Cluster Analysis for Two HoReCa Channels of a Large Grocery Retail Company

Sandra M. Aleixo[1,2(✉)], Bruno Pereira[1], Tiago Fernandes[1], José Rodrigues[3], and Miguel Pina[3]

[1] Department of Mathematics, ISEL – Instituto Superior de Engenharia de Lisboa, IPL – Instituto Politécnico de Lisboa, Rua Conselheiro Emídio Navarro, 1, 1959-007 Lisboa, Portugal
{sandra.aleixo,bruno.pereira}@isel.pt, A48695@alunos.isel.pt

[2] Faculdade de Ciências da Universidade de Lisboa, CEAUL – Centro de Estatística e Aplicações da Universidade de lisboa, Bloco C6, Piso 4, 1749-016 Lisboa, Portugal

[3] Jerónimo Martins, Rua Actor António Silva, 7, 1649-033 Lisboa, Portugal
{jose.manuel.rodrigues,miguel.pina}@jeronimo-martins.com

Abstract. This work was developed within the scope of industrial mathematics, in partnership with a large grocery retail company in the Portuguese market, Jerónimo Martins, in particular with Recheio, the largest chain of Cash&Carry stores in Portugal. The objective was to segment the Families (groups of products in the marketing structure) in Recheio's HoReCa channel, into disjoint sets that can be interpreted by the business. Initially, database manipulation, preprocessing and choice of variables to use were carried out. Then followed the segmentation phase carried out in two ways: through the total HoReCa channel and through the online HoReCa channel. A clear separation of the company's products was achieved using various Cluster Analysis methods such as K-Means, K-Medoids and Hierarchical. Different groups were obtained, based on common characteristics of the Families, with the aim of better understanding the company's market, helping to make strategic decisions and develop more effective marketing and production plans. The work involved the use of several types of software, namely Excel, SQL, Python and R.

Keywords: Cluster analysis · Database manipulation · Preprocessing · Retail · HoReCa channel · Industrial mathematics

1 Introduction

In order to strengthen their market position during recent years, the main challenge for retail companies is to maintain a sustained growth in customers and sales of both the Traditional Retail and HoReCa (Hotels, Restaurants and Catering) channels, in very competitive market conditions, due to the uncertainty

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14815, pp. 337–356, 2024.
https://doi.org/10.1007/978-3-031-65154-0_21

caused by the economic environment. This can be achieved using different marketing strategies, namely those that improve customer experience, achieve customer retention and grow their customer base. Thus, there is a growing significance placed on customer-oriented marketing. The rise of individual-focused approaches, particularly in online commerce, is rendering conventional mass marketing in this sector progressively outdated, given the feasibility of tailoring efforts to specific customer needs. This approach underscores the necessity of cultivating a profound comprehension of the customer's buying behavior instead of its demographics or location, for example.

One of the biggest companies in Portugal is Jerónimo Martins, which operates in the food distribution and specialized retail business. The group is present in Portugal, Poland and Colombia and, currently, around 70% of revenue comes from Biedronka, in Poland. The present work is based on data provided by the Market Analysis and Studies section of the Marketing Department of Recheio, one of the businesses of Jerónimo Martins, which has 4% of the revenue of the group. The company set two goals when they proposed this study: the purchasing behavior of its clients in the two HoReCa channels and the comprehension of the distinction in performance of the online HoReCa channel in comparison to the total HoReCa channel.

Instead of dividing a company's customer base into distinct and homogeneous groups or segments based on shared characteristics to develop a suitable marketing strategy [2], the pandemic has caused retailers to focus on product grouping based on customer buying pattern, in order to adapt to changes in customers lives and work, since most corporations encouraged remote work and online collaboration [3]. There are three main requirements that retailers must comply with if they want to analyze customer behavior: the ability to collect and store customer data, a suitable algorithm to characterize the collected data, and an efficient strategy to apply the extracted characterization successfully into the business model. We will focus on the first two items using suitable preprocessing methods and cluster analysis, based on the data provided by Recheio.

Cluster analysis is a data analysis technique applied in several fields of study, in which its main goal is to organize collected datasets or observations into groups or clusters based on the similarity between them, i.e. it identifies underlying groupings and patterns into subsets in a way that members within the same subset are more similar to each other than to those in other subsets [7].

The process of cluster analysis involves assigning data points to non overlapping clusters so that the within-cluster similarity is maximized, while the between-cluster similarity is minimized. This is achieved analysing how two data points or objects are alike with respect to certain attributes or characteristics such as the euclidean distance or the correlation coefficient, among others [10]. Various algorithms are used to perform cluster analysis. In this article, we focus on K-means, K-medoids, and Hierarchical clustering.

K-means clustering works by iteratively assigning data points to groups based on their distance to the cluster centroid and then minimizing the within-cluster variance by optimizing the centroids. In the end, the sum of squared distances between data points and their assigned centroids is minimum [8,11]. Since this

approach is sensible to outliers, an alternative approach designed to identify representative data points, known as medoids, can be used within each cluster [9]. Defining a predetermined number of medoids, the algorithm aims to minimize the sum of distances between those points and the remaining data, selecting, in each iteration, the optimal medoids. This algorithm offers robustness to outliers and noise in the data. Hierarchical clustering is a technique used to organize large datasets to form hierarchical groups of mutually exclusive subsets, each of which has members that have maximally affinity with respect to a specific attribute or property, in order to better consider and understand relations within those datasets [14], producing a nested set of clusters, where the relationship between them is represented in a dendrogram. This clustering method may assume that all data are in the same cluster and, in each step, separates the clusters until each data point forms its own cluster (divisive), or starts with individual data points as separate clusters and iteratively merges the most similar clusters until all data points belong to a single cluster (agglomerative).

This paper is organized as follows. In Sect. 2, we detail which variables are present in the data collected by Recheio, the preprocessing method and data refining using Python, detailing the variables and the reasoning behind the decision of considering some of the product Families as outliers, as well as establishing the optimal number of clusters. In Sect. 3, we implement the aforementioned clustering methods to valid data and obtained a clear segmentation of Recheio Families, based on descriptive metrics of their behavior. Finally, in Sect. 4, we discuss the results and draw some conclusions which helped the decision making process of more effective marketing and production plans.

2 Data Wrangling

The original data provided by Recheio for both HoReCa channels consists of transactions, made during a period of twelve months in 2022 and 2023, of each product sold by the company, which we call the Families, that can be linked to individual customers. These customers are defined in two different ways: a paying customer, who is uniquely identified by a reference (the NIF) across all Recheio stores, or as a profiled customer where an internal code is used per store to identify the customer. Hence, one paying customer can have multiple customer profiles in the same store since Recheio treats them as a different customer due to different delivery addresses, for example.

Some variables of interest are summarized in Table 1, some of which are related with the paying customer and others with the profiled customer. All variables are numerical except "family_id" and "family", which are categorical.

Due to their relevance in this work, we single out two variables. The first one is freq_payer, a ratio between the total amount of purchases from customers (payer or profiled) divided by the total number of days in the analyzed period. It represents the average amount of days between purchases in a given Family. We also define pene_cust, which is the customer penetration, a measure which gives the percentage of Recheio's customers (payer or profiled) who purchased

Table 1. Variables of the Case Study.

Name	Description
family_id	Unique value to represent each Family
family	Description of each Family
tot_sales	Total sales by Family, in euros
nm_tickets	Number of tickets, meaning purchases per Family
cust_payer	Number of paying customers per Family
spent_payer	Average spending per payer, in euros, within the specific Families
cust_profile	Number of registered customers per Family
freq_profile	Time difference, in days, between purchases per Family
spent_profile	Average spending per customer, in euros, within the specific Families

a given Family during the analyzed period. It ranges between 0 and 1, and if it is related with a payer customer, the variable is called pene_cust_payer. On the other hand, if it is associated to a profiled customer, the variable is designated by pene_cust_profile. For example, if pene_cust_payer for some Family is zero it means that no Recheio's customers payer purchased any product of this Family.

The values of the variables were obtained from the database through queries in SQL language. Then, it was necessary to preprocess them, analysing missing data, duplicate values and the existence of outliers. We found out that there were no null or repeated values, enabling the start of outlier search after the variables selection. The organized data was used in the applied methodologies. For each variable (column in the matrices), we had 155 and 124 Families (number of rows in the matrices, one for each Family), for the total and the online channel, respectively.

3 Statistical Analysis

We begin this statistical study by selecting variables that are not highly correlated, since this can help to reduce redundancy and multicollinearity in the data, improving the performance of the clustering algorithm, [6]. Subsequently, we performed a brief descriptive statistical analysis to characterize those variables and identify extreme outliers, which are under consideration for removal. Finally, we apply the K-Means, K-Medoids, and Hierarchical unsupervised machine learning clustering algorithms to obtain a segmentation of the Families of both HoReCa channels.

The focus of Recheio was on the payer customers, therefore this work will only involve the variables "cust_payer", "freq_payer" and "spent_payer", beside the general ones "tot_sales" and "nm_tickets". Since the variable "pene_cust_payer" is proportional to cust_payer, we also replace the latter by the former.

3.1 Total HoReCa Channel

Variable Selection. In cluster analysis, it is important not to use highly correlated redundant variables. The selection of variables to include in cluster analysis regulate the characteristics of the clusters themselves. If the variables exhibit significant multicollinearity, they interfere with the calculation of dissimilarity measures, namely Euclidean distance, hence the reliability of the results is not guaranteed [13].

Not all variables observed in a study are always interesting for cluster analysis. Some do not contribute to the objective of the study and others are redundant as the characteristic they represent is strongly correlated with other variables. Taking this argument into account, these redundant variables can be eliminated [1]. The inclusion of redundant variables complicates or even masks the clustering process [5].

In cluster analysis, when variables exhibit collinearity, certain variables wield greater weight than others. In instances where two variables are perfectly correlated, they are essentially similar. Consequently, they are duplicated within the dataset, doubling the weight they carry compared to other variables in cluster analysis. As a result, the resultant solution may skew disproportionately towards these variables, potentially posing new problems if this isn't taken into consideration.

Using Shapiro-Wilk test, we conclude that the normality assumption for the five variables is not verified. In fact, the histograms of the main diagonal of Fig. 1 already suggested that this was the outcome. Therefore, we opted for Spearman's coefficient to evaluate the correlation of each pair of variables among the five existing ones [13], and the associated test to the correlation significance. The p-values being close to zero leads to the rejection of the null hypothesis so we can conclude that we have significant correlations.

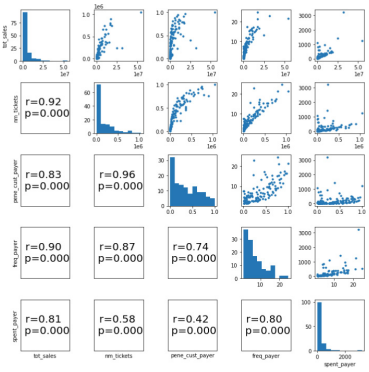


Fig. 1. Correlations Matrix.

Above the diagonal of the correlation matrix in Fig. 1, there are scatter plots which relate every possible pair of variables. In the lower triangular part, we have the value of sample correlation coefficient between each pair of variables with the p-value associated with the hypothesis test for the nullity of the population correlation coefficient, to assess the significance of their correlation.

As previously mentioned, there are two very important variables in this study from the Recheio company's perspective, namely freq_payer and pene_cust_payer. These variables were chosen to build the clusters due to their historical relevance and wide acceptance as key business metrics in internal decision business making by the company's functional teams, for example, when making decisions about the positioning in store of a specific line of products. It is also important for the company to be able to easily and consistently measure the impact of these business decisions, so having available a small set of consensual variables is crucial for that objective.

It can be assumed that two variables are highly correlated if the correlation coefficient between them exceeds 0.8 [4]. In Fig. 1, the correlation coefficient value for the aforementioned variables is 0.74, hence we conclude that they are not strongly correlated for cluster analysis. Note that the variable spent_payer is also not strongly correlated with pene_cust_payer ($r = 0.42$), but it is strongly correlated with the variable freq_payer, thus it should not be included in this analysis. Similarly, the variables nm_tickets and tot_sales are strongly correlated with the variables freq_payer and pene_cust_payer ($r = 0.87$ and $r = 0.96$, respectively, for nm_tickets; $r = 0.90$ and $r = 0.83$, respectively, for tot_sales), therefore they should also be excluded from the cluster analysis.

Descriptive Analysis. A descriptive analysis of the two variables is conducted to characterize them and to check for the existence of outliers. In Fig. 2 and Table 2 we have the boxplots and the descriptive metrics of the two variables.

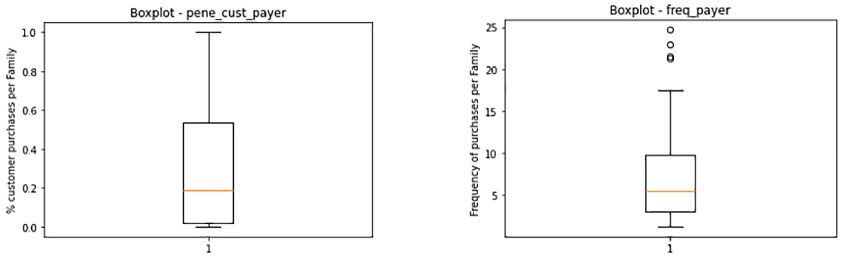


Fig. 2. Boxplots for pene_cust_payer and freq_payer for the total HoReCa channel.

Regarding the variable pene_cust_payer, the average is 0.208, which means that the average percentage of Recheio's customers payer who purchased a given Family during the analyzed period is around 21% in the total channel. Also in location terms, data round about 0.185, being 50% of the observations less

Table 2. Descriptive metrics of total HoReCa channel variables.

pene_cust_payer	value	freq_payer	value
Average	0.208		
Median	0.185	Median	5.446
Standard Deviation	0.291		
Minimum	0.000	Minimum	1.199
Maximum	1.000	Maximum	24.683
1st Quartile	0.020	1st Quartile	2.990
3rd Quartile	0.536	3rd Quartile	9.694
Inter-Quartile Range	0.516	Inter-Quartile Range	6.704
Coefficient of variation	101.301%	Coefficient of variation (resistant)	78.948%
Skewness coefficient	0.720	Bowley coefficient (resistant)	0.267
Kurtosis coefficient	−0.786	Kurtosis coefficient (resistant)	1.500

than this value and the other half greater than it. Since the median value is closer to the first quartile and the right tail is longer, the data exhibit positive skewness, confirmed by the value of the skewness coefficient, indicating that pene_cust_payer data are concentrated at lower values. For dispersion, data show a strong variability as 50% of the central values vary 0.516 at most (high IQR value length of the box), confirmed with the high value of the variation coefficient around 101%. Furthermore, in boxplot, there are no outliers. So, to characterize the sample, other nonresistant measures that use all the information contained in the data can be calculated. For the observed data, the value of the pene_cust_payer variable is on average 0.208, with a variability of 0.291 around this value. About sample location for the Freq_payer variable, the values are around 5 and a half days, with 50% of the observations being lower than this value and the rest higher. Regarding dispersion, the values present a great deal of variability, increased by the existence of outliers, and confirmed by the high value of the coefficient of variation of around 79%. The asymmetry is clearly positive since median is closer to left side of the box, right tail is much longer and there are outliers on the right, confirmed by the positive value of the Bowley coefficient. We use resistant measures to characterize the sample, since outliers exist.

Thus, as outliers are present for the variable freq_payer, a Python program was created to find which Families are outliers, to wit: fresh vegetables, fresh fish from aquaculture and basic products. Since these outliers are all mild (as they fall within the bounds of the inner fences of the boxplot) and these Families are actually very relevant to Recheio, it was decided to keep these observations in the sample. The relevance is due to the corresponding high number of sales and to follow the business strategy.

On the other hand, due to low sales (less than €10000), the company decided to remove all Families that don't reach this sales threshold, as these are indeed

insignificant for the company's business. The number of Families removed is around 30 and they are mostly related to leisure, hygiene and personal care products, food and appliances.

Cluster Analysis. We made the segmentation of Families resorting to the aforementioned cluster analysis algorithms. One of the first things we need to determine is the most appropriate number of clusters to use. There aren't clear rules, just a comprehensive set of guidelines we can use. For example, if the main goal is to identify market segments, a certain number of clusters proposed by management is the way to go. The number of clusters can also be decided using a dendogram, for hierarchical clustering, or plotting the ratio of total within-group variance against variance between each group, for non-hierarchical clustering [12]. In the present case, we will use the Elbow Method to guide us. According to Fig. 3, we can observe that the "elbow" is verified for three clusters, and the inertia is decreasing more slowly above this number, until the reduction in inertia becomes much less considerable from six clusters onwards. We are going to consider six clusters even because this number of clusters enable internal comparison in Recheio with similar cluster analysis in the companies of the group, which are focused in segmenting groups of products which also use penetration and frequency as variables. In fact, this result was in line with the view of the retail company in business context, since it is suitable for the subsequent comparison between HoReCa channels.

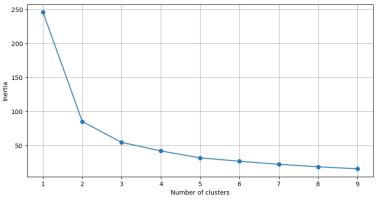


Fig. 3. Elbow method - total channel

In Fig. 4, we plot the graphical representations of the six clusters obtained by the several methods used. The main goal is to analyze which method provides a better separation of groups, which correspond to the best segmentation. Observing the scatter plots, there are two methods that seem to produce good clusters, cohesive and well separated: the Hierarchical with Ward Linkage criteria and the K-Medoids. But let's validate the results by analysing the silhouette graphs and the average silhouette score values presented in Fig. 5. Each observation or element i from each cluster has its own silhouette score $s(i)$. A large $s(i)$ (almost 1) means that this element is very well clustered, a small $s(i)$ (around 0) means that the observation i lies between two clusters and a negative $s(i)$ means that the element i is probably placed in the wrong cluster. After computing the silhouette coefficient for each point from each cluster, the average of all elements scores is the average silhouette score (red line in the Silhouette Plots on Fig. 5).

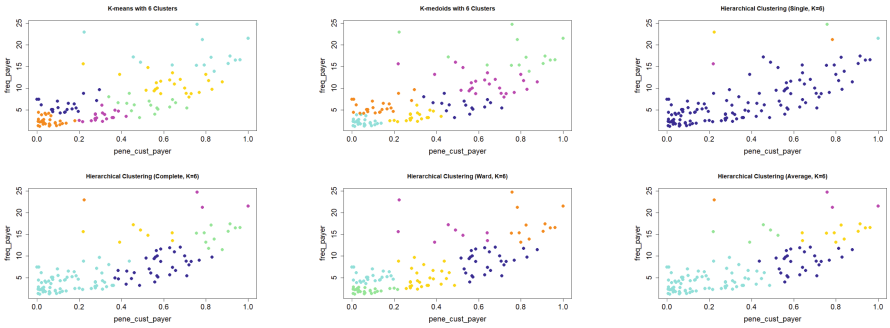


Fig. 4. K-Means, K-Medoids, Hierachical Clustering with Single, Complete, Ward and Average linkage, in the total HoReCA channel - 6 clusters.

Analysing the silhouette plots for the six methods in Fig. 5, we can take some conclusions. The Hierarchical Method with Single Linkage is the worst choice because there is a cluster with almost all elements which has many points with negative scores (misclassified) and each of the other five clusters has only one element with null score, besides that the average silhouette score is low (0.22). The K-Means Method has a low average silhouette score (0.13) and there are many misclassified elements in all clusters, so it is not a good choice. The Hierarchical Methods with Complete and Average Linkage, although having good average silhouette scores (0.48 and 0.5), higher than the Hierarchical Method with Ward Linkage, have a cluster much bigger than the others and have also some misclassified elements. From the analysis of both: the scatter plots and Silhouette plots, the two methods that produce the most cohesive and well separated clusters are the Hierarchical with Ward Linkage Criteria and the K-Medoids. For these clustering methods, all clusters have several points above the average score (red line), even if there are a few points with negative scores (misclassified) and both methods allow to obtain more homogeneous width for almost all the clusters, which means that the clusters have similar sizes. However, we consider that the clusters obtained by the Hierarchical Method with Ward Linkage are our choice in this work for the total channel, because the average silhouette score for this method (0.41) is slightly higher than for the K-Medoids (0.38).

Analysing the distribution of Families across clusters obtained by the best chosen method, it can be seen that: Cluster 1 includes the following Families: frozen fish and seafood, ready-made food products, breakfast goods, snacks and cafeteria products, accessories and cleaning products, packaging/preserving products and raw materials, and body hygiene products; Cluster 2 has Families of fish and dried cod, more specific meats (lamb and goat, small animals and special beef), meals, cold cuts, pastries, daily milk and vegetable drinks, cat/dog products, restaurant drinks and foreign wines; the Families in Cluster 3 cover body hygiene products, kitchen and household items, packaged cakes, stationery products, air fresheners, disinfection, cleaning and extermination products, and health and healthy food items; Cluster 4 aggregates grocery products, beauty

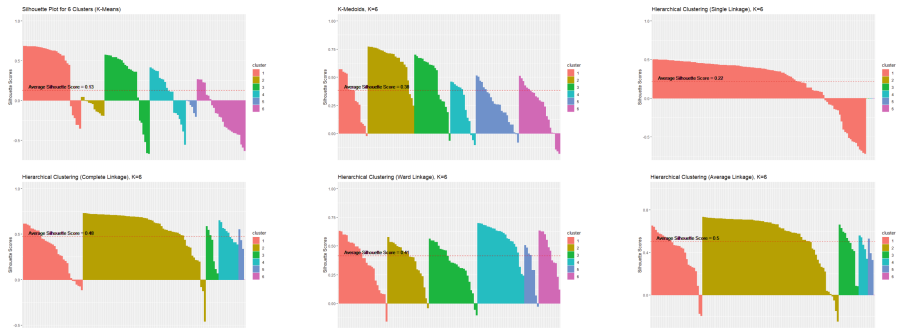


Fig. 5. Silhouette Plots for total HoReCa channel with 6 clusters, for the six cluster analysis methods.

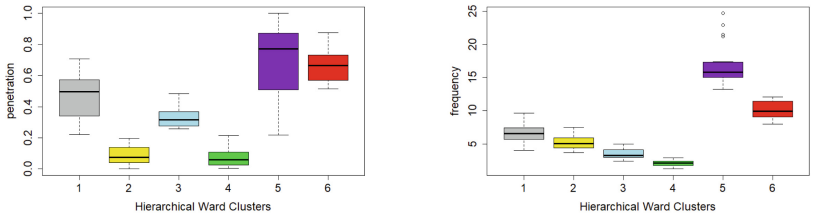
and personal hygiene products, uniforms and clothing, pet products, refreshing drinks, natural flowers and field/beach and garden items, household appliances, lighting, image and car items, and facility cleaning products; Cluster 5 has fresh products, namely poultry, beef and pork meat, fresh wild fish, fresh aquaculture fish, eggs, cold cuts, fruit and vegetables, drinks, preserved products, conservation vegetables, sweet desserts; Cluster 6 contains current and still wines, beers, spirits, coffee shop drinks, water, festive and disposable items, paper products, dish and household cleaning products, icecreams, bakery, dairy products and food fats.

For each of the six clusters, we computed statistical metrics for each variable to characterize each of these clusters and to understand their properties. The statistical metrics in Table 3 provide an overview of trends and distributions of values in each cluster. Also, in Fig. 6a, we plot two parallel boxplots about the 6 clusters concerning the variable pene_cust_payer. The Clusters 5 and 6 stand out for having the highest medians, 0.772 and 0.665, respectively. Therefore, Families belonging to these two clusters have higher values for the weight of each Family relative to the total weight of the total channel Families. On the other hand, Clusters 2 and 4 have very low medians, 0.074 and 0.059, respectively. In terms of dispersion, it is noteworthy that the Cluster 5 has the highest dispersion followed by Cluster 1. In Clusters 2, 3 and 4, the dispersion is similar. These three clusters have the median closest to the first quartile, indicating positive skewness i.e., concentration of the pene_cust_payer variable in low values. Conversely, Clusters 1, 5 and 6 have negative skewness i.e., concentration in high values. Cluster 5 has the greatest inter-quartile range equal to 0.339, meaning that 50% of the sample central values vary, at most, approximately 0.34. None of the clusters have outliers.

The parallel boxplots and statistical metrics for the variable freq_payer are presented in Fig. 6b and Table 3, respectively.

Table 3. Statistical Metrics of pene_cust_payer and freq_payer in HoReCa total

Cluster	Mean_p	Med_p	Min_p	Max_p	St Dev_p	IQR_p	Med_f	Min_f	Max_f	IQR_f
1	0.478	0.497	0.221	0.709	0.146	0.221	6.574	4.011	9.666	1.539
2	0.090	0.074	0.000	0.196	0.061	0.099	5.069	3.665	7.460	1.513
3	0.334	0.316	0.257	0.485	0.066	0.090	3.262	2.352	4.980	1.103
4	0.070	0.059	0.005	0.216	0.061	0.078	2.108	1.199	2.881	0.665
5	0.693	0.772	0.219	1.000	0.237	0.339	15.812	13.188	24.683	2.140
6	0.669	0.665	0.516	0.877	0.106	0.155	9.886	7.981	12.040	2.267



(a) total_channel - pene_cust_payer **(b)** total_channel - freq_payer

Fig. 6. Boxplots of the total channel.

Analysing Fig. 6b and Table 3, the Cluster 5 is emphasized in localization terms because it has the highest median and it is the only one that has outliers. The median value is about 16 days, which means that half of the Families in this cluster have a time between two purchases (for the payer customer) that is a maximum of 16 days and the rest have times longer than that value. Clusters 3, 5 and 6 exhibit positive skewness, meaning the data concentrates on lower values. The Clusters 5 and 6 are the ones that show the greatest inter-quartiles ranges, therefore, they are the clusters in which the variable freq_payer presents the highest variability. These inter-quartiles ranges with values of 2.140 and 2.267, respectively, mean that an half of the products related to Families in these clusters, corresponding to central values of the freq_payer variable, are purchased by payer costumers at most with a difference of around 2 days.

3.2 Online HoReCa Channel

Variable Selection. Regarding the variable selection process for the online HoReCa channel, we obtained the same results and selected the same two variables for the study: pene_cust_payer and freq_payer.

Descriptive Analysis. A descriptive analysis of the two variables is performed to characterize them and to check for the existence of outliers. Here we make only

a brief analysis because it is analogous to the other made for the total channel. In mean terms, the values of pene_cust_payer for the online channel round about 0.12, and have a high variability (coefficient of variation of 139%) and there are some outliers to the right). For the variable freq_payer, the times between product sales are around 4 days but there is also high variability (coefficient of variation of 81%). Both variables have a positive skewness, meaning a higher concentration on small values.

In Fig. 7, we show the boxplots for the two variables, and in Table 4 we have the comparable descriptive metrics.

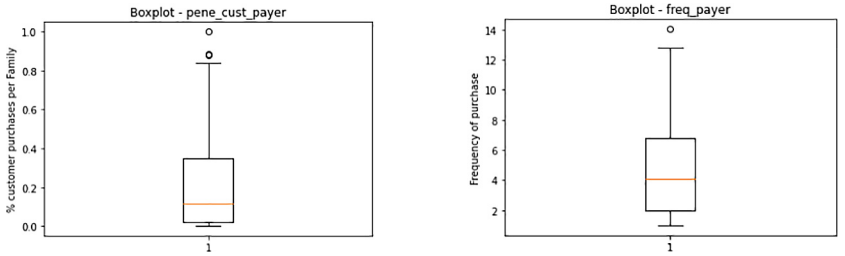


Fig. 7. Boxplots for pene_cust_payer and freq_payer for the online HoReCa channel.

Table 4. Descriptive metrics of online HoReCa channel variables.

pene_cust_payer	value	freq_payer	value
Median	0.117	Median	4.076
Minimum	0.002	Minimum	1.000
Maximum	1.000	Maximum	14.034
1st Quartile	0.020	1st Quartile	2.000
3rd Quartile	0.351	3rd Quartile	6.799
Inter-Quartile Range	0.331	Inter-Quartile Range	4.799
Coefficient of variation (resistant)	139.390%	Coefficient of variation (resistant)	81.114%
Bowley coefficient (resistant)	0.412	Bowley coefficient (resistant)	0.135
Kurtosis coefficient (resistant)	1.500	Kurtosis coefficient (resistant)	1.500

Cluster Analysis. Now with data of online HoReCa channel, we want also to group Families in separate clusters, using the same analysis algorithms used before for the data of total HoReCa channel: K-Means, K-Medoids, and Hierarchical.

Once again, the number of clusters to use will be defined based on the Elbow Method in Fig. 8 attending also to the business needs and objectives.

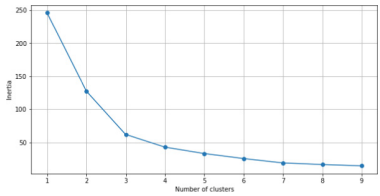


Fig. 8. Elbow method - online channel

Based on Fig. 8, the conclusion about the number of clusters is similar to the one made for the total channel. So by the same reasons presented before we choose to use six clusters, even to later be able to compare the results obtained for the two channels as intended.

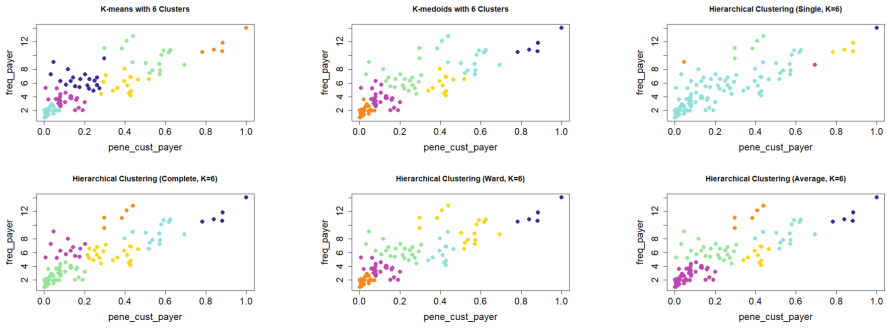


Fig. 9. K-Means, K-Medoids, Hierarchical Clustering with Single, Complete, Ward and Average Linkage, in the online HoReCA channel - 6 clusters. (Color figure online)

For the online channel, observing the scatter plots of Fig. 9, the choice of the most efficient method to obtain the best clusters is not so clear. Almost all methods seems to produce reasonable results, except the Hierarchical Method with Single Linkage. In fact, this method does not clearly produce acceptable results again, as it produces a huge cluster and the others with very few elements. However, looking more closely, the K-Means Method has some elements that do not clearly belong to a given cluster and there is a certain mix. The K-Medoids Method has two clusters (green and blue) whose elements are more dispersed. The Hierarchical Method with Ward Linkage has a group (the yellow one) that is separated in two parts with another cluster in the middle (the cyan one). The Hierarchical Method with Average Linkage is better but some elements of the yellow cluster are surrounded by cyan points of another cluster. In reality, the Hierarchical Method with Complete Linkage seems to be the one that has more cohesive clusters but well separated from each other, so it appears to be the best method for the online HoReCa data.

We will use Silhouette Plots on Fig. 10 to validate these conclusions. In this case the plot related to the Hierarchical Method with Single Linkage confirms the bad results and it is similar to the correspondent plot in the total HoReCa case. For the other five methods all the Silhouette plots are quite similar, they have a bigger cluster and some of them have clusters with many points below the red line (average silhouette score), but the values of the average silhouette score are all very alike, however it is really the Hierarchical Method with Complete Linkage which has the highest value (0.5). So we consider the clusters produced by this method for the online channel.

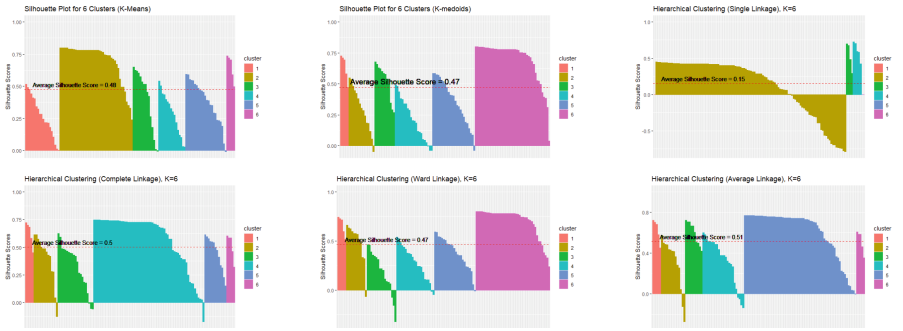
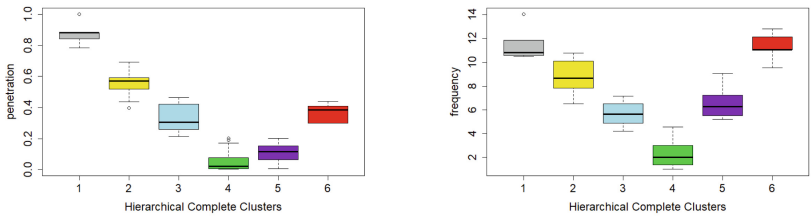


Fig. 10. Silhouette Plots for online HoReCa channel with 6 clusters, for the six cluster analysis methods. (Color figure online)

The composition of the clusters obtained from Hierarchical method with Ward Linkage is as follows: Cluster 1 has essentially Families related to fresh fruits and vegetables, preserved vegetables and some type of desserts; Cluster 2 is formed by Families of healthy food and vegetable drinks, poultry, beef and pork, appetizers, food fats, fresh aquaculture fish, packaged cakes, seafood, meals, bakery, ice-creams, still wines, beers, spirits and water. The cluster also includes cleaning accessories and products for dishes and clothes, as well as party items and packaging and conservation products; some of the Families in Cluster 3 are refreshing and warm drinks, current wines, dairy products, eggs, breakfast goods; Cluster 4 comprises a wide variety of Families, namely food Families as whole dried cod, more specific meats (small animals, special breed beef, lamb and goat), fresh wild fish and frozen fish, seafood, frozen meals, grocery and delicatessen products, pastries and milk of the day; cleaning Families such as air fresheners, disinfectants, extermination products and various types of cleaning. It also holds other types of items, specifically for the home and kitchen, pets, country/beach and garden items, hygiene, beauty and health products, uniforms and clothing, household appliances, lighting, image and automobile items; Cluster 5 owns Families of products for cats and dogs and coffee shop drinks; Cluster 6 carries Families of preserved products, basic products, condiments, milk and sweet desserts.

As was done for the total HoReCa channel, an analysis was carried out regarding the two variables under study for each of the six clusters defined for the HoReCa online channel. Firstly, for the variable pene_cust_payer, Fig. 11a and Table 5 can be observed. In terms of location, Families in Cluster 1 are those that clearly reach the highest values for the pene_cust_payer variable (median of 0.882), followed by cluster 2 (median of 0.571). For Families in Clusters 3 and 6, the pene_cust_payer values are relatively close despite being higher for Families in Cluster 6 (median of 0.384). The lowest pene_cust_payer values are very identical and smaller for Families in Clusters 4 and 5. With regard to dispersion, it can be said that the Families belonging to Cluster 3 are the ones that show greater variability in terms of the values they assume for the variable pene_cust_payer. In fact, the box length of that cluster, which corresponds to the largest IQR (around 0.16), is the biggest. Furthermore, all other clusters show low variability, which means that for the Families that belong to each of these clusters, the pene_cust_payer values are relatively close. The Families included in each cluster can be consulted as described above.



(a) online_channel - pene_cust_payer **(b)** online_channel - freq_payer

Fig. 11. Boxplots of the online channel.

In relation to the other variable freq_payer, it can be analyzed Fig. 11b and Table 5, to compare its values for the six clusters obtained for the online HoReCa channel. The Clusters 1 and 6 stand out in terms of the highest median, with about 11 days. This means that, in mean terms, products from Families in these clusters are purchased at most every 11 days by a customer payer, with half of these products being purchased within this time difference at most, while the remainder are purchased with a temporal difference greater than that number of days. On the other hand, the Cluster 4 has a very low median of 2 days, indicating that, in mean terms, products from Families in this cluster are purchased at most every 2 days by a customer payer, which means that they are purchased more frequently than all other products from Families in other clusters. When we compare the dispersion of the values of the variable freq_payer in the several clusters, there is greater variability in the temporal difference, in days, between purchases per Family in Clusters 2, 3, 4 and 5. Half of the products of Families in these clusters vary at most in two days in relation to the period of time between

two purchases of that product. In Clusters 1 and 6, the variability is similar and lower, at most around a day, for half of the products in these clusters.

Table 5. Statistical Metrics of pene_cust_payer and freq_payer in HoReCa online

Cluster	Med_p	Min_p	Max_p	IQR_p	Med_f	Min_f	Max_f	IQR_f
1	0.882	0.783	1.000	0.042	10.837	10.512	14.034	1.252
2	0.571	0.398	0.694	0.068	8.687	6.537	10.780	2.011
3	0.306	0.215	0.465	0.159	5.646	4.198	7.160	1.635
4	0.022	0.002	0.201	0.071	2.000	1.000	4.565	1.625
5	0.117	0.007	0.203	0.091	6.277	5.216	9.071	1.723
6	0.384	0.298	0.440	0.110	11.057	9.545	12.819	1.084

3.3 Comparison Between HoReCa Total and HoReCa Online

As previously stated, we choose six as the number of clusters for both HoReCa channels to ensure that a meaningful comparison that is relevant for Recheio. The clusters are numbered from 1 to 6 in decreasing order of importance related to the penetration and frequency variables. Hence, Cluster 1 will have higher values for these variables.

After obtaining the segmentations for both HoReCa total and HoReCa online data, a renaming of the clusters was performed to make it more intelligible in business context (AF1 to AF6 to substitute Clusters 1 to 6). However, it should be noted that the composition of the clusters remains the same. In Table 6, we link the AF_i groups and C_j clusters to simultaneously identify the Families belonging to each cluster and their respective variable values. AF1 represents the most important Families for the business, while AF6 represents the least important ones. Table 6a explains the renaming criteria for the Families in the HoReCa total channel, and Table 6b for the HoReCa online channel. To make things easier, we will designate the variables pene_cust_payer and freq_payers respectively as Penetration and Frequency.

Figure 12 aims to assist in the comparison between the results obtained for the two channels. On the left, we have the scatter plot and pie chart of the clusters formed using the Hierarchical Method with Ward Linkage for the HoReCa total channel. On the right, the same for the HoReCa online channel, obtained through the Hierarchical Method with Complete Linkage. We can conclude that Penetration vs. Frequency plot for the online channel indicates less frequency in this channel, and indicates fewer Families with Penetration above 50%. It is also noteworthy that in the online channel, there is a greater concentration of Families in AF6 (+32%) and a lower concentration of Families of the others AFs, specially in AF2 (−11%), AF5 (−8%), AF1 (−6%), and AF4 (−5%), when compared to the total channel. In fact, the groups AF2, AF5, AF1 and AF4 lost a significant amount of weight from total HoReCa to online HoReCa data, AF3

Table 6. Family analysis in both channels

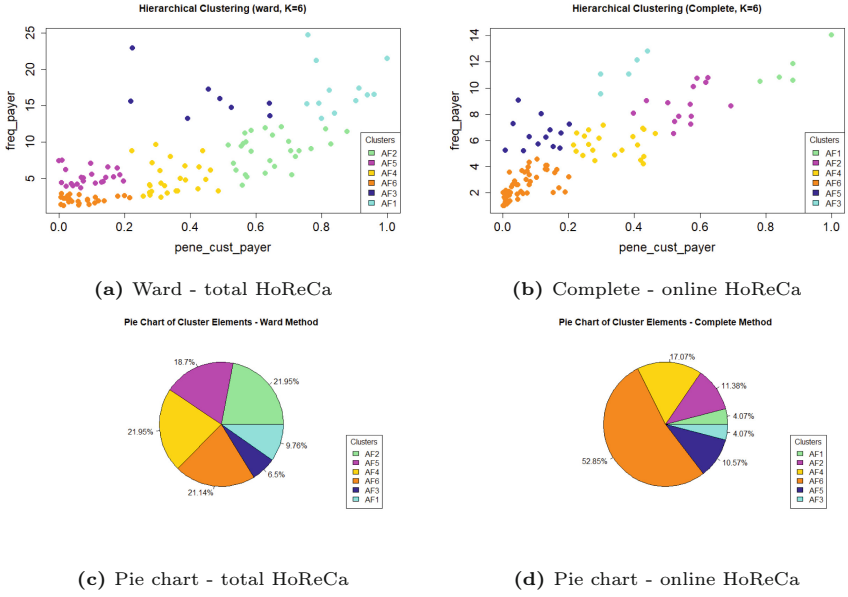
(a) Total channel

Group	Cluster	Penetration	Frequency
AF1	6	[0.750, 1.000]	[12.500, 25.000]
AF2	1	[0.515, 0.880]	[3.000, 12.500]
AF3	5	[0.220, 0.700]	[12.500, 25.000]
AF4	3	[0.200, 0.500]	[0.000, 10.500]
AF5	2	[0.000, 0.220]	[3.000, 8.000]
AF6	4	[0.000, 0.250]	[0.000, 2.900]

(b) Online channel

Group	Cluster	Penetration	Frequency
AF1	1	[0.780, 1.000]	[10.510, 14.035]
AF2	2	[0.395, 0.695]	[6.535, 10.785]
AF3	6	[0.295, 0.445]	[9.540, 12.820]
AF4	3	[0.215, 0.465]	[4.200, 7.200]
AF5	5	[0.005, 0.205]	[5.215, 9.075]
AF6	4	[0.000, 0.205]	[1.000, 4.600]

relatively maintained the same number of elements on the two channel. Finally, it should be noted that in the online channel, AF6 represents about half of the elements.



(a) Ward - total HoReCa

(b) Complete - online HoReCa

(c) Pie chart - total HoReCa

(d) Pie chart - online HoReCa

Fig. 12. Comparison between channels.

In Table 7, the defined levels for each variable are necessary to compare the behavior of Families across both channels. The quantitative levels from 1 to 5 correspond, respectively, to the qualitative levels: low, medium-low, medium, medium-high, high. These classifications contribute to a meaningful comprehension of the Penetration and Frequency of consumer activities inside each channel, enabling insightful comparisons and analyses across several customer categories.

Subsequently, we wanted to understand every substantial changes in behavior occurring between the total channel and the online channel. These changes were deemed significant if Families shifted by two or more levels in any of the variables. For example, if some Family was at level 5 in the total channel and moved to level 3 in the online channel, this would be considered a significant change. Let's present some examples of significant changes for two Families.

Table 7. Levels for both variables on each HoReCa channel.

(a) Total channel

Level	Penetration	Frequency
1	$[0, 0.2[$	$[0, 5[$
2	$[0.2, 0.4[$	$[5, 10[$
3	$[0.4, 0.6[$	$[10, 15[$
4	$[0.6, 0.8[$	$[15, 20[$
5	$[0.8, 1[$	$[20, 25]$

(b) Online channel

Level	Penetration	Frequency
1	$[0, 0.2[$	$[0, 3[$
2	$[0.2, 0.4[$	$[3, 6[$
3	$[0.4, 0.6[$	$[6, 9[$
4	$[0.6, 0.8[$	$[9, 12[$
5	$[0.8, 1[$	$[12, 15]$

Table 8. Examples of Families with significant changes between channels

Families	total channel			online channel		
	Pene	Freq	AF	Pene	Freq	AF
Fresh wild fish	0.22 (level 2)	15.62 (level 4)	AF3	0.05 (level 1)	2.63 (level 1)	AF6
Water	0.91 (level 5)	15.67 (level 4)	AF1	0.51 (level 3)	10.74 (level 4)	AF2

In Table 8, examining the transition from the total to the online channel, Family Fresh wild fish had a difference of one level in Penetration and three levels in Frequency, shifting from group AF3 to AF6, the Water Family decreases two levels in Penetration but remains in the same level for Frequency, moving from AF1 to AF2.

4 Conclusions

This study derived a comprehensive segmentation of Recheio's data, using the K-Means, K-Medoids, and Hierarchical algorithms, allowing the firm to find meaningful heterogeneous groups of products from the company's HoReCa sales channel, separating between physical and online transactions, and establishing a clear profile of its customers.

It was also possible to reveal variations in purchasing profiles when isolating transactions made through both channels. This highlights the importance of considering these two channels separately, as customers exhibit different behaviors in each channel, and of creating effective strategies for each channel separately.

The amount and nature of the variables used in the study were purposely set to be respectively small and business descriptive in order to allow quickly actionable conclusions. The groups formed with the clustering techniques are being used to: 1) discover the main differences and similarities between the pattern of purchases found in the HoReCa online and physical channel of sales and better decide what to offer in each context; 2) organize the HoReCa physical and online levels of assortment, taking advantage of the clear separation between the groups; 3) identify opportunities for recommending products to customers which have already demonstrated interest for other products within the group.
This can be a strategic advantage, allowing Recheio to make quicker and better decisions when selecting the assortment of products to be made available through each channel. Recheio can also personalize its marketing strategies, optimize product offerings, and enhance the customer experience.

Finally, some suggestions for future work are presented: 1) to carry out a similar study for the customer profile; 2) to automate the code developed to turn this study into a recurring segmentation exercise, periodically refreshing with new transactional data; 3) to apply the same methodology to other Recheio channels (non-HoReCa) to quickly segment and detect differences between the total channel and the online channel; 4) to enrich the clustering ensemble with other unsupervised methods; 5) to use this model for classifying new Families when adding/reorganizing the market structure.

Acknowledgements. Research partially funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, under the project UIDB/00006/2020, DOI: 10.54499/UIDB/00006/2020 (CEAUL) and ISEL. The authors thank the company Recheio from the Jerónimo Martins group for authorizing the publication of the information in this paper.

References

1. Branco, J.A.: Uma Introdução à Análise de Clusters. Sociedade Portuguesa de Estatística (2004)
2. Chen, Y., Guozheng, Z., Dengfeng, H., Chua, F.: Customer segmentation based on survival character. J. Intell. Manuf. **18**(4), 513–517 (2007)
3. Chen, N., Yang, Y.: The impact of customer experience on consumer purchase intention in cross-border E-commerce - taking network structural embeddedness as mediator variable. J. Retail. Consum. Serv. **59**, 102344 (2021)
4. Edwin, P.C.: Beyond the Numbers: Making Sense of Statistics. National Science Teaching Association (2012)
5. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. Wiley, Hoboken (2011)
6. Farahani, A., Rahiminezhed, H., Same, A., Immannezhed, K.: A comparison of partial least square (PLS) and ordinary least square (OLS) regressions in predicting of couples mental health based on their communicational patterns. Procedia Soc. Behav. Sci. **5**, 1459–1463 (2010)
7. Grabmeier, J., Rudolph, A.: Techniques of cluster algorithms in data mining. Data Min. Knowl. Disc. **6**, 303–360 (2002)

8. Hartigan, J.A., Wong, M.A.: A K-means clustering algorithm. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) **28**(1), 100–108 (1979)
9. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
10. Kettenring, J.R.: The practice of cluster analysis. J. Classif. **23**(1), 3–30 (2006)
11. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1: Statistics, pp. 281–297. University of California Press, Berkeley (1967)
12. Malhotra, N., Birks, D., Wills, P.: Marketing Research - An Applied Approach, 5th edn. Pearson Education, London (2017)
13. Pestana, M.H., Gageiro, J.N.: Análise de dados para ciências sociais: a complementaridade do SPSS, 6th edn. Edições Sílabo (2014)
14. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)