Analyzing Grocery Data with Association Rule Mining for Retail Optimization

Ignatia Christabelle Amadea Hardjono
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
ignatia.hardjono@binus.ac.id

Alexandra Hadiprawira
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
alexandra.hadiprawira@binus.ac.id

Raden Bagus Juan Marcelle Angelo
Dragon Devondha L.
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
raden.devondha@binus.ac.id

Henry Lucky
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
henry.lucky@binus.ac.id

***Abstract*— Retail strategies revolve around the utilization of consumer purchasing behavior, with the grocery sector being no exception. Optimization of these strategies are done with the understanding of said patterns as the foundational cornerstone, where data mining plays a pivotal role in uncovering them. In the case of groceries, commonly purchased product combination and item relationships are exceptionally important, ranging from the optimizing stock placement up to the overall customer satisfaction. In this project report, the research that was carried out applied Association Rule Mining, using namely techniques like Apriori Algorithm, Frequent Pattern Growth (FP-Growth) Algorithm, and Equivalence Class Clustering and Bottom-Up Lattice Traversal (ECLAT) to identify frequent item sets and their relationships in grocery datasets. These insights were then evaluated using support, confidence, and lift as metrics to ensure that the generated rules were valuable and weighted. These approaches demonstrated the emphasis of data mining for strategic decision making in the grocery retail district.**

Keywords—consumer purchasing behavior, association rule mining, Apriori Algorithm, Frequent Pattern Growth Algorithm, Equivalence Class Clustering and Bottom-Up Lattice Traversal, frequent item sets, support, confidence, lift

I. Introduction

The retail industry is dependent on the study of continuous historic purchasing behavior of consumers overtime to develop effective strategies [1]. This is done to optimize the overall workflow and performance, ultimately enhancing both growth and profitability of the business. Among many fields within the retail scope, the grocery sector is deemed significant due to the constant high transaction volumes and number of purchases [2]. With excellent awareness of consumer buying patterns, matters like stock management, decisive targeted promotions, and profitable product bundling will be elevated, driving both parties of the transaction exchange into advantage.

One of the most common encountered issue would be when it comes to the analysis of large-scaled transactional data [3]. The vast numbers of transaction is most likely complementary with the existence of diverse item combinations, which increases the difficulty in the identification of meaningful patterns without the help of the suitable advanced analytical methods. This is where data mining, specifically Association Rule Mining, performs a crucial role [4]. By disclosing the influential relationships between the frequently purchased items that was found, retailers will be able to resolve to data-driven decisions and consumers need will be fulfilled.

In this conducted study, we applied three widely adopted algorithms of Association Rule Mining—Apriori Algorithm, FP-Growth Algorithm, and ECLAT; due to their reputation of being proficient for their identification of frequent item sets and generation of association rules [5]. Support, lift, and confidence were also used as the evaluation metrics to ensure significance of output, placing significant emphasis on the quality of generated rules. The overall aim of this research is to show the improvements that can be done to retail strategies, primarily in grocery stores, by implementing data mining methods.

II. Theoretical Basis

A. Association Rule Mining

Association rule mining is a fundamental data mining method used to show patterns, associations, or connections among items inside huge datasets. This strategy is especially useful in transactional data, such as market basket analysis, where it recognizes associations between purchased items. The method includes the extraction of association rules, represented as A -> B, which recommends that the presence of itemset A suggests the presence of itemset B. Key metrics used in association rule mining includes Support which is the extent of transactions in the dataset that contain the itemset, Confidence which is the conditional likelihood that a transaction containing A also contains B, and Lift which is the ratio of observed support to expected support if A and B were independent [6].

B. Apriori Algorithm

The Apriori algorithm is an approach to find repeating itemsets in association rule mining. The algorithm starts by creating repeated itemsets of size 1 (single items). These repeated itemsets are then used to create bigger candidate itemsets, like sets or triples. Candidate itemsets that fail to meet the minimum support threshold are pruned, and the method is repeated until no repeated itemsets can be produced. Although efficient and straightforward, the Apriori algorithm can be computationally expensive, particularly for expansive and large datasets, due to the need to check the database repeatedly [6].

C. FP-Growth Algorithm

The FP-Growth algorithm is an alternative to Apriori. It avoids the candidate generation process by developing a compact data structure known as the FP-tree. The algorithm starts by developing an FP-tree by handling the dataset once and putting away item frequencies. It then uses the tree to extract repeated patterns through recursive pattern growth [6]. The FP-Growth algorithm is useful for large datasets because it decreases the number of databases scans and compresses the data into a compact representation.

D. Eclat Algorithm

The Eclat algorithm uses a vertical database format to mine repeating itemsets. Instead of showing data as transactions, it stores items together with their associated transaction IDs [6]. The dataset is shown in a vertical format, mapping each item to its other TIDs. Search method is used to find repeated itemsets by converging TID sets, and itemsets that don't meet the minimum support threshold, then it will be pruned. Eclat is efficient for smaller datasets or vertical formats are suitable.

E. Rstudio

The RStudio is an integrated development environment (IDE) for R, an effective programming language for statistic computing and data analysis. In association rule mining, RStudio provides tools and libraries, such as arules and arulesViz for executing algorithms like Apriori, FP-Growth, and Eclat. Features of RStudio in association rule mining include its use to assist data preprocessing, and visualization, execute repeated itemset mining and rule generation, and visualize patterns using graphs and scatter plots. RStudio improves the application of these algorithms by giving a user-friendly interface, and tools for exploratory data analysis [7].

III. Methods

A. Dataset

In this report, we utilized two datasets that were sourced from Kaggle:

1) First Dataset: This dataset provides documented grocery transactions over a period of time. There are a total of 522,064 rows of data with 7 columns, namely the bill number, name of the item, quantity of items purchased, date of transaction, price of transaction, customer ID, and the country where the transaction was done.

2) Second Dataset: This dataset also contains sales records from a grocery store, similar to the first. It differs in the total number of data as well as its attributes, consisting of 43,745 rows and 3 columns, including the date of transaction, customer ID, and description of the product.

Both datasets were selected because of their relevance and completeness to undergo exploratory data analysis, data preparation, and the application of the selected association rule mining algorithm. Table 1 below shows the dataset seen through the .csv file, alongside each of the parameters' data type.

TABLE I. Dataset Description

Parameters	***Location***	***Data Type***
BillNo	Dataset 1	nominal
Itemname	Dataset 1	numerical
Quantity	Dataset 1	numerical
Date	Dataset 1	numerical
Price	Dataset 1	numerical
CustomerID	Dataset 1	nominal
Country	Dataset 1	nominal
Transaction Date	Dataset 2	nominal
Customer ID	Dataset 2	numerical
Product Description	Dataset 2	nominal

B. Exploratory Data Analysis (EDA).

In this section, we perform an in-depth exploratory data analysis (EDA) to determine the structure and pattern of each dataset. The EDA section includes processes such as finding unique values, visualizing the most frequent items, visualizing unique transactions, and many others. This method is required to gather insights that will guide the KDD phases. The detailed EDA for each dataset is provided below.

1) First Dataset

a) Identifying Outliers in Purchase Quantities

We carried out this to analyze the data utilizing its 'Quantity' parameter as visualization. Originally, the original dataset had a vast amount of data of 522,064 rows of transaction, which was prone to the presence of outliers. The aim of this EDA was to capture data points that fall outside the normal range of variability that could potentially lead to existing patterns for better association results. The visualization was plotted as a boxplot, as seen in Fig 1. below. The thresholds of the boxplot were determined using the Interquartile Range (IQR), using the following equation:

$$Lower\ threshold = Q1 - (1.5 * IQR) \quad (1)$$

$$Upper\ threshold = Q3 + (1.5 * IQR) \quad (2)$$

The value of IQR itself is the difference between the first and the third quartile, and we decided to use 1.5 as the multiplier as it is capable in reflecting the typical variability within the dataset, meaning points that exist outside its range will most likely be the outliers we're attempting to detect. Using this method of choice, the box produced was very compressed nearing the lower end at 0, and this is due to extreme noise above the 60,000-benchmark in quantity.

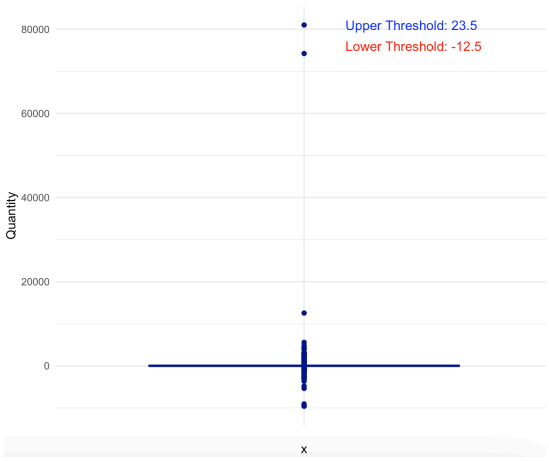

Fig. 1. Boxplot of Purchase Quantities Highlighting Outliers

b) Categorizing Transactions by Purchase Volume

The identified outliers were kept instead of removed. After acknowledging that purchase quantities can be as high as 80,000, we deduced that the sales records of this grocery dataset consist of two kinds of purchases: normal purchases (day-to-day groceries) and bulk purchases (for larger need and consumption purposes). Normal purchases are items with quantities below and equal to the upper threshold (23.5), and bulk purchases are items with quantities above the said upper threshold. Using this as baseline understanding, this next EDA step is to visualize and recognize the number of transactions that fall under the two categories. The count of both normal and bulk transactions is shown in Fig 2. below. The total number of bulk purchases was 54,399, and the number of normal purchases was 467,665.

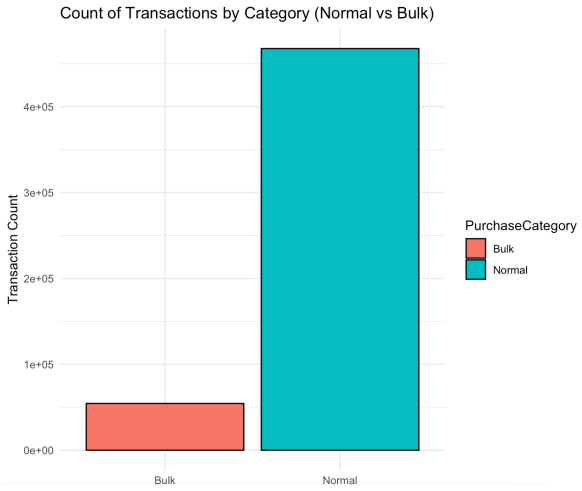

Fig. 2. Distribution of Transactions Based on Purchase Volumes

2) Second Dataset

a) Identifying the Most Purchased Items

We identified the most purchased items to highlight item purchasing trends. The analysis shows that the top most purchased items were the same in both years, with "other vegetables" and "rolls/buns" being the most frequently purchased items indicating that both items are in high demand. Other than that, "soda", "yogurt", "root vegetables", and "sausage" also emerge as popular items in 2014 and 2015. This was further illustrated in Fig 3. which presented the visualization of the most purchased items overall. Additionally, Fig 4. compared the top 5 most purchased items in 2014 and 2015, which provides an overview of the items purchased by year and giving deeper insights into the variation of customer preferences over the year.

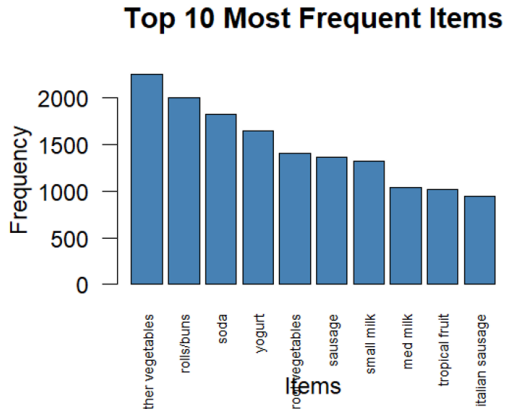

Fig. 3. Top 10 Items Sold Across All Transactions

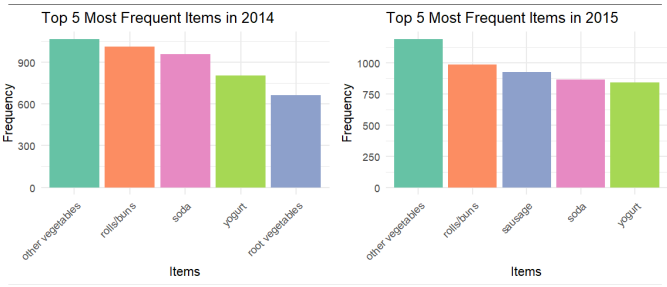

Fig. 4. Top 5 Items Sold in 2014 vs 2015

b) Categorizing Transactions by Year

To analyze the overall sales over the two years, we visualized the number of transactions for each year. This visualization provided insights into the sales performance for different years. Based on the visualization of the unique transaction for each year in Fig 5. below, it was concluded that 2014 had more transactions, with a total of 7,981 transactions compared to 6,982 transactions in 2015. This indicated that customer activity or sales volume was higher in 2014.

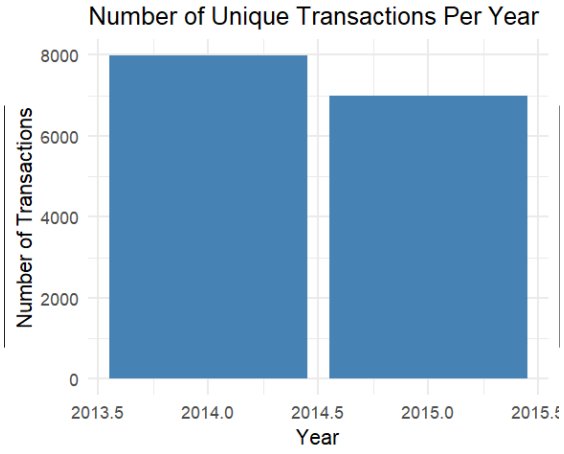

Fig. 5. Distribution of Transactions by Year (2014 vs 2015)

C. KDD (Data Preparation)

Before we implement the association rules algorithms, we need to do data preparation to ensure the dataset is cleaned, structured, and ready for further analysis. This data preparation stage involves several preprocessing steps, such as handling missing data, removing outliers, segmenting the data based on insights from the exploratory data analysis (EDA), and many more. The steps below outline the specific preprocessing tasks performed for each dataset:

1) First Dataset

a) Removing Rows with 0 and Negative Values

Through visualizing the quantities within the dataset, there were negative values identified. Purchasing an item below 1 lack logical consistency, therefore should be filtered out as it could potentially compromise the mining results. Thus, rows that contains quantity values below 1 were removed.

b) Removing Rows with Missing Itemname Values

Transactions that were missing their *Itemname* values could no longer be used for data mining in this conducted

research as pattern recognition will rely on the names of these purchased products. That being said, for rows that have no item name were filtered out and removed.

c) Identifying and Filtering Outlier Items

This accounts for the name formatting of the *Itemname* column, seeking to identify outlier items within the dataset. The uniform standard was using all capital letters (for example, "JUMBO BAG PINK POLKADOT"). Applying this knowledge, we listed out items that do not follow the said naming convention, resulting to a total of 78 items. After further scrutiny, there were items that made it into the list that did not exhibit outlier behavior. They were flagged due irregularities within the naming, but they were still legitimate items that could contribute to the association analysis. As a result, a total of 27 supposed 'outlier items' were excluded, leaving the rows that have the other 51as their item names filtered out of the employed dataset.

d) Checking for Remaining NULL Values

This is done to make sure all rows with missing values within the dataset were handled before moving forward to the data mining. Fortunately, after removing NULL values of the *Itemname* column done previously there were no missing values left to be taken care of.

e) Counting the Total Number of Unique Transactions

A transaction is denoted by its bill number and since each row only showcases one item each, there were many *BillNo* duplicates to be merged into lists of transactions. Since we've decided to carry on the analysis using the two purchasing categories we observed earlier in the EDA, the number of unique bill numbers will conclude and reaffirm how many of them will fall under each type. This will also serve as a way to double check whether all transactions have been categorized accordingly. The cleaned dataset had a total of 19481 unique transactions.

f) Detecting Mixed Purchase Category

Within the 19481 unique transactions, we hypothesized the possibility of a transaction containing both bulk and normal orders of different items in a single purchase. Testing this theory, we grouped both items from the bulk and normal category with the same bill numbers. Results obtained validated and reaffirmed our hypothesis, leaving us with a total of 10961 unique transactions with both normal and bulk orders. Considering the findings we had obtained, the dataset was now set to be categorized into three purchasing categories instead of two, namely: bulk, normal, and mixed orders.

g) Splitting the Dataset by the Purchasing Categories

The dataset was split into three according to their purchasing categories. This is done by filtering which goes into which, then creating a separate .csv file for each. The three newly generated datasets will be the ones utilized for association rule mining.

h) Dropping Uneccessary Columns

Before the datasets undergo data mining, they need to be reformatted in a way to obtain optimal results. In the case of association rule mining of this dataset, there will only be two columns needed: *BillNo* and *Itemname*. All the other columns within the three datasets will be dropped in preparation for data transformation.

i) Data Transformation

Data transformation is needed to guarantee smooth application of the selected algorithms by adjusting to their desired format. In this study, the algorithms that were chosen were the Apriori algorithm, FP-Growth, and Eclat, that required the dataset to be transformed into two transactional configurations: basket (for Apriori algorithm and Eclat) and list (for FP-Growth). The list-based format was achieved by using the split() function included in the fim4r package, grouping the items together by their attribute *BillNo*. On the other hand, the basket-based format combined all items using the *BillNo* as its initial transaction identifier before being omitted, then the arules package recognized the newly combined rows of items as unique transactions.

2) Second Dataset

a) Checking for Missing Data and Duplicates

The dataset was reviewed for missing or duplicated data. However, this dataset contains no missing or duplicated data, demonstrating the dataset's integrity.

b) Data Cleaning

The rows with *Product Description* containing the word “delete” were removed because those rows represented irrelevant data that could have negatively affected the overall analysis. After this removal, the dataset was reduced to 43,635 rows from the initial 43,745 rows.

c) Date Data Type Conversion

The *Transaction Date* column initially had a character data type. For purposes of temporal analysis, such as classifying the transactions by year, it is transformed into a date data type.

d) Sorting

The data was sorted based on the *Transaction Date* and *Customer ID* to ensure a logical order for further processing.

e) Invoice Number Generation

An *Invoice Number* was generated based on the *Transaction Date* and *Customer ID*. The *Invoice Number* was generated to help grouping the transactions that will be a necessary step for converting the data to a basket format before implementing the association rules algorithms.

f) Filtering Data by Year

Based on the EDA steps, the dataset, which had a total of 43,635 rows, was divided into two subsets: transactions from 2014 with 21,107 rows and 2015 with 22,528 rows. The dataset was separated into these two subsets to enable a comparative analysis of trends and patterns over the years.

g) Data Transformation

The data transformation for Dataset 2 was the same as Dataset 1. However, for this dataset, only basket format was utilized for all three algorithms. All items were grouped based on the *Invoice Number* as its initial transaction identifier before being removed. The basket format was used for Dataset 2 because this format eliminates the need for additional transformations which ensures a more streamlined analysis process.

D. KDD (Data Mining)

In this section, the three sub datasets of the first dataset and the second dataset undergo association rule mining employing the three different algorithm that was mentioned above: Apriori algorithm, FP-Growth, as well as Eclat, to identify patterns and relationships that were present between them.

This stage is crucial to the KDD process and is used to extract significant patterns based on previously defined support and confidence. Thresholds for the metrics differ based on their difference in analytical focus and dataset size.

1) Support

In Dataset 1, the support was set at 0.02 across both normal and mixed sub datasets, and 0.01 for the bulk sub dataset. This is due to the substantial difference of the size of the bulk dataset in comparison with the other two. In Dataset 2, the support was set at 0.002 for both 2014 sub datasets and 2015 sub datasets. We set the same support for both sub datasets since there is no significant size difference between the two sub datasets.

For a larger datasets like Dataset 1, assigning higher support filters out noise and focuses on the common and robust patterns. For a smaller dataset like Dataset 2, assigning lower support makes sure that we will be able to capture valuable niche patterns without being overly restrictive.

2) Confidence

The confidence threshold for Dataset 1 was 0.3, while Dataset 2 was set higher at 0.5. This difference reflects the adjustments done utilizing the relationship between support and confidence to balance the reliability of the rules derived from each dataset. For Dataset 1, we used a lower confidence threshold of 0.3 because the higher support (0.02) already ensures that the rules derived were from frequently occurring patterns. In contrast, Dataset 2 had a much lower support threshold of 0.002 which comes a higher risk of the number of obtained rules not being practically significant. To counter this risk, we set a higher confidence threshold of 0.5 to ensure that the derivation of rules will only be the ones with strong likelihood of occurring with one another. This provides an additional layer of reliability, making up for the low support.

IV. Result

This section presents the results of the rules generated from applying the association rule mining algorithms to the two datasets. The findings are presented into two parts which are Pattern Evaluation and Knowledge Presentation, following the KDD process. The coding that was done can be accessed through this GitHub link: https://github.com/belle000/Data-Mining-Final-Project-Submission.git

A. KDD (Pattern Evaluation)

1) First Dataset

a) Apriori Algorithm

Results that were obtained from employing this algorithm to the three sub datasets concluded a total of 91 rules generated for the Mixed Purchases category, 68 rules for the Normal Purchases category, and 12 rules from the Bulk Purchases category. Table 2, 3, and 4 presents a sample of association rules taken from each of their top 10 sorted in descending order organized by their lift values.

TABLE II. Mixed Purchases Dataset

	LHS	***RHS***	***Support***	***Confidence***	***Lift***
1	SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER PLATES	0.0203	0.8383	28.1900
2	GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.0247	0.6984	19.9386
3	PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0247	0.8914	19.9386

TABLE III. Normal Purchases Dataset

	LHS	***RHS***	***Support***	***Confidence***	***Lift***
1	SET 3 RETROSP OT TEA	SUGAR	0.0251	1.0000	39.8457
2	COFFEE, SET 3 RETROSP OT TEA	SUGAR	0.0251	1.0000	39.8457
3	GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.0207	0.7059	23.1085

TABLE IV. Bulk Purchases Dataset

	LHS	***RHS***	***Support***	***Confidence***	***Lift***
1	JUMBO BAG PINK POLKADOT, JUMBO BAG RED RETROSP OT	JUMBO BAG STRAWBE RRY	0.0135	0.9130	30.7457
2	RED HANGING HEART T-LIGHT HOLDER	WHITE HANGING HEART T-LIGHT HOLDER	0.0135	0.8077	20.5101
3	JUMBO BAG PINK VINTAGE PAISLEY	JUMBO BAG RED RETROSP OT	0.0122	0.8636	15.3767

b) FP-Growth Algorithm

From applying this algorithm, we obtained a total of 557 rules: 258 rules for the Mixed Purchased category, 282 rules for the dataset containing only normal orders, and 17 rules for the dataset containing only bulk orders. Tables 5, 6, and 7 below listed down a selection of association rules that were generated taken from the top 10 sorted by their lifts.

TABLE V. MIXED PURCHASES DATASET

	LHS	*RHS*	*Support*	*Confidence*	*Lift*
1	POPPY'S PLAYHOUSE KITCHEN	POPPY'S PLAYHOUSE BEDROOM	0.0210	0.7428	25.9282
2	POPPY'S PLAYHOUSE BEDROOM	POPPY'S PLAYHOUSE KITCHEN	0.0210	0.7357	25.9282
3	SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER PLATES	0.0227	0.8440	25.6283

TABLE VI. NORMAL PURCHASES DATASET

	LHS	*RHS*	*Support*	*Confidence*	*Lift*
1	GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.0250	0.6905	18.4444
2	PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0250	0.8970	17.5652
3	GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.0250	0.8450	16.8260

TABLE VII. BULK PURCHASES DATASET

	LHS	*RHS*	*Support*	*Confidence*	*Lift*
1	JUMBO BAG ALPHABET	JUMBO BAG APPLES	0.0103	0.6667	36.8571
2	JUMBO BAG APPLES	JUMBO BAG ALPHABET	0.0103	0.5714	36.8571
3	JUMBO BAG PINK POLKADOT, JUMBO BAG RED RETROSPOT	JUMBO BAG STRAWBERRY	0.0136	0.8400	27.6664

c) Eclat Algorithm

Despite being a different algorithmic approach, the rules that were generated using this method showed equivalent results to when Apriori algorithm was applied. Thus, samples of the rules could also be derived from Tables 2, 3, and 4 displayed previously.

Comparing the results of the three algorithms side-by-side, we deduced that both Apriori algorithm and Eclat yielded the most optimal results of strong association rules. This conclusion was encouraged by the lift values that were slightly higher across rules generated by the two algorithms, suggesting that they might be best suited for handling datasets of similar types.

2) Second Dataset

a) Apriori Algorithm

Applying the Apriori algorithm to the two subsets yielded a total of 12 rules from the 2014 dataset and 5 rules from the 2015 dataset. The association rules for both datasets are presented in Tables 8 and 9, sorted in descending order based on lift values.

TABLE VIII. 2014 DATASET

	LHS	*RHS*	*Support*	*Confidence*	*Lift*
1	{sausage, soda}	{italian sausage}	0.0035	0.6087	17.2291
2	{italian sausage, yogurt}	{sausage}	0.0027	0.8800	16.1847
3	{sausage, small milk}	{italian sausage}	0.0021	0.5484	15.5220
4	{other vegetables, sausage}	{italian sausage}	0.0025	0.5405	15.3000
5	{rolls/buns, sausage}	{italian sausage}	0.0020	0.5333	15.0960
6	{italian sausage, soda}	{sausage}	0.0035	0.8000	14.7134
7	{italian sausage, small milk}	{sausage}	0.0021	0.7727	14.2118
8	{italian sausage, rolls/buns}	{sausage}	0.0020	0.7619	14.0127
9	{italian sausage}	{sausage}	0.0263	0.7447	13.6960
10	{italian sausage, other vegetables}	{sausage}	0.0025	0.7143	13.1370

11	{fish}	{small milk}	0.0070	0.8889	11.9650
12	{domestic eggs, root vegetables}	{other vegetables}	0.0021	0.5862	4.3935

TABLE IX. 2015 DATASET

	LHS	*RHS*	*Support*	*Confidence*	*Lift*
1	{fish}	{small milk}	0.0079	0.7237	6.9800
2	{italian sausage, pastry}	{sausage}	0.0030	0.5000	3.7624
3	{canned beer, italian sausage}	{sausage}	0.0021	0.5000	3.7624
4	{herbs, root vegetables}	{other vegetables}	0.0020	0.5385	3.1624
5	{newspapers, root vegetables}	{orher vegetables}	0.0026	0.5000	2.9365

b) FP-Growth Algorithm

The FP-Growth algorithm produced the same result as the Apriori algorithm. This method also generated 12 rules for the 2014 dataset and 5 rules for the 2015 dataset. The association rules can be found in Tables 8 and 9.

c) Eclat Algorithm

Similarly, the Eclat algorithm also generated identical results to both Apriori and FP-Growth algorithms. The association rules can also be found in Tables 8 and 9.

Comparing the results of the three algorithms, it could be concluded that all three algorithms produced identical association rules for both the 2014 and 2015 datasets. This consistency indicated that each algorithm was equally effective and optimal for this analysis.

B. KDD (Knowledge Presentation)

1) First Dataset

In the previous section, findings showed that both Apriori and Eclat were the strongest overall performers. The following figures below illustrates the rules through visualization of plots to better ease understanding of the associations between the items. Figure 6 represents the visualization of rules for the mix order dataset, Figure 7 for the normal order dataset, and Figure 8 for the bulk order dataset.

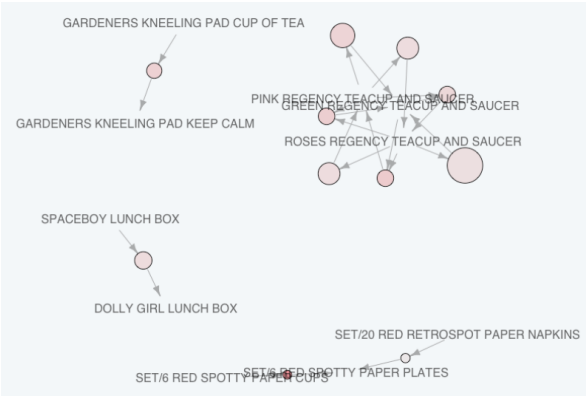

Fig. 6. Visualization of Apriori Agorithm on Mixed Order Dataset

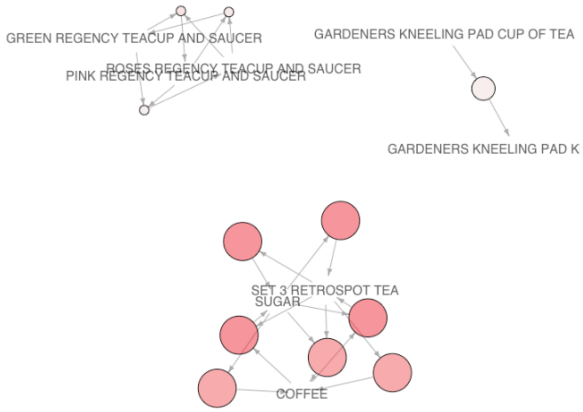

Fig. 7. Visualization of Apriori Agorithm on Normal Order Dataset

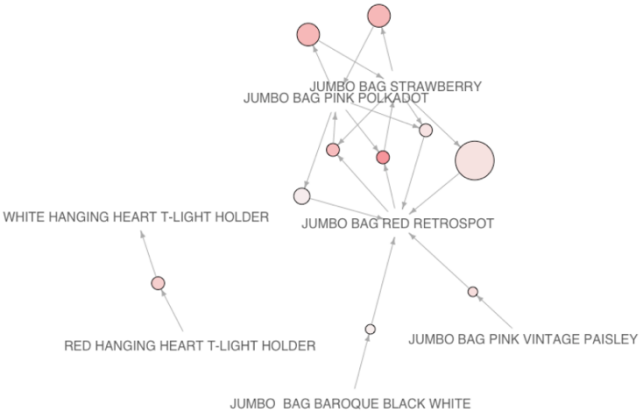

Fig. 8. Visualization of Apriori Agorithm on Bulk Order Dataset

The graph structure is composed of two components: the nodes and the arrows. The nodes represent either a single or combination of items within an association rule. The arrows indicate the direction or the relationship. This means that if an arrow is present from item A to item B, every time item A is purchased item B will be bought together with it as well.

2) Second Dataset

According to the results in the previous section, all three algorithms were equally effective and optimal. The rules generated are visualized through plots in the following figures to make it easier to comprehend the associations between each item. The rules visualization for the 2014 dataset is shown in Figure 9 and 2015 dataset in Figure 10.

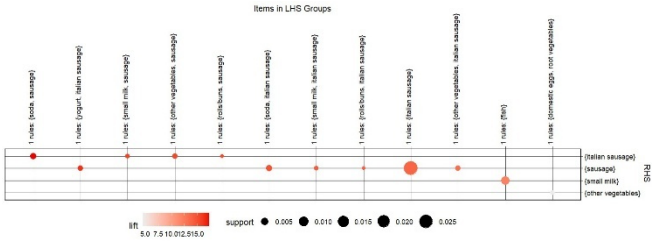

Fig. 9. Visualization of Apriori Agorithm on 2014 Dataset

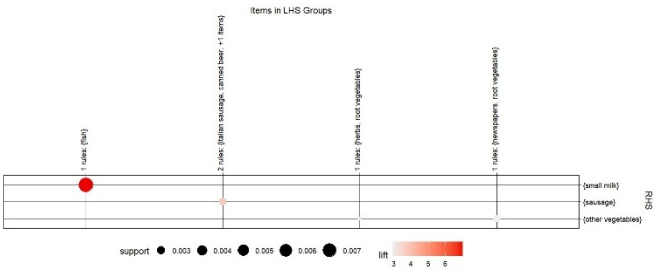

Fig. 10. Visualization of Apriori Agorithm on 2015 Dataset

This plot visualizes the association rules generated. The size of each bubble represents support (the larger the bubble, the higher the support) and the color intensity indicates the lift (the darker the color of the bubble, the higher the lift). The LHS groups are plotted along the x-axis and the RHS items are along the y-axis.

V. Conclusion

This study explored the use of association rule mining in identifying relationships between purchased items in grocery settings. Out of the three methods chosen, namely Apriori algorithm, FP-Growth, and Eclat, the strongest performer depends on the characteristics of the applied dataset as well as their analytical focuses. Visual depiction of the best suited method for each tested dataset were also attached to understand associations that were acquired. These findings gathered could be utilized in deriving actionable insights that benefits the retail field, further encouraging optimization.

VI. Suggestions

The optimization discussed above can be broken down into numerous real-life applications. The analytical approach for Dataset 1 and Dataset 2 differed from the start, in which the first reflected on everyday purchasing behaviors, niche patterns within wholesale-driven purchases, and both conditions occurring within the same transaction. Dataset 2, on the other hand, accounted for temporal patterns and changes in consumer behavior over time. Therefore, suggestions derived from the results collected from the study varies, in which some are shared and some curated specifically based on their characteristics. Shared suggestions include:

a) Bundle Promotions

To take advantage of frequently bought-together items from the association mining that had been done, those said items can be offered for a lower price when bought together as a set, and this shall be done by considering the bundles' lift values. This approach will encourage more customers to buy more of those items in complementary to each other, increasing sale numbers. For example, SET RETROSPOT TEA can be bundled together with SUGAR, or certain vegetables with herbs since they're often seen bought together.

b) Product Placement Optimization

Placing items effectively and strategically would also improve purchasing behavior. This can be done by putting associated items next to or close to one another, encouraging impulsive purchases by enhancing the convenience for shoppers when browsing through the aisles. Based on the results, placing red, roses, and pink variations of TEACUPS and SAUCERS, as well as sausages with buns together could motivate these unplanned spending tendencies.

For Dataset 1, suggestion can be in the form of the following:

a) Bulk Purchase Discount Strategy

This suggestion would especially benefit those purchasing items in bulk by providing bigger discounts the higher the quantity being checked out. Not only it would help clear out the store inventory a lot faster, but it would also attract those purchasing for needs like events and wholesale. For example, customers will be given a discount of 15% if they buy 10 or more sets of SPOTTY RED PAPER CUPS and the same themed PLATES.

For Dataset 2, suggestions can emphasize more on the ever-changing transaction trends, such as the following:

b) Targeted Marketing Campaigns

Targeted marketing campaigns can leverage temporal patterns in customer behavior to develop focused marketing strategies for specific years. For instance, in 2014, promotions could highlight sausages during peak shopping times to drive higher sales. Similarly, in 2015, marketing campaigns could prioritize bundles of small milk and fish, targeting families and home chefs through digital advertisements. Data from previous years can be used to predict upcoming trends, ensuring that marketing campaigns remain relevant and effective.

By implementing the suggestions listed above based on the type of situation accordingly, retail businesses can leverage customer purchasing patterns and behavior to achieve higher revenue. Adopting these strategies would lead to better inventory management and encourages decision making through data-driven approaches.

References

[1] C. Y. Tsai, M. H. Li, and R. J. Kuo, “A shopping behavior prediction system: Considering moving patterns and product characteristics,” *Comput Ind Eng*, vol. 106, pp. 192–204, Apr. 2017, doi: 10.1016/j.cie.2017.02.004.

[2] A. Lagorio and R. Pinto, “Food and grocery retail logistics issues: A systematic literature review,” *Research in Transportation Economics*, vol. 87, Jun. 2021, doi: 10.1016/j.retrec.2020.100841.

[3] R. Kohavi, Z. Zheng, N. Lavrač, H. Motoda, and T. Fawcett, “Lessons and Challenges from Mining Retail E-Commerce Data,” 2004.

[4] H. Xie, “Research and Case Analysis of Apriori Algorithm Based on Mining Frequent Item-Sets,” *Open J Soc Sci*, vol. 09, no. 04, pp. 458–468, 2021, doi: 10.4236/jss.2021.94034.

[5] H. Wu, “Data association rules mining method based on improved apriori algorithm,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2020, pp. 12–17. doi: 10.1145/3445945.3445948.

[6] J. Han, M. Kamber, and J. Pei, “Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems),” 2011.

[7] J. S. Racine, “RSTUDIO: A platform-independent IDE for R and sweave,” Jan. 2012. doi: 10.1002/jae.1278.