

CÂU HỎI 1:

YÊU CẦU:

- Xác định và xử lý những lỗi trùng lặp, thiếu dữ liệu, sai logic, bất nhất về định dạng của từng tập dữ liệu và toàn bộ các tập dữ liệu.
- Xác định nguyên nhân và hướng khắc phục những chi tiết không nhất quán giữa các tập dữ liệu với nhau

Để hoàn thành 2 yêu cầu này, cần:

- *Bước 1: Thực hiện EDA và tiền xử lý trên từng tập dữ liệu*
- *Bước 2: Xây dựng mô hình thực thể - mối kết hợp giữa các bảng, các dataset với nhau (câu hỏi 2)*
- *Bước 3: Kết hợp các bảng và khám phá vấn đề*

BUỚC 1: Để thực hiện EDA và tiền xử lý trên các tập dữ liệu, trước tiên cần xem xét từng tập dữ liệu

Trong tất cả các tập dữ liệu, nếu xuất hiện ID thì ID đều ở dạng số nguyên (int). Điều này không chính xác với bản chất của ID (mã định danh), nên kiểu dữ liệu của ID sẽ được chuyển đổi thành kiểu string cho phù hợp

- **TẬP DỮ LIỆU : 2017Segmentation3685case**

Lưu thông tin phân khúc khách hàng đến quán cafe vào năm 2017, với mô tả:

```

→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 4944 entries, 0 to 4943
Data columns (total 6 columns):
 #   Column      Non-Null Count Dtype  
--- 
 0   ID          4944 non-null    int64  
 1   Segmentation 4944 non-null    object  
 2   Visit        4944 non-null    int64  
 3   Spending     4944 non-null    int64  
 4   Brand        4944 non-null    object  
 5   PPA          4944 non-null    int64  
dtypes: int64(4), object(2)
memory usage: 231.9+ KB

```

Column	Description
ID	Unique identifier for each customer.
Segmentation	Customer segment label
Visit	Number of visits made by the customer during the observation period.
Spending	Total amount of money spent on the brand over a given time frame by the customer (in thousand VND).
Brand	Type of brand chosen by the customer.
PPA	Price Per Average – calculated as total spending divided by number of visits (PPA = Spending / Visit).

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
Dòng dữ liệu bị trùng lặp	Không phát hiện vấn đề
Dữ liệu thiếu	Không phát hiện vấn đề
Dữ liệu số nguyên bất thường ở cột PPA (thông thường là số thực)	Không phát hiện vấn đề
Giá trị ngoại lai bất thường	Không phát hiện vấn đề, có giá trị ngoại lai nhưng không có vẻ bất thường
Dữ liệu âm	Không phát hiện dữ liệu âm trong tất cả các cột dữ

	liệu số
Kiểu dữ liệu chưa phù hợp	ID (đã đề cập), PPA
Dữ liệu không phù hợp định dạng trong các cột phân loại	Sai chính tả trong cột Brand
Logic giữa các cột: Liên hệ giữa PPA và Segmentation	Không phát hiện vấn đề
Tên các cột chưa cùng một định dạng	Không phát hiện vấn đề

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
Kiểu dữ liệu chưa phù hợp	PPA	Chuyển PPA thành kiểu float	Đảm bảo không xuất hiện lỗi nếu lấy Spending/Visit
Lỗi chính tả	Brand	Independent chuyển thành Independent	

- **TẬP DỮ LIỆU : Brand_Image**

Lưu thông tin về hình ảnh, độ nổi tiếng của thương hiệu, có mô tả:

```

→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 643072 entries, 0 to 643071
Data columns (total 6 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   ID          643072 non-null   int64  
 1   Year         643072 non-null   int64  
 2   City          643072 non-null   object  
 3   Awareness     642675 non-null   object  
 4   Attribute     643072 non-null   object  
 5   BrandImage    643072 non-null   object  
dtypes: int64(2), object(4)
memory usage: 29.4+ MB

```

Column	Description
ID	Unique identifier for each respondent.
Year	Year of data collection.
City	The city where the respondent resides.
Awareness	The brand that the respondent is aware of.
Attribute	How the respondent perceives the brand.
BrandImage	The brand that the respondent associates with a particular image.

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
Dòng dữ liệu bị trùng lặp	Có tồn tại dữ liệu trùng lặp
Dữ liệu thiếu	Awareness
Dữ liệu có thể được tinh gọn, dữ liệu không phù hợp định dạng trong các cột phân loại	Awareness, Attribute, BrandImage

Dữ liệu âm hoặc không nguyên không hợp lệ	Không phát hiện vấn đề
Kiểu dữ liệu chưa phù hợp	ID (đã đề cập)
Tên các cột chưa cùng một định dạng	Không phát hiện vấn đề

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
ALL	Dòng dữ liệu bị trùng lặp	Loại bỏ dòng	Không thể khai thác được gì từ các dòng trùng lặp
Awareness	Dữ liệu thiếu	Loại bỏ dòng	Số lượng dòng có dữ liệu trống ít hơn 3% tổng số dòng
Awareness, Attribute, BrandImage	Lỗi chính tả, tinh gọn dữ liệu	<ul style="list-style-type: none"> - Sửa lại tên riêng. Ví dụ: 'Runam cafe': 'Runam Cafe' - Sửa lỗi chính tả: 'Indepedent Cafe': 'Independent Cafe' - Gộp các giá trị tương đồng: Ví dụ: Gom tất cả các giá trị có chữ "Other ..." về một nhãn duy nhất là "Other" 	Giảm số lượng giá trị riêng biệt

• TẬP DỮ LIỆU: Brand_Health

Lưu thông tin về tình trạng hoạt động của các thương hiệu cafe, có mô tả:

```

→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 74419 entries, 0 to 74418
Data columns (total 20 columns):
 #   Column            Non-Null Count Dtype  
--- 
 0   ID                74419 non-null   int64  
 1   Year              74419 non-null   int64  
 2   City               74419 non-null   object  
 3   Brand              74419 non-null   object  
 4   Spontaneous        30993 non-null   object  
 5   Awareness          74305 non-null   object  
 6   Trial              47330 non-null   object  
 7   P3M                28849 non-null   object  
 8   P1M                19399 non-null   object  
 9   Comprehension      26346 non-null   object  
 10  Brand_Likability  10331 non-null   object  
 11  Weekly             13382 non-null   object  
 12  Daily              7621 non-null    object  
 13  Fre#visit          19332 non-null   float64 
 14  PPA                14073 non-null   float64 
 15  Spending           14073 non-null   float64 
 16  Segmentation       14073 non-null   object  
 17  NPS#P3M            21605 non-null   float64 
 18  NPS#P3M#Group     21605 non-null   object  
 19  Spending_use       14073 non-null   float64 
dtypes: float64(5), int64(2), object(13)
memory usage: 11.4+ MB

```

Column	Description
ID	Unique identifier for each respondent.
Year	Year of data collection.
Brand	The brand being evaluated.
Spontaneous	The brand that comes first to the respondent's mind (unaided awareness).
Awareness	The brand that the respondent is aware of.
Trial	Whether the respondent has ever tried the brand.
P3M	Whether the respondent used the brand in the past 3 months.
P1M	Whether the respondent used the brand in the past 1 month.
Comprehension	How well the respondent understands or knows about the brand.
Brand_Likability	Consumer's level of affection or preference toward the brand.
Weekly	Indicates if the respondent uses the brand weekly (can be "Applicable" or "Not Applicable").
Daily	Indicates if the respondent uses the brand daily.
Fre#Visit	Number of visits made to the brand by the respondent.
PPA	Price Per Average – calculated as total spending divided by number of visits (PPA = Spending / Visit).
Spending	Total amount of money spent on the brand over a given time frame by the customer (in thousand VND).
Segmentation	Customer segment label.
NPS#P3M	Net Promoter Score over the past 3 months (P3M = Past 3 Months).
NPS#P3M#Group	NPS classification group: Promoter, Passive, or Detractor.
Spending_use	Amount of spending that was actually used or recorded.

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
Dòng dữ liệu bị trùng lắp	Không phát hiện vấn đề
Dữ liệu thiếu	Spontaneous Awareness Comprehension Trial, P3M, P1M (Xử lý theo kiểu sai mô tả) Brand_Likability (Xử lý theo kiểu sai mô tả) Weekly, Daily (Xử lý theo kiểu sai mô tả) Fre#visit PPA Spending, Spending_use Segmentation NPS#P3M, NPS#P3M#Group
Dữ liệu có thể được tinh gọn, dữ liệu không phù hợp định dạng trong các cột phân loại	Brand, Awareness, Trial, P3M, P1M, Spontaneous, Weekly, Daily,
Dữ liệu ngoại lai bất thường ở Freq#Visit	Không phát hiện vấn đề sau khi kiểm tra
Dữ liệu sai mô tả xuất hiện	Trial, P1M, P3M, Brand_Likability, Weekly, Daily
Kiểu dữ liệu chưa phù hợp	ID (đã đề cập)
Cột có cùng ý nghĩa	Spending và Spending_use
Tên các cột chưa cùng một định dạng	Brand_Likability, Fre#visit, NPS#P3M, NPS#P3M#Group, Spending_use
Logic giữa các cột	Không phát hiện vấn đề

Liên hệ giữa NPS#P3M và NPS#Group

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
Spontaneous	Dữ liệu trống	Thay bằng None	Tỉ lệ dữ liệu trống cao, chuyển thành None có nghĩa là người dùng không nghĩ đến một thương hiệu nào
Awareness	Dữ liệu trống	Loại bỏ dòng	Tỉ lệ dữ liệu trống thấp hơn 3% tổng số dòng dữ liệu
Comprehension	Dữ liệu trống	Thay bằng Unknown	Dữ liệu trống nhiều và chưa xác định được mục đích cụ thể, chưa thể xác định có nên loại bỏ cột
Fre#visit, Segmentation	Dữ liệu trống	Tách khỏi dataset gốc	Dữ liệu trống nhiều và có vai trò quan trọng, liên quan đến 2 cột PPA và Spending nên tách riêng cùng với 2 cột đó
NPS#P3M, NPS#P3M#Group	Dữ liệu trống	Thay bằng Unknown (riêng NPS#P3M thì chuyển thành kiểu string trước)	Dữ liệu trống nhiều và có vai trò quan trọng
PPA	Dữ liệu trống	Loại bỏ cột khỏi dataset gốc	PPA được tính toán dựa trên Spending/VisitFrequency nhưng VisitFrequency lại chứa quá nhiều giá trị null, không phù hợp với mục đích sử dụng. Không loại bỏ hoàn toàn mà tách riêng thành 1 dataset mới để lưu lại thông tin cho giải pháp tốt hơn

	Dữ liệu trống	Loại bỏ cột khỏi dataset gốc, riêng Spending_use thì xóa hẳn	Không phù hợp với mục đích sử dụng, số lượng dữ liệu trống quá nhiều chiếm gần 80% cột. Một số phương án khác cũng đã được đề ra nhưng không hiệu quả (xem chi tiết Notebook): tìm trong bảng 2017, thay bằng trung vị theo Segmentation, ... Không loại bỏ hoàn toàn mà tách riêng thành 1 dataset mới để lưu lại thông tin cho giải pháp tốt hơn
Trial, P1M, P3M, Brand_Likability	Dữ liệu sai mô tả	<ul style="list-style-type: none"> - Trial, P1M, P3M trên mô tả phải có dữ liệu dạng nhị phân (Yes/No hoặc tương tự), nhưng lại có dữ liệu là tên thương hiệu nên có thể nghi ngờ có vấn đề trong quá trình nhập liệu liên quan đến lệnh điều kiện khiến lặp lại giá trị Brand nếu người dùng đã từng thử/dùng thương hiệu - Trial, P1M, P3M đối chiếu với Brand, nếu bằng thì đặt lại giá trị là Yes, khác hoặc có dữ liệu trống thì đặt lại giá trị là No - Brand_Likability bị xóa (cả cột) 	<ul style="list-style-type: none"> - Trial, P1M, P3M trên mô tả phải có dữ liệu dạng nhị phân (Yes/No hoặc tương tự), nhưng lại có dữ liệu là tên thương hiệu nên có thể nghi ngờ có vấn đề trong quá trình nhập liệu liên quan đến lệnh điều kiện khiến lặp lại giá trị Brand nếu người dùng đã từng thử/dùng thương hiệu - Dữ liệu trong Brand_Likability trên mô tả không nhất thiết ở dạng nhị phân nên không thể đối chiếu so sánh, không có cách khôi phục dữ liệu thiếu
Weekly, Daily	Dữ liệu sai mô tả	Tương tự Trial, nhưng thay vì Yes/No thì là Applicable/Not Applicable	Tương tự Trial

Spending và Spending_use	Các cột tương tự về ý nghĩa	Đã xử lý xong ở phần xử lý dữ liệu trông	
Brand, Awareness, Trial, P3M, P1M, Spontaneous, Weekly, Daily,	Dữ liệu có thể được tinh gọn, dữ liệu không phù hợp định dạng trong các cột phân loại	<ul style="list-style-type: none"> - Sửa lại tên riêng. Ví dụ: 'Runam cafe': 'Runam Cafe' - Sửa lỗi chính tả: 'Indepedent Cafe': 'Independent Cafe' - Gộp các giá trị tương đồng: Ví dụ: Gom tất cả các giá trị có chữ "Other" ..." về một nhãn duy nhất là "Other" 	
Brand_Likability, Fre#visit, NPS#P3M, NPS#P3M#Group, Spending_use	Tên các cột chưa cùng định dạng	<p>Chuẩn hóa tên theo Upper CamelCase và đặt lại tên nếu cần (ví dụ: VisitFrequency)</p>	Cải thiện tính dễ đọc và tránh quá nhiều giá trị riêng biệt

Tập dữ liệu tách ra từ BrandHealth đặt tên là **BrandSegmentation** bao gồm các thông tin: ID, Brand, Spending, PPA, VisitFrequency, Segmentation

VẤN ĐỀ	CỘT
Dòng dữ liệu bị trùng lặp	Không phát hiện vấn đề
Dữ liệu thiếu	VisitFrequency PPA Spending Segmentation

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
VisitFrequency PPA Spending Segmentation	Dữ liệu trống	Thay bằng Unknown	Tỉ lệ dữ liệu trống cao, phục vụ cho phân tích sau này nên để lại nếu cần thiết có thể truy xuất thông tin

- **TẬP DỮ LIỆU : SA#VAR**

Lưu thông tin cá nhân của những người đến quán cafe, có mô tả:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11761 entries, 0 to 11760
Data columns (total 20 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   ID               11761 non-null   int64  
 1   City              11761 non-null   object  
 2   Group_size        11746 non-null   float64 
 3   Age               11752 non-null   float64 
 4   MPI#Mean          8044 non-null   float64 
 5   TOM               11761 non-null   object  
 6   BUMO              11761 non-null   object  
 7   BUMO_Previous     6096 non-null   object  
 8   MostFavourite     11761 non-null   object  
 9   Gender             11761 non-null   object  
 10  MPI#detail        8076 non-null   object  
 11  Age#group         11752 non-null   object  
 12  Age#Group#2       11752 non-null   object  
 13  MPI                8044 non-null   object  
 14  MPI#2              8044 non-null   object  
 15  Occupation         11761 non-null   object  
 16  Occupation#group  11761 non-null   object  
 17  Year               11761 non-null   int64  
 18  Col                11761 non-null   int64  
 19  MPI_Mean_Use      8044 non-null   float64 
dtypes: float64(4), int64(3), object(13)
memory usage: 1.8+ MB
```

Column	Description
ID	Unique identifier for each customer.
City	City where the respondent resides or visited the coffee shop.
Group_size	Number of people in the respondent's visit group.
Age	Age of the respondent.
MPI#Mean	Monthly personal income (numerical average, e.g., 5499 = 5.499 million VND).
TOM	Top-of-mind coffee brand mentioned by the respondent.
BUMO	Brand used most often (most used brand).
BUMO_Previous	Brand used most often previously (if any),
MostFavourite	Brand the respondent considers as their favorite.,
Gender	Gender of the respondent.,
MPI#detail	Income range in text.,
Age#group	Age group category.,
Age#group2	Alternative age group label.,
MPI	Income category.,
MPI#2	Grouped income tier.,
Occupation	Respondent's occupation.,
Occupation#group	Broad occupation group,
Year	Year of data collection.,
MPI_Mean_Use	Same as MPI#Mean; likely used for processing or reporting.,

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VĂN ĐỀ	CỘT
<i>Dòng dữ liệu bị trùng lặp</i>	Không phát hiện vấn đề
<i>Dữ liệu thiếu</i>	<ul style="list-style-type: none"> - Group_size - Age - BUMO_Previous - MPI#detail - Age#group - Age#Group#2 - MPI - MPI#2 - MPI_Mean_Use

	<ul style="list-style-type: none"> - MPI#Mean <p>Trong đó MPI, MPI#2, MPI#mean, MPI_Mean_use trông đồng bộ ở tất cả các dòng (tức là mỗi mà 1 trong các cột trên trông cũng đều có tất cả các cột trên trông)</p>
Dữ liệu có thể được tinh gọn	BUMO_Previous, TOM, BUMO, MostFavourite, Occupations, Occupation#group
Dữ liệu ngoại lai bất thường	Group_size, Group_size, MPI_Mean_Use
Dữ liệu âm hoặc không nguyên không hợp lệ	Không phát hiện dữ liệu âm trong tất cả các cột dữ liệu số, không phát hiện dữ liệu không nguyên trong các cột Group_size, Age
Kiểu dữ liệu chưa phù hợp	Group_size, Age
Dữ liệu không phù hợp định dạng trong các cột phân loại	Không phát hiện vấn đề
Các cột tương tự về ý nghĩa	<ul style="list-style-type: none"> - MPI#detail - Age#group - Age#Group#2 - MPI - MPI#2 - MPI_Mean_Use - MPI#Mean
Cột dư thừa	Col
Logic giữa các cột	Không phát hiện vấn đề

<ul style="list-style-type: none"> - Liên hệ giữa Age và AgeGroup - Liên hệ giữa MPI#Mean và MPI#detail 	
Tên các cột chưa cùng một định dạng	Group_size MPI#Mean BUMO_Previous MPI#detail Age#group Age#group2 MPI#2 Occupation#group MPI_Mean_Use

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
Group_size Age Age#group Age#Group#2	Dữ liệu trống	Loại bỏ dòng	Số lượng dòng có dữ liệu trống ít hơn 3% tổng số dòng
BUMO_Previous MPI, MPI#2	Dữ liệu trống	Thay bằng Unknown (riêng trong MPI#2 là 6. Unknown để đồng bộ)	Số lượng dữ liệu trống lớn
MPI#Mean MPI_mean_use	Dữ liệu trống	Thay bằng -1	Số lượng dữ liệu trống lớn và ở dữ liệu ở cột này có dạng số thực. Thay bằng -1 để tính toán thống kê phục vụ cho việc tìm ra các lỗi tiềm

			ản (dữ liệu âm, dữ liệu ngoại lai, xem chi tiết trong Notebook), sau đó toàn bộ dữ liệu trong cột được chuyển thành kiểu chuỗi (str) với giá trị -1 lúc này được thay bằng Unknown nhằm hạn chế nhầm lẫn về sau
BUMO_Previous	Dữ liệu có thể được tinh gọn	Thay 'Don't have any brands' bằng Unknown	Người dùng không có thương hiệu dùng nhiều nhất nên xem như không biết để đơn giản hóa
BUMO_Previous, TOM, BUMO, MostFavourite	Dữ liệu có thể được tinh gọn	<ul style="list-style-type: none"> - Sửa lại tên riêng. Ví dụ: 'Runam cafe': 'Runam Cafe' - Sửa lỗi chính tả: 'Indepedent Cafe': 'Independent Cafe' - Gộp các giá trị tương đồng: Ví dụ: Gom tất cả các giá trị có chữ "Other ..." về một nhãn duy nhất là "Other" 	Giảm số lượng giá trị riêng biệt
Occupations, Occupation#group	Dữ liệu có thể được tinh gọn	Occupation Refuse chuyển thành Other	Số lượng người dùng ghi nhận trong nhóm đó không nhiều và không thể khai thác thông tin gì từ nhóm, nên có thể đồng nhất với Other Occupations (chuyển thành Other)
Group_size, Age	Kiểu dữ liệu chưa phù hợp	Chuyển thành kiểu int	Do tuổi và số lượng người phải là số nguyên nên chuyển thành kiểu int (kiểm tra trước khi chuyển)

Group_size, MPI_Mean_Use	Dữ liệu ngoại lai bất thường	Xem lại quá trình nhập dữ liệu	
Col	Cột dư thừa	Xóa bỏ	Do không ảnh hưởng đến yêu cầu bài toán
- MPI#detail - Age#group - Age#Group#2 - MPI - MPI#2 - MPI_Mean_Use - MPI#Mean			<ul style="list-style-type: none"> - Age#Group#2 chi tiết hơn, phục vụ cho phân tích. Giữ lại cả hai cột sẽ làm tăng số chiều dữ liệu không cần thiết, nên sẽ loại bỏ Age#Group. Trong trường hợp cần thiết, vẫn có thể tạo lại Age#Group từ Age#Group#2 - MPI#detail chi tiết nhất, chuyển đổi lại thành các khoảng giá trị đơn giản hơn thay vì văn bản và giữ lại - MPI#Mean và Mean_use hoàn toàn tương tự, giữ lại 1 trong 2
Group_size MPI#Mean BUMO_Previous MPI#detail Age#group Age#group2 MPI#2 Occupation#group MPI_Mean_Use	Tên các cột chưa cùng định dạng	Age#Group#2, MPI#detail MPI#Mean	Chuẩn hóa tên theo Upper CamelCase (ví dụ: GroupSize)

• TẬP DỮ LIỆU : Companion

Lưu thông tin về những nhóm người đồng hành với người đến quán cafe, có mô tả:

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20739 entries, 0 to 20738
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   ID               20739 non-null    int64  
 1   City              20739 non-null    object  
 2   Companion#group  20739 non-null    object  
 3   Year              20739 non-null    int64  
dtypes: int64(2), object(2)
memory usage: 648.2+ KB

```

Column	Description
ID	Unique identifier for each customer.
City	City where the respondent resides or visited the coffee shop.
Companion#group	The usual type of companion the respondent has when visiting a coffee shop.
Year	Year of data collection.

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
<i>Dòng dữ liệu bị trùng lặp</i>	Có tồn tại các dòng trùng lặp
<i>Dữ liệu thiếu</i>	Không phát hiện vấn đề
<i>Dữ liệu ngoại lai bất thường</i>	Không phát hiện vấn đề
<i>Dữ liệu âm hoặc không nguyên không hợp lệ</i>	Không phát hiện vấn đề
<i>Kiểu dữ liệu chưa phù hợp</i>	Không phát hiện vấn đề
<i>Dữ liệu không phù hợp định dạng trong các cột phân loại</i>	Không phát hiện vấn đề

<i>Tên các cột chưa cùng một định dạng</i>	Companion#group
--	-----------------

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
ALL	Dữ liệu trùng lặp	Loại bỏ dòng	Không thể khai thác được các dòng trùng lặp này và xử lý cần thời gian nhiều hơn
Companion#group	Tên các cột chưa cùng định dạng	Chuẩn hóa tên theo Upper CamelCase thành CompanionGroup	Khiến các cột dễ đọc và nhất quán hơn

- **TẬP DỮ LIỆU : Competitor**

Lưu thông tin các đối thủ cạnh tranh theo thành phố, năm và số cửa hàng, có mô tả

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 234 entries, 0 to 233
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   No#         234 non-null    int64  
 1   Brand        234 non-null    object  
 2   city         234 non-null    object  
 3   Year         234 non-null    int64  
 4   StoreCount   234 non-null    int64  
dtypes: int64(3), object(2)
memory usage: 9.3+ KB
```

Column	Description
City	City where the respondent resides or visited the coffee shop.
Year	Year of data collection.
No#	Row number or entry index.
Brand	Name of the coffee brand.
StoreCount	Number of stores the brand operated in that city during the given year.

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
<i>Dòng dữ liệu bị trùng lặp</i>	Không phát hiện vấn đề
<i>Dữ liệu thiếu</i>	Không phát hiện vấn đề
<i>Dữ liệu bất thường</i>	StoreCount
<i>Định dạng của dữ liệu số thứ tự (có theo thứ tự không)</i>	Không phát hiện vấn đề
<i>Kiểu dữ liệu chưa phù hợp</i>	Không phát hiện vấn đề
<i>Dữ liệu không phù hợp định dạng trong các cột phân loại</i>	Không phát hiện vấn đề
<i>Cột không cần thiết</i>	No#

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
StoreCount	Dữ liệu bất	- Kiểm tra các dòng	

	thường	có StoreCount = 0 không phát hiện bất thường - Một số dòng có StoreCount cao bất thường cần kiểm tra lại cách nhập dữ liệu nếu cần thiết	
No#	Cột không cần thiết	Loại bỏ	Cột No# thể hiện index của hàng, nhưng giá trị này trong pandas được tạo tự động, không cần 1 cột riêng để minh họa

- **TẬP DỮ LIỆU : Dayofweek**

Lưu thông tin về những người đến quán cafe vào những ngày trong tuần, có mô tả:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39095 entries, 0 to 39094
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype  
 ---  --  
 0   ID               39095 non-null   int64  
 1   City              39095 non-null   object 
 2   Dayofweek         39009 non-null   object 
 3   Visit#Dayofweek  39041 non-null   float64 
 4   Year              39095 non-null   int64  
 5   Weekday#end       39009 non-null   object 
dtypes: float64(1), int64(2), object(3)
memory usage: 1.8+ MB
```

Column	Description
ID	Unique identifier for each customer.
City	City where the respondent resides or visited the coffee shop.
Year	Year of data collection.,
Dayofweek	Specific day of the week when the visit occurred.
Visit#Dayofweek	Number of visits made on that particular day of the week.
Weakday#end	Classification of the day as Weekdays or Weekend

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
Dòng dữ liệu bị trùng lặp	Có xuất hiện
Dữ liệu thiếu	Weekday#end, Visit#Dayofweek, Dayofweek
Dữ liệu ngoại lai bất thường	Không phát hiện vấn đề (Có xuất hiện một số giá trị ngoại lai nhưng không có vẻ bất thường)
Dữ liệu âm hoặc không nguyên không hợp lệ	Không phát hiện vấn đề
Kiểu dữ liệu chưa phù hợp	Visit#Dayofweek
Dữ liệu không phù hợp định dạng trong các cột phân loại	Không phát hiện vấn đề
Logic giữa các cột	Không phát hiện vấn đề
Liên hệ giữa Dayofweek và	

<i>Weekday#end (Ví dụ Saturday là Weekend)</i>	
<i>Tên các cột chưa cùng một định dạng</i>	Weekday#end, Visit#Dayofweek

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
ALL	Dữ liệu trùng lặp	Loại bỏ dòng	Không thể khai thác được các dòng trùng lặp này và xử lý cần thời gian nhiều hơn
Weekday#end, Visit#Dayofweek, Dayofweek	Dữ liệu trống	Loại bỏ dòng	Số lượng dữ liệu trống nhỏ hơn 3% tổng số dòng, Dayofweek trống dẫn đến Weekday#end trống
Visit#Dayofweek	Kiểu dữ liệu chưa phù hợp	Chuyển thành kiểu int	Do số lượng người phải là số nguyên nên chuyển thành kiểu int (kiểm tra trước khi chuyển)
Weekday#end, Visit#Dayofweek	Tên các cột chưa cùng định dạng	Chuẩn hóa tên theo Upper CamelCase 'Visit#Dayofweek' chuyển thành 'VisitOnDayofweek' 'Weekday#end' chuyển thành 'TypeOfDay'	Khiến các cột dễ đọc và nhất quán hơn

- **TẬP DỮ LIỆU : Daypart**

Lưu thông tin về những người đến quán cafe vào những khung giờ trong ngày, với mô tả:

```
Data columns (total 5 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   ID               19189 non-null   int64  
 1   City              19189 non-null   object  
 2   Daypart            19176 non-null   object  
 3   Visit#Daypart     18342 non-null   float64 
 4   Year              19189 non-null   int64  
 dtypes: float64(1), int64(2), object(2) 
 memory usage: 749.7+ KB
```

Column	Description
ID	Unique identifier for each customer.
City	City where the respondent resides or visited the coffee shop.
Year	Year of data collection.,
Daypart	Time range during the day when the visit occurred.
Visit#Daypart	Number of visits made during that specific time range.

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
Dòng dữ liệu bị trùng lặp	Không phát hiện vấn đề
Dữ liệu thiếu	Daypart và Visit#Daypart
Dữ liệu ngoại lai bất thường	Không phát hiện vấn đề (Có xuất hiện một số giá trị ngoại lai nhưng không có vẻ bất thường)

<i>Dữ liệu âm hoặc không nguyên hợp lệ</i>	Không phát hiện vấn đề
<i>Kiểu dữ liệu chưa phù hợp</i>	Visit#Daypart
<i>Dữ liệu không phù hợp định dạng trong các cột phân loại</i>	Không phát hiện vấn đề
<i>Tên các cột chưa cùng một định dạng</i>	Visit#Daypart

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
Daypart	Dữ liệu trống	Loại bỏ dòng	Số lượng dữ liệu trống nhỏ hơn 3% tổng số dòng
Visit#Daypart	Dữ liệu trống	Thay bằng 1	Nguyên nhân khiến dữ liệu ở Visit thiếu có thể do sai sót trong quá trình ghi nhận, người dùng với ID tương ứng có thể đã từng ghé nhưng không được ghi nhận hoặc chưa từng ghé nhưng ID của họ vẫn xuất hiện. Ta sẽ thay thế các giá trị NaN trong visit bằng 1, thể hiện rằng với mỗi ghi nhận, người dùng đã từng ghé 1 lần, giả định rằng sai lầm xuất hiện ở việc người dùng đã từng ghé nhưng chưa được ghi nhận. Do số lượng dữ liệu NaN lớn nên không thể trực tiếp loại bỏ
Visit#Daypart	Kiểu dữ	Chuyển thành kiểu	Do số lượng người phải là số

	liệu chưa phù hợp	int	nguyên nên chuyển thành kiểu int (kiểm tra trước khi chuyển)
Visit#Daypart	Tên các cột chưa cùng định dạng	Chuẩn hóa tên theo Upper CamelCase: chuyển thành VisitOnDaypart	Khiến tên cột dễ đọc và nhất quán hơn

- **TẬP DỮ LIỆU : Needstate**

Lưu thông tin về nguyên nhân những người đến quán cafe theo ngày trong tuần và theo khung giờ, có mô tả

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 75251 entries, 0 to 75250
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   ID               75251 non-null   int64  
 1   City              75251 non-null   object  
 2   Year              75251 non-null   int64  
 3   Needstates        75251 non-null   object  
 4   Day#Daypart       75251 non-null   object  
 5   NeedstateGroup    75251 non-null   object  
dtypes: int64(2), object(4)
memory usage: 3.4+ MB
```

Column	Description
ID	Unique identifier for each customer.
City	City where the respondent resides or visited the coffee shop.
Year	Year of data collection.
Needstates	Specific reason or motivation for visiting the coffee shop.
Day#Daypart	Time context for the need state.
NeedstateGroup	Broader category grouping similar need states.

Một số vấn đề có thể phát sinh, thể hiện trong bảng sau (Để xem chi tiết hơn quá trình xử lý có thể truy cập notebook đính kèm. Các quá trình đều được thực hiện bằng ngôn ngữ lập trình Python trên phần mềm Google Colaboratory) và đồng thời được kiểm chứng

VẤN ĐỀ	CỘT
Dòng dữ liệu bị trùng lặp	Có xuất hiện
Dữ liệu thiếu	Không phát hiện vấn đề
Dữ liệu ngoại lai bất thường	Không phát hiện vấn đề
Dữ liệu âm hoặc không nguyên không hợp lệ	Không phát hiện vấn đề
Kiểu dữ liệu chưa phù hợp	Không phát hiện vấn đề
Dữ liệu không phù hợp định dạng trong các cột phân loại	Needstates
Kiểm tra logic: kiểm tra xem needstate có nằm trong nhóm tương ứng không	Không phát hiện vấn đề
Tên các cột chưa cùng một định dạng	Day#Daypart

Tiếp theo đó, các vấn đề sẽ được xử lý dần dần

Cột	Vấn đề	Cách xử lý	Nguyên nhân chọn
ALL	Dữ liệu trùng lặp	Loại bỏ dòng	Không thể khai thác được các dòng trùng lặp này và xử lý cần thời gian nhiều hơn
Needstates	Dữ liệu không phù hợp định dạng trong	Phát hiện lỗi sai chính tả, xử lý lỗi sai chính tả	Đảm bảo tính chính xác

	các cột phân loại		
Day#Daypart	Tên các cột chưa cùng định dạng	Chuẩn hóa tên theo Upper CamelCase: chuyển thành DayOrDaypart	Khiến tên cột dễ đọc và nhất quán hơn

BUỚC 2: Thiết kế sơ lược lược đồ ER thể hiện mối quan hệ giữa các bảng với nhau

Bước 2 tương tự với câu 2, sẽ được trình bày ở câu hỏi 2

BUỚC 3: Kiểm tra tính nhất quán giữa các bảng với nhau

Sau khi hoàn thành bước 2, ta đã có lược đồ sơ lược về quan hệ giữa các bảng. Dựa trên lược đồ này và các tập dữ liệu đã kiểm tra, rút ra được một số vấn đề

VẤN ĐỀ	BẢNG
Số lượng đối tượng khảo sát chưa nhất quán (ID chưa đồng bộ)	Needstate & SA SA & CustomerSegmentation SA & Companion Dayofweek & SA Daypart & SA BrandHealth & SA BrandHealth & CustomerSegmentation BrandHealth & BrandImage BrandHealth & BrandSegmentation
Thông tin City và Year chưa nhất quán	Không phát hiện vấn đề
Thông tin Segmentation, Spending, PPA, VisitFrequency chưa nhất quán	BrandSegmentation & CustomerSegmentation
Tên cột chưa nhất quán	BrandHealth & CustomerSegmentation (Visit

	& VisitFrequency, Brand), BrandHealth & Competitor (cột Brand)
--	--

BẢNG	Vấn đề	Cách xử lý	Nguyên nhân chọn
Needstate & SA SA & CustomerSegmentation SA & Companion Dayofweek & SA Daypart & SA BrandHealth & SA BrandHealth & CustomerSegmentation BrandHealth & BrandImage BrandHealth & BrandSegmentation	Số lượng đối tượng khảo sát chưa nhất quán (ID chưa đồng bộ)	Nếu bảng A có các giá trị mà bảng B không có, loại bỏ các dòng đó trong bảng A và tiến hành kiểm tra ngược lại, xử lý tương tự	Số lượng dòng không trùng khớp rất ít, nỗ lực truy xuất lại thông tin ở các dòng này và thêm vào bảng còn thiếu không khả thi, hoặc quá tiêu tốn tài nguyên và thời gian

BrandSegmentation & CustomerSegmentation	Thông tin Segmentation, Spending, PPA, VisitFrequency chưa nhất quán	<p>Trích xuất các thông tin về Segmentation vào năm 2017 của bảng BrandSegmentation, nhận thấy được chỉ có giá trị Unknown, trong khi thông tin Segmentation trong bảng Customer tương đối đầy đủ.</p> <p>=> Giải pháp: Diền thông tin của bảng Customer vào bảng BrandSegmentation, chú ý chỉ diền thông tin của các ID có duy nhất 1 Segmentation, không diền thông tin của các ID có nhiều Segmentation mà để nguyên Unknown.</p> <p>Thực hiện tương tự với các thuộc tính khác</p>	Xem chi tiết trong Notebook và trong PDF
BrandHealth & CustomerSegmentation (Visit & VisitFrequency, Brand), BrandHealth & Competitor (cột Brand)	Tên cột chưa nhất quán	<p>Visit -> VisitFrequency</p> <p>Brand</p> <p>(CustomerSegmentation) -> CustomerSegBrand</p> <p>Brand (Competitor) -> CompetitorBrand</p>	

Vấn đề 1: đơn cử trong trường hợp bảng SA & CustomerSegmentation, tồn tại một số Customer có thông tin trong SA với năm 2017 không được ghi nhận trong CustomerSegmentation. Điều này có thể do:

- Trong quá trình ghi nhận thông tin, lỗi nhập liệu hoặc lỗi ở quá trình lọc dữ liệu xảy ra khiến thông tin của các user này không được đưa vào bảng CustomerSegmentation , nếu CustomerSegmentation được trích xuất từ SA
- Có thể bị nhân bản dòng, cập nhật thông tin trực tiếp vào bảng SA, hoặc lỗi nhập liệu liên quan đến năm khiến số lượng dòng có Year = 2017 trong bảng SA nhiều hơn trong bảng CustomerSegmentation, nếu CustomerSegmentation được thu thập trước sau đó mới đưa vào SA.

Giải pháp: Có 2 lựa chọn đối với trường hợp này:

- Drop các dòng có trong SA nhưng không có trong CustomerSegmentation để đồng bộ
- Thêm các dòng có trong SA nhưng không có trong CustomerSegmentation vào CustomerSegmentation để đồng bộ

Do số lượng dòng thiếu không đáng kể so với kích thước của tập dữ liệu và quá trình thêm (lựa chọn 2) cần phải truy xuất các thông tin Visit, Spending, PPA, ... tương đối phức tạp nên lựa chọn phương án 1

Vấn đề 2: đơn cử thông tin Segmentation trong bảng BrandSegmentation_2017 (lưu ý ở đây chỉ xét các thông tin có Year = 2017) và bảng CustomerSegmentation chưa nhất quán, thể hiện ở việc thông tin trong BrandSegmentation_2017 , cột Segmentation chủ yếu là Unknown, trong khi cũng cột đó, trong CustomerSegmentation, thông tin đầy đủ được chia thành các khoảng cụ thể Seg.01 đến Seg.04. Do trước đó trong cột Segmentation của BrandSegmentation_2017 tồn tại rất nhiều dữ liệu trống nên phải thay thành Unknown. Một giải pháp cho vấn đề này là truy xuất lại thông tin trong bảng BrandSegmentation_2017 bằng cách điền thông tin của bảng Customer vào bảng BrandSegmentation_2017 với cột tương ứng là Segmentation.

Tuy nhiên, trong quá trình khám phá nhận ra được rằng có một số ID trong CustomerSegmentation có nhiều Segmentation khác nhau.

```
Số lượng ID có nhiều hơn 1 Segmentation: 485
```

```
ID
```

```
89100    3  
89101    2  
89616    2  
90421    2  
90423    2
```

```
Name: Segmentation, dtype: int64
```

Nguyên nhân có thể là lỗi do nhập liệu, có thể là do có sự thay đổi theo thời gian trong 1 năm 2017, nhưng rất có thể là do Brand. Do đó ta chỉ fill những Segmentation xuất hiện đúng 1 lần trong CustomerSegmentation vào các Segmentation tương ứng theo ID trước, sau đó sẽ cân nhắc đến Brand sau.

Quá trình kiểm tra cho biết đúng là tổ hợp Brand và ID xác định giá trị cho Segmentation, một giải pháp dành cho các giá trị Unknown trong cột Segmentation của BrandSegmentation_2017 xuất hiện, đó là ta có thể kết hợp bảng BrandSegmentation_2017 và CustomerSegmentation theo ID và Brand để truy xuất thông tin điền vào những giá trị Unknown, tuy nhiên cần lưu ý là giá trị của cột Brand trong BrandSegmentation_2017 (trái) khác so với giá trị của cột Brand trong CustomerSegmentation (phải).

count	
Brand	
Street Coffee	3368
Trung Nguyên	3204
Independent Cafe	2667
Highlands Coffee	2472
Milano	951
Cộng Cà Phê	915
Starbucks	836
The Coffee House	685
Phúc Long	649
Gong Cha	591

count	
Brand	
Independent	2190
Street	1521
Chain	1225

2 nguyên nhân khiến cho giải pháp vừa đề cập không khả thi đó là:

- Số lượng giá trị không tương đồng, có thể do Brand trong CustomerSegmentation là do người dùng chọn, còn Brand trong BrandSegmentation_2017 được chọn để đánh giá, nên sẽ xuất hiện một lượng lớn thương hiệu mà người dùng không nhớ đến để có thể chọn.
- Chưa có cơ sở xác định một thương hiệu, ví dụ Milano, Aha, ... là chuỗi thương hiệu hay cửa hàng đơn lẻ, cần thêm thông tin doanh nghiệp cung cấp để xác định và phân nhóm các giá trị trong Brand của BrandSegmentation_2017 , từ đó mới có thể đồng bộ được Brand ở 2 bảng và điền giá trị khuyết

Đề xuất cải tiến: Yêu cầu thêm thông tin từ doanh nghiệp về nguyên do có sự chênh lệch giữa số lượng Brand ở 2 bảng và cơ sở phân nhóm một giá trị Brand trong BrandHealth_2017 để có thể tiếp tục kiểm tra, xử lý bằng cách điền các giá trị vào Unknown ở cột Brand của BrandSegmentation_2017.

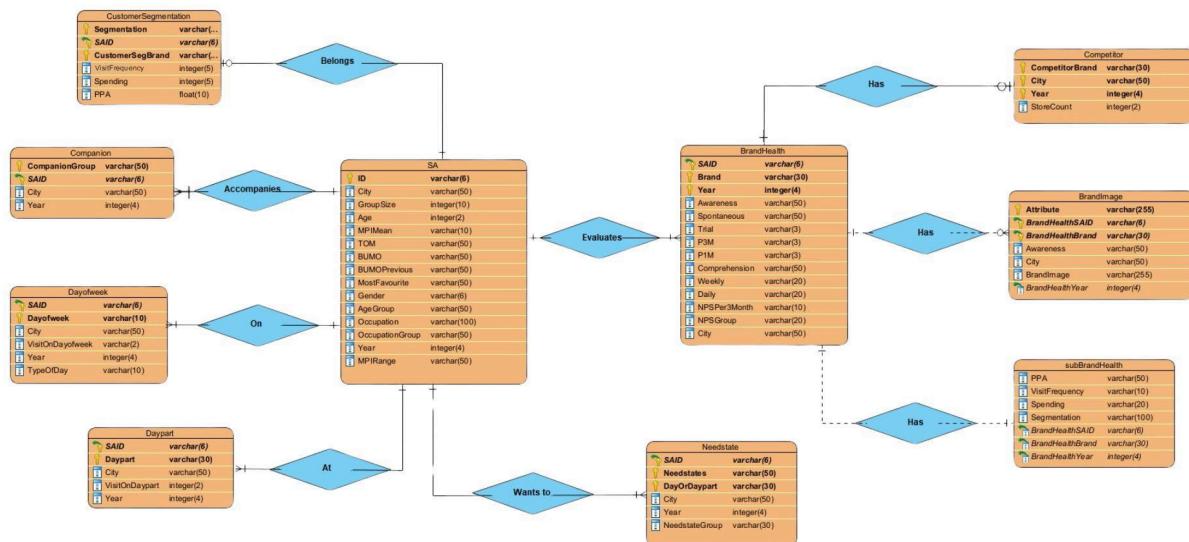
CÂU HỎI 2:

Sơ đồ mô hình dữ liệu

MỤC TIÊU

Thiết kế một **mô hình dữ liệu khái niệm** phản ánh chính xác **cấu trúc và các mối quan hệ giữa các thực thể dữ liệu** đã cho, đồng thời tích hợp kết quả từ quá trình làm sạch dữ liệu

Xây dựng sơ đồ ERD thể hiện mô hình dữ liệu đã được làm sạch, trong đó các cột dư thừa và nhãnh không nhất quán đã được xử lý.



Xác định và biểu diễn các thực thể chính

SA: *Dữ liệu thông tin của các khách hàng đến quán cafe*

- Primary key: ID - mỗi đối tượng được ghi nhận có mã ID riêng biệt, không trùng lặp
- Thuộc tính: Thể hiện các hành vi thói quen của mỗi đối tượng được ghi nhận trong bộ dữ liệu

BrandHealth: *Tình trạng hoạt động của các thương hiệu cafe*

- Primary key: Brand, Year và SAID - cặp khóa chính gồm ID mỗi người tiêu dùng và quán cafe mà họ đã trải nghiệm, mỗi cặp riêng biệt, không trùng lặp nhau
- Foreign key: SAID - được tham chiếu từ khóa chính ID của loại thực thể SA

- Thuộc tính: Các dữ liệu còn lại thể hiện hành vi của khách hàng với từng quán cafe họ đã trải nghiệm

SubBrandHealth: Thông tin về hành vi của một khách hàng với một quán cafe (Thực thể yêu**)**

- Primary key:
- Foreign key: BrandHealthSAID, BrandHealthBrand, BrandHealthYear - được tham chiếu từ khóa chính của loại thực thể BrandHealth
- Thuộc tính: Các dữ liệu còn lại thể hiện hành vi của khách hàng với từng quán cafe

BrandImage: Hình ảnh, độ nổi tiếng của quán cafe (Thực thể yêu**)**

- Primary key: BrandHealthSAID, BrandHealthBrand và Attribute - Mỗi khách hàng khi sử dụng mỗi quán cafe sẽ được ghi nhận về những nhận định hình ảnh về quán cafe riêng biệt
- Foreign key: BrandHealthYear, BrandHealthSAID và BrandHealthBrand - được tham chiếu từ kiểu thực thể BrandHealth
- Thuộc tính: các thuộc tính thể hiện thông tin về năm sử dụng và hình ảnh của quán cafe mà người tiêu dùng liên tưởng tới

Competitor: Đối thủ thương hiệu cạnh tranh theo thành phố, năm và số cửa hàng

- Primary key: Brand, City, Year. Mỗi quán cafe ở mỗi thời điểm tại mỗi thành phố sẽ có số lượng cửa hàng tồn tại khác nhau
- Thuộc tính: Thể hiện số lượng cửa hàng hiện có của quán cafe tại một địa điểm nhất định

CustomerSegmentation: Thông tin phân khúc khách hàng đến quán cafe

- Primary key: Segmentation, Brand và SAID - mỗi khách hàng được xếp vào một phân khúc cụ thể
- Foreign key: SAID - được tham chiếu từ mã khách hàng ID trong loại thực thể SA
- Thuộc tính: Thể hiện các dữ liệu liên quan đến thói quen hành vi của người tiêu dùng để phân loại phân khúc

Companion: Nhóm người đồng hành với người tiêu dùng đến quán cafe

- Primary key: CompanionGroup và SAID - mỗi người đồng hành sẽ thường xuyên đi cùng với nhóm người nhất định (gồm Friends, Family,...)
- Foreign key: SAID - được tham chiếu từ mã khách hàng ID trong loại thực thể SA

Dayofweek: Thông tin đến quán cafe vào các ngày của một tuần của người tiêu dùng

- Primary key: Dayofweek và SAID - mỗi khách hàng được ghi nhận trong ngày nào đó của 1 tuần
- Foreign key: SAID - được tham chiếu từ mã khách hàng ID trong loại thực thể SA

Daypart: Khung giờ trong ngày người tiêu dùng đến quán cafe

- Primary key: SAID và Daypart
- Foreign key: SAID - được tham chiếu từ mã khách hàng ID trong loại thực thể SA

Needstate: Nhu cầu của những người đến quán cafe theo ngày trong tuần và theo khung giờ

- Primary key: SAID, Needstates và DayOrDaypart - mỗi người tiêu dùng khi sử dụng một quán cafe ở từng thời điểm khác nhau sẽ có nhu cầu khác nhau
- Foreign key: SAID - được tham chiếu từ mã khách hàng ID trong loại thực thể SA
- Thuộc tính: thông tin về năm, địa điểm và nhóm nhu cầu của khách hàng

Gắn nhãn các thuộc tính

SA		
Thuộc tính	Label	
ID	varchar(6)	Primary Key
City	varchar(50)	
GroupSize	integer(10)	
Age	integer(2)	
MPIMean	varchar(10)	
TOM	varchar(50)	
BUMO	varchar(50)	
BUMOPrevious	varchar(50)	
MostFavourite	varchar(50)	
Gender	varchar(6)	

AgeGroup	varchar(50)	
Occupation	varchar(100)	
OccupationGroup	varchar(50)	
Year	integer(4)	
MPIRange	varchar(50)	

BrandHealth		
Thuộc tính	Label	
SAID	varchar(6)	Primary Key, Foreign Key
Awareness	varchar(50)	
Spontaneous	varchar(50)	
Trial	varchar(3)	
P3M	varchar(3)	
P1M	varchar(3)	
Comprehension	varchar(50)	
Weekly	varchar(20)	
Daily	varchar(20)	
NPSPer3Month	varchar(10)	
NPSGroup	varchar(20)	
Brand	varchar(30)	Primary Key
City	varchar(50)	

Year	integer(4)	Primary Key
------	------------	-------------

Competitor		
Thuộc tính	Label	
CompetitorBrand	varchar(30)	Primary Key
City	varchar(50)	Primary Key
Year	integer(4)	Primary Key
StoreCount	integer(2)	

BrandImage		
Thuộc tính	Label	
BrandHealthSAID	varchar(6)	Primary Key, Foreign Key
BrandHealthBrand	varchar(50)	Primary Key, Foreign Key
Attribute	varchar(255)	Primary Key
BrandHealthYear	integer(4)	Foreign Key
City	varchar(50)	
Awareness	varchar(50)	
BrandImage	varchar(255)	

CustomerSegmentation		
Thuộc tính	Label	
Segmentation	varchar(100)	Primary Key
SAID	varchar(6)	Primary Key, Foreign Key
VisitFrequency	integer(5)	
Spending	integer(5)	
CustomerSegBrand	varchar(50)	Primary Key
PPA	float(10)	

Companion		
Thuộc tính	Label	
CompanionGroup	varchar(50)	Primary Key
SAID	varchar(6)	Primary Key, Foreign Key
Year	integer(4)	
City	varchar(50)	

Dayofweek		
Thuộc tính	Label	
Dayofweek	varchar(10)	Primary Key
SAID	varchar(6)	Primary Key, Foreign Key

VisitOnDayofweek	integer(2)	
City	varchar(50)	
TypeOfDay	varchar(10)	
Year	integer(4)	

Daypart		
Thuộc tính	Label	
Daypart	varchar(30)	Primary Key
SAID	varchar(6)	Primary Key, Foreign Key
VisitOnDaypart	integer(2)	
City	varchar(50)	
Year	integer(4)	

Needstate		
Thuộc tính	Label	
DayOrDaypart	varchar(30)	Primary Key
SAID	varchar(6)	Primary Key, Foreign Key
Needstates	varchar(50)	Primary Key
City	varchar(50)	
Year	integer(4)	
NeedstateGroup	varchar(30)	

SubBrandHealth		
Thuộc tính	Label	
PPA	varchar(50)	
VisitFrequency	varchar(10)	
Spending	varchar(20)	
Segmentation	varchar(100)	
BrandHealthSAID	varchar(6)	Foreign Key
BrandHealthBrand	varchar(30)	Foreign Key
BrandHealthYear	integer(4)	Foreign Key

Giải thích các mối quan hệ giữa từng loại thực thể (dựa trên bối cảnh doanh nghiệp)

- **SA - BrandHealth (*Quan hệ một - nhiều 1:N*)**: Mỗi khách hàng có thể sử dụng hoặc trải nghiệm một hoặc nhiều quán cafe cửa hàng khác nhau. Mỗi đánh giá về một quán cafe sẽ thuộc về một khách hàng duy nhất
- **BrandHealth - BrandImage (*Quan hệ một - nhiều 1:N*)**: Mỗi khách hàng trải nghiệm quán cafe có thể có một hay nhiều hình ảnh liên tưởng về quán cafe.
- **BrandHealth - Competitor (*Quan hệ một - một 1:1*)**: Mỗi đánh giá của khách hàng về quán cafe không hay thuộc về một thương hiệu cafe ở thành phố và năm nhất định
- **SA - CustomerSegmentation (*Quan hệ một - một 1:1*)**: Một khách hàng có thể không hoặc thuộc về duy nhất một phân khúc khách hàng. Một phân khúc khách hàng của một ID khách hàng chỉ thuộc về một khách hàng duy nhất.
- **SA - Companion (*Quan hệ một - nhiều 1:N*)**: Một khách hàng sẽ có một hoặc nhiều nhóm khách hàng đồng hành. Một nhóm khách hàng đồng hành được gắn với ID của người tiêu dùng sẽ chỉ thuộc về một khách hàng duy nhất

- **SA - Dayofweek (*Quan hệ một - nhiều 1:N*):** Một khách hàng có thể sử dụng các quán cafe vào một hoặc nhiều ngày khác nhau của tuần.
- **SA - Daypart (*Quan hệ một - nhiều 1:N*):** Một khách hàng có thể sử dụng các quán cafe vào một hoặc nhiều thời điểm khác nhau của cùng một ngày.
- **SA - Needstate (*Quan hệ một - nhiều 1:N*):** Một khách hàng có thể có một hoặc nhiều nhu cầu khác nhau khi sử dụng quán cafe
- **BrandHealth - SubBrandHealth (*Quan hệ một - một 1:1*):** Mỗi đánh giá của khách hàng về quán cafe có thông tin về chi tiêu và hành vi, phân khúc của duy nhất khách hàng có ID đã được xác định

CÂU HỎI 3:

MỤC TIÊU

Có một số phòng ban chức năng, bao gồm **Hội đồng Quản trị (BOD)**, **Nhân sự (HR)**, **Tài chính & Kế toán**, **Marketing và Bán hàng**. **Phòng Bán hàng** bao gồm nhiều cấp bậc khác nhau, bao gồm **Trưởng nhóm CRM**, **Quản lý Cửa hàng**, **Quản lý Khu vực và Vận hành Bán hàng**. Các phòng ban và vai trò khác nhau yêu cầu các mức độ truy cập dữ liệu khác nhau để thực hiện nhiệm vụ của họ, đồng thời đảm bảo tuân thủ và giảm thiểu rủi ro.

Đề xuất một chiến lược toàn diện để xác định quyền truy cập dữ liệu dựa trên vai trò cho các phòng ban và vai trò khác nhau nhằm đảm bảo cả bảo mật dữ liệu và hiệu quả vận hành.

Ví dụ về nhu cầu sử dụng dữ liệu theo vai trò

Phòng ban / Vai trò	Nhu cầu dữ liệu
Hội đồng Quản trị (BOD)	Xu hướng hiệu suất tổng thể, các chỉ số sức khỏe thương hiệu và các chuẩn chiến lược
Nhân sự / Phân tích phúc lợi	Các chỉ số hiệu suất cấp cửa hàng liên quan đến thường và khuyến khích nhân

	viên.
Tài chính và Kế toán	Chi tiết giao dịch, dữ liệu cho báo cáo lãi lỗ (P&L), và dự báo ngân sách.
Marketing	Chỉ số sức khỏe thương hiệu, vị thế cạnh tranh và phân khúc khách hàng.
Trưởng nhóm CRM	Các chỉ số đo lường lòng trung thành của khách hàng (ví dụ: NPS).
Vận hành bán hàng	Phân tích liên vùng để xác định hiệu quả vận hành và các khoảng trống cần cải thiện.
Quản lý khu vực	Xu hướng và so sánh hiệu suất của tất cả các cửa hàng trong khu vực quản lý.
Quản lý cửa hàng	Các chỉ số thời gian thực liên quan trực tiếp đến cửa hàng của họ (ví dụ: doanh số, tồn kho, nhân sự).

HƯỚNG DẪN

1. Định nghĩa vai trò và phân quyền:

Xác định rõ các vai trò phù hợp và trách nhiệm công việc cho từng phòng ban và chức năng công việc.

2. Phát triển ma trận truy cập dữ liệu:

Xây dựng một **ma trận truy cập dữ liệu chi tiết** phù hợp với yêu cầu tổ chức và các nghĩa vụ bảo mật, đảm bảo quyền truy cập dữ liệu tuân theo nguyên tắc **quyền truy cập tối thiểu (least privilege)** và **nhu cầu công việc**.

- Hàng (Rows):** Các thực thể dữ liệu / bảng dữ liệu.

- **Cột (Columns):** Các vai trò được chỉ định.
- **Mức độ truy cập (Access Levels):** Cấp độ kiểm soát truy cập chi tiết theo vai trò và chức năng đã đề cập.

3. Giải trình cuối cùng theo từng vai trò:

Đối với mỗi vai trò, đưa ra một giải thích ngắn gọn nhưng có lý lẽ rõ ràng giải thích lý do tại sao **mức độ truy cập dữ liệu cụ thể** lại phù hợp. Việc giải thích nên xem xét các yếu tố sau:

- Trách nhiệm vận hành của vai trò.
- Mức độ nhạy cảm của dữ liệu.
- Cấp bậc trong cơ cấu tổ chức.
- Rủi ro tiềm ẩn nếu dữ liệu bị lạm dụng hoặc tiết lộ quá mức.

1. Định nghĩa vai trò và phân quyền:

Phòng ban	Vai trò	Phân quyền
Hội đồng Quản trị (BOD)	Giám sát và định hướng chiến lược tổng thể của doanh nghiệp, đảm bảo hiệu quả hoạt động và tuân thủ các mục tiêu dài hạn.	<ul style="list-style-type: none"> - Xu hướng hiệu suất tổng thể - Chỉ số sức khỏe thương hiệu - Các chỉ số chiến lược cấp cao
Nhân sự / Phân tích phúc lợi	Chuyên viên phân tích phúc lợi và hiệu suất nhân viên.	<ul style="list-style-type: none"> - Hiệu suất cấp cửa hàng liên quan đến lương thưởng và khuyến khích - Dữ liệu nhân sự tổng hợp
Tài chính và Kế toán	Phân tích và quản lý tài chính, kế toán và ngân sách.	<ul style="list-style-type: none"> - Chi tiết giao dịch bán hàng

		<ul style="list-style-type: none"> - Dữ liệu phục vụ báo cáo lãi/lỗ (P&L) - Dự báo và lập ngân sách
Marketing	Phân tích thị trường, thương hiệu và hành vi khách hàng.	<ul style="list-style-type: none"> - Chỉ số sức khỏe thương hiệu - Vị thế cạnh tranh - Phân khúc khách hàng và hành vi tiêu dùng
Trưởng nhóm CRM	Trưởng nhóm quản lý quan hệ khách hàng (CRM Lead)	<ul style="list-style-type: none"> - Chỉ số đo lường lòng trung thành khách hàng (VD: NPS, tần suất quay lại) - Dữ liệu phân tích hành vi cá nhân
Vận hành bán hàng	Phân tích và hỗ trợ tối ưu hiệu quả vận hành bán hàng trên toàn hệ thống.	<ul style="list-style-type: none"> - Phân tích hiệu suất theo khu vực / vùng - Xác định các khoảng trống và cơ hội cải thiện
Quản lý khu vực	Quản lý khu vực	<ul style="list-style-type: none"> - So sánh hiệu suất giữa các cửa hàng trong khu vực quản lý - Phát hiện xu hướng tăng / giảm hiệu quả
Quản lý cửa hàng	Quản lý cửa hàng	<ul style="list-style-type: none"> - Chỉ số thời gian thực tại cửa hàng (VD: doanh thu, tồn kho, nhân sự, khách

		hàng) - Dữ liệu vận hành hàng ngày
--	--	---------------------------------------

2. Phát triển ma trận truy cập dữ liệu:

Kí hiệu:

BOD: Hội đồng Quản trị (BOD)

HR: Nhân sự / Phân tích phúc lợi

F: Tài chính và Kế toán

M: Marketing

Lead CRM: Trưởng nhóm CRM

Sales Ops: Vận hành bán hàng

R Mgr: Quản lý khu vực

S Mgr: Quản lý cửa hàng

Mức truy cập	Giải thích
No	Không được phép truy cập , cột dữ liệu sẽ bị ẩn hoàn toàn với vai trò đó.
Limit	Truy cập hạn chế , chỉ thấy dữ liệu đã được làm mờ, nhóm lại hoặc ẩn bớt thông tin nhạy cảm.
View	Xem được dữ liệu ở cấp độ tổng hợp hoặc thống kê, không thấy chi tiết theo từng cá nhân.
Full	Truy cập đầy đủ , xem được toàn bộ dữ liệu gốc, chi tiết đến từng cá nhân hoặc từng dòng dữ liệu.

FULL > VIEW > LIMIT > NO

BẢNG SA: Bảng SA chứa dữ liệu chi tiết về hồ sơ nhân khẩu học và hành vi của khách ghé quán cà phê, nhằm phục vụ cho việc phân khúc và phân tích mục tiêu marketing.

	BOD	HR	F	M	Lead CR M	Sales Ops	R Mgr	S Mgr
ID	View	No	Limit	Limit	Limit	Limit	Limit	Limit
City	View	Limit	View	Full	Limit	Full	View	View
GroupSize	View	Limit	View	Full	Full	Limit	View	View
Age	View	View	View	Full	Full	View	View	View
MPIMean	View	No	Limit	Full	Full	View	View	View
TOM	View	View	View	Full	Full	View	View	View
BUMO	View	View	View	Full	Full	Full	View	View
BUMOPrevious	View	View	View	Full	Full	Full	View	View
MostFavourite	View	View	View	Full	Full	View	View	View
Gender	View	View	View	Full	Full	Limit	View	View
AgeGroup	View	View	View	Full	Full	View	View	View
Occupation	View	Limit	Limit	Full	Full	Limit	Limit	View
OccupationGroup	View	Limit	Limit	Full	Full	Limit	Limit	View
Year	View	View	View	Full	Full	Full	View	View
MPIRange	View	No	Full	Full	Full	View	Limit	Limit

BẢNG NEEDSTATE: Ghi nhận các nhu cầu tiêu dùng gắn với thời điểm trong ngày hoặc hành vi theo ngày, cung cấp hiểu biết về lý do khách hàng ghé quán cà phê vào những thời điểm cụ thể.

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
ID	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
City	View	Limit	View	Full	View	View	Limit	Limit
Year	View	View	View	Full	Full	Full	View	View
Needstates	View	View	View	Full	Full	View	Limit	Limit
Day/DayPart	View	View	View	Full	Full	Full	Limit	Limit
NeedstateGroup	View	View	Limit	Full	Full	Limit	Limit	Limit

BẢNG DAYPART: Chứa dữ liệu về tần suất ghé quán của khách hàng theo thời gian trong ngày, giúp xác định các khung giờ cao điểm của quán cà phê theo từng thành phố và năm.

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
ID	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
City	View	Limit	View	Full	View	View	Limit	Limit

Daypart	View	View	Limit	Full	Full	Full	Limit	Limit
VisitOnDaypart	View	View	Limit	Full	Full	Full	Limit	Limit
Year	View	View	View	Full	Full	Full	View	View

BẢNG DAYOFWEEK: Chứa dữ liệu về các ngày trong tuần mà người tiêu dùng thường đến quán cà phê, bao gồm tần suất ghé thăm và phân loại theo ngày trong tuần hoặc cuối tuần.

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
ID	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
City	View	Limit	View	Full	View	View	Limit	Limit
DayOf Week	View	View	Limit	Full	Full	Full	Limit	Limit
VisitOnDayOffWeek	View	View	Limit	Full	Full	Full	Limit	Limit
Year	View	View	View	Full	Full	Full	View	View
TypeOfDay	View	View	Limit	Full	Full	Full	Limit	Limit

BẢNG COMPETITOR: Chứa số lượng cửa hàng thực tế của từng thương hiệu cà phê theo thành phố và năm, dùng để đánh giá mức độ hiện diện trên thị trường và mật độ cạnh tranh.

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
No#	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
Brand	View	No	Limit	Full	Full	Full	Limit	Limit
City	View	Limit	Limit	Full	Full	Full	Limit	Limit
Year	View	View	Limit	Full	Full	Full	Limit	Limit
StoreCount	View	No	Limit	Full	Full	Full	Limit	Limit

BẢNG COMPANION: Chứa thông tin về kiểu người đi cùng phổ biến (ví dụ: bạn bè, gia đình, đi một mình) mà khách hàng có khi đến quán cà phê.

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
ID	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
City	View	Limit	View	Full	View	View	Limit	Limit
CompanionGroup	View	View	Limit	Full	Full	View	Limit	Limit
Year	View	View	Limit	Full	Full	Full	Limit	Limit

BẢNG BRAND_IMAGE: Chứa dữ liệu về mức độ nhận biết thương hiệu của người tiêu dùng, cảm nhận về các thuộc tính và liên tưởng hình ảnh thương hiệu theo từng thành phố và năm.

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
ID	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
Year	View	View	View	Full	Full	Full	View	View
City	View	Limit	View	Full	View	View	Limit	Limit
Aware ness	View	No	Limit	Full	View	View	Limit	Limit
Attrib ute	View	No	No	Full	View	View	Limit	Limit
BrandI mage	View	No	No	Full	View	View	Limit	Limit

BẢNG BRANDHEALTH: Chứa các phản hồi khảo sát chi tiết đo lường mức độ nhận biết thương hiệu, mức độ sử dụng, cảm nhận và phân khúc khách hàng giữa các thương hiệu cà phê.

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
ID	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
Year	View	View	View	Full	Full	Full	View	View
City	View	Limit	View	Full	View	View	Limit	Limit
Brand	View	Limit	Limit	Full	Full	Full	Limit	Limit

Spontaneous	View	No	View	Full	Full	View	Limit	Limit
Awareness	View	No	View	Full	Full	View	Limit	Limit
Trial	View	No	View	Full	Full	View	Limit	Limit
P3M	View	No	Limit	Full	Full	Full	Limit	Limit
P1M	View	No	Limit	Full	Full	Full	Limit	Limit
Comprehension	View	No	View	Full	Full	Full	Limit	Limit
Weekly	View	No	Limit	Full	Full	Full	Limit	Limit
Daily	View	No	Limit	Full	Full	Full	Limit	Limit
VisitFrequency	View	No	Limit	Full	Full	Full	Limit	Limit
Segmentation	View	No	View	Full	Full	Full	Limit	Limit
NPSPer3Month	View	No	Limit	Full	Full	Full	Limit	Limit
NPSGroup	View	No	Limit	Full	Full	Full	Limit	Limit

BẢNG SubBRANDHEALTH

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
PPA	View	No	Full	Full	Full	Limit	Limit	Limit
Spending	View	No	Full	Full	Full	Limit	Limit	Limit
VisitFrequency	View	No	Limit	Full	Full	Full	Limit	Limit
Segmentation	View	No	View	Full	Full	Full	Limit	Limit

BẢNG CUSTOMERSEGMENTATION

	BOD	HR	F	M	Lead CRM	Sales Ops	R Mgr	S Mgr
ID	Limit	No	Limit	Limit	Limit	Limit	Limit	Limit
Segmentation	View	No	Limit	Full	Full	View	Limit	Limit
Visit	View	No	Limit	Full	Full	Full	Limit	Limit
Spending	View	No	Full	Full	Full	Limit	Limit	Limit
Brand	View	No	Limit	Full	Full	Full	Full	Full
PPA	View	No	Full	Full	Full	Limit	Limit	Limit

3. Giải trình cuối cùng theo từng vai trò:

BOD

- **Trách nhiệm:** Đưa ra chiến lược và giám sát hiệu quả tổng thể.
- **Giải trình:**
 - + Chỉ cần View hoặc Limit, tránh truy cập dữ liệu định danh.
 - + Họ cần tổng quan để nắm xu hướng thị trường, hành vi tiêu dùng và hiệu quả thương hiệu.
 - + Không nên thấy thông tin vi mô để giảm rủi ro rò rỉ hoặc vượt quyền điều hành.

HR

- **Trách nhiệm:** Tuyển dụng, phân tích hiệu suất, phúc lợi của nhân viên.
- **Giải trình:**
 - + Hầu hết các bảng là No hoặc Limit, tránh tiếp cận dữ liệu cá nhân khách hàng.
 - + Một số cột View hỗ trợ khi cần phân tích hành vi nhân viên gián tiếp (VD: theo nhóm nhân khẩu học, vùng miền).
 - + Tránh rủi ro sử dụng dữ liệu ngoài phạm vi công việc (như profiling sai mục đích).

Finance

- **Trách nhiệm:** Phân tích lợi nhuận, dự báo, quản lý tài chính và hiệu quả chi tiêu.
- **Giải trình:**
 - + Cần View hoặc Limit dữ liệu có liên quan đến hành vi chi tiêu và thương hiệu.
 - + Không cần truy cập thông tin cá nhân để tránh rò rỉ dữ liệu định danh.
 - + Phân quyền tránh trùng lặp chức năng với Marketing/CRM.

Marketing

- **Trách nhiệm:** Phân khúc khách hàng, định vị thương hiệu, ra chiến lược truyền thông.
- **Giải trình:**
 - + Cần Full quyền truy cập gần như toàn bộ bảng dữ liệu.

- + Là bộ phận chủ chốt sử dụng các bảng để ra quyết định chiến dịch.
- + Chấp nhận rủi ro dữ liệu cao hơn nhưng cần tuân thủ chính sách bảo mật nội bộ.

CRM Lead

- **Trách nhiệm:** Giữ chân khách hàng, đo lường lòng trung thành (NPS, PPA, frequency...).
- **Giải trình:**
 - + Cần Full quyền ở các bảng segmentation, needstate, brandhealth để xây dựng mô hình chăm sóc khách hàng hiệu quả.
 - + Cần truy cập thời gian ghép quan, chi tiêu, hành vi... để cá nhân hóa hành động.
 - + Cần kiểm soát nội bộ chặt để ngăn lạm dụng dữ liệu cá nhân.

Sales Ops

- **Trách nhiệm:** Đảm bảo hoạt động vận hành hiệu quả giữa các vùng/kênh.
- **Giải trình:**
 - + Cần Limit hoặc View để xem xu hướng chung nhưng không truy cập sâu dữ liệu định danh.
 - + Tập trung tối ưu quy trình chứ không chăm sóc cá nhân nên không cần Full access.
 - + Rủi ro rò rỉ thấp hơn CRM, nhưng vẫn cần kiểm soát.

Regional Manager

- **Trách nhiệm:** Theo dõi hiệu suất vùng, hỗ trợ quản lý cửa hàng.
- **Giải trình:**
 - + Cần View các trường liên quan hành vi, phân khúc, thương hiệu để ra quyết định vùng.
 - + Không cần dữ liệu gốc nhưng cần truy cập tổng hợp theo khu vực.
 - + Phân quyền giúp tránh truy cập không liên quan đến vùng phụ trách.

Store Manager (Quản lý cửa hàng)

- **Trách nhiệm:** Quản lý hoạt động cửa hàng hằng ngày.

- **Giải trình:**
 - + Cần View hoặc Limit, chỉ giới hạn trong phạm vi cửa hàng mình phụ trách.
 - + Cần các thông tin như thời điểm khách ghé, nhóm khách chính, nhu cầu theo ngày/giờ để tối ưu vận hành.
 - + Không có quyền truy cập ID hoặc thông tin thương hiệu rộng toàn hệ thống.

CÂU HỎI 4:

Phân tích Cạnh tranh và Định vị Thương hiệu cho Highlands Coffee

MỤC TIÊU

Thực hiện phân tích toàn diện về bối cảnh cạnh tranh của Highlands Coffee và các đối thủ chính, sử dụng các dữ liệu liên quan đến **phễu thương hiệu** (brand funnel), nhận thức khách hàng và mức độ hiện diện trên thị trường.

Mục tiêu của nhiệm vụ này là **tạo ra các insight có thể hành động được** về hiệu suất thương hiệu ở các **giai đoạn tương tác khách hàng khác nhau** và các khía cạnh của **sức khỏe thương hiệu** (brand health), từ đó hỗ trợ cho việc **định vị chiến lược** và **lập kế hoạch tăng trưởng**.

HOÀN THÀNH

1. **Bảng điều khiển cạnh tranh tương tác** (Interactive Competitive Dashboard) kèm theo phân tích hỗ trợ, cho phép các bên liên quan:
 - **Theo dõi hiệu suất thương hiệu trong suốt hành trình khách hàng:** So sánh cách Highlands và các đối thủ dẫn dắt khách hàng từ giai đoạn nhận biết đến trung thành, từ đó xác định các giai đoạn bị rò rỉ (leakage) và các giai đoạn chuyển đổi cao.
 - **Chẩn đoán khoảng cách về nhận thức thương hiệu** (Brand Perception Gaps): Trực quan hóa các thuộc tính thương hiệu (ví dụ: "chất lượng cao cấp", "giá trị tốt") đang thúc đẩy sự yêu thích dành cho đối thủ, nhưng lại yếu kém ở Highlands.

- **Lập bản đồ định vị cạnh tranh (Competitive Positioning):** Xác định "khoảng trống thị trường" (white space opportunities) bằng cách vẽ đồ thị vị trí các thương hiệu theo các trực nhận thức chính (ví dụ: giá cả vs. sự tiện lợi).
- **Đánh giá sức mạnh thị trường so với mức độ trung thành:** Ưu tiên các thị trường mà Highlands có thị phần lớn nhưng mức độ giữ chân khách hàng thấp, cho thấy mức độ hiện diện không tương ứng với lòng trung thành.
- **Tích hợp các bộ lọc tương tác để khám phá Insight theo từng phân khúc khách hàng.**

2. Diễn giải và Đề xuất chiến lược cho từng biểu đồ:

- Xác định điểm mạnh/yếu ở từng giai đoạn trong hành trình khách hàng (từ nhận biết → trung thành).
- So sánh "sức khỏe thương hiệu" (brand health) giữa Highlands và đối thủ qua các chỉ số như: mức độ khác biệt hóa, mức độ liên quan (relevance).
- Khám phá các cơ hội chiến lược để:
 - + Tái định vị thương hiệu;
 - + Tối ưu hóa chuyển đổi khách hàng;
 - + Tăng cường mức độ yêu thích và giữ chân khách hàng;
 - + Mở rộng thị trường còn nhiều khoảng trống.

Tóm tắt từ câu 2:

XÁC ĐỊNH MÔ HÌNH DỮ LIỆU VÀ LIÊN KẾT CÁC BẢNG

Bảng trung tâm: SA (Survey Answer / Người tham gia khảo sát)

- Chứa thông tin nhân khẩu và hành vi cá nhân:
 - + ID: Mã định danh từng người khảo sát (khóa chính).
 - + City, GroupSize, Age, MPIMean, TOM, BUMO, BUMOPrevious, MostFavourite, Gender, AgeGroup, Occupation, OccupationGroup, Year, MPIRange.

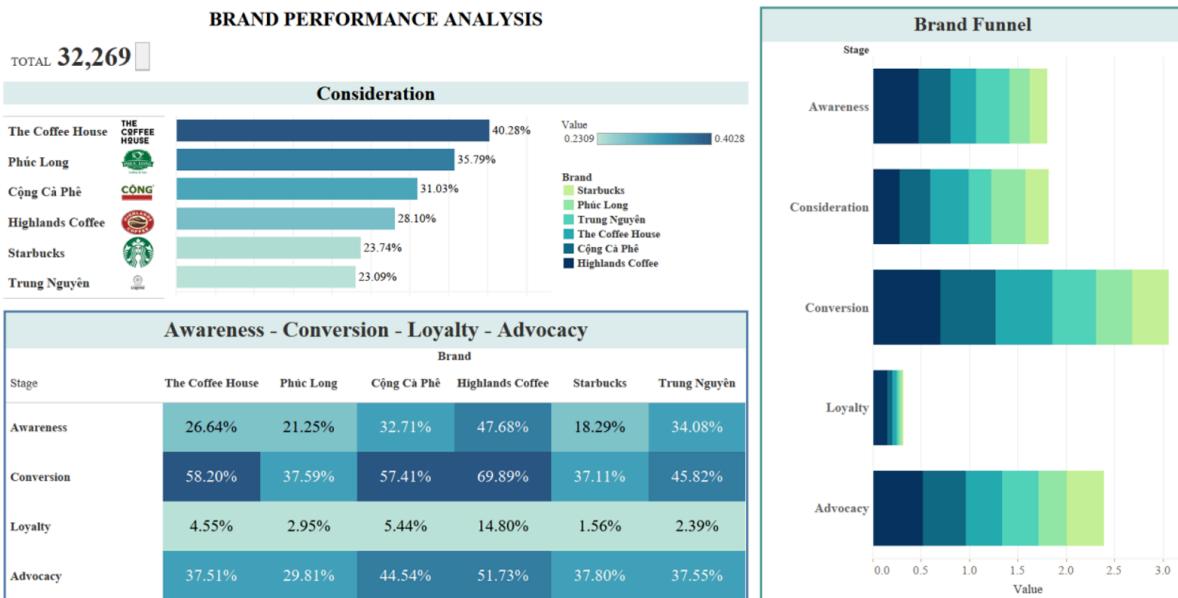
Đây là bảng gốc liên kết đến các bảng còn lại.

Bảng con	Khóa liên kết chính	Mục đích
CustomerSegmentation	SAID	Phân khúc khách hàng

Companion	SAID	Người đi cùng
Dayofweek	SAID	Thông tin đến quán cafe vào các ngày của một tuần
Daypart	SAID	Khung giờ trong ngày của người khảo sát đến quán cafe
NeedState	SAID	Nhu cầu của những người đến quán cafe theo ngày trong tuần và theo khung giờ
BrandHealth	SAID	Hiệu suất thương hiệu
Competitor	BrandHealth -> SAID	Mức độ hiện diện thị trường
BrandImage	BrandHealth -> SAID	Hình ảnh thương hiệu

DASHBOARD 1

- **Theo dõi hiệu suất thương hiệu trong suốt hành trình khách hàng:** So sánh cách Highlands và các đối thủ dẫn dắt khách hàng từ giai đoạn nhận biết đến trung thành, từ đó xác định các giai đoạn bị rò rỉ (leakage) và các giai đoạn chuyển đổi cao.



Phân Tích Toàn Diện Bối Cảnh Cạnh Tranh: Highlands Coffee vs. Đối Thủ

Tóm Tắt:

Bức tranh tổng thể cho thấy Highlands Coffee là một "Người khổng lồ" dẫn đầu thị trường về độ phủ và hiệu suất cuối phễu, nhưng lại tồn tại một "yếu huyệt" chí mạng ở giai đoạn Cân nhắc (Consideration).

- Thé mạnh tuyệt đối:** Highlands không có đối thủ về khả năng xây dựng **Nhận biết (Awareness)** và sức mạnh **Chuyển đổi (Conversion)**. Khi một khách hàng đã có ý định mua, Highlands có khả năng "chốt đơn" hiệu quả nhất. Khách hàng sau khi mua cũng thể hiện mức độ **Trung thành (Loyalty)** và **Úng hộ (Advocacy)** cao nhất.
- Thách thức lớn nhất:** Thương hiệu đang bị "rò rỉ" nghiêm trọng ở giai đoạn giữa phễu. Khách hàng "biết" Highlands, nhưng khi được hỏi "Bạn sẽ đi uống cà phê ở đâu?", họ lại "nghĩ đến" The Coffee House và Phúc Long nhiều hơn.
- Cơ hội chiến lược:** Cơ hội tăng trưởng lớn nhất không nằm ở việc thu hút thêm người mới biết đến, mà là **thuyết phục những người đã biết chọn Highlands** thay vì các đối thủ khác.

Điễn Giải và Đề Xuất Chiến Lược

1. Phân Tích Phễu Khách Hàng (Brand Funnel): Biểu đồ này trực quan hóa toàn bộ hành trình và cho thấy rõ hình dạng phễu của từng thương hiệu.

- **Giai đoạn 1: Nhận biết (Awareness)**

- + Heatmap và Brand Funnel đều cho thấy Highlands là vua về độ nhận biết (47.68%). Vấn đề "làm cho người ta biết đến mình" về cơ bản đã được giải quyết. Đây là một tài sản thương hiệu khổng lồ.
- + **Đè xuất:** Duy trì vị thế này thông qua các hoạt động thương hiệu ở quy mô lớn, nhưng có thể giảm bớt ngân sách cho các hoạt động chỉ nhằm mục tiêu tăng awareness đơn thuần.

- **Giai đoạn 2: Cân nhắc (Consideration)**

- + **Đây là "vùng báo động đỏ".** Biểu đồ Bar Chart cho thấy Highlands (28.10%) đang bị The Coffee House (40.28%) và Phúc Long (35.79%) bỏ xa. Điều này có nghĩa là dù có độ phủ lớn, Highlands không phải là lựa chọn "top-of-mind" (ưu tiên hàng đầu trong tâm trí) của khách hàng.
- + **Đè xuất:** Đây là nơi cần dồn toàn lực để cải thiện.
 - **Tái định vị thương hiệu:** Truyền thông mạnh mẽ hơn về các "lý do để đến" (reason to go) - không gian làm việc, một món nước đặc trưng, một câu chuyện thương hiệu mới... thay vì chỉ là nơi "tiện thì ghé".
 - **Tối ưu hóa trải nghiệm:** Cải thiện không gian, chất lượng dịch vụ để tạo ra một sức hút về mặt trải nghiệm, cạnh tranh trực tiếp với thế mạnh của The Coffee House.

- **Giai đoạn 3: Chuyển đổi (Conversion)**

- + Đây là điểm mạnh vượt trội và đáng tự hào nhất của Highlands. Với tỷ lệ lên đến 69.89%, Highlands có khả năng biến người cân nhắc thành người mua hàng hiệu quả nhất thị trường. Điều này cho thấy các chương trình khuyến mãi, sự tiện lợi của hệ thống, và hiệu quả của ứng dụng (app) đang hoạt động rất tốt.

- + **Đề xuất:** Duy trì và phân tích sâu hơn để tìm ra yếu tố thành công cốt lõi. Liệu có phải do chương trình "Mua 1 Tặng 1" hay một tính năng nào đó trên app? Nhận rộng các yếu tố này.
- **Giai đoạn 4 & 5: Trung thành (Loyalty) & Ủng hộ (Advocacy)**
 - + Highlands tiếp tục dẫn đầu ở hai giai đoạn cuối phễu này. Điều này chứng tỏ rằng, một khi khách hàng đã được thuyết phục mua hàng, họ có xu hướng rất hài lòng và quay trở lại. Vấn đề không nằm ở chất lượng sản phẩm hay dịch vụ ở giai đoạn này, mà là làm sao để có nhiều người hơn nữa đi đến được giai đoạn này.
 - + **Đề xuất:** Tận dụng lượng khách hàng trung thành này. Triển khai các chương trình "Referral" (Giới thiệu bạn bè) để họ trở thành những đại sứ thương hiệu, giúp kéo thêm khách hàng mới và lấp đầy khoảng trống ở giai đoạn "Cân nhắc".

2. So Sánh "Sức Khỏe Thương Hiệu" (Brand Health)

- **Mức độ liên quan (Relevance):** Highlands có mức độ liên quan cao, nhưng The Coffee House và Phúc Long đang tỏ ra "hợp thời" và "hấp dẫn" hơn đối với một nhóm khách hàng nhất định (thể hiện qua chỉ số Consideration).
- **Mức độ khác biệt hóa (Differentiation):** Highlands đang tự khác biệt hóa mình bằng hiệu suất và sự phổ biến, trong khi The Coffee House khác biệt hóa bằng trải nghiệm và không gian. Vấn đề là sự khác biệt hóa của The Coffee House đang có sức hút mạnh hơn ở giai đoạn ra quyết định.

Khám Phá Các Cơ Hội Chiến Lược

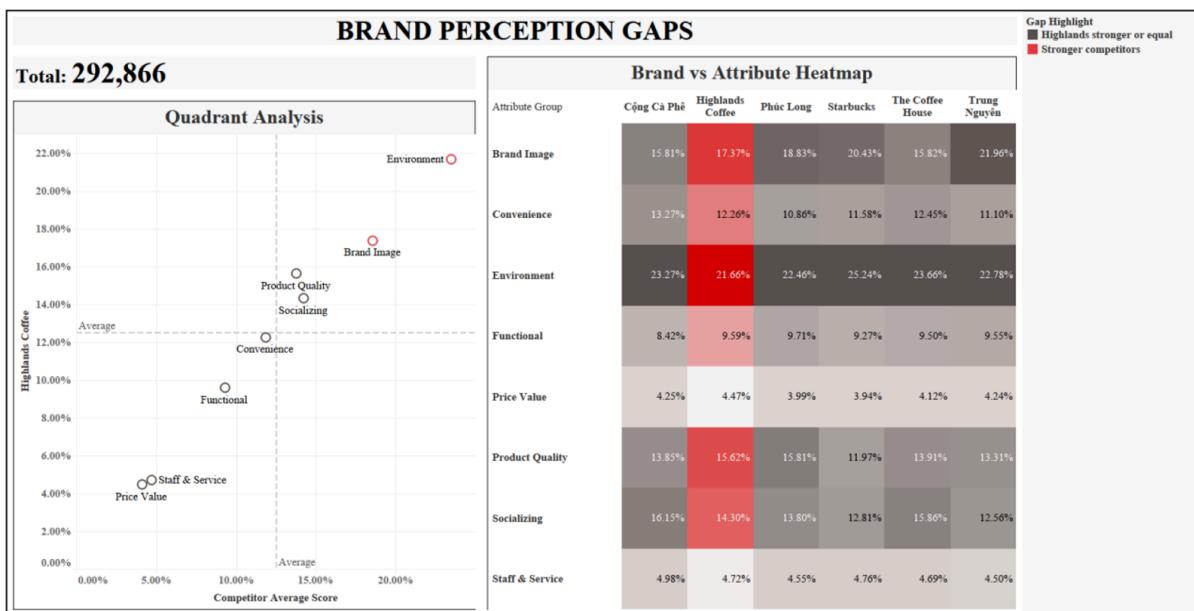
- **Tái định vị thương hiệu:** Chuyển dịch hình ảnh từ một chuỗi cà phê "tiện lợi, ở đâu cũng có" thành một "điểm đến có chủ đích" với những giá trị trải nghiệm rõ ràng hơn.
- **Tối ưu hóa chuyển đổi khách hàng:** Cân tập trung vào việc tối ưu hóa chuyển đổi từ Awareness sang Consideration. Đây là ưu tiên số một. Cần các chiến dịch quảng cáo tập trung vào việc gợi lên sự mong muốn và tạo ra lý do để ghé thăm, chứ không chỉ là nhắc nhở về sự tồn tại.
- **Tăng cường mức độ yêu thích và giữ chân:** Mặc dù Loyalty Rate đang dẫn đầu, nhưng con số tuyệt đối 14.80% vẫn còn khiêm tốn. Điều này cho thấy toàn thị trường vẫn còn rất nhiều dư địa. Việc đầu tư vào cá nhân hóa trải nghiệm cho

khách hàng thành viên sẽ giúp Highlands tạo ra một khoảng cách an toàn hơn với các đối thủ.

- **Mở rộng thị trường còn nhiều khoảng trống:** "Khoảng trống" lớn nhất không nằm ở địa lý, mà nằm ở **nhận thức**. Chưa có thương hiệu nào vừa có độ nhận biết cao như Highlands, lại vừa có sức hút về cảm nhận như The Coffee House. Đây chính là "**vùng đất vàng**" mà Highlands cần phải chiếm lĩnh để trở thành người dẫn đầu tuyệt đối trên mọi phương diện.

DASHBOARD 2

- **Chẩn đoán khoảng cách về nhận thức thương hiệu (Brand Perception Gaps):** Trực quan hóa các thuộc tính thương hiệu (ví dụ: "chất lượng cao cấp", "giá trị tốt") đang thúc đẩy sự yêu thích dành cho đối thủ, nhưng lại yếu kém ở Highland.



Phân Tích Toàn Diện: Chẩn Đoán Khoảng Trống Nhận Thức Highlands Coffee

Tóm tắt:

- Bức tranh toàn cảnh cho thấy Highlands Coffee đang đối mặt với một **thách thức lớn về định vị nhận thức**. Phân tích chỉ ra rằng thương hiệu đang **yêu thê ở chính những thuộc tính "thời thượng" và mang tính trải nghiệm** (Environment, Brand Image) mà các đối thủ đang sử dụng để xây dựng sự yêu thích.
 - + **Khoảng trống lớn nhất:** Highlands đang bị bỏ lại phía sau trong cuộc đua về "Không gian trải nghiệm" và "Hình ảnh thương hiệu hiện đại".

- + **Điểm mạnh không rõ ràng:** Đáng báo động, dashboard cho thấy Highlands **không sở hữu một điểm mạnh độc nhất** nào trong nhận thức của khách hàng (không có thuộc tính nào nằm ở góc "Vùng Độc Tôn").
- + **Cơ hội trong thách thức:** Toàn bộ thị trường, bao gồm cả các đối thủ, đều đang **chưa làm tốt các thuộc tính nền tảng** như Product Quality và Staff & Service. Đây chính là "khoảng trống thị trường" để Highlands bứt phá và tạo ra sự khác biệt bền vững.

Kết luận: Chiến lược trước mắt cần tập trung vào việc **thu hẹp khoảng cách về trải nghiệm** và cân nhắc **tái định vị để sở hữu một thuộc tính cốt lõi** mà không đối thủ nào có thể cạnh tranh.

Diễn Giải Chi Tiết

1. **Biểu Đồ "Quadrant Analysis" - Bản Đồ Chiến Lược:** biểu đồ này là công cụ chẩn đoán chính, phân loại các thuộc tính vào 4 vùng chiến lược:
 - **Góc trên bên trái - "VÙNG NGUY HIỂM" (Đối thủ mạnh, Highlands yếu):**
 - + **Insight:** Đây là yếu huyệt lớn nhất và cấp bách nhất của Highlands. Hai thuộc tính quan trọng là Environment (Không gian) và Brand Image (Hình ảnh thương hiệu) đang nằm gọn trong vùng này. Điều này có nghĩa là khách hàng cảm thấy các đối thủ mang lại trải nghiệm không gian và xây dựng hình ảnh thương hiệu tốt hơn hẳn Highlands.
 - + **Hậu quả:** Highlands có nguy cơ mất dần nhóm khách hàng trẻ, những người ưu tiên trải nghiệm và "check-in" khi chọn một quán cà phê.
 - **Góc dưới bên trái - "VÙNG CẦN CẢI THIỆN CHUNG" (Cả hai cùng yếu):**
 - + **Insight:** Đây là góc đông đúc nhất, chứa hàng loạt các thuộc tính nền tảng như Product Quality, Price Value, Staff & Service, Functional, Convenience. Điều này cho thấy **toàn bộ thị trường đều chưa thực sự xuất sắc** ở những mặt này trong mắt khách hàng. Không có một "nhà vô địch" rõ ràng.
 - + **Cơ hội:** Đây chính là "khoảng trống thị trường" lớn nhất. Nếu Highlands có thể đầu tư và cải thiện vượt bậc để trở thành người dẫn đầu ở chỉ một trong các thuộc tính này (ví dụ: "Thương hiệu có chất lượng sản phẩm đáng tin cậy nhất"), họ sẽ tạo ra một lợi thế cạnh tranh cực lớn.
 - **Hai góc còn lại - Đáng Báo Động Vì... Trống Rỗng:**

- + **Góc dưới bên phải ("VÙNG ĐỘC TÔN" - Highlands mạnh, đối thủ yếu):** Vùng này trống trơn. Điều này xác nhận Highlands **không có một thế mạnh độc nhất, khác biệt nào** trong nhận thức khách hàng.
- + **Góc trên bên phải ("VÙNG CẠNH TRANH" - Cả hai cùng mạnh):** Vùng này cũng trống. Điều này cho thấy không có thuộc tính nào mà cả thị trường cùng đang làm tốt và cạnh tranh trực tiếp ở mức độ cao.

2. Biểu Đồ "Heatmap" - Bằng Chứng Chi Tiết: heatmap cung cấp những con số cụ thể để chứng minh cho các nhận định từ Quadrant Analysis.

- **Ví dụ 1 (Chứng minh điểm yếu):** Nhìn vào hàng Environment, ô của Highlands có màu đỏ (điểm thấp, 21.66%) trong khi ô của Starbucks (25.24%) và The Coffee House (23.66%) có màu xanh (điểm cao). Điều này giải thích tại sao Environment nằm trong "Vùng Nguy Hiểm".
- **Ví dụ 2 (Chứng minh cơ hội):** Nhìn vào hàng Staff & Service, tất cả các thương hiệu đều có màu đỏ và điểm số rất thấp (quanh 4-5%). Điều này khẳng định rằng không có ai làm tốt về dịch vụ, và cơ hội đang rộng mở cho người tiên phong.

Đề Xuất Chiến Lược & Khám Phá Cơ Hội

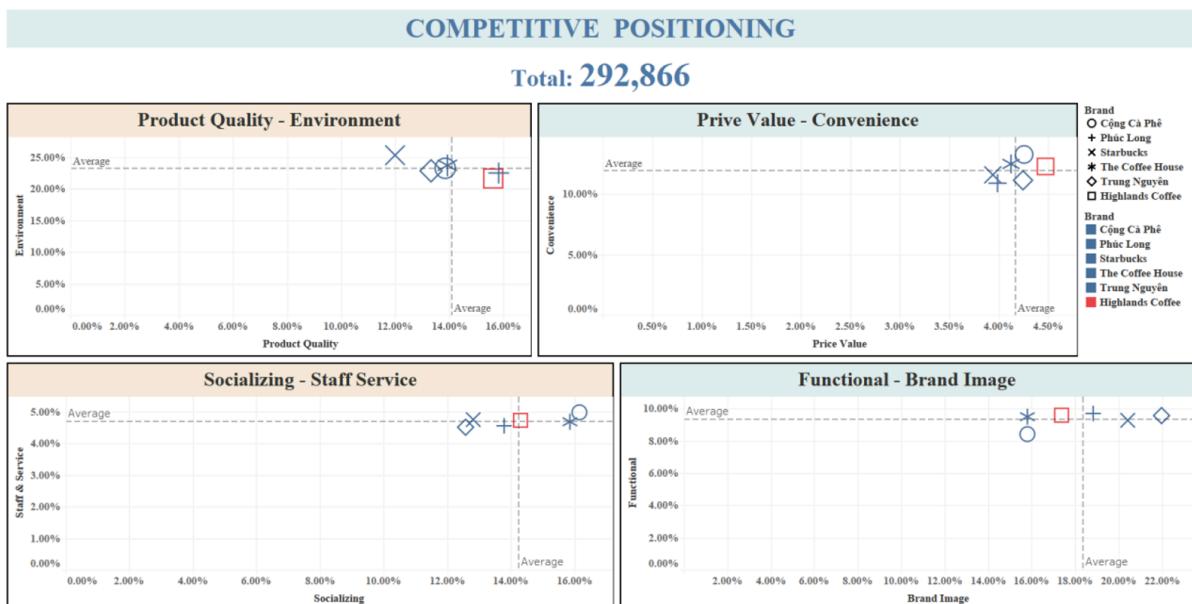
Dựa trên các phân tích trên, đây là những hành động chiến lược mà Highlands có thể cân nhắc:

- **Tái định vị thương hiệu (Brand Repositioning):**
 - + **Ưu tiên số 1: "Cải tử hoàn sinh" cho 'Environment' và 'Brand Image'.** Highlands cần một kế hoạch hành động quyết liệt: làm mới thiết kế cửa hàng theo một chủ đề nhất quán và hiện đại, tạo ra các chiến dịch xây dựng hình ảnh thương hiệu trẻ trung hơn để cạnh tranh sòng phẳng với The Coffee House.
- **Tối ưu hóa & Tăng cường (Optimization & Enhancement):**
 - + **Biến điểm yếu chung của thị trường thành thế mạnh riêng:** Thay vì chạy theo cuộc đua về "không gian" vốn đã rất tốn kém, Highlands có thể chọn một con đường khác biệt. Hãy đầu tư mạnh mẽ để trở thành thương hiệu "**Vô địch về Chất lượng Sản phẩm**" hoặc "**Vô địch về Dịch vụ Khách hàng**". Đây là những thuộc tính nền tảng, một khi đã dẫn đầu sẽ tạo ra lòng trung thành rất bền vững và khó sao chép.
- **Khai thác "Khoảng Trống Thị Trường":**

- + "Khoảng trống" ở đây không phải là một cắp thuộc tính chưa ai khai thác, mà là **cơ hội để trở thành người dẫn đầu tuyệt đối ở một thuộc tính mà tất cả đối thủ đều đang làm chưa tốt.**
- + **Chiến lược gợi ý:** Ra mắt một chiến dịch toàn quốc cam kết về "Chất lượng sản phẩm không đổi" hoặc một chương trình đào tạo dịch vụ khách hàng quy mô lớn để thực sự chiếm lĩnh nhận thức của khách hàng về các thuộc tính này.

DASHBOARD 3

- **Lập bản đồ định vị cạnh tranh (Competitive Positioning):** Xác định "khoảng trống thị trường" (whitespace opportunities) bằng cách vẽ đồ thị vị trí các thương hiệu theo các trực nhận thức chính (ví dụ: giá cả vs. sự tiện lợi).



Phân Tích Toàn Diện: Lập Bản Đồ Định Vị Cạnh Tranh & Xác Định Khoảng Trống Thị Trường

Tóm tắt:

Dashboard này khẳng định một điều rõ ràng: **Thị trường chuỗi cà phê Việt Nam không có một người dẫn đầu tuyệt đối trên mọi phương diện.** Thay vào đó, mỗi thương hiệu lớn đã thành công trong việc chiếm lĩnh một "lãnh địa nhận thức" riêng biệt trong tâm trí khách hàng.

- **Vị thế của Highlands Coffee:** Highlands đã xây dựng thành công một định vị vững chắc là một thương hiệu **đáng tin cậy, tập trung vào chức năng và chất**

lượng sản phẩm. Tuy nhiên, thương hiệu đang bị bỏ lại phía sau trong cuộc đua về các thuộc tính **trải nghiệm, cảm xúc và hình ảnh thương hiệu** so với các đối thủ chính.

- **Bối cảnh cạnh tranh:** Cuộc chiến khốc liệt nhất đang diễn ra ở các thuộc tính liên quan đến **Không gian (Environment)** và **Hình ảnh (Brand Image)**, nơi The Coffee House và Starbucks đang chiếm ưu thế.
- **Khoảng trống thị trường (Whitespace):** Cơ hội lớn nhất không nằm ở việc tạo ra một cặp thuộc tính mới, mà là trở thành "**Nhà Vô Địch**" ở **những thuộc tính nền tảng mà cả thị trường đều đang làm chưa tốt**, đặc biệt là **Dịch vụ Khách hàng (Staff & Service)** và **Giá trị Cảm nhận (Price Value)**.

Diễn giải chi tiết:

Mỗi biểu đồ scatter plot là một "bản đồ" kể một câu chuyện khác nhau về cuộc chiến giành tâm trí khách hàng.

1. Bản đồ "Product Quality vs Environment" (Chất lượng vs Không gian)

- **Câu chuyện:** Cuộc chiến giữa "uống cà phê" và "đến quán cà phê".
- **Insight:** Highlands định vị mình là nơi có **sản phẩm tốt** (nằm ở bên phải), trong khi các đối thủ như The Coffee House, Starbucks lại định vị mình là nơi có **không gian trải nghiệm tốt** (nằm ở phía trên). Hai bên đang theo đuổi hai chiến lược khác nhau.

2. Bản đồ "Price Value vs Convenience" (Giá trị vs Sự tiện lợi)

- **Câu chuyện:** Cuộc chiến giữa "giá tốt" và "sự nhanh gọn".
- **Insight:** Đây thường là bản đồ cho thấy sự phân hóa rõ rệt. Có thể không có thương hiệu nào nằm ở góc trên bên phải (vừa có giá trị tốt, vừa tiện lợi). Highlands có thể có lợi thế về Convenience do có nhiều cửa hàng, nhưng lại yếu thế về Price Value so với các thương hiệu nhỏ hơn. **Khoảng trống thị trường** ở đây có thể là một mô hình "cà phê chất lượng, giá hợp lý, mua mang đi nhanh chóng".

3. Bản đồ "Socializing vs. Staff & Service" (Giao lưu vs. Dịch vụ)

- **Câu chuyện:** Cuộc chiến về yêu tố con người và không gian xã hội.

- **Insight:** The Coffee House và Cộng Cà Phê có thể mạnh về **Socializing** (không gian phù hợp để gặp gỡ bạn bè). Tuy nhiên, **Staff & Service** rất có thể là một **diễn biến chung của cả thị trường**, không có thương hiệu nào thực sự bứt phá. Đây là một cơ hội vàng.

4. Bản đồ "Functional vs. Brand Image" (Công năng vs. Hình ảnh thương hiệu)

- **Câu chuyện:** Cuộc chiến giữa "tính hữu dụng" và "phong cách sống".
- **Insight:** Đây là bản đồ thể hiện rõ nhất sự khác biệt trong lời hứa thương hiệu.
 - + **Highlands:** Mạnh về Functional (đáp ứng nhu cầu cơ bản như một nơi làm việc, có cà phê đáng tin cậy).
 - + **Starbucks:** Mạnh về Brand Image (bán một phong cách sống, một biểu tượng toàn cầu).
 - + Chưa có thương hiệu nào dung hòa được cả hai yếu tố này một cách hoàn hảo.

Khám Phá Cơ Hội Chiến Lược

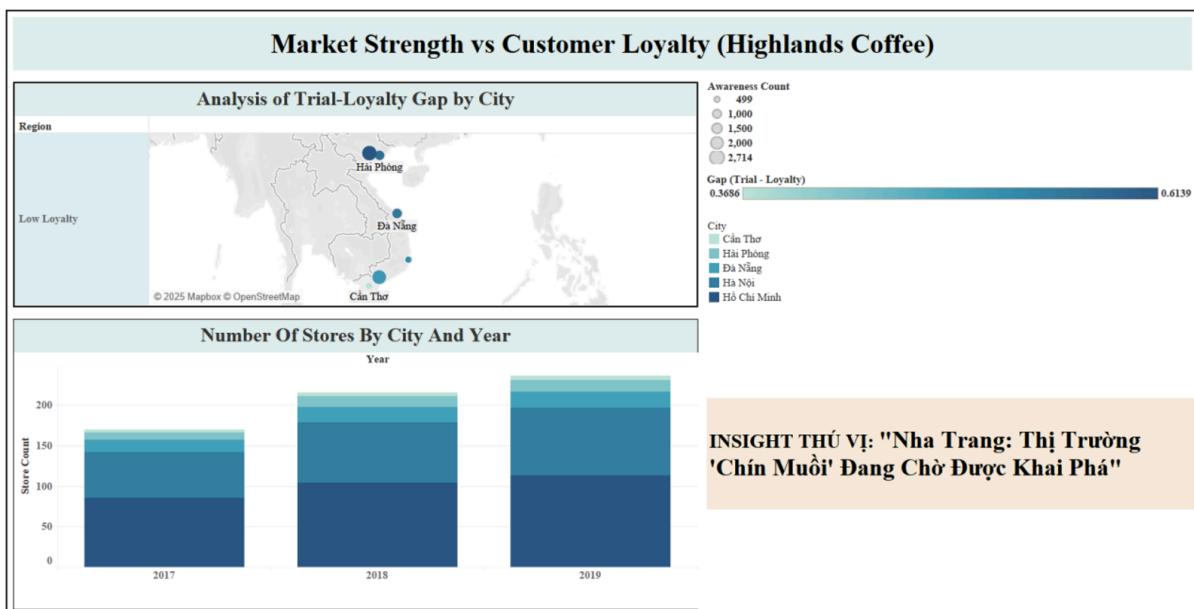
Dựa trên việc phân tích tổng hợp cả 4 bản đồ, đây là những cơ hội chiến lược dành cho Highlands:

- **Tái định vị thương hiệu (Brand Repositioning):**
 - + Highlands đứng trước một lựa chọn chiến lược:
 - **A) "Phòng Thủ":** Trở thành **nhà vô địch không thể tranh cãi** trong lãnh địa của mình. Đây mạnh truyền thông về "Chất lượng sản phẩm vượt trội" và "Không gian làm việc hiệu quả nhất".
 - **B) "Tấn Công":** Đầu tư quyết liệt để cải thiện các thuộc tính về Environment và Brand Image để cạnh tranh trực tiếp với The Coffee House và Starbucks.
- **Tối ưu hóa và Tăng cường (Optimization & Enhancement):**
 - + **Cơ hội từ "Khoảng trống Dịch vụ":** Vì cả thị trường đều yêu ở Staff & Service, Highlands có thể tạo ra một lợi thế cạnh tranh cực lớn bằng cách đầu tư vào các chương trình đào tạo nhân viên quy mô lớn, xây dựng một văn hóa dịch vụ khách hàng xuất sắc. Một khi đã dẫn đầu ở thuộc tính này, sẽ rất khó để đối thủ sao chép.
- **Mở rộng thị trường (Whitespace Opportunities):**

- + **Khoảng trống "Chất lượng với Giá trị tốt":** Có một nhu cầu tiềm ẩn cho một thương hiệu mang lại chất lượng cà phê đáng tin cậy (như Highlands) nhưng với một mức giá được cảm nhận là hợp lý hơn. Việc ra mắt một dòng sản phẩm mới hoặc một thương hiệu con (sub-brand) nhắm vào phân khúc này là một hướng đi đáng cân nhắc.
- + **Khoảng trống "Trải nghiệm Độc đáo":** Thay vì chỉ là một không gian đẹp chung chung, Highlands có thể tạo ra các cửa hàng với chủ đề độc đáo (ví dụ: cửa hàng rang xay tại chỗ, không gian nghệ thuật,...) để tạo ra một lý do ghé thăm khác biệt.

DASHBOARD 4

- **Đánh giá sức mạnh thị trường so với mức độ trung thành:** Ưu tiên các thị trường mà Highland có thị phần lớn nhưng mức độ giữ chân khách hàng thấp, cho thấy mức độ hiện diện không tương ứng với lòng trung thành.



Phân Tích Toàn Diện: Đánh Giá Sức Mạnh Thị Trường vs. Lòng Trung Thành

Tóm tắt:

Dashboard này cho thấy một câu chuyện "hai mặt" của Highlands Coffee: một mặt là sự **tăng trưởng đầy ánh tượng về quy mô hệ thống** qua 3 năm, mặt khác là một **thách thức lớn trong việc chuyển đổi sự hiện diện vật lý đó thành lòng trung thành bền vững của khách hàng**.

- **Thành công về Đầu tư:** Biểu đồ "Number Of Stores" cho thấy một chiến lược mở rộng quyết liệt và thành công, giúp Highlands gia tăng độ phủ và sự hiện diện trên toàn quốc.
- **Thách thức về Hiệu quả:** Biểu đồ bản đồ "Analysis of Trial-Loyalty Gap" lại chỉ ra rằng, tại tất cả các thị trường trọng điểm, **tỷ lệ khách hàng dùng thử luôn vượt xa tỷ lệ khách hàng trung thành** (tất cả các điểm đều có màu xanh). Điều này tạo ra một "phễu rò rỉ", cho thấy chi phí đầu tư mở rộng chưa mang lại hiệu quả tương xứng về mặt giữ chân khách hàng.
- **Kết luận chính:** Highlands đang làm rất tốt việc khiến khách hàng "thử", nhưng chưa đủ tốt trong việc khiến họ "yêu" và ở lại. Ưu tiên chiến lược lúc này cần chuyển dịch từ "**mở rộng**" sang "**đào sâu**", tập trung vào việc tăng cường lòng trung thành tại các thị trường hiện có.

Diễn Giải và Đề Xuất Chiến Lược

1. **Biểu đồ "Number Of Stores By City And Year" (Xu Hướng Đầu Tư)**
 - **Điễn giải:** Biểu đồ cột chồng này là minh chứng rõ ràng nhất cho chiến lược đầu tư và mở rộng mạnh mẽ của Highlands từ 2017-2019. Quy mô hệ thống tăng đều qua từng năm cho thấy sự cam kết và nguồn lực dồi dào trong việc chiếm lĩnh thị phần vật lý. Đây là một điểm mạnh về mặt vận hành và đầu tư.
2. **Biểu đồ "Analysis of Trial-Loyalty Gap by City" (Kết Quả Thực Tế)
 - **Điễn giải:** Đây là biểu đồ "kết quả" của sự đầu tư đó.
 - + **Kích thước vòng tròn (Awareness Count):** Cho thấy việc mở rộng đã giúp tăng mức độ nhận biết. TP. Hồ Chí Minh và Hà Nội là hai thị trường có độ nhận biết lớn nhất, tương xứng với quy mô đầu tư.
 - + **Màu sắc vòng tròn (Gap Trial - Loyalty):** Đây là nơi câu chuyện phức tạp hơn. Màu xanh càng đậm, "phễu" càng "rò rỉ". **Hà Nội và Hải Phòng** là hai nơi có mức chênh lệch lớn nhất, cho thấy đây là những thị trường "khó chiều", thu hút được nhiều người thử nhưng tỷ lệ quay lại thấp nhất.
 - **Sự kết nối giữa hai biểu đồ:** Khi đặt hai biểu đồ cạnh nhau, ta có thể suy luận: Mặc dù số lượng cửa hàng (vốn đầu tư) tăng, nhưng "chất lượng" của tệp khách hàng (lòng trung thành) lại chưa tăng tương xứng. Đây là một dấu hiệu quan trọng cho thấy cần xem xét lại hiệu quả của chiến lược mở rộng.**
3. **Hộp "INSIGHT THÚ VỊ"**

- **Diễn giải:** Đây là một "viên ngọc quý" trong phân tích của bạn. Nó chứng minh rằng **sức mạnh thương hiệu của Highlands đã vượt ra khỏi ranh giới vật lý**. Việc khách hàng ở Nha Trang biết đến và có nhận thức về Highlands dù chưa có cửa hàng là một tín hiệu cực kỳ tích cực, cho thấy tiềm năng lớn và rủi ro thấp hơn khi quyết định gia nhập thị trường này.

Khám Phá Cơ Hội Chiến Lược và Phân Loại Thị Trường

Dựa trên dashboard, chúng ta có thể phân loại các thị trường và đưa ra hành động ưu tiên:

- Ưu tiên 1 - "Vùng Cần Chăm Sóc Khẩn Cấp" (Thị trường lớn, rò rỉ cao):**
 - **Thành phố:** Hà Nội, TP. Hồ Chí Minh.
 - **Đặc điểm:** Độ nhận biết cao (vòng tròn lớn), mức chênh lệch Trial-Loyalty cao (màu đậm).
 - **Chiến lược:** Ngừng tập trung vào việc thu hút khách hàng mới. Dồn toàn lực vào các chương trình **giữ chân khách hàng (retention)**: triển khai loyalty programs theo từng địa phương, khảo sát để tìm hiểu lý do khách hàng không quay lại, cá nhân hóa ưu đãi qua ứng dụng. Cải thiện loyalty ở đây sẽ mang lại hiệu quả tức thì về doanh thu.
- Ưu tiên 2 - "Vùng Tiềm Năng Nuôi Dưỡng" (Thị trường đang phát triển):**
 - **Thành phố:** Đà Nẵng, Cần Thơ, Hải Phòng.
 - **Đặc điểm:** Độ nhận biết vừa phải (vòng tròn nhỏ), mức chênh lệch Trial-Loyalty vẫn tồn tại.
 - **Chiến lược:** Tiếp tục các hoạt động tăng nhận biết, nhưng phải **triển khai song song các chương trình loyalty ngay từ đầu**. Rút kinh nghiệm từ các thị trường lớn để xây dựng một tệp khách hàng trung thành ngay từ giai đoạn đầu.
- Ưu tiên 3 - "Khoảng Trống Chờ Khai Phá" (Whitespace Opportunity):**
 - **Thành phố:** Nha Trang.
 - **Đặc điểm:** Chưa có mặt nhưng đã có nhận biết.
 - **Chiến lược:** Đây là ứng cử viên hàng đầu cho việc mở rộng trong tương lai. Cần đưa vào kế hoạch nghiên cứu thị trường chi tiết để chuẩn bị cho việc gia nhập.

Quá trình làm: Brand Funnel Analysis (Phễu thương hiệu)

Phân tích Phễu Thương hiệu & Sức khỏe Cảnh tranh

- **Phễu marketing 5 giai đoạn (Awareness -> Advocacy)** là một mô hình chiến lược để lên kế hoạch hành động.



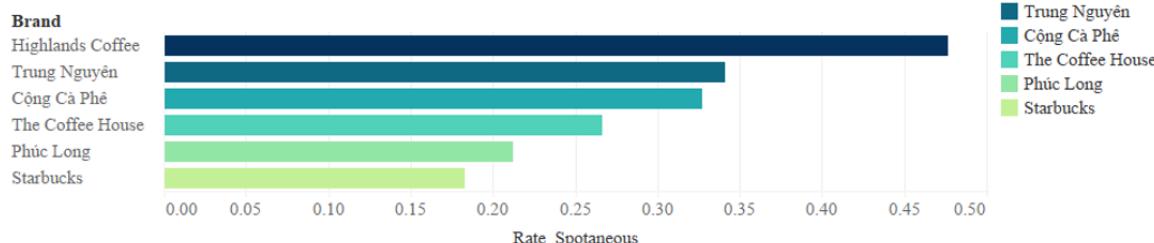
- **Hệ thống đo lường gồm 7 biểu đồ thương hiệu (Brand Funnel)** đánh giá xem các kế hoạch hành động đó có hiệu quả hay không trong việc xây dựng sức mạnh thương hiệu trong tâm trí người tiêu dùng.

Cần làm: Phân tích theo % chuyển đổi giữa các tầng của phễu, so sánh Highland và đối thủ (Phúc Long, The Coffee House, Starbucks...)

Giai đoạn 1: Awareness (Nhận biết)

Biểu đồ tương ứng: Rate_Spontaneous và Rate_Awareness.

Brand Funnel Analysis - Spontaneous



Brand Funnel Analysis - Awareness



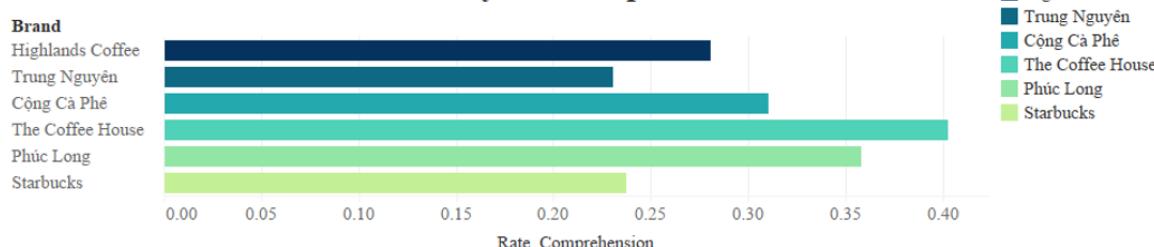
- Phân tích:

- + Các biểu đồ này cho thấy ở tầng trên cùng của phễu, Highlands Coffee đang thống trị tuyệt đối. Họ đã thực hiện xuất sắc các hoạt động marketing để tiếp cận đại chúng (độ phủ cửa hàng rộng, quảng cáo, PR...).
- + Khi một người tiêu dùng nghĩ đến "cà phê chuối", tên Highlands bật ra đầu tiên (Spontaneous cao). Hầu như tất cả mọi người đều biết đến thương hiệu này (Awareness gần 100%).
- + Các đối thủ như The Coffee House, Phúc Long... có "miệng phễu" hẹp hơn nhiều.

Giai đoạn 2: Consideration (Cân nhắc)

Biểu đồ tương ứng: Rate_Comprehension.

Brand Funnel Analysis - Comprehension



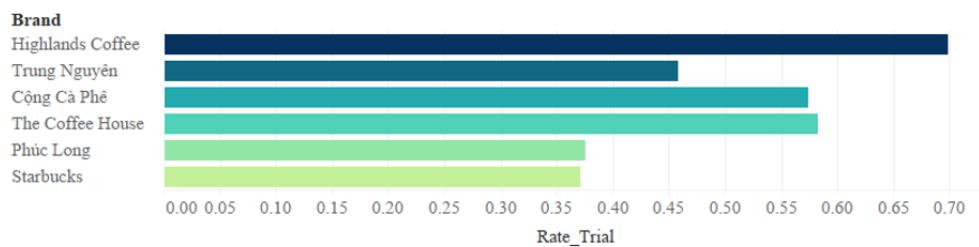
- **Phân tích:**

- + Sau khi biết đến thương hiệu, khách hàng bắt đầu tìm hiểu "Thương hiệu này là gì? Có hợp với mình không?". Đây chính là giai đoạn Thấu hiểu (Comprehension).
- + **Một insight rất thú vị: The Coffee House làm cực kỳ tốt ở giai đoạn này. Mặc dù Awareness thấp hơn, nhưng tỷ lệ người "hiểu" về định vị "ngôi nhà cà phê", không gian làm việc của họ lại rất cao, gần bằng Highlands.** Điều này có nghĩa là các hoạt động marketing của The Coffee House (email, online reviews, organic search...) rất hiệu quả trong việc thuyết phục và xây dựng lòng tin, giúp họ cạnh tranh mạnh mẽ ở giữa phễu.

Giai đoạn 3: Conversion (Chuyển đổi)

Biểu đồ tương ứng: Rate_Trial.

Brand Funnel Analysis - Trial



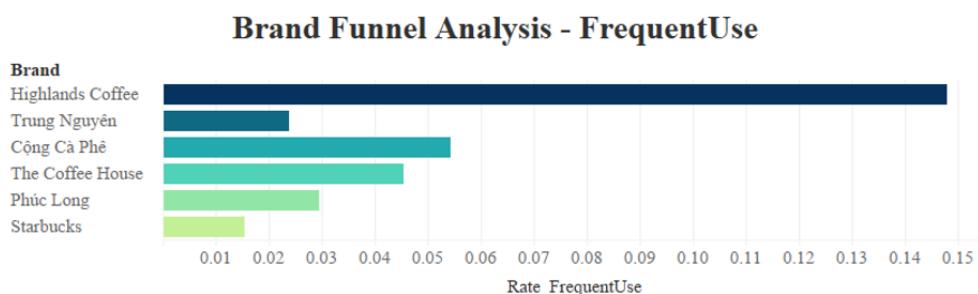
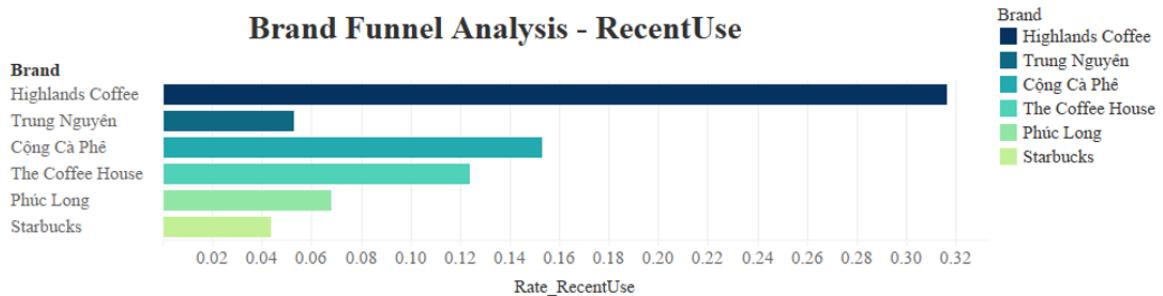
- **Phân tích:**

- + Hành động "Dùng thử" (Trial) chính là hành động chuyển đổi quan trọng đầu tiên trong ngành F&B. Nó biến một người đang cân nhắc thành một khách hàng thực sự.
- + Biểu đồ Rate_Trial cho thấy Highlands tiếp tục dẫn đầu, phần lớn nhờ vào "sức mạnh" từ 2 giai đoạn trên (nhận biết rộng, tiện lợi).

Để đánh giá hiệu quả của Giai đoạn 2 (Cân nhắc), bạn có thể tính tỷ lệ chuyển đổi từ Awareness sang Trial. Thương hiệu nào có tỷ lệ này cao chứng tỏ họ rất giỏi trong việc thuyết phục khách hàng đến thử.

Giai đoạn 4: Loyalty (Trung thành)

Biểu đồ tương ứng: Rate_RecentUse và Rate_FrequentUse.



- Phân tích:

- + Đây là giai đoạn giữ chân khách hàng sau khi họ đã "chuyển đổi" (dùng thử).
- + Rate_RecentUse (Sử dụng gần đây) cho thấy những dấu hiệu giữ chân ban đầu. Khách hàng đã quay lại ít nhất một lần.
- + Rate_FrequentUse (Sử dụng thường xuyên) là thước đo lòng trung thành thực sự. Khách hàng đã biến việc đến quán thành một thói quen.
- + Đây là "pháo đài" vững chắc nhất của Highlands và là "điểm yếu chí mạng" của nhiều đối thủ. Highlands không chỉ giỏi kéo người đến, mà còn rất giỏi giữ người ở lại, chứng tỏ các chiến dịch giữ chân (retention campaigns) của họ đang hoạt động hiệu quả.

Giai đoạn 5: Advocacy (Üng hộ/Lan truyền)

Biểu đồ tương ứng: Rate_NPSGroup

- Trước khi có biểu đồ này, ta cần xử lý về mặt dữ liệu và ý nghĩa.



```
brandhealth['NPSGroup'].value_counts()
```



count

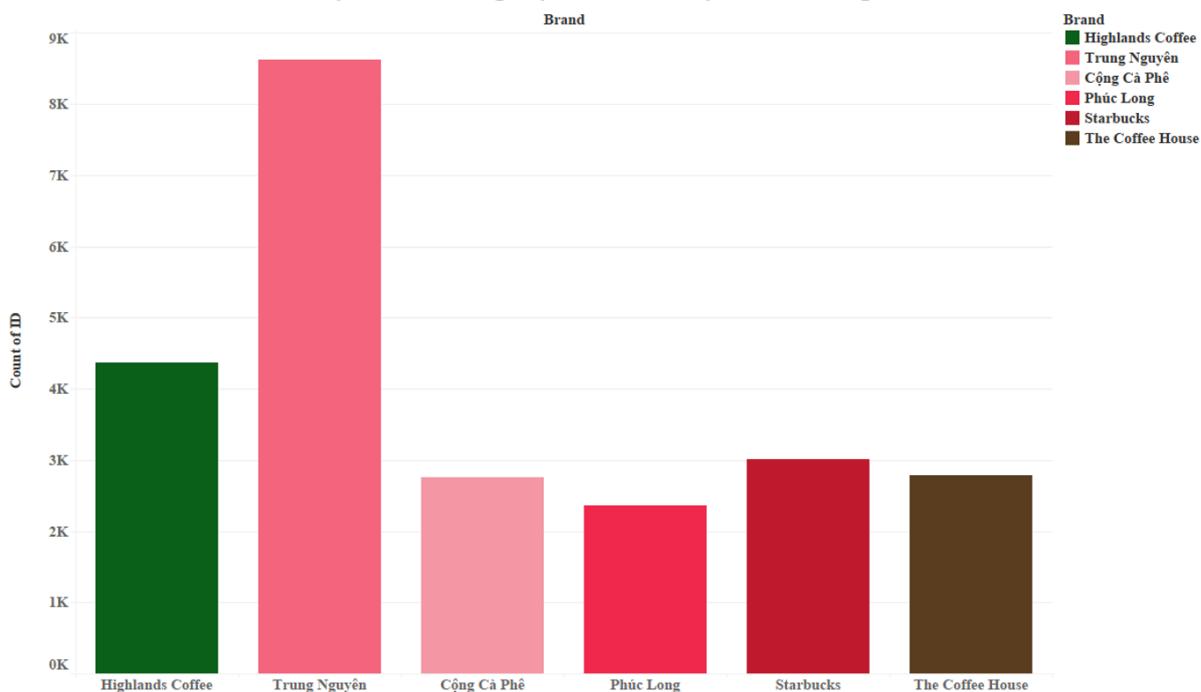
NPSGroup

Unknown	23912
Passive	3994
Promoter	3772
Detractor	591

dtype: int64

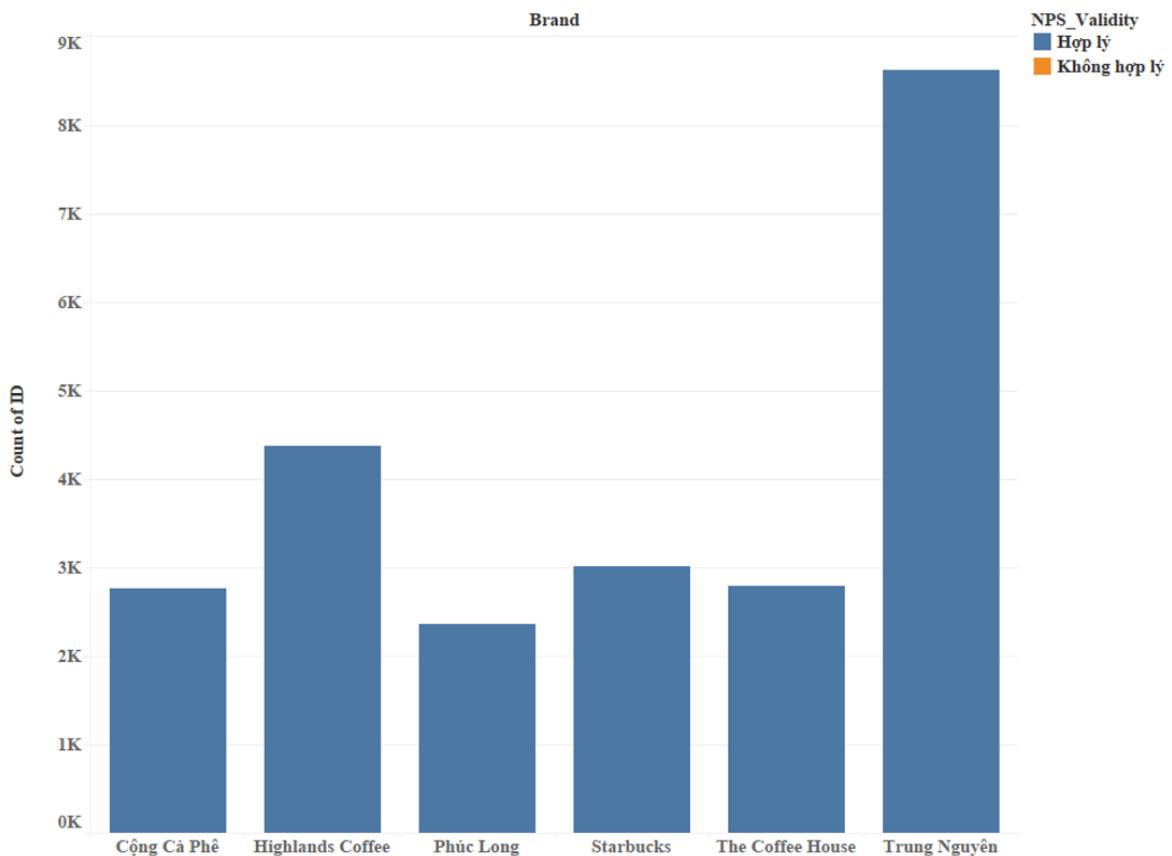
- Có thể thấy bộ dữ liệu tồn tại khá nhiều giá trị Unknown ở cột NPSGroup. Cần tìm hiểu nguyên nhân xem Unknown ở đây là hợp lý hay không hợp lý.

Tần số xuất hiện theo thương hiệu (chưa xác định NPSGroup)



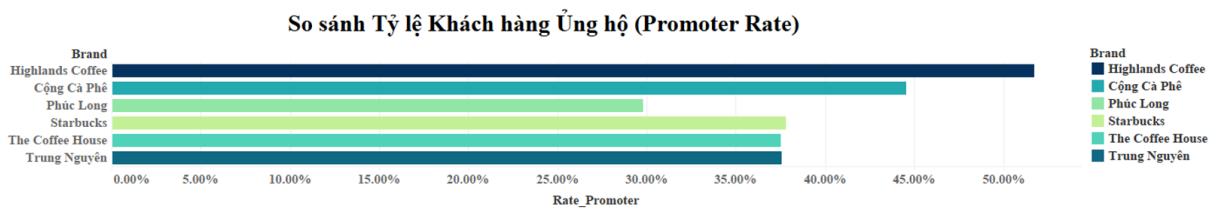
Count of ID for each Brand. Color shows details about Brand. The data is filtered on NPS Group, which keeps Unknown.

Mức độ hợp lý của phản hồi Unknown trong NPSGroup



Count of ID for each Brand. Color shows details about NPS_Validity.

- **Nhận xét: hầu hết các giá trị Unknown trong NPSGroup là hợp lệ.**
- **Tiến hành phân tích đại diện nhóm Promoter trong NPSGroup**



Rate_Promoter for each Brand. Color shows details about Brand.

"Pháo đài" Highlands Coffee - Bất khả xâm phạm nhờ Độ phủ và Lòng trung thành Cốt lõi

- **Thống trị tuyệt đối đầu phễu:** Các biểu đồ Awareness và Trial ban đầu cho thấy Highlands có một lợi thế không lồ về độ nhận diện và mức độ thâm nhập thị trường. Họ là lựa chọn mặc định và dễ tiếp cận nhất.
- **Giữ chân vượt trội:** Biểu đồ Frequent Use xác nhận họ có khả năng biến người dùng thử thành khách hàng thường xuyên tốt nhất.

Insight mới và mạnh nhất: Chất lượng lòng trung thành đỉnh cao: Biểu đồ Promoter Rate vừa tạo ra một đòn khăng định. Với tỷ lệ người ủng hộ trên 50%, Highlands không

chỉ giữ chân khách hàng bằng thói quen hay sự tiện lợi, mà họ đã tạo ra một lượng lớn những người thực sự yêu thích và sẵn sàng quảng bá cho thương hiệu.

Luận điểm chiến lược: Lợi thế cạnh tranh lớn nhất của Highlands không chỉ nằm ở số lượng cửa hàng, mà nằm ở chất lượng của tệp khách hàng trung thành. Họ đã tạo ra một "vòng lặp lan truyền" mạnh mẽ: Khách hàng trung thành trở thành Promoters -> Promoters giới thiệu khách hàng mới chất lượng cao -> Củng cố vị thế dẫn đầu. Đây chính là "con hào kinh tế" (economic moat) bảo vệ họ khỏi các đối thủ.

Các hành động Chiến lược Khả thi cho Highlands Coffee

Dựa trên các phân tích trên, đây là những đề xuất chiến lược cụ thể cho Highlands:

Khai thác "Mỏ vàng" Promoters - Chuyển từ Phòng thủ sang Tấn công bằng "Word-of-Mouth":

- **Vấn đề:** Highlands có hơn 50% khách hàng là Promoters nhưng có thể chưa khai thác hết tiềm năng này.
- **Hành động:**
 - + Triển khai Chương trình Giới thiệu (Referral Program) quy mô lớn: Sử dụng app Highlands Coffee để tạo một cơ chế "Give-Get" hấp dẫn, biến mỗi Promoter thành một "nhân viên sale".
 - + Chiến dịch "Đại sứ Highlands": Tôn vinh và kể câu chuyện của những khách hàng trung thành (UGC - User-Generated Content). Thay vì chỉ nói "chúng tôi tự hào", hãy để khách hàng nói thay bạn.

"Vá" lại Lỗ hổng của Đối thủ - Củng cố Trải nghiệm để Đón đầu:

- **Vấn đề:** The Coffee House cho thấy chỉ cần trải nghiệm không đủ tốt, khách hàng sẽ rời đi dù ban đầu rất ấn tượng.
- **Hành động:** Highlands cần tiếp tục đầu tư vào sự nhất quán và vận hành xuất sắc (operational excellence). Đảm bảo chất lượng sản phẩm, tốc độ phục vụ và sự sạch sẽ luôn được duy trì ở mức cao nhất trên toàn hệ thống. Đây là một chiến lược phòng thủ thông minh để **ngăn chặn khách hàng của mình tìm đến các lựa chọn khác.**

Phân khúc lại Khách hàng để Tăng trưởng Sâu:

- **Vấn đề:** Không phải khách hàng nào cũng giống nhau. Ngày giờ bạn đã có thể phân nhóm khách hàng một cách rõ ràng.
- **Hành động:**
 - + **Với nhóm Passives:** Chạy các chiến dịch nhỏ, có mục tiêu để "thúc đẩy" họ (ví dụ: thông báo về sản phẩm mới, một ưu đãi nhỏ bất ngờ) để chuyển họ thành Promoters.
 - + **Với nhóm Lapsed Users ("Unknown NPS"):** Đây là cơ hội tăng trưởng lớn bị bỏ quên. Hãy thiết kế một chiến dịch "Chào mừng quay trở lại" với ưu đãi thực sự hấp dẫn để kéo họ về lại với thương hiệu.

CÂU HỎI 5:

Phân tích Tỷ lệ Khách hàng Rời bỏ và Dashboard dành cho Highlands Coffee

MỤC TIÊU

Phân tích nhóm khách hàng rời bỏ đối với Highlands Coffee bằng cách xác định các yếu tố chính liên quan đến nhận diện thương hiệu, phân khúc, hành vi sử dụng, đồng hành, nhu cầu sử dụng, và nhân khẩu học thông qua trực quan hóa dữ liệu nâng cao và diễn giải chuyên sâu.

Đề xuất **chiến lược hành động** cho Highlands Coffee nhằm **giảm tỷ lệ khách hàng rời bỏ**. **Nhóm khách hàng rời bỏ:** Nhóm khách hàng được xác định rời bỏ thương hiệu Highlands Coffee là nhóm có dữ liệu về việc ghé cửa hàng Highlands Coffee trong vòng 3 tháng trước nhưng không có thông tin trong vòng 1 tháng trở lại đây.

Biểu đồ thể hiện Churn Rate theo từng tiêu chí:

1. Brand Perception:



Trong tiêu chí Brand Perception (Nhận diện thương hiệu), churn rate được biểu diễn thông qua Comprehension và NPS.

Biểu đồ Comprehension cho thấy **nhóm khách hàng rời bỏ thương hiệu tập trung chủ yếu ở nhóm “có hiểu biết ít” hoặc “hầu như không biết”** về thương hiệu cafe này, với tỉ lệ lần lượt là 65% và 77.78%. Nhóm này thường là khách hàng chưa trải nghiệm hoặc có ân tượng ban đầu mờ nhạt về Highlands.

Tiếp đến biểu đồ NPS, điều này càng được minh chứng rõ hơn qua tỉ lệ nhóm detractor của Churn. Ở đây detractor có thể hiểu là những khách hàng không đánh giá cao Highlands, và đánh giá với mức điểm thấp (từ 0 - 6) trên thang điểm 10. **Nhóm detractor trong Churn rate chiếm tỉ lệ là 72.09%, cao hơn hẳn so với 2 nhóm còn lại.**



=> **Những khách hàng có ấn tượng ban đầu mờ hồ hoặc không tốt, họ không có đủ kết nối hoặc cảm xúc để giữ chân họ tiếp tục tin dùng Highlands và khiến họ ở lại.**

Và trải nghiệm tiêu cực từ ban đầu cũng là nguyên nhân khiến họ đánh giá thương hiệu Highlands ở mức điểm thấp và rời bỏ đi.

Tóm lại, thiếu nhận biết và cảm xúc tích cực không thể giữ chân khách hàng sử dụng Highlands Coffee.

Vậy lí do vì sao dù cho Highlands có độ phổ biến và nổi tiếng nhất định, có một bộ phận khách hàng vẫn không biết hoặc đánh giá thấp về thương hiệu này?

2. Brand Image



Trong nhóm khách hàng rời bỏ, họ đánh giá Highlands với 3 tiêu chí cao nhất bao gồm: Brand Equity (4.53%), Product Quality (4.53%) và Service & People (3.92%).

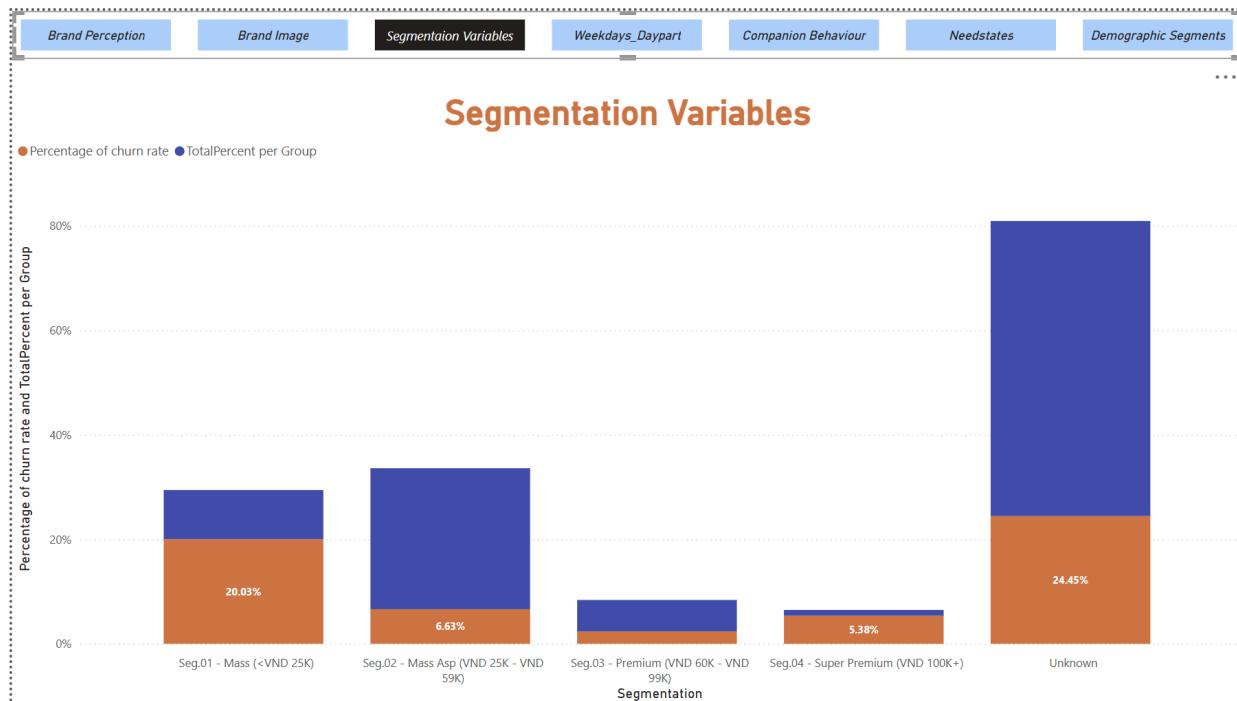
Như đã phân tích ở trên, mặc dù Highlands có độ nhận diện hiện đang đứng đầu trong lĩnh vực dịch vụ này, các thương hiệu cafe khác cũng đang dần đầy mạnh nhận diện thương hiệu, có thể kể đến Trung Nguyên. **Khoảng cách Brand Equity giữa Highlands và Trung Nguyên đang rút ngắn lại**, báo động về mức độ hiệu quả của các chiến dịch marketing, sự chững lại trong phát triển và mở rộng của chuỗi thương hiệu này, và đồng nghĩa là **khách hàng đang dần có nhiều sự lựa chọn đáp ứng được các tiêu chí mà Highlands có thể đem lại cho họ hơn**.

Bên cạnh đó, tỷ lệ Churn rate đánh giá về chất lượng sản phẩm tương đồng với Brand Equity. Nguyên nhân có thể xuất phát từ **kì vọng về sự ổn định về sản phẩm của khách hàng không được đáp ứng**, một số khách trung thành bỏ đi do chất lượng không đồng nhất (lúc ngon, lúc dở) hoặc cảm nhận chưa đổi mới trong menu. Họ mong muốn Highlands nâng tầm chất lượng để đáp ứng tính trung thành và trải nghiệm đặc biệt hơn.

3.92% khách hàng bỏ đi vì Service & People, dù cho Highlands định hình là thương hiệu hiện đại, mô hình coworking cafe với dịch vụ được đào tạo chuyên nghiệp, tuy nhiên khía cạnh này vẫn **chưa đủ xuất sắc để khiến khách hàng ấn tượng và lựa chọn tiêu dùng**.

=> Các khía cạnh của Highlands cũng đồng thời là nguyên nhân khiến cho ẩn tượng về thương hiệu trong lòng khách hàng thấp đi hoặc trở nên mơ hồ => khách hàng rời bỏ thương hiệu.

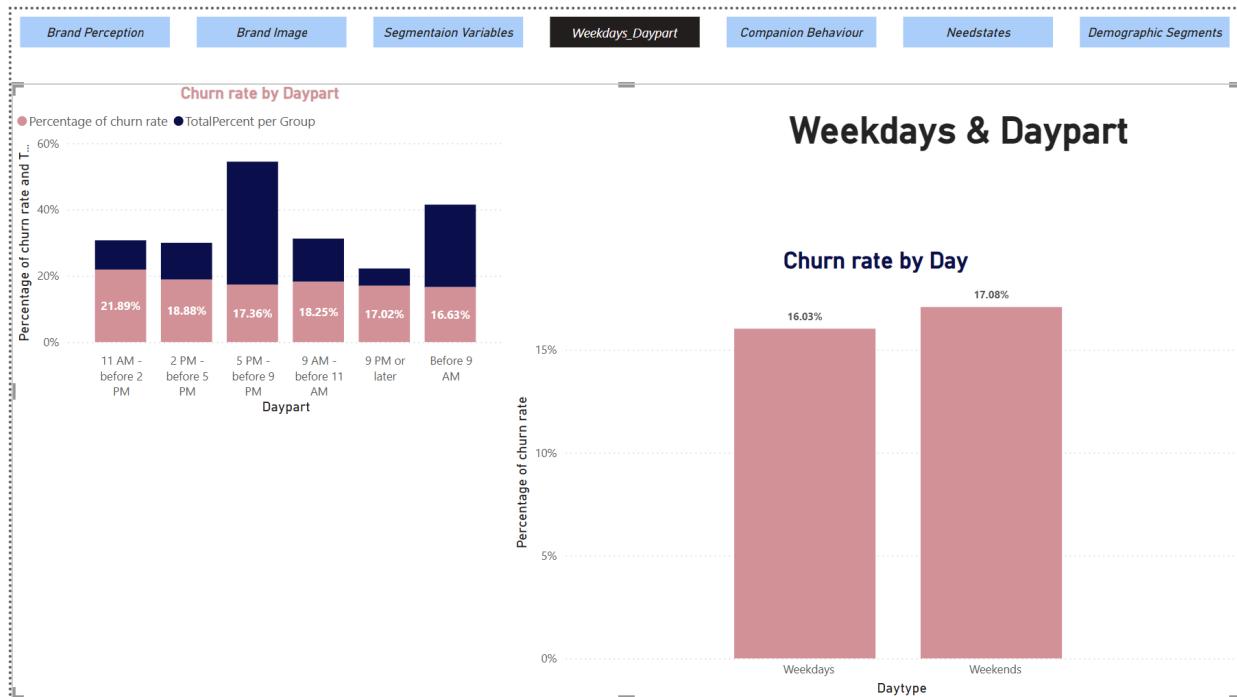
3. Segmentation Variables



Nhóm phân khúc khách hàng đầu tiên (Mass (< VND 25K)) có tỷ lệ rời bỏ cao vì nhận thấy **giá sản phẩm cao hơn tính giá trị mang lại**, có thể nhóm này là nhóm khách hàng có thu nhập vừa đủ nên việc bỏ ra một chi phí cao đồng thời cũng sẽ kèm theo đó là sự kì vọng và đánh giá gắt gao. Họ so sánh Highlands với đối thủ rẻ hơn như Milano, Cheese Coffee. Ngược lại, nhóm Premium đánh giá cao trải nghiệm và sẵn sàng trả giá cao nếu chất lượng đồng nhất.

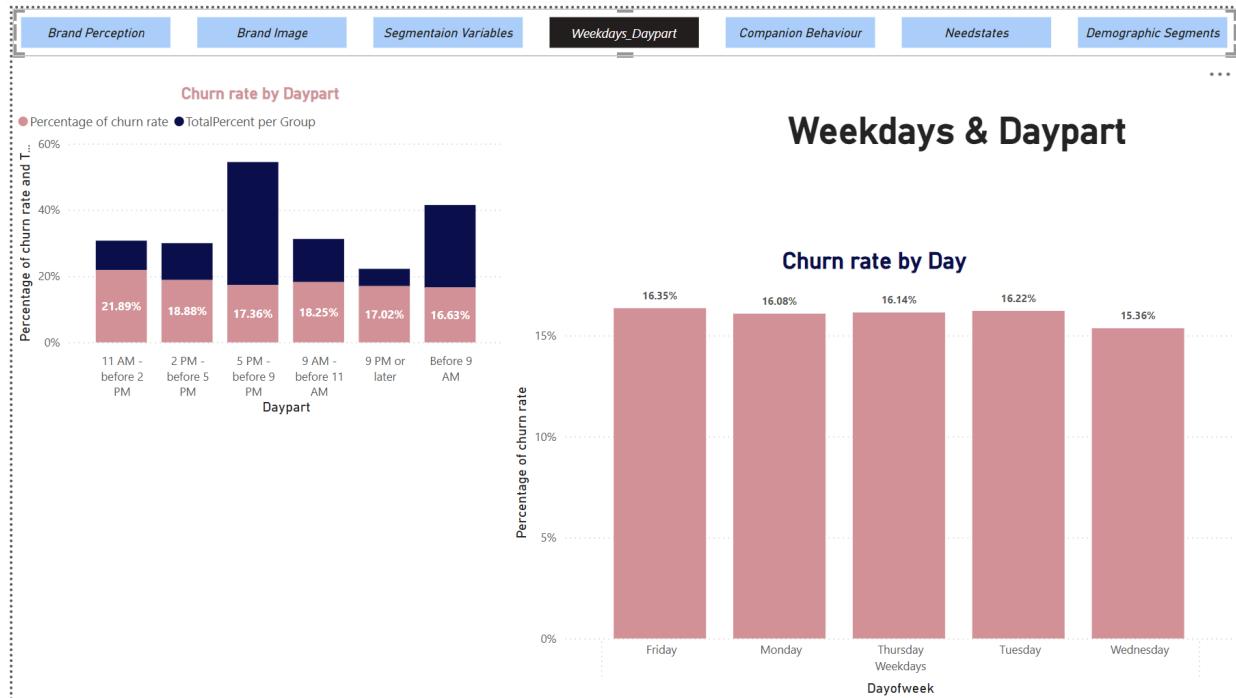
20.03% tỷ lệ Churn ở phân khúc thứ nhất có thể hiểu là khách hàng chi cho lần trải nghiệm đầu tiên và không hài lòng nên rời bỏ và 24.45% tỷ lệ Churn ở nhóm Unknown có thể bao gồm khách bất kì khách hàng nào trong nhóm còn lại, nhưng trong quá trình trải nghiệm, họ gặp thất vọng và không được đáp ứng nhu cầu nên rời bỏ Highlands.

4. Weekdays và Daypart



Ở tiêu chí về thời gian sử dụng Highlands của Churn (theo Weekdays, Dayofweek và Daypart), **Churn có xu hướng sử dụng Highlands vào khung giờ từ 11 AM - before 2 PM nhiều nhất (21.89%)**. Điều này xuất phát từ việc **nhóm khách hàng sử dụng Highlands gồm học sinh sinh viên, dân văn phòng là chủ yếu**, và khung giờ này vào trưa là thời điểm nghỉ ngơi của các công ty, trường học. Họ sử dụng Highlands như một nơi để thư giãn hoặc gặp mặt hoặc học nhóm. Do đó, việc **quá tải nhân viên, khách hàng phải đợi quá lâu, không được chăm sóc kỹ lưỡng hoặc đồng đúc, ồn ào** cũng là nguyên nhân khiến khách hàng sẽ lựa chọn một quán cafe khác để sử dụng thay vì phải xếp hàng, chờ được phục vụ => **Khách hàng rời bỏ Highlands**.

Tỉ lệ khách hàng sử dụng Highlands vào ngày cuối tuần hay ngày trong tuần là như nhau (17.08% và 16.03%). Một sự nhỉnh hơn trong hành vi sử dụng quán cafe vào ngày cuối tuần cho thấy **Churn có thể ĐÃ kỳ vọng được trải nghiệm và phục vụ nhiều hơn nhưng ĐÃ không được đáp ứng**.

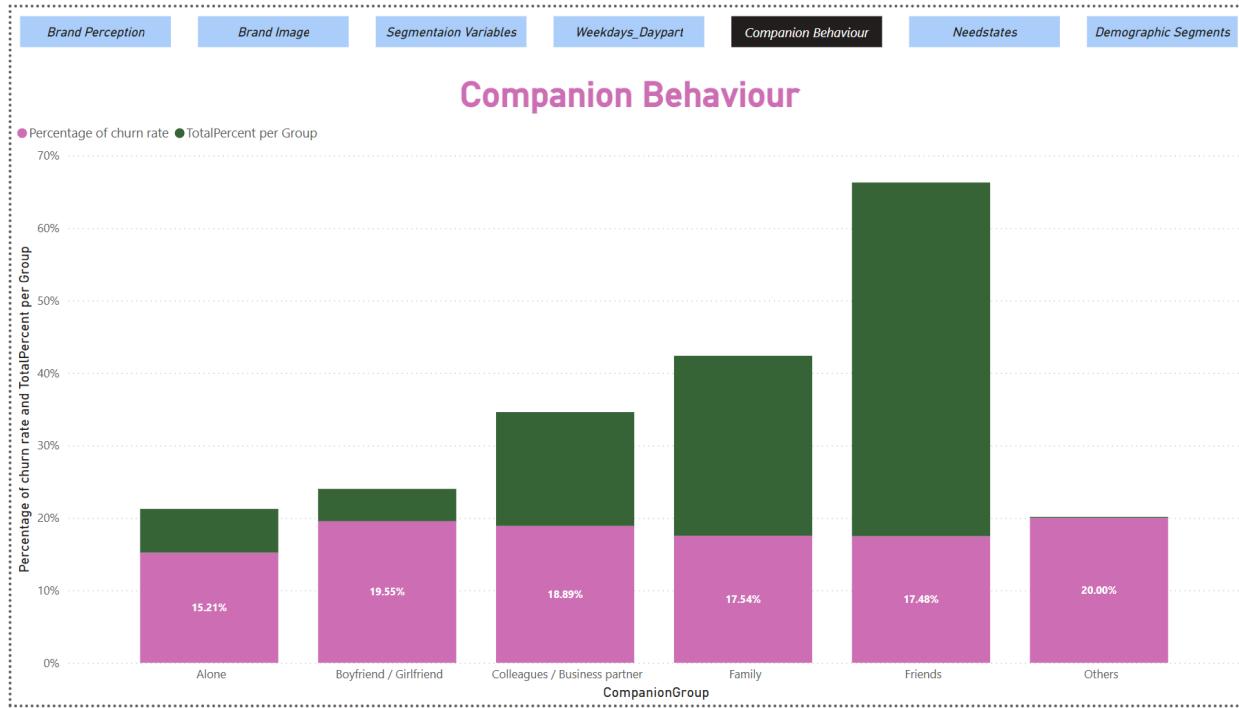


Nhìn chung, tỉ lệ Churn ở các ngày khác nhau trong tuần là như nhau, dao động từ 15% - 17%.

5. Companion Behaviour

Người đi cùng người yêu, bạn bè, hoặc nhóm “Others” có tỷ lệ churn cao hơn so với người đi một mình.

Nhóm đi với bạn bè chiếm tới gần 50% tổng số lượt đi cùng, nên dù churn ở mức trung bình (17.48%) nhưng ảnh hưởng rất lớn đến churn toàn thương hiệu. Có thể thấy nhóm **khách hàng đi theo nhóm thường nhạy cảm hơn với trải nghiệm chung: không gian, âm thanh, dịch vụ nhanh hay chậm**. Họ có xu hướng so sánh và lựa chọn giữa các thương hiệu với nhau. **Nếu không hài lòng, họ sẽ trở thành nhóm khách hàng “Detractor” và nhận diện thương hiệu của Highlands sẽ bị đánh giá ở mức điểm thấp**, và đây cũng là nguyên nhân khiến họ rời đi (như đã đề cập ở mục Brand Perception và Brand Image).



6. Needstates

Needstates cung cấp thông tin về tỷ lệ Churn thông qua các nhu cầu mục đích sử dụng quán cafe của họ, được thể hiện ở biểu đồ sau:



Trong đó, **tỷ lệ Churn ở nhu cầu Working & Business meeting là cao nhất, chiếm 11.09%**. Highlands Coffee được định hình là một thương hiệu hiện đại tuy nhiên vẫn giữ nét truyền thống, insight của Highlands là:

“Khách hàng đến Highlands không chỉ để uống cà phê, mà để tìm thấy một không gian quen thuộc, hiện đại, dễ kết nối – nơi họ có thể thư giãn, trò chuyện, làm việc, và cảm nhận bản sắc Việt trong nhịp sống đô thị.”

Do vậy, rất nhiều khách hàng tìm đến Highlands với nhiều mục đích khác nhau. Điều đó cũng dẫn đến việc không gian trở nên ồn ào, và dần không còn phù hợp với một số khách hàng. Điều đó được thể hiện ở tỷ lệ Churn có nhu cầu Working & Business meeting, họ cần sự yên tĩnh để tập trung làm việc, **họ cần một không gian chuyên nghiệp đáp ứng được điều đó nhưng Highlands không còn đáp ứng được** nên họ chuyển sang những sự lựa chọn khác để đáp ứng được kỳ vọng của bạn thân.

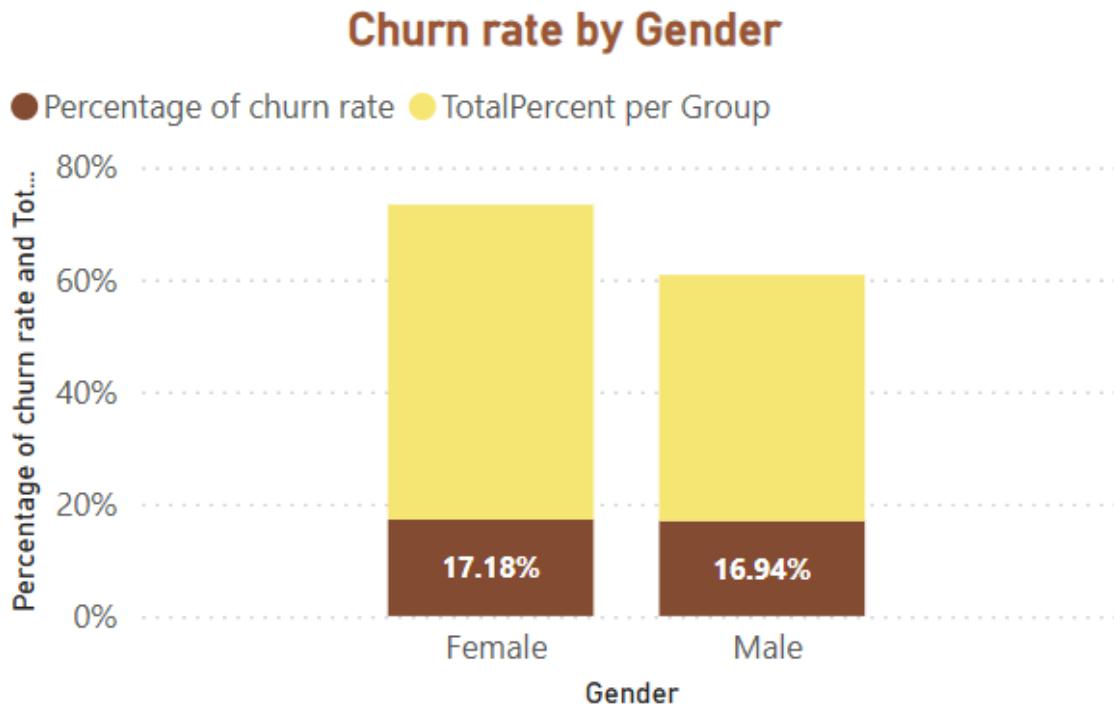
=> **Đây cũng là nguyên nhân khiến họ trở thành nhóm Churn.**

Tương tự đó, tiêu chí Studying & Others cũng diễn ra tương tự. Một tiêu chí khác có tỷ lệ **cũng khá cao là Meals & Snack (9.81%)**. Như bài báo cáo đã phân tích ở mục Brand Image, tiêu chí Product Quality cũng là một trong những nguyên nhân chính khiến khách hàng rời bỏ, bởi **chưa có sự đa dạng trong thực đơn, sản phẩm**. Cũng vì vậy, mặc dù nhu cầu Meals & Snack khá cao nhưng Highlands không đáp ứng được kỳ vọng đó

=> **khách hàng sẽ lựa chọn những thương hiệu khác có khả năng hơn.**



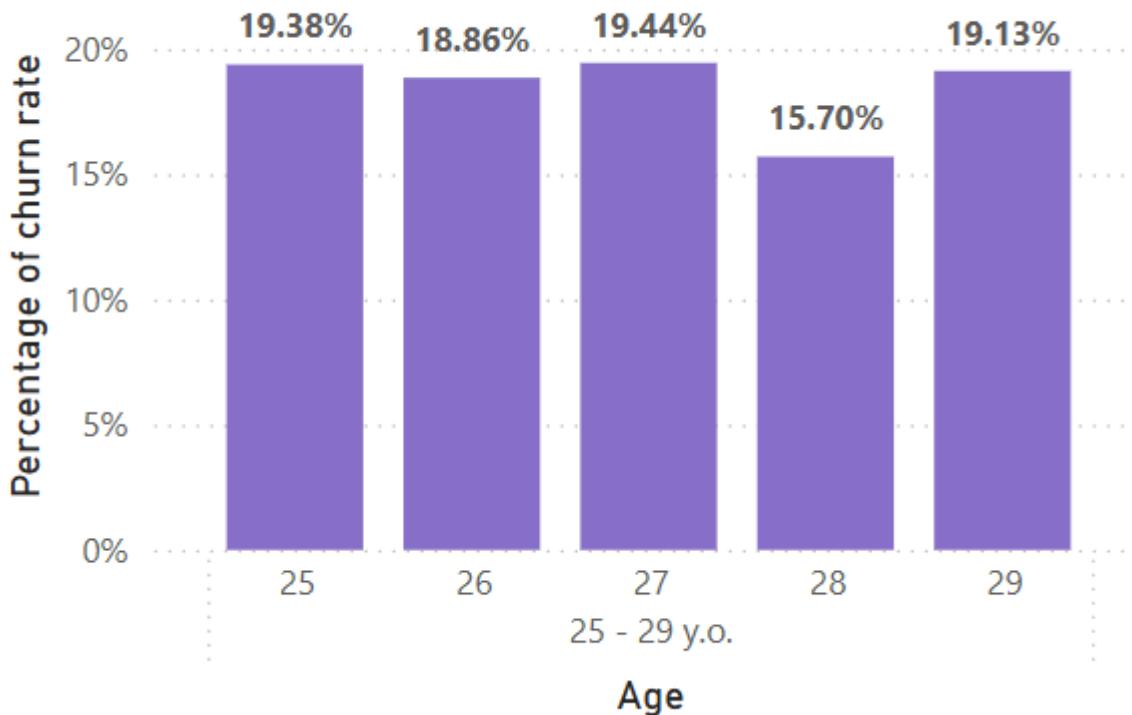
7. Demographic Segments



Nữ giới có tổng phần trăm trong nhóm khách hàng rời bỏ cao hơn \Rightarrow Highlands có nhiều khách hàng nữ hơn.

Nguyên nhân có thể xuất phát từ: Các chiến dịch marketing, sản phẩm hoặc không gian có thể thu hút nữ giới hơn. Tuy nhiên, cả hai giới đều đang rời bỏ với tỷ lệ đáng kể (17.18% và 16.94%), cần chiến lược giữ chân chung.

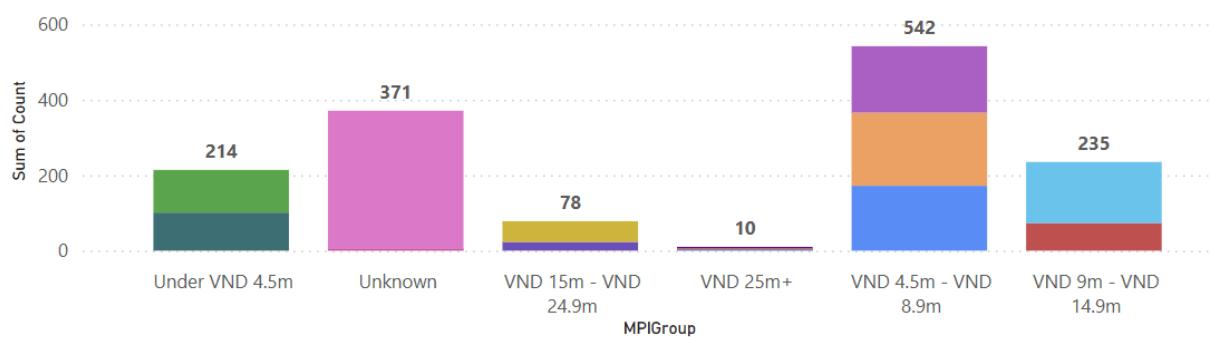
Churn rate by Age



Từ biểu đồ, tuổi 28 có churn rate thấp nhất (15.7%), thấp hơn đáng kể so với các độ tuổi khác (~19%).

Các độ tuổi còn lại khá đồng đều, dao động quanh 19%. Bởi, nhóm tuổi 28 có thể đang trong giai đoạn ổn định công việc, chi tiêu hợp lý, ưu tiên thói quen định kỳ (sử dụng cafe hàng ngày để giữ tinh táo trong công việc). Bên cạnh đó, nhóm tuổi 25–27 hoặc 29 có thể có nhiều lựa chọn thay thế hơn hoặc ít trung thành hơn với thương hiệu.

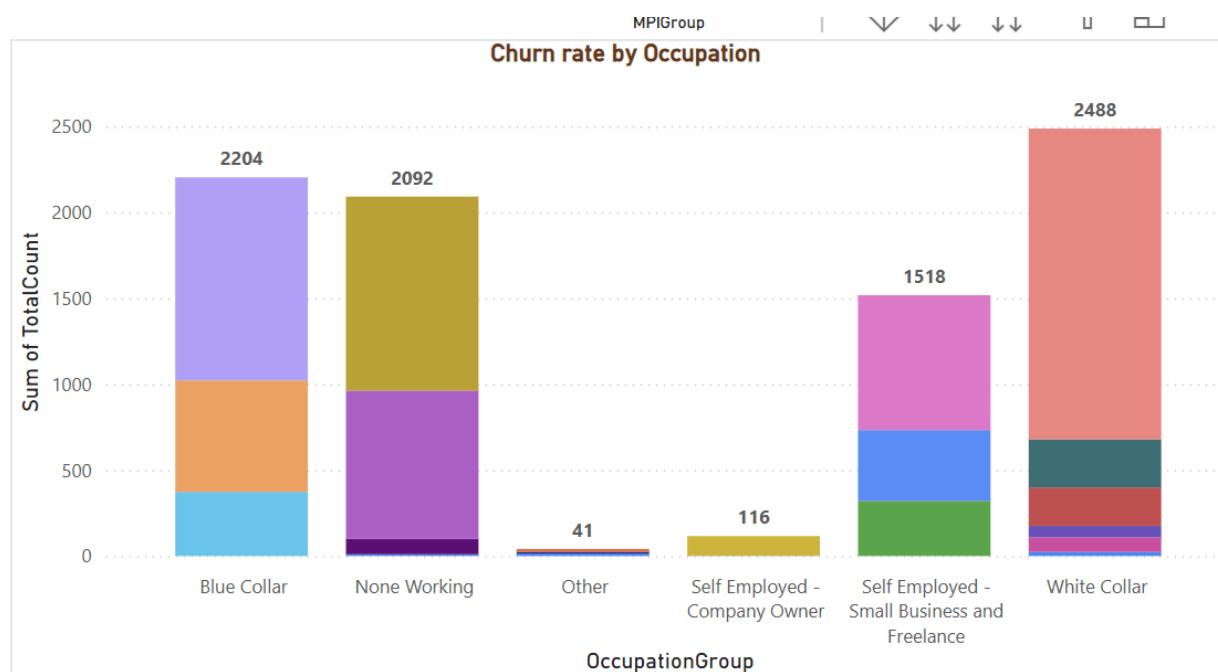
Churn rate by MPI



Có thể thấy, nhóm thu nhập trung bình thấp (4.5 – 8.9 triệu) có số lượng cao nhất (542). Nhóm thu nhập cao (15 triệu trở lên) có tỷ lệ rất thấp. Số lượng không rõ thu nhập (Unknown) cũng chiếm tỷ trọng đáng kể (371).

Như đã phân tích ở trên, phân khúc khách hàng thứ 1 (Mass VND <25K) có thể nằm trong **nhóm thu nhập trung bình thấp** nên họ sẽ rất ngại ngần khi bỏ một chi phí lớn cho cafe. Vì vậy, **giá cả** của Highlands có thể chưa phù hợp với nhóm thu nhập trung bình thấp, hoặc không đạt được sự kỳ vọng trong tâm lý của họ, họ cảm thấy không “đáng tiền” cho cốc cafe của thương hiệu này nên quyết định rời bỏ.

Ngược lại, nhóm thu nhập cao ít churn có thể vì họ ít bị ảnh hưởng bởi giá, họ sẵn sàng bỏ tiền để có trải nghiệm tốt hơn.



Từ biểu đồ “Churn rate by Occupation”, **Nhóm White Collar** (văn phòng) có tỷ lệ khách hàng rời bỏ cao nhất với 2488 khách hàng. Điều đó cho thấy đây là đối tượng **dễ thay đổi thói quen tiêu dùng, có thể do ảnh hưởng từ môi trường làm việc linh hoạt hoặc văn hóa công ty thay đổi**. Theo sau đó là nhóm **Blue Collar (2204 khách hàng)** và **None Working (2092 khách hàng)**, cho thấy khả năng rời bỏ cao có thể đến từ **độ nhạy cảm về giá hoặc điều kiện tiếp cận cửa hàng**.

Ngược lại, nhóm Chủ doanh nghiệp (Company Owner) chỉ có 116 khách, ở đây có nghĩa là mức độ trung thành của chủ doanh nghiệp cao hơn, nguyên nhân có thể xuất phát từ sự ổn định tài chính, họ không ngần ngại chi tiền cho nhu cầu này, và không bận tâm nhiều tới nó, còn có thể xuất phát từ sự ổn định trong thói quen hành vi, họ không thay đổi hành vi sử dụng thương hiệu cafe Highlands từ trước tới nay, họ không chạy theo phong trào mà ưu tiên sự ổn định.

Các chiến lược đề xuất:

1. Tái định vị thương hiệu và cải thiện trải nghiệm nhận diện:

Tăng độ nhận biết thương hiệu bằng các chiến dịch truyền thông nhắm đến nhóm khách hàng mới hoặc ít tương tác, tập trung truyền tải giá trị cốt lõi của Highlands và bản sắc thương hiệu một cách rõ ràng, dễ nhớ.

Giảm tỷ lệ detractor bằng việc tạo dựng các trải nghiệm đầu tiên tích cực như chương trình ưu đãi dùng thử, giảm giá cho lần ghé quán đầu tiên, chăm sóc khách hàng sau lần trải nghiệm.

2. Chăm sóc khách hàng phân khúc thu nhập trung bình – thấp

Phát triển các combo tiết kiệm, chương trình giảm giá theo giờ, chương trình tích điểm đổi quà được ưu tiên áp dụng ở các chi nhánh.

3. Tăng trải nghiệm cho nhóm White Collar & Working segment

Cải thiện không gian yên tĩnh, chuyên biệt dành cho làm việc nhóm, học tập – đặc biệt vào khung giờ cao điểm trưa (11AM – 2PM). Ngoài ra có thể cung cấp dịch vụ ưu tiên hoặc đặt trước chỗ ngồi cho nhóm khách văn phòng, tích hợp qua app.

Triển khai gói membership cho công ty (doanh nghiệp nhỏ & vừa) để giữ chân khách hàng theo tổ chức, hoặc phiếu voucher giảm giá cho nhân viên công ty đối tác.

4. Nâng cấp chất lượng & menu sản phẩm

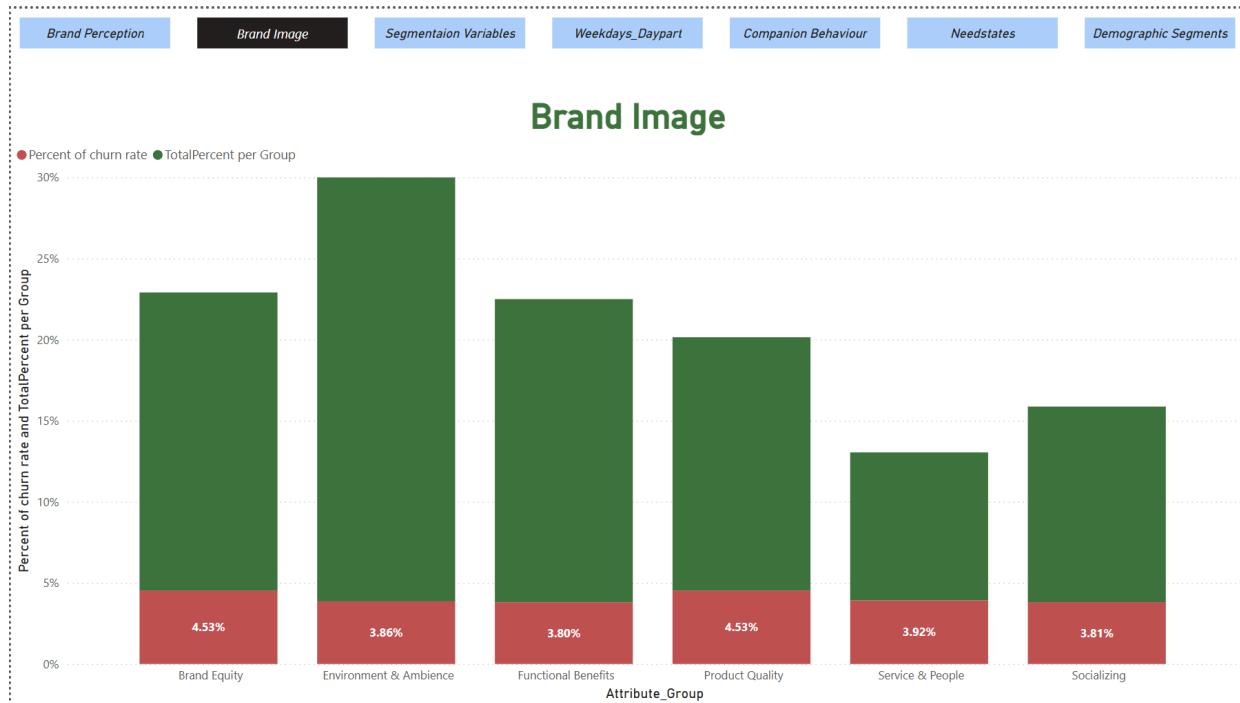
Kiểm soát chặt tính đồng nhất sản phẩm giữa các chi nhánh – đảm bảo trải nghiệm giống nhau dù khách hàng ghé bất cứ đâu. Quan trọng hơn cả là làm mới thực đơn định kỳ, đặc biệt ở nhóm snack/meals, để đáp ứng nhu cầu đa dạng hơn từ khách hàng, có thể cân nhắc cập nhật xu hướng hiện nay của giới trẻ để thu hút tầng lớp trẻ, bởi họ là đối tượng chính có thể tạo nên hiệu ứng đám đông hiệu quả nhất.

DASHBOARD

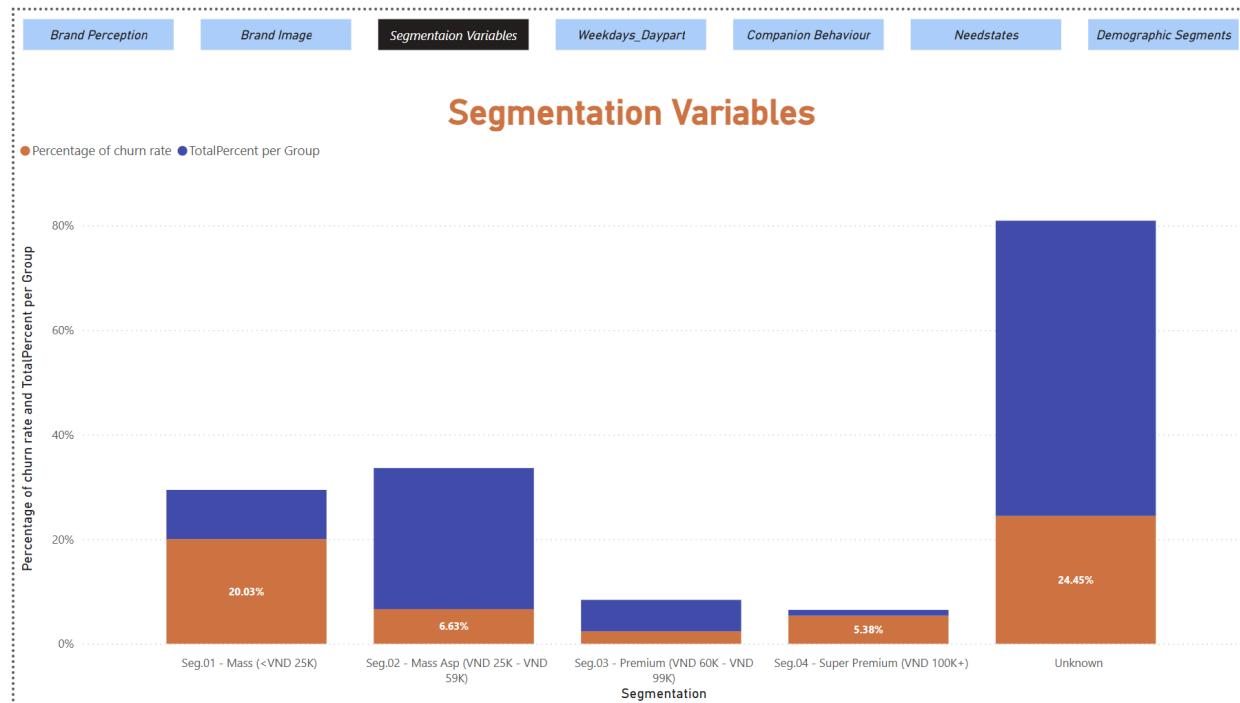
Brand Perception



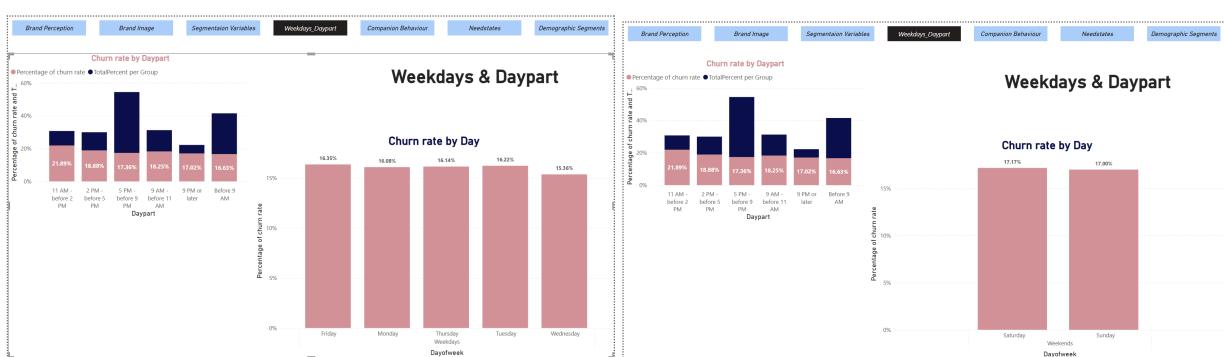
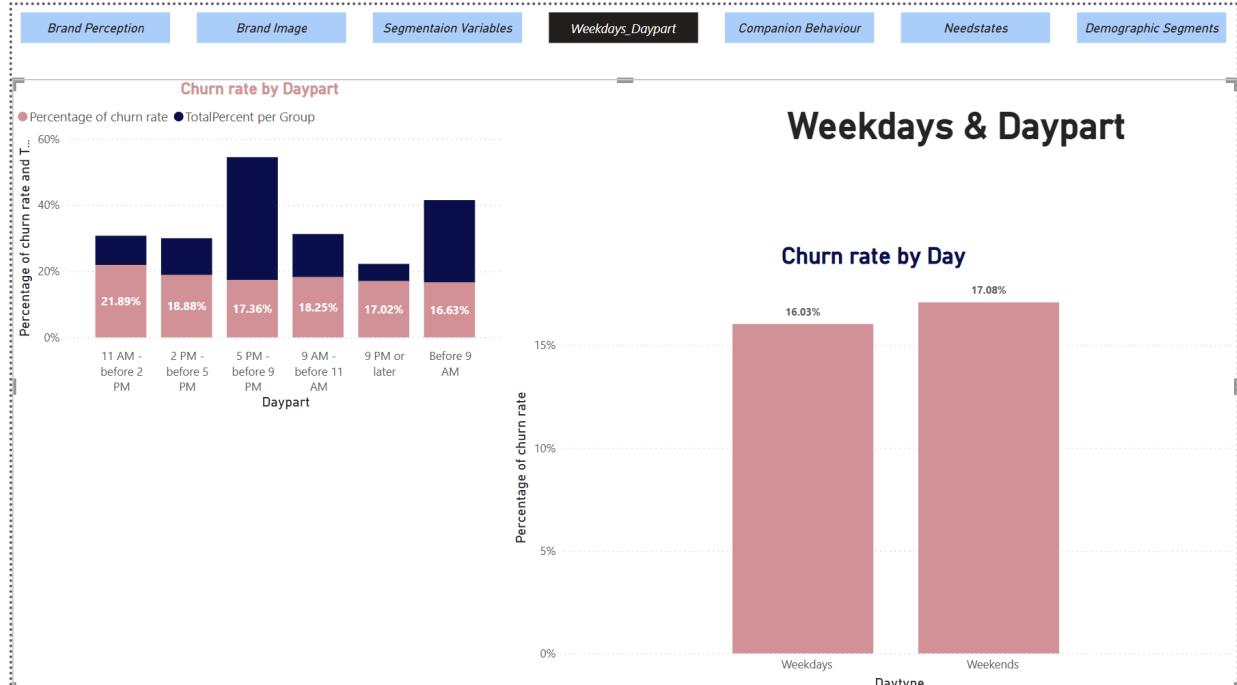
Brand Image



Segmentation Variables



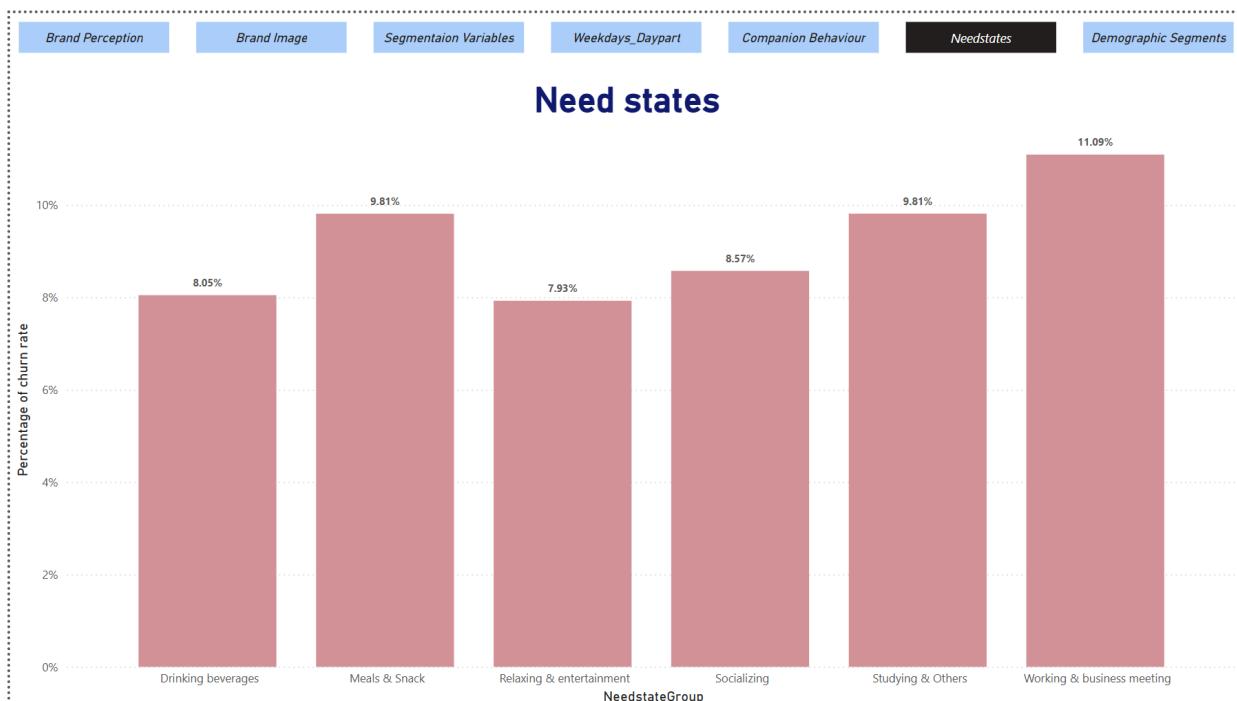
Weekday_Dayparts

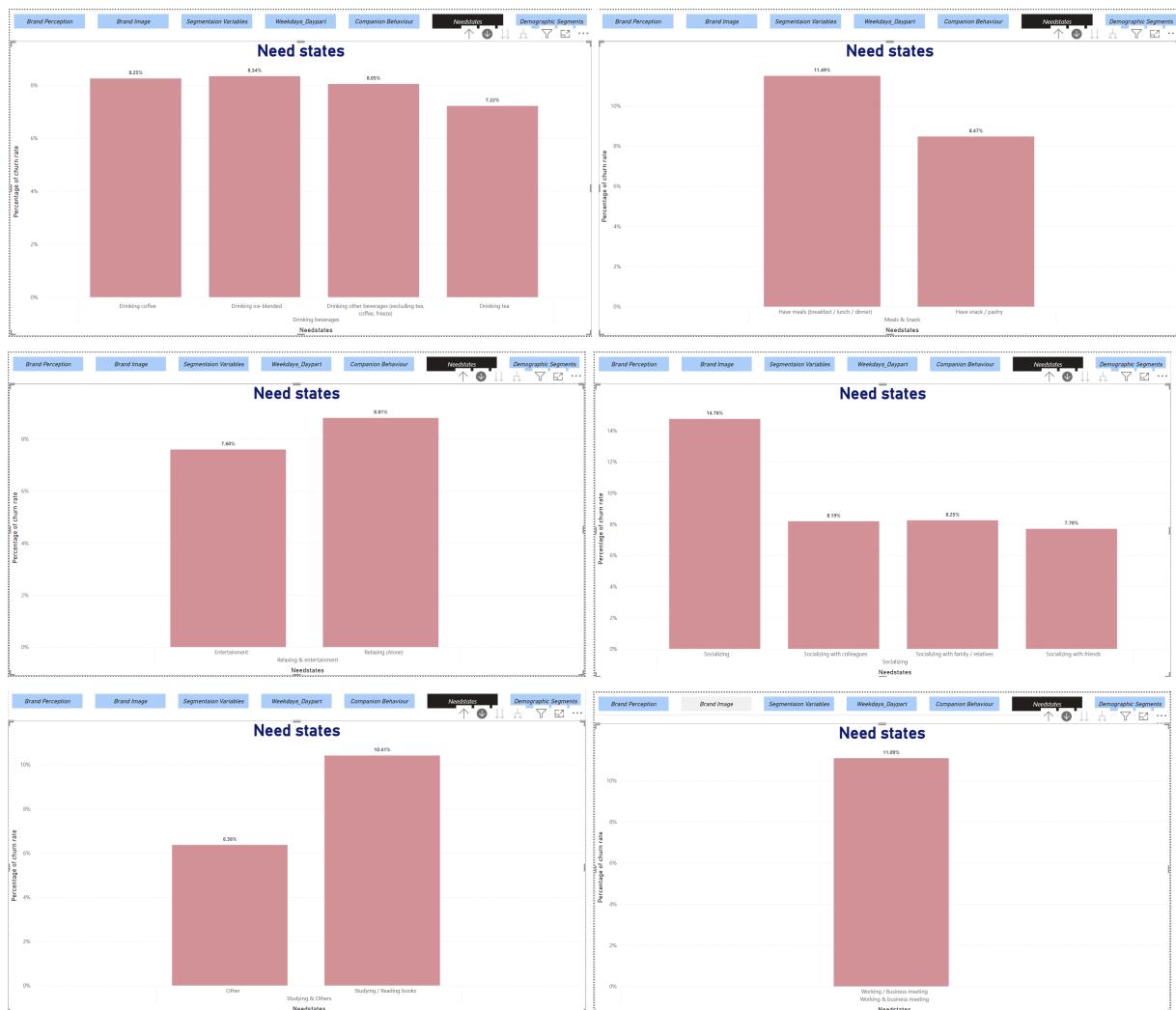


Companion Behaviour



Need States

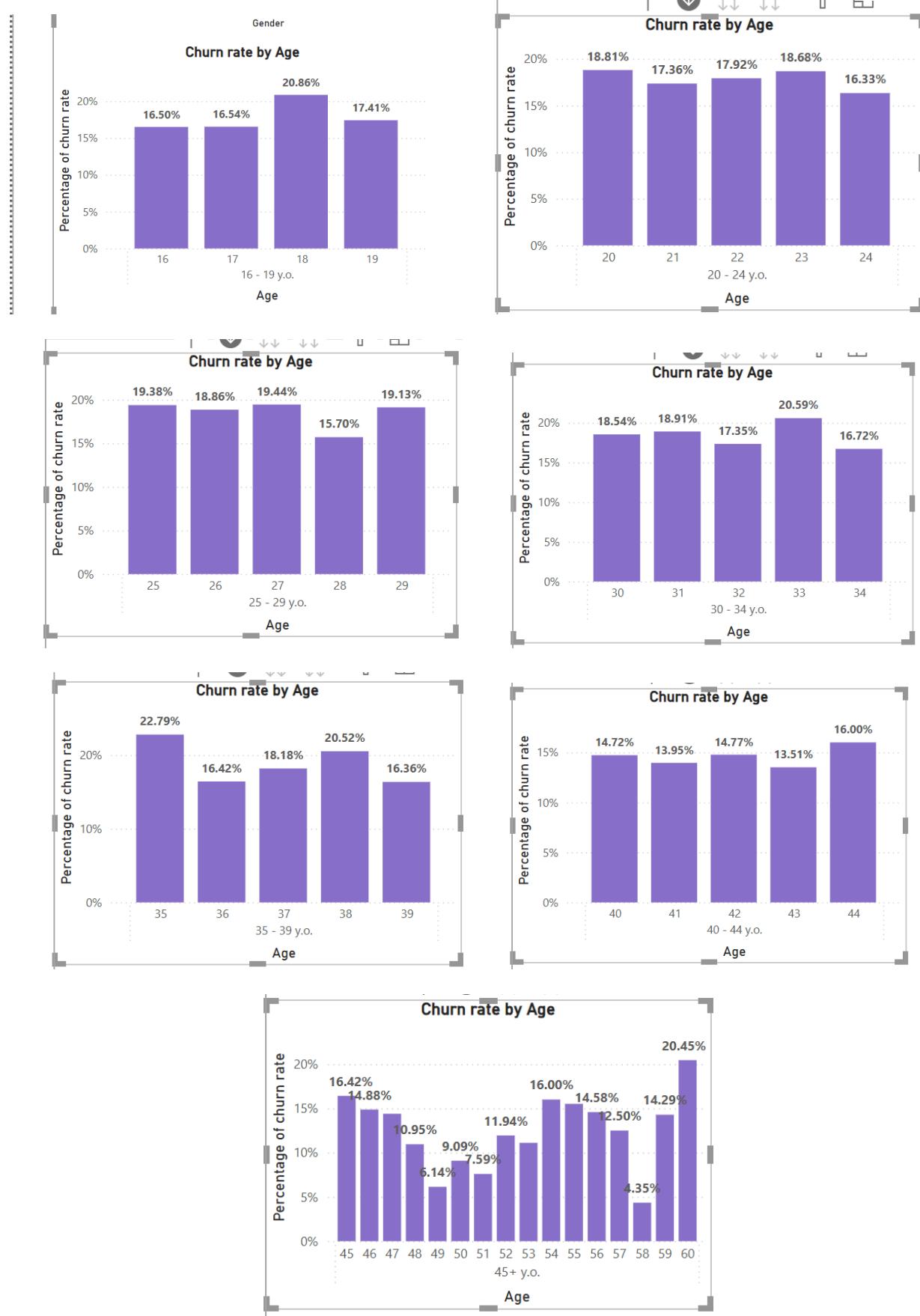




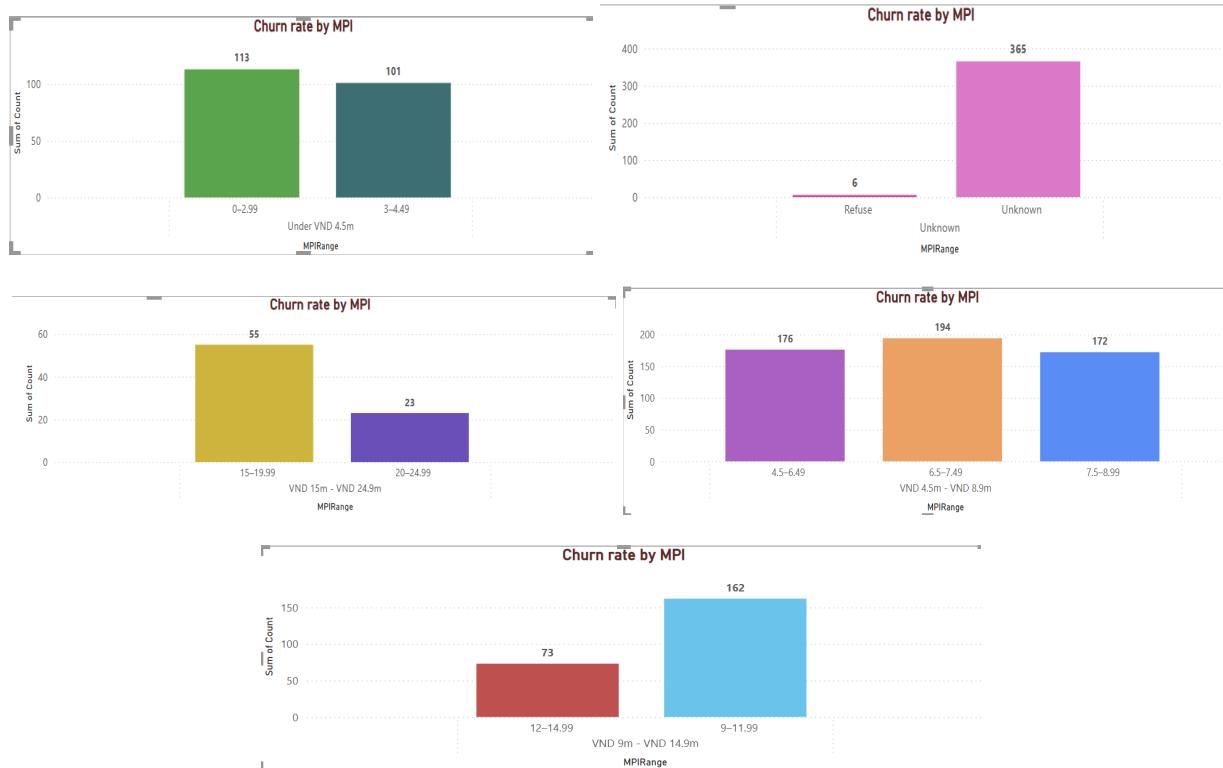
Demographic Segments



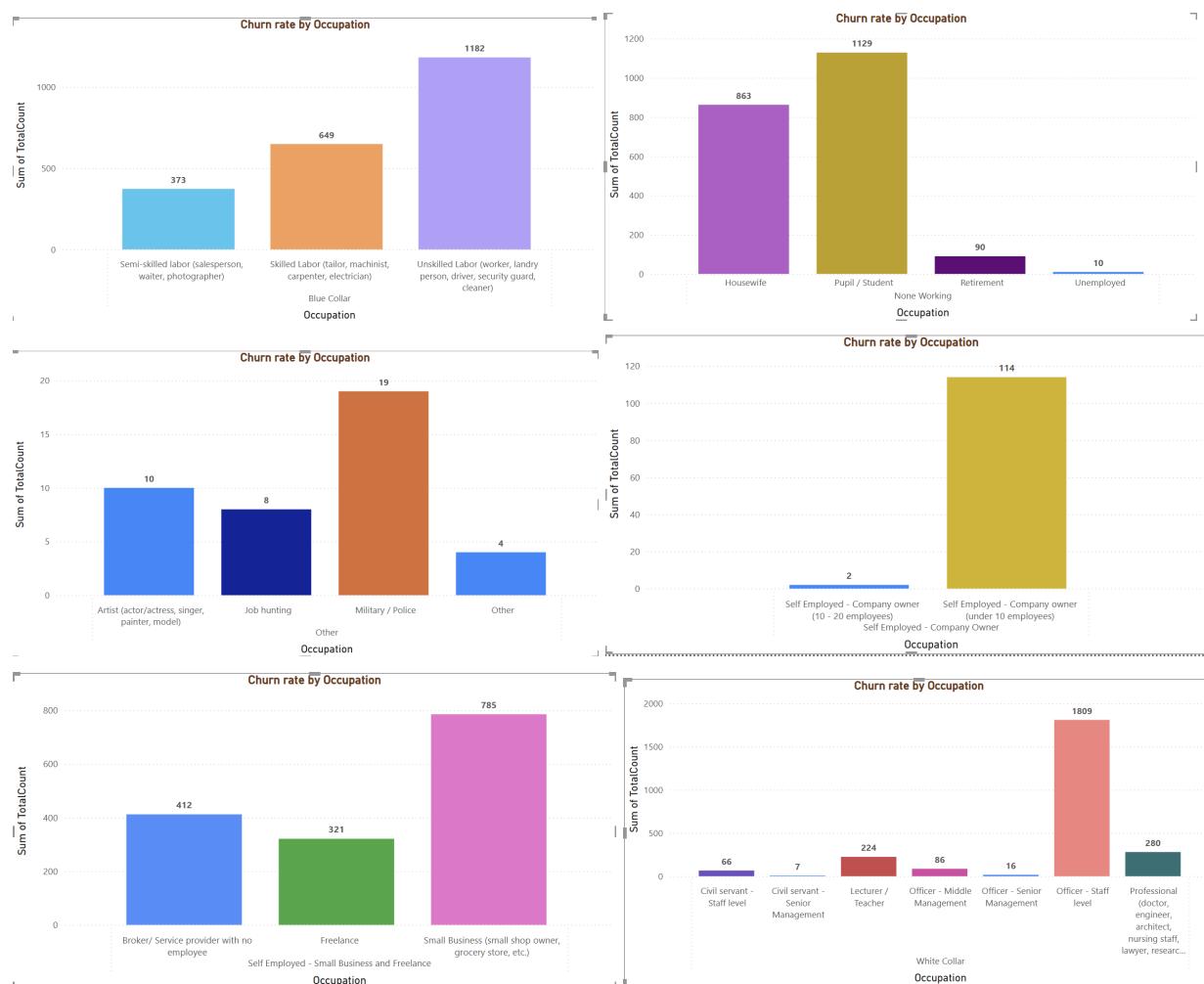
Age



MPI



Occupation



CÂU HỎI 6:

Mục tiêu: Phát triển một phương pháp dựa trên dữ liệu kết hợp phân khúc khách hàng và dự đoán rời bỏ nhằm xác định các khách hàng có nguy cơ cao. Mục tiêu là giúp Highlands Coffee triển khai các chiến lược giữ chân chủ động, theo từng phân khúc, nhằm tối đa hóa giá trị kinh doanh.

Phương pháp tiếp cận

Quy trình thực hiện được chia thành bốn giai đoạn chính như sau:

- **Giai đoạn 1: Xử lý dữ liệu**
- **Giai đoạn 2: Xây dựng mô hình phân cụm khách hàng**
- **Giai đoạn 3: Xây dựng mô hình dự đoán rời bỏ**
- **Giai đoạn 4: Đè xuất cải tiến mô hình**

Giai đoạn 1: Xử lý dữ liệu

Dựa trên các tập dữ liệu đã được tiền xử lý từ câu hỏi 1, bước đầu tiên là lựa chọn và xử lý thêm các đặc trưng phù hợp, nhằm chuẩn bị dữ liệu đầu vào cho mô hình phân cụm.

Cụ thể, nhóm tiến hành xác định các đối tượng khảo sát liên quan đến Highlands Coffee bằng cách truy xuất các ID từ tập **BrandHealth**, với điều kiện **Brand = "Highlands Coffee"**. Các đặc trưng được chọn lọc tiếp theo sẽ đóng vai trò là biến đầu vào cho mô hình phân cụm khách hàng trong giai đoạn tiếp theo.

Features	Bảng	Nguyên nhân chọn
CompanionGroup	Companion	Nhóm đồng hành có thể ảnh

		hưởng đến hành vi
NeedStateGroup	Needstate	Trạng thái nhu cầu ảnh hưởng đến mục đích người dùng đến cửa hàng, không dùng Needstate vì có quá nhiều giá trị có thể khiến phân cụm bị nhiễu, kéo lệch
Dayofweek và VisitOnDayofweek	Dayofweek	Số lượt ghé thăm trong một ngày cụ thể có thể quyết định người dùng có phải thuộc nhóm khách hàng trung thành không
Daypart và VisitOnDaypart	Daypart	Tương tự Dayofweek
AgeGroup, Gender, OccupationGroup, MPIRange	SA#var	Các thông tin về người dùng để phân cụm, lưu ý chỉ dùng Group để tránh có quá nhiều giá trị khác biệt
Comprehension, NPSPer3Month, NPSGroup	Brandhealth	Các thông tin giúp người dùng nhận diện và đánh giá thương hiệu, ảnh hưởng đến hành vi
PPA, VisitFrequency, Spending, Segmentation	BrandSegmentation (tách từ bảng Brandhealth)	Thông tin giúp đánh giá phân khúc khách hàng

Từ bảng features ở trên, ta tiến hành kết hợp các bảng liên quan để tạo thành các bảng thuộc tính. Các bảng được kết hợp dựa trên ID.

Bảng	Features
companion_features	CompanionGroup
needstate_features	NeedStateGroup

day_features	Dayofweek và VisitOnDayofweek, Daypart và VisitOnDaypart
demographic_features	AgeGroup, Gender, OccupationGroup, MPIRange
perceptual_features	Comprehension, NPSPer3Month, NPSGroup
spending_features	PPA, VisitFrequency, Spending, Segmentation
behavioural_features	spending_features + day_features

Tiếp theo, dựa trên các bảng dữ liệu đã được tạo mới, tiến hành mã hóa (encoding) các biến phân loại để chuẩn bị cho quá trình huấn luyện mô hình. Với các biến định lượng (số liên tục), không cần thực hiện bước mã hóa.

Nguyên tắc mã hóa được áp dụng như sau:

- **Các biến phân loại không có thứ bậc** (nominal) sẽ được xử lý bằng phương pháp **One-Hot Encoding**.
- **Các biến phân loại có thứ bậc** (ordinal) sẽ được xử lý bằng phương pháp **Ordinal Encoding**, đảm bảo giữ lại ý nghĩa thứ tự trong dữ liệu.

Sau bước mã hóa, toàn bộ tập dữ liệu được chuẩn hóa (standardize) bằng công cụ **StandardScaler** từ thư viện **scikit-learn** của Python, nhằm đưa các đặc trưng về cùng một thang đo, đảm bảo mô hình hoạt động hiệu quả và tránh thiên lệch do đơn vị đo lường khác nhau.

Encoding	Features
----------	----------

One-hot	CompanionGroup, NeedStateGroup, Dayofweek và VisitOnDayofweek, Daypart và VisitOnDaypart, Gender, OccupationGroup
Ordinal	AgeGroup, MPIRange, Comprehension, NPSGroup, Segmentation

Trong quá trình này sẽ phát hiện được một số vấn đề liên quan và cách khắc phục.

Vấn đề	Cách xử lý
Dữ liệu Unknown hoặc Refuse	Dữ liệu không được người dùng cung cấp không có giá trị, loại bỏ khỏi tập dữ liệu
Dữ liệu trống hoặc trùng lặp	Loại bỏ
Dữ liệu xuất hiện ở các ID khác nhau	Có một số ID xuất hiện nhiều lần khiến một số thuộc tính đa trị xuất hiện, xử lý bằng cách lấy tổng, hoặc trung bình các giá trị theo ID này nếu các cột có vấn đề đều chứa dữ liệu số. Ngược lại, nếu các cột có vấn đề chứa dữ liệu phân loại thì đếm số lần ID xuất hiện hoặc True/False

Giai đoạn 2: Xây dựng mô hình phân cụm

Do số lượng thuộc tính khảo sát khá lớn, quá trình xây dựng mô hình phân cụm sẽ được thực hiện theo từng bước, lặp lại nhiều lần với các nhóm thuộc tính khác nhau. Mục đích là để đánh giá và lựa chọn ra các đặc trưng đầu vào quan trọng, có khả năng phân biệt rõ ràng giữa các nhóm khách hàng.

Phần mềm chủ yếu được sử dụng cho toàn bộ quy trình xử lý dữ liệu và xây dựng mô hình là **Orange**, một công cụ trực quan hỗ trợ tốt cho phân tích dữ liệu. Trong giai đoạn này, hai thuật toán phân cụm được lựa chọn là:

- **K-Means**: phương pháp phân cụm truyền thống, đơn giản, dễ triển khai và trực quan hóa.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: phù hợp với dữ liệu không phân bố đều, có khả năng phát hiện nhiễu và các cụm có hình dạng bất kỳ.

Ngoài các mô hình K-Means và DBSCAN, nhóm cũng đã cân nhắc đến việc sử dụng thuật toán **Hierarchical Clustering** trong quá trình phân cụm. Tuy nhiên, do tập dữ liệu có kích thước tương đối lớn và bao gồm nhiều thuộc tính cần được phân tích đồng thời, thuật toán phân cụm phân cấp tỏ ra **không phù hợp về mặt hiệu suất và khả năng mở rộng**.

Hierarchical Clustering thường yêu cầu tính toán ma trận khoảng cách giữa tất cả các cặp điểm dữ liệu, dẫn đến độ phức tạp tính toán cao khi kích thước dữ liệu tăng lên. Bên cạnh đó, thuật toán này cũng thiếu tính linh hoạt trong việc xử lý nhiễu và không thích hợp khi dữ liệu có cấu trúc phức tạp hoặc không đồng đều.

Vì những lý do đó, nhóm quyết định **không áp dụng Hierarchical Clustering** trong phạm vi dự án này để đảm bảo hiệu quả và tính khả thi của mô hình.

Mô hình 1: Phân cụm với nhóm đặc trưng đầu tiên liên quan đến đặc trưng khách hàng

Các đặc trưng được cân nhắc ở bước này là các đặc trưng liên quan đến giới tính, tuổi, nghề nghiệp, thu nhập, nhu cầu của khách hàng và người đồng hành với họ.

Ở bước đầu tiên, nhóm tiến hành thử nghiệm phân cụm với bốn đặc trưng đầu vào: **AgeGroup**, **NeedStateGroup**, **Gender**, và **CompanionGroup**. Mô hình được lựa chọn là **DBSCAN**, nhờ khả năng nhận diện các cụm khách hàng không đồng nhất mà không cần chỉ định trước số lượng cụm. Mô hình này được gọi là mô hình 1.1.

Sau khi áp dụng mô hình, kết quả phân cụm được trực quan hóa bằng thuật toán **t-SNE**, giúp dễ dàng quan sát cấu trúc phân cụm trong không gian hai chiều:

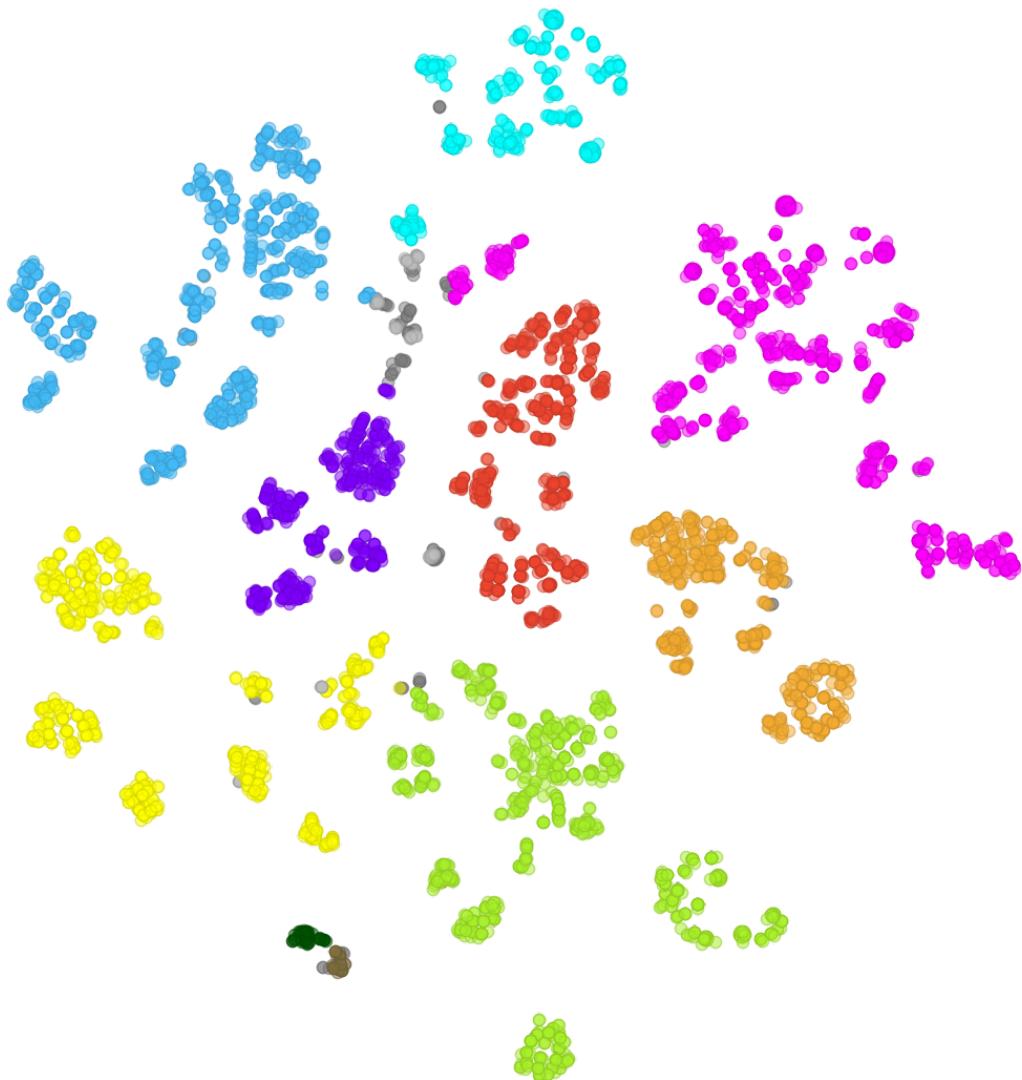


Mặc dù hai cụm khách hàng nổi bật có thể được quan sát rõ ràng trên biểu đồ trực quan, tuy nhiên vẫn tồn tại **nhiều cụm nhỏ chồng lấn nhau**, xuất hiện với **tần suất thấp**, và **khó nhận diện rõ ràng bằng mắt thường**. Điều này cho thấy dữ liệu có thể chứa nhiều hoặc các nhóm khách hàng với hành vi chưa được phân biệt rõ ràng qua các đặc trưng đã chọn.

Ở lần thử nghiệm tiếp theo, nhóm tiến hành phân cụm với các đặc trưng đầu vào sau:

MPIRange, OccupationGroup, Gender, NeedStateGroup, và AgeGroup. Mô hình này từ đây gọi là 1.2

Kết quả phân cụm tiếp tục được trực quan hóa bằng thuật toán **t-SNE**, cho thấy cấu trúc phân cụm trong không gian hai chiều như hình bên dưới:



Nhận xét

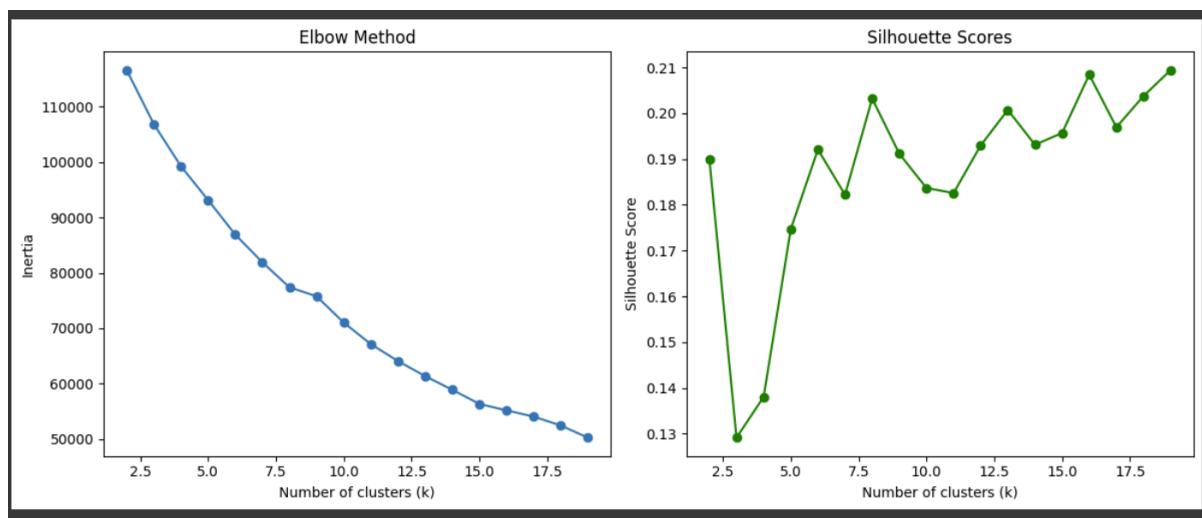
Kết quả trực quan từ mô hình cho thấy **các cụm khách hàng đã được hình thành rõ ràng hơn**, với phần lớn điểm dữ liệu được **phân tách tương đối tốt giữa các cụm**. Tuy nhiên, vẫn tồn tại **một số khu vực có sự chồng lấn**, cho thấy một phần dữ liệu vẫn chưa được phân nhóm rõ ràng hoặc có thể chứa các điểm nhiễu.

Tổng hợp các thuộc tính để xác định số cụm tối ưu

Trong bước tiếp theo, nhóm tiến hành **kết hợp toàn bộ các đặc trưng quan trọng** đã được thử nghiệm ở các mô hình trước, bao gồm: **AgeGroup**, **MPIRange**, **OccupationGroup**, **Gender**, **CompanionGroup**, và **NeedStateGroup**.

Tập hợp đặc trưng này được kỳ vọng sẽ cung cấp một cái nhìn toàn diện hơn về hành vi và đặc điểm của khách hàng, từ đó giúp đánh giá liệu mô hình phân cụm hoạt động hiệu quả hơn.

Để xác định số lượng cụm tối ưu, nhóm sử dụng biểu đồ thể hiện tiêu chí đánh giá **Silhouette Score** và **Elbow Method**, nhằm lựa chọn số cụm hợp lý cho mô hình KMeans.

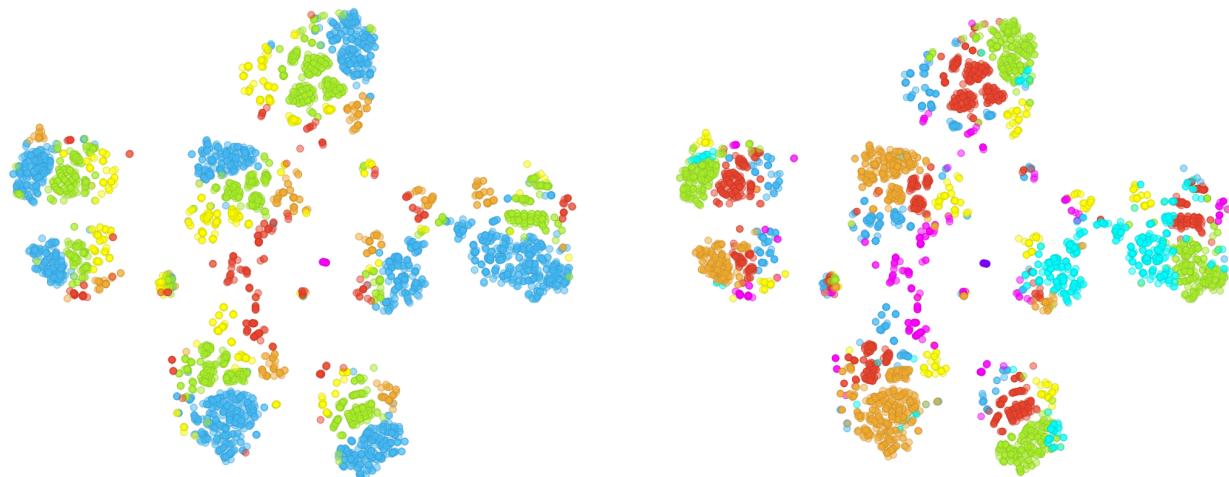


Xác định số cụm tối ưu

Dựa trên biểu đồ **chỉ số Silhouette**, có thể nhận thấy rằng các giá trị Silhouette đạt mức tương đối cao tại các mốc **6**, **8**, và **15 cụm**. Đây là những điểm gợi ý tiềm năng cho số lượng cụm tối ưu. Trong khuôn khổ báo cáo này, nhóm tập trung phân tích các trường hợp **6 cụm** và **8 cụm (mô hình 1.3 và 1.4)**, đồng thời **loại bỏ các phương án có quá nhiều cụm** do gây khó khăn trong diễn giải và ứng dụng thực tế.

Kết quả trực quan hóa bằng thuật toán t-SNE tương ứng với hai trường hợp này được trình bày dưới đây:

- **Hình 1 (trái):** Trực quan hóa kết quả phân cụm với **6 cụm**
- **Hình 2 (phải):** Trực quan hóa kết quả phân cụm với **8 cụm**



Nhìn chung, kết quả phân cụm giữa các phương án **6 cụm** và **8 cụm** không cho thấy sự khác biệt quá rõ rệt. Trong cả hai trường hợp, vẫn có hiện tượng **chồng lấn giữa các cụm**, thậm chí một số cụm có xu hướng **nằm lồng bên trong các cụm khác** khi quan sát qua biểu đồ t-SNE.

Hiện tượng này cho thấy rằng việc sử dụng **toàn bộ các đặc trưng đầu vào** không nhất thiết mang lại hiệu quả phân cụm tốt hơn. Trái lại, một số đặc trưng cụ thể có thể đang **làm loãng tín hiệu phân biệt giữa các nhóm khách hàng**, dẫn đến việc các cụm không được tách biệt rõ ràng.

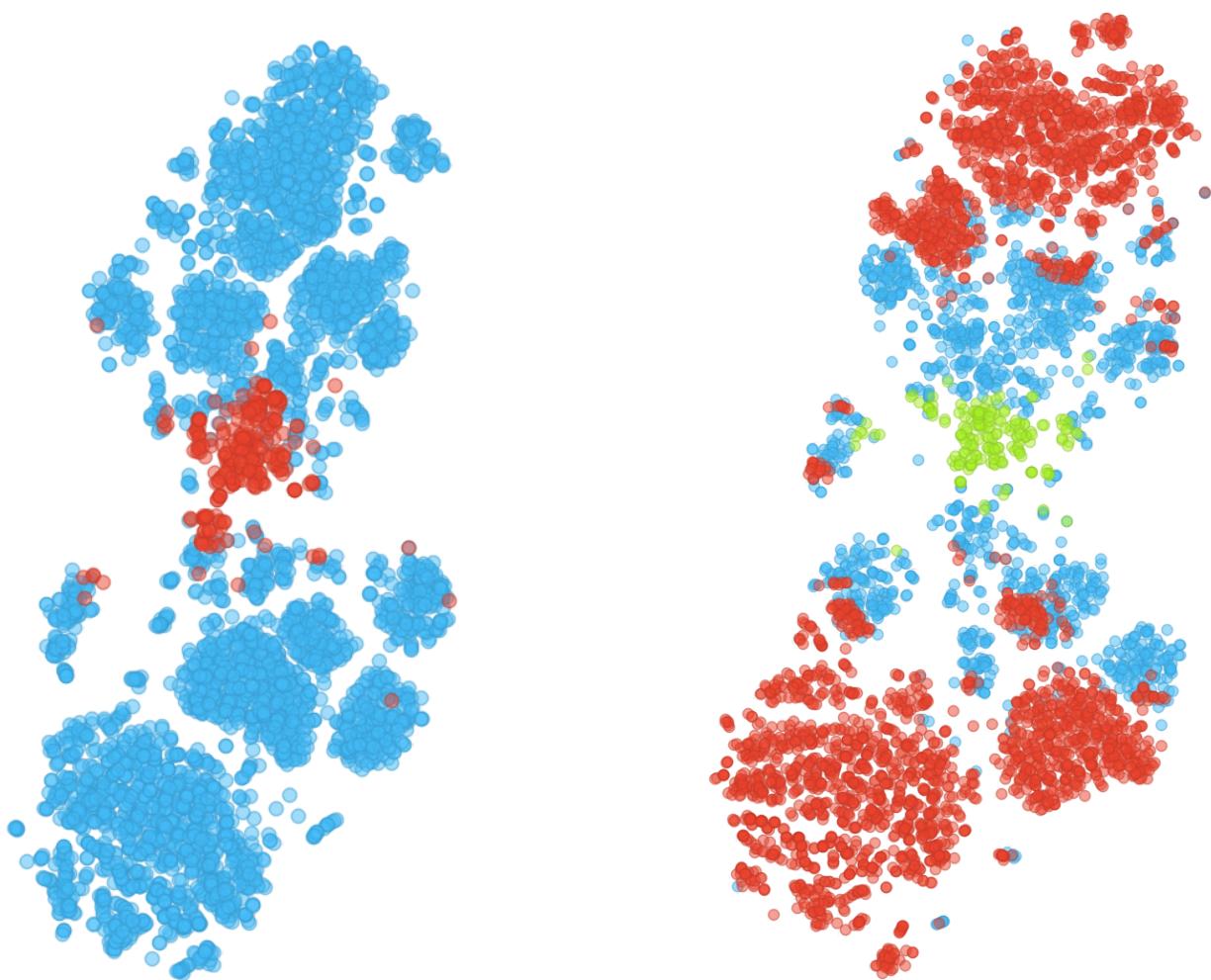
Điều này gợi ý rằng trong các bước tiếp theo, xem xét lại vai trò **đóng góp của từng đặc trưng**, từ đó **lựa chọn có chọn lọc các biến đầu vào** có thể nâng cao chất lượng phân cụm và khả năng diễn giải kết quả phục vụ mục tiêu kinh doanh.

Mô hình 2: Phân cụm với nhóm đặc trưng đầu tiên, mở rộng thêm thông tin về số lượt ghé thăm theo ngày và khung giờ

Kết quả từ nhóm đặc trưng đầu tiên, nhóm tiến hành cân nhắc các đặc trưng đầu vào liên quan đến số lượt ghé thăm theo ngày và khung giờ.

Ở bước đầu tiên, nhóm tiến hành thử nghiệm phân cụm với sáu đặc trưng đầu vào: **AgeGroup**, **NeedStateGroup**, **Gender**, **MPIRange**, **VisitOnDayofweek** và **VisitOnDaypart**, với giả định rằng các đặc trưng này có liên quan đến nhau, cùng xác định nhóm người dùng. Mô hình được lựa chọn là **KMeans**, với số cụm được chọn trước là 2 và 3. Đặt tên cho các mô hình này là 2.1 và 2.2

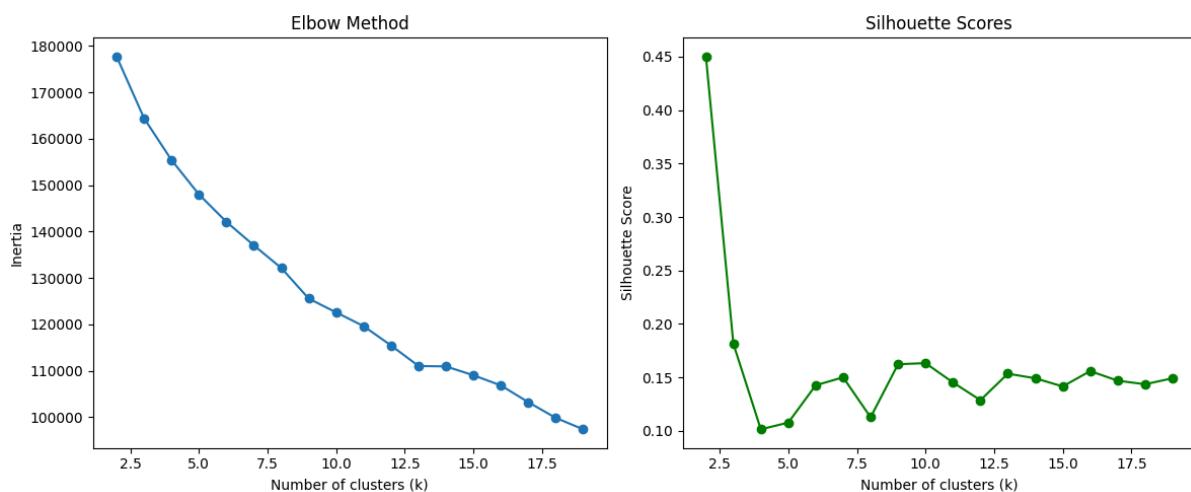
Sau khi áp dụng mô hình, kết quả phân cụm được trực quan hóa bằng thuật toán **t-SNE**, giúp dễ dàng quan sát cấu trúc phân cụm trong không gian hai chiều:



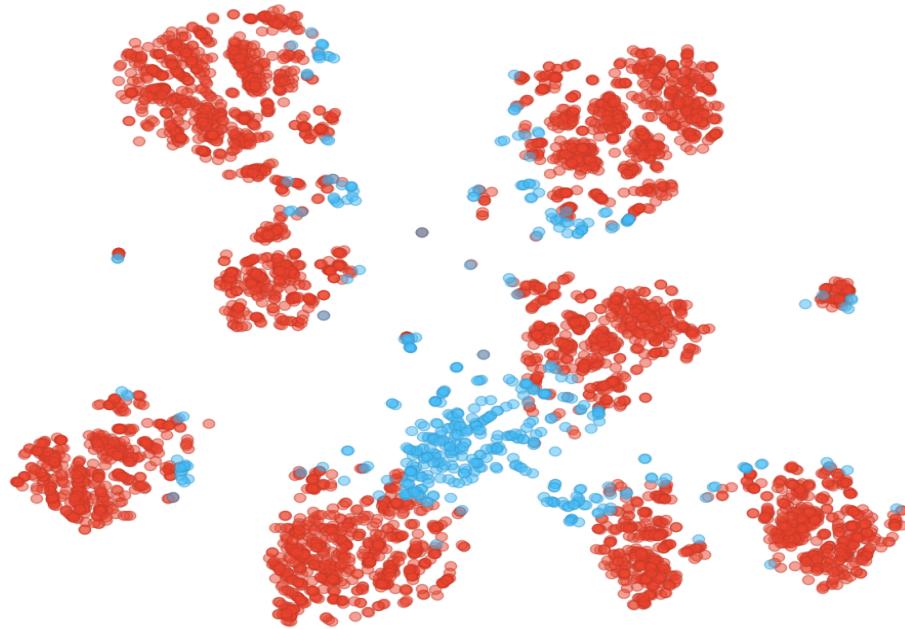
Nhận xét

Kết quả trực quan từ mô hình cho thấy **các cụm khách hàng đã được hình thành rõ ràng hơn**, với phần lớn điểm dữ liệu được phân tách tương đối tốt giữa các cụm. Tuy nhiên, vẫn tồn tại **một số khu vực có sự chồng lấn tương đối nghiêm trọng ở màu đỏ và lam**, và xuất hiện một số cụm khách hàng có số lượng áp đảo so với các cụm khách hàng còn lại, có thể trở thành những cụm khách hàng tiềm năng.

Tương tự như ở mô hình 1, bước tiếp theo, nhóm thử nghiệm mô hình trên toàn bộ các đặc trưng kế thừa trước đó, cộng thêm các đặc trưng liên quan đến số lượt ghé thăm trong từng ngày và từng khung giờ. Trước khi thực hiện mô hình, xác định số cụm cần thiết thông qua Elbow Method và biểu đồ Silhouette Score



Nhận thấy được số cụm tối ưu khoảng 2 cụm, với chỉ số Silhouette tương đối cao so với mô hình 1 ($\sim 0,45$). Kết quả phân cụm thực hiện bằng KMeans (số cụm là 2), gọi là mô hình 2.3 được trực quan hóa bằng thuật toán t-SNE, giúp dễ dàng quan sát cấu trúc phân cụm trong không gian hai chiều, như sau:



Một số nhận định chính từ biểu đồ:

- **Các cụm được hình thành rõ ràng:** Có thể quan sát thấy phần lớn các nhóm khách hàng được tách biệt tốt, thể hiện qua các vùng phân bố riêng biệt, đặc biệt là ở các khu vực rìa ngoài của đồ thị.
- **Hiện tượng chồng lấn vẫn tồn tại:** Một cụm ở trung tâm biểu đồ xuất hiện với mật độ điểm dày đặc và chồng lên nhau nhiều hơn, cho thấy tồn tại những khách hàng có hành vi và đặc điểm tương đối giống nhau khiến mô hình khó phân tách.
- **Hiệu quả của việc bổ sung đặc trưng hành vi:** Việc thêm các thuộc tính về số lượt ghé thăm theo ngày và giờ dường như đã giúp mô hình phân biệt tốt hơn các nhóm khách hàng, biểu đồ của mô hình 2.3 mặc dù còn nhiều giá trị nhiễu, nhưng ít chồng lấn hơn 2.2 và có sự tách biệt cụ thể.
- **Một số điểm nhiễu** (outliers) vẫn tồn tại rải rác, cho thấy có thể cần xử lý hoặc xem xét lại một số dữ liệu ngoại lệ.

Kết luận sơ bộ về chất lượng phân cụm

Dựa trên các thử nghiệm mô hình phân cụm với các nhóm đặc trưng khác nhau, có thể rút ra một số kết luận bước đầu như sau:

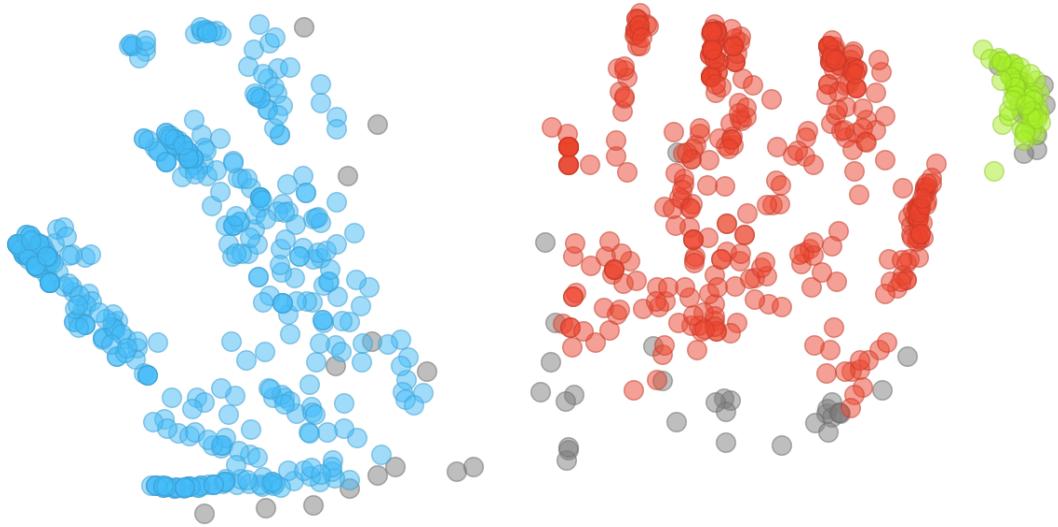
- **Mô hình 1.2 (DBSCAN):** Sử dụng các đặc trưng **MPIRange, OccupationGroup, Gender, NeedStateGroup, và AgeGroup** cho kết quả phân cụm khá rõ ràng. So với việc đưa vào toàn bộ các đặc trưng liên quan đến khách hàng, mô hình này cho thấy **hiệu quả cao hơn** về mặt trực quan và khả năng tách biệt giữa các nhóm khách hàng.
- **Mô hình 2.3 (KMeans với k = 3):** Khi sử dụng các đặc trưng kết hợp giữa đặc điểm nhân khẩu học và hành vi cho thấy **sự tách biệt cụm tương đối tốt, dẫu vẫn còn nhiều điểm cần cải thiện.**

Mô hình 3: Phân cụm với nhóm đặc trưng liên quan đến độ nhận diện thương hiệu và doanh thu từ khách hàng

Đối với mô hình này, nhóm khảo sát các đặc trưng Comprehension, NPSGroup, NPSPer3Month, Spending, PPA, VisitFrequency, Segmentation. Tuy nhiên, tồn tại một vấn đề rất lớn trong các tập dữ liệu sử dụng, đó là giá trị ‘Unknown’ ở các đặc trưng, do không thu thập được thông tin từ người dùng. Ban đầu, ghi nhận khoảng hơn 8000 khách hàng của Highlands, nhưng những người cung cấp đầy đủ các thông tin chỉ có khoảng 1000 người.

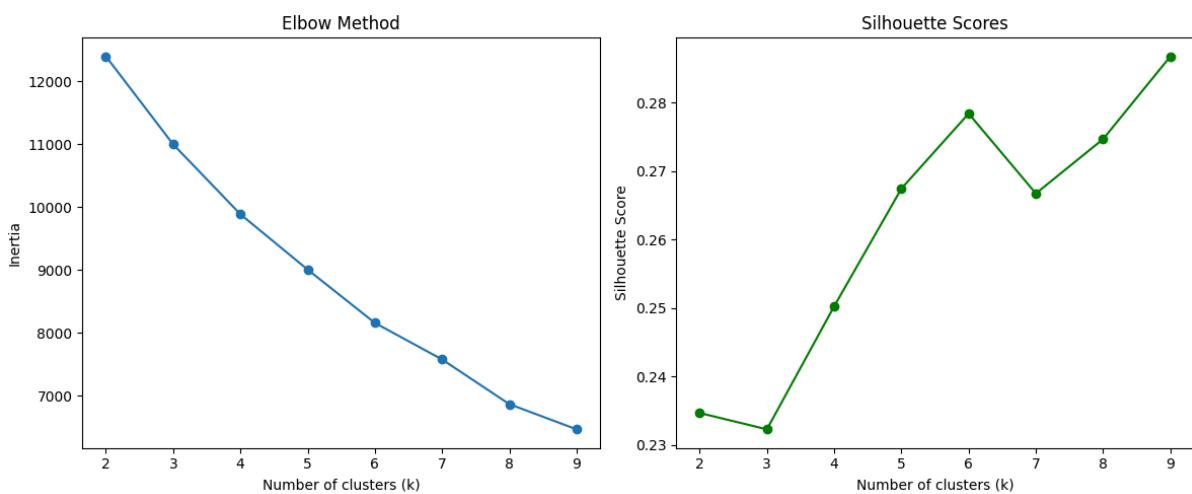
Tất cả các đặc trưng nói trên đều sẽ được xem xét đưa vào mô hình phân cụm, nhưng sẽ đồng thời được xem xét dựa trên 2 thuật toán xây dựng mô hình: DBSCAN và KMeans.

Kết quả phân cụm từ DBSCAN (đặt tên là mô hình 3.1) tiếp tục được trực quan hóa bằng thuật toán **t-SNE**, cho thấy cấu trúc phân cụm trong không gian hai chiều như hình bên dưới:



Nhận xét: Mô hình có khả năng phân tách cụm rõ ràng và dễ nhận biết: Các cụm chính có biên ranh giới tương đối rõ ràng, giúp dễ dàng nhận diện các phân khúc khách hàng khác nhau. Tuy nhiên **một số điểm dữ liệu nhiễu hoặc ngoại lệ** vẫn tồn tại và chưa được phân cụm hiệu quả, **phân bố một số cụm còn phân tán và thiếu chặt chẽ**, có thể gây khó khăn trong việc diễn giải đặc trưng cụ thể. Bên cạnh đó, **vẫn còn các cụm chồng lấn**

Kế đến, nhóm thực hiện mô hình KMeans. Trước khi thực hiện mô hình, xác định số cụm cần thiết thông qua Elbow Method và biểu đồ Silhouette Score:



Nhận thấy được số cụm tối ưu khoảng 6 hoặc 9 cụm, với chỉ số Silhouette không quá khả quan. Kết quả phân cụm thực hiện bằng KMeans (số cụm là 9), gọi là mô hình 3.2

được trực quan hóa bằng thuật toán **t-SNE**, giúp dễ dàng quan sát cấu trúc phân cụm trong không gian hai chiều, như sau

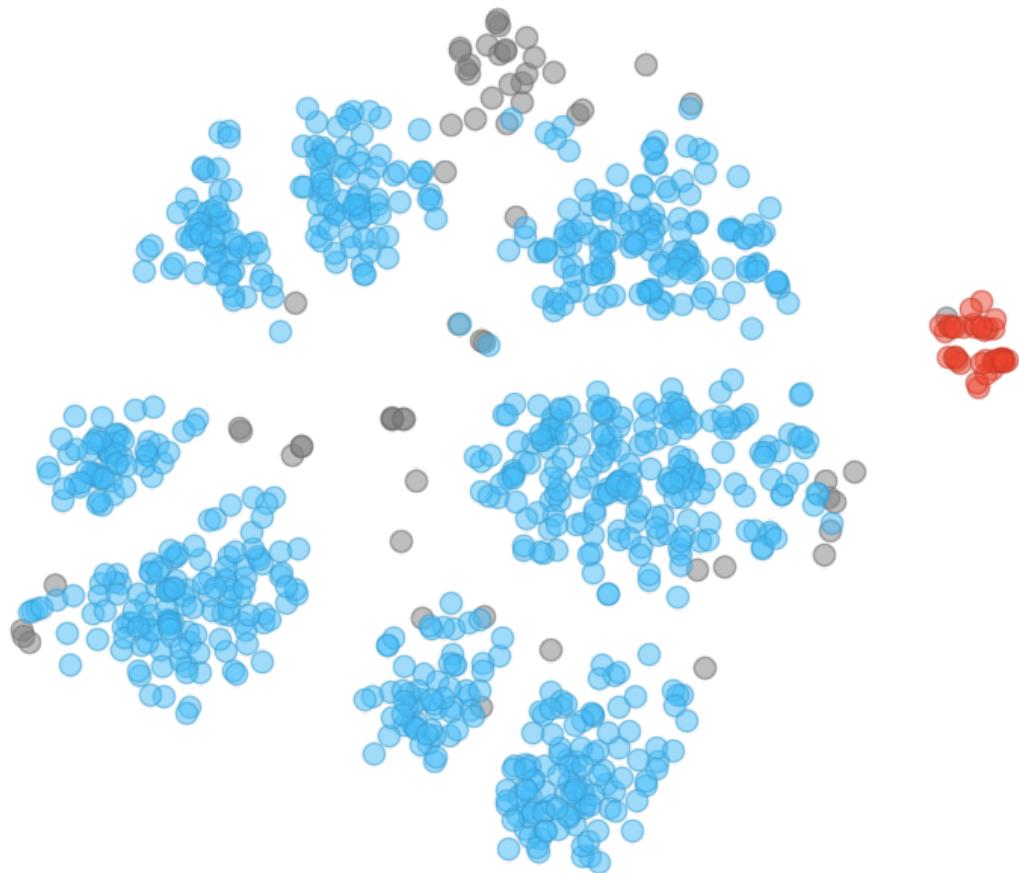


Nhận xét: Các cụm hiển thị trên biểu đồ t-SNE có khoảng cách tương đối đều và tách biệt nhau rõ ràng. Điều này cho thấy mô hình đã phân chia được các nhóm khách hàng có đặc trưng riêng biệt. Tuy nhiên vẫn còn một số hạn chế do các cụm còn chồng lấn hoặc một số giá trị trong cùng một cụm còn tương đối tách biệt nhau

Mô hình 4: Phân cụm với tất cả các đặc trưng

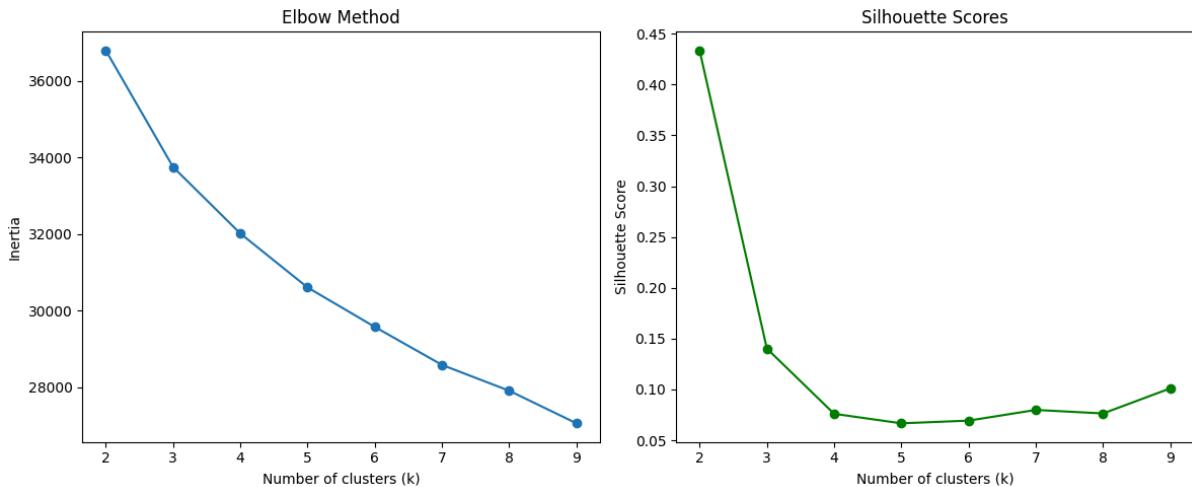
Đối với mô hình này, nhóm khảo sát tất cả các đặc trưng từng được đề cập. Do các giá trị ‘Unknown’ ở một số đặc trưng, những người cung cấp đầy đủ các thông tin có thể sử dụng chỉ còn khoảng 1000 người.

Kết quả phân cụm từ DBSCAN (đặt tên là mô hình 4.1) được trực quan hóa bằng thuật toán t-SNE, cho thấy cấu trúc phân cụm trong không gian hai chiều như hình bên dưới:



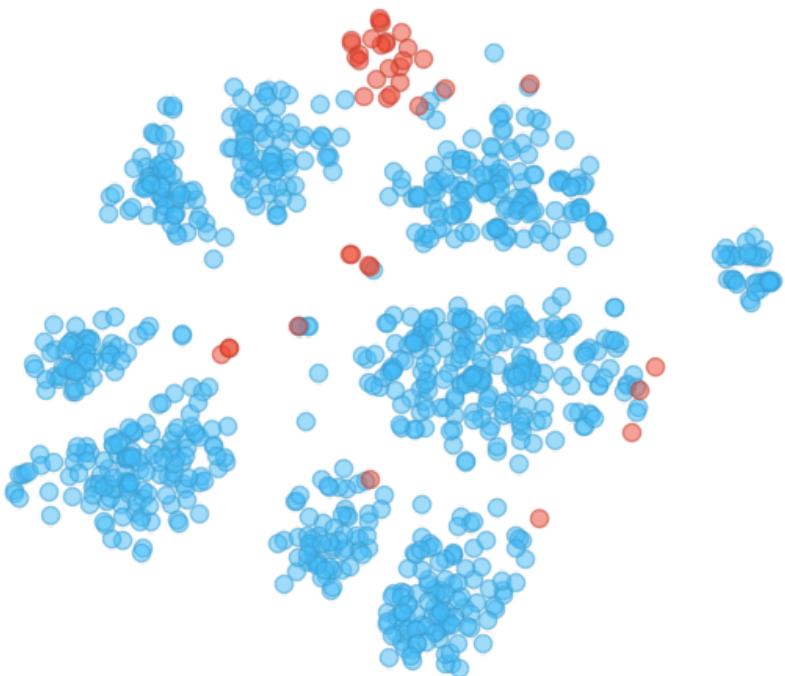
Nhận xét: Mô hình có sự tách biệt hai cụm rõ ràng và khác với các lần trước, hoàn toàn không chồng chéo lên nhau. Tuy nhiên, số lượng các đối tượng trong từng cụm mất cân bằng nghiêm trọng, chưa hẳn là điều bất thường nhưng cần phải xem xét. Bên cạnh đó, một số giá trị nhiều vẫn có thể được quan sát (thể hiện bằng màu xám).

Kế đến, nhóm thực hiện mô hình KMeans. Trước khi thực hiện mô hình, xác định số cụm cần thiết thông qua Elbow Method và biểu đồ Silhouette Score:



Chỉ số Silhouette tại $k = 2$ (tối ưu) không quá thấp, là dấu hiệu khả quan cho mô hình.

Kết quả phân cụm thực hiện bằng KMeans (số cụm là 9), gọi là mô hình 4.2 được trực quan hóa bằng thuật toán t-SNE, giúp dễ dàng quan sát cấu trúc phân cụm trong không gian hai chiều, như sau:



Nhận xét: Cũng tương tự như trực quan của mô hình 4.1, các cụm được phân tách tương đối rõ ràng, các đối tượng tuy vẫn còn trùng lặp nhưng số lượng không nhiều. Vấn đề chính của mô hình là sự mất cân bằng giữa các đối tượng của cụm, trong đó số đối tượng màu đỏ chiếm số lượng rất nhỏ so với tổng thể. Chính vì nguyên nhân này, không thể

quan sát được đặc trưng cụ thể của từng cụm thông qua các phương pháp trực quan hóa dữ liệu.

Nguyên nhân dẫn đến tình trạng trên có thể do số lượng khách hàng từ chối cung cấp các thông tin liên quan đến Spending, PPA, VisitFrequency, ... (những thông tin này ban đầu là dữ liệu trống, sau đó thay bằng Unknown) chiếm số lượng lớn (xấp xỉ 80%). Do có quá ít trường hợp có thể học nên mô hình không có tính chính xác cao, phân cụm mặc dù có chỉ số Silhouette cao nhưng không thể thu hoạch được các đặc trưng liên quan.

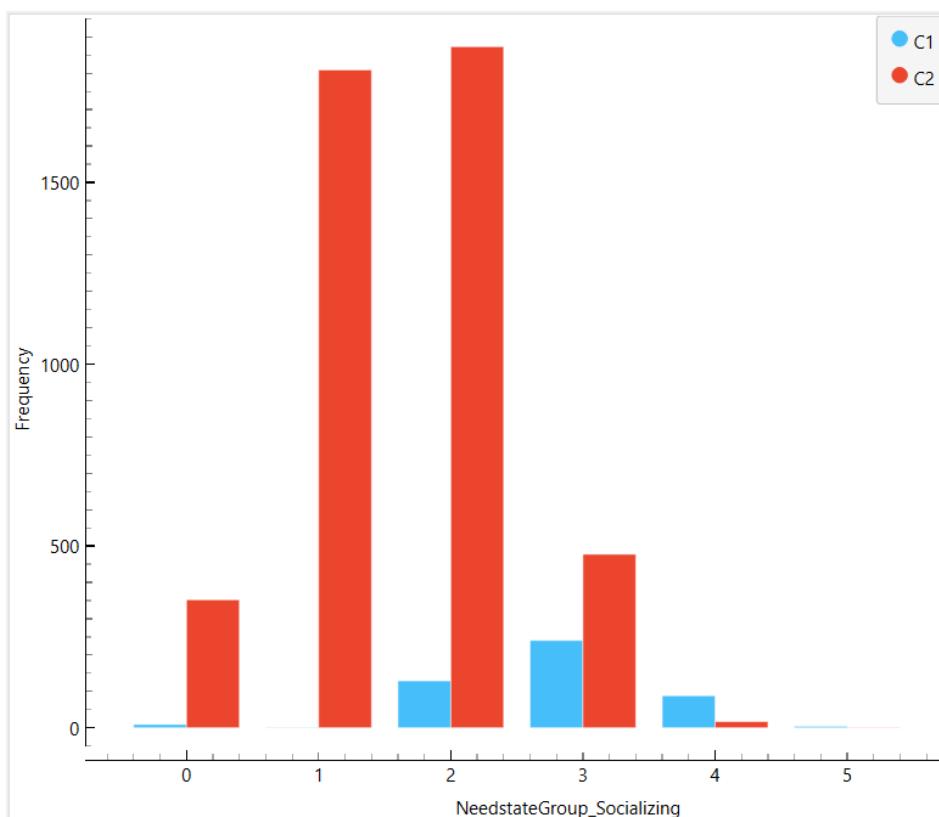
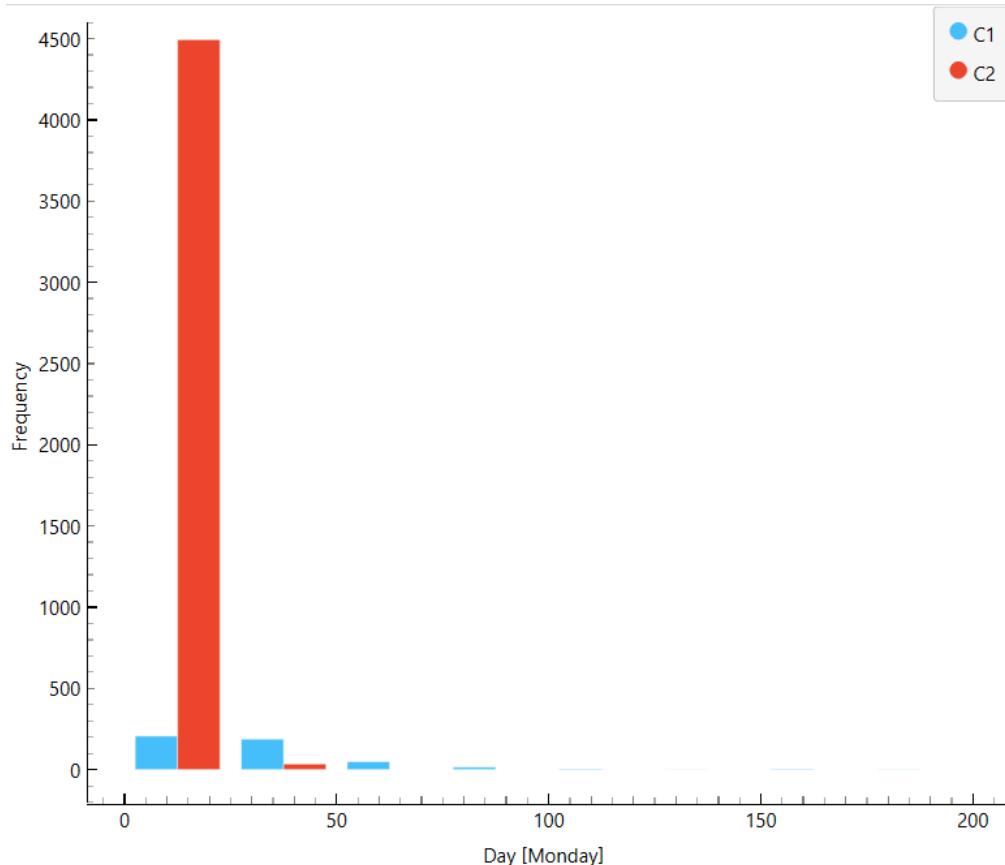
Giải pháp: Chỉ sử dụng một số đặc trưng thay vì toàn bộ các đặc trưng đề ra, cụ thể hơn chính là không sử dụng các đặc trưng Spending, Comprehension, VisitFrequency, PPA, Segmentation. Mô hình 2 chính là các mô hình sử dụng những đặc trưng khác so với các đặc trưng kể trên. Như vậy, ta sẽ thực hiện phân tích trên các cụm của mô hình 2 (mô hình 2.3 là tốt nhất).

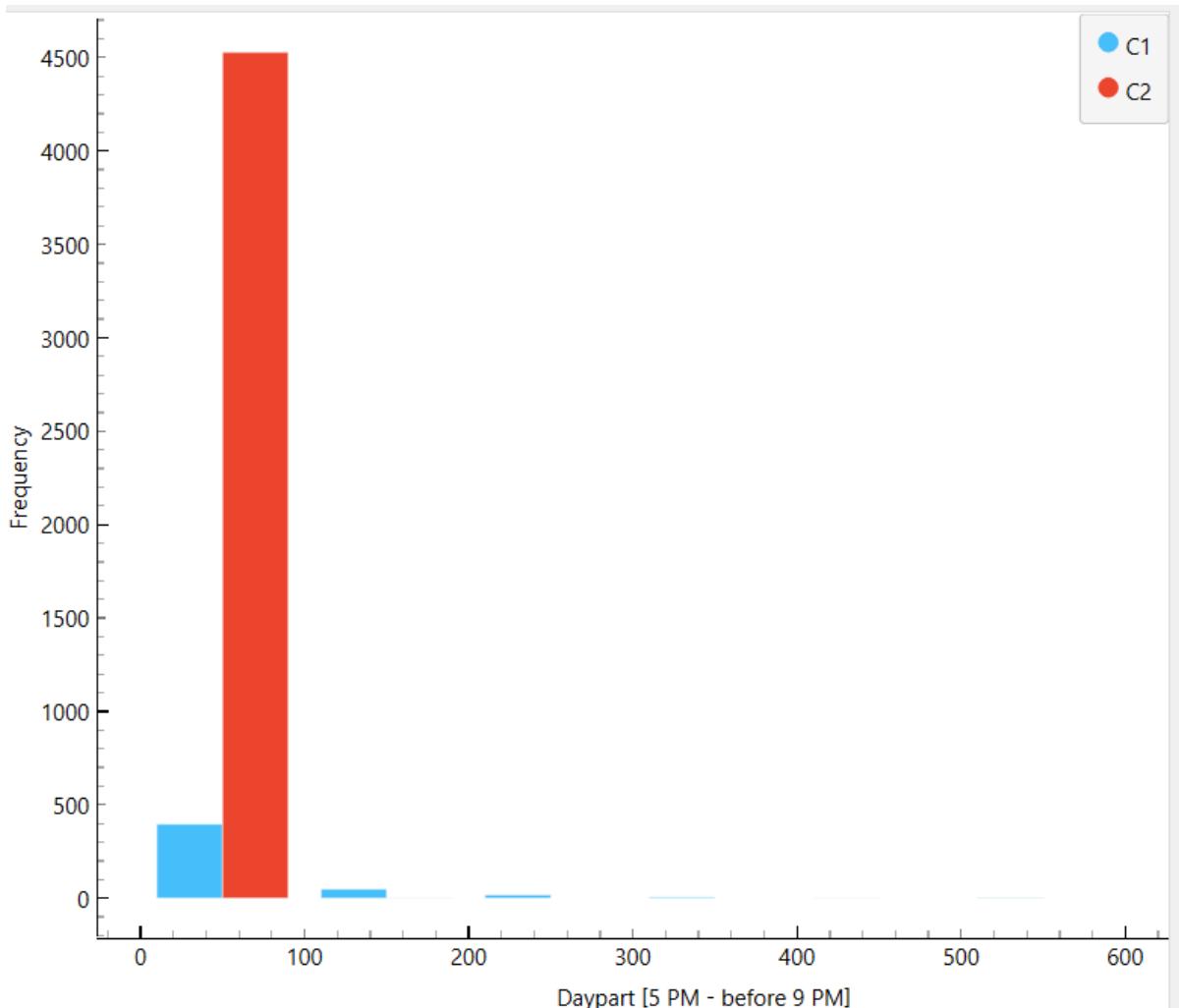
Đánh giá các đặc trưng từ mô hình

Sử dụng mô hình 2.3, phân cụm KMeans với số cụm tối ưu là 2, ta có bảng đánh giá mức độ quan trọng đối với mô hình phân cụm của các đặc trưng đầu vào

	#	Gain ratio	Gini
1	N Day [Monday]	0.163	0.090
2	N Day [Tuesday]	0.161	0.088
3	N Day [Friday]	0.149	0.085
4	N Day [Thursday]	0.135	0.071
5	N Day [Wednesday]	0.134	0.069
6	N CompanionGroup_Alone	0.133	0.038
7	N NeedstateGroup_Working & business meeting	0.100	0.025
8	N NeedstateGroup_Relaxing & entertainment	0.100	0.055
9	N Daypart [Before 9 AM]	0.093	0.053
10	N Day [Sunday]	0.086	0.050
11	N Day [Saturday]	0.084	0.044
12	N CompanionGroup_Colleagues / Business partner	0.079	0.037
13	N NeedstateGroup_Socializing	0.078	0.039
14	N NeedstateGroup_Drinking beverages	0.075	0.040
15	N NeedstateGroup_Meals & Snack	0.072	0.020
16	N CompanionGroup_Others	0.063	0.001
17	C OccupationGroup_Other	2	0.058
18	N CompanionGroup_Friends	0.056	0.029
19	N NeedstateGroup_Studying & Others	0.054	0.007
20	N CompanionGroup_Family	0.049	0.023

Từ đây nhận thấy được: các đặc trưng ảnh hưởng mạnh nhất đến định nghĩa của các cụm là các đặc trưng về số lượt ghé thăm của khách hàng theo ngày, theo khung giờ, theo nhu cầu và bạn đồng hành. Ta sẽ quan sát kỹ hơn vào các nhóm đặc trưng này.





Dựa vào các biểu đồ thu được (minh họa ở trên), có thể suy đoán các đặc trưng của 2 nhóm đối tượng C1 và C2:

- **C2:** nhóm phần lớn các khách hàng, có số lượt ghé thăm vào 1 ngày cụ thể hoặc 1 khung giờ cụ thể không cao, số lần ghé thăm của họ với từng nhóm bạn đồng hành cũng không cao, nhưng họ chiếm phần lớn trong số các khách hàng của thương hiệu
- **C1:** nhóm khách hàng ghé thăm nhiều lần hơn, vào các ngày cụ thể, các khung thời gian cụ thể. Nhìn chung, số lần ghi nhận những khách hàng này trên hệ thống khá cao, gợi ý rằng họ có thể là tệp khách hàng trung thành của thương hiệu.

Đánh giá các hạn chế và đề xuất cải tiến

Hiện tại, mô hình phân cụm còn tồn tại nhiều vấn đề tiềm ẩn, ảnh hưởng đến khả năng khai thác sâu và áp dụng thực tiễn trong việc hiểu rõ từng nhóm khách hàng. Cụ thể:

1. Chưa xác định rõ đặc trưng nội tại của từng cụm

Các cụm khách hàng được hình thành chưa được phân tích sâu để hiểu đặc điểm hành vi, nhu cầu hoặc mức độ rủi ro rời bỏ thương hiệu. Điều này khiến việc sử dụng kết quả phân cụm cho mục tiêu giữ chân khách hàng vẫn còn hạn chế.

2. Chất lượng phân cụm còn thấp

Một số đặc trưng được đưa vào mô hình có đóng góp không đáng kể, dẫn đến việc phân tách giữa các cụm chưa rõ ràng.Thêm vào đó, chỉ số **Silhouette thấp** cho thấy mô hình chưa đạt được khả năng phân tách cụm hiệu quả.

3. Phân cụm còn mang tính tổng quan

Các cụm hiện tại chủ yếu mô tả ở mức bề mặt như nhóm khách hàng trung thành hay không, chưa đi sâu vào các phân khúc hành vi cụ thể như: khách nhạy cảm giá, khách mua dịp lễ, khách đi một mình, v.v.

4. Chưa tận dụng được đầy đủ các đặc trưng quan trọng

Các đặc trưng có khả năng phân loại mạnh như **Spending, PPA**, v.v. chưa được sử dụng hiệu quả do chứa nhiều giá trị thiếu (missing values). Điều này làm giảm đáng kể khả năng mô hình phát hiện ra các nhóm khách hàng có hành vi mua sắm khác biệt.

Đề xuất cải tiến mô hình phân cụm:

1. Loại bỏ hoặc gộp các đặc trưng không có giá trị phân biệt cao

- Dựa trên chỉ số **Gain Ratio / Gini**, loại bỏ hoặc kết hợp các biến đóng góp thấp để giảm nhiễu cho mô hình.

2. Cải thiện chất lượng dữ liệu đầu vào

- Khôi phục thông tin từ các biến như Spending, PPA, ... vốn là các thông tin không được người dùng cung cấp hoặc mất đi do quá trình nhập liệu xảy ra sai sót.

- Một số thuộc tính có thể sử dụng làm đặc trưng đầu vào như TypeOfDay (Weekdays, Weekends), lưu ý nếu thêm thuộc tính này phải loại bỏ Dayofweek để tránh nhiễu. Các khung giờ cũng có thể được phân loại thành Morning, Afternoon, Evening, ... để giảm số lượng giá trị riêng biệt.

3. Thử nghiệm thêm thuật toán phân cụm khác

- Ngoài K-Means và DBSCAN, có thể áp dụng **Hierarchical Clustering** (để tìm cấu trúc dữ liệu tốt hơn), hoặc **Gaussian Mixture Models** (để phát hiện nhóm mềm - soft clustering).
- Đồng thời, đánh giá lại chỉ số **Silhouette**, **Calinski-Harabasz** hoặc **Davies-Bouldin** để so sánh chất lượng mô hình.

4. Tái phân cụm theo từng phân khúc con (subgroup)

- Phân cụm riêng cho từng nhóm khách hàng để phát hiện hành vi đặc thù theo ngữ cảnh

Giai đoạn 3: Xây dựng mô hình dự đoán rời bỏ

- Từ tập dữ liệu SA (vì mỗi ID là duy nhất và đều tham gia vào quá trình khảo sát) kết hợp với tập dữ liệu Dataset_Clustering đã được phân cụm ở câu 2 tiến hành xây dựng mô hình.
- **Tổng quan dữ liệu**
 - + Kích thước tập huấn luyện: (4000, 16)
 - + Kích thước tập kiểm tra: (1000, 16)
 - + Biến mục tiêu: Target (0: Không rời bỏ, 1: Rời bỏ)
- **Đặc trưng đầu vào**
 - + Đặc trưng phân loại (Categorical): 'City', 'MPIMean', 'TOM', 'BUMO', 'BUMOPrevious', 'MostFavourite', 'Gender', 'AgeGroup', 'Occupation', 'OccupationGroup', 'MPIRange', 'Cluster' (chọn CLuster bởi vì đây là kết quả phân cụm từ câu 2 và hợp lý nhất để đưa vào xây dựng mô hình).

- + Đặc trưng số (Numerical): 'GroupSize', 'Age', 'Year'

```
=====
KẾT QUẢ ĐÁNH GIÁ: LOGISTIC REGRESSION =====
precision    recall    f1-score   support
          0       0.82      0.99      0.90      824
          1       0.00      0.00      0.00      176

   accuracy                           0.82      1000
macro avg       0.41      0.50      0.45      1000
weighted avg    0.68      0.82      0.74      1000

ROC-AUC Score: 0.6907
```

- Mô hình Logistic Regression này đang hoạt động rất kém và hoàn toàn không hiệu quả cho bài toán của bạn. Nó đã thất bại trong việc học cách nhận diện lớp thiểu số (lớp 1).

1. Vấn đề nghiêm trọng nhất: Hoàn toàn thất bại ở lớp 1

- Đây là điểm đáng báo động nhất trong kết quả của bạn:
 - + recall = 0.00: Điều này có nghĩa là trong tổng số 176 trường hợp thực tế thuộc lớp 1, mô hình của bạn không tìm thấy được bất kỳ trường hợp nào. Nó đã bỏ lỡ 100% các trường hợp quan trọng mà bạn muốn dự đoán.
 - + precision = 0.00: Điều này có nghĩa là mô hình chưa bao giờ đưa ra dự đoán là lớp 1. Hoặc nếu có, tất cả các lần đó đều sai.
 - + f1-score = 0.00: Vì cả precision và recall đều bằng 0, f1-score (chỉ số trung bình của chúng) cũng bằng 0.

2. Chỉ số Accuracy (Độ chính xác) cao nhưng gây hiểu lầm

- Mặc dù accuracy là 0.82 (82%), con số này không phản ánh hiệu suất thực sự của mô hình. Lý do là vì bộ dữ liệu của bạn bị mất cân bằng (imbalanced data).
- Dữ liệu của bạn có 824 mẫu lớp 0 và 176 mẫu lớp 1.

- Độ chính xác 82% của bạn gần như y hệt với một mô hình không học được gì cả. Điều này khẳng định rằng mô hình của bạn đã không học được bất kỳ mẫu (pattern) hữu ích nào để phân biệt lớp 1 với lớp 0.



```
===== KẾT QUẢ ĐÁNH GIÁ: XGBOOST =====
      precision    recall   f1-score   support
0          0.84     0.93     0.88      824
1          0.33     0.16     0.21      176

accuracy                           0.79      1000
macro avg                           0.58      1000
weighted avg                          0.75      1000

ROC-AUC Score: 0.7163
```

- So với mô hình Logistic Regression trước đó, mô hình XGBoost này đã có một sự cải thiện đáng kể, tuy nhiên, hiệu suất của nó vẫn còn ở mức thấp và chưa đủ tốt để áp dụng hiệu quả vào thực tế. Nó đã bắt đầu "học" nhưng vẫn chưa "giỏi".

Phân tích chi tiết

- Điểm tích cực (The Positives)
 - + Đã bắt đầu "học" được lớp thiểu số (lớp 1): Không giống như mô hình Logistic Regression có precision/recall bằng 0, mô hình XGBoost đã có thể đưa ra dự đoán cho lớp 1. Các chỉ số precision (0.33) và recall (0.16) khác 0 cho thấy mô hình đã không còn bỏ qua hoàn toàn lớp thiểu số nữa. Đây là một bước tiến quan trọng.
 - + Điểm ROC-AUC khá (0.7163): Điểm số này cho thấy mô hình có khả năng phân biệt giữa hai lớp tốt hơn đáng kể so với việc đoán ngẫu nhiên (có điểm là 0.5). Nó chứng tỏ các đặc trưng (features) của bạn có chứa thông tin dự đoán và mô hình có tiềm năng để cải thiện.
- Điểm yếu cần khắc phục (Weaknesses to Address):
 - + Recall rất thấp (0.16) - **Vẫn đề nghiêm trọng nhất:**

- Ý nghĩa: Mô hình chỉ phát hiện được 16% trong tổng số những khách hàng thực sự sẽ rời đi. Điều này có nghĩa là bạn đang bỏ lót mất 84% khách hàng tiềm năng.
 - Tác động kinh doanh: Nếu mục tiêu là để triển khai các chiến dịch giữ chân khách hàng, việc bỏ sót 84% đối tượng mục tiêu sẽ làm cho chiến dịch gần như không hiệu quả.
- + Precision thấp (0.33):
- Ý nghĩa: Khi mô hình của dự đoán một khách hàng sẽ rời đi, nó chỉ đúng trong 33% các trường hợp.
 - Tác động kinh doanh: Điều này có nghĩa là gần 2/3 nguồn lực (thời gian, tiền bạc, khuyến mãi) dành cho việc giữ chân khách hàng sẽ bị lãng phí vào những người vốn dĩ không có ý định rời đi.
- + Accuracy thấp hơn nhưng đừng lo lắng: Điều này là bình thường và thực ra là một dấu hiệu tốt. Mô hình Logistic Regression có độ chính xác cao một cách giả tạo vì nó chỉ "lười biếng" đoán một chiều là lớp 0. Mô hình XGBoost "đúng cảm" hơn, nó cố gắng dự đoán lớp 1 và chấp nhận sai ở một vài trường hợp, do đó accuracy tổng thể giảm đi nhưng lại phản ánh đúng nỗ lực giải quyết bài toán hơn.

===== KẾT QUẢ ĐÁNH GIÁ XGBOOST VỚI SMOTE =====				
	precision	recall	f1-score	support
0	0.84	0.94	0.89	824
1	0.38	0.17	0.24	176
accuracy			0.81	1000
macro avg	0.61	0.56	0.56	1000
weighted avg	0.76	0.81	0.77	1000

Phân tích chi tiết chỉ số:

- **Ưu điểm:**
 - + Độ chính xác tổng thể (Accuracy) được cải thiện nhẹ, từ **0.79 → 0.81**

- + Mô hình vẫn giữ được khả năng phân loại tốt với lớp **không rời bỏ (0)**: Recall đạt **0.94**, F1-score **0.89**
- **Hạn chế:**
 - + Với lớp **rời bỏ (1)** – nhóm khách hàng mục tiêu cần phát hiện:
 - **Recall chỉ đạt 0.17** (tăng không đáng kể so với 0.16 trước đó)
 - F1-score vẫn thấp (**0.24**) do mô hình còn yếu trong việc nhận diện chính xác khách hàng rời bỏ
 - **Precision cao hơn (0.38)**, nghĩa là khi dự đoán có người rời bỏ, mô hình đúng hơn một chút – tuy nhiên mô hình vẫn **bỏ sót nhiều khách hàng rời bỏ**
 - **SMOTE giúp cải thiện nhẹ độ chính xác tổng thể và độ bao phủ nhóm thiểu số**, nhưng hiệu quả với lớp rời bỏ vẫn còn **hạn chế**.
 - Điều này cho thấy **SMOTE chưa đủ** để giải quyết mất cân bằng, có thể do:
 - + Dữ liệu gốc chứa tín hiệu yếu với lớp rời bỏ
 - + Mô hình cần thêm tuning (siêu tham số) hoặc kết hợp thêm kỹ thuật khác

===== KẾT QUẢ SAU KHI TINH CHỈNH =====				
	precision	recall	f1-score	support
0	0.85	0.93	0.89	824
1	0.41	0.23	0.29	176
accuracy			0.81	1000
macro avg	0.63	0.58	0.59	1000
weighted avg	0.77	0.81	0.78	1000

Cải thiện đạt được sau tinh chỉnh:

- **Lớp rời bỏ (1):**
 - + **Recall tăng rõ rệt**: từ **0.17** (sau SMOTE) → **0.23**
 - + **Precision cũng tăng**: từ **0.33** → **0.41**
 - + **F1-score tăng từ 0.21 → 0.29**, phản ánh mô hình cân bằng hơn giữa độ chính xác và độ bao phủ
- **Lớp không rời bỏ (0)** vẫn giữ hiệu quả cao (Recall 0.93, F1 0.89)

- + Accuracy giữ nguyên ở mức **0.81**, nhưng macro average và weighted average đều tăng nhẹ
- Việc tinh chỉnh siêu tham số đã mang lại hiệu quả thực sự, đặc biệt với lớp thiểu số (rời bỏ).
- Dù Recall của lớp 1 vẫn chưa cao, mô hình **bắt đầu học được tín hiệu phân biệt khách hàng rời bỏ**, điều mà mô hình gốc chưa làm được.
- Tuy nhiên, đây vẫn chưa phải kết quả lý tưởng nếu mục tiêu là **tối đa hóa khả năng phát hiện khách hàng có nguy cơ rời bỏ**.



KẾT QUẢ MÔ HÌNH GIẢM FEATURES

	precision	recall	f1-score	support
0	0.83	0.96	0.89	824
1	0.26	0.07	0.11	176
accuracy			0.80	1000
macro avg	0.54	0.51	0.50	1000
weighted avg	0.73	0.80	0.75	1000

Nhận xét chi tiết mô hình sau khi giảm số đặc trưng:

- **Ưu điểm:**
 - + Mô hình vẫn giữ được độ chính xác cao với lớp không rời bỏ (0): Recall rất cao (**0.96**), F1-score tốt (**0.89**); **độ chính xác tổng thể (Accuracy)** đạt **80%**, gần tương đương các mô hình trước
- **Hạn chế nghiêm trọng:**
 - + **Hiệu suất với lớp rời bỏ (1) giảm mạnh:**
 - + Recall chỉ còn **0.07** (rất thấp, nghĩa là mô hình gần như bỏ sót nhóm khách hàng rời bỏ)
 - + Precision thấp (**0.26**), F1-score rất thấp (**0.11**)
- **Macro average và weighted average giảm đáng kể**, phản ánh sự mất cân bằng nghiêm trọng trong hiệu suất giữa hai lớp
- Việc **giảm số lượng đặc trưng** đã khiến mô hình mất đi nhiều tín hiệu quan trọng để phân biệt lớp rời bỏ.

- Mặc dù độ chính xác tổng thể **vẫn cao**, mô hình **hầu như không còn khả năng phát hiện khách hàng rời bỏ** – điều này **nguy hiểm trong thực tiễn**, vì có thể dẫn đến mất khách mà không có cảnh báo.



KẾT QUẢ MÔ HÌNH SAU KHI ÁP DỤNG SMOTE:

	precision	recall	f1-score	support
0	0.84	0.66	0.74	824
1	0.20	0.41	0.27	176
accuracy			0.62	1000
macro avg	0.52	0.53	0.51	1000
weighted avg	0.73	0.62	0.66	1000

Điểm tích cực:

- **Recall của lớp rời bỏ (1)** đã được cải thiện **đáng kể**: Tăng từ ~0.17–0.23 trước đây lên **0.41**, nghĩa là mô hình **đã nhận diện được gần một nửa số khách hàng rời bỏ**
- Đây là một bước tiến quan trọng nếu mục tiêu là **phát hiện sớm và ngăn chặn churn**
- **Recall của lớp 0 (không rời bỏ)** vẫn khá ổn (**0.66**)

Hạn chế:

- **Độ chính xác tổng thể (accuracy)** giảm mạnh: từ ~0.81 xuống **0.62**, do:
 - + Precision của lớp 1 chỉ là **0.20** → mô hình dự đoán sai khá nhiều khách không rời bỏ thành rời bỏ
 - + F1-score lớp 1 vẫn ở mức thấp (**0.27**), thể hiện sự **đánh đổi giữa recall và precision**
- Đây là một ví dụ điển hình của **tình huống đánh đổi (trade-off)** giữa:
 - + **Accuracy tổng thể** và **Hiệu quả phát hiện lớp thiểu số (khách rời bỏ)**
 - + Mô hình này là **một bước tiến tốt**, vì nó ưu tiên **recall cho lớp 1**

KẾT QUẢ MÔ HÌNH TỐI ƯU:

	precision	recall	f1-score	support
0	0.85	0.53	0.65	824
1	0.21	0.58	0.30	176
accuracy			0.54	1000
macro avg	0.53	0.55	0.48	1000
weighted avg	0.74	0.54	0.59	1000

Điểm tích cực nổi bật:

- **Recall của lớp rời bỏ (1)** đạt **0.58**, cao nhất trong tất cả các mô hình trước đó → đây là mô hình **hiệu quả nhất nếu mục tiêu là phát hiện khách hàng rời bỏ**.
- **Recall của lớp 0** giảm xuống **0.53**, nhưng **precision vẫn cao (0.85)** → mô hình cẩn trọng khi dự đoán khách không rời bỏ.
- **Mô hình ưu tiên phát hiện churn** bằng cách **hy sinh độ chính xác tổng thể** – phù hợp với chiến lược cảnh báo sớm và can thiệp.

Hạn chế:

- **Accuracy giảm xuống chỉ còn 0.54**, cho thấy mô hình dự đoán sai khá nhiều → không phù hợp nếu doanh nghiệp đòi hỏi sự chính xác tổng thể cao.
- **Precision lớp 1 (0.21)** thấp → trong số các khách bị dự đoán là “rời bỏ”, nhiều người thực ra **không rời bỏ** → dễ gây **can thiệp sai mục tiêu** nếu không có bước lọc t

Tổng kết:

- Mô hình sau tối ưu **đã đạt được mục tiêu chính là: tăng cường phát hiện khách hàng có nguy cơ rời bỏ (churn)**, với Recall lớp 1 lên đến **58%**.
- Đây là **chiến lược phù hợp với các doanh nghiệp ưu tiên retention hơn accuracy**, sẵn sàng chấp nhận "báo động giả" để giữ khách hàng thực sự rời bỏ.
- Tuy nhiên, **chi phí can thiệp nhầm (false positives)** cần được cân nhắc trong thực tế vận hành.

Giai đoạn 4: Đề xuất cải tiến mô hình

Mục tiêu cần ưu tiên

- Trong bối cảnh bài toán dự đoán rời bỏ khách hàng, **việc phát hiện đúng những khách hàng sắp rời bỏ (recall lớp 1 cao)** thường **quan trọng hơn việc đạt accuracy cao**, vì:
 - + Doanh nghiệp có thể chủ động giữ chân khách
 - + Chi phí giữ chân có thể thấp hơn so với mất khách

Đề xuất cải tiến mô hình

- Điều chỉnh ngưỡng phân loại (threshold tuning):
 - + Không dùng ngưỡng mặc định 0.5
 - + Dịch ngưỡng về phía thấp hơn (ví dụ: 0.4, 0.35...) để **tăng recall lớp 1**, sau đó đo lại precision, F1
- Kết hợp mô hình nhiều tầng (Two-stage model):
 - + Giai đoạn 1: Mô hình nhấn mạnh **recall** để bắt hết khách có nguy cơ
 - + Giai đoạn 2: Mô hình nhấn mạnh **precision** để lọc lại, giúp **giảm báo động giả**
- Tối ưu bằng cost-sensitive learning:
 - + Áp dụng trọng số hoặc loss function ưu tiên phát hiện lớp thiểu số (churn)
 - + Trong XGBoost, có thể dùng scale_pos_weight = #neg / #pos
- Kết hợp SMOTE + kỹ thuật lọc (như SMOTE-Tomek):
 - + SMOTE giúp cân bằng, nhưng đi kèm rủi ro sinh ra nhiễu
 - + Kết hợp với Tomek Links giúp loại bỏ các điểm "gần biên", làm mô hình ổn định hơn
- Giữ lại các đặc trưng quan trọng (không giảm đặc trưng bừa bãi):
 - + Tránh loại bỏ các biến liên quan đến hành vi/thái độ như MPIMean, TOM, BUMO, Cluster
 - + Có thể dùng **SHAP** hoặc **permutation importance** để chọn đặc trưng thật sự có giá trị
- **Sử dụng các thước đo phù hợp hơn cho bài toán mất cân bằng:**
 - + F1-score lớp 1
 - + ROC AUC riêng cho lớp 1
 - + PR AUC (Precision-Recall Curve) thay vì accuracy

