Group 7 - Customer Segmentation and Market Basket Analysis

Team members: Shrey Shah, Gaurav Bhandari, Venkatasubramanian Narasimman

Abstract:
In this project, we examine a dataset of transactions from an online retailer. Using clustering techniques, we segment our customer base according to their purchasing patterns. In order to improve cross-selling opportunities, we use association rule mining to analyze co-purchased items.

Introduction and Motivation:
Data mining on e-commerce customer data is used in the project to support targeted marketing and revenue growth through customer clustering and item association discovery. With the goal of enhancing customer relations, sales, and inventory management in retail and e-commerce, an understanding of customer behavior is essential in today's competitive business environment.

Related Work:
The variables recency, frequency, and monetary basis form the basis of a typical customer segmentation model (Marisa et al., 2019). According to Zhouzhou et al. (2021), a novel methodology for real-time customer segmentation was developed using fashion bags as the product. The techniques for customer segmentation and representation that (Gomes et al. 2023) cover in detail include k-means clustering, RFM analysis, and manual feature selection. Instead of focusing on understanding consumer behavior in general, (Griva et al. 2018) employ K-means clustering to understand shopping behavior at the level of individual visits. In the study by (Mujianto et al., 2019), the Apriori algorithm is investigated. The FP-Growth algorithm, which has excellent scaling capabilities, has been studied by (Arivazhagan et al., 2022). This algorithm is used for mining association rules. The results show that while Eclat can be used effectively for small datasets, the FP-Growth algorithm requires fewer scans than the Apriori algorithm in order to generate item sets.

Methodology:

Exploratory data analysis: The features in the dataset include "InvoiceNo," "StockCode," "Description," "Quantity," and "InvoiceDate." "Country," "CustomerID," and "UnitPrice." It has a variety of data types and some negative values in the fields labeled "Quantity" and "UnitPrice," which could be return transactions. According to summary statistics, 75% of the data for "UnitPrice" is less than £4.13, and 75% of the data for "Quantity" falls between 0 and 5 units. Outliers with 80995 units and £38970 prices might need more research. Since high-value entries account for a large portion of revenue in retail data, even though there are only 241 entries with prices over £50, it is imperative to keep them for clustering analysis.

	types	counts	distincts	nulls	missing_ratio
CustomerID	float64	406829	4373	135080	24.926694
Description	object	540455	4224	1454	0.268311
Country	object	541909	38	0	0.000000
InvoiceDate	datetime64[ns]	541909	23260	0	0.000000
InvoiceNo	object	541909	25900	0	0.000000
Quantity	int64	541909	722	0	0.000000
StockCode	object	541909	4070	0	0.000000
UnitPrice	float64	541909	1630	0	0.000000

	types	counts	distincts	nulls	missing_ratio
Amount	float64	397884	2939	0	0.0
Country	object	397884	37	0	0.0
CustomerID	int64	397884	4338	0	0.0
Description	object	397884	3647	0	0.0
Internal	object	397884	2	0	0.0
InvoiceDate	datetime64[ns]	397884	17282	0	0.0
InvoiceNo	object	397884	18532	0	0.0
Quantity	int64	397884	301	0	0.0
StockCode	object	397884	3665	0	0.0
UnitPrice	float64	397884	440	0	0.0
days_since_last_purchase	float64	397884	360	0	0.0

Fig. 1: Summary statistics of the dataset before and after preprocessing with RFM features

Preprocessing: Records with negative "Quantity" or "UnitPrice" values were eliminated at a percentage of 2.15%. Furthermore, 135080 rows—or 24.6% of the dataset—that had null "CustomerID" were removed.

This stage is essential to our RFM model since it allows us to calculate the customers' and groups' monetary values, frequencies, and recency based on their IDs. Alongside null "CustomerID," null "Descriptions" were discovered. We finally have 397884 rows in our dataset, as shown in Fig. 1. The "CustomerID" data type was modified to an integer. There are discrepancies in the "Description," with several spelling variations for the same item—possibly as a result of typos. To maintain consistency, we changed inaccurate descriptions to the mode value.

Adding new features: We used an RFM (Recency, Frequency, Monetary Value) model to assess customer value because the dataset is restricted to sales records, as stated in (Marisa et al., 2019). The number of days before the reference date that a customer last made a purchase is used to calculate the Recency feature. The last day the customer made a purchase is selected as the reference date. The frequency of a customer's purchases at the store is determined by their frequency, and the product of the item quantity and unit price yields the monetary value. As shown in Fig. 2, these variables have a similar distribution with a positive skew and high kurtosis, indicating outliers. To guarantee consistent input to the algorithm, we applied the log transformation to resolve the issue.

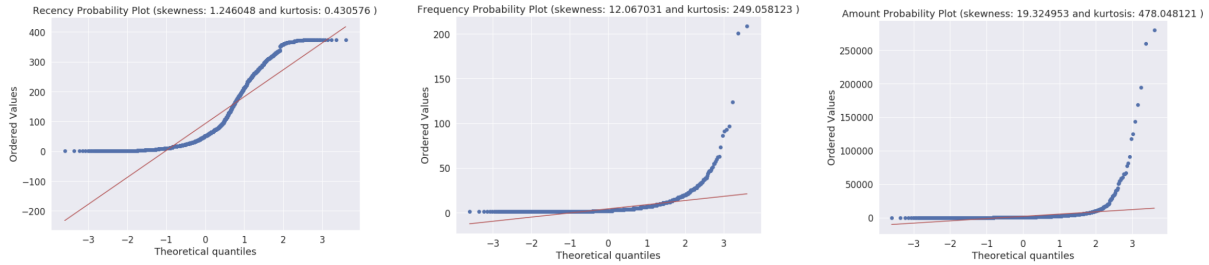

Fig. 2: QQ plot for Recency, Frequency and Monetary Value (Amount) features

Customer Segmentation: We utilized K-means++ clustering to segment customers based on log-scaled RFM features, aiming to maximize inter-cluster distance and minimize intra-cluster variance. We opted for K-means++ over K-means due to its improved centroid initialization, leading to higher-quality clusters. To determine the optimal number of clusters (k), we applied the elbow method, assessing the 'within-cluster sum of squares' for k values from 2 to n (Refer Fig 3(a)). Hyperparameters for K-means++ included clusters ranging from 2 to 7, 10 initializations to choose the best result, a maximum of 300 iterations with early stopping, a tolerance of 1e-4 for convergence, and a fixed random state (101) for consistent centroid initialization across runs. The clustering quality assessment involved the Silhouette score, Dunn Index, and Davies-Bouldin Index. Higher Silhouette scores and Dunn Index, along with lower Davies-Bouldin Index values, indicated superior cluster quality.

We experimented with Agglomerative clustering, where we used the Euclidean distance for the 'affinity' parameter. We ran hyperparameter tuning for the 'n_clusters' from 2 to 7 and tested 2 different types of 'linkage' properties: 'complete' and 'ward'. 'Ward' helps minimize the cluster variances, and 'complete' uses the maximum distances between all observations of the two sets. We have not experimented with other hyperparameters. For assessing clustering quality, we use the Silhouette score and Dunn Index.

From our observations based on K-means++ and Agglomerative clustering and our exploratory analysis, we encountered that the dataset might be noisy, and hence we decided to try the DBSCAN algorithm. For this, we need the parameters 'eps' and 'min_samples'. We decided to only tune the 'eps' and 'min_samples' hyperparameters by calculating the silhouette scores for 'min_samples' from 2 to 9. Based on the optimal values for 'eps' and 'min_samples', we performed DBSCAN clustering. To measure the clustering quality, we use the Silhouette score and the Davies-Bouldin index, since the Dunn Index is usually calculated for partitioning-based algorithms and DBSCAN is a density-based algorithm.

Association Rule Mining: We utilized the original dataset, retaining rows with missing Description and null CustomerID (24.6% of data) for critical transactional insights. Apriori served as our baseline for association rule mining, using support for frequent itemsets, and assessing rule quality with confidence, lift, and Kulczynski similarity. We employed the Kulczynski measure for its null-invariant nature, addressing null transaction sensitivity observed in Lift. The imbalance ratio with the Kulczynski measures the imbalance between two itemsets in rule implications. Kulczynski measure is a better estimate than Zhang's metric and lift as it is a null invariant measure and values closer to 1 indicate that items are common in many itemsets. The Imbalance Ratio should have small values closer to 0 indicating balanced rules.

To address Apriori's large search space, we pruned itemsets by focusing on the top 15 items contributing to high sales and having a minimum of 2 items in transactions (Table 1 in Appendix). This revealed common associations among frequently bought items. We also explored FP-Growth as a faster alternative, experimenting with support thresholds from 0.01 to 0.05 to get an estimate of the number of association rules and the mean confidence of those rules. Using a minimum support of 0.01, we considered transactions with at least 1 item, generating insightful rules across all item sets. Increasing support to 0.02 significantly reduced the number of rules (Table 2 in Appendix).

Results:

Customer Segmentation:

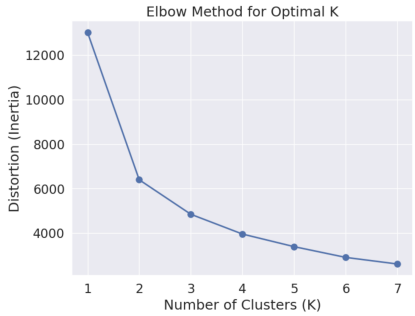

Fig. 3(a): Elbow plot for optimal k using K-Means++

Using the elbow method as seen in Fig. 3 (a), we explored k=2 to 7 and found that k=2 is optimal. The scores for this range using K-Means++ are as follows:
Silhouette scores: [**0.43,** 0.34, 0.33, 0.30, 0.31, 0.31],
Dunn Indices: [**1.472**, 1.099, 1.012, 0.888, 0.903, 0.867]
Davies-Bouldin Index: [**0.887**, 1.043, 1.0162, 1.047, 1.013, 0.982]

Fig. 3(b): Silhouette analysis and visualization of 2 clusters using K-Means++

k=2 had the highest Silhouette score and Dunn Index, and the lowest Davies-Bouldin Index, indicating better cluster quality. Plots for 2 cluster groups (k=2) show distinct customer segments based on the log (amount spent) vs. log (recency) and log (amount) vs. log (frequency) are plotted as seen in Fig. 3 (b).

Fig. 4 shows us the results for the Agglomerative Clustering when performed on the same dataset using 2 clusters. The type of linkage chosen is 'ward' linkage as it helps minimize the cluster variances. When compared with the K-Means++ results, we see that the clusters show almost similar characteristics with some misclassifications for the green cluster (negative silhouette score). We see very similar clustering with 'complete' linkage.

The scores for Agglomerative clustering (for k=2 to 7) are as follows. We can see the Silhouette score and Dunn Index is highest for k=2 clusters.
Silhouette scores: [**0.41**, 0.33, 0.23, 0.26, 0.26, 0.24],
Dunn Indices: [**1.363**, 1.019, 0.954, 0.872, 0.783, 0.718].

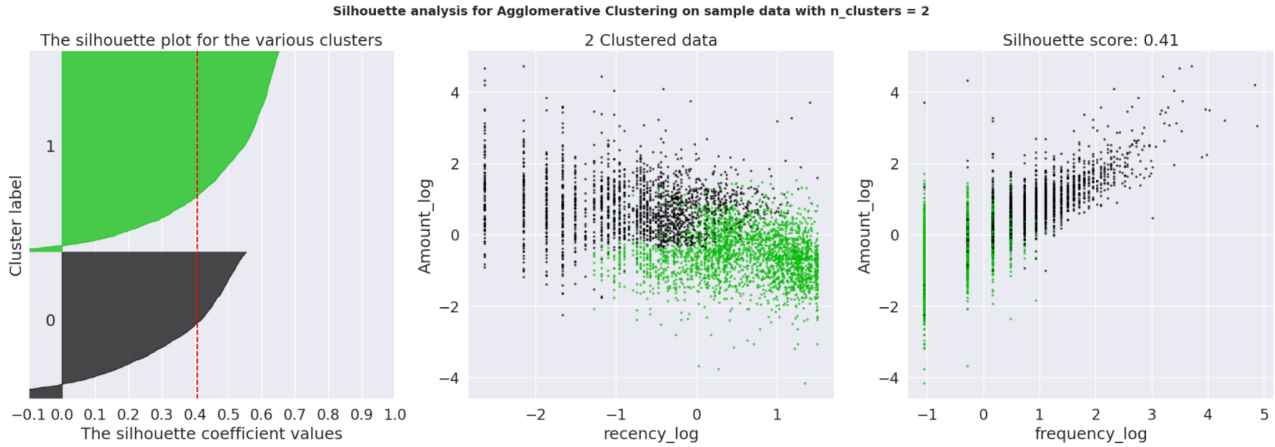

Fig. 4: Silhouette analysis and visualization of 2 clusters via Agglomerative Clustering (Ward's criterion)

We varied the 'min_samples' from 2 to 9 and calculated the silhouette scores using the DBSCAN algorithm as shown in Fig. 5 (a). The optimal value for 'min_samples' was 5 for which the Silhouette score was the highest is 0.534. Using this value we obtained the optimal 'eps' value of 0.5 using the KneeLocator function and the clusters when visualized are seen in Fig 5. (b). When we observe Fig 5(b), we see three different colours. The colours yellow and green represent the 2 clusters whereas the data points in purple indicate the noise or outliers. These indicate customers who have a transaction with a high Amount indicating high purchasing capacity.

The DBSCAN algorithm gave a silhouette score of **0.32** and Davies-Bouldin index of **1.64** (for k=2).

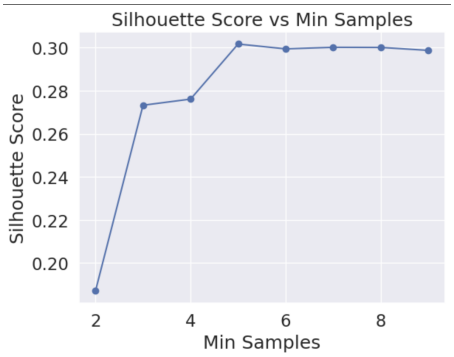

Fig. 5(a): Plot for Silhouette scores vs min_samples for DBSCAN

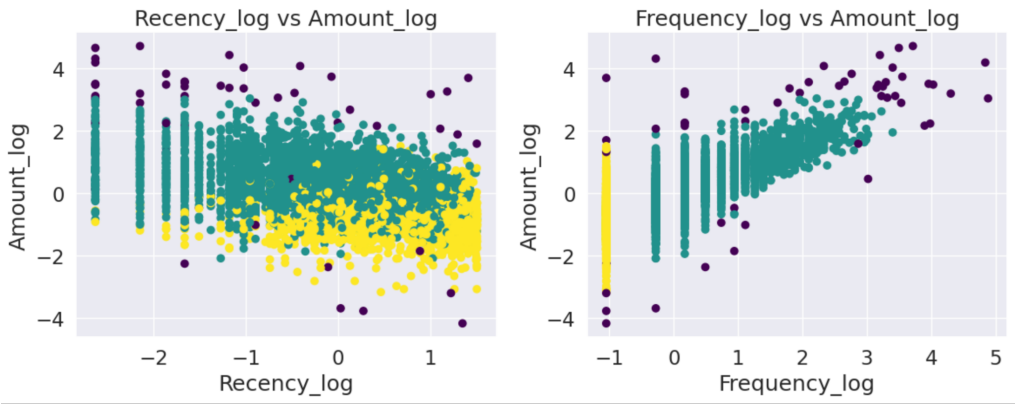

Fig. 5(b): Silhouette analysis and visualization of 2 clusters using DBSCAN

We discovered that, in all the clustering techniques we employed, they divide the customers into two major segments. Low-engagement customers with low frequency and recency make up one of the clusters, which logically leads to lower spending levels. To try and entice these customers to keep returning and shopping on the platform, promotions and discounts are the perfect fit for them. The other cluster stands for devoted, high-spending clients who make more frequent purchases and are more recent in their purchases; these clients are worthy of loyalty benefits and retention initiatives.

Association Rule Mining:

With a confidence level of 0.6 and support thresholds of 0.01 and 0.03 on the transactions, association rule mining was done. We obtained 461 frequent item sets, 164 rules, and 98 frequent item sets, 11 rules, respectively, from these settings. Based on the length of the two-item transactions and the percentage of total sales, a data pruning technique was used to retain only the most important transactional data. This made it possible to explore important associations in a meaningful way. Some of the frequent item sets observed include 'Jumbo Bag Red Retrospot', 'White Hanging Heart T-light Holder', 'Lunch Bag Red Retrospot', 'Regency Cakestand 3 Tier.'

Lift, support, confidence, and the Kulczynski measure were used to evaluate the rules. Fig. 6 shows the top 5 rules that were ranked using Kulczynski similarity. Reliability is indicated by ideal confidence levels of up to 0.7 for rules. A strong positive correlation is indicated by a lift value greater than 1. Kulczynski measure values for the top rules derived using the Apriori algorithm vary in the range of 0.43 to 0.68, indicating a low to average association between item sets. The Imbalance Ratio is also on the higher side, which tells us that the rules are imbalanced.

antecedents	consequents	confidence	kulczynski_similarity	imbalance_ratio
(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.749101	0.681563	0.334450
(JUMBO STORAGE BAG SUKI)	(JUMBO BAG RED RETROSPOT)	0.673713	0.637243	0.333795
(LUNCH BAG BLACK SKULL.)	(LUNCH BAG RED RETROSPOT)	0.600556	0.687016	0.173056
(WHITE HANGING HEART T-LIGHT HOLDER, JUMBO BAG...	(JUMBO BAG RED RETROSPOT)	0.709302	0.425883	0.823436
(WHITE HANGING HEART T-LIGHT HOLDER, JUMBO STO...	(JUMBO BAG RED RETROSPOT)	0.697368	0.432616	0.791908

Fig. 6: Top 5 association rules with support (0.01), confidence (0.6), ranked using Kulczynski similarity using Apriori algorithm

Lift, support, confidence, and the Kulczynski measure were used to evaluate the rules using FP-Growth. Fig. 7 shows the top 5 rules that were ranked using Kulczynski similarity. We observe a difference in the top 5 rules using the Apriori and FP-Growth. This is because we consider transactions with a minimum of two itemsets in the Apriori algorithm. We take all the products with a length of transaction >=1 for FP-Growth and do not prune the tree. Kulczynski measure values for the top rules derived using the FP-Growth algorithm vary in the range of 0.944 to 0.979, indicating a very strong association between item sets. The Imbalance Ratio is also on the lower side closer to 0 which means that the rules are balanced.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	kulczynski_similarity	imbalance_ratio
(REGENCY TEA PLATE GREEN)	(REGENCY TEA PLATE PINK)	0.015790	0.012845	0.011536	0.730570	56.877430	0.979935	0.172249
(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.043238	0.032766	0.026344	0.609272	18.594571	0.964436	0.210873
(POPPY'S PLAYHOUSE BEDROOM)	(POPPY'S PLAYHOUSE LIVINGROOM)	0.017426	0.013499	0.010717	0.615023	45.560193	0.952966	0.194332
(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.021558	0.017508	0.014317	0.664137	37.933374	0.947723	0.163636
(REGENCY TEA PLATE ROSES)	(REGENCY TEA PLATE GREEN)	0.018694	0.015790	0.013172	0.704595	44.623145	0.944267	0.136276

Fig. 7: Top 5 association rules with support (0.01), confidence (0.6), ranked using Kulczynski similarity using FP-Growth algorithm

Based on the rules obtained from the algorithms, we can understand the frequent items that the customers generally purchase. The antecedents of the rules give us a good estimate as to what consequent items will have a high chance of being purchased along with it. Hence such rules can be useful recommendations for those products bought together which is equivalent to placing these products together in a physical store.

Conclusion:

In summary, we utilized the RFM model with three clustering methods (K-means++, Agglomerative with ward and complete linkage, and DBSCAN) for customer segmentation, identifying low-frequency, low-revenue customers and high-value customers. Apriori and FP-Growth algorithms were employed on pruned and complete data to reveal purchasing patterns and association rules, improving customer recommendations and boosting sales. This integrated approach informs targeted marketing and enhances customer satisfaction, enabling data-driven decisions. Future extensions may involve exploring algorithms like Eclat and Prefixspan for sequential pattern mining, and integrating demographic data for more refined customer analysis based on age, income, and occupation.

References:

Arivazhagan, B., Pandikumar, S., Sethupandian, S. B., & Subramanian, R. S. (2022). Pattern discovery and analysis of customer buying behavior using association rules mining algorithm in e-commerce. *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. https://doi.org/10.1109/iceeict53079.2022.9768473

Gomes, Miguel Alves, and Tobias Meisen. "A Review on Customer Segmentation Methods for Personalized Customer Targeting in E-Commerce Use Cases - Information Systems and e-Business Management." *SpringerLink*, Springer Berlin Heidelberg, 9 June 2023, link.springer.com/article/10.1007/s10257-023-00640-4

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using Market Basket Data. *Expert Systems with Applications*, *100*, 1–16. https://doi.org/10.1016/j.eswa.2018.01.029

Marisa, F., Syed Ahmad, S. S., Mohd Yusof, Z. I., Fachrudin, F., & Akhriza Aziz, T. M. (2019a). Segmentation model of customer lifetime value in small and medium enterprise (smes) using K-means clustering and LRFM model. *International Journal of Integrated Engineering*, *11*(3). https://doi.org/10.30880/ijie.2019.11.03.018

M.R. Narasingha Rao, D., V.L Sita Ratnam, K., D.S. Prasanth, M., & Lakshmi Bhavani, P. (2018). A survey on analysis of online consumer behaviour using association rules. *International Journal of Engineering & Technology*, *7*(2.32), 206. https://doi.org/10.14419/ijet.v7i2.32.15568

Mujianto, A. H., Mashuri, C., Andriani, A., & Jayanti, F. D. (2019). Consumer Customs analysis using the Association rule and Apriori Algorithm for determining sales strategies in retail central. *E3S Web of Conferences*, *125*, 23003. https://doi.org/10.1051/e3sconf/201912523003

Yan, Zhouzhou and Zhao, Yang, "Customer Segmentation Using Real Transactional Data in E-Commerce Platform: A Case of Online Fashion Bags Shop" (2021). ICEB 2021 Proceedings (Nanjing, China). 12.
https://aisel.aisnet.org/iceb2021/12

Appendix

	item_name	item_count
0	WHITE HANGING HEART T-LIGHT HOLDER	2302
1	REGENCY CAKESTAND 3 TIER	2169
2	JUMBO BAG RED RETROSPOT	2135
3	PARTY BUNTING	1706
4	LUNCH BAG RED RETROSPOT	1607
5	ASSORTED COLOUR BIRD ORNAMENT	1467
6	SET OF 3 CAKE TINS PANTRY DESIGN	1458
7	PACK OF 72 RETROSPOT CAKE CASES	1334
8	LUNCH BAG BLACK SKULL.	1295
9	NATURAL SLATE HEART CHALKBOARD	1266
10	POSTAGE	1250
11	JUMBO BAG PINK POLKADOT	1231
12	JAM MAKING SET WITH JARS	1220
13	HEART OF WICKER SMALL	1212
14	JUMBO STORAGE BAG SUKI	1201

Table 1. Top 15 items found in the transactions

min_support	num_rules	avg_confidence
0.01	223	0.708262
0.02	22	0.691555
0.03	3	0.706136
0.04	0	NaN
0.05	0	NaN

Table 2. Summary of minimum support, association rules and average confidence