Super Sales @ The Supermarket

Add to Cart: A Market Basket Analysis

Khurush Khushrov Bengali
MComp – CS Specialization
National University of Singapore
Singapore
e1101709@u.nus.edu

Cecilia Soh
MSc Data Science and Machine Learning
National University of Singapore
Singapore
e0949465@u.nus.edu

Jing Yi Ng
MSc Business Analytics
National University of Singapore
Singapore
ng.jingyi@u.nus.edu

Poornima Sridhara
MComp - CS Specialization
National University of Singapore
Singapore
e1124722@u.nus.edu

ABSTRACT

Grocery stores like Instacart accumulate vast amounts of transactional data, offering valuable insights into both business operations and customer behavior. This report presents the analysis that could be performed on Instacart's data to gain valuable insights aimed at identifying opportunities for sales enhancement. The report starts off with an Exploratory Data Analysis to gain a better understanding of the data. Next, the report proposes the use of Apriori algorithm to identify frequent itemsets and associate rules capable of providing suggestions on product placement and bundling strategies. Apriori algorithm was also used to perform customer loyalty analysis to provide recommendations for loyalty programs. This report also delves into understanding customers' purchasing behavior through clustering, which also helped with the creation of recommender model. The recommender model was created to provide personalization recommendations to users, to encourage purchase and increase sales. Three methods of creating recommender models were compared to assess the quality of their suggestions.

KEYWORDS

Apriori Algorithm, Frequent Itemsets, Association Rules, K-Means Clustering, User-Based Collaborative Filtering, Item-Based Collaborative Filtering, Singular Value Decomposition

1 Introduction

Grocery stores encounter a variety of challenges that affect customer satisfaction and profitability, ranging from aisle organization to restocking methods and designing effective promotional campaigns. Addressing these challenges requires innovative and data-driven strategies. Therefore, this report conducts a Market Basket Analysis (MBA) of Instacart's supermarket MBA dataset to uncover relationships between products by examining customer purchase history. This report aims to tackle the problem: How can Instacart's MBA be utilized to enhance overall sales in supermarkets? Through detailed analysis, we aim to transform the large volume of transaction data into opportunities for sales growth.

The analysis begins with Exploratory Data Analysis to understand purchasing trends, followed by the application of the Apriori algorithm for identifying product associations and guiding product placement and bundling. Additionally, using the Apriori algorithm, the report investigates customer loyalty patterns to inform the development of loyalty programmes. A significant portion of the analysis is dedicated to segmenting customers based on their purchasing behaviours through clustering techniques. This will in turn inform the development of our recommender model, which aims to personalize customer recommendations and boost sales. The different models were evaluated on the quality of their suggestions.

2 Data Description

The Instacart MBA is a dataset taken from Kaggle Competition [1]. It is a relational dataset consisting of five tables detailing customers' orders throughout different periods. The tables are departments (21 rows), aisles (134 rows), orders (3,421,083 rows), products (49,688 rows), order products (32,434,489 rows). The dataset is a well-documented and highly usable dataset provided by Instacart for a predictive analytics competition on Kaggle. It offers insights into purchasing behaviour, revealing trends in product demand by day and time, as well as the duration between subsequent purchases by the same customer. Notably, the dataset was complete, with zero null values.

3 Exploratory Data Analysis

When carrying out the EDA, we found that Instacart users purchase fresh foods most frequently, with all the top 10 products being from the produce or dairy and eggs department, as seen in Figure 3-1. The fresh foods departments also dominated the top frequently ordered aisles, as seen in Figure 3-2.

When exploring the orders and basket sizes, we noted that all Instacart users made three or more orders, and the most a user ordered was 100. Most users made less than 50 orders, as seen in Figure 3-3. The basket size ranged from 1 to 145 items, with most users purchasing 5 items, as shown in Figure 3-4. We observed that most baskets were smaller in size, with most buying less than 30 items. The left-peak and long tail in the distribution of basket

size tells us that there is variability in the consumers' purchasing habits. Diving deeper into the different segments of customers will be interesting and provide insights into tailoring stock levels and promotional efforts.

Finally, we examined how duration varies between customers' orders and found a clear spike in orders every 7 days and 30 days, as observed in Figure 3-4. This indicates that there are many users who shop at Instacart at regular intervals, likely restocking their pantry on a weekly or monthly basis. This is aligned with our understanding that the most frequently bought items, fresh produce and household items, will need to be replenished regularly. Identifying such a pattern can be useful in informing our recommender system.

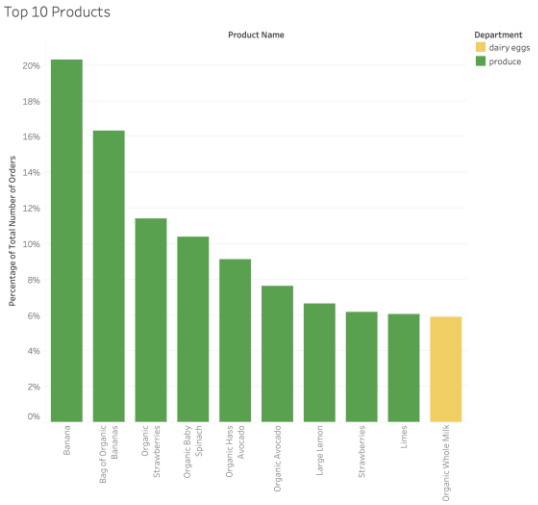

Figure 3-1: Top 10 Products Purchased

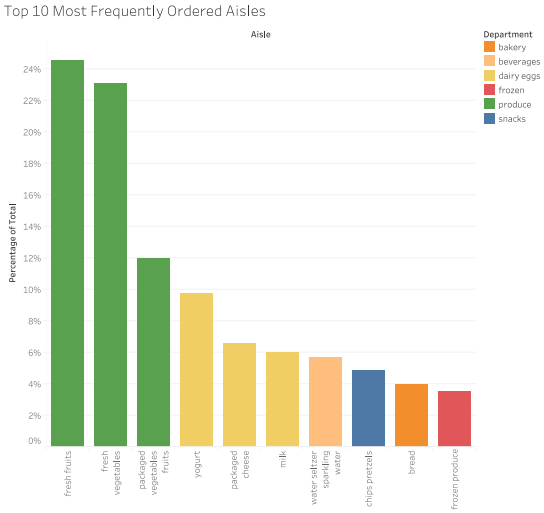

Figure 3-2: Top 10 Most Frequently Ordered Aisles

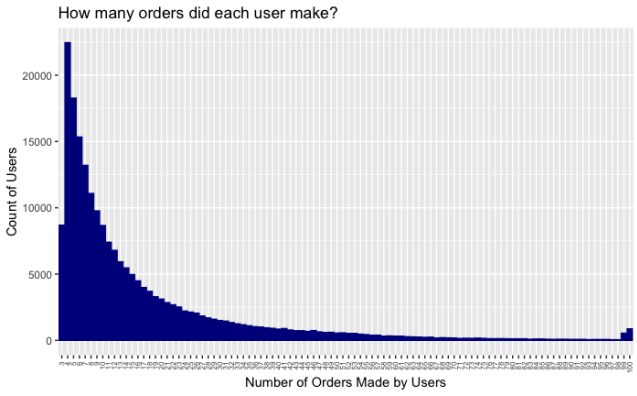

Figure 3-3: Number of Orders Made by Users

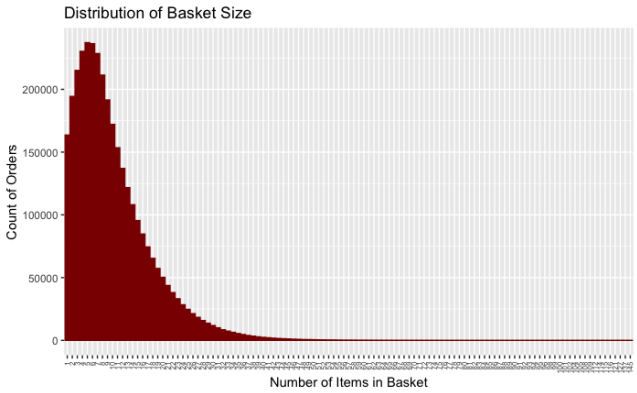

Figure 3-4: Distribution of Basket Size

4 Frequent Item Analysis

Apriori is a classic algorithm for mining frequent itemsets in transactional databases. It works by iteratively identifying frequent itemsets based on a minimum support threshold and exploiting the downward closure property. We used the Apriori algorithm to analyze transactional data per department, extracting frequent itemsets that meet a minimum support threshold, facilitating insights into product associations within each department.

4.1 Sampling Data Analysis

Given the original dataset size of 32 million, we opted to analyze a 10% sample to reduce computational complexity while still capturing significant patterns for market basket analysis. To ensure that we can preserve all the items in the selected baskets, sampling was done on the *orders* table. This downscaled dataset enables efficient processing without compromising the extraction of valuable insights regarding item associations. Upon analysis, it was discovered that the sampled dataset exhibited comparability with the actual dataset. This observation is visually depicted in Figure 4-1.

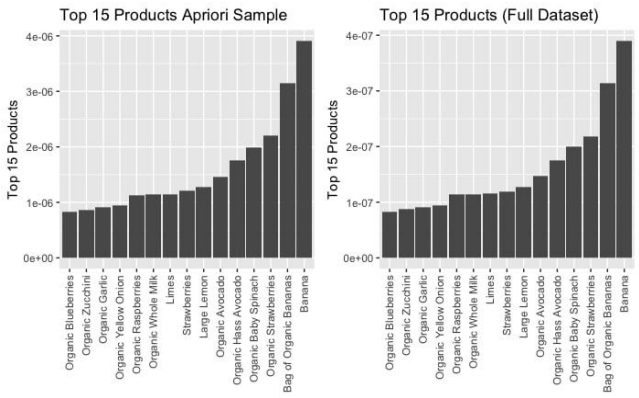

Figure 4-1 Top Products in Sample vs Full Dataset

4.2 Frequent Itemsets

Here is a table of top frequent itemsets

Support	Itemsets
0.147071	(banana)
0.118188	(Bag of Organic Bananas)
0.082568	(Organic Strawberries)
0.074689	(Organic Baby Spinach)
0.065982	(Organic Hass Avocado)

Table 4-2: Frequent Itemsets

The analysis reveals that bananas, followed by bagged organic bananas and organic strawberries, are the most frequently purchased items, indicating their popularity among customers. Notably, the inclusion of organic produce items like bagged organic bananas, organic strawberries, organic baby spinach, and organic hass avocados in the frequent itemsets suggests a strong preference for organic options among shoppers. This insight underscores an opportunity for retailers to capitalize on this demand by expanding their organic product offerings and strategically placing these items to cater to customer preferences, potentially boosting sales and enhancing customer satisfaction.

4.3 Association Rules

The association rules were mined based on confidence and lift.

4.3.1 Confidence

We first used confidence to measure the strength of the association rules. Confidence denotes the probability of the consequent item(s) being bought given the presence of the antecedent item(s) in a transaction.

$$Confidence\ (A \rightarrow B) = \frac{Support\ (A \cup B)}{Support\ (A)}$$

Below are the associations based on the Aprori algorithm using confidence.

Antecedant	Consequent	Confidence
(Org Fuji Apple)	(Banana)	0.383061
(Honeycrisp Apple)	(Banana)	0.359276
(Cucumber kirby)	(Banana)	0.325709
(Organic Avocado)	(Banana)	0.297466
(Organic Raspberries)	(Bag of organic Bananas)	0.296322

Table 4-3-1: Confidence metrics for associations

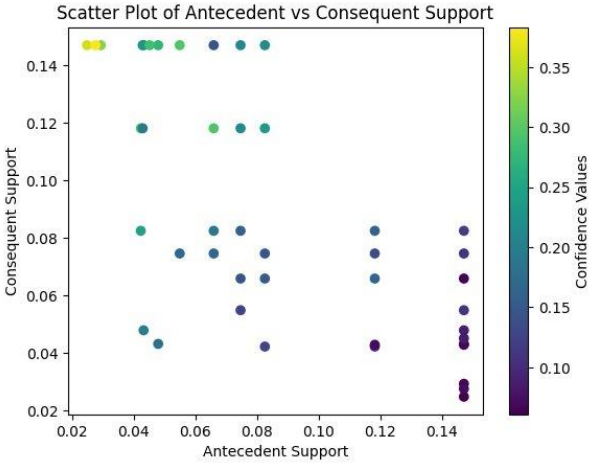

Figure 4-3-1 Supports with Confidence as a metric

Based on the above results, we recommend some useful strategies that can be adopted.

1. Highlight Organic Produce Section: Given the significant associations between organic produce items, creating a dedicated section in the store for organic fruits and vegetables would be useful for better sales. High-confidence association items like Organic Fuji Apple, Honeycrisp Apple, Organic Hass Avocado, and Organic Avocado can be placed in close proximity to banana.

2. Create Fruit Bundles: Since bananas are frequently associated with other fruits, creating bundled displays featuring bananas alongside associated fruits like Organic Strawberries, Strawberries, Large Lemon, and Honeycrisp Apple can also aid in good sales. This can encourage customers to purchase multiple fruits at once.

3. Enhance Checkout Experience: Placing small displays of high-confidence association items near the checkout counter can encourage impulse purchases. For example, Organic Fuji Apple and Honeycrisp Apple can be displayed near the checkout to entice customers to add them to their basket before leaving.

4.3.2 Lift

We also used lift to measure the association rules. Lift displays the ratio of confidence for the rule to the prior probability of having the rule prediction.

$$Lift\ (A \rightarrow B) = \frac{Confidence(A \cup B)}{Support\ (B)}$$

A lift of 1 means that the rule is no better than chance, a lift greater than 1 means that the rule is positively correlated, and a lift less than 1 means that the rule is negatively correlated.

Below are the top associations using lift. Refer to accompanying notebook for the full list of associations.

Antecedant	Consequent	Lift
(Large lemon)	(Limes)	4.17
(Limes)	(Large lemon)	4.17
(Organic strawberries)	(Organic raspberries)	3.01
(Organic raspberry)	(Organic strawberries)	3.01

Table 4-3-2: Lift metrics for associations

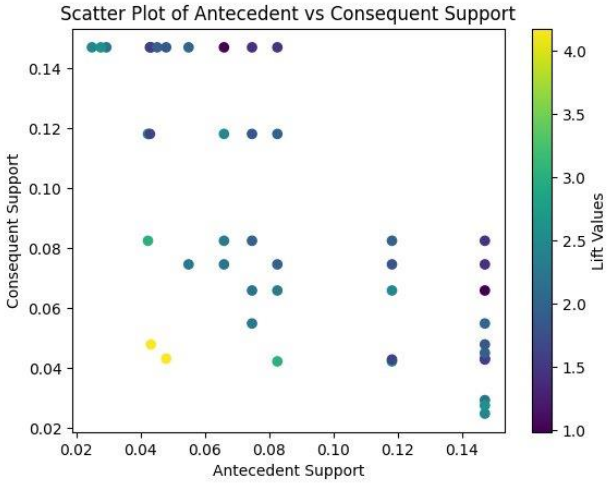

Figure 4-3-2 Supports with Lift as a metric

The association rules analysis reveals compelling insights into item relationships within the dataset.

Notably, a strong association exists between 'Large Lemon' and 'Limes', indicating that customers who purchase one are highly likely to purchase the other. Similarly, 'Organic Strawberries' and 'Organic Raspberries' exhibit a significant association, presenting opportunities for bundled promotions or cross-selling initiatives. Moreover, the moderate association between 'Organic Fuji Apple' and 'Banana' suggests potential for strategic positioning to capitalize on customer preferences. Additionally, 'Bag of Organic Bananas' demonstrates an association with 'Organic Raspberries', indicating opportunities for promoting organic fruit combinations. Below are some examples of organic fruit combinations/bundles based on the above results.

1. Organic Berry Bundle: A bundled promotion featuring 'Organic Strawberries' and 'Organic Raspberries' can be offered. Customers purchasing one pack of organic strawberries can receive a discount on a pack of organic raspberries or vice versa, capitalizing on the strong association between these items.

2. Citrus Delight Combo: Creating a bundled promotion combining 'Large Lemon' and 'Limes' would be useful. Customers purchasing a bag of large lemons can receive a complementary bag of limes or vice versa. This promotion leverages the high lift between these citrus fruits to encourage sales.

3. Organic Fruit Medley: Promoting a diverse assortment of organic fruits, including 'Bag of Organic Bananas' and 'Organic Raspberries' can be done. A special price for purchasing both items together can be offered, emphasizing the health benefits and quality of organic produce.

4.4 Customer Loyalty

Another possible approach is by grouping items based on product IDs and selecting orders which have a specified support value. Grouping orders by product IDs to identify patterns of co-purchased products within transactions aids in identifying loyal customers with high support values. We can then identify the top three customers and reward them. This approach facilitates targeted marketing strategies, personalized recommendations, and customer retention efforts tailored to the preference and behaviors of loyal customers within the analyzed department's sales data.

We query the users table to get the users corresponding to these order ids.

Support	Itemsets (order_id)	user_id
0.002886	(61355)	22906
0.002454	(2136777)	60694
0.002227	(3279252)	201268
0.002091	(2467301)	201268

We can provide different rewards for these loyal customers, such as loyalty points redeemable for free products, and exclusive discounts for future purchases. Such rewards will incentivize customers to spend more at the store and encourage the formation of shopping habits.

5 Customer Segmentation

To increase sales, it is imperative to understand the customers' purchasing behavior and introduce promotional tactics that caters to these customers. This involves segmenting customers into clusters and analyzing their characteristics.
We utilized purchase proportions across departments, time of day and day of week for clustering. Days were categorized into weekday and weekends while time was divided into morning (5am to 12pm), afternoon (12pm to 5pm), evening (5pm to 9pm) and night (9pm to 5am). Data preparation was done using map-reduce methodology for efficient processing of large datasets.

K-means clustering was used for clustering and the elbow method was used to determine optimal k value (number of clusters). This involves plotting the within cluster sum of square obtained from K-means for different values of k. The within cluster sum of square shows the amount of variance that can be explained by the clusters [2]. Figure 5-1 shows the resulting elbow where the within cluster sum of square decreases rapidly up to k=10, then starts to slow down as k increases to 20. Beyond k=20, the within

cluster sum of square starts to stabilize. Hence, k=20 was chosen for K-means.

The formed clusters reveal interesting patterns which can hint that clustering was performed well. For example, Cluster 14 stands out for its higher alcohol purchases compared to other clusters (Figure 5-2). Moreover, this cluster also purchases more from babies department (Figure 5-3), as well as products related to pets, household, and personal care (Figure 5-4 to 5-6). Customers in this cluster likely prioritize convenience, possibly due to family responsibilities, hence opting for home consumption of alcohol instead of consumption at bars. Targeted promotion such as running sales on baby products can potentially incentivize purchases from both departments. Additionally, raising the price of alcohol can be considered during the same period as further boost sales as customers may prefer the convenience of Instacart over visiting another store for alcohol.

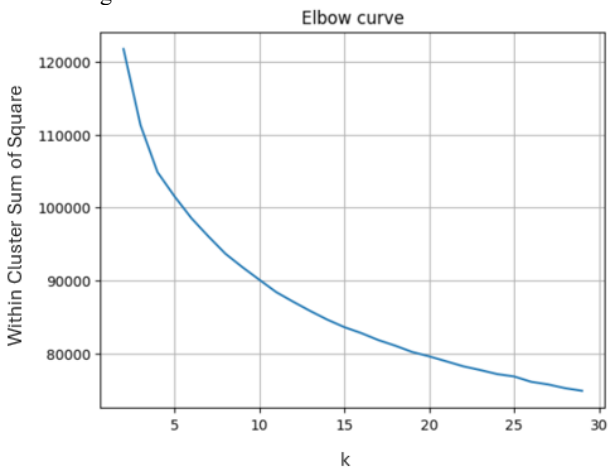

Figure 55-1: Plot of within cluster sum of square against k to get the elbow curve.

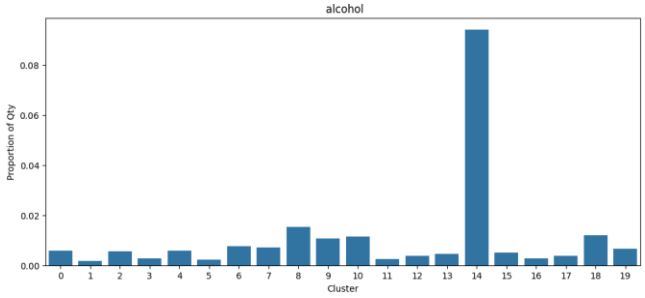

Figure 55-2: Proportion of purchase from Alcohol department for each cluster.

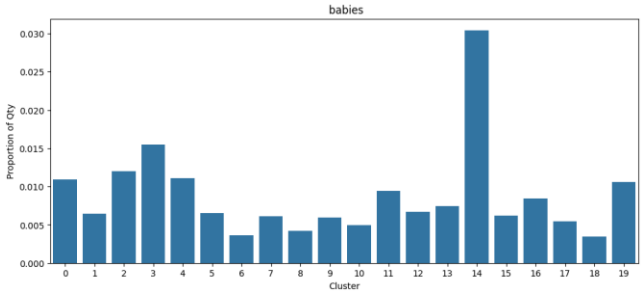

Figure 55-3: Proportion of purchase from Babies department for each cluster.

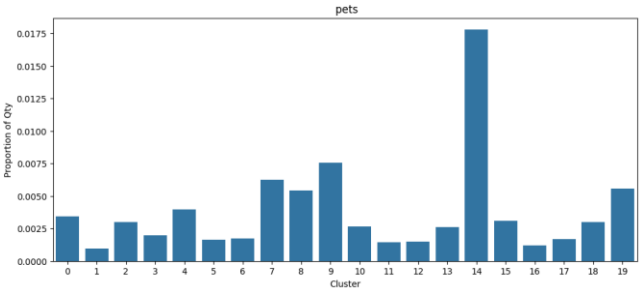

Figure 55-4: Proportion of purchase from Pets department for each cluster.

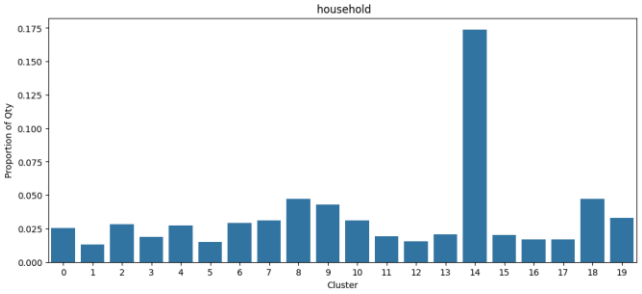

Figure 55-5: Proportion of purchase from Household department for each cluster.

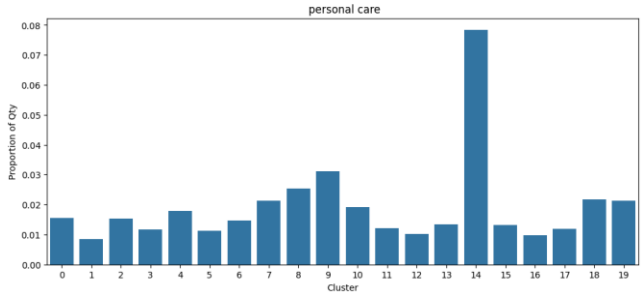

Figure 55-6: Proportion of purchase from Personal Care department for each cluster.

6 Recommender System

Another strategy to enhance sales involves providing personalized recommendations to customers based on their interests as suggested from their historical purchases. These suggestions can help customers to discover relevant products which they may not have previously considered, thereby expanding their options, and encouraging exploration. Given the scale of our transaction data, a model capable of handling huge amounts of data to provide suggestions is of great value.

The first method explored is user-based collaborative filtering which recommends products based on what similar customers are buying. However, with 206,209 customers, calculating pairwise user similarity becomes computationally challenging and memory intensive. Hence, we leveraged our clustering results to narrow the search for similar users within each cluster, reducing the amount of calculation and storage required.

The second method employed is item-based collaborative filtering, recommending items similar to what the customer has frequently purchased. Since there are smaller number of products offered by Instacart, computing item pairwise similarity becomes more manageable. Hence, item-based collaborative filtering can be implemented without the need for adjustments to accommodate our dataset size.

The third method adopted is the Singular Value Decomposition (SVD) involving factorization of the utility matrix to map items and users into vectors representing their attributes and interests respectively for a common set of latent factors [3]. This mapping is obtained by minimizing the sum of squared errors with the known entries in the training utility matrix. The missing entries of the utility matrix are then estimated using a dot product between the user's mapped vector and the item's vector.

To assess the performance of the 3 methods, 20% of the known entries in the utility matrix were randomly selected for use as test data. Root Mean Squared Error (RMSE) between the predicted and true values of the test data is tabulated. Table 6-1 shows the RMSE value for the 3 methods, and it was observed that item-based collaborative filtering outperformed the others.

Method	RMSE
User-Based Collaborative Filtering	0.0281
Item-Based Collaborative Filtering	0.0267
SVD	0.0270

Table 6-1: RMSE of the three recommender models.

We have also examined the top 10 most recommended items by each method for all users to evaluate their output quality. Figure 6-1 shows the top 10 recommendations from the user-based collaborative filtering. This method tends to recommend items that are popular among all customers as these recommendations align with the most frequently purchased products and notably from the Produce department. This indicates a lower level of personalization to customers' interests, hinting at decreased quality of suggestions, resulting in the highest RMSE among the three methods.

The recommendations generated by SVD are centered around spring water and organic bananas as shown in Figure 6-2. Each of these 2 items were recommended to more than 60,000 users. There is a lack of diversity in the SVD's recommendation possibly because the model was optimized by minimizing RMSE. Hence, the model might have prioritized fewer items that are popular such as Organic Bananas to maintain high RMSE during training. This lack of diversity implies a lower quality of recommendation provided by SVD.

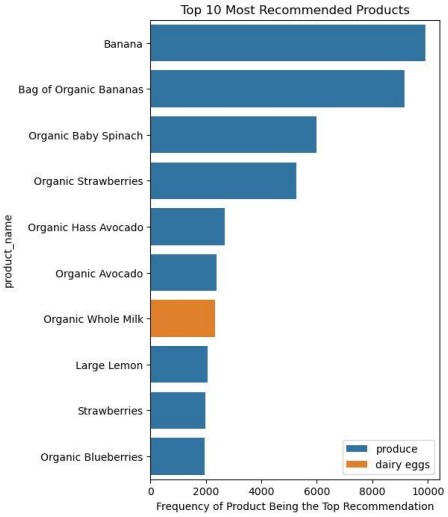

Figure 66-1: Top 10 most recommended items generated using user-based collaborative filtering.

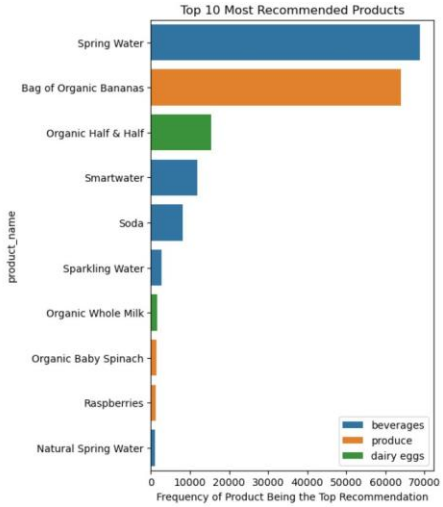

Figure 66-2: Top 10 most recommended items generated using SVD.

There is greater variety in the recommendations provided by the item-based collaborative filtering. While it may appear that recommendations are centered around Organic Fuji Apples, it was only recommended to 7,000 users, suggesting that the recommendations were spread out across other items. The top 10 recommendations span across 4 different departments further signaling the presence of diversity and customization. In addition, these recommendations align with our findings from association rules mining. We observed that Organic Fuji Apples and Honeycrisp Apples are frequently purchased alongside bananas. Considering bananas are the most purchased product, it might have resulted in Organic Fuji Apple and Honeycrisp Apple emerging as top recommendations from item-based collaborative filtering. Since it is apparent that item-based collaborative filtering performs better in product diversity, customization and RMSE, we can utilize this method to provide personalized recommendations to customers to encourage purchase and increase sales.

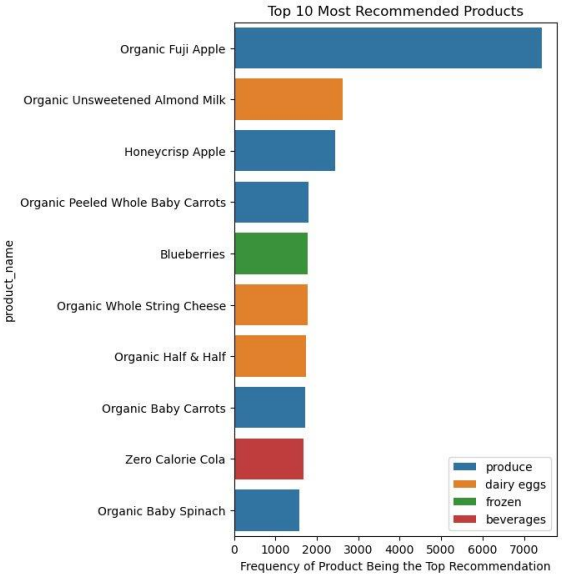

Figure 66-3: Top 10 most recommended items generated using item-based collaborative filtering.

APPENDIX

Project Source Code: GitHub

REFERENCES

[1] Instacart. 2017. Instacart Market Basket Analysis. Kaggle (2017).

[2] Purnima Bholowalia and Arvind Kumar. 2014. EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications 105, 9 (2014).

[3] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. Computer 42, 8 (2009), 30–37.