Recommendations System for Amazon Grocery and Gourmet Food Based on Association Rule Mining

Group Analysis Challenge 2

Data Scientist Team

Groceries and Gourmet Food Category

Feby Hadayani

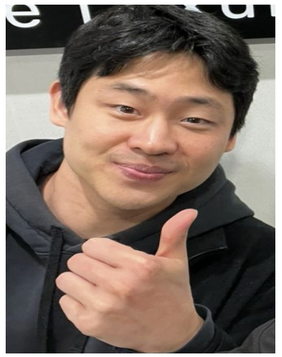
Data Scientist

Kaylie Nguyen

Data Scientist

Olivia Rumere

Data Scientist



Samuel Park

Data Scientist

Today's Agenda

Key takeaways:

- Business Problem and Scope
- Exploratory Data Analysis
- Methodology
- Model Results
- Conclusion

Business Problem and Scope

Business Problem

Personalized product recommendations enhance user experience, increase purchase likelihood, and drive revenue growth.

Scope

Identify co-occurrence of products and select best products based on the ratings of each category under consideration.

Expected Outcome

By suggesting relevant products, the recommendation system increases basket size, conversion rates, and overall customer satisfaction.

Why we choose this approach?

- Adomavicius and Tuzhilin (2005) noted that content-based systems recommend items similar to those a user has previously rated. Collaborative filtering **faces issues with sparsity**, requiring a sufficient number of ratings for accurate recommendations.
- Ahmed Alsalama (2015): proposed hybrid recommendation system using Association rules mining and content-based filtering for favourite and non-favourite item. In e-commerce, large User × Item matrices are often sparse, causing issues. Using the Apriori algorithm on these sparse matrices can result in irrelevant and poor recommendations. Easy to **implement** and **interpret.**
- Florin Stoica, Elena Pelican(2025): generate association rules from dataset containing only reviews.

Our Approach

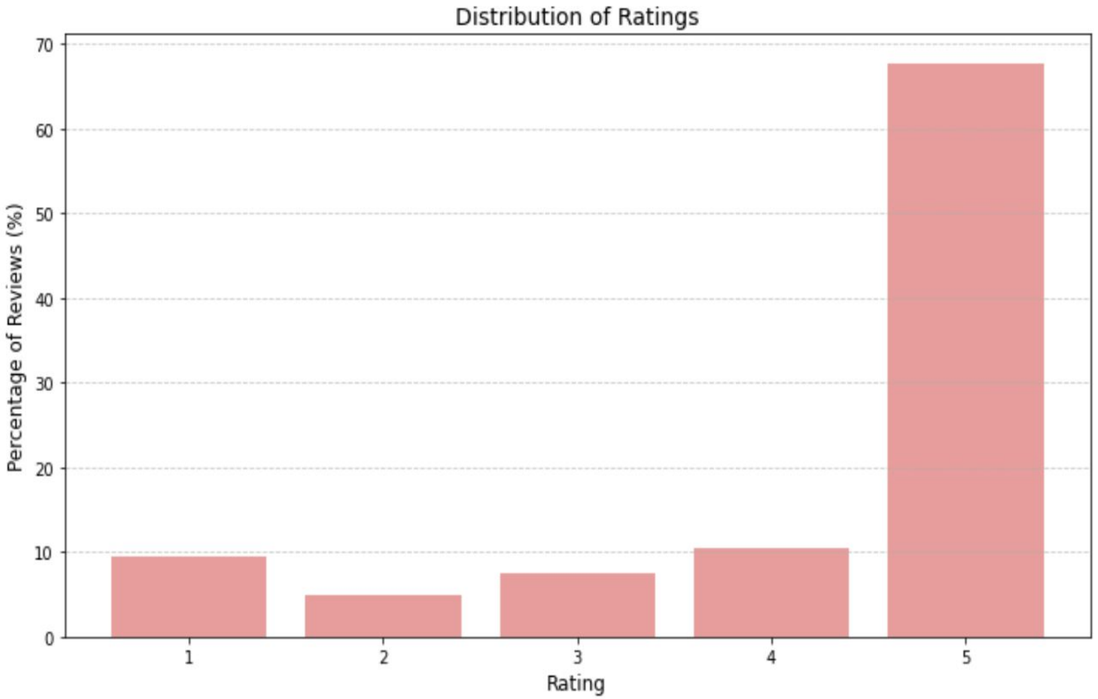
Exploratory Data Analysis

Data Summary

The data consists of **Amazon Reviews data** for **Groceries and Gourmet Food** Category covering records between 2019 and 2023 with total interaction 2,425,957.

Features	Value	Description
user_id	347,773.0	Unique identifier for the customer/reviewer
parent_asin	116,492.0	Parent ID of the product. Note: Products with different colors, styles, sizes usually belong to the same parent ID
title		Name of the product
rating	1.0-5.0	Given by the customer with
timestamp		Review timestamp (Unix)
average_rating		Rating of the product shown on the product page.
categories		Hierarchical categories of the product (in list).
rating_number		Number of ratings on the product.

Rating



The majority of ratings are concentrated at 5 stars, indicating a strong bias towards positive reviews.

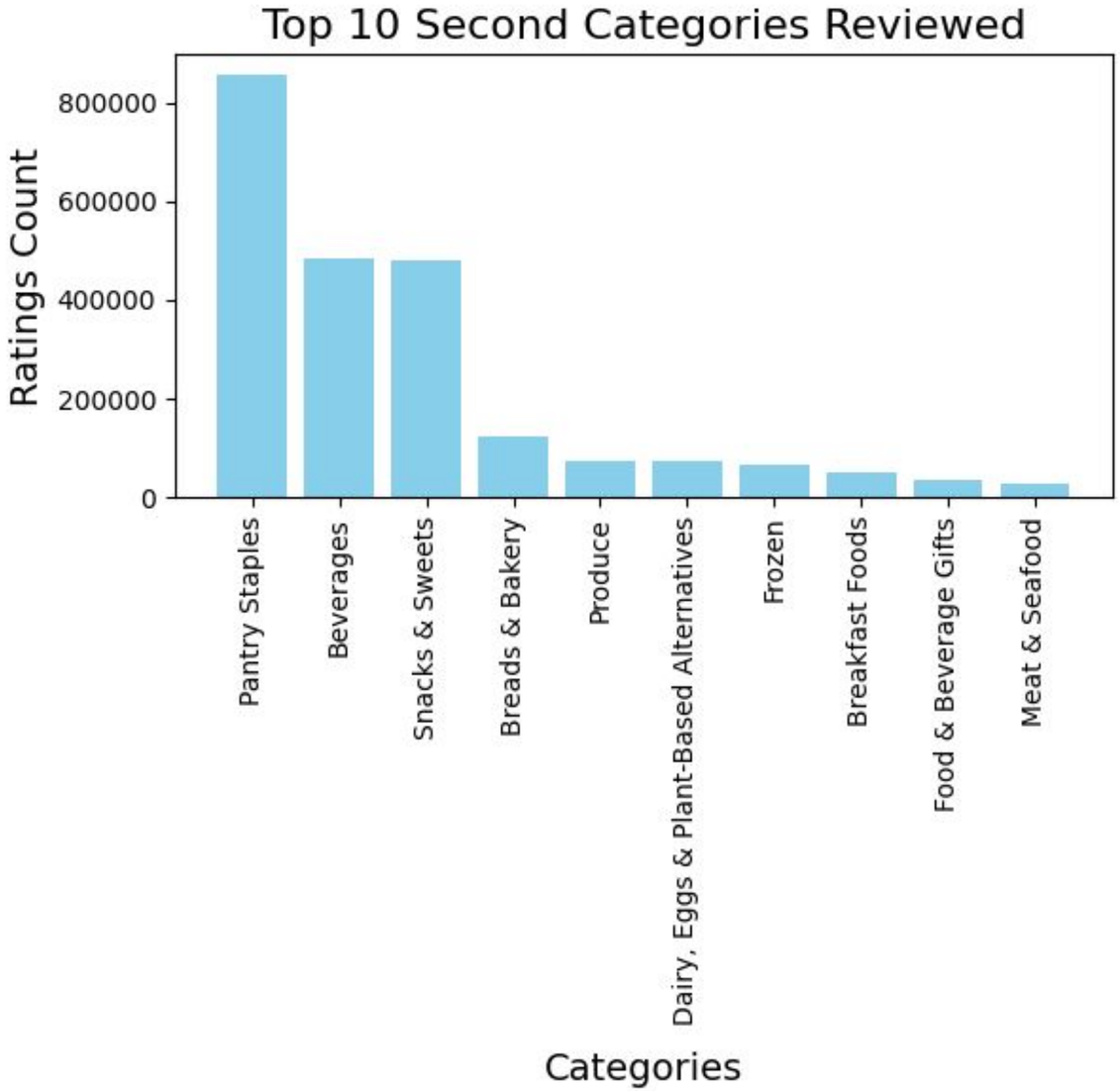
Highly-rated products tend to have a wide range of review counts, with a slight tendency for more ratings to cluster around higher average ratings.

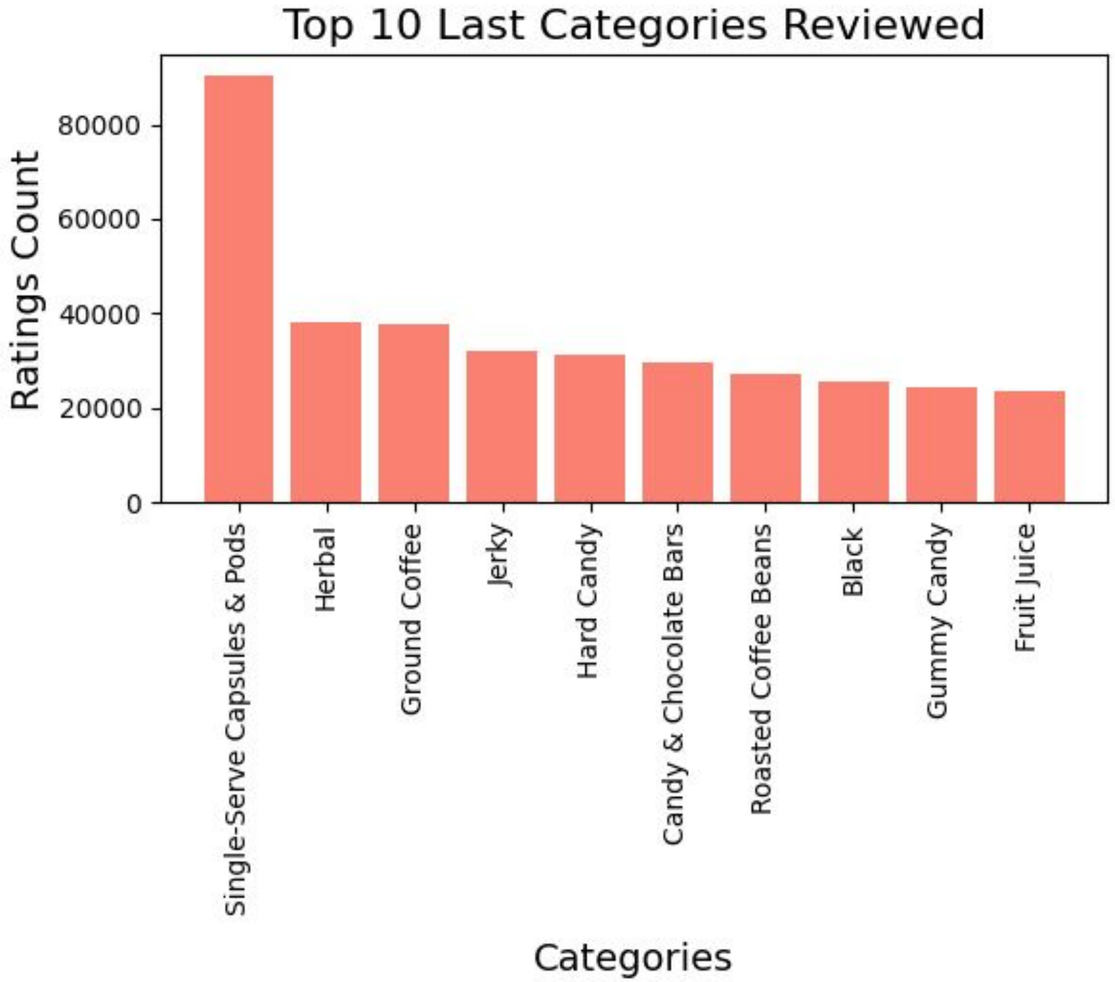
Hierarchical Categories

- Column categories is a list consist of hierarchical level of categories
- The length of the categories list vary from 1 until 7
- What subcategory should we use given the diverse length of categories across the data?
  - Check whether we can use second category (less granular option vs last category (more granular option)

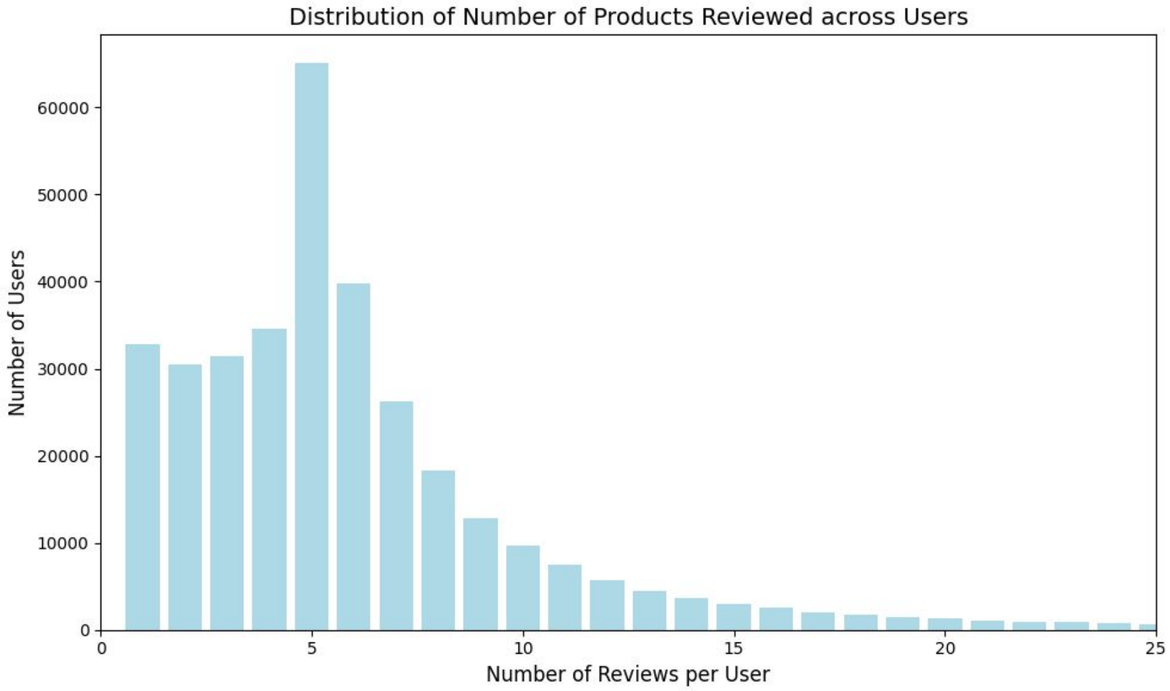
categories
[Grocery & Gourmet Food, Canned, Packaged & Baking]
[Grocery & Gourmet Food, Canned, Packaged & Baking]
[Grocery & Gourmet Food, Holiday Themed Candy]
[Grocery & Gourmet Food, Organic Groceries]
[Grocery & Gourmet Food, Food & Beverage Gifts]
...
[Grocery & Gourmet Food, Pantry Staples, Cooking & Baking, Baking Syrups, Sugars & Sweeteners, Sugars, White Sugars, White Granulated Sugar]
[Grocery & Gourmet Food, Pantry Staples, Cooking & Baking, Baking Syrups, Sugars & Sweeteners, Sugars, Brown Sugar, Jaggery]
[Grocery & Gourmet Food, Pantry Staples, Cooking & Baking, Baking Syrups, Sugars & Sweeteners, Sugars, Brown Sugar, Brown Granulated Sugar]
[Grocery & Gourmet Food, Pantry Staples, Cooking & Baking, Baking Syrups, Sugars & Sweeteners, Sugars, Brown Sugar, Jaggery]
[Grocery & Gourmet Food, Pantry Staples, Cooking & Baking, Baking Syrups, Sugars & Sweeteners, Sugars, Brown Sugar, Brown Granulated Sugar]

Top Second vs Last Categories Reviewed





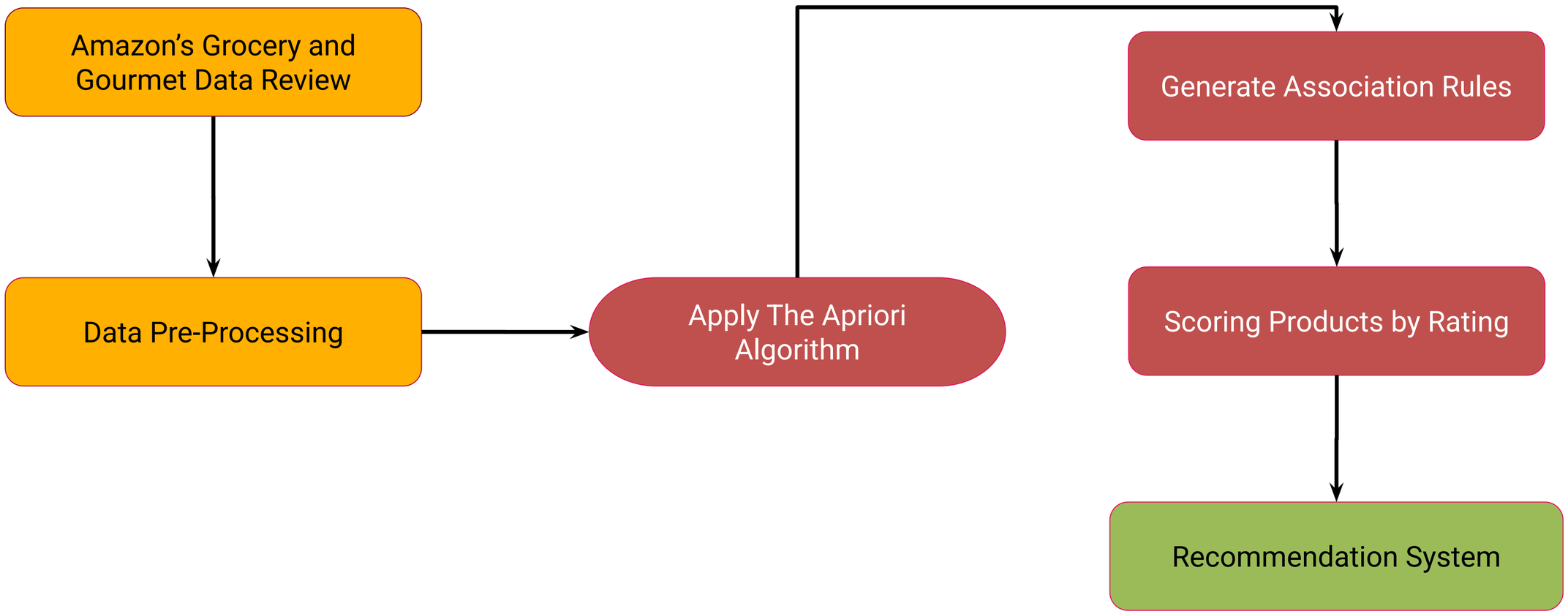
User vs Review



- Most users have reviewed only a few products compared the data size, while a small number of users contribute a large volume of reviews.
- Skewed distribution may indicate there is a marketing promotion for certain products.
- This distribution explains the sparsity of review data, which can impact recommendation performance.
- Need for grouping into identical user,date

Methodology and Analysis

Our Recommendation System Framework



Data Pre-Processing

Exclude users with unusually high number of reviews (top 0.5% quantile: 56 reviews in 5 years), as they could represent paid influencer/promotor.

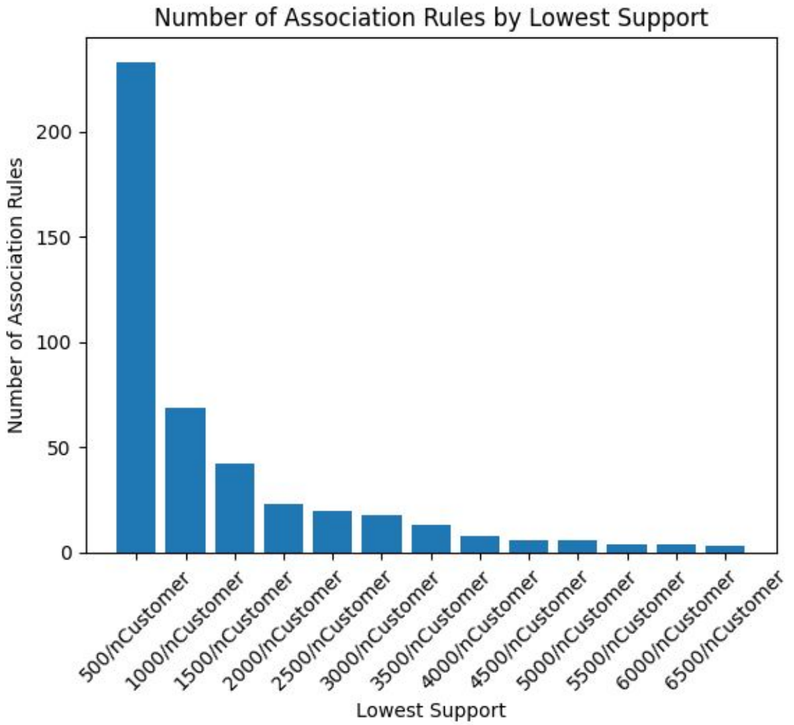
Ensure there is at least 1 transaction in a list/category.

Extract the second layer from “category” to use for Association Rule.

Transform to “Market Basket”.

Association Rules

- 0.48% as the lowest support (rules need to be applied for at least 1500 customers).
- Minimum 50% confidence: At least 50% likelihood that consequent of rule will occur given antecedent.
- Credible rules are rules with more than 1 lift, as the occurrence of antecedents boost occurrence of consequents not by chance.
- The dataset has relatively low support because of the sparse nature and larger itemsets, as the chance of a transaction containing all items in the rule decreases as the size of the itemset grows.



	support	consequent support	confidence	lift
count	42.000000	42.000000	42.000000	42.000000
mean	0.010904	0.584606	0.568148	0.984615
std	0.009608	0.061343	0.054758	0.153717
min	0.004791	0.397799	0.505657	0.836808
25%	0.005515	0.604269	0.534730	0.887633
50%	0.007423	0.604269	0.548499	0.912239
75%	0.011970	0.604269	0.575695	1.064415
max	0.049664	0.604269	0.693810	1.408517

Scoring Products by Rating

- After getting the association rules, we apply scoring by rating to recommend certain products from category.
- For each product, calculate

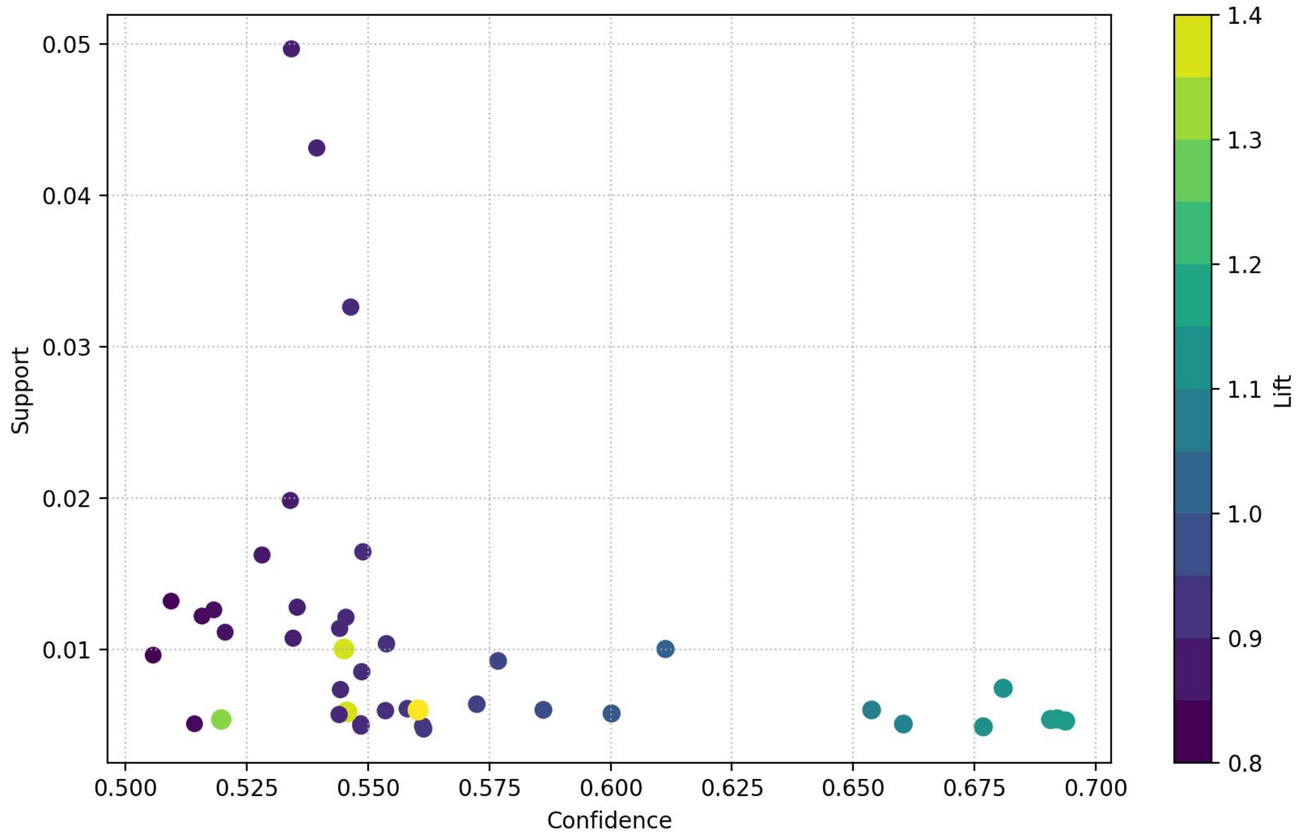
$$Scoring = \frac{average\_rating}{\max\ \_rating}\ x\ rating\_weight + \frac{num\_reviews}{\max\ \_num\_reviews}\ x\ reviews\_weight$$

- For weight, we would like to emphasize more on confidence, therefore we use rating_weight=30% and reviews_weight=70%
- Group products in each consequent category.
- Recommend the top products in each consequent category once customers buy products in the antecedents.

Model Results

Rules

Metric	Value
Number of Items	167
Number of Customers	314752
Lowest Support	1500/nCustomer= 0.48%
Max size of item sets	7
Min threshold of confidence	50%
Number of frequent itemsets	123
Number of association rules	42
Number of credible rules	12



Association Rules

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(Beverages, Pantry Staples, Breakfast Foods)	(Snacks & Sweets)	0.010774	0.397799	0.006036	0.560307	1.408517
(Breads & Bakery, Breakfast Foods)	(Snacks & Sweets)	0.010824	0.397799	0.005906	0.545641	1.371651
(Breads & Bakery, Beverages, Pantry Staples)	(Snacks & Sweets)	0.018456	0.397799	0.010059	0.545016	1.370080
(Breads & Bakery, Pantry Staples, Dairy, Eggs ...	(Snacks & Sweets)	0.010405	0.397799	0.005407	0.519695	1.306426
(Dairy, Eggs & Plant-Based Alternatives, Froze...	(Pantry Staples)	0.007647	0.604269	0.005306	0.693810	1.148179

- The consequent categories with highest lifts are Snacks & Sweets, with repetitive items in the antecedents.
- All the consequents belong to two categories: Snacks & Sweets or Pantry Staples.

Scoring of Top Products by Consequents

	title	second_cat	rating_number	average_rating	scoring
	Goldfish Cheddar Crackers, Snack Crackers, 30 ...	Snacks & Sweets	125106.0	4.8	0.969570
	Welch's Mixed Fruit, 0.9 oz, 40 Ct	Snacks & Sweets	97758.0	4.8	0.820579
	Frito-Lay Chips and Quaker Chewy Granola Bars ...	Snacks & Sweets	97009.0	4.6	0.804499

	title	second_cat	rating_number	average_rating	scoring
	[Samyang] Carbo Spicy Chicken Fried Cup Noodle...	Pantry Staples	128489.0	4.4	0.964000
	Lakanto Golden Monk Fruit Sweetener with Eryth...	Pantry Staples	112773.0	4.7	0.896380
	Organic Coconut Oil - Unrefined and Cold-Press...	Pantry Staples	96857.0	4.7	0.809671

** More details in the Appendix*

Scoring Rating

Average rating: 4.8
Number of ratings: 125,106
Score: 0.9696

Snacks & Sweets

Average rating: 4.5
Number of ratings: 74,849
Score: 0.6778

Pantry Staples

Conclusion & Recommendation

Conclusion

We have applied the concepts we learnt in this course and below are some insights:

- Association rule mining with apriori algorithm can be considered to use as a approach in building recommendation system in e-commerce.
- Threshold sensitivity: different threshold leads to different results

Limitation and Potential Modifications

- Lack of basket data.
- Association rule is not suitable for large datasets.
- Incorporate price into the model.
- Consider improve the model when we have sparsity in rating.

References

- Alsalama, A. M. (2015). A hybrid recommendation system based on association rules. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 9(1), 19-23.
- Stoica, F., & Pelican, E. (2025). A machine learning approach of enhancing eCommerce solutions. *Expert Systems with Applications*, 126, 126997. https://doi.org/10.1016/j.eswa.2025.126997
- Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734-749.

amazon

Thank You

Appendix

Data Pre-Processing Convert to Transactional Data

user_id	parent_asin	rating	timestamp	date	year	month	day	main_category	title	average_rating	rating_number	features	description	price	store	categories	details	author
3Q7A7V7UXW5JJI6UGRYQ	B0BG8M4XW7	5.0	1581313263026	2020-02-10 05:41:03.026	2020	2	10	Grocery	Kellogg's Raisin Bran, Breakfast Cereal, Origi...	4.6	704	["Crunchy, lightly sweetened breakfast cereal ...	["Wake up with Sunny and the simple goodness o...	17.87	Raisin Bran	['Grocery & Gourmet Food', 'Breakfast Foods', ...	{'Is Discontinued By Manufacturer': 'No', 'Pac...	NaN
3Q7A7V7UXW5JJI6UGRYQ	B005CD4196	5.0	1581313294965	2020-02-10 05:41:34.965	2020	2	10	Grocery	NESTLE TOLL HOUSE Cocoa 8 oz. Plastic Canister	4.7	7323	['Made with 100% pure cocoa. SNAP and EBT Elig...	['This 100% cocoa from a special blending of c...	2.78	Toll House	['Grocery & Gourmet Food', 'Pantry Staples', '...	{'Is Discontinued By Manufacturer': 'No', 'Pro...	NaN
3Q7A7V7UXW5JJI6UGRYQ	B07194LN2Z	5.0	1581313319614	2020-02-10 05:41:59.614	2020	2	10	Grocery	Oscar Mayer Bacon Bits with Hickory Smoke Flav...	4.6	521	['Product Type:Grocery', 'Item Package Dimensi...	[]	NaN	Oscar Mayer	['Grocery & Gourmet Food', 'Pantry Staples', '...	{}	NaN

```
transactions
```

	user_id	year	month	day	items
0	AE22236AFRRSMQIKGG7TPTB75QEA	2020	1	17	[B09WSC98TH]
1	AE22236AFRRSMQIKGG7TPTB75QEA	2022	3	14	[B00PGXQ68Q]
2	AE222H3FGXWLHRFUMGMS2RR57NDQ	2020	8	5	[B0057FBQTC]
3	AE222H3FGXWLHRFUMGMS2RR57NDQ	2020	11	5	[B0CDLRTR5L]
4	AE222H3FGXWLHRFUMGMS2RR57NDQ	2020	11	17	[B07YP6LYYS]

Grouped by user_id and purchase date & Category

amazon

Scoring of Top Products

	title	second_cat	rating_number	average_rating	scoring
0	Goldfish Cheddar Crackers, Snack Crackers, 30 ...	Snacks & Sweets	125106.0	4.8	0.969570
1	Welch's Mixed Fruit, 0.9 oz, 40 Ct	Snacks & Sweets	97758.0	4.8	0.820579
2	Frito-Lay Chips and Quaker Chewy Granola Bars ...	Snacks & Sweets	97009.0	4.6	0.804499
3	Quaker Chewy Granola Bars, Oatmeal Raisin, 58 ...	Snacks & Sweets	87155.0	4.7	0.756815
4	Orville Redenbacher's Gourmet Popcorn Kernels,...	Snacks & Sweets	82342.0	4.7	0.730594
5	Ring Pop Bulk Candy Lollipop Variety Party Pac...	Snacks & Sweets	70121.0	4.8	0.670015
6	Pringles Potato Crisps Chips, Snack Stacks, Lu...	Snacks & Sweets	61809.0	4.7	0.618732
7	Cheez-It Cheese Crackers, Baked Snack Crackers...	Snacks & Sweets	59702.0	4.8	0.613253
8	Ferrero Collection Premium Gourmet Assorted Ha...	Snacks & Sweets	54645.0	4.8	0.585703
9	Mott's Fruit Flavored Snacks, Assorted Fruit, ...	Snacks & Sweets	54499.0	4.7	0.578907

Scoring of Top Products

	title	second_cat	rating_number	average_rating	scoring
21819	[Samyang] Carbo Spicy Chicken Fried Cup Noodle...	Pantry Staples	128489.0	4.4	0.964000
21820	Lakanto Golden Monk Fruit Sweetener with Eryth...	Pantry Staples	112773.0	4.7	0.896380
21821	Organic Coconut Oil - Unrefined and Cold-Press...	Pantry Staples	96857.0	4.7	0.809671
21822	Maruchan Ramen Creamy Chicken Flavor, 3 Oz, Pa...	Pantry Staples	74087.0	4.7	0.685621
21823	Samyang Buldak Carbo Korean Spicy Hot Chicken ...	Pantry Staples	74849.0	4.5	0.677773
21824	Wonderful Pistachios, Roasted and Salted, 8 Ou...	Pantry Staples	64589.0	4.7	0.633877
21825	Mike's Hot Honey–Original & Extra Hot Combo 10...	Pantry Staples	55269.0	4.7	0.583102
21826	Maggi Masala 2-Minute Noodles India Snack - 24...	Pantry Staples	59627.0	4.3	0.582844
21827	Maruchan Instant Lunch Lime Chicken Flavor, 2....	Pantry Staples	54937.0	4.6	0.575293
21828	Heinz, Tomato Ketchup, 14oz Squeeze Bottle (Pa...	Pantry Staples	50263.0	4.8	0.561830