Câu 1:

- Chi tiết các lỗi trong data set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Các lỗi trong data set** | | | | |
| **Lỗi** | **Table** | **Chi tiết lỗi** | **Phương pháp xác định, xử lí** | **Phân tích nguyên nhân** |
| **Duplicate** | Brand\_Image | 13974 dòng duplicate | Dùng .duplicated().sum để phát hiện số lỗi và .drop\_duplicates() để loại bỏ trùng lắp | Do ghi nhận lặp phản hồi khảo sát hoặc người dùng thực hiện khảo sát nhiều lần |
| Companion | 799 dòng duplicate | Dữ liệu từ hệ thống phụ gửi về nhiều bản ghi giống nhau cho cùng một ID |
| Dayofweek | 37 dòng duplicate | Lỗi ghi log hoặc tổng hợp dữ liệu từ nhiều nguồn không lọc trùng |
| NeedstateDayDaypart | 72 dòng duplicate | Người dùng lặp lại hành vi trong nhiều phiên, hoặc dữ liệu sinh ra do lỗi ETL |
| **Redunancy** | SA#var | **Age#group** và **Age#group2**: đều dựa vào age để phân nhóm nhưng Age#group2 chi tiết hơn. | - Dựa vào “BI9\_Data Dictionary Round 01” để xác định ý nghĩa, unique() để kiểm tra thông tin  - Xử lí bằng drop column Age#group | Sinh biến phụ để phục vụ phân tích |
| **Col**: Không có ý nghĩa phân tích | - Dựa vào “BI9\_Data Dictionary Round 01”  - Xử lí bằng drop column | Feature thêm vào trong quá trình thu thập dữ liệu |
| **MPI#detail, MPI#2, MPI**: Trùng nội dung thu nhập | - Dựa vào “BI9\_Data Dictionary Round 01” để xác định ý nghĩa, unique() để kiểm tra thông tin từng features  - Xử lí bằng giữ MPI#2 giống MPI nhưng có thêm prefix số thứ tự, dễ sort nhóm | Dữ liệu lặp từ nhiều nguồn / Trình bày khác nhau |
| **MPI#Mean** giống **MPI#Mean\_use** do mục đích reporting | - Dựa vào “BI9\_Data Dictionary Round 01” để xác định ý nghĩa, với từng record so sáng giá trị 2 features (trừ những features cả 2 đều missing) dùng .compare()  - Loại bỏ MPI#Mean\_use | Tạo ra bản copy cho mục đích reporting |
| Brand\_Image, Brand\_Health, Companion,  Dayofweek, Daypart, NeedstateDayDaypart đều có chứa **ID**, **city, year** lặp lại hoàn toàn từ thông tin trong SA#var | | - Dựa vào **“BI9\_Data Dictionary Round 01”** để xác định ý nghĩa  - Kiểm tra bằng ánh xạ ID → (City, Year) và đối chiếu với từng bảng  - Xử lí bằng cách loại bỏ các cột chỉ giữu lại ở sheet gốc SA#var | Đây là dữ liệu dư thừa do các bảng phụ liên kết với bảng chính qua khóa ID |
| **Labeling Inconsistencies** | Brand\_Image | **Awareness, Brand Image**:  - 'Street / Half street coffee (including carts)', 'Street';  - 'Other Branded Cafe Chain', 'Other 1', 'Other 2', 'Other 3', 'Other'  - ‘Indepedent’ sai chính tả | - Dùng df[col].unique() để nhận diện  - Tạo ánh xạ chuẩn và dùng .str.lower().str.strip().replace(brand\_replace\_map) | Không chuẩn hóa từ đầu, nhập tay khác nhau |
| **Attribute:**  **-** Good place for working / business meeting, Good place for working / studying;  - Good place for socializing with friends, Good place for socializing with colleagues, Good place for socializing with family, Good place for socializing;  - Nice environment design, Comfortable and relaxing environment;  - Good coffee taste, Good tea taste, Good ice-blended taste, Good other beverages (other than coffee, tea & ice-blended);  - Delicious food, Good coffee/tea/ice-blended taste. | - Dùng df[col].unique() để nhận diện  - Nhóm các mô tả tương tự và xem độ trùng nội dung và ánh xạ về dạng chuẩn | Diễn đạt linh hoạt, không có danh mục cố định |
| Brand Health | **'Brand', 'Spontaneous', 'Awareness', 'Trial', 'P3M', 'P1M', 'Weekly':**  - 'Street / Half street coffee (including carts)', 'Street';  - 'Other Branded Cafe Chain', 'Other 1', 'Other 2', 'Other 3', 'Other'  ‘Indepedent’ sai chính tả | - Dùng df[col].unique() để nhận diện  - Tạo ánh xạ chuẩn và dùng .str.lower().str.strip().replace(brand\_replace\_map) | Không chuẩn hóa từ đầu, nhập tay khác nhau |
| **'Spontaneous', 'Awareness', 'Trial', 'P3M', 'P1M', 'Weekly':** transform các giá trị về nhị phân cho phân tích | - Dùng df[col].unique() để nhận diện  - dùng replace() để xử lí thay các giá trị đó về 1 |
| NeedstateDayDaypart | **Needstates:** - 'Drinking coffee', 'Drinking tea', 'Drinking other beverages (excluding tea, coffee, freeze)', 'Drinking ice-blended'  - 'Socializing with family / relatives', 'Socializing with friends', 'Socializing with colleagues'  - 'Working / Business meeting', 'Studying / Reading books'  - 'Relaxing (Alone)', 'Have meals (breakfast / lunch / dinner)', 'Have snack / pastry' | - Dùng df[col].unique() để nhận diện - Nhóm các cụm hành vi, rút gọn theo hành vi chính | Hành vi mô tả chi tiết không cần thiết |
| **Missing values** | Dayofweek | Visit#Dayofweek: 54 records, 0,14% | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ  - Xử lí bằng cách loại do tỉ lệ nhỏ, data nhiều | Thiếu do dẫn xuất từ Dayofweek |
| Dayofweek: 49 records, 0,13% | Có thể do lỗi sinh biến thời gian |
| Weekday#end: 49 giá trị, 0,13% | Dẫn xuất từ Dayofweek, lỗi thời gian hoặc định dạng |
| SA#var | Group\_size thiếu 15 giá trị (0.13%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ  - Điền bằng median | Dữ liệu không đầy đủ hoặc lỗi khi nhập liệu |
| Age thiếu 9 giá trị (0.08%) |
| Age#Group#2 thiếu 9 giá trị (0.08%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ  - Điền bằng suy từ Age | Thiếu do dẫn xuất từ Age chưa có |
| MPI#Mean thiếu 3,717 giá trị (31.60%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ  - So sánh 2 phương pháp  + Median cho ra một đỉnh cực mạnh tại điểm trung vị.  + Multiple Imputation tạo ra phân phối rộng hơn, phản ánh đúng độ biến động của thu nhập thực tế. => dùng Multiple Imputation | Không khai báo hoặc lỗi hệ thống, do dữ liệu nhạy cảm (thu nhập) |
| MPI#2 thiếu 3,717 giá trị (31.60%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ  - Điền bằng suy từ MPI | Biến dẫn xuất từ MPI chính, lỗi do thiếu dữ liệu nguồn |
| BUMO\_Previous thiếu 5,665 giá trị (48.17%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ  - Ghi nhận missing values thành no | Do đây là thông tin không bắc buộc |
| Brand\_Image | Awareness thiếu 397 giá trị (0.06%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ  - Ghi nhận missing values thành từ Brand Image tránh mâu thuẫn logic | Lỗi định dạng hoặc xử lý trong quá trình ETL, thiếu do lỗi khi merge dữ liệu, người trả lời bỏ qua câu hỏi |
| Daypart | Daypart thiếu 13 giá trị (0.07%) | - Sử dụng .isnull() để xác định.  - Điền bằng mode() nếu Visit#Daypart vẫn có giá trị.  - Nếu thiếu cả 2 cột → loại dòng. | - Khách không ghi rõ thời gian ghé thăm.  - Lỗi khi ghi nhận timestamp từ hệ thống POS hoặc survey. |
| Visit#Daypart thiếu 847 giá trị (4.41%) | - Điền median() của Visit#Daypart theo nhóm Daypart để giữ phân bố nội bộ.  - Nếu thiếu cả Daypart → loại bỏ dòng. | Do không ghi nhận được số lượt visit trong khung giờ đó hoặc khách không trả lời đầy đủ trong khảo sát. |
| NeedstateDayDaypart | DayGroup thiếu 19,559 giá trị (26.02%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ | Các giá trị thiếu không phải do khách bỏ trống, mà do quá trình tách cột từ trường tổng hợp chứa cả hai thông tin (DayGroup và Daypart). Nhiều bản ghi không rõ ràng hoặc định dạng không đồng nhất dẫn đến thiếu sau khi tách. |
| Daypart thiếu 55,620 giá trị (73.98%) | - Dùng .isnull().sum() để kiểm tra số lượng kết hợp df.shape[0] để tính tỉ lệ |
| Brand\_Health | Spontaneous thiếu 43,426 dòng (58.35%) | - Dùng .isnull().sum() để xác định  - Gán "0" cho các giá trị thiếu | Do đây là nhận biết tự phát → nhiều khách không nghĩ tới thương hiệu hoặc không biết đến. |
| Awareness thiếu 114 dòng (0.15%) | - Dùng .isnull().sum() để xác định  - Nếu Spontaneous có → gán Awareness = Spontaneous  - Nếu vẫn thiếu → gán "0" | Những người không có spontaneous thường cũng không có awareness. Có thể là khách mới hoặc không tiếp xúc với thương hiệu |
| P3M thiếu 45,570 dòng (61.23%) | - Dùng .isnull().sum() để xác định  -Chuyển thành nhị phân 1/0  - Nếu P1M = 1 → gán P3M = 1 | Dữ liệu hành vi gần nhất – bị thiếu nếu khách không sử dụng gần đây hoặc không nhớ. |
| Trial thiếu 27,089 dòng (36.4%) | - Dùng .isnull().sum() để xác định  - Chuyển thành nhị phân: 1 nếu không null, 0 nếu null  - Logic: nếu P3M hoặc P1M = 1 → Trial = 1 | Trial thường bỏ trống khi khách chưa dùng thử → nên null mang ý nghĩa logic (chưa từng dùng). |
| P1M thiếu 55,020 dòng (73.93%) | - Dùng .isnull().sum() để xác định  - Chuyển thành nhị phân 1/0  - Không nội suy, giữ nguyên cho churn analysis | Tần suất sử dụng trong tháng gần nhất – thường bị thiếu nếu không dùng gần đây. |
| Weekly, Daily thiếu ~82–90% | - Dùng .isnull().sum() để xác định  - Điền 0 với giá trị null | Do khách không biết đến thương hiệu nên không thể đánh giá mức độ hiểu, không thích/ biết , không thường sử dụng thương hiệu |
| - Comprehension thiếu 48,073 dòng (64.60%)  - Brand\_Likability thiếu 64,088 dòng (86.12%)  - NPS#P3M thiếu 52,814 dòng (70.97%)  - NPS#P3M#Group thiếu 52,814 dòng (70.97%) | - Dùng .isnull().sum() để xác định  - Giữ nguyên missing values do có ý nghĩa phân tích |
| Fre#visit thiếu 55,087 dòng (74.02%) | - Nếu NA → gán 0 cho fre\_visit, ppa, spending, spending\_use  - Nếu có ppa và spending → tính ngược lại | Không nhớ rõ số lần ghé thăm hoặc không có hành vi sử dụng. |
| PPA thiếu 60,346 dòng (81.09%) | - Nếu có fre\_visit + spending → tính lại  - Nếu thiếu → gán PPA = median, sau đó tính spending | Không nhớ chi tiêu mỗi lần ghé hoặc không có hành vi cụ thể. |
| Spending thiếu 60,346 dòng (81.09%) | - Nếu có fre\_visit + ppa → tính lại  - Nếu thiếu → gán spending = ppa × fre\_visit, nếu không có thì giữ NA | Tương tự PPA, không có dữ liệu thực tế hoặc khách không mua hàng. |
| Spending\_use thiếu 60,346 dòng (81.09%) | - Gán spending\_use = spending nếu NA | Tương tự PPA, không có dữ liệu thực tế hoặc khách không mua hàng. |
| **Logical integrity violations** | Brand\_Health | Có Comprehension/Likability nhưng Awareness = 'Unaware': 4 dòng | - Kiểm tra Comprehension/ Likability / Spontaneous is.notnull() nhưng Awareness = 'Unaware'  - Ghi nhận lại "Aware" | Đây là vi phạm logic hành vi người tiêu dùng, vì nếu người trả lời đã hiểu, thích hoặc tự nhớ đến thương hiệu, thì chắc chắn họ phải nhận biết thương hiệu đó.  Lỗi này có thể do nhập liệu sai, hoặc do người trả lời bỏ qua câu hỏi Awareness. |
| Có Spontaneous nhưng Awareness = 'Unaware': 90 dòng |
| **cross-table inconsistencies** | 2017Segmentation3685case | 2017Segmentation3685case là phân loại SA năm 2017 nhưng thiếu 9 giá trị | - Khi đối chiếu với Brand\_Health.csv, các ID này có các cột hành vi như Fre#visit, Spending, PPA bị **NaN**  **-** Không xử lí do đây là học họp lệ | - Thiếu dữ liệu (ví dụ: không có Spending hoặc Visit → không tính được segmentation).  - Không đủ điều kiện phân loại (chưa từng dùng thương hiệu, quá ít lượt ghé).  - Lỗi xử lý hoặc lọc dữ liệu khi làm sạch. |

**- Tóm tắt phương pháp và tiêu chí xử lý các lỗi chất lượng dữ liệu**

|  |  |  |
| --- | --- | --- |
| **Loại lỗi** | **Tổng quan cách phát hiện** | **Tổng quan xử lí** |
| Duplicate | Dùng .duplicated().sum() để đếm | Xoá các dòng trùng lặp, giữ lại bản ghi đầu tiên |
| Redundancy | Kiểm tra bằng unique(), so sánh ý nghĩa dựa vào từ điển dữ liệu | Xoá các cột trùng nội dung hoặc đã được dẫn xuất từ cột khác |
| Không nhất quán nhãn | Dùng df[col].unique() để nhận diện | Chuẩn hóa nội dung nhãn bằng ánh xạ từ điển thủ công |
| Missing | Dùng .isnull().sum() để đếm và tính tỷ lệ | Tùy tình huống: điền median/mode, nội suy từ cột khác, hoặc loại nếu tỷ lệ quá nhỏ |
| Vi phạm logic nội bộ | Kiểm tra điều kiện logic giữa các cột | Sửa lại các cột mâu thuẫn dựa trên quy luật hành vi hợp lý của người dùng |
| Không đồng nhất giữa bảng | So sánh ID và các trường chính giữa các bảng | Phát hiện và xử lý các dòng thiếu dữ liệu do lỗi khi merge hoặc thiếu điều kiện phân loại |

**- Phân tích nguyên nhân và đề xuất xử lý lỗi không đồng nhất giữa các bảng**

+ Lỗi phát hiện: Trong bộ dữ liệu năm 2017, có 9 đối tượng trong bảng SA\_var\_cleaned.csv không xuất hiện trong bảng Segmentation\_cleaned.csv, dẫn đến lỗi không đồng nhất giữa các bảng.

+ Phân tích nguyên nhân:

* Việc gán nhãn phân khúc (segmentation) được thực hiện dựa trên các biến hành vi, chủ yếu là PPA (Purchase Per Action), được tính toán từ các trường Visit (tần suất ghé thăm) và Spending (mức chi tiêu). Các trường này lấy từ bảng Brand\_Health\_cleaned.csv.
* Khi kiểm tra, cả 9 ID bị thiếu segmentation đều có giá trị NaN (thiếu) ở các trường hành vi như fre\_visit, spending, và ppa. Do đó, không thể tính được PPA, và hệ thống phân loại không thể áp dụng logic phân khúc cho các đối tượng này.
* Điều này cho thấy nguyên nhân không phải do lỗi xử lý dữ liệu, mà là do thiếu dữ liệu hành vi, có khả năng cao do những người tham gia chưa từng ghé thăm hoặc mua hàng tại thương hiệu trong khảo sát.

+ Đề xuất biện pháp xử lý:

* Đánh dấu rõ ràng các dòng này là “Không thể phân loại” (Not Segmentable)
* Giữ lại các ID này trong dữ liệu với nhãn rõ ràng như “Unassigned” để minh bạch
* Loại trừ các dòng này khỏi tập huấn luyện mô hình phân khúc hành vi nhưng có thể dùng cho các phân tích định tính khác như nhận diện thương hiệu hoặc nhận thức