BÁO CÁO DỰ ÁN  
PHÂN TÍCH GIỎ HÀNG (MARKET BASKET ANALYSIS)

Sinh viên: Nguyễn Đức Bình  
Môn: DAM501.8  
Python 3.13.6 – Thuật toán: Apriori (ID lẻ) / FP-Growth + FP-Max (ID chẵn)

# 1. Tóm tắt điều hành

Mục tiêu của dự án là khai phá tập phổ biến và luật kết hợp trên dữ liệu bán lẻ, nhằm phát hiện các nhóm sản phẩm thường mua cùng nhau để đề xuất cross-sell/upsell, thiết kế combo và hỗ trợ bố trí kệ hàng. Do dữ liệu có ít giao dịch chứa ≥2 mặt hàng (4,070/294,188), mô hình cần ngưỡng support/confidence thấp và đánh giá bằng lift.

# 2. Yêu cầu đề tài

• ID chẵn: sử dụng maxFP-growth/FP-Max để khai thác tập phổ biến tối đại; dùng FP-Growth để sinh Association Rules.  
• ID lẻ: sử dụng Apriori để khai thác tập phổ biến và sinh Association Rules.  
• Tự đưa ra giá trị min\_support, min\_confidence và giải thích cách chọn.  
• Nộp: file dataset, file notebook/code, ảnh chứng chỉ Coursera. Trình bày online theo lịch.

# 3. Dữ liệu và tiền xử lý

• Kích thước: 302,010 dòng, 30 cột. Cột giao dịch: Transaction\_ID. Cột mặt hàng: Product\_Category.  
• Số giao dịch: 294,188; giao dịch ≥2 mặt hàng: 4,070; số nhóm hàng (unique): 5.  
• Tiền xử lý: chuẩn hoá chữ (strip + upper), loại trùng (mỗi hóa đơn–mỗi mặt hàng 1 lần), gộp theo Transaction\_ID để tạo giỏ hàng; one-hot bằng TransactionEncoder.

Bảng 1. Hồ sơ cột (rút gọn):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cột | Kiểu dữ liệu | Không null | Số null | % Null | Số giá trị khác nhau |
| Transaction\_ID | float64 | 301677 | 333 | 0.11 | 294461 |
| Customer\_ID | float64 | 301702 | 308 | 0.1 | 86766 |
| Name | object | 301628 | 382 | 0.13 | 159390 |
| Email | object | 301663 | 347 | 0.11 | 52897 |
| Phone | float64 | 301648 | 362 | 0.12 | 299295 |
| Address | object | 301695 | 315 | 0.1 | 299329 |
| City | object | 301762 | 248 | 0.08 | 130 |
| State | object | 301729 | 281 | 0.09 | 54 |
| Zipcode | float64 | 301670 | 340 | 0.11 | 93978 |
| Country | object | 301739 | 271 | 0.09 | 5 |
| Age | float64 | 301837 | 173 | 0.06 | 53 |
| Gender | object | 301693 | 317 | 0.1 | 2 |
| Income | object | 301720 | 290 | 0.1 | 3 |
| Customer\_Segment | object | 301795 | 215 | 0.07 | 3 |
| Date | object | 301651 | 359 | 0.12 | 366 |

# 4. Phương pháp

• Apriori (ID lẻ): tìm tập phổ biến theo mức, sinh ứng viên và cắt tỉa nhờ tính chất kháng đơn điệu.  
• FP-Growth (ID chẵn): xây dựng FP-tree để khai phá mà không cần sinh ứng viên, nhanh hơn trên dữ liệu lớn.  
• FP-Max (ID chẵn): tìm tập phổ biến tối đại – không có siêu tập nào cũng phổ biến; dùng để tóm tắt pattern dài.  
• Association Rules: sinh luật A→B và đo lường bằng support, confidence, lift (thêm leverage, conviction).

# 5. Lựa chọn tham số

• min\_support = k / N, với N = số giỏ (N = 294,188 hoặc chỉ giỏ ≥2 món N = 4,070).  
• Quét lưới confidence: 0.8 → 0.1 (có thể hạ tới 0.01 khi dữ liệu rất thưa).  
• Lọc lift ≥ 1.0 để ưu tiên luật có sức nâng; mục tiêu số luật trình bày: 50–300.  
• Chọn cấu hình gần biên dưới để bộ luật gọn và dễ diễn giải.

# 6. Kết quả thí nghiệm

• Xuất file frequent\_itemsets.csv và association\_rules.csv từ script.  
• Chèn vào báo cáo/slide: Top-10 luật theo lift (kèm support, confidence).  
• Nếu ID chẵn: báo cáo thêm frequent\_itemsets\_maximal.csv (FP-Max).

# 7. Insight & khuyến nghị

• Dựa trên các luật có lift cao, đề xuất combo sản phẩm, cross-sell tại trang checkout.  
• Bố trí kệ: đặt gần nhau các nhóm hàng thường đi cùng.  
• Thiết kế gói khuyến mại cho bộ 2–3 nhóm hàng có liên hệ mạnh.

# 8. Hạn chế & rủi ro

• Dữ liệu thưa: chỉ 4,070/294,188 giao dịch có ≥2 mặt hàng; chỉ 5 nhóm hàng.  
• Chưa xét mùa vụ/khuyến mại/địa lý; luật phản ánh tương quan chứ không đảm bảo nhân-quả.  
• Cần kiểm tra độ ổn định của luật theo thời gian và theo phân khúc khách hàng.

# 9. Hướng phát triển

• Gom giỏ theo Customer + Date để tăng số item mỗi giỏ, giúp sinh luật rõ ràng hơn.  
• So sánh Apriori vs FP-Growth về thời gian & số lượng luật trên cùng tham số.  
• Thêm thuộc tính Brand/Type vào item để phân tích sâu hơn.  
• Triển khai A/B test cho gợi ý combo.

# 10. Tái lập & môi trường chạy

• Python 3.13.6; venv riêng. Gói: pandas, mlxtend.  
• Chạy script: python mba\_assignment\_final.py (đổi ID\_PARITY='odd'/'even' theo MSSV).