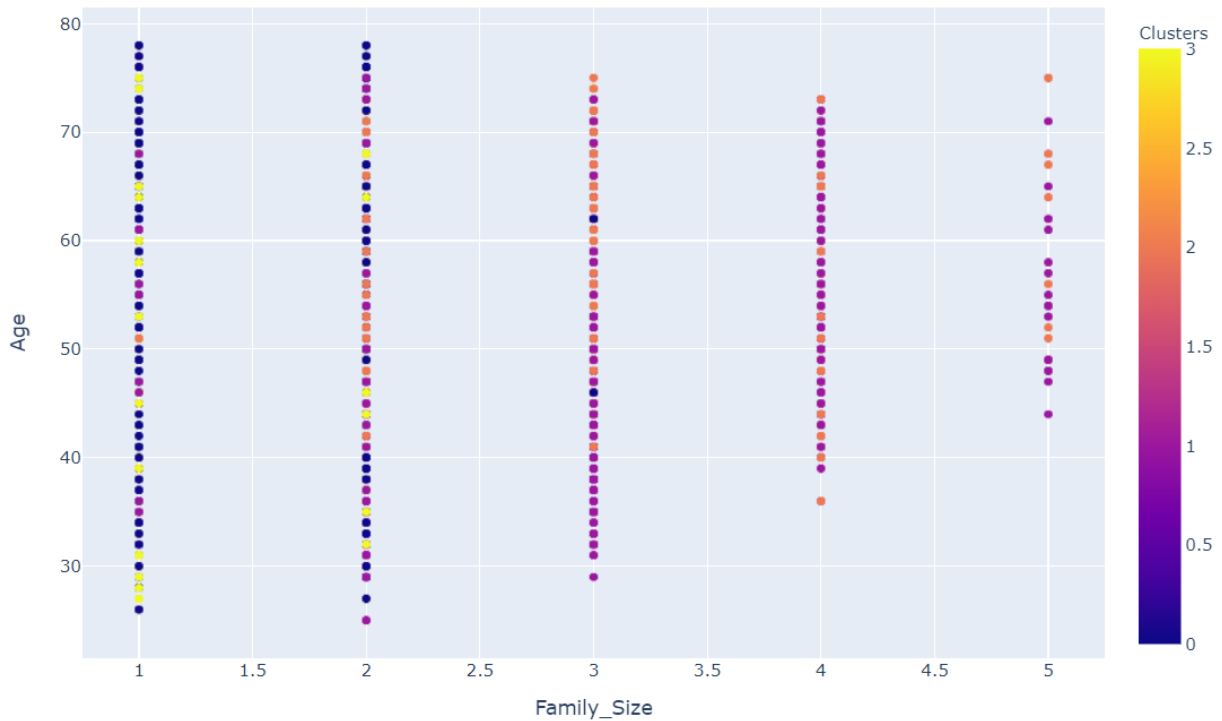


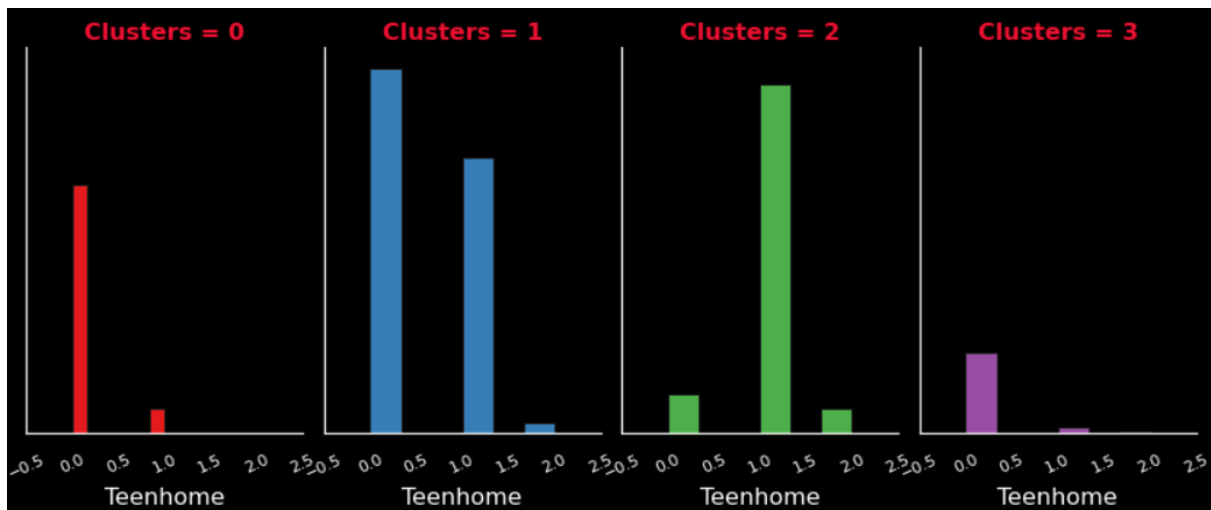
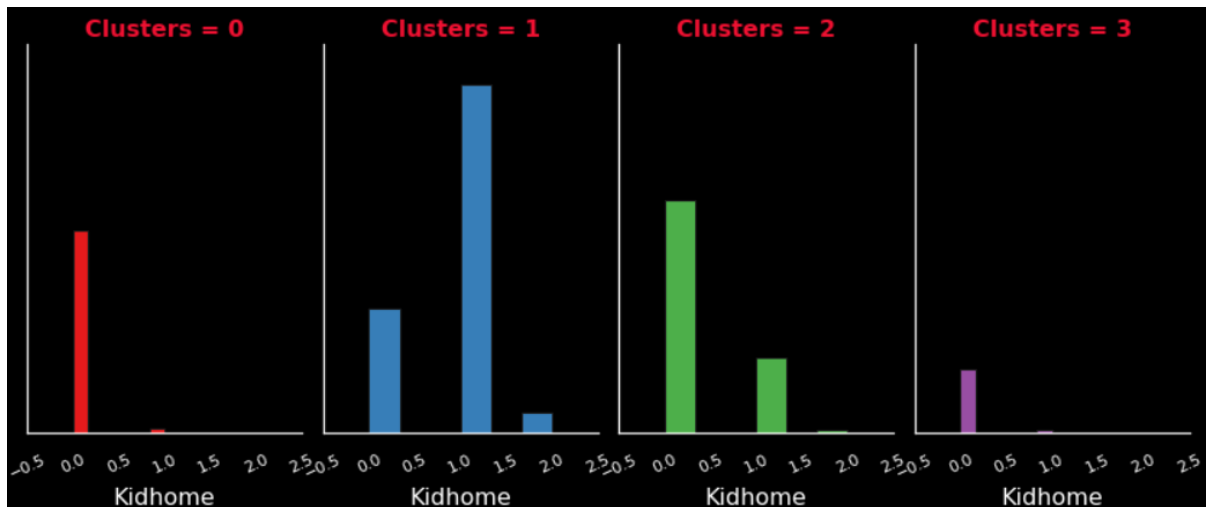
Theo biểu đồ:

- Cluster 1: Có mức chi tiêu và thu nhập thấp.
- Cluster 2: Có mức chi tiêu và thu nhập trung bình.
- Cluster 0,3: Có mức chi tiêu và thu nhập cao.

1.3.2. Dựa vào những thông tin gia đình và tuổi.

Family_Size vs Age Plot





- Nhóm 0: Khách hàng trong khoảng 30-70 tuổi, gia đình chủ yếu có từ 1-2 thành viên. Độc thân hoặc sống một mình.
- Nhóm 1: Phần lớn khách hàng có độ tuổi từ 30-60 tuổi, gia đình chủ yếu từ 3-5 thành viên (đa số khách hàng đều có 1 con nhỏ và 1 trẻ vị thành niên trong gia đình).
- Nhóm 2: Khách hàng từ 40- gần 70 tuổi, gia đình chủ yếu từ 2-3 thành viên (đa số khách hàng đều có 1 con nhỏ hoặc 1 trẻ vị thành niên trong gia đình).
- Nhóm 3: Khách hàng từ 30- 75 tuổi, gia đình có 1-2 thành viên, họ chủ yếu sống độc thân.

- **Thông tin của từng cụm:**

- 1. Nhóm 0:

- Khách hàng trong khoảng 30-70 tuổi.
 - Mỗi quan hệ: Đa số là độc thân.
 - Gia đình: có 1-2 người.
 - Nhóm có mức chi tiêu và thu nhập cao.

- 2. Nhóm 1:

- Tập trung chủ yếu khách hàng có độ tuổi từ 40-70.
 - Mỗi quan hệ: Chủ yếu đang trong mối quan hệ và chủ yếu là ba mẹ.
 - Khách hàng chủ yếu có từ 1-2 con.
 - Tối đa 5 thành viên trong gia đình và ít nhất 2.
 - Nhóm có mức chi tiêu và thu nhập thấp.

- 3. Nhóm 2:

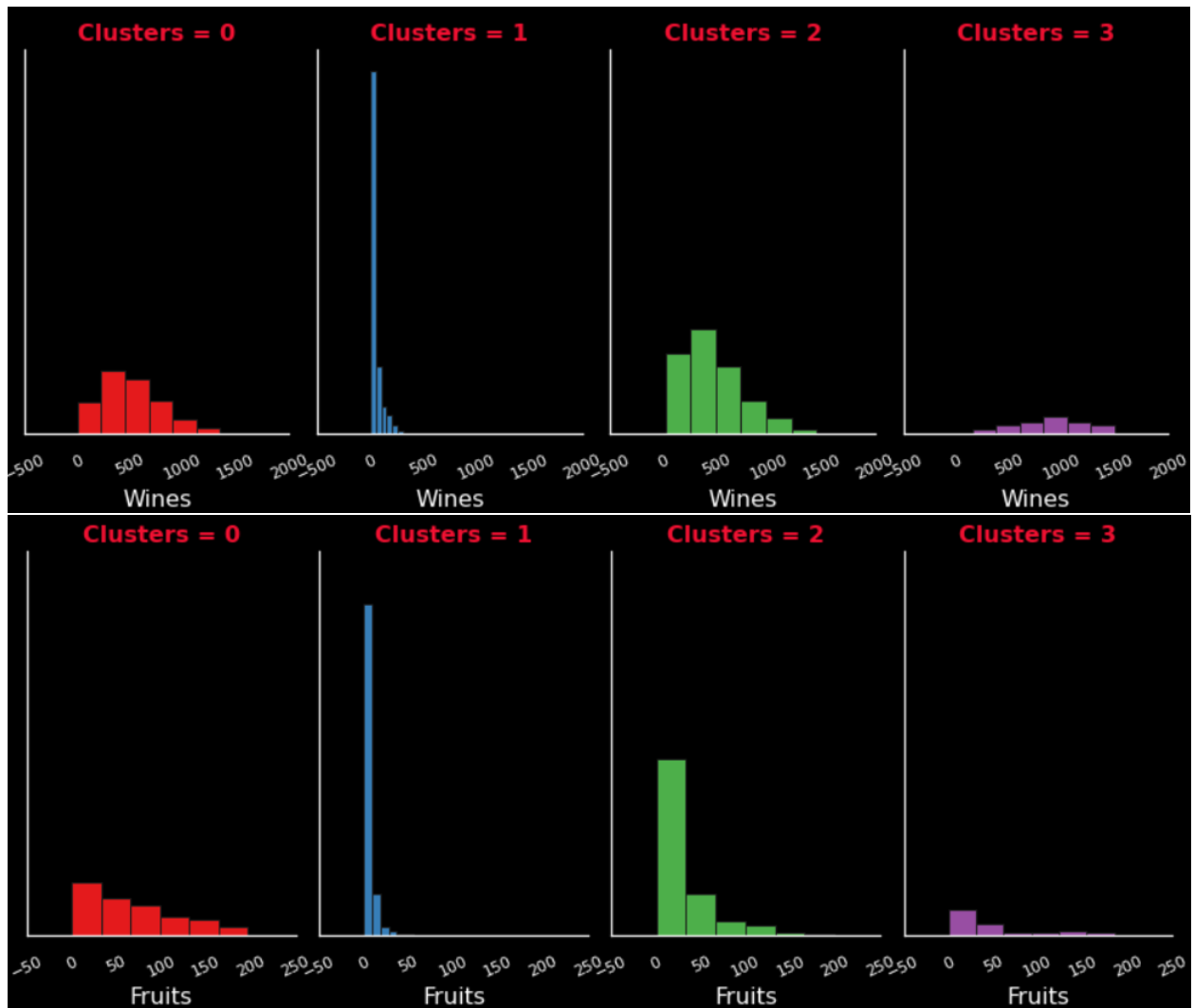
- Khách hàng từ 40- gần 70 tuổi.
 - Mỗi quan hệ: Phần lớn đã kết hôn và có con.
 - Gia đình chủ yếu từ 2-3 thành viên (đa số khách hàng đều có 1 con nhỏ hoặc 1 trẻ vị thành niên trong gia đình).
 - Nhóm có mức chi tiêu và thu nhập trung bình.

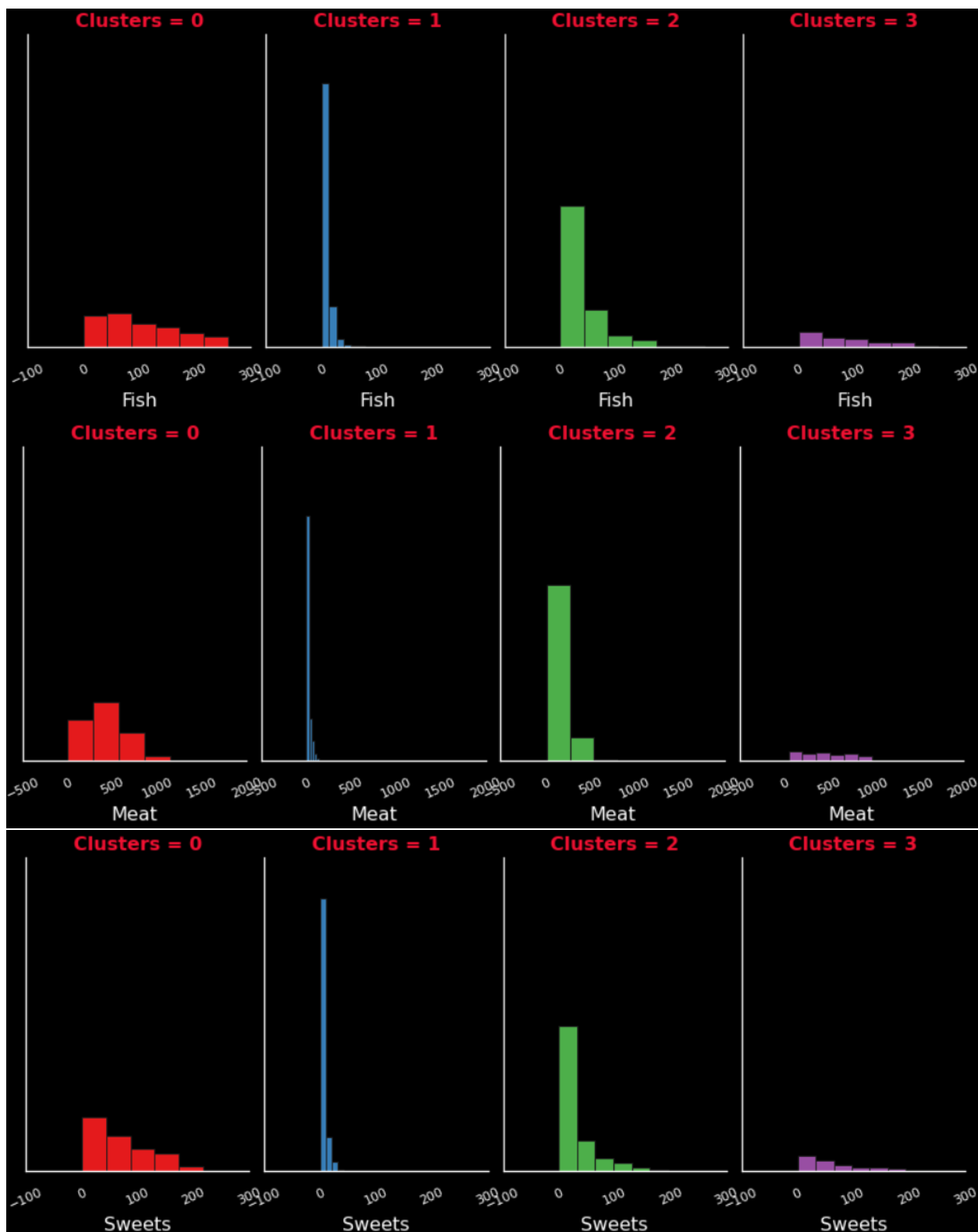
- 4. Nhóm 3:

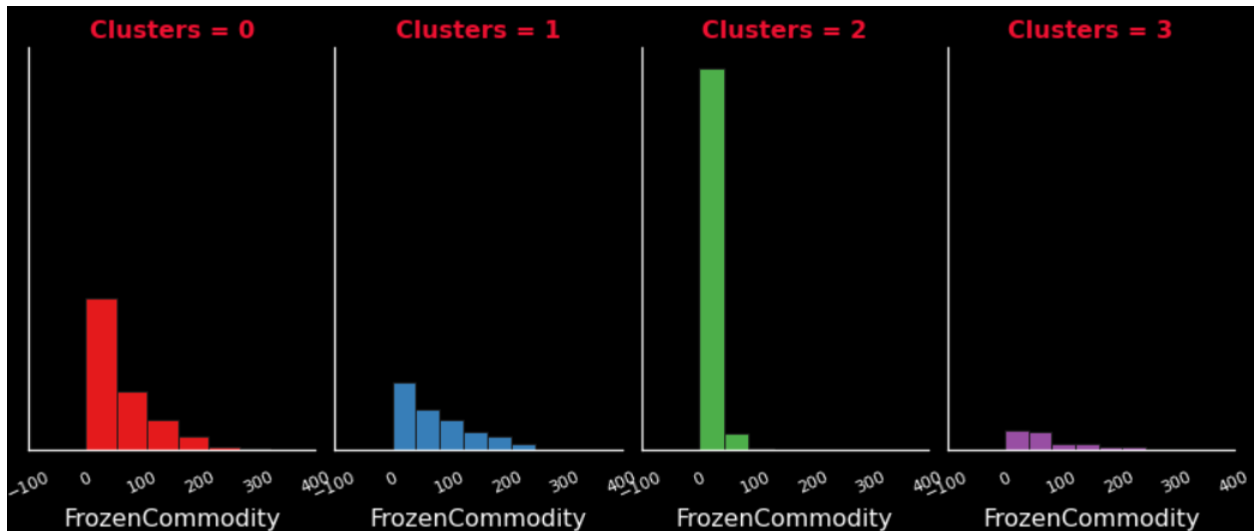
- Khách hàng từ 25- 75 tuổi.
 - Mỗi quan hệ: Độc thân.
 - Gia đình có 1 thành viên.
 - Nhóm có mức chi tiêu và thu nhập cao.

2. Xu hướng khách hàng.

2.1. Phân tích sản phẩm.





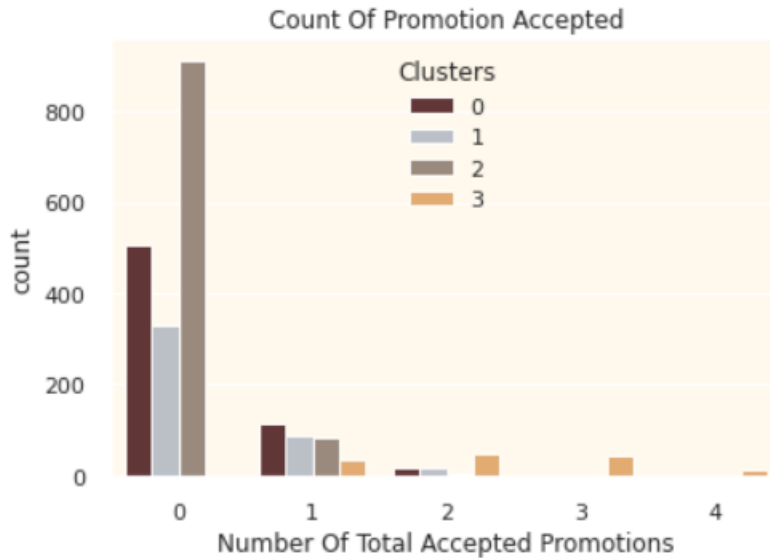


Xu hướng mua hàng của các nhóm:

- Nhóm 0 : vì độc thân lại có mức thu nhập cao nên họ chi tiêu cho thực phẩm với số tiền nhiều.
 - + Tất cả mặt hàng trong công ty họ đều mua nhưng họ chỉ cho những sản phẩm đó số tiền không quá nhiều trong 2 năm qua.
 - + Họ chi tiêu khá nhiều cho mặt hàng Thịt, Rượu và Cá.
- Nhóm 1: Họ có con cái nhưng mức thu nhập ở tầm thấp.
 - + Trong 2 năm qua, có khá nhiều người mua thực phẩm nhưng họ chi tiêu không nhiều cho những thực phẩm này.
 - + Khách hàng phân khúc 1 có số tiền chi rượu và thịt khác nhau.
 ⇒ Phỏng đoán: Có vẻ họ khá ít thời gian cho thời gian mua sắm, họ có xu hướng mua đồ đông lạnh về tích trữ.
- Nhóm 2:
 - Số người mua cho thực phẩm tươi sống, hoa quả và rượu khá nhiều nhưng mức độ chi tiêu cho các mặt hàng chỉ ở mức trung bình, họ ít chi cho đồ đông lạnh.
 ⇒ Phỏng đoán: họ thuộc kiểu gia đình khá phổ biến trong cuộc sống với mức sống trung bình và gia đình có 2-3 người, họ quan tâm đến sức khỏe, hạn chế đồ đông lạnh, mua thực phẩm tươi sống, hoa quả; và cũng có những dịp đặc biệt sẽ chi tiền cho rượu.
- Nhóm 3:
 - + Họ chi khá nhiều tiền cho thực phẩm.
 - + Sản phẩm được họ mua chủ yếu là rượu và thịt và họ chi khá nhiều tiền trong 2 năm qua cho nhóm thực phẩm này.

- + Chi ít cho mặt hàng đồ ngọt, đồ đông lạnh.
- ⇒ Phỏng đoán: Khách hàng ở phân khúc 3 có cuộc sống thành đạt, hay giao lưu bạn bè hoặc người khác giới với những bữa ăn sang trọng.

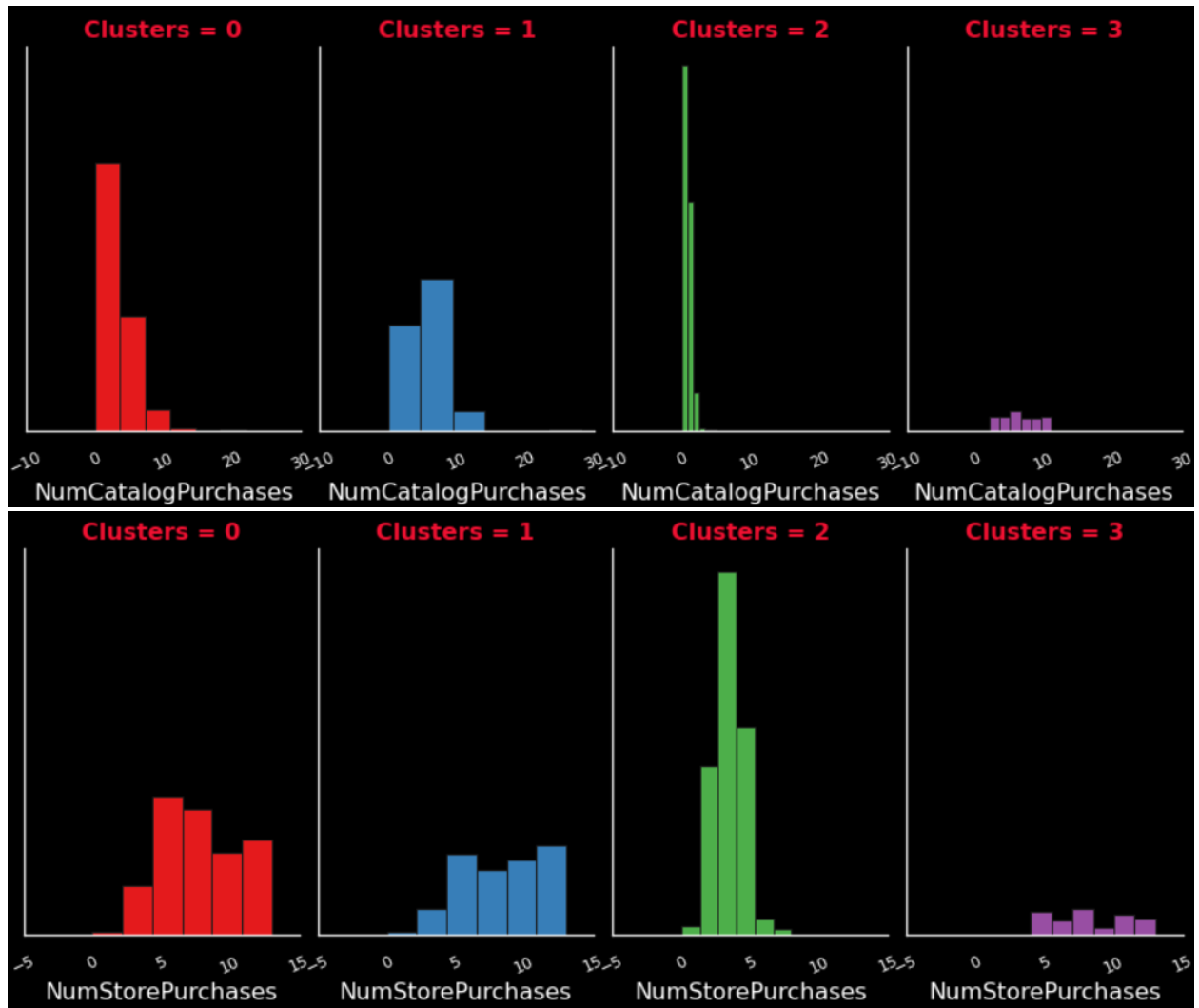
2.2. Tác động chiến dịch marketing đến khách hàng.

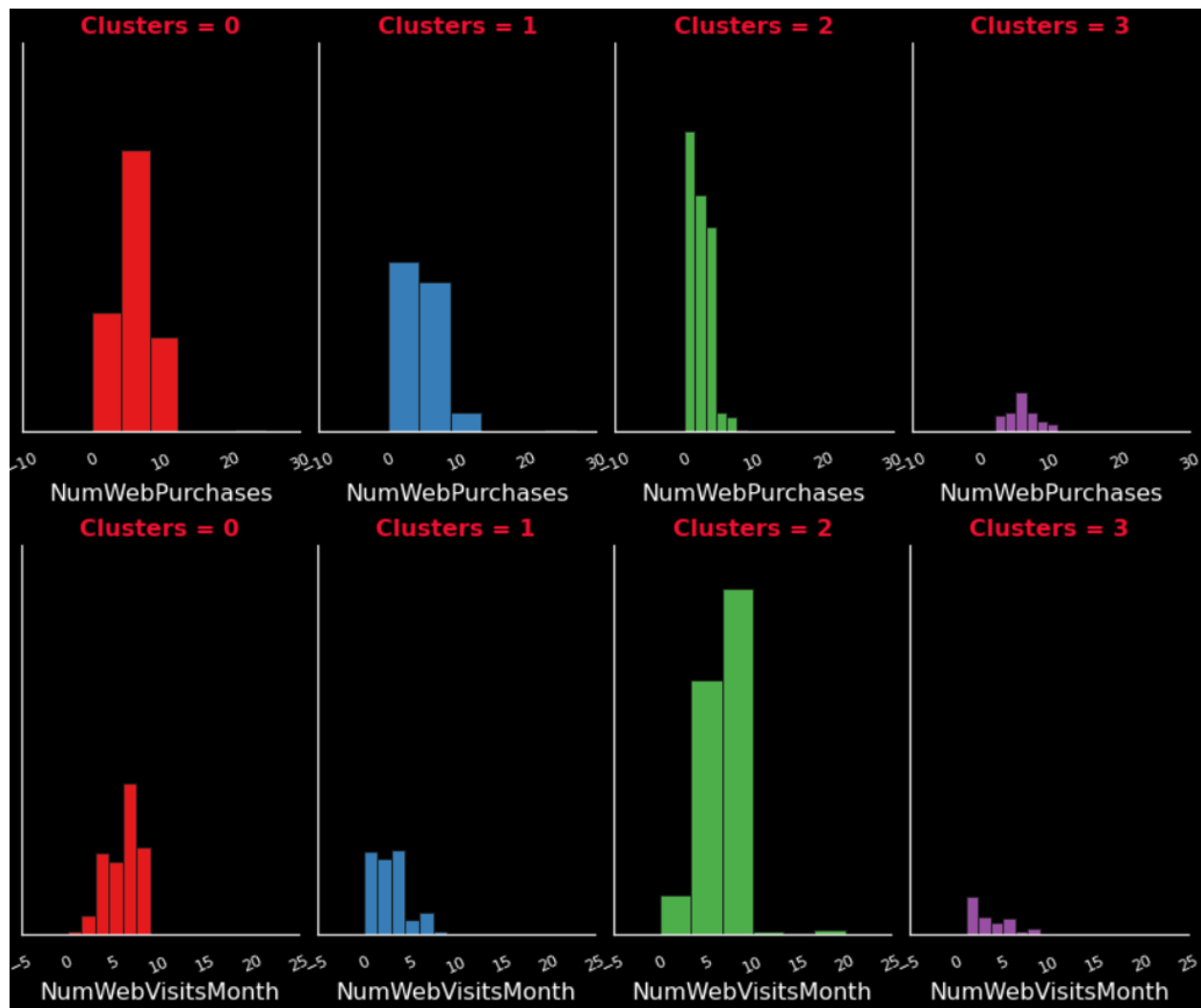


Theo biểu đồ:

- Những chiến dịch marketing không được khách hàng ưa thích: Trong số 2207 dữ liệu khách hàng, có khoảng 500 – ¼ khách hàng quan tâm đến chiến dịch marketing của công ty.
- Những người hưởng ứng chủ yếu là phân khúc khách hàng nhóm 3.
- Phân khúc khách hàng nhóm 0,1,2 chủ yếu chỉ hưởng ứng 1 chiến dịch quảng cáo.
- ⇒ Chiến dịch marketing của công ty chưa phù hợp.
- ⇒ Những chiến dịch quảng cáo cũ hiệu quả với khách hàng phân khúc 3 – phân khúc khách hàng có số lượng khách hàng chiếm tỉ lệ ít nhất.

2.3. Xu hướng mua sản phẩm của khách hàng.





- Từ năm 2017- 2019: Khách hàng vẫn chuộng mua sản phẩm tại cửa hàng và mua trên web, mua hàng qua catalog ít hơn rất nhiều.
 - Mua trên web: tỷ lệ mua sản phẩm trên web của phân khúc khách hàng nhóm 0,1,3 tương đối cao.
 - Nhóm khách hàng phân khúc 0: số người mua khi sử dụng catalog ở phân khúc nhóm 0 ít hơn so với số người mua trên web và store; họ chỉ tham gia khoảng 1-2 chiến dịch quảng cáo của công ty.
 - Nhóm khách hàng phân khúc 1 ít mua trên web mà chủ yếu mua qua cửa hàng và khi sử dụng catalog.
- ⇒ Điều này phù hợp với hoàn cảnh của họ (thu nhập thấp, gia đình đông), nhưng họ lại ít hưởng ứng những chiến dịch quảng cáo.

- ⇒ Chiến dịch quảng cáo không phù hợp với phân khúc khách hàng nhóm 3 (ví dụ: Tổng số bill > 500.000 được tặng 1 sản phẩm hoặc mua sản phẩm với giá ưu đãi).
- Số lần mua trên web và số lần thăm web của nhóm 2 trái ngược nhau; tuy có lượng mua trên web ít hơn nhưng lại có lượng vào web xem sản phẩm nhiều hơn.
- ⇒ Nhóm 2 có thể có những đồ dự định mua trước, xem trước trên web để đến cửa hàng mua.
- ⇒ Nhóm khách hàng mua hàng ít bị chia phối bởi cảm xúc. Đó cũng là nguyên nhân tại sao những chiến dịch quảng cáo từ trước tác động không hiệu quả đến chiến dịch quảng cáo từ trước đến giờ.
- Những hình thức khách hàng có thể đến mua sản phẩm, khách hàng ở phân khúc nhóm 3 đều có số lượng người mua nhiều.
- ⇒ Nhóm khách hàng thích mua sắm, có thu nhập cao. Đó là nguyên nhân khách hàng phân khúc nhóm 3 dễ dàng tiếp nhận những chiến dịch quảng cáo của công ty.

Phần 4: Kiến nghị.

- Công ty có 45.8% khách hàng ở phân khúc 1, những người có chi tiêu và thu nhập thấp, nhưng lại là nguồn khách hàng chính của công ty. Công ty nên có những chiến dịch marketing phù hợp vấn đề giá cả. Tỷ lệ họ lên web không nhiều, khi marketing ngay tại cửa hàng (Ví dụ: các biển giảm giá, những sản phẩm được tặng kèm khi mua với mức giá 200.000vnd ngay tại quầy lễ tân), và trong catalog; có thể áp dụng phát tờ rơi để dễ dàng tiếp cận đến những khách hàng ở phân khúc này.
- Với phân khúc khách hàng nhóm 2, họ biết cách chi tiêu, lượng tiền chi tiêu cho những sản phẩm của công ty trong 2 năm qua khá ổn định cho từng sản phẩm. Họ hay lên web để xem sản phẩm và số lượng mua ở cửa hàng lại lớn hơn mua ở trên web. Nên tạo những chiến dịch marketing trên app mua sản phẩm của công ty (ví dụ: tích điểm, tích xu đổi quà, hoặc chiến dịch marketing gamification).
- Bỏ loại mặt hàng đồ ngọt ra khỏi danh sách sản phẩm của công ty – loại sản phẩm ít được khách hàng ưa chuộng và mang lại ít doanh thu nhất cho công ty.
- Đây cũng là thời điểm dịch bệnh, triển khai hình thức mua online nhiều hơn, quảng cáo thông tin về web của công ty trên mạng xã hội rộng rãi hơn.

Phần kết

Cảm ơn thầy Tuấn đã cho nhóm em cơ hội làm dự án nghiên cứu phân khúc khách hàng.

Link code: <https://github.com/mailuong1241/Data-Preparation-Visualization/blob/main/EDA.ipynb>

Link data: <https://github.com/mailuong1241/Data-Preparation-Visualization/blob/main/marketing.csv.xlsx>

Tham khảo:

1. <https://www.kaggle.com/karnikakapoor/customer-segmentation-clustering?fbclid=IwAR1oJCapNnPcxwo8vSM7UmPu4a4wcJVsnIM-4gtpZA1zpUVfzIGWgXHzLhQ>
2. <https://www.kaggle.com/shakka/customer-segmentation?fbclid=IwAR3rmAMEQ3vkTqMzGWgiv8eAVd6LRLBOw2gHsivMVo4LMY-dNgT8T8DBB1A>
3. https://www.kaggle.com/miguelfzzz/store-customers-clustering-analysis?fbclid=IwAR0skcMXdyArteTt7OjfDcJKpuBbFpNq8_--N2kYAsnrBKXbZwMe1dmAco8

Đóng góp:

1. Bùi Thị Mai Lương (40%): Lên ý tưởng, đọc hiểu code, làm k-means phương pháp Silhouette Score, tổng hợp code, viết báo cáo, đánh giá biểu đồ và đưa kiến nghị, làm slide.
2. Trần Thế Phong (30%): Làm dữ liệu, đọc hiểu code, chuẩn bị thuyết trình.
3. Nguyễn Bá Đăng Khôi (30%): Đọc hiểu code, việt hóa code, làm k-means phương pháp elbow, chuẩn bị thuyết trình.