MARKET BASKET ANALYSIS

Submitted by

Name	Reg No:
ANAMIKA U	**223012**
ANAND HARIKUMAR	**223013**
ANANDHU PK	**223014**
ANITTA BIJO	**223015**

In partial fulfilment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics

Of

School of Digital Sciences

Kerala University of Digital Sciences, Innovation, and Technology

(Digital University Kerala)

Techno city Campus, Thiruvananthapuram, Kerala – 695317

July 2023

BONAFIDE CERTIFICATE

This is to certify that the project report entitled MARKET BASKET ANALYSIS submitted by:

Name	Reg No:
ANAMIKA U	223012
ANAND HARIKUMAR	223013
ANANDHU PK	223014
ANITTA BIJO	223015

in partial fulfilment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics is a Bonafide record of the work carried out at KERALA UNIVERSITY OF DIGITAL SCIENCES, INNOVATION AND TECHNOLOGY under our supervision.

Supervisor

Prof. MANOJ KUMAR TK
School Of Digital Sciences
DUK

Course Coordinator

Prof. MANOJ KUMAR TK
School of Digital Sciences
DUK

Head of Institution
Prof. SAJI GOPINATH
Vice Chancellor
DUK

DECLARATION

We**, Anamika U, Anand Harikumar, Anandhu PK** and **Anitta Bijo** students of Master of Science in Computer Science with Specialization in Data Analytics, hereby declare that this report is substantially the result of our own work, and has been carried out during the period March 2023-July 2023.

Place: TRIVANDRUM

Date:09/09/2023

ACKNOWLEDGEMENT

We want to convey my sincere and sincere gratitude to my mentor Dr. T.K. Manoj Kumar, Associate Professor, Digital University Kerala, Trivandrum, for his invaluable assistance, counsel, and support, all of which helped me to successfully complete my project.

We would also like to express my sincere gratitude to Prof. Saji Gopinath for giving me the favourable conditions, insightful advice, and educational resources that improved my capacity to carry out a project of this magnitude.

We would also like to take this chance to express my gratitude to my family and friends for their invaluable help, inspiration, and support throughout the completion of this project.

ABSTRACT

This paper introduces an efficient recommendation system for a bakery, utilizing the F-P Tree algorithm to analyse transactional data from the bakery dataset. By constructing a compact F-P Tree data structure, the system offers rapid retrieval of frequent item sets, enabling real-time personalized product recommendations. Leveraging user preferences and purchase history, the system generates recommendations based on association rules, significantly enhancing recommendation accuracy and computational efficiency compared to traditional methods. This approach caters to the evolving tastes of bakery customers, making it a valuable tool for businesses aiming to boost customer satisfaction and sales in the digital age. Moreover, the system's adaptability to changing user preferences ensures its relevance and effectiveness over time. This recommendation system, powered by the F-P Tree algorithm, offers exceptional scalability, allowing bakeries of all sizes to efficiently handle large volumes of transactional data. It excels in real-time updates, continuously adapting to changing customer preferences, thereby boosting customer engagement and loyalty through personalized product recommendations. Furthermore, it leverages data analytics and association rules to facilitate cross-selling and upselling, optimizing the average transaction value. This system not only provides a competitive edge by attracting tech-savvy customers but also seamlessly integrates with online and mobile platforms, offering valuable business insights for improved inventory management and marketing strategies.

CONTENTS

INTRODUCTION

Retailers utilize market basket analysis, a data mining approach, to boost sales by better understanding client buying habits. Large data sets, such purchase histories, must be analysed to identify product groups and items that are most likely to be bought together.

The introduction of electronic point-of-sale (POS) systems facilitated the use of market basket analysis. The digital records produced by POS systems made it simpler for apps to process and analyse massive volumes of purchase data when compared to the handwritten records retained by store owners.

Market basket analysis (MBA), also known as association-rule mining, is a useful method of discovering customer purchasing patterns by extracting associations or co-occurrences from stores' transactional data bases. It is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. For example, if you are in a supermarket and you buy a loaf of Bread, you are more likely to buy a packet of Butter at the same time than somebody who didn't buy the Bread. Another example, if you are buying a Xiaomi Power Bank in an online store, you are more likely to also buy a carrying case.

An itemset is the collection of items a consumer purchases, and MBA looks for patterns in the links between itemset purchases. A set of product association rules is what the MBA produces. We may utilize MBA to discover intriguing association rules between goods from the transaction data taken from the point of sales systems of retail establishments or the shopping carts of internet retailers. Customers often purchase product B if they purchase product A, for instance. Market basket analysis implementation necessitates a background in statistics and data science, as well as certain algorithmic computer programming abilities. There are commercial, off-the-shelf tools available for folks who lack the necessary technical abilities.

Generally speaking, this report serves as a thorough exploration dataset. We seek to identify key trends in consumer purchasing behaviour through thorough analysis utilizing the FP Tree method. We also seek to identify antecedent and subsequent products and provide insights into the trend in consumer purchasing behaviour. Understanding trends in combination of products is the ultimate goal. Join us on this fascinating trip as we reveal the secrets to market basket analysis success and provide the right consumer recommendations

DATA DESCRIPTION

The dataset employed in this project is openly accessible on Kaggle and encompasses transaction records from "The Bread Basket," a bakery located in Edinburgh's historic centre. These transactions were recorded during the period spanning from October 30, 2016, to April 9, 2017. "The Bread Basket" offers a diverse range of products, including coffee, bread, muffins, cookies, and more. Major description about the columns is:

Date: This column represents the date when a transaction occurred. It provides information about the day on which customers made purchases.

Time: This column represents the time of day when a transaction took place. It provides information about the specific timing of customer purchases.

Transaction ID: This column is a unique identifier for each transaction. It helps differentiate and track individual transactions.

Item Name: This column contains the names or codes of the items that were purchased in each transaction. Each row in this column lists the items that a customer bought during a particular transaction.

The dataset consists of a total of 21294 rows.

The sample of the dataset that is used for the analysis is given below:

	Date	Time	Transaction	Item
0	2016-10-30	09:58:11	1	Bread
1	2016-10-30	10:05:34	2	Scandinavian
2	2016-10-30	10:05:34	2	Scandinavian
3	2016-10-30	10:07:57	3	Hot chocolate
4	2016-10-30	10:07:57	3	Jam
...	...	...	...	...
21288	2017-04-09	14:32:58	9682	Coffee
21289	2017-04-09	14:32:58	9682	Tea
21290	2017-04-09	14:57:06	9683	Coffee
21291	2017-04-09	14:57:06	9683	Pastry
21292	2017-04-09	15:04:24	9684	Smoothies

21293 rows × 4 columns

METHODOLOGY

Data Inspection:

Data inspection involves a thorough examination of the dataset to ensure its quality and integrity. In the case of the dataset, this process includes checking for completeness and correctness. Key aspects of data inspection for this dataset include:

Structure and Size: Understanding how the data is organized, including the number of rows (transactions) and columns (variables).

Data Types: Identifying the data types of each column, such as 'Date' and 'Time' are properly formatted as datetime objects.

Missing Values: Detecting any missing values in the dataset, which require further investigation or handling.

Unique Values: Exploring unique values in the 'Item Name' column to get an initial sense of the types of items sold.

Summary Statistics:

Summary statistics provide an overview of the central tendencies and variability within the dataset. These steps include:

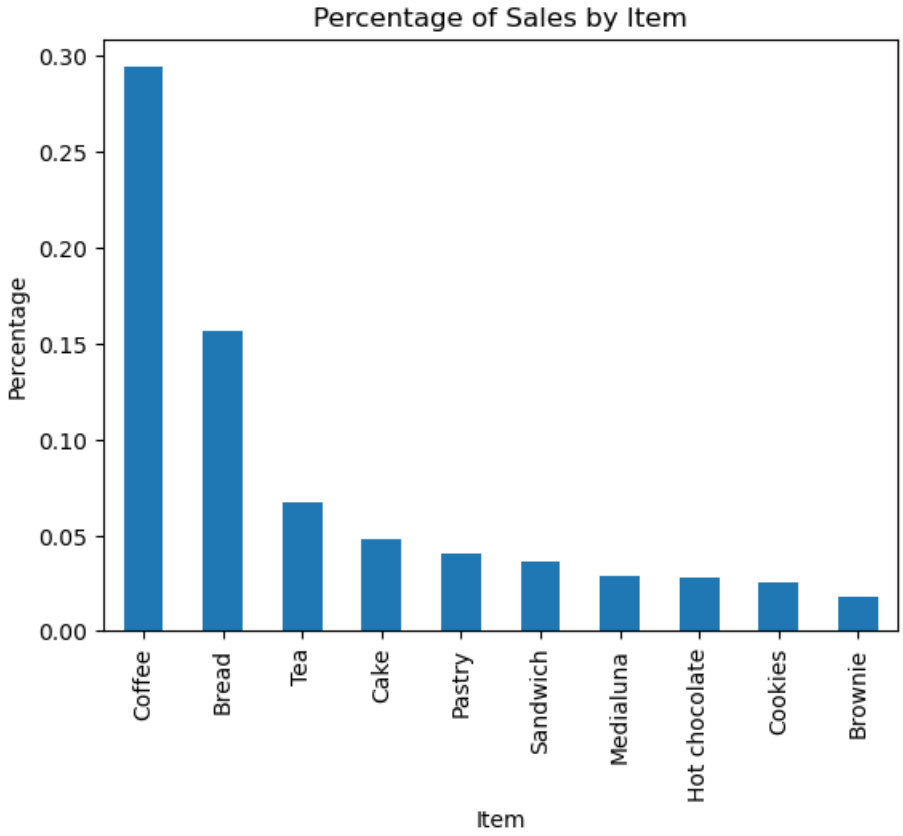
Basic Descriptive Statistics: Computing statistics like the mean, median, and standard deviation of the 'Date' and 'Time' columns to understand the distribution and variation in transaction times.

Transaction Counts: Counting the number of transactions on each date to identify patterns in customer traffic.

Item Frequency: Determining the most and least frequently purchased items by calculating item frequencies.

Data Visualization:

Bar chart to visualise percentage of sales of each item:



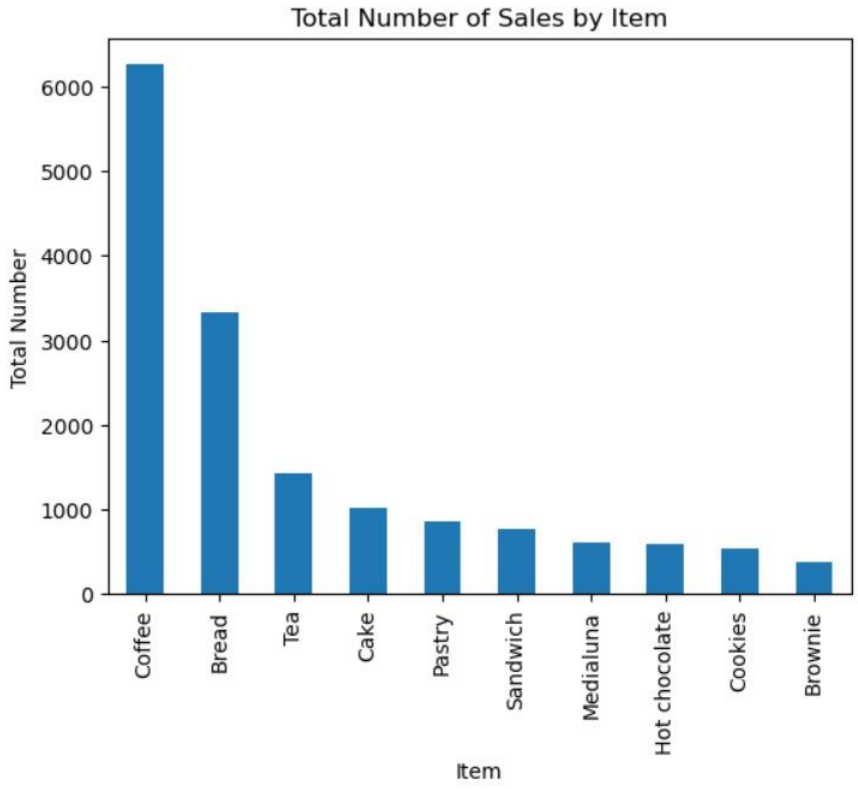
Based on the bar chart, it is evident that certain items consistently outperform others in the bakery's sales. These top-selling items not only contribute significantly to the bakery's revenue but also provide valuable insights for business decisions.

Coffee Dominance (26.7%): Coffee emerges as the unequivocal leader among all bakery items, commanding an impressive 26.7% share of total sales. This highlights the popularity of coffee among customers, indicating a strong demand that can be leveraged for further business growth.

Bread's Strong Presence (16.2%): Following closely behind coffee, bread secures the second position in sales, with a substantial market share of 16.2%. Bread's consistent performance underscores its status as a staple item, ensuring a stable revenue stream for the bakery.

Tea's Notable Contribution (7.0%): Tea, though with a smaller share compared to coffee and bread, still manages to capture a significant 7.0% of the market. This suggests a diverse customer base that appreciates both hot beverages, and tea remains a crucial component of the bakery's offerings.

Bar chart to visualise total number of sales of each item

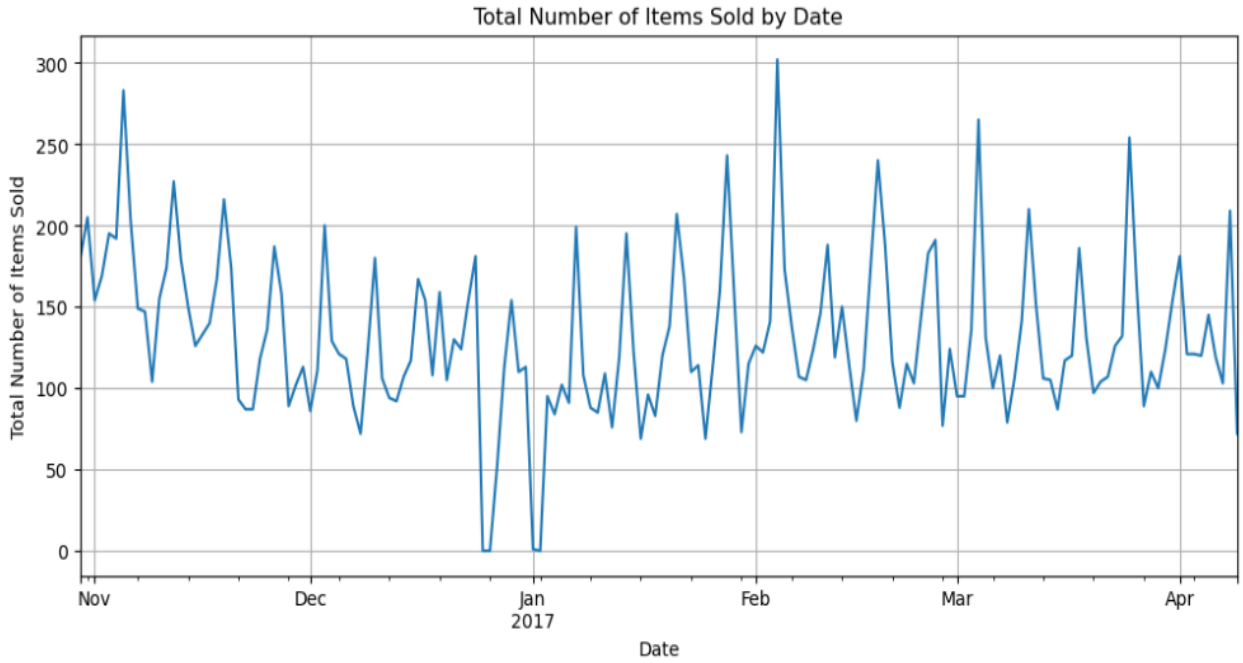


Coffee and Bread Dominance: The bar chart clearly illustrates that 'Coffee' and 'Bread' are the standout performers in terms of sales volume. 'Coffee' leads the way with a substantial number of sales, while 'Bread' follows closely behind.

Diversity of Offerings: Beyond the top two items, the bakery boasts a diverse range of products. This diversity is reflected in the distribution of sales across various items, with each contributing its share to the overall revenue.

High-Performing Specialty Items: The chart also highlights the success of certain specialty items like 'Croissant' and 'Muffin.' These items have garnered significant sales, indicating a niche market or loyal customer base.

Line chart to visualise total number of items sold by date



Daily Sales Dynamics:

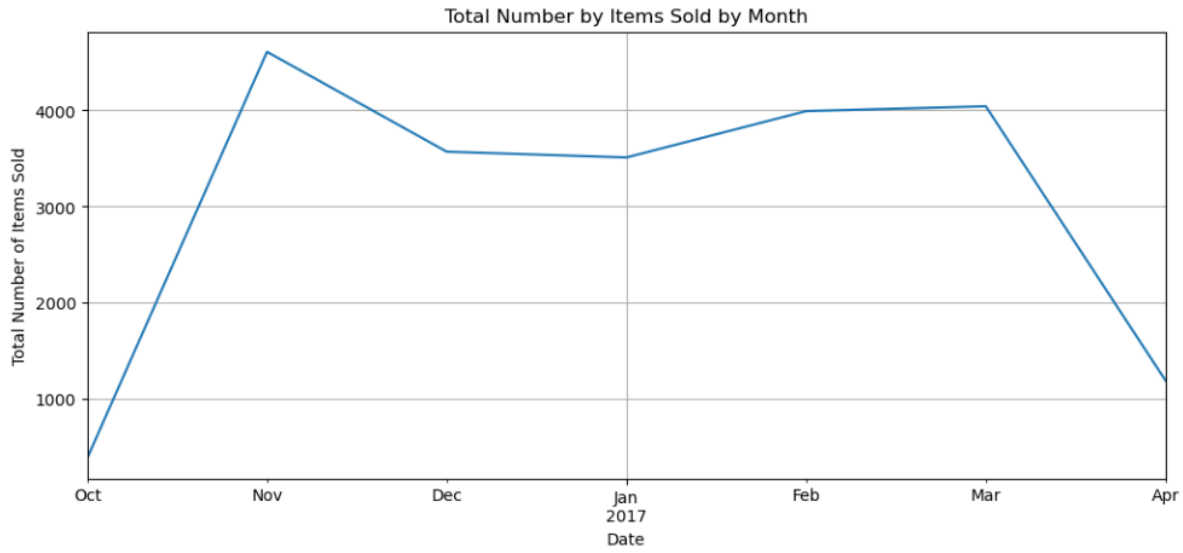
The graph representing the total number of bread items sold by date reveals interesting insights into the bakery's daily sales dynamics. These insights can inform operational decisions and strategic planning.

Seasonal Variations:

The data shows that bread sales exhibit seasonal variations. It's common to observe higher sales during specific seasons or holidays, such as holidays, festivals, or colder months.

Understanding these trends is crucial for effective inventory management**.**

Line chart to visualise total number of items sold by month



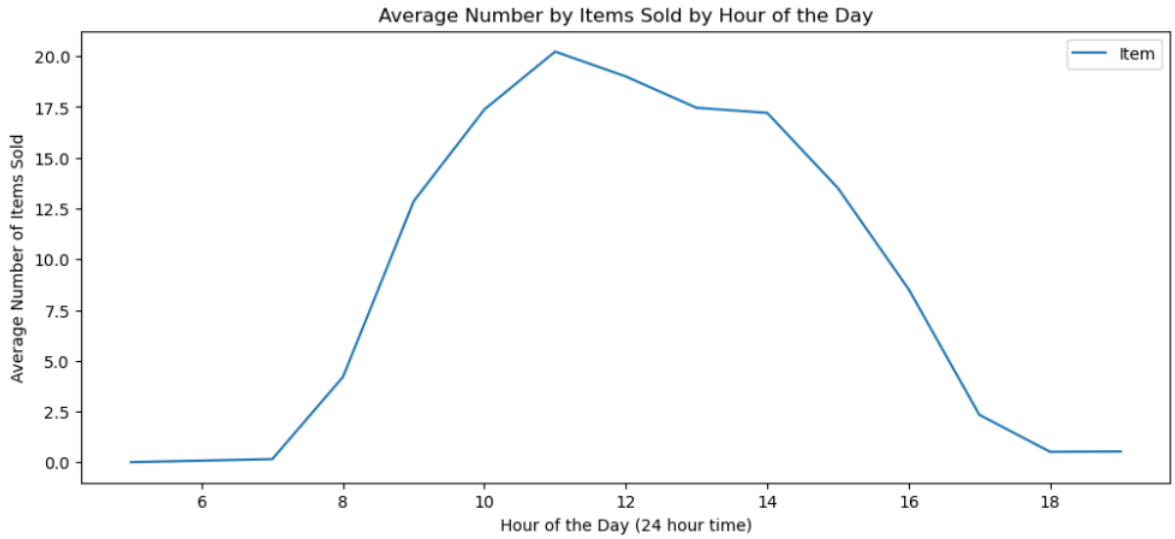
Stability in Sales During Full Months:

The graph depicting total items sold by month from November 2016 to March 2017 reveals a noteworthy trend. During the five full months within this period, there is a remarkable level of stability in sales. This observation is particularly significant given the shorter duration of the months at the beginning (October 2016) and end (April 2017).

Consistency in Customer Demand:

The relatively consistent sales volume during these full months suggests that the bakery experiences a steady and reliable level of customer demand. This consistency can serve as a valuable asset for operational planning and resource allocation.

Line chart to visualise total number of items sold in a day



Sales Patterns Throughout the Day Analysis:

Peak Business Hours:

The data clearly illustrates distinct peak hours in the bakery's sales operations. Sales start to gain momentum at 8 am, with the busiest hour occurring at 11 am. This surge in sales during the morning hours indicates a high level of customer activity and engagement during this time.

Lunchtime Sales Surge:

Notably, a significant portion of the day's sales transactions is concentrated during the lunch hours. This observation suggests that the bakery's offerings are well-received as lunch options by customers, making this period a pivotal contributor to daily revenue.

Afternoon Sales Trend:

As the afternoon progresses, there is a gradual decline in sales activity. This trend is indicative of the typical ebb and flow in customer traffic, with sales tapering off in the late afternoon.

FP-Tree algorithm

Constructing Conditional FP-Trees: The algorithm recursively creates conditional FP-trees for each item, filtering the dataset to transactions containing that item and removing it. This process is repeated until all frequent item sets are discovered.

Mining Association Rules: Finally, the algorithm mines association rules from the FP-tree, which reveal patterns of co-occurring items and their respective support and confidence levels.

Analysis and Findings

There are three metrics or criteria to evaluate the strength or quality of an association rule, which are support, confidence and lift.

1. **Support**

The percentage of transactions containing a specific mix of elements compared to the overall number of transactions is referred to as support. We need to calculate this to determine whether this combination of items is significant or insignificant. In general, we desire a high proportion, i.e., strong support, to ensure that the partnership is beneficial. Typically, we will define a criterion, for example, we will only consider a combination if it appears in more than 1% of transactions. This is called support. From the project, we take the proportion of transactions in which both duck egg and coffee are purchased.

P (Duck egg INTERSECTON Coffee)

= P (Duck egg ∩ Coffee)

$$= \frac{Number\ of\ transactions\ with\ \text{Duck egg}\ \ AND\ \ Coffee}{Total\ transactions}$$

$$= \frac{11}{9684}$$

= 0.001154

- A support value of 1.15 % signifies that 1.15% of all transactions in the dataset involve the simultaneous purchase of "Duck egg" and "Coffee". This value, exceeding the 1% threshold, confirms the presence of substantial support for this association, satisfying the initial criteria for a meaningful association between these items.

2. Confidence

Confidence measures the probability of finding a particular combination of items whenever antecedent is bought. In probability terms, confidence is the conditional probability of the consequent given the antecedent and is represented as P (consequent / antecedent). In our example, it is the probability of both Duck egg and Coffee being bought together whenever Duck egg is bought. Typically, we may set a threshold, say we want this combination to occur at least 80% of times when Duck egg is bought. Confidence (antecedent i.e., Duck egg and consequent i.e., Coffee) = P (Consequent (Coffee) is bought GIVEN antecedent (Duck egg) is bought).

P (Duck egg GIVEN Coffee)

$$= \frac{\text{P (Duck egg} \cap \text{Coffee)}}{P\ (Duck\ egg)}$$

$$= \frac{Number\ of\ transactions\ with\ Duck\ egg\ AND\ Coffee}{Number\ of\ transactions\ with\ Duck\ egg}$$

$$= \frac{11}{12}$$

= 91.6%

- With a confidence value of 91.6%, which surpasses the 80% threshold, we can assert a high level of confidence that whenever "Duck egg" is purchased, "Coffee" will also be bought. This conclusion fulfils the second criteria, affirming the strength of the association between these items.

3. Lift

Lift is a metric to determine how much the purchase of antecedent influences the purchase of consequent. In our example, we want to know whether the purchase of Coffee is independent

of the purchase of Duck egg (or) is the purchase of Coffee happening due to the purchase of Duck egg. In probability terms, we want to know which is higher, P (Coffee) or P (Coffee / Duck egg). If the purchase of Coffee is influenced by the purchase of Duck egg, then P (Coffee / Duck egg) will be higher than P (Coffee), or in other words, the ratio of P (Coffee / Duck egg) over P (Coffee) will be higher than 1.

Formula,

$$= \frac{\text{P (Duck egg } \cap \text{ Coffee)}}{P\text{ (Duck egg)}} > \text{P (Coffee)}$$

$$= \frac{\text{P (Duck egg } \cap \text{ Coffee)}}{P\text{ (Duck egg)} * P\text{ (Coffee)}} > 1$$

$$= \frac{\frac{11}{9684}}{\frac{12}{9684} * \frac{5478}{9684}}$$

= 1.790

With a lift value of 1.790, which exceeds 1, it indicates that the acquisition of " Coffee " is indeed influenced by the presence of " Duck egg " in a transaction, rather than "Duck egg" being purchased independently of " Duck egg." Furthermore, this lift value of 1.790 signifies that the purchase of " Duck egg " has a positive impact on the purchase of " Coffee," elevating the likelihood of acquiring " Coffee " by a factor of 1.79 times.

The values of support, confidence and lift obtained by using FP-Tree algorithm in the dataset for the given example is shown in following image:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
(Duck egg)	(Coffee)	0.001259	0.512013	0.001154	0.916667	1.790318	0.000509	5.855839	0.441996

The values of support, confidence and lift obtained by using FP-Tree algorithm in the dataset is shown in following image:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(Tea, Scandinavian)	(Coffee)	0.000839	0.512013	0.000734	0.875000	1.708940	0.000305	3.903893	0.415190
1	(Hot chocolate, Pastry, Medialuna)	(Coffee)	0.000734	0.512013	0.000630	0.857143	1.674063	0.000253	3.415906	0.402947
2	(Hot chocolate, Sandwich, Cake)	(Coffee)	0.000944	0.512013	0.000839	0.888889	1.736066	0.000356	4.391879	0.424386
3	(Hot chocolate, Cookies, Sandwich)	(Coffee)	0.000734	0.512013	0.000630	0.857143	1.674063	0.000253	3.415906	0.402947
4	(Bread, Sandwich, Medialuna)	(Coffee)	0.000734	0.512013	0.000734	1.000000	1.953074	0.000358	inf	0.488345
5	(Hot chocolate, Farm House)	(Coffee)	0.000839	0.512013	0.000734	0.875000	1.708940	0.000305	3.903893	0.415190
6	(Pastry, Juice)	(Coffee)	0.002308	0.512013	0.001889	0.818182	1.597969	0.000707	2.683926	0.375072
7	(Hearty & Seasonal, Sandwich)	(Coffee)	0.001469	0.512013	0.001259	0.857143	1.674063	0.000507	3.415906	0.403243
8	(Hot chocolate, Hearty & Seasonal)	(Coffee)	0.000734	0.512013	0.000630	0.857143	1.674063	0.000253	3.415906	0.402947
9	(Hearty & Seasonal, Cake)	(Coffee)	0.000734	0.512013	0.000734	1.000000	1.953074	0.000358	inf	0.488345
10	(Soup, Sandwich, Cake)	(Coffee)	0.001154	0.512013	0.000944	0.818182	1.597969	0.000353	2.683926	0.374638
11	(Mighty Protein)	(Coffee)	0.001154	0.512013	0.000944	0.818182	1.597969	0.000353	2.683926	0.374638
12	(Pastry, Coke)	(Coffee)	0.000839	0.512013	0.000734	0.875000	1.708940	0.000305	3.903893	0.415190
13	(Tea, Sandwich, Cake)	(Coffee)	0.001469	0.512013	0.001259	0.857143	1.674063	0.000507	3.415906	0.403243
14	(Pastry, Sandwich)	(Coffee)	0.001154	0.512013	0.000944	0.818182	1.597969	0.000353	2.683926	0.374638
15	(Keeping It Local)	(Coffee)	0.006610	0.512013	0.005456	0.825397	1.612061	0.002071	2.794832	0.382202
16	(Art Tray, Sandwich)	(Coffee)	0.000839	0.512013	0.000734	0.875000	1.708940	0.000305	3.903893	0.415190
17	(Tiffin, Medialuna)	(Coffee)	0.000734	0.512013	0.000630	0.857143	1.674063	0.000253	3.415906	0.402947
18	(Pastry, Toast)	(Coffee)	0.001574	0.512013	0.001364	0.866667	1.692664	0.000558	3.659899	0.409860
19	(Toast, Scone)	(Coffee)	0.000734	0.512013	0.000630	0.857143	1.674063	0.000253	3.415906	0.402947

CONCLUSION

In conclusion, the provided association rule analysis demonstrates interesting patterns and relationships within a dataset related to customer purchases in the bakery. Notably, the antecedents in all the rules seem to be strong indicators of the consequent item, which is "Coffee" in each case. The high confidence values of 91.6% indicate that when customers purchase the items listed in the antecedents, they are highly likely to also buy coffee. The lift values greater than 1 signify that these associations are more likely to occur than if the items were purchased independently. Additionally, the Zhang's metric suggests a strong association, further supporting the reliability of these patterns. These findings can be valuable for optimizing product placement, marketing strategies, and menu design to increase sales and customer satisfaction in the bakery.

REFERENCES

Aldino, A. A., Pratiwi, E. D., Sintaro, S., & Putra, A. D. (2021, October). Comparison of market basket analysis to determine consumer purchasing patterns using fp-growth and apriori algorithm. In *2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)* (pp. 29-34). IEEE.

Hossain, M., Sattar, A. S., & Paul, M. K. (2019, December). Market basket analysis using apriori and FP growth algorithm. In *2019 22nd international conference on computer and information technology (ICCIT)* (pp. 1-6). IEEE.

Liu, Y., & Guan, Y. (2008, November). Fp-growth algorithm for application in research of market basket analysis. In *2008 IEEE International Conference on Computational Cybernetics* (pp. 269-272). IEEE.

Firmansyah, F. (2021). Market Basket Analysis for Books Sales Promotion using FP Growth Algorithm, Case Study: Gramedia Matraman Jakarta. *Journal of Informatics and Telecommunication Engineering*, *4*(2), 383-392.