

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

Customer Segmentation: An application to dental medicine patients

Tiago Nobre Caldeira Gonçalves

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

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

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

CUSTOMER SEGMENTATION: AN APPLICATION TO DENTAL MEDICINE PATIENTS

by

Tiago Gonçalves

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: *Prof Doutor* Flávio Luís Portas Pinheiro

External Supervisors: Ana Sofia Ferreira Cordeiro da Câmara Pestana | Filipa Sá da Costa

DEDICATION

Aos meus pais, por me terem dado esta oportunidade, e por serem referências para mim.

À minha família e aos meus amigos, por todo o apoio que me deram, e acima de tudo por me darem a alegria e a motivação para superar este desafio.

ACKNOWLEDGEMENTS

Professor Flávio, thank you for your time and effort to provide me valuable and constructive feedback on this report. I would also like to thank you for being very approachable and easy to communicate with from the day one of the Master's, as well as for your interactive approach on teaching, which made your lectures very interesting despite the pandemic context that we all lived.

Sofia and Filipa, thank you for your guidance during the internship and the development of the project. It was an enriching first professional experience, and I've learned a lot with both of you. Sofia, thank you for having believed in me for the development of this project, and for helping me to grow professionally by following your example of hard work and organization. Filipa, thank you for being so supportive, for giving me very useful advice that helped me in developing the project and writing this report, and for our endless conversations about all imaginable topics.

To my friends, thank you for helping me achieving this goal, from the thesis sessions to help me being more productive, to our dinners and gatherings to hang out and have fun, which kept me motivated to accomplish this.

Finally, to my family, thank you for your endless support and for always being there for me, making sure that, albeit slowly, there was always some progress along the weeks while developing my thesis.

ABSTRACT

Customer segmentation allows to divide a company's customers into multiple market segments, enabling the development of customized marketing actions based on each segment's characteristics. This work describes the application of a customer segmentation approach to the patients of a Portuguese dental company. The approach taken to select the feature subset for the final model was mostly based on the LRFM (length, recency, frequency, and monetary) model, and the monetary variable was split into multiple variables according to the treatment category where the amount was spent. K-Means and Self-organizing maps were used to cluster the company's patients using these variables, and the results returned by both algorithms are compared. The final solution was obtained with K-Means, and 7 clusters of patients were identified. An overview of the 7 clusters is provided, and possible marketing actions are suggested based on their main characteristics. The results allowed the company to understand how its turnover was distributed across segments, and to develop an initiative to contact the patients belonging to a segment where most of them did not have an appointment in one of the company's clinics for a long time.

KEYWORDS

Clustering; Customer Segmentation; RFM; K-Means; Self-organizing maps

INDEX

1. Introduction.....	1
1.1. Company overview.....	1
1.2. The team and Activities.....	2
1.3. Internship Goals.....	2
2. Theoretical Framework	3
2.1. Customer Relationship Management and Customer Segmentation	3
2.2. RFM Analysis.....	4
2.3. Cluster Analysis and Algorithms	5
2.3.1. The K-Means Algorithm.....	5
2.3.2. Self-organizing maps	7
2.4. Clustering Evaluation Metrics.....	8
2.4.1. Within-cluster sum of squares and Total sum of squares.....	9
2.4.2. Silhouette Coefficient.....	9
2.4.3. Davies-Bouldin Index.....	11
2.5. Related Work – Practical Applications	11
3. Customer segmentation Project.....	15
3.1. Timeline	15
3.2. Motivation	17
3.3. Methodology	17
3.4. Data Preparation	18
3.5. Data Exploration	21
3.5.1. Data Quality.....	21
3.5.2. Data Distribution	21
3.5.3. Correlation between variables	23
3.6. Clustering – Models and Analysis.....	24
3.6.1. Baseline model	26
3.6.2. <i>RFML</i> model.....	27
3.6.3. Final model	29
3.7. Results and Discussion.....	34
4. Conclusions.....	38
4.1. Limitations	39
4.2. Future work	39
5. Bibliography.....	40

LIST OF FIGURES

Figure 2.1 – An example of the K-means algorithm.....	6
Figure 2.2 – Illustration of a Self-organizing map	8
Figure 2.3 – ‘Elbow’ example ($n = 4$)	9
Figure 2.4 - Silhouette plot example	11
Figure 3.1 – Project Roadmap: Planned vs. Real duration	16
Figure 3.2 - CRISP-DM Methodology.....	17
Figure 3.3 - Boxplots of the numerical variables.....	23
Figure 3.4 - TSS and Average Silhouette plots for the Baseline model.....	26
Figure 3.5 - TSS and Average Silhouette plots for the RFML model	28
Figure 3.6 – TSS and Average Silhouette plots for the final K-Means model	30
Figure 3.7 - TSS and Average Silhouette plots for the best SOM configurations	31
Figure 3.8 – Silhouette Plot of the final K-Means model	33
Figure 3.9 – Silhouette Plot of the final SOM model	33
Figure 3.10 - Size and Turnover Percentages of each Cluster.....	36

LIST OF TABLES

Table 2.1 – Related work summary.....	14
Table 3.1 – Connection between the Timeline Phases and the CRISP-DM Stages	18
Table 3.2 – Description of the new Treatment Categories	20
Table 3.3 – Final dataset summary	21
Table 3.4 - Percentage of Zeros for the <i>M1-M6</i> variables	22
Table 3.5 - Correlation between the numerical variables.....	24
Table 3.6 – Overview of the clusters obtained from the Baseline model	27
Table 3.7 – Clusters obtained from the RFML model	28
Table 3.8 – Best SOM configurations and their results.....	31
Table 3.9 – Summary of the clusters obtained with the final K-Means model	32
Table 3.10 – Summary of the clusters obtained with the final SOM model.....	32
Table 3.11 – Clustering evaluation metrics for the final models	34
Table 3.12 – Overview of the final model’s clusters	34

1. INTRODUCTION

Pursuing a long-term relationship with customers rather than having a strategy oriented towards transactions is, according to many relationship marketing scholars, more profitable for companies (Jayachandran et al., 1992). The constant progress in information technology has provided new ways for companies to manage their relationships with their customers – namely Customer Relationship Management (CRM) technology. One of the main goals of using CRM technology consists in segmenting the company's customers, to differentiate between profitable and unprofitable ones.

Dividing a company's customers into market segments allows the development of customized marketing strategies and communication initiatives, besides providing a descriptive analysis of its customers' base – how they are grouped according to their history with the company. Within these segments, consumers' needs are homogeneous, whereas they are heterogeneous across segments (Capon, 1982). These principles can also be applied to companies in the field of dental medicine, as dental patients have different needs, expectations and motivations based on their characteristics. Understanding how they can be grouped into different segments enables the optimization of the communication between the respective company and them, aiming to increase their satisfaction levels, retention, and maximizing their value potential. This internship report will cover the development of a customer segmentation project in a Portuguese dental medicine company, where these principles were applied to the case of dental patients. It was developed during the first half of a 12-month internship in the company.

The project had the main goal of understanding how the company's patients were distributed across different segments and having an overview on their main characteristics. The desired outcomes consisted in being able to customize the communication with the patients from each segment, aiming to increase their retention and satisfaction levels. In this report, the different phases of the development of the project will be described, from the initial analysis and preparation of the data to the results that were achieved using clustering algorithms. The two main algorithms that were used and tested were K-Means and Self-Organizing Maps (SOM), and the selection of features for clustering combined general ideas from existing variations of the Recency-Frequency-Monetary (RFM) model. The final goal consisted in achieving a quality solution, both in terms of scientific criteria and business applicability.

1.1. COMPANY OVERVIEW

The 12-month internship took place in a national dental healthcare company. The company only operates in Portugal and opened in December 2018 with only 1 dental clinic. Since then, it has been in continuous expansion, being currently responsible for a network of more than 10 dental clinics across Portugal, which are mainly located in urban areas.

The company provides dental treatments across several specialties, from simple check-up procedures to more complex ones in the fields of orthodontics or implantology. Its main purpose is to provide quality dental healthcare while adapting to the necessities and uniqueness of each patient, and it is open for everyone independently of the person being covered or not by health insurance. Being in an expansion phase, the main business strategy consists in increasing the number of patients that resort to the company's dental services, while simultaneously retaining its already existing patients.

1.2. THE TEAM AND ACTIVITIES

During the 12-month internship, I was part of the Marketing and Business Analytics team, which is responsible for all the company's marketing initiatives, and the reporting and analysis of different business key performance indicators. During approximately the first half of the internship, I was exclusively dedicated to the development of a customer segmentation project, which is the focus of the current report. In the second part my focus turned to other business analytics projects.

I was responsible for the planning and development of the customer segmentation project from the beginning to the end, in accordance with the business needs and the inputs given by the different stakeholders. It involved different types of tasks, from pre-processing the data to clustering it, which will be fully covered across this report. My main goal while developing the project consisted in achieving a quality solution both in quantitative and qualitative terms – scientific criteria and business applicability, respectively.

1.3. INTERNSHIP GOALS

The Marketing team in the company essentially had two main goals: attracting new patients and retaining the already existing ones. The development of the project mainly addressed the second goal – understanding who the already existing patients were, as well as the main characteristics of each resulting segment, would facilitate the development of customized marketing initiatives, which would then be more suited to their characteristics and needs. As previously mentioned, the main desired outcomes consisted in increasing their satisfaction levels and retention, as well as maximizing their value potential.

More specifically, understanding the differences between loyal and non-loyal patients was one of the main questions that the company wanted the project to address. For the loyal ones, the objective consisted in trying to understand what characterized them and what drove them to become loyal to the company, so their potential value could be maximized. For the non-loyal ones, it was important to find out what could be done to improve their retention levels and turn them into loyal patients.

2. THEORETICAL FRAMEWORK

In this section a literature review is presented. It provides an overview of the theoretical background that supports the developed project, as well as other research projects related to customer segmentation and their results.

2.1. CUSTOMER RELATIONSHIP MANAGEMENT AND CUSTOMER SEGMENTATION

Customer Relationship Management (CRM) is a practice that many companies and enterprises adopt which helps them managing and improving the relationships with their customers in a structured and organized way, using data and information collected from their transactions and interactions with the respective company (Xu et al., 2002). In a broad way, it can be defined as the set of rules, guidelines, and procedures that each company follows to communicate and interact with its customers (and potential ones). Overall, its main goals consist in enabling the company to build a long-term relationship with its customers, thus improving their loyalty and increasing business profits (Y.-S. Chen et al., 2012).

CRM focuses in adapting the company's behavior towards individual customers based on the interactions between both (transactions and communication) and on what the company knows about each individual, enhancing their relationships. The main premise consists in the following idea: already existing customers are more profitable than new ones, i.e., it is cheaper to sell an additional product to an individual that is already a customer than it is to attract a new one (Peppard, 2000). Therefore, improving customers' loyalty and maximizing their lifetime value to the company are the central objectives of CRM, which can be achieved by aligning the company's product offers and communication with each individual customer's expectations and needs.

From a marketing perspective, one of the main subsets of CRM is customer segmentation – assigning customers into different segments based on shared characteristics or needs (Sharma et al., 2021). Different segments end up having different sets of preferences and expectations regarding products or services – hence customer segmentation allows marketers to develop targeted campaigns and customize their communication in accordance with the characteristics and needs of these market segments to which several customers belong to, aiming to maximize the value of each individual customer to the company.

Hutt & Speh (2010) have listed the main benefits that arise from dividing a company's customer into segments. Firstly, just by segmenting the customers the marketer is pushed to become more aware of the unique needs and characteristics of each segment. Then, knowing these needs and characteristics, there are several benefits which may arise:

- Development of profitable pricing strategies;
- Selection of appropriate distribution channels;
- Targeted advertising messages;
- Development of efficient and effective business marketing strategies;
- Better allocation of marketing resources.

2.2. RFM ANALYSIS

Recency, Frequency and Monetary (RFM) Analysis is a model that was first proposed by Hughes (2005) to analyze customer behavior in transaction databases. Given its simplicity, it is one of the most frequently used segmentation techniques (Wei et al., 2010). Its three main features are usually defined as follows (Chang & Tsai, 2011):

- **R** – time interval between the customer's latest consumption and the present;
- **F** – number of consumptions/transactions within a defined period;
- **M** – amount of money spent within a defined period.

One of the main RFM scoring methods is the customer quintile scoring (Miglautsch, 2000). Customers are split into quintiles for R, F, and M, where the best is labeled with the score of 5, while the worst is labeled with 1. For each of the three variables, customers are sorted from the best to the worst – for recency, customers are sorted from the ones with the lowest number of days since the last transaction (best), to the ones with the highest number of days (worst); for frequency, they are sorted by the number of purchases, where the higher the number of purchases the better; for monetary, the same rule applies – the higher the amount of money spent, the better. Then, each customer receives a label according to the quintile in which he or she is inserted for each variable, having 3 labels in total. For instance, a customer labeled with 5,5,5 would be inserted in the “best” quintiles for R, F, and M, whereas a customer labeled with 1,1,1 would be amongst the “worst” customers. A customer labeled with 5,5,1 would be inserted on the best quintiles for both Recency and Frequency, but on the worst quintile for Monetary.

Besides its easy interpretability, one of the advantages of quintile scoring is that it generates equal sized segments – 125 in total. Nevertheless, it also faces some challenges, namely when more than 20% of the customers have the same value for one of the variables, which commonly occurs in the case of customers with only 1 recorded transaction. If more than 20% of the customers only have 1 purchase, there will be customers assigned to label 2 for Frequency, although their Frequency value is the same as the one from customers labeled with 1 (Miglautsch, 2000). Moreover, it may end up grouping together customers who have different purchasing behavior, while simultaneously separating arbitrarily customers whose purchasing behavior is similar.

Over time researchers have tried to develop new RFM models by considering additional variables, trying to improve its performance compared to the traditional ones. Chang & Tsai (2011) proposed a new framework called GRFM (Group RFM), which aims to tackle the problem of traditional RFM models not considering the characteristics of the purchased products. GRFM takes these characteristics into account so that the calculated RFM values for the customers are highly related to the products that they purchase, correctly reflecting their consumption behavior. RFM per product (RFM/P), proposed by Heldt et al. (2021), also takes the product perspective into account. First, it estimates the customer values per product, and then it aggregates them, obtaining the overall customer value.

Length RFM (LRFM) is another extended version of the RFM model, proposed by Wei et al. (2012), which takes into account the duration of the relationship between the customers and the company. This allows to differentiate between long-term customers and those that have had their first contact with the company in a recent past.

2.3. CLUSTER ANALYSIS AND ALGORITHMS

Cluster analysis consists in discovering the natural groupings in a set of unlabeled data patterns (Jain, 2010), based on the similarities and differences among them. Its main purpose is to group the data patterns into groups that are homogeneous within, i.e., patterns belonging to the same group are relatively similar to each other. In general, clustering is considered a form of unsupervised machine learning. Unlike supervised learning, where the data is labeled, on unsupervised learning the data is unlabeled (Jain, 2010), i.e., there is no ground-truth to compare the obtained results to.

Most of the existing clustering algorithms can be categorized as hierarchical or partitional. Hierarchical clustering algorithms output a clustering structure that consists of nested partitions, which can be graphically represented as a dendrogram, whereas partitional clustering algorithms create a single partition of the data, not imposing a hierarchical structure. Hierarchical algorithms can subsequently be divided into agglomerative and divisive – agglomerative ones begin with each sample as an individual cluster and build the nested sequence by merging clusters consecutively, whilst divisive ones begin with all samples assigned to the same cluster and continuously split clusters across iterations (Frigui & Krishnapuram, 1997). Partitional algorithms can also be divided into two different categories: hard and fuzzy. While on hard partitional algorithms each data point can only be assigned to one cluster, on fuzzy algorithms each data point has a quantifiable degree of membership (between 0 and 1) to each cluster (Frigui & Krishnapuram, 1997).

2.3.1. The K-Means Algorithm

The K-Means algorithm, proposed by MacQueen (1967), is an iterative partitional clustering algorithm that requires the pre-definition of K , the number of clusters. After defining this value, K seeds are initialized by randomly selecting K points from the dataset – these will be the initial cluster centers, called centroids. The following part of the algorithm is summarized in the iterative steps below:

1. Calculate the distance between every data point and each of the K centroids that were defined in the previous step and assign each point to the closest centroid.
2. Recalculate the centroids by computing the mean of the points that have been assigned to the same centroid.
3. Check if any of the stopping criteria is met – if at least one of them is met, the process terminates, and each of the final K clusters comprises the data points that are assigned to the same centroid. In case none of the criteria are met, the process goes back to step 1. The stopping criteria can be:
 - The algorithm reached a pre-defined maximum number of iterations;
 - The distance between each of the new centroids and the centroid it replaced is lower than a pre-defined constant, which means that the centroids have converged.

Figure 2.1 provides a simple example that illustrates the application of the K-Means algorithm on a 2-dimensional dataset with three clusters, divided in 5 parts:

- a) The 2-dimensional data, where three natural clusters are visible;

- b) The seeds are initialized by randomly selecting three data points, and each data point is assigned to the closest seed, forming the initial clusters;
- c) Intermediate iteration where the centroids are recalculated, and the cluster to which each point is assigned is updated;
- d) Same as c), the centroids are recalculated centroids and the cluster labels updated;
- e) Final clustering results, since the centroids converged (there were no changes in the clustering labels).

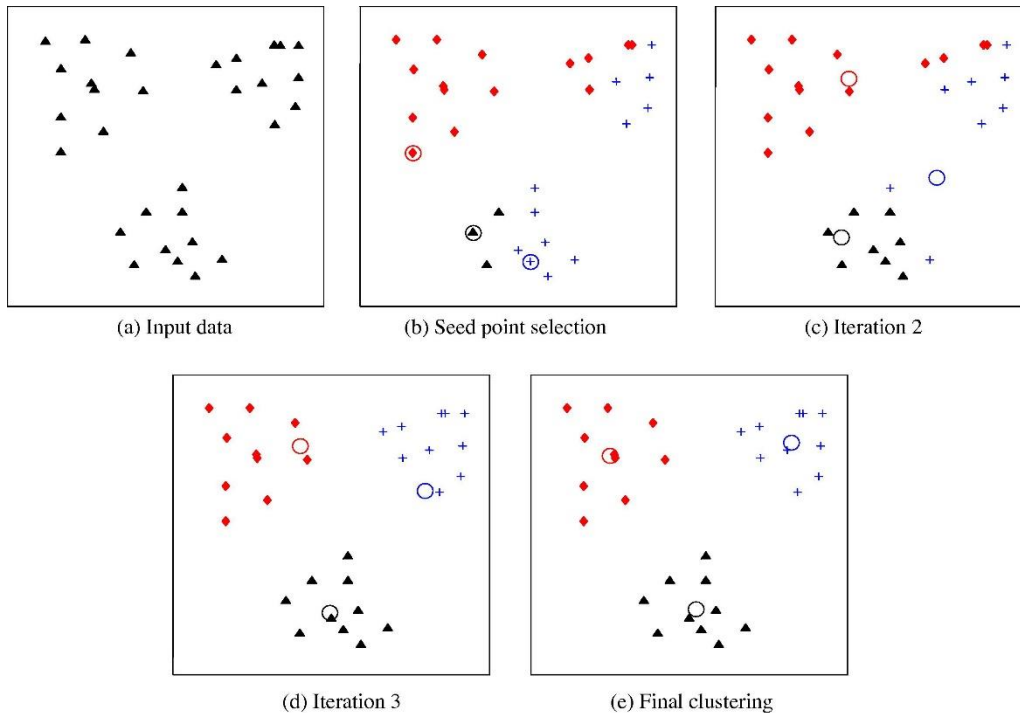


Figure 2.1 – An example of the K-means algorithm¹

After completing this iterative process, each data point is assigned to one of the K clusters. Due to its simplicity and low computational cost, K-Means is one of the most used clustering algorithms, even though it was firstly proposed more than 50 years ago (Jain, 2010), which shows the difficulty in designing a universal clustering algorithm. Nevertheless, using K-Means also comes with some disadvantages and drawbacks, one of which is the necessity of pre-defining the number of clusters, specially when there is no prior knowledge about it. Since there is no perfect mathematical criterion to define this value, there are some heuristics that can be applied to overcome this drawback, which help in assessing an adequate value for K . This value is also dependent on subjective criteria such as operational considerations, since the final number of clusters should be actionable, depending on the purpose of the application of the clustering algorithm.

Moreover, the goal of K-means is to find a partition that minimizes the squared error between the empirical mean of a cluster and the respective data points (Jain, 2010). Being a greedy algorithm, it can only converge to local optima, which results in different initializations leading to different final results. One way to overcome this problem is to run the algorithm multiple times with different initializations, selecting the result that achieved the smallest squared error.

¹ Source: (Jain, 2010)

Finally, since K-means is a distance-based algorithm, it is sensitive to noisy data and outliers - extreme values tend to have impact on the results. Furthermore, it mostly finds spherical or ball-shaped clusters in the data, having difficulties in achieving quality results where clusters have different shapes or structures.

2.3.2. Self-organizing maps

Kohonen's Self-organizing maps, henceforth referred to as SOM's, are a specific type of Artificial Neural Networks, proposed by Tuevo Kohonen in 1982. SOM's map the data patterns onto a grid of units, which is usually 1, 2 or 3 dimensional, while maintaining topological relations between these patterns (Baao et al., 2005). These data patterns are usually referred to as the input space, while the grid of units is called the output space. Although the algorithm consists of the training and testing phases, both use the same dataset (Olszewski, 2021).

Before formalizing the algorithm, some definitions are necessary:

- $\mathbf{X}$ – set of n data patterns $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$;
- $\mathbf{W}$ – $p \times q$ grid of units w_{ij} , $1 < i < p, 0 < j < q$;
- α – learning rate, which assumes values between 0 and 1;
- r – radius of the neighborhood function, $h(w_{ij}, w_{mn}, r)$.

The algorithm requires the initialization of the learning rate, the radius of the neighborhood function, and the network size (p and q), although there are no theoretical ways to obtain optimal values for these parameters. The training stage proceeds as follows (Baao et al., 2005):

1. For each data point $\mathbf{x}_n \in \mathbf{X}$, calculate its distance to each unit $\mathbf{w}_{ij} \in \mathbf{W}$, and select the unit that is closer to the data point, i.e., the unit that minimizes the distance to the point. This unit is called the winner – w_{winner} , but can also be called best matching unit (BMU). Then, update each unit $\mathbf{w}_{ij} \in \mathbf{W}$: $\mathbf{w}_{ij} = \mathbf{w}_{ij} + \alpha h(w_{ij}, w_{mn}, r) \|\mathbf{x}_k - \mathbf{w}_{ij}\|$.
2. Decrease the value of α and r .
3. If α has reached 0, the algorithm training stage ends. Otherwise, it goes back to step 1.

Since the radius of the neighborhood function decreases during training, each unit will be less affected by its neighbors as the number of completed iterations increases. Moreover, the neighborhood function usually assumes a lower value when the distance to the winning unit is higher, i.e., the more distant a unit is from the winning unit in the output space, the lower the value of the neighborhood function (Henriques et al., 2012). The value of alpha must also converge to 0 during the training process in order to ensure its stability and convergence, which is usually done linearly although other functions can be applied (Henriques et al., 2012).

One of the advantages of SOM's is that they can be used as a dimensionality reduction technique, and consequently as a data visualization tool, when the output space is one, two or three-dimensional (Olszewski, 2021). The location of the data points in the lower-dimensional output space displays the resemblance between them in the input space, and therefore the algorithm allows the user to visualize

the relationships between the different data points. Figure 2.2 illustrates a SOM where the data patterns in a multi-dimensional input space are mapped into a two-dimensional output space.

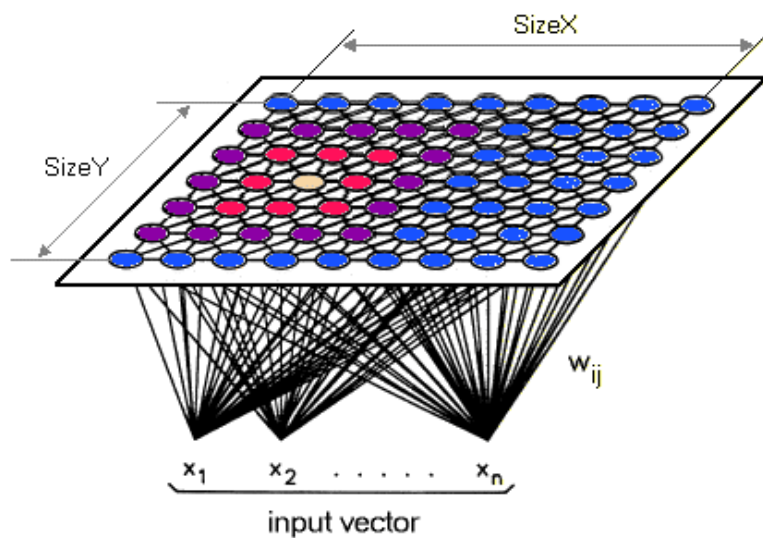


Figure 2.2 – Illustration of a Self-organizing map²

For clustering, self-organizing maps can be used by themselves or using a two-level approach, combined with other clustering algorithms. When used by themselves, each of its units is considered a cluster and each data point is assigned to its BMU. The two-level approach, described by Vesanto & Alhoniemi (2000), starts by forming a large set of prototypes using the SOM, which is much larger than the expected number of clusters in the data. In the following step, these prototypes, which can also be interpreted as “protoclusters”, are combined to form the actual clusters, which can be performed using K-Means. Since each data point set is assigned to a “protocluster”, which by its turn is assigned to a final cluster, in the end the data points are assigned to a final cluster by means of the two-level approach.

2.4. CLUSTERING EVALUATION METRICS

After clustering a dataset it is necessary to assess how good the respective clustering solution is, as well as comparing the solutions generated by different algorithms, which is often necessary to select the final clustering solution. To do so, there are several methods to measure the quality of a clustering solution. These can be divided into two broad groups – intrinsic and extrinsic methods – according to whether the ground truth of clustering is available or not (Han et al., 2012). The ground truth is often built by human experts in the respective research field – each observation is assigned to a cluster label, which can assimilate supervised learning (Han et al., 2012).

Extrinsic methods can be used when the ground truth is available, and work by comparing it with the resulting clustering solution to assess its quality. Intrinsic methods are the ones used when the ground truth is not available. These methods assess the quality of a clustering solution by considering the degree of separation between different clusters, as well as how compact each cluster is (Han et al.,

² Source: <https://codesachin.wordpress.com/2015/11/28/self-organizing-maps-with-googles-tensorflow/>

2012). In this literature review, the focus will be mainly on intrinsic methods, as the dataset used in the project did not have an available ground truth.

2.4.1. Within-cluster sum of squares and Total sum of squares

Within-cluster sum of squares (WSS) measures the variability of the data points assigned to the same cluster. It is a cohesion measure, and the lower its value, the more compact a cluster is. It can be calculated through the following expression:

$$WSS(C_k) = \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (1)$$

Where C_k is one of the n Clusters, μ_k is its centroid, and x_i is a data point belonging to Cluster C_k . It sums the squared distances of each point belonging to Cluster C_k to its centroid. The total within-cluster sum of squares (TSS) is then given by summing the WSS's of the solution's n Clusters:

$$TSS = \sum_{k=1}^n WSS(C_k) \quad (2)$$

Increasing the number of clusters in the solution will normally result in a lower TSS. In the extreme case where the number of clusters equals the number of data points, the TSS will be zero. However, clustering cohesion is being increased at the cost of adding one additional cluster to the solution. One of the methods proposed to find an appropriate number of clusters, that balances both aspects, is the “elbow method” (Liu & Deng, 2021). It is based on plotting the TSS as a function of the number of clusters n . For lower values, an increase in n leads to large decreases in the TSS, whereas when the number of clusters is higher, increasing it will only lead to marginal decreases in the TSS. The elbow method aims to identify the point where the changes in the TSS are rapidly reduced when increasing n – this point is called the ‘elbow’ and may be considered as the true number of clusters. Figure 2.3 shows an example where the ‘elbow’ is clearly visible for $n = 4$.

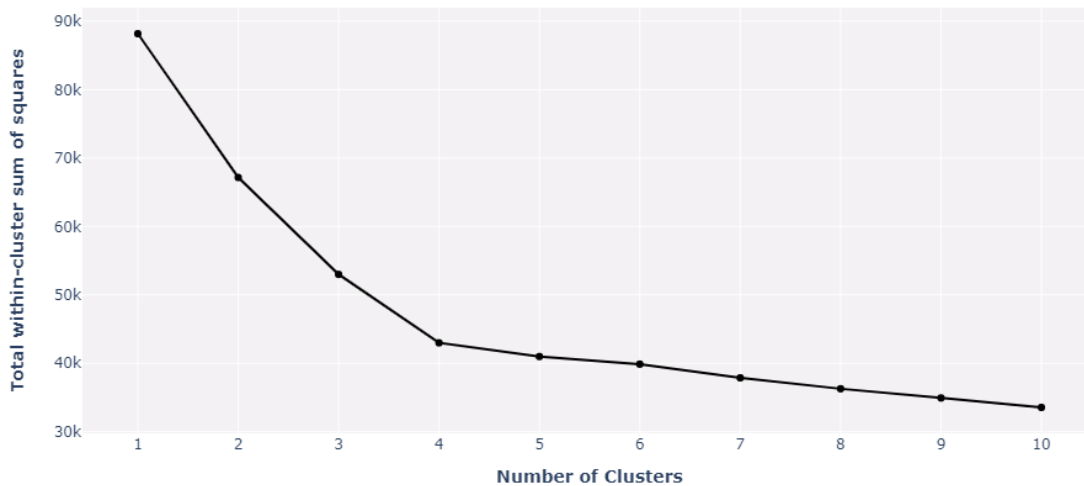


Figure 2.3 – ‘Elbow’ example ($n = 4$)

2.4.2. Silhouette Coefficient

Silhouettes were proposed by Peter J. Rousseeuw in 1987 as “a graphical aid to the interpretation and validation of cluster analysis”. The silhouette coefficient computes the similarity of each data point to its assigned cluster, compared to its similarity to the remaining clusters, thus including the effects of

both cohesion, from points assigned to the same cluster, and separation, from points assigned to a different cluster (Rousseeuw, 1987).

The silhouette coefficient value ranges between -1 and 1, and for each data point i it is calculated according to the following expression:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i) \end{cases} \quad (3)$$

where $a(i)$ is the average distance between the point and all the data points from the same cluster, and $b(i)$ is given by the minimum value of the average distances from the point i to the data points assigned to each cluster except its own, i.e., from the average distances from point i to each of the other clusters, the minimum value defines $b(i)$.

A silhouette coefficient closer to 1 means that the value of $a(i)$, the ‘within cluster’ dissimilarity, is lower than the minimum of the ‘between clusters’ dissimilarities, $b(i)$ (Rousseeuw, 1987). This indicates that the data point was well assigned to its cluster, since it is much closer to the points from its own cluster, than to the data points from the closest one. When the silhouette coefficient is near 0, $a(i)$ and $b(i)$ are approximately equal, which suggests that the data point i is close to the border between its cluster and the closest one, since it is roughly equidistant from both clusters. Finally, if the data point is much closer to the closest cluster than to its own, the value of $a(i)$ will be much larger than the one of $b(i)$, translating into a coefficient value that is close to -1 . Summarizing, a silhouette coefficient closer to 1 indicates that the point was “well clustered”, whereas a value closer to -1 indicates a “misclassification” of the point (Rousseeuw, 1987). Coefficient values that are approximately null indicate an “intermediate case”.

Since the silhouette coefficient is calculated for each datapoint, the fitness of each cluster within a clustering solution can be assessed by calculating its average silhouette coefficient. The quality of the clustering solution as a whole can be measured by calculating the average silhouette coefficient for all data points, which can also be used in a heuristic way, similarly to the elbow method, to deduce the appropriate number of clusters of a dataset (Han et al., 2012). Moreover, the silhouette values of each data point can be plotted and turned into a visualization that aids the assessment of the quality of the clustering solution – a silhouette plot, such as the one in Figure 2.4. However, having to calculate the silhouette value for each point turns it into a computationally expensive metric, specially for large datasets.

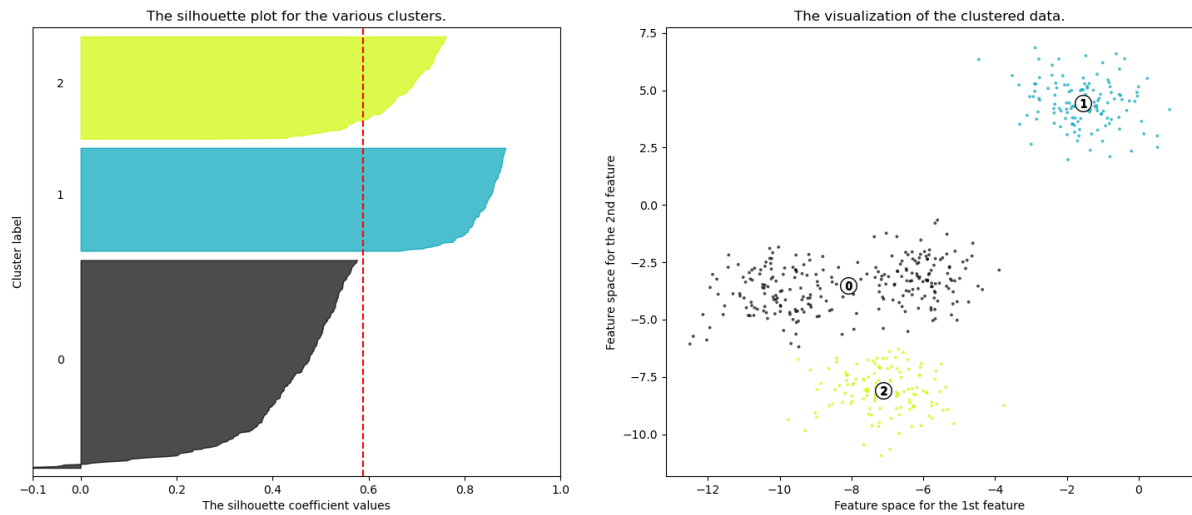


Figure 2.4 - Silhouette plot example³

Figure 2.4 shows an example of a silhouette plot, as well as a visualization of the 2-dimensional data that was clustered using K-Means. Silhouette plots show how well the data points fit into the clusters to which they were assigned, by depicting their sorted silhouette coefficients grouped by cluster. Moreover, the cluster size can be visualized from the thickness of the respective silhouette plot.

2.4.3. Davies-Bouldin Index

Proposed in 1979, the Davies-Bouldin Index is the average similarity between each cluster and the one it is most similar to (Halkidi, 2001). It is defined as:

$$DBI = \frac{1}{n} \sum_{i=1}^n \max_{i \neq j} R_{ij}; \quad (4)$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

where s_i is the average distance between the points in cluster i and its centroid, and d_{ij} is the distance between the centroids of clusters i and j . Its minimum possible value is zero and lower DBI values indicate that clusters are less similar between each other, and thus the clustering solution is better, since one of the main goals of clustering is to have clusters as distinct as possible from each other.

2.5. RELATED WORK – PRACTICAL APPLICATIONS

In this sub-section we explore the results of several customer segmentation approaches found in literature that were applied to different problems, some of which are directly related with the field of dental medicine. The purpose is to delve into the results achieved by researchers using different techniques, applying many of the theoretical concepts that have been reviewed until this point, providing a sense of the clustering results that are obtained in real-life datasets.

³ Source: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

D. Chen et al. (2012) used the K-Means clustering algorithm to segment the customers of an online retailer based on the RFM model, resulting in several meaningful segments from which the main characteristics of their respective consumers could be clearly identified. *Recency*, *Frequency* and *Monetary* were the three variables used as input for clustering, from which resulted five clusters of different sizes and characteristics. Cluster 1 was the least profitable cluster, with high values for *R* and low values for both *F* and *M*. Cluster 2 was characterized by having recent customers with high values for *F* and intermediate ones for *M*, whereas Cluster 3, being the largest cluster in size, represented ordinary customers, with intermediate values for *F* and *M*, and contained several recently registered customers who had only recently begun shopping with the retailer. Clusters 4 and 5 were the most valuable clusters in terms of profitability, with very high values for *F* and *M*, and low *R* values.

In a similar way, Anitha & Patil (2022) also applied the K-Means algorithm on the RFM variables of a retail dataset. However, the authors had clearer criteria regarding the choice of the number of clusters. K-Means was applied with $k = 3$ and $k = 5$, and the respective silhouette scores were assessed. For $k = 3$ the silhouette score was 0.362159752, whereas for $k = 5$ it was 0.3490755342. Therefore, the authors chose 3 as the number of clusters, and assessed the distribution of the RFM variables in the resulting clusters. Cluster 1 had high *R*, and low *F* and *M* values, whilst Cluster 2 had low *R*, and high *F* and *M* values, and Cluster 3 had low values for the 3 RFM variables. Although the authors had more explicit criteria to determine the number of clusters, a characterization of the obtained clusters was not performed. Moreover, the proportion of customers in each cluster is not mentioned.

Arunachalam & Kumar (2018) evaluated the performance of different approaches to cluster data from the UK hospitality industry, aiming to find profitable customer segments. Namely, the authors have clustered the data using the two-level approach of clustering of SOM (Vesanto & Alhoniemi, 2000), using hierarchical clustering. The authors have performed several cluster experiments, ensuring their validity using silhouette coefficients and the Davies-Bouldin Index. The final solution consisted of 3 clusters, which were then profiled by the authors. While Segment A was not deemed as a potential target for marketing activities, Segment C was characterized as a “potential beneficial segment”. Segment B was an intermediate segment which, as the authors mention, could be split into smaller niche sub segments should the value of *K* increase.

Wei et al. (2013) have also applied the two-level clustering approach, combining SOM and K-Means to apply on the RFM variables of a dataset related to the hairdressing industry. 11 clusters were obtained using this approach, which were then characterized by the authors – assessing for each cluster which of the *R*/*F*/*M* features were above average. However, it is not explicit how the number of clusters was determined, and no cluster evaluation metrics are mentioned.

Although research on the applications of customer segmentation in the field of dental medicine is relatively scarce, there are authors who have applied approaches akin to the ones described above, combining RFM models with clustering algorithms such as K-Means or SOM. Wei et al. (2012) developed an extended version of the RFM model, which included the Length variable (*L*) – the duration of the relationship between the customers and the company. The authors then combined the Length RFM (LRFM) model with SOM to cluster the 2258 patients of a children’s dental clinic in Taiwan. However, the monetary variable was excluded from the analysis due to Taiwan’s national health insurance program, since the co-payment and registration fees are the main sources of income for dentists in Taiwan. Once again, the authors do not mention the criteria used to determine the number

of clusters. Nevertheless, the descriptive statistics of the 12 obtained clusters are provided, as well as the number of patients assigned to each cluster. A characterization of the resulting clusters is performed based on their descriptive statistics, from which the following can be highlighted:

- Clusters 1, 5 and 10 – Loyal patients, who have visited recently and have frequent appointments, and have a long-term relationship with the dental clinic;
- Clusters 6 and 8 – Despite their lower than average values for frequency, patients from these clusters also have a long-term relationship with the dental clinic, and have had a recent appointment – hence these patients can also be characterized as loyal, having a stable relationship with the clinic;
- Cluster 11 – New patients, who have had their first appointment with the dental clinic quite recently, and still have had a low number of appointments;
- Clusters 3 and 9 – Although these are also recent patients, the main difference between them and the ones from Cluster 11 resides in the fact that they had more appointments during their short relationship with the dental clinic;
- Clusters 2, 4, 7 and 12 – Non-loyal patients, who have not had an appointment recently, and have extremely low values for frequency.

Wu et al. (2014) also clustered the patients from the same dental clinic, the main difference being the number of patients (1462) and the time frame in which the analysis took place: from July 1st 2009 to June 30th 2011, whilst Wei et al. (2012) used data from January 1st 2009 to July 15th 2010. The authors have used SOM to determine the number of clusters in the data, and then used K-Means to cluster the data. The authors also obtained 12 clusters – a summary of their characteristics is provided below:

- Clusters 2, 6 and 9 – Best patients in terms of LRF characteristics: loyal, with recent and frequent appointments and a long-term relationship with the dental clinic;
- Clusters 1, 5 and 7 – These patients have high values for Length, but their last appointment did not take place very recently. Their Frequency values are below average;
- Clusters 3, 10, 11 and 12 – These patients are described as lost patients, since their last appointment took place a while ago and their frequency values are relatively low;
- Clusters 4 and 8 – New patients, who still have a short-term relationship with the dental clinic, but have had a recent appointment. Patients in Cluster 8 have an above average value for Frequency, and thus are on a better path towards becoming loyal customers.

Overall, both studies achieved similar results in terms of cluster characteristics. Their authors also recommended some marketing initiatives based on the characterization of each cluster. For instance, for clusters with recent patients, it is recommended for the clinic to develop initiatives that aim towards maintaining their relationship with them, pushing them to become loyal patients. For clusters with loyal long-term clusters who have low frequency values, Wei et al. (2012) recommend promotions or discounts to increase their likelihood of visiting the clinic. No specific initiatives are recommended for patients characterized as loyal – the focus should be on maintaining their good relationship with the dental clinic.

Table 2.1 summarizes the customer segmentation applications that were explored in this sub-section. These studies applied the theoretical concepts that were mentioned in the literature review, using the clustering algorithms that were reviewed (K-Means, SOM, or the two-level approach) on a subset of

features, most of them related to the RFM model (or one of its variations). Some of these studies were clearer in how the quality of the results was assessed and in defining the criteria used to select the number of clusters, and others provided more complete characterizations of the resulting clusters, as well as suggesting possible marketing initiatives based on these. Overall, they helped in defining the approach taken in the project, namely applying the mentioned clustering algorithms to feature subsets related to the RFM model, and assessing the quality of the solution by using the mentioned clustering evaluation metrics. The two studies where customer segmentation was applied to dental patients also helped in understanding what kind of results could be achieved in the project, and in defining a clearer path for it.

Study	Method	Summary
D. Chen et al., 2012	Applying K-Means with RFM features in the field of retail	5 clusters with different distributions of the RFM features were identified, each with different levels of profitability.
Anitha & Patil, 2022	Applying K-Means with RFM features in the field of retail, and assessing the silhouette scores	3 clusters with different distributions of the RFM features. A characterization was not performed.
Arunachalam & Kumar, 2018	Two level approach using SOM and hierarchical clustering in the field of hospitality industry, assessing the silhouette scores and Davies-Bouldin Index	3 clusters were found and profiled by the authors. One of the clusters was characterized as a “potential beneficial segment”.
Wei et al., 2013	Two level approach using SOM and K-Means with RFM features, in the field of the hairdressing industry	11 clusters were obtained and characterized, assessing for each which of the RFM features was above average.
Wei et al., 2012	SOM with LRFM features applied to the patients of a dental clinic	12 clusters with different levels of loyalty to the clinic. Marketing initiatives based on the characteristics of each cluster were suggested.
Wu et al., 2014	Two level approach using SOM and K-Means with LRFM features applied to the patients of a dental clinic	12 clusters were also identified, from clusters with non-loyal patients to clusters with loyal patients with a long-term relationship with the clinic. Marketing initiatives based on the characteristics of each cluster were also suggested.

Table 2.1 – Related work summary

3. CUSTOMER SEGMENTATION PROJECT

As previously mentioned, the first half of the 1-year internship was exclusively dedicated to the development of a customer segmentation project, applied to the dental patients of the company – a relatively uncommon application area for this kind of projects. This section will describe what was done across the development of the project: the several phases it had, the models that were tried, as well as the final results that were achieved.

The main goal of the project consisted in understanding the distribution of the company's patients across different segments, and having an overview on the main characteristics of each of them. It was important for the company to then be able to customize the communication with the patients from each segment based on its characteristics. Furthermore, it was important to figure out what characteristics the company's loyal and non-loyal patients had. For the first group, understanding what drove them to become loyal to the company would enable the development of marketing initiatives to maximize their potential value. For the latter, the importance resided on understanding what actions could be performed to improve their retention levels, trying to convert these patients into loyal ones.

The features used for clustering were chosen by combining general ideas from several variations of the Recency-Frequency-Monetary (RFM) model, and there were two main algorithms that were extensively tested for clustering the data: K-Means and SOM. The main challenge was to achieve a quality solution both in quantitative and qualitative terms, i.e., scientific criteria and business meaningfulness and interpretability of the results. Moreover, the data that was available was mostly related to the appointments, treatments and billing of each patient, thus the segmentation process was restricted to the history of the patients with the company, not including the dental healthcare necessities of the patients (if for instance, there was structured data available regarding the dental treatment necessities of the patients).

3.1. TIMELINE

Figure 3.1 shows the planned timeline of the project, as well as the real duration of each phase. As mentioned before, the internship had a duration of 12 months, although roughly only the first half of it was dedicated to the development of the Customer Segmentation Project – which is the time frame pictured on the timeline. Its development was divided into four main phases: Project Planning (PP), Data Analysis and Preparation (DAP), Clustering – Modelling and Results Analysis (C), and Final Results – Documentation and Presentation (FR). Moreover, there were tasks that were transversal to all these phases, such as documenting the developed work and collecting inputs and feedback from the project's stakeholders across its development, in order to meet their requirements and goals.

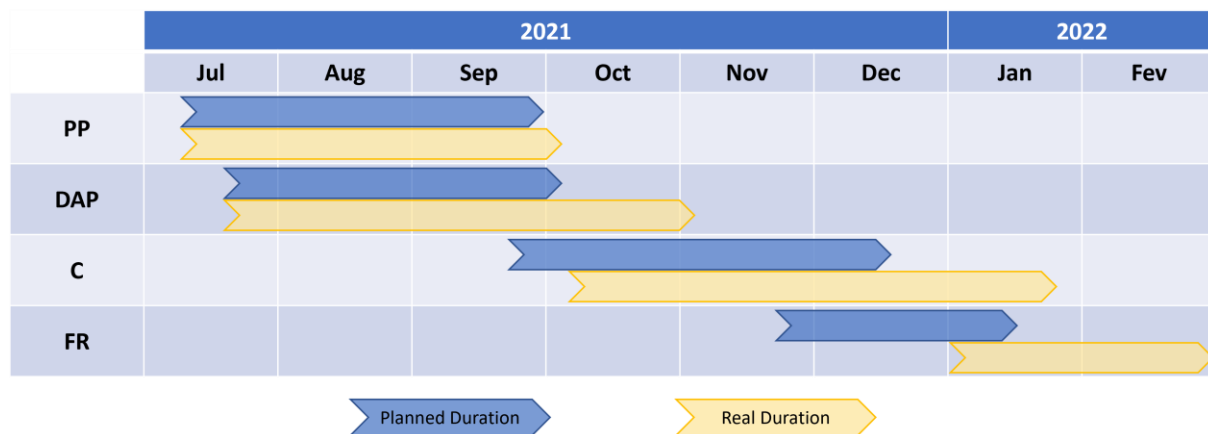


Figure 3.1 – Project Roadmap: Planned vs. Real duration

A brief description of each of the phases is provided below:

- **Project Planning (PP)** – During this phase several meetings were held with the different project’s stakeholders, aiming to understand what their main goals were and to try to plan and conduct the project in a way that would allow it to provide useful insights and information to every business area. Simultaneously, a literature review was made to understand what research has been done on the field of customer segmentation, and more specifically its applications to the case of dental medicine patients.
- **Data Analysis and Preparation (DAP)** – This phase comprised all the tasks related with getting to know and analyzing the data, assessing its quality, building features, and selecting the variables that would later be used for clustering. A dataset with information aggregated by patient was constructed, a process that will be described later on this report. The Project Planning phase overlapped with this phase during most of its duration, since the inputs received from the several stakeholders and the literature review helped defining, for instance, what kind of features would be important to build and use to achieve a quality solution.
- **Clustering - Modelling and Results Analysis (C)** – Several clustering models and approaches were tested across this phase. Different clustering models combined with different subsets of variables were extensively tested, and their results were assessed both quantitatively and qualitatively – applying scientific criteria as well as business sense, respectively. Periodically, intermediate results were presented to the stakeholders, with the goal of defining what aspects could be explored more deeply, as well as assessing the business quality of the results themselves.
- **Final Results - Documentation and Presentation (FR)** – This was the last phase of the project, which consisted in documenting and presenting the final results of the project to the stakeholders, after the final model was selected in the previous phase.

Overall, the first two phases of the project had a real duration similar to the one that was planned, although the DAP phase was slightly prolonged to allow for the exploration and engineering of different features given the results of the first tests of the Clustering phase. This phase was lengthier than what was initially planned due to the extensive testing of several models, to ensure a final quality solution was achieved. Moreover, after sharing intermediate results with the project’s stakeholders across this phase, their feedback was taken into account and originated new ideas for models to test,

some of which were not initially thought of. Naturally, the FR phase only took place once a final model was selected, and thus began later than initially planned.

3.2. MOTIVATION

The main motivation for the development of this customer segmentation project was getting to know the different kinds of patients of the company, allowing for the development of segmented marketing strategies, more suited to the necessities and characteristics of each resulting segment. The desired outcome would then be an increase in the patients' satisfaction and retention levels, thus increasing their value potential to the company.

Additionally, one of the key issues the company intended the project to focus on was understanding the distinctions between loyal and non-loyal patients. In order to achieve their potential worth, it was important to understand the loyal patients' characteristics and what motivated them to stay loyal to the company. It was also important to determine what could be done to increase the retention levels of the remaining patients, to try to turn them into loyal patients.

3.3. METHODOLOGY

The framework that was used to develop the project was mainly based on the CRISP-DM methodology, which stands for Cross Industry Process for Data Mining. It places a lot of emphasis on comprehending the business context of data (Rogalewicz & Sika, 2016), and summarizes the lifecycle of a data exploration project in 6 main stages, as represented in Figure 3.2. The arrows show the main dependencies between each of the stages, although this may vary from project to project. Furthermore, the outer circle in the figure represents the cyclic nature of the data mining process, since it does not end when a solution is deployed (Wirth & Hipp, 2000).

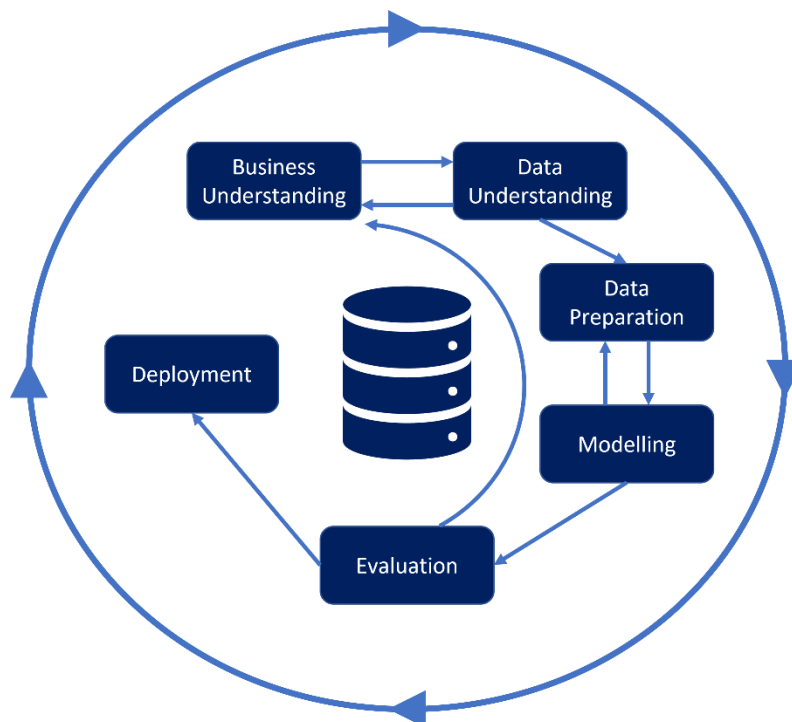


Figure 3.2 - CRISP-DM Methodology

An overview of each of the 6 stages is provided below (Wirth & Hipp, 2000):

- **Business Understanding** – This first stage focuses on comprehending the project's requirements and goals from a business standpoint. Based on this understanding, a data mining problem definition and a rough project plan are then developed.
- **Data Understanding** – This phase begins with an initial collection of the data, followed by an exploration of the data to get familiar with it, as well as identifying if there are any data quality issues, and having a first overview of the data. It is connected to the previous stage, since the definition of the data mining problem and the project planning require at least a basic level of knowledge about the existing data.
- **Data Preparation** – The Data Preparation stage is the process of creating the final dataset from the initial raw data. The final dataset is the one which will be used to feed the models that will be used. Data preparation is performed in multiple steps, such as feature selection, data cleaning, feature engineering, or data transformation.
- **Modelling** – Several models are applied during this stage, and their parameters are tuned to optimal values. This stage is interconnected with Data Preparation, since often new ideas (such as new features to engineer, for instance) arise from the modelling process.
- **Evaluation** – This stage comes after at least one high quality model is achieved. Before moving forward with its deployment, it is crucial to perform a more thorough evaluation of it and review the procedures taken to build it, to ensure it achieves the business goals. Moreover, identifying any significant business issues that have not been appropriately taken into account is one of the main goals of this stage. Finally, a choice on the application of the data mining results should be made at its conclusion.
- **Deployment** – This stage consists of gathering and organizing the knowledge that was acquired throughout the project and sharing it with the project's stakeholders. The deployment stage will be different from project to project – a report or a presentation of the achieved results may be developed, as well as the implementation of a data mining process.

Table 3.1 presents the connection between the 6 stages of the CRISP-DM methodology and the 4 phases in which the timeline of the project was divided into. Overall, the connection between the stages of the CRISP-DM methodology and the timeline's phases is clear from their descriptions.

Timeline Phase	CRISP-DM Stages
Project Planning	Business Understanding
Data Analysis and Preparation	Data Understanding Data Preparation
Clustering – Modelling and Results Analysis	Modelling Evaluation
Final Results – Documentation and Presentation	Deployment

Table 3.1 – Connection between the Timeline Phases and the CRISP-DM Stages

3.4. DATA PREPARATION

The initial stages of the project were devoted to getting to know the available data, exploring it, and assessing its overall quality. Although all the data was stored in an SQL database, it was managed by

an external software company. Given this, the only way to access the data was via Excel files that were exported from the SQL server on a regular basis. The data exploration and analysis was then conducted with Python, resorting mainly to libraries such as *Pandas* or *NumPy*.

Overall, the available data on each patient's history with the company could be divided into three main datasets: **Patients**, **Appointments**, and **Treatments and Invoicing**. Although a detailed description of their content and features cannot be provided, a brief summary of each one with the essential information needed to guide the reader through this report is presented below:

- **Patients:** this dataset mostly contained demographic information such as each patient's age, gender, ZIP code and the patient's unique internal ID, used to identify each patient, since the data was anonymized. Additional data from the patients related to other fields was also provided in this dataset. However, it was not relevant for the development of this specific project.
- **Appointments:** Information about all the existing appointments was present in this dataset, such as their dates, status (scheduled, cancelled or concluded), and the respective patient.
- **Treatments and Invoicing:** Essentially, this was where the data about each patient's treatments and respective invoicing was present. Summarizing, each row contained the patient ID, appointment ID, treatment code, treatment description, treatment category, and the monetary value paid for each specific procedure.

One of the first aspects that was immediately noticed was that there were almost no aggregations by patient about the appointments or treatments and invoicing data, which is something essential for the development of a customer segmentation project. Therefore, the main task of the Data Analysis and Preparation stage was to construct a dataset at patients' level, i.e., where each row also comprises the information about these fields for each patient. This dataset was constructed combining the already existing **Patients** dataset with engineered features built from the information present on the other two datasets. A Python script was developed to automatically perform these aggregations and export the dataset.

Concerning the **Appointments** dataset, the main focus consisted in building the Length (L), Recency (R) and Frequency (F) variables from the LRFM model (Wei et al., 2012), an extended version of the original RFM model that also considers the length of the relationship between the customer/patient and the company. Therefore, for each patient the following variables were built:

- Date of the first concluded appointment;
- Length (L), computed by calculating the difference in days between a pre-defined moment in time or the present and the date of the first concluded appointment;
- Date of the last concluded appointment;
- Recency (R), computed in the same way as Length but considering the difference to the last concluded appointment;
- Frequency (F), given by the number of concluded appointments;
- Number of scheduled appointments, as well as a Boolean variable that reflected if the patient had a scheduled appointment or not;
- Date of the next scheduled appointment (if it exists).

The **Treatments and Invoicing** dataset presented different challenges. The Monetary (M) variable, given by the total amount spent by each patient, would not reflect by itself the nature of the treatments undergone by each patient. Thus, the idea consisted in breaking the Monetary variable into multiple variables, each one given by the amount spent by the patient on a specific treatment or specific treatment category, in order to capture the product effect that other extended versions of RFM tried to capture, such as GRFM (Chang & Tsai, 2011) and RFM/P (Heldt et al., 2021). However, the original treatment categories summed up to 15, some of which with a low number of recorded transactions, which would result in multiple features with low variability – these features would then not be useful for clustering. Therefore, in accordance with the project’s stakeholders, it was decided that the 15 original treatment categories would be condensed into 6 new categories, which would allow to simultaneously capture the product effect and overall solve the low variability issue. A general description of each of the new categories is provided on Table 3.2, in order to provide an idea of the nature of the treatments assigned to each one without delving into profound medical details.

Category	Description
1. Check-up Treatments	Treatments with a check-up nature, mostly performed routinely across spaced periods in time, such as teeth scaling
2. General Treatments	General dental medicine treatments that do not have a check-up nature, and do not fall into specific categories such as replacement of missing teeth or dental alignment
3. Replacement of Missing Teeth	Treatments related to the specialties of Prosthodontics and Implantology
4. Braces and Dental Alignment	Braces and Aligners, as well as other Orthodontics’ treatments
5. Dental Aesthetics Solutions	Treatments with aesthetic purposes, such as dental whitening
6. Common Acts	Procedures that are transversal to the various categories and thus cannot be assigned to a specific one, such as X-rays

Table 3.2 – Description of the new Treatment Categories

The variables with the monetary value spent per category were then built, resulting in 6 additional variables related not only to the Monetary field, but also to the Treatment Category, allowing for the Customer Segmentation to try to capture the already mentioned product effect.

The variables constructed from the **Appointments** and **Treatments and Invoicing** datasets were then linked to the **Patients** dataset via the Patient ID. The final dataset is summarized on Table 3.3. Although it contained several other variables, some that already existed and some that were engineered, the summary provided includes only those that are directly related to the development of this project. It contained a total of 23.951 rows, each one related to a different patient.

Variables	Description
<i>PatientID</i>	Patient ID – unique identifier of each patient
<i>Birthday</i>	Birthday of the patient
<i>Age</i>	Patient's age
<i>FirstDate</i>	Date of the first appointment
<i>L</i>	Length – Days passed since the first appointment
<i>LastDate</i>	Date of the last concluded appointment
<i>R</i>	Recency – Days passed since the last appointment
<i>F</i>	Frequency – number of concluded appointments
<i>M</i>	Monetary – total monetary value spent
<i>AvgM</i>	Average monetary value spent by appointment
<i>M1-M6</i>	Monetary value spent in category 1, 2,...,6.
<i>Scheduled_Bool</i>	True if the patient has a scheduled appointment, false otherwise
<i>NextDate</i>	Date of the patient's next appointment, if it exists

Table 3.3 – Final dataset summary

3.5. DATA EXPLORATION

This sub-section describes the data exploration process that took place after the computation of the previously described variables, which started with an initial assessment of the data quality. The data distribution and the presence of outliers were also explored, as well as the correlations between variables. Although some of these aspects cannot be fully disclosed, the purpose of this sub-section is to provide the reader some general insights into the data that also support some of the decisions that were made.

3.5.1. Data Quality

Overall, no inconsistencies were verified in the final dataset, mostly because these were dealt with during its construction. Specific cases such as appointments that were registered as “concluded” (instead of “scheduled”) in future dates were found, although these inconsistencies only accounted for a residual fraction of the data. Nevertheless, these observations were fixed during the construction process so that they would not produce impossible values in the final dataset (for example, negative values in the *R* variable).

In terms of missing values, only the *Age* variable was slightly affected, with 0.9% of the entries without an assigned value for this variable. Given their small proportion, for these entries the missing values were replaced by the median value of the *Age* variable.

3.5.2. Data Distribution

Although the data distribution cannot be fully disclosed, this sub-section will provide an overview of some of its aspects, as well as an assessment of the presence of observations with extreme values for some variables, which could be regarded as candidates for outliers. Its main purpose is to provide enough context to the decisions that were made regarding the selection of features on the clustering stage.

Starting with the *M1-M6* variables, a quick analysis helps in understanding that some of these categories correspond to more general treatments, which most of the patients have done at least once, whereas other categories are associated to more specific procedures that the average patient has not done. Table 3.4 shows the percentage of patients with the value 0 assigned to each of the *M1-M6* variables, i.e., the percentage of patients that did not do a treatment of the respective category, and thus did not spend any monetary amount on it. Categories 1, 2 and 6 correspond to more general procedures, whereas categories 3, 4 and 5 correspond to more specific procedures – in the case of category 5, less than 3% of the patients have done a procedure of this category, which raises the question whether this variable should be included in the subset of features for clustering, given its low variability.

	<i>M1</i>	<i>M2</i>	<i>M3</i>	<i>M4</i>	<i>M5</i>	<i>M6</i>
Percentage of Zeros	12,62%	44,65%	86,32%	93,20%	97,22%	14,58%

Table 3.4 - Percentage of Zeros for the *M1-M6* variables

Boxplots, proposed by Tukey (1977), are a graphical technique frequently used to summarize numerical data through its quartiles. It displays, the first quartile (*Q1*), the median (*Q2*), the third quartile (*Q3*), and a set of whiskers based on the 1.5 *IQR* value, where $IQR = Q3 - Q1$. Below *Q1*, a whisker is plotted until the lowest data point that falls within a distance of 1.5 *IQR* from *Q1*, whereas above *Q3* a whisker is drawn until the largest data point that falls within a distance of 1.5 *IQR* from *Q3*. Data points outside the interval $[Q1 - 1.5 IQR; Q3 + 1.5 IQR]$ are marked on the plot and considered as “potential outliers”. Outliers are defined as observations that appear to be generated by other mechanisms and seem inconsistent with the rest of the dataset (Gao et al., 2022).

Since the data distribution cannot be fully disclosed, as previously mentioned, boxplots are used in this report as a way to provide enough context on the distributions of the numerical variables without fully disclosing them. Figure 3.3 displays the boxplots of the numerical variables, as well as the percentage of observations that fall outside the whiskers, and thus are considered to be “potential outliers”. The variables were standardized using the Z-Score formula. For each data point, its new value for the variable is given by the following formula:

$$Z_i = \frac{x_i - \bar{x}}{\sigma}, \quad (5)$$

Where x_i is the initial value the data point has for the variable, $\bar{x}$ is the mean value of the variable, and σ is its standard deviation. It is visible from the boxplots that specially for the monetary variables there are several patients that have extreme values. For the *M3*, *M4* and *M5* variables, it is possible to verify the effects of having high percentages of patients with the value 0 (before standardization) on these variables – the *IQR* value is 0, given that there is no variation in the middle 50% of the data ($Q1 = Q3$). Consequently, patients with a value different than 0 for these variables fall outside the whiskers and are considered as “potential outliers”, which does not make much sense businesswise, as these are company’s patients who have spent monetary amounts on these categories. For other variables with more variability, such as *R* or *F*, there are still data points which fall outside the whiskers – 4.2% and 7.87%, respectively. Although there are data points that fall outside the whiskers, they correspond to actual patients of the company, and their extreme values do not come from measurement errors or data quality issues. Therefore, they were not classified as outliers. Furthermore, those with extreme values on the monetary variables have a significant impact on the company’s turnover. In order to

address these business concerns, it was agreed that all the data points would be included in the analysis.

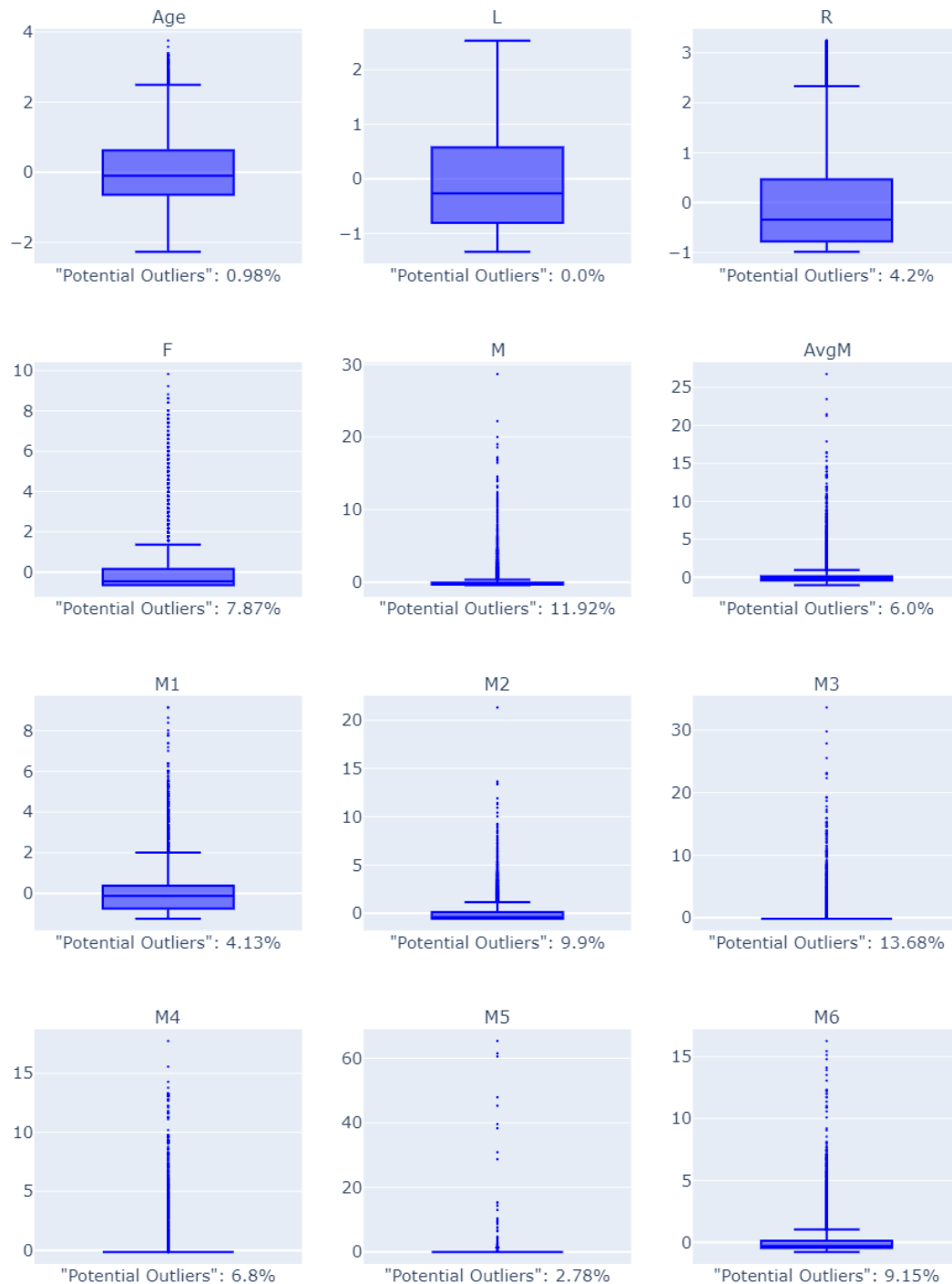


Figure 3.3 - Boxplots of the numerical variables

3.5.3. Correlation between variables

Before clustering the data, it is important to assess the correlation between the variables that are going to be used. The correlation values among the numerical variables are presented in Table 3.5. It can be verified that most of the correlation values are relatively low, although there are still some pairs of variables with considerable correlation values. In the case of the *L* and *R* variables, their higher

correlation value (0.729) can be explained by the fact that there is a large proportion of patients (35.86%) who have had only a single appointment, resulting in the same value for both *L* and *R*. The *M* and *F* variables also have a relatively large correlation value (0.683), which points to the general idea that more appointments are interrelated with more money being spent by each patient. Moreover, the *M* variable presents relatively high correlation values with the *AvgM* and *M1-M6* variables. Regarding the *M1-M6* variables, it is important to highlight that, overall, these variables have low correlation values between each other – only the *M2* and *M6* variables have a higher value (0.507). In terms of clustering, using these monetary variables may allow to capture the product effect without adding highly correlated variables to the analysis.

	<i>Age</i>	<i>L</i>	<i>R</i>	<i>F</i>	<i>M</i>	<i>AvgM</i>	<i>M1</i>	<i>M2</i>	<i>M3</i>	<i>M4</i>	<i>M5</i>	<i>M6</i>
<i>Age</i>	1											
<i>L</i>	0.069	1										
<i>R</i>	0.057	0.729	1									
<i>F</i>	0.1	0.237	-0.241	1								
<i>M</i>	0.099	0.126	-0.174	0.683	1							
<i>AvgM</i>	0.061	-0.016	-0.121	0.254	0.712	1						
<i>M1</i>	-0.036	0.252	-0.195	0.423	0.233	0.119	1					
<i>M2</i>	0.111	0.138	-0.168	0.715	0.519	0.259	0.272	1				
<i>M3</i>	0.154	0.082	-0.082	0.424	0.824	0.563	0.078	0.308	1			
<i>M4</i>	-0.089	0.03	-0.127	0.344	0.474	0.449	0.101	0.053	-0.006	1		
<i>M5</i>	0.008	0.024	-0.025	0.11	0.282	0.175	0.04	0.154	0.124	-0.001	1	
<i>M6</i>	0.084	0.143	-0.167	0.636	0.474	0.241	0.311	0.507	0.294	0.148	0.156	1

Table 3.5 - Correlation between the numerical variables

3.6. CLUSTERING – MODELS AND ANALYSIS

After the preparation and exploration of the data, the Clustering phase took place, where several tests and assessments were performed until achieving a final solution. As previously mentioned, the results were assessed resorting to both scientific and business criteria, as the main goal consisted in having a final solution that achieved quality values on different clustering evaluation metrics, but also one that could provide useful insights to the business and the project's stakeholders.

Multiple tests were performed with different feature combinations – this report will cover three main combinations:

1. A baseline model that used all the features;
2. A simple model with only the Recency, Frequency, Monetary and Length (*RFML*) variables;
3. A model with a final subset of features, chosen in accordance with business goals supported by scientific criteria.

The first and second combinations were used to explore the data and perform initial assessments of the clustering results at the earlier phases of the Clustering stage, which later led to the selection of the final subset of features. Therefore, the first two models will not be explored in this report as deeply

as the final one. The main purpose of including them is to provide some context for the subsequent decisions that were made regarding the selection of features, as well as showing the results obtained in the early stages of Clustering. The final model will then be the main focus of this sub-section.

The main algorithm used to cluster the data was K-Means, due to its simplicity and the interpretability of its results. However, for the final subset of variables, SOM's were also used for clustering, in order to assess the results obtained using a different clustering algorithm. The results obtained with both algorithms will be compared. This method was selected since it is able to find more complex patterns in the data, which may not be captured by simpler algorithms such as K-Means, while being less sensitive to noisy data and outliers (Vesanto & Alhoniemi, 2000).

To quantitatively assess the quality of the results returned by each model, as well as to select an appropriate value for the number of clusters, the following evaluation metrics were used:

- **Total Sum of Squares (TSS):** The Total Within-cluster Sum of Squares (TSS) is calculated by summing the Within-cluster Sum of Squares (WSS) of the solution's K clusters. Since increasing K will typically result in a lower TSS value, the elbow method was used to find a value for K that would balance between having a quality solution in terms of cluster cohesion and the cost of having one additional cluster on the solution. Using the elbow method allowed to perform initial assessments on the appropriate number of clusters in the data.
- **Silhouettes and Average Silhouette:** The silhouette coefficient includes the effects of both cohesion (from points assigned to the same cluster) and separation (from points assigned to a different cluster) to compute the similarity of each data point to the cluster it is assigned to. It is calculated for each data point and its value ranges between -1 and 1 – a value closer to 1 indicates that the data point is well assigned to its cluster, whereas a value closer to -1 indicates a misclassification of the point (Rousseeuw, 1987). The fitness of each cluster within a clustering solution was assessed by computing the average of the silhouette coefficients of its data points, and the overall quality of the clustering solution was measured by calculating the average silhouette coefficient for all data points. Since this metric can also be visualized in the same way as the TSS (Han et al., 2012), the average silhouette coefficients were also plotted as a function of the number of clusters, which helped in assessing appropriate values for K .
- **Davies-Bouldin Index (DBI):** DBI is the average similarity between each cluster and the one it is most similar to. Lower DBI values indicate a better clustering solution, since these mean that clusters are less similar between each other. Its minimum possible value is zero, and it was used as a complementary metric to compare between different clustering solutions.

Before clustering the data, all the variables were normalized using the already mentioned Z-Score Formula. The similarity between the data points was established using the Euclidean distance measure. Given two d -dimensional data points x_i and x_j , the Euclidean distance between both is given by the following expression:

$$D(x_i, x_j) = \sqrt{\sum_{l=1}^d |x_{il} - x_{jl}|^2}, \quad (6)$$

3.6.1. Baseline model

To have an initial overview of the data and to establish a baseline for the clustering results, a simple model that used all the numerical variables was tested. For each value of K , K-Means was run using 100 different seed initializations, and the solution that returned a lower TSS value was selected. The TSS and average silhouette plots are represented on Figure 3.4, where it is visible that for $K = 7$ the TSS value starts to plateau, and the average silhouette has a relatively high value before dropping considerably (for $K = 8$). Therefore, 7 was the selected value for K for the baseline model.

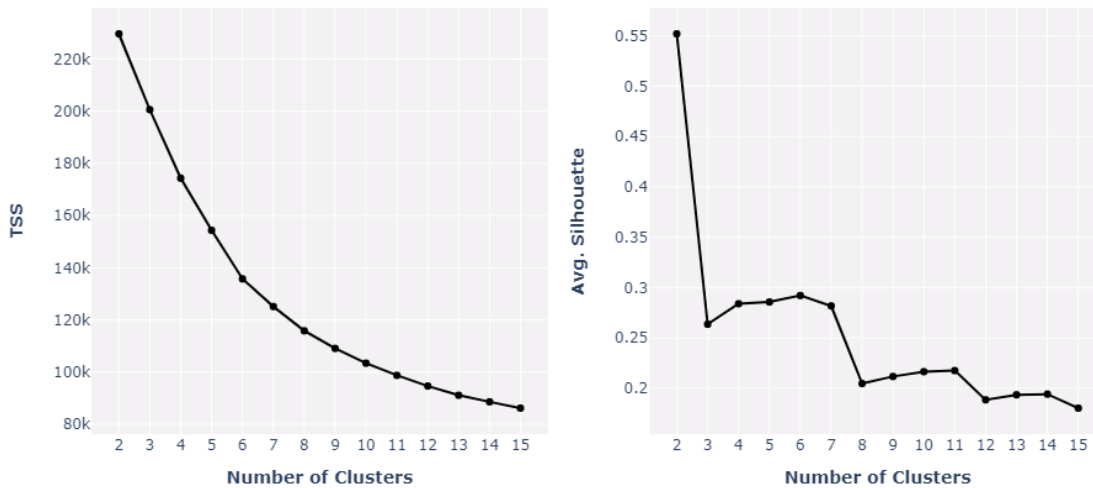


Figure 3.4 - TSS and Average Silhouette plots for the Baseline model

As Table 3.1 shows, the resulting clusters had varying sizes, from Cluster 1, with more than half of the patients (54,81%) assigned to it, to Cluster 5, which only contained a residual number of patients. In terms of the average silhouettes for each cluster, Clusters 2 and 7 had the lowest values, 0.0340 and 0.0120, respectively, whereas the remaining Clusters had average silhouettes values of at least 0.22.

Cluster	Size (%)	Avg. Silhouette	General Description
1	54,81%	0,3783	- Recent Patients - First Appointment took place quite recently - Low number of appointments and low monetary value
2	18,68%	0,0340	- Check-up Patients - Average number of appointments - Higher spendings on Treatment Category 1
3	20,04%	0,3131	- Patients who have stopped coming - High values for R and L - Low number of appointments and low monetary value
4	1,57%	0,2205	- Premium Orthodontics Patients - High number of appointments - High spendings, especially on Treatment Category 4
5	0,04%	0,5036	- Premium Aesthetics Patients - High spendings, especially on Treatment Category 5
6	0,48%	0,2309	- Premium Implantology Patients - High number of appointments - High spendings, especially on Treatment Category 3
7	4,39%	0,0120	- Intermediate Patients - Intermediate spendings and number of appointments

Table 3.6 – Overview of the clusters obtained from the Baseline model

The baseline model allowed to have an initial overview on the data's underlying cluster structure without selecting any specific subset of features. Moreover, some of the general characteristics of the resulting clusters were present in the clusters obtained using other approaches and feature subsets, including the *RFML* model and the final model.

3.6.2. *RFML* model

As previously mentioned, it was decided to test a simple model that would only use the *RFML* features, in order to assess the cluster structure in the data using only these 4 features, as well as to be able to compare the results with the ones obtained in the literature using similar approaches.

The approach that was adopted was the same as for the baseline model: for each K , the K-Means algorithm was run using 100 different initializations, from which the solution that returned a lower TSS value was selected. Figure 3.5 contains the TSS and average silhouette plots obtained for this model, where it can be visualized that 4 and 8 are two possible values for K . However, given that when $K = 4$ the TSS plot is still in a descending phase, 8 was the selected value for K , since for this value the TSS is starting to plateau and the average silhouette value is at a local maximum.

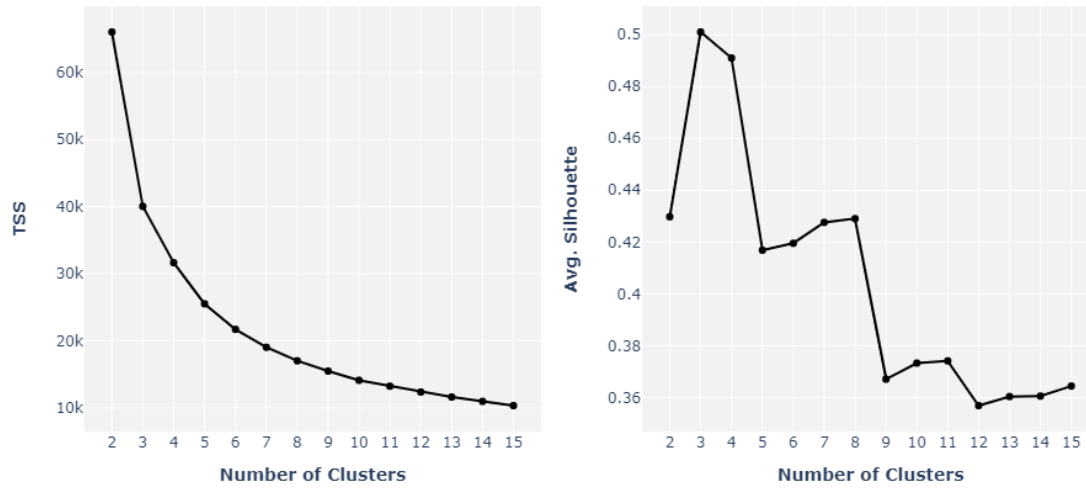


Figure 3.5 - TSS and Average Silhouette plots for the RFML model

The results (dimension, average silhouette, and centroids of each cluster) obtained for $K = 8$ are presented in Table 3.7. Once again, the clusters have varying sizes, but all of them have an average silhouette value of at least 0.20. These results cannot be directly compared with the ones from the baseline model, given the differences in the feature subset that was used to cluster the data. However, it can be verified that there are some similarities in the characteristics of certain clusters in both models, such as the ones between Cluster 3 from the baseline model and Cluster 2 from the *RFML* model (patients that have stopped coming), or between Cluster 1 of both models (recent patients). In the *RFML* model, Cluster 4 is similar to Clusters 1 and 2 in terms of the *F* and *M* variables (low number of concluded appointments and low spending values) but has values for *R* and *L* between the ones of the two mentioned Clusters. All the remaining clusters comprise of patients who have had a concluded appointment recently and have above average *F* and *M* values, which range from intermediate to very high values.

Cluster	Size (%)	Avg. Silhouette	Avg. <i>R</i>	Avg. <i>F</i>	Avg. <i>M</i>	Avg. <i>L</i>
1	39,30%	0,5324	-0,6217	-0,4233	-0,2811	-0,8685
2	13,56%	0,5456	2,0770	-0,3755	-0,2702	1,6079
3	4,00%	0,2168	-0,6327	3,0262	1,5757	0,6935
4	22,37%	0,4011	0,4021	-0,4262	-0,2785	0,1029
5	0,23%	0,3535	-0,6782	3,9797	12,4662	0,7222
6	6,69%	0,2290	-0,1769	0,5605	0,0754	1,4610
7	12,40%	0,2278	-0,6331	0,8369	0,3488	-0,2698
8	1,46%	0,2097	-0,6978	2,8385	4,7943	0,4476

Table 3.7 – Clusters obtained from the RFML model

These results were presented to the project's stakeholders at an intermediate phase of the project, aiming to obtain feedback on what potential directions could the segmentation process take. Despite the model's simplicity and interpretability, one of the main drawbacks that was identified was that it did not capture the already mentioned product effect, as it did not include any of the *M1-M6* variables, not enabling to group the patients based on the treatment categories where they spent the most. Given this, after several tests using different feature combinations, a final subset of features was

selected together with the project's stakeholders, which was used to develop the final model, described in the next sub-section.

3.6.3. Final model

As previously explained, a final subset of variables for clustering was selected together with the project's stakeholders having in consideration the business goals, supported by scientific criteria and the intermediate results that were presented. The final subset included the following variables: L , R , F , $M1$, $M2$, $M3$ and $M4$. The main rationale behind its choice consisted in capturing the effects of the *LRFM* model as well as the product effect.

Since R and F are elementary variables across several segmentation projects that are based on the *RFM* variables, these were automatically included to allow the segmentation to include the effects of the relationships between the patients and the company – allowing the recency of each patient's last appointment as well as each patient's total number of concluded ones to influence the segmentation process. Although L and R presented relatively high correlation values between them, it was decided to also include the L variable, mainly in order to differentiate between older, intermediate, and more recent patients. Regarding the variables related to the monetary field, only the $M1$ - $M4$ variables were included in the segmentation process. The $M5$ variable was not included due to its low variability, having 97.2% of the patients with a null value assigned to it, whereas the $M6$ variable was not included because it did not add useful information to the segmentation process in terms of treatment categories, since this variable is related to procedures that are common across the various fields of dental medicine. After discussing with the project's stakeholders, it was decided not to include the *Age* variable. It was considered that the information related to the treatment categories, provided mainly by the $M1$ - $M4$ variables, would be the main focus.

For this subset of variables, both K-means and SOM were used to cluster the data, and thus several tests were performed. As previously mentioned, SOM was tested to assess the results obtained using a different clustering algorithm and compare them with the ones obtained using K-Means. This extensive process of performing several tests and comparisons then lead to the selection of a final solution among the multiple ones returned by both algorithms.

3.6.3.1. K-Means using the final subset of features

Using the K-Means algorithm, the same approach that was used for the baseline and *RFML* models was adopted, running it with 1000 different initializations instead of 100 for each value of K , and then selecting the solution which returned a lower TSS value. The plots with the TSS and average silhouette values as a function of the number of clusters are represented on Figure 3.6. In this particular case, the choice of the value for K is not as clear as for the previous two models, given that any value between 5 and 7 is an acceptable choice. For $K = 5$, the average silhouette value reaches a local optimum, although the TSS plot is still at a relatively steep descending phase, whereas for $K = 6$ and $K = 7$ there are slight downgrades in the average silhouette values, but the slope of the TSS plot smoothens. Having in consideration that the selection of the value for K has an inherent subjectivity to it, since it is heuristically derived and is not limited to an exact value, the value selected for K was 7, as it would allow to have a higher number of clusters, and thus a lower TSS value, without significantly decreasing the average silhouette value.

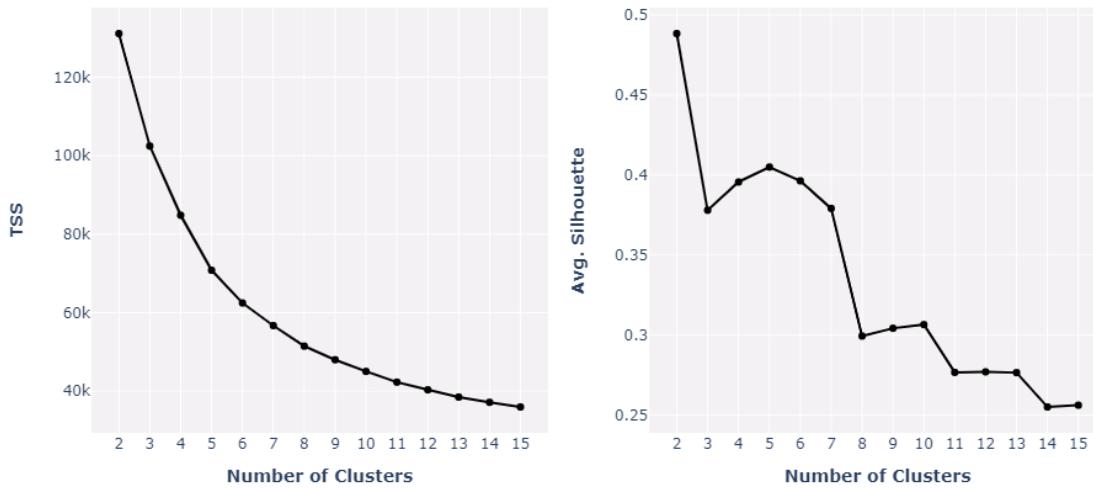


Figure 3.6 – TSS and Average Silhouette plots for the final K-Means model

3.6.3.2. Self-organizing maps using the final subset of features

Given their ability to find more complex patterns in the data, and their lower sensitivity to local optima and outliers, SOM was the algorithm selected to cluster the data and compare its results with the ones obtained with the K-Means algorithm. It was only used for the final subset of features, as its training process is lengthier and depends on more hyperparameters than K-Means. The current sub-section describes the testing process and the results that were obtained.

The process consisted in training several two-dimensional SOM's with different map sizes, and then running K-Means on top of their units, from which would result clusters composed by SOM units. The two-level approach described by Vesanto & Alhoniemi (2000) was used, and thus each data point was then assigned to the cluster of its best matching unit (BMU), resulting in a final clustering solution. The metrics used to evaluate each clustering solution were the same as the ones used for K-Means: TSS and average silhouette value. Several map sizes were tested, from rectangular to squared shaped maps, using random initializations, gaussian neighborhood function, and a hexagonal lattice. For each map size five SOM's were trained with different initializations, given that different initializations lead to different resulting SOM's for the same map size. Both the unfolding phase and the fine-tuning phase had 400 iterations each, and the final quantization error of each map was monitored. Then, K-Means was run 5 times for each value of K on top of each SOM, and the clustering results were assessed using the already mentioned metrics.

The different squared maps ranged from the size of 20x20 to 70x70, and several size combinations were tested for the rectangular maps, such as 40x20, 50x30, or 70x50. A total of 80 different SOM's were trained, with varying map sizes and different initializations, and K-Means was run 5 times on top of each SOM for each value of K , resulting in 400 different configurations that were tested. The SOM and K-Means initializations were randomized by changing the value of their seeds – this way it was possible to get random but reproducible results. This report will only cover 5 configurations that overall returned the best clustering results, from which a final one will be selected. The general characteristics and main results of these configurations are presented on Table 3.8.

Config. Nr.	Map Size	SOM Seed	K-Means Seed	TSS (K=6)	Avg. Silhouette (K=6)	TSS (K=7)	Avg. Silhouette (K=7)
181	70x70	1	0	64981,11	0,3840	58890,04	0,3762
186	70x70	2	0	65061,48	0,3911	58807,24	0,3721
274	60x30	4	3	67075,17	0,3937	59765,22	0,3752
318	60x50	3	2	65679,37	0,3985	59247,06	0,3727
347	70x30	4	1	66765,22	0,3881	60002,97	0,3725

Table 3.8 – Best SOM configurations and their results

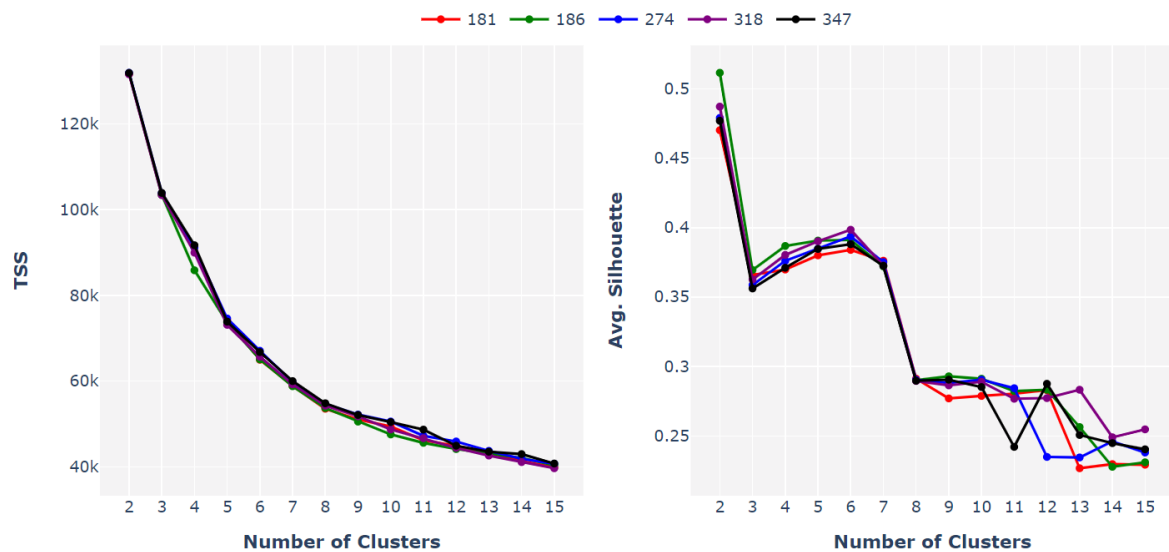


Figure 3.7 - TSS and Average Silhouette plots for the best SOM configurations

As Figure 3.7 shows, overall both the TSS and Average Silhouette plots behave in the same way for the different SOM's, with slight differences in the metrics' values among the different K values. Applying the same criteria that have been used to select a value for K in the other models, both 6 and 7 appear to be possible values for K in the different SOM's. For $K = 6$, the TSS plots can be considered to be at an 'elbow point', and the average silhouette plots are at a local optimum, whereas for $K = 7$ there is a slight decrease on the TSS values, as well as a slight decrease on the Average Silhouette values, although it allows having an additional cluster without a heavy penalization on the second metric (such as the one that happens from $K = 7$ to $K = 8$). Applying the same line of thought as the one applied for the selection of K for the K-Means model, 7 was the selected value for K , as it allowed to have an additional cluster without a heavy penalization on the average silhouette value of the solution. Moreover, it also allowed to compare 2 different solutions (K-Means and SOM) with the same number of clusters. Regarding the choice of the final configuration, it is visible on Figure 3.7 that both the TSS and Average Silhouette values are close among the different configurations. However, Table 3.8 shows that configuration 181 presents the second best TSS value and the best average silhouette value for $k = 7$. Therefore, this configuration was selected as the final SOM model.

3.6.3.3. Selection of a final solution

After having selected a model for each of the two algorithms, the final step consisted in assessing which of the two models would be selected as the final solution. The solutions provided by each model

were compared both quantitatively and qualitatively. Both models having the same number of clusters allowed for simpler and more straightforward comparisons, using the different metrics that have been used to evaluate the clustering solutions, as well as comparing their results and respective interpretations.

The summaries of the clusters produced by the K-Means and SOM models are presented on Table 3.9 and Table 3.10, respectively, which include their relative size, average silhouette values, and centroids. To simplify comparisons between the two, the clusters are ordered by their size in descending order in both tables.

In general terms, it can be verified that the clusters produced by both models have similar sizes and interpretations – there are no significant differences in the centroids when comparing clusters with the same number from one model to the other. One noticeable difference is that the smaller clusters (5, 6 and 7) are slightly larger in the SOM solution than in the K-Means one. Nevertheless, the decision regarding which of the two models would be selected as the final one would not come down to the cluster interpretations, as the results from both models were quite similar in terms of the clusters' main characteristics.

Cluster	Size (%)	Avg. Silhouette	Avg. L	Avg. R	Avg. F	Avg. M1	Avg. M2	Avg. M3	Avg. M4
1	53,03%	0,4820	-0,6503	-0,3911	-0,4251	-0,3053	-0,3785	-0,1489	-0,1288
2	19,87%	0,4139	1,2836	1,6896	-0,3891	-0,3803	-0,3169	-0,1389	-0,1493
3	12,58%	0,1307	-0,1799	-0,4837	0,8513	0,0655	1,0652	0,1395	-0,0552
4	8,91%	0,1860	0,8819	-0,3692	0,3918	1,8996	0,0111	-0,0796	-0,0962
5	3,30%	0,0615	0,7786	-0,4843	2,9548	1,4237	3,3902	1,1847	-0,0514
6	1,77%	0,2756	0,3044	-0,8614	2,5981	0,6914	0,3470	-0,0579	6,5541
7	0,54%	0,2544	0,5389	-0,6057	3,1822	0,1526	1,9718	10,6963	-0,1540

Table 3.9 – Summary of the clusters obtained with the final K-Means model

Cluster	Size (%)	Avg. Silhouette	Avg. L	Avg. R	Avg. F	Avg. M1	Avg. M2	Avg. M3	Avg. M4
1	53,01%	0,4835	-0,6636	-0,4041	-0,4087	-0,3118	-0,3544	-0,1453	-0,1386
2	20,87%	0,3857	1,2456	1,6351	-0,3770	-0,3778	-0,3055	-0,1378	-0,1493
3	10,64%	0,1563	-0,1715	-0,4882	0,8960	0,1639	1,0930	0,1248	-0,0819
4	8,82%	0,1813	0,8578	-0,3774	0,2906	1,8275	-0,0516	-0,1135	-0,1226
5	3,52%	0,0463	0,6950	-0,5023	2,6638	1,3809	3,2879	0,8062	-0,1026
6	2,21%	0,1894	0,1997	-0,8611	2,3825	0,5793	0,2939	-0,0535	5,8296
7	0,92%	0,0988	0,6011	-0,5533	3,0377	0,4386	1,9231	8,1997	-0,1352

Table 3.10 – Summary of the clusters obtained with the final SOM model

Given the previously mentioned similarities, the selection of the final model was based on clustering evaluation metrics. The silhouette plots of the K-Means and SOM models are presented on Figure 3.8 and Figure 3.9, respectively. Both models had very close average silhouette values, which are represented by the red dashed lines in the figures. The K-Means model had an average silhouette value

of 0.3792, whereas the value for the SOM model was 0.3762. However, it is visible that the SOM model's silhouette plots have a spikier shape on their left side, indicating that in this solution there are points that are clearly assigned to the wrong cluster, given their extremely low silhouette values. In the K-Means solution there are points with negative silhouette values, although these are not as extreme as in the SOM solution.

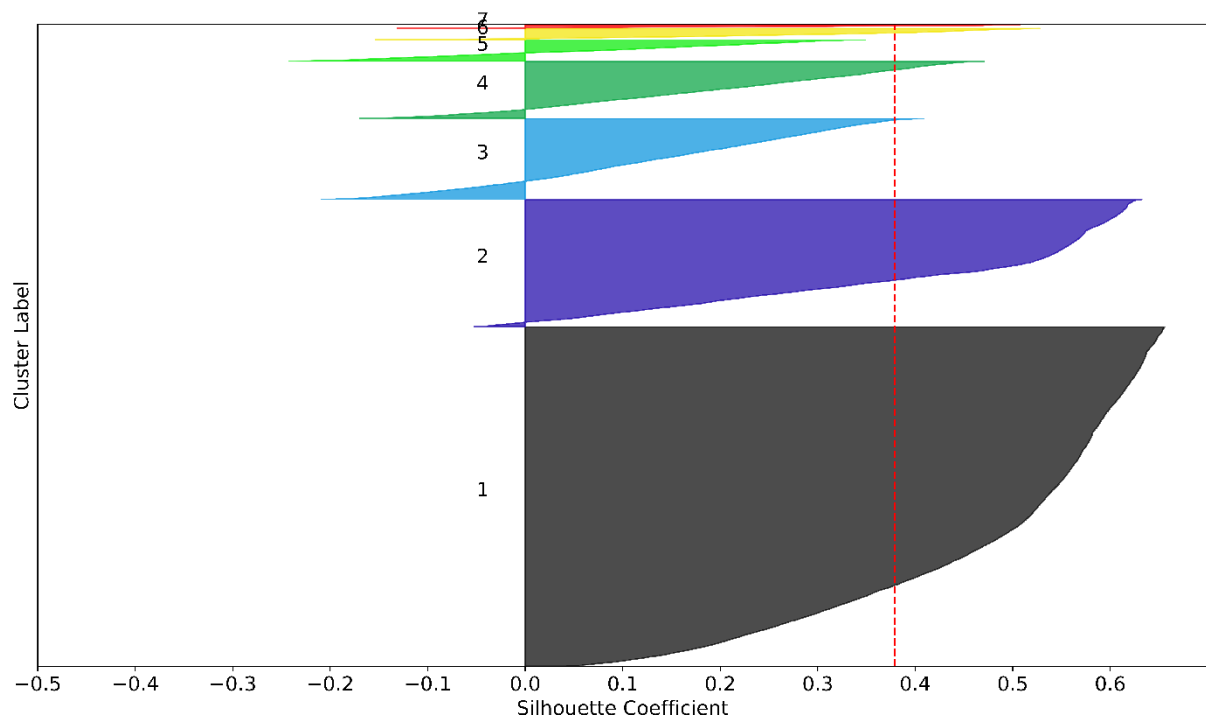


Figure 3.8 – Silhouette Plot of the final K-Means model

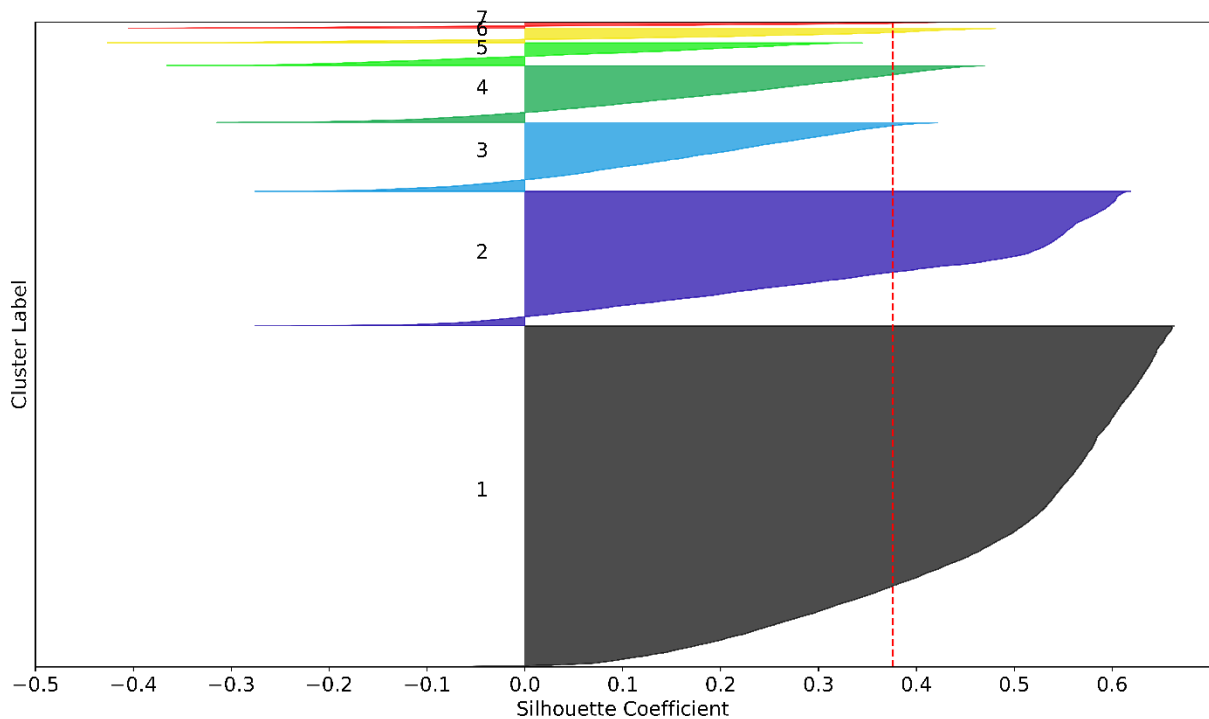


Figure 3.9 – Silhouette Plot of the final SOM model

For a final comparison between both models, their TSS and average silhouette values are displayed on Table 3.11, as well as their DBI values. Since there were no significant differences regarding the clustering structure returned by the two models, the final decision was based on which model had better values for the clustering evaluation metrics. Besides the TSS and average silhouette values, which have been used across this report, the Davies-Bouldin Index was introduced to break the tie in case one of the models had a better TSS value and the other had a better average silhouette value. However, it's clear from the table that the K-Means model outperformed the SOM model when considering all the metrics. Its TSS value is lower, which indicates a more cohesive clustering structure, given that both solutions have the same number of clusters. Its average silhouette value is slightly higher – combined with the analysis of the silhouette plots from both models, this is also considered an advantage for the K-Means model. Finally, the K-Means model also achieved a lower Davies-Bouldin Index, and the lower this value the better, since it indicates that the clusters are less similar between each other.

Model	TSS	Average Silhouette	Davies-Bouldin Index
K-Means	56724.00	0.3792	1.1001
SOM	58890.04	0.3762	1.1385

Table 3.11 – Clustering evaluation metrics for the final models

Summarizing, the two models had a similar clustering structure in terms of the sizes of each cluster and their centroids, which made the decision regarding the final model lean towards the models' evaluation metrics. The K-Means model outperformed the SOM model for the 3 evaluation metrics that were selected – TSS, Average Silhouette, and Davies-Bouldin Index. Therefore, it was selected as the final model, and its results and interpretation are explored in detail in the next section.

3.7. RESULTS AND DISCUSSION

In this sub-section the clustering results from the final model are explored and discussed, and the resulting clusters are described and analyzed. Besides showing the clusters' sizes and centroids, Table 3.12 also provides information about the turnover percentage of each cluster, as well as their percentage of patients that had a scheduled appointment.

Cluster	1	2	3	4	5	6	7
Size (%)	53,03%	19,87%	12,58%	8,91%	3,30%	1,77%	0,54%
% Turnover	17,74%	6,91%	22,62%	9,35%	17,21%	14,79%	11,38%
% Scheduled	27,03%	0,40%	37,86%	37,37%	47,22%	87,26%	60,77%
Avg. L	-0,6503	1,2836	-0,1799	0,8819	0,7786	0,3044	0,5389
Avg. R	-0,3911	1,6896	-0,4837	-0,3692	-0,4843	-0,8614	-0,6057
Avg. F	-0,4251	-0,3891	0,8513	0,3918	2,9548	2,5981	3,1822
Avg. M1	-0,3053	-0,3803	0,0655	1,8996	1,4237	0,6914	0,1526
Avg. M2	-0,3785	-0,3169	1,0652	0,0111	3,3902	0,3470	1,9718
Avg. M3	-0,1489	-0,1389	0,1395	-0,0796	1,1847	-0,0579	10,6963
Avg. M4	-0,1288	-0,1493	-0,0552	-0,0962	-0,0514	6,5541	-0,1540

Table 3.12 – Overview of the final model's clusters

A summary of each of the resulting clusters is provided below, as well as potential marketing actions and goals based on their characteristics:

1. **Recent Patients:** Cluster 1 is the largest in terms of number of patients, comprising 53.03% of the patients of the dataset, although it only has a weight of 17.74% on the company's total turnover. It is mainly characterized by patients who have had their first appointment recently, who thus have a low number of appointments and low spending levels. A similar marketing approach to the ones recommended by Wei et al. (2012) and Wu et al. (2014) for patients with these characteristics could be taken – developing initiatives with the goal of maintaining the company's relationship with patients from this cluster, so they are encouraged to become loyal patients.
2. **Defectors:** Cluster 2 comprises 19.87% of the patients and has a weight of 6.91% on the total turnover. It generally consists of patients who have had their last appointment a long time ago and have not come back since. These patients also have a low number of appointments and low spending levels – they can be characterized as “defectors”, in the sense that they stopped resorting to the company's dental medicine services. One of the aspects that helps confirming this general idea is the residual number of patients with scheduled appointments in this cluster – only 0.3%. Given its considerable number of patients, one of the marketing goals for this cluster could be to try to recover some of these patients to resort to the company's services again.
3. **Intermediate Low:** Cluster 3 consists of relatively recent patients with an intermediate number of concluded appointments, having spent intermediate monetary amounts, especially on Category 2 – General Treatments. It contains 12.58% of the total number of patients, and it is the cluster with the highest weight on the company's total turnover (22.62%). Although these patients are relatively recent, they present a higher degree of loyalty when compared with patients from Cluster 1, in the sense that they already have a considerable number of concluded appointments in a not so long period of time. Nevertheless, the company should still aim towards maintaining these patients and ensuring that they become loyal ones.
4. **Check-up/Sporadic Patients:** Patients in Cluster 4 are characterized by having relatively high consumptions on Category 1 – Check-up Treatments. These patients have an intermediate number of appointments, having had their first one a long time ago, and their last one quite recently – these patients have their appointments widely spaced in time, giving the general idea that these are patients that keep coming for dental check-up treatments (which by nature are widely spaced in time). More than a third (37.37%) of the patients belonging to Cluster 3 have a scheduled appointment. Even though the number of concluded appointments of these patients is not high, their relationship with the company is relatively long and their last appointment took place recently. This means that, despite not being patients with a high level of spendings, they can be characterized as loyal to the company, since they keep coming back to have appointments with the company.
5. **Intermediate High:** In general terms, patients belonging to Cluster 5 have a high number of concluded appointments. These patients have been resorting to the company's dental services for some time, having a high value for the *L* variable, and had their last appointment quite recently, having a low value for the *R* variable. In terms of treatment categories, these patients have higher than average monetary spendings on categories 1, 2 and 3. Although it only contains 3.30% of the total number of patients, Cluster 5 has a weight of 17.21% on the

company's total turnover. These patients can be considered as loyal patients with a high value for the company, and thus the marketing efforts should be focused on maintaining their good relationship with the company.

6. **Premium Orthodontics:** With only 1.77% of the total number of patients, Cluster 6 is mainly characterized by patients with high spendings on Category 4 – Braces and Dental Alignment. These are patients with a high number of concluded appointments, the last of which has taken place very recently. Moreover, the high percentage of patients in Cluster 6 with a scheduled appointment (87.26%) shows that most of these patients are still undergoing treatments related to Category 4, which by nature take a relatively long period of time. Despite its low number of patients, Cluster 6 manages to have a weight of 14.79% on the company's turnover.
7. **Premium Implantology:** Cluster 7 is the smallest cluster in terms of number of patients, containing only 0.5% of the company's total number of patients. However, it has a weight of 11.38% on the company's turnover, being the cluster whose patients have the largest monetary spendings. Its patients have extremely high spendings on Category 3 – Replacement of Missing Teeth, having also relatively high spendings on Category 2 – General Treatments. Moreover, these are patients who have a relatively long relationship with the company, having a very high number of appointments, the last of which has taken place quite recently. More than half of these patients (60.77%) have a scheduled appointment.

Figure 3.10 complements the information from Table 3.12 with a visualization of the distribution of the company's patients and total turnover per cluster. It is visible that, despite the different cluster sizes, the differences in the weights of each cluster in the company's total turnover are not as pronounced.

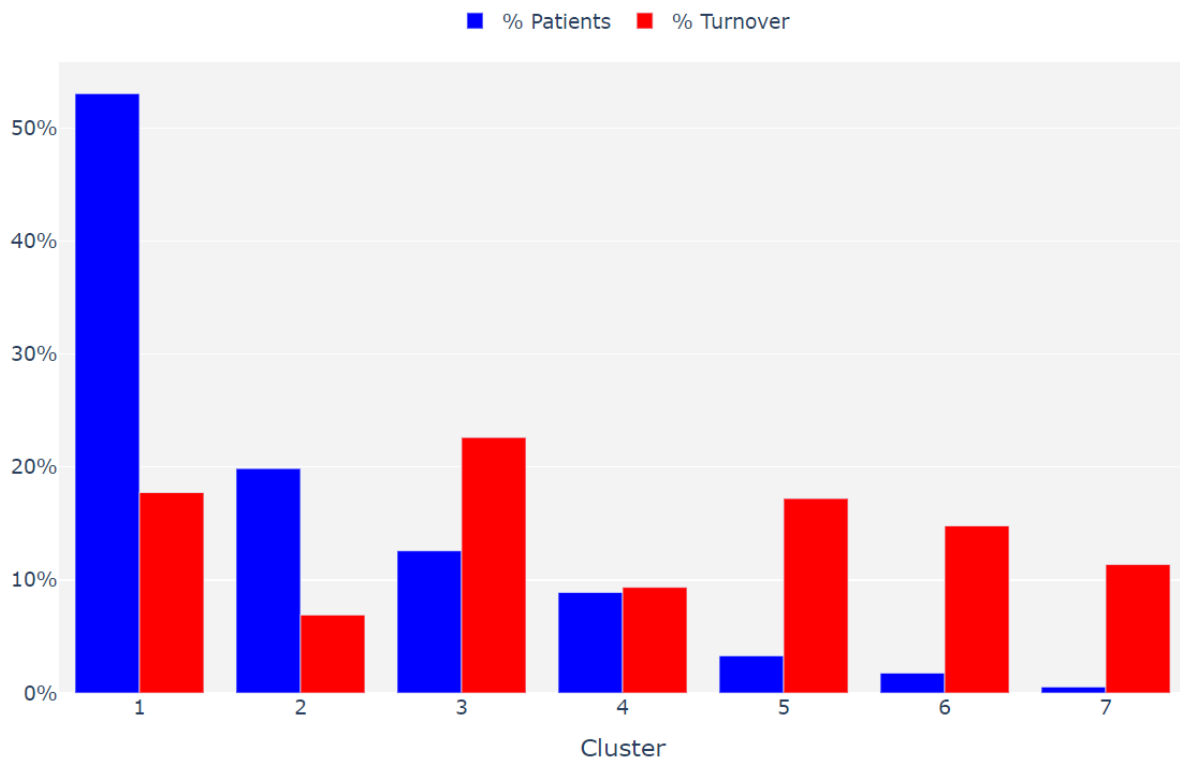


Figure 3.10 - Size and Turnover Percentages of each Cluster

Some of the main characteristics of the described clusters were already present in the clusters from the simpler *RFML* model. For instance, it already had a bigger sized cluster composed by recent patients, as well as a cluster with similar characteristics to the Defectors cluster. Nevertheless, the final model has the advantage of reflecting on which treatment categories the patients have spent monetary amounts, and thus reflects the product effect, something that the *RFML* model was not able to do.

There are also some shared characteristics with clusters from the applications of customer segmentation to the field of dental medicine found in literature. For instance, the results from Wei et al. (2012) and Wu et al. (2014) also included clusters of recent patients, and clusters whose patients had lower than average values for frequency, but already had a long-term relationship with the dental clinic and had their last appointment quite recently, similarly to the Check-up/Sporadic Patients cluster. Furthermore, the results also included clusters with characteristics similar to the Intermediate Low cluster – patients that were relatively recent to the dental clinic, but that already had a considerable number of concluded appointments. Lastly, the results also shared the two extremes in terms of desirability – clusters of lost patients, like the Defectors cluster, and clusters with loyal patients who have had frequent appointments with the dental clinic for a long time, similar to Intermediate High, Premium Orthodontics, and Premium Implantology clusters. Despite these similarities, the main difference between the results found in literature and the results from the project is that these are able to reflect the main treatment categories where each cluster spent the most. This allows understanding what kind of procedures the patients from each cluster have done with the company, and helps in assessing which marketing initiatives may make sense depending on that.

4. CONCLUSIONS

As a first note, it should be highlighted that developing a customer segmentation project is something uncommon for a dental company, where people who resort to its services are both customers and patients. The approach taken consisted mainly in equating the treatment categories to products, with the goal of having it similar to the approaches taken in other business fields, but still allowing the results to reflect that the segmentation was applied in the business field of dental medicine.

Overall, it can be considered that the project was successful. From a business standpoint, it allowed to understand how the company's patients were distributed across the several segments, as well as to have an overview on their main characteristics, which were its main goals. From a technical and scientific perspective, the results that were achieved can be considered as quality results when compared to the ones obtained in similar projects found in literature. However, the approach taken for the final model was relatively simple, consisting in applying the K-Means algorithm using 7 different features, which is something that can be discussed – if more features or another feature subset could have been used, or if other clustering algorithms would have returned significantly different results. Nevertheless, the main goal was to keep the approach simple to achieve meaningful and interpretable results, and this is something that was accomplished.

As main key points from the project's results, it can be highlighted that roughly half of the company's patients belong to the Recent Patients cluster, who have had their first appointment with the company quite recently. Retaining these patients should be a priority, since this cluster is where most of the marketing potential resides in. Identifying that almost 20% of the company's patients belonged to the Defectors cluster was also a highlight of the project. This was viewed as an opportunity to try to understand the reasons for these patients to have "defected" from the company, which could then lead to identifying whether there were services that needed improvements. Furthermore, the Check-up/Sporadic Patients cluster allowed to understand that there were loyal patients that did not necessarily have huge spending levels – naturally, their dental healthcare necessities influenced their relationship with the company. In the case of these patients, many of them just needed simple and periodical check-up treatments and kept coming to the company's clinics to perform them. The remaining clusters – Intermediate Low, Intermediate High, Premium Orthodontics, and Premium Implantology – were where most of the company's turnover was concentrated. Patients from Intermediate Low were relatively recent but already had a high number of appointments, having higher chances of becoming loyal patients. The last 3 clusters comprised of high value and loyal patients and were important to highlight the fact that a considerable proportion of the company's turnover was concentrated in a small fraction of its patients.

These key points and the project's final results were presented to its stakeholders. Some of these key points were considered as interesting and valuable information, namely the relatively high percentage of patients belonging to the Defectors cluster. Consequently, one of the marketing initiatives that resulted from the project consisted in identifying these patients and contacting them to try to understand the main reasons why they no longer resorted to the company's services, and if there were particular aspects on the company's services that could be improved.

4.1. LIMITATIONS

Overall, there were some limitations during the project although they did not have a major impact on its development. In many cases, these are aspects where there is room for improvement. It is important to highlight that, despite some data quality issues that were identified, there were no major data quality problems that deeply affected the development of the project.

As already mentioned, the external ownership of the SQL database limited the access to the data and made the feature engineering process less efficient. There were variables that were available on the SQL database that were not exported to the Excel files that were used to access the data. Exporting these variables to the Excel files needed to be requested to the external software company, which was often a lengthy and inefficient process. Moreover, it was not known if a variable that was not available on the Excel files was available on the SQL database, as these were not listed or properly documented. Having full access to the SQL database would surely enhance the efficiency of the process, and would also provide access to variables whose existence was perhaps unknown until this point, providing room for the development of new and different features.

Finally, the implementation of marketing actions for the patients from the different segments was limited to the ones that had given their marketing consent to the company. This was a minor fraction of the company's patients, which limits the targeted marketing efforts of the company, namely on marketing actions based on the segmentation results.

4.2. FUTURE WORK

Firstly, as previously mentioned, the company is still relatively recent and in a business expansion phase. Therefore, it is expected that the distribution of its patients across the different segments will evolve throughout time, as well as the clustering structure itself. It is important to monitor this evolution and assess when the results from the project are no longer meaningful, and thus clustering the data again will be necessary.

Secondly, it would be important for the company to work on making quality data on the dental healthcare necessities of the patients available. The project only used data from the patients' history with the company, resulting in clusters with clearly identifiable characteristics on this regard. However, if it could include features related to these necessities, the results could not only reflect the patients' history with the company, but also their potential in terms of what procedures they could do in the future. This could lead to clearer marketing actions for each cluster, namely the ones with more recent patients. For instance, the Recent Patients cluster could be split into multiple clusters based on the dental healthcare needs of its patients, namely recent patients with missing teeth necessities, or recent patients that need alignment treatments, among some other examples that could be provided. This would lead to much more meaningful insights that would help planning more concrete and specific marketing actions for these clusters.

Finally, K-Means and SOM were the clustering algorithms that were used in this project, and both returned similar results. However, other clustering algorithms that are not centroid-based (for instance, hierarchical or density-based ones) could be tried out to assess if these could return different yet still meaningful results, maybe giving a different perspective on the clustering structure of the data or highlighting different patterns and characteristics that the current solution did not identify.

5. BIBLIOGRAPHY

- Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5), 1785–1792. <https://doi.org/10.1016/j.jksuci.2019.12.011>
- Arunachalam, D., & Kumar, N. (2018). Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making. *Expert Systems with Applications*, 111, 11–34. <https://doi.org/10.1016/j.eswa.2018.03.007>
- Baço, F., Lobo, V., & Painho, M. (2005). Self-organizing Maps as Substitutes for K-Means Clustering. In *LNCs* (Vol. 3516).
- Capon, N. (1982). *Institutional Change in Delivery of Dental Services: A Marketing Perspective*.
- Chang, H.-C., & Tsai, H.-P. (2011). Group RFM analysis as a novel framework to discover better customer consumption behavior. *Expert Systems with Applications*, 38(12), 14499–14513. <https://doi.org/10.1016/j.eswa.2011.05.034>
- Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing and Customer Strategy Management*, 19(3), 197–208. <https://doi.org/10.1057/dbm.2012.17>
- Chen, Y.-S., Cheng, C.-H., Lai, C.-J., Hsu, C.-Y., & Syu, H.-J. (2012). Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment. *Computers in Biology and Medicine*, 42(2), 213–221. <https://doi.org/10.1016/j.combiomed.2011.11.010>
- Frigui, H., & Krishnapuram, R. (1997). CLUSTERING BY COMPETITIVE AGGLOMERATION. In *Pattern Recognition* (Vol. 30, Issue 7).
- Gao, X., Yu, J., Zha, S., Fu, S., Xue, B., Ye, P., Huang, Z., & Zhang, G. (2022). An ensemble-based outlier detection method for clustered and local outliers with differential potential spread loss. *Knowledge-Based Systems*, 258, 110003. <https://doi.org/10.1016/j.knosys.2022.110003>
- Halkidi, M. (2001). On Clustering Validation Techniques. In *Journal of Intelligent Information Systems* (Vol. 17, Issue 3).
- Han, J., Kamber, M., & Pei, J. (2012). Cluster Analysis. In *Data Mining* (pp. 443–495). Elsevier. <https://doi.org/10.1016/B978-0-12-381479-1.00010-1>
- Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, 127, 444–453. <https://doi.org/10.1016/j.jbusres.2019.05.001>
- Henriques, R., Lobo, V., & Bacao, F. (2012). Spatial Clustering Using Hierarchical SOM. In *Applications of Self-Organizing Maps*. InTech. <https://doi.org/10.5772/51159>
- Hughes, A. M. (2005). *Strategic database marketing*. McGraw-Hill Pub. Co.

- Hutt, M. D., & Speh, T. W. (2010). *Business Marketing Management: B2B* (10th Edition). Canada: Cengage Learning.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Jayachandran, S., Sharma, S., Kaufman, P., & Raman, P. (1992). The Role of Relational Information Processes and Technology Use in Customer Relationship Management. *Journal of Marketing*, 69, 177–192.
- Liu, F., & Deng, Y. (2021). Determine the Number of Unknown Targets in Open World Based on Elbow Method. *IEEE Transactions on Fuzzy Systems*, 29(5), 986–995. <https://doi.org/10.1109/TFUZZ.2020.2966182>
- MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281–297.
- Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management*, 8(1), 67–72. <https://doi.org/10.1057/palgrave.jdm.3240019>
- Olszewski, D. (2021). A data-scattering-preserving adaptive self-organizing map. *Engineering Applications of Artificial Intelligence*, 105, 104420. <https://doi.org/10.1016/j.engappai.2021.104420>
- Peppard, J. (2000). Customer Relationship Management (CRM) in financial services. *European Management Journal*, 18(3), 312–327. [https://doi.org/10.1016/S0263-2373\(00\)00013-X](https://doi.org/10.1016/S0263-2373(00)00013-X)
- Rogalewicz, M., & Sika, R. (2016). Methodologies of knowledge discovery from data and data mining methods in mechanical engineering. *Management and Production Engineering Review*, 7(4), 97–108. <https://doi.org/10.1515/mper-2016-0040>
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Sharma, D. K., Tokas, B., & Adlakha, L. (2021). Deep learning in big data and data mining. In *Trends in Deep Learning Methodologies* (pp. 37–61). Elsevier. <https://doi.org/10.1016/B978-0-12-822226-3.00002-7>
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.
- Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3), 586–600. <https://doi.org/10.1109/72.846731>
- Wei, J.-T., Lee, M.-C., Chen, H.-K., & Wu, H.-H. (2013). Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications*, 40(18), 7513–7518. <https://doi.org/10.1016/j.eswa.2013.07.053>

- Wei, J.-T., Lin, S.-Y., Weng, C.-C., & Wu, H.-H. (2012). A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Systems with Applications*, 39(5), 5529–5533. <https://doi.org/10.1016/j.eswa.2011.11.066>
- Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (2010). A review of the application of RFM model. *African Journal of Business Management*, 4(19), 4199–4206. <http://www.academicjournals.org/AJBM>
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 1, 29–39.
- Wu, H. H., Lin, S. Y., & Liu, C. W. (2014). Analyzing patients' values by applying cluster analysis and LRFM model in a pediatric dental clinic in Taiwan. *Scientific World Journal*, 2014. <https://doi.org/10.1155/2014/685495>
- Xu, Y., Yen, D. C., Lin, B., & Chou, D. C. (2002). Adopting customer relationship management technology. *Industrial Management and Data Systems*, 102(8), 442–452. <https://doi.org/10.1108/02635570210445871>



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa