CATÓLICA
CATÓLICA PORTO
BUSINESS SCHOOL
PORTO

Data Mining – Clustering Project
Group Assignment

Vasco Moutinho 355422066
Simão Magalhães 355422067
Sofia Fernandes 355422038
Tiago Silva 355422039
Ana Margarida Leite 351219008

Abstract

In the scope of the Data Mining curricular unit, it was proposed to study and work the "Online Retail" database in order to develop a clustering analysis. It contains 406,444 observations and 8 variables, providing information about the sales of products online, with the description and quantity sold supported by the respective consumer ID, the country where the purchase was made and the unit price, among other variables.

The main focus of the work was to perform an analysis of sales from the product perspective, in order to cluster the different products taking into account the variables average unit price, quantity and frequency of purchase.

We began the work by filtering the data necessary for our analysis, removing the credit notes and zero price sales, and reduced the time spectrum to 12 months (from 12/9/2010 to 12/9/2011). Next, we characterized the variables appropriately and created a column representing the revenue generated per sale ("Sales") by multiplying the unit price by the quantity sold.

Since the initial table contained several observations for the same consumer or product, that is, for the same StockCode (and its respective description) and consumer ID, since a consumer can buy the same products several times, two new databases were created (Qt_P_Agg and Qt_C_Agg). The first one expresses a product view, being present unique descriptions for each product and its repetitive aggregated values, being these, the quantity sold (aggregated), revenue generated ("Sales"). To complete the analysis, new columns were created, respective to the purchase frequency of each product and its average unit price, calculated by dividing the "Sales" by the quantity. Finally, the outliers were removed to standardize the values and to avoid a large discrepancy of data.

The second table, on the other hand, portrays the consumer's point of view, with the consumer's ID (non-repeated values) and its respective quantity purchased and revenue ("Sales"). This database allows us to analyze which are the biggest buyers and the revenue they generated.

All these steps are crucial, since they allowed us to make a descriptive analysis of the data, having calculated the focus on standard deviation, the mean, maximums and minimums, the sum of sales, the correlation between the different variables, and then a brief study of the variation of revenue per customer, product, country and date.

Before proceeding to the clustering analysis, we normalized the values and tried to find the optimum k for our analysis, using different methods (k-means, Davies Bouldin and Hierarchical Clustering). Finding the correct k (k=6), we divided our observations into 6 clusters, preceding the computation of different plots to draw conclusions about the categories appreciated in each of the clusters.

Introduction

E-commerce has become a fundamental part of modern retailing. With the development of technology and the increasing reliance of consumers on online shopping, online stores have become a viable and even preferred alternative for many shoppers.

As mentioned above, we analyzed and worked on the "Online Retail" database in order to improve our understanding of the database and the clustering techniques adopted, with a view to presenting results and conclusions that are particularly relevant for the purposes of the study at hand.

In this sense, we proposed to analyze the information contained in the database, as well as the clustering techniques employed, in order to identify patterns and trends that may offer relevant information for the development of future strategies.

Cluster analysis proves to be quite important as it allows the grouping of observations and variables, based on information gathered from the database. Thus, it is possible to obtain a description relationships between different variables, enhancing their similarities to identical values, and their differences to clusters.

Database Analysis

The database provided for this project was "Online Retail", which has 406,445 records of online sales worldwide. This database contains 8 variables (columns), which are:

- *InvoiceNo* ((Invoice number acossiated to the purchase)
- *StockCode* (Stock Code of the product bought)
- *Description* (Product Description)
- *Quantity* (How many units of that product were bought in that transaction)
- *InvoiceDate* (Date and time of purchase)
- *UnitPrice* (Unit cost of the product)
- *CustomerID* (Customer ID)
- *Country* (Customer's Country)

This database does not have any currency associated with online sales in the different countries, so it was assumed that the currency for the UnitPrice variable values would be in euros (EUR), making it easier to compare the different values.

Database Preparation

Before starting the analysis, the database was first prepared. This step is crucial, as it allows to ensure a higher quality database to perform the planned analysis.

To format the data it was decided to remove credit notes (where the invoice began with a "C", which corresponded to negative values in the Quantity column) and offers (where the "UnitPrice" value was equal to zero), since it was not considered necessary to study purchasing behavior. Thus, the focus of the analysis was only on the purchase transactions regardless of whether the products were returned or offered later.

Next, and using the "str()" command, it was verified that the InvoiceNo, StockCode, Description, Country and CustomerID columns had character type values, so these same columns were classified as being categorical, that is, with the R command "as.factor". For UnitPrice, which also had character type values, we changed "," to "." and classified it as numeric. Next, to support the descriptive analysis, we added the Sales column, which is the result of multiplying Quantity by UnitPrice. After this first phase, it was decided to study the behavior of the transactions between 12/9/2010 and 12/9/2011, that is, the time spectrum was reduced to 12 months.

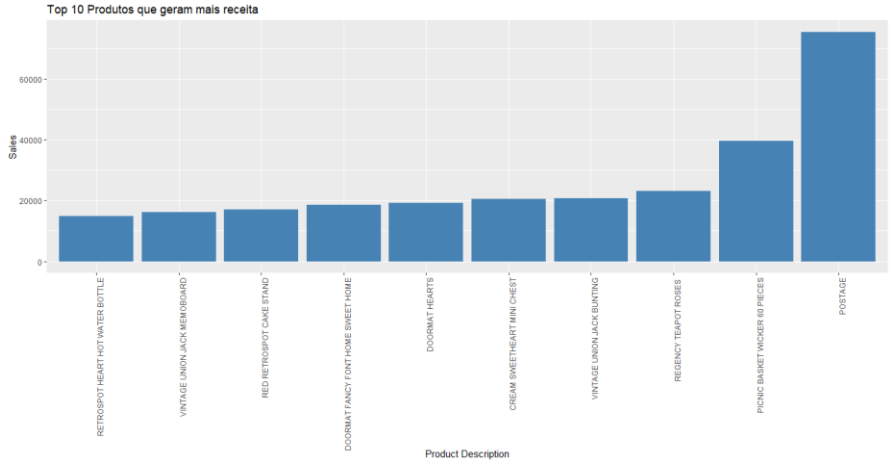
After this reduction of the database under analysis, a new data frame was created, called "Qt_P_Agg", which served as the basis for the study. This database contains fewer observations and only five variables: Description, Sales, Quantity, UnitPriceM and FreqPurchase.

To construct the "Qt_P_Agg", the aggregate code was used, and for each description (Description) of the product sold, the values of its sales (Sales) were added. The same process was applied to add the Quantity column, and so the value of the quantities sold was added to the total sales value of each product. The column with the variable "UnitPriceM" was calculated through a weighted average of the products, that is, as the product is not always sold at the same price, it was decided to add to the value of its sales and divide by the quantities sold. Finally, the column "FreqPurchase" was added, which tells how many times the product is sold in different transactions.

In order to further reduce the sample, it was decided to remove the outliers from the "Qt_P_Agg" database.

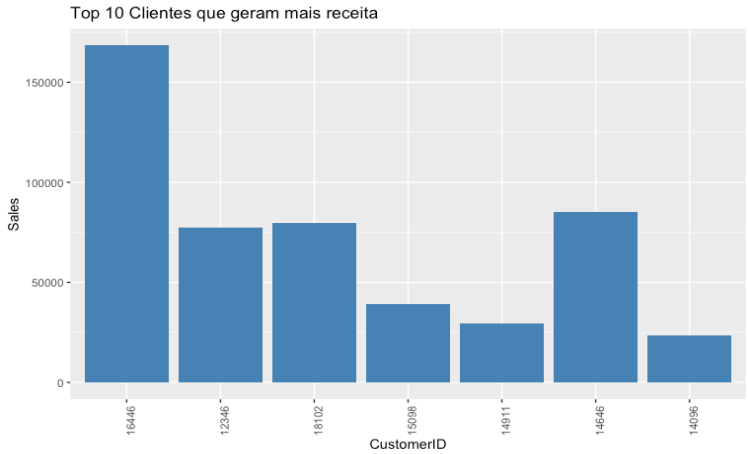
Descriptive Analysis

After preparing our data, a descriptive analysis of the numerical variables was performed. In the months under analysis, the value of total sales was 8,627,895€ and, as can be seen from the graph, the product that generated the most revenue was Postage.



Graph 1- Top 10 Product that generate most Revenue

In terms of customers, we wanted to understand which were the ten that bought the most in terms of value and not quantity. And for this, the following chart was elaborated:



Graph 2- Top 10 Clients that generate most Revenue

As can be seen from the graph below, there is a large discrepancy between the UK and the rest of the countries. France, Germany, the Netherlands, Australia and EIRE share similar values, with the remaining countries in the top 10 having slightly lower values, as can be seen in the chart:

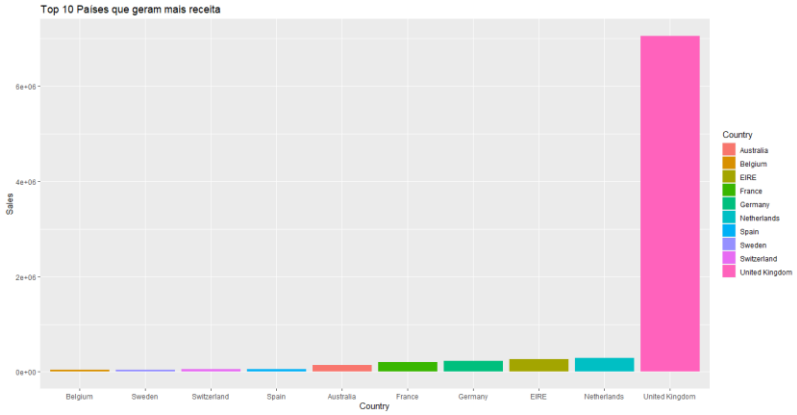


Gráfico 3- Top 10 Países que Geram mais Receita

Next, we proceeded to the statistical analysis of the variables Sales, Quantity, UnitPriceM and FreqPurchase, from the database "Qt_P_Agg" as can be seen in the table below:

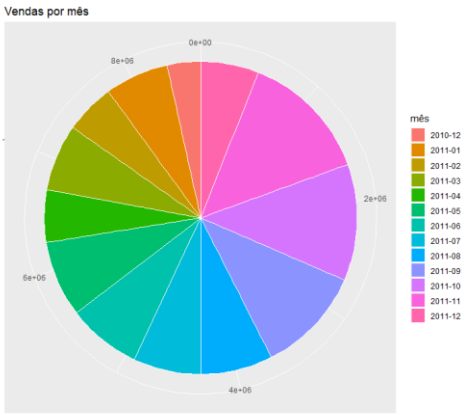
Qt_P_Agg	MIN	MEAN	MAX	Standard Deviation
Sales	0,003	1 264,92	75 454,95	2 567,44
Quantity	1,000	595,78	3 186,00	742,52
UnitPriceM	0,001	3,75	744,15	18,08
FreqCompra	1,000	66,39	1 056,00	83,76

Table 1- Maximum, minimum, mean and standard error of variables

Regarding the quantity of product sold, the average is 595.78 units, while the product with the highest quantity sold was the Pack of 12 pink paisley tissues with a value of 3 186 units sold. Analyzing the weighted average unit price of the products, it appears that the average price charged was 3.75€ and that the product sold at a higher average price was Dotcom Postage, for 744.15€.

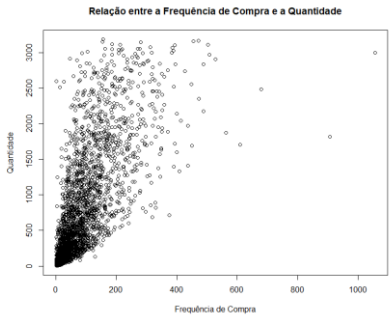
Furthermore, observing the frequency with which each product was purchased, we see an average value of 66.39 times, and we see that the product that was most often purchased was Postage, with 1,056 transactions.

Finally, regarding sales, the average value was €1,264.92 and the product with the highest revenue was Postage with a value of €75,454.96.



Graph 4- Sales per Month

As shown in graph 4, it was possible to verify that November 2011 was the month that stood out in terms of sales volume. Moreover, it is possible to observe that from August 2011 on, the sales volume has been increasing tendentially, compared to previous months.



	Quantity	FreqCompra	UnitPriceM
Quantity	X	0.7592396	-0.07204207
FreqCompra		X	-0.01609991
UnitPriceM			X

Table 2 and Plot 1- Correlation between the different variables

From the analysis of the table and plots above, which shows the correlations between the variables, we want to highlight the positive relationship between "FreqPurchase" and "Quantity", at a value of approximately 0.759, which demonstrates that products that are bought in larger quantities are also bought more frequently.

Proposed Method

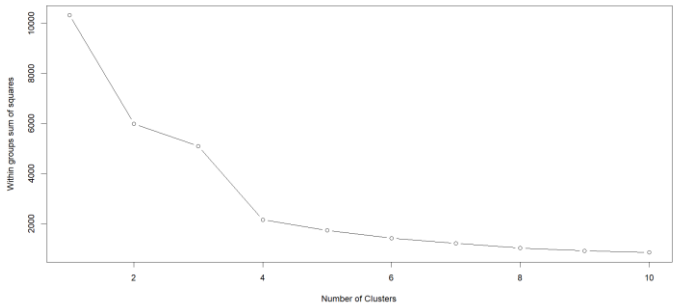
In order to perform the clusterings analysis, we used only the approaches studied in class, which are the partition and hierarchical algorithms. It should be noted that all variables used for the elaboration of the clusters are of the numerical type.

First, the heuristic method was used, specifically the k-means, in which each cluster is represented by its center, discovering which is the center of the closest cluster, using the average. Next, the distribution of clusters according to the distance to the centroid was drawn up, followed by a re-estimation of the centroid. This process ends when it is no longer possible to re-estimate the cluster points.

As the objective of this work is the analysis of product behavior, we intend to divide the sample into different clusters of products according to their buying contexts. To this end, it was found pertinent to

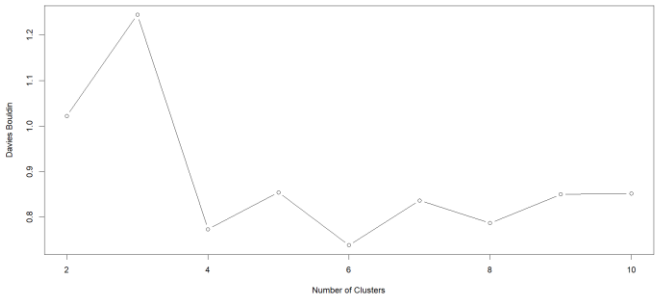
analyze the variables quantity, weighted average price and frequency of purchase of each product, that is, the number of times each product is purchased, from the "normed" data frame. The data frame mentioned above is the result of the normalization process of the numerical variables of the database "Qt_P_Agg", according to the formula "z_score_function", given in class.

In order to obtain the optimal number of clusters (k) for the variable, that is, the k that gives the best final performance, as well as, a smaller SSE and that guarantees a smaller dispersion between clusters (higher cluster quality), the Elbow method was used. We began by establishing a number of random seeds equal to 7 and 1000 optimization steps, this value being given by default in the R code, when using the "k-means" code. After viewing graph 8, realized with the "plot" function, it is possible to observe the relationship between the four variables analyzed in the cluster. However, it is concluded that no solution can be drawn from it, because the graph does not present its characteristic elbow, that is, it is inconclusive.



Graph 8- Elbow Method

As seen previously, the Elbow Method sometimes does not work properly, because it creates some ambiguity among the data, specifically in the k to be chosen. In order to solve this problem, we applied the Davies Bouldin Index method, through the use of the "clusterSim" package in R. This function generates a model that creates a graph in which it is possible to compare the Davies Bouldin Index for different numbers of k. According to DB, the optimal value of k is determined by the minimum of the ratio that relates the inter-distances between points and their respective intra-distances. After applying this method, as seen in graph 9, it proves that the optimal k is equal to 6, as this is the minimum value found in the graph, being equal to 0.7384891.



Graph 9- Davies Bouldin

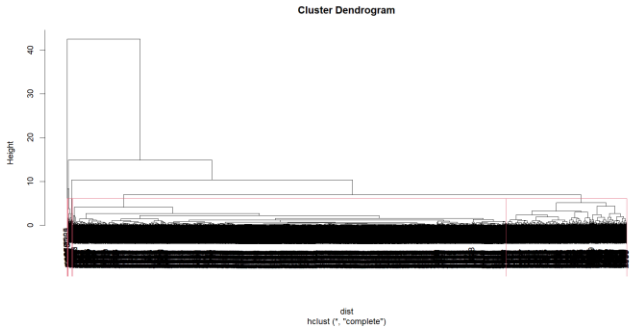
After finding the optimal number of clusters, we conclude the analysis with the partition algorithms, using the k-means command, with the value of k found. Next, two new columns were added to the "normed"

data-frame, being them the "description", which indicates which product is allocated to each cluster, and its respective "cluster".

Moving on now to the analysis of the hierarchical algorithm, this corresponds to a set of aligned clusters, which are organized as a tree. This model does not assume a particular value for k, it only indicates the similarity of the observations. Thus, the agglomerative method was chosen, in which at each step there is a merging of the pairs of clusters. The hierarchical method sorts the rows and/or columns according to their similarity. This method facilitates the visualization of correlations in the date.

It was first assumed that the proximity to be adopted would be the minimum possible between the points, through the "single link". However, no conclusions were obtained, so the "complete link" was chosen, in which the proximity used would be dictated by the most distant points belonging to different clusters, thus further reducing the weight of outliers. Thus, this process begins with each element in its own cluster, being sequentially grouped into larger clusters until all elements are grouped in a single cluster.

To perform this analysis we started by calculating the distance matrix (using the code "dist" in R), and then used the code "hclust", choosing the complete method. Subsequently, the dendogram (chart 10) was prepared and six rectangles were defined in it. We grouped into six clusters, because it is from this level that it is visible that the clusters are closer, that is, they are more similar to each other. Finally, a new column was created in the data frame "named", which indicates the cluster that each observation represents, with the objective of comparing the results obtained in the two approaches.



Graph 10- Dendrogram

Results

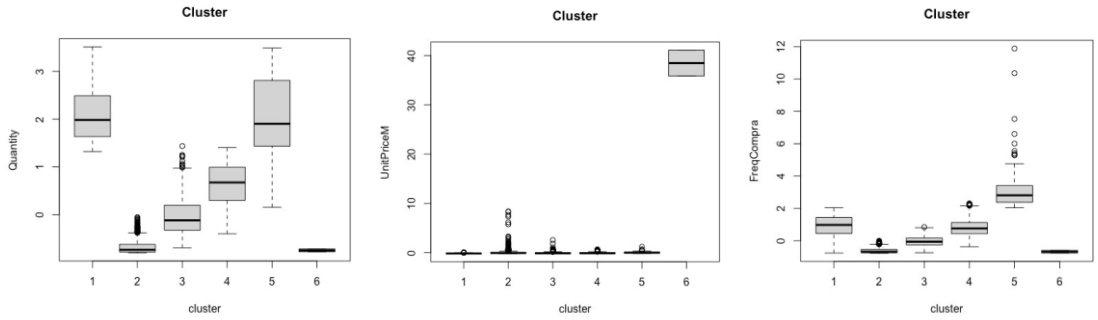
After dividing the data into 6 clusters, using the k-means method, based on three criteria ("Quantity", "UnitPriceM" and "PurchaseFreq") it was considered convenient to observe how these vary from cluster to cluster. The analysis was based on the products that stand out in terms of quantity and unit price in each cluster, in order to present a ratio that can be complemented with the frequency of purchase of these products. In this sense, we tried to identify the product behavior and enable a better understanding of the demand.

Boxplot 1 groups the products into clusters according to quantity, while boxplot 2 groups them based on the average unit price. When analyzing the first boxplot, we observe that cluster 5 stands out because of the range of values that make it up. However, through the analysis of boxplot 2, these products are sold for a low average unit price. Therefore, when analyzing boxplot 3, which investigates the frequency of purchase of the products, it was found that this cluster is purchased more frequently, reaching higher values compared to all the other clusters. This same interpretation can be applied to cluster 2, since it is

similar to cluster 5, although it shows lower values in the three criteria analyzed. Regarding clusters 3 and 4, the analysis is quite average, since the values do not stand out positively or negatively in terms of quantity, price, and frequency. Cluster 4 shows only higher levels in all three criteria compared to cluster 3.

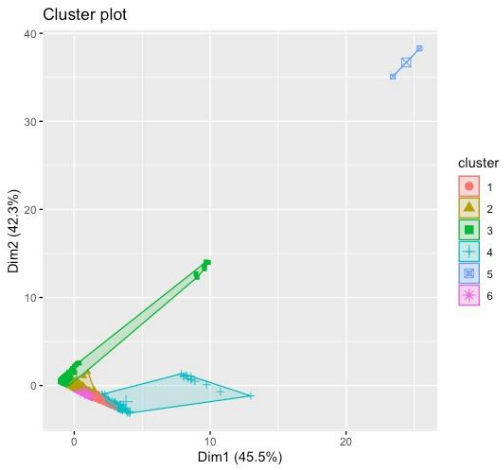
In the opposite direction, it was possible to see in boxplot 1 that cluster 6 stands out, since it groups all the products that were purchased in smaller quantities. Another aspect to highlight is the fact that these products have higher average unit prices than all other clusters. By the combination of these two aforementioned reasons, this group of products is purchased less frequently, which may indicate that they are products that are characterized as one-off purchases, as is the case of the product "DOTCOM POSTAGE".

To propose a strategy for the company, it is pertinent to further analyze cluster 2, since it stands out from all the other clusters, but by the negative aspect for presenting very low levels in the three criteria. In this sense, it would be important that the analyst verify the type of products that made up this cluster and, together with the marketing department, develop a strategy to leverage both the sales and the frequency of purchase of these products. To achieve this goal, the company could consider changing the price of the product or propose some effective marketing strategy, such as advertising campaigns or special promotions. In addition, it is necessary to investigate the possible reasons for the low performance of the products that constitute this cluster and to evaluate market consumption trends to identify opportunities for improvement and differentiation of the products offered by the company.



BoxPlots 1,2 and 3- Box Plots of each variable divided by Clusters

The following graphs are a complement to the previous ones and are intended to show the spatial location of the clusters calculated through this first method. As mentioned above, it was found that cluster 5 presents the largest size, and is not so concentrated with the other clusters.



Graph 11- Cluster Plot

Conclusion

The clustering analysis of the "Online Retail" dataset started by pre-processing the data, including adding a new column (Sales) and modifying one of the columns (Invoice Date gave rise to a column with only the transaction date). We created a new database in order to be able to work and do a better study of what we intended. Next, we performed a descriptive analysis of our data, summarizing and presenting relevant information graphically and through tables.

In terms of normalizing the data, we used the Z-Score method to normalize the values of our numerical variables. We then determined the optimal number of clusters by using different methods (k-means, Davies Bouldin and Hierarchical Clustering) and decided on 6.

Based on the analysis of the clusters and the graphs presented, we can conclude that cluster 5 is the one with the most sold products in terms of quantity and purchase frequency, although they have a lower average unit price. Cluster 6, on the other hand, groups products that are purchased in smaller quantities and with higher average unit prices than all the other clusters, indicating that these products are purchased on an occasional basis.

Cluster 2 is the one that presents a negative performance, with low values in the three criteria analyzed, being necessary to better analyze the products that comprise it to develop an effective marketing strategy to leverage the sales and purchase frequency of these products.

This analysis allowed us to conclude that clustering methods can differ in the values of k that they consider optimal, and that our model separated groups by criteria. Having considered all this, we are now prepared to apply clustering analysis to different types of data sets that we may use in the future.

Finally, it is important for the company to evaluate market consumption trends to identify opportunities to improve and differentiate the products offered in order to remain competitive and meet customer demand.