

**NOVA**

**IMS**

Information  
Management  
School

# Machine Learning II Project

Customer Segmentation:

The Key to Unlocking Business Growth and Success

## Group 24

Lourenço Passeiro - 20221838

Peter Lekszycki - 20221840

Tomás Gonçalves - 20221894

**9th June 2024**

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

# Executive Summary

This project involved conducting a detailed data analysis and customer segmentation for a retailer based in Portugal. The goal was to understand customer behaviors and preferences in order to effectively tailor marketing strategies.

The key challenge was to manage and interpret distinct datasets containing varied customer demographics, purchasing patterns, and behavioral metrics. The goal was to identify distinct customer segments to create targeted promotional strategies, addressing the specific needs of each.

## Methodology and Approach:

Our approach included several stages:

1. **Exploratory Data Analysis:** We visualized the datasets to better understand distributions, outliers, and patterns.
2. **Data Preprocessing:** This stage involved cleaning data, imputing missing values, and transforming features to better suit analytical models. Notable transformations included converting categorical data into binary formats and normalizing data sensitive to outliers. Also conducted correlation analysis to refine the features, dropping redundant or less informative ones.
3. **Customer Segmentation and Clustering:** Autoencoders were employed to reduce data complexity. We explored several clustering algorithms including K-Means, DBSCAN, and hierarchical clustering to segment the customer base. Clustering efficacy was assessed using the silhouette score and visualizations through U-MAPs and SOMs.

## Key Findings and Results:

- The final clustering solution indicated a strong alignment between K-means and Ward's hierarchical clustering approaches, suggesting robustness in the segmentation process.
- We identified seven distinct customer segments, each with unique characteristics such as purchasing power, store visit frequency, and product preferences.
- Targeted promotions and strategies were developed for each segment, generating insights from their behavior. For example, high-value customers were targeted with premium product bundles, while family-oriented segments received promotions on bulk and bundle purchases.
- Unique segments such as 'Fishermen' and 'Pet Parents' were identified with specialized strategies to address their unique needs.

# Exploratory Data Analysis and Pre-Processing

Before doing any pre-processing, we started by taking a look at our data.

- The 'lifetime spent' categories have many extreme outliers. Further on we will decide how to deal with them.

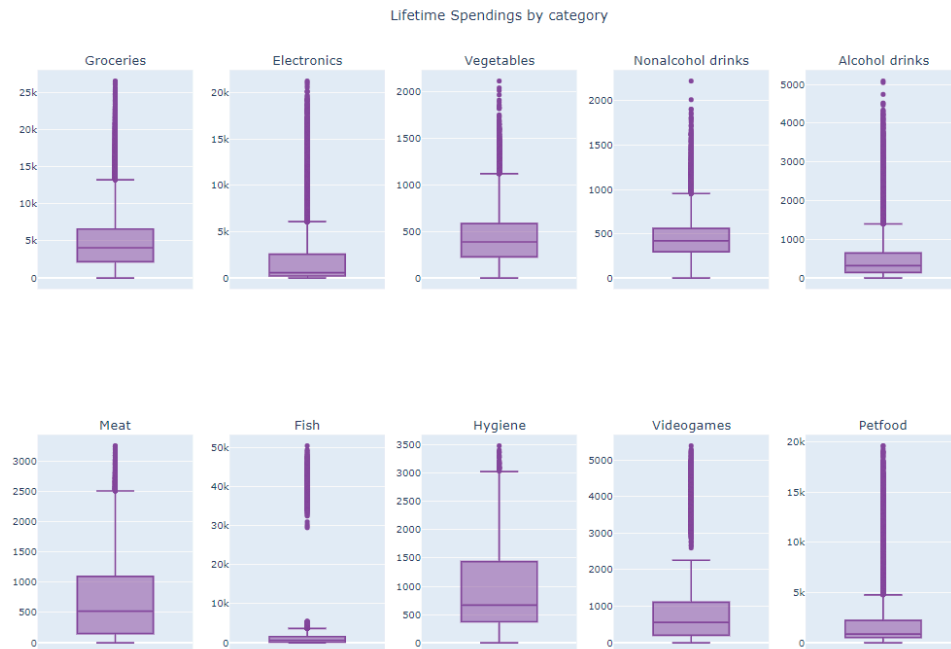


Fig.1 - Boxplot of our 'lifetime spend' categories

- *kids\_home* and *teens\_home* follow similar, right-skewed distributions, as expected;
- Customers mainly shop in the morning, but the peak hour is at 17.

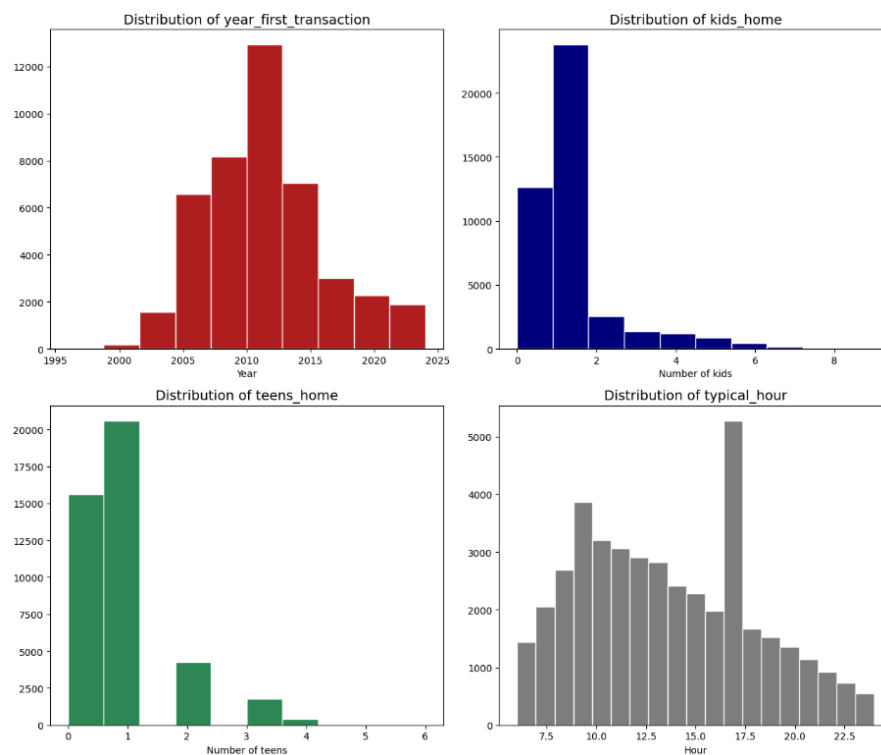


Fig.2 - Histograms of some demographical features

- Our customers **are overwhelmingly located** in the **metropolitan area of Lisbon**, with a few situated also in Ericeira and Peniche.

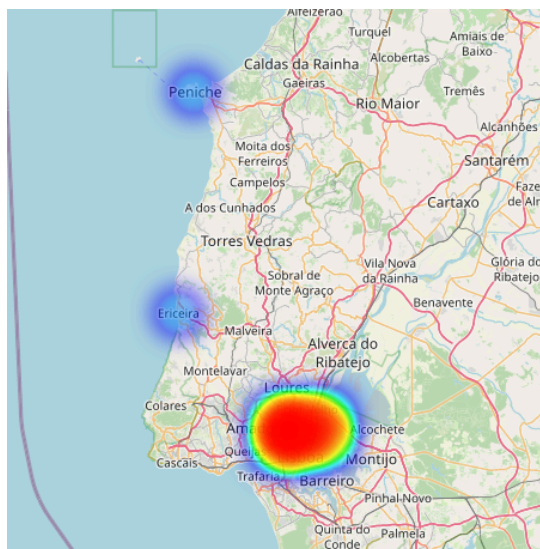


Fig.3 - Heatmap of the customers' location

- **The older customers are, the more complaints they tend to make.** It's worth highlighting the **huge increase in complaint density in the age group 80-89**, which tells us that **elderly customers make a lot more complaints** than the other groups.

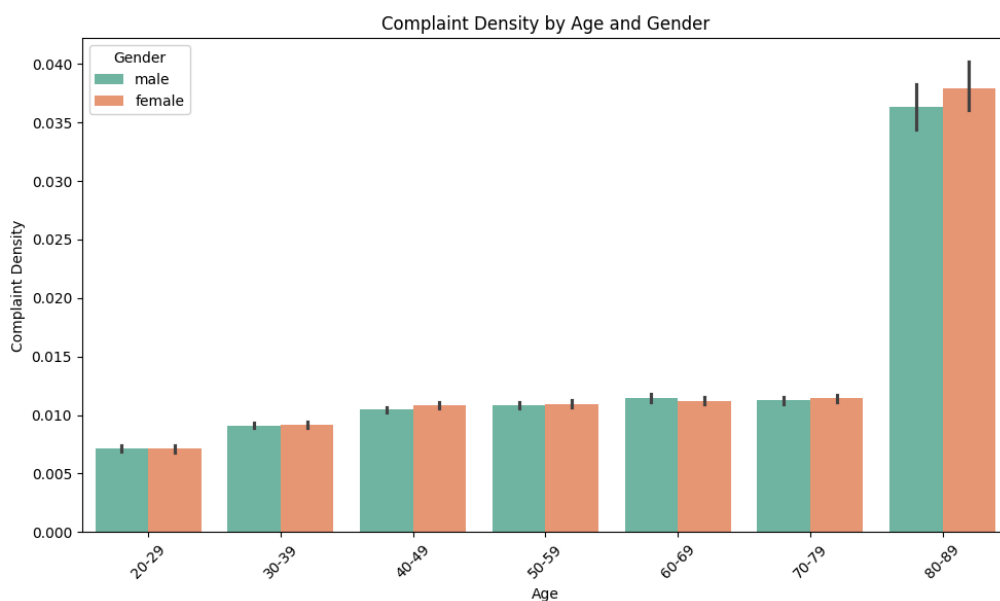


Fig.4 - Proportion of complaints divided by age and gender

- By looking at each category we can see that **the value spent is not that influenced by the card usage**, ultimately depending on the category. For instance, categories like fish, alcoholic drinks, or hygiene, are most appreciated by card users. On the other side, spending on categories like groceries, pet food, and vegetables does not depend on the loyalty card. This might be because **the loyalty card does not bring benefits to the customers of those products**, which might be a good category to implement some kind of promotion using the loyalty card.

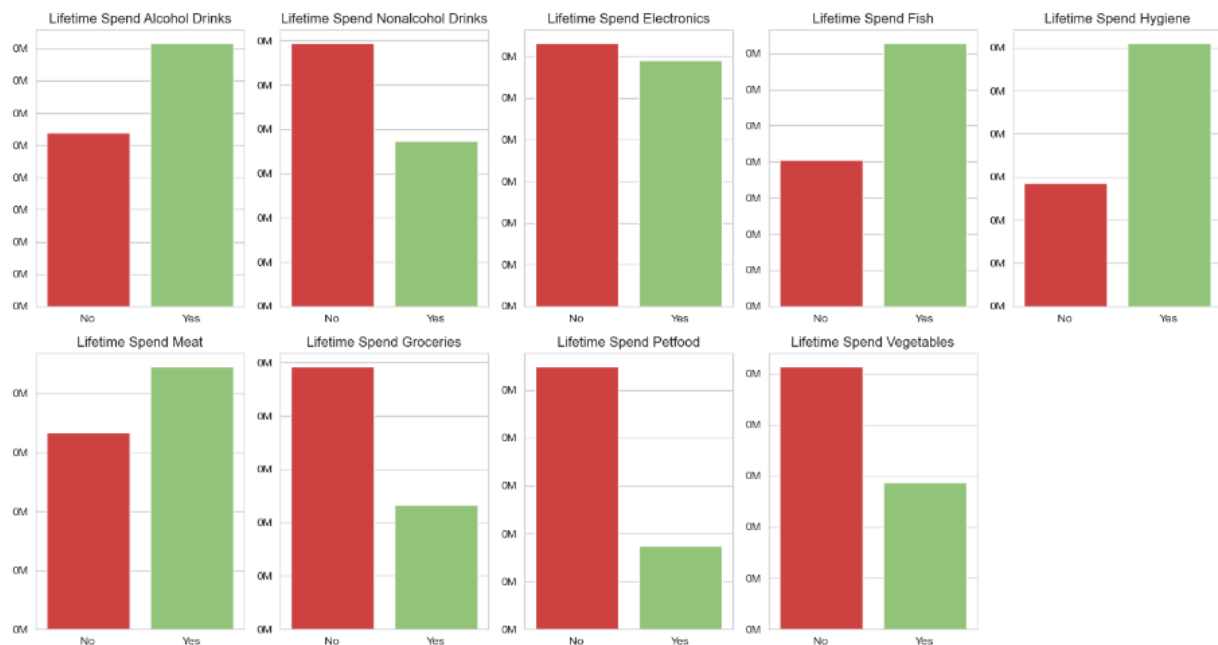


Fig.5 - Mean expenditure on each category, divided by the use of the loyalty card

**Now, let's proceed to preprocessing our data.** This is an essential step in the development of machine learning models, aiming to ensure the quality, consistency, and adequacy of data to the requirements of the algorithms.

## Data Cleaning

- customer\_basket: **No missing values**; 301 invoices **had duplicated IDs**. We dealt with it by taking dropping them, and keeping one on each group of duplicates;
- customer\_info: **No duplicates** on the customers IDs; **Relevant presence of outliers**, in several variables. For now, we won't deal with them, to not risk losing important information at such a premature stage; There was missing data in 8 variables, on which no more than 4% of the rows were missing (excluding one specific case). Before addressing them, we had to do some data transformations to our dataset;
- product\_mapping: **No missing values**; Two **products were duplicated**, and were dealt with by dropping one of them.

## Data Transformations

Here, we transformed our dataset to **make it more processable and efficient**.

The specific case of missing data on *customer\_info* was on the variable *loyalty\_card\_number*, which had **almost 50% of missing values**. By having such an amount, we could easily deduce that **it was because some customers did not have a loyalty card at all**, and so having no card number. Given this, **our approach** was to **transform it into a binary variable**, that would denote if each customer had a loyalty card. By doing so, we would lose the card number of the customers, but given that the number itself had no inherent meaning, it had no effect.

We did feature transformations on more variables:

- **Encoded** *customer\_gender*, transforming it into a **binary variable**;
- **Transformed** *customer\_birthdate* into a variable containing the **age** of customers;
- **Dropped** *customer\_name*, since it would have **no relevance** in the context of a general analysis segmentation of customers.

After these transformations, we can now address the missing values. Since we had such a **low proportion of missing values**, we thought that **imputation would be a good move**.

kids_home	1.200981
teens_home	2.340079
number_complaints	1.498934
distinct_stores_visited	3.000160
typical_hour	3.999450
lifetime_spend_vegetables	2.000871
lifetime_spend_fish	3.000160

Fig.6 - Percentage of missing data for each variable on *customer\_info*

But, before addressing them, we should consider our outliers, who were seen on the boxplots. **We tried to remove the extreme outliers**, however, after doing so, **we had a much worse imputation accuracy, giving us evidence to believe that they contain relevant information**. Given this, we decided not to do anything about them before imputation. Nevertheless, we decided to use the Robust Scaler, which is more robust to outliers than other scaling methods. It removes the median and scales the data according to the quantile range.

Before imputing, we also noticed that some variables possibly had too many possible values.


number_complaints			number_complaints	
1.0	21233		1.0	21233
0.0	13331		0.0	13331
2.0	1931		2.0	1931
3.0	68		3.0	178
4.0	45			
5.0	37			
6.0	14			
7.0	6			
8.0	6			
9.0	2			

Fig.7 - Possible values of *number\_complaints* and its frequencies, before and after the transformation

As we can see, after 3 complaints the frequency starts to decay a lot. Given this, and to facilitate predictions, we considered every customer with more than 3 complaints, to have 3. The value 3 on *number\_complaints* can now be interpreted as '3+ complaints'. We also did this transformation to *kids\_home*, *teens\_home*, and *distinct\_stores\_visited*.

For every imputation, we tried a range of 100 different Ks for each feature.

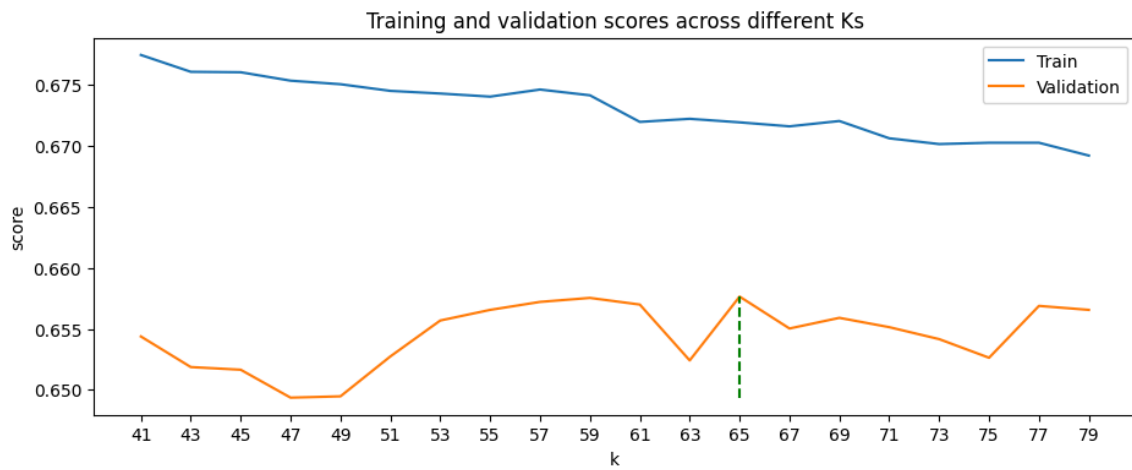


Fig.8 - Train and validation scores for each possible k, on *number\_complaints*

- The imputer of *lifetime\_spend\_fish* is **almost perfect** (suspicious). Given the fact that this variable has a huge number of outliers, we can infer that these **provide valuable information**;
- The imputer of *typical\_hour* is **horrible**;
- The rest have an **acceptable/good performance**.



Fig.9 - Performance of the imputers for each feature

Given this, we will **drop every row with missing data on *typical\_hour***, since we are **not confident** that its **imputer would yield trustable results**.

## Feature Selection

By computing the Spearman correlation matrix between our metric features, we saw that there are **high correlations** between some of these categories:

- Videogames and Electronics;
- Videogames and Meat;

- Videogames and Fish;
- Videogames and Hygiene;
- Fish and Meat (makes sense, since they serve the same purpose);
- Pet food and Groceries;
- Hygiene and Fish;
- Electronics with a lot of categories.

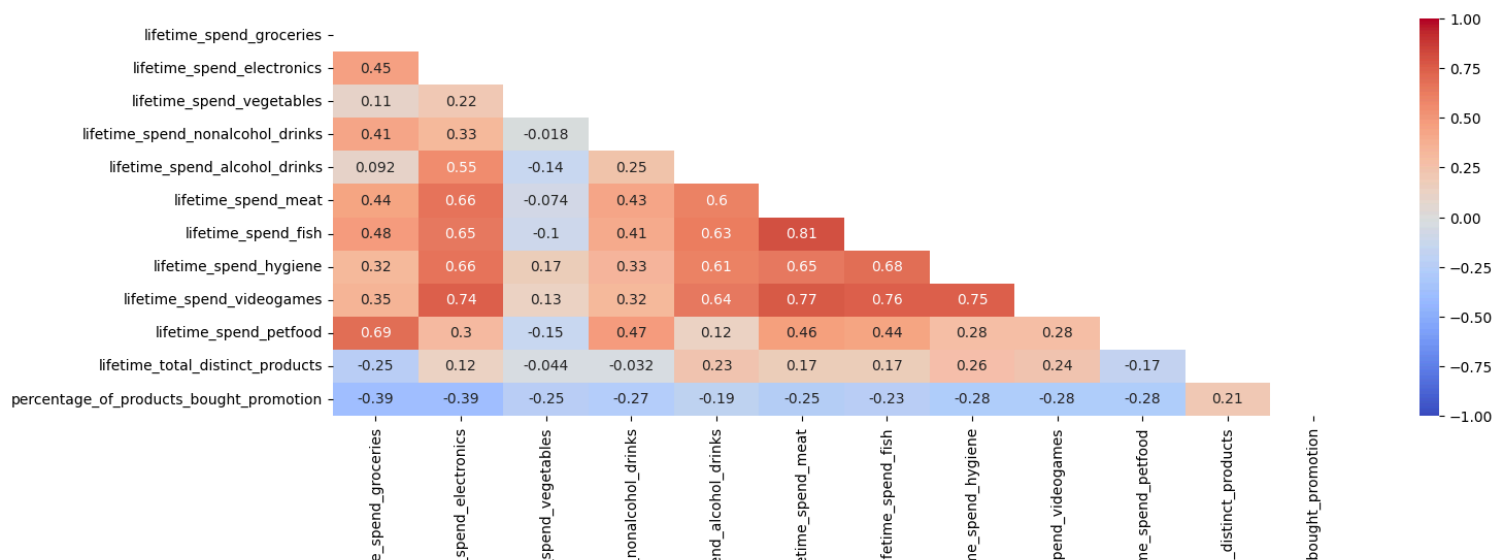


Fig. 10 - Spearman correlation of our metric features

As **Videogames** is highly correlated with a lot of categories, it isn't adding much relevant information. Given this, we removed it from the dataset.

## Customer Segmentation and Clustering

### Dimensionality Reduction

Before applying any clustering algorithm, we applied a **dimensionality reduction algorithm** to our data. We tested PCA (not very good results), but finally utilized **autoencoders**. We tested different layer sizes and activation functions, in order to get the model with the parameters that better fitted our specific data.



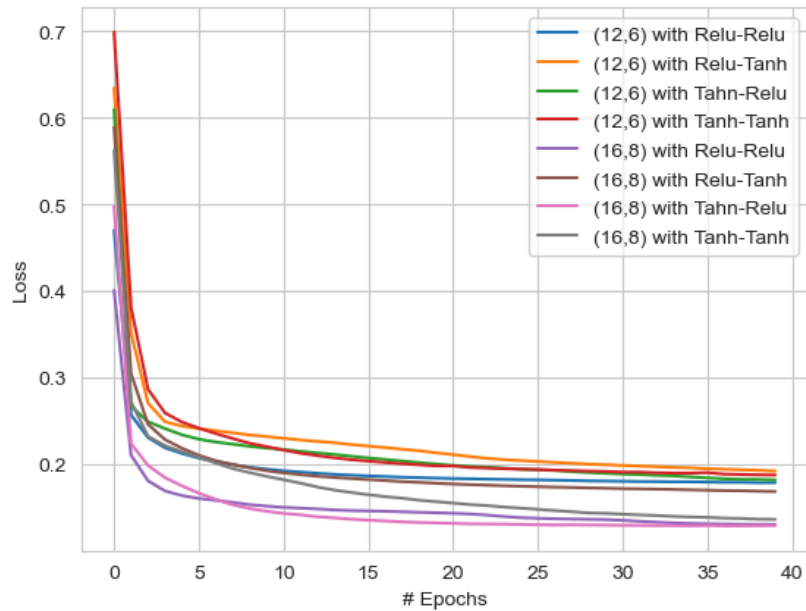


Fig. 11 - Performances of the different autoencoders.

We transformed our dataset with an autoencoder with a **hidden layer of size 16, code of size 8, Tanh** as the **encoding activation function**, and **ReLU** as the **decoder activation function** since it had the best results above. Our output decoding function was always 'linear' as our data is continuous.

Then, we **tested different clustering methods, visualized them using U-MAP and SOMs, and evaluated them with the silhouette score**. In the end, we **combined the results of the different methods to get the best solution possible**. We utilized several parameters for the U-MAP, but the main one used (that gave us the 'most real' view) was with the **50 nearest neighbors**, with a **minimum distance of 0.5**, using Euclidean distance. We applied Self-Organizing Maps with a **grid of 30x30 neurons, a sigma, and a learning rate of 1.5**. The SOM quantization error suggested that after around 800 epochs it could not be computationally worth it to continue. However, since **the training of it was so quick and inexpensive computationally**, we **trained them for thousands of epochs, to get the more robust results possible**.

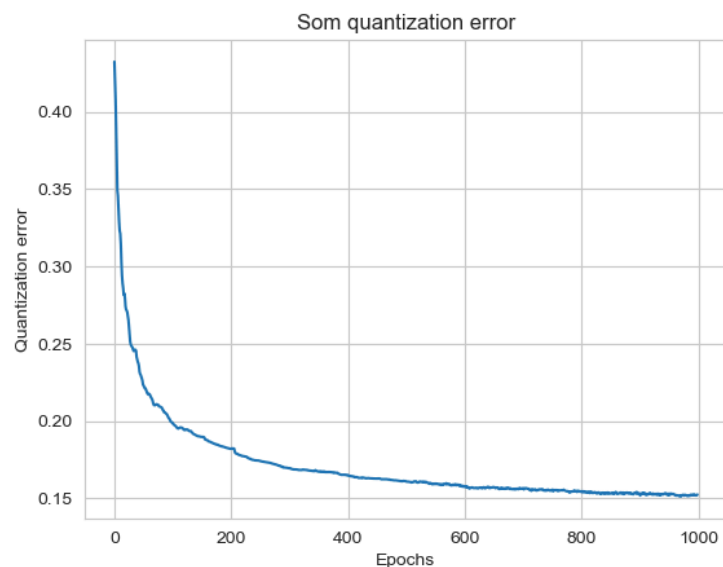


Fig. 12 - SOM quantization error plot

## Clustering Algorithms

### DBSCAN

Firstly, we started by applying the DBSCAN algorithm to our data. To define the epsilon to use we computed a K-distance line plot [2], of the 100 (our minPts) nearest neighbors. We set the **epsilon** to **0.14**, as it is a value close to the ‘elbow’ of the graphic.

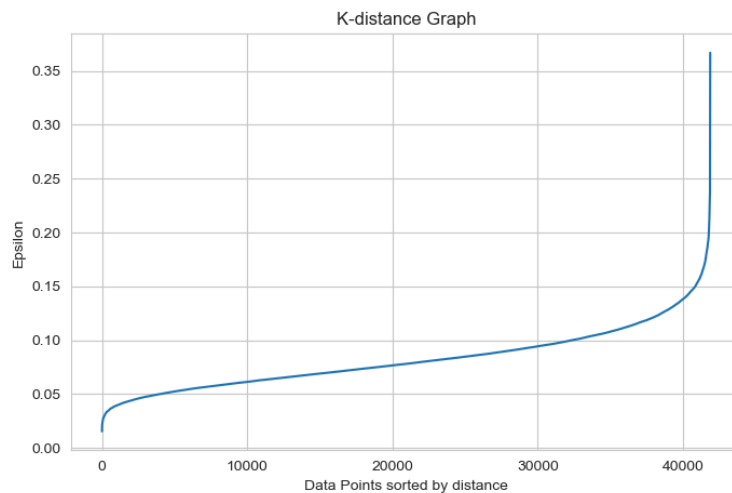


Fig. 13 - K-distance plot

**This algorithm had a terrible performance**, identifying the majority of the points as outliers.

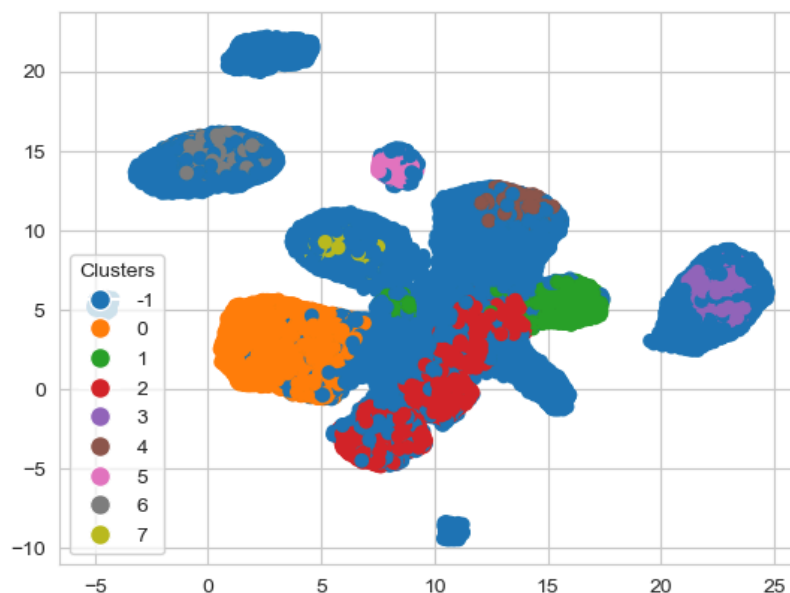


Fig. 14 - U-MAP visualization of the DBSCAN clustering on our data.

Given this, we **discarded it right away** from contributing to the optimal clustering solution.

## Mean Shift

Here, we started by estimating the optimal bandwidth for the clusters and applying it. **This method also had a terrible performance, assigning 99% of observations to the same cluster.**

Since **both density-based clustering methods had miserable performances**, we can infer that, on this problem, **they are not that useful** (clusters may be more spherical).

## K-Means

Based on the inertia plot, we decided to **create 7 clusters** (where is the red line), however, every number from 7-9 clusters could be defined as a valid solution.

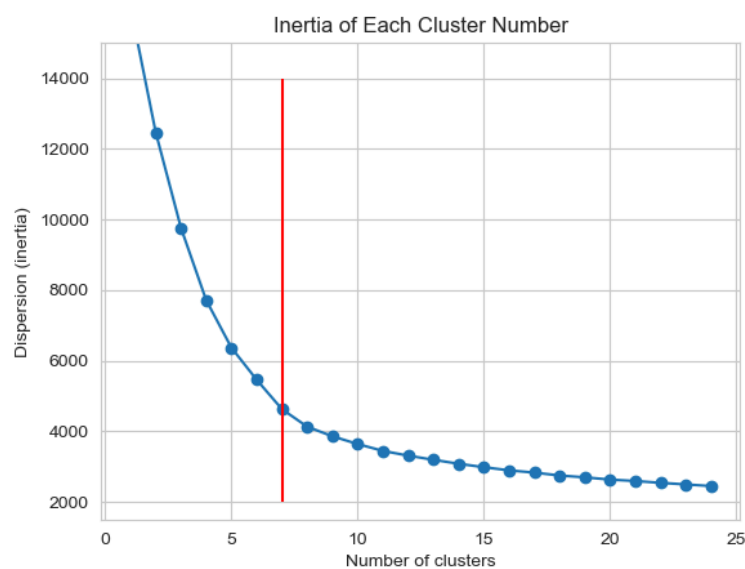


Fig. 15 - K-Means Inertia plot to find a good number of clusters.

From both visualization methods we can draw some conclusions:

- **Clusters 0, 2, 3, 5, and 6** seem to be **very well defined**;
- We can already have an idea of the form of **Cluster 4**. However, it still **needs to be refined**.
- **Cluster 1** seems to be all over the place and so **needs to be questioned or refined**;
- **Cluster 6** is very small and far away from the other clusters, what may suggest it is **consisted of outliers**.

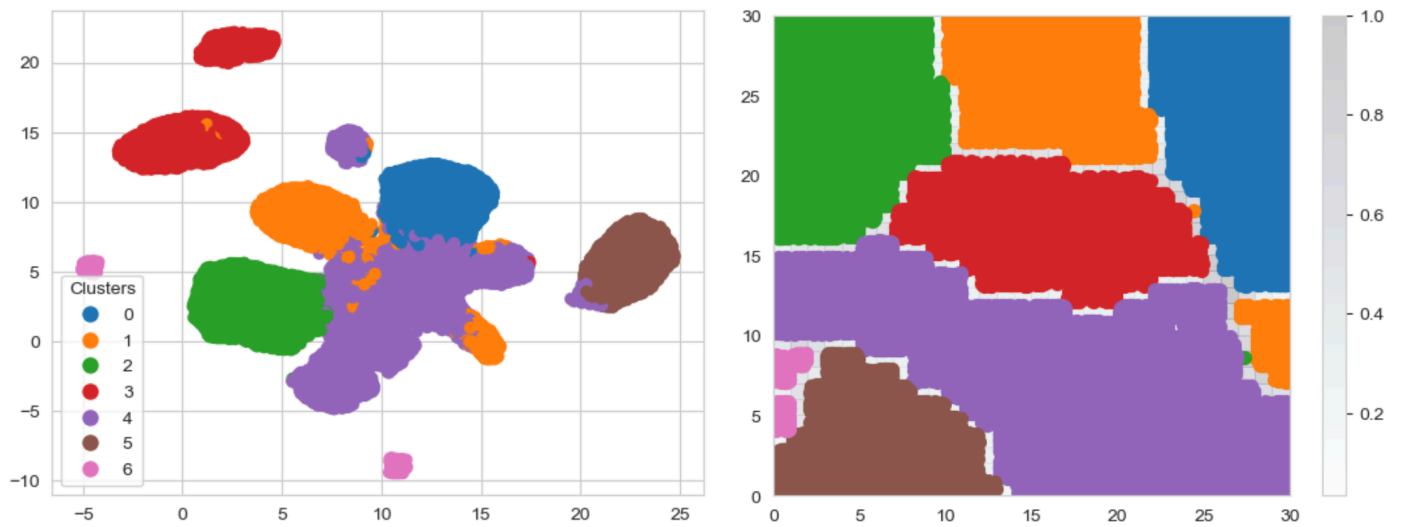


Fig. 16 - U-MAP and SOM visualizations of the K-Means clustering

## Hierarchical Clustering

In this method, **we tested the four main linkages** - single, complete, average, and ward. We started by building the dendrograms, to define the number of clusters to create. The dendrograms of complete and average linkage were all interlaced and confusing, so we discarded these two right away. The single linkage also had bad results, assigning almost every point to the same cluster.

**The remaining linkage criterion was Ward.** Based on its dendrogram, we also decided to **create 7 clusters**.

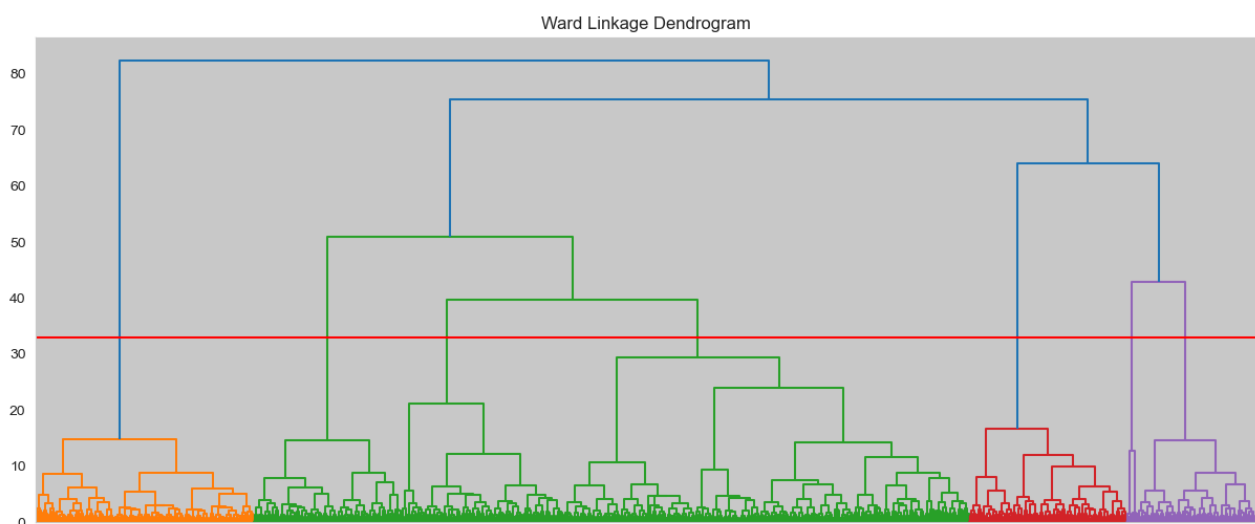


Fig. 17 - Dendrogram of Ward's hierarchical clustering

**This clustering method seemed to have a great performance, giving us a performance very similar to K-Means.** It made some well-defined clusters:

- **Clusters 1, 3, 4, 5, and 6** seem **almost perfectly defined**;
- **Clusters 0 and 2** still **need some refinement**, given their close relationship with each other;
- **Most of the clusters** are **very similar to the K-means ones**.

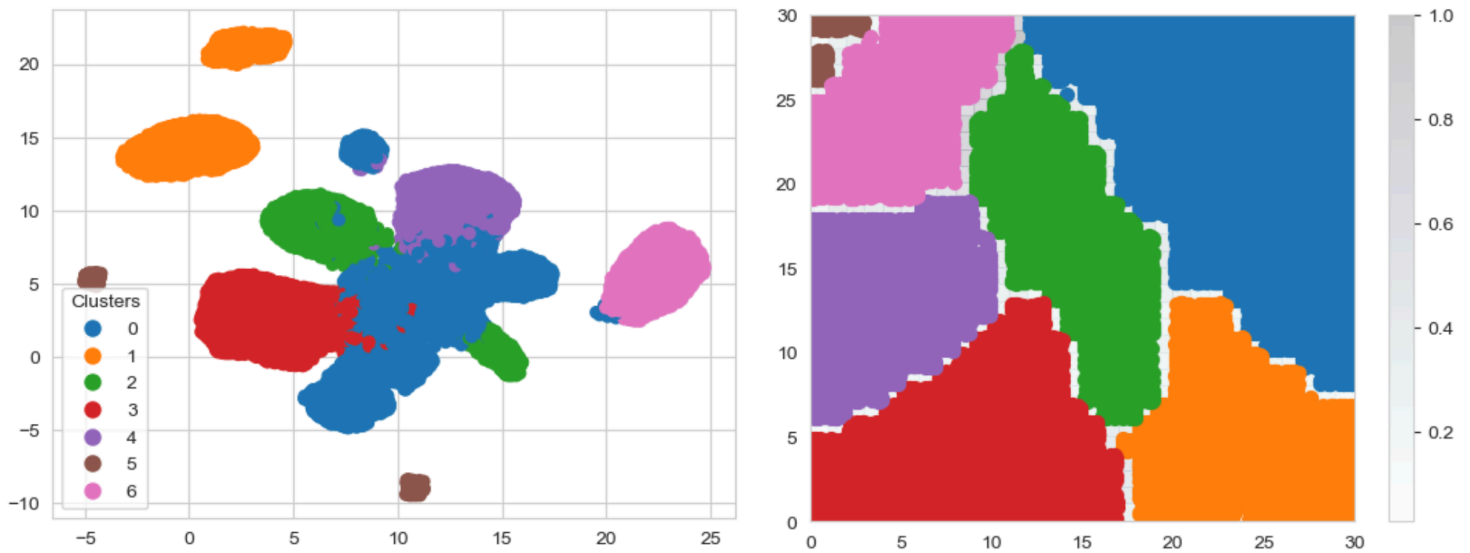


Fig. 18 - U-MAP and SOM visualizations of Ward's hierarchical clustering

## Final Clustering Solution:

To define the optimal and proposed segmentation solution, we considered the two methods that gave us good results - **K-Means and Hierarchical Clustering using Ward linkage**. So, we started by comparing the clustering solutions of both methods.

	Ward 0 Cluster	Ward 1 Cluster	Ward 2 Cluster	Ward 3 Cluster	Ward 4 Cluster	Ward 5 Cluster	Ward 6 Cluster
K-means 0 Cluster	749	0	6	4	4957	0	0
K-means 1 Cluster	455	28	5225	0	19	0	0
K-means 2 Cluster	10	0	0	7157	0	0	0
K-means 3 Cluster	6	5379	7	0	0	0	0
K-means 4 Cluster	12716	0	321	343	107	0	96
K-means 5 Cluster	0	0	0	0	0	0	3937
K-means 6 Cluster	0	0	0	0	0	364	0

Fig. 19 - Comparison of the distribution of the observations by the K-means and Ward's hierarchical clustering algorithms

We can take some insights from this confusion matrix:

- **Ward's Cluster 0** seems to be **very similar to KM's 4**;
- **Ward's Cluster 1** seems to be **very similar to KM's 3**;
- **Ward's Cluster 2** seems to be **very similar to KM's 1**;
- **Ward's Cluster 3** seems to be **very similar to KM's 2**;
- **Ward's Cluster 4** seems to be **very similar to KM's 0**;
- **Ward's Cluster 5** was **defined the same way by KM** (it makes sense since it seems to be consisted of outliers);
- **Ward's Cluster 6** seems to be **very similar to KM's 5**.

Given this information, we already have a base for every cluster, and just have to deal with the "disputed" points (for instance, the points on the intersection of Ward's 0 x KM's 0 should belong to Ward's 4 x KM's 0 or Ward's 0 x KM's 4?).

To decide on these disputes, we assigned the points to the cluster which increased the silhouette score of the solution.

After resolving the disputes we got our solution. As you can see, there are still some points that are dispersed around the plot. We fixed this by forcing some regions of the plot to be of a certain cluster:

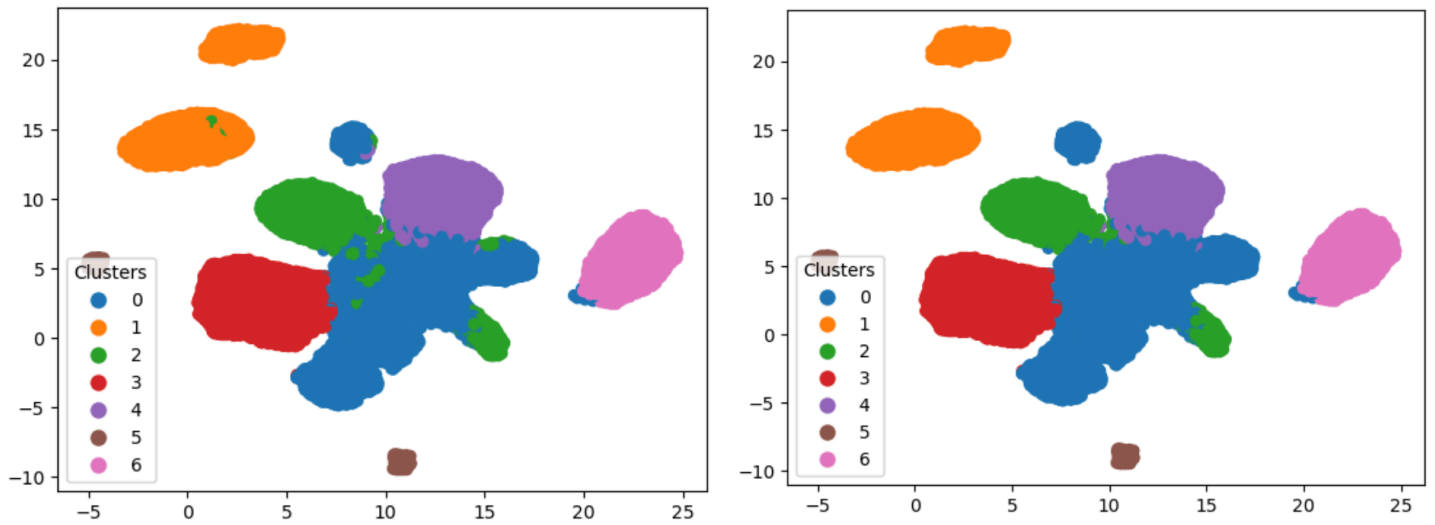


Fig. 20 - U-MAP visualization of our final clustering solution, before and after doing some corrections

We also tested other U-MAP parameters, to be sure about our clustering decisions.

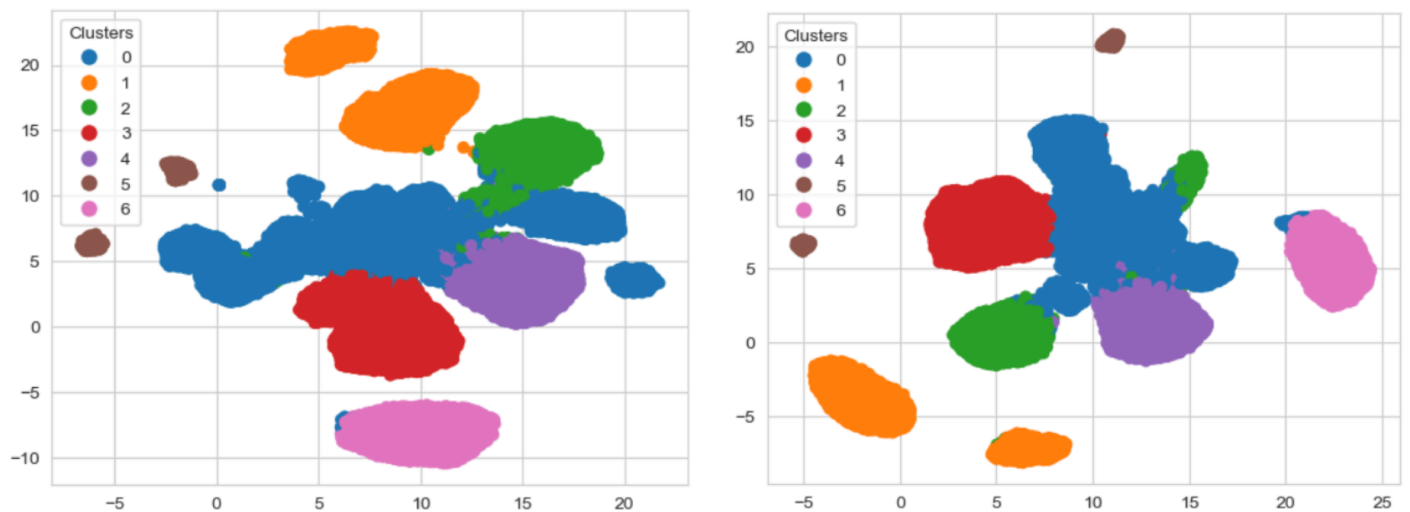


Fig. 21 - U-MAP visualization of our final clustering solution, applied with 15 and 85 nearest neighbours

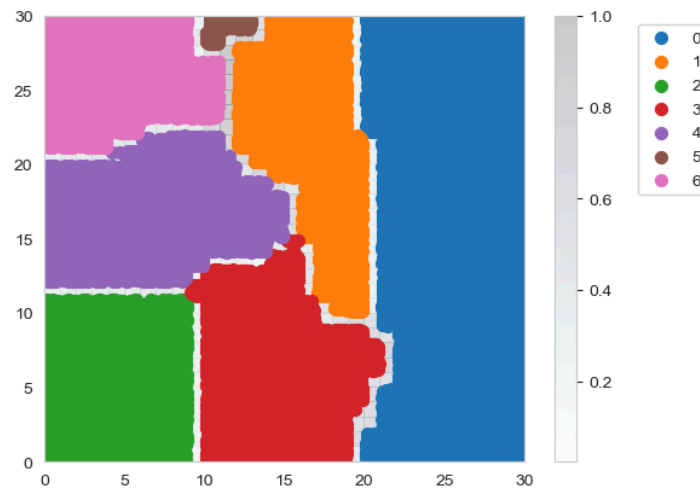


Fig. 22 - SOM visualization of our final clustering solution

Our final solution had a silhouette score of 0.36, which seems like a good value, given the nature of our problem.

## Profiling

Now that we have our final clustering solution, **we need to profile the clusters:**

- By relating the two plots below, we can see that:
  - **The segments with more customers (0 and 3) are the ones that spend the less;**
  - **Cluster 1 spends at an astonishing rate** compared to other clusters.

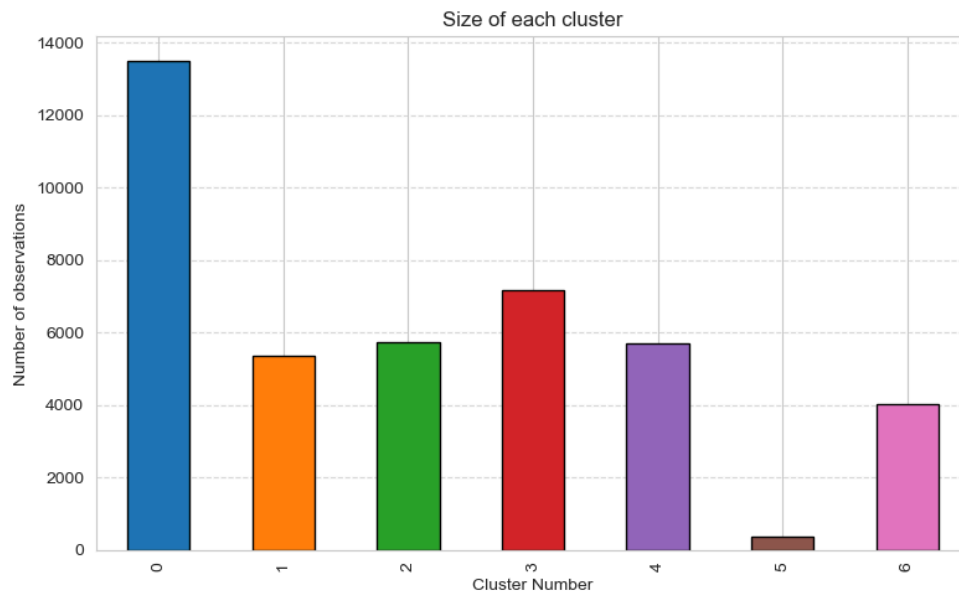


Fig. 23 - Size of each cluster of the final solution

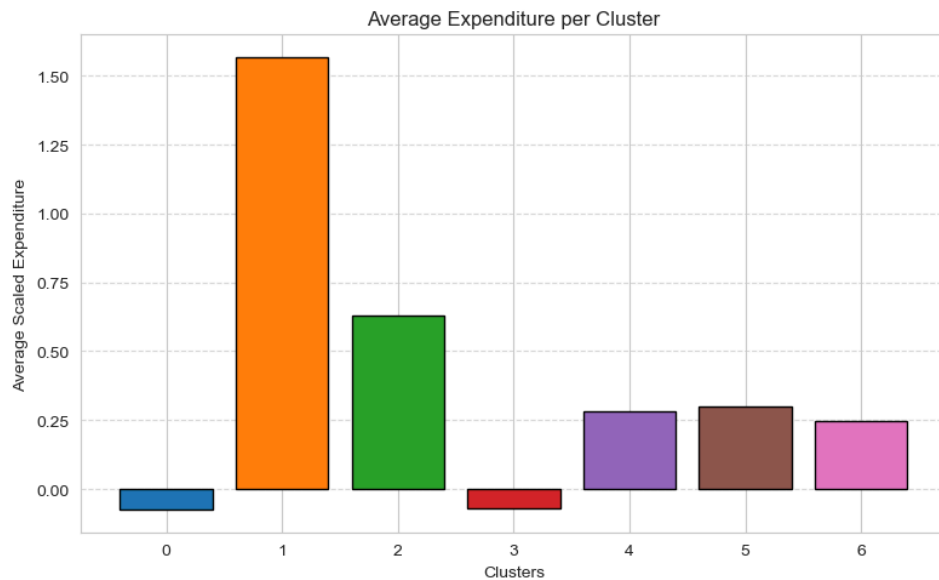


Fig. 24 - Average expenditure of each cluster of the final solution

- As seen in the preprocessing, **the overwhelming majority of our customers are located in the metropolitan area of Lisbon**. However, the **customers from Cluster 5 (the main outliers), are all located in Ericeira and Peniche**, two coastal areas.

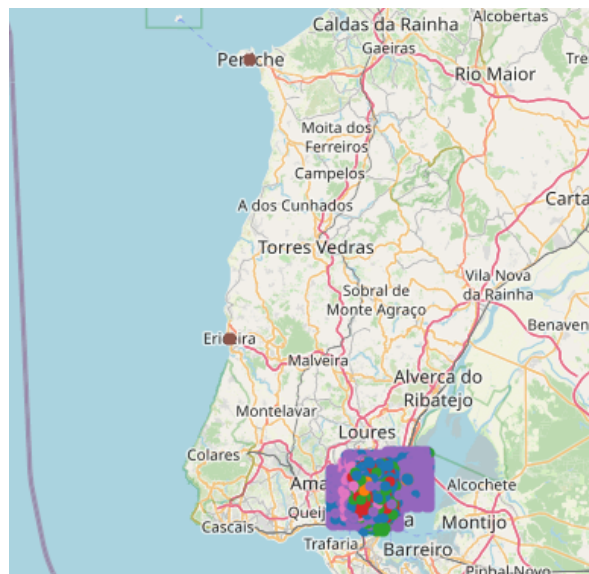


Fig. 25 - The location of our customers, marked with their respective clusters

- The radar plot shows us some **clear overall tendencies of each cluster**:
  - Cluster 0** seems to **visit a lot of distinct stores**, and also **enjoy promotions**;
  - Cluster 1** does **not enjoy promotions** at all;
  - Cluster 2** has a **dispersed behaviour**, going to a **lot of distinct stores** and **buying a lot of distinct products**;
  - Cluster 3** spends the **most on vegetables**;
  - Cluster 4** has a **lot of people at home** and complain a lot;
  - Cluster 5** has an **absurd expenditure on fish**. They also seem to **enjoy a lot of promotions**, but buy “always” the same products;



- **Cluster 6** spends **huge amounts on pet food**. However, they seem to be **detached from the store**, since they do not complain, do not have a loyalty card, and go always to the same store to buy the same products.

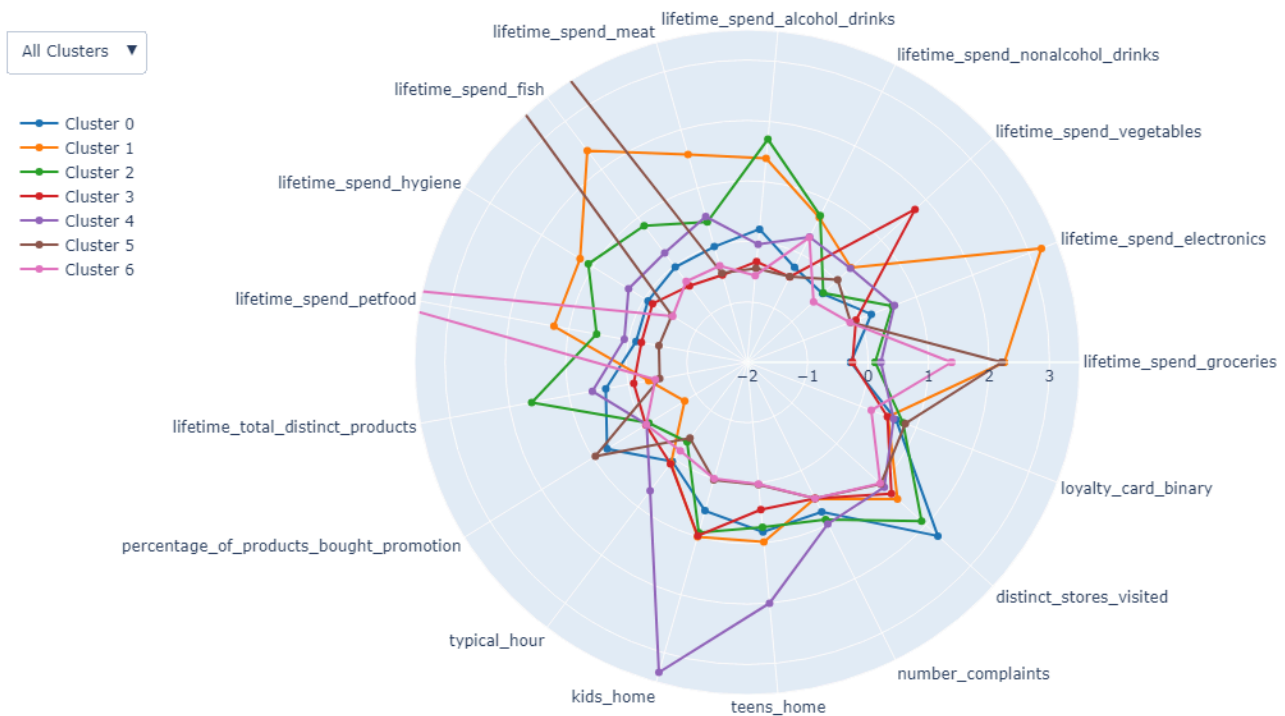


Fig. 26 - Radarplot of our clusters

The plots above used, as expected, our scaled data. However, to make a tangible interpretation of the clusters we must reverse the clustering made. **By interpreting the real values for each cluster, and with the help of the visualizations above, we defined this segmentations:**

- **Cluster 0:** Including over 30% of all customers, it is composed of low spenders, who visit a lot of different stores and enjoy promotions. Include more recent and young customers. Tend to have more teens than kids at home, so it also includes older customers. Its members are mainly **Opportunists** since they shop only when there are good opportunities (promotions);
- **Cluster 1:** Very high, consistent spenders across different types of products. Tend to not use promotions, and have a family at home - **Big, Reliable spenders**;
- **Cluster 2:** Average-high spenders (mainly in drinks and hygiene products), with very disperse behaviour, going to a lot of distinct stores and buying a lot of distinct products. Include customers who are more old to the store, tend to have a loyalty card, and complain. Shop in the morning. - **Diverse, casual**, with **nomade** behaviour, possibly including **retired** customers;
- **Cluster 3:** Very low expenditure overall, except in vegetables - composed by casual customers, that shop mainly for vegetables - may include **Vegetarians**;
- **Cluster 4:** Average spenders in a lot of distinct products. Tends to shop in the afternoon, have huge amounts of children (excluding pets), and complain regularly - composed of **Huge Families, with Karen tendencies** [2];
- **Cluster 5:** Low expenditure overall, but astronomical spending on fish. Tend to have a loyalty card, to visit always the same store, to buy the same products. Do not have kids. Their location in coastal cities famous for fishing (and their stupid expenditure on fish) suggest that they are **Fishermen**;

- **Cluster 6:** Huge expenditure on pet food, low-average in the rest. Tend to not have children (only pets), do not have a loyalty card, do not complain much, and visit always the same establishment. By this detached behaviour and preferences, we can infer that they only shop there for their pets - **Pet parents, detached.**

With these profiles, there is a good profile distinction between clusters, and could already have some ideas of promotions to apply.

## Targeted Promotions

In order to reach better and more concrete ideas for promotions, we applied association rules for each segment. We used the Apriori algorithm because, unlike ECLAT, it computes useful metrics like confidence and lift, and differentiates antecedents from consequents. Here, we searched for the relationships with a support value over 15%, and confidence over 50-60% (ideally, the confidence should be higher, but to avoid having to use always the same consequent item, we had to lower it).

From the rules created, we were able to create the following targeted promotions for each segment, using sales techniques like cross-selling, up-selling, and bundle offers [4]. We also gave other suggestions, without basing ourselves on the rules but more on their profiles:

- **Opportunists:** The customers of this segment seem to enjoy promotions, don't spend much, and visit distinct stores, so a good idea might be to offer **bundles and cross-selling** offers. For instance:
  - **For every three Oils (regular or cooking) bought, one is for free!**
  - **A bundle of Cake + Cooking Oil + Oil for 90% of the total price;**
  - **By buying 2+ units of the following items: Cake, Candy Bars, and Gums, you get an Oil bottle for 50% of its price** (one that is selling less :))
  - **Buying any Candy Bar + Cake combination gets you a 25% discount on any Oil;**
  - All the promotions above could be used in any distinct store.
    - Other ideas: Creating benefits to have the store's mobile app, which would have notifications for flash sales and promotion alerts. Subsidizing the use of ownership of a loyalty card, that would offer good promotions.
- **Big, Reliable Spenders:** The customers of this segment are consistent, loyal, and high spenders. However, do not seem to enjoy promotions. A good tactic could be to use **complementary products and upselling** techniques, in order to get them to upgrade their regular products to a "premium" version.
  - **By buying a Samsung Galaxy 10 + Bluetooth Headphones, you get 2 quality champagnes for the price of 1;**
  - **By buying any headphones (Bluetooth or AirPods), you get a Samsung Galaxy 10 for only 80% of its price;**
  - **By buying 2+ distinct units of the following: Spaghetti, Cottage Cheese, and Fromage Blanc, for each unit of premium Champagne you buy, you get a 10% discount added to the basket;**

- **Buy three tech items** from the following: **Samsung Galaxy 10, Laptop, Bluetooth headphones, iPhone 8, AirPods**, and you **get a free Champagne** (from a limited list of champagnes);
  - Other ideas: VIP member discounts, early access to new products, exclusive high-value bundles, and offers of premium quality products.
  
- **Diverse, casual customers:** These customers are good spenders, that go to distinct stores to buy diverse products. They tend to have a loyalty card, complain, and shop in the morning. Given this they could be tempted to enjoy upselling and cross-selling, to satisfy their diverse needs:
  - **From 10 to 12 o'clock**, for every **5 packages of Gums** bought, you **get an Oil bottle for 50%** of its price;
  - A **bundle** composed of **Cooking Oil, Cake, Gums, Muffins, and Oil** for **85% of the total price**;
  - **On every public holiday**, **loyalty card users** get a **25% discount on Confectionery** section items, like **Cooking Oil, Oil, Cake, Candy bars, Gums**, and more;
  - If you buy **2 bottles of Cooking Oil**, you get an **Oil bottle for half its price**;
  - All the promotions above could be used in any distinct store.
    - Other ideas: Offer loyalty card deals to move inventory, and provide complaint resolution perks (free samples, priority customer service).
  
- **Healthy, Vegetarians:** As these customers mainly shop only for vegetables, they would enjoy some **bundle offers or complementary products** to their healthy aliments:
  - **Bundle of Mashed Potatoes + Tomatoes + Asparagus** for **85% of the original price**;
  - By buying a **pack of Carrots + Asparagus**, you get a **20% off in tomatoes**;
  - By having a **basket with 3+ items** of the following: **Asparagus, Carrots, Mashed Potato, and Melons**, you get a **sample of Tomatoes for 50% off**;
  - **For every melon you buy**, you **get a 20% coupon** to use in **Asparagus**;
    - Other ideas: Healthy options, vegetarian and vegan options to replace meat and fish. If possible provide recipes, cooking tips, and health benefits of vegetables through newsletters and social media.
  
- **Huge Families, Karens:** In this segment, customers have bigger families, complain a lot, and tend to shop in the afternoon. Promotions like **bulk offers or bundle offers** (buying in bigger quantities), could facilitate their busy lives:
  - **From 17 to 19 o'clock**, each **pack of 10 units of baby food** is **25% off**;
  - By buying a **bundle of Candy Bars + Cooking oil**, you get **20% off on a pack of 4 units of baby food**;
  - A **bundle of any Cake + Cooking oil + 4 packs of Baby Food**, for the **price of the same bundle but with 2 Baby Foods**;
  - **After buying 3+ packs of 5 units of Baby Food**, every pack from then on is **15% off**;

- Other ideas: Toys and school material deals. Complaint resolution perks (free samples, priority customer service). Complementary household items.

- **Fishermen:** Given the specific and high spending on fish and the loyalty card usage in this segment, promoting complementary **high-end products and bulk packages** can be effective:
  - **By buying a package of 10+ cans of any Tuna, you get a premium quality Shrimp package with a 20% discount;**
  - **Happy Hour promotion**, twice a week - **from 9:30 to 10:30: Every high-end product from the Fish section is 10% off.** And every time **the number of products in your basket reaches a multiple of 10** (until 50), you get an **extra 10% off.**
  - **Bundle of 5+ units of at least 3 of the following products: Salmon, Tuna, Shrimp, and Seabass, for 75% of the total price;**
  - **Loyalty program rewards**, that **every of 4+ packs of canned tuna purchased, a shrimp package is offered;**
  - All the promotions should be focused mainly on the stores in Ericeira and Peniche;
    - Other ideas: Notify about promotions in advertisements in newspapers and coastal radio stations.
- **Pet parents, detached:**
  - **The dataset did not have any basket from a customer of this segment**, and thus, we could not create any specific and concrete targeted promotions. However, some ideas such as **subsidizing having a loyalty card** (since this segment tends to not have one), and **bundles of pet care products to increase their expenditure in the store** could be a good start. For instance, **creating a pet loyalty program with rewards for frequent purchases**, and offering pet care tips and product recommendations could be good moves too. This information could be passed by pet-related social media pages, pet blogs.

# Conclusion

The primary challenge of this project was to analyze and segment a complex dataset from a retail business in Portugal, characterized by diverse customer demographics and purchasing behaviors. The goal was to understand these behaviors in order to develop tailored marketing strategies that would optimize customer engagement and sales performance.

We started by exploring the datasets, in order to better understand their nature and intrinsic characteristics. Then, we had to treat and clean our data, in a way that removed all redundant information, and made it computationally workable, whilst keeping all their relevant details. We had to deal with duplicates and missing data and conduct data transformations and feature selection.

When our data was ready to work, we applied autoencoders, an efficient algorithm to handle high-dimensional data, by compressing it into lower-dimensional representations, preserving essential features while reducing noise. However, in this specific case, we thought that it may not be needed, since our dataset is not that complex and we were achieving decent results without utilizing it. Nevertheless, we thought it was a good practice to implement it, to prepare ourselves for future real-world problems.

After this, we proceeded to the segmentation part. Here, we tested every clustering algorithm that was taught and visualized them using U-MAPS and Self-Organizing Maps. The partition clustering algorithms took the lead over density-based methods, which had terrible performances on our data. The only two models that gave us good results were K-means and Hierarchical Clustering with Ward's linkage. To define the optimal clustering solution, we mixed the two methods and utilized the advantages of each, which led us to a more robust segmentation solution. Then, we created visualizations, to better profile and interpret each segment. We successfully identified and described seven distinct customer segments, each with specific characteristics and purchasing behaviors.

The final part was to create targeted promotions for each segment, given their characteristics and tendencies. Here, we used association rules, that generated some metrics on which we can rely to create interesting campaigns. We find it kind of hard to get ideas for promotions, as the Apriori algorithm generated rules on some segments on which the consequent items were always the same.

We consider that the project was successful in achieving its objectives. We ended with the impression that it was possible to reach similar results with the original dataset and that the clustering could still be more refined, but nonetheless, we are satisfied. The segmentation still allowed for the creation of focused marketing strategies that are anticipated to increase customer satisfaction and loyalty, and improve overall sales.

To conclude, this project not only provided immediate insights but also set a foundation for adaptive strategies that will be of excellent use when entering the job market.

# References

- [1] Fernando Lucas Bação, Ivo Bernardo - Machine Learning II course materials
- [2] <https://medium.com/@dilip.voleti/dbscan-algorithm-for-fraud-detection-outlier-detection-in-a-data-set-60a10ad06ea8>
- [3] [https://en.wikipedia.org/wiki/Karen\\_\(slang\)](https://en.wikipedia.org/wiki/Karen_(slang)).
- [4] <https://www.axa.co.uk/business-insurance/business-guardian-angel/upselling-cross-selling-bundles-retail/>
- [5] <https://www.fool.com/the-ascent/small-business/retail-management/retailer-promotions/>
- [6] <https://chatgpt.com/?model=gpt-4o>



**NOVA Information Management School**  
**Instituto Superior de Estatística e Gestão de Informação**  
Universidade Nova de Lisboa