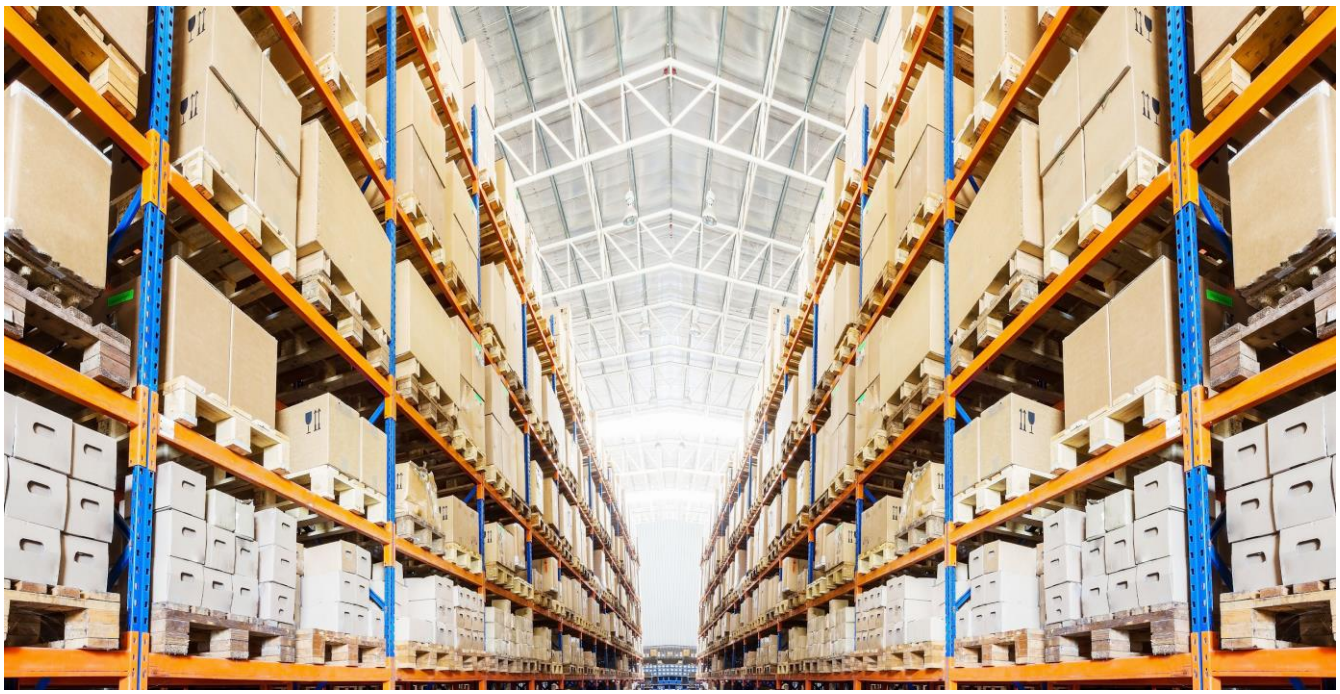
CATÓLICA
CATÓLICA PORTO
BUSINESS SCHOOL
PORTO

Wholesale Customers

Clustering Analysis and its Application in Management



André Macedo da Costa nº 355419075
Andrew El Matni nº 355419096
José Miguel Teixeira nº 355419031
Pedro Henrique Meireles nº 355419086
Sanghoon Han nº 357919153

Católica Porto Business School
Masters in Management – Data Mining
Professor: Vera Lúcia Migueis Oliveira e Silva
Porto, 10/04/2020

Index

1. Introduction

This study is based on a dataset called “Wholesale Customer”, which contains 440 observations and is composed by 8 variables, of which Channel and Region are categorical variables serving as labels to understand the Channel with which each customer buys and so called as factors. The other variables are Fresh, Milk, Grocery, Frozen, Detergents, Paper and Delicatessen, which indicate the annual spending (monetary units) of each of these products.

The goal of studying this dataset is to cluster the different wholesale clients and group them by the similarities they have when buying the different kinds of products, based on the spending relatively to each other, for example, if they buy more Fresh and less Frozen or if the spending they have in Grocery and Detergents are similar.

Firstly, an initial approach to the data analysis will be performed, starting by justifying the differentiation between the use of categorical and numerical variables, specifying their description, frequency, minimum and maximum observations in each one of the variables.

Later, insights about relations between variables will be collected, such as Revenue per Product Category, assessing the revenue per channel and the distribution of revenue and products per region.

This first dive into the data will set our first understandings and conclusions about the variables and each of their impact on the wholesale company. This way, we will get detailed information about each variable and then proceed to analyse them one on one, understanding the influence of the presence of outliers and removing them to compare with the first models.

Posteriorly, we will proceed with the clustering analysis, finding the right number of clusters through the elbow method and the Davies Bouldin. This will set up the right number of clusters in order to then segment and analyse the results of a Partitioning algorithm and also of a Hierarchical algorithm.

All of the analysis and processes will set up the pillars of the insights we will drive to the management context, where we give our thoughts about what this clustering analysis can bring to management in order to increase revenue, segment customers, channels and regions of this wholesale company. This whole study will be supported by graphs and plots with the goal of understanding better and make information more visible and easily understandable in each context, from the first analysis to the clustering.

2. Data Base

The Wholesale Customers data set refers to the clients of a wholesale distributor. They are from diverse regions of Portugal. This database includes the annual spending in monetary units (m.u.) on diverse product categories per client.

The database is from the business area, representing the amounts of money spent by 440 clients of the wholesale distributor and therefore the revenue of this one at the same time, in 2014. The number of attributes, that help us "characterize" the clients, is 6, each of which represents a category of products and the spending made by the clients in that same category.

2.1. Attributes

Categorical Attributes		
Name of the attribute	**Description of the attribute**	**Frequency**
CHANNEL	Channel used by the consumers: 1. HORECA (Hotels/Restaurants/Coffee Places) 2. Retail channel	HORECA: 298 Retail Channel: 142
REGION	Customers` region: 1. Lisbon 2. Oporto 3. Other	Lisbon: 77 Oporto: 47 Other: 316

Tabela 1-Categorical Attributes

Numeric Attributes			
Name of the attribute	**Description of the attribute (in m.u,)**	**Min. Value**	**Max. Value**
FRESH:	annual spending on fresh products	3	112151
MILK	annual spending on milk or milk products	55	73498
GROCERY	annual spending on grocery products	3	92780
FROZEN	annual spending on frozen products	25	60869
DETERGENTS_PAPER	annual spending on detergents and paper products	3	40827
DELICATESSEN	annual spending on and delicatessen (fine, unusual or foreign) products	3	47943

Tabela 2-Numeric Attributes

2.2. Stat Analysis

When accessing the products categories, we computed the main statistical calculations (mean, standard deviation, median, min and max) in order to compare them in terms of the revenue the Wholesale would make. This way, we concluded that the product that generates the most revenue are the Fresh products, then comes Grocery, Milk, Frozen, Detergents&Paper and in the end, with the lowest mean of revenue comes the Delicassen, easily illustrated by the following Pie Chart:

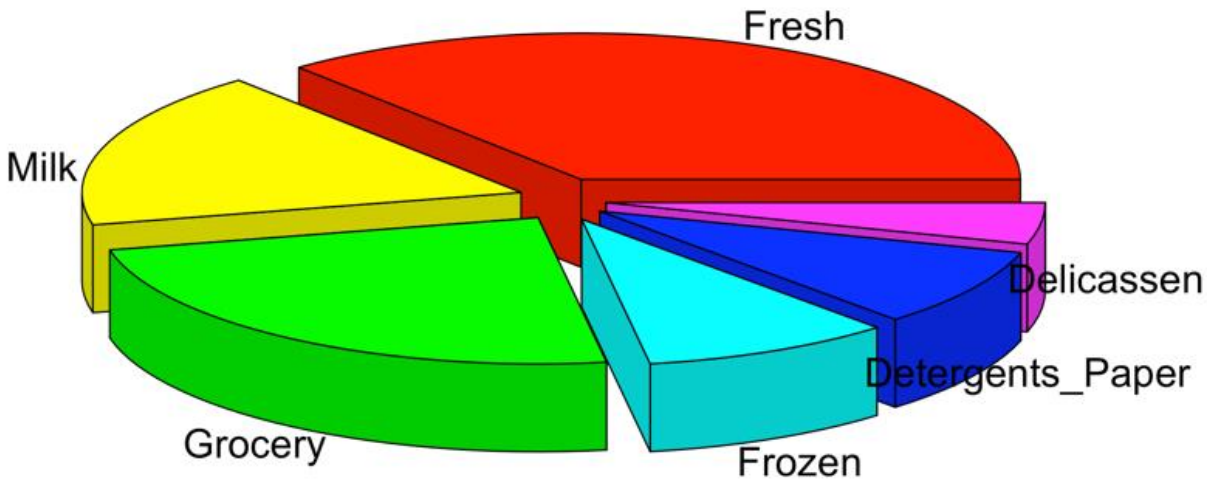


Figure 1-Revenue per Product Category

Posteriorly, we proceeded with the study of the mean of revenue of each Product per Region (Lisbon, Oporto or Other). Firstly, we noticed that we were comparing two of the biggest cities (Lisbon and Oporto) with the Other, which is the whole rest of the country. We can see that products like Grocery, Frozen and Detergents & Paper sell a lot more in Porto than in the rest of the country, this can bring us managerial insights in order to understand where to focus on distribution of this kind of products. Furthermore, we also noticed that, in the rest of the products (Fresh, Milk and Delicassen) we observed that they have more or less the same distribution: more revenue comes from the Other, then it comes from Lisbon and Porto has the least mean of revenue. This is the pattern of the product revenue per region, although the means of revenue differ from Product to Product.

After assessing the revenue per region and per product, we aimed to analyze it per channel of distribution, this way we conclude that the Retail channel is used way more than HORECA, with a mean of 44865.8 compared to 25428.24 in the HORECA. This analysis also giving valuable managerial insights in order to segment and decide in which Channel to invest more.

3.Analysis Method

3.1. Data cleaning and preparation

The analysis of the attributes that generate the dataset, started with the process of data cleaning where we identified the most important attributes which basically are the numeric ones that represent the revenue of the Wholesale/spending of customers in different categories of products and then we removed the categorical ones, once they were not important in what our clustering process concerns.

In order to assess the information within the same scale and not to be biased by the different measurements, we normalized the data to achieve more data integrity and reduce data redundancy. If the concern was the total spending across all product categories, then the raw data could be used for clustering. On the other hand, if the product categories should be weighted equally, then we needed to normalize the features. This led us to compare the data before and after the normalization and this being said we came up with the following box plot with the data after normalization:

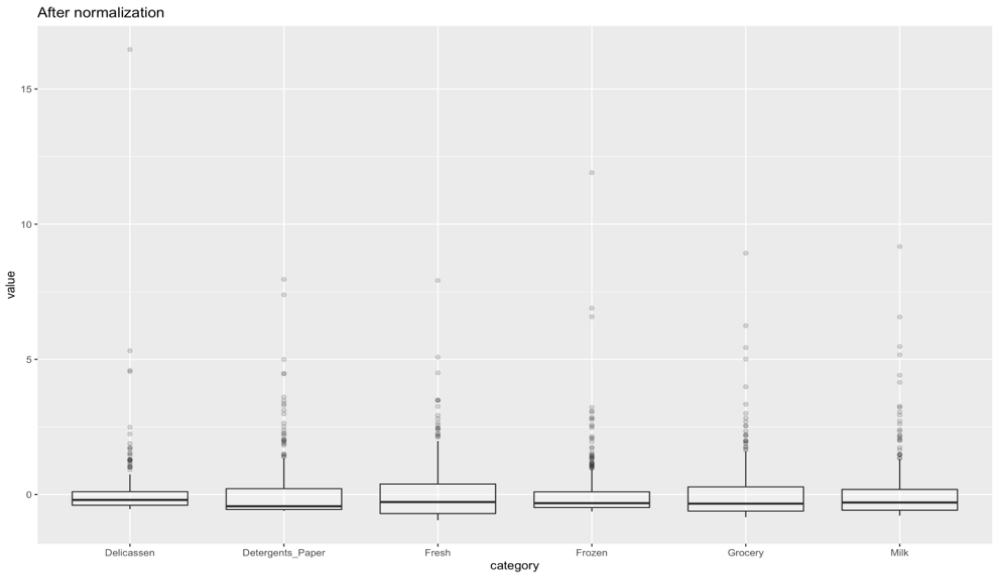


Figure 2-Data distribution per Product Category

Here we can already detect some possible outliers and assess the distribution of the data by each product category. This way, we see that in general all the Products are positively skewed, since the mean of the observations is higher that the median (which shows in the box plots) and tell us that most of the observations are above the median. But this can be affected by the presence of outliers, and it is, and for that reason, the next step is to approach them.

In the matter of outliers, we removed them with the function Zscore and we created box plots to analyze the influence of removing outliers with a 95%, 99% and 99.5% confidence interval and the result was that as the normalization increased, the distributions and the median of the product categories boxplots decreased, this way we can see that the presence of outliers influenced the distribution and the median in the previous box plots. With the 97,5%

normalized data that we will frequently use on this analysis, there were 66 abnormalities. But these 66 (top 2.5% clients) anomalies are clients who spend more than average consumers.
To access this change, we used the following commands:

```
#Draw boxplot of normalized data

drawboxplot(Perc97.5,"95% Normalized")
drawboxplot(Perc99,"97.5% Normalized")
drawboxplot(Perc99.5,"99% Normalized")
```

Figure 3- Boxplots of Normalized data

3.2. Choice of the Number of Clusters

Choosing the ideal number of clusters is a real important part for any business and so, we need to identify and divide each object in such a way that objects in the same cluster (group) are more similar to each other than to those in other clusters.

In order to achieve the ideal number of clusters we proceeded by doing two tests: Elbow Method and Davies Bouldin test.

3.2.1. Elbow Method

The idea of the Elbow curve method, in order to be possible to realize the k-means clustering is to define the ideal number of clusters where the total within-cluster sum of square errors (SSE) or WSS is minimum.

The value of SSE or WSS is viewed by the elbow method as function, where the number of clusters is chosen when adding another cluster doesn't improve much better the value of SSE.

In this case, we decided to run our data in three ways: the first one using a confidence interval of 97,5%, removing the outliers which were 2.5% above. Secondly, we did the same for the next 2 plots, but working with respectively 99% and 99.5% confidence.

The percentage removed from the total data was fulfilled with outliers which were causing damage to our results.

In this case, we are going to use a confidence interval of 97,5%, which gives us a k equal to 4 to be our ideal number of clusters.

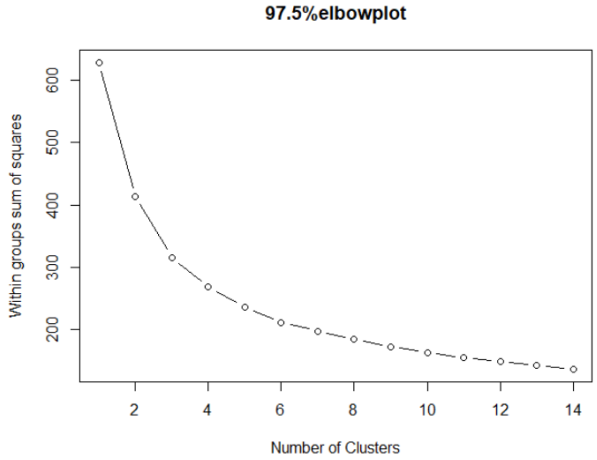


Figure 4- Elbow Plot

3.2.2. Davies Bouldin

Davies Bouldin Index (DBIndex) is a method used for measuring the validity of the cluster. According to DB, the optimal value of k is determined by the minimization of the ratio that relates the inter-distance between the points and their respective intra-distance.

To analyse this index, we used the same confidence intervals that were used in the Elbow Method test. (97,5%, 99% and 99,5%).

In the same way, we chose to use the 97,5% confidence interval, which was the value that reflected the best result. By using this method, the value of k will be equal to 4.

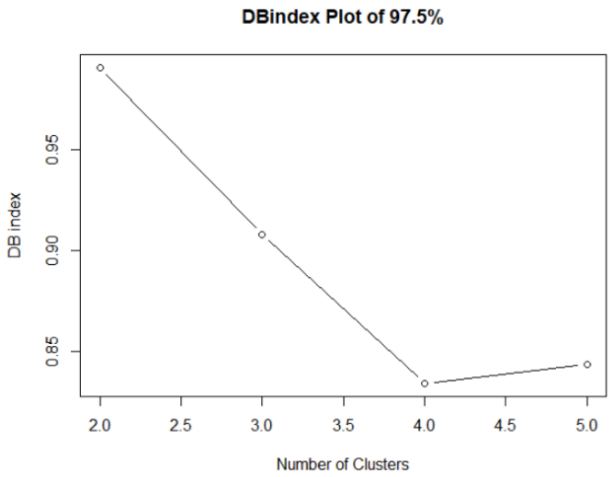


Figure 5-Davies Bouldin Plot

3.3. Major clustering approaches

The process of calculating the number of the wholesale`s customers contained in each cluster was developed according to two main methods: partitioning clustering and hierarchical clustering. These clusters were based on all numeric attributes of our dataset, and each attribute represents a category of products where the mean of the spending of each customer was inputted.

3.3.1. Partitioning algorithm

Regarding the partial clustering, it was used the **k-means algorithm**, according to which the definition of the size of each cluster aims to minimize the distance between the various observations and the cluster centre closest to each of the same.

As we referred before, for the k-means algorithm, we have to define the number of clusters. Using the elbow curve and Davies Bouldin, the number of cluster centres we defined to help us group our data was **4** (a, b, c and d from the following graphs) with the 97.5% data used.

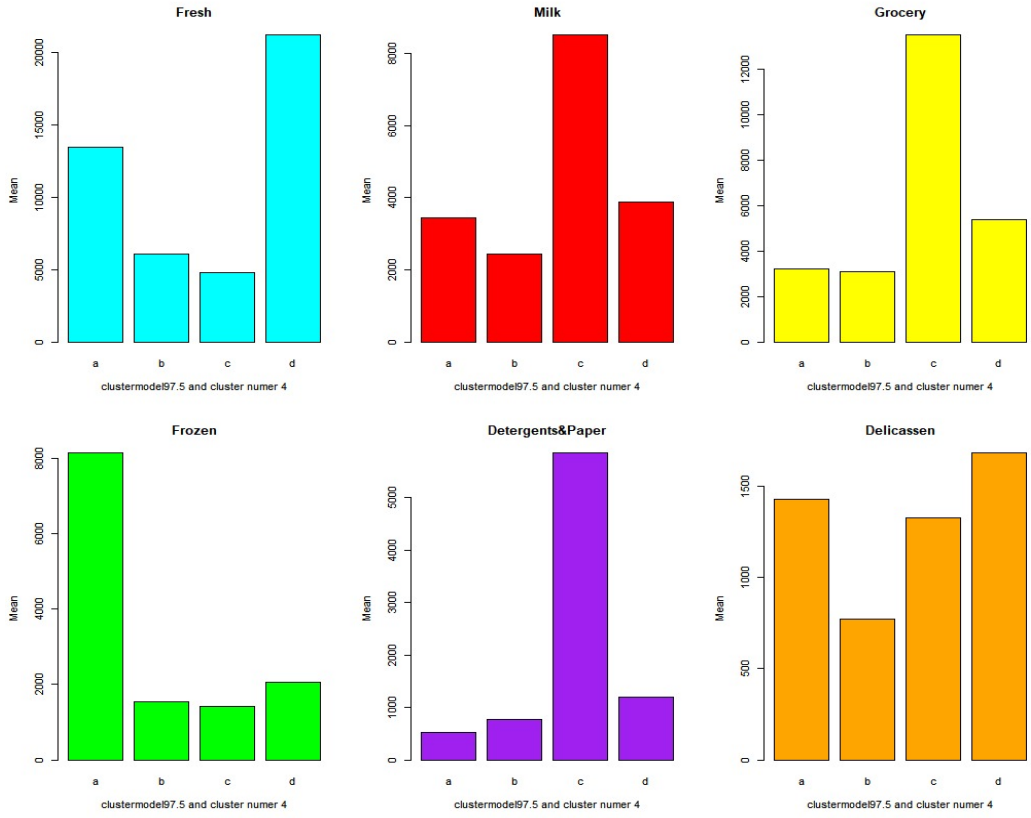


Figure 6-Partitioning algorithm plots

3.3.2. Hierarchical algorithm

As k-means is susceptible to initialization problems and outliers, we resort to hierarchical clustering also. Here, at the beginning of the process, each element is in its own cluster (individually). Subsequently, the clusters are sequentially combined into larger clusters until all elements are conducted for the same cluster.

We were aware that there are three main methods of hierarchical clustering, however the single link method exhibited a chaining phenomenon due to combining, at relatively low thresholds, observations linked by a series of close intermediate observations, so it was considered defective. In order to obtain the best approach, we opted for a comparison between the other two different agglomerative hierarchical techniques: **MAX (complete link)** and **Group Average Proximity**.

We concluded that the two methods could be considered efficient for the clustering task, as we compared the differences and the common subsets (103) created by the two methods in the following tanglegram:

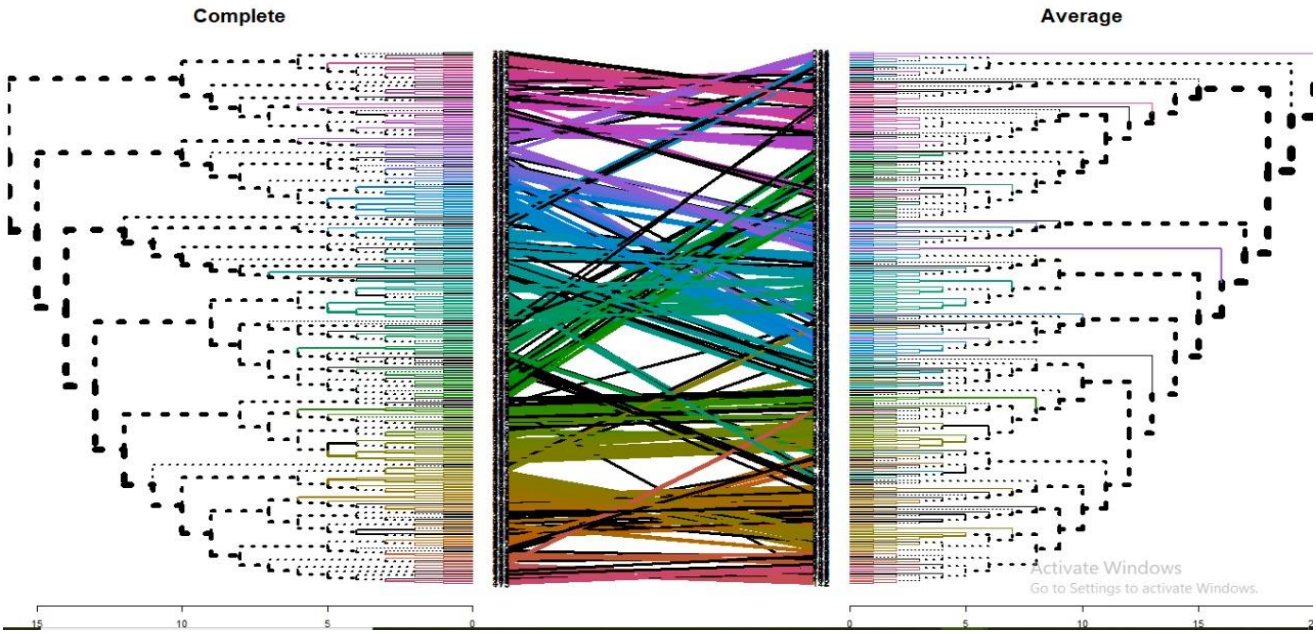


Figure 7-Hierarchical algorithm - Tanglegram

After all, we end up choosing, to proceed with the analysis, the complete link method and ultimately created **3** different clusters from this approach, once this number of clusters is the one, looking from this approach, that differentiates in the best way the groups of clients:

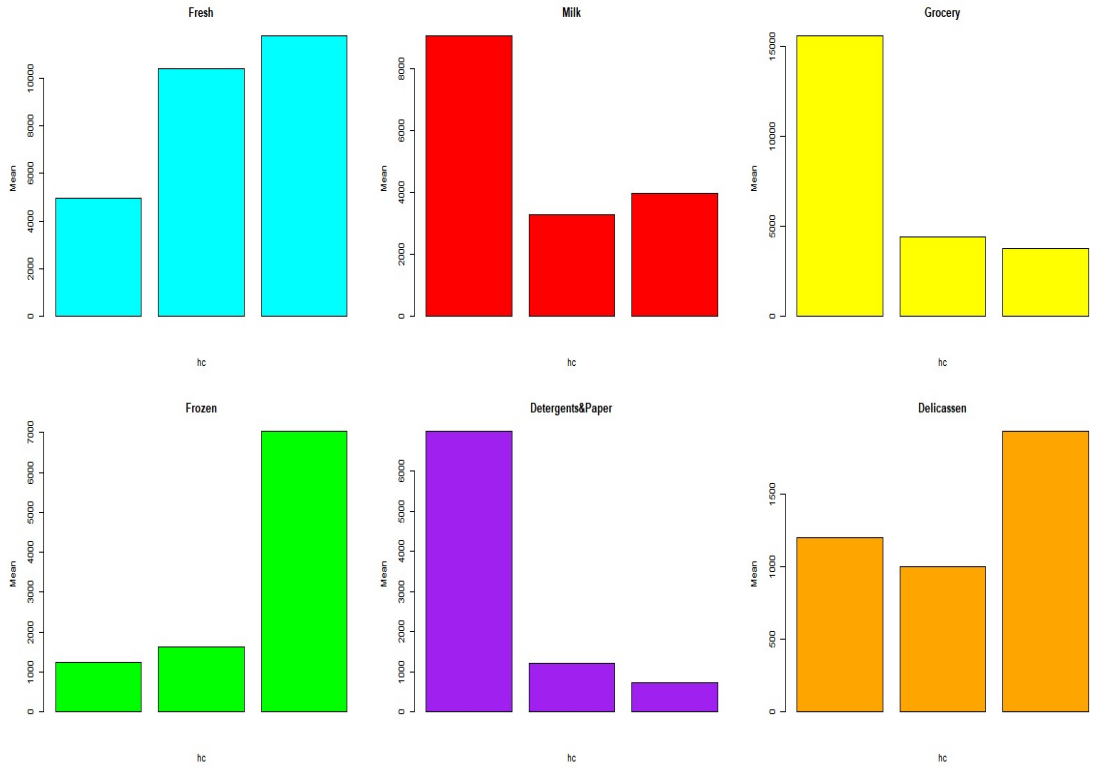


Figure 8-Distribution of observations in complete hierarchical clusters

4.Wholesale`s management context

This phase of the project has proved to be crucial to establish a context, in the area of Management, that favoured the practical application of the results obtained with the creation of the clusters. The context in cause translates into the existence of a Wholesale company that pretends to analyse the spending patterns of its customers in different categories of products, that would translate into its revenues. In that order, it is fundamental to the business practises and tactics of the company to establish the different types of clients that buy its products and what categories of products those clients tend to acquire the most.

Once both the Elbow Curve and Davies Bouldin show that the best number of clusters is four, we will take as example the clustering developed by using the k-means approach for defining the number of different types of clients that the company has, without the outliers (66 abnormalities) which were deleted. These are the 4 type of clients the Wholesale has based on their profiles:

Type I (cluster a): 164 clients. The spending/buying pattern of this clients is directed towards Fresh, Frozen and Delicassen products. This is the type of client who spends the most money in Fresh products compared to the other ones, once the average of money spent in this

products by this type of client is more than 8.000 m.u and all of the other types of clients each don`t spend more than 2.000 m.u in average in this category of products.

Type II (cluster b): 41 clients. This type of client, compared to the other ones, is the one who spends in average the lowest amount of money in every category of products, not having a special type of products that he prefers to buy in mass.

Type III (cluster c): 78 clients. The spending/buying pattern of this clients is directed towards Milk, Grocery, Detergents & Paper and Delicassen products. This type of client of the Wholesale is especially important for the wholesale in the Milk, Grocery and Delicassen products once is by far the one who spends more money in this category of products.

Type IV (cluster d): 91 clients. The spending/buying pattern of this clients is directed towards Fresh and Delicassen products, being the ones that in average spend more money in these two categories.

The clustering method is fundamental to distinguish the type of clients, based on the whole portfolio, that the Wholesale has and to implement business and selling strategies oriented to each type of client`s preferences of products. It could be very important also to know which region or distribution channel each type of clients is inserted in, so that a difference distribution strategy could be set by the company, in order to avoid waste of money and to increase the profit to the maximum. We can recommend that the Wholesale focus on sending to the Type I clients, Type III and Type IV clients some special discounts and the latest promotions regularly on Frozen, Detergents & Paper and Fresh products, respectively. This type of clients is fundamental to the operation of the company on each of these categories, because they are the ones who generate the greatest amount of revenue on this particular set of products. This could raise the engagement of the clients with the company, generate more sales in the future and customize the offer significantly.

5.Conclusion

Through the development of this report, it was possible to obtain an in-depth knowledge about the database and the variables studied, as well as the existing relationships between them.

Afterwards, a cluster analysis was developed, where, after estimating two clustering methods, the group realized that the best for the study in question end up being k-means, with k= 4, having moved to a graphical analysis in order to provide the best possible understanding.

Regarding the management context, any management team would like to understand the types of customers who are, or could be, doing business with their company. Each type of customer has a different value for the company, so, they should be managed in different ways. The aim to have a customer portfolio analysis is to divide customers into mutually exclusive clusters in order to identify profitable and valuable customers. Therefore, the company can apply marketing tactics to preserve and develop valuable present and future customers.

In conclusion, with the development of this clustering project it was possible to upgrade our programming skills on R and understand the importance of a clustering analysis for a company, on a practical level.