**TRƯỜNG ĐẠI HỌC KINH TẾ**

**ĐẠI HỌC ĐÀ NẴNG**

**KHOA THỐNG KÊ – TIN HỌC**



**BÁO CÁO**

**Đề tài: Ứng dụng luật kết hợp trong phân tích giỏ hàng**

**Môn:** Phân tích dữ liệu bằng Python

***Giáo viên hướng dẫn*** *:* TS. *Lê Diên Tuấn*

***Sinh viên thực hiện*** *:* 48K21.2-Nhóm 7

Doãn Diệu Diễm: 0367674438 (100%)

Hà Thị Quỳnh Anh (100%)

Trịnh Thị Khánh Huyền (100%)

Nguyễn Kiều Khánh Hạ (100%)

Nguyễn Anh Thư (100%)

*Đà Nẵng 03/2025*

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn chân thành đến thầy **Lê Diên Tuấn**, người đã tận tình hướng dẫn và đồng hành cùng nhóm chúng em trong suốt quá trình thực hiện dự án. Sự tận tâm và những góp ý hữu ích của thầy không chỉ giúp nhóm hoàn thiện dự án mà còn mang lại cho chúng em nhiều bài học giá trị. Những sự hỗ trợ này là nguồn động lực to lớn, giúp chúng em vượt qua khó khăn và đạt được kết quả như mong đợi.

**LỜI CAM ĐOAN**

Nhóm chúng em xin cam đoan bài Báo cáo môn học Phân tích dữ liệu bằng Python là hoàn toàn tự nhóm chúng em nghiên cứu và thực hiện. Mọi nội dung trong bài báo cáo đều trung thực, khách quan, chính xác và không sao chép của bất kỳ ai. Các thông tin đều được tìm hiểu và thu thập từ các nguồn uy tín và đã được kiểm chứng. Bài báo cáo này chưa từng được công bố hay sử dụng cho bất kỳ mục đích nào. Chúng em xin chịu hoàn toàn trách nhiệm về toàn bộ nội dung bài báo cáo này.

**MỤC LỤC**

[**LỜI CAM ĐOAN** ii](#_Toc197133361)

[**DANH MỤC HÌNH ẢNH** v](#_Toc197133362)

[**DANH MỤC TỪ VIẾT TẮT** vi](#_Toc197133363)

[I TỔNG QUAN 1](#_Toc197133364)

[1.1 Giới thiệu 1](#_Toc197133365)

[1.2 Ứng dụng của luật kết hợp 1](#_Toc197133366)

[1.3 Tại sao chủ đề có liên quan đến học phần phân tích dữ liệu 1](#_Toc197133367)

[II CƠ SỞ LÝ THUYẾT 2](#_Toc197133368)

[1.1 Luật kết hợp 2](#_Toc197133369)

[1.1.1 Đặc điểm của luật kết hợp 2](#_Toc197133370)

[1.1.2 Các chỉ số đánh giá 2](#_Toc197133371)

[1.1.3 Các bước thực hiện 3](#_Toc197133372)

[1.1.4 Thuật toán Apriori/ FP-Growth 3](#_Toc197133373)

[1.1.5 Tại sao nên sử dụng thuật toán Apriori cho đề tài 5](#_Toc197133374)

[1.2 Phân tích bài toán: Khám phá mối quan hệ mua hàng từ dữ liệu giao dịch tại tiệm bánh 5](#_Toc197133375)

[1.2.1 Bối cảnh 5](#_Toc197133376)

[1.2.2 Vấn đề đặt ra 5](#_Toc197133377)

[1.2.3 Mục tiêu 6](#_Toc197133378)

[1.2.4 Yêu cầu cụ thể 6](#_Toc197133379)

[III KẾT QUẢ MINH CHỨNG 7](#_Toc197133380)

[1.1 Mô tả dữ liệu 7](#_Toc197133381)

[1.1.1 Giới thiệu về dữ liệu 7](#_Toc197133382)

[1.1.2 Cấu trúc dữ liệu 7](#_Toc197133383)

[1.1.3 Nguồn dữ liệu: The Bread Basket 8](#_Toc197133384)

[1.1.4 Mô tả dữ liệu 8](#_Toc197133385)

[1.1.5 Quy trình nghiên cứu 8](#_Toc197133386)

[1.2 Kết quả, nhận xét code 9](#_Toc197133387)

[1.2.1 Kiểm tra và tiền xử lý dữ liệu 9](#_Toc197133388)

[1.2.2 Trực quan hoá dữ liệu 15](#_Toc197133389)

[1.2.3 Encoded transaction matrix 18](#_Toc197133390)

[1.2.4 Luật kết hợp 19](#_Toc197133391)

[IV KẾT LUẬN 26](#_Toc197133392)

[1.1 Kết luận chung 26](#_Toc197133393)

[1.2 Hướng đi của đề tài 27](#_Toc197133394)

[1.3 Hạn chế 27](#_Toc197133395)

[1.3.1 Giới hạn về phạm vi dữ liệu 27](#_Toc197133396)

[1.3.2 Ảnh hưởng của sản phẩm phổ biến 27](#_Toc197133397)

[V TÀI LIỆU THAM KHẢO 28](#_Toc197133398)

**DANH MỤC HÌNH ẢNH**

[Hình 1: Cấu trúc dữ liệu dataset "The Bread Basket" 7](#_Toc196948344)

[Hình 2. Mô tả dữ liệu 8](#_Toc196948345)

[Hình 3: Framework 8](#_Toc196948346)

**DANH MỤC TỪ VIẾT TẮT**

Conf: Confidence

Sup: Support

Min\_conf: Minimum Confidence

Min\_sup: Minimum Support

FP-Tree: Frequent Pattern Tree

# TỔNG QUAN

## Giới thiệu

Luật kết hợp là mối quan hệ giữa các tập thuộc tính trong cơ sở dữ liệu, là phương tiện hữu ích để khám phá các mối liên kết trong dữ liệu. Một luật kết hợp là một mệnh đề kéo theo có dạng X → Y, trong đó X, Y ⊆ I, thỏa mãn điều kiện X giao Y =Ø. Tập X gọi là nguyên nhân [Tiền đề], tập Y gọi là hệ quả [1].

Nghiên cứu luật kết hợp nhằm khám phá các mối liên hệ tiềm ẩn trong dữ liệu, tạo nền tảng cho việc hiểu sâu về dữ liệu và hỗ trợ đưa ra những quyết định hiệu quả trong thực tiễn.

## Ứng dụng của luật kết hợp

Phân tích giỏ hàng :Giúp tìm sản phẩm hay được mua cùng nhau.

Hệ thống gợi ý : Gợi ý sản phẩm, phim, dịch vụ dựa trên thói quen người dùng

Phân tích hành vi người dùng: Hiểu chuỗi hành động phổ biến trên web, app

Y tế: Liên hệ giữa triệu chứng và bệnh, thuốc và tác dụng phụ

Tài chính: Phát hiện mẫu giao dịch gian lận bất thường

Sinh học: Tìm mối liên hệ giữa các gen hoặc đột biến

Phân tích văn bản: Phát hiện các từ hay đi kèm trong tài liệu.

## Tại sao chủ đề có liên quan đến học phần phân tích dữ liệu

Chủ đề tìm luật kết hợp có liên quan chặt chẽ đến học phần Phân tích dữ liệu, vì bản chất của phân tích dữ liệu là khai thác và phát hiện các quy luật ẩn trong dữ liệu nhằm phục vụ cho các mục tiêu thực tiễn.

Việc tìm ra các luật kết hợp giúp nhận diện những mối quan hệ tiềm ẩn giữa các đối tượng trong tập dữ liệu, từ đó hỗ trợ dự đoán xu hướng, hiểu rõ hành vi người dùng và xây dựng các chiến lược tối ưu hóa hoạt động kinh doanh.

# CƠ SỞ LÝ THUYẾT

## Luật kết hợp

### Đặc điểm của luật kết hợp

Phát hiện các mẫu ẩn: Tìm ra mối quan hệ thường xuyên giữa các mục trong dữ liệu (ví dụ: sản phẩm thường mua cùng nhau).

Dễ triển khai: Dựa trên các thuật toán đơn giản như Apriori, FP-Growth, dễ áp dụng bằng các thư viện Python.

Xử lý dữ liệu lớn hiệu quả: Có khả năng mở rộng, xử lý tốt khi dữ liệu có hàng ngàn giao dịch hoặc mục.

Không yêu cầu dữ liệu theo thứ tự: Chỉ cần biết các mục cùng xuất hiện, không quan trọng thứ tự xảy ra.

Có thể sinh ra nhiều luật: Cần dùng ngưỡng như minsup, minconf để lọc ra các luật có ý nghĩa.

Ứng dụng đa dạng: Dùng trong phân tích giỏ hàng, hệ thống gợi ý, phân tích hành vi, phát hiện gian lận...

### Các chỉ số đánh giá

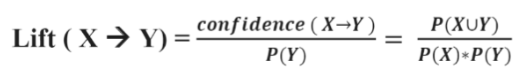
Độ hỗ trợ ( support): Trong khai thác luật kết hợp là tỷ lệ các giao dịch trong cơ sở dữ liệu chứa một tập hạng mục (itemset) cụ thể.



Độ tin cậy (confidence): Đo lường xác suất mặt hàng B được mua khi đã mua mặt hàng A, hay còn gọi là độ tin cậy của luật kết hợp A→B



Lift: Đánh giá mối quan hệ độc lập giữa A và B. Nếu Lift > 1, có mối liên hệ dương; nếu Lift < 1, có mối liên hệ âm; Lift = 1 thì A và B độc lập.



Khai thác luật kết hợp:

* Chọn ra độ phổ biến tối thiểu (minsup) và độ tin cậy tối thiểu (minconf)
* Bài toán khai thác LKH là tìm tất cả các luật dạng X → Y thỏa mãn độ phổ biến và độ tin cậy tối thiểu
* sup(X → Y) ≥ minsup
* conf(X → Y) ≥ minconf

→ Những luật kết hợp thỏa mãn độ hỗ trợ tối thiểu (min\_sup) và độ tin cậy tối thiểu (min\_conf) được gọi là các luật mạnh

### Các bước thực hiện

**Bước 1:** Tiền xử lý dữ liệu (Đưa về dạng encoding)

**Bước 2:** Tìm tất cả các tập phổ biến (theo ngưỡng minsup)

**Bước 3:** Tạo ra các luật từ các tập phổ biến

* Mỗi tập phổ biến X, tạo các tập con khác rỗng của X
* Đối với mỗi tập con (khác rỗng) A của X, B=X-A
* A→B là luật kết hợp nếu conf(A→B)>=minconf
* conf(A→B)=sup(A ∪ B)/sup(A)=sup(X)/sup(A)

### Thuật toán Apriori/ FP-Growth

#### Thuật toán Apriori

1. **Giới thiệu:**

Do R. Agrawal và R. Srikant giới thiệu năm 1994

Khai phá các tập mục thường xuyên cho các luật kết hợp dạng Boolean.

Chiến lược lặp của Apriori: các k-itemset được sử dụng để khảo sát các (k + 1)-itemset [2]

1. **Phương pháp:** Tìm tất cả các tập phổ biến 1 hạng mục

**Bước 1:** Tạo các tập ứng viên k+1 hạng mục từ tập phổ biến có k hạng mục

**Bước 2:** Kiểm tra độ phổ biến của ứng viên trên cơ sở dữ liệu. Chỉ lấy các Large Items làm tập ứng viên, bỏ qua Small Items (tức là lấy những items nào >=minsup cho trước).

**Bước 3:** Dừng khi không tạo ra được tập phổ biến hay tập ứng viên

#### Thuật toán FP-Growth

1. **Giới thiệu:**

Được đề xuất nhằm cải thiện hiệu năng so với Apriori trong việc khai phá tập mục thường xuyên.

Sử dụng một cấu trúc dữ liệu gọi là FP-Tree để biểu diễn cơ sở dữ liệu giao dịch theo cách nén gọn.

Không sinh tập ứng viên như Apriori, giúp giảm thiểu chi phí tính toán đáng kể.

Áp dụng phương pháp chia để trị để khai thác tập phổ biến từ FP-Tree một cách đệ quy.

1. **Phương pháp:**

**Bước 1:** Khởi tạo FP-Tree: Ban đầu, FP-Tree chỉ chứa duy nhất một nút gốc, được ký hiệu là null.

**Bước 2:** Lần duyệt 1 (quét dữ liệu):

* Duyệt toàn bộ cơ sở dữ liệu giao dịch để tính độ hỗ trợ (support) của từng mục.
* Loại bỏ các mục không đạt ngưỡng hỗ trợ tối thiểu (minsup).
* Các mục còn lại được sắp xếp theo thứ tự giảm dần về độ hỗ trợ.

**Bước 3: Lần duyệt 2 (xây cây)**

* Với mỗi giao dịch:
* Chỉ giữ lại các mục thường xuyên, sắp xếp theo thứ tự đã tính.
* Chèn vào FP-Tree từ nút gốc, tạo ra các đường đi (path) đại diện cho giao dịch.
* Các giao dịch có chung phần tử sẽ chia sẻ đường đi, giúp nén dữ liệu.

**Bước 4: Khai thác tập phổ biến từ FP-Tree:**

* Phân tích từng nút theo bottom-up (từ lá lên gốc).
* Tạo các FP-Tree điều kiện (conditional FP-Tree).
* Khai thác đệ quy các tập mục phổ biến mà không cần sinh tập ứng viên. [3]
* Dừng khi cây con không còn mục nào đạt minsup.

### Tại sao nên sử dụng thuật toán Apriori cho đề tài

Thuật toán Apriori được lựa chọn để áp dụng cho bộ dữ liệu giao dịch bán bánh vì bản chất bài toán là khai phá các mối quan hệ mua hàng - một bài toán kinh điển của khai phá luật kết hợp.

Với quy mô dữ liệu vừa phải, Apriori vận hành hiệu quả nhờ nguyên lý chống đơn điệu (anti-monotonicity), giúp loại bỏ sớm các tổ hợp mặt hàng không phổ biến, tiết kiệm tài nguyên tính toán.

Ngoài ra, Apriori cho ra các chỉ số như support, confidence và lift rất rõ ràng, dễ giải thích, hỗ trợ trực tiếp cho việc đề xuất các chiến lược kinh doanh như gợi ý sản phẩm hoặc tạo combo khuyến mãi.

Đặc biệt, các thư viện như mlxtend trong Python cung cấp sẵn công cụ triển khai Apriori, giúp việc thực hiện và trực quan hóa kết quả trở nên nhanh chóng và hiệu quả.

## Phân tích bài toán: Khám phá mối quan hệ mua hàng từ dữ liệu giao dịch tại tiệm bánh

### Bối cảnh

Trong bối cảnh ngành ẩm thực và đồ uống ngày càng cạnh tranh, việc thấu hiểu thói quen mua hàng của khách hàng đóng vai trò then chốt trong việc tối ưu hóa chiến lược kinh doanh.

Một tiệm bánh nổi tiếng đã thu thập dữ liệu chi tiết về các giao dịch bán hàng hàng ngày, bao gồm danh sách sản phẩm được mua, thời gian giao dịch và phân loại ngày bán (ngày thường/cuối tuần).

### Vấn đề đặt ra

Làm thế nào để khai phá các mẫu hành vi mua hàng tiềm ẩn, phát hiện những nhóm sản phẩm thường xuyên được mua cùng nhau?

Từ đó, đề xuất các chiến lược như:

* Gợi ý bán chéo sản phẩm.
* Tạo combo ưu đãi hợp lý.
* Tối ưu hóa cách bố trí sản phẩm tại cửa hàng.
* Nâng cao trải nghiệm khách hàng và tăng doanh thu.

### Mục tiêu

Phát hiện các nhóm sản phẩm thường xuyên được mua cùng nhau trong các giao dịch buổi sáng cuối tuần.

Đưa ra các khuyến nghị kinh doanh dựa trên các mẫu được phát hiện như: gộp các sản phẩm lại thành combo, bố trí gần nhau trên kệ, khuyến mãi chéo.

### Yêu cầu cụ thể

#### Chuẩn bị dữ liệu

Làm sạch dữ liệu

Định dạng: Chuyển dữ liệu thành danh sách giao dịch, mỗi giao dịch là tập hợp các mặt hàng (Item) được mua cùng nhau, dựa trên cột Transaction và Item.

#### Phân tích luật kết hợp

Thuật toán: Sử dụng Apriori.

Ngưỡng chỉ số:

* Minimum Support: ≥ 0.01 (1%) để lọc các tập hợp mặt hàng phổ biến.
* Minimum Confidence: ≥ 0.2 (20%) nhằm mục tiêu khám phá đa dạng các mối quan hệ tiềm năng giữa các sản phẩm
* Lift: Ưu tiên các luật có lift > 1 (mối quan hệ không ngẫu nhiên).

Xử lý kết quả: Sắp xếp luật theo lift hoặc confidence, loại bỏ luật dư thừa.

#### Kết quả đầu ra

Danh sách các luật kết hợp với các chỉ số: support, confidence, lift.

Khuyến nghị kinh doanh dựa trên các luật: Gợi ý xây dựng combo bán hàng hoặc khuyến mãi chéo dựa trên các cặp sản phẩm có lift cao.

#### Công cụ thực hiện

Python với thư viện pandas, mlxtend (Apriori, association rules), matplotlib/seaborn cho trực quan hóa.

# KẾT QUẢ MINH CHỨNG

## Mô tả dữ liệu

### Giới thiệu về dữ liệu

Dataset "The Bread Basket" ghi lại các giao dịch mua hàng tại một tiệm bánh ở Anh từ 30/10/2016 đến 09/4/2017.

### Cấu trúc dữ liệu

Số dòng (bản ghi): 21,293

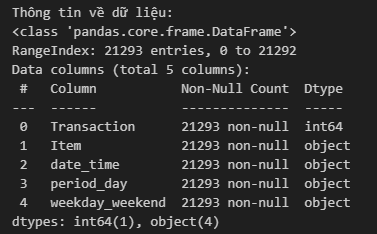
Số cột: 5

|  |  |  |
| --- | --- | --- |
| **Cột** | **Kiểu dữ liệu** | **Mô tả** |
| Transaction | int64 | Mã giao dịch (có thể nhiều món hàng trong 1 giao dịch) |
| Item | object | Tên mặt hàng được mua |
| date\_time | object | Ngày giờ mua hàng |
| period\_day | object | Thời điểm trong ngày (morning, afternoon, evening, night) |
| weekday\_weekend | object | Phân loại là ngày cuối tuần hay trong tuần(weekday hoặc weekend) |

Hình 1: Cấu trúc dữ liệu dataset "The Bread Basket"

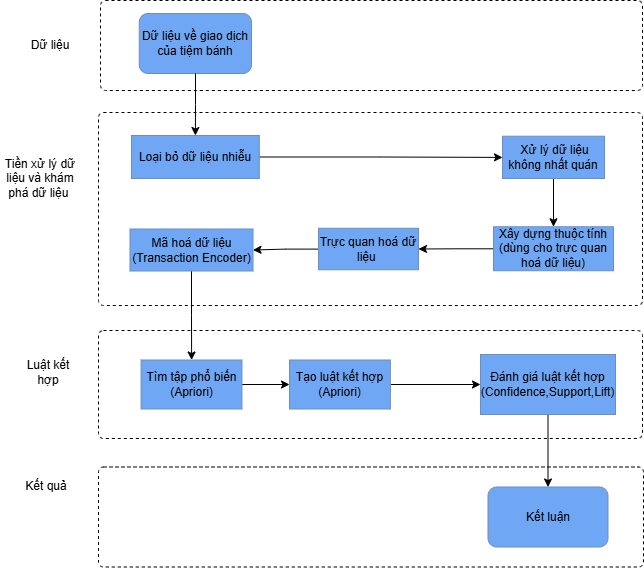
### Nguồn dữ liệu: [The Bread Basket](https://www.kaggle.com/datasets/mittalvasu95/the-bread-basket)

### Mô tả dữ liệu

****

Hình 2. Mô tả dữ liệu

### Quy trình nghiên cứu



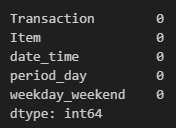
Hình 3: Framework

## Kết quả, nhận xét code

### Kiểm tra và tiền xử lý dữ liệu

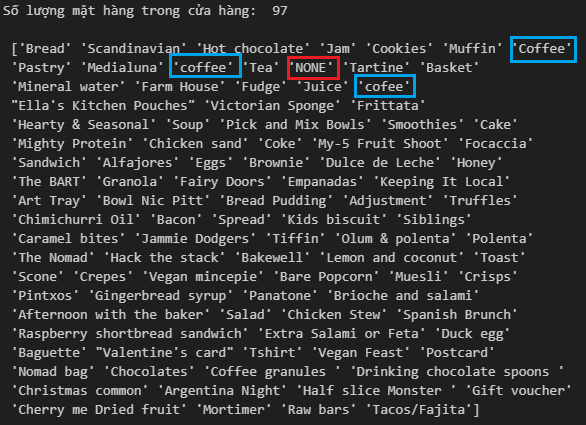
#### Dữ liệu nhiễu

1. Kiểm tra dữ liệu



**Nhận xét:** Qua kiểm tra, dữ liệu hiện tại không có giá trị bị thiếu (null) ở bất kì thuộc tính nào. Điều này cho thấy tập dữ liệu đầy đủ và sạch, đảm bảo rằng tất cả các biến đều sẵn sàng sử dụng cho các bước phân tích hoặc xây dựng mô hình sau này. Nhờ đó, không cần thực hiện các kỹ thuật xử lý giá trị thiếu như loại bỏ bản ghi, thay thế giá trị hoặc ước lượng giá trị còn thiếu. Đây là một yếu tố thuận lợi giúp tiết kiệm thời gian tiền xử lý và đảm bảo độ chính xác cao hơn cho các mô hình học máy hoặc phân tích thống kê tiếp theo.

1. Số lượng mặt hàng

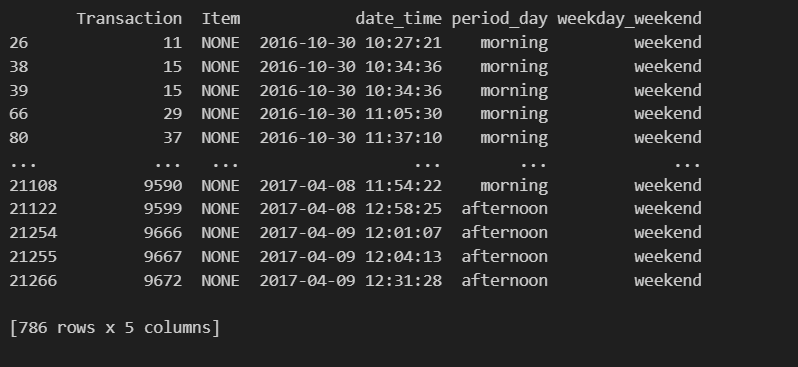


**Nhận xét:** Kết quả cho thấy cửa hàng có 97 mặt hàng. Tuy nhiên ta thấy có một số vấn đề như sau:

* Có item “NONE” không đại diện cho item nào cửa cửa hàng. Điều này cho thấy rằng trong danh sách các mặt hàng, “NONE” có thể là một giá trị không hợp lệ hoặc được sử dụng để chỉ các mục không xác định, không phù hợp với mục đích bán hàng. Để đảm bảo tính chính xác trong việc quản lý và phân tích dữ liệu, cần loại bỏ các mục “NONE” khỏi danh sách.
* Định dạng không thống nhất: có item viết hoa chữ cái đầu chữ sau viết thường nhưng có item lại viết hoa chữ cái sau (ví dụ: 'The BART'), hoặc viết thường(ví dụ: coffee). Mặt hàng coffee được lưu trữ ở dạng không thống nhất (coffee, Coffee, cofee)

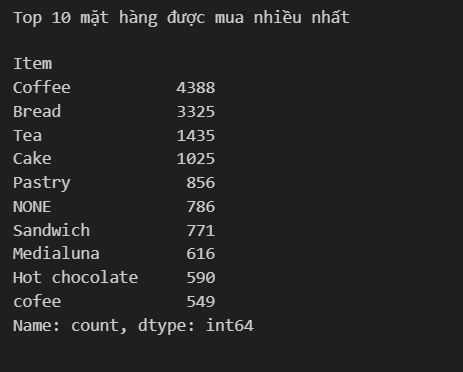
1. Loại bỏ dữ liệu nhiễu

**Liệt kê các sản phẩm có giá trị NONE**



**Nhận xét:** Trong tập dữ liệu, có tổng cộng 786 giao dịch chứa mặt hàng 'NONE', chiếm một tỷ lệ đáng kể so với tổng số giao dịch. Để đánh giá mức độ ảnh hưởng của giá trị này, ta tiến hành phân tích 10 mặt hàng bán chạy nhất trong tập dữ liệu và xác định vị trí của 'None' trong danh sách này.

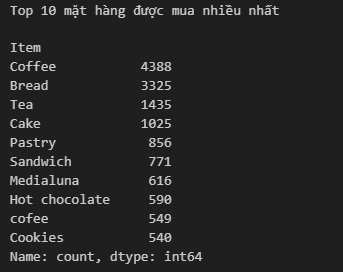
**Top 10 mặt hàng được mua nhiều nhất**

****

**Nhận xét:** Ta thấy mặt hàng "NONE" nằm trong top 6 mặt hàng có tần suất xuất hiện cao nhất, với tổng cộng 768 giao dịch. Tuy nhiên, "NONE" không đại diện cho một mặt hàng thực tế mà có thể là kết quả của lỗi nhập liệu hoặc quy trình ghi nhận giao dịch chưa chuẩn hóa. Vì vậy, nhóm quyết định loại bỏ toàn bộ các item "NONE" ra khỏi tập dữ liệu, dựa trên các lý do sau:

* Không thể suy luận chính xác giá trị thực tế:
* Không có cách nào chắc chắn để xác định sản phẩm thực sự tương ứng với "NONE". Việc sử dụng các phương pháp điền giá trị giả như thay thế bằng món phổ biến nhất, giá trị trung bình, hoặc dự đoán đều tiềm ẩn nguy cơ làm sai lệch bản chất thực tế của giao dịch.
* Ảnh hưởng tiêu cực đến quá trình khai thác tập mục:
* Giao dịch có sản phẩm "NONE" sẽ gây nhiễu trong quá trình tìm tập mục phổ biến và khai thác luật kết hợp, vì "NONE" không phải là một sản phẩm thực sự. Nếu giữ lại, "NONE" có thể xuất hiện trong các tập mục và luật khai thác, từ đó làm giảm giá trị thực tiễn và độ tin cậy của phân tích.

**Top 10 mặt hàng sau khi xóa giá trị NONE**



**Kết quả:** Nếu các giá trị 'NONE' đã được xóa hoàn toàn khỏi dữ liệu, thì các kết quả nhận được cho top 10 mặt hàng là hợp lệ và phản ánh các mặt hàng phổ biến trong cơ sở dữ liệu sau khi loại bỏ các bản ghi không xác định. Số bản ghi sau khi loại bỏ item”NONE” là 20507

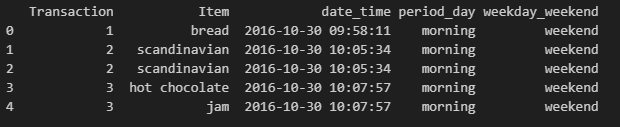
https://lh7-rt.googleusercontent.com/docsz/AD_4nXeeUUSExcpueYJYz91o0xV6cRAKlohVWLRYY6d6ym1pB8z-Wmn8qDEL6y6r3W7eW9ub_9PZhvCLOfWa-zdo9R88k0qhjInFf-IjSGgR0xFR4rjSNxoYPyv5OgQ6FNjB5fgdZ98bCA?key=jJes8ZTUoTBw5xxSnCCUJTqf

#### Xử lý dữ liệu không nhất quán

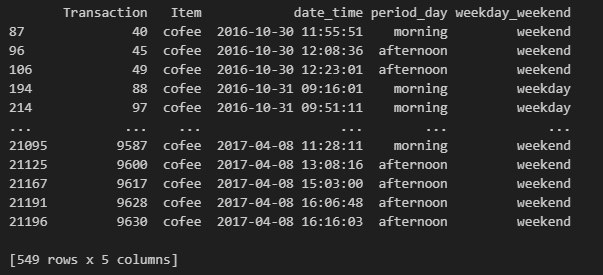
Sau khi loại bỏ các giao dịch chứa sản phẩm "NONE", nhóm tiếp tục xử lý vấn đề về định dạng tên mặt hàng. Cụ thể, dữ liệu ghi nhận không thống nhất: một số item được viết hoa toàn bộ, một số chỉ viết hoa chữ cái đầu, và một số để chữ thường hoàn toàn. Nếu giữ nguyên, sự khác biệt này có thể dẫn đến việc hệ thống nhận diện cùng một sản phẩm thành nhiều mục khác nhau, làm sai lệch quá trình khai thác tập mục phổ biến và luật kết hợp.

Để đảm bảo tính đồng nhất và chính xác trong phân tích, nhóm đã chuẩn hóa toàn bộ tên mặt hàng về chữ viết thường.

**Kết quả**

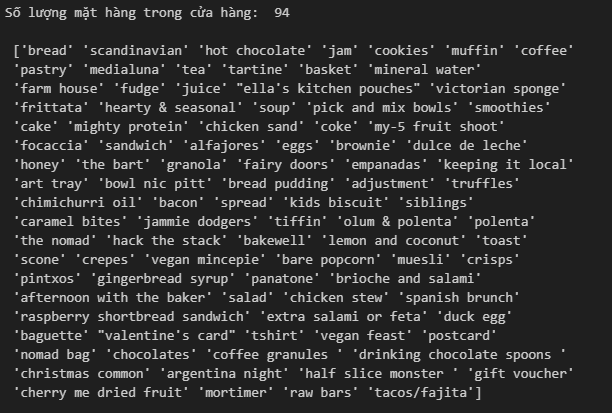
****

**Chuyển đổi định dạng các mặt hàng “cofee” thành “coffee”**

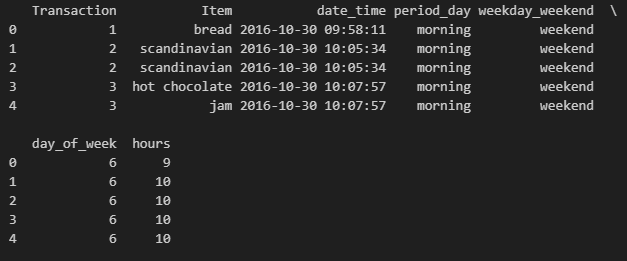


**Nhận xét:** Trước khi thực hiện thay đổi, có tổng cộng 549 dòng trong cột 'Item' có giá trị là "cofee". Sau khi sử dụng phương pháp thay thế để chuyển "cofee" thành "coffee", kết quả cho thấy số lượng các dòng có giá trị "cofee" đã được thay thế hoàn toàn, giúp chuẩn hóa tên mặt hàng và đảm bảo tính nhất quán trong dữ liệu.

**Kết quả số lượng mặt hàng sau khi làm sạch dữ liệu**



#### Xây dựng thuộc tính của dữ liệu (Dùng cho trực quan hoá dữ liệu)

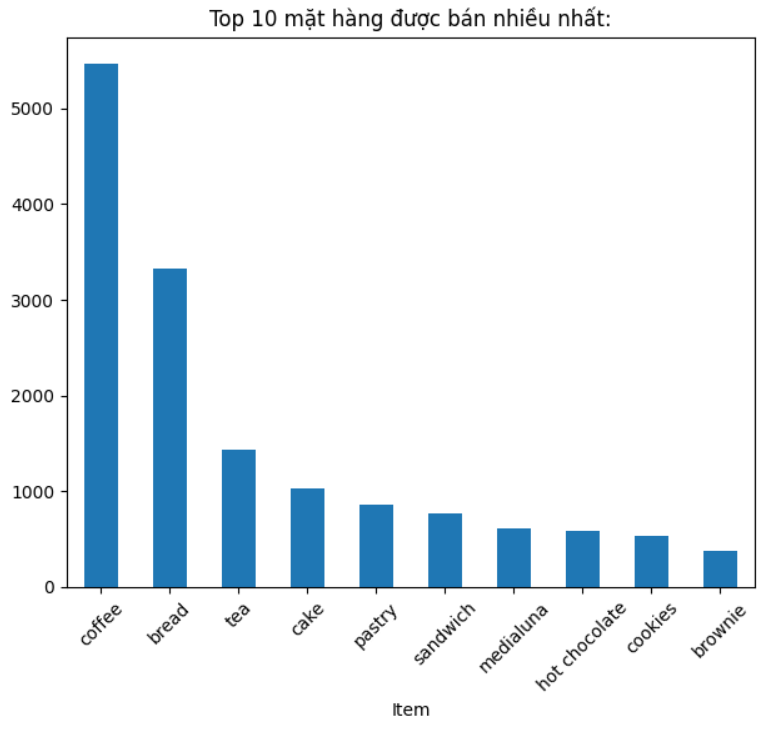


**Nhận xét:** Thêm các cột day\_of\_week, hours giúp hỗ trợ phân tích đa chiều.

* Cột day\_of\_week hỗ trợ phân tích sự khác biệt trong hành vi mua hàng giữa các ngày trong tuần, ví dụ như cuối tuần lượng giao dịch có thể tăng mạnh.
* Cột hours giúp nhận diện các khung giờ cao điểm và thời gian thấp điểm trong ngày, từ đó hỗ trợ tối ưu hóa hoạt động bán hàng, nhân sự và chương trình khuyến mãi.

### Trực quan hoá dữ liệu

#### Top 10 mặt hàng bán chạy nhất



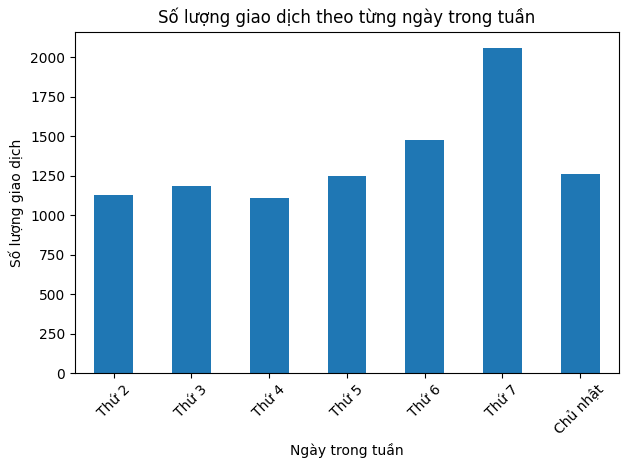
**Nhận xét:**

* Coffee là mặt hàng được bán nhiều nhất, với số lượng vượt trội so với các mặt hàng còn lại.
* Bread giữ vị trí thứ hai, tuy có doanh số thấp hơn coffee nhưng vẫn ở mức cao so với phần còn lại.
* Các mặt hàng từ tea đến brownie có số lượng bán thấp hơn nhiều và giảm dần đều, cho thấy sự phân hóa mạnh về nhu cầu tiêu dùng.
* Khoảng cách lớn giữa hai sản phẩm dẫn đầu (coffee, bread) và các sản phẩm khác phản ánh xu hướng tập trung mua sắm vào một số ít mặt hàng chủ lực.
* Các sản phẩm như brownie, cookies có doanh số khá thấp trong top 10, cho thấy mức độ ưa chuộng hạn chế.

**Kết luận**

 Biểu đồ cho thấy coffee và bread là hai sản phẩm chủ lực, chiếm tỷ trọng lớn trong tổng doanh số bán hàng. Các sản phẩm còn lại có mức tiêu thụ thấp hơn nhiều, gợi ý rằng doanh nghiệp nên tiếp tục duy trì thế mạnh ở các mặt hàng chủ lực đồng thời tìm kiếm giải pháp thúc đẩy tiêu thụ ở các mặt hàng còn lại.

#### Số lượng giao dịch ra theo từng ngày trong tuần



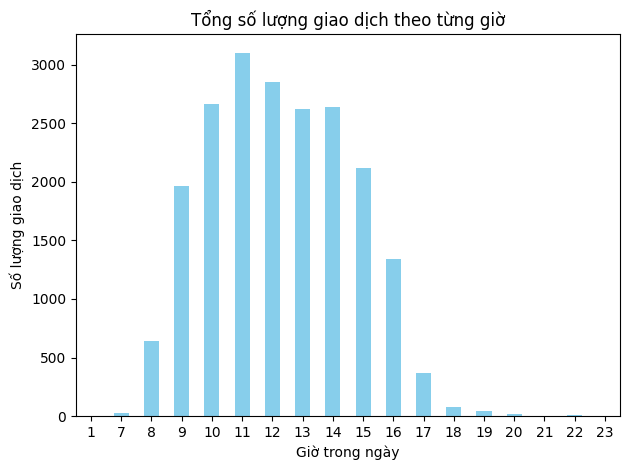
**Nhận xét:**

* Số lượng giao dịch trong tuần khá ổn định từ thứ 2 đến thứ 6, dao động quanh mức 1100–1500 giao dịch mỗi ngày.
* Thứ 7 ghi nhận lượng giao dịch cao nhất trong tuần, vượt mốc 2000 giao dịch.
* Chủ nhật có số lượng giao dịch thấp hơn thứ 7 nhưng vẫn cao hơn hầu hết các ngày trong tuần, ngang ngửa với thứ 5 và thứ 6.

**Kết luận:**

Giao dịch có xu hướng tăng mạnh vào cuối tuần, đặc biệt là thứ 7. Doanh nghiệp nên tận dụng thời điểm này để đẩy mạnh hoạt động kinh doanh như tổ chức các chương trình khuyến mãi hoặc ưu đãi đặc biệt vào thứ 7 nhằm tối ưu doanh số.

#### Tổng số lượng giao dịch theo từng giờ



**Nhận xét:**

* Giao dịch bắt đầu tăng mạnh từ 8 giờ sáng và đạt đỉnh vào khoảng 11 giờ trưa.
* Từ 11 giờ đến 14 giờ, số lượng giao dịch duy trì ở mức cao và khá ổn định.
* Sau 15 giờ, lượng giao dịch bắt đầu giảm nhanh và gần như chấm dứt sau 17 giờ.
* Các khung giờ từ 18 giờ trở đi có số giao dịch rất thấp, gần như không đáng kể.

**Kết luận:**

Giao dịch chủ yếu tập trung vào khoảng thời gian từ 8 giờ sáng đến 15 giờ chiều, đặc biệt cao nhất vào 11 giờ. Doanh nghiệp nên tập trung nhân lực, hàng hóa và các chương trình ưu đãi trong khung giờ vàng này để tối ưu doanh thu.

### Encoded transaction matrix



**Nhận xét:**

* Dữ liệu được trình bày dưới dạng bảng nhị phân (giá trị True/False) với 94 cột tương ứng với 94 sản phẩm khác nhau.Mỗi dòng tương ứng với một giao dịch, trong đó:
  + True thể hiện mặt hàng đó đã được mua trong giao dịch.
  + False nghĩa là mặt hàng đó không được mua trong giao dịch.
* Quan sát 10 dòng đầu tiên, phần lớn các giá trị đều là False, cho thấy trong mỗi giao dịch, khách hàng chỉ chọn mua một số ít sản phẩm (không phải mua nhiều loại cùng lúc).
* Dữ liệu dạng này thường được sử dụng để phân tích giỏ hàng (Market Basket Analysis), áp dụng các thuật toán như Apriori nhằm tìm ra các nhóm sản phẩm thường được mua cùng nhau.

**Kết luận:**

Tập dữ liệu đang ở dạng thích hợp để khai thác các mối liên kết giữa sản phẩm (association rules mining), từ đó đưa ra gợi ý mua hàng chéo (cross-selling) hoặc bố trí sản phẩm trong cửa hàng hợp lý hơn.

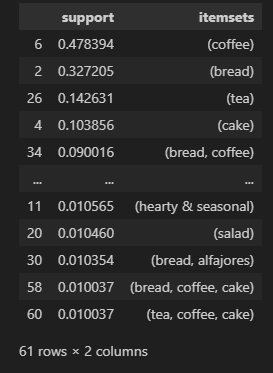
### Luật kết hợp

#### Tìm tập phổ biến

Để xác định các tập mục phổ biến trong phân tích giỏ hàng, trước tiên cần thiết lập ngưỡng hỗ trợ tối thiểu (min\_sup). Trong nghiên cứu này, ngưỡng min\_sup được chọn là 1%, tương ứng với yêu cầu một tập mục phải xuất hiện trong ít nhất khoảng 94 giao dịch để được coi là phổ biến. Lựa chọn ngưỡng này được cân nhắc dựa trên đặc điểm của dữ liệu và mục tiêu phân tích.

Dữ liệu giỏ hàng thường có tính phân tán cao, với nhiều sản phẩm chỉ được mua cùng nhau trong một số lượng nhỏ giao dịch. Nếu đặt ngưỡng min\_sup quá cao, chẳng hạn 5% (tương đương khoảng 470 giao dịch), các tập mục tiềm năng, bao gồm những cặp sản phẩm ít phổ biến nhưng có giá trị kinh doanh có thể bị loại bỏ. Ngược lại, ngưỡng min\_sup 1% là một giá trị tương đối thấp, cho phép giữ lại nhiều tập mục phổ biến hơn để phân tích, đồng thời loại bỏ các tập mục quá hiếm (xuất hiện dưới 94 giao dịch) nhằm đảm bảo ý nghĩa thống kê và thực tiễn.

1. **Kết quả**

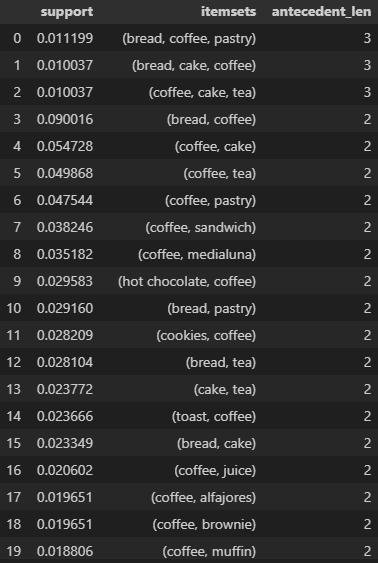


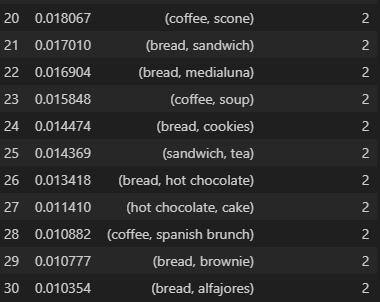
1. **Nhận xét**

Với min\_sup = 0.01 thì ta có 61 tập thường xuyên. Các tập mục 1-itemset phổ biến nhất là coffee (support=0.48) và bread (support = 0.33), cho thấy đây là hai sản phẩm được mua nhiều nhất

Chỉ những tập mục có từ 2-itemset trở lên mới tạo được luật kết hợp, ta tiến hành lọc ra những tập thường xuyên có 2-itemset trở lên

1. **Kết quả của 2-itemset trở lên**





1. **Nhận xét**

Sau khi lọc các tập mục thường xuyên có độ dài từ 2 trở lên, thu được 31 tập mục.

Các tập 3-itemset nổi bật gồm:

* {bread, coffee, pastry} có support 0.011199,
* {bread, cake, coffee} và {coffee, cake, tea} đều có support 0.010037.

➔ Tuy có support nhỏ, nhưng cho thấy xu hướng kết hợp mua thực phẩm ăn nhanh kèm đồ uống.

Các 2-itemset có support cao nhất:

* {bread, coffee} với support 0.090016,
* {coffee, cake} đạt support 0.054728,
* {coffee, tea} đạt support 0.049868.

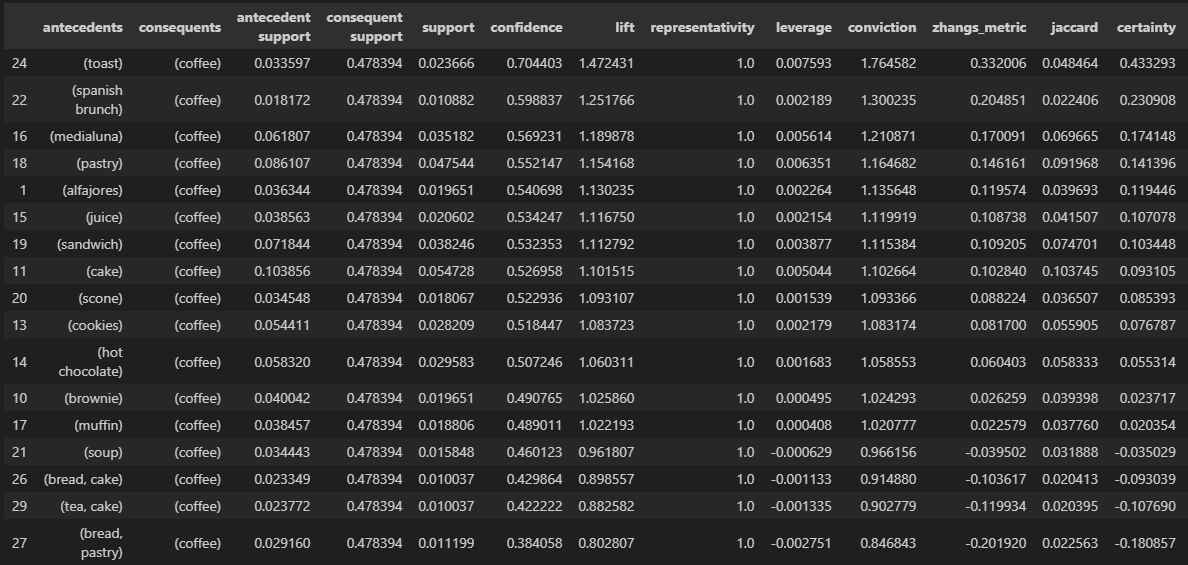
➔ Các kết quả này phản ánh mối quan hệ mua sắm rất mạnh giữa cà phê và các sản phẩm khác như bánh mì, bánh ngọt, trà.

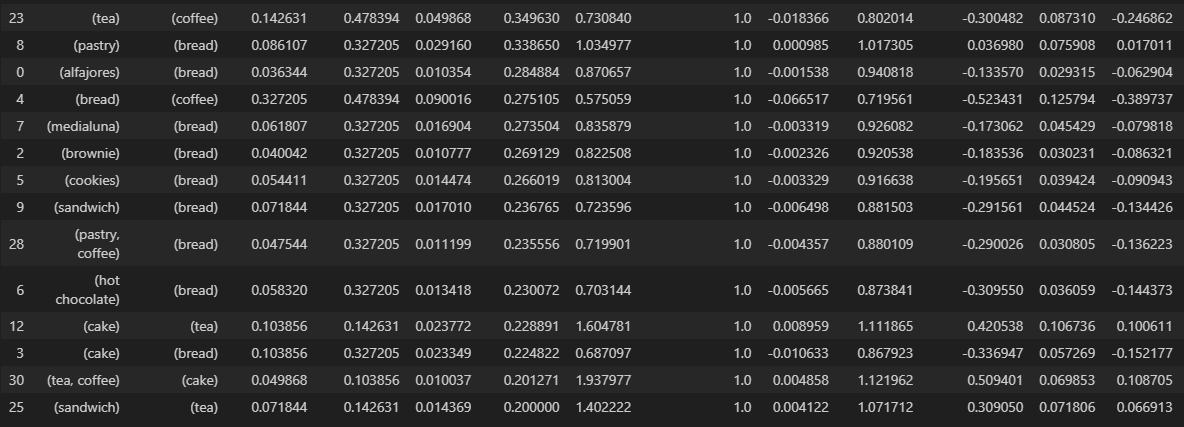
Ngoài ra, nhiều tổ hợp 2-itemset khác xoay quanh coffee và bread, cho thấy cà phê và bánh mì là hai sản phẩm chủ đạo của cửa hàng

#### Tạo ra luật kết hợp

Từ các tập mục phổ biến, ta thiết lập ngưỡng độ tin cậy tối thiểu (min\_confidence) là 0.2 để tạo các luật kết hợp, đảm bảo rằng mỗi luật có xác suất xảy ra ít nhất 20%. Sử dụng hàm association\_rules từ thư viện mlxtend, các luật được tạo ra và sắp xếp theo độ tin cậy giảm dần để ưu tiên các luật có mối quan hệ mạnh nhất.

1. **Kết quả**





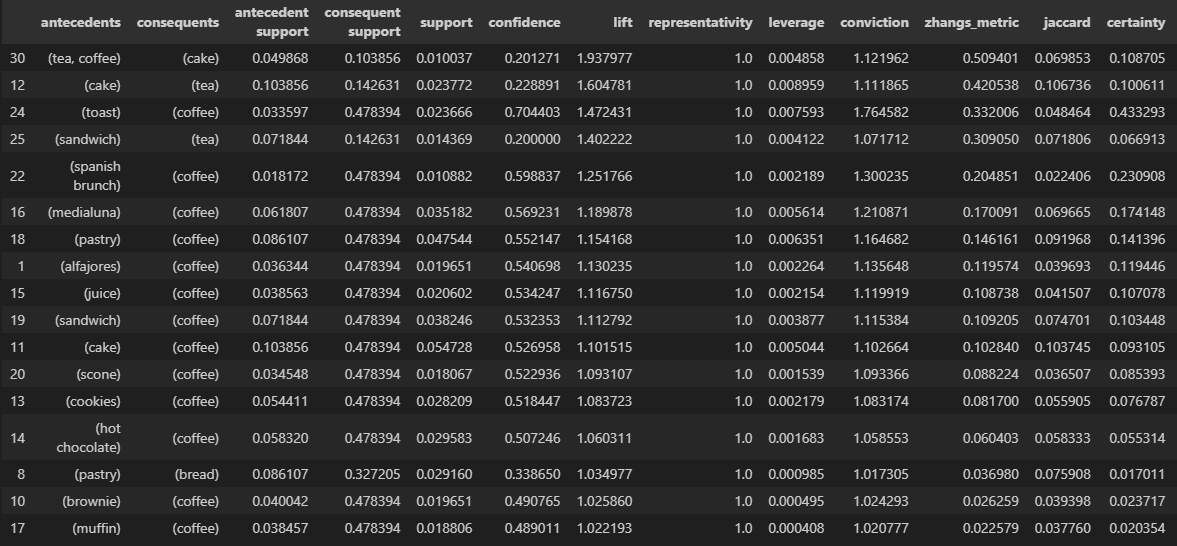
**Nhận xét:**

* **Mặt hàng coffee:** Luật có độ tin cậy cao nhất là (toast) → (coffee) (confidence 0.704403, lift 1.472431), cho thấy 70.4% giao dịch mua toast cũng mua coffee, với lift > 1 (1.47) xác nhận mối quan hệ tích cực, mạnh hơn so với ngẫu nhiên. Các mặt hàng như spanish brunch, medialuna, pastry cũng có xu hướng mạnh dẫn đến mua coffee (confidence 0.55–0.60, lift > 1).
* **Mặt hàng bread:** Các luật như (pastry) → (bread) (confidence 0.338650, lift 1.034977) và (alfajores) → (bread) (confidence 0.284884, lift 0.870657) cho thấy bread không có xu hướng được mua cùng các mặt hàng khác một cách nổi bật. Với lift gần 1 hoặc nhỏ hơn 1, mối liên kết giữa bread và các sản phẩm như pastry hay alfajores chỉ ở mức trung bình hoặc yếu, nghĩa là việc mua bread không phụ thuộc nhiều vào việc mua các mặt hàng này, và chúng thường không xuất hiện cùng nhau nhiều hơn so với kỳ vọng ngẫu nhiên.
* **Mặt hàng tea và cake:** Luật (cake) → (tea) (confidence 0.228891, lift 1.604781) và (tea, coffee) → (cake) (confidence 0.201271, lift 1.937977) có lift cao (> 1.5), chỉ ra mối quan hệ đáng kể, đặc biệt khi khách hàng mua tea và coffee, họ có xu hướng mua thêm cake.

➔ Một số luật như (bread) → (coffee) (confidence 0.275105, lift 0.575059) có lift < 1, cho thấy mối quan hệ yếu hơn so với ngẫu nhiên, dù bread và coffee đều phổ biến. Điều này nhấn mạnh rằng tần suất cao không đảm bảo mối quan hệ mạnh, cần dựa vào lift để đánh giá.

* Để đảm bảo chỉ giữ lại các luật kết hợp phản ánh mối quan hệ thực sự giữa các mặt hàng, ta sử dụng độ đo lift làm tiêu chí lọc chính. Chỉ số lift > 1 cho tất cả các luật cho thấy rằng giữa các sản phẩm tồn tại mối quan hệ mua cùng nhau mạnh mẽ hơn ngẫu nhiên. Ngược lại, các luật có lift ≤ 1 bị loại bỏ hoàn toàn, do chúng biểu thị mối quan hệ yếu hoặc các mặt hàng độc lập với nhau.

1. **Kết quả khi lọc lại với lift > 1**

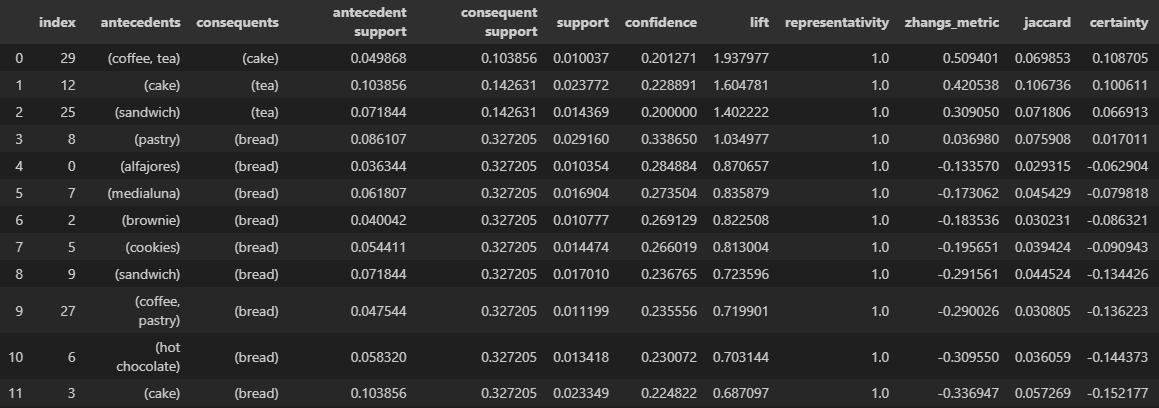


**Nhận xét:**

* Quy tắc có lift cao nhất: {tea, coffee} → {cake}, lift = 1.937977
* Khi khách hàng mua đồng thời {tea, coffee}, xác suất họ mua thêm {cake} tăng gấp 1.94 lần so với ngẫu nhiên (support = 0.010036, confidence = 0.201271). Điều này cho thấy sự kết hợp giữa {tea, coffee} và {cake} rất phổ biến, phù hợp để xây dựng các combo khuyến mãi.
* Đối với {toast} → {coffee}, lift = 1.472431
* Khách hàng mua {toast} có xác suất mua {coffee} cao hơn 1.47 lần so với ngẫu nhiên (support = 0.023666, confidence = 0.704403). Mối quan hệ này phản ánh thói quen tiêu dùng bữa sáng nhẹ kết hợp với {coffee}.
* Đối với {cake} → {tea}, lift = 1.604781
* Khách hàng mua {cake} có xu hướng mua thêm {tea} cao hơn 1.60 lần (support = 0.023772, confidence = 0.228891), cho thấy {tea} là đồ uống ưa thích khi dùng kèm {cake}.
* Các quy tắc khác đáng chú ý:
* {medialuna} → {coffee}, lift = 1.189878 và {pastry} → {coffee}, lift = 1.154166: Các mặt hàng bữa sáng nhẹ như {medialuna} và {pastry} thường được mua cùng {coffee}, với xác suất lần lượt cao hơn 1.19 và 1.15 lần.
* {hot chocolate} → {coffee}, lift = 1.060311: Sự kết hợp giữa {hot chocolate} và {coffee} cũng có mối liên hệ nhẹ (lift > 1), nhưng không quá mạnh.
* Các quy tắc với lift > 1 cho thấy {coffee} là mặt hàng chủ đạo, thường được mua cùng các món ăn nhẹ hoặc đồ uống khác.
* Vì {coffee} xuất hiện quá thường xuyên (47% giao dịch), nó có xu hướng xuất hiện trong hầu hết các luật liên kết, làm lu mờ các mối quan hệ giữa các mặt hàng khác. Điều này khiến các luật như {toast} → {coffee} hay {medialuna} → {coffee} trở nên kém ý nghĩa, vì {coffee} vốn đã là mặt hàng mà khách hàng có khả năng mua cao bất kể họ mua gì.
* Confidence cao không đồng nghĩa với một mối liên hệ mạnh, mà chỉ phản ánh tần suất cao của {coffee}.
* Lift không đủ để lọc các luật quan trọng: Mặc dù lift > 1 cho thấy có sự liên kết, nhưng giá trị lift của các luật liên quan đến {coffee} thường không đủ cao để chứng minh một mối quan hệ đặc biệt (ví dụ, lift của {hot chocolate} → {coffee} chỉ là 1.06).

➔ Do đó, để có cái nhìn rõ ràng hơn về các mối quan hệ thực sự giữa các mặt hàng khác, chúng ta nên loại bỏ các luật có {coffee} là consequent. Điều này giúp tập trung vào các quy tắc ít hiển nhiên hơn, chẳng hạn như {cake} → {tea} (lift = 1.604781) hoặc {tea, coffee} → {cake} (lift = 1.937977)

1. **Kết quả khi bỏ các luật có {coffee} là consequent**



**Nhận xét:**

* Sau khi loại bỏ các luật có {coffee} là consequent, mối quan hệ giữa các loại mặt hàng khác trở nên rõ ràng hơn, giúp làm nổi bật các xu hướng tiêu dùng không bị chi phối bởi sự phổ biến của {coffee}. Các quy tắc liên kết còn lại cho thấy những mối quan hệ đáng chú ý, đặc biệt giữa {tea}, {cake}, và các món ăn nhẹ như {sandwich}:
* **Luật kết hợp {tea, coffee} → {cake}:** Quy tắc này có support = 0.010036, confidence = 0.201271, và lift = 1.937977. Khi khách hàng mua đồng thời {tea, coffee}, xác suất họ mua thêm {cake} tăng gấp 1.94 lần so với ngẫu nhiên, cho thấy sự kết hợp giữa đồ uống nóng và {cake} rất phổ biến.
* **Luật kết hợp {cake} → {tea}:** Quy tắc này có support = 0.023772, confidence = 0.228891, và lift = 1.604781. Khoảng 22.9% khách hàng mua {cake} sẽ mua thêm {tea}, cao hơn 1.6 lần so với ngẫu nhiên. Điều này cho thấy {tea} là lựa chọn đồ uống ưa thích khi dùng kèm {cake}, thể hiện mối liên hệ mạnh mẽ giữa hai mặt hàng này.
* **Luật kết hợp {sandwich} → {tea}:** Với support = 0.014369, confidence = 0.200000, và lift = 1.402222, quy tắc này chỉ ra rằng 20% khách hàng mua {sandwich} sẽ mua thêm {tea}, cao hơn 1.4 lần so với ngẫu nhiên. Kết quả này cho thấy {tea} cũng là một lựa chọn đồ uống phổ biến khi khách hàng tiêu dùng các món ăn nhẹ như {sandwich}.

# KẾT LUẬN

## Kết luận chung

Trong nghiên cứu này, chúng em đã sử dụng thuật toán Apriori để khai phá các luật kết hợp giữa các sản phẩm từ bộ dữ liệu BreadBasket Bakery. Thuật toán Apriori là một công cụ mạnh mẽ để tìm ra các tập mục phổ biến và các luật kết hợp trong dữ liệu giao dịch. Để áp dụng thuật toán này, dữ liệu đã được xử lý và chuẩn hóa về dạng giao dịch phù hợp. Sau khi thiết lập các ngưỡng min\_support = 0.01 và min\_confidence = 0.2, nhóm đã tiến hành khai phá các luật kết hợp.

Kết quả phân tích cho thấy:

* Các sản phẩm như {toast}→ {coffee} (confidence 70.4%, lift 1.47) và {tea, coffee} → {cake} (lift 1.94) là những mối liên kết mạnh mẽ, phản ánh thói quen tiêu dùng điển hình của khách hàng.

Tuy nhiên, quá trình khai phá cũng chỉ ra rằng các chỉ số support và confidence không đủ để đánh giá độ mạnh của một luật, đặc biệt khi có sự xuất hiện thường xuyên của các sản phẩm phổ biến như “coffee”. Do đó, việc bổ sung tiêu chí lift > 1 là rất cần thiết để lọc ra các mối liên kết thực sự có giá trị.

Nhằm làm nổi bật các xu hướng tiêu dùng ít bị ảnh hưởng bởi coffee, nhóm đã lọc bỏ các luật có hậu tố (consequent) là coffee. Kết quả sau lọc giúp làm rõ những mối liên hệ khác giữa các sản phẩm, ví dụ:

* {cake} → {tea} (confidence 22.9%, lift 1.60): khách mua cake có xác suất cao hơn 1.6 lần để mua thêm tea.
* {sandwich} → {tea} (confidence 20%, lift 1.40): cho thấy tea cũng là lựa chọn phổ biến khi mua các món ăn nhẹ như sandwich.

Tổng kết lại, bài báo cáo này đã cho thấy việc áp dụng thuật toán Apriori không chỉ giúp khám phá các mối quan hệ sản phẩm hữu ích, mà còn nhấn mạnh tầm quan trọng của việc lựa chọn chỉ số đánh giá phù hợp (đặc biệt là *lift*) trong quá trình khai phá. Đồng thời, việc xử lý các mặt hàng phổ biến như coffee giúp mang lại cái nhìn rõ nét hơn về hành vi mua sắm thực tế của khách hàng.

## Hướng đi của đề tài

Trong nghiên cứu này, chúng em đã áp dụng thuật toán Apriori để khai phá các luật kết hợp giữa các sản phẩm trong bộ dữ liệu BreadBasket Bakery. Để phát triển và mở rộng hơn nữa đề tài, nhóm đề xuất một số hướng nghiên cứu tiếp theo như sau:

* **So sánh hiệu quả các thuật toán khai phá luật kết hợp**
* Nghiên cứu sẽ được mở rộng bằng việc thử nghiệm thêm các thuật toán khác như FP-Growth để so sánh hiệu quả về thời gian xử lý và chất lượng luật kết hợp so với Apriori. Qua đó, lựa chọn phương pháp tối ưu phù hợp với đặc điểm bộ dữ liệu.
* **Phân tích xu hướng mua sắm theo thời gian**
* Đề tài sẽ phân tích các xu hướng mua sắm theo khung thời gian như giờ trong ngày, ngày trong tuần và các mùa vụ để phát hiện những biến động và đặc điểm tiêu dùng đặc thù, từ đó hỗ trợ việc dự báo và xây dựng chiến lược bán hàng phù hợp.
* **Phân nhóm khách hàng và cá nhân hóa gợi ý sản phẩm**
* Sử dụng các kỹ thuật phân cụm để nhóm khách hàng dựa trên hành vi mua hàng và khai thác các luật kết hợp trong từng nhóm, nhằm xây dựng hệ thống gợi ý sản phẩm cá nhân hóa, nâng cao trải nghiệm người dùng và tăng doanh thu.

## Hạn chế

### Giới hạn về phạm vi dữ liệu

Bộ dữ liệu The Bread Basket chỉ ghi lại các giao dịch trong khoảng thời gian từ tháng 10/2016 đến tháng 4/2017, do đó kết quả phân tích có tính chất tạm thời và chưa thể khái quát được xu hướng tiêu dùng lâu dài hoặc biến động theo các mùa vụ khác nhau.

### Ảnh hưởng của sản phẩm phổ biến

Sản phẩm như coffee chiếm tỷ lệ giao dịch rất cao, làm xuất hiện nhiều luật kết hợp liên quan đến mặt hàng này, gây ảnh hưởng đến việc phát hiện các luật quan trọng khác và có thể làm lu mờ những mối quan hệ tiềm năng giữa các sản phẩm ít phổ biến hơn.

Link video thuyết trình: <https://www.youtube.com/watch?v=S7LNAXvdB0s>

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | 28 4 2025. [Online]. Available: https://viblo.asia/p/research-on-association-rules-EbNVQANBVvR. |
| [2] | X. Hong. [Online]. Available: https://ongxuanhong.wordpress.com/2015/08/23/khai-thac-luat-tap-pho-bien-frequent-itemsets-voi-thuat-toan-apriori/. [Accessed 2 5 2025]. |
| [3] | [Online]. Available: https://viblo.asia/p/khai-pha-du-lieu-va-lop-bai-toan-khai-thac-cac-tap-pho-bien-p2-m68Z0W06KkG. [Accessed 2 5 2025]. |