FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Customer Segmentation and Distribution Network Optimization

A Case Study in the Automotive Industry

Sofia Marques Cruz

Mestrado Integrado em Engenharia e Gestão Industrial

Supervisor: Professor António Miguel da. F. F. Gomes

June 28, 2020

Customer Segmentation and Distribution Network Optimization

A Case Study in the Automotive Industry

Sofia Marques Cruz

Mestrado Integrado em Engenharia e Gestão Industrial

June 28, 2020

Abstract

Amidst an ever-growing globalization, where the pressure among competitors is building up, and customers are getting increasingly more demanding, businesses must keep up with these developments, or they risk losing their edge and their customers' interest.

On the one hand, companies must strive to leverage on the massive amount of data collected to understand their customers' interests and needs. A proper customer segmentation based on customer's buying patterns will allow a better understanding on how valuable they are to the company and adequately defining the service level required.

On the other hand, companies involved in the distribution of goods face the increased challenge of designing a Distribution Network that not only allows the satisfaction of all customers but also minimizes the company's transportation cost. Being capable of doing so may determine whether or not a company survives.

This study proposes the usage of an analytical customer segmentation, based on Customer Lifetime Value to segment customers. Given the complexity of logistics operations of the Automotive Aftermarket sector, considering consumer behaviour variables alone is not enough to properly design a company's service model. To acknowledge the real-life difficulties that exist in delivering to different geographic locations this project suggests the application of a density-based clustering algorithm to cluster customers based on their geographical location. Once the two techniques are applied, a proper Service Model can be defined.

Lastly, a mathematical model is presented that aims at designing a Delivery Network that not only allows satisfying every customer's need, but also minimizes transportation costs.

Keywords: Segmentation, Clustering, K-Means, DBSCAN, Distribution Network, Optimization, Integer Linear Programming

Resumo

Num contexto de crescente globalização, onde a pressão entre concorrentes aumenta e os clientes estão cada vez mais exigentes, as empresas devem acompanhar os desenvolvimentos ou correm o risco de ficar para trás e perder o interesse dos consumidores.

Por um lado, as empresas devem ser capazes de aproveitar a enorme quantidade de dados que todos os dias recolhem, para entender os interesses e necessidades dos seus clientes. Uma segmentação adequada dos seus clientes, com base em padrões de compra, permitirá entender o valor que estes representam para a empresa e, assim, definir o nível de serviço adequado a cada um deles.

Por outro lado, as empresas envolvidas em atividades de distribuição enfrentam o desafio acrescido de desenhar uma Rede de Distribuição que não só permita satisfazer todos os clientes, como minimizar o custo de transportes. A sobrevivência de uma empresa pode estar dependente da sua capacidade de o conseguir fazer.

Este trabalho propõe o uso técnicas analíticas para segmentar os clientes, com base no seu valor de vida útil (*Customer Lifetime Value*). Dada a complexidade das operações de logística do setor de *Aftermarket* Automóvel, considerar apenas as variáveis comportamentais não é suficiente para corretamente desenhar o modelo de serviço destas empresas. Assim, para considerar o esforço inerente à entrega de produtos em diversas localizações geográficas, propõem-se o uso de um algoritmo de agrupamento baseado em densidade para agrupar clientes pela sua localização geográfica. Uma vez aplicadas as duas técnicas, o Modelo de Serviço poderá ser corretamente definido.

Por fim, é apresentado um modelo de otimização matemático que visa estabelecer uma Rede de Distribuição que consiga não só satisfazer as necessidades dos clientes como minimizar o custo de transportes.

Palavras-chave: Segmentação, Cluster, K-Means, DBSCAN, Rede de Distribuição, Otimização, Programação linear inteira

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my colleagues, Rafael Henriques and Tiago Abreu, for always challenging and helping me throughout this project's development. To Joel Queirós and the rest of Kaizen Institute for the support and encouragement.

I am very thankful to my supervisor, Professor Miguel Gomes, for the help and availability throughout these past months.

To one of my biggest friends, Adriana, who accompanied me throughout this journey. Without you these past five years would not have been the same. To my friends Beatriz, Joana, Margarida and Marta, my sincere thanks for the many years of laughter and friendship.

To my caring and supportive parents, Cristina and José Miguel, my deepest gratitude. Thank you for all the sacrifices, for preparing me for the future and encouraging me to always follow my dreams.

Finally, to my sisters, Marta and Francisca, for the unconditional love and support.

Sofia Marques Cruz

"If you work hard enough and assert yourself, and use your mind and imagination, you can shape the world to your desires."

Malcolm Gladwell

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Objectives . . . 2
- 1.3 Methodology . . . 3
- 1.4 Dissertation Structure . . . 3

2 Literature Review **4**
- 2.1 Service Model . . . 4
  - 2.1.1 Customer Segmentation . . . 5
  - 2.1.2 Geographical Clustering . . . 12
- 2.2 Distribution Network . . . 13
  - 2.2.1 Vehicle Routing Problem . . . 13
  - 2.2.2 Stochastic Vehicle Routing Problem . . . 15

3 Automotive Industry's Case Study **16**
- 3.1 Company Presentation . . . 16
- 3.2 Delivery System Analysis . . . 17
- 3.3 Customer Segmentation . . . 20
  - 3.3.1 What impacts customer satisfaction? . . . 21
- 3.4 Historical Data Analysis . . . 21
  - 3.4.1 Which Warehouse is serving which client? . . . 22
  - 3.4.2 Which system is more profitable? . . . 23
  - 3.4.3 Do Dedicated customers use all the offered deliveries? . . . 26
- 3.5 Conclusion . . . 27

4 Service Model **29**
- 4.1 Proposed Approach . . . 29
  - 4.1.1 Business Understanding . . . 30
  - 4.1.2 Data Understanding & Preparation . . . 30
  - 4.1.3 Modelling . . . 32
  - 4.1.4 Evaluation & Deployment . . . 41
- 4.2 Conclusion . . . 44

5 Distribution Network **45**
- 5.1 Context . . . 45
- 5.2 Model . . . 46
  - 5.2.1 Formulation . . . 47
  - 5.2.2 Solution Approach . . . 50

5.3 Analysis & Results 51
5.4 Conclusion 52

6 Conclusions and Future Work **55**

References **57**

Appendix **62**

List of Figures

2.1 Clustering example, adapted from Tan et al. (2013) . . . 8

3.1 Delivery Systems representation . . . 20
3.2 Pareto Analysis of the Causes for Service Level Complaints . . . 21
3.3 Real Warehouse Distribution . . . 22
3.4 Dedicated vehicles' volume utilization . . . 25
3.5 Average number of destinations per route . . . 26
3.6 Average time per route & % of routes with < 1h, 1h-1h45 and > 1h45 . . . 26
3.7 Average customer usage of the Dedicated system . . . 27

4.1 Proposed Approach, based on Shearer (2000) and adapted from Carneiro and Miguéis (2020) . . . 30
4.2 Boxplot of the RFM variables . . . 33
4.3 Matrix of correlation of the RFM variables . . . 34
4.4 Elbow Curve . . . 35
4.5 3D Clusters Representation . . . 36
4.6 Clustering Decision Tree . . . 37
4.7 Elbow Curve . . . 39
4.8 3D Sub-Clusters Representation . . . 39
4.9 Sub-clustering Decision Tree . . . 40
4.10 k-NN distances plot . . . 42
4.11 DBSCAN Clustering Results . . . 43

5.1 Grouping Representation . . . 46
5.2 Histogram of the number of clients that order per day . . . 47
5.3 Optimization Model solution for Setúbal . . . 53
5.4 Optimization Model solution for Porto . . . 54
5.5 Optimization Model solution for Lisbon . . . 54

List of Tables

2.1 Examples of application of K-Means algorithm in customer clustering 9
2.2 Examples of application of the VRP 14
2.3 Examples of application of the SVRP 15

3.1 Cutoffs available for route R 18
3.2 Logistic Partners' Competence Matrix 19
3.3 Current customer segmentation 20
3.4 Causes for wrong deliveries 23
3.5 Summary of the Delivery Systems (June 2019 - February 2020) 24

4.1 Sample of the original transaction dataset 31
4.2 Sample of the obtained dataset 32
4.3 Attributes in the obtained dataset 32
4.4 Summary of the variables 32
4.5 Summary of the normalized variables 34
4.6 Results of the Average Silhouette and Davies-Bouldin Methods 35
4.7 Sum of Square by Cluster 35
4.8 Normalized RFM Variables per Cluster 36
4.9 AHP scale, adapted from Liu and Shih (2005a) 37
4.10 AHP questionnaire for the pairwise comparison 38
4.11 CLV Ranking 38
4.12 Sales Amount per cluster 38
4.13 Normalized RFM Variables per Sub-Cluster 39
4.14 Customers Buying Behaviour per Sub-Cluster and CLV Ranking 40
4.15 Customer Segments Summary 41
4.16 Customer Segments Summary 41
4.17 Classification scale in terms of delivery effort 43
4.18 Future Service Level for each Customer Segment, given the required Delivery Effort 44

5.1 Short term Maximum # Deliveries per Customer Segment 51
5.2 Short term Service Level for each Customer Segment, given City's Delivery Effort 51
5.3 Number of Customers and Customer's Groups with 3 Deliveries, per District . . 52
5.4 Optimization Model Results 53

1 List of cities eligible for 3 deliveries per day 62

List of Algorithms

1 K means algorithm, adapted from Tan et al. (2013) 10
2 Elbow Method, adapted from Carneiro and Miguéis (2020) 11
3 Average Silhouette Method 11
4 Auxiliary Routing Algorithm 25

Abbreviations

CLV	Customer Lifetime Value
CRM	Customer Relationship Management
IAM	Independent Aftermarket
ILP	Integer Linear Programming
KDD	Knowledge Discovery in Databases
LP	Linear Programming
OEM	Original Equipment Manufacturer
RFM	Recency, Frequency and Monetary
SKU	Stock Keeping Unit
SSE	Sum Square Errors
VRP	Vehicle Routing Problem
VSD	Value Stream Design

Chapter 1

Introduction

The Automotive Industry includes all the companies that are involved in designing, manufacturing, selling and repairing vehicles. It not only is one of the most significant economic forces in the world, but it largely impacts today's social life and habits. According to the European Commission, this industry represents 6.1 percent of global EU employment, providing direct and indirect jobs to more than 13 million Europeans.

The Automotive Aftermarket sector concerns the manufacturing, distribution, retailing and installation of all vehicle parts, chemicals, equipment, and accessories, after the automobile's sale. Automobile parts may be Original Equipment Manufacturer (OEM) or Independent aftermarket (IAM) and belong to one of the two following categories (U.S. Department of Commerce, 2009): replacement parts - that aim to replace OEM parts when they become worn or damaged - or accessories that are designed for add-on after the original sale of the vehicle.

1.1 Motivation

In an increasingly globalized world, where competitors pressure is exponential, and customers demand more with every second, businesses must strive to achieve an efficient distribution network, than can both be cost-efficient and meet the ever-changing needs of customers. The Automotive Industry is no exception.

In the Aftermarket sector, in particular, this has become crucial, as fundamental changes to vehicle technology, customer behavior and industry dynamics occur. Therefore, businesses who do not embrace change will lose their competitive edge and fail to meet their customers needs.

Over the last years, the automobile industry has faced many changes. In the past, economies of scale were achieved through the standardization of components that allowed mass production. Today, given the industrial and technological breakthroughs, both in machinery, automation and operations, standardization is not a requirement, and it is possible to achieve economies of scale

even with diversity and highly customized products. Given these changes in the business environment, car parts manufacturers continually increase their products references in an attempt to sell the car that best suits their customers' preferences.

Moreover, when compared to a typical manufacturing supply chain, the aftermarket inherently faces a higher uncertain environment. In addition to managing a wider range of stock-keeping units (SKUs) and taking care of all the reverse logistics (return, repair and disposal of failed components), demand is far more unpredictable and challenging, requiring a response in one day or even hours (Cohen et al., 2006)). When an automobile parts dealer makes a request to its suppliers, the right component is expected to be delivered at the right place, time and cost. It is important to keep in mind that at the end of the automotive aftermarket supply chain cycle is a customer with a vehicle waiting for a repair part (Robinson, n.a.).

In addition to this, the geographical dispersion of the several delivery points and the smaller dimension of the orders have resulted in a significant increase in the complexity of logistics operations.

Having said so, it is vital that companies operating in this sector revise their Service Model, aiming to achieve a sustainable model that allows gaining competitive advantages in conquering new market share through leveraging the service level as a differentiation strategy.

Today, the massive amount of data that companies collect daily can and should be used to identify and figure out what do their customers want. Hence the increasing popularity of data mining techniques in Customer Relationship Management (CRM) (Carneiro and Miguéis, 2020). A proper customer categorization based on customer's patterns and behaviours will allow offering them solutions that exceed their expectations. Nevertheless, most companies find it hard to understand their customers' needs.

This study proposes the usage of analytical customer segmentation based on Customer Lifetime Value to segment customers. Given the complexity of logistics operations of the Automotive Aftermarket sector, considering consumer behaviour variables alone is not enough to properly design a company's service model. To acknowledge the effort that exists in delivering to different geographic locations, this project suggests the application of a density-based clustering algorithm to cluster customers based on their geographical location. Once the two techniques are applied, a proper Service Model can be defined. Lastly, a mathematical model is presented that aims at designing a delivery network that not only allows satisfying every customer's need, but also minimizes transportation costs.

1.2 Objectives

This dissertation was carried out as part of a Kaizen Institute project, in a client Company[1]. Hence, this project aims at addressing the following three main goals of the Company:

- Redesign the Service Model in a profitable and sustainable way, oriented towards conquering of new market share:

[1]The name of the company will not be revealed for confidentiality reasons

  - Increase customer satisfaction and loyalty
  - Redefine the number of deliveries and Lead Time per customer
- Optimize the Distribution Network
  - Calculate the number of vehicles required to satisfy the demand
  - Define the allocation of customers to the vehicles

1.3 Methodology

The methodology adopted was divided into four major stages.

Initially, a Preparation phase was undertaken to collect reliable data that support the modelling of the current state of the Service Model and Distribution Network of the company. Both quantitative and qualitative data was gathered, and furthered discussed.

Secondly, the data was analyzed to understand the Company's initial situation and to have an integrated view of its macro processes and interactions between players in the supply chain. Besides assessing the current state of the processes it is critical to evaluate how these could be improved to increase the operations' quality and efficiency.

Having analyzed how the organization works, the Company's Service Model was redefined. Subsequently, the Distribution Network was redesigned to match the new Service Model and minimize transportation costs.

1.4 Dissertation Structure

The structure of the thesis is as follows. This chapter describes the motivation and aim of the present project, as well as the methodology adopted.

Chapter 2, Literature Review, introduces the studied literature on two topics of interest for the project: Customer segmentation and related data mining concepts and techniques; and Vehicle Routing Problem.

Chapter 3 presents the Case Study. In particular, the current situation of the Company where the project was developed is described, providing an extensive analysis of its critical processes and the results of the analysis carried out.

Chapter 4 and 5 explain the process of reshaping the Service Model and Distribution Network.

Lastly, chapter 6, Conclusions and Future Work, summarizes the content of this dissertation, as well as suggesting future work developments.

Chapter 2

Literature Review

This study tackles two different research topics: Service Model Redesign and Distribution Network Optimization, applied in the specific context of an Automotive Aftermarket player.

In this chapter the theoretical concepts behind the project's development are introduced. Section 2.1 introduces all the relevant topics to properly define the Company's Service Model. Section 2.2 presents topics of interest for the creation of an optimization model for the Company's Distribution Network.

2.1 Service Model

As previously explained, to properly design a company's service model, companies must be able to leverage on the massive amount of data to properly understand their customers' worth. This study proposes the usage of clustering techniques to create customer segments based on Customer Lifetime Value (CLV). Being able to do so will provide relevant information for developing new Customer Relationship Management (CRM) strategies. Subsection 2.1.1 explains the concepts behind CRM and CLV, as well as all the models and data mining techniques used to segment customers.

Nevertheless, to properly design the Company's Service Model, it is vital to account for the complexity of logistics operations of the Automotive Aftermarket sector and consider the real-life difficulties that exist in delivering to multiple geographic locations. Subsection 2.1.2 presents an overview of a few density-based clustering algorithms that could be used to take the customers' location into consideration.

2.1.1 Customer Segmentation

As Payne and Frow (2005) describes it,

> "CRM is a strategic approach that is concerned with creating improved shareholder value through the development of appropriate relationships with key customers and customer segments. (...) CRM provides enhanced opportunities to use data and information to both understand customers and concrete value with them. This requires a cross-functional integration of processes, people, operations, and marketing capabilities that is enabled through information, technology, and applications."

CRM focuses on applying data mining techniques in the data collected and kept in data warehouses, to analyze customers characteristics and behaviours. The goal is to strengthen the relationship with customers, thus improving customer acquisition, retention, loyalty and profitability.

Kotler and Armstrong (2006) defend that, whilst acquiring customers is very important, losing a customer means losing its Customer Lifetime Value (CLV), that is, "the value of an entire stream of purchases that the customer would make over a lifetime of patronage". Hwang et al. (2004) agree, stating that the lack of information on new customers makes customer acquisition a harder task. In this paper, Hwang et al. (2004) present several definitions for CLV or Lifetime Value (LTV), as they call it. For the authors, the main goal of this measure is to estimate the impact that each customer has and therefore the effort that the company should place on each other's relationship.

As companies grow, it becomes unbearable to study every customer individually (Maskan, 2014). When that happens, segmentation processes can be used to group customers with similar characteristics. Customers can be segmented using numerous criteria, which can be demographic, psychographic, geographic and behavioural. Having said so, the success of CRM is extremely dependent on the correct customer segmentation and respective CLV evaluation (Hwang et al., 2004).

One of the most well-known models used to calculate CLV is RFM. The RFM analytical model, proposed by Hughes (1996), differentiates customers based on three variables associated with customer behaviour: Recency, Frequency and Monetary. As Bult and Wansbeek (1995) explain,

Recency is the time period since the last purchase. The lower this time interval, the higher the R part;

Frequency is the number of purchases within a certain time period. The higher the frequency, the higher the value of F part;

Monetary is the amount of money spent during a certain period. The higher the monetary value, the higher the value of M part.

In the traditional RFM approach, equal weights are allocated to the three variables. However, different weights can be given. Each organization has its own specifications, and so deciding their

own priority of each one of the RFM variables allows to obtain results that are tailored to their needs. Several studies have proposed the use of Weighted RFM models (WRFM) to compute the CLV of each cluster. WRFM has been used in several case studies: Maskan (2014) used it in an internet service provider company, Carneiro and Miguéis (2020) in a food industry company, Liu and Shih (2005b) in a hardware retail company and Khajvand et al. (2011) in a health and beauty company.

With this model, the CLV of each cluster can be calculated, as seen in equation 2.1.

$$CLV^i = w_R \cdot C_R^i + w_F \cdot C_F^i + w_M \cdot C_M^i \tag{2.1}$$

where:

- w_R, w_F, w_M are the relative importance of each variable (adapted to the company's needs)
- ${C_R}^i$, ${C_F}^i$, ${C_M}^i$ are the normalized values obtained for each cluster i

One of the most popular techniques used to assess the relative weights is the Analytical Hierarchy Process (AHP), introduced by Saaty (2000). In his book "Fundamentals of Decision Making and Priority Theory", he explains that AHP is based on the natural human capability to access the relative weight of several factors through pairwise comparisons, based on intuition, judgment and experience.

Over time, many studies have started to introduce other variables to the RFM model. Among other variables, there are studies incorporating the length of the relationship with the customer in the RFM model (LRFM), first introduced by Wei et al. (2012). Since then, Li et al. (2011) used LRFM to cluster the customers of a textile company and Kao et al. (2013) in an outfitter company, for example. The customer geographic location is also quite used. Carneiro and Miguéis (2020) used a geographic segmentation followed by the RFM segmentation model to cluster customers of a food industry company.

2.1.1.1 Data Mining and Its Techniques

So what exactly is data mining? Data mining is the process of applying data analysis techniques to large data sets, in order to find anomalies, patterns and correlations that will be used to predict outcomes.

Typically, data mining comprises three main phases: pre-processing, data task application and post-processing. Preprocessing comprises all functions related to data capturing, organizing and preparation. This phase aims at selecting, cleaning and preparing the data before it is used in the next phase. Post-processing relates to the treatment and re-arrangement of the knowledge obtained from data mining, making its interpretation easier (Goldschmidt and Passos, 2005).

Data Preprocessing

Whilst this step has always been important, nowadays it places a vital role in data quality as it never has before. In fact, the huge size and complexity of today's databases make it much more susceptible to incomplete, incompatible and noisy data collection. As Han et al. (2011) explain, "Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining". This book describes the concepts and methodologies for the application of the various preprocessing techniques.

Data Cleaning involves rectifying data quality problems, including filling in missing data, smoothing noisy data, detecting and correcting inconsistencies.

Data Integration consists of combining data from multiple sources into a coherent data store, in order to give the user a unified view of these data.

Data Reduction aims at obtaining a reduced representation of the dataset that only contains data that impacts the analysis. It may involve dimension reduction, feature aggregation and/or elimination or clustering.

Data Transformation aims at converting and consolidating data into appropriate forms for mining. Among others, it may involve building new attributes from old ones, data smoothing, data normalization and data discretization.

Data Discretization is the process of transforming continuous numeric data into discrete variables with a small number of values (intervals or labels).

Task Application

Amongst several tasks identified by Goldschmidt and Passos (2005), Association, Classification and Clustering are the most widely used. Depending on the task, the techniques and algorithms to be used are decided.

Association tasks aim at analysing items that frequently occur simultaneously and finding relations between those attributes, with a certain level of confidence.

Classification tasks are meant to classify an item according to predefined categorical labels, in order to identify to which category an object belongs (Saraiva, 2018).

Clustering aims at organizing items into unknown groups. As Saraiva (2018) highlights, these classes of objects - clusters - are attained from similarity metrics or probability density models.

In this specific CRM context, clustering techniques will allow grouping together customers with similar buying patterns and separate them if otherwise. As Nimbalkar and Shah (2013) outlines, the clusters obtained should maximize intra-group and minimize inter-group similarity. To better understand this concept, consider figure 2.1, that shows two different ways of attributing 12 customers to clusters.

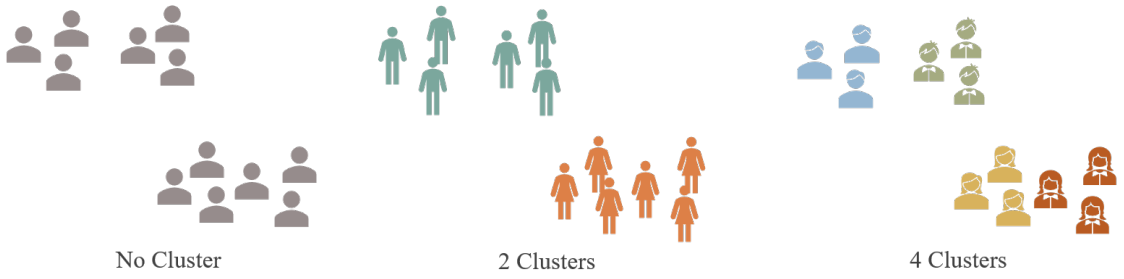

Figure 2.1: Clustering example, adapted from Tan et al. (2013)

There are numerous clustering techniques. Tan et al. (2013) explain the various types of clustering:

- Hierarchical versus Partitional: In partitional (unnested) methods, given a certain criterion, data is divided into k non-overlapping subsets (clusters). Each object belongs to only one of these clusters. On the contrary, in hierarchical (nested) methods, each object belong to one cluster and several sub-clusters. Basically, it works like a series of partitional clusterings where, in each iteration, the clusters created are divided into sub-clusters.
- Exclusive versus Overlapping versus Fuzzy: In exclusive methods, each object belongs to a single cluster, while in overlapping methods they may belong simultaneously to various groups. In fuzzy clustering, each item belongs to all groups with a certain membership weight that varies between 0 (absolutely does not belong) and 100% (absolutely belongs).
- Complete versus Partial: whilst complete clustering assigns every item to a cluster, partial considers that not all objects belong to a well-defined group.

2.1.1.2 Techniques Used

Assessing Clustering Tendency

Prior to applying any clustering technique, it should be analyzed whether the dataset has a non-random structure (Tan et al., 2013). The reason why this must be done is that the clustering algorithm may return clusters even for random structure dataset (a set of uniformly distributed points, for example). In cases like these, despite returning clusters, they are random and meaningless.

The Hopkins statistic (Lawson and Jurs, 1990) is a spacial statistic that assesses the spacial randomness of the data. It is calculated as seen in equation 2.2.

$$H = \frac{\sum_{i=1}^{n} dist(q_i)}{\sum_{i=1}^{n} dist(p_i) + \sum_{i=1}^{n} dist(q_i)} \tag{2.2}$$

where:

- D is the data set
- *dist(p_i)* is the distance between n sample points p_i from D and their nearest real neighbours in D
- *dist(q_i)* is the distance between n generated points q_i uniformly distributed in the space of D and their nearest real neighbours in D

Basically, if D is uniformly distributed, then $\sum_{i=1}^{n} dist(p_i) \approx \sum_{i=1}^{n} dist(q_i)$ and H would be about 0.5. On the other hand, if there is a clustering tendency, $\sum_{i=1}^{n} dist(p_i)$ would be much higher than $\sum_{i=1}^{n} dist(q_i)$ and H would tend to 0.

Having said so, the null hypothesis is that D is uniformly distributed (H = 0.5). We reject the alternative hypothesis (D is not uniformly distributed, containing meaningful clusters) if H > 0.5.

Clustering Techniques

There are numerous clustering techniques. Given the desired output of this analysis, partitional clustering methods are the ones to be considered.

One of the most popular algorithms is K-Means, a partitional clustering technique that separates n items in K user-specified number of clusters (Tan et al., 2013). As a prototype based technique, it creates a single level partitioning of the objects. This method is one of the oldest and most widely-used clustering algorithms. Table 2.1 shows examples of case studies that have used K-Means to cluster customers and estimate their CLV.

Table 2.1: Examples of application of K-Means algorithm in customer clustering

Case Study	Reference
Food Industry Company	Carneiro and Miguéis (2020)
Hardware Retailer	Shih and Liu (2003)
Internet Service Provider	Maskan (2014)
Health and Beauty Company	Khajvand et al. (2011)
Online Retail Industry	Chen et al. (2012)
Iranian Private Bank	Khajvand and Tarokh (2011)
Office Supplements Products Seller	Taher et al. (2016)

K-Means algorithm presents a few drawbacks (Tan et al., 2013): it is only appropriate for quantitative data; it is affected by outliers (reason why outlier detection and removal is advised); it

cannot handle non-globular clusters or that have different sizes and densities; it can only be used for data where the notion of a centre (centroid) makes sense.

On the other hand, in addition to its intuitiveness, simple implementation and wide applicability (Tan et al., 2013), the main benefit of K-Means algorithm is its efficiency in clustering large data sets (Huang (1998)).

Alternatively, there are other partitional methods that could be considered. Han et al. (2011) explain a few of them:

- K-Median: As stated earlier, K-Means is extremely sensitive to outliers because the mean value of a cluster will take into consideration the items that are far away from the majority of the data. K-Median is another variation of K-Means, that aims to tackle its sensitivity to outliers, by using the median instead of the mean.

- K-Medoids: Like the previous algorithm, K-Medoids overcomes the outlier sensitivity of K-Means, by using an actual object of the cluster to represent it. Each remaining item is assigned to the clusters based on the similarity to the representative item of the cluster. Despite this advantage, this algorithm is much more complex and therefore, time-consuming and costly.

For the purpose of this dissertation, given its wide usage in similar studies and advantages, K-Means algorithm will be used. The rationale behind the algorithm is explained by Tan et al. (2013). After specifying K number of clusters, K initial centroids are chosen. Each cluster is created by attributing each one of the objects in the dataset to the closest centroid. For each one of these clusters, the centroid will be re-calculated and the objects re-assigned to the new closest centroid. This is an iterative process that only stops when the centroids remain the same.

Algorithm 1 K means algorithm, adapted from Tan et al. (2013)

```
1: Select K points as initial centroids
2: repeat
3:     Form K clusters by allocating each data point to the cluster whose centroid is the closest
4:     Recompute the centroid of each cluster
5: until data points do not change
```

Variables Normalization

As the variables used in the RFM model have different ranges, it is required to proceed to their normalization so that they all have the same weight in the final model. To do so, the min-max normalization is one of the most used methods. According to Han et al. (2011), considering that min_X and max_X are the minimum and maximum values of an attribute X, then v' (the new value of v) is calculated in equation 2.3, such that, in this case, $newmin_X$ is 0 and $newmax_X$ is 1. The results can be seen in table 4.5.

$$v' = \frac{v - min_X}{max_X - min_X} \cdot (newmax_X - newmin_X) + newmin_X \tag{2.3}$$

Defining K

There are several methods to estimate the number of clusters, K. One of the most popular ones is the **Elbow Method**. This method looks at the percentage of variance explained as a function of the number of clusters (Bholowalia and Kumar, 2014). The idea is to analyze the percentage of additional variance explained by adding one extra cluster. To discover K, we must plot the percentage of variance explained against the number of clusters and find the "elbow", that is, the point where the marginal gain is no longer significant. Algorithm 2 details the process.

Algorithm 2 Elbow Method, adapted from Carneiro and Miguéis (2020)

1: Apply the chosen clustering algorithm to different values of K
2: For each K, compute the within-cluster sum of square errors (WSS)
3: Plot the WSS curve
4: The number of clusters K should correspond to a curve in the graph (elbow)

Sometimes is not possible to get a clear elbow. When this happens, there are other solutions. The **Average Silhouette Method** (Kaufman and Rousseeuw, 2009) considers that the right K is the one that maximises the average silhouette over several possible values for K. The silhouette evaluates how similar an object is to the others in its cluster compared to the other clusters.

Algorithm 3 Average Silhouette Method

1: Apply the chosen clustering algorithm to different values of K
2: For each K, calculate the mean observations, s(i)
3: Plot the curve
4: The number of clusters K corresponds to the maximum value

The method is defined by the following equation:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.4}$$

where:

- *a(i)* represents the compactness of the cluster to which *i* belongs (measure of cohesion) - the smaller, the more compact
- *b(i)* indicates how much *i* is separated from the other clusters (measure of separation) - the larger, the more separate
- *s(i)* ranges from -1 to 1. A value close to 1 requires $a(i) \ll b(i)$, meaning that data is appropriately clustered.

Alternatively, **Davies-Bouldin Method** could be used. Consider S_i the within cluster scatter for cluster i and $M_{i,j}$ the separation between clusters i and j. The Davies-Bouldin index is computed as seen in equation 2.5. Essentially, the lower the DB value, the better the clustering.

$$DB = \frac{1}{N}\sum_{i=1}^{N} D_i \tag{2.5}$$

where:

$$Di = \max_{i \neq j} R_{i,j} \quad \text{and} \quad R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \qquad \forall i \in N,\ j \in N \tag{2.6}$$

Evaluating the Clustering Results

An important step in the segmentation process is to evaluate the quality of the clustering results. These results can be evaluated by comparing inter-cluster homogeneity with intra-cluster heterogeneity.

Among other measures, the Sum of Squared Error (SSE) can be used to evaluate how compact each cluster is. As Han et al. (2011) explain, the quality of a certain cluster C_i can be measured by SSE between its centroid c_i and each one of the other objects, p, of that cluster. The smaller the cluster SSE, the higher the quality. Having said so, the SSE for all objects in a dataset is computed as seen in equation 2.7, where *dist* is the Euclidean distance.

$$SSE = \sum_{i=1}^{K} \sum_{p \in C_i} dist^2(p, c_i) \tag{2.7}$$

2.1.2 Geographical Clustering

The clustering algorithms discussed so far (K-Means, K-Medians and K-Medoids) are based on the idea that clusters should be formed in such a way that objects in the same cluster should be similar to each other and distinct to objects belonging to other clusters. By doing so, it is assumed that a similarity measure (distance) can be computed.

Density-based clustering methods create clusters assuming that these are dense regions in the data space, separated by areas of low point density. As Hahsler et al. (2019) highlights, these methods are able to discover clusters of non-globular shape, something that partitioning algorithms can not do. Additionally, density-based clustering methods are able to identify outliers in low density regions. These properties are quite advantageous for several problems, such as trajectory data processing (Chen et al., 2014).

Density-Based Clustering Based on Connected Regions with High Density (DBSCAN)

Tan et al. (2013) explain a few existing density-based algorithms. One of the most popular density-based algorithms is DBSCAN (Density-Based Clustering Based on Connected Regions

with High Density), proposed by Ester et al. (1996). This algorithm finds core objects (that have dense neighbourhoods) and links them to their neighbours to form dense regions as clusters. To do so, two parameters must be specified: the search radius (Eps [1]) and the minimum number of points (MinPts) within a circular area defined by the radius. An object is a core object if it contains at least MinPts items in this Eps neighborhood.

Fan et al. (2019) used DBSCAN algorithm to "identify consumer clusters in cities and more effectively analyze the structure of urban consumer space".

Other Density-Based Algorithms

To overcome the difficulty in using one set of global parameters (Eps and MinPts) a cluster analysis method called OPTICS (Ordering Points to Identify the Clustering Structure) was proposed (Ankerst et al., 1999). Instead of producing explicit clusters, this algorithm outputs a cluster ordering, that can later be used to extract clustering information.

Another density-based clustering algorithms is DENCLUE (Clustering Based on Density Distribution Functions), proposed by Hinneburg and Keim (1998). While in DBSCAN and OPTICS density is defined through the number of objects in a neighborhood of a given radius, DENCLUE uses a Gaussian kernel to estimate density based on the given set of objects to be clustered. As explained by Tan et al. (2013), the idea behind kernel density estimation is that each observed object is an indicator of high-probability density. The probability density depends on the distances from this point to the other observed objects.

2.2 Distribution Network

Logistics' distribution costs represent a substantial cost structure component of companies engaged in the delivery of goods. For these companies, finding an efficient distribution network is crucial as it will highly impact their profitability. Thus, the second part of this study, aims at developing a model that allows companies to define the number of vehicles required to satisfy all customers' orders.

One of the most commonly used solution approaches for these problems is the Vehicle Routing Problem (VRP). The following section presents a non-exhaustive overview of the VRP. In fact, in the Automotive Aftermarket customers do not order everyday. Thus, the delivery locations visited vary from day to day, leading to one of the VRP's variants: VRP with Stochastic Customers (SCVRP). For that reason, this variant will be further explained in subsection 2.2.2.

2.2.1 Vehicle Routing Problem

Firstly introduced by Dantzig and Ramser (1959) as the "truck dispatching problem", the Classical Vehicle Routing Problem (CVRP) is one of the most popular combinatorial optimization

[1] *epsilon*-neighborhood of x

problems (Barnhart and Laporte, 2006). It aims at finding the optimal set of routes to serve a set of customers, given a fleet of homogeneous vehicles. Each customer is visited once and the routes start and end at the depot. In "An overview of vehicle routing problems", Toth and Vigo (2002) present an extensive review of the VRP.

Given its wide applicability, the popularity of the VRP keeps growing. In fact, Eksioglu et al. (2009) showed that, between 1959 and 2008, 1021 journal articles had VRP as their main topic.

Since Dantzig and Ramser (1959)'s article, several variants have been introduced (Barnhart and Laporte, 2006). The VRP with Time Windows (VRPTW) considers that each customer presents a fixed schedule for receiving the goods; The Inventory Routing Problem (IRP) manages routing decisions and inventory control together; The Stochastic Vehicle Routing Problem (SVRP) considers that some of the VRP's components are random, instead of deterministic.

Table 2.2 provides a few examples of the different applications of the various VRP variants to real-life problems, over the years.

Table 2.2: Examples of application of the VRP

Case Study	Reference
Vehicle scheduling and routing compliant with the regulations for drivers' working hours	Goel (2009)
Transport of end-of-life consumer electronic goods	Kim et al. (2009)
Waste collection	Benjamin and Beasley (2010)
Incorporate cross docking in vehicle routing	Liao et al. (2010)
'Environmental-friendly' vehicle routing	Bektas and Laporte (2011)
Distribution of highly perishable food products	Amorim and Almada-Lobo (2014)

There are several methods for solving a VRP, from exact algorithms to approximate solution methods. As Maini and Goel (2017) explain, exact algorithms yield the optimal solution, while heuristics and meta-heuristics provide solutions that are close to the optimal.

There are numerous exact algorithms. As highlighted by Maini and Goel (2017), some of the most popular ones are Branch and X (Branch and Bound algorithm (Lawler and Wood, 1966), Branch and Cut algorithm (Gomory, 1958), Branch and Price algorithm (Barnhart et al., 1998)), Dynamic Programming (Bellman, 1954) and Linear Programming.

Despite guaranteeing the optimality of the solution, exact algorithms are time expensive (Jourdan et al., 2009). Therefore, they are not considered appropriate for real-life VRPs, given its complexity. For problems where finding the optimal solution is impracticable, heuristics and meta-heuristics can be used. Heuristics are used to find a satisfactory solution within a certain time limit. Meta-heuristics can be used to guide the heuristics and keep them from being trapped in a local optimum. To solve VRP, Genetic algorithms, Simulated Annealing, Ant Colony and Tabu Search are some of the most common meta-heuristics used (Maini and Goel, 2017).

2.2.2 Stochastic Vehicle Routing Problem

SVRP arises when some of the problem's components are random. Aguilella et al. (1996) explain the most common variations:

- VRP with Stochastic Customers: Customers have deterministic demands and present a certain probability of being present;
- VRP with Stochastic Demand: Customer's demand is random;
- VRP with Stochastic Travel Times: The time required to serve a certain customer and travel times between two points are random.

Over the years, many authors have used SVRP for a wide variety of applications. Table 2.3 provides a few examples.

Table 2.3: Examples of application of the SVRP

Case Study	Reference
Delivery of meals	Bartholdi et al. (1983)
Delivery of home heating oil	Dror et al. (1989)
Sludge disposal	Larson (1988)
Forklift routing in Warehouses	Bertsimas (1992)
General pickup and delivery operations	Hvattum et al. (2006)
District design	Haugland et al. (2007)

As Barnhart and Laporte (2006) explain, Stochastic VRPs are commonly solved with stochastic programming, in a two stages approach. Initially, an a priori solution is computed considering that all customers are served once. Then, once the random variables become known, the initial solution is corrected. This recourse action may cause additional cost (if after knowing the actual values the vehicle needs to go back to the depot more often, for example) or provide a lower cost solution (if the actual route is easier than the a priori). The choice of the most appropriate recourse policy depends on the time at which the random variables become know. Dror et al. (1989) present an extensive discussion on this topic.

VRP with Stochastic Customers typically arises in courier companies (FedEX, DHL, etc) operations. Each vehicle is assigned a district that contains both a set of regular customers and some additional occasional ones. Lei et al. (2012) introduced a heuristics for solving the vehicle routing and districting problem with stochastic customers. Groër et al. (2009) highlighted that consistently assigning the same sets of customers to drivers can improve service.

Chapter 3

Automotive Industry's Case Study

In order to guarantee the viability and sustainability of the improvement solutions, it is vital to start by carefully assessing the current state of the processes and how these could be improved to increase the operations' quality and efficiency.

During this chapter, the current situation of the Company's Service Model will be analysed. The main goal of this section is to deliver an integrated view of the Company's macro processes. These analyses will have a big impact on the final solution as one of the key activities is to identify improvement opportunities that will later be worked on.

3.1 Company Presentation

The Company's story begins in 1930 when their owner started a franchised business of an international luxury vehicles brand in Portugal. In 2000, they launched their heavy-duty Aftermarket activity, with the acquisition of an Aftermarket distributor and today the Company is present in 16 countries with more than 3800 employees across the world.

Their activity is developed in four major sectors: Original Equipment Solutions, Integrated Aftermarket Solutions, Recycling Solutions and Safekeeping Solutions. To ensure a long relationship with the customer, that went beyond the moment of sale, the Integrated Aftermarket Solutions business includes importation and distribution of heavy and light vehicle parts, the latter being the focus of the present thesis.

Despite the Company's stable position in the market, it recognizes the volatility and competitiveness of the business environment, understanding the need for a culture of action and speed in adapting to new market trends and customer needs. As highlighted in last year's financial statement,

> "Our vision of the future is one of growth and development but also of alignment and sustainability of our operations and businesses, globally, and this is what made it essential to redesign the governance model".

In the 2018 financial report, the company explains how its approach to innovation relies on three pillars: defend their business position, seek new ones and be disruptive to change the industry. As such, it invests in a numerous amount of projects. In 2017, they started a transformation project that includes all areas of the business. It aims at strengthening and uniforming the processes, creating a solid structure to face a series of current and future challenges and opportunities. The first implementation of the project was in the Aftermarket segment, in Portugal and Spain.

In 2019, as part of this project, Portugal was the pilot market for the implementation of a project that aimed at optimizing the Aftermarket logistics.

As far as 2020 is concerned, two more projects regarding customer relations and after-sales processes are being developed.

3.2 Delivery System Analysis

Several factors contribute to the complexity of the business's supply chain. Firstly, the wide variety of products offered (over 100 000 Stock Keeping Units (SKUs)) belonging to several categories, such as brakes and traction control; cooling, heating and climate control; electrical and lighting; engine management; filters and PVC; among many others. Secondly, the wide variety of delivery locations (over 5000) and each customer's delivery requirements. The last and one of the most critical aspects for the purpose of this study has to do with the existence of multiple daily deliveries. This need arises from the fact that is impossible for an automobile repair shop to keep stock of the wide range of SKUs offered today. Besides the obvious space problem, there is a working capital barrier. That being said, the only way to offer a good service to the end customer (a person whose mobility is limited because he just delivered his car at the repair shop) is through multiple daily deliveries to the repair shops. In this way, the repair shop does not keep stock and only buys once the end consumer asks for a repair and the aftermarket distributor ensures the quickest arrival by delivering several times per day.

The business has two warehouses (one in Porto and another one in Seixal) and four stores (one in each warehouse, one in Lisbon and one in Faro). The focus of this analysis will be the warehouses, as they serve the majority of the Company's clients.

Receiving and Processing an Order

When a customer makes a request, a buying order is created and then converted to a warehouse order. This conversion will assure the stock existence and all the information necessary for the warehouse (location, amount of stock, etc). This warehouse order is then automatically distributed to a team of pickers. After the picking activity, the items are subjected to a conference activity that checks if the products and quantities picked are compliant with the customer order. Afterwards, the items are packed and prepared for the expedition.

The expedition of the customer order is dependent on the time that the order was placed. Here, the Company uses what they call a "Cutoff System": each cutoff has a start time, an end time and a

transport exit time. All orders placed between the start and end time will be aggregated in a single delivery.

Each customer is assigned to a single delivery route, and each route has a certain number of cutoffs that corresponds to the number of deliveries offered to the customers in that route. Each route is assigned to a vehicle. As an example, please consider table 3.1 that corresponds to the cutoffs available for the exemplary route R, assigned to vehicle V. If a customer assigned to this route makes an order between 10:30 and 13:00, those items will leave the warehouse at 14:15 and will be delivered until 16:00. For route R, where customers have four deliveries/day, vehicles have exactly 120 minutes [1] for loading (which takes 15 minutes), delivering the items and coming back to the warehouse to load the next cutoff's orders and leave for the next round (16h30). It should be noted that the start time of the first cutoff (20:00) concerns the previous day.

Table 3.1: Cutoffs available for route R

Start Time	End Time	Loading	Transport Exit	Delivery to Customer
20:00	10:30	11:00 - 11:15	11:15	11:15 - 13:00
10:30	13:00	14:00 - 14:15	14:15	14:15 - 16:00
13:00	16:00	16:00 - 16:15	16:15	16:15 - 18:00
16:00	20:00	9:00 - 9:15	9:15	9:15 - 11:00

Since there are several warehouses and stores, each customer is allocated to one of them as their preferential centre, based on proximity. Therefore, each client will always be served by it, unless there is a stockout of the ordered item.

Order Delivery

The Company has no transportation of their own so the deliveries are outsourced in two different ways. For some deliveries, a "Dedicated" system is in place where, each month, they rent several "delivery vans" of which they have full control. For the rest, a "Shared" system is used where they pay each delivery individually. Whilst in the Dedicated system the costs per month are fixed, in the Shared system, they vary accordingly to the number of deliveries and additional weight per delivery.

As the Company's business grew, so did customers located in areas where they never delivered before. Given that the delivery area of each logistics partner is limited, the company was forced to start working with more and more partners in order to respond to these clients. Besides the geographical coverage constraints, there are also some limitations regarding specific products' transportation (body components/plates, oils and batteries) and delivery cost charging (for clients with monthly budgets agreements).

Today, the Company works with a total of 13 logistics partners [2], each one with its own price table. Right away, this is something identified as a weakness, as working with fewer would allow obtaining economies of scales, thus reducing the overall transportation costs. To understand

[1] 120 minutes corresponds to dividing the the vehicle's driver working hours - 8 hours - by 4 deliveries.

[2] Logistics partners' names hidden for confidentiality reasons.

	Price Table		Competences				
Logistics Partner	**Dedicated vehicle cost**	**Shared service cost**	**Geographic Coverage**	**Maximum deliveries/day**	**Body Parts**	**Oils & Batteries**	**Cost charge**
LP1		3,75 €	Interior North	1	✓	✓	
LP2	2300 €	3,02 €	Porto, Center, Lisbon & Algarve	2 (Shared) 4 (Dedicated)	✓	✓	✓
LP3		3,73 €	Porto, Center, Lisbon & Algarve	2	✓	✓	
LP4		4,75 €	National	1			✓
LP5	2117 €	3,70 €	Porto & few zones Interior North	2	✓	✓	✓
LP6		3,80 €	Porto & few zones Interior North	2	✓	✓	
LP7	2000 €		Porto & Lisbon	4	✓	✓	✓
LP8		5,70 €	Interior North	2	✓	✓	
LP9		4,80 €	Alentejo & Algarve	1	✓	✓	
LP10		5,59 €	Spain	1	✓	✓	✓
LP11		6,58 €	National	1	✓	✓	✓
LP12		4,80 €	Interior North	2	✓	✓	
LP13		4,80 €	Interior North	2	✓	✓	

Table 3.2: Logistic Partners' Competence Matrix

what improvement opportunities might exist in this area the delivery areas of each partner were analysed, as well as their price table and other constraints. Table 3.2 summarizes the results.

Given the wide range of partners, it is likely that there is room for improvement. Once the Service Model and Distribution Network are designed and future needs assessed, this should be worked on.

Regarding the two delivery systems, not only do they differ in terms of pricing but also in the process (figure 3.1). In the Dedicated system, the car picks up the order from the warehouse and delivers it directly to the customer, ensuring a lower delivery time and, ultimately, a higher service level. In the Shared system, the logistics partner consolidates several orders in a transit point before doing the delivery. Thus, the delivery time is higher and, as there is more handling of the parts, the risk of damage increases.

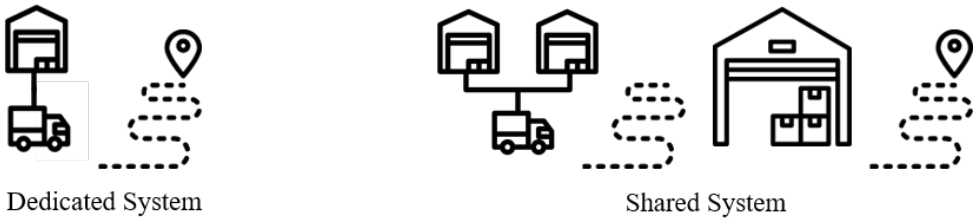

Figure 3.1: Delivery Systems representation

At the moment, each month, the Company hires 24 Dedicated vehicles: 7 for Porto's warehouse and 17 for Seixal's. In fact, the Company does not present a structured process for evaluating the required number of Dedicated vehicles. Basically, customers are grouped together based on proximity and whenever a certain vehicle is unable to deliver to all its customers, they hire a new one. This process is very naive and lacks consolidating data to support decision making. Creating an optimization model that would calculate the exact number of dedicated cars needed might just be what the Company needs to reduce transportation costs and ensure the best service level.

3.3 Customer Segmentation

The Company divides its customers into six segments: White, Blue, Silver, Gold, Platinum and Diamond. This differentiation affects not only the number of deliveries offered but also the discounts provided. Regarding the first topic, this differentiation is much more simple: Gold, Platinum and Diamond are offered four deliveries per day, except if they are located more than 45 minutes away (it would take 1h30 minutes just to go there and come back, making it impracticable to deliver to several customers and therefore economically unviable); Silver, Blue and White only have two deliveries per day. The latter are charged delivery fees unless the purchase value of the first order of the cutoff [3] is higher than 150€. Table 3.3 summarizes the above description.

Table 3.3: Current customer segmentation

Customer Segment	Number of Customers	Number of daily deliveries offered
Diamond	2	4
Platinum	269	4
Gold	437	4
Silver	324	2
Blue	693	2
White	3796	2

When asked about the segmentation process, it was clear that it was something done years ago that has never been completely reevaluated ever since. Only sporadic changes are made, that rely on experience and commercial expertise. Hence it is missing crucial information about how

[3] Customers may place several orders for the same cutoff.

valuable each customer is to the Company. On June 2019, the Company implemented SAP system that captures much more data, in a much more organized way. A proper customer segmentation analysis using the new data would allow offering just the right service level to each customer, thus providing a cost-effective, while remaining competitive in the market.

3.3.1 What impacts customer satisfaction?

Besides knowing what the most important aspects of the Service Model impacting customer satisfaction are, it is critical to understand which ones actually affect the Company's clients.

To do so a Pareto analysis was conducted to assess the causes for customer complaints, considering all customer's complaints from January 2019 to January 2020. The results can be seen in figure 3.2. Right away, it is evident that delivery delays is one of the aspects that the company must master. On the other hand, the Company believes that most of the other causes could be handled by increasing the share of Dedicated deliveries over Shared deliveries, of which they have less control.

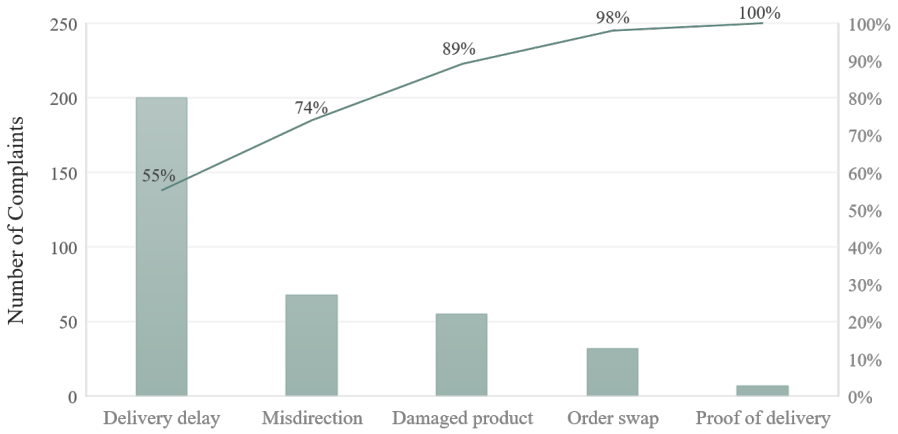

Figure 3.2: Pareto Analysis of the Causes for Service Level Complaints

3.4 Historical Data Analysis

Besides mapping the Company's processes, further clarification of a few more questions is necessary. The first analysis - subsection 3.4.1 - aims at understanding whether customers are being given the promised delivery time. Then, it is critical to identify what factors drive transportation cost reduction and evaluate the differences in the profitability of the Dedicated and Shared systems. This is done in subsection 3.4.2. Lastly, it is important to evaluate whether the Company customers are taking full advantage of the deliveries offered, as it might provide critical insights on customer trends, which should help decide the service level definition. Subsection 3.4.3 analyzes this topic.

Since the Company installed a new Enterprise Resource Planning system (ERP) in June 2019, the analysis done focuses on the time horizon from the installation until February 2020.

3.4.1 Which Warehouse is serving which client?

The first analysis made was to the warehouse that served each order, that is, where are the customers' orders actually leaving from? Answering this question will allow a better evaluation of whether customers' delivery time is being jeopardized because their order is not leaving from their preferential warehouse.

In order to do so, the past year's billing data was studied, from June 2019 until February 2020. All the deliveries were mapped in a Logistics Graph Map (figure 3.3) and coloured accordingly to the warehouse or store they left from. As it can be seen, both warehouses have Portugal-wide deliveries[4].

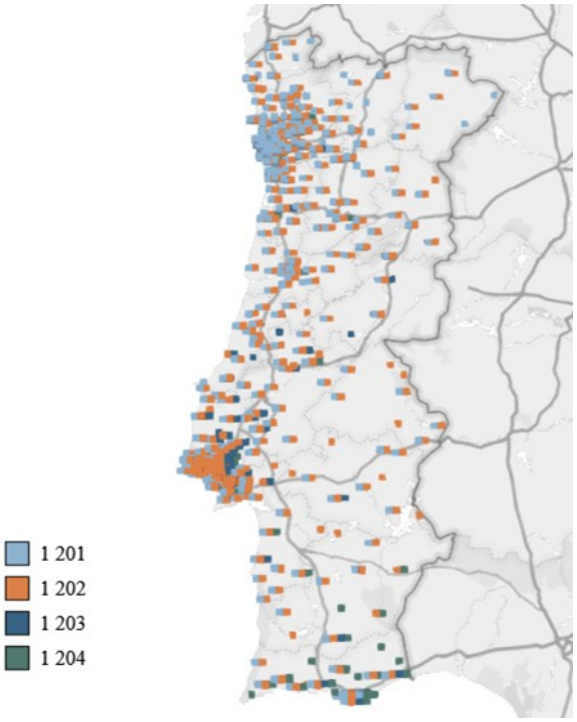

Figure 3.3: Real Warehouse Distribution

Although it was expectable to see Porto's deliveries from north to south of Portugal (because body parts' stock is centralized in Porto's warehouse), the same can not be said of the others. To further understand why this was happening, all the February 2020 deliveries which were not leaving from the preferential warehouse and respective causes were evaluated. Table 3.4 shows the causes for these situations.

In a total of 172 misallocations, 82 happened because the customer went directly to a store that gets its products from a different warehouse. This is something the Company cannot control. Only three cases were attributable to employees' mistakes.

More critical is the fact that 85 cases originated from stockout of the item ordered in the preferential centre. After further analysis, the patterns showed that each stockout was not being

[4]Each point in the map represents one delivery location

Table 3.4: Causes for wrong deliveries

Cause	Number of Cases	Percentage
Stock	85	49%
Customer	82	48%
Employee	5	3%

solved, as the items would only be replaced in the warehouse they left from, being the cause of more wrong deliveries.

This analysis intended to understand whether the management model for transports implemented made sense and although no inconsistencies were found at the distribution level, a relevant problem of stockout was identified. Even though stock management falls outside the scope of this project, this is a critical problem that the Company must tackle in the near future.

3.4.2 Which system is more profitable?

Another important analysis is the profitability of each delivery system. The Dedicated system behaves like a fixed cost, that is, the more deliveries it does, the lower the cost per delivery, up to the maximum capacity of the leased fleet. On the other hand, the Shared system is a variable cost, as the Company only pays for the deliveries it hires. Therefore, the bigger the share of deliveries that the Dedicated system is responsible for, the lower the number of extra hired deliveries and the lower the overall distribution costs.

The idea behind this study is the following: as previously explained, each customer is assigned to a route, and each route is either Dedicated or Shared. The analysis was made by comparing the monthly revenue of the customers that belong to a certain type of route with the monthly costs of that type of route. Table 3.5 summarizes the comparison of the two systems.

Surprisingly, the results are not that different. While the Dedicated system transportation costs represent 4.5% of overall sales, the Shared system shows a value of 5.4%. Therefore, several strategies were identified. For the Dedicated System, the goal is to maximize their utilization because it represents a fixed cost (maintaining the number of cars). This means that, unless the Dedicated vehicles are being fully utilized, there is improvement margin and the possibility to transfer more routes or customers into the Dedicated system, reducing the number of extraordinary Shared deliveries. To evaluate such a possibility we must analyse the volume and time occupation of the Dedicated vehicles. In respect to the Shared system, it is obvious that it must be resorted to it in two situations: firstly, for customers placed in isolated areas, distant from the warehouses; secondly, when the cost of using it is inferior to the cost of an additional Dedicated car because there are not enough deliveries to compensate for it.

Volume and Time Occupation of the Dedicated Vehicles

In order to understand how well the Company is occupying its Dedicated vehicles, it is necessary to study its volume and time occupation.

	Cost Model	Nº Destinations	Profitability	Strategy
Dedicated System	2000€ - 2400€ monthly cost per vehicle	771	**0,9 M€** Sales/month **53 k€** Costs/month **4,5%** Costs/Sales	Ensure vehicles maximum occupancy
Shared System (internal costs)	Cost per delivery depends on the address, cutoff & additional weight (each Logistics Partner has its price table)	903	**1 M€** Sales/month **75 k€** Costs/month **5,4%** Costs/Sales	Since cost does not depend on the travelled distance, it is highly profitable in secluded areas, distant from the warehouse
Shared System (customer pays delivery)	N.A.	218	**0,5 M€** Sales/month	Evaluate if they can be added to the Dedicated system to make it more profitable

Table 3.5: Summary of the Delivery Systems (June 2019 - February 2020)

Starting with the volume occupation, the vehicles' utilized volume was calculated as seen in equation 3.1.

$$Vol = \sum_{i=1}^{N} (NumItems_i \cdot AvgContainerSize) \tag{3.1}$$

where:

- NumItems is the number of orders in that route
- AvgContainerSize is the average size of a product's container (m^3), which was calculated by using the box consumption records

As it is hard to fully utilize the vehicle's truck volume with carton boxes, a 30% vehicle's volume waste (from $4m^3$ to $2{,}8\ m^3$) was assumed. Nevertheless, as it can be seen in figure 3.4, Dedicated vehicles are not even close to being fully utilized. Therefore, the volume will never be a restrain for allocating more customers to a certain route. This conclusion is in agreement with the warehouse managers and drivers' empirical knowledge.

Regarding time occupation, the January 2020 billing information was used to figure out which clients were visited by a certain route. Then, to recreate the vehicle routes and get the vehicles travelling distance, a simple routing algorithm was created. The following assumptions were

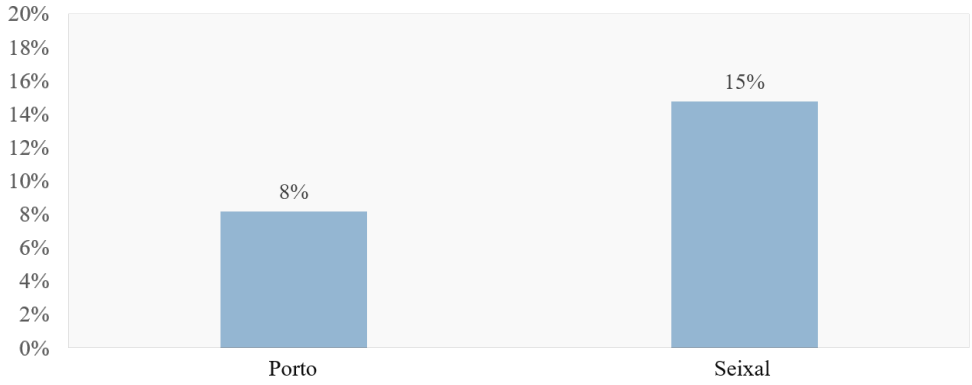

Figure 3.4: Dedicated vehicles' volume utilization

made: average speed of 50 km per hour, a 2-minute stop [5] in each customer location and 15 minutes to load the vehicle in the warehouse.

Given a set of customers assigned to a route with a certain preferential warehouse, the algorithm works as seen in algorithm 4.

Algorithm 4 Auxiliary Routing Algorithm

1: Travel time = 15 min
2: Select the nearest customer to the warehouse (W) $\rightarrow$ A
3: Travel time = Travel time + ((vector distance between W and A) $\div$ 50 km/h) $\cdot$ 60 min
4: **repeat**
5: Select the closest customer to A $\rightarrow$ B
6: Travel time = Travel time + ((vector distance between A and B) $\div$ 50 km/h) $\cdot$ 60 min
7: Customer B $\rightarrow$ A
8: **until** All clients are in the route
9: Travel time = Travel time + ((vector distance between B and W) $\div$ 50 km/h) $\cdot$ 60 min

As it can be seen in figure 3.5, there is great variability between the average number of destinations per Dedicated vehicle, suggesting that it might be possible to increase its utilization. In addition, looking at figure 3.6, it is possible to conclude that the Dedicated vehicles time utilization is very low, as many vehicles present an average travelling time inferior to 1 hour (much smaller than the lowest available time of 1h45 [6]).

Additionally, it became clear that there are several vehicles performing the same route. In Porto, there are 7 vehicles for 5 routes, and in Seixal there are 17 vehicles for 8 routes. This means that vehicles' routes are overlapping - a clear sign of process inefficiency.

All in all, these analyses made it possible to understand that time will be the restraining variable when it comes to defining the number of customers assigned to each Dedicated vehicle and not the orders' volume. Moreover, the vehicles' routes should be improved to avoid inefficiencies such as overlapping routes.

[5] To know the vehicles' unloading time, a car was followed for one day and the unloading times registered. The result was an average of 2 minutes per stop.

[6] Considering 8 working hours per day, divided in 4 cutoffs and subtracted 15 minutes to load the car

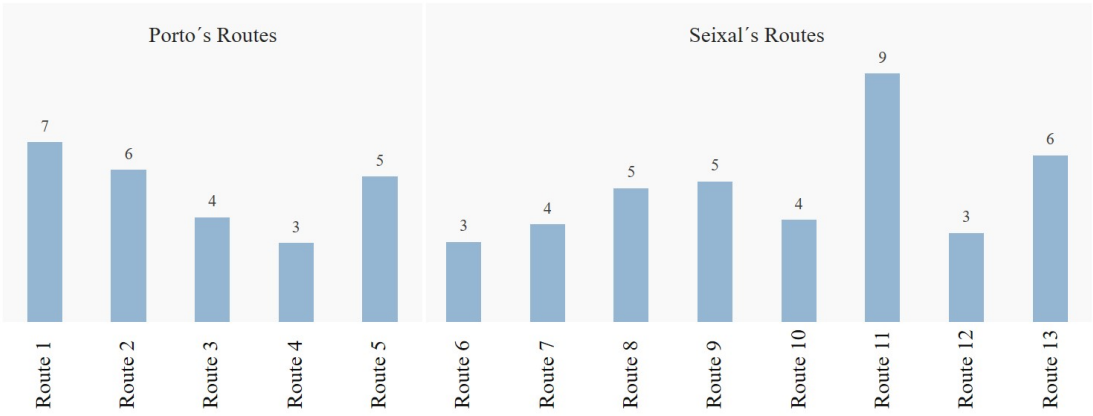

Figure 3.5: Average number of destinations per route

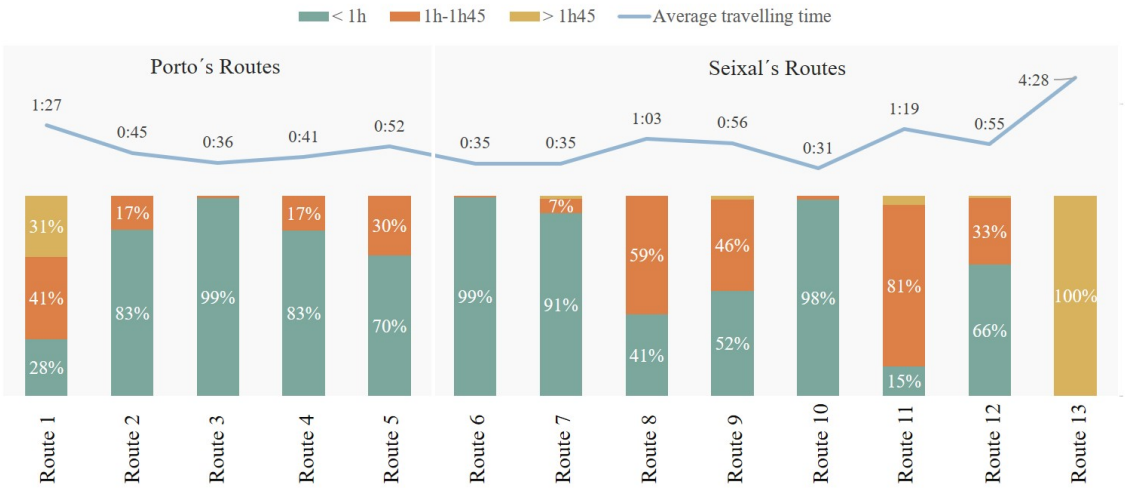

Figure 3.6: Average time per route & % of routes with < 1h, 1h-1h45 and > 1h45

3.4.3 Do Dedicated customers use all the offered deliveries?

As stated earlier, the automotive aftermarket is very competitive in terms of logistics and service offered, meaning that, among other factors, the clients will often choose a certain supplier accordingly to the number of deliveries offered per day and the lead time of each delivery. Or, at least, this is something that the Company's Executive Director of the light vehicle parts believes in.

On the other hand, the higher the number of deliveries offered, the higher the logistics effort and cost it represents to the Company. Therefore, before assuming that increasing the number of daily deliveries will have a positive impact in the Company's profitability, it is crucial to analyse whether the customers actually use the deliveries already offered. To do so January and February 2020 billing information was used to estimate which Dedicated customers [7] were visited in a certain route for each cutoff per day.

[7] Customers served by the Dedicated system

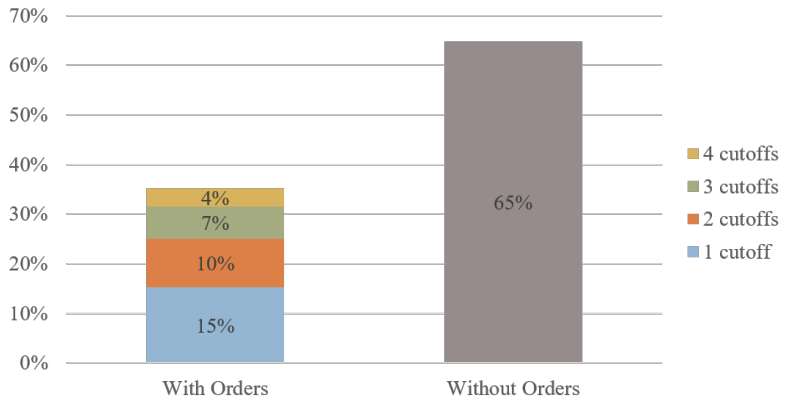

Figure 3.7: Average customer usage of the Dedicated system

From figure 3.7 , some conclusions can be drawn. On average per day, only 35% [8] of customers with Dedicated service place orders. Of those, 70% [9] of the time they merely take advantage of 1 or 2 deliveries. This raises the question: *Until what point does offering more deliveries per day increases customer service level? Should the Company focus on offering less deliveries, but more reliable ones?*

3.5 Conclusion

Throughout this chapter, it was possible to highlight the most important inefficiencies of the process and so the biggest improvement opportunities that will lead to the next steps of the dissertation. The conclusions were that the stock management process should be revised and the Company should leverage the information collected by the new ERP to segment their customers correctly. Regarding the Distribution Process, it became clear that the Company should focus on the Dedicated system. Its efficiency could be improved by increasing the vehicles' time occupation. This would not only impact service level, but it could potentially reduce the number of extraordinary Shared deliveries. Lastly, the Company should consider eliminating the 4 deliveries/day service, as it is seldom used.

That being said, it seems logical that the first step is to perform an analytical customer segmentation based on Customer Lifetime Value, as the current segmentation made by the Company is outdated and mostly updated based on intuition and commercial expertise. This will allow the Company to further understand its customers and the actual value that they represent.

After doing so, it is vital to define the future service model offered to each customer segment, never losing sight of the geographical limitations that might exist. Here, a key insight must not be forgotten: customers are not taking full advantage of the service level offered.

[8] 35% = 15% + 10% + 7% + 4%
[9] 70% = (15% + 10 %)/35%

Moreover, the analysis done highlighted the need for a transportation management tool. Designing and implementing a tool that would allow further knowledge on the optimal way to serve each customer would bring a lot of savings in transportation costs, something vital in this business.

Chapter 4

Service Model

As mentioned in the previous chapter, the Company currently defines the service offered to each customer based on empirical knowledge and commercial expertise. While that might have worked when it only had a few customers, it becomes unbearable to design a clear strategy for each one of the many customers it currently has. It should be noted that we are talking about delivering over 10 000 items to more than 500 delivery locations from north to south of Portugal, to be distributed over one, two, three or four deliveries per day. In fact, given the dimension and complexity of this problem, there is no efficient and profitable way of doing this without using analytical methods.

Having said so, it seems only logical that re-designing the Service Model is the first step. All the information collected and analysis performed that relate to this topic are described in this chapter. Initially, the proposed approach is presented, and then the analysis undertaken and respective results and conclusions are discussed.

4.1 Proposed Approach

For the purpose of this analysis, a Cross Industrial Standard Process for Data Mining (CRISP-DM) methodology (Shearer, 2000) will be followed. As Shearer (2000) describes, this model has six phases:

1. Business understanding - focuses on understanding the objectives from the business point of view and developing a plan to achieve them.
2. Data understanding - involves data collection and description, as well as exploring it to verify its quality.
3. Data preparation - corresponds to the preprocessing phase described in subsection 2.1.1.1.
4. Modelling - involves the selection of data mining techniques to solve the problem. There may be several structures to do so, therefore, it might be necessary to go back to the data preparation task if they have different requirements.

5. Evaluation - consists of the evaluation of the model in order to check if it served the purpose of reaching the business objectives.
6. Deployment - involves discussing the results and organizing the knowledge gained in a way that the company can use it.

The proposed approach is illustrated in figure 4.1.

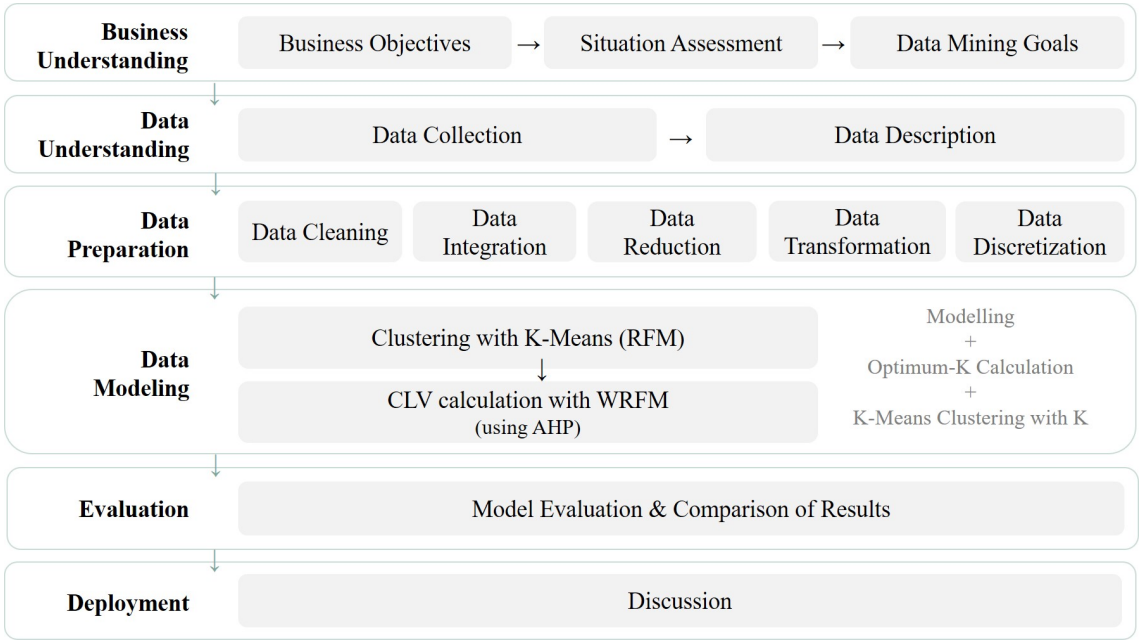

Figure 4.1: Proposed Approach, based on Shearer (2000) and adapted from Carneiro and Miguéis (2020)

4.1.1 Business Understanding

Given the increasingly competitive pressure of the Automobile Aftermarket Industry, the Company believes that leveraging on the Service Model offered to customers plays a vital role to remain relevant in this market. Hence the need for a proper customer value segmentation, which will help understand the buying patterns of different customers. The results of this study will then be used to establish CRM strategies for different customer groups.

4.1.2 Data Understanding & Preparation

The dataset used was provided by the Company. It relates to customer information and 9 months of customer transactions (table 4.1), from June 2019 until the end of February 2020. Prior to any data preprocessing, the dataset includes 1 133 527 transactions belonging to 5 249 different

customers [1]. The time span chosen for the analysis relates to the months after the implementation of SAP system in the Company.

Table 4.1: Sample of the original transaction dataset

Transaction Date	Customer ID	Customer Name	Product Code	Product Description	Quantity	Transaction Value
02/02/2020	1000000093	CUST1	A-642	Mirror	1	16,97 €
02/02/2020	1000000094	CUST2	MO-110	Bearings Kit	1	7,16 €
03/03/2020	1000000099	CUST3	AU-110	Valve Cover	1	1,75 €

In order to conduct the required RFM analysis, the provided dataset must be preprocessed. The main steps performed here, where:

1. Eliminate transactions that were registered twice.
2. Eliminate transactions that did not have an associated customer.
3. Select attributes of interest from the dataset and eliminate any unnecessary information. In this case, Customer Id, Customer Name, Transaction Date, Transaction Value, Customer Country, Customer Postal Code and Customer Coordinates were chosen.
4. Match each customer ID to their transactions and geographic location (country, postal code and coordinates).
5. Filter transactions where:
   (a) The Customer's Country is not Portugal
   (b) The Customer's Postal Code is empty or invalid
   (c) The Customer's ID does not correspond to an actual customer but to a logistics's partner (transactions related to damaged products)
   (d) The Transaction Value is zero
6. Merge the several transactions made in a day by the same customer into one.
7. Transform the dataset, as seen in table 4.2. Recency was calculated as the number of days since the last purchase until 29/02/2020 [2]; Frequency as the number of days with purchases; Monetary as the average transaction value. Table 4.3 shows the attributes that compose the dataset.

After these steps, the dataset obtained contains a total of 1 006 893 transactions that belong to 5191 different customers.

Table 4.4 presents a brief summary of the RFM analysis, performed in RStudio. From this brief preliminary overview, some insights stand out immediately. Unlike the other two variables,

[1] It is worth noting that the customers referred to during this analysis do not necessarily correspond to a company, but to the actual points of deliveries it has. As an example, consider customer A that has two repair shops. For the purpose of this analysis, each repair shop (point of delivery) will be analysed separately

[2] Last day of the analysis

Table 4.2: Sample of the obtained dataset

Customer ID	Customer Name	Customer PostalCode	Customer Coordinates	Recency	Frequency	Monetary
1000000093	CUST1	2620135	41,25; -8,64	2	107	82
1000000094	CUST2	4100320	38,48; -9,12	2	80	147
1000000099	CUST3	4410014	37,07; -8,00	142	4	176

Table 4.3: Attributes in the obtained dataset

Variable	Data Type	Data Format	Description
Customer ID	Nominal	Integer	Correspond to each distinct customer number
Customer Name	Nominal	Text	Correspond to the customer name
Customer PostalCode	Nominal	Integer	Correspond to the customer 7 digits postal code
Customer Coordinates	Nominal	Text	Correspond to the customer *latitude, longitude*
Monetary	Numeric	Integer	Average amount spent per customer transaction
Recency	Numeric	Integer	Number of days between the customer last transaction and the last day of analysis (29/02/2020)
Frequency	Numeric	Integer	Number of transactions made within the analysis time interval

the mean of the frequency variable is higher than the 3rd quartile, meaning that less than 25% of the customers have the highest number of deliveries. Secondly, the maximum values of frequency and monetary variables are quite distant from the 3rd quartile, meaning that it is likely that a few customers have atypical consumption behaviours in terms of these characteristics.

Table 4.4: Summary of the variables

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Recency	0,00	9,00	61,00	86,18	151,00	271,00
Frequency	1,00	1,00	2,00	26,93	15,00	227,00
Monetary	1,00	35,00	68,00	99,17	114,00	6044,00

4.1.3 Modelling

After spending time preparing the data, it is now ready to be modelled. Subsection 2.1.1.2 presents a detailed description of all the techniques used to do so. The entire analysis was performed using RStudio.

For the purpose of this dissertation, given its wide usage in similar studies and advantages, K-Means algorithm will be used. As mentioned in the Literature Review, K-Means algorithm is

very sensitive to outliers or variables that have incompatible scales. For that reason, the segmentation process was subdivided in several parts. Firstly, an outlier detection analysis was computed. Secondly, the correlation between the variables was evaluated. The last step, previous to the actual clustering, was the normalization of variables. Nevertheless, the first thing to do is verify whether the data has a cluster tendency or not.

4.1.3.1 Assessing Clustering Tendency

To assess cluster tendency, the Hopkins statistic was calculated[3]. The Hopkins value for a sample of 1000 data points was 0.0103, meaning that the null hypothesis is rejected and the data is highly clusterable. After having verified this, we can proceed with the analysis.

4.1.3.2 Outlier Analysis

Examining the boxplots of the RFM variables (figure 4.2) it is possible to verify that some instances have atypical Frequency and Monetary values as they are outside the interquartile range.

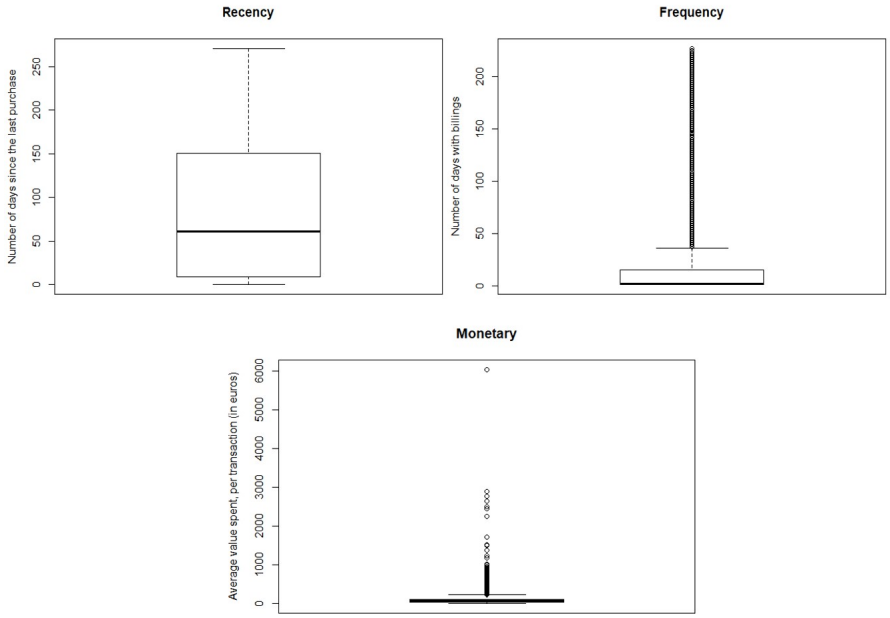

Figure 4.2: Boxplot of the RFM variables

Further analysis of the interquartile range of each variable revealed that there are a total of 356 Monetary outliers and 989 Frequency outliers. As some of these coincide, the total number of outliers is 1 151, representing 22.1% of the total customer base.

From the business point of view, these are valid instances, as they represent real transactions. In fact, these customers that have extraordinary Monetary and Frequency values are, in most likelihood, the Company's most valuable customers. For that reason, these values were considered for the rest of the analysis.

[3]Using R studio's *hopkins* function in *clustertend* package

4.1.3.3 Correlation Analysis

Next, the Person correlation coefficient was computed to evaluate the linear correlation between each pair of the RFM variables. The results can be seen in figure 4.3.

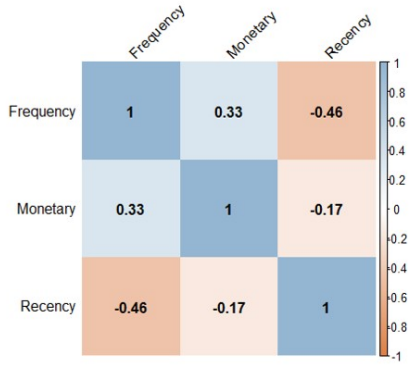

Figure 4.3: Matrix of correlation of the RFM variables

Evans (1996) proposed the following guide for evaluating the strength of the correlation: absolute values $\in$ [0.00, 0.19] are considered “very weak”, [0.20, 0.39] “weak”, [0.40, 0.59] “moderate", [0.60, 0.79] “strong” and [0.80, 1] are considered “very strong”. Considering this scale and given that the absolute results range between 0.17 and 0.46, it is possible to say that the variables do not present a strong correlation.

4.1.3.4 Variables Normalization

The next step is to proceed to the variables normalization, as they are not on comparable scales (Recency $\in$ [0, 271], Frequency $\in$ [1, 227] and Monetary $\in$ [1, 6044]). To do so, the min-max normalization method was used, as described in section 2.1.1.2. The results can be seen in table 4.5. Looking at the variables' median and mean it is evident that, despite the normalization, the values' distribution is not the same.

Table 4.5: Summary of the normalized variables

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Recency	0,0000	0,0332	0,2251	0,3180	0,5572	1,0000
Frequency	0,0000	0,0000	0,0044	0,1146	0,0650	1,0000
Monetary	0,0000	0,0056	0,0111	0,0163	0,0187	1,0000

4.1.3.5 Clustering

As aforementioned, K-Means algorithm will be used. There are several methods to find the optimal number of clusters, K. The first metric used was the Elbow Method. Figure 4.4 shows the Elbow curve, that is, the percentage of variation explained as a function of the number of clusters,

considering a maximum of 20 clusters. Looking at the graph the curve does not show a clear elbow. In fact, any k between 3 and 5 seems acceptable.

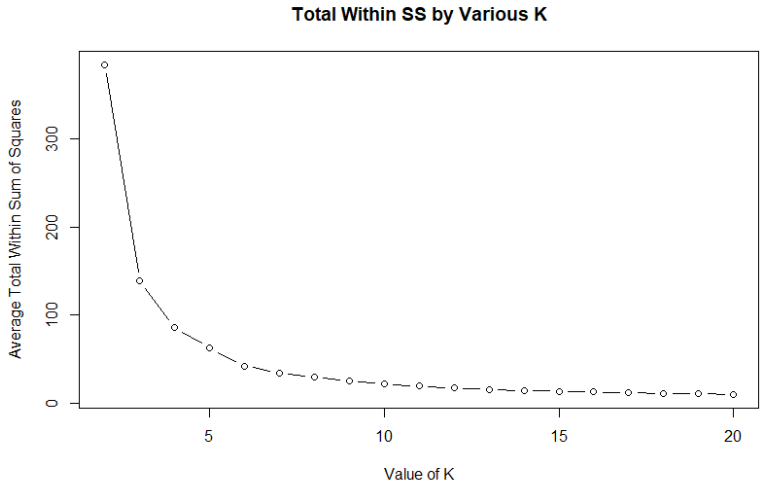

Figure 4.4: Elbow Curve

As this method did not reach plain results, other two metrics were tested: Average Silhouette and Davies-Bouldin. It should be noted that while in the Average Silhouette Method the aim is to maximize the average total within sum of squares, s(i), in the Davies-Bouldin the goal is to minimize the DB value. The outcome can be seen in tables 4.6. Both methods proposed using a k value of 3.

Table 4.6: Results of the Average Silhouette and Davies-Bouldin Methods

Number of Clusters	2	3	4	5	6	7	8	...	20
Davies Bouldin	0,838	**0.566**	0,731	0,610	0,695	0,665	0,663	...	0,794
Average Silhouette	0,507	**0.609**	0,519	0,564	0,524	0,530	0,510	...	0,439

Given the results of these three methods, K-Means clustering was conducted dividing the customer in three clusters that cover 82,40% of the dataset variance (table 4.7). Cluster C3 has the highest cohesion, while C2 has the highest dispersion.

Table 4.7: Sum of Square by Cluster

	C1	C2	C3
Within-cluster sum of square	49,27	64,46	24,41
Between SS / Total SS		82,40%	

The results can be seen in table 4.8. Figure 4.5 shows the visual representation of the clusters obtained[4].

[4]Using R studio's *plot3d* function in *rgl* package

Table 4.8: Normalized RFM Variables per Cluster

Cluster	#Customers	%Customers	Recency	Frequency	Monetary
C1 - Dark green	1759	34%	0,6978	0,0036	0,0115
C2 - Orange	2751	53%	0,1521	0,0450	0,0147
C3 - Blue	681	13%	0,0071	0,6824	0,0348

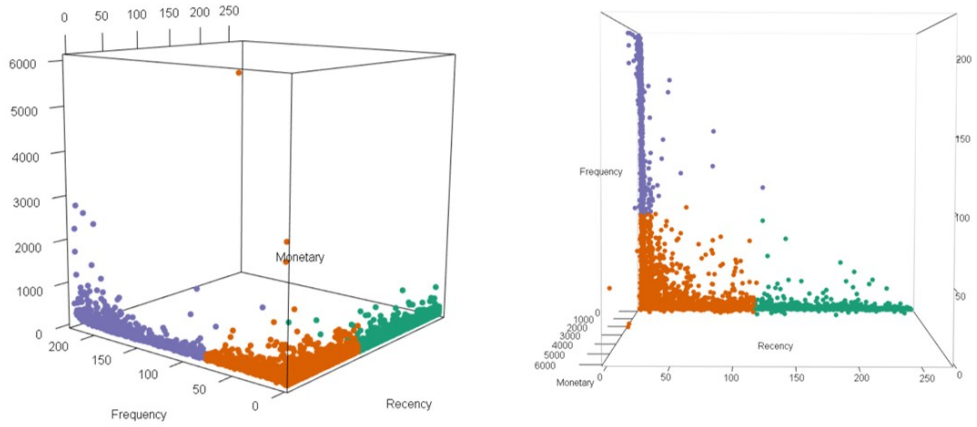

Figure 4.5: 3D Clusters Representation

Each cluster represents one type of consumer. The first cluster C1, representing 34% of the Company's customers, is characterized for a very high Recency value, low Frequency and lowest Monetary. The average C1 customer has not bought anything for 189 days [5], only made 1.8 purchases from June 2019 to February 2020 and spent an average of 70.7 €/purchase. The second cluster C2, representing half of the Company's clients, has an "intermediate" behaviour. The average customer in this cluster does not buy something for 40 days, makes an average of 11 purchases in 9 months and spends around 89.6 €/purchase. The last cluster C3 represents the best customers (only 13% of the entire customer base), who have an average of 153 days with invoices in the 234 [6] working days under analysis. These customers spend an average of 211.3 €/purchase.

Additionally, a decision tree was constructed[7] (figure 4.6) to further understand the behaviour of the customer's belonging to the same cluster. C1 are customers that do not buy anything from the Company for more than 4.5 months. These customers either bought once or have abandoned the Company. C1 are *One-Time-Only* customers. C2 show more recent purchases, but their frequency is extremely low. These are *Occasional* customers. The most important customers are those belonging to C3, as their purchase frequency is the highest and recency the lowest.

4.1.3.6 Estimating Customer Lifetime Value

Prior to estimating the Customer Lifetime Value of each cluster, it was necessary to assess the relative importance of each one of the three criteria. Recently, Analytical Hierarchy Process (AHP)

[5] As stated before, only invoices until the 29th of February 2020 were considered
[6] 6 working days a week
[7] Using RStudio's *rpart* package with default parameters

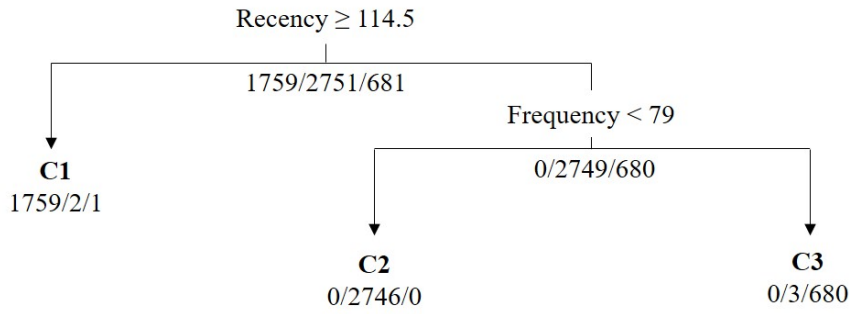

Figure 4.6: Clustering Decision Tree

is one of the most popular methodologies to access the relative importance of each one of the three RFM criteria. This method asks the Company's decision-makers to intuitively evaluate the relative weight of these parameters.

Accordingly to Shih and Liu (2003), AHP can be used to compute the relative weights of the variables in three steps:

1. Evaluators compare each pair of variables (Recency with Frequency; Frequency with Monetary; Monetary with Recency), using the scale proposed by Liu and Shih (2005a) (Table 4.9);
2. Calculate the consistency ratio (CR) to evaluate the consistency of the judgements made. Accordingly to Saaty, only an index inferior to 10% should be accepted. If it surpasses this limit steps 1 and 2 should be repeated;
3. Compute the relative weights of the variables - w_R, w_F and w_M, according to the pairwise comparison matrixes. This can be done using eigenvector's theory, proposed by Saaty (1977).

Table 4.9: AHP scale, adapted from Liu and Shih (2005a)

Value	Description
1	Equally important
2	between 1 and 3
3	Weaker importance of one factor
4	between 3 and 5
5	Vital importance of one factor
6	between 5 and 6
7	Demonstrated importance of one factor
8	between 7 and 9
9	Absolute importance of one factor

The Company's Executive Director of the light vehicle parts and Marketing Manager were asked to evaluate how important each criterion was when compared to another, using table 4.10's questionnaire.

After this it is necessary to evaluate the consistency of the judgements. The comparisons made present consistency ratios of 0.7% and 6,8%, considerably inferior to the 10% limit. Given

Table 4.10: AHP questionnaire for the pairwise comparison

Criterion 1	With respect to AHP priorities, how much more important is one criterion over the other?																	Criterion 2
Recency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Frequency
Frequency	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Monetary
Monetary	9	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	Recency

the assessments made, the relative weight of the RFM variables is w_M = 70.49%, w_F = 21.09% and w_R = 8.41%, ordered by priority.

Once the relative weights of the RFM variables are defined, the Customer Lifetime Value of each cluster can be computed. Table 4.11 shows the results. As expected, C3 is the most valuable cluster, followed by C2 and then C1.

Table 4.11: CLV Ranking

Cluster	CLV	Score
C1	0,2148	3
C2	0,6084	2
C3	0,8468	1

Nevertheless, from the business point of view, it seems like having three clusters is not enough to properly evaluate the Company customers because it could be suspected that an enormous portion of the Company's sales is obtained from cluster C3. To validate the total value spent by all customers in a given cluster was calculated, as well as the total daily deliveries, during the period under analysis. The results, seen in table 4.12, confirm the assumption.

Table 4.12: Sales Amount per cluster

Cluster	Total Sales Amount (€)	% of Overall Sales	Total Number of Days with deliveries	% of Daily Orders
C1	281 813	1%	195	2%
C2	3 567 389	12%	30 748	22%
C3	24 985 220	86%	105 707	76%

Given this conclusion, sub-clusters for cluster C3 were created in order to further understand the consumer behaviour of the different 681 customers belonging to that cluster, as they are the most valuable ones. This decision was supported by the business' interest in having workable clusters, which adequately explain the difference between customer segments.

The previous process was repeated. Firstly it is necessary to find the optimal number of clusters. To do so, the Elbow Method was used, suggesting a k value between 3 and 6 (figure 4.7).

Once again, the Average Silhouette and Davies-Bouldin metrics were tested suggesting a k value of 2. However, dividing into 2 clusters would only covering 67.2% of the dataset variance. Therefore, it was opted for a k value of 5, that covers 87.7%. The results can be seen in table 4.13 and figure 4.8.

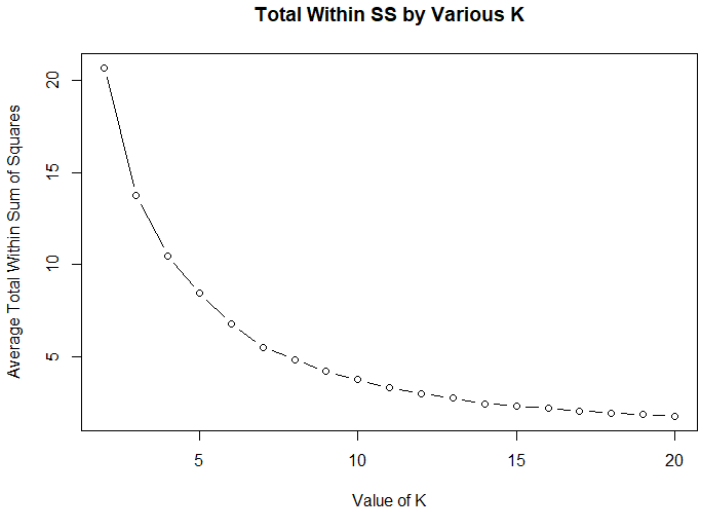

Figure 4.7: Elbow Curve

Table 4.13: Normalized RFM Variables per Sub-Cluster

Sub-cluster	#Customers	% C3 Customers	% Overall Customers	Recency	Frequency	Monetary
SC1 - Dark green	159	34%	4,5%	0,0162	0,3965	0,0366
SC2 - Orange	168	36%	4,8%	0,0333	0,1265	0,0317
SC3 - Blue	134	29%	3,8%	0,0031	0,8711	0,1014
SC4 - Pink	212	41%	5,4%	0,0102	0,6611	0,0621
SC5 - Light green	8	2%	0,3%	0,0031	0,9141	0,7041

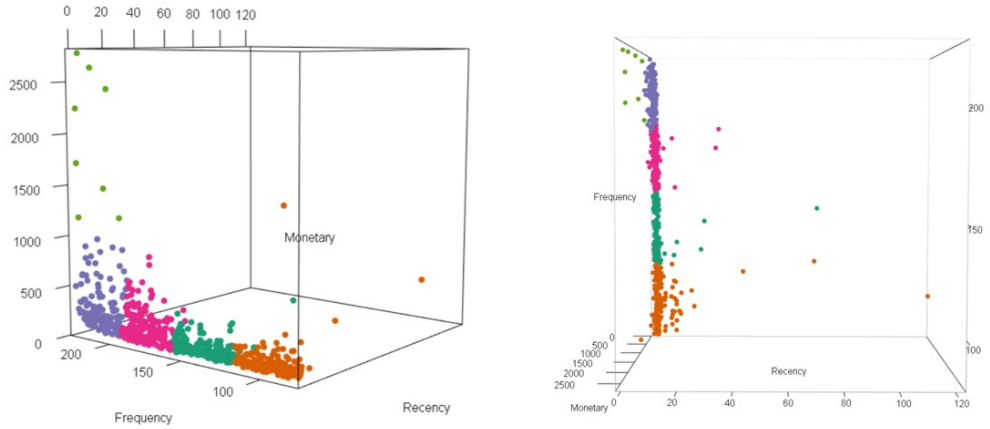

Figure 4.8: 3D Sub-Clusters Representation

Once again, each sub-cluster represents one type of customer. Their buying behaviour and CLV ranking can be seen in table 4.14.

Table 4.14: Customers Buying Behaviour per Sub-Cluster and CLV Ranking

Sub-cluster	Recency	Frequency	Monetary	CLV	Score
SC1	2	138	136	0,7802	4
SC2	4	99	123	0,7108	5
SC3	0	208	314	0,8950	2
SC4	1	177	206	0,8424	3
SC5	0	214	1966	0,9548	1

The decision tree was once again constructed (figure 4.9).

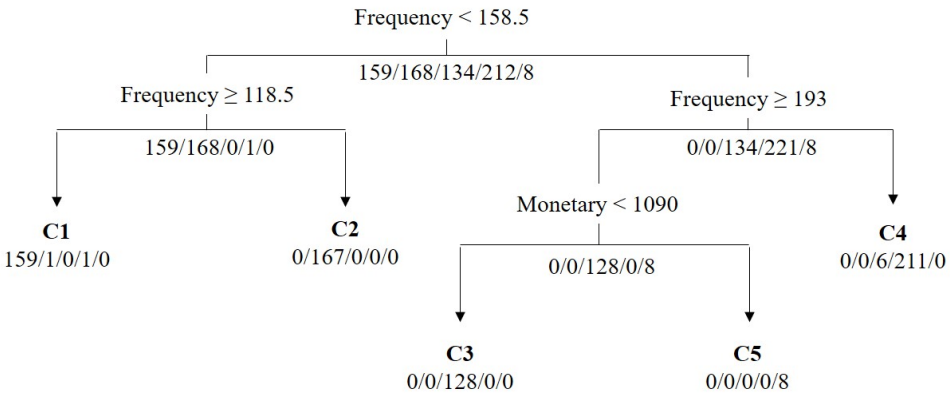

Figure 4.9: Sub-clustering Decision Tree

SC5 customers stand out from the others for their atypical average transaction value. These are the *Big Spenders*.

From SC3 (second best) to SC2 (worst), the biggest difference concerns the frequency values. SC3, like SC5, buys at least 5 days per week, but their monetary values are much lower. These are *Champion* customers because they buy almost every day. SC4 typically has between 4 and 5 days with transactions per week (bought between 158 and 193 of the 234 days under analysis). Despite not being as good as the first two clusters, these customers are *Loyal* to the Company. SC1 bought between 118 and 158 of the days, translating to 3 to 4 days with transactions per week. The Company should try to understand their needs to transform them into *Loyals*. For now, they are *Potential Loyalists.*

Lastly, SC2 bought between 79 and 118 days, meaning that their number of days with transactions per week varies between 2 and 3. SC2 are *Promissing* customers in the sense that they present a high potential for growth because they already buy on a weekly basis. [8]

[8]The "personas" created are merely exemplary. They were created for a better understanding of the data mining techniques used, when presenting the results to the Company.

4.1.4 Evaluation & Deployment

The customer clustering allowed obtaining seven customer segments (table 4.15).

Table 4.15: Customer Segments Summary

Customer Segment	Cluster	CLV Score	#Customers	Recency	Frequency	Monetary
Big Spenders	SC5	1	8	0	214	1966
Champions	SC3	2	134	0	208	314
Loyal	SC4	3	212	1	177	206
Potential Loyalists	SC1	4	159	2	138	136
Promising	SC2	5	168	4	99	123
Occasional	C3	6	2751	41	11	90
One Time Only	C1	7	1759	189	2	71

At this point, it is now essential to understand if the business objectives are met. To do so, the Company was involved in assessing the models and examining the segments obtained and verify whether they make sense from the business point of view.

For that, it is paramount to compare the results with the Company's current customer segmentation. In an optimum situation, table 4.16 would show a green diagonal across the table, meaning that the customer segments that the company currently has matched the ones retrieved from the analytical analysis. In fact, it would be acceptable to have some level of unfitness as a few customers that belong to the same company are usually given the same service level as the best customer of that group, for commercial reasons.

Table 4.16: Customer Segments Summary

Customer Segment	Big Spenders	Champions	Loyal	Potential Loyalists	Promising	Occasional	One Time Only
Diamond		50%	50%				
Platinum	3%	25%	26%	12%	8%	22%	4%
Gold		9%	16%	11%	11%	45%	8%
Silver		10%	14%	9%	15%	44%	8%
Blue		1%	7%	10%	11%	60%	11%
White						56%	44%

Even so, this analysis highlights the urge to review the current customer segmentation, as so many disparities were found.

The next step is to develop a business strategy, in terms of deliveries, based on this customer segmentation analysis. As one may imagine, a pure customer segmentation based on consumer behaviour variables is not enough. To properly define the number of deliveries offered to each customer it is vital to consider their geographical location. To incorporate this factor a density-based clustering algorithm will be used to cluster customers based on their geographic coordinates.

4.1.4.1 Creating Geographical Clusters

For this analysis, the density-based algorithm chosen was DBSCAN [9]. As explained in the Literature Review, DBSCAN algorithm requires two input parameters: Eps and MinPts. MinPts is the minimum number of points required to make a cluster which, from the business point of view, made sense to be equal to the average number of points a Dedicated vehicle visits per cutoff. Therefore, MinPts was defined as 7.

For the neighbourhood radius, Eps, Hahsler et al. (2019) suggests plotting the points kNN's distances [10] in decreasing order and search for a knee. The concept behind this is that since points in the same cluster will be close together, they will have a low k-nearest neighbour distance, while noisy isolated points will have high kNN distances. The knee corresponds to a threshold where a sharp change occurs along the k-distance curve. As it can be seen in figure 4.10, a knee is visible at around a 7-NN distance of 0.085 [11].

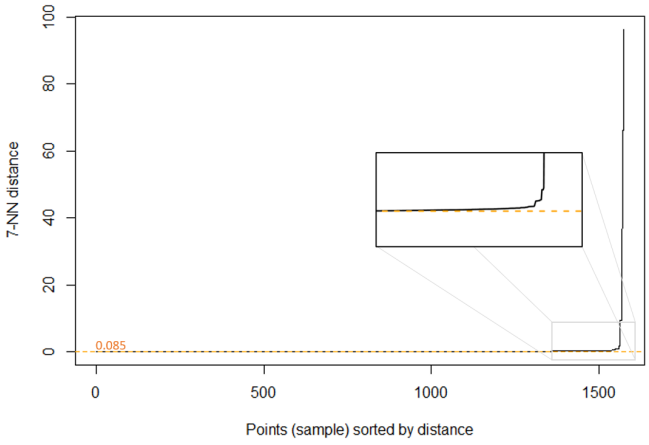

Figure 4.10: k-NN distances plot

Given these inputs, the model proposed 24 clusters that include 1325 out of the 1573 [12] customers under analysis. The remaining 248 were considered outliers (grey points in figure 4.11). From the business point of view, the results are perfectly aligned with what was expected. There are two big clusters: *Porto & Guimarães* with 419 customers and *Lisboa & Setúbal* with 598. Clusters *Ericeira*, *Leiria*, *Faro* and *Guarda* have 55, 24, 20 and 20 customers, respectively. All the other clusters have between 7 and 18 customers. Notice that the customer's preferential warehouse will be redefined based on the geographical cluster they belong to.

4.1.4.2 Defining the Service Level

The number of deliveries per day that is possible to offer depends on the clusters' centroid distance to the warehouse. The rationale is the following. Delivering 4 times a day means that the vehicles

[9] Other algorithms considered, DBSCAN (dis)advantages and its applications are described in Chapter 2

[10] Distance to the k^{th} nearest neighbour

[11] Horizontal line manually included for reference

[12] For the purpose of this analysis, service desk customers that only buy in stores were excluded as the Company does not deliver to them

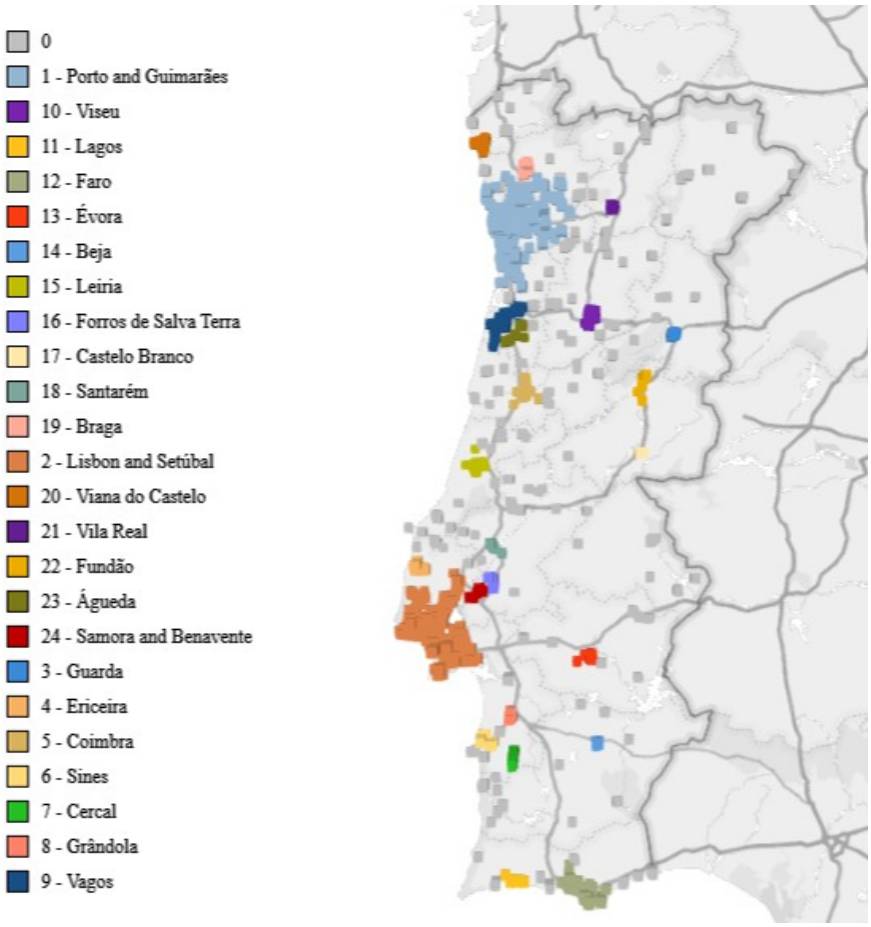

Figure 4.11: DBSCAN Clustering Results

have 1h45min (2h minus 15 min to load the car) to deliver the goods. Therefore, this system is not cost-effective for any cluster placed more than 45 minutes away. The same exercise was undertaken for the rest of the deliveries, together with the Company's Executive Director and sales team, using table 4.17's scale.

Table 4.17: Classification scale in terms of delivery effort

Score	Distance to the Warehouse	Nº of possible deliveries per day
Low	< 45min	4
Medium	45-1h30min	2
High	>45 min	1

The results are as follows:

- Low : Porto & Guimarães, Lisbon & Setúbal, Benavente and Ericeira
- Medium : Coimbra, Sines, Grândola, Vagos, Viseu, Lagos, Faro, Évora, Beja, Leiria, Forros de Salvaterra, Santarém, Braga, Viana do Castelo, Vila Real and Águeda
- High : Guarda, Cercal, Castelo Branco and Fundão

Then, given these results, the Company's Executive Director of the light vehicle parts and sales team decided on the future number of deliveries offered to each customer segment, on a given geographical cluster (table 4.18).

Table 4.18: Future Service Level for each Customer Segment, given the required Delivery Effort

	Delivery Effort		
Customer Segment	Low	Medium	High
Big Spenders	4	2	1
Champions	4	2	1
Loyal	4	2	1
Potential Loyalists	3	2	1
Promising	3	2	1
Occasional	2	2	1
One Time Only	2	2	1

As all customers must be served, the outliers were assigned to the closest geographical cluster based on the cluster centroid coordinates and given the respective service level.

4.2 Conclusion

Customer segmentation can help companies to better understand their customers, allowing them to effectively satisfy customer's needs. Particularly in the Automotive Afterparts industry, where logistics costs have a huge impact on companies' profitability, it is vital to offer the exact service level that allows companies to remain both competitive and cost-efficient.

In this chapter, the Company's customers were segmented accordingly to three consumer behaviour variables (Recency, Frequency and Monetary), using K-Means clustering. As a result, customers were grouped into 7 segments (5 sub-clusters of cluster 3 - the most relevant cluster, cluster 2 and cluster 1). To account for specific industry characteristics, AHP was used to assess the relative weights of the RFM variables, based on the Company's decision makers point of view. Thus, the clusters' Customer Lifetime Value was calculated as the weighted sum of normalized RFM values and used to assess the importance of each segment.

Additionally, a density-based customer clustering was performed to account for customers' geographic location, which may restrict the number of deliveries the Company can offer.

The results of the two clustering techniques were used to redefine the number of deliveries offered to each customer. All this was achieved permanently baring in mind the business perspective.

In the future, the new customer segmentation should be used to reconsider Customer Relationship Management strategies beyond the number of deliveries, such as the discounts offered to each customer. The best segments should receive premium treatment to maintain customer value.

Chapter 5

Distribution Network

Having defined a future service model for each customer segment in a given geographical cluster, it is now important to properly build a mathematical tool to design a delivery network that not only allows the satisfaction of all customers' orders but also minimizes the Company's transportation cost.

> "In today's changing and competitive industrial environment, the difference between ad hoc planning methods and those that use sophisticated mathematical models to determine an optimal course of action can determine whether or not a company survives." (Hoffman and Ralphs (2013))

This chapter describes the optimization model designed to calculate the optimal number of Dedicated vehicles and the assignment of customers to each car, as well as the results obtained.

5.1 Context

Before presenting the developed model, a few things must be considered. Looking at the typical Vehicle Routing Problem's variations, the one that comes closest to the Company's problem is the Vehicle Routing Problem (VRP) with Stochastic Customers. Even so, the Company's problem presents a key difference: as the Company outsources the Dedicated vehicles to logistics partners, it presents no control over the vehicles' routes. Apart from deciding which customers are assigned to each vehicle, it currently cannot impose any route for serving them. Since routing highly increases the problem's complexity and as the company has no control over the routes, it is was decided to lighten the problem, thus making it more computable and scalable.

Therefore, despite sharing many assumptions and concepts behind the Stochastic VRP, the proposed model resembles an allocation model in terms of problem formulation and objective: allocating customers to vehicles, not defining the route.

Additionally, by the time of the initial stages of the model's development, the entire world was living amidst unprecedented times. COVID-19 negatively impacted most economic sectors, including the Automotive Afterparts. Given such circumstances, the Company was forced to make cost reductions and decided to decrease the number of daily deliveries to 2 deliveries per day in Porto, Lisbon and Setúbal. All the other regions became automatically supplied by the Shared system only.

As things got back to normal, customers started to demand more deliveries and the Company decided that, starting from the 1st of July, it would increase Porto, Lisbon and Setúbal's offering to a maximum of 3 deliveries per day, depending on the customer segment and city. Considering the situation, the model was designed to find the solution that would minimize the total transportation cost for the upcoming months, as such an evaluation was never so critical.

Given the pressing nature of the problem, the model was designed to decide on the optimal number of Dedicated vehicles and respective customer allocation, without deciding on the optimal route (as this is currently decided by the logistics partners). Additionally, as the Shared system is unable to deliver 3 times per day, finding the best way of performing the 3 deliveries per day with the Dedicated system was the initial focus of the model.

5.2 Model

To solve the Company's allocation problem, an integer linear programming model (ILP) was designed that will not only assess the optimal number of Dedicated vehicles but also allocate each customer to one of the cars. All the variables considered and their bounds are presented next, as well as the restrictions used and the objective function of the problem. Here are some of the underlying assumptions.

Given the huge dimension of the problem, customers were grouped based on the first four digits of their postal code (exemplified in figure 5.1). By doing so, it is assumed that whenever a postal code is assigned to a certain vehicle, it becomes responsible for all customers that belong to that postal code.

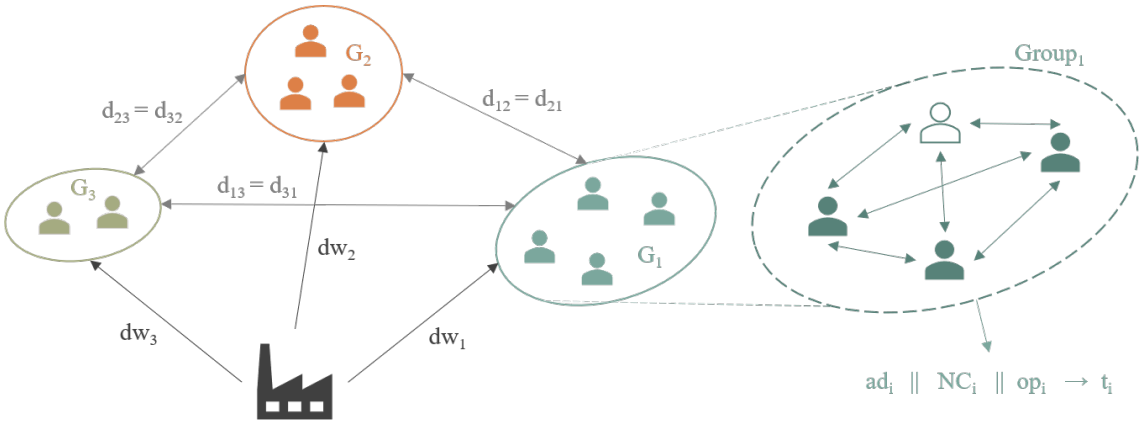

Figure 5.1: Grouping Representation

Additionally, as previously explained, customers do not order every day. To take this into consideration billing data (from July 2019 to February 2020) was analysed to find the actual number of customers that order per day. Figure 5.2 represents the histogram of the number of clients that order per day versus the percentage of days it represents. As it can be seen, the distribution curve is skewed to the right side, thus considering the average number of customers that order per day could imply not being able to fulfil many orders in high demand days. For that reason, it was decided to consider the demand (in number of customers with orders) that covers 95% of the days, insuring a high service level instead of considering the average demand of the data set. In this way, the Company only faces the risk of not being able to deliver in 5% of the days.

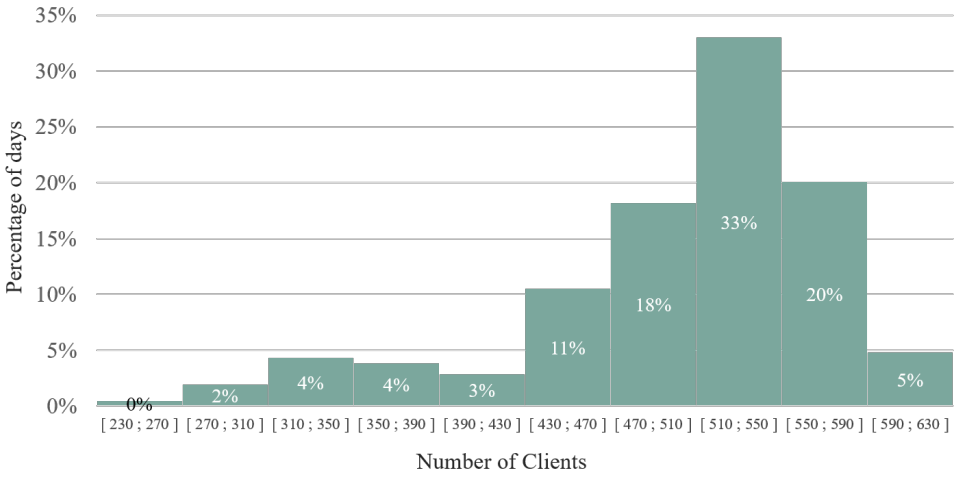

Figure 5.2: Histogram of the number of clients that order per day

Distances between customers, customer's groups and their respective warehouse were computed using Google's API and assumed to be symmetric. To know the vehicles' unloading time a car was followed for one day and the unloading times registered. The result was an average of 2 minutes per stop.

As seen in chapter 3, given the small volume of the products, capacity is not a constrain.

The objective of the problem is to determine the optimal customer allocation and the respective number of Dedicated vehicles for each Warehouse, which minimizes the total transportation cost while guaranteeing the best customer service.

5.2.1 Formulation

The problem formulation is as follows.

Index

I: set of customers' groups, $i \in I, k \in I$

J: set of Dedicated vehicles, $j \in J$

Parameters

NC_i: Number of customers in customers' group i

ad_i: Average distance between customers in customers' group i

op_i: Order probability for customers' group i

t_i: Time (minutes) to satisfy all clients belonging to customers' group i

ND_i: Total number of daily deliveries of customers' group i

d_{ik}: Travel time (minutes) between customers' groups i and k

dw_i: Travel time (minutes) between customers' group i and it's preferential warehouse

NG_j: Number of customers' groups assigned to vehicle j

TC_j: Average travel time (minutes) between customers' groups assigned to vehicle j

TW_j: Average travel time (minutes) between customers' groups assigned to vehicle j and their preferential warehouse

TI_j: Time (minutes) required to satisfy all customers' groups assigned to vehicle j

C_D: Daily cost (€) of renting a Dedicated vehicle

C_{KM}: Cost (€) per kilometer

CT: Cutoff available time (minutes)

M: Big M

Decision variables

$$V_j = \begin{cases} 1, & \text{if vehicle j is open} \\ 0, & \text{otherwise} \end{cases}$$

$$Y_{ij} = \begin{cases} 1, & \text{if customers' group i is assigned to vehicle j} \\ 0, & \text{otherwise} \end{cases}$$

$$Z_{ikj} = \begin{cases} 1, & \text{if customers' groups i and k are assigned to vehicle j} \\ 0, & \text{otherwise} \end{cases}$$

Objective Function

$$\min \quad C_D \cdot \sum_{j=1}^{J} V_j \quad + \quad C_{KM} \cdot (50/60) \cdot \sum_{j=1}^{J} TC_j \;/\; I \tag{5.1}$$

Constraints

$$\sum_{j=1}^{J} Y_{ij} = 1 \qquad \forall i \in I,\ j \in J \tag{5.2}$$

$$V_j * M^{1} \geq \sum_{i=1}^{I} Y_{ij} \qquad \forall j \in J \tag{5.3}$$

$$Z_{ikj} \geq Y_{ij} + Y_{kj} - 1 \qquad \forall i \in I,\ k \in I,\ j \in J,\ i \neq k \tag{5.4}$$

$$Z_{ikj} \leq Y_{ij} \qquad \forall i \in I,\ j \in J \tag{5.5}$$

$$Z_{ikj} \leq Y_{kj} \qquad \forall k \in I,\ j \in J \tag{5.6}$$

$$CT \geq 2 \cdot TW_j + TC_j + TI_j \qquad \forall j \in J \tag{5.7}$$

$$Y_{ij},\ Z_{ikj} \quad \text{are binary} \qquad \forall i \in I,\ k \in I,\ j \in J \tag{5.8}$$

where:

$$TW_j = \sum_{i=1}^{I} (tw_i \cdot Y_{ij}) \ / \ NG_j \qquad \forall i \in I,\ j \in J \tag{5.9}$$

$$TC_j = (NG_j - 1) \cdot (d_{ik} \cdot Z_{ikj}) \ / \ NG_j \qquad \forall i \in I,\ k \in I,\ j \in J \tag{5.10}$$

$$NG_j = \sum_{i=1}^{I} Y_{ij} \qquad \forall i \in I,\ j \in J \tag{5.11}$$

$$TI_j = \sum_{i=1}^{I} t_i \cdot Y_{ij} \qquad \forall i \in I,\ j \in J \tag{5.12}$$

$$t_i = op_i \cdot NC_i \cdot UnloadTime + (op_i \cdot NC_i - 1) \cdot ad_i \qquad \forall i \in I \tag{5.13}$$

The objective function (5.1) not only minimizes the cost of the Dedicated vehicles but also promotes the allocation of closer customers' groups in the same car. The first part of the function [$C_{\mathrm{D}} \cdot \sum_{j=1}^{J} V_{\mathrm{j}}$] promotes the minimization of the number of necessary vehicles, by considering the increase in the daily cost of having an additional car. The second part [$C_{\mathrm{KM}} \cdot (50/60) \cdot \sum_{j=1}^{J} TC_{\mathrm{j}} \ / \ I$] enhances the allocation of closer groups of customers to the same car. To do so the sum of the travel time between clusters is converted to distance - assuming an average speed of 50 km/h - and multiplied by the cost per kilometre[2].

Regarding the problem's restrictions, constrain 5.2 guarantees that all customer's groups are served by a Dedicated vehicle. Constrain 5.3 assures that if any group of customers i is assigned to a given Dedicated vehicle j, then vehicle j cost is taken into consideration. Constrains 5.4, 5.5, 5.6 ensure that the distances between any two customer's group i and k are only considered in vehicle's j average travel time between customer's groups ($\mathrm{TC_j}$) if both groups are allocated to this vehicle. Additionally, 5.7 assures that the total length of vehicle's j route is lower or equal to the

[1] For the proposed model, a value of M = 99999 is enough to ensure the constraints

[2] It was assumed a $\mathrm{C_{KM}}$ of 0.20€/km

time between two consecutive cutoffs, for a given service level. Finally, constrain 5.8 indicates that those variables are binary.

Equations 5.9-5.13 refer to additional calculated variables. Equation 5.9 calculates the average travel time to the warehouse for vehicle j, where NG_j is the number of customer groups assigned to this vehicle (equation 5.11). Equation 5.10 calculates the total travel time between customer groups assigned to vehicle j as the average distance between all customer groups $[(d_{ik} * Z_{ikj})/NG_j]$ times the number of travels between them (NG_j-1). Equation 5.12 calculates the time spent inside vehicle j's customer groups, where t_i has two components (equation 5.13): time spent unloading on each customer and time spent travelling between customers (similarly to the rationale behind TC).

5.2.2 Solution Approach

Since the company is already used to working with this program, the model described above was developed in Matlab (version R2020a) and solved using *intlinprog* solver, designed for mixed integer linear problems (MILP). The strategy behind this solver is as follows (The MathWorks (2016)):

1. Use preprocessing techniques to reduce the size of the problem by "eliminating redundant variables and constraints, improve the scaling of the model and sparsity of the constraint matrix, strengthen the bounds on variables, and detect the primal and dual infeasibility of the model" (The MathWorks (2016));
2. Solve the equivalent relaxed (noninteger) problem, that is, the linear programming problem (LP) of the MILP (same objective and constraints, but without the integer restrictions). In a minimization problem, the obtained value is the minimum among all feasible points;
3. Perform Mixed-Integer Program Preprocessing. This involves rapidly preexamining and eliminating a portion of the pointless subproblem candidates that branch-and-bound would otherwise consider.
4. Perform cuts - additional linear inequality constraints - that further restrict the admissible region of the LP relaxations so that their solutions are closer to integers.
5. Use heuristics to find integer-feasible solutions, to get an upper bound on the objective function.
6. Perform the Branch and Bound method to search systematically for the optimal solution. This algorithm develops a sequence of subproblems with updated bounds on the optimal objective function value, step-by-step converging to a solution of the MILP.

The model was run on a computer with an Intel(R)Core(TM)i1-6500 U CPU 2GZ processor, 8 GB installed physical memory (RAM), with a time limit of four hours.

5.3 Analysis & Results

Information of all the customers belonging to Porto, Lisbon and Setúbal was organized and, considering the new customer segmentation, the Company decided the maximum number of deliveries that would be offered in the upcoming months. Table 5.1 summarizes the results.

Table 5.1: Short term Maximum # Deliveries per Customer Segment

Customer	Short term Maximum # Deliveries
Big Spenders	3
Champions	3
Loyal	3
Potential Loyalists	2
Promising	2
Occasional	2
One Time Only	2

Once again, the geographical location presents a key factor for deciding the service level offering and the Company decision-makers evaluated the maximum possible number of deliveries per city. Table 1 (appendix 6) shows the lists of cities that can have up to 3 deliveries per day. All the other cities belonging to these regions can have a maximum of 2 deliveries per day.

Similarly to what was done in the last part of chapter 4, segmentation and geographical restrictions were crossed. The results are shown in table 5.2

Table 5.2: Short term Service Level for each Customer Segment, given City's Delivery Effort

	Delivery Effort	
Customer Segment	Low	High
Big Spenders	3	2
Champions	3	2
Loyal	3	2
Potential Loyalists	2	2
Promising	2	2
Occasional	2	2
One Time Only	2	2

Given current contractual conditions, the Company must offer 3 deliveries per day to some customers belonging to lower Customer Lifetime Value segments. Table 5.3 presents a summary of the number of customers and customers' groups per segment and region, that will be offered 3 deliveries/day. It should be noted that, as previously explained, it was assumed that customers were grouped based on the first four digits of their postal code. Each postal code corresponds to a customers' group.

Lastly, the model was run three times, once for each region, with the customers that will be offered 3 deliveries.

Figures 5.3, 5.4 and 5.5 correspond to the model's solution. Each point in the map represents a customers' group, painted in the same color as the vehicle to which it is allocated.

Table 5.3: Number of Customers and Customer's Groups with 3 Deliveries, per District

Segment	Porto	Lisbon	Setúbal
Big Spenders	1	1	0
Champions	24	36	23
Loyal	13	45	19
Potential Loyalists	13	44	14
Promising	17	37	26
Total #Customers	68	163	82
Total #Customers' Groups	25	57	22

As can be seen, the model proposed renting 5 for Setúbal (figure 5.3), 5 Dedicated vehicles for Porto (figure 5.4) and 11 for Lisbon (figure 5.5). Table 5.4 presents a summary of the vehicle's indicators [3]. Looking at the solutions' representation, it becomes obvious that the model is capable of optimizing the customer's groups allocation, as closer groups belong to the same vehicle. This can be verified looking at the average travel time between groups (TC_j), as the values were much higher when the model was run without the second part of the objective function.

Additionally, to test whether the number of vehicles is being minimized, the model was tested imposing a maximum number of Dedicated vehicles equal to *NVehicles - 1*, that is, minus one than the proposed solution. For Porto and Setúbal the model could not find any feasible solutions. Thus, it is assumed that the optimal number of vehicles was found. For Lisbon, as the time limit was reached before finding a feasible solution optimality can not be guaranteed. This has to do with the fact that Lisbon's problem is much more complex than the other two regions. As it considers much more customers' groups, the model could not test all possible combinations of customers' groups allocation to vehicles withing the time limit[4].

5.4 Conclusion

During this chapter, the optimization model designed was presented. Given the current extraordinary circumstances and time limitation, the model does not focus on fully optimizing the Company's Distribution Network. Instead, it seeks to solve the problem of deciding the optimal number of Dedicated vehicles and respective customer allocation.

The model proposed renting 21 Dedicated vehicles: 5 for Porto, 11 for Lisbon and 5 for Setúbal. The results proved to be quite promising and should be validated once the Company starts to increase the number of deliveries.

When compared to the Company's current situation, the outcome of the model proposes outsourcing three less vehicles per month. This is equivalent to a saving of 2139€ (average monthly price of one Dedicated vehicle) x 3 = 6417 €/month. This value corresponds to 11% of the total monthly Dedicated vehicles cost.

[3] $\text{Total Time}_j = TC_j + 2 \times TW_j + TI_j$; $\text{Slack}_j = 150 - \text{Total Time}_j$

[4] The three regions where tested with a time limit of four hours

Figure 5.3: Optimization Model solution for Setúbal

Table 5.4: Optimization Model Results

City	V_j	# Customers	NG_j	TC_j	TW_j	TI_j	Total $Time_j$	$Slack_j$
Porto	V_1	15	5	46	14	72	146	4
	V_2	10	5	51	36	40	126	24
	V_3	19	5	38	16	92	145	5
	V_4	17	6	56	22	69	147	3
	V_5	7	4	74	47	28	149	1
Lisbon	V_1	13	8	69	44	33	146	4
	V_2	13	7	55	17	22	95	55
	V_3	11	7	63	43	22	128	22
	V_4	17	8	55	44	25	125	25
	V_5	23	11	55	32	51	137	13
	V_6	21	13	67	23	50	140	10
	V_7	10	5	23	26	18	67	83
	V_8	21	10	59	36	44	139	11
	V_9	16	8	51	37	33	121	29
	V_{10}	8	5	67	59	15	141	9
	V_{11}	10	5	37	66	23	126	24
Setúbal	V_1	18	5	60	35	44	140	10
	V_2	25	5	38	33	78	149	1
	V_3	11	5	44	27	23	94	56
	V_4	13	3	54	45	36	135	15
	V_5	15	4	55	28	55	138	12

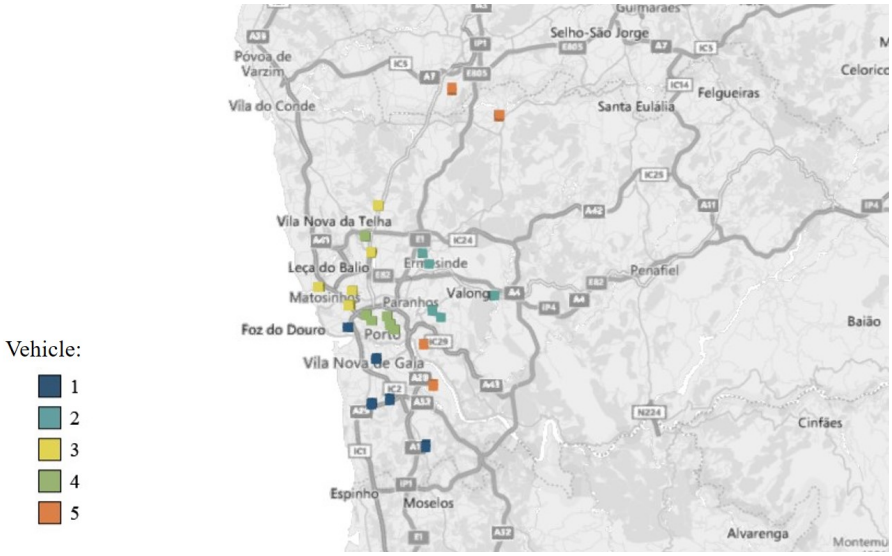

Figure 5.4: Optimization Model solution for Porto

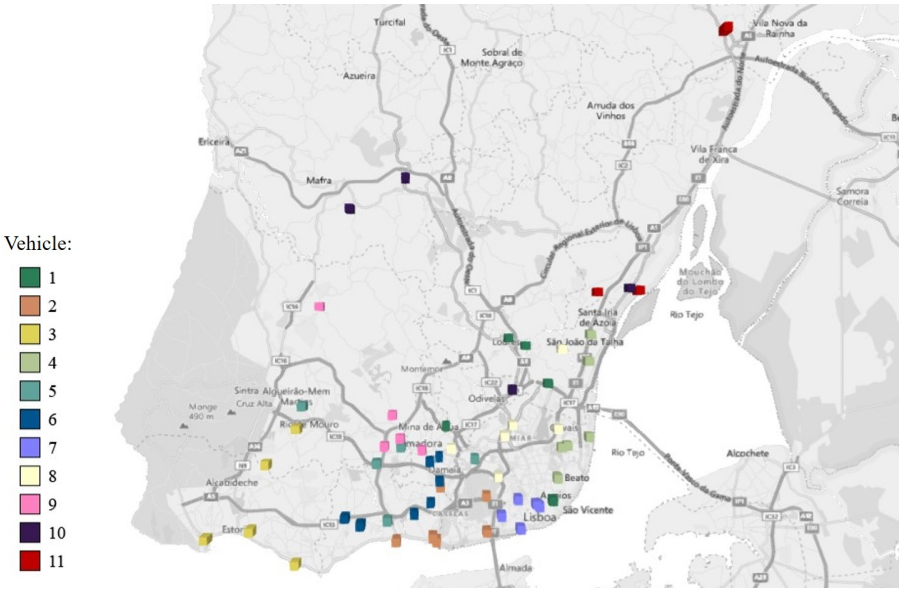

Figure 5.5: Optimization Model solution for Lisbon

Chapter 6

Conclusions and Future Work

As this work comes to an end, this chapter presents the most relevant conclusions of this thesis.

This study tackles two different research topics: Service Model Redesign and Distribution Network Optimization, applied in the specific context of an Automotive Aftermarket player. When combined, the worked developed on the two topics allowed for a complete redesign of the companies' Service Model.

For the first topic, the work was focused on redesigning the way the Company segments their customers through the application of data mining techniques. To the best of my knowledge, it combines an innovative way of considering partitioning and density-based clustering techniques for redefining a service level (in terms of number of deliveries per day) that not only takes into account the Customer Lifetime Value, but also the logistical complexities inherent to a real-life problem.

Regarding the Distribution Network, an optimization model was developed that supports companies in making decisions regarding the number of vehicles they should outsource and decide on the exact vehicle's customer allocation. This approach could be extended to companies with their own fleet.

When applied to the specific case study, the work developed allowed the design of strategies tailored to each of the seven customer segments, adapted to their needs and worth. In the future, it is expected that the present customer segmentation can provide a sales increase, if correctly used in Customer Relationship Management. This sales increase should be quantified once the new CRM strategies are put in place.

The optimization model provided a cost saving of 11% of the Dedicated vehicles' cost, vital for the Company's survival during this time of crisis. Furthermore, the Company now has a tool that provides them with the possibility of finding the number of vehicles they should outsource and correspondent allocation, in a flexible and autonomous way. Instead of taking such an impactful decision based on empirical knowledge, that could jeopardize customer service and the Company's profitability, it can now ensure a much more reliable delivery.

The objectives initially proposed of *redefining the number of deliveries and lead time per customer*, *calculating the number of vehicles required to satisfy the demand* and *defining the allocation of customers to the vehicles* are considered to have been achieved. The increase in customers satisfaction and loyalty can only be measured in the long term.

As for future work, some additional tasks remain to be done. Firstly, a deep analysis to the logistics partners' competences and comparison with the Company's needs should be performed. Reducing the number of partners could increase the Company's bargain power and the Company's control of the service level. Secondly, the stock management problem reported in chapter 3 should not be overlooked, and it should lead to a review of the stock management model.

On the more complex side, a few improvements for the proposed transportation model were identified. Recently, an analysis of the last months' transactions (September 2019 - January 2020), highlighted an unevenness in the number of orders per cutoff. Therefore, the possibility of uneven cutoff delivery time should be carefully assessed together with the Company's decision makers. Such a possibility could be incorporated in the model to evaluate whether it could provide additional cost savings.

Moreover, the designed optimization model could be extended to optimize the entire Distribution Network and define the most cost-effective way of implementing the Service Model designed in Chapter 4. This includes considering outsourcing Dedicated vehicles for customers currently served by the Shared system.

Lastly, this study could be evolved into a Vehicle Routing Problem with Stochastic Customers. Having control over their routes could immensely increase customer service level, as the Company would have higher control over their Service Model.

References

Aguilella, V., Aguilella-Arzo, M., Ramírez, P., Canon-Tapia, E., Walker, G., Herrero-Bervera, E., Gendreau, M., Laporte, G., and Séguin, R. Stochastic vehicle routing. *European Journal of Operational Research*, 88, 1996. doi: 10.1016/0377-2217(95)00050-X.

Amorim, P. and Almada-Lobo, B. The impact of food perishability issues in the vehicle routing problem. *Computers & Industrial Engineering*, 67:223–233, 01 2014. doi: 10.1016/j.cie.2013.11.006.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. Optics: Ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, 1999. ISSN 0163-5808. doi: 10.1145/304181.304187.

Barnhart, C. and Laporte, G. *Handbooks in Operations Research and Management Science: Transportation*. ISSN. Elsevier Science, 2006. ISBN 9780080467436.

Barnhart, C., Johnson, E., Nemhauser, G., Savelsbergh, M., and Vance, P. Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, 46:316–, 01 1998.

Bartholdi, J., Platzman, L., Collins, R., and Warden, W. A Minimal Technology Routing System for Meals on Wheels. *Interfaces*, 13:1–8, 1983. doi: 10.1287/inte.13.3.1.

Bektas, T. and Laporte, G. The Pollution-Routing Problem. *Transportation Research Part B: Methodological*, 45:1232–1250, 09 2011. doi: 10.1016/j.trb.2011.02.004.

Bellman, R. The theory of dynamic programming. *Bull. Amer. Math. Soc.*, 60(6):503–515, 11 1954. URL https://projecteuclid.org:443/euclid.bams/1183519147.

Benjamin, A. and Beasley, J. Metaheuristics for the waste collection vehicle routing problem with time windows, driver rest period and multiple disposal facilities. *Computers & OR*, 37:2270–2280, 12 2010. doi: 10.1016/j.cor.2010.03.019.

Bertsimas, D. A Vehicle Routing Problem with Stochastic Demand. *Operations Research*, 40:574–585, 1992. doi: 10.1287/opre.40.3.574.

Bholowalia, P. and Kumar, A. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 105:17–24, 2014.

Bult, J. and Wansbeek, T. Optimal Selection for Direct Mail. *Marketing Science*, 14:378–394, 1995. doi: 10.1287/mksc.14.4.378.

Carneiro, F. and Miguéis, V. Applying Data Mining Techniques and Analytical hierarchy Process to the Food Industry: Estimating Customer Lifetime Value. In *Proceedings of the International Conference on Industrial Engineering and Operations Management*. IEOM Society International, 2020. Accepted for publication.

Chen, D., Sain, S., and Guo, K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19, 2012. doi: 10.1057/dbm.2012.17.

Chen, W., Ji, M., and Wang, J. T-DBSCAN: A spatiotemporal density clustering for GPS trajectory segmentation. *International Journal of Online Engineering (iJOE)*, 10:19–24, 2014. doi: 10.3991/ijoe.v10i6.3881.

Cohen, M., Agrawal, N., and Agrawal, V. Winning in the Aftermarket. *Harvard Business Review*, 84:129–138, 05 2006.

Dantzig, G. and Ramser, R. The Truck Dispatching Problem. *Management Science*, 6:80–91, 10 1959. doi: 10.1287/mnsc.6.1.80.

Dror, M., Laporte, G., and Trudeau, P. Vehicle Routing with Stochastic Demands: Properties and Solution Frameworks. *Transportation Science*, 23, 08 1989. doi: 10.1287/trsc.23.3.166.

Eksioglu, B., Vural, A., and Reisman, A. The vehicle routing problem: A taxonomic review. *Computers & Industrial Engineering*, 57:1472–1483, 11 2009. doi: 10.1016/j.cie.2009.05.009.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

Evans, J. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole Publishing Company, 1996. ISBN 9780534231002.

Fan, T., Guo, N., and Ren, Y. Consumer clusters detection with geo-tagged social network data using DBSCAN algorithm: a case study of the Pearl River Delta in China. *GeoJournal*, 2019. doi: 10.1007/s10708-019-10072-8.

Goel, A. Vehicle Scheduling and Routing with Drivers' Working Hours. *Transportation Science*, 43:17–26, 02 2009. doi: 10.1287/trsc.1070.0226.

Goldschmidt, R. and Passos, E. *Data mining: um guia Prático*. Elsevier Editora, 2005. ISBN 9788535218770.

Gomory, R. Outline of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society*, 64:275–278, 09 1958. doi: 10.1090/S0002-9904-1958-10224-4.

Groër, C., Golden, B., and Wasil, E. The Consistent Vehicle Routing Problem. *Manufacturing & Service Operations Management*, 11:630–643, 10 2009. doi: 10.1287/msom.1080.0243.

Hahsler, M., Piekenbrock, M., and Doran, D. dbscan : Fast Density-Based Clustering with R. *Journal of Statistical Software*, 91:1–30, 10 2019. doi: 10.18637/jss.v091.i01.

Han, J., Pei, J., and Kamber, M. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011. ISBN 9780123814807.

Haugland, D., Ho, S., and Laporte, G. Designing delivery districts for the vehicle routing problem with stochastic demands. *European Journal of Operational Research*, 180:997–1010, 08 2007. doi: 10.1016/j.ejor.2005.11.070.

Hinneburg, A. and Keim, D. A. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 58–65. AAAI Press, 1998.

Hoffman, K. and Ralphs, T. Integer and Combinatorial Optimization. *Encyclopedia of Operations Research and Management Science*, 01 2013. doi: 10.1007/978-1-4419-1153-7_129.

Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Min. Knowl. Discov.*, 2:283–304, 1998. doi: 10.1023/A:1009769707641.

Hughes, A. Boosting response with RFM. *Marketing Tools*, 5:4–7, 01 1996.

Hvattum, L. M., Løkketangen, A., and Laporte, G. Solving a Dynamic and Stochastic Vehicle Routing Problem with a Sample Scenario Hedging Heuristic. *Transportation Science*, 40:421–438, 11 2006. doi: 10.1287/trsc.1060.0166.

Hwang, H., Jung, T., and Suh, E. An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications*, 26:181–188, 02 2004. doi: 10.1016/S0957-4174(03)00133-7.

Jourdan, L., Basseur, M., and Talbi, E.-G. Hybridizing exact methods and metaheuristics: A taxonomy. *European Journal of Operational Research*, 199:620–629, 12 2009. doi: 10.1016/j.ejor.2007.07.035.

Kao, Y.-T., Wu, H.-H., Hsuan-Kai, C., and Chang, E.-C. K. A case study of applying LRFM model and clustering techniques to evaluate customer values. *Journal of Statistics and Management Systems*, 14:267–276, 2013. doi: 10.1080/09720510.2011.10701555.

Kaufman, L. and Rousseeuw, P. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470317488.

Khajvand, M. and Tarokh, M. J. Analyzing customer segmentation based on customer value components (case study: A private bank) (technical note). *Advances in Industrial Engineering*, 45(Special Issue):79–93, 2011. ISSN 2423-6896.

Khajvand, M., Zolfaghar, K., Ashoori, S., and Alizadeh, S. Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3:57 – 63, 2011. ISSN 1877-0509. doi: 10.1016/j.procs.2010.12.011.

Kim, H., Yang, J., and Lee, K.-D. Vehicle routing in reverse logistics for recycling end-of-life consumer electronic goods in South Korea. *Transportation Research Part D: Transport and Environment*, 14:291–299, 07 2009. doi: 10.1016/j.trd.2009.03.001.

Kotler, P. and Armstrong, G. *Principles of Marketing*. Pearson Prentice Hall, 2006. ISBN 9780131469181.

Larson, R. Transporting Sludge to the 106Mile Site: An Inventory/Routing Model for Fleet Sizing and Logistics System Design. *Transportation Science*, 22:186–198, 08 1988. doi: 10.1287/trsc.22.3.186.

Lawler, E. L. and Wood, D. E. Branch-and-bound methods: A survey. *Operations Research*, 14 (4):699–719, 1966. doi: 10.1287/opre.14.4.699. URL https://doi.org/10.1287/opre.14.4.699.

Lawson, R. and Jurs, P. New Index for Clustering Tendency and Its Application to Chemical Problems. *Journal of Chemical Information and Computer Sciences*, 30:36–41, 02 1990. doi: 10.1021/ci00065a010.

Lei, H., Laporte, G., and Guo, B. Districting for routing with stochastic customers. *EURO Journal on Transportation and Logistics*, 1:67–85, 06 2012. doi: 10.1007/s13676-012-0005-x.

Li, D.-C., Dai, W.-L., and Tseng, W.-T. A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business. *Expert Syst. Appl.*, 38:7186–7191, 2011. doi: 10.1016/j.eswa.2010.12.041.

Liao, C.-J., Lin, Y., and Shih, S. Vehicle routing with cross-docking in the supply chain. *Expert Syst. Appl.*, 37:6868–6873, 10 2010. doi: 10.1016/j.eswa.2010.03.035.

Liu, D. and Shih, Y. Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management*, 42(3):387–400, 2005a. doi: https://doi.org/10.1016/j.im.2004.01.008.

Liu, D.-R. and Shih, Y.-Y. Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. *Journal of Systems and Software*, 77:181–191, 08 2005b. doi: 10.1016/j.jss.2004.08.031.

Maini, R. and Goel, R. Vehicle routing problem and its solution methodologies: a survey. *International Journal of Logistics Systems and Management*, 28:419, 01 2017. doi: 10.1504/IJLSM.2017.10008188.

Maskan, B. H. H. Proposing a Model for Customer Segmentation using WRFM Analysis (Case Study: an ISP Company) - TI Journals. *International Journal of Economy, Management and Social Sciences*, 2014.

Nimbalkar, D. and Shah, P. Data mining using RFM Analysis. *International Journal of Scientific & Engineering Research*, 4:940–943, 2013. doi: 10.13140/RG.2.2.24229.04328.

Payne, A. and Frow, P. A Strategic Framework for Customer Relationship Management. *Journal of Marketing*, 69(4):167–176, 2005. doi: 10.1509/jmkg.2005.69.4.167. URL https://doi.org/10.1509/jmkg.2005.69.4.167.

Robinson, A. 6 Competitive Advantages from Effective Transportation Management for the Automotive Aftermarket Supply Chain (And Any Shipper). Available at https://cerasis.com/automotive-aftermarket-supply-chain/, n.a. Accessed: 2020-04-19.

Saaty, T. A Scaling Method for Priorities in Hierarchical Structures. *Journal of Mathematical Psychology*, 15:234–281, 1977. doi: 10.1016/0022-2496(77)90033-5.

Saaty, T. *Fundamentals of Decision Making and Priority Theory With the Analytic Hierarchy Process*. AHP series. RWS Publications, 2000. ISBN 9781888603156.

Saraiva, A. R. M. Aplicação de Marketing Analítico. Master's thesis, Católica Porto Business School, Porto, 2018.

Shearer, C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehouse*, 5: 13–22, 2000.

Shih, Y.-Y. and Liu, C.-Y. A method for customer lifetime value ranking — Combining the analytic hierarchy process and clustering analysis. *The Journal of Database Marketing & Customer Strategy Management*, 11:159–172, 2003. doi: 10.1057/palgrave.dbm.3240216.

Taher, N., Elzanfaly, D., and Salama, S. Investigation in Customer Value Segmentation Quality under Different Preprocessing Types of RFM Attributes. *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, 4:5, 2016. doi: 10.3991/ijes.v4i4.6532.

Tan, P., Steinbach, M., and Kumar, V. *Introduction to Data Mining: Pearson New International Edition*. Pearson Education Limited, 2013. ISBN 9781292038551.

The MathWorks, I. Optimization toolbox^{TM} user's guide, 2016.

Toth, P. and Vigo, D. *The Vehicle Routing Problem*. Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics, 2002. ISBN 9780898715798.

U.S. Department of Commerce. U.s. automotive parts industry annual assessment. Available at https://partschecklive.files.wordpress.com/2012/08/oem-parts-supplie-pressures.pdf, 2009. Accessed: 2020-03-05.

Wei, J.-T., Lin, S.-Y., Weng, C.-C., and Wu, H.-H. A case study of applying LRFM model in market segmentation of a children's dental clinic. *Expert Syst. Appl.*, 39:5529–5533, 04 2012. doi: 10.1016/j.eswa.2011.11.066.

Appendix

Table 1: List of cities eligible for 3 deliveries per day

District	City
Lisbon	Lisboa
Lisbon	Vila Franca de Xira
Lisbon	Odivelas
Lisbon	Loures
Lisbon	Cascais
Lisbon	Sintra
Lisbon	Oeiras
Lisbon	Mafra
Lisbon	Alcabideche
Lisbon	Marvila
Lisbon	Seixal
Lisbon	Amadora
Lisbon	São Jorge de Arroios
Lisbon	Alenquer
Lisbon	Lumiar
Porto	Porto
Porto	Vila Nova de Gaia
Porto	Gondomar
Porto	Maia
Porto	Valongo
Porto	Matosinhos
Porto	Santo Tirso
Setúbal	Almada
Setúbal	Barreiro
Setúbal	Moita
Setúbal	Seixal
Setúbal	Palmela
Setúbal	Sesimbra
Setúbal	Setúbal
Setúbal	Montijo
Setúbal	Alcochete