U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

A TEMPORAL ANALYSIS OF FREQUENT PATTERNS – THE IMPACT ON MANAGEMENT

Teresa Raquel do Monte Ramos

Dissertation
Master in Management

Supervised by
Professor Doutor João Manuel Portela da Gama

2021

Acknowledgements

This work would not be possible without the contribution and support of others. As a result, I would like to express my gratitude to everyone who helped along this journey. They all proved to be essential for the accomplishment of this work.

First, I would like to thank my supervisor, Professor Doutor João Manuel Portela da Gama that since the first day supported and guided me. Thank you for always being available, for the motivation, comments and suggestions that made it possible to achieve the goals of this paper. Without his supervision, I could not realize this work with the same rigour and quality presented.

I also want to express my sincerest gratitude to my friends, who were always by my side and always willing to cheering me up and encourage me.

I cannot finish without thanking my family for all the opportunities that proportionated me, making this journey possible, and for always believing in me.

Abstract

The present work aims to research the application of Calendar-Based Temporal Association Rules in a retail business and how to apply the retrieved knowledge in Management.

Nowadays, data is created and stored at an alarming pace, making Knowledge Discovery in Databases of crucial importance, in particular, Data Mining tools. Furthermore, the retail industry is very seasonal, which means that it faces several changes due to modifications in customer behaviour. The aim of this report is then to study association rules using Market Basket Analysis. We explore a vast volume of transactions of a retail business to examine period-characteristic frequent patterns of consumption in a specific time.

The first approach includes the analysis of association rules in the Christmas and Easter seasons. However, this approach has a limitation and profit is not considered. For that reason, we present two complementary analyses. We study high utility itemsets, and we also attempt to discover high utility rare itemsets.

Combining the three approaches' results of the two different seasons makes it possible to understand customers' purchasing trends and improve the company's performance by incorporating the new knowledge in the firm's strategies.

Keywords: Data Mining, Association Rules, Market Basket Analysis, Frequent Itemsets, Apriori Algorithm, Calendar Based Temporal Association Rules, High Utility Itemsets, Rare Itemsets

Resumo

O objetivo do presente trabalho é estudar a aplicação de Regras de Associação temporais baseadas em eventos de calendário numa empresa de retalho e como aplicar o novo conhecimento em estratégias de Gestão.

Hoje em dia, os dados são criados e armazenados a um ritmo alarmante, é, por isso, crucial a descoberta de conhecimento com base numa base de dados, em particular a utilização de ferramentas de Data Mining. O setor do retalho é muito sazonal, o que significa que sofre várias mudanças devido a alterações no comportamento dos clientes. Assim, este relatório tem como propósito estudar as regras de associação usando uma técnica conhecida por "Market Basket Analysis". Um vasto volume de transações de uma empresa de retalho é explorado a fim de estudar os padrões frequentes de consumo que são característicos de um período de tempo específico.

Uma primeira abordagem inclui a análise das regras de associação na época do Natal e da Páscoa. No entanto, essa abordagem tem uma limitação, uma vez que o lucro de cada transação não é considerado. Por esse motivo, também apresentamos duas análises complementares. São estudados itens com alta utilidade e também realizamos uma tentativa de descobrir itens raros de alta utilidade.

Combinando os resultados dos três estudos aplicados a dois períodos diferentes, é então possível entender as propensões de compra dos clientes e melhorar o desempenho da empresa, incorporando o novo conhecimento nas estratégias da organização.

Palavras-chave: Data Mining, Regras de Associação, Market Basket Analysis, Itens Frequentes, Algoritmo Apriori, Regras de Associação temporais baseadas em eventos de calendário, Itens de Elevada Utilitdade, Itens Raros

Table of Contents

List of Figures

List of Tables

List of Equations

Glossary

ARM – Association Rule Miming

CRISP-DM - Cross-Industry Standards Process for Data Mining

DDD - Data-driven decision-making

DM – Data Mining

FIM – Frequent Itemset Mining

HUIM -High Utility Itemset Mining

KDD – Knowledge Discovery in Fatabases

MBA - Market Basket Analysis

RIM – Rare Itemset Mining

1. Introduction

In recent years, with the evolution of technology, enormous amounts of data are available, creating opportunities and challenges for businesses to leverage new knowledge acquired. However, the availability of large databases and their increasing complexity makes it hard to explore and extract useful information (Shaw et al., 2001).

Industry 4.0 has forced firms to rethink the way of performing in the market. Indeed, with the digital transformation, companies need to acknowledge the ever-increasing value and importance of data. However, the information itself does not have any meaning if we do not explore it in a way that allows amplifying its value (McKinsey & Company, 2015). Additionally, consumers are willing to share their information and are also more informed, making customers more demanding and empowered. It is therefore critical for businesses to study customer behaviour and how it changes.

Data Mining techniques allow to access patterns in data and improve the decision-making process. In particular, Market Basket Analysis has proven to be a technique of great relevance in the study of transactions, and its purpose is to find dependency relationships between articles sold. It allows to describe consumption behaviour of customers using the information on customer's purchasing patterns, granting companies the possibility to gain competitive advantage. In fact, a MBA is a crucial tool for retail since it monitors purchasing patterns and enhance customer satisfaction (Chen et al., 2005).

Over the past two decades, Data Mining has received plenty of attention. Indeed, we cannot undermine the applications of Data Mining tools. It is essencial for any business to monitor its activities and rapidly adjust future plans. Besides, retail is a highly seasonal business, making it of even higher importance to understand how customer's preferences change with time.

To what concerns the literature, there is plenty of information referring to association rules and basket analysis, especially algorithms' efficiency and interestingness measures. However, there is little information regarding the study of calendar-based temporal association rules. The goal of this report is then to analyse what products are usually bought together, but also if there are seasonal characteristic frequent patterns for consumption. For example, although products like bread and milk are purchased all year, other products have a higher frequency of consumption during a specific period. To investigate that, we study association rules in two different periods, Christmas and Easter.

Subsequently, we evaluate whether the rules are characteristic of a certain period and whether rules remain frequent in all considered periods. To complement this study, we also perform an analysis of high utility itemsets and rare itemsets, and the focus of these is the revenue for the company.

The technique in analysis in this research allows managers to incorporate the new knowledge in business decisions to serve the customer needs more effectively and efficiently, creating a competitive advantage.

More precisely, this report intends to answer the following three questions: "What type of products are usually acquired together?", "How a season can affect customer purchase behaviour?" and "How can companies exploit the new knowledge about consumer changing preferences and use it in the business?". The first two interrogations relate to customer behaviour, while the last one refers to the company's opportunity to exploit the new knowledge.

1.1 Motivation

Nowadays, we do not witness any more intimate relationships between companies and clients as it happened before when customers were known by their name. The focus on mass production led companies to interact less with their customers and deliver more standardized products. Furthermore, fierce competition among retailers due to the development of technology and the growth of online channels made it mandatory for companies to find out how to guarantee sustainability within the market (Leninkumar, 2017). At the same time, customers are also more informed, meaning that they are more empowered than ever before. It is then mandatory for enterprises to focus on customer engagement more than ever. As a result, companies can provide better customer service by understanding customer needs and adapting these quickly (Leppäniemi et al., 2017).

Customized service in retail markets is of primordial importance (Chen et al., 2005). The authors also reported that the detection and prediction of changes in customers behaviours allow managers to create longer and pleasanter relationships with them.

Relationship marketing theory centres on the fact that customer needs and preferences evolve. A crucial aspect of the concept includes understanding the key drivers that influence firms' outcomes and requires companies to have a customer-centric approach (Grönroos, 2011). Indeed, a better understanding of the process will allow companies to tailor better their services and improve their customer portfolio value. Organizations can achieve this status by providing outstanding value and satisfaction to the point that exceeds the clients' expectations.

Moreover, customer relationship management focuses on customer retention. It is less costly for companies to retain existing customers than to acquire new ones since expenses with customers decrease with time (Lam et al., 2004) is then crucial to deliver a relationship based on quality, and that we cannot find in the competition. According to Shaw et al. (2001), customer relationship management is only possible when new information is incorporated in marketing strategies, allowing to know each customer better, rather than the customers as a group.

Not only anticipating customer needs leads to loyalty, but also, it is an indicator of companies' future profits. A satisfied customer is less price-sensitive, buys more products and spend less time choosing them. At the same time, it is less influenced by competition (Yoo & Park, 2016).

Ultimately, it is possible to achieve relationship quality with DM techniques such as Market Basket Analysis, which we can use to recommend products (Phan & Vogel, 2010). Also, the creation of defensive policies and communication of market offerings allows companies to increase profit by using strategies such as cross-selling (Kassim & Abdullah, 2010).

1.2 Document Structure

The document is divided into four chapters and structured as follows. The first chapter concerns the framework, motivation and structure of the work. Chapter two provides a general overview of the most relevant topics in the literature concerning association rules. In chapter 3, the methodology is presented and outlined, the case study realised, and the results evaluated. This case study is split into three frameworks: frequent itemsets mining, rare itemset mining and high utility itemset mining. Finally, we discuss the results in chapter 4 in the form of practical applications. In the same chapter, we present a summary, the limitations of the study and future work. Finally, this work finishes with the references that supported it.

2. Literature Review

2.1 Data Mining Models

A lot of attention has been devoted to Data Mining (DM) process models since they guide organizations in the implementation of Data Mining projects. However, even though several models are available, they all include sequential steps to support Data Mining tasks (Shafique & Qaiser, 2014).

Brammer (2013) defines Knowledge Discovery in Databases (KDD) as the "non-trivial extraction process of implicit, previously unknown and potentially useful information from data". KDD is a complex method and a multidisciplinary activity to extract high-quality knowledge from low-level data (Fayyad et al., 1996).

Referring to Pitta (1998), the procedure of knowledge discovery may involve the participation of a technical expert in the first phase of the method and then marketing interpretation due to the complexity of the databases nowadays. The importance of the technique for marketing and its implication is discussed further in this work. KDD applications are vital since they allow organizations to reap several benefits, including lower costs, increased profitability, and improved service quality (Hemalatha, 2012). Even though data mining is commonly misinterpreted as KDD, data mining is part of the KDD process and a fundamental step in an iterative sequence of stages (Han et al., 2012). According to Fayyad et al.(1996), KDD is an iterative process that might require significant interaction and involves five different phases. The different steps are illustrated in Appendix 1.

In the initial phase of the process, the data is selected and then cleaned to remove noise and inconsistent data. The next step is the data transformation, in which the data is converted and consolidated in the best format to be processed by Data Mining techniques/algorithms. Following the previously described phases, intelligent methods are applied to obtain frequent patterns and discovered information that ordinarily would not be visible or hardly found, corresponding to the Data Mining process. Subsequently, the new knowledge acquired must be evaluated regarding its interestingness, using adequate measures. In the last phase, the new evidence is presented to the user and stored as new knowledge. However, the authors believe that the KDD procedure may not follow all these steps every time. This means that the process may involve loops between any two steps.

Although the KKD approach has played an important role in the implementation of Data Mining projects, it has performed a subordinated position (Schafer et al., 2018), and the Cross-Industry Standards Process for Data Mining, CRISP-DM, is the most used methodology in DM tasks (Saltz, 2020).

This method translates business problems into data mining tasks independently of the domain field or technologies used (Huber et al., 2019). It includes 6 phases, and these are more comprehensive than in the KDD process. The first one is the business understanding, in which business and data mining goals are defined. The data understanding step includes the collection, exploration, and quality assessment of the database. The data is then selected and processed for the modeling step, which employs various parameters and data mining techniques. Subsequently, it is crucial to compare the outcomes to the data mining objectives, evaluate the results and assess how to use the new knowledge. This fifth step is called evaluation. The last task, deployment, consists of implementing the techniques defined in the previous step and monitoring the results. This process can be visualized in Appendix 2.

Table 1 presents a comparative analysis of the two procedures.

Data Mining Models	**KDD**	**CRISP-DM**
Name of the phases	---	Business Understanding
	Data selection	Data Understanding
	Data Pre-processing	
	Data transformation	Data Preparation
	Data Mining	Modeling
	Interpretation /Evaluation	Evaluation
	Using the Discovered Knowledge	Deployment

Table 1. A comparative study of the KDD and CRISP-DM models
Source: Own creation

Even though Data Mining tools have been available for a long time, the progress made in technology allowed it to become more attractive. Data mining is an essential phase in the process of discovering information in databases, and it is continually evolving as the amount of data created grows and changes. Data mining techniques are rising in popularity and importance since they may be used in practically any research sector, as well as in the modernization of company management (Sarra, 2020).

Fayyad et al. (1996) refer to Data Mining as "the non-trivial process of identifying new, valid, potentially useful and, mainly understandable by observing the data contained in a database of data". Data mining tasks include dependency analysis, class identification, concept description, deviation detection, and data visualization, and are used to extract patterns from big data sets (Shaw et al., 2001). A representation of the data mining tasks can be found in Appendix 3. For the purpose of this research, we will mainly focus on Dependency Analysis, more specifically on the discovery of frequent patterns. Given that Data Mining techniques enable to uncover hidden information, it makes the study of the information extracted particularly useful for marketing purposes.

Data-driven decision-making (DDD) refers to the process of relying on data for decision-making rather than on pure intuition (Provost & Fawcett, 2013). Indeed, it was proven by Brynjolfsson et al. (2011) that DDD is associated with an increase in productivity and market value. Besides, the study conducted by Tambe (2014) demonstrated that using big data leads to improved efficiency and higher returns on performance.

Ultimately, the employment of Data Mining technology in the retail industry can be particularly beneficial if its applications and results are incorporated in the companies' strategies, allowing them to be more successful in a competitive market (Dongre et al., 2014).

2.2 Itemsets Mining

2.2.1 Frequent Itemsets

Frequent itemset mining (FIM) can be defined as the process of identifying groups of products that are often purchased together by clients (Fournier-Viger et al., 2017). Its goal is to discover interesting associations between items, given a transaction database. Different methods can be used to determine the interestingness of a given pattern. In FIM, the support is traditionally the metric used for that purpose.

The support allows one to know how often a rule is encountered in a given dataset (Han et al., 2012). The support of the rule $X \rightarrow Y$ is the percentage of transactions that contain X and Y jointly, over the total number of transactions in a database (D):

$$sup(X \rightarrow Y) = \frac{\text{Number of transactions containing both X and Y}}{\text{Total number of transactions in the D}}$$

Equation 1 - Support of an association rule (X→Y)

An itemset is considered frequent if its support is greater or equal to a user-defined minimum support threshold.

Generally, a data set containing k objects will produce up to $2^k - 1$ frequent itemsets. Therefore, a common problem in FIM is to reduce the number of frequent patterns without sacrificing the loss of information (Rodríguez-González et al., 2018). To mitigate the problem, restrictions in the traditional frequent itemset mining process were suggested (Borgelt, 2012). A closed frequent itemset exists if there is not another itemset with the same frequency. On the other hand, an itemset is called maximal frequent itemset if none of its supersets is frequent (Sutha & Dhanaseelan, 2017). Maximal frequent itemsets are also closed frequent itemsets, as it is represented in Appendix 4.

2.2.2 Rare Itemsets

Association rule mining (ARM) has primarily concentrated on detecting frequent patterns to be used in modelling and prediction, not capturing events that are rare in the dataset. Nevertheless, events that occur rarely may be more interesting and profitable, since they allow to discover lesser-known phenomena (S. Liu & Pan, 2018; Shrivastava & Johari, 2017).

Intuitively, rare itemsets are those occurring together in few transactions or a small percentage of the total transactions. Algorithms designed for regular itemset mining are inefficient for extracting rare association rules, which means that the extraction of this type of itemsets represents a challenge and a more complex process (Szathmary et al., 2007). One of the challenges with rare events is that they can be hard to detect since the rarity of the objects is relative. The occurrences may not be rare in an absolute sense but rare compared to other cases.

Rare itemset mining (RIM) is implemented by comparing the support of an itemset with the defined threshold to mine rare itemsets (S. Liu & Pan, 2018). Here, the threshold has also to be carefully chosen. In frequent pattern mining, a minimum support threshold is defined, and the pattern is interesting if it satisfies the threshold. However, if we select a maximum support, we obtain rare patterns (Adda et al., 2007). On the one hand, if minimum support is too high, it will not be possible to find sporadic events. On the other hand, if the threshold is defined low, as a result, we will have not only rare itemsets but also frequent ones.

Shrivastava and Johari (2017) proved that frequent itemsets do not guarantee to deliver the most profitable solution. Indeed, an infrequent itemset can generate more revenue than a frequent itemset. More recently, the concept of high-utility rare itemsets has emerged in the literature. According to Goyal et al. (2015), this type of itemsets aim to improve decision-making by emphasizing rare itemsets that generate high revenues. The utility of the itemset is considered a function of its profit value and quantity.

2.2.3 High-Utility Itemsets

The concept of high utility itemset mining (HUIM) is an extension of the traditional frequent pattern mining. Though frequent pattern mining may be useful, it is based on the assumption that frequency is sufficient to determine the actual utility of an itemset, even though this is not necessarily the most profitable. High utility itemset presents a new approach in which not only quantities purchased of a product are considered, but also the profit margin plays a role when deciding the most noteworthy itemsets. The concept of utility can refer to profit, sales value, or other user preferences, but in the end, it will always allow quantifying the usefulness of the itemset (Ninoria & Thakur, 2017).

The utility of an itemset X, u(X), is calculated by adding the utilities of the itemset X in all transactions that contain this item. The primary goal of high utility itemset mining is to identify all itemsets that have utility greater than or equal to a minimum utility threshold specified by the user (Pillai & Vyas, 2010). In the case of a market basket analysis, the purpose is to find the itemsets that produce a profit greater or equal to the minimum defined.

Nonetheless, the challenge of HUIM is more complicated than the FIM for two reasons. The first explanation is that the search space can be enormous depending on the number of different products, how similar the transactions are and how large is the range of values for the utility and the minimum value threshold. The number of possible itemsets can then be huge even if the database does not contain many transactions (Fournier-Viger et al., 2019). The second reason is that high utility itemsets are usually dispersed in the database. As a result, many itemsets have to be considered before finding the actual high utility itemsets. Besides, utility measures do not benefit from the monotonicity propriety. This means that effective methods to reduce the search space used in FIM cannot be used to specifically address this issue in high utility mining.

Overcoming the referred problems, the non-frequent itemsets have numerous applications and may contribute to a significant part of the company total earnings. It is, then, in the retailer's best interest to acknowledge customers who contribute a large portion of the company profit. Indeed, mining high utility itemsets can assist in finding consumers who are typically linked with high-profit products but are rarely seen in the bulk of transactions. Companies may profit from this data mining model since it can be easily applied to commoditized resources (Y. Liu et al., 2010).

Undeniably, this model provides the retailer more information about the customer buying behaviour, allowing to implement more targeted campaigns. It is then of the marketing professionals' interest to promote sales of itemsets with high utility. This goal can be achieved by creating reduced price campaigns or discounts targeting the most valuable clients, regardless of how frequent the itemsets are (Y. Liu et al., 2010).

2.3 Association Rules

The concept of association rules was first introduced by Agrawal et al. (1993). This technique allows us to find out if the presence of a set of items in the records of a database implies its appearance in another set of items (Agrawal & Srikan, 1994).

Being D a database consisting of a set of transactions, where a transaction is defined as T = {t1, t2, ..., tm}, each set of items I corresponds to a transaction where I = {i1, i2, ..., im}. An association rule is represented as an implication in the form LHS ⇒ RHS, in which LHS ⊆ I and RHS ⊆ I and LHS ∩ RHS =∅. LHS and RHS are respectively the precedent (left-hand side) and the consequent (right-hand side) of the rule. Both the antecedent and the consequent of a rule of association can be formed by sets containing one or more items (Agrawal & Srikan, 1994).

As the first studies focused on the analysis of data relating to shopping baskets in a supermarket in order to identify products that are usually purchased together, this method became known as Market Basket Analysis. Supermarket basket data consists of transactions where each transaction is a set of items purchased by a customer.

Association rules are a data-mining operation that has aroused great interest both in the academic field and practical applications. These are not restricted to dependency analysis in the context of retail applications. Currently, the exploration of Association Rules has increasingly been studied and applied in several domains. The technique has been used in the manufacturing industry (Lin et al., 2019), where association rules are employed to determine the major causes of irregularities in various production lines. Kamsu-Foguem et al. (2013) showed how the technique can be applied to improve the efficiency of a production process. In the banking industry there also several applications. Aggelis (2004) studied association rules between different types of e-banking payments offered by a bank. On the other hand, Sánchez et al. (2009) explored the technique application in transactional credit card databases to detect and prevent fraud. Moreover, Ordonez et al. (2001) explored the discovering of association rules in medical data, particularly to predict heart diseases.

It is then clear the importance of this data mining technique. Within the scope of data exploration, some research fields have evolved through the use of DM techniques and algorithms that allow the creation of knowledge.

2.3.1 Importance of MBA in Management

As already referred, creating solid relationships with customers is of primordial importance. Cross-selling can be one of the strategies adopted by the company to achieve that goal. As the customer acquires more products from a given vendor, the higher is their connection and, consequently, the switching costs for the customer are higher (Kamakura et al., 2003). It is therefore not only advantageous to the seller, but also to the customer, who buys a wider variety of goods from the same company. Schmitz (2013) defined cross-selling as a customer management process that consists of selling to customers an additional product (different from the first) that may be related to the customer's purchasing behaviour in the past. Another strategy that can be used by companies is up-selling, which consists in increasing the number of units sold of a product or by upgrading to a more expensive version of a product (Kamakura, 2008). A Market Basket Analysis can then support businesses in implementing these types of strategies.

Association rule mining has also been used to analyse the historical and evolutionary connection between itemsets that can then be used in recommendation systems (Moonen et al., 2016). More and more people depend on online sites for their daily purchases, making recommendation systems of increased relevance for the website owners since they can offer tailored services to clients. Recommendation systems attempt to present to the user items in which the customer may be interested. Amazon.com, Netflix and Youtube are three examples of companies, among many others, that use the information on consumers patterns to create product recommendations, increasing this way their sales (Smith & Linden, 2017).

Since the objective of a company is to boost sales, it is important to arrange the products in the store in a way that allows increasing the likelihood of an article being acquired. With this in mind, the company can not only use the conclusions from the MBA to place items with a high level of confidence together or decide to place these products not close to each other and, this way forcing client to go through a long distance and see more products.

Another application of the association rules is related to discounts strategies. When a company decides to apply promotions to products, the final goal is to make customers visit the store and buy more than they usually do. The promoted items usually have a small margin of profit, and, for that reason, companies must make sure that these products are bought with other goods with a higher profit margin. It is then important to strategically decide which products will be placed next to discounted products.

2.3.2 Association Rule Mining

Association rules can be discovered by finding all the rules in a database that the support is greater *minsup* (support threshold) and confidence greater than *minconf* (confidence threshold).

Confidence allows measuring the times a specific item (or itemset) appears together with a specific item (or itemset) out of the total number of transactions that this particular item (or itemset) appears in the entire database. Confidence of the rule $X \rightarrow Y$ gives the proportion of transactions that include X and Y among the number of transactions that contain X, indicating the conditional frequency of Y given X:

$$conf(X \rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)}$$

Equation 2 - Confidence of an association rule (X→ Y)

The higher the confidence, the more likely is one item/itemset to be present in a transaction that contains the other item/itemset.

The process of association rule discovery is usually decomposed into two subtasks. The first one corresponds to the generation of all frequent itemset, that is the discovery of all rules that are no less than a defined support threshold. The second one refers to the extraction of all rules that satisfy the confidence threshold from all the rules discovered in the previous step.

An extension of association rule mining is the multi-level association mining where several hierarchical levels are analysed to extract as much interesting information as possible. An approach to explore multiple-level association rule mining is to directly apply the algorithm in each hierarchical level (Eavis & Zheng, 2009). This type of analysis can be advantageous for organizations since they usually benefit from a product structure that can be exploited to analyse the relation between products at different hierarchical levels.

2.3.3 Evaluation Measures

As the number of potential association rules increases exponentially with the number of objects, it is crucial to choose the "interesting" rules. For example, the strength of an association rule can be measured in terms of its support and confidence, which are the most used measures in the literature (Djenouri et al., 2014; Martínez-Ballesteros et al., 2014). Both measures are used to exclude rules that are low in importance, reducing the total number of associations to consider and, at the same time, the processing power required for the association mining algorithm. This procedure is done by defining minimum desired levels of support and confidence (minimum support and minimum confidence) allowing to drop associations that are less likely to occur in the early stages of the calculation.

These measures are important because they allow accessing the interestingness of a rule. A rule with low support can occur only by change, which means that it may not be profitable for the company to promote the items that the customer bought together.

However, they can also be an object of a random choice or be a false discovery. Rules with high confidence can also occur by chance, implying that confidence and support alone may not be enough to access the interest of a rule. This limitation in support and confidence inspired the creation of many other objective measures to evaluate the quality of association patterns.

One of the main issues that data mining faces is how to evaluate the process of knowledge discovery and how to find interesting rules. The application of interestingness measures is essential for reducing the number of discovered rules and limiting the number of uninteresting patterns. Various methods for identifying significant rules have been presented, and they can be classed based on statistical significance (objective measures) or user subjectivity (subjective measures), which incorporates the user's domain knowledge (Sethi & Shekar, 2019).

Over the past few years, many measures have been proposed to assess the quality and importance of the rules generated by the algorithms, the main difficulty in applying them is to understand and select which is the most appropriate to employ in different situations. In most circumstances, to analyse the interest of the rules generated by the algorithms, a single measure alone, may not be enough.

An objective measure is a data-driven technique to evaluate the quality of a relationship that takes into account statistical characteristics (Yongmei & Fuguang, 2015). It is domain-independent and requires minimal input from the users. They are based on the quality and similarity between rules, rather than taking into account the user believes (Al-hegami, 2004). By contrast, subjective measures, as the name itself suggest, are user-driven and domain-dependent. They depend on the knowledge of the user who examines the rules.

A well-known measure and widely used to filter non-interesting association rules is the lift. It measures how far from independence are X and Y (Azevedo & Jorge, 2007). Given the rule X → Y, the lift will reveal how much more frequent is Y when X happens.

$$lift(X \rightarrow Y) = \frac{\text{conf}(X \ \rightarrow \ Y)}{\text{sup }(Y)}$$

Equation 3 - Lift of an association rule (X→ Y)

The lift can take values between 0 and infinity, and an association rule is interesting if it has a value greater than 1.

If lift (X → Y) = 1, then X and Y are independent, and the rule is not interesting.

If lift (X → Y) > 1, then X and Y are positively dependents.

If lift (X → Y) < 1, then X and Y are negatively dependents.

The lift can be interpreted as the greater the value of the lift, the more interesting is the rule. It measures co-occurrence and it is symmetric concerning antecedent and consequent.

Two main subjective measures are unexpectedness and actionability (Al-hegami, 2004). Considering the concept of unexpectedness, a rule is interesting if it is unknown or if it contradicts the user believes. On the other hand, a rule is actionable if the user can take an action to his/her advantage based on the rule.

Both objective and subjective measures are complementary. Objective measures can be used as a first filter to select interesting rules. Additionally, subjective rules can be used as a second filter to find the desired interesting rules.

2.3.4 Calendar based temporal association rules

According to Verma and Vyas (2005), we may find distinct association rules while considering different periods. Nowadays, the retail business suffers plenty of changes due to the alterations in customer's buying preferences. It is an industry highly affected by seasonal swings, and customer's behaviour is also season oriented (Ayu et al., 2019). Different factors characterise each period, which means that each season the target audience changes and, consequently, association rules of products also change. Hence, the company must analyse the distinctive patterns of each period, allowing the company to incorporate, in its strategy plans, the alterations in customers' behaviour and create better seasonal strategies to drive the market.

Lonlac et al. (2020) argue that periodic trends have recently attracted considerable attention. Indeed, the frequency with which the rules occur, whether sporadically or consistently, may be used to determine how interesting the rules are. Therefore, it is possible to discover new information when association rules are combined with calendar information, also known as calendar-based temporal association rules (Verma & Vyas, 2005).

Traditional frequent pattern mining is only concerned with the frequency of occurrence of the rules, meaning that it neglects the pattern occurrence behaviour (Tanbeer et al., 2009). However, discovering patterns that are exclusive of a given period or that typically occur in a specific period might be important to understand client behaviour. Furthermore, due to the vast number of potential patterns for a particular calendar schema, this sort of association rule can be hard to detect.

A set of calendar-based patterns is defined by a calendar schema (Y. Li et al., 2001) For example, it is possible to have calendar patterns such as every 31 of October and every 24 of December. Let us consider the following example of 5 frequent patterns.

Frequent Itemsets
A
A,B,C
C, E
B, D

Table 2 - Frequent Itemsets example

Even though the above itemsets may be frequent in the database, some of them may not be frequent when considering only one period of the database. For example, {A} and {A, B, C} maybe appear as frequent itemsets regardless of the period considered. In contrast, {C, E} may occur more frequently at the beginning of the year, and {B, D} at the end of the year. Therefore, the latter patterns can be more noteworthy to study. By analysing the discovered association patterns on each period, the company may apply diverse marketing strategies and offer different products in distinct periods.

2.4 Algorithms

2.4.1. Frequent items

Several algorithms have been developed to study ARM. One of the most popular algorithms is the Apriori, which presents a solution to reduce the number of candidates itemsets. It was first defined by Agrawal and Srikan (1994) and, it represents an improvement of the AIS algorithm developed by Agrawal et al. (1993), where rules could only contain one item in the consequent. With the *Apriori,* that characteristic no longer applies, and itemsets are allowed as a consequence of a rule. This algorithm has the property that if one itemset is frequent, then each of its subsets will also be frequent. This implies that if {A,B,C} is a frequent itemset, then {A,C}, {A,B}, {B,C}, {A}, {B}, {C} must also be frequent. As a result, an itemset' support never surpasses the support for its subsets (anti monotone property) (J. Li et al., 2013).

In the first step, the support of individual items (one item) is determined, and the frequent items selected (those that satisfy the minimum support). The non-frequent items are then eliminated (prune step). In the first phase, the algorithm generates potential frequent itemsets, known as candidate itemsets, which refers to the *apriori-gen* function. In the second interaction, the items proposed in the first phase are used to generate new proposed frequent items by joining the sets of individual items suggested in the previous step. Frequent 2-itemsets are then generated and will be used to create frequent 3-itemsets. The algorithm continues with the same process until it cannot find more frequent items.

Other algorithms have also been developed to deal with frequent pattern mining. The FP- Growth allows the discovery of frequent itemsets without the generation of candidates. It builds a compact data structure known as FP- Tree, and extract from this the frequent itemsets. The Eclat uses a vertical database, and for that reason, it only needs to scan the data once. Besides, this algorithm does not compute the confidence, only calculates the support.

2.4.2. Rare itemsets

It is interesting to note that because of the growing importance of rare association mining, several algorithms have already been modified, based on the apriori algorithm, to mine infrequent itemsets.

The MSapriori (Multiple Supports Apriori) is an algorithm that adopts multiple minimum support thresholds, by modulating the support of an itemset taking into account the support of each item, to successfully discover rare itemsets (B. Liu et al., 1999).

Two other well know algorithms are the Apriori-Rare and Apriori-Inverse. The Apriori Rare uses a subroutine called support count to calculate the support of the itemsets, and it generates maximal frequent itemsets and minimal rare itemsets (Sadhasivam & Angamuthu, 2011).

The Apriori-Inverse was proposed by Koh and Rountree (2005). It inverts the Apriori algorithm's downward-closure principle since we look for the subsets that are under maximum support and not only above minimum support, which allows discovering perfectly sporadic rules (Koh & Rountree, 2005). This algorithm aims to find rare rules concerning dispersed itemsets, with low support but high confidence, by defining a minimum and maximum support. For that reason, we can avoid the discovery of a large number of rules, and it is widely faster (Fournier-Viger et al., 2017).

2.4.3 High-Utility Itemsets

Several experiments have been conducted to develop HUIM algorithms. Most of the algorithms adopt a two phases approach. First, it generates the high utility candidates. Secondly, it computes the exact utility to identify the high utility itemsets (M. Liu & Qu, 2012). Algorithms such as IHUP and UPGrowth use this technique. However, one-phase algorithms are faster and generally more memory efficient than two-phase algorithms. Due to the complexity of the problem, they are still computationally expensive and are space and time-consuming.

HUI-miner was one of the earliest attempts to build an efficient one-phase method. It was revolutionary since it allowed users to find high utility itemsets without generating candidates (M. Liu & Qu, 2012).

More recently, in 2015, the EFIM, a single-phase algorithm, was introduced, and it showed significant progress in terms of efficiency (Ninoria & Thakur, 2017). It relies on two new concepts, the sub-tree utility and local utility, to prune the search space. This algorithm has proven to overperform IHUP as well as other recent algorithms, including FHM (2014), HUP-Miner (2014) and HUI-Miner (2012) (Zida et al., 2017).

The EFIM takes as input a transaction database and a minimum utility threshold.

3. Case Study in a Retail Dataset

In this work, through the method case study, we intend to find the groups of more frequent articles and derive association rules for each time considered, according to the values established for the support and reliability parameters. To assess the proposed methods, we utilize “The Complete Journey”, a publicly available database from Dunnhumby, which was used in a Kaggle competition, and that includes real-world transactions of regular clients at a retailer. We intend to understand which frequent patterns under the analysis period are not frequent patterns when considering all the transactions globally and propose marketing strategies to apply the new knowledge. To guide this data mining project, we chose the CRISP-DM model.

This project is split into three studies. Taking advantage of the temporal disposition of the data, a first analysis focus on detecting association rules to find the relation between the articles found. Then, we study, as a complementary analysis, rare itemsets and high-utility items. To conclude, we combine the three analyses to provide a full understanding of the consumers buying preferences, as well as recommendations for making more effective long-term decisions.

R program is used to analyse the data, particularly the *Arules*, an extension package of R. *Arules* provides the structure required to build and modify input data sets for mining algorithms and evaluate the resulting itemsets and rules (Hahsler et al., 2011). For that reason, this package needs to be installed at an early stage. For the software to read the data correctly, it is necessary to guarantee that the information is in the correct format to apply the algorithms. In this case, the data was in the form {T; x1} {T; x2}…{T; xn} where T represents the transaction id and x the name of the hierarchy of the articles sold and where each set is on a line. In addition, we also use the SPMF to support the study of rare and high utility itemsets.

3.1. Business Understanding

In this first phase of the CRISP-DM, it is important to define the data miming techniques that will be used to solve our problem. As already stated in chapter 1, this work focuses on a dependency analysis, particularly on the identification of frequent patterns.

It is also vital to understand the business, which implies identifying the problem under study. As previously indicated, retail is a very seasonal industry, making it crucial to comprehend how the clients' choices vary over time. The study of frequent itemsets and association rules not only can support the company in recognizing how the customer decisions change with time, but also assists in deciding the best products to be target by the marketing campaigns, and which strategies would provide the highest profit. Among the existing methods, it stands out the cross and upper selling, which entails selling an additional product to the clients, the recommendation systems that present to clients potential interesting products, based on past consumer patterns and discount strategies, which make clients visit the store more frequently and spend more money than they normally would.

Lastly, we identified as a goal to increase the revenue of the company, and for that reason, marketing strategies will be suggested with the intention of boosting profits.

3.2 Data Understanding

The Complete Journey database includes data of the purchases of 2500 households who are regular shoppers at a retailer over the course of 102 weeks. This database has a total of 275 889 transactions which comprehends a total of 92 010 different products, within a set of 43 different departments, 307 commodities and 2373 sub-commodities, and an average of 9.3 items purchased per transaction.

From the initial database, two different datasets were used. The product.csv dataset contains information on each product such as type of product, manufacturer, brand, and product size. The transaction_data.csv dataset contains the type of information that would be found on a store receipt. To have more information about the hierarchy of the products bought in each transaction, the two datasets product.csv and transaction_data.csv were joined by the product_id identifier, appendix 5 describes this process.

In the new dataset, each transaction has a unique identifier number, basket_id, and corresponds to the purchase of one or more items made by a customer at a given point in time. Besides, products are organized in a hierarchy, so each product is assigned a department, a commodity, a sub-commodity and a product ID. An example of this hierarchy is presented in figure 1.

Examining the data, one may conclude that the department is not an interesting hierarchy to study, since most of the products in a transaction are under “Grocery”, “Drug Gm” or “Produce”, as can be concluded by looking at appendix 6. Besides, relevant information for commercial purposes would not be found.

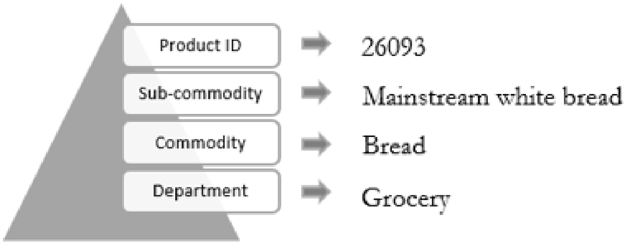


Figure 1 - Example of the product hierarchy

Looking at the product ID level, only a code is provided. Although this information could be complemented with information concerning the sub-commodity and commodity, we established that a study at this level would not create interesting information and, therefore, noteworthy recommendations could not be suggested. Following this review, we decided to study association rules at commodity and sub-commodity levels.

3.3. Frequent Itemset Mining

3.3.1. Data Preparation

In this initial state, data cleaning and data selection are expected since it resembles one of the stages of the CRISP-DM process. This procedure is critical because it is the cleaned and selected data that will be analysed and used to reach new findings.

3.3.1.1 Data Cleaning

This phase has the aim of reducing the dataset and choosing only the variables that will be used in the data mining process. The following changes were made to turn the file into the transactional database that served as the starting point for the other files that enable us to mine association rules for two different hierarchical levels.

- Products with quantity equal to zero were removed from the transaction.

- Non-relevant variables for this study were eliminated: quantity, sales_value, store_id, coupon_match_disc, coupon_disc, retail_disc, trans_time, manufacturer, brand, curr_size_of_product. Although some of these variables will be required at a later stage, only the critical ones were chosen at this time.

- Elimination of lines of text without any useful information for the analysis (for example, blank values, the commodities "unknown" and "no commodity description, "corp use only", "tickets" and "donations" and coupons related categories).

3.3.1.2 Data selection

Since our final goal is to study calendar-based temporal association rules, the previously created database was divided into three other datasets, two corresponding to Christmas, in different years, and the other to Easter. Given that the database does not provide information on the Christmas and Easter periods, a first analysis was realised to understand which weeks were more interesting to study. For this purpose, it was evaluated in which weeks the commodity "Christmas" and "Easter" had more sales.

Looking to the distribution of items bought under the commodity "christmas seasonal" through the 102 weeks, it was concluded that the week 40 and 92 would be the best ones to analyse the Christmas period. The same study was performed for the "easter" commodity. In this case, week 56 stood out the most, and for that reason was the one selected to analyse the Easter period. These results can be consulted in Appendix 7.

At this point, we have four datasets. A dataset with the purchase behaviour of the consumers over the 102 weeks, and three others with information comprising two Christmas and one Easter.

Since it is irrelevant the number of times an article appears in a transaction, the elimination of repeated references, a step from the data cleaning process, was still missing to be performed and, for convenience, it was decided to complete it after having all the necessary files. Thus, all duplication of structure in the same transaction was eliminated by creating two different files for the commodity and sub commodity levels. At the end of this process, we have eight datasets, and for each of them, an analysis of the association rules was performed. In appendix 8, a summary of this process is presented.

3.3.2. Modelling

In this first phase, using the Apriori algorithm, it is intended to find the groups of most frequent articles and derive association rules, according to the values defined for the support and confidence parameters.

A minimum threshold that is too high may exclude interesting items with low support, while a minimum support that is too low may produce too many rules. Therefore, a test of sensitivity to the parameters was carried out since it is important to set values that will allow us to find the most relevant results given the database and the purpose of this study. Table 3 summarizes the number of closed items for each support and hierarchy chosen.

Support	Commodity	Sub-commodity
0,001	1446946	41155
0,005	21549	1523
0,008	6202	626
0,01	3505	415
0,05	72	19
0,06	52	11

Table 3 - Number of closed itemsets for each support and hierarchical level defined.

The support selected for the commodity analysis was 0.008 since it was the one that generated more interesting rules within an acceptable time. The same reasoning was applied to the sub-commodity, hence the support chosen was 0.001, which is the one that generates more itemsets.

Defined the support for each hierarchy, another test was performed to define the confidence, meaning that to different levels of confidence (same support) an analysis of the volume of rules was performed.

Support	Conf = 0.05	Conf = 0.01	Conf = 0.005	Conf = 0.001
Supp=0.008 (commodity)	18076 rules	18178 rules	18178 rules	18178 rules
Supp=0.001 (sub-commodity)	98898 rules	108456 rules	108724 rules	108792 rules

Table 4 - Number of non-redundant rules with lift >1 generated for different values of confidence.

Table 4 shows the number of non-redundant and interesting rules (lift >1) for each support and confidence selected. As we can see, the more detailed the category, the more rules are generated, since there is a greater number of products to be crossed. Considering the results obtained, we chose a confidence of 0.001, for both studies, which, despite being low, is the one that allows us to obtain more association rules. After having selected and cleaned the database and revising all parameters and input constraints, we reach the stage where we evaluate the results.

3.2.3. Evaluation

At a first glance, we chose to analyse the commodity level, before discovering the association rules at the sub-commodity level.

As mentioned in the previous section, the minimum support used for this hierarchical level was 0.008. The top 10 closed frequent itemsets for the commodity level are represented in table 5, for the 102 weeks analysis.

When looking at the frequent items considering all the database, we can confirm something that it was already expected, which is the fact that most of the products bought are under the category Grocery. The commodity "soft drinks" is the one with higher support and it is present in 28% of the transactions realised. The second commodity with higher support is "fluid milk products", and the third is "baked bread/buns/rolls". These results are not surprising since they correspond to basic products that one buys regularly during the year. Indeed, if we look at the top 10, none of the present commodities is a revelation.

itemsets	**support**
soft drinks	0.283
fluid milk products	0.274
baked bread/buns/rolls	0.239
cheese	0.186
bag snacks	0.166
beef	0.145
tropical fruit	0.129
baked bread/buns/rolls, fluid milk products	0.123
eggs	0.110
refrgratd juices/drnks	0.103

Table 5 - 10 itemsets with highest support at commodity level over 102 weeks

This conclusion confirms the importance of removing the items that are frequent during all year for a better assessment of the items that are characteristic of a certain period. Undeniably, it is not pertinent to discover, for example, that people during Christmas and Easter also buy milk and bread, since this is an obvious observation that does not create new knowledge. Therefore, to study the seasonal frequent items at the commodity level, all the commodities present in the results of frequent items performed over the entire database for the support of 0.008 were removed from the Christmas and Easter datasets. The results can be found in table 6 and 7.

itemsets	support
christmas seasonal	0.166
canned milk	0.066
candles/accessories	0.059
nuts	0.052
disposible foilware	0.042
toys and games	0.040
film and camera products	0.037
misc wine	0.033
bookstore	0.028
prepaid wireless&accessories	0.028

Table 6 - 10 itemsets with highest support at commodity level in the Christmas (week=40) dataset

Observing the results for week 40, a more diversified type of products is found. “Grocery” is not the most common department, but instead “drug gm”. The commodity “christmas seasonal” leads with a support of 0.166, which means that it is present in approximately 16% of the transactions under analysis realised during the Christmas week. This commodity, together with “toys and games”, was already predicted to be among the most frequent. We can also find in this top the commodity “disposable foilware”. This finding was not predictable. However, it can be justified for the fact that people usually cook and bake more at Christmas. Commodities related to technology are also represented in this top 10, namely “film and camera products” and “prepaid wireless&accessories”, which seems to be related to the fact that people usually wait for the Christmas season to buy and offer more expensive products. It is also interesting to notice that wine is present in this ranking and may suggest that people usually drink more in this season. Besides, the commodity “bookstore” indicates that people usually read more books in this season or buy to offer as a gift.

A parallel analysis was performed for the second Christmas, but no significant differences were found. For that reason, it was decided that a detailed analysis would not be necessary.

Analysing the information obtained from the Easter period, it is not surprising that “easter” is the commodity present in more transactions (150 transactions) and has a support of 0.188. In the second place appears “toys and games” which, together with “nuts”, were already commodities present in the Christmas analysis. Surprisingly, the fourth most frequent commodity is “hair care accessories”, which may mean that people tend to take more care of their hair in this season. "Floral-flowering plants" is the next in this ranking with a support of 0.038. Indeed, when looking for the most frequent itemsets in the easter dataset there seems to be a relation with good weather since commodities such as "floral-flowering plants" and "spring/summer seasonal" are also present.

itemsets	**support**
easter	0.188
toys and games	0.071
charcoal and lighter fluid	0.050
hair care accessories	0.045
floral-flowering plants	0.038
kitchen gadgets	0.030
frzn jce conc/drnks	0.029
nuts	0.029
frzn fruits	0.028
spring/summer seasonal	0.026

Table 7 - 10 itemsets with highest support at commodity level in the Easter dataset

It will be interesting to see, if, in fact, there are association rules that relate to these categories. The presence of the commodities "frzn jce conc/drnks" and "frzn fruits" may also be related to the good weather and the preparation of cold drinks.

The same analysis was realised at the sub-commodity level and the results over the 102 weeks are presented in table 8. These results were complemented with the respective commodity.

itemsets	**support**
fluid milk white only - fluid milk products	0.244
bananas - tropical fruit	0.121
mainstream white bread - baked bread/buns/rolls	0.107
soft drinks 12/18&15pk can car - soft drinks	0.102
sft drnk 2 liter btl carb incl - soft drinks	0.094
shredded cheese - cheese	0.084
potato chips - bag snacks	0.072
candy bars (singles)(including - candy - checklane	0.065
dairy case 100% pure juice – o - refrgratd juices/drnks	0.065
bananas - tropical fruit, **fluid milk white only** - fluid milk products	0.062

Table 8- 10 itemsets with the highest support at the sub-commodity level

Once again, when looking for the frequent items considering the 102 weeks, it was expected to find general products that are part of the daily life of the consumers, and are not characteristic of a particular part of the year. Actually, even when analysing at the sub-commodity level, there are sub-commodities with a relatively high support. This conclusion confirms the fact that there are specific products that are frequently bought during the year.

The results are also not surprising for the fact that the sub-commodities with the greatest support are all under the most frequent commodities. “Fluid milk white only” is the most frequent sub-commodity by being present in nearly 24% of the transactions. “Bananas”, a fruit usually available all year, is the second on the list, with a support of 0.121. Not surprisingly mainstream white bread closes the top 3, presenting a support of 0.107. As already concluded in the commodity analysis, several soft drinks are also within the most frequent sub-commodity. As previously stated, cheese is also one of the most frequent items. “Potato chips” and “candy bars” are unexpected commodities to be within the most frequent, but it will be interesting to see if there are other candy-related sub-commodities that will emerge in the Christmas and Easter analysis.

itemsets	**support**
misc. seasonal items - christmas seasonal	0.032
hams-half/port bone-in - smoked meats	0.027
frzn bread dough - frozen bread/dough	0.024
ribbons and bows - christmas seasonal	0.022
baking cocoa - baking needs	0.021
wrap - christmas seasonal	0.018
nuts inshell - nuts	0.017
bandana/scarves - fall and winter seasonal	0.017
hams-spiral - smoked meats	0.017
bread:party breads - bread	0.017

Table 9 - 10 itemsets with highest support at the sub-commodity level in the Christmas 1 dataset (week=40)

In table 9, we can see the itemsets with the highest support for the Christmas 1 dataset. Among the most frequent itemsets in the Christmas period, we can see a dominant presence of the commodity “christmas seasonal”. The most common sub-commodity in the Christmas dataset, after removing the most frequent sub-commodities, is “misc. seasonal items” with a support of 0.032, which means thar this commodity is present in roughly 32 % of the transactions considered in this period. “Ribbons and bows” and “wrap” are also among the most frequent, which is not a surprise, since these items are used in Christmas’ presents. The “smoked meats” commodity also has a strong presence, being represented by “hams-half/port bone-in” and “hams-spiral”. “Baking cocoa” reinforces the predisposition to bake more in this season. “Bandana/scarves” is associated with the characteristic weather of this season. Besides, not only “nuts inshell” seems to be a traditional product in this season, but we can also see that there are two types of bread that people buy more frequently in this season “frzn bread dough” and “bread:party breads”.

Table 10 presents the 10 itemsets with the highest support at the sub-commodity level for the second Christmas. Except for the sub-commodity "dogs and cat accessories" no big differences were found. Turkey, which is present in this top 10, it is also present in the top 12 of the first Christmas under analysis.

itemsets	**support**
wrap - christmas seasonal	0.049
ribbons and bows - christmas seasonal	0.034
outside vendors gift cards - misc sales tran	0.027
hams-spiral - smoked meats	0.023
turkey breast bone in - meat	0.023
xmas plush - christmas seasonal	0.018
hams-spiral - smoked meats	0.018
gift-wrap seasonal - greeting cards/wrap/party sply	0.015
dog & cat accessories - pet care supplies	0.014
hams-half/port bone-in - smoked meats	0.014

Table 10 - 10 itemsets with highest support at the sub-commodity level in the Christmas 2 dataset (week=92)

The results of the Easter analysis are presented in table 10. The frequent itemsets found are not that interesting since most of the itemsets include only Easter products. The commodity "easter" had a strong position as we saw in the commodity analysis, it is therefore not surprising that most of the sub-commodities belong to this commodity. Hams-half/port bone-in is not only present in Christmas, but also Easter, and is the only sub-commodity that is frequent in two considered periods. Also here, we see a category "pansies" associated with the typical weather of this season. It will be interesting to see if the association rules discovered include all the frequent items.

itemsets	**support**
easter egg coloring - easter	0.087
grass/shred - easter	0.057
easter fill eggs - easter	0.049
easter baskets - easter	0.046
hams-half/port bone-in - smoked meats	0.036
grass/shred - easter **, easter baskets** - easter	0.027
hams-spiral - smoked meats	0.027
easter basket stuffers- easter	0.024
easter plush - easter	0.021
pansies – garden center	0.018

Table 11 - itemsets with highest support at the sub-commodity level in the Easter dataset

After having found the most frequent itemsets at commodity and sub-commodity levels, which allowed us to have an idea of the most commercialized typologies of products, we will now derive the association rules to find out if we obtain strong associations between products. As already stated, we used the Apriori algorithm, and it was applied a minimum support of 0.008 and minimum confidence of 0.001 for the commodity analysis. To have only the more relevant and strong association rules, the lift was the criteria used and for that reason, only association rules with lift higher than 1 were studied. In the case there is a rule {X → Y} with the same lift as {Y→X}, the first rule was the only one considered, even if with different confidence, since the second rule does not provide extra information.

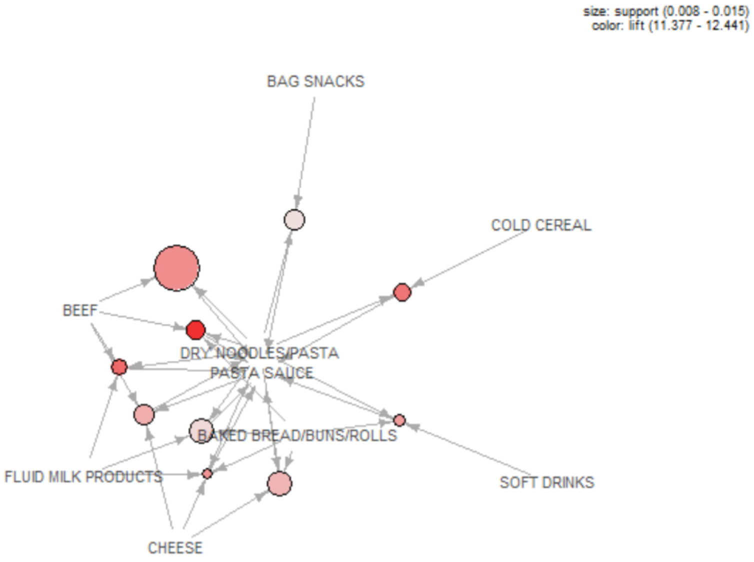


Figure 2 – Graphical representation of the 10 association rules with the highest lift at the commodity level

In figure 3, we can see the top 10 generated association rules by lift, when considering all the database. They all have a lift higher than 1, meaning that the items are positively dependent. The association rule with the highest lift, represented by a stronger colour, is the rule {beef,cheese,dry noodles/pasta} → {pasta sauce}. It has a lift of roughly 12 and a support of 0.010, which is symbolized by the size of the circles corresponding to the rules. This rule is a surprise for the fact that the two categories were not present among the most frequent ones, however, the fact that the products are usually bough together is not a revelation. It is also not a surprise that items like "fluid milk products","soft drinks", "bag snacks", "cheese" and "baked bread/buns/rolls" appear in the association rules with the greatest lift since they were already included in the top 10 of the frequent items. Besides those items, the association rules focus on the categories "pasta sauce", "beef" and "dry noodles/pasta".

It can be observed that the association rules with the highest lift are not diversified, since most of the rules have 2 or 3 commodities in common. However, in this phase, our goal was not to find rules that could represent knowledge for the retailer, since most of the items are part of the daily groceries of all clients, and therefore, not surprising. The aim was to give a better overview of the type of association rules that can be found when considering all the database.

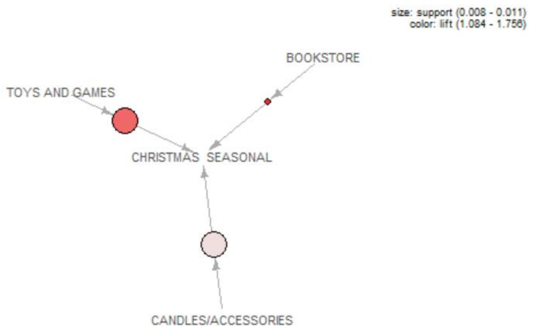


Figure 3 - Association rules with lift greater than 1 for the Christmas dataset at the commodity level

The association rules with lift greater than 1 for the Christmas dataset at commodity level are presented in figure 4, in the form of a network of products. As already stated, a stronger rule is represented by a stronger colour and the biggest the support, by the biggest the circle connecting the products.

There are 3 rules with a lift higher than 1, which means that only 3 rules represent commodities that are positively correlated. Whenever someone does buy "chrismas seasonal", he is very likely to buy "bookstore" as well, as inferred from the lift value of 1.756. These commodities appeared already in the top 10 of most frequent commodities and it is not shocking that these products are bought together since, as already stated, people can buy books to offer in this season. The second rule associates "christmas seasonal" and "toys and games" with a lift of 1.594. This rule also has a confidence of 0.265 that means that it is expected that 26.5% of the transactions with "toys and games" implies the purchase of an item under the commodity "christmas seasonal". This is the most obvious rule and, for that reason, not so interesting since it is already of the retailer knowledge. A weaker association exists between "christmas seasonal" and "candles/accessories", where the lift is 1.084. The support of this rule is 0.01, which means that both items are present in 1% of the considered transactions. However, it represents an interesting rule, not only for the fact that they were both among the most frequent commodities but also because "candles/accessories" is a commodity that is not necessarily associated with the Christmas season.

In figure 6, we can see the association rules of commodities, with a lift greater than 1, in the easter dataset. Additional information is also provided in appendix 9. There are also 3 rules with a lift higher than 1 and we can already conclude that the commodity "easter" is the consequent in the 3 rules. This was already expected considering its strong presence in the Easter dataset. The strongest rule contains the commodity "hair care accessories" and has a lift of 2.667, which means that there is more chance of occurrence of the commodity "easter" given that "hair care accessories" is also present in the transaction. This rule can be explained by a bigger concern with the hair and that people usually take more care of the hair in this season. A good strategy would, therefore, be to position both commodities close to each other in a store.

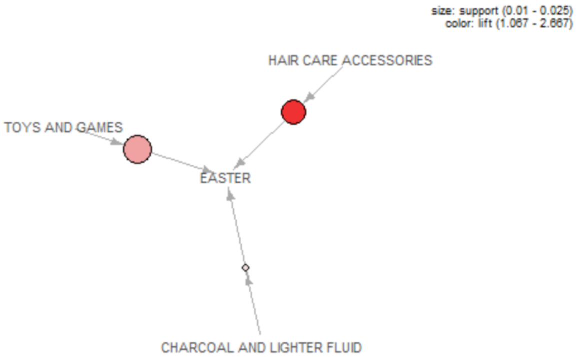

size: support (0.01 - 0.025)
color: lift (1.067 - 2.667)
HAIR CARE ACCESSORIES
TOYS AND GAMES
EASTER
CHARCOAL AND LIGHTER FLUID

Figure 4 – Association rules with lift greater than 1 for the easter dataset at the commodity level

"Toys and games" commodity also appear related to Easter. This rule has the biggest support, as we can see by the size of the circle that relates both commodities, which is 0.025, and means that both commodities are present in 2,5% of the considered transactions in Easter. It can be related to the fact that there is the tradition to offer presents in the Easter period, particularly to kids. The last rule {charcoal and lighter fluid} → {easter} has the lowest lift and support, and a confidence of 0.2, which means that there is 20% more probability of a product under "easter" to be bought when an item of the commodity "charcoal and lighter fluid" is also bought. It can be explained by the fact that usually in easter there is good weather and usually people start doing barbecues. The placement of this commodity close to the easter section could also be a strategy adopted by the retailer to increase the sales of the commodity.

Although at this point of our study we can already conclude that there are different patterns of purchase when considering different seasons in the year, a similar analysis was performed for the sub-commodity level, to look for associations with higher lifts and more interesting rules.

To discover association rules at the sub-commodity level we used a support of 0.001 and confidence of 0.001. The 10 association rules with the highest lift according to each period in the analysis are presented in figures 6,7,8 and 9, considering all dataset, the two Christmas and Easter, respectively. More information regarding these rules can also be found in Appendix 10.

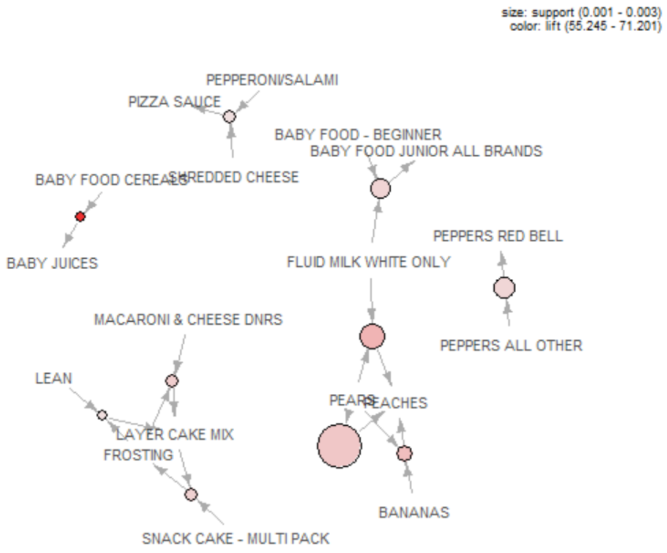


Figure 5 – 10 association rules with lift greater than 1 at sub-commodity level

Looking at the top 10 association rules, we can see that with exception of "fluid milk white only", "bananas" and "shredded cheese", all the other itemsets were not among the most frequent, but all the values for the lift are higher than 55, we can then conclude that all the association rules, for all levels of confidence, represent a strong relationship of dependence between the subcategories presented.

The rule {baby food cereals} → {baby juices} stands out as the strongest rule and with a lift of 71, which means that the 2 commodities are strongly positively correlated. In addition, the rule that relates milk and the two sub-commodities of baby food is also a rule easy to anticipate rule since these products are usually used together, and for that reason also bought together. The rules that connect the fruits bananas, pears and peaches, and milk, are among the ones with the highest lift. Although these items are usually not consumed together, they are usually present at the consumer's home, it is then also not a surprise. However, it could be an opportunity for the retailer to create packs for smoothies or shakes, for example.

All the other rules presented in figure 6, are not a surprise, since they relate products under the same category, especially the one that connects "layer cake mix", "snack cake-multi pack" and "frosting". In this case, if this knowledge is still not being used, it could be exploited by creating a pack with a product of each of the sub-commodities. The same reasoning applies to the rule {pepperoni/salami", shredded cheese} → {pizza sauce}.

When looking at the Christmas rules with the highest lift, at sub-commodity level, the fact that most of the rules had high lifts catched our attention. After carefully analyse the results, we could conclude that this phenomenon was because there were rules that only happened once, and had a low support, making for that reason a lift extremely high. Taking this information into account, the lift was used as a first filter, but the support was used as a second filter since we want to look at rules that happened at least more than once.

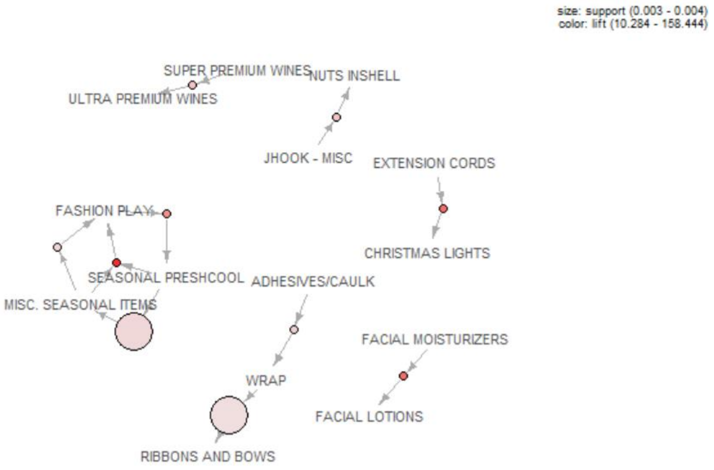


Figure 6 - Top 10 association rules in the Christmas 1 (week 40) dataset by lift and support at the sub-commodity level

Analysing the results in figure 7, the strongest rule, with a lift of roughly 158 and a confidence of 0.667, relates "misc. seasonal items", "seasonal preschool" and "fashion play". These are high values that almost guarantee that the products under these categories are bought together. This rule could be considered a surprise since one would not expect preschool items to be sold in the Christmas season. Although it can be explained by the fact that people usually buy this type of items as presents for children in this period. For that reason, this type of products should also be close to the Christmas department in the store. The second rule with the highest lift relates "extension cords" and "christmas lights" and has a confidence of 1, which means that 100% of clients who bought "christmas lights" in this period, also bought "extension cords". Given that "extension cords" is under the commodity "electrical supplies", this is a not so obvious rule that could beneficiate the retailer if the two products were next to each other in stores. The rule with the third-highest lift can also be considered an interesting one since facial lotions and moisturizers are available all year. However, considering that these products are already next to each other in the stores, to position them close to the Christmas department could result in an uplift of sales, since people would be more exposed to the products, and is also a type of products that are offered as presents during this season.

Furthermore, the rule that relates “super premium wines” and “ultra premium wines”, although relating the same type of products, is interesting for the retailer since premium wines are usually more expensive and can represent an opportunity to increase revenue.

In addition to the points examined, it is worth noting that each rule under analysis associates mostly products of the same family, however, when looking at all the association rules, we observe that several sub-commodities are present. Further in this work, we will propose methods to exploit this information.

In figure 8 is presented the 10 association rules in Christmas 2 (week=92) with the highest support and lift. Given the analysis of frequent itemsets, it was already expected to find the sub-commodity “dog & cat accessories”. Besides that, items under the commodities “baby foods” and “continuity” also differ from the first Christmas.

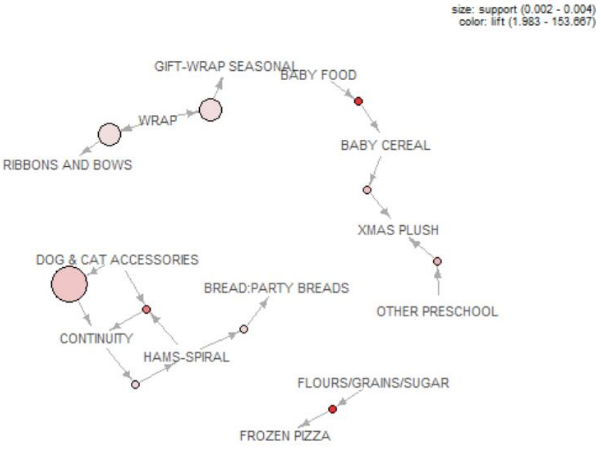


Figure 7 - Top 10 association rules in the Christmas 2 (week=92) dataset by lift and support at the sub-commodity level

We can then conclude that although some rules are similar between Christmas, others are not. This can be explained by the fact that people usually keep products from one year to the other, as can be the case of the Christmas lights.

To study the association rules for the Easter period, the same approach was used as for the Christmas periods. At a first glance, the lift was used in order to have just the more interesting rules, and after that, just the rules with the highest support were presented.

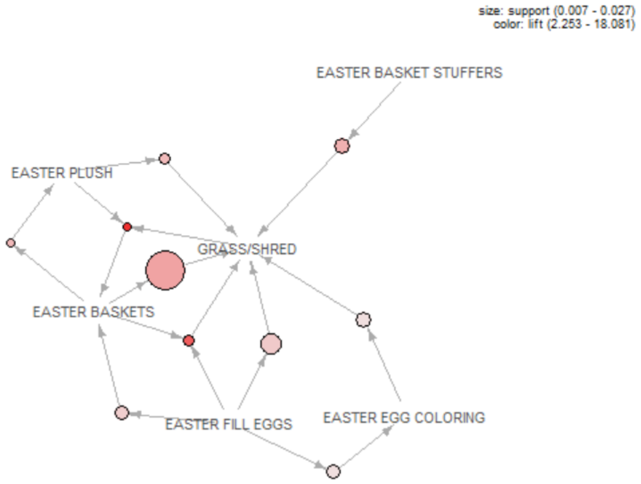


Figure 8 - Top 10 association rules in the easter dataset by lift and support at the sub-commodity level

As it was already expected from the study of the frequent itemsets, most of the subcategories are under the category “easter”. The fact that this commodity does not have a strong relationship with others does not mean that it should not be analysed. Indeed, under the same commodity, strong rules can still be found, as is the case of the rule {easter plush, grass/shred} → {easter baskets}, with a lift of 18. In fact, all the rules have a lift higher than 1, which means that the products that belong to the same rule are more likely to be bought together. The rule with the highest support and the third-highest lift relates “easter baskets” and “grass/shred”. Here assuming that the company does not know about these relations, it can still be interesting to take advantage of this information with the creation of Easter baskets, for example. Later in this work, we will suggest techniques to leverage this information. It is also noteworthy to refer that the first rule involving not only easter products, relates “hams-spiral” and “easter egg coloring”, with a lift of 3.

The study of association rules at commodity and sub commodity levels allowed us to identify in an objective way how the various categories relate with themselves. As already mentioned, an analysis at the product ID level will not be performed due to the high number of items that would generate. However, it would also imply choosing low values for the support and confidence, which would be time-consuming.

3.4. Rare Itemset Mining

The database under analysis contains items with several frequencies. Some items are bought almost every week, and appear thousands of times, like bread and milk, and there are the items that appear rarely in transactions, as can be the case of premium liquors or technology devices. Rare events, in real life, can be of extreme importance and value. The same can happen with Big Data. For that reason, we perform an analysis of rare itemsets, in this section. The same methodology will be used as in the frequent itemsets analysis. In a preliminary phase, all the database will be used to discover the rare itemsets, but since we want to discover the rare itemsets specific to each season and that are not rare in all database, the overall rare itemsets will be removed from the Christmas and Easter datasets.

3.4.1 Data Preparation

The algorithm used in this step is the Apriori-Inverse and its advantages were already explained in chapter 2.4.2. Since it is not available in R software, the discovery of rare itemsets was supported by the SPMF software, an open Data Mining library in Java language. As one of the conclusions in the previous chapter was that an analysis at the sub-commodity level provides more detailed and actionable information, this study will focus on the sub-commodity hierarchical level.

The software uses a very specific file as an input, and for that reason, our data had to be adapted in an eligible way to be accepted by the program. In the first phase, it was necessary to convert the sub-commodity names to numbers since the program only reads items in the form of positive integers. The next step consisted of adapting the database to a transactional format. This process was performed using SQL software, in a way to be directly uploaded in SPMF. The cleaning process was already done in the initial phase of the study, which means that it was not necessary to delete duplicated codes in the same transaction.

In the end, as an input file, we have a text file, in which each line represents a transaction, and each item of the transaction is sorted in ascending order and separated by a single space. As already mentioned, the analysis will once again be conducted for four datasets, for that reason, four input files were prepared.

3.4.2. Modeling

Since the Apriori-Inverse uses a minimum and maximum support, it was necessary to perform some tests to understand which would be the best thresholds for our analysis. The minimum support used was 0.00001 and the maximum support was 0.00005. In appendix 11, a print of the interface inputs is shown. These support values were used four times since a change in the support would manipulate our results. For each time, as an output, a text file is generated where each line represents a rare itemset. In each line, it is then possible to see the rare itemsets, followed by an integer that indicates its support. The number of rare itemsets in each analysis, the running time and the memory consumption, can be consulted in Appendix 12. The results are presented in the next section.

3.4.3. Evaluation

After having the list of the rare itemsets for the selected parameters, all the codes that were used by SPMF were again converted to their sub-commodities descriptions.

itemsets	Absolute support	Sales Value
air conditioners	1	99.99
indoor sports	1	89.99
patio furniture	2	67.48
umbrellas	2	64.99
new designer fragrance	1	49.99
home audio	2	44.44
tricycles/wagons/etc	1	39.98
peripherals	1	39.98
holiday dinner (hot)	2	39.98
promotional furniture	1	33.49

Table 12 - 10 rare itemsets with the highest sales value

Since we are talking about a grocery store, where the average expenditure is 29.21$, marketing strategies to promote rare itemsets with low profit would not represent an advantage of the retailer. Therefore, an average price per sub-commodity was calculated and added to the results given by SPMF. Our attention focused on the most expensive items since these are the ones expected to have a higher profit margin. Table 12 shows the 10 rare itemsets with the highest sales value.

The support of the discovered rules is low, mainly with an absolute support of 1 or 2, but this was already expected since we are trying to find associations between rare items.

It is noteworthy to refer that the itemsets "air conditioners", "indoor sports", "patio furniture" and "umbrellas" belong to the commodity "spring/summer seasonal". Although these transactions were not realized during the easter weeks under analyse, this information could be used by the retailer to create marketing campaigns during the spring and summer periods. The other sub-commodities do not concern seasonal categories and for that reason, to know which high-utility rare itemsets we have during Christmas and Easter, the approach used to discover calendar-based temporal association rules, was again used, implying that all the sub-commodities that were present in our first output file were removed from the Christmas and Easter datasets. The rare itemsets with the highest sales value, in each analysis, can be consulted in Appendix 13.

In table 13, we can find the association rules between rare itemsets in the Christmas dataset, for the minimum support of 0.001%, maximum support of 0.005% and confidence of 0.01%. Once again, the average sales value was added since we are interested in the profit of each transaction. Although the confidence and lift values were not computed by default in the SPMF, they are present in the tables of results for greater uniformity in the presentation of results.

Antecedent	Consequent	Absolute support	Confidence	Lift	Sales value
area rugs	collegiate	1	1	2529	12.89
flavored rum (42 under proof)	canadian whiskey(42 under pro	1	1	2529	9.66
cooking wines	breast - bone-in (frz)	1	1	2529	8.81
vases&containers	home decor	1	1	2529	8.16
candy boxed chocolates w/flour	rib - stk/chp/slc	1	1	2529	8.38

Table 13 - 5 rare itemsets in Christmas with the highest sales value

When looking at the confidence of the association rules, we can already infer that none of the subcategories appears in a transaction without the other, for that reason the confidence is 1 for all the rules. Therefore, the lift of the association rules is also equal for all of them. The average sales value was added and the itemsets ordered by sales value.

The association rule with the highest average price is only 12 dollars, which although may seem not a lot, when considering the average value spent by customers, it is almost half, so it may represent an opportunity for the retailer. A category that stands out is the wines/alcoholics drinks. Looking at the second rules, {flovered rum} → {canadian whiskey}, we see an opportunity to the retailer, not only for the fact that is the second-highest when considering the sales value but also because one of the association rules that appeared in the Christmas dataset connected premium wine. The creation of a 3 products pack with two frequent sold premium wines and a rare alcoholic drink with a high margin for the retailer can be considered a good strategy to increase the sale of these products. It is also interesting to notice that the fourth rule relates products under the commodity "Floral-hard good", which is not surprising since, because of the weather, people do not tend to buy this type of products during Christmas.

Antecedent	Consequent	Absolute support	Confidence	Lift	Sales value ($)
diet control tablets/capsule	hair nets and caps	1	1	3080	9.83
traditional vodka	traditional rum	1	1	3080	8.85
nut supp-c vitamins	nut supp-homeopathy	1	1	3080	8.10
power toothbrush/refills	diet/nutrition vitamin (adult	1	1	3080	7.97
toast/griddles	puzzles	1	1	3080	6.96

Table 14- Top 5 rare itemsets in Christmas2 (week=92) with the highest sales value

In table 14, the top 5 rare itemsets in Christmas 2 (week=56) are presented. Once again, the commodity "liquor" stands out and it is present in the antecedent and consequent of rule with the second highest average sales value. The rare association rules in Christmas 2, unlike Christmas 1, has present the commodities "cold and flu" and "vitamins". This may indicate that this type of products is not bought every year or are bought in different periods each year. It would be interesting to perform the same analysis for another 2 years to confirm if these suspicions are confirmed.

Antecedent	Consequent	absolute support	confidence	lift	sales value
deli tray:meat and cheese	standard annuals (outdoor)	1	1	2723	14.53
whitening systems	kids	1	1	2723	12.72
healing garden	whitening systems, kids	1	1	2723	11.20
fem hygn douches	deli tray:meat and cheese, standard annuals (outdoor)	1	1	2723	10.44
stuffed/mixed pork	corned beef	1	1	2723	8.19

Table 15 -Top 5 rare itemsets in easter with the highest sales value

Looking at table 15, we can find the 5 rare association rules with the highest sales value in the easter dataset. Once again, all the items appear only once in the dataset. The association rule between rare itemsets with the highest average sales value relates "deli tray: meat and cheese" and "standard annuals(outdoor)". The sub-commodity "whitening systems", part of the commodity "oral hygiene products" appears associated with "kids". Since these two products both refer to care/hygiene products, it could be a great opportunity for the retailer to sell the two products together.

3.5. High Utility Itemset Mining

The goal of each retailer is of course to maximize revenue. For that reason, it is extremely important to take into account the price or margin of a product. It can be more advantageous for a retailer to sell a lower quantity of products with a high margin than a high quantity of product with low value.

The detection of frequent itemsets and association rules, although useful to understand how items relate, presents a disadvantage. It does not consider the quantities bough of an item or the profit that a product can provide, which is rather important to consider when defining marketing strategies. In the previous section, the sales value was considered when deciding the most interesting rare association rules, however, we were just looking at rare itemsets, which means that they had low support. What if there are itemsets that are not rare and provide high utility to the company? To answer this question we perform a complementary study to discover high utility itemset mining in this chapter. Not only products bought more than once will be considered, but also the sales value will play an important role. The "sales value" represents the dollars received by the retailer on the sale of the specific product, taking any discount into account Here, the assumption that products with higher sales value have also a higher margin for the retailer was made.

3.5.1 Data Preparation

The algorithm chosen, for the reasons presented in chapter 2.4.3, is the EFIM. This algorithm is not available in R software, and for that reason the Data Mining library SPMF was once again chosen to support this study.

This time the initial database was used as an initial point, but we did not delete the variables "sales value" and "quantity". All the other steps performed in the data cleaning phase were realized again. Since the system only accepts files with numbers, all the sub-commodities descriptions were converted to codes, that posteriorly were converted again to descriptions.

The file used as an input is also a text file, however this time the data had to be disposed differently to be eligible to be accepted by the program. After adapting the database to a transactional format, an SQL query was used to perform the modifications missing. In the input file, each line represents a transaction, and this is split into three sections, divided by the symbol ":". First, in each line, appears the products that are part of the transaction. In the second section, we can find the total utility of the transaction, which, in this case, is the sum of the sales value for each transaction. The total utility is followed by the utility of each product in the transaction, separated by singles spaces. An example of the input file can be seen in Appendix 14.

The same methodology was applied for four datasets, meaning that 4 input files were prepared. However, this time, only the high utility itemsets of 1 week were removed from the other databases, since if we considered all the database nothing would be left in the seasonal databases. Week 15 was chosen as representative of the dataset because it was one of the weeks with more transactions but also because it is distant from Christmas and Easter weeks.

3.5.2. Modeling

Considering that the software faced issues reading such long transactions, and since our goal is to access which itemsets have high utility, only items with sales value higher than 10 were considered. The minimum utility used was 30 in the 4 analyses. In Appendix 15, a print of the interface input is shown.

For each analysis, an output text file is generated, in which each line has the high utility itemset followed by its utility. Once again, all the information regarding the number of high utility itemsets, the running time and memory used can be consulted from appendix 16. The results are presented and discussed in the following section.

3.5.3. Evaluation

The outputs of SPMF were loaded into Excel and the codes used for the sub-commodity were matched with their correspondent descriptions. Table 16 shows the 5 itemsets with the highest utility in week 15.

Itemsets	Utility ($)
beeralemalt liquors - beers/ales	1870
select beef - beef	572
baby diapers - diapers & disposables	546
natural beef - beef	310
loin - stk/chp/slc - beef	260

Table 16 - 5 itemsets with the highest utility in week 15

According to the table above, the itemset with the highest sales value is "beeralemalt liquors" with a utility of 1870 $. Its commodity was already present in the most frequent itemsets analyse. We can also see that the category "beef" has a significant presence in this analysis as it is present in 3 out of 5 high utility itemsets. In third place, we can see the baby diapers that did not appear in any of the other analyses realised.

Once again, we can conclude that the products we found are available all the year and we cannot say that is not typical of a specific season. For that reason, the same methodology as before will be used here again. All the sub-commodities found in the output file of the SPMF, when analysing week 15, will be removed from the Christmas and Easter datasets. For the calendar-based association rules, only sets composed of at least 2 items will be discussed, but the high utility itemsets with the highest utility are presented in Appendix 17. In addition, objective measures such as confidence and lift will be included to complement the previous work.

Antecedent	**Consequent**	**Confidence**	**Lift**	**Utility ($)**
mastercard gift card	electronic gift cards activati	0.50	7	77
premium 1.5lt wines	popular 1.5lt wines	1.00	37	66
fruit baskets	electronic gift cards activati	0.33	5	54
fruit baskets	hams-whole boneless	0.33	2	44
angus beef	deli tray:sandwiches	0.50	56	42

Table 17 - 5 itemsets with the highest utility in Christmas 1 (week=40)

All the above itemsets present a lift higher than 1, which means that all the rules are strong. Besides, as it was our goal, they all have high utilities associated. The first rule, with a utility of 77, relates to two different types of gift cards. Indeed, it is not a surprise that the two products have the highest utility in this period since this type of products usually has a minimum value to be activated, and besides that, they are more common at Christmas, since people can offer it. Premium and popular wines are the itemsets characteristic of the period that provide the second highest utility. It is interesting to notice that the rule{super premium wines} → {ultra premium wine}does not appear in this top 5 because the "ultra premium wines" was one of the sub commodities removed. This means that, although the itemsets are usually bought together during the first Christmas 1, the "ultra premium wines" also generates a considerable amount of income in other periods. Also noteworthy is the rule {angus beef} → {deli tray:sandwiches}, that although it has the lowest utility in this ranking, it has the biggest lift. Therefore, these are the itemsets that are more positively correlated among this top.

Antecedent	Consequent	Confidence	Lift	Utility
electronic gift cards activati	outside vendors gift cards	0.20	2	189
deli tray:sandwiches	outside vendors gift cards	0.50	6	67
turkey breast bone in	hams-whole boneless	0.14	2	63
reading glasses, designer fragrances	angus beef	1.00	31	58
fm-nfl (apparel)	outside vendors gift cards	1.00	12	57

Table 18 - 5 itemsets with the highest utility in Christmas 2 (week=92)

Looking at the itemsets with the highest utility for the second Christmas, in table 18, we can see a lot of similarities with the first Christmas. There is an abundance of sub-commodities related to gift cards and the rule with the highest utility also relates two types of gift cards. The third rule relates "turkey breast bone in" and "hams-whole boneless". This last one already appeared in the first Christmas, but relating with "fruit baskets". In fact, the turkey sub commodity also appears in the top 15 of the high utility of the first Christmas. However, in the second Christmas, it is usually bough with other itemsets with high sales values. Indeed, the confidence is lower for this rule, indicating that the antecedent is related to a large number of items.

Antecedent	Consequent	Confidence	Lift	Utility
fruit party tray	hams-spiral	0.50	5	36
coff shop: retail pack beverag	hams-spiral	1.00	11	34
reading glasses	hams-whole bone-in	0.50	11	33
hardback/trade best seller	bourbon/tn whiskey (42 under p	0.50	34	33
seafood-frz-raw shlfsh-other	angus beef	0.14	3	31

Table 19 - 5 itemsets with the highest utility in Easter

At a first glance, when looking at table 19, we can already conclude that the utility generated during the easter weeks is lower than in the Christmas weeks. Besides, hams seem to play an important role in what concerns utility in week 56. The combination of this information with the fact that the sub-commodity "hams-half/port bone-in" was among of the most frequent can be exploited by the retailer. Besides, "fruit party tray" is associated with a ham category in this top, but also in the analysis of the first Christmas. The rule with the highest lift relates a type of book with a whiskey.

To finish, also interesting, the fifth rule relates to seafood and beef. Although these products are not usually consumed together, they represent high utility for the retailer when bough together, and, for that reason, some marketing actions should be taken in this regard. From this analysis, it can be concluded, that it is in the retailer interest to consider the utility of an itemsets and not only objective measure such as the lift to decide which rules are the interesting ones.

4. Conclusions

4.1 Results discussion

The discovery of association patterns, in each period, is of great use to companies in domains of marketing and logistics, allowing companies to apply different marketing strategies and product offering to distinct periods, and enhance revenues by making recommendations based on data and not in what is thought to be more popular.

Given that the association rules under analysis are seasonal, a practical application that could be explored by the retailer is the store layout. The products can be arranged in a strategic way that allows to increase profit. On the one hand, if the goal is to minimize the time a client needs to find all the products needed, the proposal includes the disposable of the items that belong to association rules with high lift next to each other in order to influence the client by the convenience of the two products be close to each other. For example, during the Christmas period, "extension cords" should be placed next to "Christmas lights". Additionally, during the Easter period, a small fridge with products of the sub-commodity "hams-spiral" could be placed close to the Easter periods. On the other hand, the retailer could also choose to arrange the products with high confidence far from each other since these products are dependent form each other and would make the client exposed to more products. Moreover, the position of the products in the storage could also be arranged to increase efficiency when picking the products.

Although consumers can have different buying patterns when comparing physical and online shopping, it can be assumed that if two products are usually bought together in a physical store, there is also a high probability of the itemsets being bought together in an online store. Therefore, the conclusions discussed above could be incorporated into an online platform in the form of recommendation systems. Every time a product would be added to the online store cart, the platform would suggest 3 other products that could be interesting for the client. This suggestion would be based on the most frequent bought together, for example, if a "super premium wine" was added, an "ultra premium wine" would be recommended. It would also recommend a high utility rare itemset of another sub-commodity, for example, a product under the sub-commodity "bourbon/tn whiskey". The third product would be a high utility item, for instance, a product under "Hams-whole boneless". The same reasoning can be applied to the Easter period. Let's consider the example of a client who adds an “easter egg coloring”, then it could be recommended an “easter fill eggs” item, a “premium 1.5lt wine” (rare high-utility itemset) and a product under “hams-whole boneless” (high-utility itemset).

Taking advantages of the seasons and the fact that people are usually willing to spend more money during these periods, a good strategy of cross-selling would be the creation of “baskets” with a small reduction of the price. These baskets should include frequently bought together items but should also be combined with rare and high utility, allowing us to take advance of our discoveries and increase revenue for the retailer. A possible Christmas basket can include products of the following sub-commodities “facial lotions”, “facial moisturizers”, “nuts inshell”, “candy boxed chocolates w/ flour”, “rib - stk/chp/slc”, “premium 1.5lt wines” and popular “1.5lt wines”. This basket would, this way, combine the information discovered from the frequent itemsets analysis, and the three association rules studies. Another recommendation, considering the easter period this time, would be to create an Easter basket with all the most frequent bought Easter products. This type of strategies can make clients buy products that they usually would not just because the products are sold together and advertised as a cheaper choice.

The results discovered can also be useful for the company in the case there are launches of new products or products not performing as well as expected. So, if the item does not achieve the sales target, but the product is usually sold with a popular product, the company can apply a discount to the popular product with the condition that the worst-performing product is bought at the same time. This technique not only would boost the sales of the product less sold, but also increase sales of the popular product. For example, during Christmas, if “facial moisturizers” has a bad performance, a discount should be applied to “facial lotions”, which is a product frequently bought in this season when people also buy “facial moisturizers”. An alternative could be the creation of a pack with the two products. In the Easter period, the same approach could be used when one of the easter products is not performing as good as expected, a discount could be offered when people would also buy “pansies”, a popular product during this period.

Furthermore, the retailer may provide promotional incentives to convince a client who went on a certain shopping trip and purchased items that largely belong to a specific cluster to return to the store and purchase items from a different cluster. This strategy could be applied by sending coupons via phone or email to targeted consumers.

4.2 Summary

Data Mining techniques are applied to huge datasets and increasingly allow to explore patterns and remove knowledge for business management and decision making. A Data Mining tool that can be used for this purpose is the Market Basket Analysis. Indeed, it has proven to be of enormous relevance, since it helps companies to provide a customized service, which in the end allows to create a longer relationship with customers. Traditional frequent pattern mining neglects pattern occurrence behaviour. However, different association rules can be uncovered while considering different periods. Therefore, to understand customer behaviour, it is desirable to discover patterns that usually occur in a certain period considered.

In this work, using the CRISP-DM model, a methodology to discover calendar-based temporal association rules was proposed. The analyse focused on the Christmas and Easter periods, by removing all the itemsets that were considered to be present during all year, and for that reason not relevant for our analysis.

This study comprises three different analyses. In the first one, the frequent itemsets, for each period under analysis, were detected and discussed, for the commodity and sub-commodity hierarchy levels, followed by the study of association rules. However, this study had a limitation since it only considered the frequency of occurrence of the events. To overcome this constraint, two other studies were also presented at the sub-commodity level. We realised an attempt to discover rare high utility itemsets, which allowed considering less frequent events, but only looking to the most profit for the retailer since rare itemsets with low utility are not interesting for the retailer, given that there are items with low sales value. In the last analysis, we realised a study of high utility itemsets which allowed us to take only into account the utility of an itemset.

It can be concluded that the consumer behaviour at Christmas is very different from the Easter behaviour for the 3 analyses realised. The consumer's buying patterns can also vary in two different Christmas, especially when considering frequent and rare association rules. It would be interesting to understand if this behaviour changes every year or if it is constant every two years, for example.

The analysis of the results of the 3 studies elaborated in the previous chapter, when merged, allows extracting interesting and valuable knowledge that permits the proposal of techniques easily applicable by retailers, with the purpose of attracting customers and boosting sales. In the previous chapter, several techniques that could be applied by the retailer were proposed. Though the study provides an exhaustive picture of the client's behaviour in two seasons, it had some limitations, but also space for further enhancement.

4.3. Limitations

The fact that the database used is public allows one to replicate the work developed, which represents an advantage. Besides, this database provides a wide range of transactions and important information that was used in this work. However, the lack of information of the company and more detailed information of the products restricted the quality of the conclusions and suggestion, since a deeper understanding of the business and its context, would provide more insightful conclusions.

In addition to the database constraints, the discovery of interesting rules was sometimes hard since redundant rules were not eliminated by the algorithm or objective measures. This is considered a limitation due to the limited knowledge of the database, which prevents knowing which associations are already known and exploited by the retailer.

4.4 Future Work

A possible solution for the redundancy problem suggested in the previous section could be the application of other objective measures besides the lift, which could enable the discovery of interesting rules that were ignored in this study. Since the database provides information on the different stores' id, transaction time and clients profile, it would be interesting to apply a clustering analysis. This type of study would allow us to understand if certain associations happen more frequently at a store ID or at a specific range of time. Also, the use of the customer's profile would allow to cluster them by groups, the company could this way optimize marketing resources by defining better its target for marketing campaigns. In addition, the calendar-based temporal association rules could be complemented by a sequential rule mining analysis. This study would allow discovering if when a product is bought in a certain period, it is usually followed by a purchase of another product in the future. It has immense applications and includes valuable insight into consumer's complex behaviour.

References

Adda, M., Wu, L., & Feng, Y. (2008). Rare Itemset Mining (pp. 73–80). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/icmla.2007.106

Aggelis, V. (2004). Association rules model of e-banking services. In *Management Information Systems* (Vol. 10).

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, 207–216. https://doi.org/10.1145/170035.170072

Agrawal, R., & Srikan, R. (1994). Fast Algorithms for Mining Association Rules. *Proc. 20th Int. Conf. Very Large Data Bases*, *VLDB*(1215), 487–499.

Al-hegami, A. (2004). Subjective Measures and their Role in Data Mining Process. In *In Proceedings of the 6th International Conference on Cognitive Sytems*.

Ayu, S. K., Surjandari, I., & Zulkarnain, Z. (2019). Mining Association Rules in Seasonal Transaction Data. In *Proceedings - 2018 5th International Conference on Information Science and Control Engineering, ICISCE 2018*. IEEE. https://doi.org/10.1109/ICISCE.2018.00074

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

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 437–456. https://doi.org/10.1002/widm.1074

Brammer, M. (2013). Basic principles of data mining. In *Principles of Data Mining*. https://doi.org/10.1007/978-1-4471-4884-5

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

Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. In *Expert Systems with Applications* (Vol. 28, Issue 4). https://doi.org/10.1016/j.eswa.2004.12.033

Djenouri, Y., Gheraibia, Y., Mehdi, M., Bendjoudi, A., & Nouali-Taboudjemat, N. (2014). An efficient measure for evaluating association rules. In *6th International Conference on Soft Computing and Pattern Recognition, SoCPaR 2014*. https://doi.org/10.1109/SOCPAR.2014.7008041

Dongre, J., Prajapati, G. L., & Tokekar, S. V. (2014). The role of Apriori algorithm for finding the association rules in Data mining. In *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014*. https://doi.org/10.1109/ICICICT.2014.6781357

Eavis, T., & Zheng, X. (2009). Multi-level frequent pattern mining. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 5463, pp. 369–383). https://doi.org/10.1007/978-3-642-00887-0_33

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases* (Vol. 17, Issue 3). https://doi.org/10.1609/aimag.v17i3.1230

Fournier-Viger, P., Chun-Wei Lin, J., Truong-Chi, T., & Nkambou, R. (2019). *A Survey of High Utility Itemset Mining* (pp. 1–45). https://doi.org/10.1007/978-3-030-04921-8_1

Fournier-Viger, P., Lin, J. C. W., Vo, B., Chi, T. T., Zhang, J., & Le, H. B. (2017). A survey of itemset mining. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 7, Issue 4). https://doi.org/10.1002/widm.1207

Goyal, V., Dawar, S., & Sureka, A. (2015). High Utility Rare Itemset Mining over Transaction Databases. In: *Chu W., Kikuchi S., Bhalla S. (eds) Databases in Networked Information Systems. DNIS 2015. Lecture Notes in Computer Science, vol 8999. Springer, Cham.* https://doi.org/10.1007/978-3-319-16313-0_3

Grönroos, C. (2011). A service perspective on business relationships: The value creation, interaction and marketing interface. In *Industrial Marketing Management* (Vol. 40, Issue 2). Elsevier Inc. https://doi.org/10.1016/j.indmarman.2010.06.036

Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets. *Journal of Machine Learning Research,* 12, 2021–2025.

Han, J., Micheline Kamber, & Jian Pei. (2012). Data mining : concepts and techniques. In *Elsevier Inc.* https://doi.org/https://doi.org/10.1016/C2009-0-61819-5

Hemalatha, M. (2012). Market basket analysis - A data mining application in Indian retailing. In *International Journal of Business Information Systems* (Vol. 10, Issue 1). https://doi.org/10.1504/IJBIS.2012.046683

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. In *Procedia CIRP* (Vol. 79). Elsevier B.V. https://doi.org/10.1016/j.procir.2019.02.106

Kaggle - The Complete Journey. (n.d.). Retrieved December 5, 2020, from https://www.kaggle.com/frtgnn/dunnhumby-the-complete-journey?select=transaction_data.csv

Kamakura, W. A. (2008). Cross-Selling. *Journal of Relationship Marketing, 6*(3–4), 41–58. https://doi.org/10.1300/J366v06n03_03

Kamakura, W. A., Wedel, M., de Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. In *International Journal of Research in Marketing* (Vol. 20, Issue 1). https://doi.org/10.1016/S0167-8116(02)00121-0

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. In *Expert Systems with Applications* (Vol. 40, Issue 4). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2012.08.039

Kassim, N., & Abdullah, N. A. (2010). The effect of perceived service quality dimensions on customer satisfaction, trust, and loyalty in e-commerce settings: A cross cultural analysis. In *Asia Pacific Journal of Marketing and Logistics* (Vol. 22, Issue 3). https://doi.org/10.1108/13555851011062269

Koh, Y. S., & Rountree, N. (2005). Finding sporadic rules using Apriori-Inverse. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 3518 LNAI*. https://doi.org/10.1007/11430919_13

Lam, S. Y., Shankar, V., Erramilli, M. K., & Murthy, B. (2004). Customer Value, Satisfaction, Loyalty, and Switching Costs: An Illustration From a Business-to-Business Service Context. *Journal of the Academy of Marketing Science*, *32*(3), 293–311. https://doi.org/10.1177/0092070304263330

Leninkumar, V. (2017). The Relationship between Customer Satisfaction and Customer Trust on Customer Loyalty. *International Journal of Academic Research in Business and Social Sciences*, *7*(4). https://doi.org/10.6007/ijarbss/v7-i4/2821

Leppäniemi, M., Karjaluoto, H., & Saarijärvi, H. (2017). Customer perceived value, satisfaction, and loyalty: the role of willingness to share information. In *International Review of Retail, Distribution and Consumer Research* (Vol. 27, Issue 2). Routledge. https://doi.org/10.1080/09593969.2016.1251482

Li, J., Le, T. D., Liu, L., Liu, J., Jin, Z., & Sun, B. (2013). Mining causal association rules. *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*, 114–123. https://doi.org/10.1109/ICDMW.2013.88

Li, Y., Ning, P., Wang, X. S., & Jajodia, S. (2001). Discovering calendar-based temporal association rules. In *Proceedings of the International Workshop on Temporal Representation and Reasoning*. https://doi.org/10.1109/TIME.2001.930706

Lin, H. K., Hsieh, C. H., Wei, N. C., & Peng, Y. C. (2019). Association rules mining in R for product performance management in industry 4.0. In *Procedia CIRP* (Vol. 83). Elsevier B.V. https://doi.org/10.1016/j.procir.2019.04.099

Liu, B., Hsu, W., & Ma, Y. (1999). Mining Association Rules with Multiple Minimum Supports. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 337–341.

Liu, M., & Qu, J. (2012). Mining high utility itemsets without candidate generation. In *ACM International Conference Proceeding Series*.https://doi.org/10.1145/2396761.2396773

Liu, S., & Pan, H. (2018). Rare itemsets mining algorithm based on RP-Tree and spark framework. In *AIP Conference Proceedings* (Vol. 1967). https://doi.org/10.1063/1.5039144

Liu, Y., Li, J., Liao, W.-K., Choudhary, A., & Shi, Y. (2010). High Utility Itemsets Mining. *International Journal of Information Technology & Decision Making*, *09*(06), 905–934. https://doi.org/10.1142/S0219622010004159

Lonlac, J., Doniec, A., Lujak, M., & Lecoeuche, S. (2020). *Extracting Seasonal Gradual Patterns from Temporal Sequence Data Using Periodic Patterns Mining*. http://arxiv.org/abs/2010.10289

Martínez-Ballesteros, M., Martínez-Álvarez, F., Troncoso, A., & Riquelme, J. C. (2014). Selecting the best measures to discover quantitative association rules. In *Neurocomputing* (Vol. 126). Elsevier. https://doi.org/10.1016/j.neucom.2013.01.056

McKinsey & Company. (2015). *Industry 4.0 How to navigate digitization of the manufacturing sector*. https://www.mckinsey.com/~/media/McKinsey/Business Functions/Operations/Our Insights/Industry 40 How to navigate digitization of the manufacturing sector/Industry-40-How-to-navigate-digitization-of-the-manufacturing-sector.ashx

Moonen, L., Di Alesio, S., Binkley, D., & Rolfsnes, T. (2016). Practical guidelines for change recommendation using association rule mining. In *ASE 2016 - Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*. https://doi.org/10.1145/2970276.2970327

Ninoria, S. Z., & S. S. Thakur. (2017). A Survey on High-utility Itemsets Mining. *International Journal of Computer Applications*, *175*(4), 43–50. https://doi.org/10.5120/ijca2017915521

Ordonez, C., Omiecinski, E., De Braal, L., Santana, C. A., Ezquerra, N., Taboada, J. A., Cooke, D., Krawczynska, E., & Garcia, E. V. (2001). Mining constrained association rules to predict heart disease. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/icdm.2001.989549

Phan, D. D., & Vogel, D. R. (2010). A model of customer relationship management and business intelligence systems for catalogue and online retailers. In *Information and Management* (Vol. 47, Issue 2). Elsevier B.V. https://doi.org/10.1016/j.im.2009.09.001

Pillai, J., & Vyas, O. P. (2010). Overview of Itemset Utility Mining and its Applications. In *International Journal of Computer Applications* (Vol. 5, Issue 11). https://doi.org/10.5120/956-1333

Pitta, D. (1998). Marketing one-to-one and its dependence on knowledge discovery in databases. *Journal of Consumer Marketing*, *15*(5), 468–480. https://doi.org/10.1108/EUM0000000004535

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. In *Big Data* (Vol. 1, Issue 1). https://doi.org/10.1089/big.2013.1508

Rodríguez-González, A. Y., Lezama, F., Iglesias-Alvarez, C. A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & de Cote, E. M. (2018). Closed frequent similar pattern mining: Reducing the number of frequent similar patterns without information loss. In *Expert Systems with Applications* (Vol. 96). https://doi.org/10.1016/j.eswa.2017.12.018

Sadhasivam, K. S. C., & Angamuthu, T. (2011). Mining rare itemset with automated support thresholds. *Journal of Computer Science*, *7*(3), 394–399. https://doi.org/10.3844/jcssp.2011.394.399

Saltz, J. (2020). *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects*. https://www.datascience-pm.com/crisp-dm-still-most-popular/

Sánchez, D., Vila, M. A., Cerda, L., & Serrano, J. M. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications*, *36*(2 PART 2), 3630–3640. https://doi.org/10.1016/j.eswa.2008.02.001

Sarra, C. (2020). Data Mining and Knowledge Discovery. Preliminaries for a Critical Examination of the Data Driven Society. In *Global Jurist* (Vol. 20, Issue 1). https://doi.org/10.1515/gj-2019-0016

Schafer, F., Zeiselmair, C., Becker, J., & Otten, H. (2018). Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes. *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 190–195. https://doi.org/10.1109/ITMC.2018.8691266

Schmitz, C. (2013). Group influences of selling teams on industrial salespeople's cross-selling behavior. In *Journal of the Academy of Marketing Science* (Vol. 41, Issue 1). https://doi.org/10.1007/s11747-012-0304-7

Sethi, R., & Shekar, B. (2019). Subjective Interestingness in Association Rule Mining: A Theoretical Analysis. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 21). https://doi.org/10.1007/978-3-319-93940-7_15

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). In *International Journal of Innovation and Scientific Research* (Vol. 12, Issue 1).

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. In *Decision Support Systems* (Vol. 31, Issue 1). https://doi.org/10.1016/S0167-9236(00)00123-8

Shrivastava, S., & Johari, P. K. (2017). Analysis on high utility infrequent itemsets mining over transactional database. In *2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings*. https://doi.org/10.1109/RTEICT.2016.7807958

Smith, B., & Linden, G. (2017). Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing*, *21*(3), 12–18. https://doi.org/10.1109/MIC.2017.72

Sutha, M. J., & Dhanaseelan, F. R. (2017). Mining frequent, maximal and closed frequent itemsets over data stream - A review. In *International Journal of Data Analysis Techniques and Strategies* (Vol. 9, Issue 1). https://doi.org/10.1504/IJDATS.2017.10004000

Szathmary, L., Napoli, A., & Valtchev, P. (2007). Towards rare itemset mining. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* (Vol. 1). https://doi.org/10.1109/ICTAI.2007.30

Tambe, P. (2014). Big Data Investment, Skills, and Firm Value. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2294077

Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S., & Lee, Y.-K. (2009). Discovering Periodic-Frequent Patterns in Transactional Databases. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8302 LNCS* (pp. 242–253). https://doi.org/10.1007/978-3-642-01307-2_24

Verma, K., & Vyas, O. P. (2005). Efficient calendar based temporal association rule. *ACM SIGMOD Record*, *34*(3), 63–70. https://doi.org/10.1145/1084805.1084818

Venkatesan, R. (2007). The Complete Journey. Dunnhumby. Retrieved December 5, 2020, from https://www.dunnhumby.com/source-files

Yongmei, G., & Fuguang, B. (2015). The Research on Measure Method of Association Rules Mining. In *International Journal of Database Theory and Application* (Vol. 8, Issue 2). https://doi.org/10.14257/ijdta.2015.8.2.23

Yoo, J., & Park, M. (2016). The effects of e-mass customization on consumer perceived value, satisfaction, and loyalty toward luxury brands. In *Journal of Business Research* (Vol. 69, Issue 12). Elsevier B.V. https://doi.org/10.1016/j.jbusres.2016.04.174

Zida, S., Fournier-Viger, P., Lin, J. C. W., Wu, C. W., & Tseng, V. S. (2017). EFIM: a fast and memory efficient algorithm for high-utility itemset mining. *Knowledge and Information Systems*, *51*(2), 595–625. https://doi.org/10.1007/s10115-016-0986-0

Appendix

Appendix 1 - The Steps That Comprise the KDD Process

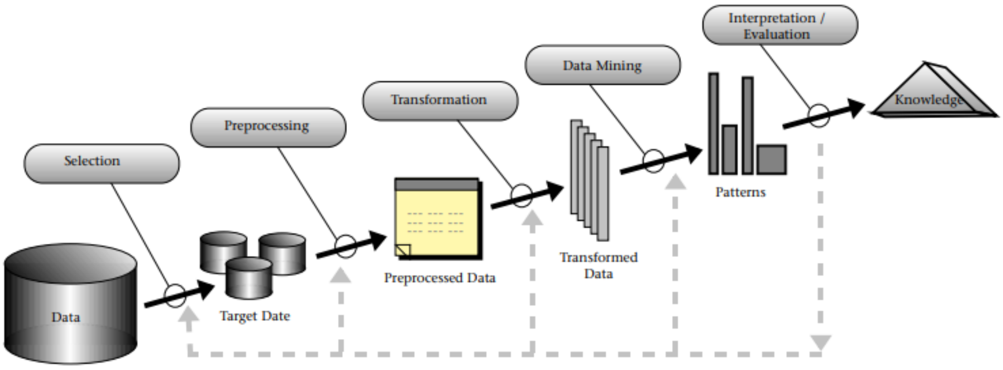


Figure A 1 - The Steps That Comprise the KDD Process

<u>Source:</u> Fayyad, Piatetsky & Smyth (1996)

Appendix 2 - Phases of the CRISP-DM Process Model for Data Mining

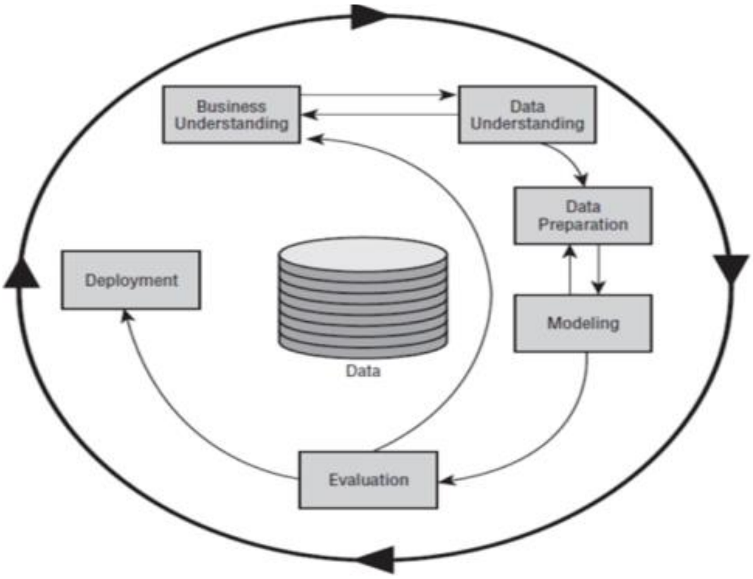


Figure A 2 - Phases of the CRISP-DM Process Model for Data Mining

Source: Shafique & Qaiser, 2014)

Appendix 3 – Data Mining tasks

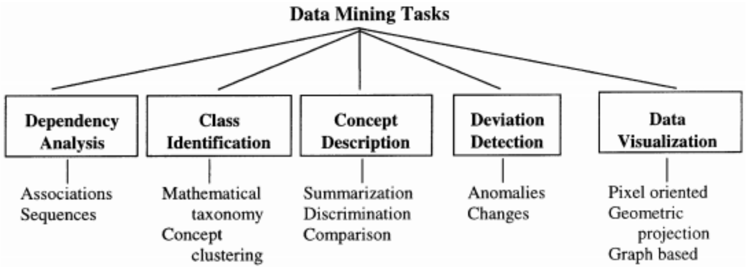


Figure A 3 – Scheme of the Data Mining tasks
Source: Shaw et al.,2001

Appendix 4 – Relation between frequent, closed and maximal itemsets

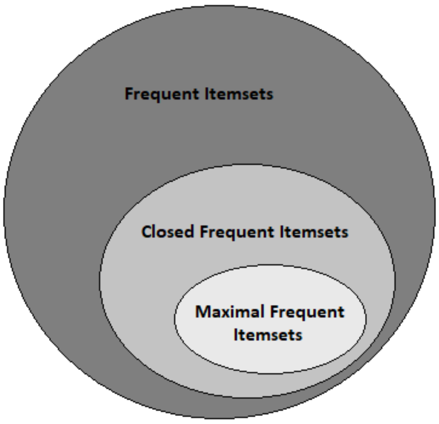
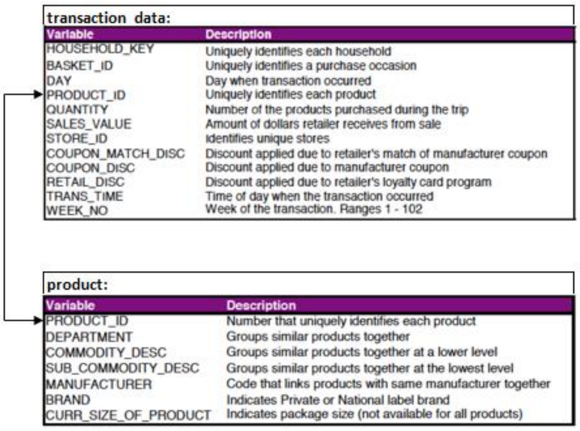


Figure A 4 - Relational scheme between frequent, closed and maximum itemsets

Appendix 5 – Datasets aggregation



transaction_data:

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
BASKET_ID	Uniquely identifies a purchase occasion
DAY	Day when transaction occurred
PRODUCT_ID	Uniquely identifies each product
QUANTITY	Number of the products purchased during the trip
SALES_VALUE	Amount of dollars retailer receives from sale
STORE_ID	Identifies unique stores
COUPON_MATCH_DISC	Discount applied due to retailer's match of manufacturer coupon
COUPON_DISC	Discount applied due to manufacturer coupon
RETAIL_DISC	Discount applied due to retailer's loyalty card program
TRANS_TIME	Time of day when the transaction occurred
WEEK_NO	Week of the transaction. Ranges 1 - 102

product:

Variable	Description
PRODUCT_ID	Number that uniquely identifies each product
DEPARTMENT	Groups similar products together
COMMODITY_DESC	Groups similar products together at a lower level
SUB_COMMODITY_DESC	Groups similar products together at the lowest level
MANUFACTURER	Code that links products with same manufacturer together
BRAND	Indicates Private or National label brand
CURR_SIZE_OF_PRODUCT	Indicates package size (not available for all products)

Figure A 5 – Scheme of the aggregation between the "transaction_data" and "product" datasets

Appendix 6 – Histogram of the most frequency departments

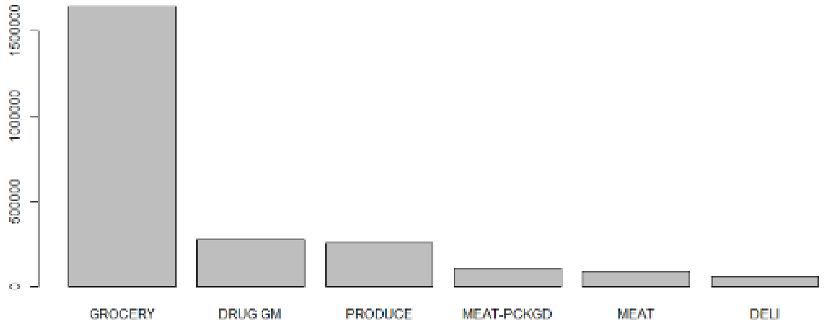


Figure A 6 – Histogram of the 6 most frequency departments

Appendix 7 – Distribution of the commodities “christmas seasonal” and “easter”

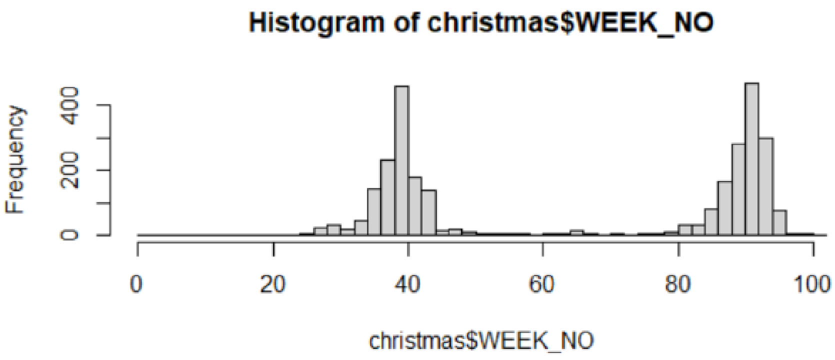


Figure A 7 – Distribution of the commodity “christmas seasonal” by the 102 weeks

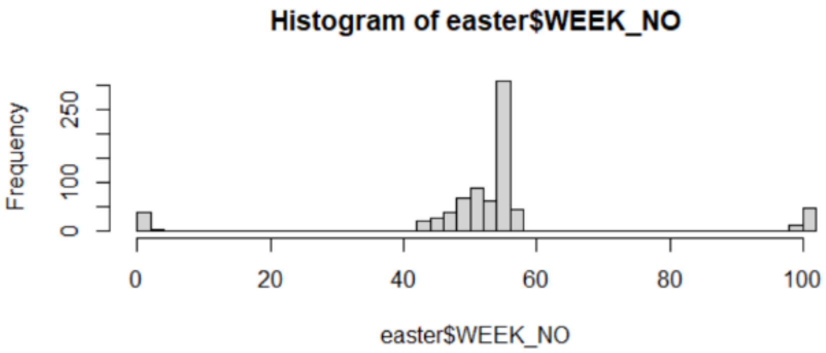


Figure A 8 - Distribution of the commodity “easter” by the 102 weeks

Appendix 8 – Scheme of the creation of the different files

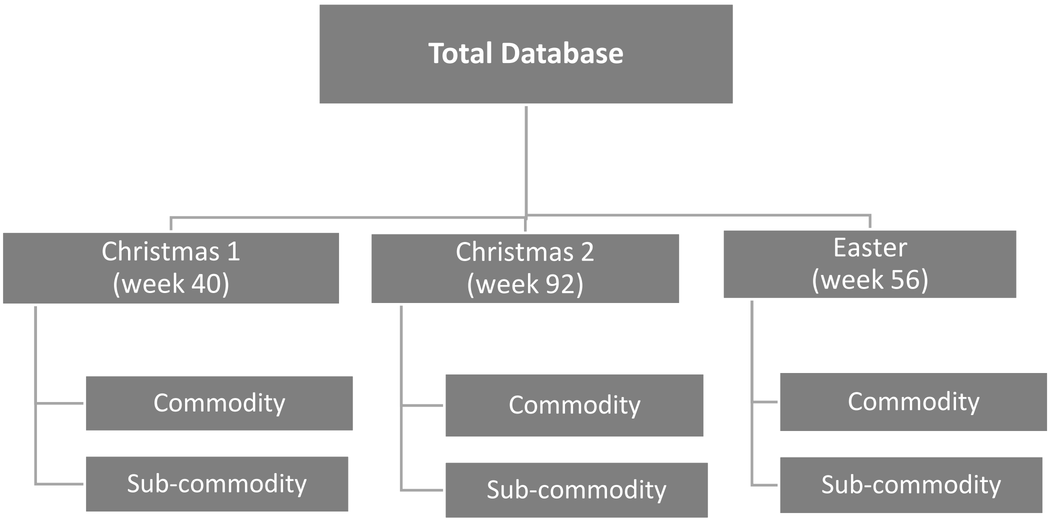


Figure A 9- Scheme of the creation of the different files for the commodity and sub-commodity analysis

Appendix 9 – Association rules at the commodity level

Association rules	support	confidence	coverage	lift
bookstore => christmas seasonal	0.008	0.292	0.028	1.756
toys and games => christmas seasonal	0.011	0.265	0.040	1.594
candles/accessories => christmas seasonal	0.011	0.180	0.059	1.084

Table A 1 - Association rules in the Christmas dataset (week=40) with lift>1, at the commodity level

Association rules	support	confidence	coverage	lift
disposible foilware => spring/summer seasonal	0.010	0.250	0.040	6.206
bookstore => toys and games	0.008	0.200	0.040	3.747
bookstore => christmas seasonal	0.015	0.375	0.040	2.046
toys and games => christmas seasonal	0.013	0.245	0.053	1.338

Table A 2 - Association rules in the Christmas 2 dataset (week=92) with lift>1, at the commodity level

Association rules	support	confidence	coverage	lift
hair care accessories => easter	0.023	0.500	0.045	2.667
toys and games => easter	0.025	0.351	0.071	1.871
charcoal and lighter fluid => easter	0.010	0.200	0.050	1.067

Table A 3 - Association rules, in the Easter dataset with lift>1, at the commodity level

Appendix 10 – Association rules at the sub-commodity level

association rules	support	confidence	coverage	lift
baby food cereals => baby juices	0.001	0.269	0.004	71.201
fluid milk white only,pears => peaches	0.002	0.493	0.004	60.778
bananas,pears => peaches	0.001	0.484	0.003	59.684
pears => peaches	0.003	0.474	0.006	58.472
frosting,macaroni & cheese dnrs => layer cake mix	0.001	0.827	0.001	57.324
layer cake mix,snack cake - multi pack => frosting	0.001	0.699	0.002	57.001
baby food - beginner,fluid milk white only => baby food junior all brands	0.001	0.420	0.003	56.738
peppers all other => peppers red bell	0.002	0.421	0.004	56.337
pepperoni/salami,shredded cheese => pizza sauce	0.001	0.207	0.005	55.451
frosting,lean => layer cake mix	0.001	0.797	0.001	55.245

Table A 4 – Top 10 association rules with the highest lift at sub-commodity level

Association rules	support	confidence	coverage	lift
seasonal preshcool => misc. seasonal items	0.004	0.600	0.007	18.600
wrap} => ribbons and bows	0.004	0.231	0.018	10.284
misc. seasonal items,seasonal preshcool => fashion play}	0.003	0.667	0.004	158.444
extension cords => christmas lights	0.003	1.000	0.003	118.833
facial moisturizers => facial lotions	0.003	0.500	0.006	118.833
fashion play => seasonal preshcool	0.003	0.667	0.004	95.067
super premium wines => ultra premium wines	0.003	0.400	0.007	40.743
jhook - misc => nuts inshell	0.003	0.667	0.004	39.611
adhesives/caulk => wrap	0.003	0.500	0.006	27.423
misc. seasonal items => {fashion play	0.003	0.087	0.032	20.667

Table A 5 - Top 10 association rules with the highest support and lift, in Christmas (week=40), at sub-commodity level

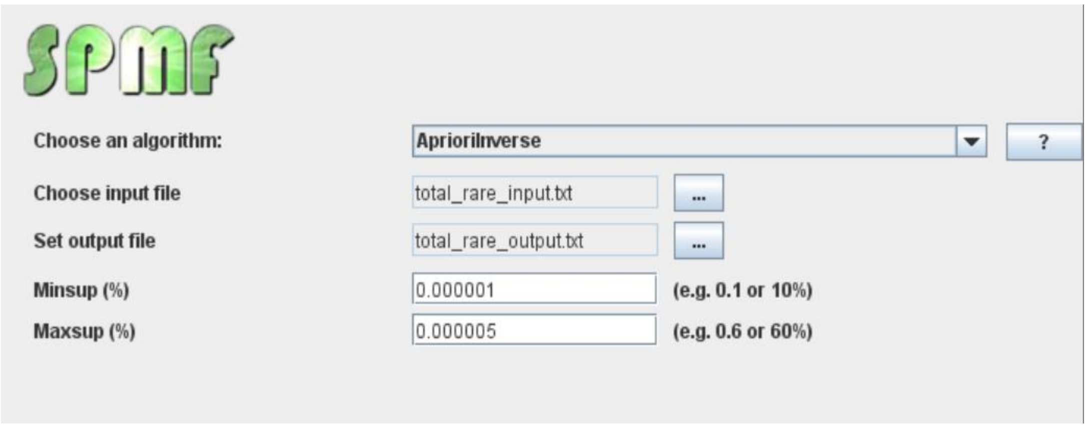
Association rules	support	confidence	coverage	lift
{dog & cat accessories} => {continuity}	0.004	0.308	0.014	31.521
{wrap} => {gift-wrap seasonal}	0.003	0.067	0.049	4.390
{wrap} => {ribbons and bows}	0.003	0.067	0.049	1.983
{baby food} => {baby cereal}	0.002	0.500	0.004	153.667
{flours/grains/sugar} => {frozen pizza}	0.002	0.667	0.003	153.667
{dog & cat accessories,hams-spiral} => {continuity}	0.002	1.000	0.002	102.444
{other preschool} => {xmas plush}	0.002	1.000	0.002	57.625
{baby cereal} => {xmas plush}	0.002	0.667	0.003	38.417
{hams-spiral} => {bread:party breads}	0.002	0.095	0.023	12.544
{continuity} => {hams-spiral}	0.002	0.222	0.010	9.757

Table A 6 - Top 10 association rules with the highest support and lift, in Christmas (week=92), at sub-commodity level

Association rules	support	confidence	coverage	lift
{easter baskets} => {grass/shred}	0.027	0.576	0.046	10.055
{easter fill eggs} => {grass/shred}	0.014	0.286	0.049	4.990
{easter basket stuffers} => {grass/shred}	0.011	0.471	0.024	8.218
{easter egg coloring} => {grass/shred}	0.011	0.129	0.087	2.253
{easter fill eggs} => {easter baskets}	0.010	0.200	0.049	4.339
{easter fill eggs} => {easter egg coloring}	0.010	0.200	0.049	2.310
{easter baskets,easter fill eggs} => {grass/shred}	0.008	0.857	0.010	14.969
{easter plush} => {grass/shred}	0.008	0.400	0.021	6.985
{easter plush,grass/shred} => {easter baskets}	0.007	0.833	0.008	18.081
{easter plush} => {easter baskets}	0.007	0.333	0.021	7.232

Table A 7 - Top 10 association rules with the highest support and lift, in Easter (week=56), at sub-commodity level

Appendix 11 – SPMF interface - Apriori inverse application



Choose an algorithm: AprioriInverse

Choose input file total_rare_input.txt

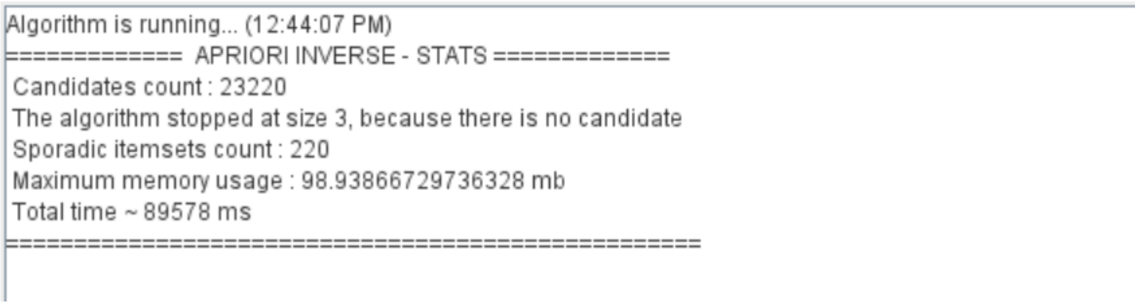
Set output file total_rare_output.txt

Minsup (%) 0.000001 (e.g. 0.1 or 10%)

Maxsup (%) 0.000005 (e.g. 0.6 or 60%)

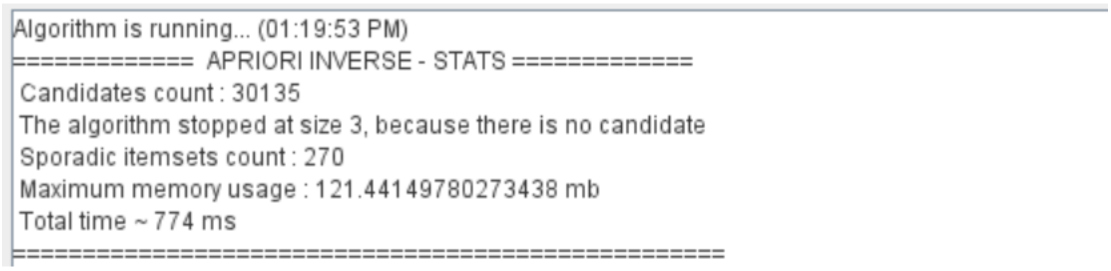
Figure A 10 - Input of the Apriori inverse application in SPMF

Appendix 12 - Software SPMF – Results summary of the Apriori inverse



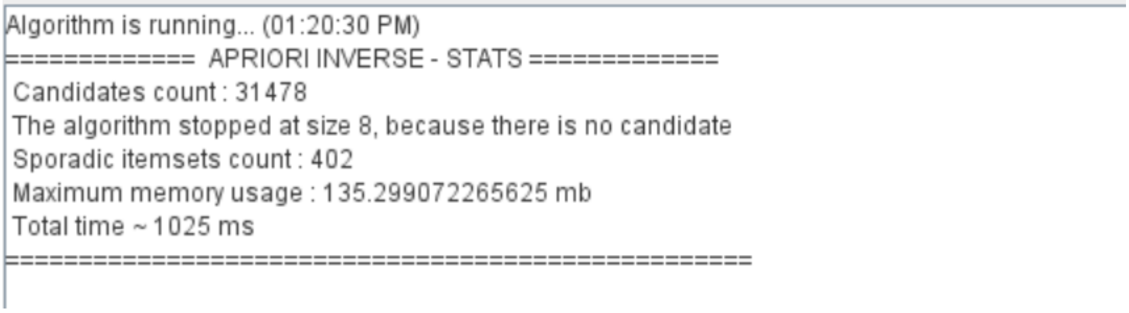
```
Algorithm is running... (12:44:07 PM)
=============  APRIORI INVERSE - STATS =============
 Candidates count : 23220
 The algorithm stopped at size 3, because there is no candidate
 Sporadic itemsets count : 220
 Maximum memory usage : 98.93866729736328 mb
 Total time ~ 89578 ms
===================================================
```

Figure A 11 - Results summary of the Apriori inverse application in all dataset



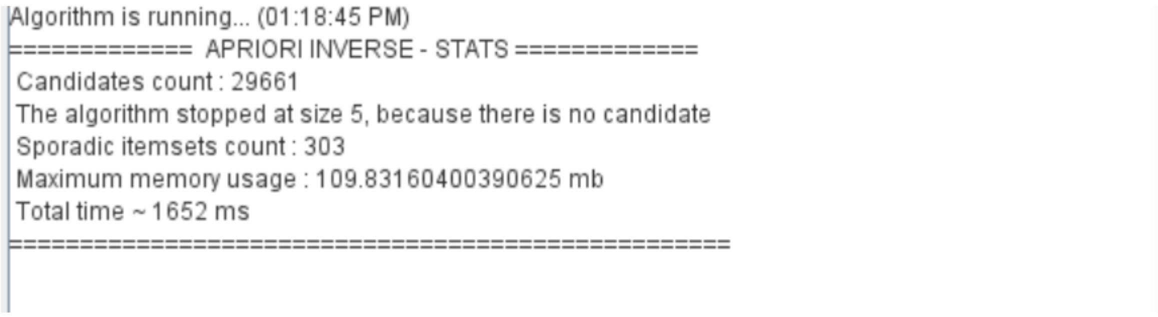
```
Algorithm is running... (01:19:53 PM)
=============  APRIORI INVERSE - STATS =============
 Candidates count : 30135
 The algorithm stopped at size 3, because there is no candidate
 Sporadic itemsets count : 270
 Maximum memory usage : 121.44149780273438 mb
 Total time ~ 774 ms
===================================================
```

Figure A 12 - Results summary of the Apriori inverse application in the Christmas (week=40) dataset



```
Algorithm is running... (01:20:30 PM)
=============  APRIORI INVERSE - STATS =============
 Candidates count : 31478
 The algorithm stopped at size 8, because there is no candidate
 Sporadic itemsets count : 402
 Maximum memory usage : 135.299072265625 mb
 Total time ~ 1025 ms
===================================================
```

Figure A 13 - Results summary of the Apriori inverse application in the Christmas (week=92) dataset



```
Algorithm is running... (01:18:45 PM)
=============  APRIORI INVERSE - STATS =============
 Candidates count : 29661
 The algorithm stopped at size 5, because there is no candidate
 Sporadic itemsets count : 303
 Maximum memory usage : 109.83160400390625 mb
 Total time ~ 1652 ms
===================================================
```

Figure A 14 - Results summary of the Apriori inverse application in the Easter (week=56) dataset

Appendix 13 – Rare itemsets by descending order of sales value

Rare itemsets	Absolute Support	Sales Value ($)
televisions	1	93.67
wireless phones	1	65.99
misc carrier	1	46.66
smoking cessation	1	38.36
holiday arrangements	1	22.61
carpet cleaners	1	19.22
bourbon/tn whiskey	1	18.20
sphe dvds	1	17.90
persnl appl: ft bth/massgrs	1	17.78
area rugs	1	17.20

Table A 8 – Top 10 rare itemsets in the Christmas (week=40) dataset ordered by sales value

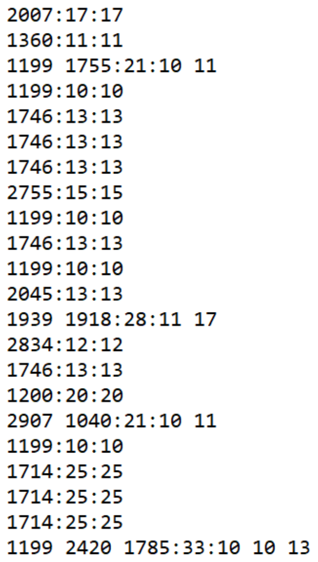
Rare itemsets	Absolute Support	Sales Value ($)
message center/signing	1	77.18
authentic thai foods	1	38.36
misc. pet supplies	1	26.03
adult incontinence undergarmen	1	25.57
honey/syrup	1	24.33
metal polish&rust removers	1	24.32
vegetables/dry beans	1	23.36
foot care - medicated corn/cal	1	22.39
healing garden	1	20.90
bourbon/tn whiskey	1	19.10

Table A 9 - Top 10 rare itemsets in the Christmas (week=92) dataset ordered by sales value

Rare itemsets	Absolute Support	Sales Value ($)
patio sets	1	173.40
outside vendors gift cards	1	42.52
folding furniture	1	26.03
premium 1.5lt wines	1	18.57
fm-mlb	1	17.08
deli tray:meat and cheese	1	16.65
everyday medium	1	16.43
cd-r	1	15.43
whitening systems	1	15.24
liqueurs/specialties (42 under	1	15.18

Table A 10 - Top 10 rare itemsets in the Easter (week=56) dataset ordered by sales value

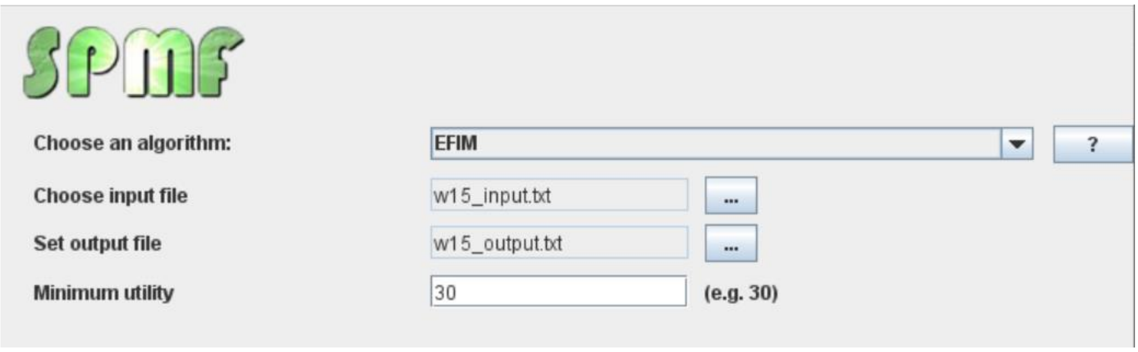
Appendix 14 – Example of the SPMF input file



```
2007:17:17
1360:11:11
1199 1755:21:10 11
1199:10:10
1746:13:13
1746:13:13
1746:13:13
2755:15:15
1199:10:10
1746:13:13
1199:10:10
2045:13:13
1939 1918:28:11 17
2834:12:12
1746:13:13
1200:20:20
2907 1040:21:10 11
1199:10:10
1714:25:25
1714:25:25
1714:25:25
1199 2420 1785:33:10 10 13
```

Figure A 15 - Excerpt of the SPMF input file for the EFIM application

Appendix 15 - SPMF interface - EFIM application



Choose an algorithm: EFIM

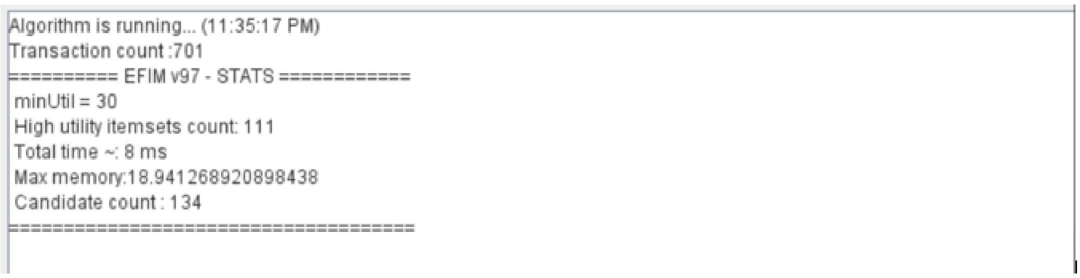
Choose input file: w15_input.txt

Set output file: w15_output.txt

Minimum utility: 30 (e.g. 30)

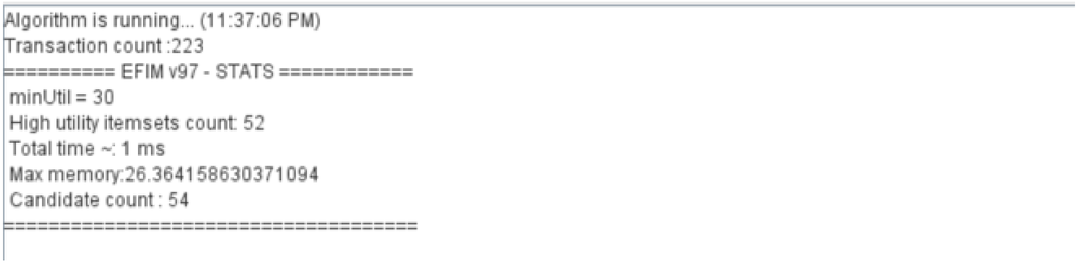
Figure A 16 - Input of the EFIM application in SPMF

Appendix 16 – Summary of outputs



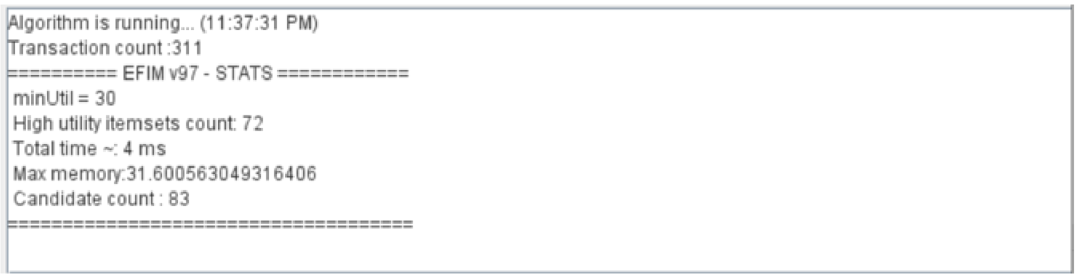
```
Algorithm is running... (11:35:17 PM)
Transaction count :701
========== EFIM v97 - STATS ============
 minUtil = 30
 High utility itemsets count: 111
 Total time ~: 8 ms
 Max memory:18.941268920898438
 Candidate count : 134
======================================
```

Figure A 17 - Results summary of the EFIM application in all dataset



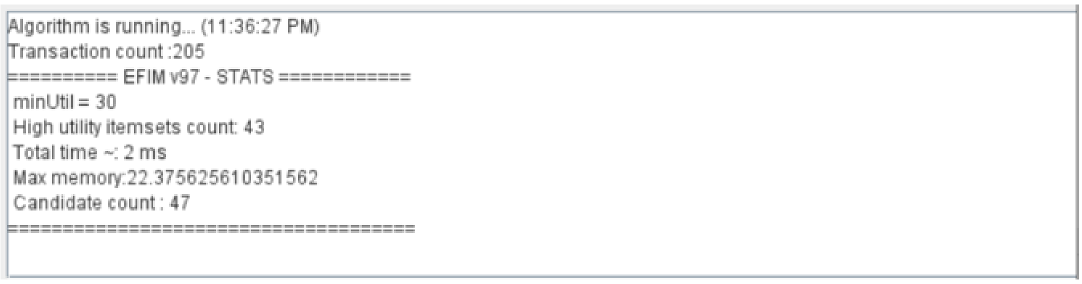
```
Algorithm is running... (11:37:06 PM)
Transaction count :223
========== EFIM v97 - STATS ============
 minUtil = 30
 High utility itemsets count: 52
 Total time ~: 1 ms
 Max memory:26.364158630371094
 Candidate count : 54
======================================
```

Figure A 18 - Results summary of the EFIM application in the Christmas 1(week=40) dataset



```
Algorithm is running... (11:37:31 PM)
Transaction count :311
========== EFIM v97 - STATS ============
 minUtil = 30
 High utility itemsets count: 72
 Total time ~: 4 ms
 Max memory:31.600563049316406
 Candidate count : 83
======================================
```

Figure A 19 - Results summary of the EFIM application in the Christmas 2 (week=92) dataset



```
Algorithm is running... (11:36:27 PM)
Transaction count :205
========== EFIM v97 - STATS ============
 minUtil = 30
 High utility itemsets count: 43
 Total time ~: 2 ms
 Max memory:22.375625610351562
 Candidate count : 47
======================================
```

Figure A 20 - Results summary of the EFIM application in the Easter (week=56) dataset

Appendix 17 – Top high utility itemsets by utility

Itemsets	**Utility ($)**
hams-whole boneless	330
electronic gift cards activati	320
hams-spiral	264
hams-whole bone-in	242
holdiay dinner (cold)	230
whole - hens (15 lbs & under f	176
outside vendors gift cards	172
mastercard gift card	114
fruit baskets	102
whole - tom (16 lbs & over frs	96

Table A 11 - Top 10 high-utility itemsets in the Christmas 1 (week=40) dataset with the highest utility

Itemsets	**Utility ($)**
outside vendors gift cards	1118
hams-spiral	462
hams-whole bone-in	352
mastercard gift card	342
electronic gift cards activati	300
hams-whole boneless	270
turkey breast bone in	231
fruit baskets	204
whole hen (over 15lbs)	200
electronic gift cards activati, outside vendors gift cards	189

Table A 12 - Top 10 high-utility itemsets in the Christmas 2 (week=92) dataset with the highest utility

Itemsets	**Utility ($)**
hams-spiral	418
hams-whole boneless	270
pansies	221
hams-whole bone-in	198
patio sets	173
angus beef	162
electronic gift cards activati	120
premium flowering plants	117
seafood-frz-raw shlfsh-other	91
seafood-frz-rw-all	60

Table A 13 - Top 10 high-utility itemsets in the Easter (week=56) dataset with the highest utility