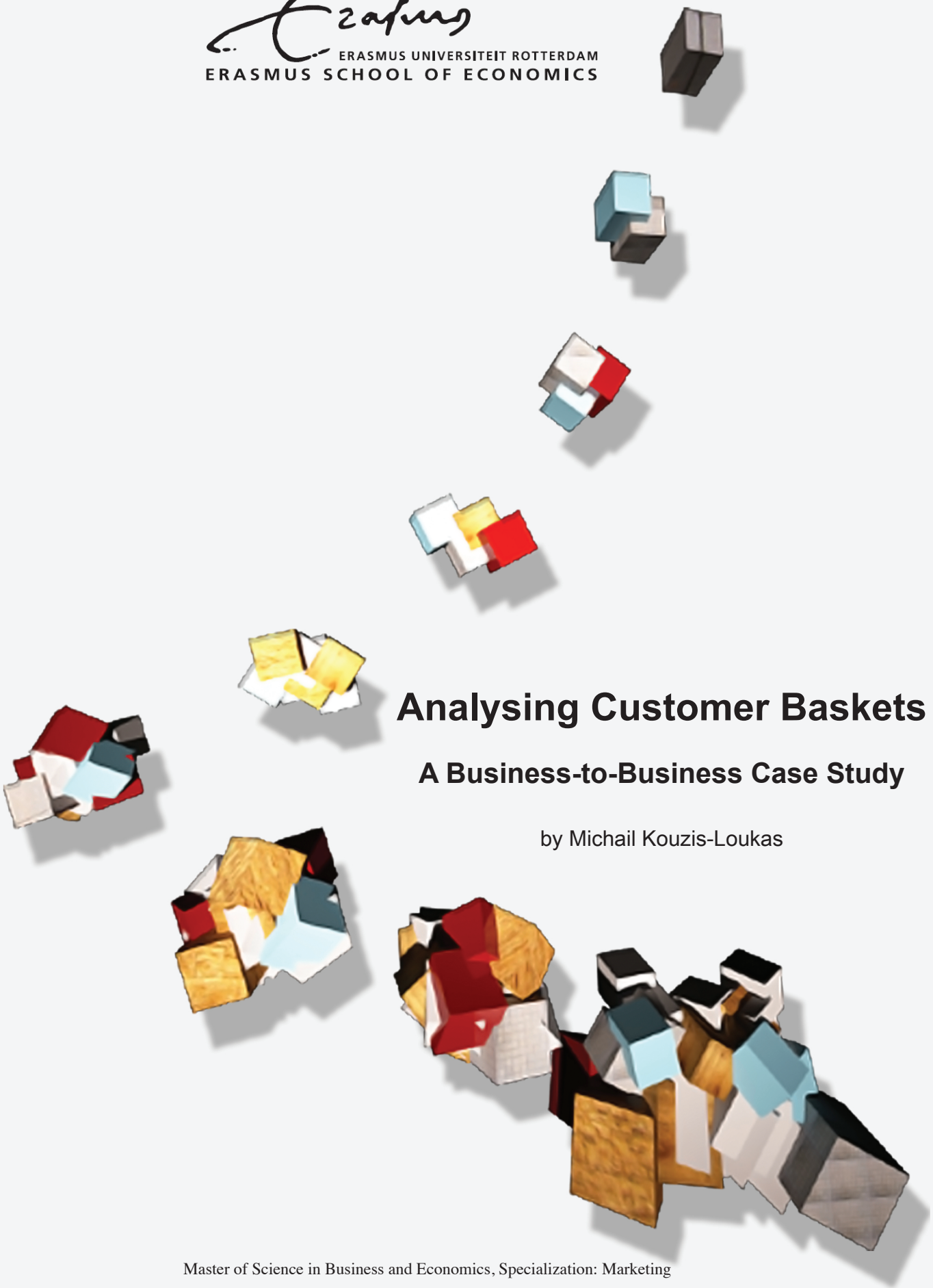
ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS



Analysing Customer Baskets

A Business-to-Business Case Study

by Michail Kouzis-Loukas

Master of Science in Business and Economics, Specialization: Marketing
Erasmus School of Economics
Supervisor: MSc Bruno Jacobs
Student Number: 383314
Date: 9.2014

Acknowledgements

First of all, I would like to thank my supervisor, Bruno Jacobs, for his indispensable guidance and help on the theoretical and technical issues that arose during the composition of this thesis and of course for giving me the flexibility to build on my own ideas and inspirations.

I would also like to express my many thanks to my brother Dimitrios Kouzis-Loukas. His insightful comments gave important depth to this dissertation.

At this point I would also like to thank the thesis team members and all of my colleagues and university friends, Dimos, Maria, Sophia and many more. Without them, composing this thesis and the overall experience would be no fun at all.

I would also like to thank my dear friends Nikos, Orestis, Panagiotis and Spyros for their continuous moral support.

Last and most importantly, I would like to thank my parents for being supportive with my life decisions.

Abstract

This research is an attempt to provide with a robust way to analyse customer baskets in a business to business environment. The methodology can be generalized to many similar cases and can be extended in various different directions depending on the needs.

We present transformation techniques for vectors of independent variables and databases by using Python scripts. We then study models with two independent variables that are more accurate and reliable.

Three different software packages, Microsoft Office Excel, IBM SPSS Modeler and IBM SPSS Statistics have been recruited in order to obtain our final results. We chose the Apriori algorithm for basket analysis and binary logistic regression as our main analysis tools.

The results are straightforward and lead to the acceptance of our hypothesis confirming our initial intuitions. Products included in frequently purchased itemsets are usually coming from the same or complementary product divisions. Customer orders include more frequently purchased itemsets during the summer. This indicates a significant seasonality on the probability of an order to contain a frequent itemset. Last, we prove the significant negative effect of product price, positive effect of purchased quantity and promotional discount rate on the probability of a product being part of a frequent itemset. This way we construct a model that connects order and product attributes with the cross-selling opportunity as discovered with our association rule mining.

This research contributes to the body of business-to-business literature by presenting an approach for discovering high quality cross-selling candidates by analysing available customer basket records. The techniques we applied and the managerial insights derived are useful for large wholesalers who can use them to grow their customer base and build stronger customer relationships.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 OVERVIEW AND STUDY MOTIVATION

21st century is the era of innovative technological breakthroughs and computer science. Technology and computer-based applications became part of our everyday life and this brought profound changes on our understanding of the world and the way we make decisions. Vast amounts of data are generated every minute and companies have long realized the potential of utilizing them. Modern companies strive to study and understand both their internal and external environment by mining their database "goldmines" to extract knowledge which will help them maximise profitability and optimize operations. Data-mining and statistical methods have been recruited by corporations of all scales in order to make the most out of their data.

Academic research contributed by providing the analytical tools and methodologies in order to make it possible for companies and organizations to extract valuable information from their databases. Various data-mining techniques were developed based on the implementation specifications, desirable outcomes and technical constraints. One specific category of algorithms particularly interesting for our purposes are those algorithms that help us study customer purchasing behaviour and identify purchasing patterns in order to optimize selling techniques, marketing actions and strategies. The anticipated benefits are so significant that companies are forced to restructure their operations in a way that embraces data-mining techniques.

Data-mining techniques were early adopted by virtually all classes of retailers and are widely affecting their relationship with their customers. Widening adoption of the Internet and online retailing as a sales channel gave online companies instant access to massive and highly accurate databases. Data-mining innovators gained competitive advantage, competition became fiercer and this quickly led to the adoption of data-driven marketing as the main paradigm. Companies on business-to-consumers (B2C) markets were the first to realize the value of data-mining and implement their practices. Business-to-business (B2B) companies followed but still data-driven marketing adoption in B2B environments remains thin. This is due to many reasons but the two most important are the following:

Data-mining dates back to the early nineties when the first theories concerning data-mining were developed (for more details skip to the literature review section). In business there is a lag between the time a theory is shaped and the time real-world testing and widespread adoption takes place. Communication between academics and business is relatively slow and this is one of the main reasons literature lacks case studies of data-mining applied in business-to-business environments.

Even more important is the fact that due to significant vendor-customer relationships, academic literature and companies were mostly focused on different approaches for developing these relationships. One could characterize B2B relationships as a very constrained type of the ones found in a B2C setting, mainly involving contractual deals and seasonal projects. This created a significant amount of certainty and less need for high reflexes comparing to B2C companies. Nevertheless, understanding customers' needs in order to develop relationships plays a key role in B2B settings as well. As executives of B2B companies started to realize the potentials of increasing cross-selling, their multiplicative effect on profits and the close relationship between cross-selling rates and data-

mining findings, they started pushing sales directors and marketing managers to incorporate data-mining techniques into the standardized back-office and marketing processes. This also happened due to the tremendous success on B2C environments.

Identifying the key components of customer buying behaviour is one of the main goals of data-mining. This can be achieved by extracting rules that associate products based on their purchase frequency as pairs, triplets, quadruplets etc. This way, businesses can shade light on cross-selling opportunities that aren't obvious otherwise. Discovering association rules among products can also provide improved layouts for catalogues and commercial web sites. Understanding how product attributes (such as price) and other characteristics such as time of purchase and product quantities purchased affect the probability of product belonging in a frequently purchased itemset, can significantly help in many ways. First, companies can optimize prices of products in order to achieve higher cross-selling rates. Additionally by focusing on high-margin products, it is possible to achieve high-revenue streams while using fewer resources (e.g. smaller sales force). Providing customers with the right levels of discounts might further increase cross-selling rates as well as limit losses from underpriced products. Understanding when is the right time to provide a cross-selling offer based on purchasing patterns can save time and money. It can also increase the odds that a product will be part of a cross-sell. Studying how purchasing quantities affect cross-selling opportunities can provide directions for product ordering. This can be used by on-line recommendation systems that provide customers cross-selling offers only when certain quantity thresholds are met. This is way more common in a B2B environment than a B2C.

Investigating the above in a business-to-business setting is the main goal of this research. For this endeavorment we will use a daily sales database for one of the leading wholesalers worldwide. This wholesaler specializes in energy product solutions and services for houses, businesses and factories such as energy control systems, solar panels, cooling systems and capacitors. Our study uses company's sales database for the entire Greek market for 2013.

1.2 RESEARCH STRUCTURE

This study is consisted of six main chapters:

First we review the literature foundations within three sections. The first section refers to the drivers that determine the relationship between businesses and customers in a business-to-business setting. Building on that, we extensively describe how businesses rely on cross-selling techniques in order to expand and strengthen their relationships with customers as well as the main categories of cross-selling techniques. Next, we introduce the reader to the main data mining concepts and processes.

Second, we present the research questions and hypothesis for this research as well as the conceptual frameworks for each of these questions.

On the next part, a thorough description of the databases used and the data preparation process is presented. In this part we also explain the product taxonomy for this specific case study and the aggregation patterns used.

In chapter 5 we explain in-depth all the data-mining techniques used (from database transformation to Apriori implementation and basic notations) and statistical methods we applied in this research.

At the end of this chapter we examine some additional statistics used to investigate multicollinearity.

In the next chapter we present our analysis. It consists of three parts corresponding to our three research questions. In the first part we analyse our Apriori algorithm results and whether the itemsets we mine consist of products from the same or complementary product categories. Then we investigate the seasonality effect on the probability of a customer order to include a frequently purchased itemset. At the end, after transforming two of our independent variables into logarithmic scales and checking for possible multicollinearity issues, we explore the relationship between product promotional discount, purchased quantity and price per piece and the probability that a product is part of a frequent itemset.

Our findings, conclusions as well as research limitations and future directions are enlisted in the 6th and last chapter.

2. LITERATURE FOUNDATIONS

2.1 DEVELOPING BUSINESS-TO-BUSINESS RELATIONSHIPS

Consumer goods literature is very rich of studies on the key drivers underpinning the relationship between customers and retailers. By studying areas ranging from consumer psychology and behaviour on the individual level to the more macro-level theories of customer and consumer dynamics of the masses, marketing scholars have provided modern businesses and organizations with a wide range of tools and theory on how to manage their customers in a strategic and efficient, way. Customer relationships management and consumer marketing in a business-to-customer (B2C) environment provides the foundations on which initial attempts of managing relationships in a business-to-business (B2B) setting were made. Their approach tries to maximize profit throughout customer development and acquisition as classic theory dictated.

Pete Naude and Christopher Holland claim this is an over-simplistic, adversarial and short-term approach which results in the implementation of common practices such as the manipulation of certain marketing-mix variables (i.e. the 4Ps) with an aim to maximize the markets' return (Naude P. & Holland C., 1996). This approach was not able to elicit substantial gains mainly due to the short-term orientation which contradicts the long-term perspective of most B2B relationships as well as the lack of acknowledgment of the complexity of B2B relationships which are vastly different to the B2C ones (Naude P. & Holland C., 1996).

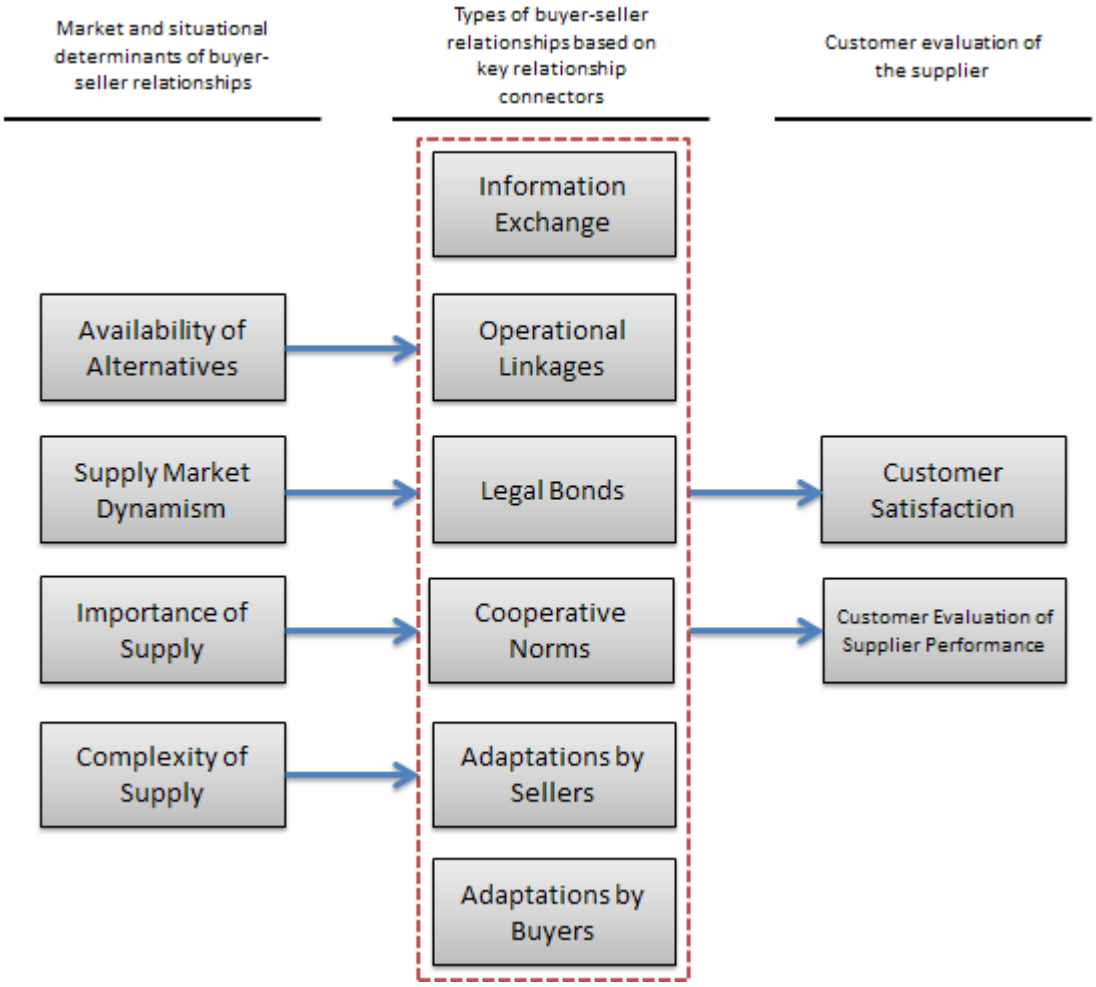
Scanzoni (1979) claims that relationships evolve through five general phases: awareness, exploration, expansion, commitment and dissolution. Each phase represents different ways of how both parties regard one another and interact. This research focuses on the phase of expansion, where interdependence is boosted between exchange partners (Dwyer, Schurr, & Oh, 1987), by looking into customer basket analysis techniques. The later contribute to the continual growth of business strategies for market penetration and product development (Ansoff 1957) helping, at the same time, to develop the commitment phase where relational continuity is pledged (Dwyer et al., 1987).

Productive and enduring relationships between business suppliers and customers are the main focus of innovative managers in a global scale nowadays according to (Cannon & Jr., 1999). Many manufacturing firms are decreasing the number of business partners they maintain but empower the relationships with the remaining ones (Emshwiller 1991). This is also necessary because of just-in-time inventory systems and computerized order placements (Anderson and Narus 1990; Frazier, Spekman, and O'Neal 1988).

New insights on the drivers of B2B partners' interrelations such as the way trust and commitment, uncertainty and dependence affects the characteristics of the relationship as well as the way these influence key performance outcomes, has suggested that it is necessary to characterize this relationship in various different ways (Cannon & Jr., 1999). Consequently, businesses can be related or connected with formal contracts or trusting agreements, openly share information and implement common communication systems or choose to disclose information, have a shared sense of cooperation or be totally independent (Cannon & Jr., 1999). These connectors can form multivariate relationship profiles and it is assumed that they are not necessarily correlated with one another (Cannon & Jr., 1999). As the schematic overview of key constructs relevant to the practice of buyer-seller relationships suggests (figure 2.1), there are six relationship connectors which constitute the way business buyers and sellers interrelate: information exchange, operational linkages, legal bonds, cooperative norms, and relationship-specific adaptations by buyers and sellers

(Cannon & Jr., 1999). Market and situational factors reflect key conditions in which relationships form; outcomes of a relationship is customer evaluations of the supplier which break down into customer satisfaction and customer evaluation of supplier performance.

Figure 2.1: Schematic Overview of Key Constructs Relevant to the practice of Buyer-Seller Relationships



The above conceptual framework described by (Cannon & Jr., 1999) consitutes the “tangible” and more measurable sector of B2B relationships, the core structure and foundations on which other, more “intangible” aspects of firms’ interrelations are build upon. Many authors studied these “intangible” aspects from the perspective of long-term orientation that business-to-business relationships must always incorporate. (Oliver C., 1990; Pfeffer and Salancik, 1978) suggested that there two main motivators exist that forces firms to enter relationships with their suppliers: minimizing uncertainty and the need to increase dependence. (Lages, Lancastre, & Lages, 2008) introduced the B2B Relationship Performance Scale and Scorecard as a way to bring relationship marketing into business-to-business practice. This Relationship Performance Scale suggested that relationship performance planning, implementation, and control can be achieved by using a scale of five distinct, yet related, dimensions: relationship orientation, relationship commitment, trust, mutual cooperation and relationship satisfaction (Naude P. & Holland C., 1996).

A very important study by (Ganesan, 1994) sade light on the determinants of long-term orientation in B2B relationships by describing it as a function of two main factors: mutual dependence and the extent of established trust. Dependence and trust reduce environmental uncertainty, facilitate transaction-specific investments and increase satisfaction between business partners. Credibility and

benevolence, as the main ingredients of trust, help to reduce the perception of risk associated with opportunistic behaviors by vendors, increases confidence of the retailers that short-term inequities will be resolved over a long period and reduces the transaction costs in exchange relationships according to (Williamson, 1975, 1981). Finally all the above play a very key role in increasing the willingness to rely on a specific vendor to whom a retailer has confidence in and results from perceived expertise, reliability and intentionality (Moorman, Zaltman & Deshpande 1992).

Based on the above, it can be infered that building relationships with existing customers is more important than just acquiring new customers when a common customer development strategy is formed by organizations in a B2B long-term-oriented setting. As (Ford, 1980) also suggests, there must be a distinction between the "strategic management" of relationships and "operational management" of a single relationship mainly because strategic management covers a portfolio of relationships and involves the assessment of any one relationship within the company's strategy in a particular market or markets.

Moreover, expanding already existing customers' "share of wallet" can vastly improve profitability and customer value; this requires a proper customer management as well as the right selling techniques. As (Knox, 1998) mentions, the implementation strategies may differ from market to market. The most common strategy though, is to build loyalty amongst preferred customers which in turn allows a more effective alignment of the organisation's resources and skill base. Another effective technique is cross-selling of multiple products or services that enhances customer retention rates. This is mainly due to the increase of already high switching costs in a B2B setting (lock-in effect), contributing at the same time in the development of a dependence relationship amongst the different business parties, as also described in the previous paragraphs.

In the next section we will mainly focus on the cross-selling literature related to the B2B environment.

2.2 CROSS-SELLING AND CUSTOMER RELATIONSHIPS

As discussed in the previous section, cross-selling can be an effective solution for improving profitability by contributing to the development of a stronger relationship between the two business partners. Hereunder we will take a closer look at cross-selling as a technique, its positive and negative implications and how we can facilitate cross-selling by utilizing information technology.

For the purpose of our research, we will make use of (Schmitz, 2012) definition of cross-selling:

> "cross-selling is a customer management process that involves the sale of additional products or services that are not the same and that can be related or unrelated to those that a customer has purchased or declared a desire to buy previously"

This definition is representative because it incorporates the concept of scrutinizing information on past transactions which might not have been initiated by a specific customer we might be studying. These transactions might be related to purchases made by other customers and as a result can be identified as possible cross-selling opportunities for the current customer. Overall, every customer has a cross-buying potential. The degree to which this potential is utilized is called *customer cross-selling performance* (Schmitz, 2012).

Cross-selling needs to be distinguished from the concept of up-selling which involves increasing the purchasing volume by selling more units of the same purchased product or upgrading into a more

expensive version of the purchased product (Kamakura, 2008). There are five main precepts that characterize the understanding of cross-selling techniques according to (Schmitz, 2012):

1) The product sold might be a physical product or a service or a bandle including combinations of those
2) cross-sold products are items which have not been purchased nor are the same as previously purchased items
3) it is sufficient of a customer to only have indicated previous intent or desire to buy an item in order for this item to be considered to have a cross-selling potential
4) the potential cross-sold products might be related or unrelated to the originally purchased and their orizination might be either from a third party or the company itself
5) products and services can be cross-sold together at the same time or successively.

Overall, (Kamakura, 2008) underlines that there are three main drivers for firms to incorporate cross-selling. First, they desire to increase the customer's share of wallet, secondly they aim to broaden the relationship scope and third, to increase customer retention rates.

There has been imminent evidence from business applications as well as a wide range of studies that "offering current customers with additional products or services that can provide added value for them" (as cross-selling was defined in (Zboja & Hartline, 2012)) is connected with increased sales, greater levels of customer loyalty and higher overall customer spending. (Lynn, 1999) highlight the role of selling agents as solution providers whose goal is to cross-serve customers, creating as many exit barriers as possible, setting it at the same time impossible to establish similar relationships with other suppliers (Lynn, 1999).

More specifically, (Coyles & Gokey, 2005) found that potential customer value can be improved as much as ten times when marketing strategies focus on increasing customer share of wallet. Furthermore, (Malms & Schmitz, 2011) pointed out the importance of leveraging customer realtionships by cross-selling, a selling technique which is described as a low-risk initial investment tool. High customer switching costs, lower customer churn rates, and leveraged distribution systems can be established by this realtionship expansion.

Based on (Kamakura et al., 2003), "increasing the number of products a customer uses from three to four product lines doubles the firm's profitability". (Weese, 1997) reports that a one-product relationship equals to a 10% retantion rate in a 5-year timeframe, a two-product relationship increases retantion rate to the level of 45% and a more-than-three-products relationship increases the chance of retaining a customer up to 80%. (Reichheld & Sasser, 1990) demonstrated that profits can be increased by 25% to 85% with a 5% increase in customer retention rates. Adding to that, as (Gupta, Lehmann, & Stuart, 2014) mentions, profits in contractual settings increase over the customer's lifetime with a 1% improvement in margins (such as from cross-selling) resulting in a 1% increase in customer value.

Cross-selling from the side of the vendor and cross-buying from the side of the customer can provide important benefits to both sides. According to (Tuli et al., 2007 and Kumar et al., 2008), cross-buying can reduce the number of suppliers a customer has, substantially decrease the total cost of buying, provide an increased buying convenience, increase the purchased volumes from each vendor which can also be translated into higher rebates and bonuses. (Akçura, Özdemir, & Altınkemer, 2009) also identify that cross-selling strategies provide customers with lower prices and vendors with significant strategic advantages while simultaneously raising customer satisfaction as market gets broader. (Weese, 1997) mentions low price perceptions as one of the drivers for cross-buying; (Netessine, Savin, & Xiao, 2006) describe the perceived fairness of a package as a "fair trade" as long

as its price does not exceed the sum of components' price. From the vendor's perspective, cross-selling to already existing customers can cost five times less than acquiring and serving a new customer (Rothfeder, 2003), response rates from cross-selling efforts are up to 5 times higher than cold sales (Andrews, 1999), cross-selling not only contributes to the increasing of customer "share of wallet" but also leads to increased firm "share of mind" (Kamakura, 2008), it increases the actual and psychological switching-costs thus improving retention rates (Kamakura et al. 2003), firms gain an information advantage and deeper knowledge concerning the customer's needs and preferences improving the former's ability to successfully market their products and services.

On the other hand, realizing the potential of cross-selling technique is not trivial and often fails to show expected or even positive results. If a firm's salesperson is highly motivated to engage into cross-selling activities, she or he may make so many offers that customers become annoyed and put the salesperson's expertise and reliability under question (Malms & Schmitz, 2011). (DeGabrielle 2007) reports that cross-selling initiatives fail with a rate of 70% or higher. Research on the German financial market has shown that finacnial service providers exploit 33% of their cross-selling potentials (Homburg & Schäfer, 2001). A relevant survey has shown that 75% of German bank managers are unsatisfied with the success of their cross-selling actions. Another factor of decreasing rates of cross-selling activities especially for niche markets is the high costs of customization required to reach additional customers (Akçura et al., 2009) while firm targeting broader markets do not seem to face that issue.

Moreover and most importantly according to (Erich & James, 1987), there seems to be three key reasons why companies do not engage into cross-selling: 1) simplistic thinking 2) organization inertia and 3) compromise costs. Simplistic thinking refers to companies' false mentality that one salesperson should sell additional products. This strategy is successful only when it is supported by lead generation schemes, training and supervision of the sales people. Organization inertia stands for the necessity of innovative organization structure which is a key to successful cross-selling marketing programs. Furthermore, many companies adopt vertical, product-defined organizational structures which on the one hand facilitate accountability, on the other hand, since each product-defined profit center is charged with achieving high performance, there is little incentive to accommodate sales communication across profit centers. (Erich & James, 1987) define that as the unwillingess to compromise (or compromise costs) which produces less than optimum results for the companies.

(Kamakura, 2008) is also tackling the issue of compromise costs which is described here as a "silo mentality where employees associated with one product line do not feel responsible for other lines, regardless of the fact that they might all be serving the same customer" which results in a product-centric structure that constricts profitability. This cross-divisional orientation (Malms & Schmitz, 2011) (CDO) is necessary for the realization of the cross-selling potential and the number of product divisions a salespersons sells, thus influencing salesperson's ability and motivation to engage in cross-selling acts which finally sets the cross-selling process as successful (Malms & Schmitz, 2011). (Malms & Schmitz, 2011) study the relationship between CDO and cross-selling and prove that a CDO has a positive effect on cross-selling success which is also important for this current study.

In the next three paragraphs a less parsimonious approach is going to be attempted to cross-selling from the perspective of information technology management. This will be detrimental in the unfolding of the literature review as we move into part three and a more detailed analysis on customer baskets and data mining methods.

Increasing competition in modern global B2B markets intensified the pressure to target for more efficient and effective marketing efforts, leading firms to seek ways to utilize this critical function

(Cannon & Jr., 1999). Fast technological improvements started to massively invade the business environment. This fact combined with advanced business practices and more complex economic conditions, highlighted the need to deal with marketing more as an information-handling issue (Naude P. & Holland C., 1996, Cannon & Jr., 1999). Advances in information technology, especially the incorporation of the Internet in the internal business processes and external interactions with customers create incentives and motivate a firm to engage into cross-selling activities. Based on (Akcura & Srinivasan, 2014), "successful cross-selling requires customer intimacy and detailed information on customer demographics and preferences"; once such information is available, data can be leveraged and identification of cross-selling opportunities is possible (Ansari & Mela 2003, Kamakura et al. 2003). (Milne and Boza 1999) also provide with evidence that advanced information management practices affect purchasing volumes. The more information customers reveal the higher cross-selling revenues a firm can obtain while gaining a competitive advantage of achieving lower prices (Akcura & Srinivasan, 2014). Identifying customers who are likely to cross-buy has become a very critical issue for both B2B and B2C companies of all sizes and across different industries.

Based on (Kamakura, 2008), there are two main groups of analytical tools that make cross-selling possible in a customer relationship management context: Acquisition Pattern Analysis and Collaborative Filtering. Acquisition Pattern Analysis studies the customer's patterns of previous acquisitions as well as other customers' respective patterns in order to identify what the former's next logical purchasing step is. On the other hand, collaborative filtering focuses on "association patterns among purchases across customers to identify suggestions of other items that would go along with the purchased one" (Kamakura, 2008).

Applications of collaborative filtering for item recommendations are also widely used on the Internet (Kamakura, 2008). The most popular machine learning technique for item recommendations is that of extracting sets of association rules describing the most common associations among products and services. (Netessine, Savin & Xiao, 2006) argued that in the on-line markets, effectiveness of cross-selling recommendations is increasing when items are bundled at a discount. (Netessine, Savin & Xiao, 2006) added that this must always take into account the risk of future stock-outs due to the increased purchasing volumes and might set as more profitable an option of selling the bundled products separately.

After analysing the concepts of customer relationship management in a B2B environment along with the key role of cross-selling as a selling technique, the next section is dedicated to big data, customer basket analysis and association rules which are used in our study.

2.3 MINING THE BIG DATA

Within the last 15 years, massive amounts of data get generated on daily basis with almost 90% of the data existing today created within the last two years. Google CEO, Eric Scmit remarked that "there was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days, and the pace is increasing", exponentially one could possible add. Based on (Manyika J., et. al., 2011), the term "Big Data" refers to datasets the size of which exceeds the abilities of typical database software tools to capture, store, manage and analyse. Leveraging Big Data analytics in order to achieve maximum utilization of firms' "information assets" can be considered as a matter of gaining an "unfair advantage" over competitors, increasing organization's chance to be among leaders within their industry by 20% (Teradata Labs, Silicon Valley). (Xu & Walton 2005) also describe acquisition of customer knowledge as the main strategic tool for gaining competitive advantage in the modern globalized markets. After acquiring and storing

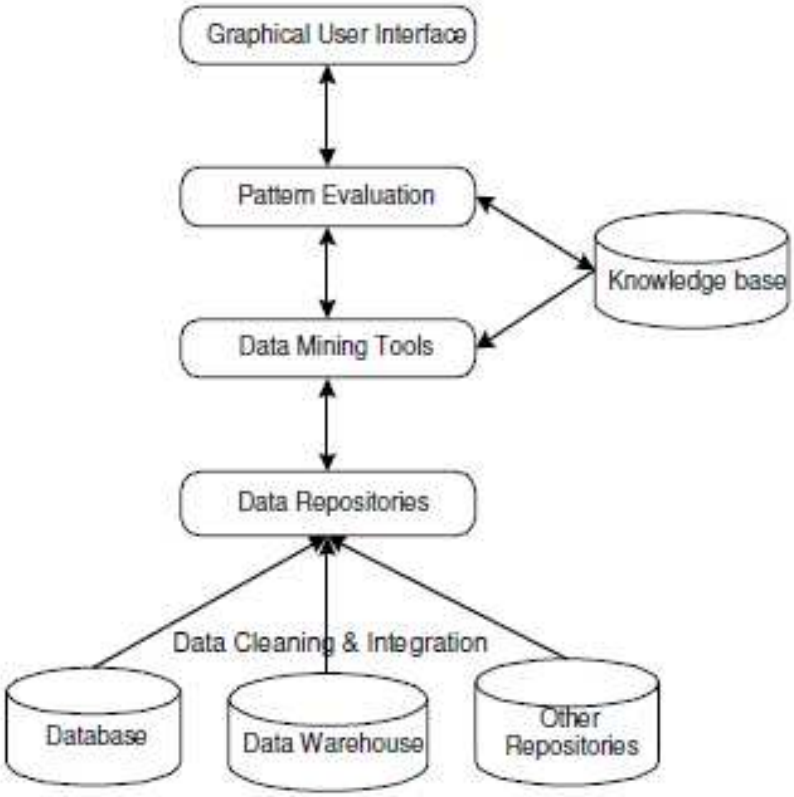
customer knowledge in large information repositories such as relational databases, data warehouses etc, data mining techniques can be applied in order to extract meaningful insights with strategic importance.

Data mining is defined by (Chen et al 1996) as the process of extracting interesting information or patterns from large information repositories and is frequently considered synonymous to concepts such as Knowledge Discovery in Database (KDD) and machine-learning. There is no doubt that machine learning practitioners borrow some data mining techniques to train their systems on performing specific tasks but yet, the core difference with data mining is that the former is used to discover previously unknown properties of the data while the latter is mainly focused on prediction based on known properties extracted from the data at hand. Furthermore, based on (Han & Kamber 2000), data mining is the core process of KDD's three stages:

1) preprocessing which includes data cleaning, integration, selection and transformation
2) data mining process where various algorithms are applied to uncover hidden knowledge
3) postprocessing where evaluation and presentation of the mining results is taking place

More specifically, (Zhao & Bhowmick, 2003) present a more detailed conceptual framework of these three processes where first the data gathered from different sources is cleaned, integrated and stored in data repositories, and then the data which is most relevant to our task is selected and transformed into a format that is ready to be mined. Last but not least, various data mining techniques are applied, results are evaluated and approved based on certain rules and are displayed as raw data, 3D graphics, decision trees etc (figure 2.2).

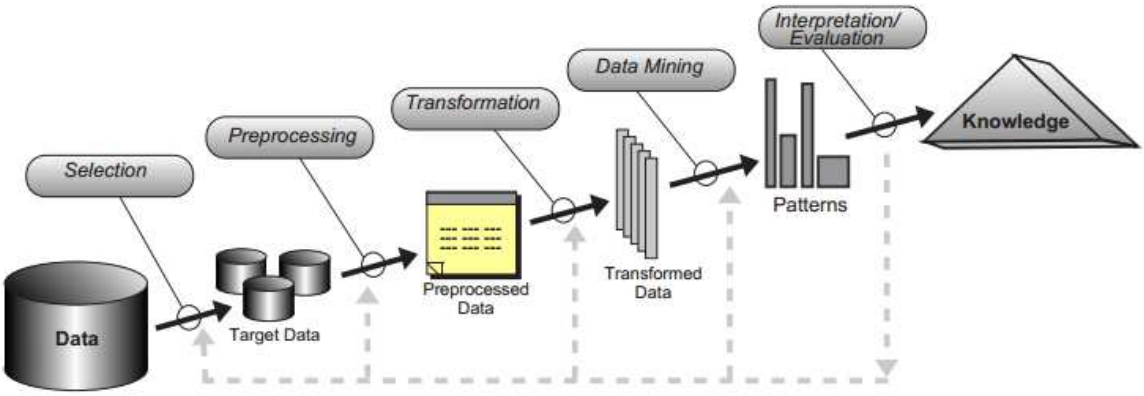
Figure 2.2: Knowledge Discovery in Database processes



Source: (Zhao & Bhowmick, 2003)

Fayyad and Shapiro, pioneers in data mining, describe KDD process as iterative and interactive including nine discrete stages and many user decision-making for each step (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). These stages involve identifying the goal of KDD by looking into previous knowledge and domain applications, creating a target data set, cleaning and pre-processing the data, reducing and projecting the data, matching KDD goals with data mining methods, selecting the data mining algorithms, adopting the right set of representational forms to facilitate pattern searching, interpreting patterns discovered and finally consolidating, documenting or reporting the knowledge produced (figure 2.3).

Figure 2.3: An overview of the steps comprising the KDD process



Source: [Fayyad et al., 1996]

(Zhao & Bhowmick, 2003) identify three types of data mining techniques; association rule mining (ARM), classification and clustering. Association rule mining was first proposed by (Agrawal et al., 1993) and deals with extracting casual structures, correlations, associations and frequent patterns among itemsets (products, services etc) in transactional databases or other data sources. Classification refers to the mapping of a data item or a class of objects into several predefined classes based on common attributes or characteristics so as to predict the classification of objects whose class is unknown. Clustering is similar to classification with the difference that there are no predefined classes and objects are grouped together based on their similarities. (Fayyad et al., 1996) also add to the data mining literature by describing techniques such as regression, summarization, dependency modeling, change and deviation detection, link analysis and sequence analysis. For the purpose of this research we will mainly focus on association rule mining for customer baskets and more specifically on the discovery of frequent itemsets whose items have high correlations also known as frequent itemsets mining (FIM) (Lee, Park, & Moon, 2013).

Marketing literature classifies methods for analysing market baskets into *explanatory* and *exploratory*. Explanatory models aim to identify and quantify causal relationships between choices, marketing variables and product attributes. Exploratory methods, on the other hand, do not include consumer-related information and marketing mix variables. Their goal is the summarisation of vast amount of data into fewer rules and measures by uncovering complex cross-category interdependency structures and relations. Exploratory methods are not appropriate to predict future consumption or perform root-cause analyses. They are rather computational simple methods, such as association rules generation, useful to reveal unknown relationships between items in a database such as patterns in the purchase frequency of items or among different product categories (Mild & Reutterer, 2003). Such exploratory methods involve association rule generation for customer baskets, collaborative filtering etc.

A market basket consists of customer purchased itemsets during a single shopping occasion. Market basket data analysis refers to the methodological toolbox used to study the composition of market baskets or bundles in order to extract meaningful marketing insights. Such methodological toolbox involves association rule mining and frequent itemsets mining which have found wide application in customer basket analysis of on- and off- line retailers who are interested in forming micro-marketing strategies and derive targeted cross-selling programs. Recommendation systems make use of association rules to predict the customers' preferences and provide them with customized offers based on their past preferences or preferences of people who bought similar products. (Han et al., 2000) define an association rule (AR) as "a probabilistic rule that if a set of data attributes occur together, then some other disjoint set of attributes is also likely to occur". In the same article it is stated as a fundamental challenge of ARM that a data set which includes N attributes, has 2^N-1 candidate patterns. Collaborative filtering, first introduced by Goldberg and colleagues in 1992, is one of the most famous set of algorithms used by on-line retailers such as Amazon.com, E-bay.com etc. in order to fine-tune their offerings and facilitate cross-buying behaviour.

As sales are becoming more and more preformed on-line, intensified data collection, urgency of analysis of different types of data, increasing computational complexity as well as increasing demand for computational power due to exponential growth of data attracted the interest of marketing researchers, computer scientists and programmers providing them with new challenges concerning the development of advanced association rule algorithms dealing with these issues. For that reason, many association rule mining algorithms have been developed such as Apriori, AprioriTid by (Agrawal & Ramakrishnan, 1994), AprioriHybrid and AIS by (Agrawal et al., 1993), SETM by (Houtsma & Swami, 1993), GSP, PrefixSpan, SPADE and ISM by (Cheng H., Han J., 2004), AOG by (Cheung, Fu & Han, 1996), Count Distribution(CD), Data Distribution(DD), Intelligent Data Distribution(IDD) by (Han, Karypis, Kumar, 2000), Frequent Pattern Growth by (Han et al., 2000) etc.

The main differences among the above association rules have to do with the computational strength and speed, performance and execution. (Gantz J. & Reinsel D., 2011) mention the existence of a new generation of technologies and architectures, designed to economically extract valuable insights from very large data volumes. "General purpose sensemaking" is one of these latest emerging architectures and data mining techniques where new transactions and observations integrate with previous transactions and permits the system users to "do something about whatever is happening while it is still happening" (Cavoukian A. & Jonas J., 2012). A significant volume of research recommends that trying and comparing different data mining techniques can significantly improve the efficiency of association rule implementation since different mining results might tackle the weaknesses coming along with these association rules.

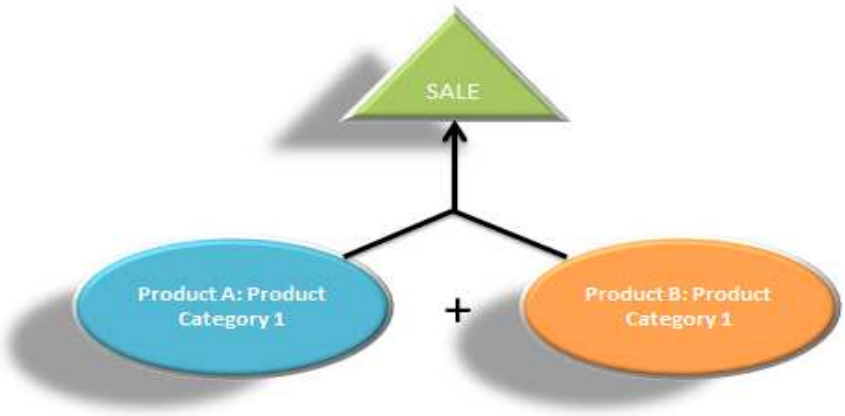
3. RESEARCH QUESTIONS AND CONCEPTUAL FRAMEWORKS

Based on the literature foundations, a B2B relationship is mainly build upon trust, commitment, mutual dependence, credibility, benevolence and customer satisfaction. These basic elements are playing a vital role in cultivating a long-term relationship between the two parts and help develop the relationship throughout its various levels. For a B2B vendor, expanding the customer relationship, in the sense of increasing the cross-selling opportunities, is of vital importance and can leverage profitability. In our effort to understand the drivers of cross-selling opportunities in a B2B environment, we shall investigate the relationship between products purchased together in frequently purchased itemsets. Thus, the 1st hypothesis of this research accrues:

Research Question no 1: Which product categories are being sold together more frequently in a business-to-business setting?

- H1: Products of the same product category have high affinity to be purchased together.

Figure 3.1: Conceptual Framework for Hypothesis 1



- H2: Products across complementary product categories have high affinity to be purchased together.

Figure 3.2: Conceptual Framework for Hypothesis 2

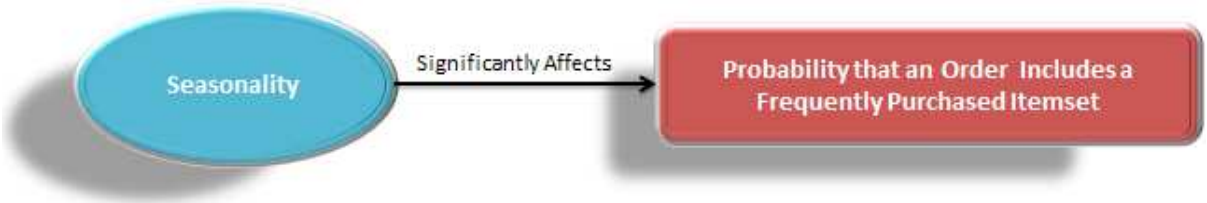
It is important to point out that seasonality might have a significant effect on sales of products. After identifying and comparing product categories which are parts of frequently purchased itemsets, it is interesting to investigate whether seasonality plays an important role in the inclusion of a product in

a frequently purchased itemset. Since seasonality is directly connected to sales orders and frequent itemsets consist of products frequently purchased together, the second research question is formed accordingly:

Research Question no 2: Does seasonality influence the probability that an order will include a frequently purchased itemset?

- H3: The seasonality effect on the probability that an order will include a frequently purchased itemset is significantly different than zero across different months of the year.

Figure 3.3: Conceptual Framework for Hypothesis 3



The final step in order to understand and expand the list of drivers of cross-selling opportunities in a B2B environment is to investigate the relationship between product attributes and frequently purchased itemsets. Building on that, the 3rd research question is formed:

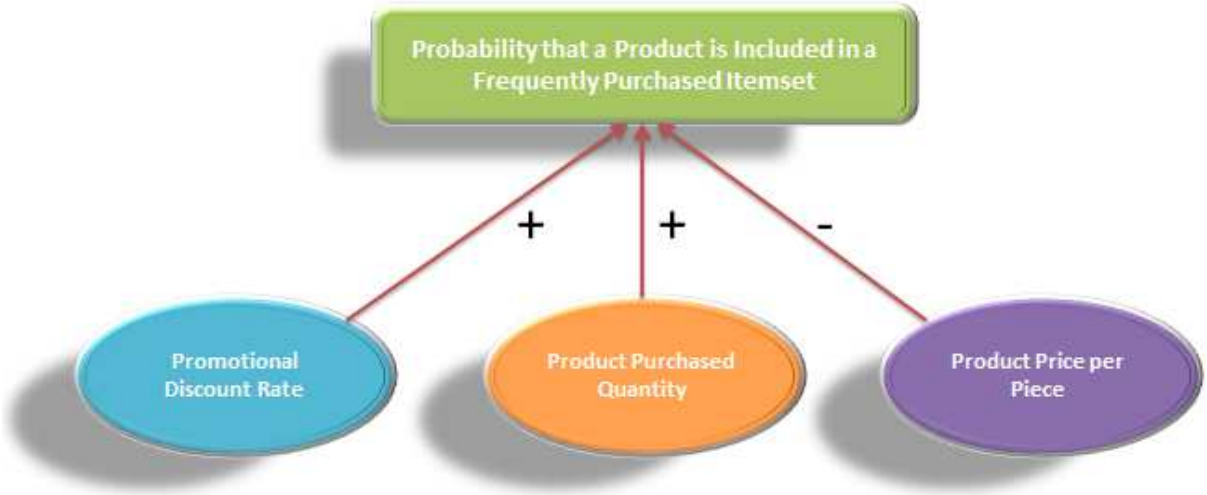
Research Question no 3: Does promotional discount rate, initial catalogue price per piece and product purchased quantity have significant effect on the probability of a product to be included in a frequent itemset in a business-to-business setting?

- H4: Discount rate positively influences the probability that a product will be included in a frequently purchased itemset.

- H5: Initial catalogue price negatively influences the probability that a product will be included in a frequently purchased itemset.

- H6: Product purchased quantity (also referred as order quantity) positively influences the probability that a product will be included in a frequently purchased itemset.

Figure 3.4: Conceptual Framework for Hypothesis 4, 5 and 6



In order to proceed to the necessary analyses to answer our research questions and relevant hypothesis, we will group our efforts into three different phases:

- The first phase involves the generation of frequently purchased itemsets with the implementation of Apriori algorithm on the tabular database with customer sales.
- The second phase identifies the seasonal effect on the probability of a customer order to include a frequently purchased itemset.
- The third phase studies the effect of product attributes on the probability of a product to be part of a frequent itemset.

In order to answer the two last research questions we will need to create two different dependent variables, one for the orders that include frequently purchased itemsets and one for products that are part of a frequent itemset. In the first case we only have one independent variable which is the month of purchased order while in the second case we have three different independent variables: promotional discount rate, product purchased quantity and product price per piece. Product purchased quantity and product price per piece will be transformed into logarithmic scales for reasons that will be further described later on in this study.

Variables

- 1st Dependent Variable: Order includes a frequently purchased itemset (Binary; Values: 0-1)
- 1st Independent Variable: Month of sales (Categorical; Dummy)
- 2nd Dependent Variable: Product in order line is part of a frequently purchased itemset (Binary; Values: 0-1)
- 2nd Independent Variable: Discount rate (Ratio), natural logarithm of initial catalogue price per piece(Ratio), natural logarithm of product purchased quantity (Ratio)

Table 3.1 summarizes the hypotheses for all research question enlisted above:

Table 3.1: Hypothesis Review

Hypothesis
H1: Products of the same product category have high affinity to be purchased together.
H2: Products across complementary product categories have high affinity to be purchased together.
H3: The seasonality effect on the probability that a product will be included in a frequently purchased basket is significantly different than zero across different months of the year.
H4: Discount rate positively influences the probability that a product will be included in a frequently purchased basket.
H5: Initial catalogue price negatively influences the probability that a product will be included in a frequently purchased basket.
H6: Product purchased quantity positively influences the probability that a product will be included in a frequently purchased basket.

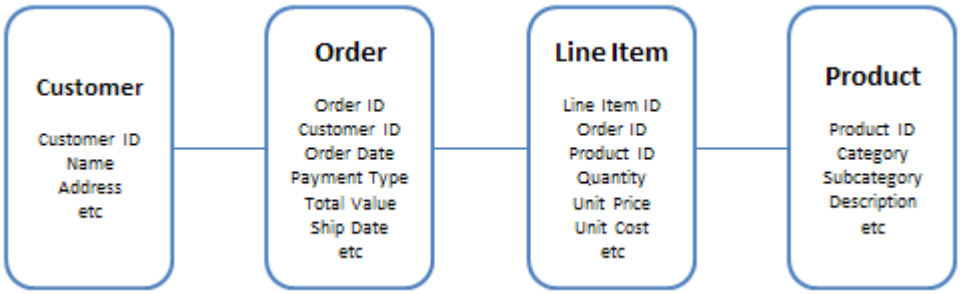
4. DATA

4.1 DATA DESCRIPTION AND DATA PREPARATION

The database that we use on our research was provided to us by one of the top suppliers of electrical equipment and energy solutions in Greece. This company is a wholesaler with a customer base of over 250 customers who are businesses and retailers located all over the country. The company retains a leading market position.

The initial raw database is of transactional format including 281 customer codes, 179,705 order lines for the financial year of 2013, 9,317 unique products from 45 product categories, 8 business units (definition in section 4.2) and 55 million Euros of total sales volume. Moreover, it combines in one Microsoft Office Excel sheet all the four fundamental tables a typical market basket database should consist of according to (Berry M. & Linoff G., 2004):

Figure 4.1.1: The Four Fundamental Tables for Basket Analysis



Source: (Berry M. and Linoff G., 2004)

Annual sales of 2013 for each customer and product on a daily basis was selected as the ideal dataset, taking into account the limited available computational capacity of the commercial personal computer used which was already pushed to its limits due to the computational complexity of Apriori algorithm and the limited memory capacity required to process Microsoft Excel files of this magnitude.

There are two main general database formats used in the association rule modeling; transactional and tabular databases. A reference to both database formats will be made since both where used to facilitate different analysing purposes. Transactional format database (also known as till-roll format) usually contains a separate record for each transaction or, in our case, different order lines for each product, connected to each other with a customer order ID:

Figure 4.1.2: Format of Transactional Database

Customer ID	Item
1	1
1	1
1	3
2	2
2	3
3	1

Tabular format database (also known as basket or truth-table data) contains items and flags for the presence (TRUE) or absence (FALSE) of each specific item, each line representing a complete set of associated items in a customer's basket:

Figure 4.1.3: Format of Tabular Database

Customer ID	Item 1	Item 2	Item 3
1	TRUE	FALSE	TRUE
2	FALSE	TRUE	TRUE
3	TRUE	FALSE	FALSE

Transactional database is converted into tabular format database (16,584 baskets, 11 products per basket on average) by using a Python script (refer to section 5, Methodology, for further details). The initial transactional database was used to create all types of conventional analyses, tables and charts as well as to run the second logistic regression of this research. Tabular database was used as an input in SPSS Modeler software package in order to create the association rules. It was also used to run the first binary logistic regression which has to do with studying the seasonality effect.

The transactional database was extracted from the company's SAP databases which is one of the most widely used ERP (Enterprise Resource Planning) software packages. Since some of the fields necessary for our research (e.g. product category per product) were not possible to be downloaded together with customer order lines in one excel spreadsheet, different excel files were downloaded and the target information was added in the initial sales Excel spreadsheet with the help of the appropriate Excel functions.

A variety of calculations needed to be done in order to offset missing data. For example, total order line value was divided by total order line quantity in order to create the catalogue price for each product unit for each order line. Moreover, products with catalogue price = 0 were removed. These were exceptional cases where, for example, gifts were given away to the customers for certain strategic marketing purposes. In total, our transactional database included the 16 fields presented below.

Figure 4.1.4: Transactional Database Fields

Fields in Transactional Database	
1. Order Creation Date	8. Product Line*
2. Creation Date (Month)	9. Product Description
3. Customer ID	10. Order Quantity
4. Customer Name	11. Unit of Measurment*
5. Sales Order ID	12. Price per Piece
6. Business Unit*	13. Total Catalogue Price*
7. Product Category*	14. Promotional Discount Rate*

Not self-explanatory terms are:
*6. Highest aggregated level of taxonomy (More details in section 3.2)
*7. Second highest aggregated level of taxonomy (More details in section 3.2)
*8. Least aggregated level of taxonomy based (More details in section 3.2)
*11. Pieces or meters, based on the type of product
*13. Customer order line total value
*14. Discount rate applied on total catalogue price during promotional activities

A 15th field named “Frequent Item” is added as a new column in the initial transactional database, based on the results provided by the association rule mining analysis. This is the binary dependent variable for the model for the third phase of this analysis. This constructed variable takes values 1 or 0 depending on whether the product is included in a frequently purchased customer itemset or not. A similar column was added as a last column in the tabular database, representing whether an order includes a frequently purchased itemset or not. This is the dependent variable for the second phase of our analysis.

The Microsoft Excel database extracted in the first phase of the analysis includes seven fields (Figure 4.1.5) whose definition and contribution will be further discussed in the next chapter.

Figure 4.1.5: Tabular Database Fields

Fields in Association Rule Mining Database
1. Antecedent 1
2. Antecedent 2
3. Antecedent 3
4. Consequent
5. Support %
6. Confidence
7. Lift

This database is further enriched with thirteen additional fields which are necessary for calculating the total expected profit of the association rule mining analysis (Figure 4.1.6).

Figure 4.1.6: Tabular Database Additional Fields

Additional Fields in Association Rule Mining Database
8. Basket Type*
9. Times Sold Together*
10. Times Sold Together (no Consequent)*
11. Price Consequent*
12. Price Antecedent 1*
13. Price Antecedent 2*
14. Price Antecedent 4*
15. Total Set Value*
16. Total Set Value (no Consequent)*
17. Total Set Value*Times Sold Together (no Consequent)
18. % Increase in Basket Value*
19. Profit*

Not self-explanatory terms are:

*8. Pair, Triplet, Quadruplet based on the number of Antecedents

*9. Times a frequently purchased itemset is purchased

*10. Times a frequently purchased itemset is purchased without the consequent

*11, 12, 13, 14. Price of each product within a frequently purchased itemset

*15. Total itemset value

*16. Total itemset value without the consequent

*18. Basket value increase: (Total Set Value/Total Set Value (no Consequent))-1]

*19. Potential profit: (Total Set Value*Times Sold Together (no Consequent)*% Increase in Basket Value)

4.2 PRODUCT TAXONOMY AND DATA AGGREGATION

Product taxonomy is a hierarchical classification of products based on specific and pre-defined characteristics. Their use is of significant importance as they provide with instant access to the right information in aggregated or disaggregated level, allowing enterprises to better utilize their information assets and organize their functions and processes around products and customers. Product taxonomies are widely used in information systems and more specifically in data mining techniques as a way to extract meaningful and interpretable insights from massive unstructured data.

In this research, there are three different levels of product hierarchy based on product aggregation. 9,317 unique products are classified into 45 distinct product categories based on their product description, general characteristics and application. Product categories are further classified into 8 super-categories, which will be called business units, representing the eight different markets the company operates in (Figure 4.2.1).

Figure 4.2.1: Product Taxonomy

Different types of results can be extracted by focusing on each hierarchical level while accuracy of the results decreases as analysis is moving up to higher levels. In addition, as the number of items in customer basket analysis increases, while moving down to more disaggregated levels of hierarchy, computational complexity grows exponentially. Therefore, the lowest hierarchy level, product level, was used in this customer basket analysis with an aim to achieve maximum accuracy and high-quality results in exchange for interpretational convenience and high computational capacity needs. The interpretational convenience issue is tackled later on in this analysis by combining information taken by the initial transactional database and the final results after applying the Apriori algorithm. More specifically, the products identified as frequently purchased together are matched with their initially assigned product categories and business units thus further facilitating a more aggregated level of analysis.

Based on a disclosure agreement, signed between the researcher and the company at the beginning of this research, all product categories, business units and product codes are renamed in order to ensure confidentiality. Each business unit is labelled randomly as BU 1, BU 2 etc and each product

category is labelled based on the business unit they belong to as PC 1.1 for a product category belonging to BU 1, PC 2.1 for BU2 etc as indicated in figure 4.2.2. Product codes, discovered throughout the rule mining phase of our research, are labelled randomly “Product 1”, “Product 2” and so on. Figure 4.2.3 describes the counts of unique products each product category consists of:

Figure 4.2.2: Taxonomy BUs vs PCs

BU 1	BU 2	BU 3	BU 4	BU 5	BU 6	BU 7	BU 8
PC 1.1	PC 2.1	PC 3.1	PC 4.1	PC 5.1	PC 6.1	PC 7.1	PC 8.1
PC 1.2	PC 2.2	PC 3.2	PC 4.2	PC 5.2	PC 6.2	PC 7.2	
PC 1.3		PC 3.3	PC 4.3	PC 5.3	PC 6.3	PC 7.3	
		PC 3.4	PC 4.4	PC 5.4	PC 6.4	PC 7.4	
		PC 3.5	PC 4.5	PC 5.5	PC 6.5	PC 7.5	
		PC 3.6	PC 4.6		PC 6.6		
		PC 3.7	PC 4.7		PC 6.7		
		PC 3.8			PC 6.8		
		PC 3.9			PC 6.9		
					PC 6.10		
					PC 6.11		
					PC 6.12		
					PC 6.13		

Figure 4.2.3: Taxonomy BUs vs Products

Product Category	Unique Products	Product Category	Unique Products	Product Category	Unique Products
PC 1.1	1	PC 4.2	5	PC 6.5	576
PC 1.2	138	PC 4.3	18	PC 6.6	359
PC 1.3	33	PC 4.4	56	PC 6.7	64
PC 2.1	2	PC 4.5	1	PC 6.8	665
PC 2.2	14	PC 4.6	17	PC 6.9	151
PC 3.1	58	PC 4.7	1	PC 6.10	42
PC 3.2	9	PC 5.1	1	PC 6.11	3
PC 3.3	35	PC 5.2	20	PC 6.12	605
PC 3.4	248	PC 5.3	1	PC 6.13	2
PC 3.5	207	PC 5.4	24	PC 7.1	1
PC 3.6	850	PC 5.5	18	PC 7.2	285
PC 3.7	1182	PC 6.1	101	PC 7.3	146
PC 3.8	7	PC 6.2	812	PC 7.4	922
PC 3.9	189	PC 6.3	991	PC 7.5	110
PC 4.1	24	PC 6.4	294	PC 8.1	29

5. METHODOLOGY

5.1 INTRODUCTION

For the purpose of this study, a three-step analysis was conducted. First we created all frequently purchased itemsets by applying the Apriori algorithm on our initial transactional database after transforming it into tabular format. Based on these results, we attempt to study the effect of seasonality by constructing our first binary dependent variable which represents whether an order includes a frequently purchased itemset or not. This dependent variable is created based on the tabular database and then, a binary logistic regression is carried out between this artificial binary dependent variable and the categorical independent variable: month of sales order. The third phase of our research involves the construction of our second binary dependent variable: whether a product is part of a frequently purchased itemset or not. In this part we study the effect of promotional discount rate, product purchased quantity and price per piece on the probability of this product to be part of a frequently purchased itemset. Successive binary logistic regressions are then carried out in order to identify the model with the best fit. In the next couple of chapters we are going to get deeper into detail, elaborating more on the methods used.

5.2 MARKET BASKET ANALYSIS AND ASSOCIATION RULES

As mentioned in chapter 2, Literature Foundations, market basket analysis (MBA) involves a whole set of methodological toolbox, part of which is association rule mining (ARM) and frequent itemset mining. The discovery of association rules and frequently-purchased-together patterns among products, included in customer baskets, provides marketers with a totally new perspective and different angle on how the products interrelate to each other. Completely different products/services and product categories merge with each other, creating a new approach of customer-oriented product mapping.

The usefulness and interestingness of an association rule, however, depends on contextual factors such as its application, level of aggregation and interpretability. For example an association rule like: “if a customer purchases a chicken burger, then that customer will also purchase a chicken burger box case” might not be as useful as an association rule like: “if a customer purchases a chicken burger, then that customer will also purchase a coke light” which might imply specific consumption patterns and customer preferences. In that direction, (Berry M. & Linoff G., 2004) have specified three different types of association rules: the actionable, trivial and inexplicable rules.

An actionable rule is the one providing with high-quality information concerning the relationship among products or services. This high-quality information can be utilized and easily translated into specific courses of action. In a B2B context, an actionable rule may lead a supplier to selectively provide customers with promotional discounts on specific bundles of products in order to boost cross-selling efficiency. Moreover, it can shade light on specific product outliers in the advertising decision process or be incorporated into the customer purchasing platforms as a recommendation providing system.

The importance of utilizing actionable rules is of such magnitude that they can be conceived as a powerful tool and a competitive advantage. This will be further supported with an example: an actionable rule has been discovered between product A and product B; when product A is

purchased, there is a high chance that product B is purchased as well. These two products are not directly connected to each other in a sense that they do not belong to the same product category or share common attributes but a specific behavioral pattern is connecting their sales. Product A's profitability is 4% while product B's profitability is 10% thus a significant difference between the two levels of profitability exists. If the company decides to, for example, bundle the two products together, "sacrifice" part of product A's profitability so as to provide with a lower bundle price compared to the sum of the products' individual prices, then there is a high chance that the extra profit generated by product B's boosted sales together with the discounted profit of product A, is outperforming the previously individual profits. This is one of the many selling techniques that demonstrate the importance of actionable rules which is also the main focus of this research.

The second and third category of association rules is trivial and inexplicable rules. As the name also implies, a trivial rule is a rule which is self-evidenced and already known by the company. In our case study for example, a rule that a grey plug is sold together with a grey plug socket is not adding any additional useful information. A seemingly actionable association rule might fall into this category of rules because it might be the result of previous marketing actions and the measurement of already acted-upon results. Trivial rules are useless rules except in one case, when impairments in the business operation, data collection and processing need to be identified. Last, inexplicable association rules are rules that are not interpretable, are of relatively low value and do not suggest any course of action.

In this research, actionable, trivial and inexplicable association rules have been identified with the guidance of the company's product managers and specialized customer correspondents.

5.3 TRANSACTIONAL TO TABULAR FORMAT: CONVERSION PROCESS

The main software package used for extracting association rules is IBM's SPSS Modeler, a software application used in data mining and text analytics that provides with a wide variety of robust analyses by implementing statistical and data mining algorithms without having to make use of any programming skills. SPSS Modeler input can be both transactional and tabular databases of various different sources. For convenience purposes, the initial transactional Microsoft Excel (.xlsx) database was used, as also described in chapter 4.1. All unnecessary fields were erased and only the two fields including customer order lines and respective product codes were kept. The final format of our file is also described in chapter 4.1.

Unfortunately, SPSS Modeler could not process the transactional database provided and after detailed investigation, no specific cause could be identified for the software's non-responding behavior. For that reason a decision was made to try out the alternative route of converting the transactional database to tabular format (described in chapter 4.1 which is also the most "natural" and easy-to-read database format for customer basket analysis). Since the available resources were very limited (relatively low processing power and memory capacity compared to the database size), the conversion could not be made by using Microsoft Excel which was systematically crashing down in efforts to do so. Consequently, a programming tool had to be conscripted: Python programming language.

Python is a high-level programming language with an abstract structure, making the programing process much easier and understandable. The syntax used in Python allows programmers to use fewer line codes than in other programming languages such as C and C++. It is considered as a

dynamic language which enables automatic memory management and is often used as a scripting language. Python also allows for reading files in CSV (comma-separated values) format which is a format used for exchanging and converting databases between different types of spreadsheet programs. For example, one line from the transactional Excel file has a default format of:

Table 5.3: Transactional Database Format

Column 1	Column 2	Column 3
Order 1	Product 1	Product 2

When stored as a .CSV file, the Excel's values change into comma-separated values [Order 1, Product 1, Product 2] and stored as plain-text. After the creation of the CSV file, the database can be opened and read by Python programming platforms such as ActivePython. ActivePython enables the user to write single modules that can manipulate and transform such data. For the purposes of this research the following Python code sequence was used:

Figure 5.3.1: Python Coding Sequence for Database Transformation

```
import csv
from collections import defaultdict

orders = defaultdict(set)
pructs = set()

first = True
with open('a.csv', 'rb') as csvfile:
    spamreader = csv.reader(csvfile, delimiter=',')
    for row in spamreader:
        if first:
            first = False
        else:
            products.add(row[1])
            orders[row[0]].add(row[1])

products = list(products)

with open('b.csv', 'wb') as csvfile:
    spamwriter = csv.writer(csvfile, delimiter=',',
quoting=csv.QUOTE_MINIMAL)
    spamwriter.writerow(['ORDER'] + products)
    for order in orders.keys():
        mo = orders[order]
        spamwriter.writerow([order] + ["1" if product in mo else "0" for
product in products])
```

The above code reads the CSV Microsoft Excel file, writes all product codes in one single pivot table, maps individual product reservations for each order line in the table, filling-in with ones [1] if product is included in the customer order or zeros [0] alternatively. Then it creates an output CSV file which includes all the data generated (further discussion on Python language and coding is out of the scope of this study). The output file is of tabular format and can be processed by both Microsoft Excel and SPSS Modeler. After using the newly generated file as an input, SPSS Modeler successfully provided us with the results for the first phase of our analysis: extracting the most frequently purchased itemsets.

5.4 MODELING ASSOCIATION RULES: BASIC CONCEPTS

In this section we are going to dive deeper into the details of the basic concepts for association rule mining processes applied in this research.

To define an association rule we first have to elaborate on the concepts of itemsets and transactions as found in our transformed database:

- As an itemset I we define a set of products with distinct k attributes, thus I= { I_1, I_2, …, I_k }. In this research, itemset I includes all products coming from the top row in our tabular database as described in chapter 4.1

- A transaction set t_i is a subset of itemset I including some or all items I_k and represents one customer order. In our database, each of the transactions t_i is represented by each different line and has the format of a binary vector, with t[m] = 1 if customer purchased the item I_m, and t[m] = 0 if customer did not purchase the item.

- The complete record of transactions t_i comprises the entire tabular database D.

- A transaction t_i "satisfies" a set of some items X from I if for all k items included in X, t[k] = 1, thus $X \subseteq I$ and $X \subseteq t_i$. Itemset X is defined as a basket and used hereunder to further define an association rule.

The most general type of an association rule has the following form: $X \rightarrow I_j$ (X and I_j as defined above and $X \cap I_j = \emptyset$). The implication of this rule is that if some of the k items from itemset I= {I_1, I_2, …, I_k} are included in a basket X then item I_j is likely to appear in the same basket as well. Itemset X and the items included in it are called antecedents and item I_j is called as consequent. The SPSS Modeler output used in that research is consisted of a maximum of three antecedents and one consequent, forming pairs (one antecedent-one consequent), triplets (two antecedents-one consequent) and quadruplets (three antecedents-one consequent) of items. An example taken from our database is that if a customer basket includes Product 24, Product 5 and Product 2 then it is also likely to include Product 1 (products are coded as described in chapter Product Taxonomy and Data Aggregation).

Two of the most basic concepts for finding interesting association rules and large or frequent itemsets are *support threshold* and *confidence threshold*. Both are used in order to exclude rules which are of low importance to the user and reduce the number of associations, provided by the association mining algorithm, to a manageable size. As a consequence, the user can predefine the desired levels of support and confidence thus dropping, in the early calculation stages, uninteresting item associations which are less likely to occur, reducing at the same time, the processing power required to produce the final results. These predefined support and confidence levels are called minimum support (or minsup) and minimum confidence (or minconf).

Support of an association rule is the percentage of transactions that contain itemset X and item I_j over the total number of transactions in a database D:

$$Support\ (X\ I_j) = \frac{\text{Number of transactions containing both X and } I_j}{\text{Total number of transactions in D}} \qquad (5.1)$$

In other words, support is the probability that a specific rule is encountered in a specific database and can also be considered as the statistical significance of an association rule. The algorithm applied to extract the association mining rules, increases the count for each item by one whenever this item

is confronted in a transaction t_i of database D during the scanning process. This process is indifferent to the quantities of items as it only takes into account the times that specific itemsets are found in the database. For example:

In a database of total 100 transactions, the following rule was found in 13 of the transactions: {Product A, Product B, Product C} → {Product D}. The support of this rule is 13/100= 0.13 or 13% and does not depend on the product purchased quantities.

A support value of 1 or 100% means that an association rule is found in all the transactions t_i while a support value of 0.01 or 0.1% means that this association rule is found in only a fraction of the transactions setting that rule as unimportant. Setting the right level of support as a minimum threshold for an association rule to be valid (thus interesting) is up to the user's discretion however, as the total number of transactions increases, the probability that an association rule above high minsup thresholds decreases. In addition, discovering association rules for expensive and not so frequently purchased items might still be interesting due to, for example, high margins contributed. As a result, setting a lower level of minimum support threshold is sometimes recommended. For this research a minimum support threshold of 1.0% and 1.5% was selected as also recommended for large databases in (Rajaraman, Leskovec, & Ullman, 2014).

Confidence of an association rule is the percentage of transactions that contain itemset X and item I_j over the number of transactions that contain itemset X:

$$Confidence\ (X \mid I_j) = \frac{\text{Support (X } I_j)}{\text{Support (X)}} \tag{5.2}$$

The importance of confidence level is quite significant as it measures the strength of an association rule. Confidence is used to measure the times a specific itemset is found together with a specific item out of the total times this specific itemset is found in the entire database. If the item is encountered together with the itemset all of the times (a confidence level of 100%), that means that the probability of finding this item is 100% in case the respective itemset is encountered. In other words, we can be 100% sure that a customer will buy product A if s/he also purchases a specific combination of products within itemset X. For example:

In a database of total 100 transactions, the itemset {Product A, Product B, Product C} was found 20 times in total. 13 out of 20 times was found together with {Product D} providing us with the rule: {Product A, Product B, Product C} → {Product D}. The confidence of this rule is 13/20= 0.65 or 65% meaning that this rule is relatively strong.

Setting a minimum level of confidence is also recommended as it contributes to the pruning of the results, reduction of the mining algorithm's calculation complexity and derivation of the most powerful association rules. For this research, a minimum confidence threshold was set to the level of 70%, excluding all rules below that level.

An additional measure of performance used in that research is that of the *lift* of an association rule. *Lift* of an association rule is defined as the confidence of the association rule over the unconditional probability of the consequent (also described as the support of item I_j):

$$Lift\ (X\ I_j) = \frac{Confidence\ (X \mid I_j)}{\text{Support } (I_j)} \tag{5.3}$$

Lift is used to describe how much more likely it is to find the association rule's consequent together with a specific itemset compared to the entire population of the database. Lift can take values between 0 and infinity and also makes an association rule more useful if it takes values greater than 1. High levels of lift mean that the consequent is scarcer within the population and more frequent within the specific itemset, setting the rule as more unique, interesting and useful for predicting the consequent in future uses. For example:

In a database of total 100 transactions, the itemset {Product A, Product B, Product C} was found 20 times in total, 13 out of which were together with {Product D} providing us with the rule: {Product A, Product B, Product C} → {Product D}. The confidence of this rule is 13/20= 0.65 or 65%. In addition, {Product D} was found 18 times within the 100 transactions; 18/100= 0.18 or 18%. The lift of this association rule is 65%/18%= 3.61 meaning that it is 3.62 times more probable to find Product D together with itemset {Product A, Product B, Product C} than in the entire population.

A lift of 2 implies that the consequent is found together with a specific itemset twice more often than expected while a lift of 1 is a sign that the occurrence probabilities of the antecedent and consequent are independent. Six rules had to be excluded from our database, regardless of their support and confidence levels, just because their lift was equal to 1.

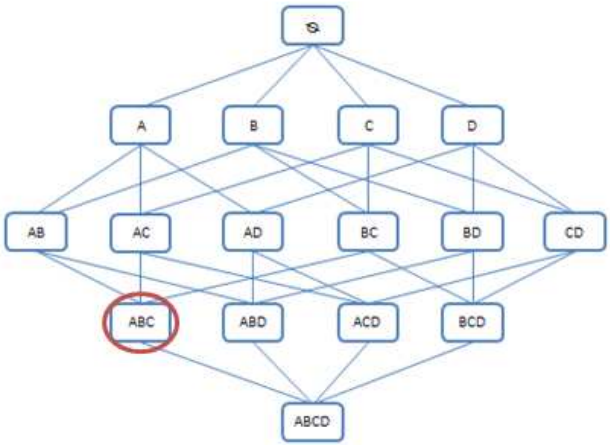
5.5 MODELING ASSOCIATION RULES: GENERATING FREQUENT ITEMSETS

After defining the basic notions of customer basket analysis and association rule mining, we can proceed to the next step of discovering frequent association rules. The purpose of association rule discovery is to mine rules that exceed or are equal to a pre-defined minimum level of support (rule support ≥ minsup) and confidence (rule confidence ≥ minconf). This goal is fulfilled in two steps:

- Step A: Generating Frequent Itemsets

The first phase of association rule mining is to discover the itemsets whose occurrence is greater than the predefined threshold (minimum support) from the given database. These itemsets are called frequent or large itemsets. This step can be broken down into two more sub-steps: A. Generating candidate large itemsets and B. Generating frequent itemsets. Candidate itemsets are those who are expected to exceed a minimum support level and frequent itemsets are those who actually do exceed or are equal to this threshold. The number of itemsets is increasing exponentially as items increase. More specifically, a database of k items can generate maximum 2^k-1 frequent itemsets (1 is subtracted for the null set) in case no support threshold has been specified.

Figure 5.5.1: Generating Itemsets (4 items)



As figure 5.5.1 illustrates, four items A, B, C and D can generate 2^4-1=15 different combinations of items. By selecting the minimum support levels virtually we reduce the itemsets selected by our algorithm based on their observed frequencies.

- Step B: Generating Strong Association Rules

The second phase of association rule mining is to generate association rules from the frequent itemsets provided by step A, based on a specific confidence threshold (minimum confidence). These association rules are called strong association rules. If there are k items included in a frequent itemset, 2^k-2 candidate association rules can be generated (2 is subtracted because rules with empty antecedents and consequents are omitted). For example, after identifying the frequent itemset {ABC} in figure 5.5.1, this itemset can generate 2^3-2=6 candidate association rules: {A,B}→{C}, {A,C}→{B}, {B,C}→{A}, {A}→{B,C}, {B}→{A,C} and {C}→{A,B}. These candidate association rules will be further pruned by the minimum confidence threshold, separating the strong association rules from the weak ones.

5.6 MODELING ASSOCIATION RULES: THE APRIORI ALGORITHM

Chapter 2.3 enlists the most common algorithms used for association rule mining. For the purpose of this research one of the most basic algorithms is going to be put in action: Apriori algorithm. This algorithm is part of the Apriori series algorithms and was first introduced by Agrawal in (Agrawal & Srikant, 1994). Apriori differed from other existing algorithms because it employed a different approach for generating candidate itemsets and pruning of the frequent ones. Apriori algorithm was applied with the help of IBM SPSS Modeler as already mentioned.

Apriori follows the rule called itemset monotonicity (Rajaraman et al., 2014) also known as Apriori Property. This rule presupposes that in order an itemset to be frequent, every subset of this itemset must be frequent as well. Consequently, the support of an itemset will never exceed the support of its subsets. A direct implication of monotonicity is that a frequent itemset, for example, a triplet is including three different frequent pairs but also three different frequent pairs can be included in an infrequent triplet. All infrequent itemsets are pruned as the algorithm executes multiple passes over the dataset until itemsets satisfying the minimum support condition reach the maximum possible number of item combinations. As already mentioned, in our research a maximum number of 4 items in an itemset was selected, providing us with frequent pairs, triplets and quadruplets.

Figure 5.6.1 illustrates the two main processing steps of mining strong frequent association rules as described in the previous section. It also describes in detail how these steps are applied within the Apriori algorithm framework.

Step A:

- ✓ Apriori scans the database and calculates the support S for each item.
- ✓ For all items with support S greater than minimum pre-defined support threshold, the frequent 1-itemsets L_1 are generated.
- ✓ k-frequent candidate itemsets are generated by using frequent itemsets L_{k-1} of size k-1 and joining their items to create item supersets.
- ✓ The support of each new k-frequent itemset is calculated. Itemsets with support ≤ minimum are pruned.

- ✓ This process is repeated until no other candidate itemsets can be generated.
- ✓ Then Apriori moves on to step B.

Step B:

- ✓ For each itemset L, all nonempty subsets are generated (all possible combinations of items).
- ✓ Confidence for each subset J is calculated
- ✓ All subsets with confidence greater than predefined minimum confidence are added to strong rules.

Figure 5.6.1: Apriori Algorithm, Conceptual Framework

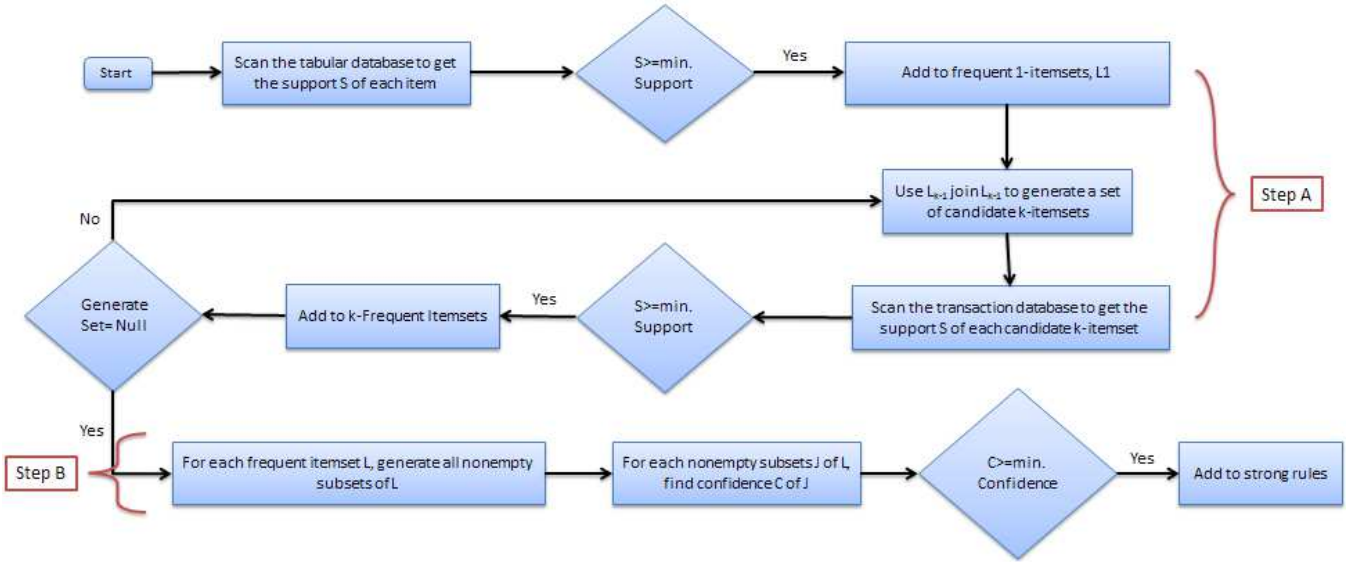
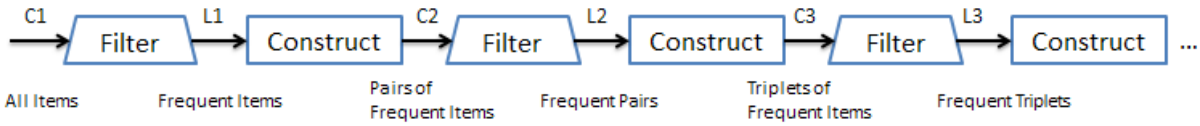


Figure 5.6.2 demonstrates how Apriori moves from candidate itemsets to frequent itemsets increasing, at the same time, the itemset size by 1 in every successive pass:

Figure 5.6.2: Apriori alternation between candidates C and frequent itemsets L



Last but not least, it is more than obvious that Apriori algorithm is a robust way to extract association rules. However, as one can see in the above description, there are two main drawbacks for Apriori. First is the relatively complicated candidate generation process that consumes a great amount of time, space and memory. Second is the multiple scanning of the database.

5.7 BINARY LOGISTIC REGRESSION

5.7.1 PRELIMINARY INFORMATION

Throughout the first phase of our research all frequently purchased itemsets were identified for both cases of 1.0% and 1.5% support thresholds with a minimum confidence level of 70%. A total of

332 association rules were generated in the former case while 1,164 rules were generated in the latter. 13 and 19 unique products were characterized as consequents respectively; based on these consequent items, the two dependent variables of the second model in our research is constructed.

Based on the association rule mining results, two different binary dependent variables were constructed. The first binary dependent variable was created in order to study the effect of seasonality on itemsets while the second binary dependent variable was created to investigate the relationship between product price, purchased quantity and discount rate on products found in frequently purchased baskets. The creation of two dependent variables was made because the relationship between month of sales and customer orders was considered to be more straightforward than the one between month of sales and products. Adding to that, studying seasonality of frequently purchased itemsets on product level would provide with qualitatively poor results since the model-generation process would weight higher the orders with more products than those with less. Moreover, the researcher decided that the seasonal effect would add too much noise in a model where eleven binary variables (twelve months minus one reference month) and three continuous variables (price per piece, purchased quantity and discount rate) would co-exist. This happens mainly due to the fact that regression in general, is considered to perform better when continuous, rather than binary, explanatory variables are used.

Building on the above, the first binary dependent variable was created so as to represent whether one customer order included a frequent basket or not. The second dependent variable demonstrated whether the product was found in a frequently purchased basket or not. In the former case, we made use of our tabular database, where each line represented a complete customer order. In the latter case, our initial transactional database was used. The structure of the transactional database facilitated the analysis on product level since it included each order's products in separate lines. The tabular database included 16,584 customer orders while the transactional database included 179,705 order lines. In the first case, an additional column was created at the end of the tabular database, taking values "1" or "0" if a customer order included a frequently purchased itemset or not respectively. The latter case introduced an additional column at the end of the transactional database taking values of "1" if a product was part of a frequently purchased itemset or "0" otherwise. To construct the two dependent variables, Python programming language was used. Python coding sequences can be found in Appendix C; explaining the meaning behind these sequences goes beyond the purpose of this research.

This process was carried out for both the results provided for support thresholds of 1.0% and 1.5% allowing us to compare the efficiency for each one of the models. 2,932 out of 16,584 orders were labelled with a 1 for a 1.0% support threshold and 2,682 out of 16,584 orders were labelled with a 1 for a 1.5% support threshold. A total of 27,552 out of 179,705 order lines and 20,486 out of 179,705 were labelled with a 1 for a 1.0% and 1.5% support threshold respectively, providing us with sufficient amount of cases to proceed to the next steps.

The second and third phase of our research necessitated the use of a statistical method called binary logistic regression or binary logit regression. Binary logistic regression is a probabilistic statistical classification model which describes the relationship between a binary categorical dependent variable and the predictor variables. By making use of the binary logistic regression, we manage to calculate the odds and probabilities of an outcome occurring (model output=1). Hereunder, we are going to get into more detail concerning the basic assumptions and notions for the binary logistic regression.

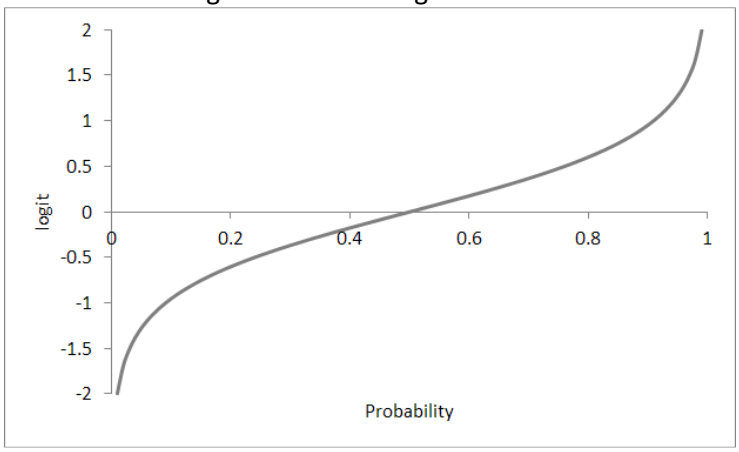
5.7.2 THEORETICAL BACKGROUND

Our dependent variable is a binary variable while our independent variables take any possible real value. The logit model maps the real range to the [0,1] range. More specifically, let's name our real variable π_i and the output variable $logit(\pi_i)$. The logit function:

$$logit(\pi_i) = log\frac{\pi_i}{1-\pi_i} \tag{5.4}$$

maps $\pi_i \in [0,1]$ to $logit(\pi_i) \in \mathbb{R}$ as presented in Figure 5.7.1.

Figure 5.7.1: The logit Function

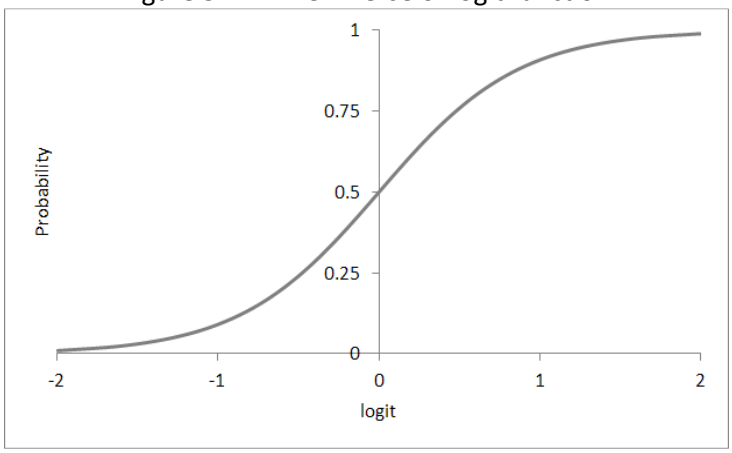


by solving for π_i we get:

$$\pi_i = \frac{\exp\{logit(\pi_i)\}}{1+\exp\{logit(\pi_i)\}} \tag{5.5}$$

The probability π_i as defined in equation 5.5 is the probability that our binary variable will take the value of 1 and it follows the logistic distribution shown in figure 5.7.2.

Figure 5.7.2: The inverse of logit function



As Figure 5.7.2 shows, if the logit takes values below zero, the probability is below 50% while if logit lies above zero, the probability is above 50%.

We use OLS (ordinary least squares) method to estimate the $logit(\pi_i)$ based on our input data. The model we fit is:

$$logit(\pi_i) = x_i'\beta \tag{5.6}$$

By combining equation 5.5 and 5.6 we have an expression of π_i as a function of our observations:

$$\pi_i = \frac{\exp\{x_i'\beta\}}{1 + \exp\{x_i'\beta\}} \tag{5.7}$$

Probability π_i as defined in equation 5.7 is the probability that a certain binary outcome will take the value of 1 and it follows a logistic distribution described in figure 5.7.2 based on the values of the explanatory variables and logit model. The probability that the binary outcome will be 0 is 1- π_i. The ratio $\frac{\pi_i}{1-\pi_i}$ is called odds ratio and is defined as the ratio of the probability of success over the probability of failure. The right part of equation 5.4 ($log\frac{\pi_i}{1-\pi_i}$) is called log-odds ratio.

By combining equations 5.4 and 5.6 we have an expression of log-odds and odds ratio as a function of our observations:

$$log\frac{\pi_i}{1-\pi_i} = x_i'\beta \rightarrow \frac{\pi_i}{1-\pi_i} = e^{x_i'\beta} \rightarrow Odds_i = e^{x_i'\beta} \tag{5.8}$$

OLS (ordinary least squares) estimation calculates the regression coefficients β in a binary logistic regression that minimize the log-likelihood function:

$$logL(\beta) = \sum\{y_i \log(\pi_i) + (n_i - y_i)\log(1-\pi_i)\} \tag{5.9}$$

We run this process by using the SPSS statistical package. Its functions calculate the β coefficients, odds ratios, standard errors and significance levels for each β coefficient. Moreover, SPSS calculates the -2LogLikelihood statistic which measures how well the model fits the training set.

We also check for multicollinearity by creating and analysing the Pearson's R correlations table of independent variables, in order to avoid inflated standard errors and type 2 errors (failing to reject the null hypothesis while it is false).

The relative quality of our statistical models will be measured and compared by making use of the Akaike Information Criterion (AIC) in order to select the best model. AIC formula is given below:

$$AIC = -2LogLikelihood + 2k \tag{5.10}$$

where k is the number of model's parameters plus the model's constant. AIC is a metric that "penalizes" models with large number of parameters (large k) since they may (if properly calibrated) overfit to the input data. On the other hand it rewards good fitting (-log (L)). In other words, given two models with the same in-sample predictive performance it will indicate as better (lower AIC) the one with the lowest number of parameters k. Given two models with the same number of parameters it will indicate as better (lower AIC) the one which "fits the data" best (i.e. lower log (L)).

Since AIC is not conclusive we also use classification rates to evaluate the performance of our models. The higher the number of correct classifications is, the higher the (in-sample) predictive performance of our model will be.

6. RESULTS AND ANALYSIS

6.1 PHASE ONE: BASKET ANALYSIS

The purpose of the first phase of our analysis is to identify all frequently purchased itemsets. This is achieved by transforming our initial transactional database into tabular format and applying Apriori algorithm for 1.0% and 1.5% support threshold levels and 70% for confidence. At the 1.0% support threshold, Apriori mined 1,164 association rules comprised of 4,130 products in total and 64 unique products. At the 1.5% support threshold, 332 association rules were constructed for 1,092 products in total and 41 unique products. Table A.1 in Appendix A demonstrates a sample of the final excel output. Table 6.1.1 and table 6.1.2 summarize the number of consequents, antecedents and unique products in each case per type of itemset (pair, triplet and quadruplet). As expected, more association rules were generated as we decreased the support threshold (3.5 times more association rules for a 0.5% support threshold decrease) and more unique products are added to the itemsets.

Table 6.1.1: No of Consequents, Antecedents per Itemset Type (1.0% Support)

In total	1164 Consequents 2966 Antecedents	19 Unique Consequents 64 Unique Antecedents
Pairs	23 Consequents 23 Antecedents	19 Unique Consequents 62 Unique Antecedents
Triplets	480 Consequents 960 Antecedents	19 Unique Consequents 64 Unique Antecedents
Quadruplets	661 Consequents 1983 Antecedents	19 Unique Consequents 64 Unique Antecedents

Table 6.1.2: No of Consequents, Antecedents per Itemset Type (1.5% Support)

In total	332 Consequents 760 Antecedents	13 Unique Consequents 38 Unique Antecedents
Pairs	19 Consequents 19 Antecedents	5 Unique Consequents 13 Unique Antecedents
Triplets	198 Consequents 396 Antecedents	10 Unique Consequents 34 Unique Antecedents
Quadruplets	115 Consequents 348 Antecedents	11 Unique Consequents 15 Unique Antecedents

Based on the above tables and figures 6.1.1 and 6.1.2, we can infer that the majority of the association rules were consisted of either 3 or 4 products (triplets, quadruplets) with the number of quadruplets increasing as we decrease the support threshold. This is happening due to the nature of the orders, since most of the customer order lines include 11 products in average. Thus, we expect that orders with fewer products are less frequent assuming normality in the shape of the frequency distribution of products per order as also indicated in Figure A.1, Appendix A. Adding to that, by decreasing the support threshold, we allow our model to consider more candidate rules as valid. As a result, by increasing the amount of valid rules, the chances that a basket is consisted of less than 2 or 3 products decreases.

Figure 6.1.1: Type of Itemset (1.0% Support)

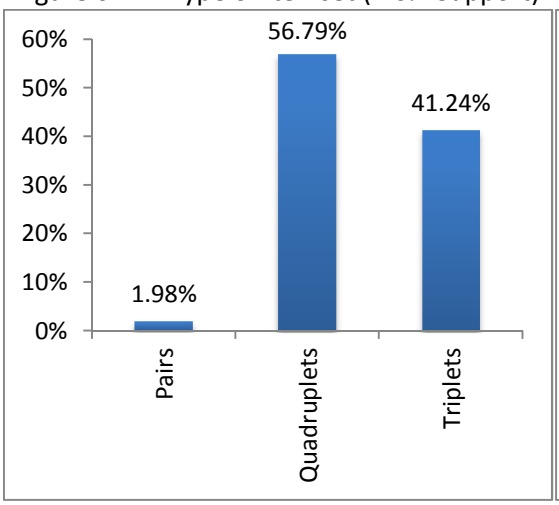
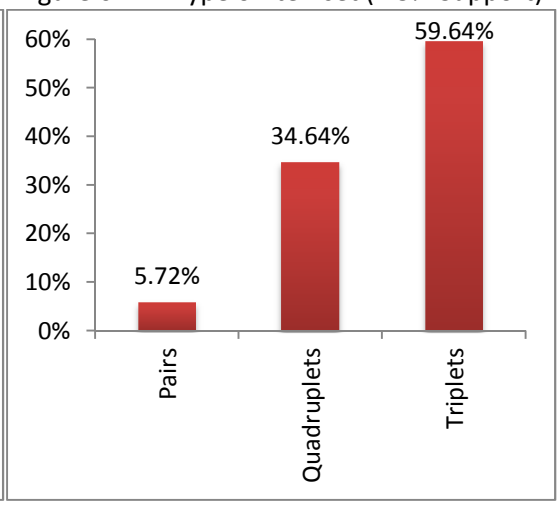


Figure 6.1.2: Type of Itemset (1.5% Support)



After aggregating our product results based on their business unit we surprisingly observe that only two out of eight business units, business units 6 and 7, were included in frequently purchased itemsets for both cases of 1.0% and 1.5% support. This might be partially explained by the high volume of sales for these two business units as demonstrated in figure A.2, Appendix A. In combination with that, most products in the generated frequent itemsets turned out to be complementary products as will be further demonstrated later on in our analysis. Business unit 7 accounts for more than 80% of products found in itemsets for both 1.0% and 1.5% support thresholds while business unit 6 accounts for approximately 16% on average (figure 6.1.3 and 6.1.4).

Figure 6.1.3: Business Units (1.0% Support)

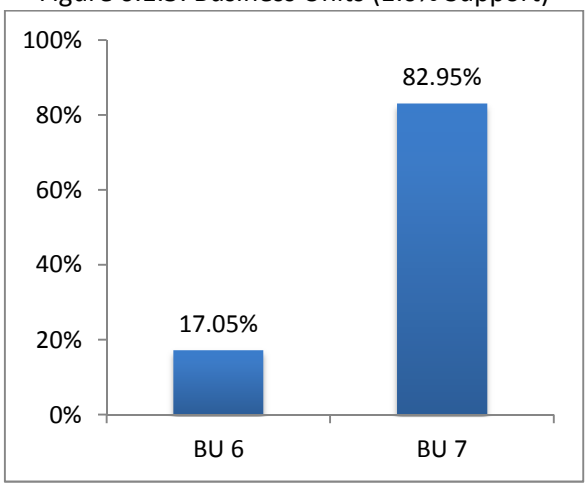
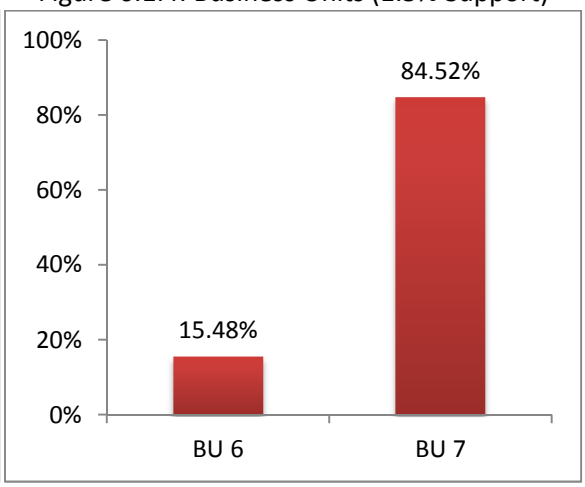


Figure 6.1.4: Business Units (1.5% Support)



By using aggregation techniques and text-merging formulas we managed to characterize each frequently purchased itemset based on whether both its antecedents and consequent belong to the same business units. Based on the results of this analysis, figure 6.1.5 and 6.1.6 was constructed. These figures demonstrate that the majority of rules include product combinations coming from the same business unit. Itemsets comprised of products from different business units increase from 6.63% to 13.92% as we decrease the support threshold from 1.5% to 1.0% respectively. This indicates that the product synthesis of itemsets might change as we vary our model's support thresholds.

Figure 6.1.5: Same vs Different Business Unit (1.0% Support)

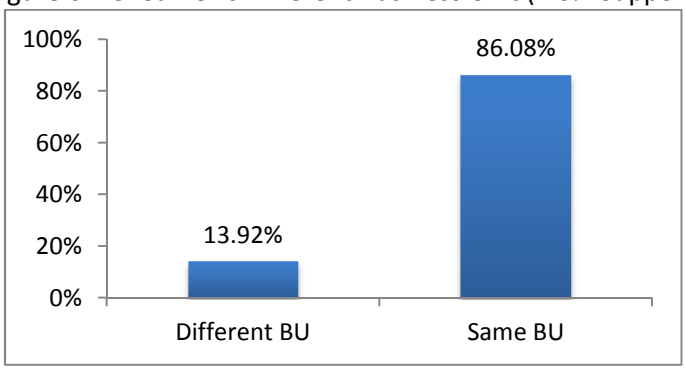
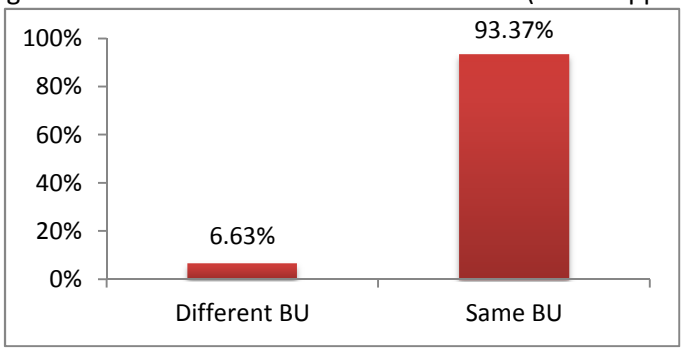


Figure 6.1.6: Same vs Different Business Unit (1.5% Support)



86.08% and 93.37% of the itemsets include products from the same business unit. This is a strong indication that complementary products are more frequently sold together than non-complementary ones. Nevertheless, we have to take a more careful look at the product category level in order to make more accurate assessments (Table A.2, Appendix A for business unit and product category index). Our analysis indicates that frequent itemsets include products orienting from 7 out of 45 product categories at the 1.0% support threshold (figure 6.1.7). Moreover, only 5 out of 45 products categories are included in frequent itemsets at the 1.5% support threshold (figure 6.1.8). Figures 6.1.7 and 6.1.8 validate the strong relationship between product categories 7.4, 6.5, 6.4 and products included in frequently purchased itemsets since they account for 97.02% and 97.99% of the products.

Figure 6.1.7: PC in Itemsets (1.0% Support)

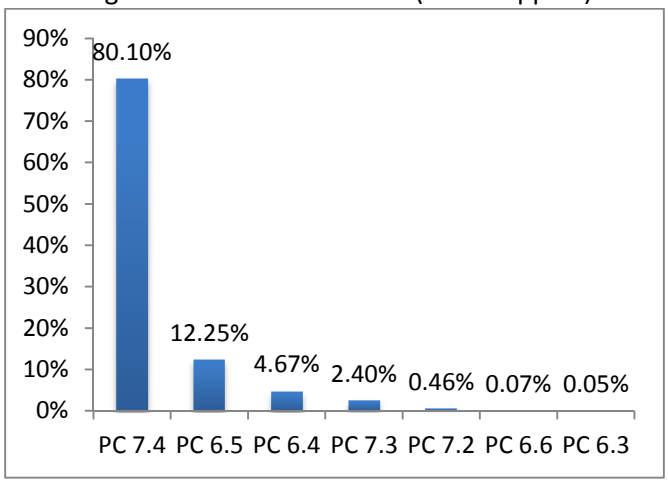
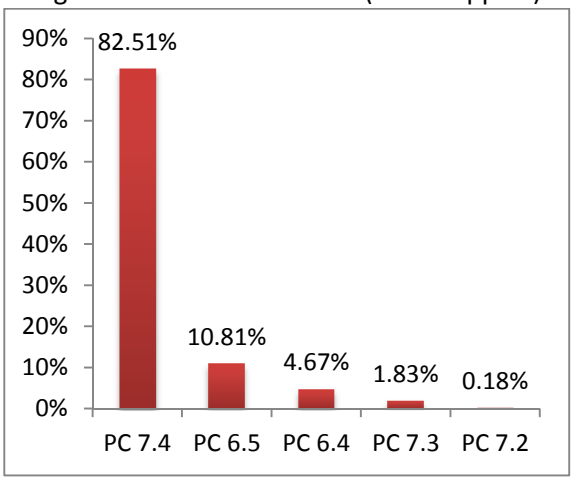


Figure 6.1.8: PC in Itemsets (1.5% Support)



As far as the product synthesis of the itemsets' consequents is concerned, it is of paramount interest that the vast majority of products marked as consequents, orient from product category 7.4 for both support thresholds (figure 6.1.9 and 6.1.10). This implies that if a basket is characterized as frequently purchased, there is an 84% chance that it will include consequents from PC 7.4 (and BU7 consequently).

Figure 6.1.9: PCs for Consequents (1.0% Support)

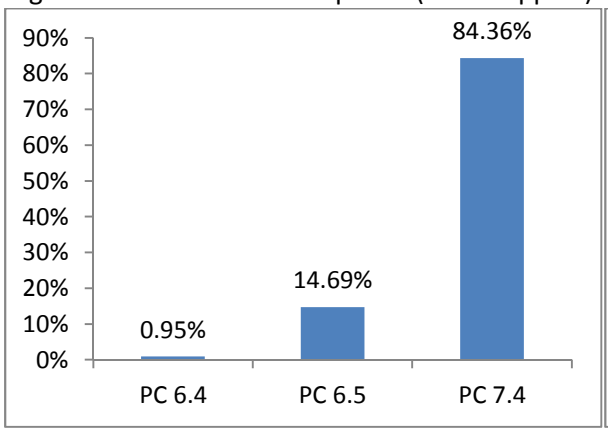
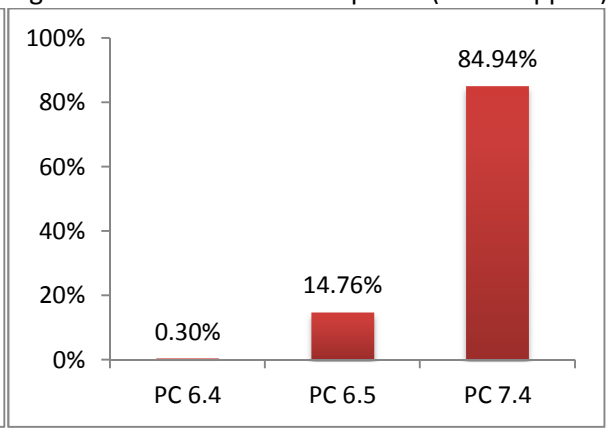


Figure 6.1.10: PCs for Consequents (1.5% Support)



Following the same logic as above there is a last step that needs to be taken in order to complete this first phase of our analysis. By using the same aggregation and text-merging techniques, we managed to identify whether all itemset products come from the same product category. After that, we counted the times a frequently purchased itemset is comprised of products of the product category. As figure 6.1.11 and 6.1.12 demonstrate, frequently purchased itemsets are expected to be comprised of products from the same product category 67.53% and 77.71% of the times for a support threshold of 1.0% and 1.5% respectively.

Figure 6.1.11: Same vs Different PC (1.0% Sup.)

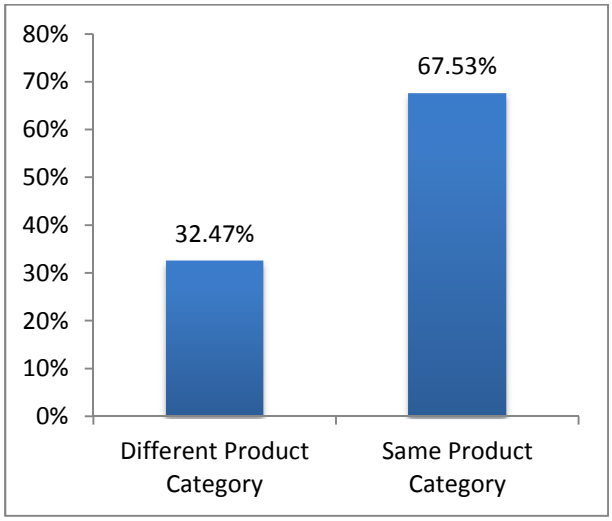
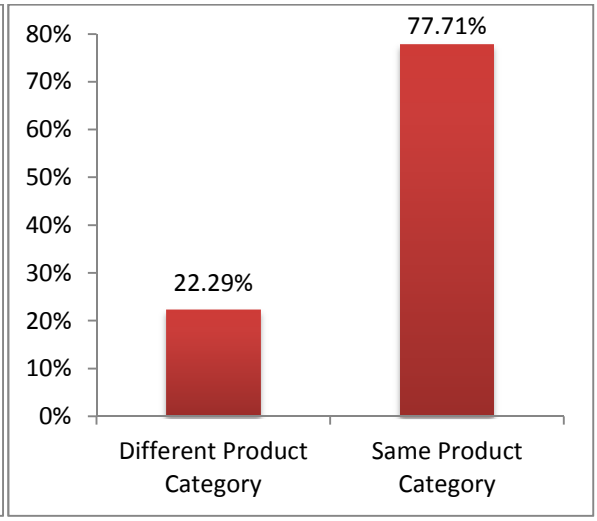


Figure 6.1.12: Same vs Different PC (1.5% Sup.)



The same results are observed, if we take a careful look at figures A.3 and A.4, Appendix A, where the combination of product categories found in frequent itemsets is shorted out from the most frequent to the less frequent. These tables underline the massive effect of product category 7.4 on frequently purchased itemsets since quadruplets and triplets of products from this product category account for a 61% and 65% of total product category combinations for 1.0% and 1.5% support thresholds respectively.

Figures A.5-A.10 in Appendix A have been constructed for completeness purposes since they summarize the percentage of rules assigned to various support, confidence thresholds and lift levels. Figures A.5 and A.6, Appendix A, illustrate that the more we increase support threshold, the less weak rules our model will provide us with. In addition, most of the rules have a support level between 1% and 2%. The proportion of rules having a support greater than 2% is substantially decreasing as we move to higher support levels. The same interpretation, more or less, counts for confidence levels (figures A.7 and A.8, Appendix A) where we see that rules with strong confidence levels greater than 90% vary between 4.22% and 5.24% of total rules. Figures A.9 and A.10, Appendix A, follow the same logic; increasing lift levels account for fewer rules than lift levels lower than 30. It is important thought to underline that 37.11% and 40.66% of the itemsets had a lift level greater than 10 and less than 30. That means that many of the rules discovered were unique rules found only within these specific baskets and not in our database in general. This might be due to the discovery of many trivial rules as well as unique customer purchasing patterns.

The above results were presented to the company under study and provided with strong evidence that products of the same product category are most probable to be sold together. A meeting was set between the researcher and the three product managers responsible for the products within the itemsets. A thorough discussion was made, after carefully investigating the association rules, concerning the relationship among the products under investigation. The outcome of this meeting was that the vast majority of the products included in frequently purchased itemsets as well as the relationship between antecedents and consequents, is strongly complementary.

Figure 6.1.13: Interdependencies between Product Categories

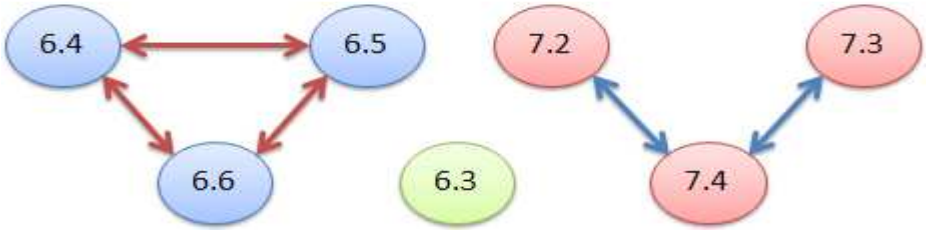


Figure 6.1.13 illustrates the complementarity of all product categories included in the given itemsets for both cases of 1.0% and 1.5% support thresholds. All three product categories 6.4, 6.5, 6.6 are highly correlated in the sense that the products included within these product categories, are expected to be sold together since their attributes share complementary characteristics (e.g. white socket and white socket frame). Likewise, product categories 7.2, 7.3 and 7.4 were described by the responsible product managers as complementary product categories with the difference that category 7.2 and 7.3 were not related to each other. Product category 6.3 was considered as independent but we decided to pay no further attention to that since it only accounts for 0.05% of the customer frequent itemsets (Figure 6.7).

The above analysis in combination with the inferences made based on figures 6.1.11 and 6.1.12 will be the main outcomes justifying our final conclusions connected to our research questions and hypothesis.

6.2.1 PRELIMINARY RESEARCH ON SEASONALITY

On retail baskets we would expect itemsets to have a strong seasonal component (e.g. Christmas products) but on a B2B environment we would expect this effect to be weaker. In order to examine the seasonality effect we are going to group sales and sales that contain one or more frequently purchased itemsets from our previous analysis (for more information concerning the construction of the dependent variable, please refer to chapter 5, Methodology). By aggregating the data, we get the following results:

We have two sets of itemsets, the 1.0% support itemset and the 1.5% support itemset. Here are the results (months are marked with numbers e.g. February=2, July=7 etc):

Figure 6.2.1: Sales orders per month (1.0% & 1.5% support)

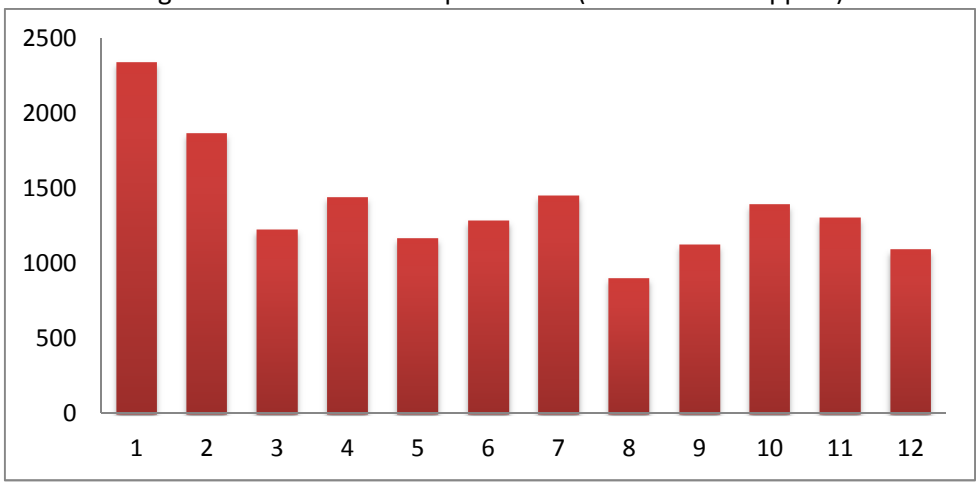


Figure 6.2.1 shows that there is strong seasonality with sales on January, February, April and July which is most likely connected with sales for the end of the year or with demand due to the summer season of the Greek market. If we plot the percentage of orders that contain a frequently purchased itemset as a function of the month for the two different support levels, we get the below figures:

Figure 6.2.2: % of orders containing one or more frequent itemsets per month (1.0% support)

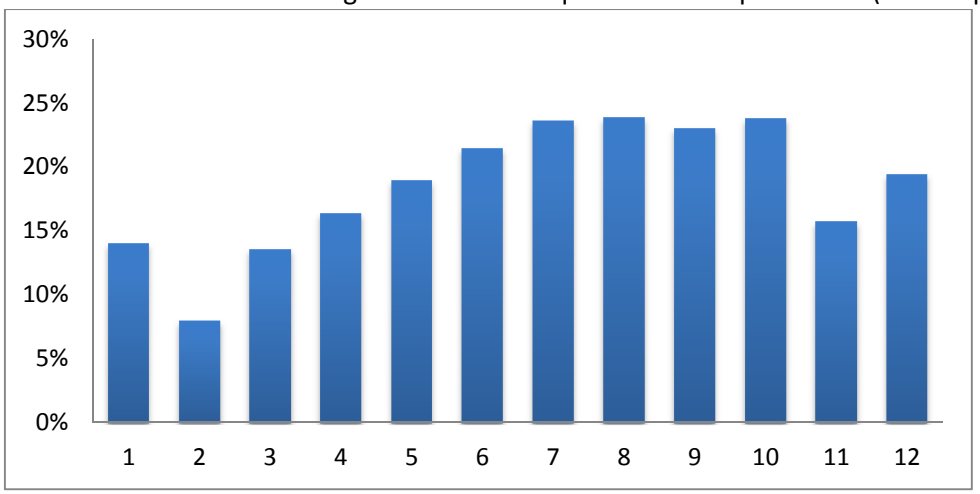
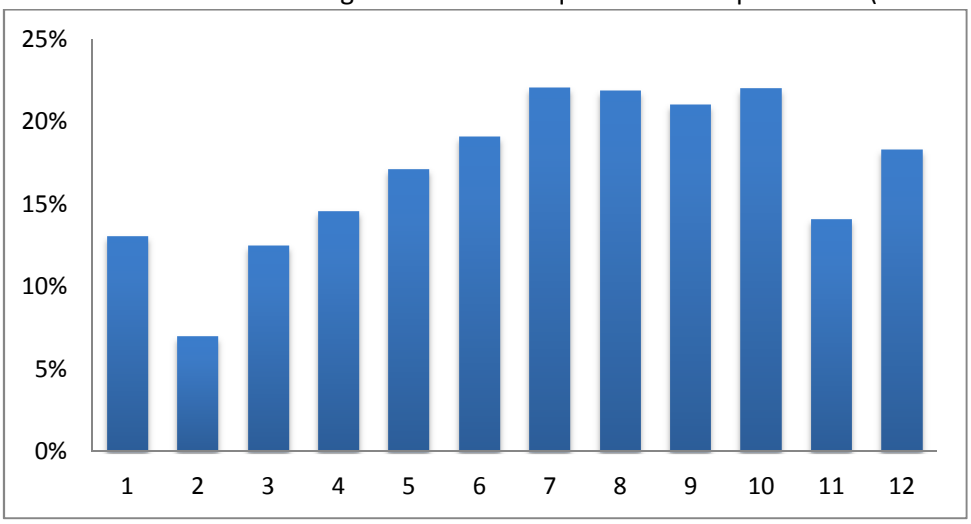


Figure 6.2.3: % of orders containing one or more frequent itemsets per month (1.5% support)



Figures 6.2.2 and 6.2.3 clearly illustrate the seasonal component and the fact that it is quite strong. The percentage of sales that contain one or more itemsets ranges from 7% to 21% (3x), depending on the month with July, August, September and October being the strongest no matter if we use 1.0% or 1.5% support levels.

It is interesting to observe that this pattern is different to the one of the Sales (see figure 6.2.1 at the beginning of this section). In the sales we also observe a variance of 3x between the strongest and the less strong months, but August for example is one of the weakest months in terms of sales and one of the strongest in terms of sales that contain frequent itemsets. It's clear that the forces that drive sales of itemsets are different to the ones that drive general sales.

We will get back to the subject while examining the results of the logistic regression. It's also interesting to research and find the business justification behind those results.

6.2.2 SEASONALITY AND BINARY LOGISTIC REGRESSION

In order to model seasonality on the frequently purchased itemsets, we use binary logistic regression. Our binary dependent variable represents the presence or absence of a frequently purchased itemsets on an order. Our independent variable is a categorical variable, months, which was treated as eleven dummy variables, one for each month and one as a reference variable. Our general model equation has the below form:

$$logit(\pi_i) = x_i'\beta \rightarrow Ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(January) + \beta_2(February) + \beta_3(March) + \beta_4(April) + \beta_5(May) + \beta_6(June) + \beta_7(July) + \beta_8(August) + \beta_9(September) + \beta_{10}(Octeber) + \beta_{11}(November) \quad (6.1)$$

Ratio $Ln\left(\frac{\pi_i}{1-\pi_i}\right)$ stands for the log-odds ratio (as calculated by SPSS) which was more extensively presented on chapter 5.7. Variables (January), (February),…, (November) can only take values 1 and 0. A basic assumption of this specific model is that months are mutually exclusive, meaning that if one variable takes the value of 1, the rest of the values automatically become 0. If all variables take the value 0 then that means we are studying the effect of the reference variable, (December), on the

binary dependent variable. An important attribute of the mutual exclusiveness of months as an independent variable is that we have no multicollinearity issues since co-occurrence of independent dummy variables is impossible.

The results provided by SPSS are partly controversial but confirm, up to a specific degree, our expectations from the previous chapter's analysis on seasonality. The controversy originates mainly from the fact that we did not get sufficient proof of the predictive strength for the above model. As we observe on table 6.2.1, the chi-square statistic is 314.2 on 11 degrees of freedom, significantly beyond 0.001. Omnibus Tests of Model Coefficients (table 6.2.1) is a test of the null hypothesis that adding the 11 dummy variables to the model has not significantly increased our predictive performance for the training set, a hypothesis which is rejected at the 0.001 significance level.

Table 6.2.1: Omnibus Tests of Model Coefficients

Chi-square	df	Sig.
314.200	11	0.000

The -2Loglikelihood statistic (table 6.2.2) measures how well the model predicts the binary dependent and has a value of 15,158. The lower the -2Loglikelihood statistic, the better a model is considered to be. Adding the dummy variables into the model reduced the value of -2Log likelihood statistic by the above chi-square statistic meaning that a model with only the intercept would have a -2Loglikelihood of 15,158 +314=15,472. The -2Loglikelihood of a model with only the intercept is higher than a model including the intercept and the dummy variables so as a result we can assess that adding the dummy variables improve the in-sample predictive performance of the model.

Moreover, we observe very low values for Cox & Snell R^2 (can be interpreted like R^2 in a multiple regression, but cannot reach a value of 1) and Nagelkerke R^2 (can reach a maximum of 1) that shows that the data might not fit well our statistical model.

Table 6.2.2: Model Summary

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
15158.724	.019	0.031

The above analysis shows a significant improvement in the in-sample predictive performance of our model compared to a model with only the intercept but rather discouraging results based on the values of the two coefficients of determination. Hereunder we will take a look at the classification rates table for our binary logistic regression at the 1.0% support threshold.

Based on the probability function 5.7, regression coefficients (table 6.2.4) and our input data we can classify each case based on the predicted probability π_i. By default, SPSS classifies a case as "1" if the predicted probability is greater than 0.5 and "0" otherwise. The predicted and observed cases are plotted together in a table called classification table (table 6.2.3). Default cut-off value of 0.5 is not always preferable since the case of an event occurring (taking the value "1") might be relatively rare as it happens in our case (our dependent variable is taking the value "1" less than 20% of the total cases). This can lead to misclassification of "1" cases as "0" based on their predictive probability. This is why we have to set a cut-off value that provides us with the most correctly classified cases. By following a technique called Receiver Operating Characteristic (ROC) technique, we can approximate

the best cut-off value. The purpose of this technique is to maximize both cases correctly predicted as "0" and "1". Further explanation of ROC concepts goes beyond the scope of this thesis.

By running ROC in SPSS, we got a suggested cut-off value of 0.20 for the seasonality model. Based on that, SPSS provided us with classification table 6.2.3. As this table illustrates, we manage to successfully classify 65.3% of the "0" cases and 48.5% of the "1" cases. The overall successful classification rate is 62.3% and is also known as sensitivity of prediction. The sensitivity of prediction of our model has significantly decreased compared to the sensitivity of prediction for a model with only the intercept (82.3%). Nevertheless, a model with only the intercept completely fails to successfully classify the cases where an event occurs (takes values "1") leading us to prefer the model including the dummy variables.

Table 6.2.3: Classification Table (1.0% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	8917	4735	65.3
	1	1510	1422	48.5
		Overall Percentage:		62.3

*cutoff value: .200

Table 6.2.4 provides the estimated β coefficients of the model. We observe results almost similar to our initial seasonality analysis. First, we have to reject the null hypothesis that the difference between May's and December's effect on the probability of a customer order to include a frequently purchased itemset, is significantly different from zero (ceteris paribus), since May's β coefficient is insignificant at the 0.774 significance level. The same inferences can be made for June since β coefficient is 0.223. By keeping all other parameters constant (ceteris paribus) we imply that all other dummy variables automatically take a value of 0 since dummy variable January has taken the value 1. The rest of the months have a significant effect at the 0.05 significance level which means they differ from December significantly.

Table 6.2.4: Estimated Beta Coefficients

Explanatory Variables	B	Sig.
January	-.394	.000
February	-1.026	.000
March	-.429	.000
April	-.209	.046
May	-.031	.774
June	.125	.223
July	.249	.011
August	.264	.016
September	.216	.038
October	.261	.009
November	-.257	.017
Constant	-1.426	.000

The logit model for January has the below form (according to equation 6.1):

$$logit(\pi_i) = x_i'\beta \rightarrow Ln\left(\frac{\pi_i}{1-\pi_i}\right) = -1.426 - 0.394(1) - 1.026(0) - 0.429(0) - 0.209(0) - 0.31(0) - 0.125(0) - 0.249(0) - 0.264(0) - 0.216(0) - 0.261(0) - 0.257(0) = -1.82 \quad (6.2)$$

Likewise we can construct the logit modes for every month separately. Setting all dummy variables equal to 0 provides us with a model with only the constant. This model provides us with the log-odds ratio for December which was used as a reference month for the estimation of betas for the remaining eleven months.

Moreover, we can make use of regression's betas in order to construct the probabilities and odds for each month based on equations 5.7 and 5.8 respectively. Calculation of odds and probabilities for all months provides us with the results as demonstrated on table 6.2.5. Odds, as used for this specific research, show that e.g. an order placed on January is 0.162 more likely to include a frequently purchased itemset than not to include one, ceteris paribus. According to equation 5.7, the formula to calculate odds ratio for January is:

$$Odds_{January} = e^{(\beta_0+\beta_1)} = e^{(-1.426-0.394)} = 0.162 \tag{6.3}$$

where β_0= Constant= -1.426 and β_1= Regression coefficient for January= -0.394.

As table 6.2.5 demonstrates, the highest odds ratios are July's, August's, October's and December's. All months' odds ratios are below 1.0 meaning that throughout the entire year, it is more likely that an order will not include a frequently purchased itemset than to include one. May's and June's coefficients are insignificant so we do not take them into account.

The probability for January can be calculated as demonstrated below:

$$Odds_{January} = \frac{\pi_{January}}{1-\pi_{January}} \rightarrow (1-\pi_{January})Odds_{January} = \pi_{January} \rightarrow$$
$$Odds_{January} = \pi_{January}(1+Odds_{January}) \rightarrow \pi_{January} = \frac{0.162}{1+0.162} = 13.94\% \tag{6.4}$$

This probability shows that e.g. there is 13.94% likelihood that an order placed on January will include a frequent itemset. In other words, it is expected that 13.94% of January's customer orders will include a frequently purchased itemset. The same interpretation holds for the remaining months; β coefficients are calculated by taking December as a reference month. If all our dummy variables equal to 0 then we get a model with only the constant. By using this model we calculate the odds and probability for December (0.240 and 19.38% respectively).

Table 6.2.5: Odds and Probabilities

Month	Odds	Probabilities
January	.162	13.94%
February	.086	7.93%
March	.156	13.53%
April	.195	16.32%
May	.233	18.90%
June	.272	21.40%
July	.308	23.57%
August	.313	23.84%
September	.298	22.98%
October	.312	23.78%
November	.186	15.67%
December	.240	19.38%

We observe that the probabilities on table 6.2.5 follow exactly the same trend line as figures 6.2.2 and 6.2.3, confirming our initial claims that the seasonal component is strong and reaches its peak on July, August, September and October. The probabilities that an order placed on these months will include a frequently purchased itemset exceed 23%.

The above analysis was conducted at the 1.0% support threshold. Tables A.3-A.7 in Appendix A summarise the results for 1.5% support threshold. The conclusion for the 1.5% support threshold is similar to the one for the 1.0% with the only difference that beta coefficient for September is also insignificant and that the probabilities for July, August and October are approximately 22% (December as a reference month).

6.3 PHASE THREE: BINARY LOGISTIC REGRESSION RESULTS AND ANALYSIS

6.3.1 INDEPENDENT VARIABLES: PREPARATORY STAGE AND TRANSFORMATION

After conducting the first stage of the binary logistic regression, we can proceed to the next stage of our analysis where the behaviour of our second binary dependent variable will be studied. Moreover, it should be mentioned that our dependent variable and the database used for this phase of our research, differs from the one used in the previous chapter (for more explicit explanation please refer to chapter 5.7). Adding to that and in contradiction with the previous stage, there might be a multicollinearity issue among our independent variables. The chances that such an issue exists among our independent variables are quite high since they are conceptually related to each other. For example, expensive products are usually not sold in large quantities so we do expect a negative relationship between quantity and price. The same counts for quantity and promotional discount rates since the higher the discount rates, the higher the expected quantity of products purchased.

Before running the binary logistic regression, it is necessary to understand the characteristics of each one of our independent variables as well as the relationship between them. After running a linear regression between purchased quantity and price per piece, the results were quite controversial. Table B.1 in Appendix B shows a significant negative relationship between price per piece and product order quantity, meaning that the more the price of a product is increased, the more the purchased quantity is decreased. Moreover, by taking this model's R^2 statistic into account (table B.2, Appendix B) we can imply that this linear model explains almost none of the variability of the response data around its mean since its value is very close to 0 (R^2= 0.001). In order to investigate further the relationship between these two variables we constructed a scatterplot including both variables' values. Figure B.1 in Appendix B clearly depicts a non-existing relationship between price and quantity.

After taking an even closer view to the histograms of our independent variables, we discovered two rather inconvenient and extremely positively-skewed histograms as figures B.2 and B.3 in Appendix B clearly demonstrate. These figures reveal a highly dense concentration of products around low quantities (figure B.2) and low prices per piece (figure B.3). Moreover, figures B.2 and B.3 indicate that the prices and quantities range significantly in this B2B setting. More specifically, price ranges from €0.04 to €16,000 and order quantity ranges from 1 piece to 2,500 pieces.

Instead of using the price and quantity themselves, we use their logarithms. Our motivation is the following. By estimating our weights on the original price (or quantity) we implicitly give the same

effect of e.g. a +€10 price change on a €10 product and a €10.000 product. It is obvious thought, that this price change does not have the same effect on people's decisions in these two different cases. For that reason we decided to transform our initial independent variables into logarithmic ones in order to incorporate the non-linear effect of price and quantity perceptions. This way, for example a 10% discount on a €10 product and a €10.000 product will, after a linear transformation, have similar effect which is closer to what really happens when people decide if they want to buy or not. The main purpose of this decision is to allow the logistic regression to "see" more details on low prices and quantities (where they are more significant) and less on high.

Towards that direction, we transformed our independent variables into logarithmic ones: LnPriceperPiece and LnOrderQuantity. Figures 6.3.1 and 6.3.2 depict the histograms for the two constructed variables.

Figure 6.3.1: Histogram: LnPriceperPiece

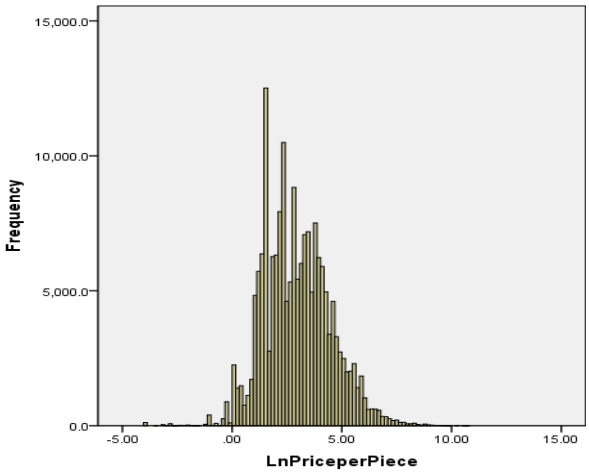
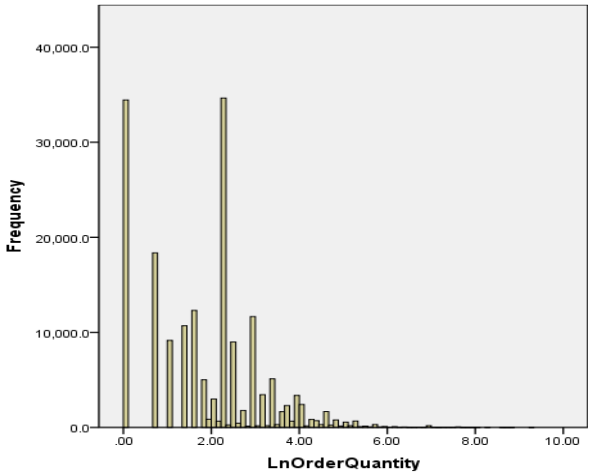
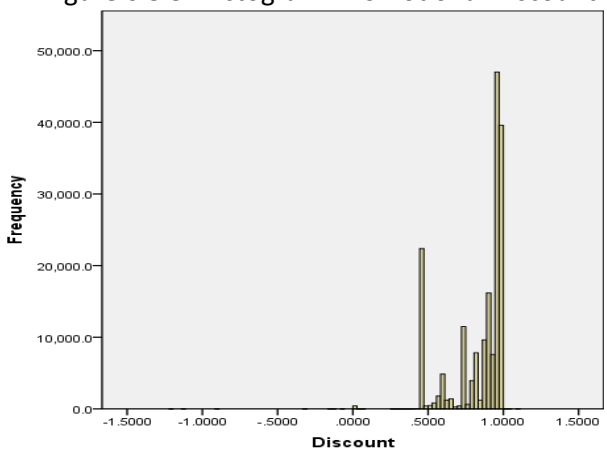


Figure 6.3.2: Histogram: LnOrderQuantity



As we notice from the above figures, the histograms of our transformed variables approximate the shape of a normal distribution (figure 6.3.1) and a skewed multimodal distribution (figure 6.3.2). Last, as we notice on figure 6.3.3, the histogram for the discount rate follows the shape of a slightly asymmetric and negatively skewed frequency distribution but no extreme skewness is observed as in the case of our untransformed independent variables. More information on the kurtosis and skewness statistics of the independent variables can be found on table B.3, Appendix B. We can now proceed further with our analysis (the same transformation process was followed for 1.5% support threshold with similar results).

Figure 6.3.3: Histogram: Promotional Discount Rate



6.3.2 INDEPENDENT VARIABLES: MULTICOLLINEARITY

The purpose of this sub-section is to investigate possible multicollinearity issues occurring among our three independent variables. As we already explained in the previous chapter, we expect that a linear relationship exists among the untransformed variables. The only difference now is that we have logarithmic prices and quantities rather than our initial prices and quantities. This multicollinearity control will be attempted by calculating and comparing Person Correlations as well as the Variance Inflator Factors (VIFs).

6.3.2.1 PEARSON CORRELATIONS

The results on table 6.3.1 have been provided by SPSS and are significant at the 0.01 significance level for both 1.0% and 1.5% support thresholds. By making use of the rule of thumb, Pearson Correlations higher than 0.70 (max 1) show a very strong positive relationship and below 0.50 (min 0) show a moderate relationship between the variables under testing. Negative values refer to negative relationship between the two variables.

Table 6.3.1: Pearson Correlations

	Discount	LnOrderQuantity	LnPriceperPiece
Discount	1	0.829	-0.560
LnOrderQuantity		1	-0.659
LnPriceperPiece			1

As clearly depicted, Discount is strongly correlated to LnOrderQuantity with a Pearson correlation coefficient of 0.829 and negatively correlated to LnPriceperPiece with a correlation coefficient of moderate magnitude (-0.560). Moreover, LnOrderQuantity is negatively correlated to LnPriceperPiece with a correlation coefficient of -0.659, demonstrating a moderate to high correlation between the two variables. Table 6.3.1 confirms the previous section's expectations that the higher the price of a product is, the lower the order quantity for this product will be. In addition, we confirm our intuition that high discount rates positively influence the product quantities a customer orders. Last, expectations of higher promotional discount rates translating into lower product prices are confirmed.

6.3.2.2 VARIANCE INFLATOR FACTOR

In this section, the Variance Inflator Factor (VIF) will be used as an alternative for testing our independent variables for multicollinearity issues. VIF is a statistic that quantifies the severity of multicollinearity and provides us with an index of how much the variance of the estimated regression coefficients is increased because of collinearity. If VIF value is greater than 5, there is high multicollinearity among the independent variables. VIF's performance increases if the variables under study are intervals or rations as in that specific case.

After running three successive linear regressions, each time rotating variables in order to get all possible combinations of pairs of independent and dependent variables), SPSS provided us with Tables 6.3.2, 6.3.3 and 6.3.4. VIFs for all combinations of variables were low comparing to number 5 meaning that high multicollinearity does not exist. It is worthwhile to say that when Discount and

LnOrderQuantity were paired as independent variables, our model output provided us with the highest VIF values, once again showing the very strong correlation between these two variables, also indicated by the previous sub-section's high Pearson Correlation (0.829).

Table 6.3.2: VIF, Linear Regression, 1st pair

Model	B	Sig.	VIF
(Constant)	.638	.000	
LnOrderQuantity	.113	.000	1.767
LnPriceperPiece	-.003	.000	1.767

*Dependent Variable: Discount

Table 6.3.3: VIF, Linear Regression, 2nd pair

Model	B	Sig.	VIF
(Constant)	4.591	.000	
Discount	-.375	.000	3.199
LnOrderQuantity	-.721	.000	3.199

*Dependent Variable: LnPriceperPiece

Table 6.3.4: VIF, Linear Regression, 3rd pair

Model	B	Sig.	VIF
(Constant)	4.591	.000	
LnOrderQuantity	-.375	.000	3.199
Discount	-.721	.000	3.199

*Dependent Variable: LnPriceperPiece

At this point of our research it is very important to mention that even though it is expected that the high collinearity among the independent variables will inflate the results of the binary logistic model, the researcher chooses to proceed with the binary logistic regression, always bearing in mind the possibility of distorted outcomes. Adding to that, it is a positive sign that none of the variance inflator factors was substantially high.

6.3.3 MODEL SELECTION

Before proceeding to the analysis of the results for the binary logistic regression, we will attempt to identify the best possible model in order to provide with the highest quality of results. This will be achieved by creating five different binary logistic models: one including all three variables before the logarithmic transformation, one with all three variables (promotional discount, price per piece, and order quantity) and three reduced models with paired combinations. The process of comparing full and reduced models, in combination with performance diagnostic statistics, such as Loglikelihood

statistic and Akaike Information Criterion (AIC), will direct us towards the selection of the proper models for the needs of this analysis. The next chapter addresses the process of model selection for both 1.0% and 1.5% support levels.

6.3.3.1 LOGLIKELIHOOD AND AKAIKE INFORMATION CRITERION

IBM SPSS provides us with all the information necessary to make the appropriate comparisons between the models. First, SPSS output calculates the loglikelihood for each binary model which is the statistic that measures how poorly the model predicts the binary outcome within the sample database. The closer this statistic it is to zero, the better the model predicts the outcome. Afterwards, the AIC statistic will indicate the best possible model. AIC statistic is a function of the loglikelihood but it also includes a penalty for increasing explanatory variables as described in section 5.7.2. The lower the AIC statistic is the more preferred the model. After calculating the AIC statistics for every different combination of variables, column "Δ(AIC)" is calculated which is the percentage difference of the model with the lowest AIC statistic (model 1) and AIC statistic for each one of remaining models.

Table 6.3.5: AIC Statistic Calculation (1.0% Support Threshold)

A/A	Model Explanatory Variables	-2 Log Likelihood	No of Predictors+Constant	AIC	Δ(AIC)
1	Discount, OrderQuantity, PriceperPiece	123815.096	3	123821.096	
2	Discount, LnOrderQuantity, LnPriceperPiece	124774.414	4	124782.414	0.78%
3	Discount, LnPriceperPiece	125258.973	3	125264.973	1.16%
4	LnOrderQuantity, LnPriceperPiece	127134.498	3	127140.498	2.65%
5	Discount, LnOrderQuantity	127307.762	3	127313.762	2.75%

As table 6.3.5 demonstrates, the lowest AIC statistic belongs to the model including explanatory variables Discount, OrderQuantity and PriceperPiece while the model that includes all three variables with the logarithmic transformations comes 2nd in performance with a 0.78% difference from the 1st one. Selecting a model with untransformed independent variables does not follow section's 6.3.1 argumentation. In this section we supported that logarithmic transformation was necessary in order to construct a more representing model. For that reason we will make use of other model quality measurements in order to identify the most efficient one.

6.3.3.2 CLASSIFICATION TABLES

In section 6.2.2 we introduced the basic concepts of classification tables in the logistic regression and ROC analysis. This analysis approximated an effective cut-off value of 0.250. As we can see on table 6.3.6, the 1st model successfully classifies a product as part of a frequently purchased itemset 75.1% of the times. The correct classification of a product not being a part of a frequent itemset is 73.8%, providing us with a 74.0% sensitivity of prediction.

Table 6.3.6: Classification Table for Model 1 (1.0% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	112353	39800	73.8
	1	6849	20703	75.1
	Overall Percentage:			74.0

*cutoff value: .250

Table 6.3.7 illustrates the 2nd model's classification table. The correct classification of cases as "1" increases to 83.0% and the correct classification of cases as "0" decreases to 59.9%. Nevertheless, we notice a significant improvement in the sensitivity of analysis by 5.4 percentage points up to the level of 79.4%.

Table 6.3.7: Classification Table for Model 2 (1.0% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	126260	25893	83.0
	1	11056	16496	59.9
		Overall Percentage:		79.4

*cutoff value: .250

The above classification tables provide us with vital insights in order to continue with the selection of a proper model. We notice that even though the 1st model –consisted of the untransformed variables- has the lowest Loglikelihood rates and AIC values of all, we should also take into serious consideration the in-sample predictive performance of our model. Consequently, we are forced to reject model 1 as an optimal option since it has the lowest in-sample predictive performance (sensitivity of prediction is 74.0%). The classification tables for the rest of the models are enlisted in Appendix B (tables B.4- B.6). Moreover, for the proper model selection, we need our choices to meet three requirements:

1) Have the lowest AIC value possible.
2) Have the highest sensitivity of prediction possible.
3) Include as many independent variables as possible.

Based on the above criteria, model 4 seems to be one of the best candidates since it has the highest sensitivity of prediction (80.6%). Nevertheless, it has the 2nd lowest AIC value and only includes two independent variables. Model 2 seems to perfectly meet all the above criteria since it has the 2nd lowest AIC value, 2nd highest sensitivity of prediction (79.4%) and it includes all three independent variables under study.

Based on the above, we conclude to the selection of model 2 at the 1.0% support threshold. The same rationale and conclusions count for the 1.5% support threshold and tables B.7 to B.12 in Appendix B. The logistic function for model 2 is:

$$logit(\pi_i) = x_i'\beta \rightarrow Ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(Discount_i) + \beta_2 Ln(OrderQuantity_i) + \beta_3 Ln(PriceperPiece_i) \quad (6.4)$$

6.4 RESULTS AND ANALYSIS

This section discusses the binary logistic regression results for our database. As already addressed, the binary dependent variable is whether a product is part of a frequently purchased basket or not. The explanatory variables are the promotional discount rate, the natural logarithm of product quantity purchased and the natural logarithm of product catalogue price per piece. In the beginning of this research, we stated that there are expectations that a positive relationship exists between the discount rate and the probability of a product to be part of a frequent itemset. Moreover, as the price per piece increases, the probability of a product being part of an itemset should decrease. Increasing purchased quantities for a product should translate into higher probability of this product being sold in a frequent itemset. Hereunder we will test these hypotheses and validate them whenever possible.

The analysis on section 6.3.3 indicated model 2 as the most efficient of all candidate models in order to study the probability functions for our binary logistic regression. Based on our database for the 1.0% support threshold, SPSS estimated the beta coefficients of our model which are presented on table 6.4.1. Table 6.4.1 also includes the significance levels for each beta coefficient.

Table 6.4.1: Beta Coefficients, Model 2 (1.0% Support)

Model	B	Sig.
Discount	7.088	.000
LnOrderQuantity	.233	.000
LnPriceperPiece	-.340	.000
Constant	-7.911	.000

As we notice, all beta coefficients are significant at the 0.01 significance level. The logistic regression function for model 2 becomes:

$$logit(\pi_i) = x_i'\beta \rightarrow Ln\left(\frac{\pi_i}{1-\pi_i}\right) = \\ -7.911 + 7.088(Discount_i) + 0.233Ln(OrderQuantity_i) - 0.340Ln(PriceperPiece_i) \quad (6.7)$$

Based on equation 6.7 and 5.8 we calculate the Odds function:

$$Odds_i = e^{x_i'\beta} \rightarrow Odds_i = e^{-7.911+7.088(Discount_i)+0.233Ln(OrderQuantity_i)-0.340Ln(PriceperPiece_i)} \quad (6.8)$$

Since $Odds_i = \frac{\pi_i}{1-\pi_i}$, by solving 6.8 for π_i we calculate probability function:

$$\pi_i = \frac{\exp\{x_i'\beta\}}{1+\exp\{x_i'\beta\}}, where\ x_i'\beta\ as\ defined\ in\ (6.7) \quad (6.9)$$

Equations 6.8 and 6.9 provide us with the odds and probability values for all combination sets of independent variables. For example, if we set Discount=1, LnOrderQuantity=0 and LnPriceperPiece=0 then we get an odds value of 0.44 and probability value of 30.51% (first line, table 6.4.2). If we set Discount=0%, LnOrderQuantity=1 and LnPriceperPiece=0 then we get an odds value of 0.000463 and probability value of 0.046% (second line, table 6.4.2). Setting Discount=0%, LnOrderQuantity=0 and LnPriceperPiece=1, we get an odds value of 0.000261 and probability value of 0.026% (third line, table 6.4.2).

Table 6.4.2: Odds and Probabilities, Model 2 (1.0% Support)

Odds	Probabilities
.44	30.51%
.000463	0.046%
.000261	0.026%
.000367	0.037%

A model with only the constant provides us with an odds value of 0.000367 and a probability value of 0.037% (fourth line, table 6.4.2). That means that a product with LnPriceperPiece=0, LnOrderQuantity=0 (PriceperPiece=1, OrderQuantity=1) and Discount=0 is approximately 0.04% times more likely to be part of a frequent itemset than not to be.

We can plot probabilities as a function of logarithmic quantities and prices, setting the discount rate at prespecified fixed levels. In figures 6.4.1, 6.4.2 and 6.4.3 we plot the probabilities π_i, LnOrderQuantity_i and LnPriceperPiece_i for discount rates of 0%, 50% and 95% respectively.

At this point, it is important to mention that a 1 point increase in the logarithmic values of quantity or price -as we see it on the horizontal and depth axis of the below figures- corresponds to a 172% increase in the actual values of quantities and prices. This is more visible in the below equation:

$$Ln(PriceperPiece) + 1 = Ln(PriceperPiece) + Ln(e) = Ln(e \cdot PriceperPiece) \qquad (6.10)$$

Relationship 6.10 shows that each time we increase variable LnPriceperPiece by 1 unit it corresponds to a proportional multiplication by e≈2.72 for a 1 unit increase of its logarithmic counterpart. This is a 172% increase. The same applies for LnOrderQuantity and OrderQuantity.

Figure: 6.4.1: 3-D scatterplot: π_i, LnOrderQuantity_i, LnPriceperPiece_i, Discount= 0% (1.0% Support Threshold)

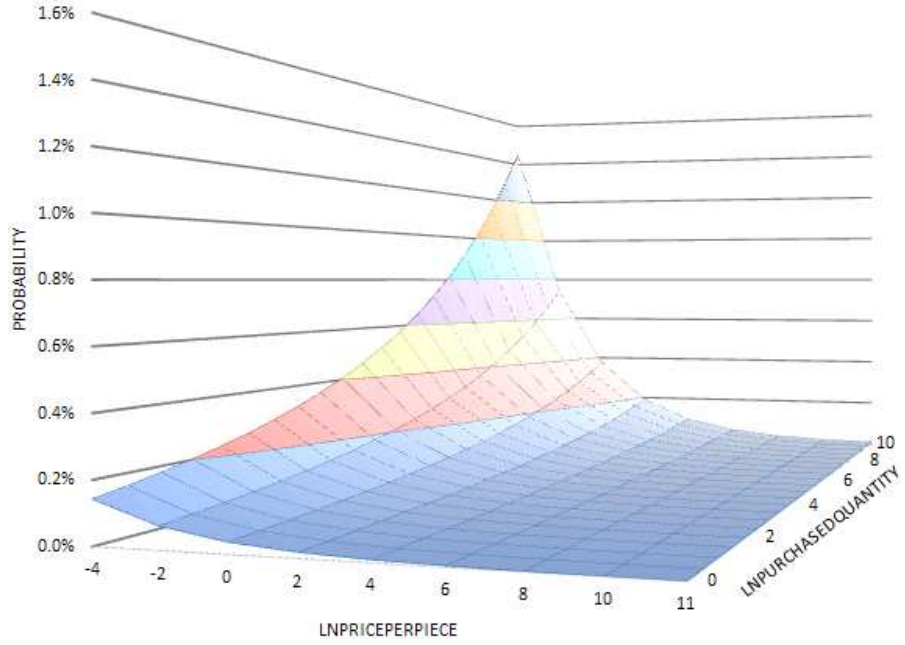


Figure: 6.4.2: 3-D scatterplot: π_i, LnOrderQuantity$_i$, LnPriceperPiece$_i$, Discount= 50% (1.0% Support Threshold)

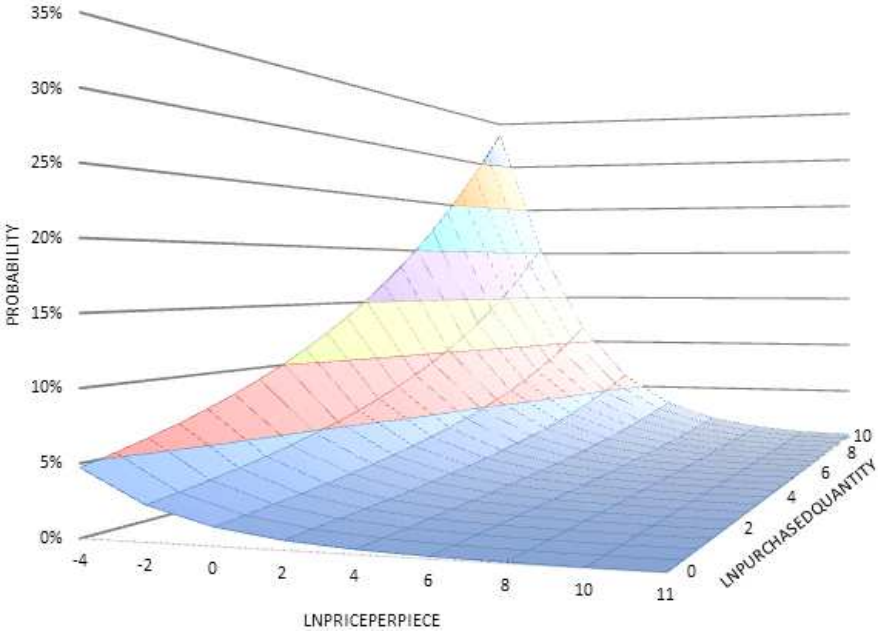
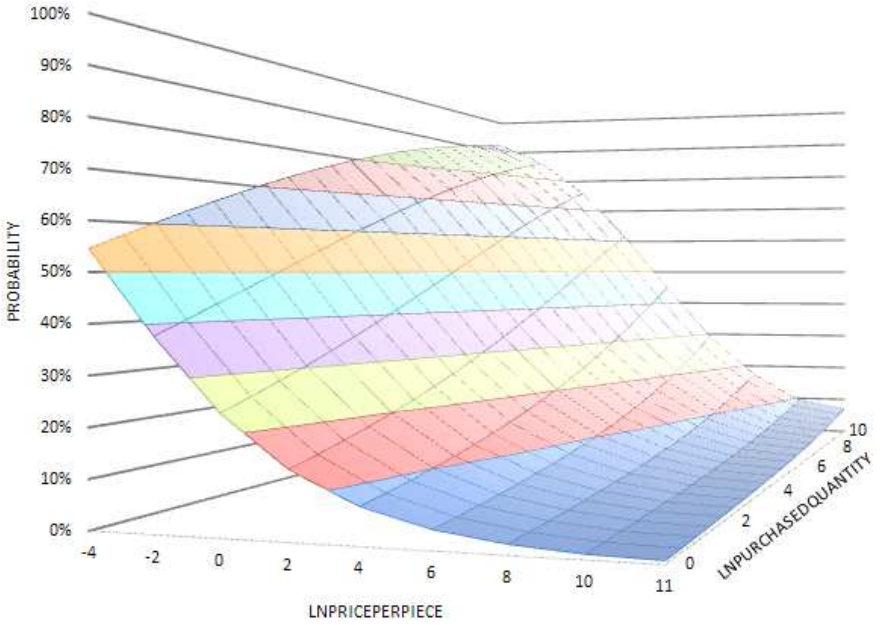


Figure: 6.4.3: 3-D scatterplot: π_i, LnOrderQuantity$_i$, LnPriceperPiece$_i$, Discount= 95% (1.0% Support Threshold)



The above three scatterplots were constructed in order to graphically demonstrate the relationship between all three independent variables and the dependent variable. We can easily observe that increasing discount rates greatly increase the probability that a product is part of a frequent itemset. When discount rate is set to the level of 0%, probability ranges from 0%-1.45% (figure 6.4.1). Setting discount rate to the level of 50% provides with a probability range of 0%-33.69% (figure 6.4.2). A discount rate of 95% dramatically increases the probability range, resulting in a range of 0%-92.5% (figure 6.4.3).

By making use of the probability function 6.9 we calculate the percentage change in the probability for changes in the values of the independent variables. Let's take the example where Discount=50%, LnPriceperPiece=0 and LnOrderQuantity=0. If we increase discount rate to the level of 95%, probability that a product is part of a frequent itemset also increases by 0.2230 points, keeping all

other parameters constant (equation 6.11). This corresponds to a +1,780% probability increase from 1.25% up to the level of 23.55%.

$$d\pi_i = \frac{e^{-7.911+7.088*0.5+0.233*0-0.340*0}}{1+e^{-7.911+7.088*0.5+0.233*0-0.340*0}} - \frac{e^{-7.911+7.088*0.95+0.233*0-0.340*0}}{1+e^{-7.911+7.088*0.95+0.233*0-0.340*0}} = 0.2355 - 0.0125 = 0.2230 \quad (6.11)$$

Figures 6.4.1, 6.4.2 and 6.4.3 clearly demonstrate that products with lower prices have a higher probability to be part of a frequent itemset (taking equation 6.10 into account). For example, in the case where discount=50%, LnPriceperPiece=0 and LnOrderQuantity=0 (figure 6.4.2), a 1 unit increase in the LnPriceperPiece decreases probability that a product is part of a frequent itemset by 0.0079 points. This corresponds to a 63% probability decrease from 1.25% to 0.46%, keeping all other variables constant.

$$d\pi_i = \frac{e^{-7.911+7.088*0.5+0.233*0-(0.340*0+1)}}{1+e^{-7.911+7.088*0.5+0.233*0-(0.340*0+1)}} - \frac{e^{-7.911+7.088*0.5+0.233*0-0.340*0}}{1+e^{-7.911+7.088*0.5+0.233*0-0.340*0}} = 0.0046 - 0.0125 = -0.0079 \quad (6.12)$$

Based on the analysis so far, we can interpret this result as following: Given a product with 50% promotional discount rate, 1 euro price per piece and an order quantity of 1 piece, we can decrease the probability that a product is part of a frequent itemset from 1.25% to 0.46% (63% decrease) by increasing its price by 172% (from 1 euro to 2.72 euros).

Similarly, products sold in higher quantities have a higher probability to be part of a frequent itemset. Using the same example as above, a product with 50% promotional discount rate, 1 euro price per piece and 1 piece of purchased quantity has a 2.54% probability to be included in a frequently purchased itemset. An increase in its purchased quantity by 100 pieces results in a probability increase by 0.490 points up to the level of 7.44% (+193%), keeping all other parameters constant.

$$d\pi_i = \frac{e^{-7.911+7.088*0.5+(0.233*0+4.615)-0.340*0}}{1+e^{-7.911+7.088*0.5+(0.233*0+4.615)-0.340*0}} - \frac{e^{-7.911+7.088*0.5+0.233*0-0.340*0}}{1+e^{-7.911+7.088*0.5+0.233*0-0.340*0}} = 0.0744 - 0.0254 = 0.490 \quad (6.13)$$

The same inferences can be made for various discount rates, price per piece and order quantity levels.

Likewise, table 6.4.3 shows the beta values for model 2 at the 1.5% support threshold. All beta coefficients are significant at the 0.01 significance level so there is sufficient evidence to reject the null hypothesis and accept that the effect of independent variables on the dependent is different from zero. For economy reasons, we will skip the detailed analysis of the effect for each one of the regressors as it can be directly assessed by following the exact same logic as extensively described above.

Table 6.4.3: Beta Coefficients, Model 2 (1.5% Support)

Model	B	Sig.
Discount	8.179	.000
LnOrderQuantity	.295	.000
LnPriceperPiece	-.309	.000
Constant	-9.586	.000

The logistic regression function for model 2 at the 1.5% support level is:

$$logit(\pi_i) = x_i'\beta \rightarrow Ln\left(\frac{\pi_i}{1-\pi_i}\right) = -9.586 + 8.179(Discount_i) + 0.295Ln(OrderQuantity_i) - 0.309Ln(PriceperPiece_ice) \quad (6.14)$$

Figures B.4, B.5 and B.6 (Appendix B) illustrate the probabilities π_i, $LnOrderQuantity_i$ and $LnPriceperPiece_i$ for fixed discount rate levels of 0%, 50% and 95% respectively at the 1.5% support threshold. The main difference between plots for 1.0% and 1.5% support thresholds is that the latter provide with lower maximum probabilities for the various discount rates providing us with shorter probability ranges. Interpretation and main outcomes remain the same as in the 1.0% case.

6.5 SUMMARY OF RESULTS

Throughout the chapters of this part of the research, we gradually unfolded a structured analysis on customer baskets. We broke down the main research questions into many hypotheses and we attempted to analyse all possible factors in a multidimensional way. All initial hypotheses were accepted except one which was partially accepted and initial intuitions were validated. In table 6.5 we present the initial hypothesis and the current status for each one of them. Hypothesis 3 is partially accepted since beta coefficients for May and June at the 1.0% support level and betas for May, June and September at the 1.5% support level were insignificant. The betas for the remaining months were significant and there seemed to be a seasonal effect, reaching its peak during the summer period.

Table 6.5: Hypothesis Status

Hypothesis	Status
H1: Products of the same product category have high affinity to be purchased together.	Accepted
H2: Products across complementary product categories have high affinity to be purchased together.	Accepted
H3: The seasonality effect on the probability that a product will be included in a frequently purchased basket is significantly different than zero across different months of the year.	Partially Accepted
H4: Discount rate positively influences the probability that a product will be included in a frequently purchased basket.	Accepted
H5: Initial catalogue price negatively influences the probability that a product will be included in a frequently purchased basket.	Accepted
H6: Product purchased quantity positively influences the probability that a product will be included in a frequently purchased basket.	Accepted

7. CONCLUSIONS AND MAIN FINDINGS

7.1 PROCESS OVERVIEW

In the first phase of the analysis, Apriori algorithm is implemented on a tabular database in order to acquire the desirable combination of products that are most frequently purchased together. Next, these results are extensively analysed in order to investigate the products' interrelations. Adding to that, it is realized that only two out of eight business units, 7 out of 45 product categories for the 1.0% and only 5 out of 45 product categories for the 1.5% support threshold, are included into frequently purchased itemsets. Both support thresholds are high in concentration of itemsets with products coming from the same business unit (86.08% and 93.37% for 1.0% and 1.5% support thresholds) as well as products coming from the same product category (67.53% and 77.71% for 1.0% and 1.5% support thresholds). After discussing the results with the management of the company and supporting them with additional analysis, it is confirmed that all of the results are consisted of products from complementary product categories providing us with strong evidence to accept hypothesis 1 and hypothesis 2.

The second phase of our research is dealing with the seasonality effect on the inclusion or not of a frequently purchased itemset in a customer order. At first, our analysis indicates that a strong seasonality effect (on whether an order includes a frequent itemset) is present during all of the summer months and early autumn. This is further supported after getting the binary logistic regression results for our first independent variable and a categorical variable which represents the month that an order is placed. Thus, the initial intuitions from the preliminary research are further supported, leading us to partially accept hypothesis 3. The partial acceptance of hypothesis 3 in based on the fact that beta coefficients were insignificant for May and June at the 1.0% support level and May, June, September at the 1.5% support level. Despite these insignificant results, it is still obvious that seasonality greatly affects the probability that an order includes a frequently purchased itemset. Judging by the probabilities' trend line (table 6.2.5) and the preliminary research on seasonality (chapter 6.2.1), we can indirectly infer -though cannot prove- a strong seasonal effect on the probability for these months.

In the third phase of this analysis two out of three independent variables are transformed into logarithmic ones. The new logarithmic indexes contribute to the creation of a most representing final model (as also assessed by the improved success rates of classification tables for the logarithmic models versus the untransformed ones). Multicollinearity issue is studied by calculating and comparing Pearson correlations and variance inflator factors among the independent variables. As it was initially expected, a moderate to strong multicollinearity issue is detected but it was decided not to deter us from continuing our analysis. Discount rate seems to be highly correlated with logarithmic order quantity and moderately correlated with logarithmic price per piece. Logarithmic price per piece and order quantity are moderately correlated.

After constructing five different versions of logistic models, AIC values are calculated and classification tables are generated. One out of five models is selected as the most efficient one since it combines a high level of AIC value and sensitivity of prediction as well as the highest number of independent variables. This model includes independent variables: Discount rate, LnOrderQuantity and LnPriceperPiece.

Beta coefficients for the logistic regression model are generated and the logistic function of our model is created based on these beta coefficients. Then, probabilities are calculated as a function of

the two logarithmic independent variables and various fixed levels of discount rates. The calculated probabilities are plotted together with the two logarithmic variables in three different 3-D scatterplots, each one corresponding to fixed levels of discount rate (0%, 50% and 95%). These scatterplots together with some examples clearly demonstrate the strong positive effect of discount rate on the probability of a product to be part of a frequent itemset leading us to accept hypothesis 4. Hypothesis 5 is also accepted since it is proved that lower prices translate into increasing probabilities for a product to be included in an itemset. Higher product quantities purchased correspond to higher probabilities for a product, leading us to accept our final hypothesis, hypothesis 6.

It is important to mention that throughout the three stages of our overall analysis, all results are cross-validated by simultaneously running the same analysis for two different support thresholds: 1.0% and 1.5%. No significant deviations are observed between those two cases.

7.2 GENERAL CONCLUSIONS AND MANAGERIAL IMPLICATIONS

The main findings of this research is that customers frequently purchase itemsets which are comprised of products similar and/or complementary to each other, or at least coming from the same or complementary product category in a business-to-business environment. Furthermore, it is more probable for a customer to place orders that include frequently purchased itemsets during the summer period than it is for the winter period. This indicates that season in a B2B setting can determine the purchase of frequent itemsets. Adding to our general conclusions, a product is more likely to be included in a frequently purchased itemset when it is sold in higher discount rates. Products sold in higher quantities are more likely to be sold in a frequent itemset and products with higher prices are less likely to be included in frequent itemsets.

The above findings provide with an insightful way of dealing with cross-selling opportunities in a business-to-business environment. One of the most important conclusions of this study is that a company in a B2B environment can increase its total cross-selling rates and partner dependency not only by improving the trust and commitment levels between partners, but also by adjusting factors directly connected to products and sales such as price, promotional discount rates and time of sale. The above conclusion always refers to the business-to-business setting, a fact that the author regards as innovative based on the content of the existing literature for B2B relationships. In addition, we provide with enough evidence that a cross-divisional orientation is hard to be achieved in a business-to-business environment since most of the products included in frequent itemsets are of the same business unit. There is some space for improvement though throughout careful study of the results and identification of product combinations made of cross-divisional products.

Moreover, seasonality should be taken into seriously account since as we proved, suppliers' actions for improving cross-selling rates may result into differently weighted outcomes based on the season. This implies that a wholesaler can intensify the cross-selling efforts during the seasons with higher probabilities for cross-selling as indicated from a similar analysis. As a result, wholesalers can significantly improve the cross-selling rates for specific periods within the year as well as improve marketing or sales actions' efficiency for these specific months.

The results of this analysis can also be incorporated in the selling process in the form of a recommendation system. This recommendation system can be online in the form of pop-up recommendations in an online purchasing platform. It can also be offline in the form of product recommendations within the hardcover catalogues or recommendations made by the salespersons when closing the deals with the customer.

7.3 LIMITATIONS AND FUTURE DIRECTIONS

One of the most important limitations for this research is that the results should be compared and cross-validated by running similar analyses on various companies and markets in a B2B setting.

Moreover, even though Apriori algorithm is considered to be one of the most important algorithms in the history of data mining so far, it is of lower performance compared to its contemporaries. It would be interesting to investigate whether the same conclusions hold for more advanced algorithms.

Furthermore, we suspect that the basket analysis results were inflated in favour of business units with very high percentage of sales. We noticed that the association rules provided included products of the two business units whose sales account for the 77% of the total sales. Based on this intuition, we understand the importance of running the same analysis, this time excluding products from these two business units. In that way we would discover all association rules for the rest of the business units, utilizing our database to the fullest.

Another important limitation is that seasonality has to do with the market and company attributes. In order to make sure that our conclusions can be generalized, it is necessary to test whether they hold for various companies and different markets in a business-to-business environment.

Moreover, the inclusion of products in frequent itemsets is most probably based on other various explanatory factors which might be irrelevant to the product and customer order attributes. This is also highlighted in our literature review and the theory on cross-selling in a business-to-business environment.

Insights concerning the limitations were also provided after visiting the company and presenting the results. The products most frequently included together in the majority of itemsets were associated in an obvious way since most of them were complementary products. In that direction we assume that from one hand, the majority of rules might be trivial, on the other hand a potential improvement is underlying based on the confidence of the rules. For example, product A is frequently accompanied by a complementary product B of the same colour; as a result we extracted the relevant rule for this pair of products with confidence threshold e.g. 85%. That means that 15% of the times product A is sold, it might be sold together with product C of a different colour than product A's. Nevertheless, we do not have sufficient clues to prove that a strong relationship between product A and C exists. We are also unable to come up with straightforward conclusions for further increase in the profits generated by the sale of product A and B. This issue can be tackled by running the customer basket analysis in a more aggregated level, making generalized inferences about product categories (since product A will be grouped in a different product category than products B and C).

Complementarity of the products included in a frequently purchased itemset in a B2B environment is also inferred by the empirical fact that most of business-to-business relationships are contractual as already mentioned in the literature foundations chapter. Sales are done mainly based on projects with certain needs and strong incentives (in the form of high discount or gifts) are given to the customers in order to reach and/or exceed specific level of sales. Different purchasing relationships can lead to different conclusions. Adding to that, most of the B2B customers operate in very specific

markets with needs for specific products, contributing to the existence of products which are frequently purchased together in an itemset.

7.4 ADDITIONAL COMMENTS

After discussing separately with the product managers, the logistics manager and two vice presidents, we observed that product managers and the logistic manager (lower management hierarchy) were more indifferent in the results presented and resistant to the processes applied during the various steps of this thesis. Vice presidents, on the other side, were much more interested and could better comprehend the concepts of this research as well as draw action lines regarding the implementation and utilization it. Product managers were mostly interested in identifying opportunities across business units and product categories, since this was the only part where they felt more insecure and had less experience on. This can be explained if we take into account the different perspectives and goals of each level of hierarchy; perspectives that also shape the mentality of each hierarchical group. For product managers, this research was mostly informative, mostly providing them with already known information, while for the vice presidents this research could be an improvement that has the potential of adding millions of euros in cash if implemented in the proper way. We must also recognise that cross-selling theory and practices are not widely adopted yet in the business-to-business world even though it is slowly starting to become a trend.

It is also worthwhile to mention that in this specific case study, even though the company is one of the leaders in the Greek market, parts of its purchasing platforms are outdated and most of the times they do not support the implementation of recommendation systems in various steps of purchasing process. Consequently, the results extracted and analysis followed for the purposes of this research, were not considered -at first sight- as an easy-to-adopt way for analysing customer baskets. Further analysis, though, revealed that utilization of the results could be achieved by thoroughly investigating the selling process. This more detailed analysis indicated that information created by our research could be communicated to the sales representatives along with some training material on various cross-selling methods. Adding to that, since most of the company's clients are using either hard-cover or on-line catalogues, the material of this research could be incorporated inside these catalogues.

BIBLIOGRAPHY

1. Agrawal, Rakesh, and Ramakrishnan Srikant. 1994. “Fast Algorithms for Mining Association Rules.” IBM Almaden Research Center 1–32.

2. Akçura, M. Tolga, Zafer D. Özdemir, and Kemal Altınkemer. 2009. “Privacy, Customization, and Cross-Selling of Information.” Journal of Organizational Computing and Electronic Commerce 19(2):112–32.

3. Akcura, Tolga, and Kannan Srinivasan. 2014. “Research Note : Customer Intimacy and Cross-Selling Strategy.” Managemenr Science 51(6):1007–12.

4. Anderson J., Narus J. 1990. “A Model of Distributor Firm and Manufacturer Firm Working Partnerships”. Journal of Marketing 54: 42-58.

5. Andrews, Rick L., and Imran S. Currim. 2002. “Identifying Segments with Identical Choice Behaviors across Product Categories: An Intercategory Logit Mixture Model.” International Journal of Research in Marketing 19(1):65–79.

6. Andrews, Kelly J. 1999. “Extending Up-Selling and Cross-Sellign Efforts”. Target Marketing, 22(11): 36-40.

7. Ansari A., Mela C. F. 2003. “E-Customization”. Journal of Marketing Research 40(2):131-145.

8. Asim, Ansari, Essegaier Skander, and Kohli Rajeev. 2014. “Internet Recommendation Systems.” 37(3):363–75.

9. Berry MJ., Linoff G. 2004. Data mining techniques: for marketing, sales, and customer relationship management. Publisher John Wiley & Sons, Inc.

10. Cannon, Joseph P., and William D. Perreault Jr. 1999. “Buyer-Seller Relationships in Business Markets.” Journal of Marketing Research 36(4):439.

11. Cavoukian A., Jonas J. 2012. “Privacy bu Design in the Age of Big Data”. Information and Privacy Commissioner.

12. Chen M., Han J., and Yu P. 1996. “Data mining: an overview from a database perspective”. IEEE Transactions On Knowledge and Data Engineering 8: 866-883.

13. Cheng H., Yan X., Han J. 2004. “IncSpan: incremental mining of sequential patterns in large database. Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining.

14. Cheung D.W., Han J. 1996. “Maintenance of discovered association rules in large databases: an incremental updating technique”. Data Engineering.

15. Coenen, Frans. 2011. “Data Mining: Past, Present and Future.” The Knowledge Engineering Review 26(01):25–29.

16. Coyles S., Gokey TC. 2005. “Customer Retention in not Enough”. Journal of Consumer Marketing 22(2): 101-105.

17. Dwyer, F. Robert, Paul H. Schurr, and Sejo Oh. 1987. “Developing Buyer-Seller Relationships.” Journal of Marketing 51(2):11.

18. Erich, W., and W. James. 1987. “Cross-Selling : The Unfulfilled Promise”.

19. Eui-Hong, Han, George Karypis, and Vipin Kumar. 2000. “Scalable Parallel Data Mining for Association Rules.” IEEE Transactions on Knowledge and Data Engineering.

20. Fayyad, Usama, Gregory Piatetsky-shapiro, and Padhraic Smyth. 1996. “From Data Mining to Knowledge Discovery in.” 17(3):37–54.

21. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. “The KDD Process for Extracting Useful Knowledge from Volumes of Data.” Communications of the ACM 39(11):27–34.

22. Ford, David. 1980. “The Development of Buyer-Seller Relationships in Industrial Markets.” European Journal of Marketing 14(5/6):339–53.

23. Frazier, Spekman, O’Neal. 1988. “Just-in-time exchange relationships in industrial markets. The Journal of Marketing.

24. Ganesan, Shankar. 1994. “Determinants of Long-Term Orientation in Buyer-Seller Relationships.” Journal of Marketing 58(2):1.

25. Gantz J., Reinsel D. 2011. “Extracting Value from Chaos”. IDC IVIEW. http://www.emc.com/digital_universe.

26. Gupta, Sunil, and Donald R. Lehmann. 2003. “Customers as Assets.” Journal of Interactive Marketing 17(1):9–24.

27. Gupta, Sunil, Donald R. Lehmann, and Jennifer Ames Stuart. 2014. “Valuing Customers.” Journal of Marketing Research 41(1):7–18.

28. Harlam, Bari A., Aradhna Krishna, and Donald R. Lehmann. 1992. “Impact of Bundle Type , Price Framing and Familiarity on Purchase Intention for the Bundle.” Journal of Business Research 2963(1990).

29. Han J., and Kamber M. 2000. “Data mining concepts and techniques”. Morgan Kaufmann Publishers.

30. Hruschka, Harald, Martin Lukanowicz, and Christian Buchta. 1999. “Cross-Category Sales Promotion Effects.” Journal of Retailing and Consumer Services 6(2):99–105.

31. Houtsma M., Swami A. 1993. “Set-Oriented Data Mining in Relational Databases”. IBM Research Report. Data & Knowledge Engineering 17(3): 245-262.

32. Ibm, Agrawal, and Tomasz Almaden Road. 1993. “Mining Association Rule between Sets of Items in Large Databases.” IBM Almaden Research Center 207–16.

33. Jaroszewicz, Szymon, Dan A. Simovici, and Morrissey Blvd. 2004. “Interestingness of Frequent Itemsets Using Bayesian Networks as Background Knowledge.” 178–86.

34. Julander, Claes-Robert. 2014. “Basket Analysis: A New Way of Analysing Scanner Data.” (May).

35. Kamakura, Wagner A. 2008. “Offering the Right Product to the Right Customer at the Right Time.” Journal of Relationship (January 2014):37–41.

36. Kamakura, Wedel, De Rosa. 2003. “Cross-selling through database marketing: a mixed data factor analyser for augmentation an prediction”. International Journal of Research in Marketing 20(1):45-65.

37. Kim, Byung-Do, Kannan Srinivasan, and Ronald T. Wilcox. 1999. “Identifying Price Sensitive Consumers: The Relative Merits of Demographic vs. Purchase Pattern Information.” Journal of Retailing 75(2):173–93.

38. Knox, Simon. 1998. “Segmentation and the Customer Development Process.” European Management Journal 16(6):729–37.

39. Knox S. 1998. “Loyalty-based Segmentation and the Customer Development Process”. European Management Journal 16(6): 729-737.

40. Kumar, V., M. George, and J. Pancras. 2008. “Cross-Buying in Retailing: Drivers and Consequences.” Journal of Retailing 84(1):15–27.

41. Lages, Luis Filipe, Andrew Lancastre, and Carmen Lages. 2008. “The B2B-RELPERF Scale and Scorecard: Bringing Relationship Marketing Theory into Business-to-Business Practice.” Industrial Marketing Management 37(6):686–97.

42. Latusek, Wojciech Peter. 2010. “B2B Relationship Marketing Analytical Support with GBC Modeling.” Journal of Business & Industrial Marketing 25(3):209–19.

43. Lee, Dongwon, Sung-Hyuk Park, and Songchun Moon. 2013. “Utility-Based Association Rule Mining: A Marketing Solution for Cross-Selling.” Expert Systems with Applications 40(7):2715–25.

44. Lynn, M. 1999. “Use Cross-Selling to Increase Client Retention.” National Underwriter.

45. Malms, Oliver, and Christian Schmitz. 2011. “Cross-Divisional Orientation: Antecedents and Effects on Cross-Selling Success.” Journal of Business-to-Business Marketing 18(3):253–75.

46. Manyika J., Chui M., Brown B. 2011. “Big Data: The next frontier for innovation, competition, and productivity”. McKinsey Global Institute.

47. McAfee, Andrew, and Erik Brynjolfsson. 2012. “Big Data: The Management Revolution.” Harvard business review 90(10):60–66, 68, 128.

48. McNicholas, P. D., T. B. Murphy, and M. O’Regan. 2008. “Standardising the Lift of an Association Rule.” Computational Statistics & Data Analysis 52(10):4712–21.

49. Mild, Andreas, and Thomas Reutterer. 2003. “An Improved Collaborative Filtering Approach for Predicting Cross-Category Purchases Based on Binary Market Basket Data.” Journal of Retailing and Consumer Services 10(3):123–33.

50. Milne G.R., Boza G.R. 1999. “Trust and concern in consumer’s perceptions of marketing information management practices”. Journal of Interactive Marketing 13(1):5-24.

51. Mulhern, Francis J., and Robert P. Leone. 2014. “Implicit Price Products : Bundling of Retail A Multiproduct Approach to Maximizing Store Profitability.” 55(4):63–76.

52. Moorman C., Deshpande R., Zaltman G. 1992. “Relationships between Providers and Users of Market Research: The Dynamics of Trust”. Journal of Marketing 57.

53. Naude P, Christopher H. 1996. "Business-to-Business Relationships". 41-56.

54. Natessine, Serguei, Sergei Savin, and Wenqiang Xiao. 2006. “Revenue Management Through Dynamic Cross Selling in E-Commerce Retailing.” Operations Research 54(5):893–913.

55. Natessine S., Savin S. and Xiao W. 2006. “Revenue management through dynamic cross-selling in e-commerce retailing”. Operations research.

56. Oliver C . 1990. "Determinants of Interorganizational Relationships: Integration and Future Directions". The Academy of Management Review, 15(2): 241-265.

57. Piatetsky-Shapiro, Gregory. 1996. “Knowledge Discovery and Data Mining : Towards a Unifying Framework Usama Fayyad Knowledge Discovery and Data Mining : Towards a Unifying Framework 2 KDD , Data Mining , and Relation to Other Fields.”

58. Pfeffer J., Salancik G. 1978. "The External Control of Organizations: A Resource Dependence Perspective". New York: NY. Haper and Row Publishers.

59. Rajaraman, A., Jure L., and Jeffrey D. Ullman. 2014. “Mining of Massive Datasets.”

60. Rodriguez, G. 1978. “Logit Models for Binary Data.” (Revised: September 2007).

61. Rothfeder, J. 2003. “Trend: Cross-Selling”. Ziff Davis CIO Insight.

62. Reichheld, FF, Sasser WE. 1990. “Zero defections: quality comes to services”. Harvard Business Review. 105-111.

63. Scanzoni J. 1979. “Social Exchange and Behavioral Interdependence”. New York: Academic Press, Inc.

64. Schmitz, Christian. 2012. “Group Influences of Selling Teams on Industrial Salespeople’s Cross-Selling Behavior.” Journal of the Academy of Marketing Science 41(1):55–72.

65. Tuli K., A.K. Kohli, and S.G. Bharadwaj. 2007. “Rethinking customer solutions: From product bundles to relational process”. Journal of Marketing 70(3):1.

66. Värlander, Sara, and Ali Yakhlef. 2008. “Cross-Selling: The Power of Embodied Interactions.” Journal of Retailing and Consumer Services 15(6):480–90.

67. Walters, Rockney G. 2014. “Assessing the Impact of Retail Price Promotions on Product Substitution, Complementary Purchase, and Interstore Sales Displacement.” Journal of Marketing 55(2):17–28.

68. Weese, Samuel H. 1997. “Cross-Selling Can Help Bolster Customer Retention.” National Underwriter.

69. Xu M. and Walton J. 2005. “Gaining customer knowledge through analytical CRM”. Industrial Management & Data Systems 105(7): 955-71.

70. Zboja, James J., and Michael D. Hartline. 2012. “An Examination of High-Frequency Cross-Selling.” Journal of Relationship Marketing 11(1):41–55.

71. Zhao, Qiankun, and Sourav Bhowmick. 2003. “Association Rule Mining : A Survey.” Technical Report, CAIS, Nanyang Technological University, Singapore.

APPENDIX A

Table A.1: Example of Final Excel Output Format

Antecedent 1	Antecedent 2	Antecedent 3	Consequent	Basket Type	Support %	Confidence %	Lift %	Profit
Product 5	Product 1	Product 2	Product 6	Quadruplet	1.64	70.22	10.84	24845.30
Product 7	Product 19		Product 9	Triplet	3.55	73.34	10.01	8064.16
Product 8	Product 19		Product 9	Triplet	2.92	70.72	15.19	6886.55
Product 8	Product 16		Product 9	Triplet	2.58	71.50	10.80	6011.44
Product 7	Product 14	Product 16	Product 9	Quadruplet	2.47	70.17	8.87	5852.98
Product 8	Product 14		Product 9	Triplet	2.37	75.06	10.13	5257.45
Product 8	Product 18		Product 9	Triplet	2.18	72.85	9.00	4975.89
Product 18	Product 7		Product 9	Triplet	2.26	75.73	8.04	4972.29
Product 8	Product 7		Product 9	Triplet	2.34	80.15	11.08	4860.88
Product 8	Product 9	Product 14	Product 7	Quadruplet	1.78	70.17	10.20	4848.21
Product 12	Product 9	Product 14	Product 7	Quadruplet	1.75	70.45	11.46	4763.65
Product 7	Product 14	Product 19	Product 9	Quadruplet	2.20	77.53	9.79	4727.28
Product 48	Product 9		Product 7	Triplet	1.82	73.84	9.09	4716.47
Product 18	Product 14	Product 19	Product 9	Quadruplet	1.98	70.21	10.02	4705.35
Product 7	Product 16	Product 19	Product 9	Quadruplet	2.17	77.78	10.15	4647.93
Product 41	Product 7		Product 9	Triplet	2.03	72.62	9.35	4646.23
Product 11	Product 7		Product 9	Triplet	1.95	72.84	9.31	4466.74
Product 7	Product 3		Product 9	Triplet	1.91	71.29	11.22	4465.01
Product 11	Product 9	Product 19	Product 7	Quadruplet	1.65	71.79	9.72	4385.08
Product 8	Product 14	Product 16	Product 9	Quadruplet	1.97	74.92	9.14	4382.70
Product 12			Product 14	Pair	4.86	73.08	9.07	4368.46
Product 8	Product 16	Product 19	Product 9	Quadruplet	1.85	75.82	10.26	4052.91

Figure A.1: Distribution of Number of Products per Order

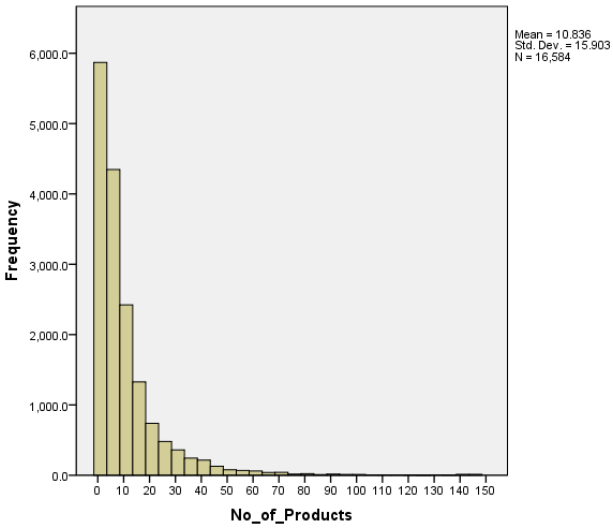


Figure A.2: % of Order Lines per Business Unit

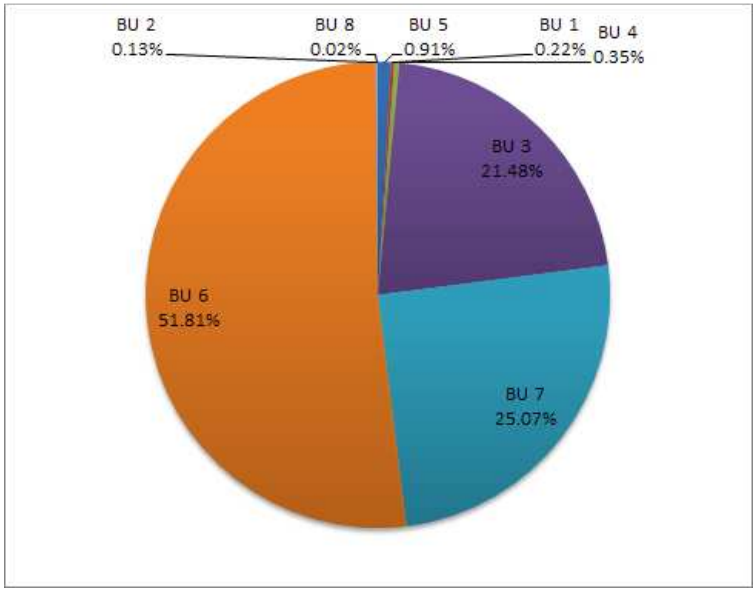


Table A.2: Product Categories (PC) per Business Unit (BU)

BU 5	BU 1	BU 4	BU 2	BU 3	BU 7	BU 6	BU 8
PC 5.1	PC 1.1	PC 4.1	PC 2.1	PC 3.1	PC 7.1	PC 6.1	PC 8.1
PC 5.2	PC 1.2	PC 4.2	PC 2.2	PC 3.2	PC 7.2	PC 6.2	
PC 5.3	PC 1.3	PC 4.3		PC 3.3	PC 7.3	PC 6.3	
PC 5.4		PC 4.4		PC 3.4	PC 7.4	PC 6.4	
PC 5.5		PC 4.5		PC 3.5	PC 7.5	PC 6.5	
		PC 4.6		PC 3.6		PC 6.6	
		PC 4.7		PC 3.7		PC 6.7	
				PC 3.8		PC 6.8	
				PC 3.9		PC 6.9	
						PC 6.10	
						PC 6.11	
						PC 6.12	
						PC 6.13	

Figure A.3: % of Baskets per Combinations of Product Categories (1.0% Support)

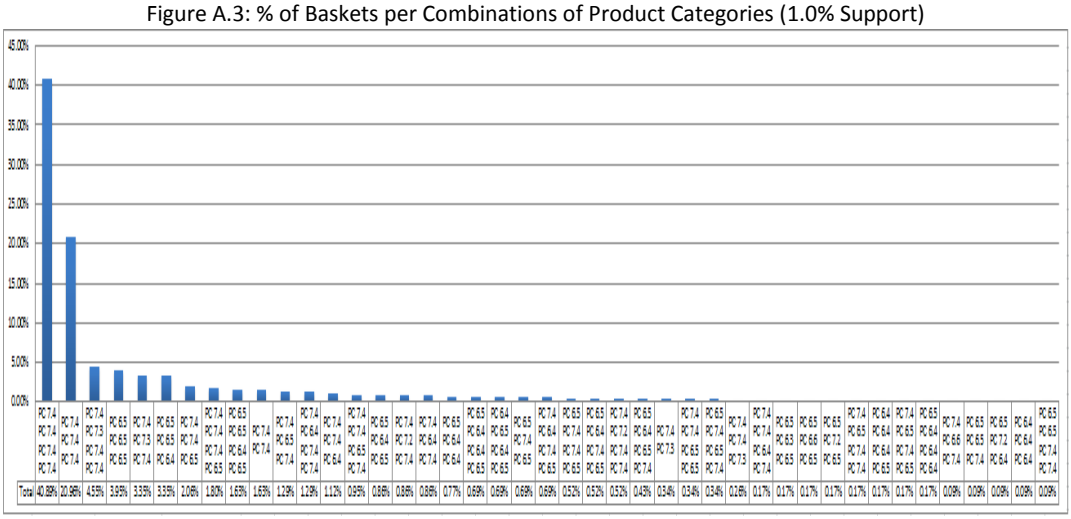


Figure A.4: % of Itemsets per Combinations of Product Categories (1.5% Support)

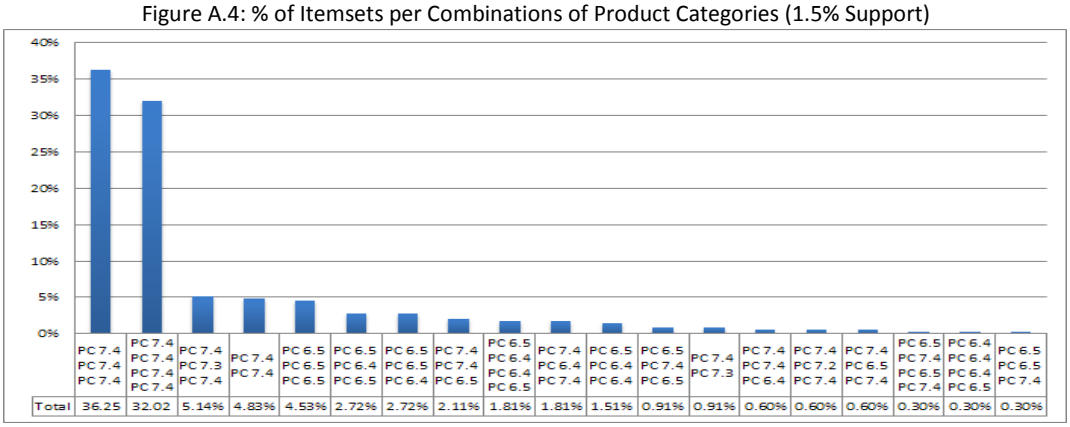


Figure A.5: Support Threshold Occurrence (1.0% Support)

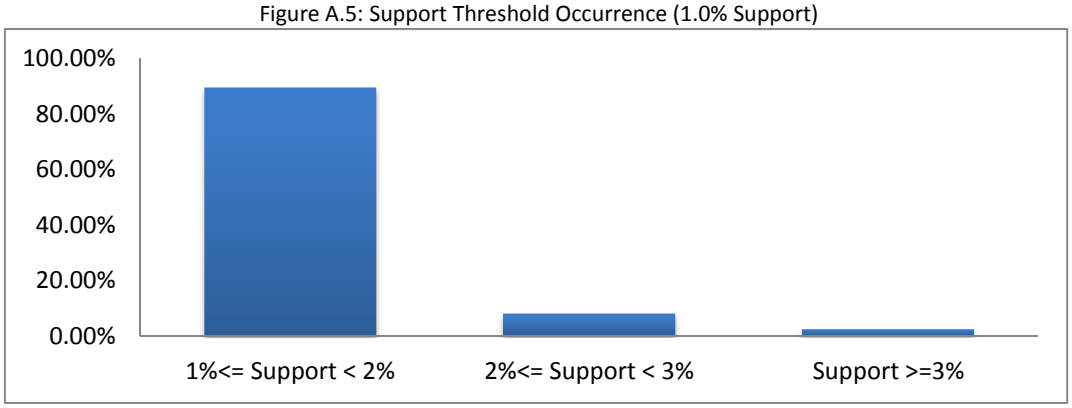


Figure A.6: Support Threshold Occurrence (1.5% Support)

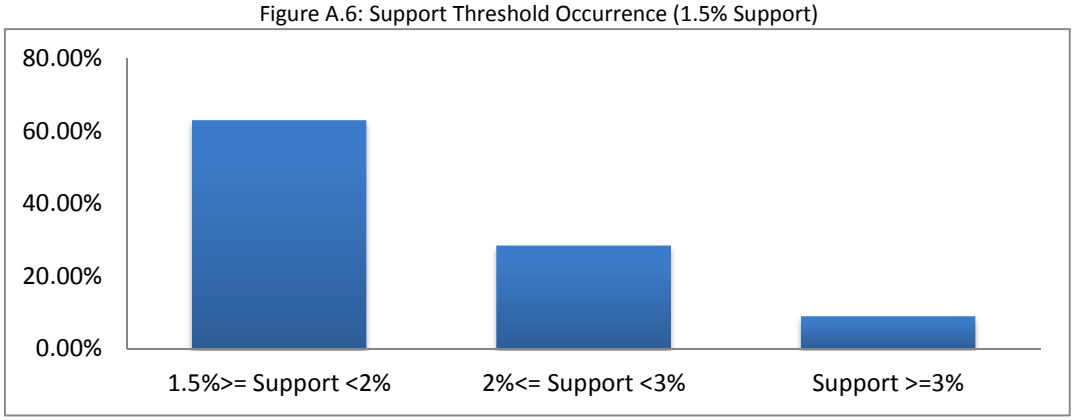


Figure A.7: Confidence Threshold Occurrence (1.0% Support)

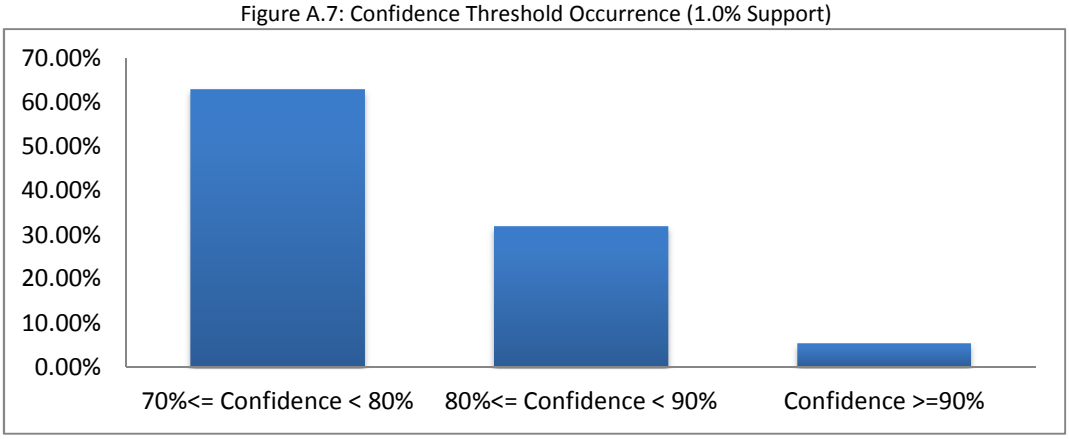
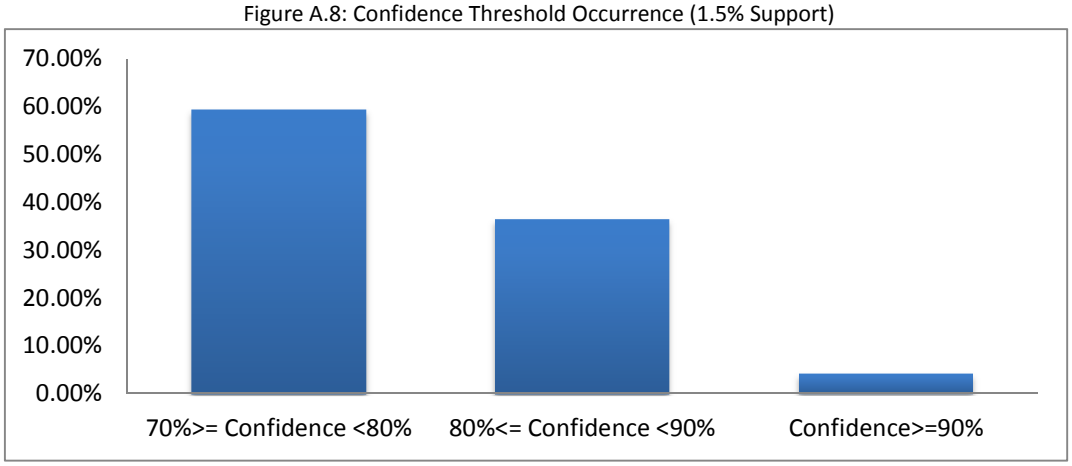


Figure A.8: Confidence Threshold Occurrence (1.5% Support)



Figures A.9: Lift Threshold Occurrence (1.0% Support)

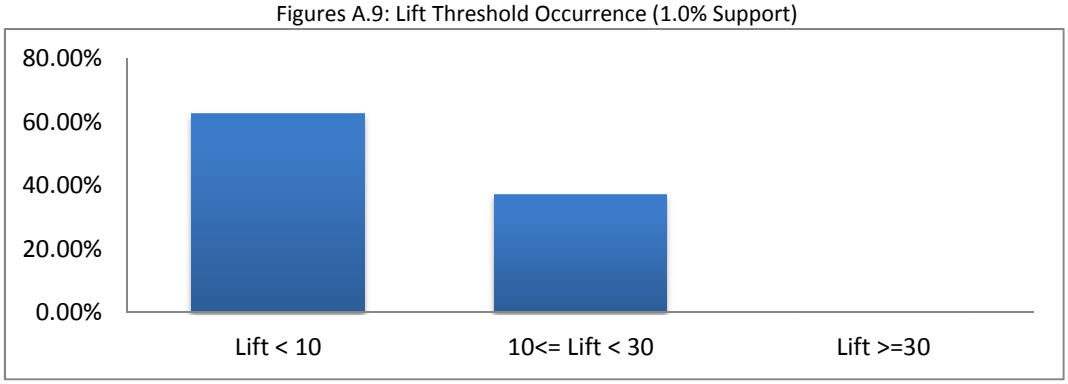


Figure A.10: Lift Threshold Occurrence (1.5% Support)

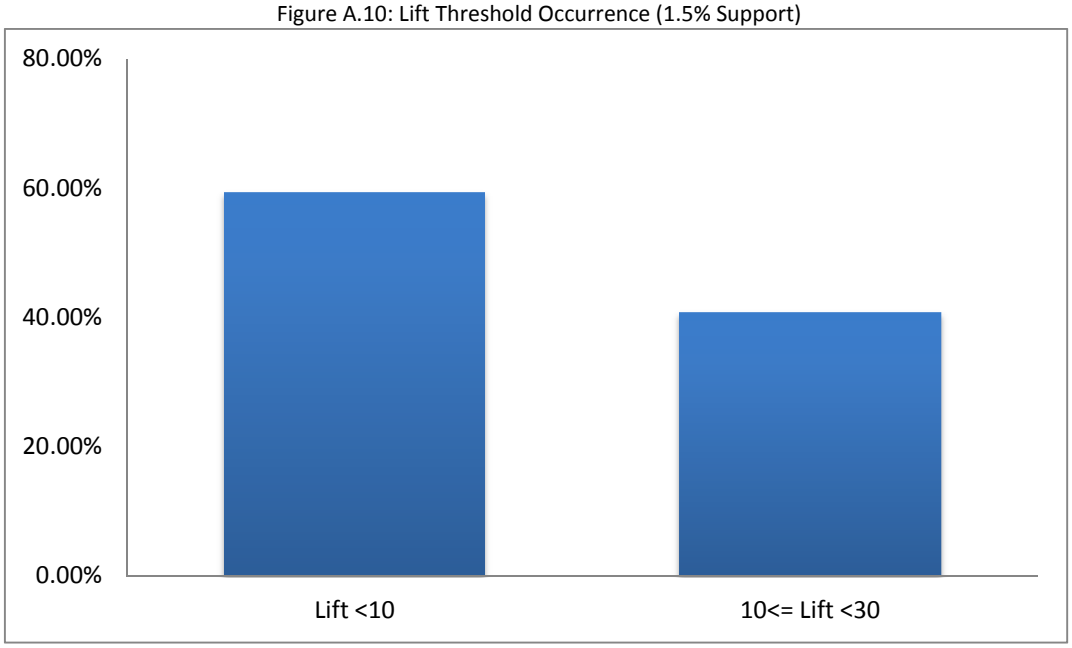


Table A.3: Omnibus Tests of Model Coefficients (1.5% Support)

Chi-square	df	Sig.
297.562	11	0.000

Table A.4: Model Summary (1.5% Support)

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
14379.762	.018	0.30

Table A.5: Classification Table (1.5% Support)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	10090	3812	72.6
	1	1622	1060	39.5
		Overall Percentage:		67.2

*cutoff value: .200

Table A.6: Estimated Beta Coefficients (1.5% Support)

Explanatory Variables	B	Sig.
January	-.403	.000
February	-1.094	.000
March	-.451	.000
April	-.275	.011
May	-.081	.464
June	.052	.622
July	.235	.020
August	.223	.047
September	.175	.103
October	.231	.023
November	-.313	.005
Constant	-1.498	.000

Table A.7: Odds and Prob/ies (1.5% Support)

Month	Odds	Probabilities
January	.149	13.00%
February	.075	6.96%
March	.142	12.47%
April	.170	14.51%
May	.206	17.10%
June	.236	19.07%
July	.283	22.05%
August	.279	21.84%
September	.266	21.03%
October	.282	21.98%
November	.164	14.06%
December	.224	18.27%

APPENDIX B

Table B.1: Linear Regression (OrderQuantity, PriceperPiece)

Model	B	Sig.
Order Quantity	-.148	.000
Constant	82.724	.000

*Dependent Variable: PriceperPiece

Table B.2: R, R^2, Std. Error: Linear Regression (OrderQuantity, PriceperPiece)

R	R Square	Std. Error of the Estimate
0.03326	.001	399.320

Figure B.1: Scatter Plot: PriceperPiece / OrderQuantity

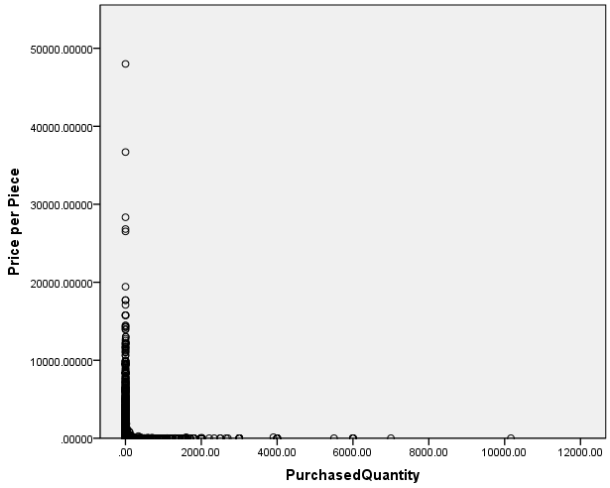


Figure B.2: Histogram: OrderQuantity

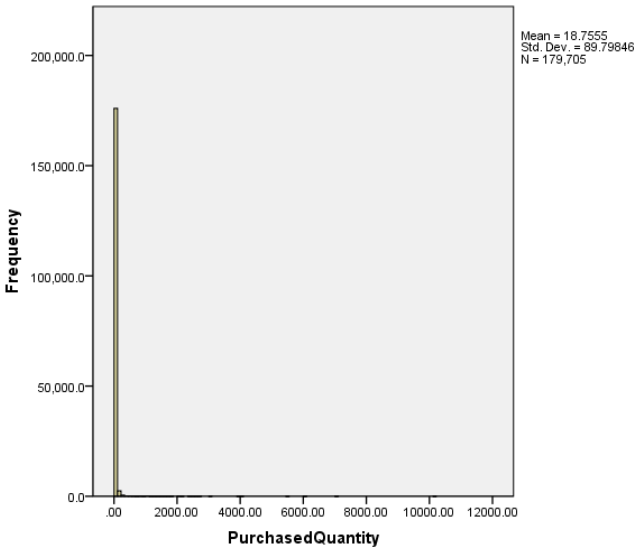


Figure B.3: Histogram: PriceperPiece

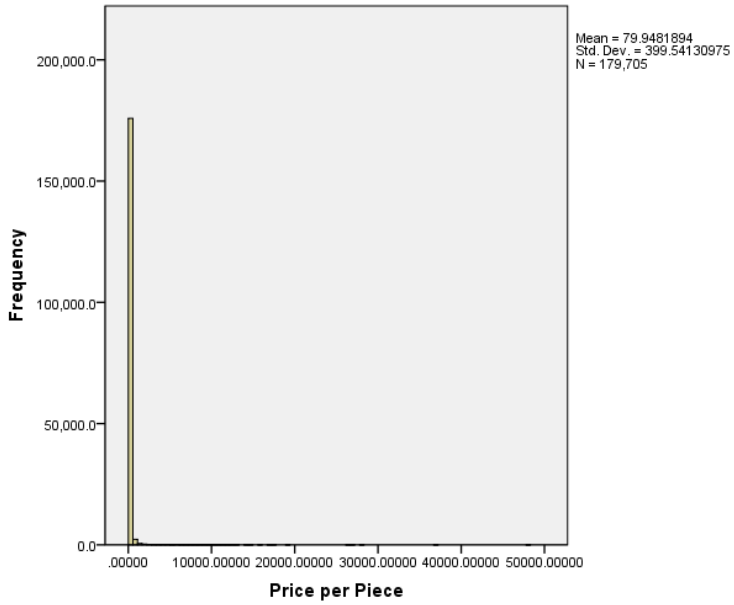


Table B.3: Skewness and Kurtosis: Discount, LnOrderQuantity, LnPriceperPiece

	Discount	LnOrderQuantity	LnPriceperPiece
Skewness	-1.355	.423	0.288
Std. Error of Skewness	.006	.006	0.006
Kurtosis	.963	.031	0.455
Std. Error of Kurtosis	.012	.012	0.012

Table B.4: Classification Table for Model 3 (1.0% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	121970	30183	80.2
	1	9593	17959	65.2
	Overall Percentage:			77.9

*cutoff value: .250

Table B.5: Classification Table for Model 4 (1.0% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	131958	20195	86.7
	1	14722	12830	46.6
	Overall Percentage:			80.6

*cutoff value: .250

Table B.6: Classification Table for Model 5 (1.0% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	126609	25544	83.2
	1	13445	14107	51.2
	Overall Percentage:			78.3

*cutoff value: .250

Table B.7: AIC Statistic Calculation (1.5% Support Threshold)

A/A	Model Explanatory Variables	-2LogLikelihood	No of Predictors+Constant	AIC	Δ(AIC)
1	Discount, OrderQuantity, PriceperPiece	103060.698	3	103066.698	
2	Discount, LnOrderQuantity, LnPriceperPiece	103119.920	4	103127.920	0.06%
3	Discount, LnPriceperPiece	103742.487	3	103748.487	0.66%
4	Discount, LnOrderQuantity	104775.633	3	104781.633	1.65%
5	LnOrderQuantity, LnPriceperPiece	105014.514	3	105020.514	1.86%

Table B.8: Classification Table for Model 1 (1.5% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	141194	18025	88.7
	1	10436	10050	49.1
	Overall Percentage:			84.2

*cutoff value: .250

Table B.9: Classification Table for Model 2 (1.5% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	144508	14711	90.8
	1	12215	8271	40.4
	Overall Percentage:			85.0

*cutoff value: .250

Table B.10: Classification Table for Model 3 (1.5% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	144111	15108	90.5
	1	11858	8628	42.1
	Overall Percentage:			85.0

*cutoff value: .250

Table B.11: Classification Table for Model 4 (1.5% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	143969	15250	90.4
	1	12278	8208	40.1
	Overall Percentage:			84.7

*cutoff value: .250

Table B.12: Classification Table for Model 5 (1.5% Support Threshold)

		Order Has Basket (Predicted)		Percentage Correct
		0	1	
Order Has Basket (Observed)	0	146728	12491	92.2
	1	13101	7385	36.0
		Overall Percentage:		85.8

*cutoff value: .250

Figure B.4: 3-D scatterplot: π_i, LnOrderQuantity$_i$, LnPriceperPiece$_i$, Discount= 0% (1.5% Support Threshold)

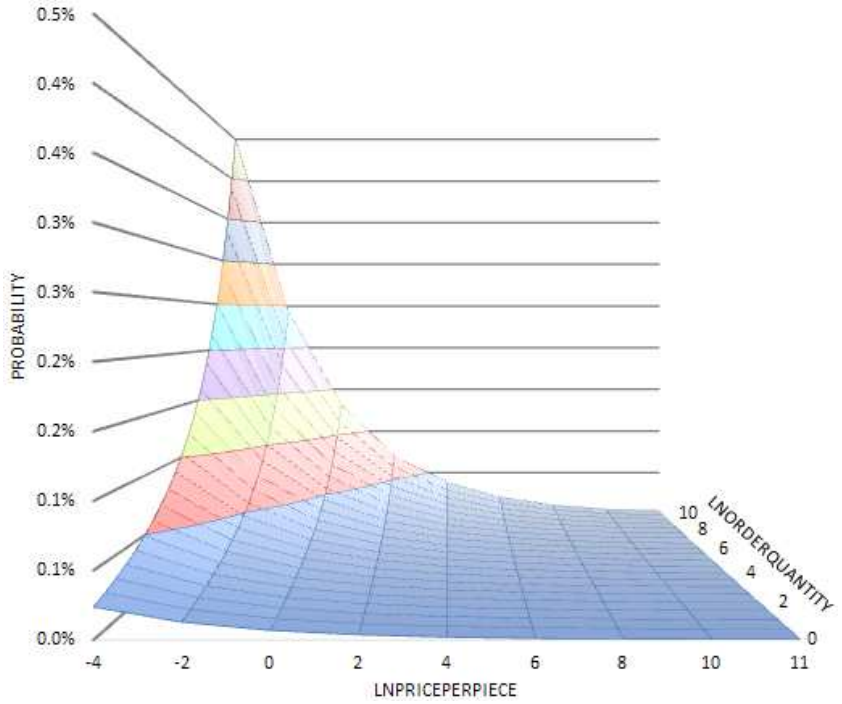


Figure B.5: 3-D scatterplot: π_i, LnOrderQuantity$_i$, LnPriceperPiece$_i$, Discount= 50% (1.5% Support Threshold)

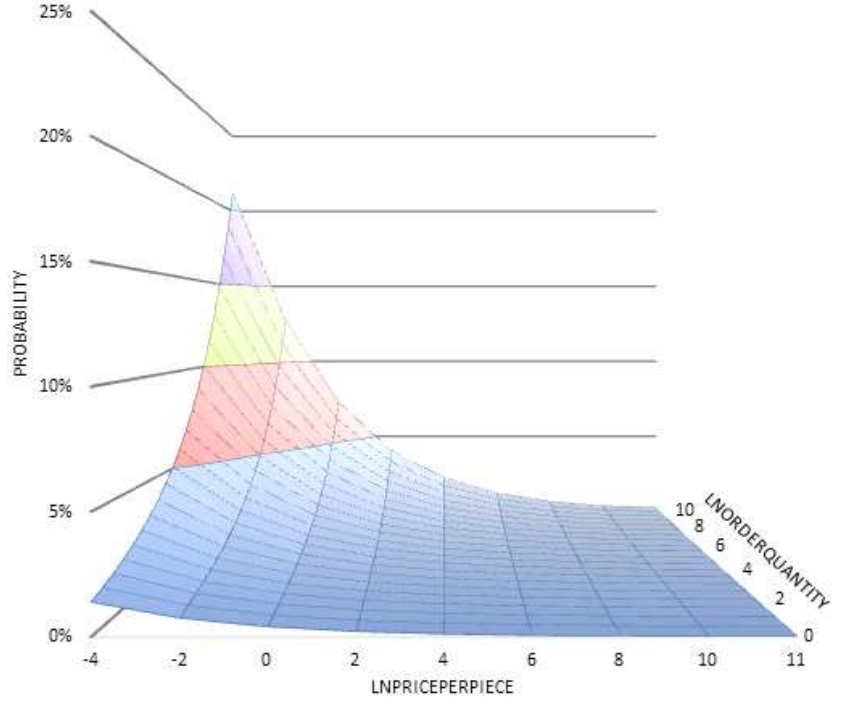
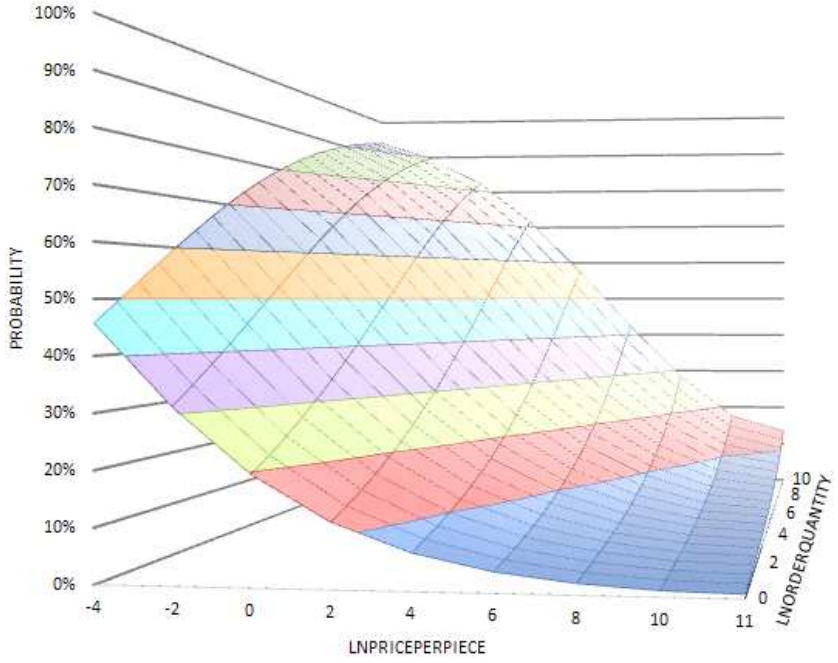


Figure B.6: 3-D scatterplot: π_i, LnOrderQuantity$_i$, LnPriceperPiece$_i$, Discount= 95% (1.5% Support Threshold)



APPENDIX C

```
import csv, sys
from collections import defaultdict

version = '15'
version = '1'

basket_file = 'baskets_%spc.csv' % version
outfile = 'orders_that_have_basket_with_%spc.csv' % version
outfile2 = 'products_that_have_basket_with_%spc.csv' % version
outfile3 = 'products_per_order_thatt_have_basket_with_%spc.csv' % version

allbaskets = []
with open(basket_file, 'rb') as csvfile:
    spamreader = csv.reader(csvfile, delimiter=',')
    for row in spamreader:
        basket = set()
        for i in range(4):
            if row[i]:
                basket.add(row[i])
        allbaskets.append(basket)


orders = defaultdict(set)
months = defaultdict(int)
products = set()
sum_order_quantity = defaultdict(float)
sum_product_price = defaultdict(float)
sum_discount = defaultdict(float)
sum_product_hits = defaultdict(float)
first = True
with open('all_data.csv', 'rb') as csvfile:
    spamreader = csv.reader(csvfile, delimiter=',')
    for row in spamreader:
        if first:
            first = False
        else:
            product = row[2]
            order = row[1]
            products.add(product)
            orders[order].add(product)
            months[order] = int(row[0])
            #row[3,4,5] = Order Quantity,Price per Piece,Discount
            sum_order_quantity[product] += float(row[3].replace(',',''))
            sum_product_price[product] += float(row[4])
            sum_discount[product] += float(row[5])
            sum_product_hits[product] += 1.0


# Re-open the main file and extract the detailed - per order -
# properties and in-basket
with open('all_data.csv', 'rb') as csvfile, open(outfile3, 'wb') as
out_detailed:
    spamreader = csv.reader(csvfile, delimiter=',')
    spamwriter = csv.writer(out_detailed, delimiter=',',
quoting=csv.QUOTE_MINIMAL)
    first = True
    for row in spamreader:
        if first:
            first = False
            spamwriter.writerow(row + ['in_basket'])
        else:
            product = row[2]
            order = row[1]
            mo = orders[order]
            has_basket_product = False
            for basket in allbaskets:
                if product in basket and (basket.issubset(mo)):
                    has_basket_product = True
                    break
            spamwriter.writerow(row + ['1' if has_basket_product else
'0'])
out_detailed.close()

products = list(products)

product_hit = defaultdict(int)
product_in_basket = defaultdict(int)
with open(outfile, 'wb') as csvfile:
    spamwriter = csv.writer(csvfile, delimiter=',',
quoting=csv.QUOTE_MINIMAL)
    spamwriter.writerow(['order', 'months', 'has_basket_product'])
    for order in orders.keys():
        mo = orders[order]
        has_basket_product = False
        all_products_that_are_in_basket_in_this_order = set()
        for basket in allbaskets:
            if (basket.issubset(mo)):
                has_basket_product = True
                all_products_that_are_in_basket_in_this_order =
all_products_that_are_in_basket_in_this_order.union(basket)
        spamwriter.writerow([order, months[order], has_basket_product])
        for product in mo:
            product_hit[product] += 1
        for product in all_products_that_are_in_basket_in_this_order:
            product in basket[product] += 1
```