ASSIGNMENT 3: GROUP REPORT

Group 18

Sharlene Rodrigues A1894739

Tanveer Akram A1904700

Yurou Qiu A1893163

Executive Summary

Our project focuses on optimizing product placement and boosting sales at our grocery store by analyzing transaction data from our loyalty program to uncover common purchasing patterns. This analysis identifies products that customers frequently buy together, providing a basis for enhancing store layouts and promotional strategies to better align with customer behavior. The main goal is to arrange the store so that items that are often bought together are placed near each other. This makes shopping more convenient for our customers, encouraging them to buy more during their visits, which in turn boosts our sales. Furthermore, by understanding which items are commonly purchased together, our marketing team can target promotions more effectively, offering deals on combinations known to be popular. Another benefit of this project is improving how we manage our stock. With better insight into which products are in high demand, we can make sure these items are always available, reducing the risk of stockouts during peak periods. This efficient management helps us save on costs and ensures customer satisfaction. The approach we've taken is designed to grow with our store. As we collect more data and our customer base expands, our system can handle the increase, ensuring that the insights we gain remain relevant and useful. In summary, this project uses customer purchase data to make our store layout more intuitive, promotions more appealing, and stock levels more reliable.

Introduction

The report outlines a project aimed at enhancing the shopping experience and increasing sales in a grocery store by optimizing product placements and improving online recommendation systems. By analyzing transaction data from our loyalty program, which includes approximately one million customer interactions, the project seeks to identify common patterns of items frequently purchased together. The insights gained from these patterns will guide strategic decisions about product arrangement within the store, aiming to create a more intuitive shopping environment that encourages additional purchases. Moreover, the project will refine our online recommendation system by integrating two key methodologies: frequent itemset mining and collaborative filtering. This dual-framework approach will help develop a sophisticated recommendation engine tailored to individual customer preferences. This report will detail the methodologies employed to uncover

purchasing patterns, describe the development of the recommendation systems, and discuss their implementation. Through this data-driven approach, the project intends to improve operational efficiency, enhance customer satisfaction, and drive sales growth.

Exploratory Analysis

The exploratory analysis of product sales (see Figure 1) reveals that items like "whole milk," "other vegetables," and "rolls/buns" are consistently popular among customers, indicating their status as essential staples. This pattern suggests that these items should be prominently displayed to facilitate impulse buys and enhance sales.

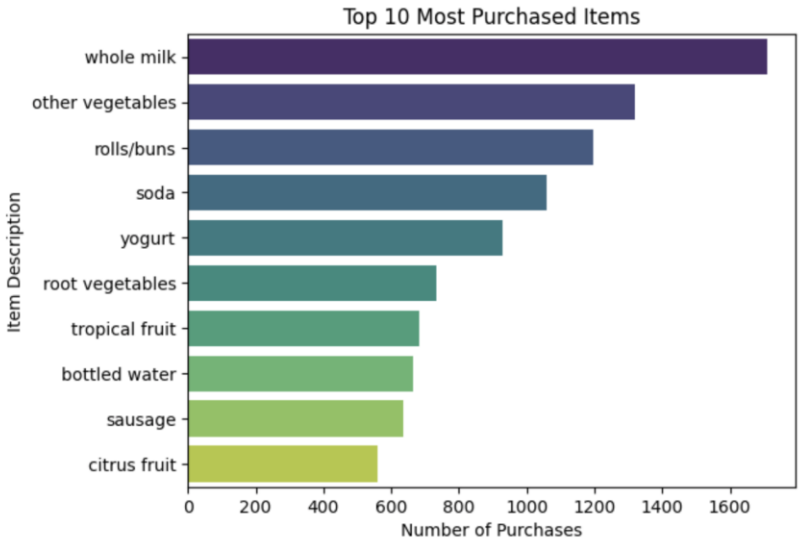

Figure 1. Most Purchased Items

Further analysis of purchase trends (see Figure 2) across different days of the week shows a stable flow of customers, with minor day-to-day fluctuations. This steady customer traffic supports the need for inventory management to ensure that stock levels are continuously aligned with demand.

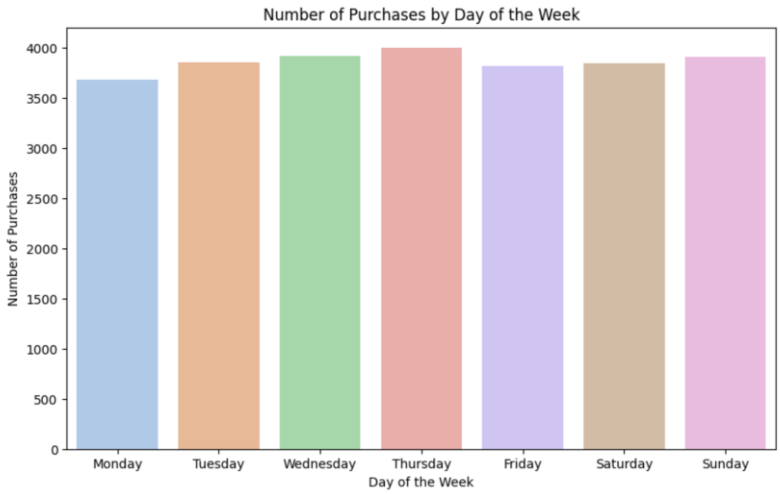

Figure 2. Frequency of Purchases by Weekday

Additionally, customer loyalty data (see Figure 3) reveals that a significant majority, 96.1%, of purchases are made by repeat customers, highlighting the effectiveness of existing loyalty programs and the potential for targeted marketing campaigns to further leverage these buying patterns. The exploratory analysis highlights the importance of maintaining sufficient inventory to meet constant demand and devising strategies to transform occasional buyers into regular customers.

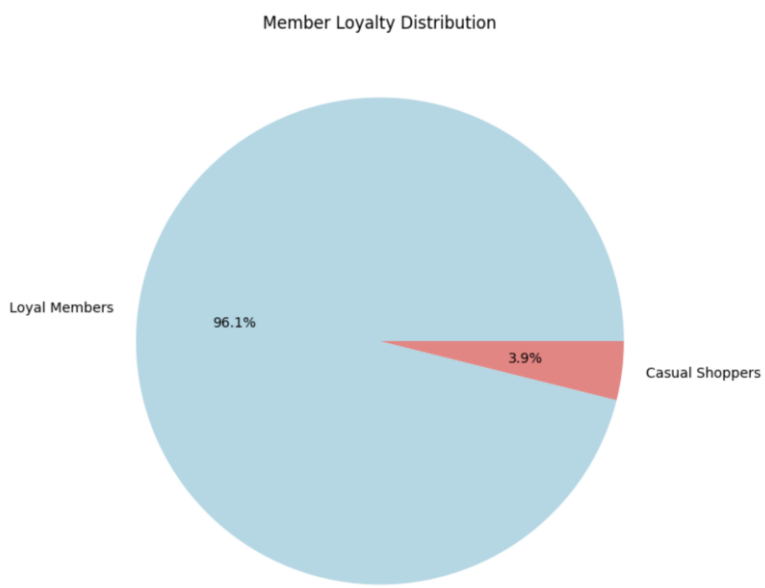

Figure 3. Distribution of Customers

Based on the insights gained from the exploratory analysis, we have chosen the Apriori algorithm to analyze our transaction data. This decision stems from the algorithm's proficiency in identifying frequent itemsets, which directly corresponds with our observation of staples like "whole milk"

and "vegetables" frequently appearing in customer purchases. The Apriori algorithm's reliance on support and confidence metrics makes it particularly suitable for our dataset, which is characterized by high customer retention and consistent daily transactions. These metrics enable us to generate reliable association rules that are instrumental in guiding strategic product placement and targeted marketing strategies. Additionally, the scalability and straightforward implementation of the Apriori algorithm allow it to efficiently manage our increasing data volumes, providing clear and actionable insights that store management can readily utilize to enhance operational efficiency and customer experience.

Building on this analytical foundation, as illustrated in our system diagram (see Figure 4), we extend the use of the Apriori algorithm's insights to refine our store operations further. By integrating these findings with content-based and collaborative filtering techniques, we can optimize product placements and generate personalized recommendations.

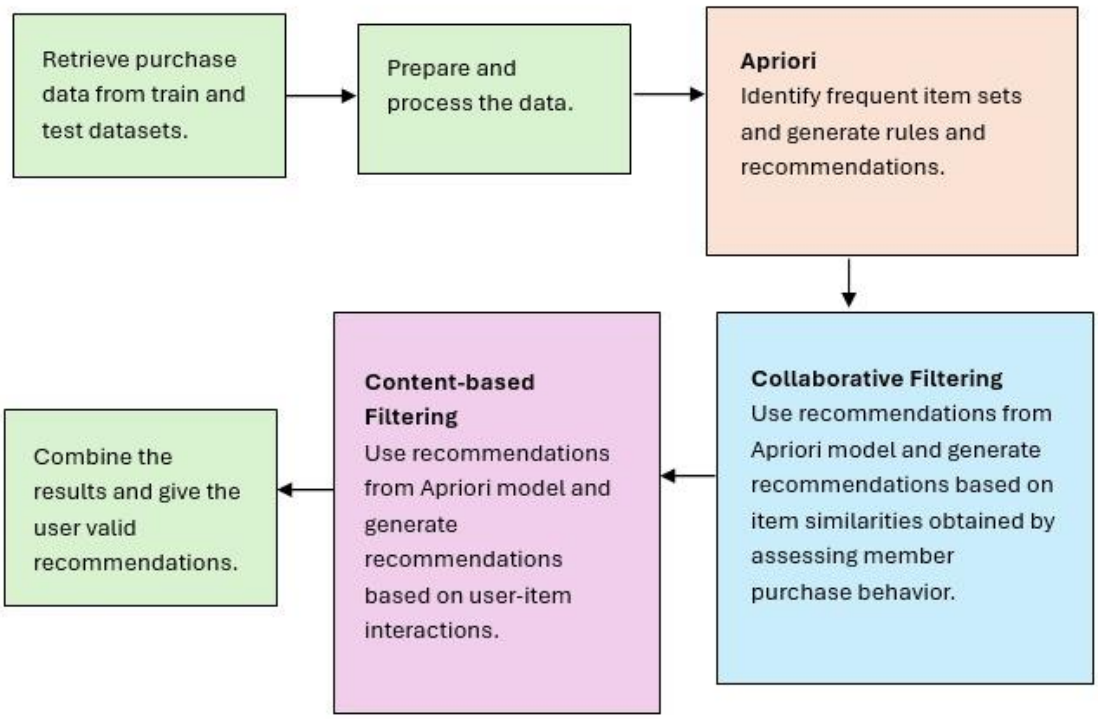

Figure 4. System Diagram

Frequent pattern mining

The frequent pattern mining process starts with data preprocessing, where transaction data is organized into a binary matrix format suitable for pattern recognition. The Apriori algorithm identifies frequent itemsets, which are groups of items commonly purchased together, with a set minimum support level to ensure relevance. From these frequent itemsets, association rules are generated, which help uncover relationships between products, such as determining that customers who buy milk may also often buy bread. Further, a personalized recommendation system is developed. This system utilizes the association rules to suggest products to customers based on their previous purchases, aiming to enhance the shopping experience and increase sales. The user-item interaction matrix, a core component of this system, tracks which items each customer buys and is subsequently augmented with predictions about other products they might like. Finally, the output of the user-item matrix provides a visual representation of customer purchasing behaviors, allowing for easy identification of popular items and potential opportunities for targeted marketing (Luna et. al. 2018, p. 2860).

In our analysis, we've set specific parameters for the Apriori algorithm to ensure effectiveness in identifying purchasing trends. The support threshold is fixed at 1%, which means an itemset must be present in at least 1% of all transactions to be considered significant. This helps in focusing our marketing and inventory efforts on the most impactful product combinations. Additionally, we use a confidence threshold of 10% for creating association rules. The rationale behind these thresholds is to balance precision and relevance: a 1% support threshold filters out less common itemsets, focusing on those that can genuinely drive sales, while a 10% confidence level ensures the rules we derive are robust enough to base our marketing decisions on, avoiding actions based on rare or random occurrences.

Collaborative Filtering

Collaborative filtering is a type of recommender that assumes that similar users have similar likes and hence uses the past user interactions to generate recommendations. (Kotu & Deshpande, 2019). Collaborative filtering can be either user-based or item-based. User based filtering recommends items based on similarities between user profiles whereas item-based filtering makes recommendations based on item profiles.

Item based collaborative filtering is usually preferred as it is much easier to compare and group items than users because most times, there are more users than items. In cases where the number of items exceeds the number of users, user based collaborative filtering is preferred. Analyzing the dataset provided shows that there are significantly more users than items.

	count
Members	3872
Items	167

Figure 5. Number of Members and Items in the Dataset

Figure 5 shows that there are 3872 members as compared to 167 items. Hence, item based collaborative filtering would be the ideal choice in this case.

Item-based collaborative filtering groups items based on a single user's interaction with them, considering an item to be like another where the user interaction is similar (Kotu & Deshpande, 2019). This is then performed on all user item pairs and a consolidated similarity score matrix is generated based on which the recommendations are made.

A matrix indicating which items have frequently been bought together (co-purchase matrix) is created. After this, a similarity matrix for all items using the similarity function. The metric used for similarity is cosine similarity. This matrix contains the similarity value between every item pair. A user item interaction matrix which records the items purchased by each user is then generated and for each item that does not have a rating, the rating is predicted by computing the weighted average of the ratings of the most similar items, where the weights are the similarities. The top N items with the highest predicted ratings are then provided as recommendations.

Evaluation Metrics

A key part of the collaborative filtering algorithm is generating the similarity matrix. This must be done considering a similarity metric. There are various metrics that can be used. Some of the common metrics used are Jaccard similarity, Cosine similarity, and Pearson correlation coefficient (Kotu & Deshpande, 2019). In this case, the cosine similarity metric has been considered.

The following formula can be used to obtain the value for cosine similarity between 2 vectors.

$$\text{Cosine similarity } (|x \,.\, y|) = \frac{x\,y}{||x||\;||y||}$$

Figure 6. Cosine Similarity Formula (Kotu & Deshpande, 2019**)**

Using this formula, a function to calculate this value is built and a matrix of similarity values between every pair of items is generated.

	frankfurter	pork
frankfurter	0.000000	0.131262
pork	0.131262	0.000000

Figure 7. Similarity Matrix

For example, in the matrix generated above, the cosine similarity score between frankfurter and pork is 0.131262. This score is generated based on the co-purchase matrix, which records how many times a pair of items have been bought together. This means that a member that purchases frankfurters is ~13% likely to buy pork in the same transaction.

This algorithm can be analyzed and improved using other metrics such as Jaccard similarity or Pearson correlation.

Recommendation method from frequent patterns

The model used here is a hybrid system that combines item-based collaborative filtering and content-based filtering, receiving frequent pattern recommendations from the Apriori algorithm.

The Apriori algorithm generates rules based on past transactions made by members, identifying item sets that frequently appear together. The collaborative filtering model utilizes a co-purchase matrix to find items that are frequently bought together, providing insights into item similarities and generating recommendations based on these associations. The content-based filtering model generates recommendations based on items that members have previously interacted with, leveraging item attributes and user preferences. The hybrid model combines recommendations from both the collaborative and content-based models to deliver more accurate suggestions. Due to limited data availability, both models are given equal weightage. The frequent patterns are integrated into the user-item matrix, which serves as input to each filtering model, enhancing the overall recommendation process.

Discussion of results

The code employs collaborative filtering, using the co-purchase matrix to differentiate item associations among users, with cosine similarity as the key metric for pattern identification. It predicts ratings for unrated items, prioritizing the top-N recommendations based on these ratings. Content-based filtering, using TF-IDF vectors and cosine similarity, tailors recommendations to

users' preferences. Hybrid recommendations merge collaborative and content-based approaches, ranking patterns and recommendations by evaluating support, confidence, predicted ratings, and similarity scores. Overall, patterns and recommendations are ranked based on these metrics, ensuring relevance and effectiveness in user-specific contexts.

Table 1. Frequent Patterns from Training Set

Serial Number	Item	Pattern	Support	Confidence
1	herbs	whole milk	0.012397	0.436364
2	UHT-milk	tropical fruit	0.011105	0.210784
3	rolls/buns	beef	0.025826	0.098135
4	chicken	tropical fruit	0.010589	0.145907
5	brown bread	tropical fruit	0.019370	0.189873

Table 2. Frequent Patterns from Test Set

Serial Number	Item	Pattern	Support	Confidence
1	frankfurter	root vegetables	0.020919	0.206107
2	butter	canned bee	0.012913	0.137363
3	rolls/buns	root vegetables	0.010376	0.075510
4	hole milk	tropical fruit	0.019069	0.095506
5	frankfurter	whole milk	0.010376	0.221557

Table 3. 10 examples of recommendations

Member Number	Items Bought	Items Recommended
1889	Herbs, UHT-milk, candles, rolls/buns, chicken	Soups, tropical fruit, sweet spreads, canned beer, berries, long life bakery product, bottled beer, frozen meals, whole milk, cake bar
1076	whole milk, bottled water, frankfurter, ham, brown bread, butter, chicken	Tropical fruit, root vegetables, canned beer, hygiene articles, snack products, bottled beer, candles, butter, whole milk, brown bread

```
Top 3 rules based on support:
     support  confidence      lift         antecedents         consequents
32  0.037297    0.186798  1.235846        (whole milk)  (other vegetables)
33  0.037297    0.246753  1.235846  (other vegetables)        (whole milk)
48  0.028043    0.140449  1.022128        (whole milk)        (rolls/buns)

Top 3 rules based on confidence:
     support  confidence      lift         antecedents   consequents
61  0.016545    0.258772  1.296040     (shopping bags)  (whole milk)
1   0.010095    0.255319  1.278747              (beef)  (whole milk)
33  0.037297    0.246753  1.235846  (other vegetables)  (whole milk)

Top 3 rules based on lift:
     support  confidence      lift         antecedents         consequents
21  0.013741    0.217778  1.440808         (pip fruit)  (other vegetables)
20  0.013741    0.090909  1.440808  (other vegetables)         (pip fruit)
8   0.010376    0.141762  1.319908     (bottled water)            (yogurt)
```

Fig 9. Frequent Patterns based on Support, Confidence and Lift

Support: These rules highlight the most frequently occurring associations between items in the dataset. For example, the first rule suggests that 'whole milk' and 'other vegetables' are frequently bought together, with a support of 0.0373, indicating that approximately 3.73% of transactions contain both items.Support is used to identify how frequently an itemset appears in the dataset, helping to uncover popular item combinations.

Confidence: These rules emphasize the strength of the association between antecedent and consequent items. For instance, the first rule indicates that when 'shopping bags' are purchased, there's a 25.88% chance that 'whole milk' will also be bought. Confidence measures the reliability of the association between items, indicating the proportion of transactions containing the antecedent where the consequent also occurs.

Lift: These rules highlight the significance of the association between antecedent and consequent items, considering the support of the items individually. For example, the first rule suggests that the likelihood of buying 'pip fruit' and 'other vegetables' together is 1.44 times higher than if their occurrence was independent. Lift measures how much more often the antecedent and consequent occur together than if they were independent of each other. It helps to identify associations that are more significant than random chance.

```
User ID: 3486
Items previously bought:
- hamburger meat
- candy
- soap
- whole milk
- pip fruit
- citrus fruit
- spread cheese
- yogurt
Recommendations:
- whole milk
- soap
- male cosmetics
- tropical fruit
- candy
- pip fruit
- jam
- baking powder
- yogurt
- ice cream
```

```
User ID: 3486
Items previously bought:
- hamburger meat
- candy
- soap
- whole milk
- pip fruit
- citrus fruit
- spread cheese
- yogurt
Recommendations:
- canned fish
- whole milk
- instant coffee
- dessert
- salty snack
- tropical fruit
- citrus fruit
- zwieback
- dishes
- candy
```

Figure 10. Recommendations for a sample from train set without and with patterns

```
User ID: 2672
Items previously bought:
- frozen meals
- liquor (appetizer)
- soda
Recommendations:
- organic sausage
- ham
- sausage
- cake bar
- yogurt
- cleaner
- pastry
- light bulbs
- beverages
- organic products
```

```
User ID: 2672
Items previously bought:
- frozen meals
- liquor (appetizer)
- soda
Recommendations:
- whole milk
- dessert
- flour
- potato products
- sausage
- waffles
- ham
- yogurt
- decalcifier
- pastry
```

Figure 11. Recommendations for a sample from test set without and with patterns

The time taken for running the model on the test dataset containing 11766 transactions is ~237 seconds (about 4 minutes). Based on further calculation, it is estimated that 1 million transactions will take ~21000 seconds (about 6 hours).

Conclusion & Recommendations

In this project, we transformed raw transaction data into meaningful patterns and generated recommendations. Using the Apriori algorithm, we identified frequent itemsets and extracted association rules based on support, confidence, and lift metrics. These patterns provided insights

into customer purchase habits. Additionally, we implemented content-based and collaborative filtering models to make personalized recommendations. By combining these approaches in a hybrid model using a weighted-sum method, we leveraged **recommendations from frequent patterns** to generate accurate results.

For practical implementation, we recommend using the hybrid model that leverages frequent pattern recommendations along with collaborative and content-based filtering. This approach combines the strengths of all methods, providing a comprehensive recommendation system that can adapt to different user behaviors and preferences (Kim et. al. 2006).

Scaling up this system involves integrating it into a real-time recommendation engine, which would significantly enhance the customer experience by providing personalized suggestions. This can lead to increased customer satisfaction, higher sales, and improved customer retention. Additionally, using a hybrid approach ensures that the system remains robust and flexible as more data is accumulated.

For future improvements, it is recommended to enhance the hybrid recommendation system by incorporating additional contextual information, such as temporal and demographic data, to further personalize recommendations. Implementing more advanced machine learning techniques, such as neural collaborative filtering, can also improve prediction accuracy.

Reflection

Throughout this project, we've gained valuable insights into the use of metrics such as support, confidence, and lift in mining frequent patterns from transactional data. These metrics help us understand how often items are purchased together (support), the strength of the relationship between them (confidence), and whether their association is significant or just by chance (lift). Employing the Apriori algorithm has been effective for discovering these item sets in large datasets by eliminating infrequent combinations early on. Additionally, we utilized collaborative filtering to predict user preferences based on behaviors of similar users, and content-based filtering to recommend items by matching item attributes with user profiles.

The foundation of our analysis was a comprehensive exploratory analysis, which emphasized the importance of meticulous data preparation. This initial phase was critical in identifying key trends, such as frequently bought product combinations and high levels of customer loyalty, setting a solid base for the application of advanced analytical techniques like the Apriori algorithm. The insights obtained from the analysis were instrumental in guiding further strategic decisions. This project has underscored the power of combining multiple analytical methods to extract meaningful patterns and drive business decisions effectively.

One main area for improvement next time would be to enhance the data preparation process by incorporating additional features such as time-based trends and customer demographics. This would allow for more nuanced and accurate recommendations, improving the insights and strategic value derived from the analysis.

References

Kim, BM, Li, Q, Park, CS, Kim, SG & Kim, JY 2006, 'A new approach for combining content-based and collaborative filters', *Journal of Intelligent Information Systems*, vol. 27, no. 1, pp. 79–91.

Kotu, V & Deshpande, B 2019, *Data science : concepts and practice*, Second edition., Morgan Kaufmann Publishers, an imprint of Elsevier, Cambridge, MA.

Luna, JM, Padillo, F, Pechenizkiy, M & Ventura, S 2018, 'Apriori Versions Based on MapReduce for Mining Frequent Patterns on Big Data', *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 2851–2865.

Appendices

Contribution Appendix

Task	Sharlene Rodrigues	Tanveer Akram	Yurou Qiu
Code Integration	✅	✅	
Executive Summary			✅
Introduction			✅
Exploratory Analysis			✅
Frequent Pattern Mining			✅
Collaborative Filtering	✅		
Recommendation Method from Frequent Patterns	✅		
Discussion of Results		✅	
Conclusion and Recommendations		✅	