

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/377926263>

Understanding Customer Buying Patterns Through Business Analytics

Thesis · February 2024

CITATIONS

0

READS

442

1 author:



[Kingsley I. Obiegbo](#)

University of Pécs

4 PUBLICATIONS **0** CITATIONS

SEE PROFILE

**University of Pécs,
Faculty of Business and Economics**



***UNDERSTANDING CUSTOMER BUYING
PATTERNS THROUGH BUSINESS ANALYTICS***

OBIEGUO Ifeanyi Kingsley

Pécs, 2023

**University of Pécs,
Faculty of Business and Economics**

***UNDERSTANDING CUSTOMER BUYING
PATTERNS THROUGH BUSINESS ANALYTICS***

By

OBIEGUO Ifeanyi Kingsley

Instructor: Prof. Ferenc Kruzslicz | Associate professor

**University of Pécs
Faculty of Business and Economics
MSc in Business Development**

Pecs, 2023

ACKNOWLEDGEMENT

I sincerely give glory, honor, and adoration to the Almighty God for He alone has made this thesis and my program a success and a reality, through thick and thin. He has been my strength and source of inspiration. May His name be praised forever, Amen.

My profound gratitude to my supervisor Professor Ferenc Kruzslicz who constructively proffered suggestions, and observations and guided me in the construction of this thesis. His support and encouragement in the limited time we had to complete this thesis was amazing.

I want to acknowledge Gerda Doczi of the study department for her unflinching encouragement and motivation to ensure I persevere until the end of the development of my thesis, I am forever grateful.

I wish to express my profound gratitude to my parent Mr. and Mrs. Obieguo Martin Chukwuji for their consistent prayers, support, and faith in me, and for their guidance and always motivating me in the right direction. You guys are the real MVP.

My sincere appreciation to my siblings namely: Obieguo Calister, Rev. Sis. Obieguo Judith, Obieguo Maryjane and Obieguo Victoria for their unflinching support. In the same vein, my colleagues in the business development program are not left out; I respectfully acknowledge your companionship, advice, and willingness to assist during the time of academic needs. I cannot sufficiently express my gratitude for your prayers, team spirit, and oneness. I am indeed grateful. May all your efforts be crowned with success in the name of our Lord Jesus Christ.

To Kopf Andras, Nahid, Clement, Bence, and Leap for being true friends throughout my study period. I really appreciate you guys from my heart.

Finally, I wish to acknowledge that it is a privilege to be used as an instrument in the hands of God; Almighty Father, I, am honored.

Table of Contents

ACKNOWLEDGEMENT	iii
LIST OF FIGURES.....	3
LIST OF TABLES.....	3
LIST OF APPENDICES.....	4
CHAPTER ONE	5
INTRODUCTION	5
1.0 Background of the Study.....	5
1.2 Problem Statement	7
1.3 Justification for the study	9
1.4 Research Aim and Objectives	9
1.5 Research Question.....	10
1.6 Purpose of the Study	11
1.7 Scope of the Study	11
1.8 Limitation of the Study	11
1.9 Structure of the Study.....	12
CHAPTER TWO	14
CONCEPTUAL MODEL	14
2.0 Introduction	14
2.1 Concept of Buying Patterns	14
2.2 Types of Buying Patterns	15
2.3 Store Layout	17
2.4 Inventory Management	17
2.5 Concept of Data Mining.....	19
2.6 Data mining techniques.....	20
2.6.1 The Predictive Model.....	21
2.6.2 The Descriptive Model	22
2.7 Association Rule Mining (ARM).....	23
2.8 Measures of Interestingness	27

2.9	Searching Frequent Itemset.....	30
2.9.1	Apriori Algorithm	31
2.9.2	Frequent Pattern Growth Algorithm (FP-Growth)	32
2.10	Empirical review	34
2.9	Conclusion.....	36
CHAPTER THREE.....		37
METHODOLOGY		37
3.0	Introduction	37
3.1	CRISP-DM.....	37
3.1.1	Business Understanding.....	38
3.1.2	Data Understanding	39
3.1.3	Data Preparation	40
3.1.4	Data Modeling	41
3.1.5	Data Evaluation.....	41
3.1.6	Data Deployment	41
3.2	Analytical tools	41
3.3	Conclusion.....	42
CHAPTER FOUR.....		43
STATISTICAL OVERVIEW AND ASSOCIATION RULE MINING		43
4.0	Introduction	43
4.1	Statistical Overview	43
4.1.1	Understanding the Price and Product Distribution.....	43
4.1.2	Price and Product Analysis	45
4.1.3	General Product Structure	46
4.2	Association rule mining	48
4.2.1	Data Modeling	50
4.2.2	All Confidence Measure	54
4.2.3	Sensitivity analysis of support threshold	57
4.3	Conclusions	58
CHAPTER FIVE		59
CONCLUSION AND RECOMMENDATIONS.....		59

5.0	Conclusion.....	59
5.1	Recommendations	61
5.1	Contribution to Knowledge.....	62
5.3	Suggestions for further research.....	62
	REFERENCES.....	64
	APPENDIX.....	74

List of Figures

Figure 1: Thesis framework	10
Figure 2: Data Mining Model	21
Figure 3:Steps of CRISP-DM Methodology	38
Figure 4: Rank order of customers based on price (1).....	45
Figure 5: Rank order of customers based on price (2).....	46
Figure 6: Customer Segmentation based on Price.	47
Figure 7: Average Unit Price of the Basket	47
Figure 8: Data Preparation	49
Figure 9: Rapid Miner Process Design for the association rule mining	50
Figure 10:Visuals of the Association rules based on the Support and Confident Thresholds..	56

List of Tables

Table 1:Structure of the Study	122
Table 2:Compressed Comparison of Apriori and F-P Growth Algorithms.....	33
Table 3: A summarizes a review of related studies in descending order of year.	34
Table 4: First 10 rows of the original dataset	40
Table 5: Price and Product Distribution	44
Table 6: Data Transformation.....	48
Table 7: Most Frequent two-item set combinations.	52
Table 8: Most Frequent three-item set combinations.	52
Table 9: All Confidence Evaluations.....	54
Table 10: Result of Association Rule Mining	55
Table 11: Highest Lift Rules.....	57

Table 13: Lowest Lift Rules	57
-----------------------------------	----

List of Appendices

Appendix 1: Price and Product Distribution.....	
Appendix 2: Data Transformation.....	
Appendix 3: Frequent Item Combinations	
Appendix 4: All Confidence Evaluation.....	
Appendix 5: Sensitivity Analysis Table	

CHAPTER ONE

INTRODUCTION

1.0 Background of the Study

As global businesses take a new turn, an intimate understanding of the customers to predict their ultimate wants becomes the supreme goal of retail stores all around the world (Raymond, 2019). Studies conducted by Tanja (2015); Pinakshi (2019); and AlShamsi (2022) reveal that understanding customers' expectations is the most critical challenge retailers face today, with consumers having an uncountable list of options and product varieties available on a large scale almost independently in all domains (Monerah and Ahmed, 2021), and the recent development in information technology and globalization, exponentially raising these list of options to a different level (Sohaib, 2019), such that consumers can now choose between a huge variety of products and their variances. This growth in the retail sector has created a major shift in recent years, with major retail players jumbling to grow or at least retain their market share. The shifts included a more frequent splitting of customer purchases between retailers like ever before, tighter competition and disruptive practices amongst grocery retailers, increasing online retailing, and the frequent use of smaller shopping trips due to compact household budgets, etc. (Raymond, 2019). Given this effect caused by new entrants, retail stores all around the world are now compelled to seek progressive and robust marketing strategies at a rapid pace to effectively acquire new customers and to at least retain their existing customers (Monerah and Ahmed, 2021).

While a handful of traditional marketing strategies have been in practice for ages to help retail outlets understand, and anticipate customers' needs, the era of information and communication technology introduced newer and more robust techniques and strategies that rendered the traditional method not just obsolete but also inefficient to use in understanding customers behavior for decision making (Pinakshi, 2019). Several previously conducted studies (Barbera, Amato, and Sannino, 2016; Tingting, William, Ling, and Yan, 2019; AlShamsi, 2022), reported how different and new technologies have facilitated the prediction of customers' purchase behavior, enhanced product placement, and provided insight vision based on customers demand (AlShamsi, 2022). One of the most highly mentioned tools in their studies that facilitated these

processes was a data mining tool also known as a knowledge discovery and data process (KDD), a process that is associated with the identification and uncovering of patterns from large data sets to gain insights and other valuable information for business decision making.

It is highly important to mention that in the last two decades, there has been explosive growth in data, with massive amounts of data collected and stored every day about customers such as the “customer profiles, periodic and lifetime transactions, sales data, customers everyday life activities and even their future desires and expectations”. Despite the increasingly large amounts of customer data available for use, many organizations are still in a maze as to how to unscramble the power behind the data to unleash its full potential (Nayyar, 2019). Pinakshi (2019) suggested that to extract only the relevant data from these huge datasets, hidden patterns are required to be identified. One of the most well-known methods for detecting underlying patterns hidden in large transactional data sets is association rule mining, generally known for its use of machine learning techniques to first, identify frequently purchased product combinations and secondly, to discover concealed associations among the products (Kuisma, 2019). This machine learning technique is recognized as a very powerful technology because it can capture a massive quantity of consumer data in real-time, and with its predictive capabilities, it can assist businesses in making effective decisions in sales and promotion strategies and enhancing marketing activities after proper analyzes of sales pattern and consumer behavior to generate profitable outcome for the business. (Kurnia, Yohanes, Yo, Aditiya, & Riki, 2019).

Since the main focal point of interest for retail stores has been to obtain information about the preferences, motives, tastes, purchases, choices, habits, and demands of customers based on their behavior (Monerah and Ahmed, 2021), to create some sort of understanding of their needs so that these needs can be satisfied effectively, discovering all the concealed information and patterns within the entirely large amount of collected from the varied collections of sources becomes practically inevitable to make effective decisions that can at least lead to the expected satisfactions. It is against this backdrop that this study seeks to understand customer buying patterns through business analytics for effective decision making, by establishing novel data mining models that grocery retail stores can adopt to predict customer behaviors and

specifically understand how customers split their purchases across multiple grocery product categories.

Additionally, there has been only a few published literature on the precise descriptions of the analytical techniques and tools used by individual retailers, various algorithms, models, and frameworks have been developed by several researchers in the hope that grocery retailers will likely leverage these solutions to facilitate their businesses, and by other business practitioners to enhance their business operations in their respective domains and specializations (Nathan, Yuchi, Xueming, and Xiaoyi, 2016). It is important to note that when measuring interesting by this businesses, the major determinant to consider is the context or domain in which they are being applied, hence the need for plethora of measures and the specificity of application domain. This study is therefore framed on this theory and ideology to generate new association rules that might be helpful to a gift retail store and other retail stores in identifying the best item combinations and customers to target and to support the retail stores with the prediction about customer behavior and more specifically on how customer split their purchases across multiple options of retail stores. By understanding the various faces of customers' unstable behavior, retail stores might be able to enhance their efforts in providing a unique customer experience through tailored marketing promotions and campaigns thereby: (1) enhancing the effectiveness of, and benefitting from the expenses made on marketing, (2) enticing more customers to remain while reducing the level of switch to other retail stores, where customers split their purchases, and (3) enhancing a sustaining loyalty amongst its current customers (Raymond, 2019).

1.2 Problem Statement

From the traditional roadside stores of the 1900s to the modern superstores of today, the shopping experience has changed drastically (Monerah and Ahmed, 2021). The transition created to a new era of international competition and business opportunities with customer adoption becoming even more challenging than maintaining a long-term relationship with the existing ones (Mehmet & Robert, 2022). This has created the need for retail stores to begin searching for a new approach to reach their target audience and to keep the existing ones. The most proven and efficient approach has been through customer understanding (Tung & Carlson, 2015). Studies conducted by Tanja (2015); Pinakshi (2019); and AlShamsi (2022) have revealed

that one of the most critical challenges facing retailers today is proper comprehension of customers' expectations and needs. It may seem new and interesting, but surprisingly, many retail stores are already familiar with this challenge considering the large amount of customer data available these days to deal with. While in the past a customer interested in purchasing items could only choose a product from the catalog, the new era of information and globalization has provided a list of options to customers that increases exponentially and limitlessly (Sohaib, 2019), thereby influencing the way customers think, and feel, the choices and purchases they make, their habits, motives, preferences, tastes, demands and their overall buying behavior.

With these possibilities made available, many retail stores grapple with the fact that customers' buying patterns are never stable, switching from one competitor to another in response to an attractive competitive offer (Wagner, Michel, Fernando, & Jose, 2023). They seem to acknowledge the fact that it is impossible to understand consumer behavior completely because it is closely connected to the human mind so much so that even the consumers themselves might not recognize why and how their behavior changes (Tanja, 2015). However, there has been an increase in search by retailers for various strategies to unravel this complex puzzle posed by customer's changing behavior. A large number of retailers have resorted to the adoption of a data mining approach, through market basket analysis (association rule mining) since it has proven to be worthwhile. One cornerstone for the adoption of data mining is data (*which nowadays is considered to be the new oil*) to understand customers' behaviors and how to satisfy them.

According to Sohaib (2019), the last two decades have experienced exponential growth in the data, both relevant and irrelevant. The proportion of data generated daily has increased ten times greater than what can be utilized by retail stores globally, such that the biggest challenge faced by retailers today is how to uncover the hidden knowledge deposited in huge amounts of data available from varying collections of sources (Anindita, 2016). To extract only the most important and relevant data, identifying hidden pattern are a key requirement, this is where data analytics comes in handy as it is recognized as a very powerful technology for capturing massive quantities of consumer data in real-time (Pinakshi, 2019); this process of uncovering and mining useful information and patterns from data is what is called data mining originally known as

knowledge discovery and data process (KDD) (Shashi, Ajai, Neetu & Rachna, 2018), with several mining techniques, such as clustering, association, classification, forecasting, regression, visualization, and sequence discovery used in performing one or more modeling (Ngai, Xiu, & Chau, 2009)

Although several systems have put into practice customer behavioral analytics, it's still an upcoming and unexplored research area that has greater potential for better advancements. Also, there is little published literature on the exact explanation of the analytical techniques used by individual retailers. However, researchers consistently developed algorithms, models, and frameworks, in most cases literarily independent of any retail store inputs, with the hope that retail stores generally would likely leverage these solutions to enhance their business operations in their various domains (Nathan, Yuchi, Xueming, and Xiaoyi, 2016). It is against this backdrop that this study seeks to understand customer buying patterns through business analytics by establishing novel data mining models that can help grocery retail stores predict customer behaviors and specifically understand how customers divide their spending across multiple grocery product categories.

1.3 Justification for the study

Several individualistic assessments have been conducted by various researchers on customer buying behavior and business analytics across diverse areas and disciplines, but none has investigated the Understanding of Customer Buying Patterns through Business Analytics to Structure Promotion and Uncover Cross Demands for retail stores. In the same vein, several systems have conducted research on customer behavior analytics, and developed models, frameworks, techniques, and even algorithms, in several cases independent of the retail stores, yet there still exist some gaps and unexplored areas of interest to be researched upon and that has greater potential for advancement. This study is an attempt to build upon a body of reliable and verifiable information about business analytics and customer behavioral understanding.

1.4 Research Aim and Objectives

This study aims to understand customer buying patterns through business analytics. In doing this, the study underpinned market basket analysis to uncover new association rules that may

be applicable in different business environments. The result of this study contributes to identifying interesting insights about customers and provides actionable recommendations to case companies that they can rely upon to succeed in the demanding market conditions.

1.5 Research Question

The research question of this study is defined based on the established backgrounds and motivations of this study. The study intends to answer the following research questions:

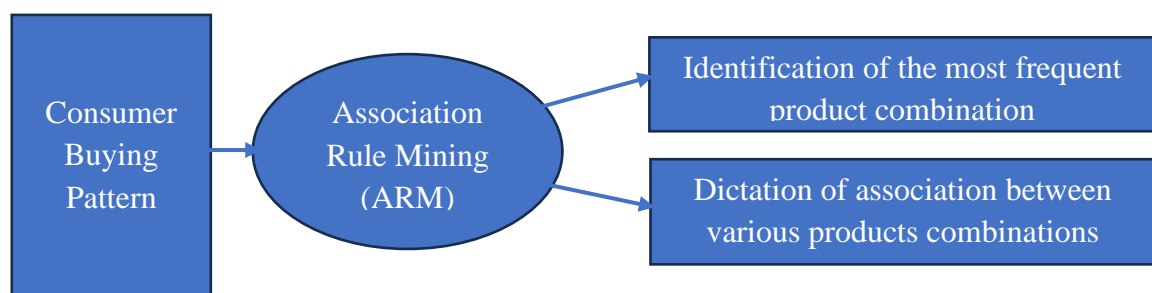
1. Which is the most frequent product combination?

This question aims to identify the most purchased product combinations. The question seeks to provide added value to the retail store as it may unveil previously unknown information and insights. To answer this question, items or products with the most frequent combinations will be extracted from the customer transaction data used in the study.

2. What sort of associations can be identified between the different products and their combinations?

Once the first research question is answered the next is to answer the second research about the sort of association that is detected. To do this, the process of association rule mining comes in handy. Additionally, the quality of the association will be assessed using various metrics.

Figure 1: Thesis framework



Source: Authors Constructs (2023)

1.6 Purpose of the Study

Understanding customer buying patterns has become a capstone of interest for many retailers recently, this is a result of the benefits it presents to them in the face of the fierce competitive environment. This thesis is purposed to help retailers (a hypothetical gift store) understand consumer buying patterns through business analytics for the effective decision making, underpinning market basket analysis to uncover new association mining rules. To reach the purpose of the thesis, the FP-Growth Algorithm was utilized to find patterns from the online retail transactional data in France adopted for the study.

1.7 Scope of the Study

The primary focus of this study is understanding consumer buying patterns in retail stores, within the context of business analytics. It covers various data related to customers' behavior, and buying patterns and considers these buying patterns from the lens of customer market basket in the retail store to uncover interestingness. The study covers market basket analysis as a cornerstone for developing new association rules that could be applied in various fields, especially in the retail industry.

1.8 Limitation of the Study

Research is a result of human effort and since imperfection is human, it is probably not going to be surprising to identify some failing in this research work (Ezuma, 2010) this study has encountered some setbacks and limitations which include:

Difficulty in accessing real-time data of most organizations and retail stores, probably due to GDPR being a topic of discussion as many customers are now concerned about how their data is being utilized. Even though the sort of data needed for this study does not require the personal details of the customer, getting data for this study was a major challenge.

Additionally, gaining insights about, and utilizing a new methodological tool or statistical technique requires a bit of hands-on experience. Although this study described how to use market basket analysis through an expository dataset, there were several drawbacks when

practically analyzing the data, several iterations were required to get better and more effective results.

1.9 Structure of the Study

As depicted in the table below, this study has been structured into two parts and five chapters.

Part 1 presents the foundational and preliminary knowledge which is required to understand this study.

Table 1: Structure of the Study

Part 1 Foundational and Preliminary Knowledge	Chapter 1 <i>Introduction</i>	Chapter 2 <i>Concept of Customer Buying Pattern and Market Basket Analysis</i>	Chapter 3 <i>Data presentation, description, and exploration</i>
Part 2 Conceptual Model	Chapter 4 <i>Statistical overview and association rule mining</i>		Chapter 5 <i>Conclusion, and Recommendations</i>

Source: Authors Construct (2023)

Chapter 1 (*Introduction*) established the context of this study and detailed the research problems. It also set out the research objective and the research questions addressed by the study.

Chapter 2 (*Concept of Customer Buying Pattern and Market Basket Analysis*) detailed the concept of customer buying patterns as illustrated in related literature and the phenomenal market basket analysis, highlighting the different data mining techniques and association rule mining.

Chapter 3 (*Data presentation, description, and exploration*) Present the data and the methodology for the data mining underpinning market basket analysis to achieve the objective set in the study.

Part 2 elaborates on the conceptual model underpinning this study and the threshold for measuring the interestingness of the pattern in the study.

Chapter 4 (*Statistical Overview and Association Rule Mining*) elaborates on the model for analysis as well as the development of a set of metrics for the measurement components defined in the study objectives.

Chapter 5 (*Conclusion, and Recommendations*) presents an analysis of the extent to which the research questions defined in the previous chapters have been answered by the thesis, summarizes its contribution, and makes some recommendations.

CHAPTER TWO

CONCEPTUAL MODEL

2.0 Introduction

This chapter presents a review of relevant literature associated with this study. It includes the conceptual issues on customer buying patterns and the phenomenal market basket analysis, highlighting the different data mining techniques and association rule mining. Strong attention was also given to the review of current literature related to the study.

2.1 Concept of Buying Patterns

In today's ever-changing and dynamic business environment, it has become highly essential for retail managers to not only have a glimpse of their customer's needs but also to clearly understand and predict how distinct types of consumers behave when buying various products and services to satisfy their needs (Jalal, 2017). According to the study conducted by Styvén, Foster & Wallström (2017), it could be deduced that a detailed understanding of the purchasing behavior of customers by the retail stores not only offers the opportunity for sales but also assists in identifying the right promotional strategy to win consumer (Jalal, 2017). While in the past, marketing of products was focused on a product-specific point of view, the trend has changed recently to a consumer-focused approach (Bansude, & Vispute, 2022), in other words, marketing has shifted from a manufacturer to a client strategy (Rashed, 2022) where more attention is now been paid to the desires, wants, and satisfaction of the customers (Basant, 2021), than on product-specific features, hence the need for customer behavioral understanding. Although there are assertions portraying consumer behavior as a complex subject to describe, several definitions exist, but all the definitions converge on the consumer as a human being (Eric & Krishna, 2019) whose behavior involves feelings, thoughts, and behaviors that are often complicated, and full of disputes and inconsistencies (Suregka & Hema, 2022).

Prasad and Jha (2014) defined consumer behavior as a concept that involves the process by which people engage in actions that involve them to think about, want to buy, decide to buy, use, or consume goods and services. Similarly, Tshepo (2021) defined consumer behavior as

the purchasing patterns of the ultimate consumer i.e., how he chooses, purchases, and consumes products and services to fulfill his or her desires. In the same vein, Priyabrata & Dhananjay (2022), presented the concept of consumer behavior as the science of how people individually or in groups select, buy, use, and dispose of products or services, ideas, or experiences to fulfill their needs and wants. A study by Rasheed (2022) suggested that customers who use a product often leave traces of their actions that suggest how their behavior is most likely to be in the future such that the retail stores can easily extract insight that can suggest the customers' buying patterns from these actions. A buying pattern is a common way by which consumers purchase goods or utilize a service based on the frequency, quantity, duration, timing, etc. demonstrated by them. It alludes to the normal manner by which consumers purchase products or benefit administrations (Pooja, 2019). Finding predictive patterns in this granular consumer behavioral data has proven to be relevant for companies. By understanding the buying patterns of consumers, retail stores can effectively design their product and service offerings, develop a targeted marketing campaign, and customize or personalize their customer's experience to meet the needs and preferences of their target audience (Šostar & Ristanović, 2023).

2.2 Types of Buying Patterns

Often, a customer's buying behavior is largely influenced by the needs, preferences, and tastes of the consumers. While buying behavior patterns and buying habits may seem to be synonymous, it is important to understand that both are different. A habit is a natural inclination toward a specific action that over time becomes second nature, in contrast, a buying pattern shows a predictable mental design. It is expected that every customer should have distinct buying habits. However, buying behavior patterns are often a collective intertwine of activity and hence offer an exclusive set of characterization to retailers. Following the studies conducted by Ulaikere, Asikhia, Adefulu, & Ajike (2020); Arnar, Dadi, & Kyrre (2021); Tsuji, Shibata, Terasawa & Umeda (2021), it could be deduced that patterns are mostly grouped into five (4) categories: Place of purchase, Product type and Quantity, Frequency and time of purchase, and method of purchase.

- *Place of Purchase:* Customers are often prone to visit several stores to compare prices and other offers before making their purchase. In most cases, they are often inclined to

split their purchases between several stores, even if all items they need are available in one store. In such a case, one could conclude that a customer is not loyal to a store. Hence the need to study the customer behavior in terms of their choice of store, could assist retail stores in making decision about the location of the stores, the nature of their merchandise and the availability of stocks.

- *Product Type and Quantity:* Analyzing and understanding a shopping basket can present to retail store great insights into the trends and pattern of specific products or group of products in the market. The filled customer basket can also convey insights about the customer. Although several indices influence the type and quantity of products purchased by customers (perishability, availability, purchasing power, demand, unit of sale, and price), understanding these indices is relevant for decision-making.
- *Frequency and time of purchase:* The number of times a customer purchases a product or group of products is especially important in determining a customer's buying pattern. A customer may purchase a product or group of products daily, weekly, bi-weekly, or even monthly, they may also purchase products based on factors like weather conditions, seasonal variations, and location. Retail stores can use this trend to present interesting offers to customers, based on the time and frequency of the customer's purchase and also generate insights that could assist in optimizing the retail store inventory.
- *Method of Purchase:* How a customer chooses to approach his purchase desires speaks volume about the customer. While some customers prefer to walk into a store, buy a product or group of products of need and pay by cash, others may prefer to make order online and pay via credit or with debit card, some customers may prefer to walk to the store to collect their items, others may prefer to have their ordered items delivered to their doorsteps. Often the method of purchase can induce less or more spending from the customer, gathering relevant information's and understanding this information's may present an interesting strategy for the retail store to effectively influence the customer to perform a repeated purchase and with a higher value for the retail store.

2.3 Store Layout

With a great variation of products and user buying behaviors, the shelf on which products are being displayed is one of the most important resources in the retail environment. Garaus, Wagner & Kummer (2015) empirically found that appropriate store layouts increase the amount of time spent in-store by customers and most often a pleasurable experience too. This is strongly correlated to the study carried out by Yoon and Park (2018), which suggests that product layout does not only influence customers' purchase of desired products, but also the shopping experiences of customers before purchase, such as finding the product they want, and interacting with several store personnel along the way. Since store layout is crucial to optimizing the customer shopping experience and increasing revenue and profitability for the retail store, acquiring detailed knowledge, and developing an understanding of customer movement patterns inside the store has become a topic of interest to retailers with a particular focus on creating a customer-centric layout that leverages the various touchpoints of a “moment of truth” during a customer’s shopping experience (Hyunwoo, Jonghyuk, Zoonky & Soyeon, 2017).

It has been shown that changes in product placement could result in a change in the shopping path of shoppers for a given planned purchase list (Ballester, Guthrie, Martens, Mowrey, Parikh & Zhang, 2014), this often leads to stumbling on other products that may be of interest to them. Ballester et al (2014) also pointed out that, when product placement of associated products that drive traffic is changed to a different location, the traffic densities it drives change with it. For example, if a popular milk product is relocated or placed differently in the retail store, the associated traffic it drives will follow it to the new location, (Evren, Alvin & Jiefeng, 2022), hence retail stores need to know the products that drive traffic and which are often bought together, at a particular time or season to determine which products should be placed differently and those that should be placed next to each other. The subsequent sub-chapters shall elaborate on this business process.

2.4 Inventory Management

Inventory is a central management function, it is the most important asset held by many organizations, especially retail stores, representing as much as half of its expenses, or even half

of its total capital investment, thus, controlling it is critical to operational success and performance (Munyaka & Yadavalli, 2022). According to Render, Stair, & Hanna, (2016), Inventory is defined as a stored resource used to satisfy demand, current and future. It is the company's assets that are intended for use in the production of goods or services made for sale, are currently in the production process, or are finished products held for sale in the ordinary course of business (Macharia & Mukulu, 2016). In the same vein, Kamelija & Janka, (2023) defined inventory as a category of assets in which capital is invested, and which, like all assets, requires adequate management to achieve the general goals of the operation. Too little or too much inventory in a store could be very problematic for the retailer. If a retail store has too much inventory in stock, it can lead to the tying up of too many financial assets that are not in circulation, causing a lack of storage space, and leading to various defects of inventories or their destruction. Conversely, if a retail store has too little inventory it may hinder the regular operation of the retail store thus, negatively impacting on the consumer perception of the retail store (Kamelija & Janka, 2023). To achieve the target that materials or items are available when they are required while the holding costs are minimized, Mengying, (2015) suggested that Inventory planning should be done thoroughly, by considering which inventories should be kept in larger quantities and which should be reduced, however, this is easier said than done. Since it is difficult to predict the amount of goods that are needed in the market, to prevent situations in which there is a limited product in stock, Kamelija & Janka, (2023) posited that retailers should create inventories, this process of creating inventory is called inventory management. According to Khobragade, Selokar, Maraskolhe, & Talmale, (2018), inventory management, also known as materials management, is defined as the organization, securing, storage, and distribution of the right materials, in the right quantity, of the right quality, in the right place and at the right time, to coordinate, and facilitate the sales processes. It involves maintaining some stock levels at a minimized cost while improving the value-adding measures of customer satisfaction, which are useful measures of performance (Nemtajela & Mbohwa, 2017). Milusheva (2019) also posits that inventory management is the direction of activity to get the right inventory in the right place at the right time, in the right quantity, and determining when to replenish it (Mengying, 2015). It is often suggested that optimal inventories are those that are between the minimum (the minimum number of products that can satisfy demand) and the maximum inventory (the maximum number of products that can be in inventory) or stocks that

allow a full and regular supply of customers while minimizing storage and other related cost. Although There are three types of inventories: raw materials, work-in-progress, and finished goods (Kowo & Vareckova, 2023), It Is important to note that in the context of this study, inventories are considered as finished and finished goods or stocks retailers placed in their stores for sales.

2.5 Concept of Data Mining

Due to the immense progress made in the technological domain of information and communication technology, coupled with the growing customer demand, and changing needs caused by the introduction of more and new products into the market (Hajar, Elnaz, Isaline, Nhan & Mourad, 2023), massive amounts of data are continuously being collected from the day-to-day operations of businesses and stored in databases. Sağın & Ayvaz (2018) made an estimation of the amount of data that is predicted to be collected and suggested that the amount would double every twenty months. This huge amount of data collected accompanying the need for a powerful data analytical tool to unravel meaning from the data has been described as a data-rich but information-poor situation. One very pressing challenge at the front of corporations and retailers has been the heavy investment in data collection with little or no idea on how to extract significant insight from the massive repository of customers information to achieve a competitive edge (Praveena, Jahnavi, & Sunayana, 2022). Since customer purchase behavior analysis is crucial to inform the retailer about the purchase intentions of their customer while aiding business strategic business design (Ebrahimi, Hamza, & Zarea, 2020), many companies are becoming interested in extracting useful information or discovering knowledge through data mining (association mining), about consumer purchase behavior from the large amount of customer data that have been collected from the customers and store, to boost their profitability. For example, the uncovering of interesting associations from the huge amounts of business transaction records to create catalog design, loss leader analysis, cross-marketing and other business decision-making processes (Arpitha, 2017). This process of discovering interesting associations is what is called the data mining process.

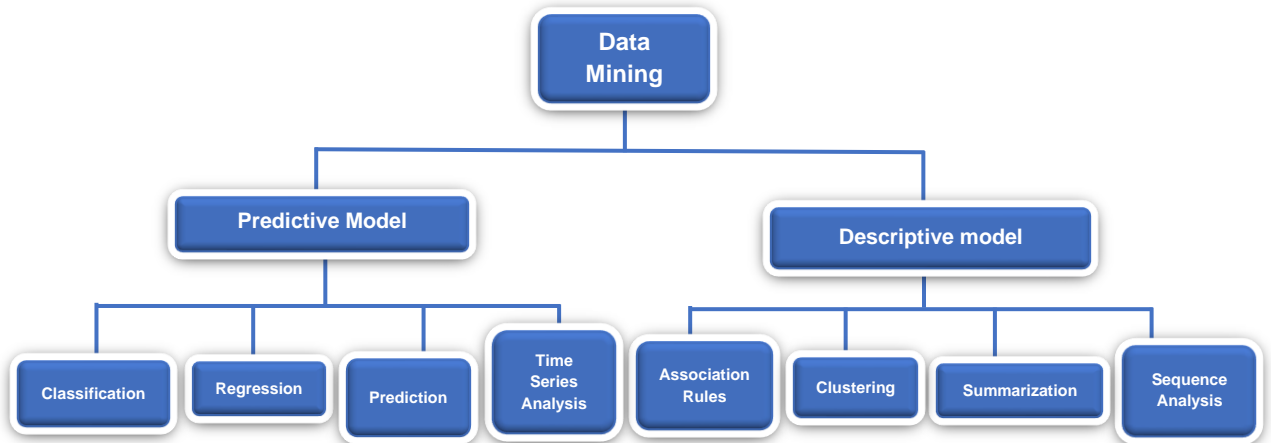
Data mining is a knowledge discovery process. Many other terms convey similar or slightly different meanings to data mining such as knowledge extraction, knowledge mining from

database, data archaeology, data/pattern analysis and data dredging. Data mining is useful in dealing with large amounts of stored dataset to uncover the desired information and knowledge needed for business decision making. According to Suntoro (2019), data mining is conceptualized as a process of extracting data that is considered useless into useful and meaning information or knowledge from large source of data, it is a process of analyzing large data sets to uncover significant patterns and rules and the analysis of (often large) observational data sets to find unprecedented relationships while summarizing the data into novel ways that are both understandable and useful to the data owner (Arpitha, 2017). while others suggest that data mining is only a mere essential step in the process of knowledge discovery, Nithya, Sivapriya, and Rajshree (2020), redefined data mining from a broad perspective as the process of uncovering patterns and knowledge that are interest from a large amount of data. The source of the data could include the web, data warehouses, databases, other information repositories, or dynamically streamed data into a system. This definition suggests a broad application of data mining in various other fields of specialization like – Law, Astronomy, Medicine, Industrial process control, etc. (Istrat & Lalić, 2017). And various domains like market analysis, production control, science exploration, fraud detection and customer retention (Arpitha, 2017). Howbeit, this study has only focused interest on market basket analysis, one of the promising and widely used data mining techniques in the retail store industry for increasing sales through appropriate understanding of customer purchasing patterns.

2.6 Data mining techniques

According to Kanimozhi (2019), data mining is a method of analyzing a large relational database from different angles or perspectives to gain useful insight that can be harnessed to increase the revenue of businesses. It enables backend processors to analyze data from diverse dimensions, categorize and summarize the association identified. Categorically, data mining processes and methods can be categorized into different ways. Some methods and processes are traditional and already established while some are relatively recent. Data mining goals can be fulfilled by modeling it either as predictive or descriptive (Gangurde, Kumar & Gore, 2017). Below we briefly review this nature and categorize the data mining techniques.

Figure 2: Data Mining Model



Source: Authors elaboration based on (Ünvan, 2021)

2.6.1 The Predictive Model

The predictive model is a model that makes predictions about the values of data using known results found from different datasets (Gangurde, Kumar, & Gore, 2017). It uses historical records to predict the future outcomes of values in a record or in a database with already-established answers. The task of the predictive model includes times series, classification, regression, and prediction (Laxmi, Kavitha & Nagarani, 2017).

- **Classification:** This approach is often considered the best understood of all data mining approaches in the predictive model. The data mining goal in classification involves the assignment of items in a collection or categories or classes for instance high class, medium class, low class, or in another instance into roll number, student names, father name, etc., there are three common characteristics of the classification approach and they include: (i) A supervised learning (ii) categorical nature of the independent variable (iii) The model built can assign new data to one of a set of well-defined classes.
- **Regression:** This approach creates a model to predict attribute values for new cases by analyzing the dependence of some attribute values on the values of other attributes in the same item. It is a supervised learning approach and when being used, it could predict the value of a given (continuous) variable based on the values of other variables in the data, assuming a linear or nonlinear model of dependency. For instance, to forecast the

possibility of fraud for new transactions, given a set of credit card transactions the regression approach could be used.

- **Prediction:** This approach is purposed to determine future outcomes rather than current behavior. Its functionality is based on predicting some unknown or missing attribute values based on a known attribute value, or other information. The output of this approach can take the form of categorical or numerical values. For instance, making a forecast of the sales value of next week based on the current information at hand or predicting the probability that a certain transaction is fraudulent based on a prediction model of credit card transactions.
- **Time Series Analysis:** This approach uses one or more time-dependent attributes to predict problems. The approach is usually associated with predicting outcomes of numeric value such as prices of individual stock in the future.

2.6.2 The Descriptive Model

The goal of the descriptive model has been to identify patterns or associations between the attributes presented by data. It divides a large amount of data into smaller ones so that patterns and the correct ones are easily identified from them. The descriptive model does not predict new properties but rather explores the properties of the data that has been examined previously and it is generally purposed to enable a specific understanding of the system, what it does, and how it does it. This model includes tasks like Association Rules, Clustering, Summarizations, and Sequence Analysis (Gangurde, Kumar, & Gore, 2017).

- **Sequence Analysis:** This task is sequential in its approach and assists in determining patterns that are sequential in nature. The main goal of this approach is to examine associations between items over time based on a time sequence of actions. In sequence analysis, item analysis is based on the items purchased overtime in contrast to market basket analysis in which items analysis is based on items purchased at the same time in a transaction hence, basically the utmost relationship in sequence analysis is based on time.
- **Summarization:** This approach could be associated with simple descriptions like Median, Mode, Median, Standard Deviation, and Variance. It maps out data into subsets and can be used as a summarization approach.

- **Clustering:** This approach is appropriate for identifying groups of items with similarities. By putting comparable data points together, clustering can help assign points that are like each other in the same group and as different from the points in the other groups. A simple example is the identification of a subgroup of customers that have similar buying behavior from a large data set of customers.
- **Association rules:** This is probably the most interesting approach in the descriptive model designed to uncover frequent patterns, associations, and correlations from data sets that are available from different databases. The main thrust of the association rule is underlined in the establishment of relationships between items that exist in the market such that the presence of one pattern suggests the presence of another pattern. It is simply a cause-and-effect approach i.e., to what extent one item is related to another. A typical example of an association rule is market basket analysis used to explore possible combinations of potentially interesting product groups.

2.7 Association Rule Mining (ARM)

Association rule mining (ARM) is an important branch of knowledge-mining research. It was first applied in the buying performance analysis of a supermarket using the Apriori algorithm propounded by Agrawal and Srikan in 1994. From then onwards, association rule mining has not only played an important role in commercial data analysis but also in finding interesting associations and patterns in many other fields (Yuan, 2017). Association rule mining forms the basis of market basket analysis (Hamid and Khafaji, 2021), and is purposed to discover relative correlations from commercial transaction data sets (Najafabadi, Mahrin, Chuprat, & Sarkan, 2017). The ideology behind association rule mining is the examination of all possible if-then relationships between products and product groups and the selection of only the most probable indicators of the dependency relationship between products. Although the association rule examines the if-then rules, the rules are determined from transactional data, and are probabilistic in nature, unlike the if-then rules of logic. The term antecedent usually represents the “if” part of the rule and the consequent represents the “then” part of the analysis. Antecedent i.e., the “if” part, and consequent i.e., the “then” part is often a set of items called the item sets or group of products that are disconnected or disjoint i.e., they do not have any items of commonality or

relationship altogether. Nithya et al (2020) Suggested that association rules highlight the relationship between disjoint item sets and are used to discover frequent associations and patterns among sets of items in a database that is relational or transactional or other data repositories. A typical form of this relationship is given thus: $X \rightarrow Y$ i.e., Y happens if X does and vice versa.

According to Khasanah (2020), one well-known and appreciated data mining method is association rule mining used to identify associative linkages amongst multiple data sets. It often works by establishing rules. These rules are formed by analyzing frequent patterns and using the support and confidence factors. Martinez and Escobar (2021) have defined association rule mining as a typical data mining process in which items frequently occurring together in transactions are found inside each dataset. In the same vein, Dogan (2023) described association rule mining (ARM) as a widespread and effective approach to learning consumer preferences. It investigates customer-purchase item sets, or recorded transactions and generates product rules that define an association between the products to enable recommendations to customers based on what is relevant to them. Oyeboode and Agbalaya (2022) also posited that the rules produced by association, analysis need to be interpreted with caution. The conclusion from an association rule may not necessarily imply causality but rather it may suggest a robust co-occurrence association between the items in the rule's antecedent and consequent. Usually, causality entails linkages that occur across time and requires knowledge about the cause-and-effect properties in a data set.

In general, association rule mining assists in excavating interesting correlations, patterns, and associations among products or groups of products in a data set. It is composed of two primary metrics, support, and confidence, each having a minimum threshold that can be influenced by the user or domain consultants for optimum results and only the rules that satisfy the minimum support and confidence thresholds are relevant and useful (Praveena et al, 2022). Since association rules mining identifies interesting associations amongst item sets, they can assist retailers in making reasonable business decisions by understanding the customer's buying habits (Wang, Zeng, Zhou, Li, Iyengar, & Shwartz, 2018). A broader view of the association rules is in the form: $X \rightarrow Y$, where X and Y are item sets and $X \cap Y = \emptyset$. If X occurs, Y is probably going to occur (Neysiani, Soltani, Mofidi & Nadimi-Shahraki, 2019). Association rule derivations

from databases usually take place in two stages (Sağın, & Ayvaz, 2018; Neysiani et al, 2019), as seen below.

- ***Finding Frequent Itemset:*** This stage entails the identification of frequently purchased items from the basket of customers and finding associations, casual structures, or correlations among products or groups of products in a transactional data set. The main thrust underpinning this stage is the minimum support threshold which helps to determine this frequency in the item sets. Hence for this stage to be complete, the minimum support threshold must be satisfied.
- ***Establishment of strong association rules from frequent items:*** At this stage, the main object of interest is discovering associations and correlations among the products or groups of products that customers place in their market carts. The association rules are generated from the frequent item sets identified in the first stage which defined the minimum trust thresholds. Here the association rules are established from the frequent items set in the first step which establishes the minimum trust thresholds. This stage usually can be satisfied only if the first stage is completed. The main thrust of interest in this stage is the minimum confidence threshold which must be fulfilled for this stage to be completed. Once the identification of the frequent item sets in the first stage is satisfied it is often easy to extract the highest confidence rule to formulate the association rule.

Market basket analysis is a typical example of the association rule. Its name stems from the practice of having the customer fill up their shipping basket with all they purchase while shopping in the supermarket. The analysis investigates patterns in customers' purchases by discovering associations among various products that customers place in their shopping carts. This way, retailers devise means to expand marketing strategies by gaining insight into which items are frequently purchased by customers. Market Basket Analysis (MBA) also known as association rule mining or affinity analysis, is a data-mining approach that originated in the field of marketing but has recently found widespread applications in other fields., such as bioinformatics, education field, nuclear science, cyber security, etc. (Alqahtani, 2022). The ultimate goal of an MBA in marketing is to provide relevant information and insights to retailers that could enable an understanding of the purchase behavior of the customer to influence correct and effective decision making (Kaur & Kang, 2016) while market basket analysis (MBA) is

applicable in many fields and defined differently by various researchers based on their unique area of studies it has mostly been considered an essential part of retail businesses.

Ghassani, Jamaludin, and Irawan (2019) defined market basket analysis (MBA) as a technique of analyzing consumer behavior to find out specific purchasing patterns by grouping items that often appear or are purchased simultaneously in transactions. Nithya et al (2020) describe MBA as a knowledge-mining technique that is generally used to uncover customer patterns such that if a customer buys a certain group of items, then customers are likely to buy another group of items. It is a very useful technique for identifying correlations of items and co-occurrence of items in consumer shopping carts. Similarly, Rana and Mondal (2021) defined market basket analysis as a mining methodology that is observational in nature and used to investigate consumer buying patterns in retail supermarkets. It examines the customer's cart and investigates the association among products to make useful decisions regarding the design of store layouts, as well as various strategic plans, and merchandising decisions that have a great impact on retail marketing and sales. The first step in market basket analysis is the frequent item sets mining. The association rules mining uncovers the co-occurrence among products by looking at what products the customers frequently purchase together. In addition, Ghous, Malik & and Rehman (2023) defined market basket analysis (MBA) as a data mining technique that assists retailers in defining the customer's buying habits to make new marketing decisions that satisfy the buyer's frequently changing desires and expanding needs. All the definitions above are summarized into one key objective of the retailer which is to improve market and sales plan efficiency.

There are several benefits of market basket analysis the most common ones are price determination, store layouts, and product display optimization, computation of product associations, designing and personalizing promotions, determining market trends for inventory management purposes. Prediction of future demands of products, merchant development, designing of marketing strategies sales influence highlights, improved decision making, gaining marketing shares, and increase the profitability of the retail store (Ankita & Shobha, 2022; Alghanam, Al-Khatib & Hiari., 2022). MBA can specifically be used to increase sale volume by arranging or displaying complementary products closer to each other on the market shelf or by providing steep discounts for such items. The engine behind most recommendation is the

market basket analysis or recommendation engine which collects information about people's habits and understand the reasons behind the scenario “When a customer buy bread, they are usually also going to buy butter or cheese” (Kanimozhi, 2019). An early illustration of this is the legend of lager and diapers, which was first publicly disclosed approximately two decades ago in a 1998 Forbes magazine and serves as an early example of this. "A retail store placed all its checkout counter information into a giant advanced distribution center and set the circle drive rolling. A completely unexpected association between diaper deals and beer emerged. It seems obvious that young fathers would go to the shop in the middle of the night to grab some Bud Light and Pampers. The store combined the disparate items after receiving the disclosure and deals exploded. Thus, a new application, of the market basket analysis was observed: the arrangement of goods within a store space, as the case may be (Anurag, 2021). Through Market basket analysis various association rules could be established from transactional data of the retail database. Although not all rules are interesting, strong rules identified from the transactional data using measures of interest such as support, confidence, and lift are often considered useful when employing market basket analysis for associate rule mining (Aakanksha, Aditya & Ramesh, 2023).

2.8 Measures of Interestingness

The concept of an interestingness measure is an essential aspect of data mining or the extraction of interesting patterns from a database irrespective of the kind of pattern that is being identified. The measures are employed to identify interesting patterns and rank them based on the user's needs (Prajapati, Garg & Chauhan, 2017). It is often established that a pattern interesting can be validated if it fulfills the following conditions: if (1) humans can easily understand it (2) can be validated on novel or test data with some degree of certainty (3) potentially usable, and (4) if it is a new discovery (Selva, 2018). Since not all association rules that are generated by the association mining process are useful or usable to the organization, the need then arises to filter out the uninteresting rules for effective decision-making. While the most widely used interesting measures, support, confidence, and Lift ratio have been established, different kinds of knowledge may require different interestingness measures (Prajapati et al, 2017).

- **All-confidence:** This measure explains that all the rules generated from item X have at least a confidence of all confidence (X). All-confidence means that all rules that can be generated from item set X have at least a confidence of all-confidence(X). It satisfies the property of downward closed closure and hence, can be used effectively for interesting association rule mining. All-confidence is defined on item set (not rule) as

$$\text{All - confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Max}(\text{Support}(X), \text{Support}(Y))}$$

- **Cosine Similarity:** The cosine similarity emphasizes on the closed unit interval that no constant value is achieved when the antecedent and consequent are statistically independent of one another. i.e., A value of cosine (X/Y) close to 1, indicates more transactions containing item X also contain item Y and vice versa. In the same vein, if a value of cosine (X/Y) close to 0 indicates more transactions containing X items without Y and vice versa. Cosine similarity is defined as:

$$\text{Cosine}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\sqrt{\text{Support}(X) * \text{Support}(Y)}}$$

- **Conviction:** This metric measures the strength of the implication of the rules from a statistical point of view and compares the probability that item X appears without item Y if both are dependent on the actual frequency of the appearance of X without Y. Conviction is given thus:

$$\text{Conviction}(X \rightarrow Y) = \frac{1 - \text{Support}(Y)}{1 - \text{Confidence}(X \rightarrow Y)} = \frac{P(X) * P(\hat{Y})}{P(X \cup \hat{Y})}$$

Where P(Y) is the likelihood that Y does not appear in a transaction

- **Validity:** The difference between the probability of item “X” and “Y” occurring together and the occurrence probability of “Y” without “X” occurring in database D is what is known as validity. Because the value range of P(XY) and P(XY) is [0,1], the value range of validity is obviously [-1, 1]. Validity is given thus:

$$\text{Validity}(X \rightarrow Y) = P(YX) - P(\bar{X}Y).$$

- **Improve:** Based on the description of the defects of the traditional interestingness measurement method, the Improve was proposed as a new interestingness measurement rule. Improve defines the difference between the conditional probability $P(Y|X)$ and the probability of “Y”. Improve is given thus:

$$\text{Improve } (X \rightarrow Y) = [P(Y|X) - P(Y)].$$

- **Support:** The support measures the degree of frequent occurrence of collection items in an association rule. Sometimes, it is expressed as a percentage of the total number of records in the database and represented in metrics of 0 to 100. It is often suggested that a support rule with a low score, especially lower than 10, is practically uninteresting from a business point of view and sometimes eliminated. Simply put, support is the default popularity of a product or group of products, and it is estimated by dividing the number of transactions containing a specific products or groups of products by the total number of transactions. Support is given thus:

$$\text{Support } (X) = \frac{(\text{Transactions containing } (X))}{(\text{Total Transactions})} \text{ (Gurudath, 2020)}$$

- **Confidence:** An index reliability of association rule is called Confidence. It defines the conditional probability $P(X|Y)$ that is, the probability of event X happening given that event Y already happened, and details the proportion of transactions containing X, that also contain Y. Both Confidence and Support are similar given that they are expressed as a percentage metric ranging from 0-100. However, Support measures the frequency of occurrence of X and Y in the transactional record of a database, while Confidence measures the accuracy of the rule (Arpitha, 2017). The confidence metric is given thus:

$$\text{Confidence } (X \rightarrow Y) = \frac{(\text{Transactions containing both } (X \text{ and } Y))}{(\text{Transactions containing } X)}$$

- **Lift:** Originally known as interest (Kanimozhi, 2019) is a metric that measures frequency X and Y together if both are statistically independent of each other. The lift focuses its functionality on measuring the strength of a rule over the randomness of the occurrence of item X and item Y together (Swati, Mahesh & Joshi, 2018). Mathematically, Lift compares the confidence for all rules divided by the confidence benchmark. It is often suggested that

a lift with a higher value has a greater strength of association (Suharjo & Wibowo, 2020) i.e. When the lift of a rule is higher than 1, it means that the lift of a rule is greater than 1, it means that when item X is purchased the chancing of purchasing item Y also increases. The support is expressed as:

$$\text{Lift}(X \rightarrow Y) = \frac{(\text{Confidence}(X \rightarrow Y))}{(\text{Support}(Y))} \quad (\text{Oyebode \& Agbalaya, 2022})$$

A typical evaluation of the Lift association rule discovered from the study of Martinez & Escobar, (2021) and, Praveena, et al, (2022) can be expressed as follows:

- If $\text{Lift}(X \rightarrow Y) = 1$, then the item occurrence of “X” is independent of the item occurrence of “Y” and vice versa, and X and Y occur together only by chance
- If $\text{Lift}(X \rightarrow Y) > 1$, then the item occurrence of “X” influences the probability that the item occurrence of “Y” will happen, which means X and Y occur more often than random
- If $\text{Lift}(X \rightarrow Y) < 1$, then the item occurrence of “Y” influences the probability that the item occurrence of “X” will not occur, which means x and y occur together less often than random

A typical evaluation of the support and confidence association rule can be given thus (Praveena, et al, 2022):

- If a rule has a high support and confidence score, this rule is obvious and tends to be uninteresting.
- If a rule has a reasonably high support score but a low confidence score, the sale of items X and Y may be higher than the support threshold, but not all transactions that contain item X also contain item Y. Such a low-confidence rule tends to be uninteresting.
- If a rule has a low support and confidence score, such a low-confidence rule tends to be uninteresting.
- If a rule has a low support but a high confidence score, such a rule tends to be interesting.

2.9 Searching Frequent Itemset

For decades, many association rule mining algorithms have been in use for different mining functionalities. Some of the widely known and earliest-used association rule mining algorithms

include the Apriori proposed by Agrawal and Srikant, (1994), FP-Growth by Han Pei & Yin, (2000), Eclat by Han and Kamber (2001), K-Nearest Neighbor by Larose and Larose (2005), Naive Bayes by Kamruzzaman and Rahman (2010), K-Apriori by Annie and Kumar (2011), and K-Means by Liu et al (2014),. This study is based on the widely adopted algorithm, Frequency pattern growth (FP-Growth) Algorithm to address market basket analysis and identify address market basket analysis and identify products that combine well together for effective decision-making. The Apriori algorithm (the first developed algorithm) shall be discussed also as a prelude to the FP-Growth algorithm.

2.9.1 Apriori Algorithm

Apriori is a popular, essential, and scalable approach for mining frequently occurring item sets and association rules. It was considered the first association mining algorithm and was introduced by Agrawal and Srikant in 1994. The name comes from the word “prior”, and it is seen as the core of various algorithms for data mining problems and analyses of data sets to determine which product combinations occur together frequently (Joshi, 2018). Many retail stores such as shopping malls, general stores, grocery stores, and so on, employ the Apriori algorithm for transactional operations in real-time applications by collecting the items purchased by customers over time so that frequent items can be generated (Praveena et al, 2022). Apriori uses an approach known as a level-wise search, where k -itemsets are used to explore $(k+1)$ -itemsets and $(K+2)$ itemsets to explore $(K+1)$ and so on. The approach works such that, the first database is scanned to identify all the frequent items and then counts each of them to capture the minimum support threshold item sets. It is required to scan the entire database until the algorithm cannot find frequent items any longer (Sagin & Ayvaz, 2018). Minimum support and confidence thresholds are required by the Apriori algorithm, first to check if the items are higher than or equal to the minimum support, and second to check for frequently used item sets. The Apriori algorithm is applied as follows (Idris, Ardhana & Manapa, 2022):

- First, scan the database to find a 1-item set.
- Second, expressed the frequent item sets found among the 1-itemsets as K_1
- Thirds, create 2 item sets using K_1 .
- Fourth, create K_2 by determining the frequent occurrences among the 2-itemsets.

- Fifth, repeat the process so that K3, K4, K5, and so on can be found.

Based on the Apriori characteristics, K-item sets that do not satisfy the minimum support thresholds are eliminated from each iteration and only the item sets that satisfy the minimum threshold would be moved to the next cycle (Anas, Rumui, Roy & Saputro, 2022). By employing the Apriori algorithm, the association mining process gives interesting rules however, as the size of the database increases, the performance of the result decreases this is contingent on the fact that the Apriori algorithm usually scans the entire database every time it scans the transaction to identify frequent items (Dilrukshi, 2021).

2.9.2 Frequent Pattern Growth Algorithm (FP-Growth)

The Frequent pattern growth algorithm also known as the FP-Growth Algorithm was propounded by Han in 2000 as a scalable and efficient method that mines frequent item sets without costly process (Alyoubi, 2020). It could be said that the FP-Growth is an improved version of the Apriori, in other words, a development algorithm from Apriori. It came into the limelight due to the two major demerits of the Apriori algorithm. First, the generation of a huge number of candidates sets, and second, the continuous scanning of the database to identify frequent item sets (Anas et al, 2022). When the Apriori algorithm is employed, large databases will often require hundreds of scans which often lead to huge time loss, using the FP-Growth the database is first compressed into a structure that looks like a tree called the FP-tree, containing the association information of the item (Nasyuha, Jama, Abdullah, Syahra, Azhar, Hutagalung and Hasugia, 2020), which is then divided into the conditional data structure, where each of these databases is mined independently and is linked to a common object. The FP-Growth algorithm solves the problem of scanning the transactional records multiple times to find large common item sets by providing a solution that searches the minors and then merges the suffixes. (Hu, Liang, Qian, Weng, Zhou, and Lin, 2021).

By applying the FP-tree structure, the FP-Growth algorithm can obtain frequent itemsets directly, so the FP-Growth algorithm is faster than in the Apriori algorithm system (Hu, Liang, Qian, Weng, Zhou, and Lin, 2021). Making a comparison of FP-Growth and the Apriori algorithm, Alyoubi (2020) suggests that FP-Growth is much more powerful in the case of efficiency. By applying the FP-Growth algorithm in a step-by-step process they can help

remove unnecessary data and improve the performance of the overall process. Similarly, in terms of speed, although the accuracy of Apriori rules is higher than FP-Growth, due to repeated scanning, the speed is slower (Salam, Zeniarja, Wicaksono, & Kharisma, 2018). Additionally, the FP-Growth algorithm can find more association rules in the process than the Apriori algorithm (Zaki and Meira, 2018),

A typical FP-Growth algorithm execution process as suggested by (Shabtay, Fournier-Viger, Yaari & Dattner, 2021) is given thus:

1. First, scan the database and identify items that are equal to or greater than the threshold
2. Second, list the support values of frequent item sets in order of size (big to small).
3. Third, Create a tree with only roots
4. Fourth, scan the database again, and for each sample.
 - a) Add to the tree items taken from the sample ensuring only the frequent items listed in step 2 are added.
 - b) Repeat the step until all the samples are processed

Table 2:Compressed Comparison of Apriori and FP-Growth Algorithms

Metrics	Apriori Algorithm	FP-Growth Algorithm
Frequent pattern	Pattern selected based on fulfilling the minimum support criteria	Pattern selected using the conditional FP-tree growth construct
No. of Scan	Data is scanned each time a candidate set is generated.	The database is scanned at most twice
Speed	Speed is slower due to repeated scanning time	High speed because data is scanned two times only.
Memory	Saves singletons, pairs, triplets, etc.	Stored in compact versions
Techniques	Breath first Search	Divide and conquer
Application	Used in case of large data	Best used for closed item sets

Source: Authors elaboration based on Ünvan (2021).

2.10 Empirical review

Various topics on Association mining and consumer buying behavior have been extensively discussed conceptually and empirically in previous studies. Hence the table below presents a summary of the related literature reviewed in descending order of year.

Table 3: A summarized review of related studies in descending order of year.

S/N	Author, Year	Purpose of study	Dataset	MBA Algorithm	Result
1	Idris, Ardhana & Manapa (2022)	To make comparisons between the Apriori, Apriori TID, and FP-Growth algorithms in determining consumer transaction behavior	Retrieved from Kaggle.com with three attributes, namely, i. member number, ii. Date, iii. Item description.	Apriori, Apriori TID, and FP-Growth Algorithm	The FP-Growth was found to have the best performance among the algorithm compared but consumes more memory compared to the other algorithms.
2	Oyebode & Agbalaya (2022)	To optimize the market basket using a sales pattern	Customer transaction records of a supermarket of 20 different products and 7500 transactions	Apriori Algorithm	The support, confidence, and lift threshold gave the association rules from the customers
3	Praveena Jahnavi & Sunayana (2022)	To combine a hashing function with the Apriori algorithm, for optimization	Website grocery data, including a list of the things that customers purchased	Apriori Algorithm	When the hash table was employed, the temporal complexity of the Apriori algorithm was reduced.
4	Tripathi & Pandey (2022)	To analyze the market basket of cosmetic products.	Copies of purchase bills at the Bachat Store, Vishal Mart, Buy Chance, and Army Canteen in Pithoragarh. A total of 50 transactions. 5 transactions- were considered invalid	Using the RStudio software, the Apriori Algorithm was implemented.	The analysis produced three association rules depicting a positive relation between antecedents and consequents
5	Arora, Bhateja, Goswami, Kukreja & Rajput (2022)	To identify a trend for customers buying patterns in 3 countries	13 months daily order sheets from a Super Mart. December 2010 - December 2011.	Apriori Algorithm	Generated three results for each of the 3 countries (France, Germany, and the UK)

		(France, Germany, and the UK)	541911 rows and 8 columns.		based on a high Lift value
6	Martinez & Escobar (2021)	To predict which products are sold together and will make the most profit.	A total of 1565415 records of sales transactions from 2008-2017	F-PGrowth algorithm using the orange tool	The result shows that bed sets, headers, mattresses, mattresses, night tables, and pillows have a strong association with each other.
7	Rao, Kiran & Poornalatha (2021)	To find frequent diseases that occur together in a certain place	Online repository of healthcare Records from different countries. Information about patients' diseases, year of infection, etc.	Apriori approach	Three frequent disease item sets are found. 1 (Amoebiasis & Botulism) 2. (Amoebiasis, Botulism, Brucellosis) 3. (Amoebiasis, Botulism, and HIV)
8	Rehman & Ghous (2021)	To provide direction for future research by reviewing MBA using deep learning and association rule	70 research papers in the field of retail	Deep learning algorithm	27 papers based on association rules of MBA; 23 papers based on offline datasets; 4 papers based on online data.
9	Unvan (2021)	Using WEKA software to analyze sales data from any supermarket received from the Vancouver Island University website.	225 products	FP-Growth Algorithm	The top 10 rules obtained were according to the conviction value
10	Anurag (2021)	To examine the customer buying pattern during the lockdown	Open access market basket optimization dataset on Kaggle. aggregate of 7501 exchange records	Apriori Algorithm	Specific items are related to each other more and association between them along with the importance.
11	Rana & Mondal (2021)	To extract the seasonal associations among products by uncovering the hidden seasonal item sets.	Bangladeshi supermarket dataset that contains 3 variables namely,	FP-Growth Algorithm	An additional 642 seasonal frequent patterns were generated with the 636

			transaction number, transaction date, and list of items for a customer transaction. 99760 transactions for 2382 unique items.		overall frequent patterns for the user defined support threshold of 0.1%
--	--	--	--	--	---

Source: Author (2023)

2.9 Conclusion

While understanding consumer buying behavior has become a major issue of concern for many retailers, it has been established that association rule mining is one option outside to conventional way to understand customers that retailers can resort to. In this chapter, the study discussed the concept of buying patterns, and how understanding the buying patterns of customers can help retailers make effective decisions. The study has also discussed the term market basket analysis as one of the techniques of data mining that is useful in uncovering associations among the large transactional data set of a retail store. In the following chapter, we shall discuss the methodology this study would adopt to achieve its set goal.

CHAPTER THREE

METHODOLOGY

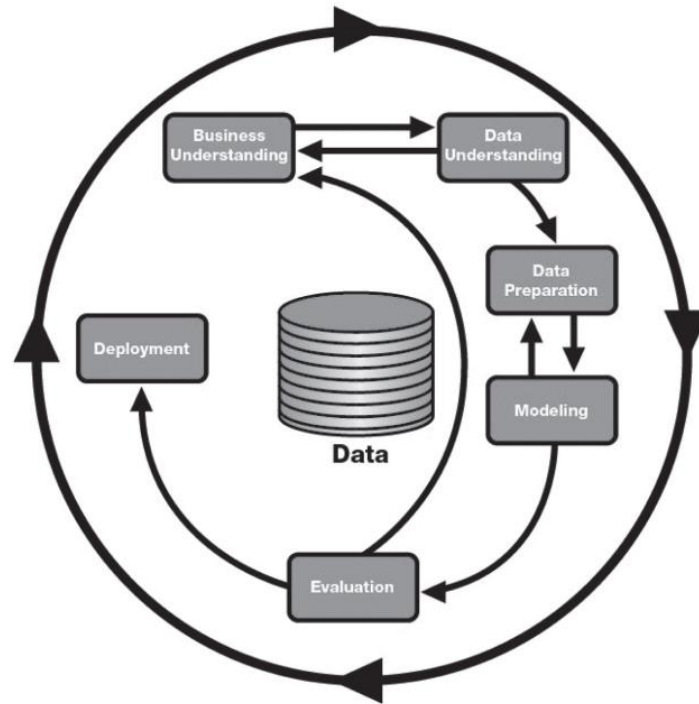
3.0 Introduction

This chapter presents the data adopted for the study and gives a proper description of the data set. The study explored the data set using the FP-Growth Algorithm of the association ruling mining to uncover patterns amongst products in the market basket of the transactional data set of customers in a gift retail store in France. To achieve the study purpose, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was employed, which is a development of the KDD methodology. It is worth stating that the study did not comply strictly with the sequential structure of the CRISP-DM, but the overall analysis was conducted in the light of this methodology.

3.1 CRISP-DM

Today, several data mining methodologies exist; the most widespread and frequently used of them is the CRISP-DM. The CRISP-DM methodology suggested by Chapman (2000) was adopted for this study. This methodology has six phases, namely, Business understanding, Data understanding, Data preparation, Data Modeling, Evaluation, and Deployment, with each phase also divided into various tasks. While this methodology follows a particular order or sequence of stages, it allows for some flexibility and iterations. The general sequence of phases is shown in the figure below.

Figure 3: Steps of CRISP-DM Methodology



Source: (Chapman, 2000)

3.1.1 Business Understanding

The first phase of the CRISP-DM methodology is the business understanding, it entails understanding the project objective and demands from the business point of view and transforming this knowledge into a characterization of a data mining problem, and the main goal is to construct a preliminary plan to achieve the business objective. This study has reviewed several works of literature in the study area and gathered relevant information about the current trends in customer buying patterns and association rule mining to gain background knowledge about the current business problems. Since the data employed for the study is an open-source dataset without any prior understanding of the specific business problem of the transactional data, this study focuses on generating association rule related to a gift store retailer, but in some cases applicable independent of the gift store retailers, with the aim that the conclusion this study arrived at, would be leveraged by gift store retailers' and other practitioners to enhance their business operations in their respective domains.

3.1.2 Data Understanding

The second phase of the CRISP-DM methodology is data understanding. This phase begins with the initial collection of data and proceeds with other activities that relate to the data discovery processes such as identifying data quality problems, generating insights about the data, and recognizing interesting subsets to form hypotheses that can be considered to uncover the hidden information within the data collected for the study.

In this study, data has been retrieved from an open-source dataset created by the UCI Machine Learning Repository that contains 8557 sales transactions dated from 01-12-2010 until 09-12-2011 (UCI, 2018). The data set was downloaded from github.com, a predictive modeling platform used for competitions that store open-source datasets uploaded by companies or users (GitHub, 2019). The data was downloaded and unzipped manually in .csv format in the working directory of the project. The data set contains transactional data of a French-based gift store retailer that sells unique all-occasion gifts to wholesalers and end users. The data variables are explained below:

- Invoice No: (ID - Categorical) Nominal. Unique transactional identifier. Code starting with the letter “C” indicates a cancellation and should be discarded.
- Stock Code: (ID – Categorical) Nominal. Unique Product identifier
- Description: (Categorical) Nominal. Product name.
- Quantity: (Integer) Numeric. The quantities of each product per transaction.
- Invoice Date: (Date) Numeric. Invoice Date and Time. The day and time each transaction was generated.
- Unit Price: (Continuous) Numeric. Product price per unit in Euro.
- Customer ID: (Categorical) Nominal. Customer number. A 5-digit integral number assigned uniquely to each customer.
- Country: (Categorical) Nominal. Country name. The name of the country where each customer resides.

The first 10 rows of the original datasets are shown in the table below:

Table 4: First 10 rows of the original data set

Invoice No	Stock Code	Description	Quantity	Invoice Date	Unit Price	CustomerID	Country
536370	22728	ALARM CLOCK BAKELIKE PINK	24	12/1/2010 8:45	3.75	12583	France
536370	22727	ALARM CLOCK BAKELIKE RED	24	12/1/2010 8:45	3.75	12583	France
536370	22726	ALARM CLOCK BAKELIKE GREEN	12	12/1/2010 8:45	3.75	12583	France
536370	21724	PANDA AND BUNNIES STICKER SHEET	12	12/1/2010 8:45	0.85	12583	France
536370	21883	STARS GIFT TAPE	24	12/1/2010 8:45	0.65	12583	France
536370	10002	INFLATABLE POLITICAL GLOBE	48	12/1/2010 8:45	0.85	12583	France
536370	21791	VINTAGE HEADS AND TAILS CARD GAME	24	12/1/2010 8:45	1.25	12583	France
536370	21035	SET/2 RED RETRO SPOT TEA TOWELS	18	12/1/2010 8:45	2.95	12583	France
536370	22326	ROUND SNACK BOXES SET OF 4 WOODLAND	24	12/1/2010 8:45	2.95	12583	France

Source: GitHub (2019).

3.1.3 Data Preparation

This phase of the methodology consists of all necessary activities required to set up the final dataset from the initial raw data, for modeling purposes. The process is usually carried out repeatedly and in no specific order. Since different datasets are likely to reveal new problems, selecting the appropriate data mining algorithm tools and techniques at this phase is essential to produce the best results from the available data. Often, distinct subsets of data are expected to display dependencies among different categories of attributes, it is also possible not to use all the variables for analysis and the modeling process this is because most variables are often not relevant or have missing values. The utmost goal of the data preparation phase is to filter the dataset to ensure only the datasets needed for the analysis are left and irrelevant data are either removed or replaced. In this phase, suitable attributes for prediction are also determined.

3.1.4 Data Modeling

The modeling phase includes deciding on an appropriate modeling technique or building the model to run the association rule mining process. The modeling phase also involves the selection of appropriate techniques for problem-solving and the refinement and assessment of the model whenever necessary to meet the requirements and other constraints and to ensure the best results are generated from the analysis. It also involved the process of parameters calibrated to optimal values. In this study, the FP-Growth algorithm of association rule mining a descriptive approach was employed.

3.1.5 Data Evaluation

This phase performs a thorough examination of the model developed and revises the steps executed in the modeling phase to create and compare the model with the established business objectives. The data evaluation phase could be likened to the thesis, antithesis, and synthesis methodology. Where a proposition is made by the thesis, the antithesis negates the proposed thesis, and the proposition is then synthesized to solve the conflicts. The steps in the data evaluation phase include the result evaluation, process review, and determination of the next steps. In this study confidence, support and lift were the indices used for the evaluation of the association rule mining techniques and algorithm.

3.1.6 Data Deployment

This phase is the last in the CRISP-DM methodology. In this phase, the model knowledge acquired from the analysis is transferred for implementation purposes. This phase involves the creation of visuals or reports and the actual application of a repeatable data mining process throughout the organization. This study aims to generate results that satisfy the questions raised in the study.

3.2 Analytical tools

This study employed Rapid Miner as the tool to perform the data mining simulation using the FP-Growth algorithm and association rules method (Brilliant, DwiHandoko & Sriyanto, 2017). According to Elvitaria & Havenda (2017), RapidMiner is an open-source application or

software that employs data mining, text mining, and predictive analysis methodologies to perform data analytic procedures, and it is designed to run on any information system given that it is a Java-based application (Ghassani, Jamaludin & Irawan, 2021).

3.3 Conclusion

This chapter has presented the analytical tool employed in the study, it also highlights the CRISP-DM Methodology and the various steps involved which do not follow any specific order when it is employed. The chapter also mentioned the parameters for measuring or evaluating the quality of the association rules that were mined during the study.

CHAPTER FOUR

STATISTICAL OVERVIEW AND ASSOCIATION RULE MINING

4.0 Introduction

This chapter presents the results of the findings from the data deployed in the study per the research

goal and methodology laid out in the previous chapters. To ensure coherence and speed in the analytical process the data set was prepared accordingly to remove and filter unwanted data sets. Different scenarios were compared by analyzing the quantity and prices of items purchased by customers in a transaction and their lifetime from the retail store. The aim of the comparison for a lifetime customer transaction was to collect information about the basket size and to better understand the transactional data set for proper association rule mining. The lifetime analysis was carried out to understand the overall dataset and the transactional analysis was conducted for the association rule mining. The study utilized the RapidMiner software to perform the data mining process using the FP-Growth algorithm and association rule method on the transactional dataset.

4.1 Statistical Overview

4.1.1 Understanding the Price and Product Distribution

To understand the data, the study has undertaken a series of analyses. Firstly, the study aggregated the customer's lifetime purchases to capture all the transactions a customer made during their lifetime in the retail store in terms of quantity and prices irrespective of the period the transaction was undertaken. This aggregation aims to collect information about the most important customers in the retail stores in terms of basket quantity and basket price.

Table 5: Price and Product Distribution

Row No.	InvoiceNo	CustomerID	sum(Quantity)	sum(UnitPrice)	Row No.	InvoiceNo	CustomerID	sum(Quantity)	sum(UnitPrice)
1	536370	12583	449	55.290	450	C574512	12577	-6	10.400
2	536852	12686	107	22.810	451	C574946	12656	-2	4.250
3	536974	12682	132	67.070	452	C576908	12670	-6	3.300
4	537065	12567	611	287.190	453	C578377	14277	-360	2.160
5	537463	12681	585	109.230	454	C578743	12681	-3	18
6	537468	12567	167	73.210	455	C579127	12674	-1	2.950
7	537693	12441	121	40.710	456	C579192	12657	-390	47.420
8	537897	12683	107	54.390	457	C579532	12494	-5	27.950
9	537967	12494	9	31.900	458	C579562	12553	-2	3.730
10	538008	12683	557	100.590	459	C580161	12700	-2	18
11	538093	12682	344	92.200	460	C580263	12536	-149	23.230
12	538196	12731	418	58.500	461	C581316	12523	-3	19.200

ExampleSet (461 examples,0 special attributes,4 regular attributes)

The first 12 transactional datasets

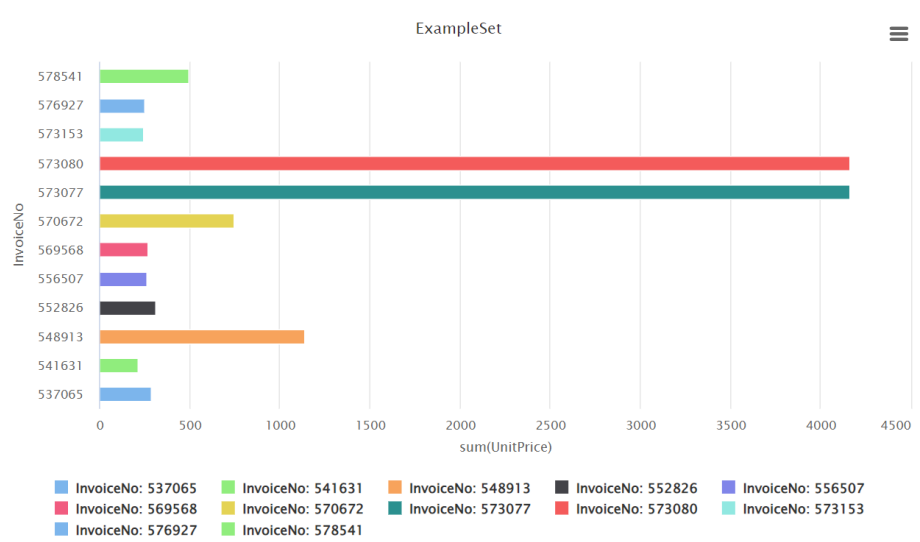
The last 12 transactional datasets

Source: Author (2023)

The above table shows the price and product distribution of customers. It contains the first and last 12 transactional datasets from the comparison made between the customers based on their lifetime transactions in the retail store. Codes starting with the letter “C” indicate a cancellation and were discarded from the analysis as there are negative values from these transactions. From the analysis, we could see a fluctuation in the unit price of items bought by the different customers which indicates that many items in the basket do not necessarily mean a high value for the retail store, for instance, customer ID 12686 with invoice no. 536852 shows a high number of items (107 items) in the customer basket compared to customer ID 12494 with invoice no. 537969 with (9 items) in the customer basket but a comparison of the lifetime unit price of the former customer to the later customer shows that the later offer a higher value to the retail store than the former, by understanding this comparison the retail store may generate information for targeting their customer in terms of their marketing offers.

4.1.2 Price and Product Analysis

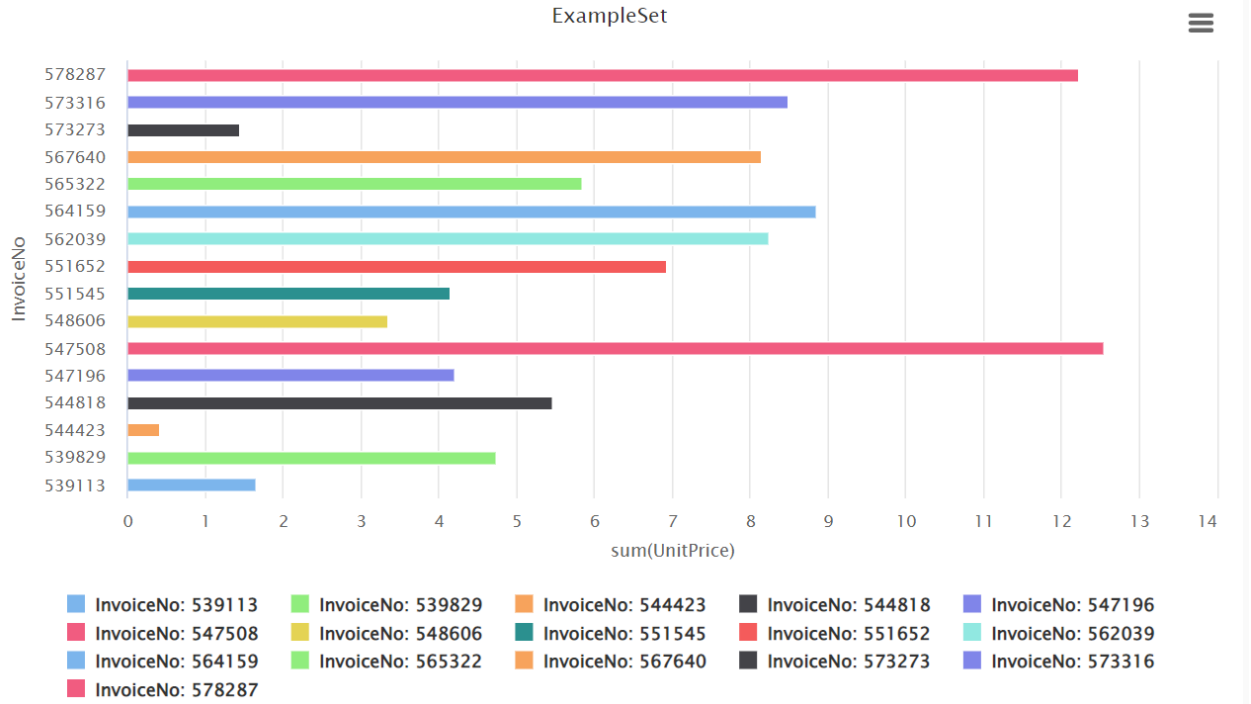
Figure 4: Rank order of customers based on price (1).



Source: Author (2023)

The above figure displays the rank order of the most important customers based on their unit price and irrespective of the quantity of items purchased in their lifetime basket. By setting the minimum unit price high to 200 euros, we could see that customers with Invoices No 573077 and 573080 are ranked top in the list of customers in the retail store with the same unit price value of 4161.060 euros, and then the other customers followed suit. It is additionally noticeable that only a few of the retail store customers are categorized in this segment. This could be because of the shared market focus the retail store has, i.e., a focus on both business-to-business (B2B) operations and business-to-customer (B2C) operations. From this analysis, the retail store can collect information about the customers that generate higher revenue for the retail store and influence decision-making regarding product offerings and marketing campaigns.

Figure 5: Rank order of customers based on price (2).



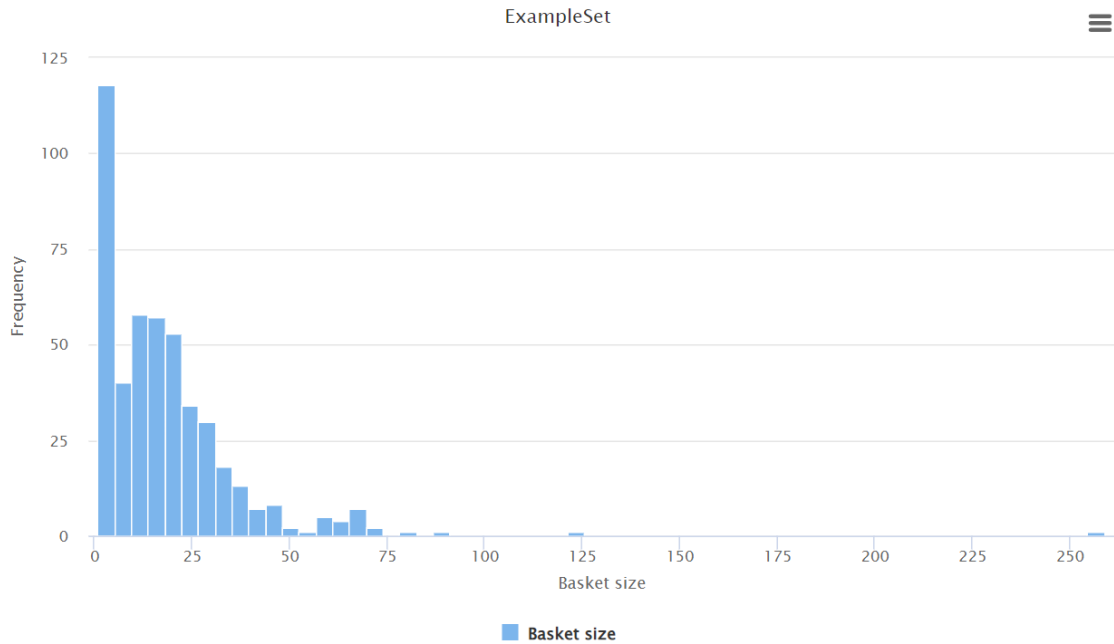
Source: Author (2023)

In a reverse analysis, setting the minimum unit price lower to 15, it could be seen from the figure above that most of the retail store customers fall in this segment. The retail store might want to consider re-strategizing its market focus to wholesale customers only, however, if the store wants to continue to focus on the retail market it could be that they are interested in capturing more market share.

4.1.3 General Product Structure

A separate experiment was carried out to understand the product structure with respect to customer segments and price. The result shows that very few customers are found in the high-unit price segment based on the market basket the rest are clustered in the low-unit price segment, which suggests that these categories of customers should be discarded as they only form a segment of the retail market as shown in the figure below.


Figure 6: Customer Segmentation based on Price.



Source: Author (2023)

Additionally, the generate aggregation operator was employed to make aggregation-by-products to calculate the average unit price of a basket as show below.

Figure 7: Average Unit Price of the Basket

Name	Type	Missing	Statistics		Filter (4 / 4 attributes): <input type="text" value="Search for Attributes"/>
CustomerID	Nominal	0	Least 12736 (4)	Most 12567 (325)	Values 12567 (325), 12681 (278), ...[86 r
UnitPrice	Real	0	Min 0	Max 4161.060	Average 3.850
Description	Nominal	0	<div><p>Least ZINC T-L [...] LARGE (1)</p><p>Most POSTAGE (80)</p><p>Open visualizations</p></div>		
Basket analysis	Numeric	0	Min 0	Max 4161.060	Average 3.850

Source: Author (2023)

It could be deduced that the minimum unit price of a basket was zero and the maximum unit price was placed at 4161.060 euros which is very high and could be associated with wholesale transactions. The average unit price of the basket analysis as shown in the figure was 3.850, which is the average selling price of the items of a transaction in a given period more like a one-time transaction and not a transaction of a lifetime. Hence the retail store may employ the information obtained from these findings to compare the prices across the different product offerings and categories and gain insight about customers for planning marketing, events, and inventory throughout the year.

4.2 Association Rule Mining

The study has prepared the data following three main steps namely, data cleaning, reduction, and transformation. The first, data cleaning, requires that the data set values were processed in advance with missing and incomplete data identified, filtered, and managed. The second, which is the data reduction requires the trimming of the data set to suit the mining process of common materials and requirements, and the last, which is the data transformation requires that the data be transformed into binary numbers of 0 and 1, since the association rule only works with binary codes to simplify the algorithm for technicality purposes and to make the data mining process faster while reducing lagging time.

Table 6: Data Transformation

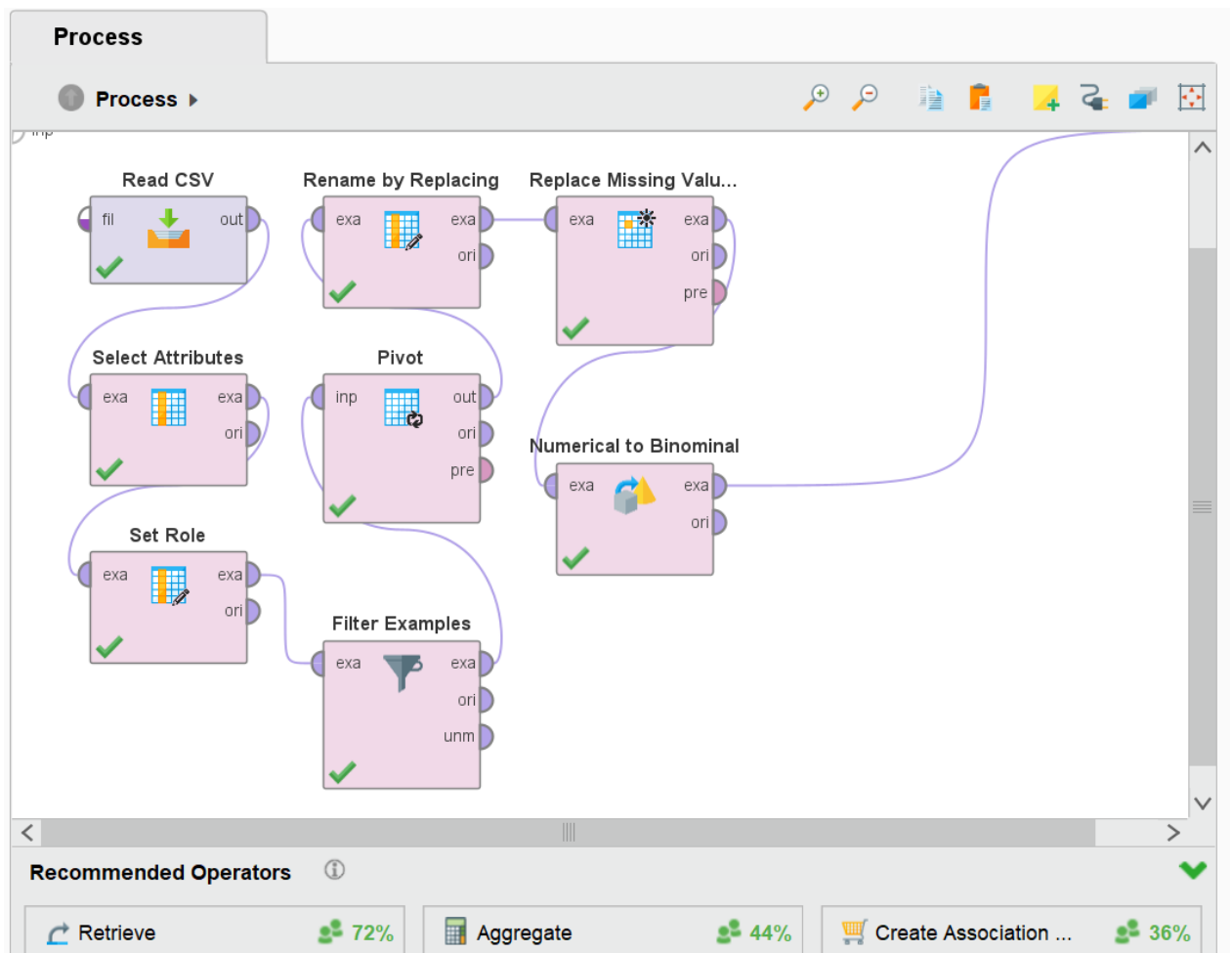
Row No.	InvoiceNo	10 COLOUR...	12 COL... ↓	12 EGG HO...	12 MESSAG...	12 PENCIL ...	12 PENCILS...	12 PENCILS...	12 PENCILS...	12 PENCILS...
59	544115	true	true	false	false	false	true	true	false	false
80	546678	false	true	false	false	false	false	false	false	false
88	547504	false	true	false	false	false	false	false	false	false
132	553411	false	true	false	false	false	false	false	false	false
202	561795	false	true	false	false	false	false	false	false	false
373	579792	false	true	false	false	false	false	false	false	false
1	536370	false	false	false	false	false	false	false	false	false
2	536852	false	false	false	false	false	false	false	false	false
3	536974	false	false	false	false	false	false	false	false	false
4	537065	false	false	false	false	false	false	false	false	false
5	537463	false	false	false	false	false	false	false	false	false
6	537468	true	false	false	false	false	false	false	false	false
7	537693	false	false	false	false	false	false	false	false	false

ExampleSet (392 examples, 1 special attribute, 1,563 regular attributes)

Source: Author (2023)

The table above shows the data set imported into the rapid miner. With some operators used to process the data, we could identify some missing values that were replaced by zero with the replaced missing value operator such that non-missing values were transformed into one and the missing values were transformed into zero. The study also performed the transformation using the “numerical to binominal” operator to modify the data in a way that values that are zero were converted to false and values greater than zero were made true. Since the focus of the rule mining was on the customers' transaction at a period and not the lifetime transaction, InvoiceNo was made an ID as it identifies the behavior of the customers based on their purchasing pattern. The data preparation stage was completed with the following operators as shown in the model below.

Figure 8: Data Preparation

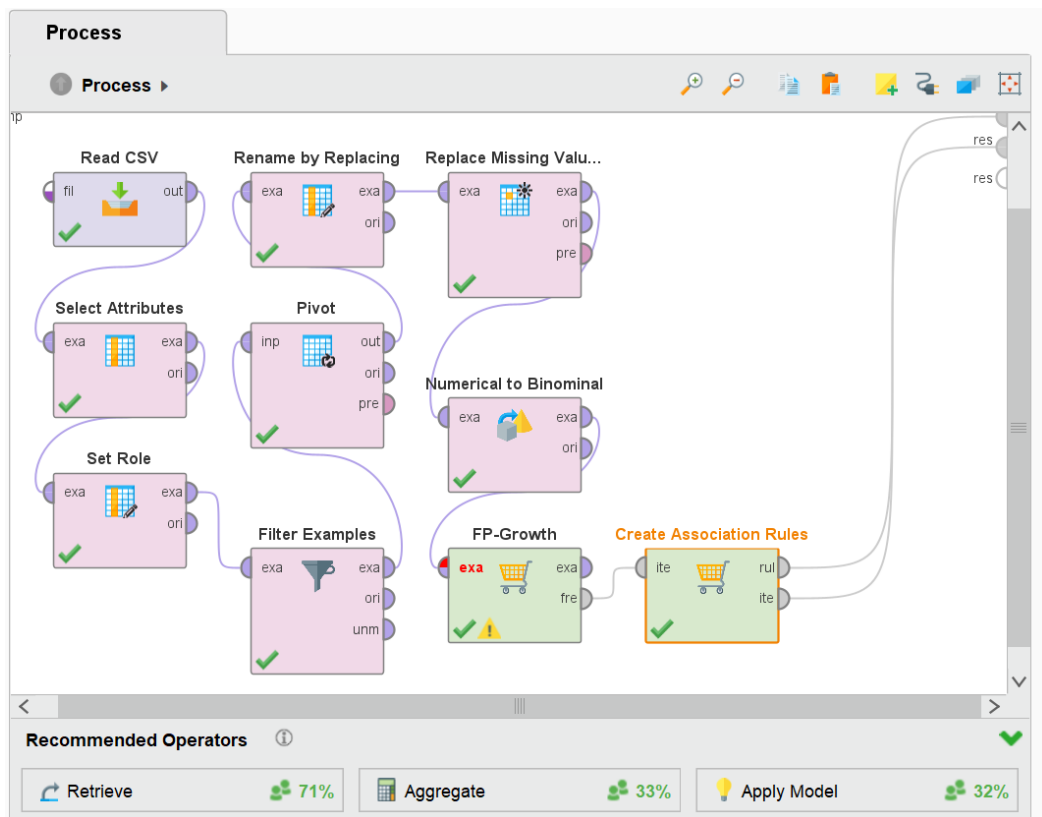


Source: Author (2023)

4.2.1 Data Modeling

After the data preparation and understanding are completed, data modeling is the next step of action which involves the application of the algorithm. The study employed the FP-Growth algorithm which is used to create frequent item sets and to see the associations between the different product categories in the transaction database. The model for the FP-Growth requires a dataset where all the columns are binary, hence, the first column which is the Invoice No was excluded by making it an ID using the set role operation to ensure only the itemset needed to be transformed into binary values were featured. Some parameters for the FP-Growth algorithm were established in this study. These parameters are *min_support*, *min_confidence*, and max number of antecedents and lift score, and the results were analyzed to give room for a min no of antecedent=1. *Min_support* and *min_confidence* were considered in generating the rules and the lift value was the main determinant for evaluating the rule. Below is the rapid miner process design for the data modeling.

Figure 9: Rapid Miner Process Design for the association rule mining



Source: Author (2023)

FP-growth is considered the most effective method for determining frequently co-occurring items in a transaction database. The market basket analysis of this FP-growth algorithm is based on two metrics derived from the transaction record of the retail store namely, support and confidence. The first metric, which is the *support* is described as the probability of two items being purchased together for small values of support, the appearance of an association may be the result of a few coincidences. To decrease the chances of an unreliable or contrived association result, a minimum support value greater than or equal to 10% i.e., $min_support \geq 10\%$ is typically adopted in practice. The second metric, which is the *confidence* describes the conditional probability that an item is purchased if another item is first purchased. The *confidence* value can only be considered valuable or beneficial if the value is high. In this study, the minimum support tested was 15% to 55%, while the minimum confidence tested was 75% to 95%. The minimum support and confidence thresholds usually influence the rule generation. Initially, the study set a high value for the minimum thresholds of confidence and support, but the pattern generated was minimal. In this case, the values were decreased until the study found a value that generated enough patterns.

To answer the first research question that states: “*Which is the most frequent product combination?*”, The product quantity was subjected to ensure reliability when comparing the most frequent product combinations. The parameters set the combination frequency > 10 such that only transactions showing records of items bought more than 10 times together by a customer were considered. This is to ensure the reliability of the result when making comparisons. The support threshold of the FP-Growth algorithm was set at 15% to ensure enough patterns were generated for comparison. The result of the analysis is given as seen below.

Table 7: Most Frequent two-item set combinations.

No. of Sets: 127				
Total Max. Size: 3				
Min. Size: <input type="text" value="1"/>				
Max. Size: <input type="text" value="3"/>				
Contains Item:				
<input type="text"/>				
<input type="button" value="Update View"/>				
Size	Sup... ↓	Item 1	Item 2	Item 3
1	0.122	STRAWBERRY LUNCH BOX WITH CUTLERY		
2	0.122	SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER PLATES	
1	0.120	LUNCH BAG SPACEBOY DESIGN		
1	0.117	LUNCH BAG WOODLAND		
1	0.107	ROUND SNACK BOXES SET OF 4 FRUITS		
1	0.105	MINI PAINT SET VINTAGE		
2	0.105	PLASTERS IN TIN WOODLAND ANIMALS	PLASTERS IN TIN SPACEBOY	
1	0.102	ALARM CLOCK BAKELIKE PINK		
1	0.102	PACK OF 72 RETROSPOT CAKE CASES		
2	0.102	PLASTERS IN TIN WOODLAND ANIMALS	PLASTERS IN TIN CIRCUS PARADE	
2	0.102	SET/6 RED SPOTTY PAPER CUPS	SET/20 RED RETROSPOT PAPER NAPKINS	
2	0.102	SET/20 RED RETROSPOT PAPER NAPKINS	SET/6 RED SPOTTY PAPER PLATES	
1	0.099	DOLLY GIRL LUNCH BOX		

Source: Author (2023)

The highlighted item sets show the most frequent two-item set combinations bought from the retail store. It could be seen that “Set/6 Red Spotty Paper Cups” is bought most frequently with “Set/6 Red Spotty Paper Plates (12.2%) and with “Set/20 Red Retro Spot Paper Napkins (10.2%). Additionally, Plaster in Tin Woodland Animals is frequently bought with Plasters in Tin Space Boy (10.5%) and Plasters in Tin Circus Parade (10.2%) and finally “Set/20 Red Retro Spot Paper Napkins is often bought with “Set/6 Red Spotty Paper Plates (10.2%)

Table 8: Most Frequent three-item set combinations.

No. of Sets: 127				
Total Max. Size: 3				
Min. Size: <input type="text" value="1"/>				
Max. Size: <input type="text" value="3"/>				
Contains Item:				
<input type="text"/>				
<input type="button" value="Update View"/>				
Size	Support	Item 1	Item 2	Item 3
2	0.064	LUNCH BAG APPLE DESIGN	LUNCH BAG SPACEBOY DESIGN	
2	0.054	LUNCH BAG APPLE DESIGN	LUNCH BAG WOODLAND	
2	0.056	SPACEBOY LUNCH BOX	LUNCH BAG SPACEBOY DESIGN	
2	0.071	SPACEBOY LUNCH BOX	DOLLY GIRL LUNCH BOX	
2	0.064	LUNCH BAG SPACEBOY DESIGN	LUNCH BAG WOODLAND	
2	0.051	LUNCH BAG SPACEBOY DESIGN	LUNCH BAG DOLLY GIRL DESIGN	
2	0.074	ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE GREEN	
2	0.074	ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE RED	
2	0.079	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	
2	0.064	CHILDRENS CUTLERY DOLLY GIRL	CHILDRENS CUTLERY SPACEBOY	
2	0.051	PACK OF 6 SKULL PAPER CUPS	PACK OF 6 SKULL PAPER PLATES	
3	0.069	PLASTERS IN TIN WOODLAND ANIMALS	PLASTERS IN TIN CIRCUS PARADE	PLASTERS IN TIN SPACEBOY
3	0.099	SET/6 RED SPOTTY PAPER CUPS	SET/20 RED RETROSPOT PAPER NAPKINS	SET/6 RED SPOTTY PAPER PLATES
3	0.064	ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED

Source: Author (2023)

The Table above shows the most frequent three-item set combinations bought from the retail store. It could be seen that *“Set/6 Red Spotty Paper Cups”* is bought most frequently with *“Set/20 Red Retro Spot Paper Napkins* and *“Set/6 Red Spotty Paper Plates (9.9%)*, *“Plaster in Tin Woodland Animals* is frequently bought with *Plasters in Tin Circus Parade* and *Plasters in Tin Space Boy (6.9%)* and finally, *Alarm Clock Bake Like Pink* is frequently bought with *Alarm Clock Bake Like Green* and *Alarm Clock Bake Like Red (6.4%)*

A quick observation and understanding of these combinations explained above shows that in the most frequent two-item combinations, the items are well complementary in nature. Paper cups and paper plates are commonly used together in social settings and events and when customers purchase *“Set/6 Red Spotty Paper Cups”* they may have a preference for a personalized theme or design leading them to also buy *“Set/6 Red Spotty Paper plates”* to create a cohesive work for their social event. This is the same for the other combinations. Additionally, in the most frequent three-item combinations, the items tend to be more substitute items, especially for the last combination as the only significant difference between the *Alarm Clock Bake Like Pink*, *Alarm Clock Bake Like Green*, and *Alarm Clock Bake Like Red* is the color but the core functionalities remain the same. This becomes interesting as it is expected that substitute products should be competitive, however in this study, although, the percentage value of the support as recorded from this analysis is low, the combination of these products as a frequently occurring product combination could be of interest to the retail store to consider. Several factors could influence frequent combinations, probably the product price, the season, etc. Hence, the retail store should consider looking into this product combination as there is an interesting pattern in the product combinations for decision-making in areas of product shelving, pricing, or other influencing factors in the retail store.

4.2.2 All Confidence Measure

Table 9: All Confidence Evaluations

Open in

Turbo Prep
 Auto Model
 Interactive Analysis

Filter (127 / 127 examples): all

Ro...	Items	Size	Frequency	Support	Score ↓
127	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKELIKE RED	3	25	0.064	68.307
126	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS , SET/6 RED SPOTTY PAPER PLATES	3	39	0.099	42.684
125	PLASTERS IN TIN WOODLAND ANIMALS, PLASTERS IN TIN CIRCUS PARADE , PLASTERS IN TIN SPACEBOY	3	27	0.069	17.375
124	PACK OF 6 SKULL PAPER CUPS, PACK OF 6 SKULL PAPER PLATES	2	20	0.051	14.255
123	CHILDRENS CUTLERY DOLLY GIRL , CHILDRENS CUTLERY SPACEBOY	2	25	0.064	12.963
122	ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKELIKE RED	2	31	0.079	8.643
121	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE RED	2	29	0.074	7.681
120	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE GREEN	2	29	0.074	7.479
112	SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY PAPER PLATES	2	48	0.122	6.969
113	SET/20 RED RETROSPOT PAPER NAPKINS , SET/6 RED SPOTTY PAPER PLATES	2	40	0.102	6.031
117	SPACEBOY LUNCH BOX , DOLLY GIRL LUNCH BOX	2	28	0.071	5.744
111	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS	2	40	0.102	5.584
119	LUNCH BAG SPACEBOY DESIGN , LUNCH BAG DOLLY GIRL DESIGN	2	20	0.051	5.055

ExampleSet (127 examples,0 special attributes,5 regular attributes)

Source: Author (2023)

To support the analysis above, the study performed an *All-confidence* analysis which is a measure of the frequent items set by using the “item set to data” operator. This measure helps to better understand the importance of the product structure and shows the item combinations that are strictly together. The table above shows that the All-confidence of 2 product combinations shows the highest score. “*Alarm Clock Bake Like Pink, Alarm Clock Bake Like Green and Alarm Clock Bake Like Red* (68.307%) and “*Set/6 Red Spotty Paper Cups, Set/20 Red Retro Spot Paper Napkins, Set/6 Red Spotty Paper Plates*” (42.684). Although the All-confidence measure emphasizes the importance of frequent product combinations, knowing the kind of association between these product combinations is greatly important, hence the need to answer the second research question.

To answer the second research question that states “*What kind of associations can be detected between different products and product combinations?*”, We generated different association rules using the create association rule operator. The minimum confidence level was set at a default of 95%. An initial test with 100% confidence showed no rule, probably due to the strength of the confidence. The common interpretation when a confidence score is 100% is the absence of occurrences in the dataset where the antecedent is associated with some other

consequence. In this study, the confidence % was gradually decreased at each test trial until twenty-one (17) association rules were generated at a 75% confidence level. The table below shows the twenty-one (27) association rules generated and ordered by the Lift metric.

Table 90: Result of Association Rule Mining

No.	Premises	Conclusion	Support	Confidence	Lift ↓
10	PACK OF 6 SKULL PAPER CUPS	PACK OF 6 SKULL PAPER PLATES	0.051	0.800	14.255
19	PACK OF 6 SKULL PAPER PLATES	PACK OF 6 SKULL PAPER CUPS	0.051	0.909	14.255
18	CHILDRENS CUTLERY DOLLY GIRL	CHILDRENS CUTLERY SPACEBOY	0.064	0.893	12.963
20	CHILDRENS CUTLERY SPACEBOY	CHILDRENS CUTLERY DOLLY GIRL	0.064	0.926	12.963
15	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	0.064	0.862	9.133
16	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN	0.064	0.862	8.893
13	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	0.079	0.816	8.643
14	ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN	0.079	0.838	8.643
11	ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE PINK	0.064	0.806	7.903
8	ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE PINK	0.074	0.784	7.681
7	SET/6 RED SPOTTY PAPER PLATES	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS	0.099	0.780	7.644
22	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS	SET/6 RED SPOTTY PAPER PLATES	0.099	0.975	7.644
23	SET/20 RED RETROSPOT PAPER NAPKINS , SET/6 RED SPOTTY PAPER PLATES	SET/6 RED SPOTTY PAPER CUPS	0.099	0.975	7.078
17	SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER PLATES	0.122	0.889	6.969
21	SET/6 RED SPOTTY PAPER PLATES	SET/6 RED SPOTTY PAPER CUPS	0.122	0.960	6.969
12	SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY PAPER PLATES	SET/20 RED RETROSPOT PAPER NAPKINS	0.099	0.812	6.125
9	SET/6 RED SPOTTY PAPER PLATES	SET/20 RED RETROSPOT PAPER NAPKINS	0.102	0.800	6.031

Source: Author (2023)

The table above shows the result of the market basket analysis containing the rules generated from the transactional data sets. The study placed the minimum threshold values for support and confidence at 15% and 75% respectively and generated 17 association rules from the market basket analysis. The lift was used as the main metric for evaluation since it specifies the strength of the rule. It is often suggested that a high lift value indicates a stronger association (Suharjo & Wibowo, 2020) that is, when the lift of a rule is greater than 1, it means that the purchase of *X item* increases the probability of purchasing *Y item*.

In the result generated, we could see that there are 17 association rules found with *min_support* = 0.051 and *min_lift* = 6.031. An order by support shows that “*Set/6 Red Spotty Paper Cups → Set/6 Red Spotty Paper Plates*” has the highest support, appearing in 12.2% of the total orders but the confidence is quite low compared to some other product combinations. An illustration of this is given thus: When customers order *Set/6 Red Spotty Paper Cups* (support=12.2%), more than 80% of them will also buy *Set/6 Red Spotty Paper Plates* as It is 6.969 times more

lift values of the 17 generated rules are greater than 1, it shows that the item occurrence in one instance influences the probability of item occurrence in the other instance.

4.2.3 Sensitivity analysis of support threshold

A sensitivity analysis was conducted to evaluate the level of change in the output model influenced by the changes in the *min_support* thresholds. By performing some changes on the *min_support* threshold, the study determined which input parameter is more important or sensible to achieve accurate output values. The first *min_support* threshold applied for the analysis was 45% and it generated only a total of 54 frequent item sets. When the *min_support* threshold was reduced to 35% there was a slight increase in the frequent item sets generated to 83 and finally, the threshold was reduced to 15% to generate more frequent item combinations and was the established *min-support* threshold applied in this study, the frequent item sets generated at the application of this support threshold increased to 127 which is a 44 increase in the frequent item sets generated. These new frequent item sets produced, in turn, delivered a total of 27 new associations compared to the 11 and 15 rules respectively, generated by the *min_support* threshold of 45% and 35%. The top and bottom rules based on *lift* are listed below, excluding redundant rules.

Table 11: Highest Lift Rules

	Rule	Lift
10	Pack of 6 Skull Paper Cups → Pack of 6 Skull Paper Plates	14.255
19	Pack of 6 Skull Paper Plates→ Pack of 6 Skull Paper Cups	14.255
18	Children Cutlery Dolly Girl→ Children Cutlery Space Boy	12.963

Source: Author (2023)

Table 12: Lowest Lift Rules

	Rule	Lift
5	Set/6 Red Spotty Paper Plates→ Set/6 Red Spotty Paper cups	6.969
3	Set/6 Red Spotty Paper cups, Set/6 Red Spotty Paper Plates → Set/20 Red Retro Spot Paper Napkins	6.125
2	Set/6 Red Spotty Paper Plates → Set/20 Red Retro Spot Paper Napkins	6.031

Source: Author (2023)

The above tables show a significant increase in the calculated *lift* scores for this new set of rules. When this kind of analysis is performed, it is often necessary and important to combine the *lift* values with the *confidence* of the consequent(s) occurring, given the antecedent(s), to make justified recommendations. It is notable to state that *lift* scores below one usually detects a negative association between product combinations and indicate a repulsion, however, in this study, the *lift* scores were all above one which indicates a positive association for most of the item sets. However, *high lift scores* indicate a stronger association between item sets and are most likely to be bought together compared to product combinations with *lift* scores that are low.

4.3 Conclusions

This chapter analyzed and evaluated the transactional dataset used in the study from the lens of CRISP-DM methodology. The chapter first presented a statistical overview of the data set to better understand the data for analysis. A market basket analysis was done using the FP-Growth and the create association rule operators to generate 21 association rules from the datasets. The 21 association rules generated were evaluated using the minimum threshold value of support at 35% and confidence at 85%. Due to the inconsistencies in the values in the support and confidence results, the lift metric was used to evaluate the association rules generated. It is worthy of note to state that the thresholds and metrics used in this study were based on the author's discretion as deemed fit for effective analysis, caution should be taken as to the general application of the result from the study. Specifically, the association rules generated could be applicable independently of a specific gift store retailer, for decision-making in areas of cross-selling and upselling, targeted market promotion, store layout design, etc. for-profit maximization.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.0 Conclusion

In today's overly competing global economy, maintaining a business position requires immediate decision capability. This immediate decision capability requires quick analysis of both timely and relevant data that are available to businesses for accurate and useful conclusions. One of the most important goals of businesses, especially retail stores, is to improve their relationship with their customers not just to acquire new customers, but to detect valuable customers from the already existing ones and retain them to increase their long-term profits. Data mining functionalities such as association mining have proven to be a useful tool for detecting these valuable customers.

Generally, current businesses have a deficiency of past usage in developing customer life cycle business strategies. Market basket analysis is one set of data mining techniques that can assist in digging deep into large data sets to bring information to the surface when the data is effectively manipulated. An understanding of the past and present often leads to the decision as to the steps to take in both the present and future and this is what the market basket analysis presents.

In this study, the goal has been to employ market basket analysis to uncover the most frequent item set combinations and to derive association rules from a large hypothetical retail data set from France for strategy settings in areas of product shelving, targeted marketing campaigns, decisions on the market segments to focus operation, and even inventory management in this highly competitive environment, and the FP-Growth algorithm has been employed for mining the association rules in the large database.

The fundamental contribution of this study is the comprehensive demonstration of leveraging the open-source customer transaction data, to reveal interesting information to shopping malls and other retail companies.

The results of the analysis attempt to answer the research question raised in the study. But before answering the questions a proper understanding of the data was performed to gain directions on the approach to adopt for analyzing the data to generate meaningful and valid answers to the research questions. Statistical overview using descriptive analysis was employed to understand the data structure in terms of price and product distribution, customer ranking based on price and quantity of product purchased, and market segmentation. The study then proceeds to answer the research questions of the study. To answer the first research question that aims to identify the most frequent product combinations, out of the total of 8 557 transactional data used for the study, 127 frequent item sets were recorded from the analysis, out of the pairs of item sets showed the most frequent two product combinations, “Set/6 Red Spotty Paper Cups and Set/6 Red Spotty Paper Plates (12.2%), Plaster in Tin Woodland Animals and Plasters in Tin Space Boy (10.5%), Plaster in Tin Woodland Animals and Plaster in Tin Circus Parade (10.2%) “Set/6 Red Spotty Paper Cups and Set/20 Red Retro spot Paper Napkins (10.2%) and Set/20 Red Retro spot Paper Napkins and Set/6 Red Spotty Paper Plates (10.2%). For frequent product combinations of three-item sets, the most frequent combinations were “Set/6 Red Spotty Paper Cups “Set/20 Red Retro Spot Paper Napkins and “Set/6 Red Spotty Paper Plates (9.9%), “Plaster in Tin Woodland Animals, Plasters in Tin Circus Parade and Plasters in Tin Space Boy (6.9%) and Alarm Clock Bake Like Pink, Alarm Clock Bake Like Green and Alarm Clock Bake Like Red (6.4%).

This pattern revealed that “Set/6 Red Spotty Paper Cups “Set/20 Red Retro Spot Paper Napkins and “Set/6 Red Spotty Paper Plates were the most frequent items purchased from the retail store and it is most likely to be bought with other items set if well placed in the store, Hence attention should be given to these products to better understand how the products work together to make appropriate and strategic business decisions.

Additionally, the study performed analysis to understand the kind of associations between these product combinations by setting a certain minimum threshold based on support, confidence, and lift metrics and arrived at a finding that, a frequent combination with a lift score higher than 1 has a strong association rule and would likely be bought together in most of the cases and lift score below 1 has a negative association and indicate repulsion. However, the higher the lift and confidence score, the higher the chances that if one product is bought there is a probability

that the other would be bought. The finding from this study shows that these frequent combinations, Pack of 6 Skull Paper Cups → Pack of 6 Skull Paper Plates (lift score: 14.255), Pack of 6 Skull Paper Plates → Pack of 6 Skull Paper Cup (lift score: 14.255), Children Cutlery Dolly Girl → Children Cutlery Space Boy (lift score: 12.963) and Children Cutlery Dolly Girl → Children Cutlery Dolly Girl (lift score: 12.963) has the highest chances of being purchased together in every case and these combinations Set/6 Red Spotty Paper Plates → Set/6 Red Spotty Paper cups (lift score: 6.969), Set/6 Red Spotty Paper cups, Set/6 Red Spotty Paper Plates → Set/20 Red Retro Spot Paper Napkins (lift score: 6.125) and Set/6 Red Spotty Paper Plates → Set/20 Red Retro Spot Paper Napkins (lift score: 6.031) has the lowest chances of being purchased together even though their lift score is above 1. The study did not dictate any negative associations amongst the variables, but the differences in the strength of association amongst the various combinations based on the lift scores can provide insight to the retail store when making decisions as to which product combinations might produce the highest sales for the store and how to present them to their customers to gain competitive advantage.

5.1 Recommendations

Based on the findings and conclusion drawn from this study, the following recommendations were postulated:

In making decisions about frequent item combinations, the retail store should consider item sets with high support scores. However, the decision depends greatly on the result of the analysis. Results with low support scores for the product combinations do not necessarily mean the products are not a good combination, but high support value can help the retail store understand and identify the rules worth considering for further analysis as low support values often do not provide enough information on the relationships between itemset and hence difficult to make conclusions from the rule.

Additionally, the result of item sets after the market basket analysis sometimes might show a strong association, it is greatly important to conduct a sensitivity analysis of the thresholds to ensure the results generated are valid since thresholds are often manually set by the user.

Furthermore, a thorough analysis of the functional aspects of the core product range might be important to perform. The result from most studies conducted in this research area has proven that often certain products or product combinations that were initially thought of as substitutes to each other may incorporate complementing features. Rules do not extract an individual's preference but rather find relationships between the set of elements of every transaction.

Lastly, the result from this study should be understood as a result that not only applies to the case company but to other retail stores aspiring to understand their customers much better and on a deeper level and aiming to increase their store sales. It is also important to note that the approach used in this study is specifically targeted at Market Basket Data, it may perhaps be extended to other areas.

5.1 Contribution to Knowledge

The study contributes to the body of knowledge in the field of business development and strategic decision-making through business analytics and contributed specifically to:

1. Literature on Market basket analysis of retail stores
2. Literature on sales and marketing strategy for retail stores and companies
3. Literature on consumer buying behavior and understanding

5.3 Suggestions for further research

Due to the limitations and deficiencies summarized in the study and based on interesting directions of research that have emerged during the process of this study, the need to provide suggestions for further research becomes imperative. Additionally, the study has suggested areas for further research to broaden the knowledge of studies related to market basket analysis and customer buying behavior as shown below.

One of the major shortcomings of this work was the access to real-time data probably due to the general data protection regulations (GDPR) in Europe, hence this study was limited to the analysis of a hypothetical data set. The proposed matter for further research would involve the use of a real-time data set that is coherent and has a business-side understanding in other to

provide specific findings channeled to specific organizations as against the general findings provided in this study due to limited access to real-time data.

As most of the circumstances bounding the analytical dexterity of this thesis are related to the source data, the proposed matters for further research would involve acquiring data of better quality. In this research only transaction data of 8557 transactions was used, expanding the range of transactions to cover items dispersed across different retail stores or geographical areas could increase the number of valid data points remaining after the data cleaning. This will not only improve the probability of finding interesting customer behavior patterns but also, reinforce the already found associations by confirming their existence in a larger population. Hence, it is hoped that future research can use more transaction data as this provides more robust results.

This research has been completed by applying the FP-Growth algorithm of the market basket analysis. For further research, it is hoped that other algorithms can be used to enable comparison and perhaps generate new kinds of findings.

Finally, this research has been conducted using the RapidMiner tools as the main instrument for data analysis. For further research, it is hoped that other analytical tools could be used to generate findings that might not have been discovered using the RapidMiner tool.

References

- Aakanksha J, Aditya J & Ramesh D.J (2023) Association Rule Mining in Retail: Exploring Market Basket Analysis with Apriori Algorithm (May 27, 2023). Available at SSRN: <https://ssrn.com/abstract=4461121> or <http://dx.doi.org/10.2139/ssrn.4461121>
- Agrawal R, Imielinski T and Swami A, (1993) "Mining association rules between sets of items in large database", Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, ACM Press, Dec. 1993, pp. 207-216
- Alghanam, O. A., Al-Khatib, S. N. and Hiari, M. O. (2022). Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context. *International Journal of Advanced Computer Science and Applications*. 13(2):421-428
- Alqahtani A.Y (2022) Market Basket Analysis in Polymers Industry: Power BI Case. Proceedings of the 3rd South American *International Industrial Engineering and Operations Management Conference*, Asuncion, Paraguay, July 19-21, 2022
- AlShamsi, A Y, (2022) "Understanding Customer Behaviour in Restaurants based on Data Mining Prediction Technique" (2022). Thesis. Rochester Institute of Technology.
- Alyoubi, K. H. (2020). Association Rule Mining on Customer's Data Using Frequent Pattern Algorithm. *IJCSNS International Journal of Computer Science and Network Security*, 20(5).
- Anas S, Rumui N, Roy A, & Saputro P. H (2022) Comparison of Apriori Algorithm and FP-Growth in Managing Store Transaction Data. *International Journal of Computer and Information System (IJCIS)* Peer Reviewed – International Journal. Vol. 03, Issue 04, October 2022 e-ISSN: 2745-9659
- Anindita A.K (2016) Performing Customer Behavior Analysis using Big Data Analytics. 7th International Conference on Communication, Computing and Virtualization 2016. *Procedia Computer Science* 79 (2016) 986 – 992.
- Ankita T, & Shobha P (2022) Market Basket Analysis of Cosmetic Products using Apriori Algorithm. *Vimarshodgam Journal of Interdisciplinary Studies (VIMJINS) (National, Annual, Bilingual, Interdisciplinary, Peer-Reviewed, Open Access, Online Journal) Volume 2, No. 1, August 2022 ISSN: 2583-228X*
- Annie M.C.L.C., Kumar D.A., (2011) Frequent Item set mining for Market Basket Data using K-Apriori algorithm, in *International Journal of Computational Intelligence and Informatics*, Volume 1, No. 1, pp.14-18
- Anurag S (2021) Implying Association Rule Mining and Market Basket Analysis for Knowing Consumer Behavior and Buying Patterns in Lockdown - A Data Mining Approach. *Preprints* (www.preprints.org) doi:10.20944/preprints202105.0102.v1.

- Arnar B, Dadi K, and Kyrre R., (2021) Habits in Frequency of Purchase Models: The Case of Fish in France Article in Applied Economics · March 2021 DOI: 10.1080/00036846.2021.1883541
- Arora Y, Bhateja N, Goswami V, Kukreja R & Rajput A (2022) Market Basket Analysis using Apriori Algorithm. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)* ISSN: 2347-5552, Volume-10, Issue-3, May 2022 <https://doi.org/10.55524/ijircst.2022.10.3.12> Article ID IRV1036, Pages 62-66 www.ijircst.org
- Arpitha. P (2017) Market Basket Analysis For Data Mining: concepts and techniques *International Journal of Latest Research in Engineering and Technology (IJLRET)* ISSN: 2454-5031 www.ijlret.com || Volume 03 - Issue 01 || January 2017 || 15-20
- Ballester, N., Guthrie, B., Martens, S., Mowrey, C., Parikh, P.J., & Zhang, X., (2014) Effect of retail layout on traffic density and travel distance, in: IIE Annual Conference. Proceedings, Institute of Industrial and Systems Engineers (IISE). p. 798
- Bansude, S., & Vispute, J. (2022). A study on consumer buying pattern on retail stores: A literature review. *International Journal of Health Sciences*, 6(S2), 8872-8882. <https://doi.org/10.53730/ij>
- Barbera F.L, Amato M and Sannino G., (2016), "Understanding consumers' intention and behaviour towards functionalized food", *British Food Journal*, Vol. 118 Iss 4 pp. 885 – 895 Permanent link to this document: <http://dx.doi.org/10.1108/BFJ-10-2015-0354>
- Basant K.S (2021) Changing Dynamics of Consumer Buying Behaviour of Indian Customers: A Review. *Journal of Management Science, Operations & Strategies*, (E- ISSN 2456-9305) Vol. 5, Issue, 01. 75-83p, Nov. 2021
- Brilliant M, DwiHandoko & Sriyanto (2017) Implementation of Data Mining Using Association Rules for Transactional Data Analysis 3rd International Conferences on Information Technology and Business (ICITB), 7th Dec 2017 :177
- Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide.
- Dilrukshi R.S.N (2021) Design product placement layout and personalized discount based on Customer Travel Path. A Thesis Submitted for the Degree of Master of Computer Science. University of Colombo School of Computing.
- Dogan O (2023) Market Basket Analysis with Statistically Improved Association Rules Considering Product Details. licensed under a Creative Commons Attribution 4.0 International License. <https://doi.org/10.21203/rs.3.rs-2581178/v1>.
- Ebrahimi, P., Hamza, K. A., & Zarea, H. (2020). Consumer Knowledge Sharing Behavior and Consumer Purchase Behavior: Evidence from E-Commerce and Online Retail in Hungary. *Sustainability*, 13(18), 10375. <https://doi.org/10.3390/su131810375>

- Elvitaria, L., & Havenda, M. (2017, July). Predicting the Level of Extracurricular Interest in Abdurrah Health Analysis Vocational School Students Using the C4.5 Algorithm (Case Study: Abdurrah Health Analysis Vocational School). *Universal Journal of Information Technology and Systems*, Vol. 2
- Eric E. M and Krishna K. G (2019). Antecedents to consumer buying behavior: the case of consumers in a developing country. *Innovative Marketing*, 15(3), 99-115. doi:10.21511/im.15(3).2019.08
- Evren G, Alvin L, & Jiefeng X (2022) Retail Store Layout Optimization for Maximum Product Visibility. arXiv:2105.09299v1 [math. OC] 19 May 2021.
- Ezuma, M. (2010). *The Role of Non-Governmental and Non-Profit Multinational Organizations in the Alleviation of Rural Poverty: The Nigerian Experience*. University of Nigeria, Enugu Campus, Nigeria.
- Gaikwad, P., Kamble, S., Thakur, N. V. and Patharkar, A. S. (2023) Evaluation of apriori algorithm on retail market transactional database to get frequent itemset," in *RICE*, 2017, p. 187{192, ACSIS, Vol. 10 ISSN 2300-5963.
- Gangurde, R., Kumar, B., & Gore, D. (2017). Building Prediction Model using Market Basket Analysis. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(2) SS
- Garaus, M., Wagner, U., & Kummer, C. (2015). Cognitive fit, retail shopper confusion, and shopping value: Empirical investigation. *Journal of Business Research*, 68(5), 1003-1011
- George A. & Binu D. (2012) An approach to product placement in supermarkets using Prefix Span algorithm. *Journal of King Saud University – Computer and Information Sciences* (2013) 25, 77–87
- Ghassani F Z, Jamaludin A & Irawan A S Y (2021) *Market Basket Analysis Using Algorithm Fp-Growth In Determining Cross-Selling. JIP (Polinema Informatics Journal)*_Volume 7, Issue 4, pg. 49-54 ISSN: 2614-6371 E-ISSN: 2407-070X
- Ghous, H., Malik, M., & Rehman, I (2023). Deep Learning-based Market Basket Analysis using Association Rules. *KIET Journal of Computing and Information Sciences*, 6(2), 14-34. <https://doi.org/10.51153/kjcis.v6i2.166>
- Gurudath S (2020) Market Basket Analysis & Recommendation System Using Association Rules Submitted in partial fulfillment for the degree of Master of Science in Big Data Management and Analytics Griffith College Dublin.
- Hajar N, Elnaz AY, Isaline B, Nhan Q.N, & Mourad T (2023) Decision-making in the context of Industry 4.0: Evidence from the textile and clothing industry. *Journal of Cleaner Production*, 2023, 391, pp.136184. (10.1016/j.jclepro.2023.136184).

- Hamid, Z. and Khafaji H. K. (2021). A General Algorithm of Association Rule-Based Machine Learning Dedicated for Text Classification. *Journal of Physics: Conference Series* 1-11. <https://raw.githubusercontent.com/gitganeshnethi/Datasets/master/storeddata.csv>
- Han, J., Kamber, M., (2001) *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA.
- Han, J., Pei, H., Yin, Y., (2000) Mining Frequent Patterns without Candidate Generation. *Proc. Conf. on the Management of Data SIGMOD'00*, ACM Press, New York, NY, USA.
- <https://github.com/pycaret/pycaret/blob/master/datasets/france.csv?plain=1>
- Hu S, Liang Q, Qian H, Weng J, Zhou W and Lin, P (2021) "Frequent-pattern growth algorithm-based association rule mining method of public transport travel stability," *International Journal of Sustainable Transportation*, vol. 15, no. 11, pp. 879-892, 2021.
- Hyunwoo H, Jonghyuk K, Zoonky L and Soyeon K (2017) Store layout optimization using indoor positioning system. *International Journal of Distributed Sensor Networks* 2017, Vol. 13(2) DOI: 10.1177/1550147717692585.
- Idris A.1, Ardhana V.Y.P, & Manapa E.S (2022), Comparison of Apriori, Apriori-TID and FP-Growth Algorithms in Market Basket Analysis at Grocery Stores. *International Journal of Informatics and Computer Science* Vol 6 No 2, July 2022, Page 107-112 DOI 10.30865/ijics.v6i2.4535
- Istrat V, Lalić N. (2017) Association Rules as a Decision-Making Model in the Textile Industry. *Fibres & Textiles in Eastern Europe* 2017; 25, 4(124): 8-14. DOI: 10.5604/01.3001.0010.2302
- Jalal R. H (2017) An examination of the factors affecting consumer purchase decisions in the Malaysian retail market. Faculty of Business and Management, *PSU Research Review: An International Journal* Vol. 2 No. 1, 2018 pp. 7-23
- Joshi S. M (2018) Market basket analysis using apriori algorithm in data mining. *International Research Journal of Engineering and Technology (IRJET)* e-ISSN: 2395-0056 Volume: 05 Issue: 04 | Apr-2018
- Kamelija T & Janka D (2023) Methods and policies for inventory management. *Journal of Economics (1857-9973)* 2023, Vol. 8 Issue 1, p29-44. 16p.
- Kamruzzaman, S. M., Rahman, C. M., (2010) Text Categorization Using Association Rule And Naïve Bayes Classifier
- Kanimozhi A (2019) Personalized market basket prediction with temporal annotated recurring sequences. *Journal of Emerging Technologies and Innovative Research* June 2019, Volume 6, Issue 6 pg. 120-143. www.jetir.org (ISSN-2349-5162)

- Kaur, M, and Kang S. (2016)"Market Basket Analysis: Identify the changing trends of market data using association rule mining." *Procedia computer science* 85, no. Computational Modeling and Security (2016): 78-85.
- Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B. and Draheim, D. (2021)"A systematic assessment of numerical association rule mining methods." *SN Computer Science* 2.5 (2021): 1-13.
- Khasanah A U A (2020) Implementation of Market Basket Analysis based on Overall Variability of Association Rule (OCVR) on Product Marketing Strategy. IOP Conf. Series: Materials Science and Engineering 722 (2020) 012068 IOP Publishing doi:10.1088/1757-899X/722/1/012068
- Khobragade, P., Selokar, R., Maraskolhe, R. and Talmale, M. (2018) Research paper on inventory management system. *International Research Journal of Engineering and Technology (IRJET)*, 5(4): pp. 252-254. www.irjet.net.
- Kowo, S.A., Vareckova, L. (2023). Correlate of Inventory Management and Organizational Performance, *Ekonomicko-manazerske spektrum*, 17(1), 1-13 [dx.doi.org/10.26552/ems.2023.1.1-13](https://doi.org/10.26552/ems.2023.1.1-13)
- Kurnia, Y, Yohanes I, Yo C G, Aditiya H, & Riki. (2019). Study of the application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm. *Journal of Physics: Conference Series*.
- Larose, D.T., Larose, P.D.T., (2005) Discovering knowledge in data: An introduction to data mining. New York: Wiley-Interscience.
- Laxmi B.D, Kavitha B., Nagarani M. (2017) Descriptive and Predictive Data Mining Techniques to Improve Student Academics and Employability. *International Journal of Innovative Science and Research Technology. Volume 2, Issue 12, December– 2017*
- Liu, G., Huang, S., Lu, C. and Du, Y., (2014) An improved k-means algorithm based on association rules, *International Journal of Computer Theory and Engineering*, 6(2), pp. 146–149. doi: 10.7763/ijcte.2014.v6.853.
- Macharia, S. M., & Mukulu, E. (2016). Role of Just-In-Time in Realization of an Efficient Supply Chain Management: A Case Study of Bidco Oil Refineries Limited, Thika. *The Strategic Journal of Business & Change Management*, 3(6), 123-152
- Maksim P. (2021) Study On Customer Behavior Analysis Using Machine Learning, Turku University Of Applied Sciences. Degree Programme: Information and Communications Technology 2021 | 43
- Martinez M & Escobar B (2021) Market basket analysis with association rules in the retail sector using Orange. Case Study: Appliances Sales Company CLEI *Electronic Journal*, Volume 24, Number 2, Paper 12, July 2021

- Mehmet S.K & Robert R (2022): Predicting customers' cross-buying decisions: a two-stage machine learning approach, *Journal of Business Analytics*, DOI: 10.1080/2573234X.2022.2128447
- Mengying F (2015) Inventory Optimization – Based on Purchasing Activities Analysis. Degree Programme Degree Programme in Logistic Engineering Technology, communication, and transport Jamk University of Applied Sciences.
- Milusheva, P. (2019). Some aspects of the decision to buy, not to produce parts and components. *electronic journal "Economics and computer science"*, issue 2, 2019 2, 64-67.
- Monerah M. A., and Ahmed M. B (2021) A Survey on Methods and Applications of Intelligent Market Basket Analysis Based on Association Rule. *Journal of Big Data* DOI: 10.32604/jbd.2022.021744. pg1-26
- Munyaka J.B. & Yadavalli V.S.S (2022) Inventory management concepts and implementations: a systematic review *South African Journal of Industrial Engineering Jul 2022 Vol 33(2), pp 15-36.*
- Najafabadi, M.K M., Mahrin, M.N.R., Chuprat, S and Sarkan, H. M. (2017) "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," *Computers in Human Behavior*, vol. 67, pp. 113-128, 2017
- Nathan M.F, Yuchi Z, Xueming L, and Xiaoyi W (2016) Targeted promotions and cross-category spillover effects. *Fox School of Business Research Paper No. 16-035* DOI: [10.2139/ssrn.2847635](https://doi.org/10.2139/ssrn.2847635)
- Nasyuha, A.H Jama J, Abdullah R, Syahra Y, Azhar Z, Hutagalung J and Hasugian B. S. (2020), "Frequent pattern growth algorithm for maximizing display items," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 2, pp. 390-396, 2020.
- Nemtajela, N. & Mbohwa, C. (2017) Relationship between inventory management and uncertain demand for fast-moving consumer goods organizations. 14th Global Conference on Sustainable Manufacturing, GCSM, 3-5 October 2016, Stellenbosch, South Africa. *Procedia Manufacturing*, 8(1), pp. 699–706.
- Neysiani B S, Soltani N, Mofidi R, & Nadimi-Shahraki M.H (2019) Improve Performance of Association Rule-Based Collaborative Filtering Recommendation Systems using Genetic Algorithm. *International Journal of Information Technology and Computer Science*, 2019, 2, 48-55 Published Online February 2019 in MECS (<http://www.mecspress.org/>) DOI: 10.5815/ijitcs.2019.02.06
- Nayyar, T., (2019) "Analyzing Customer Buying Behavior" Creative Components Project Report 336. Iowa State University Capstones, Theses and Dissertations <https://lib.dr.iastate.edu/creativecomponents/336>

- Nithya P, Sivapriya K, & Rajshree M (2020) An Overview of Market Basket Analysis Using Apriori Algorithm. *International Journal of Advance Research in Science and Engineering*. Vol No.9 Issue No.03, March 2020.
- Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2 PART 2), 2592–2602. <https://doi.org/10.1016/j.eswa.2008.02.021>
- Online Retail. (2015). UCI Machine Learning Repository. <https://doi.org/10.24432/C5BW33>.
- Oyebode E. O & Agbalaya M. O. (2022) Market-Basket Optimization using Sales Pattern of Supermarket. *London Journal of Research in Computer Science and Technology Volume 22 / Issue 3 / Compilation 1.0 ISSN: 2514-8648*
- Pinakshi K (2019) The contribution of Data Analytics in predicting the future purchase intentions of consumers MSc in Management National College of Ireland
- Pooja (2019) Analyzing the Consumer Purchasing Pattern by Defining the Level of Satisfaction in Rural Consumers with Special Reference to Organized Retail Kirana Store. *Journal of Advances and Scholarly Research in Allied Education Vol. 16, Issue No. 4, March 2019, ISSN 2230-7540*.
- Prajapati D.J., Garg S, Chauhan N.C. (2017) Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment. *Future Computing and Informatics Journal* 2 (2017) 19e30 <http://www.journals.elsevier.com/future-computing-and-informatics-journal/>
- Prasad, R. K., & Jha, M. K. (2014). Consumer buying decisions models: A descriptive study. *International Journal of Innovation and Applied Studies*, 6(3), 335.
- Praveena M, Jahnavi V.S.S, & Sunayana P (2022) Market Basket Analysis using Apriori Algorithm. *Journal of Current Research in Engineering and Sciences Volume 5- Issue 2, Paper 47 August 2022*
- Priyabrata R & Dhananjay D (2022) Theory and Models of Consumer Buying Behaviour: A Descriptive Study Parishodh *Electronic Journal ISSN NO:2347-6648 Volume XI, Issue VIII, August/2022 ISSN NO:2347-664*
- Rana S. and Mondal M. N. I (2021) A Seasonal and Multilevel Association Based Approach for Market Basket Analysis in Retail Supermarket. *European Journal of Information Technologies and Computer Science* DOI: <http://dx.doi.org/10.24018/ejcompute.2021.1.4.31> Vol 1 | Issue 4 | October 2021 pp 9-15
- Rao, AB, Kiran, JS & Poornalatha G, (2021), 'Application of market–basket analysis on healthcare', *International Journal of Systems Assurance Engineering and Management*. <https://doi.org/10.1007/s13198-021-01298-2>

- Rashed I. K (2022) Buyer Prediction Through Machine Learning A Graduate Paper/Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies in Data Analytics Rochester Institute of Technology RIT Dubai
- Raymond S. M (2019) Enhancing the Prediction of Missing Targeted Items from the Transactions of Frequent, Known Users. A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy in the Faculty of Computing, Engineering, and Media, De Montfort University Leicester, United Kingdom.
- Rehman, I & Ghous, H. (2021). Structured Critical Review on Market Basket Analysis using Deep Learning & Association Rules *International Journal of Scientific & Engineering Research*, 12(1), pp:1-24.
- Render, B. Stair, R.M. and Hanna, M.E. (2016) *Quantitative analysis for management*, 9th ed. Pearson Prentice Hall. <https://www.pearson.com/us/higher-education/program/Render-Quantitative-Analysis-for-Management-with-CD-9th-Edition/PGM304065.html> (Accessed in 12 October 2023)
- Sağın A N, & Ayvaz B (2018) Determination of Association Rules with Market Basket Analysis: An Application in the Retail Sector. Southeast *Europe Journal of Soft Computing*. VOL.7 NO1 March 2018 - ISSN 2233 – 1859. Available online: <http://scjournal.ius.edu.ba>
- Salam, A., Zeniarja, J., Wicaksono, W., & Kharisma, L. (2018). Search for Association Patterns for Arranging Goods Using a Comparison of the Apriori and FP-Growth Algorithms (Case Study of the Epo Store Pematang Distro). *Journal of Dynamics*, Vol. 23
- Selva M.G (2018) Lecture note IT1101 - Data warehousing and Data Mining Unit-III LTTC 3003 UB 812 SRM University, Chennai
- Shabtay, L., Fournier-Viger, P., Yaari, R., & Dattner, I. (2021). A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data. *Information Sciences*, 553, 353-375. <https://doi.org/10.1016/j.ins.2020.10.020>
- Shashi P.S, Ajai K, Neetu Y & Rachna A (2018) Data Mining: Consumer Behavior Analysis 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT-2018), MAY 18th & 19th 2018
- Sohaib Z. A (2019) Market Basket Analysis: Trend Analysis Of Association Rules In Different Time Periods. Dissertation presented as a partial requirement for obtaining the master's degree in Statistics and Information Management NOVA Information Management School Instituto Superior de Estatística e Gestão de Informação Universidade Nova de Lisboa.
- Šostar, M. & Ristanović, V. (2023) Assessment of Influencing Factors on Consumer Behavior Using the AHP Model. *Sustainability* 2023, 15, 10341. <https://doi.org/10.3390/su151310341>

- Styvén, M. E., Foster, T., & Wallström, Å. (2017). Impulse buying tendencies among online shoppers in Sweden. *Journal of Research in Interactive Marketing* Vol. 11 No. 4, 2017 pp. 416-431 DOI 10.1108/JRIM-05-2016-0054
- Suharjo R.A & Wibowo A (2020) Customer Relationship Management in Retail Using Double Association Rule. *International Journal of Emerging Trends in Engineering*. Volume 8. No. 5 Available Online at <http://www.warse.org/IJETER/static/pdf/file/ijeter23852020.pdf>
<https://doi.org/10.30534/ijeter/2020/23852020>
- Suntoro, J. (2019). Data Mining Algorithms and Implementation with PHP Programming. <https://doi.org/10.17605/OSF.IO/ZJ5C3>.
- Suregka K. F & Hema M. N, (2022) A Study on Consumer Buying Behavior Towards Organized Retail Stores In Tiruchirappalli District. *Journal of Positive School Psychology* 2022, Vol.6, No.4, 2074-2084 <http://journalppw.com>
- Swati Mahesh Joshi (2018): Market basket analysis using apriori algorithm in data mining. *International Research Journal of Engineering and Technology (IRJET)* e-ISSN: 2395-0056
- Tanja L (2015) Factors affecting consumers' buying decision in the selection of a coffee brand. Bachelor's Thesis 2015 Saimaa University of Applied Sciences Faculty of Business Administration, Lappeenranta Degree Programme in International Business
- Tingting Z, William Y.C.W, Ling C, & Yan W, (2019) "The role of virtual tryon technology in online purchase decision from consumers' aspect", *Internet Research*, <https://doi.org/10.1108/IntR-12-2017-0540>
- Tripathi A & Pandey S (2022) Market Basket Analysis of Cosmetic Products using Apriori Algorithm. *Vimarshodgam Journal of Interdisciplinary Studies (VIMJINS) (National, Annual, Bilingual, Interdisciplinary, Peer-Reviewed, Open Access, Online Journal)* Volume 2, No. 1, August 2022 ISSN: 2583-228X pg 14-28
- Tshepo T (2021) The Impact of Store Layout on Consumer Buying Behaviour: A Case of Convenience Stores from a Selected Township in Kwazulu Natal. *International Review of Management and Marketing*, 2021, 11(5), 1-6. ISSN: 2146-4405 available at <http://www.econjournals.com>
- Tsuji K, Shibata M, Terasawa Y and Umeda S (2021) Products with High Purchase Frequency Require Greater Inhibitory Control: An Event-Related Potential Study. *Front. Psychol.* 12:727040. doi: 10.3389/fpsyg.2021.727040
- Tung, B., & Carlson, J. (2015). Examining determinants of cross-buying behaviour in retail banking. *International Journal of Quality & Reliability Management*, 32(8), 863–880. <https://doi.org/10.1108/IJQRM-11-2012-0148>

- Ulaikere S A O, Asikhia O. U., Adefulu, A D. & Ajike, E. O (2020) Consumer Shopping Behaviour Affectors and Purchase Frequency of Selected Online Students Buyers in Lagos State, Nigeria *International Journal of Advanced Studies in Economics and Public Sector Management* | IJASEPSM p-ISSN: 2354-421X | e-ISSN: 2354-4228 Volume 8, Number 1 February 2020
- Unvan Y. A (2021) Market basket analysis with association rules," *Communications in Statistics – Theory and Methods*, 2021, Vol. 50, No. 7, 1615–1628 <https://doi.org/10.1080/03610926.2020.1716255>.
- Wang, Q., Zeng, C., Zhou, W., Li, T., Iyengar, S. S. & Shwartz, L (2018), "Online interactive collaborative filtering using multi-armed bandit with dependent arms," *IEEE Transactions on Knowledge and Data Engineering*, 2018
- Wagner A. K, Michel W, Fernando de R, Jose A.M (2023) Cross-selling through database marketing: a mixed data factor analyzer for data augmentation and prediction *Intern. J. of Research in Marketing* 20 (2003) 45–65
- Yoon, S., Park, J.E. (2018), Tests of in-store experience and socially embedded measures as predictors of retail store loyalty. *Journal of Retailing and Consumer Services*, 45, 111-119.
- Yuan X (2017) An improved Apriori algorithm for mining association rules. *Advances in Materials, Machinery, Electronics I AIP Conf. Proc.* 1820, 080005-1–080005-6; doi: 10.1063/1.4977361
- Zaki M. J. and Meira, W. J. (2018) Data Mining and Analysis, in *Data Mining and Analysis*, Cambridge University Press, 2018.

APPENDIX

Appendix 1: Product Distribution by quantity

Source: Author's (2023)

Row No.	InvoiceNo	CustomerID	sum(Quantity)	Row No.	InvoiceNo	CustomerID	sum(Quantity)
1	536370	12583	449	19	539688	12678	1
2	536852	12686	107	20	539727	12678	25
3	536974	12682	132	21	539829	12734	88
4	537065	12567	611	22	540178	12681	362
5	537463	12681	585	23	540239	12682	373
6	537468	12567	167	24	540351	12735	263
7	537693	12441	121	25	540365	12413	180
8	537897	12683	107	26	540455	12583	491
9	537967	12494	9	27	540463	12489	105
10	538008	12683	557	28	540521	12651	78
11	538093	12682	344	29	540642	12681	635
12	538196	12731	418	30	540688	12736	173
13	539050	12577	134	31	540789	12643	578
14	539113	12494	3	32	540824	12728	122
15	539407	12726	333	33	540835	12724	115
16	539435	12691	65	34	540851	12523	43
17	539551	12721	129	35	540972	12437	134
18	539607	12681	604	36	540976	12652	505

Appendix 2: Data Transformation

Source: Author's (2023)

Row No.	InvoiceNo ↑	10 COLOUR...	12 COLOUR...	12 EGG HO...	12 MESSAG...	12 PENCIL ...	12 PENCILS...	12 PENCILS...	12 PENCILS...
1	536370	false	false	false	false	false	false	false	false
2	536852	false	false	false	false	false	false	false	false
3	536974	false	false	false	false	false	false	false	false
4	537065	false	false	false	false	false	false	false	false
5	537463	false	false	false	false	false	false	false	false
6	537468	true	false	false	false	false	false	false	false
7	537693	false	false	false	false	false	false	false	false
8	537897	false	false	false	false	false	false	false	false
9	537967	false	false	false	false	false	false	false	false
10	538008	false	false	false	false	false	false	false	false
11	538093	false	false	false	false	false	false	false	false
12	538196	false	false	false	false	false	false	false	false
13	539050	false	false	false	false	false	false	false	false
14	539113	false	false	false	false	false	false	false	false
15	539407	false	false	false	false	false	false	false	false
16	539435	false	false	false	false	false	false	false	false
17	539551	false	false	false	false	false	false	false	false
18	539607	false	false	false	false	false	false	false	false
19	539688	false	false	false	false	false	false	false	false
20	539727	false	false	false	false	false	false	false	false
21	539829	false	false	false	false	false	false	false	false
22	540178	false	false	false	false	true	true	false	false
23	540239	false	false	false	false	false	false	false	false
24	540351	false	false	false	false	false	false	false	false
25	540365	true	false	false	false	false	false	false	false
26	540455	false	false	false	false	false	false	false	false
27	540463	false	false	false	false	false	false	false	false
28	540521	false	false	false	false	false	false	false	false
29	540642	false	false	false	false	false	false	false	false
30	540688	false	false	false	false	false	false	false	false
31	540789	false	false	false	false	false	false	false	false
32	540824	false	false	false	false	false	false	false	false
33	540835	true	false	false	false	false	false	false	false
34	540851	false	false	false	false	false	false	false	false
35	540972	false	false	false	false	false	false	false	false
36	540976	false	false	false	false	false	false	false	false

ExampleSet (392 examples,1 special attribute,1,563 regular attributes)

Appendix 3: Frequent Item Combinations

Source: Author's (2023)

No. of Sets: 127				
Total Max. Size: 3				
Min. Size: <input type="text" value="1"/>				
Max. Size: <input type="text" value="3"/>				
Contains Item: <input type="text"/>				
<input type="button" value="Update View"/>				
Size	Sup... ↓	Item 1	Item 2	Item 3
1	0.122	STRAWBERRY LUNCH BOX WITH CUTLERY		
2	0.122	SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER PLATES	
1	0.120	LUNCH BAG SPACEBOY DESIGN		
1	0.117	LUNCH BAG WOODLAND		
1	0.107	ROUND SNACK BOXES SET OF 4 FRUITS		
1	0.105	MINI PAINT SET VINTAGE		
2	0.105	PLASTERS IN TIN WOODLAND ANIMALS	PLASTERS IN TIN SPACEBOY	
1	0.102	ALARM CLOCK BAKELIKE PINK		
1	0.102	PACK OF 72 RETROSPOT CAKE CASES		
2	0.102	PLASTERS IN TIN WOODLAND ANIMALS	PLASTERS IN TIN CIRCUS PARADE	
2	0.102	SET/6 RED SPOTTY PAPER CUPS	SET/20 RED RETROSPOT PAPER NAPKINS	
2	0.102	SET/20 RED RETROSPOT PAPER NAPKINS	SET/6 RED SPOTTY PAPER PLATES	
1	0.099	DOLLY GIRL LUNCH BOX		

No. of Sets: 127				
Total Max. Size: 3				
Min. Size: <input type="text" value="1"/>				
Max. Size: <input type="text" value="3"/>				
Contains Item: <input type="text"/>				
<input type="button" value="Update View"/>				
Size	Support	Item 1	Item 2	Item 3
2	0.064	LUNCH BAG APPLE DESIGN	LUNCH BAG SPACEBOY DESIGN	
2	0.054	LUNCH BAG APPLE DESIGN	LUNCH BAG WOODLAND	
2	0.056	SPACEBOY LUNCH BOX	LUNCH BAG SPACEBOY DESIGN	
2	0.071	SPACEBOY LUNCH BOX	DOLLY GIRL LUNCH BOX	
2	0.064	LUNCH BAG SPACEBOY DESIGN	LUNCH BAG WOODLAND	
2	0.051	LUNCH BAG SPACEBOY DESIGN	LUNCH BAG DOLLY GIRL DESIGN	
2	0.074	ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE GREEN	
2	0.074	ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE RED	
2	0.079	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	
2	0.064	CHILDRENS CUTLERY DOLLY GIRL	CHILDRENS CUTLERY SPACEBOY	
2	0.051	PACK OF 6 SKULL PAPER CUPS	PACK OF 6 SKULL PAPER PLATES	
3	0.069	PLASTERS IN TIN WOODLAND ANIMALS	PLASTERS IN TIN CIRCUS PARADE	PLASTERS IN TIN SPACEBOY
3	0.099	SET/6 RED SPOTTY PAPER CUPS	SET/20 RED RETROSPOT PAPER NAPKINS	SET/6 RED SPOTTY PAPER PLATES
3	0.064	ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED

Appendix 4: All Confidence Evaluation

Source: Author's (2023)

Ro...	Items	Size	Frequency	Support	Score ↓
127	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKELIKE RED	3	25	0.064	68.307
126	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS , SET/6 RED SPOTTY PAPER PLATES	3	39	0.099	42.684
125	PLASTERS IN TIN WOODLAND ANIMALS, PLASTERS IN TIN CIRCUS PARADE , PLASTERS IN TIN SPACEBOY	3	27	0.069	17.375
124	PACK OF 6 SKULL PAPER CUPS, PACK OF 6 SKULL PAPER PLATES	2	20	0.051	14.255
123	CHILDRENS CUTLERY DOLLY GIRL , CHILDRENS CUTLERY SPACEBOY	2	25	0.064	12.963
122	ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKELIKE RED	2	31	0.079	8.643
121	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE RED	2	29	0.074	7.681
120	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE GREEN	2	29	0.074	7.479
112	SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY PAPER PLATES	2	48	0.122	6.969
113	SET/20 RED RETROSPOT PAPER NAPKINS , SET/6 RED SPOTTY PAPER PLATES	2	40	0.102	6.031
117	SPACEBOY LUNCH BOX , DOLLY GIRL LUNCH BOX	2	28	0.071	5.744
111	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS	2	40	0.102	5.584
119	LUNCH BAG SPACEBOY DESIGN , LUNCH BAG DOLLY GIRL DESIGN	2	20	0.051	5.055
118	LUNCH BAG SPACEBOY DESIGN , LUNCH BAG WOODLAND	2	25	0.064	4.533
89	PLASTERS IN TIN WOODLAND ANIMALS, PLASTERS IN TIN SPACEBOY	2	41	0.105	4.442
97	PLASTERS IN TIN CIRCUS PARADE , PLASTERS IN TIN STRONGMAN	2	23	0.059	4.269
114	LUNCH BAG APPLE DESIGN, LUNCH BAG SPACEBOY DESIGN	2	25	0.064	4.255
93	PLASTERS IN TIN CIRCUS PARADE , PLASTERS IN TIN SPACEBOY	2	35	0.089	3.850
99	ROUND SNACK BOXES SET OF4 WOODLAND , ROUND SNACK BOXES SET OF 4 FRUITS	2	25	0.064	3.763
116	SPACEBOY LUNCH BOX , LUNCH BAG SPACEBOY DESIGN	2	22	0.056	3.745
115	LUNCH BAG APPLE DESIGN, LUNCH BAG WOODLAND	2	21	0.054	3.652
109	LUNCH BOX WITH CUTLERY RETROSPOT , STRAWBERRY LUNCH BOX WITH CUTLERY	2	25	0.064	3.646
106	LUNCH BAG RED RETROSPOT, LUNCH BAG SPACEBOY DESIGN	2	26	0.066	3.614
87	PLASTERS IN TIN WOODLAND ANIMALS, PLASTERS IN TIN CIRCUS PARADE	2	40	0.102	3.546
90	PLASTERS IN TIN WOODLAND ANIMALS, PLASTERS IN TIN STRONGMAN	2	19	0.048	3.474
104	LUNCH BAG RED RETROSPOT, LUNCH BAG APPLE DESIGN	2	26	0.066	3.467
102	ROUND SNACK BOXES SET OF4 WOODLAND , ALARM CLOCK BAKELIKE RED	2	20	0.051	3.418
110	PLASTERS IN TIN SPACEBOY, LUNCH BAG SPACEBOY DESIGN	2	21	0.054	3.243
101	ROUND SNACK BOXES SET OF4 WOODLAND , ALARM CLOCK BAKELIKE GREEN	2	19	0.048	3.161
107	LUNCH BAG RED RETROSPOT, LUNCH BAG WOODLAND	2	22	0.056	3.125
100	ROUND SNACK BOXES SET OF4 WOODLAND , ALARM CLOCK BAKELIKE PINK	2	19	0.048	3.003
96	PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK BAKELIKE PINK	2	19	0.048	2.821
105	LUNCH BAG RED RETROSPOT, SPACEBOY LUNCH BOX	2	21	0.054	2.800
108	LUNCH BOX WITH CUTLERY RETROSPOT , RED RETROSPOT MINI CASES	2	20	0.051	2.593
95	PLASTERS IN TIN CIRCUS PARADE , LUNCH BAG SPACEBOY DESIGN	2	19	0.048	2.401
98	ROUND SNACK BOXES SET OF4 WOODLAND , RED RETROSPOT MINI CASES	2	20	0.051	2.342

Appendix 5: Sensitivity Analysis Table

Source: Author's (2023)

No.	Premises	Conclusion	Support	Confidence	Lift ↓
10	PACK OF 6 SKULL PAPER CUPS	PACK OF 6 SKULL PAPER PLATES	0.051	0.800	14.255
19	PACK OF 6 SKULL PAPER PLATES	PACK OF 6 SKULL PAPER CUPS	0.051	0.909	14.255
18	CHILDRENS CUTLERY DOLLY GIRL	CHILDRENS CUTLERY SPACEBOY	0.064	0.893	12.963
20	CHILDRENS CUTLERY SPACEBOY	CHILDRENS CUTLERY DOLLY GIRL	0.064	0.926	12.963
15	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	0.064	0.862	9.133
16	ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN	0.064	0.862	8.893
13	ALARM CLOCK BAKELIKE GREEN	ALARM CLOCK BAKELIKE RED	0.079	0.816	8.643
14	ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN	0.079	0.838	8.643
11	ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE PINK	0.064	0.806	7.903
8	ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE PINK	0.074	0.784	7.681
7	SET/6 RED SPOTTY PAPER PLATES	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER NAPKINS	0.099	0.780	7.644
22	SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETROSPOT PAPER N...	SET/6 RED SPOTTY PAPER PLATES	0.099	0.975	7.644
23	SET/20 RED RETROSPOT PAPER NAPKINS , SET/6 RED SPOTTY PAPE...	SET/6 RED SPOTTY PAPER CUPS	0.099	0.975	7.078
17	SET/6 RED SPOTTY PAPER CUPS	SET/6 RED SPOTTY PAPER PLATES	0.122	0.889	6.969
21	SET/6 RED SPOTTY PAPER PLATES	SET/6 RED SPOTTY PAPER CUPS	0.122	0.960	6.969
12	SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY PAPER PLATES	SET/20 RED RETROSPOT PAPER NAPKINS	0.099	0.812	6.125
9	SET/6 RED SPOTTY PAPER PLATES	SET/20 RED RETROSPOT PAPER NAPKINS	0.102	0.800	6.031

DECLARATION OF ORIGINALITY

I, OBIEGUO Ifeanyi Kingsley (JNLJTK) declare, that the content of this thesis titled: **“Understanding Customer Buying Patterns through Business Analytics”** is the result of my own work and I only included proper citations in the development of this work following all relevant regulations of the University of Pécs.

I am aware that my work may be checked by the University of Pécs for potential plagiarism and I declare permission to upload my work to the TURNITIN software with all the necessary details (name, title of the work, and its content) and I further accept other academic works to be compared to my work.

Pécs, 21st November 2023

Signed: OBIEGUO Ifeanyi Kingsley