Market Basket Analysis

Our Team

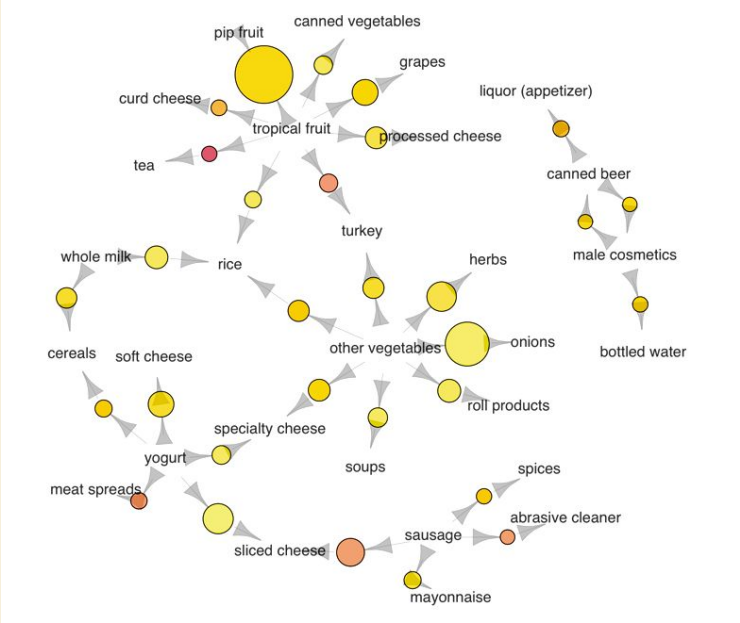
Harry Tan

Thanh Dinh

Table of Contents

Background & Purpose

- Goal: Finding frequent itemsets and come up with promotional strategies and recommendation for customers
- Process: Identify frequent itemsets using the Apriori algorithm
  - Extract valuable insights from our transactional data
  - Formulating personalized recommendations for our customers
  - Identify meaningful associations between items and complementary products



Data Overview

Granularity

- Each row in the dataset corresponds to a single respondent
- Each columns provide various attributes and responses related to their online grocery shopping habits and demographics
- The granularity allows for insights into shopping preferences and behaviors .

	PID	Bạn đã sử dụng ứng dụng đi chợ trực tuyến nào trong vòng 6 tháng gần đây?	Trung bình, mức độ thường xuyên đi chợ trực tuyến của bạn:	Hình thức thanh toán bạn thường sử dụng:	Lý do bạn lựa chọn hình thức đi chợ online?	Số tiền trung bình bạn trả cho một lần mua hàng:	Các mặt hàng bạn thường lựa chọn:	Giới tính của bạn	Độ tuổi của bạn	Tình trạng hôn nhân của bạn	Số thành viên là con cái trong gia đình của bạn	Ngành nghề của bạn:	Mức thu nhập tháng của bạn nằm trong khoảng
0	PID001	Shopee Food,	2-3 lần/tuần	Ví điện tử	Khuyến mãi của nhà cung cấp, Khuyến mãi của các hình thức thanh toán (ví điện tử, thẻ tín dụng,...), Thuận tiện trong việc mua sắm	100.000 – 300.000 VND	Thực phẩm ăn liền (bánh mì, cháo, đồ ăn nhanh, mì miến,...), Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,...), Nước giải khát (nước khoáng, nước ngọt,...), Sữa (tươi, chua,...) và các chế phẩm từ sữa (bơ, phô mai,...)	Nữ	18-22 tuổi	Độc thân	Không có	Sinh viên	5-10 triệu

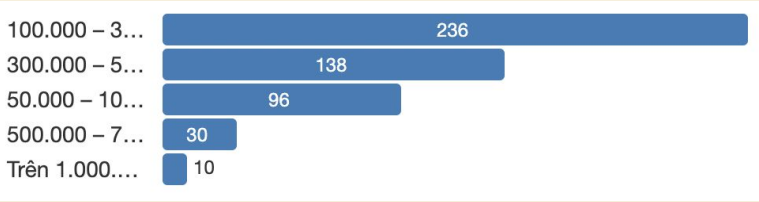
Basic Stats

- 13 columns (4 texts and 9 categorical)
- 518 rows (single respondent)
- No nulls value or missing cells.
- Age is moderately correlated with occupations. (r = 0.54)
- Gender and marriage status is highly imbalanced.

	PID	Bạn đã sử dụng ứng dụng đi chợ trực tuyến nào trong vòng 6 tháng gần đây?	Trung bình, mức độ thường xuyên đi chợ trực tuyến của bạn:	Hình thức thanh toán bạn thường sử dụng:	Lý do bạn lựa chọn hình thức đi chợ online?	Số tiền trung bình bạn trả cho một lần mua hàng:	Các mặt hàng bạn thường lựa chọn:	Giới tính của bạn	Độ tuổi của bạn	Tình trạng hôn nhân của bạn	Số thành viên là con cái trong gia đình của bạn	Ngành nghề của bạn:	M... t... nh... thấ... c... b... n... tro... khoả...
0	PID001	Shopee Food,	2-3 lần/tuần	Ví điện tử	Khuyến mãi của nhà cung cấp, Khuyến mãi của các hình thức thanh toán (ví điện tử, thẻ tín dụng,...), Thuận tiện trong việc mua sắm	100.000 – 300.000 VND	Thực phẩm ăn liền (bánh mì, cháo, đồ ăn nhanh, mì miến,...), Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,...), Nước giải khát (nước khoáng, nước ngọt,...), Sữa (tươi, chua,...) và các chế phẩm từ sữa (bơ, phô mai,...)	Nữ	18-22 tuổi	Độc thân	Không có	Sinh viên	5-... tr...

Exploratory Data Analysis

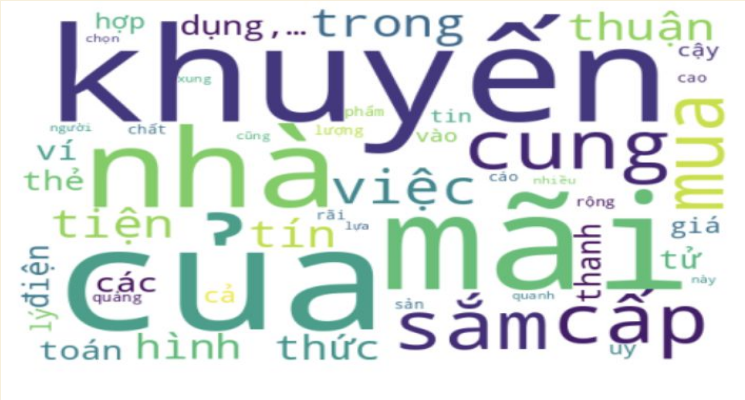
Online Grocery EDA (Part 1)



1. The respondent's self-reported frequency of online grocery shopping.

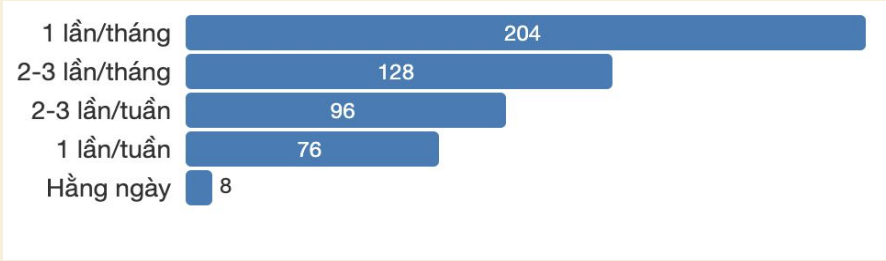
2. The respondent's choice of online grocery shopping applications used in the last 6 months.

Online Grocery EDA (Part 2)



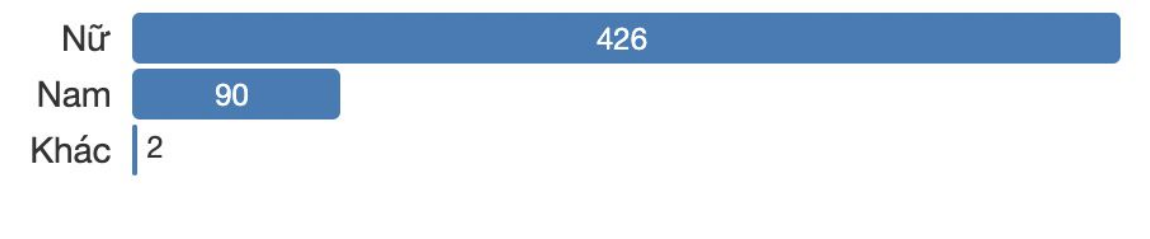
3. The reasons why the respondent chooses to shop for groceries online.

4. The respondent's self-reported frequency of online grocery shopping.

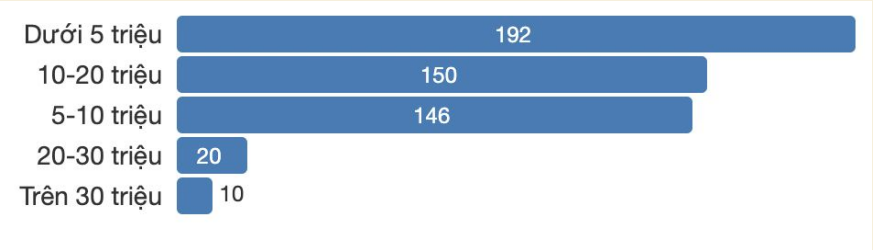


Demographic EDA (Part 1)

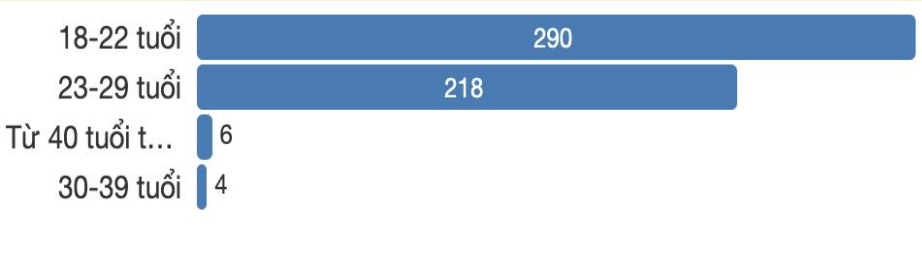
1. The respondent gender



2. The respondent ranges of monthly income



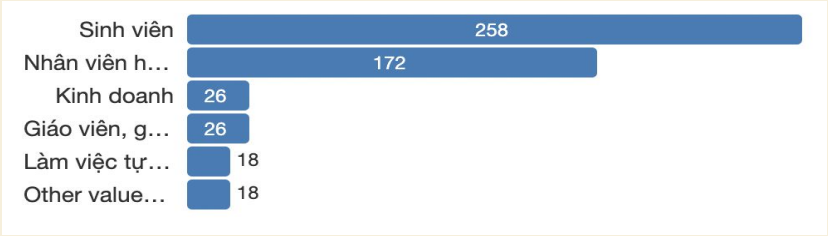
3. The respondent ranges of ages



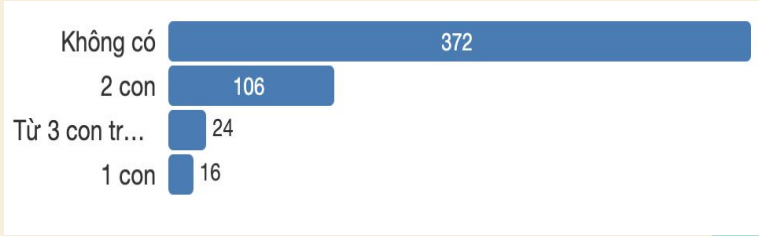
Demographic EDA (Part 2)

4. The respondent marital status

5. The respondent occupation



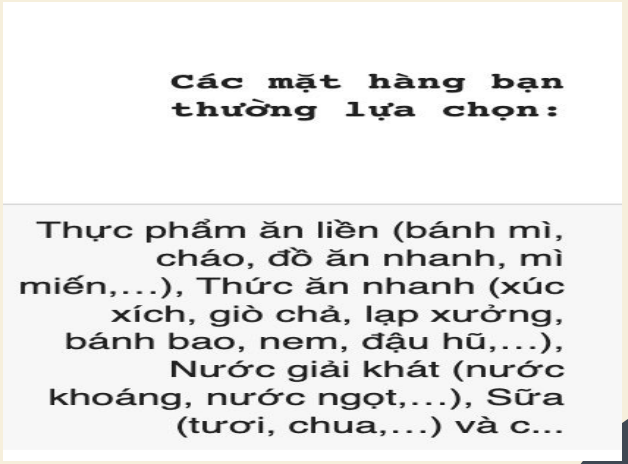
5. The respondent family size



Data Cleaning & Pre-processing

Pre-Processing

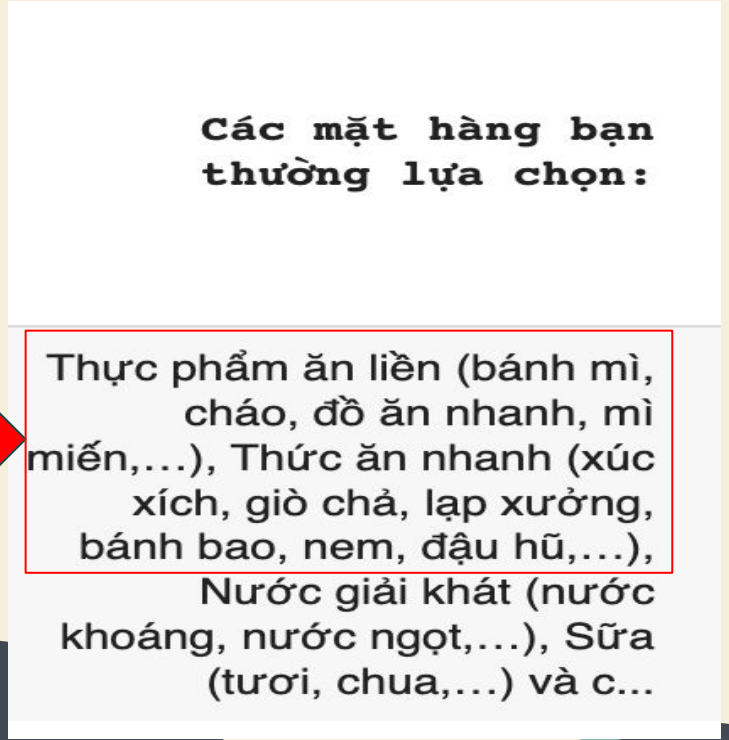
- Given dataset was a survey with groups of item
  - Splits the "Products you often choose" column into separate groups of items using a regular expression
  - Create a new row for each item in the list of products, effectively transforming each group of items into single transactions.



PID	Products
PID001	Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…)
PID001	Sữa (tươi, chua,…) và các chế phẩm từ sữa (bơ, phô mai,…)
PID001	Nước giải khát (nước khoáng, nước ngọt,…)

Removing/Replacing Duplicates

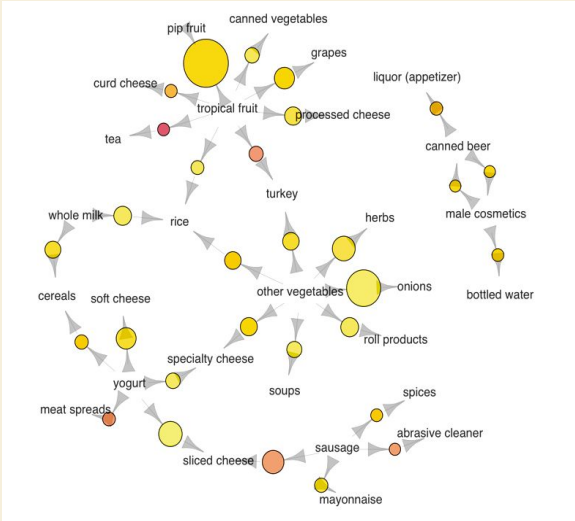
- Not much of a difference between fast food and instant food relating to online grocery
- If a customer choose both fast food and instant food -> only include fast food
- Replace all instant food with fast food to keep the data consistent



Apriori Algorithm

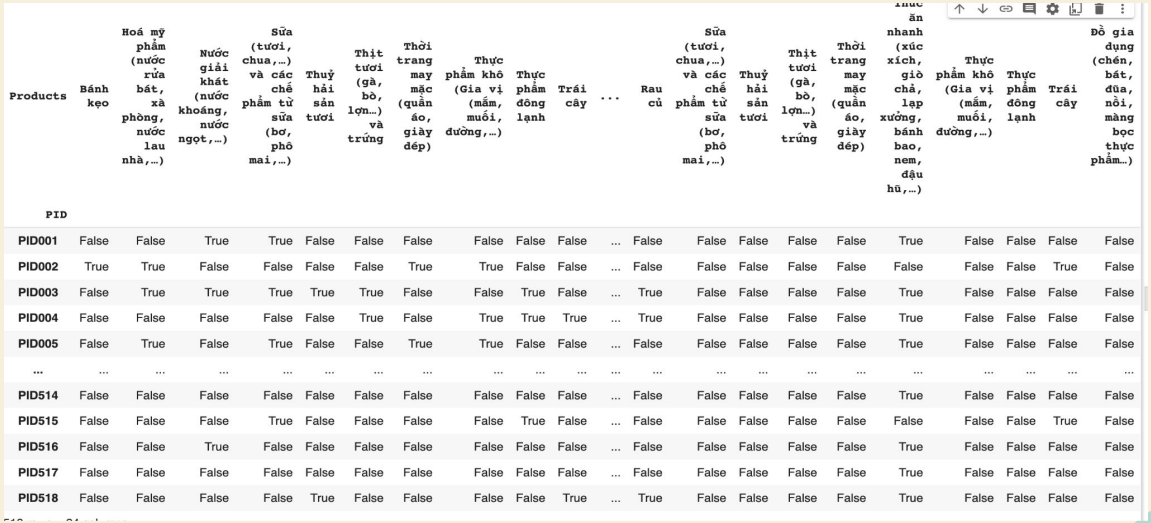
Apriori Algorithm Theory

- ML algorithm that widely used method for association rule mining and mining frequent itemsets in transactional data.
  - Uses an iterative approach to find frequent itemsets based on the "**Apriori property**."
  - If {Milk, Bread} is a frequent itemset (appearing in at least X transactions), then all its subsets ({Milk} and {Bread}) must also be frequent
- <u>Benefits:</u>
  - Simplicity and ease of implementation.
    - Efficient pruning technique for reducing computational complexity.
    - Scalability to handle large datasets.
  - Widely used in industry & studied in academia.



Implementation

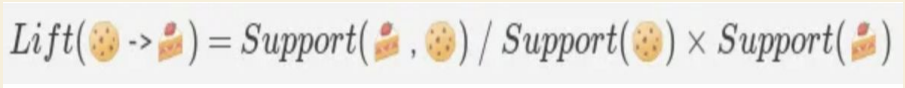
- Group item purchases into baskets of items.
- Use Transaction Encoder to encode the transactions into a format that is suitable for the Apriori function.



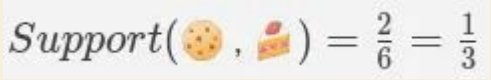
Products	Bánh kẹo	Hoá mỹ phẩm (nước rửa bát, xà phòng, nước lau nhà,…)	Nước giải khát (nước khoáng, nước ngọt,…)	Sữa (tươi, chua,…) và các chế phẩm từ sữa (bơ, phô mai,…)	Thuỷ hải sản tươi	Thịt tươi (gà, bò, lợn…) và trứng	Thời trang may mặc (quần áo, giày dép)	Thực phẩm khô (Gia vị (mắm, muối, đường,…)	Thực phẩm đông lạnh	Trái cây	...	Rau củ	Sữa (tươi, chua,…) và các chế phẩm từ sữa (bơ, phô mai,…)	Thuỷ hải sản tươi	Thịt tươi (gà, bò, lợn…) và trứng	Thời trang may mặc (quần áo, giày dép)	Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…)	Thực phẩm khô (Gia vị (mắm, muối, đường,…)	Thực phẩm đông lạnh	Trái cây	Đồ gia dụng (chén, bát, đũa, nồi, màng bọc thực phẩm…)
PID																					
PID001	False	False	True	True	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	False	False
PID002	True	True	False	False	False	False	True	True	False	False	...	False	False	False	False	False	False	False	False	True	False
PID003	False	True	True	True	True	True	False	False	True	False	...	True	False	False	False	False	True	False	False	False	False
PID004	False	False	False	False	False	True	False	True	True	True	...	True	False	False	False	False	True	False	False	False	False
PID005	False	True	False	True	False	False	True	True	False	False	...	False	False	False	False	False	True	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
PID514	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	False	False
PID515	False	False	False	True	False	False	False	False	True	False	...	False	False	False	False	False	False	False	False	True	False
PID516	False	False	True	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	False	False
PID517	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	True	False	False	False	False
PID518	False	False	False	False	True	False	False	False	False	True	...	True	False	False	False	False	True	False	False	False	False

Parameters for my models

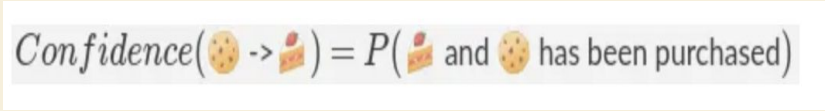
- Filter for Lift value > 1.0



- Min_support > 0.2



- Confidence level >= 0.5



Results on entire dataset

Entire Dataset

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
9	(Trái cây)	(Rau củ)	0.297297	0.401544	0.297297	1.000000	2.490385	0.177919	inf	0.851648
1	(Nước giải khát (nước khoáng, nước ngọt,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.274131	0.675676	0.254826	0.929577	1.375775	0.069602	4.605405	0.376289
5	(Thịt tươi (gà, bò, lợn…) và trứng)	(Rau củ)	0.281853	0.401544	0.254826	0.904110	2.251581	0.141650	6.241037	0.774031
7	(Thực phẩm khô (Gia vị (mắm, muối, đường,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.235521	0.675676	0.204633	0.868852	1.285902	0.045497	2.472973	0.290833
3	(Sữa (tươi, chua,…) và các chế phẩm từ sữa (bơ, phô mai,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.332046	0.675676	0.266409	0.802326	1.187442	0.042054	1.640700	0.236324
8	(Rau củ)	(Trái cây)	0.401544	0.297297	0.297297	0.740385	2.490385	0.177919	2.706707	1.000000
4	(Rau củ)	(Thịt tươi (gà, bò, lợn…) và trứng)	0.401544	0.281853	0.254826	0.634615	2.251581	0.141650	1.965454	0.928837
2	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	(Sữa (tươi, chua,…) và các chế phẩm từ sữa (bơ, phô mai,…))	0.675676	0.332046	0.266409	0.394286	1.187442	0.042054	1.102754	0.486715
0	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	(Nước giải khát (nước khoáng, nước ngọt,…))	0.675676	0.274131	0.254826	0.377143	1.375775	0.069602	1.165386	0.842172
6	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	(Thực phẩm khô (Gia vị (mắm, muối, đường,…))	0.675676	0.235521	0.204633	0.302857	1.285902	0.045497	1.096588	0.685535

Customer Segmentation

Unraveling Customer Insights among 19-22-Year-Old Student Women"

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
9	(Trái cây)	(Rau củ)	0.254902	0.372549	0.254902	1.000000	2.684211	0.159938	inf	0.842105
1	(Nước giải khát (nước khoáng, nước ngọt,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.303922	0.705882	0.294118	0.967742	1.370968	0.079585	9.117647	0.388732
5	(Thịt tươi (gà, bò, lợn…) và trứng)	(Rau củ)	0.254902	0.372549	0.245098	0.961538	2.580972	0.150135	16.313725	0.822105
7	(Thực phẩm khô (Gia vị (mắm, muối, đường,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.264706	0.705882	0.245098	0.925926	1.311728	0.058247	3.970588	0.323200
3	(Sữa (tươi, chua,…) và các chế phẩm từ sữa (bơ, phô mai,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.352941	0.705882	0.294118	0.833333	1.180556	0.044983	1.764706	0.236364
8	(Rau củ)	(Trái cây)	0.372549	0.254902	0.254902	0.684211	2.684211	0.159938	2.359477	1.000000
4	(Rau củ)	(Thịt tươi (gà, bò, lợn…) và trứng)	0.372549	0.254902	0.245098	0.657895	2.580972	0.150135	2.177979	0.976250
0	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	(Nước giải khát (nước khoáng, nước ngọt,…))	0.705882	0.303922	0.294118	0.416667	1.370968	0.079585	1.193277	0.920000
2	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	(Sữa (tươi, chua,…) và các chế phẩm từ sữa (bơ, phô mai,…))	0.705882	0.352941	0.294118	0.416667	1.180556	0.044983	1.109244	0.520000
6	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	(Thực phẩm khô (Gia vị (mắm, muối, đường,…))	0.705882	0.264706	0.245098	0.347222	1.311728	0.058247	1.126408	0.808000

Unraveling Customer Insights among 23-29-Year-Old Women"

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
17	(Trái cây)	(Rau củ)	0.348315	0.404494	0.348315	1.000000	2.472222	0.207423	inf	0.913793
20	(Thịt tươi (gà, bò, lợn…) và trứng, Trái cây)	(Rau củ)	0.213483	0.404494	0.213483	1.000000	2.472222	0.127130	inf	0.757143
25	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…), Trái cây)	(Rau củ)	0.213483	0.404494	0.213483	1.000000	2.472222	0.127130	inf	0.757143
13	(Thịt tươi (gà, bò, lợn…) và trứng)	(Rau củ)	0.280899	0.404494	0.258427	0.920000	2.274444	0.144805	7.443820	0.779212
6	(Nước giải khát (nước khoáng, nước ngọt,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.269663	0.651685	0.235955	0.875000	1.342672	0.060220	2.786517	0.349451
15	(Thực phẩm khô (Gia vị (mắm, muối, đường,…))	(Thức ăn nhanh (xúc xích, giò chả, lạp xưởng, bánh bao, nem, đậu hũ,…))	0.258427	0.651685	0.224719	0.869565	1.334333	0.056306	2.670412	0.337879
16	(Rau củ)	(Trái cây)	0.404494	0.348315	0.348315	0.861111	2.472222	0.207423	4.692135	1.000000
18	(Rau củ, Thịt tươi (gà, bò, lợn…) và trứng)	(Trái cây)	0.258427	0.348315	0.213483	0.826087	2.371669	0.123469	3.747191	0.779904

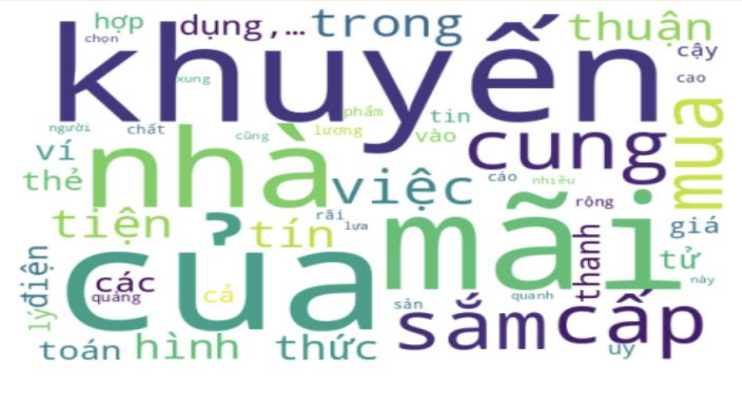
Marketing Recommendations and Insights

Promotional Strategies

1. Limited-Time Flash Sales based on the Apriori Algorithm product recommendations for each customer segment.

2. Tailor email campaigns

3. Cross-Selling Bundles and Dynamic In-App Recommendations

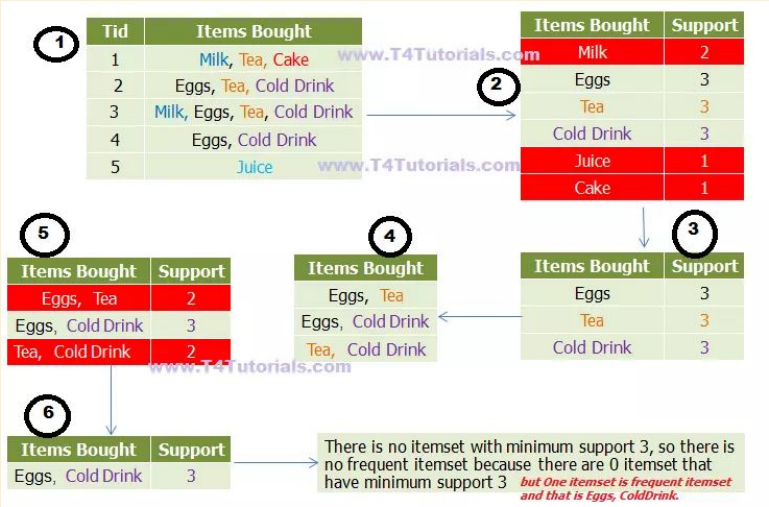


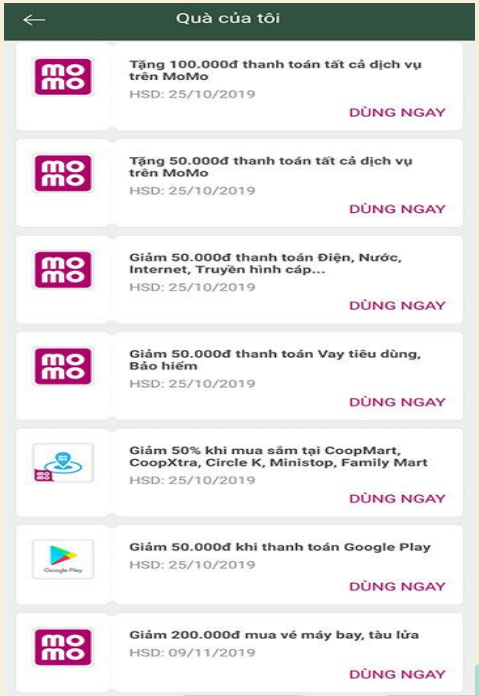
Business Strategies

1.) Continuous Algorithm Refinement

2.) Customer Segmentation Insights

3.) Collaboration with online banking/e-wallet





Limitations of Data and Algorithm

1. Data Limitations:

Data Size and Coverage: The dataset used for the analysis may have limitations in terms of size and coverage, potentially leading to incomplete or biased results.

Data Quality: Since the survey is self reported it may affect the accuracy and reliability of the Apriori Algorithm outputs.

Limited Historical Data: The dataset's time span might be limited, restricting the ability to capture long-term trends and patterns accurately.

2. Algorithm Limitations:

Scalability: The Apriori Algorithm's efficiency can decrease with large datasets, impacting its performance in real-time or high-volume scenarios.

Threshold Selection: The choice of minimum support and lift thresholds can influence the number and significance of generated frequent itemsets and association rules.

Conclusion

Conclusion

- The Apriori Algorithm provided valuable insights into customer behavior, preferences, and purchasing patterns.
- Promotional strategies, such as personalized email campaigns and dynamic in-app recommendations, were designed based on the algorithm's results.
- Collaborating with online banking and utilizing user feedback enhanced customer trust and satisfaction.
- Continuous algorithm refinement ensures recommendations align with changing customer preferences and market trends.
- The algorithm's customer segmentation insights facilitated targeted marketing efforts for 19-22 years old women and 23-29 years old women.
- The promotional strategy aims to increase customer engagement, satisfaction, loyalty, and overall customer lifetime value.
- By leveraging data-driven approaches, we hope to provide an exceptional online grocery shopping experience.

QUESTIONS?