The beauty or the beast inside retail stores? A Market Basket Analysis of a cosmetic company

Rita Sebastiana Vaz Gonçalves

Master in Computer Science and Business Management

Supervisor:

Doctor Sérgio Miguel Carneiro Moro, Assistant Professor (with Aggregation),

Iscte- Instituto Universitário de Lisboa

Co-Supervisor:

Doctor Pedro de Paula Nogueira Ramos, Associate Professor (with Aggregation),

Iscte – Instituto Universitário de Lisboa

November,2020

The beauty or the beast inside retail stores? A Market Basket Analysis of a cosmetic company

Rita Sebastiana Vaz Gonçalves

Master in Computer Science and Business Management

Supervisor:

Doctor Sérgio Miguel Carneiro Moro, Assistant Professor (with Aggregation),

Iscte-Instituto Universitário de Lisboa

Co-Supervisor:

Doctor Pedro de Paula Nogueira Ramos, Associate Professor (with Aggregation),

Iscte-Instituto Universitário de Lisboa

November,2020

Acknowledgements

First and foremost, I have to thank my supervisors, Dr Sérgio Moro and Dr Pedro Ramos, without their support and assistance this thesis would not have been accomplished.

I would also like to show gratitude to the cosmetics company that provided the data. Without their contribution, this analysis would not have been possible.

Abstract

Nowadays, companies have in storage a large quantity of customer data and do not take advantage of this. Given the state of the market, in order to stay competitive, retailers should know and adapt to the needs of their customers. One way to accomplish such task is to perform a Market Basket Analysis to the available customer data. A Market Basket Analysis examines transactional data to find which items are related to each other enabling retailers to create new marketing strategies such as store layout, product placement, pricing and promotions. In this study, a market basket analysis was performed to transactional data provided by a Portuguese cosmetics company with the purpose of finding which items are put in the same "basket" by their current customers. This analysis resulted in the generation of several association rules about customers with similar demographic variables such as, gender, age and geographical location. For example, this analysis allowed to discover the items that customers belonging to a certain gender and age group are likely to buy. For example, through this analysis it was possible to discover that women aged over 60 years old that buy the hair removal service {87880} will also buy the hair removal service {114325} with a confidence of 94%. It was also discovered that women aged between 35 and 44 years old that visit one particular store in the greater Lisbon area buy the makeup foundation 85945 and the cream 99282 will buy the cream 108741 with a likelihood of 89%. One other example of the association discovered is that customers that visit another store in the greater Lisbon area have a very high inclination to buy together the men's' perfume {73933}, the aftershave {84539} and the deodorant {105213}.

Keywords: Market Basket Analysis, Data Mining, Cosmetics, Marketing, Association Rules, Customer Clustering;

Resumo

Hoje em dia, as empresas têm uma grande quantidade de dados de clientes armazenada que não é aproveitada e, tendo em conta a situação atual do mercado, para se manterem competitivas, as empresas têm que conhecer e adaptar-se às necessidades dos seus clientes. Uma maneira de fazer isso é realizar uma *Market Basket Analysis* usando os dados do cliente. A *Market Basket Analysis* examina os dados transacionais para descobrir que itens estão relacionados, o que permite que as empresas criem novas estratégias de marketing, como o layout de loja, a alocação de produtos, preços e promoções. Neste estudo, será realizada uma *Market Basket Analysis* recorrendo a dados transacionais fornecidos por uma empresa portuguesa de cosméticos com o objetivo de descobrir que itens são colocados no mesmo carrinho pelos seus clientes atuais. Esta análise resultou na geração de diversas regras de associação sobre clientes com variáveis demográficas semelhantes, como o género, a idade e a localização geográfica. Por exemplo, ao realizar esta análise foi possível descobrir que mulheres com mais de 60 anos que compram o serviço de depilação {87880} vão adquirir também o serviço de depilação {114325} com uma confiança de 94%. Foi também descoberto que as mulheres com idades compreendidas entre os 35 e os 44 anos que visitam uma determinada loja da grande Lisboa que compram a base {85945} e o creme {99282} comprarão também o creme {108741}, com uma probabilidade de 89%. Outro exemplo do que foi descoberto é que os clientes que visitam uma determinada loja na zona da grande Lisboa têm uma tendência muito elevada para comprarem juntos o perfume masculino {73933}, o *after-shave* {84539} e o desodorizante {105213}.

Palavras-chave: Análise de Cesto de Compras, *Data Mining*, Cosmética, *Marketing* , Regras de Associação, *clustering* de consumidores;

Index

Table Index

Figure Index

List of Abbreviations

MBA	Market Basket Analysis
DM	Data Mining
RFM	Recency, Frequency, Monetary
CRM	Customer Relationship Management

Chapter 1: Introduction

Currently, a retailer must know the needs of customers to adapt to them in order to stay competitive [1]. Retailers collect and store voluminous and several types of data about their customers daily, ranging from customer demographics, to data that indicate how customers move into the physical or web stores, what products they put in their baskets, etc [2]. One of the biggest challenges that companies have nowadays is how to extract relevant information from those vast databases to gain a competitive advantage. The consumer data needs to be analysed in order to reach the competitive advantage, which can be done using various Data Mining (DM) techniques. DM discovers interesting patterns in databases some of them being, association rules, correlations, clusters and many others of which association rule mining is one of the most popular. This technique discovers valuable associations and correlations relationships contained in a dataset [1]. Data mining techniques can also be applied in the marketing field [3]. An example of the application of association rule mining in the marketing field is called Market Basket Analysis (MBA) [4]. While association rule mining extracts associations from any large amount of data [4], MBA focuses on extracting associations about purchasing patterns and customer behaviour, specifically from stores' transactional data [1] .

One way to gain a competitive advantage is to perform a MBA [5]. Since this technique originated in the marketing field and its purpose is to understand which items are put in the same "basket", it adopted the name "Market Basket Analysis" [6]. Although being originated in the marketing field, the application of MBA has recently, branched out to other fields, such as, nuclear science [7], pharmacoepidemiology [8], immunology [9], and geophysics [10]. The application of this technique is based on the possibility that sales of different products are correlated. The associations discovered are then used to make new marketing strategies [11]. Such marketing strategies include, store layout, product placement, pricing and promotions. The application of MBA will ultimately, upgrade the business in terms of sales and will also, improve the relationship with the customer [4, 13].

MBA, also known as association rule mining, is a good way to provide scientific decision support for retail market by mining relationships among items people purchased together [13]. Aguinis et al. [6] defined MBA as a DM technique that focuses on finding associations between items, products or categories. MBA analyses the composition of the basket of items purchased by a customer on a single visit to the store and assesses the extent to which the items co-occur. This assessment makes it possible to find customer buying patterns through generating association rules between the products placed on the shopping baskets [4, 9].

For a better understanding of this technique, figure 1 [4] shows an example of a typical MBA question, "*Which items are frequently bought together by the customers?*" [4].

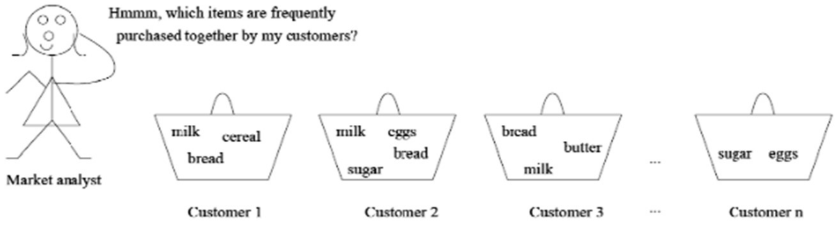

FIGURE 1- MARKET BASKET ANALYSIS [4]

This system helps analysts find out the sets of items that customers are likely to purchase together and can be carried out using all available data from customer transactions. As said previously, the discovered associations will guide the business when planning the marketing or advertising strategies.

Lee, Liu and Mu [14] applied an MBA to transactional data provided by a Korean shopping mall that mainly sells food, goods, cosmetics and others. The transactional data consisted of 51 080 transactions from June 2015 until June 2016 and contained customer data such as, age, gender, costumer ID and others. The first step of this study was to segment the customer data into VIP and non-VIP trough the method of RFM. After this is done, the VIP data is then put through an MBA to identify the most effective rules and patterns. The authors show the top 10 association rules that showed confidence above 90% and in four of them, cosmetics were present. To better understand the association rules obtained, the authors analysed the main purchase categories of VIP clients and found that 16% of VIP transactions accounted for cosmetics, making this group of items the focus of VIP clients. The purpose of this study is to take advantage of the discovered information to create new CRM strategies.

Hidayat et al. [15] performed an MBA using transactional data from Breillant, an Indonesian online cosmetic store on the Shopee platform, which is an e-commerce online shopping platform. The transactional data consisted of 34 sales transactions carried out in the month of November of 2018. The transactions contained 47 different types of beauty products, making the total number of products within the transactions was 126. The authors don't specify the total number of rules generated by the Apriori algorithm, however they shared that a combination of products with a strong support and confidence is, customers who bought Original Liquid Bleaching Seeds and Harva Peeling Gel will also buy Castor oil with 30% confidence and 8,8% support. The aim of this study was to increase revenue for the shop owners by improving sales strategies and product promotion based on the associations discovered.

Abdullah et al. [16] resorted to an MBA with the purpose of assisting PureGlow, an online cosmetic retailer, in making decisions regarding the choice of stock quantity. PureGlow provided transactional data, specifically 69 transactions, to which the authors applied the Apriori algorithm. This resulted in the generation of a number of association rules and frequent itemsets that the authors do not specify. However, they provide the association of products that has the highest confidence, which is, customers that buy DeepClensingMilk and DeepClensingToner will also buy Whitening Soap with a confidence of 93%.

The goal of this thesis is to understand which products are put in the same "basket" by current consumers, meaning, not only discover the associations between the products, but also, find out what products are drivers for buying other products within the same basket. To address this objective, a DM approach was adopted, specifically, a MBA. There are several papers regarding the application of an MBA in retail stores like, supermarkets [1], [17], grocery stores [18], and even retail of sporting goods [19, 20]. However, as can be seen by the articles above, there are very few papers regarding the application of an MBA in the retail of cosmetic, given that only two articles were found. On top of that, while searching for related work, there were not found studies based on data from physical stores, which presents an advantage to this study because the data provided to this analysis is from a Portuguese cosmetic brands' various physical stores throughout Portugal.

The studies above [16, 17] aimed only at association rule generation. In comparison, this study, also leverages the company's knowledge by providing associations rules between the products. However, these associations are based on data split by data variables, specifically, customers' age and gender. Similar research adopting this type of analysis in the cosmetic store area has received little attention by scholars, given that it was not specified in any of the articles analysed that the examined data had any kind of condition, for example, being split by gender or age. These new patterns can for example, help specialists to propose a new store layout and place items that are regularly purchased together close to each other and thus, promote the sales of the items. It also allows to create more personalised content for customers, which ultimately, improves customer satisfaction.

Chapter 2: Literature Review

Marketing is everything that is done to place a particular product or service in the hands of customers. It includes sales, public relations, pricing, packaging, distribution, among others, and it has evolved immensely over time [22, 23]. This evolution is mainly due to recent digital transformation, namely, the emergence of the Internet and e-commerce [22].

In the traditional form of marketing, data analysis is conducted using small databases with limited analytics platforms and implementation capacity [22]. It uses strategies like direct sales, TV, radio, mail, print ads in newspapers or/and magazines and printed materials to reach as many consumers as possible [21]. The traditional marketing communication model for media [23, 24] holds that communication is a one-to-many process whereby a company transmits content through a medium to a large group of consumers that are considered to be homogeneous in their tastes with respect to the information being transmitted. This communication is static, which means there is no interaction between the consumers and the companies [25].

Traditional marketing is still used in some occasions, for example, to attract older consumers who have limited access to digital devices [26]. However, studies show that this type of marketing has become obsolete due to the new ways to shop and pay, intense competition between companies, additional sales channels and the declining effectiveness of advertising [27]. Additionally, consumers have become tired of advertising that is unconnected to their interests [27].

These days, consumers want to have a one-to-one relationship with companies and, on the other hand, total customer relationship is the reigning paradigm for companies [28]. The retail companies are now realizing that it is possible to gain a competitive advantage by utilising DM. Given the information that the retailer has been collecting through the years, they can now take advantage of DM techniques to find useful information [29]. If done thoroughly and properly, the process will increase profits by the exclusion of the non-profitable customers and continuous rewarding of those profitable customers over an extended period of time [21, 30, 31].

DM techniques are able to predict future trends and behaviours by analysing the company’s databases. It discovers hidden patterns and new information that experts may have missed. Based on this new information, companies can make knowledge-driven and proactive decisions [32].

From the last decade, DM has received more attention due to its significance in decision making, and it has become an essential component in various industries. The field of DM has been prospered and posed into new areas such as manufacturing, insurance, medicine, retail [32].

One of the most widely used areas of DM in the retail area is in marketing and, there has been an increasing number of retail companies using DM for marketing purposes and benefiting from it [30, 34].

The next line exhibit some examples of Data Mining implementation in the retail area [6, 33]:

- MBA: this analysis reveals items that consumers tend to buy together. This knowledge allows companies to improve in-store layout strategies, promotions and stocking strategies.
- Sales Forecasting: analysing time-based patterns helps retailers make stocking decisions.
- Merchandise planning and allocation: When retailers add new stores, they can improve merchandise planning and allocation by examining patterns in stores with similar demographic characteristics. Retailers can also use DM to determine the ideal layout for a specific store.
- Acquiring and Retaining Customers: By using Data mining techniques, the retailer can retain existing customers by providing discounts or offer, attract customers and acquire customers.
- Customer Segmentation and Target Marketing: Data mining can be used in grouping or clustering customers based on their behaviour [34]. This type of information is useful to define similar customers in a cluster, holding on the right customers and identify likely responders for target marketing.

2.1 Application of Market Basket Analysis in the retail area

Given the changes mentioned in the first section, a retailer must be aware of the costumer's needs and adapt to their needs [29]. However, one of the challenges that companies face is how to extract relevant information from their vast databases to gain a competitive advantage [35].

One way to be able to adapt to the consumers' needs and gain a competitive advantage is by applying DM techniques. One of these techniques is called MBA. Studies show that leading retailers apply an MBA to better understand their customers' habits and adjust operations to obtain maximum success. Additionally, the MBA is also allowing retailers to rapidly adjust and is enabling the organisations to work smarter [36].

The main idea behind MBA is that consumers rarely make purchase decisions that are isolated. For example, when shopping in a supermarket, customers rarely buy only one product; they are far more likely to purchase an entire basket of products and typically from different product categories [37]. One way that retailers influence their customers to buy more and hence, increasing sales, is by carefully analyse their purchases [38].

One possible definition of MBA is "the study of retail stock movement data recorded at a point of sale to support decisions on shelf-space allocation, store layout, and product location and promotion effectiveness" [39].

The input of an MBA is a transactional database, meaning, a dataset of customer transactions [29]. A market basket is composed of the items bought by a customer in a single store visit. In this analysis, the quantity and the price of the items are ignored and, the most important attributes are the transaction identification and attribute identification. Each transaction represents a market basket, which occurred in a specific time and place. This purchase can be associated with an identified customer or a non-identified customer [40].

The purpose of this analysis is to extract associations or co-occurrences between products and determine which are most frequently bought together and which of the products are drivers for purchasing certain products, to afterwards plan strategically for marketing decisions. Such decisions are related to product placement, pricing, product promotion [2, 42].

Association rule mining is based on association rules, as it is possible to understand by the name, and it was first introduced by Agrawal et al. [42]. In DM, an association rule is an expression X -> Y, where X and Y are sets of items. X is termed the left-hand-side and is the conditional part of an association rule. On the other hand, Y is called the right-hand-side and is the consequent part [43]. The general idea of an association rule is "A1 ..... An → B", meaning customers who buy product A also have a considerable opportunity to buy product B [15].

The mathematical model for association rules is as follows: Let I = {i1, i2,…, in} be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that T ⊆ I. Each transaction is associated with a unique identifier TID. A transaction T is said to contain X, a set of items in I if X ⊆ T. An association rule is an implication of the form “X -> Y”, where X ⊆ I; Y ⊆ I, and X ∩ Y = ∅. The rule X -> Y has support in the transaction set D if s% of the transactions in D contain X U Y. It is said that the rule X -> Y holds in the transaction set D with confidence c if c% of transactions in D that contain X also contain Y [44]. Support measures how many times the transactional record in the database contains both X and Y and confidence measures the accuracy of rule.

The standard formulas for support and confidence, respectively, are as follows [45]:

$$\text{Support of (X -> Y)} = \frac{Number\ of\ transactions\ containing\ (X\ U\ Y)}{Total\ number\ of\ transactions} \quad (1)$$

$$\text{Confidence of (X -> Y)} = \frac{Number\ of\ transactions\ containing\ (X\ U\ Y)}{Number\ of\ transactions\ containing\ X} \quad (2)$$

One very important aspect of association rules is that they are only valuable if they satisfy both a minimum support and a minimum confidence. This values can be set by users or domain consultants [4]. The set of items that respect the minimum support value are called Frequent Itemsets. The support value indicated the proportion of transactions which contain said itemset. Frequent Itemsets can be generated by using DM algorithms like Apriori, FP-Growth Algorithm and K-Apriori [1]. On the other hand, confidence is the measure of assurance associated with the discovered patterns. The association rules are derived from the itemsets depend on the minimum confidence value imposed [1].

A good example to easily understand the concepts of support and confidence is described in the work by Aguinis, Forcum and Joo [6]. The authors found that in 10 000 transactions, 30 consisted of soda and orange juice together, with soda being purchased in 33 and orange juice in 497 transactions. The support of the association rule between soda and orange juice is 30/10 000= 0.3. This means that 30% of transactions included both orange juice and soda. The confidence is calculated by $\frac{30/10000}{33/10000}$= 0.91. This means that 91% of transactions that included soda, also included orange juice. These results may serve as an indicator that the purchase of soda causes an inclination to buy orange juice [6].

One other variable to have in consideration is lift. This variable supplies a decision-making tool in order to determine if an association can be explained by chance and it also helps to screen out rules that are not relevant. For example, considering an association between two items, A and B. If the lift is greater than 1, this means that the relationship is positive in nature, meaning the items are complementary. The presence of A is associated with the presence of B. On the other hand, if the lift is lower than 1, this means that the association has a negative nature and the items are substitutes. In this case, the presence of item A is associated with the absence of item B. The last case is when the lift value is close to 1. This means that the purchase was made by chance and there is no relevant association between the items. These rules should not be taken into consideration [6]. The formula for lift is as follows:

$$\text{Lift} = \frac{P(A \cap B)}{P(A) * P(B)} \qquad (3)$$

In 1994, Agrawal & Srikant developed the *Apriori* algorithm, which is the most well-known and common used algorithm today to determine association rules [47, 48]. This algorithm performs two steps. First, it determines the frequent itemsets. These are sets of items that have at least the minimum support. The second step consists of generating the association rules from the itemsets created in the first step [48]. Despite being the most used, this algorithm has one major flaw in that it is time consuming due to its breadth-first computational approach [38]. Given such drawback, researchers have developed other algorithms as an attempt to solve this problem.

One of these algorithms is called ECLAT, and it was developed by Han, Pei, Yin and Mao [49]. Unlike the Apriori algorithm, this algorithm performs a depth-first scan of the database to identify all frequent 1-item sets and then uses this result and the Apriori principle to generate larger frequent itemsets [38].

Another of these algorithms is the *FP-Growth* algorithm developed by Zaki [50]. This algorithm adopts a recursive elimination scheme. In simple words, it deletes all items from the transactions that are not frequent individually, meaning items that do not appear in a user-specified minimum number of transactions [51]. This algorithm which uses a tree structure may be seen as a hybrid approach, with a breadth-first scan to establish nodes (e.g. frequent 1-item sets) followed by a depth-first scan to find all subsequent frequent itemsets [38].

When comparing the three algorithms, it is possible to conclude that the ECLAT and the FP-growth algorithm have increased speed at generating rules than the Apriori algorithm. However, they can be more memory intensive on larger databases, this is particularly important to the retail area, given the significant size of a retailers' database [50, 53]. Additionally, a study performed by Yildiz and Ergenç [53] shows that the Apriori Algorithm outperforms the FP-Growth algorithm. Even though the FP-Growth was sometimes faster at generating rules than the Apriori algorithm, it sometimes failed to generate some rules with a high confidence value. These factors make the Apriori algorithm the most reliable choice to determine association rules [46].

MBA can only provide a profile of customers' purchasing affinities. It provides the combinations that are in customer's purchasing carts but cannot give an explanation as to why they are bought together. Most times, the association is apparent, like detergent and fabric softener. However, sometimes the reason why some products are bought together is not so evident and easy to explain, like the famous association between beer and diapers or bottled juice and cold remedies. *Apriori* algorithm provides a possible answer to these questions by investigating spatial relationships between displayed products and their impact on sales that result from the visual effects of adjacency on impulse buying and cross-selling. It means the *Apriori* algorithm performs best as compare to other traditional techniques [54].

2.2 Advantages and Disadvantages of an MBA

The main advantage of an MBA is its flexibility for analysing many kinds of data patterns using mild assumptions which can usually be met by most datasets [37]. Besides that, the MBA offers several other significant advantages [13, 33, 42]:

- Performing an MBA is useful for shelf design, deciding the location by means of combination so that customers can quickly locate the items.
- Based on the MBA results, it is possible to design more effective pricing and promotion strategies. Retailers can use price promotion strategies to positively impact the sales of products bought together;
- MBA discovers hidden consumer preferences patterns. Through analysing these patterns, it is possible to discover the consumer's behaviour and get a better market segmentation.
- Through this analysis it is possible to find out which products should be cross-sold, and which have no significant related items;

However, the MBA also shows some limitations, such as [41]:

- MBA is not suitable to handle association rules for products purchased together within a period of time, but not necessarily at the same time. The data analysed does not consider temporal information.

- When analysing large databases, MBA generates a vast number of rules. Analysing all the rules is complicated and time-consuming.

The MBA is widely used in the retail area because it allows retailers to obtain essential and useful information easily. This new knowledge then allows them to make the right decisions that ultimately lead to a profit increase [12].

2.3 Mining Customer Behaviour

Initially, the purpose of association rules was to identify associations between products and its most common application is an MBA . However, in addition to discovering product associations, it is possible to use association rules to mine changes in customer behaviour to acquire new information about their needs [43]. Businesses can benefit from analysing customer data, meaning, purchase patterns of individual customers and groups, to determine their preferences and thus improve support for marketing decisions, like one-to-one marketing strategies [57, 59]. From the marketers' new perspective, different customers have different needs, even if they purchase identical products or services. Therefore, market segmentation is necessary [57].

Several studies [44, 59, 60] have integrated RFM variables, demographic variables and association rule mining to analyse and predict customer behaviours.

The authors begin by segmenting the customers in clusters with similar RFM values and assign each customer to an appropriate segment. The concept of RFM was first established by Bult and Wansbeek [59]. RFM relies on Recency, Frequency, and Monetary measures which are three important purchase-related variables that influence customers' future buying possibilities [58]. Recency refers to the period since the last purchase. Marketers have stated that most-recent purchasers are more likely to purchase again than less-recent purchasers, meaning a lower value in this field corresponds to a higher probability of the customer making a repeat purchase [59, 61]. Frequency indicates the number of transactions that a customer has made within a certain period of time. The higher the value in this field, the higher the customers' loyalty [58, 60]. Lastly, Monetary relates to the amount of money spent by a customer within a period of time. A higher value in this field suggests that the company should invest more on the customer [58]. Clustering customers into different groups based on RFM is an effective way of not only improving the quality of recommendations of products to customers, but it also helps in the decision-making to develop more effective marketing strategies [57].

The assumption behind the process of clustering customers based on their RFM values is that customers with similar purchasing behaviours, are likely to have similar RFM values [57].

One example of the application of an association rule algorithm to RFM based clusters is the study of Liu and Shih [57]. In their study, after segmenting the customers based on their RFM values, they applied an association rule algorithm in order to extract frequent purchase patterns from each cluster, rather than from all customer transactions. In their approach, products are recommended to customers based on frequent purchase patterns of customers with similar purchases. They concluded that by knowing each clusters' preferences, the decision-makers can now target these customers groups and develop marketing strategies to satisfy their needs and consequently, increase the market share of the company.

However, analysing recency, frequency and monetary factors is not the only of finding clusters if customers with similar purchasing patterns, as consumer buying behaviour is influenced by one's individual physical and social surroundings [60]. There are many different processes within the customer behaviour and various factors that can affect consumer purchasing behaviour. These factors consist of demographic factors, such as gender, age, income and education, geographic factors and psychological factors [61], [62].

The gender of the customer is a demographic factor that affects the consumer behaviour and it is divided in two categories: male or female. Women tend to purchase most of the household goods and men still make the most purchasing decisions related to consumer durables such as, refrigerators, cars and TVs [62]. Researchers say that males and females want different products and are very likely to have different preferences [63].

One other demographic behaviour is age. The age of the consumer is one of the most important demographic factors that influence purchase habits [61]. Consumer's tastes and preferences change with time, which then, translates on their purchasing habits that changes as the time passes, for example, young adults and seniors have different wants and needs [3, 6].

The personal income of an individual is a determinant demographic factor that has a considerable impact on the buying behaviour of a consumer [64]. It is known that people with low income tend to spend the majority of their income in food, rent, clothing among other essential items. As the income increases, consumers are more likely to buy more non-essential products, like durable goods and luxuries that improve the level of comfort and lifestyle of an individual [62].

Finally, the occupation of a consumer is another demographic factor that affects the purchasing behaviour. For the most part, the more educated a person is, the more demanding they are as a buyer [62]. One example of this is that a marketing manager is more likely to buy business suits and the low level worker of the same company is more likely to buy durable work clothes [64].

For some goods and services, geographical variations can represent a significant factor regarding the consumers' purchasing habits. The geographic area represents a cultural factor, that implies that the people that live in the same area share set of beliefs that influences their shopping decisions [61]. By segmenting the customers using this variable the marketers are able to design products and strategies in line with the needs of the geographic group [64]. For example, in a hot geographic area, the request for refrigerators will be higher than in an cooler area [62].

Lastly, the purchasing habits of consumers are also influenced by phycological factors and the most significant ones are motivation and perception [61]. Every individual has different needs and some of these needs are more pressing than others. So, a need becomes a motive the more pressing there is to satisfy it. Perception is the capacity of selecting, organizing and interpreting information and create a relevant experience and it can be categorized into 3 categories: selective attention and the retailer must try to attract the attention of the customer; selective distortion and selective retention; in these cases, the retailers in order to capture the customers' attention has to adapt to the customers beliefs [64].

In [65] the data of 300 clients of an insurance company is split using demographic variables such as, age, gender, occupation, education level, marital status, place of residence and clients' income. The next step was to perform an MBA to discover hidden association rules within the insurance industry. The authors believe that the new association rules can be used to targeting new and appropriate customers for the insurance company.

In their study, Chen et al. [43] applied the Apriori algorithm to the dataset of purchases made by the cluster of their most valuable customers in order to identify the association between a customer profile and product items purchased. In this perspective, the left-hand-side of the rule chooses customer profile variables that include gender, age, yearly income, etc. and the right-hand side consists of the products bought [43]. Based on this knowledge, they were able to establish marketing strategies to adapt to the preferences of their most valuable customers, thus increasing customer value. For example, they found out that male customers purchase vegetables. This pattern can lead marketing decision-makers to enforce marketing efforts in promoting vegetables to male customers.

2.4 An alternative to Association Rule mining

The majority of applications of MBA involve the method association rule mining. However, recently, some researchers have argued that in some cases, this is not the most suitable method. In this study is it said that when given a large volume of sales transactions with a high number of products, the data matrix to be used for association rule mining usually ends up large and sparse. This means more time is needed to process data.

In their study, Tan and Lau applied cluster analysis instead of association rule mining to perform an MBA. While association analysis aims at identifying groups of products, cluster analysis focuses on identifying groups of similar records (i.e., sales transactions). Time-series clustering of sales transactions requires data to be summarised as time series, which can result in a substantially smaller data set that requires less time to process and it can be used to discover which products are commonly purchased across a period of time which represents the advantage over the association rule method [66].

They conclude by saying that market basket data can be more easily analysed using time series clustering instead of association analysis. However, time-series clustering is a poorly tested method compared to the association rules method. [66]. When searching for both approaches, the time series clustering only has 85 results. On the other hand, when searching for association rules, about 9500 results are shown.

2.5 The Importance of Physical Stores

Although a rapid grow of cosmetics sales over the internet can be observed, the traffic to beauty websites has been decreasing year after year [68, 69]. As technologies develop, the cosmetics industry is trying to take advantage of the new opportunities and interact with consumers on a more "practical" and personal level. This happens, for example, through Social Media [68]. Social Media can be characterised as “a group of Internet-based applications that build on the ideological and technological foundations of Web

2.0, and that allow the creation and exchange of user-generated content" [69]. Some examples of these applications are Facebook, Instagram and YouTube.

Nowadays, beauty/cosmetics brands increasingly use Social Media to give consumers a more actively engaging brand experience [70]. Kane et al. [71] say that social media presents users and companies with capabilities that were not available in traditional offline marketing.

The rapid growth in the popularity of online shopping over the last decade has forced retailers to rethink their physical stores [72]. However, in spite of its rising popularity, physical retailing continues to be the leading choice in many product categories, such as high quality clothing and groceries [73].

Retailers now understand that it is necessary to combine modern technology with traditional and basic values, such as staff and store layout [74]. Online shopping is becoming more popular among consumers because it is more convenient than physical stores, for example, it saves time by preventing a trip to the store and so, retailers are now devoting time and money to create a more pleasant and convenient customer experience when buying online. Online shopping not only saves time, but it also provides almost boundless inventory [72]. However, studies show that consumers are willing to buy some products online but not others [74]. One major limitation of online shopping is related to the fact that through online shopping is not possible to see and touch the products before purchasing it and the judgement is made by looking at photos and videos, which can lead to let-downs when the product arrives and does not match the expectations [4, 6]. This disappointment later leads to higher rates of product returns which represents a major loss to the retailer [72].One other factor that gives physical stores advantage is the one-to-one interaction between the customer and the employee of the physical store, as studies show that this is factor for preferring purchasing in physical store instead of online [74, 75]. Issues related to payment by credit card and security issues are also related to the preference of physical stores [75]. Studies show that the main reason that customers prefer physical stores is that they have access to the purchased product much faster given that when bought online the products takes days to arrive [74].

Nevertheless, against the common assumption that online shopping damages sales from physical stores, studies show that it does not [76]. In fact, retailers that have both online and physical stores must take advantage of both services, for example, the option of ordering online and pick up the product from the store [73]. Both physical and online channels show challenges on many aspects and it is necessary to coordinate both channels' strategies [72]. In order to succeed, retailers must understand the complexities of these channels and take advantage of their main capabilities [73].

2.6 Related Work

There are several papers regarding the application of the MBA in the retail area (supermarkets, convenience stores, grocery stores), as shown in the articles in table 1. This is where this technique is most used, primarily, to enhance the store layout to make sure the customers find the associated products easily and in the promotion of products. However, this kind of analysis is poorly researched in the field of cosmetics. As shown in the examples above, MBA analysis has been used in retail stores containing a cosmetic area, like supermarkets, but it has been done very few times in specifically cosmetics stores, only two papers were found. Both of these papers aimed at discovering product associations. It is also important to notice the lack of research regarding the application of an MBA using data from a physical cosmetic store. In fact, when searching for articles that show the application of an MBA in a physical cosmetic store, there were not found any studies regarding this kind of application. This absence of studies in this field creates a gap in the research that must be filled. This analysis will contribute to filling this research gap because it will not only provide association rules between the products, but it will also be able to explain which consumer groups define said patterns while using data provided from a physical cosmetic store.

Reference	Market Area	Main goal	Data Used	Method	Results
[16]	Online retail of cosmetics	Assist the company, PureGlow, in making decisions regarding the choice of stock quantity	Dataset of transactions provided by PureGlow; 69 transactions	Trough applying the Apriori algorithm to the transactions provided, they were able to discover associations between the products	The total number of associations is not shown, and it is only shared the rule with highest confidence: if a customer buys DeepClensingMilk and DeepClensingToner then they will also buy Whitening Soap with a confidence of 93%.
[15]	Online retail of cosmetics	Improve sales strategies and promotion of products	Dataset of monthly sales of November 2018 provided by Breiliant Online Store; 34 transactions	The discovery of associations within the product was conducted by applying the Apriori algorithm to the transactional database provided.	The authors do not show the whole process of discovering association rules, nor the total number of rules they discovered, they only share the best association discovered: Original Liquid Bleaching Seeds, Harva Peeling Gel and Castor oil and it had 8.8% support and 30% confidence.

[1]	Supermarket	Produce new and personalized promotions to customers and create a new and convenient store layout	Transactional database provided by Anantha stores; from July 2011 until January 2012 with an average 962 transactions/day.	The k-Apriori algorithm was applied to all transactions and then to customers groups information separately, to generate association rules	The authors were able the generate 253 682 association rules and 33 263 frequent itemsets.
[17]	Grocery store	Create a new store design to give more shopping convenience to customers	Transactional database provided by the retail store; 1049 transactions	The process of identifying the related products and product categories was done by using the Apriori algorithm.	They were able to discover 5 product category associations. In order to obtain more specific results, succeeding processing of the category association was conducted. This originated in 14 subcategory association rules.

TABLE 1-RELATED WORK

Chapter 3: Methodology

3.1 Data Characterisation

The accomplishment of the dissertations' objective is supported on a transactional dataset provided by a Portuguese cosmetics company. This dataset consists of more than 6 million anonymized real transactions carried out between January 2017 until February 2020 by customers in their 161 physical stores throughout the Portuguese territory. The transactional database consists of the following information (as shown in table 2).

Name	Type	Analysed	Name	Type	Analysed
ID	Integer	Yes	Store Typology	Category	No
Nlines	Integer	No	Store locality	Integer	Yes
Terminal	Integer	No	Store age	Integer	No
OprRegisto	Integer	No	Store closing date	Date	No
Total	Numeral	No	Store area	Integer	No
Discount value	Numeral	No	Store Insignia	Integer	No
Amount paid	Numeral	No	Store Type	Category	No
Date	Date	Yes	Client Distance	Numeric	No
Hour	Hour	No	Line	Integer	No
TaxFree	Boolean	No	Code	Integer	Yes
DocType	Integer	No	Quantity	Integer	No
IdWithApp	Boolean	No	Selling price	Numeric	No
Nº Client	Integer	No	Discount Line	Numeric	No
Client Gender	Category	Yes	Total Line	Numeric	No
ClientHasPhone	Boolean	No	Liquid	Numeric	No
ClientHasAdress	Boolean	No	Total net line	Numeric	Yes
ClientHasTaxid	Boolean	No	Id Operator	Integer	No
ClientHasEmail	Boolean	No	Brand	Integer	No
ClientAge	Numeric	Yes	Range	Integer	No
ClientAntiquity	Numeral	No	Shaft	Integer	No
Store nº	Integer	Yes	Intro Date	Date	No

TABLE 2- VARIABLES IN THE DATABASE

The "analysed" table column indicates whether the variable was part of the analysis, so if the value is "Yes", the variable entered in the analysis and were used to generate association rules. On the other hand, if the value is "No", then the variable did not enter in the analysis. The variables were chosen based on the literature that investigates which variables influence the customer purchasing decisions. The variables that have a great impact on the consumers' choice include gender, age and location and given that the provided database contained these variables were chosen to be analysed. Other variables such as ClientHasMail, ClientHasPhone or ClientAntiquity were not included in the analysis because it was considered that they did not contribute with useful information regarding the consumer that would influence their purchasing decisions. For example, the variable "ClientHasPhone" is a Boolean that indicates if the customer has a phone number associated. It was considered that this variable does not provide useful information about the customers' purchasing decisions and therefore was not considered to be influential and was not further analysed.

Before starting the analysis, it is necessary to clean the database to get rid of data that is not suitable to be analysed and so, should not to be considered. This was also the case with the provided database. Each entry of the database represents a line of a certain invoice, and some aspects should be taken into consideration when looking at the Selling Price and the Total Net Line. The entries of the database in which the Selling Price is 0.00€ are to be dismissed because, they represent promotional codes or offer codes and so, they will generate rules with 100% confidence. However, given that these are offers provided to the customer for free, the rules are not valuable. One other aspect to be taken into account is the variable Total Net Line. Only entries with a positive value for this variable are taken into analysis, because when the value is negative it indicates that the item was returned and so, should not be considered when looking for associations in the purchased items. These invalid entries were about one million but given that the total number of entries is about 10 million, eliminating these invalid ones does not represent a threat.

The company that provided the data made sure that all the information concerning the transactions was anonymous and so, the database was given with every variable represented by numbers with the exception of "Date" and "Hour". For example, the variable store typology is composed by categories, but those categories were not revealed, and nothing can be assumed. However, the company disclosed some information about the database. They revealed the categories of the gender variable, the identity of store location nº188 and the name to which some product codes corresponded.

Given the format that the dataset was given, it was necessary to aggregate the transactions. The format in which the database was delivered with only the analysed variables is shown in table 3.

Tr.ID	Tr.Nº Lines	Cl.Gender	Cl.Age	Pr.Code	Tr.Date	Tr.Total
6094757	2	2	49	88919	16/06/2018	45.04
6094757	2	2	49	67519	16/06/2018	45.04
6094775	3	2	76	95794	24/03/2018	109.48
6094775	3	2	76	67786	24/03/2018	109.48
6094775	3	2	76	100392	24/03/2018	109.48

TABLE 3-EXAMPLE OF 5 OBSERVATIONS FROM THE DATABASE

As it is possible to observe in table 3, the transactions are in an atomic format, this means that each entry of the database represents the purchase of one product and entries with the same Transaction ID, belong to the same transaction/purchase. For example, the transaction 6094757 consisted in the purchase of 2 items, 88919 and 67519.

In order to find the sets of items that are bought together and generate the associations rules, it is necessary to aggregate the dataset in the "basket" format. In the basket format, the products that belong to the same transactions are in the same entry of a data frame. To do this, first it is necessary to aggregate the dataset based on the transaction ID, "Tr.ID". This is possible by using the ddply() function of the plyr package in RStudio. This function splits the data frame based on the transaction ID and Date and pastes the code of the products purchased on that transaction. In other words, the algorithm looks for which transactions have the same ID and date and joins the codes of the products present in those transactions. Then it creates another data frame with this data, as shown in table 4.

Tr.ID	Tr.Date	Pr.Code
6094757	16/06/2018	88919, 67519
6094775	24/03/2018	95794, 67786, 100392

TABLE 4-EXAMPLE OF AGGREGATED DATABASE

In this format, each entry of the database corresponds to a transaction with all the purchased items present in the Pr.Code cell, as can be observed in table 4. After aggregating the transactions, it was discovered that the database contained about 6 million transactions that were fit to be analysed.

In order to apply the Apriori algorithm, the transactions must be in the “basket” format and so, the next step is to make the variables “Transaction ID” and “Date”, null. This is necessary because the Apriori algorithm only analyses the codes of the items in the transaction, the date and ID of the purchases are not relevant in the analysis. Next, the function read.transactions() from the “arules” package is applied, and the transactions are now in the “basket” format, as shown in table 5.

Pr.Code
88919,67519
95794,67786,10032

TABLE 5-EXAMPLE OF "BASKET" FORMAT

The “basket” format means that each entry corresponds to the items from a transaction. For example, as shown in the table above (table 5), the first transaction corresponded to the purchase of two items, specifically the items 88919 and 67519 and the second transaction concerned the purchase of the items 95794,67786 and 10032

After having the transactions in “basket” format, it is possible to apply the Apriori algorithm to generate the association rules present in the transactions.

3.2 Filters used in the analysis

There are different processes involved in the consumer behaviour. When making a purchasing decision, the customer is influenced by several different factors, specificities and characteristics. Such factors include cultural, social, personal and other factors, shown in figure 2 [61].

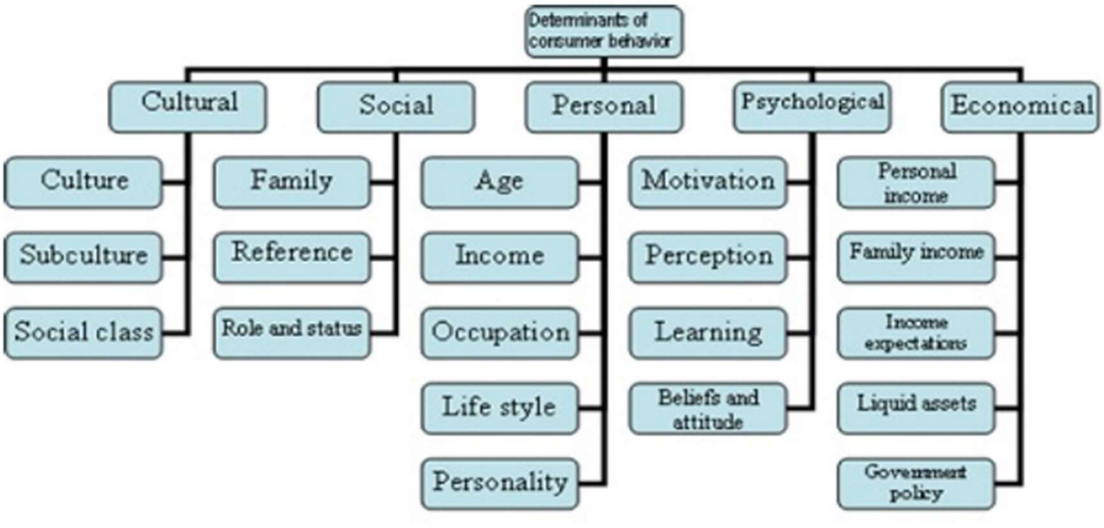

FIGURE 2 - DETERMINANTS OF CONSUMER BEHAVIOUR [61]

The variables analysed in this study are included in these factors. The first analysed variable is gender and as said before, it is a demographic and also a personal factor, as can be seen in the figure above (figure 2) . Gender is not only a biological concept; it is also a market segmentation variable and it has a strong influence on decisions. There are many differences in attitudinal and behavioural aspects between men and women because of psychological and physiological differences, for example, men are more likely to make purchases based on immediate. On the other hand, women consider purchases as a long-term decision [63]. In the database, this variable has 3 possible values: 0,1 and 2, where 0 represents the customers that did not to indicate a gender, 1 represented the male customers and 2 the female customers. Given that the female gender represents 80% (about 4 853 000) of total transactions, it was analysed in more detail than the other genders. And so, this gender was divided into the age groups and then analysed in order to generate more accurate and personalized.

The second variable added to the analysis was age and, as gender, it is a demographic and personal factor (figure 2). The age of the customer is one of the most important factors influencing purchases. What people buy changes over time because their taste and preferences also change; [61]. In order to analyse the data using this variable, it was divided into 5 age groups. The first age group included transactions from customers with ages between 18 and 24 years old. The second included transactions from 25 until 34 years old. The third, between 35 and years old. The fourth contained transactions with ages between 45 and 59 years old and finally the fifth age group covered the transactions from clients aged 60 years or over. However, it should be noticed that some transactions in the database had invalid numbers for this variable and naturally, these transactions were not taken into consideration when analysing by age. One other aspect that should be noticed is that the age groups begin at 18 years old, so only transactions made by adults were analysed.

Lastly, the third variable added was store location. This variable is a cultural factor, more specifically, a sub-cultural factor. Sub-cultural factors are a group of beliefs shared by groups of people that belong to a certain culture such as, religions, racial groups and geographic regions. This creates key market segmentations and companies have to design products and create marketing strategies tailored to their needs [61]. In total there are 77 different store locations and 161 different stores. To choose which stores to analyse, a count was made within the locations to ascertain the total number of stores in each location. Then, by assorting the total number of stores of each location, it was discovered that the location containing more stores is location 188 and it has 32 stores. Because analysing all 32 stores was very time consuming, the 5 stores with the most entries in the database belonging to locality 188 were chosen to be analysed. In reality, location 188, corresponds to the Greater Lisbon area.

Understanding the purchase decision making allows companies to gain more knowledge about their customers and it can create a foundation to creating more effective marketing strategies targeted specifically to their customers [63].

3.3 Analysis Process

The algorithm chosen to perform this analysis was the Apriori algorithm.This was the chosen algorithm because, it is the most known algorithm for mining transactional databases since it is exceptionally efficient at discovering all the frequent itemsets and association rules from a set of transactions [46, 78]. The Apriori algorithm performs in two steps. First, it discovers the frequent occurrence of single items and secondly, it discovers combinations of said single items while also checking if the occurrence of the combination is equal or above the defined support level This process is repeated until the largest frequent itemsets are found and the association rules are found by using all the subsets of the frequent itemsets [33]. In spite of being the most common algorithm used to perform MBA, the Apriori algorithm has a setback, the number of possible combinations enlarges as the number of items within the itemsets and rules increases which can make this method more expensive in terms of time spent analysing the output of the algorithm [78], which was also the case in this analysis. It was noted that the more transactions the cluster analysed had, the more association rules were generated. The Apriori algorithm was applied to discover association rules. In order to generate the association rules, it is necessary to choose a minimum value for the variable "support". This value variates depending on the number of transactions being analysed. The more transactions there are, the smaller the support has to be. For example, when analysing the transactions made by female customers, the value of support was 0,00001 because there were being analysed millions of transactions. To generate association rules, as well as being necessary to specify the minimum support, it also required to define a value for confidence. Such as the case of the support, the confidence also depends on the number of transactions being analysed.

Among the generated rules, were “duplicated rules”. Two rules are said to be duplicated when the right-hand side and left-hand side of the rules are interchanged. In other words, they represent the same rule, but with the items interchanged. An example of these rules is as shown in table 6.

Rule	Support	Confidence	Lift
A -> B	S	C1	L
B -> A	S	C2	L

TABLE 6- EXAMPLE OF DUPLICATE RULES

These rules have to be removed before analysing the rest of the generated rules [79].

One other aspect to have in consideration is the length of rules generated which was 2. This means that the minimum number of items in the association rules and rules was 2. If the minimum length was kept with the default value, 1, rules with only 1 item would be generated. These rules show which products customers buy regardless of the purchase of others, which is not useful to discover sets of items that are bought together. Therefore, the chosen minimum value for the variable was 2.

Regarding the generation of association rules, the first cluster to be analysed was the data split only by gender, male and female. The algorithm was applied to each of the clusters and generated several association rules. After the rules were generated for both genders, male and female, the next clusters to be analysed are the 5 age groups of female customers. Next, a search was conducted to discover the store location with the most stores and it was discovered that the location 188, which corresponds to the greater Lisbon area, had the most stores, more specifically, 32. The data of the most significant stores of the greater Lisbon area, mentioned in the previous section 3.2, Filters used the analysis are analysed, specifically, the stores 264, 316, 355, 377 and 401._After generating rules for each store, the next step was to analyse them in order to realize if there were any similarities in the patterns. It was also decided to analyse the store that had the most transactions, store 377. Thus, the data of this store was split by gender and age groups and the algorithm was applied to generate association rules.

After applying this algorithm and cleaning the generated results, meaning removing the duplicate and redundant rules, it is now possible to draw valuable conclusions as to, not only, which items are bought together but also, the demographic characteristics of the customers who purchase them. Based on these conclusions, the retailer can create new marketing strategies that will improve the relationship with the customer, increase sales and ultimately, increase the revenue.

Chapter 4: Experimental Results and Discussion

4.1 Generating Association Rules based on customers' gender

It was decided to investigate according to gender because it was considered that it would create more useful information than not applying any filters to the original data. The rules generated without applying any filters would create too broad information because they do not take into consideration any personal and demographic information. It was considered more convenient to know which products are bought by customers belonging to the same gender. Based on the database provided, it is possible to do this, by filtering the data by the customers' gender.

4.1.1 Generating association rules for male customers

After filtering the data using the gender variable, we can conclude that about 20% (about 1 196 992 transactions) of the population are male customers. The application of the Apriori algorithm generated 8 914 association rules and the 10 most valuable are represented in table 7. Each entry of the table represents a rule and each rules involves the codes of the associated items. For example, the first entry says that the items with the codes {106115} and {102822} are associated.

Rule	Support	Confidence	Lift
{106115}=>{102822}	0.0007	0.41	135.83
{96762} => {102822}	0.0003	0.68	229.34
{96762} => {106115}	0.0002	0.43	250.59
{114305} => {90748}	0.0002	0.56	990.09
{99152} => {88435}	0.0002	0.41	116.55
{81661} => {98933}	0.0001	0.55	1471.06
{106115,96762}=>{102822}	0.0001	0.56	188.13
{114235} => {117819}	6.02e-05	0.54	2646.75
{98494} => {107782}	5.59e-05	0.72	4636.28
{89019} => {110310}	4.01e-05	0.4	2712.48

TABLE 7- TOP10 ASSOCIATION RULES FOR MALE CUSTOMERS

It is worthy of noting that within these rules, four rules involve the items 102 822 and 96762. These items correspond to gift checks. Hence, one of the best rules, with high support and high confidence, {96762} => {102822}, involves both gift checks.

Given that the generated rules with the best support did not show high values of confidence, rules with lower support but with high confidence were also chosen. For example, it was discovered that 72% of the male customers that bought the product {98494}, also bought the product {107782}.

One other rule that presents good values in support and confidence is the rule {114305} => {90748}. This rule states that male customers that consume the product {114305} which corresponds to a hair removal service will also purchase the product {90748}, which corresponds to another hair removal service, with a likelihood of 68%.

All the rules in the table above present high values of lift. This means that the items are complementary and so, when the sales of one item increase, the same will happen to the sales of the other.

After analysing the rules, an average was calculated using the maximum prices for each of the items included in the rules above. Regarding the male customers, indicated that they spent in each item an average of 33€.

Among the generated rules, there are rules with very low support but very high confidence. These rules should not be discarded because although presenting a low value of support they show high confidence. Some examples of these rules are present in table 8.

Rule	Support	Confidence	Lift
{110194,89056}=>{89034}	8.35e-06	1	2468.03
{110194,89056}=>{75124}	7.52e-06	0.9	1173.52
{115359} => {85149}	6.68e-06	0.8	47879.72
{108744,66527}=>{97242}	6.68e-06	0.8	13299.92
{86811,94978}=>{111102}	6.68e-06	0.89	861.53

TABLE 8- LOW SUPPORT RULES FOR MALE CUSTOMERS

With the new information, the retailer can promote products associated with male transactions to the customers that identify with that gender. This will result in effective and more personalized promotions which will lead to revenue increase and higher customer satisfaction. One other advantage of this information is if there is a store zone for male products, then that area can now be re-designed according to the generated association rules and thus, facilitating the shopping experience for the customers.

4.1.2 Generating association rules for female customers

Regarding the transactions made by female customers, about 80% (4 748 268 transactions) of the population belong to this gender. The Apriori algorithm generated 29 500 association rules and the 10 most valuable are shown in table 9.

Rule	Support	Confidence	Lift
{114305} => {90748}	0.0023	0.52	101.07
{81661} => {98933}	0.0013	0.69	269.95
{96762} => {102822}	0.0003	0.64	242.25
{92843} => {99749}	0.0002	0.48	649.97
{89074} => {114305}	0.0002	0.57	127.29
{116255} => {98257}	0.0002	0.73	920.41
{76503} => {109582}	0.0002	0.79	1827.93
{66776} => {97708}	0.0002	0.78	703.33
{106177}=>{111742}	0.0001	0.61	2106.24
{75720} => {97708}	0.0001	0.93	835.15

TABLE 9-TOP10 RULES FOR FEMALE CUSTOMERS

When analysing the rules above it is noteworthy to notice that as the male customers, the female customers also have a rule that associates two gift checks, in the case of female customers, the rule states that 64% of the customers that bought the gift check {96762} also bought the gift check {102822}.

Within these rules, is it interesting to notice the rules that show the higher value of confidence, for example, the rule {66776} => {97708} has a value of confidence of 78%. Both of the items involved in this rule correspond to hairdressing services. This rules says that 78% of the female customers that consumed the hair dressing service {66776} also consumed the hairdressing service {97708}.

Another rule that presents a high value of confidence is the rule {76503} => {109582}. This rules claims that 79% of the customers that bought the product {76503} also bought the product {109582}. The product {76503} corresponds to another hairdressing service and the product {109582} corresponds to an eyes service.

Lastly, the rule that presented the highest value of confidence, within the rules in table 15, was the rule {75720} => {97708}. These items also refer to hairdressing services and the rules claims that 93% of the female customers that purchased the hairdressing service {75720}, also purchased the hairdressing service {97708}.

Just as in the case of male customers, an average using the prices of the items included in the rules was also conducted and in the case of female customers, the mean was of 14€.

Once again, all the rules associated with female customers also present high values of lift, indicating that all the products involved in each rules are complementary.

It is also interesting to analyse the generated rules that although having low values of support, have high values of confidence. Some of these rules are shown in table 10.

Rule	Support	Confidence	Lift
{105180,84132}=>{98632}	1.05e-06	1	13263.32
{89158,90610}=>{102565}	1.05e-06	1	1687.97
{83669,96498} => {93771}	1.26e-06	1	8942.12
{80103,82862}=>{108982}	1.47e-06	1	6190.70
{117596} => {85887}	1.89e-06	1	38292.49

TABLE 10-LOW SUPPORT RULES FOR FEMALE CUSTOMERS

4.2 Generating Association Rules based on customers' gender and age

As said before, 80% of the population is female. Thus, the generation of association rules based only on gender is useful however, too broad. So, it is preferable to add more variables to limit the population and thus obtain more accurate information. As said in the methodology section, the chosen variable was the age of the customer and so, the transactions made by the gender 2 were divided into 5 age groups.

4.2.1 Generating association rules for age group 1

The first age group contains all transactions made by people aged between 18 and 24 years. There were about 400 000 transactions in this group and 3 770 association rules were generated. The 10 most relevant rules are represented in table 11.

Rule	Support	Confidence	Lift
{114305} => {90748}	0.0044	0.56	54.74
{81661} => {98933}	0.0028	0.74	131.32
{116255} => {98257}	0.0004	0.88	511.41
{106177}=> {111742}	0.0003	0.61	935.91
{76503} => {109582}	0.0003	0.76	975.01
{96762} => {102822}	0.0003	0.62	325.1
{109919}=> {111742}	0.0002	0.41	628.21
{75720} => {97708}	0.0001	0.95	1348.53
{89515} => {97708}	0.0001	0.87	1231.34
{89074} => {114305}	0.0001	0.47	59.84

TABLE 11-TOP10 ASSOCIATION RULES FOR FIRST AGE GROUP

Within the rules represented on the table above, it is interesting to notice that 3 of them present values of confidence above 80%. One of these rules is the rule {89515} => {97708}, this rules involves two hairdressing services and it claims that 87% of the customers that bought the {89515} hairdressing service, also bought the hairdressing service {97708}.

One other rule among the best is the rule {116255} => {98257}. This rule involves two nail services and it states that 88% of the customers that bought the nail service {116255} also bought the nail service {97708}.

Finally, the rule that has the highest number of confidence is the rule {75720} => {97708}. This rule implies that people that consume the hairdressing service {75720} will also consume the hairdressing service {97708} with a likelihood of 95%.

The rule {109919}=> {111742} stands out because it is not present in any other table of age groups. Suggesting that this rule is particular to this age group.

Among the 7728 generated association rules regarding this age group, were rules that show low values of support but high values of confidence. Some of these rules are shown in table 12.

Rule	Support	Confidence	Lift
{75116}=>{70255}	9.98e-06	1	80131.8
{73582}=>{103232}	9.98e-06	1	66776.5
{93072} => {90224}	9.98e-06	1	80131.8
{106754}=>{102342}	7.49e-06	1	266.92

{104665,94966}=>{102342}	7.49e-06	1	266.93

TABLE 12-LOW SUPPORT RULES FOR AGE GROUP 1

4.2.2 Generating association rules for age group 2

The next age group in study involved transactions made by customers aged between 25 and 34 years old. This group contained 911 468 transactions and 8 535 association rules were generated. Within the rules generated the 10 most valuable are shown in table 13.

Rule	Support	Confidence	Lift
{114305} => {90748}	0.0035	0.52	68.67
{81661} => {98933}	0.0025	0.71	155.94
{114235}=> {117819}	0.0008	0.52	276.85
{96762} => {102822}	0.0004	0.64	250.17
{76503} => {109582}	0.0004	0.85	1083.61
{116255} => {98257}	0.0003	0.71	621.53
{75720} => {97708}	0.0002	0.98	633.79
{89515} => {97708}	0.0001	0.83	538.77
{66776} => {97708}	0.0001	0.76	489.97
{104991,114235}=>{117819}	0.0001	0.80	462.22

TABLE 13-TOP10 ASSOCIATION RULES FOR SECOND AGE GROUP

Within the generated rules for this age group represented in table 13, it is possible to notice that 4 of them present a value of confidence above 80%. One of these rules is the rule {104991,114235}=>{117819}. The items 104991 and 114235 relate to hair removal services and the item 117819 relates to eyes services. This rules states that 80% of the customers of this age group that purchased the hair removal services {104991} and {114235} also bought the eyes service {117819}. This rule stands out also because it is not present in any other table regarding age groups.

One other important rule related to this age group is the rule {76503} => {109582}. This rule involves two hairdressing services and it states that 85% of the customers of this age group that bought the hairdressing service {76503}, also bought the other hairdressing service {109582}.

The rule that presents the highest value of confidence is the rule {75720} => {97708}. As said before, the items involved in this rule relate to hairdressing services and it claims that 98% of the customers that purchased the hairdressing service {75720}, also purchased the other hairdressing service {97708}. It is to notice that this rule is also present in table 12 that shows the most valuable rules for female customers with ages between 18 and 24, also with a high value of confidence, specifically, 95%.

The generated rules that despite of presenting low values of support, have good confidence values, are show in table 14.

Rule	Support	Confidence	Lift
{103779,104592}=>{74796}	9.87e-06	0.9	5542.71
{114264,74645}=>{111352}	9.87e-06	0.9	22170.86
{72325,91262}=>{96446}	7.68e-06	1	11993.01
{84475,88007} => {66846}	6.58e-06	1	32552.46
{87490} => {93771}	5.49e-06	1	11252.70

TABLE 14-LOW SUPPORT RULES FOR AGE GROUP 2

4.2.3 Generating association rules for age group 3

The third age group to be studied covers the transactions made by customers aged between 35 and 44 years old. This group consists of 1 068 176 transactions and about 9 393 association rules regarding this group were generated, the 10 most valuable being represented in table 15.

Rule	Support	Confidence	Lift
{114305} => {90748}	0.0025	0.51	89.07
{81661} => {98933}	0.0013	0.65	271.09
{92843} => {99749}	0.0004	0.52	612.91
{96762} => {102822}	0.0004	0.65	238.48
{66776} => {97708}	0.0003	0.76	642.15
{89074} => {114305}	0.0002	0.64	132.19
{89515} => {97708}	0.0002	0.74	619.71
{93152} => {97708}	0.0002	0.81	677.01
{76503} => {109582}	0.0002	0.77	1856.27
{106177}=>{111742}	0.0001	0.59	2759.3

TABLE 15-TOP10 ASSOCIATION RULES FOR AGE GROUP 3

Regarding the rules of this cluster, the rule that showed the highest value of confidence was the rule {93152} => {97708}. This rule involves two hairdressing services and it indicates that 81% of the customers belonging to this cluster that bought the hairdressing service {93152}, also bought the hairdressing service {97708}.

One other rule with good values of support and value is the rule {76503} => {109582}. This rule claims that 77% of the customers that bought the item {76503}, which corresponds to a hairdressing service, also purchased the item {109582}, which corresponds to an eye service.

Concerning this age gorup, 9 393 association rules were generated. Among them were rules with low support but with high confidence. Some of these rules are shown table 16.

Rule	Support	Confidence	Lift
{117596}=>{85887}	8.43e-06	1	66761.06
{94576,98494}=>{66776}	8.43e-06	1	2863.74
{67194,93423}=>{73610}	8.43e-06	0.9	3547.45
{104508,93473}=>{78410}	8.43e-06	0.9	1680.7
{91720,93473} => {78410}	6.55e-06	1	1867.44

TABLE 16-LOW SUPPORT RULES FOR AGE GROUP 3

4.2.4 Generating association rules for age group 4

The next age group in study covers the transactions mady by customers aged between 45 and 59 years old, containing 1 033 810 transactions. After aplying the apriori fuction, 8 380 association rules were generated and the 10 most important are shown in the table 17:

Rule	Support	Confidence	Lift
{99749} => {92843}	0.0004	0.42	612.91
{96762} => {102822}	0.0004	0.65	238.48
{89074} => {114305}	0.0002	0.64	132.19
{93152} => {97708}	0.0002	0.81	677
{76503} => {109582}	0.0002	0.77	1856.27
{75720} => {97708}	0.0001	0.94	789.1
{79539} => {97708}	0.0001	0.63	525.83
{66986,98933}=>{81661}	0.0001	0.71	358.15
{117819,66986}=>{114235}	0.0001	0.66	502.49
{75232} => {97708}	0.0001	0.67	563.76

TABLE 17-TOP10 ASSOCIATION RULES FOR AGE GROUP 4

Among the rules represented above in table 17, it is possible to observe that 2 rules have confidence values above 80%. The rule {93152}=>{97708} states that 81% of the customers belonging to this age group that bought the product {93152}, also bought the product {97708}. Both of these items relate to hairdressing services.

The other rule that shows high values of confidence is the rule {79539} => {97708} and the items involved in this rule also relate to hairdressing services. This rule states that 94% of the people in this cluster that bought the hairdressing service {79539}, also bought the hairdressing service {97708}.

The rule involving the products {66986}, {98933} and {81661} stands out because it does not appear in any other regarding female age groups. This rule states that 71% of the customers that bough the products {66986} and {98933} together, also bought the product {81661}.

Among the generated rules, there were rules with low support but with high confidence. Some of these rules are shown table 18:

Rule	**Support**	**Confidence**	**Lift**
{88996,96146}=>{75232}	9.36e-06	1	6634.64
{117596} => {85887}	8.42e-06	1	66761.06
{117596} => {71720}	8.42e-06	1	3317.32
{78410,85302}=>{104508}	6.55e-06	1	3513.74
{91720,93473} => {78410}	6.55e-06	1	1867.44

TABLE 18-LOW SUPPORT RULES FOR AGE GROUP 4

4.2.5 Generating association rules for age group 5

The last age group in study involves customers aged above 60 years old. This group contains 376 505 transactions and 5 422 association rules were generated. The 10 most valuable are represented in the following table 19.

Rule	Support	Confidence	Lift
{90748} => {114305}	0.0009	0.56	301.26
{87880} => {114235}	0.0005	0.94	621.34
{96762} => {102822}	0.0004	0.60	176.01
{89074} => {114305}	0.0003	0.73	393.61
{96762} => {106115}	0.0003	0.42	220.83
{81661} => {98933}	0.0002	0.79	2442.77
{66986} => {114235}	0.0002	0.43	287.52
{112104}=>{103069}	0.0002	0.42	310.47
{92843} => {99749}	0.0002	0.48	957.68
{116255} => {98257}	0.0001	0.98	1852.65

TABLE 19-TOP10 ASSOCIATION RULES FOR AGE GROUP 5

Among the rules represented in the table above, there are 2 rules that stand out from the others due to their confidece values. The first one being the rule {87880} => {114235} and both items relate to hair removal services. The rule states that 94% of the people that bought the hair removal service {87880}, also bought the hair removal service {114235}.

The other rule that presents a high value of confidence is the rule {116255} => {98257}. This rule involves items related to nail services and it implies that 98% of the customers that bought the nail service {116255}, also bought the nail service {98257}.

Some of the rules generated the in spite of showing low support valules, showed high confidence values are shown in the table 20 bellow.

Rule	Support	Confidence	Lift
{98070}=>{71296}	7.97e-06	1	94126.5
{113641}=>{78182}	7.97e-06	1	75301.2
{85587}=>{117535}	7.97e-06	1	47063.25
{69753}=>{108579}	7.97e-06	1	94126.5
{113203}=>{107993}	7.97e-06	1	41834

TABLE 20-LOW SUPPORT RULES FOR AGE GROUP 5

Based on these association rules, the retailer can now implement new efficient strategies. By knowing which products are related, the retailer can create a new store design and place the related item near each other. Another example of the usage of this information is the capability to create new pricing strategies by promoting one item and increase the price of its' related item(s). The association rules based on gender and age of the customers enables the retailer to create new, personalized and effective promotions that will lead to higher customer satisfaction and ultimately, revenue increase.

4.3 Generating Association Rules by store location

As said before in section 3.2, Filters Used in the analysis, consumer patterns can change depending on geographical locations [61]. Given this, it was chosen to use the variable store location. This allows retailers not only to understand the similarities in consumer patterns of stores in the same locality but also, the differences in consumption patterns between stores in different locations. In this section, 5 stores of the region containing the highest number of stores, which corresponded to the Greater Lisbon metropolitan area, were analysed,

4.3.1 Store 316

This store had 163 168 transactions in the database and the 10 most important association rules for the store are represented in table 21.

Rule	Support	Confidence	Lift
{96762} => {102822}	0.0003	0.54	205.7
{96762} => {106115}	0.0003	0.49	277.5
{114305} => {90748}	0.0003	0.48	696.06
{106177}=>{111742}	0.0003	0.63	1513.56
{89019} => {115658}	0.0002	0.43	837.93
{109919}=>{111742}	0.0001	0.47	1123.19
{99485} => {115924}	0.0001	0.40	909.69
{100049,68712}=>{92338}	0.0001	0.68	309.28
{82926} => {67823}	0.0001	0.41	786.69
{100049,91649}=>{92338}	0.0001	0.68	309.93

TABLE 21- TOP10 ASSOCIATION RULES OF STORE 316

By analysing the table above, there are certain rules that stand out due to their confidence values. The first one being the rule {106177}=>{111742}. The items involved in this rule are creams and the rule states that people that 63% of the customers that frequent this store that bought the cream {106177}, also bought the cream {111742}.

The other 2 rules that show values of confidence of 68% have two items in common. The items 100049 and 92338 are in both rules and they relate to hand creams. The items that are different in each rule, {68712} and {91649}, are also hand creams. However, they form two different rules when associated with the hand cream {100049}. The rule {100049,68712}=>{92338}, says that 68% of the customers that bought the hand creams {100049} and {68712}, also bought the hand cream {92338}. The other rule presents the same value of confidence but the drivers for the purchase are the hand creams 100049 and {91649}. This suggests that the items {100049} and {92338} are popular within this store.

It is also notable that the rules {114305} => {90748} and {96762} => {102822} are present in this table. As said before, the items {114305} and {90748} concern hair removal services and the items {96762} and {102822} are gift checks.

4.3.2 Store 377

One other store that was analysed was store 377, this store had 181 056 transactions. Among the association rules that were generated, the 10 most important association rules are represented in table 22.

Rule	Support	Confidence	Lift
{114305} => {90748}	0.0005	0.44	185.07
{117819} => {114235}	0.0002	0.51	964.93
{96762} => {102822}	0.0002	0.65	299.05
{106177} => {111742}	0.0002	0.82	2481.14
{110310} => {115658}	0.0002	0.39	619.06
{89019} => {115658}	0.0002	0.42	677.01
{115065} => {73610}	0.0002	0.58	1553.18
{89074} => {90748}	0.0002	0.67	282.02
{105839} => {100049}	0.0001	0.42	314.33
{110310,89019}=>{115658}	0.0001	0.8	1281.81

TABLE 22- TOP10 ASSOCIATION RULES OF STORE 377

When comparing the confidence values of the rules above, two stand out. The first one being the rule {106177} => {111742}. The items in this rule are two creams and the rule states that 82% of the customers that visited this store and purchased the cream

{106177}, also bought the cream {111742}. It also interesting to notice that this rule is also present in table 21, which shows the most important rules regarding store 316.

The other rule that attracts attention for its confidence value is the rule {110310,89019}=>{115658}. The product {110310} corresponds to a moisturizing cream, the product {89019} is a cleansing cream and the product {115658} is a cream. This rule states that 80% of the customers that purchased the moisturizing cream {110310} and the cleansing cream {89019} also bought the cream {115658}.

Once again, the rules {114305} => {90748} and {96762} => {102822} are also present in this table.

4.3.3 Store 264

The next store to be analysed was store 264 and it had 78 868 transactions. In table 23 are represented the 10 most valuable association rules for this store.

Rule	Support	Confidence	Lift
{90748} => {114305}	0.0014	0.55	212.05
{96762} => {102822}	0.0007	0.70	251.66
{96762} => {106115}	0.0005	0.46	222.89
{115894} => {114305}	0.0004	0.65	247.26
{89074} => {114305}	0.0003	0.85	323.96
{89074} => {115894}	0.0002	0.46	758.36
{89074} => {90748}	0.0002	0.46	186.67
{101705} => {91885}	0.0001	0.55	1057.1
{115894,89074}=>{114305}	0.0001	0.92	350.95
{82078} => {105213}	0.0001	0.56	1848.49

TABLE 23- TOP10 ASSOCIATION RULES OF STORE 264

Regarding the association rules generated for this store represented in the table above, two rules present confidence values above 80%. One of these rules is the rule {89074} => {114305}. This rule says that 85% of the customers that bought the item 89074 also bought the item 114305. The other rule with the highest confidence value, is the rule {115894,89074}=>{114305}. This rule involves the same two items as the rule mentioned previously but it also associates them with the item {115894}. Regarding this rule it claims that 92% of the customers that visited this store and bought both products

{115894} and {89074}, also bought the product {114305}. The product {115894} is an eye service and the products {89074} and {114305} are hair removal services.

It is also interesting to notice the number of rules in which the hair removal service {114305} appears, which are 4 and that the rule {96762} => {102822} is present in this table. As mentioned before, the items in this rule are gift checks.

4.3.4 Store 335

Store 335 had 133 490 transactions in the database, and it was the fourth store to be analysed. In table 24, shown below, it is possible to observe the top 10 association rules regarding store 335.

Rule	Support	Confidence	Lift
{106115} => {102822}	0.0005	0.4	211.69
{114305} => {90748}	0.0002	0.60	1877.10
{79695} => {95874}	0.0002	0.46	634.81
{107403} => {90033}	0.0001	0.45	918.76
{106177} => {111742}	0.0001	0.70	3202.18
{102405} => {88435}	0.0001	0.43	165.35
{114950,69803}=>{114962}	0.0001	0.68	310.64
{100519} => {114168}	0.0001	0.42	249.48
{68368} => {100711}	0.0001	0.4	1089.72
{74380} => {109919}	0.00009	0.76	2686.35

TABLE 24- TOP10 ASSOCIATION RULES OF STORE 335

Regarding the rules mentioned in table 24, the rule that is noticeable due to its confidence value is the rule {74380} => {109919}. This rule presented the highest value of confidence when comparing it with the others present in the table. The items involved in this rule are both creams and according to the rule, 76% of the customers that bought the cream {74380}, also bought {109919}.

It is interesting to notice that the rules {106177} => {111742} and {114305} => {90748} are also present in this table and both show good values of confidence, 70% and 60%, respectively.

4.3.5 Store 401

The last store to be analysed was store 401 and it had 56 950 transactions. The 10 most valuable association rules related this store are shown in table 25.

Rule	Support	Confidence	Lift
{114305} => {90748}	0.0009	0.60	182.44
{106115} => {102822}	0.0008	0.43	112.82
{66275} => {67450}	0.0005	0.61	491.10
{92843} => {99749}	0.0003	0.68	1016.98
{100711} => {92338}	0.0002	0.42	243.70
{68368} => {68712}	0.0002	0.43	413.69
{77365} => {70522}	0.0001	0.73	2301.05
{77336} => {81103}	0.0001	0.8	1980.90
{73933,84539}=>{102370}	0.0001	0.89	1488.92
{91350} => {105213}	0.0001	0.86	5423.90

TABLE 25-TOP10 ASSOCIATION RULES OF STORE 401

When analysing the table above, the rule that shows the highest value of confidence is the rule {73933,84539}=>{102370}. This rule states the 89% of the customers of this store that bought both products, {73933} and {84539}, also bought the product {102370}. Regarding the items involve in this rule, the item {73933} is a men's perfume, the product {84539} corresponds to an after shave and the product {102370} is a deodorant.

One again, it is to noteworthy to notice that the rule {114305} => {90748} is also present in this table.

With this information the retailer can create new marketing strategies and adapt to the needs of the customers that shop in this location, which is ultimately improve the relationship with the customers and increase revenue.

4.4 Deeper analysis of Greater Lisbon

Given that the Greater Lisbon contains more than 1 million entries in the database and has 32 stores, it would be interesting to analyse in more depth this location. So, given the importance of this area, the most important store in terms of number of transactions, store 377, was analysed by filtering by gender and age groups. However, due to the fact that the percentage of transactions made by male customers, which is 21% (38 610 transactions), it was chosen to analyse only the transactions made by female customers, which represents about 78% the database (140 924 transactions).

As before, the first age group is composed by female customers with ages between 18 and 24 years old and it contained 14 454 transactions. The 10 most important association rules related to this cluster are represented in table 26.

Rule	Support	Confidence	Lift
{99259} => {98257}	0.0011	0.52	28.15
{66986} => {114235}	0.001	0.7	459.93
{114305} => {90748}	0.001	0.47	75.79
{115894} => {88996}	0.001	0.45	112.55
{117819} => {114235}	0.0008	0.92	602.29
{106177} => {111742}	0.0005	1	1606.11
{84167} => {89266}	0.0005	0.54	370.64
{101705} => {91885}	0.0003	0.71	1032.5
{105839} => {100049}	0.0003	0.67	507.19
{67703} => {115286}	0.0003	0.8	1927.33

TABLE 26- TOP10 ASSOCIATION RULES OF AGE GROUP 1 IN STORE 377

Three rules stand out when looking at the confidence values. The rules being {106177}=>{111742}, {117819} => {114235} and {67703} => {115286}.

The rule {106177}=>{111742} presents the highest confidence value and it claims that 100% of the customers belonging to this cluster that bought the cream {106177}, also bought the cream {111742}.

The second rule that associates two services, more specifically, the product {117819} is an eye service and the product {114235} is a hair removal service. This rule states that 92% of the customers that bought the eye service also bought the hair removal service.

Lastly, the rule {67703} => {115286} has a confidence value of 80%. The product {67703} is an eye cream and the item {115286} is a cream. The rule claims that 80% of the customers that bought the eye cream also bought the cream.

The next age group to be analysed was customers with ages between 25 and 34 years old and there were 24 240 transactions regarding this group. The top 10 most important association rules regarding this cluster are represented in table 27.

Rule	Support	Confidence	Lift
{114305} => {90748}	0.0015	0.78	149.38
{117819} => {114235}	0.0007	0.64	577.17
{90033} => {107403}	0.0006	0.44	232.49
{115065} => {73610}	0.0005	0.65	492.4
{89074} => {90748}	0.0005	0.86	163.61
{70273} => {90748}	0.0005	0.55	104.11
{96762} => {102822}	0.0003	1	515.76
{98821} => {115871}	0.0003	0.57	865.75
{109503} => {80605}	0.0003	0.78	1450.32
{115384} => {80396}	0.0003	0.58	565.62

TABLE 27- TOP10 ASSOCIATION RULES OF AGE GROUP 2 IN STORE 377

Regarding the rules in the table above it is noticeable that 100% of the customers of this cluster that bought the item {96762} also bought the item {102822}. As said previously, these items correspond to gift checks.

One other interesting rule to analyse regarding this cluster, is the rule {89074} => {90748}. This rule says that 86% of the customers that consumed the hair removal service {89074}, also consumed the hair removal service {90748}.

The age group of customers with ages between 35 and 44 years old was the next age group to be analysed and it contained 27 659 transactions. The top 10 most important association rules are represented in table 28.

Rule	Support	Confidence	Lift
{69820} => {71548}	0.0005	0.57	236.87
{81989} => {76444}	0.0004	0.71	673.24
{114305} => {90748}	0.0004	0.67	335.26
{99282} => {108741}	0.0004	0.59	229.15
{99282} => {85469}	0.0003	0.53	542.33
{85916} => {94304}	0.0003	0.8	713.78
{85469,99282}=>{108741}	0.0003	0.89	346.28
{101177} => {109050}	0.0003	0.88	346.28
{115894} => {88996}	0.0003	0.78	551.60

{96762} => {102822}	0.0003	0.7	254.75

Table 28- Top10 association rules of age group 3 in store 377

The rule that stands out when analysing the confidence values of the rules above is the rule {85469,99282}=>{108741}. Regarding the identity of the items involved in this rule, the item 85469 a foundation and the items {99282} and {108741} are creams. This rule indicates that 89% of the customers contained in this cluster that purchased both the makeup foundation {85469} and the cream {99282}, also bought the cream {108741}.

The next group to be analysed was the group regarding purchases made by female customers with ages from 45 to 59 years old. This cluster contained 29 721 transactions, this being the cluster with more transactions. The 10 most important association rules regarding this group are represented in table 29.

Rule	**Support**	**Confidence**	**Lift**
{106659} => {98257}	0.0012	0.6	130.42
{90748} => {98257}	0.0009	0.44	97.13
{80187} => {89372}	0.0004	0.87	559.98
{115894} => {80367}	0.0003	0.53	80.47
{106411} => {108909}	0.0003	0.47	777.05
{90194} => {89372}	0.0003	0.44	287.17
{86067} => {99232}	0.0002	0.58	1083.61
{103617,106659}=>{98257}	0.0002	1	218.54
{106140} => {89372}	0.0002	0.67	430.75
{106177} => {111742}	0.0002	1	4246

Table 29- TOP10 Association Rules of Age Group 4 in Store 377

It is noteworthy that two rules in the table above present the maximum value for confidence. The first one being {103617,106659}=>{98257}. This rule states that 100% of the customers that purchased both items {103617}, which is an eye service, and {106659},which corresponds to a nail service, also bought the product {98257}, that also represents a nail service.

The other rule that presents the highest possible value of confidence is the rule {106177}=>{111742}. As said before, the items {106177} and {111742} represent two creams and according to the rule, 100% of the customers belonging to this cluster that bought the cream {106177}, also bought the cream {111742}.

The last age group to be analysed was the group of customers with ages equal and above 60 years old and it had 10 721 transactions, making this cluster the smallest of all.

The 10 most valuable association rules related to this age group are represented in table 30.

Rule	Support	Confidence	Lift
{90748} => {80367}	0.0041	0.83	54.95
{114305} => {80367}	0.0028	0.43	28.78
{98996} => {98257}	0.0012	0.93	207.42
{98996} => {114305}	0.0011	0.86	133.19
{106115} => {102822}	0.0007	0.42	94.05
{91390} => {93630}	0.0006	0.86	612.69
{96762} => {102822}	0.0006	0.75	167.53
{100429} => {87859}	0.0005	0.71	186.79
{75945} => {97165}	0.0003	1	564.32
{99045} => {68304}	0.0003	0.8	1429.6

TABLE 30-TOP10 ASSOCIATION RULES OF AGE GROUP 5 IN STORE 377

Regarding the rules associated with the last cluster, one rule presents the highest possible value of confidence. This rule states that 100% of the customers that bought the product {75945}, which is a cream, also bought the item {97165}, which represents a night cream.

One other rule that stands out from the other rules, is the rule {98996} => {98257}. This rule indicates that 93% of the customers that consumed the foot service {98996}, also consumed the nail service {98257}.

4.5 Discussion

An MBA was performed to investigate the existence of association rules within the transactional data provided by a cosmetic retail company and display what said associations consisted of. The analysis resulted in several association rules between products purchased by customers with similar demographic and geographic factors, such as gender, age and location.

The first step of the analysis consisted of cleaning the original data and removing the transactions that were not suitable for analysis. Next, the data was divided by customers' gender, male and female and, the first cluster to be analysed was the cluster composed of transactions carried out by men, which consisted of 1 196 992 transactions. It was noticed that male customers have an inclination to buy the items {102822} and {96762}, which correspond to two gift checks. These two items were present in four out of the ten most significant rules generated for this cluster. The fact that these items are so present in the rules suggests that male clients have a high likelihood of buying gift checks. To even further this suggestion, one of the best rule regarding this cluster, says that 68% of men that bought the gift check {96762}, also bought the gift check {102822}, which shows a tendency of men to buy both gift checks. One other appealing rule regarding this gender was the rule that associates the products {98494} and {107782}. This rule states that 72% of the male customers that purchased the item 98494, also bought the item 107782.

The next cluster to be analysed was the one that covered transactions made by female customers, which contained 4 748 268 transactions. It was noticed that, like men, the female customers also have a tendency of buying gift checks. The rule that suggests this claims that 64% of the female customers that bought the gift check {96762}, also bought the gift check {102822}. Given the importance of this cluster, it was decided that it would be interesting to split the data by age to obtain more personalized information, so the data was split into five age groups, and the algorithm was applied to each of these clusters. It was discovered that among the generated rules, some seemed to be particular of each cluster. For the cluster covering female customers aged between 18 and 24, it was discovered that they had a slight inclination to buy the cream {111742} when also buying the cream {109919}. This rule stated that 41% of the customers of this cluster that bought the cream {109919}, also bought the cream {111742}. Although not showing a very high confidence value when compared to the other rules, this rule was considered to be characteristic of this cluster because it did not appear in any other Top10 of any other age group. Regarding the cluster covering transactions made by female customers with ages between 25 and 34 years old and it was discovered that these customers tend to purchase the items {104991},{114325} and {117819} together. These items correspond to hair removal and eye services, and the rule states that 80% of the customers that bought both the hair removal services {104991} and {114235}, also bought the eye service {117819}. Regarding the transactions made by female customers aged between 45 and 59 years old, it was discovered that they have an inclination to buy the product {81661}, when also buying the products {66986} and {98933}. The rules states that 71% of the customers belonging to this cluster that bought both the products {66986} and {98933}, also bought the product {81661}. The last cluster to be analysed was composed of transactions made by female customers over 60 years of age, and it was discovered that this cluster tends to buy the hair removal services {87880} and {114325} together. This is suggested by the rule that states 94% of the customers of this cluster that purchased the hair removal service {87880}, also bought the hair removal service {114325}. It is useful to notice that the hairdressing service {97708} associated with other hairdressing services is present in the rules of four out the five age groups. The only cluster that does not have any rule involving this item is the cluster covering transactions made by female customers aged above 60 years old. The fact that this item is present in several association rules generated made by women and shows high values of confidence in all of them is an indicator that this service is very popular among women. One other interesting aspect of the rules involving this item is that all of them present confidence values of at least 60%, which means that there is a significant inclination to buy this product. For example, one of the rules generated for the cluster that covered the transactions made by female customers aged between 18 and 24 years old involving the hairdressing service {97708}, states that 95% of the customers of this cluster that bought the hairdressing service {75720}, also bought the hairdressing service {97708}.

One other aspect that was investigated was the average price of the items involved in the rules. It was concluded that male customers spend on average 30€ per item, and female customers spend an average of 14€, on each item. It is interesting to notice that, although not shopping as much as women, men tend to spend more on each item, than women. In fact, the male average corresponds to more than double the women average, which can be significant in terms of sales volume.

In a world where the market is increasingly competitive, it important to be able to stand out from the competition and with this knowledge about customers' purchases, the retailer can apply it and create new and efficient strategies that will increase the relationship with customers, such as a new store layout or item promotions, which will, in the long run, increase revenue.

One other variable that was analysed was "store location". In this part of the analysis, the location with more stores was found by searching the database. It was discovered that the location that contained more stores was location 188, which corresponds to the area of the greater Lisbon, and it incorporates 32 stores and has more than one million entries in the database. After examining the 32 stores, the five that had more than 100 000 entries in the database were chosen to be analysed, more specifically, stores 264, 316, 335, 377 and 401. The next step consisted of analysing the transactions of each store and examine the results in order to discover if there were any resemblance in the purchasing patterns.

It was found that the rule, involving two creams, {106177} => {111742} appears in stores 316, 335 and 377 and all of them show good values of confidence, 63%, 70% and 82%, respectively. It is also important to notice the presence of the rule that involves two gift cheks, more specifically, {96762} => {102822}, which is present in stores 264, 316 and 377. This rule presents good values of confidence in all stores, specifically, 70%, 54% and 65%, respectively. Given that these two rules are present in three out of the five analysed stores, it suggests an inclination of the consumers, who frequent the greater Lisbon area, to consume the creams {106177} and {111742} and the gift checks. One other rule that was discovered as being common to the analysed stores is the rule {114 305} => {90748}, which is present in stores 316, 335, 377 and 401, this rule involves two hair removal services. In the store in which this specific rule does not appear, the contrary rule is present {90748} => {114305}, meaning that these items are present in all stores._Although showing lower confidence values when comparing with other rules mentioned above, it is interesting to notice the presence of this rule in the analysed stores. Being present in all stores, shows interest, even if not very high when compared to the others above, of the customers that frequent/reside in the greater Lisbon area, in the consumption of these items.

Now regarding the stores individually, it was discovered that some rules seemed to be specific to each store. For example, regarding store 316, the rule {100049,91649}=>{92328} seemed to be specific to this store because it did not appear as significant in any other analysed store. The items involved in this rule are three hand creams, and the rule states that 68% of the customers that bought both the {100049} and {91649} hand creams, also bought the hand cream {92328}. This show an inclination of the customers that frequent this store, towards the purchase of the hand cream {92328} when also buying the two other hand creams, {100049} and {91649}. Now speaking in terms of the store 377, the rule {110310,89019}=>{115658} appears to be specific, and it involves the moisturising cream {110310}, the cleansing cream {89019} and the cream {115658}. This rule indicates 80% of the customers who purchased both the moisturising and the cleansing creams also bought the cream {115658}. The next store to be analysed was store 264, and the rule that is specific to this store is the rule {115894, 89074}=>{114305}. This rule suggests that the people that frequent this store and purchase the eye service {115894} and the hair removal service {89074}, also tend to purchase the hair removal service {114305}, with 92% confidence. Now concerning store 335, the rule that seems to be particular is the rule that involves two creams specifically, 74380 and 109919.The rule states that 76% of the customers that visit this store and bought the cream {74380}, also bought the cream {109919}. The last store to be analysed was store 401 and the rule {73933, 84539}=> {102370} is particular to this store. This rule indicates that the customers of this store that buy the men's perfume {73933} and the aftershave {84539} also tend to buy the deodorant {102370}, with 89% confidence. Based on these rules, the retailer can create new marketing strategies that promote the consumption of the associated items. For example, by knowing that the men's perfume, the aftershave and the deodorant are associated, these items can be placed closer to each other in the store, which will improve the consumer experience when searching for these items.

One other aspect regarding the stores that was analysed was the average price of the items involved in the rules. After the calculations, it was discovered that the store with the highest average price was store 316 and the average price was 50€ per item, and the store with the lowest was store 264 with 8,3€. This indicates that the customers that frequent and purchase the store 316 are willing to spend more money per item, which suggests that this store is located in an area of the greater Lisbon where the socio-economic level is high. For example, it is interesting to notice that one of the rules connected to this store, {82926} => {67823}, involves two creams with very high prices, 279,83€ and 422,30€, respectively. On the other hand, store 264 suggests the opposite. The average price suggests that this store belongs to a part of Lisbon in which the population has a lower socio-economic level and is not able to pay much money for each item and choose to buy cheaper products.

Given the significance of the greater Lisbon area, it was considered interesting to the analysis, to examine in more depth the store located in greater Lisbon that had the highest number of transactions which was store 377 and contained 181 055 transactions. Given that the purchases made by female customers represented about 80% of total transactions (140 924) of this store, only transactions made by this gender were analysed. The transactional data was then, divided by the same age groups as before, and it was noticed that there were rules that seemed to be particular for each age group. The first age group consisted of female customers with ages between 18 and 24 years old and one interesting rule that is suggested to be characteristic of this age group was the rule {67703} => {115286}. This rule states that 80% of the customers of this cluster that bought the eye cream {67703}, also bought the cream {115286}. This suggests the female customers aged between 18 and 24 years old that frequent this store, have a tendency of buying the cream {115286} when also buying the eye cream {67703}, which means that these items are frequently bought together by the customers belonging to this cluster. One other exciting rule regarding this cluster is that 100% of the customers that bought the cream {106177}, also bought the cream {111742}. The next cluster to be analysed contained female customers aged between 25 and 34 years old. Regarding this cluster, it was noticed that costumers of this cluster tend to purchase the hair removal services {89074} and {90748} together. The rule that associates these services claims that 85% of the customers of this cluster that bought the hair removal service {89074}, also bought the hair removal service {90748}. One other rule that stood out from the other is the rule {96762} => {102822}. This rule involves two gift checks, and it states that 100% of the customers belonging to this cluster that bought the gift check {96762}, also bought the other gift check {102822}. The third cluster analysed covered transactions made by female customers with ages between 35 and 44 years old. It was found that customers belonging to this cluster tend to buy the foundation {85469} and both creams {99282} and {108741}, together. The rule stated that 89% of the customers that bought both the cream 99282 and the foundation {85496}, also bought the cream {108741}. The next cluster to be analysed consisted of transactions made by female customers aged between 45 and 59 years old. It was found that the customers of this cluster buy the eye service {103617} and the nail services {106659} and {98257} together. The rule that shows this claims that 100% of the customers that purchased the eye service and the nail service {106659}, also bought the other nail service {98257}. This rule shows that these three items are always bought together by this cluster. It is also interesting to notice that, just as the cluster involving female customers aged between 18 and 24 years old, this cluster also buys the creams {106177} and {111742} together. The rule involving these two creams states that 100% of the customers that bought the cream {106177}, also bought the cream {111742}. The last cluster to be examined covered transactions carried out by female customers aged over 60 years old. The analysis discovered that the customers of this cluster buy the cream {75945} and the night cream {97165} together. The rule showed that 100% of the customers the cream {75945}, also bought the night cream {97165}.

When reviewing existing body of knowledge for similar work, meaning, i.e where MBA was applied in a physical cosmetic store, no articles were found. The only articles found regarding the application of an MBA in a cosmetic store, only two articles were found and both concerning online stores. This represents an opportunity to carry out this analysis, since the lack of work in this specific area, that is a cosmetic physical store, justifies its realization.

In their study, Roodpishi and Nashtaei [65], the data of 300 clients of an insurance company was split using demographic variables such as age, gender, occupation, education level, marital status, place of residence and clients' income. Then, an MBA was applied to each of the clusters to discover hidden patterns within the insurance industry. An example of what the authors discovered is that the cluster composed by clients with ages between 35 and 50 years old, with 12 to 14 years of education and with an income of $5x10^6$-$20x10^6$ rials use their life insurance generally between 2 and 5 years and they usually renew their insurance in the third quarter of the year.

In their work, Abdullah et al, [16] carried out an MBA with the purpose of helping the online cosmetic store PureGlow with stock decisions. To do this, they applied the apriori algorithm to transactional data provided by the company and discovered that the best rule discovered had 93% confidence.

In their work, Hidayat et al [15], applied an MBA to transactional data of cosmetics provided by an online store, the Breillant Store. The purpose of this analysis was to understand the customers' attitudes. After applying the algorithm and generating rules, the authors discovered that the best rule discovered had 30% confidence and 8.8% support.

Chapter 5: Conclusions

5.1 Contributions

To date, a considerable body of research has sought to discover relationships between items sold in various kinds of stores in the retail area, both in online and physical stores, such as, groceries stores [17], online cosmetic stores [16, 17], supermarkets [1], among others. This research is carried out by performing an MBA to transactional data and it has provided retailers with a number of important insights, for instance, the ability to promote a better relationship with customers through a more personalized service for each customer. One other example, that is especially important in physical stores, is the capability to create an ideal store design that creates a better customer experience during their visit to the store.

However, the research on the application of MBA in physical cosmetic stores is notably lacking. During the research for the application of Market Basket Analysis on transactional data from cosmetic stores, only two articles were found regarding this subject and both were about online stores. In the searching for papers related to the application of an MBA in a physical cosmetics stores, no articles are found. The main and more important contribution of this study to the research community is that it fills this gap by applying the same technique, an MBA, and examine real transactions from several physical stores belonging to a cosmetics company in order to obtain associations between the items through generating association rules that discover the consumption patterns of their customers. This study also benefits the cosmetic company that can now take advantage of this knowledge and create new strategies, such as promotions and a new store layout, that ultimately will contribute to the success of the company.

5.2 Practical Implications

The most important practical implication this study brings is its contribution to the company that provided the data. Since this study performed an MBA on real transactional data, the results provide real insight into the consumption patterns of their customers. The company could benefit from this insight because new marketing strategies can be created based on the results. The analysis resulted in the generation of several association rules for each of the analysed clusters. Among these rules, some were found to be specific of certain analysed clusters. For example, it was found that female customers aged between 24 and 35 years old have a tendency for buying together the hair removal services

{104991} and {114235} and the eye service {117819}. By knowing this, the company can start to create new strategies such as item promotions and campaigns that encourage the purchase of these products to this female age group, making these strategies more effective. These strategies help the company decrease losses and increase gains, given that the likelihood of their success is high because it is based on the rule that indicates that the customers have a very high chance of buying those products together.

Another analysed factor was store location. The location containing the highest number of stores was found to be the greater Lisbon metropolitan area and the five stores with the highest number of transactions were selected for analysis. This part of the analysis revealed that there were certain association rules that were common to the analysed stores, suggesting that customers that frequent the stores of the greater Lisbon metropolitan area share preferences. Based on these rules, the company now has the opportunity to adapt to the needs of the customers that frequent that location and can promote the items involved in the common rules more strongly both in the form of promotions, as well as in the way the items are displayed in the store. Specific rules to each store were also found. As an example, it was found that customers who visit store 401 usually buy the men's perfume {73933}, the aftershave {84539} and the cream {109919} together. By knowing this, the company can develop a new store design where the associated products are placed near each other, in order to create a more pleasant and effective shopping experience allowing the customer to find the wanted items more easily and even persuade the customer to consume the related items.

5.3 Limitations and Future Research

This study has some limitations worthy of being discussed. The first limitation is related to possible issues with the provided database. Given that this study analyses transactional data from a database provided by a cosmetics company, the results depend on the state of the reliability of the provided database. This means that if the database contains errors the end results will not correspond to the wanted results, which are, accurate and reliable association rules. In other words, the association rules discovered would not reflect the real consumer patters of the customers.

The second and crucial limitation is related to the data's anonymity. As all the data was given anonymous, it was more difficult to understand if the associations were making sense or not. For example, one way to understand if the algorithm is working as it is supposed to, is to check the names or categories of the associated items and confirm if the associations are making sense. Although there are associations that are not so obvious, one way to perceive if the association rules generated are making sense is to check if the ones that are more obvious are within the results. Given that this information was anonymous, it was impossible to have this “control” over the results throughout the analysis. This limitation made it necessary to repeat the analysis because there were items that were being analysed that should not have been, items like offer codes and promotional codes. These items would have been easily identified if their identity had been revealed from the start. Although the company later revealed some information about certain items involved in the rules, it was only when requested. So, with this is mind, future studies should be able to perform this analysis with access to certain information from the star, such as, store location, name of the products, range of products and others to overcome this limitation and not have any setbacks that come from it. However, customer data privacy must be taken into consideration, but specially data concerning store information should be available.

References

[1] L. C. M. C. Annie and A. D. Kumar, “Market Basket Analysis for a Supermarket based on Frequent Itemset Mining,” *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 257–264, 2012.

[2] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos, “Retail business analytics: Customer visit segmentation using market basket data,” *Expert Syst. Appl.*, vol. 100, pp. 1–16, Jun. 2018.

[3] M. Kaur and S. Kang, “Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining,” in *Procedia Computer Science*, 2016, vol. 85, pp. 78–85.

[4] S. Gupta and R. Mamtora, “A Survey on Association Rule Mining in Market Basket Analysis,” *Int. J. Inf. Comput. Technol.*, vol. 4, no. 4, pp. 409–414, 2014.

[5] C. Rygielskia, J.-C. Wangb, and D. C. Yen, “Data Mining Techniques for Customer Relationship Management,” *J. Phys. Conf. Ser.*, vol. 910, no. 1, pp. 483–502, 2002.

[6] H. Aguinis, L. E. Forcum, and H. Joo, “Using Market Basket Analysis in Management Research,” *J. Manage.*, vol. 39, no. 7, pp. 1799–1824, Nov. 2013.

[7] A. Hibino and Y. Niwa, “Graphical representation of nuclear incidents/accidents by associating network in nuclear technical communication,” *J. Nucl. Sci. Technol.*, vol. 45, no. 5, pp. 369–377, 2008.

[8] S.-C. Hsieh, J.-N. Lai, C.-F. Lee, F.-C. Hu, W.-L. Tseng, and J.-D. Wang, “The prescribing of Chinese herbal products in Taiwan: a cross-sectional analysis of the national health insurance reimbursement database,” *Pharmacoepidemiol. Drug Saf.*, vol. 17, no. 6, pp. 609–619, Jun. 2008.

[9] Y. Kanagawa, S. Matsumoto, S. Koike, and T. Imamura, “Association analysis of food allergens,” *Pediatr. Allergy Immunol.*, vol. 20, no. 4, pp. 347–352, Jun. 2009.

[10] R. Yang, J. Tang, and M. Kafatos, “Improved associated conditions in rapid intensifications of tropical cyclones,” *Geophys. Res. Lett.*, vol. 34, no. 20, p. L20807, Oct. 2007.

[11] G. J. Russell and A. Petersen, “Analysis of cross category dependence in market basket selection,” *J. Retail.*, vol. 76, no. 3, pp. 367–392, 2000.

[12] M. Svetina and J. Zupančič, “How to Increase Sales in Retail with Market Basket Analysis,” *SYSTEMS INTEGRATION.* pp. 418–428, 2005.

[13] W. X. Xie, H. N. Qi, and M. L. Huang, “Market basket analysis based on text segmentation and association rule mining,” in *Proceedings of the 1st International Conference on Networking and Distributed Computing, ICNDC 2010*, 2010, pp. 309–313.

[14] R.-Q. Liu, Y.-C. Lee, and H.-L. Mu, “Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from Distribution Big Data of Korean Retailing Company,” *Knowl. Manag. Res.*, vol. 19, no. 4, pp. 59–76, 2018.

[15] A. A. Hidayat, A. Rahman, R. M. Wangi, R. J. Abidin, R. S. Fuadi, and W. Budiawan, "Implementation and comparison analysis of apriori and fp-growth algorithm performance to determine market basket analysis in Breiliant shop," *J. Phys. Conf. Ser.*, vol. 1402, p. 077031, Dec. 2019.

[16] D. Abdullah *et al.*, "Data Mining to Determine Correlation of Purchasing Cosmetics With A priori Method," *J. Phys. Conf. Ser.*, vol. 1361, p. 012056, Nov. 2019.

[17] I. Surjandari and A. C. Seruni, "DESIGN OF PRODUCT PLACEMENT LAYOUT IN RETAIL SHOP USING MARKET BASKET ANALYSIS," *MAKARA Technol. Ser.*, vol. 9, no. 2, Oct. 2010.

[18] V. Kavitha and C. Issac Davanbu, "MARKET BASKET ANALYSIS USING FP GROWTH AND APRIORI ALGORITHM: A CASE STUDY OF MUMBAI RETAIL STORE," *BVIMSR's J. Manag. Res.*, pp. 56–63, 2016.

[19] W. F. Abbas, N. D. Ahmad, and N. B. Zaini, "Discovering purchasing pattern of sport items using market basket analysis," in *Proceedings - 2013 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2013*, 2013, pp. 120–125.

[20] H. Kaur and K. Singh, "Market Basket Analysis of Sports Store using Association Rules," *Int. J. Recent Trends Electr. Electron. Engg*, vol. 3, no. 1, pp. 81–85, 2013.

[21] S. K. Das and D. G. S. Lall, "Traditional marketing VS digital marketing: An analysis," *Int. J. Commer. Manag. Res.*, vol. 2, no. 8, pp. 05–11, 2016.

[22] J. F. Hair, D. E. Harrison, and J. J. Risher, "Marketing Research in The 21ST Century: Opportunities And Challenges," *Brazilian J. Mark.*, vol. 17, no. 5, 2018.

[23] H. D. Lasswell, "The structure and function of communication in society ," 1948.

[24] E. Katz and Lazersfeld, "Personal Influence," *Glencoe, Free Press*, 1955.

[25] D. L. Hoffman and T. P. Novak, "A New Marketing Paradigm for Electronic Commerce," *Glencoe, Free Press*, vol. 13, no. 1, pp. 43–54, 1997.

[26] R. D. Todor, "Blending traditional and digital marketing," *Bull. Transilv. Univ. Braşov Ser. V Econ. Sci.*, vol. 9, no. 58, 2016.

[27] S. H. Liao, C. M. Chen, C. L. Hsieh, and S. C. Hsiao, "Mining information users' knowledge for one-to-one marketing on information appliance," *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 4967–4979, 2009.

[28] W. J. Hauser, "Marketing analytics: the evolution of marketing research in the twenty-first century," *Direct Mark. An Int. J.*, 2007.

[29] A. M. Hormozi and S. Giles, "Information Systems Management Data Mining: A Competitive Weapon for Banking and Retail Industries," *Inf. Syst. Manag.*, vol. 21, no. 2, pp. 62–71, 2004.

[30] P. F. Nunes and J. Merrihue, "The Continuing Power of Mass Advertising," *MIT Sloan Manag. Rev.*, vol. 48, no. 2, pp. 63–71, 2007.

[31] J. Wind and V. Mahajan, "Digital Marketing," *Symphonya. Emerg. Issues*

Manag., no. 1, pp. 43–54, 2002.

[32] B. M. Ramageri and B. L. Desai, “ROLE OF DATA MINING IN RETAIL SECTOR,” *Int. J. Comput. Sci. Eng.*, 2013.

[33] M. Hemalatha, “Market basket analysis - A data mining application in Indian retailing,” *Int. J. Bus. Inf. Syst.*, vol. 10, no. 1, pp. 109–129, May 2012.

[34] M. Wedel and W. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*. Springer Science & Business Media, 2000.

[35] B. Sabitha, N. G. Bhuvaneswari Amma, G. Annapoorani, and P. Balasubramanian, “Implementation of Data Mining Techniques to Perform Market Analysis,” *Int. J. Innov. Res. Comput. Commun. Eng. (An ISO*, vol. 3297, no. 11, 2014.

[36] L. Gordon, “Leading practices in market basket analysis: How top retailers are using market basket analysis to win margin and market share.” Factpoint Group, 2008.

[37] D. Solnet, Y. Boztug, and S. Dolnicar, “An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue,” *Int. J. Hosp. Manag.*, vol. 56, pp. 119–125, Jul. 2016.

[38] R. Moodley, F. Chiclana, F. Caraffini, and J. Carter, “A product-centric data mining algorithm for targeted promotions,” *J. Retail. Consum. Serv.*, 2019.

[39] R. J. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, “Mining Business Databases,” *Commun. ACM*, vol. 39, no. 11, pp. 42–48, 1996.

[40] L. Yongmei and G. Yong, “Application in market basket analysis based on FP-growth algorithm,” in *2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009*, 2009, vol. 4, pp. 112–115.

[41] M. M. Mostafa, “Knowledge discovery of hidden consumer purchase behaviour: A market basket analysis,” *Int. J. Data Anal. Tech. Strateg.*, vol. 7, no. 4, pp. 384–405, 2015.

[42] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 1993, pp. 207–216.

[43] M.-C. Chen, A.-L. Chiu, and H.-H. Chang, “Mining changes in customer behavior in retail marketing,” *Expert Syst. Appl.*, vol. 28, no. 4, pp. 773–781, 2005.

[44] B. Ramasubbareddy, A. Govardhan, and A. Ramamohanreddy, “Mining positive and negative association rules,” in *ICCSE 2010 - 5th International Conference on Computer Science and Education, Final Program and Book of Abstracts*, 2010, pp. 1403–1406.

[45] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” *VLDB ’94 Proc. 20th Int. Conf. Very Large Data Bases*, 1994.

[46] A. N. Sagin and B. Ayvaz, “Determination of Association Rules with Market Basket Analysis: Application in the Retail Sector,” *Southeast Eur. J. Soft*

Comput., vol. 7, no. 1, May 2018.

[47] W. Y. Chiang, “To mine association rules of customer values via a data mining procedure with improved model: An empirical case study,” *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1716–1722, Mar. 2011.

[48] C. Borgelt and R. Kruse, “Induction of Association Rules: Apriori Implementation,” in *Compstat*, Physica-Verlag HD, 2002, pp. 395–400.

[49] J. Han, J. Pei, Y. Yin, and R. Mao, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach,” *Data Min. Knowl. Discov.*, vol. 8, no. 1, pp. 53–87, Jan. 2004.

[50] M. J. Zaki, “Generating Non-Redundant Association Rules,” in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 34–43.

[51] C. Borgelt, “An implementation of the FP-growth algorithm,” in *In Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations (OSDM ’05)*, 2005, pp. 1–5.

[52] C. Borgelt, “Frequent item set mining,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 6, pp. 437–456, Nov. 2012.

[53] B. Yildiz and B. Ergenç, “Comparison of two association rule mining algorithms without candidate generation,” in *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications, AIA 2010*, 2010, pp. 450–457.

[54] J. Vohra and E. Jyoti, “Data Mining Approach for Retail Knowledge Discovery,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 3, pp. 740–742, 2016.

[55] D. Peppers and M. Rogers, *The one to one future : building business relationships one customer at a time*. London: Piatkus, 1993.

[56] J. D. Wells, W. L. Fuerst, and J. Choobineh, “Managing information technology (IT) for one-to-one customer interaction,” *Inf. Manag.*, vol. 35, no. 1, pp. 53–62, 1999.

[57] D. R. Liu and Y. Y. Shih, “Integrating AHP and data mining for product recommendation based on customer lifetime value,” *Inf. Manag.*, vol. 42, no. 3, pp. 387–400, Mar. 2005.

[58] D. Birant, “Data Mining Using RFM Analysis,” in *Knowledge-oriented applications in data mining*, InTechOpen, 2011.

[59] J. R. Bult and T. Wansbeek, “Optimal Selection for Direct Mail,” *Mark. Sci.*, vol. 14, no. 4, pp. 378–394, Nov. 1995.

[60] G. Vani, M. Ganesh Babu, and N. Panchanatham, “Consumer Buying Behaviour The Controllables & Uncontrollables,” *Int. J. Exclus. Manag. Res.*, vol. 1, no. 1, pp. 1–12, 2011.

[61] A. Mohamed and N. Ramya, “Factors affecting consumer buying behavior,” vol. 2, no. 10, pp. 76–80, 2016.

[62] R. Kumar, “Impact of Demographic Factors on Consumer Behaviour - A

Consumer Behaviour Survey in Himachal Pradesh," *Glob. J. Enterp. Inf. Syst.*, vol. 6, no. 2, p. 35, Jul. 2014.

[63] S. Bakshi, "IMPACT OF GENDER ON CONSUMER PURCHASE BEHAVIOUR," *J. Res. Commer. Manag.*, vol. 1, no. 9, pp. 1–8, 2012.

[64] N. B. Gajjar, "Factors Affecting Consumer Behavior," *Int. J. Res. Humanit. Soc. Sci.*, vol. 1, no. 2, pp. 10–15, 2013.

[65] M. V. Roodpishi and R. A. Nashtaei, "Market basket analysis in insurance industry," *Manag. Sci. Lett.*, vol. 5, no. 4, pp. 393–400, 2015.

[66] S. C. Tan and J. P. S. Lau, "Time series clustering: A superior alternative for Market Basket Analysis," in *Lecture Notes in Electrical Engineering*, 2014, vol. 285 LNEE, pp. 241–248.

[67] B. Shen and K. Bissell, "Social Media, Social Me: A Content Analysis of Beauty Companies' Use of Facebook in Marketing and Branding," *J. Promot. Manag.*, vol. 19, no. 5, pp. 629–651, Nov. 2013.

[68] A. Łopaciuk and M. Łoboda, "Global Beauty Industry Trends In the 21st Century," in *Management, knowledge and learning international conference*, 2013, pp. 19–21.

[69] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[70] G. Wilson and B. Ozuem, *Competitive social media marketing strategies*. IGI Global, 2016.

[71] G. Kane, M. Alavi, J. Labianca, and S. Borgatti, "Integrating social networks and information systems: A review and framework for research," *MIS Q.*, vol. 38, no. 1, pp. 275–304, 2014.

[72] F. Gao, S. Cui, and V. V Agrawal, "The Effect of Multi-Channel and Omni-Channel Retailing on Physical Stores," 2018.

[73] D. Grewal, R. Krishnan, and J. Lindsey-Mullikin, "Building Store Loyalty Through Service Strategies," *J. Relatsh. Mark.*, vol. 7, no. 4, pp. 341–358, Dec. 2008.

[74] G. J. Browne, J. R. Durrett, and J. C. Wetherbe, "Consumer reactions toward clicks and bricks: investigating buying behaviour on-line and at stores," *Behav. Inf. Technol.*, vol. 23, no. 4, pp. 237–245, Jul. 2004.

[75] A. Rashad Yazdanifard, "The Review of Physical Store Factors That Influence Impulsive Buying Behavior," *Int. J. Manag. Account. Econ. Vol.*, vol. 2, no. 9, pp. 1048–1054, 2015.

[76] E. Biyalogorsky and P. Naik, "Clicks and mortar: The effect of on-line activities on off-line sales," *Mark. Lett.*, vol. 14, no. 1, pp. 21–32, Feb. 2003.

[77] M. R. Wick and P. J. Wagner, "Using market basket analysis to integrate and motivate topics in discrete structures," in *Proceedings of the Thirty-Seventh SIGCSE Technical Symposium on Computer Science Education*, 2007, pp. 323–327.

[78] N. Cerpa and L. Kenneth, “Support vs Confidence in Association Rule Algorithms.” Proceedings of the OPTIMA Conference, Curicó, Chile, 2001.

[79] R. Bali and D. Sarkar, *R machine learning by example*. Packt Publishing, LDA., 2016.