Profiling online and physical supermarket customers using Factor and Clustering Methods

Eleftheria Matta and George Stalidis[0000-0003-4544-2171]

International Hellenic University, Thessaloniki, Greece
stalidgi@ihu.gr

Abstract. Profiling of customers allows businesses to address their needs with precision and to perform effective marketing actions. Profiling methods can be applied on questionnaire-based surveys or customer history data found in databases or log files. Machine learning techniques are able to capture consumer behavior and automatically perform profiling and targeted marketing actions, while data analytics and statistical analysis methods are more suitable as decision support tools. The aim of this paper was to extract the profiles of supermarket customers from their purchase history, giving emphasis in understanding their behavior and linking data-driven findings with known profiles from marketing theory. The analysis was conducted on a dataset that derived from the purchase records of 61K supermarket customers over a rolling year. Data from both physical stores and e-shop were integrated with demographic data available through the loyalty program of the supermarket chain. The core methods utilized were a combination of Multiple Correspondence Analysis (MCA) and Hierarchical Cluster Analysis on Principal Components (HCPC). These methods were chosen for their excellent ability to discover trends and build easily explainable profiles, as well as to identify clusters based on a large number of qualitative variables. The analysis identified six supermarket customer profiles, which were associated with product preference patterns and features such as level of spending, loyalty and promo-hunting. The profiles extracted by our data-driven methods were associated to profiles documented in consumer behavior research, suggesting potential marketing implications.

Keywords: Profiling, supermarket customer behavior, MCA, clustering.

1 Introduction

The study of consumer profiles has a long history, dating back several decades, with the earliest references to consumer behavior appearing in magazines of the 1950s (Martineau, 1958). The field is rapidly evolving nowadays, following the remarkable progress in machine learning technologies. Profiling methods allow businesses to address the needs of their customers with precision and to perform effective marketing actions. In the field of supermarkets and e-Grocery, competition has led businesses to adopt personalized marketing techniques, develop mobile apps and advanced loyalty programs. The goal of profiling methods is to identify distinct segments within a larger population based on common traits, behaviors, preferences, or demographics. In supermarkets, profiling offers insights into the customer needs and preferences, leading to increased customer satisfaction, higher sales, better inventory management, and a more competitive position in the retail market.

Besides its multiple benefits, profiling also creates many challenges, especially when it involves the collection and analysis of personal data. Individuals may feel uncomfortable with the potential misuse of their personal data and certain groups may receive unfair treatment. A negative public perception of profiling practices can lead to reputational damage for organizations. Finally, overreliance on automated profiling algorithms can lead to decisions that lack human oversight and nuance, revealing the importance of striking a balance between automation and human judgment.

Consumer profiles can vary depending on the information they are built upon, encompassing demographic, psychographic, behavioral (Durga, 2018) or geographic aspects (Jenneson et al., 2022). Methods are divided into two major categories: supervised and unsupervised, while the data can be collected through questionnaire-based surveys or can be originating from databases or log files. Questionnaire-based methods contribute to marketing theory and offer valuable knowledge to marketing planners. However, they have their limitations, such as response/non-response, small samples with high cost, and social desirability bias. They may also have limited depth and context, poor wording or cause fatigue (Choi & Pak, 2005). On the other hand, data-driven methods employing machine learning and data science, offer objective decision making based on empirical evidence, utilization of large datasets, personalization relying on the customer's behavior and preferences and market competitiveness by making data-informed decisions (Erevelles et al., 2016). Machine learning techniques (mainly based on deep learning) are able to capture consumer behavior and automatically perform profiling and targeted marketing actions, while data analytics and statistical analysis methods are more suitable as decision support tools.

The aim of this paper was to extract the profiles of supermarket customers from their purchase history and loyalty program data, giving emphasis in understanding their behavior and linking data-driven findings with known profiles from marketing theory. Our further goal was to develop an advanced customer classification mechanism that provides information to the digital marketing platform used by a supermarket chain. For this purpose, we applied statistical methods from the family of multidimensional data analysis (Benzecri, 1992), namely a combination of Multiple Correspondence Analysis (MCA) and Hierarchical Clustering on Principal Components (HCPC) (Husson, 2017). The chosen data-driven approach involves powerful explorative statistical methods, which are promising for discovering patterns and building profiles, as well as identifying clusters of customers based on their purchasing behavior (Greenacre, 2017).

2 Related work

In this section we briefly review the literature on supermarket customer profiling. Initially, we present methods applied on questionnaire and transactional data. In this work, the metrics extracted from transactions are about the amount spent and the frequency of purchases, focusing on customer loyalty and spending. We subsequently examine studies that explore transactional data more extensively, taking into account product attributes such as their group, price, and whether they were offered as part of a promotional campaign.

2.1 Profiling Supermarket customers through transactions and questionnaires

One of the most fundamental methods applied in the supermarket (SM) sector, is the Recency-Frequency-Monetary (RFM) analysis (Frasquet et al., 2021). Clustering methods are also widely used. In (Hiziroglu et al., 2012), the authors, using transactional data from supermarket clients, evaluated the results of two different clustering methods: crisp clustering and fuzzy clustering and concluded that the latter provides better results. Rokaha B. et al. (2018), utilizing Hierarchical Clustering, distinguish five groups of supermarket customers (High profit: high income and SM expenditure, High standard: Average income and SM expenditure, Low-risk: Low income and SM expenditure, High focus: High income but low SM expenditure, Low-care: Low income and spend as much as their income). Lingras et al. (2005) conducted temporal and non-temporal analysis based on conventional and modified Kohonen self-organizing maps (SOM) - a type of unsupervised neural network. The modified Kohonen SOM created interval set representations of clusters using properties of rough sets. The paper compared the above methods in studying customer loyalty and identified 5 clusters: Loyal big spenders, Loyal moderate spenders, Semi-loyal moderate spenders, Semi-loyal potentially big spenders, Infrequent customers. Although interval set clusters and crisp clusters were similar, the interval set representations of customers provided a warning of potential transition from a more desirable cluster to a less desirable one. Theodoridis & Chatzipanagiotou (2009), using confirmatory factor analysis, distinguished four types of SM buyers in Greece: Typical, Unstable, Social and Occasional, while Mahalakshmi et al. (2020) distinguished seven types of consumers: Lookers, Discount Hunters, Buyers, Researchers, New Customers, Dissatisfied Customers and Loyal Customers.

Loyalty is an important issue in marketing. In the supermarket sector, via a questionnaire-based survey, it was determined that the most significant factors contributing to loyalty included emotional commitment, satisfaction with the environment, and the value offered by the visit experience (Vieira, 2007). Omar et al. (2009) combined geographical and behavioral characteristics and concluded that students studying in big cities are brand-conscious and consider price as an indicator of quality, while students whose university is located away from a large city are Recreational-Shopping consumers. Harris et al. (2017) conducted a survey on customers who made grocery purchases both online and offline. They grouped store and online customers separately, considering their perceptions of the advantages and disadvantages associated with each shopping channel. The cross-tabulation of these customer clusters suggested that the decision to shop online or in-store might not necessarily hinge on the perceived advantages of one channel over the other, but rather on the desire to avoid the greater disadvantages of the alternative.

2.2 Profiling Supermarket customers based on purchased products characteristics.

In their research, Oliveira & Ara (2022) adapted a modified RFM model and used Gaussian mixture models to cluster the data. The RFM model was enriched with the average item price, the ratio of items on sale and diversity. The average item price was an indicator of how premium a customer was, while diversity was measured as the number of different product categories in a transaction. The segmentation resulted in six customer profiles: frequent, specific, regular, opportunity, prime, and large shoppers. Nguyen (2021) designed a segmentation model based on a combination of a deep neural network which attempts to compress the information of the input variables into a reduced dimensional space, and a self-supervised probabilistic clustering technique. His results showed four clusters: Customers who mostly buy daily groceries and fresh food, Non-food cosmetics customers, customers that make small purchases (mainly convenience products) and customers who buy canned food, processed food, beverages and confections.

Dogan et al. (2021) highlighted that boundary data which are close to more than one segment may be assigned incorrect classes. So, they proposed an intuitionistic fuzzy clustering algorithm applied to supermarket consumers' data, according to the amount spent in eight main product categories. The results indicated that the intuitionistic fuzzy c-means produces more reliable and applicable marketing campaigns than conditional fuzzy c-means and k-means segmentation methods. Focusing on efficiency in big data, Huang & Zhou (2017) designed a parallel algorithm of k-means based on Spark and validated it with sales data of a supermarket. By using the distributed system parallel computing, they improved the execution efficiency of the massive data operation.

(Lingras et al., 2014) implemented an iterative meta-clustering through granular hierarchy of supermarket customers and products. Information retrieved from transactional data about customers and products was represented as static granules. Subsequently, clustering was applied separately to both static granules. The proposed algorithm feeds the clustering profiles from one level of granularity to augment the information granules at the other level of granularity and vice versa. Also, in a later research (Lingras, 2015) performed a supermarket customers' clustering utilizing Kohonen neural networks, using as criteria the numbers of categories, subcategories and items the customers purchased as well as the value of groceries, the number of visits and discounts. The time series values of these six variables, over a thirteen-week period, were employed to represent the customers. Comparing the results to their previous work (Lingras & Young, 2001) concluded that the value of groceries provides an indication of spending potential, the number of visits is a reasonable surrogate of customer loyalty and discounts represent the value consciousness of the customer.

3 Methods

3.1 Overview of analysis process

The aim of the analysis process applied in this paper was to synthesize customer profiles by extracting patterns from their purchase history and identifying the main factors that explain their behavior. The input data were the log files of purchase history from both physical stores and e-shop, integrated with demographic data available through the loyalty program of the supermarket chain.

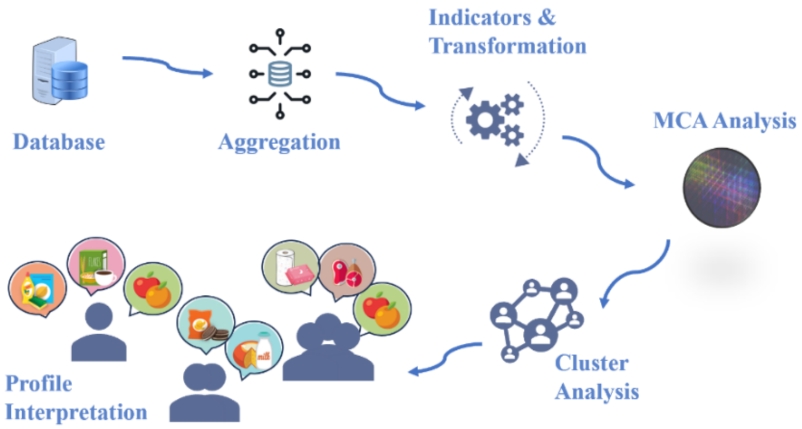

Fig. 1. The data analysis process scheme.

Our approach initially involved identifying distinct customer profiles based on their product group preferences. Subsequently, the elucidation of these profiles was enriched with demographic information and distinctive behavioral indicators, encompassing their preferred shopping channels (physical stores and e-shop), the temporal patterns of their shopping activities, their chosen payment methods, their dispositions towards private label products, the proportion of food-to-non-food product selections, as well as the ratio of products purchased on promotion. The customers' purchase history was analyzed by combining factor and cluster analysis: explorative factor analysis using MCA uncovered shopping trends and profiles of purchase preferences, while Hierarchical clustering using Benzecri's chi square distance and Ward's linkage criterion was then applied to group customers with similar behavior. It should be noted that our analysis was primarily centered around product groups rather than individual items. This approach was driven by the supermarket chain's marketing objective, to identify the product groups that customers preferred when shopping at the particular chain, rather than focusing on specific brand preferences. Following that, the identified clusters were linked to preferences for specific product groups and various behavioral variables, which enabled the formulation of comprehensive customer profiles. Ultimately, marketing experts provided interpretations of these profiles, drawing connections to profiles documented in consumer behavior research and discerning potential marketing implications. The data analysis process scheme is presented in Fig. 1.

3.2 Data collection and preparation

The dataset was derived from a rolling year's sales (1/4/2021–31/3/2022) of a large supermarket chain in Greece. The reference period starts 15 months after the outbreak of the Covid-19 pandemic. At that time, businesses and schools had reopened after the quarantine and the normal daily activity of citizens was slowly being restored. For the purposes of the analysis, raw sales data were captured from both physical stores and the company's eShop, integrated with the loyalty progam data and aggregated at customer level. The supermarket's clientele consisted of hundreds of thousands of customers. Our dataset was a small sample of the company's clientele that consisted of 61.895 individuals (physical store, online and hybrid customers). The number of baskets in the dataset was in the order of 2 million.

Table 1. Active Variables

Variable Name	Description	Variable Name	Description
Cold Cuts UnP	Bulk cold cuts	**CannedFood**	Canned food
Cold Cuts P	Packaged cold cuts	**Creams Pastry R**	Creams and pastry from the refrigerator
Refreshments	Refreshments	**Butchery**	Bulk butchery
Home Accessories	Home accessories	**Butchery Frz**	Frozen packaged meat
Dish Deter	Dish detergents	**Vegetables Frz**	Frozen vegetables
Cloth Deter	Laundry detergents	**Greengrocery**	Greengrocery
Bakery P	Packaged bakery	**Cookies Snacks**	Cookies and snacks
Egg Butter	Eggs and butter	**Alcohol Drinks**	Alcohol drinks
Milk and substit R	Milk and substitutes of milk from the refrigerator	**Diapers Sanit Napkins**	Diapers and sanitary napkins
Yogurt	Yogurt	**Breakfast**	Breakfast
Pers Hygiene	Personal hygiene products	**Oil Vinegar Sauces**	Oil, vinegar and sauces
Household	Household products	**Cheese P**	Packaged cheese
Food Dough Frz	Frozen food and dough	**Cheese Unp**	Bulk cheese
Ready Meals	Ready meals	**Bakery Fresh**	Fresh bakery
Confectionery	Confectionery	**Papers**	Papers
Sugary	Sugary	**Juices Water**	Juices and water
Pasta Pulses	Pasta and pulses	**Beverages R**	Beverages from the refrigerator
House Cleaners	House cleaners		

The initial step in data preparation involved the integration of data from both the physical store database and the e-Shop. The purchase data were then aggregated at customer and product group levels, taking into account both quantity and value. Additionally, a set of indicators was assessed for each customer, including preferences such as the type of store they typically visited and their propensity to purchase Private Label products, among others. It's worth noting that the product categorization hierarchy and the formulation of customer indicators were based on the existing work carried out by the company's analytics department.

Table 2. Supplementary Qualitative Variables

Variable Name	Categories
customerShopType	eCustomer/ Hybrid/ StoreCustomer
Employee	Employee/ Not Employee
PLCustCategory	DislikePL/ IndifferentPL/ Like PL/ Really Like PL/ PL Lovers
TimeOfDayCategories	Morning/ Noon/ Evening
DailyWeekend	Daily/ Weekend
PaymentType	Cash/ Plastic/ Other
AppUser	AppUser/ Not appUser
DiscountType	Leaflet/ Instore/ TV/ No Discount
StoreType	GrandSM/ LargeSM/ MediumSM/ SmallSM
Prefecture	Central Greece/ Central Macedonia/ East Macedonia and Thrace/ West Macedonia/ North Aegean/ Thessaly/ eshop
fnfc	Food Customer/ Non Food Customer/ Balanced FNF
CustPremStatus	Very Premium/ Medium Premium/ Not Premium
biocustomer	BioCustomer/ Not Bio Customer
HaveChild	Yes/ No
HaveElders	Yes/ No
PromoHunter	Very PH/ More than normal PH/ Normal PH/ Not PH
CatBasket	< 50/ 50 and 99.99/ 100 and 149.99/ 150 and 199.99/ 200 and 249.99/ 250 and 349.99/ 350 and 499.99/ > 499.99

After several cycles of experimentation in refining the feature set to be used, we concluded to a set of 52 qualitative variables:

(a) 35 qualitative variables we utilized to capture each customer's purchasing history with respect to the products they predominantly buy (Table 1). Each variable corresponded to a specific product group and derived from the total quantity of items purchased by the customer within that product group. These product quantities were transformed into attraction coefficients, normalized based on each customer's purchases and all customers' purchases collectively, and then categorized into three levels. As a result, we obtained 35 qualitative product preference variables. Level 1 indicates that a customer buys significantly fewer or no items from a particular product category, level 2 signifies average quantities, and level 3 indicates a distinct preference, meaning larger quantities compared to other products in their basket and to other customers. These variables serve as active variables in both Multiple Correspondence Analysis (MCA) and cluster analysis.

(b) 17 qualitative variables depict the shoppers' characteristics. These variables were generated from statistical indicators computed from the customer's purchasing history and are converted into categorical variables using thresholding. The variables and their respective categories are presented in Table 2.

The descriptive statistics showed that the customers in the dataset shopped on average 33 times per year and their average basket value was 22.30€. They purchased 1 out of 4 products on promotion and 18% of the products they bought were Private Label. Also 96.8% of them purchased only from physical stores, 1% purchased only online and 2.2% from both channels. In addition, 55.8% of them prefer to pay by credit card, 41.5% in cash and 2.7% with other types of payment. They preferred to shop at noon by 55%, in the morning by 24% and in the evening by 21% and 10% of them purchased only on weekends. The tech savvy customers that were using the company's application reached 5%. The non-food customers made up 31%, the food customers 33% and the balanced between food and non-food products 36%. About 60.3% of the customers were characterized as moderate premium, 21.4% were very premium and 18,4% not premium at all. Finally, 30% of the customers were very promo hunters, 31.5% more than normal, 29.5% normal and 8.8% were not promo hunters.

3.3 Factor and cluster analysis

The core method utilized in our research was Multiple Correspondence Analysis (MCA), a dimensionality-reduction method, similar to factor analysis (FA) and Principal Component Analysis (PCA), that is not limited to quantitative variables but is particularly suitable to datasets with a large number of categorical variables (Greenacre, 2013). It is a highly intuitive, graphical method for estimating and visualizing complex relations among qualitative features, as well as discovering trends in customer behavior (Manca et al., 2018). MCA is commonly used to analyze data from surveys (Husson et al., 2017) but can also be applied to a wide range of datasets of different nature, including logs and any type of frequency data. While it has been applied to marketplace data for profiling problems (Bejaei et al., 2020), the application of MCA to supermarket purchase history data is limited (Stalidis, 2019). The analysis was applied on the generalized contingency table (Burt) that was formed from the product group variables. The behavioral variables were used as supplentary variables, i.e. they did not participate in the estimation of the factors but were only projected on the factorial planes in order to depict associations among product preferences and bahavioral indicators.

Hierarchical Cluster Analysis on Principal Components (HCPC) was applied on the results of MCA for further analysis and visualization in order to explore patterns or groupings in the data based on the dimensions created by MCA. HCPC grouped individuals who shared similar characteristics according to a set of complex variables and built a tree structure that showed how individuals were progressively grouped (Philippe et al., 2019). The clusters were then projected on the factorial planes and were associated with purchase behavior, revealing customer profiles. The analysis was performed using the FactoMineR R package.

4 Analysis results

4.1 MCA Results

The inertia distribution of the MCA results showed that the 1st factor expressed 33.4% of the total inertia of the analyzed table. The 2nd factor explained 4.7%, the 3rd one 3.9%, and the 4th one 2.6% of the total inertia. The empirical criterion for selecting the number of factors with useful information (inertia percentage greater than the 0.95-quantile of the inertia percentages distribution obtained by simulating n>500 data tables of equivalent size on the basis of a uniform distribution), suggested to consider up to the 7th factor. The first 7 dimensions accounted for 50.2% of the total inertia, which is considered satisfactory. It was notable that a very large percentage of intertia was concentrated on the 1st dimension, whereas all other dimensions expressed single digit, smoothly distributed percentages. This was an indication that the 1st factorial axis reflected a foundamental phenomenon, while the other axes expressed finner contrasts in customer behavior.

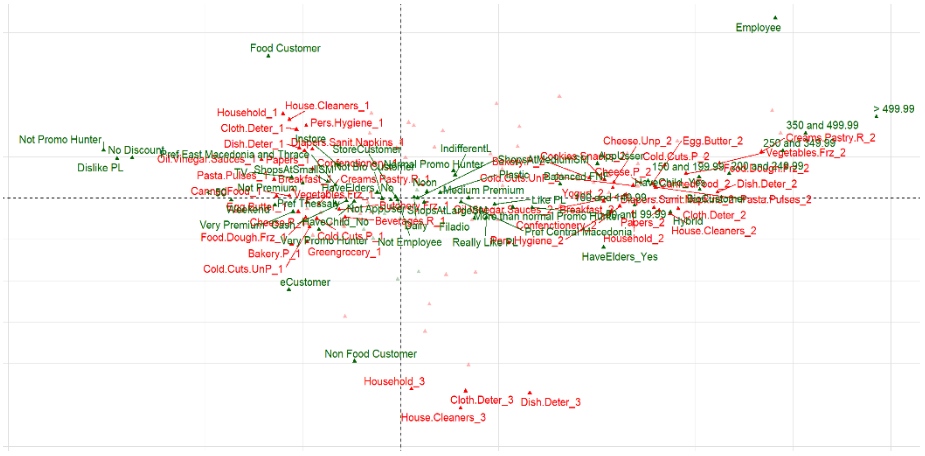

Fig. 2. Representation of active and supplementary categories on the plane F1 x F2.

The factorial plane F1 x F2 is illustrated in Fig.2. It is noted that categories with suffix “1” correspond to no purchasing or purchasing in small quantities of a product group, suffix “2” corresponds to average purchasing, and suffix “3” high purchasing. In this figure, the categories of product group variables are printed in red font, while the categories of supplementary variables in green.

On the left side of the diagram, we observed the low purchasing categories (i.e. level 1) of all the product group variables, while on the right side the corresponding average purchasing categories. It was clear that the 1st factor (33.4% of inertia) expressed the contrast between limited occasional buys and the average purchases of regular customers. By examining the supplementary variables, we discovered that the occasional profile was associated with no preference for private label, and with being either very promo hunter or not at all promo hunter. In contrast, the regular profile was linked to a preference for private labels, a higher-than-average inclination towards promotional offers, a medium-to-premium spending pattern, having elder family members, and being a mobile app user. It was also noticeable that along this factor towards the right direction, the level of spending escalated from 100-150€ per month (at the center of the regular profile) up to more than 500€ per month at the extreme edge of the axis. Along the vertical axis (2nd factor - 4.7% of inertia) high purchasing of **non-food** product groups (bottom side) are juxtaposed from **food** product groups (top side). The 2nd dimension was thus the food vs non-food factor.

From a business perspective, the factorial plane F1 x F2 as a whole, including the supplementary categories, was interpreted as follows: The first group (top left) was the profile of those who purchase small quantities, appear to be food customers with small monthly baskets below 50€, that shop mostly on weekends from small stores and pay mostly in cash. They tend to be either very premium or not at all. The same applies to promotions, they are either very promo hunters or not at all. A few of them that are not promo hunters also show no preference for private label products. In this group a few individuals were purely e-customers that did not purchase a lot and had not visited brick and mortar stores in the examined period. If they purchased products with discount, their preferable discount was TV and instore offers. They were mostly residents of the prefectures of East Macedonia, Thrace and Thessaly. They were not employees of the supermarket chain and they did not purchase products for kids or elders.

The second group (top right) purchased larger quantities of food and non-food products. They appeared to be balanced FNF meaning that ±80% of their amount was spent on food and ±20% on non-food categories and a lot of them had children. Children might be the reason why many of them purchased organic products. Also, many of them were app users suggesting that this group is tech savvy and probably not very old. In addition, in this profile, customers tend to be hybrid, meaning that they purchase from brick-and-mortar stores, but they also order from the company's e-shop. They also tend to buy products for elders; these people might be old or younger family members who shop on their behalf. Their average baskets are between 100€ and 350€ per month and increase as we move far from the axis origin to the right. Employees seem to be great customers with large baskets spending more on food product categories. The third group (bottom) consists of non-food customers that purchase significant quantities of household products, cleaners, laundry and dish detergents.

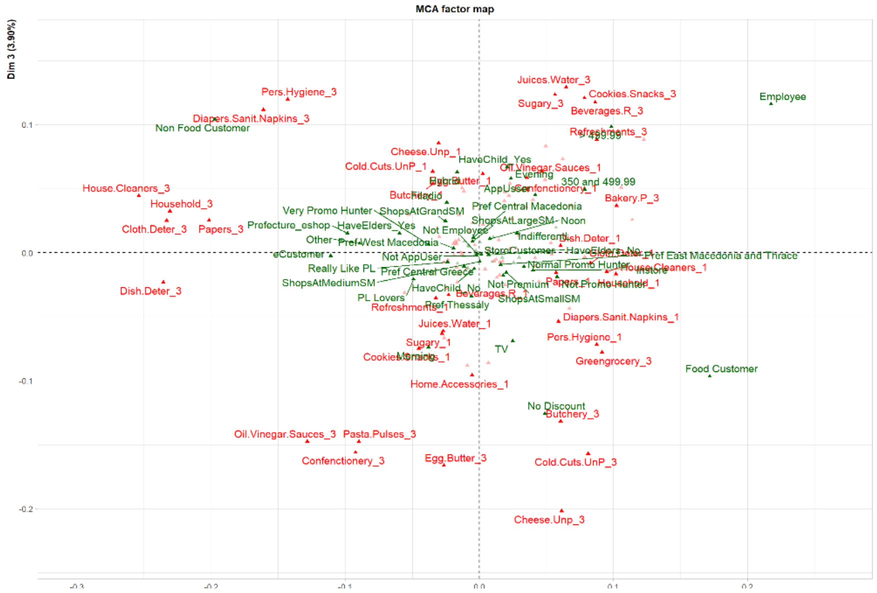

Fig. 3. The factorial plane F2 x F3 associates the food vs non-food factor with the cooking vs ready-made factor.

In Fig. 3, the 2nd factor is projected together with the 3rd factor, explaining in total 8.1% of inertia. While the 2nd factor (x-axis) juxtaposes the purchasing of non-food categories (left side) with food categories (right side), the 3rd factor (y-axis) juxtaposes the purchasing of ingredients for cooking (bottom side) from preference for snacks and ready meals (top side). The 3rd factor could be labeled as the home cooking vs ready-made. The group formed at the top right of the scheme consists of preference for snacks and ready-made food, high spending, tendency to shop in the evening, usage of the company's application, indication that the customer has children and is employee of the supermarket chain. The group formed at the top left shows customers that purchase non-food products like cleaners and personal grooming products. The group formed at the bottom left indicates the preference for grocery categories for cooking, whereas at the bottom right we found the profile of customers that buy fresh food categories like butchery, bulk cheese and cold cuts and greengroceries. They often purchase butchery and cold cuts advertised on TV with discount but due to inelastic demand, they seem to purchase these categories even without discount. Both customer profiles at the bottom side of the 3rd factor can be characterized as those who prefer to cook.

Continuing the interpretation of factorial planes, up to the 7th dimension, we found that the 4th factor (2.6% of inertia) juxtaposed the preference for bulk cheese and cold cuts from the packaged ones, while the 5th factor (2%) juxtaposed the purchasing of breakfast products from fresh food and alcohol. The 6th factor (2%) differentiated the preference for dairy products like milk and yoghurt plus cold beverages from alcohol, refreshments and groceries, and finally, the 7th factor (1.6%) differentiated the preference for frozen meat and vegetables from cheese, cold cuts and cold beverages. The interpretation of the factorial planes formed by the above factors revealed a few more interesting profiles associated with preferences for certain combinations of products. These profiles were however difficult to interpret from a business perspective.

4.2 HCPC Results

The optimal number of clusters was defined by the inertia gain at the point where it starts to decrease with a slower rate. In our research, although the suggested number of clusters was 3, we chose to analyze 6 clusters, in order to delve deeper into the differences among the clusters. Table 3 displays the descriptive statistics for the 6 identified clusters.

Table 3. Cluster descriptive statistics.

Cluster	Cust Count	%Cust	Total Transactions	Monthly Basket	%Items on promotion	%Private Label Products
1	11.661	19%	4,8	23,3	39,5%	15%
2	13.444	22%	16,2	45,6	24,7%	18%
3	9.334	15%	33,7	73,9	23,3%	18%
4	10.992	18%	45,1	85,8	23,1%	18%
5	6.833	11%	26,0	76,5	25,0%	20%
6	9.631	16%	73,5	148,5	25,2%	19%
Total	61.895	100%				

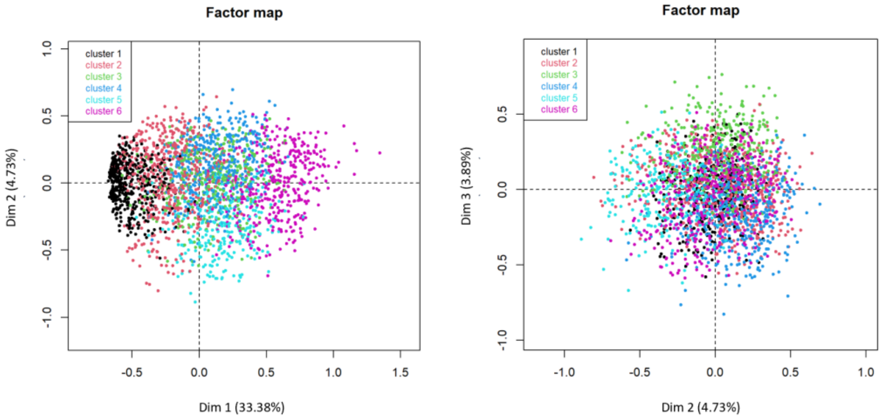

Fig. 4. Projection of clustered individuals on the factorial planes.

In order to associate clusters with customer profiles, the clusters were projected on the factorial planes F1 x F2 and F2 x F3. As mentioned above, factor 1 (x-axis in Fig.4a) separated limited occasional purchases from average regular purchases, with increasing level of spending, and factor 2 (y-axis in Fig 4a and x-axis in Fig. 4b) separated high purchasing of non-food vs food products. Factor 3 (y-axis in Fig. 4b) separated preference for cooking materials vs ready-made products. Keeping these in mind, as well as the clusters' characteristics from Table 3, we developed an interpretation for each cluster. Notice that clusters 3, 4 and 5 overlap in factorial plane F1 x F2 but are clearly distinguished in F2 x F3.

Cluster 1: **Not regular customers who only make a few occasional purchases**, either food or non-food products. They seem not to care about private label and are a mixture of customers with a strong inclination towards seeking promotions and customers showing no interest in them at all. Cluster 1 comprised 19% of the sample.

Cluster 2: **Low spenders who prefer bulk cheese and cold cuts**. Cluster 2 comprised 22% of the sample.

Cluster 3: Moderate spenders who prefer snacks over cooking. They prefer packaged cheese, packaged cold cuts, cookies & snacks, home accessories, cold beverages and refreshments. Their preferable discount items seem to be the ones advertised on the company's leaflet. They prefer to purchase snacks and light food over cooking. Cluster 3 comprised 15% of the sample.

Cluster 4: Moderate spenders, mostly food customers who seem to cook. They purchase greengroceries, eggs, butter, confectioneries among other food categories from grocery. The food categories they purchase can be either fresh, packaged or frozen. Cluster 4 comprised 18% of the sample.

Cluster 5: Moderate spenders, non-food customers. They purchase house cleaners, laundry & dish detergents, household products, diapers and sanitary napkins plus papers and personal hygiene products. Cluster 5 comprised 11% of the sample.

Cluster 6: High spenders, balanced FNF customers (±80% of their amount is spent on food and ±20% on nonfood categories) who purchase all the product groups. Cluster 6 comprised 16% of the sample.

5 Discussion

The primary objective of this study was the identification of supermarket customer profiles. We utilized MCA and multidimensional clustering that are not commonly applied on this kind of data but are highly promising in discovering interpretable behavior patterns. Our analysis was performed on a dataset consisting of 61.895 supermarket customers' purchases within a rolling year. The method was applied to the product groups the customers purchased and was enriched with the shoppers' characteristics, provided by the supplementary variables. The MCA results were subsequently used in the Hierarchical Cluster Analysis on Principal Components (HCPC). The HPCP identified six clusters based on the dimensions created by MCA, specifically, (1) Occasional customers who dislike PL products and are very or non promo hunters, (2) Low spenders who prefer bulk cheese and cold cuts, (3) Moderate spenders who prefer snacks over cooking, (4) Moderate spenders mostly food customers who prefer cooking, (5) Moderate spenders that are non-food customers, and (6) High spenders that are balanced FNF customers.

Our findings are comparable with numerous previous studies, align with the existing literature and shed more light on the products that different segments prefer to purchase and how these are related to behavioral features. In (Lingras et al., 2005) the cluster “infrequent customers” is analogous to our “Occasional customers” and the cluster “Loyal big spenders” corresponds to our “Big spenders”. (Theodoridis, P. K., & Chatzipanagiotou, K. C., 2009) like us, use the term “Occasional’. Mahalakshmi et al. (2020) identified a segment named “Offer Hunters” which is a quality that partly characterizes our “Occasional”, while their “Loyal Customers” correspond to our “High Spenders”. In (Oliveira & Ara, 2022) the cluster “Specific” which consists of customers that shop selected items and have low on-sale item ratio, is similar to our “Low spenders” who purchase specifically bulk cheese and cold cuts and also present low “items on promotion” percentage. Also, their “Large Shoppers” correspond to our “High Spenders”. The clusters identified in (Nguyen, 2021) bare many similarities to ours. The first one that consists of customers who mostly buy daily groceries and fresh food corresponds to our cluster 4 that is “Moderate spenders mostly food customers who prefer cooking”. The “Non-food cosmetics customers” corresponds to our Cluster 5 that is “Moderate spenders, non-food customers”. The “customers that make small purchases (mainly convenience products)” correspond to our cluster 2 and “customers who buy canned food, processed food, beverages and confections” corresponds to our cluster 3 that is “Moderate spenders who prefer snacks over cooking”.

6 Conclusion

In this research, we performed Multiple Correspondence Analysis on a dataset of 61.895 supermarkets customers. The estimation of principal dimensions was based on the product groups that the customers purchased and was enriched with supplementary qualitative features. We then applied hierarchical clustering on principal components. The results revealed six clusters.

Over 40% of the company’s customers (cluster 1 & 2) appear to be non-loyal, as they have very few transactions over the year. The company could apply attraction marketing strategies similar to what Tesco did, to entice lower income shoppers (Disney, 1999). On Clusters 3, 4 & 5, who are moderate spenders, retention and growth marketing strategies could be implemented. Finally, regarding high spenders, the company should focus on customers retention strategies (Myler, 2016).

While our study provided rich findings supported by a large real-world dataset, we note some limitations. The results of MCA can be sensitive to the binning process. Different coding schemes can lead to different results, so it's important to carefully consider how to discretize the variables. Converting purchased quantities of product groups into scores 1-3 was crucial for the interpretability of the results but also introduced some roughness in the differentiation among moderate, high and very high spenders. Interpreting MCA results can be challenging, especially when dealing with a large number of categorical variables or categories. Extracting meaningful insights from the plots and results can require expertise and domain knowledge. Despite these limitations, MCA can be a valuable tool for analyzing categorical data and uncovering patterns and associations, especially when used in combination with other techniques. Future research could explore the findings from product groups in conjunction with an RFM analysis or a loyalty clustering. Moreover, our methods can be compared to alternative ones, such as Latent Class Analysis and neural network based learning algorithms. Finally, a more specialized research could focus solely on FMCG product categories, thereby allowing for a finer analysis of a reduced number of categories.

References

1. Bejaei, M., Cliff, M. A., & Singh, A. (2020). Multiple correspondence and hierarchical cluster analyses for the profiling of fresh apple customers using data from two marketplaces. Foods, 9(7). https://doi.org/10.3390/foods9070873
2. Benzecri, J-P: Correspondence Analysis Handbook. New-York: Dekker, P. (1992)
3. Choi, B. C. K., & Pak, A. W. P. (2005). A catalog of biases in questionnaires. Preventing Chronic Disease, 2(1), 1–13.
4. Disney, J. (1999). Customer satisfaction and loyalty: The critical elements of service quality. Total Quality Management, 10(4–5), 491–497. https://doi.org/10.1080/0954412997442
5. Dogan, O., Hiziroglu, A., & Seymen, O. F. (2021). Segmentation of Retail Consumers with Soft Clustering Approach. Advances in Intelligent Systems and Computing, 1197 AISC, 39–46. https://doi.org/10.1007/978-3-030-51156-2_6
6. Durga, P., & Durga, B. P. (2018). Customer profiling and segmentation using transactional utilities.
7. Frasquet, M., Ieva, M., & Ziliani, C. (2021). Online channel adoption in supermarket retailing. Journal of Retailing and Consumer Services, 59. https://doi.org/10.1016/j.jretconser.2020.102374
8. Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001
9. Greenacre, M. (2017). Correspondence analysis in practice. CRC press.
10. Greenacre, M. J. (2013). Correspondence analysis. The Oxford Handbook of Quantitative Methods. Statistical Analyses, 2, 142-153.
11. Harris, P., Dall'Olmo Riley, F., Riley, D., & Hand, C. (2017). Online and store patronage: a typology of grocery shoppers. International Journal of Retail and Distribution Management, 45(4), 419–445. https://doi.org/10.1108/IJRDM-06-2016-0103
12. Hiziroglu, A., Patwa, J., & Talwar, V. (2012). Customer portfolio analysis: Crisp classification versus fuzzy classification - Based on the supermarket industry. Journal of Targeting, Measurement and Analysis for Marketing, 20(2), 67–83. https://doi.org/10.1057/jt.2012.5
13. Huang, Q., & Zhou, F. (2017). Research on retailer data clustering algorithm based on Spark. AIP Conference Proceedings, 1820. https://doi.org/10.1063/1.4977378
14. Husson, F., Lê, S., & Pagès, J. (2017). Exploratory Multivariate Analysis by Example Using R. CRC Press https://doi.org/10.1201/b21874
15. Industrial Psychology, Annual Review of Psychology, Vol. 9:243-266 (Volume publication date February 1958) https://doi.org/10.1146/annurev.ps.09.020158.001331
16. Jenneson, V., Clarke, G. P., Greenwood, D. C., Shute, B., Tempest, B., Rains, T., & Morris, M. A. (2022). Purchasing Behaviour Using Supermarket Transaction Data. 1–18.
17. Lingras, P. (2015). Selection of Time-Series for Clustering Supermarket Customers.

18. Lingras, P., Elagamy, A., Ammar, A., & Elouedi, Z. (2014). Iterative meta-clustering through granular hierarchy of supermarket customers and products. Information Sciences, 257, 14–31. https://doi.org/10.1016/j.ins.2013.09.018
19. Lingras, P., Hogo, M., Snorek, M., & West, C. (2005). Temporal analysis of clusters of supermarket customers: Conventional versus interval set approach. Information Sciences, 172(1–2), 215–240. https://doi.org/10.1016/j.ins.2004.12.007
20. Mahalakshmi, V., Jhoncy, A., & Geetha, A. (2020). A study on customer profiling. Malaya Journal of Matematik, S(2), 4584–4586. https://doi.org/10.26637/MJM0S20/1182
21. Manca, F., D'Uggento, A. M., & Convertini, N. (2018). Customer segmentation through multiple correspondence analysis. 2018 110th AEIT International Annual Conference, AEIT 2018, October 2020. https://doi.org/10.23919/AEIT.2018.8577279
22. Martineau, P. (1958), "The personality of the retail store", Harvard Business Review, Vol. 36, pp. 47-56
23. Myler, L. (2016), "Acquiring new customers is important but retaining them accelerates profitable growth", Forbes, June 8, available at: https://www.forbes.com/sites/larrymyler/2016/06/08/ acquiring-new-customers-is-important-but-retaining-them-accelerates-profitable-growth/ #7efff0546671
24. Nguyen, S. P. (2021). Deep customer segmentation with applications to a Vietnamese supermarkets ' data. Soft Computing, 25(12), 7785–7793. doi:.10.1007/s00500-021-05796-0
25. Oliveira, W. V, Araujo, D. S. A., & Bezerra, L. C. T. (2022). Supermarket customer segmentation: A case study in a large Brazilian retail chain. Proceedings - 2022 IEEE 24th Conference on Business Informatics, CBI 2022, 1, 70–79. https://doi.org/10.1109/CBI54897.2022.00015
26. Omar, M. W., Mohd Ali, M. N., Hussin, Z. H., & Rahim, H. A. (2009). Decision Orientations towards Shopping and Buying among Young-Adult Malays in the Universities. International Journal of Business and Management, 4(7). https://doi.org/10.5539/ijbm.v4n7p26
27. Owolabi, O. O., Adeleke, Y. S., & Abubakar, K. (2013). Technology Enabled Customer Relationship Management in Supermarket Industry in Nigeria. American Journal of Industrial and Business Management, 03(02), 222–228. https://doi.org/10.4236/ajibm.2013.32027
28. Philippe, J., Malet-damour, B., Harimisa, M., Fontaine, L., & Rivière, G. (2019). GIS-based approach to identify climatic zoning : A hierarchical clustering on principal component analysis. Building and Environment, 164(March), 106330. https://doi.org/10.1016/j.buildenv.2019.106330
29. Rokaha, B., Gautam, B. P., & Ghale, D. P. (2018). Enhancement of Supermarket Business and Market Plan by Using Hierarchical Clustering and Association Mining Technique; Enhancement of Supermarket Business and Market Plan by Using Hierarchical Clustering and Association Mining Technique. https://doi.org/10.1109/NaNA2018.2018.00075
30. Stalidis, G., & Diamantaras, K. (2019). Offers just for you : intelligent recommendation of personalised offers employing multidimensional statistical models. 7th International Conference on Contemporary Marketing Issues, July, 2–5.
31. Theodoridis, P. K., & Chatzipanagiotou, K. C. (2009). Store image attributes and customer satisfaction across different customer profiles within the supermarket sector in Greece. European Journal of Marketing, 43(5–6), 708–734. https://doi.org/10.1108/03090560910947016
32. Vieira, V. A., & Damacena, C. (2007). Loyalty in the supermarket. BAR - Brazilian Administration Review, 4(3), 47–62. https://doi.org/10.1590/s1807-76922007000300005