Market Basket Analysis:

Comparative Study of Apriori and FP-Growth Algorithms in Market Basket Analysis

Author: Thant Thiha,
2025178@student.cct.ie
Higher Diploma in Data Analytics for Business,
CCT College Dublin

1. EXECUTIVE SUMMARY

This report presents a comprehensive market basket analysis of retail grocery transactions using two prominent frequent pattern mining algorithms: Apriori and FP-Growth. The analysis examined 130,769 transactions from 1,762 households across 25,994 unique products, filtered to 291 commodity categories over a 3-month period.

Key Findings:

- Both algorithms identified identical patterns: 2,477 frequent itemsets and 2,220 association rules
- Apriori demonstrated superior computational efficiency (1.81 seconds vs 86.19 seconds)
- Strong complementary purchasing patterns discovered, particularly around meal components (pasta, sauce, beef)
- Highest lift value of 14.16 indicates powerful cross-selling opportunities

The number of word count in this report is 742 from Introduction to Limitations (excluding Titles, Subtitles, Tables, Figures, Captions, References, Citations).

2. INTRODUCTION

2.1 Background

Market Basket Analysis (MBA) is a data mining technique used to discover associations between products purchased together (Han *et al.*, 2011). Retailers use these insights for:

- Product placement optimization
- Cross-selling and promotional strategies
- Inventory management
- Personalized recommendations

2.2 Project Objectives

- Implement and compare two industry-standard algorithms: Apriori and FP-Growth
- Identify significant product associations in retail grocery data
- Evaluate algorithmic performance differences
- Provide actionable business recommendations

2.3 Dataset Characteristics

The dataset is the same as the one used across all projects in Integrated CA2.

Table 1: Dataset Characteristics

Metric	Value
Unique Households	1,762
Unique Products	25,994
Total Transactions	130,769
Average Basket Size	9.2 items
Media Basket Size	5 items
Analysis Period	Weeks 1-14 (90 days)
Categories	291
Matrix Dimensions	14,219 x 291
Matrix Spasity	97.68%

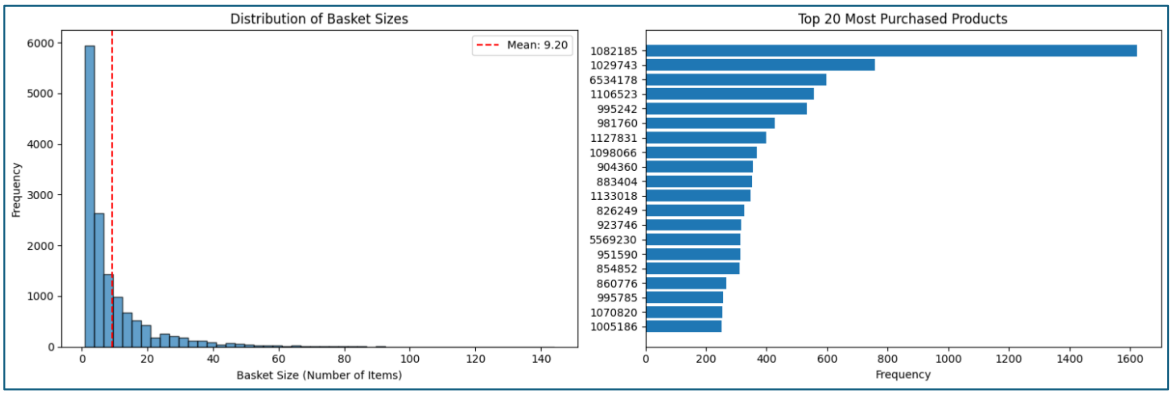

Figure 1:Distribution of the basket sizes and top products

The high sparsity (97.68%) indicates that most product combinations are not purchased together, which is typical in grocery retail where customers purchase specific subsets from a large catalog (Linoff and Berry, 2011).

3. METHODOLOGY

3.1 Algorithm Selection

Apriori Algorithm approach is breadth-first search through the matrix and generate candidates level by level by using Apriori principle (if an itemset is infrequent, all its supersets are infrequent). FP-Growth Algorithm approach is divide-and-conquer by using pattern growth method and building compact FP-tree data structure (Han *et al.*, 2000).

The analysis used moderately conservative thresholds to ensure high-quality, actionable rules.

Table 2: Parameter Selection

Parameter	Value	Justification
Minimum Support (X) = (count(X)/total_transactions)	0.01 (1%)	Captures patterns appearing in >131 transactions, balances rare vs common patterns
Minimum Confidence (X -> Y) = Support(XuY)/Support(X)	0.50 (50%)	Ensures rules are correct at least half the time, higher reliability for business decisions
Minimum Lift (X -> Y) = Confidence(X->Y)/Support(Y)	2.0	Requires associations to be at least 2 times better than random chance, filters trivial patterns

These parameters were intentionally stringent to produce high-quality, reliable rules suitable for strategic business decisions.

3.2 Data Preprocessing

We only used the first 90 days of the original 2 years dataset for computational efficiency on local machine. Data were grouped by BASKET_ID and COMMODITY_DESC and converted to binary matrix (purchase/no purchase). Dimension is then reduced by aggregating product-level data to commodity categories (25,994 to 291) (Larose, 2014).

4. PERFORMANCE EVALUATION RESULTS

4.1 Frequent Itemset Discovery

Both algorithms discovered **2,477 frequent itemsets** meeting the minimum support threshold of 1%.

Table 3: Top 10 Most Frequent Itemsets

Rank	Itemset	Support	Transactions
1	SOFT DRINKS	27.55%	3,916
2	FLUID MILK PRODUCTS	23.78%	3,380
3	BAKED BREAD/BUNS/ROLLS	21.34%	3,047
4	CHEESE	15.66%	2,227
5	BAG SNACKS	15.28%	2,172
6	BEEF	13.98%	1,988
7	TROPICAL FRUIT	12.36%	1,757
8	REFRIGERATED JUICES/DRINKS	10.10%	1,436
9	FLUID MILK + BREAD	10.09%	1,435
10	CANDY – CHECKLANE	9.99%	1,421

Key Insights: Staple Items such as soft drinks, milk, and bread are purchased in >20% of transactions. First Multi-Item Pattern (Milk + Bread) appears at 9th position with 10.09% support. Category diversity is seen with a mix of beverages, dairy, bakery and snacks in top 10.

4.2 Association Rules

2,220 association rules were generated meeting all threshold criteria.

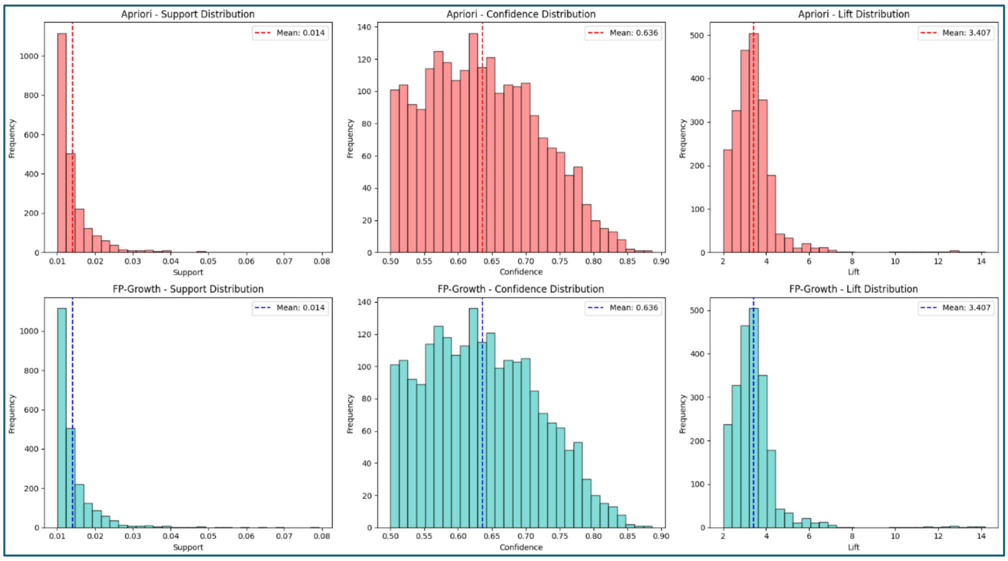

Figure 2: Distributions of Support, Confidence and Lift

Table 4: Statistical Summary of the Rules

Metric	Mean	Std Dev	Min	25%	Median	75%	Max
Support	1.41%	0.57%	1.01%	1.08%	1.23%	1.50%	7.93%
Confidence	63.62%	8.16%	50%	57.02%	63.03%	69.03%	88.62%
Lift	3.41	1.14	2.00	2.81	3.28	3.73	14.16

Table 5: Top 10 Association Rules (by Lift)

Rank	Antecedent	Consequent	Support	Confidence	Lift
1	MILK + CHEESES	DELI MEATS + BREAD	1.01%	54.37%	14.16
2	DRY NOODLES + BEEF	PASTA SAUCE	1.41%	58.60%	14.10
3	DRY NOODLES + SOFT DRINKS	PASTA SAUCE	1.09%	56.99%	13.71
4	PASTA SAUCE + BEEF	DRY NOODLES	1.41%	64.42%	13.51
5	CHEESE + DRY NOODLES	PASTA SAUCE	1.36%	54.65%	13.15
6	PASTA SAUCE +CHEESE	DRY NOODLES	1.36%	61.59%	12.92
7	BREAD + DRY NOODLES	PASTA SAUCE	1.40%	53.64%	12.91
8	BREAD + PASTA SAUCE	DRY NOODLES	1.40%	60.49%	12.69
9	MILK + DRY NOODLES	PASTA SAUCE	1.31%	52.39%	12.61
10	MILK + PASTA SAUCE	DRY NOODLES	1.31%	58.86%	12.34

9 out of 10 strongest rules involve pasta-related items (noodles, sauce, beef) showing meal component synergy. High lift values were seen as top rules show 12-14 times improvement over random chance. Moderate support with rules affecting 1-1.5% of transactions and strong confidence of average 63.62% indicating reliable predictions. Cross-category patterns were also observed as the rules span multiple departments (dairy, meat, pasta, bread).

4.3 Performance Comparison

Table 6: Comparative Summary: Apriori vs FP-Growth

Metric	Apriori	FP-Growth	Difference
Execution Time	1.81 seconds	86.19 seconds	FP-Growth 47.64 times slower
Frequent Itemsets	2,477	2,477	Identical
Association Rules	2,220	2,220	Identical
Memory Efficiency	High	Moderate	Apriori better
Scalability	Moderate	High (theoretically)	Dataset-dependent

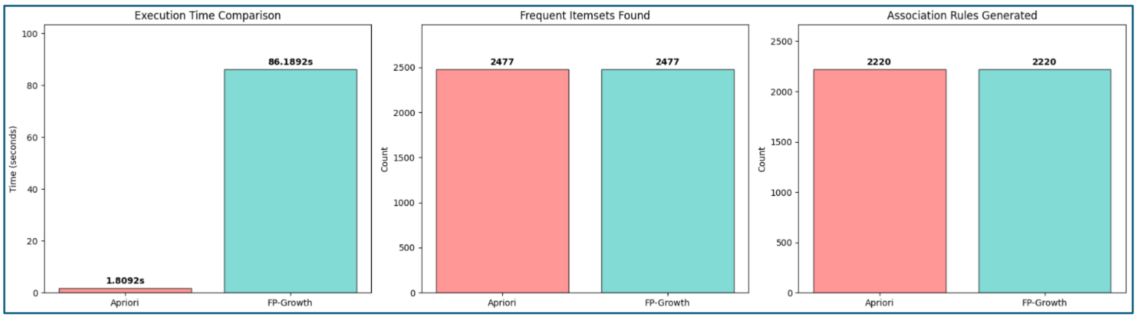

Figure 3: Algorithm Comparison

Both algorithms demonstrated perfect agreement in their results with same frequent itemsets, same association rules, same support values, and same confidence and lift. These must be because both find all frequent itemsets (no false negatives) and only frequent itemsets (no false positives). Both solve the exact same mathematical problem (although Apriori and FP-Growth algorithmic approaches differ in search strategy, candidate generation, data structure, and database scans) and since the problem definition is identical and both are proven correct, they must produce identical results. In terms of business value, both algorithms are suitable for production use and results are trustworthy and reproducible (Aggarwal, 2015).

Although FP-Growth should be faster on large datasets, it is observed that Apriori was 47.64 times faster in this analysis. This could be because Apriori's pruning is highly effective on sparse data and the dataset is not large enough for FP-Growth advantages to dominate and Apriori's simplicity pays off at this scale (Aggarwal, 2015).

5. BUSINESS INSIGHTS & LIMITATIONS

5.1 Strategic Product Association

	Finding	Business Implications	Recommendations	Expected Impact
Insight 1: Pasta Meal Ecosystem	Strongest associations involve pasta, sauce, and beef (lifts 12-14)	Customers buying pasta ingredients are highly predictable and 58-64% confidence means 6 out of 10 customers follow these patterns	**Store Layout**: Place pasta, sauce, and beef in proximity or sight lines **Promotions**: Bundle pasta + sauce + beef at discounted price **Recipe Cards**: Display easy pasta recipes near these sections **Inventory**: Maintain synchronized stock levels for these items	10-15% increase in basket size for pasta purchasers
Insight 2: Dairy + Deli + Bakery Connection	Milk + Cheese -> Deli Meats + Bread (lift 14.16, highest)	Sandwich/breakfast makers show strong cross-category purchasing with small market (1.01% support) but very high quality (14× lift)	**Cross-Department Marketing**: "Complete Breakfast" campaigns **Sampling Stations**: Deli samples near dairy section **Meal Kits**: Pre-assembled sandwich ingredient packages **Digital Coupons**: Triggered offers when scanning milk/cheese	5-8% increase in deli department sales
Insight 3: Staple Items (Milk, Bread, Soft Drinks)	20-28% of baskets contain these high-support items	Traffic drivers that attract regular shoppers with lower lift values (not surprising combinations)	**Loss Leader Strategy:** Competitive pricing on staples to drive store traffic **Strategic Placement**: Position at store perimeter to maximize exposure **Upsell Opportunities**: Place premium/complementary items nearby **Loyalty Programs**: Rewards tied to staple purchases	Maintain customer retention and store visits

5.2 Limitations & Considerations

Data Limitations: the dataset is only 90-day window and seasonal patterns are not captured. Also commodity-level masks brand-specific patterns and cannot link to demographics due to customer anonymization (Witten *et al.*, 2017).

Analytical Limitations: Association doesn’t mean causation and rules may change over time. Additionally, Promotions, holidays are not accounted for in this study.

6. CONCLUSION

This analysis demonstrates that in data science, **elegant theory meets messy reality**. While FP-Growth is theoretically superior for large-scale mining, Apriori's simplicity and effectiveness on sparse data made it the superior choice for this dataset. The perfect agreement between algorithms—2,477 identical itemsets and 2,220 identical rules—validates both the implementation and the fundamental mathematical equivalence of these approaches.

The discovered associations, particularly the pasta meal ecosystem (lifts 12-14) and dairy-deli-bakery connection (lift 14.16), represent genuine opportunities for revenue growth through targeted bundling, promotions, and store layout optimization.

Ultimately, this comparative study reinforces a key principle: **the best algorithm is the one that delivers correct results efficiently for specific data**. Theory guides, but empirical testing decides (Tan, Steinbach and Kumar, 2018).

REFERENCES

Han, J., Kamber, M. and Pei, J. (2011) *Data mining: concepts and techniques*. 3rd edn. Waltham: Morgan Kaufmann.

Linoff, G. and Berry, M. J. A. (2011) *Data mining techniques: for marketing, sales, and customer relationship management*. 3rd edn. Indianapolis: Wiley.

Han, J., Pei, J. and Yin, Y. (2000) 'Mining frequent patterns without candidate generation', in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. Dallas, TX: ACM, pp. 1-12.

Larose, D. T. (2014) *Discovering knowledge in data: an introduction to data mining*. 2nd edn. Hoboken, NJ: Wiley.

Aggarwal, C. C. (2015) *Data mining: the textbook*. Cham, Switzerland: Springer.

Tan, P. N., Steinbach, M. and Kumar, V. (2018) *Introduction to data mining*. 2nd edn. Boston: Pearson.

DataCamp - Market Basket Analysis in Python