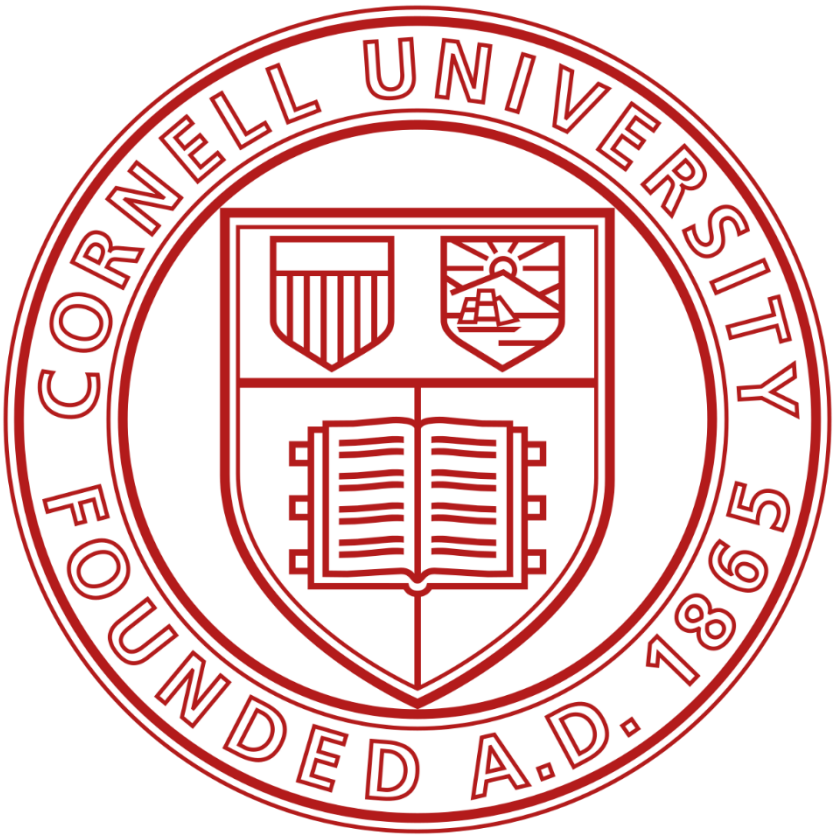


Data Analytics (CEE 4930)

Project 1 Market Basket Analysis Document

Team Members

Sarp Sertturk (ss4476)

Mahima Arhanth (ma2466)

Prabhujith Eshwar Aravindan (pa425)

Anuprita Kaple (ak2837)

Table of Contents

Introduction

Project Title: Understanding Customer Preferences by Market Basket Analysis

Problem:

Retailer players are struggling to capture the insights from their customers' purchases. This issue creates challenges in formulating product bundles, campaigns, and cross-selling opportunities.

Project Goal:

Understanding consumer behavior is essential to developing effective marketing strategies in the dynamic world of business and commerce. Our goal in this project is to use a large dataset to examine the nuances of consumer purchasing behavior via the prism of market basket analysis. With the help of this project, we investigated possible product combinations & sales opportunities from our dataset.

Approach:

Three different algorithms have been used to capture insights from the dataset. The algorithms that have been used are as follows:

- Equivalence Class Clustering and Bottom-Up Lattice Traversal (ECLAT)
- FP-Growth
- K-means clustering

The programming languages that have been used to perform coding: R & Python.

Dataset:

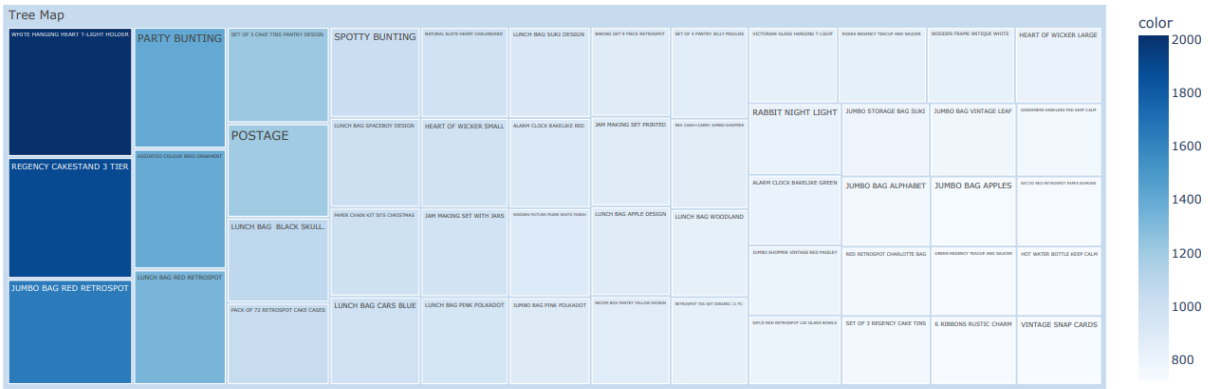
https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis

ECLAT Market Basket Analysis

Reflection on the results:

The ECLAT-driven market basket analysis of the Online Retail dataset uncovers distinct purchasing trends characterized by a small number of frequently purchased items and significant co-purchase relationships within certain product categories. Upon data cleaning and transforming each invoice into a compilation of purchased items, the binary transaction matrix illustrated that only a small portion of the 3,944 unique products appear in at least 2% of transactions, with the WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER, and JUMBO BAG RED RETROSPOT standing out as the most commonly bought individual items. Using a minimum support threshold of 2% for mining frequent itemsets, the analysis revealed just a few product pairs that achieve this threshold, indicating that co-purchases are significant yet relatively selective. The most prominent pairings included ROSES REGENCY TEACUP AND SAUCER & GREEN REGENCY TEACUP AND SAUCER, JUMBO BAG PINK POLKADOT & JUMBO BAG RED RETROSPOT, and ALARM CLOCK BAKELIKE RED & ALARM CLOCK BAKELIKE GREEN, showcasing a consistent behavior where customers tend to buy multiple variants of similar items and particularly color or style variations within the same product range.

Visualization of the most frequent items



These trends suggest that the store's selection fosters complementary purchasing, especially in areas related to home décor, gifting, craft supplies, and themed household items. The prevalence of a limited number of SKUs (Stock Keeping Units) indicates a high rate of inventory turnover and implies that promoting these items in marketing campaigns or cross-selling efforts could greatly enhance overall sales. Furthermore, the observable clustering around theme-based or variant-based products highlights strong potential for strategic bundling

(such as teacup color sets or jumbo bag assortments), improved shelf organization, and tailored recommendation strategies that emphasize matching or coordinating products. In summary, the analysis reveals that customer purchases are both highly focused and driven by patterns, providing clear avenues for decision-making in merchandising, marketing strategies, and inventory management aligned with identified co-purchase behaviors.

Overview of the ECLAT algorithm

Although ECLAT is particularly suitable for discovering frequent itemsets in transactional retail datasets, its application in this analysis presents both benefits and drawbacks. On a positive note, ECLAT is computationally efficient for sparse, high-dimensional data like this one because it utilizes a vertical data format and set intersections instead of iterative candidate generation, making it especially effective at recognizing frequent combinations when the dataset features numerous items yet relatively few in each transaction. Its straightforwardness and speed facilitate the clear identification of high-support individual items and pairs, as shown by the results. However, there are significant limitations of using ECLAT in this situation: it does not inherently provide confidence or lift metrics, which constrains the interpretability of the associations in comparison to algorithms like Apriori or FP-Growth that generate complete association rules. Furthermore, since ECLAT's efficacy significantly diminishes with longer combination lengths, the analysis was limited to pairs rather than delving into more extended, intricate shopping patterns. The algorithm's sensitivity to support thresholds also implies that meaningful but less frequent combinations might be missed if the support is set too high. In summary, ECLAT delivers efficiency and simplicity for identifying frequent itemsets, but it offers a more limited analytical scope and fewer explanatory metrics compared to other association-mining techniques.

For ECLAT, we have used the pyECLAT library in the Python programming language. Due to the large number of code blocks, we have not included our code blocks for this algorithm.

FP-Growth Market Basket Analysis

Key Findings and Insights

In this project, the FP-Growth algorithm was applied to a large retail transaction dataset to identify actionable product associations, co-purchase patterns, and strategic merchandising opportunities. Following extensive preprocessing, the final dataset had 387,985 cleaned records across 18,159 transactions and 3,846 unique products. The average transaction

consisted of 20.81 items, while the median was 15, showing that most consumers have an average purchase size, with some making more sizable purchases.

The FP-Growth algorithm was a desirable solution because of speed and scalability. The model concluded its execution in 8.93 seconds and identified 968 frequent itemsets and 568 association rules, within the user-specified parameters of 1% for support, 30% for confidence, and a lift of 1.2 or greater. The analysis demonstrated that FP-Growth is acceptable for a retail analytics environment where turning insights around quickly is a key consideration.

1. Most Common Items in the Dataset

The analysis of individual items' frequency uncovered a cohort of very frequently purchased products that were present in the majority of transactions. The most frequently purchased item was the White Hanging Heart T-Light Holder, which appeared in 10.57% of all transactions. It was followed closely by Regency Cakestand 3 Tier (8.96%) and Jumbo Bag Red Retrospot (8.66%). The high frequency of these products evidences their broad appeal to customers, and should, therefore, be highlighted in-store, online, or purchasing decisions kept constant to ensure stock availability. Purchases that occurred in 5% or more of transactions are generally be considered as strong candidates for sales opportunities in high traffic merchandising locations.

2. **Strongest Association Rules (Lift > 40)**

The FP-Growth analysis revealed very strong associations between several individual product pairs and small product groups. Very high lift values (> 40) suggest that these items are purchased together in significant frequency compared to random occurrence. These associations are particularly valuable as part of any cross-selling effort, packaging design considerations, or store layout planning.

Top high-lift rules include:

Regency Milk Jug Pink → Regency Sugar Bowl Green (Lift: 56.59)
Regency Tea Plate Green → Regency Tea Plate Roses (Lift: 50.76)
Poppy's Playhouse Livingroom ↔ Bedroom/Kitchen (Lift: 48.00–46.07)
Red Spotty Paper Cups ↔ Red Spotty Paper Plates (Lift: 47.34)
Children's Cutlery Spaceboy ↔ Dolly Girl (Lift: 45.29)
Scandinavian Paisley Picnic Bag ↔ Pink Vintage Paisley Bag (Lift: 41.01)

These rules are the strongest co-purchase signals in the data set. For example, the "Regency" line of dinnerware shows exceptionally high association strength, indicating customers purchase these items together as matching items. Children’s playhouse items and themed party supplies also have similar patterns; customers consistently purchase complementary items together.

3. **Patterns of Co-Purchasing Common to Categories**

There are some categories where strong co-purchase patterns can be observed:

A. Coordinated Tableware Sets

There was a strong lift for Regency-set items (milk jug, sugar bowl, tea plates), which means customers simply love to purchase items in a coordinated set, and are prime candidates for bundled purchases and displays.

B. **Children's Toys and Crafts**

We are continually seeing Poppy's Playhouse items and Spaceboy/Dolly Girl cutlery sets co-occurring, which would support theming gift sets and coordinated displays.

C. Party Supplies

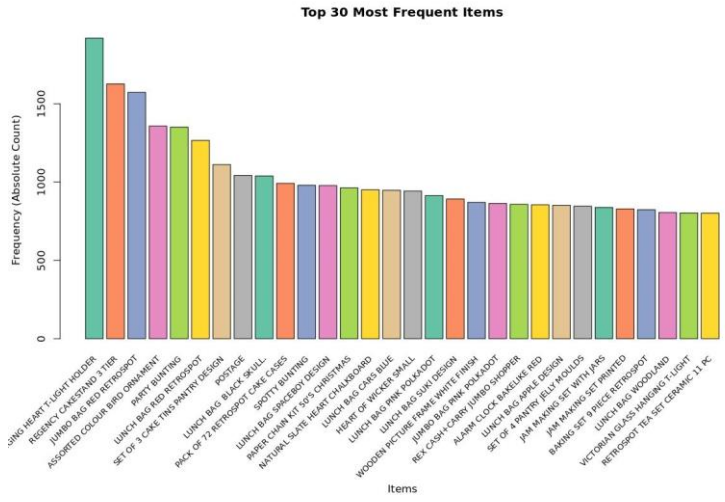
Red spotty cups, plates, and decorations display strong co-purchase patterns, signalling demand for pre-made party packs.

Additional insights, including all visualizations, diagnostic plots, and the full FP-Growth code, are presented in the Appendix for reference.

Top 10 Most Frequent Items:

Rank	Item Name	Support	Frequency
1	WHITE HANGING HEART T-LIGHT HOLDER	10.57%	1,919
2	REGENCY CAKESTAND 3 TIER	8.96%	1,627
3	JUMBO BAG RED RETROSPOT	8.66%	1,573
4	ASSORTED COLOUR BIRD ORNAMENT	7.48%	1,358
5	PARTY BUNTING	7.44%	1,351
6	LUNCH BAG RED RETROSPOT	6.97%	1,266
7	SET OF 3 CAKE TINS PANTRY DESIGN	6.12%	1,112
8	POSTAGE	5.74%	1,043
9	LUNCH BAG BLACK SKULL	5.72%	1,039
10	PACK OF 72 RETROSPOT CAKE CASES	5.46%	992

Figure: Item Frequency Analysis (Absolute)



Interpretation:

This bar chart shows the total absolute purchased counts of the top 30 chosen items. The White Hanging Heart T-Light Holder significantly leads the count by 1,919 purchases, followed by the Regency Cake stand 3 Tier (1,627), and Jumbo Bag Red Retro spot (1,573) completes the top three. The drop off of frequency in terms of items purchased after these top three show the long-tail distribution that is often seen within a retail environment: a higher percentage of sales from a small number of item

Association Rules Analysis

Metric	Minimum	Maximum	Mean	Median
Support	0.0101	0.0170	0.0119	0.0113
Confidence	0.3006	0.8971	0.5847	0.5952
Lift	1.2001	56.588	13.35	10.12
Count	184	309	216	207

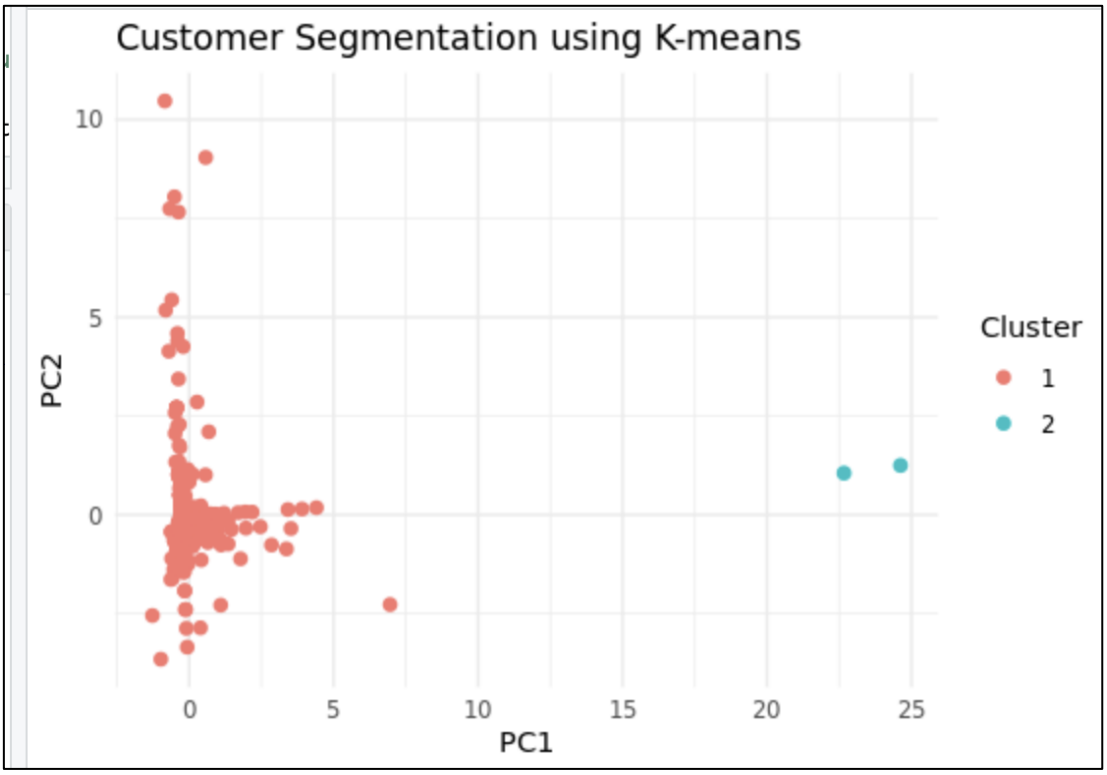
Top Association Rules

The following table presents the top 20 association rules ranked by lift, representing the strongest product associations discovered in the analysis:

Rank	Rule (LHS → RHS)	Support	Confidence	Lift
1	Regency Milk Jug Pink → Regency Sugar Bowl Green	0.0101	0.751	56.59
2	Regency Sugar Bowl Green → Regency Milk Jug Pink	0.0101	0.764	56.59
3	Regency Tea Plate Green → Regency Tea Plate Roses	0.0114	0.841	50.76
4	Regency Tea Plate Roses → Regency Tea Plate Green	0.0114	0.688	50.76
5	Poppy's Playhouse Livingroom → Bedroom	0.0109	0.811	48.00
6	Poppy's Playhouse Bedroom → Livingroom	0.0109	0.645	48.00
7	Set/6 Red Spotty Paper Cups → Plates	0.0126	0.824	47.34
8	Set/6 Red Spotty Paper Plates → Cups	0.0126	0.725	47.34
9	Poppy's Playhouse Livingroom → Kitchen	0.0115	0.852	46.07
10	Poppy's Playhouse Kitchen → Livingroom	0.0115	0.619	46.07
11	Childrens Cutlery Spaceboy → Dolly Girl	0.0112	0.656	45.29
12	Childrens Cutlery Dolly Girl → Spaceboy	0.0112	0.776	45.29
13	Poppy's Playhouse Bedroom → Kitchen	0.0135	0.801	43.31
14	Poppy's Playhouse Kitchen → Bedroom	0.0135	0.732	43.31
15	Scandinavian Paisley Picnic Bag → Pink Vintage	0.0107	0.628	41.01

16	Pink Vintage Paisley Bag → Scandinavian	0.0107	0.698	41.01
17	Small Marshmallows Pink Bowl → Orange Bowl	0.0122	0.779	39.73
18	Small Dolly Mix Orange Bowl → Pink Bowl	0.0122	0.624	39.73
19	Feltcraft Cushion Rabbit → Butterfly	0.0107	0.614	39.25
20	Feltcraft Cushion Butterfly → Rabbit	0.0107	0.683	39.25

K-Means Clustering Market Basket Analysis:



Dataset initialization

A dataset comprising 698 separate customer IDs and purchase volumes for several product categories (more than ten different products) was used for the analysis. Plotting of the generated clusters on the first two principal components (PC1 and PC2) was done using the K-means technique with k = 2 clusters (with the help of the Elbow Method).

Key Findings

1) Cluster Distribution: The great majority of clients are concentrated close to the origin in Cluster 1, suggesting comparable buying habits.
2) Only a small number of clients are located far along PC1 in Cluster 2, indicating extreme purchasing behaviour (perhaps bulk buying or outliers).

3) Principal Component Spread: PC1, which is mostly influenced by purchase volume, captures the most diversity in purchasing behavior.
4) PC2 exhibits less spread, suggesting that overall purchase quantity has a greater impact than product mix variation.

Inference:

When we use K-means clustering with two clusters on your dataset, the results indicate a clear division of clients based on how they buy things and their purchasing behaviour. The majority of clients—696 in total—are found in Cluster 1. These consumers have comparable purchasing habits, which are defined by comparatively modest and steady purchase amounts for a variety of goods. Regular retail consumers who probably make modest but frequent purchases are represented by this category. Marketing techniques like loyalty programs, tailored discounts, and focused promotions would work well for these clients to keep them interested and promote repeat business.

However, there are just two customers in Cluster 2, denoted by Customer IDs 12931 and 16422. Compared to the rest of the customer base, these clients stand out as high-value purchasers since they probably make bulk purchases or have much greater transaction volumes. Their behavior is very different, as evidenced by their distance from the main cluster in the PCA visualization. To improve their relationship with the company, these clients should be handled as premium accounts and given exclusive deals, special discounts, or dedicated account management.

The tight grouping of Cluster 1 near the origin indicates a high level of homogeneity among customers, suggesting that most exhibit similar purchasing behaviors. This uniformity implies limited differentiation opportunities for marketing strategies within this group. Additionally, the sparse spread along PC2 shows that secondary factors, such as diversity in product selection, have minimal influence compared to overall purchase volume, reinforcing that the primary driver of segmentation is purchase quantity rather than product variety

Conclusion

This project was a useful journey to discover the retail dynamics by applying cutting-edge machine learning techniques to derive insights from the retail activities.

Key takeaways:

- The data cleaning process is crucial to getting better insights from the data

- Different algorithms can have different RAM usage, which could impact the efficiency of the analysis process
- Using the different tools can be useful to capture the hidden aspects of the data

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94) (pp. 487-499). Morgan Kaufmann.

Chen, Y. L., Chen, J. M., & Tung, C. W. (2006). A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. Decision Support Systems, 42(3), 1503-1520. https://doi.org/10.1016/j.dss.2006.01.001

Hahsler, M., Grün, B., & Hornik, K. (2005). arules - A computational environment for mining association rules and frequent item sets. Journal of Statistical Software, 14(15), 1-25. https://doi.org/10.18637/jss.v014.i15

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann Publishers.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. ACM SIGMOD Record, 29(2), 1-12. https://doi.org/10.1145/335191.335372

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. Addison-Wesley.

Note: ChatGPT has been used for code debugging purposes.

Appendix

(R- code for k-means)

Load required libraries

```
install.packages("tidyr")
```

```
library(readxl) library(dplyr) library(tidyr) library(ggplot2)

data <- read_excel("CleanedData.xlsx") names(data)

#Prepare data for clustering

Remove CustomerID for clustering

#customer_matrix <- customer_product %>% select(-CustomerID)

#Scale the data

scaled_data <- scale(customer_product)

#Determine optimal number of clusters (Elbow Method)

wss <- sapply(1:10, function(k){ kmeans(scaled_data, centers = k, nstart = 10)$tot.withinss
})

plot(1:10, wss, type = "b", pch = 19, frame = FALSE, xlab = "Number of clusters K", ylab =
"Total within-clusters sum of squares")

#Apply K-means clustering (choose K based on elbow plot)

set.seed(123) k <- 2 # Example: 5 clusters kmeans_result <- kmeans(scaled_data, centers = k,
nstart = 25)

#Add cluster labels to original data

customer_product$Cluster <- kmeans_result$cluster

View cluster summary

print(table(customer_product$Cluster))

#Visualize clusters using PCA

pca <- prcomp(scaled_data) pca_df <- data.frame(pca$x[,1:2], Cluster =
factor(kmeans_result$cluster))

ggplot(pca_df, aes(PC1, PC2, color = Cluster)) + geom_point(size = 2) + labs(title =
"Customer Segmentation using K-means") + theme_minimal()
```