**MỤC LỤC**

[BÁO CÁO PROJECT 2](#_Toc163050833)

[I. Giới thiệu bài toán 2](#_Toc163050834)

[Đặt vấn đề 2](#_Toc163050835)

[Giới thiệu 2](#_Toc163050836)

[II. Nghiên cứu liên quan 4](#_Toc163050837)

[III. Phương pháp 7](#_Toc163050838)

[3.1. Các kỹ thuật nền tảng 7](#_Toc163050839)

[3.1.1 Luật kết hợp (Association rules) 7](#_Toc163050840)

[3.1.2 Độ hỗ trợ (Support) 8](#_Toc163050841)

[3.1.3 Độ tin cậy (Confidence) 8](#_Toc163050842)

[3.1.4 Độ nâng (Lift) 8](#_Toc163050843)

[3.1.5 Độ nhận thức (Conviction) 9](#_Toc163050844)

[3.2. Đề xuất phương pháp 10](#_Toc163050845)

[3.2.1 Phương pháp tìm kiếm theo chiều rộng - Thuật toán Apriori (1994) 10](#_Toc163050846)

[3.2.2 Phương pháp phát triển mẫu – Thuật toán FP-Growth (2000) 11](#_Toc163050847)

[IV. Thực nghiệm 14](#_Toc163050848)

[4.2. Tiền xử lý dữ liệu 14](#_Toc163050849)

[4.3. Các độ đo đánh giá hiệu năng 17](#_Toc163050850)

[4.3.1 Thang đo Support 17](#_Toc163050851)

[4.3.2 Thang đo Confidence 18](#_Toc163050852)

[4.3.3. Thang đo Lift 18](#_Toc163050853)

[4.3.4. Thang đo Conviction: 20](#_Toc163050854)

[4.4. Các tham số và môi trường cài đặt 21](#_Toc163050855)

[4.5. Các phương pháp cơ sở 25](#_Toc163050856)

[4.6. Phân tích, so sánh các kết quả 27](#_Toc163050857)

[V. Kết luận 30](#_Toc163050858)

# BÁO CÁO PROJECT

# I. Giới thiệu bài toán

## Đặt vấn đề

Hành vi khách hàng là một lĩnh vực nghiên cứu đầy hấp dẫn, tập trung vào việc giải mã cách thức mà cá nhân, nhóm hay tổ chức lựa chọn, sử dụng và loại bỏ sản phẩm, dịch vụ, ý tưởng hay trải nghiệm để thỏa mãn nhu cầu và mong muốn của họ. Hiểu được hành vi khách hàng là chìa khóa giúp doanh nghiệp đưa ra những quyết định kinh doanh sáng suốt, từ đó phát triển sản phẩm phù hợp, tăng hiệu quả marketing, cải thiện trải nghiệm khách hàng và ultimately, gia tăng doanh thu và lợi nhuận. Nghiên cứu hành vi khách hàng cung cấp cho doanh nghiệp một bức tranh toàn cảnh về hành trình mua sắm của khách hàng, từ khi nhận thức nhu cầu cho đến khi sử dụng và đánh giá sản phẩm. Bằng cách áp dụng các phương pháp nghiên cứu như khảo sát, phỏng vấn, nhóm tập trung hay phân tích dữ liệu, doanh nghiệp có thể thu thập thông tin quý giá về sở thích, nhu cầu, động lực và hành vi của khách hàng. Nghiên cứu hành vi khách hàng mang lại nhiều lợi ích cho doanh nghiệp như phát triển sản phẩm và dịch vụ phù hợp, tăng hiệu quả marketing, cải thiện trải nghiệm khách hàng và tăng doanh thu và lợi nhuận. Nói tóm lại, hành vi khách hàng là một lĩnh vực nghiên cứu đầy tiềm năng và mang lại lợi ích to lớn cho doanh nghiệp. Doanh nghiệp nên đầu tư vào việc nghiên cứu hành vi khách hàng để đưa ra những quyết định kinh doanh sáng suốt, gia tăng lợi thế cạnh tranh và gặt hái thành công trong thị trường đầy biến động hiện nay.

## Giới thiệu

Ngành thương mại điện tử đang phát triển mạnh mẽ trên toàn cầu. Trước thách thức và cơ hội từ sự bùng nổ này, nghiên cứu được tập trung vào phân tích giỏ mua hàng của một cửa hàng E-commerce hướng tới việc phục vụ khách hàng toàn cầu, nhưng chủ yếu tập trung vào thị trường nội địa Vương quốc Anh. Mục tiêu chính của nghiên cứu là khám phá những mẫu mua sắm tiềm ẩn trong dữ liệu giao dịch, từ đó hiểu rõ hơn về hành vi của khách hàng và đưa ra các chiến lược kinh doanh có hiệu quả cao.

Để đạt được mục tiêu này, nghiên cứu sử dụng hai thuật toán phổ biến trong việc khai phá dữ liệu là Apriori và FP-Growth. Apriori được sử dụng để xác định những nhóm sản phẩm thường được mua cùng nhau, trong khi FP-Growth sử dụng cấu trúc cây FP để biểu diễn dữ liệu giao dịch, vượt trội hơn Apriori trong việc phát hiện những mẫu mua sắm phức tạp.

Dữ liệu nghiên cứu là tập dữ liệu giao dịch của một cửa hàng E-commerce, chứa thông tin chi tiết về sản phẩm, số lượng, giá cả, thời điểm mua sắm và thông tin khách hàng nếu có. Sự rộng lớn của dữ liệu này giúp nghiên cứu phản ánh chân thực và đa dạng của hành vi mua sắm trực tuyến. Việc hiểu rõ hơn về hành vi của khách hàng giúp xác định những sản phẩm thường đi kèm, theo dõi xu hướng mua sắm theo thời gian và khu vực. Từ đó, chiến lược marketing có thể được cá nhân hóa tốt hơn, với khả năng đề xuất sản phẩm phù hợp và nhắm mục tiêu quảng cáo một cách hiệu quả.

Ngoài ra, việc tối ưu hóa quản lý hàng tồn kho dựa trên dự đoán nhu cầu sản phẩm từ dữ liệu cũng là một ưu điểm quan trọng. Điều này giúp cửa hàng E-commerce đảm bảo nguồn cung sản phẩm luôn đáp ứng đúng nhu cầu của khách hàng, đồng thời giảm thiểu tình trạng tồn kho dư thừa.

Cuối cùng, nghiên cứu này không chỉ hỗ trợ việc quản lý hiệu quả cho cửa hàng E-commerce mà còn là nguồn thông tin quý giá để phát triển sản phẩm mới. Bằng cách xác định chính xác nhu cầu thị trường, cửa hàng có thể đưa ra các sản phẩm mới phù hợp với mong muốn và yêu cầu của khách hàng.

Tóm lại, phân tích giỏ mua hàng với Apriori và FP-Growth không chỉ là một công cụ hiệu quả để hiểu rõ hơn hành vi của khách hàng mà còn là cơ sở để xây dựng những chiến lược kinh doanh đầy hiệu quả cho cửa hàng E-commerce. Mặc dù nghiên cứu này tập trung vào thị trường Vương quốc Anh, những phương pháp và kết quả có thể áp dụng cho nhiều thị trường khác với những điều chỉnh và thích ứng phù hợp.

# II. Nghiên cứu liên quan

Phân tích Giỏ Hàng (Market Basket Analysis) là kỹ thuật khai thác dữ liệu giúp các doanh nghiệp hiểu rõ hơn về cách khách hàng mua sắm, từ đó giúp các doanh nghiệp mở khóa tiềm năng trong việc hiểu rõ hơn về hành vi mua sắm của khách hàng. Kết quả phân tích này chủ yếu được sử dụng để cải thiện doanh số bán hàng của các cửa hàng bán nhiều sản phẩm bằng cách nâng cao vị trí sắp xếp sản phẩm dựa trên thói quen mua sắm của người tiêu dùng [1]. Kỹ thuật này hoạt động dựa trên việc khám phá các mối liên quan giữa các sản phẩm thường được mua cùng nhau trong giỏ hàng của khách hàng. Hiện nay, Phân tích Giỏ Hàng ngày càng được ứng dụng rộng rãi, mở ra cánh cửa cho vô số kỹ thuật và thuật toán khai thác dữ liệu độc đáo. Trong số đó, nổi bật nhất là những cái tên sau: Thuật toán Apriori, AIS, Thuật toán SETM, FP-Growth,…. Tuy ngày càng có nhiều kỹ thuật khai thác dữ liệu mới được áp dụng trong Phân tích Giỏ Hàng, hai phương pháp được áp dụng nhiều nhất vẫn là Thuật toán Apriori và FP-Growth.

Thuật toán Apriori là một công cụ phổ biến và được biết đến rộng rãi trong khai thác luật kết hợp, đặc biệt là trong Phân tích Giỏ Hàng. Các thuật toán như AI và SETM cho rằng Apriori có độ chính xác cao. Apriori giúp tìm các tập sản phẩm thường xuyên (frequent itemsets) trong giao dịch và xác định mối quan hệ (association rules) giữa các sản phẩm này. Thuật toán Apriori được áp dụng nhiều trong các nghiên cứu về nhiều lĩnh vực khác nhau, có thể kể đến như: Phân tích Giỏ Hàng trên nền tảng Thương mại Điện tử [6], Phân tích Giỏ Hàng tại Cửa hàng Tiện lợi [2], Phân tích Giỏ Hàng trong Y tế [3], Phân tích Giỏ Hàng tại Cửa hàng Máy tính [4],… Thuật toán Apriori đóng vai trò nền tảng trong các hệ thống cơ sở dữ liệu quan hệ, được sử dụng chủ yếu để khám phá các mẫu sản phẩm thường được mua cùng nhau thông qua quy tắc kết hợp (association rule learning) [1]. Thuật toán hướng tới mục tiêu chính là xác định các mặt hàng xuất hiện thường xuyên trong cơ sở dữ liệu, sau đó kết hợp tăng dần để hình thành các tập sản phẩm lớn hơn, với điều kiện các sự kết hợp này xuất hiện nhất quán trong dữ liệu. Các tập sản phẩm thường xuyên được Thuật toán Apriori xác định chính là cơ sở cho việc tạo ra các luật kết hợp, đóng vai trò quan trọng trong việc phát hiện các xu hướng chi phối trong cơ sở dữ liệu. Điều này đặc biệt hữu ích cho các tác vụ như Phân tích Giỏ Hàng. Thuật toán Apriori xem xét mỗi giao dịch như một tập sản phẩm riêng biệt. Kỹ thuật này sử dụng phương pháp "từ dưới lên", bắt đầu với các sản phẩm xuất hiện thường xuyên riêng lẻ, rồi kết hợp chúng một cách có hệ thống bằng cách tham chiếu thông tin từ dữ liệu. Quá trình kết hợp dừng khi không tìm được sự kết hợp mới nào có ý nghĩa [1].

Thuật toán Apriori được biết đến là thuật toán lâu đời và đơn giản nhất, tuy nhiên lại tốn nhiều thời gian tính toán và dung lượng bộ nhớ để tìm kiếm các tập sản phẩm (itemsets) [6]. Điều này là do thuật toán phải quét dữ liệu lặp đi lặp lại nhiều lần. Thuật toán này được sử dụng để xác định mối tương quan giữa các mặt hàng mà khách hàng quan tâm, lưu trữ trong cơ sở dữ liệu. Sau khi thu được các tập sản phẩm được mua thường xuyên, một quy tắc được chọn và sau đó các nhà nghiên cứu phân tích, so sánh về tốc độ thực thi, quá trình tạo quy tắc và độ chính xác của quy tắc từ thuật toán [6]. Phân tích giỏ hàng của dữ liệu giao dịch bán hàng sử dụng các giai đoạn của Thuật toán Apriori sẽ tạo ra thông tin dưới dạng các quy tắc kết hợp với giá trị hỗ trợ tối thiểu là 50% và độ tin cậy tối thiểu là 60% [7]. Đây có thể là tài liệu tham khảo trong việc sắp xếp các mặt hàng trên kệ cửa hàng bằng cách tham khảo sự kết hợp các mặt hàng thường được người tiêu dùng mua đồng thời. Thuật toán Apriori là một công cụ hữu ích để khai thác tri thức từ dữ liệu, đặc biệt là trong lĩnh vực phân tích giỏ hàng (market basket analysis). Tuy nhiên, cần lưu ý đến những hạn chế của thuật toán này khi áp dụng cho các tập dữ liệu lớn hoặc thưa thớt.

Vì Thuật toán Apriori tốn nhiều thời gian tính toán và dung lượng bộ nhớ để tìm kiếm các tập sản phẩm nên nhiều nhà phân tích đã lựa chọn Thuật toán FP-Growth (Frequent Pattern Growth) vì nó xử lý dữ liệu lớn nhanh hơn [9]. FP-Growth là thuật toán khai thác dữ liệu được sử dụng để tìm các tập mục thường xuyên xuất hiện trong tập dữ liệu giao dịch. Thuật toán này có 3 bước chính: Xây dựng cây FP (FP-Tree Construction), Phát triển FP-Growth (FP-Growth Generation), Tập mục thường xuyên (Frequent Item Set) [9]. FP-Growth là thuật toán được sử dụng để tìm kiếm các tập dữ liệu thường xuyên xuất hiện trong tập dữ liệu giao dịch. So với Apriori, FP-Growth có nhiều ưu điểm nổi trội như hiệu quả xử lý dữ liệu lớn nhanh hơn và sử dụng ít bộ nhớ hơn. Tuy nhiên, thuật toán này cũng có một số nhược điểm nhất định như gặp khó khăn trong việc xử lý dữ liệu thưa thớt và yêu cầu nhiều thời gian để xây dựng cây FP-Tree [8]. FP-Growth là thuật toán khai thác dữ liệu hiệu quả, đặc biệt hữu ích cho việc xử lý tập dữ liệu giao dịch lớn. Tuy nhiên, cần lưu ý đến những hạn chế của thuật toán này khi áp dụng cho các tập dữ liệu thưa thớt.

Ngoài việc áp dụng đơn lẻ hai Thuật toán trên, nhiều nhà nghiên cứu còn ứng dụng cả 2 Thuật toán vào việc Phân tích giỏ hàng. Một công ty dược phẩm ở Indonesia đã áp dụng các kỹ thuật khai thác dữ liệu, đặc biệt là khai thác quy tắc kết hợp sử dụng thuật toán Apriori và FP-GROWTH, cho phân tích giỏ hàng trên PT. XYZ [12]. Kết quả cho thấy: FP-GROWTH có thời gian thực thi nhanh nhất là 84.655 giây. Tuy nhiên, việc sử dụng bộ nhớ cho thuật toán Apriori là thấp nhất ở mức 482,32 MiB, với mức tăng: 0,21 MiB. Đối với các quy tắc được tạo, cả hai thuật toán (Apriori và FP-GROWTH) đều tạo ra số lượng quy tắc, giá trị hỗ trợ, độ tin cậy, độ nâng, độ hỗ trợ Bi, độ tin cậy Bi và độ nâng Bi tương tự nhau. Từ đó, các nhà nghiên cứu kết luận: Apriori được đề xuất cho các tập dữ liệu bán hàng nếu việc sử dụng bộ nhớ và tính dễ triển khai là quan trọng. Tuy nhiên, nếu tốc độ thực thi và khối lượng dữ liệu lớn được xem xét, FP-GROWTH là lựa chọn tốt hơn vì thời gian thực thi nhanh hơn cho khối lượng dữ liệu lớn [12]. Nghiên cứu này khảo sát vai trò của khoa học dữ liệu trong việc hiểu hành vi khách hàng và nâng cao doanh số bán hàng, tập trung cụ thể vào việc áp dụng thuật toán Apriori và FP-Growth tại cửa hàng bán lẻ Deli Point ở Labuan Bajo [14]. Nghiên cứu làm sáng tỏ tác động của "dữ liệu rác" đối với việc phân tích dữ liệu giao dịch, nhấn mạnh sự cần thiết của các quy trình làm sạch dữ liệu mạnh mẽ để đảm bảo kết quả chính xác. Sử dụng thuật toán FP-Growth nhanh hơn, nghiên cứu đã phân tích hiệu quả các mô hình mua hàng của khách hàng để xác định sự kết hợp sản phẩm tối ưu nhằm cải thiện doanh số bán hàng. Nghiên cứu này không chỉ đóng góp những hiểu biết có giá trị về hành vi của người tiêu dùng bán lẻ và các chiến lược đặt sản phẩm hiệu quả mà còn nhấn mạnh vai trò chuyển đổi của khoa học dữ liệu trong việc tối ưu hóa doanh số và thúc đẩy khả năng cạnh tranh trong lĩnh vực bán lẻ [15]. Nhìn chung, việc kết hợp Apriori và FP-Growth có thể mang lại nhiều lợi ích cho phân tích giỏ hàng, nhưng cũng cần cân nhắc đến những thách thức đi kèm. Chúng ta cần lựa chọn phương pháp phù hợp phụ thuộc vào mục tiêu nghiên cứu, kích thước dữ liệu và tài nguyên sẵn có.

# III. Phương pháp

## 3.1. Các kỹ thuật nền tảng

Phân tích giỏ thị trường (MBA) là một phương pháp khai thác dữ liệu được sử dụng để tìm ra các sản phẩm mà khách hàng có xu hướng mua cũng nhau. Doanh nghiệp sử dụng các hệ thống POS (Point of sale) để quản lý quá trình giao dịch, dữ liệu hàng hóa và bán hàng. Những dữ liệu này sẽ được làm sạch, xử lý thông qua các thuật toán dựa trên Quy tắc Kết hợp để tạo ra các thông tin có giá trị giúp doanh nghiệp đưa ra quyết định.

### 3.1.1 Luật kết hợp (Association rules)

Khai thác luật kết hợp là cơ sở để phân tích giỏ hàng thị trường. Nhiệm vụ chính là tìm ra các mối liên hệ giữa các mặt hàng khác nhau được khách hàng chọn mua thông qua việc phân tích sự xuất hiện đồng thời của các mặt hàng đó trong hóa đơn mua tại cửa hàng. Hiểu đơn giản, luật kết hợp giúp doanh nghiệp nắm được khi khách hàng mua sắm nhóm các sản phẩm A thì sẽ có khả năng mua sản phẩm Y với một xác suất nào đó. Có thể minh họa như sau:

X => Y where X ⸦ I, Y ⸦ I and X ∩ Y=0

Ví dụ, quy tắc {chees,onion}={pizza} được tìm thấy trong việc bán hàng siêu thị sẽ chỉ ra rằng nếu khách hàng mua pho mát và hành cùng nhau đồng nghĩa rằng khách này có thể sẽ làm pizza từ chúng và sẽ thường mua cococola uống kèm. Thông tin này có thể được sử dụng làm cơ sở cho quyết định trong các hoạt động thị trường, ví dụ như tạo combo sản phẩm hoặc giảm giá. Quy tắc kết hợp không chỉ hỗ trợ cho việc phân tích giỏ thị trường mà nó còn đóng một vai trò quan trọng phát hiện xâm nhập, tin sinh học và khai thác web.

### 3.1.2 Độ hỗ trợ (Support)

**Định nghĩa:** Chỉ số support thể hiện tần suất xuất hiện của một sản phẩm X hoặc nhóm sản phẩm X và Y trong tổng số các giỏ hàng, được tính như sau:

Conviction thực chất đánh giá luật kết hợp theo hướng ngược lại với chỉ số Lift.

**Công thức tính Supp:**

Supp(X=>Y)=

Giả sử để tính tần suất xuất hiện của nhóm sản phẩm gồm gà rán và cocacola trong giao dịch của khách hàng, chúng ta sẽ lấy số các giao dịch có mua cả gà rán và cocacola chia tổng số giao dịch.

Support thể hiện độ phổ biến của tập các mặt hàng trong tập dữ liệu.

### 3.1.3 Độ tin cậy (Confidence)

**Định nghĩa:**

Chỉ số confidence thể hiện tỷ lệ % số lần Y xuất hiện trong những giỏ hàng có nhóm sản phẩm X.

Conviction thực chất đánh giá luật kết hợp theo hướng ngược lại với chỉ số Lift.

**Công thức tính Conf:**

Conf(X=>Y)=

Ví dụ: gà rán xuất hiện 6/10 giỏ hàng, gà rán và coca xuất hiện đồng thời 4/10 giỏ hàng. Khi đó:

Conf(Chicken=>Coca) = = 2/3 ~ 67%\

### 3.1.4 Độ nâng (Lift)

**Định nghĩa:**

Trong khai thác dữ liệu và phân tích mối quan hệ giữa các mục (item) trong dữ liệu, kỹ thuật đo lường Lift thường được sử dụng để đo lường mức độ phụ thuộc giữa các mục. Lift thường được sử dụng trong việc xác định mối quan hệ giữa các mục trong quy tắc kết hợp (association rules), nơi bạn đánh giá mức độ ảnh hưởng của một mục (hoặc tập hợp các mục) lên một mục khác**.**

**Công thức tính Lift:**

Công thức tính của kỹ thuật được biểu diễn như sau:

Trong đó:

): xác suất xuất hiện của sự kết hợp của X và Y trong tất cả các giao dịch.

là xác suất của tập hợp X xuất hiện trong tất cả các giao dịch. Nó được gọi là "support" của X.

là xác suất của tập hợp Y xuất hiện trong tất cả các giao dịch. Nó được gọi là "support" của Y.

**Ý nghĩa của Lift:**

**Lift > 1:** Điều này cho thấy sự phụ thuộc giữa các thuộc tính X và Y. X xuất hiện thường xuyên hơn cùng với Y so với trường hợp chúng độc lập. Nói cách khác, quy tắc kết hợp là đáng chú ý và có thể hữu ích trong việc dự đoán sự xuất hiện của Y dựa trên sự hiện diện của X.

**Lift = 1:** Điều này cho thấy X và Y độc lập với nhau. Sự xuất hiện của X không ảnh hưởng đến khả năng xuất hiện của Y. Quy tắc kết hợp không hữu ích cho việc dự đoán.

**Lift < 1:** Điều này cho thấy sự đối lập giữa các thuộc tính X và Y. X ít xuất hiện cùng với Y so với trường hợp chúng độc lập. Quy tắc kết hợp có thể hữu ích trong việc dự đoán sự *không*  xuất hiện của Y dựa trên sự hiện diện của X.

### 3.1.5 Độ nhận thức (Conviction)

**Định nghĩa:**

Conviction là một độ đo được sử dụng trong quy tắc kết hợp (association rules) để đo lường tần suất mong muốn X xảy ra mà không cần Y. Nó thường được sử dụng cùng với các độ đo khác như support và confidence để phân tích mối quan hệ giữa các mục. Conviction đo lường xác suất xuất hiện mục X mà không có mục Y nếu chúng độc lập so với xác suất xuất hiện mục X và không có mục Y trong thực tế. Nếu Conviction càng lớn thì Y càng phụ thuộc mạnh mẽ vào X.

Conviction thực chất đánh giá luật kết hợp theo hướng ngược lại với chỉ số Lift.

**Công thức tính Conviction:**

Trong đó:

là tần suất không xuất hiện của mục *Y* trong tất cả các giao dịch.

là mức độ tin cậy rằng mục *Y* không phụ thuộc vào mục *X* trong một quy tắc kết hợp.

**Ý nghĩa:**

**Giá trị Conviction > 1:** Cho biết Y phụ thuộc mạnh mẽ vào X.

**Giá trị Conviction < 1:** Cho biết Y ít phụ thuộc vào X.

**Giá trị Conviction = 1**: Cho biết X và Y độc lập.

## 3.2. Đề xuất phương pháp

### 3.2.1 Phương pháp tìm kiếm theo chiều rộng - Thuật toán Apriori (1994)

Thuật toán Apriori là thuật toán được R. Agrawal và R. Srikant giới thiệu vào năm 1994, nhằm mục đích tìm các tập phổ biến chạy trên một tập hợp dữ liệu. Trong phép lặp k, bạn sẽ tìm thấy tất cả các tập mục có k mục, được gọi là tập mục k. Đặc điểm chính của thuật toán apriori là tất cả các tập con của tập phổ biến cũng là thành viên của tập phổ biến.Thuật toán tăng trưởng Apriori là một trong những thuật toán quan trọng được sử dụng trong việc phát hiện các mẫu kết hợp (frequent itemsets) trong dữ liệu, đặc biệt là trong lĩnh vực khai phá dữ liệu (data mining) và phân tích dữ liệu. Thuật toán này được đặt tên là Apriori vì nó dựa trên một nguyên tắc quan trọng có tên là "Apriori principle".

Thuật toán Apriori cũng là một kỹ thuật khai thác dữ liệu để tìm ra các quy tắc kết hợp giữa một tổ hợp các mục. Thường được sử dụng để phân tích việc mua hàng trong siêu thị với mục đích tìm hiểu khả năng người tiêu dùng mua một loại mặt hàng này cùng với các loại hàng hóa khác, để công ty có thể tổ chức các chiến lược quảng cáo và bố trí các hàng hóa liên quan đặt gần nhau. cùng nhau và xác định chiến lược định giá khuyến mại cho một số mặt hàng nhất định có mối liên hệ với nhau.

Thuật toán apriori được chia thành nhiều giai đoạn gọi là tường thuật hoặc truyền.

Gồm 2 giai đoạn:

* Tìm tập L gồm tất cả các Itemset phổ biến ( có độ hộ trợ không nhỏ hơn min supp)
* Phát sinh các luật từ phổ biến I

Cách chạy của thuật toán:

Giai đoạn 1:

* Chạy duyệt Database: C1: {1}: 2; {2}: 3; {3}: 3; {4}: 1, {5}: 3. Sau khi đếm xong, duyệt dựa trên minsup=50%=2

=> loại {4} do 1<2

L1: {1}: 2; {2}: 3; {3}: 3;, {5}: 3

* Phát sinh C2 bằng cách ghét đôi lần lượt không trùng tập với nhau:

C2: {1,2}, {1,3}, {1,5}, {2,3}, {2,5}, {3,5}

duyệt database lần 2: C2: {1,2}: 1, {1,3}: 2, {1,5}: 1, {2,3}: 2, {2,5}: 3, {3,5}: 2

* Sau khi đếm xong, duyệt tiếp ms=2

=> L2: {1,3}: 2, {2,3}: 2, {2,5}: 3, {3,5}: 2

* Tương tự như vậy:

C3: {1,2,3}:1 , {1,3,5}: 1, { 2,3,5}: 2

=> L3: { 2,3,5}: 2

### 3.2.2 Phương pháp phát triển mẫu – Thuật toán FP-Growth (2000)

FP-Growth là một thuật toán được sử dụng để xác định tập dữ liệu thường xuyên xuất hiện (tập mục thường xuyên) trong một tập dữ liệu. Thuật toán này sử dụng khái niệm xây dựng cây, thường được gọi là FPTree, trong tìm kiếm các tập mục thường xuyên thay vì sử dụng tạo ứng viên được thực hiện trong thuật toán Apriori.

Thuật toán FP-Growth được chia thành ba bước chính như sau:

1. Giai đoạn tạo cơ sở mẫu có điều kiện

Cơ sở mẫu có điều kiện là dữ liệu con chứa đường dẫn tiền tố (đường dẫn ban đầu) và mẫu hậu tố (mẫu hậu tố). Tạo cơ sở mẫu có điều kiện thu được thông qua Cây FP đã được xây dựng trước đây.

Đầu tiên, thuật toán duyệt CSDL lần 1 để tính độ hỗ trợ (supp) của từng mục. Chúng ta cần đưa ra độ hỗ trợ tối thiểu để loại những mục không đủ độ hỗ trợ. Các mục còn lại sẽ được sắp xếp theo thứ tự giảm dần độ hỗ trợ vào 1 bảng để tiếp tục cất vào FP-Tree trong bước tiếp theo

Cụ thể, từ bảng dữ liệu bao gồm TID và Items sau:

|  |  |
| --- | --- |
| **TID** | **Items** |
| 1 | f , a, c, d, g, i, m, p |
| 2 | a, b, c, f, l, m, o |
| 3 | b, f, h, j, o |
| 4 | b, c, k, s, p |
| 5 | a, f, c, e, l, p, m, n |

Đầu tiên, tìm các item mức 1 có sup >=3 và sắp xếp theo thứ tự sup giảm dần sẽ được bảng các mẫu phổ biến như sau:

|  |  |
| --- | --- |
| **items** | **sup** |
| f | 4 |
| c | 4 |
| a | 3 |
| b | 3 |
| m | 3 |
| p | 3 |

Sau đó, ta sắp xếp các mục phổ biến mức 1 vừa tìm được theo thứ tự giảm dần về tần số trong mỗi giao dịch.

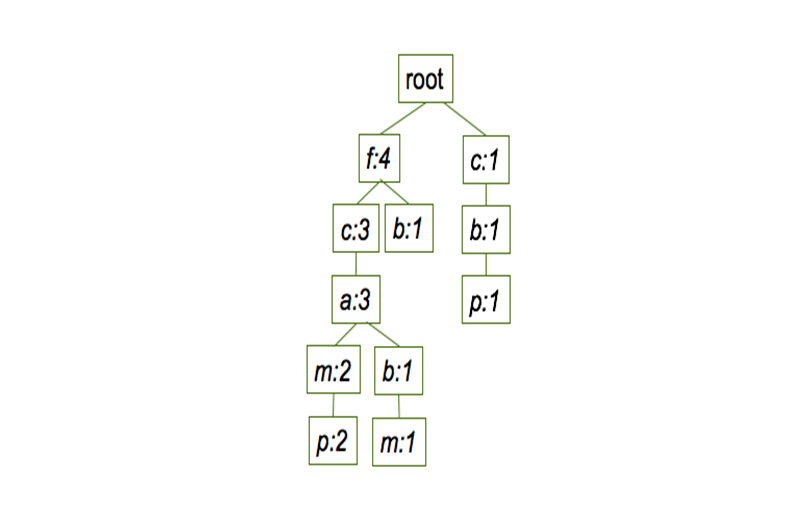
|  |  |  |
| --- | --- | --- |
| **TID** | **items** | **items phổ biến** |
| 1 | f, a, c, d,f , i, p | f, c, a, m, p |
| 2 | a, b, c, f, l, m, o | f, c, a, b, m |
| 3 | b, f, h, j, o | f, b |
| 4 | b, c, k, s, p | c, b, p |
| 5 | a, f, c, e, l, p, m, n | f, c, a, m, p |

2. Giai đoạn tạo FP-Tree có điều kiện

Ở giai đoạn này, số lượng hỗ trợ của từng mục trong mỗi cơ sở mẫu điều kiện được cộng lại, sau đó mọi mục có nhiều hỗ trợ hơn lớn hoặc bằng số lượng hỗ trợ tối thiểu sẽ được tạo bằng FP-Tree có điều kiện.

Với mỗi giao dịch được xây dựng lại từ bước 1, chúng ta sẽ xây dựng một đường đi tương ứng trong FP-Tree bằng cách:

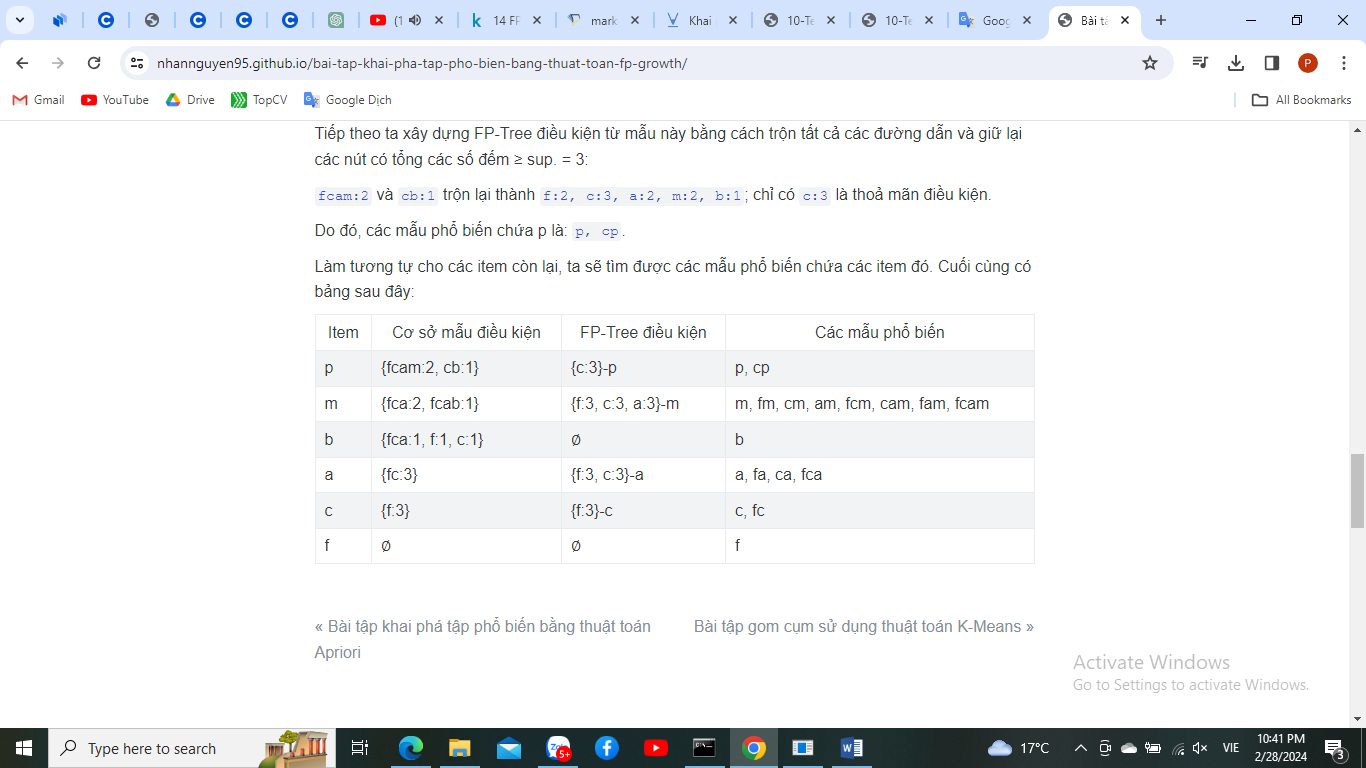
* Nếu một đường đi có tiền tố chung tổn tại, tăng tần số của các nút trên đường đi và nối thêm hậu tố (ví dụ TID 1,2,3,5 có chung item f nên tần số sẽ là 4 và nối thêm các nhánh bên dưới theo item chung tiếp theo)
* Ngược lại, tạo một nhánh mới (TID 4 không có item f nên tạo nhánh bắt đầu bằng item phổ biến nhất là c)

Ta được cây FP sau:

3. Giai đoạn tìm kiếm các mẫu phổ biến

Ở bước này, duyệt từng mục phổ biến theo thứ tự tăng dần của tần số, xây dựng cơ sở mẫu điều kiện và các cây FP-tree có điều kiện tương ứng với nó. Trong ví dụ trên, bắt đầu từ item p, cơ sở mẫu điều kiện của nó là tất cả các đường đi tiền tố của cây FP-Tree khi duyệt từ nút gốc Root đến nút p (fcam:2 và cb:1) với số theo sau là tần suất xuất hiện của nút p tương ứng với mỗi tiền tố.

Sau đó, xây dựng cây FP-Tree có điều kiện bằng cách trộn tất cả các đường đi và chỉ giữ lại các nút có tần số >=3. Cụ thể, đối với item p, trộn đường đi fcam:2 và cb:1 sẽ thu được {f:2,c:3,a:2,m:2,b:1}, chỉ giữ lại nút có tần số >=3 là c:3. Do đó, các mẫu phổ biến chứa p là: p và cp.

Tương tự với các item còn lại, ta sẽ tìm được các mẫu phổ biến chứa các item đó. Cuối cùng ra được bảng sau đây:

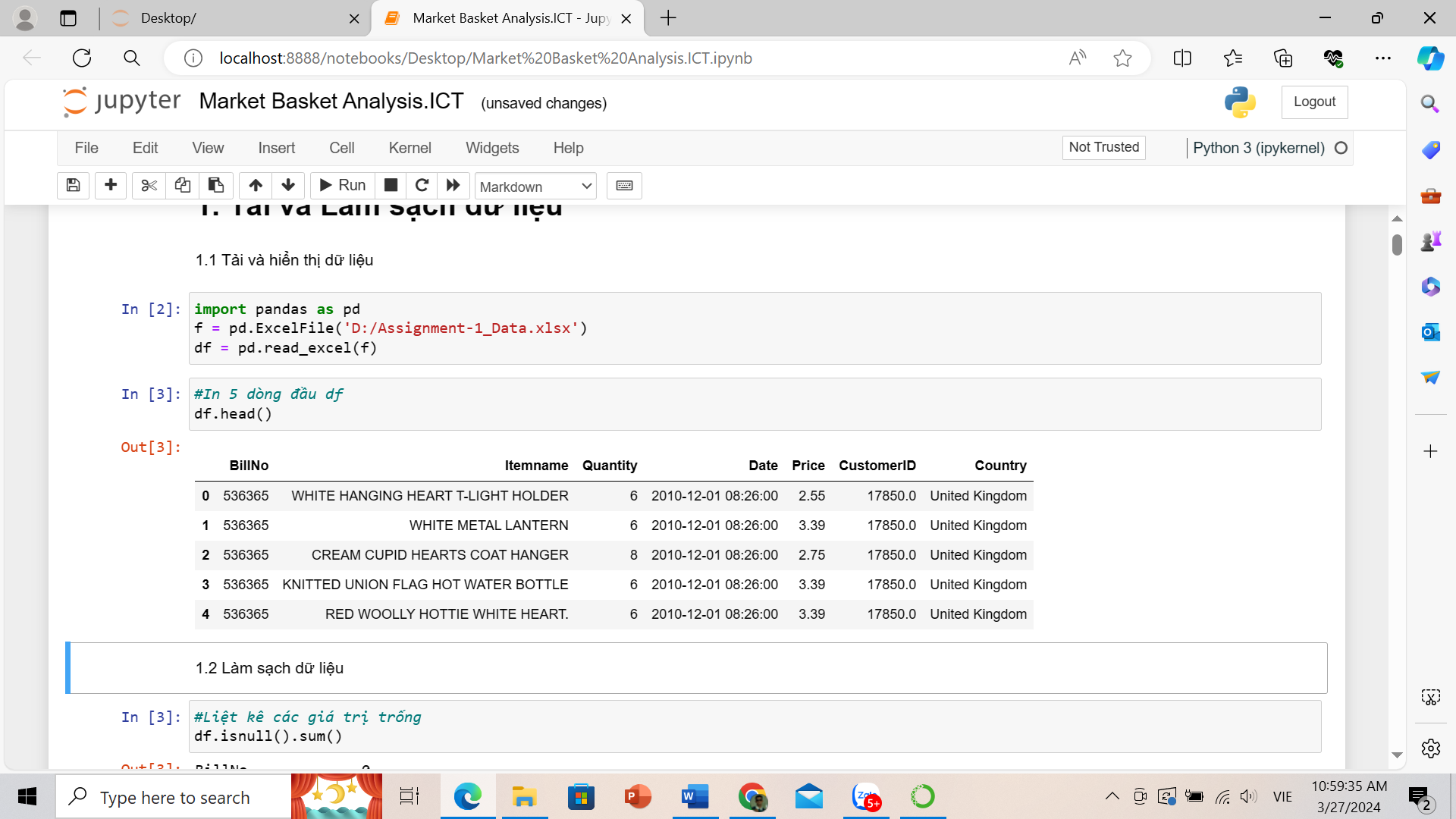
Kết quả thuật toán cho ra bảng gồm các mẫu phổ biến tương ứng theo từng item. Doanh nghiệp có thể sử dụng kết quả trên để đưa ra các chiến lược kinh doanh.

# IV. Thực nghiệm

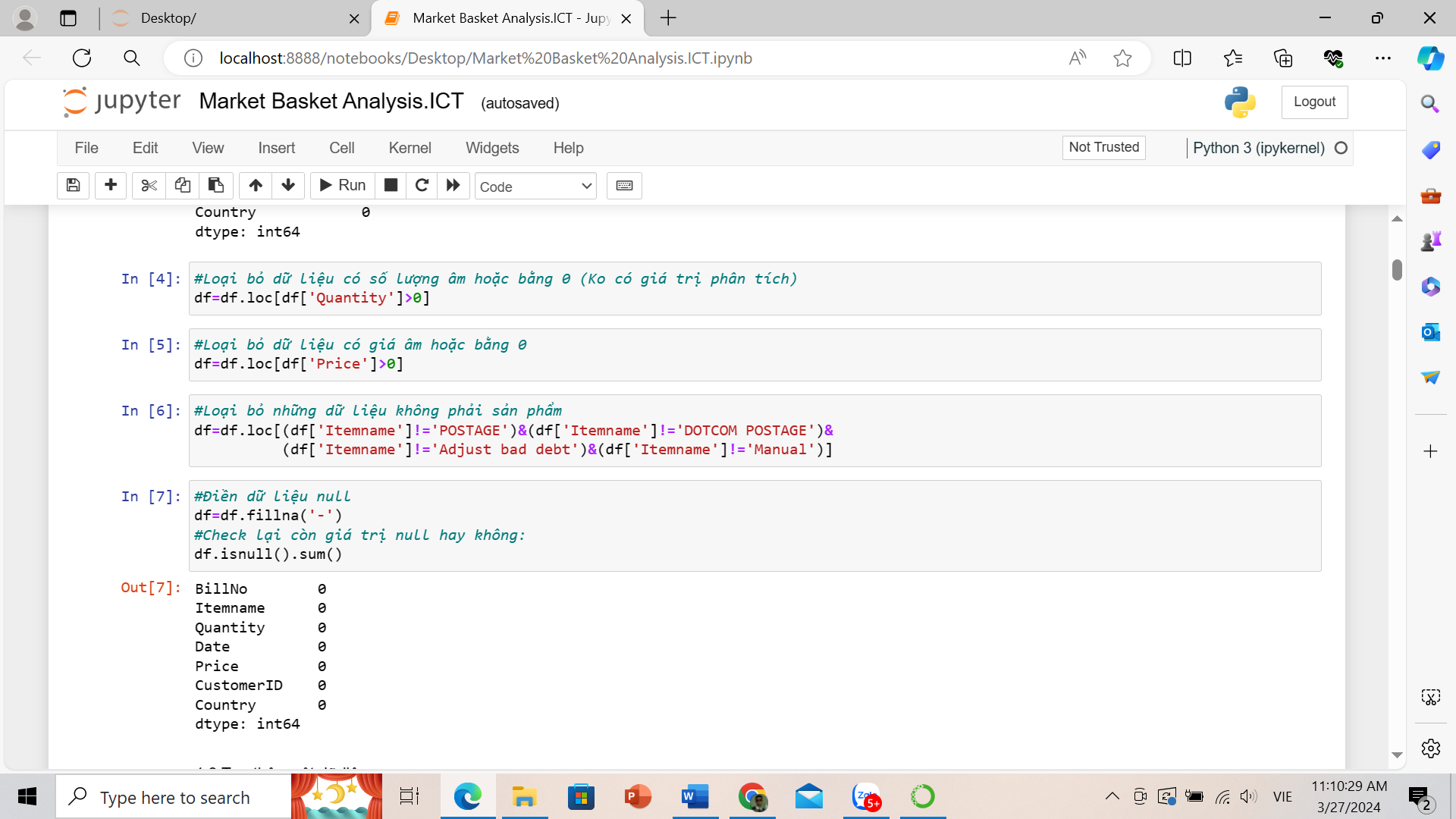
## 4.2. Tiền xử lý dữ liệu

Dữ liệu hiện tại đang có là dữ liệu thô. Tiến hành sử dụng jupyter notebook để xử lý dữ liệu.

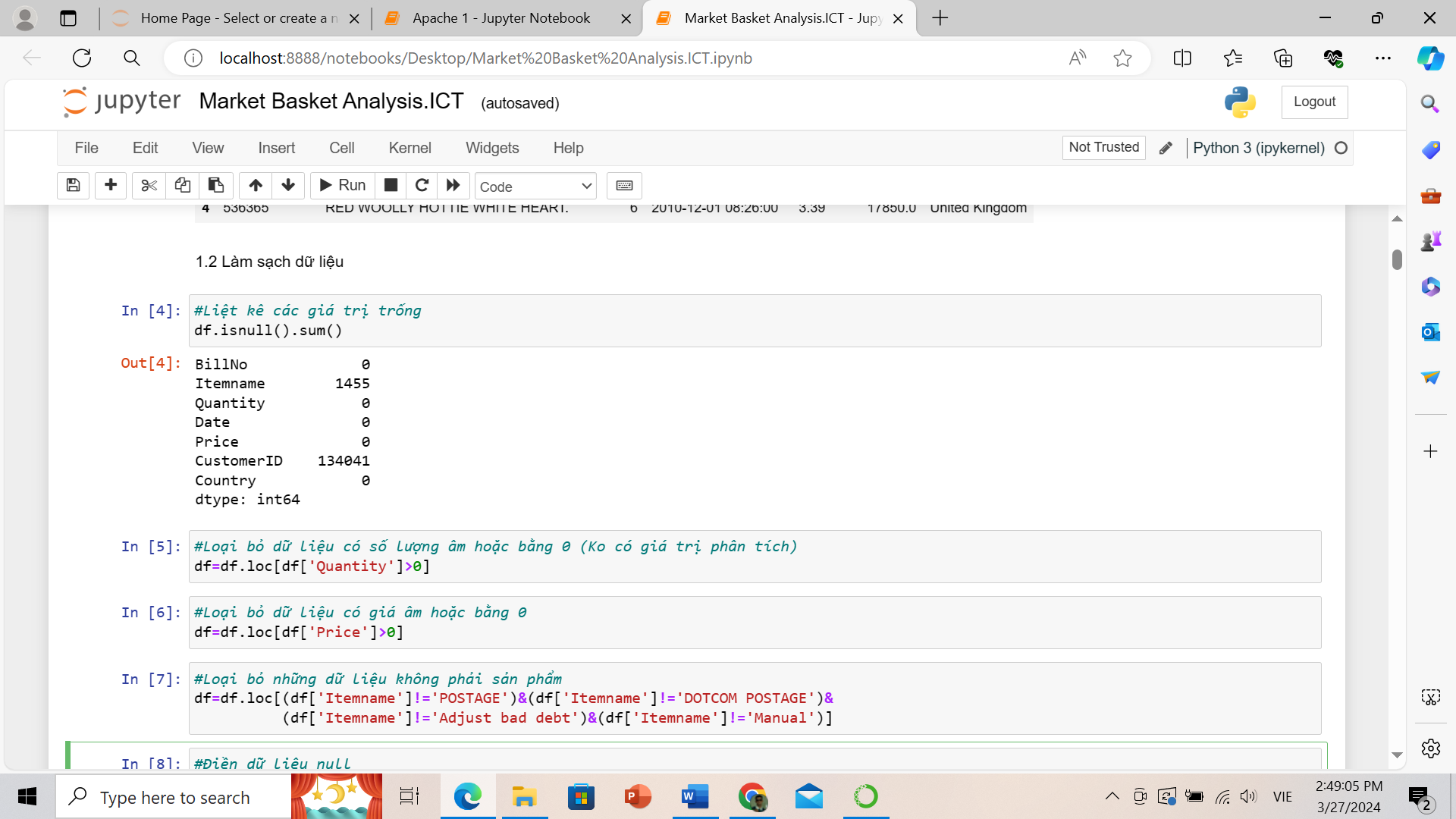
* Tiến hành tải và hiển thị dữ liệu: Đây là bước đầu tiên trong tiến trình tiền xử lý dữ liệu. Việc nhìn qua dữ liệu (có thể là 5 dòng đầu) giúp hình dung được cơ bản cấu trúc của dữ liệu, thông tin về dữ liệu

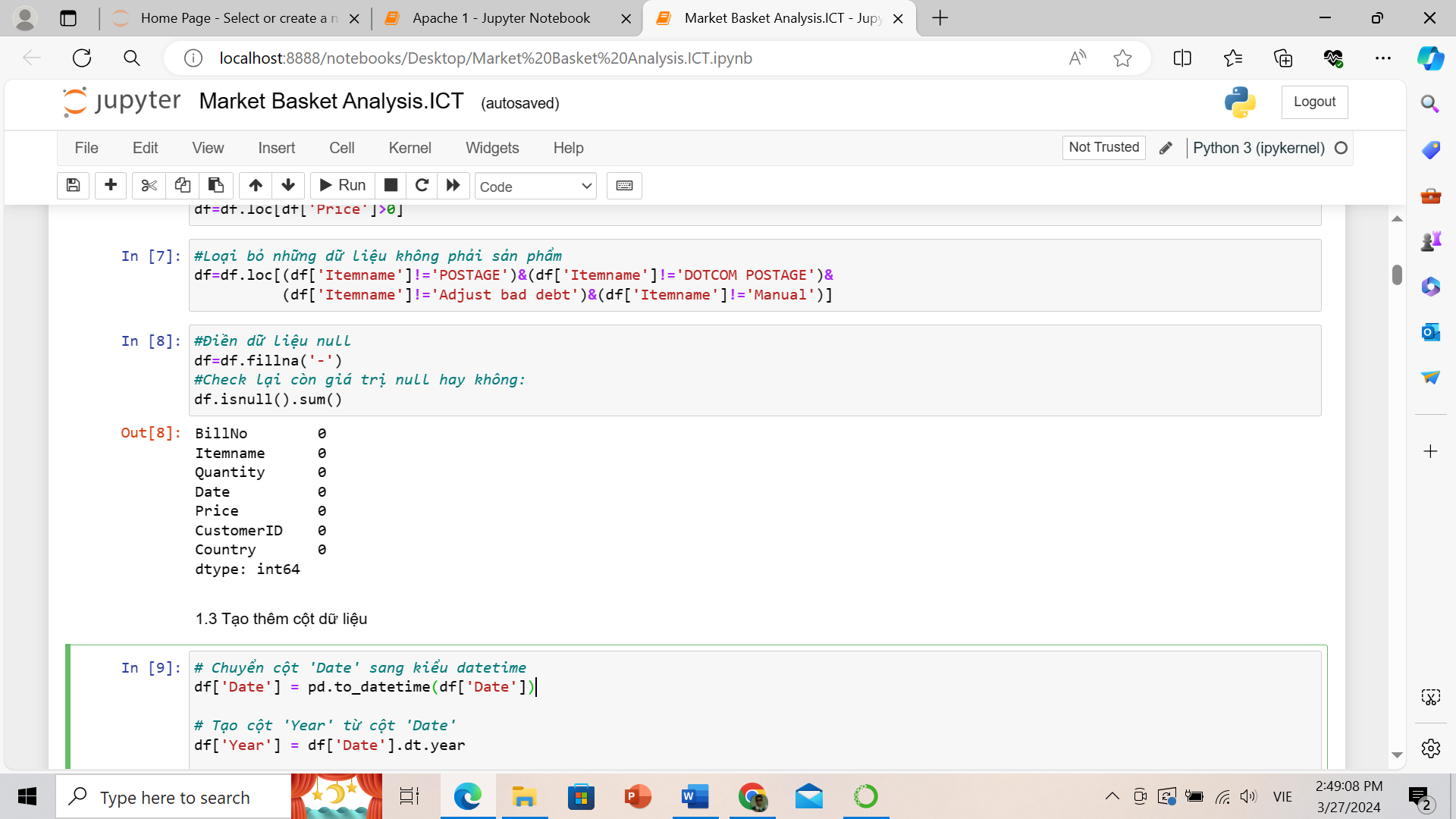
****

* Loại bỏ dữ liệu có số lượng âm hoặc bằng 0 (Ko có giá trị phân tích) và những dữ liệu không phải sản phẩm.

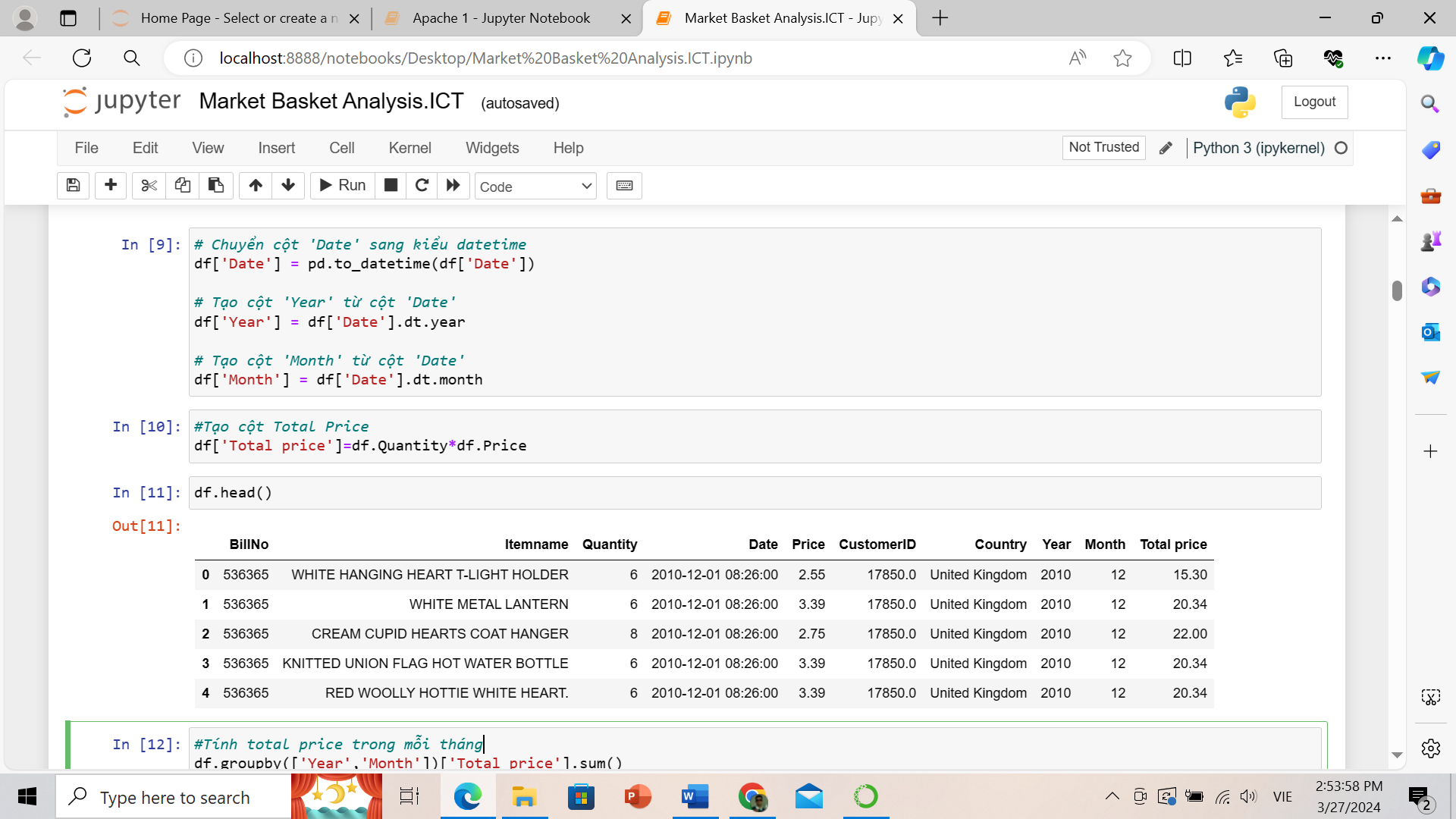
****

* Điền các dữ liệu còn trống

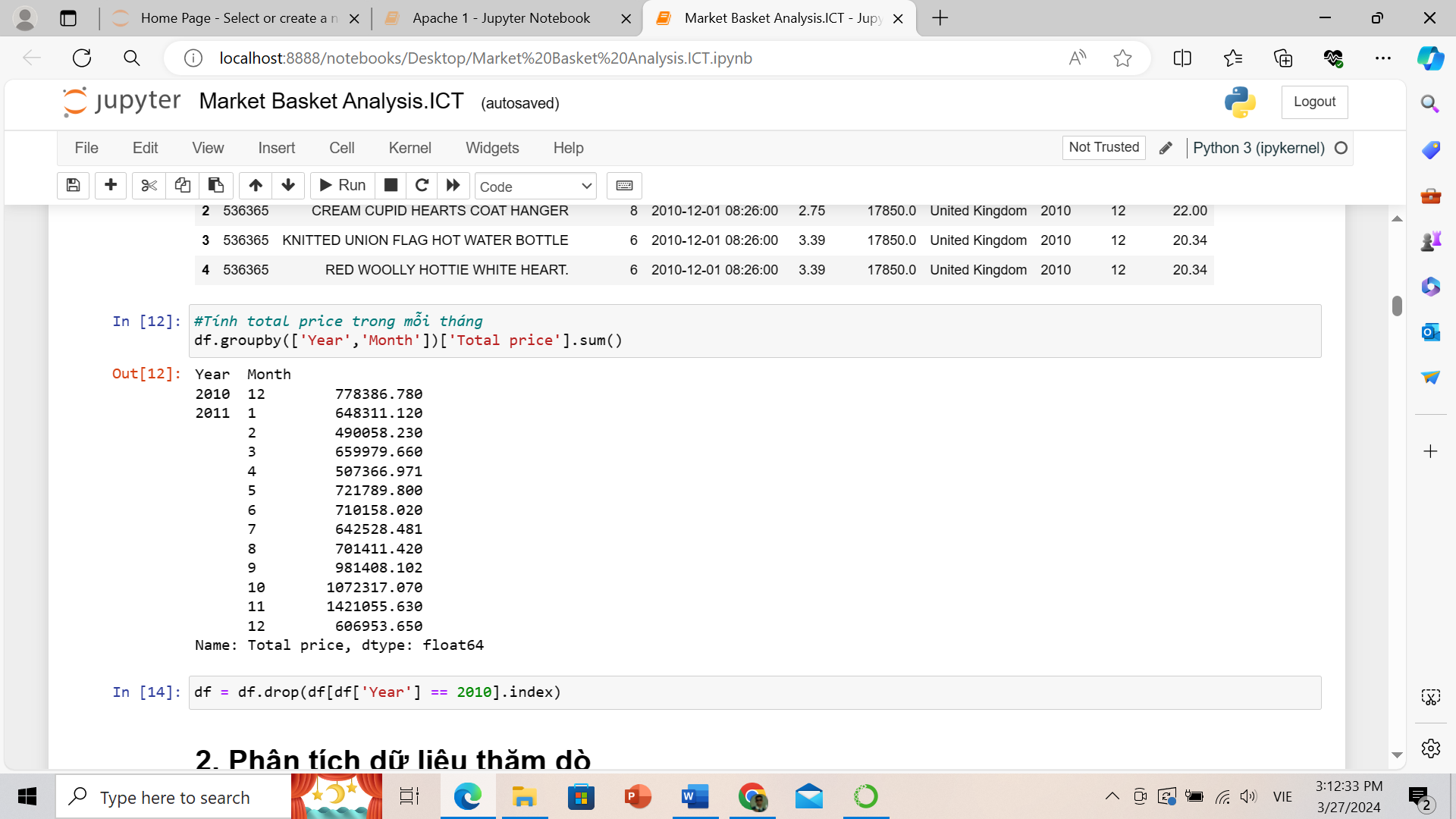
****

****

* Chuyển cột “Date” sang kiểu “datetime”. Sau đó tạo thêm các cột “Year” và “Month” từ cột “Date”

****

* Tính doanh thu theo từng tháng. Nhận thấy năm 2010 chỉ có dữ liệu của tháng 12 nên xóa dữ liệu năm 2010

****

## 4.3. Các độ đo đánh giá hiệu năng

### 4.3.1 Thang đo Support

Theo các chỉ số trong bài toán “Market Basket Analyst”, chúng ta có kết quả của thang đo Confident cho 1 vài trường hợp sau:

*Alarm Clock Bakelike RedAlarm Clock Bakelike Green*

= 0.052938

0.049845

*Jumbo Bag Red Retrospot Jumbo Shopper Vintage Red Paisley*

0.034845

= 0.106237

0.060103

*Lunch Bag Pink Polkadot Lunch Bag Red Retrospot*

0.030567

= 0.055309

0.079433

### 4.3.2 Thang đo Confidence

Theo các chỉ số trong bài toán “Market Basket Analyst”, chúng ta có kết quả của thang đo Confidence cho 1 vài trường hợp sau:

*Alarm Clock Bakelike RedAlarm Clock Bakelike Green*

= 0.052938

0.049845

*RG* = =0.61636

*Jumbo Bag Red Retrospot Jumbo Shopper Vintage Red Paisley*

0.034845

= 0.106237

0.060103

*RP* =

*Lunch Bag Pink Polkadot Lunch Bag Red Retrospot*

0.030567

= 0.055309

0.079433

PR = =0.55266

### 4.3.3. Thang đo Lift

Theo các chỉ số trong bài toán “Market Basket Analyst”, chúng ta có kết quả của thang đo Lift cho 1 vài trường hợp sau:

*Alarm Clock Bakelike RedAlarm Clock Bakelike Green*

= 0.052938

0.049845

*RG* = 12.365410>1

* **Mức độ liên quan:** Mức độ liên quan giữa hai thuộc tính “Alarm Clock Bakelike Red” và “Alarm Clock Bakelike Green” cao.
* **Ý nghĩa:**

**Có sự phụ thuộc:** "Alarm Clock Bakelike Red" và "Alarm Clock Bakelike Green" có ảnh hưởng đến nhau. Khách hàng mua sản phẩm “Alarm Clock Bakelike Red” cũng có khả năng mua sản phẩm “Alarm Clock Bakelike Green”.

*Jumbo Bag Red Retrospot* *Jumbo Shopper Vintage Red Paisley*

0.034845

= 0.106237

0.060103

*RP* = 5.457225 > 1

**Mức độ liên quan:** Mức độ liên quan giữa hai thuộc tính “Jumbo Bag Red Retrospot” và “Jumbo Shopper Vintage Red Paisley” cao.

* **Ý nghĩa:**

**Có sự phụ thuộc:** "Jumbo Bag Red Retrospot" và "Jumbo Shopper Vintage Red Paisley" có ảnh hưởng đến nhau. "Jumbo Bag Red Retrospot" xuất hiện thường xuyên cùng với "Jumbo Shopper Vintage Red Paisley" so với trường hợp hai thuộc tính này độc lập.

Khách hàng mua “Jumbo Bag Red Retrospot” có xu hướng mua “Jumbo Shopper Vintage Red Paisley” cao hơn so với khách hàng mua lẻ mỗi sản phẩm.

*Lunch Bag Pink Polkadot* *Lunch Bag Red Retrospot*

0.030567

= 0.055309

0.079433

*PR* = 6.957514 > 1

* **Mức độ liên quan:** Mức độ liên quan giữa hai thuộc tính “ Lunch Bag Pink Polkadot” và “Lunch Bag Red Retrospot” cao.
* **Ý nghĩa:**

**Có sự phụ thuộc:** "Lunch Bag Pink Polkadot" và "Lunch Bag Red Retrospot" có ảnh hưởng đến nhau. "Lunch Bag Pink Polkadot" xuất hiện thường xuyên cùng với " Lunch Bag Red Retrospot" nhiều hơn so với trường hợp hai thuộc tính này độc lập.

### 4.3.4. Thang đo Conviction:

Theo các chỉ số trong bài toán “Market Basket Analyst”, chúng ta có kết quả của thang đo Conviction cho 1 vài trường hợp sau:

1.:

= 1 0.947835

= 1 0.249743

**Ý nghĩa:** Conviction > 1 chứng tỏ sản phẩm “Roses Regency Teacup And Saucer” phụ thuộc mạnh mẽ vào sản phẩm "Green Regency Teacup And Saucer".

2.:

= 1 1.055309

= 1 0.615185

* **Ý nghĩa:** Conviction > 1 chứng tỏ sản phẩm “Lunch Bag Red Retrospot” và “” có ảnh hưởng tới nhau, khách hàng mua “Lunch Bag Red Retrospot” có khả năng họ có mua sản phẩm “” .

## 4.4. Các tham số và môi trường cài đặt

Môi trường cài đặt cho bài toán "Market basket analysis" thường bao gồm các công cụ và thư viện phổ biến được sử dụng cho xử lý dữ liệu và khai thác quy tắc kết hợp từ tập dữ liệu mua sắm. Dưới đây là các thành phần cần thiết cho bài toán này:

Ngôn ngữ lập trình: Nhóm sử dụng ngôn ngữ Python vì ngôn ngữ này cung cấp các thư viện mạnh mẽ cho xử lý dữ liệu và phân tích.

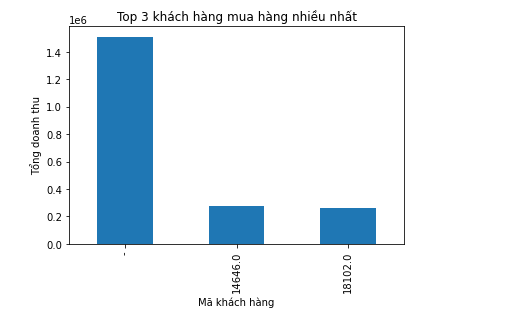
Thư viện phân tích: Trong Python, chúng ta có thể sử dụng hàm read\_excel () từ thư viện pandas để đọc file dữ liệu.



Tiếp theo chúng ta cần tải các thư viện cần thiết cho việc phân tích dữ liệu và trực quan hóa dữ liệu bằng các biểu đồ như: matplotlib, seaborn, geopandas, mlxtend.

Thư viện Matplotlib được sử dụng để trực quan hóa dữ liệu, bao gồm biểu đồ đường, biểu đồ cột, biểu đồ tròn,…

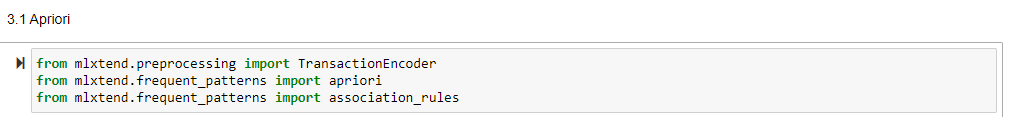


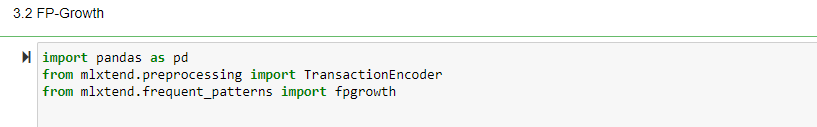


Thư viện Seaborn được cung cấp để vẽ biểu đồ box plot



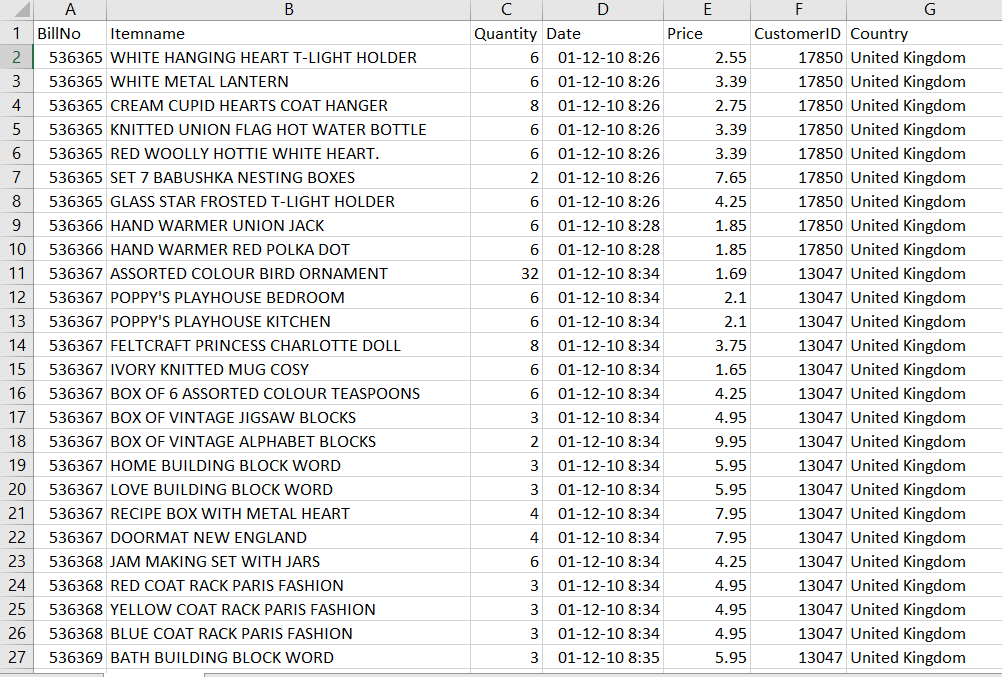
Thư viện MLxtend cung cấp các công cụ để tính toán các tập hợp mục thường xuyên và quy tắc kết hợp từ dữ liệu phổ biến, bao gồm Apriori và FP-Growth.





Các Tham số được sử dụng trong bài toán “Market basket Analyst”:

Tập dữ liệu (Dataset): Đây là file dữ liệu về các giao dịch mua sắm, trong đó mỗi giao dịch được biểu diễn bởi một danh sách gồm ID người mua, tên sản phẩm, đơn giá, số lượng sản phẩm, thời gian mua và nơi mua trong một lần giao dịch.

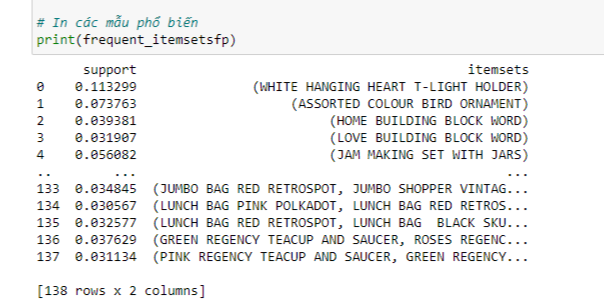


Tập hợp mục (Itemset): Là một tập hợp các mặt hàng có thể xuất hiện cùng nhau trong một giao dịch hoặc một nhóm giao dịch.

Tập hợp mục thường xuyên (Frequent Itemset): Là một tập hợp các mặt hàng xuất hiện cùng nhau trong một số lượng giao dịch lớn hơn hoặc bằng một ngưỡng được xác định trước.

Ví dụ, giả sử có một tập dữ liệu mua sắm và ta quan sát rằng trong 80% số lần mua, cà phê (coffee) và sữa (milk) được mua cùng nhau. Trong trường hợp này, tập hợp {cà phê, sữa} sẽ được coi là một frequent itemset. Tần suất của frequent itemset thường được đo bằng một độ đo gọi là "support".

Frequent itemset là một phần quan trọng trong quy trình khai thác quy tắc kết hợp trong market basket analysis, vì nó cho phép chúng ta xác định những mối quan hệ phổ biến giữa các mặt hàng trong các giao dịch mua sắm.



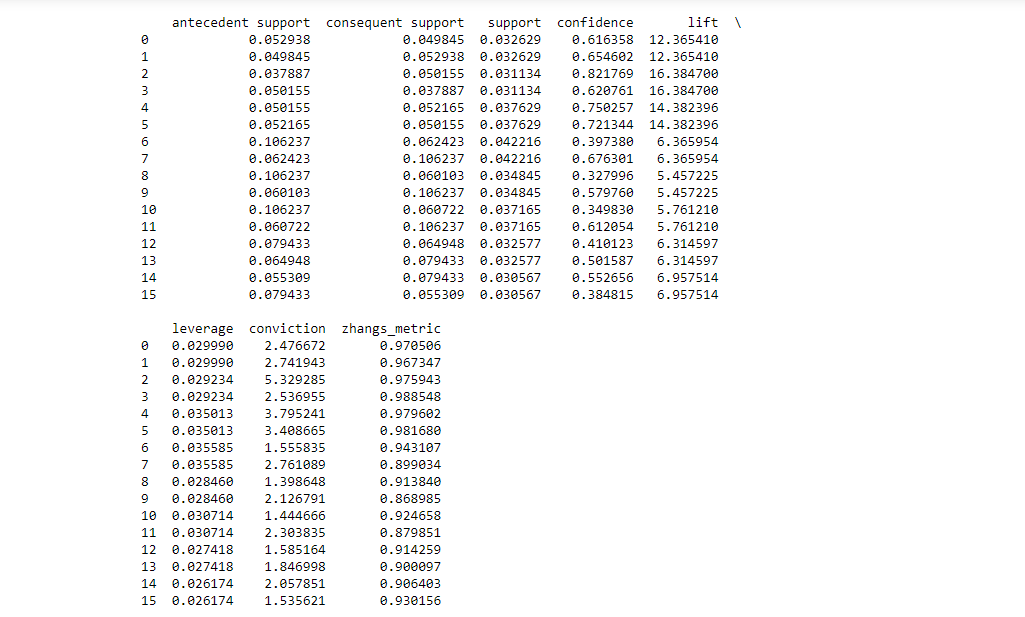
Quy tắc kết hợp (Association Rules): Là các quy tắc mô tả mối quan hệ giữa các mặt hàng trong tập dữ liệu, ví dụ như "nếu mặt hàng A được mua, thì mặt hàng B cũng thường được mua".

Support (Hỗ trợ): Là xác suất để một tập hợp mục xuất hiện trong tất cả các giao dịch, thường được đo lường bằng tỷ lệ giữa số lượng giao dịch chứa tập hợp mục và tổng số giao dịch.

Confidence (Độ tin cậy): Là xác suất để quy tắc kết hợp đúng khi mặt hàng A được mua, được tính bằng tỷ lệ giữa số lượng giao dịch chứa cả A và B và số lượng giao dịch chỉ chứa A.

Lift (Tăng cường): Đo lường mức độ mà mặt hàng B được mua cùng với mặt hàng A so với việc mặt hàng B được mua mà không cần A. Lift cao hơn 1 có nghĩa là quy tắc kết hợp có ý nghĩa.

Ngưỡng support và confidence: Là các ngưỡng được xác định trước để quyết định xem một tập hợp mục có thể coi là thường xuyên và một quy tắc kết hợp có thể coi là ý nghĩa.



## 4.5. Các phương pháp cơ sở

Luật kết hợp là một hướng quan trọng trong khai phá dữ liệu, giúp ta tìm được các mối liên hệ giữa các mục dữ liệu/thuộc tính (items) của dữ liệu, tìm các luật kết hợp “quý hiếm” và mang nhiều thông tin từ cơ sở dữ liệu tác nghiệp là một trong những hướng tiếp cận chính của lĩnh vực khai phá dữ liệu.

Các hướng tiếp cận trong khai phá luật kết hợp

* Luật kết hợp nhị phân
* Các items chỉ được quan tâm là có hay không xuất hiện trong cơ sở dữ liệu giao tác chứ không quan tâm về mức độ hay tần xuất xuất hiện
* Thuật giải Apriori
* Luật kết hợp có thuộc tính số và thuộc tính hạng mục
* Dùng các phương pháp rời rạc hóa chuyển về dạng nhị phân để có thể áp dụng các thuật giải đã có
* Luật kết hợp tiếp cận theo hướng tập thô
* Tìm kiếm luật kết hợp dựa trên lý thuyết tập thô
* Luật kết hợp nhiều mức
* Với cách tiếp cận luật kết hợp thế này sẽ tìm kiếm thêm những luật có dạng: mua máy tính PC 🡪 mua hệ điều hành Window.
* Luật kết hợp mờ
* Với những khó khăn gặp phải khi rời rạc hóa các thuộc tính số, luật kết hợp mờ khắc phục hạn chế đó và chuyển luật kết hợp về một dạng gần gũi hơn
* Luật kết hợp với thuộc tính được đánh trọng số
* Các thuộc tính được đánh trọng số theo mức độ xác định nào đó
* Nhờ vậy, thu được những luật “hiếm” (tức là có độ hỗ trợ thấp nhưng mang nhiều ý nghĩa)
* Luật kết hợp song song
* Nhu cầu song song hóa và xử lý phân tán là cần thiết vì kích thước dữ liệu ngày càng lớn

Quy tắc kết hợp

* Cho là một tập các mục (mặt hàng,...)
* Cho D là một tập các giao dịch mà mỗi giao dịch T là một tập các mục,
* Mỗi giao dịch có một mã định danh riêng gọi là TID
* Cho A là một tập các mục (mặt hàng). Một giao dịch T được gọi là chức A khi và chỉ khi A T
* Một luật kết hợp được diễn đạt dưới hình thức A B, với A I, B I,A B =
* Ý nghĩa: Khi xuất hiện A thì B cũng xuất hiện (với xác xuất nào đó)

VD: Mô tả cơ sở dữ liệu tác vụ A, C, D, T, W là các mục: là các tác vụ

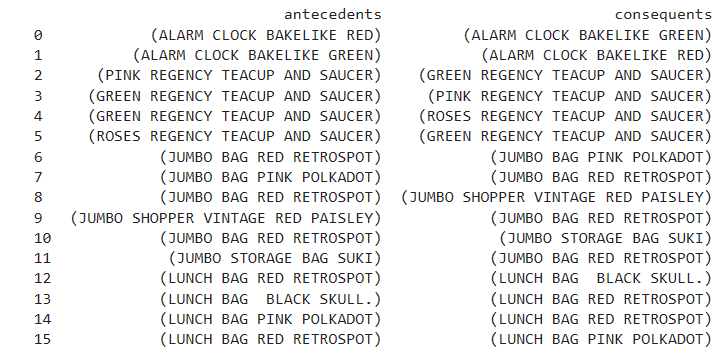
|  |  |
| --- | --- |
| Tập các tác vụ | Các mục dữ liệu |
| 1 | A C T W |
| 2 | C D W |
| 3 | A C T W |
| 4 | A C D W |
| 5 | A C D T W |
| 6 | C D T |

Mỗi giá trị của mục dữ liệu (Item) thể hiện thuộc tính xuất hiện (nhận giá trị 1) hay không xuất hiện (nhận giá trị 0) trong tác vụ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TID | A | C | D | T | W |
|  | 1 | 1 | 0 | 1 | 1 |
|  | 0 | 1 | 1 | 0 | 1 |
|  | 1 | 1 | 0 | 1 | 1 |
|  | 1 | 1 | 1 | 0 | 1 |
|  | 1 | 1 | 1 | 1 | 1 |
|  | 0 | 1 | 1 | 1 | 0 |

## 4.6. Phân tích, so sánh các kết quả

Phân tích, so sánh các kết quả:

Chúng tôi đã sử dụng 2 phương pháp là Apriori và FP-Growth để áp dụng với dữ liệu được cung cấp, kết quả thu được là như trên và như nhau.

Chúng tôi đã tìm được 15 nhóm sản phẩm phổ biến nhất dựa trên danh sách hóa đơn của khách hàng đã cung cấp với những hệ số như trên, cần chú ý tới chỉ số tin cậy là trên 60% và thước đo trên 86%.

Do kết quả là như nhau vì vậy có thể đánh giá 2 phương pháp trên đã áp dụng hiệu quả trên bộ dữ liệu là như nhau, sau đây là đánh giá so sánh về tính tối ưu giữa 2 phương pháp:

Dưới đây là một so sánh về ưu nhược điểm giữa hai đoạn mã:

Đối với phương pháp Apriori:

* Phương pháp:
* Sử dụng defaultdict(list) để nhóm các mục theo giá trị của cột “BillNo”
* Tạo một từ điển mới để lưu trữ dữ liệu nhóm
* Sử dụng zip để lặp qua dữ liệu, điều này phù hợp khi DataFrame ta tạo mới ở trên “df1” có cùng kích thước với cột “BillNo” và “Itemname”.
* Tận dụng các cấu trúc dữ liệu python cơ bản như danh sách và từ điển
* Ưu điểm:
* Dễ hiểu và dễ triển khai
* Phù hợp với các bộ dữ liệu nhỏ và đơn giản
* Không cần thư viên bên ngoài
* Nhược điểm:
* Khả năng mở rộng và xử lý dữ liệu bị giới hạn
* Không cung cấp các thuật toán tối ưu hóa cho việc tìm mẫu phổ biến và quy tắc kết hợp

Đối với phương pháp FP – growth:

* Phương pháp:
* Sử dụng groupby để nhóm các mục theo giá trị của cột “BillNo”
* Sử dụng mlxtend để thực hiện phân tích mẫu phổ biến và quy tắc kết hợp sử dụng FP – growth
* Có thể sử dụng cho các bộ dữ liệu lớn với hiệu suất cao hơn so với các cách tiếp cận cơ bản.
* Ưu điểm:
* Cung cấp một phương pháp hiệu quả để tìm kiếm mẫu phổ biến và quy tăvs kết hợp trong các bộ dữ liệu lớn.
* Sử dụng thuật toán FP – growth làm nền tảng, giúp tăng hiệu suất tính toán
* Nhược điểm:
* Yêu cầu sử dụng thư viện bên ngoài (mlxtend), đòi hỏi cài đặt thêm
* Có thể khó hiểu hơn, đặc biệt đối với những người mới bắt đầu hoặc không quen với các thuật toán khai thác dữ liệu.

Kết luận: Nếu bạn làm việc với các bộ dữ liệu nhỏ và đơn giản, hoặc chỉ cần nhóm các mục theo một cột cụ thể, thì phương pháp Apriori có thể là lựa chọn phụ hợp. Tuy nhiên, nếu bạn làm việc với các bộ dữ liệu lớn và cần tìm kiếm mẫu phổ biến và quy tắc kết một cách hiệu quả, thì FP – growth là lựa chọn tốt hơn.

# V. Kết luận

Trong báo cáo này, chúng tôi đã áp dụng hai phương pháp phổ biến là Apriori và FP-Growth để phân tích dữ liệu giỏ thị trường và qua đó tìm hiểu hành vi của khách hàng.

Từ việc thực hiện phân tích này, chúng tôi đã có cái nhìn sâu sắc hơn về cách mà khách hàng tương tác với các sản phẩm trong cửa hàng. Cụ thể, chúng tôi đã xác định được các mẫu mua sắm phổ biến, gợi ý sản phẩm có thể được bày bán cùng nhau để tăng doanh số bán hàng và cải thiện trải nghiệm mua sắm của khách hàng.

Ngoài ra, chúng tôi cũng đã nhận thấy rằng việc sử dụng FP-Growth có thể mang lại hiệu suất cao hơn so với Apriori đối với các tập dữ liệu lớn. Điều này làm cho việc phân tích giỏ thị trường trở nên hiệu quả hơn và giúp chúng tôi đưa ra các quyết định kinh doanh dựa trên thông tin được thu thập từ dữ liệu.

Tóm lại, việc áp dụng Apriori và FP-Growth đã cung cấp cho chúng tôi cái nhìn toàn diện về hành vi mua sắm của khách hàng và đã đóng góp vào việc tối ưu hóa chiến lược kinh doanh của cửa hàng. Chúng tôi hi vọng rằng các phân tích và kết quả từ báo cáo này sẽ hữu ích cho quyết định kinh doanh và phát triển tiếp theo của tổ chức.