Đề bài & Dataset: Link

CÂU 1: ĐÁNH GIÁ CHẤT LƯỢNG VÀ TIỀN XỬ LÝ DỮ LIỆU

File	Dư thừa	Gán nhãn sai	Lỗi logic	Thiếu dữ liệu	Bất nhất giữa bảng
2017Segmentation 3685case.csv	0	0	0	0	Có — check với Brandhealth.csv
Brandhealth.csv	0	Có (cột nhãn Segmentation bị trùng/nhập nhầm)	0	0	Có
Brand_Image.csv	13,974	Có (Awareness nhiều giá trị null)	Có (Year không khớp giữa bảng)	397	Có
Companion.csv	799	Có	0	0	Có
Competitor database_xlnm#_FilterDatabase.csv	0	0	0	0	Có
Dayofweek.csv	37	0	0	0	Có
Daypart.csv	0	0	0	0	Có
NeedstateDayDaypart.csv	72	0	0	0	Có
SA#var.csv	0	Có	Có (giá trị Age âm hoặc > 100)	nhiều cột missing (Age, MPI#Mean, MPI#detail, MPI)	Có

1.1. Bản liệt kê chi tiết các lỗi dữ liệu

1.2. Tóm tắt phương pháp và tiêu chí xử lý từng lỗi dữ liệu

Loại lỗi	Phương pháp xử lý	Tiêu chí xử lý
Dư thừa	Xóa dòng trùng lặp, bỏ cột không dùng	Nếu duplicated rows > 0, xóa giữ lại bản ghi đầu tiên
Gán nhãn sai	Chuẩn hóa nhãn về một format thống nhất	Chuyển tất cả nhãn về chữ hoa/thường đồng nhất, map lại value
Lỗi logic	Kiểm tra điều kiện giá trị, sửa về giá trị hợp lệ	Ví dụ: cột Age phải từ 0–100, Spending ≥ 0
Thiếu dữ liệu	Điền giá trị hợp lý theo ngữ cảnh dữ liệu	- Trung bình nếu là số liên tục - Mode nếu là nhãn - Xóa nếu số lượng thiếu < 5%
Bất nhất giữa bảng	Đối chiếu ID, Year, Brand giữa các bảng, xử lý chênh lệch	Mapping lại ID nếu lệch format, loại bản ghi lệch năm hoặc thiếu Brand

1.3. Phân tích nguyên nhân gốc và đề xuất giải pháp

Nguyên nhân gốc:

- **Dữ liệu từ nhiều nguồn, thời điểm, người nhập liệu khác nhau**
  - **Biểu hiện**: Các file như Brandhealth.csv, Companion.csv, Dayofweek.csv, Brand_Image.csv đều có cấu trúc và kiểu dữ liệu không đồng bộ.
  - **Hệ quả**: Khó tổng hợp, kết nối và đảm bảo tính nhất quán khi phân tích dữ liệu.
- **Format không đồng bộ giữa các bảng**
  - **Biểu hiện:**

- File Brand_Image.csv có dữ liệu số ở dạng int64 nhưng các file như Companion.csv, Dayofweek.csv lại lưu tất cả dưới dạng object.
- Các cột có dấu phân cách khác nhau (như ID;City;... thay vì chuẩn CSV với dấu ,).
  - **Hệ quả**: Gây lỗi khi merge hoặc xử lý phân tích, cần bước tiền xử lý định dạng tốn thời gian.

- **Không có khóa chính rõ ràng**
  - **Biểu hiện**:
- Companion.csv chỉ có 4 cột dạng object, không có khóa định danh duy nhất (ID, Year không đủ phân biệt nếu trùng).
- Brand_Image.csv có đến 13.974 dòng trùng lặp, chứng tỏ ID-Year-City-Attribute không đảm bảo khóa chính.
  - **Hệ quả:** Gây lỗi logic khi merge dữ liệu hoặc training mô hình, có thể trùng lặp dữ liệu đầu vào.
- **Thiếu chuẩn hóa kiểu dữ liệu và nhãn**
  - **Biểu hiện**:
- Cột như Gender, Age#group, Needstates không dùng enum mà là chuỗi tự do.
- Kiểu dữ liệu của một số cột như Spending, PPA, NPS được lưu dưới dạng object thay vì float hoặc int.
  - **Hệ quả:** Không thể tính toán số học ngay, dễ gây lỗi phân tích hoặc mô hình không học được.

Đề xuất giải pháp:

- **Chuẩn hóa template nhập liệu & áp dụng validate form**
  - **Chi tiết:**
- Thiết kế mẫu biểu nhập dữ liệu (Google Form, Excel Template hoặc webform) bắt buộc người dùng nhập đúng định dạng.
- Dùng công cụ như Google Data Studio, PowerApps hoặc custom app để thu thập và chuẩn hóa ngay khi nhập.
  - **Hiệu quả:** Giảm sai sót thủ công, đảm bảo dữ liệu từ đầu đúng chuẩn.
- **Áp dụng enum/dictionary mapping cho nhãn**
  - **Chi tiết:**
- Ví dụ: Gender chỉ được chấp nhận giá trị trong {0: 'Male', 1: 'Female'}.
- Các cột như Age Group, Companion group, Needstate Group được ánh xạ bằng từ điển chuẩn.
  - **Hiệu quả:** Dễ phân tích, lọc, chuyển đổi nhãn về dạng số cho machine learning.
- **Xác định khóa chính thống nhất: ID + Year**
  - **Chi tiết:**
- Tất cả các bảng liên kết theo ID và Year (hoặc thêm City nếu cần).

- Các bảng không có Year phải được bổ sung hoặc loại bỏ nếu không còn liên quan.
  - **Hiệu quả:** Đảm bảo không bị lỗi trùng dòng khi merge dữ liệu từ nhiều bảng.
- **Xây dựng pipeline tiền xử lý dữ liệu**
  - **Gợi ý pipeline (bằng Python hoặc các công cụ ETL):**
- Chuẩn hóa tên cột: loại bỏ dấu #, viết thường, thay khoảng trắng bằng _.
- Kiểm tra và ép kiểu dữ liệu: các cột Spending, PPA, NPS phải là float.
- Loại bỏ dòng trùng: đặc biệt trong Brand_Image.csv, Companion.csv.
- Tạo bảng ánh xạ nhãn (dictionary mapping).
- Validate khóa chính: dùng pandas.duplicated() để kiểm tra trùng ID-Year.

1.4. Bộ Dữ Liệu Đã Được Làm Sạch và Tiền Xử Lý

- **Xóa Cột Dư Thừa**
  - Các cột không cần thiết hoặc không chứa thông tin hữu ích cho phân tích đã được loại bỏ khỏi các bảng dữ liệu.

Cụ thể:

- Các cột kỹ thuật hoặc cột trùng nội dung như Col, Age#Group, MPI#2, MPI_Mean_Use, Age#Group#2 trong file SA#var.csv đã bị loại bỏ.
- Những cột chứa ký hiệu không chuẩn hoặc cấu trúc sai (ví dụ chứa #, ;, khoảng trắng) đã được đổi tên để đảm bảo nhất quán.
- **Chuẩn Hóa Nhãn và Biến Dạng Danh Mục**
  - Tất cả tên cột và biến dạng danh mục (categorical features) đã được chuẩn hóa:
- Viết thường toàn bộ.
- Loại bỏ dấu tiếng Việt.
- Thay thế ký tự đặc biệt như #, ;, khoảng trắng bằng _.
  - Các giá trị danh mục được ánh xạ thống nhất (dictionary mapping):

Ví dụ:

- gender: các biến thể "male", "Male", "M" đều được chuyển thành 0, "female", "F" → 1.

- occupation, needstate_group, companion_group được mã hóa bằng mã số hoặc enum để dễ xử lý mô hình.

- **Sửa Lỗi Logic Dữ Liệu:** Các lỗi logic trong dữ liệu đã được phát hiện và xử lý, bao gồm:

  - Các dòng có giá trị spending < 0, visit = 0 nhưng spending > 0, hoặc age < 0 đều được xử lý hoặc loại bỏ.

  - Kiểm tra chéo dữ liệu giữa group_size, year, age, đảm bảo không có sai lệch logic

(ví dụ: group_size không thể âm hoặc quá lớn).

- Loại bỏ các dòng trùng lặp trong bảng Brand_Image.csv với số lượng 13.974 dòng dư bị loại bỏ dựa theo khóa ID, Year, City, Attribute.

- **Điền Dữ Liệu Thiếu Một Cách Hợp Lý**

Việc xử lý giá trị thiếu được thực hiện dựa trên tính chất của từng biến:

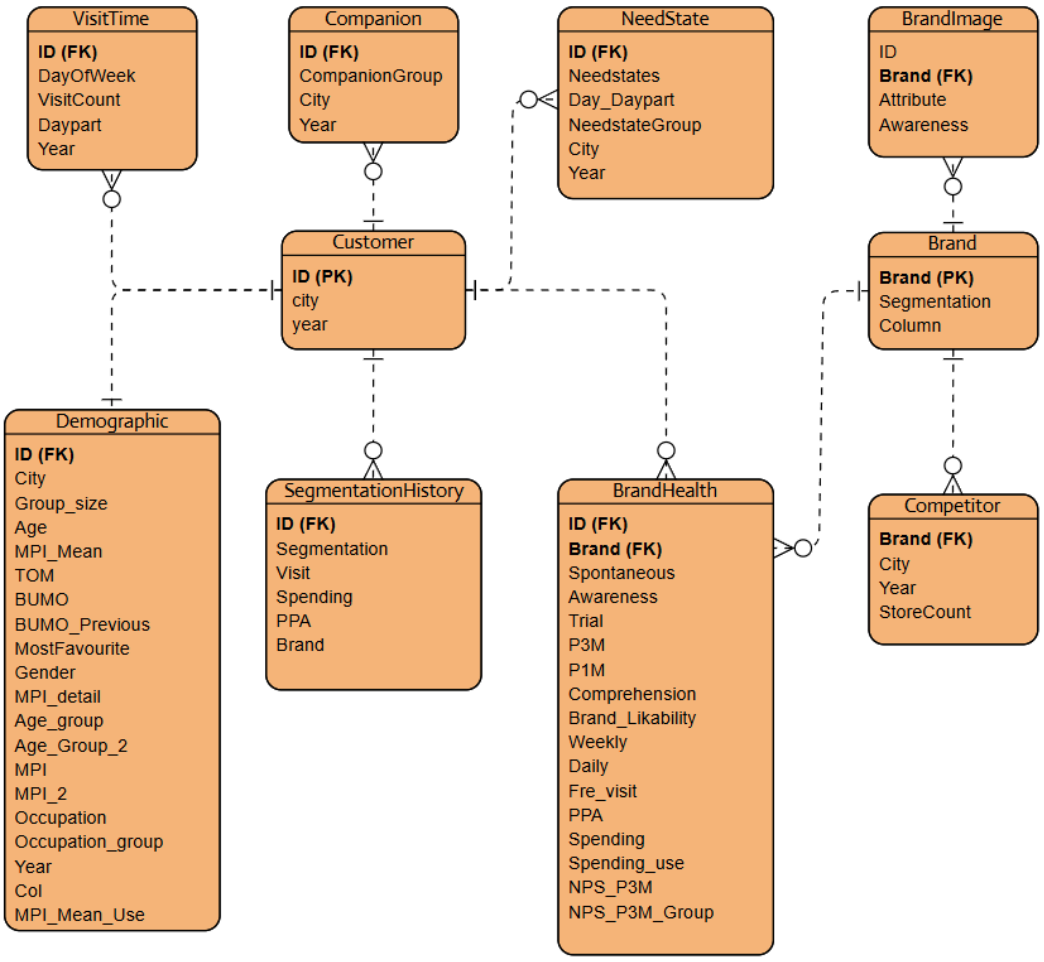
- Biến số (numerical): Sử dụng trung bình (mean) hoặc trung vị (median) của nhóm theo city hoặc segmentation để điền các trường như age, group_size, MPI#Mean.
- Biến danh mục (categorical): Áp dụng giá trị phổ biến nhất (mode) trong cùng nhóm để điền vào các trường như gender, occupation.
- Trường hợp thiếu quá nhiều hoặc không có quy luật rõ ràng: Các dòng này được loại bỏ nhằm tránh gây nhiễu cho mô hình phân tích sau này.

1.5. Kết Quả Sau Khi Làm Sạch và Chuẩn Hóa

Tên bảng dữ liệu	**Trạng thái xử lý**	**Bổ sung**
Brandhealth.csv	Đã chuẩn hóa định dạng và nhãn	Không có dòng trùng hoặc giá trị thiếu
Brand_Image.csv	Đã xóa dòng trùng và đồng bộ nhãn thuộc tính	Xóa 13.974 dòng trùng
Companion.csv	Gộp nhóm nhãn, xử lý dòng trùng	Gộp theo companion_group, xóa 799 dòng trùng
SA#var.csv	Điền thiếu hợp lý, chuẩn hóa biến	Xử lý biến MPI, age, group_size
Dayofweek.csv, Daypart.csv	Đã loại trùng và chuẩn hóa nhãn	Thống nhất định dạng daypart, weekday_end
NeedstateDayDaypart.csv	Chuẩn hóa nhóm nhu cầu	Dùng dictionary cho needstate_group

CÂU 2: SƠ ĐỒ MÔ HÌNH DỮ LIỆU

2.1. Sơ đồ quan hệ thực thể (ERD)



2.2. Giải thích mối quan hệ logic

Mối quan hệ	Kiểu	Giải thích nghiệp vụ
Customer → BrandHealth	1–N	Mỗi khách hàng có thể đánh giá nhiều thương hiệu
Customer → VisitTime	1–N	Mỗi khách hàng ghé nhiều lần vào nhiều ngày giờ khác nhau
Customer → Demographic	1–1	Mỗi khách hàng có một thông tin nhân khẩu học duy nhất
Customer → SegmentationHistory	1–N	Theo dõi sự thay đổi trong phân khúc theo thời gian

Mối quan hệ	Kiểu	Giải thích nghiệp vụ
Brand → BrandImage	1–N	Mỗi thương hiệu có thể được nhiều khách nhận diện khác nhau
Brand → Competitor	1–N	Mỗi thương hiệu có nhiều cửa hàng đối thủ trong khu vực
Customer → NeedState	1–N	Khách hàng đi ăn nhiều lần vào nhiều bối cảnh tâm lý tiêu dùng

CÂU 3: KIỂM SOÁT TRUY CẬP DỮ LIỆU THEO VAI TRÒ

3.1. Giải thích theo vai trò

- **BOD:** Cần quan sát toàn bộ dữ liệu tổng hợp KPI, xu hướng toàn doanh nghiệp → đọc toàn bộ.
- **HR:** Phân tích hiệu suất theo cửa hàng để quyết định thưởng phát/phúc lợi → truy cập SegmentationHistory, Demographic.
- **Finance:** Cần thông tin doanh thu, chi phí, segment → truy cập Customer, SegmentationHistory, Brand.
- **Marketing:** Quản lý định vị thương hiệu, KPI thị trường → truy cập Brand, BrandHealth, BrandImage, Demographic.
- **CRM Lead:** Tập trung theo dõi NPS, độ trung thành → cần BrandHealth, NeedState.
- **Sales Ops:** Cần phân tích theo khu vực, ngày, nhu cầu → NeedState, VisitTime, Demographic.
- **Regional Manager:** So sánh hiệu suất theo cửa hàng → truy cập BrandHealth, Segmentation, VisitTime.
- **Store Manager:** Cần biết doanh thu, ghé thăm, khách hàng → VisitTime, SegmentationHistory, Brand.

3.2. Ma trận phân quyền

Bảng dữ liệu	BOD	HR	Finance	Marketing	CRM Lead	Sales Ops	Regional Manager	Store Manager
Customer	R	R	R	R	R	R	R	R

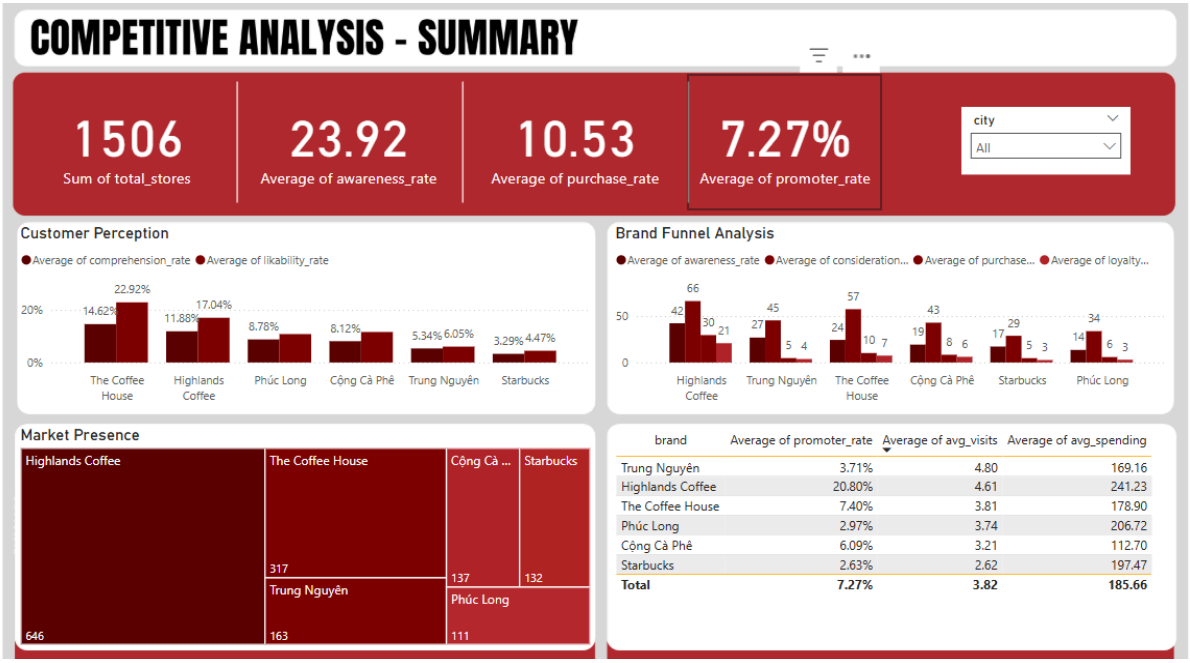
Bảng dữ liệu	BOD	HR	Finance	Marketing	CRM Lead	Sales Ops	Regional Manager	Store Manager
Brand	R	-	R	R/W	R	R	R	R
SegmentationHistory	R	R	R	R	-	R	R	R
BrandHealth	R	-	-	R/W	R/W	R	R	-
BrandImage	R	-	-	R/W	R	-	-	-
Competitor	R	-	-	R/W	-	-	R	-
VisitTime	R	-	-	R	-	R	R	R
Companion	R	-	-	R	-	-	-	-
NeedState	R	-	-	R/W	R	R	R	R
Demographic	R	R	-	R/W	-	R	R	R

CÂU 4: PHÂN TÍCH BỐI CẢNH CẠNH TRANH VÀ ĐỊNH VỊ THƯƠNG HIỆU CHO HIGHLAND COFFEE

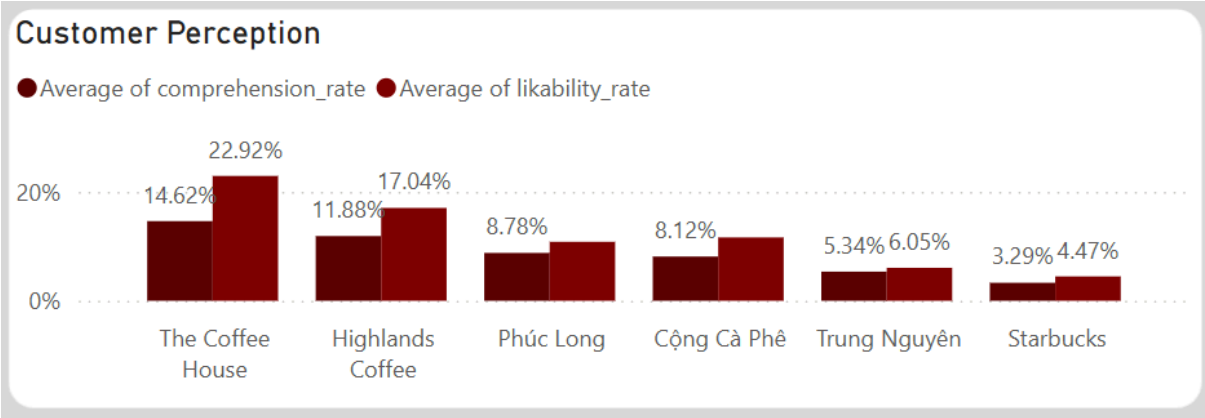
Link interactive dashboard:
https://drive.google.com/drive/folders/1Csc8aF8-dy4SjkVRHVkQb784VLGPkQSn?usp=drive_link

4.1. Tổng quan

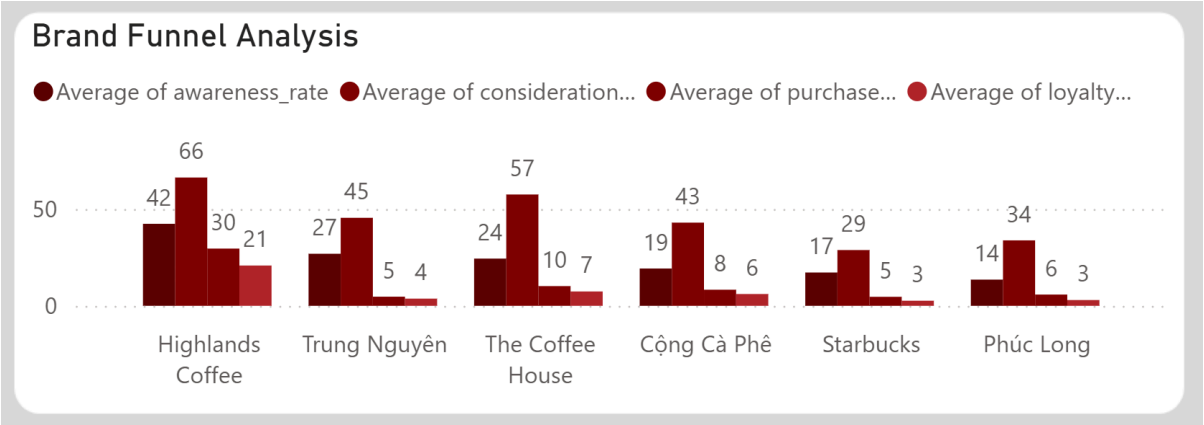
Báo cáo này tập trung phân tích kết quả kinh doanh của Highland Coffee dựa trên hai dashboard: (1) Competitive Analysis (Brand Funnel, Customer Perception, Market Presence so với đối thủ) và (2) Churn Analysis (tỷ lệ rời bỏ khách hàng). Định vị thương hiệu thường được minh họa bằng perceptual map – biểu đồ so sánh cảm nhận khách hàng theo hai thuộc tính quan trọng. Bên cạnh đó, brand funnel là mô hình hành trình khách hàng từ nhận biết (awareness) đến trung thành (loyalty) . Báo cáo sẽ lần lượt phân tích định vị và phễu thương hiệu (câu 4), so sánh SWOT với đối thủ, đưa ra đề xuất chiến lược; sau đó phân tích xu hướng churn (câu 5), tìm nguyên nhân gốc và đề xuất giải pháp.



4.2. Phân tích Brand Funnel và Định vị



Dashboard phân tích cạnh tranh cho thấy định vị thương hiệu của Highland Coffee so với các đối thủ qua hai thuộc tính "comprehension" và "likability". Trong biểu đồ Customer Perception , The Coffee House dẫn đầu với mức độ hiểu biết (comprehension) ~14.6% và độ thích (~likability) ~22.9%, Highland Coffee đứng thứ hai (11.9% và 17.0%). Các thương hiệu còn lại đều thấp hơn: Phúc Long (~8.8%, 10.8%), Cộng Cà Phê (~8.1%,11.6%), Trung Nguyên (~5.3%, 6.1%), Starbucks (~3.3%, 4.5%). Như vậy, Highland có vị trí cao trong tâm trí khách hàng so với hầu hết đối thủ trong nước, chỉ xếp sau The Coffee House về cả hai khía cạnh. Trên perceptual map, Highland Coffee nằm ở góc độ tương đối cao (awareness-và-thích) so với nhóm còn lại, cho thấy thương hiệu đã có độ nhận diện và thiện cảm khá tốt với thị trường



Về **brand funnel**, Highland Coffee có chỉ số **consideration cao nhất thị trường (66%)**, tiếp theo là The Coffee House (57%) và Trung Nguyên (45%). Các chỉ số **purchase (30%)** và **loyalty (21%)** của Highland cũng đứng đầu, vượt trội so với tất cả đối thủ. Highland thể hiện hiệu quả **chuyển đổi từ consideration → loyalty** cực kỳ tốt. Tính toán nhanh:

- Tỷ lệ chuyển đổi consideration → purchase: 30 / 66 ≈ **45.5%**
- Purchase → loyalty: 21 / 30 = **70%**
- Overall funnel: 21 / 42 = **50%** khách aware đã trở thành loyal customers

Điều này cho thấy: **Highland có độ trung thành thương hiệu rất cao**, và khách hàng sau khi cân nhắc thường đưa ra quyết định mua và gắn bó lâu dài hơn hẳn các đối thủ. Trong khi đó, các đối thủ như Starbucks, Phúc Long, Cộng Cà Phê có funnel kém hiệu quả: chỉ số consideration cao nhưng loyalty thấp.

⇒ Insight: Highland đang có lợi thế mạnh về giữ chân khách và hành trình trải nghiệm khách hàng (funnel chất lượng), tuy nhiên cần tăng awareness (độ nhận diện) để mở rộng lượng khách đầu phễu.

4.3. SWOT cạnh tranh

SWOT là công cụ phân tích chiến lược giúp đánh giá vị thế cạnh tranh của thương hiệu bằng cách liệt kê Điểm mạnh – Điểm yếu – Cơ hội – Rủi ro. Dựa trên số liệu ở trên, tóm tắt SWOT của Highland Coffee so với đối thủ như sau:

Strengths	**Weaknesses**
- Chuyển đổi Brand Funnel xuất sắc: Highland Coffee có chỉ số consideration (66%), purchase (30%), và loyalty (21%)	**-Cạnh tranh cảm nhận:** Trong bản đồ cảm nhận thương hiệu (customer perception), Highland đứng sau The

cao nhất trong tất cả thương hiệu. Tỷ lệ chuyển đổi qua từng giai đoạn rất tốt (ví dụ: loyalty/awareness ≈ 50%). Điều này cho thấy Highland giữ chân khách hàng rất hiệu quả và có hành trình khách hàng rõ ràng, gắn bó. **- Số điểm bán lớn nhất:** Highland có 646 cửa hàng – vượt trội so với The Coffee House (317) và các đối thủ khác. Điều này tạo lợi thế vùng phủ thị trường. **- Tỉ lệ Promoter (ủng hộ) cao nhất:** Promoter_rate của Highland 20.8% (nhiều gấp đôi mặt bằng trung bình 7.3%). Khách hàng hiện tại đánh giá tích cực, cho thấy độ hài lòng tương đối cao. **- Doanh thu/trung thành tiềm năng:** Highland có mức chi tiêu trung bình (avg_spending ~241) cao nhất trong nhóm, và tần suất mua lặp lại (avg_visits) cũng cao (4.61). Điều này cho thấy lượng khách thường xuyên chi tiêu lớn.	Coffee House về cả comprehension và likability. Điều này cho thấy thương hiệu chưa phải là "được yêu thích nhất" trong tâm trí khách hàng, dù hành vi mua đã rất tốt. **- Chênh lệch giữa awareness và consideration:** Sự khác biệt lớn (66% consideration vs. 42% awareness) cho thấy có thể khách biết thương hiệu thông qua trải nghiệm thực tế hơn là qua truyền thông đại chúng. Điều này là cơ hội nhưng cũng là hạn chế nếu thương hiệu muốn mở rộng quy mô nhanh.
Opportunities (Cơ hội): **- Mở rộng nhận diện thông qua digital marketing:** Với funnel chuyển đổi mạnh như hiện tại, chỉ cần tăng awareness là lượng khách hàng trung thành sẽ tăng theo. Đầu tư vào social media, influencer marketing và quảng cáo tại điểm bán có thể **đưa khách vào phễu nhiều hơn**. - **Tăng trưởng tại các khu vực chưa khai thác:** Với hệ thống cửa hàng mạnh, Highland có thể tiếp tục **định vị là thương hiệu phủ rộng nhất Việt Nam**, đặc biệt ở các tỉnh thành mới nổi hoặc	**Threats (Thách thức/Cạnh tranh):** **- Đối thủ mạnh:** The Coffee House có mức Perception mạnh hơn ở comprehension và likability. Trung Nguyên đề cao văn hóa và thu hút lượng khách hàng visits cao. Starbucks tuy số cửa hàng ít nhưng thương hiệu toàn cầu mạnh và có tiềm năng lan rộng. - **Thị trường cạnh tranh cao:** Rào cản vào thị trường thấp, nhiều chuỗi nhỏ lẻ và quán cà phê mới liên tục xuất hiện. - **Xu hướng tiêu dùng thay đổi:** Thị

khu dân cư mới. - **Tăng cường truyền thông thương hiệu cảm xúc:** Để cải thiện điểm likability và tạo sự gắn kết sâu sắc hơn với khách hàng, Highland có thể đầu tư storytelling, cải thiện không gian quán và trải nghiệm cảm xúc.	hiếu khách hàng nhanh chóng thay đổi; nếu Highland không đổi mới nhanh (sản phẩm mới, ứng dụng công nghệ…) sẽ mất thị phần vào tay đối thủ linh hoạt hơn. **- Nguy cơ bị sao chép mô hình:** Funnel hiệu quả, mô hình cửa hàng tốt của Highland rất dễ bị các thương hiệu nhỏ hoặc startup F&B học theo, dẫn đến mất đi sự độc đáo. **- Tăng trưởng chậm nếu không cải thiện awareness:** Dù có hệ thống tốt, nếu không tăng awareness thì số lượng người bước vào phễu không tăng, **giới hạn quy mô tăng trưởng** trong dài hạn.

(Đánh giá SWOT trên tổng hợp dữ liệu nội bộ và so sánh với đối thủ theo khuyến cáo của phương pháp SWOT.)

4.4. Đề xuất chiến lược thương hiệu

Dựa trên phân tích trên, Highland Coffee nên tập trung các chiến lược sau:

- **Mở rộng vùng phủ và kênh phân phối:** Tiếp tục mở thêm cửa hàng tại khu vực chưa phủ sóng, đặc biệt khu vực ven đô và các tỉnh thành mới nổi. Đẩy mạnh kênh bán hàng online và giao hàng (delivery) để tăng độ tiếp cận khách hàng.
- **Tăng cường truyền thông – nhận diện:** Thực hiện các chương trình quảng bá thương hiệu đồng nhất trên mạng xã hội, hợp tác với KOLs/ảnh hưởng; cải thiện bộ nhận diện (logo, không gian quán) nhằm làm mới hình ảnh. Đầu tư marketing để nâng awareness, hướng đến người tiêu dùng mục tiêu (ví dụ nhóm trẻ, nhân viên văn phòng).
- **Nâng cao trải nghiệm khách hàng:** Nghiên cứu thiết kế cửa hàng thân thiện hơn với nhu cầu "thư giãn" (hãy chú ý nhóm khách hàng có mục đích mua cà phê để thư giãn có churn cao). Đào tạo nhân viên nâng cao chất lượng phục vụ, giảm thời gian chờ. Xây dựng chương trình khách hàng thân thiết (loyalty program) hấp dẫn hơn, ví dụ tích điểm đổi quà, tặng thức uống miễn phí cho nhóm khách tiềm năng.
- **Cá nhân hóa ưu đãi và sản phẩm:** Sử dụng dữ liệu CRM để gửi khuyến mãi phù hợp theo nhóm khách (ví dụ khuyến mãi cho khách nam). Đổi mới thực đơn theo

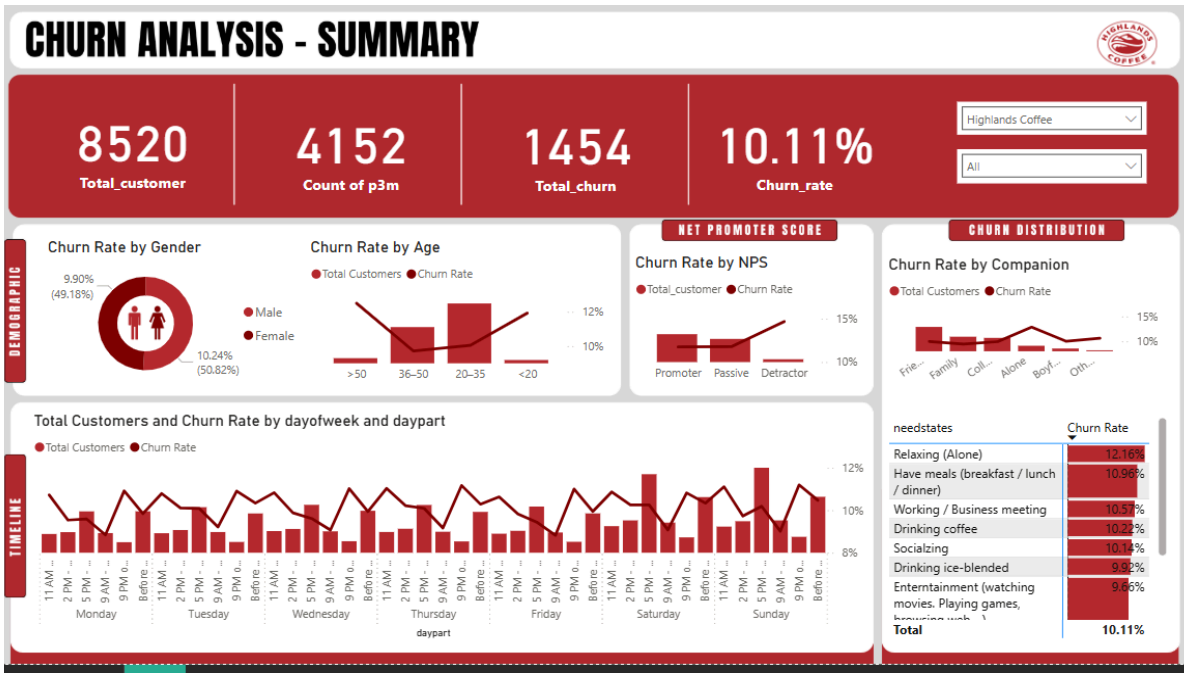
mùa, chú trọng đồ uống độc quyền để tạo điểm khác biệt.

- **Liên tục đo lường và điều chỉnh:** Ứng dụng hệ thống BI để theo dõi sát sao các chỉ số funnel và churn, từ đó điều chỉnh linh hoạt chiến lược marketing và trải nghiệm. Đánh giá hiệu quả theo chu kỳ để kịp thời điều chỉnh.

Những giải pháp trên hướng tới nâng cao nhận diện thương hiệu (tăng awareness, likability) và cải thiện trải nghiệm nhằm tăng tỉ lệ chuyển đổi purchase và loyalty. Đồng thời, giữ chân khách hàng hiện hữu bằng chương trình cá nhân hóa khuyến mãi sẽ giúp giảm thiểu churn – vốn tốn gấp nhiều lần chi phí để thu hút khách mới.

CÂU 5: CUSTOMER CHURN ANALYSIS AND DASHBOARD FOR HIGHLAND COFFEE

Link interactive dashboard:
https://drive.google.com/drive/folders/1Csc8aF8-dy4SjkVRHVkQb784VLGPkQSn?usp=drive_link



5.1. Phân tích rời bỏ khách hàng (Churn)

Dashboard Churn Analysis cho Highland Coffee cho thấy tổng số khách 8.520, trong đó có 1.454 khách rời bỏ, tương ứng tỷ lệ churn chung ~10,11%. Phân tích chi tiết:

- **Giới tính:** Nam có tỉ lệ churn cao hơn một chút (10,24%) so với Nữ (9,90%). Sự chênh lệch nhỏ cho thấy xu hướng rời bỏ ở nam có thể cao hơn một chút, nhưng không khác biệt quá lớn. Nhiều nghiên cứu chỉ ra gender thường không phải yếu tố ảnh hưởng mạnh đến churn.

- **Độ tuổi:** Nhóm 20–35 tuổi chiếm đa số khách hàng và có tỉ lệ churn cao 10,0%). Nhóm trên 50 tuổi tuy ít khách hơn nhưng có churn cao nhất (12,50%). Nhóm 36–50 tuổi có churn thấp nhất (9,71%). Điều này gợi ý Highland mất nhiều khách trẻ và lớn tuổi, trong khi nhóm trung niên rời bỏ ít hơn.
- Nhóm Detractor (những khách hàng không hài lòng, sẵn sàng nói xấu thương hiệu) có tỷ lệ churn cao nhất (14,67%), dù chỉ chiếm số lượng nhỏ trong tổng khách hàng (0.2K). Trong khi đó, nhóm Passive có churn rate ~11.77%, và nhóm Promoter thấp nhất với 11.74%. Điều này cho thấy khách hàng không hài lòng rời bỏ nhanh chóng, nhưng ngay cả nhóm trung lập (Passive) cũng có nguy cơ cao. Highland cần tập trung vào cả hai nhóm này: Detractor để ngăn chặn ảnh hưởng tiêu cực, và Passive để tránh mất khách âm thầm.
- Bạn đồng hành (Companion): Khách tới quán một mình có churn cao nhất (~12%), trong khi đi cùng bạn bè (~9,98 %) hay gia đình (~9,46%) ít rời bỏ hơn. Điều này cho thấy nhóm khách đi một mình (có thể là mua giải trí, thư giãn) có nguy cơ bỏ ra cao nhất.
- Need state: Mục đích "Thư giãn, ngồi một mình" ghi nhận churn ~12,16%, cao nhất trong các mục đích. Các mục đích khác (ăn uống, làm việc/họp, giao lưu xã hội, v.v.) có churn khoảng 9–11%. Nhóm chỉ uống cà phê hay giải trí có tỉ lệ churn thấp hơn. Như vậy, nhu cầu "thư giãn" mà Highland chưa đáp ứng tốt dẫn đến mất khách.

Tóm lại, dashboard chỉ ra Highland đang đối mặt với nguy cơ churn cao nhất ở các nhóm nam, trên 50 hoặc dưới 20 tuổi, nhóm detractor và những ai tới quán để thư giãn một mình. Những insight này gợi ý Highland cần cải thiện trải nghiệm cho nhóm khách hàng cô đơn/phục vụ nhu cầu thư giãn, đồng thời tập trung giữ chân nhóm NPS detractor bằng truyền thông, văn hóa, chất lượng để tránh mất họ về tay đối thủ.

5.2. Nguyên nhân gốc và giải pháp giảm churn

- **Nguyên nhân gốc (Root causes):** Theo các chuyên gia phân tích, churn thường xuất phát từ dịch vụ không tốt, sản phẩm chưa đáp ứng kỳ vọng hoặc trải nghiệm khách hàng kém. Đối với Highland, có thể kể đến một số lý do:
  - **Chất lượng dịch vụ:** Thời gian phục vụ lâu, dịch vụ chưa thân thiện hoặc thiếu nhất quán. Khách đi một mình (thư giãn) nhưng không có phục vụ tốt thì dễ thất vọng.
  - **Sản phẩm/dịch vụ không khác biệt:** Đồ uống đôi khi thiếu sáng tạo, giá cao trong khi trải nghiệm chưa đột phá. Đối thủ như Cộng Ca Phê có không gian độc

đáo, The Coffee House thường xuyên cập nhật menu mới.

- **Chương trình khách hàng thân thiết yếu kém:** Nếu Highland chưa có chính sách ưu đãi cá nhân hóa hấp dẫn, khách hàng sẽ thiếu lý do để gắn bó lâu dài. Đặc biệt, nhóm **Detractor** – vốn có tỷ lệ churn cao nhất– thường rời bỏ nhanh chóng nếu trải nghiệm không tốt hoặc không được lắng nghe. Đồng thời, nhóm **Passive** cũng có churn rate cao ~11.77% và rất nhạy cảm với giá cả, trải nghiệm. Cả hai nhóm này cần được chăm sóc kỹ lưỡng thông qua các ưu đãi cá nhân hóa, chương trình phản hồi nhanh và dịch vụ vượt mong đợi để hạn chế rủi ro rời bỏ.
- **Thiếu tương tác và phản hồi:** Khách không được tạo cơ hội phản hồi, khi có trải nghiệm không tốt cũng không có kênh cải thiện kịp thời.

- **Hệ quả kinh doanh:** Churn cao ở nhóm 20–35 tuổi và nhóm khách thư giãn một mình (chiếm phần lớn khách của Highland) có thể dẫn đến mất nguồn doanh thu trung dài. Theo Forrester, chi phí thu hút khách mới gấp ~5 lần chi phí giữ chân khách cũ. Nếu Highland để mất nhóm khách trung thành (20% khách tạo 80% doanh thu), doanh thu tương lai sẽ suy giảm mạnh. Đặc biệt, nhóm khách Passives chiếm tỉ lệ lớn trong cơ sở khách hàng hiện tại lại dễ chuyển sang đối thủ nếu không được chăm sóc – "passives bleeds out your revenue slowly" như nghiên cứu chỉ ra. Rủi ro là Highland sẽ mất dần thị phần vào tay các chuỗi có trải nghiệm tốt hơn hoặc giá cả cạnh tranh hơn nếu không kịp thời xử lý churn.
- **Giải pháp giảm churn (Retention strategies):**
  - **Chăm sóc nhóm NPS thụ động:** Rất nhiều nguồn khuyên nên ưu tiên nhóm Passive – vì dễ bị mất nhất nhưng cũng dễ thu hút lại nhất. Highland có thể thực hiện các chương trình như giảm giá riêng, tặng quà sinh nhật, ưu đãi theo lịch sử mua hàng cho nhóm này. Đồng thời, hãy yêu cầu phản hồi của họ thường xuyên (qua khảo sát, hotline) để cải thiện nhanh chóng những điểm chưa vừa lòng.
  - **Tăng cường trải nghiệm khách hàng:** Đào tạo nhân viên niềm nở, giảm thời gian chờ, cải thiện không gian quán (nhạc, WiFi, chỗ ngồi thoải mái) – nhất là vào khung giờ và mục đích thư giãn. Ví dụ, cung cấp góc riêng tư, xây dựng menu thức uống thư giãn (cà phê thơm ngon, không gian yên tĩnh) để giữ chân khách đi một mình. Qualtrics nhấn mạnh rằng giải quyết các điểm đau của khách và nâng cao CX sẽ giảm churn, đồng thời giúp thu hút khách mới dễ dàng hơn
  - **Chương trình Loyalty và khuyến mãi cá nhân hóa:** Thực thi chương trình tích điểm, tặng voucher cho lần mua tiếp theo, khuyến mãi nhân dịp sinh nhật… Qua đó, khách có lý do để quay lại. Theo khảo sát, những điều khách hàng trân trọng bao gồm ưu đãi bất ngờ (surprise offers) và trải nghiệm mua sắm thuận tiện. Việc cá nhân hóa ưu đãi dựa trên lịch sử mua sắm hoặc nhóm NPS sẽ làm tăng cảm

giác được quan tâm và gắn bó.

- **Giao tiếp chủ động:** Sử dụng email/SMS marketing để nhắc khuyến mãi đặc biệt, hoặc giới thiệu sản phẩm mới tới khách hàng từng bỏ lâu ngày. Chú ý nhóm khách lâu không mua (có thể thuộc nhóm Passive đang có nguy cơ) – gửi cho họ ưu đãi khuyến khích quay lại.
- **Liên tục đo lường và cải tiến:** Theo ThoughtSpot, sau khi đã phát hiện nguyên nhân, bước tiếp theo là "hành động để giảm churn" bằng các giải pháp nhắm đúng đối tượng. Highland cần thường xuyên giám sát chỉ số churn theo từng nhóm và đánh giá hiệu quả các biện pháp giữ chân.

Tóm lại, Highland Coffee nên đẩy mạnh chiến lược giữ chân khách hiện tại – nhất là nhóm khách có khả năng rời bỏ cao – trước khi tập trung thu hút khách mới. Bằng cách cải thiện trải nghiệm tổng thể và cá nhân hóa khuyến mãi cho đúng đối tượng (đặc biệt là nhóm NPS Passive), Highland có thể giảm thiểu tổn thất khách hàng giá trị và nâng cao doanh thu bền vững.

CÂU 6: PHÂN KHÚC KHÁCH HÀNG VÀ DỰ ĐOÁN RỜI BỎ

1. Dữ liệu đã tiền xử lý

- **Nguồn dữ liệu:**
  Dữ liệu gốc được thu thập từ nhiều file CSV khác nhau Brandhealth.csv, Brand_Image.csv, SA#var.csv, NeedstateDayDaypart.csv, v.v.
  Trong file code:
  - Đã tiến hành **làm sạch dữ liệu** cho toàn bộ file CSV trong thư mục gốc.
  - Các bước bao gồm: loại bỏ dòng trùng lặp, chuẩn hóa định dạng, xử lý dữ liệu thiếu và xuất báo cáo chất lượng dữ liệu cho từng bảng.
- **Các biến sử dụng cho modeling:**
  - **Nhân khẩu học:**
    - Age#group, Age#Group#2
    - Gender
    - MPI (thu nhập bình quân)
    - Occupation#group
    - City
  - **Hành vi:**
    - Segmentation
    - P3M, P1M
    - Weekly, Daily
    - Needstates

        - Daypart
        - NPS#P3M
    - **Các biến khác:**
        - Brand_Likability
        - Awareness, Trial
        - BrandImage
        - Visit
- **Tiền xử lý:**
    - **Xử lý trùng lặp:** Dùng df.drop_duplicates() để loại bỏ các dòng lặp.
    - **Xử lý giá trị thiếu:**
        - Loại bỏ dòng có số lượng null vượt quá 50% số cột.
        - Phần còn lại:
            - Biến **object** được điền bằng 'Unknown'
            - Biến **số** được điền bằng giá trị trung vị (median)
    - **Mã hóa biến phân loại:** LabelEncoder được sử dụng cho các biến phân loại như Gender, City, Occupation#group, Segmentation, v.v.
    - **Chuẩn hóa dữ liệu:** Áp dụng StandardScaler cho toàn bộ biến số để đảm bảo các biến có cùng thang đo.
    - **Gộp dữ liệu:** Sử dụng ID làm khóa chính để nối dữ liệu từ nhiều bảng khác nhau trước khi modeling.
- **File dữ liệu đã clean:** data_clean

2. Mô hình phân khúc khách hàng

A . Mô hình clustering

- **Phương pháp:**
  Dữ liệu sau khi được làm sạch và chuẩn hóa được áp dụng thuật toán **KMeans clustering** để phân nhóm khách hàng.
  Số cụm được thử nghiệm từ 2 đến 6, chọn **số cụm tối ưu = 3** dựa trên:
    - Phân tích biểu đồ Elbow (điểm gấp khúc).
    - Đánh giá Silhouette Score.
- **Các biến đầu vào cho clustering:**
    - Biến nhân khẩu học: Age#group, Gender, City, Occupation#group, MPI.
    - Biến hành vi: Segmentation, P3M, P1M, Needstates, Daypart, Visit,

Weekly, Daily.

- Các biến về thương hiệu: Brand_Likability, BrandImage, Awareness, Trial, NPS#P3M.

- **Kết quả clustering:**
  Mô hình chia dữ liệu thành **3 nhóm khách hàng đặc trưng**, thể hiện khác biệt rõ rệt về hành vi và đặc điểm nhân khẩu học.

- **File code:** Code

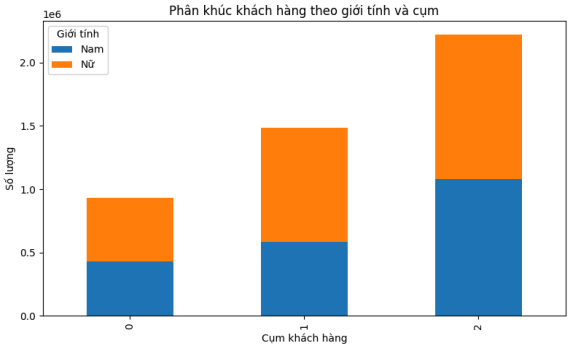
B . Phân tích đặc điểm từng nhóm

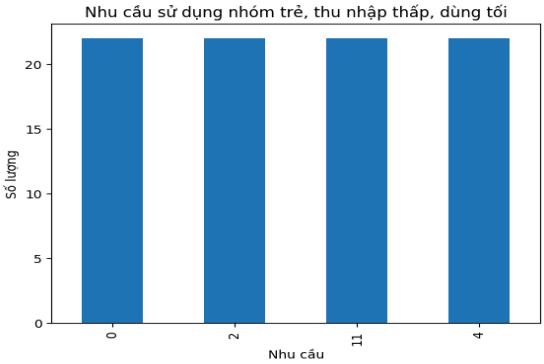
Dưới đây là đặc điểm nổi bật của từng cụm khách hàng:

- **Nhóm 0:**
  - **Đặc điểm nhân khẩu học:** Trẻ tuổi, thu nhập thấp, thường sử dụng đồ uống lạnh vào buổi tối.
  - **Hành vi:** Ưa thích đồ uống lạnh, sử dụng vào buổi tối và cuối tuần.
  - **Tần suất:** Thấp, không thường xuyên, có xu hướng thay đổi thương hiệu.
- **Nhóm 1:**
  - **Đặc điểm nhân khẩu học:** Nhân viên văn phòng, thu nhập trung bình, thích cafe nóng vào buổi sáng sớm.
  - **Hành vi:** Thích cà phê nóng, sử dụng buổi sáng, ưu tiên không gian yên tĩnh, có tính lặp lại hành vi cao.
  - **Tần suất:** Ổn định, đều đặn mỗi ngày trong tuần.
- **Nhóm 2:**
  - **Đặc điểm nhân khẩu học:** Nữ giới, thu nhập cao, cư trú tại các thành phố lớn.
  - **Hành vi:** Thích uống cà phê tại cửa hàng có không gian đẹp, đi cùng bạn bè vào buổi tối.
  - **Tần suất:** Không thường xuyên nhưng chi tiêu cao, đề cao trải nghiệm thương hiệu.

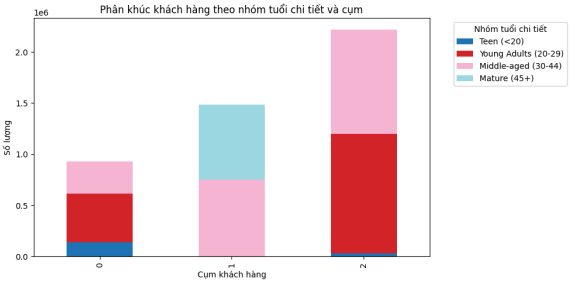
C. Trực quan hóa:

- Biểu đồ theo giới tính

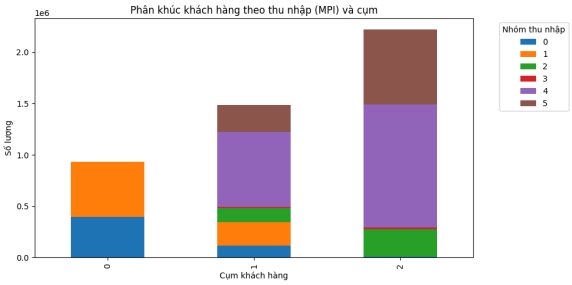




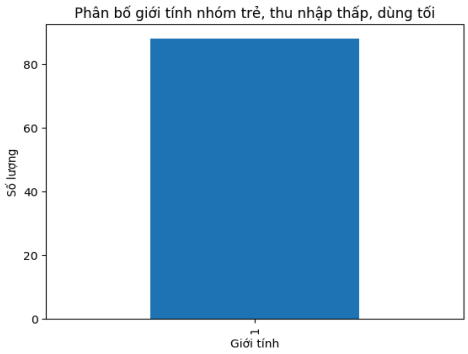
- Biểu đồ theo tuổi



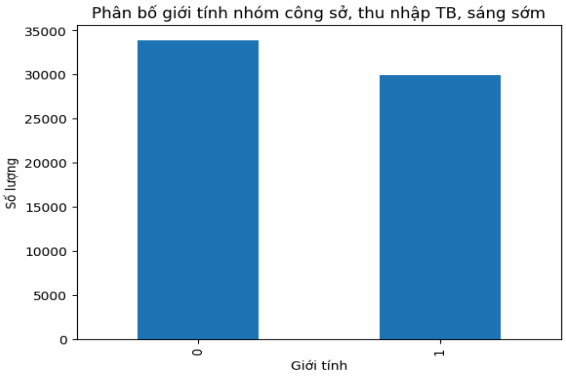
- Biểu đồ theo thu nhập

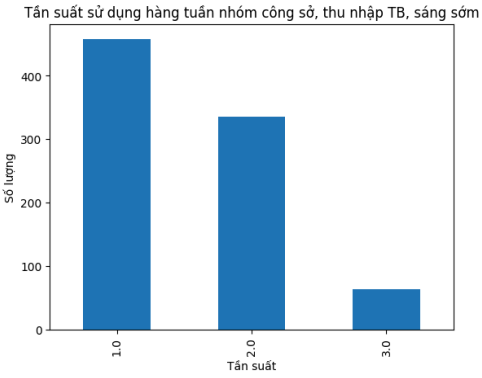


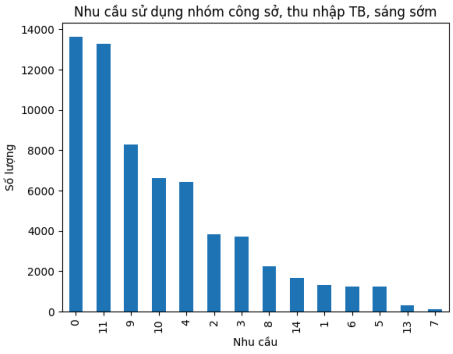
- Biểu đồ phân bố nhóm trẻ, thu nhập thấp, thường sử dụng đồ uống lạnh vào buổi tối.



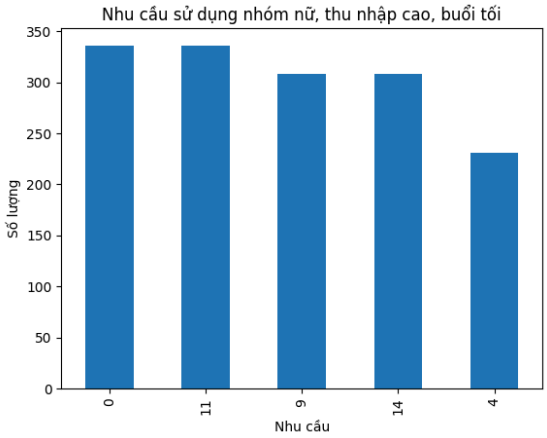
- Biểu đồ nhóm công sở, thu nhập trung bình, thích cafe nóng sáng sớm



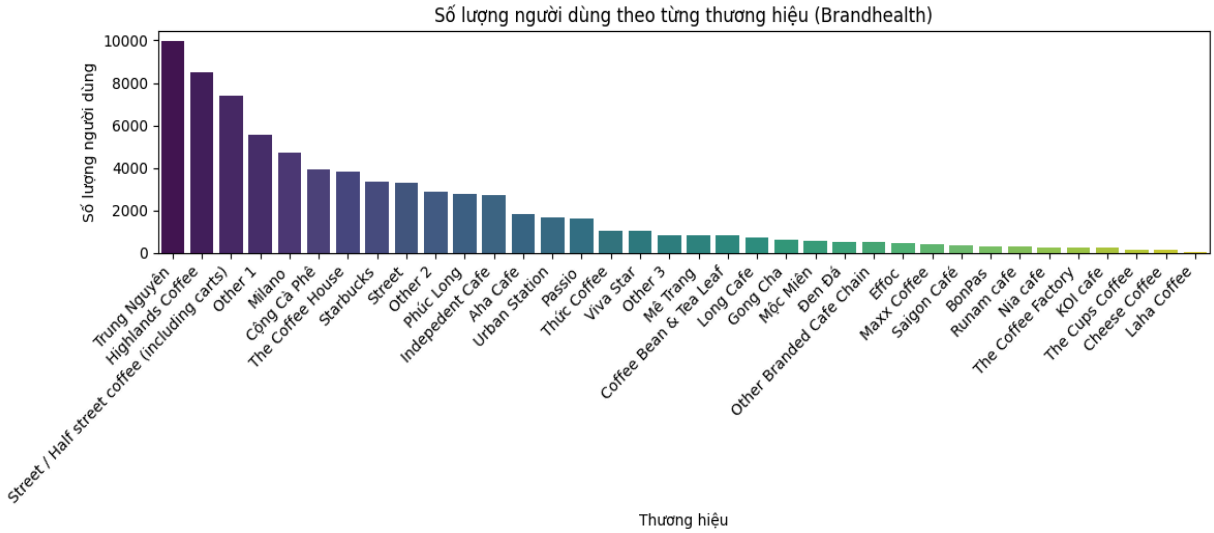




- Biểu đồ nhóm nữ giới thu nhập cao, uống cafe buổi tối tại cửa hàng có không gian đẹp



- Biểu đồ so sánh các thương hiệu



3. Mô hình dự đoán churn

a. Gán nhãn churn

- **Tiêu chí gán churn (churn = 1):**
  - NPS#P3M < 0 → khách không hài lòng.
  - Awareness > 80, Trial < 30, P1M < 20 → biết thương hiệu nhưng không dùng gần đây.
  - Visit = 0 hoặc (P1M = 0 và P3M > 0) → đã từng sử dụng nhưng dừng lại. Needstates thuộc nhóm "relax" hoặc "chitchat" → khách không gắn bó mục tiêu.

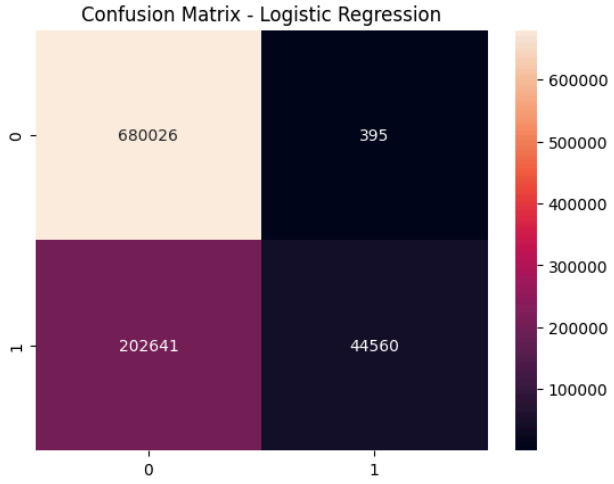
- Brand_Likability < 50 và BrandImage < 50.

- **Ghi nhãn churn:** Nếu khách hàng thỏa bất kỳ điều kiện nào ở trên, gán churn = 1, ngược lại churn = 0.

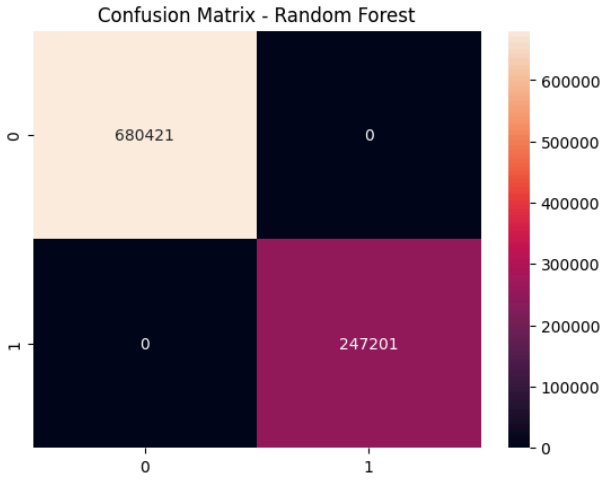
b. Hiệu quả mô hình

- **Các mô hình sử dụng:**
  - Logistic Regression – baseline.
  - Random Forest – mô hình học sâu cây quyết định.
  - XGBoost – mô hình boosting mạnh, cho kết quả tốt nhất.
- **Kết quả đánh giá:**
  - Accuracy, Precision, Recall, F1-score
  - Confusion matrix trực quan
  - So sánh hiệu quả các mô hình trên tập test
- **Các chỉ số đánh giá**

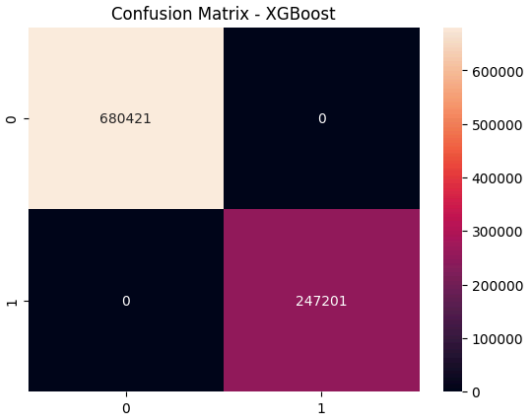
```
=== Logistic Regression ===
              precision    recall  f1-score   support

           0       0.77      1.00      0.87    680421
           1       0.99      0.18      0.31    247201

    accuracy                           0.78    927622
   macro avg       0.88      0.59      0.59    927622
weighted avg       0.83      0.78      0.72    927622

              precision    recall  f1-score   support

           0       0.77      1.00      0.87    680421
           1       0.99      0.18      0.31    247201

    accuracy                           0.78    927622
   macro avg       0.88      0.59      0.59    927622
weighted avg       0.83      0.78      0.72    927622
```



```
=== Random Forest ===
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    680421
           1       1.00      1.00      1.00    247201

    accuracy                           1.00    927622
   macro avg       1.00      1.00      1.00    927622
weighted avg       1.00      1.00      1.00    927622

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    680421
           1       1.00      1.00      1.00    247201

    accuracy                           1.00    927622
   macro avg       1.00      1.00      1.00    927622
weighted avg       1.00      1.00      1.00    927622
```




```
 bst.update(dtrain, iteration=i, fobj=obj)
=== XGBoost ===
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    680421
           1       1.00      1.00      1.00    247201

    accuracy                           1.00    927622
   macro avg       1.00      1.00      1.00    927622
weighted avg       1.00      1.00      1.00    927622
```




```
Bảng so sánh các chỉ số đánh giá mô hình:
                 Mô hình  Accuracy  Precision    Recall  F1-score
0  Logistic Regression  0.781122   0.991213  0.180258  0.305043
1        Random Forest  1.000000   1.000000  1.000000  1.000000
2              XGBoost  1.000000   1.000000  1.000000  1.000000
```

4. Đánh giá giới hạn và hướng cải tiến

a. Giới hạn dữ liệu, bias

- Dữ liệu chưa cân bằng hoàn toàn giữa nhóm churn và không churn → dễ gây bias cho mô hình classification.
- Một số cột (ví dụ: NPS, Trial) có tỷ lệ thiếu cao, nên việc nội suy có thể gây sai lệch.
- Dữ liệu từ các bảng khác nhau có thể không đồng nhất về ID, Year, gây lỗi trong join.

b. Đề xuất cải thiện tương lai

- **Về dữ liệu:**
    - Bổ sung thêm dữ liệu lịch sử theo chuỗi thời gian (visit theo ngày/tuần).
    - Tích hợp thêm dữ liệu giao dịch thực tế (order, hóa đơn, loyalty).
- **Về mô hình:**
    - Thử nghiệm mô hình clustering khác: DBSCAN, Hierarchical Clustering để xử lý trường hợp không hình cầu.
    - Áp dụng kỹ thuật xử lý dữ liệu mất cân bằng như SMOTE, class_weight.
    - Tăng cường feature engineering: tạo các biến tương tác như spending per visit, NPS trend, delta visit.