

Exploring the Depths of Market Basket Analysis: A Comprehensive Guide to Transaction Analysis with FP-Growth and Apriori Algorithms

I Wayan Pio Pratama^{1*}

¹Information Technology, Politeknik eLBajo Commodus
Gang Ame Nompos, Labuan Bajo, Nusa Tenggara Timur, Indonesia-86754

*Corresponding author: pio.pratama@poltekelbajo.ac.id Doi: <https://doi.org/10.24036/invotek.v23i2.1094>

This work is licensed under a Creative Commons Attribution 4.0 International License



Abstract

This research investigates the role of data science in understanding customer behavior and enhancing sales, focusing specifically on the application of Apriori and FP-Growth Algorithms at a retail store, Deli Point, in Labuan Bajo. It illuminates the impact of 'rubbish data' on transactional data analysis, emphasizing the need for robust data cleaning procedures to ensure accurate results. Utilizing the faster FP-Growth Algorithm, the study effectively analyzed customer purchasing patterns to identify optimal product combinations for sales improvement. It discovered that 'parsley local' and 'mint flores' items had the highest support with a value of 0.036, indicating that strategic placement of these items together could enhance sales. The rule between chicken leg bone, orange sunkist, and chicken breast boneless was found to have a high confidence value and a lift value higher than 1, implying a higher potential for these items to be sold when positioned near each other. This study contributes to understanding consumer behavior and provides insights for enhancing sales and competitiveness in the retail industry. An association rule involving 'chicken leg bone', 'orange sunkist', and 'chicken breast boneless' demonstrated high confidence and a lift value above one, suggesting significant sales potential when these items are grouped together. This study not only contributes valuable insights into retail consumer behavior and effective product placement strategies but also underscores the transformative role of data science in optimizing sales and boosting competitiveness in the retail sector.

Keywords: Apriori, FG-Growth, Association Rule Analysis, Sales Enhancements, Retail Store.

1. Introduction

In recent years, big data analytics has helped retail industries in improving their customer experience and make better decisions in terms of sales of their products [1]. One specific area of focus for retail companies is the placement of items within their stores. In retail stores, the placement of items on the shelf space significantly impacts the sales of items [2]. Additionally, the understanding of item association in retail business can also assist in the organization of inventory management [3].

As described in previous research, such as [4], the placement of items can significantly impact sales. This is supported by studies such as [5], that placing fruits and vegetables near store entrances should be considered alongside policies to limit the prominent placement of unhealthy foods. Additionally, another study by [5] found that organizing product categories according to their consumption goal leads to increased purchases and expenditures, particularly for less involved consumers and those with less specific shopping goals. Furthermore, practical research studies [6], [7], and [8] have demonstrated that the apriori algorithm can successfully be applied to solve problems related to prediction or to organize the items in a shop. Other research methods for conducting frequent item set analysis are also discussed in [9] and [10] with FP-Growth.

The algorithm was utilized to uncover the most suitable combinations of items to understand the customer's needs and determine strategies to improve sales based on the analysis of transactional data. The Apriori Algorithm is known for its simplicity in implementation. However, it has been found to encounter difficulties when applied to large datasets, resulting in prolonged execution time [11].

Previous studies have reported that the FP-Growth Algorithm demonstrates superior performance in terms of processing speed compared to the Apriori Algorithm [12]–[14]. As the behavior of algorithms can vary concerning different datasets [15], both FP-Growth and Apriori algorithms were employed to examine their performance differences on the given dataset. Further, [14] and [16] emphasize the importance of context-specific evaluation of these algorithms, pointing out that factors such as the number of transactions and the average transaction width can dramatically influence algorithm performance. Despite these insights, there remains a lack of comprehensive comparative studies that examine these performance differences across a diverse range of datasets, especially in emerging applications like real-time analytics and big data environments.

A notable novelty of this study lies in its application of both the Apriori and FP-Growth Algorithms to a retail context to uncover strategic insights about product placement and sales optimization. This offers a unique addition to existing literature where these algorithms have been used individually or not in direct comparison. Furthermore, this study advances our understanding of item association rules in a retail environment, contributing to inventory management strategies.

2. Research Methodology

This study utilized two association rule mining algorithms, namely FP-Growth and Apriori, to assess their effectiveness in facilitating transaction analysis at Deli Point. A series of sequential steps were undertaken to evaluate these algorithms effectiveness. Firstly, data collection and preprocessing were performed before applying the aforementioned association rule mining algorithms. The performance of the algorithms was then evaluated using various metrics such as support, confidence, and lift. Furthermore, rules that produced only one item were trimmed as they were deemed to be not informative. Figure 1 shows the steps carried out in this study. The process starts with data collection, data preprocessing, apriori, and rules selection.

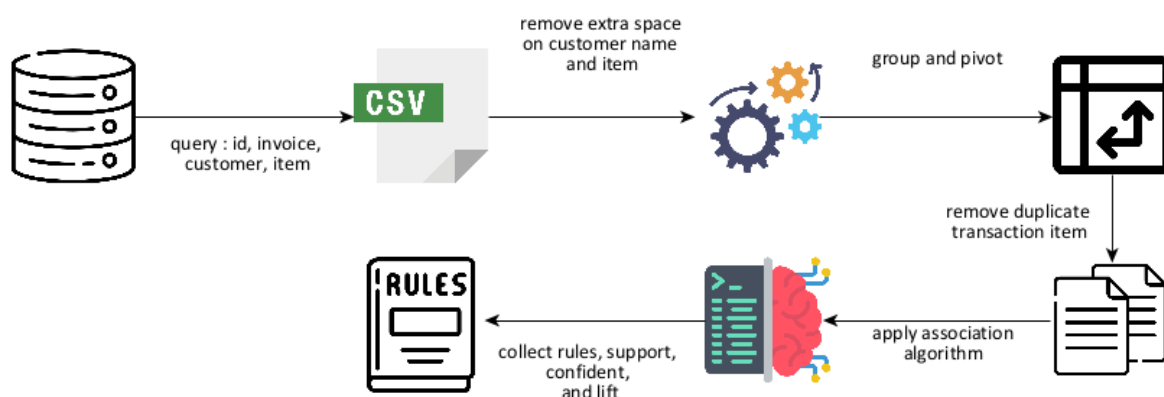


Figure 1. Workflow Diagram

2.1 Data Collection

To gather the data for this study, we first extracted transaction records from a MySQL database. The data included information such as customer name, transaction date, invoice number, and the item purchased. However, certain columns such as price, discount, quantity, and total price were removed as they were not relevant to the analysis. In total, the data used in this research was 71,537.

2.2 Data Preprocessing

In the data cleaning process, we first removed any extra spaces in the customer name and item columns to ensure accurate analysis. Additionally, we applied case folding to the customer’s name and item. After that text cleaning is applied to the customer’s name. After cleaning the data, we grouped the data by the transaction date and invoice number and pivoted the item column. This was done to ensure that the data was organized in a way that would allow for accurate analysis. We then removed any duplicate transactions that were present in the dataset to further ensure the accuracy of the results

2.3 Training Process

The algorithm was implemented using the "mlxtend" library in Python, with minimum support of 0.01 and minimum confidence of 0.02, as there is no sufficient rule produced with minimum confidence and support as in [1]–[4].

2.4 Rule Matrices

There are three metrics to be considered to evaluate rules produced by association algorithms. There are support, confidence, and lift. Using those metrics then we can decide which rule to pick based on our specific needs. Support is a measure of how frequently an item set appears in the transactional data. It is calculated as the number of occurrences of the item set divided by the total number of transactions. The minimum support threshold is a user-defined parameter that determines the minimum number of occurrences an item set must have to be considered frequent [5].

$$Support(X) = \frac{Transactions\ Containing\ X}{Total\ Transactions} \tag{1}$$

Confidence is a measure of the strength of the association between an antecedent item (i.e., the "if" part of the rule) and a consequent item (i.e., the "then" part of the rule) [5]. It is calculated as the ratio of the support of the item set (i.e., the antecedent and consequent items together) to the support of the antecedent item. A confidence value close to 1 indicates that the consequent item is highly likely to be purchased when the antecedent item is purchased

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} \tag{2}$$

Lift is a measure of the strength of the association between two items. It is calculated as the ratio of the support of the item set (i.e., the items together) to the product of the support of the individual items [5]. A lift value greater than 1 indicates that the two items are more likely to be purchased together than would be expected based on their supports.

$$Lift(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X).Support(Y)} \tag{3}$$

2.5 Rule Selection

The FP-Growth and Apriori Algorithm generates a large number of candidate sets, which requires pruning for further analysis [6]. To achieve this, the resulting rules are filtered to include only those with a minimum length of 2. These filtered rules are then weighted based on their proportion of total transactions in each customer group, which is obtained by dividing the support of each rule by the overall total transactions. The rules are sorted according to support, confidence, and lift metrics, allowing for the optimal placement of closely related items to be determined.

3. Result and Discussion

3.1 Result

3.1.1 Data Collection

The required data was procured from a relational database using a SQL query. The query was executed to extract transaction data from the "transactions" table and item data from the "items" table. The query employed a JOIN operation to associate the transactions and items based on the "id_item" field. The resulting data was then exported to a CSV file through the utilization of the "INTO OUTFILE" clause. The fields within the CSV file were separated by a comma and enclosed within double quotation marks, while each line was terminated by a new line character. This SQL query enabled the efficient retrieval and export of the transaction data necessary for the analysis performed in this study. Figure 2 shows the query to extract the data.

```
SELECT tt.id, tt.invoice, tt.customer_name, it.item INTO OUTFILE 'file.csv' FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\n' FROM transactions tt INNER JOIN items it ON tt.id_item=it.id_item;
```

Figure 1. Query to Extract the Data

Using the query in Figure 2, where 71,537 data can be extracted in the range of February 2019 to October 2022. During a specified period, there are 755 unique item transactions taking place. Figure 3 shows the five most popularly sold items in Deli Point.

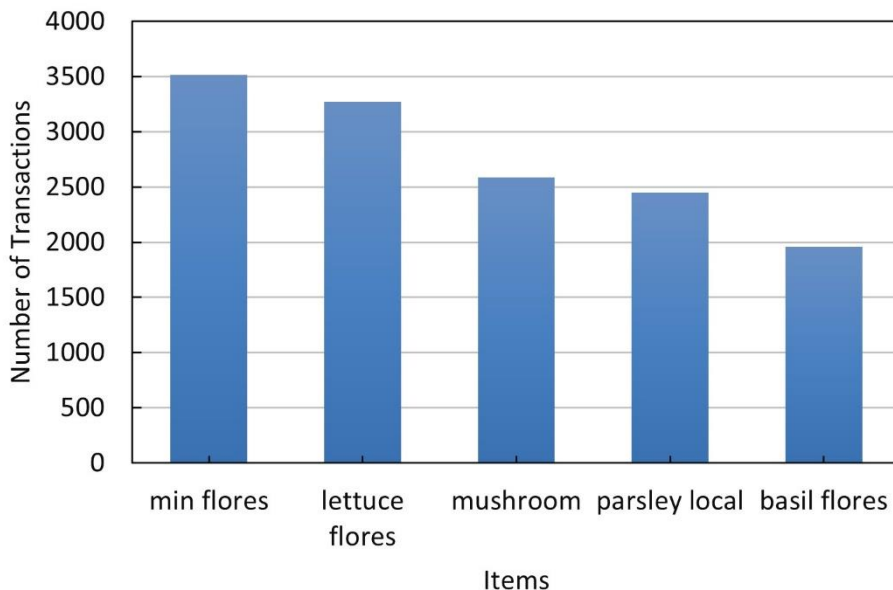


Figure 3. Top 5 Most Popularly Sold Items

Figure 3 provides a frequency analysis of the sales transactions for the top five items at Deli Point, a retail establishment in Labuan Bajo. The 'min flores' item is the most frequently purchased, as evidenced by its highest bar, representing the greatest number of transactions. It is followed, in decreasing frequency, by 'lettuce flores', 'mushroom', 'parsley local', and 'basil flores'. These findings are derived from a comprehensive analysis of the store's sales data, highlighting which items are most often chosen by customers. This frequency analysis is essential for making informed decisions on inventory management, marketing, and sales strategies, such as the placement of the most popular items in prominent locations to maximize visibility and encourage additional purchases. It is a fundamental component of understanding and catering to consumer demand to optimize sales performance.

3.1.2 Data Preprocessing

Before delving into the specifics of the data preprocessing steps undertaken in this study, it is crucial to understand the importance of this phase in the overall data analysis process. Data preprocessing is a set of procedures aimed at transforming raw data into a clean dataset. In the context of our research, this means refining the transactional data from Deli Point to ensure that the subsequent frequency analysis reflects accurate customer behavior. These procedures typically include the removal of extra spaces, cleaning of text data to eliminate unwanted characters, and case folding to standardize text format. Each step is designed to address different aspects of data cleanliness and structure, which are fundamental for reliable analysis. The following points provide detailed examples of each preprocessing step implemented:

- Extra Space Removal

The aim of the extra space removal step in data preprocessing is to standardize the formatting of the text data, which facilitates accurate data analysis and pattern recognition. Inconsistent spacing can introduce errors or bias in the analysis, particularly when dealing with textual data where space characters can significantly alter string matching algorithms and analytical outcomes. The objective is to ensure that all items in the dataset are uniformly formatted with appropriate spacing, thus eliminating any variations solely due to spacing errors. Table 1 shown extra space removal result example demonstrates the effectiveness of this process by showing examples of the original text and the results after the extra spaces have been removed. These examples highlight how the removal of leading, trailing,

and excessive inter-word spaces can clean and normalize the data, preparing it for further analysis such as frequency counts and the application of data mining algorithms.

Table 1. Extra Space Removal Result Example

Original Text	Result
Lbajo Komodo *	Lbajo Komodo *
Guest Bajo (January)	Guest Bajo (January)
eLBajo Komodo	eLBajo Komodo

▪ Text Cleaning

This process is a crucial component of data preprocessing with the aim to purify the dataset by removing extraneous characters such as (, *, # were removed that could potentially skew the analysis. The objective of this step is to ensure that the dataset contains only meaningful and relevant characters, thereby enhancing the accuracy of the analysis. Table 2 showcases the results of this text cleaning process. For instance, the original entry 'Lbajo Komodo *' is stripped of the asterisk to yield 'Lbajo Komodo', and 'Guest Bajo (January)' has the parentheses and content within removed to become 'Guest Bajo January'. These transformations are representative of the broader cleaning process that was applied to the entire dataset. By removing these irrelevant characters, the dataset is standardized, reducing noise and enabling more precise data mining techniques. This cleaned dataset forms the foundation for subsequent analytical tasks, such as identifying trends, patterns, and associations within the data.

Table 2. Text Cleaning Result Example

Original Text	Result
Lbajo Komodo *	Lbajo Komodo
Guest Bajo (January)	Guest Bajo January

▪ Case Folding

This is a fundamental data preprocessing technique with the objective to bring about uniformity in textual data. It involves converting all letters in the dataset to lowercase [7], which is a crucial step for ensuring consistency, especially in cases where text data is case-sensitive. This uniformity is important because many data processing and text analysis algorithms treat letters with different cases (uppercase vs. lowercase) as distinct characters, which can lead to discrepancies in the results. Table 3 displays the results of applying the case folding process to the dataset. For instance, the customer name 'Lbajo Komodo' from the original text remains the same as it does not contain uppercase letters beyond the initial character. However, 'Guest Bajo January' is transformed to 'Guest Bajo January', with all letters standardized to lowercase. This demonstrates the case folding effect on the dataset, ensuring that all entries are treated equally in subsequent analysis stages. The process aids in eliminating any variations that arise solely due to case differences, thereby facilitating more accurate matching and comparison of text strings. As a result, this step is vital for analytical processes that follow, such as text mining and pattern recognition, as it enhances data integrity and contributes to the reliability of the analysis.

Table 3. Case Folding Result Example

Original Text	Result
Lbajo Komodo	Lbajo Komodo
Guest Bajo January	Guest Bajo January

Additionally, the process of trimming customer names is performed to retain only the first name of the customer. This is done to avoid issues that may arise during the research process when customers are differentiated based on the time of transaction. For example, in this research, customers with the name "Anaya January" are considered the same as customers with the name "Anaya February." This

step ensures that customer names are consistently represented throughout the transactional data, improving the accuracy of the association rule analysis.

3.1.3 Group Pivoting

In the process of conducting association rule analysis, it is necessary to transform the raw transactional data into a structured format that can be interpreted by computational algorithms. To facilitate this data reshaping, the data is initially transformed into a pandas data frame, which provides a convenient and flexible framework for data manipulation. Subsequently, a grouping operation is performed based on the date and invoice number attributes, and the resulting data is subjected to a pivot transformation to arrange it into an appropriate format suitable for input into the "mlxtend" library, which implements the association rule mining algorithm.

The data shape depicted in Table 4 before the application of the pivot transformation reveals that the data is organized on a per-customer basis, with each transaction represented as a separate row, even when multiple transactions are performed simultaneously. This structure may not be ideal for association rule mining, as the relationships between items in a single transaction may not be effectively captured in this format. The pivot transformation aims to remedy this by reorganizing the data more suitably for association rule analysis. After this process, the number of transactions was drastically reduced to 54,744.

Table 4. The Shape of the Original Data Example

Date	Invoice	Customer	Item
2019-04-01 17:17:31	2019-02-07 11:51:529	Anaya	Anchovies fillet 720 gr
2019-04-01 17:17:31	2019-02-07 11:51:529	Anaya	Aust unsalted butter 10 kg

The results in Table 5 demonstrate the new format of the data, where transactions with the same invoice and date values are combined into a single row, and the items are merged into a single line of data. This transformation enables the analysis of the relationships between items within a single transaction, facilitating a more comprehensive understanding of the underlying associations within the transactional data.

Table 5. Shape After Group and Pivoting

Date	Invoice	Customer	Item
2019-04-01 17:17:31	2019-02-07 11:51:529	Anaya	Anchovies fillet 720 gr, Aust unsalted butter 10 kg

3.1.4 Remove Duplicate

Data consistency is improved by resolving the problems like bad data or duplicate data and instead of finding the whole dataset, focusing on finding associations in the filtered dataset [8]. Following both of these processes, the data was further reduced to 29,893. Duplicate data removal is an essential step in data preprocessing aimed at ensuring the uniqueness of each data entry. Duplicate records can occur due to various reasons such as data entry errors or data integration processes, and they can significantly distort statistical analysis by giving undue weight to repeated entries. The goal of this step is to identify and eliminate redundancy in the dataset to maintain the accuracy of the analysis and the integrity of the dataset. Table 6 shows the dataset before the process of duplicate data removal. It includes an example of a redundant record where the same transaction appears twice with identical timestamps, invoice numbers, customer names, and purchased items. This repetition does not reflect additional transactions but is rather an unnecessary replication.

Table 6. Before Duplicate Data Removal

Date	Invoice	Customer	Item
2019-04-01 17:17:31	2019-02-07 11:51:529	Anaya	Anchovies fillet 720 gr, Aust unsalted butter 10 kg

2019-04-01 17:17:31	2019-02-07 11:51:529	Anaya	Anchovies fillet 720 gr, Aust unsalted butter 10 kg
---------------------	----------------------	-------	---

After duplicate date was removed, the result become what is shown in [Table 7](#).

[Table 7](#). After duplicate data is removed

Date	Invoice	Customer	Item
2019-04-01 17:17:31	2019-02-07 11:51:529	Anaya	Anchovies fillet 720 gr, Aust unsalted butter 10 kg

3.1.5 Apriori Algorithm

The Apriori Algorithm was applied to preprocessed data, with the difference being that the data was not grouped by customer but instead was kept in a single list variable which was fed to the Apriori library. In this process, we used minimum support of 0.01 and minimum confidence of 0.02.

In [Table 8](#), each rule is evaluated using three key metrics: Lift, Confidence, and Support. Lift is a measure of how much more frequently the items in the rule (antecedent and consequent) are purchased together than would be expected if they were independent; a lift value greater than 1 indicates a positive relationship. Confidence provides an estimate of the probability that the consequent item is bought when the antecedent is purchased, with higher values indicating stronger rules. Support quantifies how common the rule is within the dataset, represented as the proportion of transactions containing both the antecedent and consequent. These metrics collectively help in assessing the strength and relevance of each rule, providing insights into customer purchasing patterns. For example, the rule 'parsley local → mint flores' with a lift of 2.56, confidence of 0.31, and support of 0.036 suggests that transactions including 'parsley local' are more likely to include 'mint flores' as well, and this occurs in 3.6% of all transactions. Similarly, the rule 'Chicken leg bone in & Orange sunkist → Chicken breast whole boneless' has a very high lift of 27.84, indicating a strong association between the purchase of Chicken leg bone in & Orange sunkist and Chicken breast whole boneless, with this rule being true in 96% of cases where the antecedents are bought, and appearing in 1.2% of all transactions. Meanwhile, the highest lift value was obtained for the transaction “green bell peppers & yellow bell peppers → red bell peppers”. A total of 239 rules were formed.

The rules in this table have been selected for their significance in terms of these metrics, indicating potential patterns of customer purchasing behavior. Understanding these patterns can be valuable for strategic decision-making in areas such as marketing, inventory management, and product placement.

[Table 8](#). Interesting Rule Example

Rule	Lift	Confident	Support
Parsley local → Mint flores	2.56	0.31	0.036
...
Chicken leg bone in & Orange sunkist → Chicken breast whole boneless	27.84	0.96	0.012
...
Green bell peppers & Yellow bell peppers → Red bell peppers	35.03	0.89	0.014
...
Chicken leg bone in → Chicken breast whole boneless	19.90	0.69	0.017
...
Chicken breast whole boneless → Chicken leg bone in & lemon import	26.20	0.29	0.010

3.1.6 FP-Growth Algorithm

[Figure 4](#) compares the speed of the Apriori and FP-Growth Algorithms in rule generation, with FP-Growth showing superior performance over Apriori. Both algorithms were configured with equal minimum support and confidence levels and generated similar rules. The faster performance of FP-

Growth is theoretically explained by its design; it constructs an FP-tree, a compressed tree structure that represents the database without needing to generate candidate sets or perform repetitive database scans. This approach is more efficient than Apriori's candidate generation method, which requires multiple database scans and is less efficient with increasing itemsets [9], [10]. FP-Growth's divide-and-conquer strategy, which splits the database into smaller conditional databases and builds an FP-tree for each, enables it to quickly find frequent itemsets, making it especially suitable for large datasets. This practical advantage confirms the theoretical benefits of FP-Growth and supports its use in scenarios where computational speed is paramount.

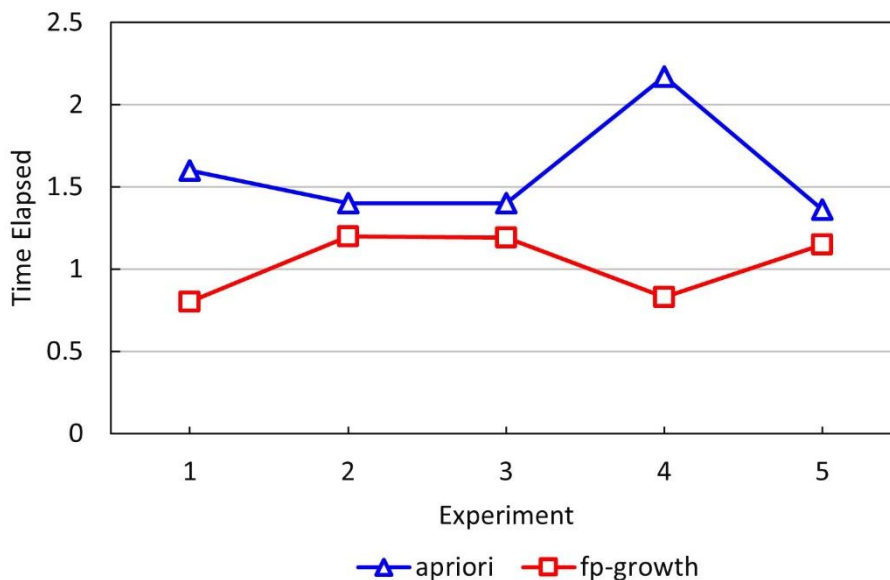


Figure 4. Time Elapsed Apriori and FP-Growth Algorithms

3.2 Discussion

The results obtained indicate that both algorithms successfully discovered interesting patterns from the transactions that have occurred at Deli Point. Notably, we found that both algorithms produce the same results on our dataset, suggesting that they can provide organizations with similar insights and benefits. Items with high transaction frequencies, such as Mint Flores, also appear in the rules with the highest support values. However, overall, the transactions of Parsley Local and Mint Flores only achieve a support value of 0.036, indicating that only 3.6% of the total transactions include both of them in the rule.

One intriguing finding is that when combined with orange sunkist, lemon import, and chicken leg bone exhibits a high confidence value. This is intriguing, as chicken and orange Sunkist belong to different product segments, i.e., meat and fruit. This could potentially be leveraged for cross-selling techniques, where a secondary, cheaper product is offered in addition to the main product [11]. On the other hand, the rule between chicken leg bone and chicken breast boneless suggests the possibility of using up-selling techniques [12].

The rule between Chicken Leg Bone In, Orange Sunkist, and Chicken Breast Whole Boneless with a high confidence value (0.96) and a lift value higher than 1, reaching 27.84, indicates that these items have a higher potential to be sold when positioned close to each other. Other items such as green, yellow, and green peppers also have potential as pairs of items. This is reasonable considering they are all in the same category.

In terms of speed, the FP-Growth Algorithm clearly shows better performance than apriori. This can be seen in Figure 4 where in 5 trials with the same environment, fp-growth outperforms apriori. This is consistent with what has been described in [17]. Other than that, accelerated algorithms for rule generation have the potential to aid organizations in swiftly processing greater volumes of data, a particularly pertinent issue in the context of retail transactions where large datasets are common. This expedited rule generation may facilitate the identification of patterns and trends within the data, enabling organizations to promptly adapt their strategies and make better-informed decisions.

4. Conclusion

In conclusion, the results of this research provide valuable insights into the transaction patterns at Deli Point. Both Apriori and FP-Growth Algorithms were found the same rules. The high confidence value of the rule between chicken leg bone in, orange sunkist, and chicken breast whole boneless, along with the higher lift value, suggests that these items have a higher potential for co-purchasing behavior. Additionally, the speed performance of the FP-Growth algorithm was clearly better than the Apriori algorithm, so this can help organizations to decide quickly based on data transactions. The implications of this study are not limited to inventory management but extend to creating more effective marketing campaigns that target the customer's purchasing habits. The insights gained could inform bundle offers, discounts, or loyalty programs that cater to the established buying patterns, thus enhancing customer satisfaction and loyalty. These findings can be used to inform marketing and product placement strategies to optimize sales at Deli Point. Future research can be developed by searching for associations based on customers or sessions, which will provide a richer perspective.

References

- [1] M. Y. Ardianto, S. Adinugroho, and I. Indriati, "Penentuan Tata Letak Produk menggunakan Algoritma FP-Growth pada Toko ATK," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 5, no. 9, pp. 3826–3832, 2021, [Online]. Available: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/9741>
- [2] W. P. Nurmawati *et al.*, "Market basket analysis with apriori algorithm and frequent pattern growth (Fp-Growth) on outdoor product sales data," *Int. J. Educ. Res. Soc. Sci.*, vol. 2, no. 1, pp. 132–139, 2021, doi: <https://doi.org/10.51601/ijersc.v2i1.45>.
- [3] W. W. Ariestya, W. Supriyatn, and I. Astuti, "Marketing strategy for the determination of staple consumer products using FP-growth and apriori algorithm," *J. Ilm. Ekon. Bisnis*, vol. 24, no. 3, pp. 225–235, 2019, doi: <http://dx.doi.org/10.35760/eb.2019.v24i3.2229>.
- [4] F. Z. Ghassani, A. Jamaludin, and A. S. Y. Irawan, "MARKET BASKET ANALYSIS USING THE FP-GROWTH ALGORITHM TO DETERMINE CROSS-SELLING," *J. Inform. Polinema*, vol. 7, no. 4, pp. 49–54, 2021, doi: <https://doi.org/10.33795/jip.v7i4.508>.
- [5] E. Hikmawati, N. U. Maulidevi, and K. Surendro, "Minimum threshold determination method based on dataset characteristics in association rule mining," *J. Big Data*, vol. 8, no. 1, pp. 1–17, 2021, doi: <https://doi.org/10.1186/s40537-021-00538-3>.
- [6] L. Shi and Q. Zhu, "Association Rule Analysis of Influencing Factors of Literature Curriculum Interest Based on Data Mining," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–8, 2022, doi: <https://doi.org/10.1155/2022/6866134>.
- [7] E. R. Kaburuan, Y. Sartika, and I. Agustina, "Sentiment Analysis on Product Reviews from Shopee Marketplace using the Naïve Bayes Classifier," *LONTAR Komput.*, vol. 13, no. 3, pp. 150–159, 2022, doi: [10.24843/LKJITI.2022.v13.i03.p02](https://doi.org/10.24843/LKJITI.2022.v13.i03.p02).
- [8] A. Sam, B. Denney, C. Haid, and R. Knight, "Simple tools for examining and cleaning dirty data." 2021. [Online]. Available: <https://github.com/sfirre/janitor>
- [9] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM sigmod Rec.*, vol. 29, no. 2, pp. 1–12, 2000, doi: <https://doi.org/10.1145/335191.335372>.
- [10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, 1994, vol. 1215, pp. 487–499. [Online]. Available: https://www.it.uu.se/edu/course/homepage/infoutv/ht08/vldb94_rj.pdf
- [11] N. K. D. Y. Utami, T. Trianasari, and P. I. Rahmawati, "THE EFFECT OF RELATIONSHIP MARKETING AND CROSS SELLING ON THE MARKETING PERFORMANCE OF INSURANCE PRODUCTS PT SUN LIFE INDONESIA SALES OFFICE SINGARAJA," *JMM UNRAM-MASTER Manag. J.*, vol. 11, no. 3, pp. 179–191, 2022, doi: [10.29303/jmm.v11i3.722](https://doi.org/10.29303/jmm.v11i3.722).
- [12] F. Purwaningtias, "Strategi Up Selling Pada Website Penjualan," *Simetris J. Tek. Mesin, Elektro*

- dan Ilmu Komput.*, vol. 9, no. 1, pp. 109–120, 2018, doi: <https://doi.org/10.24176/simet.v9i1.1910>.
- [13] M. Kavitha and S. T. Selvi, “Comparative study on Apriori algorithm and Fp growth algorithm with pros and cons,” *Int. J. Comput. Sci. Trends Technol.*, vol. 4, no. 4, pp. 161–164, 2016, [Online]. Available: <https://www.ijcstjournal.org/volume-4/issue-4/IJCST-V4I4P28.pdf>
- [14] J. Heaton, “Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms,” in *SoutheastCon 2016*, 2016, pp. 1–7. doi: 10.1109/SECON.2016.7506659.
- [15] P. L. Ginting, N. Dengen, and M. Taruk, “Comparison of Piori and FP-Growth algorithms in determining association rules,” in *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, 2019, vol. 6, pp. 320–323. doi: 10.1109/ICEEIE47180.2019.8981438.
- [16] R. Ranjan and A. Sharma, “Evaluation of frequent itemset mining platforms using apriori and fp-growth algorithm,” *arXiv Prepr. arXiv1902.10999*, pp. 1–6, 2019, doi: <https://doi.org/10.48550/arXiv.1902.10999>.
- [17] A. Bala, M. Z. Shuaibu, Z. KaramiLawal, and R. I. Y. Zakari, “Performance analysis of apriori and fp-growth algorithms (association rule mining),” *Int. J. Comput. Technol. & Applications*, vol. 7, no. 2, pp. 279–293, 2016.