NOVA IMS Information Management School

MDDM

Master's Degree Program in
Data-Driven Marketing

Data Science in Business: Understanding Growth from a Data-Driven Perspective

A Case Study of a Lisbon-Based Bakery

Marc Jerschov

Project Work
presented as partial requirement for obtaining a Master's Degree in Data-Driven Marketing

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Data Science in Business: Understanding Growth from a Data-Driven Perspective

A Case Study of a Lisbon-Based Bakery

By Marc Jerschov

Master Project Work presented as partial requirement for obtaining the Master's degree in Data-Driven Marketing, with a specialization in CRM & Market Research

Supervised by

Joao Caldeira, Invited Assistant Professor, NOVA IMS

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Marc Jerschov

[Lisbon, 12st of July, 2024]

DEDICATION

To Wofi, Ki, Christine, Niki, Mariana and Igor.

ACKNOWLEDGEMENTS

I would like to thank and express my appreciation to my supervisor, Professor João Carlos Palmela Pinheiro Caldeira, PhD, for all the support and knowledge he shared with me throughout this process. I want to thank him for the kind words that helped me push through the difficult parts of this work, his commitment to making me succeed, and his professionalism. I also want to especially thank Adjunct Professor Vasco Jesus and Mind over Data, who believed in me, provided me with amazing and current data, and always responded to every question I had throughout this process.

Also, I want to thank all my colleagues, professors, friends and family that I could share my little victories and temporary frustrations throughout this process with. They were:

- My professors from my earlier studies and from Nova IMS, who are the central pillars of my modest academic career and who majorly influenced my work.

- Baptiste Larrouy, who provided many helpful and joyful moments, that made this time so very enjoyable, both professionally and personally.

- Jannick Schulte and Hannah Langenbach, who were an endless source of friendship, knowledge, and amazing vegetarian food.

- My family members, to whom I tried to explain my thesis on the phone on a weekly basis. They were steady listeners to my ideas, regardless of how good the call connection was, and supported me every step on the way.

- Kamran Sattari, who was the best emotional support.

- Waleed Sabir, who was the best companion for this data analytical journey, professionally and of course also personally.

- And all the amazing friends I made along the way during my time here in Lisbon.

ABSTRACT

In the highly competitive food industry, data analytics has become an essential tool for driving strategic growth and expansion. Leveraging data insights allows businesses to make informed decisions, optimize operations, and enhance customer experiences, thereby building a strong foundation for business growth. This work analyzes customer and sales data from a renowned bakery in Lisbon, Portugal, which contains data from 16 points of sale across four years. The primary objective of this research is to leverage advanced analytics to identify underlying patterns in the data that help the brand to grow. In order to achieve this, four main categories were identified that build the foundation of growth: profit generation, cost reduction, risk mitigation, and improving innovation life cycles. The insights of the analysis aim to facilitate and optimize strategic business decisions. The underlying patterns were identified by using association rule mining, utilizing the apriori algorithm, with which every shop and every year of that bakery was analyzed. Subsequently, the identified rules were further explored through a Meta-Analysis, utilizing K-means clustering to investigate similarities across points of sale based on their association rules. Additionally, Kruskal-Wallis tests were employed to assess the significance of seasonal variations, followed by Dunn's post-hoc test for a more in-depth analysis. The clusters were further analyzed by examining the coefficient of variation within each cluster to gain deeper insights. The analysis revealed various association patterns that were consistently found across different shops and years. Significance could be found in sales behaviour across the seasons, weekdays, and months. Significant seasonal variations were also observed for the metrics of the association rules. The clusters formed based on the association rules did not exhibit significant differences in sales growth rates. A deeper analysis of the clusters unveiled a more complex structure, underscoring the need for careful decision-making across clusters. Furthermore, the analysis identified business-relevant patterns, such as strong product bundles and resulted in the development of a recommendation system. The results could be integrated into four foundational fields for growth: profit generation, cost reduction, risk mitigation and the improvement of innovation life cycles. The findings of this work were culminated in a strategic framework designed to help businesses leverage their data through an incremental and visual association rule mining approach. Thus, the outcomes can be utilized by other brands to enhance their business processes as demonstrated in this work.

KEYWORDS:

Business Growth, Advanced Data Analytics, Association Rule Mining, K-means Clustering, Strategic Decision Making, Python, Seasonal Analysis

Sustainable Development Goals (SDG):

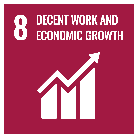

TABLE OF CONTENTS

Statement of Integrity 1
Dedication 2
Acknowledgements 3
Abstract 4
List of Figures 8
List of Tables 9
1. Introduction 10
2. Literature Review 12
3. Study Setup 15
3.1. Conceptual Model 15
3.1.1. Hypotheses 16
3.2. Methodology 16
3.2.1. Research Approach 16
3.2.2. Data Handling 18
3.2.3. Evaluation Metrics 18
4. Results and Discussion 21
4.1. Exploratory Data Analysis 21
4.1.1. Performance of Different Shops 21
4.1.2. Products Sold 22
4.1.3. Seasonal Products 23
4.1.4. Temporal Analysis 24
4.1.5. Pastry Sales 25
4.1.6. Item Rationalization 26
4.2. Association Rules 27
4.2.1. Association Rules (also Redundant Rules) Node Size Based on Indegree 28
4.2.2. Rules (also Redundant Rules) Node Size Based on Lift 29
4.2.3. Non-Redundant Rules: Based on Conviction 31
4.2.4. Top 30: Node Size Based on Indegree, Edge Width Based on Conviction 33
4.2.5. Final Step, Monetary Rule Analysis 34
4.2.6. Commentary on the Remaining Analysis 35
4.3. Meta-Analysis 36
4.3.1. Metric Comparison (All Shops All Years) 37
4.3.2. Top 5 Rules 38
4.3.3. Rules Changing Over Time 39

4.3.4. Seasonal Metric Change 40
4.3.5. K-Means Clustering 43
4.3.6. Hierarchical Clusters 45
4.3.7. Network of Shops (Colored by K-means Clusters) 46
4.3.8. Testing Clusters 48
4.3.9. High Potential Low Sales Analysis 50
4.4. Recommendation System 50
4.5. Threats to Validity 53
4.5.1. Construct Validity 53
4.5.2. Internal Validity 53
4.5.3. External Validity 53
4.5.4. Conclusion Validity 54
5. Managerial Implications 55
5.1. Strategic Implications for Profit Generation 55
5.1.1. Seasonal Adjustments and Product Offerings 55
5.1.2. Promotion and Discount Strategy 56
5.1.3. Recommendation System 56
5.1.4. Utilizing Top Globally Performing Rules 57
5.2. Strategic Implications for Cost Reduction 57
5.2.1. Cluster Analysis for Tailored Strategies 57
5.2.2. Strategic Product Testing 58
5.2.3. Product Rationalization 58
5.3. Strategic Implications for Innovation Life Cycles 58
5.3.1. Product Line Expansion 58
5.3.2. Eco-Friendly Packaging 59
5.3.3. New Seasonal Pairings 59
5.3.4. Optimize Shop Layouts Across Clusters 59
5.3.5. Personalized Marketing 59
5.3.6. Monetizing Data 59
5.4. Strategic Implications for Risk Mitigation 60
5.4.1. Temporal Sales Analysis and Early Risk Assessment 60
5.4.2. Seasonal Analysis and Overstocking 60
5.4.3. Stability-Focused Strategy 60
6. Conclusions and Future Research 62
6.1. Overall Benefit 62
6.1.1. SAFARI-Framework 63
6.2. Limitations Of The Framework 64

6.3. Future Research 65
Bibliographical References 66
7. Appendix 70

LIST OF FIGURES

Figure 1 - Process Model 15
Figure 2 – Quarterly Development of Total Value 2020-2023 for each Shop 21
Figure 3 – Average Quarterly Total Value across all Shops 22
Figure 4 – Number of Different Products 22
Figure 5 – Seasonal Distribution of Pastry Sales 25
Figure 6 – Seasonal Distribution of non-Pastry Sales 26
Figure 7 – Presence of Products in “worst list” for each shop (Clustered) 26
Figure 8 – Zoom in Figure 10 28
Figure 9 – Zoom in Figure 10 28
Figure 10 – Network Graph with all Rules, Node size based on Indegree 28
Figure 11 – Zoom in Figure 12 29
Figure 12 – Network Graph with all Rules, Node Size based on Lift 29
Figure 13 – Zoom in Figure 12 29
Figure 14 – Zoom in Figure 18 31
Figure 15 – Zoom in Figure 18 31
Figure 16 – Zoom in Figure 18 31
Figure 17 – Zoom in Figure 18 31
Figure 18 – Rules after Redundancy Elimination, Node Size based on Conviction 31
Figure 19 – Zoom in Figure 20 33
Figure 20 – Top 30 Rules, Node Size Based on Indegree, Edge Width based on Conviction 33
Figure 21 – Heatmap Support (All Shops All Years) 37
Figure 22 – Support Rate Changes over Time 39
Figure 23 – Lift Rate Changes over Time 39
Figure 24 – Confidence Rate Changes Over Time 40
Figure 25 – PCA Cluster Plot with Shop Labels 44
Figure 26 – Hierarchical Clustering Dendrogram 45
Figure 27 – Network of Shops Based on Rule Similarities (colored by Clusters) 46
Figure 28 – Output of Recommendation System 1 51
Figure 29 – Output of Recommendation System 2 52
Figure 30 – Output of Recommendation System 3 52
Figure 31 – Heatmap Confidence (All Shops All Years) 70
Figure 32 – Heatmap Lift (All Shops All Years) 71

LIST OF TABLES

Table 1 – Best Seasonal Products Spring 23
Table 2 – Best Seasonal Products Summer 23
Table 3 – Best Seasonal Products Autumn 23
Table 4 – Best Seasonal Products Winter 24
Table 5 – Outstanding Rules Shop 1 33
Table 6 – Top Monetary Rules by Unit Price, Shop 1 2020 34
Table 7 – Unit Price of ID 44 and ID 60 34
Table 8 – Top Monetary Rules by Total Sales Value, Shop 1 2020 35
Table 9 – Association Rule Findings 36
Table 10 – Top 5 Global Rules 38
Table 11 – Top 5 Global Rules, 2023 38
Table 12 – Kruskal Wallis Test Hypothesis 2 40
Table 13 – Temporal Variation Findings 42
Table 14 – Results Unique Rules Proportion 47
Table 15 – Results Unique Rules Proportion (without Shop 7) 48
Table 16 – Kruskal-Wallis Results for Hypothesis 3 49
Table 17 – Results of Coefficient of Variation Test 49
Table 18 – Results of High Potential and Low Sales Analysis 50

1. INTRODUCTION

In today's competitive food industry, especially in the aftermath of Covid-19, restaurants, bakeries, and cafés have been compelled to integrate digital technologies into their decison making. The integration of digital strategies, including data analytics, has become a key element in their business approaches (Almansour, 2022). Leveraging data insights allows businesses to make informed decisions, optimize operations, and enhance customer experiences, thereby building a solid foundation for business growth (Troisi et al., 2020). To explore how strategic decisions can be derived from data science insights, this work dives into customer and sales data from a premium bakery in Lisbon, revealing underlying patterns and uncovering growth opportunities. The respective bakery operated in 16 points of sale at the moment of the analysis, offering high-quality products and pursuing ambitious growth and expansion goals. The bakery aims to optimize its resources and establish a data-driven foundation for future decision-making. Particularly during periods of growth the bakery aimed to adopt "lean" methodologies to reduce costs, enhance profit generation, and drive impact with maximum efficiency.
By harnessing machine learning techniques and analyzing shop similarities, the bakery can lay the foundation for strategic decision-making and the development of broader business strategies. Furthermore, the goal of this work is to create a comprehensive business analytics framework, providing firms with guidance on effectively analyzing their data and utilizing these insights to drive strategic growth and informed decision-making. Additionally, this work aims to enhance traditional business efforts through the application of data science techniques, to explore important temporal variations, and to identify the most important rules for business use.
To guide the analysis and ensure comprehensive insights, the following research questions have been formulated. These questions are crucial as they address the core aspects of the data and provide a structured approach to uncovering valuable business insights.

RQ1: How can data science techniques be utilized to enhance traditional business efforts?

RQ2: What are the most important association rules in the bakery sales data?

RQ3: What are the most important temporal variations?

RQ4: Can the findings from individual bakery shops be generalized into a broader growth analysis framework that is applicable to similar businesses in the retail sector?

To address these questions, this dissertation begins with an exploratory data analysis, followed by association rules mining. Subsequently, a Meta-Analysis was conducted, which further explored the patterns discovered through association rules and clustered the shops based on these rules. This revealed insights into the various points of sale and their similarities to each other. The uniqueness of this work lies within the fact that it aims to generate new insights into growth strategies by an innovative combination of advanced data analytics, growth marketing literature, real world empirical data, and business knowledge.
Specifically, it examines the underexplored potential of data analytics to enhance traditional business strategies (Troisi et al., 2020).
The expected contribution of this work lies in the novel insights gained from a unique perspective, demonstrating how data science methodologies, particularly association rule mining, can enhance business processes. Specifically, the association rule mining, based on a visual analysis, provided insights into each shop and year combination, resulting in 25 sets of top-performing rules. These sets were analyzed through a meta-analysis. By synthesizing these insights and translating them into actionable business strategies, this work introduces the

Strategical Application Framework of Association Rule Integration (SAFARI). The SAFARI framework significantly contributes not only to the academic field but also to decision-makers and strategists aiming for similar objectives as outlined in this work. The framework serves as the core contribution of this work.
While the effectiveness of the outcomes cannot be quantitatively measured within the scope of this work, it contributes significantly to the academic field by examining growth strategies through the lens of data science and business knowledge. This approach, applied to a highly specific company type, has the potential to reveal novel insights, enriching current understandings with new perspectives.

2. LITERATURE REVIEW

Business strategies serve as crucial guides for companies aiming to enhance their competitive standing and influence their performance within the industry. Incorporating data science into business processes has gained significant traction over the years (Brynjolfsson et al., 2011). In contemporary times, there is a consensus in academia advocating for the integration of big data analytics into business decision-making processes (Troisi et al., 2023). This includes enhancing managerial openness to analytics, conducting research, and building a strong infrastructure to support the technical aspects of data collection and integration (Troisi et al., 2020) .

Furthermore, data science applies advanced analytics methods and scientific concepts to derive useful business insights from data, with a focus on understanding patterns to predict future outcomes. While basic analytics offer a description of data in general, advanced analytics provides a deeper understanding and detailed analysis of granular data (Sarker, 2021a). This work emphasizes advanced analytics to generate business value.

A study utilizing constructivist grounded theory, based on interviews with 20 Italian start-up founders, identified the critical intersection of technological, managerial, and human dimensions, which the researchers termed "data-driven orientation." The findings demonstrated that the integration of these three dimensions is essential for enhancing a start-up's competitive edge. Consequently, the primary conclusion of the study was encapsulated in the phrase "Think Human, Act Digital.” (Visvizi et al., 2021).

Another notion from this study is that big data can be strongly biased and thus, foster preconceptions. However, when utilized correctly, a data-driven approach enhances decision-making, allows for more flexible strategies, and improves business processes and models (Visvizi et al., 2021). Furthermore, there is a perspective that data itself can be an output of the innovation process, transforming it into a significant competitive advantage. This underscores the necessity of a data-driven mindset and the integration of data into strategic processes (Trabucchi & Buganza, 2019). This approach aligns with the concept of creating knowledge by leveraging a "Client-As-a-Source" perspective, where client interactions and feedback directly contribute to the innovation and enhancement of services and products (Trabucchi et al., 2017).

Machine learning has the potential to help businesses plan and coordinate their activities more effectively (Aluri et al., 2019). It allows for better resource allocation, thus reducing costs and operational improvements by providing insights from data, which are also critical for maintaining competitiveness and viability (Walcott & Ali, 2021).

Furthermore, academia advocates that for a company that aims to grow, integrating technology and data into business processes is essential to foster that process (Ribeiro-Navarrete et al., 2021). Growth is defined as “an increase in the size or the importance of something” (Garcia-Martinez et al., 2023). In general, growth can be achieved with for example new products and services, new market development, new marketing strategies, new market niche exploitation, innovation or external resources (Garcia-Martinez et al., 2023). A comprehensive study analyzing detailed survey data from 179 large publicly traded firms in 2011 revealed that companies adopting data-driven decision-making exhibit output and productivity levels that are 5-6% higher than anticipated. This indicates that the implementation of data-driven decision-making, along with investments in information technology and data science, significantly surpasses expected performance metrics (Brynjolfsson et al., 2011).

Business growth is not solely aiming to achieve new milestones or expanding market reach; it's crucial for maintaining a company's current market position as well. Growth equips businesses with competitive advantages, enabling them to stand strong against rivals and navigate through challenges more effectively (Durmaz & Ilhan, 2015). Moreover, firms that exhibit high growth rates are typically more adaptable and able to quickly respond to market changes. This agility

is manifested through enhanced product flexibility and quality, setting them apart from their competitors (Braguinsky et al., 2021).
Furthermore, growth can be divided into two types; quantitative and qualitative growth within a business. Quantitatively, growth can be observed through increases in output, sales revenue, product range, resources such as the number of employees and capital, and investments. On the other hand, qualitative growth focuses on enhancing the quality of business elements, such as services, products, and effectiveness of business operations, which often cannot be measured in numerical terms (Durmaz & Ilhan, 2015). Academia suggests that growth showed to be also driven by the development of new products or services (Ribeiro-Navarrete et al., 2021).
This focus on innovation is crucial to this work. The academic body considers customer orientation as an important basis for innovation (Wang et al., 2016). Furthermore, innovation is seen as a crucial catalyst for business growth. Consequently, this research emphasizes these three steps, from customer orientation through innovation to business growth as key points in strategic planning (Bekata & Kero, 2024).
A comprehensive study surveyed 416 Italian firms operating in export markets and employed confirmatory factor analysis to explore the relationship of customer orientation, innovation, and firm growth.. The analysis yielded significant results and demonstrated that customer orientation and customer relationship management are distinct constructs that positively influence firm innovativeness. This innovativeness, in turn, significantly impacts firm growth, indicating that innovativeness mediates the relationship between customer orientation and firm growth (Tuominen et al., 2023). Customer orientation involves a focus on customer needs, preferences, and customer relationship management, characterized by the establishment of enduring relationships with customers. Both customer orientation and relationship management can foster innovation and play a crucial role in driving growth within a company (Tuominen et al., 2023). This work will primarily concentrate on customer orientation by analyzing buying behaviors, rather than focusing on relationship management. Companies can develop novel goods and services, increase operational effectiveness, and gain a competitive advantage thanks to innovation (Rakhsitha et al., 2023).
This work focuses on innovation, and on also three more areas: revenue maximization, cost minimization, and risk mitigation.
A big part of quantitative growth is revenue maximization, which is a crucial objective for firms, because it enables them to re-invest in the company to improve efficiency. This can increase return on investment, help the firm to stay competitive, foster innovation and thus, secure long-term growth (Djuraeva, 2021). Businesses can increase their revenue by selling more products, raising prices, or introducing new goods and services (Rakhsitha et al., 2023).
Having a better financial situation, also allows reducing risk, which is crucial for uprising companies. Growing always has an inherent risk and reducing risk to some degree is an integral part of most growth strategies (Rakhsitha et al., 2023).
The third focus point of this work is minimizing Costs. This goes hand in hand with improving revenue, thus maximizing profit. By increasing operational effectiveness, employing cost-efficient production techniques, taking unsuccesful product lines off production and outsourcing non-core activities, businesses can cut costs (Rakhsitha et al., 2023). Businesses must, however, make sure that their efforts to cut costs do not come at the expense of quality or customer satisfaction (Wirtz & Zeithaml, 2018).
The fourth focus point is risk mitigation. In a stable environment with little uncertainty and risk about the demand for products, firms can efficiently manage production scheduling, finished goods inventory management, and the timing and amounts of supplies of raw materials and labor. Firms can thus realize numerous cost savings (Graves, 2011). Risk mitigation is also important for investors. Investments are more likely to be made in safer environments, particularly regarding growth and expansion into new locations and geographies. Investors are

often willing to accept lower returns in exchange for lower business risk (Rakhsitha et al., 2023).
Next to those mentioned aspects of growth, there are a lot of other concepts that improve growth, like alliances with other companies, age as a factor, or governmental subsidies, that are not regarded in this work (Garcia-Martinez et al., 2023).

3. STUDY SETUP

3.1. CONCEPTUAL MODEL

The work is structured to analyze the bakery sales data across multiple shops over four years to identify interesting association rules, similarities between different points of sale, and seasonal variations. This includes a deep shop-by-shop analysis and a Meta-Analysis for overarching business strategies.
Given the exploratory nature of this research, a traditional conceptual model and research design were not applicable. Instead, a process model is presented to illuminate the thought process and the various concepts considered throughout this work. This model, shown in Figure 1 - Process Model provided a structured overview of the methodologies and analytical steps taken to achieve the research objectives.

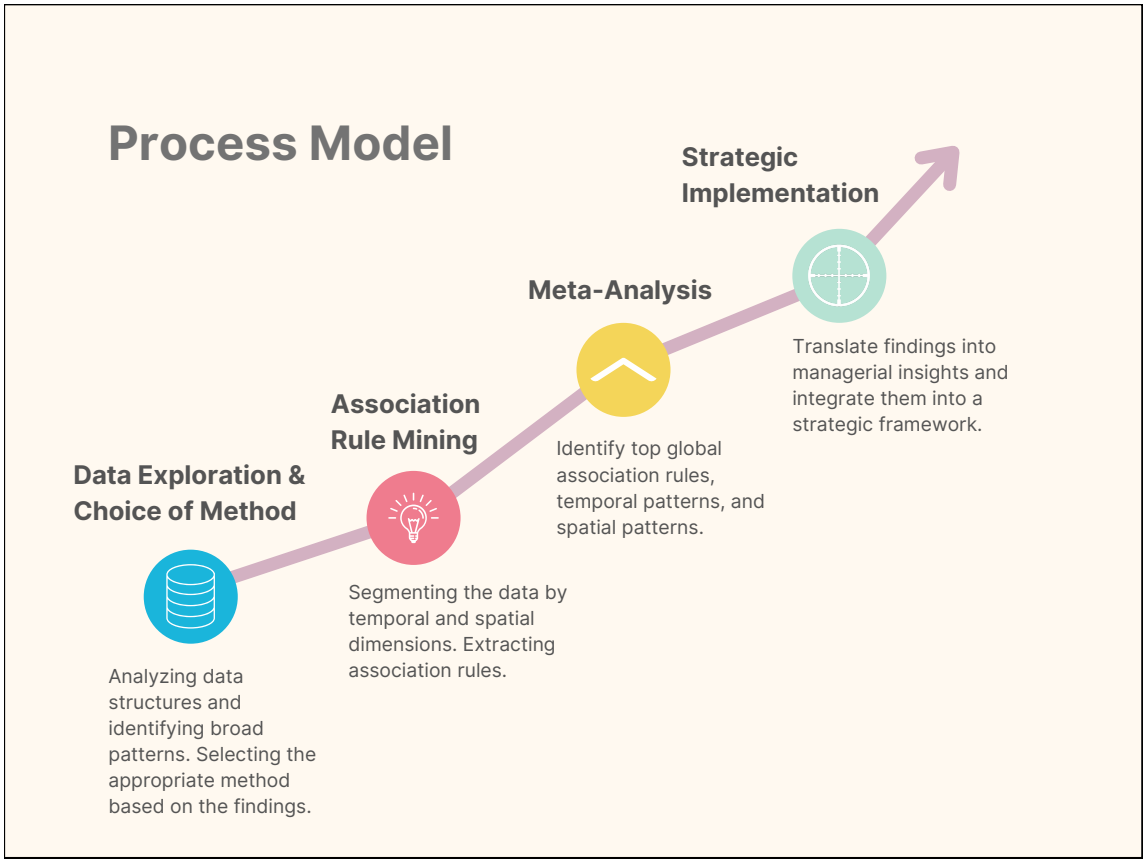

Figure 1 - Process Model

3.1.1. HYPOTHESES

The hypotheses aim to identify significant findings from various approaches. The temporal variations tested in H1a to H1c and H2a to H2c hold great value for bakery managers because they can adapt their various strategies according to these temporal variations. Understanding these patterns allows managers to optimize inventory, staffing, and marketing strategies accordingly. The clustering of shops, as examined in H3, presents a valuable concept for managers, allowing them to explore the applicability of knowledge transfer and decision-making across a cluster.

H1a: There are significant variations in sales across seasons.

H1b: There are significant variations in sales across the week days.

H1c: There are significant variations in sales across months.

H2a: There are significant variations in support across seasons.

H2b: There are significant variations in confidence across seasons.

H2c: There are significant variations in lift across seasons.

H3: Shops with similar customer purchase patterns (based on association rules) show different sales growth rates.

3.2. METHODOLOGY

3.2.1. RESEARCH APPROACH

The goal of this project is to utilize data science techniques to enhance traditional business efforts, which is reflected in RQ1. However, there was no clear method or algorithm chosen prior to the analysis in the research design. Only after completing the exploratory data analysis was the decision made to utilize the unsupervised learning method, association rules.
Association rule learning is a rule-based machine learning approach used to discover interesting relationships, “IF-THEN” statements, associations, and co-occurences between variables (Sarker, 2021b). Furthermore, it is highly scalable and works well for sets of items in large databases (Azevedo & Jorge, 2007). Association rules always consist of an antecedent, also known as the left-hand-side, and a consequent, known as the right-hand-side. In this context, the antecedent consists of one or more products that lead to the purchase of another item. Various metrics can be used to evaluate the strength of this association. The direction of a rule describes the fact that the antecedent leads to the consequent. For example, if products 7 and 8 are already in the basket, product 9 is added as a consequent.
This method was chosen due to the high amount of data entries and the presence of the transaction ID variable, which enabled basket analysis. In healthcare, physicians can use association rules to perform patient diagnosis by assessing the conditional likelihood of illnesses through symptom associations (Sarker, 2021a). In marketing, association rules can help optimize the strategies, sales plan, and improve efficiency of marketing efforts through

association of products. Furthermore, association rules are used for consumer behavior analysis, prediction, customer market analysis, and recommendation systems (Sarker, 2021a).
There are several possible algorithms that can be used to perform association rule mining, amongst the most common and fundamental are Apriori and FP-Growth (Agrawal & Srikant, 1994), (Han et al., 2000), (Sarker, 2021a). The Apriori algorithm works by first identifying all frequent itemsets in the dataset, meaning sets of items that occur together with a frequency above a defined threshold. The algorithm then generates association rules from these itemsets that satisfy a user-defined minimum confidence level. It uses a "bottom up" approach, building from small groups to larger ones, adding one item at a time (Sarker, 2021b).
The Apriori algorithm is the most classic association rules method (Cui, 2021). It is commonly used for large datasets, that are low dimensional (Sarker, 2021b). It is considered the most widely tested algorithm across various domains for association rule mining and is extensively studied and approved. Apriori is very reliable and accurate, since it finds all the frequent itemsets and starts filtering out only after. Furthermore, it is very flexible and can be used for different types of data sets and variable types, while, for example, FP-Growth can only work with binary attributes. In order to use FP-Growth with non-binary attributes, additionally preprocessing is required.
Apriori's main drawback is its computational intensity, this is usually where FP-Growth outperforms Apriori (Sarker, 2021b). However, despite the large volume of data entries, the dataset is segmented into subsets by year and by shop. This approach results in relatively small amounts of data being used at a time to calculate frequent itemsets. Consequently, the algorithm's computational drawbacks are minimized in this context, allowing for effective analysis. Its business purpose is to understand the buying behavior, add additional items to the baskets of customers and thereby increase profit.
Throughout the analysis, different thresholds were gradually increased, and filtering was primarily conducted through visual representation of different rule metrics, primarily using network graphs. Visualizing the data helps especially with large amounts of rules and makes the analysis more accessible (Hahsler & Chelluboina, 2011).
The project also includes a Meta-Analysis that compares and examines patterns across all considered shops and time frames, as well as variations over time. Interpreting the Meta-Analysis aims to answer RQ2 and RQ3
Additionally, the project explores the concept of similarity, by employing a Cluster Analysis based on the results of the Association Rule Mining. The found clusters are then analyzed and tested. This can help the bakery in the future to improve operational efficiency and mitigate risk.
Finally, combining all the findings and consolidating them into a business framework addresses RQ4.
Furthermore, the information about what stands behind the product codes was not provided, hence, the analysis was conducted blindly without any real world knowledge. Although this made the analysis more abstract and more difficult, it strongly reduced any real world biases.
Following an applied research approach, this work intended to address a particular challenge faced by a specific company, using it as a case study. More specifically, the research follows a theory-building case study approach, which is primarily concerned with generating new theories from the ground up by exploring, analyzing, and synthesizing complex data without a pre-established theory (Dul & Hak, 2007). [1]

[1] The whole analysis was conducted in Python.

3.2.2. DATA HANDLING

The data was automatically collected at the poins of sale through the transactions and it was provided by the company “Mind Over Data”. The data contains the purchases in the bakeries and the customers were identified through their anonymized tax number. The dataset contains 3.856.874 records and featured various quantitative and qualitative variables such as:

- *transaction date,*
- *quantity,*
- *unit price*
- *pastry (y/n)*
- *product code,*
- *family,*
- *sub family,*
- *shop ID,*
- *customer ID,*
- *transaction ID*

The data was anonymized by the company "Mind Over Data", ensuring confidentiality and ethical compliance. This is a standard procedure for many data science projects to protect the user or customer privacy (Majeed & Lee, 2021). While having real-world information would provide deeper insights, the data remains classified due to its relevance, still offering significant potential for pattern recognition and value, even without knowing the specific products.
Pre-processing of the dataset involved minimal cleaning, as no significant data inconsistencies were identified. However, there were alterations to data types and feature engineering performed, including the addition of variables for time (such as day, month, and season) and total value (calculated as quantity multiplied by price). Notably, a margin variable was not available in the dataset.

3.2.3. EVALUATION METRICS

Before diving into the findings of the analysis, it is essential to clarify the metrics and terminology. To make the understanding easier, an example of a cookie and a coffee is going to be used throughout this work, where the coffee is (x) the antecedent and the cookie is (y) the consequent.

1) Support

Support measures how often the items in an association rule appear alone or together in the dataset. There is a support for the antecedent, the consequent, and for both together. It is calculated by dividing the number of transactions that include those items by the total number of transactions (Hoque, 2024). The formula for support is:

$$Support(x) = \frac{|x|}{|D|}$$

Equation 1 – Support

where $|x|$ is the number of transactions containing the itemset x, and $|D|$ is the total number of transactions in the dataset. Applied to the example of the cookie and the coffee, this reflects in

the following questions: what percentage of all transactions include coffee? Or what percentage of all transactions include a cookie? Or what percentage of all transactions include coffee and a cookie?

2) Confidence

Confidence measures the strength of the relationship between items in a rule. It indicates the percentage of transactions that include both the antecedent (the initial items) and the consequent (the resulting items) relative to the transactions that include only the antecedent (Hoque, 2024). The formula for confidence is:

$$\text{Confidence}(x \Rightarrow y) = \frac{\text{Support}(x \cup y)}{\text{Support}(x)} = \frac{|x \cup y|}{|x|}$$

Equation 2 – Confidence

Here, $|x \cup y|$ represents the support of both x and y, indicating the number of transactions that contain both the antecedent x and the consequent y, relative to all transactions. Similarly, $|x|$ represents the support of the antecedent, representing the proportion of transactions that contain only the antecedent x, compared to all transactions.
How often do the coffee and the cookie appear together in relation to how often the coffee appears alone? This metric focuses on the antecedent (coffee) and examines, how often based on the coffees appearance, the cookie appears as well.

3) Lift

Lift measures co-occurrence (not implications), meaning it is non-directional and symmetric for antecedent and consequent (Azevedo & Jorge, 2007). This means that the antecedent and consequent share the same lift value, as lift measures the strength of their mutual association rather than the impact of the antecedent on the purchase of the consequent.
Lift essentially measures the statistical co-occurrence of items. It compares the likelihood of items appearing together in the same basket by chance to how often they actually do appear together. By doing so, lift provides insights into the strength of the association between items, indicating whether the presence of one item increases the probability of purchasing the other item beyond what would be expected by random chance.
This is often referred to as the likelihood of the items appearing together compared to if they were independent. Lift indicates the strength of the association between the items (Hoque, 2024). The formula for lift is:

$$\text{Lift}(x, y) = \frac{\text{Confidence}(x \Rightarrow y)}{\text{Support}(y)} = \frac{\text{Support}(x \cup y)}{\text{Support}\,(y) * \text{Support}\,(x)}$$

Equation 3 – Lift

where Confidence ($x \Rightarrow y$) is the confidence of the rule and Support (y) is the support of the consequent y. By substituting confidence with its formula, it becomes evident that lift essentially relates all occurences. Therefore, it is purely statistical and non-directional, providing a measure of the co-occurrence of items without implying any direction between them.

Lift examines how exceptional the co-occurrence of coffee and cookies is, compared to the general frequency of cookie and coffee appearances.

4) Conviction

Conviction reflects the strength of association between antecedent and consequent items in a rule. It also encompasses the concept of independence of the antecedent and consequent, similar to lift. It puts support and confidence in relation, but unlike lift, conviction is sensitive to the direction of the rule. Lift measures co-occurrence only and is symmetric, whereas conviction considers implication (direction) and the support of both X and Y. Furthermore, conviction is highly sensitive to strong confidence values.
A conviction value of 1 indicates that the items are independent. The higher the conviction value gets, the stronger is the dependency. This value has no upper limit.
The conviction value C of a rule means "if a customer buys coffee, they also get a cookie" is C times more likely to be true than if the two items were independent from each other.

$$Conviction\ (X\ \Rightarrow Y) = \frac{\big(1 - Support\ (Y)\big)}{1 - Confidence\ (X \Rightarrow Y)}$$

Equation 4 – Conviction

4. RESULTS AND DISCUSSION

In the following section, the analysis results will be discussed. Initially, a comprehensive exploratory data analysis will be presented, followed by a detailed examination of the association rules analysis for one specific shop. This will demonstrate the methodology used to generate rules for each shop annually, providing a blueprint for understanding the analytical processes involved.

4.1. EXPLORATORY DATA ANALYSIS

4.1.1. PERFORMANCE OF DIFFERENT SHOPS

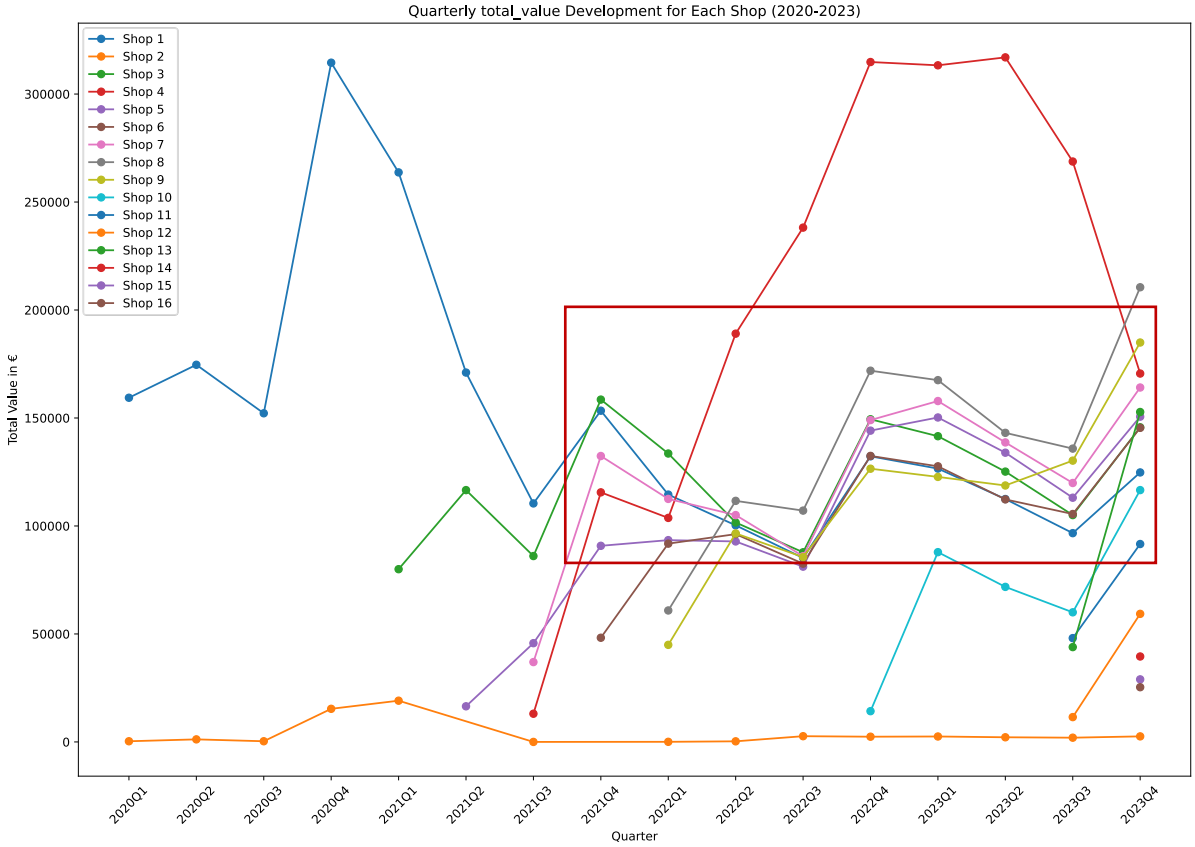

Figure 2 – Quarterly Development of Total Value 2020-2023 for each Shop

As illustrated in Figure 2 – Quarterly Development of Total Value 2020-2023 for each Shop, several shops exhibited comparable behavior in terms of total value.
Given the strong similarities observed in the total value graph for shops 3 through 9 (excluding shop 4), it was particularly interesting to investigate the area highlighted by the red box to determine if these shops exhibited similar patterns in their association rules. Additionally, shop 1 seemed to follow the same pattern as most of the shops after 2022.
Shop 2 was analyzed but appeared to be very different and was therefore excluded from further analysis, the same applied for shop 10. Shops 14, 15, and 16 only had data for 2 months, which could provide a distorted picture, and were thus also disregarded. The data for 2024 was not

analyzed in this work, as it did not cover a full quarter or season and did not include special days like Easter. This could have skewed the results, so it was ignored.

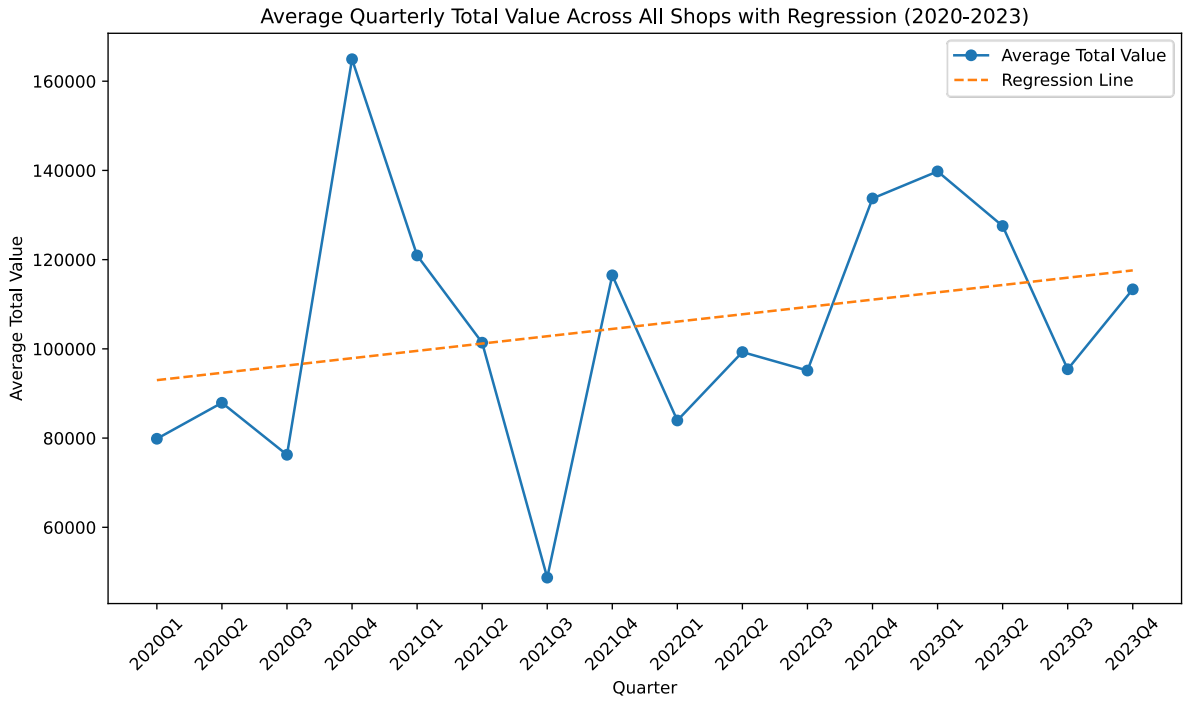

Figure 3 – Average Quarterly Total Value across all Shops

The observed trend in Figure 3 – Average Quarterly Total Value across all Shops, indicated a progressive increase in the total regression line across all shops. Quarterly summations of total values exhibited an upward trend. Furthermore, another analysis showed that the number of monthly transactions, aggregated across all shops, also demonstrated a consistent rise throughout the years. Notably, a heightened total value was consistently observed in Q4 of each year compared to the quarters before, which already indicated a seasonal pattern.

4.1.2. PRODUCTS SOLD

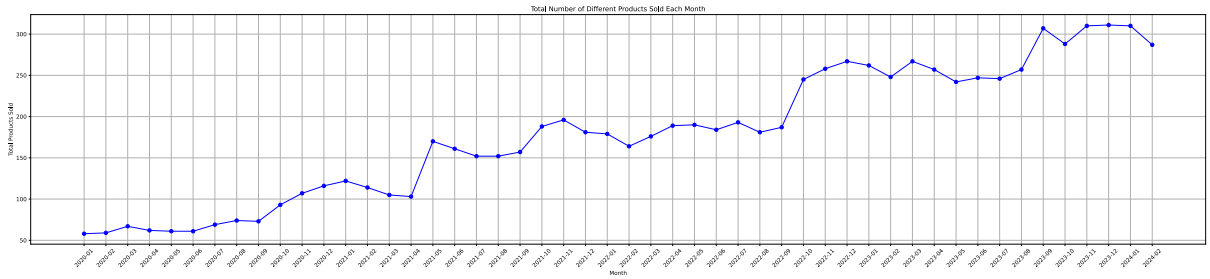

Figure 4 – Number of Different Products

Figure 4 – Number of Different Products showed that there was an increase in the number of different products sold each month. It was noteworthy that product availability varied across different shops and time periods. Not all products were consistently available across all shops, and even within a single shop, product availability varied over time.

4.1.3. SEASONAL PRODUCTS

Products with a unique season count of 1 were selected to ensure that only those exclusively available in a single season were considered for the analysis.
The *total_value* variable was summed up for each product across each year. Afterwards, the *total_value* for all products, all shops, and all years was summed up and then divided by each of the top 5 product's *total_value* to analyze what percentage each product contributed to the *total_value* of all products.
A second analysis was done, that focused only on the year 2023 to provide up-to-date insights. The two analyzes offered two important insights: first, which products were the most important in 2023 and hence, could be potential money makers for the coming year. Secondly, historically important products were identified that performed very well in the past and could be potentially rebranded and reintroduced into the bakeries.

Spring:

2023			All Years		
Product Code	Total Value	Percentage of Total %	Product Code	Total Value	Percentage of Total %
ID 267	9596.58	0.167559	ID 267	27100.00	0.207856
ID 583	5823.09	0.101673	ID 583	5823.09	0.044663
ID 525	3665.53	0.064001	ID 525	3665.53	0.028114
ID 569	1678.27	0.029303	ID 569	1678.27	0.012872
ID 584	1387.95	0.024234	ID 144	1562.17	0.011982

Table 1 – Best Seasonal Products Spring

Summer:

2023			All Years		
Product Code	Total Value	Percentage of Total %	Product Code	Total Value	Percentage of Total %
ID 648	9245.34	0.161426	ID 648	9245.34	0.070911
ID 597	3428.88	0.059869	ID 221	3493.93	0.026798
ID 632	2482.35	0.043342	ID 597	3428.88	0.026299
ID 634	1473.40	0.025726	ID 234	3362.79	0.025792
ID 575	1048.80	0.018312	ID 632	2482.35	0.019040

Table 2 – Best Seasonal Products Summer

Autumn:

2023			All Years		
Product Code	Total Value	Percentage of Total %	Product Code	Total Value	Percentage of Total %
ID 716	7160.40	0.125022	ID 716	7160.40	0.259789
ID 618	5275.53	0.092112	ID 618	5275.53	0.093336
ID 717	2420.10	0.042256	ID 717	2420.10	0.072193
ID 726	1814.20	0.031676	ID 193	2195.97	0.067745
ID 645	1764.15	0.030803	ID 726	1814.20	0.039380

Table 3 – Best Seasonal Products Autumn

Winter:

2023			All Years		
Product Code	Total Value	Percentage of Total %	Product Code	Total Value	Percentage of Total %
ID 731	14878.89	0.259789	ID 719	21575.70	0.165485
ID 719	5345.60	0.093336	ID 731	20701.98	0.158783
ID 729	4134.70	0.072193	ID 761	10414.17	0.079876
ID 718	3879.94	0.067745	ID 759	7803.70	0.059854
ID 728	1984.22	0.034645	ID 729	4134.70	0.031713

Table 4 – Best Seasonal Products Winter

Tables 1 to 4 show the top 5 products for each season for all shops in 2023 on the left and across all years on the right.
The differences between "all years" compared to the ones for 2023 were minimal but noteworthy. For example, Product 731, a product sold in winter, which can be seen in Table 4, achieved a significantly higher Percentage of Total Value in 2023 (0.256%) compared to its value for all years (0.158%). This meant it was a product that had been increasing in sales and could be a business opportunity in the future. Interestingly, Product 761, which was placed third in winter season across all years, did not appear in the top 5 in 2023 at all. This product could potentially be reintroduced and rebranded for the coming winter. Similar patterns were found in all four seasons, indicating great potential for generating more profit.

4.1.4. TEMPORAL ANALYSIS

In this section RQ3 will be partially answered. This is especially interesting for strategical planning across time frames.

RQ3: What are the most important temporal variations?

To explore the temporal patterns further, a significance test was employed to address Hypotheses H1a, H1b and H1c of this work.

H1a: There are significant variations in sales across seasons.

The null hypothesis posited that there are no significant variations in sales across different seasons. First, a Kolmogorov-Smirnov normality test was performed and the data did not show a normal distribution.
To prepare the data, it was grouped by the four seasons. Following this, the means and variances were calculated and compared. For a comparison of three or more groups ANOVA (Analysis of Variance) is the most commonly used method (Lee, 2022). However, it requires normal distribution, which was not the case. If this assumption is not met, the Kruskal-Wallis test can be performed for three or more groups (Barnett et al., 2022).
The results were clear, showing that winter has the highest mean transactional value of 3.64, meaning, on average, each transaction had a value of €3.64. The Kruskal-Wallis test indicated a highly significant difference in sales across the seasons, with a p-value of 0.00 and a Kruskal-Wallis test statistic value of 8514.04. This suggested that seasonal variations had a substantial impact on sales performance. Hence, the null hypothesis was rejected. The ANOVA and Kruskal-Wallis Test compare the means between groups and conclude whether the results of all group means show differences or not. However, they do not give any specific information

about which pairs are significant and in which way they differ. To explore these differences across the seasons even further, a post-hoc test was employed. When the non-parametric Kruskal-Wallis Test was used, Dunn's test can be performed to examine pairwise comparison (Dinno, 2015). The findings from the Dunn test showed that winter exhibits significant pairwise differences and emerges as the season with the highest total value. Specifically, winter's average total value per transaction (3.637852) is significantly higher compared to spring (3.314504, p = 0.0), summer (3.229797, p = 0.0), and autumn (3.232852, p = 0.0). Hence, the null hypothesis can be rejected.
Furthermore, it was interesting, that autumn and summer did not show a pairwise difference ($p > 0.05$).

H1b: There are significant variations in sales across the week days.

Following the same method, the Kruskal-Wallis test showed a test statistic of 10619.50 and a p-value of 0.00 (<0.05), indicating significant variations in profitability across the weekdays. The findings from the post-hoc Dunn test showed that Saturday exhibits significant pairwise differences and emerges as the day with the highest total value. Specifically, Saturday's average total value (3.547970) is significantly higher compared to Monday (3.229104, p = 0.0), Tuesday (3.248827, p = 0.0), Wednesday (3.277151, p = 0.0), Thursday (3.364255, p = 0.0), Friday (3.487038, p < 0.000001), and Sunday (3.411307, p < 0.000001). Monday, Tuesday, and Wednesday did not show any significant pairwise differences ($p > 0.05$). Hence, the null hypothesis can be rejected. Furthermore, Sunday and Friday also exhibited very similar values to Saturday, meaning that they should also be considered as well-performing days.

H1c: There are significant variations in sales across months in 2023.

The following analysis only considered data from 2023. Following the same method, the Kruskal-Wallis test yielded a test statistic of 3633.62 and a p-value of 0.00 (<0.05), indicating significant variations in profitability across the months in 2023. The findings from the post-hoc Dunn test revealed that December stands out with significant pairwise differences, making it the month with the highest total value. Specifically, December's average total value (4.009108) is significantly higher compared to January (3.318087, p = 0.00), February (3.287685, p = 0.00), March (3.313329, p = 0.00), April (3.444881, p = 0.00), May (3.421913, p < 0.00), June (3.382154, p < 0.00), July (3.448076, p < 0.00), August (3.331399, p = 0.00), September (3.324926, p = 0.00), October (3.309013, p = 0.00), and November (3.344771, p < 0.00). Hence, the null hypothesis can be rejected. Furthermore it was interesting that no significant differences were found among January, February, and March ($p > 0.05$).

4.1.5. PASTRY SALES

Exploring seasonality further, an analysis of pastry sales revealed a slightly different distribution across the seasons, depicted in Figure 5 – Seasonal Distribution of Pastry Sales. Winter accounted for over 35% of pastry sales, followed by fall at approximately 25%, spring at around 22%, and summer with 18%. Specifically comparing with Figure 6 – Seasonal Distribution of non-Pastry Sales, it became clear that pastries were better sold during winter and spring, while non-pastries were favored during summer and fall. These findings highlighted seasonal variations in total values, with autumn and

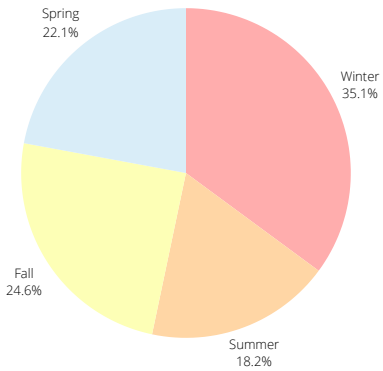

Figure 5 – Seasonal Distribution of Pastry Sales

summer emerging as preferable seasons for non-pastry items, while spring and winter demonstrated higher total values for pastries. This was especially important since 79.6% of the total value in 2023 were derived through pastry sales.

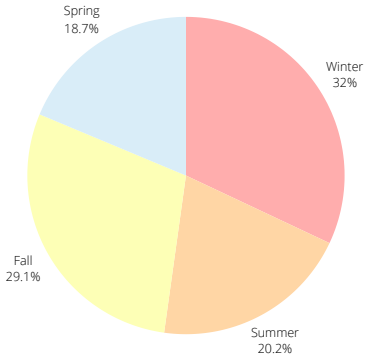

Figure 6 – Seasonal Distribution of non-Pastry Sales

4.1.6. ITEM RATIONALIZATION

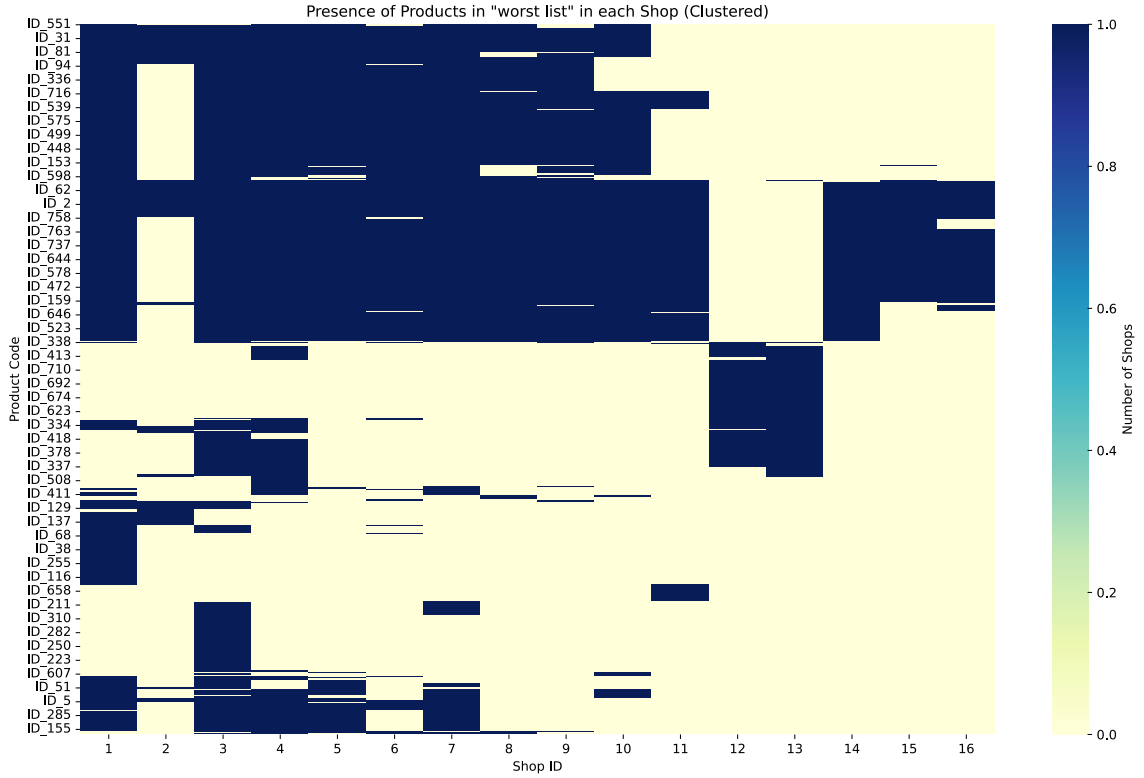

Figure 7 – Presence of Products in “worst list” for each shop (Clustered)

As can be seen in Figure 7 – Presence of Products in “worst list” for each shop (Clustered), upon generating a list of the ten worst-performing products for each shop, a heatmap was constructed to visually represent the presence of these products across multiple shops. Each product is assigned a value of either 0 or 1, depending on whether it appears in the list of worst products for each shop. The rules were clustered in their order on the y-axis for better readability. The analysis of the heatmap revealed a pattern of consistent poor revenue performance for certain products across multiple shops. These products exhibited a "block-like" behavior, where they consistently appeared on the list of worst-performing products across different shops. Considering the widespread poor performance of these products, they deserved consideration for removal from the assortment. Products flagged across multiple shops as consistently underperforming represented opportunities for optimization of the assortment to enhance overall revenue performance and save costs, aligning with one of the main goals of the analysis.

4.2. Association Rules

In the following section, the analysis of shop 1 for a single year, 2020, will be presented as an illustrative example of the methodology employed. The analysis was extended to the shops examined across their respective years, including shop 1, shop 3, shop 4, shop 5, shop 6, shop 7, shop 8, shop 9, shop 11, shop 12, and shop 13. Shop 2 and shop 10 have been excluded because of very specific behavior. Shop 14, 15, and 16 have been excluded due to the lack of enough data. They contain less than two months of recorded sales.

Notably, certain aspects of the project will not be extensively discussed due to the high level of granularity and the absence of domain-specific knowledge regarding the real-world information behind the product codes. While numerical patterns were the result of this analysis, the main goal was to translate these patterns into top-level business insights that can be valuable not just for bakery managers but for a broader audience as well.

The following analysis of the association rules began with grouping transactions by transaction ID and product code. Next, frequent itemsets were identified using a minimum support threshold of 0.001, meaning the items in these rules are bought together in at least 0.1% of all transactions. Subsequently, rules are generated with a minimum confidence level of 0.1, leading to the identification of 553 rules for shop 1 in the year 2020.

The analysis had three main stages: i) an examination of all rules; ii) a focus on non-redundant rules; iii) a selection of the top 30 rules based on confidence. This multi-step approach avoided overly restrictive filtering in one iteration.

The analytical approach employed graphical representations, particularly network graphs, to visualize the relationships between the items and rules. This approach followed the guidelines outlined in the paper “Visualizing Association Rules” (Hahsler & Chelluboina, 2011). While some parameters were set differently in the following analysis, it used the concept of the Fruchterman-Reingold layout, a force-directed algorithm that positions nodes in a way that minimizes edge crossings and evenly distributes nodes (Hahsler & Chelluboina, 2011). Although various visualization methods exist, the network graph-based approach was particularly well-suited due to the large number of rules considered in this analysis.

4.2.1. Association Rules (also Redundant Rules) Node Size Based on Indegree

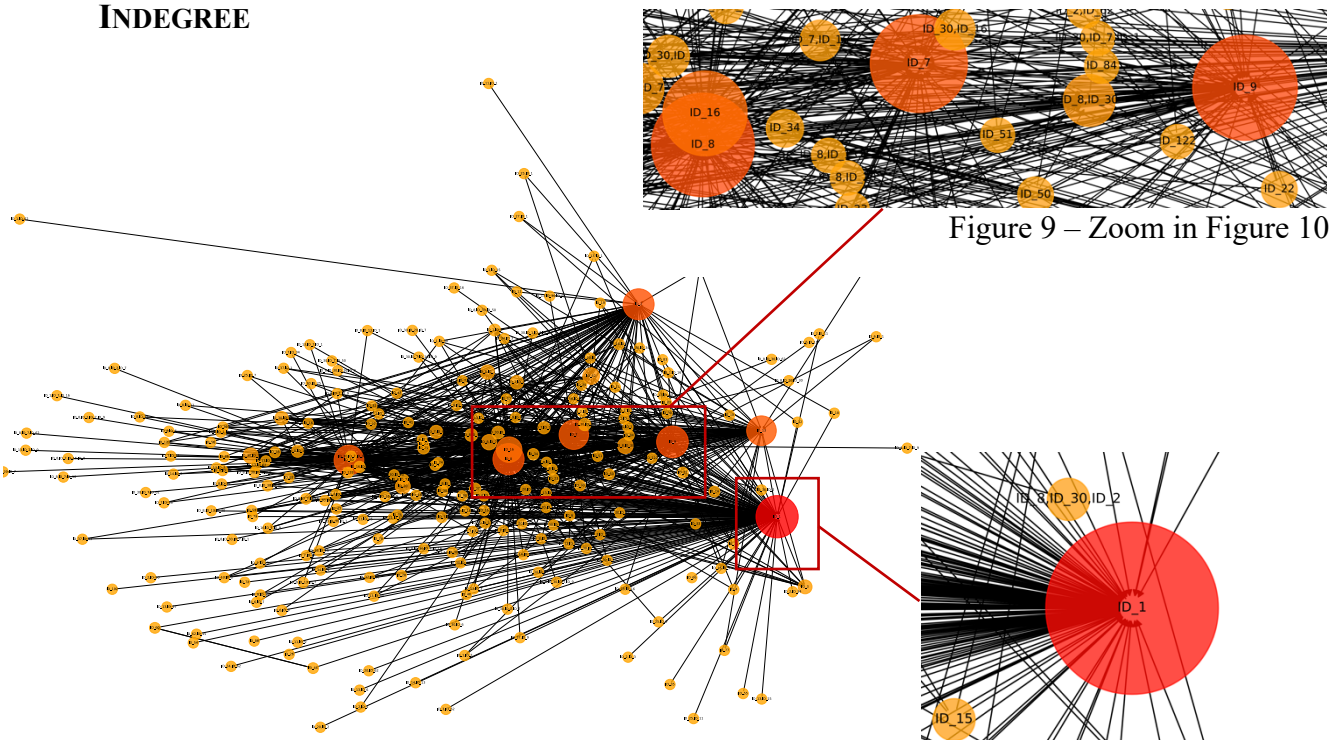

Figure 9 – Zoom in Figure 10

Figure 8 – Zoom in Figure 10

Figure 10 – Network Graph with all Rules, Node size based on Indegree

The purpose of Figure 10 – Network Graph with all Rules, Node size based on Indegree was to give an initial overview of the analysis. The node size in the network was determined by its indegree, which indicates the number of rules that have this item as a consequent. Essentially, the more rules that concluded with this item, the larger the node appeared. Figure 9 – Zoom in Figure 10 and Figure 8 – Zoom in Figure 10 are excerpts from Figure 10 – Network Graph with all Rules, Node size based on Indegree. As can be seen in Figure 8 – Zoom in Figure 10, the analysis highlighted the significance of the products ID 1, which appeared to be particularly interesting and dominant within the dataset.
Figure 9 – Zoom in Figure 10 shows a group of products ID 7, ID 8, ID 9, and ID 16 emerging as noteworthy. Furthermore, the products ID 13 and ID 30 also seem interesting after this first assessment. They gathered in the middle of the graph, building a cluster of bigger nodes. These products exhibited a certain level of importance through their node size and their proximity to each other, suggesting potential patterns or associations. It was advisable to maintain focus on these products as the analysis progressed, as they provided valuable insights into behavior and purchasing patterns. It is important not to emphasize products based solely on one metric; it is crucial to explore them through various lenses and different metrics for a comprehensive analysis.

4.2.2. RULES (ALSO REDUNDANT RULES) NODE SIZE BASED ON LIFT

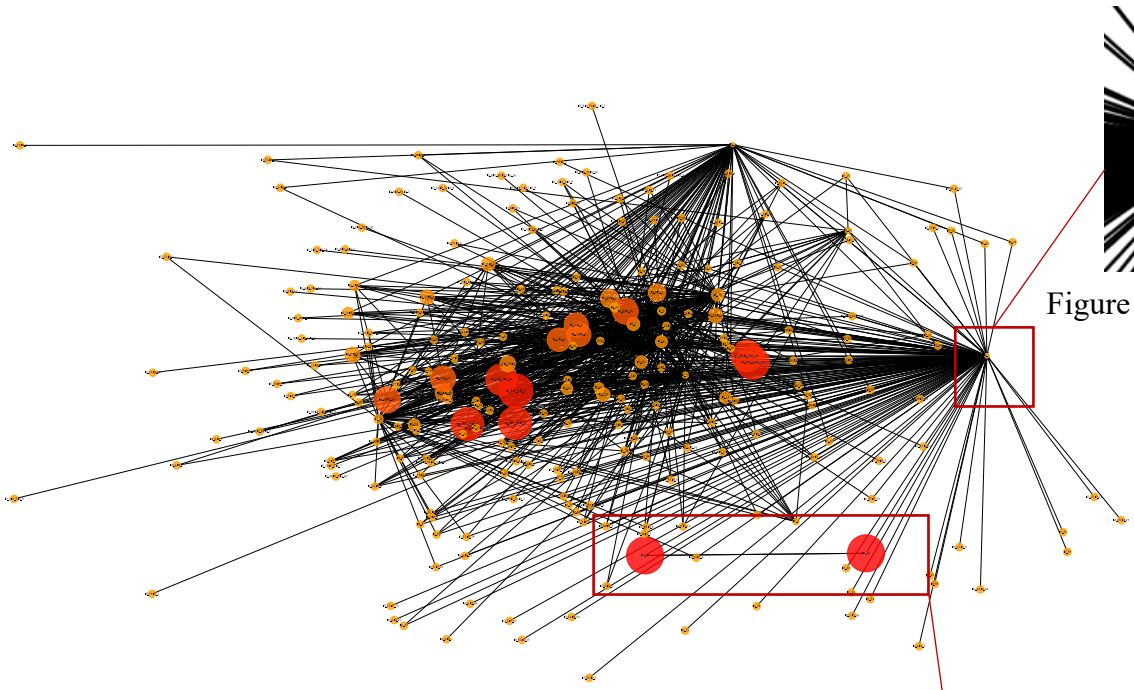

Figure 11 – Zoom in Figure 12

Figure 12 – Network Graph with all Rules, Node Size based on Lift

Figure 12 – Network Graph with all Rules, Node Size based on Lift shows a graph where the node size is based on the calculation of the average Lift for each rule. This provided valuable insights into the associations between products. Figure 11 – Zoom in Figure 12 illustrated the size of the node representing ID 1 when the node size is determined by the average lift. Notably, despite the frequent purchase of Product 1 as a consequent and its high indegree, the low average lift suggested a low exceptionality of the co-occurrence with any other product.

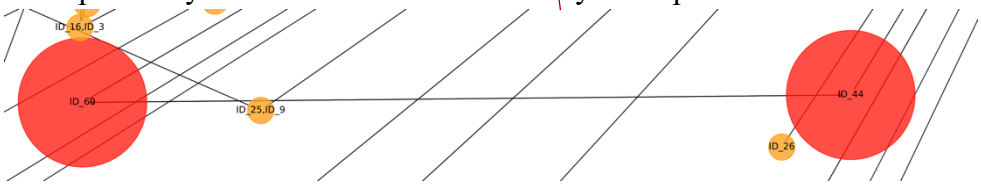

Figure 13 – Zoom in Figure 12

These findings suggested that while Products ID 1 was popular and frequently bought, it did not present optimal opportunities for promotion. Their associations with various items lacked distinct patterns or niches that could be leveraged for targeted marketing efforts. As such, other products with higher average lift values offered more promising business opportunities for promotion and strategic initiatives.

The central cloud of products with increased node size emphasized the importance of the previously identified items: ID 7, ID 8, ID 9, ID 13, ID 30, and ID 16. These products formed groups within the rule sets, appearing both as antecedents and consequents. While their specific roles within the rules were yet to be determined, their consistent presence suggested a strong association and potential for further exploration. Figure 13 – Zoom in Figure 12 revealed a crucial finding in the sequential visual analysis: the products ID 60 and ID 44 exhibited a notably high lift despite having low support. This indicates that, although these items were not frequently purchased individually, they appeared together in baskets significantly more often than would be expected if their purchases were independent.

A confidence level of 0.22 was rather low, but still high enough to be considered important, meaning that 22% of the transactions containing the one item, led to the purchase of the other. However, a high lift ratio between the two products of 47.63 in the context of rather low confidence was particularly intriguing as it illustrated a significant and interesting relationship.

To illustrate this relationship, let's use the example of coffee (antecedent) and cookies (consequent) again.
The confidence level between coffee and cookies indicates that when coffee is purchased, it leads to the subsequent purchase of cookies only 22% of the time. This is represented by Equation 5 – Confidence.

$$Confidence(X \Rightarrow Y) = \frac{0.001042}{Support(X)} = 0.22$$

Equation 5 – Confidence

However, the high lift value indicates that when cookies are bought with coffee, it occurs far more often than expected by chance. Lift measures the exceptionality of the co-occurrence of items.
Specifically, in this case, it is 47.63 times more likely that cookies and coffee will appear together than if their purchases were independent. This can be seen in Equation 6 – Lift.

$$Lift(X,Y) = \frac{\text{Support}(\text{x} \cup \text{y})}{\text{Support (y)} * \text{Support (x)}} = \frac{0.001042}{0.004654 * 0.004699} = 47$$

Equation 6 – Lift

The lift formula compares the observed co-occurrence of the items (represented by support of x ∪ y) to the occurrence of the consequent (Support y) and the occurrence of the antecedent (support x). Although the confidence was not very high (0.22), the occurrence of the cookie and the coffee were very low, resulting in this high lift.
Lift heavily favors low support values. When the supports of X and Y decrease, their impact on lift increases exponentially. This sensitivity arises because, statistically, it is highly improbable for two rarely occurring products to appear together by chance. Consequently, the unlikelihood of their joint occurrence grows exponentially as their individual supports diminish.
However, coffee leads to the purchase of a cookie in only 22% of cases, which does not make coffee a good predictor for the purchase of cookies. This means that lift indicates exceptionality rather than prediction of purchase, representing two distinct concepts. This distinction should prompt managers to ask questions. For example, if the products were typically bundled, the confidence should be higher. Therefore, the exact nature of the connection between these two products remains unclear, but it offers a valuable source for further business strategies and a significant learning opportunity. This discovery underscored the importance of conducting a multi-perspective analysis that considers multiple metrics. By incorporating lift as a complementary metric, valuable associations were identified that may not have been immediately apparent from traditional confidence-based analyses.
Redundancy occurs when a rule does not provide additional information beyond what other rules already convey. A function was deployed that evaluates each rule for redundancy by comparing it against the rest of the rules. It examined if there's another rule with the same consequent, equal or higher confidence, and a subset of the antecedents. If such a rule existed, the function identified that rule as redundant, allowing the analysis to proceed with a refined set of rules. Then, the network analysis based on lift and indegree was repeated. Through this step, the previously identified rules became even clearer. However, very few new discoveries emerged. Hence, this work continues with the introduction of a new metric to the analysis - conviction.

4.2.3. NON-REDUNDANT RULES: BASED ON CONVICTION

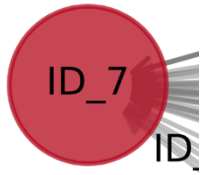

Figure 15 – Zoom in Figure 18

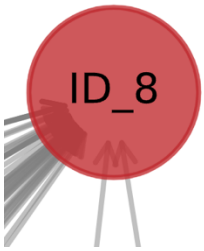

Figure 14 – Zoom in Figure 18

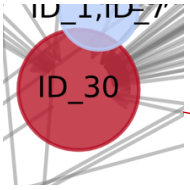

Figure 17 – Zoom in Figure 18

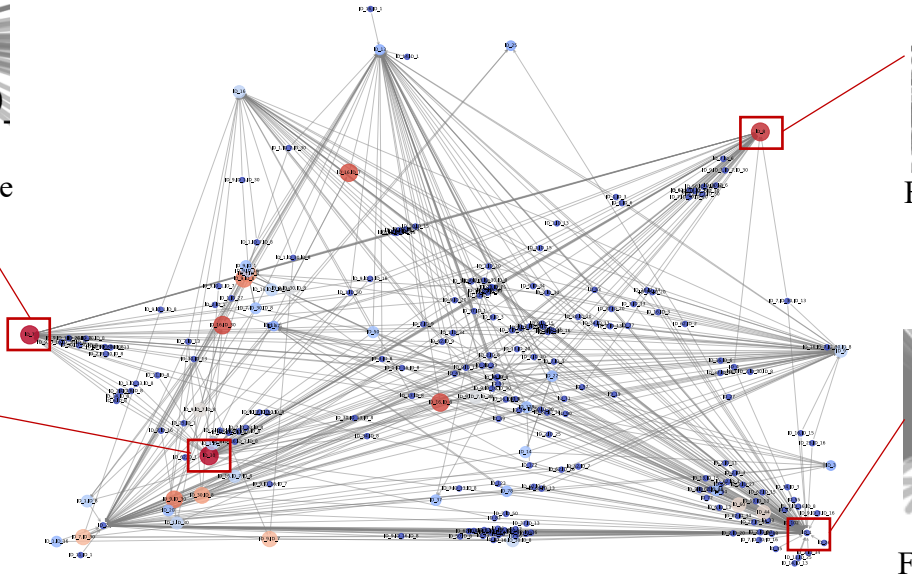

Figure 18 – Rules after Redundancy Elimination, Node Size based on Conviction

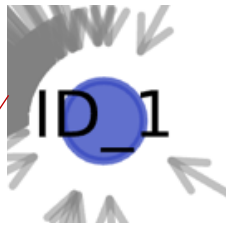

Figure 16 – Zoom in Figure 18

Figure 18 – Rules after Redundancy Elimination, Node Size based on Conviction displayed a graph where node size was based on average conviction as a consequent, introducing another crucial metric into the analysis. Moreover, nodes with a higher degree of redness indicate greater conviction values, while blue indicates low conviction values.

As mentioned earlier, conviction reflects the strength of the predictive power of the antecedent to consequent items in a rule. However, it penalizes very frequently occurring consequents and favors strong confidence. This is why product ID_1 was not regarded as interesting and appeared very small on the graph, which can be seen in Figure 16 – Zoom in Figure 18. Furthermore, conviction relates to support and confidence, but unlike lift, it is sensitive to the direction of the rule. Figure 14 – Zoom in Figure 18, Figure 15 – Zoom in Figure 18, and Figure 17 – Zoom in Figure 18, painted a clearer picture of what was visible before. Products ID 7, ID 8, and ID 30 exhibited a very high conviction, which aligned with earlier findings as they were also already exhibiting high confidence and increased lift values. However, it is interesting that product ID 44 and product ID 60 did not exhibit a very high conviction value, while they showed an extremely high lift.

The confidence level was rather low (0.22), and the support of both items was very low, around 0.00465. These low support levels made the random co-occurrence extremely unlikely und resulted in a high lift. Looking at the formula for conviction, it was evident that because of the rather low confidence, the conviction value was rather low, too. As can be seen in Equation 7 – Conviction, conviction takes into account the low support of y, but it heavily penalizes low confidence values.

$$Conviction(X \Rightarrow Y) = \frac{1 - Support(Y)}{1 - Confidence(X \Rightarrow Y)} = \frac{1 - 0.004654}{1 - 0.221} = 1.27885$$

Equation 7 – Conviction

Low conviction and confidence values imply that the products are not good predictors for the subsequent purchase of the consequent. However, there is the notion that longer rules, which

employ more conditions, generally exhibit stronger confidence as they are more niche and consequently lead to the consequent more often. In this context, the pattern identified is a one-to-one pattern, which typically has a lower confidence compared to longer rules. This suggests that the confidence and conviction values might be higher in this particular context than they initially appear. Given the strong anomaly presented by this one-to-one pattern within the data, it warrants further investigation to understand the underlying reason of this anomalie, offering a valuable opportunity for deeper insights and strategic development. This highlighted a nuanced decision point: a low conviction value might mislead an analyst into thinking there is a weak and not releveant association, whereas, in reality, there is a strong association. However, that association was statistical, highlighting the exceptionality of the co-occurrence of the two items, rather than providing one item with strong predictive power for the purchase of the other.

This case also demonstrated how the combined metrics of lift and conviction must be interpreted carefully to avoid incorrect conclusions.

In contrast, products ID 7, ID 8, ID 9, and ID 16, which have been previously identified as items of interest, displayed high indegree, lift, and conviction values. This signified that these items were frequently purchased as a result of other items being bought and were integral to certain shopping patterns. Each of these products appeared as antecedents and consequents of each other. The high lift indicated a strong likelihood of these items being bought together, while their high conviction underscored a strong dependency on preceding purchases.

In summary, these products not only frequently co-occured with other items but also strongly drove additional sales. Emphasizing customer orientation, these insights can guide strategic initiatives such as cross-selling strategies, promotions, and product placement optimizations. By aligning these concepts with an understanding of customer buying behaviors and preferences, the bakery can improve overall sales performance and customer engagement in the shops.

4.2.4. Top 30: Node Size Based on Indegree, Edge Width Based on Conviction

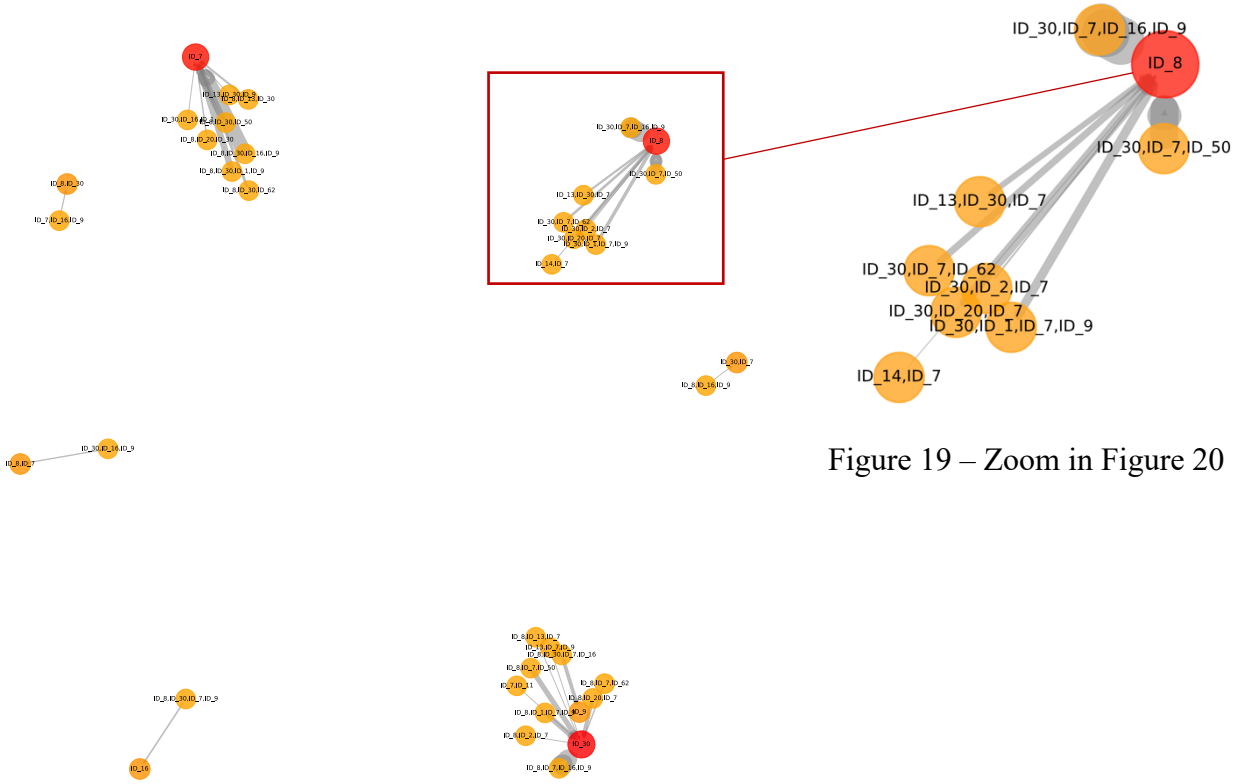

Figure 19 – Zoom in Figure 20

Figure 20 – Top 30 Rules, Node Size Based on Indegree, Edge Width based on Conviction

Figure 20 – Top 30 Rules, Node Size Based on Indegree, Edge Width based on Convictionfurther supported the importance of products ID 30, ID 8, and ID 7 as central products within the network. This visualization clarified this by emphasizing the importance of these key items based on their connectivity within the dataset.
However, it's notable that the filtering of the top 30 rules based on confidence resulted in the deletion of the rule involving ID 44 and ID 60, which exhibited a very high lift. While this rule may have been excluded due to lower confidence, its removal underscored the trade-off between confidence and lift in rule selection. As seen in Figure 19 – Zoom in Figure 20, the rules having ID 8 as a consequent showed different conviction values, which measured the likelihood that the purchase of ID 8 was directly influenced by the presence of these item combinations. Table 5 – Outstanding Rules Shop 1 showed the two rules that stood out the most.

Antecedent	Consequent
(ID 30, ID 7, ID 16, ID 9)	ID_8
(ID 30, ID 7, ID 50)	ID_8

Table 5 – Outstanding Rules Shop 1

Therefore, these findings implied a strategic opportunity for businesses to leverage these dependencies by strategically promoting or bundling ID 8 with item groups that exhibited the highest conviction values. By capitalizing on these strong associations, businesses can enhance

sales performance and customer satisfaction through targeted marketing efforts and optimized product offerings.

4.2.5. FINAL STEP, MONETARY RULE ANALYSIS

In the next step of the analysis, the total price for each rule was analyzed. Given the absence of information about margins, two assumptions were made. First, rules with a higher total unit price could be top revenue rules. Secondly, rules with a higher total sales value in that year could be top revenue rules. The total sales value was determined by exact matching of how often that rule was bought in the exact way it is depicted. However, it is clear that without knowing the costs of these products, this part of the analysis can only provide a rough estimate. Table 6 shows the top rules sorted by total unit price. This approach allowed for the prioritization of rules based on their potential revenue contribution, providing valuable insights for resource allocation and strategic decision-making. By focusing on rules with higher total unit prices, marketing efforts, product placements, and promotions for these products can be prioritized by businesses to maximize revenue generation and profitability.

Antecedents	Consequents	Total Unit Price
ID 7, ID 11	ID 30	8.34
ID 1, ID 9, ID 8, ID 30	ID 7	6.15
ID 1, ID 7, ID 9, ID 9	ID 30	6.15
ID 1, ID 7, ID 9, ID 30	ID 8	6.15
ID 2, ID 7, ID 8	ID 30	6.04
ID 2, ID 7, ID 30	ID 8	6.04

Table 6 – Top Monetary Rules by Unit Price, Shop 1 2020

The unique prices for ID 44 and ID 60 can be seen in Table 7 – Unit Price of ID 44 and ID 60. The analysis revealed that this rule possessed an equally high value of around 8€, presenting a promising business opportunity due to its strong lift. This pattern consistently occurred across all shops and years with different pairs of products, underscoring the importance of this pattern.

Product code	Unit price
ID 44	4.06
ID 60	4.06

Table 7 – Unit Price of ID 44 and ID 60

Another approach can be seen in Table 8 – Top Monetary Rules by Total Sales Value, Shop 1 2020. The idea was to filter out all the sales generated by the top rules, by filtering out instances where customers bought exactly this combination of products. By employing a strict matching approach, the exact sales value attributed to this rule for the year were determined. This calculation provided a precise monetary measure of the rule's contribution to revenue generation, enabling comparisons across different rules. It is important to note that this calculation is highly shop-specific.

Antecedents	Consequents	Total Sales Value
(ID_7, ID_30, ID_13)	ID 8	2670.58
(ID_9, ID_16, ID_7, ID_30)	ID 8	2631.42
(ID_2, ID_7, ID_30)	ID 8	2580.50
(ID_50, ID_7, ID_30)	ID 8	2225.84
(ID_1, ID_30, ID_16)	ID 7	2189.87
(ID_14, ID_7)	ID 8	2110.69
(ID_62, ID_30, ID_8)	ID 7	1938.02
(ID_9, ID_1, ID_30, ID_8)	ID 7	1784.48

Table 8 – Top Monetary Rules by Total Sales Value, Shop 1 2020

This selection highlights the rules recommended for further investment in shop 1, distinguished by their robust lift, high confidence, and the highest total sales value. As illustrated in Table 8 – Top Monetary Rules by Total Sales Value, Shop 1 2020, there are only two distinct consequents: products ID_8 and ID_7. Promoting either of these products could potentially yield the highest monetary results, as the rules with these products as consequents achieved the highest total sales values in 2020. It is important to note the limitation that the costs associated with these products are not known, and thus the analysis only reflects revenue values, not profit margins. Additional findings suggest that rules involving ID_7 and ID_8 consistently appear with high sales values, indicating strong customer demand. The frequent occurrence of common antecedents, such as ID_7 and ID_30, across multiple rules suggests these combinations are popular among customers. This highlights the potential for bundled promotions and cross-selling strategies. Furthermore, maintaining adequate inventory of these key items would be essential for meeting customer demand and maximizing sales.[2]

4.2.6. COMMENTARY ON THE REMAINING ANALYSIS

This process was repeated 25 times, covering all mentioned shops and their respective years. By diving into one shop for a specific year, valuable insights into the data sources, rule filtering mechanisms, and the underlying patterns of shop behavior were gained. Through this in-depth analysis, recurring patterns were identified, such as the presence of unique high-lift rules between products like ID 44 and ID 60 across multiple shops.

However, many more patterns can be found for each shop for each year, from which business strategies can be derived and additional insights generated. Recognizing the importance of these patterns, it became clear that each shop required its own in-depth analysis to understand product purchasing behaviors. Additionally, deeper analysis involving more variables, such as product family and subfamily, could be beneficial.

While conducting these individual analyses for each shop and year provided detailed insights into specific behaviors, the high granularity and lack of real-world context behind product codes limited their usefulness. Furthermore, this level of detail was regarded as too extensive for the scope of this work. Instead, this work aimed to gain a top-level understanding of trends across all shops and years. The next section, the Meta-Analysis, tackled the identification of overarching patterns, common trends, and key insights that transcend individual shop analyses, providing a comprehensive view of the data and enabling informed decision-making at a broader scale. Finally, also the findings from the Association Rule Mining analysis can be presented in Table 9 – Association Rule Findings.

[2] It should be noted that this analysis should be repeated if a margin variable were available.

Finding 1	**Product Availability and Performance**: There was an increase in the total number of different products sold each month, but product availability varied across different shops and time periods. Not all products were consistently available, and some products showed potential for rebranding and reintroduction based on historical performance.
Finding 2	**Node Size Based on Lift**: Products ID 1, despite being frequently purchased, showed very low lift values, indicating weak situational associations and limited potential for targeted promotions.
Finding 3	**Identification of Key Products**: Certain products, such as ID 7, ID 8, ID 9, ID 13, ID 16, and ID 30, consistently showed high importance across various metrics (indegree, confidence, lift, and conviction). These products were frequently bought together and drove additional sales, making them central to strategic initiatives like cross-selling and promotions.
Finding 4	**High-Lift Rules with Low Confidence**: The rule involving products ID 60 and ID 44 exhibited a very high lift despite low confidence. This indicated a strong, though infrequent, statistical association, suggesting significant potential for targeted marketing despite their low occurrence.
Finding 5	**Low Conviction doesn't mean low Association:** Products ID 44 and ID 60 had a strong lift but low conviction, indicating that even though conviction and confidence both suggested this rule is not highly relevant, lift emphasized the exceptionality of this co-occurrence. This highlights the need for further investigation, as it is intriguing how something can statistically occur so much more often than expected, yet still be almost irrelevant for conviction.
Finding 6	**Confidence vs. Lift Trade-Off**: The exclusion of the rule involving ID 44 and ID 60, due to lower confidence despite a high lift, highlighted the trade-off between using confidence and lift in rule selection. This pattern like ID 44 and ID 60 was consistently found across different shops and years, underscoring its importance.
Finding 7	**Multi-Perspective Analysis**: The analysis underscored the importance of considering multiple metrics, such as lift, confidence, and conviction, to identify valuable associations that might not be apparent through confidence-based analyses alone.

Table 9 – Association Rule Findings

4.3. META-ANALYSIS

The Meta-Analysis aimed to identify broad patterns across all shops, highlighting similarities and anomalies to facilitate decision-making across multiple shops simultaneously. Furthermore, the first year, 2020, of shop 1 was deleted since it exhibited very different behavior. The Meta-Analysis focused on whether the more recent data from all shops, including shop 1, showed similarities to the other shops.

The Meta-Analysis started with exploring metrics of the resulting rules of all the shops and years by visualizing them in heatmaps. Then, a scoring system was deployed, to evaluate the top 5 global rules across all years and all shops. Furthermore, the development of the rules over time was analyzed and the seasonal metric change was being tested. The section continued with hierarchical and k-means clustering and concluded with a high potential, but low sales analysis. The whole Meta-Analysis was done two times, once with the top 30 rules per shop per year sorted by confidence and once sorted by lift. The repetition based on lift did support most of

the findings of the confidence-based Meta-Analysis. However, the clustering process had slightly different results, which was expected.

4.3.1. METRIC COMPARISON (ALL SHOPS ALL YEARS)

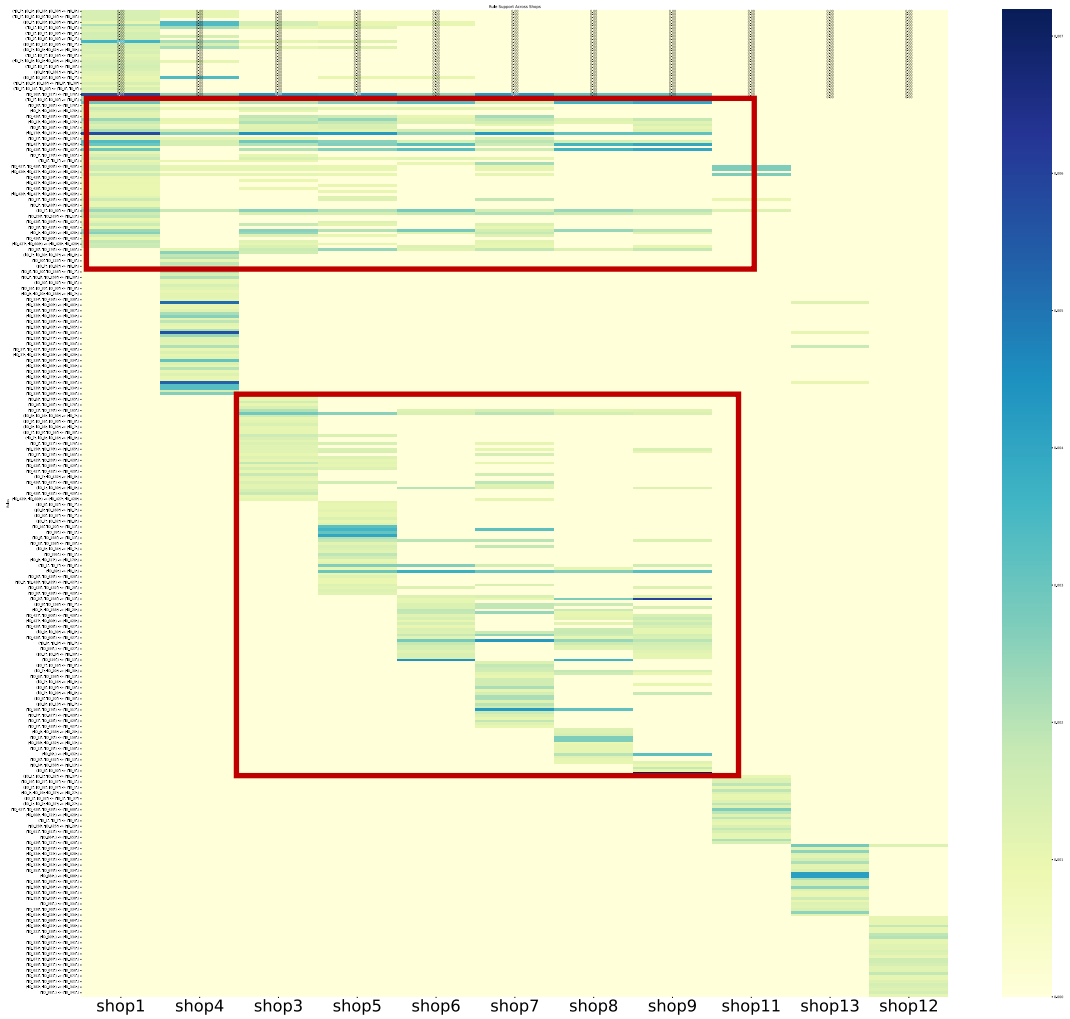

Figure 21 – Heatmap Support (All Shops All Years)

A heatmap was created that displayed the average value based on the respective metric, visually depicting variations and trends. The most important insight from this heatmap is the initial insight into the similarities among the shops.
Analyzing the heatmap based on support across all shops with the X-axis showing the shops and the Y-axis representing the Association Rules, two key areas emerged, as can be seen in Figure 21 – Heatmap Support (All Shops All Years). In the top red square, the same rules appear to be important, creating a darker area in their cells for shops 1 through 9. In the bottom red square, shop 3 to shop 9 displayed similar patterns.
This suggested that shops that showed similar sales patterns in the beginning of the analysis shared similar association rules. Additionally, certain rules consistently exhibited strong support levels across multiple shops and years, indicating important revenue generation opportunities for the company. Furthermore, the shops 11, 12, and 13 showed almost no shared support for the rules across the three shops.
A similar pattern observed in the Support Heatmap was also evident in the heatmaps for Confidence and Lift. They can be seen in the appendix, Figure 31 – Heatmap Confidence (All Shops All Years) and Figure 32 – Heatmap Lift (All Shops All Years).

4.3.2. TOP 5 RULES

The next part of the analysis evaluated the rules according to each metric across all years and shops to identify the top 10 rules for lift, confidence, and support. Scores were assigned based on their position in the ranking, and then the rankings were added up together. This process was used to determine the top 5 globally best rules. The highest-ranked rule in each metric received ten points, the second received nine points, and so forth. Consequently, the maximum possible total score for a rule was 30 points. This methodology allowed for identifying the most important rules overall across all shops and years, based on these three metrics.
Table 10 displays the top 5 rules resulting from this analysis. This analysis addressed RQ2 to some extent, which focused on identifying the most important association rules in the bakery's sales data.

RQ2: What are the most important association rules in the bakery sales data?

To add relevance, the same analysis was done again with rules generated from sales data only from 2023. The results were almost identical, as can be seen in Table 11 – Top 5 Global Rules, 2023. Reflecting the concept of the most important rules, this work could not pinpoint specific rules. The top 5 rules showed the highest scores overall, but this did not necessarily mean, that they were the most important ones, simply because they exhibited high lift, confidence and support. It could be argued that the most important rules should be the most profitable ones. However, this would need to be determined specifically for each shop and can be found in the monetary analysis of each shop, with the limitation that there was no margin variable and it was determined based on price and total value.
Despite this, these five rules carry importance for the bakery as a whole, as they had the highest total scores across all shops and all years. When defining importance through high values in the metrics, these five rules were the answer and can be used for communication for the whole bakery or the brand image of the entire bakery.

Rule	Score Lift	Score Support	Score Confidence	Total Score
ID 427, ID 429 => ID 428	9	8	9	26
ID 165, ID 317 => ID 176	4	10	10	24
ID 427, ID 428 => ID 429	10	7	0	17
ID 176, ID 317 => ID 165	5	9	3	17
ID 428, ID 429 => ID 427	6	6	0	12

Table 10 – Top 5 Global Rules

Rule	Score Lift	Score Support	Score Confidence	Total Score
ID 427, ID 429 => ID 428	9	8	9	26
ID 165, ID 317 => ID 176	5	10	10	25
ID 176, ID 317 => ID 165	7	9	5	21
ID 427, ID 428 => ID 429	10	7	0	17
ID 428, ID 429 => ID 427	8	6	0	14

Table 11 – Top 5 Global Rules, 2023

4.3.3. RULES CHANGING OVER TIME

The following part of the project analyzed the development of the rules over time. This is the second part of the answer of RQ3:

RQ3: What are the most important temporal variations?

Figure 22 – Support Rate Changes over Time, Figure 23 – Lift Rate Changes over Time, and Figure 24 – Confidence Rate Changes Over Time show the most important temporal evolution, unveiling trends. The graphs show the change rate between the first and the last year. Also, a filter was deployed, which only allowed rules to be depicted, that changed by more than 10%.

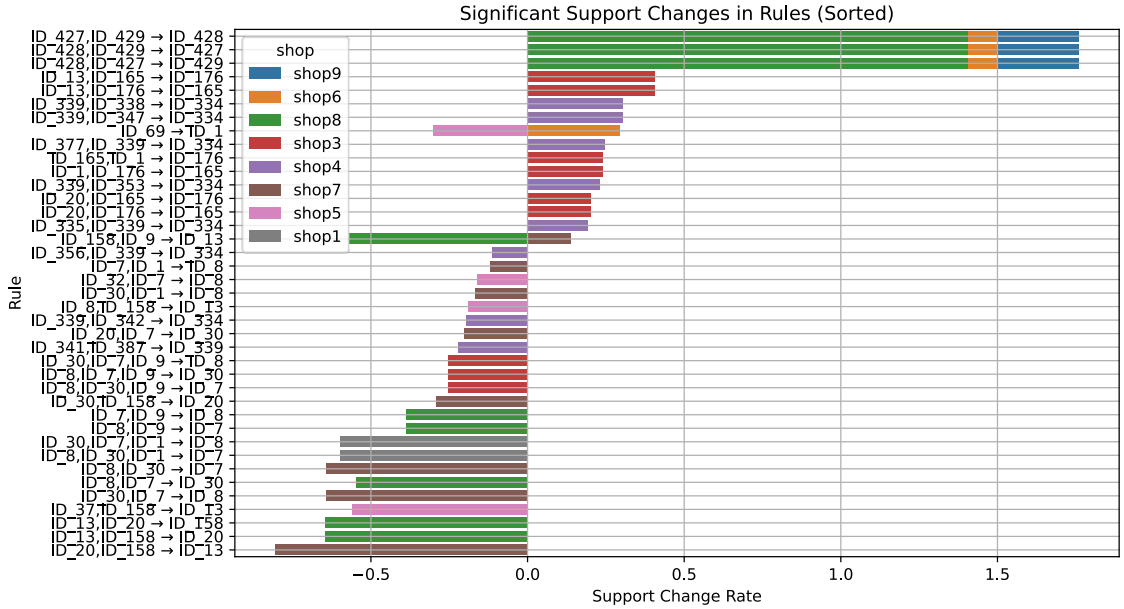

Figure 22 – Support Rate Changes over Time

Figure 22 – Support Rate Changes over Time revealed a significant increase in support for the rule involving products 427, 429, and 428, which was determined to be one of the most important rules. However, Figure 23 – Lift Rate Changes over Time illustrated a strong decline in lift for the same rule.

This shift suggested a change in customer purchasing patterns, possibly indicating a shift in consumer preferences or marketing strategies. The rule did gain more popularity over time, hence its rise in support. However, it is possible that the items were being bought more frequently individually or with other items, hence the drop in lift.

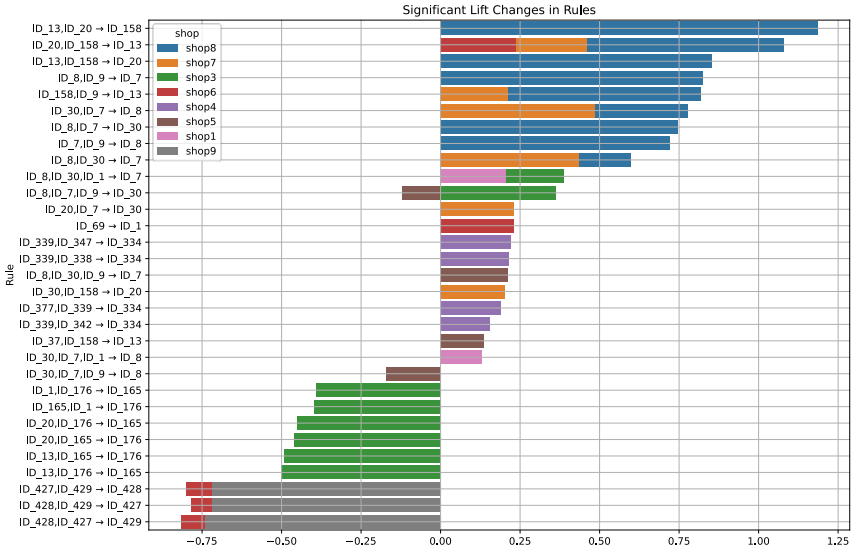

Figure 23 – Lift Rate Changes over Time

Further analysis and knowledge of the products behind the product codes would be needed to uncover the underlying reasons behind these changes and their implications for business strategies. However, since these were among the most important rules, the decline in lift seems very important.

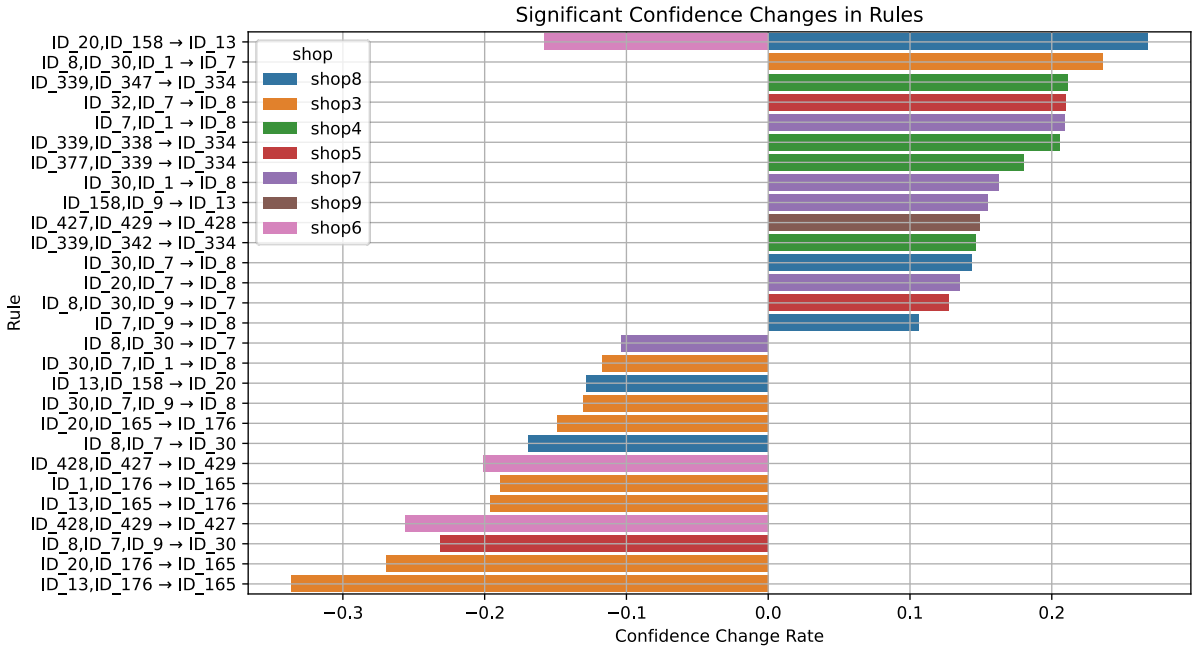

Figure 24 – Confidence Rate Changes Over Time

Looking at Figure 24 – Confidence Rate Changes Over Time, where the changes in confidence levels over time were visualized, it became clear which rules show stable and positive confidence levels. Such stability in confidence suggested that these rules can be reliably utilized when the bakery would opt for a risk-free strategy in the future.
Another very interesting observation was that shop 6, had a strong decline in confidence in the top performing rule for shop 8. If the assumption was correct that these two shops could be clustered together, there would be transferable knowledge between them based on their very different behavior for rule (ID 158, ID 20 => ID 13) based on confidence.

4.3.4. Seasonal Metric Change

An additional concept was the variation of metrics throughout the seasons. The following analysis addresses hypotheses H2a, H2b, and H2c.

H2a: There are significant variations in Support across seasons.

H2b: There are significant variations in Confidence across seasons.

H2c: There are significant variations in Lift across seasons.

Metric	Test Statistics	P-value	Reject
Support	21.64	0.000077	True
Confidence	1.304615	0.728	False
Lift	16.998	0.000707	True

Table 12 – Kruskal Wallis Test Hypothesis 2

First, a Kolmogorov-Smirnov normality test was performed and the data did not show a normal distribution. Afterwards, the rules were grouped by season, similar to the approach in the first hypothesis and the Kruskal-Wallis test was used. Instead of comparing total values, the respective metrics were analyzed season by season. The results are depicted in Table 12 – Kruskal Wallis Test Hypothesis 2. Interestingly, while confidence did not show significant differences between the seasons, significant variations were found for H2a and H2c, indicating that support and lift vary significantly across the seasons. The rejection of the null hypotheses for H2a and H2c underscored the importance of these seasonal changes. Although confidence did not show significant differences for all years and all shops, the variations in lift and support across seasons had strong managerial implications. The importance of the findings lied in their potential applications for strategic and operational improvements. Businesses can leverage the seasonal variations in support and lift to design targeted promotions, optimize inventory management, mitigate risk, and enhance product placement.

Furthermore, the Dunn test for H2c showed, that winter shows a significant rise in lift values compared to summer ($p = 0.0081$) and a higher lift value compared to spring ($p = 0.003325$). The average lift in winter was 39.12, while it was only 14.28 in spring and 15.22 in summer. This demonstrated a significant difference and had managerial implications. However, there was no significance found in the comparison between autumn and winter.

Furthermore, the increase in lift during winter, indicating stronger product associations, provided valuable insights for strategic planning. This suggested that targeted promotions and tailored marketing campaigns were particularly effective in winter, leading customers to purchase more bundled items. The finding aligned with the notion of increased basket size during winter, as evidenced by the higher lift values observed in this season. The observed trend highlighted an affinity for bundles during the winter months, raising the question of whether bundling inherently works better in winter or represents a missed opportunity in other seasons that can be further leveraged. Exploring this potential could uncover significant opportunities for enhancing sales throughout the year.

Interestingly, the Dunn test for H2a showed that winter exhibits a significant decrease in support values compared to summer ($p = 0.0083$) and a lower support value compared to spring ($p = 0.000135$). The average support in winter was 0.003238, while it was 0.020490 in spring and 0.042274 in summer. However, there was no significance found in the comparison between autumn and winter.

The pattern found for support is exactly the opposite as the pattern for lift. The decrease in support indicated that products were present in a smaller proportion of baskets relative to the total number of baskets, suggesting a decrease in basket size during winter. An explanation for this phenomenon could be, that there are more products available in winter, which led to an overall smaller support. In winter (640) there were indeed almost 30% more products available than in summer (506). This indicates that during winter, there is a broader variety of products that appear in bundled transactions. Management needs to be pay extra attention to the stocking of these bundles, as they seem to drive revenue. The increased product variety, coupled with stronger demand, contributes substantially to overall revenue. Thus, careful inventory management and strategic planning are essential to capitalize on this seasonal trend.

Finally, the findings regarding RQ3 can be presented in Table 13 – Temporal Variation Findings.

RQ3: What are the most important temporal variations?

Finding 1	**Pastry sales**: were highest in winter (over 35%), followed by fall (25%), spring (22%), and summer (18%), indicating that pastries were better sold during winter and spring, while non-pastries were favored during summer and fall.
Finding 2	**Seasonal Sales Variations**: Winter exhibits the highest average total value, significantly outperforming spring, summer, and autumn, indicating it is the most profitable season.
Finding 3	**Weekly Sales Variations**: Saturday has the highest average total value compared to other weekdays, making it the most profitable day of the week.
Finding 4	**Monthly Sales Variations in 2023**: December stands out with the highest total value among all months in 2023, significantly higher than other months.
Finding 5	**Support Rate Changes Over Time**: The rule involving products 427, 429, and 428 showed a significant increase in support, indicating rising popularity over time.
Finding 6	**Lift Rate Changes Over Time**: The same rule (products 427, 429, and 428) experienced a strong decline in lift, suggesting a shift in customer purchasing patterns.
Finding 7	**Confidence Rate Stability**: Certain rules maintained stable and positive confidence levels over time, suggesting their reliability for risk-free strategies.
Finding 8	**Decline in Confidence for Specific Shops**: Shop 6 showed a strong decline in confidence for a top-performing rule in shop 8, indicating different behaviors and potential areas for knowledge transfer.
Finding 9	**Seasonal Variations in Support and Lift**: Significant variations in support and lift across seasons were identified, highlighting the need for season-specific strategies for promotion, forecasting demand, and inventory.

Table 13 – Temporal Variation Findings

4.3.5. K-MEANS CLUSTERING

To understand if and how knowledge can be transferred among the shops, the next section of the analysis aimed to identify similarities between different shops. This would enable the grouping of shops with comparable purchasing patterns. To segment the shops, a cluster analysis was employed, using the k-means algorithm. K-means clustering is an unsupervised learning, which is fast, robust, and straightforward. The algorithm delivers reliable results when datasets are well-separated (Tabianan et al., 2022). K-means assigns data points to clusters by minimizing the squared distance between the data points and the cluster centroid. The K-Means algorithm was selected due to its efficiency, scalability and effectiveness in grouping data into clusters with minimized within-cluster variance (Hoque, 2024).

The feature matrix for the clustering was created by first constructing a dictionary containing the top 30 rules based on confidence from all years and all shops. The DataFrames inside the dictionary included the shop names, the rules, and their corresponding confidence values. This dictionary was then expanded into one big DataFrame. A pivot table was generated, where the rows represented shops, the columns represented rules, and the values were the confidence levels for each rule in each shop. Finally, this pivot table, called the feature matrix, was normalized so that all values fell between 0 and 1.

The number of clusters (k) was determined using the elbow method, which identifies a point where adding more clusters do not significantly improve the within-cluster sum of squares (SSE) (Patil et al., 2021). This analysis revealed that 4 clusters were optimal. Furthermore, the K-means algorithm used the Euclidean distance.

The question was raised as to whether confidence was the best metric to use in the pivot table to perform clustering. Lift could be well suited as well. As mentioned earlier, the entire Meta-Analysis was performed twice, with the second iteration utilizing the top 30 rules based on lift for each shop and year. Clustering was also executed using a feature matrix containing the lift values for each shop and each rule. However, lift is particularly effective for identifying niche relationships, associations, and strong dependencies between items. Confidence on the other hand reveals more general patterns and doesn’t penalize frequent consequent items (add-on items) that appeared frequently also in various baskets. In this specific case of comparing the shops, a top-level approach was preferred to identify general similarities in the patterns. Consequently, confidence was chosen for the analysis, as it effectively highlights these broader trends across different shops.

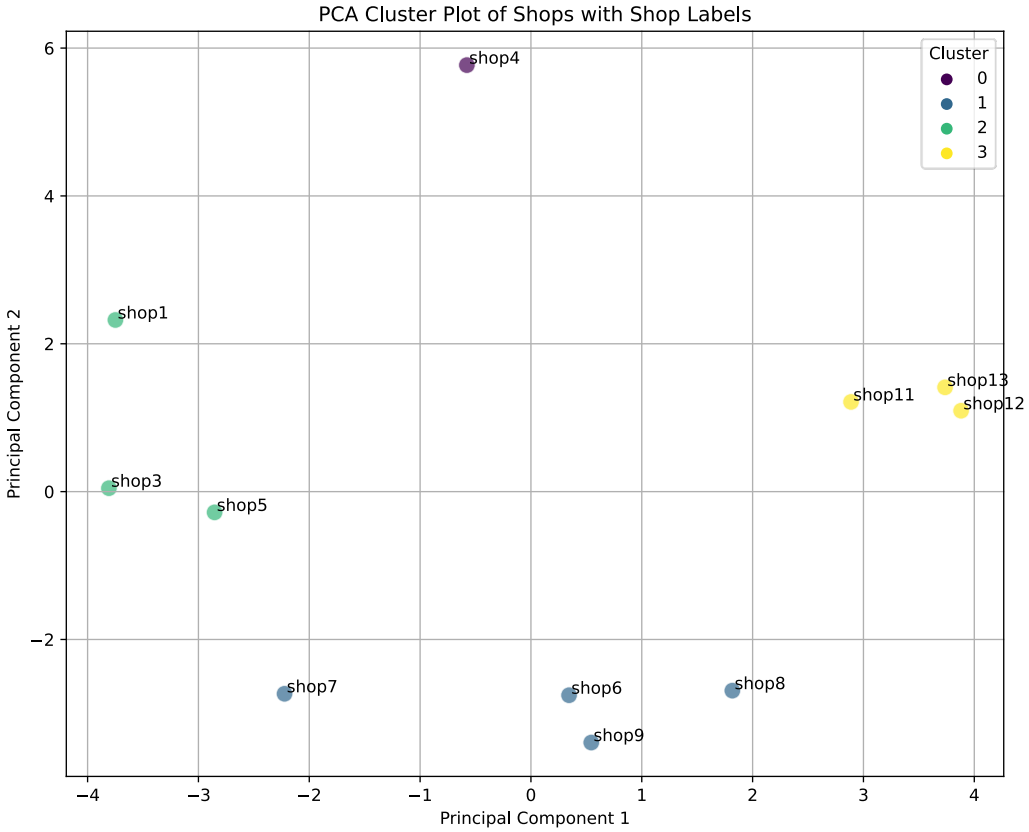

Figure 25 – PCA Cluster Plot with Shop Labels

Figure 25 – PCA Cluster Plot with Shop Labels illustrated a coordinate system where the axes were created using the Principal Component Analysis (PCA) with 2 components. The PCA was used to reduce the data complexity to two dimensions, allowing for the visualization of clusters on a coordinate system using a scatter plot. The results indicated that shops 1, 3, and 5 exhibited similarities, clustering closely together. Furthermore, shops 6, 7, 8, and 9 formed another distinct cluster, while shops 11, 12, and 13 also showed similar patterns and grouped into the same cluster. Conversely, shop 4 appeared to be distinct from the other clusters, indicating unique characteristics. This clustering analysis, visualized through a scatter plot with PCA components, highlights the relationships and similarities between different shops, providing valuable insights for strategic decision-making and operational improvements.

4.3.6. HIERARCHICAL CLUSTERS

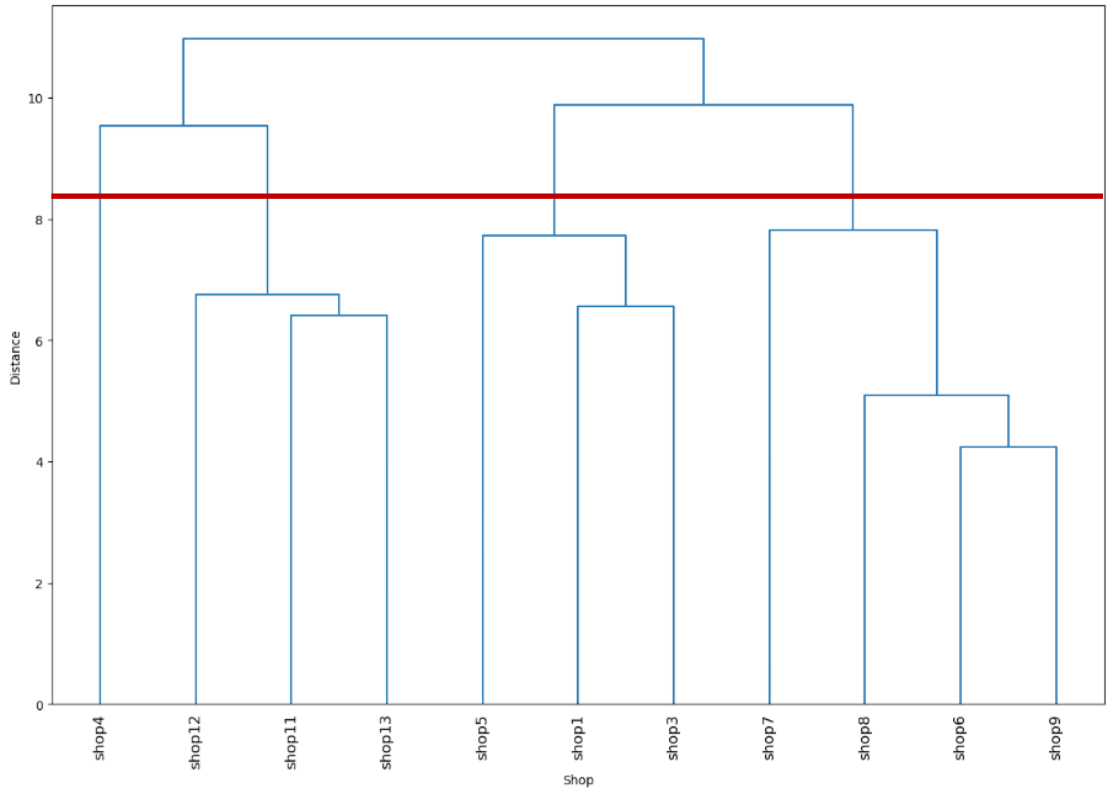

Figure 26 – Hierarchical Clustering Dendrogram

Figure 26 – Hierarchical Clustering Dendrogram illustrated the hierarchical agglomerative clustering process, where the algorithm initiated with 11 distinct clusters, each representing an individual data point. Subsequently, the algorithm evaluated the similarity between these clusters and progressively merged them, ultimately forming larger clusters based on their degree of similarity (Ran et al., 2023). This iterative "bottom-up" process continued until only one cluster remained, effectively grouping all entities together. A dendrogram is commonly used to display the clusters (Shetty & Singh, 2021). In this hierarchical clustering process, the distance, represented on the y-axis of the dendrogram, was based on a measure of similarity or dissimilarity, with the Ward distance chosen. The greater the distance between merged clusters along the y-axis, the more distinct the newly joined clusters were. Thus, the height at which clusters were merged in the dendrogram reflected the level of dissimilarity between them, with larger vertical distances indicating greater differences between the clusters (Ran et al., 2023).
Interpreting the dendrogram revealed that the most significant gap after the first iteration of joining appeared around the y-value between 8 and 9, where the red line was drawn. This suggested that an optimal cut at this point supported the decision to form four distinct clusters. Additionally, it was noteworthy that shop 5 and shop 7 joined the clusters at a later stage, which implied that these shops exhibited unique characteristics and should be approached with extra caution when making cluster-wide business decisions.

4.3.7. NETWORK OF SHOPS (COLORED BY K-MEANS CLUSTERS)

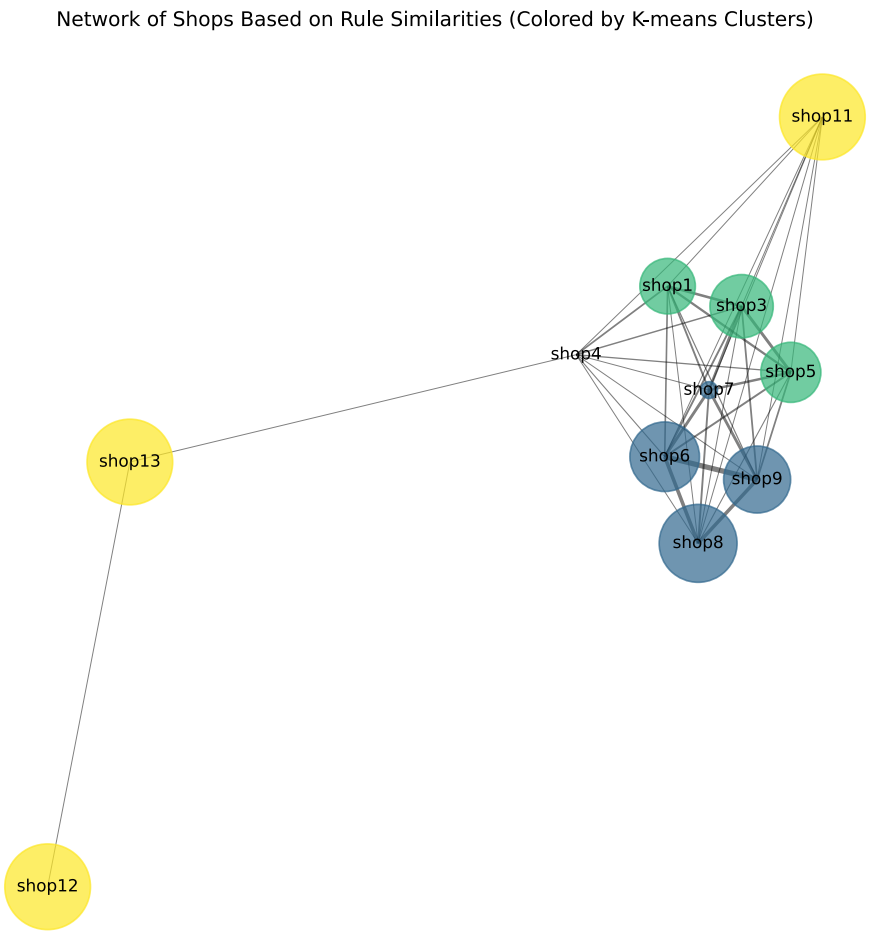

Figure 27 – Network of Shops Based on Rule Similarities (colored by Clusters)

Finally, the plot depicted Figure 27 – Network of Shops Based on Rule Similarities (colored by Clusters) displays a network graph, showing the relation between the shops.
The thickness of lines represented similarity between shops, based on the Jaccard similarity. Jaccard similarity is defined by the number of items that two sets have in common divided by the number of unique items in both sets (Hahsler & Chelluboina, 2011). The observation that shops 11, 12, and 13 exhibit only a minimal connection to each other, in contrast to the stronger similarities among the other shops, is consistent with earlier findings. To further analyze the similarities and differences, an examination of the node sizes provides additional insights.
The node size was calculated by the formular shown in Equation 8 - Unique Rules Proportion. This calculation involved subtracting the set of rules of a specific shop from the total set of rules in the entire cluster. The result was then divided by the total number of rules of the shop.

$$Node\ Size = Unique\ Rules\ Proportion = \frac{Len((all\ rules\ in\ cluster) - (set(shop\ rules[shop]))}{Len((all\ rules\ in\ cluster)}$$

Equation 8 - Unique Rules Proportion

Shop ID	Unique Rules Proportion
Shop 1	0.45
Shop 3	0.51
Shop 5	0.49
Shop 6	0.55
Shop 7	0.33
Shop 8	0.60
Shop 9	0.53
Shop 11	0.66
Shop 12	0.66
Shop 13	0.66
Shop 4	0.0

Table 14 – Results Unique Rules Proportion

By dividing the number of rules in the cluster that are not present in the current shop by the total number of rules in the cluster, the result was a proportion that represents how many of the cluster's rules were not found in the current shop. A value of 0.5 means that 50% of the rules in the cluster were not present in the current shop, indicating that half of the cluster's rules are unique to the other shops. A value of 0 would mean that all of the cluster's rules are present in the current shop, while a value of 1 would mean that none of the cluster's rules are present in the current shop. Shop 4 was in a cluster by itself, hence its value was 0. The bigger the nodes were, the more different they were from the rest of the cluster. The results can be seen in Table 14 – Results Unique Rules Proportion.

It was very interesting to note that Shop 7 had a very small node, with a unique rules proportion of 0.33. This means that only 33% of the rules in the cluster were not present in Shop 7, indicating that it did not have a very strong individual character. In contrast, Shop 8 held a value of 0.6, making it quite different from the rest of the cluster. An intriguing pattern emerged in this analysis. Shop 7 exhibited a high degree of rule overlap, with more than 60% of the cluster's total rules being present within this shop. Conversely, the remaining shops in the cluster displayed a more distinct set of rules. This raised the question of whether Shop 7 possessed a significant number of unique rules while also sharing a substantial portion of the cluster's rules. This dual characteristic made Shop 7 appear highly similar to other shops in the cluster, while simultaneously contributing to the overall heterogeneity of the cluster.

To find that out, a second time the “Unique Rules Proportions” were calculated, this time excluding Shop 7.

Shop ID	Unique Rules Proportion
Shop 6	0.33
Shop 8	0.41
Shop 9	0.30

Table 15 – Results Unique Rules Proportion (without Shop 7)

Table 15 – Results Unique Rules Proportion (without Shop 7) showed the results from this analysis. Indeed, Shop 7 demonstrated this paradoxical behavior. This indicated that patterns observed in other shops were likely to appear in Shop 7 due to its similarity to the others. However, the other shops were more similar to each other than to Shop 7, making Shop 7 an unreliable indicator for the rest of the cluster. It was not surprising that the differences between the shops decreased, when a shop was removed from the cluster. However, removing any of the other shops resulted in a maximum change in the Unique Rules Proportion of only 0.07, whereas removing Shop 7 led to a change of up to 0.23.
Interestingly, the cluster that encompasses shop 1, 3, and 5 showed a stronger similarity. In contrast, shops 11, 12, and 13 appeared different from each other, with weaker or no connections represented by thinner lines. Despite being in the same cluster, these shops displayed quite strong differences based on the analysis. Their unique rules proportion value was 0.66, which aligns with the observation that each of these shops possessed 30 rules, indicating that they did not include 66% of the rules present within the cluster. This finding supported the observations from the heatmaps, where shops 11, 12, and 13 shared almost no rules in terms of support, lift, or confidence. This lead to the assumption that this cluster was formed based on the number of rules or as a "new clusters".

4.3.8. TESTING CLUSTERS

Examining the clusters and their development in total sales, as shown in Figure 1, raised an intriguing question: do clusters formed based on similarities in association rules also exhibit similar growth rates? The following subchapter addressed this topic, thereby also exploring Hypothesis 3:

H3: Shops with similar customer purchase patterns (based on association rules) show different sales growth rates.

Cluster one contained shop 1, 3, and 5. Cluster two contained shop 6, 7, 8 and 9. Cluster three contained shop 11, 12, and 13. The fourth cluster, which only contained shop 4, was dropped for this analysis.
To test for different sales growth rates within the clusters, a Kruskal-Wallis test was deployed, because the data was not normally distributed. The test was deployed three times, for all three clusters. The results can be seen in Table 16 – Kruskal-Wallis Results for Hypothesis 3. The test did not show any significant differences in quarterly growth rates within the clusters. Hence, the alternative hypothesis, that there are differences between the shops in the sales growth rates within the clusters, was rejected.

Cluster	p-value	Test statistic
1	0.32	2.28
2	0.836	0.86
3	0.56	1.14

Table 16 – Kruskal-Wallis Results for Hypothesis 3

Furthermore, it was very interesting that the p-value of the second cluster was 0.836, which was very high and gave a hint that the shops in that cluster could be indeed very similar.
To explore this further, the coefficient of variation (CV) for each cluster was calculated. Each cluster had a mean CV and each shop had a specific coefficient of variation.

Shop	Cluster	CV	Mean CV
1	1	102.85	37.1
3	1	6.17	37.1
5	1	2.30	37.1
6	2	3.74	3.2
7	2	3.361	3.2
8	2	2.985	3.2
9	2	2.7093	3.2
11	3	5.11	2.99
12	3	1,765	2.99
13	3	2.09	2.99

Table 17 – Results of Coefficient of Variation Test

Cluster 1 exhibited the highest average coefficient of variation (CV), indicating a strong variability in quarterly growth rates. This high variability was primarily driven by Shop 1, which showed extreme fluctuations, while Shops 3 and 5 displayed moderate to low variability, respectively. This suggested that Cluster 1 experienced considerable inconsistency in growth patterns, necessitating focused strategies to manage and mitigate volatility.
In contrast, Cluster 2 had the lowest average CV, with all shops showing low variability in their quarterly growth rates. Shops 6, 7, 8, and 9 demonstratede consistent and stable growth, making this cluster the most predictable and reliable in terms of sales performance. This stability can be leveraged for strategic planning and resource allocation, as the predictable growth patterns allowed for more precise forecasting and decision-making.
Cluster 3 showed low to moderate variability, with shops 11, 12, and 13 generally maintaining stable growth, though shop 11 exhibited some fluctuations. This cluster's relative stability, combined with occasional variability, suggested that while most shops can be managed with standard strategies, some may require additional attention to address specific behaviors. Overall, these insights highlighted the potential for tailored business strategies across clusters, focusing on volatility management in Cluster 1, leveraging stability in Cluster 2, and addressing mixed stability in Cluster 3.
While cluster-wide decisions are feasible, the inherent complexity required a nuanced approach. It is essential to carefully address and account for variations within the cluster when applying insights across different shops.

4.3.9. HIGH POTENTIAL LOW SALES ANALYSIS

The final step of the Meta-Analysis focused on rules with high confidence and low sales values to identify opportunities for sales improvement. The analysis started by establishing a low sales threshold, defined as the lowest 10% of total sales values, followed by ranking the rules based on confidence.
This process returns 10 rules that represent opportunities for sales enhancement. These rules, shown in
Table 18 – Results of High Potential and Low Sales Analysis served as valuable indicators for optimizing product assortments or implementing targeted sales strategies to capitalize on untapped potential.

Rule	Shop	Lift	Confidence	Total Sales Value	Support
(ID_187) => ID_167	shop7	102.165	0.851	2581.380	0.001344
(ID_684, ID_532) => ID_686	shop12	145.783	0.909	10512.275	0.001075
(ID_686, ID_532) => ID_684	shop12	172.240	0.833	10512.275	0.001075
(ID_505) => ID_403	shop4	62.396	0.692	12755.850	0.005163
(ID_384, ID_614) => ID_368	shop13	7.583	0.737	39179.970	0.002520
(ID_350, ID_723) => ID_628	shop13	14.674	0.684	44874.050	0.001170
(ID_674, ID_669, ID_671) => ID_621	shop12	10.746	0.67	46148.680	0.001075
(ID_416) => ID_338	shop4	16.453	0.903	46890.440	0.002104
(ID_338, ID_505) => ID_403	shop4	60.447	0.671	59038.500	0.001018
(ID_673, ID_392) => ID_338	shop13	5.028	0.673	60096.900	0.001485

Table 18 – Results of High Potential and Low Sales Analysis

4.4. RECOMMENDATION SYSTEM

The last part of this project was the recommendation system. Association rules are widely utilized in the development of recommendation systems. This approach leveraged the ability to discover interesting relationships and patterns within large datasets, which can then be used to suggest products or services to customers based on their previous behaviors and preferences (Cui, 2021).
The algorithm behind can take one or more products as input and recommends five different products based on the association rules.
Association rules can have the purpose of maximizing profit by increasing the number of products customers add to their baskets. An analysis of all transactions showed that the average number of products per basket was 1.98 throughout all the years and 1.96 in 2023. This number should be the central KPI to track the success of the recommendation system.
The system iterated through all the years and shops, analyzing all the rules to optimize product recommendations and boost sales.
It then checked if the antecedents of each rule included one of the input products. If a match was found, the function appended the metrics; lift, confidence, support, and conviction to a recommendations dictionary for the respective consequent. The four metrics were normalized and then added together to a “total metric”. In this process, lift and conviction were given a

weight of 2, because they are the most important metrics to determine dependency between products. Conviction favors strong confidence values, highlighting very strong predictive connections. Lift, on the other hand, highlights the exceptionality of co-occurrence, ensuring that the algorithm does not just propose the most sequentially bought product but identifies the statistically most exceptional product that shows a significant connection. The shared support of the entire rule is also considered by the algorithm to correct for instances where very strong lift values might occur.
There was a threshold of 1 implemented for the total metric for a product in order to be recommended. If no matches were found, the function returns an empty list. Otherwise, it aggregated the recommendations by calculating the total metric for each consequent and sorted the list based on that in descending order. The algorithm also contained a rule that prevented recommending a product that was part of the input. This was important because when there were many different products in the basket, it was possible that one of the input products would also be a top output product. However, since the customer had already purchased this item, there was no need to recommend it again. The function then returned the top five recommendations with a color gradient. The color intensity and length of the bars in the graphical representation indicated the recommendation power, reflected by the total metric value. The graphical representation was designed to facilitate and speed up the interpretation process. A practical application of this would be at the cashier's counter, where the cashier could see the top five product recommendations each time a customer made a payment. This would enable the cashier to suggest additional products, particularly when the recommendation strength was high. This system is most effective when the data science insights are combined with the staff's human intuition and judgment, which refers to the concept of “Think human, act digital” (Visvizi et al., 2021). Nevertheless, even new employees, without prior experience, could make informed suggestions using the recommendation system, thus ensuring consistent and knowledgeable customer interactions.
The data used for the recommendation system was derived from the initial iteration of association rules for each year and each shop. This means the algorithm utilized unfiltered and redundant rule sets to maximize the available information.

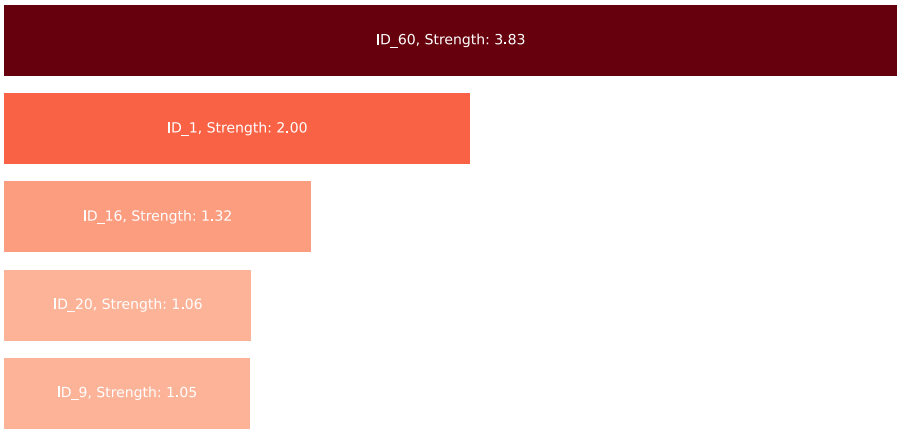

Figure 28 – Output of Recommendation System 1

In the output depicted in Figure 28 – Output of Recommendation System 1, ID 44 was used as input. As previously determined, product ID 44 and product ID 60 formed a high-lift product

pair. Additionally, product ID 1, identified early on as a probable add-on product, appeared to perform exceptionally well.

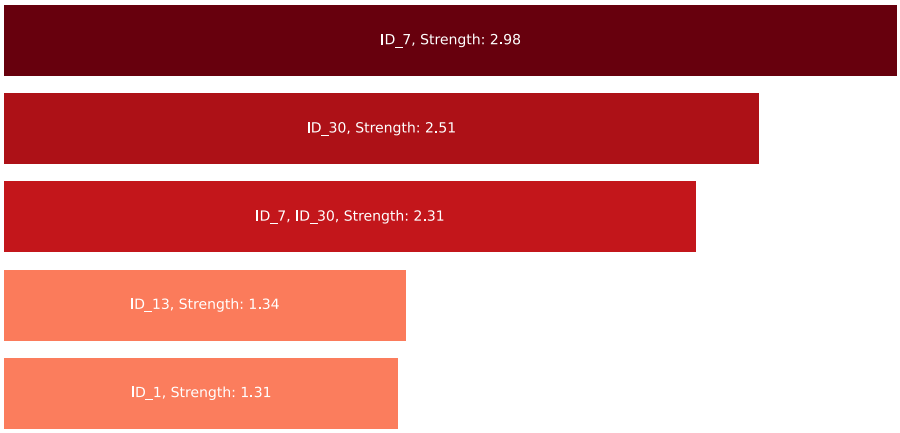

Figure 29 – Output of Recommendation System 2

Figure 29 – Output of Recommendation System 2 shows the results of a random combination of products ID 27, ID 400, and ID 95. Here, interesting patterns were found, where product ID 37, together with product ID 1 exhibited a higher total metric than just product ID 37. However, product ID 1 alone did not exceed the threshold, hence, only four items appeared on the graph. Furthermore, it is advised to reflect on those recommendations with real world knowledge.

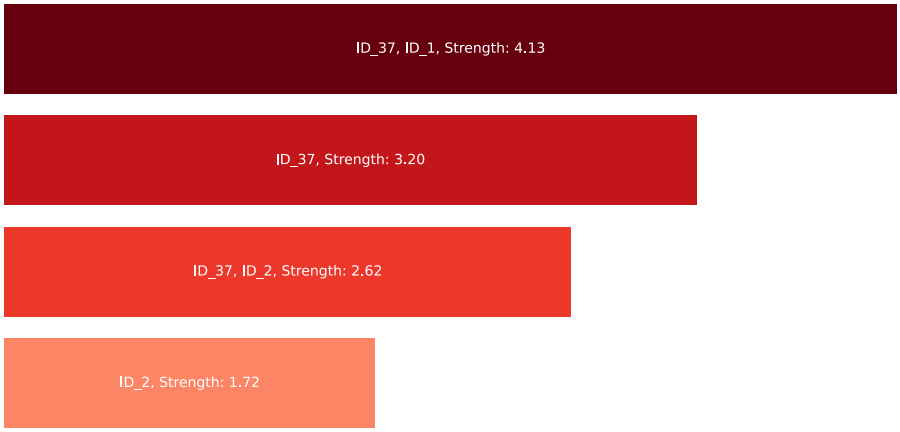

Figure 30 – Output of Recommendation System 3

Figure 30 – Output of Recommendation System 3 shows the recommendations for the input products ID 8, ID 16, and ID 9. The results aligned with expectations from the earlier analysis of these products.

4.5. Threats to Validity

While this work has provided a nuanced understanding of customer purchasing patterns and product associations unique to this bakery, it is important to recognize the threats to validity inherent in this approach.

4.5.1. Construct Validity

Construct validity refers to the extent to which the measurements and procedures used in a study accurately reflect the theoretical concepts they are intended to represent.
In this study, a significant issue with construct validity arose from the fact that the data was received from the company "Mind over Data", not collected by the researcher. This lead to potential problems with data quality and completeness. Additionally, since the data was not collected by the researcher, there may be inherent biases in how it was gathered and recorded.
To mitigate these risks, meetings with responsible data personnel were conducted to understand their data collection processes and ensure alignment with the research needs. During these meetings, detailed questions were answered to verify the accuracy and consistency of the data provided. Furthermore, various significance tests were applied to assess and ensure the reliability and validity of the results.

4.5.2. Internal Validity

Internal validity refers to the extent to which it can be confidently established that a causal relationship exists between variables in a study, without being influenced by other factors or variables. The issue of not collecting but receiving the data also posed a threat to internal validity. Furthermore, confounding variables, such as seasonal events, can influence the validity of the findings.
To address this, data was gathered from multiple locations of the same brand, and significance tests were conducted. Additionally, the confounding factors of seasonality were analyzed and found to be significant. This discovery is valuable in its own right and can now provide useful insights for the business.

4.5.3. External Validity

External validity refers to the degree to which the results of a study can be transferred or generalized to other contexts or groups. This study faces several challenges in this regard:
First, the project was based on data from a single industry (bakery), which limits the ability to generalize findings to other industries. Additionally, since the data was sourced from only one company, it may not represent the broader market or different types of businesses. The data was only collected from Portugal-based locations, which might not be representative of other geographic areas or markets.
To mitigate these issues, the use of multiple locations was emphasized to provide a broader perspective. The study acknowledges the limitations of focusing on a single industry and suggests that further research be conducted in various industries to validate the findings. Replication of the study in different geographic areas is encouraged to test the applicability of the results in diverse markets. Additionally, there is a highlighted need for further studies in businesses with different operational strategies to determine the generalizability of the findings.

4.5.4. CONCLUSION VALIDITY

Conclusion validity refers to the ability to draw statistically correct conclusions based on the measurements. While a common threat to conclusion validity is sample size, a sufficiently large number of subjects were secured to achieve statistical significance.
Association rules were employed, which may present issues, particularly when dealing with blind data where only product IDs are known, but no real world knowledge is provided. The choice of metrics (support, lift, confidence, conviction) could influence the conclusions drawn from the association rules, and potential biases in the provided data might affect the accuracy and reliability of the conclusions. To mitigate these challenges, the most popular metrics were used and multiple metrics were applied and compared consistently. A large sample size was ensured for significant results and hypothesis testing, such as Kruskal-Wallis and Dunn's tests, were conducted to further validate the findings. Recognizing the limitations of using blind data, it was suggested that future studies include detailed product information to enhance the depth and accuracy of the analysis.

5. MANAGERIAL IMPLICATIONS

As mentioned earlier, growth is essential for businesses to remain competitive. Companies that grow constantly are more stable, can react faster to market changes, and manage risk more effectively (Durmaz & Ilhan, 2015). This work states that data science methods, like association rule mining, can enhance strategic decision-making for business growth, by leveraging the insights gained from the analysis, which leads to RQ1.

RQ1: How can data science techniques be utilized to enhance traditional business efforts?

To address research ruestion number three, the managerial implications of integrating insights from association rule mining into existing growth strategies will be explored.
The focus will lie on the earlier mentioned four critical areas of growth: profit generation, cost minimization, improving innovation life cycles and risk mitigation. Each aspect will be discussed individually to highlight how data-driven decisions can significantly impact the operational efficiency and strategic direction of the bakery.

5.1. Strategic Implications for Profit Generation

The exploratory monetary analysis of the bakery indicated a positive trend, suggesting an opportune moment for growth for the bakery, because a high profit is needed for reinvestment into business opportunities (Rakhsitha et al., 2023).
Each shop has identified key rules that drove the most total sales value, and focusing on these can optimize revenue generation. Additionally, unique to each shop are the rules characterized by high lift but low support and confidence, presenting another avenue to increase monetization. Furthermore, add-on products like product ID 1 and ID 2 could be used for pricing strategies. The price could be slightly increased to boost profit or lowered to create an attractive discount strategy. However, pricing strategies are a distinct and complex subject that will not be explored further in this work.

5.1.1. Seasonal Adjustments and Product Offerings

The data revealed significant seasonal variations, with the shops generally performing better in winter and some shops experiencing a slight peak in spring around Easter. The analysis of Hypotheses H1a, H1b, and H1c highlighted significant temporal variations. The results indicated that Saturday consistently generated the highest sales among weekdays, December emerged as the most profitable month, and winter was identified as the season with the greatest total sales value. Additionally, as hypotheses H1a, H2a, H2b, and H2c demonstrated, there are statistically significant differences between the seasons in terms of sales, support, and lift. These findings suggest a recommendation to place greater emphasis on these seasonal purchasing patterns to optimize sales strategies.
Pastry sales were highest in winter (over 35%), followed by fall (25%), spring (22%), and summer (18%), indicating that pastries were better sold during winter and spring. Non-pastry sales analysis revealed that summer and fall exhibited a significant increase, indicating that non-pastry items are better sold during these seasons. This can guide the bakery in designing its product offerings to align with seasonal preferences, ensuring that each shop promotes items that are more likely to be purchased during these times.
Table 1 – Best Seasonal Products Spring to Table 4 – Best Seasonal Products Winter highlight the top 5 products for each season across all shops in 2023, providing valuable insights into

seasonal trends. For example, oroduct 731 showed an increased percentage in 2023, suggesting a rising demand, thereby providing an opportunity for further promotion of the product.
Another interesting observation was for product 761, which, despite its historical importance, did not make it into the top 5 in 2023, indicating potential for rebranding and reintroduction. While a boost in profit can be achieved by selling new products (Rakhsitha et al., 2023), a look into the past, which leads to the reintroduction of these old high performing products can be another way of improving the sales.
A third way is the high potential rules, identified in
Table 18 – Results of High Potential and Low Sales Analysis. These rules hold opportunities for sales and marketing because of high confidence and lift values, but low sales.
Furthermore, the analysis of H2a revealed a greater product variety in winter. This insight provides a foundation for further exploration. For instance, expanding product lines in summer could potentially generate more profit, or alternatively, reducing them even further could help cut costs. Identifying such patterns can help managers optimize product offerings and enhance profitability by focusing on high-demand items and reintroducing previously successful products.

5.1.2. PROMOTION AND DISCOUNT STRATEGY

By leveraging sales data, the bakery can design promotions or bundle offers that capitalize on products frequently purchased together, thus being based on high-support rules. Promotions can also be based on strong lift values. Since lift indicates unexpected patterns and strong connections, often from low-support items, these combinations can be an excellent source for new promotion strategies to push those product pairings. This strategy should be specifically tailored to each shop for optimal effectiveness. Although these promotion offers are based on high-lift rules, they must be carefully monitored, as they might also generate high-lift or high confidence rules. Promotion packages will naturally exhibit high lift or high confidence and appear compelling in the analysis, but this can create a self-fulfilling prophecy. Therefore, they need to be analyzed separately.

5.1.3. RECOMMENDATION SYSTEM

A further development of the promotion concept was to utilize data-driven insights into commonly paired items to implement effective recommendation strategies at each shop to add items to the basket. Training sales staff to utilize the recommendation system creates opportunities for suggesting complementary products at the checkout. This can significantly increase the effectiveness of these upselling strategies.
This approach can also be beneficial for onboarding new staff, enabling them to recommend the right products without extensive work experience. Tracking the duration of the onboarding process can serve as another KPI for assessing the effectiveness of the recommendation system. Furthermore, implementing digital recommendations (*"customers who purchased product A also purchased product B")* prompts based on the created algorithm at self-service stations can also guide customers towards making additional purchases, effectively using the potential of data analytics to enhance traditional marketing strategies and their effectiveness (Troisi et al., 2020).
The analysis revealed that the average number of items per basket is around 1.96 for 2023. One significant way association rules can increase revenue is by adding products to the basket. Monitoring changes in basket size over the years can serve as a valuable KPI for tracking the success of the recommendation system.

5.1.4. Utilizing Top Globally Performing Rules

The top five rules across all shops identified from the analysis and shown in Table 10 could be prominently featured in global marketing efforts such as social media or email campaigns to drive awareness and sales.
Furthermore, if these products are not currently available in every shop, the reasoning behind this and shop-specific performance of these top products needs to be evaluated. However, it is important to remember that the relevance of the top products is highly specific to each shop, and needs to be managed accordingly.

5.2. Strategic Implications for Cost Reduction

As stated earlier, reducing costs is the second rule of profit maximization and is therefore crucial for growth (Rakhsitha et al., 2023). Implementing efficient processes and leveraging technology can lead to significant cost savings. Additionally, cost reduction strategies enhance the financial stability of a business, allowing for reinvestment into growth initiatives and innovation

5.2.1. Cluster Analysis for Tailored Strategies

Identifying clusters of shops with similar customer behaviors facilitates more accurate and broader applicable marketing strategies. This similarity allows for the standardization of promotional activities across these shops, leading to cost efficiencies and enhanced campaign effectiveness. If decisions and strategies can be deployed for an entire cluster and do not have to be tailored for each shop, time and cost savings become very valuable. Specifically, the time and cost are reduced by a factor of one divided by the number of shops in the cluster. This efficiency could result in time cuts to 50% or even 25%.
Hypothesis H3, which posited that shops with similar customer purchase patterns show different sales growth rates, was tested using the Kruskal-Wallis test. The results indicated no significant differences in quarterly growth rates within the clusters, further supporting the feasibility of cluster-wide decision-making despite the inherent complexity.
The coefficient of variation (CV) analysis provided additional insights. Cluster 1 exhibited the highest average CV, indicating strong variability in quarterly growth rates, primarily driven by Shop 1. Cluster 2 had the lowest average CV, demonstrating consistent and stable growth, making it the most predictable and reliable in terms of sales performance. Cluster 3 showed low to moderate variability, suggesting a mix of stable growth with occasional fluctuations.
However, cluster-wide decisions must be approached with caution. Between 40% and 50% of the rules were consistent across the shops in clusters 1 and 2, indicating that 50% of the rules were unique to individual shops within these clusters. However, since the growth rates of the shops within these clusters did not show significant differences and the coefficients of variation were quite low, it can be assumed that decision-making across a cluster is possible. Nevertheless, it remains fairly complex and must always be approached carefully, considering the strategic context.
Furthermore, the paradoxical behavior of Shop 7 needs to be considered, indicating that patterns cannot be blindly transferred from one shop to another. This complexity requires a more nuanced approach when applying insights across different shops within the cluster.

5.2.2. Strategic Product Testing

By testing new products in a single shop before rolling them out across similar shops, the bakery can minimize costs associated with broader testing. This approach allows for an evaluation of customer reception and sales performance in a controlled environment, enabling refined adjustments before wider implementation.

5.2.3. Product Rationalization

Regular analysis of sales performance across all shops can identify underperforming products. Discontinuing these items can streamline operations and focus resources on products that generate higher sales, enhancing overall profitability. Especially regarding the worst rules per shop, an issue tackled in chapter 4.1.7 and depicted in Figure 7 – Presence of Products in "worst list" for each shop (Clustered), shop or cluster wide decisions can be made to remove certain products from the assortment for cutting costs.

5.3. Strategic Implications for Innovation Life Cycles

Reflecting on the established concept that customer orientation can foster innovation (Tuominen et al., 2023) and that innovation drives growth (Ribeiro-Navarrete et al., 2021), the following section will explore possibilities for innovation based on the buying behavior. Furthermore, as mentioned earlier, the number of different products sold each month has been steadily increasing, indicating the bakery's continuous innovation. The following section will explore the various ways in which these innovations can be derived from the findings of the association rules.

5.3.1. Product Line Expansion

Utilizing insights from association rule mining, the bakery can identify potential new product lines or variations that align closely with current customer purchasing behaviors. This method serves as a solid foundation for innovation. This approach to product development is not only strategic but essential for growth, enabling the creation of offerings that are precisely targeted and highly effective in meeting customer needs.

Based on strong associations, the bakery could consider introducing new variations of popular items. For instance, if croissants commonly sell with coffee, introducing flavored croissants or premium fillings could appeal to the same customer base. Innovation in the right direction can give the bakery a competitive edge, since innovation can lead to a unique position in the market (Rakhsitha et al., 2023).

Moreover, leveraging customer behavior insights, especially the affinity between high-lift products like ID 60 and ID 44, is crucial. If these items are complementary, it would be beneficial to enhance features that make use of their connection or to develop new product variations that use their charactistic link in a similar way. This strategy has the potential to directly address and satisfy customer needs. However, it is difficult at this point to make precise suggestions without knowing the actual products behind the codes.

5.3.2. ECO-FRIENDLY PACKAGING

Due to increasing concern about environmental damage, people have begun transforming their environmental awareness into commitments to purchase 'green' items, a behavior known as green purchasing behavior (Sheng et al., 2019).
Based on the analysis of frequently bought-together products, the bakery can implement sustainable practices by strategically bundling items to minimize packaging waste. By encouraging the use of reusable containers and designing bundles that reduce the need for excess packaging, the bakery can effectively appeal to environmentally conscious consumers.

5.3.3. NEW SEASONAL PAIRINGS

Recognizing patterns around seasonal peak events like Easter and Christmas presents opportunities for seasonal product innovations, aligning product offerings with customer expectations during these key shopping periods. The increase in lift during winter, indicating stronger product associations, suggests that bundling is particularly effective during this season. Additionally, it is important to investigate whether the lower lift during summer represents a missed opportunity for bundling, which could potentially be leveraged to improve summer sales. Since there was a substantial difference between the seasons, this could have a strong benefit for the business.

5.3.4. OPTIMIZE SHOP LAYOUTS ACROSS CLUSTERS

By standardizing store layouts in shops or clusters according to successful product placement strategies derived from item association rules, the bakery can enhance customer purchasing behavior efficiently across multiple locations. By strategically placing complementary products together, customers are more likely to make additional purchases, thereby increasing overall sales. This approach not only boosts revenue but also enhances the customer experience by optimizing the shop layout based on items that are frequently bought together.

5.3.5. PERSONALIZED MARKETING

A possible starting point for personalized marketing would be a checkout app, where a barcode is scanned before paying for the products. This approach allows the bakery to gather more demographic information about their customers, enabling them to tailor product offerings more effectively. Personalized discounts and advertisements for new products are excellent ways to leverage personal data. Currently, the recommendation system only uses product-related variables. However, once user data is gathered through the app, the recommendation system can become smarter by incorporating user-centric patterns into the algorithm.
Furthermore, this aligns with the academic body that emphasizes data itself as a strategic asset and an output of innovation (Trabucchi & Buganza, 2019).

5.3.6. MONETIZING DATA

Finally, the bakery could establish a new revenue stream by encrypting and selling data instead of using it only for internal consumption. This could help other researchers and marketers understand the industry better, thus creating value and income.
Additionally, by positioning themselves as a leader in data-driven decision-making within the industry, the bakery could offer consultancy services to similar companies, thereby expanding

their influence and expertise. This strategic move would not only enhance their market presence but also establish them as a key player in promoting data-centric business practices.

5.4. Strategic Implications for Risk Mitigation

Minimizing risk is crucial for growing companies, as investing in expansion often increases a company's vulnerability during the process. The following sub-chapter presents various ways in which association rules can help companies reduce risk.

5.4.1. Temporal Sales Analysis and Early Risk Assessment

The bakery can utilize detailed sales data to explore temporal patterns. This analysis is particularly important for shops identified as part of the same cluster, as similar sales trends can be anticipated. For instance, if a decline in certain product sales is observed in one shop, similar trends can be expected in others within the cluster. In chapter 4.3.3, the analysis of rules changing over time showed that shop 6 exhibited a strong decline for a top rule. The other shops in that cluster should be alarmed, and management should investigate the decline and react accordingly. Patterns like that enable the company to have an early risk assessment.

5.4.2. Seasonal Analysis and Overstocking

Another helpful concept is the seasonal and weekly analysis, which shows certain patterns that can help with demand per weekday and season of the year. With this analysis, the risk and costs of overstocking can be reduced especially for perishable goods. Considering the fact that not all products are available at all times, and the demand for each product rises depending on the season. Regarding Hypotheses 2a and 2c, which identified significant seasonal variations in metrics, management should account for these rule-specific and shop-specific fluctuations when making decisions about inventory and marketing campaigns. This strategic consideration will help mitigate risk effectively. Furthermore, regarding H2a, it is noteworthy that winter not only has the highest amount of sales but also the highest variety of products. This finding necessitates special attention to overstocking and inventory management during this season to optimize operations and avoid excess inventory.
While this study has made strides in understanding and managing inventory levels, it has not fully addressed the issue of overstocking and demand forecasting, leaving room for further exploration to mitigate this risk. Further research could dive deeper into more granular data or employ more sophisticated predictive analytics to refine inventory forecasts and reduce overstocking.

5.4.3. Stability-Focused Strategy

A key strategy involves prioritizing the stability of high-confidence rules across various times and conditions. Although there was no significance for the change of confidence across all rules across the seasons, it is still recommended to focus on rules that show constant confidence levels across the years.
These rules can be found in Chapter 4.3.3 and are depicted in Figure 24 – Confidence Rate Changes Over Time. By focusing on products or combinations that demonstrate consistent performance, the bakery can ensure steady revenue streams. Regular tracking of changes in the confidence levels of association rules, particularly during seasonal shifts or varying market

conditions, helps maintain focus on stable and reliable product offerings. Although high confidence is important for rule stability, a consistent sales count and constant support are also strong indicators of stability in products. To track that stability, a stability index can be implemented that is based on the change in metric analysis. This index would quantify the variations in key metrics, such as sales growth, support, lift, and confidence, over defined periods.

When focusing on single products, it is recommended to emphasize rule sets that have a strong conviction value for that product as a consequent. Meaning, the rule has a strong association towards that product and predicts the purchase of the consequent better. This also helps reduce risk. These high conviction rules are shop specific and can be found in the end of each association rules analysis. In a stable environment, certain decisions can be made with more confidence and safety, which leads to cost savings in numerous cases like inventory management and timings (Rakhsitha et al., 2023).

6. CONCLUSIONS AND FUTURE RESEARCH

In the introduction, it was proposed to explore how data science methods could enhance traditional marketing strategies within the context of a case study of a bakery chain.
The methods used in the project were not entirely clear and only became defined as the analysis progressed. Later, it was determined that the focus would lie on identifying key association rules, analyzing seasonal variations, clustering shops based on purchasing patterns, and developing a recommendation system.
Through a detailed analysis of each interesting shop and their respective years, important association rules highlighting customer purchasing behavior were identified. Initially, it was unclear whether the same thresholds would apply across different years and shops. Therefore, the analysis was conducted individually for each shop and year, resulting in a substantial amount of code. Retrospectively, it was observed that the thresholds were nearly identical across all shops used in the final analysis, suggesting that a loop through the shop and year combinations would have been a more efficient method. However, this incremental approach was necessary to determine the appropriate methodology. For future coding, a more systematic approach is recommended. This approach should automatically analyze different shop and year combinations and return information on whether the required number of rules are generated.
However, different kinds of seasonal variations were evident and could be determined as significant through Kruskal-Wallis tests, emphasizing the importance of targeting different products at different times of the year.
Shops were clustered to reveal similar patterns across certain groups, enabling more efficient knowledge transfer, product testing, risk assessment, and strategic alignment. A Kruskal-Wallis test showed that there were no significant difference in the growth rate within the clusters and the clusters showed strong similarities through the coefficient of variation analysis.
Hence, it can be concluded that this work produced significant results and established a strong foundation for the respective bakery to grow and for future research at the intersection of business and data science. A recommendation system, based on the identified association rules, has shown potential in increasing basket size and overall sales. It could be used to help employees onboard faster and to make better informed recommendations to the customers.
Finally, this work offers numerous benefits for the business, providing valuable insights for growth by implementing the findings into four categories: reducing cost, improving revenue, reducing risk, and facilitating innovation. Given the fact that the product codes were not known, this work provides a high-level approach to identifying interesting patterns and potential strategic implementations of these findings. Without delving into specifics due to the lack of real-world knowledge, it highlights avenues for future strategic decision-making and outlines clear pathways for the business to follow.

6.1. OVERALL BENEFIT

The work examined the potential of data science methods to enhance traditional market strategies, which is, according to the academic body an underexplored area (Troisi et al., 2020). To achieve this, the work proposes a framework, which leads to the third research question:

RQ4: Can the findings from individual bakery shops be generalized into a broader growth analysis framework that is applicable to similar businesses in the sector?

The work seeks to address this question by culminating the insights and learnings into a comprehensive framework designed to benefit not only the company or sector but also marketers in general who aim to foster growth for their organizations.

6.1.1. SAFARI-FRAMEWORK

The **S**trategical **A**pplication **F**ramework of **A**ssociation **R**ule **I**ntegration (SAFARI) is designed to systematize the process of extracting and implementing insights from association rule mining across various aspects of business operations.

1. **Data Collection and Analysis Initiation**
    - Begin by collecting comprehensive transaction data across different points of sale and different points in time.
    - Use association rule mining to uncover patterns and trends that indicate customer preferences and purchasing behaviors across the defined sections.
2. **Network Analysis of Rules**
    - Implement network graphs to visualize and understand product relationships based on their association rules. This iterative process involves evaluating the data using metrics like indegree, support, lift, confidence, and conviction, and filtering the rules to identify key products and combinations.
3. **Meta-Analysis for all Rules**
    - Identify product pairs, strong Association Rules and interesting patterns. Strong Product combinations can be used for promotion and discount strategies and interesting patterns can be used as a foundation for innovation life cycles.
    - Identify products that are performing poorly across all shops and remove them from the assortment to reduce costs.
    - Identify the global top rules across all shops and all years for broad communication and potentially social media.
    - Develop a recommendation system using the identified association rules to suggest products to customers. Potentially implement more sophisticated approaches like purchasing history and cluster based recommendations.
4. **Meta-Analysis for Different Points of Sale**
    - Conduct a Meta-Analysis across different shops to identify common patterns and clusters based on the top association rules found in each shop in each year.
    - Utilize these insights to understand similarities between shops to carefully standardize operations across similar shops and tailor strategies to unique market demands where necessary. Use these findings for cost reduction.
5. **Meta-Analysis for Different Points of Time**
    - Analyze the change of confidence, support and lift over time, evaluating the stability of identified association Rules. Identify risk factors by examining rule instability and use this for early risk assessment and mitigation. Focus on rules with high stability across different times and conditions for consistent implementation. Use these findings for risk mitigation.
    - Find Seasonal patterns, high performing products, historically important products for potential re-introduction and high potential rules, rules with strong metrics. Use these findings for profit generation.

6. **Feedback and Continuous Improvement**

Track the following KPIs to evaluate this framework. These KPI's could be culminated in a dashboard to create a comprehensive overview of the performance based on this framework.

- **Profit Generation:**
    1. Average basket size
    2. Cross and upselling rates of the recommendations
    3. Bundle offer conversion rate
    4. Revenue increase from seasonal promotions
    5. Seasonal sales growth
    6. ROI of marketing efforts
- **Cost Reduction:**
    1. Onboarding duration for new staff because of the recommendation system
    2. Cost savings because of cluster-wide marketing campaigns
    3. Time savings because of cluster-wide marketing campaigns
- **Innovation:**
    1. New product introduction rate
    2. Sales growth through innovative products
    3. Product development life cycle time
    4. Customer Feedback
- **Risk Mitigation:**
    1. Risk incident frequency (per month)
    2. Early warning indicators (per month)
    3. Number of risks identified through association rule-based mitigation (per month)
    4. Number of risks identified through other means (per month)
    5. Stability index of key products
- **Other:**
    1. Change of coefficients of variation over time for clusters
    2. App engagement metrics
    3. Conversion rate of digital marketing campaigns
    4. Social media engagement rate (especially for top global rules)

Finally, continuously refine and expand the analytical models to adapt to changing market conditions and consumer trends.

6.2. Limitations Of The Framework

The Strategical Application Framework of Association Rule Integration may face limitations because of its reliance on historical data. It may not fully capture emerging trends and sudden market shifts. However, its primary goal is to provide insights for mid to long-term growth strategies rather than reacting ad hoc to immediate market movements. Additionally, analyzing each shop, each year, and the meta data correctly may require an expert with a strong understanding of both business and data science.
Lastly, the quality of the outcomes, performance of the recommendation system, and also applicablity of the framework heavily rely on the quality of data. Continuous data validation and improvement processes are essential to maintain the accuracy and effectiveness of the framework.

6.3. Future Research

For future research, a critical task is to test the framework, monitor the KPIs, and evaluate the efficacy of the association rules concepts. Additionally, further analysis can include customer segmentation, demand forecasting, and exploring variables like product families and sub-families.

The marketing strategies that emerge from these findings need to be tracked and analyzed to determine if the insights provided are effective. This testing should start within the bakery that this project is based on but also extend to other companies in the food sector. Expanding the research to other industries and countries will help explore differences and make the framework more stable and adaptable. This broader application will allow for the comparison of results and further validation of the findings. Additionally, testing the framework across multiple studies will address various threats to validity.

Furthermore, it would be very valuable to conduct a similar analysis with real-world knowledge to implement practical insights into the findings. Future research should also explore psychological factors influencing customer decisions and compare these with the association rules. Understanding these dynamics can significantly enhance the recommendation system and provide more depth to marketing strategies.

The recommendation system could be further enhanced by incorporating customer-based variables like demographic variables and customer segments based on their behavior. Another interesting concept would be to consider the specific shop where the product is purchased. Currently, the recommendation system operates globally, similar to an online shop, without considering specific locations. In reality, a customer is always at a specific point of sale in the bakery context, so the system could include the location of the customer, for instance, shop 6. The recommendation algorithm could then prioritize rules from the cluster that includes shop 6, while still learning from other shops in other clusters. Since it is known that knowledge transfer works better within clusters, this concept could be integrated into the recommendation system.

By addressing these areas, future research can build on the foundational work laid out in this study, expanding its relevance and improving its utility for broader business applications. This continuous cycle of analysis, testing, and refinement is essential for evolving business strategies to keep pace with a rapidly changing market landscape.

BIBLIOGRAPHICAL REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proc. of 20th International Conference on Very Large Data Bases, {VLDB'94}*.

Almansour, M. (2022). Food start-ups: leveraging digital marketing and disruptive information systems innovations to survive in the post-COVID environment. *European Journal of Innovation Management*. https://doi.org/10.1108/EJIM-07-2022-0370

Aluri, A., Price, B. S., & McIntyre, N. H. (2019). Using Machine Learning To Cocreate Value Through Dynamic Customer Engagement In A Brand Loyalty Program. *Journal of Hospitality and Tourism Research*, *43*(1). https://doi.org/10.1177/1096348017753521

Azevedo, P. J., & Jorge, A. M. (2007). Comparing rule measures for predictive association rules. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4701 LNAI*. https://doi.org/10.1007/978-3-540-74958-5_47

Barnett, M. J., Doroudgar, S., Khosraviani, V., & Ip, E. J. (2022). Multiple comparisons: To compare or not to compare, that is the question. In *Research in Social and Administrative Pharmacy* (Vol. 18, Issue 2). https://doi.org/10.1016/j.sapharm.2021.07.006

Bekata, A. T., & Kero, C. A. (2024). Customer orientation, open innovation and enterprise performance, evidence from Ethiopian SMEs. *Cogent Business and Management*, *11*(1). https://doi.org/10.1080/23311975.2024.2320462

Braguinsky, S., Ohyama, A., Okazaki, T., & Syverson, C. (2021). Product Innovation, Product Diversification, and Firm Growth: Evidence from Japan's Early Industrialization. *American Economic Review*, *111*(12). https://doi.org/10.9767/BCREC.17.1.12392.65-77

Brynjolfsson, E., Hitt, L., & Kim, H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? *International Conference on Information Systems 2011, ICIS 2011*, *1*. https://doi.org/10.2139/ssrn.1819486

Cui, Y. (2021). Intelligent Recommendation System Based on Mathematical Modeling in Personalized Data Mining. In *Mathematical Problems in Engineering* (Vol. 2021). https://doi.org/10.1155/2021/6672036

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *Stata Journal*, *15*(1). https://doi.org/10.1177/1536867x1501500117

Djuraeva, L. (2021). Importance of the Innovative Business Models for the Future Success of the Company. *SHS Web of Conferences*, *100*. https://doi.org/10.1051/shsconf/202110001013

Dul, J., & Hak, T. (2007). Case Study methodology in business research. In *Case Study Methodology in Business Research*. https://doi.org/10.4324/9780080552194

Durmaz, Y., & Ilhan, A. (2015). Growth Strategies in Busınesses and A Theoretical Approach. *International Journal of Business and Management*, *10*(4).

Garcia-Martinez, L. J., Kraus, S., Breier, M., & Kallmuenzer, A. (2023). Untangling the relationship between small and medium-sized enterprises and growth: a review of extant literature. *International Entrepreneurship and Management Journal*, *19*(2). https://doi.org/10.1007/s11365-023-00830-z

Graves, S. C. (2011). Uncertainty and production planning. In *International Series in Operations Research and Management Science* (Vol. 151). https://doi.org/10.1007/978-1-4419-6485-4_5

Hahsler, M., & Chelluboina, S. (2011). Visualizing Association Rules: Introduction to the R-extension Package arulesViz. *R Project Module*.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, *29*(2). https://doi.org/10.1145/335191.335372

Hoque, Engr. M. J. (2024). Optimizing Decision-Making Through Customer-Centric Market Basket Analysis. *Journal of Operational and Strategic Analytics*, *2*(2), 72–83.

Lee, S. W. (2022). Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee. *Life Cycle*, *2*. https://doi.org/10.54724/lc.2022.e1

Majeed, A., & Lee, S. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2020.3045700

Patil, S., Khan, H., Mehta, S., & Mandawkar, U. (2021). STUDY OF CUSTOMER SEGMENTATION USING k-MEANS CLUSTERING AND RFM MODELLING. *Journal of Engineering Sciences*, *12*(6).

Rakhsitha, Sonegara, S., Shreya, R., Kumawat, S., Singh, S., Raj, S., & Shukla, S. (2023). Profit Maximization Principles for Business Growth in the Modern World. *International Journal of Development Research*, *13*(04).

Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review*, *56*(8). https://doi.org/10.1007/s10462-022-10366-3

Ribeiro-Navarrete, S., Botella-Carrubi, D., Palacios-Marqués, D., & Orero-Blat, M. (2021). The effect of digitalization on business performance: An applied study of KIBS. *Journal of Business Research*, *126*. https://doi.org/10.1016/j.jbusres.2020.12.065

Sarker, I. H. (2021a). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. In *SN Computer Science* (Vol. 2, Issue 5). https://doi.org/10.1007/s42979-021-00765-8

Sarker, I. H. (2021b). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). https://doi.org/10.1007/s42979-021-00592-x

Sheng, G., Xie, F., Gong, S., & Pan, H. (2019). The role of cultural values in green purchasing intention: Empirical evidence from Chinese consumers. *International Journal of Consumer Studies*, *43*(3). https://doi.org/10.1111/ijcs.12513

Shetty, P., & Singh, S. (2021). Hierarchical Clustering: A Survey. *International Journal of Applied Research*, *7*(4). https://doi.org/10.22271/allresearch.2021.v7.i4c.8484

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability (Switzerland)*, *14*(12). https://doi.org/10.3390/su14127243

Trabucchi, D., & Buganza, T. (2019). Data-driven innovation: switching the perspective on Big Data. *European Journal of Innovation Management*, *22*(1). https://doi.org/10.1108/EJIM-01-2018-0017

Trabucchi, D., Buganza, T., & Pellizzoni, E. (2017). Give Away Your Digital Services: Leveraging Big Data to Capture Value. *Research Technology Management*, *60*(2). https://doi.org/10.1080/08956308.2017.1276390

Troisi, O., Maione, G., Grimaldi, M., & Loia, F. (2020). Growth hacking: Insights on data-driven decision-making from three firms. *Industrial Marketing Management*, *90*. https://doi.org/10.1016/j.indmarman.2019.08.005

Troisi, O., Visvizi, A., & Grimaldi, M. (2023). Digitalizing business models in hospitality ecosystems: toward data-driven innovation. *European Journal of Innovation Management*, *26*(7). https://doi.org/10.1108/EJIM-09-2022-0540

Tuominen, S., Reijonen, H., Nagy, G., Buratti, A., & Laukkanen, T. (2023). Customer-centric strategy driving innovativeness and business growth in international markets. *International Marketing Review*, *40*(3). https://doi.org/10.1108/IMR-09-2020-0215

Visvizi, A., Troisi, O., Grimaldi, M., & Loia, F. (2021). Think human, act digital: activating data-driven orientation in innovative start-ups. *European Journal of Innovation Management*, *25*(6). https://doi.org/10.1108/EJIM-04-2021-0206

Walcott, T. H., & Ali, M. (2021). Machine Learning for Smaller Firms: Challenges and Opportunities. *Proceedings - 2021 International Conference on Computing, Electronics and Communications Engineering, ICCECE 2021*. https://doi.org/10.1109/iCCECE52344.2021.9534852

Wang, Q., Zhao, X., & Voss, C. (2016). Customer orientation and innovation: A comparative study of manufacturing and service firms. *International Journal of Production Economics*, *171*. https://doi.org/10.1016/j.ijpe.2015.08.029

Wirtz, J., & Zeithaml, V. (2018). Cost-effective service excellence. *Journal of the Academy of Marketing Science*, *46*(1). https://doi.org/10.1007/s11747-017-0560-7

7. APPENDIX

The code used for the analysis is available both as a Jupyter notebook and as a PDF online, allowing the reviewers of this work to follow the steps comprehensively. The complete source code for the analysis can be found here:

https://github.com/MarcJerschov/MasterThesis

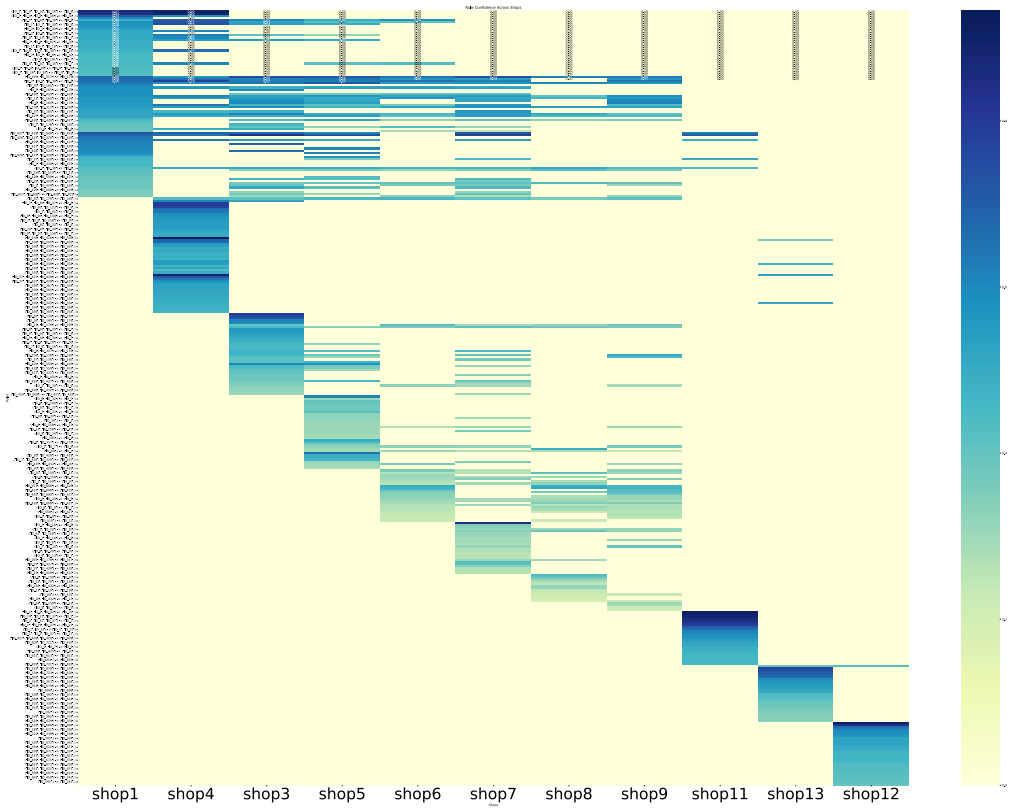

Figure 31 – Heatmap Confidence (All Shops All Years)

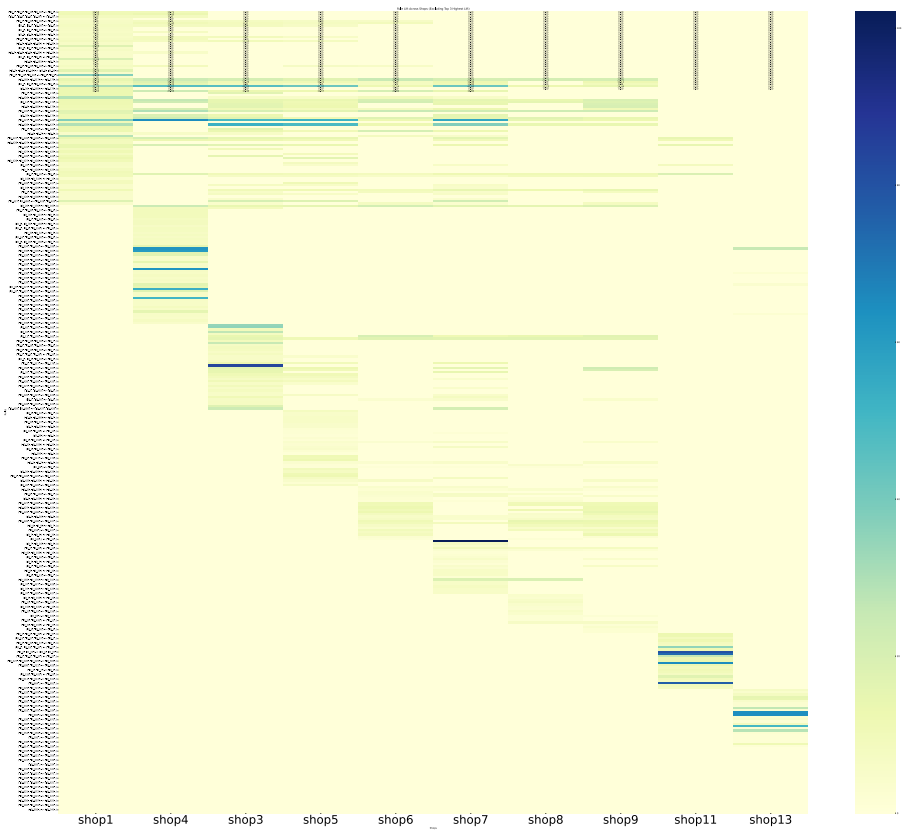

Figure 32 – Heatmap Lift (All Shops All Years)

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa